②

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION	1b. RESTRICTIVE MARKINGS
UNCLASSIFIED	

2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT
	Approved for public release; distribution unlimited.

AD-A195 731

5. MONITORING ORGANIZATION REPORT NUMBER(S)

AFOSR·TR· 88-0569

6a. NAME OF PERFORMING ORGANIZATION	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION
Univ of California		AFOSR/NE

6c. ADDRESS (City, State, and ZIP Code)	7b. ADDRESS (City, State, and ZIP Code)
College of Engineering Berkeley, CA 94720	Bldg 410 Bolling AFB, DC 20332-6448

8a. NAME OF FUNDING / SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR/NE	NE	F49620-87-C-0041

8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS			
Bldg 410 Bolling AFB, DC 20332-6448	PROGRAM ELEMENT NO. 61102F	PROJECT NO. 2305	TASK NO. A9	WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Joint Services Electronics Program

12. PERSONAL AUTHOR(S)
Dr Oldham

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)
Annual Progress	FROM 01May87 TO 31Dec 87	Feb 88

16. SUPPLEMENTARY NOTATION

17.	COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

As part of a project to construct, model and compare very high performances Si and GaAs devices, deep-submicron Si MOSFETs have been investigated. using a novel optical lithography process, these extremely high-gain MOSFETs with varing oxide thicknesses and channel lengths have been fabricated. Figure 1 shows a Scanning Electron Microscope (SEM) picture of a transistor cross-section with an effective channel length 0.2 μm, an oxide thickness of 86 A, and junction depth of 0.2 μm. Devices with effective channel lengths as small as 0.15 μm and oxide thickness as thin as 36 A have been successfully fabricated and operated.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT	21. ABSTRACT SECURITY CLASSIFICATION
☐ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL	22b. TELEPHONE (Include Area Code)	22c. OFFICE SYMBOL
Witt	(202) 767-4931	NE

DD FORM 1473, 84 MAR   83 APR edition may be used until exhausted.   SECURITY CLASSIFICATION OF THIS PAGE

All other editions are obsolete.

AFOSR-TR- 88-0569

Report No. UCB/ERL-87/1

JOINT SERVICES ELECTRONICS PROGRAM

ANNUAL PROGRESS REPORT
(Contract F49620-87-C-0041)
(1 May 1987 — 31 December 1987)

W. G. Oldham

29 February 1988

Prepared for:

Air Force Office of Scientific Research (NE)
Bolling Air Force Base
Washington, DC 20332

88 5 16 109

ELECTRONICS RESEARCH LABORATORY
College of Engineering
University of California, Berkeley, CA 94720

TABLE OF CONTENTS

PART A - DIRECTOR'S OVERVIEW

Beginning May 1, 1987, a new JSEP three-year program was initiated at Berkeley. Two major themes were identified: High Speed Wide Band Elements for High Frequency Electronics and New Architectures for Parallel Computation. The first contains four work units and a supplement and the second, three work units.

In an effort to coordinate the separate parts of the program and to encourage collaboration between the various investigators, a couple of internal reviews have been held. On September 21, 1987, there was a one-day conference, with presentations by 14 Research Assistants then supported by the program. It was decided after that meeting that a specialists' meeting within the two team areas would be fruitful. On February 5, 1988, a half-day meeting was held to cover all the activity under new architectures for parallel computation. A similar review of the high-frequency thrust will be held in the near future. We find these meetings useful and essential for managing the nudges in the direction of the program.

At present there are three or four more Research Assistants supported by this program than can be continued with the combination of inflation and reduction in funding. Thus, it is critical to identify the most fruitful areas for emphasis. The joint service electronics program continues to be an important factor in the ability of this laboratory to respond to important new research opportunities. The partial funding of the supplement on Electro Optic probing and the GaAs devices and materials research of our new faculty member, Professor Steve Smith, are positive examples.

