Future Trends in Microelectronics

Reflections on the Road to Nanotechnology

edited by

Serge Luryi
Department of Electrical Engineering,
State University of New York,
Stony Brook, NY, U.S.A.

Jimmy Xu
Department of Electrical & Computer Engineering,
University of Toronto,
Toronto, Ontario, Canada

and

Alex Zaslavsky
Division of Engineering,
Brown University,
Providence, RI, U.S.A.

Kluwer Academic Publishers

Dordrecht / Boston / London

Published in cooperation with NATO Scientific Affairs Division

Proceedings of the NATO Advanced Research Workshop on
Future Trends in Microelectronics: Reflections on the Road to Nanotechnology
Ile de Bendor, France
July 17–21, 1995

A C.I.P. Catalogue record for this book is available from the Library of Congress

ISBN 0-7923-4169-4

Published by Kluwer Academic Publishers,
P.O. Box 17, 3300 AA Dordrecht, The Netherlands.

Kluwer Academic Publishers incorporates the publishing programmes of
D. Reidel, Martinus Nijhoff, Dr W. Junk and MTP Press.

Sold and distributed in the U.S.A. and Canada
by Kluwer Academic Publishers,
101 Philip Drive, Norwell, MA 02061, U.S.A.

In all other countries, sold and distributed
by Kluwer Academic Publishers Group,
P.O. Box 322, 3300 AH Dordrecht, The Netherlands.

Printed on acid-free paper

Printed in the Netherlands

MASTER COPY KEEP FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	Nov 96	Final

4. TITLE AND SUBTITLE

Future Treands in Microelectronics. Reflections on the Road to Nanotechnology

5. FUNDING NUMBERS

DAAH04-95-1-0081

6. AUTHOR(S)

Serge Luryi (principal investigator)

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)

State Univ of New York at Stony Brook
Stony Brook, NY 11790

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park,, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARO 34402.1-EL-CF

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

ABSTRACT NOT AVAILABLE

14. SUBJECT TERMS

15. NUMBER IF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as; prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NORFORN, REL, ITAR).

- **DOD** - See DoDD 4230.25, "Distribution Statements on Technical Documents."
- **DOE** - See authorities.
- **NASA** - See Handbook NHB 2200.2.
- **NTIS** - Leave blank.

Block 12b. Distribution Code.

- **DOD** - Leave blank
- **DOE** - Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports
- **NASA** - Leave blank.
- **NTIS** - Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only).*

Block 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.

Future Trends in Microelectronics

NATO ASI Series

Advanced Science Institutes Series

A Series presenting the results of activities sponsored by the NATO Science Committee, which aims at the dissemination of advanced scientific and technological knowledge, with a view to strengthening links between scientific communities.

The Series is published by an international board of publishers in conjunction with the NATO Scientific Affairs Division

A Life Sciences **B Physics**	Plenum Publishing Corporation London and New York
C Mathematical and Physical Sciences **D Behavioural and Social Sciences** **E Applied Sciences**	Kluwer Academic Publishers Dordrecht, Boston and London
F Computer and Systems Sciences **G Ecological Sciences** **H Cell Biology** **I Global Environmental Change**	Springer-Verlag Berlin, Heidelberg, New York, London, Paris and Tokyo

PARTNERSHIP SUB-SERIES

1. Disarmament Technologies	Kluwer Academic Publishers
2. Environment	Springer-Verlag / Kluwer Academic Publishers
3. High Technology	Kluwer Academic Publishers
4. Science and Technology Policy	Kluwer Academic Publishers
5. Computer Networking	Kluwer Academic Publishers

The Partnership Sub-Series incorporates activities undertaken in collaboration with NATO's Cooperation Partners, the countries of the CIS and Central and Eastern Europe, in Priority Areas of concern to those countries.

NATO-PCO-DATA BASE

The electronic index to the NATO ASI Series provides full bibliographical references (with keywords and/or abstracts) to more than 50000 contributions from international scientists published in all sections of the NATO ASI Series.
Access to the NATO-PCO-DATA BASE is possible in two ways:

- via online FILE 128 (NATO-PCO-DATA BASE) hosted by ESRIN,
Via Galileo Galilei, I-00044 Frascati, Italy.

- via CD-ROM "NATO-PCO-DATA BASE" with user-friendly retrieval software in English, French and German (© WTV GmbH and DATAWARE Technologies Inc. 1989).

The CD-ROM can be ordered through any member of the Board of Publishers or through NATO-PCO, Overijse, Belgium.

Series E: Applied Sciences - Vol. 323

This book contains the proceedings of a NATO Advanced Research Workshop held within the programme of activities of the NATO Special Programme on *Nanoscale Science* as part of the activities of the NATO Science Committee.

Other books previously published as a result of the activities of the Special Programme are:

NASTASI, M., PARKING, D.M. and GLEITER, H. (eds.), *Mechanical Properties and Deformation Behavior of Materials Having Ultra-Fine Microstructures.* (E233) 1993 ISBN 0-7923-2195-2

VU THIEN BINH, GARCIA, N. and DRANSFELD, K. (eds.), *Nanosources and Manipulation of Atoms under High Fields and Temperatures: Applications.* (E235) 1993 ISBN 0-7923-2266-5

LEBURTON, J.-P., PASCUAL, J. and SOTOMAYOR TORRES, C. (eds.), *Phonons in Semiconductor Nanostructures.* (E236) 1993 ISBN 0-7923-2277-0

AVOURIS, P. (ed.), *Atomic and Nanometer-Scale Modification of Materials: Fundamentals and Applications.* (E239) 1993 ISBN 0-7923-2334-3

BLÖCHL, P. E., JOACHIM, C. and FISHER, A. J. (eds.), *Computations for the Nano-Scale.* (E240) 1993 ISBN 0-7923-2360-2

POHL, D. W. and COURJON, D. (eds.), *Near Field Optics.* (E242) 1993 ISBN 0-7923-2394-7

SALEMINK, H. W. M. and PASHLEY, M. D. (eds.), *Semiconductor Interfaces at the Sub-Nanometer Scale.* (E243) 1993 ISBN 0-7923-2397-1

BENSAHEL, D. C., CANHAM, L. T. and OSSICINI, S. (eds.), *Optical Properties of Low Dimensional Silicon Structures.* (E244) 1993 ISBN 0-7923-2446-3

HERNANDO, A. (ed.), *Nanomagnetism* (E247) 1993. ISBN 0-7923-2485-4

LOCKWOOD, D.J. and PINCZUK, A. (eds.), *Optical Phenomena in Semiconductor Structures of Reduced Dimensions* (E248) 1993. ISBN 0-7923-2512-5

GENTILI, M., GIOVANNELLA, C. and SELCI, S. (eds.), *Nanolithography: A Borderland Between STM, EB, IB, and X-Ray Lithographies* (E264) 1994. ISBN 0-7923-2794-2

GÜNTHERODT, H.-J., ANSELMETTI, D. and MEYER, E. (eds.), *Forces in Scanning Probe Methods* (E286) 1995. ISBN 0-7923-3406-X

GEWIRTH, A.A. and SIEGENTHALER, H. (eds.), *Nanoscale Probes of the Solid/Liquid Interface* (E288) 1995. ISBN 0-7923-3454-X

CERDEIRA, H.A., KRAMER, B. and SCHÖN, G. (eds.), *Quantum Dynamics of Submicron Structures* (E291) 1995. ISBN 0-7923-3469-8

WELLAND, M.E. and GIMZEWSKI, J.K. (eds.), *Ultimate Limits of Fabrication and Measurement* (E292) 1995. ISBN 0-7923-3504-X

EBERL, K., PETROFF, P.M. and DEMEESTER, P. (eds.), *Low Dimensional Structures Prepared by Epitaxial Growth or Regrowth on Patterned Substrates* (E298) 1995. ISBN 0-7923-3679-8

MARTI, O. and MÖLLER, R. (eds.), *Photons and Local Probes* (E300) 1995. ISBN 0-7923-3709-3

GUNTHER, L. and BARBARA, B. (eds.), *Quantum Tunneling of Magnetization - QTM '94* (E301) 1995. ISBN 0-7923-3775-1

PERSSON, B.N.J. and TOSATTI, E. (eds.), *Physics of Sliding Friction* (E311) 1996. ISBN 0-7923-3935-5

MARTIN, T.P. (ed.), *Large Clusters of Atoms and Molecules* (E313) 1996. ISBN 0-7923-3937-1

DUCLOY, M. and BLOCH, D. (eds.), *Quantum Optics of Confined Systems* (E314). 1996. ISBN 0-7923-3974-6

ANDREONI, W. (ed.), *The Chemical Physics of Fullereness 10 (and 5) Years Later. The Far-Reaching Impact of the Discovery of C_{60}* (E316). 1996. ISBN 0-7923-4000-0

NIETO-VESPERINAS, M. and GARCIA, N. (Eds.): *Optics at the Nanometer Scale: Imaging and Storing with Photonic Near Fields* (E319). 1996. ISBN 0-7923-4020-5

LURYI, S., XU, J. and ZASLAVSKY, A. (Eds.): *Future Trends in Microelectronics: Reflections on the Road to Nanotechnology* (E323). 1996. ISBN 0-7923-4169-4

RARITY, J. and WEISBUCH, C. (Eds.): *Microcavities and Photonic Bandgaps: Physics and Applications* (E324). 1996. ISBN 0-7923-4170-8

CONTENTS

USLI MICROELECTRONICS: CHALLENGES AND FUTURE DIRECTIONS

SYSTEM AND ARCHITECTURE EVOLUTIONS AND DEVICE LIMITATIONS

NANO AND QUANTUM ELECTRONICS

SIMULATIONS AND MODELING

NEW MATERIALS AND DEVICE TECHNOLOGIES

SYSTEMS AND CIRCUITS

Preface

Ever since the invention of the transistor and especially after the advent of integrated circuits, semiconductor devices have kept expanding their role in our life. For better or worse, our civilization is destined to be based on semiconductors.

The microelectronics industry is now at a crossroad; the hardware side of microelectronics - that which concerns devices and technologies - is going through breathtaking ups and downs. We are at a turning point in the logical evolution of the giant VLSI industry, which, of course, is and will remain the dominant force in microelectronics. The celebrated Si technology has known a virtually one-dimensional path of development: reducing the minimal size of lithographic features. In the meantime, the investment in manufacturing facilities has doubled from generation to generation. There is a widespread fear that this path has taken us to the point of diminishing return. This fear has slowed the pace of new hardware technology development and encouraged investment in software and circuit design within existing technologies.

There is no shortage of opinion about what is and will be happening in our profession. Some, haunted by the specter of steel industry, believe that the semiconductor microelectronics industry has matured and the research game is over. Others believe the progress in hardware technology will come back roaring, based on innovative research.

Identifying the scenarios for the future evolution of microelectronics is the key to constructive action today.

Perhaps this can be dismissed as "fortune telling" or, at best, viewed as a high risk undertaking. Indeed, prediction is hard to make, especially when it is about the future. And, too often did forecasts by well-informed and authoritative sources prove wildly wrong. A few examples can be readily found in the field of computers - the most valued "customer" of microelectronics:

"I think there is a world market for maybe five computers."

- Thomas Watson, chairman of IBM, 1943.

"Computers in the future may weigh no more than 1.5 tons."

- Popular Mechanics, 1949.

"I have traveled the length and breadth of this country and talked with the best people, and I can assure you that data processing is a fad that won't last out the year."

- An editor for Prentice Hall, 1957.

"There is no reason anyone would want a computer in their home."

- Ken Olson, chairman of Digital Equipment Corp., 1977.

However, advocates for new approaches and optimists of departure from the existing path of established technologies usually fare no better. Critics of adventurous new approaches, though not often cited, are frequently in the right. Even worse, we have all seen superior technologies fail for reasons entirely unrelated to technical merits...

Still, as we shed our illusions, we can not afford actions without vision.

What is needed is critical assessment of where we are, what lies ahead, where new opportunities and/or alternative paths might be and what the limiting factors are...

It is in this spirit that we organized the NATO Advanced Research Workshop on "Future Trends in Microelectronics - Reflections on the road to nanotechnologies", which took place at the Ile de Bendor, France, July 17-21, 1995. The main purpose of the Workshop was to provide a rare forum for the gathering of leading professionals in industry, government and academia; and to promote a free-spirited debate on the future of microelectronics, to discuss recent developments, to identify the main road blocks and to explore future opportunities. The main topics of discussion were:

- What is the technical limit to shrinking devices? Is there an economic sense in pursuing this limit? In the memory market? In the microprocessor market?
- Review of nanoelectronics. Where is it heading? Are quantum-effect devices useful? Is mass production of nanodevices technologically and economically feasible?
- Are there green pastures beyond the traditional semiconductor technologies? What can we expect from combinations with superconducting circuits? Molecular devices? Polymers?
- What kind of research does the silicon industry need to continue its expansion? What are the anticipated trends in lithography? Modeling? Materials? Can we expect a "display revolution"? Will wide-area electronics be integrated with VLSI?
- What are the limits to thin film transistors? Do we need three-dimensional integration? SOI?

- To what extent can we trade high speed for low power? Is adiabatic computing in the cards?

- Is there a need for (possibility of) integrating compound semiconductor IC's into Si VLSI? What are the merits and prospects of hybrid schemes, such as heteroepitaxy and packaging? What are the most attractive system applications of optoelectronic hybrids?

- What are the possible implications of opto-electrical-microwave interactions? Are on-chip phased-array antenna systems feasible? Desirable? What can we expect from photonic bandgap structures?

- What is happening in system and architecture evolutions (or revolution)?

- Review of the recent progress in widegap semiconductor technologies, electronic and photonic. What are current problems and ultimate goals in optical disk memories? Automotive electronics? Other potential markets?

- What is happening in narrow gap semiconductors? Are intersubband devices a viable alternative? What are the potential applications of the unipolar laser?

The format of the Workshop included prepared invited presentations, ad hoc contributions and uninhibited exchange of views and rebuttals, in an attempt to reach some consensus on these critical issues. Many dominant figures of our profession with pioneering contributions to their credit came to share their opinions and to lead the discussions. In keeping with our goal of providing a forum for promoting free-spirited exchange and debate of ideas over the five days of the Workshop, each day started with formal presentations by key speakers on subjects in one or two chosen themes, and concluded with an evening panel session that began with two lead (and intentionally provocative) presentations followed by debates among the five panelists and the audience. The oral presentations, discussions and debates were complemented by afternoon poster sessions.

This book is a result of this exercise. It is a reflection of the issues and views debated at the workshop and a summary of the technical assessments and results presented.

No holds were barred at the Workshop, however understandably, and perhaps also wisely, some of the participants elected not to venture all of their opinions in writing.

The Workshop was organized by a committee which, in addition to the co-editors, included Francois Arnaud d'Avitaya, Jacques Derrien, Sergei Gurevich, Hans Rupprecht, and Claude Weisbuch. Much helpful advice was provided ex-officio by Alec Broers, Gerald Iafrate, Jan Slotboom, Klaus von Klitzing and Michel Voos. Alix Arnaud d'Avitaya and Sandra Craig-Hallam had the all important tasks of local

arrangement and administrative support. In addition to NATO, the Workshop was sponsored by a number of agencies, including ARO (US), DRET (France), ONR (US), PHANTOMS Network (EU), Motorola (US), Nortel (Canada), and the US DOD European Office.

For all those who came together to share ideas, and for all the prospective readers, we hope that this publication will serve as a useful reference and a springboard for new ideas.

Long Island, New York, USA — Serge Luryi

Toronto, Canada — Jimmy Xu

Providence, Rhode Island, USA — Alex Zaslavsky

ALL THAT GLITTERS ISN'T SILICON

Or "Steel and Aluminum Re-Visited"

HERBERT KROEMER
ECE Department, University of California
Santa Barbara, CA 93106, USA

1. Introduction

At the 1974 International Electron Devices Meeting (IEDM), Marty Lepselter gave an invited talk under the title "*Integrated circuits – The New Steel*" [1]. His message was that the emerging integrated circuit technology was likely to play the same central role in the industrial revolution of the late-20th century that steel played in the great industrial revolution of the early-19th century.

I have always found this an extraordinarily apt analogy, and it has long been my conviction that this analogy can be extended to an important general analogy between *structural* metallurgy in general and *electronic* metallurgy in general [2].

In structural metallurgy, steel has, for some two centuries, been the dominant structural metal—and is likely to remain so for the foreseeable future. Without steel, we would have no modern industrial society. But we also would not have such a society if we relied on steel alone. Modern society depends vitally on the diversity contributed by such additional materials as aluminum, magnesium, titanium, etc. While we continue to build automobiles and ships and other "heavy goods" (mostly) from steel, if steel were the only structural metal available, we would still build airplanes from wood, and we would not build spacecraft (and communication satellites) at all.

Similarly, in electronic metallurgy, Silicon is, without ant doubt, the dominant electronic material—and is likely to remain so. But a mature electronic technology, too, calls for a great diversity, more than can be provided by Si technology alone. Enter other materials, such as GaAs and beyond.

Structural metallurgists divide metallurgy into ferrous and nonferrous metallurgy. The analogy to Si and compound semiconductor technology is obvious. In a very real

S. Luryi et al. (eds.), Future Trends in Microelectronics, 1–12.

sense, GaAs may be viewed as the Aluminum of electronic metallurgy—or should I say: Aluminum is the GaAs of structural metallurgy?

I believe that, by taking this analogy between structural and electronic metallurgy seriously, we can get a good insight where exactly non-Si technology fits in with Si, and what the future role of non-Si technology is likely to be. This is the central theme of my presentation.

2. New Device Concepts: Where Do They Fit In?

2.1. FROM THE NAYSAYERS' DICTIONARY

Anybody involved in new device concepts inevitably encounters a number of counter-arguments from what I call the *Naysayers' Dictionary*, with the three most common arguments listed here in order of increasing deviousness.

"It can't be done"

New device concepts often call for new technology, and the first argument against them is: There is no technology that would be able to do this. This actually tends to be true—and largely irrelevant. History shows that, given a strong incentive (and a sound physical basis for the concept), technologies tend to develop to meet the needs. Be suspicious of claims that some current difficulties are "fundamental" and hence cannot be overcome. Historically, many such supposedly fundamental difficulties were not so fundamental after all (or even were blessings in disguise); if they are solvable, the solutions tend to be discovered by those actually working on the problems, rather than by the naysayers.

If you make a strong case that the technology can and should be developed, the naysayers' argument shifts:

"It can't compete with Silicon"

Probably true—and indeed deadly if your application is one that can already be done with silicon. But suppose it isn't. Then you will very likely encounter the ultimate defense:

"There are no applications for this"

It is the most insidious and fallacious argument of them all, and much of my presentation deals with this issue. The negative claim actually flies in the face of historical experience: Historically, almost any sufficiently new and innovative technology always

has created economically viable new applications that draw on the new technology. I like to express my own counter-argument in the form of a "lemma":

2.2. THE "LEMMA OF NEW TECHNOLOGY"

I claim the following:

> The principal applications of any sufficiently new and innovative technology always have been — and will continue to be — applications *created* by that technology.

The use of the new technology to obtain merely better *quantitative* improvements in applications for which a technology already exists always has been—and will continue to be—a secondary consequence of the success of the new technology in *new* applications, usually as a result of cost reductions brought about by the extensive use of the new technology in the new applications.

The pattern of new science creating new devices that create their own applications is likely to continue well into the next century.

2.3. EXAMPLES

2.3.1. *The Transistor*

Perhaps the most important example of this central historical lesson is the transistor itself. Initially viewed simply as a replacement for electron tubes, it ultimately *created* the modern computer and the new industrial revolution that followed it.. I have been told that portable radios and hearing aids were also cases of new applications that preceded IC's and computers, and they played an important role in getting semiconductor technology started. True, but those earlier applications could never have formed the basis for a true industrial revolution.

2.3.2. *The Semiconductor Laser*

Another example of a device creating its own application was the double-heterostructure laser. Having been the originator of this concept [3, 4], I recall painfully that I was told in 1963 that there was no point in developing a technology for this new concept, because this device would *never* become useful, because of its the low anticipated power and a relatively poor spectral purity. If those skeptics had been right, we would today not have optical fiber communications, nor compact discs. In fact, the optoelectronics that developed in the wake of the DH laser is likely to be one of the "driving engines" for device development well into the next century.

2.3.3. *The HEMT*

A third example—of a different kind—is the HEMT. Although probably of lesser importance than my two examples above, I like to mention it for some other lessons it contains. It was initially widely hailed as a device for high-speed RAM's. If everything else had been the same, the higher mobilities in GaAs would have given it a considerable speed advantages over Si-RAM's. But everything else just wasn't the same, and GaAs-HEMT's never could compete with RAM's based on Si-CMOS—ultimately not even on speed. In terms of our analogy to structural metallurgy, it was as ill-guided an attempt as the use of aluminum in the superstructure of warships (remember the *Sheffield*?). What happened instead was that HEMT's turned out to be superb low-noise devices for the direct reception of TV signals from satellites, practically creating the industry of those small (if ugly) dishes seen outside many windows worldwide. It would in principle have been possible to do that with Si-FET's, but the better noise performance of HEMT's permitted the use of much smaller dishes, and this created a large economic *leverage* that more than made up for the higher cost of the FET itself—not to mention the much better customer acceptance of the smaller dishes. Please keep that concept of *leverage* in mind; I will return to it shortly.

My list of examples could easily be extended *ad nauseam*, but I think I have made my case.

2.4. LESSONS

If it is indeed true that the principal applications of any sufficiently new and innovative technology will be applications *created* by that new technology, then this has the far-reaching consequence that all of us must take a long-term look when judging the potential of any new technology:

2.4.1. *How NOT to Judge New Technology*

New technology evidently must not be judged simply by how it might fit into already existing applications, where the new discovery may have little chance to be used in the face of competition with already-existing and entrenched technology. And it must not be dismissed on the grounds that it has no realistic existing applications. Such actions only stifle progress towards those applications that *will* grow out of that technology.

None of this is intended to relieve the researcher of the obligation to *look* for near-term applications of his/her research, and if *credible* near-term applications can indeed be identified, so much the better. But often that will not be the case. In this event it should be made a part of the researcher's obligation to consider what kind of totally new applications might be created by the research. This is may be harder than simply

trying to squeeze everything into existing applications, and more often than not it will not succeed—or run the risk of sliding off into irresponsible science fiction. In fact, experience shows, that, more often than not, the applications of a new research discovery are found by someone other than the original researcher, and this is likely to remain so. Nevertheless, we should at least try.

Quite frankly, I do not think we can realistically predict which new devices and applications may emerge, but I believe we can create an environment encouraging progress, by not always asking immediately what any new science might be good for (and cutting off the funds if no answer full of fanciful promises is forthcoming—a worldwide problem. We must make it an acceptable answer to the quest for applications if the researcher has sincerely tried to identify credible applications—near-term or long-term—but has failed to do so.

What is *never* acceptable—and what researchers must refrain from doing—are attempts to justify the research by promising credibility-stretching mythical improvements in *existing* applications. Most such claims are not likely to be realistic, are easily refuted, and only discredit the research they were intended to justify.

2.4.2. *A Fable: Friedrich Wöhler and the Discovery of Aluminum*

Let me illustrate my homily with a little fable. In the year 1827 it came to pass that the German chemist Friedrich Wöhler published the discovery of a new element, which he called alum-in-ium, because he had extracted it from the mineral alum. Flushed by his triumph, he applied for a grant for follow-up research. But his application was turned down on the grounds that the new material had no conceivable applications: Being far too soft, with little structural strength, and it oxidizing and corroding like crazy, it could not possibly ever compete with steel. Wöhler was downcast, but a fairy sent him a dream, and the next day he called his funding agency that aluminum would be the metal from which "aircraft" would be built (he knew better than to admit it was just a dream). "Aircraft, what's that?" was the reply. "Well, you know, flying machines, things in which people can fly." Now, in 1827, the closest thing known to a flying machine was a hot-air balloon; so Wöhler's proposal was re-evaluated for its promise of great progress in hot-air balloon technology: one could build the balloons from thinly rolled-out aluminum sheet metal. This would have two advantages: (a) Aluminum was less likely to catch fire than the balloon fabrics used at the time, an important issue with hot-air balloons, and (b) it wouldn't get soaked in rain, thereby making the balloon too heavy and forcing it to land. Of course, the proposal got turned down again, and aluminum remained an obscure metal for more than another quarter-century (and aircraft made from aluminum had to wait for a whole century).

My fable is obviously ridiculous—and is intended to be so. But it describes exactly what we are doing today when we try to force a researcher to tell us how a new research direction on, say, quantum devices, fits into a CMOS world.

2.4.3. *Everything isn't a Computer*

In the wake of the triumphs of Si IC's, and of CMOS specifically, too much of a tendency has developed—this workshop is no exception—to judge everything in the context of digital IC's, especially high-density memories, and of the specific problems associated with the voracious appetite for increasing complexity in computers, leading to increasing density. No matter how important digital signal and data processing are, they aren't going to be the only application that will be around. Precisely because so much progress has been made already, other areas (for example, photonics) just might turn out to be bigger beneficiaries of new technology than digital IC's.

At least one such direction is already under way: Just as there is no foreseeable limit to the appetite for more and more complexity in IC's, I see no limit to the appetite for more and more bandwith in communications. This calls for even more speed than can be achieved in highly-integrated CMOS structures, and the push for ever-higher speed, even at low levels of integration, will be an increasingly important driving force in the decades to come. But there are bound to be many others.

2.5. TECHNOLOGICAL DARWINISM: THE ROLE OF LEVERAGE

Too many attempts to look at the future of semiconductor judge new device concepts by whether they can be mass-produced at the huge volumes and low cost that are characteristic of Si integrated circuit technology. This is of course appropriate for concepts that are indeed intended to find their application in the same market as Si integrated circuits, where it is indeed extraordinarily difficult to compete with the existing technology.

But remember that the applications of new concepts are more likely to be applications that get *generated* by the new concepts than pre-existing applications, and here the economics is an altogether different one. What matters for the economic viability of the new technology is simply whether the added value of the new application can support the R&D cost and the manufacturing cost of that technology. If a new technology has enough of that crucial *economic leverage* I referred to earlier in the context of HEMT's, it may be economically viable even at a low manufacturing volume and a high attendant cost per device. For example, if a new but expensive-to-make $1000 device makes possible a new $20,000 instrument that simply cannot be built without that device, and if there is enough demand for the enhanced capability of that instrument to permit a recovery of the cost of making each device, then the technology for making the device becomes self-supporting, and has a chance of surviving—never mind that the increase in cost over, say, silicon technology is huge: The latter cannot do the job. Recent history abounds with examples of such high-leverage devices, and one of my predictions is that we will see much more of this, especially in the instrumen-

tation and sensor field, and that high-leverage applications in these fields will be amongst the driving engines of device technology for the next century.

The number of such devices for any single such application, and even their associated money value, may be minuscule compared to the number and money volume of Si IC's, but this does not in any way diminish the attractiveness of the devices to those working on them: Working on high-leverage special-purpose devices may, in fact, be an attractive career path for a young scientist or engineer. Moreover, it is an excellent way for universities to prepare future scientists and engineers for the technologies of the future. Nor are such high-leverage activities negligible from the point of view of the economics of an entire nation: While each individual example might indeed have a negligible impact on that economics, the *cumulative* effect of the very large number of such activities can be huge.

3. The Role of the Universities: Research as Part of Education

3.1. THE NEED FOR OPEN-ENDED RESEARCH —AND INDUSTRY'S RETREAT FROM IT

The idea that the principal applications of new technology will be applications created by that technology, calls for an assessment of the role of *open-ended research*, not tied to a specific application.

What I call *open-ended research* is often referred to simply as *long-term research*, but long-term research need not be open-ended, as some of the discussion at this workshop on the development of CMOS technology past the year 2010 demonstrates. Some call it *curiosity-driven*, as opposed to being applications-driven, but this, too, does not hit the mark: While the motivation of the individual researcher may very well be pure curiosity, those of society at large, which supports this activity, are not: Ultimately, society *does* expect a payoff even from open-ended research, be it direct or indirect. Society is simply willing to leave it open what that payoff might be, based on the experience that there always has been such a payoff, not necessarily on every project, but certainly collectively. This specifically includes the recognition that the payoff has often been indirect, through its impact on subsequent research one or more research generations down the road.

One of the developments of the last decade that has deeply influenced and even shocked all of us—and continues to do so—is the retreat of industry from this *open-ended research*. I shall not analyze here the reasons why this happened, nor bemoan it, but simply take it as a given that is likely to remain with us, and look at some of the consequences, and specifically on the impact of this development on the universities: It may very well turn out that the only places where open-ended research can be conducted in the future on a significant scale will be the universities. Let us turn to this issue.

3.2. THE ROLE OF THE UNIVERSITIES

3.2.1. *Education versus Research?*

The need of society for open-ended research, stated above, has not changed, hence the retreat by industry from this field evidently puts a much larger responsibility for open-ended research on the universities. However, we should not take it for granted that society at large will automatically realize this and treat those of us who are at universities accordingly: In the eyes of most of society, the *primary* function of the universities—and the primary reason for supporting them—is education, not research. Put bluntly:

> The principal product of universities is people — highly educated people.

I actually agree with this idea wholeheartedly, but:

> The open-ended research conducted at universities not only meets a need of society for such research, but it is also an essential ingredient in our educational mission, an ingredient without which we could not fulfill that mission.

I believe it is essential that those of us who are engaged in this kind of work tale a more active role in bringing this last point to the attention of everybody else involved—so they don't hear on;y the other side.

I will say relatively little about that part of my assertion that implies that society needs a healthy open-ended research activity, if not in industry, then somewhere else. Superficially, it appears that society's own recognition of this need has not changed. For example, we all hear such buzzwords as *industrial competitiveness*, presumably acknowledging a recognition of this need. But much of what I hear and see seems to be more lip service than evidence of positive action; in fact, there is plenty of talk that actually conflicts with a true recognition of that need. For example, the loud clamor for more "relevance" even in non-industrial research can be safely translated into a clamor for *less* open-ended research—and sometimes for less research of any kind.

Let me concentrate on the other part, that such research is an essential ingredient of the educational mission of the university.

In some circles, the research done by the universities is not even viewed as meeting an essential need of society, much less as an ingredient in education, but as a luxury embarked on by an elite fraction (presumably a derogatory term) of faculty at those

universities that we call research universities—themselves only a fraction of all universities. Being a member of that elite (and not at all ashamed of it), I will be the first to admit that I am gratified to be amongst those who are able to embark on this supposed luxury. But is it really a luxury?

3.2.2 *The Educational Role of Ph.D.-Level Research*

The educational mission of universities spans a wide range, from the 4-year bachelor's level to the full Ph.D. level. When talking about research, we are primarily talking about the Ph.D. end of this scale. Taking it for granted that the Ph.D.-level education itself meets an essential need of society for future leaders in science and technology, I claim the following:

> There is no known better way to fulfill the university's educational mission at the Ph.D. level than through research, with a significant fraction of that research being open-ended.

Furthermore, even though the total number of Ph.D.'s is only a fraction of the total number of university graduates, even at most Ph.D.-granting research universities, this fraction has a huge leverage, not only in industry, but also on *all* levels of education.

The Ph.D. as Future Teacher. The overwhelming majority of the teaching faculty in science and technology at all universities, including at the undergraduate level, are themselves products of Ph.D. programs. I stated earlier that the primary product of universities is people; on the Ph.D.-level, one of the most important groups of those people are those whose mission is as future educators at *all* university levels.

It is absolutely essential that those individuals must themselves have an education that is up to the state of the art in their field. But just as important, they must be provided with the intellectual resources that permit them to *stay* at this forefront throughout much of their careers, rather than having become obsolete carriers of past knowledge before the midpoint of that career. A research Ph.D. provides this *obsolescence reserve* by providing the student with a knowledge the fundamentals that will be part of the foundations of future developments as well. A Ph.D. research project does that more thoroughly than mere coursework ever can, and it provides an opportunity to develop the skills of actually applying those fundamentals in a real context. It doesn't really matter much what the particular set of fundamentals is that is drawn upon in a given research dissertation: It is the acquisition of a *methodology* that matters: *The method is the message*!

The Ph.D. in Industry. All the above applies just as well to a career in industry, but there is more. Going beyond research per se, open-ended research provides a "live education" in what I like to call the *management of uncertainty*, the need to make rational decisions in the face of very incomplete knowledge. This is a skill important to any future leader in science and technology, including the teacher at a university, but it becomes indispensable in industry, where a scientist or engineer often has to make decisions about major commitments to future technology, in the face of precisely the kind of very incomplete knowledge that characterizes open-ended research. The experience provided by the latter is likely to lead to more rational decisions than simply being conservative (missing valuable opportunities), or following a "gut feeling," or following the bandwagon "consensus" of equally-uninformed others. An additional merit of an education involving open-ended research is that it tends to create people with a personal stake in innovation, rather than in the maintenance of the status quo. Now we all pay a lot of lip service to the need for innovation, often under such national banners as industrial competitiveness, but much of it is again just lip service. I know only one way to "institutionalize" innovation, and that is by bringing in young people and relying on that most reliable of all human motivations, self-interest: They have nothing to gain by maintaining the status quo, and everything by being the leaders of innovation.

3.3 INDUSTRY AND UNIVERSITY: THEIR COMPLEMENTARY ROLES IN SOCIETY

Properly understood and managed, the retreat of industry from open-ended research can lead to a partnership in which both parties understand and serve their complementary roles in society as partners rather than adversaries. But this partnership is not to come about without an understanding of each other. There are problems on both sides.

The most serious threat coming from industry is the wide-spread clamor for "more relevance" in university education. The reason for this is of course the desire of many industrial managers to hire university graduates who make a positive contribution to the "bottom line" from day-1, without requiring any further in-house training—let the long term be damned! On the bachelor's level, economic realities may make this, to some extent, inevitable (if deplorable). But it has very little, if any, validity on the Ph.D. level, where it is nothing other than a thinly disguised call that universities abandon the long-term research needs of society just as industry is abandoning them—and short-changing their students along with society at large in the process.

We must resist these calls—but we must also communicate to society why! In the last analysis, our obligations as educators are to our students and to society, not to short-term-driven accountants at corporate headquarters, who—as recent years have amply demonstrated—are increasingly viewing even the skills of their R&D staff as

short-term commodities rather than as valuable long-term investments. *If* career changes are to be an essential ingredient in the future of industrial Ph.D.'s, we have even more obligation to give them the broad long-term education that enables them to make such changes. But that does not mean sacrificing a research education—to the contrary.

3.4. THE OTHER SIDE OF THE COIN: UNIVERSITY'S OBLIGATIONS

Having said the above in defense of our research/teaching mission, I turn to the list of things where we ourselves must do a better job as part of the bargain.

Probably the first need is that the university needs to communicate its role better to society at large—much better. Because otherwise, society, and the political decision makers whose decisions presumably reflect society's will, will only hear the other side, which tends to be rather vocal.

Next in line is a need to recognized the changed role of the research Ph.D. itself. This has several aspects.

We must stop pretending that we are educating our Ph.D.'s students for "pure" research careers, academic or otherwise. Only a minority of them will enter such careers; most will enter teaching careers or applied R&D careers (with more 'D' than 'R'), and we must make this clear to ourselves and our students:

> The research Ph.D. is an education in rational problem solving, rather than a preparation for a research career.

A closely related need is to abandon a value system that rates "pure" research somehow higher than applied research. What industry demands,—and justly so!—is that we provide our students with a motivation towards the kind of shorter-term and more applied work that is the very nature of industrial R&D, rather than creating people who do that sort of thing only reluctantly. This is going to be one of the hardest changes to make, because it is ultimately a cultural rather than scientific change, but it is necessary all the same to make applied research at least as respectable as "pure" research. This attitude seems to be highly developed in Japan, which goes a long way towards explaining Japan's industrial success. A change in social values amongst the rest of us might, in turn, go a long way helping us retaining our global competitiveness.

Finally, we must recognize that our Ph.D.'s will change their work many times in the course of their careers, and that their education must enable them to handle such changes. This calls for more breadth than is contained in a dissertation in which a student spends too many years studying a single (usually rather narrow) topic in "infinite-depth," often far beyond the depth the topic deserves. The Ph.D. should be a

certificate that says, in effect: "This individual has proven that he/she is capable to perform independent high-quality engineering or scientific work, has adaptability to a wide range of needs, and the ability to make, within a broad strategic context, the decisions about how to conduct that work." This is far more useful than, say: "This individual has spent over four years studying the low-temperature optical absorption of sowhatnium, has honed the technique involved to perfection, and knows more about this specific topic than anyone else in the world."

4. References

1. Lepselter, M. (1974) Integrated Circuits—The New Steel, *IEDM Digest*.
2. Kroemer, H. (1982) Heterostructure Bipolar Transistors and Integrated Circuits, *Proc. IEEE* **70**, 13-25.
3. Kroemer, H. (1963) A Proposed Class of Heterojunction Lasers, *Proc. IEEE* **51**, 1782-1783.
4. Kroemer, H. (1967) Solid State Radiation Emitters, *U.S. Patent* 3,309,553.

Si-MICROELECTRONICS: PERSPECTIVES, RISKS, OPPORTUNITIES, CHALLENGES

12 Statements

ARMIN W. WIEDER
Siemens Corporate Research and Development
Microelectronics
Otto-Hahn-Ring 6
81730 Munich, Germany

Abstract

The impressive progress in microelectronics in the last two decades has generated enormous computational power and huge storage capacity at ever decreasing cost per function. To carry on progress requires overcoming enormous challenges and likewise will take advantage of great advances and opportunities. This translates into physics overcoming all kinds of technical and technological limits. It requires technologists to cleverly increase productivity and creatively define new products. Furthermore, it requires overcoming the economic hurdles of exploding R&D costs and increasing manufacturing investment. New opportunities, e.g. in the field of "system on chip", will creatively integrate logic, memory and other functions on the same chip. New schemes for cost-effective R&D and manufacturing will impose tremendous challenges, especially on R&D personnel. These needs will require education, skills and creativity, with much broader levels of expertise and knowledge ("research to production") in the increasingly overlapping fields of devices, processes, circuits, chip architectures, product definitions and manufacturing cleverness. The complex situation is condensed into 12 major statements.

1. Technology Perspectives

It took about 10 years after the invention of the transistor in 1948 to invent planar technology and the integrated circuit: in 1959 the evolution of bipolar technology started. It took another 10 years to overcome the interface problem of MOS devices, so that the evolution of MOS circuits started in 1969. Since then every 3 years a new generation of MOS technology emerges with 4 times the transistor count (complexity) of the previous generation.

Today the leading semiconductor manufacturing companies are *producing* 4 Mbit and 16 Mbit DRAM's and logic IC's with critical dimensions of 0.7 μm and 0.5 μm respectively. *Development* in these companies is concentrating on 64 Mbit and 256 Mbit DRAM structures and the corresponding logic, with the 64 Mbit DRAM currently entering the pre-production phase. Fully functional samples of the 256 Mbit DRAM's

S. Luryi et al. (eds.), Future Trends in Microelectronics, 13–21.

are available in stacked as well as in trench implementations. R&D is successfully focusing on 1 Gbit and 4 Gbit devices with 0.1 μm structures and *exploratory research* is evaluating concepts for even higher complexity levels and finer structures < 0.1 μm.

The optimism for believing in the manufacturability of such complex devices basically results from three sources. Firstly, it is based on the extensive use of *competent simulation tools* indicating that today's MOS structures can be scaled down to about 20 nm without reaching their functional limit. Secondly, exploratory devices fabricated by "*dirty processing*" have confirmed these simulations: transistor structures with critical dimensions down to 30 nm operate reasonably well for all kinds of binary applications. Last but not least, R&D is pressured to achieve the results desired: about 15% of sales is being spent for research world wide. Therefore, the scientific and industrial community is convinced that progress of microelectronics will not encounter fundamental roadblocks for the next 15 years. Moreover the decrease of performance due to V_{DD}-scaling will partially be compensated by ballistic transport (effective below 80 nm). Likewise V_{DD}-reduction will effectively help to overcome degradation problems.

Further progress sees no "show stoppers" for the next 15 years (1)

Nevertheless enormous challenges have to be met to make this happen. Research in processing will no longer focus on the device level, but more and more on the metallization and/or storage systems on top. Progress here requires substantial *material innovations*. Metals need improving on electromigration, especially on stress migration. Low ε, high ε and ferro-dielectrics have to be explored. The new materials required in Si microelectronics are collected in Fig. 1.

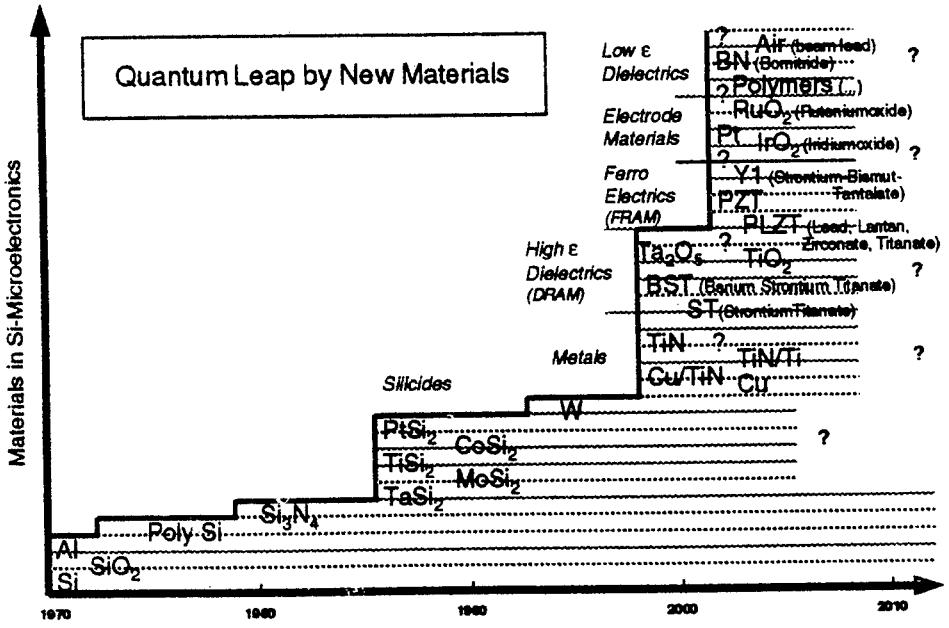


Figure 1.

Further progress increasingly needs material innovations (2)

The evolution of bipolar technology has resulted in integrated circuits with gate delays below 20 ps, complexities of 10^5 components per chip and communication IC's with data rates up to 40 Gbit/s and operating frequencies up to 30 Ghz. Further progress will not encounter fundamental technical roadblocks, but bipolar technology is challenged by the competing mainstream CMOS technology. *Performance migration and cost advantages of CMOS* force bipolar more and more into analog and high frequency and likewise force GaAs (and other III-V's) into super performance and optoelectronic applications (Fig. 2).

CMOS technology wins in complexity, power dissipation and system speed (3)

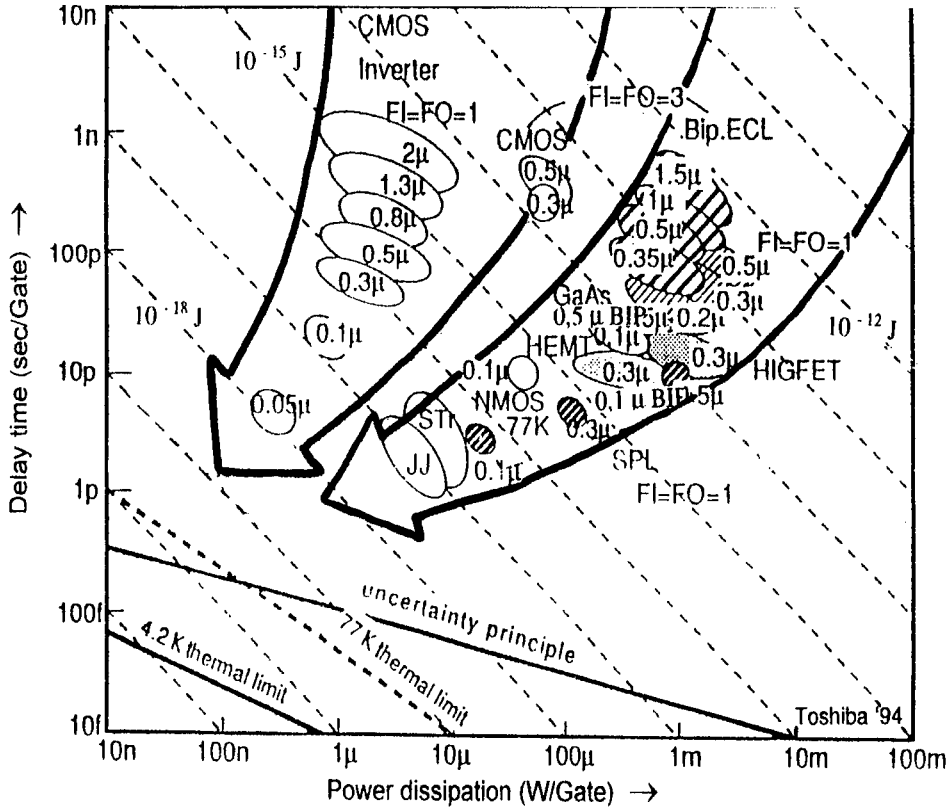


Figure 2. Trends in power-delay-products.

Fundamental considerations show that there is "microelectronic life" beyond CMOS. Chip complexity continues to increase by a factor of 4 every 3 years (exponential growth). Power consumption per chip, however, has to be limited to some reasonable value (e.g. 1 W for portable electronic devices and 100 W for high

performance systems). This leads to the consequence that the number of active electrons per transistor has to be decreased, step by step, because it is directly proportional to the power consumption of the chip. Calculation shows that between year 2010 and 2020 switching devices must have only a very small number of active electrons (e.g. 10). A *fundamental change of the functional concept* is approaching: instead of controlling the average behavior of a large number of electrons, control of a small number of individual electrons is necessary; this consequently leads to single electron devices (also known as quantum effect devices, or QED's). The potential functionality of those devices already has been successfully demonstrated in nanoelectronic research.

QED's have the potential to carry on microelectronics beyond year 2020 (4)

The enormous know-how of mainstream microelectronics is driven by memory (volatile and non-volatile) and logic applications (µP, µC, ASIC's, analog, DSP's, *etc.*) and more and more by "system on chip"-type applications with "on-chip" memory and logic functions. This knowledge base cost-effectively (*nearly for free*) will be used more and more for "spin-off-technologies" like "smart power". The integration of "power and intelligence" opens up large fields of monolithic solutions. Likewise the knowledge will be used to cost-effectively realize *CMOS-compatible microsystems*. Monolithic integration of "intelligence" and sensors & actuators will establish itself parallel to the mainstream of microelectronics. Cost-effectiveness can be guaranteed by restricting technology, equipment and manufacturing facilities to the current generation of microelectronics (Fig. 3).

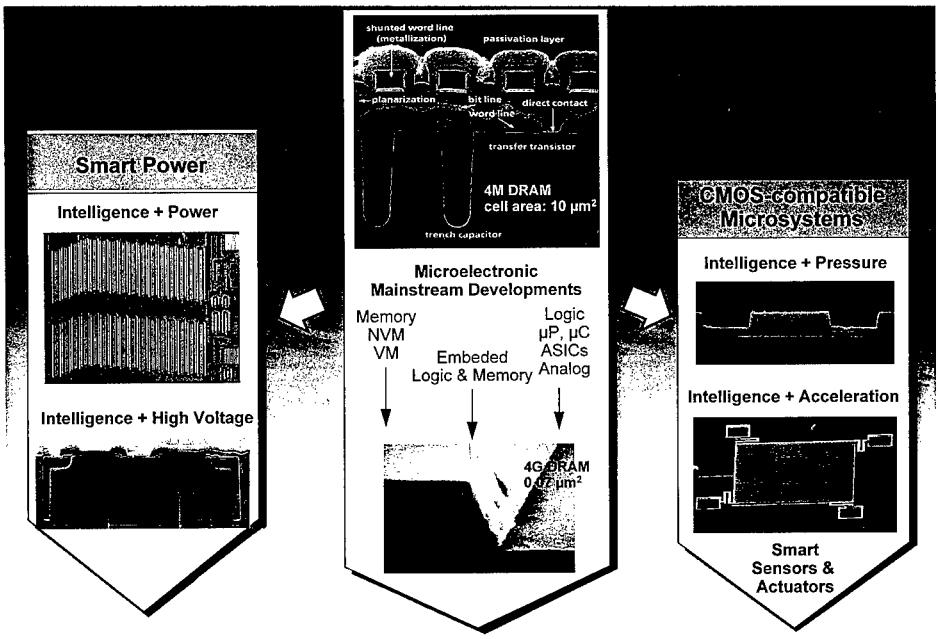


Figure 3. Microelectronic mainstream and spin-offs

This also holds for the increasing need to develop *novel packaging concepts*, passive as well as active, eventually ending up in vertically integrated circuits (cubic integration) with higher packing density (with chip size becoming the same as the packaging size), lower power consumption, and increased system speed.

Microelectronic spin-offs become important: smart power, sensors, packaging, micro-machining, ... (5)

2. Economic Perspectives

The investment in manufacturing facilities has doubled from generation to generation. This has resulted in dramatic increase of manufacturing as well as R&D costs. Cost generating factors in this context primarily are the decreasing lithographic dimensions, the likewise decreasing critical defect sizes, the increasing wafer size, the increasing number of processing steps and, last but not least, the increased size of the manufacturing facilities. To meet this economical challenge, *smart fabrication* and *smart R&D* schemes will have to be developed, leading to more cost-effectiveness at all levels. In the case of processing, this will lead to design and process development with simpler, fewer, and more controllable process steps. On the equipment side it will lead to *standardization*, *reuse* and *modularity* of equipment. The equipment will be a 100% controlled unit with *in-situ-monitoring and sensors*, as well as *embedded simulation* (equipment/process simulation). It will lead to higher equipment specific yields, as well as higher utilization levels due to improved process control. Shorter cycle times will be the key factor for reducing cost and increasing productivity. Instead of mass volume pre-manufacturing, *accelerated learning* with smaller number of wafers and shorter cycle times will become the reality. To this end, computer-aided failure analysis and integrated simulation tools (virtual R&D, virtual manufacturing) look most promising. A key factor for increasing control will be the gradual move to *single wafer processing*. This will be motivated by improved control and compatibility with super-sized wafer diameters. Also it facilitates accelerated learning and therefore in the long run will become the most cost-effective way to develop and produce microelectronics.

Smart (accelerated) fabrication & development become the # 1 economic challenge (6)

Last but not least, decreased time to market and time to volume production require *concurrent engineering* for development (whereby product development and process development are completed simultaneously) as well as for *concurrent manufacturing design*, ranging from equipment development to process development, product development, facility design, and manufacturing operations.

3. Opportunities and Applications

The key performance yardsticks of modern microelectronics are 16 Mbit DRAM's with 40 ns access time; 300 k, 100 ps gate arrays; 0.5 μm CMOS with 10^7 components per chip; and so on. This leads to figures of merit like computational power of about 200 MIPS. Future requirements of speech recognition and speech processing (e.g. translation) with a vocabulary of 10^4 words in real time, however, will need about 10^5

MIPS. Even more computational power will be needed for high quality (HDTV) real time image/video recognition and processing: 10^6 MIPS. This also holds for "intelligent" robots, expert systems, future man-machine-interface solutions, and other sophisticated applications in information technology. All of these applications need an increase of performance of 3 to 4 orders of magnitude.

This cannot be achieved by technological improvements alone. Novel architectures are needed. All of them, however, need a massive increase in components per chip. Therefore, by virtue of its superiority in yield, low power dissipation, design ease, noise immunity, *etc.*, CMOS will be the winning technology. Follow-up technologies will have to be better in all aspects and not in just some of them.

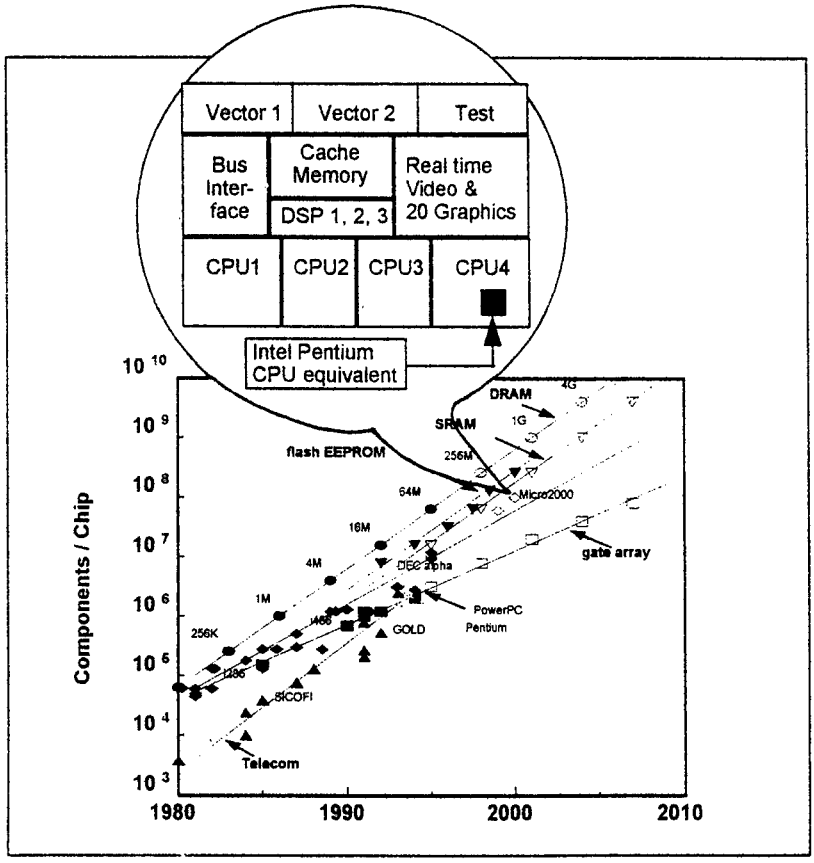


Figure 4. Forward integration of Si-microelectronics into systems.

The possibility of having *on-chip memory and logic* will lead to rationalized board solutions of current products and also to novel systems not feasible today with "quantum leaps" in performance. The latter will be realized by massive parallelism in space (complexity) and time (pipelining). Neural networks, for instance, look best for handling the complex problems of real-time man-machine interfaces. The trend to on-

chip memory and logic functions will cause microelectronics to go on integrating forward into systems with increasing system knowledge and value added in silicon. This will enforce much more *intensive cooperation* of system and semiconductor manufacturers than in the past, as illustrated in Fig. 4 where future devices are classified by system content rather than, say, minimum feature size in μm.

Microelectronics integrate forward into systems (system on chip) (7)

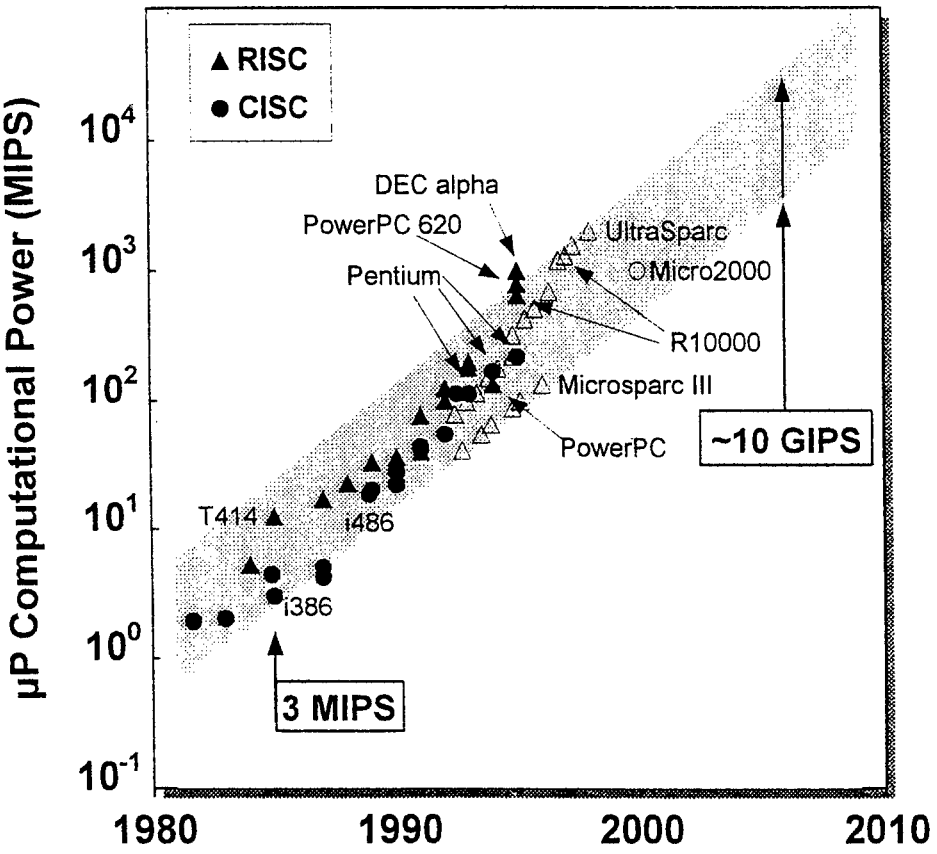


Figure 5. Trends of chip computational power.

The trend to the decentralized "intelligence" of computers (PC/WS networks) and to "intelligent" communication networks (ATM network management) likewise is driving complexity, computational power and especially transfer rates of the operating networks. All this, however, can only become true, when the *#1 technological challenge*, the reduction of power dissipation per chip can be successfully met. Power dissipation of today's IC's has to be reduced in order to allow for full portability of the electronic devices (the forecast that by the year 2000 about 50% of electronic devices will be portable was made already at the 1993 IEDM). This holds even when the

performance/functions will be increased by orders of magnitude. The power dissipation needed per speed/function, therefore, will become the figure of merit for the Giga-era. It will be met by harmonized contributions at various levels of development, including the technological level (physical scaling instead of V_{DD}-scaling), circuit level (static CMOS, carry save & pointer techniques), layout level (maximal local wiring), CAD level (minimization of wiring & duty cycles) and the very important architectural level (massive parallel in space (for suitable problems) and time (pipelining concepts)).

Low power dissipation becomes the # 1 technological challenge (8)

Under these conditions systems on-chip with "quantum leaps" in performance will be manufactured. The dramatic progress can best be demonstrated by the prognosis that in the year 2005 we will have the computational power of a CRAY-2 machine on a single chip (Fig. 5)!

The computer-aided design and manufacturing (CAx) systems have to overcome major challenges, too. On the one hand, deep submicron parasitics require CAx-systems to deeply *dig into physics*, and on the other hand, the increased level of complexity (Mega to Giga) requires solutions at *higher abstraction* levels. Furthermore, the aid of computers and simulation tools is needed for questions ranging from physical, to electrical, to economic: e.g. equipment, process, device, circuit and logic simulation, accelerated yield learning, cost simulation, yield prediction, and so on all the way to virtual factory and management tools. Here *intensive cooperation* between the technology designer, the system architect, the system designer, the manufacturing specialist, the manager, the business administrator, and the controller are mandatory in order to provide relevant and cost-effective CAx-systems: ready for use simultaneously with the corresponding microelectronic technology and the manufacturing facility.

Integrated CAx-systems from physics to economics become crucial (9)

4. Conclusions and Megatrends

The major changes in computer industry from 1980 to 1995 which led to the decentralization of computers were primarily caused by the cost reduction of computational power due to the progress of microelectronics. Today it appears that further cost reduction and miniaturization of communication technologies, even greater decentralization and networking of "computer intelligence" (PC's/workstations) and, last but not least, the further reduction of cost of computational power will lead to the *redefinition of a new powerful industry* emerging from communication, data and consumer industry. TV's will become "intelligent" and gain communication capabilities, the phone will be upgraded by "intelligent" displays and portability, whereas the computer will get additional features like communications and portability.

Key competencies for this industry will be *batteries*, *displays* and *low power IC* techniques like µP's, DSP's, µC's, DRAM's, E^2PROM's, telecommunications IC's, ..

Interdisciplinary creativity becomes the #1 educational challenge (10)

To be successful in this industry, personnel will need a broad level of expertise, ranging from processing and circuits, to systems and software, and cybernetics.

Interdisciplinary creativity e.g. for optimal hardware/software co-designs of complex systems, therefore, will be a crucial success factor for future IC innovations, when microelectronics migrates more and more into systems.

According to analysts, the world IC market is growing more than twice as fast as the world electronics market. The dependence of the electronic industry on microelectronics, therefore, will become even stronger. Microelectronics will continue growing for another 10 to 15 years at the same pace as in the past.

Microelectronics will grow twice as fast as the world electronics market (11)

This, however, implies exponential growth for another 15 years: an *exponential growth of "technical intelligence"* (complexity, computational power, transfer rates, speed, ..). This gigantic increase in "technical intelligence" could offer the means for solving the serious problems in various areas of our life: environment, traffic, medicine, office, manufacturing, limited resources, energy, telecommuting, *etc.* This will become reality by the shift of paradigms in the fields of *technology* (where the #1 challenge will be low power electronics), *economics* (where the #1 challenge will be low-cost fabrication) and *education* (where the #1 challenge will be interdisciplinary creativity). It is no longer the device, it is the system that counts. And it will be in silicon.

Microelectronics and software will change the world more than any other technology in history (12)

MASS PRODUCTION OF NANOMETRE DEVICES

ALEC N BROERS
Churchill College
Cambridge CB3 0DS

1. Device and Circuit Requirements

It has become clear that the benefits of miniaturising electronic devices will continue until dimensions enter the nanometre region, that is, below 0.1 μm. This is beyond the resolution limit of the ultra-violet (UV) projection method used to manufacture today's integrated circuits so a replacement will have to be found. The major challenge for the new method is not resolution but the need to meet the extreme image complexity and accuracy requirements at acceptable cost. Microchip images already contain more than a billion pixels and if nanometre devices are to be cost-competitive, the pixel count in a single 'chip' will have to approach a trillion. A pixel is defined here as being four times smaller than the minimum feature size. Acceptable cost is essential if there is to be commercial justification for further development. The requirements of cost and pattern complexity will remain no matter what device technology is used. The only way to avoid them would be to have the structures self-assemble but as yet there are no methods for making contact to self-assembled structures nor are materials that self-assemble suitable for electronics devices.

At present, in order to meet throughput requirements, circuit images must be projected or shadow-printed from a mask. The mask can be written at a slower pace with a scanning beam. The scanning beam cannot be used directly because it is too slow. The fastest scanning systems, which are those that use variably shaped electron beams, only run at an exposure rate of around 10^9 pixels per second and would take about 200 seconds to write one level of a 1G bit chip. A UV projection camera exposes the same pattern in one or two seconds. The cost of the scanning systems is typically five times that of the UV cameras ($15M versus $3M) so the electron beam approach is hundreds of times more expensive. The cost of exposing the images is already becoming excessive so any increase in the exposure cost per pixel is unacceptable.

In this paper the various lithography methods capable of producing nanometre dimensions are assessed for mass-production. Ultra-violet light projection is discussed first to justify the conclusion that it is incapable of producing nanometre dimensions.

2. Ultra-violet (UV) Light Projection

It is not possible to predict precisely the resolution limit for UV projection but the ultimate camera might operate at a wavelength (λ) of 160 nm and have a lens with a Numerical Aperture (NA) of 0.75. Using the standard expression for the minimum

S. Luryi et al. (eds.), Future Trends in Microelectronics, 23–34.

feature size, $k\lambda/NA$, where k is typically 0.8, this camera would reproduce 0.17 μm lines. k depends on the mask type, the illumination and the resist contrast. With phase-shift masks and off-axis illumination as described for example by Okazaki[1], combined with proximity-effect correction and a resist in which only a shallow surface layer needs to be exposed[2,3], a value of 0.5 might be reached for k and the feature size reduced to 0.1 μm. This would allow the 1G bit chip and its equivalents to be made with UV projection and perhaps the 4 Gigabit generation as well. The depth of focus, $\lambda/(NA)^2$, would fall to 0.28 μm, but methods may be available for effectively increasing this, for example, by using dual masks spaced along the optical axis.

In addition to small depth of focus, difficulties will remain with sources and with the fabrication of optical components for the sub-200 nm region. Of these, densification and colour centre formation in fused quartz, and the lack of suitably reliable laser sources, have already been identified. The performance of lasers in the vacuum UV regime will also have to be improved.

Overall, it will be extremely difficult to reach 0.1 μm with UV projection lithography, and to produce smaller dimensions would seem impossible.

3. Soft X-ray Projection and X-ray Lithography

The most obvious way to penetrate the 0.1 μm barrier is to retain the basic concepts of UV projection but to use much shorter wavelength radiation. Optical components for wavelengths below 100 nm are difficult to make but reasonable performance has been obtained with multi-layer mirrors at 13 nm and this has led to proposals for projection cameras[4,5,6]. The major difficulty is that the mirrors must be fabricated with an accuracy of a few tenths of a nanometre and which is impossible today but may be achievable in the future.

The alternative method for using X-rays is shadow-printing[7,8]. This is known as X-ray lithography and it is a technique that has received much more attention than soft X-ray projection. The shadow-mask consists of a thin membrane with adequate transparency to the soft X-rays and a thick metal layer to absorb the X-rays. Much shorter wavelength (0.5 nm to 2 nm) X-rays are used than with the projection method to avoid diffraction blurring. The accuracy requirements of the mask can only just be met for 0.2 μm today and for nanometre dimensions may be impossible, particularly because of the stresses induced by the thick metal absorber.

With X-ray lithography, the mask and wafer are separated by a gap that is large enough to prevent them touching each other but not so wide that image degradation due to diffraction is unacceptable. The minimum linewidth is approximately given by the simple expression $\sqrt{\lambda g}$, where λ is the average wavelength of the X-rays and g is the gap between mask and sample. $\sqrt{\lambda g}$ is the linewidth at which the intensity in a narrow line first reaches that of the background with a transparent mask, assuming single

wavelength, coherent, illumination. In practice, the situation is more complex because the source produces a spread of wavelengths, the absorber is not completely opaque and the substrate is not completely transparent. In fact both introduce phase shifts and there is the possibility of using these phase shifts to improve image contrast. Despite this complexity, $\sqrt{\lambda g}$ still gives a useful estimate of minimum linewidth. In the limit for chip lithography, a gap of 5 μm might be acceptable which, used in combination with a wavelength of 1 nm, would permit 0.07 μm dimensions to be reproduced. In the laboratory, intimate contact between mask and resist allows dimensions of a few tens of nanometres to be reached but this would be impractical for production.

A major difficulty with X-ray lithography for mass-production is that it is necessary to use an electron synchrotron storage-ring source rather than a simple point-source. This is because point-sources do not produce a large enough flux of X-rays when placed far enough away from the sample to produce acceptably small divergence. Divergence produces distortion when the gap between mask and wafer varies as it invariably does because mask and wafer cannot be perfectly flat. The source could be placed closer to the sample if an efficient collimator lens could be made but techniques for making efficient lenses are not yet available. Work continues on increasing the output of electron bombardment, laser-induced plasma and spark-induced plasma point-sources but so far their output remains at least ten times too low. An alternative is to improve the sensitivity of the resist to 1 mJ/cm^2 as compared to the practicable limit of 10-20 mJ/cm^2 but so far attempts to accomplish this have failed.

The only source that provides an adequate flux of X-rays is the electron synchrotron storage-ring. Electrons in a storage-ring emit electromagnetic radiation as they are constrained in their circular orbit by the magnets that comprise the ring. Tens of wafers per hour can be exposed with the X-rays produced from a few degrees of the 360° orbital path of a few hundred milliampere, 500 MeV to 1 GeV, beam. The divergence of the fan of X-rays leaving the ring is so small that the beam may be considered to be parallel and the placement errors, that occur because the source is divergent, become negligible.

Storage-rings reliably produce adequate intensity for practicable lithography although concerns remain about cost and about the manufacturing logistics of using a single high output, high cost, source. Steppers for use with storage-rings have been built which nominally meet the requirements for minimum features of 0.18 μm (NTT and IBM have both demonstrated the capability for 0.2-0.25 μm groundrules) and it is reasonable to assume that steppers capable of reaching the diffraction limit of 0.07 μm will eventually be built.

In summary, it should be possible to use X-ray lithography for the mass-production of devices with dimensions down to about 0.07 μm, that is, marginally within the nanometre regime. However, the problem of manufacturing with a single source, which illuminates many exposure stations simultaneously, will have to be tolerated and the

difficulty of maintaining mask dimensional stability overcome. Projection printing with X-rays is far from practicable realisation because the fabrication of the mirrors does not look feasible for many years to come.

4. Direct-writing with Scanning Electron Beams

Electron beams are used successfully for three applications; mask-making, 'quick-turn-around' production of ASICs (Application Specific Integrated Circuits), and the fabrication of devices in research and development. As already mentioned, they are too slow and expensive for mass-production. Electron beams provide higher resolution than any other method except local probe atom manipulation, allowing the size of devices to be reduced to the practicable limit of resist processes, which at present falls at about 10 nm.

To increase the throughput of scanning electron beam systems, methods are being explored in which more and more of the pattern is exposed at each beam flash. The slowest systems write the pattern one pixel at a time with a round electron beam whose diameter is equal to the pixel size. This is very flexible and accurate but is extremely slow. The next fastest systems use a beam which can take on any shape up to about four times the minimum feature size[9,10]. This allows tens of pixels to be exposed at each flash and makes it possible to manufacture ASICs. Finally, pattern cells containing tens of shapes and thousands of pixels are projected in a single flash in what are called 'character' or 'block' projectors. These are the only scanning systems capable of potentially reaching realistic throughput for pattern replication[11,12,13].

Figure 1 shows estimates of the throughput (as a function of beam current density) for the three types of scanning electron beam systems scaled to 0.05 μm dimensions. The following parameters were used in the estimates; minimum feature size 0.05 μm, pixel size 0.013 μm, pixel delay time for round beam 0.5 ns (this will only be possible if the pattern is treated as a series of shapes with >20 pixels per shape so that the delay per shape is $\geq \sim 10$ ns), shape and character/block delay time 30 ns for shaped beam and character projection systems, sub-field delay time 3 μs (it is assumed that there are only two levels of deflection although three levels will probably be needed to meet the required accuracy), chip delay time 100 ms, wafer change time 20 s, fraction of chip area exposed 30% for round and shaped beam cases and 100% for character/block exposure. Because the assumptions about electronic delay times are optimistic, the throughputs in all cases are dominated by writing time and not by overheads and must be considered to be higher than are likely to be achieved in practice.

In practice the maximum usable current density is set by electron-electron interactions and therefore depends upon the total beam current. The electron-electron interactions increase the velocity spread in the beam and therefore the beam blurring due to the chromatic aberrations of the focusing lenses and deflection systems. To minimize the interactions, the electron beam columns are made as short as possible.

At present for dimensions of 0.2 μm, current densities above 1000 A/cm^2 are feasible for round beams, but for the shaped and character systems the maximum current density is only about 100 A/cm^2 and 20 A/cm^2 respectively.

The most serious problem with character/block projection is that the pattern must be printed with a finite number of characters or blocks and yet customisation of design rules to accommodate peculiarities of lithography systems has been considered unacceptable in the past. A further difficulty is that proximity effects, which are discussed later, can only be corrected by changing the shapes of the individual pattern elements. The exposure dose has to be the same for all elements in a character or block. Proximity effects are more easily corrected with single beams where it is possible to change the dose as well as the shape.

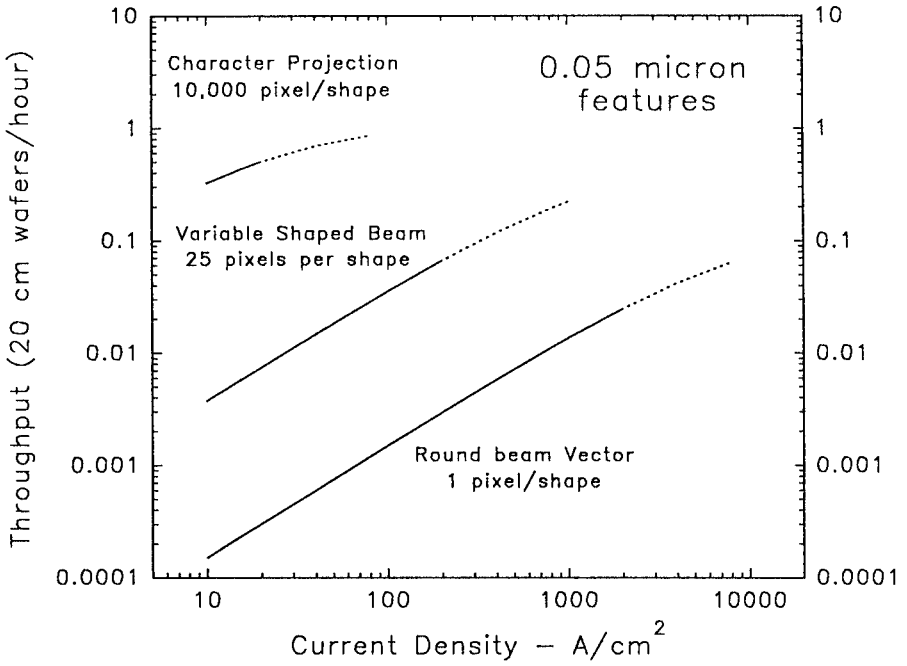


Figure 1. Throughput for 0.05 μm scanning electron beam lithography systems.

5. Multiple Electron Beams

An alternative to exposing many pixels at each beam flash is to use multiple beams. Two distinctly different approaches are being taken to implement this idea. One is a version of the block projection concept in which the character mask is replaced by a multi-aperture mask that in effect produces an array of about 1000 beams[14]. The other is to use ten or more miniature electron beam columns to write on a single wafer. Each

column has a field emission or thermal field emission cathode and addresses several chips[15]. There are practical difficulties to be overcome with both approaches. With the large array of beams there is the difficulty of building a multi-deflection unit that can blank and deflect each beam individually and at the same time keep them precisely registered with respect to each other. The electronic drive unit for this multi-deflection unit must produce relatively high voltages and operate at very high speed. Control of the beams will be difficult to achieve as the degree of electron-electron interaction in the column will change with the number of beams that are active. The miniature electron optical columns promise very high electron optical performance because of their extremely short focal length, very low aberration, lenses but severe mechanical tolerances will have to be met to achieve these low aberrations. It will also be difficult to keep the field emission cathodes emitting uniformly and to fabricate the dense array of electron detectors that will be needed to keep track of the individual beams. Overall, there are many problems to be solved before either approach will be a contender for production lithography.

6. Electron Beam Full Pattern Projection

The final possibility for achieving competitive throughput with electron beam exposure is to build the electron equivalent of the UV step and repeat camera. This was first attempted in the early 1970s[16]. The resolution of the projected image for small field sizes can approach that of an electron microscope, that is a few tenths of a nanometre, and the demagnification can be high enough to make mask fabrication practicable. The first systems demagnified the mask 10 to 20 times. Thin metal stencil masks were used in the early systems and it was proposed to resolve the difficulties with unsupported and/or fragile features by using two masks for each level. More recently a very thin supporting membrane has been used and the electrons that are scattered by the membrane are removed by the projection lens stopping aperture[17]. Provided the membrane is thin, that is less than 0.1 μm, and the accelerating voltage is kept relatively high, for example >100 kV, the loss of current due to scattering can be reduced to about 50%.

The use of a membrane solves some of the mask difficulties but the scattering reduces the target current density and means that the beam current in the upper column has to be increased thereby increasing electron-electron interactions. As with the shaped beam scanning systems, the increase in energy spread blurs the image and this effect will ultimately limits the exposure rate. Preliminary estimates suggest that the exposure speed should be adequate, at least for dimensions down to 0.1 μm, but below 0.1 μm this difficulty may be insoluble.

The absorber for the mask can be a relatively thin (<0.1 μm) layer of a high atomic weight material such as gold that is easy to fabricate. This is a significant advantage over X-ray lithography where the absorber must be up to 0.4 μm thick.

A key difficulty with electron beam projection is pattern distortion. Distortion arises in the mask and in the projection imaging. It should be possible to keep mask distortion at an acceptable level through the use of supporting ribs which can be fabricated with silicon anisotropic etching, but image distortion will have to be corrected. It is proposed to accomplish this with dynamic corrections. The mask will be illuminated with a small beam that will only cover a fraction of the image. Distortion will then be corrected by tilting this beam as it is scanned over the mask to complete the exposure. Variable axis condenser and projector lenses will have to be used and the signals to the correction coils in these lenses will have to be extremely accurately synchronised. The size and current in the illuminating beam will be as large as electron-electron interactions allow.

Another major increase in the complexity of electron projection systems arises because, for realistic chip sizes, it will be necessary to scan the mask and the sample to avoid the lenses becoming impractically large. Large lens diameter increases column length and is intolerable because of electron-electron interactions, as discussed by Berger[18].

There will therefore be two scanning mechanisms operating simultaneously. The illumination will be scanned to correct for distortion and the mask and wafer will be scanned to allow the lens diameters to remain practicable. Both condenser and projector lenses will have to be of the axis type because the field sizes will none the less remain relatively large. The overall complexity will be formidable. It has taken many years to resolve the difficulties with variable axis immersion lenses in scanning systems where the beam diameter is small, there is no mechanical scanning, and there is no need to scan the illumination. The situation with the projector will be many times more complex.

In summary, electron beam reduction projection potentially offers higher throughput than character or block projection systems, but to suppress distortion, the optical system will have to be extremely complex and will take many years to accomplish and it will be difficult to keep the cost within acceptable limits.

7. Ion Beams

Lithography systems that use ions have been investigated for many years but there are a number of factors that place them at a disadvantage compared with their electron equivalents. Firstly, they have to use electrostatic lenses which have higher aberrations than the magnetic lenses used with electrons. It is also not possible with electrostatic lenses to overlap focusing and deflection as can be done with magnetic lenses and coils. This overlapping has been found to be essential with electron beam systems if adequate field size and beam aperture are to be obtained. The disadvantage of the higher aberrations is offset somewhat when diffraction limits are encountered because of the shorter wavelengths of the ions but this is unlikely to allow te performance of 0ver-

lapping lens/deflection systems to be equalled. Secondly, the effects of particle to particle Coulomb interactions are exaggerated. Thirdly, ion sources have poorer performance in terms of chromatic spread, brightness and/or total current, and finally, ions do not penetrate resist layers to detect alignment marks as do electrons.

There is one significant advantage and that is that exposure with ions does not suffer from the deleterious proximity effect encountered with electron exposure. Ions do not penetrate the substrate deeply and hence are not backscattered through the resist into areas away from their intended point of exposure. However, the lack of penetration means that the resist is not exposed uniformly throughout its depth unless high (>100 kV) accelerating voltages are used. If the energy of the ions is increased until the ions do penetrate through the resist into the substrate then unacceptable damage of the sample may occur. This is not a problem for mask writing, and for direct-writing with surface imaging resists, but to date few have considered the advantage of proximity free exposure to be significant enough to override the shortcomings of ion optical systems.

The ultimate resolution of ion optics is more than 20 times worse than it is for electron optics, ~ 5 nm versus ~ 0.2 nm because of the higher aberrations of the electrostatic lenses and the larger energy spread of the sources. This difference is only marginally important for conventional resist exposure, as the resolution of resist itself is about 10 nm, but for inorganic resists, where dimensions well below 10 nm can be obtained, the ion optical resolution limit will become the fabrication limit.

Ion beam systems have been shown to be valuable for mask and circuit repair, particularly for the removal of unwanted material, but for direct-write lithography the systems proposed do not seem to offer any advantage over their electron equivalents either in scanning or projection configurations.

8. Local Probes

The difficulty of writing the complex patterns for tomorrow's nanostructures becomes quite clear when one considers what would be required if local probes (as used in Atomic Force and Scanning Tunnelling microscopes) were to be used for this task. It has been shown that individual atoms can be manipulated and gold deposited selectively with local probes and it has been suggested that this method might be used for nanostructure patterning. It would certainly solve all difficulties with respect to resolution, provided of course that suitable processes could be identified, but the writing rate would present severe difficulties. At present the maximum patterning rate has only been a few points per second but supporters of the concept suggest that it might be possible to write at lateral tip velocities approaching 1 cm/second, corresponding to 10^6 pixel/sec for 10 nm pixels. A nanometre device 'chip' will have about 4×10^{12} pixels, therefore, to expose it in 2 seconds, which is the rate required for a throughput of 30 20 cm wafers per hour, 2×10^6 tips will have to be used simultaneously. A 20 cm

wafer will contain about 60 chips, so if there are $2x10^6$ tips per wafer 33,000 will be availably to write each chip. The difficulty will be that the tips will have to be fabricated on 30 μm centres. Even more difficult would be the need to keep the cost per tip, including the control and data-handling system, to less than \$10 so that the overall system cost would not exceed \$20M. Neither of these requirements would seem to be feasible today.

The likelihood that local probes will prove useful for mass-production of nanometre scale devices would therefore seem remote. For scientific work, however, their unique ability to manipulate single atoms makes them ideal for fabricating simple test structures. For example, Eigler and his coworkers have created 'quantum corrals' by placed single atoms in a ring with an accuracy of less than a tenth of a nanometre[19]. This remarkable resolution could not have been approached by any other method.

9. Conclusions

If electronic devices with nanometre dimension are to be manufactured in the volumes necessary to rival silicon integrated circuits then a new lithography approach is needed. A wide variety of methods are being explored which meet some of the cost and pattern complexity requirements but X-ray lithography alone appears capable of satisfying them in the foreseeable future and even then only for dimensions down to about 0.07 μm.

Electron projection methods can meet the resolution requirements easily but the system complexity required to achieve practicable throughput will probably make this approach too expensive and in any cases such systems will take years to develop. Ion beam methods will be even more expensive and difficult with the difficulties out-weighing their only advantage, which is lack of long-range proximity effect. In the very long range, X-ray projection cameras may produce 0.05 μm capability but it may take decades to perfect the methods needed to fabricate the mirror lenses. Multi-electron beam columns may also offer a solution but precise control of hundreds of beams presents a multitude of challenges.

In summary, it is not possible to predict that it will ever be possible to mass-produce nanometre-scale structures particularly below the cut-off for X-ray lithography which falls at about 0.07 μm.

Table 1 summarises the resolution limits of the different lithography methods. The 'Practical Limit' is the limit that should be achieved in production, and the 'Ultimate Limit' is that which may be achieved in the laboratory under ideal conditions. Table 2 compares the different electron beam nanofabrication and nanostructuring methods.

The work discussed in this paper has been described by many workers in publications that are too numerous to cite comprehensively. Further reading is available

in the proceedings of the International Symposia on Electron Ion and Photon Beams[20] published each year in the Journal of Vacuum Science and Technology, the proceedings of the International Conferences on MicroProcess published each year in the Japanese Journal of Applied Physics[21] and the proceedings of the Microcircuit conferences published by Elsevier[22].

10. References

1. Okazaki, S. (1991) *J. Vac. Sci. Technol.* **B 9** (6), pp. 2829-2833
2. Coopmans, F. and Roland B. (1987) *Solid State Technology*, June, p 93.
3. Maaike Op de Beeck and Luc Van den Hove, (1992) *J. Vac. Sci. Technol.* **B 9 (6)**
4. Tennant, D.M. et al., *J. Vac. Sci. Technol.*, (1991) **B 9** (6), pp. 3176-3183
5. Kurihara, K., Kinoshita, H., Mizota, T., Haga, T. and Torii, Y. (1991) *J. Vac. Sci. Technol.* **B 9** (6), pp. 3189-3192
6. White, D.L. (1991) *Solid State Technology,* July 1991, p. 37
7. Spears, D.L. and Smith, H.I. (1972) *Electronics Lett.*, **8**, p. 102-104
8. Smith, H.I. and Schattenburg, M.L. (1992) *Proc. SPIE Symposium on Microlithography,* March 8-13, San Jose, CA.
9. Pfeiffer H.C. (1978) *J. Vac. Sci. Technol.*, **15**, p 887
10. Pfeiffer H.C. and Groves, T.R. (1991) *Microelectronic Engineering 13*, p. 141
11. Sohda, Y. et al. (1991) *J. Vac. Sci. Technol.* **B 9** (6), pp. 2940-2943
12. Hattori K. et al. (1993) *J. Vac. Sci. Technol.* **B 11** (6), pp. 2346-2351
13. Sakamoto, K. et al. (1993) *J. Vac. Sci. Technol.* **B 11** (6), pp. 2357-2361
14. Yasuda, H. (1993) *Proc. Microprocess Conference*
15. Chang, T.H.P., Kern, D.P. and Muray, L.P. (1992) *J. Vac. Sci. Technol.* **B10 (6)**, pp. 2743-2748
16. Heritage, M.B. (1995) *J. Vac. Sci & Technol.*, **12**, p. 1135
17. Berger, S.D. and Gibson, J.M. (1990) *Appl. Phys. Lett.*, **57**, p. 153
18. Berger, S.D. et al. (1993) *J. Vac.Sci. Technol.*, **B 11** (6), pp. 2294-2298
19. Crommie, M.F., Lutz, C.P. and Eigler, D.M. (1993) *Nature,* **363**, 524-527
20. Proceedings of the International Symposia on Electron, Ion and Photon Beams, *J. Vac. Sci. Technol.*, American Institute of Physics, New York.
21. Proceedings of the International Symposia on MicroProcess, *Jpn. J. Appl. Phys.*, published by Japanese Journal of Applied Physics, Tokyo.
22. Proceedings of the Microcircuit Engineering Conferences, Elsevier, Amsterdam, London, New York and Tokyo.

Table 1. Present and ultimate capability for systems under development for chip lithography.

LITHOGRAPHY TYPE	PRESENT CAPABILITY Minimum feature size Throughput (8" waf/hr) & Cost	ULTIMATE CAPABILITY Minimum feature size Throughput (8" waf/hr) & Cost	ABILITY TO MASS-PRODUCE NANOSTRUCTURES
U.V. Projection	0.35 μm features 20 - 40 8" wafer/hour \$1.5M - \$5M	0.1 - 0.15 μm features Adequate throughput >\$5M	None, because of inadequate resolution
Extended-UV Projection	0.1 μm features Negligible throughput Laboratory test equipment	0.04 μm - 0.06 μm Adequate throughput ?? \$10M?	Must await demonstration of extreme mirror tolerances (~0.1 nm)
X-Ray Proximity printing	0.2 μm features ~20 8" wafer/hour/station ~\$80M for 20 station facility	0.05 - 0.07 μm features Adequate throughput \$10M? per station	Should eventually be capable for dimensions >0.05 μm
Electron - Scanning Round beam	0.01 μm - 0.25 μm features 10^{-4} - 0.1 8" wafer/hour \$1.5M - \$5M	0.01 μm unless higher resolution fabrication processes are discovered	None, because of low throughput and high cost
Electron - Scanning Shaped beam	~0.25 μm features ~2 8" wafers/hour \$4M -\$12M	?0.04 μm features <0.1 8" wafers/hour ?	May eventually be capable but only at very low throughput
Electron - Scanning Character projection	~0.25 μm features ~10 6" wafers/hour ?\$10M	?0.04 μm features <1 8" wafers/hour ?	May eventually be capable but only with a limited menu of shapes
Electron - Projection	0.1 μm features Negligible throughput Laboratory test equipment	?0.04 μm features ?Marginal throughput ?High cost	May eventually be capable but extreme complexity will lead to high cost
Ion - Projection	0.1 μm features Negligible throughput Prototype equipment only	?0.04 μm features ?Marginal throughput ?High cost	May eventually be capable but extreme complexity will lead to high cost

LITHOGRAPHY TYPE	PRACTICAL LIMIT	ULTIMATE LIMIT
U.V.Light Contact and Proximity	2.5 μm Fresnel diffraction at minimum practicable gap and wavelength Wavelength 200 nm, Gap 25 μm	0.125 μm Fresnel diffraction for contact print with 100 nm thick resist Wavelength 160 nm, Gap 100 nm
X-ray Proximity	0.1 μm Absorber aspect ratio and Fresnel diffraction at practicable gap for step and repeat operation Wavelength 10 nm, Gap 10 μm	0.01 μm Fresnel diffraction for contact print and resist resolution limit Wavelength 1 nm, Gap 100 nm
Ultra-violet Light Projection	0.15 μm Fraunhofer Diffraction at N.A. set by fabrication and field-size limits With phase shift mask etc. Wavelength 157 nm, NA 0.75 Depth of focus 0.3 μm	0.1 μm Fraunhofer diffraction at wavelength set by transparency of resist and optical materials Wavelength 157 nm, NA 0.9 Depth of focus 0.2 μm Field size < 200 μm
Soft X-ray Projection	0.05 μm Fraunhofer diffraction at NA set by achievable tolerances on optical components Wavelength 13-15 nm, NA 0.1 Depth of focus 1.4 μm	0.01 μm Fraunhofer diffraction and resist resolution limit Wavelength 2-5 nm, NA 0.2 Depth of focus 0.07 μm Field size ??
Electron Beam	0.03 - 0.05 μm Lateral scattering of electrons in resist and/or interaction range of exposure process Accelerating voltage 100 kV (In practice resolution is more often limited by the beam size/current as determined by throughput needs)	0.007 - 0.02 μm Resist Delocalisation of exposure (2ndary electrons/inelastic range) 0.001 - 0.005 μm Direct exposure/sublimation Combination of electron interaction range and electron optical limits (diffraction & spherical aberration)
Ion Beam	0.03 - 0.05 μm Ion optical limits (chrom. aberr.) and interaction range with resist (In practice resolution is more often limited by the beam size/current as determined by throughput needs)	0.01 - 0.02 μm Resist Delocalisation of exposure (2ndary electrons/inelastic range) and ion optical limits (Chromatic aberration) 0.01 μm Ion Milling Delocalisation of sputtering process

Table 2. Practical and ultimate resolution limits for lithography methods.

Active Packaging: a New Fabrication Principle for High Performance Devices and Systems

SERGE LURYI
State University of New York at Stony Brook
Stony Brook, NY, 11794-2350 USA

1. Introduction

In this work I shall discuss a new device fabrication principle which I would like to refer to as *active packaging* (AP). The meaning of this term is that certain essential fabrication steps (lithography, etching, metallization, etc.) are performed *after* the partially processed device or circuit is packaged onto a host platform.

One of the most important goals of the AP concept is the combination of dissimilar materials (notably, III-V compound semiconductors) with silicon integrated circuitry (IC) on a single Si substrate [1]. This goal, now widely recognized as an important research direction in microelectronics, is shared by other emerging technologies, such as those based on *heteroepitaxial* and *thin-film transfer* techniques [2]. At the same time, AP *widens* significantly the class of device structures that can be manufactured. Our ultimate goal is not only to "teach the old dog new tricks" but also to greatly expand the assortment of tricks available.

It is worth stressing that the word "packaging" is used somewhat unconventionally in the AP context. Active packaging is a device fabrication technique, intended to implement devices on a foreign (not necessarily even semiconductor) platform that perform better than conventionally fabricated devices on their natural semiconductor substrates. In many instances, AP enables the implementation of structures that cannot be realistically obtained in another way, such as those requiring lithography on *opposite* sides of a thin semiconductor film.

The principle of active packaging will be illustrated in the instance of a heterostructure bipolar transistor (HBT) structure, schematically shown in Fig. 1. Such a structure would reduce the parasitic capacitance between the base and the collector electrodes, enabling ultrafast operation with oscillation frequencies in the range of 300-400 GHz and even higher. This in turn would open up the possibility of implementing on-chip millimeter-wave phased-array antenna systems [3].

S. Luryi et al. (eds.), Future Trends in Microelectronics, 35–43.

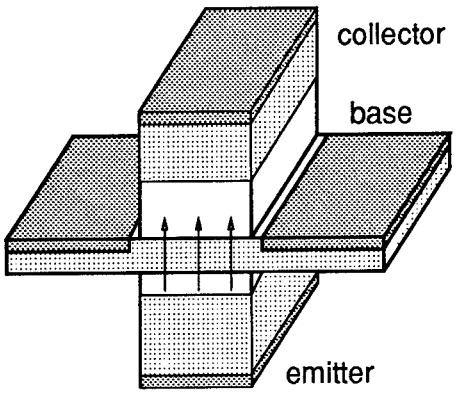


Fig. 1. Idealized cross-section of a heterostructure bipolar transistor to be fabricated by active packaging. Dotted pattern indicates the conducting (undepleted) layers, shaded pattern the contact metal. The emitter and the collector stripes are defined by independent lithographic steps and aligned to each other. If the upward direction is defined as that outward from the surface on which the lithography is performed, then both the emitter and the collector can be considered "up".

2. Active Packaging HBT Process

The process is based on flip-chip packaging, removal of the substrate, and backside lithography. Fabrication begins with the following epitaxial structure:

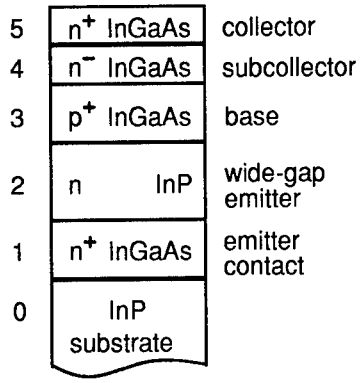


Fig. 2. Epitaxial structure of InGaAs/InP HBT to be fabricated by active packaging. The emitter contact layer 1 is also an etch stop for InP. For etch-stop reliability it may be convenient to include a pair of sacrificial InP-on-InGaAs layers between layers 0 and 1, so that the etching is performed in steps and results in uncovering an ideal flat surface of the emitter contact.

Top side processing includes etching of the collector stripe down to the base layer, evaporation of self-aligned contacts to the base, deposition of a passivating dielectric, etching of via holes and metallization. At this point all the base and collector contacts are connected in a circuit with lines running over the passivating dielectric. The circuit is then covered by another ("interlevel") dielectric layer and a relatively small number of selected points of the circuit are connected to "top" metal pads through a second set of via holes. The top ("communication") pads may be relatively wide (e.g., $\gtrsim 100\ \mu m$). The interlevel dielectric (e.g., polyimyde) may be planarized. No attempt is made at this stage to contact the emitter.

Flip-chip mount; "consulator" film. The circuit is then mounted on a "carrier" wafer which has a mirror pattern of metallic communication pads. The carrier may be any substrate, including glass, ceramics, etc., but first and foremost a silicon wafer that has already undergone the integrated circuit processing. Connection between communication pads is established with the help of an anisotropically conducting film with electrical properties, illustrated in Fig. 3. The film must provide a short between overlapping contacts and an open circuit otherwise. Such vertically conducting and laterally insulating films, which may be called "consulators", can be prepared in a variety of ways. The primary purpose of the consulator, besides providing vertical electrical connections, is to provide a stable mechanical support for the packaged chip – support that will become crucial when the InP substrate is removed.

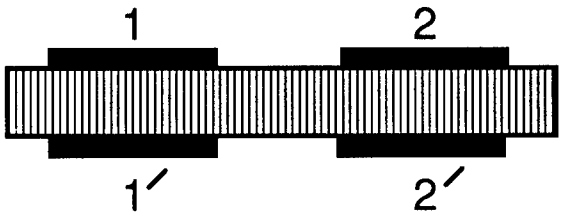


Fig. 3. Required "consulator" film. For overlapping electrodes the film provides a short, $R(11') \approx R(22') \lesssim 10\,\Omega$, while nonoverlapping electrodes form an open circuit, $R(12) \approx R(12') \gtrsim 10^8\,\Omega$.

In principle, a perfectly adequate consulator can be provided by the solder-bump technology. One needs an adhesive dielectric that can flow to fill the narrow spacing between the chip and the carrier wafer and then stiffen to provide the necessary mechanical support. Another possible approach is to use the existing packaging technology of anisotropically conductive adhesive films, used in liquid crystal display assemblies. These materials are not intrinsically anisotropic, they conduct in a preferred direction only after having been processed, Fig. 4.

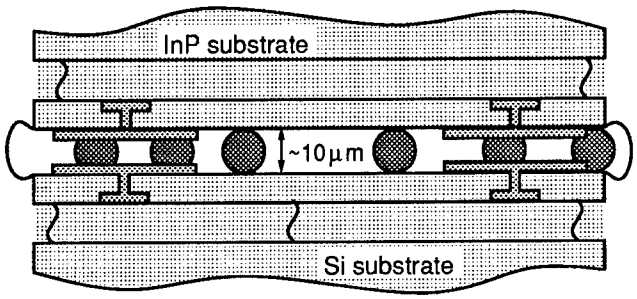


Fig. 4. Anisotropically conductive adhesives (after Ref. [4]). Small conducting spheres are dispersed in an epoxy matrix at a concentration below the percolation threshold for electric conductivity. The assembly is compressed until hard spheres touch both surfaces (extra epoxy oozes out). Electrical connection is established only vertically, since neighboring spheres rarely touch.

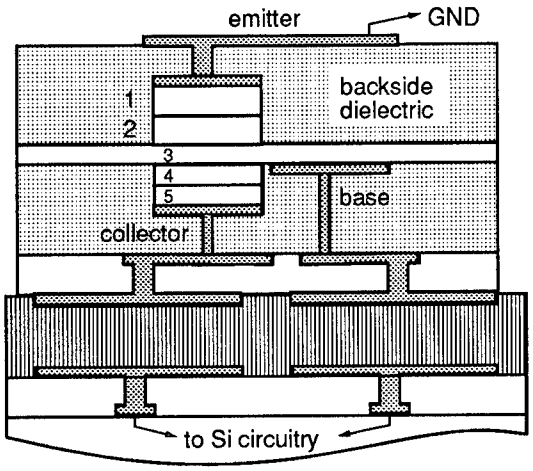


Fig. 5. Cross-section of the final assembly after active packaging. Emitter stripe is aligned to and slightly overlaps the collector stripe. It is assumed for simplicity that all HBT's are used at common emitter, so that the emitter final metal is connected to the ground of the Si circuit. Of course, this may not be true in general, and the emitter final metal pattern may be more complicated. Connection to the rest of the circuit may go across the periphery of the chip, or via the consulator film.

Substrate removal and back-side processing. It is quite possible to etch the entire InP substrate down, stopping at a 0.1 μm InGaAs layer. This step is based on the well-known extreme selectivity between the etch rates for InP and InGaAs in hydrochloric acid solutions. It is essential that the uncovered surface of layer 1 is uniformly flat, adequate for performing a fine line optical lithography. A large hole etched from the substrate side would not do, because there would be problems with focal depth. To make lithographic alignment to the base contact level, the contact metal should be seen with a sufficient contrast through layers 1-3. If this proves to be inconvenient, then special topography features for the back-side alignment must be provided at the top-side processing stage.

The final assembly is illustrated in Fig. 5. The emitter contact is established by a standard lift-off evaporation of a suitable metal. It is well known that ohmic contacts to n^+ InGaAs are good without alloying. No elevated temperature procedures should be contemplated after the chip has been mounted, because of the limited thermal stability that can be expected of a consulator film and the need to preserve the integrity of fully processed Si integrated circuits on the carrier wafer.

3. Advantages

The reduced base-collector capacitance offers significant advantages for microwave performance of HBT. An enhancement of the maximum oscillation frequency f_{max} by a factor of 2 to 3 has been predicted [5,6] over optimized collector-down structures. Moreover, with a suppression of the extrinsic collector capacitance C_{Cx} it becomes possible to implement HBT structures with *coherent* effects in the base [7,8], resulting in a power gain above the conventional cutoff frequencies. Figure 6 shows the modeled microwave characteristics of a collector-up HBT, in which the base bandgap is graded so that the total base propagation delay τ is much shorter than the diffusive delay in a flat base of the same width [8]. The magnitude of the base transport factor $\alpha = |\alpha| \exp(-2\pi i f \tau)$ decreases so slowly with increasing frequency f that it becomes feasible to activate transit-time resonances far above $f_T \approx 1/2\pi\tau$. The fundamental peak in U occurs near πf_T [7,8].

The coherent transistor can be designed to have the first high-gain peak at any desired frequency, provided the effect is not destroyed by the parasitics. For frequencies below 100 GHz it is possible to use conventional structures (Fig. 6a), where typically the extrinsic (parasitic) collector capacitance C_{Cx} is about twice the intrinsic (useful) capacitance C_C. In order to push the peak into a sub-millimeter wave range (Fig. 6b) it is essential to reduce C_{Cx} below C_C. The AP process opens a way to substantially reduce the extrinsic capacitance.

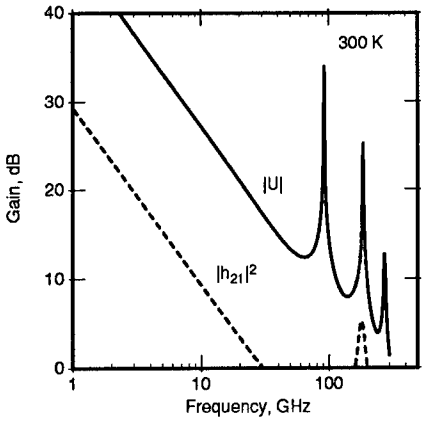


Fig. 6a. Common-emitter current gain $|h_{21}|$ and the unilateral power gain $|U|$ of a model coherent transistor with a special graded-gap base design, optimized for stable oscillation at 94 GHz. Base total width $W = 1\,\mu m$. Transistor is assumed loaded with the parasitics with state-of-the art equivalent circuit parameters, e.g. $C_{Cx} = 2C_C$. Conventional current-gain cutoff is $f_T \approx 32\,GHz$, however the transistor also exhibits a range of current gain $|h_{21}| > 1$ at $f \approx 2\pi f_T$ (near the second peak in U). The fundamental peak in U occurs near πf_T (after Refs. [7] and [8]).

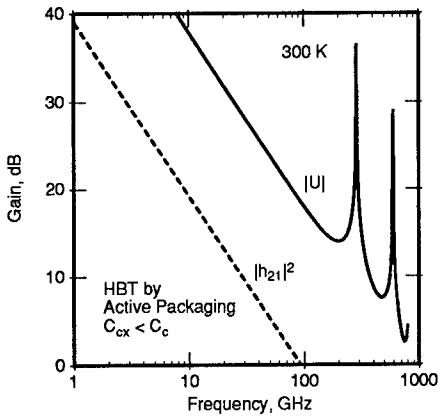


Fig. 6b. Reduction in the parasitic capacitance C_{Cx} by the active packaging process enables a more aggressive design of the coherent transistor, optimized for stable oscillation at near 300 GHz. Base total width $W = 0.3\,\mu m$. Transistor is assumed loaded with the parasitics with state-of-the art equivalent circuit parameters, except C_{Cx} which is assumed small, $C_{Cx} \lesssim 0.5\,C_C$. The minority-carrier diffusivity $D \approx 25\,cm^2/s$. The base transit time (by drift) is $\tau_B = 1\,ps$ and collector transit time $\tau_C = 0.75\,ps$, resulting in a conventional $f_T \approx 100\,GHz$.

An interesting further advantage of an AP HBT is the possibility of accommodating a Schottky collector, Fig. 7. An important parasitic resistance in small area devices is due to the metal semiconductor junction. Since the resistance of an ohmic contact scales with its area, at small enough dimensions it must dominate other resistances that scale with the contact perimeter. It therefore makes sense to dispense with the n^+ doped semiconductor layer in the collector (layer 5 in Fig. 2). Such an approach has been successfully used in the fabrication of submicron resonant tunneling diodes [9]. The high thermal conductivity of a metallic layer and its proximity to the n^- base-collector field region, where most of the heat is generated, is another important advantage of a Schottky collector. Needless to say, a Schottky collector is realistic only in a collector-up configuration. Its implementation is entirely compatible with the AP process.

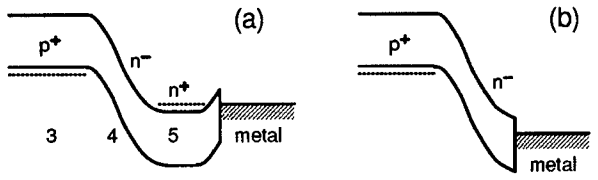


Fig. 7. Illustration of a conventional ohmic semiconductor-metal contact (*a*) and a Schottky collector (*b*). Numbers in (*a*) refer to the layer labeling scheme in Fig. 2. In a Schottky-collector device, layer 5 would be eliminated altogether and a suitable metal deposited directly on the lightly-doped layer 4, completely depleted of mobile carriers.

4. Applications

Based on AP HBT technology it is entirely feasible to implement local oscillators and amplifiers that operate at millimeter and even submillimeter wavelengths. One obvious application of such devices would be for satellite communication systems in the atmospheric transmission window of 345 GHz.

Another extremely attractive application [3] is the possibility of fabricating millimeter-wave *phased arrays on a silicon chip*. A $\lambda/2$ spaced linear array of 20 elements radiating at 300 GHz would be about a centimeter long. The advantage of having transistor oscillators is that the millimeter-wave beam can be electrically steered off broadside by controlling the relative amplitudes of different oscillators, while their relative phases are locked together by the evanescent wave interaction. The point is that most available phase shifters used in centimeter wave phased array systems are bulky elements that cannot be used in on-chip designs. Instead, we should use electronic beam steering by controlling the *amplitude* of constant-phase

array elements [10]. As far as I am aware, this idea has not been employed in practical phased-array antenna systems, perhaps because at centimeter wavelengths it is more efficient to control the relative phases of array elements. In the millimeter and submillimeter wavelength range amplitude steering appears to be the only realistic way to implement purely electronic beam steering. Three-terminal devices are ideally suited for this purpose. On-chip focal plane antenna arrays should have important applications as steerable radar systems in avionics, automated manufacturing, and especially in automobile collision avoidance and early warning systems.

5. Conclusion

Generality of the active packaging principle transcends microwave transistor applications. The new degree of freedom in manufacturing – lithography on opposite sides of a thin film – permits the implementation of a variety of new devices and functions. Of the many possible examples, let me mention the possibility [11] of fabricating a collector-up charge injection transistor with the channel-defining trench etched in side 2 and aligned to the collector stripe on side 1. To indicate the scope of contemplated applications, let us note that the technique makes feasible an active directional coupler in which two edge-emitting laser resonators overlap in a portion of their length. It also simplifies many schemes for integrating electronic and photonic devices into a single functional unit to be placed within the integrated circuitry on a silicon chip.

Vertical cavity surface emitting lasers (VCSEL) should have an important role in this program, because such elements enable interchip communication directly from the chip interior. Active packaging technology offers several advantages in the integration of VCSELs with silicon VLSI. For example, it allows to use *non-epitaxial* Bragg mirrors (such as stacks of ZnS and SiO_2 layers) not only for the top but also for the bottom mirror of a VCSEL cavity. Also it permits the implementation of *tandem systems* in which one VCSEL (master) works under electrical injection of carriers while the other (slave) is optically pumped by the former.

I believe that most significant applications of compound semiconductor electronics will be associated with its use in silicon electronics. In terms of the old debate on Si vs GaAs, my view is that silicon is the ultimate customer for GaAs. The logic of industrial evolution will motivate new paths for a qualitative improvement of system components, other than the traditional path of a steady reduction in fine-line feature size. The principle of active packaging, illustrated in the present work using the instance of implementing ultra-high performance InP HBT on a silicon chip, will become one of the central design principles of future microelectronics.

References

1. Luryi, S. and Sze, S. M. (1984) The Future of Microelectronics — Hybrid Material Systems of IV/III-V Compound Semiconductors, *AT&T Bell Laboratories Technical Memorandum* 52111-840920-01.

2. Deboeck, J. and Borghs, G. (1993) III-V on Si — heteroepitaxy versus lift-off techniques, *J. Cryst. Growth* **127**, 85-92.

3. Luryi, S. (1994) How to make an ideal HBT and sell it too, *IEEE Trans. Electron Devices* **TED-41**, 2241-2247.

4. Lyons, A. M. and Dahringer, D. W. (1993) Electrically Conductive adhesives, in K. Mittal and A. Pizzi (eds.) *Handbook of Adhesives Technology*, Marcel Dekker, New York.

5. Kroemer, H. (1982) Heterostructure bipolar transistors and integrated circuits, *Proc. IEEE* **70**, 13-25.

6. Fonstad, C. G. (1984) Consideration of the relative frequency performance potential of inverted heterojunction n-p-n transistors, *IEEE Electon Dev. Lett.* **EDL-5**, 99-100.

7. Grinberg, A. A. and Luryi, S. (1993) Coherent Transistor, *IEEE Trans. Electron Devices* **ED-40**, 1512-1522.

8. Luryi, S., Grinberg, A. A., and Gorfinkel, V. B. (1993) Heterostructure bipolar transistor with enhanced forward diffusion of minority carriers, *Appl. Phys. Lett.* **63**, 1537-1539.

9. Allen, S., Reddy, M., Rodwell, M., Smith, R., Liu, J., Martin, S., and Muller, R. (1993) Submicron Schottky-collector AlAs/GaAs resonant tunnel diodes, 1993-IEDM *Tech. Digest*, 407-410.

10. Costas, J. P. (1981) An antenna beam steering technique comprised of constant-phase array elements, *Proc. IEEE* **69**, 745-747.

11. Luryi, S. (1994) Article Comprising a Real-Space Transfer Semiconductor Device and Method of Making the Article, US Pat. **5,309,003**.

THE WIRING CHALLENGE: COMPLEXITY AND CROWDING

T.P. SMITH III, T.R. DINGER, D.C. EDELSTEIN,
J.R. PARASZCZAK, T.H. NING
IBM Research Division, T. J. Watson Research Center,
P. O. Box 218
Yorktown Heights, NY 10598

Abstract

The complexity of state-of-the-art silicon technology places great demands on the performance capability of each of the components of the circuit, from cell isolation to wire-bond pad. High on the list of critical performers is the interconnect wiring, which does not benefit from miniaturization as do devices. In fact, wiring performance worsens with ground rule reduction and die size increases. We summarize state-of-the-art on-chip wiring process technology as it pertains to microprocessor performance, suggest future trends for microprocessor performance, and indicate both technology and design solutions for the wiring bottlenecks created by today's most demanding processor performance requirements. In addition, we address the feasibility of novel future wiring technologies for solving issues which plague conventional VLSI designs.

1. Introduction

The economics of the semiconductor industry are such that increases in productivity have come at a rate experienced by no other industry in history except perhaps those transformed by the advent of the steam engine. This exponential rate of productivity growth has been driven by the availability of new generations of lithography tools which have ratcheted down device ground rules (G/R's) at regular intervals.

An important ingredient of this scenario is the capacity of VLSI technology to show overall performance improvements with the trend toward increasing miniaturization. A key component which has not benefited from this miniaturization is VLSI interconnect technology, which is becoming increasingly hard-pressed to deliver device performance in the face of its own inability to provide performance improvement through scaling and increasing complexity brought on by ever increasing circuit densities and chip sizes.

These problems are not new, but they are becoming debilitating to VLSI designers as microprocessors approach 3 ns cycle times [1]. In fact, much of the proposed solution has been perceived for many years [2]. The sobering fact remains that nothing short of a revolution in VLSI technology will provide a solution to the interconnection dilemma. New materials such as copper wiring and low-epsilon dielectrics can and do provide some relief, but a renewable source of performance improvement for VLSI interconnects is not on the horizon. Our intention is to examine current interconnect

S. Luryi et al. (eds.), Future Trends in Microelectronics, 45–56.

technology in light of trends occurring and expected to occur within the semiconductor industry and to discuss the potential of those alternative and future technologies which may relieve some of the current wiring burden.

2. Current VLSI Wiring Technology Capability

Much of the robustness of current VLSI technology can be credited to a very solid and extendable multi-level metal (MLM) technology. While this technology has benefited immensely from rather recent breakthroughs in planarization (e.g. chemical-mechanical polishing), many of its components are old stalwarts that have been reengineered through several generations.

Some of the attributes that have benefited the Al/SiO_2 interconnect technology are generally low line resistance, highly stable and rather low permittivity dielectric, metal electromigration resistance, and overall compatibility with device processing limitations.

2.1 PROCESS DESCRIPTION

Shown in Figure 1 is a schematic cross-section of a state-of-the-art VLSI technology. An interesting note is that, although the figure is not drawn to scale, an inordinate amount of the volume (and therefore cost) of the structure is dominated by the interconnect structures (shown as solid and shaded areas). It is estimated that for a complex VLSI technology, such as that used for a CMOS microprocessor, fully half of the manufacturing cost and more than half of the yield loss can be attributed to the so-called "back end of the line" (BEOL) or interconnect process.

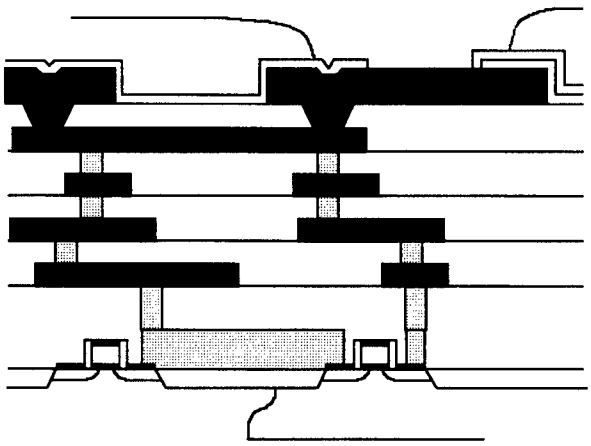


Figure 1. Cross-section of planarized multi-level VLSI interconnect technology.

A rather state-of-the-art description of the process sequences needed to construct such a 3-dimensional wiring structure is given in Table 1. Such a process is typically

called a TACT-TiN process in reference to the materials set involved: Al(Cu) alloy line metallurgy (AC), Ti layers (T) over and under this line for electromigration resistance, and TiN antireflective coating. The process summarized in Table 1 also includes tungsten studs lined with PVD Ti/TiN, and a PECVD TEOS SiO_2 dielectric.

It should be pointed out that the rather long sequence of process steps associated with the line/stud portion of the build may need to be repeated 4-5 times for a high performance microprocessor design. It is quite easy to understand where the cost and yield loss due to the BEOL originates.

Table 1. TACT-TiN Wiring Process Sequence

• PECVD TEOS PSG Deposition • Lithography • Oxide RIE Etch • Ti/TiN PVD Liner Deposition • CVD W Deposition/CMP	Contact Levels
• PVD Ti Deposition • PVD Al(Cu) Deposition • PVD Ti/TiN Deposition • Lithography • Metal RIE • Metal Anneal • PECVD TEOS Oxide Deposition/Etch • Insulator CMP • Lithography • Insulator RIE • PVD Ti/TiN Deposition • CVD W Deposition/CMP	Line/Stud Levels
• Terminal Metals Processing	Repeat

2.2 RC AND CURRENT DENSITY ISSUES

It is interesting to remind ourselves of where the interconnect performance bottleneck due to scaling originates. This has been treated by a multitude of authors, but bears repeating [1, 3-4].

Ground rule scaling from one VLSI generation to the next can be described by the scaling factor S ($S > 1$). Independent of this G/R scaling are die size increases which have typically been occurring with time due to manufacturing improvements. These die size increases can be described by a separate scaling factor ($S_c > 1$). The scaling factors influence interconnect performance in several ways.

First and foremost is the impact on interconnection RC delay which is increased by the factor $S^2S_c^2$ due to the shrinking cross-sectional dimension of the wire and its increasing length. Similarly, current density in the interconnect scales by the ideal scaling factor S. Since reliability for interconnects (resistance to electromigration) is given as a mean-time-to-failure (MTTF) related to current density (J) and interconnect dimensions (W_{int}, H_{int}) by the relation:

$$MTTF \propto W_{int} H_{int} e^{E_a/k_B T} / J^n \quad (1)$$

where n is found experimentally to typically be 2, we find that the reliability of interconnects to electromigration is degrading from generation to generation by a factor of S^4! This result is cause for alarm.

3. Future Trends for Microprocessors

It is fair to say that complexity in the design of state-of-the-art high performance microprocessors is approaching a point of diminishing returns. The numbers of circuits required has risen steadily in the past several generations of machines to the point where reducing the number of cycles per instruction requires an unjustified increase in the total circuit count. This situation taxes the CAD tools available to designers and increases the likelihood of fatal design flaws.

It seems likely that this set of circumstances will lead to the following trends in microprocessor design. First, the overall circuit counts in the leading-edge processors, which is used as a gauge of machine complexity, will begin to plateau [1]. If significant functionality were missing from today's processors, this trend would not obtain. However, we believe that sufficient functionality is present in today's processors to cause designers to focus their energy at cycle time reductions in order to improve performance.

This leads us to the second trend that we expect in future CMOS microprocessors. Assuming that the current trend in overall machine performance continues, which it no doubt will, time-of-flight delays in interconnect wiring will begin to be of importance. In order to drive cycle time down, designers will be forced to consider smaller chip footprints in order to keep time-of-flight delays manageable.

Requirements for on-board cache will continue to increase in order to drive cycle times down. This obviously tends to increase chip footprints so that optimizing design for on-board cache *vs.* chip size will be necessary. We believe that in general, chip sizes will begin to plateau in size even as additional on-board cache is required.

Finally, the market for portable computers will drive the industry to lower voltage technologies, as is already occurring. While this requires extreme tolerance control in the devices themselves, it also in creates constraints on the linewidth control for interconnect wiring. Cross-chip and cross-wafer linewidth, difficult to control in today's technology, will become even more difficult control as throughput increases and linewidth tolerances decrease.

4. Limits of Wiring Technology

Much of what is present in today's VLSI wiring technology is perfectly adequate for the bulk of designs. As microprocessors become more and more pervasive in consumer goods this trend will increase. It is only with the highest performing microprocessors, those limited to a very small percentage of the overall application space, that we run into performance difficulties due to the interconnect technology. Unfortunately, or perhaps not, most semiconductor manufacturers attempt to cover all of their design space with a single technology, since semiconductor technology is fairly expensive

either to license or develop. For companies with a very narrow range of products, this may not be a problem. For companies with large product scopes, this situation typically results in a high performance technology being developed which is eventually amortized by incorporation into lower performance chip designs. The following discussion will treat such a high performance technology and will emphasize those technology elements which act as bottlenecks to overall system performance.

4.1 LOSS OF SCALING

Even in the best of circumstances, we have seen that ideal scaling results in RC delays that are proportional to $S^2S_c^2$. This loss of performance due to scaling occurs purely from line length increase and cross-section reduction. The really bad news is that today's interconnect technologies, which contain parallel path conductors (cladding) for redundancy to safeguard against metal voiding, do not even scale this well! This is simply the result of the interconnect dimensions decreasing to the point where the resistivity of the cladding layer itself becomes important. Figure 2 illustrates the cladding situations encountered for both aluminum and copper wiring and the effective resistivity that results as G/R's are shrunk. The curves labeled "future" are predictions of where the technology might end up as new or thinner cladding layers are implemented.

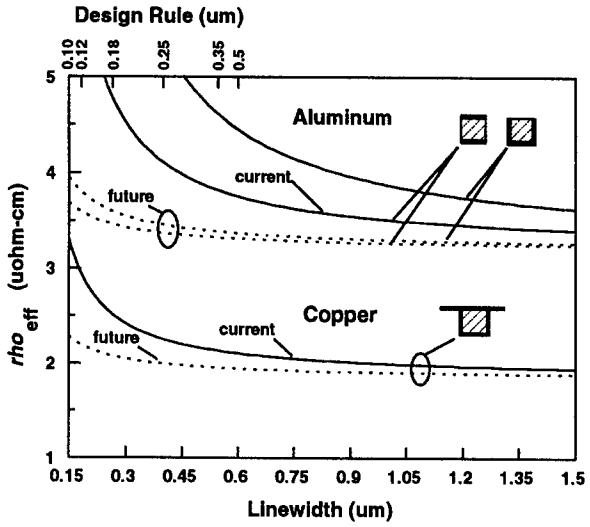


Figure 2. Effective resistivity (*rho*$_{eff}$) *vs.* linewidth for square cross-section aluminum and copper interconnects.

Another interesting ramification of increasing circuit counts in high performance processors is that the capacity of a given technology to wire a given design comes into question. This topic has been treated by many authors and will be discussed elsewhere in this volume. Suffice it to say that as complexity increases, more and more wiring levels will be needed to ensure the wireability of the design. Such an effort has diminishing returns, however, as the lower-lying levels block wiring paths to the devices so that the

upper levels have reduced utility. In addition, increasing the number of levels in an interconnect technology does not come free. Increasing production costs combined with lower overall chip yields will cause any production manager worth his salt to cast a very suspicious glance at technologies which support increasing numbers of wiring levels.

Another result of loss of resistivity scaling is that the aspect ratio of VLSI wiring has tended to increase in order to compensate for decreased horizontal G/R's. While this results in obvious benefit to the wiring resistance, the effects on process complexity, wiring capacitance, and noise immunity are not so beneficial. The most noticeable of these effects is that in-plane cross-talk increases. As noise tolerances will decrease with the push to low-power technologies, high-aspect-ratio interconnects will demonstrate increasingly unacceptable performance.

Finally, the requirement for 4-5 levels of wiring makes a planarizing technology extremely attractive. Depth-of-focus constraints in current lithography tools make planarization desirable. Unfortunately, planarization of topography can result in a more or less debilitating situation upon subsequent planar metal deposition ... where did the alignment marks disappear to? The better the planarization process, the worse the subsequent alignment process through the blanket metal. Additional (expensive) processing either before or after metal deposition makes the process possible, but this is clearly a limitation of the current RIE technology and one that will disappear with the advent of a metal damascene approach.

4.2 WIRE DELAY

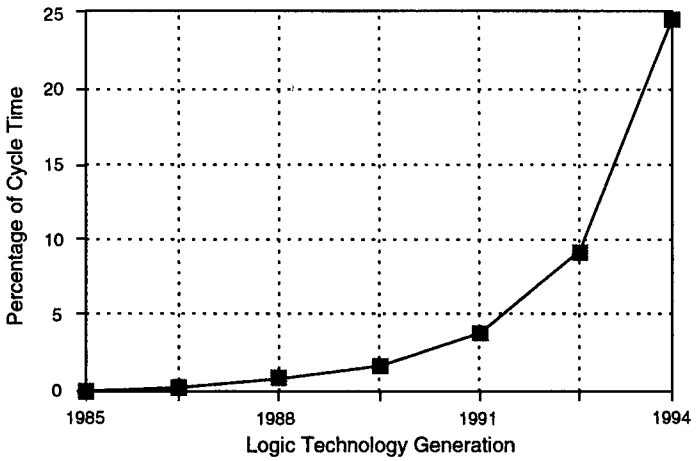


Figure 3. Global wire RC delay as a percentage of the total microprocessor cycle time for the past decade.

As Fig. 3 would suggest, the fraction of the overall microprocessor cycle time that is attributed to wire delay has been increasing at an alarming rate over the past decade. The cause for this rapid rate is obvious. As already discussed, interconnect RC delays are increasing as $S^2S_c^2$ with each successive technology. At the same time, device

performance is scaling most effectively with miniaturization. Interconnect performance, constrained by the laws of physics, just cannot keep up and becomes more and more of the overall system performance constraint. There is no end in sight to this trend without the implementation of technological band-aids as discussed in the next section.

5. Technology Solutions to Wiring Issues

While not very innovative, it is the belief of the authors that there is no white knight on the horizon to alleviate the VLSI wiring bottleneck. Several novel future technologies which provide hope for the very distant future will be mentioned in the following section. But to address the requirements of the next couple of generations of CMOS microprocessor design, we need to provide several shots of technological adrenaline, and soon!

5.1 ADVANCED TECHNOLOGY

The most straightforward of the materials improvements which can be made is to provide a copper conductor. While this may appear utterly straightforward, many hurdles must be overcome to establish copper technology in the VLSI fab, not the least of which is the over twenty years' worth of entrenched experience with aluminum. From Figure 4, it is possible to see two paths of implementation which will increase overall interconnection performance through the introduction of copper. The first path is to provide copper at the same G/R's as aluminum within an SiO_2 dielectric. The effect in RC is dramatic though the overall capacitance increases slightly in the copper case due to details of the processing sequence.

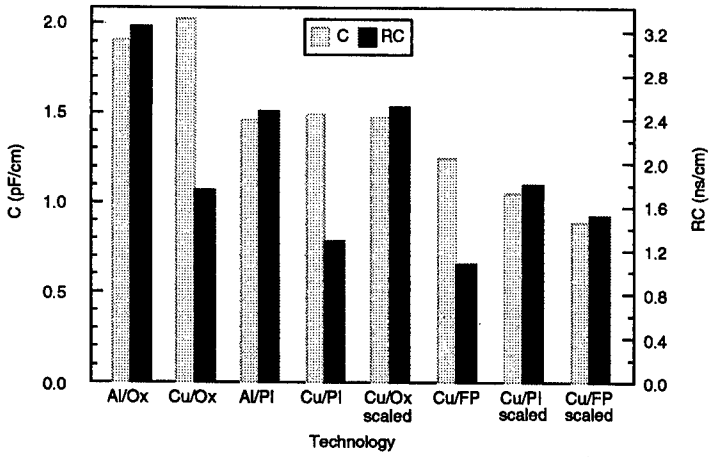


Figure 4. Three-dimensional capacitance and RC for multilevel 0.5 μm interconnects made in various technologies.

A second path of implementation would be to introduce copper at reduced G/R's such that the scaled resistivity is the same as that for a larger G/R aluminum line. The net result, as shown by the Cu/Ox (scaled) data points is a nearly 25% reduction in both capacitance and RC. Various technologies incorporating polyimide (PI) and fluorinated polymers are also shown for comparison. One must remember, however, that switching the conductor metallurgy is a one-time bonus. There are no other serious contenders for this role besides copper.

Changeover to a copper damascene metallurgy has several other potential advantages to the current aluminum RIE process. Among these is simplicity of lithographic alignment through blanket dielectric, potential for cost reduction through simultaneous line/via deposition, commonly referred to as dual damascene processing, and several orders of magnitude improvement in electromigration resistance [5].

The situation regarding future dielectric choices is, arguably, less clear. Many are possible ... few will ultimately make the grade. Most current candidates have warts of one type or another, but almost all suffer from similar problems such as poor thermal stability and hardness. The exception to this rather general statement may be several of the spin-on-glass (SOG) materials which do not appear to display the same via poisoning problems that affect most SOG and polymeric materials when used in conjunction with a hot W stud process. Among these materials is hydrogen silsesquioxide, which has an overall dielectric constant of approximately 3.0. It should be noted that transition to a copper damascene technology would also make the entry barrier lower for many of the polymeric dielectrics, since the processing conditions are not nearly so severe.

A further limitation to current interconnection technology is the general inability to provide unscaled interconnections. Wiring theory suggests that any microprocessor design will contain a small number of chip-length lines which will tend to be found in cycle limiting paths. As Fig. 3 suggests, the effect of these long lossy lines on the overall machine cycle time has been increasing. As chip sizes increase, G/R's decrease, and clock rates increase (and all are happening), the effects of these long lossy lines become more pronounced.

5.2 WIRING HIERARCHY

While materials changes can be used to provide performance improvement, they cannot provide sufficient RC delay improvement to treat the long lossy line problem. An obvious solution presents itself, that of providing additional levels of unscaled wiring to specifically treat those circuits which are potentially cycle time limiting [1]. These same "fat wire" levels can also by used for power bussing and clock distribution as is currently done with the upper levels. The distinction between "fat wire" levels and what is currently available is that "fat wires" have more cross-sectional area. In practice, this is a fairly difficult structure to fabricate. Metal RIE is particularly unforgiving when it comes to etching metal several microns thick.

Fortunately, copper damascene processing may provide an easier route to fabrication of these large conductor structures [5]. Indeed, many of the types of structures required have already been demonstrated using copper damascene technology. Further discussion of the utility of unscaled wiring levels can be found later in this volume [6].

5.3 MULTIPLE CHIP AND INTERPOSER SOLUTIONS

Another solution to the long lossy line problem, but one which is much more controversial and slightly less elegant, is to provide long cross-chip wiring in the form of interposer wiring on a single- or multi-chip module. Such solutions will be particularly attractive for multi-processor designs.