In the following we list a number of significant accomplishments which, because of the recent start of the new three-year program, include research begun under the three-year proposal ending April 30, 1987. Finally, in Part C, we describe the work under the individual work units and cite the publications under JSEP sponsorship.

PART B - ACCOMPLISHMENTS

High Speed/Wide Band Silicon and GaAs Devices
Professor Chenming Hu and Professor Ping Ko

As part of a project to construct, model and compare very high performance Si and GaAs devices, deep-submicron Si MOSFETs have been investigated.

Using a novel optical lithography process, these extremely high-gain MOSFETs with varying oxide thicknesses and channel lengths have been fabricated. Figure 1 shows a Scanning Electron Microscope (SEM) picture of a transistor cross-section with an effective channel length 0.2 Uum, an oxide thickness of 86 A, and junction depth of 0.2 um. Devices with effective channel lengths as small as 0.15 um and oxide thickness as thin as 36 A have been successfully fabricated and operated.

Figure 2 displays a plot of linear transconductance (a measure of device speed) versus the effective channel length. MOSFETs with 36 A gate oxides and 0.15 um effective channel lengths exhibit record room-temperature transconductances of 460 mS/mm. For these devices intrinsic cut-off frequency is 73 GHz. NMOS ring oscillators have been fabricated with delays as small as 22 $ps/stage$.

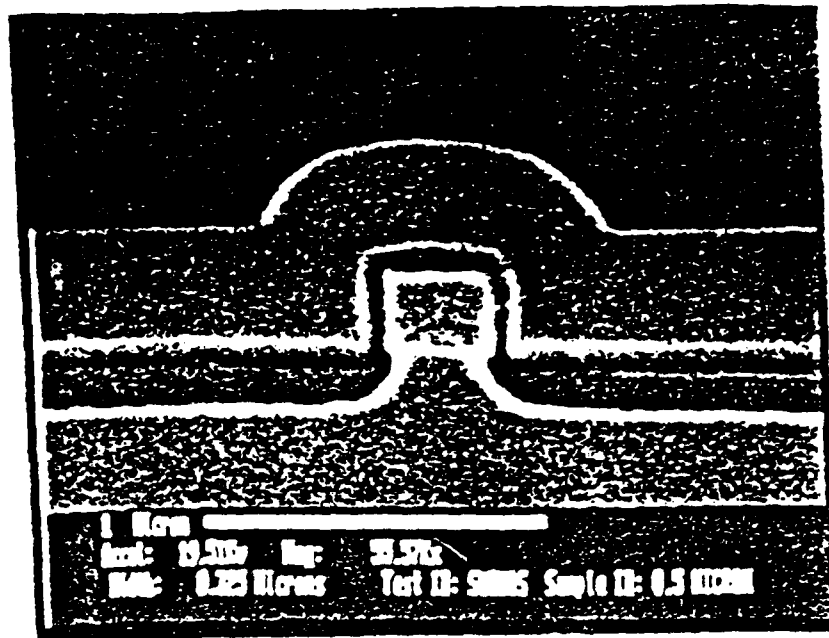Figure 1. SEM picture of a transistor cross-section with an effective channel length of 0.2 μm, an oxide thickness of 86 Å, and a junction depth of 0.2 μm.
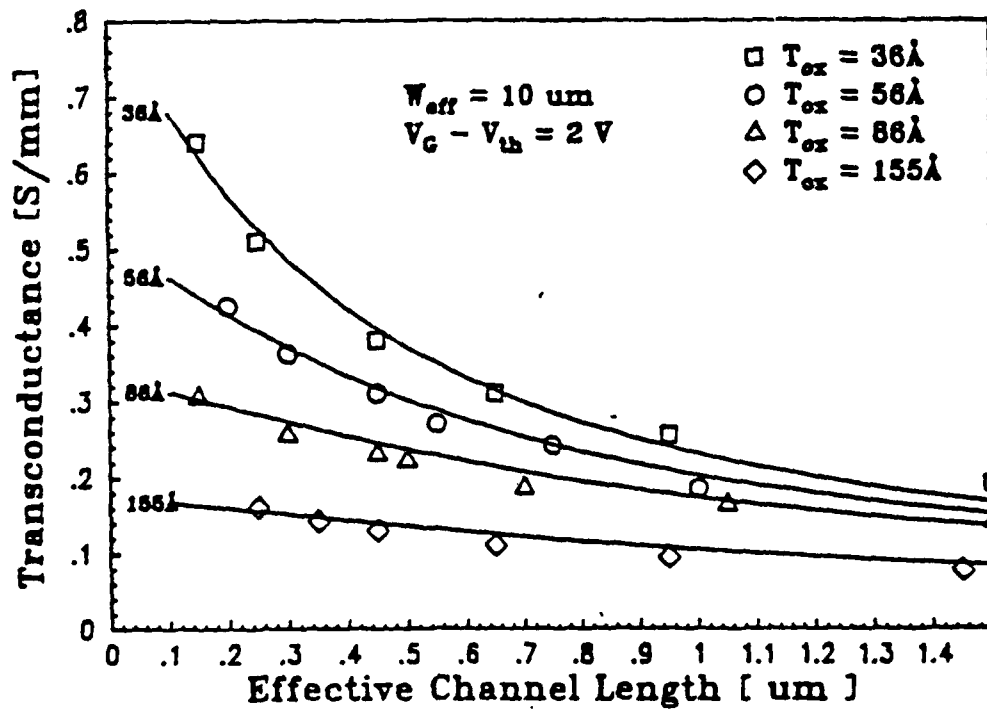