Current packaging technology is available to effect such a solution so the real decision is as much an economic one as it is a technological one. Large chips (greater than 15 mm) can benefit by as much as 20% in RC delay by going to thin film wiring as opposed to using double-wide 0.35 G/R on-chip wiring. If the die size plateau is not reached, as we are projecting, the use of alternative low resistance, cross-chip wiring, such as is available in the form of thin-film wiring on SCM's or MCM's, will be a viable alternative to additional BEOL wiring levels.

6. Future Technology

Very few viable technologies which directly impact the VLSI wiring bottleneck loom shimmering on the horizon. There are glimmers of hope, such as room temperature superconductors, which might make things interesting, but if we ground ourselves in reality and make critical evaluations of what is available, the undeniable conclusion is that materials technology is the only relatively near-term remedy for the interconnection bottleneck, and that remedy only treats the symptoms. Some of the more exotic technologies which offer some hope are the following.

6.1 MULTI-LEVEL LOGIC

Multi-level or multi-state logic is a concept which is currently available in many respects. The most important of these is that CMOS devices built using currently available technology can be used for multi-state operation. This situation is beneficial since it removes the need to fight the large capital infrastructure in place in the semiconductor industry. Unfortunately, system design using base R, where R is the number of distinct logic states available, is not currently practiced and there is very little momentum available to establish a new standard. Wiring reductions by a factor of $\log_2(R)$ may be available [7], but this may not be enough to push the industry to consider the considerable investment needed to move to a base R computing standard.

Another issue which is frequently addressed with regard to multi-state logic is that of signal tolerance. A binary system is clearly very robust from this standpoint, but as lower power technologies become available, it is clear that we are able to engineer lower noise margins. At 2.5 V we are already equivalent to a three state technology at 5 V. It is not clear that signal tolerance is an insurmountable hurdle with multi-state logic and, of the alternative technologies conceived, this approach may be the most realistic. Overcoming the entrenched binary standard will, however, be extremely difficult.

6.2 ALTERNATIVE INTERCONNECTIONS

Several materials alternatives for VLSI wiring processes have already been discussed. The purpose of this section is to address the feasibility of the more exotic forms of physical interconnection available. Throughout the discussion we continue to consider

the current Si-based VLSI technology as the base with which any conceived alternative must compete and ultimately displace.

6.2.1 Optical.
At this time there is no implementable scheme for optical interconnects in Si-based VLSI technology. The more unfortunate point is that there is currently no obvious path to implementation or that if a path was available, whether it would be practical [8]. One might state that on-chip (Si) interconnects are not in the realm of reality for optical interconnects and be pretty safe. It is apparent that the more easily obtainable and, perhaps, more useful role for optical interconnects is as intra-MCM and, in the future, as inter-MCM connections.

6.2.2 Superconducting.
For many years after the discovery of the high temperature superconducting cuprates, the goal of producing usable on-chip (Si) interconnects was viewed as highly desirable. Such an implementation still has many hurdles to cross [9], since many of the high temperature superconductors are metastable phases in their own right, and more seriously, since they react quite quickly with Si, SiO_2, polymers, etc. at the temperatures required to deposit them in thin-film form. Another serious limitation is their inability to carry the high current-densities required of on-chip interconnections. Unless new discoveries occur, the usage of high T_c superconductors for on-chip wiring appears completely out of the question.

6.3 CELLULAR AUTOMATA

One of the most exciting prospects for alternative computing which has major impact on the wiring aspects of system design is the concept of quantum cellular automata [10]. This concept is clearly many, many years away from usable products, but it does contain some clearly valuable attributes.

Of these attributes, the most astonishing is that Coulombically coupled quantum cellular automata do not require wires at all! Wires are constructed of the devices themselves. In the most concrete suggestions for implementation [11], devices exist as islands of nano-phase material patterned using an STM tip. While this fabrication concept and the actual usage of such a cellular automaton scheme is clearly immature [12], the ability to conceive of an implementation scheme is undoubtedly a step in the right direction.

7. Summary

Modern VLSI technology is faced with a challenge in terms of its ability to reasonably wire highly complex microprocessors and, at the same time, provide sufficient performance in those interconnections to keep pace with steadily improving device characteristics and system cycle times.

The current generation of VLSI wiring technology is superb. It has been reengineered through many generations and is still capable of delivering on both its reliability and density requirements. Wiring delay as a percentage of system cycle time, however, has been increasing. Today's highest performing microprocessors are clearly at

the mercy of wire delay constraints. Increasing chip sizes and decreasing G/R's are tightening these constraints.

Unfortunately, the avenues available to alleviate the wiring problem are not numerous. Materials substitution in the form of copper metallurgy and low epsilon dielectrics can provide some relief; there is increased activity and emphasis in the industry to adopt these. In addition, it is expected that chip sizes and overall machine complexity (circuit count) will begin to plateau. Design methodology will begin to concentrate on cycle time reduction which will drive chip size and interconnect length down, but will also demand extremely low RC wiring paths (fat wires) in order to reduce time-of-flight delays.

While several novel technological concepts exist to address the wiring issue, none of these seem sufficient to impact the industry in the near- to medium-term. This being true, the semiconductor industry is left with the options of introducing new materials technologies which give one-time boosts to wiring performance, introducing unscaled wiring to treat cycle limiting paths, and awaiting the arrival of less complex and smaller chip-size microprocessor designs.

8. Acknowledgments

The authors would like to thank the following people for both technical assistance and critical comments: Rolf Landauer, John Heidenreich, J. Frank White, John Hummel, Steve Greco, C.-K. Hu, George Sai-Halasz, and Kerry Bernstein.

9. References

1. Sai-Halasz, G.A. (1995) Performance trends in high-end processors, *Proc. IEEE* **83**, 20-36.

2. Bakoglu, B. (1990) *Circuits, Interconnections, and Packaging for VLSI*, Addison-Wesley Publishing Company, New York.

3. Keyes, R.W. (1982) The wire-limited logic chip, *IEEE J. Solid-State Circuits* **17**, 1232-1233.

4. Taur, Y. *et al.* (1995) CMOS scaling into the 21st century: 0.1 μm and beyond, *IBM J. Res. Develop.* **39**, 245-260.

5. Edelstein, D.C. (1995) Advantages of copper interconnects, *IBM Corp. Internal Research Report No. RC 20097.*

6. Sai-Halasz, G.A. (1996) Processor performance scaling, in this volume.

7. Etiemble, D. (1992) On the performance of multivalued integrated circuits: past, present and future, *IEICE Trans. Electron.* **E76-C**, 364-371.

8. Chen, R.T. (1994) Guided-wave optoelectronic interconnects: their potential and future trends, *SPIE Optoelectronic Interconnects II* **2153**, 196-199.

9. Solomon, P.M. (1988) The need for low resistance interconnects in future high-speed systems, *Proc. SPIE* **947**, 104- 116.

10. Lent, C.S., Tougaw, P.D., and Porod, W. (1994) Quantum cellular automata: the physics of computing with arrays of quantum dot molecules, *Proceedings of Workshop on Physics and Computing,* IEEE Computer Society Press, Dallas.

11. Bandyopadhyay, S., Das, B., and Miller A.E. (1994) Supercomputing with spin-polarized single electrons in a quantum coupled architecture, *Nanotechnology* **5**, 113-133.

12. Landauer, R., (1990) Advanced technology and truth in advertising, *Physica A* **168**, 75-87.

PHYSICS, MATERIALS SCIENCE, AND TRENDS IN MICROELECTRONICS

H. VAN HOUTEN
Philips Research Laboratories
Professor Holstlaan 4
5656 AA Eindhoven, The Netherlands

1. INTRODUCTION

From the end of world war II to the demise of the cold war, the expenditure on basic research in solid state physics and materials science has often been justified by reference to the invention of the transistor, and the resulting revolution in electronics [1]. It is the stated aim of this conference to study the future of microelectronics, beyond the time when the shrinkage of CMOS feature sizes will have come to a stop. While major western industries are losing confidence in the business generating potential of basic research in physics and materials science, government funding agencies are sponsoring research programmes on esoteric subjects such as quantum devices, single electron tunneling, molecular electronics, atom manipulation, or self-assembly [2]. In doing this, they are guided by the idea that breakthroughs in these fields will shift the physical limits of the miniaturization trend to the nanometer scale, and thereby ensure decades of continued growth for the electronics industry. The fact that this would require "Revolutionary chip architectures which remove the current interconnection limitation to functional density", as well as "Revolutionary devices which make use of physical phenomena on a much smaller scale than transistors" does not seem to deter the scientific community [3]. Indeed, similar revolutionary ambitions abound in the fields of computing and data storage (the all-optical computer, the biochip, scanned probe data storage,...).

This paper presents the view point that those of us trying to defend the strategic role of research in physics and materials science should not advocate such trendy goals, which remind us of failures from the past, for example the Josephson computer [4]. This type of desired breakthrough is

S. Luryi et al. (eds.), Future Trends in Microelectronics, 57–70.

indicated in Fig. 1 by a small cloud, situated at the intersection of a main technology trend and its (perceived) fundamental performance limit. While taking an ambitious - but naive and rather unimaginative - aim far beyond such limits, these endeavours often have intrinsic flaws, do not fulfill all the requirements, or do not address the real limitations or bottlenecks of the existing solutions [5].

Instead, we ought to identify more realistic goals for our strategic physics and materials research. Goals that make sense from a technical and business point of view can be found, even in a world dominated by dreams of multimedia systems, software, and services. To help assess such goals, it is useful to distinguish between "trend feeding" and "trend setting" innovation. [1] The following definitions will be used in this paper.

- **Trend feeding innovation** is characteristic of an established product/market combination. Support is given to an existing trend of improvement versus time of key performance parameters (data rate, data density, instructions per second, packing density, cost per bit, etc). It can be based on evolutionary or more radical technology changes (see arrows in Fig. 1), but *all* technical and economical compatibility requirements characteristic of the main trend must be met simultaneously.
- **Trend setting innovation** addresses a new product/market combination. The performance level of existing trends is often not met, at least initially. One or two outstanding features of a new technology are exploited, to create the initial market, and to set the new trend.

The nature of these two types of innovation will be described in more detail in the next two sections. We also present some examples, which clarify the difference, and which may point the way to some more general themes for innovation based on physics and materials science.

2. TREND FEEDING

2.1. SYSTEM TRENDS

System engineers tend to think in terms of hardware options available today. Yet, system trends can provide directions for hardware research. The Japanese company NEC has been guided over decades by their vision of the merging of computers and communication (C&C)[7]. Today, several companies are convinced that a further merging of these two product areas with consumer electronics (CC&C) is imminent. Technical driving forces behind this unification are digitalization, and the need for interconnect-

[1] In an interesting paper, Bower and Christensen recently made a somewhat similar distinction between disruptive and non-disruptive technologies[6]

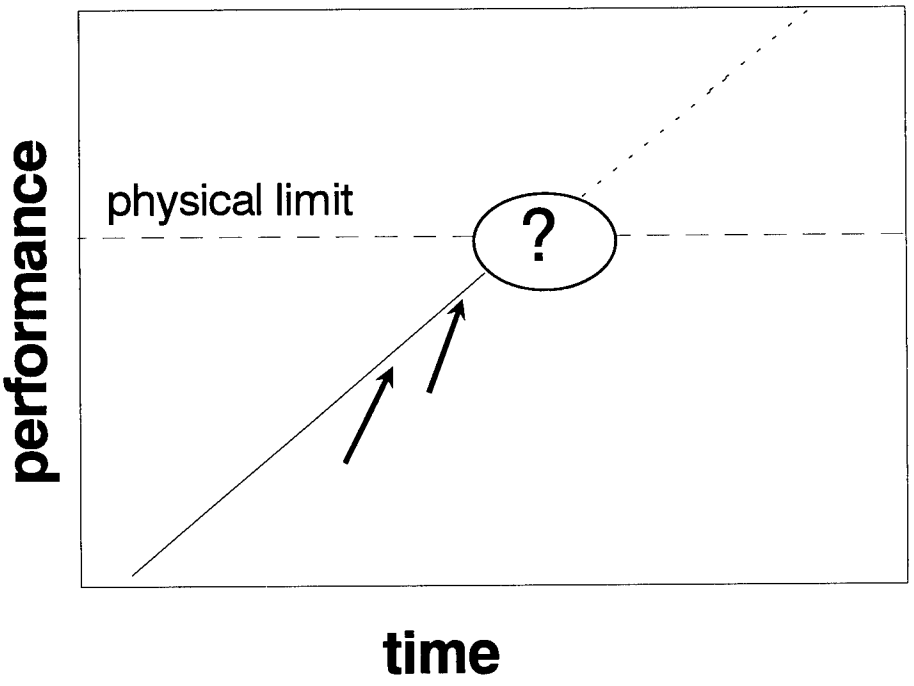


Figure 1. Typical exponential trend in performance versus time, fed by evolutionary or more radical technology changes (as indicated by arrows). According to conventional wisdom, physics and materials science should create breakthroughs (indicated by the small cloud) that would dramatically shift the perceived physical limits of the existing technology (dashed line), thereby ensuring a continuation of the same trend for further decades.

ing the traditional stand alone electronic "boxes" (such as a TV, a PC, a telephone) using networks. Quite important is also the recognition that a common layered system architecture may underpin quite different functions. A service (a movie, or a telephone conversation) is distributed over a general purpose utility (e.g. a cable network). The system is ruled by an operating system, governed by (perhaps downloadable) software. The hardware is typically build in a modular fashion, with a microprocessor, a hierarchy of storage modules, a display function, a user interface, etc. A fashionable concept on this emerging multimedia battle ground is that of the value chain. In the trajectory from hardware component to service rendered, each of the parties involved will strive for the most profitable slice of the pie. While some see no compelling reason to remain active in hardware, wishing instead to concentrate on the (more profitable) software and services, others see interesting opportunities for "key components" or "key

modules", which constitute the core of the hardware (and which may have added value through embedded software). The drive for increased functionality of key components is reflected in the slogans "a system on a chip", and "a system on a display". The power balance between manufacturers of components (e.g. a microprocessor) and set makers (e.g. a PC) is uncertain. The vagueness of the demarcation line between components, modules, and systems has a significant impact on packaging issues.

2.2. TECHNOLOGY TRENDS

Supporting the system trends discussed above are some dominant technology trends. The relentless progress made in the fields of data storage, processing, and communication is one of the most remarkable phenomena of the twentieth century. For example, the areal density of information, data rate, and cost per bit of magnetic hard discs, and of silicon based dynamic random access memories have improved as an exponential function of time over several decades. This development has not yet come to a stop.

This thrust is based in part on heavy investments in evolutionary technology development (such as successive generations of silicon wafer diameters, clean room equipment, and CMOS processes). The significance of this development is self-evident. Yet, it would be a mistake to think that fairly radical technology changes are not required as well to feed the ongoing trends (perhaps they *seem* less radical from the basic physicist's point of view). The *common aspects of both evolutionary and more radical trend feeding innovations are that they must meet all of the demanding technical requirements characteristic of the main trend* (in terms of performance, manufacturability, cost, and compatibility with system requirements). Typically, such innovations lead to a next generation product on existing or similar markets. The nature of the technology step (radical or evolutionary) is often not noticeable to the end user. The decision on the introduction of the new technology will therefore be taken based on an evaluation of critera such as technical figures of merit, manufacturability, investment needed, and expected benefits.

2.3. EXAMPLES

Radical technology changes, intended to feed main trends, have to meet the *same* severe compatibility requirements (in terms of manufacturing, standards, etc.) as more evolutionary ones – on top of a demonstrated compelling advantage in one or several aspects. A boundary condition usually is that the increase in product cost should be negligible. Proving that all of these demands can be met simultaneously is a major challenge for industrial research, requiring a sustained commitment over an extended

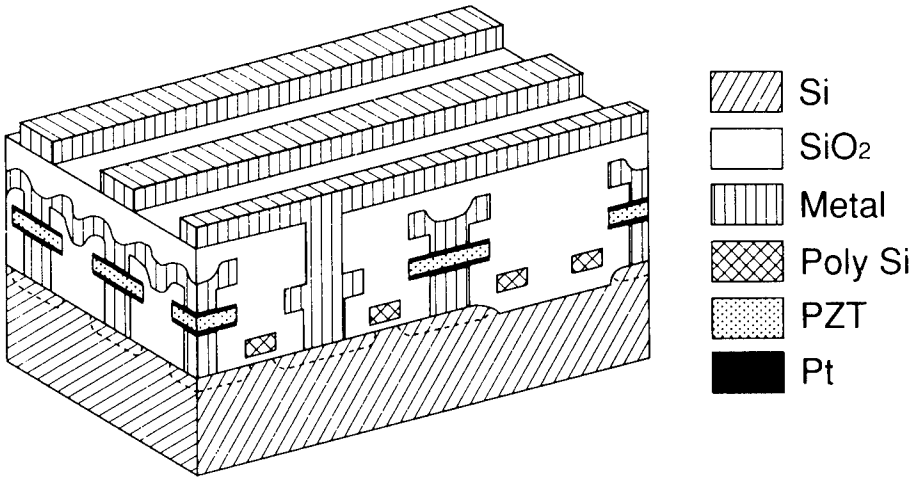


Figure 2. Non-volatile ferroelectric memory cell consisting of a thin film ferroelectric capacitor integrated on top of an MOS transistor.

period of time. Below, we will give some examples from the fields of embedded memories, packaging, and magnetic data storage. These examples were selected to illustrate four rather general themes for "trend feeding" innovation based on physics and materials science:

- The enhancement of the performance or functionality of silicon based microelectronics through new materials.
- Miniaturization enabled by sophistication in packaging.
- Opto-electronic technologies providing solutions for systems demanding high performance in terms of processing (using electronics) as well as communication (using photonics).
- The migration of LSI technology to application areas which are served today by non-planar (i.e. bulk like) technologies.

2.3.1. *Ferroelectric Memories*

Solid state data storage addresses the high-price, high-speed, low-capacity segment of the memory hierarchy, with magnetic or optical recording being complementary in terms of these performance indicators. Solid state memories are perhaps most familiar as stand-alone products. The dominant innovation trend is of course the ever larger memory size per chip, at a cost per Mbyte that decreases as an exponential function of time. In addition to this dominant trend, another very interesting market exists for

non-volatile memory, stand-alone or embedded in microprocessors. Today, this memory is typically based on the electrically programmable read-only memory (EPROM). An EPROM stores information without consuming any electrical power – albeit at the expense of a limited number of read-write cycles, and a low write speed. The reason is that writing requires the transfer of (hot) electrons across an insulating layer to a gate electrode buried in silicon dioxide. An interesting alternative option is depicted in Fig. 2, which shows a ferroelectric memory cell, consisting of a ferroelectric capacitor integrated on top of an MOS transistor. Bits are stored at zero voltage as one of the two possible states of remanent polarization of a ferroelectric thin film (typically lead-zirconium-titanate). A read-out can be effected by applying a poling pulse, and monitoring the charge transfer (restoring the information subsequently by means of a write pulse).

The idea to use ferroelectrics for a non-volatile memory dates back to the 1950s, but it has taken a major research effort in oxidic thin film deposition, patterning, and interface engineering to come to a viable technology that is compatible with CMOS processes. Philips Research has played an active role[8]. The ferroelectric memory technology has now proven to possess a number of very attractive features: fast switching, long retention, high endurance, low switching voltage, and scalability to small cell sizes. It is this combination of properties that makes ferroelectric random access memories a likely candidate to replace embedded memories in CMOS-based circuits (such as the EPROM)[8]. Once the hurdle of introducing this technology into CMOS factories will have been taken, a range of additional possibilities emerge, such as integrated pyroelectric sensors, or smart piezoelectric actuators and sensors [9]. In addition, one can think of high dielectric constant materials for capacitors, or for gate dielectrics.

2.3.2. *Integrated Components Module*

The continued miniaturization and enhanced functionality of (portable) electronic equipment is a major thrust for innovation, but smaller feature sizes and larger VLSI chip sizes are only one of the enabling technology trends. What is needed as well is an increased sophistication in packaging. To realize the significance, consider the example of an 8 mm video cassette recorder, which today (1995) contains as many as 70 integrated circuits, 1400 discrete resistors and capacitors, and 200 discrete transistors. The miniaturization of passive components helps – but their manufacture, handling, placing, and soldering becomes increasingly untractable. A possible solution investigated in Philips Research (see Fig. 3) and also at AT&T Bell Laboratories is the integration of a large number of passive components on a single substrate (e.g. silicon), using thin film technology. Provided a dedicated production line is used, the properties of advanced materials can be

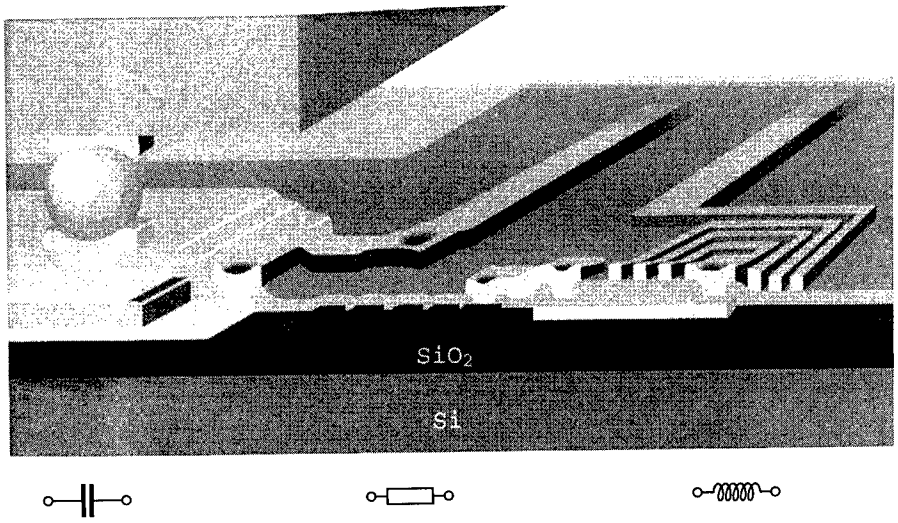


Figure 3. Technology for an Integrated Components Module. From left to right one sees an IC mounted upside down onto a Si substrate, containing a resistor, a capacitor, and an inductor.

exploited without worrying about cross contamination of CMOS processes. Thus, alloys (e.g. NiCrAl) can be used to tailor resistors with a vanishing temperature coefficient, and oxides with a high dielectric constant can be used to make small capacitors. CMOS integrated circuits can be mounted on the same substrate with minimum interconnect lengths using a flip-chip bonding technique, yielding improved high frequency perofmance and electromagnetic compatibility. The resulting hybrid is referred to as an Integrated Components Module.

2.3.3. *Board Level Optical Interconnects*

The discovery of efficient photo-luminescence from porous silicon spurned a frenzy of scientific activity, tempered only moderately by the fact that efficient electroluminescence proved elusive [10]. This activity was motivated by the idea that a Si based light emitting diode would enable *intra*-chip optical interconnects between logic gates, and thereby alleviate the so-called "interconnect bottleneck". This bottleneck arises because down-scaling of CMOS feature sizes leads to faster gates, while the RC time associated with the charging of the interconnect line is constant. In contrast, optical interconnects have a delay time determined by the response times of source and detector, and the propagation time of the light. However, a straight-

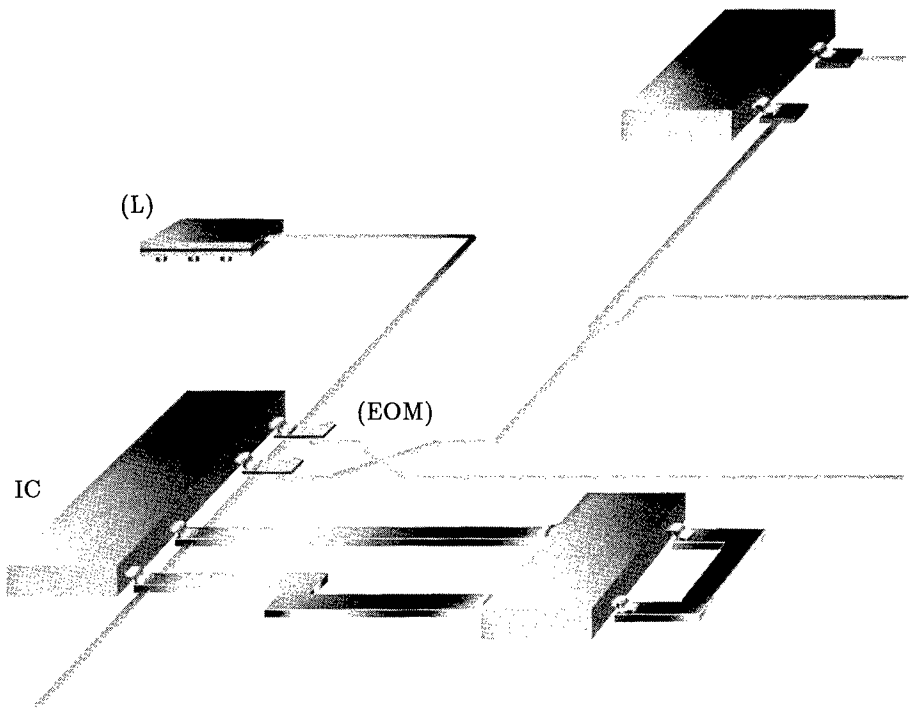


Figure 4. Concept of a hybrid technology for optical interconnects on a printed circuit board (PCB). A single laser (L) feeds a polymer waveguide acting as a bus (thin grey line). Electro-optical modulators (EOM), addressed by the electical output pins of a mounted IC, couple modulated optical signals into signal waveguides. These waveguides may intersect each other without interference (thin grey lines). At the receiving end a detector is used to convert the optical signal back into an electrical one. In addition to the optical links, also conventional copper lines are used on the PCB.

forward analysis of the signal to noise requirements for a point-to-point optical interconnect at the chip level shows that a Si-compatible microlaser would be required, capable of modulation at GHz frequencies, and with an output power of at least 10 micro Watt[11]. Demanding requirements have to be met by the detector as well. These practical difficulties seem unsurmountable at present.

Better opportunities for optical interconnect may arise at the next level in the interconnect hierarchy, which is the board or module level. In this case, the bottleneck is related to high digital transfer rates between IC's, at distances on the order of 1–10 cm. Digital signal processing of video signals

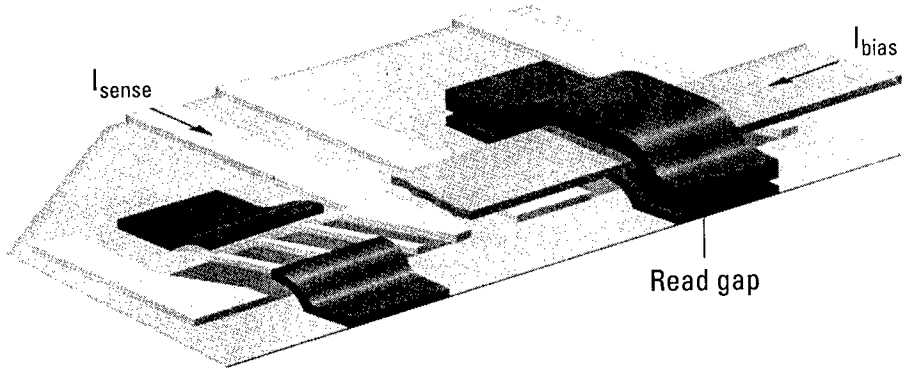


Figure 5. Schematic view of the play back part of two of the digital channels of a DCC head. For clarity the left channel shows only the magnetoresistive element (with the barber pole) and the lower half of the magnetic yoke. For the right channel the complete read yoke is shown. In practice, a write yoke is processed on top of this.

will occur at even higher clock frequencies (exceeding a GHz in the near future). This gives rise to significant electromagnetic compatibility (EMC) problems, which may conflict with government directives. An interesting approach is to integrate the optical waveguides, a laser source, modulators, and detectors on the substrate of a single "multi-chip module", using flip-chip bonding to connect the IC's [12]. Such a hybrid technology has the advantage that standard IC's can be used. An alternative low-cost approach under investigation in Philips Research is to add a few optical links to a printed circuit board (PCB). In addition to the benefits mentioned before, optical interconnects on a PCB could eliminate the need for multilayer electrical interconnects. The PCB environment also provides enough space, and conventional electro-optical technology (e.g. polymer waveguides) can be used. This approach, which focuses on studying a systems concept rather than a single device, is illustrated in Fig. 4.

2.3.4. *Thin film recording heads*

In magnetic data storage, the longitudinal recording density is increasing by a factor of 10 per decade [13]. The required narrow- and multi-track recording heads can no longer be manufactured using a discrete inductive

ferrite head technology. Instead, thin film heads using magnetoresistive elements for the read function are being introduced[14]. Fig. 3 shows part of the play-back section of a multi-track thin film head. Such a head, used in the Philips Digital Compact Cassette (DCC) player, contains nine parallel digital channels as well as two analog channels that enable also the playback of standard analog compact cassettes[15]. The magnetoresistive element is part of a soft-magnetic yoke that guides flux from the magnetic medium. A linear response is ensured in this case by forcing the read current under a 45 degree angle to the magnetisation direction, using the "barberpole" shaped conductive strips visible in the left part of Fig. 3, in combination with a biasing magnetic field established by the current indicated in the right part of Fig. 3. A second function of the barberpole is that the skew sense current induces a small stabilization field along the element that suppresses Barkhausen noise. Whereas this head currently exploits the anisotropic magnetoresistance effect in NiFe, it is expected that in the future the giant magnetoresistance effect in metallic multilayers will be introduced. The linearity of this effect will eliminate the need for complicated biasing schemes, and the sensitivity will enable even higher bit densities[15]. Other applications of integrated magnetic devices should be expected to follow. One imagines, for example, miniaturized high frequency magnetic circuitry in the front end of mobile phone systems, or automobile sensors for control of car functions or navigation.

3. TREND SETTING

3.1. NEW PRODUCTS, NEW MARKETS

New combinations of existing technologies, and occasionally radically new technologies, may also give rise to entirely new trends, addressing new product or new market opportunities. At least initially, there is often no possibility of matching the performance or requirements of existing product trends. Instead, a new product or product family arises which builds on one or two strengths which the established technology did not possess. Thus, the personal computer could not be compared to a mainframe in terms of performance, but became a success because of the appeal of individual ownership, and convenient size so that it could be put on a desk in the office or at home. The development of applications such as text processing and computer games were an important factor as well. The transistor was a success because it enabled new battery driven miniaturized products operating at low voltage and with low power consumption, while its performance could not compete with that of the vacuum tube. Flat panel liquid crystal displays cannot (as yet) compete with cathode ray tubes in terms of brightness, or viewing angle. Instead, they entered the datagraphic display

market because they enabled the development of the lap-top computer.

It does not make sense to map trend setting innovations on existing charts of performance versus time. Traditional market research (asking your customers) is also of little use in trying to establish their potential in an early stage. Similarly, cost prize analysis is a poor guide, as the cost of manufacturing for the consumer market can often be made amazingly low, as long as large volumes are made and sold. The successes of the video cassette player and the compact disc player prove these points[16].

Researchers may have to consider quite different sources of information to identify opportunities. Thus, one may think about the (possibly latent or emerging) needs of the *end user*, the *product designer*, the *manufacturer*, or the *distributor*. It is of little use to try to formulate general statements about such needs. Instead, we will illustrate this point of view by an example.

3.2. EXAMPLE: PLASTIC ELECTRONICS

The success of molecular biologists and organic chemists in synthesizing highly complicated organic molecules has opened up the possibility to engineer materials with interesting properties. The ambitions in this field vary from the mundane (engineering a new type of plastic) to the exotic (inventing a transistor based on a single molecule, "molecular electronics"). Where - between these two extremes - should one position the idea that thin films of conjugated polymers can serve as the active layer in electronic devices? Functional accumulation-type thin film transistors and Schottky diodes (making use of heavily doped polymers) have been demonstrated. Philips research laboratories is one of the active players[17]. In addition, efficient light emitting diodes based on a sandwich of an intrinsic polymer layer between electron- and hole-injecting electrodes have been manufactured succesfully by various groups[18]. The technical problems faced by these devices are considerable. Thus, issues such as reproducibility, lifetime, and degradation are only beginning to be addressed. The understanding of the underlying device physics is also rather rudimentary at present. It is reasonable to expect significant improvements in these areas in the coming years. Yet, one should be prepared to answer the sceptic's "so-what?", because it is quite unlikely that conjugated polymer devices will be able to compete in terms of performance, or reliability, with existing products based on anorganic semiconductors. The answer could be that this is a trend-setting technology, which may address new product/market combinations.

To substantiate this statement, one should look at the potential intrinsic "selling points" of organic technology. Two aspects come to mind. Firstly, polymer layers can be deposited in patterns using printing technology (off-

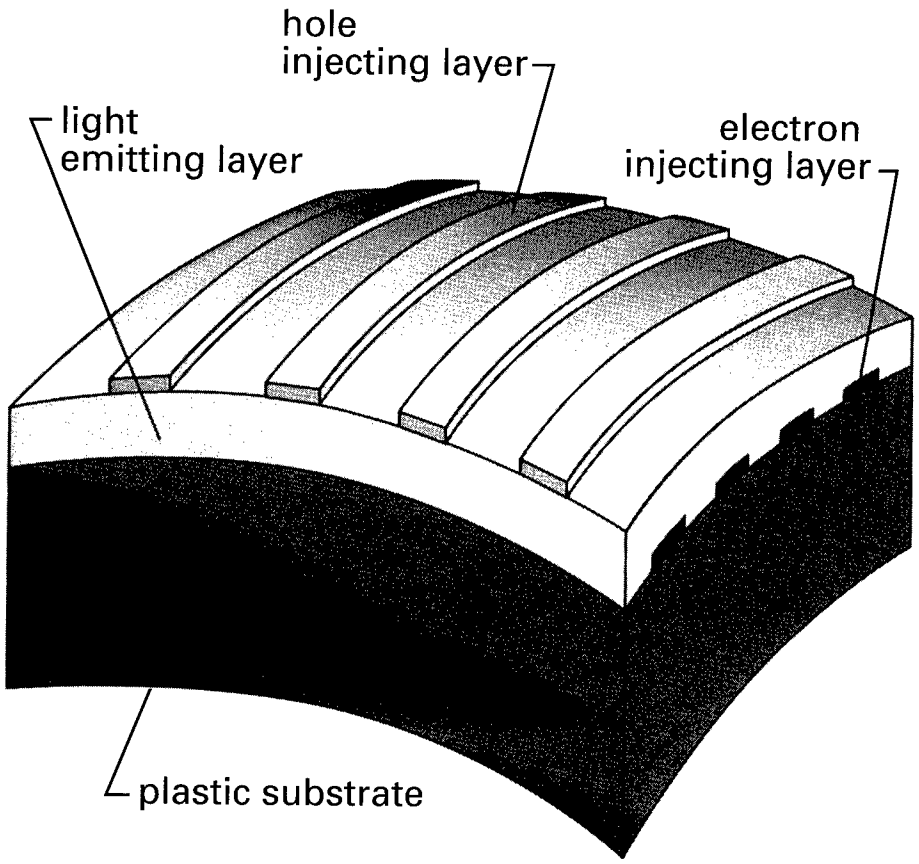


Figure 6. Concept of an electroluminescent display based on conjugated polymer films serving as electron and hole injecting layers, and as light emitting layer. Cheap printing technologies may be used to manufacture this device. Flexibility may be achieved by using a plastic foil as a substrate.

set or silk-screen), which could yield a dramatic improvement in terms of cost compared to standard photo-lithography[19]. Secondly, polymer layers can be deposited on flexible plastic substrates (foils). This would pave the way to replace batch-type manufacturing with a continuous flow-type process (leading to a low-cost large area electronics). In addition, it would give product designers enhanced freedom of design. The non-planar display sketched in Fig. 6 is an example. On the longer term, one might even conceive of fully flexible electronic portable equipment. It is important to

realize that a dramatic cost reduction might also create new markets ("low-end electronics"). One may think of printed plastic electronics replacing bar codes for identification purposes, to track the flow of goods, and to enable flexible pricing (depending on the time of day). Similarly, anti-shop lifting buttons might in the future be replaced by very cheap electronic solutions. Further interesting ideas have been described in the literature[20].

4. CONCLUSIONS

We have argued in this paper that strategic roles can be found for physics and materials science, in support of continued growth of the microelectronics industry. Instead of defending the mainstream vision that physics and materials science should attack perceived fundamental limits of downscaling by adventurously trying to invent yet another successor of the transistor (which could be labelled as "trend fighting"), we have sketched two alternative routes.

Firstly, there is room for "trend feeding" innovation. This may be based on fairly radical technology, but it must be compatible with existing technologies, and it must meet all requirements characteristic of the main trend.

Secondly, there is room for "trend setting" innovation. Here one should not try to compete with the dominant trend. Instead, the focus should be on a new product/market combination, based on one or two outstanding properties (with a reasonable performance in other aspects).

We stress that these two types of innovation do not rely on technology only: they also offer challenging subjects for physics and materials research. Examples are the phenomenon of giant magnetoresistance in metallic and oxidic multilayers, and the interplay of transport and luminescence in disordered polymer films. Last but not least, an entrepreneurial attitude is a prerequisite for success in both cases.

5. Acknowledgement

I thank my colleagues at Philips Research Laboratories for useful discussions. I am grateful in particular to Rob Fastenau, Wiep Folkerts, Poul Larsen, Dago de Leeuw, Nick Pulsford, Coen Liedenbaum, Emile Staring, Robert-Jan Visser, and Gerjan van de Walle for providing me with information and with the figures used in this paper. I have also benefitted from the views of Gerard Vos from Philips Key Modules.

References

1. *Physical Review: Centenary-from basic research to high technology,* Physics Today, October 1993.

2. *Engineering a Small World, From Atomic Manipulation to Microfabrication*, Science **254** (1991) 1300-1342.
3. R.T. Bate, G.A. Frazier, W.R. Frensley, J.W. Lee, and M.A. Reed, *Prospects for quantum integrated circuits*, SPIE Vol. 792 (1987) 26-35.
4. W. Anacker, *Computing at 4 degrees Kelvin*, IEEE Spectrum May 1979, 26-37.
5. R. Landauer, *Advanced Technology and Truth in Advertising*, Physica A **168** (1990) 75-87.
6. Joseph L. Bower and Clayton M. Christensen, *Disruptive Technologies: Catching the Wave*, Harvard Business Review, January-February 1995, p. 43-53.
7. W. Aspray, Interview with Koji Kobayashi, in *Engineers as Executives* (IEEE Press, New York, 1995).
8. P.K. Larsen, G.A.C.M. Spierings, R. Cuppens, and G.J.M. Dormans, *Ferroelectrics and high permittivity dielectrics for memory applications*, Microelectronic Engineering **22**, (1993) 53-60.
9. O. Auciello and R. Waser, *Science and Technology of Electroceramic Thin Films*, NATO ASI Series E, Vol. 284 (Kluwer, Dordrecht, 1995).
10. L.T. Canham, Appl. Phys. Lett. **57**, 1046 (1990).
11. N.J. Pulsford, unpublished results.
12. T.E. van Eck, G.F. Lipscomb, A.J. Ticknor, J.F. Valley, and R. Lytel, Appl.Optics **31**, 6823 (1992).
13. Y. Miura, *Advances in magnetic disc storage technology*, J. Magnetism and Magnetic Materials **134** (1994) 209-216.
14. J.L. Simonds, *Magnetoelectronics today and tomorrow*, Physics Today April 1995, 26-32.
15. W. Folkerts, *Magneto-resistive thin film heads for tape recording: past, present, and future*, Read/Write **18** (1994) 8-11. W. Folkerts, J.C.S. Kools, M.C. de Nooijer, J.J.M. Ruigrok, L. Postma, K.-M.H. Lenssen, G.H.J. Somers and R. Coehoorn, *Performance of Yoke type GMR heads*, Proc. Intermag. '95, San Antonio.
16. P. Ranganath Nayak and J.M. Ketteringham, *Breakthroughs*, Arthur D. Little Inc.
17. See, e.g., A.R. Brown, A. Pomp, D.M. de Leeuw, D.B.M. Klaassen, and E.E. Havinga, *Precursor route pentacene metal-insulator-semiconductor field effect transistors*, Appl. Phys. Lett. to be published.
18. D. Braun, A.J. Heeger, *Visible light emission from semiconducting polymer diodes*, Appl. Phys. Lett. **58**, 1982 (1991); J.H. Burroughes, D.D.C. Bradley, A.R. Brown, R.N. Marks, R.H. Friend, P.L.Burn, and A.B. Holmes, *Ligh emitting diodes based on conjugated polymers*, Nature **347**, 539 (1990).
19. F. Garnier, R. Hajlaoui, A. Yassra, Pratima Srivastava, *All-polymer field effect transistor realized by printing techniques*, Science **265**, 1684 (1994).
20. R. Friend, D. Bradley, and A. Holmes, *Polymer LEDs*, Physics World November 1992, p. 42; P. May, *Polymer Electronics - fact or fantasy?*, Physics World, March 1995, p. 52.

GROWING UP IN THE SHADOW OF A SILICON 'OLDER BROTHER': TALES OF AN ABUSIVE CHILDHOOD FROM GaAs AND OTHER NEW TECHNOLOGY SIBLINGS!

Lessons from GaAs technology development.

PAUL R. JAY
Microwave Modules Group
Northern Telecom Ltd.
P. O. Box 3511, Station 'C'
Ottawa, Ontario, Canada K1Y 4H7

1. Introduction

Not surprisingly, the gallium arsenide community has been, for some years, the object of focused aggression and resentment from their colleagues in the silicon industry. Family psychologists will tell us that such types of "sibling rivalry" are not uncommon, and that ultimately the older offspring will come to accept and even value the younger upstart that has created such a distraction within the semiconductor family!

Perhaps this change of attitude is beginning to happen as GaAs technology is establishing itself as an authority in certain applications areas, and the silicon protagonists realize that the new technology challenges only certain niche areas, as opposed to undermining the whole basis of the silicon economy. Indeed, it is becoming apparent that certain aspects of technology developed to cope with the particular issues facing GaAs (such as rapid thermal annealing) can be usefully applied to Si processing.

Having established this platform for a provocative discussion in a workshop atmosphere, this paper aims to summarize some of the recent applications achievements of GaAs technology, and to examine the history of the emergence of GaAs to determine what lessons can be learned from the pitfalls encountered *en route*. This allows us to look for some sort of "selection rules" that might be useful to test a proposed new technology and estimate its right to survival. The analysis also begs the question presented by continued progress down the paths towards further miniaturization of existing technologies for higher performance and greater speed: "Is a fundamentally new approach to device technology required?"

2. Background of GaAs Applications

In order to illustrate the fact that GaAs technology is now able to make useful contributions to the electronics marketplace, this section will describe a few examples of applications at different volume levels, many drawn from a workshop presented a few years ago at the IEEE GaAs IC Symposium [1]. The range of successful applications now runs from relatively low volumes of complex, high-performance IC's to substantial

S. Luryi et al. (eds.), Future Trends in Microelectronics, 71–86.

volumes (in excess of millions of IC's per year) usually involving IC's optimized for minimum size and cost. It is interesting to observe that whereas some of the earlier applications fulfilled by GaAs involved very high speed digital functionality, the major markets currently creating a renewed interest in GaAs are principally driven by the needs of microwave functions in telecommunications and wireless transmission.

2.1 HEWLETT-PACKARD TEST INSTRUMENTATION

For many years now, Hewlett-Packard have been employing linear and microwave GaAs IC's in their high speed test and measurement instruments [2]. The introduction of GaAs chips was usually done on a case-by-case basis where functional benefit was evident from the prototype demonstrations in terms of a simplified overall architecture and an associated performance enhancement. The types of series volumes involved are not necessarily large, but the value of developing an application-specific IC (ASIC) for one product line can be justified even for product runs of 5,000-10,000 units, where an NRE cost of $100k would involve only $10-20 additional cost per unit. In order to evaluate and introduce these codes, HP sustains an in-house production line to supply their commercial needs. Typical functions involved are microwave amplifiers, switches, and high frequency dividers, and a number of the applications have enjoyed additional advantages as simplifying retrofits in field service situations.

2.2 DIRECT-BROADCAST SATELLITE RECEIVERS

In 1989 a UK company (Continental Microwave Technology Ltd.) designed a small GaAs microwave chip to down-convert 11-12 GHz signals received from a 30 cm square satellite dish that is lightweight enough to be mounted directly on the outside wall of a consumer's premises. The resulting 0.9-1.75 GHz IF signal is then fed directly to the set-top unit of the TV receiver. The chip concerned integrates onto a single die all the high-frequency RF-IF operations of the receiver front-end, without the need for costly assembly and tuning of multiple components.

This DBS (direct broadcast satellite) chip is still manufactured by Anadigics' New Jersey GaAs foundry, in volumes of more than a million a year, at less than $5 per part. The capability of this type of technology has now stimulated interest in telecommunications applications of a similar approach, aimed at bringing voice and data capability to the consumer without the need for costly cable laying and network hook-up. Quite obviously these types of market are taking GaAs manufacturing and packaging technology beyond the traditionally perceived low-volume markets, and a new type of aggressive cost-sensitive development is taking place.

2.3 MOTOROLA 'ALTAIR' LAN APPLICATION

Building on the successful experience derived from programs such as the 'ALTAIR' Wireless Ethernet project [3], Motorola have in recent years established a $100 million GaAs IC facility handling 4" wafers, and are currently preparing to run 6" GaAs through their process as substrates of the required quality become routinely available.

The ALTAIR system uses 18 GHz wireless links to provide up to 15 Mb/s of data connectivity in a LAN (local area network). Key to the low cost and flexibility of this application is the implementation of the microwave functions on GaAs MMIC (monolithic microwave integrated circuit) chips operating at between 2 and 20 GHz. In

ramping up to volumes of several thousand assemblies, Motorola commented on various lessons learned, some of which will be discussed in the next section. With their extensive Si experience, especially in statistical process control (SPC) of manufacturing processes, and the added advantage of vertically integrated access to radio system developers, Motorola are well positioned to couple their major fabrication capability with the exploding wireless communications business in a very competitive consumer marketplace.

2.4 WESTINGHOUSE PHASED ARRAY RADAR SYSTEMS

The whole field of GaAs technology in North America (and to a large extent also in Europe) owes a major acknowledgment to the support received over the years from military funding programs. Although recent years have seen a considerable swing away from that type of application, there are still several companies pursuing the application of GaAs to the perennial problem of cost-effective (and weight-effective) manufacture of large arrays (around 3000 RF modules per face) for use in beam-steerable radar systems. The advantage of this type of design is that each module embodies individually-controllable transmit and receive functions, such that the focal point of the radar beam may be steered electronically rather than mechanically, thus offering a much greater degree of agility to deal with multiple hazards at different heights and distances.

Especially for airborne systems, the integration possibilities and performance advantages of GaAs IC's enable the required functionality to be concentrated into a minimum volume, and Westinghouse Electric have been able to apply their long experience in all aspects of GaAs technology [4,5] to this subject. Among topics currently being addressed in their planned evolution, the use of HBT (heterojunction bipolar transistor) technology offers improved linearity and power efficiency across the very large number of output stages in a given system. Recent announcements indicate that Westinghouse is further enhancing investment in their GaAs capability to supply commercial MESFET, HBT and p-HEMT IC's.

2.5 SUPERCOMPUTER APPLICATIONS

The supercomputer industry has had an interest in GaAs that focuses more on high level integration of high speed digital functionality, and has met with rather more severe challenges than the microwave side of the business. This type of need has driven work aimed at high performance in a context of reproducibility and uniformity as required for VLSI GaAs. Early in 1991, Convex Computer Corporation announced the development of a 2 Gigaflop supercomputer based on GaAs IC's in combination with other high speed technologies. By October of that year they had successfully delivered their first C3800 machines based on an architecture using as many as 30 different GaAs IC designs with up to 45,000 gates/chip. The Convex experience grew out of confidence gained in GaAs substitutions for high speed Si bipolar IC's in earlier machines, which enabled equivalent performance but for lower dissipated power. The IC's were realized through a close and successful collaboration with Vitesse Semiconductor, whose self-aligned gate (SAG) FET process with direct coupled FET logic dissipated the low power levels that enabled the C3800 to be air-cooled.

In contrast to the air-cooled Convex machine, the CRAY-3 machine from Cray Computer Corporation was based on an innovative packaging approach using total liquid immersion of several thousand MSI (medium scale integration) GaAs chips for

effective heat removal. This architecture, together with the 100 ps gate delay of the GaAs logic chips gave the CRAY-3 its 2 nanosecond clock speeds; more than twice the overall performance per CPU as compared to its CRAY-2 predecessor. The first CRAY-3 went into service in 1993, and work at Cray was focusing on more highly integrated GaAs chips using a SAGFET technology for the 1 nanosecond clock CRAY-4, until the recent announcement of the unfortunate demise of the Cray Computer Corporation. At the time of writing, efforts are underway to sustain the 4" captive GaAs facility developed at Cray, with a view to helping meet the current increased demand for microwave GaAs.

Both the Cray and Convex programs have suffered from the drop in market share of the supercomputer manufacturers rather more than from inadequacies of the GaAs technology *per se* although it is clear that the demands of highly integrated GaAs digital circuitry require a greater degree of process maturity and uniformity than is sufficient for microwave IC's. In addition to this constraint, both of the supercomputer applications pushed GaAs into a ground breaking domain of high-speed interconnect and packaging. At least in the Cray case, the packaging issues rather than the GaAs capability seem to have been the limiting factor in the commercial viability of the project.

2.6 GaAs IN TELECOMMUNICATIONS: NORTHERN TELECOM

In 1989 Northern Telecom introduced a family of fiber transmission products called "FiberWorld" which has become a world-leader in high-capacity transmission systems, and makes extensive use of GaAs IC technology. The system architecture is based on optical links running at rates from 150 Mb/s (OC3) to 600 Mb/s (OC12) and as high as 2.488 Gb/s (OC48) using the SONET protocols for signal multiplexing and assembly. The "FiberWorld" family also includes Digital radio links running at 4—8 GHz, frequently used to complement the fiber capability in regions where fiber installation may be difficult or inconvenient (for example in mountainous terrain).

Decisions of where to use GaAs, and why, varied from one application to another. In the case of the 2.4 Gb/s digital multiplex and demultiplex functions, as well as the clock-recovery circuitry, the GaAs technology was the only solution available in that time frame that offered all of the following: design margin sufficient for state-of-the-art manufacturability; sufficient integration levels; acceptable power dissipation; and appropriate packaging technology. For the microwave circuits the semi-insulating substrate was a key factor enabling the cost-effective integration onto the GaAs chips of key components such as balanced inductors, digital switches and microwave amplifier and mixer stages. In all, about 12 different GaAs chip designs were used in various parts of the FiberWorld products.

More recent Northern Telecom products have used GaAs MMIC's in a set of microwave modules used to perform transmit and receive functions in a 900 MHz cellular radio basestation. This "Dual-mode radio" product offers the capability of analog or digital interfacing for cellular users into the rest of the network capability of Northern's transmission and switching systems, and requires upwards of 100,000 GaAs IC's/year.

Sourcing for these parts includes both internal and external suppliers (TriQuint, Vitesse, Anadigics, Harris/Samsung, Fujitsu), and, as noted by other shared-source users (e.g. Motorola), a close relationship between chip developer and supplier is essential to a smooth introduction of the new technology into volume product. Now, as the products mature, further generations of GaAs capability are being used to improve

performance, reduce power dissipation, increase the integration level of separate functions, and in some cases (e.g. at 150—600 Mb/s), functions that previously needed GaAs can now be done with high performance silicon technology. Recent advances in SiGe bipolar technology are enabling some impressive laboratory level performance figures, however most of the transmission technology manufacturers developing OC192 (10 Gb/s) systems seem to opting for the more generous performance margins offered by GaAs-based heterojunction bipolar technology.

3. Lessons Learned in the Introduction of GaAs Technology

3.1 UNDERSTANDING THE MATERIALS CHARACTERISTICS

The materials technology of the GaAs system is inherently more complex than that of Si, and this inevitably also contributes to a process technology requiring control of more degrees of freedom. GaAs technologists have learned many valuable lessons concerning the material purity, its stability under thermal treatments, and especially the sensitive nature of its surface and interface chemistry. Even armed with this new knowledge, the Si technology still has a major advantage of many years of high volume manufacture, and the authority and empirical understanding conferred by that history carry considerable weight in comparisons with the III-V systems.

As an example, the Si circuit designer knows from detailed statistics that for a given geometry, the performance of his design medium will follow a well-defined set of rules, within quite close tolerances. Even the degradation of the tolerances at progressively smaller dimensions is relatively predictable. The GaAs designer, on the other hand, has to contend with a much smaller body of statistics for his prior data, and even now is dealing with relatively small batches of "identical" substrates from each boule. As a result, to ensure that his circuit will operate over the required range of operating conditions and process parameters, the designer must expend extra performance margin, power budget, and even real estate, to guarantee the objectives. This maturity penalty for GaAs IC's is eroding more rapidly now as large volume applications become established, but it has severely narrowed the predicted performance gap between GaAs and Si for a number of situations.

Such an argument admits that GaAs is still on the learning curve, and this is both good and bad news. It is good in that the performance margin against Si can be expected to widen as volumes further increase, but bad in that a major area for improvement is the relationship between starting materials and wafer processing technology, which is all too often a "vendor-supplier" one, rather than a true team effort with open understanding on both sides. Nevertheless, the performance/cost combination still leaves much encouragement for GaAs, such as illustrated by the comparative extrapolations of cost vs. yield for GaAs IC's by Skinner [6], who demonstrates the extent to which (for large high speed IC's), packaging and testing dominate the cost much more than the contribution from the cost of the processed wafers.

3.2 THE "DEEP TRAP RAT-HOLE"

One of the key beneficial aspects of GaAs technology is the semi-insulating nature of the high resistivity substrates which enables the close proximity of high frequency components with low parasitics and minimum interaction between adjacent devices.

Assuring the stability of the insulating condition and especially its electrical completeness at all frequencies, has been a major subject of study over the last 15-20 years. The unusual physics of this deep-level system has enabled the database to be substantially enriched by research programs from many related disciplines and has provided the materials growers with a wealth of models and solutions by which they may understand and control the seemingly delicate semi-insulating situation.

Unfortunately, the subject of deep level studies became so interesting in its own right that the burgeoning body of data eventually added little to the knowledge of how best to control the material, and perhaps even started to cloud the issue. Ultimately the solution evolved more out of practical adaptation based on statistical experiments at a manufacturing level, than out of a detailed understanding of the processes involved.

A feature of this is that, even now, many different GaAs fabs use substantially different "recipes" for dealing with implant annealing, backgating etc., and this translates to different demands presented to the substrate suppliers, and thereby hinders the convergence of requirements of the base material, and so sustains higher substrate costs.

3.3 SHARED RESPONSIBILITY OF DESIGN/WAFER FAB

It is inevitable that during the parallel development of a new IC technology and the circuit designs that will use it, some mismatch will occur between the needs of the designers and the readiness or optimization of the wafer fab. At this point, especially in a race for market share, the pressures to deliver can easily result in situations that are less productive than "perfect teamwork", especially if the two groups are parts of different organizations.

Many new product developers have commented from their experiences that the notion of IC design on the basis of "Foundry PCM (process control monitor) Parameters" is fundamentally flawed in that no PCM set can perfectly guarantee operation of the circuit, and reciprocally, no foundry would wish to be constrained as tightly as the spreads of an "ideal" PCM set would require.

The resultant compromise has to be an atmosphere of close cooperation between designer and wafer fab, and this is most often achieved by the two functions being part of the same organization. In many cases, development groups have found that a mutual training activity in the form of in-house "User-design courses" serves both to sensitize product and system designers to the new technology's capabilities and limitations, and to create a shared sense of ownership in advance of the issues that are bound to arise.

Most of the original GaAs foundries have now moved towards a basis of more "selected customers", and have recognized that the notion of "uncommitted ASIC multi-project chip designs" tends to show a relatively low yield of true product developments, whilst requiring substantial levels of engineering support. For this reason, most new product developments for GaAs now take place either as customized ASIC programs, (with the associated NRE costs) or as in-house developments in a vertically integrated organization.

3.4 FULL EVALUATION OF VULNERABILITY/RELIABILITY

In spite of the fact that they are more proof against damage from radiation effects (because of the absence of an oxide layer), GaAs devices and IC's are potentially sensitive to ESD (electrostatic discharge) damage, as many manufacturers have

discovered by experience. An oft repeated adage is that *"you don't realize that you had a major ESD loss problem until you have installed all the measures to eliminate it!"* This is due in part to the latent nature of many forms of ESD damage, which will only show at some later stage as a failure apparently due to something else. Operating practices and disciplines need to be upgraded and constantly reinforced to eliminate this problem. A typical example frequently encountered in system development labs is the use of heat guns to either test the temperature sensitivity of a suspect GaAs part, or even to remove it from the board. Since heat guns are an excellent source of charged ions, the suspicions about the GaAs part were often proven apparently true as a result of this barbaric treatment.

True reliability data on a brand-new technology can only be obtained after the requisite number of hours of in-the-field operation on a statistically significant batch of samples. Nevertheless, useful predictive data can be derived from studies based on accelerated life-tests, although the variety of activation energies quoted for similar GaAs technologies suggests that some subjectivity is still involved. In general, for GaAs FET's using recessed-gate technologies (usually with Pt or Pd and Au gate electrodes), the predominant failure mechanism over time is referred to as a "sinking-gate" mechanism, and this is generally slow enough not to be an issue in most applications. In technologies using refractory metal gates (such as SAG processes with WSi gates) the ability of the gate metal to withstand the high implant temperatures means that it is very stable at typical operating temperatures, and the failure mechanism of the Ohmic contacts then appears as the next "layer of the onion".

More recent studies have highlighted an additional mechanism that occurs when recessed-gate devices (using Pt or Pd gates) are exposed to hydrogen-containing atmospheres in a confined enclosure. This can occur when a recessed-gate GaAs IC is hermetically sealed in a Kovar package, since the outgassing of H2 from the walls of the package is sufficient to provoke an effect [7]. Although this effect is now understood and largely under control, its recent appearance highlights the need for exhaustive reliability evaluation of new technologies before they are introduced to field applications.

3.5 IMPLICATIONS OF PACKAGING REQUIREMENTS

The spectacular operation of a new high speed technology is of little value if it cannot be communicated to the rest of the system, and yet for the most part the packaging of a GaAs IC is frequently treated as a design afterthought. Inappropriate termination of digital input/output cells or poorly-matched RF ports on a microwave IC can seriously degrade the simulated performance of the circuit, and both cases have caused GaAs designers to address issues of high speed packaging necessary to deliver their commodity.

Indeed, one of the strengths of a multi-chip-module (MCM) approach has been that it encompasses all the "troublesome" components into a single well-controlled unit that can be tested for total functionality before being put onto a circuit board. It is likely that, as individual component yields improve, and as the tolerances and losses achievable with chips auto-placed on high-quality printed circuit boards permit, then the need for separately testable enclosures will face serious competition on a cost basis.

Another of the issues addressed by the packaging environment is that of heat dissipation, as an important factor in determining the correct conditions of operation, and especially for maintaining the correct junction temperature from a reliability

viewpoint. Key to the understanding of this facet of IC design is the ability to correctly model the thermal characteristics of the circuit, including the package and heatsinking hardware. In spite of the rigorous demands of 'mil-spec' quality requirements, the need to satisfy commercially-competitive environmental specifications in adverse customer premises applications (e.g. usable on the wall of a house in either Alaska or Africa) is driving the packaging technologies to a new domain of cost-effective solutions.

3.6 REALISM OF PERFORMANCE CLAIMS

It is fair to say that some of the early GaAs technology proponents oversold their claims, frequently promising performance advantages based on isolated devices, but without taking account of manufacturing and operating margins. A consequence of this enthusiasm was the disappearance of a number of the early GaAs enterprises, and a major rethink for some of the survivors. As mentioned in section 3.1 above, a realistic designer is obliged to trade some of the performance margin to compensate for non-uniformity of devices within a circuit as well as interconnect losses, *etc.* This cautionary comment is intended to warn against comparison of technologies in differing contexts. A research result on an isolated device, or a carefully chosen and "tweaked" circuit does not necessarily represent what can be achieved on a manufacturable basis. It is, however, curious to note that, whereas a few years ago Si defenders were pointing out that GaAs results were only obtained on laboratory specimens, the reverse situation now prevails, where GaAs IC's are being assailed by highly-tuned Si circuits from university research groups! A sad fact is that a number of otherwise promising devices have been culled from the route to development because they were unable to meet their original expectations, even though in some cases the basic design contained intrinsic merit.

3.7 "APPLES AND ORANGES"

This section considers issues of comparability, somewhat as discussed in the preceding section, but particularly in respect of digital GaAs. This is an area where the Si community has sustained the strongest opposition to GaAs, even though as demonstrated by Vitesse, it is possible to deliver higher speed performance at lower power dissipation than for bipolar Si. The pitfall here is that for GaAs to achieve system insertions normally occupied by Si, there is usually some modification of the system architecture needed to optimize the overall performance to the power supplies, gate counts, and switching speeds appropriate to GaAs. There seem to be very few cases where the full accommodation is allowed, with the result that most of the GaAs demonstrations are disadvantaged in some way.

On a basis of cost analysis, the sheer scale of Si technology (wafer size, fab throughputs, *etc.*) renders a dubious comparison with GaAs, even though for many Si technologies the number of process steps (or mask levels) considerably exceeds the simplicity of a GaAs MESFET process. Again, reference to the article by Skinner [7] illustrates that an 80,000 sq. mil. digital IC made in GaAs would cost only 15% more than the same device made in Si at the same yields. Whether or not yields can be comparable is a relevant question at this stage, although for GaAs there is still much progress to be made in manufacturing maturity and materials costs/wafer sizes, whereas Si is in a much flatter part of the learning curve. The relative simplicity of the GaAs

process (largely a consequence of the semi-insulating substrate) is clearly one of its strongest redeeming assets.

My opinion on the future of VLSI GaAs is that if it can survive the next few years without getting "killed off" by pressure from silicon technologies, then the increasing maturity, volume and stability of the GaAs technology will help to diminish the remaining barriers, and will provide an opportunity for the GaAs designers to realize a true advantage of the high speed and low power capability at reasonable (100k gates/chip and above) levels of integration.

4. Selection Rules for a New Technology to Reach Commercial Viability

It is valid to question at this point whether or not commercial viability should be taken as an objective for selecting future technologies, especially given that some new technology ideas which are not in themselves commercially viable may nevertheless spark off a more cost effective version that does provide a better end-result. A stimulating discussion of the subject by Kroemer [8] highlights the fact that pushing towards applications solutions and cost effective mass-production tends to limit creativity and stifle original approaches (such as described in the example of VLSI digital GaAs development mentioned above).

Nevertheless, my aim here is to consider what type of features of a technology may enhance its chances of survival to the point of usefulness, in the hopes that fatal setbacks experienced by other technologies could be averted, or at least anticipated. Many of these features go beyond the normal technical considerations that might be addressed in a research paper, and embrace significant aspects such as economic factors, timing and competition, related supporting technology demands, *etc.*

4.1 "TWENTY QUESTIONS ANALYSIS"

The following set of "twenty questions" could stimulate a "pre-graduation" analysis of a proposed new technology. Such an analysis could of course reject a very interesting technology simply because there is no interest for it *at that time.* As in the analogy of show business, timing and marketing can dominate the effects of talent!

The questions are then summarized in Table 1, and (very subjectively) applied to a comparative analysis of the GaAs MESFET against the Permeable Base Transistor (PBT) and the type of multi-state resonant tunneling devices referred to as Quantum Functional Devices (QFD).

4.1.1 *Are there multiple ways to make this device?*

Since the first publications on GaAs FET's, different variants have co-existed (e.g. junction FET's, SAGFET's, heterojunction FET's, recessed gate FET's, power FET's *etc.*) and have learnt from one another. As in husbandry, this type of cross fertilization can strengthen the species.

4.1.2 *Are there multiple applications for this device?*

GaAs was fortunate in being able to address microwave, digital and analog applications with only slight variants of the same technology. At different times in its history, predictions of the "winner" have varied. Currently it appears to be microwaves, but that

could change, and meanwhile the manufacturing base is growing and maturing. A "single-application" technology is very vulnerable to the competition. As in show-business, versatility helps!

4.1.3 *Is the technology simpler than that of its mature competition?*
As mentioned in section 3.2, GaAs technology uses only 11-12 mask levels, as compared to 18-22 mask levels required for Si bipolar or BiCMOS technologies. This simplicity has helped to offset the increased costs of the raw material and low yields during development. Most commercial GaAs FET technologies are also implantation-based, rather than dependent upon epitaxial material.

4.1.4 *Are there applications for which this device is the only solution?*
This question depends very much on context; for Northern Telecom, GaAs was the only solution for 2.4 Gb/s digital IC's in 1989, and around that time the same choice was the only cost-effective way for Continental Microwave to meet their microwave requirement for DBS. The issue here is one of driving force.

4.1.5 *Is there enough performance advantage to trade some of for manufacturing margin?*
For GaAs, many of the early claims of performance advantage had to be scaled back somewhat to meet the demands of production margins, non-uniformity, packaging losses etc. If a new technology only has a marginal advantage over other alternatives, this may disappear altogether in manufacture.

4.1.6 *Is there a packaging/interconnect solution for this device?*
Especially at higher frequencies, the ability of a device to drive interconnect is crucial to its survival, but especially for a device seeking to move into a new frequency domain, the developers must prepare a suitable package with which to interface their new function to the outside world.

4.1.7 *Does the technology benefit from conventional processing equipment?*
Much to the advantage of GaAs, its processing was largely based on conventional process steps already developed for Si. This continues to be of value as many 4" GaAs fabs equip themselves with ex-Si machines at reduced prices. A new technology that requires substantial process technology development will have a very expensive (and perhaps prohibitive) learning curve.

4.1.8 *Are there fully descriptive device models ready for IC design use?*
Assuming that the aim of a new electronic device is to become part of an integrated circuit, it is necessary to provide a representative set of (scalable) credentials for the designer to manipulate and simulate into a circuit. In most cases, if the models and design system are not there, the system designer will look for another technology, and the designer's opinion counts heavily !

4.1.9 *Does the technology allow for testing prior to process completion?*
Though not essential, it is very helpful if a device's functionality can be verified to some extent, at some intermediate point in the process sequence. This can reduce the cost of processing failed wafers, and expedite the feedback needed to correct certain process steps.

4.1.10 *Are costing estimates for this device technology complete?*
When offered better performance, system developers will ask "what does it cost me?" A smoother journey to commercial viability is assured if the technologist can offer a

description of the costs that anticipates *all* the potential contributors: materials, fabrication, yields, testing, packaging, test and package development, reliability programs, and especially how these costs will evolve as the technology matures.

4.1.11 *Does the device have a weak point, fatal flaw or Achilles' heel?*
If there is some type of operating condition that is highly unfavorable to the device, it is better to be "up-front" about it, and to propose design considerations or a development program to alleviate the impact of the potential weakness.

4.1.12 *Does the device need unusual or extra power supplies compared to competitors?*
Certain digital GaAs architectures have found difficulty as "drop-in" replacements for Si, simply because the logic family needed different power levels. Compliance is better unless it seriously undermines the performance advantage.

4.1.13 *Will the device need to push the edge of some associated technology envelope?*
As in 4.1.6 above, the technology developer may also be faced with developing some aspect of a related technology on which the new device depends, such as special epitaxy (e.g.: overgrowth of metal arrays for the PBT), lithography (for nanodimensional devices), special packaging solutions, cooling technologies for cryogenic devices etc.

4.1.14 *Are there environmental factors e.g. toxic chemicals, limited raw materials?*
Early allegations suggested that the world supply of metallic gallium would limit the volume use of GaAs devices; this has not proved to be a problem yet. The toxicity of As does not emerge as a significant problem either, although some of the metal-organic chemicals used in epitaxy of certain III-V compounds are being studied to find more acceptable substitutes.

4.1.15 *Are there major technical or commercial risks involved?*
This aspect could reflect instances where several technology areas (e.g. 4.1.6, 4.1.7, 4.1.13) could compound to represent a very high level of necessary investment to ensure success. If this is coupled with an uncertain marketplace and a unique application, then risk is high.

4.1.16 *Are there any potential reliability exposures or sensitivities?*
The first heterostructure devices were faced with proving that their atomically thin layers were not a major liability. Nevertheless, some surprises can occur in a new technology, e.g.: the "purple plague" in Si IC's, or the H2/gate issue mentioned in 3.4 above.

4.1.17 *Does the device require any special operating conditions (e.g. cryogenic cooling)?*
Josephson junction devices are a clear example of a technology that had to support a major "selling" exercise to market the ultra-fast functionality available. Even with a solution in hand, the unfamiliarity of a major constraint can seriously undermine the confidence of investors!

4.1.18 *Is the operation difficult to explain, or the name hard to remember?*
This may seem trivial, but the new idea will have to be "sold" to prospective supporters of a non-technical background, and if the barrier of comprehension and retention is too high, then the chance of obtaining the required support may be reduced.

4.1.19 *Does the device face comparable responses from competing technologies?*
One of the best things to happen to Si over the last 20 years is GaAs. The competition of GaAs pushed Si technology into new areas that might otherwise have taken longer to

emerge. By the same token, the advances of Si sometimes surprised the otherwise complacent GaAs engineers!

4.1.20 *Are there any patents limiting free competitive development?*
Given a marginal choice between two competing technologies, the one which is constrained by legal limitations may face a battle on a different playing field, at least until the rights of the party concerned expire, or the limitations are removed.

4.2 *Comparative table of technology assessment*

TABLE 1. Questions applicable to the evaluation of a new technology; for 1-10, 'Yes' scores 5% and 'No' scores zero; vice versa for 11-20. 'PBT' refers to Permeable Base Transistor [9] and 'QFD' to Quantum Functional Devices [10].

QUESTION	***GaAs FET***	***PBT***	***QFD***
1. Multiple ways to make?	Y	N	Y
2. Multiple applications?	Y	Y	N
3. Simpler technology than competition?	Y	N	N
4. Only solution for some applications?	Y	N	N
5. Performance headroom to trade for mfg margin?	Y	Y	N
6. Is there a packaging solution?	Y	Y	Y
7. Uses conventional processing?	Y	N	N
8. Are there models for IC design?	Y	Y	N
9. Testable before process completion?	Y	N	Y
10. Does costing account for everything?	N	N	N
11. Any weak point or "Achilles heel"?	Y	Y	Y
12. Need unusual power supplies?	Y	N	Y
13. Pushing envelope of support technology ?	Y	Y	Y
14. Any environmental limitation?	N	N	N
15. Major technical or commercial risks?	Y	Y	Y
16. Potential reliability exposures?	Y	Y	Y
17. Need special operating conditions?	N	N	Y
18. Complex operation or description?	N	N	Y
19. Significant competition?	Y	Y	N
20. Patent constraints?	N	N?	?
Scores:	65%	45%	25%

This attempt to quantify the judgments listed above is not intended to be either exhaustive or exact; for example there is no real assessment of the magnitude of technical advantage of a proposed technology, nor is there any relative weighting of the various aspects in generating a "score". The two examples chosen for comparison against the GaAs FET are the PBT [9] which came into prominence around the 1980s, and the QFD-type of devices, as described for example in a recent review article [10].

The PBT is essentially a vertical field-effect transistor where the gate is made as an array of electrodes buried within an epitaxial structure. Pinch-off is then between adjacent fingers of the gate electrode (and therefore more effective than between a gate and a hi-resistivity interface), and the gate length is determined by the thickness of a deposited metal layer, and is therefore potentially very short indeed. Excellent simulation work showed the potential for this device to offer very high frequency performance, but in spite of the elegant nature of the structure, the practical problems associated with the overgrowth eventually lead to conclusions that the device would be overly difficult to manufacture. Variants of the structure (e.g. by implantation of the gate regions) offer some of the useful features, but sacrifice many of the advantages.

Quantum Functional Devices refers to the various types of III-V (or Si-based) devices that aim to use tunneling between well-defined states to create I-V characteristics with regions of NDR (negative differential resistance). With such a multi-state characteristic, circuit designers could potentially realize certain logic functions or frequency-doubling operations with a single device where normally several devices would be needed. The potential savings in device count, interconnect, and chip size could be significant, however the battle is currently to achieve the requisite functionality at room temperature and above, and in a manufacturable architecture. Many of the "answers" in the table are arguable, and their status will certainly improve given the level of activity on these types of devices.

5. Trends in the Development of New Technologies

The discussion of the previous section highlights the fact that a new technology usually needs a *substantial* performance advantage if it is to overcome the various other hurdles *en route* to industrialization. Typical advances now tend to be factors of 2 or 3 in speed or power or size. The QFD family may ultimately provide a factor of 5-6 in integration, and possibly some associated speed/power advantages, but only with a much more complex technology. Clearly the stage is set for a major change that could offer improvements of 10-100 times, but that sort of advance does not seem to be on the path of current semiconductor developments.

5.1 SELECTION OF DEVICE STRUCTURES

New devices are usually selected for further development on the basis of reported characteristics of peak performance, but only rarely for intrinsic advantages of easier or cheaper manufacturability. Since the answer to future requirements may not always mean faster or smaller, it is helpful to reconsider the criteria. Most devices now stem from evolutions of the basic 3-terminal function, whether bipolar or field-effect in operation, however an ideal device would have totally flexible coupling from input to output, with no committed common terminal. Possible new architectures could perform powerful single-device operations (such as the QFD) and have multiple inputs and

outputs, with options for optical interaction or programming of interconnect. Of course the right technological environment is necessary to encourage such progress.

5.2 THE TREND TO MEGAFABS

Some may take exception to the recently-quoted phrase "Real men have Megafabs!" Apart from the political incorrectness, this suggests that the only solution to cost-effective manufacturing is scale, and that the only successful technologies are the very high volume ones. It is easy to show that wafer production facilities operating with large batches of 12" wafers are inhibiting to the development of small prototype circuits for low-volume applications. Even with 4" GaAs, the opportunities for, say, university groups to exercise design ideas in the medium are severely constrained by the needs of the big bill payers, and flexibility of design rules or rapid turn-around are not generally available. In the pursuit of lower costs in a given medium, the megafab philosophy may be creating a monster so large as to find difficulty in maneuvering for survival when a different, much cheaper technology presents itself.

5.3 MINIATURIZATION AND MICRO-ORGANIZATION

Spurred on by ingenious developments in micro-lithography, the scaling of all semiconductor devices offers a route to faster performance and higher levels of integration, both of which can translate to lower cost. As the various technologies compete by sharpening their gate lengths or increasing the sophistication of their materials structures, the number of carriers per active region (electrons/channel etc.) becomes smaller, to the point where the information contained in a given device state becomes of questionable stability in relation to thermal or electromagnetic upset influences.

At the same time, the smaller devices have to be connected together and yet the ratio of interconnect to device is becoming larger and more of a proportional burden on the device's drive capability. This tendency to increased organization at smaller physical dimensions is heading towards an era of diminishing returns. It would now be appropriate for device designers to consider ways of working with larger arrays of less-well specified (even potentially redundant) devices and the statistical handling of such operations for effective computation *etc.* This may involve a more malleable medium of interconnect than the current multi-level metal/poly schemes, and should certainly pursue the development of the third dimension and address the associated issues of heat extraction. The notion of reorganizational capability within a circuit suggests some of the attributes of neural networks.

5.4 NEW DEVICES AND MATERIALS

Neural network technology has many attractive features, but generally involves very high component counts. Most of the development has therefore tended to focus on architectures that may be realized within conventional VLSI technology, and relatively little work has looked at the possibilities of a substantially new technology specific to neural functions. Features such as variable interconnect, variable thresholds and weighted responses are key to the adaptability and learning incentives of neural networks, but are difficult to achieve in hard-wired semiconductor technology. Nevertheless, the mix of functionality possible within, say, the III-V family (optical,

digital, analog) presents some interesting options for e.g. integrated retinal detection/processing, and some operations such as optimization of digital signal recovery from optical signal streams might be more efficiently done using a neural approach than in the conventional manner. It seems unlikely however that such architectures will really "take off" until they are given a more sympathetic medium in which to operate. When that happens, the learning derived in the current technology will be able to bloom.

Perhaps a candidate for the necessary material for neural architectures will be polymer electronics. Although the mobilities of conjugated polymers are substantially less than we use in semiconductors, the cost is potentially much lower, the volume usage more efficient, and the potential for applications more flexible (in the true sense of the word!). It is in this context that one should imagine how the semiconductor industry would respond to competition from a medium that was so low cost that volume became no object, and could be conformally applied to the walls/support structures of any mechanical equipment. There is also potential for an integrated marriage between the technologies whereby slower operating layers of e.g. polymer memories/neural routing networks could be deposited onto conventional higher speed circuits to offer hybrid functionality and intelligent interconnect.

6. Conclusions, Ideas, Provocations

An overriding theme of the discussion in this paper has been towards application-oriented development. It is tempting to say that the applications must therefore be clearly identified up front, but history shows that many technologies have to be developed and matured before their best applications are found. This suggests that pursuing the trend to cheaper functionality for larger volume applications (because that pays the bills) may eclipse some valuable new initiatives. Perhaps by focusing further out on more radically different technologies, we can create some extra flexibility to distract developers from the limiting trends currently underway. Key features for future emphasis should be:

- Anticipate the "end-of-the-road" for miniaturization of current devices.
- Avoid the technology development being dominated by "megafabs".
- Pursue the development of small less-precise devices and statistical architectures.
- Encourage true 3-dimensional device and interconnect topologies.
- Develop some cross-disciplinary bridges to e.g. polymer electronics to enhance the intermediate benefits and lead into the next generation.
- Be aware of the advantages of an easily-manufacturable technology.
- In taking a new technology through to industrialization, expend more up-front effort on the surrounding issues to smooth the transition.

7. References

1. Jay, P.R. *et al.* (1991) *IEEE GaAs IC Symposium: Short Course Handbook.*

2. Peterson, V. (1988) Applications of GaAs IC's in instruments, IN *Proc. 10th IEEE GaAs IC Symposium, Nashville, TN*, pp. 191-4.

3. Malone, H. R. *et al.* (1991) High volume GaAs MMIC applications, in *Proc. 13th IEEE GaAs IC Symposium, Monterey, CA*, pp. 135-138.

4. Thomas, R. N. (1980) Large-diameter, undoped S-I GaAs for high mobility direct ion-implanted FET technology, in G.J. Rees (ed.), *Proc. 1st International Conf. on Semi-Insulating III-V Materials, Nottingham, 1980*, Shiva Publishing, UK, pp. 76-82.

5. Shannon, L. C. (1990) Increased yield of microwave devices due to subsurface damage reduction in SI GaAs wafers, in A.G. Milnes and C.J. Miner (eds.), *Proc. 6th International Conf. on Semi-Insulating III-V Materials, Toronto, 1990*, Inst. of Physics Publishing, Bristol, UK, pp. 359-366.

6. Skinner, R. D. (1991) What GaAs chips should cost, in *Proc 13th IEEE GaAs IC Symposium, Monterey, CA*, pp. 273-276.

7. Hu, W. W. *et al.* (1994) Reliability of GaAs p-HEMT under hydrogen containing atmosphere, in *Proc. 16th IEEE GaAs IC Symposium, Philadelphia, PA*, pp. 247-250.

8. Kroemer, H. (1994) Devices for the Future: a peek into the next century, in *Ext. Abstracts of 1994 International Conf. on Solid State Devices & Materials, Yokohama*, pp. 397-399.

9. Nichols, K. B. *et al.* (1985) Fabrication and performance of GaAs permeable base transistors, in *Proc. IEEE/Cornell Conf on Advanced Concepts in High Speed Semiconductor Devices & Circuits, Cornell*, pp. 61-71.

10. Goronkin, H. *et al.* (1994) Progress in quantum functional devices to overcome barriers to ULSI scaling, in *Proc. 16th IEEE GaAs IC Symposium, Philadelphia, PA*, pp. 9-12.

COMMENTS ON THE NATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

JAMES F. FREEDMAN
Semiconductor Research Corporation
79 Alexander Dr., Bldg. 4401, Suite 300
Research Triangle Park, NC 27709

Abstract

The SIA National Technology Roadmap for Semiconductors (NTRS) [1] represents a coordinated effort by leading technologists from all sectors of the U.S. semiconductor infrastructure. Led by industry and coordinated by the Semiconductor Research Corporation (SRC) and SEMATECH, this effort involves industry, academia and government agencies in defining a unified description of the semiconductor technology requirements for ensuring advancements in the performance of integrated circuits. The NTRS provides a fifteen-year horizon, extending through the year 2010 and covering an anticipated six generations of future product and technology needs.

1. NTRS Methodology

Building on the experience gained from the first roadmap [2, 3], the current revision uses a structured approach to assure unification. A Roadmap Coordinating Group was established to define the overall strategy and a standardized template. This resulted in the creation of eight Technology Working Groups (TWG's). Each TWG represents a technology area critical to semiconductor product development and manufacturing operations and is assigned the responsibility of identifying the current technology status, the roadmap of technology needs, potential solutions, other dependencies and potential paradigm shifts. The eight TWG's (technology areas) are: 1) Design and Test; 2) Process Integration: Devices and Structures; 3) Environment, Safety and Health; 4) Lithography; 5) Interconnect; 6) Materials and Bulk Processes; 7) Assembly and Packaging; and 8) Factory Integration.

Since each TWG has a limited number of members, an open workshop was held to solicit a broad critique of each TWG roadmap prior to publication and release. There was nation-wide representation and the open discussion resulted in a consensus document.

2. NTRS Purpose and Scope

A common misconception of the roadmap is that it is a literal forecast of the integrated circuit (IC) industry's future technology. However, the document identifies only the problems and possible approaches, but not the solutions, and is intended not only to

S. Luryi et al. (eds.), Future Trends in Microelectronics, 87–91.

preserve, but to encourage innovation. In the document, the evolution of technology is graphically described through an analogy with paved roads, unpaved roads and footpaths. Basic research, now concentrated heavily in universities, provides the first explorations of the technology "countryside" by providing an understanding of the lay of the land and creating the first footpaths. The main purpose of the NTRS is to adequately define the industry's future needs so that the basic research will be selected to solve specific problems, therefore defining where the footpaths are located, as well as their direction. This "problem-driven" motivation is key since it is a well-known historical characteristic of roads that those chosen to be developed and improved are those most widely used.

3. NTRS Framework

To provide a unified framework for the eight TWG's, an Overall Technology Characteristics Chart (OTCC) was defined that projects the key elements of the technology areas. The framework of the current Roadmap, besides being extended to 2010, was broadened over the 1992 version by segmenting the characteristics according to the IC product design, since different applications drive different aspects of the technology. The OTCC major market areas addressed are: memory, including dynamic random-access memory (DRAM) and flash memory; high-volume microprocessors; and low-volume application specific integrated circuits (ASIC's). In all three areas, the primary driving assumption is Moore's Law [4]. A condensed version of the OTCC is shown in Table 1, which summarizes shows the projection algorithm assumed and the requirement in 2010 for the driving market area. For more detailed parameter projections, refer directly to the NTRS.

While the end of technology scaling as defined by Moore's Law has been forecast many times in the past, technology progress has continued unabated. Nevertheless, the key question still remains: when will the industry forecast deviate from Moore's Law?

TABLE 1. Moore's Law Trend Projections

		2010
Die Area	1.5X every 3 years	14 cm^2 DRAM
Min. Feature Size	30% reduction every 3 years	0.07 μm
Transistors/Die	4X every 3 years	64 Gb DRAM
On-Chip Circuit Clock	1.5X every 3 years	1 GHz μP
Cost/Transistor	>50% reduction every 3 years	$0.00001 (Logic)
Fab Cost	2.3X every 3 years	>$10 B

4. Analysis

Although technology scaling has been the driving force for technology advancement, it is the improved cost per function that has generated the market growth and revenues that

have fueled the research and development (R&D) costs. For example, of the 4X improvement in DRAM capacity per generation, 2X has been attained from lithography (1.4X per linear dimension), 1.5X from chip size increases driven by improved manufacturing, and the remaining 1.3X from innovation related to cell size reduction and other technological and architectural advances. Although it is recognized that continuing improvements in lithography are required to meet the OTCC cadence, these alone are inadequate to continue the 30% cost-per-function improvement that has fueled the global growth of the industry.

Consider the following simplified manufacturing cost-per-transistor analysis:

$$C_w = K_1 (FC/(FT \times Y)) \quad (1)$$

Where

C_w = wafer cost
K_1 = constant
Y = yield
FC = capital cost of factory
FT = factory throughput

Therefore,

$$C_b = C_w/(N_c \times N_b) \quad (2)$$

Where

C_b = bit cost
N_c = no. of chips on a wafer
N_b = no. of bits on a chip

Now,

$$N_c = K_2 D^2/(N_b \times L^2 \times PE) \quad (3)$$

Where

K_2 = constant
D = wafer diameter
L = resolution level (lithographic)
PE = packaging efficiencies

Thus,

$$C_b = K (FC/(FT \times Y)) \times L^2/D^2 \times PE \quad (4)$$

Equation (4) demonstrates the importance of the factory (including the use of large diameter wafers) in addition to lithography and packaging (or design) efficiency.

Factory productivity assessments [5] from 1970-1990 show that wafer starts per month increased from 10,000 wafers/mo. to 40,000 wafers/mo., yield increased from 30% to the 80-90% range, and wafer diameter increased from 75 mm to 200 mm, which is the norm in modern factories. The result was a total throughput improvement of 65X when the total factory cost increased by only 10-15X, resulting in a real capital cost decrease. The fifteen-year period covered by the NTRS indicates a potential new (and negative) factory paradigm, where there will be increased capital costs per silicon unit. Clearly, the yield factor has been almost completely exploited. Factory productivity demands the movement to not only 300 mm wafers but also to 400 mm wafers mid-way through the projection period. With the larger wafers and the current equipment utilization, it is questionable whether a further increase in the volume of wafers per month will have any beneficial effects, especially when nanometric shrink technology is driving the need for unit wafer, *in-situ* processing.

5. Challenges

Considerations such as these were factored into the NTRS, resulting in the acknowledgment that the semiconductor industry faces new challenges as it moves to production of feature sizes that are less than 150 nm.

Clearly, the first challenge is to develop the cost-competitive lithographic processes that will allow exposures of sub-150 nm lines. It is generally recognized that optical (deep-UV) lithography will be extended and used down to 0.18 μm. However, this extension requires complex enhancements with a large development expense — both for the tool and for the process. Critical dimension control and overlay are becoming crucial, placing an increasing demand on metrology tools and masks. The challenge is to contain the manufacturing cost associated with this increased complexity. In addition, there is no clear manufacturing solution for dimensions below 100 nm.

Furthermore, the NTRS identified four challenges that span the entire spectrum of technology and require a major effort to resolve. Termed the "Grand Challenges," they are:

1. Productivity improvement: Previous agreements were presented to argue the need for a new paradigm to compensate for the diminished contributions of factory productivity. The demands on "innovation" to increase the cleverness factor must be increased to provide new approaches to satisfy the 30% per-year per-function cost reduction.

2. Complexity management: The historical perspective shows that process complexity increases to accommodate the appropriate manufacturing and design window required for smaller dimensions. This is clearly expected to continue as the wireability levels are increased and dimensions continue to shrink. As complexity increases, a competitive advantage can be generated by developing the engineering support tools that can sustain this increasing complexity. This is also true for device and circuit computer-aided design tools if one expects to control the non-recurring design cost that will allow the design of ASIC chips containing tens of millions of devices. The challenge is to define and solve the problem *before* a crisis.

3. Advanced technology programs: In the United States, global competitiveness has drastically changed the contributions of industrial research laboratories like IBM and AT&T. Universities have been called upon to fill the gap, but many aspects of semiconductor research require a large investment in facilities or equipment to allow state-of-art research to occur. This escalating cost is occurring simultaneously with a decrease in funds and sponsors. The competition for resources places a high demand on immediate applicability of the results, diminishing the support of truly long-term research. The identified challenge is to provide the required distribution of funds and the programs to meet the most demanding technology requirements.

4. Technology funding: The cost of new factory capital and the escalating cost of R&D can only be funded out of current profits. The pressure has already driven the industry to establish cost-sharing consortia like SRC and SEMATECH and to establish individual partnerships and other relationships. The trend towards common research funding will and must continue as solutions are defined to address the issues discussed

above. The reality is that the available funds will never be adequate to address all the issues raised as the industry moves to nanotechnology. The challenge is to implement a funding strategy that covers all *critical* needs and to demonstrate the viability of new concepts prior to funding the development of commercializable solutions.

6. Summary

By focusing on needs rather than solutions, the National Technology Roadmap for Semiconductors provides a broad description of the technology challenges requiring resolution in order to extend the integrated circuit revolution into the 21st century. It realistically addresses the major barriers to this extension, recognizing that the cadence will not continue if these barriers are not removed. The challenges are great, but the promise is even greater.

7. References

1. The National Roadmap for Semiconductors, SIA, 1994.
2. Semiconductor Technology, Workshop Conclusions, SIA, 1993.
3. Semiconductor Technology, Workshop Working Croup Reports, SIA, 1993.
4. Moore, G.E. (1975) Progress in digital integrated circuits, *IEEE IEDM Tech. Digest*, 11-13.
5. Finan, W. and Vardaman J., Semiconductor clean room technology in the 1990's, *1990 Dataquest Report.*

CRITIQUE OF REVERSIBLE COMPUTATION AND OTHER ENERGY SAVING TECHNIQUES IN FUTURE COMPUTATIONAL SYSTEMS

PAUL M. SOLOMON
IBM Research Division
T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

Abstract

The capability of silicon technology has increased and cost of doing computation has decreased to the point where vastly expanded personal computational facilities become available to a large class of users. In this environment the energy cost of computation becomes a critical issue, especially for portable applications, but even in desktop, household and office environments. The capability of the upcoming technologies to deliver performance for some future yielded chip will vastly exceed the power budget allotted to the application desirous of using that performance. The concept of "excess capacity" is introduced to describe this situation and this excess capacity may be traded for power in different ways to realize different system solutions. This paper will discuss the nature of the different kinds of trade-offs, and the classes of system solutions realized.

1. Introduction: Setting the Stage

The semiconductor industry has seen unprecedented growth over the past three decades, forming the basis a great social revolution, akin to the industrial revolution of two hundred and fifty years ago. Since its advent we have seen a million-fold increase in the scale of integration and in the capability of electronic systems, which have become increasingly available, through their reduced cost, for use by the average citizen.

The pattern of development of computing systems, since the invention in the 1970's of the single chip microprocessor, has been a hierarchy of large, fast and expensive mainframe computers (high end) contrasted with small, slow, yet cheap microcomputers (low end). As technology has progressed, the capabilities at both the low and high end have increased both in terms of speed and numbers of circuits. The number of memory circuits has increased much faster than the logic, with slower memory chips serving fast logic chips. The objective, at the low end, has been to fit an entire CPU (central processing unit) on one chip, and this has been achieved by reducing the number of bits being processed at a time, starting at 4b in the early 70's and progressing to 8b, 16b and 32b as the technology advanced, compared to 64b in a

S. Luryi et al. (eds.), Future Trends in Microelectronics, 93–109.

typical mainframe. Now with 64b CPU chips becoming available, we see a convergence between the low and high end in terms of single CPU capability.

As a result of this phenomenal growth, the average user in his office or home now has personal access to computational power that was once reserved for mainframes, and a plethora of applications has developed to exploit this capability. Fortunately the electrical power per unit of computation has also decreased dramatically, permitting these computers to operate at power levels consistent with home or office use. Indeed, the power has been reduced so dramatically that an entirely new class of computers, the portable computer, has arisen. This accelerates the trend toward more computation at ever reduced power levels and leads to the first postulate for the present critique: *The power budget for the most important computer applications will decrease over time.* The assumption here is that home, office, and portable applications will be the most important (in terms of market share), and forces driving the power budget down are primarily increased battery life for portable applications, and energy savings for home and office computers. Indeed, it has been estimated [1] that today, computers consume ≈ 5% of the nation's electrical energy!

Historically, power has not been an important constraint for low-end processor design within the boundaries imposed by complementary metal oxide silicon (CMOS) design practice. Compared to bipolar and n-channel FET technologies, which consume power even when idle, CMOS offered such large power savings for systems running at low clock rates that technology was able to advance, on a fairly conservative scaling path, limited by speed and area constraints, while keeping the power dissipation of the CPU chip fairly small. The only concern related to power was the ability to cool the chip, since most of the power consumption of the desk-top computer lay in the display. Since its introduction, CMOS technology has been able to proceed toward higher densities, without a major collision with power constraints because of a combination of circuit innovations and the application of scaling principles. Indeed, according to the scaling principles laid out by Dennard *et al.* [2] more than two decades ago, FET integrated circuits can be scaled to smaller dimensions, higher densities and higher speeds without increasing power density. This is easily derived from the dependence,

$$P = N_{\mathrm{ckt}} C V^2 f \tag{1}$$

of the power on the number of circuits switching, capacitance per circuit and voltage being switched, and switching frequency, where capacitance, voltage and logic delay (inverse of the maximum switching frequency) scale linearly with dimension. For several generations the voltage had remained constant at 5 V, to maintain compatibility with the old TTL logic interface circuitry, and power had been kept in check by combination of circuit, layout, and architectural innovations.

Further progress in scaling has demanded a return to scaling the power supply voltages, hence the new 3.3 V standard, as well as 2.5 V and 1.8 V supply voltages proposed for the future. At some point voltage scaling becomes harder to do, as we will discuss further, and the voltages will be constrained to be larger than some minimum which depends on various system parameters.

At a constant voltage, and with no further innovation in layout, circuit techniques, and architecture, the power density will increase rapidly with scaling as the inverse square of the dimension. If past trends are followed, chip size will continue to increase and the total power per system would increase dramatically for systems at the cutting edge of speed and integration capabilities of the technology. This leads to the second

postulate of this study: *The power dissipation of a chip designed for maximum speed and integration potential of the current technology will increase dramatically with time.* Now one may argue, as is done in Ref. [3], that chip size need not increase and will actually decrease with time once a full 64b CPU and its cache can be fit onto one chip. This is difficult to dispute, but flies in the face of the economic incentive to exploit the new technology to its fullest extent.

The two postulates of our study, as illustrated in Fig. 1, lead to a dramatic divergence between the power budget for the major applications and the power dissipation of a chip operating at maximum capacity. This trend is also document in recent work by Stork [4]. In order to reconcile the two, the processing capacity of the chip has to be reduced until the power fits within the power budget. We will call the ratio between the maximum capability of the technology and the actual processing capacity at a given power budget the *excess capacity* . Using reasonable trends [4, 5] we find an excess capacity of between one and three orders of magnitude for desk-top and portable computer applications for the 0.1 μm technology generation.

Given this excess capacity, the question becomes how to trade it for lower power consumption in the most efficient manner. This will be the subject of the rest of this paper.

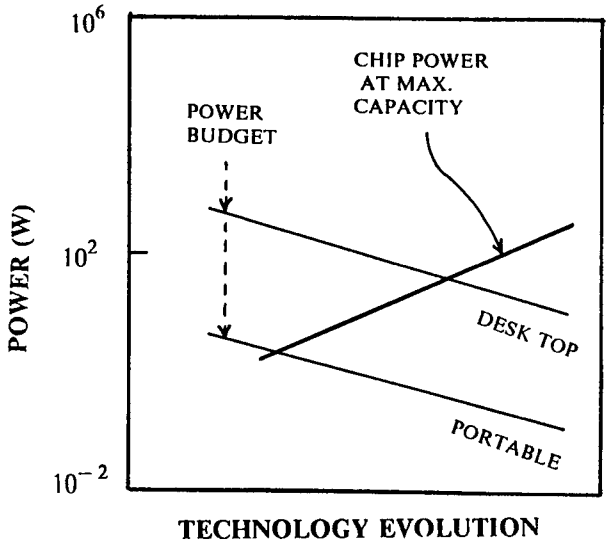


Figure 1. Divergence of chip power dissipation from the power budget for typical applications as a function of technology evolution (time).

2. Adiabatic vs. Switched DC Logic Implementations

Conventional CMOS logic can be considered to be a complex network of voltage controlled switches (transistors) which establish a '0' or '1' logic level by switching a given node to ground or to a DC power supply. Under these circumstances, Eq. 1 is valid without regard to the details of the switch design, with the technology only setting

the capacitance, voltage and switching frequency.

Landauer [6], by contrast, showed that energy dissipation was not fundamental to the computational process — only erasure of information necessitates dissipation. It was not clear until recently [7] how these proofs could be applied to CMOS circuits, although Bennett [8] showed the means by which computation could be made reversible, hence dissipationless, in hypothetical systems. Recent research has investigated practical ways of implementing reversible or quasi-reversible logic circuits using CMOS.

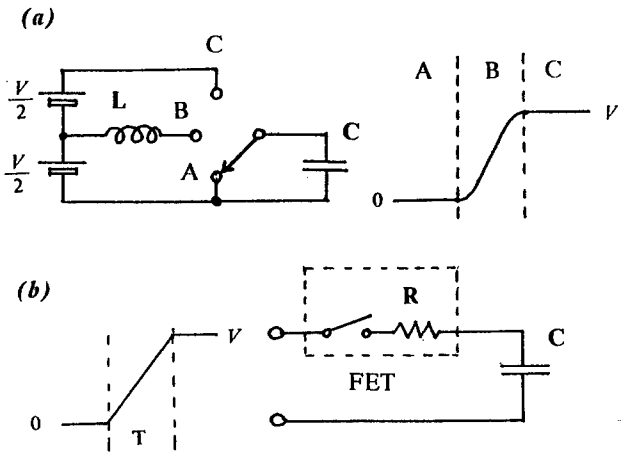


Figure 2. (a) Charging of capacitor C via switched inductor. (b) Charging capacitor via resistive FET.

The circuit of Fig. 2 (a) illustrates how a node may be charged and discharged dissipationlessly, assuming that the circuit elements were lossless, clearly showing that the paradigm represented by Eq. 1 is not universal. In this figure the switch represents an FET, the gate of which is controlled by other adiabatic circuits. A simpler representation is shown in Fig. 2 (b) where the switched FET is represented by a resistor, R, and the capacitor C represents the gate plus wiring capacitance of the circuits being driven. The circuit is fed by a ramped power supply of rise time T. For this example, the energy dissipation per transition,

$$U = CV^2(RC/T) \tag{2}$$

is less by the factor of (RC/T) compared to the conventional case.

A simple example of an adiabatic circuit using conventional CMOS circuitry and following the scheme of Hall [9], is shown in Fig. 3, where the boxes represent CMOS pass gates. For the circuits to operate in an adiabatic manner, the control (gate) voltage on the switching FET's has to be established before voltage is applied to the current (source and drain) terminals. This leads to the concept of the "retractile cascade" where the power supply (clock) voltages are applied at increasing delays as one goes down the logic chain, then withdrawn in reverse sequence.

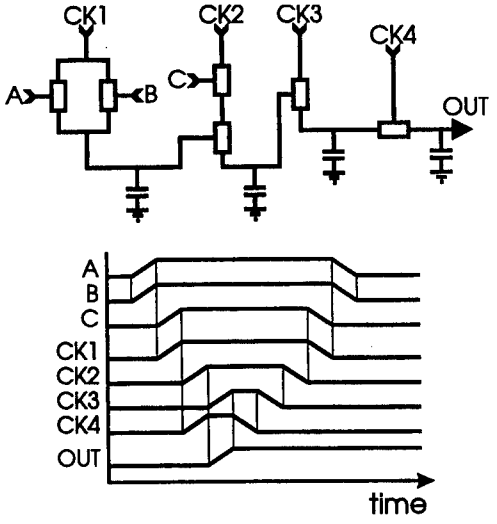


Figure 3. Retractile cascade adiabatic circuit [9].

While this is one of the simplest ways of applying adiabatic concepts to CMOS logic it is by no means the only way [10]. For instance, one can achieve a reversible pipeline by using inverse logic functions to implement a reverse pipeline which adiabatically resets the logic stages in the wake of the forward propagating signals [11].

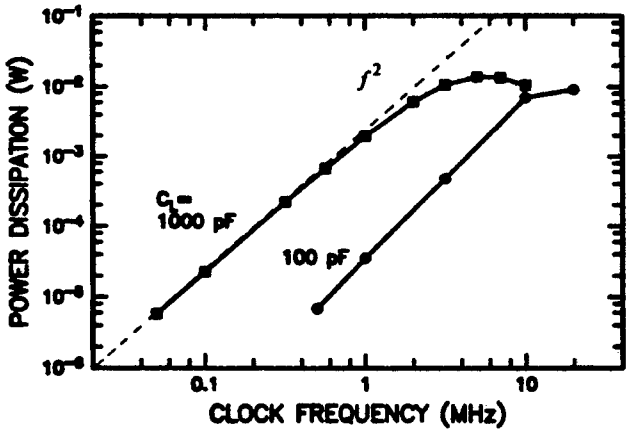


Figure 4. Measurements of on-chip power dissipation of an adiabatic buffer with different off-chip load capacitors.

The property of adiabatic circuits is that for a given V_{DD} the power is proportional to the square of the frequency. This was verified by us (see Fig. 4) in the case of a

simple adiabatic buffer circuit, by measuring the small temperature rise of the chip caused by the dissipated power. This means that at a low enough frequency the adiabatic circuits will dissipate less power than conventional circuits.

Fig. 2 (a) represents a possible on-chip solution for powering adiabatic circuits requiring a low-loss on-chip inductor [12]. Unfortunately such inductors are impractical to make on a VLSI chip. Apart from the considerable problems of incorporating micron sized magnetic components on chip, the Q of a given inductor design is reduced linearly with dimension so that conventional inductor designs would not work well at micron sized dimensions. More practical schemes use capacitor networks [13] but these have limited potential for energy savings.

3. Cost of Information Erasure

Practical adiabatic computers must include latches to store intermediate results. Otherwise the system will suffer too large a temporal penalty, as in the retractile cascade scheme, or too large a penalty in terms of extra circuits (in the reversible pipelined approach). Latch erasure necessitates energy dissipation of at least kT per bit, but this is still about $1x10^8$ less than the energy dissipated in a typical CMOS circuit. Merely to approach this limit would be very desirable.

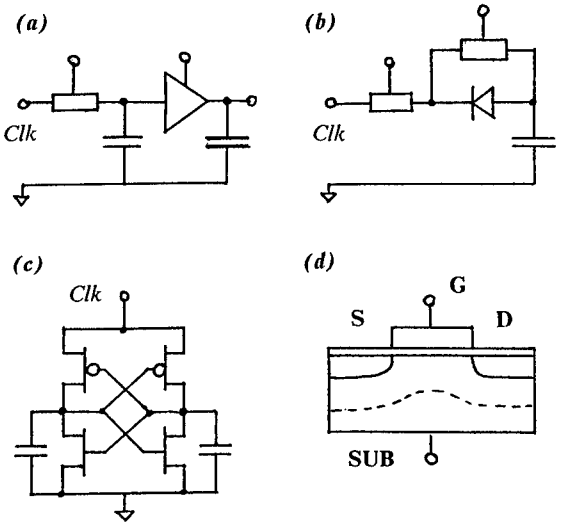


Figure 5. Various methods of information erasure.

Practical CMOS latch circuits, each in succession approaching that limit more closely, are shown in Fig. 5 (a-d). The first achieves efficiency by minimizing the capacitance of the latch and buffering it from the larger load capacitance with an adiabatic buffer. The second employs a series diode, making the cost of energy erasure $\approx CV_{DD}V_{diode}$, so the energy saved is V_{diode}/V_{DD}, where $V_{diode} = (kT/e)\ln(I_{on}/I_{off})$. When the 'on/off' current ratio is optimized for minimum power dissipation, V_{diode} =

$(kT/e)\ln(eV_{DD}/\eta kT)$ is the diode for ward voltage drop, and η ($\eta < 1$) is an activity factor. A further power saving may be achieved [14] using a cross-coupled latch for data storage (Fig. 5 (c)). For reasonably fast erasure the last V_T of voltage on the internal nodes must be erased non-adiabatically leading to a penalty proportional to CV_T^2. For extremely slow erasure the energy cost will ultimately be reduced to several times $(kT/e)^2C$. This same reduced cost can be achieved in a reasonably short time by ramping down the threshold voltage during latch erasure [15], for instance by changing the transistor well bias locally as shown in Fig. 5 (d). We can relate the above result to the more fundamental result of Landauer by expressing it as nkT where n is the number of electrons stored on the capacitor C. Landauer's limit is therefore approached in this kind of latch when the FET's are small enough to store just a single electron!

Because partially adiabatic latches can be made so much more energy-efficient than conventional CMOS, even including a large number of latches into an adiabatic system can result in substantial power savings. As an extreme case, the scheme of Denker *et al.* [16] uses every gate as a latch and realizes a 3:1 power reduction.

4. Practical Adiabatic Systems

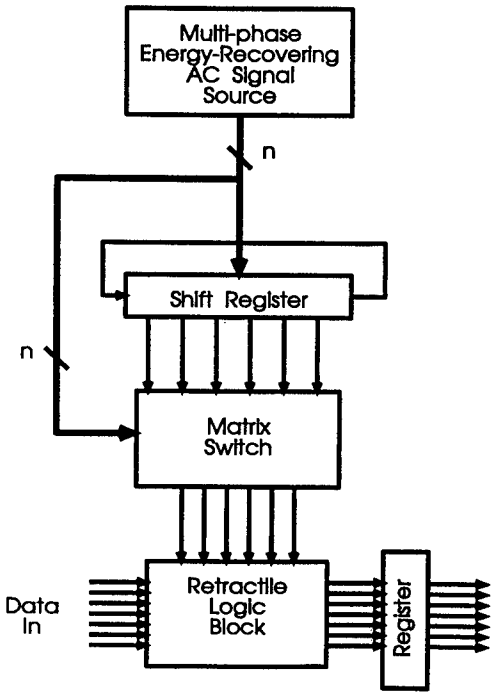


Figure 6. Hypothetical adiabatic logic chip, from Frank and Solomon [18].

A hypothetical adiabatic system is shown in Fig. 6. The system is powered by an efficient sinusoidal AC power supply [17]. The power supply may be on or off-chip and the energy storage in the power supply may be inductors, switched capacitors, or some

other resonator capable of delivering and accepting reactive power. An adiabatically switched network is included on chip [18] to convert the sinusoidal power supply waveform into flat topped waveforms.

The logic is transferred through the logic block and stored on output registers, where it may be erased irreversibly.

5. Voltage Reduction

In expanding the possibilities of digital circuits by including adiabatic circuits we altered the power-frequency trade-off from f to f^2. A much easier way to get a better trade-off is simply to reduce the voltage. This can be done either at a constant V_T or by reducing V_T with V_{DD}. At a constant V_T the voltage range is limited to V_{DD} being greater than about $2V_T$. Below this value the delay increases [19], and the noise tolerance decreases rather sharply.

At a constant V_T/V_{DD} ratio, the transfer curves and relative noise margins of the circuits would not change appreciably over a wide range of V_{DD}, yet the delay will decrease with increasing voltage due to the increasing velocity of the carriers, and eventually saturate at higher voltages. CMOS circuits find themselves at the edge of the velocity saturation regime at the highest voltages, so that reducing V_{DD} at a constant V_T/V_{DD} increases the delay inversely with V_{DD} at the lower voltages. With this approach, reducing V_{DD} and reducing the clock frequency proportionately (keeping the ratio of circuit speed to clock frequency roughly constant) would reduce the power as the cube of the clock frequency. This is a very strong factor and voltage reduction is the method of choice for reducing power.

Since voltage reduction is such an effective means of reducing power dissipation, much thought has been devoted to the question of the minimum voltage for CMOS logic operation [20, 21]. The smallest voltages can be attained when the FET is in its sub-threshold region [21] since voltage nonlinearities are then maximized. For a simple CMOS inverter, the minimum power supply at which it exhibits gain is $2\ln 2(kT/e)$ (36 mV at room temperature). Real logic needs at least a fan in of two, and this requirement raises the minimum voltage to $2.27(kT/e)$ (59 mV) for a CMOS NAND gate [21].

Complex logic functions have been demonstrated on CMOS circuits operating on a power supply of just 0.2 V [22]. That work proved that such low voltage operation is possible, but also that special techniques are needed to adjust threshold voltages dynamically, through controlling the back bias of the p and n-wells of the FET's.

In large scale commercial application of CMOS logic circuits, two factors will combine to increase minimum power supply voltages considerably higher than those quoted above: standby power and tolerances.

The ratio of 'off' to 'on' current (except under special circumstances involving impact ionization) is at least:

$$\frac{\text{off current}}{\text{on current}} \geq \exp\left(\frac{e\Delta V_{GS}}{mkT}\right) \tag{3}$$

where ΔV_{GS} is the gate voltage swing between the 'on' and 'off' states ($\Delta V_{GS} = V_{DD}$ for conventional CMOS) and m is the sub-threshold slope factor ($m \geq 1$). Typically $m \approx 1.4$ for bulk CMOS and approaches unity for silicon on-insulator (SOI).

The total power dissipation consists of both dynamic power ($\overline{C}V_{DD}{}^2f$) and the static power due to the $N_{ckt}\,I_{off}V_{DD}$ 'off' current. The two can be related through Eq. 3 and the condition that the maximum switching time t_{max} be less than the clock period. Here $C_{\max}$ is the capacitance of the slowest circuit, as opposed to the average capacitance $\overline{C}$ used above. The optimum power supply voltage $V_{DD,opt}$ to minimize the total power is then:

$$V_{DD,opt} = \frac{mkT}{e} \times \ln\left[\frac{N_{ckt}C_{\max}F_I F_{I,max}}{N_{on}\overline{C}ft_{\max}}\left(\frac{eV_{DD,opt}}{mkT} - 1\right)\right] \qquad (4)$$

where N_{ckt} is the total number of circuits, N_{sw} is the number of circuits being switched, F_I is the average fan-in, and $F_{I,max}$ is the fan-in associated with the slowest circuit.

Tolerances become extremely important at low supply voltages. Of the tolerances we will consider only the V_T tolerance (assumed equal for p and n-channel FET's for the sake of simplicity). Other tolerances, such as those of the power supplies, are important, but not strategically so, since it is quite feasible to regulate the power supply on-chip, even at multiple points. The simplest way to include the V_T variations is increase the power supply voltage by them:

$$V_{DD,min} = V_{DD,opt} + \Delta V_T{}^+ + \Delta V_T{}^- \qquad (5)$$

The reason for adding both $\Delta V_T{}^+$ and $\Delta V_T{}^-$ is that $\Delta V_T{}^-$ subtracts from V_T, requiring a higher nominal ΔV_T, while $\Delta V_T{}^+$ raises V_T above nominal, requiring an even higher power supply voltage to achieve the speed objectives.

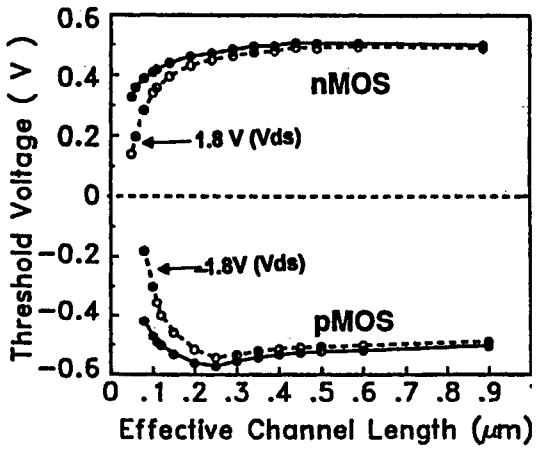


Figure 7. Threshold voltage *vs.* gate length for 0.15 μm CMOS, from Davari *et al.* [5].

Threshold voltage variation is due to many different factors, including doping profiles, work function, well bias and temperature, *etc.*, but we will consider only channel length since this is the most difficult to control. Threshold voltage varies with channel length because of the interpenetration of source and drain fields into the channel region and their screening effect on threshold control implants. Typical V_T roll-off curves are shown in Fig. 7 for a 0.1 μm technology. In a sense, these curves define the technology, both in terms of channel length control and in terms of vertical profile, so that for a given state of the technology there is a clear trade-off between channel length and V_T control. In order to get the tighter V_T control necessary for low voltage operation, the channel length for a given technology would have to be increased.

The parameters in Eq. 4 reflect many possible choices in system architecture, design methodology, and layout. Such parameters are the ratio of switching to total circuits, the ratio of maximum switching time to clock period, and the ratio of average to maximum capacitance. A range of choices is covered in Table 1, the first row representing a large general purpose chip, while the second a specialized function optimized for low voltage. The range of $V_{DD,min}$ (0.2—1 V) represented by these choices is considerable and is an indication that future chips may well depart from the single, standard power supply voltage that has been used up to now.

Portable systems run off fixed voltage batteries of course, and lithium batteries, at ~3 V, are increasingly being used because of their high storage capacity. This voltage must be regulated down to whatever power supply voltage is appropriate. The losses in the regulator will introduce extra factors into the optimization of V_{DD}, resulting in somewhat higher values of $V_{DD,min}$.

TABLE 1. Extreme Scenarios for Determining minimum V_{DD}.

$\frac{mkT}{e}$	$\frac{N_{ckt}}{N_{sw}}$	$\frac{C_{max}}{\overline{C}}$	ft_{max}	$\overline{F_I}$	$F_{I,max}$	ΔV_T^+	ΔV_T^-	$V_{DD,min}$
$\left(\frac{mV}{Dec}\right)$						(mV)	(mV)	(V)
90	10^4	100	0.05	3	4	100	150	1.1
60	10	3	0.5	2	3	25	50	0.28

6. Voltage Scaling For Adiabatic Circuits

At a constant V_T it can be shown [11, 17] that power dissipation is minimized at $V_T/V_{DD} \approx 3.4$. This is when the effect of the resistance increase of R due to reduced gate voltage just balances that of the reduced CV^2.

When the V_T/V_{DD} voltage ratio is held constant, adiabatic circuits yield the same square law dependence (CV_{DD}^2RC/T) of switching energy on voltage as nonadiabatic circuits, assuming that T and R are both inversely proportional to $1/V_{DD}$. To find the minimum power supply voltage for adiabatic circuits we revert to our sub-threshold model and optimize the power supply voltage using similar procedures as before:

$$V_{DD,opt} = \frac{2mkT}{e} \ln\left[\frac{N_{ckt}}{2N_{sw}} \left(\frac{T}{R\overline{C}} \right)^2 \right] \quad (6)$$

where R is effective average channel resistance of the 'on' FET's at a source voltage of half the power supply voltage. This resistance can be adjusted by choosing an appropriate threshold voltage. Now Eq. 6 resembles Eq. 4 the main differences being the factor of 2 outside and the $R\overline{C}/T$ replacing the ft_{max} inside the logarithm. The resulting optimal power supply voltage is therefore larger by approximately a factor of two than in the dissipative case. Tolerances still have to be added, as in Eq. 5. One might argue that adiabatic circuits sense average rather than extreme values because power dissipation, rather than delay, is the issue. This is true in principle, yet in the sub-threshold regime extreme values tend to dominate due to the exponential dependencies.

7. Pass Gates and Static Memory

The V_{DD} limits we have derived so far apply only to simple static logic circuits which are very robust in terms of noise margin. Other circuits generally require higher supply voltages, especially circuits involving pass gates. Such circuits include the ubiquitous static memory cell.

Pass transistors require approximately an extra V_T of supply voltage because two G/S voltage drops are in series. In the sub-threshold regime we can derive the extra voltage by requiring that the delay penalty of the pass transistor and the following logic gate be equal to the delay of the heavily loaded circuit discussed above, resulting in:

$$\Delta V = V_{DD'} - \frac{mkT}{e} \ln\left(\frac{V_{DD} C_{\max}^2}{mkTC_{IN}C_{out}} \right) \quad (7)$$

where $V_{DD''}$ is the value of V_{DD} for the system with no pass transistors. Inserting typical values, $\Delta V_{DD} = V_{DD'}$ - 0.25 V. The penalty is therefore much larger under conditions that require a larger V_{DD} in the first place and approaches doubling of the original V_{DD} .

In a static memory cell, as shown in Fig. 8, the bit line access transistors (n-channel) act as pass transistors; however, the cell is read and written differentially so that at least one of these transistors will always have its source close to ground potential during the entire operation. For low power operation it is feasible to design the cell and sense amplifier such that the access time does not depend critically on the access transistor on the ungrounded side. For the READ this need not depart much from some modern designs [23] in which sensitive sense amplifiers can detect small differential signals. The WRITE operations will require some changes to the standard design practice. While pull-down via conventional sense amplifier-decoder is satisfactory, the pull-up via the pass transistors will be very weak, so that the p-channel transistor in the cell would have to be strong enough to provide this function within a WRITE access time.

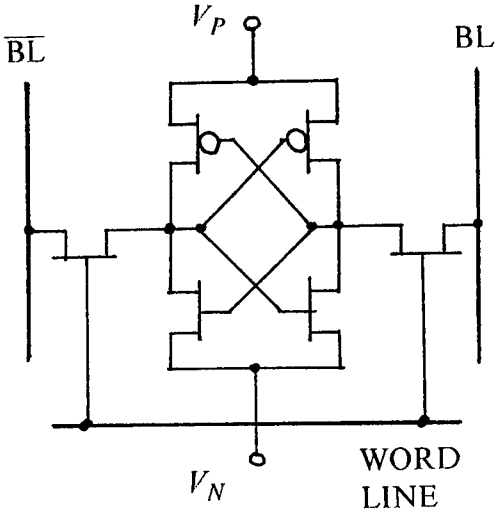


Figure 8. Static memory cell with separate array power supplies.

The above discussion illustrates how the voltage penalty due to the pass transistors in static memory can be circumvented by means of appropriate design technique. Some common practices, such as the use of area-saving very high impedance pull-up devices, could prove unsuitable for low-voltage design.

Pass transistors for LOGIC are not strictly necessary. For instance any combinatorial logic function can be built from simple NAND gates. Therefore the CPU and static RAM could in theory be built without incurring the voltage penalty of pass transistors. The lesson to be learned, however, is that an additional trade-off in low-power (voltage) design is the necessity to abandon certain circuit techniques, which are often used to obtain higher speed at a conventional V_{DD}. This becomes just another part of the speed-power-density trade-off.

8. Sleep Modes

An increasingly used low power design technique is the power down of large blocks of circuitry that are not being used for a particular calculation. This eliminates the DC leakage current associated with that circuitry, but of course requires that the power-down circuitry itself have low leakage current. We see the utility of this technique from Eq. 4, where the threshold voltage could be reduced (to increase I_{on}) and the power supply voltage reduced proportionally to the ratio of N_{sw} to N_{ckt}, where now N_{ckt} is the number of (low V_T) transistors in the *non-sleep* mode, and the transistors in the power-down circuitry have high V_T to reduce leakage.

One can easily see how such a technique may be applied to logic, but for memory the stored information may be lost if the array is powered down. On the other hand, the need for a sleep-mode is especially acute for static memory which may comprise the bulk of the circuitry on a modern CPU chip.

Low standby power static memory requires transistors in the cell which have high threshold voltages; and this conflicts with the goal of low-voltage operation. This dilemma may be resolved if the voltage powering the array be higher, and the internal voltage swings be larger than those of the peripheral circuits. This is possible because the static memory cell has internal gain. Keeping the voltage swing on the peripheral circuits low, including bit lines, is beneficial because these circuits consume most of the dynamic power of the array. In order to complete this picture the voltage swing on the word lines must also be kept low, but this conflicts with leakage requirements as seen from Eq. 3. To resolve this dilemma we must introduce two modes of operation of the array, ACTIVE and SLEEP mode. During active mode the array operates with normal voltage swings on the gates of the access transistors and the leakage current is tolerated. During SLEEP mode this leakage path is shut off by applying an appropriate bias. An alternative, but technologically more difficult, method would be to increase the threshold voltages during the SLEEP mode.

This hierarchy of modes is particularly suited to the already existing hierarchy in the memory organization where a small, fast cache is served by a much larger, slower cache, parts of which may be in a SLEEP mode.

9. Comparison of Trade-Offs

As we have seen, a simple reduction of clock rate does not reduce the energy required for a given computation. As noted by Horowitz [24], this energy is reduced when other system parameters are changed in the course of the optimization toward low power operation. These parameters may be voltage (which we have discussed at length), device size, circuit design, *etc.* Horowitz noticed that the energy time product (Ut) was approximately constant over a range of many parameters. Table 2 summarizes the behavior of FET circuits for some of these variations. Included in this list are adiabatic circuits, where the clock frequency is the variable.

TABLE 2. Constant of scaling *vs.* size of scaling parameter.

	V_{DD}	FET width	adiabatic clk freq
small	U	U	U/t
med	Ut^2	Ut	Ut
large	t	t	-

It is seen that as one varies the parameter of interest that at one extreme (high voltage, large area) the system exhibits little flexibility in speed with respect to changes in power dissipation, while at the other extreme there is little trade-off of energy for speed. At intermediate values, and for adiabatic circuits, there is considerable flexibility in trading speed for energy dissipation, and a region exists which has an approximately constant Ut product. The constant Ut region (also called *action*), while having no

particular theoretical underpinnings, is nevertheless a useful heuristic with which to characterize the effects of design changes.

For the different techniques listed in Table 2 the total energy savings realized in traversing the trade-off region are very different: less than one order of magnitude for the case of size variations, about two orders of magnitude for voltage, and potentially many orders of magnitude for adiabatic computation.

10. Use of Parallelism

In the trade-off region discussed above, energy may be traded for delay, so that by slowing down the computation its total energy cost is reduced. This may also be achieved by partitioning the same computation among several processors, if the problem so lends itself, resulting in the same or even greater computational throughput but at reduced energy cost.

At a constant *Ut* product, and a desired total throughput, the total energy required for a given computation will be inversely proportional to the number of processors in the system. As discussed in the introduction, this choice will become increasingly available for future technologies where many complete processors could fit onto a single chip. As an example, let us consider the technology choices outlined by Davari *et al.* [5]. At the 0.1 μm level this technology will be able to support a RISC processor on an area of about 18 mm^2, so that a 2 cm^2 chip would support about 10 such processors. For this example, parallelism reduces the chip power by a factor of 10 while maintaining the same throughput.

With this approach, the granularity of the larger system on a chip will be that of a single CPU and its associated cache. Other large CPU-sized special purpose functional blocks will be included in this mix. An advantage of this partitioning, apart from the benefits discussed above, is that for work loads not requiring all of the CPU's, the unused CPU's can be put into the SLEEP mode with all of the attendant power savings.

11. Technology Evolution

In the introduction we presented the thesis that the driving force behind the quest for low power was the advance of technology. A technology road map into the next century is presented, for instance, by Davari *et al.* [5]. In this section we will consider these options in more detail.

The primary driving force of the technology is the increase in density which, as we have discussed previously, will allow for multiple processors and large static RAM's on chip. To achieve the goals for low power, the voltages will be driven down toward their lower limit and SLEEP mode will be used extensively. DC-DC converters will find common usage on-chip to regulate the battery or external power supply voltage down to the chip power supplies.

Device scaling will be pushed to the limit with MOSFET's having effective channel lengths of less than 0.1 μm. At such short channel lengths the carrier velocities in the channel would be saturated at voltages low as 0.5 V, so that reducing voltages toward their minimum would not necessitate a severe delay penalty. In a constant voltage scaling scenario, adiabatic circuits fare better than conventional circuits because the channel voltage is always small and velocities do not saturate. Anything done to

improve the channel mobility, such as the inclusion of a germanium alloy, will therefore help adiabatic more than non-adiabatic circuits.

The crucial issue for scaling and for voltage reduction is V_T. Further progress on the latter front is very difficult to achieve due to the technological difficulty of controlling the dopant profile, but there are no fundamental limitations to V_T control down to dimensions of at least 0.05 μm effective channel length [25].

12. Wire Resistance

Wire resistance will be an important limiting factor to achieving very high performance in future high speed circuits. This trend will reinforce the move toward parallelism, since the slower individual processors will be much less limited by the wiring delay.

The scaling potential of adiabatic chips is limited by wire resistance. The power dissipated in the wires, P_w is proportional to $f^{5/2}/\lambda^2$ for a constant-sized chip [17], where λ is the scaling parameter. This relationship reverses the otherwise favorable trend for adiabatic circuits compared to conventional circuits.

As an example we estimate the magnitude of this effect using the same scaled version of Davari's processor [5] as before, where the capacitance C of the processor is derived from his data on power dissipation to be 5 nF. Assuming an AC power distribution bus of 10 μm thickness (the skin depth at 130 MHz) fed in from the side, the power dissipated in the wires is 80 mW for the single processor. If a larger chip were used and 16 processors were to operate simultaneously at 1/16 the frequency (and with a 40 μm thick power distribution bus), the power dissipation in the bus would be reduced to 1 mW. Note that the thick power bus could be situated off chip and the AC voltage transmitted to the chip via an aerial array.

The above example illustrates that the wire resistance limit to the scaling of adiabatic circuits is several generations beyond today's technology, but will be encountered before the scaling of FET technology has run its course.

13. Conclusions

We have compared different approaches to obtaining low power in future computational systems. The easiest and most potent technique is to reduce the supply voltage, the implementation of which requires the use of both high and low threshold voltage FET's, the extensive use of the SLEEP mode, and different power supply voltages for logic and memory. Using a combination of these techniques it is feasible to use power supply voltages as low as 0.5 V for a future technology capable of supporting large, multiple CPU chips. The attainment of this voltage depends critically on attainment of sufficient threshold voltage control and may necessitate longer channel lengths than would otherwise be the case for a given state of technology evolution.

The power levels for the scaled technology are low enough that adiabatic techniques will not be needed in the CPU except in specialized applications. These techniques could, however, be very useful in the peripheral, input-output areas (display drivers, cameras, memory busses, *etc.*) where data rates are not so high and substantial energy is expended driving large capacitances.

13. References

1. Lemnios, Z. and Gabriel, K. (1993) *ARPA Low Power Electronics Presentation.*
2. Dennard, R.H., Gaensslen, F.H., Yu, H.N., Rideout, V.L., Bassous, E., and LeBlanc, A.R. (1974) *IEEE J. Solid-State Circuits* **9**, 256.
3. Sai-Halasz, G.A. (1995) *Proc. IEEE* **83**, 20.
4. Stork, J.M.C. (1995) *Proc. IEEE* **83**, 607.
5. Davari, B., Dennard, R.H., and Shahidi, G.G. (1995) *Proc. IEEE* **83**, 607.
6. Landauer, R. (1961) *IBM J. Res. Develop.* **5**, 183.
7. See articles in *Proc. PhysComp'92 Workshop on Physics of Computation* (Dallas, Texas, Oct. 1992).
8. Bennett, C.H. (1988) *IBM J. Res. Develop.* **32**, 16.
9. Hall, J. S. (1992) Proc. of the 4th International Conf. on Computing and Information, ICCI'92.
10. See articles in *Proc. ACM-SIGDA* and *IEEE-CAS 1994 Int'l Workshop on Low Power Design* (Napa, CA).
11. Younis, S.G. and Knight, Jr., T.F. (1994) Proc. of 1994 Int'l Workshop on Low Power Design (Napa, CA), p. 177.
12. Athas, W.C., Svensson, L.J., Koller, J.G., Tzartzanis, N., and Chou, E. (1994) Proc. of 1994 Int'l Workshop on Low Power Design (Napa, CA), p. 189; Athas, W.C., *et al.* (1994) *IEEE Trans. VLSI Systems* **2**, 398.
13. Svensson, L.J. and J.G. Koller, in *Proc. of 1994 Int'l Workshop on Low Power Design* (Napa, CA), p. 159.
14. Koller, J.G. and Athas, W.C. (1992) in *Proc. PhysComp'92 Workshop on Physics of Computation* (Dallas, TX).
15. Solomon, P.M. and Frank, D.J., unpublished
16. Denker, J.S., Avery, S.C., Dickenson, A.G. Kramer, A., and Wik, T.R. (1994) in *Proc. of 1994 Int'l Workshop on Low Power Design* (Napa, CA), p. 183.
17. Solomon, P.M. and Frank, D.J. (1994) in *Proc. of 1994 Int'l Workshop on Low Power Design* (Napa, CA), p. 93.
18. Frank, D.J. and Solomon, P.M. (1995) in *Proc. of 1995 Int'l Symp. on Low Power Design* (Laguna, CA).
19. Taur, Y., Mii, Y-J., Frank, D.J., Wong, H-S., Buchanan, D.A., Wind, S.J., Rishton, S.A., Sai-Halasz, G.A., and Nowak, E.J. (1995) *IBM J. Res. Develop.* **39**, 245.
20. Swanson, R.M. and Meindl, J.D. (1972) *IEEE J. Solid-State Circuits* **7**, 146; Liu, D. and Svensson, C. (1993) *IEEE J. Solid-State Circuits* **28**, 10.

21. Frank, D.J., unpublished.

22. Burr, J.B. and Shott, John (1994) in *Digest of Tech. Papers ISSCC94 IEEE Int'l Solid-State Circuits Conf.*, p. 84.

23. Itoh, K., Sasaki, K., and Nakagome, Y. (1995) *Proc. IEEE* **83**, 524.

24. Horowitz, M., Indermaur, T., and Gonzalez, R. (1994) in *IEEE 1994 Symp. on Low Power Electronics* (San-Diego, CA) p. 8.

25. Frank, D.J., Laux, S.E., and Fischetti, M.V. (1992) in *IEDM Digest of Tech. Papers*, p. 553.

ARCHITECTURAL FRONTIERS ENABLED BY HIGH CONNECTIVITY PACKAGING

STEVE NELSON
Steve Nelson and Associates
6706 N. Lakeshore Drive
Chippewa Falls, WI 54529

1. Introduction

Absolute performance and price/performance of computing systems have advanced at an extremely rapid rate since the early 1950's. The rapid decrease in the cost per operation has caused the computer industry's designers to begin thinking of "computers as commodities". High volume manufacturing and mass markets have even challenged mission-oriented government agencies to consider commercial off-the-shelf technology as the best path to ever higher performance at a reasonable cost. The reduction of printable line widths for integrated circuits has been primarily responsible for this trend. While the projected improvements in line width continue, there has been a decline in the rate of minimum feature improvement. Short of a revolutionary breakthrough, it is generally believed that only modest gains (relative to the historical trends) will come with higher gate and clock speeds alone. Rather, higher density active devices coupled with scalable architectures and three-dimensional packaging technologies will provide the new high performance solutions.

Traditional "high-multiple-device" CPU implementations, such as the classic vector supercomputers from Cray Research, are a passing reality. Even the highest performing versions of these processors will fit on one, or perhaps a very small handful, of integrated circuits. Clock frequencies, while increasing, will not provide dramatic increases in performance. There is, rather, new motivation to effectively connect large numbers of parallel CPU resources and to tightly couple coherent cache memories into the architecture. New I/O requirements will drive toward two-dimensional arrays of I/O pads on integrated circuits that can only be connected with technology that closely couples substrates to the circuits.

What other architectural dimensions could be enabled by a cost-effective three-dimensional interconnect approach? Higher levels of parallelism are possible within a traditional CPU design by adding large numbers of vector pipes and associated functional units. The standard interface design for dynamic random access memories is being seriously challenged by new designs which provide both architectural and physical improvements to dramatically increase device bandwidth. More challenging is the exploration of radically different ways of building processors. There is now renewed interest in *very long instrucrion word* (VLIW) CPU's. Another promising architecture which has languished due to the difficulty of cost-effectively interconnecting massively large numbers of active devices is the *data flow machine.* The time may now be right to

S. Luryi et al. (eds.), Future Trends in Microelectronics, 111–123.

assemble the technology to build the first practical versions of these highly concurrent data processing architectures.

2. Integrated Circuit Trends as a Packaging Driver

In 1993 the Semiconductor Industry Association (SIA) projected the future capabilities of commodity integrated circuit devices into the next century. Table 1 charts line width reductions of a factor of about seven during a sixteen year period.

TABLE 1. Semiconductor Industry Association roadmap (1993).

		1992	1995	1998	2001	2004	2007
Feature size (μm)		0.5	0.35	0.25	0.18	0.12	0.1
Gates / chip		300K	800K	2M	5M	10M	20M
Bits / Chip	- DRAM	16M	64M	256M	1G	4G	16G
	- SRAM	4M	16M	64M	256M	1G	4G
Wafer processing cost ($ / cm^2)		$4.00	$3.90	$3.80	$3.70	$3.60	$3.50
Chip size (mm^2) -	- logic / microproc	250	400	600	800	1000	1250
	- DRAM	132	200	320	500	700	1000
Wafer diameter (mm)		200	200	200-400	200-400	200-400	200-400
Defect density (defects / cm^2)		0.1	0.05	0.03	0.01	0.004	0.002
No. of interconnect levels - logic		3	4 - 5	5	5-6	6	6-7
Maximum power (W / die) - high perf.		10	15	30	40	40-120	40-200
	- portable	3	4	4	4	4	4
Power supply voltage (V) - desktop		5	3.3	2.2	2.2	1.5	1.5
	- portable	3.3	2.2	2.2	1.5	1.5	1.5
Number of I/Os		500	750	1500	2000	3500	5000
Performance (MHz)	- off chip	60	100	175	250	350	500
	- on chip	120	200	350	500	700	1000

The message here is that traditional "high-multiple-device" CPU's are a passing reality. However, integration of a single chip solution with multiple levels of cache is not always possible either. The multichip implementation of the next generation microprocessor from Intel (the "P6") is a good example.

Consider the following six data points from the SIA numbers:

	1995	2001	2007
Chip size (logic, mm^2)	400	800	1250
Number of I/O's	750	2000	5000

If the area is to double (from 1995 to 2001), the IC perimeter will only increase by a factor of about 1.4. Yet the number of projected I/O pads is projected to increase by a factor of 2.66. I/O requirements will drive two dimensional arrays of I/O pads on IC's which can only be connected with technology that closely couples the substrates to the

IC's. This has the effect of more tightly coupling the substrate wiring to the device wiring. In fact, pad-to-pad connections on the same IC will even be sometimes routed the board-level substrate to provide configurability, to solve a tough on-chip routing problem, or even to reduce propagation delay. The packaging technology which allows this sandwich of silicon and substrate can be the base of an even more aggressive three dimensional packaging system. The motivation to connect large numbers of integrated circuits to build a traditional high performing single CPU (as used in the CRAY-1) will be history. This will, however, be replaced by new motivation to effectively connect large numbers of parallel CPU resources and to tightly couple coherent cache memories into the architecture.

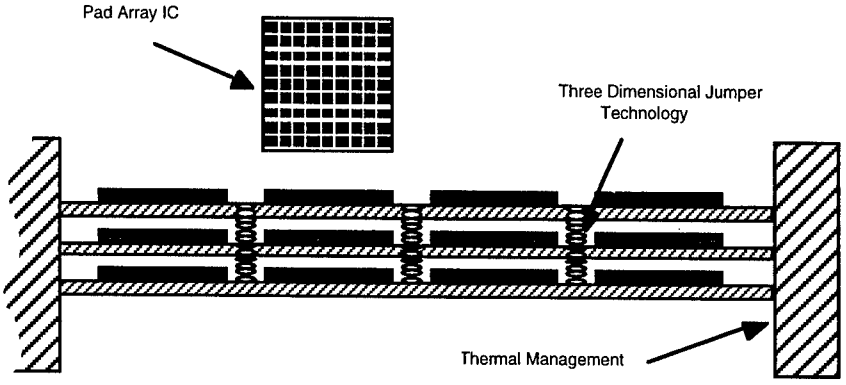


Figure 1. Three dimensional packaging with area array IC pads.

3. Improving Performance of Cache Memories with 3D Packaging

Cache is introduced into system architectures to reduce latencies to main memory. But adding additional levels of hierarchy to the memory subsystem increases system complexity dramatically. Virtual memory address translation and multiprocessor configurations provide additional challenges which must be met. It is fair to say that no system has been built to date that can provide a simple cache coherency model, high performance, and scalability to very large numbers of processors. At least one of these objectives is invariably compromised.

3.1 CACHE ASSOCIATIVITY

Associativity describes the degree of flexibility the system has in placing cache blocks (lines) from main memory into the cache. Direct-mapped is the easiest to implement but provides for only one "slot" where a main memory block can be moved to the cache. This generates a higher possibility of conflict among memory blocks and can cause a debilitating characteristic called "thrashing," an effect where blocks move inefficiently back and forth between cache and main memory. At the other extreme, a fully associative organization provides a cache with slots that can receive any memory block. While the performance of a fully associative cache is almost always superior, it

is typically prohibitive in cost. This cost is primarily borne in the address translation and cache tag identification circuitry. Quickly locating a cache block requires a highly parallel translation and a search usually using content-addressable memories (CAM). Large scale content-addressable memories are very expensive to build due to the extra active devices required for each memory cell and added wiring complexity.

The n-way set associative cache is a compromise solution, where n is the number of cache block locations that any main memory block copy can occupy. This greatly restricts the size of the CAM. A classic study of cache miss rates (the frequency of CPU references which do not find the data already in cache) versus cache size and associativity was made by Hill. This study showed that as the degree of associativity increases, the miss rate decreases. In addition, for large caches with higher levels of associativity, improvements in the miss rate become vanishingly small. Looking at such cache miss data, a computer architecture designer would probably conclude that eight-way associativity is enough. However, the cache miss table data is badly skewed away from typical numerical workloads which exhibit a very different address pattern than the mixed workloads used in the study. Due to underlying array data structures these codes generate frequent non-unit stride memory address reference patterns as shown in the two examples of Fig. 2.

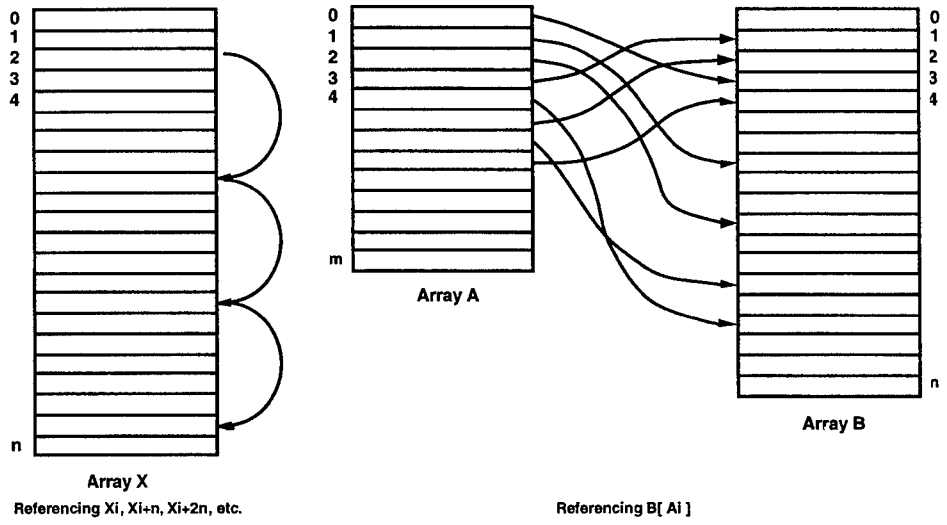


Figure 2. Difficult array reference patterns for cache memory systems.

Non-unit stride single-word references have the effect of greatly increasing the relative cost of a large block (line) cache load. If only 8 bytes are required but an entire line, of perhaps 64 bytes, is moved the memory port is operating at only 1/8 efficiency for this pattern. This is the primary reason why vector architectures have traditionally eschewed cache memories.

How might three dimensional packaging address this problem? Numerical codes on scalar architectures would do much better if the cache were organized with very small line sizes (probably only one word — 8 bytes — in length). Additionally, the cache

should be very large. If it cannot be made fully associative, it should at least be "many-way" set associative to avoid serious contention and excessive miss rates. Such a cache subsystem will contain a very high degree of wiring complexity to adequately drive the many sets and to implement the large content addressable memory.

3.2 CACHE COHERENCE IN MULTIPROCESSOR SHARED-MEMORY SYSTEMS

A shared-memory multiprocessor system with cache memory on each processor must solve the consistency problem for shared variables. Coherency is at risk when one processor updates a shared variable that has previously been copied into another processor's cache.

"Bus snooping" is the most common way to handle coherency in a moderately parallel multiprocessor system. Each processor independently monitors the shared memory bus and invalidates any cached copies of main memory variables that are modified by other processors. Bus snooping does not scale well beyond roughly a dozen processors due to loading effects on the shared bus wires. Buffering can help, but this adds latency. Three dimensional interconnect would reduce parasitics on the bus so that bus snooping could be extended by a factor of perhaps two or three.

While not insignificant, this technique, by itself, is probably not the way to build a highly parallel implementation. More promising approaches for the support of large systems are data migration and address remapping techniques such as the ALLCACHE™ technology implemented by Kendall Square. Migrating data to the point of use has a lot of promise for reducing latency in multiprocessor systems. Unfortunately the ring structure of the KSR systems add significant latency of its own and much of the potential benefit is lost. A higher efficiency topology, such as a tree structure or an n-dimensional torus could significantly improve the situation, especially if wide data paths were employed. Wide data paths with a three-dimensional topology (or higher) could make the difference.

4. Improving RAM Performance in a 3D Environment

Whereas the arguments for cache subsystem improvement concentrate on reducing latency, the argument for improving RAM technology is related to bandwidth. While memory chip densities have grown dramatically over the years, the story of memory device performance improvements is far less compelling (Fig. 3).

Clearly the performance challenge is being lost by conventional DRAM design approaches. There are several ways to improve this situation:

i) Dramatically increase the width of the DRAM channel.

ii) Increase the speed of the signals for the DRAM data I/O.

iii) Internally buffer the DRAM cells with internal cache and speed up the I/O.

Memory in high performance systems has one other important characteristic — there is usually a lot of it. Managing very wide word DRAM I/O and/or very high speed DRAM I/O when many devices are involved is difficult without dense packaging technology. Wide I/O means many more "wires" between the memory chips, data

buffers, and address drivers. Very high speed transmission lines require low parasitics (throw the package away!), as short as possible wire lengths, and termination resistors or active termination devices. (It would be best to fully integrate the termination resistors within the three dimensional packaging environment.)

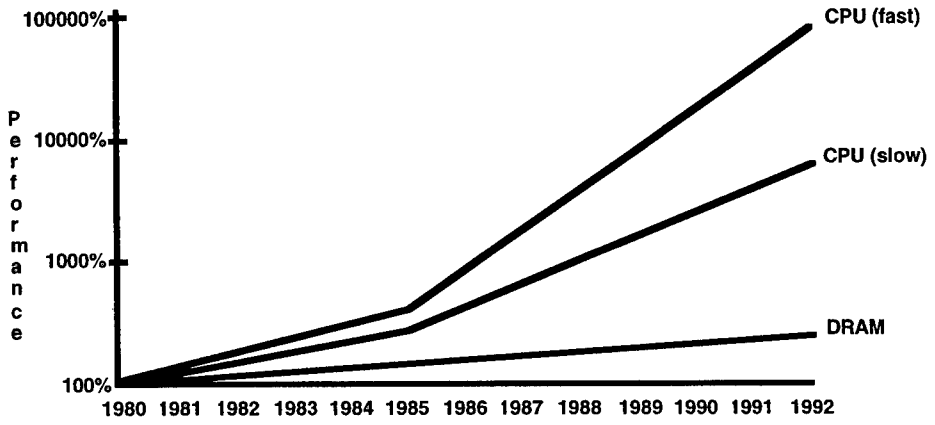


Figure 3. Historical DRAM and CPU performance (from Patterson, D. and Hennessy, J. (1990)*Computer Architecture: a Quantitative Approach* , Morgan Kaufman, San Mateo).

Leading production DRAMs:

1 bit in 60 nsec => 16.7 Mbits / sec
4 bits in 60 nsec => 66.7 Mbits / sec

For 16 Meg DRAM, 4096 bits could theoretically be made available by the row select at approximately one half the access time. The "not to exceed" number becomes:

4096 bits in 30 nsec => 137 Gbits / sec

More practical solution:
Feed 128 bits or 256 bits to secondary cache and allow 5 nsec for selecting:

128 bits in 35 nsec => 3.66 Gbits / sec

256 bits in 35 nsec => 7.31 Gbits / sec

Factors of over a hundred improvement in DRAM bandwidth are conservatively available with an area interconnect / 3D packaging approach.

Reference point:

Leading production SRAMs

18 bits in 9 nsec => 2 Gbits / sec
36 bits in 36 nsec => 4 Gbits / sec

(Motorola BurstRAMS)

Figure 4. Improving RAM bandwidth with area interconnect.

4.1 BROADSIDE RAM ACCESS

The internal aggregate data rate of the memory array sense amps is much higher than that transmitted through typical l-bit or 4-bit wide DRAM packages. Figure 4 demonstrates how much room there is for improvement if this resource could be tapped. But bringing 128 or 256 bits off of the DRAM would easily overwhelm a traditional packaging environment. Only area die interconnect and three dimensional stacking of devices are likely to succeed.

4.2 HIGHER SPEED DRAM I/O

Rambus is a company that has been successfully licensing a 500 MByte/sec DRAM I/O technology. The Rambus DRAM channel requires high-speed drivers and receivers both within the DRAM's and the memory subsystem ASIC's or the microprocessor.

The Rambus technical guide states, "The physical length of one Rambus Channel is presently limited (to approximately 10 cm) by the 2 nanosecond propagation time of signals from one end to the other. This length can accommodate up to 32 RDRAMs or up to ten memory RModules, or combinations of the two. Ten memory RModules hold up to 320 DRAM'S, giving a total of 160 MBytes of memory capacity using 4 Mbit RDRAMs." (16 Mbit devices will be offered soon to allow 640 MBytes of storage.)

This total package is roughly 16 in^2 (4x4x1 inch) and contains about 30% silicon. It would appear that a more aggressive stacking technology (but with improved thermal management support) would allow an increase in bit density of about three and still obey the 10 cm rule. The other possibility is a potential increase in the clock rate with a smaller configuration by using improved packaging technology.

4.3 MERGED CACHE/RAM

Ramtron has introduced a merged device that provides cache buffering of slower but more dense RAM cells. However the actual sizes of the individual cache and RAM components are quite modest. OKI Electric has just reported on a 32-Bank 256 Mbit DRAM with cache and tag. A three dimensional packaging technology combining many broadside-I/O DRAM'S and SRAM's acting as cache could extend the goals of the Ramtron and OKI devices for single processor systems. Then, adding coherency support between multiple modules would add significant performance to symmetrical multi-processor nodes. Again this would require high wiring density and close spacing.

5. Multiple-Pipe Vector Opportunities using 3D Packaging

One way to improve parallelism in a vector processor is to add multiple vector pipes. The CRAY C90 uses dual pipes as shown in Fig. 5. Simultaneously, two operands pairs are extracted from the vector registers and routed through dual functional units, such as the multiply units in the figure. More pipes, more parallelism. (It is important to realize that the benefits for short vectors will be less pronounced, due to vector start-up time.) But for large matrices and long vectors multiple pipes can greatly improve performance. The price to pay, however, will be a large increase in wiring density. This is one example of expanding a state-of-the-art high-performance-processor in a way that necessarily requires multiple integrated circuits. Very high density wiring to and from

the memory subsystem will be needed to adequately feed the large number of vector operations.

For an eight- or sixteen-pipe design the signal count becomes very high indeed. And every data path is latency sensitive so each must be as short as possible. Only a three dimensional technology could meet all of these requirements.

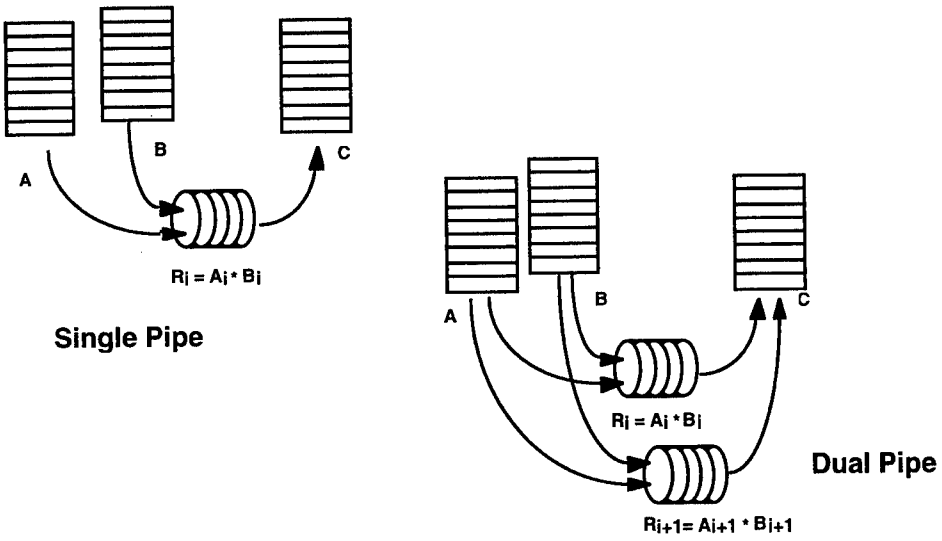


Figure 5. Increased parallelism with multiple pipe vectors.

Vector processors have the reputation of being high-cost solutions. However, looking forward, there is no fundamental reason why the CPU silicon has to be very much more expensive than a high performance microprocessor, the difference being solely in the manufacturing volume. And if anything, the utilization of that silicon is higher. The memory devices can be similar in cost, as recent low-end DRAM vector systems have shown. The traditional cost factor which needs to be improved is that associated with packaging.

6. A Fine-Grained MIMD System

Advances in integrated circuit density as applied to DRAM technology will very soon provide a watershed opportunity for large system design. This will occur at about the time DRAM densities achieve 256 Mbit integration levels. Then a single chip node will be possible which contains 8 or 16 MBytes of storage, sufficient cache memory, a complete 64-bit microprocessor with multiprocessor synchronization support, and a three dimensional router (Fig. 6). Due to the design point of the DRAM and power considerations, this will probably not be the fastest processor of its generation, but for certain applications performance can be made up by using very large numbers of processors. Consider a large system with more than 10,000 64-bit processing elements.

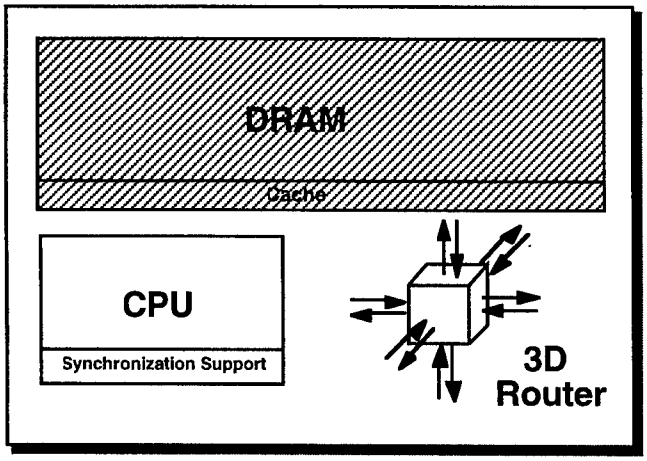


Figure 6. Single-chip fine-grained MIMD node.

Using only modestly aggressive I/O driver circuitry, very respectable I/O rates over each router path are possible: 100 MBytes/sec on each of the twelve paths would be a reasonable starting design point. The three-dimensional router would allow a folded torus topology.

This FG-MIMD system would introduce some important tradeoffs that affect the selection of application suites which would run efficiently:

i) Because the processors would not be the fastest of the generation, a larger number would be required to achieve a given peak performance level. This means that communication latencies would be longer.

ii) In a single chip per node implementation the amount of memory per processor would be modest.

There is no reason why a processor node could not be augmented by additional memory devices and such a trade-off should be studied in detail. A homogeneous system with essentially only one part type seemed compelling for early consideration, however.

7. Data Flow Architectures

Jack Dennis from MIT described the data flow concept in a 1979 paper at the First International Conference on Distributed Computing Systems entitled "The varieties of data flow computers". He said: "Fundamentally, the data flow concept is a different way of looking at instruction execution in machine level programs -- an alternative to the von Neumann idea of sequential instruction execution. In a data flow computer, an instruction is ready for execution when its operands have arrived -- there is no concept of 'control flow', and data flow computers do not have program location counters. A consequence of data-activated instruction execution is that many instructions of a data

flow program may be available for execution at once. Thus highly concurrent computation is a natural accompaniment of the data flow idea".

Figure 7 graphically shows the machine code for a simple arithmetic instruction. With many values being created and presented as input tokens that "fire" when the data arrives, there is a huge opportunity for concurrency and, therefore, high performance.

AVERAGE: **real** := 0.87 * (A + B + C + D).

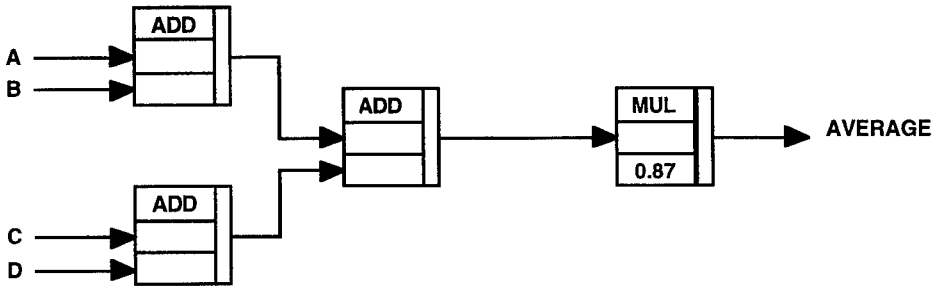


Figure 7. Data flow machine code for an arithmetic expression.

The original data flow model proposed by Jack Dennis at MIT is now known as the static data flow model. An alternative, the dynamic data flow model was proposed by Arvind, a collaborator with Dennis in early data flow research. The detection of matching tokens in data flow computer requires very large amounts of content addressable memory. In addition there is a problem of allocating resources when a code-block is mapped to a processor mapping unit. If this resource becomes overcommitted, the processor can deadlock. As already described, content-addressable memory requires more wiring than directly addressable memory. A planar array of silicon devices embedded within a three dimensional wiring array to transfer and buffer the tokens to be matched would be an important enabler for the data flow architecture, which has never delivered on its promise because of serious implementation obstacles.

8. Very Long Instruction Word (VLIW) Architectures

VLIW is an alternative to superscalar architectures which assigns responsibility for multiple issues of instructions to the compiler. Superscalar hardware has difficulty with efficiently issuing more than three or four instructions at once because data dependencies are difficult to resolve. Conceptually the number of instructions components that could be issued in a VLIW system is just a function of how much the dependency analysis will allow.

While VLIW architectures can provide high levels of concurrency with appropriate compilation, they can cause unwieldy code expansion. Whenever a concurrent operation cannot be scheduled, a no-op instruction must be substituted. This strains the cache resource needs which are already expanded due to the added length of each instruction. If multiple VLIW devices are to be networked into a multiprocessor with an effective cache coherency policy, the interconnect requirements become very high.

The wiring paths to be optimized involving VLIW cache are, of course, extremely latency-sensitive. A compact three-dimensional packaging approach would reduce wiring lengths and allow increased fan-out for synchronization and coherency signaling.

Example Instruction Word Format

integer	integer	floating point	floating point	memory ref	memory ref	branch

"A VLIW instruction might include two integer operations, two floating-point operations, two memory reference, and a branch. An instruction would have a set of fields for each functional unit--perhaps 16 to 24 bits per unit, yielding an instruction length of between 112 and 168 bits."

Figure 8. Very Long Instruction Word (VLIW) architectures (from Patterson, D. and Hennessy, J. (1990) *Computer Architecture: a Quantitative Approach*, Morgan Kaufman, San Mateo).

9. Chip-to-Chip Latency Reduction in a Superconducting Supercomputer

Building a competitive supercomputer with superconducting integrated circuits in the mid-1990's would absolutely require very close three-dimensional connection of multiple integrated circuits. The level of integration available for each circuit would be at about the level used by the CRAY Y-MP team (circa 1985,86) -- at best!

Could 3 dimensional packaging make up for lagging levels of integration?

Most recent reported results:

"A Subnanosecond Clock Josephson 4-bit Processor," Kotani, Imamura, Hasuo, 1990

"An 8-bit Josephson Digital Signal Processor," Kotani, Inoue, Imamura, Hasuo, 1990

"A 2-kbit Superconduction Memory Chip," Yuh, 1993

"The Design for a Josephson Micro-Pipelined Processor," Harada, Hioe, Takagi, Kawabe, 1994

If full 64-bit supercomputers were to be built on this technological base any time soon the design would necessarily consist of multiple devices for each CPU. Extremely dense packaging using 3 dimensional approaches would be mandatory to reduce the latency of intraprocessor communication paths.

Free space propagation delay: 3.4 ps per millimeter

Die size for 4-bit processor above: 2.2 mm X 2.5 mm
Gate delay in 4-bit processor: 8.7 ps
Propagation delay across chip: 8 ps / millimeter

Propagation in medium with dielectric constant of 3: 5.87 ps per millimeter

Figure 9. Superconducting technology.

Poor yields and high support costs for a super-cooled system have slowed advances (and investment) in this technology to a snail's pace. But the technology continues to set a high water mark for performance, with typical gate delays on a 4 bit-sliced processor design of 8.7 ps. Also, the design for a 12-bit micro-pipelined Josephson processor with ALU, multiplier and 16 registers has been reported, but actual circuits have not yet apparently been built.

A rough rule of thumb for a multi-chip CPU is that the average wiring latency between devices should not exceed the total on-circuit gate and wiring delay within each clock period. Using an estimate of about eight gates per clock at 8.7 ps per level, the wiring delay should not exceed 70 ps. In a transmission medium with a dielectric constant of 3 (having a propagation delay of 5.87 ps/mm), this would translate into 12 mm of wiring delay (Fig. 9). While this doesn't represent a hard and fast criterion, if the interconnect structure provides much slower transit times one has to question the inefficient use of the fast circuits.

10. Conclusions

This paper describes several architectural concepts which could be dramatically advanced by three-dimensional packaging. The most intriguing of these is probably the exploitation of data flow architectures by building large content-addressable memories for token matching. Fine-grained MIMD systems, while not as general purpose as one would like given the relative immaturity of today's software, will probably be the most cost effective way to achieve tera-ops of performance for those algorithms which do fit. Cache has been the main ingredient to drive microprocessor scalar performance up to, and even beyond, traditional supercomputers. But the next performance leap will surely have to come from multiprocessor systems of these devices. Cache coherency that scales well will certainly be on every designers "to do" list. Very low latency must be offered by many tightly coupled devices. Area array pads and three-dimensional interconnect would greatly increase DRAM bandwidth. Any very high speed device technology which does not support VLSI and ULSI densities will only deliver system-wide high performance via a very dense interconnect scheme.

11. References

Dally, W.J. and Seitz, C.L. (1987) Deadlock-free message routing in multiprocessor interconnection networks, *IEEE Trans. Computers* **36**, 547-553.

Kessler, R.E. and Schwarzmeier, J.L. (1993) CRAY T3D: a new dimension for Cray Research, in *Proc. COMPCON 1993*, pp. 176-182.

Hill, M.D. and Smith, A. (1989) Evaluating associativity in CPU caches, *IEEE Trans. Computers* **38**, 1612-1630.

Hill, M.D. (1987) Aspects of cache memory and instruction buffer performance, Ph.D. Thesis, Univ. of California at Berkeley, Tech. Rep UCB/CSD 87/381, p. 489.

Temam, O., Fricker, C., and Jalby, W. (1993) Impact of cache interferences on usual numerical dense loop nests, *Proc. IEEE* **81**, 1103-1115.

Chaiken *et al.* (1991) LimitLESS directories: a scalable cache coherence scheme, in *Proc. 4th Intern. Conf. on Architectural Support for Programming Languages and Operating Systems*, pp. 224-234.

Lenoski *et al.* (1991) Overview and status of the Stanford DASH multiprocessor, in *Proc. 5th Annual ACM Symp. on Principles of Distributed Computing*, pp. 229-239.

Tanoi, S., Tanaka, Y., Tanabe, T., Kita, A., Inada, T., Hamazaki, R., Ohtsuki, Y., and Uesugi, M. (1994) A 32 bank 256-Mb DRAM with cache and TAG, *IEEE J. Solid-State Circuits* **29**, 1330-1335.

Dennis, J. (1979) The varieties of data flow computers, in *Proc. 1st Intern. Conf. on Distributed Computing Systems*, pp. 430-439.

Lee, B. and Hurson, A. R. (1994) Data flow architectures and multithreading, *Computer* **27**, August 1994, 27-39.

Gray, J., Naylor, A., Abnous, A., and Bagherzadeh, N. (1993) VIPER: a VLIW integer microprocessor, *IEEE J. Solid-State Circuits* **28**, 1377-1382.

Kotani, S., Inoue, A., Imamura, T., and Hasuo, S. (1990) A subnanosecond clock Josephson 4-bit processor, *IEEE J. Solid-State Circuits* **25**, 1518-1525.

Harada, Y., Hioe, W., Takagi, K., and Kawabe, U. (1994) The design for a Josephson micro-pipelined processor, *IEEE Trans. Appl. Superconductivity* **4**, 97-106.

PROCESSOR PERFORMANCE SCALING

G. A. SAI-HALASZ
IBM, T. J. Watson Research Center
Yorktown Heights, N.Y. 10598

Technology dependent trends are projected for high performance processors. There are opposite demands placed on the system's area stemming from a need to reduce the proportion of interconnection capacitance and to send signals across the processor. Delays resulting from wiring capacitance decrease if processor area increases, while signal travel considerations favor reducing area. This trade-off for bipolar processors is governed by power density, while for CMOS the processor size primarily is determined by wiring considerations. Judicious planning of interconnections to avoid a so called "RC crises" is necessary to achieve the potential inherent in a technology. The performance limits of bipolar and room temperature CMOS uni-processors are very similar. The highest performance existing technology is liquid nitrogen temperature CMOS. It is not obvious how alternate technologies will fit into the picture of future general-purpose high-end systems.

1. Introduction

A long range view is presented of how evolving technological progress in device design, processing capabilities, and packaging will manifest itself in processor performance. The principal aim is to identify the fundamentals which determine the system level performance of both CMOS and bipolar technologies. We'll look at the point where each technology will be encountering its limits, and see if there are feasible alternates in the offing. The projections are based on a new cycle-time model[1]. The absolute values of the simulated cycle-times are somewhat subject to the assumptions. However, by applying the same assumptions to identical processors embodying differing technologies, cycle-time ratios and trends can be captured quite well. Existing technology and design practices are not necessarily followed; rather the emphasis is on what is accomplishable.

We are dealing exclusively with the performance issues of a single, or uni-, processor. The merits and the promise of various multi-processing strategies are beyond the scope of our discussion. It is believed, however, that without major breakthroughs in general purpose massively parallel computing, there always will be a premium on the highest performing uni-processor.

S. Luryi et al. (eds.), Future Trends in Microelectronics, 125–138.

2. High-end processor properties and modeling

All high performance processors have some common properties stemming from the requirements that they have to be simultaneously complex and fast. Performance is measured in million instruction per second (MIPS). MIPS has two components, cycle time (CT) and cycles per instruction (CPI): MIPS = 1/(CT×CPI). The lower a CPI desired, the more complex the CPU must be. From a practical point of view it means utilizing a vast number of circuits. But, complexity has its limits too. If the addition of more circuits hurts CT more than the improvement achieved in CPI, then it is clearly a loosing trade-off. Also, designing ever more complicated systems may become too complex a problem for worthwhile pursuit. Hitherto circuit count in high performance machines has been increasing steadily, a trend which at the uni-processor level is not expected to continue.

Predicting the CT of a high end processor is actually becoming simpler with technological advancement. The simultaneous requirement of complexity and speed leads to certain general constrains that must be obeyed. High circuit count requires power and large area, while signal propagation favors short distances. These two opposites lead to some well defined optimal configurations, which are relatively insensitive to details.

In the case of complex bipolar mainframes it has been well established how much hardware is needed to achieve a state of the art CPI. Taking this as guidance, we have chosen to analyze a processor which in its bipolar incorporation consists of 4×10^5, 3-input emitter-coupled-logic (ECL) cells, including the peripheral circuits for the memory arrays. It is also assumed that this CPU has a total of 2Mb of static random access memory (SRAM) for cache, various directories, translation tables, etc., and approximately 0.5Mb of read only storage (ROS). The SRAM cells, even in the bipolar machines, are assumed to be 6 device CMOS cells; otherwise, due to the low density of bipolar memories the CPU would have to do with a much smaller cache. Finally, it is assumed that such a processor has 800 signal input/output ports communicating with the outside world. For the CMOS machine, the system has an identical design, and same CPI to the bipolar one, except that the logic operations are carried out by circuits containing MOSFETs rather than bipolar transistors.

To model cycle time, one has to decide how many logic stages to fit into one cycle. Since in general purpose systems the pipeline is usually 3 to 5 deep, we are taking 12 CSEF stages worth of logic as one cycle, which is consistent with breaking the logic path into about four segments. The exact number 12 is not a critical factor. A different number of stages, anywhere between 10 and 15, to first order would only lead to a uniform up or down shift in all of the presented trend curves in the figures.

A generic critical path has additional components besides circuit delays. Delays due to long wires on the chip, and package delay have to be taken into account as well. Accordingly, in the modeling, 11 of the 12 stages drive the average wire lengths of the critical path, while one circuit drives a wire of chip-edge length. The delay through the 11 ECL stages, interconnected by average wires, plus the 1 driving a long wire, is the "on chip" portion of the cycle time. A "package" delay to be added consists of two terms. One is propagation through the package, which we take as having good quality

transmission line properties, giving ~7ps/mm time-of-flight delay. The assumed distance is a "Manhattan" type route across the package as shown on Fig. 1. Two, an additional delay is assigned for signals having to exit and enter chips. The sum of the "on chip" and "package" delay gives a nominal system cycle time. If, as in the case of some CMOS processors, the whole system can fit on a single chip the "package" delay is avoided; however, the on-chip "long" wire delay is added twice in the cycle-time, as illustrated on Fig. 1. This treatment captures the fact that the inter-CPU communication requirements do not depend on how many chips make up the system. Consequently, when the whole CPU consists of only a single chip the on-chip wiring has to take over that communication burden which was carried by the package interconnections in earlier multi-chip processors.

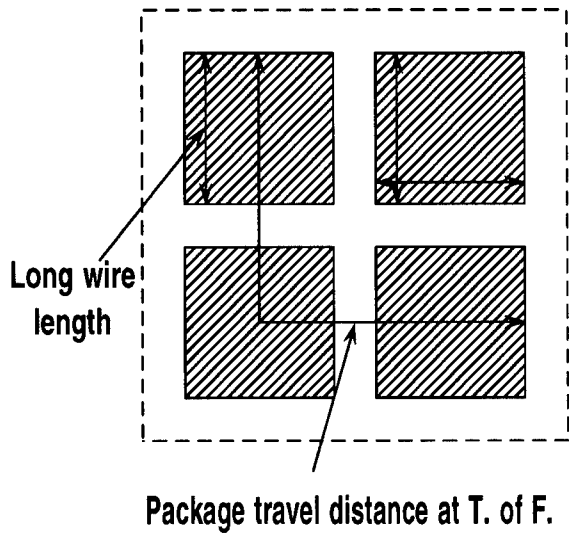


Fig. 1. Schematic depiction of a processor on 4 chips. Package and "long" wire elements of the critical path are indicated

To arrive from the nominal cycle time to an actual one, one should add the time due to various tolerances, clock skews etc.. These however are not addressed here although are critically important. They depend on considerations which are different than those the rest of this paper is dealing with, and would deserve a whole separate treatment. However, with proper attention devoted to tolerances they can be expected to scale with technology in a similar manner as the nominal cycle times, and in percentage terms to remain roughly constant.

Before introducing the modeling and presenting results it will be useful to discuss two particular topics which are of importance in understanding performance issues. One has to do with the physical size of the processor, the second relates to the importance of dealing with interconnect resistance.

2.1. SYSTEM PHYSICAL SIZE AND PERFORMANCE

With advances in technology leading to smaller features and more powerful devices, processors are becoming simultaneously smaller and faster. As a result of this trend it is commonly assumed that smaller area, or equivalently higher circuit density, by itself results in less circuit delay. In reality this is not the case. To understand this, one must look at the role of circuit density alone, and not as an accompanying feature of technology improvement. It is then not difficult to see that under such conditions increasing circuit density, in fact, increases delay. Delay stems from the time it takes the current of the driver circuit to charge the combined capacitance of the receiving circuits and interconnections. Circuit speed improves with increasing area because for optimized circuit layouts the current delivered by the driving circuit is proportional to area, while the interconnection part of the load goes up only as the square root of the area.

The reason that current scales with area is a different one for ECL and for CMOS circuits. For ECL circuits the current is proportional with area because the power level of their operation is determined by the thermal energy (heat) that can be removed from the chip. If a circuit occupies $X\mu m^2$, and the cooling capability is Y watts/cm^2, then the average power at which the circuit can run is $X \times Y \times 10^{-5}$ mW. The output current is proportional to the power level. Since for ECL circuits most of the capacitance is due to wires and not the devices, loading scales as $X^{1/2}$, giving a delay proportional with $X^{-1/2}$.

Let us now look at the situation with CMOS circuits. One expects that in the high-end arena reasonable cooling capability will be affordable, in which case CMOS circuits are not power density limited. The reason that the delivered current is linearly proportional to the circuit area comes from layout considerations. Suppose we have a maximum area of X available for laying out a certain circuit, and we end up with an average device width of x. Let's now assume that the available area for the same circuit changed to X. What we will find is that to the first order the average device width has changed to x. The net is that, at given design-rules, area directly translates into device width. Since delivered current is proportional to device width, circuit current is proportional to area. The argument for the wire loading is the same as for ECL circuits. Since CMOS does not suffer from the power density limitations of the bipolar circuits, it is easy to reach a point where the wire load is less than the loading arising from the receiving circuits. At this point the intrinsic, unloaded, speed of the CMOS circuit places a floor under delay improvements. Circuit delay as function of processor area goes as $(\sim area^{-1/2} + D)$, where D stands for the intrinsic delay of the circuit. With proper power expenditure, in high end processors, one can get to the situation where D becomes the dominant term. At this point decreasing circuit density leads to no further delay improvements.

The advantage gained by increasing area is only part of the picture. The other has to do with signal time-of-flight delays. The reason we'll not increase processor areas indefinitely, is that we have to be able to communicate across the system. When one is faced with cycle times of only a few nano-seconds, time-of-flight delays across

the CPU eventually render area increases detrimental to performance. The next section deals with this aspect of the systems.

2.2. INTERCONNECTIONS AND PERFORMANCE

The delay stemming from wire resistance, commonly referred to as RC delay, to first order is: $R \times L_W \times (C_l + 0.5 \times C \times L_W)$, and has to be included with other delay components. Here, R and C are the resistance and capacitance of unit wire length, L_W is the wire length, and C_l is the load at the end of the wire. The part in the RC delay due to the wire alone does not decrease in spite of scaling to smaller dimensions[2]. The factor that improves the $0.5 \times R \times C \times L_W^2$ term through shorter L_W is negated by the increase in R due to wire cross section shrinkage. Wire capacitance in the meantime remains constant, around 0.2pF/mm for minimum width wires with oxide dielectric[3]. As long as one is dealing with cycle-times over 3-4ns, the wire RC delays are barely noticeable. However, as we'll see later, for ultra high performance CPUs the resistance in the wiring can be important for ECL systems, and critical for CMOS processors. The use of repeaters to regenerate the signal along the way helps, since it decreases the dependence of interconnect delay on wire length from square to linear. But repeaters alone do not solve the problem. Delays would still remain unacceptable, and repeaters entail additional power consumption and design complexity.

If one looks carefully at roles various interconnects play, a better approach suggest itself[1]. High performance processors need two kind of wires. First, there are the wires that serve the vast majority of interconnects. Let's call them "short" ones. For CMOS processors they are typically up to 1-2mm in length. They are mainly responsible for making the chip "wirable" by providing sufficient number of interconnections. Here the RC delay plays no appreciable role. Such "short" wires should follow the minimum lithography features of the available technology. Second, there is a need for "long" wires, where density is secondary to delay considerations. They run between distant parts of the chip, and their characteristic length is that of a chip-edge. A good scaling gauge for such "long" wires is that the time of signal propagation on them should be only a small fraction of the cycle-time. From such considerations immediately follows that the cross section of these wires and insulators, cannot follow minimum lithography features. This type of interconnects will be referred to as "fat" wires. Fig. 2 shows in cross section an example of the interconnection scheme needed by ultra high performance processors. It features a hierarchy of 3 x-y wiring level pairs. The bottom 2 levels are at the finest pitch of which the device technology can take advantage of. Here lines and spaces should be almost at minimum design rules. The next 2 levels' dimensions already pay attention to the RC problem, and finally the top 2 can serve to run signals to full chip-edge length, or longer, distances. With this type of wiring, where conductor and dielectric cross sectional dimensions are scaled together, capacitance per unit length stays constant for each level, while resistance decreases proportionally with wire cross section increase. In Fig. 2, RC in the second x-y plane-pair is 1/4th, and in the third 1/36th, that of the bottom plane.

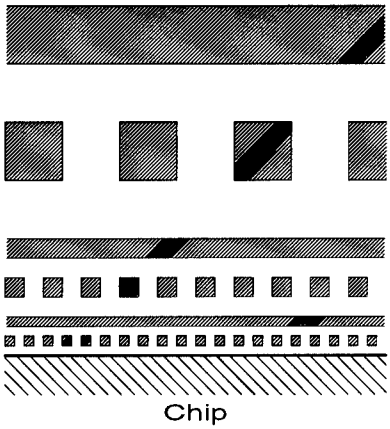


Fig. 2. Example of wiring scheme needed by future high performance processors. Three x-y wiring plane pairs are shown in cross section.

One consequence of having low RC wires is that one will observe transmission line characteristics not only on the package, but also on the chips themselves. With an oxide insulator the minimum delay that a signal can achieve due to the finite velocity of electromagnetic wave propagation is ~7ps/mm. For example, on a 15mm long wire the signal flight time cannot be less than 105ps. This is significantly longer than the switching time of drivers in the considered technologies. When the input of a wire is driven with a faster signal than the travel time down that line, delays are necessarily dominated by transmission line characteristics, and finite signal propagation speed must be taken into account.

The net result is that with the *proper kind of wiring one can avoid a so called "RC crisis"*. The "fat" wire scheme reduces the problem to coping with time-of-flight delays, which for CMOS at least is a much less severe restriction on performance than the RC delay would be.

2.3. MODELING DETAILS

The modeling is described in detail in reference[4]. Here only a few of the more important point will be sketched.

It might seem that the wiring scheme depicted in Fig. 2 wastes too many wiring channels in comparison with having all levels at minimum dimensions. The difference between the number of wiring-channels offered by the two cases is not as large as it might first appear. The reason is that wiring levels block one-another. Thus, a "fat" wire on the top provides less wiring capability than one at fine pitch, but if the fine pitch wire were fully utilized, it would impact more severely the number of available wiring channels in all of the lower levels. If all the pitches are identical, it is estimated that a level blocks ~12 -15% of the wiring capacity of every layer underneath it. This means that one cannot make systems indefinitely smaller by adding wiring levels. There exists a minimum size defined by the interconnections. Indeed, if one uses

efficient wiring protocols and is able to make good use of the available tracks, then 6 or 7 wiring levels are about the maximum useful number.

Besides wiring, the simulation deals with the number of circuits, the size of the memory arrays, the number and sizes of the chips, the interconnection capacitance and resistance, the signal propagation speed in the package, input-output needs, and the time penalty for chip crossing. Circuit timings are done with simple, linearized, equations. The coefficients of these equations capture the essential properties of device scaling. For CMOS effective FET resistance in digital applications is best characterized by the device's large-signal-transconductance[5]. Since large-signal-transconductance has already been experimentally obtained[6,7] for devices from 0.28μm gates down to 0.07μm, one does not need to explicitly deal with intrinsic device parameters, such as mobility, velocity saturation, and the like.

The capacitive load and wire length in the critical paths are obtained from the average net length. Resistive and time-of-flight delays, calculated separately, are combined with the capacitive and intrinsic delay components.

3. Bipolar ECL Processors

The two main indicators in the degree of advancement in a bipolar ECL processor are the power density, and the speed of the bipolar transistors at a given current level.

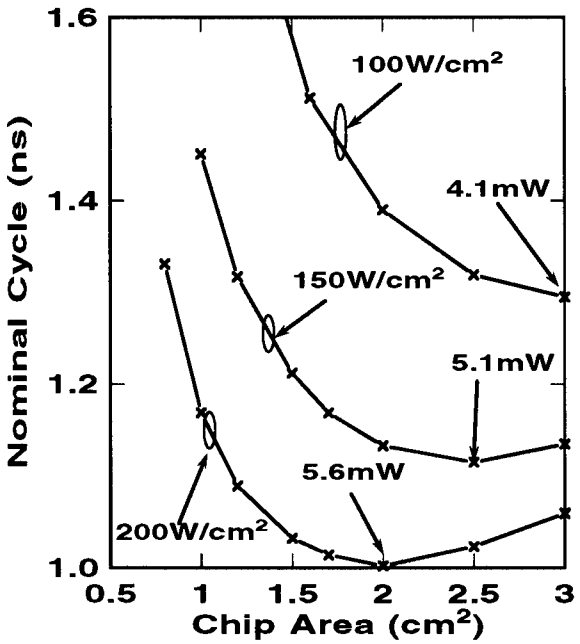


Fig. 3. "Ultimate" bipolar ECL CPU embodying. Power densities and device power in critical paths as indicated.

Fig. 3 and Table 1 show the type of bipolar ECL system which one can regard as an ultimate goal. Device speed is characterized by 14ps delay at 5mW and a 400mV signal swing. Lithographic capability is 0.25μm and we assume that the CPU occupies 4 chips. Clearly, such advances can translate into system speed only if power density is increased as well. The initial sharp delay improvement in Fig. 3 with increasing area is the result of the fundamental behavior of area vs speed we discussed earlier. The more interesting parameters of the 200W/cm^2, 2cm^2 chips ECL processor are given in Table 2.

Table 1. Characteristic parameters in a possible "ultimate" bipolar uni-processor.

Parameters/Characteristics	"Ultimate" bipolar CPU
CPU arrangement	4 chips, 2cm^2 each
Circuits/Technology	ECL, SiGe transistors 14ps unloaded delay @ 5mW
Signal swing	400mV
Lithography	0.25μm
Power density in logic	200W/cm^2
Circuit power in critical paths	5.6mW
Wiring	4 levels: 2 @ 1.6μm pitch 2 @ 6.4μm pitch
Average net length in in critical paths	0.9mm
Average stage delay in critical paths	38ps
On-chip long wires and package delay	500ps
Nominal cycle time	1ns
Logic power consumption	1550W

Even at 200W/cm^2 the system is not capable of making use of the density that lithography would allow. Intrinsic device speed has little to do with ultimate system performance. Indeed, even if in this processor we were to assume zero intrinsic device delay, nominal cycle time would improve only ~15%. Most delays stem from the size of the CPU, which due to power density limitations, cannot be shrunk sufficiently. The problem is that in bipolar ECL circuits steady current is flowing between voltage levels that are related to the silicon bandgap, and thus must stay roughly constant in spite of miniaturization.

4. CMOS Processors

To break the barrier of power density limitation, a technology is needed where power consumption is limited to charging/discharging of capacitances during logic operation. It is also important to be able to decrease power supply voltages along with dimensions. CMOS is such a technology.

Modeling performance in a CMOS processor is more difficult than in a ECL one. Complications primarily arise in calculating wirability for the processors. Wiring statistics[8] involves some empirical factors, which for CMOS processors are not yet well established. Theoretically, the Rent exponent for random logic it is 2/3, while for memories it is about 1/2. For high density CMOS, where a whole processor, made up of functional blocks having relatively little communication with each other, can be integrated on a single chip, the Rent exponent is not well established. We assumed that a processor consists of a small number of building blocks within which a the Rent exponent is high, but the interconnections between the blocks is governed by a smaller exponent. Many cases were investigated to determine what influence such choices have on performance. As it turns out, they have practically no impact. However, they

strongly influence power consumption. Another potential complication in projecting performance is that custom design circuit approaches are increasingly being implemented. Clearly one can't anticipate as yet to be invented design developments, but this is not really needed to project trends. To first order all circuits take equal advantage of progress in technology. A differing design methodology would yield a vertically shifted cycle-time curve roughly in parallel to the presented ones in Fig. 5.

As was indicated earlier, it is assumed that the ECL CPU is mapped into CMOS. It is assumed that 2.5 CMOS NAND gates are in average equivalent to 1 fan-in of 3 ECL circuit. Accordingly, our system under investigation is assumed to consist of 1 million 2-way NAND gates. This means 4 million FET transistors in the logic, and over 12 million in the memory cells. The CMOS critical path also has 2.5 times as many stages, consisting of 30 NAND gates. Again, all but one of the stages are connected with the average wire length of the critical paths. One stage drives a chip-edge-length wire through an inverter buffer. If there is no package delay the inverter and long wire are counted twice. Choosing 2-way NANDs as basic building blocks doesn't mean real processors will be implemented this way, but such a system serves modeling purposes well. With $4x10^6$ logic transistors it captures complexity, and through the 2-way NAND it has a basic circuit whose performance parallels that of any other circuit in reflecting technology improvements.

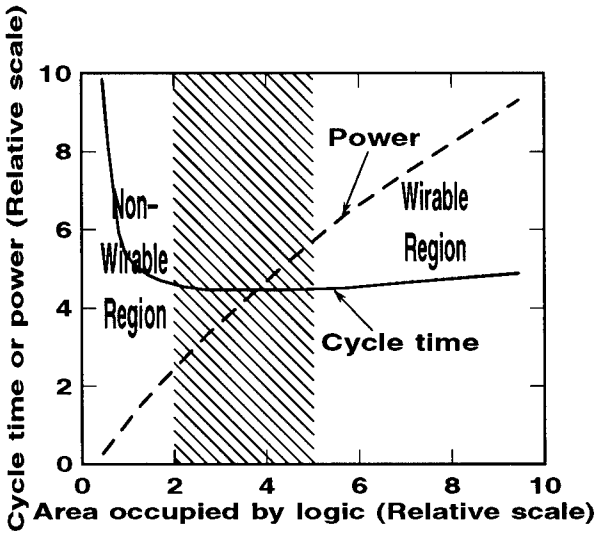


Fig. 4. Performance optimized CMOS processor behavior

The followed simulation method results in a processor optimized for performance. One starts out by assuming a circuit count, for instance, the one million 2-way NAND gates for our example CPU. Next, the minimum wirable processor area is computed based on the circuit count and the assumed wire levels and pitches. Once the processor size has been obtained, the area occupied by the memory cells is subtracted. The remaining area in principle is all available for logic circuits. In reality, layout constrains limit the useful area to some fraction of this total. In all cases this fraction is set to 50%. The area to be populated with devices has to accommodate the off-chip drivers/receivers, and the drivers, possibly repeaters as well, of long on-chip wires. The widths of these drivers are calculated in relation to the characteristic impedance of the interconnects. After allowing for the drivers and receivers, the remaining area

is divided evenly among the NAND gates. Finally, based on this allotted area and the assumed design rules, device widths are made out to be as large as possible.

Fig. 4, showing cycle time and power as function of area, serves to illustrate the situation in high performance CMOS processors. It is presented in relative units since the picture is qualitatively the same in all cases when the the processor can fit into one or two chips. For a given circuit count the larger the area, the wider can the devices be. For the smallest areas, devices are so narrow that the wire load overwhelms them. Cycle time improves rapidly as the fraction of loading stemming from wires decreases, and the circuits are nearing their unloaded intrinsic speed. This improvement is counteracted by the long line delays. However, in absolute size, CMOS systems are small hence propagation delays are less harmful. Accordingly there is a broad range of CPU sizes that allow for essentially identical cycle times.

In the shaded region the processor may or may not be wirable. As discussed, at this stage it is difficult to precisely assess wirability. The boundary of the shaded region toward the left is characterized by an internal Rent exponent of 0.55, and the one toward the right by a 2/3 one. These considerations are based on 6 wiring levels, the kind pictured in Fig. 2. As we can see, by the time a CPU occupies a large enough area to be wirable, it is also large enough that the devices can be made sufficiently wide not to be choked off by wire load. The main price to pay for ever larger areas is power. From one side of the shaded region to the other, cycle time remains practically unchanged, but only because we are more than doubling power consumption.

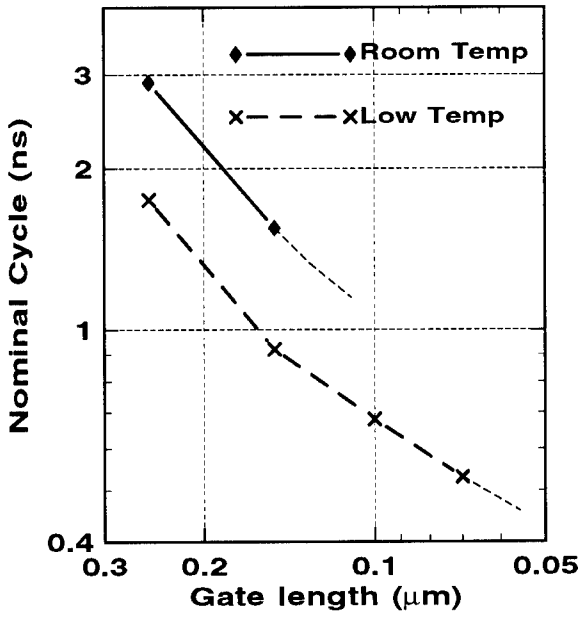


Fig. 5. CMOS high-end processor scaling path

With this introduction we can now look at the expected CMOS CPU performance progress with improving technology given in Fig. 5. The nominal cycle is shown as function of gate length. The gate, instead of the channel, was chosen as measure because it is a straightforward indicator of the state of lithographic capability at any given time. For both the RT and LT processors, the overall lithographic ground-rules are assumed to be more relaxed than the gate length itself. Progressing with a single lithographic level ahead of the overall capability is technically entirely feasible. The

density of device gates is quite small, which means that even electron-beam lithography can be used if the shortest gates were problematic by other lithographic methods. In Fig. 5, at the 0.25μm gate length symbols 0.5μm overall process ground-rules were assumed. The finest wiring pitch was set to to 1.2μm with 0.6μm of lines, spaces, and interlevel separations. At the 0.15μm gate length point ground-rules are 0.25μm, and minimum wire pitch 0.6μm. Such values were chosen as the most probable ones based on how technology is likely to progress.

The RT improvement between the 0.25 and 0.15μm points is a factor of ~1.8, with the 0.15μm gate length cycle time approaching 1.5ns. The improvement results mainly from faster devices as gate length shortens, and as gate oxide goes from 7nm to 5nm. Junction capacitances are also significantly reduced with ground-rules shrinking by a factor of two. With finer wiring the chip size can be reduced by almost by a factor four, and time-of-flight decreases.

To give an indication of the weight with which various parameters influence performance we take a detailed look at an intermediate point: 0.18μm gate length and 0.35μm ground-rules with 2.0V power supply. This processor has 1.8ns nominal cycle at 55W of power consumption, on a 2.5cm^2 size chip. Table 2 shows cycle time sensitivity to some parameters.

We are looking at a processor optimized for performance. All the devices are of maximum width limited only by area considerations. In the critical path circuits loading comes 65% from devices and 35% from wires. The circuits are only 25% slower than their intrinsic unloaded speed. For such a system gate length is the single most important parameter: a 10% change results in 8% cycle time improvement. Since wire loading is minimized already through the use of maximum width devices, 10% improvement in the dielectric constant would bring only 2.5% return in cycle time. However, the dielectric constant can be leveraged to save power. If one uses the same 10% dielectric constant reduction to remain at the original speed, but with narrower devices, one obtains a 15% power reduction. Finally, we come to wire resistance. Performance insensitivity to this parameter results from the use of the "fat" wire scheme. If we did not have the low RC interconnects, but only 6 layers at the minimum pitch fitting this device technology, about 0.85μm, we'd have suffered a 60% performance loss, even if we had been allowed to use very wide wires when needed. Also, adding to complexity, the long wires would be in need of 4 to 6 repeaters, without which cycle time would deteriorate a factor of about 2.5.

Table 2. Parameter sensitivities of an advanced CMOS processor

10% variation in	Resultant variation in cycle time
Gate length	8.0%
Oxide thickness	3.0%
Ground-rules	2.5%
Wire capacitance	2.5%
Wire resistance	0.5%
Power	2.0%

Returning to the discussion of Fig. 5, below 0.15μm gate length we ceased improving ground-rules, keeping them at 0.25μm. There are several reasons for this. First, true 0.25μm lithography is already stretching the limits of technology, and to go below 0.25μm would be difficult indeed. Second, there is not much to gain by shrinking ground-rules further. Third, progressing below 0.25μm can actually be detrimental to performance. The reason is contact resistance. Even a contact resistance value close to theoretical limits, like $2\times10^{-7}\Omega/cm^2$, which is not likely to be ever attained in millions of contact holes in the CPU, is a significant factor at 0.25μm. Shrinking contacts further counteracts other advantages derived from a smaller system.

On Fig. 5, the last RT gate-length point is 0.15μm. The dashed line continuing to 0.12μm reflects the view that somewhere around 0.15μm is the RT gate limit in performance oriented CPUs. Shorter gate-length devices had been made and operated quite successfully at room temperature. But it is one thing to fabricate a few thousand devices, and quite another to have a viable design for a system with millions of transistors, where tolerances play a pivotal role. The main problem with RT operation is the inability to turn off devices sufficiently. For this reason a relatively high threshold of ~0.5V is the lowest viable at RT. The high threshold then requires a power supply around 2V. To live with such a high power supply one has to introduce corrective measures into the device design which are detrimental to performance. And, performance would have been the reason to shorten gates to begin with. In summary, in spite of a recent surge in optimism regarding a 0.1μm RT technology, we are doubtful that in the highest performance CPUs we will see nominal gate lengths much below 0.15μm. On the optimistic side, let's not forget that such a 0.15μm gate-length processor would still be many times as powerful as todays mainframes. With some help from custom circuit design it might reach, or even better, a 1ns cycle time, which would at least equal what an "ultimate" ECL bipolar processor could do.

Finally, lets turn our attention to the LN_2 temperature situation. Looking at Fig. 5, LT obviously offers the highest performance of all systems. The performance advantages of FET LN_2 operation have been recognized and advocated for a long time[9].. However, it appears that until performance improvements can be made at RT, LN_2 operation will remain a matter of discussion only. The aim in presenting Fig. 5 is to contend that shortly LN_2 temperature CMOS will have to be taken seriously because it is the only avenue open toward higher performance processors. As apparent from Fig. 5, LT offers 1.6-1.7 times the performance of a RT processor in the same technology generation. This is mainly due to better device properties, and somewhat to the lower, ~1/5, wire resistance. The main point, however, is that at LT gate lengths can be shortened at least down to 0.07μm[7] with spectacular performance gains. The fundamental reason for scalability of FETs at LN_2 is that devices can be turned off much more readily than at RT. This fact allows for a whole different low threshold, low voltage design space from which RT operation is excluded. Comparing the room and low temperature performances at their probable limits of 0.15 and 0.07μm gate-lengths, we see that LN_2 temperature offers an almost three fold advantage over what is achievable at room temperature. It is worth contrasting the parameters of these two systems, and comparing them with the 1ns cycle-time ECL bipolar processor. Table 3 summarizes the important parameters of the 0.55 Rent exponent scenario.

Table 3. Selected parameters of possible "ultimate" CMOS processors at room and LN_2 temperatures

Parameters/Characteristics CPU with optimistic wiring assumptions	Room Temp. CMOS CPU Single, 0.8cm^2 area chip	LN_2 Temp. CMOS CPU Single, 0.8cm^{2}area chip
Overall lithography design rules	0.25μm	0.25μm
Gate lithography	0.15μm	0.07μm
Gate oxide	5nm	2.8nm
Power supply	1.8V	0.8V
Wiring	6 levels; 2 @ 0.6μm pitch	6 levels; 2 @ 0.6μm pitch
("fat" scheme)	2 @ 1.2μm pitch	2 @ 1.2μm pitch
	2 @ 2.4μm pitch	2 @ 2.4μm pitch
Average net length in critical paths	80μm	80μm
Average stage delay in critical paths	40ps	10ps
Cycle time portion caused by long wires	300ps	220ps
Nominal cycle time	1.5ns	520ps
Logic power consumption	18W	10W

In this technology at LN_2 a logic stage takes only ~10ps; hence the long wire delay plays proportionally a more important role than in slower technologies. Unfortunately, contact resistivity plays a larger role as well. If it could be eliminated it would lead to ~90ps, more than 15% cycle-time improvement with the 0.25μm design-rules. Furthermore, without contact resistivity problems it would be worthwhile to reduce design-rules to, let say 0.15μm, and one could have a 370ps cycle-time processor on a $0.4cm^2$ chip at 10W.

5. Discussion

It has been shown that performance directions at the high end are quite predictable. The time of CMOS in the performance arena has arrived. However, performance potentials are not limitless. A combination of obstacles, mainly due to device behavior, are identifiable.

At the chip size where a given CMOS processor becomes just wirable total wire capacitance scales with design-rules, while total device capacitance scales with $(\text{design-rules})^2 \times C_{ox}$. This means that eventually CMOS too, could find itself overwhelmed by wires. However, this would occur only at dimensions an order of magnitude below the probable device scaling limits. Accordingly, well designed, high performance CMOS processors should be immune to wiring imposed limits. Since wire capacitance per unit length does not scale with lithography, any technology which tries to replace CMOS by shrinking dimensions beyond those of CMOS, will have to face the wire loading problems, and provide the necessary current drive capability to deal with it. We are not aware of any such technology on the horizon.

Contact resistivity between a semiconductor and metal, as discussed earlier, becomes a performance crippling effect at deeply submicron dimensions. Again, any semiconductor technology would have to surmount this obstacle. Along these lines, a deeply scaled Schottky source/drain CMOS might be the path out of contact resistance difficulties.

One has to say a few words about the communication needs of high end processors, either in stand-alone, or in multi-processing configurations. To make full use of future processor performance a communication bandwidth of several hundred Gbit/s should be provided, together with memory capacity measured in Tbits. The communication angle is one where optical techniques based on compound semiconductors could find probable applications.

The considerations above relate to general purpose processors. If one is satisfied to serve only one specific task, for instance, a machine dedicated to matrix inversion, or for data searches, performance for that particular application can probably greatly exceed the ones arrived at in this article. Hardware being ever cheaper, one can envision the end of general purpose machines in the highest end domain. The obstacle is the time spent and the difficulty of system and circuit design. It is worth to consider a research effort into design systems, which could result in fast delivery of a systems dedicated to a specific task. This may ultimately be the most cost effective path toward performance beyond the present projections.

In summary, the performance potentials at both RT and LT CMOS, coupled with the relatively low cost of FET technology, is quite exiting. The real challenge will be for society to absorb what such technologies can deliver.

6. References

1. Sai-Halasz, G.A. (1992) Directions in future high-end processors, *ICCD Digest, p. 230.*

2. Meindl, J.D. (1987) Opportunities for gigascale integration, *Solid State Technology, vol. 30, p. 84.*
3. Edelstein, D.C. (1990) 3-D capacitance modeling of advanced multilayer interconnection technologies, *Proc. of SPIE vol. 1389, p. 352.*
4. Sai-Halasz, G.A. (1995) Performance Trends in High-End Processors, *Proc. IEEE, vol. 83, p. 20.*
5. Solomon, P.M. (1982) A comparison of semiconductor devices for high-speed logic, *Proc. IEEE, vol. 70, p. 489.*
6. Sai-Halasz, G.A., Wordeman, M.R., Kern, D.P., Rishton, S., and Ganin, E. (1988) High transconductance and velocity overshoot in NMOS devices at the 0.1μm-gate-length level, *IEEE Electron Device Lett., vol. EDL-9, 464.*
7. Sai-Halasz, G.A., Wordeman, M.R., Kern, D.P., Rishton, S., and Ganin, E., Chang, T.H.P., and Dennard, R.H. (1990) Experimental technology and performance of 0.1μm-gate-length FETs operated at liquid-nitrogen temperature, *IBM J. Res. Develop., vol. 34, p. 452.*
8. Donath, W.E. (1979) Placement and average interconnection lengths of computer logic, *IEEE Trans. on Circuits and Systems, vol. CAS-26, p. 272.*
9. Gaensslen, F.H., Rideout, V.L., Walker, E.J., and Walker, J.J. (1977) Very small MOSFETs for low temperature operation, *IEEE Trans. Electron Dev., vol. ED-24, p. 218.*

Quantum Devices for Future CSICs

Herb Goronkin
Motorola
2100 East Elliot Road
EL508
Tempe, AZ 85284

1. Scaling of VLSICs

My assignment for this presentation was to convey a view of future electronics that would stimulate lively dialog. I will concentrate on the scaling of ultradense microelectronic integrated circuits. Predicting the future is a high risk behavior and invariably succeeds in stimulating debate. Those most skilled in the art usually choose a time frame that is short enough for reasonable extrapolation of data or long enough to guarantee the predictor will be out of the picture when the results are tabulated. I do not take the view that because a .300 batting average is pretty good in baseball, why not in the game of prediction? Since the purpose of this presentation is to stimulate discussion and ideas, I will begin by describing a few barriers to CMOS scaling, describe the consequences of such scaling and offer a view of what might become the ultimate smallest transistor.

So let's start with some predictions that fall into the category of being nearly safe. In figure 1, the transconductance of heterostructure FETs and silicon MOSFETs is plotted against the gate length. All of the data are taken from the literature. Even if we ignore the underlying physics of the trends of the data, we can predict that when the gate length is less than about 70nm, silicon FETs will have larger transconductance than III-V HFETs. The underlying reason for the superior scaling of silicon is the dependence of transconductance on gate insulator thickness and the dependence of tunnel current on both the energy gap and thickness. The tunnel current is proportional to

S. Luryi et al. (eds.), Future Trends in Microelectronics, 139–149.

$$\exp(-2\sqrt{\frac{2m^*\Delta E}{h^2}}T_{ox}) \qquad (1)$$

where ΔE is the majority carrier band offset and T_{ox} is the insulator thickness. The transconductance, g_m, is approximated by

$$g_m = \frac{\varepsilon v Z}{T_{ox}} \qquad (2)$$

where v is the channel velocity and Z is the gate width. Figure 2 shows the scaling trend of gate length and gate oxide thickness based on extrapolation of introduction dates of DRAM families. In order to maintain approximately the same magnitude of transconductance as the gate width shrinks and in order to avoid short channel leakage as the gate length scales, the gate oxide thickness must scale with the gate length. In silicon, for gate oxide thickness below about 3nm and supply voltage of about 1.0 volt, the tunneling current will be several orders of magnitude larger than the channel current and the gate cannot control the channel charge. In III-V HFETs the insulator thickness must be larger to compensate for the much smaller ΔE which is on the order of 0.3 - 1.0 volt compared to 3.5 volts at the silicon-silicon dioxide interface.

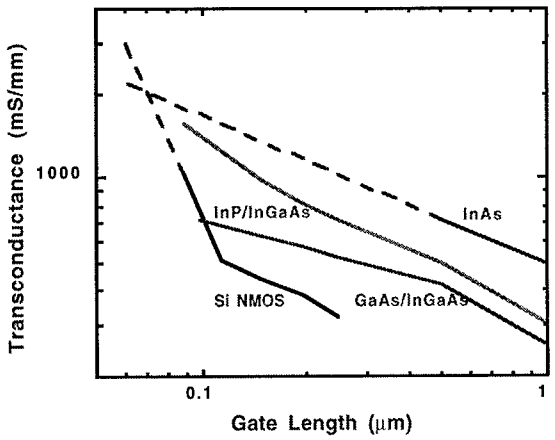


Figure 1. Trends of experimental transconductance as a function of gate length for FETs of various materials.

As a consequence of maintaining the tunnel current at a much smaller level than the channel current, the thickness of AlGaAs or InAlAs, etc. is usually greater than 20nm and it is maintained at that thickness as the gate length shrinks. For this reason, the transconductance slope is much smaller for the III-V FETs than for silicon and even though the drive capability of III-V is far superior to silicon for large gate lengths, vertical scaling is limited by tunnel leakage current.

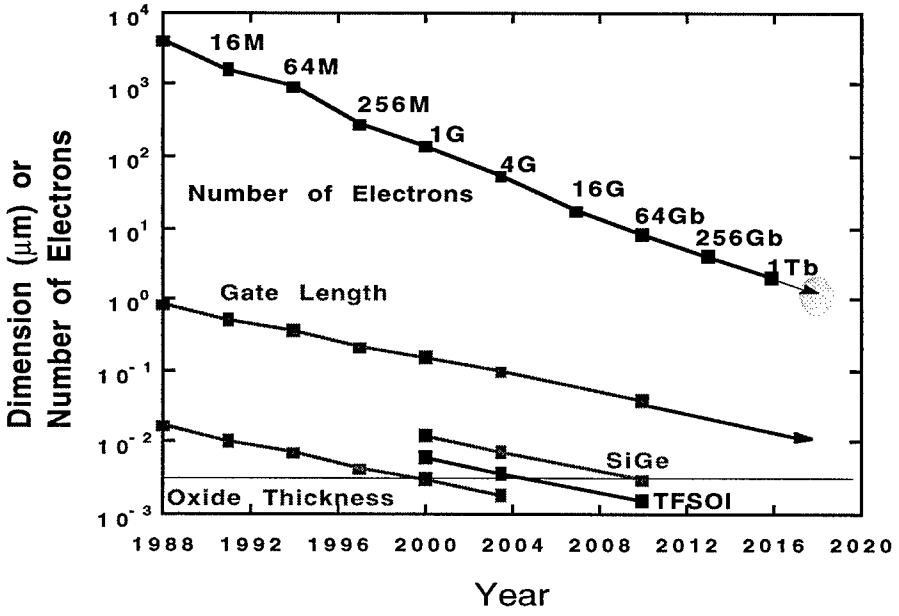


Figure 2. Extrapolated historical trends of DRAM scaling.

Using Figure 2, we can also predict that the number of electrons in the DRAM channel will decrease as the density of transistors increases. This arises from the need to keep the chip from melting during operation and to maintain a reasonable chip cost.. The projected number of electrons in the 64Gb DRAM channel is about 10. We might expect somewhat unusual I-V characteristics as one or two electrons, which represents a large fraction of the total number of electrons, enter or leave the channel. Let's examine this possibility in a little more detail.

Figure 3 illustrates the concept of Coulomb blockade. When the device and associated capacitance are sufficiently small, an electron entering the

channel causes the channel or quantum well ground state to shift by an energy equal to

$$E = \frac{q^2}{C} \tag{3}$$

where E is the energy and C is the device input capacitance. The value of C determines whether the I-V characteristics will be smooth as is in classical transistors or stepwise. Accordingly, if we consider the normal operating range of 0-80C, E must be greater than 4kT giving the result that C must be less than 0.6aF. At this temperature a larger value of capacitance will cause the steps to become smeared and continuous. For practical device operation a more stringent condition of 40kT should be used to achieve device stability.

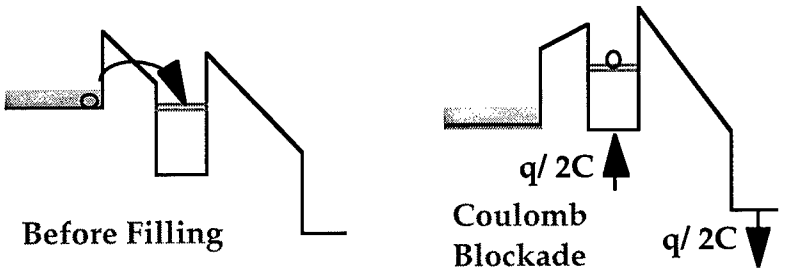


Figure 3. The electron transferred into the quantum well raises the energy level by an amount $q^2/2C$ and a voltage equal to q/2C must be applied to pull the eigenstate back into alignment with the electron gas in the emitter.

Using the scaling trend, let's examine whether the 64Gb DRAM is likely to exhibit normal or step-like I-V characteristics at normal operating temperatures. Figure 4 shows the gate capacitance, number of electrons and corresponding DRAM density as a function of gate length. From the capacitance requirement for 300K operation the arrow points to the maximum value of capacitance for stepwise current increase. Continuing to use the 64Gb DRAM as an example, although the number of electrons scales to about 10, the capacitance scales to about 70aF. In order for clear single electron behavior to be observable, the temperature needs to be reduced to 3.3K. Therefore, we can predict that if traditional scaling continues, the 64Gb DRAM will have smooth I-V characteristics.

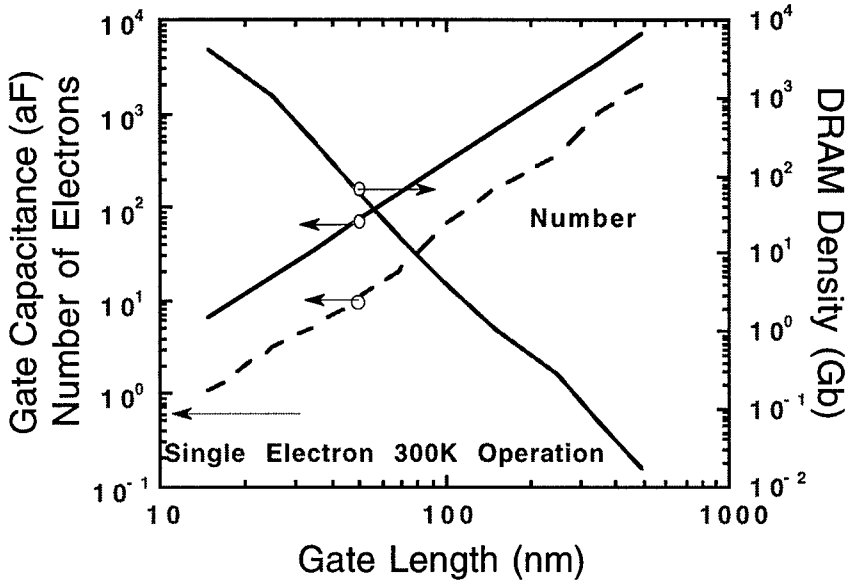


Figure 4. Scaling of gate capacitance and number of electrons with DRAM density. The arrow points to the maximum value of capacitance for observation of sharp stepwise I-V characteristics.

From statistical thermodynamics, we can obtain an expression for the fluctuations in the number of electrons in a perfect gas open system, i.e. one in which the number of electrons can change while the volume and pressure are constant is given by

$$\frac{\Delta N}{N} = \frac{1}{\sqrt{N}} \tag{4}$$

Consider a transistor with 100 electrons. The fluctuation, $\Delta N/N$, in electron number will be about 10%. If the transistor is used as a switch in a conventional inverter and the load transistor has a comparable electron number, the variation in the low state will be substantially increased as shown in Figures 5 and 6. Consequently, the noise margin will substantially diminish or vanish. In order to reduce the variation in V_{Low} arising from electron number fluctuations, the refresh rate can be increased. The rule of thumb for refresh rate is about 1ms/Mb. As this rate is increased, perhaps as much as a factor of 100, the refresh circuits dissipates a proportional amount of power. The logic low state uncertainty can be improved with the use of a resonant tunnel transistor load as shown

in the same figure. When the transistor scales to 10 electrons, the fluctuations will increase to about 30%. In this range, even though smooth I-V characteristics can be obtained, it is not at all clear that conventional transistors can operate in a practical voltage and temperature range, i.e. 300K and 1 volt. If the current scaling trend is followed, according to Figure 2, this difficulty will have to be faced in about 2004.

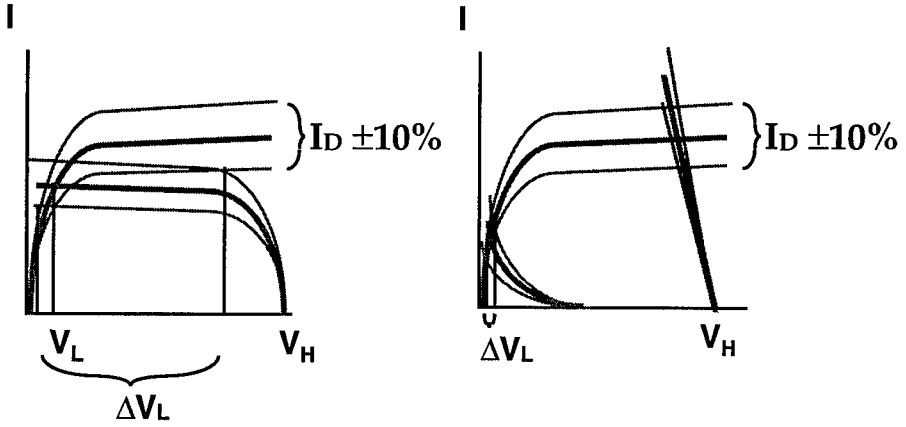


Figure 5. Inverter characteristics for conventional transistors and a resonant tunnel transistor load, each having 10% variation in current due to electron fluctuations.

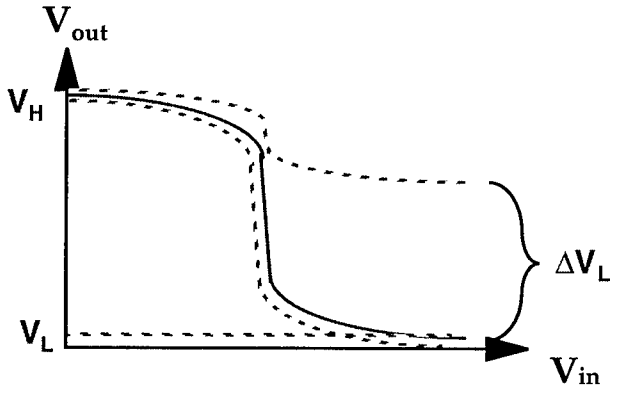


Figure 6. Transfer curve illustrating the effects on the transfer curve of ±10% level of electron fluctuation in the switch and load transistors.

2. CSIC-Collosal Scale Integrated Circuits

Let us start with a working definition of CSIC.

Density	1Tb
Voltage	0.25V
Power dissipation	10W
Access time	10ns
Max operating temperature	80C
Chip size	2.5x2.5 cm

These specifications lead to the following device requirements:

Average power dissipation/bit	10pW
Cell Size	25x25 nm
Cell transistor size	5x5 nm
Transistor capacitance	1.4 aF
Number of electrons/cell	2

If we suppose that scaling can take us to the CSI regime, the above discussion leads to the conclusion that in about 2010 we will have to have a nanometer-scale fabrication technology in manufacturing labs in order to meet the introduction date of 2016. Since the electron density scales to 2 electrons per cell, conventional transistors will have to be replaced by a more deterministic structure. In addition, with scale lengths on the order of ones to tens of atoms, there is presently no conceivable lithography that can perform pattern transfer to the accuracy of atomic dimensions.

3. The Molecular Transistor

In order to open a dialog on the problems of scaling to small electron numbers and the associated small dimensions, we should try to envision the smallest possible transistor. Of course, we are assuming that the smallest practical device will actually be a transistor. We propose that the ultimate transistor will be a three terminal field effect molecular transistor fabricated using molecular self assembly. Although the proposed structure can be fabricated using semiconductor quantum wells, we focus here on the smallest limit. The structure combines the approaches of Likharev[1] and Lent/Porod[2]. It uses coulomb interaction in a gate region to control the charge in a central region of a single electron structure. Figures 5, 6 and 7 illustrate the principles.

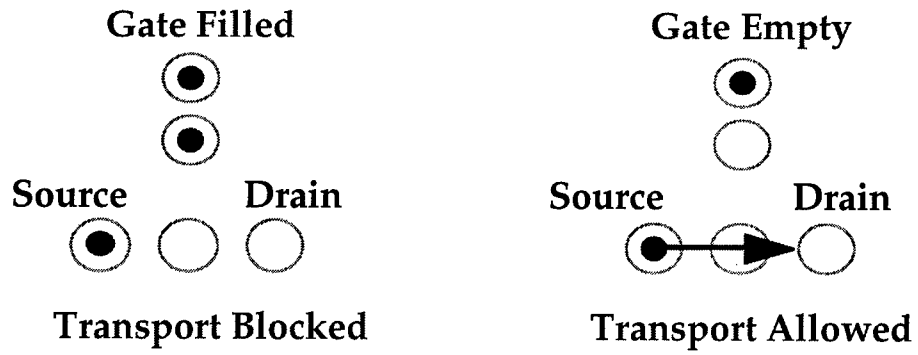


Figure 5. Schematic of molecular transistor. Charging of gate atom repels electron from middle atom of the source-drain molecule.

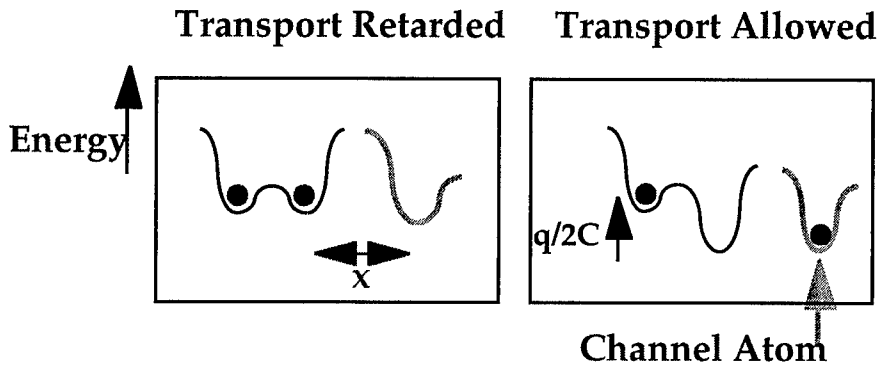


Figure 6. Transfer of charge in gate circuit retards or allows charge transfer in channel.

The source-drain circuit is a molecular backbone along which charge is transferred by a bias potential. As the charge moves from atom to atom, the potential energy of the atom changes by an amount equal to q/2C. The energy levels along the short chain (Figure 7) shift up or down according to whether the states are full or empty.

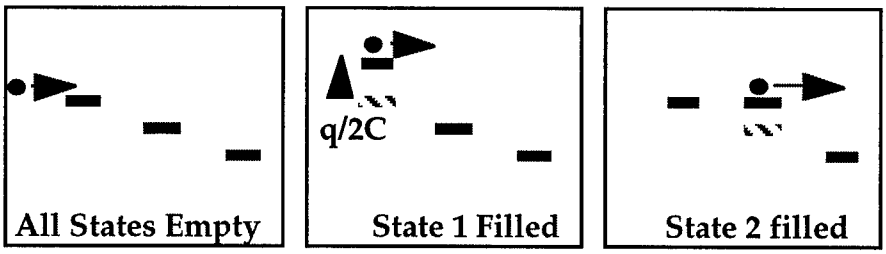


Figure 7. Energy levels of atoms in the source-drain circuit. The energy level depends on the occupancy of each state.

The gate can contain one or two electrons. If the gate contains one electron in the position nearest the channel, a channel electron can repel the gate electron. The second electron holds the bias electron in place. The bias electron raises the second atom state and this retards the flow of charge into that stage. We can derive a transconductance for this structure

$$g_m = \frac{\varepsilon}{\tau x^2} \qquad (5)$$

where ε is the permittivity and τ is the single electron transfer time.

4. Technology for CSICs

This section is for handwaving. It is here that the problems of using the smallest transistor begin to emerge and the solutions to those problems have not been adequately addressed. So I will make some predictions and hope they will serve as a starting point for further discussion that may actually lead to useful approaches.

Silicon factories will probably continue to be the mainstream manufacturing media. The introduction of self assembled molecular transistors into CSICs will initially involve hybrid fabrication in which the input and output devices are silicon and the dense memory and logic regions are molecular transistors. Molecular self assembled (MSA) species will be specifically synthesized with unique head and tail units to provide local interconnects as well as nucleating points. The silicon circuit will be composed of local molecular ICs which are interconnected to form the total IC. The arrangement will be controlled by a combination of prepared surface regions for selective self assembly and the interaction between MSA species. In this way, two and three dimensional local ICs can be fabricated.

The merging of conventional and molecular processing requires a scheme for transmitting signals between the molecular transistors and the I/O buffers and other silicon logic. This will probably be accomplished using self assembled molecular transitions in which the molecules have multiple head and tail groups for reacting with specifically prepared silicon regions or MSA transistor tail groups.

The accuracy of computation with sparse electron systems will depend on the magnitude of electron fluctuations in the self assembled molecules as well as fluctuations in the connecting molecular chains. Errors in the IC increase the entropy. The total free energy, G = U - TS, where U is the internal energy of the system, can thus decrease to the point where the internal memory or logic is lost. Thermal fluctuations will be moderated by the stronger bonding of charge carriers to atoms in the molecular transistors and may offer a means of reducing errors.

Carriers will move more slowly than in semiconductor devices. In order to take advantage of the small transistor size and the automatic formation of local interconnects as well as to deal with the inherent slowness of molecular transistor action, it will be necessary to utilize massive parallel processing to exploit molecular electronics and achieve the goal of 10ns access time.

5. Summary

Traditional scaling of silicon transistors inexorably leads to barriers in fabrication and utilization which arise from the discrete nature of the crystalline lattice and the sparse electrons which must carry information without substantial loss of data. Silicon scales more favorably than III-V semiconductors because the large bandgap of SiO_2 allows vertical scaling to continue to about 3nm before significant tunnel current occurs. However, even though scaling to multi-megabit DRAM levels may be possible using extentions of current fabrication technology, when the 64Mb level is reached, the number of electrons in the channel will scale to about 10. Thermal fluctuations in charge will be on the order of 30% and can seriously upset logic levels.

New technologies can intercept and revector the scaling trend. We proposed a molecular transistor that is fabricated by self assembly. Such transistors will initially be merged with conventional silicon devices and wafers.

The time frame in which new technology will be needed, according to the trend chart, in the range of 2015-2020. Long before it is introduced into the marketplace, the technology must be developed in research labs. The

complexity of CSICs will require a plethora of new approaches to fabrication, architecture and utilization. If we are to hit the market window, the time to begin development is now.

6. References

[1] Likharev, K.K. (1991) Single Electronics: Correlated transfer of single electrons in ultrasmall junctions, arrays and systems, Granular Electronics. D.K. Ferry, J.R. Barker, C. Jacobini, Eds. New York: Plenum, 371-391

[2] see for example, Tougaw, D. P., Lent, C.S. (1994) Logical devices implemented using quantum cellular automata, J. Appl. Phys **75** (3), 1818-1825

Challenges and Trends for the Application of Quantum-Based Devices

Gerald J. Iafrate and Michael A. Stroscio
U.S. Army Research Office
Research Triangle Park, North Carolina 27709
U.S.A.

1. Introduction

As semiconductor technology continues to drive the scaling of electronic device dimensions into the ultrasubmicron, nanodimensional regime, many ultrasmall and ultrafast concepts and phenomena will continue to be put forth for notional consideration. The stunning achievements of nanofabrication in the last decade now allow for band-engineering and atomic-level structural tailoring not heretofore available or explorable except through naturally occurring atomic and molecular processes. Indeed, the techniques of atomic layer epitaxy facilitate the growth of structures atomic layer by atomic layer; as well, advanced lithographic techniques are capable of defining "lateral" structures with an accuracy of about 50 Angstroms.

These revolutionary trends and nanofabrication techniques have opened the way to fabricating quantum wells, quantum wires and quantum dots that may provide the basic building blocks for future nanoelectronic and mesoscopic (quantum) device technologies. As well, these trends provide new opportunities for realizing quantum-based information processing devices but many challenges must be addressed and intensive international basic research is essential for the full exploitation of these revolutionary devices.

2. Highlights of Selected Challenges

A major challenge in the application of quantum-based devices arises from the molecular feature-size considerations implicit in nanoelectronic and mesoscopic technology; indeed, many interesting questions arise concerning fluctuations, tolerances, robustness, and other statistical considerations which might conceivably wash-out many of the seemingly fragile characteristics of nanodevices. There are many illustrative examples in which inherent statistical variation in composition and device dimensions produce substantial deviation from the desired nanostructure electrical response. These examples include: minimum metal-oxide-semiconductor transistor size as determined by a combination of gate oxide breakdown, drain-source punch-through and substrate doping fluctuations [1]; minimum planar bipolar transistor size as determined by a combination of collector junction breakdown, base punch-through and base doping fluctuations [2]; effects of structural and alloy disordering on the electronic states in quantum wires [3];

S. Luryi et al. (eds.), Future Trends in Microelectronics, 151–157.

and the effect of fabrication-related dimensional variations on carrier scattering rates in quantum wires [4,5].

Clearly, a challenge for nanoelectronic and mesoscopic device communities is to explore concepts and designs that optimize robustness and suppress device fragility. In meeting this challenge it will be essential to circumvent statistical phenomena which typically exacerbate the scaling of conventional device technologies to reduced feature sizes; as well, it will also be essential to exploit phenomena specific to the nanoscale regime which are known to suppress noise and related fluctuations in "small" and mesoscopic devices. On a more positive note, a related challenge for nanoelectronic and mesoscopic device communities is to explore concepts and designs that optimize robustness and suppress device fragility.

3. Illustrative Research Trends and Requirements

Conceptual options for achieving robustness are an essential ingredient for the successful exploitation of quantum-based information processing devices and systems. Indeed, several phenomena which play critical roles in nanoelectronic and mesoscopic devices provide important opportunities to realize robustness in such nanodimensional devices.

Promising avenues for achieving robustness include the application of Coulomb blockade effects, design through quantum control theory, and emulation of biological and chemical systems where phenomena such as neuron networking and self-organization finesse disordering processes. For example, potential applications of quantum control theory to "small" and mesoscopic electronic devices are motivated based on past uses of robust optimal control theory for the selective excitation of quantum mechanical vibrational states of molecules [6-10]. Additionally, a number of important, recent developments in the field of Coulomb blockade are highly encouraging; these developments include: the observation of Coulomb blockade effects at temperatures which are an appreciable fraction of room temperature [11]; theoretical prescriptions for enhancing the reliability of single-electron switches operating on the basis of Coulomb effects [12]; and recent progress in understanding how mesoscopic Coulomb blockade effects may be used to greatly suppress noise in electron emission processes in p-i-n semiconductor junctions as well as in p-n microjunctions [13]. The observation of Coulomb blockage effects at 100 degrees Kelvin [11] have been extended recently by the principal authors of [11] to 110 degrees Kelvin for the case of holes and 170 degrees Kelvin for electrons. As a means of expanding on the progress made in some of these areas, this paper places particular emphasis on two specialized topics: the simulation of the capacitance of quantum dots [14]; and the tailoring of deformation potential and piezoelectric scattering in mesoscopic devices in order to maintain de Broglie wave coherence [15].

The impressive trends in both the fabrication and design of quantum-based devices motivate and lead to the need for circuit design tools for integrated circuits with

quantum-based component devices. In the recent past there has been substantial progress in the development of design tools for ultrafast and compact circuits using heterojunction bipolar transistors and negative differential resistance (NDR) devices such as resonant hot-electron transistors, resonant tunneling transistors, and resonant tunneling diodes; examples of these works are summarized in [16]. Indeed, these efforts are realized, in part, based on a new circuit simulator, NDR-SPICE, which has been developed by extending the Berkeley SPICE simulator to the domain of NDR circuits. To date such tools have been used to design a wide variety of circuits including: multiple-valued multiplexers and demultiplexers; totally parallel multiple-valued logic adders; four-valued up/down counters; analog-to-digital converters; and a 32-bit parallel correlator.

Recent successes in defining quantum logic gates that are potentially suited as the building blocks of quantum logic networks [17-19] portend new vistas and enormous payoffs through the realization of quantum computers. This possibility has not been the focus of extensive attention within the microelectronics and nanoelectronics communities but these communities have created fabrication and design technologies that may provide the basis for realizing quantum computers.

4. Quantum Capacitance

Through the dramatic advances in nanofabrication technology it is now feasible, indeed routine, to produce nanoscale devices that manifest electrical properties that are determined principally by the laws of quantum physics rather than classical physics. From the point of view of possible applications of quantum-based devices, it is essential that the electronic properties of such nanoscale devices be understood fully. Enlightening simulations of the capacitance of quantum dots [14] reveal a rich atomic-like structure for the capacitance of two-dimensional circular quantum dots modelled on the basis of a self-consistent solution of the Schroedinger equation; in these studies many-body effects were included using a local density approximation as well as the optimized Kreiger-Li-Iafrate (KLI) exchange potential. Figure 1 depicts the capacitance versus electron number for a two-dimensional gallium arsenide quantum dot with a radius of 200 nm (empty squares), 100 nm (solid circles), and 50 nm (empty circles) in the presence of a conducting backgate located at a distance of 1/10 of the radius. At such short distances, it was demonstrated [14] that the presence of a backgate significantly increases the capacitance. As is evident from Figure 1, the capacitance exhibits pronounced minima corresponding to shell filling just as in the case of an atom. These shell-like grouping were shown [14] to be due to degeneracies introduced by symmetry and the dips are the consequence of the increased difference between the chemical potential values for consecutively accumulating electrons as each new set of degenerate orbitals starts to be filled. These shell-like groupings are illustrated clearly in Figure 2 where the capacitive energy calculated with the KLI exchange potential is

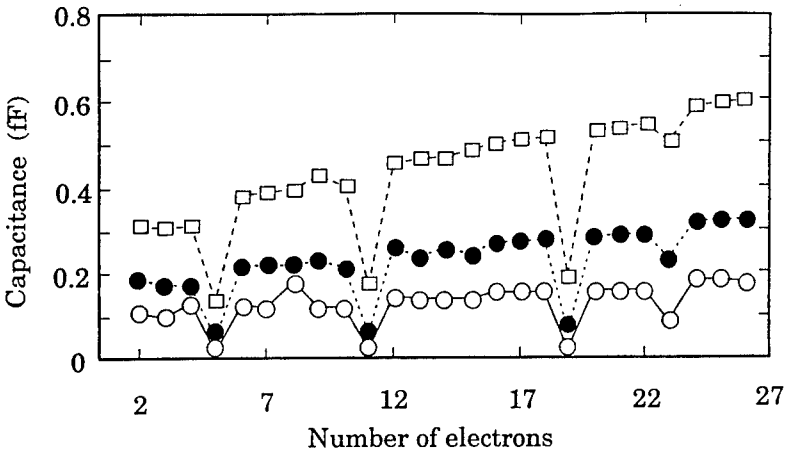


Figure 1. Capacitance versus electron number for a two-dimensional GaAs quantum dot with dimensions as defined in the text.

plotted as a function of the electron number of atoms with nuclear charge number Z. These non-classical effects in the quantum capacitance should be important in a variety of applications of quantum-effect devices including quantum-dot memory cells for compact high-density data-storage devices.

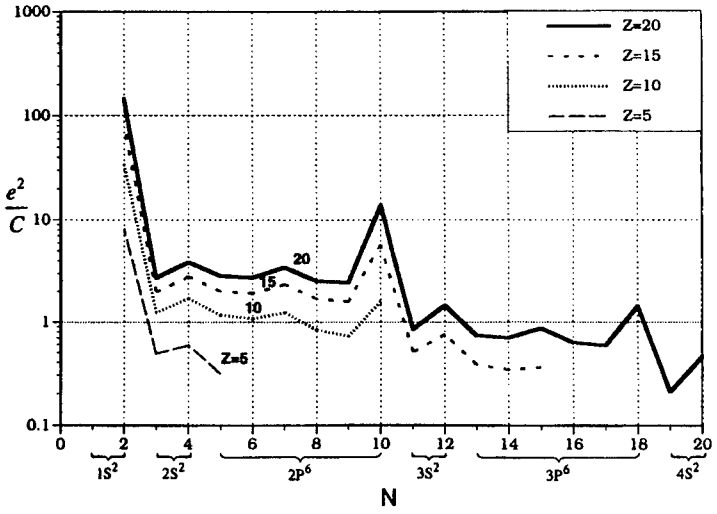


Figure 2. Capacitive energy versus electron number for atoms with nuclear charge Z.

5. Tailored Scattering Rates in Mesoscopic Devices

In order to tailor the strengths of deformation potential and piezoelectric scattering in mesoscopic devices it is essential that the spectrum of acoustic phonons in these mesoscopic devices be understood. Acoustic phonons have been quantized for a variety of nanoscale and mesoscopic structures in order to assess the role of electron--acoustic-phonon scattering in limiting the performance of nanoscale and mesoscopic electronic devices. These structures include quantum wells, quantum wires with cylindrical and rectangular cross sections, and quantum dots with spherical, cylindrical and rectangular boundaries [20]. These quantized phonons have been studies for the two cardinal boundary conditions of classical acoustics: free boundaries (open boundaries) where the phonon displacements are unrestricted and allowed to balance all normal traction forces to zero; and clamped boundaries (rigid boundaries) where phonon displacements are required to vanish at the boundaries.

For the case of quantum wires, scattering rates have been calculated only for the case of infinitely long quantum wires and, as appropriate for this case, the acoustic phonons have been quantized in only the lateral dimensions. However, for realistic mesoscopic device designs, the quantum wire input and output "leads" as well as the active regions of the devices with quantum-wire geometries have finite lengths. Accordingly, deformation and piezoelectric scattering rates must be based on acoustic phonons that are quantized in all three spatial dimensions.

The international community does not appear to have considered the role of three dimensional confinement of acoustic phonons in mesoscopic devices but it is clear from the solutions of classical acoustics that boundary conditions imposed at the ends of wire-like regions can have a profound effect on the properties of acoustic modes. Based on our current understanding of such finite wire-like structures, it should be possible to "engineer" mesoscopic structures so that electron--acoustic-phonon scattering is reduced as a result of reducing the overlap of the electronic wavefunctions and the phonon standing modes. As a straightforward example, the standing-mode pattern for acoustic modes in a free-standing quantum wire with cylindrical cross section and finite length has a simple analytical form [15] which facilitates the selection of device geometries and dimensions that minimize electron--acoustic-phonon scattering rates. Such an "engineered" reduction is likely to be most important in mesoscopic devices which operate on the basis of "coherent" electron-wave interference effects.

6. Conclusion

The challenges before the nanoelectronic and mesoscopic device communities are --- on the one hand --- to explore concepts and designs that optimize robustness and --- on the other hand --- to exploit quantum effects that open the way to unique means of information processing that have no counterparts in the classical domain. Through this dual approach the nanofabrication technology revolution can be exploited fully by the

architects of future generations of quantum-based information processing devices and systems. As predicted by the famed physicist Richard Feynman in his 1959 talk entitled "There's Plenty of Room at the Bottom," we shall realize the "bottomless" possibilities of "manipulating and controlling things on a small scale."

7. References

1. Hoeneisen, B. and Mead, C. A. (1972) Fundamental limitations in microelectronics - I. MOS technology, *Solid-State Electronics* **15**, 819-829.

2. Hoeneisen, B. and Mead, C. A. (1972) Fundamental limitations in microelectronics - II. Bipolar technology, *Solid-State Electronics* **15**, 891-897.

3. Singh, Jasprit (1991) Effect of structural disorder on electronic states in GaAs/AlGaAs quantum wires, *Appl. Phys. Lett.* **59**, 3142-3144.

4. Mickevicius, R., Mitin, V. V., Kim, K. W., Stroscio, Michael A., and Iafrate, Gerald J. (1992) Electron intersubband scattering by confined and localized phonons in real quantum wires, *J. Phys. Condens. Matter* **4**, 4959-4970.

5. Mickevicius, R., Mitin, V. V., Kim, K. W., and Stroscio, Michael A. (1992) Electron high field transport in multisubband quantum wire structures, *Semicond. Sci. Technol.* **7**, B299-B301.

6 Warren, Warren S., Rabitz, Herschel, and Dahleh, Mohammed (1993) Coherent control of quantum dynamics: The dream is alive, *Science* **259**, 1581-1589.

7. Shi, Shenghua, and Rabitz, Herschel (1992) Optimal control of selectivity of unimolecular reactions via an excited electronic state with designed lasers, *J. Chem. Phys.* **97**, 276-287.

8. Beumee, Johan G. B., and Rabitz, Herschel (1992) Robust optimal control for selective vibrational excitation in molecules: A worst case analysis, *J. Chem. Phys.* ***97*** 1353-1364.

9. Rabitz, Herschel (1992) Optimal control of molecular motion, in A. D. Bandrauk and S. C. Wallace (eds.) *Coherence Phenomena in Atoms and Molecules in Laser Fields*, Plenum Press, New York, pp. 315-331.

10. Dahleh, M, Pierce, A. P., and Rabitz, H. (1992) Design challenges for control of molecular dynamics, *IEEE Control Systems* **12**, 93-94.

11. Leobandung, Effendi, Guo, Lingjie, Wang, Yun, and Chou, Stephen Y. (1995) Observation of quantum effects and Coulomb blockage in silicon quantum dot transistors at temperatures over 100 Kelvin, *J. Appl. Phys.*, in press.

12. Imamoglu, A. and Yamamoto, Y. (1993) Noise suppression in semiconductor p-i-n junctions: Transition from macroscopic squeezing to mesoscopic Coulomb blockade of electron emission processes, *Phys. Rev. Lett.* **70**, 3327-3330.

13. Imamoglu, A., Yamamoto, Y., and Solomon, P. (1992) Single-electron thermionic-emission oscillations in p-n microjunctions, *Phys. Rev. B* ***46***, 9555-9563.

14. Macucci, M., Hess, Karl, and Iafrate, G. J. (1995) Simulation of electronic properties and capacitance of quantum dots, *J. Appl. Phys.* **77**, 3267-3276.

15. Stroscio, Michael A. and Kim, K. W. (1994) Piezoelectric scattering of carriers from confined acoustic modes in cylindrical quantum wires, *Phys. Rev. B* **48**, 1936-1941.

16. Mohan, S., Mazumder, P., Haddad, G. I., Mains, R. and Sun, S. (1991) Ultrafast pipelined adders using resonant tunneling transistors, *IEE Electronics Letters* **27**, 830-831; Mohan, S., Mazumder, P., and Haddad, G.I. (1991) Subnanosecond 32-bit multiplier using negative differential resistance devices, *IEE Electronics Letters* **27,** 1921-1931; Mazumder, P. (1994) Picosecond pipelined adder using 3-terminal devices, *IEE Proceedings E, Computers and Digital Technics* **141**, 104-110.

17. Barenco, Adriano, Deutsch, David, and Ekert, Artur (1995) Conditional quantum dynamics and logic gates, *Phys. Rev. Lett.* **74**, 4083-4086.

18. Sleator, Tycho and Weinfurter, Harald (1995) Realizable universal quantum logic gates, Phys. Rev. Lett. **74**, 4087-4090.

19. Cirac, J. I. and Zoller, P. (1995) Quantum computations with cold trapped ions, *Phys. Rev. Lett.* **74**, 4091-4094.

20. Stroscio, Michael A., Kim, K. W., Yu, SeGi, and Ballato, Arthur (1994) Quantized acoustic phonon modes in quantum wires and quantum dots, *J. Appl. Phys.* **76**, 4670-4675.

WIRE AND DOT RELATED DEVICES

E. GORNIK, V. ROSSKOPF, P. AUER, J. SMOLINER, C. WIRNER,
W. BOXLEITNER, R. STRENZ[+], G. WEIMANN[+]
Institut für Festkörperelektronik, Floragasse 7, A-1040 Wien, Austria
[+]Walter Schottky-Institut, TU München, D-85748 Garching,

Abstract

We report measurements of electronic excitations in the confinement regime between zero and two dimensions. FIR transmission spectroscopy has been successfully used to detect 2D plasmons dispersion, localized plasmons and depolarization shifted 1D intersubband transitions. The far infrared response of arrays of periodic quantum wires has been investigated by cyclotron resonance transmission and photoconductivity (PC) measurements. Due to narrow geometrical dimensions (300 nm) quantum confinement arises and leads to the formation of one-dimensional electronic subbands with a typical energy separation of 1-3 meV in the case of the heterostructures and up to 9 meV in the case of the quantum wells. The far infrared transmission spectra of the quantum wire structures show one strong resonance, which can be described by an harmonic oscillator model, assuming that the confining potential is of parabolic shape. Depending on the intensity of bandgap illumination a well pronounced transition from an one-dimensional electronic system behaviour to a modulated 1D system and finally to a pure 2 dimensional system can be achieved for the heterostructure samples. In addition, results of Smith-Purcell-type far infrared emission in one dimensional quantum wires and tunneling experiments on low dimensional systems are also decribed.

1. Introduction

The characterization of nanostructures is a topic of vital interest, since the dimensions of commercially available semiconductor devices are now in a range, where quantum size effects become evident. In this paper methods are presented to characterize low-dimensional structures. Far infrared transmission and photo conductivity measurements are used to determine the low-dimensional properties, e.g. subband energies[1].

S. Luryi et al. (eds.), Future Trends in Microelectronics, 159–169.

Much insight into the electronic properties of low-dimensional electron systems is gained from investigations of their FIR excitations. It has been demonstrated that in etched [2] as well as field-effect [3] induced low-dimensional electron systems the bare confinement potential V_{ext} is of parabolic shape in a very good approximation. Taking the generalized Kohn theorem [4] into account, the FIR conductivity exhibits a single dimensional resonance at a frequency that corresponds to the characteristic frequency of the bare potential. The meaning of "bare potential" here is, the potential is induced by all external charges contributing to the confinement as e.g. charges on a gate pattern, in surface states or image charges on the semiconductor interfaces, but not the electrons occupying the quantum wire. Experiments [5] as well as model calculations demonstrate that the electron systems usually are smaller than the lithographic defined patterns on the surface due to lateral depletion widths of the order of several some 10 nm. Therefore, once the wires in a superlattice are defined, they can be regarded as decoupled in a very good approximation since wide barriers separate them.

In the present paper, FIR transmission and photo conductivity measurements are used to investigate the 1D subband spacings of low dimensional structures fabricated on GaAs-AlGaAs heterostructures and quantum wells.

2. Basic Properties of quantum wires

Much insight into the electronic properties of one-dimensional electronic systems is gained from investigations of their far infrared radiation (FIR) excitations.The heterostructure samples show a 2D carrier concentration of $4x10^{11}$ cm^{-2} with a mobility of 1.5 10^6 cm^2/Vs at 4.2K. Then, laser holographic gratings having a typically period a between 475 nm and 630 nm, were fabricated on the samples. The geometrical dimensions of the structured sample are 3 x 3 mm^2. The photoresist patterns are transferred step by step into the GaAs by wet chemical etching down into the modulation doping area in the case of the heterostructures. This results in a periodic potential with a modulation energy in the range 1 - 3 meV for the heterostructure samples. In these samples, the transverse conductivity vanished and separated wires with several 1D subbands occupied are expected (quasi 1-dimensional electron gas). By bandgap illumination with a red LED the filling of the wires could be tuned. In the region of shallow etching, the electronic system could be changed from electrically isolated wires to a density modulated system and back to an almost perfect two dimensional system with increasing duration of illumination. A schematic view of the samples is shown in the inset of Fig.1. The patterning of the samples produces a periodic electrostatic potential and leads to a lateral type II superlattice. Electrons and holes are separated in spatially adjacent regions.

The experimental setup is the standard configuration of a commercial FIR gas laser, connected via a light pipe to a helium cryostat with two superconducting solenoids. In the centre of the upper magnet the sample and in the lower magnet FIR-detectors for transmission measurements were located. For the detection of the chopped FIR radiation narrowband InSb and high purity GaAs layers were used. The photoconductivity has been measured along the wires, therefore two indium contacts were alloyed into the wire array and a current of typically some 50 nA was passed along the wires. Together with the chopper signal, the voltage drop at the contacts of the sample, or in case of transmission measurements, at the detector was used for low noise lock-in-technique. All measurements were performed at 4.2 K. The advantage of the high FIR power of the laser was especially used for the photoconductivity measurement and the PC-behaviour was compared with the transmission data. It can be shown, that the high power excitation of the magnetoplasmon-resonances in the PC is shifted to lower energies with respect to the low power excitation with the Fourier spectrometer.

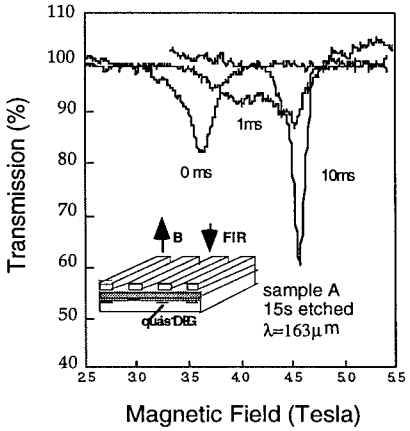


Fig. 1. Transmission of a 630nm period quantum wire sample for different durations of the above bandgap illumination. The inset shows the geometry of the sample.

In figure 1 the experimental FIR spectra of a one layered shallow etched (15s etched) quantum wire structure of a heterostructure sample measured at a fixed laser wavelength of 163.6μm in transmission geometry with different band gap illumination times are shown. The resonance position of the non illuminated quantum wire structure is shifted to lower magnetic fields with respect to the strong illuminated sample, which appears at the position of the unperturbated cyclotron resonance, i.e. it is shifted to higher resonance energies.

The spectrum of the short illuminated sample shows one resonance peak at the CR position and one at a towards higher energies shifted position. In case of electrically separated wires, the FIR excitation should measure the confined plasmon mode [6]:

$$\omega_p^2 = \frac{e^2}{2\varepsilon\varepsilon_0 m^*} n_{2D}(j+\alpha)\frac{\pi}{w}, \qquad j = 1,3,5 \qquad (1)$$

with the effective two dimensional carrier density in the wires n_{2D}, the effective mass of the carriers m* and the electrical wire width w and α, which denotes a correction for the

phase shift occurring due to the reflection of the electrons at the wire walls. In the FIR experiments we clearly identify the plasmon shifted CR with the first odd mode j=1.

In a density modulated system, one expects to find two resonances [7], the unperturbated CR and an extended plasmon, which is a better description, when coupling between the electrons in the different parallel wires becomes important. Then ω_p is dominated by the grating period a [8]:

$$\omega_p^2 = \frac{e^2}{2\varepsilon\varepsilon_0 m^*} n_{2D} \frac{2\pi}{a} \qquad (2)$$

By increasing the illumination time, the potential modulation disappears at all, and one finds the pure two dimensional behaviour of the system. By subsequently etching the sample, the resonance position shifts to higher energies, and is not affected anymore by above bandgap illumination. Due to surface charge depletion effects, the effective electron concentration decreases rapidly and therefore the amplitude of the resonance decreases also.

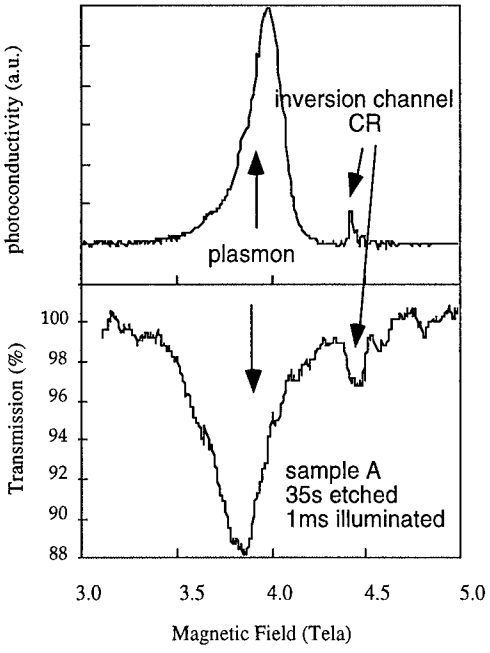


Fig. 2 Typical spectra obtained with the FIR-laser in photoconductivity and transmission measurements of sample A.

In Fig.2 a FIR photoconductivity spectrum is shown taken at a laser wavelength of 163μm. Comparing the transmission data with the data obtained from the photoconductivity measurements performed on the same sample, one clearly sees, that the resonances appear at the position of the confined plasmon. The plasmon energy is easily determined from a plot like fig.3. by extrapolating the point of intersection of the line of the plasmon energies with the y-axis.

In addition, there is a small, but very sharp resonance (FWHM = 0.2 cm^{-1}, an effective mass of 0.067 m_0 and an effective 2D electron density of $4 \cdot 10^{10}$ cm^{-2}) near the position of the unperturbed 2D CR. It can also be seen in non-structured samples and its position is not affected by the etching process of the wires. So we expect the origin of this resonance is due to an inversion channel deep in the sample, as well known from similar grown samples [9].

The spectrum taken with the FIR laser is in good agreement with the bolometric model of the PC [10], [11], where the PC grows both with the absolute resistance of the system and its temperature dependence. The evaluation of the resonance energies for PC and transmission measurements of the sample A with a=630 nm and an etching time of 35s are given in a double squared plot (fig.3.) With good agreement between the different methods, the plasmon resonance follows the quadratic dispersion relation

$$\omega^2 = \omega_p^2 + \omega_c^2 \qquad (3)$$

which is obtained for the collective plasmon excitation of the electrons in an magnetic field coupling quadratically to the cyclotron resonance energy ω_c. Assuming an harmonic oscillator model for the confining potential

$$V_{ext} = \frac{1}{2} m^* \omega_0^2 x^2 \qquad (4)$$

the same relation for the dispersion of the 1D resonance energy as eq.(3) can be obtained. For the non illuminated sample A (35s etched) we derive a subband spacing $\omega_0 = 3$ meV and an electrical width w = 380 nm (dark) and w = 510 nm (illuminated), calculated from eq. 1 with $\varepsilon = 12.9$. Differences in the resonance position between transmission and PC values are attributed to changes in the population of the wires, which may have different causes: heating of the electron system by the high FIR intensity as well as heating and charging of the wires by the current.

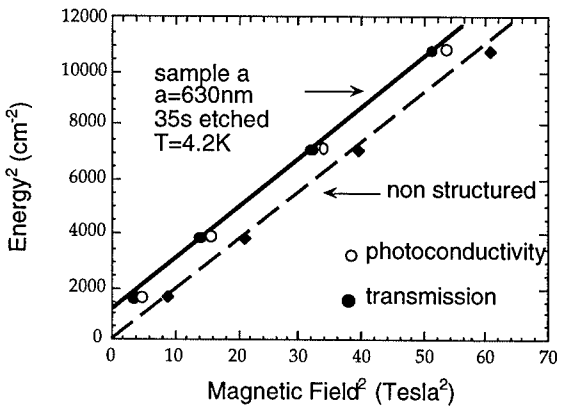


Fig.3. Squared energy of the evaluated resonance position (transmission and photoconductivity results) versus squared magnetic field (sample AThe dotted line corresponds to the unperturbed 2D CR with a mass of $m^* = 0.07 m_0$

It is shown, that via above band gap illumination the electronic properties of the shallow etched wire structured sample could be changed from quasi 1D to a density modulated 2D system and finally to a pure 2D system. The PC is strongly correlated to the position of the localized plasmon in FIR transmission. Recent PC measurements [11] were successfully performed on a tiny sample of 80 x 200 μm^2 with 130 wires. The big advantage of the PC measurements is, that there is no need for a large structured area as in transmission experiments. One disadvantage of this quantum wire array (some

hundreds) is till now the very bad detectivity, but combining holographic and electron beam lithography, it is possible to prepare photoconductive detectors consisting of very few wires with much more improved properties.

3. Smith-Purcell type FIR radiation in one dimensional wires

The coupling of the carriers via a grating momentum $q = 2\pi/a$ to a weak harmonic potential results in energy loss via photon emission. From conservation rules one obtains for $q << k$ (k = electron momentum) the frequency ω of this Smith-Purcell emission [12] which is directly proportional to the electron velocity v and the grating momentum q and given by $\omega = v \cdot q$.

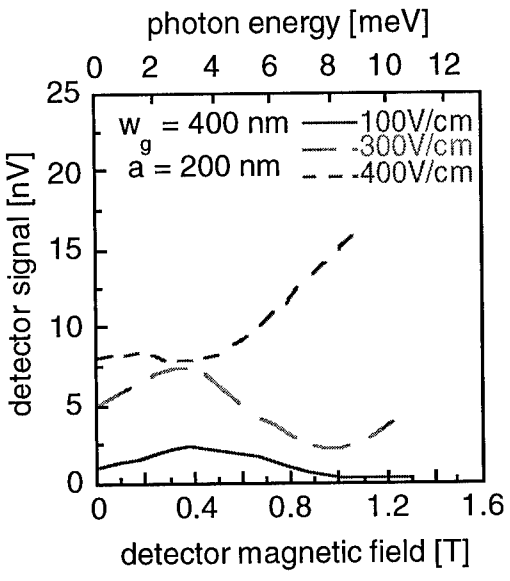


Fig. 4: Experimental Smith-Purcell emission signals in a 1DEG for different electric fields F = 100 V/cm (solid line), F = 300 V/cm (dashed line) and F = 400 V/cm (dotted line).

Observation of Smith-Purcell emission in a two dimensional electron gas of high mobility samples has been reported by Wirner et al. [13]. It should also be possible to observe Smith-Purcell emission in one dimensional structures. However, the mobility in the 1DEG is drastically reduced in comparison to the 2DEG. On the other hand the mobility increases strongly with decreasing wire length. This occurs for wire lengths smaller than 2μm which indicates that the interactions with the walls and potential fluctuations are strongly reduced below this length scale. For this reason we also expect for longer wires a mean free path which is large enough for the electrons to couple to a grating with a period length well below this length scale. The samples used are described in detail in reference [13]. The original 2DEG has an electron concentration of $n = 6 \times 10^{10}$ cm^{-2} and a mobility of $\mu = 8 \times 10^5$ cm^2/Vs.

A weak periodic grating potential of $V_0 \approx 0.6$ meV with different period lengths of a = 200 nm, a = 300 nm, and a = 400 nm is produced by nanostructuring the GaAs top layer with laser holography and shallow wet chemical etching. The one dimensional wires are produced by an additional etching process. The geometrical wire width is $w_g = 400$ nm whereas the wirelength is 100 μm. Analyzing the Shubnikov-deHaas oscillations (assuming a parabolic wire potential) we get a one dimensional

subband spacing of $\Delta E = 0.5$ meV. The one dimensional electron density is determined to be $n_{1D} = 20\times10^6$ cm^{-1}. The emitted Smith-Purcell type FIR radiation is detected by an optimized magnetic field tunable InSb cyclotron-resonance detector at 4.2 K [14]. The experimental setup is described in detail in reference [15].

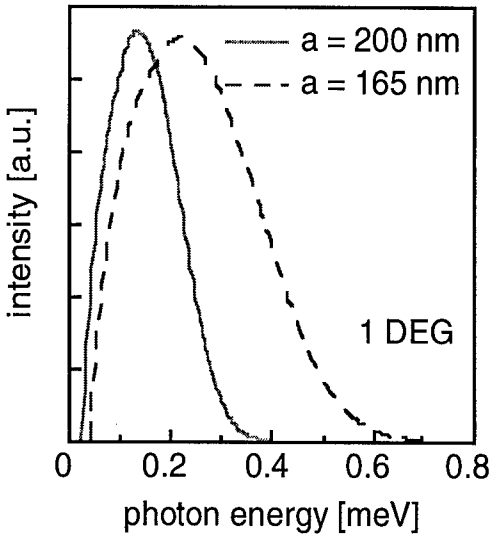


Fig. 5: Theoretical Smith-Purcell emission in the 1DEG including the detector characteristic for two different grating periods a = 165 nm and a = 200 nm. The underlying distribution function is a shifted Fermi-distribution for T = 10 K and a drift velocity $v_D = 5\times10^6$ cm/s (determined experimentally).

The experimental results (Fig. 4) clearly confirm our assumptions. In analogy to the two dimensional case the FIR-emission shows two contributions: a) free carrier emission (i.e. the background signal) and b) the Smith-Purcell-emission sitting on this background. Due to the reduced emitting power in the 1DEG and increased contact effects the background is larger compared to the 2DEG. As a consequence of the reduced mobility the applied electric field has to be higher to observe Smith-Purcell-emission. The spectral width decreases with increasing electric field indicating a longer mean free path of the electrons for higher electric fields. However for electric fields exceeding 400 V/cm the background emission becomes dominant. Our interpretation in terms of Smith-Purcell-emission is confirmed by the study in ref. [16], which shows that for our configuration the emission is clearly no plasmon emission.

4. TUNNELING SPECTROSCOPY OF 0D STATES

The samples for 0-2D tunneling consist of a nominally undoped GaAs layer $\left(N_A < 1\times10^{15}\,\mathrm{cm}^{-3}\right)$ grown on a semi-insulating substrate, followed by an $Al_xGa_{1-x}As$ barrier of a total thickness of 200 Å [50 Å spacer, 50 Å doped $\left(N_D = 3\times10^{18}\,\mathrm{cm}^{-3}\right)$, 100 Å spacer; x=0.36], and a *n*–doped GaAs layer [d= 800 Å, $N_D = 1.2\times10^{15}\,\mathrm{cm}^{-3}$]. An additional GaAs cap layer was highly *n*–doped [d= 150 Å, $N_D = 6.4\times10^{18}\,\mathrm{cm}^{-3}$]. This structure provides two 2D systems separated by a barrier of only 200 Å. From Shubnikov-de Haas measurements it was deduced that in both

2DEG systems only one subband is occupied having electron concentrations of $n_s^{inv} = 6.0 \times 10^{11}\,cm^{-2}$ and $n_s^{acc} = 5.5 \times 10^{11}\,cm^{-2}$, respectively.

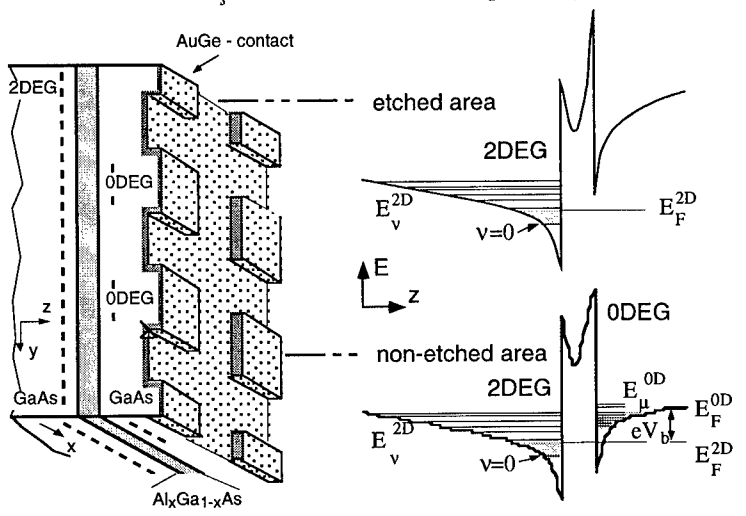


Fig.6 : A schematic view of the sample is shown on the left hand side. The corresponding conduction band profile for the etched and non etched areas of the samples is shown on the right.

The sample geometry is shown in Fig.6(a). First, bar shaped mesas were etched and Ohmic AuGe contacts were aligned to both channels. Then, a holographic photoresist dot patterns with a period of a=350nm was fabricated on the mesas, which were etched wet chemically into the GaAs cap layers. AuGe was evaporated on the total area of the quantum dot system. Finally, the GaAs layers around the top tunneling contact were removed selectively, yielding independent contacts to both the multiple quantum dot (MQD) system and the 2DEG system. By applying a voltage V_b, the 0D states are shifted energetically by $\Delta E = e\Delta V_b$ with respect to 2DEG [17] The bandstructure of the sampleis shown in Fig.6(b) for both the etched and the non–etched regions (upper and lower part, respectively). All measurements were made using a 4-terminal conductance bridge [18]

The experimental results are plotted in Fig.7. The lower part (a) of this figure shows the dI/dV_b characteristics of the nano-structured (0D-2D) sample, where the temperature is varied between T=1.7K [curve (1)] and T=40K [curve (12)]. For reference reasons, the dI/dV_b characteristics of a not-nanostructured sample are plotted in the upper part (b) of Fig. 7 for the two temperature values T=1.7K [curve (1)] and T=40K [curve (2)]. The comparison of the two characteristics [a(1)] and [b(1)] shows that the nano-fabrication process leads to a multitude of new resonances which exist within the whole

voltage range considered. All resonance peaks of the (0D–2D) dI/dV_b characteristic show a strong dependence on the temperature. Above T=4.2K [curve a(2)], only about half of the resonances can still be resolved. A further increase of the temperature results in a monotone broadening of all resonance structures, accompanied by a monotone decrease of peak amplitudes.

The tunneling probability for transitions between the subbands of the 2DEG system and the fully quantized states of the MQD system are calculated in analogy to the the selection rules for 1D–2D tunneling processes [19] Details are published elsewhere [20] To compare the calculated results with the experiment, the overlap integral between the wave functions on the 2D and 0D side were calculated as a function of bias voltage. The overlap integral is directly proportional to the tunneling current. A cosine shaped potential best describes the experimental situation in the dots and the main peaks in the wave tunneling probability are always expected when $w=i\pi/k$ where i is an integer, k is the wave vector of the tunneling electron and w is the width of the potential at the coresponding energy. Note that the overlap integral can be regarded as a Fourier transform of the wave functions.

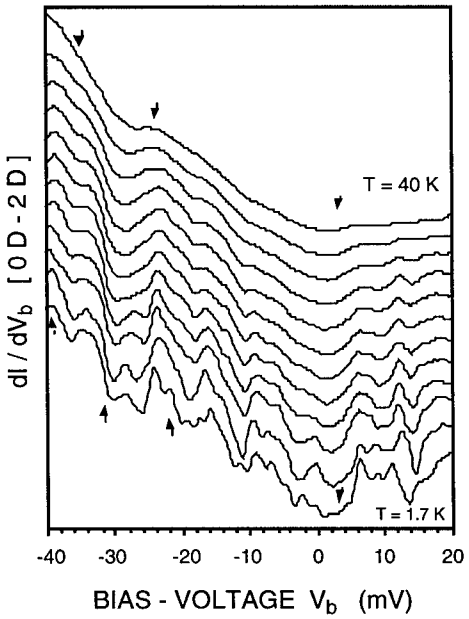


Fig.7 : (a) Measured dI/dV_b curves in the temperature range between 1.7 K and 40 K. (b) : dI/dV_b curves of unstructured samples

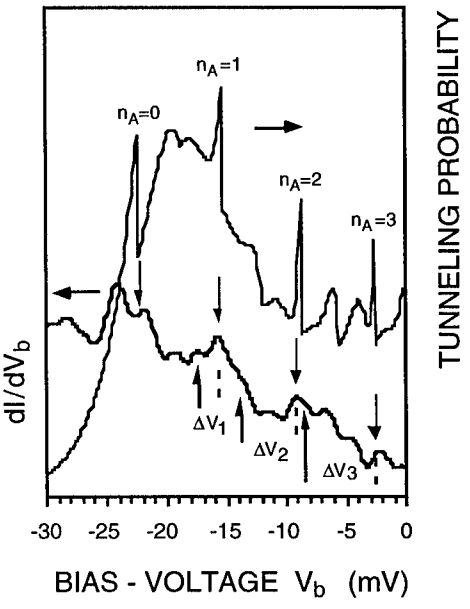


Fig.8: Calculated tunneling probability and the measured dI/dV_b curve at 1.7 K. The downward arrows mark the resonance peaks. The upward arrows mark the k=0 positions for each subband.

Fig.8 shows a comparison between the calculated tunneling probability and the dI/dVb curve measured at 1.7 K. As one can see, the experimental results agree well with the calculated peak positions. The lowest three 0D subbands are determined as $\Delta E01 \approx 7$ meV, $\Delta E12 \approx 6$meV,$\Delta E23 \approx 5$ meV. The relation between the width of the potential, w, and the wave vector, k , w=ip /k, be used to check the order of magnitude of w for the lowest 0D subbands. The upward arrows mark the position, where k=0 for all 0D subbands. and thus, the voltage spacing DV can be used to determine k using $e\Delta V=\Delta E=h2k2/2m^*$ which then yields a value of w=94 nm, 102 nm, and 120 nm for the nA=1,2,3 subbands. These values are in good agreement with the geometrical diameter of the dot of dgeo≈160 nm if the surface depletion is taken into account.

5. Conclusion

Two different FIR spectroscopy techniques on quantum wires have been presented. It is shown, that the electronic properties of the shallow etched wire structured sample can be changed by above band gap illumination from quasi 1D to a density modulated 2D system and finally to a pure 2D system. The photoconductivity is strongly correlated to the position of the localized plasmon in FIR transmission and has a possible application for detectors. The quantum well samples with the shallow etched wire structures show a strong magnetoplasmon resonance with an confinement potential up to $\hbar\omega_0$=9.3meV and effective electrical wire width down to 120nm.

We have studied Smith-Purcell-emission from modulated high mobility GaAs/AlGaAs heterostructures. The results of Smith-Purcell emission in the 1DEG show a smaller spectral width than in the 2DEG indicating a narrower electron distribution in k-space.

Tunneling experiments between 2D and 0D states can be used to determine the subband spacings. In contrast to 2D-2D tunneling, the IV-curves turn out to be the Fourier transform of the wave funktions in the 0D states, which can be used to extract information on the confining potential.

Acknowledgment - This work was supported by the Öesterreichische Nationalbank (Jubiläumsfond), BMWF and the GMe, Austria.

References

1. For a review see W.Hansen, J.P.Kotthaus, U.Merkt, *Semicond. and Semimetals*, **35,** 279 (1990).
2. T.Demel, D.Heitmann, P.Grambow, and K.Ploog, *Superlattices and Microstructures* **9, 285** (1991).

3. W.Hansen, in Proc. NATO ASI *"Quantum Coherence in Mesoscopic Systems"*, ed. B.Kramer, Series B: Vol.**254,** (Plenum Press New York 1991), p.23.
4. S.K. Yip; *Phys. Rev.***B43**, 1707 (1991) and references therein.
5. R.Strenz, V.Roßkopf, F.Hirler, G.Abstreiter, G.Böhm, G.Tränkle and G.Weimann, *Sem. Sci. Techn.* **79**, 399 (1994).
6. W.L.Schaich and A.H.MacDonald, *Solid State Commun.* **83** (1992) 779.
7. T.Demel, D.Heitmann, P.Grambow and K.Ploog, *Phys. Rev.* **B38,** 12732, (1988).
8. D.Heitmann and U.Mackens, *Superlatt. Microstruc.* **4**, 503 (1988).
9. M.Besson, E.Gornik, G.Böhm, G.Weimann, *Surf. Sci.* **263,** 650 (1992).
10. F.Thiele, E.Batke, J.P.Kotthaus, V.Dolgopolev, G.Gusov, G.Weimann, W.Schlapp, *Solid-State Electron.* **32** (1989) 1503.
11. C.M.Engelhardt, V.Roßkopf, E.Gornik, M.Aschauer, R.Strenz, G.Böhm, G.Weimann, *Proceedings of the 11th Int. Conf. on High Magn. Fields in Semicond. Physics, Cambridge, MA, USA, August 8-12, 1994.*
12. S.J. Smith and E.M. Purcell, Phys. Rev. **92**, 1069, (1953).
13. C. Wirner, C. Kiener, W. Boxleitner, M. Witzany, E. Gornik, P. Vogl, G. Böhm, and G. Weimann, Phys.Rev.Lett. **70**, 2609 (1993).
14. G. Strasser, K. Bochter, M. Witzany, and E. Gornik, Infrared Phys. **32**, 439 (1991).
15. E. Gornik, W. Müller, and F. Kohl, IEEE Trans. Microwave **22**, 12 (1974).
16. R. Strenz, V. Roßkopf, F. Hirler, G. Abstreiter, G. Böhm, G. Tränkle and G. Weimann, Sem. Sci. Techn. **79**, 399 (1994).
17. W. Demmerle, J. Smoliner, G. Berthold, E. Gornik, G. Weimann, and W. Schlapp, Phys. Rev. B **44**, 3090 (1991).
18. R. Christanell and J. Smoliner, Rev. Sci. Instrum. **59**, 1290 (1988).
19. W.Demmerle, J Smoliner, E.Gornik, G.Böhm, G.Weimann, Phys. Rev. B**47**, 13574 (1993).
20. Proc. 8.th Int. Winterschool, Mauterndorf 1994, to be published in Semicond. Sci. Technol. (1994).

NONLITHOGRAPHIC FABRICATION AND PHYSICS OF NANOWIRE AND NANODOT ARRAY DEVICES - PRESENT AND FUTURE

A. A.TAGER[1], D. ROUTKEVITCH[2], J. HARUYAMA[1],
D. ALMAWLAWI[2], L. RYAN[2], M. MOSKOVITS[2] and J. M. XU[1]

University of Toronto,
[1]Dept. of Electrical & Computer Engineering, [2]Dept. of Chemistry and
[1,2]Laser & Lightwave Research Centre; Toronto, M5S 1A4 CANADA

1. Introduction

It has been the "proven truth" for some decades that the performance of electronic devices should rise with decreasing dimensions. The continuos advance of lithographic technology in the past made possible serious exploration of the fascinating world of low-dimensional structures and was in turn propelled by the results of this exploration. Today, the submicron lithography in combination with different growth techniques such as MBE and MOCVD is routinely used for fabrication of quantum-well based devices with superior characteristics.

The situation is changing however as researchers move toward one- and zero-dimensional structures. Although there have been a number of successful efforts, e.g. fabrication of free-standing Si or Se nano-wells or nano-pillars by the less controllable thinning techniques after conventional lithography or e-beam lithography, the now celebrity lithographic technology seems to approach its resolution limits here, making the cost of controllable manufacturing of high-density arrays of 1-D and 0-D structures almost prohibitively high for industrial implementation as the dimensions of individual units fall to 10 - 100 nm.

The situation has forced us to look for new approaches for the nano-manufacturing. One way is to use different methods of direct writing of desirable structures on semiconductor or dielectric wafers, e.g., the focused ion beam technology or the atomic force microscope (AFM) and perhaps STM. The method looks promising for relatively small structures, but becomes unrealistic when a large area or a large number of devices are needed. One alternative is to make use of spontaneous island formation in planar growth of some lattice-mismatched semiconductor systems, that gives more or less disordered 2-D arrays of nano-sized clusters [1]. It is not yet clear however whether the usefulness and cost efficiency of this method are sufficiently high to outweigh the. considerable sacrifice in the ability to control the structure parameters in this case.

S. Luryi et al. (eds.), Future Trends in Microelectronics, 171–183.

Another very different approach is non-subtractive direct fabrication by selective deposition in the openings of some sort of *template*. This of course could be a regular lithographic structure, with all the above mentioned limitations.

On the other hand, one can use different self-organised regular nano-structured (nano-porous) templates, such as zeolites and molecular sieves [2], polymer nuclear trucks membranes [3] or porous anodic aluminum oxide (AAO) films [4]. This is a particularly attractive approach as it circumvents the lithographic limitations altogether by making use of naturally occurring nanotemplates (working with Mother Nature!).

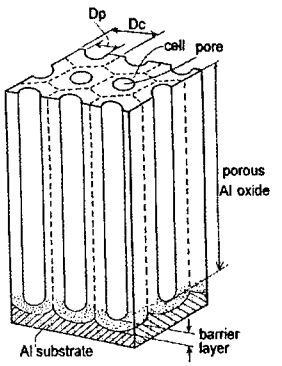


Figure 1. Idealised cross-section of the porous anodic aluminum oxide.

In making a choice of template one must take into account its stability, insulating properties, minimal diameter, density and uniformity of the pores as well as the ability to integrate the template into a device or chip. The latter include the possibility to form a template film of good quality on desired substrate and to make electrical or optical contacts to the nano-structures inside the template. The most attractive candidate from this point of view is AAO which has a highly anisotropic porous structure (Fig. 1) with very uniform and nearly parallel pores [5]. We are able to produce AAO with pore diameter, d_p, ranging from 8 to 200 nm, pore length from 1 to 50 μm and pore density in the range 10^9-10^{11} cm^{-2}. All these make it almost ideal for fabrication of fairly monodispersed nanometer-scale wires and dots. The smallest pore diameter is of the order of the Bohr diameters of some bulk semiconductor excitons, suggesting that quantum size effects might be observed. Moreover, the use of AAO templates readily allows scale-up to very large surface area systems and is amenable to continuous production.

Providing that AAO is formed on conductive substrate, electrochemical (EC) methods, which can be used to grow materials of a wide range of morphologies, structures and composition, seem to be the logical choice for selective deposition into the pores. Electrodeposition (ED) has previously been used successfully for the formation of ceramic [6], metallic [7] and semiconductor [8] superlattices. Attempts were also made to use sequential underpotential electrodeposition of monolayers to grow epitaxial compound semiconductor films [9].

Electrodeposition of metal nano-wire arrays in porous aluminum oxide has been used to produce deposits with interesting magnetic [10,11], catalytic [12] and optical [13] properties, at times differing from those of the bulk metals. The wires faithfully reproduce the shape of the pores [14]. Other porous films such as polymeric membranes, have also been used for the ED of ferromagnetic Ni and Co [15] as well as for multilayered Cu-Co [16] nano-wires.

Below we will first discuss our results in what we believe to be the most promising of the existing direct fabrication methods - the electrochemical deposition of metals and semiconductors into porous AAO films, which produces densely-packed uniform

2-D nanowire arrays with diameters down to 8 nm or even less [4]. In sections 2 and 3 we will address the technology of the array fabrication, and present experimental results on some effects ascribable to anisotropy, electron confinement and room-temperature single electron tunnelling (SET) in nano-wire based devices. In section 4 we present theoretical results for the effect of interwire coupling on SET characteristics in such systems. Finally, we will give a preliminary and hence incomplete account of the interesting possibilities in device applications and research which the new method and structures promise.

2. Fabrication and Characterisation of the Nanowire Structures

The preparation procedure for porous AAO films with different pore diameters is described in [4]. AC electrolysis was used for the deposition of metals (Ni, Co, Fe and others) and compound semiconductors (CdS, CdSe, CdS_xSe_{1-x} and GaAs) resulting in the formation of nano-wire arrays in the pores of anodic alumina films. Deposition of other semiconductors is the subject of our current efforts. Electrodeposition was performed at room temperature for metals and GaAs; and at $100 - 160^oC$ for the $A^{II}B^{VI}$ systems. Metals and GaAs deposition was carried out from aqueous acidic electrolytes, whereas, $A^{II}B^{VI}$ were deposited from solutions containing metal salt and elemental chalcogen in dimethylsulphoxide [17]. Although different in execution, the technique is similar in scope to the one used to deposit 200 nm diameter cadmium selenide/telluride microwires in the pores of Anopore membranes [18]. However, the large pore diameter of most available Anopore films precludes observation of low-dimensional effects.

Take the fabrication of CdS nano-wires as an example. Electron microphotographs (fig. 2) show that the deposited material fill the pores uniformly. Quantitative electron microprobe analysis indicates a 1:1 stoichiometry to within the detection tolerance of the technique. XRD data of the AAO\CdS nano-wires have diffraction patterns corresponding to the hexagonal. The unit cell of AAO\CdS was found to have smaller parameters and greater density than what is reported for the bulk. Annealing at 500^oC for 1 hour successfully relieved the lattice distortion except for samples with $d_p \leq 12$ nm.

The X-ray diffraction data suggest that the spatial confinement of the deposition affects the crystallite orientation. The relative intensity of the (002) diffraction peak of the AAO\CdS is considerably higher than those recorded for the Pt\CdS (DC deposited) sample or a crystalline "bulk-like" CdS powder therefore indicating that crystallites growth with the c-axis oriented along the pore. Annealing only slightly enhances the diffraction patterns of AAO\CdS, preserving its existing (002) texture. The contributions of other primary diffraction planes ((100), (101), (110)) are suppressed significantly. Yet another template-induced structural feature of the AAO\CdS is that crystal domains size changed very little after annealing, suggesting that most of the deposited CdS is crystalline and the CdS domains span the widths of the pores. We are tempted to conclude that the main factor determining the observed

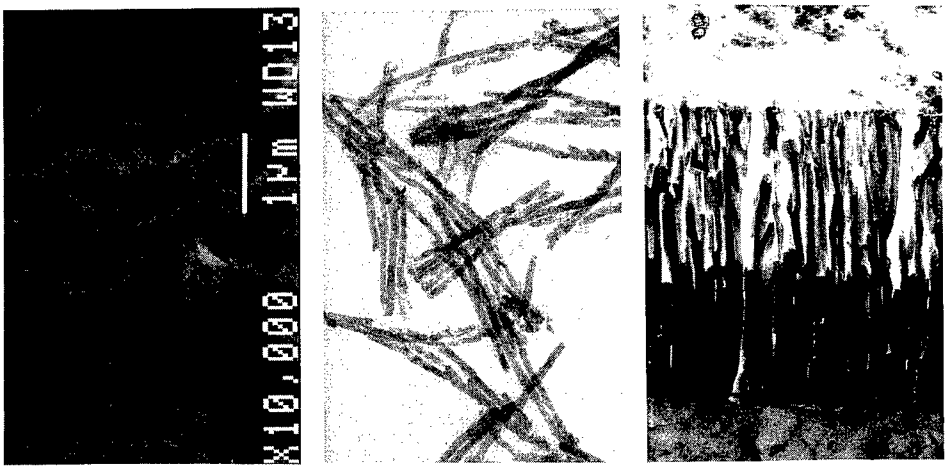


Figure 2. (a.) SEM micrographs of CdS nano-wires partially exposed by dissolution of the oxide film; **(b.)** TEM micrographs of the CdS nano-wires after complete dissolution of AAO supporting matrix; **(c.)** TEM micrograph of a microtomed cross section of the Al\AAO\CdS film.

structural features of the AAO\CdS, and especially its texture, is the spatial confinement inside the pores.

3. Measurement Results on the Metal and Semiconductor Nano-wire Arrays

3.1. SIZE EFFECT IN CdS NANOWIRES

Polarised resonance Raman spectroscopy (RRS) was used to investigate qualitatively the effect of the nano-wire diameter on the bandgap energy of the CdS quantum wires. The Bohr diameter of the 1S exciton of CdS (6 nm, [19]) is close to the smallest AAO pore diameter that we can produce. Accordingly, some indication of quantum confinement effects might be observable.

The measured RRS spectra are dominated by overtones of the LO phonon. The overall intensity of the Raman scattering and the number of overtones observed increased dramatically with annealing, making quantitative spectra analysis possible [20]. This behaviour is probably due to the fact that the intensities of the RRS lines are very sensitive to any improvement in the crystalinity of the wires. Furthermore, Raman scattering might also benefit from the annealing of surface defects at the CdS\AAO interface or at grain boundaries.

The intensities of the overtones in the Raman spectra of semiconductor clusters depend strongly on particle size [20,21] and are related to the value of the band gap. Polarised RRS of AAO\CdS nano-wires excited with the electric field vector both along and across the wires showed that the ratio of the first overtone intensity to that of the

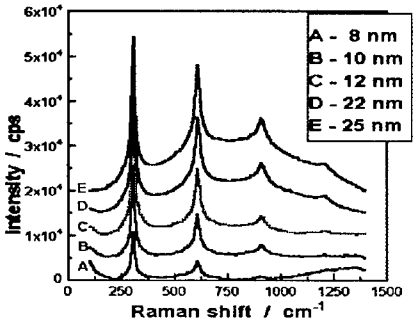


Figure 3. Resonance Raman spectra of AAO\CdS nanowires arrays with light polarized across the wires

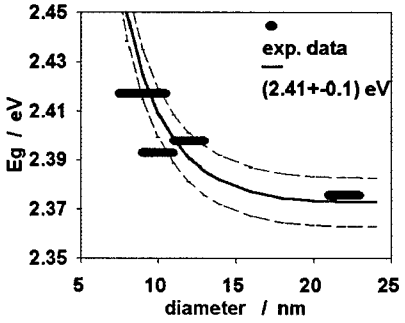


Figure 4. Band gap as a function of CdS nano-wires diameter. Symbols - experiment, line - calculation.

fundamental mode decreases with decreasing wire diameter, suggesting that the quantum size effect regime has been approached (Fig. 3). Based on this dependence we calculated band gap energies and other parameters by fitting experimental data to the expression which describes the ratios of the overtone intensities as a function of excitation energy [20,21].

The band gaps determined with polarisation across the wires ranged from 2.376 eV for the large pore diameters to 2.417 eV for the smallest pore diameter and are well behaved as a function of the nano-wire diameter (Fig. 4) determined by the expression for *spherical particles* [22] when bulk values are used for the effective masses of the electron and hole, and the optical dielectric constant; 2.41 eV is assumed for the band-gap energy of bulk CdS. The calculated overtone intensity ratios values for ss-polarisation are consistent with the experimental data (Fig. 4).

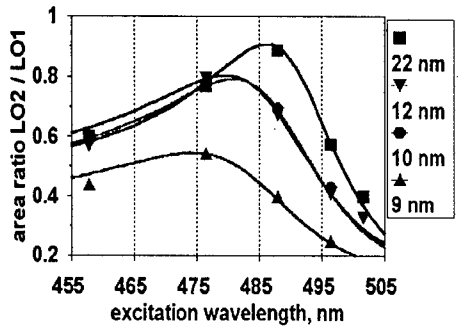


Figure 5. RRS overtone ratio as a function of excitation wavelength for different CdS wire diameter. Symbols - experiment, lines - calculation.

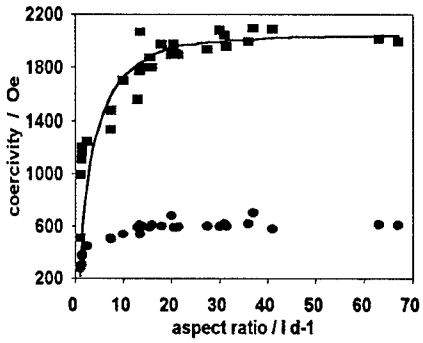


Figure 6. Measured perpendicular and parallel coercivities of AAO\Fe wires as a function of wire aspect ratio. Squares - $H_{c\perp}$, circles - $H_{c\parallel}$.

3.2. MAGNETIC EFFECTS IN NANOWIRES.

Iron nano-wire arrays with varying diameters were fabricated [11] and their magnetic properties determined. The coercivity was found to be highly anisotropic and dependent on the aspect ratio of the particles (Fig. 6). The functional dependence of H_c on the aspect ratio suggested that the metal deposit consists of a cylindrical assembly of fused single-domain particles.

3.3. SINGLE ELECTRON TUNNELLING IN METAL NANOWIRES

Room temperature operation of devices based on single-electron tunnelling (SET) requires extremely small capacitance tunnel junctions, lithographic fabrication of which is difficult and expensive. We fabricated a two-junction device from 10 nm nickel nano-wire array where one junction was the (bottom) barrier oxide layer (~10-20 nm) and the other (top) was a thin (~ 8 nm) nickel oxide film grown on the tips of the nickel nano-wires. The device, which can be considered as a 2-D-array of double-junction systems with common "source" and "drain" contacts [23], at room temperature exhibited very complex behaviour including conductance oscillations and staircase I-V characteristics. Periodic conductance oscillations were observed in similar structures with CdS nanowires (Fig. 7), and recently in «AAO+Ni-wires\CdS film» array devices [24]. Their origin rises many fundamental questions but without any clear explanation in sight.

The I-V stairs of metal wires (an example of which is shown in Fig. 8) on the other side strongly resembled the Coulomb staircase as predicted by the standard theory of single-electron tunnelling (SET) in case of strong asymmetry of drain-source junctions [25,26]. Simple estimation of junction capacitances for a single wire in the array gives values in the order of 10^{-19} to 10^{-18} F [23], that together with different thickness of the oxide layers and the observed stair width seem to support the hypothesis of SET. These arrays however are essentially different from those considered previously [26], where

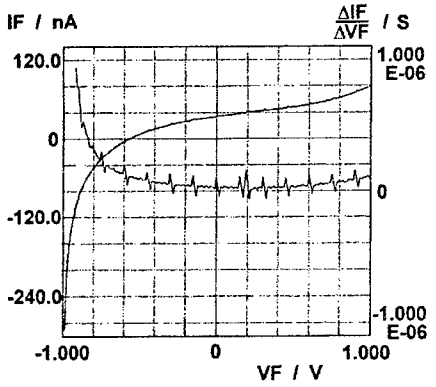


Figure 7. Periodic conductance oscillations of AAO\CdS nanowire array.

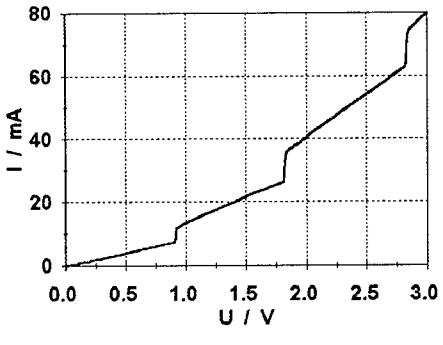


Figure 8. Current-voltage characteristics of the double tunnelling junction Al/AAO-BL/Ni-wires/NiO/Ag.

the tunnelling occurs sequentially in the plane of the array of islands connected in matrix, and coupling capacitances are of the order of the junction capacitances. Adequate understanding of the novel system must take into account a number of unique features and effects including the very strong electrostatic coupling between nanowires due to small interwire spacing (20 to 50 nm) as compared to the wire length (>1 μm in this case).

4. Single Electron Tunnelling through Coupled Nanowires: Spontaneous Polarisation

As a first step toward a good understanding, we investigated possible effects of interwire coupling on the single-electron tunnelling in coupled double-junction systems considering a two-wire system (Fig. 9) which is a building block of the device.

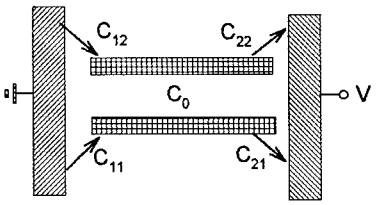


Figure 9. Two-wire double-junction system. Arrows show the prevailing directions of tunneling for V>0.

Using the "global rule" of SET [27] to calculate the effect of wire charging on the tunnelling probabilities, we performed Monte-Carlo simulations of the electron transport through the system assuming a low-impedance environment. Typical numerical values used in modelling were: junction capacitances $C_{ij} = (1.6 \pm .2) \cdot 10^{-19}$ F, coupling capacitance $C_0 = (0 \text{ to } 200) \cdot C_{ij}$, junction resistances $R_{11} = R_{12} = 50\text{k}\Omega$ (source), $R_{21} = 200 \cdot R_{11}$ and $R_{22} = (1 \text{ to } 100) \cdot R_{21}$ (drain). We assumed zero conductance between wires.

4.1 RESULTS

Analysis that uses the "global rule" ideology shows that in the case of very strong coupling ($C_0 >> C_{ij}$) the whole array seems to respond as a single double-junction system wherever the electron actually tunnels, and the relative contribution of the particular wire charge to the system "charging energy" is proportional to a small ratio C_{ij}/C_0. The apparent consequence is the reduction of the Coulomb-blockade region in proportion to the number of participating wires. Attempt to apply this finding directly

to our experimental system is impeded by present lack of knowledge of what specific wires are participating in the charge transport at any given moment, knowing a finite dispersion in oxide thickness and a large dispersion in junction resistances.

The interesting part is that strong coupling also leads to *spontaneous polarisation* of neighbouring wires, when an accumulation of excessive electrons on one wire is partly compensated by the hole accumulation on another one, yielding very low-energetic "excitonic excitation" of the system. The polarisation is conveniently described by the difference in charges on individual wires $P = N_2 - N_1$ (in units of electron charge e). It can considerably exceed the total charge $N = N_1 + N_2$ on both wires, which at low enough temperature remains rather strictly determined by the applied voltage.

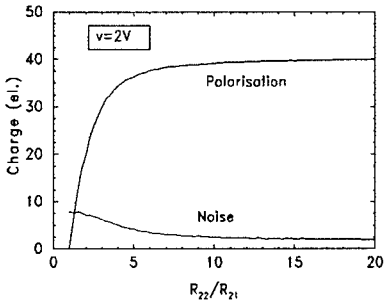


Figure 10. Average polarization and polarization dispersion vs. drain resistance asymmetry; T=0, $C_0/C_{ij} \cong 100$.

For identical wires P stochastically oscillates in time around zero average value $\langle P \rangle$, with increasing with coupling dispersion $D_P = \sqrt{\langle (\Delta P)^2 \rangle}$. The polarisation becomes however quite regular in the case of *asymmetrical wires*, when electrons can tunnel, say, from one of the wires "faster" than from the other. Figure 10 illustrates

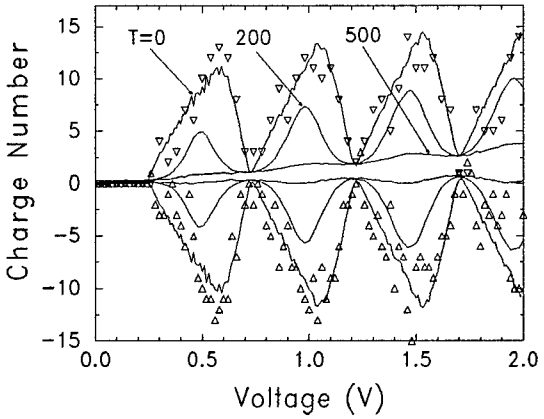


Fig. 11. Average wire charges $\langle N_j \rangle$ vs. voltage at three different temperatures. Top 3 curves represent the "slower" wire, other 3 - the "faster" one. "Δ" and "∇" show the "faster" and "slower" wire charges at a particular time instant for T=0 K

rapid increase of $\langle P \rangle$ with increasing asymmetry of the drain resistances, with a maximum saturation level in the order of $C_0/(C_{11}+C_{21})$ (determined by the "faster" wire).

The effect is voltage- and temperature-sensitive. It practically disappears at the critical voltages corresponding to steps on the *I-V* curves. At these points the maximum total charge on both wires N increases by one. This brings about a new phenomenon which can be called the Coulomb blockade of polarisation. The wire charges, $\langle P \rangle$ and D_P thus periodically oscillate with increasing voltage (Fig. 11).

The strong system polarisation in-between these critical voltages is a SET-induced effect which does not exist for high temperatures when $kT \geq e^2/2C_j$ and the thermal fluctuations control the charge statistics. For higher temperatures the average wire charges approach the Kirchhoff's values at all voltages and vary almost linearly with the voltage (500K- curves in Fig. 11). For lower temperatures the polarisation statistics is governed instead by the shot noise in combination with the Coulomb-blockade effect, which suppresses the fluctuations of the total charge on both wires but increases the anticorrelated fluctuations of the individual wire charges.

This behaviour is illustrated in Fig. 12 which shows how $\langle P \rangle$ and D_P depend on temperature for "peak polarisation" (left graph) and "low polarisation" (right graph) voltages. In the first case the almost constant average polarisation is accompanied by the increasing noise as temperature increases. What is interesting and counter-intuitive is that not only the average polarisation, but also *the polarisation noise* decrease at first as the temperature rises, and only at higher temperatures the polarisation noise start rising again due to the thermal contribution.

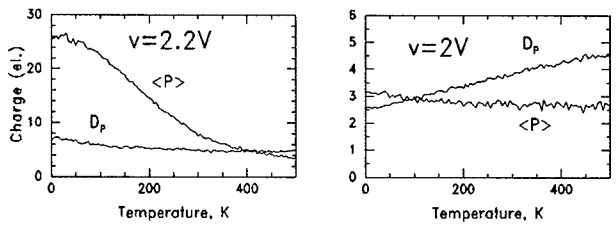


Figure 12. Average polarization and polarization noise vs. temperature for different voltages.

This anomalous temperature dependence of the polarisation fluctuations is even more evident if the noise spectral characteristics are considered (Fig. 13) since the spectral bandwidth of the SET-induced polarisation noise, which at T= 0 K is determined by the characteristic "charging time" of the coupling capacitor $\tau_c \sim RC_0$, is much smaller than that of the pure shot noise. The bandwidth however increases with temperature, and the low-frequency polarisation noise decreases much faster than the polarisation dispersion as the temperature is increasing - in our case more than 20 times when the temperature increases from 0 K to 300 K.

Similarly, relatively low fluctuation bandwidth of the SET-induced polarisation noise makes the difference between the noise intensity of the total charge number N

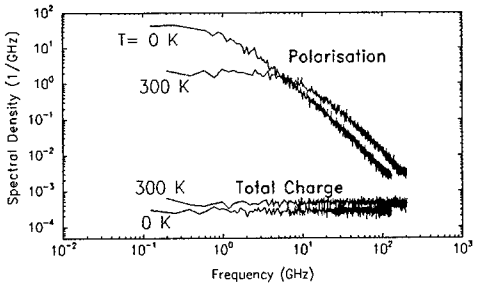


Figure 13. Spectral density of fluctuations for the wire charge polarisation and the total wire charge at two temperatures.

and that of the charge polarisation P much more striking - at low temperatures $\langle \Delta P^2 \rangle(\omega = 0)$ exceeds $\langle \Delta N^2 \rangle(\omega = 0)$ by 5 orders of magnitude.

Thus, under the SET conditions the wire coupling leads to strong interwire polarisation of charges, which for just two wires stochastically oscillates in time and periodically - as function of the applied voltage. In the much more complex case of multi-wire systems and nanowire arrays, this SET-induced spontaneous self-polarisation may lead to a number of (quasi)static and dynamic phenomena. Among others, the prospect of "phase transition" between random polarisation and ordering with changing of system parameters or an applied field, and the possibility of the self-sustained "polarisation waves" are being investigated.

5. Prospects of the Nonlithographic Nanowire and Nanodot Arrays

The above described effect of SET-induced spontaneous polarisation is only one example of all the opportunities which the nano-arrays provide for both research and device engineering. *On the basic research front,* they inspire and provide experimental ground for investigations of different "collective" effects (such as those described above, or that based on magnetic properties of wires) in 2-D or 1-D arrays, as well as studies of intrinsic properties of individual 1-D (or 0-D) objects - e.g. carrier localisation at low-temperatures, magneto-conductance, superconductance in low-dimensional structures etc. It can be accomplished either "on wafer" - e.g. using a conductive AFM tip to locate and electrically access a wire, or by dissolving the template material and working with individual wires directly.

On the front of device development, the high density, uniformity and very small wire diameter of the arrays make them very promising for such hot applications as *flat-panel electroluminescent displays* [28], *magnetic memory* [11,29] and charge storage memory. In the first case, very high electric field near the wire ends which exists even for relatively low voltage between the wire and the contact layer, can lead to very efficient electroluminescence if a thin layer of appropriate semiconductor is placed in-between the wires and the (transparent) contact layer.

For magnetic storage, the nanowire arrays offer numerous advantages - perpendicular magnetic anisotropy hence perpendicular reading/writing, high density, almost quantum nature of writing/reading due to single domain nature of the wires. These structures are also promising for giant magnetoresistance (GMR) sensors with the current perpendicular to the plane (CPP) geometry as an alternative to the current

in plane (CIP) configuration [30]: CPP does not require as thin layers as CIP, control over coercivity field is possible by changing layers thickness and wires diameter, antiparallel alignment can be enforced by depositing magnetic layers with different thicknesses.

In less traditional directions, one can think of using the nanowire (or better yet, nano-dot, when the "wires" are made short enough) arrays as "charge storage" media (in analogy with magnetic storage) when operating in the SET regime. Fabrication of highly-ordered nano-arrays [31] should make easier reading the information (charging states) from individual wires. That can be accomplished, at least in laboratories, by AFM-like writing tips and reading sensors sensitive to electrostatic forces.

Certainly very - if not the most - attractive would be to bring these structures into microelectronics, and possibly their integration with more traditional lithographic technology. The arrays of semiconductor and/or metal nanowires can be used to create 2-D matrixes of either more conventional *p-n*, hetero- or Schottky diodes with "in-wire" junctions [32] which should benefit from their reduced dimensions, or a new type of "point-contact" diodes with quasi-0-D to 3-D junctions between the wire ends and a semiconductor layer deposited on top of the array, with a possibility of either 2-D barrier or "inverse" shell layers in-between. The first attempt to fabricate such kind of metal-semiconductor anti-dot arrays from *Ni* nanowire templates gave structures with rather complex but interesting behaviour including periodic and anomalous conductance oscillations at room temperatures [24]. Other applications might include photodiode arrays of photo-detectors and photo-luminescent diodes, arrays of SET diodes and transistors [32], as well as single- or few-wire devices such as trance-impedance nano-crosses, or based on superconducting wires and Josephson junctions.

What we described above is a very preliminary and incomplete account of the new possibilities which are opened up by this method of nonlithographic fabrication of nano-structured nano-materials and nano-sized elements, mainly in the area of solid-state physics and electronics. Exciting results are expected also from application of these structures in other areas ranging from electrochemistry itself as nano-electrode arrays, to bioencapsulation and drug delivery [3]. We believe that many other ideas of their application both in science an industry will emerge as the fabrication technology, which is still in its childhood now, evolves to give highly-ordered, defect free nanostructures with controllable parameters.

6. References

1. Leon, R., Petroff, P.M., Leonard, D. and Fafard, S. (1995) Spatially resolved visible luminescence of self-assembled semiconductor quantum dots, *Science* **267**, 1966-1968.
2. Ozin, G. A. (1992) Nanochemistry: synthesis in diminishing dimensions, *Adv. Mat.* **4**, 612-648.
3. Martin, C. R. (1994) Nanomaterials: a membrane based synthetic approach, *Science* **266**, 1961-1966.
4. Al-Mawlawi, D., Liu, C. Z. and Moskovits, M. (1994) Nanowires formed in anodic oxide nanotemplates, J. Mater. Res. 9, 1014-1018; Al-Mawlawi, D., Douketis, C., Bigioni T. et al (1995) Electrochemical Fabrication of Metal and Semiconductor Nano-wire Arrays, *Proc. 187 Meet. Electrochem. Soc., May 21-26, Reno, Nevada*, Pennington, in print.

5. O'Sullivan, J.P. and Wood, G. C. (1970) The morphology and mechanism of formation of porous anodic films on aluminium, *Proc. Roy. Soc. Lond. A.* **317**, 511-543.
6. Switzer, J. A., Hung C. J., Breyfogl B. E. et al (1994) Electrodeposited defect chemistry superlattices, *Science* **264**, 1573-1576.
7. Ross, C. A. (1994) Electrodeposited multilayer thin films, *Annu. Rev. Mater. Sci.* **24**, 159-188.
8. Streltsov, E.N. Osipovitch, N.P. Routkevitch, D. L. and Sviridov, V.V. (1995) Electrochemical deposition of compositionally modulated bismuth chalkogenide films, *in preparation.*
9. Gregory, B. W. and Stickney, J. L. (1991) Electrochemical atomic layer epitaxy (ECALE), *J. Electroanal. Chem.* **300**, 543-561.
10. Kawai, S. and Ueda, R. (1975) Magnetic properties of anodic oxide coatings on aluminum containing electrodeposited Co and Co-Ni, *J. Electrochem. Soc.* **122**, 32-36; Shiraki, M., Wakui, Y. and Tsuya, N. (1985) Perpendicular magnetic media by anodic oxidation method and their recording characteristics, *IEEE Trans. Magn.* **21**, 1465-1467;
11. Al-Mawlawi, D., Coombs, N. and Moskovits, M. (1991) Magnetic properties of Fe deposited into anodic aluminum oxide pores as a function of particle size, *J. Appl. Phys.*, **70** 4421-4425; Dunlop, D. J., Xu, S., Ozdemir, O., Al-Mawlawi, D. and Moskovits, M. (1993) Magnetic properties of oriented iron particles as a function of particle size, shape and spacing, *Phys. Earth Planet. Inter.*, **76**, 113.
12. Haruma, M., Kobayashi, T., Sano, H. and Yamada, N. (1987) *Chem. Soc. Jap. Chem. Lett.*, 407; Miller, D. and Moskovits, M. (1989) Separate pathways for the synthesis of oxygenates and hydrocarbons in Fisher-Tropsch reaction, *J. Am. Chem. Soc.*, **111**, 9250.
13. Foulke, D. G. and Stoddard, W. B. (1963) in F. A. Lowenhiem (ed.) *Modern Electroplating,* Wiley, New York, 632; Preston, C. K. and Moskovits, M. (1988) New Technique for the determination of metal particle size in supported metal catalyst, *J. Phys. Chem.*, **92**, 2957-2960, ; Preston, C. K. and Moskovits, M. (1993) Optical characterization of anodic aluminum oxide films containing electrochemically deposited metal particles, *J. Phys. Chem.* **97**, 8495-8503; Saito, M., Kirihara, M., Taniguchi, T. and Miyagi, M. (1989) Micropolarizer made of the anodized alumina film, *Appl. Phys. Lett.*, **55**, 607.
14. Pontifex, G. H., Zhang, P., Wang, Z., Haslett, T. L., Al-Mawlawi, D. and Moskovits, M. (1991) STM imaging of the surface of small metal particles formed in anodic oxide pores, *J. Phys. Chem.* **95**, 9989-9993.
15. Whitney, T. M., Jiang, J. S., Searson, P. C. and Chien, C.L. (1993) Fabrication and magnetic properties of arrays of metallic nanowires, *Science* **261**, 1316;
16. Nagodawithana, K., Searson, P.C., Liu, K. and Chien, C.L. (1995) Processing and properties of electrodeposited Cu-Co multilayered nanowires, *Proc. 187 Meet. Electrochem. Soc., May 21-26, Reno, Nevada*, Pennington, in print.
17. Baranski, A. S. and Fawcett, W. B. (1980) The electrodeposition of metal chalcogenides, *J. Electrochem. Soc.* **127**, 766-767.
18. Klein, J. D., Herrick, R. D., Palmer, II, D., Sailor, M. J., Brumlik, C. J. and Martin, C. R. (1993) Electrochemical fabrication of cadmium chalcogenide microdiode arrays, *Chem. Mater.*, **5**, 902-904.
19. Brus, L. E. (1984) Electron-electron and electron-hole interactions in small semiconductor crystallites: the size dependence of the lowest exited electronic state, *J. Chem. Phys.* **80**, 4403;
20. Routkevitch, D., Ryan, L. and Moskovits, M. (1995) in preparation.
21. Shiang, J. J., Risbud, S.H. and Alivisatos, A.P. (1993) Resonance Raman studies of the ground and lowest electronic excited state in CdS nanocrystals, *J. Chem. Phys.* **98**, 8432.
22. Wang, Y. and Herron, N. (1991) Nanometer-sized semiconductor clusters: materials synthesis, quantum size effects, and photophysical properties, *J. Phys. Chem.* **95**, 525-532; Sweeny, M. and Xu, J.M. (1989) Hole energy levels in zero-dimensional quantum balls, *Solid State Comm.*, **72**, 301.
23. Al-Mawlawi, D., Moskovits, M., Ellis, D., Williams A. and Xu, J.M. (1993) Working with Mother Nature: Nano-wire diode arrays made by novel techniques and functioning at 300K, *Proc. of 1993 Int. Device Research Symposium*, Virginia, USA, 311.
24. Haruyama, J., Moskovits, M., Routkevich, D., Tager, A.A. and Xu, J.M. (1995) Periodic conductance oscillations in novel nano-wire/semiconductor anti-dot arrays, *Submitted to the IEDM'95 Conference.*
25. Mullen, K., Ben-Jacob, E., Jaklevic, R.C. and Schuss, Z. (1988) I-V characteristics of coupled ultrasmall-capacitance normal tunnel junctions, *Phys.Rev.B* **37**, 98-105.

26. Mooij, J.E. and Schön, G. (1992) Single charges in 2-dimensional junction arrays, in H.Grabet and M.H.Devoret (eds.), *Single Charge Tunnelling, Coulomb Blockade Phenomena in Nanostructures*, NATO ASI Ser. B **294**, Plenum Press, New York., pp.275-310.
27. Averin, D.V. and Likharev, K.K. (1991) in B. Altshuler, P. A. Lee, and R. A. Webb (eds.), *Mesoscopic Phenomena in Solids*, Elsevier, Amsterdam, Chapt.6.
28. Misuki, I., Yamamoto, Y., Yoshino, T. and Baba, N. (1987) *J. Met. Finish. Soc. Jap.* **38**, 561.
29. Chou, S.Y., Wei, M.S., Krauss, P.R. and Fisher, P.B. (1994) Single-domain magnetic pillar array of 35 nm diameter and 65 Gbits/in^2 density for ultrahigh density quantum magnetic storage, *J. Appl. Phys.* **76**, 6673-6675.
30. Piraux, L., George, J.M., Despres, J.F., Leroy, C. et al (1994) Giant magnetoresistance in magnetic multilayered nanowires, *Appl. Phys. Lett.* **65**, 2484-2486; Blondel, A., Meier J.P., Doudin, B. and Ansermet, J.-Ph. (1994) Giant magnetoresistance of nanowires of multilayers, *Appl. Phys. Lett.* **65**, 3019-3021.
31. Masuda, H. and Fukuda, K. (1995) Ordered Metal Nanohole Arrays Made by a Two-Step Replication of Honeycomb Structures of Anodic Alumina, *Science* **268**, 1466-1468.
32. Moskovits, M. and Xu, J. M. (1994) Nano-electric devices, *USA Patent application.*

TAMING TUNNELLING EN ROUTE TO MASTERING MESOSCOPICS

M J Kelly and V A Wilkinson,
Department of Physics,
University of Surrey,
Guildford GU2 5XH,
United Kingdom.

Abstract:

A single tunnel barrier in an asymmetric doping environment exhibits some superior properties as a microwave detector diode, when compared with any competing device. The manufacturability of even this simple structure is not yet established. One cannot yet procure semiconductor epitaxial material that routinely meets the tight tolerances on the specification required to produce diodes with a sufficiently small spread of dc I-V and microwave characteristics. Wafer-scale uniformity and wafer-to-wafer reproducibility are key issues, as are the techniques used for qualifying wafers. Proposals for devices that exploit mesoscopic systems will appear all the more realistic if tunnelling devices are established as manufacturable.

1. Introduction:

With miniaturisation as the driver in achieving ever greater performance within mainstream silicon microelectronics, the need to build devices and circuits with nanometre-scale design rules is not too far over the horizon. It is not only a matter of fabricating devices with suitable uniformity and sufficient yield. The service life of circuits based upon such devices can not be any shorter than we accept today, and the cost will have to fall well short of the extrapolation of the exponential growth curve that characterises the cost of new silicon fabrication facilities being built during the 1990s.

The quest for ultra-small devices suitable for computation a decade or two from now may not appear too urgent at the present time of tight R&D budgets. It is fortunate that one prerequisite for the successful application of small structures in the future has applications close at hand. If we are to confine carriers within small volumes, and to control their passage from one volume to another, we are going to have to master tunnelling phenomena. Tunnelling through thin barriers will be a primary source of leakage, but tunnelling may also be one of the required forms of carrier motion when that motion is needed.

In the area of III-V semiconductor devices, tunnel structures occupy an interesting position. Some remarkable figures of merit have been demonstrated over the last decade. The fastest purely electronic form of solid-state oscillator is based on the resonant tunnelling diode comprising two narrow barriers separated by a well (each of the three

S. Luryi et al. (eds.), Future Trends in Microelectronics, 185–195.

layers being typically 2-3nm thick). Fundamental mode operation has been reported at 420GHz in the GaAs/AlAs materials system [1], and 712GHz in the GaSb/InAs system [2]. The output power levels are derisory at these highest frequencies (sub-μW), but a very high efficiency of dc-to-rf power conversion (~50% at 2GHz), combined with useful power levels (>20mW) has been achieved in the InGaAs/AlInAs/GaAs system [3], by adding a ~0.15μm transit region to the anode side of the double barrier diode structure. In the 1990s, this work has been enhanced by the demonstration of (i) quantum barrier varactors as efficient triplers from 70GHz to 210GHz [4], (ii) asymmetric spacer layer tunnel detectors with comparable sensitivity to, but a rather lower sensitivity to ambient temperature than, a Schottky barrier, a planar-doped-barrier or other microwave detector diode [5], and (iii) resonant tunnelling mixers with conversion gain [6]. All these results refer to room temperature operation. In addition to all these two-terminal devices, resonant tunnelling hot electron transistors [7], resonant tunnelling bipolar transistors [8] and other three-terminal tunnel-based structures [9], when incorporated into circuits, result in reduced circuit complexity by exploiting the internal complexity of the device characteristics.

The new device results are one-offs and best efforts. In discussion with potential exploiters, the major hurdle between these research results and the prospect of manufactured systems based on them is the ability or otherwise to reproduce the dc and rf properties of devices that rely for their operation on critical layer thicknesses of tunnel barriers that are specified with sub-nanometre precision. Can the required wafers be reproduced as a matter of routine and with an adequate uniformity? Despite the claims that are made for both the molecular beam epitaxy and metal-organic chemical vapour deposition techniques, the reproducibility issue only once has been addressed in the context of tunnel devices [10]. The planar-doped-barrier diode [11], in production with GEC Plessey Semiconductors in the UK and Hewlett-Packard in the USA has a critical p+ layer whose doping-thickness product is very tightly constrained. If this device can be made (usually by molecular beam epitaxy, but also by metal-organic chemical vapour deposition), so too should the tunnel devices, but there are no publications establishing the manufacturability of heterojunction-based tunnel devices. In this paper, we review some progress in establishing this manufacturability in the context of high-performance microwave components. If we cannot succeed in this exercise, then the prospects for devices based on mesoscopic components are bleak.

2. The Asymmetric Spacer Layer Tunnel (ASPAT) Diode

2.1 DESIGN, SIMULATION, TOLERANCING AND MATERIALS QUALIFICATION

In Figure 1 is sketched the simplest possible heterojunction tunnel device, a single AlAs barrier in an asymmetric doping environment in GaAs, which has useful rectifying characteristics. This structure has been tested as a microwave detector at 9.4GHz and compared directly with a Schottky barrier and a planar-doped-barrier in the same hybrid test circuit. All the conventional figures of merit (e.g. the transfer function - the terminal voltage as a function of input microwave power, the rf parasitic impedances, etc) match the earlier and standard devices (c.f. Figure 1). The attractive feature of this

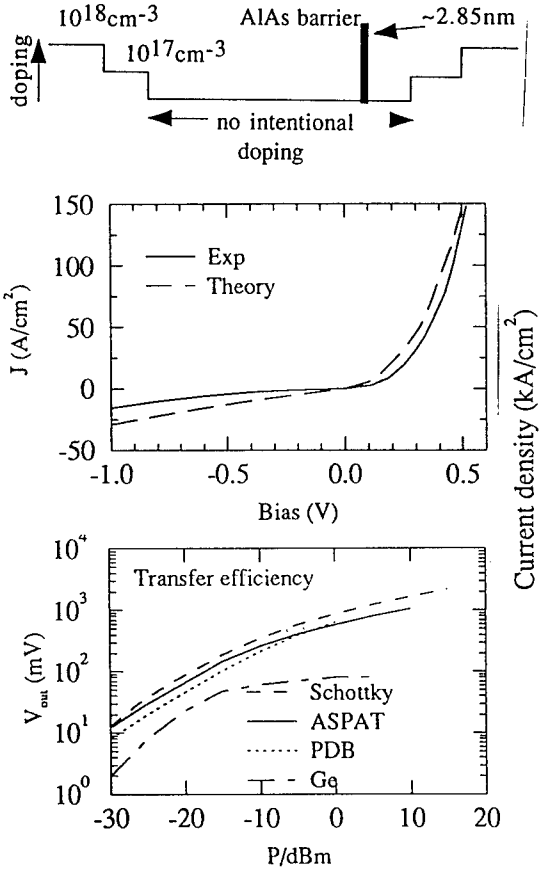


Figure 1: The ASPAT Diode, showing the general design, the I-V characteristics against a model calculation, and the transfer function of a microwave detector compared with Schottky, planar-doped-barrier and Ge backward diodes [5].

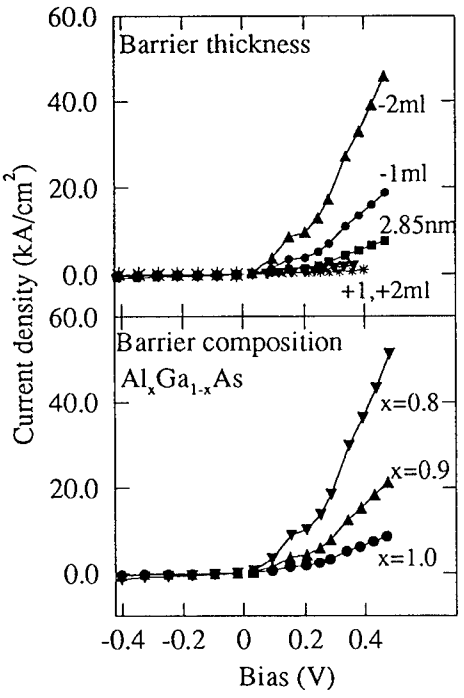


Figure 2: The sensitivity of the dc characteristics to (a) monolayer differences in layer thickness and (b) a shortfall in the the aluminium composition in the barrier layer.

device is that the current-limiting mechanism is tunnelling through a barrier rather than thermionic emission over a barrier. As a direct consequence, the sensitivity to ambient temperature of the microwave properties is greatly reduced [5]. It has a wider dynamic range than the Ge backward diode which relies on interband tunnelling, and is still used because of the temperature independence of its performance. In low-cost communication systems, the ability to dispense with temperature compensation circuits is attractive. The temperature-dependence that remains comes from residual thermionic emission over the barrier.

A simulation package has been developed for help in the design of the device [5].

This incorporates a self-consistent solution of the Poisson equation in the n-i-n doping structure (including an accumulation layer on the cathode side of the tunnel barrier), and uses the Thomas-Fermi approximation to obtain a starting potential profile. The current density passing through the tunnel barrier is then calculated by standard tunnelling theory techniques. There are some further important issues of detail, such as the determination of the Fermi energy throughout the structure. While this is easy at zero bias, the way E_f varies between the two contacts by an amount eV when a bias V is added is non-trivial. It is approximated here by being assumed constant though the undoped regions of GaAs from the appropriate contacts, and taking an average value within the barrier. No account is taken of tunnelling through indirect gap states in the AlAs or any inelastic processes during tunnelling or roughness of the tunnel barrier profile.

In designing to a specification, the incompleteness of the simulation package itself is a problem. If devices that are claimed to be grown within specification give different I-V characteristics, how much of the difference is due to the inadequate model and how much to variances between grown structure and the specification? We return to this problem below. In Figure 2, we show the tolerancing on our diode structure by virtue of monolayer errors in the barrier thickness and shortfalls in aluminium fraction in the barrier. The sensitivity to these variances is indicative of the exponential sensitivity of the current to the width and height of the barrier. This sensitivity has been the major hurdle in establishing manufacturable tunnel devices. Are crystal growers working in the commercial, as opposed to research, sphere able to meet specifications, such as shown in Table 1 below, with sufficient yield to provide diodes costing 20¢, as required for automobile radar at 76GHz or ozone monitoring at 170GHz? Multiple regrowths to hit the target, or the need to test individual devices in the event of device-to-device fluctuations, simply rule out this device as uncompetitive. There are other sensitivities in the design associated with the level of doping in the contact layers, and the length of the shorter undoped GaAs layer, but these are less critical than the details of the barrier layer.

TABLE 1: Layer Specification for ASPAT Diode

Layer	Material	Thickness (μm)	Doping (cm^{-3})
7	GaAs	0.75	3E18 Si
6	GaAs	0.04	1E17 Si
5	GaAs	0.005	nid
4	AlAs	0.00285	nid
3	GaAs	0.20	nid
2	GaAs	0.04	1E17 Si
1	GaAs	0.75	3E18 Si
0	GaAs	2" substrate	~1E18 Si

Thickness uniformity:
±1% over central 40mm
Thickness accuracy:
within 10% of spec.
Doping uniformity:
±1% over central 40mm
Doping accuracy:
within 20% of spec
Composition accuracy:
99% Al in barrier

There is an important issue of metrology that runs in parallel with the design and growth to a tunnel device specification. There are no wafer-scale, non-destructive, methods for qualifying wafers bought in from a commercial supplier. Furthermore, all conventional materials assessment methods, photoluminescence (PL), transmission electron microscopy (TEM), secondary ion mass spectroscopy (SIMS), and even X-ray diffraction (XRD) fall short in one or more ways in being able to confirm whether an as-

grown wafer is within specification [12]. Whereas TEM can specify chemical composition of layers with good detail, the data is valid only over a small region of the wafer surface (of linear dimension ~0.1μm), and nothing can be inferred about the doping profile. SIMS allows the doping and composition profiles to be registered with respect to each other, but not with the resolution demanded by the tolerancing. Both techniques are destructive and unsuitable for wafer-mapping. XRD can map out the uniformity of the GaAs layers above the AlAs, but it does not routinely have the resolution to determine that the total amount of Al in the AlAs layer is both laterally uniform and within the specified composition profile.

2.2 PROGRESS TO DATE ON MOCVD-GROWN MULTILAYERS

Wafers in batches of five have been ordered from commercial suppliers of MBE and MOCVD epitaxial multilayers, against the specification given in Table 1. Of the 13 suppliers approached only half quoted, the others not being able or willing to meet either the uniformity target or even the specification itself. (The MBE wafers were grown sequentially, while the MOCVD layers were grown simultaneously in a multi-wafer machine.) Representative data taken from MOCVD grown wafers by TEM, SIMS and XRD is collated and shown in Figure 3 as a radial map. The TEM gives an accurate thickness of the AlAs at some point on the wafer, and its depth from the top surface. Lateral variations, including the distribution of steps between layers of monolayer thickness differences, can also be determined on a length scale of ≤0.1μm. With the help of detailed simulations of the electron diffraction patterns, the aluminium composition profile through the barrier can also be determined. The SIMS allows the dopant (Si) distribution to be determined, and its general registry with respect to the Al concentration. Here there is a specific problem: the mass of AlH is the same as that of Si^{28}, and sputtered Al which picks up hydrogen from the background contaminates the Si signal, preventing an accurate registry of the doping profile with respect to the Al profile, and obliterating part of the Si profile. A more demanding Si^{29} SIMS analysis is required. The depth resolution of SIMS is insufficient in some cases even to detect some of the thinner layers in our structures. Absolute calibration of SIMS concentrations remains difficult, and in any event only the *dopant* concentration, not the *doping* concentration, is measured. The information about the size and shape of the Al profile (and *a fortiori* its wafer-scale uniformity) is contained in the relative amplitude of the successive peaks in the large angle X-ray data, and the signal-to-noise in a standard instrument is found to be insufficient - a full triple axis spectrometer may be required. The 'noise' in the high angle X-ray data (from which the Al profile would be reconstructed) may be intrinsic to the high doping level in the contact layers in our sample. The differences in thicknesses of layers 5-7 inferred by the different techniques is noted, and needs resolution.

Viewed as a radial map, the uniformity of the thicknesses of the layers seems satisfactory. It appears that the GaAs layers above the AlAs layer, and the AlAs layer itself are both slightly thinner than specified and there may even be systematic variations in the thicknesses as interpreted by the various techniques.

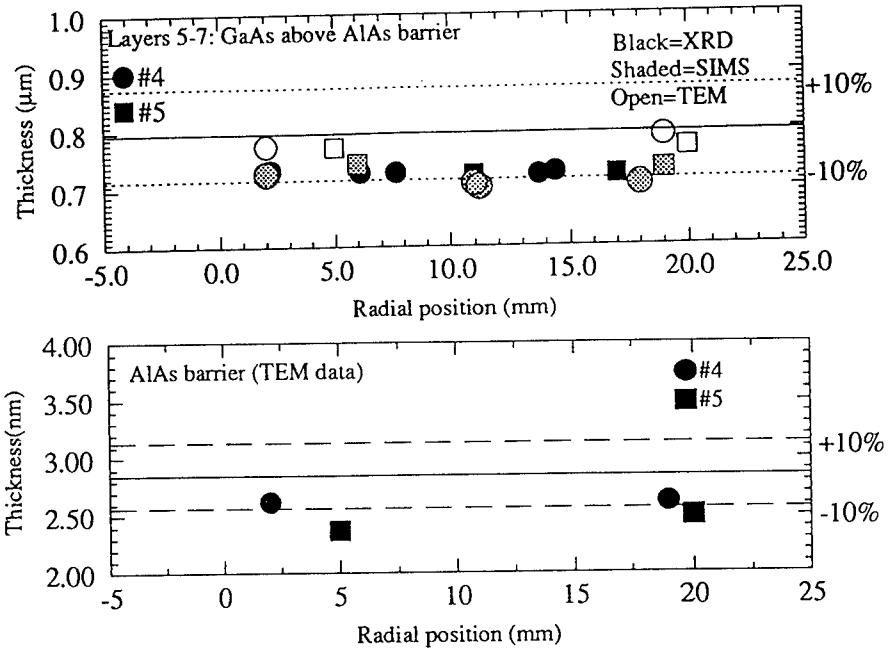


Figure 3: The radial uniformity of the MOCVD samples as implied by TEM, SIMS and XRD data (a) the GaAs layers above the AlAs barrier layer and (b) the AlAs barrier layer itself. It appears that all layers are systematically thinner than the target specificaton.

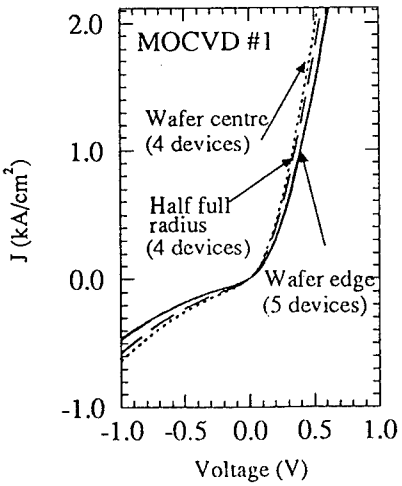


Figure 4: The radial variation in dc I-V characteristics taken from an MOCVD wafer.

In Figure 4, we show the range of dc I-V characteristics taken from another MOCVD-grown wafer in the same batch. Here the picture is potentially more hopeful. It appears that the uniformity of the I-V characteristics as averaged over 60μm diameter mesas from wafer #1 is quite adequate for production (a 6% spread from centre to mid-way out, and 16% from centre to edge). A 4Ω series resistance gives quantitative agreement between simulation and experimental I-V characteristics. The results from wafer #2 (with 20μm diameter diodes) seem quite unsatisfactory, with a 100% spread in the I-V data without a systematic trend, and the surface appearance suggests that this may be due to process-induced artifacts. Results on wafer #3 are similar to those shown in Figure 4.

The performance of this type of diode as a microwave detector has already been established [5]. As a part of this exercise, the uniformity of the rf parameters will be correlated with the dc parameters. Accelerated life testing will be undertaken: the active region of the device is single crystal material, so the device is likely to be robust [11].

2.3 PROGRESS TO DATE ON MBE-GROWN MATERIAL

The MBE wafers were grown sequentially. The same type of characterisation techniques have been applied, and the structural results inferred from TEM, XRD and SIMS analysis of two wafers are shown in Figure 5. All three techniques indicate a radial variation of the thickness (of layers 5-7) not present in the MOCVD data, and this has been seen in previous exercises [12]. Again there is a significant difference between the results inferred from the different analytical techniques. The SIMS data indicates that the doping level within the contacts is within specification, although the doping in layer 2 is not. The thickness of the AlAs barrier is within specification, but slightly thicker than the target figure. (The MOCVD layers were thinner than the target figure, and this difference has been noted before [5].) There is no electrical data as yet.

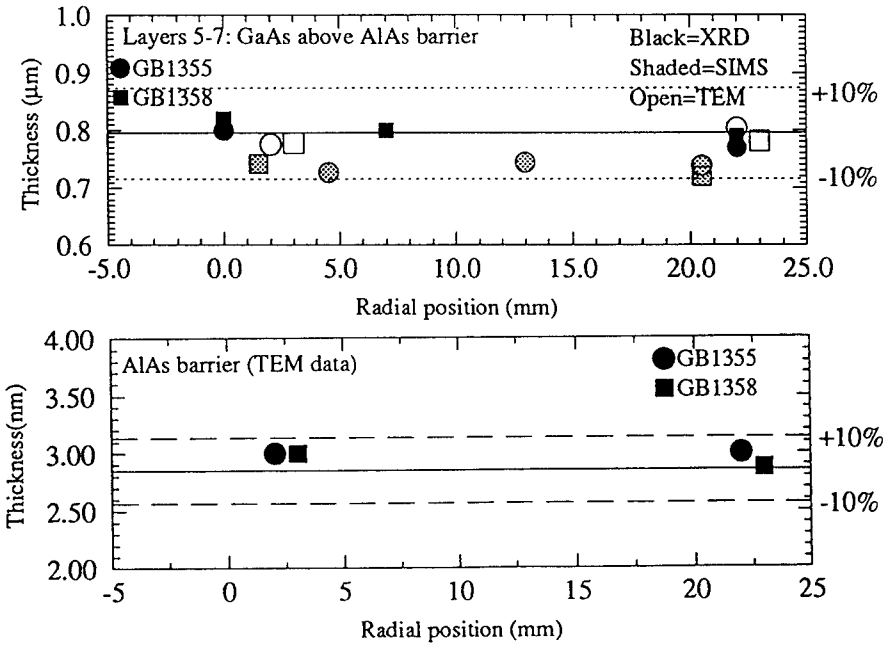


Figure 5: The radial uniformity of the MBE samples as implied by TEM, SIMS and/or XRD data (a) the GaAs layers above the AlAs barrier layer and (b) the AlAs barrier layer itself.

2.4 PRESENT STATUS

We have several challenges to meet before tunnel devices can be considered

manufacturable, meaning that we can produce devices to a specification with a sufficiently small spread in performance parameters.

Even if the dc and rf performances of the remaining wafers prove to have a narrow range of performances parameters, the methods for qualifing our wafers are unsatisfactory, and so we are unable to undertake any form of reverse engineering. We need to be much better able to attribute discrepancies between theory and experiment to shortcomings in the theory or to inaccuracies in the structure. Temperature and magnetic field dependences of the I-V characteristics may be a help.

A further issue is the length-scales of any fluctuations of composition and doping profile across a wafer, and how these are reflected in the device performance: this relates directly to the physics of mesoscopic systems. Does the electrostatic repulsion associated with current crowding limit the effect of increased tunnelling probability in sizeable areas of thinner/lower barrier? Given the encouraging results on wafers #1 and #3, is the modelling and simulation too sensitive to detail?

3 Implications for Mesoscopic Systems

3.1 THE CONVENTIONAL TECHNOLOGIES TAKEN FORWARD

The ability to fabricate very thin layers with adequate precision and uniformity for tunnel-based microwave diodes would be encouraging for the feasibility of using mesoscopic systems as elements of future computational systems. Our 20μm diameter microwave diodes may still be averaging the properties of the tunnel barriers, or even exploiting regions of thinnest and lowest Al composition. Even so, the modest level of variability (and the systematic variation across the wafer) of the diode device performance is encouraging.

If one continues to retain the conventional methods for device and circuit fabrication based on epitaxy and lithography at the level of mesoscopic devices (with active volumes and contact areas characterised by 10-100nm sides), then a comparable exercise involving electron-beam lithography is needed. The resistances of individual members of a series of wires of sub-0.1μm diameter of ~10μm length will have to show a small standard deviation, whether those wires be in metals, silicides or doped semiconductors.

The much tougher exercise of integrating both technologies at the mesoscopic level has started in a limited number of physics-based contexts. Starting with the work of Reed et al [13], a number of exercises have sought one-dimensional resonant tunnelling in ~0.1μm diameter pillars. Tewordt et al [14] have produced results with an enormous variability in I-V characteristics between diodes of comparable feature size. Indeed, some workers [15] have used the variability to map out the strategic locations of donor levels that play a controlling role in the level of current that a diode can transmit. It is a matter of urgency that an exercise comparable to that reported in section 2 is undertaken to ascertain the uniformity of I-V characteristics that can be achieved, and their stability under accelerated life-tests. A major difficulty here is the absence of any self-limiting process in operation while etching is performed: if such a process could be invented, it would have a major impact.

3.2 RADICAL ALTERNATIVE TECHNOLOGIES

There are many alternative approaches to the assembly of mesoscopic systems, although none have anything like the accumulated R&D investment of Si and GaAs in the context of computation and communication. The first example is the wealth of organic semiconducting molecules, often in self-assembled structures, that show applicable electronic or optical properties, although without the prospect of ever competing on the basis of speed with inorganic semiconductors - the electron-phonon interaction in hydrocarbons is simply too great [16]. The discovery of C_{60} should have been less a surprise than it was, as magic number clusters are a frequent occurrence in small molecular systems. These form identical size starting units, but the methods for ordering and connecting to and between such elements in giant three-dimensional arrays is not established.

It is possible to use natural lithography to form highly ordered arrays of 3D microstructures. The intercalation of the ~0.7nm diameter channels of zeolites with metals [17] gives rise to a dense bundle of 1D wires. The ~0.1μm diameter interstices of opal minerals can be filled in with semiconductors [18] to form 3D arrays of microstructures joined by narrower wires. This seems the most promising way forward for mastering mesoscopic fluctuations. The ability to tailor the arrays (so as to access and route information) has not yet been established. The ability to intercalate such structures with variable periodicity (the analogue of *staging* in graphite intercalation [19]) would represent a major step forward. Selective filling of different pore sizes might yet prove possible.

These radical alternatives are even more remote than tunnel structures from manufacturability, and are unlikely to have an impact on Si and GaAs technology in the foreseeable future, even if progress on these latter technologies were to come to a halt.

4. Summary

We are attempting to establish the manufacturability of tunnel structures, with a view to exploitation in electronic devices. Until we succeed, the prospects remain poor for ever exploiting mesoscopic devices in any way resembling more conventional devices.

5. Acknowledgements

This work is supported by the Engineering and Physical Sciences Research Council. The device fabrication and evaluation was undertaken by Mr M Carr at GEC Plessey Semiconductors. We thank Dr W M Stobbs (University of Cambridge) for the TEM results, Dr A Chew (Loughborough University of Technology) for the SIMS analysis and Dr P Kidd (University of Surrey) for the XRD measurements. The MOCVD layers were grown by Epitaxial Products International Ltd (Cardiff) and the MBE samples by GEC Marconi Materials Technology Ltd (Caswell).

6. References

1. Brown, E.R., Sollner, T. C. L. G., Parker, C. D., Goodhue, W. D. and Chen, C. L. (1989) Oscillation up to 420GHz in GaAs/AlAs resonant tunnelling diodes, *Applied Physics Letters* **55** 1777-9.
2. Brown, E. R., Soderstrom, J. R., Parker, C. D., Mahoney, L. J., Molvar K. M. and McGill, T. C. (1991) Oscillations up to 712GHz in InAs/AlSb resonant tunnelling diodes, *Applied Physics Letters* **58** 2291-4.
3. Javalagi, S., Reddy, V., Gullapalli K. and Neikirk, D. (1992) High efficient microwave diode oscillators, *Electronics Letters* **28** 1969-701.
4. Rydberg, A., Gronqvist, H. and Kollberg, E. (1990) Millimeter- and submillimeter-wave multipliers using quantum-barrier varactor (QVB) diodes, *IEEE Electron Device Letters* **EDL11** 373-5.
5. Syme, R. T., Kelly, M. J., Smith, R. S., Condie, A. and Dale, I. (1991) A tunnel diode with asymmetric spacer-layers for use as a microwave detector, *Electronics Letters* **27** 2192-4, and Syme, R. T. (1993) Microwave detection using GaAs/AlAs tunnel structures, *GEC Journal of Research* **11** 12-23.
6. Higgs, A. W. and Smith, G. W. (1991) Conversion gain at room temperature in a GaAs/AlAs double-barrier resonant-tunnelling mixer, in K E Singer (ed), *Gallium Arsenide and Related Compounds 1990*, Institute of Physics Conference Series **112** pp. 495-7, and Hayes, D. G., Higgs, A. W., Wilding, P. J., and Smith, P. J. (1993) Conversion gain at 18GHz from resonant tunnelling diode mixer operated in fundamental mode, *Electronics Letters* **29** 1370-2.
7. Yokoyama, N., Imamura, K., Ohnishi, H., Mori, T., Muto, S. and Shibatomi, A. (1988) Resonant Tunnelling Hot Electron Transistor, *Solid State Electronics* **31** 577-82, and Imamura, K., Takatsu, M., Mori, T., Adachihara, T., Muto, S., and Yokoyama, N. (1992) A Full Adder Using Resonant-Tunneling Hot Electron Transistors (RHETs), *IEEE Transactions on Electron Devices* **39** 2707-10.
8. Capasso, F., Sen, S., and Beltram, F. (1990) Quantum-Effect Devices, in S. M. Sze (ed.), *High-Speed Semiconductor Devices*, Wiley, New York, pp. 465-530.
9. Sen, S., Capasso, F., Cho, A. Y. and Sivco, D. (1987) Resonant Tunneling Device with Multiple Negative Differential Resistance: Digital and Signal Processing Applications with Reduced Circuit Complexity, *IEEE Transactions on Electron Devices* **ED-34** 2185-90.
10. Mars, D. E., Yang, L., Tan, M. R. T and Rosner, S. J. (1993) Reproducible growth and applications of AlAs/GaAs double barrier resonant tunneling diodes. *Journal of Vacuum Science and Technology B* **11** 965-8
11. Kearney, M. J., Kelly, M. J., Davies, R. A., Kerr, T. M., Condie A. and Dale, I. (1989) Asymmetric planar doped barrier diodes for mixer and detector applications, *Electronics Letters* **25** 1454-6, and Kearney, M. J. and Dale, I. (1990) GaAs planar doped barrier diodes for mixer and detector applications, *GEC Journal of Research* **8** 1-12.
12. Rimmer, N., Syme, R. T., Frost, J. E. F., Ritchie, D. A., Jones, G. A. C., Kelly, M. J. and Stobbs, W. M. (1991) Wafer uniformity of an MBE grown $Al_xGa_{1-x}As$ tunnelling structure, in A. G. Cullis and N. J. Long (eds), *Microscopy of*

Semiconducting Materials, Institute of Physics Conference Series **117** pp. 577-80

13. Reed, M. A., Randall, J. N., Aggarwal, R. J., Matyi, R. J., Moore, T. M. and Wetsel, A. E. (1988) Observation of discrete electronic states in a zero-dimensional nanostructure, *Physical Review Letters* **60** 535-8.
14. Tewordt, M., Law, V. J., Nicholls, J. T., Martin-Moreno, L., Ritchie, D. A., Kelly, M. J., Pepper, M., Frost, J. E. F., Newbury, R. and Jones, G. A. C. (1994) Single-electron tunneling and Coulomb charging effects in ultrasmall double-barrier heterostructures, *Solid-State Electronics* **37** 793-9 and references therein.
15. Dellow, M. W., Beton, P. H., Langerak, C. J. G. M., Foster, T. J., Main, P. C., Eaves, L., Henini, M., Beaumont, S. P. and Wilkinson, C. D. W. (1992) Resonant tunnelling through the bound states of a single donor atom in a quantum well, *Physical Review Letters* **68** 1754-7.
16. Burroughes, J. H., Bradley, D. D. C., Brown, A. R., Marks, R. N., Mackay, K., Friend, R. H., Burns, P. L. and Holmes, A. B. (1990) Light-Emitting Diodes Based on Conjugated Polymers, *Nature* **347** 539-41
17. Romanov, S. (1993) Electronic Structure of Minimum-diameter Tl, Pb and Bi Quantum Wire Superlattices, *Journal of Physics: Condensed Matter* **5** 1081-90
18. Romanov, S. (private communication, 1995, and to be published).
19. Suematsu, H. (1992) Structural Phase Transitions and Kinetic Processes in Graphite Intercalation Compounds, in (H. Aoki et. al. eds.) *New Horizons in Low-Dimensional Electron Systems*, Klewer Academic Publishers, Dordrecht, pp. 25-44

PROSPECTS FOR QUANTUM DOT STRUCTURES APPLICATIONS IN ELECTRONICS AND OPTOELECTRONICS

R. A. SURIS
A. F. Ioffe Physical-Technical Institute
194021 Polytechnicheskaya 24, St. Petersburg, Russia

1. Introduction

Current technological achievements promise to make imaginable manufacturing of macroscopic arrays of Quantum Dots (QDs). This is why it is reasonable to discuss some prospects of their applications for electronics and optoelectronics.

The main feature of QDs is real discrete energy spectrum of electrons and holes. In quantum well (QW) carriers are confined only in one direction (*Figure 1*) while in the plane of QW their motion is free and is characterized by continuous energy spectrum. In quantum wire (QWR) carriers freely move along its axis. It means those carrier energy spectra in QWs and QWRs are continuous and there is overlapping of energies of confined states and delocalized states. Such overlapping results in very fast processes of carrier transition between confined and delocalized states as well as between ground and excited states.

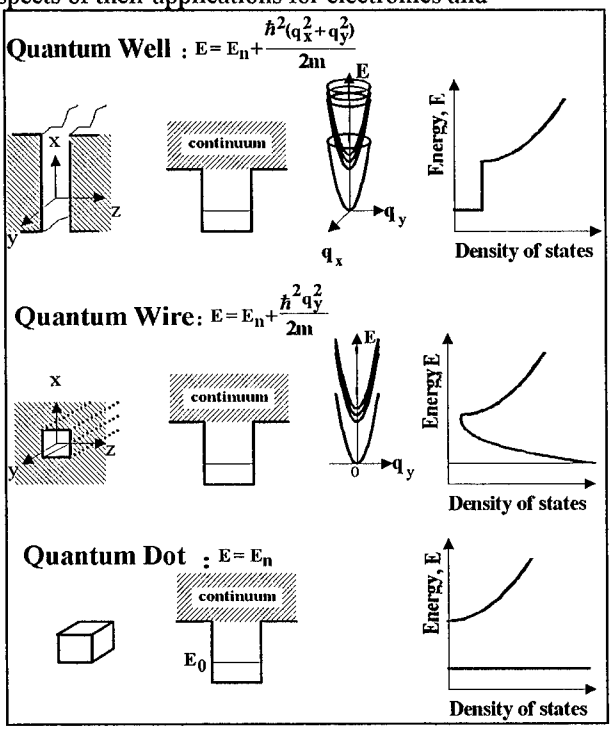


Figure 1

S. Luryi et al. (eds.), Future Trends in Microelectronics, 197–208.

We shall discuss how these features manifest themselves in photodetectors based on QWs and QDs. We will briefly consider some application possibilities of QD structures for cascade lasers. Results of analysis of threshold current reduction for QD injection lasers will be presented. Finally, we will briefly discuss some peculiarities of Bloch oscillations in 2D and 3D arrays of QDs.

2. Semiconductors with QDs as Material for IR Photodetectors

The efficiency of a photodetector is controlled by the product of the carrier lifetime, τ, and the carrier photoexcitation rate, G: $G \cdot \tau$. Let us consider photodetectors based on multiple QW structures or superlattices (SL). The carrier photoexcitation rate, G, is proportional to the QWs number per a unit length, or to the reciprocal distance between QW centers, 1/L, and to the photoexcitation probability which is proportional to the surface concentration of carriers filling localized states in QW and it depends on QW shape.

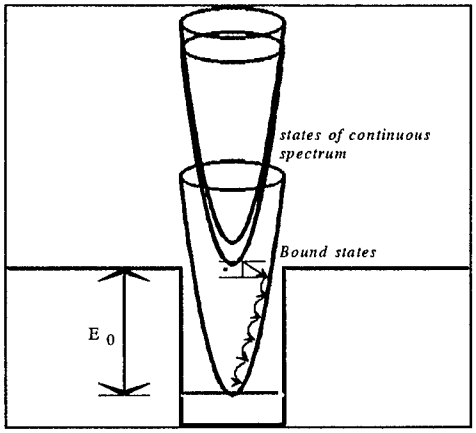


Figure 2

The carrier lifetime is controlled by the process of the carrier capture by QWs. This process can be divided into two stages (Figure 2):
(i) Carrier transitions from the states of continuos spectrum of longitudinal motion into those localized in QWs with a large lateral momentum, and
(ii) localized carrier "cooling" down by the lateral motion energy due to the phonon emission, Ref.[1]

If the temperature is small enough, an electron being transferred in the localized states due to the optical phonon emission practically has no chances to go back. Therefore, the capture is controlled by the first process.

This process is naturally characterized by the capture velocity, $S_{k,q}$. This velocity is a part of the normal component of current density per free carrier with the longitudinal motion energy $E_k = \hbar^2 k^2/2m$ and the lateral momentum $\mathbf{q}$. This part feeds carrier transitions from the states of continuous spectrum of longitudinal motion into the localized states due to the phonon emission:

$$S_{k,\mathbf{q}} = \frac{2\pi}{\hbar} \sum_{\mathbf{Q}} |V_{\mathbf{Q}}|^2 \left| \langle \psi_k | e^{iQ_z \cdot z} | \psi_0 \rangle \right|^2 \left(N_{\mathbf{Q}} + 1 \right) \cdot \delta \left(\frac{\hbar^2 k^2}{2m} + \frac{\hbar^2 \mathbf{q}^2}{2m} + E_0 - \frac{\hbar^2 (\mathbf{q} - \mathbf{Q}_t)^2}{2m} - \hbar \omega_{\mathbf{Q}} \right)$$

Here the matrix element of the electron - optical phonon interaction is

$$V_Q = \frac{4\pi e e^*}{\sqrt{Q_z^2 + \mathbf{Q}_l^2}} \sqrt{\frac{\hbar N}{2M\omega_Q \Omega}}$$

where e* and M are effective charge and ion reduced mass, ω_Q is longitudinal phonon frequencies (the phonon wavenumber $\mathbf{Q} = (Q_z, \mathbf{Q}_l)$, Ω is the total system volume. ψ_k is electron wave function of the continuous spectrum, ψ_0 and E_0 are the wave function and ionization energy of the localized state and $N_{\mathbf{Q}} = \left(\exp\left(\hbar\omega_{\mathbf{Q}} / T\right) - 1\right)^{-1}$.

We shall estimate the parameter $S_{k,q}$, neglecting phonon dispersion and supposing δ - like potential for QW[1]. Taking into account that in the usual situation the QW ionization energy E_0 is much more than the energies of phonons and incident carriers and using the following equations

$$\left|\left\langle \psi_k \left| e^{iQ_x \cdot x} \right| \psi_0 \right\rangle\right|^2 \approx 16 \frac{\kappa k^2 Q_z^2}{\left(\kappa^2 + Q_z^2\right)}$$

where $\hbar\kappa \equiv \sqrt{2mE_0} >> \hbar k$, we obtain[2]

$$S_{k,\mathbf{q}} = \pi \cdot \frac{\varepsilon_0 - \varepsilon_\infty}{\varepsilon_\infty} \cdot \frac{e^2}{\hbar\varepsilon_\infty} \cdot \frac{k^2}{\kappa^2} \cdot \frac{\hbar\omega}{E_0} \cdot \left(N_{\mathbf{Q}} + 1\right).$$

Here ε_0 and ε_∞ are the dielectric constant values at low and high frequencies.

Averaging $S_{k,\mathbf{q}}$ with Boltzman distribution function gives

$$S = \pi \cdot \frac{\varepsilon_0 - \varepsilon_\infty}{\varepsilon_\infty} \cdot \frac{e^2}{\hbar\varepsilon_\infty} \cdot \frac{\hbar\omega}{E_0} \cdot \frac{T}{E_0} \cdot \left(N_{\mathbf{Q}} + 1\right)$$

The capture velocity S as a function of ionization energy E_0 and temperature T is presented in Figure 3.

If the distance between QWs, L, is less than carrier free path length, l_{fp}, we can introduce the carrier lifetime $\tau = L/S$ (Figure 3). If $L > l_{fp}$, we should solve drift - diffusion equations equating the carrier current density on the QW surface to $S \cdot n_b$, where n_b is the carrier boundary concentration. Nevertheless, this estimation for τ is still correct if $\tau > L^2 / D$ (D is the diffusion coefficient).

Therefore, we see that the efficiency of photodetectors based on multiple QWs (or SL) is limited by a comparatively high carrier capture rate[3]. For L = 100

[1] For the sake of simplicity we shall not consider the effect of resonant states on the capture process.

[2] Naturally, this perturbation theory result is valid if $S_{k,q}$ is less than the velocity in the z-direction, $\hbar k/2m$.

[3] By the way, the parameter $G\tau$ does not depend on the distances between QWs because $G \propto 1/L$ while $\tau \propto L$.

Å the characteristic scale of lifetime is 10^{-6}s. This is due to overlapping the energy spectra of free electrons and of electrons localized in QWs in the z-direction.

The situation can be radically improved by replacing of QWs by QDs. In this case, we have a real discrete spectrum of localized states. The spectra of localized and free electrons have no overlapping and the capture process is strongly suppressed. One can suppose the capture cross sections of QDs should be of the same order of magnitude as the capture cross sections of impurity states and the carrier lifetime have to be large.

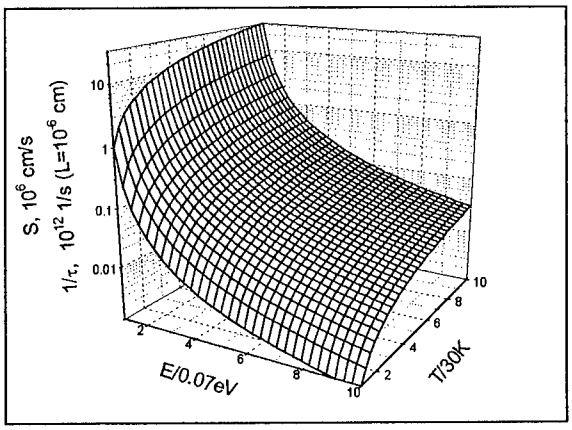


Figure 3

Consequently, the photoelectric characteristics of such structures should be similar to those of impurity photoconductors. A great advantage of the material with QDs would be the possibility to cover the desirable wavelength range by manufacturing of semiconductor structures with the QDs of corresponding sizes and using heterojunctions with the corresponding band offsets. Here we would deal with engineering artificial atoms.

3. Semiconductors with QDs as Material for Cascade Lasers

The problem of spectrum overlapping remarkably manifests itself in the cascade lasers, Ref[2], Ref.[3] . Just this overlapping resulting in very high rate of electron transitions from the excited state 2 into the ground state 1 inside each QW is the main obstacle against the creation of inverse population between the quantization energy levels 1 and 2 inside of QWs, Ref.[1] (Figure 4). It is why in Ref. [2] we proposed to create the inverse population between the ground state of a QW and the excited state of adjacent QW. Price for this is a small oscillator strength and, consequently, a small gain.

In semiconductor with 3D array of QDs this problem can be easily resolved. Again, the exclusion of spectrum overlapping results in a strong suppression of electron transitions from the excited state 2 into the ground state 1 of QD. This suppression allows us to obtain inverse electron population within each QW. It should be noted that this statement is almost

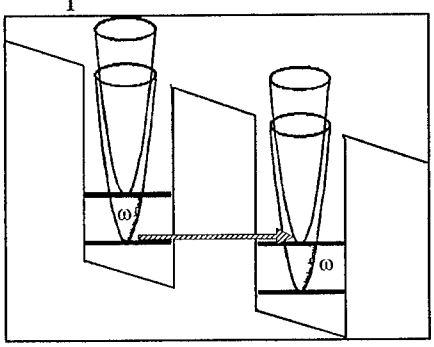


Figure 4

trivial: the similar situation is ordinary for atoms and here we deal with artificial atoms. However, there are two important distinctions in the case of semiconductor with QDs.

The first one is a conceivable possibility to make these artificial atoms have the desirable spectrum. It can be done by choosing the appropriate sizes of QDs and band offsets.

The second remarkable and very important distinction consists in the possibility of dc current pumping inherent in semiconductor heterolasers. This is a current of consequent electron tunneling between the neighboring QDs that is providing by tunable thickness and height of the potential barrier between QDs.

Let us try to imagine some thinkable schemes of cascade lasers with QDs.

The first one uses four-level system (Figure 5). The electric field is applied

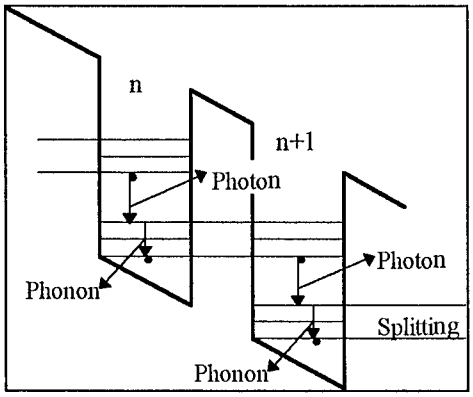


Figure 5

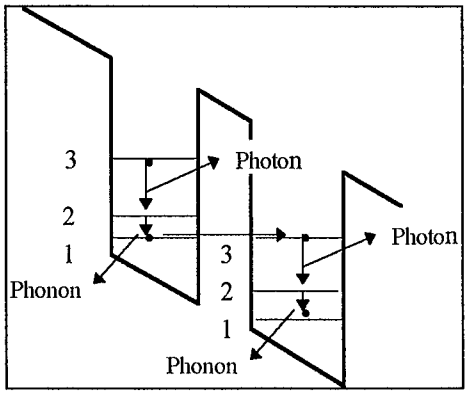


Figure 6

to the semiconductor with 3D array of QDs. When the field provides the resonance between the ground electron state of n-th QD and the first excited states of the neighboring (n+1)-th QD, their energy levels are splitted into two ones (see Figure 5). If the splitting value is close to the phonon energy and temperature is low, the filling factor of upper splitted level is less than filling factor of the lower one. Therefore, we have the inverse occupation in the system of hybridized states of the neighboring pairs of QDs (Figure 5). This picture is valid while the splitting value is large as compared with $\hbar/\tau_r$ where τ_r is the characteristic time of electron transitions between the splitted levels. It implies a large value of overlapping integrals of wave functions of neighboring QDs and low temperature.

Other scheme uses three-levels in each QW (Figure 6). This scheme would work when the rate of electron transitions $2 \Rightarrow 1$ inside a QD exceeds the rate of the tunnel transitions $1 \Rightarrow 3$ between the neighboring QDs. In this case we have the inversion between the levels 3 and 2 inside each QD.

4. Injection Lasers with QDs

Another direction of exploitation of carrier discrete spectrum in the QD

structures is their application to injection lasers. The main expected advantage of QD lasers over the conventional quantum well lasers is a lower threshold current density, Ref.[4] The main reason why the QD lasers should have very low threshold currents is a δ-function like energy spectrum of carriers in QDs (*Figure 1*). A schematic picture of QD laser along with the band diagram is presented in *Figure 7*. Transitions between the electron and hole levels in QDs are analogous to those between the exactly discrete levels of individual atoms as it was discussed above. Accounts have appeared quite recently about the early fabrications of the QD lasers, Ref. [5].

We did not discuss the question of the technological noise effect on device characteristics up to now. It is naturally to attempt to estimate whether the expected advantages of QD lasers could be practically used or the technological noise "kills" the effect. Dispersion of QD sizes broadens the laser line and this will eventually result in increasing the threshold current. Below we discuss some results of the theoretical analysis of the threshold current density of a QD laser, taking account of the inhomogeneous line broadening caused by the QD-size dispersion (Ref.[6]). The theoretical estimations presented confirm the possibility of a significant reduction of the threshold currents of QD lasers as compared with the conventional quantum well lasers.

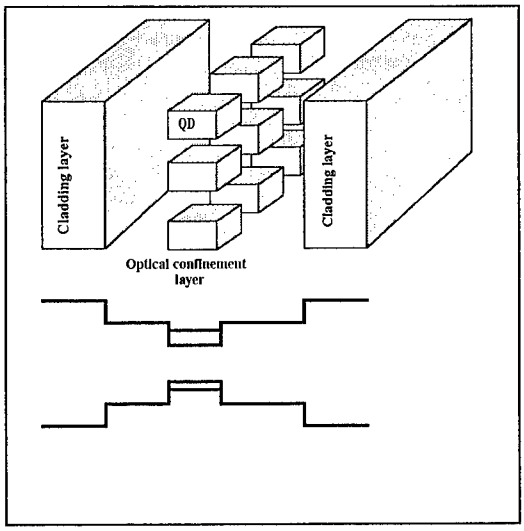


Figure 7

Fortunately, carrying out the analysis is simplified by the following fact. The minimum threshold current density and optimum parameters of the laser (surface density of QDs and thickness of the waveguide region) can be presented as universal functions of the one dimensionless parameter that describes the inhomogeneous line broadening. This parameter is the ratio of the stimulated transition rate in QDs to the spontaneous transition rate in the waveguide region.

Generally, the following cases are realized (depending on the temperature T, QD-size fluctuations and conduction and valence band offset at the QD-narrow-gap region heteroboundary $\Delta E_{c,v}$):

A) Equilibrium filling of QDs (relatively high temperatures and/or shallow potential wells) and narrow line of the quantized energy distribution. That is the case when (i) the characteristic times of thermally excited escapes of an electron and hole from a QD are small compared with the radiative lifetime in QDs τ_{QD} and

(ii) the inhomogeneous line broadening (due to fluctuations in QD-sizes) $(\Delta\varepsilon)_{in\,hom}$ is less than the temperature.

The threshold current density is

$$j_{th} = \frac{eN_S^{min}}{\tau_{QD}} \frac{< f_n >< f_p >}{< f_n > + < f_p > -1} + ebBn_1p_1 \frac{< f_n > + < f_p > -1}{(1- < f_n >)(1- < f_p >)},$$

where N_S is the QD surface density, b is the thickness of the optical confinement layer (OCL) (the thickness of the narrow-gap region), B is the radiative constant for this region, $n_1 = N_c \exp\left((\varepsilon_n - \Delta E_c)/T\right)$, $p_1 = N_v \exp\left((\varepsilon_p - \Delta E_v)/T\right)$, $\varepsilon_{n,p}$ are the quantized energy levels of an electron and hole in QD (measured from the band edges), $< f_{n,p} >$ are the filling factors (averaged over QDs) of these levels, satisfying the threshold condition $< f_n > + < f_p > -1 = N_S^{min}/N_S$, $N_S^{min} = (1/\xi)(a/\Gamma_y)(F/\tau_{ph})$ is the minimum QD surface density required for the lasing at given losses in waveguide β and inhomogeneous line broadening, a is the mean size of QDs, Γ_y is the optical confinement factor, F is the photon mode density, $\tau_{ph} = \sqrt{\varepsilon}/(c\beta)$ is the photon lifetime, ξ is the numerical constant (equal to $/\pi$ and $/\sqrt{2\pi}$ for the Lorentzian and Gaussian functions of QD-size distribution, respectively).

The minimum threshold current density is expressed as

$$j_{th}^{min} = \left(\sqrt{\frac{eN_S^{min}(b^{opt})}{\tau_{QD}}} + \sqrt{eb^{opt}Bn_1p_1}\right)^2 = eb^{opt}Bn_1p_1\left(1 + \sqrt{\frac{a}{b^{opt}\Gamma_y(b^{opt})}s}\right)^2,$$

where the dimensionless parameter s is the ratio of the stimulated transition rate in QDs at the lasing threshold to the spontaneous transition rate in the narrow-gap region

$$s = \frac{(1/\xi)(a/\Gamma_y)(F/\tau_{ph})(\Delta\varepsilon)_{in\,hom}}{Bn_1p_1}.$$

The optimum thickness of the OCL, b^{opt}, and the optimum surface density of QDs, N_S^{opt}, are, in their turns, the functions of the dimensionless parameter s. By way of illustration we considered the following double-heterostructure laser structure: the materials of wide-gap regions (cladding layers), narrow-gap region (OCL) and QDs are InP, $Ga_{0.21}In_{0.79}As_{0.46}P_{0.54}$ and $Ga_{0.47}In_{0.53}As$ respectively, the latter two being lattice-matched to InP. For the total

By way of illustration we considered the following double-heterostructure laser structure: the materials of wide-gap regions (cladding layers), narrow-gap region (OCL) and QDs are InP, $Ga_{0.21}In_{0.79}As_{0.46}P_{0.54}$ and $Ga_{0.47}In_{0.53}As$ respectively, the latter two being lattice-matched to InP. For the total losses in the waveguide β =10 cm^{-1} and 10 percent relative QD-size fluctuations, we obtained $j_{th}^{min} \approx 8.3\,A/cm^2$, $N_S^{opt} \approx 6.2\times10^{10}cm^{-2}$. Optimal surface density of QDs and minimum threshold current density are presented on *Figure 9* as functions on QD size relative fluctuations, δ, and waveguide losses, β.

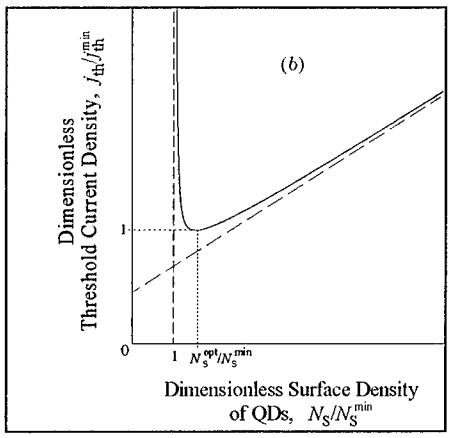


Figure 8

B) Equilibrium filling of QDs (relatively high temperatures and/or shallow potential ells) and wide line of the quantized energy distribution. This is the case when the condition (i) of section *A*) satisfies and (ii) the inhomogeneous line broadening (due to fluctuations in QD-sizes) is larger than the temperature. This case is of little importance in view of the fact that the threshold current density is high as compared with threshold current density in the other two cases discussed.

C) Nonequilibrium filling of QDs (relatively low temperatures and/or deep potential wells). That is the case when the characteristic times of thermally excited escapes of an electron and hole from a QD are large as compared with the radiative lifetime in QDs.

The threshold current density is

$$j_{th} = \frac{eN_S}{\tau_{QD}} f_n f_p + \frac{ebB}{\nu_n \nu_p \tau_{QD}^2} \frac{f_n^2 f_p^2}{(1-f_n)(1-f_p)},$$

where $f_{n,p}$ are the filling factors (common to all QDs in this case) satisfying the threshold condition $< f_n > + < f_p > -1 = N_S^{min}/N_S$, $\nu_{n,p} = \sigma_{n,p}\nu_{n,p}$, $\sigma_{n,p}$ are the cross sections of electron and hole capture into a QD, $\nu_{n,p}$ are the thermal velocities of electrons and holes. Here the ratio of the stimulated transition rate in QDs at the lasing threshold to the spontaneous transition rate in the narrow-gap region, $s = (1/\xi)(a/\Gamma_y)(F/\tau_{ph})(\Delta\varepsilon)_{inhom}/Bn_1p_1$, plays the role of a universal dimensionless parameter controlling the magnitudes of the minimum threshold current density, optimum surface QD density and the optimum thickness of the OCL.

Therefore, the threshold currents of injection lasers can be significantly reduced as compare with convenient QW lasers by using active layers with QDs. The threshold currents strongly depend on inhomogeneous broadening caused by the dispersion of QD parameters.

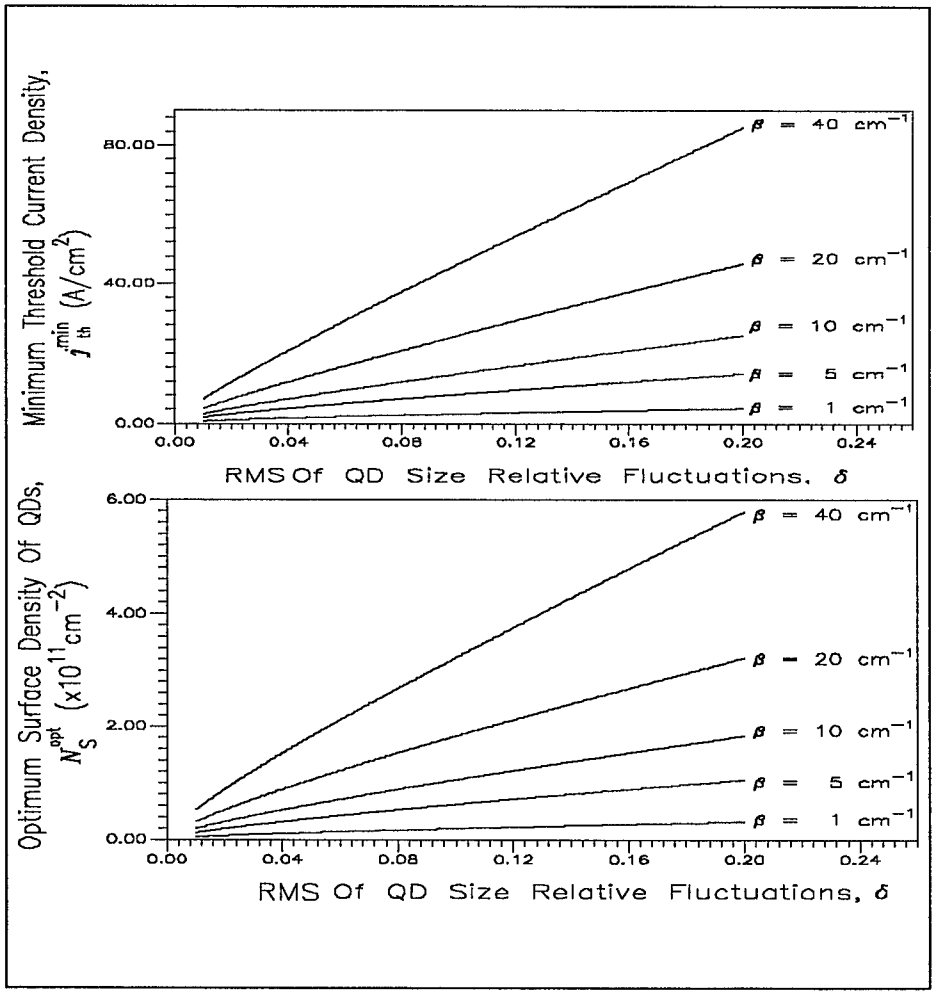


Figure 9

In conclusion of this Section, it is necessary to make an important remark.

As we have discussed in Section 2, the carrier capture by QDs is strongly suppressed as compared with QWs. On the other hand, the process of carrier thermal excitation from QDs is slow too, and the more QD ionization energy, the lower the thermal excitation rate. Naturally, it should results in slowing down the dynamic processes in lasers based on QDs.

Other consequence of the capture and excitation suppression is the suppression of carrier space redistribution between QDs. This peculiarity should

result in "hole burning" in gain near the maxima of light intensity in the laser cavity. Obviously, it gives low values of the threshold current for multifrequency generation, as it should take place for the laser generation on the transitions: conducting band ⇒ acceptor, Ref.[7].

4. Bloch oscillations in QD arrays

Let us consider the 1D SL consisting of the QW periodical array. In the tight binding approach the electron energy spectrum is

$$E(k,\mathbf{q}) = \Delta\cdot(1-\cos ka) + {\hbar^2\mathbf{q}^2}/{2m}$$

where k and **q** are the electron wavenumbers along the SL axis and in QW plane correspondingly, Δ is the miniband width and a is the SL period. In the framework of semiclassical approach the electron movement in electric field F directed along SL axis can be described with the following equations

$$\text{for momentum } \frac{d}{dt}\hbar k(t) = eF, \quad \text{and for velocity } v(t) = \frac{\partial}{\partial\hbar k}E(k,\mathbf{q}) = \frac{\Delta a}{\hbar}\cdot\sin\left(k(t)\cdot a\right)$$

Therefore, abrupt switching on of the electric field results in electric current oscillations:

$$j = e\frac{\Delta a}{\hbar}\cdot n\cdot\left\langle\cos\left(k_o a\right)\right\rangle\cdot\sin(\Omega t)$$

Here $\Omega \equiv eFa/\hbar$ is the Bloch oscillation (BO) frequency, n is the electron density and the brackets $\langle..\rangle$ mean averaging over initial wavenumber distribution.

The BO can manifest themselves as microwave and far IR radiation. However, the BO are very fast smoothed by electron scattering processes and after the relaxation the voltage dependence of current is described by the N-shaped characteristics (Ref.[8]):

$$j = e\frac{\Delta a}{\hbar}\cdot n\cdot\left\langle\cos k_o\right\rangle\frac{\Omega\tau_{sc}}{1+\left(\Omega\tau_{sc}\right)^2}$$

with a scattering time τ_{sc} that is of order of 10^{-13}... 10^{-12} s.

The eigenfunctions of an electron in SL under the electric field F are centered on the N-th SL cell with the characteristic length of localization $a\cdot(\Delta/eFa)$. The eigenenergy of this state is −eFaN. The eigenenergies form so called Stark ladder. Using

this picture we can consider BO as time evolution of the initial wave pocket $\delta(k-k_o)$. One can easily obtain presented above BO of the current, expanding this function in set with basis formed by eigenfunctions of the electron in SL with the electric field[4], $\Psi_N(n) = J_N(\Delta/eFa)$, where J_N is the Bessel function of N-th rank. The picture of BO is smoothed over due to mixing of different states of the Stark ladder caused by scattering processes. Again, these processes are very intensive in the SL consisting of the QW array due to overlapping energy spectra of the lateral movement of the Stark ladder states with different numbers N (Fugure 10).

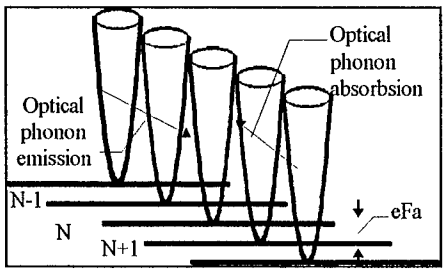


Figure 10

The scattering processes would be suppressed in the SLs formed as 2D or 3D QD periodic structures. In this case, in the tight binding approximation the electron spectrum is[5]

$$E(k_x,k_y,k_z) = \Delta_x\cdot(1-\cos(a_x k_x)) + \Delta_y\cdot(1-\cos(a_y k_y)) + \Delta_z\cdot(1-\cos(a_z k_x)),$$

where a_x, a_y and a_z are QD SL periods.

In the electric field with the components (F_x,F_y,F_z) the electron spectrum is completely discrete and consists of the triple Stark ladder:

$$E_{N_x,N_y,N_z} = -eF_x a_x\cdot N_x - eF_y a_y\cdot N_y - eF_z a_z\cdot N_z,$$

where N_x, N_y and N_z are integers. Electrons are localized in 3 dimensions in contrast to 1D SL consisting of QWs, where they are localized only in direction of SL axis while the motion in the QW plane remains free.

Elastic scattering processes can not cause transitions between states with different (N_x, N_y, N_z) due to the energy difference[6]. As for inelastic transitions with the phonon emission or absorption, these transitions are impossible while phonon frequencies are smaller than Bloch frequency, Ω. If phonon frequencies, ω, are much more than Ω, there are possible transitions only between states that lies far away apart. The distance is as large as $(\Omega/\omega)a$ and probability of these transitions exponentially decreases with (Ω/ω).

A remarkable peculiarity of BOs in 2D or 3D periodical structures of QDs is the strong dependence of the BO spectrum on electric field orientation. Using

---

[4] This result can be easily obtained from more general expression for eigenfunctions presented in Ref[2]

[5] Here for the sake of simplicity we suppose the QDs forming a rectangular lattice with periods a_x, a_y, a_z.

[6] In principle, when the rations of F_x, F_y, and F_z are rational there are sets of different N_x, N_y and N_z values giving the same energy value. In this case the resonant transitions are possible. Here we do not consider this case and restrict ourselves to the following remark. If all of this rations are far from unity the probabilities of these transitions are exponentially small.

semiclassical approach we obtain the following equations for the current density components

$$j_\alpha = e\frac{\Delta_\alpha a}{\hbar}\cdot n\cdot\left\langle\cos\left(k_{o,\alpha}a_\alpha\right)\right\rangle\cdot\sin\left(\Omega_\alpha t\right), \qquad \Omega_\alpha = eF_\alpha a_\alpha/\hbar, \quad \alpha = x,y,z$$

Therefore, the current projection on the certain direction is the sum of three harmonics with the frequencies that can be tuned by applied field rotation.

One can expect that this feature should significantly broaden functionality of the BO applications.

5. References

[1] R. Kazarinov and R. Suris, Theory of electrical properties of semiconductors with superlattices, *Sov. Phys.- Semiconductors*, 1973, 7, no. 3, p. 347

[2] R. Kazarinov and R. Suris, Possibility of amplification of electromagnetic waves in a semiconductor with superlattice, *Sov. Phys.- Semiconductors,* 1971, **5**, no. 4, p. 707; Electric and electromagnetic properties of semiconductors with a superlattice, *Sov. Phys.- Semiconductors*, 1972, **6**, no. 6, p. 120,

[3] J. Faist, F. Capasso, D. L. Sivco, C. Sirtori, A.L. Hatchinson and A.Y. Cho, Quantum cascade laser, *Science* 1994, **264,** 553; *Electron. Lett.* 1994, **30,** 865

[4] Y. Arakawa, H. Sakaki, Multidimensional quantum well laser and temperature dependence of its threshold current, *Appl. Phys. Lett.*, 40, 939, (1982).

[5] N. Kirstaedter, N. Ledentsov, M. Grundmann, D. Bimberg, V. Ustinov, S. Ruvimov, M. Maximov, P. Kop'ev, Zh. Alferov, U. Richter, P. Werner, U. Gö1sele, J. Heydenreich. *Electron. Lett.*, 30, 1416 (1994). Low threshold, large T_0 injection laser emission from (InGa)As quantum dots, Electron. Letters, 30, 1416, (1994).

[6] R. Suris and L. Asryan, "Quantum-Dot Laser: Gain Spectrum Inhomogeneous Broadening and Threshold Current", *Proceedings* of SPIE's 1995 International Symposium on Optoelectronic, Microphotonic & Laser Technologies. PHOTONICS WEST'95, 4-10 February 1995. San Jose, California USA, v. 2399, pp.433-444.

L. Asryan and R. Suris, "Linewidth Broadening and Threshold Current Density of Quantum-Box Laser", *Proceedings* of International Symposium Nanostructures: Physics and Technology. June 20-24, 1994, St.Petersburg, Russia. pp. 181-184.

[7] R. Suris and S. Shtofich, Role of impurities in the appearance of multifrequency emission from injection semiconductor lasers, *Sov. Phys.- Semiconductors(July 1983)* **17**, no 7, 859

[8] L. Esaki and R. Tsu, Superlattice and Negative differential conductivity, *IBM J. Res. and Dev.* **14,** 1970, p 61

ARCHITECTURES FOR NANO-SCALED DEVICES

LEX A. AKERS
Center for Solid State Electronics Research
Arizona State University
Tempe, AZ 85278-5706

1. Introduction

The desire for higher performance, low cost electronic systems seem insatiable. We have witnessed in the latter part of the 20th century the introduction of microcomputers and cellular telecommunications systems. These systems have undergone continued enhancements in computational power, memory, and special features. The demand for further system improvements, lower power consumption, anywhere-anytime access to data and communications, multimedia, and portable personalized digital assistants will continue[1]. Also many real-time applications such as vision and speech recognition, robotics, and numerous other interactive control and signal processing applications will require hundreds of gigaflops of processing speed[2]. The scaling of device feature sizes into nanometer dimensions can conceivably allow systems made with these components to fulfill the performance improvements desired.

The advantages of scaling devices to smaller sizes are overwhelming since the area per function, the energy needed to switch a device, and the energy needed to be stored to represent information scale with the physical size of the device. However, the common assumption that improvements in device performance will automatically translate into improved system performance is not true. We believe one major challenge to implementing such systems using nano-scaled devices is finding appropriate system architectures to host these devices. The modularity and locality of the architecture and the length of the connections between devices and between subsystems determine if the improvement in device performance predicted by scaling theory is reflected in improvement in system performance. A discussion of these issues is presented, followed by recommendations on the types of architectures that allow designers to better capitalize on the characteristics of nano-scaled devices.

2. Architectural Issues

Contemporary computer architectures do not scale well. The total length of connections needed to interconnect devices in these architectures is rapidly increasing as design rules are scaled[3]. Figure 1 shows the predicted total interconnection wire length summed from all layers as the design rules scale from .35μm to .07μm. As will be discussed below, this increase in connection length causes an interconnect capacitance increase. The system clock has to drive much of this additional capacitance. While

S. Luryi et al. (eds.), Future Trends in Microelectronics, 209–214.

clock drivers have been designed to handle large capacitance loads[4], there is a power dissipation penalty. As a compromise, the future clock speeds[3] have been slowed below what one would calculate using scaling theory. The traditional Von Neumann architecture is not allowing full use of the device speed-up with scaling.

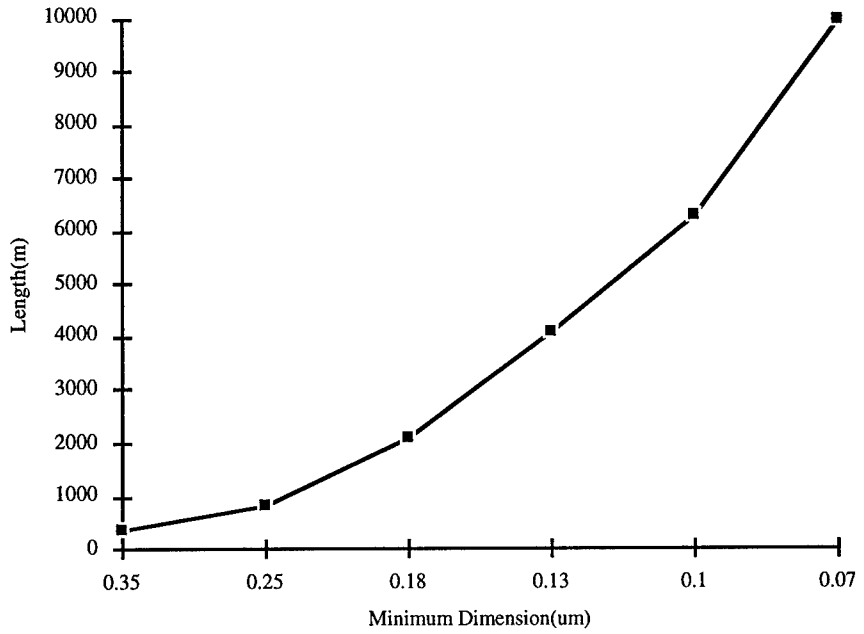


Figure 1. Total interconnection length vs. minimum feature size

Another characteristic of traditional architectures implemented with CMOS devices is as the design rules are scaled into nanometer dimensions, the device switching speed becomes dominated by connection capacitance rather than device capacitance. Device capacitance is composed of two terms. The first term is the gate capacitance. As device size is scaled, the gate area decreases by the square of the scaling factor. However, the oxide thickness also decreases resulting in a $1/\alpha$ decrease in gate capacitance, where α is the scaling factor. Twice the gate capacitance value was used in these calculations since the load circuit element assumed is a complimentary pair inverter. The second factor is the drain diffusion capacitance. The drain diffusion capacitance is composed of a junction capacitance and a periphery capacitance. The drain area was assumed to scale down as the square of the design rules. Values of .1ff per μm^2 for the junction capacitance and .2ff per μm for the periphery capacitance were used in the calculations.

For the connection capacitance calculation, a capacitance per unit area of .03ff per μm was used. This is an average of routing capacitance between various metal and metal to poly and poly to substrate levels. The average connector width was held constant at 1μm for all design rules. A constant connection line width was used since while line widths will scale, larger connection capacitances will require some wires to increase in width to handle the current needed to charge the large capacitor loads. As

shown in Fig. 2, the connection capacitance becomes the dominate capacitance in device to device connections for small design rules. While increasing device drive will overcome this increased load, the cost is rapidly increased interconnect power dissipation.

Power dissipation results from both dynamic and static operations. The dynamic power dissipation is reduced per device as the device is scaled since the drain and gate capacitances and the operating voltage are reduced. The increase in clock frequency will in part counter some of this reduction. The total power dissipation for all the devices on the die increases slightly as more and more devices are placed on the die. However as shown in Fig. 3, the power dissipated in the connection capacitances and the power dissipated in driving the outputs consume most of the power. The calculation assumes the output drivers source a 50 pf load, and that only 10% of all the gates and I/O switch at one time. Total power dissipation of over 50 watts per package, an unacceptable power dissipation level for air-cooled and portable systems, will occur. In these calculations the static power dissipation was neglected. However, small size effects and low threshold voltages will result in insufficient swing to fully turn a device off and hence allow large subthreshold currents to flow in the device in the off state. This non-negligible off current will increase the static power dissipation to unacceptable levels for portable computers and telecommunication systems.

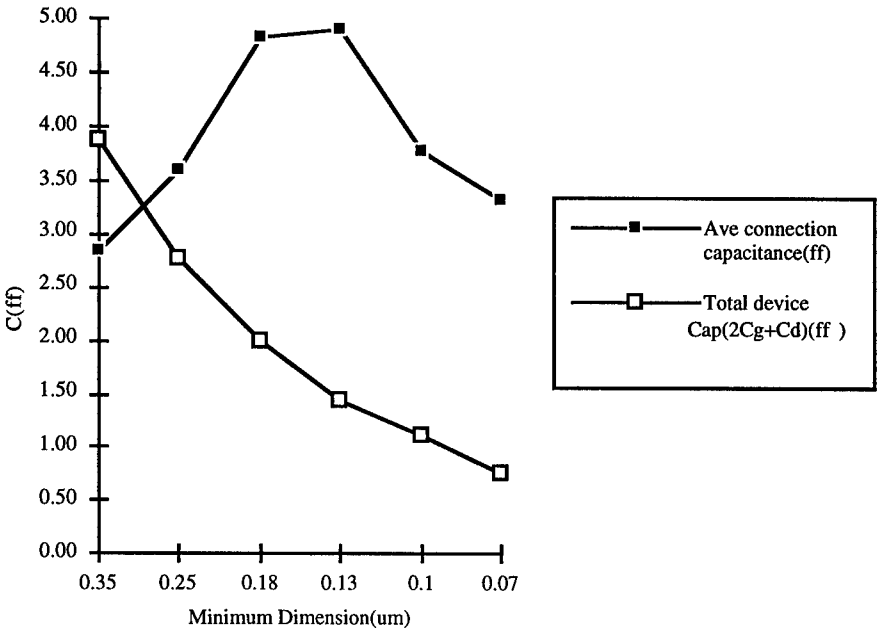


Figure 2. Connection and device capacitance vs. minimum feature size

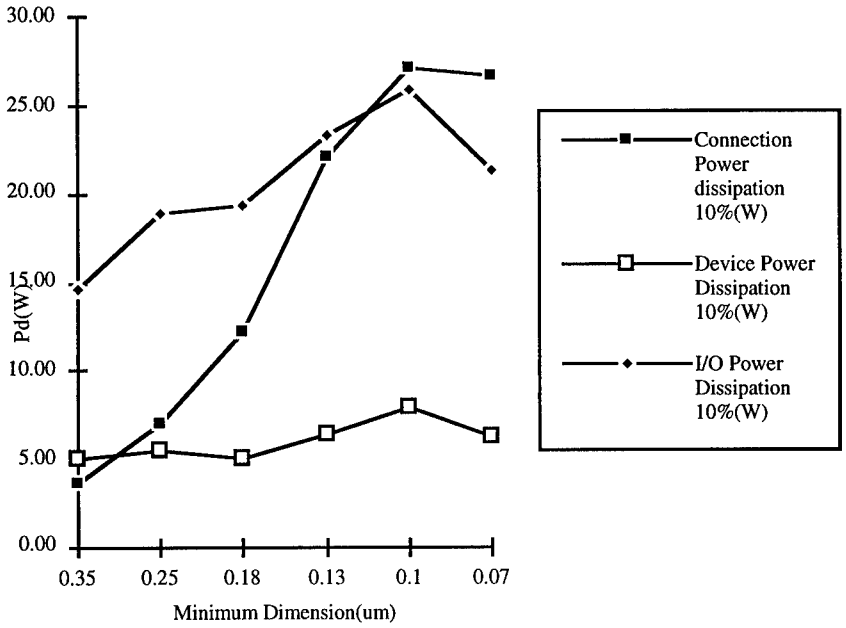


Figure 3. Power dissipation vs. minimum feature size

3. Locally Interconnected Architectures

To overcome these problems, we propose the use of locally interconnected architectures such as cellular automata[5], cellular neural networks[6], and locally adaptable neural networks[7]. The total connection capacitance for these architectures is lower than that found in traditional architectures. This results from the locality of computation and data in these architectures. Locally interconnected architectures feed information to nearest neighbors and have memory stored at the computation site instead at a distance site as in Von Neumann architectures. Therefore, these systems don't require long information busses. While these architectures have been proven to do universal computation, they are most suited for specialized computations such as image processing, feature detection, speech and vision recognition, sensor fusion, and various other real-time applications. These architectures also offer the advantage of parallel computation. While the potential of parallel computation has been known for some time, its practical application has been difficult to achieve. However, recent applications have demonstrated its capability[8]. Parallel operation allows reducing the clock speed and the supply voltage making significant reductions in power dissipation while simultaneously keeping the overall system speed constant.

As discussed earlier, I/O is a real power consumer. Functionally partitioned[9] and locally interconnected architectures allows the number of I/O to be much smaller than

the predicted I/O needs in conventional architectures[3]. This difference is shown in Fig. 4. These architectures provide system performance that will improve with device performance while keeping the overall power dissipation, shown in Fig. 5, down to acceptable levels.

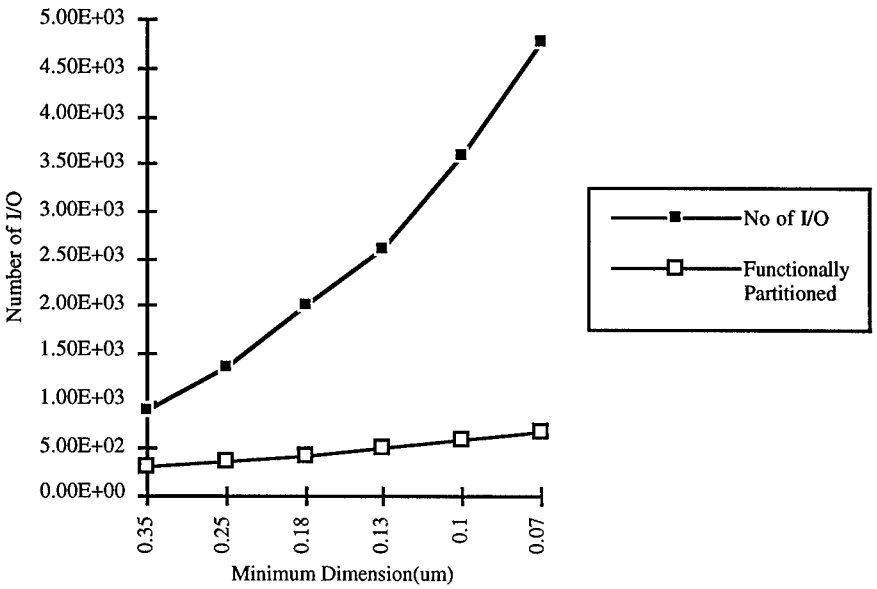


Figure 4. I/O for traditional and locally interconnected architectures vs. minimum feature size

4. Conclusion

Scaling will allow devices to shrink to nanometer dimensions and result in vast improvements in their performance. However, to translate this improvement into system improvements will require system architectures which use local rather than global information and communication for computations.

5. References

1. Pahlavan, K, and Levesgue, A. (1994) Wireless Data Communications, *Proceedings of the IEEE* 82, 9, 1398-1430.
2. Koch, C. (1995) Smart Vision Chips: An Overview, An Introduction to Neural and Electronics Networks, Academic Press, 315-333.
3. The National Technology Roadmap for Semiconductors, (1994) SIA Report.
4. Weste, N., and Eshraghian, K., (1994) *Principles of CMOS VLSI Design*, Addison-Wesley.
5. Biafore, M., (1994) Cellular automata for nanometer-scaled computation, *Physica D* 70, 415-433.
6. L. O. Chua, L. Yang, and K. Krieg, (1991) Signal Processing Using Cellular Neural Networks, *J. VLSI Signal Processing* 3, 25.

7. Akers, L.A., Walker, M., Ferry, D.K., and Grondin, R., (1989) A Limited-Interconnect, Highly Layered Synthetic Neural Architecture, *VLSI for Artificial Intelligence*, Kluwer Academic Press.
8. Matsuzawa, A. (1994) Low-Voltage and Low-Power Circuit Design for Mixed Analog/Digital Systems in Portable Equipment, *IEEE J. Solid-State Circuits* 29, 4, 470-480.
9. Ferry, D.K., Akers, L.A., and Greeneich, E., (1988) *Ultra Large Scale Integrated Microelectronics*, Prentice-Hall.

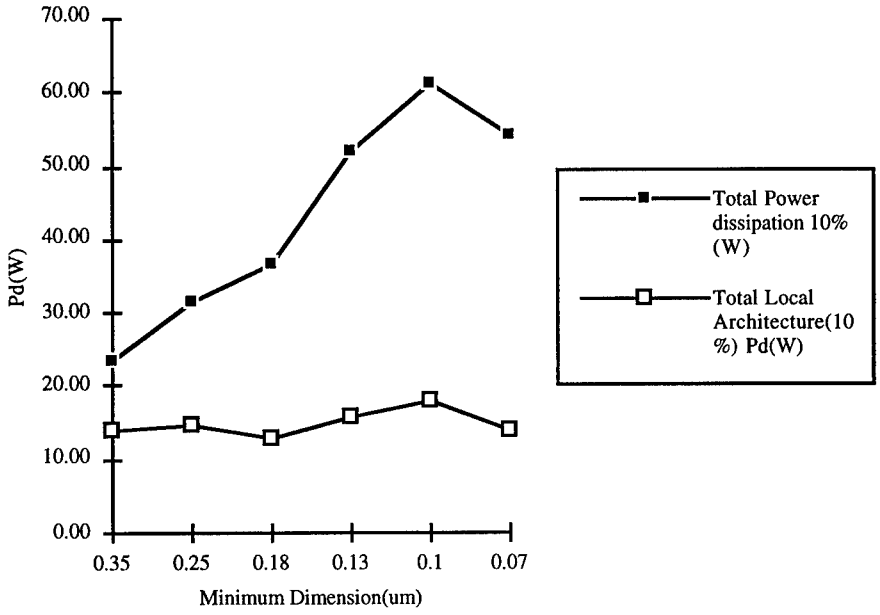


Figure 5. Total power dissipation for the traditional and locally-interconnected architectures vs. minimum feature size

SIMULATING ELECTRONIC TRANSPORT IN SEMICONDUCTOR NANOSTRUCTURES

K. HESS, P. VON ALLMEN, M. GRUPEN AND L.F. REGISTER
The Beckman Institute
University of Illinois
405 N. Mathews Ave.
Urbana, IL 61801

1. Introduction

Simulation of semiconductor devices has reached some maturity for device structures that can be described by the system of device equations given by Shockley [1, 2]. High energy transport, including hot electron effects such as impact ionization and gate currents, can be correctly simulated by full band Monte Carlo approaches [3] (solving Boltzmann-type equations) and has also matured into the realm of engineering; the missing pieces being mainly standardization and numerical efficiency. Correspondingly, commercial packages, which solve the Shockley equation system (even in three dimensions) and feature full band Monte Carlo post processors, are available or are in the final stages of development. Complex full band Monte Carlo device codes are also available [4].

The treatment of (abrupt) heterojunctions in devices has not yet matured to the desirable degree and is almost certainly required to be understood in the future in great detail. One can make a case, and current developments point strongly to it, that heterolayers must more and more replace the dilute donor/acceptor doping configurations as the sizes of devices decrease toward the typical donor/acceptor spacing. The simulation of abrupt heterostructures invariably involves quantum mechanics and solutions of the Schrödinger equation. The quasi-two-dimensional electron gas of the metal-oxide semiconductor transistor is very well investigated and understood [5] and work on quantum wires and dots is in progress.

Complex problems are involved, however, in the coupling of these quantum regions to the classical "Shockley regions". This coupling needs to be

S. Luryi et al. (eds.), Future Trends in Microelectronics, 215–225.

accomplished on various levels. On the physical side much progress has been made with Landauer-Büttiker type of coupling which in many instances is a generalization (and specialization with regard to dimensionality) of Bethe's thermionic emission theory. From a numerical point of view the quantum region necessitates multigrid approaches, where in the simpler cases the grid describing the quantum region is typically reduced in dimension (by one) compared to the reservoirs. In this way one can construct a multiscale approach that combines classical and quantum regions in one simulation entity. We have performed such a task for semiconductor quantum well lasers diodes and will report this as an example below. While these simulations deal mainly with current and resistance concepts, the quantization also influences the concept of capacitance and inductance and examples of atomistic features of quantum dot capacitance have been given [6]. The knowledge of electronic band structures also becomes increasingly important as shown by full band Monte Carlo simulations, calculations of optical transitions in lasers, density of states calculations, etc.

The next step in scaling down, however, involves the mesoscale and molecular scale and requires still more elaborate quantum mechanical methods to obtain insight into the electronic structure. One might, for example have in mind a quantum dot containing only a very limited number of constituting atoms, a tunneling tip, etc. At this scale the dynamics of atoms also assumes a special importance as can be seen from experiments of force and tunneling microscopes that move and switch atoms. The method of Car and Parrinello [7] emerges as a powerful tool to simulate this scale and in turn can feed back to the bigger scales giving information about surface and interface structure and dynamics.

A complete coupling of these methods to understand questions on all scales is still in the future. However, some coupling of those methods has already been achieved and representative examples are given in the following sections.

2. Semiconductor Laser Simulation–A Multiscale Problem

Laser diodes can be useful examples for studying nanostructures such as quantum wells, wires, and dots coupled to classical regions. The photons produced in these active regions convey information about the nanostructure. Careful simulation of the optical output shows the multiscale nature of nanostructures, which depend on the quantum regions as well as their coupling to surrounding bulk regions.

To demonstrate, we have simulated the quantum well laser structure shown schematically in Figure 1. Current pumping requires carrier injection over the cladding and separate confinement heterostructure (SCH).

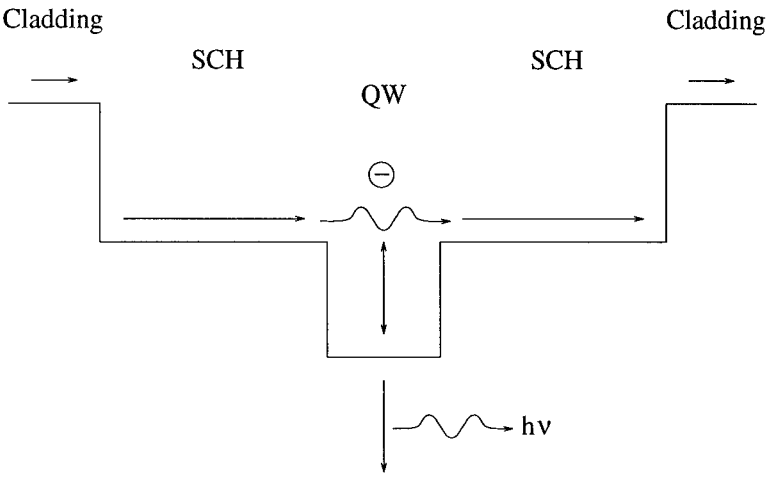


Figure 1. Schematic conduction band edge for a quantum well laser diode

These are approximately classical 3-D regions, and transport can be modeled with the Shockley equations. Carriers move ballistically over the abrupt heterojunctions and thermalize in roughly a mean free path. Since the quantum well (QW) can be on the order of the mean free path or much smaller, injected carriers may traverse the well, reflect, resonate above the well, or inelastically scatter into bound states. The coupling of this quantum region to the classically behaving diode is a complicated mesoscopic problem that must be treated by a Bethe-Landauer type of approach. Bulk drift-diffusion is coupled in our simulation [8] to ballistic injection into 3D (continuum) states above the well. Continuum carriers can then transfer to 2D "bound states" through a net capture rate given by [9]:

$$U_{\text{cap}} = s_{\text{3D,2D}} \left[\left(\int_{\text{bound}} \frac{g_{\text{2D}}}{L_{\text{qw}}} dE \right) n_{\text{3D}} - n_{\text{2D}} n_{\text{3D}} \right] \left[1 - \exp\left(\frac{F_{\text{2D}} - F_{\text{3D}}}{kT/q} \right) \right] \tag{1}$$

where the density of states g_{2D} is integrated only over energies within the well, and $F_{\text{2D(3D)}}$ is the quasi-Fermi level for bound (continuum) carriers.

The sensitivity of the laser output to capture and the scattering parameter $s_{\text{3D,2D}}$ is shown by solving the transport model self-consistently. Figure 2 shows a set of small signal modulation responses for an 80 Å $In_{0.2}Ga_{0.8}As$ quantum well laser. The curves are labeled by capture times, defined as $\tau_{\text{cap}} = [s_{\text{3D,2D}} \int_{\text{bound}} g_{\text{2D}}/L_{\text{qw}} dE]^{-1}$. As τ_{cap} increases, the responses roll off at low frequencies. This is due to carriers that are not captured in the well but instead are deposited as minority carriers in the SCH [10]. Their ef-

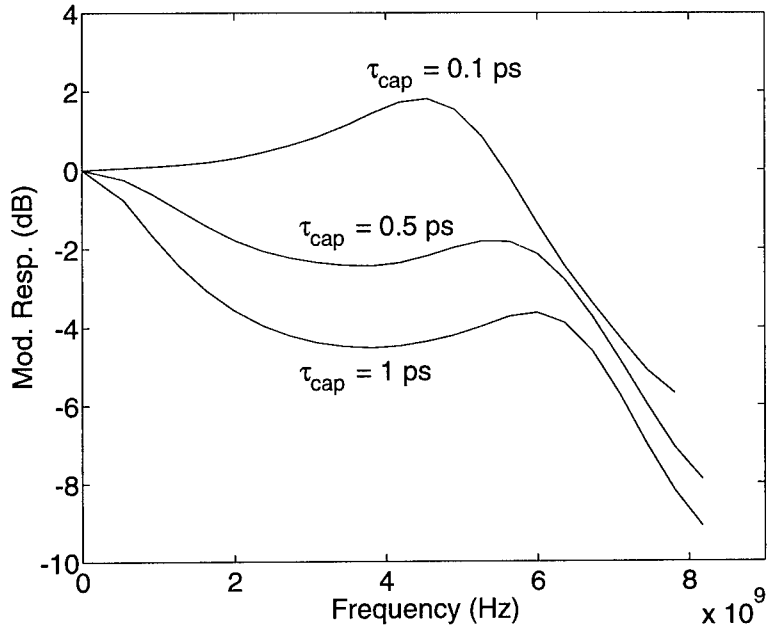


Figure 2. Modulation of a laser (80 Å $In_{0.2}Ga_{0.8}As$ well; 1500 Å $Al_{0.1}Ga_{0.9}As$ SCH) using different $\tau_{cap} = [s_{3D,2D} \int_{bound} g_{2D}/L_{qw} dE]^{-1}$. Bias is 10 mA; output power is 4.63 mW.

fect can only be calculated through multiscale simulation, coupling carrier capture with bulk transport in the SCH.

Comparing Figure 2 with experiment puts some limits on τ_{cap}. A laser with this quantum well and comparable SCH should show negligible rolloff up to powers of 27 mW [11]. Therefore, simulation shows that capture times for this quantum well should be subpicosecond for electrons and much less for holes, thus predicting that the laser output depends critically on an accurate calculation of capture efficiency.

Initial attempts to model carrier capture in quantum wells relied on classical theory [12]. Later, via *Fermi Golden Rule* based calculations, it was demonstrated that quantum interference effects in the free as well as bound carrier states could, in principle, significantly affect capture [13, 14, 15]. However, inelastic scattering is inherently accompanied by a loss of phase coherence in the carrier wave functions, and strong inelastic scattering is required for efficient carrier capture. The inability to directly model this loss of phase coherence and the corresponding transition between quantum and classical transport in such *Golden Rule* based calculations has lead to ambiguities in the reported capture rates [13, 14, 15] and perhaps overesti-

mates of the effects of quantum resonances and anti-resonances on capture under lasing condition. Simulation of capture on a more sophisticated level appears therefore necessary if one wants to describe in detail the transitions between classical and quantum regions.

To model the physics of relatively strong inelastic scattering on transport in mesoscopic structures, we turn to a simulation technique we refer to as "Schrödinger equation (based) Monte Carlo" (SEMC) [16, 17] which represents the next level in our multiscale approach. SEMC is specifically formulated to allow first-principles simulation of dissipative quantum transport and is rigorously quantum mechanical. However, the numerical algorithm has much in common with semiclassical Monte Carlo methods. In brief, for carrier-phonon interactions, a set of Schrödinger equations is solved simultaneously (and deterministically) for the carrier, with the individual equations corresponding to the discrete initial or "trunk" state and various final or "branch" states of the phonon system, with energies separated by plus or minus the phonon energy for phonon absorption and emission, respectively. Only the non-local coupling potentials between the trunk and branch states are obtained by Monte Carlo sampling of (the spatial correlation functions of) the true carrier-phonon interactions, as describe in detail in [16]. In this way, probability current, provided to the trunk state in the boundary conditions, flows from the trunk state to branch states via the coupling potentials, and, of critical importance, the trunk state is altered self-consistently by this interaction. Here, to continue with the laser system, a version of SEMC currently being optimized for geometries that vary in 1-D only and a toy short-range interaction to low energy phonons, much like acoustic phonon scattering, with a variable coupling strength are employed. However, quantitatively accurate calculations of polar-optical-phonon scattering have been demonstrated in 2-D geometries [16], and the formalism is amenable to simulation of other phonon scattering process and carrier-carrier scattering as well. Figure 3 shows a SEMC calculation of electron capture by a 10 nm wide, 237 meV deep quantum well, with the bound states of the well clearly exhibited among the branch states. Without scattering, the trunk electron state exhibited resonant (near 100%) transmission over the well; however strong inelastic scattering reduces the electron phase coherence and, thus, destroys the resonance condition. Indeed, the most probable outcome in this example is reflection from the well, followed by capture.

To model such losses of phase coherence more efficiently, if less rigorously, complex "absorbing" potentials can be used within Schrödinger's equation in lieu of the branch states to represent capture. Figure 4 exhibits an example of such a calculation for the same structure as Figure 3, and, again, the destruction of the quantum resonance is apparent. While the

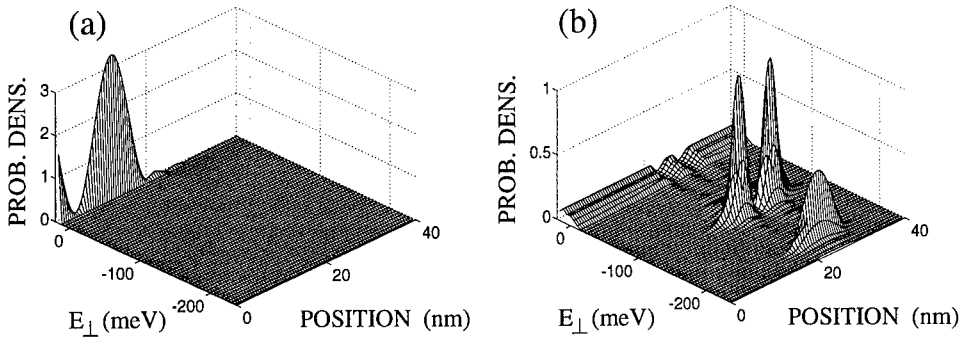


Figure 3. Probability density obtained by SEMC for the quantum well system of Figure 1 in the (a) normally incident and nominally resonant trunk state and (b) corresponding branch (scattered) states with strong inelastic scattering. $E_\perp$ is the component of energy associated with motion normal to the plane of the quantum well. The well is located between 20 nm and 30 nm.

"scattering rates" required may appear high, the transmission probabilities shown here may be, if anything, large. For example, assuming only a bulk-optical-phonon-like scattering rate of 0.01/fs for electrons above the well, would suggest that a significant majority of the electrons would be transmitted across the well. However, the internal quantum efficiency of quantum well lasers is typically quite high, often approaching 100% [18], which implies that few electrons are transmitted beyond the quantum well(s). Further, direct comparison with the capture rates mentioned above is difficult because of the previously mentioned ambiguities in the reported capture rates and the possibility that capture under lasing conditions is dominated by carrier-carrier scattering from bound carriers, which would best be characterized by capture cross sections instead of rates. Nevertheless, even at a rate of 0.01/fs, a significant drop in the differential capture probability with increases in the scattering rate becomes apparent for low energy electrons in this system (a 27% drop for 14 meV electrons, approximately the average component of kinetic energy of the carriers in any one direction at room temperature). At a scattering rate of 0.05/fs the net capture probability for low energy electrons approaches saturation. Such rates are not uncommon for holes and even for electrons in II-VI compounds. Further decreases in the transmission probability are compensated for predominately by increases in the reflection probability, even at the resonance energy. Such behavior, of course, is not possible to model via *Golden Rule*-based calculations.

The preceding simulation examples clearly show the various layers of simulation techniques that are necessary for a detailed understanding of quantum well laser diodes. While we can deal with the various levels and also have a physical theory and numerical method of connecting them,

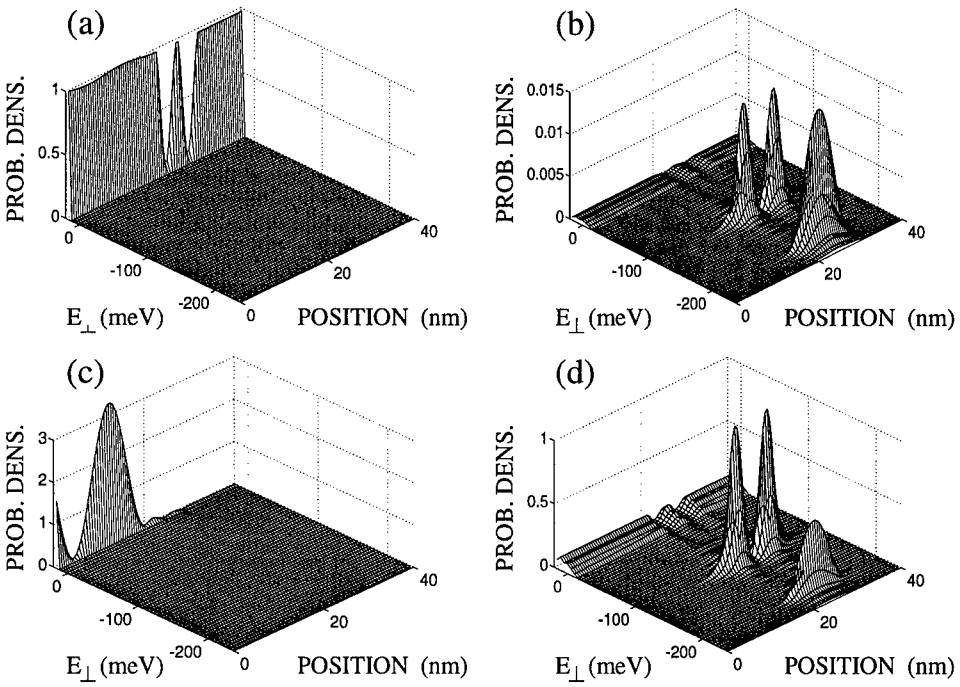


Figure 4. (a) Probability density in the nominally resonant incident electron state as a function of position and imaginary absorbing potential $-i\hbar R/2$, where R is the scattering rate, and (b) transmission, (c) reflection and (d) transmission probabilities as a function of incident energy and scattering rate R. The quantum well is located between 20 nm and 30 nm in (a).

a complete and connected numerical treatment currently is beyond the capability of even the largest computational resources. Nevertheless, even this approach leaves many questions untouched. For example, what are the effects of alloy disorder at quantum well boundaries, how are electronic states really connected across semiconductor interfaces, what is the capture rate for quantum dots instead of wells, etc.? Resolution of such questions necessitates a still finer resolution of feature sizes down to the molecular and atomic scale. Examples for the mesoscale can be found in the literature [6]; the approach to the atomic scale is described next.

3. Atomic-Scale Dynamics

As feature sizes approach the atomic scale, we can no longer rely on the effective mass approximation and other such continuum approximations. Rather, ions are modeled by pseudopotentials that determine their electronic orbitals. Thus, with present levels of computing power, this down-

ward scaling restricts us to simulation of only relatively small systems. In most cases, however, this restriction is not prohibitive since only a limited number of atoms need be considered in such atomic-scale systems. Further, at this scale, defects and disorder can be treated in a straightforward manner, and dynamic effects related to the motion of the ions, such as their thermal vibration, can be studied at the most basic level.

In our simulations, electrons are described by density functional theory, and the motion of the ions obeys Newton's equations with forces given by the gradient of the total energy of the system. Both the electrostatic repulsion between the ions and the contribution from the electronic clouds are included. In this way, the many-body effects for the electrons are greatly simplified. This method has been proven to be extremely reliable for calculation of ground state properties such as the total energy and the equilibrium configuration of the ions. For properties, such as optical absorption and transport, that involve excited electronic states, extensions of the method, e.g. the Generalized Gradient Approximation, greatly improve the density functional theory.

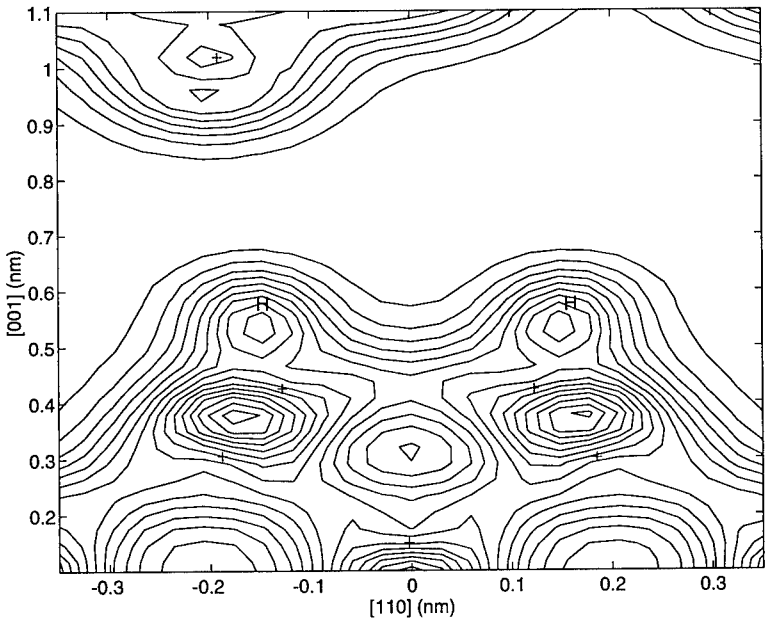


Figure 5. Total electronic density for the tip-surface system. The crosses indicate the positions of the silicon atoms.

Ultimately, the numerical problem reduces to solving the one particle Schrödinger equation and Poisson's equation self-consistently while mini-

mizing the total energy of the system. Various sophisticated methods have been devised for solving both equations [20]. We are currently concentrating on a minimization approach that appears to be quite natural to density functional theory. We have tested several minimization methods including the linear search algorithm and the conjugate gradient method, and we have considered the functional form for the total energy that requires explicit orthonormalization of the electronic orbitals as well as the form that avoids this constraint. We have found that a modification of the simple linear search procedure for the functional form requiring explicit orthonormalization is the most efficient algorithm for the systems of moderate size that we have studied so far. In this method, part of the total energy gradient is integrated exactly over one iteration step. However, for larger systems we expect that the functional for non-orthonormal orbitals will be more efficient.

To study the dynamic properties, we use the method introduced by R. Car and M. Parrinello [7]. They have shown that it is possible to write equations of motion for the coupled system of the electrons and the ions such that the electrons oscillate around their instantaneous ground state while the ions obey Newton's equations of motion. In other words, the system oscillates close to the Born-Oppenheimer surface [19], and it is not necessary to compute the electronic ground state for each ionic configuration. This method is very efficient, and a vast literature has demonstrated its applicability to a broad range of problems.

Combined, this set of first-principles tools has been tested for bulk materials, and excellent agreement is routinely achieved for the equilibrium lattice constant, the bulk modulus, and the bulk phonon frequencies computed both within the frozen phonon approximation and with molecular dynamics. We also obtain a reconstruction of the clean and of the H-passivated silicon (001)-(2×1) surface in agreement with recent highly converged results in the literature.

As an example of such atomic-scale calculations, the interaction of a STM tip with a H-passivated silicon surface has been simulated. The surface is modeled as a slab of eight layers of silicon atoms with the atoms in the four upper layers relaxed to their equilibrium positions. The STM tip is represented by a single silicon atom adsorbed to the back of the slab. Figure 5 shows the distribution of the total electronic density before the thermostat is switched on. The hydrogen passivation induces a reconstruction of the surface with the formation of symmetric dimers in the [110] direction. When the thermostat is switched on (the bath is set at room temperature), the atoms start to oscillate, and Figure 6 shows the corresponding variation of the tunneling barrier height between the tip atom and the surface. This fluctuation is larger than the thermal energy spreading, and a

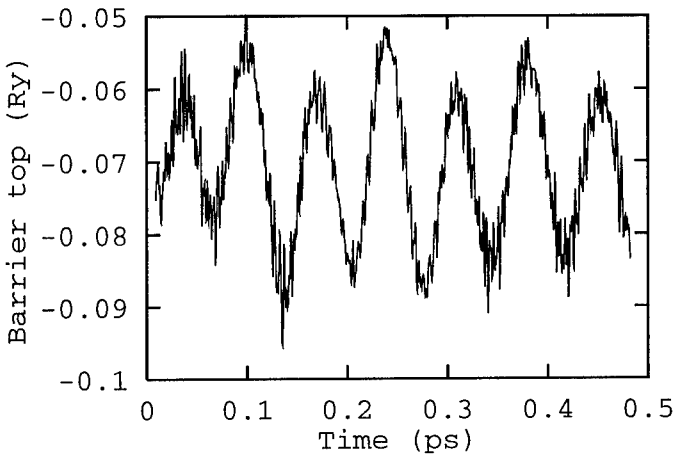


Figure 6. Fluctuation of the top of the tunneling barrier due to the thermal motion of the tip and surface atoms.

corresponding major fluctuation in the tunneling current results. Because this fluctuation occurs at very high frequency, it can not be measured in a standard STM experiment. However, it is expected that these fluctuations have a significant effect on dynamic behavior like the hydrogen-desorption mechanism. This example demonstrates the sophisticated simulation tools that are needed when the atomic scale is approached. It also shows that to achieve a detailed understanding of seemingly straightforward experiments, such as measurement of STM tunneling currents, may require high level simulation approaches.

Acknowledgement

This work was supported by the Office of Naval Research (N00014-89-J-1470, N00014-92-J-1519 KH) and by the Army Research Office (DAAL03-92-G-0271).

References

1. Streetman, B.G. (1980) *Solid State Electronic Devices*, Prentice-Hall, Englewood Cliffs, NJ.
2. Hess, K. (1988) *Advanced Theory of Semiconductor Devices*, Prentice-Hall, Englewood Cliffs, NJ.
3. Bude, J. (1991) Scattering Mechanisms for Semiconductor Transport Calculations,

Monte Carlo Device Simulation: Full Band and Beyond, ed. K. Hess, Kluwer Academic Publishers, Norwell, Mass.

4. Laux, S.E. and Fischetti, M.V. (1988) Numerical Aspects and Implementation of the *DAMOCLES* Monte Carlo Device Simulation Program, *Monte Carlo Device Simulation: Full Band and Beyond*, ed. K. Hess, Kluwer Academic Publishers, Norwell, Mass.
5. Ando, T., Fowler, A.B., and Stern, F. (1982) Electronic Properties of Two-Dimensional Systems, *Review of Modern Physics*, **54**, 466.
6. Macucci, M., Hess, K., and Iafrate, G.J. (1995) Simulation of Electronic Properties and Capacitance of Quantum Dots, *J. Appl. Phys.*, **77**, 3267-3276.
7. Car, R., and Parrinello, M. (1985) Unified Approach for Molecular Dynamics and Density-Functional Theory, *Phys. Rev. Lett.* **55**, 2471-2474.
8. Grupen, M., Ravaioli, U., Galick, A., Hess, K., and Kerkhoven, T., (1994) Coupling the Electronic and Optical Problems in Semiconductor Quantum Well Laser Simulations, *Proc. SPIE OE/LASE Conf.*, **2146**, Los Angeles, CA, 133-147.
9. Grupen, M., Kosinovsky, G., and Hess, K. (1993) The Effect of Carrier Capture on the Modulation Bandwidth of Quantum Well Lasers, *1993 International Electron Device Meeting Technical Digest*, 23.6.1-23.6.4.
10. Grupen, M. and Hess, K. (1994) Self-Consistent Calculation of the Modulation Response for Quantum Well Laser Diodes, *Appl. Phys. Lett.* **65**, 2454-2456.
11. Nagarajan, R., Mirin, R.P., Reynolds, T.E., and Bowers, J.E. (1993) Experimental Evidence for Hole Transport Limited Intensity Modulation Response in Quantum Well Lasers, *Electron. Lett.* **29**, 1688-1690.
12. Shichijo, H., Kolbas, R. M., Holonyak, N., Dupuis, R. D., and Dapkus, P. D. (1978) Carrier Collection in a Semiconductor Quantum Well, *Solid State Communications* **27**, 1029.
13. Brum, J. A. and Bastard, G. (1986) Resonant Carrier Capture by Semiconductor Quantum Wells, *Phys. Rev. B* **33**, 1420-1423.
14. Sotirelis, P. and Hess, K. (1994) Electron Capture in GaAs Quantum Wells, *Phys. Rev. B* **49**, 7543-7547.
15. Preisel, M. (1994) *Carrier Capture and Carrier Kinetics in Biased Quantum Well Devices*, Tele Danmark Research, Hørsholm, Denmark.
16. Register, L.F. and Hess, K. (1994) Numerical Simulation of Electron Transport in Mesoscopic Structures with Weak Dissipation, *Phys. Rev. B* **49**, 1900-1906.
17. Hess, K., Register, L.F., and Macucci, M. (1994) Toward a Standard Model in Nanostructure Transport Problems Including Dissipation, *Proceedings of the 2nd International Symposium on Quantum Confinement: Physics and Applications* **94-17**, 3-17.
18. Zory, P.S. (1993) *Quantum Well Lasers*, Academic Press, San Diego.
19. Pastore, G., Smargiassi, E., and Buda, F. (1991) Theory of *Ab Initio* Molecular-Dynamics Calculations, *Phys. Rev. A* **44**, 6334-6347.
20. Payne, M.C., Teter, M.P., Allan, D.C., Arias, T.A., and Joannopoulos, J.D. (1992) Iterative Minimization Techniques for *Ab Initio* Total-Energy Calculations: Molecular Dynamics and Conjugate Gradients, *Rev. of Mod. Phys.* **64**, 1045-1097.

MONTE CARLO SIMULATION FOR RELIABILITY PHYSICS MODELING AND PREDICTION OF SCALED (100 NM) SILICON MOSFET DEVICES

R. B. HULFACHOR, J. J. ELLIS-MONAGHAN*, K. W. KIM,
and M. A. LITTLEJOHN
North Carolina State University
Department of Electrical and Computer Engineering
Raleigh, NC 27695-7911
(* Current Address: IBM, Essex Junction, VT 05452)

1. Introduction

Since the early 1970's, silicon integrated circuit technology has been propelled by continual and successful efforts to reduce the active channel length of MOSFET devices [1 - 3]. This exercise in scaling provides the framework that has produced increases in the density of devices on a chip, increases in device frequency response and operating speed, and increases in the precision required to achieve more complex systems with greater functionality and performance. Today, devices with channel lengths well below 100 nm have been produced in many research laboratories, and the downward scaling trends of the past twenty years are expected to persist at the same pace until at least the 40 nm generation in manufacturing [1], or about the year 2015. This level of technology is expected to correspond to 128 GBit DRAMs and 28 Ggate microprocessors with five times the clock frequency and 1/9 the power consumption of today's devices, all operating at a power supply voltage of around 0.5 V.

Historically, silicon MOSFETs designed for increased performance through scaling and scaling-related approaches to drain and channel engineering have also produced compromises in device reliability. It was anticipated that as the effective channel length was scaled below the electron mean free path (L_{eff} < about 200 nm), non-local transport effects (such as velocity overshoot and quasi-ballistic transport) would actually enhance device performance. Furthermore, by scaling the power supply voltage below the Si/SiO_2 interface barrier height (i.e., $qV_{DD} < \emptyset_B = 3.1$ eV), it was expected that many device reliability problems could be suppressed or eliminated. A new regime for device operation was predicted where electron energy distribution functions would be "cooler" than those deduced from models based on the local electric field. However, drain current degradation was observed in recent experiments on 150 nm floating gate MOSFETs biased with a drain voltage as low as 1.5V, and no discontinuities in degradation were observed as the voltage was reduced below 3.1 V [4]. These experiments confirmed that the degradation was due to non-tunneling current, via carrier injection over the interface barrier and into the oxide. For such low bias, the channel electrons cannot gain enough energy from the drain electric field alone to surmount the barrier; thus, those rare electrons that do inject must experience some type of "energy

S. Luryi et al. (eds.), Future Trends in Microelectronics, 227–236.

gaining" interactions to obtain the required energy. Physical mechanisms that produce this effect are not fully understood, but Monte Carlo simulations indicate that electron-electron interactions near the drain edge play a significant role [6]. For 100 nm devices biased at low voltages, high-energy phenomena such as injection into the oxide are extremely rare and their effects are difficult to measure experimentally. Therefore, attempts to understand fundamental limits of silicon technology for next generation devices will need to rely heavily on accurate simulation and modeling methods that can: (a) identify those phenomena that occur at high energies; (b) predict the consequences of design and processing on device performance, and (c) analyze device and circuit reliability and operational lifetime.

2. Reliability Simulation, Modeling, and Prediction for Next-Generation Devices

Figure 1 illustrates the general technology and design environment for down scaling of MOSFET devices. This schematic shows relationships between simulation, modeling, and reliability evaluation/prediction (right side) to basic elements of the scaling process, manifested through process technology, device structure choices, and experimental evaluation (left side). In this context, it is important to recognize the many challenges for process technology, device structure choices, and experimental methodology required to fabricate 100 nm MOSFETs [7]. For example, use of highly-controlled processes that achieve necessary dopant distributions for critical drain and channel engineering regions can also accentuate limits to device scaling and reliability [8]. Thus, the difficult physical levels of MOSFET scaling on the left side of Fig. 1 must be ameliorated by more sophisticated process and simulation models that are valid for devices with critical dimensions of 100 nm (and below). Moreover, difficulties in measuring critical phenomena such as gate and substrate currents in devices with low bias voltages will hinder the identification and development of physical understanding of reliability issues [9]. Physical descriptions of non-local, high energy transport phenomena will transcend the use of lucky electron models which dictate that degradation mechanisms conform to the maximum value of lateral electric field [10]. High energy tails in electron energy distributions cannot be predicted by lucky electron or hydrodynamic models [7,11]; thus, Monte Carlo methods will likely be essential to support both device and process simulation and modeling for dimensions in the 100 nm regime [12].

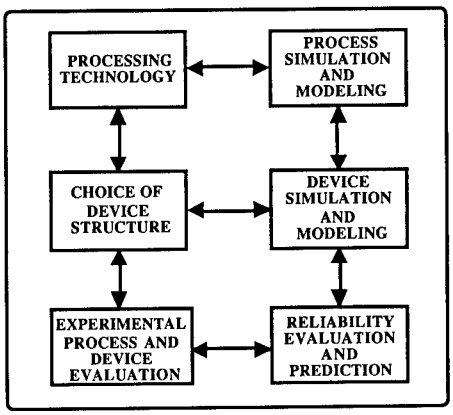


Figure 1. Schematic of the relationship of process technology (left side) to device and process simulation, modeling, and reliability prediction.

These emerging challenges to resolving process technology and device characterization issues in 100 nm devices lead us to the conclusion that comprehensive simulation, modeling, and reliability prediction will play a significant role in guiding

future experimentation and new physical interpretation [2, 9]. Fortunately, the concurrent rapid development of computer technology will allow this high degree of simulation and modeling of reliability physics to be implemented on a personal workstation in a computing and network environment for integrated circuit design that has become known as technology computer-aided design (TCAD) [2,8].

3. Approach For MOSFET Reliability Physics and Modeling

Figure 2 shows a schematic for elements of an initial attempt to implement the TCAD design environment discussed in Section 2. The goal of this approach is to simulate the "hot electron problem" in small silicon MOSFETs at dimensions of 100 nm and below. Hot electron phenomena in the channel is the focal point for most device level reliability issues. Important elements of the hot electron problem include: (a) electron heating by transport through the channel; (b) highly-localized lateral and vertical electric field distributions between gate and drain regions; (c) electron injection into the SiO_2 gate insulator with a nominal injection barrier of 3.1 eV; (d) hot-carrier-induced generation and local charging of Si/SiO_2 interface states; (e) electron transport in the gate insulator; (f) degradation of the gate insulator during injection and transport; (g) modification of the electrostatic potential and channel mobility by charges in the oxide and at the interface; (h) induced changes in device characteristics, such as threshold voltage, subthreshold slope, drain current drive, transconductance, etc.; and (i) imminent expiration of device lifetime and circuit failure due to device characteristic changes. The approach shown in Fig. 2 accounts for most of these effects, except for transport in the oxide. Recent reports have shown that this effect can be considered using drift and diffusion models in TCAD approaches for large devices [7]; however, our focus here is on the role of Monte Carlo simulation in the development of accurate reliability models for very small, next generation devices.

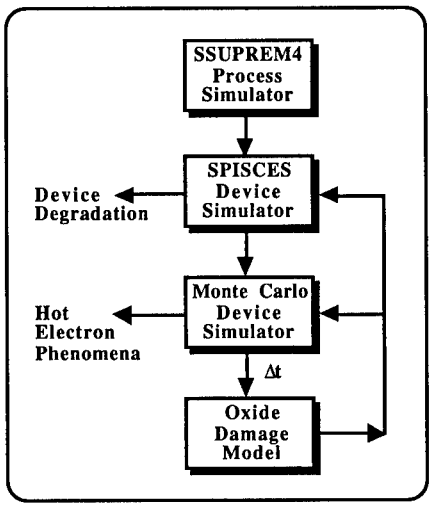


Figure 2. MOSFET reliability physics simulation and modeling environment.

As shown in Fig. 2, SSUPREM4 is used as the process simulator to design two-dimensional device structures. Data from SSUPREM4 (e.g., realistic doping profiles, process-dependent physical parameters, etc.) are provided as input for the device simulator, SPISCES, which provides standard electrical characteristics and device parameters for later determination of device degradation. These device parameters are also incorporated as initial estimates (e.g., initial electric field distributions, carrier distributions, etc.) for the Monte Carlo simulator. The Monte Carlo method is used to (particularly) model the high-energy tails in the electron energy distribution and the resultant hot electron effects [6,13].

The Monte Carlo method [6] employs a realistic silicon band structure for the two lowest conduction bands that are calculated from the pseudopotential method. A 2-D solver for Poisson's equation is coupled to the Monte Carlo method, and the solver is rapidly updated with a time step of $\Delta t = 0.1$ fsec, providing a dynamic electric field distribution for calculating long-range Coulomb interactions. The scattering routines include mechanisms for acoustic and intervalley phonons, ionized impurities, impact ionization, interface scattering, and electron-electron (short-range Coulomb) interactions. An enhanced particle statistics algorithm generates an ensemble of about 30,000 superparticles to provide necessary details for the high energy component of the electron energy distribution function. This algorithm is an essential feature to the overall success of the model.

Prediction of hot-electron-induced oxide damage is performed by adapting Yasuda's emperical model for interface state generation [6,14]. This uniform carrier injection model is modified to account for conditions of non-uniform injection near the drain that are typical for MOSFET devices. Model parameters are determined from data obtained from charge pumping experiments, and the validity of the approach was demonstrated for a 1000 nm device [6]. This approach is utilized to obtain distributions of interface states for 100 nm devices since, to our knowledge, experimental techniques are not able to resolve the spatial distribution of interface states for small MOSFETs. The resulting induced device degradation is determined by incorporating these interface states into the SPISCES simulator (Fig. 2). We assume that interface states are uniformly distributed in energy throughout the bandgap, and that they are donor-like below midgap and acceptor-like above midgap. The location of the quasi-Fermi levels at the interface determines the fraction of filled states [15]. Our approach, including Yasuda's model, has recently received some additional affirmation from simulations of pMOSFETs [7].

4. Role of Monte Carlo Simulation

A variety of device structures have been studied using the approach described in Sect. 3. These include conventional n-MOSFETs, drain- and channel-engineered n-MOSFETs, and SOI n-MOSFETs. The effects of constant field scaling and generalized scaling on hot electron phenomena have been investigated for conventional devices with channel lengths from 1000 nm down to 100 nm and applied source-to-drain voltages as low as $V_{DS} = 1.4$V. As the applied voltage was reduced below the Si/SiO_2 barrier height, no conditions were observed for which abrupt changes in slope or discontinuities occurred in simulated measures of hot carrier phenomena. This is consistent with recent experimental observations [5]. Thus, charge injection into the oxide is observed for voltages less than half the barrier height. Our simulations indicate that the lateral electric field may be increased with each scaling generation for an equivalent rate of hot electron injection. For applied voltages as low as 1.5V and channel lengths as low as 100 nm, our simulations show that lucky electron models and average electron temperature models are inaccurate for predicting the total rate and spatial distribution of hot electron effects in this regime of device scaling. Monte Carlo simulations demonstrate that, due to electron-electron interactions near the drain edge, energy distributions of hot electrons entering the drain of small devices do not cool as rapidly as the average electron temperature nor as rapidly as the electric field decays. Thus, it seems that hydrodynamic and/or energy balance models will have to be modified (if

possible) to include electron-electron interactions in order to predict hot electron behavior at small applied voltages at room temperature. Previously, electron-electron scattering was considered to be a minor effect for room temperature device operation. However, it now appears that even at room temperature, electron-electron scattering is important for modeling high-energy phenomena in devices under low-voltage operation. For example, Figure 3 shows Monte Carlo calculations of the electron energy distribution functions at the channel position of maximum electron injection into the oxide for 120 nm scaled silicon MOSFETs with three bias conditions of $V_{DS} = 2V_{GS}$ between 4.95V and 1.8V, both with (solid curves) and without (dotted curves) short range electron-electron scattering. Here, electron-electron scattering has a dominant effect on the distribution function for an applied voltage of $V_{DS} = 1.8V$.

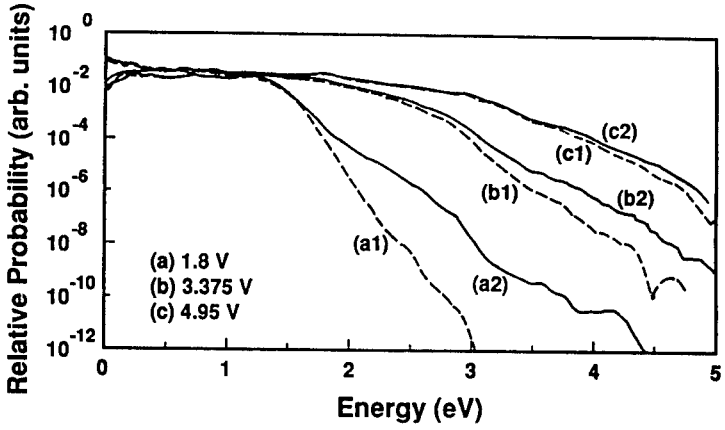


Figure 3. Energy distrib. functions for Si MOSFETs. Curves c1, c2, b1, b2 use constant field scaling; a1 and a2 use generalized scaling [6]. The applied voltages are as shown. Dashed curves are without electron-electron scattering; solid curves include electron-electron events.

These issues are further exemplified in Figure 4, which demonstrates the effects of high-energy carrier transport in a 100 nm MOSFET under low-voltage bias conditions (V_{DS}=2V_{GS}=1.5V). Fig. 4 displays the (normalized) lateral electric field, the average electron energy, and the distribution of electron injection for the channel region between source and drain. The physical channel is located from 160 nm to 260 nm, and most of the spatial variations in the quantities of interest occur in the drain region. First, it is observed that the average electron energy is spatially retarded from the electric field, with the peak in average energy occurring 5 nm "ahead" of the peak electric field. Furthermore, the peak in electron injection is also retarded another 15 nm ahead of the peak in average electron energy--a location where the average energy is well below its peak value. Hydrodynamic and/or energy balance models will not predict such retardation effects. The predicted total spatial retardation of the injection is 20 nm--a significant 20% of the 100 nm channel length. Monte Carlo scaling analysis has shown that this retardation effect is independent of channel length; thus, retardation will become an increasing fraction of the channel length as devices are scaled below 100 nm. Thus, the resultant hot electron and spatial retardation effects in electron injection will most likely become very important in future scaled MOSFETs.

Figure 5 shows electron distribution functions calculated from the Monte Carlo simulations at positions A, B, C, and D in the source-drain region and along the channel

for the device shown in Fig. 4. Point A is located in the source and point D is located well into the drain region, and both points are in regions of small variations in electric field intensity, average electron energy, and electron injection. These are "normal" distributions in the sense that they can be approximately described as drifted Maxwellian. Yet, at point B (the peak of electric field) and point C (the peak of electron injection), the electron distribution functions have extreme high energy tails extending well above the energy that can be gained from the applied bias (qV_{DS} = 1.5 eV). Note also that the ample supply of electrons in the drain region greatly reduces the average electron energy while at the same time it allows electrons to interact among themselves quickly enough to produce a high energy tail before inelastic scattering reduces the energy distribution to a Maxwellian form as electrons drift further into the drain. The effects of electron-electron interactions play a significant role in the spatial retardation shown in Fig. 4, and these retardation effects serve as further evidence of the necessity of using Monte Carlo simulation for modeling scaled MOSFET device technologies with channel lengths of 100 nm and below.

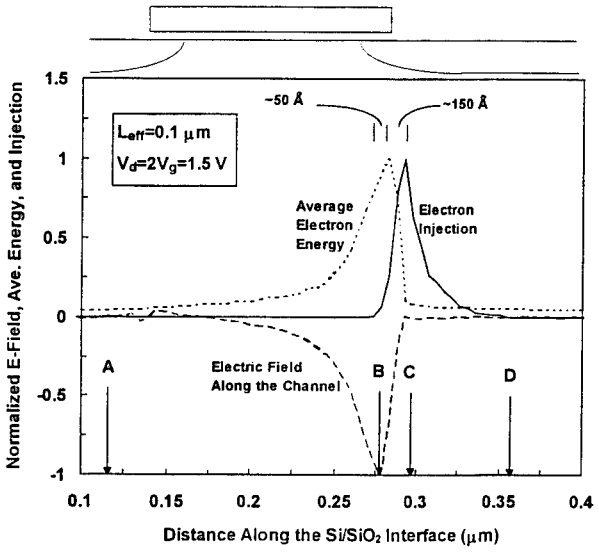


Figure 4. High energy carrier transport in a 100 nm device for low-voltage biasing conditions.

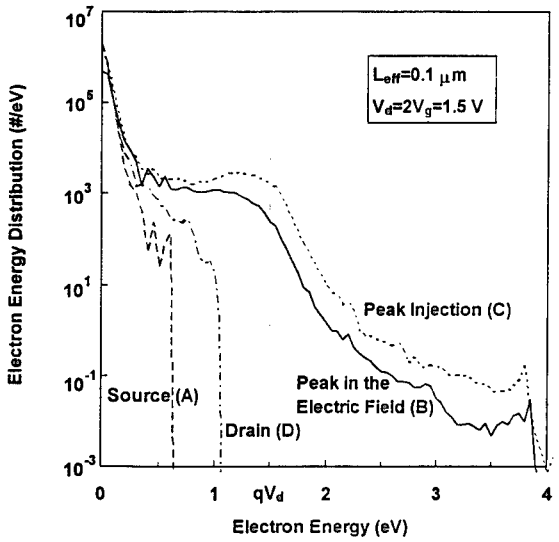


Figure 5. Electron energy distribution functions at four locations along the channel.

5. Device Reliability Predictions

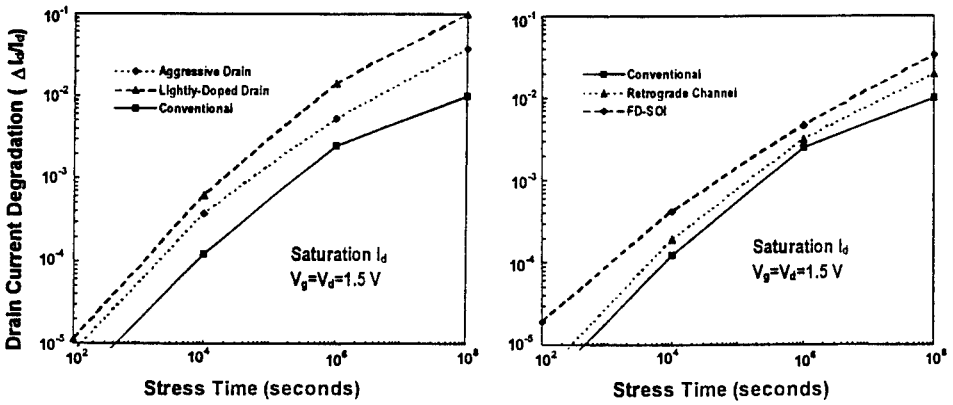
In order to evaluate the method for reliability physics and prediction described in Section 3, we have studied a selection of design concepts that are under consideration as candidates for future scaled 100 nm device structures. These devices include

a n-MOSFET using conventional processing techniques and other devices based on more novel concepts for drain- and channel-engineering [16]. Table I lists the device parameters for an array of five structures used in this comparative reliability study. The first two columns are for drain-engineered structures and the last two columns are for channel-engineered structures. The conventional structure is used as a baseline device.

Table 1. Design Parameters for 100 nm Silicon MOSFET Structures [16].

	Device Structures				
	[Drain-Engineered Devices]			[Channel-Engineered Devices]	
Device Parameter	**Aggressive Drain**	**Lightly-Doped Drain**	**Con-ventional**	**Retrograde Channel**	**Fully-Depleted (FD) SOI**
Junction Depth (nm)	35	30	50	46	30
Source/Drain Extension Doping (cm^{-3})	Large Dopant (Sb) and In Halo	7×10^{18}	degenerate	degenerate	degen-erate
Channel Doping (cm^{-3})	6.5×10^{17} (uniform)	6.5×10^{17} (uniform)	6.5×10^{17} (uniform)	Retrograde from 10^{16} to 10^{18} in 30 nm	6.5×10^{17} (uniform)
Gate Oxide Thickness (nm)	4.0	4.0	4.0	4.0	4.0
Threshold Voltage (V) at V_{DS} = 0.05V	0.41	0.38	0.40	0.41	0.41

As shown in Fig. 2, the Monte Carlo simulator is used to calculate high energy transport properties such as hot electron injection into the oxide for each n-MOSFET design for $V_{DS}=2V_{GS}=1.5V$ stress conditions. This injection data is then coupled with oxide damage models to produce distributions of interface states. These interface states are inserted into the SPISCES device simulator as interface charge determined by the electron quasi-Fermi level at the interface (see Sect. 3) to evaluate hot-electron-induced degradation of device characteristics. For example, Figures 6a and 6b illustrate simulated saturation drain current degradation ($\Delta I_D/I_D$) versus time for devices listed in Table 1. Our reliability simulation method predicts some interesting trends for these designs. First, variations in the drain design have a more pronounced effect on overall



(a) (b)

Figure 6. Simulated hot-electron-induced drain current degradation for the array of (a) drain-engineered and (b) channel-engineered device designs listed in Table I.

device reliability than variations in the channel design. Next, the lightly-doped-drain (LDD) provides little electric field reduction over the conventional design; thus, the LDD experiences comparable electron injection. On the other hand, due to the LDD's reduced drain doping, it cannot efficiently screen interface charge, and thus suffers considerably greater degradation than the conventional design. Last, among the channel-engineered designs, the fully-depleted (FD) SOI design suffers the greatest drain current degradation. This result can be mainly attributed to interface coupling effects due to substantial electron injection and subsequent oxide damage in both the front and back oxides. It is also worthy to note for the array of designs that the hot electron injection distributions have relative magnitudes that correlate directly to the relative magnitudes in peak parallel electric field, and the injection distributions are displaced about 20 nm beyond their respective highly-localized electric field distributions, as also demonstrated for the 100 nm device in Fig. 4. The results of Fig. 6 predict that the conventional design is the most reliable device in the array of designs considered. This outcome indicates that although there is a fundamental change in the underlying mechanisms that produce high-energy carriers as device size and applied voltage are scaled down (i.e., electron-electron interactions become more significant), the trade-off between device performance and reliability is still quite significant.

Another point from Fig. 6 is that the drain current does not conform to the model that predicts a square root dependence of degradation on the stress time [see discussion in Section 3 of Ref. 6]. The slope of the degradation curves is particularly dependent on the interface state density and the resultant carrier recombination and trapping effects that can occur. In particular, in Fig. 6, the FD-SOI device appears to have a smaller curvature, which could be due to the fact that both oxide interfaces participate in the degradation process. In an attempt to further evaluate this degradation signature, we studied two additional 100 nm FD-SOI designs. These include a heavily-doped channel (SOI-1) and a lightly-doped channel (SOI-2) device design, with a uniform channel doping of 1 x 10^{18} cm^{-3} and 1 x 10^{16} cm^{-3}, respectively. The silicon layer thickness, front oxide thickness, and back oxide thickness are 30 nm, 4 nm, and 80 nm,

respectively, for both devices. The gate workfunction is adjusted to produce equal threshold voltages of 0.4V for both devices. Figure 7 shows the simulated drain current degradation in both the linear and saturation regions of the current-voltage characteristics of SOI-1 and SOI-2. The SOI-2 design experiences greater degradation mainly due to increased coupling of interface states through the lightly-doped channel. It is also observed that these two devices experience greater degradation rates than the 100 nm designs listed in Table 1. This is consistent with the interpretation that the degradation of drain current for SOI devices is proportional to hot electron injection and subsequent oxide damage in both the front and back oxides.

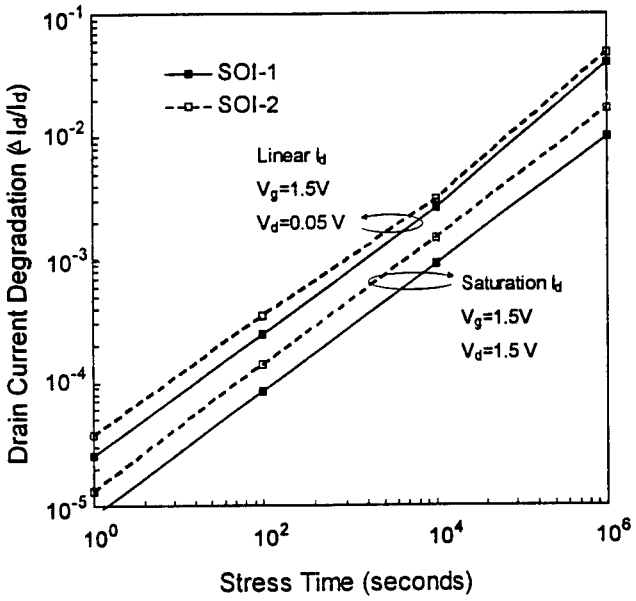


Figure 7. Drain current degradation characteristics for SOI-1 and SOI-2.

6. Conclusion

A Monte Carlo simulator has been developed and combined with a process simulator, a device simulator, and oxide damage models to investigate hot electron phenomena and device reliability for deep submicron (100 nm) n-MOSFETs and SOI devices designed for low-power applications. Monte Carlo simulations indicate that non-local transport and two-dimensional effects in the drain current and electric field distributions influence hot electron injection into the oxide(s) for both bulk MOSFET and SOI devices. Also, electron-electron interactions play an important role in the creation of the high energy tail of the electron energy distribution function for devices under low-voltage bias conditions ($qV_{DS} < \varnothing_B$). The Monte Carlo results are linked to the reliability simulation method to investigate hot-electron-induced device degradation. This method predicts that as devices are scaled down to the 100 nm regime, aggressive drain designs can provide increased performance while reducing the trade-off of reduced reliability. The LDD design concept, while improving short channel effects, no longer reduces device degradation when compared to the conventional design. Finally, ultrathin-film, fully-depleted SOI designs greatly improve short channel effects; however, they experience considerably greater device degradation than their bulk MOSFET counterparts due to hot-carrier-induced damage to both the front and back oxides.

7. List of References

1. Hiroshi Iwai, Hisayo Sasaki Momose, Masanobu Saito, Mizuki Ono, and Yasuhiro Katsumato (1995) The Future of ultra-small geometry MOSFETs beyond 0.1 micron, *Microelectronics Engineering* **28**, 147-154.
2. Armin W. Weider (1995) Si-Microelectronics: Technology Perspectives-Risks, Opportunities, and Challenges (This Volume).
3. G. Baccarani, M. R. Wordeman, and R. H. Dennard (1984) Generalized Scaling Theory and Its Application to a 1/4 Micrometer MOSFET Design, *IEEE Trans. Electron Devices* **ED-31**, 452-459.
4. J. E. Chung, M.-C. Jeng, J. E. Moon, P.-K. Ko, and C. Hu (1990) Low-Voltage Hot-Electron Currents and Degradation in Deep-Submicrometer MOSFET's, *IEEE Trans. Electron Devices* **ED-37**, 1651-1658.
5. David Esseni, Luca Selmi, Roberto Bez, Enrico Sangiorgi, and Bruno Ricco (1994) Bias and Temperature Dependence of Gate and Substrate Currents in n-MOSFETs at Low Drain Voltage, *Proceedings of the International Electron Devices Meeting (EDM,),* 307-310.
6. John J. Ellis-Monaghan, K. W. Kim, and Michael A. Littlejohn (1994) A Monte Carlo study of hot electron injection and interface state generation model for silicon metal-oxide-semiconductor field-effect transistors, *J. Appl. Phys.* **75**, 5087-5094.
7. A. v. Schwerin and W. Weber (1995) 2-D Simulation of pMOSFET hot-carrier degradation, *Microelectronic Engineering* **28**, 277-284.
8. M. Rodder, A. Amerasekera, S. Aur, and I. C. Chen (1994) A Study of Design/ Process Depaendence of 0.25 μm Gate Length CMOS for Improved Performance and Reliability, *Proceedings of the IEDM,* 71-74.
9. Andrea Ghetti, Luca Selmi, Enrico Sangiorgi, Antonio Abramo, and Franco Venturi (1994) A Combined Transport-Injection Model for Hot-Electron and Hot-Hole Injection in the Gate Oxide of MOS Structures, *Proceedings of the IEDM,* 363-366.
10. S. Jallepalli, C.-F. Yeap, S. Krishnamurthy, X. L. Wang, C. M. Maziar, and A. F. Tasch Jr. (1994) Application of Hierarchical Transport Models for the Study of Deep Submicron Silicon MOSFETs, *VLSI Symposium Proceedings*, 91-94.
11. Marco Mastrapasqua and Jeff D. Bude (1995) Electron and Hole Impact Ionization in Deep Sub-micron MOSFETs, *Microelectronic Engineering* **28**, 293-300.
12. K. Taniguchi, M. Yamaju, K. Sonoda, T. Kunikiyo and C. Hamaguchi (1994) Monte Carlo Study of Impact Ionization Phenomena in Small Geometry MOSFETs, *Proceedings of the IEDM,* 355-358.
13. R. B. Hulfachor, K. W. Kim, M. A. Littlejohn, and C. M. Osburn (1995) Non-Local Transport and 2-D Effects on Hot Electron Injection in Fully-Depleted 0.1 μm SOI n-MOSFETs Using Monte Carlo Simulation, *Microelectronic Engineering* **28**, 175-182.
14. N. Yasuda, H. Nakamura, K. Taniguchi, and C. Hamaguchi (1989) Interface State Generation Mechanism in N-MOSFETs, *Solid State Electronics* **32** , 1579-1586.
15. M. V. Fischetti and S. E. Laux (1988) Monte Carlo Analysis of Electron Transport in Small Semiconductor Devices Including Band-Structure and Space-Charge Effects, *Phys. Rev. B* **38**, 9721-9730.
16. R. B. Hulfachor, K. W. Kim, M. A. Littlejohn, and C. M. Osburn (1995) A Monte Carlo Study of Drain and Channel Engineering Effects on Hot Electron Injection and Induced Device Degradation in 0.1 mm n-MOSFET's, *Fifty-Third Annual Device Research Conference (DRC) Digest*, 14-15.

SUPERCONDUCTOR-SEMICONDUCTOR DEVICES

HERBERT KROEMER
ECE Department, University of California
Santa Barbara, CA 93106, USA

1. Introduction

1.1 THE PREMISE

It has long been recognized that electronic devices operating at reduced temperatures—including both semiconductor and superconductor devices—can often offer much higher performance (by several criteria) than room-temperature devices. But the need for cooling has greatly retarded their use, and there exists an almost-universal persistent belief that low-temperature devices just don't have a chance to find significant practical applications.

My presentation is based on the premise that this belief is a myth, and that the future of electronics is likely to draw increasingly, within the next decade or two, on low-temperature devices, at least in applications such as high-performance workstations and scientific and medical instrumentation, where increasing performance requirements can justify the additional cost of the cryogenics, which is itself decreasing

However, the performance-to-cost relation is by no means the only issue: No matter how favorable that relation is, no system engineer is going to fool around in a "real" commercial system with cryogenics under conditions that resemble those of a research laboratory. What is absolutely essential is "user-friendly" cryogenics! The enabling technology for the widespread actual use of cryogenic electronics is likely to be the increasing availability of small self-contained closed-cycle refrigerators. The development of the latter (mainly Stirling-cycle machines), originally driven by IR detector technology, has more recently found increasing use in high-T_c superconductor applications. It is rapidly approaching the point that we may begin to view such a refrigerator as just another module inside a piece of electronic equipment, somewhat analogous to, say, a fancy high-voltage power supply.

S. Luryi et al. (eds.), Future Trends in Microelectronics, 237–250.

Suppose I offered you a self-contained box, about 2-3 liters in volume, drawing less than 100 Watts, and I would provide inside this box a volume of about $100cm^3$ inside which I guarantee a temperature T, of say, 77K, with a cooling capacity of, say 3-4 Watts. Given a reasonable cost, such a box would evidently meet our demand for user-friendly cryogenics. The above specifications are not fictitious, they are those of actual hardware about to go into production, interestingly by a company whose business is in the field of high-T_c superconductors, and which has found it necessary to provide integrated system solution to its customers, solutions that include a user-transparent cryogenics package (Superconductor Technologies, Santa Barbara, CA).

The principal bottleneck to their more widespread use is their cost, but this is likely to follow the classical pattern of dramatic cost reduction in the wake of building up mass production. Furthermore, the specifications are likely to improve with time, including rapid progress to lower temperatures with time, at least to about 20K, the practical limit of the Stirling cycle, with slower progress below that.

1.2. SUPERCONDUCTOR-SEMICONDUCTOR DEVICES

1.2.1. *Hybrids With Buffer Layers*

The devices that very likely will emerge in the wake of this development will not only be superconducting devices using high-T_c superconductors, and conventional devices such as FET's explicitly designed to operate at low temperatures, but also integrated super-semi hybrids. The first class likely to emerge are high-T_c superconductors integrated on-chip with semiconductor devices, like a superconducting SQUID integrated with GaAs or InAs electronics. Because of processing compatibility limitations, such devices require a buffer layer between the two kinds of materials, a technology in which much progress has been made recently [1, 2]. But, being devices operating at temperatures within easy range of the Stirling cycle, such devices should emerge relatively soon.

1.2.2. *Monolithic Integration without Interface Barrier*

As "practical" temperatures get pushed lower, we will also see devices in which a low-T_c superconductor, such as Nb, has been integrated with a semiconductor, such as InAs, without an intervening layer, in such a way that the electrons can cross the interface while retaining the phase information that is the essence of superconductivity, thereby inducing superconductivity in the semiconductor. New kinds of Josephson devices based on this principle are rapidly emerging, offering advantages over more conventional Josephson devices. In fact, much of my presentation—all of Section 2—will deal with this particular combination, as a look far ahead at a branch of low-temperature transport physics that is likely to become important over the long term.

For the near-term future (< 10 years), the need for operating temperatures below the Stirling-cooler range (< 20K) implies more elaborate cryogenic techniques, and these devices may, for some time, remain restricted to two kinds of applications environments: (a) Environment where cryogenic temperatures are available in any event, and where cryogenic electronics can be piggy-backed on the existing cryogenics with minimal additional cost. (b) Large-scale "ultimate-performance" computer mainframes the cost of a helium liquefier would represent only a small fraction of the cost of the overall machine.

1.3. ON NOT REPEATING THE PAST

Anybody invoking this last scenario as a realistic one for the future must address him- or herself to the fact that a huge effort of precisely this kind was undertaken by IBM during the 70-s, only to be abandoned in 1983. The failure of this project had a terribly discouraging effect on the whole field of low-temperature electronics, and anybody re-considering this approach is in danger of running afoul of Santayana's famous dictum that "those who do not remember the past are condemned to repeat it."

It has been argued persuasively by Likharev [3] that this failure was due, not to the need for liquid-helium temperatures, but to two quite unrelated reasons: (a) The use of a unsuitable non-refractory metallurgy based on lead as a superconductor, which was not sufficiently stable under thermal cycling. The resulting reliability problems would have been avoided by using niobium as a superconductor. (b) The use of a logic principle, employing voltage-state logic, that was basically too imitative of semiconductor logic, and which had inherent power dissipation limits that negated much of the speed advantage of Josephson junctions. As Likharev points out, a much more suitable form of superconducting logic would be one that is based on the unique property of superconductors that magnetic flux in superconducting loops is quantized, and which shuffles single flux quanta rather than shuffling voltage states. Likharev's own presentation at this workshop reviews the present state

It would constitute a major breakthrough for superconductor-semiconductor devices if a high-temperature superconductor could be found that is technologically compatible with existing semiconductors, especially III-V semiconductors. As it stands now, all the high-T_c superconductors are oxides that must either be deposited, or require a post-deposit anneal, in a high-temperature oxidizing atmosphere that will simply destroy any of the semiconductors it is in atomic contact with, thereby eliminating barrier-free structures. Current research on high-T_c superconductors stresses the achievement of higher critical temperature, rather than elimination of the need for a high-temperature oxidizing environment. From the point of view of super-semi devices, the achievement of semiconductor-compatible materials would be a far more valuable goal, even if it meant a drastic reduction in critical temperature, say, to 40K.

2. Semiconductor-Coupled Superconducting Weak Links

2.1. INTRODUCTION

As the title of my presentation indicates, its objective is restricted to low-temperature devices in which superconductors and semiconductors are monolithically integrated into a common device, ignoring both "pure" superconductor devices—such as Josephson tunnel junctions—that do not involve a semiconductor, and pure semiconductor devices that just happen to be specifically designed for low-temperature use. In fact, my presentation concentrates on what I consider the potentially most interesting form of monolithic superconductor-semiconductor integration, namely, *semiconductor-coupled superconducting weak links*. Much of contents of this section is based on a recent longer introductory review of this topic by Professor Hu and myself [4], where the interested reader may find additional details and additional references. An earlier elementary introductions is found in [5].

The term *weak links* refers to superconducting devices in which two superconducting "banks" are coupled through another *conducting* medium, as opposed to Josephson *tunnel* junctions, in which the current flow is by Cooper pair tunneling through an *insulating* barrier. In the case of interest here, the conducting medium is a semiconductor rather than a metal. More specifically, it is a narrow (~15nm) InAs quantum well with AlSb barriers, forming a short (<1μm) conducting link between two Nb superconducting banks, schematically shown in Figure 1. For reasons I will discuss below, this combination has emerged as a particularly promising one.

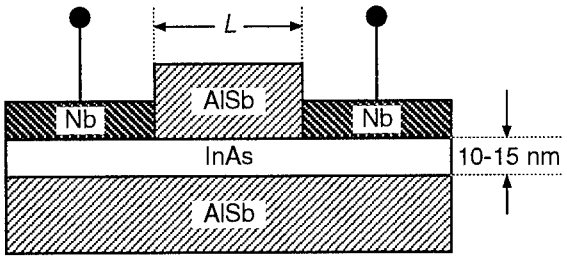


Figure 1. Semiconductor-coupled superconducting weak link based on an InAs-AlSb quantum well forming a conducting link between two superconducting Nb electrodes.

Like Josephson tunnel junctions, weak links exhibit a pronounced *Josephson effect*, manifested by a current-voltage characteristic as in Figure 2, which shows data from a semiconductor-coupled weak link of the kind shown in Figure 1. The characteristic feature of the Josephson effect is the existence of a current range inside which a resistance-less *supercurrent* can flow between the two superconducting banks, up to a

certain critical current I_c. Only when this current is exceeded does a voltage appear between the superconducting terminals.

Compared to tunnel junctions, weak links have a much larger inter-electrode separation between the two superconducting banks, which leads two potential major advantages: (a) much lower capacitances, an important consideration for the use of these devices as high-speed devices (b) a much smaller sensitivity of the characteristics to variations in the electrode separation.

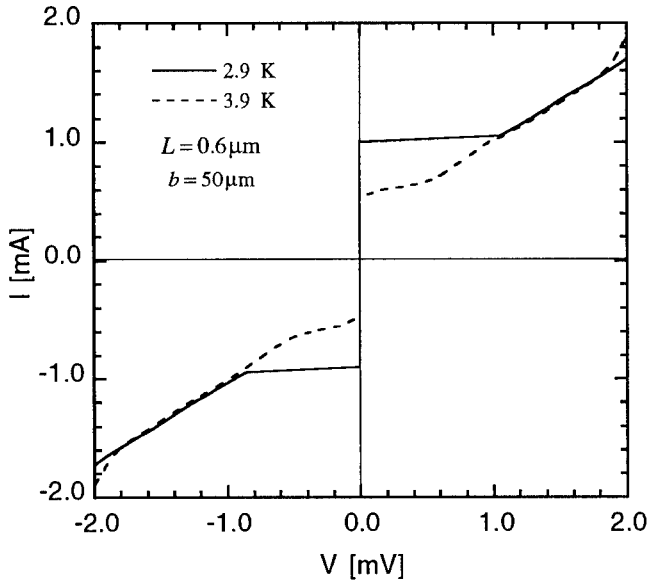


Figure 2. Josephson-type *I-V* characteristics of a device as shown in Figure 1, with 0.6μm electrode separation, at two temperatures [6].

I will not address myself here to the actual *applications* of semiconductor-coupled weak links. In principle, weak links are candidates for all applications for which Josephson tunnel junctions are candidates, with the advantage of higher potential speed, and a technology that lends itself naturally to integration with semiconductor circuitry, including monolithic integration in which the latter operates at the temperature of the weak link itself. One specific application is of course in computers based on the Josephson effect, a topic where I gladly defer to Likharev's presentation at this workshop. However, as I have stated in my earlier presentation at this workshop, the principal applications of any sufficiently new technology tend to be applications *created* by the new technology—which at this time must be left open to speculation.

2.2. BASIC WEAK LINK PHYSICS: A TUTORIAL

2.2.1. *Current-Phase and Phase-Voltage Relations*

An understanding of weak links requires at least a rudimentary understanding of the basic physics underlying Josephson junctions in general, and weak links in particular. I summarize her the basic facts, without justifications or derivations, for which I must refer to relevant texts (see, for example, refs. [7-10]).

The Pair Wave Function and its Phase. The essence of superconductivity is the existence of a common pair wave function for the Cooper pairs in the superconductor, which may be written

$$\psi(\mathbf{r}) = |\psi(\mathbf{r})| \cdot \exp[i\theta(\mathbf{r})]. \tag{1}$$

Here, the magnitude $|\psi(\mathbf{r})|$of the pair wave function is related to the local Cooper pair density $n(\mathbf{r})$ via

$$|\psi(\mathbf{r})|^2 = n(\mathbf{r}), \tag{2}$$

and $\theta(\mathbf{r})$ is a phase. The key point is that this phase is coherent over macroscopic distances, and, in the absence of a current, it is the same throughout the entire superconductor.

Supercurrent as a Function of the Phase Difference. In a weak link, two superconductors are coupled through another conducting medium, through which electrons can pass in such a way that the phase of the electrons is preserved in the process. If the phases of the two superconductors are the same, there will be zero net current, but if there is a phase *difference* between the two superconductors, a resistanceless Josephson supercurrent can flow from one superconductor to the other, the magnitude of which is a function of the phase difference $\theta_2 - \theta_1$. For Josephson *tunnel* junctions the functional relationship is simply sinusoidal,

$$I = I_c \cdot \sin(\theta_2 - \theta_1). \tag{3}$$

Here, with the ordering of the two phases as given, a positive current designates a flow of Cooper pairs from bank #2 to bank #1. Because the pairs carry a negative charge $-2e$, the *electrical* current is in the opposite direction, from bank #1 to bank #2. In weak links, more complicated relations may occur, but $I(\theta_2 - \theta_1)$ is always an odd function

of the phase difference, and inasmuch as phase differences have a physical meaning only modulo 2π, the $I(\theta_2 - \theta_1)$ relation is necessarily a periodic one, with a period 2π.

A.C. Josephson Effect. In the absence of a bias voltage between the two superconducting banks, whatever phase difference $\theta_2 - \theta_1$ may be present, will not change with time, hence the current will continue to flow—which is why it is called a supercurrent. If an external bias voltage is present, the difference becomes time-dependent according to the simple law

$$\frac{d}{dt}(\theta_2 - \theta_1) = \frac{2e}{\hbar} \cdot (V_2 - V_1). \tag{4}$$

The supercurrent-vs.-phase relation $I(\theta_2 - \theta_1)$ remains valid in the presence of such a voltage, but the supercurrent now oscillates about zero, with the Josephson frequency

$$\nu_J = \frac{2e}{h} \cdot (V_2 - V_1), \tag{5}$$

where $h = 2\pi\hbar$ is Planck's constant.

2.3. ANDREEV REFLECTIONS

2.3.1. *Semiconductor-Coupled Weak Links as "Clean" Weak Links*

The weak-link physics of Sec. 2.2 holds independently of the nature of the mechanism that preserves the phase of the electrons. In the semiconductor-coupled weak links discussed here, the mean free path of the electrons tends to be larger than the inter-electrode separation, in which case the mechanism for the phase transfer tends to be dominated by the phase-coherent flow of *ballistic* electrons between the banks. In the jargon of superconductivity, such weak links are called "clean" weak links, in contrast to the more common "dirty" weak links extensively studied in the past, in which the electron transport is diffusive. Unfortunately, much of the literature on weak links, including Likharev's classical review of weak links [11], is still dominated by considerations of dirty weak links.

The mean free path that matters for the phase transfer is not the elastic mean free path that determines the low-field mobility, but the *inelastic* mean free path that is responsible for any de-phasing of the electron waves, and which is typically much longer than the elastic mean free path. For example, in impurity scattering the phase of the scattered wave is coherent with the phase of the incident wave, and while such scattering may create a chaotic wave front, this does not constitute phase-incoherence

in the sense of weak link theory: there is still a fixed phase relation between any two points in the wave field.

Given an inelastic mean free path much longer that the inter-electrode spacing, the dominant phase-altering process for the electrons becomes the scattering, not inside the semiconductor, but at the semiconductor-superconductor interface, between the electrons in the semiconductor and those in the superconductor. Now, electron-electron scattering is normally a phase-destroying process. However, at a super-semi interface, at sufficiently low temperatures, the only electrons available for participation in scattering on the superconductor side are the Cooper pairs. But, as we saw earlier, the Cooper pairs all have the same well-defined phase. As a result, the scattering interaction of electrons in the semiconductor with electrons in the superconductor becomes itself a phase-coherent process. It is universally referred to as Andreev scattering or, more commonly, as *Andreev reflections* (AR's), in honor of the man who discovered the possibility of such a process in 1964 [12].

Although postulated over thirty years ago, Andreev reflections have received major attention only during the last few years, when it became clear that their understanding is central to the understanding of clean-limit weak links. As a result of this belated recognition, they have not yet found their way into current textbooks on superconductivity. Even the 1979 weak-ink review by Likharev, written just before clean weak links became technologically realizable, mentions Andreev reflections only in passing. In fact, on page 132 of his paper [11], Likharev explicitly lists a number of experimental observations that are not consistent with the then-existing theoretical understanding, all of which find their explanation via Andreev scattering. I therefore provide here the necessary background on this topic.

2.3.2. *Andreev Reflections: Basic Concept*

The basic idea behind Andreev reflections is simple. Consider an interface between a degenerately doped semiconductor and a superconductor. As shown in Figure 3a, a superconducting energy gap has opened up on the superconductor side. If now an electron with an energy $\mathcal{E}$ *above* the Fermi level (but still inside this gap) is incident on the interface from the semiconductor side, the absence of single-particle states within the gap prevents that electron from entering the superconductor as a *single* electron, and one might expect this electron to be reflected, and the electrical resistance to current flow across the interface actually to increase at the onset of superconductivity in the metal.

However, the electron may pair up with a second electron at the same energy $\mathcal{E}$ *below* the Fermi level, forming a Cooper pair, which *can* enter the superconductor, causing a doubling of the current compared to that in the absence of superconductivity, rather than a suppression. The electron removed from the semiconductor below the Fermi level leaves behind a hole in the Fermi sea. The generally accepted jargon

associated with this phenomenon is to say that the incident electron is *reflected as a hole.*

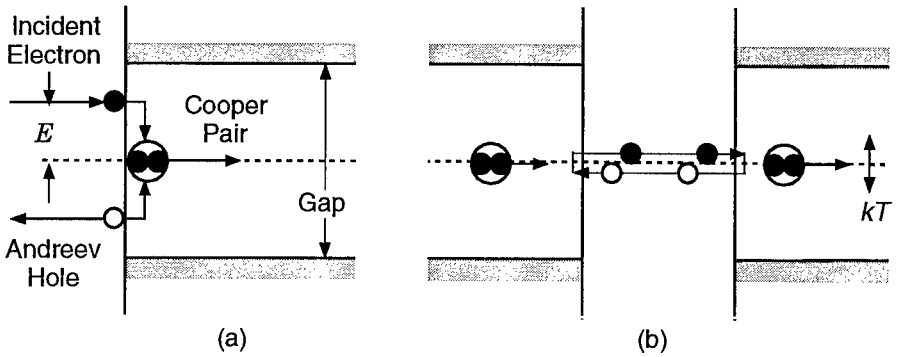


Figure 3. Andreev reflections. (a): Basic concept. (b): Persistent current flow by multiple Andreev reflections.

In a semiconductor with a large mean free path for the electrons, the Andreev hole left behind has a large mean free path itself, roughly equal to that of the original electron, and theory shows that the hole travels back into the semiconductor along a trajectory that essentially re-traces the trajectory of the original incident electron. If its mean free path is sufficiently large, the hole will eventually reach the opposite superconducting electrode. In the absence of any bias across the structure, the energy of the hole is still within the superconducting gap on that side. Such a hole cannot enter the superconductor, but it can be annihilated by breaking up a Cooper pair inside the other superconductor: One of the electrons of the pair annihilates the hole, the other electron takes up the annihilation energy, and is injected into the semiconductor as a ballistic electron *above* the Fermi level, at an energy exactly equal to that of the initial electron. This process, illustrated in Figure 3b, can evidently be repeated: The Andreev reflections act as what I like to call a "Cooper pair pump," annihilating Cooper pairs on one side, and re-creating them on the other. If *all* reflections of electrons and holes were Andreev reflections rather than "ordinary" reflections, and if there were no other kinds of scattering processes, the result would be a persistent current.

However, perturbations are always present, and we would expect simply an enhancement of the conductivity by a factor equal to the number of ballistic traverses before an unfavorable reflection or collision event randomize either the electron or the hole flow in this chain reaction. Also, even if no unfavorable reflection and collision events took place, the diagrams in Figure 3b show only half the story: For each state with a given direction of arrows there exists another state with all current flows reversed. If both states of such a pair are occupied, their currents will cancel. To understand how a supercurrent can arise, we must go beyond the pure ballistic particle

picture of Figure 3b, and must take into account the wave properties of the unpaired electrons and holes, and of the Cooper pairs [13].

2.3.3. *Andreev Supercurrents*

Waves have phase, and even in the absence of any scattering events, the simple current-carrying state illustrated schematically by Figure 3b is a quantum-mechanically allowed stationary state only if the round-trip phase shift along the electron-hole loop is an integer multiple of 2π,

$$\Delta\phi_{RT} = 2\pi n. \tag{6}$$

For every value of n, there will actually be two states, corresponding to opposite directions of the flow arrows in Figure 3b.

Up to a point, the above is exactly the same condition as for the bound states in an "ordinary" one-dimensional semiconductor quantum well. These, too, are states for which the round-trip phase changes are the different multiples of 2π. In fact, with regard to the spatial confinement of the *unpaired* electrons and holes inside the semiconductor portion of the structure, the stationary states may indeed be viewed as a new kind of bound states [13], the difference being that the "heterojunction" barriers are now formed, not by the conventional energy gap of another semiconductor, but by the superconducting energy gap of the two superconducting electrodes.

However, there are two decisive differences. The first is that for an AR state confined by superconducting energy gap barriers, the phase on one of the two traverses is carried by an electron, on the other traverse by a hole. This means that these kinds of bound states actually carry a current across the semiconductor, in contrast to the current-less conventional bound states in a conventional quantum well. The two states belonging to a given n belong to opposite directions of that current flow.

A second difference is the following. As in a conventional quantum well with barriers of finite height, the round-trip phase shift contains a contribution from the reflections at the two superconductor barriers. In a semiconductor quantum well, these contributions simply represent the finite penetration of the wave function into the barrier, and they are responsible for lowering the bound state energies with decreasing barrier height.

But in the case of Andreev reflections there is an *additional* phase shift at each bank, equal in magnitude to the phase of Cooper pair wave function in that bank, but with a sign depending on whether an electron or a hole is reflected: When an electron is reflected at a superconductor with phase θ, the wave function of the hole resulting from the reflection acquires an additional phase shift by $-\theta$. This can be readily understood by realizing that the Andreev reflection of an incident electron creates an additional

Cooper pair with phase θ. The phase shift $-\theta$ of the reflected hole simply compensates for the phase of the new Cooper pair.

Conversely, if a hole is reflected, the wave function of the resulting electron acquires the phase $+\theta$, with a similar interpretation. What matters for the Andreev bound states is of course the *net* round-trip phase shift. If the two superconducting banks have the same phase, the phase shifts by $\pm\theta$ at the two banks cancel, but if there is a phase difference between the two banks, it will make a contribution

$$\Delta\phi = \pm(\theta_2 - \theta_1) \tag{7}$$

to the round-trip phase shift, with the following sign rule: If, in Figure 3b, the left-hand bank is bank #1, the minus-sign applies, otherwise the plus sign.

In order to retain the round-trip condition (6) in the presence of the phase shift contribution $\Delta\phi$, the latter must be compensated for by an opposite change in the phase shift contribution associated with the ballistic flight through the semiconductor itself. But this leads to a change of the energy of the Andreev bound states: A positive contribution to the round-trip phase shift requires a lowering of the ballistic phase contribution, and hence a lowering of the bound-state energy, while a negative contribution raises the latter. Because of the sign difference in (7), in the presence of a nonzero phase difference $\theta_2 - \theta_1$, the energies of the bound states will depend on the direction of current flow in each state, in such a way that the states with a current flow in the direction proper for a Josephson supercurrent will have a lower energy and hence a higher thermal occupation probability, than those with a current flow in the opposite direction. Hence, in this case there will be a thermodynamically stable net current flow, even in the presence of scattering events.

Recall finally that a time-independent phase difference corresponds to zero bias voltage. Hence the stable current is a true zero-resistance supercurrent, with a certain maximum value, the critical current, for some particular value of the phase difference $\theta_2 - \theta_1$.

When the current through the device exceeds the critical current, a bias voltage develops across the semiconductor, leading to the bending-over of the *I-V* characteristic seen in Figure 2. This dissipative regime contains itself a rich variety of physical phenomena, the discussion of which would again go beyond the scope of this paper; the interested reader is referred to the literature, probably starting with a few existing elementary review papers [4-6], which contain extensive references to key original papers, including specifically to papers on the detailed theory for the various phenomena.

2.4. SELECTED RECENT RESULTS ON INAS/ALSB QUANTUM WELLS-COUPLED WEAK LINKS

2.4.1 *InAs, and the JOFET Concept*

I return now to the specific case of weak links of the kind shown in Figure 1 based specifically on InAs-AlSb quantum wells with Nb electrodes. A full discussion of the reasons for this preference would go far beyond the scope of this paper; the interested reader is referred to recent review papers on these structures [4-6], and I can give here only a very condensed summary. Of the various semiconductors, InAs is preferred principally because of its well-known property that the Fermi level pinning at metal-to-InAs interfaces takes place inside the InAs conduction band [14]; hence such interfaces do not form Schottky barriers impeding the flow of electrons between the superconductor. The choice of AlSb as the barrier material then follows naturally, because of its low lattice mismatch to InAs (1.3%), and a favorable band lineup, with a large conduction band offset of 1.35eV [15]. The latter permits the use of modulation doping to achieve very high electron sheet concentrations (approaching $10^{13}/cm^2$), while maintaining high mobilities (approaching $10^5 cm^2/V{\cdot}s$). Finally, Nb is the natural candidate as the superconductor, because of its high critical temperature. Obviously, the use of a high-T_c superconductor would be more desirable, and remains a goal for the future, but none of the currently known high-T_c superconductors is technologically compatible with the known semiconductors.

Historically, the idea of to combine InAs with Nb is not new, but was proposed already in the two 1978 and 1980 papers by Silver et al. [16] and by Clark et al. [17] that stimulated much of the current interest in semiconductor-coupled weak links. These two papers actually went beyond two-terminal weak links; they proposed the possibility of *three*-terminal FET-like devices, called *Josephson Field Effect Transistors*, now commonly referred to as a JOFET's. The central idea was simple: One of the most characteristic features of semiconductors is that their electron concentration is not a fixed quantity, but can be modulated by a gate electrode. Hence, in a semiconductor-coupled weak link, it should be possible to modulate the current-voltage characteristics by such a gate electrode, leading to an FET-like current-voltage characteristic. These authors, especially Silver et al., already recognized the importance of the contact barrier problem, and hence proposed the use of InAs as the preferred semiconductor.

It would go beyond the scope of the present article to give a review of these earlier development. The interested reader is referred to two recent papers by the Japanese NTT group [18, 19], which report the best true three-terminal JOFET's to-date, employing InAs quantum wells, with confinement barriers made, not from AlSb, but from (Ga,In)As/(Al,In)As. These papers contain extensive references to related work can be found.

However, work on true (three-terminal) JOFET's forms only a small fraction of the overall recent work on semiconductor-coupled weak links that was stimulated by the original JOFET proposal. None of the JOFETs actually reported to-date have shown the kind of performance that offers promise for practical applications. One of their most severe problems is that the obtainable drain-to-source voltage swings are typically much less (<< 1mV) than the gate voltage swings (>> 1mV) required to achieve significant drain current changes. These devices therefore have painfully low voltage gains, which appear to be inherent in their physics. Even the recent NTT devices just barely achieve a voltage gain of unity under optimal loading conditions. It remains to be seen whether or not future developments will overcome this problem. As it stands now, a more likely application of JOFETs is as current-routing switches in superconducting networks, drawing on the fact that a JOFET is an FET with a true zero-resistance on-state, something no pure semiconductor device can offer.

2.4.2 *Multi-Gap Grating Structures*

We ourselves have found it useful to go beyond a single-gap device geometry of Figure 1, and to study series-connected periodic arrays, prepared by laser holography, involving a large number (≥ 300) of gaps, shown schematically in Figure 4.

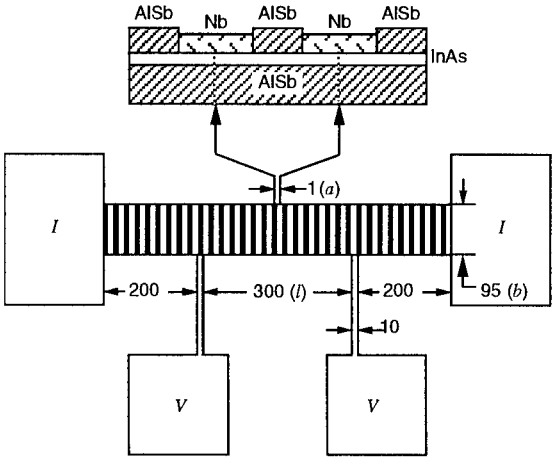


Figure 4. Overall layout (bottom) of Nb grating structure, along with (top) a schematic cross-section through a pair of Nb lines separated by a narrow stripe of InAs-AlSb quantum well. All dimensions are in μm.

We consider such grating structures particularly promising for future applications, and extensive studies of such structures are currently underway, to be reported in due course. Initial results are found in [6] and [4].

References

1. Chang, L. D., Tseng, M. Z., Fork, D. K., Young, K. H., and Hu, E. L. (1992) Epitaxial MgO buffer layer for $YBa_2Cu_3O_{7-x}$ thin films on GaAs, *Appl. Phys. Lett.* **60**, 1753-1755.
2. Tseng, M. Z., Jiang, W. N., and Hu, E. L. (1994) Measurements and analysis of Hall effect of a two dimensional electron gas in the close proximity of a superconducting $YBa_2Cu_3O_{7-x}$ film, *J. Appl. Phys.* **76**, 3562-3565.
3. Likharev, K. K. and Semenov, V. K. (1991) RSFQ Logic/Memory Family: A new Josephson-Junction Technology for Sub-Terahertz-Clock-Frequency Digital System, *IEEE Trans. Appl. Supercond.* **1**, 3-28.
4. Kroemer, H. and Hu, E. (1996) "Semiconducting and Superconducting Physics and Devices in the InAs/AlSb Materials System," in *Nanotechnology*, G. Timp, Ed., New York, AIP Press. In the press.
5. Kroemer, H., Nguyen, C., and Hu, E. L. (1994) Electronic Interactions at Superconductor-Semiconductor Interfaces, *Solid-State Electron.* **37**, 1021-1025. (Proc. MSS-6, Garmisch-Partenkirchen, Germany, Aug. 1993).
6. Kroemer, H., Nguyen, C., Hu, E. L., Yuh, E. L., Thomas, M., and Wong, K. C. (1994) Quasiparticle transport and induced superconductivity in InAs-AlSb quantum wells with Nb electrodes, *Physica B* **203**, 298-306. (Proc. NATO Advanced Research Workshop on Mesoscopic Superconductivity, Karlsruhe, 1994).
7. Feynman, R. P., Leighton, R. B., and Sands, M. (1965) *The Feynman Lectures on Physics; Vol. 3: Quantum Mechanics*, Addison-Wesley, Reading. See Sec. 21-9.
8. Kittel, C. (1986) *Introduction to Solid State Physics*, Wiley, New York.
9. Tinkham, M. (1975) *Introduction to Superconductivity*, McGraw-Hill, New York.
10. de Gennes, P. G. (1966) *Superconductivity of Metals and Alloys*, Benjamin, New York.
11. Likharev, K. K. (1979) Superconducting weak links, *Revs. Mod. Phys.* **51**, 101-158.
12. Andreev, A. F. (1964) The thermal conductivity of the intermediate state in superconductors, *Sov. Phys. JETP* **19**, 1228-1231.
13. van Houten, H. and Beenakker, C. W. J. (1991) Andreev reflection and the Josephson effect in a quantum point contact, *Physica B* **175**, 187-197.
14. Mead, C. A. and Spitzer, W. G. (1964) Fermi Level Position at Metal-Semiconductor Interfaces, *Phys. Rev.* **134**, 713-716.
15. Nakagawa, A., Kroemer, H., and English, J. H. (1989) Electrical properties and band offsets of InAs/AlSb *n-N* isotype heterojunctions grown on GaAs, *Appl. Phys. Lett.* **54**, 1893-1895.
16. Silver, A. H., Chase, A. B., McColl, M., and Millea, M. F. (1978) Superconductor-Semiconductor Device Research, *Future Trends in Superconductive Electronics*, Charlottesville, VA, J. B. S. Deaver, C. M. Falco, H. H. Harris, and S. A. Wolf, Eds., Am. Inst. Phys. Conf. Ser., vol. 44, Am. Inst. Physics, pp. 364-379.
17. Clark, T. D., Prance, R. J., and Grassie, A. D. C. (1980) Feasibility of hybrid Josephson field effect transistors, *J. Appl. Phys.* **51**, 2736-2743.
18. Takayanagi, H., Akazaki, T., Nitta, J., and Enoki, T. (1995) Superconducting Three-Terminal Devices Using an InAs-Based Two-Dimensional Electron Gas, *Jpn. J. Appl. Phys.* **34**, 1391-1395.
19. Akazaki, T., Nitta, J., and Takayanagi, H. (1995) Superconducting Junctions using a 2DEG in a Strained InAs Quantum Well Inserted into an InAlAs/InGaAs MD Structure, *IEEE Trans. Applied Supercond.* **5**, 2887-2891.

FIELD EFFECT TRANSISTOR AS ELECTRONIC FLUTE

M. I. DYAKONOV and M. S. SHUR*
A. F. Ioffe Physico-Technical Institute, St. Petersburg, 194021, Russia
Department of Electrical Engineering
**University of Virginia, Charlottesville, VA 22903-2442, USA*

Abstract.

When electron-electron collisions are more frequent than electron collisions with impurities and phonons, electrons are described by hydrodynamic equations. Many new interesting physical phenomena, such as wave instability, shock waves, turbulence, and choking, should occur in this electron fluid. Plasma effects in is a High Electron Mobility Transistor should allow us to design a new family of solid state devices - a FET emitting far infrared radiation, an electronic flute, and a terahertz detector and mixer. These devices should be able to push a three terminal device operation into a much higher frequency range than has been possible for conventional, transit time limited regimes of operation.

1. Introduction.

The development of silicon technology has led to a dramatic reduction in device sizes (from 10 μm or so in the nineteen sixties to sub-0.1 μm in short silicon MOSFETs in the late nineteen nineties). This reduction resulted in a decreasing number of electrons in the device channel. However, the electron **density** has actually increased because the gate voltage swing does not scale proportionally to the decreasing thickness of the gate oxide. This is illustrated by Fig. 1, which shows the calculated dependence of the surface and volume electron densities in the MOSFET channel on the gate length.

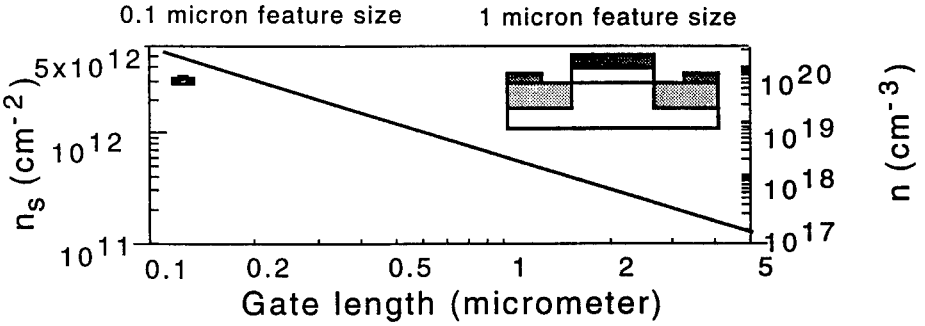


Fig. 1. Surface and volume electron concentration in Si MOSFETs versus gate length. Parameters used in the calculation: gate length to oxide thickness ratio $L/d = 25$, gate voltage swing, $U = 1$ V, temperature $T = 300$ K. Also shown are the relative sizes of 0.1 micron and 1 micron gate MOSFETs.

S. Luryi et al. (eds.), Future Trends in Microelectronics, 251–261.

As seen from the figure, in a 0.1 μm Si MOSFET, the volume carrier concentration is on the order of 10^{20} cm^{-3}, which corresponds to a weakly degenerate and highly non-ideal electron gas at room temperature. At room temperature, the electron thermal velocity in silicon, $v_{th} = (3k_BT/m) \approx 1.1 \times 10^7$ cm/s, is of the same order as the Fermi velocity. (In this estimate, we used the effective mass of density of states in Si, $m \approx 1.1\ m_e$, where m_e is the free electron mass.). The average distance between electrons is $R_s = n_s^{-1/2} \approx 40$ Å (of the same order as the Bohr radius), and the electron-electron collision time is on the order of $\tau_{ee} = R_s / v_{th} \approx 0.4 \times 10^{-14}$ s, much smaller than the collision time with impurities and phonons (i. e. the momentum relaxation time, $\tau = \mu_n m/e \approx 0.2$ ps at room temperature). Here e is the electronic charge and μ_n is the low field mobility. Under such conditions, electrons in the MOSFET channel should behave as a two dimensional (2D) electron fluid, i. e. they should be governed by hydrodynamic equations.

The gate electrode in a FET (see Fig. 2) is separated from the channel by the gate insulator [which is a silicon dioxide layer in a MOSFET and a doped or undoped wide band gap semiconductor, such as AlGaAs in a typical High Electron Mobility Transistor (HEMT)]. The surface concentration, n_s in the FET channel is given by

$$n_s = CU/e \qquad (1)$$

where C is the gate capacitance per unit area. Eq. (1) represents the usual gradual channel approximation [2] which is valid when the characteristic scale of the potential variation in the channel is much greater than the gate-to-channel separation, d.

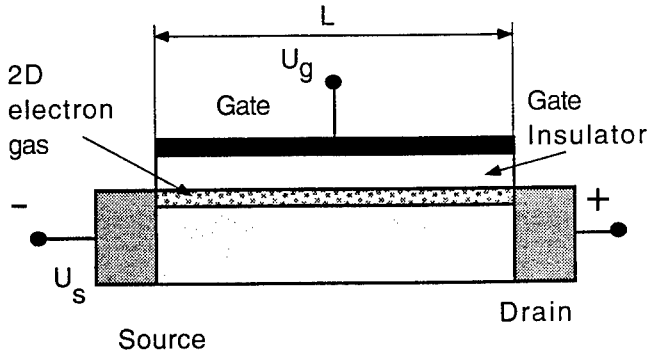


Fig. 2. Schematic structure of a FET. [1] In a Ballistic FET, the gate length, L, is much smaller than the mean free path, λ, but much longer than the mean free path for electron-electron collisions, λ_{ee}.

The equation of motion (the Euler equation) is

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = -\frac{e}{m}\frac{\partial U}{\partial x} \qquad (2)$$

where $\partial U/\partial x$ is the longitudinal electric field in the channel, $v(x, t)$ is the local electron velocity, and m is the electron effective mass. Eq. (2) has to be solved together with the usual continuity equation which (taking Eq. (1) into account) can be written as:

$$\frac{\partial U}{\partial t} + \frac{\partial (Uv)}{\partial x} = 0 \qquad (3)$$

We notice that eqs. (2) and (3) coincide with the hydrodynamic equations for shallow water (see, for example, [3]). Hence, the 2D electron fluid in a Ballistic FET should behave like shallow water. In this hydrodynamic analogy, v corresponds to the fluid velocity, and eU/m corresponds to gh where h is the shallow water level and g is the free fall acceleration.

This analogy has profound consequences for understanding of interesting physics of 2D electrons in the Ballistic FET. Phenomena similar to wave and soliton propagation, hydraulic jump, and the "choking" effect [4, 5] should take place in this hydrodynamic electron fluid. The effects of collisions, surface scattering, changes in the channel cross section, and others may be also understood using this analogy.

The waves propagating in this 2D electron fluid are plasma waves, and their dispersion law (which can be easily obtained from linearized equations (1) and (2)) is similar to that for shallow water waves (or sound waves):

$$\omega = sk, \text{ where } s = \sqrt{\frac{eU}{m}} \qquad (4)$$

Here ω is frequency, k is the wave vector, $U = U_{gc}(x) - U_T$ is the gate voltage swing, $U_{gc}(x)$ is the local gate-to channel voltage, U_T is the threshold, and s is the wave velocity. For comparison, the velocity of shallow water waves is $(gh)^{1/2}$, where h is the water level, and g is the acceleration of a free fall.

Allen et al. [6] observed the plasma waves by measuring infra-red absorption in a silicon FET with a grating metal structure on the gate. Tsui et al. [7] reported similar studies which detected infrared emission. The frequency of the infrared radiation was equal to the frequency of the plasma waves which were excited by the drain-to-source current and were tuned by the gate bias. The measured dependence of the plasma wave frequency on the gate voltage swing was in excellent agreement with eq. (4).

In GaAs and related compounds, where electrons have a small effective mass, the velocity, s, of the plasma waves can considerably exceed typical electron drift velocities. (Typical values of s in GaAs are on the order of 10^8 cm/s, ten times larger than typical electron drift velocities.) This speed advantage can be used in a new generation of high speed and ultra-high frequency electronic devices.

In this paper, we review four such novel proposed devices - an "electronic flute" (which uses a resonant structure for plasma waves) [8], a terahertz HEMT oscillator [1], a detector operating in terahertz range [9], and a terahertz frequency mixer. [9]

2. Electronic Flute.

Since, as discussed above, the behavior of the plasma waves in 2D systems is governed by the same equations as for sound waves, resonant structures, similar to those in musical instruments, may be realized for the plasma waves, and these waves can be excited by a direct current just like wind musical instruments are excited by air jets. However, the plasma wave velocity is much higher than the sound velocity and the FETs are very small. As a consequence, the plasma wave frequencies are in the terahertz range. These plasma waves are accompanied by a variation of a dipole moment created by charges in the FET channel and mirror image charges in the gate and, hence, should cause the emission of far infrared (terahertz range) electromagnetic radiation.

Plasma waves are similar to shallow water waves or to sound waves since

they have a linear dispersion law. In turn, shallow water behavior is similar to the dynamics of a gas with pressure proportional to the square of the density, (see, for example, [3]). Thus, the nonlinear hydrodynamic equations for the 2D electron fluid are similar to (but not identical with) the equations for a real gas, such as air. However, the linearized equations describing small-amplitude plasma waves in a FET and sound waves in a gas are identical. Since the linearized equations determine the instability threshold for a steady flow (i. e. the wave generation threshold), the instability conditions for a real gas and for a 2D electron fluid should be similar provided that the Reynolds numbers and quality factors of resonance cavities are the same. Below, we will show that these parameters for a HEMT structure may be of the same order of magnitude as for a conventional flute.

Fig. 3a (from Ref. [10]) shows the schematics of a jet driven wind musical instrument of a flute family. An air jet excites a resonant cavity formed by the pipe closed at both ends. A similar structure can be realized using a modulation doped gated device (see Fig. 3b). Hence, we call the part of this device shown below the dashed line in Fig. 3b an electronic flute. The remaining gated portion of the device connected to the drain is similar to the outside air space for a jet driven musical instrument. In the electronic flute, a direct current flow excites plasma waves in the resonance cavity in the same way as an air jet excites sound waves in an acoustic cavity.

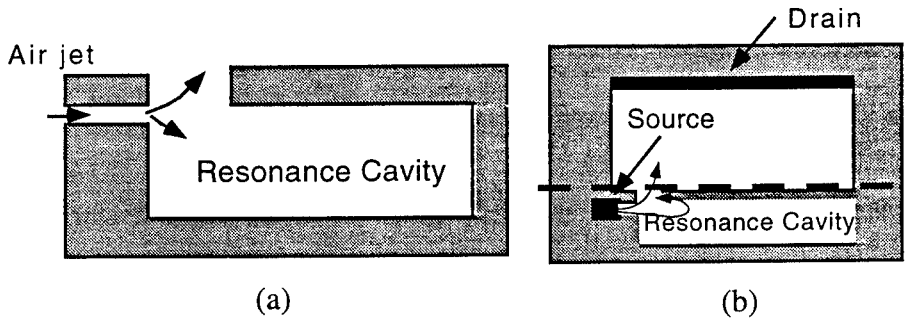


Fig. 3 a. Simplified diagram of a jet driven pipe musical instrument of a flute family. Arrows show the direction of air flow.
b. Electronic flute. White areas show the gated region of the device with the 2D electron fluid in the channel. The part of this device shown below the dashed line is the electronic flute. The top gated portion of the device connected to the drain is similar to the outside air space for a jet driven musical instrument. Arrows show the streamlines of the electron current, which flows between the source and the drain. [8]

Unfortunately, the plasma waves are confined within the electron fluid and cannot be directly enjoyed by a grateful audience. However, the plasma oscillations in the cavity are coupled to electromagnetic radiation. Indeed, the plasma oscillations lead to a variation of the dipole moment created by the electron charge in the channel and the positive mirror charge in the gate metal. Since the device dimensions are small compared to the wavelength, λ_{em}, of an electromagnetic wave corresponding to the plasma oscillation frequency ($\lambda_{em} \approx 100$ μm), this dipole behaves like a point dipole emitting electromagnetic radiation. The measurements of this radiation should provide the means of detecting the excited plasma waves. A small size of the dipole leads to a relatively weak coupling. However, this coupling can be greatly improved by using special antenna structures or, better still, phase locked device arrays as discussed below.

Let us now compare the relevant parameters, which are the Reynolds numbers and quality factors for a conventional flute and for an electronic flute (see Table 1). For a conventional wind instrument, a critical dimension is on the order of 1 cm (which is either the pipe diameter or the size of the embouchure hole). The characteristic dimension for the electronic flute is on the order of 1 micron. We choose a relatively small flow velocity of 10 cm/s for the conventional case. The upper bound for the electron flow velocity is limited by the electron saturation velocity (below than approximately 10^7 cm/s). The sound velocity in air is about $3x10^4$ cm/s while the typical value of the plasma wave velocity in a FET is on the order of 10^8 cm/s.

Parameter	Musical Instrument of Flute Family	Electronic Flute
relevant dimension	1 cm	10^{-4} cm
flow velocity	10 cm/s	10^7 cm/s
wave velocity	$3x10^4$ cm/s	10^8 cm/s
viscosity	0.15 cm^2/s	15 cm^2/s
frequency	100 - 10^4 Hz	10^{11} - 10^{13} Hz
Reynolds number	60	60
Quality factor	3-100	10

Table. 1. Parameter comparison for conventional and electronic flutes.

Another important parameter is viscosity. The viscosity of the 2D electronic fluid is of the order of $v_F\lambda_{ee}$ where v_F is the Fermi velocity and λ_{ee} is the mean free path for the electron-electron collisions. As was discussed in [1], for a highly non-ideal electron gas in AlGaAs/GaAs heterostructures with the electron surface concentration, n_s, on the order of 10^{12} cm^{-2} at 77 K, the thermal energy, the Fermi energy, and the Bohr energy are of the same order magnitude, and λ_{ee} is on the order of the inter-electronic distance, $n_s^{-1/2}$. Then we obtain the following estimate for the viscosity of the electron fluid: $\nu=\hbar/m \approx 15$ cm^2/s for GaAs. For comparison, the viscosity of air is about 0.15 cm^2/s. For these parameters, the Reynolds numbers of the electronic flute and the conventional flute are very close. However, we must bear in mind that we chose a relatively low value of the air flow velocity and a relatively high electron flow velocity, and the Reynolds numbers under normal operating conditions may be somewhat higher for real musical instruments than for the electronic flute.

For the electronic flute, the quality factor of the resonance cavity is on the order of $s\tau/L$ where s is the plasma wave velocity, τ is the electron collision time with phonons and impurities, and L is the size of the cavity. For comparison, the parameter that determines whether a FET behaves as a ballistic device is $v\tau/L$ where v is the drift velocity. Since v may be much smaller than s, the criterion $v\tau/L >> 1$ is much harder to meet that the condition of a high quality factor $s\tau/L >> 1$. Hence, the electronic flute can have much larger dimensions than a ballistic device. For a high mobility modulation doped structure, the momentum relaxation time at cryogenic temperature may be larger than 10 ps. For $s = 10^8$ cm/s, $s/\tau = 10$ micron. Hence, a 1 micron size cavity will have a quality factor of approximately 10. For this plasma wave velocity and this cavity size, the resonant frequency, $s/(4L) \approx 250$ GHz. With a 0.25 micron

cavity, the electronic flute should operate in a terahertz range. (Quality factors for wind musical instruments in Table 1 are from Reference [8].)

In the device shown in Fig. 4, the plasma wave oscillations are excited in the resonance cavities by the current flowing between the source and drain. For the optimum coupling, the basic structure can be repeated many times and this device may be made as an array with dimensions equal to a quarter of the wave length of the electromagnetic radiation with the same frequency. (Just as for musical instruments, the optimum design of the electronic flute is a matter of skill and art.)

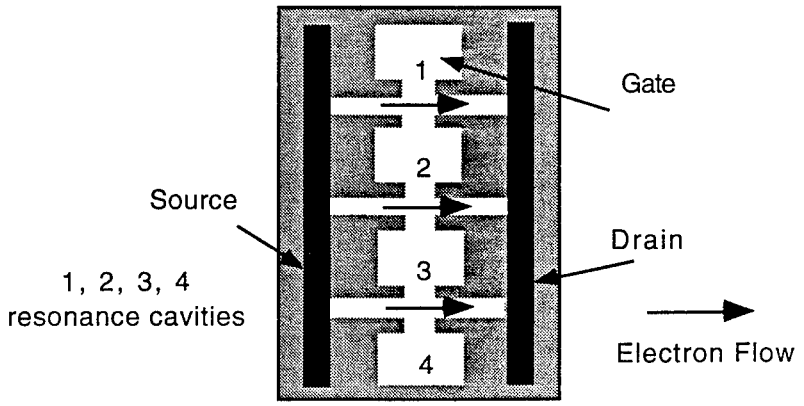


Fig. 4. Array of electronic flutes with a more efficient coupling of plasma waves to electromagnetic radiation. The plasma wave oscillations are excited in the peripheral resonance cavities by 2D electrons flowing from the source to the drain. [8]

Thus, a complete similarity between the plasma waves in a FET and sound waves led us to believe in the possibility of realizing an electronic flute based on the excitation of plasma waves by a direct current in gated modulation doped structures. This electronic flute should operate in a terahertz range of frequencies and emit far infrared radiation.

3. Shallow Water Wave Instability. Ballistic FET as a Terahertz Oscillator.

Early theories of ballistic transport in semiconductors considered planar *n-i-n* structures where the electron ballistic motion was somewhat similar to that in electron tubes [11,12]. Experimental studies of ballistic effects primarily involved vertical devices called hot electron transistors [13]. Ideally, in this mode of ballistic transport, electrons travel across the active region of a ballistic device with no collisions.

As was shown in [1], the situation is quite different in a short field effect transistor where electrons experience practically no collisions with phonons and/or impurities during the transit time (we call such a device a Ballistic FET) but where a high electron concentration results in many electron-electron collisions. In this case, as was mentioned in the introduction, individual electrons cannot be considered as ballistic particles but the two dimensional (2D) electron gas as a whole will exhibit interesting hydrodynamic behavior. As we showed in [1], the steady state of a current-carrying Ballistic FET is unstable.

Let us consider an AlGaAs/InGaAs High Electron Mobility Transistor similar to one described in [14] At 77 K and 300 K, the momentum relaxation time in a 2D electron gas in InGaAs (where ionized impurity scattering is suppressed) is $\tau_p \approx 10^{-11}$ s and $3.5\text{x}10^{-13}$ s, respectively. For the electron drift velocity of 10^7 cm/s, the electron transit time is 10^{-11} s, 10^{-12} s, and $3\text{x}10^{-13}$ s for a 1 μm, 0.1 μm, and 0.03 μm gate length, respectively. Hence, the electron transit time can definitely be made shorter than the momentum relaxation time at 77 K and perhaps even at 300 K. However, for a typical surface carrier concentration of the 2D electron gas $n_s = 10^{12}$ cm^{-2}, the mean free path for electron-electron collisions is only of the order of the average distance between electrons, i. e. of the order of 100 Å since the average distance between electrons at this concentration is close to the Bohr radius (≈ 100 Å) and hence the electron gas is highly non-ideal. Thus, the number of electron-electron collisions during the transit time is large. We also assume that the electron gas is not degenerate since electron-electron collisions are suppressed in a strongly degenerate electron gas (when the Fermi level is more than four thermal energies above the bottom of the subband) because of the Pauli principle.

Let us first assume that the gate voltage swing is fixed at U_o and the channel current is zero. The wave dispersion law, $k = \pm\omega/s$, corresponding to the well known shallow water waves is readily obtained from the linearized system of eqs. (2) and (3).

If the electrons move with a velocity v_o corresponding to the electron flux per unit width $j = n_s v_o = CU_o v_o/e$, this dispersion relation becomes $k = \omega/(v_o \pm s)$. This change in the dispersion relation means that the waves are carried along by the flow.

We now consider the situation when the source and drain are connected to a current source and the gate and source are connected to a voltage source, U_{gs}. This corresponds to the constant value of $U = U_o$ at the source ($x = 0$) and to the constant value of the current at the drain ($x = L$). The *ac* variation of the electric current at source side of the channel is possible even for a constant external current since the *ac* current at the source is short circuited to the gate by the *dc* voltage source. These boundary conditions correspond to a zero impedance at the source and an infinite impedance at the drain and are analogous to those for a transmission line, short circuited at one end and open at the other end. However, in contrast to the transmission line, the wave velocities in our system differ for the opposite directions of propagation.

As was shown in [1], this velocity difference leads to the instability of the steady electron flow with respect to plasma wave generation. Indeed, let us consider the temporal behavior of a small fluctuation superimposed on a steady uniform flow.

We let $v = v_o + v_1\exp(-i\omega t)$, $U = U_o + U_1\exp(-i\omega t)$, linearize eqs. (2), (3) with respect to v_1 and U_1, and use the boundary conditions $U_1(0) = 0$ and $\Delta j(L) = 0$ (i. e., $U_o v_1(L) + v_o U_1(L) = 0$) as discussed above. This procedure leads to the following expressions for the real and imaginary parts of $\omega = \omega' + i\omega''$:

$$\omega' = \frac{\left|s^2 - v_o^2\right|}{2Ls}\pi n \qquad \omega'' = \frac{s^2 - v_o^2}{2Ls}\ln\left|\frac{s+v_o}{s-v_o}\right| \tag{5}$$

where n is an odd integer for $|v_o| < s$ and an even integer for $|v_o| > s$. Eq. (5) shows that for positive v_o, the steady flow is unstable if $v_o < s$. and stable if $v_o > s$. If v_o is negative (or, in other words, if the boundary conditions at the source and drain are

interchanged), the flow is stable if $|v_o| < s$ and unstable if $|v_o| > s$. The wave increment in units of $s/(2L)$ depends only on the Mach number, $M = v_o/s$. For $M << 1$, $\omega'' = v_o/L$ which is the inverse electron transit time.

The reason for the instability becomes clear if we consider the wave reflection from each boundary of the FET channel. The solution of linearized eqs. (2) and (3) shows that the reflection does not change the wave amplitude at $x = 0$ (where the voltage is fixed), while at $x = L$ (where the current is fixed), the amplitude ratio of the reflected and oncoming waves is $(s+v_o)/(s-v_o)$. Hence, the reflection from the boundary with the fixed current results in the wave amplification for $v_o < s$. Let $\tau = L/(s+v_o) + L/(s-v_o)$ be the time during which the wave travels from the source to the drain and back. During time t, the wave amplitude grows in $[(s+v_o)/(s-v_o)]^{t/\tau}$ times, since t/τ is the number of wave round passages during time t. Equating $[(s+v_o)/(s-v_o)]^{t/\tau}$ to $\exp(\omega'' t)$, we obtain the same expression for ω'' as in eq. (5). Thus, the proposed new mechanism of plasma wave generation is based on the amplification of the wave during its reflection from the boundary where the current is kept fixed.

We are unaware of any observations of such an instability of low velocity flows in shallow water. (The required boundary condition of a fixed flow at the drain end of a water channel is an unusual one.)

There are two decay mechanisms which oppose the wave growth: external friction related to electron scattering by phonons or impurities, and internal friction caused by the viscosity of the electron fluid. The external friction can be accounted for by adding the term $-v/\tau_p$ into the right-hand side of eq. (2). This leads to the addition of the $-1/(2\tau_p)$ term to the wave increment. Hence, the wave grows only if the number of scattering events during the transit time is small. The viscosity, ν, of the electron fluid causes an additional damping with the decrement of νk^2 where k is the wave vector. Hence, the viscosity is especially effective in damping higher order modes. Comparing ω'' with νk^2 for the first mode, we find that the effect of the viscosity for $v_o \le s$ is small when the Reynolds number $\mathrm{Re} = Lv_o/\nu$ is much greater than unity. In a highly non-ideal electron gas where the Bohr energy, thermal energy, and Fermi energy are of the same order which roughly corresponds to the surface electron concentration of 10^{12} cm^{-2} at 77 K, the viscosity of the electron fluid, $v_F\lambda_{ee}$ (where v_F is the Fermi velocity) is on the order of $\hbar/m$ which is approximately 15 cm^2/s (comparable to that of castor oil or glycerin at room temperature). The Reynolds number of our electron fluid may be estimated as $\mathrm{Re} = mv_oL/\hbar \approx 12$ for $v_o = 10^7$ cm/s and $L = 0.2$ μm.

For a sample with $L = 0.2$ μm at 77 K, assuming $\tau_p \approx 10^{-11}$ s, the increment v_o/L exceeds the decrement $1/(2\tau_p)$ caused by the collisions when $v_o > 10^6$ cm/s. For the same sample, the decrement caused by viscosity, $\nu(2\pi/L)^2/16$, is smaller then the increment v_o/L when $v_o > \pi^2\nu/(4L) \approx 1.8\times10^6$ cm/s. Hence, the threshold velocity for the instability is well below the peak velocity in GaAs.

Once the electron velocity exceeds the threshold, the plasma waves grow. Since no other steady states exist for $v_o < s$, this growth should lead to oscillations for

which the plasma wave amplitude is limited by non-linearity.

Let us now discuss possible applications of this instability. The plasma oscillations result in a periodic variation of the channel charge and the mirror image charge in the gate contact, i. e. to the periodic variation of the dipole moment. This variation should lead to electromagnetic radiation. The device length is much smaller than the wavelength of the electromagnetic radiation, λ_R, at the plasma wave frequency. (The transverse dimension, W, may be made comparable to λ_R.). Hence, the Ballistic FET operates as a point or linear source of electromagnetic radiation. Many such devices can be placed into a quasi-optical array for power combining. The maximum modulation frequency is still limited by the transit time (≈ 2 ps in our example).

4. Detection and Mixing of Terahertz Radiation by Two Dimensional Electronic Fluid

As we discussed above, plasma waves may be coupled to electromagnetic radiation. Conversely, electromagnetic radiation can excite plasma waves, and, therefore, a FET has a resonance response at the plasma wave frequency. The half width of the resonance curve is determined by the inverse momentum relaxation time. As we discussed in [1] and above, the appropriate boundary conditions for the plasma waves are the short circuited source and the open circuit drain of the FET channel. In [8], we showed that these boundary conditions lead to the resonance detection and mixing of electromagnetic radiation at terahertz frequencies.

Fig. 5. shows an equivalent circuit of the FET detector or mixer.

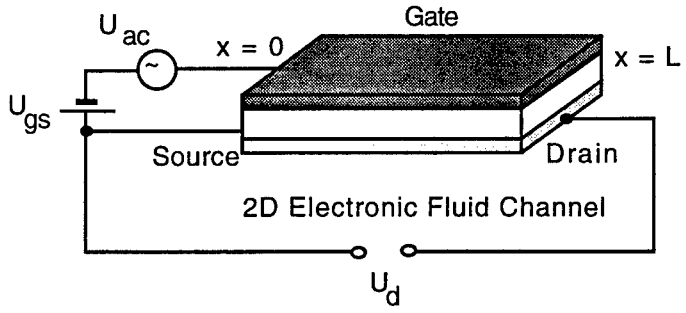


Fig. 5. Schematic geometry of FET operating in detector mode.

Fig. 5 corresponds to a slot antenna design providing an ac voltage, U_{ac}, between the source and gate. The applied constant gate-to-source bias, U_{gs}, determines the velocity of the plasma waves, s. As a function of frequency, the output dc voltage U_d has a typical resonance response with the resonance frequency $\omega_o = \pi s/(2L)$ where L is the channel length. The quality factor of the resonance, $Q = s\tau/L$. For comparison, the parameter that determines whether a FET behaves as a ballistic device is $v\tau/L$. Since v may be much smaller than s, the criterion $v\tau/L >> 1$ is much harder to meet that the condition of the high quality factor for plasma waves, $s\tau/L >> 1$. In a HEMT structure, the momentum relaxation time at cryogenic temperature may be larger than 10 ps (for GaAs, $\tau \approx 12$ ps corresponds to the electron mobility of 300,000 cm^2/Vs).

For $s = 10^8$ cm/s, $s\tau = 10$ micron which corresponds to the quality factor of 50 for L = 0.5 micron.

The viscosity of the electronic fluid may also contribute to damping. The kinematic viscosity is on the order of $\hbar/m \approx 18$ cm^2/s, for $m = 0.063\ m_e$. The corresponding quality factor $Q_v = sLm/\hbar$. For $s = 10^8$ cm/s and $L = 0.5$ micron, $Q_v \approx$ 270.

For the equivalent circuit shown in Fig. 5, the solution of eqs. (1) - (3) shows that for the frequencies, ω, such that $|\omega - \omega_o| << \omega_o$, where $\omega_o = \pi s/(2L)$ is the fundamental mode frequency:

$$\frac{U_d}{U_o} = \left(\frac{QU_a}{U_o}\right)^2 \frac{\gamma^2/4}{(\omega-\omega_o)^2+\gamma^2/4} \qquad (6)$$

Hence, an electronic fluid in a FET should behave as a resonant detector of the electromagnetic radiation with the resonance quality factor $s\tau/L$. A similar resonant response may be shown to exist at odd harmonics of the fundamental frequency ω_o.

In practical systems, mixing of weak incoming signal, $U_s \cos\omega_s t$ with a strong local oscillator signal, $U_{loc} \cos\omega_{loc} t$, is often more desirable because of a much higher sensitivity. As discussed in [8], the equation describing the electron fluid mixer response coincides with eq. (6), where U_a^2 is replaced with $U_s U_{loc}$, provided $|\omega - \omega_o| << \gamma$.

Fig. 6 shows the fundamental resonant frequency versus the gate voltage swing. Fig. 7 shows the calculated responsivity as a function of frequency.

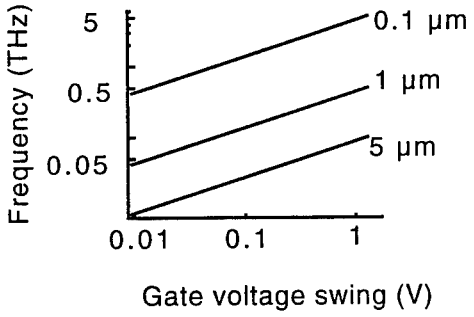


Fig. 6. Resonant frequency versus the gate voltage swing for 0.1 μm, 0.25 μm, 0.5 μm, and 1 μm HEMTs. Parameters used in the calculation: $\tau = 10$ ps, $m = 0.063\ m_o$. [8]

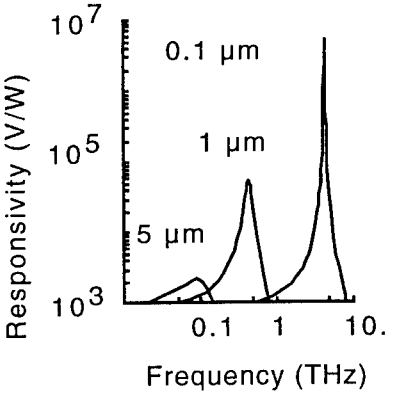


Fig. 7. Responsivity as a function of frequency for 0.1 μm, 0.25 μm, and 0.5 μm HEMT detectors. Parameters used in the calculation: $\tau = 10$ ps, $m = 0.063\ m_o$. [8]

These figures clearly show that this detector/mixer has the highest responsivity when the device length is small, and the device operates in a terahertz range. As seen, the HEMT responsivity may greatly exceed the responsivities of Schottky diodes (on the order of 10^3 V/W) used as detectors and mixers in a terahertz range.

5. Conclusions.

In small FETs with a high concentration of electrons in the channel, the electron behavior is governed by hydrodynamic equations. Devices using plasma waves propagating in this electron fluid, such as an oscillator, an "electronic flute", a detector, and a mixer, should operate at much higher frequencies than those possible for conventional, transit-time limited regimes of operation.

6. Acknowledgment.

The authors are grateful to Professor Robert Weikle for useful discussions and comments. The work at the Ioffe Institute has been partially supported by the Russian Government, by the International Science Foundation, and by the US Army through its European Research Office. The work at the University of Virginia has been partially supported by the US Army Research Office (Project Monitor Dr. John Zavada) and by the Office of the Naval Research (Project Monitor Dr. Yoon Soo Park).

7. References.

1. Dyakonov M. I. and Shur M. S. (1993) Shallow Water Analogy for a Ballistic Field Effect Transistor. New Mechanism of Plasma Wave Generation by DC Current, *Phys. Rev. Lett.*, **71**, 2465
2. Shur M. S. (1990) *Physics of Semiconductor Devices*, Prentice Hall, New Jersey
3. Landau L. D. and Lifshitz E. M. (1966) *Fluid Mechanics*, Pergamon, New York
4. Streeter V. L. and Wylie E. B. (1985) *Fluid Mechanics*, ch. 7, McGraw Hill, New York
5. Dyakonov M. I. and Shur M. S. (1995) Choking of Electron Flow - A Mechanism of Current Saturation in Field Effect Transistors, *Phys. Rev.* **B51**, 14341
6. Allen, Jr., S. J., Tsui D. C., and Logan R. A. (1977) *Phys. Rev. Lett.*, **38**, 980
7. Tsui D. C., Gornik E., and Logan R. A. (1980) *Solid State Comm.*, **35**, 875
8. Dyakonov M. I. and Shur M. S. (1995) Two Dimensional Electronic Flute, *Appl. Phys. Lett.*, August 21,
9. Dyakonov M. I. and Shur M. S. (1995) Detection and Mixing of Terahertz Radiation by Two Dimensional Electronic Fluid, *in the Proceedings of 22d International Symposium on GaAs and Related Compounds*, Cheju, Korea, Aug. 28- Sep. 2
10. Fletcher N. H. and Rossing T. D. (1991) *The Physics of Musical Instruments*, Springer-Verlag, New York
11. Shur M. S. and Eastman L. F. (1979) *IEEE Trans. Electron Devices*, **ED-26**, 1677
12. Shur M. S. (1987) *GaAs Devices and Circuits*, Plenum, New York
13. Heiblum, M., Nathan M. I., Thomas D. C., and Knoedler C. M. (1985) *Phys. Rev. Lett.*, **55**, 2200
14. Chao P. C., Shur M. S., Tiberio R. C., Duh K. H. G., Smith P. M., Ballingall J. M., Ho P., and Jabra A. A. (1989) *IEEE Trans. Electron Devices*, **ED-36**, 461

HETERODIMENSIONAL TECHNOLOGY FOR ULTRA LOW POWER ELECTRONICS

M. S. SHUR, W. C. B. PEATMAN*, M. HURT, R. TSAI, T. YTTERDAL, and H. PARK
University of Virginia, Charlottesville, VA 22903-2442, USA
**Advanced Device Technologies, Inc. ,Charlottesville, VA 22903*

Abstract.

We describe novel heterodimensional devices utilizing Schottky barriers to a 2D electron gas. These devices include a 2D-3D Schottky diode, a 2D AlGaAs/GaAs Schottky Gated Resonant Tunneling Transistor, a 2D AlGaAs/InGaAs MESFET, and a Coaxial MESFET. These devices hold promise of ultra low power high speed operation. The 1 micrometer wide 2D MESFET, which has a very low output conductance and steep subthreshold slope, exhibited the highest transconductance of any 1 micron wide device.

1. Introduction.

All semiconductor devices utilize interfaces between different regions -- ohmic, *p-n* junctions, Schottky barrier junctions, heterointerfaces, interfaces between a semiconductor and an insulator. Typically, these interfaces are planes separating different regions. However, recently a new generation of semiconductor devices has emerged. These devices utilize interfaces between semiconductor regions of different dimensions and are called heterodimensional devices. An example of such an interface is a Schottky barrier between a three dimensional (3D) metal and a two dimensional electron gas. Other possible configurations include the interface between a two dimensional electron gas and a two dimensional Schottky metal, an interface between a one dimensional electron gas and a two dimensional Schottky metal, and an interface between a one dimensional electron gas and a three dimensional Schottky metal.

Different heterodimensional Schottky contacts have several features in common - smaller capacitance because of a smaller effective cross section and a wider depletion region, a high carrier mobility related to properties of the two dimensional (2D) electron gas, a smaller electric field, and a higher breakdown voltage. (A wider depletion region is caused by fringing electric field streamlines.) These features make these devices very promising for applications in ultra-high frequency varactors and mixers.

In this paper, we describe a new generation of devices utilizing Schottky contacts between a metal and a 2Dimensional electron gas (2DEG). This new high speed heterodimensional contact has unique characteristics which are particularly promising for applications in the fields of millimeter wave electronics and high speed, ultra low power integrated circuits. These devices include a new two terminal

S. Luryi et al. (eds.), Future Trends in Microelectronics, 263–268.

heterodimensional Schottky diode (see Fig. 1 a) and three new heterodimensional transistors which utilize a side-gate formed by plating gate metal into a trench etched through the plane of the 2DEG.

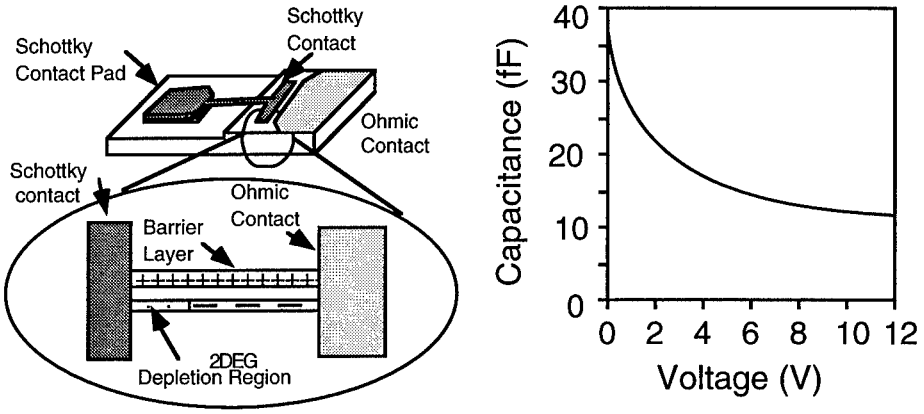


Fig. 1. Schematic structure (a) and measured C-V characteristics (b) of a heterodimensional Schottky diode. [1]

Our heterodimensional transistors have a very small size (i. e. 0.5x0.8 μm^2). These small sizes, small gate-to-channel capacitances, and low parasitic capacitances result in a nearly ideal performance with a very small number of electrons in the channel (see Fig. 2).

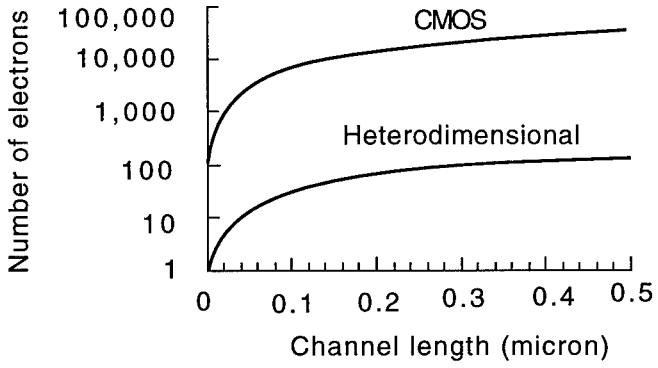
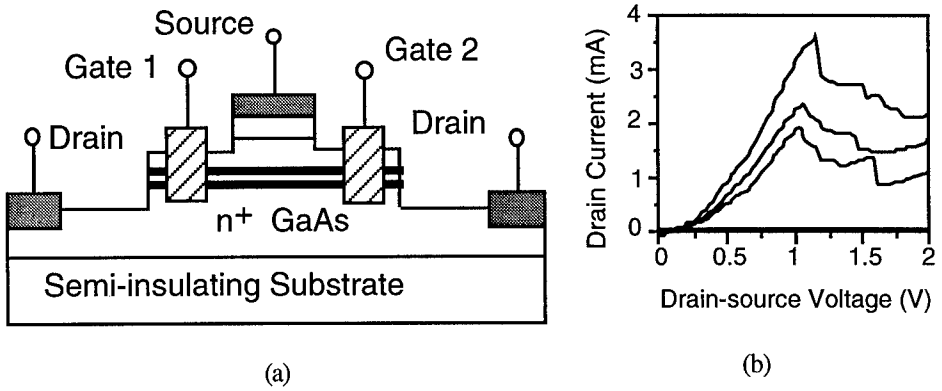


Fig. 2. Number of electrons, N, versus channel length for gate voltage swing, V_{gt} = 0.5 V. For CMOS, $N = C_i WLV_{gt}/q$. For Heterodimensional Transistors, $N = C_h LV_{gt}/q$, where $C_i = \varepsilon_o\varepsilon_i/d$ is the gate capacitance per unit area. C_h is the gate capacitance per unit length, W, L, and d are the channel width and length and the gate dielectric thickness, respectively, $\varepsilon_o\varepsilon_i$ is the silicon dioxide permittivity, *and* q is the electronic charge. In this calculation, we assumed $d = L/20$, $W = 10\ L$, $C_h = 10^{-10}$ F/m.

2. Heterodimensional Transistors.

We fabricated the Schottky-gated resonant tunneling transistor (SG-RTT), which has demonstrated high transconductance at room temperature, see Fig. 3 [2].

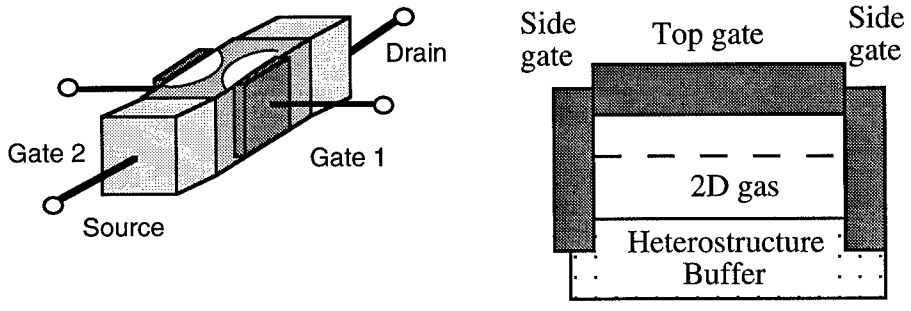


(a) (b)

Fig. 3. Schematic structure of a Resonant Tunneling Transistor (RTT) (a) and RTT current-voltage characteristics 3 (b). [2]

The second transistor is a novel 2Dimensional metal-semiconductor field effect transistor (2D MESFET), see Fig. 4 a. A 2D MESFET is similar to an HFET except that the gates are placed on the sides of the conducting layer. In our device, the conducting layer was a 2D electron gas which was formed in a pseudomorphic AlGaAs/InGaAs heterostructure. The Schottky side gates modulate the width of the 2DEG channel, and therefore the current, between source and drain. The gates are formed by etching through the plane of the conducting layer and by electroplating Pt/Au onto the walls using resist as the mask (metal thickness is easily varied by adjusting the plating parameters). Otherwise, conventional HFET processing techniques are used.

A related device is a coaxial MESFET schematically shown in Fig. 4 b.



(a) (b)

Fig. 4. Schematic Diagram of 2D MESFET (a) and coaxial MESFET (b). [3,4]

In the coaxial MESFET, the electron gas is constricted from four sides and controlled from three sides - by the top Schottky gate and by the two side Schottky gates. This should allow us to achieve a very precise control of the electron density from a 2D gas to a 1D gas to a level of a few electrons in the channel. These devices are particularly promising for low power, high speed integrated circuit applications because it eliminates a narrow channel effect. [3] They also may achieve a high speed because of reduced parasitic capacitance. In a conventional FET, the gate capacitance can be reduced by either decreasing the device area or by increasing the gate-to-channel separation. Both changes in the device dimensions lead to a substantial increase of C_p/C since the parasitic fringing capacitance decreases roughly in proportion to the gate periphery and C decreases roughly in proportion to the gate area. Once the power delay product becomes limited by $C_p\Delta V^2$, where ΔV is the voltage swing, a further decrease in C only leads to the deterioration in speed without any benefit for reducing the power consumption. Hence, we conclude that the decrease in both parasitic and gate capacitance is needed to achieve a low power technology. We also expect a coaxial MESFET to have a very sharp pinch-off and an extremely small leakage current.

Fig. 5 compares qualitative distributions of the electric field streamlines in a conventional HFET, in a 2D-MESFET, and in a coaxial MESFET. As can be seen from this figure, most of the streamlines for a 2D MESFET and even more so for a coaxial MESFET terminate on the gate electrode. Hence, the parasitic capacitance of a 2D-MESFET is smaller than that of a conventional device. This parasitic capacitance is even smaller for a coaxial MESFET. These distribution shows that the heterodimensional devices greatly reduce the detrimental narrow channel effect and, hence, allow us to use narrower and lower power devices, reducing the gate capacitance without a commensurate increase in the relative importance of the parasitic capacitance. Of course, this does not solve the problem of driving the interconnects. The circuit layout for low power electronics must have short interconnects, except for a few long interconnects which have to be driven by special drivers. If the number of such long interconnects is not large, then the share of the drivers in the total power budget can be small or, at least, manageable.

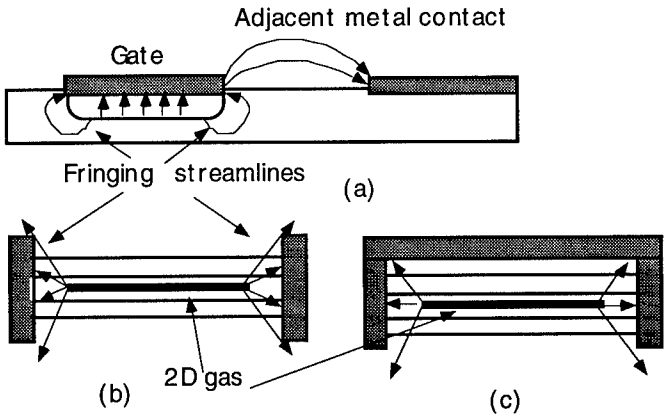


Fig. 5. Device cross sections and electric field streamlines in conventional HFET (a), 2D-MESFET (b), and coaxial MESFET (c). [5]

We demonstrated a 1 μm wide AlGaAs/InGaAs 2D MESFET having a peak drain current and transconductance of 210 mA/mm and 210 mS/mm and the subthreshold slope of 75 mV/decade corresponding to an ideality factor of 1.3 [3] (this value is comparable to that in the state-of-the-art 10 μm wide HFET). From this we estimated the cutoff frequency to be about 21 GHz which is comparable to that of the best 1 μm long HFETs.

By eliminating the narrow channel effect and reducing parasitic capacitances, this new technology enables scaling the gate width to submicron dimensions to achieve a large reduction in the power consumption without loss of speed performance. The Schottky side gates modulate the width of the 2DEG channel and the current between the source and drain. The gates are formed by etching through the plane of the conducting layer and by electroplating Pt/Au onto the walls using resist as the mask. Further details of the fabrication are described in [3]. We fabricated both depletion mode and enhancement mode devices. For a 0.5 μm wide channel device, the threshold voltage was 0.0 V while the knee (ON) voltage was about 0.2 V. [6] Such device should operate at less than 1 V bias. Based on our charge control model of the 2D MESFET, we estimated the power-delay product of 0.1 fJ which is an order of magnitude smaller than state-of-the-art technologies. We also observed a nearly zero threshold voltage shift with temperature and almost total absence of DIBL in these devices. [6]

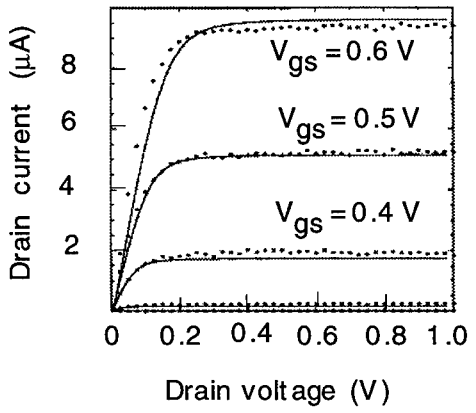


Fig. 6. Measured and simulated current-voltage characteristics of a 0.5 μm 2D MESFET. [6]

3. Quasi-three-dimensional modeling.

The design, optimization, and circuit simulation of 2D MESFETs require a three dimensional (3D) modeling. Such modeling is a prerequisite to the development of 2D MESFET integrated circuits. To our knowledge, no three dimensional device simulators are capable of simulating heterostructure devices. To achieve a 3D simulation of 2D MESFETs, we perform a two dimensional numerical simulation of the device cross-section under different gate biases using ATLAS-II 2D device simulator capable of modeling heterostructures. [7] We then use the interpolated results of the 2D numerical simulation as input to the analytical theory describing the potential

distribution and current in the third dimension. A good agreement with the measured data confirms the validity of this new simulation approach. This general approach of quasi-3D simulation is ideologically similar to the old quasi-2D FET models and can be also applied to short and narrow channel CMOS, short and narrow channel Thin Film Transistors, Surrounding Gate Transistor [8], DELTA transistor [9], and other devices where 3D modeling is required.

4. Conclusion.

Our experimental results and device and circuit simulations show that the heterodimensional technology holds promise of ultra low power operation at a relatively high speed. It should also allow us to study the behavior of single electrons in a wide temperature range.

5. Acknowledgment.

This work at UVa was supported by ONR, Contract N00014-90-J-4006 (Project Monitor Dr. Yoon Soo Park). The work at ADT, Inc. was supported by ONR, Contract N00014-94-C-0260 (Project Monitor Dr. Alvin M. Goodman). The authors thank Quantum Epitaxial Designs, Inc. for providing the MBE material.

6. References.

1. W. C. B. Peatman, T. W. Crowe, and M. Shur, A Novel Schottky/2DEG Diode for Millimeter and Submillimeter Wave Multiplier Applications, IEEE Electron Device Letters, vol. 13, No. 1, pp. 11-13, Jan. (1992)
2. W. C. B. Peatman, E. R. Brown, M. J. Rooks, P. Maki, W. J. Grimm, and M. Shur, Novel Resonant Tunneling Transistor with High Transconductance at Room Temperature, IEEE Electron Device Letters, vol. 15, No. 7, pp. 236-238, July (1994)
3. W. C. B. Peatman, H. Park, and M. Shur, IEEE Electron Dev. Lett., vol. 15, No. 7, pp. 245-247, July (1994)
4. W. C. B. Peatman, B. Gelmont, W. L. Grimm, H. Park, M. Shur, E. R. Brown, and M. J. Rooks, Heterodimensional Schottky-Gate Devices, in Proceedings of 2D International Semiconductor Device Research Symposium, Charlottesville, VA, December, pp. 427-430 (1993)
5. M. Shur, W. C. B. Peatman, H. Park, W. Grimm, and M. Hurt, Novel Heterodimensional Diodes and Transistors, Solid State Electronics, Sep. (1995)
6. W. C. B. Peatman, R. Tsai, T. Ytterdal, M. Hurt, H. Park, J. Gonzales, and M. S. Shur, Sub-half-micron Width 2D MESFET , unpublished
7. M. Hurt, M. S. Shur, W. C. B. Peatman, and P. B. Rabkin, Quasi-three-dimensional Modeling of a Novel 2D MESFET, IEEE Trans. Electron Devices, accepted for publication
8. H. Takato, K. Sunouchi, N. Okabe, N. Nitukyama, K. Hieda, F. Horiguchi, and F. Masuoka, High Performance CMOS Surrounding Gate Transistor (SGT) for Ultra High Density LSIs, IEDM, pp. 222-225 (1988)
9. D. Hisamoto, T. Kaga, Y. Kawamoto, and E. Takeda, A Fully Depleted Lean-Channel Transistor (DELTA) - A Novel Vertical Ultrathin SOI MOSFET, IEEE Electron Dev. Lett., vol. 11, No. 1, pp. 36-38 (1990)

LATERAL CURRENT INJECTION LASERS - A NEW ENABLING TECHNOLOGY FOR OEICS

D.A. SUDA and J.M. XU

University of Toronto
Department of Electrical Engineering
10 Kings College Rd.
Toronto, Ontario, M5S 1A4, Canada

1. Introduction

At present, almost all semiconductor lasers use designs based on the vertical injection of current into the active region. While this scheme has proven successful for discrete laser diodes, it has several intrinsic drawbacks. First of all, there is an incompatibility between vertical injection and the preference for lateral integration. Just as in electronics, the evolution from discrete components to integrated circuits necessitates a change from vertical to lateral devices.

Another drawback concerns the present use of a highly conductive substrate associated with the vertical injection scheme. This makes device-to-device isolation difficult and can result in high parasitic losses and delays. In addition, the injected current must pass through the cladding layers which provide optical and electrical confinement. The resistance inherent to these wide bandgap layers is detrimental to the lasing threshold and lasing efficiency, particularly under high injection conditions. This necessitates a compromise when choosing the cladding layer material and thickness in order to provide adequate optical confinement while avoiding excessive resistance.

Overcoming the resistance becomes critical in the case of Vertical Cavity Surface Emitting Lasers (VCSELs), where the cladding layers are replaced by high reflectivity Distributed Bragg Reflectors (DBRs). Since vertical current injection precludes the use of insulating materials in the DBRs, many alternating layers of semiconductors with relatively small differences in refractive index are needed. This results in a high resistance path for the injected current and much lower lasing efficiency than is theoretically possible. Efforts have been made to alleviate this problem in VCSELs by using elaborate etching and recessed contacts at the expense of further complicating the fabrication. Distributed FeedBack (DFB) lasers also suffer, since the current injection path passes through the corrugated grating region of high interface defect density.

Lateral Current Injection (LCI) lasers have the potential to overcome many of these limitations by making use of the under-explored lateral degree of freedom. In an LCI laser, the planar geometry is well suited for OEIC applications. In addition, current does not pass through the optical cladding layers or the substrate, allowing both these regions

S. Luryi et al. (eds.), Future Trends in Microelectronics, 269–278.

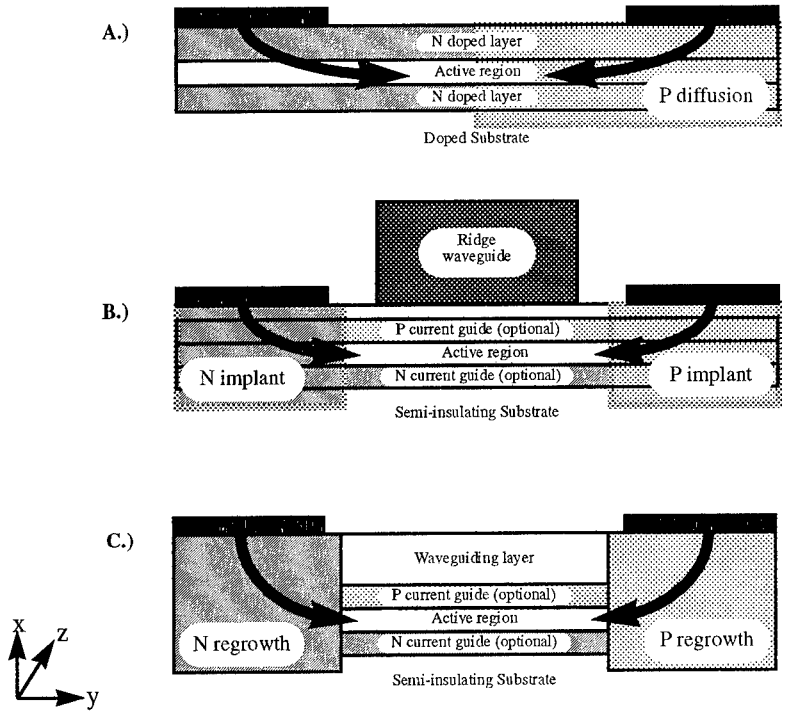


Figure 1. Various LCI laser designs. A.) is a Transverse Junction (TJ) LCI laser, B.) is a PIN structure using a ridge waveguide and implanted contact regions, and C.) is a Buried Heterostructure LCI (BH-LCI) design.

tc be designed for optimal optical confinement and made of wide bandgap semi-insulating material for reduced parasitic loss and capacitance. In this configuration the electrical and optical designs are de-coupled and the cladding layers can be optimized separately from the current injecting regions.

Despite these potential advantages and a considerable amount of experimental effort devoted to this type of laser [1-26], the progress of LCI laser development has been rather slow. In a large part this is due to the predominance of trial-and-error experimental attempts with very little backing from theoretical analysis. The best results from the most recent trials are still inferior to that of state of the art vertical injection designs. Typically, the threshold current is high while the efficiency starts out low and then decreases with increasing current.

It is difficult to determine the origins of these shortcomings from experimental measurements alone. Many interdependent factors are involved, some of them uniquely new. Starting from fundamental laser physics the job is no easier. First principle considerations may be the same as for vertical injection, but new aspects of the electro-optic interaction are involved which require self-consistent analysis. This is what we have done using our 2D self-consistent Finite Element Light Emitter Simulator (*FELES*) [27].

In this paper we will first present a brief summary of the history and current state of

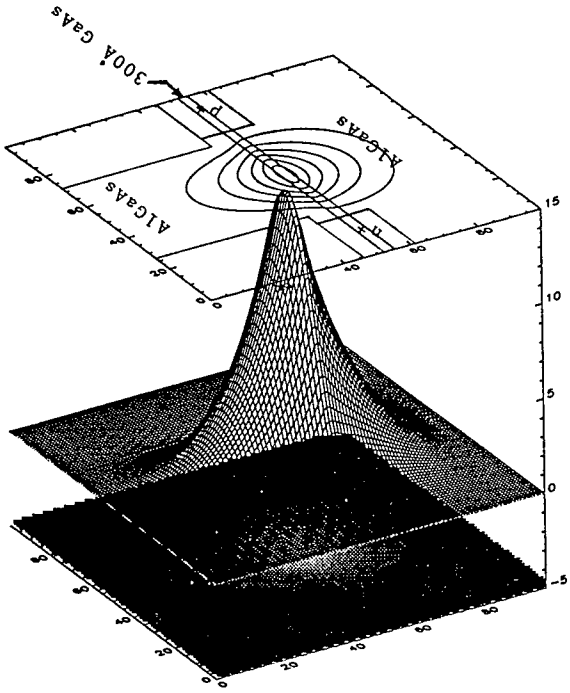


Figure 2. Fundamental optical mode and its contour plot on the facet of PIN ridge waveguide LCI laser.

LCI laser design, followed by the results of our theoretical investigation, and then concluding with future prospects for this type of device in OEICs.

2. Progress to Date

Since lateral injection was first proposed in 1974 [1], LCI laser design has evolved with the capabilities of semiconductor growth and processing technology. Several device configurations have been proposed in the literature (Figure 1). The earliest designs were based on a planer Transverse Junction (TJ) approach, while more recent work has concentrated on PIN structures in combination with either lateral heterojunctions or a ridge for lateral optical confinement.

In the basic transverse homojunction design (Figure 1A), epitaxial growth is used to create an N-type GaAs active layer sandwiched between N-type AlGaAs cladding layers. A P-type dopant is then implanted or diffused over half the structure, resulting in a transverse junction extending the length of the laser. The best measured performance of this design is poor, with a threshold current (I_{th}) of 17 mA [11], mainly because of weak optical confinement in the lateral direction and high losses across the electrically unconfined junction region. TJ VCSELs have also been fabricated, but the reported performance has not been encouraging: $I_{th} = 48$ mA [21].

The performance can be improved considerably by using two implanted or diffused regions, one n-type and one p-type, to inject carriers into an undoped active region (Figure 1B). In this type of design the impurity induced disordering [28] in the doped regions provides optical and carrier confinement in the active region composed of either Multi-Quantum Well (MQW) or bulk material. Separating the active region from the junctions increases the radiative recombination and reduces the extension of the optical mode into the doped regions, thereby reducing losses compared to the basic TJ structure. Our simulations predict that this type of LCI laser, when combined with a semiconductor or dielectric ridge waveguide to strengthen the optical confinement, is capable of adequate performance with threshold currents on the order of a few mAs (Figure 2) [29]. While the simplicity and relative ease of fabrication of this type of LCI laser are appealing, car-

rier confinement in the lateral direction is still rather weak. This problem can be reduced by incorporating many QWs so that the carrier concentration in each well remains relatively low and thus the leakage out of the wells is suppressed.

A design which provides strong carrier and optical confinement in the lateral direction is the Buried Heterostructure LCI (BH-LCI) laser (Figure 1C). In this type of device the implanted or diffused contact regions are replaced with n and p-type regrowth regions. The drawback to this design is the technical difficulty associated with doing two regrowth steps but, as the functional devices already reported illustrate [25,26], this is a difficulty that can be overcome.

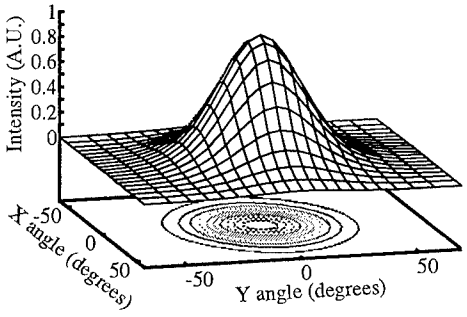


Figure 3. Far field pattern from a *FELES* simulation of the BH-LCI laser reported in [25].

The variations discussed so far are all true lateral injection lasers, that is both electrons and holes are injected into the active region from the sides rather than from the top and bottom. Lasers have also been fabricated where the substrate contact has been moved to the top surface of the device, but the current injection path is still strictly vertical [20,22], or in which one carrier is injected laterally and the other vertically [16,17]. The drawbacks to both these approaches are that the layers above the active region must still be made of highly conductive material and there is often a current crowding problem.

3. Theoretical Analysis of Selected Structures

Much of our work on LCI lasers has involved the theoretical analysis of various designs using *FELES* [30]. In this section, we will present the results of our investigation of the underlying physics of two recently reported BH-LCI structures that illustrate the unique internal operating mechanisms of LCI lasers.

One of the advantages of BH-LCI lasers is the good carrier and optical confinement achievable in both transverse directions. Since the cladding layers outside the active region are not electrically active and serve only to guide the wave, the compositions and widths of these layers as well as that of the active region can be designed to produce a single transverse mode that exhibits an almost perfectly circular far field pattern (Figure 3) without compromising the electrical characteristics of the device.

Our simulations also reveal a key design issue that is rooted in fundamental device physics unique to LCI lasers. The lateral gain profile in the QWs can be rather asymmetric (Figure 3). This problem, which can only be seen clearly through simulation, originates from the disparity in mobilities between electrons and holes. To maintain quasi charge neutrality, the higher mobility electrons tend to be pulled over to meet the lower

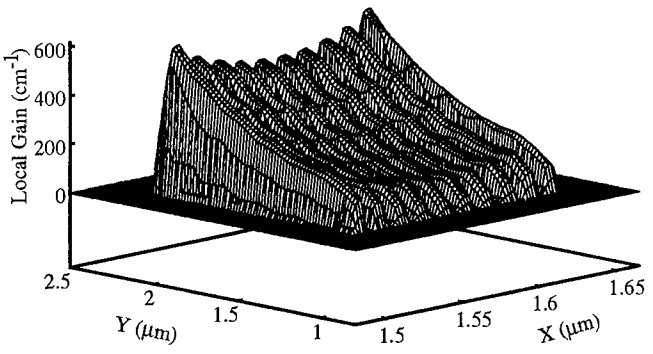


Figure 4. Non-uniform gain profile in a BH-LCI laser due to the difference in electron and hole mobilities.

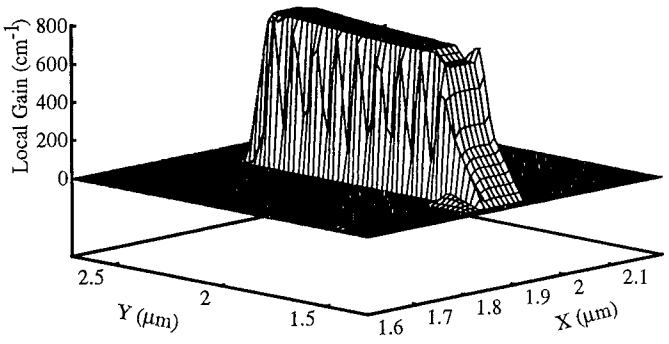


Figure 5. Improved gain uniformity through the use of "current guiding" layers above and below the active region.

mobility holes and to "pile up" near the p-doped/active region interface. When combined with stimulated recombination, this results in a highly asymmetric gain profile which decreases the overlap of the optical mode and the gain peak and ultimately increases threshold and lowers efficiency. In a structure which is multi-mode at zero bias (or becomes multi-mode due to bias dependent changes to the refractive index), the non-uniform gain profile will selectively pump higher order modes and could even cause them to lase before the fundamental mode. This was confirmed in one of our experimental trials.

A promising design approach for minimizing this problem was implemented in a BH-LCI laser recently reported in the literature [26]. In this device, regrown n and p-doped InP regions inject carriers directly into an InGaAsP bulk active layer as in previous LCI laser structures, but in addition carriers are also injected through doped InGaAsP "current guiding" layers above and below the active region. The addition of these current guiding layers results in a design which is a lateral/vertical injection hybrid. Current is not only injected from the ends of the active region as in typical lateral injection designs, but also along the length of the channel as is the case for vertical injection lasers. While the injection paths are now more similar, the differences in fabrication and contacts remain. No current flows through the top cladding layer or the substrate and the geometry is still planar, so the benefits of lateral injection are retained. Our simulations revealed that the use of these thin current guiding layers indeed produces a more uniform gain profile (Figure 3), but at the cost of increased leakage. This is a problem that is specific to particular structures and simulations of design variations can be used to study this trade-off and optimize the design.

We simulated the external characteristics of the structure reported in [26] using *FELES* and found the predicted single mode operation, threshold current, and L-I characteristics to be in good agreement with experiment (Figure 6). An especially striking feature of the experimental results is the steady decrease in differential efficiency with increasing current. This leads to output power saturation at a relatively low level: only 10mW at a current of 150mA. Speculations based on inspection were made regarding the possible causes of this roll-off. Two natural suspects are high non-radiative recombination at the regrowth interface and carrier leakage in the lateral direction.

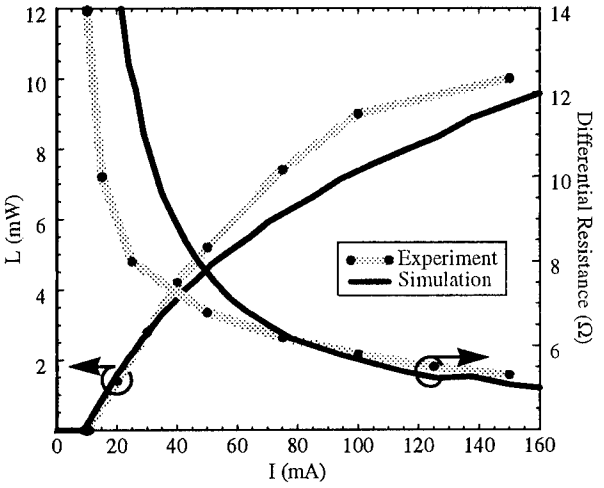


Figure 6. Comparison of *FELES* and experiment for L-I and differential resistance results of the BH-LCI laser reported in [26]

To test the possibility that the steady roll-off was due to a high defect density at the regrowth/active region interfaces, narrow (0.1μm) regions with high non-radiative recombination rates were added to the simulations at these interfaces. These results showed a slight increase in threshold and a decrease in peak optical output power, but even recombination lifetimes as short at 10ps did not significantly alter either the threshold or the efficiency. This can be explained by the fact that the actual area of the vertical regrowth interface through which the current flows is very small and is perpendicular to the direction of current flow. This makes its total effect much smaller than that observed in conventional DFB or buried heterostructure lasers for a given defect density.

We next considered the possibility of carrier leakage out of the active region in the lateral direction. Simulations allowed us to examine the current components in the device and revealed that at a bias of 150mA, less than half of the terminal current was passing through the active region. Since more than half of the current was never even getting into this area of the device, leakage out of this region could not possibly be the explanation of the decrease in differential slope efficiency.

Further simulations showed that leakage current was the major cause of the roll-off, but not the type of out-of-well leakage that is typical of vertical injection lasers. Instead, this decrease in efficiency with bias is mostly due to the activation of parallel conduction paths through the guiding layers and the InP buffer. These layers form wide bandgap diodes which are electrically in parallel with the active region. Since the turn-on voltages of these diodes, especially the PIN diode formed between the regrowth regions through the InP buffer layer (~1.35V), are much larger than the turn-on voltage across the active

region (~0.85V), one might not expect leakage to be a problem. Indeed, this is true for injected currents up to a few mAs as attested to by the low threshold currents of both the experiment and simulation. Under high injection, however, the voltage drop becomes sufficient to bias these parallel diodes near their turn-on voltages. This activates these parasitic conduction paths and results in an exponential increase in leakage current and a corresponding decrease in efficiency.

When the leakage problem is minimized, our analysis indicates that BH-LCI lasers should be comparable to vertical injection lasers in terms of threshold and slope efficiency. Another consideration which naturally arises is the device speed. This is a complex and challenging subject involving such issues as 3D to 2D capture processes and ambipolar transport in 2D space. Our initial investigation into this question indicates that LCI lasers have, at least, the advantage of low parasitic delay.

Since the substrate can be made semi-insulating, the effective area of the n and p contacts is small, and the separation between the contacts is relatively large, the capacitance of LCI lasers should be much lower than vertical injection designs. In confirmation of this, *FELES* simulations predict a zero bias capacitance of 0.8pF for the structure reported in [26], compared to the 0.5pF measured experimentally. While it is difficult to experimentally measure the capacitance under large forward bias, it was easily determined from *FELES* simulations and found to be approximately 15pF with the device biased at twice the threshold current.

In addition, perhaps contrary to one's intuition, both experiment and simulation results exhibit a low differential resistance (Figure 6). This can be explained by the fact that although the conduction path through the active region is narrow, it is composed of high mobility, low bandgap material. In contrast, the conduction path in conventional lasers is quite wide but passes through the large bandgap, low mobility cladding layers.

4. Future Directions

As the basic design problems associated with LCI lasers are solved, many new and exciting research possibilities will open up. The ridge waveguide, freed from the current conduction requirements of vertical injection lasers, can be adapted to a variety of tasks. Technological compatibility with FETs, HBTs, photoreceivers, and waveguides multiplies the number of options.

A dielectric ridge waveguide, in addition to its role in the optical confinement of the laser mode, can also act as "photonic wiring". The laser output could be coupled into these dielectric-loaded waveguides and, when extended over the wafer, they would act as optical interconnects. These interconnects could route the optical output of LI lasers to photoreceivers, phototransistors, and cross-connection networks. But can a dielectric ridge of relatively low index deposited on high index semiconductor layers provide an effective waveguiding structure? Via a series of investigations, we have found that the answer is yes provided that the proper structure and layer designs are in place first. In fact, the waveguiding ability of a dielectric ridge is so wavelength sensitive that it can be a valuable new addition to our "tool box" of wavelength and mode control techniques [31].

The basic structure can be extended to create a VCSEL by replacing the ridge with a DBR. Insulating dielectric material can now be used for the DBR to achieve high reflectivity with fewer layers than in vertical injection designs. Adding Bragg gratings to the base of the ridge would produce a DFB laser. This grating can be fabricated by traditional methods or optically. For the former, the stringent requirement for grating interface quality is removed since no current will flow through the region. For the latter with a photorefractive dielectric ridge, the interference pattern created by two high power laser beams incident on the ridge would "burn-in" the Bragg Reflector. This method results in low absorption and recombination losses and a laser operating wavelength that could be chosen after the main part of fabrication is complete. Carrying this idea a bit further, even the type of laser (cleaved facet, DFB, DBR, VCSEL, etc.) could be chosen after the material growth is complete by depositing dielectric material and then patterning the appropriate reflector (Figure 7).

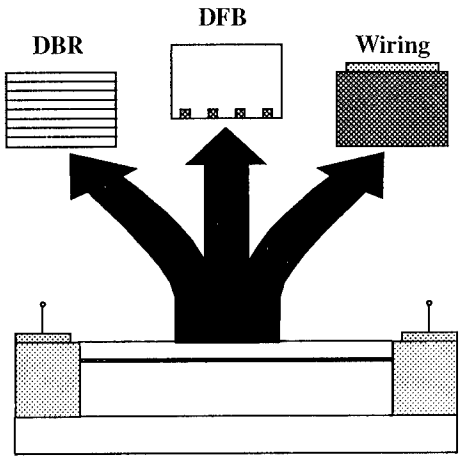


Figure 7. Illustration of the post-processing choices for an LCI laser. Since the ridge need not be electrically conductive, it can be made of dielectric material deposited after the semiconductor growth is complete. The dielectric ridge can form a DBR for a VCSEL, and DFB grating, or photonic wiring.

By adding a metal gate on the top or on the side of the ridge waveguide, both the emission wavelength and the power of the laser can be controlled. The former opens up the possibility of yield enhancement by tuning. Devices on the same wafer inevitably vary in wavelength because of material and/or process variations. With the ability to tune the gain spectrum by gate field via the Stark Effect, we may be able to tune devices fabricated outside the desired frequency range instead of discarding them. The success of this approach could translate into a significant cost reduction in production. Once integrated with logic circuitry, the gate could also be used to compensate for the change in wavelength with operating temperature. The output wavelength could be monitored by a circuit that would control the gate voltage to lock the laser output to a specified wavelength. Therefore lasers using the same design and fabricated on a single wafer could be set to different wavelengths for Wavelength Division Multiplexed (WDM) communications.

The possibility of gate-controlled emission power is even more exciting, as it may realize the long-sought capacitive modulation (gain switching) of lasers via a third terminal [32]. Implementing a gate in the space between the ridge and one of the contact regions effectively turns the structure into an integrated laser-FET module. This type of structure offers the potential of increased modulation speed due to minimized series resistance.

By having the current flow in the lateral dimension, we have opened up the vertical degree of freedom to many additional design possibilities for creating novel devices. For instance, one can easily realize that multi-section lasers can readily be created along the longitudinal direction without compromising the LCI design. This opens the possibility of having multiple electrodes so that the gain or loss of each section can be tuned individually.

5. Summary

As the primary application of photonic devices evolves from discrete components to OEICs, the need for a planar laser technology becomes critical. Lateral current injection lasers have the potential to fill this role and to overcome some of the limitations inherent to vertical current injection structures. Despite a slow but continuous improvement in LCI performance over the past 20 years, this potential has yet to be realized.

One reason for this slow progress has been a lack of theoretical support. Our investigation of the shortcomings of these experimental devices using *FELES* simulations has produced explanations and possible solutions for several of the observed problems. Through a cycle of device fabrication and analysis, it should be possible to make LCI lasers which are competitive with state of the art vertical injection structures.

LCI lasers which are free from these performance limitations will open up many new possibilities in OEICs. These include photonic wiring, post-processing selection of the laser characteristics, novel lasers of high performance, and integrated laser-FET modules.

References

1. Namizaki, H., Kan, H., Ishii, M., and Ito, A., (1974) "Transverse-Junction-Stripe-Geometry DH Lasers with Very Low Threshold Current", *J. Appl. Phys.*, **45**, pp. 2785-2786.
2. Namizaki, H., (1975), "Transverse-Junction-Stripe Lasers with a GaAs p-n Homojunction", *IEEE J. Quantum Elec.*, **11**, pp. 427-431.
3. Nagano, M. and Kasahara, K., (1977), "Dynamic Properties of Transverse Junction Stripe Lasers", *IEEE J. Quantum Elec.*, **13**, pp. 632-637.
4. Kumabe, H., Tanaka, T., Namizaki, H., Ishii, M., and Susaki, W., (1978), "High-Temperature Single Mode CW Operation with a Junction-Up TJS Laser", *Appl. Phys. Lett.*, **33**, pp. 38-39.
5. Ury, I., Matgalit, S., Yust, M., and Yariv, A., (1979), "Monolithic Integration of an Injection Laser and a Metal Semiconductor Field Effect Transistor", *Appl. Phys. Lett.*, **34**, pp. 430-431.
6. Nita, S., Namizaki, H., Takamiya, S., and Susaki, W., (1979), "Single-Mode Junction-Up TJS Lasers with estimated Lifetime of 10^6 Hours", *IEEE J. Quantum Elec.*, **15**, pp. 1208-1209.
7. Yang, Y.J., Lo, Y.C., Lee, G.S., Hsieh, K.Y., and Kolbas, R.M., (1986), "Transverse Junction Stripe Laser with a Lateral Heterobarrier by Diffusion Enhanced Alloy Disordering", *Appl. Phys. Lett.*, **49**, pp. 835-837.
8. Isshiki, K., Kaneno, N., Kumabe, H., Namizaki, H., Ikeda, K., and Susaki, W., (1986), "Ten-Thousand-Hour Operation of Crank Transverse-Junction-Stripe Lasers Grown by Metal-Organic Chemical Vapor Deposition", *J. Lightwave Tech.*, **4**, pp. 1475-1481.
9. Suzuki, Y., Mukai, S., Yajima, H., and Sato, T., (1987), Transverse Junction Buried Heterostructure (TJ-BH) AlGaAs Diode Laser", *Electronics Lett.*, **23**, pp. 384-386.
10. Ohta, J., Kuroda, K., Mitsunaga, K., Kyuma, K., Hamanaka, K., and Nakayama, T., (1987), "Buried Transverse-Junction Stripe Laser for Optoelectronic Integrated Circuits", *J. Appl. Phys.*, **61**, pp. 4933-4935.

11. DeFreez, R.K., Puretz, J., Orloff, J., Elliot, R.A., Namba, H., Omura, E., and Namizaki, H., (1988), "Operating Characteristics and Elevated Temperature Lifetests of Focused Ion Beam Micromachined Transverse Junction Stripe Lasers", *Appl. Phys. Lett.*, **53**, pp. 1153-1155.
12. Shimoyama, K., Katoh, M., Noguchi, M., Inoue, Y., Gotoh, H., Suzuki, Y., and Satoh, T., (1988), "Transverse Junction Buried Heterostructure (TJ-BH) Laser Diode Grown by MOCVD", *J. Crystal Growth*, **93**, pp. 803-808.
13. Shimoyama, K., Katoh, M., Suzuki, Y., Satoh, T., Inoue, Y., Nagao, S., and Gotoh, H., (1988), "CW Operation and Extremely Low Capacitance of TJ-BH MQW Laser Diodes Fabricated by Entire MOVPE", *Jpn. J. Appl. Phys.*, **27**, pp. L2417-2419.
14. Furuya, A., Makiuchi, M., Wada, O., and Fujii, T., (1988), "AlGaAs/GaAs Lateral Current Injection Multiquantum Well (LCI-MQW) Laser Using Impurity-Induced Disordering", *IEEE J. Quantum Elec.*, **24**, pp. 2448-2453.
15. Ahn, D. and Chuang, S.L., (1988), "A Field-Effect Quantum-Well Laser with Lateral Current Injection", *J. Appl. Phys.*, **64**, pp. 440-442.
16. Yasuhira, N., Suemune, I., Kan, Y., and Yamanishi, M., (1990), "Selectively Doped Double-Heterostructure Lateral Current Injection Ridge Waveguide AlGaAs/GaAs Laser", *Appl. Phys. Lett.*, **56**, pp. 1391-1393.
17. Honda, Y., Suemune, I., Yasuhira, N., and Yamanishi, M., (1990), "A New Optoelectronic Device Based on Modulation-Doped Heterostructure: Demonstration of Functions as Both Lateral Current Injection Laser and Junction Field Effect Transistor", *IEEE Photonics Tech. Lett.*, **2**, pp. 881-883.
18. Sin, Y., Hsieh, K.Y., Lee, J.H., and Kolbas, R.M., (1991), "Surface and Bulk Leakage Currents in Transverse Junction Stripe Lasers", *J. Appl. Phys.*, **69**, pp. 1081-1090.
19. Honda, Y., Suemune, I., Yasuhira, N., and Yamanishi, M., (1991), "Continuous-Wave Operation of a Lateral Current Injection Ridge Waveguide AlGaAs/GaAs Laser with a Selectively-Doped Heterostructure", *Jpn. J. Appl. Phys.*, **30**, pp. 990-991.
20. Zou, W.X., Law, K.K., Merz, J.L., Fu, R.J., and Hong, C.S., (1991), "Laterally Injected Low-Threshold Lasers by Impurity-Induced Disordering", *Appl. Phys. Lett.*, **59**, pp. 3375-3377.
21. Schaus, C.F., Torres, A.J., Cheng, J., Sun, S., Hains, C., Malloy, K.J., Schaus, H.E., Armour, E.A., and Zheng, K., (1991), "Transverse Junction Vertical-Cavity Surface-Emitting Laser", *Appl. Phys. Lett.*, **58**, pp. 1736-1738.
22. Beyler, C.A., Hummel, S.G., Chen, Q., Osinski, J.S., and Dapkus, P.D., (1991), "Low Threshold Current Lateral Injection Lasers on Semi-Insulating Substrates Fabricated Using Si Impurity-Induced Disordering", *Electronics Lett.*, **27**.
23. Hihara, M., Hirata, T., Suehiro, M., Maeda, M., and Hosomatsu, H., (1991), "Fabrication of GaAs/AlGaAs Lateral Current Injection Quantum Well Laser", *Extended Abstracts of the 1991 Int. Conf. on Solid State Dev. and Mat., Yokohama*, pp. 735-736.
24. Evaldsson, P.A., Taylor, G.W., Cooke, P., Burrus, C.A., and Tell, B., (1992), "Small Signal and Continuous Wave Operation of the Lateral Current Injection Heterostructure Field-Effect Laser", *Appl. Phys. Lett.*, **60**, pp. 1697-1699.
25. Kawamura, Y., Noguchi, Y., and Iwamura, H., (1993), "Lateral Current Injection InGaAs/InAlAs MQW Lasers Grown by GSMBE/LPE Hybrid Method", *Electronics Lett.*, **29**, pp. 102-104.
26. Oe, K., Noguchi, Y., and Caneau, C., (1994), "GaInAsP Lateral Current Injection Lasers on Semi-Insulating Substrates", *IEEE Photonics Tech. Lett.*, **6**, pp. 479-481.
27. Tan, G.L, Lee, K., and Xu, J.M., (1993), "Finite Element Light Emitter Simulator (FELES): A New 2D Software Design Tool for Laser Devices", *Jpn. J. Appl. Phys., part 1*, **32**, pp. 583-589.
28. Laidig, W.D., Holonyak, N., Camras, M.D., Hess, K., Coleman, J.J., Dapkus, P.D., and Bardeen, J., (1981), "Disorder on an AlAs-GaAs Superlattice by Impurity Diffusion", *Appl. Phys. Lett.*, **38**, pp. 776-778.
29. Tan, G.L., Xu, J.M., and Shur, M., (1993), "GaAs/AlGaAs Double-Heterostructure Lateral P-I-N Ridge Waveguide Laser", *Optical Engineering*, **32**, pp. 2042-2045.
30. Suda, D.A., Lu, H., Makino, T., and Xu, J.M., (1995), "An Investigation of Lateral Current Injection Laser Internal Operation Mechanisms", to be published in *IEEE Photonics Tech. Lett.*.
31. Pavlidis, D., Sweeny, M., Anis, H., and Xu, J.M., (1995), "Bragg Reflectors for Mode Control in Directions Orthogonal to the Bragg's Periodicity", to be presented at the *1995 Canadian Semiconductor Conference*.
32. Sun, C.C. and Xu, J.M., (1989), "Observation of Capacitive Modulation of Bipolar Current in Poly-Si Gated Lateral PIN Structures", *Appl. Phys. Lett.*, **54**, pp. 1875-1877.

WIDE BAND GAP SEMICONDUCTORS. GOOD RESULTS AND GREAT EXPECTATIONS.

M. S. SHUR
Department of Electrical Engineering
University of Virginia, Charlottesville, VA 22903-2442, USA

Abstract.

We will review properties of wide band gap semiconductors, which make them superior materials for many electronic and optoelectronic applications. These semiconductors should allow us to achieve a very high on-to-off ratio in transistors, implement nonvolatile solid state memories, and develop new optoelectronic and optical devices for visible and ultraviolet ranges as well as electronic and optoelectronic systems operating in a harsh environment and/or at elevated temperatures. Technological difficulties, relatively low mobility values, and problems related to contacts and traps make the realization of this great potential a challenge. We show that many of these difficulties can be alleviated in AlGaN/GaN Heterostructure Field Effect Transistors (HFETs), which use superior transport properties of the two dimensional electron gas in wide band gap semiconductors. AlGaN/GaN HFET's, which have been fabricated on a transparent sapphire substrate, are very sensitive to ultraviolet light. Hence, they can be also used as solar blind ultraviolet photodetectors.

1. Introduction.

Most applications of semiconductor materials use amazing their amazing ability to change the electric conductivity in an extremely wide range. Metals are good conductors of electricity, dielectrics have a very high resistance for an electric current, but a semiconductor sample may change its conductivity by 10 orders of magnitude or more. In a Field Effect Transistor (FET), the channel conductivity is changed by the applied gate bias from a high (above threshold) value to a very low (below threshold) value. This on-to-off ratio is one of the most important FET characteristics. A high on-to-off ratio allows us to combine low power consumption in the off-state with high switching speed, since, in the on-state, a large current is available for charging device and circuit capacitances in a transient switching process. In the on-state, the current is determined by the device geometry, the gate voltage swing, and by the semiconductor transport properties. In the off-state, the current is controlled by the barrier separating the FET source and drain (see Fig. 1). The height of this barrier is on the order of one half of the energy gap, E_g, and, hence, the on-to-off ratio is proportional to $\exp(-E_g/2k_BT)$ where k_B is the Boltzmann constant and T is temperature. Therefore, the wide band gap semiconductors should have a much higher on-to-off ratio than silicon or even GaAs (see Fig. 2).

S. Luryi et al. (eds.), Future Trends in Microelectronics, 279–290.

Since amorphous silicon has a relatively wide energy gap (the energy gaps are 1.12 eV for silicon, 1.42 eV for GaAs, and 1.71 eV for amorphous silicon), amorphous silicon thin film transistors have the largest on-to-off ratio, in spite of a very low on-current. Wide band gap semiconductors (such as GaN with the energy gap of 3.4 eV) should have very low leakage currents [since the intrinsic carrier concentration, n_i, is proportional to $\exp(-E_g/2k_BT)$]. They have extremely large breakdown voltages (since in order to cause an impact ionization event, an electron or a hole has to obtain an energy in excess of E_g from the electric field). These materials are thermally and chemically stable and uniquely suited for applications in electronic circuits and systems operating in a harsh environment and/or at high temperatures.

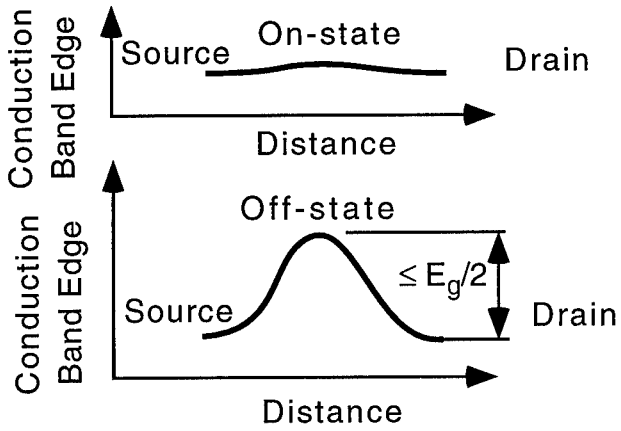


Fig. 1. Band diagram for a Field Effect Transistor.

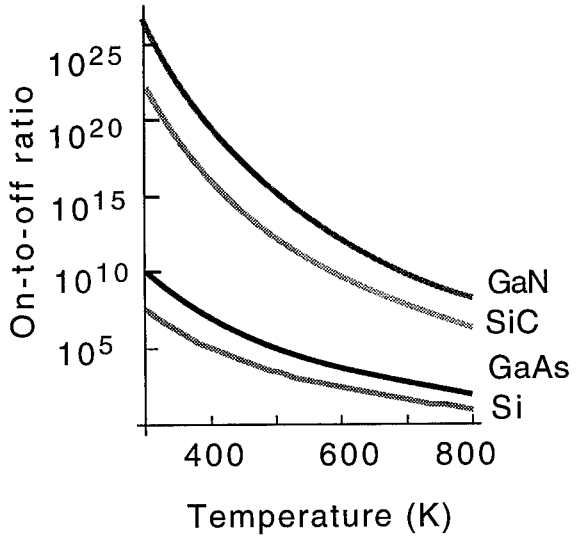


Fig. 2. On-to-off ratio versus temperature for FETs from different semiconductors (calculated assuming that the off-current is proportional to the intrinsic carrier concentration.

The elementary theory of p-n junctions yields the following expression for the reverse saturation current density, j_R, for a p^+n junction (see, for example, [1]):

$$j_R = q\sqrt{\frac{D_p}{\tau_p}}\frac{n_i^2}{N_d} + \frac{qn_iW}{\tau_e} \tag{1}$$

where q is the electronic charge, n_i is the intrinsic carrier density, N_d is the doping density in the n-type region, D_p is the hole diffusion coefficient, τ_e is the effective lifetime, and

$$W = \sqrt{\frac{2\varepsilon_s(V_{bi}-V)}{qN_d}} \tag{2}$$

is the thickness of the depletion region. Here V_{bi} is the built-in voltage and ε_s is the dielectric permittivity of the semiconductor.

In wide band gap materials, the second term in the right-hand part of eq. (1) is dominant. In cubic SiC at room temperature, the theoretical value of n_i is as low as 10^{-6} cm^{-3}! For an estimate, let us assume certain values for m_n and m_p, independent of the energy gap, since this assumption will not change the order of magnitude of the resulting concentration. Let us take $m_n = 0.3m_e$ and $m_p = 0.6m_e$, to be specific. Then we find

$$n_i\left(\mathrm{m}^{-3}\right) = 1.34\times10^{21}\times T^{3/2}\exp\left(-\frac{E_g}{2k_BT}\right)(\mathrm{K}) \tag{3}$$

(see Fig. 3).

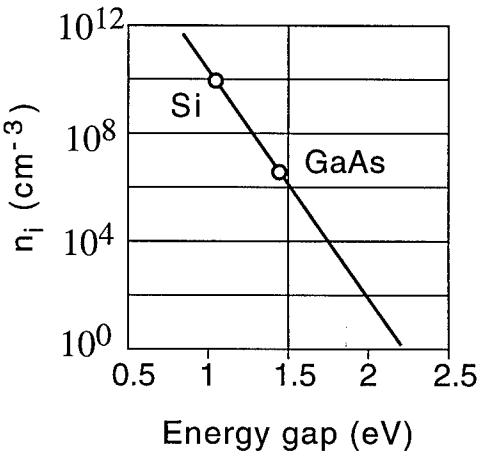


Fig. 3. Intrinsic carrier concentration versus energy gap at room temperature. (The values of n_s for Si and GaAs are approximate since we used the same values of m_n and m_p in the calculation of this dependence.) [1]

Choosing the device volume to be $1 \times 0.1 \times 10\ \mu m^3$ and assuming a generation time of 1 ns, we obtain

$$I_{leakage}(\mathrm{A}) = 2.14 \times 10^{-7} \times T^{3/2}(\mathrm{K}) \times \exp\left(-\frac{E_g}{2k_B T}\right) \tag{4}$$

The resulting dependence of the minimum leakage current on the energy gap for room temperature is shown in Fig. 4. However, experimentally, such low values of the saturation current density have not yet been observed, and the conventional theory of a *p-n* junction is not even valid when n_i is very small. Nevertheless, these estimates clearly illustrate the potential of wide band gap semiconductors for applications in nonvolatile memories and integrated circuits operating both at room temperature and at elevated temperatures.

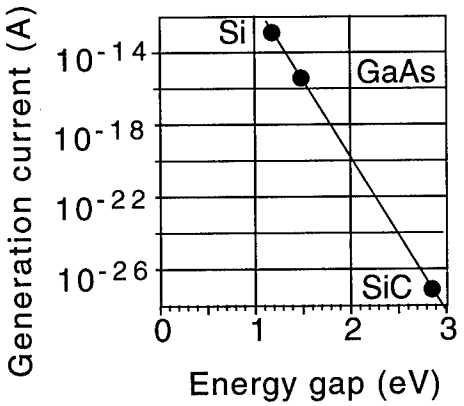


Fig. 4. Generation current (per 1 μm^3 of the depletion region volume) versus energy gap (at room temperature). [1]

Many applications utilize the optoelectronic properties of semiconductors. Usually, the peak of a semiconductor sensitivity to electromagnetic radiation corresponds to the valence-to-conduction band transitions. Using the impurity level-band transitions or subband transitions in superlattices or quantum wells, it is also possible to obtain response at photon energies smaller than the energy gap.

Fig. 5 compares the energy gaps of semiconductors with the spectral sensitivity of human eye. The energy gaps of Si, GaAs, and InP correspond to the infrared range. (Si is indirect gap semiconductor, which has an inferior optical response, and GaAs and InP have direct energy gaps.). The wide band gap semiconductors, such as SiC or GaN, can cover the visible and even the ultraviolet range. Therefore, these materials can be used for displays, photodetectors, and imagers in the visible and ultraviolet ranges.

Photosensitive semiconductors, such as amorphous silicon or selenium have been used for optically recorded information storage and retrieval. In this application, the minimum area of a semiconductor surface required to record one bit is determined by the wavelength of the electromagnetic radiation. Fig. 6 shows the number of bits per cm^2 as a function of the photon energy and the corresponding wavelength (assuming the minimum pixel area of $10\ \lambda^2$). As seen, wide band gap semiconductors, such as

GaN, which demonstrated an excellent photoresponse, should allow us to achieve an extremely dense information storage using an ultraviolet light.

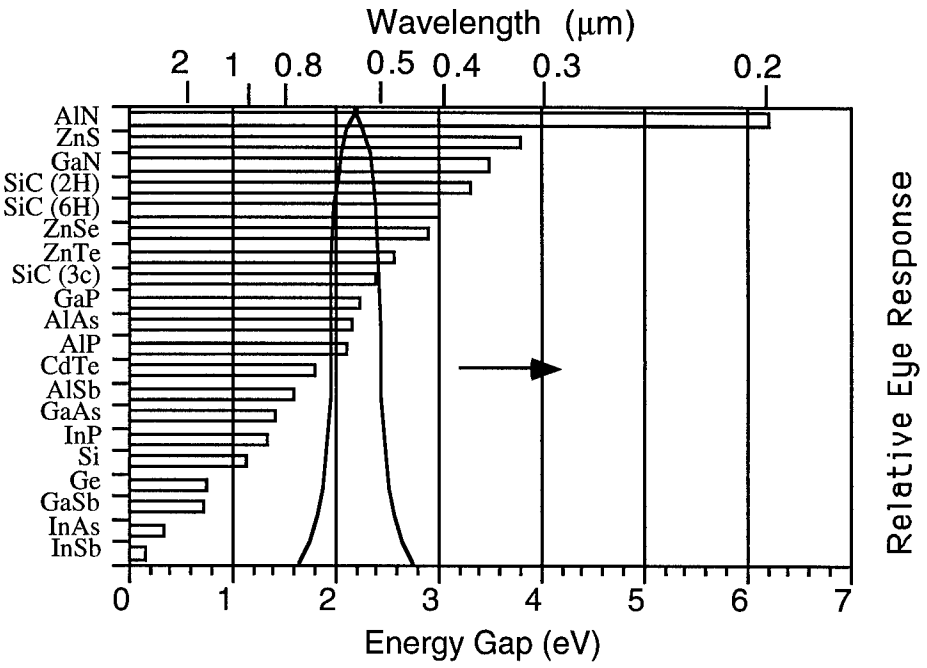


Fig. 5. Energy gaps of semiconductor materials compared with spectral sensitivity of human eye. [1]

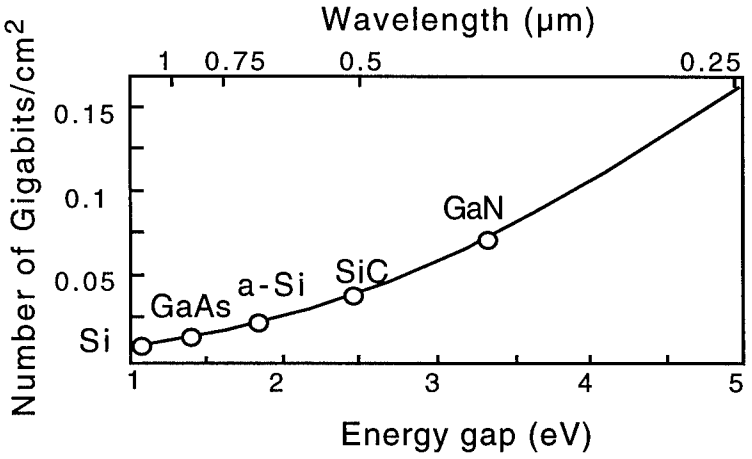


Fig. 6. The number of bits per cm^2 as a function of the photon energy and the corresponding wavelength, λ, (assuming the minimum pixel area of $10\lambda^2$).

In order to take advantage of the properties of wide band gap semiconductors, one has to develop the material, device, and integrated circuit technologies. So far, this ambitious goal has been fully achieved only for one semiconducting material - silicon - and partially achieved for gallium arsenide. Since Si has a huge cost and integration scale advantage over GaAs, the jury is still out on whether GaAs will be able to win a fairly large market share in high speed and low power applications or will remain an important niche technology. Since GaAs has been studied intensively since the nineteen fifties, we can expect that the realization of the full potential of wide band gap semiconductors will take at least a few years if not a few decades.

At least four families of wide band gap semiconductors have attracted interest of researchers worldwide - diamond, II-VI semiconductors such as ZnS or ZnSe, SiC and related materials, and GaN and related materials.

Silicon carbide is one of the first semiconductor materials to be discovered. As early as 1907, Round observed electroluminescence in silicon carbide, which he reported in his article published in *Electrical World*, Vol. 19, p. 309, 1907. In 1955, Lely developed a new technique of SiC growth, and the studies of silicon carbide started to develop on a more sound basis. However, only recently, have practical applications of silicon carbide devices become a reality with the development of modern epitaxial techniques, primarily by groups in Russia (LETI and A. F. Ioffe Institutes in Sankt Petersburg), US (NCSU and CREE research), and in Germany (Erlangen).

Silicon carbide exists in more than 170 different polytypes. The properties of the polytypes are so different that, in fact, SiC may be more accurately considered as a group of closely related materials. Depending on the polytype crystal structure, the energy gap of silicon carbide varies from 2.2 to 3.3 eV. It has a predicted electron saturation drift velocity, v_S, of 2×10^7 cm/s (approximately two times larger than in silicon), a breakdown field larger than 2,500 to 5,000 kV/cm (compared to 300 kV/cm for silicon), and a high thermal conductivity of 3.5 W/cm°C (compared to 1.3 W/cm°C for silicon and 0.5 W/cm°C for GaAs). These properties make SiC important for potential applications in high-power, high-frequency devices as well as in devices operating at high temperatures and/or in a harsh environment. Applications of SiC include high-power devices, microwave devices (both avalanche diodes and microwave field effect transistors), and optoelectronic devices such as light-emitting diodes covering the visible electromagnetic spectrum and even the ultraviolet range.

Recently, new solid-state solutions of AlN/SiC/InN/GaN have been demonstrated. This exciting development opens up the possibility of a new generation of heterostructure devices based on SiC. Solid-state solutions of AlN-SiC are also expected to lead to direct gap ternary materials for UV and deep blue optoelectronics, including the development of visible light-emitting diodes and lasers.

All basic device elements – from ohmic contacts to Schottky diodes and *p-n* junctions – and most of the semiconductor devices – from field effect transistors to bipolar junction transistors, thyristors, and light-emitting Diodes – have been demonstrated in materials ranging from different polytypes of SiC [2] to the GaN/AlN material system. [3]

In this lecture, we will limit ourselves to GaN, since this material and related semiconductor can excel in both electronic and optoelectronic applications.

2. Transport properties of GaN.

The Monte Carlo calculations show that the peak field in GaN is very high (on the order of 100 kV/cm compared to approximately 3.5 kV/cm in GaAs); see Fig. 7. The reason for such a shape of the velocity-field characteristic is a large intervalley separation and a very large energy of polar optical phonons in GaN (nearly three times higher than in GaAs). In high electric fields, the electron velocity in GaN is only weakly dependent on temperature (see Fig. 7 b).

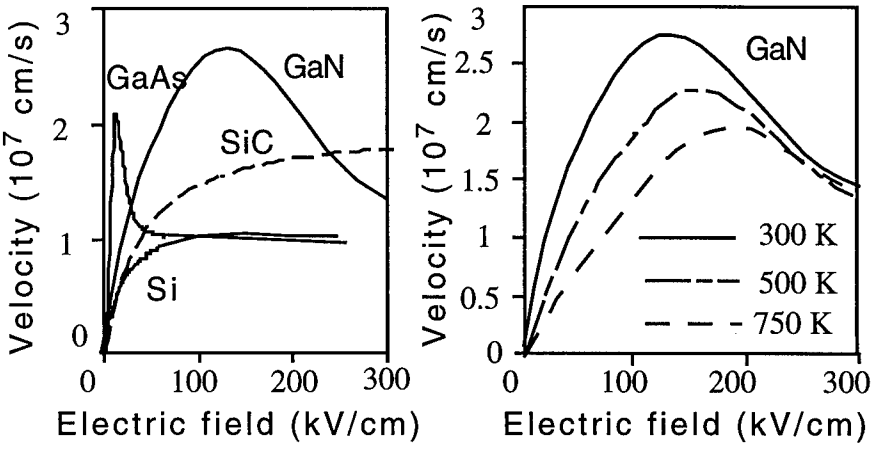


Fig. 7. (a) Electron drift velocity at 300 K in GaN, SiC, and GaAs. (b) Electron drift velocity in GaN at 300 K, 500 K, and 750 K. [4]

Compared to GaAs, GaN has a relatively low mobility. The most important scattering mechanisms in GaN are polar optical scattering, ionized impurity scattering, and piezoelectric scattering. [5] In principle, the ionized impurity scattering may be nearly totally suppressed in the two dimensional electron gas (2DEG) in a modulation doped heterostructure, since the donors are located in a barrier layer, away from the 2DEG. In practice, this is very difficult to achieve in an AlGaN/GaN modulation doped structure. However, the calculation of the electron mobility limited by combined optical polar, piezoelectric, and acoustic scattering gives an upper bound for the 2DEG mobility. The results of such a calculation of the electron drift mobility are shown in Fig. 8. [5] (The Hall mobility is a higher than the drift mobility.)

Fig. 9 shows the dependencies of the electron drift mobility in GaN on the electron carrier concentration for different compensation levels at 300 and 77 K. As can be seen from the figure the electron mobility in 2DEG can be substantially enhanced due to the screening of the piezoelectric scattering. This enhancement is a feature specific for GaN, which is a very strong piezoelectric. We notice large calculated values of the drift mobility for uncompensated material.

In our recent paper [6], we reported on the measurements of the Hall mobility in bulk GaN and in the 2DEG at the GaN/AlGaN heterointerface. These data clearly demonstrate the difference between the electron transport in 2DEG and bulk GaN. Fig.

10 shows the temperature dependence of the electron Hall mobility in GaN for the 2DEG and bulk GaN. Solid lines and open dots show the calculated dependence and our experimental data, respectively. For the bulk calculation, we used $n = 2\text{x}10^{16}$ cm^{-3}, $N_T = 2\text{x}10^{17}$ cm^{-3} and for the 2DEG calculation, $n = 5\text{x}10^{17}$ cm^{-3}, $N_T = 6.5\text{x}10^{16}$ cm^{-3}.

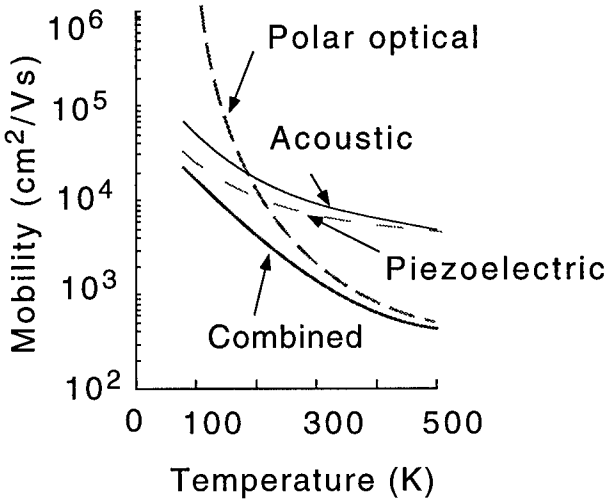


Fig. 8. Mobility limited by polar optical, piezoelectric, and acoustic scattering versus temperature· [5]

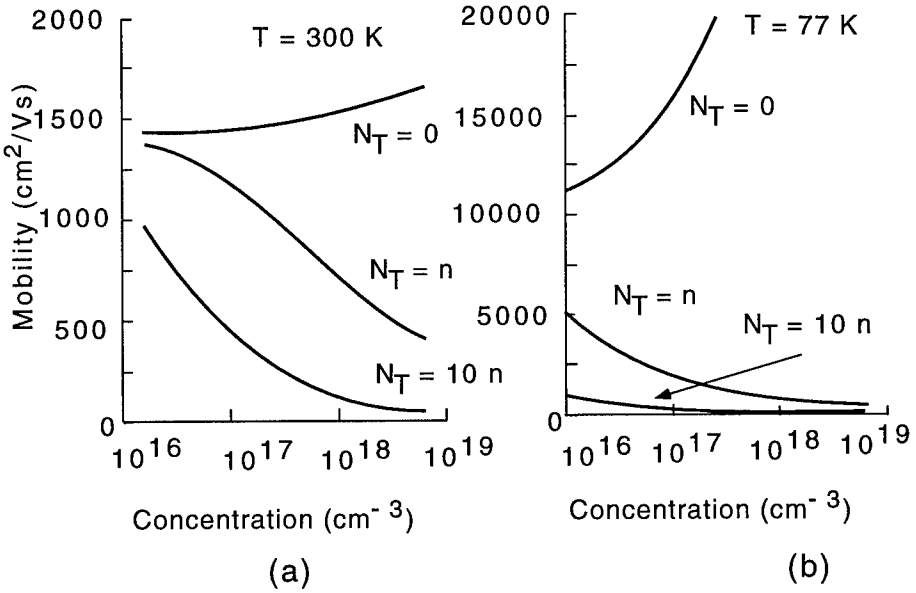


Fig. 9. Electron drift mobility in GaN versus electron carrier concentration for different compensation levels at 300 K (a) and 77 K (b). [5]

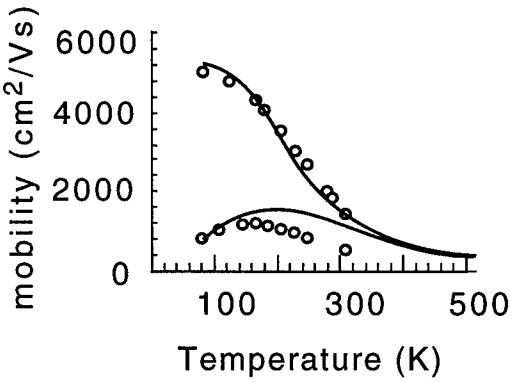


Fig. 10. Measured and calculated Hall electron mobility in bulk GaN and in the 2DEG. [7]

The results of the calculation are in good agreement with our experimental data for the 2DEG. For the bulk GaN, the theory overestimates the electron mobility at elevated temperatures where the polar optical polar scattering should be the dominant scattering mechanism. As was pointed by Professor Mishra [7], the most likely reason for this disagreement is the nonuniform distribution of the electron concentration and compensation ratio in the GaN film, with a strongly compensated layer near the channel/buffer interface. However, the qualitative agreement with experimental data is good enough to illustrate the mobility enhancement in the 2DEG

3. AlGaN/GaN Heterostructure Field Effect Transistors.

A FET is probably the most important electronic device. All in all, the GaN electron velocity at elevated temperatures is large enough for achieving high FET transconductance, especially in short channel devices. Since the electron velocity saturation in an FET channel plays a larger role in higher electric fields, the scaling of the gate length in GaN FETs should lead to larger relative improvements in device characteristics for GaAs FETs.

The device structure of AlGaN/GaN Heterostructure FETs (HFETs) is shown in Fig. 11. These devices have operated at temperatures up to 300 °C (see Fig. 12). Recently reported experimental data on the microwave operation of GaN/AlGaN HFETs indicate a fairly high maximum frequency of oscillations and cutoff frequency (f_{max} > 70 GHz, f_T > 20 GHz at room temperature) for these devices. [8] However, their performance still falls far short of the theoretically predicted performance. It is limited by a small intrinsic gate-voltage swing and can be dramatically improved by reducing the source series resistance and by optimizing the device design. Our analysis [4] shows that transconductances over a hundred mS/mm should be achievable in submicron AlN/GaN HFETs with a relatively small drop at elevated temperatures (see Fig. 12, 13, and 14). These results show that, in spite of a smaller thermal conductivity than for SiC, the excellent transport properties of GaN at elevated temperatures make these devices a viable alternative for high temperature microwave and digital applications.

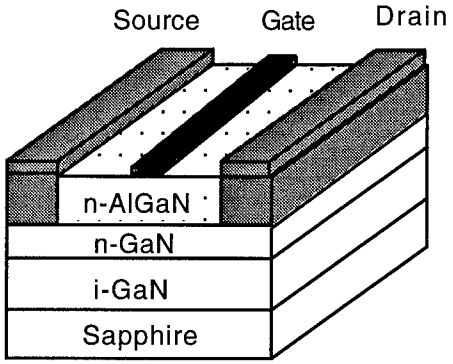


Fig. 11. Device structure of AlGaN/GaN HFETs. [5]

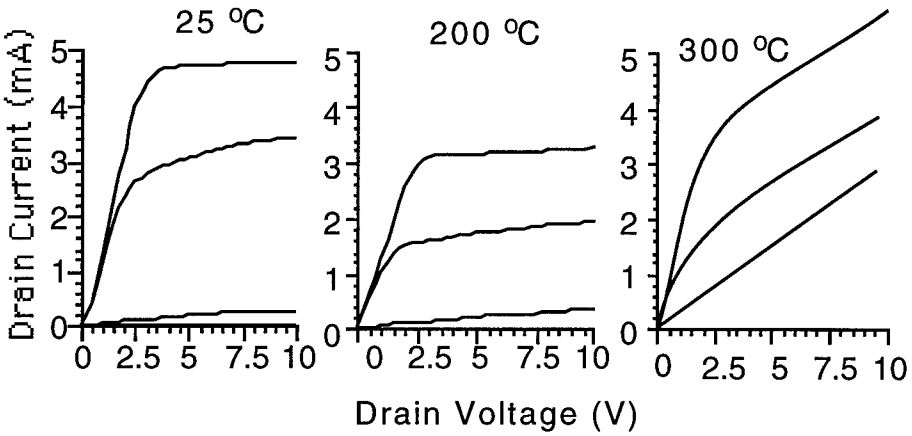


Fig. 12. I-V characteristics of AlGaN/GaN HFETs at different temperatures. [8]

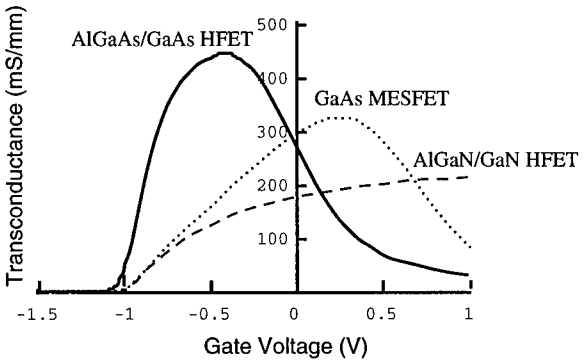


Fig. 13. Computed device transconductance in the saturation region versus gate bias for three devices: 0.5 μm gate GaAs MESFET, 0.5 μm gate AlGaAs/GaAs HFET, and 0.25 μm AlGaN/GaN HFET. [4]

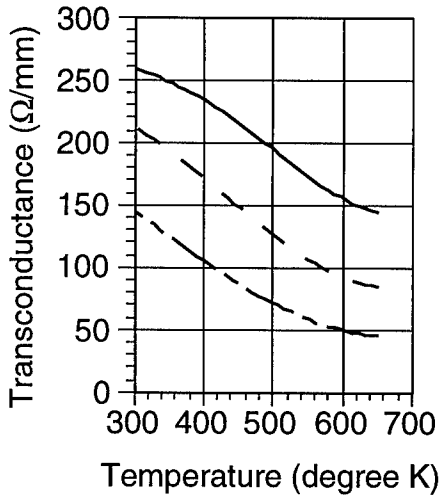


Fig. 14. Computed transconductance for 0.5 μm, 1 μm, and 2 μm AlGaN/GaN HFETs.

4. Optoelectronic AlGaN/GaN HFETs.

Unique optical and electronic properties of the GaN/AlGaN material system open up numerous opportunities for visible-blind optoelectronic devices. These devices have a high sensitivity and a large gain-bandwidth product and can be integrated with GaN/AlGaN field effect transistors which have already demonstrated an operation at microwave frequencies. A transparent sapphire substrate makes AlGaN/GaN HFETs well suited for optoelectronic applications.

The HFET photodetector is based on a 0.2 micron gate AlGaN/GaN HFET [9] and utilizes a shift in the threshold voltage caused by the light generated carriers (see Fig. 15). These results show that unique optical and electronic properties of GaN/AlGaN material system open up numerous opportunities for visible-blind optoelectronic devices. These devices could have a high sensitivity and a large gain-bandwidth product and can be integrated with GaN/AlGaN field effect transistors for applications in optoelectronic integrated circuits.

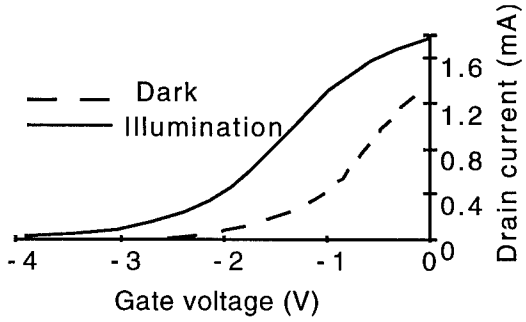


Fig. 15. I-V characteristics of AlGaN/GaN HFETs in the dark and under light. [9]

5. Conclusions.

Wide band gap semiconductor devices in general and GaN-based devices, in particular, have already demonstrated an impressive performance. Their excellent transport and optoelectronic properties should allow us to achieve excellent performance in harsh conditions and/or at elevated temperatures. Large band discontinuities in these materials should allow us to obtain very high densities of 2DEG (up to $2x10^{13}$ cm^{-2} for GaN). 2DEG should have much better transport properties than bulk electrons. This has already been confirmed by a good performance of AlGaN/GaN HFETs.

6. Acknowledgment.

The work at the University of Virginia has been partially supported by the ONR (Project Monitor Max Yoder).

7. References.

1. Shur, M. (1995) *Introduction to Electronic Devices*, John Wiley and Sons, New York
2. Kelner, G. and Shur M. (1995) SiC Devices, in "*Properties of Silicon Carbide*", G. Harris, Editor, M. Faraday House, IEE, England
3 Strite, S. and Morkoç, H. (1992) *J. Vac. Sci. Technol.* **B10** 1237
4 Shur, M. and Khan, M. A. Electronic and Optoelectronic AlGaN/GaN Heterostructure Field Effect Transistors , unpublished
5 Shur, M., Gelmont, B., and Khan, M. A., High Electron Mobility in Two Dimensional Electrons Gas in AlGaN/GaN Heterostructures, unpublished
6 Khan, M. A., Chen, Q, Sun, C. J., Shur, M. S., and B. Gelmont, B, 2D-Electron Gas in GaN-AlGaN Heterostructures Deposited Using TMAA as the Aluminum Source in Low Pressure Metalorganic Chemical Vapor Deposition, unpublished
7 Mishra, U. (1995) Private communication, June 21
8 Khan, M. A., Shur, M. S., Kuznia, J. N, Burm J., and Schaff W. (1995) Temperature activated conductance in GaN/AlGaN heterostructure field effect transistors operating at temperatures up to 300 ^{o}C, *Applied Physics Letters*, **66**, 1083
9. Khan, M. A., Shur, M. S., Chen, Q., Kuznia, J. N., and Sun C. J. (1995) *Electronics Letters*, **31**, 398

GaN and Related Compounds for Wide Bandgap Applications

Dimitris Pavlidis

Department of Electrical Engineering and Computer Science
The University of Michigan, 1301 Beal Ave.,
Ann Arbor, MI 48109-2122. USA

1. Introduction

GaN and related compounds are wide bandgap semiconductor materials with great potential for optoelectronic applications from blue to ultraviolet wavelengths, and high-power, high-temperature devices. GaN can be crystallized in either hexagonal (wurtzite) or cubic (zincblende) structure depending on the substrate symmetry and growth conditions. In certain cases both structures may co-exist because of the small difference in energy of formation. High-quality wurtzitic GaN has been grown successfully on a variety of substrates, in particular on the basal plane of sapphires. However, cubic structures possess in principle superior electronic properties i.e. doping efficiency and high-speed transport and allow easy cleaving, as necessary for devices such as lasers [1],[2]. Although the traditional substrate used for nitride material growth is sapphire, it is consequently desirable to explore the possibility of using substrates such as silicon or GaAs. In addition to allowing cubic material growth, this could lead to reduction of interfacial defects and impurities as well as, integration of GaN with Si or GaAs-based devices.

This paper reviews the characteristics of nitride materials as obtained by various growth techniques [3],[4],[5]. Their structural, optical and electrical properties are described. The use of specially grown buffers appears to play a major role in the quality of the layers [6],[7],[8]. Growth of AlN, InN, GaN and their ternaries offers the possibility of bandgap engineering but imposes material challenges. Understanding and controlling of the background and intentional doping characteristics is crucial due to the relatively high background levels and difficulties in efficient doping of these materials. Possibilities of obtaining cubic material are discussed, as demonstrated by MOCVD growth on GaAs [9]. Various optical i.e. LEDs [10]-[14]and

S. Luryi et al. (eds.), Future Trends in Microelectronics, 291–302.

electronic i.e. FETs, HEMTs [15]-[17] devices are also examined. Experimental demonstration of blue LEDs confirms the high potential of nitrides and related compounds for optoelectronic applications. Recent work confirms the feasibility of realizing GaN MESFETs and GaN/AlN HEMTs with promising high frequency characteristics. Currently obtained characteristics and expected performance of nitride materials and devices built with them are finally presented

2. Substrates and Growth Techniques

The key binaries of nitrides are GaN, AlN and InN. Their bandgap energy varies from 1.9eV for InN, to 6.2eV for AlN as shown in the diagram of Fig. 1 where the bandgap energy is plotted for various materials as a function of the lattice constant. Shown in the same figure are other commonly used III-Vs such as GaAs and InP. As one can see the lattice constant of nitrides is much smaller than that of III-Vs or Silicon and none of the traditionally used semiconductor substrates is compatible with them from the point of view of lattice matching. The major difficulty encountered in the development of high quality nitride films lies in fact on the lack of suitable substrates which leads to the need for heteroepitaxial solutions.

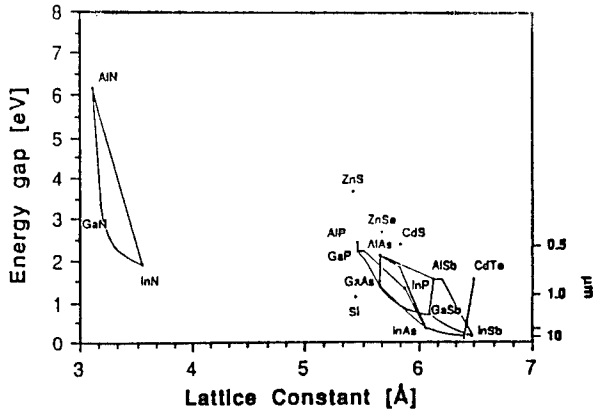


Figure 1. Bandgap energy vs. Lattice constants of nitrides and other commonly used semiconductor materials

The most popular substrate used for nitride growth is sapphire (a=4.758Å, c=12.99Å). It is stable up to the very high temperatures (950°C to 1050°C) normally used for nitride growth but has a rather low thermal conductivity of

0.5W/cmK compared with other substrates such as 3C-SiC or 6H-SiC (4.9W/cmK) and Si (1.5W/cmK). By way of comparison GaAs has similar thermal conductivity as sapphire and the corresponding values for GaN, AlN are 1.3W/cmK, 2.0W/cmK. These differences make sapphire not an optimum choice for applications where heat needs to be dissipated very efficiently, as for example high-power/high -temperature electronics. Differences also exist between the values of coefficients of thermal expansion of the nitrides and the substrates on which they are deposited, leading to difficulties upon cooling down at the end of their growth.

Nitrides grown on sapphire have hexagonal crystal structure. Sapphire (0001) and $(01\bar{1}2)$ substrates have been reported to lead to (0001) and $(2\bar{1}\bar{1}0)$ GaN growth. GaN films grown on GaAs or Si can have either hexagonal or cubic features depending on the initial substrate. For example, GaAs (100) and (111) results in (001) and (0001) GaN respectively. Similar trends are observed for growth on Silicon. The availability of cubic GaN offers a tremendous advantage in terms of the possibility of cleaving which is not feasible for hexagonal material such as nitrides grown on sapphire. Moreover, cubic structures possess in principle superior electronic properties for device applications [1],[2] such as higher electron velocity and more efficient doping. In addition to GaAs [9] and Si, cubic SiC and Mg have also been employed for demonstrating cubic GaN.

The lack of suitable substrates for nitride growth prompted research on the development of buffer technologies. GaN or AlN can be used for this purpose. A study of the role of AlN buffer [6] showed that following the low temperature growth of a 500Å thick AlN layer, GaN nucleation sites are generated with the same orientation as the substrate. This promotes the lateral growth of GaN due to the decrease of interfacial free energy between the substrate and the epitaxial GaN film, leading eventually to good quality bulk GaN. The optimum buffer thickness has been reported to be small (~200Å) in case of GaN buffers. The use of a buffer layer helps in improving the crystal quality as demonstrated by the reduction of Full-Width-Half-Maximum (FWHM) values of X-Ray Diffraction (XRD) spectra of GaN grown with and without buffer layers [7]. Moreover the introduction of a GaN or AlN buffer leads to reduction of the residual carrier concentration and improvement of the mobility of bulk GaN grown on top [8].

Various growth techniques have been employed for the growth of nitrides. These include Molecular Beam Epitaxy (MBE), Metalorganic Chemical Vapor Deposition (MOCVD) and Hydride Vapor Phase Epitaxy (HVPE). Electron Cyclotron Resonance (ECR) used in conjunction with MBE allows activation of molecular nitrogen N_2^+ and low temperature growth under ultra high vacuum conditions leading to reduced autodoping [3]. MOCVD growth is in most cases carried out at low pressure of ~0.1atm using NH_3 gas and TMGa. Growth temperatures for nitrides on sapphire substrates are in the order of 1050°C, while lower temperatures (~600°C) are sufficient for growth on GaAs. A common feature of all nitride films is their relatively high background concentration which has for long being associated with nitrogen vacancies (V_N). As the growth temperature increases the background concentration is normally found to decrease but V_N increases [4]. Thus, V_N alone cannot account for the observed high background electron concentration and other effects such as impurity incorporation i.e. O or Si may be present. HVPE has been reported to permit growth of "bulk-like" GaN substrates opening therefore the possibility of homoepitaxial GaN growth [5]. A sputtered ZnO buffer layer has been employed in this case between the sapphire substrate and the GaN and was removed at the end of growth leading to thick (~600μm) GaN films. Growth took place at high growth rates (80-130μm/hr) and high crystalline quality GaN was grown on top of these substrates.

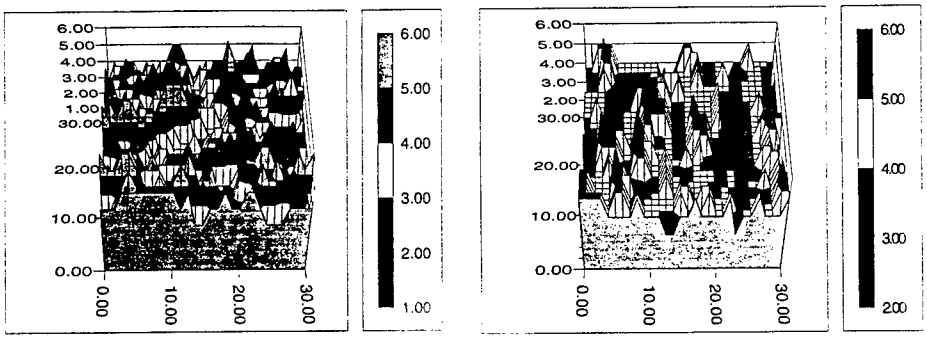


Figure 2. 3D growth front contours of GaN grown at 400°C and 500°C (V/III= 10) as simulated by Monte Carlo. The legend on the right hand provides the number of surface monolayers.

Most nitride developments have up to now been based on empirical approaches. To improve material quality it is, however, essential to gain deeper insight into the growth mechanisms. An atomistic model consistent with a

variety of experimental observations was developed for this purpose [18]. The model is based on MBE-like conditions of GaN growth and employs a Monte Carlo approach to study the impact of substrate temperature, Ga flux and V/III ratio on growth rate and growth front quality. The growth rate was found to increase with the V/III ratio reaching a saturation value which is determined by the Ga flux. The quality of the growth front improves by using a smaller Ga flux for a fixed temperature and V/III ratio or by reducing the V/III ratio at a given temperature. A consideration of the growth kinetics suggests that GaN grown surfaces are likely to be Ga stabilized. These theoretically estimated trends are evidenced by 2D and 3D growth front contours evaluated under various growth conditions as shown in Fig.2 for the case of GaN grown under two different temperatures. The 3-D plots of the growth front are shown here at 400°C and 500°C. The bottom legend in each figure indicates the number of monolayers of the growth front. As one sees, the growth front is 3 monolayers (ML) at 500°C, and 4 MLs at 400°C. This suggests that the growth front is improved as the temperature is increased.

p-doping of nitrides has primarily been based on the use of Cp_2Mg sources. The Mg dopant introduced in this way was reported to require activation by Low-Energy-Electron-Beam-Irradiation (LEEBI) and hole concentrations in the order of $3x10'18cm^{-3}$ have been demonstrated in this way [19]. Hydrogen acceptor compensation was proposed as possible reason for the high resistivity observed with Mg doping. LEEBI allows conversion of Mg-H complexes to the expected Mg doping. Moreover Mg seems to be more sensitive to LEEBI than other group II metals due to the lack of d-electrons and thus possibility of resulting in shallow rather than deep acceptors. Carbon doping has also been attempted more recently using CCl_4 sources and hole concentrations of $3x10'17cm^{-3}$ have been achieved [20]. n-doping is usually achieved by SiH_4 or Si_2H_6 and high electron concentrations of mid $10'19cm^{-3}$ have been demonstrated with smooth, mirror like surfaces.

3. Transport and Electrical Characteristics

The large (~1.5eV) satellite valley separation from the central valley minimum in GaN permits large peak velocities to be attained. Monte-Carlo simulations of electron transport in GaN suggested peak velocity values as high as 2.7x10'7cm/sec for GaN doping levels of $\sim 10'17cm^{-3}$ and showed that intervalley transitions in high electric fields play an important role in spite of

the large separation between the central and upper valleys [21]. This leads to the possibility of realizing electronic devices with reasonably high-frequency characteristics unlike what was initially expected for nitride semiconductors.

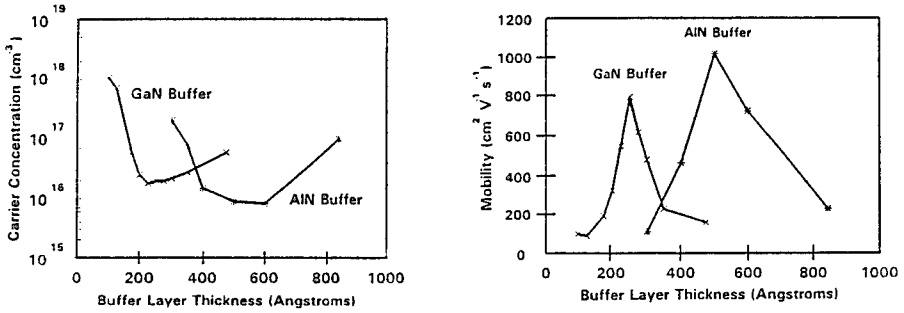


Figure 3. Residual carrier concentration and mobility of GaN films as a function of nitride film thickness.

Reasonably high electron mobilities have been demonstrated experimentally in MOCVD grown GaN. Experiments with GaN thickness varying between 1 and 8μm showed (see Fig. 3) that the mobility increases from 100 to more than 500cm^2/Vsec for 4μm thick layers [8]. Further increase of film thickness lead to mobility degradation. The background carrier concentration showed similar trends, being highest (~10'19cm^{-3}) for 1μm thick films and decreasing to ~10'17cm^{-3} for 4 to 5μm thick GaN. The experimentally obtained mobility values are close to the theoretically expected ones provided that compensation is considered to be present in the samples. They suggest the possibility of using nitrides for electronic device applications such as FETs of various designs.

Electrical contacts on nitrides are still in their infancy but significant progress has been made during the last few years in the understanding of contact formation and their quality improvement. Al is commonly used for ohmic and Au for Schottky contacts [22]. The Schottky barrier height achieved using Au is 0.8 to 0.9eV and ideality factors close to 1 are possible. Annealing at 575°C has, however, been reported to lead to contact degradation as manifested by change to ohmic behavior and is probably related to the presence of Au diffusion [22] Specific contact resistivities of 8'10-6Ωcm^2 have been achieved using Ti/Al [23]. The low contact resistance was speculated to be due to the solid phase reaction between Ti and GaN forming TiN. If the nitrogen is then

out-diffused from the GaN, N vacancies and thus heavy doping would be present at the surface allowing realization of tunneling contacts.

4. Optical Device Applications

Nitride developments have impacted tremendously the area of optical devices providing the possibility of realizing Light-Emitting-Diodes (LEDs) and Lasers in the blue spectrum. These developments will have a tremendous impact on future commercial and military applications such as displays and indicator lights. First reports on blue LEDs employed Mg-doped, LEEBI treated p-type GaN in conjunction with non-intentionally doped, n-type GaN [10]. The diodes showed a threshold voltage of 3V and UV emission (~375nm) was related to band-to-band transitions in the n-layer due to recombination of electrons with injected holes from the p-layer. Violet-blue (VB) emission observed from the same devices at ~425nm was associated with transitions in Mg-related luminescence centers in the p-layer. Double heterostructure AlGaN(p-doped by Mg)/GaN(n-doped by Si) LEDs have also been demonstrated and their electroluminescence spectrum showed UV emission at 372nm related to band-to-band transitions and blue emission at 423nm associated with Mg centers in the p-AlGaN [11],[12].

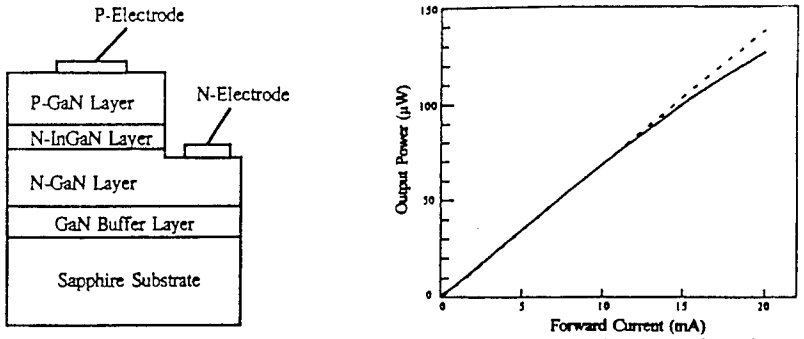


Figure 4. Cross sectional view of blue GaN LED and its associated output power spectrum.

Very promising characteristics for blue LEDs have been presented by Nichia, Japan [13]. A double heterostructure consisting of p-GaN/n-InGaN/n-GaN has been used for this purpose. Its cross section together with the output power characteristics of the diode are shown in Fig.4 [13]. Emission was detected at 430nm. A maximum output power of about 125μW at 20mA can be seen from these characteristics which is higher than the 60μW achieved by Zn(S,Se) at the

same current. The quantum efficiency of the GaN LED was also found to be superior to that of ZnSe; 0.22% and 0.10% for GaN and Zn(S,Se) respectively.

Work towards laser diode development has led to the demonstration of UV emission from AlGaN/GaN/AlGaN double heterostructures [14], as well as, 1 to 3μm thick GaN films deposited on sapphire [4]. Pumping by Argon laser (λ=2750-3050Å, power=40mW) can be used for excitation. The PL spectra of AlGaN/GaN(0.2μm)/AlGaN structures at 30K showed emission at 3350Å and 3600Å from AlGaN and GaN respectively. By replacing the "bulk", 0.2μm thick GaN film by a thin, 300Å GaN quantum-well, peak emission shifted by 38.7meV due to quantum size effects. Vertical cavity and edge emission was demonstrated from GaN (1-3μm)/AlN buffer(500Å)/sapphire films at room temperature. Stimulated emission at 368.2nm was also shown by optical pumping of AlGaN/GaN/AlGaN. These results set up the basis for future laser development using nitrides.

Other optical device applications of nitrides include the development of MSM photodetectors operating in the UV spectrum. Compared with GaAs MSM, the use of GaN permits smaller leakage and thus better signal-to-noise ratios, as well as, higher voltages which would lead to reduction of the Schottky depletion capacitance and thus possibility of higher operation frequencies. Monte Carlo simulations of 0.25μm GaN photodetectors have in fact predicted high cutoff frequencies of 100GHz. These compare with operation cutoff frequencies using similar size GaAs MSM diodes.[24].

5. Electronic Device Applications

Various electronic device applications can be envisaged using nitrides. These include Surface Acoustic wave (SAW) devices, MESFETs and High Electron Mobility Transistor (HEMT) applications. AlN is for example a very promising candidate for SAW applications due to its very high Rayleigh wave velocity (5650m/s) and low dispersion, as necessary for signal processing at high frequencies [15]. The electromechanical coupling constant (k^2) of AlN is reasonably high and can be increased by increasing the growth temperature. The k^2 value is in fact critical for SAW device development since it's proportional to the piezoelectric constant of the film; larger piezoelectricity values can be obtained for larger k^2 values. The immunity of nitrides to high operation temperatures is another feature which renders the use of nitride SAW

devices very useful in applications such as automotive industry and other harsh environments.

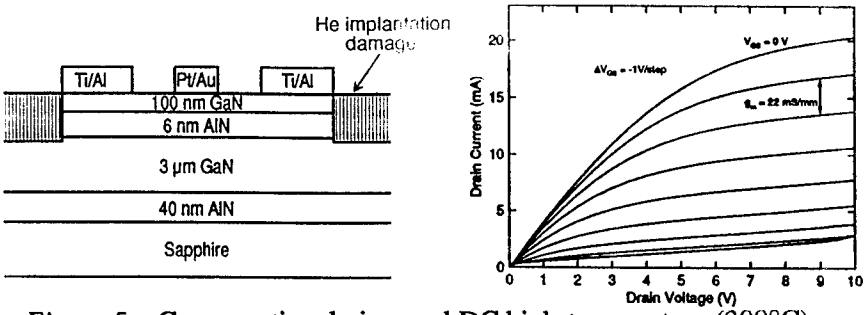


Figure 5. Cross-sectional view and DC high-temperature (300°C) characteristics of AlN/GaN HEMT.

GaN MESFETs have been demonstrated using n-GaN films deposited on sapphire with an intermediate AlN buffer [16]. The devices were isolated by proton implantation. Ohmic contacts were made by Ti/Au (25Å/1500Å) and silver was used for the fabrication of 4µm long gates. The DC characteristics of these devices showed modest performance with transconductances of 23mS/mm at Vg of -1V. More recent results on heterostructure-based approaches showed very promising characteristics as confirmed by both DC and high frequency data [17]. Fig.5 shows the cross-section of a GaN HFET and the DC characteristics of the device at 300°C [17]. The transconductance of 1µm HFETs was 45mS/mm at 30°C and 22mS/mm at 350°C, while the f_t and f_{max} were 8 and 22GHz respectively.

6. General trends and conclusions

The results discussed in the preceding sections show that nitride research has tremendously progressed during the last few years. Material quality has been improved and promising device characteristics have been demonstrated for both optical and electronic components. A major obstacle in nitride development has been the lack of lattice matched substrates. Various possible solutions exist to face this difficulty. The traditional approach used up to now is the development of suitable buffer technology to allow good heteroepitaxial growth. Other alternatives would be the development of growth techniques for bulk GaN substrates and the study of "compliant" substrate technologies which would confine dislocation within an initial "soft" buffer layer.

Growth techniques may be possible to take advantage of the availability of sources such as "supersonic jets" in order to control the energy of injected species and thus improve the quality of the layers by achieving a smaller window of energy spread and lower growth temperatures. Improvements are also still necessary in terms of residual doping control and intentional doping. Basic studies are necessary in this area to fully understand the basic physical properties of nitrides and their relation to the observed electrical and optical characteristics.

Although encouraging results have been obtained for both optical and electronic devices, significant improvements are still necessary before use of such components can be made. This applies to LEDs, lasers, photodetectors operating in the blue spectrum, as well as, MESFETs and HEMTs built on nitrides for high-power/high-temperature operation. Material quality improvement and controlled doping are parameters that impose difficulties in such developments. Basic transport property studies and experiments on heterostructure formation and control are essential for the success of such developments.

New applications may emerge from the above discussed nitride developments. For example, photoemission studies of AlN showed negative electron affinity (NEA) characteristics [25]. These could be extremely interesting for the development of electron emitters. A possible application of them could be microwave tubes where cold cathodes could be realized based on the NEA characteristics of nitrides.

Overall, the success of recent nitride developments during the last five years sets up a solid base for future developments of optical devices in the blue spectrum and electronic components for high-power/high-temperature applications. The synergy between material, chemistry and electrical/electronic engineering expertise is expected to lead to further progress, as necessary for practical use of such components in various applications.

7. Acknowledgment

The author would like to acknowledge the contributions of C.H. Hong, K. Wang, Y. Park and J. Singh on the experimental studies of GaN growth by

MOCVD and the theoretical simulation of its growth. This work was supported by ONR Contract No: N00014-92-J-1552.

8. References

1. J. I. Pankrove, Perspective on gallium nitride, MRS Symp. Proc., (Materials Research Society. Symposia Proceedings) 162(1990) 515
2. K. Das and D.K.Ferry, Hot electron microwave conductivity of wide bandgap semiconductors, Solid-State Electron., 19(1976) 851
3. C.R.Eddy, Jr., T.D. Moustakas, J. Scanlon, Growth of gallium nitide thin films by electron cyclotron resonance microwave plasma-assisted molecular beam epitaxy, J. Appl. Phys., 73, 448, 1993
4. M.A. Khan, J.M.Van Hove ,D.T. Olson, S.Krishnankutty, R.M.Kolbas Growth of high optical and electrical quality GaN layers using low pressure metalorganic chemical vapor deposition, Appl. Phys. lett 58, 526, 1991
5. T.Detchprohm , K. Hiramatsu, N.Sawaki, I. Akasaki The homoepitaxy of GaN by metalorganic vapor phase epitaxy using GaN substrates, J. Crystal Growth 137, 170(1994)
6. K.Hiramatsu, H. Amano, I. Akasaki, H.Kato, N. Koide and K. Manabe MOVPE growth of GaN on a misoriented sapphire substrate , J.Crystal Growth 107(1991)
7. C.H. Hong, K. Wang and D. Pavlidis, Epitaxial growth of cubic GaN on (111) GaAs by Metalorganic Chemical Vapor Deposition, Journal of Electronic Materials, Vol. 24, No.4, 1995, pp.213-218
8. J.N.Kuznia, M.A.Khan, D.T. Olson, R. Kaplan and J. Freitas, Influence of buffer layers in the deposition of high quality single crystal GaN over sappire substrates, J. Appl. Phys., 73, 4700, 1993
9. C.H. Hong, K. Wang and D. Pavlidis, Epitaxial growth and structural properties of cubic GaN on (100) and (111) GaAf grown by metalorganic chemical vapor deposition, Presented at Int. Symp. Compound Semicond., San Diego, 18-22 September 1994, Inst. Phys. Conf. ser. No 141, Chapter 2, pp.107-112
10. I.Akasaki, and H.Amano,, High efficiency UV and blue emitting devices prepared by MOPVE and low enegy electron beam irradiation treatment SPIE Vol 1361,1990
11. I.Asasaki and H.Amano, Conductivity control of AlGaN fabrication of AlGaN/GaN multi-Heterostructure and their application to UV/blue light emitting devices, Mat.. Res. Soc. Symp., Proc., 1992

12. I.Akasaki and H.Amano, Widegap column-III nitride semiconductors for UV/blue light emitting devices , J. Electrchem Soc, Vol 141 No.8 2266,1994
13. S.Nakamura, M.Senoh and T. Mukai, High-Power InGaN/GaN double -heterostructure violet light emitting diodes, Appl. Phy. letters 62(19) 2390, 1993
14. H.Amano, N.Watanabe, N.Koide and I.Akasaki, Rome-temperature low-threshhold surface stimulated emission by optical pumping from $Al_{0.1}Ga_{0.9}$/GaN double Hetereostructure , Jpn Apl. Phys. Vol 32(1993) pp. L1000 1993
15. K.Tsubouchi,K.Sugai andN.Mikoshiba, High-frequency and low-dispersion characteristics of surface acousitic waves on AlN/Al_2O_3
Jpn J. Appl Phys. Vol 19, pp. L751, 1980
16. M.A.Khan, J.N.Kuznia, A.R.Bhattarai and D.T.Olson, Metal semiconductor field effect transistor based on single crustal GaN, Appl. Phys. Letter, 62(15), pp.1787, 1992
17. S.C. Binari, L.B. Rowland, G. Kelner, W. Kruppa, H.B. Dietrich, K. Doverspike and D. K. Gakill, DC, microwave, and high-temperature characteristics of GaN FET structures, Presented at Int. Symp. Compound Semicond., San Diego, 18-22 September 1994, Inst. Phys. Conf. ser. No 141, Chapter 4, pp.459-462
18. K. Wang, J. Singh and D. Pavlidis, Theoretical Study of GaN Growth: A Monte Carlo Approach", Journal of Applied Physics, Vol. 76 (6), 15 September 1994, pp.3502-3510
19. S.Nakamura, M.Senoh and T. Mukai, Highly P-typed Mg-Doped GaN Films Grown with GaN buffer layers , Jpn. J. Appl. Phys. 30, L1708(1991)
20. C.R.Abernathy , J.D.Mackenzie and S.J. Pearton, W.S. Hobson
CCl_4 doping GaN grow by metalorganic molecular beam epitaxy. Appl. Phys. Letters, 66, 1969, 1995
21. B. Gelmont, K.Kim, M.Shur, Monte Carlo simulation of electron transpot in gallium nitride, J. Appl Phys., 74, 1818, 1993
22. J.S.Foresi,and and T.D. Moustakes , Metal contacts to gallium nitride Appl. Phys. Lett, 62, 2859, 1993
23. M.E. Lin, Z. ma, F.Y. Huang, Z.F. Fan, L.H. Allena nd H. Morkoç, Low resistance ohmic contacts on wide band-gap GaN, Appl.
Phys. Lett., 64,(8), 21 February 1994, pp.1003-1005
24. R.P. Joshi, A.N.Dharamsi and J. McAdoo Appl. Phys. lett. 64(26), pp.3611, 1994, Simulations for the high-speed response of GaN metal-semiconductor-metal photodetectors
25. M.C. Benjamin, Chang Wang, R. F. Davis, R.J. Nemanich, Observation of a negative electron affinity for heteroepitaxial AlN on alpha(6H) SiC(0001) , Appl.Phys.lett 64(24) p.3288, 1994

PROSPECTS IN WIDE-GAP SEMICONDUCTOR LASERS

Arto V. Nurmikko and R.L Gunshor*
Division of Engineering and Department of Physics
Brown University, Providence RI 02912, USA

1. Introduction

Compact blue and green semiconductor lasers will impact significantly on such technologies as optical storage and multicolor displays. The availability of a blue semiconductor laser near 450 nm, in conjunction with improvements in encoding topology and superresolution (transcending diffraction limits), is expected to increase the bit density in a compact disk by at least tenfold from that available in today's consumer electronic equipment. Disk storage densities of up to 10 GBit/in^2 are projected over the next decade, comparable to expectations with magnetic disks. Motion pictures, multimedia presentations etc will be among the direct beneficiaries for a portable medium such as the CD-ROM, linked closely to personal computer technology. Optical disk technology in general is destined for explosive growth; for example, compact terabyte data storage has been demonstrated in an optical disk "jukebox".

Two main approaches exist for achieving a compact blue laser source. Using the presently available high power infrared diode lasers (GaAs technology), nonlinear conversion by second harmonic generation has provided tens of mW of coherent blue emission, power levels required for optical writing (reading is performed at mW level). Such equipment, which is becoming commercially available, is however quite expensive due to the complex optoelectronic engineering designs which are also excessively bulky e.g. for a CD-ROM. The second, direct approach is to utilize a wide bandgap semiconductor such as ZnSe or GaN as the basis of an eletrical injection laser - a task which until recently appeared rather hopeless.

There are fundamental and practical reasons why wide bandgap semiconductors, both of the II-VI and III-V variety, met with striking lack of success in attempts to incorporate them in optoelectronic devices and were largely abandoned for such applications (and left largely for dead, or for academics, or for dead academics). In 1991, however, following a half a decade development in molecular beam epitaxial methods for II-VI compounds, first diode laser demonstrations under pulsed conditions at cryogenic temperatures were achieved [1],[2]. Today one can find laboratory demonstrations of room temperature ZnSe-based cw diode lasers at several research facilities and a bright GaN-based blue LED is a commercial reality. Hence a question which had no factual basis of being asked a few years ago: (to

* Purdue University, School of Electrical Engineering, West Lafayette IN 47907 USA

S. Luryi et al. (eds.), Future Trends in Microelectronics, 303–313.

paraphrase) "is a technologically viable blue semiconductor laser really to be or not?"

A wide bandgap semiconductor would generally prefer to be an insulator, unless coaxed into ambipolar electrical activity by insightful synthesis procedures, bordering until recently on nearly theological approaches. In spite of substantial advances on the theoretical front about doping and significant successes in e.g the p-type doping of both ZnSe and GaN

within specific epitaxial growth schemes, microscopic understanding of these processes to guide experiment towards more flexible pn-junction designs is still far from complete (with different reasons for ZnSe and GaN). We will survey this subject below from the perspective of the requirements for blue and near UV diode lasers, including the vital issues of defects and device degradation. The ability to 'bandstructure engineer' a semiconductor by tailored superlattice structrures has found a particularly fertile application in the ZnSe diode lasers - for example the contacts to metal electrodes are configured in this manner.

On the other side of the seesaw, the wide gap semiconductors enjoy fundamentally large interband optical cross sections - an obvious asset for blue light emitters. Excitons and related many electron-hole complexes play an important role in facilitating gain in a ZnSe quantum well, while impurity states coupled strongly to the lattice appear to be of direct benefit to the GaN LED in spite of a highly defected crystal microstructure. In the II-VIs particularly, flexible heterostructure designs have produced, in addition to the advances with the conventional edge emitting lasers, recent demonstrations of vertical cavity surface emitting lasers and evidence for very large normal mode (vacuum Rabi) splittings in a microcavity in the strong coupling regime. Apart from some fascinating science in photonic nanostructures, these developments add to the design flexibility in the basic optical/electronic "infrastructure" for wide gap optical semiconductors, hence further increasing the odds for an eventual emergence of technologically viable blue diode lasers.

In the epitaxy of common II-VI heterostructures ZnSe forms the 'hub' for blue-green lasers. Lattice constant mismatch is a serious design constraint for the II-VI materials and the choice of the substrate is not obvious, either. To date,

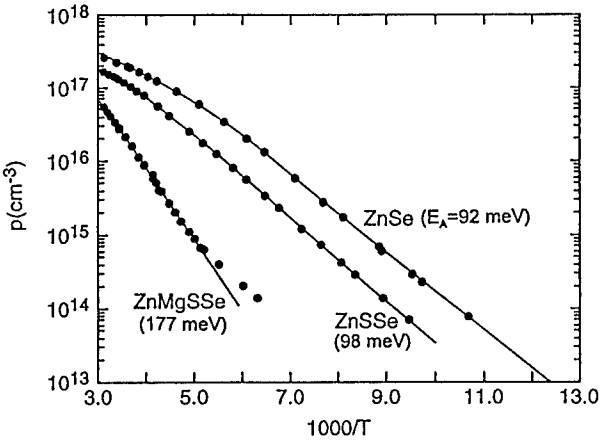


Figure 1: Temperature dependence of the free hole density in p-ZnSe:N, p-Zn(S,Se):N, and (Zn,Mg) (S,Se):N (Ref. 6).

most light emitters have been grown on GaAs substrates/buffer layers which present an 0.25% lattice mismatch to ZnSe, but can be fully lattice matched to Zn(S,Se) and (Zn,Mg)(S,Se) of particular compositions. Results based on ZnSe homoepitaxy are just beginning to appear. Complicating the epitaxy of the II-VI ternaries and quaternaries is the fact that none of the elements have an (even close to) unity sticking coefficient on the growth surface, in strong contrast with most III-V cases.

2. Doping and Transport of ZnSe and its Wide Bandgap Alloys

(a) p-type doping of ZnSe and related compounds

A major advance in the perennial difficulty to dope ZnSe p-type occurred in 1990 when Park et al., and independently, Ohkawa et al. showed how using nitrogen as the acceptor, net acceptor concentrations in excess of 10^{17} cm^{3} could be achieved [3]. This made the fabrication of pn-junctions a reality and led to the demonstration of the diode laser as described in the following section. However, basic questions remain about attaining substantially higher hole densities, especially for the wider gap bandgap alloys. In ZnSe, n-type doping with a shallow donor such as Cl leads readily to electron concentrations of 10^{19} cm^{3} and beyond; on the other hand, p-type doping with column V elements other than nitrogen (such as As and P) has been unsuccessful. The physical pictures describing such restricted doping are based on either lattice relaxation effects for the formation of deep levels [4], or issues of solid solubility [5]. In the former, an unsuccessful acceptor candidate distorts the tetrahedral bonding arrangement strongly (including bondbreaking) leading to a deep level state; qualitatively, the position of the valence band maximum (and such physical parameters as the covalent radius and electronegativity of the anion) give some guidance for the anticipated trends towards deep level formation among different host compounds. When comparing ZnSe:N and ZnTe:N, for example, the latter may be predicted to have a shallower acceptor ground state. For p-ZnSe, calculated trends among the different column V (substitutional) acceptors suggest that the lightest element, nitrogen, is least likely to form a deep level, localized state. Experimentally, a series of observations have been made recently on epitaxial layers

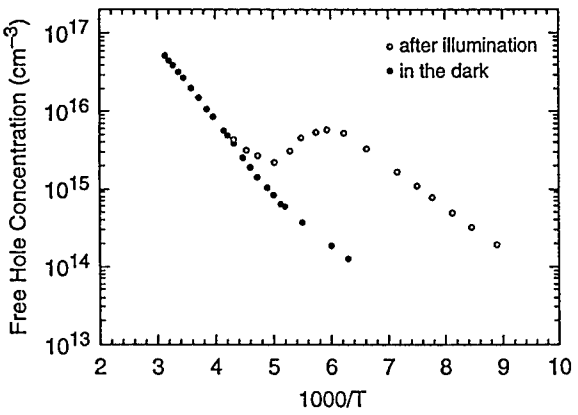


Figure2: The persistent photoconductivity effect in p-(Zn,Mg)(S,Se);(Ref. 7).

under nominally the same nitrogen doping conditions for p-ZnSe:N, p-Zn(S,Se):N, and (Zn,Mg)(S,Se):N, that lend support to this view [6]. **Figure 1** displays the hole concentration as a function of inverse temperature. The hole concentration decreases distinctly as the polarity (and disorder) increases, while the acceptor state deepens from about 90 meV to nearly 180 meV. In fact, closer examination of the Hall data for the quaternary shows a second slope, indicative of a second acceptor state.

The nature of the deeper acceptor state in p-ZnMgSSe is elucidated in **Figure 2** which shows the persistent photo-conductivity observed in p-(Zn,Mg)(S,Se) in Hall measurements [7]. The resistivity of the epitaxial layer increases by several orders of magnitude upon illumination with visible light. The transport experiments support the proposition that this acceptor state is configurationally lattice relaxed. The situation is analogous to the DX center in (Ga,Al)As except that holes are at issue. These observations show that the challenges to p-type doping may increase substantially with increasing Mg and S concentrations in ZnMgSSe, as one attempts to design light emitters for deep blue wavelengths. At present, II-VI lasers have been operated to about 460 nm, while the best devices are still near the 500 nm wavelength. The occurrence of lattice relaxation is, of course, not unexpected in the relatively polar wide gap II-VI semiconductors. One should also expect comparable effects in the wide gap nitrides, given the large electron (hole)-phonon interaction strength in these III-V semiconductors.

(b) Low Resistance Electrical Contacts

Progress in p-doping of ZnSe notwithstanding, large Schottky barriers are immediately observed if direct contacting to a metal is attempted. (The first diode lasers required working voltages on the order of 20 V). However, once high levels of p-doping in ZnTe:N were achieved, with free-hole concentrations reaching $p=10^{19}$ cm-3 [8], and since e.g. palladium was found to form a low-resistance contact to such ZnTe epilayers [9], a 'bandstructure engineered' low resistance contact became feasible. In particular, a p-Zn(Se,Te) graded bandgap

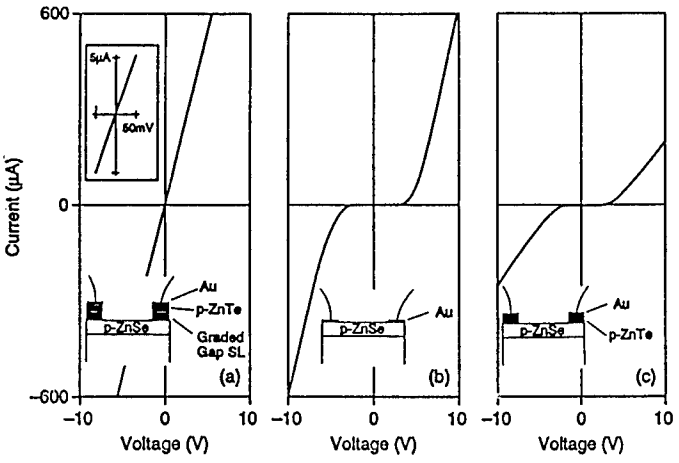


Figure 3: Current-voltage characteristics of a p-ZnSe epitaxial layer, contacted by the ZnSeTe graded bandgap scheme (from Ref. 8).

scheme was designed and implemented by Fan et al.; specifically to reduce the space charge and potential energy barriers which would otherwise present a large electrical impedance at an abrupt p-ZnTe/p-ZnSe hetero-interface [10]. (For reasons of precise control of the Te concentration, a short period p-type, highly strained ZnTe/ZnSe superlattice of 20Å per period was designed, with the ZnTe and ZnSe layer thicknesses in each cell varying to approximate a graded bandgap material). A very similar structure has been discussed in terms of resonant tunneling of holes by researchers at Sony Laboratories [11]. **Figure 3** shows how the overall contact to p-ZnSe is quite ohmic; the specific contact resistance was determined to be in the range of $2x10^{-4}$ Ωcm^2. The graded bandgap contact scheme is now a standard in the laser diode devices and has also been instrumental in facilitating a range of transport measurements.

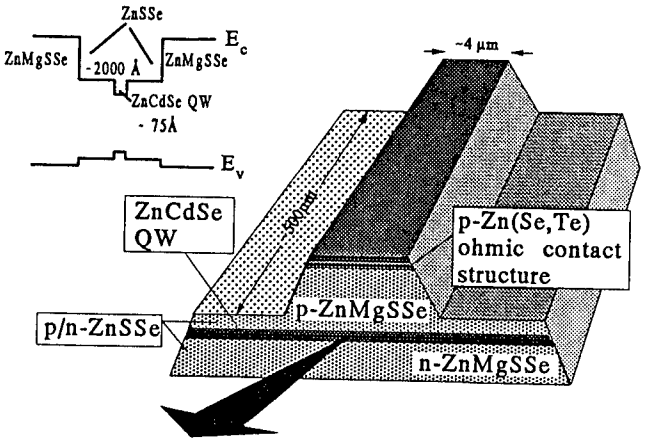


Figure 4: Schematic of the ZnCdSe/ ZnSSe/ZnMgSSe SCH design

3. Diode Lasers

(a) Device Design and Performance

In any semiconductor laser, the chief design criteria is to combine electronic and optical confinement by enveloping a quantum well with an optical waveguide within a multilayer heterostructure. Due to lattice mismatch constraints, which impose a severe penalty if strain relaxation is allowed to take place (in the form of misfit dislocations), the early blue-green diode lasers which featured layered arrangements of (Zn,Cd)Se, ZnSe, and Zn(S,Se) were not optimally configured (while also lacking a useful contact scheme to the topmost p-type II-VI layer. The demonstration at Sony Laboratories [12] and at Philips Laboratories in 1993 [13] that the quaternary (Zn,Mg)(S,Se) could be grown, doped in an ambipolar fashion (with the caveats above), and incorporated to a diode laser, made the design and fabrication of a pseudomorphic SCH finally possible. **Figure 4** shows the schematic of such a heterostructure with the Cd-concentration typically $x \approx 0.20$, sulfur concentration $y \approx 0.07$, and the Mg-concentration $z \approx 0.12$ for a laser emitting at around 500 nm.

Significant improvement of the SCH-QW device performance has been realized in index guided ridge waveguide structures to reduce the injection current. **Figure 5** shows the *continuous-wave* output/input characteristics at λ=508 nm of three nominally identical 4.2μm wide devices from the authors' laboratories, where the ZnSeTe graded bandgap ohmic contacts and high reflectivity facet coatings were also incorporated [14]. Threshold voltages well below 6V and currents below 10mA have now been reached, with differential efficiencies up to 60%. Single transverse mode characteristics have been obtained in the ridge waveguide geometry for both single and multiple quantum well devices. Initial devices sustained the cw operation less than a minute before failure but the groups at Sony laboratories and at 3M/Philips have recently extended the longevity to the scale of hours, illustrative of the rapid progress in the field (when compared e.g. with the evolution of the GaAs injection laser in the 1970s). The problem of device degradation and lifetime of the II-VI lasers is clearly a vital one, whose solution is imperative for technological applications of these devices. Given the mechanical properties of these relatively polar semiconductors (suggesting more vigorous dislocation dynamics), coupled with the low MBE growth temperature, the control of defects and associate device degradation poses an inherently larger challenge than in GaAs lasers.

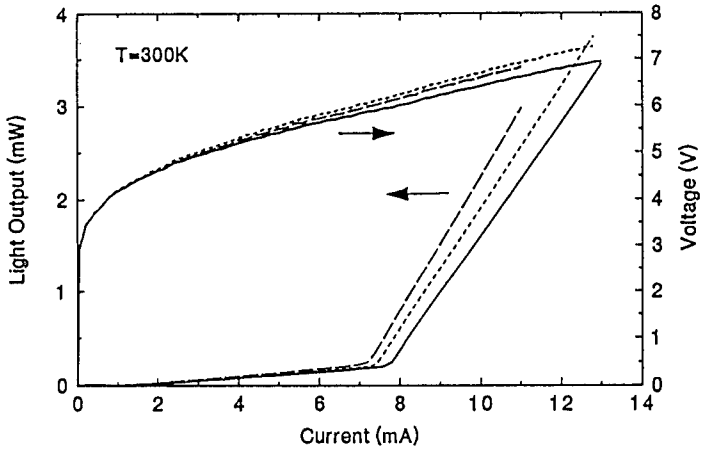


Figure 5: Room temperature, continuous-wave operating characteristics of three nominally identical index guided SCH lasers at λ = 508 nm.

Recently, researchers at 3M, Philips and the authors' group have conducted systematic studies of the microstructure associated with the device failure [15],[16]. We now know that *extrinsic* morphological defects forming during epitaxy act as launching centers for dislocation networks. The defects are predominantly stacking faults created at the GaAs/ZnSe heterovalent interface, and lead to the presence of a finite initial threading dislocation density in the strained active QW layer of the laser. Nonradiative recombination in the QW at nearby point defects induces further dislocation activity such as the generation of dislocation loops and leads to enetual optical and electrical degradation of the QW. The dislocation multiplication proceeds at a finite rate, even under relatively modest current densities ($\approx$ 100 A/cm^2 as in a

LED). Several groups are now focusing their attention to improving the epitaxy of laser quality material, with specific emphasis on the control of stacking faults and point defects. For example, the stacking fault density has been reduced from about 10^6 cm^{-2} a couple of years ago to about 10^4 cm^{-2} presently; correspondingly device lifetime has increased thousandfold.

(b) Physics of Gain and Stimulated Emission

In formulating the microscopic description of the gain supplied by an electron-hole gas in a QW, it is important to note the differences between GaAs-based and ZnSe-based heterostructures. The fact that the exciton binding energy (Coulomb interaction) in ZnSe-based QWs can reach the condition $E_x > h\omega_{LO}$, kT (at room temperature) raises a question about the nature of many-body electron-hole states in a II-VI blue-green laser. In optical pumping experiments at cryogenic temperatures (typically 10-100K), it has been demonstrated directly that stimulated emission is dominated by exciton-like process in the ZnSe-based QWs [17]. Here we give an example of gain spectroscopy at room temperature for a (Zn,Cd)Se/ ZnSe/Zn(S,Se) index guided device [18] while probing the optical constants near the n=1 HH QW exciton resonance, performed by a method first introduced by C.H. Henry et al. [19]. The technique makes use of fundamental and very general relationships between spontaneous and stimulated emission. **Figure 6** shows gain/absorption spectra near the n=1 interband transition of the active 75 Å thick ZnCdSe QW at T=300K. Under low current injection, both the n=1 heavy hole (HH) and light hole (LH) exciton absorption peaks are well defined. With increasing injection, both absorption peaks begin to saturate due to many-body effects. At laser threshold (I≈300 mA in gain guided devices), the LH exciton peak remains well defined, and even the HH exciton resonance remains discernible, while supplying gain on its low energy side. Hence it appears that electron-hole pairwise Coulomb correlation remains important in the diode laser operating conditions. Such a

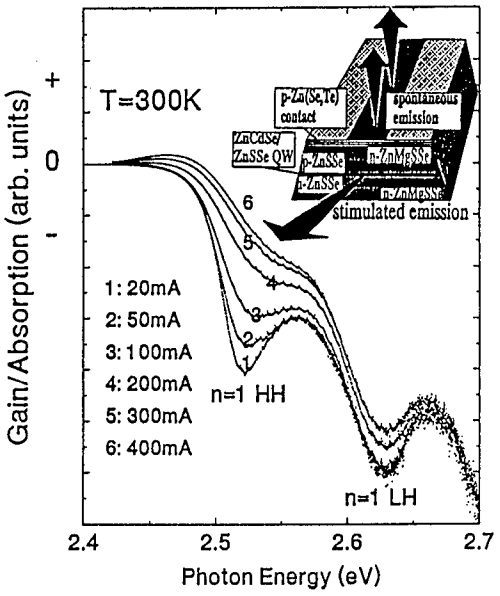


Figure 6: Experimental gain/loss spectra of a ZnCdSe/ZnSSe/ZnMgSSe SCH diode laser as a function of injection current, including the n =1 QW exciton resonance region (Ref. 18)

correlation would result in an enhancement in the electron-hole wavefunction overlap when compared to a one-electron picture. A corresponding enhancement in the optical transition matrix element has been measured in the spontaneous recombination rate.

4. Other Geometries: Vertical Cavities and Microcavity Physics in the Blue-Green

In parallel to the development of the edge emitting diode lasers, there are ample reasons to explore other resonator configurations such as the vertical cavity surface emitting laser (VCSEL). At room temperature the gain of a ZnCdSe QW SCH diode laser is approximately $g_{th} \approx 1500$ cm^{-1} (at a current density I~500 A/cm^2). In designing a VCSEL structure, this implies a requirement for the mirror reflectivity of R=0.997 for a three QW thick gain medium ($L_w \approx$ 75Å). While the direct epitaxial growth of distributed feedback Bragg (DBR) reflectors is in principle possible, the small differences in the indexes of refraction e.g. for ZnSSe and ZnMgSSe requires a very large number of λ/4 layer pairs (~100), grown with high precision. We have pursued the alternative of dielectric coating materials in the form of SiO_2/TiO_2 DBR stacks.

Figure 7(a) shows a schematic of the device arrangement for optical pumping studies in which the 3QW gain segment, part of a ZnCdSe/ZnSSe/ZnMgSSe SCH pseudomorphic structure, was placed at an antinode (maximum optical field) in a 5λ long vertical cavity [20]. **Figure 7(b)** shows the spectral characteristics at T= 200K and 300K of the emission recorded in the vertical direction of a typical laser structures both below ($0.8W_{th}$) and above the threshold ($1.2W_{th}$). The onset of laser emission was confirmed visually by the emergence of a distinct beam of circular cross section emerging from the devices at a typical angle of divergence of 4-5°. Conversion

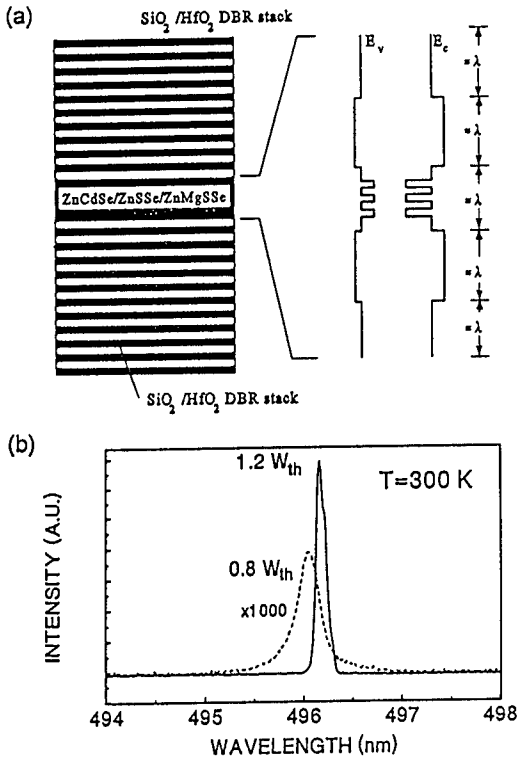


Figure 7: (a) schematic of blue-green VCSEL; (b) emission spectrum below and above threshold at T=300K (Ref. 20)

efficiencies of about 20% were measured below room temperature. As expected, already below the lasing threshold the impact and the presence of the vertical cavity is profound in producing strong linewidth reduction in the spontaneous emission, given here by $\Delta\lambda \approx 0.26$ nm ($\Delta E \approx 1.65$ meV). Above threshold, the laser emission was always <110> polarized. Work is presently underway to explore the fabrication of electrically injected VCSELs.

Short vertical resonator structures, planar "quantum" microcavities ($L \sim \lambda$), offer an opportunity to explore the coupling of electromagnetic radiation with electronic resonances in semiconductors at a fundamental level. As for applications, predictions range from high brightness LEDs to 'thresholdless' lasers. Both linear and nonlinear optical experiments have been performed on III-V microcavities to show a diversity of phenomena ranging from enhanced/inhibited spontaneous emission to the observation of the so-called vacuuum Rabi splittings (in analog to work with atom optics in high Q resonators). The II-VI compound QWs are particularly suited for such investigations due to the large exciton oscillator strengths that can readily lead to the strong coupling regime, defined approximately by the condition $\Omega > \gamma, \gamma_c$, where Ω is the Rabi frequency (proportional to the square root of the oscillator strength) and γ, γ_c^{-1} are the exciton linewidth broadening and the photon cavity lifetime, respectively. Recent work has demonstrated microcavity effects in a II-VI system; here we show an example from linear spectroscopy applied to an undoped

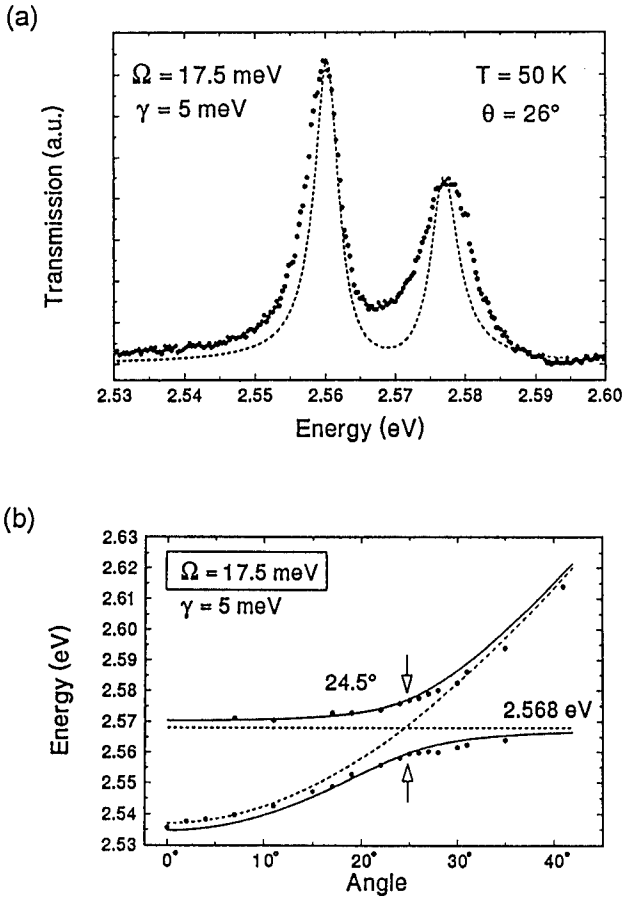


Figure 8: (a) Rabi-splitting in the transmission spectrum of a ZnCdSe QW microcavity at the n=1 exciton state; (b) the normal mode frequencies over a range of detuning (21)

ZnCdSe/ZnSSe/ZnMgSSe SCH structure with three 75 Å thick ZnCdSe QWs [21]. To tune the frequency separation of the cavity mode and the electronic (exciton) resonance, we resorted to angle tuning (also temperature tuning) to generate transmission spectra such as shown in **Figure 8**(a) which demonstrates the normal mode splitting when the uncoupled oscillator frequencies are made equal to each other, here at an angle of θ=26° at T=50 K. Note that the minimum splitting, the Rabi frequency, at this "closest approach" of the mixed exciton-cavity mode is $\Omega \approx 17.5$ meV. This value not only exceeds the substantial inhomogeneous linewidth in the (Zn,Cd)Se QWs but is also a sizable fraction of the HH exciton binding energy. The anti-crossing feature of the coupled electromagnetic-excitonic oscillators is shown further in **Figure 8(b)**. Both figures also includes a theoretical fit, subject to the following parameters: mirror reflectivity R=0.994, cavity linewidth $\gamma_c=c(1-R)/(2n_{cav}L_{eff})=2.1$ meV , nonradiative broadening γ=5.0 meV, and the "radiative linewidth" of the exciton $h\Gamma_o$=0.3 meV for the total three QWs. The "bare" exciton resonance was fixed at E_x=2.568 eV with an effective index of refraction n_{cav}=2.68.

The observed Rabi splitting is much larger than obtained in the GaAs microresonators. While the work with the II-VI systems is only beginning in this field, it is possible to envision microresonator arrangements in which $\Omega > kT$ can be achieved at room temperature. This implies a very high degree of intermixing of the electronic and photonic waves and suggests many new and fundamentally distinct possibilities for blue-green light emitters.

5. Summary Remarks

In this overview we have touched on current work on the new blue-green II-VI lasers. The field is moving rapidly and any overview is subject to annual revision. Clearly, further advances need be made e.g. for improving the vertical transport, the p-type doping, and QW designs for larger gain. Better control of crystalline defects, including choices of substrate and buffer layers, and their identification in the device degradation process are crucial issues now when technological prospects of these diode lasers need be evaluated realistically and competitively. Much innovation and basic research is also awaiting the researchers in this expanding field. Extension deeper into the blue and near UV is being pursued, as well as studies of new lasers such as small vertical cavity emitters.

Acknowledgement ~ The authors wish to acknowledge the following members of their group: J. Ding, H. Jeon, M. Hagerott, P. Kelkar, V. Kozlov, A. Salokatve, and M. Hovinen at Brown, and D. Grillo, Y. Fan, J. Han, Li He, and M. Ringle at Purdue. The research was supported by ARPA (N00014-92-J-1893), NSF (at Brown DMR-9112329 and MRG/DMR-9121747; at Purdue DMR-9202957) and by AFOSR (F49620-92-J-0440).

References:

[1] M. Haase, J. Qiu, J. DePuydt, and H. Cheng, (1991) *Appl. Phys. Lett.* **59**, 1272
[2] H. Jeon, J. Ding, W. Patterson, A.V. Nurmikko, W. Xie, D.C. Grillo, M. Kobayashi, and R.L. Gunshor, (1991) *Appl. Phys. Lett.* **59**, 3619.
[3] R.M. Park, M.B. Troffer, C.M.Rouleau, J.M. De Puydt, and M.A. Haase, *Appl. Phys. Lett.* **57**, 2127 (1990); K. Ohkawa, T. Karasawa, O.Yamazaki, (1991) *Jpn J. Appl. Phys.* **30**, L152.
[4] D.J. Chadi, (1994) J. *Cryst. Growth* **138**, 295.
[5] D.Laks, C.G. Van de Walle, G. Neumark, and S.Pantelides, (1993) *Appl. Phys. Lett.* **63**, 1375
[6] J. Han, Y. Fan, M. Ringle, L. He, D. Grillo, R. Gunshor, G. Hua, and N. Otsuka, (1994) *J. Cryst. Growth* **138**, 464.
[7] J. Han, M. Ringle, Y. Fan, R Gunshor, and A Nurmikko, (1994) *Appl. Phys. Lett.* **65**, 3230
[8] J. Han, T. Stavrinides, M. Kobayashi, M. Hagerott, and A.V. Nurmikko, (1993) *Appl. Phys. Lett.* **62**, 840.
[9] H. Okuyama, T. Miyajima, Y. Morinaga, F. Hiei, M. Ozawa and K. Akimoto, (1992) *Electr. Lett.* **28**, 1798.
[10] Y. Fan, J. Han, L. He, J. Saraie, R.L. Gunshor, M. Hagerott, H. Jeon, A.V. Nurmikko, (1992) *Appl. Phys. Lett.* **61**, 3160.
[11] M. Ozawa, F. Hiel, A. Ishibashi, and K. Akimoto, (1993) *Electr. Lett.* **29**, 503-504.
[12] N. Nakayama, S. Itoh, K. Nakano, H. Okuyama, M. Ozawa, A. Ishibashi, M. Ikeda, and Y. Mori, (1993) *Electr. Lett.* **29**, 1488.
[13] J.M. Gaines, R.R. Drenten, K.W. Haberern, T. Marshall, P. Mensz, and J. Petruzzello, (1993) *Appl. Phys. Lett.* **62**, 2462-2464.
[14] A. Salokatve, H. Jeon, J. Ding, M. Hovinen, A. Nurmikko, D.C. Grillo, J. Han, H. Li, R.L. Gunshor, C. Hua, and N. Otsuka, (1993) *Electr. Lett.* **29**, 2192
[15] S.Guha,J.DePuydt, M. Haase, J.Qiu, and H.Cheng, (1993) *Appl. Phys. Lett.* **63**, 3107
[16] G.C. Hua, N. Otsuka, D.C. Grillo, Y. Fan, M.D. Ringle, R.L. Gunshor, M. Hovinen, and A. Nurmikko, (1994) *Appl. Phys. Lett.* **65**, 1331
[17] J. Ding, T. Ishihara, M. Hagerott, H. Jeon, and A.V. Nurmikko, (1992) *Phys. Rev. Lett.* **60**, 1707-1710, (1993) *Phys. Rev.* **B47**, 10528-10537 .
[18] J. Ding, M. Hagerott, P. Kelkar, A.V. Nurmikko, D.C. Grillo, Li He, J. Han, and R.L. Gunshor, (1994) *Phys. Rev.* **B50**, 5787.
[19] C.H. Henry, R.A. Logan, and F.R. Merritt, (1980) *J. Appl. Phys.* **51**, 3042.
[20] H. Jeon, V. Kozlov, P. Kelkar, A.V. Nurmikko, D.Grillo, J. Han, M. Ringle, and R.L. Gunshor, (1995) Electronics Lett. **31**, 106; (1995) Appl. Phys. Lett. **67**, 1668
[21] P. Kelkar, V. Kozlov, H. Jeon, A. V. Nurmikko,C.-C. Chu, D. C. Grillo, J. Han, C.G. Hua, and R. L. Gunshor, (1995) Phys.Rev.**B52**, R5491

ORGANIC TRANSISTORS — PRESENT AND FUTURE

G. HOROWITZ

Laboratoire des Matériaux Moléculaires
C. N. R. S., 2 rue Henry-Dunant
94320 Thiais
France

1. Introduction

Organic materials are almost everywhere in electronic devices. They are used for instance in lithography and encapsulation. They are everywhere, but at the very heart of the device, upon which silicon still imposes its dictatorship. Nevertheless, organic semiconductors do exist, and have indeed been largely studied since the early fifties [1]. It has been shown that metal-semiconductor (MS) and metal-insulator-semiconductor (MIS) structures can be realized with these organic materials. Practical applications were even thought to be within reach in 1978, when a photovoltaic cell made with merocyanine — an organic dye — was claimed to present a power efficiency close to 1 % under AM1 solar illumination [2]. Unfortunately, this yield, which is still one order of magnitude too small, has not been improved to date, but the domain is still active [3-5].

The second chance for organic semiconductors came in the early eighties, with the emergence of conjugated polymers and oligomers. Conjugated polymers present the unique property of having their conductivity increased by several orders of magnitude upon doping. Highly conducting polyacetylene has been claimed to present a conductivity close to that of metals [6]. It has been recognized more recently that the non intentionally doped ("undoped") form of conjugated polymers and oligomers constitute a new class of organic semiconductors. A tremendous amount of work has been done on light-emitting diodes made with organic conjugated polymers, [7-9] after it appeared that these devices could compete with their mineral counter-part in term of efficiency and extent of colors that can be produced.

The field-effect transistor (FET) is another device where conjugated polymers and oligomers have proved useful [10-16]. Of these, polythiophenes [10,12,14,16] and oligothiophenes [13-15] seem to be the most promising. Polythiophenes present a field-effect mobility ranging from 10^{-5} to 10^{-4} $cm^2V^{-1}s^{-1}$, whereas that of oligothiophenes, and more precisely sexithiophene (6T) and some of its derivatives, is two orders of magnitude higher. Although these values are far below those of mineral semiconductors, they are approaching that of hydrogenated amorphous silicon (a-Si:H), which would make organic materials particularly suitable for large area devices, e.g., flat panel displays. The electrical properties of these polymers and oligomers have been found to

S. Luryi et al. (eds.), Future Trends in Microelectronics, 315–326.

be much dependent on the purity of the material and method of film preparation [17,18]. They can also be modulated through chemical derivations.

In the present paper, we describe the fabrication and operating mode of organic field-effect transistors (OFETs), and review devices published to date. We will then focus on oligothiophenes, and show that a strong correlation can be found between their electrical and structural properties. Accordingly, the device performance can be optimized by both chemical and physical means.

2. Fabrication and operating mode of OFETs

Figure 1 gives a schematic view of the most commonly used structure. OFETs are fabricated according to the thin film transistor (TFT) architecture, which was first described thirty years ago [19]. Because an insulating layer is not easy to deposit on top of an organic semiconductor, the structure is inverted, i.e., it is built over the gate electrode. Most often, the insulator is silicon oxide thermally grown on a silicon wafer. The Si substrate, which plays no active role in the device, serves as the gate electrode. Two gold source and drain electrodes are evaporated on the silicon oxide before the deposition of the active organic semiconducting layer. They form ohmic contacts to the semiconducting layer.

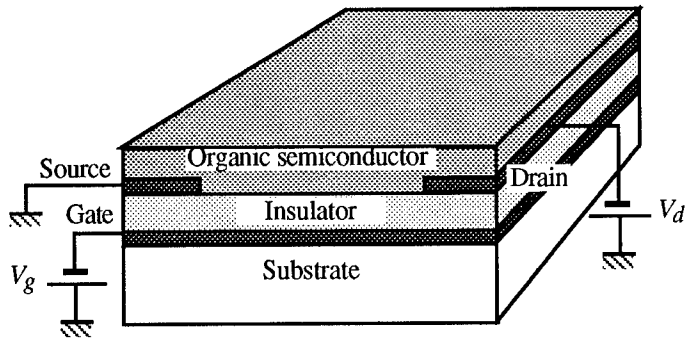


Figure 1. Schematic view of an organic field-effect transistor.

We have recently developed an all-organic structure in which the insulator is a spin-coated polymer — generally poly(methyl-methacrylate) (PMMA) [20,21]. The aluminum gate is evaporated on the substrate before the deposition of the insulating layer. Gold source and drain and organic semiconductor are then deposited in sequence. A main advantage of this structure is the possibility to built it on a flexible substrate.

The characterization of OFETs consists of measuring the drain current I_d as a function of the source-drain voltage V_d for various source-gate voltages V_g. A set of such I_d-V_d curves is shown in Figure 2.

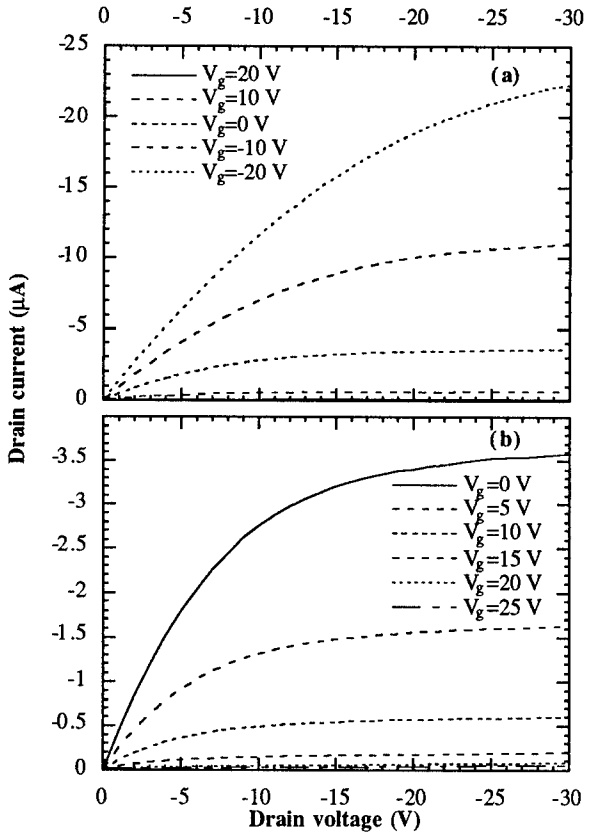


Figure 2. Current voltage characteristics of an organic field-effect transistor. The semiconductor is a sexithiophene derivative, and the insulator PMMA. The device operates either in the enrichment (top curves, V_g<0) or the depletion (bottom, V_g>0) regime. The geometrical parameters of the device are L=50 μm, W=5 mm and C_i=10 nF/cm^2.

The device can operate either in the enrichment mode, Figure 2a, or in the depletion mode, Figure 2b. The signs of both V_d and V_g are consistent with a p-type semiconductor (here 6T). In the enrichment mode, the curves can be described by Equations (1) [22].

$$I_d = \frac{W}{L} C_i \mu_{FET} \left[(V_g - V_t) V_d - \frac{V_d^2}{2} \right] \qquad (V_d < V_g) \qquad (1a)$$

$$I_{d,sat} = \frac{W}{2L} C_i \mu_{FET} (V_g - V_t)^2 \qquad (V_d > V_g) \qquad (1b)$$

Here, W and L are the channel width and length, respectively, C_i is the insulator capacitance (per unit area), μ_{FET} is the field-effect mobility and V_t the threshold voltage. Curves in the enrichment mode can thus be used to determine the field-effect mobility.

In the depletion mode, the curves show that the semiconducting layer can be fully depleted. (The small remaining ohmic current is due to leakage through the insulator.) This occurs when the width of the depletion layer equals that of the semiconducting layer, that is when the gate voltage equals the pinch-off voltage V_p, Equation (2) [22,23].

$$V_p = \frac{qNd^2}{2\varepsilon_s\varepsilon_0}\left(1 + 2\frac{C_s}{C_i}\right) \tag{2}$$

Here, q is the electron charge, N the dopant concentration, d the thickness of the semiconductor, ε_s its dielectric constant, ε_0 the permittivity of free space, and C_s the dielectric capacitance of the semiconducting layer ($C_s=\varepsilon_s\varepsilon_0/d$). Equation (2) can be used to estimate the doping level of the organic semiconductor.

3. Materials used in OFETs

The organic materials used in field-effect transistors can be sorted into two main groups: polymers and molecular materials.

3.1. CONJUGATED POLYMERS

These are trans-polyacetylene (PA), [11] polythiophene (PT) [10,12,14,16] and poly(thienylene-vinylene) (PTV) [24], Figure 3.

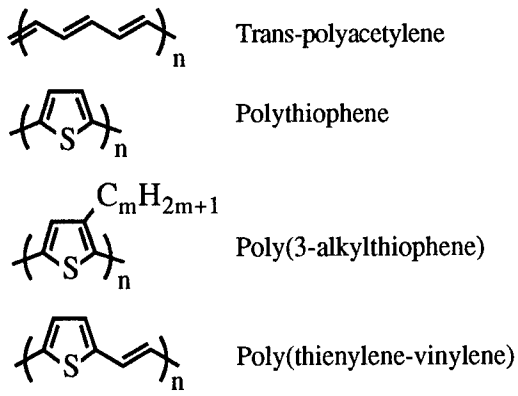


Figure 3. Chemical structure of the conjugated polymers used in OFETs.

Most of the early conjugated polymers were infusible and insoluble. They were therefore difficult to obtain as thin films. Films of PT could be grow on the pre-deposited source and drain electrode by electrochemical polymerization. Such obtained films are poorly defined, and the performance of the devices was limited. More recently, spin-cast films of conducting polymers have been realized, either from a soluble precursor polymer, or from a conjugated polymer substituted with solubilizing groups, e.g., poly(3-alkylthiophene), Figure 3.

3.2. MOLECULAR MATERIALS

By this we mean materials made of smaller molecules than polymers. They comprise conjugated oligomers, and other conjugated molecules. We note that the early organic semiconductors, e.g., anthracene, were members of this class of materials.

3.2.1. *Conjugated oligomers*

These are almost exclusively the thiophene oligomers (*n*T, where *n* stands for the number of thiophene units). Figure 4 shows the molecular structure of sexithiophene, the most widely used oligothiophene, and of two of its derivatives. Thiophene oligomers are deposited by vacuum evaporation. Longer oligomers substituted by solubilizing pendent alkyl groups (up to 16T [25]) have been isolated recently. As oligothiophenes beyond 8T tend to crack when heated under vacuum, these compounds are spin cast.

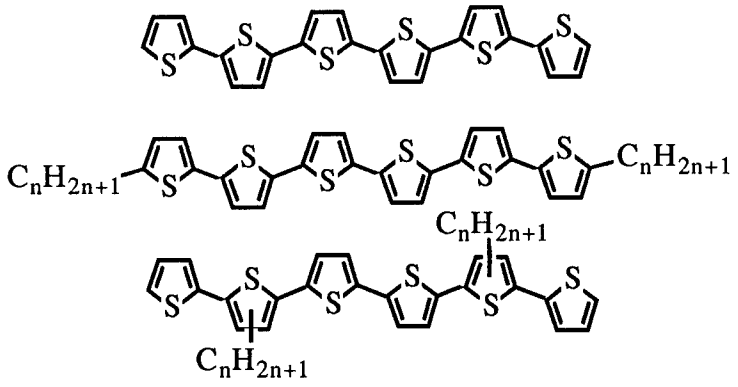


Figure 4. Molecular structure of sexithiophene (top) and its derivatives substituted with alkyl chains at end (middle) and pendent (bottom) positions.

3.2.2. *Other conjugated molecules*

OFETs have been made with various diphthalocyanines [26,27], which are all p-type. More recently, some n-type organic semiconductors have been studied. These compounds are electron donors, such as tetracyanoquinodimethane (TCNQ) [28], and fullerene C_{60} [29].

4. Performance of the organic FETs

4.1. "AMORPHOUS" ORGANIC SEMICONDUCTORS

The electrical parameters of "amorphous" organic semiconductors have been recently reviewed by Brown and coworkers. They found a "universal" relationship between the conductivity σ and mobility μ, of the form $\mu \propto \sigma^{\delta}$, which means that an increase of the mobility is always associated with an increase of the conductivity. Data are gathered in Table 1. We note that an OFET based on a Langmuir-Blodgett film of the charge-transfer material (N-octadecylpyridinium)-Ni(dmit)$_2$ has been recently claimed to present a field-effect mobility of 0.18 $cm^2V^{-1}s^{-1}$ [32]. Unfortunately, the conductivity (1 S/cm) is also very high, which considerably limits on-off ratio, i.e., the ratio of the saturation drain current in the accumulation regime, to that when the gate bias is zero.

TABLE 1. Field-effect mobility and conductivity of variously doped "amorphous" organic semiconductors (unless otherwise stated, all materials are p-type)

Material	Deposition technique	Conductivity S/cm	Mobility $cm^2V^{-1}s^{-1}$	Ref.
Poly(hexylthiophene)	Langmuir-Blodgett	4×10^{-7}	7×10^{-7}	14
Quinquethiophene	Langmuir-Blodgett	3×10^{-7}	1×10^{-5}	14
Sc-diphthalocyanine	vacuum evaporation	5×10^{-6}	10^{-3}	26
Lu-diphthlocyanine	vacuum evaporation	10^{-5} - 10^{-3}	10^{-4} - 10^{-3}	27
Tm-diphthalocyanine	vacuum evaporation	10^{-4} - 10^{-3}	10^{-4} - 10^{-2}	27
poly(alkylthiophene)	spin coating	10^{-8} - 10^{-5}	10^{-8} - 10^{-5}	31
TCNQ (n-type)	vacuum evaporation	10^{-10} - 10^{-6}	10^{-10} - 10^{-4}	28
poly(DOT)$_3$	spin coating	10^{-8} - 10^{-5}	10^{-6} - 10^{-3}	30
C_{60} (n-type)	vacuum evaporation	10^{-8} - 10^{-2}	10^{-5} - 10^{-2}	29
(*N*-octa')-Ni(dmit)$_2$	Langmuir-Blodgett	1	0.18	32

4.2. OLIGOTHIOPHENES

Oligothiophenes deviate from the general trend observed above. This is probably because of their self-assembling behavior, which leads to a crystalline structure of the evaporated films.

4.2.1. *Effect of the length of the oligomer*

Table 2 gives the conductivity and field-effect mobility of vacuum evaporated oligothiophenes, from terthiophene (3T) up to octithiophene (8T). The most remarkable feature of these data is the huge increase of the mobility up to the sexithiophene. We note that the mobility increases must faster than the conductivity. The mobility of octithiophene 8T is lower than that of 6T, and compares with that of polythiophene. We

also note that the conductivity of *n*Ts is anisotropic, the charge transport being favored in the direction parallel to the film.

4.2.2. *Effect of substitutions*

Substitution on the oligomer can take place either at ends of the molecule [15,33], or as pendent groups [34]. The effect of these substitutions is shown in Table 3 and 4, respectively. End substitution results in an increase by a factor of 10 to 100 of the field-effect mobility, whereas the conductivity is practically unchanged. Substitution by pendent groups gives quasi insulating materials. Measurable conductivity and mobility are only obtained in very long oligomers (12T), where they compare to that of the parent polymer.

TABLE 2. Conductivity and field effect mobility of vacuum evaporated non substituted oligothiophenes *n*T

Oligomer	Conductivity (S/cm)		Mobiliy ($cm^2V^{-1}s^{-1}$)
	perpendicular	parallel	
3T	10^{-10}		$<10^{-7}$
4T	10^{-9}		2×10^{-7}
5T	10^{-8}	10^{-6}	$2\text{-}5\times10^{-5}$
6T	2×10^{-7}	1×10^{-6}	2×10^{-3}
8T	10^{-7}		2×10^{-4}

TABLE 3. Conductivity and field-effect mobility of oligothiophenes substituted with alkyl chains at end positions. DE: diethyl, DH: dihexyl

Oligomer	Conductivity (S/cm)		Mobiliy ($cm^2V^{-1}s^{-1}$)
	perpendicular	parallel	
DE3T			2×10^{-7}
DE4T			5×10^{-5}
DE5T			9×10^{-4}
DH6T	5×10^{-7}	6×10^{-5}	4×10^{-2}
DH8T	5×10^{-7}	4×10^{-4}	1×10^{-2}

TABLE 4. Conductivity and field-effect mobility of oligothiophenes substituted with pendent alkyl chains. DD: didecyl, TD: tetradecyl

Oligomer	Conductivity (S/cm)	Mobility ($cm^2V^{-1}s^{-1}$)
DD6T	10^{-13}	$<10^{-7}$
TD12T	10^{-9}	5×10^{-6}

5. Crystal structure of sexithiophenes

We shall now focus on 6T and its derivatives. We have recently resolved the crystal structure of non substituted 6T. The unit cell, Figure 5, is monoclinic, and contains four molecules arranged in a herringbone close-packing [35].

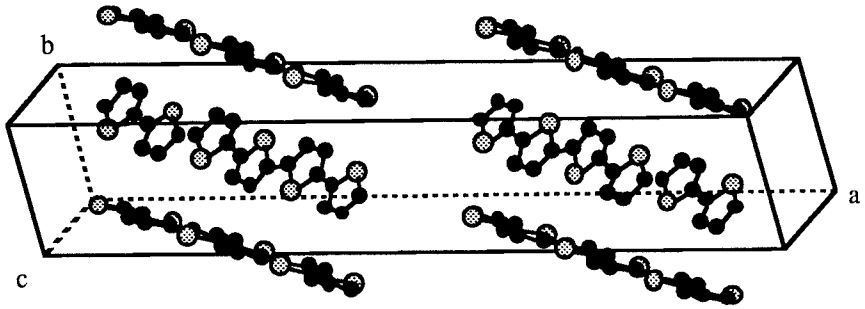


Figure 5. Crystal structure of sexithiophene

This crystal arrangement presents some remarkable features. Molecules are rigorously planar and strictly parallel to each other. Moreover, the π-orbital overlap between nearest neighbor molecules is maximum, which can account for that, unlike conjugated polymer, the charge transport is favored along the π-stack direction.

The crystal structure of sexithiophene substituted by pendent butyl groups has been resolved recently [36]. The main effect of the pendent group is to shift the nearest neighbor thiophene chains, which results in a dramatic decrease of the π-overlap. This would account for the very low conductivity and mobility of that compound.

End substituted oligothiophene present a two dimensional structure, where layers of thiophene chains alternate with alkyl layers. Up to the diethyl-quinquethiophene (DE5T), the unit cell has been claimed to be orthorombic [37]. In spite of great efforts, we have not been able to grow single crystals of dihexyl-sexithiophene (DH6T), in which both the thiophene and alkyl chains are longer. However, X-ray diffraction on evaporated films of DH6T showed a two dimensional structure extending over large distances, that would compare to that of a liquid crystal.

6. Charge transport in sexithiophene

Like for conjugated polymers, the field-effect mobility of oligothiophenes is thermally activated. In conjugated polymers, this is generally ascribed to a hopping transport. We have shown that in the case of sexithiophene, this could be explained in the frame of a

multiple thermal trapping and release model [38]. Data are consistent with an exponential distribution of traps, that compares to that found in a-Si:H. The traps are attributed to grain boundaries, and the higher mobility of the end-substituted DH6T can be ascribed to its particular liquid-crystal-like structure, with a very low density of grain boundaries in the direction parallel to the film. Our model can also account for the enhanced mobility of 6T vacuum evaporated on heated substrates, in which electron microscopy showed an increase of the crystal grain size [39].

The deposition on a heated substrate also results in a decrease of the conductivity, which can be attributed to an improvement of the purity of the film. An even lower conductivity was measured on single crystals. In that case, the mobility is close to 0.1 $cm^2V^{-1}s^{-1}$, and the doping level, estimated from the pinch-off potential, is 3×10^{14} cm^{-3}, which corresponds to a molar ration of 0.2 ppm. Table 5 compares the electrical parameters of polythiophene to that of sexithiophene in its different forms.

TABLE 5. Conductivity, mobility and doping level of polythiophene and different forms of sexithiophene

Material	Conductivity (S/cm)	Mobility ($cm^2V^{-1}s^{-1}$)	Doping level (cm^{-3})
Polythiophene	10^{-7} - 10^{-6}	10^{-5} - 10^{-4}	10^{17}- 10^{18}
6T (substrate at 25°C)	1.5×10^{-6}	0.003	1×10^{18}
6T (substrate at 280°C)	1.2×10^{-7}	0.025	5×10^{16}
6T (single crystal)	$<10^{-9}$	0.075	3×10^{14}
DH6T	6×10^{-5}	0.04	3×10^{17}

These results indicate that, in contrast to "amorphous" organic compounds, large on-off ratios can be obtained with sexithiophene.

7. Conclusions

Field-effect transistors have been made with thin films of a number of organic semiconductors, including conjugated polymers and oligomers. Two categories of behavior can be differentiated. First, in most of the conjugated polymers, and in a great number of "amorphous" molecular materials, conduction is governed by a hopping mechanism. This result in a field-effect mobility that depends on the doping level. Accordingly, high mobility is only obtained with materials that also present a high conductivity. FETs made with these materials present an inherently poor on-off ratio. The second category is that followed by the oligothiophenes. In these materials, the charge transport obeys a thermal trapping mechanism, and is hence only dependent on the density of traps, whereas the conductivity depends on the doping level. High mobility can thus be associated with low conductivity by using highly ordered and very pure materials. Ordering can be obtained by chemical means, as it is favored by substitution at ends of the molecule. Highly crystalline films can also be realized by adjusting the deposition parameters.

Practical applications of the organic FETs can be envisioned in fields where large areas are needed. Another interesting property of these devices is the possibility to make them on flexible substrates. We finally note that the fabrication of OFETs requires soft techniques and very low temperatures.

8. References

1. Pope, M. and Swenberg, C.E. (1982) *Electronic Processes in Organic Crystals*, Oxford University Press, New York.
2. Ghosh, A.K. and Feng, T. (1978) Merocyanine Organic Solar Cells., *J. Appl. Phys.* **49**, 5982-5989.
3. Tang, C.W. (1986) Two-Layer Organic Photovoltaic Cell., *Appl. Phys. Lett.* **48**, 183-186.
4. Hiramoto, M., Fujiwara, H. and Yokoyama, M. (1991) Three-Layered Organic Solar Cell with a Photoactive Interlayer of Codeposited Pigments, *Appl. Phys. Lett.* **58**, 1062-1064.
5. Antoniadis, H., Hsieh, B.R., Abkowitz, M.A., Jenekhe, S.A. and Stolka, M. (1994) Photovoltaic and Photoconductive Properties of Aluminum Poly(p-Phenylene Vinylene) Interfaces, *Synthet. Metal.* **62**, 265-271.
6. Basescu, N., Liu, Z.X., Moses, D., Heeger, A.J., Naarman, H. and Theophilou, N. (1987) High Electrical Conductivity in Doped Polyacetylene, *Nature* **327**, 403-405.
7. Burroughes, J.H., Bradley, D.C.C., Brown, A.R., Marks, R.N., McKay, K., Friend, R.H., Burns, P.N. and Holmes, R.B. (1990) Light-Emitting Diodes Based on Conjugated Polymers, *Nature* **341**, 539-541.
8. Braun, D. and Heeger, A.J. (1991) Visible Light Emission from Semiconducting Polymer Diodes, *Appl. Phys. Lett.* **58**, 1982-1984.
9. Suzuki, H., Meyer, H., Simmerer, J., Yang, J. and Haarer, D. (1993) Electroluminescent Devices Based on Poly(Methylphenylsilane), *Advan. Mater.* **5**, 743-746.
10. Tsumura, A., Koezuka, H. and Ando, Y. (1988) Polythiophene Field-Effect Transistor: its Characteristics and Operation Mechanism, *Synthet. Metal.* **25**, 11-23.
11. Burroughes, J.H., Jones, C.A. and Friend, R.H. (1988) Polymer Diodes and Transistors: New Semiconductor Device Physics, *Nature* **335**, 137-141.
12. Assadi, A., Svensson, C., Willander, M. and Inganäs, O. (1988) Field-Effect Mobility of Poly(3-hexylthiophene), *Appl. Phys. Lett.* **53**, 195-197.
13. Horowitz, G., Fichou, D., Peng, X.Z., Xu, Z.G. and Garnier, F. (1989) A Field-Effect Transistor Based on Conjugated Alpha-Sexithienyl, *Solid State Commun.* **72**, 381-384.
14. Paloheimo, J., Kuivalainen, P., Stubb, H., Vuorimaa, E. and Yli-Lahti, P. (1990) Molecular Field-Effect Transistors Using Conducting Polymer Langmuir-Blodgett Films, *Appl. Phys. Lett.* **56**, 1157-1159.

15. Akimichi, H., Waragai, K., Hotta, S., Kano, H. and Sakati, H. (1991) Field-Effect Transistors Using Alkyl Substituted Oligothiophenes, *Appl. Phys. Lett.* **58**, 1500-1502.
16. Xie, Z., Abdou, M.S.A., Lu, X., Deen, M.J. and Holdcroft, S. (1992) Electrical Characteristics and Photolytic Tuning of Poly(3-Hexylthiophene) Thin Film Metal Insulator Semiconductor Field-Effect Transistors (MISFETs), *Can. J. Phys.* **70**, 1171-1177.
17. Abdou, M.S.A., Lu, X.T., Xie, Z.W., Orfino, F., Deen, M.J. and Holdcroft, S. (1995) Nature of impurities in pi-conjugated polymers prepared by ferric chloride and their effect on the electrical properties of metal insulator semiconductor structures, *Chem. Mater.* **7**, 631-641.
18. Dodabalapur, A., Torsi, L. and Katz, H.E. (1995) Organic transistors: Two-dimensional transport and improved electrical characteristics, *Science* **268**, 270-271.
19. Weimer, P.K. (1962) The TFT - A New Thin-Film Transistor, *Proc. IRE* **50**, 1462-1469.
20. Peng, X.Z., Horowitz, G., Fichou, D. and Garnier, F. (1990) All-Organic Thin-Film Transistors Made of Alpha-Conjugated Sexithienyl Semiconducting and Various Polymeric Insulating Layers, *Appl. Phys. Lett.* **57**, 2013-2015.
21. Garnier, F., Hajlaoui, R., Yassar, A. and Srivastava, P. (1994) All-polymer field-effect transistor realized by printing techniques, *Science* **265**, 1684-1686.
22. Sze, S.M. (1981) *Physics of Semiconductor Devices*, John Wiley, New York.
23. Horowitz, G., Deloffre, F., Garnier, F., Hajlaoui, R., Hmyene, M. and Yassar, A. (1993) All-Organic Field-Effect Transistors Made of pi-Conjugated Oligomers and Polymeric Insulators, *Synthet. Metal.* **54**, 435-445.
24. Fuchigami, H., Tsumura, A. and Koezuka, H. (1993) Polythienylenevinylene Thin-Film Transistor with High Carrier Mobility, *Appl. Phys. Lett.* **63**, 1372-1374.
25. Bauerle, P., Fischer, T., Bidlingmeier, B., Stabel, A. and Rabe, J.P. (1995) Oligothiophenes - Yet longer? Synthesis, characterization, and scanning tunneling microscopy images of homologous, isomerically pure oligo(alkylthiophene)s, *Angew. Chem. Int. Ed.* **34**, 303-307.
26. Clarisse, C., Riou, M.T., Gauneau, M. and Le Contellec, M. (1988) Field-effect transistor with diphthalocyanine thin film, *Electron. Lett.* **24**, 674-675.
27. Guillaud, G., Al Sadoun, M., Maitrot, M., Simon, J. and Bouvet, M. (1990) Field-Effect Transistors Based on Intrinsic Molecular Semiconductors, *Chem. Phys. Lett.* **167**, 503-506.
28. Brown, A.R., Deleeuw, D.M., Lous, E.J. and Havinga, E.E. (1994) Organic n-Type Field-Effect Transistor, *Synthet. Metal.* **66**, 257-261.
29. Hoshimono, K., Fujimori, S., Fujita, S. and Fujita, S. (1993) Semiconductor-Like Carrier Conduction and Its Field-Effect Mobility in Metal-Doped C_{60} Thin Films, *Jpn. J. Appl. Phys. Pt 2* **32**, L1070-L1073.
30. Brown, A.R., Deleeuw, D.M., Havinga, E.E. and Pomp, A. (1994) A universal relation between conductivity and field-effect mobility in doped amorphous organic semiconductors, *Synthet. Metal.* **68**, 65-70.

31. Holland, E.R., Bloor, D., Monkman, A.P., Brown, A., Deleeuw, D., Bouman, M.M. and Meijer, E.W. (1994) Effects of Order and Disorder on Field-Effect Mobilities Measured on Conjugated Polymer Thin-Film Transistors, *J. Appl. Phys.* **75**, 7954-7958.
32. Pearson, C., Gibson, J.E., Moore, A.J., Bryce, M.R. and Petty, M.C. (1993) Field-Effect Transistor Based on Organometallic Langmuir-Blodgett Film, *Electron. Lett.* **29**, 1377-1378.
33. Garnier, F., Yassar, A., Hajlaoui, R., Horowitz, G., Deloffre, F., Servet, B., Ries, S. and Alnot, P. (1993) Molecular Engineering of Organic Semiconductors - Design of Self-Assembly Properties in Conjugated Thiophene Oligomers, *J. Am. Chem. Soc.* **115**, 8716-8721.
34. Delabouglise, D., Hmyene, M., Horowitz, G., Yassar, A. and Garnier, F. (1992) Electrochemical Coupling of Dialkylated Sexithiophene, *Advan. Mater.* **4**, 107-110.
35. Horowitz, G., Bachet, B., Yassar, A., Lang, P., Demanze, F., Fave, J.L. and Garnier, F. (1995) Growth and Characterization of Sexithiophene Single Crystals, *Chem. Mater.* in press.
36. Herrema, J.K., Wildeman, J., Vanbolhuis, F. and Hadziioannou, G. (1993) Synthesis and Crystal Structures of 2 Dialkyl-Substituted Sexithiophenes, *Synthet. Metal.* **60**, 239-248.
37. Hotta, S. and Waragai, K. (1991) Alkyl-Substituted Oligothiophenes - Crystallographic and Spectroscopic Studies of Neutral and Doped Forms, *J. Mater. Chem.* **1**, 835-842.
38. Horowitz, G., Hajlaoui, R. and Delannoy, P. (1995) Temperature dependence of the field-effect mobility of sexithiophene. Determination of the density of traps, *J. Phys. III France* **5**, 355-371.
39. Servet, B., Horowitz, G., Ries, S., Lagorsse, O., Alnot, P., Yassar, A., Deloffre, F., Srivastava, P., Hajlaoui, R., Lang, P. and Garnier, F. (1994) Polymorphism and charge transport in vacuum-evaporated sexithiophene films, *Chem. Mater.* **6**, 1809-1815.

MICROCAVITY EMITTERS AND DETECTORS

BEN G. STREETMAN, JOE C. CAMPBELL, DENNIS G. DEPPE
Microelectronics Research Center
The University of Texas at Austin
Austin, TX 78712 USA

Abstract

Modern crystal growth methods allow multilayer heterostructures to be incorporated in a variety of novel and useful devices. For example, the use of distributed Bragg reflectors (DBR's) with high reflectivity designed for a specific wavelength has led to microcavities for both light emitters and detectors. This has revolutionized the design of semiconductor lasers, which now have resonant cavities on the order of a single wavelength of light. Photodetectors also have been changed by incorporating DBR's to form microcavities for absorption. The resonant-cavity photodiode structure in effect decouples the quantum efficiency from the transit-time. It is also possible to introduce additional periodicities in the mirror design to achieve reflectivity at two (or four) separate wavelengths. These wavelength-selective mirrors should have a variety of applications in wavelength division multiplexing. Lasers and detectors employing resonant cavities on the wavelength scale will play an important role in a variety of future optoelectronic applications, and for optical interconnects. By using techniques such as selective oxidation of AlAs layers, it is possible to define cavities in the lateral plane as well as vertically between the DBR's. As a result of these advances in the design of microcavities, new approaches to low-dimensional confinement of photons are possible, in analogy to the study of electron confinement.

1. Introduction

The development of III-V epitaxial growth techniques such as molecular beam epitaxy (MBE) and metal-organic chemical vapor deposition (MOCVD) has made possible the design and reproducible growth of a broad range of novel optoelectronic and high speed electronic devices based on ultra-thin layers. High electron mobility transistors and heterojunction bipolar transistors offer substantial improvements over traditional devices [1]. Other applications of heterostructure growth provide new types of devices requiring ultrathin layers, such as quantum wells for lasers, resonant tunneling diodes, *etc.* In much of this work the thin layered structures are designed to confine carriers, or to realize thin potential barriers to take advantage of quantum tunneling effects. Among the advantages of high growth control are the ability to do modulation doping and delta doping, with control on the scale of a monolayer [2].

Semiconductor quantum wells (QW's), achievable through control of layer thickness with monolayer accuracy, have been a rich source of insight into

S. Luryi et al. (eds.), Future Trends in Microelectronics, 327–335.

semiconductor physics and have led to novel structures for electronic and photonic device applications. The interaction between photons and carriers in a quantum well, the optical transition, is governed by the fact that only discrete energy states are allowed for electrons and holes in the well. Not only do quantum wells provide transition energies different from the bulk material, but also population inversion is achieved at a lower threshold current in a QW laser. Confinement of carriers can be substantial in δ-doped quantum wells. For example, we have reported confinement of holes on a scale of about 5 Å in a 50Å QW δ-doped with Be [2]. Carrier confinement can lead to a number of interesting low-dimensional effects, including those related to quantum wires and quantum boxes. The quantum-confined Stark effect is an example of a useful effect (for light modulators) that can only be achieved with these modern growth techniques. In the usual quantum well structure, where electrons and holes are well-confined in the potential well region, transition energies occur near the energy gap of the well layer. For applications requiring higher transition energies, a number of quantum well structures have been proposed and synthesized, such as those using narrow wells, high-gap well layers, and superlattices in the well region. From the earliest days of MBE, there has been an interest in artificial periodicities available by growth of multilayer heterostructures. Both compositional and doping periodicities can lead to new "miniband" or "subband" conduction of electrons and holes. Superlattice quantum wells (SLQW's) can be used to achieve high-energy transitions, which can be varied over a few hundred meV using different AlAs and GaAs layer thicknesses in the SLQW's. It is also possible to use aperiodic layer thicknesses, including random-period superlattices, to further tailor the properties of the quantum wells [3].

2. Distributed Bragg Reflectors

An extension of the periodic growth of heterostructures is the ability to grow alternating layers of thicknesses corresponding to fractions of a wavelength of light (e.g., λ/4) in the material. It is therefore possible to simulate mirrors within the device by incorporating distributed Bragg reflectors (DBR's) with high reflectivity designed for a specific wavelength. Recently, mirrors have been made with additional periodicities to achieve the proper phase change and/or delay such that the DBR exhibits reflectivity at two separate wavelengths [4]. We have also achieved two- and four- wavelength mirrors in our lab [5]. These wavelength-selective mirrors should have a variety of applications in wavelength division multiplexing. Of particular interest is the use of DBR's to form microcavity structures in which the cavity length is adjustable on the scale of a wavelength of the light being used. This has revolutionized the design of semiconductor lasers, and has led to the use of microcavities in both lasers and detectors, as discussed below.

3. Photodiodes

Recently we have applied the Bragg reflectors available by MBE growth to take advantage of resonant absorption of photons in microcavities. Such resonant cavities can have enormous impact on traditional photodiodes.

The PIN photodiode is the most widely deployed photodetector for photonic applications. The light that enters the photodiode is attenuated exponentially with

distance into the absorbing layer, and photogenerated electrons and holes give rise to a photocurrent that is proportional to the incident intensity. There is a tradeoff between the responsivity and the bandwidth of the PIN structure, since to achieve high quantum efficiency a relatively thick absorption layer is required, which in turn requires a longer time to collect the photogenerated carriers. This is the origin of the transit time limit to the bandwidth.

We have demonstrated a novel resonant-cavity photodiode, shown in Fig. 1, that circumvents the quantum efficiency bandwidth tradeoff due to transit time effects [6]. This structure increases the absorption through multiple reflections between two parallel mirrors in a Fabry-Perot cavity whose length is typically one wavelength. The lower mirror is an integrated Bragg reflector consisting of alternating $\lambda/4$ epitaxial layers, having a reflectivity >99%. The top mirror is usually a high reflectivity dielectric stack that can be deposited after fabrication and initial characterization.

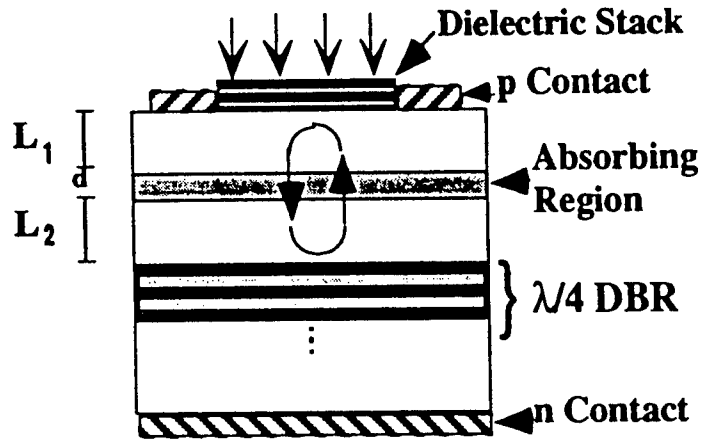


Figure 1. A resonant cavity PIN photodiode with distributed Bragg reflectors to define a cavity one wavelength long.

For resonant-cavity photodiodes having device areas ~50 μm^2, bandwidths greater than 100 GHz can be achieved without sacrificing responsivity. The photon buildup time in such a short cavity corresponds to a bandwidth in excess of ~10^{12} Hz and poses no limitation on the speed. This illustrates one advantage of the resonant-cavity approach, namely, that the quantum efficiency can be effectively decoupled from the transit-time [6]. While the improved bandwidth is accomplished at the expense of a narrower spectral response, almost all photonic systems operate within a very narrow wavelength range. In fact, this may be used to advantage for applications such as wavelength-division multiplexing in fiber optic systems. Resonant-cavity structures promise performance enhancements in several photodetector applications [7].

4. Resonant-Cavity Avalanche Photodiode with Separate Absorption and Multiplication

It is well known that the internal gain of avalanche photodiodes (APD's) can provide substantial improvements in signal-to-noise compared to PIN photodiodes. The gain is achieved through carrier multiplication, a consequence of impact ionization at high electric fields. Two of the crucial performance characteristics of APD's, the gain-bandwidth product and the excess noise arising from the random nature of the multiplication process, are determined primarily by the electron and hole ionization coefficients (α and β, respectively) or, more specifically, by k which is defined as the ratio of the ionization coefficients. For low noise and high gain-bandwidth products, a large difference in the ionization rates ($k << 1$) is desirable [8].

In a properly designed APD only the carrier with the highest ionization rate (either the electrons or holes, depending on the material) is injected into the high-field multiplication region. This can be accomplished with separate absorption and multiplication (SAM) structures. The SAM APD's have been widely deployed in long-wavelength, high-bit-rate optical transmission systems. As these transmission systems have progressed to higher and higher bit rates, however, the bandwidth of the SAM APD's has become a limitation. Therefore, the resonant-cavity structure has become an attractive alternative to achieve high speed without sacrificing quantum-efficiency.

Recently, we have successfully incorporated a SAM APD into a resonant-cavity structure [9]. A schematic cross section of the device is shown in Fig. 2. It has a 500 Å

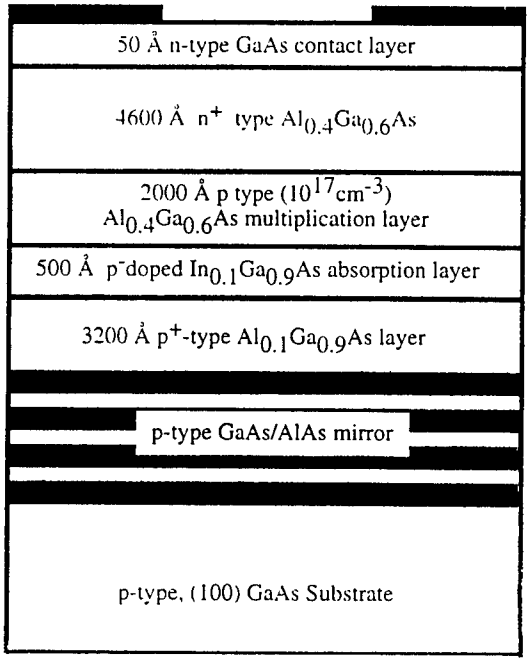


Fig. 2. A resonant cavity enhanced avalanche photodiode with separate absorption and multiplication regions [9].

$In_{0.1}Ga_{0.9}As$ absorption region and a 2000 Å-thick $Al_{0.4}Ga_{0.6}As$ multiplication region. The lower mirror is made up of twenty quarter-wavelength pairs of AlAs(755 Å)/GaAs(625 Å) and the top mirror is a quarter-wave-stack of $ZnSe/CaF_2$ pairs. Initial devices have produced very encouraging results: peak external quantum efficiency ~75%, the dark current at 90% of breakdown < 10 nA, and multiplication values in excess of 30. The quantum efficiency is much higher than the value of 4% that would have been achieved in a single pass without the resonant-cavity structure.

Two intriguing characteristics of these APD's are their extremely low operating voltage and low multiplication noise. Conventional APD's require bias voltages in the range 50 to 100 V. Our resonant-cavity APD's exhibit breakdown at < 14 V representing almost an order of magnitude reduction in power dissipation. The value of k for these devices is about 0.3, which is smaller than the value for the long-wavelength InP/InGaAs SAM APD's that are widely used for long-distance telecommunications (typically in the range $0.4 < k < 0.5$) [10]. A great deal of research is necessary in order to fully understand the physics of the multiplication process in these APD's and to achieve even lower noise. In lattice-mismatched systems, there may be a strain-induced enhancement in α/β for tensile strain in the quantum well and compressive strain in the barrier [11]. One might introduce strain into the resonant-cavity SAM APD's by utilizing a multiplication region of alternating layers of AlGaAs and AlInGaAs.

5. Vertical Cavity Surface Emitting Lasers

The vertical-cavity surface-emitting laser has recently gained much attention in the research literature. This type of structure represents a minimum volume laser, since the Fabry-Perot cavity length is on the order of the lasing wavelength, and thus it has the greatest potential of any semiconductor laser structure for ultra-low threshold current. The device's vertical geometry is tailor-made for large scale integration, since light is emitted normal to the epitaxial surface, and thus is also compatible with large area two-dimensional operation of phased arrays for high power operation. From a manufacturing standpoint, wafer scale testing is possible with the vertical-cavity laser. This is a major advantage over the traditional edge emitter, which must be cleaved and mounted for testing.

A common problem with VCSEL structures is that current is injected through the same region from which the light is emitted, unlike edge emitters in which current and light emission are along different directions. Thus, optimizing the top Bragg reflector mirror has in the past led to high series resistance for the current, and a resulting high bias voltage at the threshold for lasing. Current funneling may be achieved through the use of MBE regrowth over a patterned n-type current blocking layer placed within the p-type top mirror near the active region. However, these thin n-type layers pass a considerable amount of leakage current. In addressing this problem, we began by incorporating a layer of AlGaAs grown at a reduced temperature as a semi-insulating region to allow current funneling from the top contact to the active region of the laser [12]. This AlGaAs region is lattice-matched to the rest of the structure, and we have developed methods for selective etching and regrowth to provide the buried semi-insulating current funneling region in these devices. Post growth processing consists of a Cr/Au metallization and deposition of a quarter-wave $ZnSe/CaF_2$ top DBR.

Commercial VCSEL fabrication is based largely upon the process of proton implantation. The proton implantation is used to damage the epitaxial semiconductor

crystal below its surface, greatly decreasing its conductivity by trapping charge carriers, while leaving an electrically conducting path near the crystal surface. By performing the implantation into selected regions of the crystal, electrical current can be funneled into the VCSEL active region, thus exciting the electron-hole pair density necessary to achieve lasing. This proton implantation process is a carry over from an older edge-emitting laser diode technology, but one which is relatively cheap, well-characterized, and is adequate in realizing reasonable performance VCSEL's. In both the buried semi-insulating AlGaAs method and the proton bombardment, electrical isolation is achieved with little change in refractive index. Therefore, lateral optical confinement requires another approach.

6. Selective Oxidation of AlAs

In 1990 Holonyak and co-workers at the University of Illinois noted that a single crystal film of AlAs could be selectively converted to high quality Al_xO_y using "wet" oxidation at about 400°C [13]. In January of 1994 it was demonstrated by Deppe and co-workers at UT-Austin that the selective oxidation process can produce very low threshold VCSEL's [14,15]. The benefit of this process in VCSEL fabrication is two-fold: it serves as a buried insulator to tightly confine the injected current to a small area active region, and it provides a lateral index-guide for optical confinement.

Figure 3 (a) shows a schematic cross-section of a full-wave cavity layered structure, and Fig. 3 (b) shows a scanning electron microscope (SEM) image looking down on a selectively oxidized epitaxial heterostructure of AlAs-GaAs InGaAs [14]. In Fig. 3 (b) the Al_xO_y oxidation front has diffused laterally under a GaAs mesa (see Fig. 3 (a)) leaving a 4 μm square AlAs region which becomes the active part of a fully processed device. For the selectively oxidized VCSEL the major advantage of the half-wave cavity [16] as compared to the full-wave cavity [14] is the ability to place the Al_xO_y layer close to the light emitting quantum well region and the lasing mode intensity peak (at the center of the vertical-cavity). The half-wave cavity, therefore, is an important design feature for maximizing optical index confinement, and therefore minimizing threshold.

The selectively oxidized half-wave cavity VCSEL is unique in that it is an index-guided laser which requires no epitaxial regrowth step and yet provides strong optical confinement as well as excellent current confinement. Our research group was first to demonstrate the application of the native oxide process to VCSEL fabrication [14], and recently we have demonstrated that the use of a half-wave cavity spacer can lead to extremely low threshold currents [16]. Our recent work has used a combination of epitaxially grown AlAs/GaAs and post-growth deposited dielectric CaF_2/ZnSe Bragg reflectors. The result due to the optical and electrical confinement has been the lowest threshold VCSEL's yet realized, with room-temperature threshold currents under 100 μA and a record low threshold current of 59 μA at an optimized temperature of 250K [16]. Besides the more fundamental interest in studying ultra-small cavity lasers and exploring the limits of minimizing threshold for a semiconductor laser, it is likely that the important area of digital signal transmission will benefit in terms of switching speed by reducing the laser diode threshold current well into the sub-100 μA regime [17]. Present commercial VCSEL technology based on proton implantation confinement is limited to threshold drive currents in the 2-5 mA range.

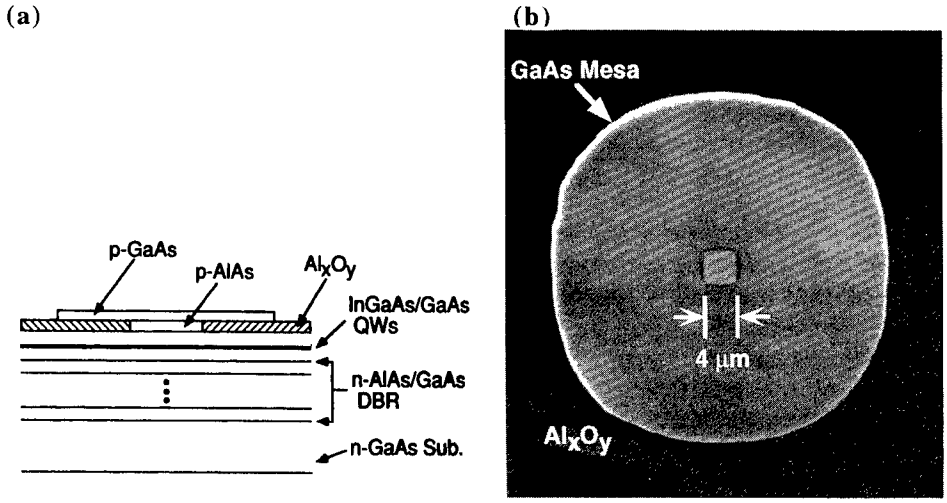


Figure 3. (a) Schematic cross section of the buried ring contact VCSEL structure showing the role of the lateral oxidation of the AlAs underneath the GaAs mesa defining the device active region [14]. (b) SEM photograph of a native-oxide defined 4 μm square AlAs region buried beneath a 30 μm diameter GaAs mesa. The oxide layer surrounding and beneath the GaAs mesa provides device isolation for current injection [14].

Since the initial demonstration in 1994, first reported as a late paper at the Conference on Lasers and Electro-Optics [15], the impact of selective oxidation on VCSEL research has been impressive [14-16, 18-20] and has led to several record results [16, 18, 19]. The native-oxide defined VCSEL has yielded the lowest threshold currents (sub-100 μA) yet achieved in a laser [16, 19], as well as the highest power conversion (wall-plug) efficiency of >50% [18]. For comparison, typical threshold currents of proton implanted VCSEL's are in the 2-5 mA range as mentioned above, with the highest reported wall-plug efficiency being ~20% [18]. An important demonstration was made recently by Sandia workers who showed device reliability of over 28 days of continual operation with little degradation for an all epitaxial structure [21], giving strong evidence that the Al_xO_y defined VCSEL can be made to operate reliably.

7. What Does the Future Hold?

Clearly, the use of microcavity devices employing DBR's will play an important role in the emitters and detectors of the future. Furthermore, we have enough evidence to believe that lateral confinement of light will likewise contribute significantly to vertical cavity laser structures. An intriguing question remains regarding what new physics will emerge from the very recent ability to "confine" light in three dimensions on the scale of a wavelength. The confinement vertically is quite good, and can be easily dimensioned to a wavelength or a fraction such as $\lambda/2$. In the lateral dimension, the recent use of selective oxidation allows us to define the cavity within a few wavelengths

and in the future it may be possible to choose lateral cavity dimensions with even more accuracy. Obviously, the refractive index change between the semiconductor and Al_XO_Y does not lead to complete confinement. Therefore, an analogy with the confinement of carriers is not accurate. On the other hand, this is an example of a new regime for experiments that did not exist before, and one expects surprises when such things happen.

8. Acknowledgments

This work was supported by the Joint Services Electronics Program under contract F49620-92-C0027, and by the Texas Advanced Research and Technology Programs.

9. Bibliography

I. Sze, S.M. (1990) *High-Speed Semiconductor Devices*, Wiley Interscience, New York.

2. Shih, Y.C. and Streetman, B.G. (1991) Modulation of carrier distributions in delta-doped quantum wells, *Appl. Phys. Lett.* **59**, 1344-1346.

3. Shih, Y.C., Sadra, K., and Streetman, B.G. (1994) Random-period superlattice quantum wells, *J. Vac. Sci. Technol.* B **12**, 1082-1985.

4. Lee, C.P., Tsai, C.M., and Tang, J.S. (1993) Dual-wavelength Bragg reflectors using GaAs/AlAs multilayers, *Electron. Lett.* **29**, 1980-81.

5. Anselm, K.A., Murtaza, S.S., Campbell, J.C., and Streetman, B.G. (1995) Four wavelength Bragg mirror using GaAs/AlAs, *Optics Lett.* **20**, 178-179.

6. Murtaza, S.S., *et al.* (1994) SiGe/Si resonant cavity photodetector, *IEEE Dev. Research Conf.*, Boulder, CO.

7. Murtaza, S.S., *et al.* (1995) High-efficiency, dual-wavelength, wafer-fused resonant-cavity photodetector operating at long wavelengths, *IEEE Photonics Technol. Lett.* **7**, 679-681.

8. McIntyre, R.J. (1966) Multiplication noise in uniform avalanche diodes, *IEEE Trans. Electron. Dev.* **13**, 164.

9. Murtaza, S.S., Anselm, K.A., Hu, C., Nie, H., Streetman, B.G., and Campbell, J.C. (1995) Resonant cavity enhanced (RCE) separate absorption and multiplication (SAM) avalanche photodetector (APD), to appear in *IEEE Photonics Technol. Lett.*

10. Campbell, J.C., *et al.* (1989) *J. Lightwave Tech.* **7**, 473.

11. Chandramouli, V., and Maziar, C.M. (1993) Monte Carlo analysis of band structure influence on impact ionization in InP, *Solid State Electron.* **36**, 285-290.

12. Rogers, T.J., Lei, C., Deppe, D.G., and Streetman, B.G. (1993) Low threshold

voltage cw vertical- cavity surface-emitting lasers, *Appl. Phys. Lett.* **62**, 2027-2029.

13. Dallessasse, J.M., *et al.* (1990) Hydrolization oxidation of AlGaAs-AlAs-GaAs quantum well heterostructures and superlattices, *Appl. Phys. Lett.* **57**, 2844-2846.

14. Huffaker, D.L., Deppe, D.G., Kumar, K., and Rogers, T.J. (1994) Native oxide defined ring contact for low threshold vertical-cavity lasers, *Appl. Phys. Lett.* **65**, 97-99.

15. Deppe, D.G., Huffaker, D.L., Lin, C.C., and Rogers, T.J. (1994) Nearly planar low threshold vertical-cavity surface-emitting lasers using high contrast mirrors and native oxide, *Conference on Lasers and Electro-Optics 1994 Technical Digest Series* **8**, CPD2-1/3-6/8, May 8-13, Anaheim, CA.

16. Huffaker, D.L., Deppe, D.G., and Shin, J. (1995) Threshold characteristics of planar and index-guided microcavity lasers, *Appl. Phys. Lett.* **67**, 4-6.

17. Cutrer, D.M., and Lau, K.Y. (1995) Ultralow power optical interconnect with zero-biased, ultralow threshold laser - how low a threshold is low enough, *IEEE Photonics Technol. Lett.* **7**, 4.

18. Lear, K.L., Choquette, K.D., Schneider, R.P., Kilcoyne, S.P., and. Geib, K.M. (1995) Selectively oxidized vertical-cavity surface emitting laser with 50% power conversion efficiency, *Electron. Lett.* **31**, 208.

19. Hayashi, Y., *et al.* (1995) A record low threshold index-guided InGaAs/GaAlAs vertical-cavity surface-emitting laser with a native oxide confinement structure, *Electron. Lett.* **31**, 560.

20. Huffaker, D.L., Shin, J.L., Deng, H., Lin, C.C., Deppe, D.G., and Streetman, B.G. (1994) Improved mode stability in low threshold single quantum well native-oxide defined vertical-cavity lasers, *Appl. Phys. Lett.* **65**, 2642-2644.

21. Choquette, K.D., *et al.* (1995) Cavity characteristics of selectively oxidized vertical-cavity lasers, *Quantum Optoelectronics 1995 Topical Digest* **14**, 191.

Optical Amplification, Lasing and Wavelength Division Multiplexing Integrated in Glass Waveguides

R. L. HYDE, D. BARBIER, A. KEVORKIAN
GeeO, 46 rue Félix Viallet, F-38031 Grenoble Cedex, France.

J-M. P. DELAVAUX
AT&T Bell Laboratories, 9999 Hamilton Boulevard., Breinigsville, PA 18031, USA.

J. BISMUTH, A. OTHONOS, M. SWEENY, J.M. XU
Ontario Laser and Lightwave Research Center, University of Toronto, 10 King's College Road, Toronto, ONT M5S 1A4, Canada.

1. Introduction

Micro-electronic components became feasible with the invention of the semi-conductor devices which superceded the electronic valve in the late 1950s. At the same time the first coherent source of optical radiation was invented. It is remarkable that up till that time the optical physicist had worked with incoherent radiant sources which would be regarded by the electronic community as a noise generator. Today the optical and electronic communities work together thanks to the laser and optical amplifier.

Research in optical integration dates back to the mid-1960s, but a major catalyst for this activity was the re-invention of the optical fibre amplifier in the mid-1980s, and the semi-conductor pump laser, both quickly adopted by the telecommunications industry. Optical integration and micro-optical components are now in development and pre-production stages.

In this paper, which is divided in two main parts, we review our work on optical amplification, lasing and WDM devices integratable in doped glass waveguides. The first part addresses integrated lasers and amplifiers implemented in rare earth doped phosphate glass substrates. In the second part, a new type of WDM devices is investigated both theoretically and experimentally with a first WDM implementation in germanium doped glass waveguide and successful inscription of gratings in Er-doped phosphate glass.

It is evident that these two components are essential for a fully integrated optical device. The active and passive demonstrators described in each part of this paper enables us to proceed to full integration of optical devices for telecommunication, instrumentation and sensor applications (Fig. 1).

2. Optical Amplifiers & Lasers in Rare Earth Doped Phosphate Glass

An integrated laser system (as in an optical fibre laser system) obtains a major advantage from the significant spatial overlap of the pump radiation mode with that of the output radiation mode, as they propagate in a waveguide. Consequently two factors

S. Luryi et al. (eds.), Future Trends in Microelectronics, 337–351.

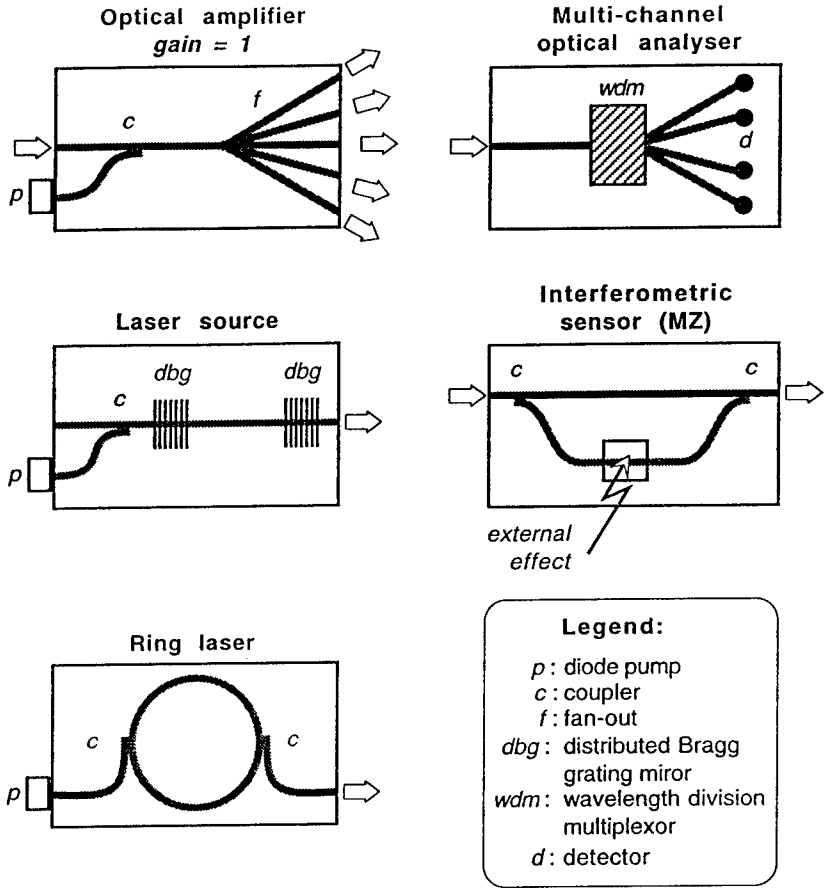


Figure 1: Basics elements to be associated for the realization of fully integrated devices for multi-channel telecommunication, instrumentation and sensor applications.

are in play, the spectroscopy of the active ions in the host medium, in this case the lanthanides or rare-earths in a glass host, and the opto-geometrical properties of the waveguide, i.e., its numerical aperture and modal diameters. Our goal was therefore, to make a low-loss micro-waveguide in a glass host containing a small percentage of rare-earth, whilst maintaining the desired spectroscopic properties. To this end we have spent several years in finding the correct glass host and the optimum ion-exchange conditions.

2.1 THE ION-EXCHANGE TECHNOLOGY

The ion exchange process itself consists of immersing a glass wafer, the desired form of the wave-guide suitably defined by a photo-lithography, into a molten bath of an

alkaline salt. The unprotected part of the glass is then subject to a migration of ions, for example Na ↔ K or Ag, and a channel of higher optical refractive index (≈ 1 %) is produced. A second stage effectively buries the guid
at a depth of a few micrometres below the surface to reduc
losses. This process was pioneered at the Institut National Polytechnique de Grenoble [1] and in Japan [1a].

We have found that the ion exchange process does not change the spectroscopic properties from that of the bulk material to any significant degree.

2.2 THE RARE-EARTH DOPANTS

We have chosen two rare-earth laser systems, neodymium, and erbium co-doped with ytterbium.

Neodymium has a strong transition at 1055 nm which terminates at a level above the ground state. The erbium transition at 1530 nm terminates at the ground state and therefore has an absorption band which must be overcome before net gain is reached.

On the other hand the lifetime of the excited metastable state is much longer in erbium making it a strong candidate for stimulated emission and in addition the erbium transition lies in the 1530 nm band, an optimum propagation window in optical fibre and therefore widely used in the telecommunications industry.

The absorption bands vary only slightly with different glass hosts, but more significantly, the radiative time constants can vary considerably depending on the phonon energy spectrum of the glass host. This is more advantageous in some glasses and we have chosen to use a phosphate oxide glass as distinct from a silicate glass.

A necessary condition for exciting the active ion is a pump absorption band which corresponds to the wavelength of an available powerful semi-conductor laser diode. In neodymium there is a strong absorption at 800 nm which we exploit. In erbium the preferred pump band is at 980 nm but the absorption is not strong so we have co-doped it with ytterbium, which has a stronger band in this region. In this co-doped system the ytterbium ion is excited (pumped) and its energy is transferred to the erbium ion. A shorter waveguide will then be the optimum.

2.3 MODELLING

The complex marriage between the spectroscopic properties and the waveguide geometry may be modelled to some degree and eventually one can calculate the optima for the performance of a device. We can also model second order parasitic phenomena and thereby seek to improve the concentrations of the basic materials.

Our mathematical model consists of solving the coupled differential equations of propagation, having first determined the steady state carrier densities at each significant energy level by solving the rate equations. As the optical powers in the guide are also radially dependent, there is an additional requirement to solve an intensity dependent spatial integral.

A comparison of the model and the experimental results has shown that second order phenomena, probably due to co-operative energy transfer between dopant ions, are important and that significant improvement in performance may be expected by optimising the doping and the geometry of the amplifiers and lasers.

2.4 OPTICAL AMPLIFICATION AND LASING AT 1054 nm [2]

We have made waveguides by ion exchange in a 3wt-% neodymium-doped phosphate glass. A guide was placed into a resonant cavity and when pumped at 797 nm oscillation occured at the wavelength 1054 nm. The threshold for lasing is 8.2 mW and the slope efficiency is ≈ 12.7%. Using this laser as a source we characterized the optical amplification in other waveguides.

2.4.1 *The neodymium laser*

The laser was made using a 38 mm long waveguide. Dielectric mirrors were butted at the two polished end faces of the glass chip and maintained by capillarity with an index matching liquid. The reflectivity of the mirrors is 71% at 1054 nm and 5% at 800 nm. The pump radiation was focused into the waveguide core through a microscope objective. Laser action was observed for an absorbed pump power of 8.2 mW at 797 nm. At this threshold, lasing occurs in our structure as a single band at 1053.8 nm. With increased pump power a second band appears at 1056 nm. The measured width of these bands is ≈ 1 nm, limited by the resolution of the monochromator.

2.4.2 *The neodymium amplifier*

Using the above laser as a signal source we characterized the optical amplification at 1054 nm in 41.5 mm long waveguides. The pump power, at a wavelength of 797 nm, and a signal were injected into the core waveguide via a duplexor. The theoretical coupling losses between this fibre and the integrated waveguide are 0.6 dB. We have measured the gain as a function of the injected pump power. The net gain is 7.1 dB for 52 mW of absorbed pump power.

We have developed a model which takes into account the up-conversion due to the high doping concentration. The rate equation of the metastable level population of neodymium ions may be written as follows :

$$\frac{dN_m}{dt} = R_{fp}.N_f - A_{mf}.N_m - W_{mf}.N_m - C.N_m^2 \qquad (1)$$

where $R_{fp}.N_f$ is the absorption term of the ions in their fundamental level, $A_{mf}.N_m$ and $W_{mf}.N_m$ are respectively the spontaneous and stimulated emissions from the metastable level. C is an up-conversion coefficient which is independent of the rare-earth concentration.

For $C \approx 10^{-23}$ m^3s^{-1}, we obtain a reasonable agreement between the experimental and theoretical curves. On eliminating up-conversion ($C = 0$) the model predicts a gain of 15 dB for the same pump power.

2.5 AMPLIFICATION IN ERBIUM:YTTERBIUM DOPED MICROGUIDES [3, 3a]

We have adapted our ion-exchange process, using silver ions, to our phosphate glass doped with 2% by weight of erbium and 4% by weight of ytterbium. This gives us the possibility of making short 44 mm microguides, single mode at both pump and signal wavelengths, 980 nm and 1540 nm, respectively. The average mode diameter (@ 1/e) of

these guides for the two wavelengths are respectively 4.7 μm at 980 nm and 7.5 μm at 1540 nm. These guides are buried 3.5 μm below the surface to avoid surface losses and enable high coupling efficiency with fibres of up to 89 %. We have measured the net gain obtained from these structures operating as travelling wave amplifiers. The small signal net gain at 1537 nm for a pump power of 100 mW was 11 dB for a double pass and 6 dB for a single pass. Large signal tests indicated that the -3 dB saturation point occured at an output power of 10 dBm.

An optimisation of the waveguide diameter and the pump wavelength, together with an optimisation of the rare-earth concentration, will allow us to make efficient optical amplifiers on compact glass substrate of less than 2.5 cm^2.

2:6 TUNABILITY OF Er:Yb INTEGRATED OPTICAL LASERS [4]

2.6.1 *Introduction*

We report here our first results of an integrated optical laser made by ion-exchange in Er:Yb phosphate glass. The structures show good slope efficiencies in double pump configurations and tunability when used with an external fibre bragg reflector.

2.6.2 *Gain Experiments and Results*

The lasing characteristics of the waveguides were tested with the double-pump configuration. A wide band dielectric mirror centered around 1540 nm was glued onto the connector of the output port of a wavelength division multiplexor. The reflectivity of this mirror was 98.4 % over 50 nm. For the output reflector we used fibre bragg

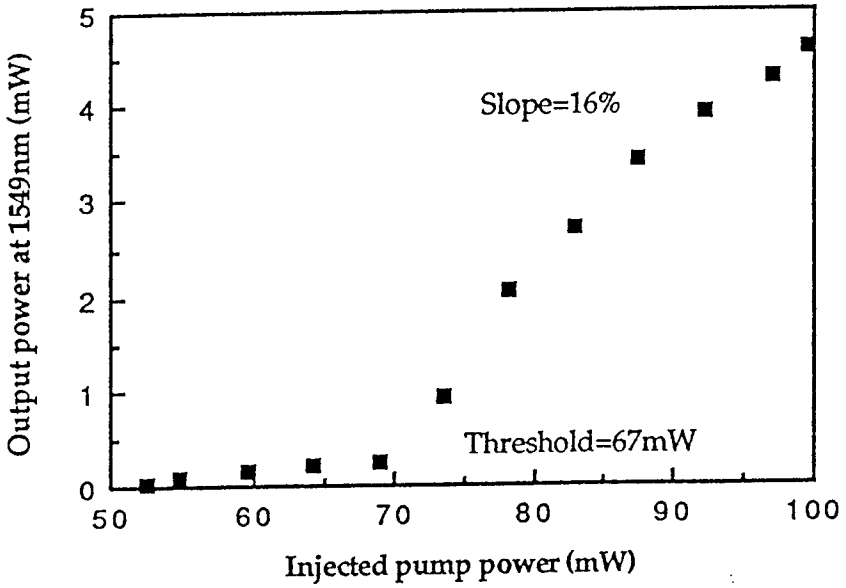


Figure 2: Output characteristics of the integrated optical laser made by ion-exchange in Er:Yb phosphate glass.

reflectors spliced between a second wavelength division multiplexor and the output face of the waveguide. Two different bragg reflectors were used. The first one has a reflectivity of 20 % at 1549 nm and the second has a reflectivity of 78 % at 1534 nm, both of them have a band pass of 0.25 nm.

We measured output powers of this laser configuration with a photodetector placed at the output port. We observed that a laser operation at 1549 nm is obtained with a slope efficiency of 16 % above a threshold of 67 mW of injected pump power (see Figure 2).

We have also analysed the tunability of our laser by dilating the fibre bragg reflector with a heater. When heated, from 25 to 450°C, the grating shows a shift in its center wavelength towards a higher wavelength without presenting any change in its reflectivity. We have shown a tunability of 4 nm without major change in output power. On the other hand with a second bragg reflector the tunability was shown to be greater than 7 nm, but with an increase in output power of more than a factor 2 between emission at 1534 nm and emission at 1542 nm.

3. Superimposed Bragg Gratings WDMs in Doped Glass Waveguides

Having discussed the signal generation and amplification part, we now move onto the signal retreving part by WDM technique which is particularly important for enhancing the transmission capacity of fiber optic transmission lines. Specifically, we present our work on new WDM devices in doped glass.

The ideal characteristics of a WDM device include a high wavelength selectivity, a large channel fan-out and a low loss. In the meantime, a base technology that allows good fiber connection (or low insertion loss) and is compatible with the other elements of the transmission line, i.e., laser sources, modulators, amplifiers and photodetectors is also an important consideration. In attempts to achieve some of these ideal characteristics, various configurations of integrated WDMs have been designed and demonstrated on various technologies. The curve diffraction grating WDM, the Mach-Zehnder WDM and the arrayed-waveguide WDM are among the most studied.

The curve diffraction grating (or echelle grating) WDM [5] is an integrated copy of conventional bulk-type spectrometer. It is well suited for numerous wavelength channels with relatively large wavelength spacing (a few nm). Its main weak point is the high insertion loss, typically of about –15 dB. The Mach-Zehnder WDM [6] uses the wavelength selective coupling property of a 2-wave interferometer. It is a good candidate for very small channel spacing of a few Å. Its drawback is the maximum fanout of only two channels. To obtain larger fanout, it is then necessary to cascade them which increases the size and the losses of the device. The arrayed-waveguide WDM [7] is in fact a hybrid of the two previous kinds of WDMs. The diffraction grating is replaced by the array of channel waveguides that acts like a multi-wave Mach-Zehnder interferometer.

In search of an alternative WDM device technology that is relatively simple, compact, high performance and is potentially implementable and integratable with the amplification and lasing elements discussed in the previous sections, we have theoretically and experimentally explored a totally different type of WDM concept—superimposed gratings.

The first successful implementation of multiple sets of gratings was demonstrated

in polymer waveguides by Wang *et al.* [8]. More reports of further experimental successes, also in polymer, have come out subsequently from the same group [9]. Although it is no secret that many are rather skeptical of the use of polymer in fiber-optical communications primarily because of its poor tolerance to high temperature environment which is viewed as inevitable in most processing steps required in the OEIC integration and packaging etc, these results, as proof-of-concept, are rather encouraging and important. It is also true that the large wavelength and angular spacings between adjacent channels in these reported implementations imply the absence of inter-grating coupling effects, thus, a likely neglect of such considerations in the designs. The inclusion of inter-grating coupling effects in design is, however, necessary only when we want to enter the very dense WDM domain, as we have found.

3.1 PHYSICAL CONCEPT OF SUPERGRATING WDM

The basic concept of a supergrating WDM with multiple sets of gratings superimposed on a planar waveguide is not difficult to understand. The base structure comprises a planar waveguide, made of dielectric or semiconductor films, and multiple sets of Bragg gratings photo-inscribed or etched into a portion of the guide. Each set of gratings is designed to diffract one wavelength in one direction for a common input angle. To carry out the actual device design, we should answer several questions pertinent to the superposition of multiple gratings.

Would the light of a particular wavelength get bounced back and forth by all the gratings, "see" the combined effect and come out in a way other than that is expected of a single set of gratings? If yes, what would be the combined effects on the diffraction direction? efficiency? and wavelength selectivity? It is not a trivial task to answer these questions because of the possible couplings between the gratings. To study this coupling between superimposed gratings we have developed a theoretical (coupled-mode) model [10-12]. With this model, we were able to show that there indeed could be significant couplings between gratings, that the effects of the couplings could greatly affect all aspects of the diffraction characteristics, and that it is possible to find WDM configurations for which the gratings are quasi-decoupled. In regimes of very weak coupling, the effects of the other gratings, co-existing with the grating designed to diffract a particular wavelength, are that of largely non-coherent scatterings off distributed "centers" of small index perturbations. Such effect manifests itself mainly as a rise in the background "noise" level, namely, a background loss to the efficiency of diffraction of a given light. In these weak coupling regimes, the task of designing a supergrating WDM device is greatly simplified –one can essentially design each set of grating as if the other gratings were absent except for an overall efficiency loss. Of course, this is limited to relatively large wavelength and angular spacings (i.e., a relatively small number of channels).

3.2 MODELLING

The Bragg and near Bragg angle of incidence, as well as the multiple-scattering in successively recorded volume holographic gratings have been theoretically investigated [13]. In most of the cases the coupled mode theory was used. Although the results obtained are useful for us, these models are not suitable for the case of multiple gratings

in planar waveguides, particularly for WDM applications. In the case of multiple volume holographic gratings for data storage applications, the light beams used to write and read-out the gratings are the same, they travel perpendicularly to the recording media and are at a fixed wavelength (the effect of multi-gratings on the wavelength selectivity of each of them was not studied), whereas in our case the gratings are read-out by a multi-wavelength beam travelling within the guide itself. By extending the treatment for the single grating in guiding structures [14] to the cases of multiple gratings and by including the couplings, we developed a model for superimposed gratings in planar waveguide [10-12]. Further improvements to the model have been made and will be presented in detail elsewhere.

The model allows us to compute the wavelength and angular selectivities and diffraction efficiencies of multiple (and possibly closely spaced) wavelengths by multiple gratings either inscribed in the volume or etched on the surface of the guide. The gratings are characterized by their orientation, their periodicity and their strength which is calculated taking into account the properties of the guided propagation (polarization and mode confinement). Mathematically, the coupled differential equations of propagation that describe the energy exchange between the input beam and the diffracted ones are numerically solved over the interaction (grating window) area.

3.3 WDM DESIGN

3.3.1 *Conditions for quasi-decoupled gratings*

With the aforementioned model, we are able to quantify different regimes of coupling between superimposed gratings. Generally speaking, the smaller the wavelength separation or the angular separation between channels, the stronger the coupling. In the case of strong coupling, superimposed gratings show very large deviations from the responses expected of individual sets of gratings. In this case it is difficult –if not impossible– to construct a conventional WDM. It is interesting, however, to explore the new opportunities opened up by strongly-coupled gratings to manipulate (or engineer) pass-bands and/or stop-bands. In the weak coupling case, one can essentially decouple each grating from the others. The criterion of quasi-decoupling, we found, can be stated as:

$$\Delta\lambda_d \geq 1.5\ \Delta\lambda_0 \tag{2}$$

where $\Delta\lambda_d$ is the wavelength separation between the adjacent channels and $\Delta\lambda_0$ is the total width at the first two zeros of the isolated grating spectrum, (i.e. calculated as if the grating was the only one present in the guide. See inset of Figure 3).

In the quasi-decoupled regime, the analysis (and design) for a N-channel WDM is relatively straightforward. For example, from [15] we can extract the expression for $\Delta\lambda_0$ as function of the key structure and operation parameters. Assuming a perfectly collimated input beam, we have:

$$\frac{\Delta\lambda_0}{\lambda_d} = A\,\frac{\cos(\theta_d)}{1 - \cos(\theta_d - \theta_i)} \tag{3a}$$

$$A = \sqrt{3}\,\frac{\lambda_d}{n_e\,Lc} \tag{3b}$$

with

where Lc is the so-called "coupling length" of the (single) grating, n_e the guide effective index, θ_i the input beam incidence angle, λ_d the diffracted wavelength and θ_d the angle of diffraction. This equation shows that grating selectivity depends on the angle of diffraction. As shown in Figure 3, it increases ($\Delta\lambda_0$ diminishes) when θ_d increases.

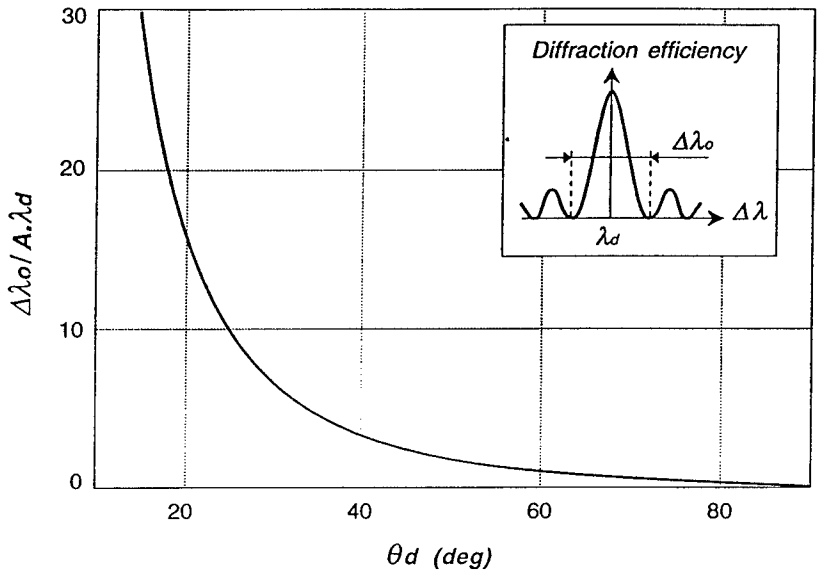


Figure 3: Wavelength selectivity of a single grating versus diffraction angle ($\theta_i = 0$).

Now, if we consider the design of a WDM with evenly spaced channels of the quantity $\Delta\lambda_{WDM}$ one can see there is a minimal diffraction angle under which the gratings will be coupled and will present a distorted spectrum. This minimal diffraction angle, θ_{dmin}, is so that $\Delta\lambda_0(\theta_{dmin}) = \Delta\lambda_{WDM}$. For $\theta_i = 0$, which is the most convenient configuration, one finds:

$$\theta_{dmin} = \cos^{-1}\left(\frac{1}{1 + 1.5\,A\,\lambda_d/\Delta\lambda_{WDM}}\right) \tag{4}$$

For a central wavelength λ_d, a channel spacing $\Delta\lambda_{WDM}$ and guiding structure given by the effective index, θ_{dmin} will depend on the Lc of the grating. The longer the grating, the smaller the minimal diffraction angle. For example for λ_d = 1.3 μm, $\Delta\lambda_{WDM}$ = 2 nm and n_e = 1.5 we calculate $\theta_{dmin} \approx$ 66, 39 and 29 deg for Lc = 1, 5 and 10 mm (see Figure 4).

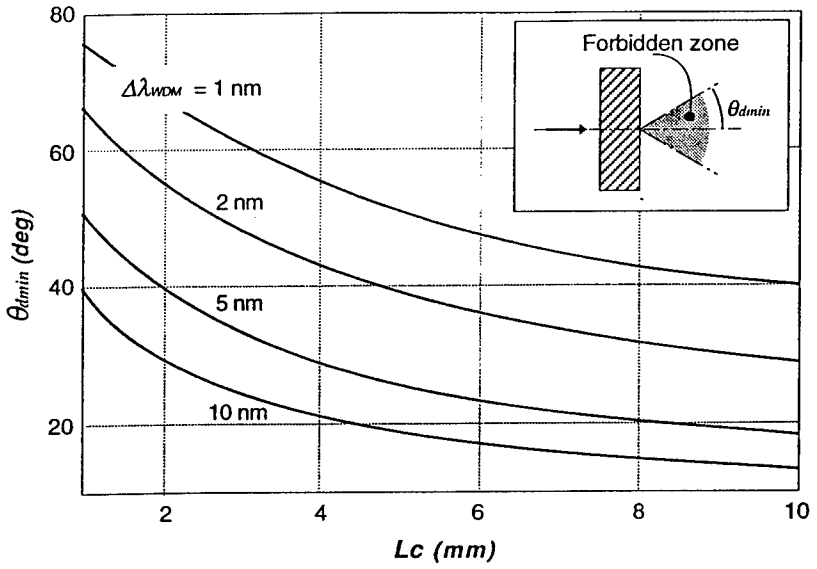


Figure 4: Minimal diffraction angle versus grating length for different values of WDM wavelength spacing. Calculation parameters: $\theta_i = 0$, $\lambda_d = 1.3$ μm and $n_e = 1.5$.

These considerations about the single grating wavelength selectivity show that both angular ranges $[\theta_{dmin}, \pi/2]$ and $[-\theta_{dmin}, -\pi/2]$ can be used for a superimposed gratings WDM to double the fanout capacity.

3.3.2 *Conditions for high diffraction efficiencies*

The first condition for a maximal diffraction efficiency is that the width of the grating window has to be equal to the coupling length Lc which depends on both the angle of diffraction and the grating strength κ. For $\theta_i = 0$, we have:

$$Lc = \frac{\pi}{2} \frac{\sqrt{\cos(\theta_d)}}{\kappa} \tag{5}$$

For example when θ_d varies in the range 30-80 deg, $\sqrt{\cos(\theta_d)}$ varies within a factor of 2. Then if we want to use the entire window of diffraction angles, the strength or the length of each grating will have to be adjusted independently to match Lc.

The second condition on which the effective diffraction efficiency depends is related to the overlap between the divergence of the input beam and the grating angular selectivity. Only the part of the input beam inside the cone of acceptance of the grating will be diffracted. When limited by diffraction the divergence of the input beam is typically of a few milliradians, depending on its width and the wavelength. On the other hand the grating angular selectivity is given by (see Figure 5):

$$\Delta\theta_{i0} = A \frac{\cos(\theta_d)}{\sin(\theta_d - \theta_i)} \tag{6}$$

For the numerical values used previously and for a beam width of 1 or 2 mm, we find that the angular selectivity is one order of magnitude smaller than the input beam divergence, that is an equivalent loss of about 10 dB. One way to limit the input beam divergence, thus reduce the loss, is to enlarge its width by inserting a large angle taper between the input channel waveguide and the grating area. Another way is to not use a collimated beam but to focus it on the grating area instead. In such a configuration one takes advantage of the fact that at the focus point a gaussian beam is a plane wave. The beam width and the focal length used will then have to be carefully calculated.

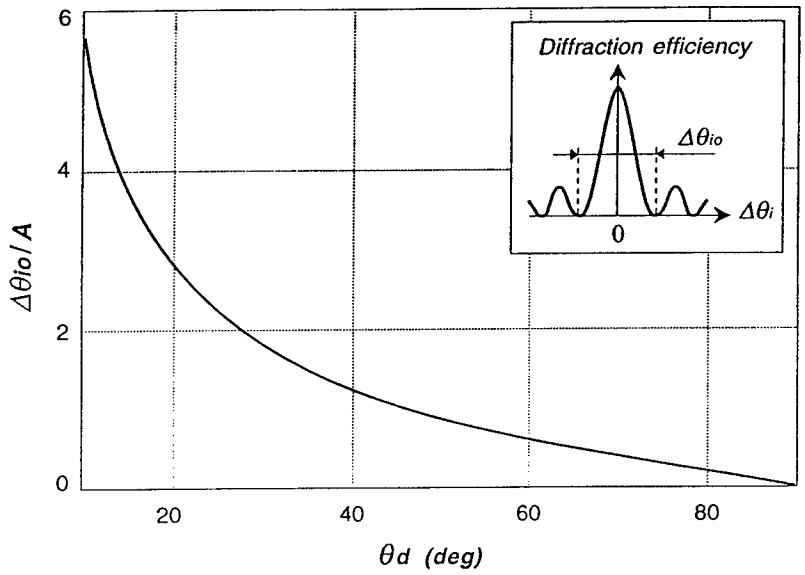


Figure 5: Angular selectivity of a single grating versus diffraction angle. Calculation for $\theta_i = 0$.

3.3.3 *Estimation of the minimal angular spacing between channels*

One way to determine the minimal angular spacing, $\Delta\theta_{dmin}$, is to consider directly the superimposed grating case. In [12], we calculated the effect of the inter-grating coupling that can eventually occur for very close diffraction angles. By doing this we found the minimal angle spacing between the channels, above which the coupling can be neglected, is in the order of a few milliradians.

From this, the angular channel density is found of a few channels per degree. This means that over the entire window of possible diffraction angles, the maximal number of channels with which a superimposed grating WDM can deal is a few hundreds. Being an order of magnitude smaller than the estimates previously made in the literature that

ignored the coupling effects, this is still a huge number by the present standard of WDM technology, suggesting a great potential of this new type of device. In practice, the number of channels will therefore depend mainly on the technology used to implement the designs, i.e., the material and the type of grating (corrugated or index gratings). In the case of corrugated gratings the maximal number of gratings will depend on the partial etching depth and the thickness of the structure as well as the lithographic resolution, whereas for index gratings it will depend on the available range of index change.

3.3.4 *The possible technologies for implementation*

The option of corrugated gratings on semiconductor waveguides is very attractive as it allows the integration of both the laser sources or detectors and the WDM on the same semiconductor substrate. But, direct superposition of many sets of gratings in semiconductors can be technically very difficult. Using index gratings inscribed in polymer deposited on semiconductor substrate is also attractive because of the large range of available index change, but the material instability, particularly thermal, can be a serious problem. Another very promising material is the photosensitive germanosilicate glass [16]. This material can be compatible with rare earth doped glasses developed for amplifiers and lasers described in the first part of this paper. It can also be deposited or grown directly on silicon substrate that has already shown very good compatibility with optical fiber connection and allows the integration with photodetectors as well. One can also consider implementation on photorefractive materials such as BSO, $BaTiO_2$ or $LiNbO_3$ for dynamic multi-grating applications.

3.4 FIRST EXPERIMENTAL RESULT ON Ge:SiO_2/SiO_2/Si WAVEGUIDE

In this section we report for the first time the realization of a 4-channel superimposed gratings WDM on photosensitive Ge-doped silica planar waveguide on silicon. The guiding structure is shown on the inset in Figure 4. Only the Ge doped guiding layer is photosensitive at 250 nm wavelength. The implementation process starts first with the high pressure hydrogenation of the sample (1500 PSI at room temperature for a couple of days) to increase its photosensitivity. Then the gratings are successively written with a KrF excimer laser holographic writing set-up similar to the one described in [17]. The four superimposed gratings have been designed to diffract λ_i = 830, 840, 850 and 860 nm at θ_{di} = +32, -26, -30 and +28 deg respectively. The choice of the diffracted wavelengths has been dictated by the range of tunability of the laser used for the characterization. Of course, the design can readily be transposed to the useful telecommunication wavelength windows at 1.3 μm or 1.5 μm. The gratings common length is 5 mm. They have been exposed for 15 mn with a ~1.8 cm^2 laser beam of ~100 mJ/pulse at 30 Hz, i.e., a total energy exposure of ~25 W.mn/cm^2. This exposure condition has been determined after a few writing and testing processes in an attempt to optimize the diffraction efficiencies.

Once the device was fabricated, we used two different methods to characterize it. The first one consists of the measurement of the WDM device characteristics. The collimated beam of a Ti/Sapphire tunable laser was coupled into the TE0 guided mode with a prism and the measurements were taken from the output edge of the waveguide. The diffraction efficiency vs. input wavelength was then recorded at each θ_{di}. Figure 6

shows the results of these measurements. The diffraction efficiencies vary from 62% to 21% while the FWHMs are around 3 nm, about 6 times broader than the theoretical values given by (2), and the cross-talks are between –10 and –15 dB.

The broader than expected FWHMs prompted us to perform a second kind of experiment on the fabricated samples to inspect the uniformity of the index modulation profiles. By illuminating the sample with a coherent visible beam from the top (non-guided light), we generated scattered beams, corresponding to the different orders of the Raman-Nath diffraction through the thin layer of the grating (6 μm guiding layer thickness). Because the intensity of each of them is proportional to the local index modulation of the grating the use of a large diameter incident beam allows the observation of the index modulation profile over the whole grating area. As suspected, highly non-uniform gratings have been observed in the samples fabricated. We attribute

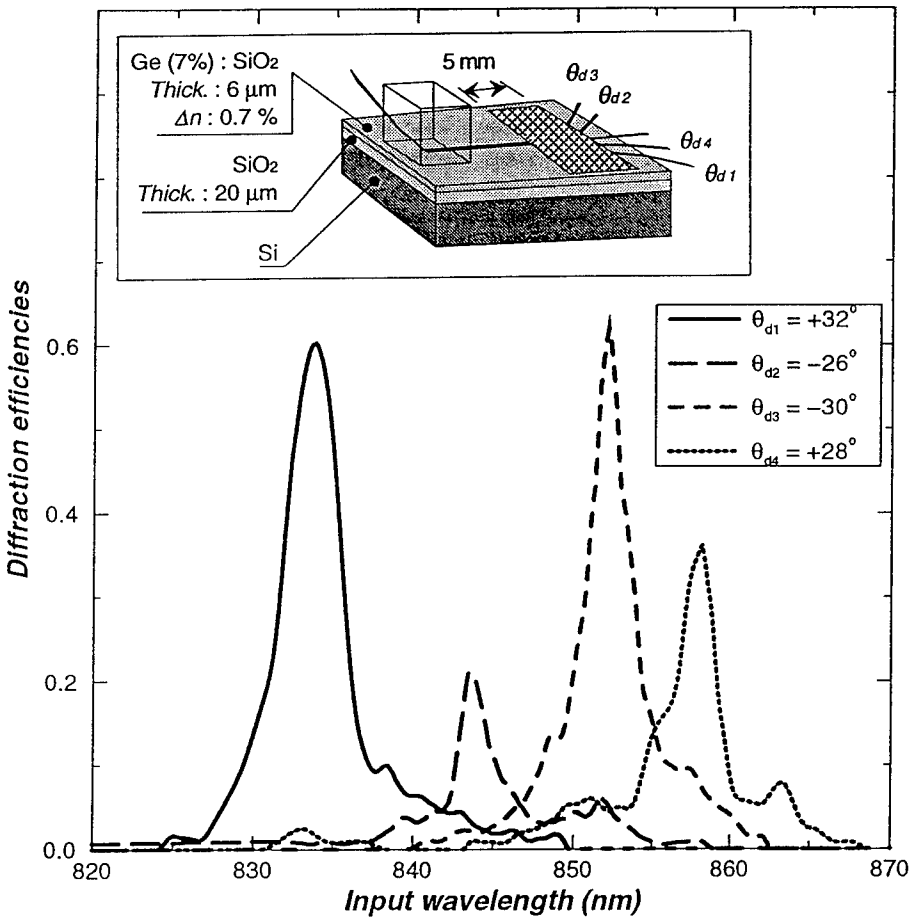


Figure 6: Measured diffraction spectrum of each port of the 4-channel WDM. The inset shows the guiding structure and the WDM geometry.

this to the non-uniformity of the excimer laser beam itself. Indeed, the particular laser we have in our facility was working at just above the threshold because a Fabry-Perot etalon was installed inside the cavity to obtain the minimal coherence length needed for the holographic writing.

The non-uniformity of the index modulation over the gratings area explains the broadening of the spectral response and also the relatively high level of cross-talk. It also partially explains the relatively low diffraction efficiencies, which should be improved with a better control of the gratings strength, i.e., of the exposure.

However, as a proof-of-concept demonstration, the 4-channel WDM implemented in silica glass exhibited satisfactory functionality of wavelength demultiplexing.

4. Conclusion & perspectives

The feasibility of integration of passive and active optical devices in glass has been demonstrated. Future projects will extend the integration.

A realistic goal for the near future on the laser front is an amplifier in an optical cavity, formed by a distributed Bragg grating written onto the guide and including directional couplers, wavelength multiplexors, optical fibre connectors and isolators. Another possibility is a system of 3 dB splitters making a 1 x 8 fan-out with zero insertion loss. Switching, using pump induced refractive index changes has been demonstrated elsewhere [18].

On the WDM front, the concept of supergrating WDM in a planar waveguide, built in special glasses, photorefractive materials or semiconductors and the device design method can be further extended to various new configurations and implementation schemes. Further improvements to the device performance are shown to be possible through both theoretical analysis and experiments. Implementations in other material systems, including the Er:Ytterbium doped phosphate glass, are being carried out with preliminary success.

5. Acknowledgements

This collaborative project was made possible by supports from:

- The International Co-operation Program between the Région Rhône-Alpes (France) and the Province of Ontario (Canada).
- The French "Ministère de l'Enseignement Supérieur et de la Recherche".
- K. OGAWA of *AT&T Bell Laboratories*, W. WESTWOOD and G. CHIK from *Bell-Northern Research*, G. DUCK of *JDS Fitel* and I. TCHAPLIA of *ITS Electonics*.

6. References

1. G. H. Chartier, P. Jaussaud, A. D. de Oliveira, O.Parriaux, *Electron. Lett.*, 1977, Vol. 13, pp. 763.

1a. H. Aoki, O. Maruyama, Y. Asahara, "Glass waveguide laser", *IEEE Phot. Tech. Letters*, 1990, Vol. 2, No. 7, pp. 459-460.

2. D. Barbier, J. Hubner, J.M. Jouanno, A. Kévorkian, A. Lupascu, B.Hyde, "Waveguide amplifiers in rare-earth doped glasses: fabrication, characterisation, modelling and

prospects", Post-Deadline Proceeding of the *ECIO'93 conference*, April 18-22, 1993, Neuchâtel, Switzerland & Poster at *ECIO'95*, April 3-6, 1995, Delft, Holland.

3. D. Barbier, J.M. Delavaux, A. Kévorkian, P. Gastaldo, J.M.Jouanno, "Yb/Er integrated optics amplifiers in phosphate glass in single and double-pass configuration", *OFC'95*, Postdeadline papers, PD3, February 23-March 3, 1995, SanDiego, California.

3a. D. Barbier, P. Gastaldo, B. Hyde, J.M. Jouanno, A. Kévorkian, "Amplification in Erbium Doped Microguides Realised on Phosphate Glass", *ECIO'95*, 1995, Delft, Holland.

4. D. Barbier, J.M. Delavaux, R.L. Hyde, J.M. Jouanno, A. Kévorkian, P.Gastaldo, "Tunability of Yb/Er Integrated optical lasers in phosphate glass", *OAA'95*, Postdeadline papers, PD3-2, June 15, 1995, Davos, Switzerland.

5. for example, M. Fallahi *et al.*, "Grating demultiplexer integrated with MSM detector array in InGaAs/AlGaAs/GaAs", *IEEE Photonics Techn. Lett.*, Vol. 5, No. 7, July 1993.

6. for example, B.H. Verbeek *et al.*, "Integrated four-channel Mach-Zehnder multi/demultiplexer fabricated with phosphorous doped SiO2 waveguides on Si", *J. of Lightwave Techn.*, Vol. 6, No. 6, June 1988.

7. for example, H. Takahashi *et al.*, "Transmission characteristics of array waveguide *N* x *N* wavelength multiplexer", *J. of Lightwave Techn.*, Vol. 13, No. 3, June 1995.

8. M.R. Wang, R.T. Chen, G.J. Sonek and T. Jannson, "Wavelength-division multiplexing and demultiplexing on locally sensitized single-mode polymer microstructure waveguides", *Optics Letters*, Vol. 15, No. 7, April 1990.

9. for example, R.T. Chen, H. Lu, D. Robinson and M.R. Wang, "Ten channel single-mode wavelength demultiplexer in near IR", *SPIE Vol. 1583, Integrated Optical Circuits*, 1991.

10. V. Minier, A. Kévorkian and J.M. Xu, "Diffraction characteristics of superimposed holographic gratings in planar optical waveguides", *IEEE Photonics Techn. Lett.*, Vol. 4, No 10, October 1992.

11. V. Minier, A. Kévorkian and J.M. Xu, "Superimposed phase gratings in planar optical waveguides for wavelength demultiplexing applications", *IEEE Photonics Techn. Lett.*, Vol. 5, No 3, October 1993.

12. V. Minier and J.M. Xu, "Coupled-mode analysis of superimposed phase gratings guided-wave structures and intergrating coupling effects", *Optical Engineering*, Vol. 32, No. 9, September 1993.

13. see [12] for list of references.

14. for example, A. Yariv and M. Nakamura, "Periodic structure for integrated optics", *IEEE, J. of Quantum Electronics*, Vol. QE-13, No. 4, April 1977.

15. H. Nishihara, M. Haruna and T. Suhara, "Optical integrated circuits", (1989), *McGraw-Hill Book Cie*, Chap. 4.

16. F. Bilodeau *et al.*, "Photosensitization of optical fiber and silica-on-silicon/silica waveguides", *Optics Letters*, Vol. 18, No. 12, June 1993.

17. C.G. Askins *et al.*, " Fiber Bragg reflectors prepared by a single excimer pulse", *Optics Lett.*, Vol. 17, No. 11, June 1992.

18. R.H. Pantell, M.J. Digonnet, R. Sadowski, H.J. Shaw, "Analysis of non-linear optical switching in an Erbium-doped Fiber", *J. of Lightwave Technology*, Vol.11, No. 9, pp. 1416, Sept.1993.

ULTIMATE PERFORMANCE OF DIODE LASERS IN FUTURE HIGH-SPEED OPTICAL COMMUNICATION SYSTEMS

S. A. GUREVICH
A. F. Ioffe Physico-Technical Institute, RAS
26, Politekhnicheskaja, St Petersburg, 194021, Russia

1. Introduction

During last two decades the transmission capacity of fiber-optical communication systems has been increased for about one order of magnitude in each five years. Now we are at the edge of 10 Gbit/s. Further increase in transmission rates, up to 40 Gbit/s, is considered as a goal to be reached at the end of this century. The main purpose of such systems is to establish local and global computer networks capable to operate with huge amount of information. This is, of course, rather challenging engineering problem.

Two general approaches are useful in realization of extremely high transmission rates in fiber-optical communication links. They are time division multiplexing (TDM) and wavelength division multiplexing (WDM). TDM implies very broad operation bandwidth of diode lasers, photodetectors and driving electronic components, the integrated optoelectronic transmitter and receiver to be the only practical design in this case. In WDM systems, the precise control of multiple wavelength in operation is of grate demand. Regarding diode lasers as a major light source for TDM and WDM systems, the question is whether these lasers can meet the specific requirements imposed by high operation frequencies and desired high degree of wavelength control. In diode lasers directly modulated by pumping current the modulation bandwidth is limited by several factors like differential and nonlinear gains, carrier drift time across the waveguide, carrier capture time into the active layer, the maximum reported bandwidth being 33 GHz [1]. Characteristic for directly modulated diode lasers is also the dynamic chirping of the emission wavelength [2]. At transmission rates above 10 Gbit/s chirp may create severe problems if the length of line is about 100 km. Besides, there is temperature related drift of emission line (about 1Å/K), which prevents from precise wavelength control. This is especially worth effect when WDM is used.

However, several new ideas are circulating which may be useful in designing of diode lasers free from the above discussed shortages. Practical realization of these propositions might be a breakthrough for extremely high transmission rates.

S. Luryi et al. (eds.), Future Trends in Microelectronics, 353–364.

Recently, dual modulation technique has been proposed [3], which implies that pumping current is modulated simultaneously with one of the parameters controlling the optical field in the laser cavity. Possible candidates are material gain, optical confinement factor or mirror reflectivity. With appropriate choice of amplitudes and waveforms of two laser controlling signals, the frequency dependent laser output response may be substantially flat, promising the modulation bandwidth up to 100 GHz. Dual modulation results also in radical suppression of chirp by rejecting carrier concentration oscillations inside the laser active layer. Below, we will examine in some detail the advantages of laser operation under dual modulation of the pumping current (J) and optical confinement factor (Γ). In the consideration of dual J&Γ modulation scheme we refer to the specific laser structure proposed earlier for confinement factor control [4]. Besides, we will look at new design of the diode laser which allows for efficient control of the cavity losses via governing the laser mirror reflectivity. In some cases, dual modulation by pumping current and mirror reflectivity, J&R modulation, may turn to be even more promising for high speed operation when compared to J&Γ modulation.

The spectral position of the laser emission line may, in general, be stabilized by using extended cavity lasers, where the gain area is only a small part of the hole resonator length. In this paper we are considering new laser design based on integrated optic technology, which allowed earlier to effectively match on a common substrate the heterostructure laser waveguide with dielectric-film waveguide [5]. In composite cavity laser, having integrated distributed Bragg reflector on dielectric-film waveguide and extremely short gain region, both the emission line drift and chirp are supposed to be reduced by two orders of magnitude compared with common diode lasers.

2. Possible High Speed and Low Chirp Operation of a Diode Laser Modulated by Dual Pumping Current and Optical Confinement Factor Control

An important factor limiting the modulation bandwidth of a diode laser is that the influence of pumping current on the photon density is not straightforward but takes place via the variation of nonequilibrium carrier density in the laser active layer. As a result, even in the ideal case of instantaneous carrier drift across the waveguide, fast carrier capture into QW active and lack of circuit parasitics, the intrinsic output modulation response of the laser drops rather abruptly with the increase of modulation frequency. The situation may be quite different and the modulation bandwidth can be substantially expended if the pumping current is controlled simultaneously with the confinement factor. Demonstration of enhanced operation frequencies and chirp suppression by dual modulation of optical confinement factor and pumping current is the subject of ongoing research work.

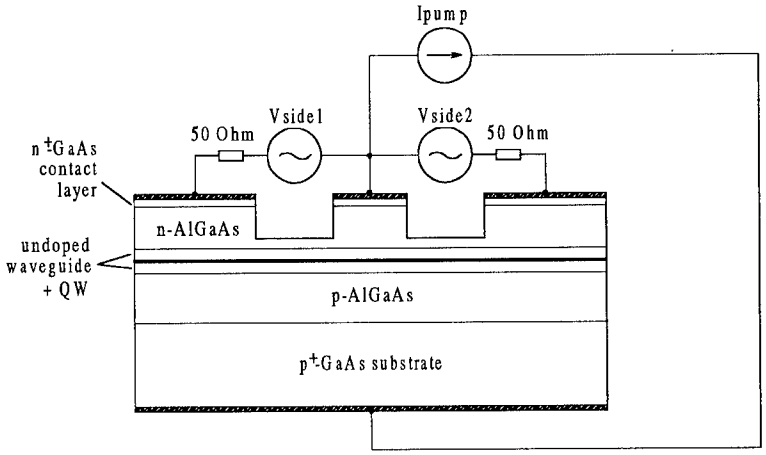


Figure 1. Four-terminal ridge guide laser structure and its connection scheme.

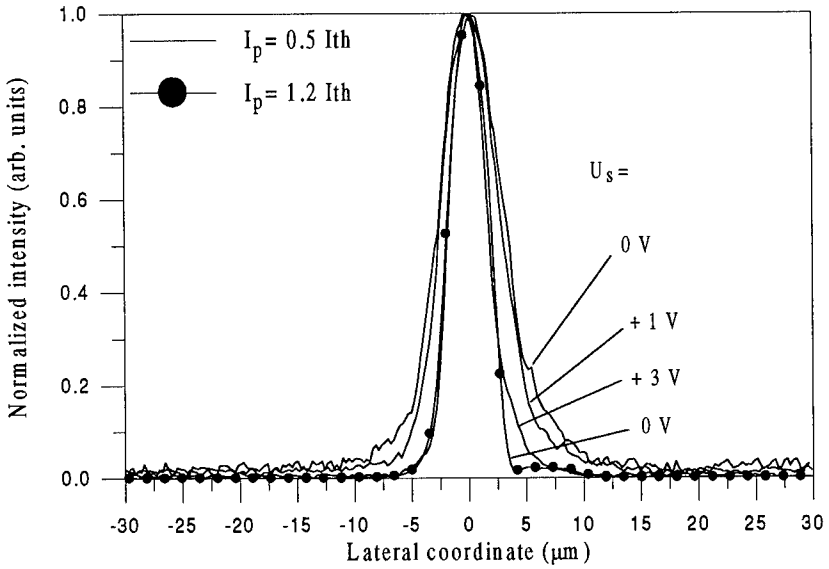


Figure 2. Normalized near field distributions registered at various side voltages below and above the laser threshold.

2.1. FOUR-TERMINAL RIDGE-GUIDE LASER STRUCTURE CAPABLE FOR CONFINEMENT FACTOR MODULATION

In the following we consider the laser structure of a ridge guide type [4], having four electrodes (Fig. 1). Single quantum well AlGaAs/GaAs separate confinement heterostructures were used for the device fabrication. Two grooves were etched in parallel to define the ridge and to separate the ridge from the side areas. Two contact terminals placed on the ridge top (5 μm wide) and on the substrate surface were used for laser

pumping, whereas two electrodes situated on the top surface of the side areas were for governing the potential distribution in the structure. The laser cavity length was normally 500 μm. Fig. 1 shows also the laser connection. The current generator, I_{pump}, is connected between the ridge and the substrate. Two voltage generators, U_{side1} and U_{side2}, are connected between the ridge and side contacts to produce additional bias of the structure. With this, near field intensity distributions were measured, scanning the image of the laser mirror in lateral direction, parallel to the junction plane. In these experiments the subject of main interest was the dependence of near field patterns on the voltage applied to the side contacts.

Fig. 2 shows normalized near field patterns recorded at different U_s voltages and pumping currents. Below the laser threshold, FWHM of near field distributions depends on side voltage if it is in the range U_s=1÷3 V. It means that U_s controls the current spread in the structure and thus governs the lateral gain/loss profile. Above the threshold, U_s voltage was found not to influence FWHM of the near field pattern corresponding to the lateral mode shape. This is because the lateral mode shape is determined by index guiding ridge. Consequently, U_s voltages influences the overlap of the gain and optical mode which is the mechanism of optical confinement factor control.

2.2. FORMULATION OF THE MODEL

To analyze the dynamic behavior of our four-terminal laser we carried out small signal analysis of laser rate equations. Basing on previous results, we assume that at sufficiently high U_s the size of inverse population area w is controlled solely by the side voltage. In the laser with the cavity length l, the current density J is given by:

$$J = \frac{I}{w \cdot l} \tag{1}$$

where I is the overall pumping current. The variation of (1) gives:

$$\frac{\delta J}{J_0} = \frac{\delta I}{I_0} - \frac{\delta w}{w_0} \tag{2}$$

The optical confinement factor Γ should also be considered as a function of w. Using the effective refractive index approximation, the optical confinement factor can be represented as a product of its lateral and transverse parts:

$$\Gamma = \Gamma_l \cdot \Gamma_t. \tag{3}$$

Assuming carrier density and gain are constant over the inverse population area of width w, we have:

$$\Gamma_l(w) = \frac{\int_0^w |E(x)|^2 \, dx}{\int_{-\infty}^{\infty} |E(x)|^2 \, dx}, \tag{4}$$

where $E(x)$ is the lateral mode field, x - axis is paralleled to the junction plane. We will consider I and U_s as driving forces, causing the variations of w, Γ_t, J and, consequently, of carrier and photon densities. The variation of optical confinement factor can be derived from (3) and (4) as:

$$\frac{\delta\Gamma}{\Gamma_0} = \frac{|E(x)|^2 \delta w}{\int_0^w |E(x)|^2 dx}. \tag{5}$$

Defining the mean value of optical mode intensity as:

$$\overline{E^2} \cdot w = \int_0^w |E(x)|^2 dx, \tag{6}$$

and introducing the ratio of optical field intensity at the point $x=w$ to its mean value:

$$\xi = \frac{|E(w)|^2}{\overline{E^2}}, \tag{7}$$

we obtain:

$$\frac{\delta\Gamma}{\Gamma_0} = \xi \frac{\delta w}{w_0}. \tag{8}$$

The parameter ξ characterizes the efficiency of optical confinement factor modulation. As it is clear from (7), $\xi \approx 1$ when w is small compared to lateral size of the optical mode and $\xi \to 0$ if w is comparable or even large with respect to mode size. The value of ξ is determined by side voltage as well as by the parameters of ridge waveguide. In our case $\xi \approx 0.1$ when $U_s \approx 1.0$ V (see Fig. 2). Further, neglecting the influence of parasitics in the laser structure we assume that temporal variation of w immediately follows the variation of U_s. With this, we have:

$$\frac{\delta w}{w_0} = -\zeta \frac{\delta U_s}{U_{s0}}. \tag{9}$$

Using the results of Fig. 2, one can estimate $\zeta \approx 0.36$ at $U_s \approx 1.0$ V.

Peculiar feature of the task under consideration is that in our laser the carrier and photon densities can be governed by simultaneous variations of pumping current and side voltage. Various operation regimes are available in this case. We are going to concentrate on one of them, more particularly, we will find a solution of the rate equations, setting the variation of carrier density to be equal to zero ($\delta n \equiv 0$). As it was first discussed in [3], this doesn't mean the photon density will be constant. In fact, the variations of pumping current and confinement factory may compensate each other to keep $\delta n=0$, still causing appreciable variation of photon density. Obviously, output modulation with $\delta n=0$ leads to suppression of chirp. With $\delta n=0$ neither optical cavity length nor spectral position of gain peak are the functions of time. The other advan-

tages of this modulation regime is that relaxation oscillations are suppressed and modulation bandwidth is very broad.

Providing δn=0, the linearized rate equations are:

$$v g_0 \delta S = \frac{\delta J}{ed}, \tag{10a}$$

$$\dot{\delta S} + \Delta \delta S = (v g_0 S_0 + \frac{\beta n_0}{\tau_s}) \delta \Gamma, \tag{10b}$$

were Δ is the difference between modal gain and loss:

$$\Delta = \frac{1}{\tau_p} - v \Gamma_0 g_0,$$

others notations are common. Remarkable feature of equation (10a) is that the variation of photon density is just proportional to the variation of current *density*. If the amplitude of the current density variation δJ is kept constant by appropriate choice of δI and $\delta \Gamma$, the output response of the laser will be *frequency independent*. It means that in practice the modulation bandwidth will be limited only by the structure and circuit parasitics.

Eliminating δS from equations (10a) and (10b), we have:

$$\gamma(\omega) = \frac{i\omega + \Delta}{v g_0 (v g_0 S_0 + \frac{\beta n_0}{\tau_s})} \frac{j(\omega)}{ed}, \tag{11}$$

where $\gamma(\omega)$ and $j(\omega)$ are complex amplitudes of $\delta \Gamma$ and δJ variations, respectively, and ω is the modulation frequency. Frequency dependent relation (11) should be maintained in order to keep δn=0.

We will consider the case when side voltage and pumping current are modulated so that the relation between the amplitudes of these signals obey the following relation:

$$\frac{i(\omega)}{I_0} + \zeta \frac{u(\omega)}{U_{s0}} = \frac{j}{J_0} = \text{const} \tag{12}$$

To fulfill (11), the following additional relation should be satisfied:

$$u(\omega) \xi \zeta \frac{\Gamma_0}{U_{s0}} = -\frac{i\omega + \Delta}{v g_0 (v g_0 S_0 + \frac{\beta n_0}{\tau_s})} \frac{j}{ed} \tag{13}$$

2.3. RESULTS OF NUMERICAL SIMULATIONS

To illustrate the possibility of chirp and relaxation oscillation suppression, we carried out numerical simulations of laser response in case pumping current and side voltage are sinusoidal signals at the frequency equal to 20 GHz. Fig. 3a displays the specific choice of δI and δU_s signals made in accordance with relations (12) and (13). Fig. 3b shows how photon density S and carrier concentration n oscillate under these conditions. The oscillation of photon density displayed in terms of $(S\text{-}S_0)vg_0$ and indicated by

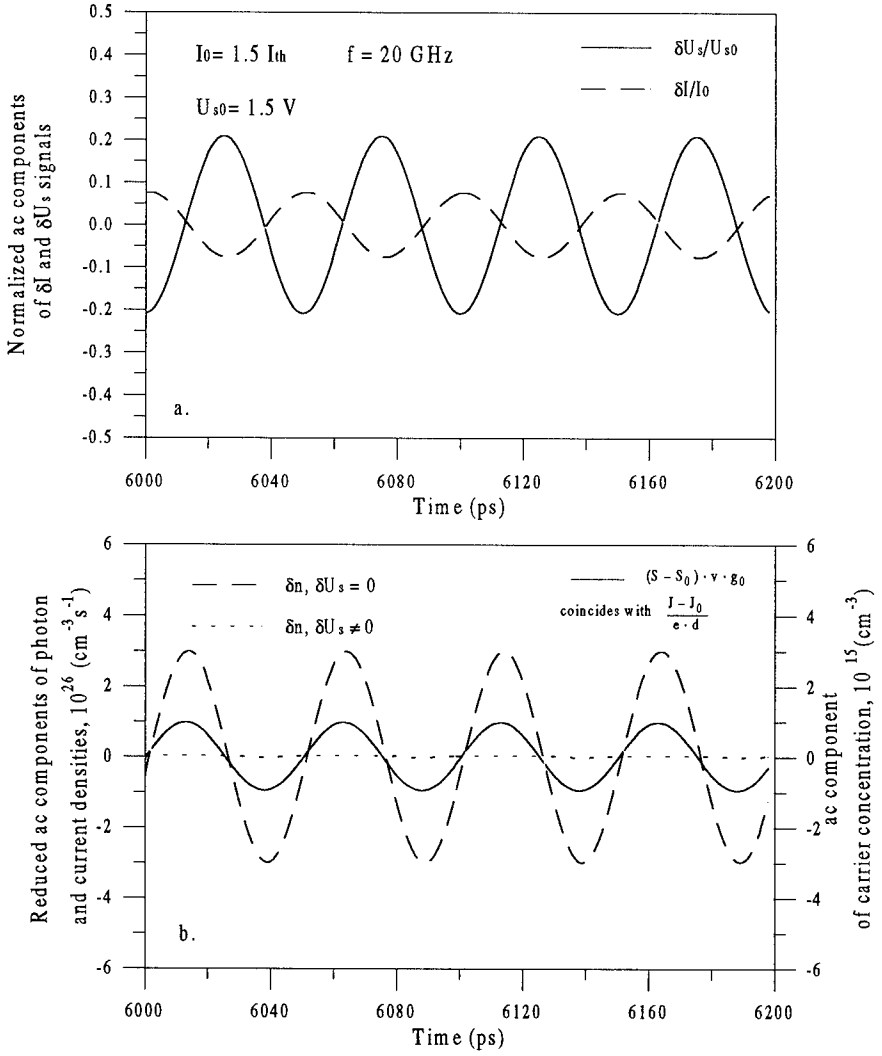


Figure 3. a. Normalized ac components of pumping current and side voltage signals.
b. Calculated responses of photon and current densities (solid curve) and of carrier concentration displayed under dual modulation ($\delta U_s \neq 0$, dotted curve) and direct modulation ($\delta U_s = 0$, dashed curve).

solid curve just merged with the variation of $(J-J_0)/ed$. This coincidence is an indication that equation (10a) is fulfilled. Dotted curve shows the excursion of carrier density from its steady state value. For comparison, dashed curve represents the oscillation of carrier density obtained with δU_s signal switched off. The last case corresponds to common direct modulation. Strong suppression of carrier density variation by dual modulation is evident from this example.

3. Dual Modulation of the Diode Laser, Having Y-Branch Tunable Mirror

In this section, we consider the possibility of the laser output modulation by controlling the pumping current and the mirror reflectivity. Compared to the laser structure studied for optical confinement factor modulation, new laser structure is even more tailored for high speed operation. To the best of our knowledge, there was only one paper [6] where the laser with tunable reflectivity mirror was described That was surface emitting laser with the reflectivity of DBR controlled by electrooptic variation of the layer refractive indexes. However this structure is rather complicated in fabrication and can hardly be considered as practical one.

The laser structure having integrated Y-branch tunable mirror is shown schematically in Fig. 4. When the phase difference between two waves coming back after propagation through Y-branch reflector is equal to odd integer multiplied by π, destructive interference takes place in the input waveguide. In this case, the optical power is radiating into substrate modes at the Y-point and the reflection from Y-branch section is near to zero. Contrary to this, if the phase difference is even integer multiplied by π, the reflection from ideal Y-branch mirror is equal to unity. The variation of phase difference in Y-branch reflector and thus the modulation of its reflection coefficient can be accomplished by using the dependence of the refractive index on carrier concentration or electric field which are sufficiently strong effects in QW containing A^3B^5 heterostructure materials.

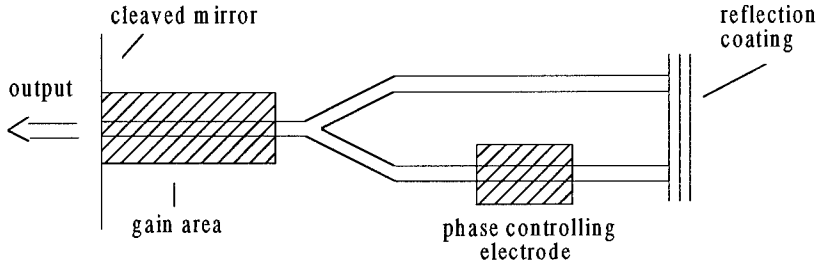


Figure 4. Schematic drawing of the laser with integrated Y-branch tunable mirror.

To eliminate the optical losses in the tunable mirror, the waveguide in Y-branch area can be formed by using impurity induced disordering, ion beam disordering or selective growth techniques. Next, dielectric is deposited on the structure surface and the windows are aligns and opened on the top of the gain and phase controlling areas

and contacts are formed. Reflection coating is produced at the end of Y-branch waveguides. It is to note that integrated Y-branch reflector placed inside the laser cavity can be used as tunable mirror in combination with DFB structure in the gain area.

As most of electrooptic devices, integrated Y-branch reflector can be tuned at rather high frequencies. Being limited by only RC parameters of the circuit, the operation bandwidth of the reflector itself may be about several dozens of GHz. Note that, the above described design insures strong electrical separation of gain and phase controlling areas. Owing to this, the electrical signals governing the gain and phase controlling areas will not intermix. This feature is important when dual modulation scheme is applied. The above described structure may be superior over that designed for optical confinement factor modulation, where the maximum speed of operation is, in practice, more complicated function of structure parasitics. Thus, in the laser with tunable Y-branch reflector all the advantages of dual modulation scheme can be realized.

4. Frequency Stabilized Diode Laser with Monolithically Integrated Dielectric Film Waveguide DBR

In single frequency DFB and DBR heterostructure diode lasers the stability of the spectral position of single longitudinal mode is an important parameter. There are two sources of emission line drift in DFB and DBR lasers. One is related to the variation of carrier concentration occurring under hf output modulation. This is the reason for dynamic line chirp. The other source of line drift is associated with temperature dependence of the refractive indexes of waveguiding layers. To some extent it could be eliminated by introducing the electrical feedback loop to the laser mount cooler but this will enhance complexity and cost of the laser module. Dynamic chirp in DFB and DBR laser can be eliminated by appropriately choosing coupling coefficients and grating phase shifts or by using external modulators. However, the chirp reduced in these ways is still fare above the transfer limited one [7]. Lasers operating free from both excess dynamic chirp and temperature related line drift would be very useful for high capacity communication systems.

Our proposition for the lasing line stabilization is based on the results of previous work [5], where we have first developed the integrated optic technology which makes it possible efficient end-to-end matching of heterostructure laser waveguide with thin-film dielectric waveguide sputtered on a common semiconductor substrate. Using this technology, single frequency heterostructure lasers with monolithically integrated DBR on corrugated dielectric waveguide have been fabricated. Because of very small temperature coefficients of refractive indexes of dielectric waveguiding layers, the spectral position of DBR reflection band was extremely stable. The width of DBR reflection band being $\approx$ 2 Å, its temperature drift was as small as $\approx$ 0.01 Å/K. Owing to high waveguide matching efficiency, low optical loss in dielectric DBR and because of high regularity of DBR parameters, the laser output efficiency was high, 32% per cleaved

output mirror. Besides selection of single longitudinal mode, dielectric film DBR offered strong selection of fundamental lateral mode.

However, the stabilization of DBR reflection band is necessary but not sufficient for the emission line stabilization. Indeed, in DBR lasers discussed above the variation of device temperature led to the variation of refractive index of heterostructure gain area and due to this the longitudinal cavity modes drifted through narrow and position fixed DBR reflection band. As a result, mode switching was observed under the temperature variation. To avoid mode switching one should, in addition, fix the position of the cavity mode.

This may be achieved in composite cavity DBR laser fabricated by using the above discussed integration technique. The proposed laser structure is shown schematically in the Fig. 5. In this laser the heterostructure gain area has relatively small length l and heterostructure waveguide is monolithically and-to-end joint with dielectric waveguide. The corrugation of the dielectric waveguide starts at the distance L from the waveguide matching interface. Thus, DBR laser has composite cavity with the length of regular waveguide area equal to l+L. For the laser to operate in single frequency mode, the length l+L should be taken so that the corresponding intermode distance be comparable to the width of DBR reflection band. On the other hand, to make the emission line position insensitive to the variations of both temperature and carrier concentration, the condition l/L << 1 should held. The other important condition involved is that the light reflection at the waveguide matching interface should be small. This condition seems to be easily fulfilled by using previously developed integration technology.

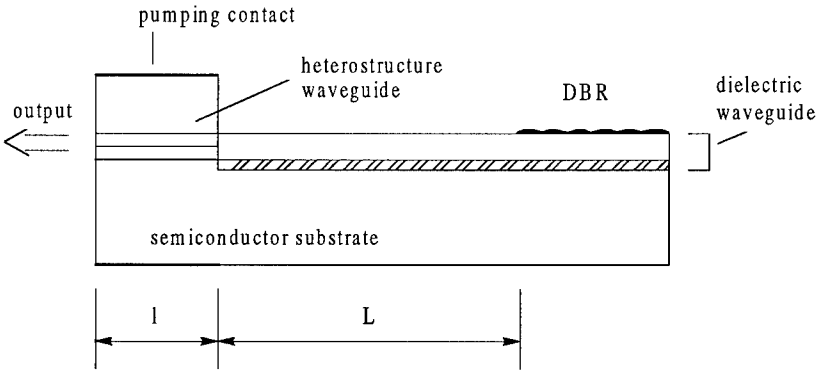


Figure 5. Schematic drawing of the laser with integrated dielectric film waveguide DBR.

To evaluate possible laser parameters, we take l=50μm which is reasonable choice for minimum length of heterostructure gain region comprising MQW active layers. Taking L=5.000μm and the refractive index of dielectric material n ≈ 2.0, the separation between the composite cavity modes will be about 0.5 Å. Owing to low optical loss in dielectric film DBR, its reflectivity is high and the width of reflection band can be controlled by varying the corrugation depth. In particular, it may be set to be equal to 0.5 Å. This will ensure single frequency operation of the proposed laser.

Further, with l = 50μm and L = 5.000μm, the ratio $l/L = 10^{-2}$. With this, only small part of the cavity is heterostructure waveguide where the refractive index depends on temperature and carrier concentration, whereas the rest part is dielectric waveguide have practically constant index. Consequently, in our case, the longitudinal mode of the composite cavity will have the drift rate 10^{-2} Å/K. This value just concides with the drift rate of DBR reflection band expected in dielectric film DBR. Under such condition, the mode switching will not be observed in the laser and the overall drift of the lasing line will be 10^{-2} Å/K. In this laser chirp will be reduced by two orders of magnitude. It is to note, that due to small l/L ratio the capacitance of pumped active area will be ten times smaller that in usual lasers which will be beneficial for high speed operation.

5. Summary

In this paper we have discussed several possibilities for the diode lasers to have considerably expanded modulation bandwidth, low chirp and high terminal stability of the laser line spectral position. These are the features of main concern in the design of future very high-bit-rate fiber-optical communication systems. It was shown that dual modulation of the optical confinement factor and laser pumping current may result in operation bandwidth up to 100 GHz and in very low chirp. New laser structure with tunable Y-branch mirror is proposed for realization of dual modulation by governing the pumping current and cavity loss, which can be even more prospective for high speed operation. Novel structure of integrated composite cavity DBR laser is considered for suppression of dynamic chirp and thermal drift of the lasing line, promising the improvement of these parameters by two orders of magnitude.

Acknowledgments

The author would like to thank M. Shatalov, Dr. V. I. Skopina and E. Tanklevskaya for their essential assistants in this work, O. Utkina for the help in preparation of the manuscript. Useful discussions with Prof. R. A. Suris are greatly acknowledged. This work was supported in part by ISF grant NU 9000.

References

1. Ralston J.D., Eisele, K., Sah, R.E., Larkins, E.C., Weisser, S., Rosenzweig, J., Fleissner, J., and Bender, K. (1994) Low-Bias-Current Direct Modulation up to 33 GHz in GaAs-Based Pseudomorphic MQW Ridge-Waveguide Lasers Suitable for Monolithic Integration, *Conference Digest of 14th ISLC*, Maui, USA, 211-212.
2. Bowers, J.E. (1987) High speed semiconductor laser design and performance, *Solid- State Electronics* **30**, 1-11.

3. Gorfinkel, V.B., Camacho, F., and Luryi, S. (1993) Dual Modulation of Semiconductor Laser, *Proceedings of 1993 International Semiconductor Device Research Symposium (ISDRS)*, Charlottesville, Virginia, USA, 723-726.
4. Gorfinkel, V.B., Kompa, G., Gurevich, S.A., Shtengel, G.E., and Chebunina, I.E. High Frequency Modulation of a QW Diode Laser by Dual Modal Gain and Pumping Current Control, *Proceedings of 20th International Symposium on GaAs and Related Compounds*, Freiburg, Germany, 41-42.
5. Alferov, Zh.I., Gurevich, S.A., Karpov, S.Yu., Portnoi E.L., and Timofeev, F.N. (1987) Monolithically-Intergrated Hybrid Heterostructure Diode Laser with Dielectric-Film Waveguide DBR, *IEEE Journal of Quant.Electron.* **QE-26,** 869-881.
6. Blum O., Zucker, J.E., Chiu, T.H., Divino, M.D., Jones, K.L., Chu, S.N.G., and Gustafson, T.K. (1991) InGaAs/InP multiple quantum well tunable Bragg reflector, *Appl.Phys.Lett.* **59**, 2971-2973.
7. Jonson, J.E., Tanbun-Ek, T., Chen, Y.K., Fishman, D.A., Logan, R.A., Morton, P.A., Chu, S.N.G., Tate, A., Sergent, A.M., Sciortino, P.F., Wetch, K.W., and Jr. (1994) Low-Chirp Integrated EA-Modulator/DFB Laser Grown by Selective-Area MOVPE, *Conference Digest of 14th ISLC*, Maui, USA, 41-42.

INCREASED-FUNCTIONALITY VLSI-COMPATIBLE DEVICES BASED ON BACKWARD-DIODE FLOATING-BASE Si/SiGe HETEROJUNCTION BIPOLAR TRANSISTORS

Z. S. GRIBNIKOV
Institute of Semiconductor Physics
Ukrainian Academy of Sciences
Kiev-28, Ukraine

S. LURYI
Dept. of Electrical Engineering
SUNY at Stony Brook
Stony Brook, NY, U.S.A. 11794

A. ZASLAVSKY
Div. of Engineering
Brown University
Providence, RI, U.S.A. 02912

1. Introduction

As modern semiconductor circuits progress towards greater complexity and ever smaller feature sizes at ever greater processing cost, increasing the functionality of logic devices is becoming a primary direction in microelectronics research and development. Several classes of semiconductor devices promising higher functionality than standard transistor logic have been intensively studied over the past decade, including multiterminal real-space transfer[1] and resonant tunneling[2] heterostructures. However, the vast majority of these devices were based on the bandgap engineering possibilities of compound semiconductor heterostructures and required low temperatures for effective operation, making them ill-suited for current digital technology. Recently, there has been a demonstration of a class of increased functionality devices based on multi-emitter heterojunction bipolar transistor (HBT's) fabricated in the InGaAs/InP material system.[3] We propose to investigate the performance of an analogous, multi-terminal backward-diode HBT structures in the Si/SiGe material system,[4] which would utilize the advantageous properties of Si/SiGe heterojunctions and, more importantly, would feature full

S. Luryi et al. (eds.), Future Trends in Microelectronics, 365–370.

compatibility with the silicon technology that will dominate microelectronics for the foreseeable future.

The increased functionality HBT's rely on the incorporation of a backward diode into the emitter-base junction of a bipolar transistor together with multiple contacts to the emitter region. The base contact is dispensed with altogether, which simplifies the fabrication process. The emitter-base backward diode characteristic allows emitter contacts to extract the base majority carriers under reverse bias, so the emitter region contacts can perform either "emitter" or "base" electrode functions. This structural symmetry leads to higher logic functionality of a single device, leading up to a ten-fold demonstrated reduction in device elements for some (but not all) logic functions.[3] Since our variant of the multi-emitter increased-functionality HBT can be implemented using a single Si/SiGe heterostructure operating at room temperature, the proposed class of devices will be fully compatible with Si VLSI technology. Furthermore, freed of the "base" contacting problem, the proposed devices will be ideally suited to BiCMOS circuits, where the simplified bipolar fabrication will simplify integration with standard multi-mask CMOS designs at no penalty to the bipolar performance.

2. Device operation

In this section we illustrate the principle of operation[3,4] of increased-functionality HBT devices built in silicon. Consider a Si/SiGe/Si *npn* bipolar transistor with a backward diode[5] emitter-base (E-B) *np* junction (created by the appropriate doping of the emitter and base regions) and a standard base-collector *pn* junction. The emitter region is contacted by two (or more) trench-isolated electrodes, while the base is left floating, as shown schematically in Fig. 1. If the E-B junction is forward-biased, minority electrons are injected into the base and extracted by the collector, resulting in ordinary bipolar transistor operation with current gain. However, because of the backward-diode characteristic shown in Fig. 2, when the E-B junction is reverse-biased it acts as an efficient Ohmic contact to the base.

If in the structure of Fig. 1 both of the emitter contacts are grounded ($V_{e1} = V_{e2} = 0$) and the collector is biased high ($V_c >> kT$), a small collector current will flow as in a standard bipolar transistor with a floating base. Conversely, if one of the emitter contacts is grounded ($V_{e1} = 0$) and the other is biased $V_{e2} < V_c$, a large hole current will flow through the reverse-biased E-B junction under the second emitter (which acts as a "base" contact), leading to a large output current. If the current densities of the forward and reverse-biased junctions are J_1 and J_2, and corresponding ponding junction areas are A_1 and A_2 respectively, in the active transistor regime with high common-emitter current gain ($\beta >> 1$)

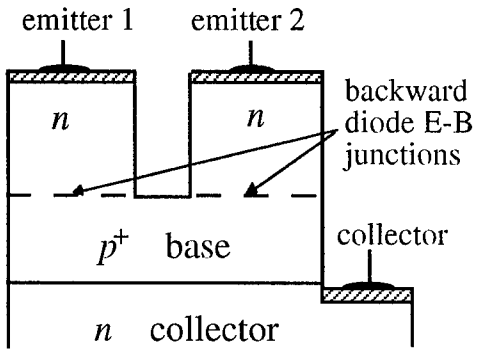


Fig. 1. Schematic of a structure with two emitter contacts (E-B junctions are backward diodes).

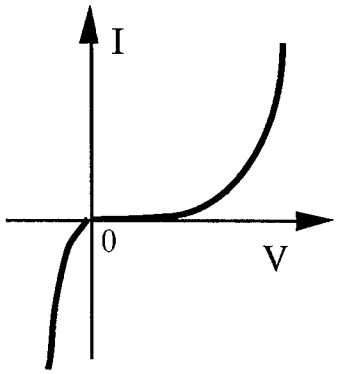


Fig. 2. I(V) characteristic of a backward diode in silicon (from Ref. 4).

we have:

$$I_c \approx A_1 J_1 = \beta A_2 J_2 \qquad (1)$$

Given the backward-diode E-B junction characteristics shown in Fig. 2, one can determine the transconductance characteristic $I_c(V_{e2})$ using the graphical construction shown in Fig. 3. By replotting the reverse bias characteristic of the backward diode multiplied by $-\beta A_2/A_1$, one can read off the biasing difference V_{e2} between the emitter contacts for a given collector I_c as the horizontal (voltage) distance between the forward bias curve and the rescaled reverse bias curve, as shown in Fig. 3.

Since the emitter contacts are fully symmetric, the single device of Fig. 1 possesses full *xor* logic functionality. By fabricating more than two contacts to the emitter region, more complicated logic functions can be implemented. For example, an analogous three-emitter contact structure can perform the *ornand* logic function in a single device: room-temperature operation of such a three-emitter structure was demonstrated recently.[3]

The utility of the proposed class of devices hinges on two requirements. First, the increased functionality of the device cannot come at the expense of high-speed operation typical of standard bipolar transistors. Second, since the range of high-level logic functions that can be implemented using the multiple emitter contact geometry of Fig. 1 is limited, the devices must in principle be compatible with BiCMOS logic circuits. The second requirement would appear to rule out the already demonstrated compound semiconductor versions of the device,[3] but both of these requirements can be addressed by implementing the devices using modified Si/SiGe/Si heterojunction bipolar transistor (HBT) designs.[6]

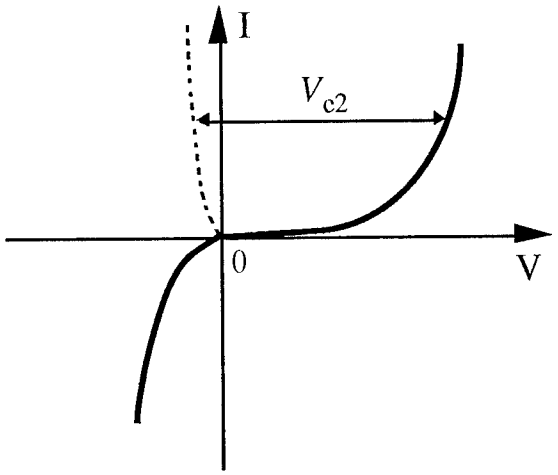


FIG. 3. Graphical determination of the transistor characteristic $I_c(V_{e2})$. The dashed line shows the reverse bias backward diode characteristic rescaled by $-\beta A_2/A_1$. For a given value of V_{e2} the output current I_c can be read off as shown.

High-speed operation of the device in Fig. 1 is governed by the standard HBT rules — high emitter efficiency (suppressed hole injection from base to emitter) in forward bias, high electron transfer coefficient through the base, and low base sheet resistance — combined with low backward diode resistance of the E-B junction in reverse bias. All of these rules can be simultaneously satisfied by making the heavily doped p-type base from epitaxial $Si_{1-x}Ge_x$ (with Ge content graded to $x \approx 0.25$ or $x \approx 0.3$ at the E-B junction) and then growing an n^+-Si emitter. The band discontinuity in the $Si/Si_{1-x}Ge_x$ occurs entirely in the valence band ($\Delta E_V \approx 200$ meV for $x = 0.2$)[7] conferring the usual heterostructure advantage of permitting high base doping (and hence low base sheet resistance) without sacrificing emitter efficiency. In fact, the absence of a barrier to forward injection of minority carriers into the base confers an additional advantage to the Si/SiGe implementation of the device over its compound semiconductor counterparts, as illustrated in Fig. 4 which shows the band diagram of the proposed *npn* Si/SiGe/Si HBT in the backward-diode (V_{e1} low;V_{e2} high) regime.

The very high base doping available in Si epitaxy reduces the tunneling resistance of the reverse-biased E-B backward diode, which is unaffected by the additional valence band barrier since the current is carried by Zener tunneling across the (narrower) bandgap. As a result, the proposed Si/SiGe/Si *npn* structures should have equivalent performance to state-of-

the-art Si/SiGe heterojunction bipolar transistors and yet perform high-level logic functions in a single device.

3. Conclusion

The implementation of proposed backward-diode floating-base HBT's in Si/SiGe heterostructures is advantageous from both the processing and VLSI integration standpoints. Since contacting the base is typically the most demanding task in bipolar transistor fabrication, the proposed structures would greatly reduce the number of processing steps. The integration of these bipolar devices with standard CMOS designs will benefit from the fabrication simplicity, leading to easy implementation of such promising BiCMOS circuits as CMOS logic integrated with high current-drive bipolar Si/SiGe input-output devices.

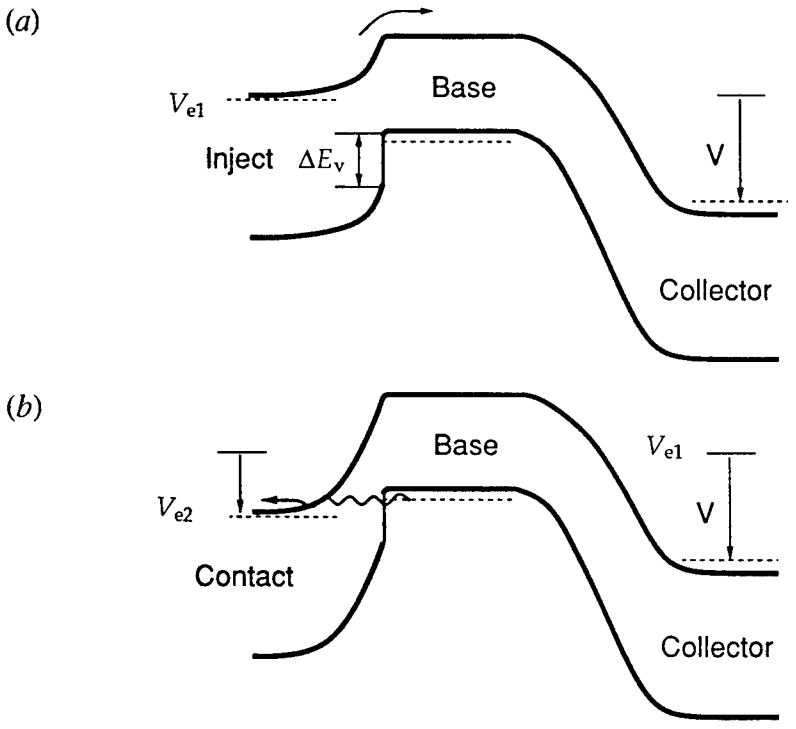


FIG. 4. Band diagrams of the device under the two emitter contacts under the V_{e1} = high (*a* — injecting emitter) and V_{e2} = low (*b* — contact emitter). The base is taken to be $Si_{1-x}Ge_x$ with the Ge fraction $x = 0.2$ (graded up from $x = 0$ in the low-doped base-collector junction).

4. References

1 Hess, S., Morkoç, H., Shichijo, H., and Streetman, B. G. (1979) *Appl. Phys. Lett.* **35**, 649; Luryi, S., Mensz, P. M., Pinto, M. R., Garbinski, P. A., Cho, A. Y., and Sivco, D. L. (1990) *Appl. Phys. Lett.* **57**, 1787.

2 Sen, S., Capasso, F., Cho, A. Y., and Sivco, D. L. (1987) *IEEE Trans. Electron. Dev.* **ED-34**, 2185; Seabaugh, A. C., Kao, Y.-C., and Yuan, H.-T. (1992) *IEEE Electron Dev. Lett.* **13**, 479.

3 Imamura, K., Takatsu, M., Mori, T., Bamba, Y., Muto, S., and Yokohama, N. (1994) *Electron. Lett.* **30**, 459; Imamura, K. *et al.*, Extended Abstracts 1994 ICSSDM, Yokohama, pp. 467-469.

4 Gribnikov, Z. S. and Luryi, S. (1994) Article comprising a bipolar transistor with a floating base, AT&T Bell Laboratories patent item 1-32, filed August, 1994.

5 Sze, S. M. (1981) *Physics of Semiconductor Devices*, 2nd ed., Wiley, New York, p. 537.

6 For a review of Si/SiGe heterostructures see, for example, Bean, J. C. (1992) *Proc. IEEE* **80**, 571.

7 Robbins, D. J., Canham, L. T., Barnett, S. J., Pitt, A. D., and Calcott, P. (1992) *J. Appl. Phys.* **71**, 1407.

REAL-SPACE-TRANSFER OF ELECTRONS IN THE INGAAS/INALAS SYSTEM

W. TED MASSELINK
Humboldt-Universität zu Berlin,
Institut für Physik,
10099 Berlin, Germany

1. Introduction

Monte-Carlo analysis [1] indicates that as FET gate lengths shrink to the sub-100 nm regime, the electric fields in a field-effect transistor become large enough that the electrons occupy most of the Brillouin zone. Because carriers outside of the Γ valley behave similarly in most semiconductors, transistor performance in this small-gate-length limit is predicted to be rather material-independent.

A noted exception to this generalization is $In_{0.53}Ga_{0.47}As$ (i.e., InGaAs lattice-matched to InP), which is predicted to significantly outperform most other semiconductors. The advantage in InGaAs is that the energy difference between the lowest-lying conduction-band Γ valley and the higher-lying L and X valleys is much greater than that in most other semiconductors. Thus the electrons in InGaAs tend to remain in the Γ valley to larger electric fields, resulting in a higher peak velocity and therefore in a higher average velocity with correspondingly smaller transit time in the transistors.

Ref. [1] assumes that each semiconductor is used in a MOSFET structure, in which the channel is implanted p-type and the gate is insulated with SiO_2, whose conduction band lies much higher than the conduction band of the semiconductor comprising the channel. InGaAs FETs, however, are usually of a heterostructure design, in which the electrons are confined by the wider-bandgap $In_{0.52}Al_{0.48}As$. Because the conduction band of the $In_{0.52}Al_{0.48}As$ lies only 0.5 eV above the conduction band of the $In_{0.53}Ga_{0.47}As$ [2], while the L valleys of the $In_{0.53}Ga_{0.47}As$ lie 0.55 eV above the Γ valley [3], scattering of channel electrons out of the channel into

S. Luryi et al. (eds.), Future Trends in Microelectronics, 371–376.

the InAlAs gate insulator (real-space transfer [4]) may be more important than the intervalley scattering. The present experimental results demonstrate that the peak electron velocity in $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ modulation-doped heterostructures is smaller than in bulk $In_{0.53}Ga_{0.47}As$ and indicate that the additional scattering mechanism responsible for this behavior is real-space transfer.

The use of pseudomorphic InGaAs with enhanced In content can improve the transport characteristics of InGaAs-containing heterostructures by decreasing the electron mass and increasing the energy separation between the InGaAs Γ valley and the InAlAs Γ valley. Further results of material transport and FET characteristics in pseudomorphic structures indicate that strain can be used to increase the bandgap compared to unstrained material, thus allowing higher breakdown voltages, while simultaneously improving the electron mobility and peak velocity.

2. Experimental Results and Discussion

Structures were grown using gas-source molecular beam epitaxy (GSMBE) on InP:Fe semi-insulating substrates. High-purity solid sources were used for the In, Al, and Ga, as well as for the Si, which serves as the n-type dopant. Thermally-cracked high-purity arsine supplied the arsenic. The growth rate was typically 2.5 Å/s and the In mole fraction y of the $In_yGa_{1-y}As$ was reproducibly controlled to within ± 0.003 as determined from the x-ray rocking curve data. Background doping in $In_{0.53}Ga_{0.47}As$ is n-type, approximately $7 \times 10^{15}\ cm^{-3}$ and can be lowered still further by using a lower hydride cracker temperature.

Electron mobility was measured using low-field Hall measurements in the van-der-Pauw geometry. Electron velocity as a function of electric field was measured up to about 6 kV/cm using a 35-GHz technique, which allows velocity measurements to be made without causing domains of non-uniform electron concentration [5].

Fig. 1. depicts the velocity–electric field characteristics for similarly-doped $In_{0.53}Ga_{0.47}As$ and GaAs, measured as described in Ref. [5]. Due to the lower effective mass in InGaAs, the low-field mobility of electrons in this material is higher than it is in equivalently-doped GaAs. Because of both the lower effective mass and the larger energy difference between the Γ and L valleys $\Delta E_{\Gamma-L}$ in InGaAs compared to GaAs, the electron velocity is higher not only at low electric fields (what the mobility measures), but for all electric fields. The peak velocity of the InGaAs sample is consistent with measurements of similar material based on the analysis of Gunn devices [6]. Both the GaAs and the InGaAs velocity curves are consistent with the Monte-Carlo calculation of Fischetti [7].

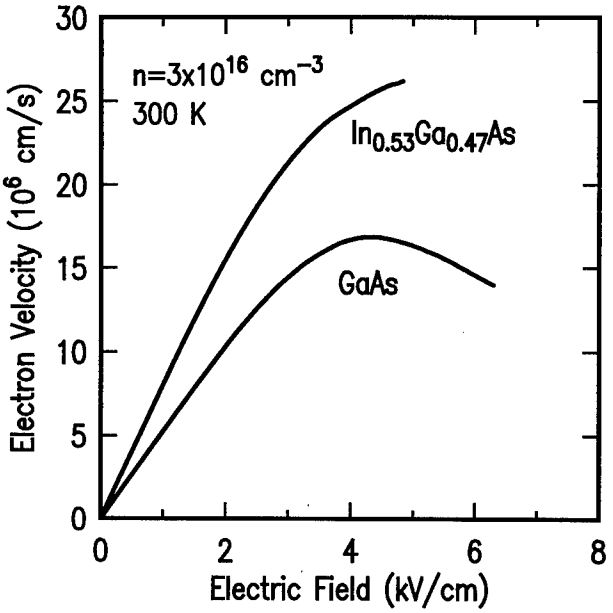


Figure 1. Comparison of electron velocity in lightly doped $In_{0.53}Ga_{0.47}As$ with that in similarly-doped GaAs.

Modulation-doped heterostructures were also prepared as a part of a larger series of variously-doped $In_{0.53}Ga_{0.47}As/In_{0.52}Al_{0.48}As$ heterostructures for application as FET structures [8]. One such sample has the following structure beginning after the InP substrate: a 400-Å $In_{0.52}Al_{0.48}As$ buffer layer, 300 Å of undoped InGaAs, a 50-Å spacer layer of undoped InAlAs, a delta-doped sheet of Si (2.4×10^{12} cm^{-2}), 300 Å of undoped InAlAs, and finally a 50-Å cap layer of n-doped InGaAs to assist in ohmic contact formation. The two-dimensional electron concentration in this structure was 2.0×10^{12} cm^{-2} and the low-field mobility of the electrons was 9320 cm^2/Vs at room temperature. The bulk InGaAs sample used for comparison consists simply of a 1-μm film of $In_{0.53}Ga_{0.47}As$ doped with Si so that n=3.5×10^{16} cm^{-3}. It had a room temperature mobility of 8100 cm^2/Vs. Fig. 2 shows the velocity versus electric field for these two samples. In both samples, the electron transport is in lightly-doped $In_{0.53}Ga_{0.47}As$ so that at low fields the mobilities are quite similar. At higher electric fields, however, the electrons in the heterostructure reach their peak at a significantly lower electric field. We propose that this reduced peak velocity occurring at a smaller electric field in the heterostructure is due partially to real-space transfer and partially to the total absence of density of electron states below the ground-state subband formed by the heterostructure and the conduction band bending in the InGaAs.

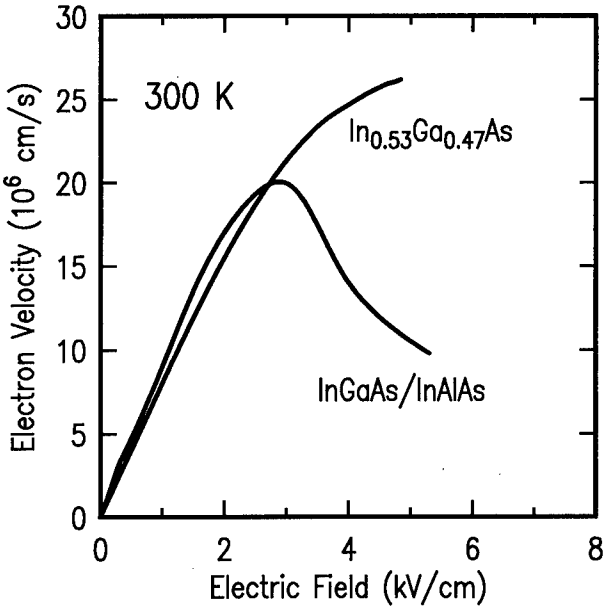


Figure 2. Comparison of electron velocity in lightly doped $In_{0.53}Ga_{0.47}As$ with that in a InGaAs/InAlAs modulation-doped heterostructure.

A similar effect has been observed in the GaAs/AlGaAs system [9, 5] in which the presence of the AlGaAs Γ valley allows scattering out of the GaAs Γ valley at a lower electric field. In the GaAs/AlGaAs system, the peak velocity depends on the alloy composition of the $Al_xGa_{1-x}As$, with x=0.3 resulting in a 20% lower velocity than in bulk GaAs. The presently-observed lowering of peak velocity in the InGaAs/InAlAs system appears to be analogous to that observed in the GaAs/AlGaAs system.

In the bulk $In_{0.53}Ga_{0.47}As$, the L valleys lie 0.55 eV higher than the Γ valley [3]. This is significantly higher than the 0.32 eV in bulk GaAs [7], accounting for the higher electron velocities in InGaAs. The presence of the adjacent InAlAs in the modulation-doped structure, introduces alternative scattering into the Γ valley in the InAlAs. Taking the band offset to be 0.5 eV [2], the InAlAs Γ valley is only 0.5 eV above the InGaAs conduction band.

In the heterostructure, however, the electrons cannot populate the energies immediately at the conduction band edge, but first find state to occupy at the first subband. This subband can be estimated to lie E_0=140 meV above the conduction band edge. The L valleys are not so effected because the electron masses are much larger there. Therefore, the Γ-L separation in the InGaAs in the heterostructure is reduced from what it is in bulk by the subband energy E_0. Together with the addition density of states into

which to scatter in the InAlAs, we expect these two effects to account for the peak velocity in the heterostructure occurring at a lower electric field and therefore smaller magnitude when compared to bulk. Simple numerical calculations similar to those of McCumber and Chynoweth [10] also support the hypothesis that the added scattering channel from the two-dimensional electron gas (2DEG) into the InAlAs can account for the reduced peak velocity [11].

One implication of this result is that in order to attain the ultimate performance potential in InGaAs-based FETs, a wider-bandgap insulator in place of the $In_{0.52}Al_{0.48}As$ would be advantageous to reduce real-space transfer. Because of the difficulty in depositing a wide-bandgap insulator such as SiO_2 onto the InGaAs and simultaneously maintaining a defect-free interface, the best compromise may be to use pseudomorphic $In_yAl_{1-y}As$ with excess Al ($y < 0.52$), thereby increasing the conduction band discontinuity [12].

Alternately, or in combination with the use of a wider-bandgap insulator, the structure can benefit from the use of pseudomorphic InGaAs with excess In. This material has a lower-lying Γ valley, resulting in an increased Γ-L energy separation. In related work [8], we have optimized one such variation of this idea with a strained InAs quantum well in the middle of a (lattice-matched) InGaAs well. Unstrained InAs has a bandgap of only 0.36 eV with L valleys lying 1.08 eV above the Γ valley. When the InAs is pseudomophically grown on the InP lattice constant, the bandgap will be approximately 0.55 eV. Further, in a 30-Åwell width, the confinement effects will effectively increase the InAs bandgap to approximately 0.65 eV. The InAs first electron subband is still lower than the Γ valley in the InGaAs so that electron transfer (either k-space into the L valleys or real-space into the adjacent InAlAs) is reduced through the presence of the InAs. Additionally, the relatively wide effective bandgap of the InAs quantum well ensures that under high electric fields, impact ionization will not be a limiting scattering mechanism.

3. Conclusion

Measured velocity versus electric field for electrons in bulk $In_{0.53}Ga_{0.47}As$ and in InGaAs/InAlAs modulation-doped heterostructures show that the heterojunction causes in a lower peak electron velocity, which is explained through real-space transfer of electrons from the InGaAs Γ valley into the InAlAs at lower fields than the k-space transfer of the electrons into the InGaAs L valleys. This effect ultimately limits the electrons in the heterosystem to lower velocities than those of electrons in bulk InGaAs. One solution to avoid this "premature" electron transfer is to use an insula-

tor whose conduction band discontinuity relative to that in InGaAs is greater than that between the InGaAs and the InAlAs, such as pseudomorphic InAlAs with excess Al. Another solution is to use narrower bandgap, strained InGaAs or InAs as the transport channel.

Acknowledgments – The author gratefully acknowledges the support of his colleagues at the IBM T.J. Watson Research Center in Yorktown Heights and especially the assistance of John Zahurak in fabricating some of the structures.

References

1. M.V. Fischetti and S.E. Laux, IEEE Trans. Electron Devices **38**, 650 (1991).
2. R. People, K.W. Wecht, K. Alavi, and A.Y. Cho, Appl. Phys. Lett. **43**, 118 (1983).
3. K.T. Cheng, A.T. Cho, S.B. Christman, T.P. Persall, and J.E. Rowe, Appl. Phys. Lett. **40**, 423 (1982).
4. K. Hess, H. Morkoç, H. Shichijo, and B.G. Streetman, Appl. Phys. Lett. **35**, 469 (1979).
5. W.T. Masselink, Semicond. Sci. Technol. **4**, 503 (1989).
6. D. Hahn and A. Schlachetzki, J. Electronic Mat. **21**, 1147 (1992).
7. M.V. Fischetti, IEEE Trans. Electron Devices **38**, 634 (1991).
8. J.K. Zahurak, A.A. Iliadis, S.A. Rishton, and W.T. Masselink, J. Appl. Phys. **76**, 7642 (1994); J.K. Zahurak, A.A. Iliadis, S.A. Rishton, and W.T. Masselink, IEEE Elec. Device Lett. **15**, 489 (1994).
9. W.T. Masselink, N. Braslau, W.I. Wang, and S.L. Wright, Appl. Phys. Lett. **51**, 1533 (1987).
10. D.E. McCumber and A.G. Chynoweth, IEEE Trans. Electron Devices **13**, 4 (1966).
11. W.T. Masselink, Appl. Phys. Lett. **67**, (to be published, 1995).
12. S.R. Bahl, W.J. Azzam, and J.A. del Alamo, IEEE Trans. Electron Devices **38**, 1986 (1991).

CHARGE INJECTION TRANSISTOR AND LOGIC ELEMENTS IN $Si/Si_{1-x}Ge_x$ HETEROSTRUCTURES

M. MASTRAPASQUA
Eindhoven University of Technology
600 MB Eindhoven, The Netherlands

C.A. KING, P.R. SMITH, and M.R. PINTO
AT&T Bell Laboratories
600 Mountain Ave
Murray Hill, NJ, 07974 U.S.A.

1. Introduction

The charge injection transistor (CHINT) [1] concept refers to a class of devices based on the principle of real space transfer (RST) [2] of hot carriers between two independently contacted conducting layers. One of these layers, the emitter, has source and drain contacts, while the other, collector, layer is separated by a heterostructure barrier. When an electric field is induced between the drain and the source, the carriers in the channel become *"hot"* and can overcome the barrier. The RST effect manifests itself in two ways. The collector current I_C, at a constant collector bias V_C, increases at a sufficiently high source and drain bias. Simultaneously, the drain current decreases showing a negative differential resistance (NDR) in the current voltage characteristic.

A fundamental property of RST transistor is that the collector current does not change if the source and drain contacts are interchanged. Thus the device exhibits an exclusive-OR (*xor*) dependence of the output current on the input voltages regarded as binary logic signals. Even more powerful logic functionality is obtained in a CHINT device with three input terminals [3]. This device, which we shall refer to as the ORNAND gate, has a cyclic three-fold symmetry. Depending on the logic value, *high* or *low*, of one of the three electrodes, the output current behaves as an *nand* or *or* function of the other two electrodes.

The charge injection transistor has been successfully implemented in a number of III-V systems [1, 4-6]. A monolithic optoelectronic ORNAND device has also been demonstrated in a InGaAs/InAlAs heterostructure material [7]. However, there is great interest in realizing the charge injection transistor in a silicon-based heterostructure, making it possible to integrate CHINT devices with traditional VLSI silicon logic and memory circuits. Hot hole RST has been observed in Si/SiGe heterostructures by Mensz *et al.* [8]. Recently, RST of electrons in a strained silicon layer has been observed in an n-type Si/SiGe heterostructure [9]. The present work reviews our recent results on the realization of CHINT, and a monolithic ORNAND function logic device with an epitaxial layer structure containing strained $Si_{0.7}Ge_{0.3}$ and Si [10].

S. Luryi et al. (eds.), Future Trends in Microelectronics, 377–383.

2. Epitaxial Structure and Device Fabrication

Figures 1 and 2 show schematic cross-sections of the device structures discussed below: the CHINT is illustrated in Fig. 1 and the ORNAND gate in Fig. 2. We grew the $Si/Si_{1-x}Ge_x$ by rapid thermal epitaxy (RTE) on a 125 mm p-type Si substrate. The Ge fraction (x) in the emitter channel and in the collector layer was 0.3, as determined by Rutherford backscattering spectrometry. The strained layer thickness was about 15 nm, as measured by transmission electron microscopy (TEM) and secondary ion mass spectrometry (SIMS), which is less than the mechanical equilibrium critical thickness for this composition.

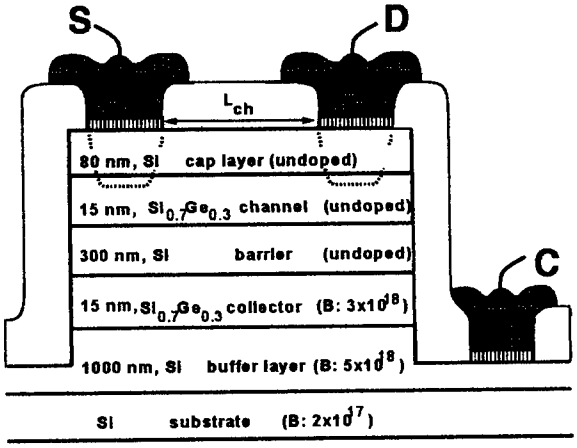


Figure 1. Schematic cross section of the charge injection transistor structure. The distance between source and drain, L_{ch} = 0.5 μm and the width W =40 μm.

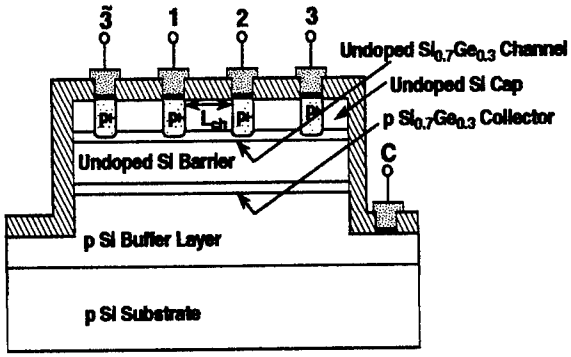


Figure 2. ORNAND logic gate structure. Each of the three channels, defined by the p^+ diffusion in the silicon cap layer, has length L_{ch} =1.0 μm and width W = 40 μm. Cyclic symmetry, required for the *ornand* logic operation, results from the periodic boundary condition on V_3.

Shallow source and drain ohmic contacts are the most critical processing step for CHINT devices, because the contact must be deep enough to penetrate through the cap layer into the 15 nm channel without reaching the underlying barrier layer. We used RTE to grow a B-doped Ge layer to act as a diffusion source [11]. The Ge layer grows selectively only on the exposed silicon surface and not on the oxide. After growth, we used rapid thermal annealing for 12 min. at 800 ^{o}C to form the junction contacts with an estimated final depth < 90 nm and active surface concentration $>10^{20}$ cm^{-3}.

Finally 10 nm Ti, 100 nm TiN and 500 nm of aluminum were deposited on the front of the wafer and patterned to form the source, drain and collector contacts. Since most of the band gap discontinuity between the strained $Si_{0.7}Ge_{0.3}$ channel layer and the Si barrier falls into the valence band, the devices employ RST of hot holes.

3. Device Characteristics

Figure 3 shows the room temperature current-voltage characteristic of the Si/SiGe CHINT. The drain current, Fig. 3 (a), shows a strong NDR for V_D above 1 V, with a peak to valley ratio (PVR) that increases with the collector bias. Simultaneously, the collector current increases, as illustrated in Fig. 3 (b). Prior to the onset of RST (V_D<1 V) a small I_C is present due to the thermionic emission at the lattice temperature of *"cold"* holes from the channel into the collector. For V_C = -5.5 V, a drain current PVR > 2 is observed. A further increase of the collector bias will bring only an apparent increase of the PVR in the drain current. In fact, as V_C increases the leakage of *"cold"* holes also increases, primarily because of hole accumulation at the heterointerface and the escalating role of tunneling. As a result, the I_D curve shifts down, almost rigidly, towards positive current, hence the apparently increasing PVR. We have characterized the high-frequency operation of the SiGe CHINT using a vector network analyzer. For a L_{ch} =0.5 μm device we found a short circuit current gain cutoff frequency of 6 Ghz.

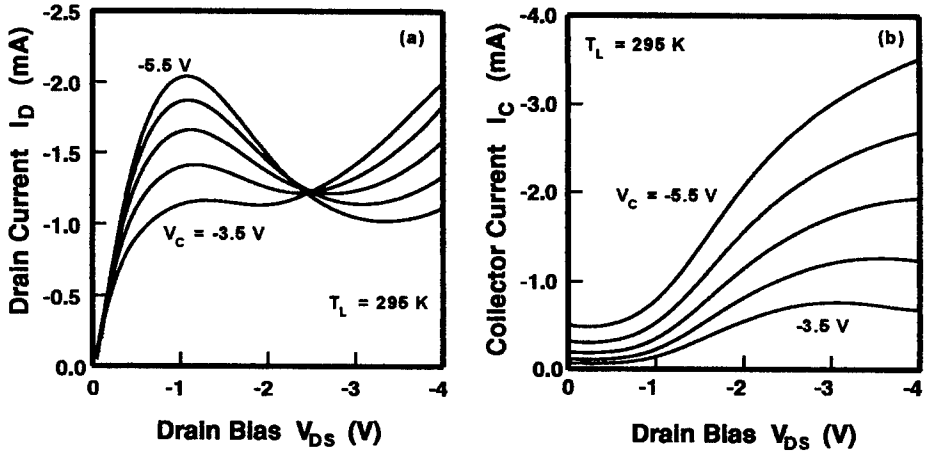


Figure 3. Room temperature characteristics of the drain (a) and collector (b) current at different collector biases for a 0.5 μm x 40 μm Si/SiGe CHINT.

4. Logic Functions

Figure 4 demonstrates the *xor* logic operation at two different lattice temperatures. We see that I_C = xor(V_D,V_S) at all temperatures. The on/off ratio is 10 dB at room temperature and increases to 65 dB at cryogenic temperature due to diminishing leakage current in the (0,0) logic state.

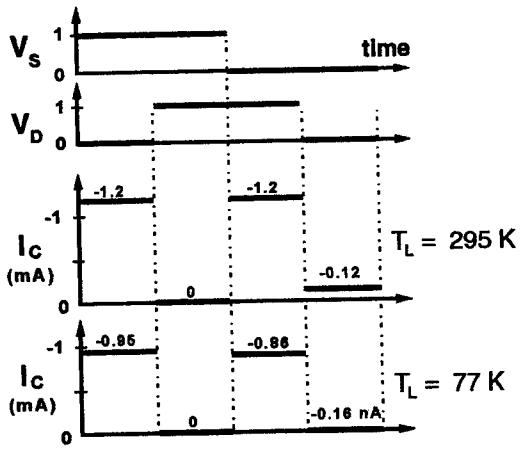


Figure 4. Exclusive-OR operation of a 0.5 μm x 40 μm Si/SiGe CHINT device. The collector bias is fixed at -4 V, and the binary values "logic-0" and "logic-1" of the input signals V_S and V_D correspond to 0 and -4 V respectively.

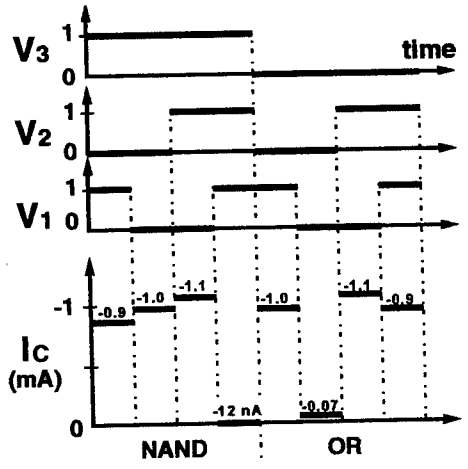


Figure 5. Logic operation of a 1.0 μm x 40 μm Si/SiGe ORNAND gate at a lattice T = 77 K. The collector bias is fixed at -5 V, and the binary values "logic-0" and "logic-1" of the input signals V_1, V_2, and V_3, correspond to 0 and -3 V respectively.

Figure 5 demonstrates the logic operation of the ORNAND multiterminal logic device. The collector bias is fixed at V_C = -5 V and the input signal V_1 and V_2 are varied between *low* = 0 and *high* = -3 V, while the split electrode, chosen as the control, is fixed either at the low value V_3 = 0 for the *or* function or at the high value V_3 = -3 V for the *nand* function. Since at room temperature the leakage of *"cold"* holes in the (0,0,0) state is comparable to the current in the on state, the operation of this device is demonstrated only at 77 K. However, if the channel length is reduced to 0.5 μm, decreasing the total area of the device, room temperature operation is expected.

5. Simulations

In order to study how the device performance of the Si/SiGe CHINT can be improved, we have used the device simulator PADRE [12] which solves Poisson, drift-diffusion and energy balance equations for an arbitrary heterostructure device in 1D/2D/3D spatial dimensions.

First, we have simulated the device with a structure as shown in Fig. 1, with the doping value and layer thickness noted above. Figure 6 compares the measured electric characteristic (dashed line) with the one simulated by PADRE (dotted line). It is important to stress that in these simulations we have not adjusted any parameters to fit the experimental data. Clearly, a large part of the discrepancy from simulations and measurements may simply be due to a parasitic resistance at the source, drain and collector contacts.

Figure 6 also shows how the drain and collector characteristic of a SiGe CHINT would change by changing some of the parameters in the design of the structure:

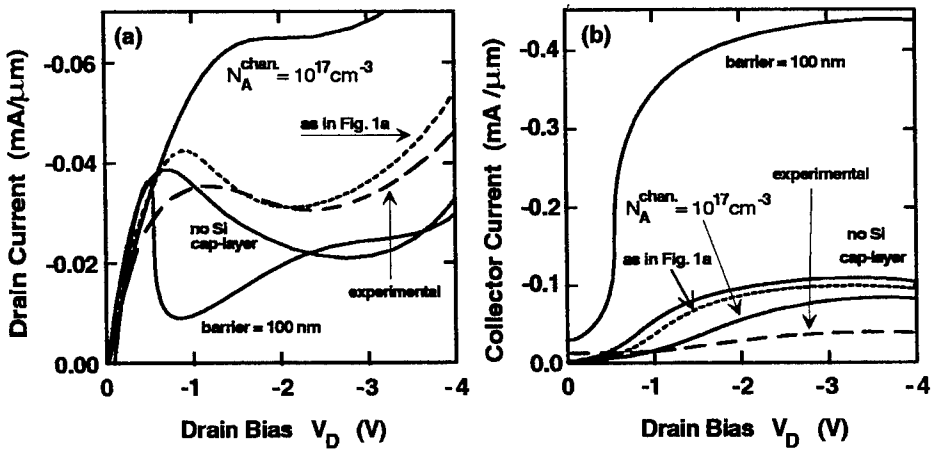


Figure 6. Drain (a) and collector current (b) as it is simulated by device simulator PADRE for different design parameters of the SiGe CHINT structure with L_{ch} =O.5 μm. The dotted line is the simulation for the device structure as shown in Fig. l. The dashed line is the experimental data from Fig. 2, V_C = - 4 V.

i) Increasing the doping level in the channel is a detriment. The channel mobility decreases due to impurity scattering. This results in a decrease of the carrier temperature and therefore a decrease of RST current.

ii) The absence of a silicon cap layer improves the device performance. The silicon cap layer constitutes an alternative current path for the drain current, especially at high drain bias. However, the use of a silicon cap layer is desirable for other reasons, although it could be thinner than 80 nm. From a point of view of the device fabrication, the cap layer helps to relax the requirements on the source/drain diffusion-contacts from being abrupt and shallow to only abrupt. From the crystal growth point of view, the presence of the cap layer promotes stability of the strained structure.

iii) A thinner barrier layer increases the overall device performance due to the increase of the electric field over the barrier.

6. Conclusions

Room temperature operation of a $Si/Si_{0.7}Ge_{0.3}$ CHINT is demonstrated. The drain current characteristic shows negative differential resistance with a peak-to-valley ratio greater than 2. The device shows an *xor* property with on/off ratio of 10 dB at 295 K and 65 dB at 77 K. A monolithic multiterminal logic device that functions at 77 K as an *ornand* gate is also demonstrated. Simulations have shown that the device performance can be further improved by reducing the thickness of the barrier layer and of the silicon cap layer. We believe that with the above changes and with a distance between the emitter contacts shorter than 1 μm, room temperature operation for the ORNAND device could be obtained. These functional devices may be integrated into silicon VLSI technology offering the possibility to employ the *xor* and *ornand* logic properties or the NDR current-voltage characteristic in silicon based circuits.

7. Acknowledgments

The authors would like to thank Ray Cirelli for his help in device processing and Serge Luryi for helpful discussions.

8. References

1. Luryi, S., Kastalsky, A., Gossard, A.C., and Hendel, R.H. (1984) Charge injection transistor based on real space hot-electron transfer, *IEEE Trans. Electron Dev.* **31**, 832-839.

2. For a review of real-space transfer phenomena see Gribnikov, Z.S., Hess, K., and Kosinovsky, G.A (1995) Nonlocal and nonlinear transport in semiconductors: real-space transfer effects, *J. Appl. Phys.* **77**, 1337-1373.

3. Luryi, S., Mensz, P., Pinto, M.R., Garbinski, P.A., Cho, A.Y., and Sivco, D.L. (1990) Charge injection logic, *Appl. Phys. Lett.* **57**, 1787-1789.

4. Mastrapasqua, M., Luryi, S., Capasso, F., Hutchinson, A.L., Sivco, D.L., and Cho, A.Y. (1993) Light-emitting transistor based on real-space transfer: electrical and optical properties, *IEEE Trans. Electron Dev.* **40**, 250-258.

5. Belenky, G.L., Garbinski, P.A., Smith, P.R., Luryi, S., Cho, A.Y., Hamm, R., and Sivco, D.L. (1993) Microwave studies of self-aligned top collector charge injection transistor, *IEDM Tech. Dig.*, 423-426.

6. Lai, J.T. and Lee, J.Y. (1994) Enhancement of electron transfer and negative differential resistance in GaAs-based real-space transfer devices by using strained InGaAs channel layers, *Appl. Phys. Lett.* **76**, 1965-1967.

7. Mastrapasqua, M., Luryi, S., Belenky, G.L., Garbinski, P.A., Cho, A.Y., and Sivco, D.L. (1993) Multi-terminal light emitting logic device electrically reprogrammable between OR and NAND functions, *IEEE Trans. Electron Dev.* **40**, 1371-1377.

8. Mensz, P.M., Luryi, S., Bean, J.C., and Buescher, C.J. (1990) Evidence for real-space transfer of hot holes in strained GeSi/Si heterostructures, *Appl. Phys. Lett.* **56**, 2663-2665.

9. Zhou, G.L., Huang, F.Y., Fan, F.Z., Lin, M.E., and Morkoç, H. (1994) Observation of negative differential-resistance in strained n-type Si/SiGe MODFETs, *Solid-State Electronics* **37**, 1687-1689.

10. Mastrapasqua, M., King, C.A., Smith, P.R., and Pinto, M.R. (1994) Charge injection transistors and logic elements in $Si/Si_{1-x}Ge_x$ heterostructures, *IEDM Tech. Dig.*, 385-388.

11. Park B.G., King, C.A., Eaglesham, D.J., Sorsch, T.W., Weir, B., Luftman, H.S., and Bokor, J. (1993) Ultrashallow p^+-n junctions formed by diffusion from an RTCVD-deposited B:Ge layer, *Proc. of SPIE* **2091,** 122-131.

12. Pinto, M.R. (1991) Simulation of ULSI device effects, in Cellar, G. and Andrews, J. (eds.), *Electrochemical Society Proceedings* **91-11,** 43-51.

New Ideology of All-Optical Microwave Systems Based on the Use of Semiconductor Laser as a Down-Converter.

V. B. GORFINKEL,[*] M.I. GOUZMAN[**], S. LURYI[*] and E.L. PORTNOI[***]

[*] *State University of New York, Stony Brook, NY 11794-2350*
[**] *Bally Wulff GmbH, D-34125, Kassel, Germany*
[***] *Ioffe Institute, 194026, St. Petersburg, Russia*

Abstract

We propose a novel all-optical structure of phase lock loop for locking two semiconductor lasers with a stable microwave offset for use in phased-array antenna systems.

1. Motivation

Optical control of millimeter wave antenna arrays is an important technological challenge. This goal requires several key elements, currently under intense development worldwide. One of these elements is a dual optical beam source capable of producing a stable millimeter wave beat frequency when mixed at an antenna site. Such a source is important in the implementation of virtually any architecture for optical distribution of microwave phase between antenna array oscillators.

Development of a universal source based on semiconductor lasers faces several essential barriers. Semiconductor lasers have a relatively broad linewidth, and moreover the line center slowly drifts around due to environmental influences. This places a particularly stringent demand on the stabilizing phase lock loop that ties the dual beam difference frequency to a desired offset. Currently existing opto-electronical phase lock loops (OEPLL) based on electronic components (mixers, amplifiers, etc.), can be used for locking only modest microwave offset frequencies up to about 20 GHz [1]. This practical limit arises due to the unavoidable loop propagation delay in an electronic system (see Fig. 1, left). It would be highly desirable to implement a phase lock loop in which the intermediate frequency (IF) signal, corresponding to the deviation of the optical-beat frequency from the standard frequency of an offset generator, is itself coded onto an optical carrier. This would minimize the loop delay time and at the same time enable a dispersion-free and cross-talk-free delivery of the broadband IF feedback signal to the correcting block of the PLL (see Fig.1, right).

The key to a successful implementation of an all-optical PLL is our recent breakthrough in the conceptualization and experimental demonstration of a laser mixer which manipulates the microwave envelopes of optical carriers.

S. Luryi et al. (eds.), Future Trends in Microelectronics, 385–389.

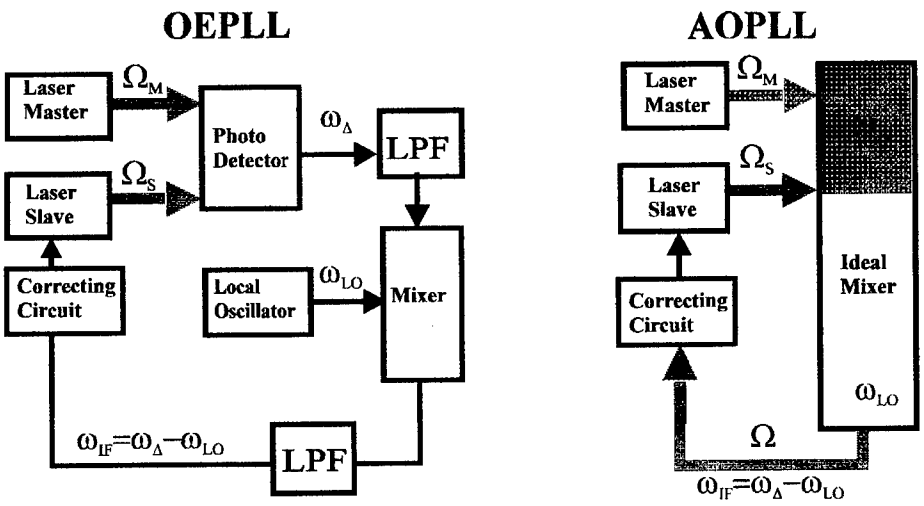


Fig. 1. From opto-electronic to all-optical PLL.
The ideal mixer combines the functions of a photodetector, a local oscillator and a mixer. Black and gray arrows correspond to electrical and optical signals respectively.

2. Laser Mixer

Recently, we proposed and experimentally demonstrated [2] a novel type of optoelectronic mixer, the laser-mixer. In this device the microwave signal is encoded in the form of an intensity modulation of the laser optical output. The input signals to be mixed can themselves be either lightwaves modulated in amplitude at mm wave frequencies or conventional microwaves (or both). The essential physics underlying the concept of the laser-mixer is external modulation of "material" parameters of the laser controlling the propagation of light in the laser cavity, such as the modal gain or optical losses. Consider the rate equation for photon density in the laser cavity:

$$\frac{dS}{dt} = S(\Gamma g - \alpha_{loss})$$

where g is the material gain, Γ is the confinement factor, α_{loss} is the cavity loss. Modulating two quantities, say, S and α_{loss}

$$S = S_0 + S_1 \sin\omega_1 t$$
$$\alpha = \alpha_0 + \alpha_1 \sin\omega_2 t$$

we obtain an optical response modulated at the sum and difference frequencies,

$$\omega = \omega_1 \pm \omega_2 .$$

Functionally, this physics is entirely analogous to that in conventional electronic mixers, where mixing is accomplished due to the variation of material parameters, controlling the electric output of the device (capacitance, resistance, etc.). The idea of a parametric control of laser output, combined with a pumping current variation, is relatively new [3,4]. Experimentally, an efficient parametric control was demonstrated in a three-terminal quantum-well laser structure, realized in two material systems: GaAs/AlGaAs at the Ioffe Institute and InGaAs/InGaP/GaAs at Bell Laboratories[5,6].

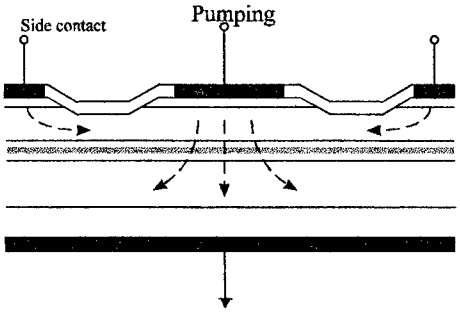


Fig. 2. Three-terminal laser structure with dual modal gain and pumping current control.

This three-terminal structure (Fig. 2) lends itself to a natural use as a laser-mixer with electric inputs. The electric impedance of the parametric input circuit, controlling the modal gain of this laser structure is 50 ohm, which is very attractive for microwave applications. Another experimentally demonstrated parametric laser-mixer [2] was based on the variation of cavity loss in a laser with saturable absorber (Fig. 3).

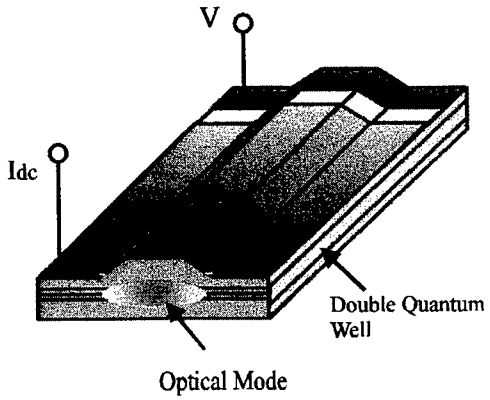


Fig. 3. Laser with saturable absorber in self-pulsating regime is capable of mixing the envelope w_D of two optical frequencies ($\omega_\Delta = \Omega_1 - \Omega_2$,) with its self-pulsation at frequency ω_{LO}. Resulting laser output at optical frequency Ω is modulated at difference frequency $\omega_{IF} = \omega_\Delta - \omega_{LO}$.

Because of the short (picosecond) recovery time available in saturable absorbers, this type of laser-mixer can take as input optical signals modulated in intensity at frequencies well over 100 GHz. Moreover, the optical nature of the input signal closes the entire loop in the optical medium and permits the implementation of all-optical circuits, operating entirely with microwave signals coded as the envelope of an optical carrier.

It should be clearly understood that parametric mixing in semiconductor laser is different in principle from the heterodyne coherent mixing of optical waves. The latter technique is based on the square-law dependence of the optical transition rate on the lightwave amplitude $|\mathbf{A}|$. In contrast, with respect to the lightwave intensity $|\mathbf{A}|^2$ the semiconductor laser is a highly linear element. Mixing of different channels in an optical communication system - which manifests itself as an unwelcome intermodulation distortion - is an exceedingly minor effect (typically less than -70 dBc) which arises mostly due to higher-order parametric phenomena [7].

3. Operation principle of the AOPPL

Let us return to the AOPLL structure shown in Fig. 1 and discuss its operation principle. Part of the radiation from both the master and the slave lasers is focused on the fast saturable absorber, modulating its transparency at the difference frequency

$$\omega_\Delta=\Omega_M-\Omega_S$$

and phase ϕ_Δ. This frequency is mixed with that of mode-locked pulsation in the laser cavity at the the latter frequency ω_{LO} can be as high as hundreds of GHz [2,3]. It may be further stabilized by an electronic generator. The intermediate frequency

$$\omega_{IF}= \omega_\Delta- \omega_{LO}$$

is then transmitted to the correcting circuit as an envelope of the optical output of the laser mixer. The inherited phase of the intermediate signal equals $\phi_{IF} =\phi_\Delta-\phi_{LO}$. This signal is received by a fast photodiode integrated in the correcting circuit, which tunes the slave laser so as to keep ϕ_{IF} constant.

As shown in Fig.1, the electric part of the PLL the correcting circuit is localized in the vicinity of the slave laser. The entire optical circuit can be integrated on a single silicon substrate. At the same time the AOPLL architecture permits a relatively remote positioning of the master and slave lasers, which is advantageous for their stability. We would like to emphasize the universality of the all-optical PLL architecture, rooted in the fundamentally dispersionless transmission of the microwave envelope via optical waveguides. Not only this implies the broad loop bandwidth but it also means that changing the operating microwave frequency does not entail any major revision of the passive optical circuit.

4. Conclusion

The laser-mixer whose optical output is modulated in amplitude at the intermediate frequency between two microwave or millimeter wave signals (themselves either optical or electrical) is a new element with a considerable potential for the implementation of optically interconnected systems operating in a very broad range of microwave frequencies. These systems may have a cascade structure and produce simultaneously microwave signals at multiple frequencies. As an example of the new possibilities, consider a system (Fig. 4.) generating a grid of equidistant optical

frequencies spaced apart by a given offset frequency. In this system the slave laser of

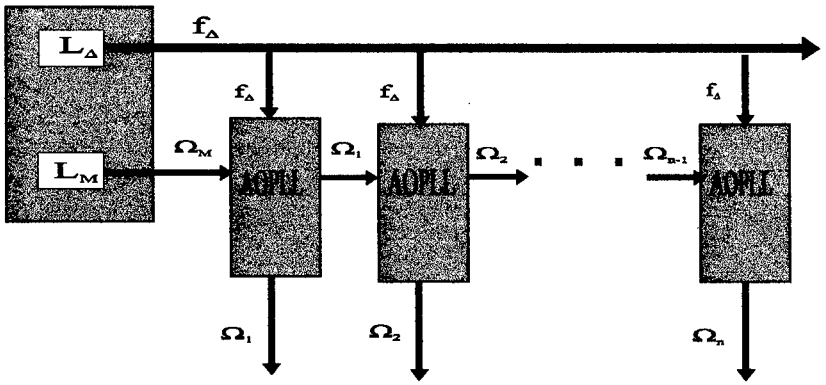


Fig. 4. Block diagrame of the source of multiple optical frequencies

the n-th AOPLL cascade plays the role of master for the following (n+1)-st cascade. A particular example of a system could be an array of optically coherent lasers operating at the same wavelength. In this case, an optical signal emitted by a single laser-master would be split between a number of AOPLL's which control the laser-slave array, thus keeping a given unique offset frequency between the laser-master and each of the laser-slaves.

References

1. Seeds, A.J. (1993) "Optical Technologies for Phased Array Antennas", *IEICE Trans. Electron.* **E76-C**, 198-206.
2. Portnoi E.L. , Gorfinkel, V.B. , Avrutin, E.A., Thayne, I., Barrow, D.A., Marsh, J.H. and Luryi, S (1995) "Optoelectronic Microwave-Range Frequency Mixing in Semiconductor Lasers", *IEEE J. Select. Top. Quant. Electron.*, **1**, 451-460.
3. Gorfinkel, V.B. and Luryi, S. (1994), "Dual modulation of semiconductor lasers", *Physics and Simulation of Optoelectronic Devices, ed. by M. Osinski, Proc. SPIE* **2146**, 204-209.
4. Gorfinkel, V.B. and Luryi, S. (1994), "Article that comprises a semiconductor laser, and method of operating the article", **US Pat. 5,311,526**.
5. Gorfinkel, V.B., Kompa, G., Novotny, M., Gurevich, S., Shtengel, G.,Chebunina, I. (1993), "High-frequency modulation of a QW diode laser by dual modal gain and pumping current control" *1993-IEDM Tech. Digest*, 933-937.
6. Frommer,A., Luryi, S., Nichols, D.T., Lopata, J. and Hobson, W.S. (1995) "Direct modulation and confinement factor modulation of semiconductor lasers", *Appl. Phys. Lett.*, **67** (Sept 18,).
7. Gorfinkel, V.B. and Luryi, S. (1995), "Fundamental limits for linearity of CATV lasers", *IEEE J. Lightwave Technol.*, 252-260.

MICROTECHNOLOGY - THERMAL PROBLEMS IN MICROMACHINES, ULSI & MICROSENSORS DESIGN[1].

Andrzej NAPIERALSKI
Division of Microelectronics & Computer Sciences
Institute of Electronics, Technical University of Łódź
Stefanowskiego 18/22, 90-924 Łódź, POLAND
e-mail: napier@j-23.p.lodz.pl

Abstract: In this paper some thermal problems related to modern microtechnology will be discussed. First example presents a design process of an integrated micropump. The general thermomechanical model is proposed. As the second example, the possible sources of heat generation in silicon micromotor will be discussed. The third example concerns the integrated CMOS thermal sensor designed in order to detect the high temperature rise in VLSI or Smart Power devices.

1. Introduction

The main object of this paper is to present the new problem in the modern microtechnology which is the high power dissipation density caused by the smaller dimensions of circuits and higher operating speeds. In the near future the thermal problems will be the bottleneck for integration in microelectronics. Actually in the modern silicon devices the thermal transient phenomena are as fast as electrical and the new methods for temperature computation must be developed. The transient heat equation can be presented in the following form:

$$\lambda \nabla^2 T = C_v \frac{\partial T}{\partial t} \tag{1}$$

where: T - temperature, t - time, λ - thermal conductivity,
Cv - the thermal capacity per unit volume (J/m^3K).

After introduction of scaling factor α, this equation can be presented under the following form:

$$\lambda\left[\frac{\partial^2 T}{\partial(\alpha x)^2}+\frac{\partial^2 T}{\partial(\alpha y)^2}+\frac{\partial^2 T}{\partial(\alpha z)^2}\right] = C_v \frac{\partial T}{\partial \alpha^2 t} \tag{2}$$

where: x, y, z - co-ordinates

[1]This work was supported by European Community Basic Research Project ESPRIT (no. 8173 - BARMINT), and European Community Project COPERNICUS (no. CP-940922 - THERMINIC)

S. Luryi et al. (eds.), Future Trends in Microelectronics, 391–396.

Therefore, if the dimensions are reduced by a factor α, the time scale has to be reduced by a factor α^2. If for example heat needs 1 second to spread over a 1 cm^3 silicon block, all transistors of 1 μm ($\alpha = 10^{-4}$) can be heated in 10^{-8} second (10ns). This is comparable to the actual electrical time constants in electronic VLSI circuits. In the case of 0.1 μm technology, the thermal time constants will be as small as 100ps and it will be absolutely necessary to consider thermal phenomena as equally fast as the electrical!

2. Thermomechanical design of silicon micropump

In modern micro-electronics, the thermal phenomena are the basis of the concept of many new micromachines. A good example is the silicon micropump. To make such a pump work, a periodic power pulse has to be put into the heating resistor. During a heating period a temperature inside the air cavity and the resulting pressure increase. When a value of the pressure is sufficiently high the membrane deflects and the outlet valve opens. Then the small volume of the fluid is removed from the fluid cavity. During a cooling process the membrane returns to its steady-state position closing output valve and opening the input one. A fluid gets into the fluid cavity. As it can be seen from this short description, the operating principles involve thermal, mechanical and also hydraulic phenomena. Therefore it is necessary to find out and model the mutual dependencies between all the physical variables in the system. This problem can be easily illustrated by the Fig. 1:.

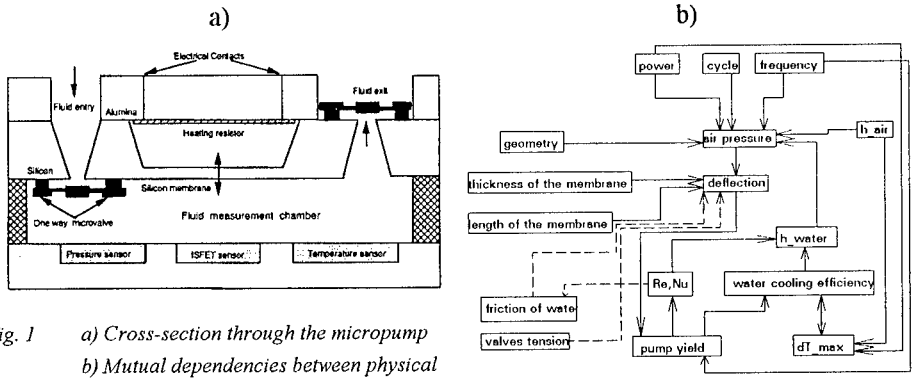


Fig. 1 *a) Cross-section through the micropump*
b) Mutual dependencies between physical variables in the micropump

As it can be seen there are plenty of mutual dependencies that require iterative solving and communication between different (thermal and mechanical) parts of the software. The exact modelling of this kind of micromachine is very complicated, and generally not possible by an analytical solution [6]. The CAD program has to take into account all thermal and mechanical properties of the designed structure.

3. Electromechanical and thermal modelling of IC processed micromotors.

As the second example of the importance of thermal considerations, the thermal model of the micromotor will be presented. As an example of case study, a complete thermal analysis of rotary IC-processed silicon micromotor, fabricated at MIT [1, 2, 3], will be presented. It is a variable-capacitance, radial-gap (side-drive) micromotor presented in Fig. 2. The basic dimensions are depicted in Fig. 3: the rotor radius R_0 is 50 μm, the gap between aligned rotor and stator pole G is 1.5 μm or 2.5 μm, the rotor and stator pole face thickness T is 2.2 μm. The motor design has 12 stator and 8 rotor poles. Every third stator pole is electrically connected to form three stator pole sets or phases, to which excitation voltages are supplied in a sequence.

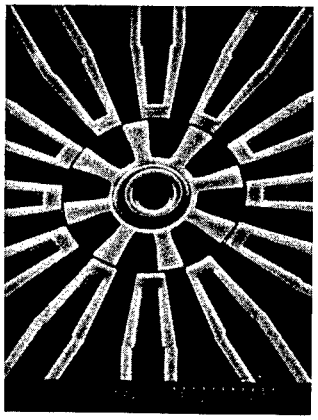


Fig. 2. A scanning electron micrograph of an IC-processed electrostatic micromotor, fabricated at MIT [2] (rotor diameter = 100 μm)

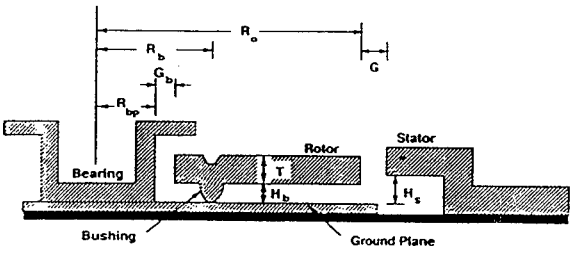


Fig. 3. a)Cross-section schematic of the micromotor from Fig. 2 [2]; b) details of acting fields and forces.

There are three possible sources of heat generation related to the micromotor movement:

1. side-pull friction , 2. bushing friction, and 3. viscous drag. The heat generation can be calculated assuming that all mechanical losses are converted into the heat.

$$P_{heat} = P_{mechanical} = T \bullet \omega = (T_{sp} + T_b + T_{visc}) \bullet \omega = P_{sp} + P_b + P_{visc} \qquad (3)$$

where: T - friction torque, ω - angular speed of rotor, P_{sp}, P_b, P_{visc} - the power components for side-pull friction, bushing friction and viscous drag respectively.

Using the *MICROMOTOR* program with the mathematical model of micromotor dynamics the maximum angular velocity of rotor has been calculated as: $\omega_{max} = 43\,633$ rad/s. The three power components for this maximum angular velocity have been calculated and they are respectively: $P_{sp} = 3.8\ 10^{-8}$, $P_b = 2.5\ 10^{-8}$, $P_{visc} = 4.2\ 10^{-7}$. As it can be seen from these results, the biggest amount of energy possibly converted into heat is lost as an effect of viscous drag.

To calculate the temperature rise in the motor volume, a the 3-D thermal model must be used. As input data for such model, spatial distribution of power generation must be provided. To perform the thermal analysis the 3-D RC equivalent network model of heat transfer has been applied [4, 5, 7]. The obtained temperature results show that the generated power values (eq. 3) cause only very minor increase of temperature, less than 0.1 K in any point of the device. As the result, friction heating does not appear as a dangerous phenomenon in the operation of such micromotors.

4. Integrated CMOS Thermal Sensors Design

One of the important question in the field of VLSI systems and power electronics is:

How to perform the thermal monitoring of the silicon wafer, containing semiconductor devices, in order to indicate the overheating situations?

First method consists of placement of many sensors everywhere on the chip, then their output can be read simultaneously and compared with the reference voltage recognised as the overheating level (Fig.4).

The idea of the next method, which is proposed in this paper, is to measure the temperature gradient along the given distance, in a few places only on the monitored wafer (Fig.5) and evaluate obtained information in order to achieve the temperature of the heat source [8]. This problem is known in the literature, especially in the field of modelling of the temperature distribution, as an *inverse problem*. It means that we know the results of investigated phenomena (e.g. values of the temperature in some places on the wafer's surface) and we try to find the parameters of the phenomena's source (e.g. the temperature of the heat sources). In general case the solution of the inverse problem requires powerful computations and not always gives the proper result. The reason is that a small inaccuracy in the initial conditions can cause unacceptable inaccuracy in the final result. In case of an IC or a power semiconductor device, there is no place on the layout for the complicated unit performing such computations, but there is also no need for it, as we want only detect the overheating situations. Moreover, in most cases, the overheating occurs only in one place.

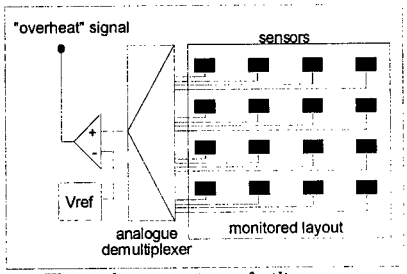


Fig.4. Thermal monitoring of silicon structure - method 1.

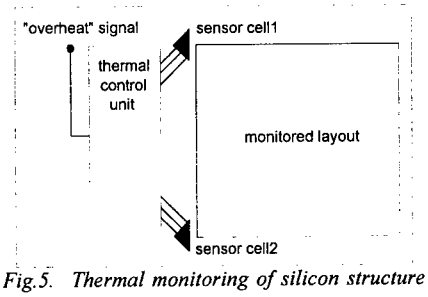


Fig.5. Thermal monitoring of silicon structure - method 2.

For the proposed method four assumptions have been made:

A1-there is only one punctual heat source on the monitored wafer.
A2-the temperature is linearly distributed over the surface of silicon wafer[2].
A3-temperature sensors used in this method give output voltage linearly proportional to temperature.
A4-the sensors in the sensor cell are placed sufficiently close one to another, that isotherms crossing them can be represented by straight lines.

Three sensors placed in distance a create the sensor cell (Fig.6). The actual value of the temperature gradient can be calculated if we know the output voltages from two sensors and the value of cosα. In order to obtain the information about angle α we introduce the third sensor. In order to obtain the temperature of a single, punctual heat source we have to calculate the distance between the sensor and the heat source. Two sensor cells are required for this purpose. The cells are placed in a given distance (H) and each of them gives the information about the angle α (α_1 and α_2) to the heat source (Fig.6). The heat source and cells form the triangle in which the length of one side and values of the angles adjacent to this side are known. This means that we can calculate the distances between the heat source and sensors. Now we can calculate the temperature gradient along the known distance. By adding it to the temperature of the sensor we obtain the temperature of the heat source.

The two sensor cells A_1,B_1,C_1 and A_2,B_2,C_2 can be placed in two corners of monitored layout in the distance H. The angle α of each sensor cell (Fig.7) covers as much of the monitored layout area as possible. The Fig.7 shows this arrangement. Only part of the layout in form of the isosceles triangle with the base of H and height of 0.3H is not monitored by the cells. Such configuration has been chosen for simplifying the control circuit.

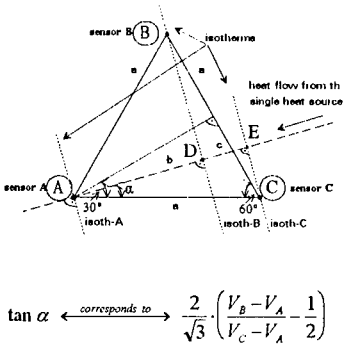


Fig.6 The idea of the sensor cell

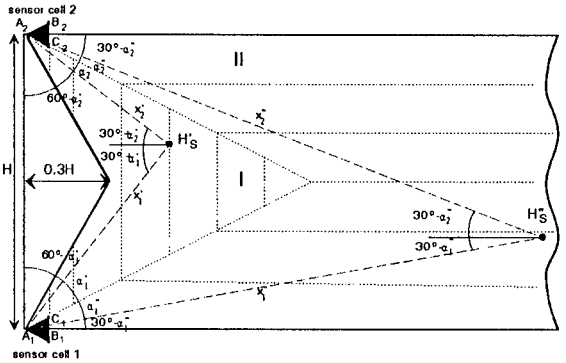


Fig.7 The proposed arrangement of two sensor cells .

[2] In this paper we will consider this assumption, however the function describing the real temperature distribution is not linear. For the given substrate one can find very precisely this function by application of the PYRTHERM software [9]. During calculations of the heat source temperature the found function can be transformed to the linear one.

The presented method can be very useful in the case of indication the overheating situations on the surface of an IC or other semiconductor devices. It can be implemented directly on the monitored surface without disturbing the integrity of the device and requires only a few sensor cells placed in any way out of the monitored area.

5. Conclusion

The main object of this paper was the presentation of the new problem in the modern microtechnology which is the high power dissipation density caused by the smaller dimensions of circuits and higher operating speeds. Actually in the modern silicon devices the thermal transient phenomena are as fast as electrical and the new methods for temperature computation must be developed. In this paper some thermal problems related to modern microtechnology have been discussed.

The three examples of the modern devices based on the thermal phenomenon have been presented. As the conclusion, one can say, that in the next future much more modern devices based on the thermal phenomenon will be certainly designed.

6. References

1. Bart, S.F. (1990) Modeling and Design of Electroquasistatic Microactuators, Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge,
2. Bart, S.F., Mehregany, M., Tavrow, L.S., Lang, J.H., Senturia, S.D. (1992) Electric Micromotor Dynamics, *IEEE Trans. Electron Devices*, **39**, 566-575
3. Mehregany, M., Senturia, S.D., Lang, J.H., Nagarkar, P. (1992) Micromotor Fabrication, *IEEE Trans. Electron Devices*, **39**, 2060-2069
4. Pacholik, J., Napieralski A., Grecki, M., Turowski, M. (1994) Temperature computation in semiconductor devices using 3D network concept, *Proc. of PEED, SPEEDAM'94*, Taormina, Italy, Appendix 1-4
5. Turowski, M., Jabłoński, G., Napieralski, A., Wiak, S. (1995) Electromechanical and Thermal Modelling of IC-processed Micromotors, *Proc. of IInd Workshop: MIXED-VLSI DESIGN*, Kraków, Poland, 343-348
6. Pacholik, J., Furmańczyk, M., Jabłonski, G., Turowski M., Grecki M., Napieralski A. (1995) Thermomechanical Design of Silicon Micropump", *Proc. of IInd Workshop: MIXED-VLSI DESIGN*, Kraków, Poland, 373 - 378
7. Furmańczyk M., Jabłoński, G., Napieralski, A., Pacholik, J. (1995) The 3D transient thermal simulation with arbitrary border conditions, *Proc. of International THERMINIC Workshop*, Grenoble 25-26/09/1995
8. Wójciak, W., Napieralski A. (1995) The distance dependent method for temperature measurement of single heat source in semiconductor devices - the first approach, *Proc. of International THERMINIC Workshop*, Grenoble 25-26/09/1995
9. Napieralski A., Dorkel J.M., Leturcq Ph.,: *"PYRTHERM - Manuel de l'utilisateur - Version 4.0 November 1990"*, rapport LAAS Nr. 90 380, (132 pages).

EMERGING AND FUTURE INTELLIGENT AVIATION & AUTOMOTIVE APPLICATIONS OF MIMO ASIM MACROCOMMUTATORS AND ASIC MICROCONTROLLERS

B.T. FIJALKOWSKI
Cracow University of Technology
Cracow, Poland
pmfijalk@cyf-kr.edu.pl

1. Introduction

Current trends in aviation and automotive macro-, meso-, micro- and nanoelectronics are towards intelligent electronically-commutated (reciprocational and rotational) electromechanical actuators (electrical machines) and other devices with much higher performance and greater compactness, not only at the interconnection stage. The increasing complexity of intelligent electronically-commutated electromechanical actuators (electrical machines) and other devices is leading to larger chip sizes and the greater degree of integration means that the power dissipated by the working chip is also raising. Increasingly, the tendency is package several chips together in the form of multichip module, using techniques such as wire bonding, tape automated bonding and flip chip bonding to reduce the interconnections length and hence the overall size.

Continuing demands for high-performance intelligent multi-input/multi-output (MIMO) macroelectronic AC-to-AC, AC-to-DC-to-AC, AC-to-DC/DC-to-AC, DC-to-AC and DC-to-AC-to-DC converter commutators, called the AC-to-AC, AC-to-DC-to-AC, AC-to-DC/DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutators, and microelectronic neuro-fuzzy (NF) computer (processor) controllers, called the NFmicrocontrollers are leading to integrated matrixers (IM) or application specific integrated matrixers (ASIM) and integrated circuits (IC) or application specific integrated circuits (ASIC) of far greater complexity than before. Artificial neural networks and fuzzy logic (NF) are increasingly being incorporated in MIMO control systems to provide robust and effective control in a wide range of aviation and automotive applications. The method is based on a many-valued logic which enables general principles and expert knowledge to be used to provide control rules and procedures. The intrinsic non-linearity and variability of operating conditions make NF control on ideal method for this area.

The author is actively pursuing the integration of matrixery and circuitry on the same transparent substrate as the intelligent MIMO ASIM AC-to-AC, AC-to-DC-to-AC,

S. Luryi et al. (eds.), Future Trends in Microelectronics, 397–405.

AC-to-DC/DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutators and ASIC NF microcontrollers, respectively. The most obvious advantage of integrating matrixery and circuitry on the same transparent substrate is the reduction in the 'work-horse' components that are mono- and polycrystalline as well as amorphous-Si or GaAs super- and/or semiconductor, fast-switching 'discrete' and/or 'continuous' uni- and/or bipolar electrical valves [diodes, transistors, charge injection transistors (CHINTs), CMOS-transistors, heterojunction bipolar transistors (HBTs), high electron mobility transistors (HEMTs), insulated gate bipolar transistors (IGBTs), MES field effect transistors (MESFETs), metal insulator-semiconductor FETs (MISFETs), MOSFETs, organic FETs (OFETs), bipolar quantum interference transistors (QUITs), static induction transistors (SITs), bipolar superconductor-base hot electron transistors (SUPER-HETs), gate-turn-off (GTO) thyristors, light-triggered-and-quenched (LTQ) thyristors, MOS-controlled thyristors (MCTs), ovonics etc], and in external connections.

The 'discrete' electrical valves suffer from having a performance that is strongly dependent upon temperature. The author have therefore carried out a detailed programme of research into ASIMs using 'continuous' electrical valves, in which electrical conduction is much less temperature-dependent than in 'discrete' electrical valves. The external connections are also one of the dominating causes of ASIM and ASIC failure. ASIMs and ASICs will be lighter and more compact because a proportion of the size and mass of them. It is difficult to estimate the cost and reliability of these ASIMs and ASICs as they are not yet in manufacture. Although research, development and initial manufacture costs are high, the leading electronic manufacturers will be investigating significantly in this key technology and the intelligent MIMO ASIM AC-to-AC, AC-to-DC-to-AC, AC-to-DC/DC-to-AC , DC-to-DC and DC-to-AC-to-DC macro-commutators capable of true inaudible (> 20 kHz) operation, and ASIC NF micro-controllers will soon extend their application from the current aviation and automotive intelligent electronically-commutated electromechanical actuators (electrical machines) and other devices markets in to other areas and they have every chance of completely revolutionizing the ASIM world.

The last few year have witnessed a rapid progress in macroelectronics (integrated high-power electronics), mesoelectronics (integrated medium-power electronics), micro-electronics (integrated low-power electronics) and nanoelectronics (integrated ultra-low-power electronics), above all when it comes to the development that have taken place in the field of avionics (aviation electronics) and automotronics (automotive electronics). New applications for macro-, meso-, micro- and nanoelectronics have thus evolved and established applications have undergone further advancements.

In 1973 I have forecasted a potential revolution in in applications of macro- and microelectronicss as high-performance intelligent MIMO mono- and/or polycrystalline as well as amorphous super- and/or semiconductor ASIM AC-to-AC, AC-to-DC-to-AC, AC-to-DC/DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutators (conceived and widely popularized by the author in the papers presented at the 1st Nat. RAILWAY VEHICLES Conference, Krakow-Zawoja, Poland, October 1973; and the First European Conference on POWER ELECTRONICS AND APPLICATIONS in Brussels, Belgium, October 1984), and ASIC NF microcontrollers improve the economics of reshaping

electric power to accomplish various tasks not only in aviation and automotive sectors but also in other applications.

Intelligent MIMO ASIM macrocommutators are the key to overcome the contactless (brushless), ie sparkless commutation (changing over the way of electric-current flow), and so to the static conversion of one kind of electrical energy into another, with the application of the phenomena inducing uncontrolled or controlled electric current conductivity (carrying of electric charges by positive or negative carriers into a definite medium under the action of electric field).

Intelligent MIMO ASIM macrocommutators have announced the era of newly designed energy-saving AC- and/or DC-powered elctromechanical actuators (electrical machines), enabling continuous (stepless) and contactless control of their torque and velocity or angular speed (during motoring), or their voltage and current (during generating).

Intelligent MIMO ASIM macrocommutators with uni- and/or bipolar commutating ASIMs constitute a successive breakthrough in the develiopment of AC- and/or DC-powered electronically-commutated electromechanical actuators (electrical machines), and multiply their possibilities enormously. There exist fantastic perspectives which are probable to become reality as soon as at the end of this century. But not in this does the scientific sense of concepts conceived by the author over twenty years ago consist.

The most essential value of this new concept was the achievement of cardinal breakthrough in the thinking itself, on the static conversion of one kind of electrical energy into another, with the aid of 'discrete' and/or 'continuous' uni- and/or bipolar electrical valves - a change of paradigm, that was a kind of thinking standard in the field of power electronics.

2. Intelligent MIMO ASIM Macrocommutators

In 1820, Le Marquis De Laplace in his work: *"Theorie Analytique des Probabilities"* written: "An intelligent being who knew a given instant all forces by which nature is animated and possessed complete information on the state of matter of which nature consists provided his mind were powerful enough to analyse these data - could express in the same equation the motion of the largest bodies of the universe and the motion of the smallest atoms. Nothing would be uncertain for him, and he would see the future as well as the past at glance." This real world's image had great impact in the systems approach (systems thinking) in the 1980s. Nowadays, scientists believe that in nature chaos develops in most phenomena and the future is not unpredictable in the long term. Controlling chaos has major importance from the scientific point of view.

Scientific revolution in the field of power electronics has been taken place very rapidly - due to the overcoming of subsequent stereotypes of ideas valid in this field of knowledge. All the progress is stimulated by thinking free from dogmas opposed to the indiscriminate approach to existing concepts. It is also well known that the most perfect practical solutions are born out of ingenious theory. "There is nothing more practical than a good theory" - said once L Boltzmann.

The concept of an ASIM macrocommutator with commutating ASIM, realized on 'discrete' and/or 'continuous' uni- and/or bipolar electrical valves was born due to the mathematical matrix notation of systems approach (systems thinking) on dynamics for the formation methodics of mathematical models of AC- and/or DC-powered electronically-commutated electromechanical actuators (electrical machines).

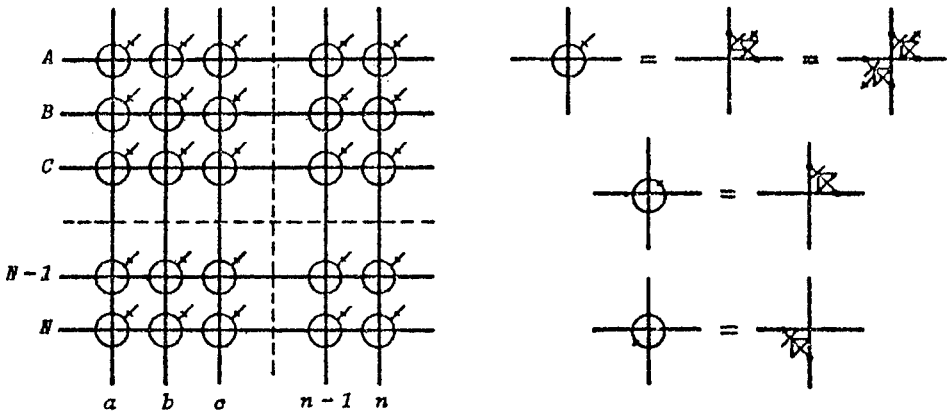


Figure 1. A generalized intelligent MIMO AC-to-AC, AC-to-DC-to-AC, AC-to-DC/ DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutator with an optimal commutating non-integrated matrixer or ASIM (Generalized physical model)

The super- and/or semiconductive ASIM macrocommutator is a macroelectronic valvular-converter commutator, being at the same time the changer of voltage, frequency and number of phases - that is a static unit for converting one kind of electrical energy (eg, with the voltage - u^P, frequency - f^P and number of phases - m^P) into another (eg, with the voltage - u^s, frequency - f^s and number of phases - m^s), performing the definite matrix function, and having at least one commutating non-integrated or integrated matrixer, in which the finite number of 'discrete' and/or 'continuous' uni- and/or bipolar electrical valves (switches) is inseparately connected in a solid (continuous medium) or on its surface. The ASIM macrocommutator being made in this manner can be neither changed nor repaired.

In Figure 1 there is a matrix diagram of the generalized valvular-converter commutator with an optimal commutating non-integrated matrixer or ASIM, containing among the others also an ASIM macrocommutator with a commutating ASIM, realized on 'discrete' and/or 'continuous' uni- and/or bipolar electrical valves.

In macroelectronics the ASIM macrocommutator is a simple integral valvular-converter commutator just like in microelectronics - ASIC microcomputer or microprocessor, and in electronics - a diode, triode (transistor, thyristor etc) or tetrode (spacistor, tetristor etc) with respect to manufacturing manner hybrid and monolithic ASIM macrocommutators can be distinguished.

Assuming that extortions of an ASIM macrocommutator are given in a form of the matrices of voltage sources between the primary (input) and secondary (output) collectors (Fig. 2), the set of differential equations of dynamics can be expressed as follows:

$$\begin{bmatrix} E_{AN} \\ E_{BA} \\ E_{CB} \\ \cdot \\ E_{N(N-1)} \\ E_{an} \\ E_{ba} \\ E_{cb} \\ \cdot \\ E_{n(n-1)} \end{bmatrix} = \begin{bmatrix} Z_{AA} & Z_{AB} & Z_{AC} & \cdot & Z_{A(N-1)} & Z_{AN} & Z_{Aa} & Z_{Ab} & Z_{Ac} & \cdot & Z_{A(n-1)} & Z_{An} \\ Z_{BA} & Z_{BB} & Z_{BC} & \cdot & Z_{B(N-1)} & Z_{BN} & Z_{Ba} & Z_{Bb} & Z_{Bc} & \cdot & Z_{B(n-1)} & Z_{Bn} \\ Z_{CA} & Z_{CB} & Z_{CC} & \cdot & Z_{C(N-1)} & Z_{CN} & Z_{Ca} & Z_{Cb} & Z_{Cc} & \cdot & Z_{C(n-1)} & Z_{Cn} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ Z_{NA} & Z_{NB} & Z_{NC} & \cdot & Z_{N(N-1)} & Z_{NN} & Z_{Na} & Z_{Nb} & Z_{Nc} & \cdot & Z_{N(n-1)} & Z_{Nn} \\ Z_{aA} & Z_{aB} & Z_{aC} & \cdot & Z_{a(N-1)} & Z_{aN} & Z_{aa} & Z_{ab} & Z_{ac} & \cdot & Z_{a(n-1)} & Z_{an} \\ Z_{bA} & Z_{bB} & Z_{bC} & \cdot & Z_{b(N-1)} & Z_{bN} & Z_{ba} & Z_{bb} & Z_{bc} & \cdot & Z_{b(n-1)} & Z_{bn} \\ Z_{cA} & Z_{cB} & Z_{cC} & \cdot & Z_{c(N-1)} & Z_{cN} & Z_{ca} & Z_{cb} & Z_{cc} & \cdot & Z_{c(n-1)} & Z_{cn} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ Z_{nA} & Z_{nB} & Z_{nC} & \cdot & Z_{n(N-1)} & Z_{nN} & Z_{na} & Z_{nb} & Z_{nc} & \cdot & Z_{n(n-1)} & Z_{nn} \end{bmatrix} \begin{bmatrix} i^{AN} \\ i^{BA} \\ i^{CB} \\ \cdot \\ i^{N(N-1)} \\ i^{an} \\ i^{ba} \\ i^{cb} \\ \cdot \\ i^{n(n-1)} \end{bmatrix} \quad (1)$$

Coefficients

$$Z_{kl} = \frac{\partial E_{k(k-1)}}{\partial i^{l(l-1)}} \qquad (i^{i(i-1)} = 0;\ i \neq l) \qquad (2)$$

have the dimension of impedance.

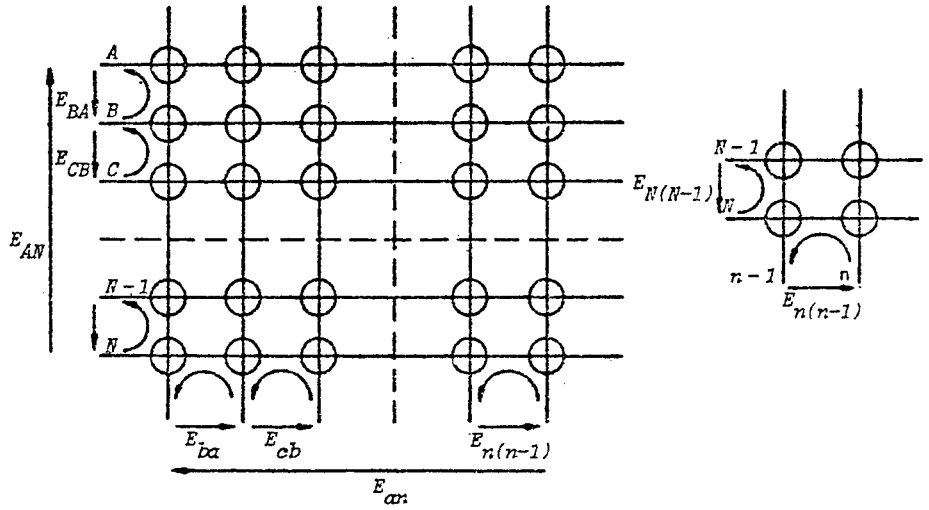


Figure 2. A generalized intelligent MIMO AC-to-AC, AC-to-DC-to-AC, AC-to-DC/ DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutator with voltage sources (Physical model I)

Assuming that extortions of an ASIM macrocommutator are given in a form of the matrices of current sources flowing in the primary (input) and secondary (output) collectors (Fig. 3), the set of differential equations of dynamics can be expressed as follows:

$$\begin{bmatrix} I^A \\ I^B \\ I^C \\ . \\ I^N \\ I^a \\ I^b \\ I^c \\ . \\ I^n \end{bmatrix} = \begin{bmatrix} Y^{AA} & Y^{AB} & Y^{AC} & . & Y^{A(N-1)} & Y^{AN} & Y^{Aa} & Y^{Ab} & Y^{Ac} & . & Y^{A(n-1)} & Y^{An} \\ Y^{BA} & Y^{BB} & Y^{BC} & . & Y^{B(N-1)} & Y^{BN} & Y^{Ba} & Y^{Bb} & Y^{Bc} & . & Y^{B(n-1)} & Y^{Bn} \\ Y^{CA} & Y^{CB} & Y^{CC} & . & Y^{C(N-1)} & Y^{CN} & Y^{Ca} & Y^{Cb} & Y^{Cc} & . & Y^{C(n-1)} & Y^{Cn} \\ . & . & . & . & . & . & . & . & . & . & . & . \\ Y^{NA} & Y^{NB} & Y^{NC} & . & Y^{N(N-1)} & Y^{NN} & Y^{Na} & Y^{Nb} & Y^{Nc} & . & Y^{N(n-1)} & Y^{Nn} \\ Y^{aA} & Y^{aB} & Y^{aC} & . & Y^{a(N-1)} & Y^{aN} & Y^{aa} & Y^{ab} & Y^{ac} & . & Y^{a(n-1)} & Y^{an} \\ Y^{bA} & Y^{bB} & Y^{bC} & . & Y^{b(N-1)} & Y^{bN} & Y^{ba} & Y^{bb} & Y^{bc} & . & Y^{b(n-1)} & Y^{bn} \\ Y^{cA} & Y^{cB} & Y^{cC} & . & Y^{c(N-1)} & Y^{cN} & Y^{ca} & Y^{cb} & Y^{cc} & . & Y^{c(n-1)} & Y^{cn} \\ . & . & . & . & . & . & . & . & . & . & . & . \\ Y^{nA} & Y^{nB} & Y^{nC} & . & Y^{n(N-1)} & Y^{nN} & Y^{na} & Y^{nb} & Y^{nc} & . & Y^{n(n-1)} & Y^{nn} \end{bmatrix} \begin{bmatrix} u_A \\ u_B \\ u_C \\ . \\ u_N \\ u_a \\ u_b \\ u_c \\ . \\ u_n \end{bmatrix} \tag{3}$$

Coefficients

$$Y^{kl} = \frac{\partial I^k}{\partial u_l} \quad (u_i = 0;\ i \neq l) \tag{4}$$

have the dimension of admittance.

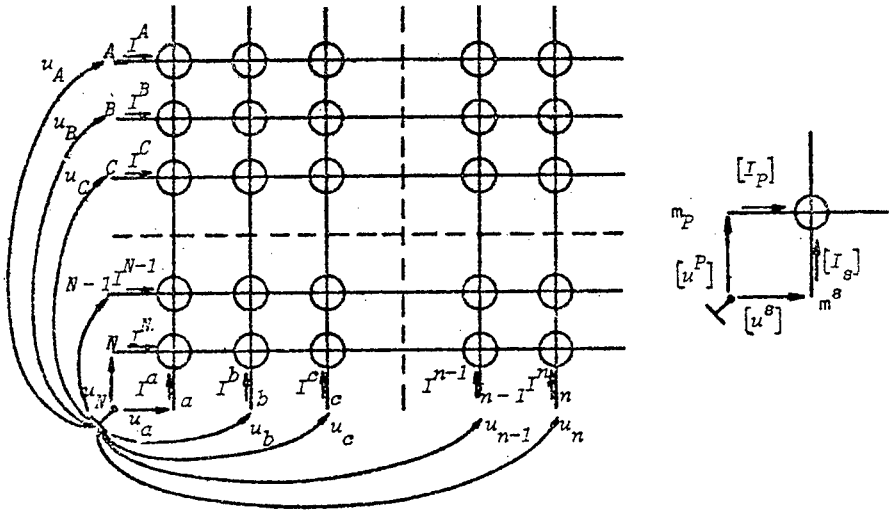


Figure 3. A generalized intelligent MIMO AC-to-AC, AC-to-DC-to-AC, AC-to-DC/ DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutator with current sources (Physical model II)

4. Conclusions

The hybrid ASIM macrocommutator with a commutating ASIM realized on 'discrete' and/or 'continuous' uni- and/or bipolar electrical valves (switches), established a mediate element amidst the poly-valvular electrocommutator

with a commutating non-integrated matrixer, realized also on 'discrete' uni- and/or bipolar electrical valves, and the monolithic ASIM macrocommutator with a commutating ASIM, realized on 'continuous' uni- and/or bipolar electrical valves. Reactive components of commutating ASIM, and so primary (input) and secondary (output) collectors, ie the connections between 'discrete' uni- and/or bipolar electrical valves are made on ceramic or glassy base with the vacuum evaporation, printing or injection casting technique. Active components of commutating ASIM, and so 'discrete' uni- and/or bipolar electrical valves: diodes, triodes (transitors, thyristors etc) and tetrodes (spacistors, tetristors etc) without package are soldered or pressed into the commutating ASIM manufactured in this manner.

The hybrid ASIM macrocommutator, especially made in the thick-film technique, has a number of essential advantages:

- Not-too-long time of designing and preparation of production)\;
- Moderate cost of production, even in small serries;
- Possibility of integration of 'discrete' uni- and/or bipolar electrical valves and protection circuits;
- High overcurrent capability;
- High overvoltage capability;
- High output power.

By way of example the 'rotative' ASIM macrocommutators, designed to be built into the rotor of the AC-to-DC macrocommutator mechanoelectrical generator, which work in the most severe conditions are nowadays, made most frequently in the hybrid technique.

The monolithic ASIM macrocommutator with a commutating ASIM, realized on 'continuous' uni- and/or bipolar electrical valves (switches) is a macroelectronic converter commutator, in which all active components ('continuous' uni- and/or bipolar electrical valves) are made in a single wafer of the crystalline or amorphous material in the shape of single chips. This technique has assured high packing density of active components. According to the number of individual single active components integrated on the single wafer of crystalline or amorphous material, commutating ASIMs with a small scale of integration comprising several to 36 'continuous' uni- and/or bipolar electrical valves, or commutating ASIMs with a medium scale of integration from 24 to hundreds 'continuous' uni- and/or bipolar electrical valves can be distinguished.

The monolithic ASIM macrocommutator with the commutating ASIM, realized on semiconductor mono-crystalline 'continuous' uni- and/or bipolar electrical valves used in the static converter technique and solid-state commutator electromechanical actuators (electrical machines) shows so far series of limitations, namely:

- Difficulties with the integration of 'continuous' uni- and/or bipolar electrical valves;
- Limited resistance against overvoltages;
- Low cost only for series production on a large scale (100,000 of commutating ASIMs yearly);
- Relatively low output power.

Recently in hybrid and monolithic ASIM macrocommutators two basic types of commutating ASIMs can be used: unipolar and bipolar.

For over twenty years the author has been applying full-diffusion technology with loose mono- and/or polysilicon wafer for manufacturing MOS-controlled thyristor (MCT) type ASIMs for intelligent MIMO ASIM AC-to-AC, AC-to-DC-to-AC,

AC-to-DC/DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutators. Today, neutron-doped silicon of the so-called FZ (Float-Zone) type is used mainly. Neutron doping means that the silicon rod is irradiated with neutrons in a nuclear reactor prior to being sliced into silicon wafers. A number of silicon atoms are then converted into phosphorus atoms and the and the material becomes weakly n-doped. The a major advantage, however, is that the silicon rod after heat treatment will have a very homogenenous resistivity. This is valid for achieving a high-voltage capability of the ASIMs.

To obtain an acceptable yield in the ASIM manufacturing, ie, the number of MCT type ASIMs fulfilling the specification in relation to the number of MCT type ASIMs manufactured, extreme cleanless requirements are made on the environment.

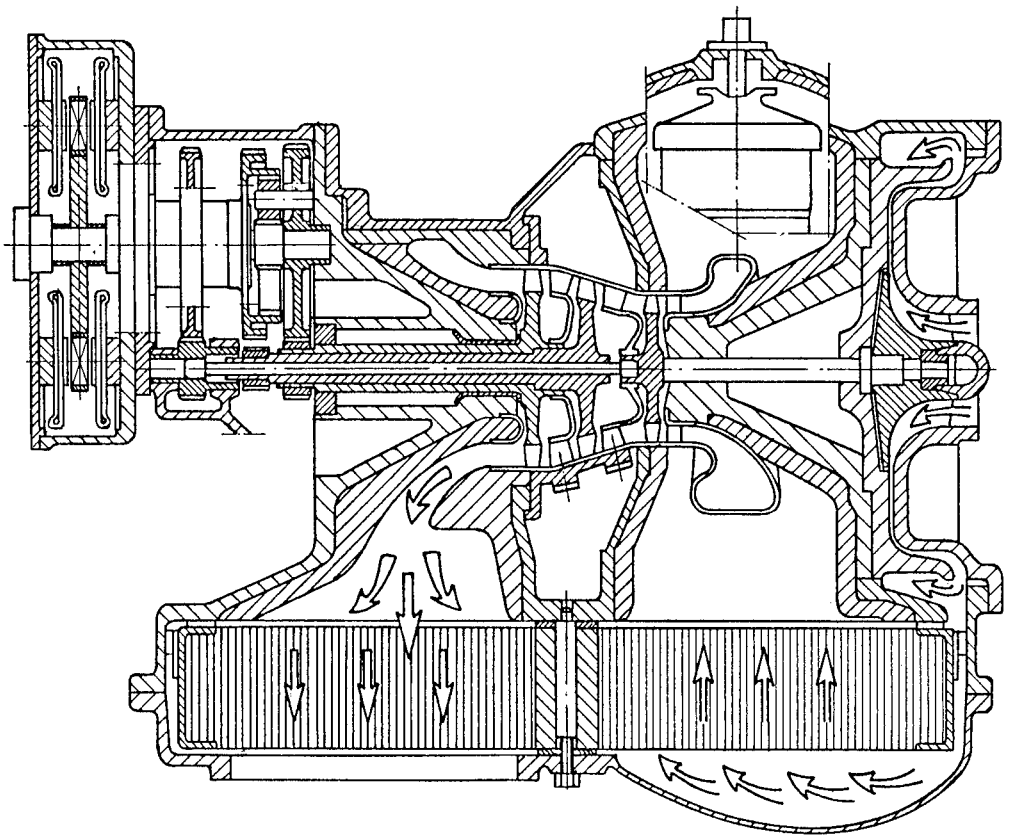
Diffusion technology means that the silicon wafer is coated with galium, aluminum or phosphorus. These substances are then allowed to diffuse into the silicon wafer in diffusion furnaces under high temperature. In certain ASIM manufacturing stages only parts of the silicon wafer surface are coated for subsequent diffusion. For this purpose, the photolitography is utilized. After the completion of the diffusion and the coating of an aluminum contact, the silicon wafer is sliced into ASIM chips with the aid of a high-power laser.

Depending on the silicon wafer area of the chip, the intelligent MIMO ASIM AC-to-AC and DC-to-AC macrocommutators, eg, for AC-to-AC macrocommutator electrical dynamotors (generators/motors) incorporated 3×3 ASIMs, and DC-to-AC macrocommutator electrical dynamotors incorporated 2×3 ASIMs.

By controlling the life time of the charge carriers, it is possible to determine the position of the MCT type ASIM on the recovery charge - minimum on-state voltage drop curve. For this purpose, the gold irradiation is applied, which gives considerably closer tolerances. With electron irradiation the ASIM elements are bombarded with high-energy electrons ($>$ 10 MeV) from electron guns.

On the basis of the large ASIM manufacturing volume of primarily unipolar diode or thyristor type ASIMs for MIMO ASIM AC-to-AC, AC-to-DC-to-AC, AC-to-DC/ DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutators that occurred at the beginning of the 1980s, it has been possible to introduce more reliable dimensioning criteria. In combination with closer ASIM manufacturing tolerances, this has substantially raised above all the maximum reverse voltage/off-state voltage. Lowering the on-state voltage drop, thus reducing the on-state charge and con-sequently high power losses in protective circuits around the intelligent MIMO ASIM AC-to-AC, AC-to-DC-to-AC, AC-to-DC/DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutators. Besides, a conception of bipolar commutating ASIM, realized on 'continuous' bipolar power superconductor-base hot electron transistors (SUPER-HETs) on the base of GaAs/Nb/InSb; 'continuous' bipolar power quantum interference transistors (QUITs). The bipolar commutating ASIMs of that type can be used, owing to specific conditions of the electric-current super- and/or semiconductivity, as basic trivalent-logic functors of the superconductive ASIM macrocommutator, which seems to be the simplest and most effective static converter for conversion of one kind of electrical energy into another. This trivalent logic is based on the Lukasiewicz-Tarski algebra.

Semi- and/or superconductive ASIM AC-to-AC, AC-to-DC-to-AC, AC-to-DC/ DC-to-AC, DC-to-DC and DC-to-AC-to-DC macrocommutators are nowadays indispensable in modern electromechanical actuator engineering and for energy converting purposes, especially in aviation and automotive sectors.



Trends in Thermal Management of Microcircuits

Vladimír Székely and Márta Rencz
Dept. of Electron Devices, Fac. of Electr. Eng. & Infor., TU Budapest, Goldmann György tér 3. H-1521 Budapest XI., Hungary

and

Bernard Courtois
TIMA, 46 Avenue Fèlix Viallet, 38031 Grenoble cedex, France

1. Introduction

With silicon microtechnology we intend to realize electrical networks: these are the integrated circuits. This goal however can never be obtained solely – a thermal network is also generated necessarily. The electrical parts dissipate heat, this will be the source of the thermal network. As a result the temperature of the chip will increase, changing the electrical parameters. In some cases this can even result in burning out the elements. With the decreasing chip feature sizes and package dimensions, with the increasing integration density the heat production per unit volume increases – continuously enlarging the severity of these problems.

That is why during the design of an integrated circuit concentrating only to the electrical operation is not sufficient. We have to keep in hand the thermal network as well, overheating and electrical/thermal cross couplings have to be impeded. For that we need methods to calculate and measure the temperature distribution on the chip. In some cases even continuous monitoring of the temperature distribution is necessary.

It is interesting to note that thermal effects are not always only problem sources, sometimes they are exploited to realize special functions.

2. Thermal considerations in the IC design phase

During the design of electrical circuits thermal considerations have continuously to be kept in mind. This is why it is an important requirement for the microcircuit design program systems to be equipped with thermal simulators. The simplest such programs calculate the surface temperature distribution of the chip, considering isotherm platform. This is sufficient to find the so called hot spots. These are such regions of the chip where there are numerous strongly dissipating elements resulting in extremely high local temperature. By rearranging the circuit elements we have to find such a placement where the temperature increase is more uniform and hot spots are avoided.

S. Luryi et al. (eds.), Future Trends in Microelectronics, 407–412.

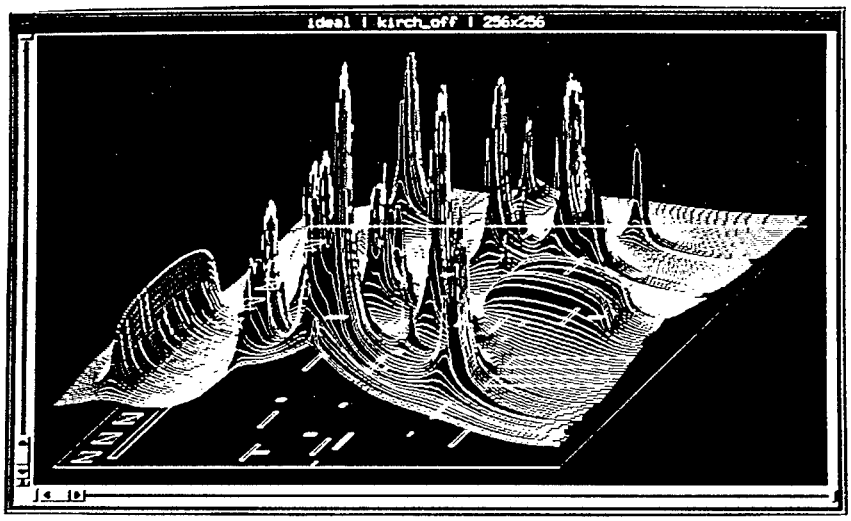


Fig. 1. Simulated surface temperature distribution of an IC. On the floor of this pseudo-3D image the chip layout is shown. The curved surface represents the temperature distribution: the height of this surface above each point is proportional to the temperature rise of that point

Fig. 1. shows a temperature map calculated by such a thermal simulation program.

It is often necessary to calculate the thermal distribution of more complicated structures, considering the chip, the platform, the package with different geometry and material parameters. For such cases FEM (finite element method) programs are usually used, (e.g. such as ANSYS [1]).

Between the electrical and the thermal network couplings may exist. In the case of an analogue power amplifier as an example, the temperature of the entire chip, including the input stage, changes corresponding to the dissipation of the output stage. Since the transistors of the input stage are temperature dependent, this acts as an additional control. Already a change of some thousandth of centigrade on the input transistors results in considerable feedback and consequently in the altered behaviour of the circuit. Because of the extremely small sizes, this effects are considerable on relatively high frequencies, already at 10 or 100 kHz.

To examine the electrical/thermal coupling effects, we need such a simulator which can consider the effects of thermal coupling [2]. The appearance of extensions to the well known circuit simulation programs, which make them appropriate for coupled electrical/thermal simulation, can be noticed.

3. Package design and the thermal problems

In IC packaging DIL-like solutions were used for about twenty years. In the last decade however new solutions are intensely investigated (QFP, PGA, flip-chip, tape-bond, solder-bump, MCM, Cubic structures etc.). In all these the two main goals are

- the highest possible pin number on the highest possible raster density,
- the best heat conductance.

We meet extreme dissipation values (e.g. $2000\,W$ from a single MCM) and unusual solutions for cooling (water cooling package etc.) [3]. Heat removal often occurs directly from the active face to avoid even the thermal resistance contribution of the die itself. 3D packaging result in higher element density, consequently even more serious thermal problems.

4. Thermal transient recording

Thermal transient recording methods gain increasing role in the thermal investigation of IC structures and their packaging. In this method the thermal characteristics of the structure is measured by a thermal step response recording. A typical step-response function is shown in Fig. 2. This describes both the static and the dynamic thermal behaviour. Many research papers deal with the problem of finding methods for not only to indicate but also to locate heat sinking defects based on these measurements. As a result of subsequent data processing the heat flow map of the packaging structure can be obtained, on which the individual sections of the heat sinking path can be separated and the possible weak points or anomalies can be located.

5. Obtaining surface temperature map

In the case of thermally critical designs the temperature distribution has to be measured on some experimental samples. This needs thermal measurements with some tenth of centigrade accuracy on the surface with a resolution of some micrometers.

The classic method for this purpose is the infrared thermography, which gives the thermal map of the surface with the speed of a TV image [4]. On the other hand this method is clumsy and expensive, the infrared detector has to be strongly cooled, and the scanning is usually mechanical. The verification of the temperature scale gives notable difficulties.

In these years some methods based on new principles are investigated. In one of these methods the IC surface is covered by liquid crystal and the temperature

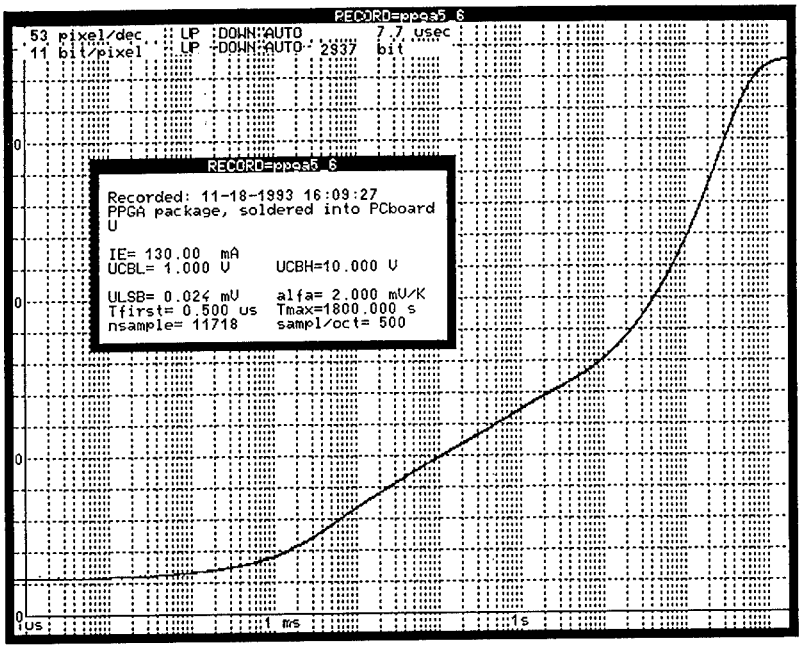


Fig. 2. Thermal step-response function of an IC package. The x-axis is the time, and the y axis is the temperature rise. Note, that the response has been obtained for an extremely wide time range, from $1\,\mu s$ to $1000\,s$. The thermal time constants of the system are expected to be distributed in this range

dependent phase transition is observed. The thermal map is obtained by computer aided image acquisition and processing methods. In an other method the surface is covered by fluorescent material, the emission of which is thermal dependent, so that after UV exposure the intensity of the fluorescentia draws a thermal map. A further interesting method exploits the temperature dependence of the surface optical reflectance. Scanning the surface with a laser beam the reflection rate is temperature dependent – however only weakly, and the dependence is influenced by the material as well.

Fig. 3. shows a temperature map obtained by liquid crystal measurement and subsequent image processing.

6. Using built-in sensors: thermal monitoring

In the practice of the integrated circuit design it is usual to add some excess circuitry serving to support the design for testability (DfT). This method can be

Fig. 3. Measured temperature map of a small region on the IC surface. The result is obtained by liquid crystal method. The different colors (different shadings on this b-w picture) correspond to regions of different temperatures. The temperature step between two neighboring regions is 0.1 °C

extended for thermal testing. Recent experiments deal with the question of building one or more thermal sensors into the chip and use them for thermal testing.

Chips supplied with thermal sensors can be thermally tested after fabrication (DfTT, Design for Thermal Testability) [5]. With some additional circuitry the temperature of the IC is readable also if the circuit is already built in an equipment. This way the inner temperature of the IC can be controlled in the operating equipment, preventing overheating and degradation. This is the basic principle of thermal monitoring. The current problem of development is to find how can be this problem solved with minimal extra expenses (chip area, excess pins). The best solution seems to be the connection with the other electrical test circuits (e.g. boundary scan) [6].

7. Thermal stabilization and protection

In the case of delicate circuits the parameters can be improved if the circuit is kept on constant temperature in a thermostat. The circuit itself can be designed

such that, it behaves as a kind of a micro thermostat. For this purpose a thermal sensor and a dissipator are required on the chip surface, with some control circuitry. The latter controls the dissipator to assure constant (e.g. 80 °C) chip surface temperature. Such circuits are called TSS (temperature-stabilized surface) ICs.

In the case of higher dissipation circuits protection against overheating can be useful. A temperature sensor observes the chip temperature and inhibits the operation in the case of overheating, or reduces the dissipation e.g. in the case of CMOS VLSI circuits by decreasing the clock frequency.

8. Thermal functional units

Exploiting the thermal effects different functional units can be realized in an integrated circuit. As an example a dissipating resistor and a temperature gradient sensor placed beside it form a true RMS (root-mean-square) measuring unit. A dissipating element and the thermal sensor placed close to it can be used as a flow meter. Based on similar principle the microelectronic version of the Pirani vacuum meter can be constructed. Infrared sensor can be built by dropping IR radiation on a (from the ambient thermally well isolated) target and measuring the temperature increase of the target [7]. It has to be noted however that for the fabrication of most of these units excess technological steps are required beyond the usual IC process steps. At some places of the chip thin membranes or cantilevers have to be fabricated by etching, new surface layers of special material are needed etc. These latter however belong already to another category: to the integrated microsystems.

References

1. ANSYS – Engineering Analysis System: Theoretical Manual. Swanson Analysis Systems Inc.
2. V. Székely and A. Poppe: Novel tools for thermal and electrical analysis of circuits. *Electrosoft*, 1(4):234–249, 1990. Computational Mechanics Publications.
3. R. E. Simons: The Evolution of IBM High Performance Cooling Technology. In *Proc. of the 11th IEEE SEMI-THERM Symposium*, pages 102–114, 7-9 Febr. 1995.
4. G. Gaussorgues: *Infrared thermography.* Chapman & Hall, London, 1994.
5. V. Székely *et al.*: Design for Thermal Testability (DfTT) and a CMOS realisation. In *Proc. of the THERMINIC Workshop*, Grenoble, France, 25-26 Sept 1995. (accepted for publication).
6. V. Székely and M. Rencz: Thermal Test and Monitoring. In *Proc. of the ED&T Conference '95, Paris*, page 601, 7-9 March 1995.
7. S. Middelhoek and S. A. Audet: *Silicon Sensors.* Academic Press, London, 1989.

CONTRIBUTORS

Lex A. Akers
Dept. of Electrical Engineering
Arizona State University
Box 875706
Tempe, AZ 85287-5706
U.S.A.

Alec Broers
Department of Engineering
University of Cambridge
Trumpington Street
Cambridge, CB2 1PZ
England

Misha Dyakonov
A.F. Ioffe Physico-Tech. Inst.
Polytechnicheskaya 26
St. Petersburg, 194021
Russia

Bogdan T. Fijalkowski
Dept. of Mechatronics
Cracow University of Technology
ul. Baluckiego 5A/1
PL 30-318, Krakow
Poland

Jim Freedman
Semiconductor Res. Corp.
P.O. Box 12053,
79 Alexander Dr., Bldg. 4401
Research Triangle, NC 27709
U.S.A.

Vera B. Gorfinkel
Dept. of Electrical Engineering
State University of New York
Stony Brook, NY 11794-2350
U.S.A.

Erich Gornik
Inst fuer Festkoerperelektron
TU-Vienna
Gushausstrasse 27-29
A-1040, Vienna
Austria

Herb Goronkin
Phoenix Corporate Research
Laboratories
Motorola Inc.
2100 East Elliot Road,
MD EL 508
Tempe, AZ 85284
U.S.A.

Sergei Gurevich
A.F. Ioffe Physico-Tech. Inst.
Polytechnicheskaya 26
St. Petersburg, 194021
Russia

Karl Hess
Dept. of Electrical Engineering
University of Illinois
Urbana-Champaign
1101 West Springfield Avenue
Urbana, IL 61801-3082
U.S.A.

Gilles Horowitz
Laboratoire des Materiaux Moleculaires
CNRS
2, rue Dunant
94320, Thiais
France

Robert Hyde
Alp Optics
38220, Laffrey
France

Paul Jay
Microwave Module Group
Northern Telecom Limited
P.O. Box 3511, Station C
Ottawa, Ontario K1Y 4H7
Canada

Michael Kelly
Department of Physics
University of Surrey
Guildford, GU2 5XH
England

Herbert Kroemer
Dept. of Electrical &
Computer Engineering
University of California
Santa Barbara
Engr I, Room 4113
Santa Barbara, CA 93106
U.S.A.

Michael A. Littlejohn
Dept. of Electrical Engineering
North Carolina State University
Raleigh, NC 27695-7911
U.S.A.

Serge Luryi
Dept. of Electrical Engineering
State Univ. of NY - Stony Brook
Stony Brook, NY 11794-2350
U.S.A.

W. Ted Masselink
Department of Physics
Humboldt University in Berlin
Invalidenstr. 110
10115, Berlin
Germany

Marco Mastrapasqua
Physics Department
Eindhoven Univ. of Technol.
Room NL a2.10, P.O. Box 513
5600 MB, Eindhoven
The Netherlands

Andrzej Napieralski
Division of Microelectronics &
Computer Sciences
Inst. of Electr., Tech. Univ. of Lodz
ul. Stefanowskiego 18/22
90-924, Lodz
Poland

Steve Nelson
Steven Nelson and Associates
6706 N. Lakeshore Drive
Chippewa Falls, WI 54729
U.S.A.

Arto Nurmikko
Center for Advanced Materials Research
Brown University
182 Hope Street
Providence, RI 02912
U.S.A.

Dimitris Pavlidis
Solid State Electronics Laboratory,
Dept. of EE
University of Michigan
1301 Beal Avenue
Ann Arbor, MI 48109-2122
U.S.A.

Marta Rencz
Dept. Electron Devices
Technical Univ. - Budapest
Goldmann Gyorgy ter. 3
H-1521, Budapest
Hungary

George Sai-Halasz
IBM T.J. Watson Res. Center
P.O. Box 218
Yorktown Heights, NY 10598
U.S.A.

Michael Shur
Dept. of Electrical Engineering
University of Virginia
Thornton Hall
Charlottesville, VA 22903-2442
U.S.A.

Theoren P. Smith
Director, Physical Sciences
IBM T.J. Watson Res. Center
P.O. Box 218
Yorktown Heights, NY 10598
U.S.A.

Paul M. Solomon
IBM T.J. Watson Res. Center
P.O. Box 218
Yorktown Heights, NY 10598
U.S.A.

Ben G. Streetman
Dept. of Electrical Engineering
University of Texas at Austin
P.O. Box 7728
Austin, TX 78712
U.S.A.

Michael A. Stroscio
US Army Research Office
P.O. Box 12211
Research Triangle,
NC 27709-2211
U.S.A.

Robert Suris
A.F. Ioffe Physico-Tech. Inst.
Polytechnicheskaya 26
St. Petersburg,
Russia

Henk van Houten
Philips Research Lab. Holland
P.O. Box 80.00
Eindhoven, JA 5600
The Netherlands

Armin W. Wieder
Director, Silicon Microelectronics
Siemens Research Labs
8 Munich 83/Zt Zie ME 4,
Otto Hahn Ring 6
8000, Munich 83
Germany

Jimmy Xu
Dept. of Electrical &
Computer Engineering
University of Toronto
10 King's College Road
Toronto, Ontario M5S 1A4
Canada

Alex Zaslavsky
Div. of Engineering
Brown University
182 Hope Street, Box D
Providence, RI 02912
U.S.A.

Index