Figure 2. Plot of linear transconductance versus the effective channel length.

Defect-Engineering:

Proximity Gettering with Mega-electron-volt Carbon Implantation

Hing Wong and Professor Nathan Cheung

The effectiveness of impunity gettering in semiconductors can be enhanced by placing the gettering centers *close* to the active device region while maintaining crystal perfections *within* the device region. This **proximity gettering** idea is illustrated in Figure 1. With Mega-electron-volt (MeV) ion implantation, a buried gettering layer can be readily formed without creating significant defects in the top active device region. Recently, we have demonstrated that a strong proximity gettering layer with excellent thermal stability can be formed with MeV carbon implantation in silicon.

MeV carbon ions with doses ranging from $5 \times 10^{-14} cm^{-2}$ to $5 \times 10^{-16} cm^{-2}$ were implanted into silicon wafers contaminated with gold. The Au was intentionally diffused into the wafers as a marker impurity. The wafers were then annealed in an N_2 ambient at $1000^\circ C$ for one hour. Secondary Ion Mass Spectroscopy (SIMS) was used to measure the depth profiles of C and the gettered Au. Shown in Figure 2 (a) is the SIMS profile of $2 \times 10^{-16} cm^{-2}$ 2.4 MeV carbon implant after annealing. A large amount of Au ($5 \times 10^{13} cm^{-2}$) was trapped in the C implanted region. Complexing of the C with the background oxygen is suggested by the O peak at the same depth as the carbon. A dose dependence study showed that the gettering efficiency of implanted carbon is linear with dose. The thermal stability of the carbon gettering effect was tested for annealing up to 12 hours at $1000^\circ C$. The amount of gettered Au increased with annealing time, indicating that gettering is still in diffusion limited regime under such annealing conditions.

Cross-section transmission electron microscopy (XTEM) was used to study the structure of the carbon implanted gettering layer. Figure 2 (b) is an XTEM micrograph obtained on the same wafer whose SIMS profiles are shown in Figure 2 (a). Unlike ions, implanted carbon does not create any visible extended defects, such as dislocations or large precipitates. The only feature in the carbon implanted layer is a vague dark band with a thickness of about $0.5um$. This finding suggests that the gettering effect of carbon is through point defects or their clusters, not through dislocations or precipitate interfaces.

Four MeV carbon ions were also implanted into diode structures to check the device quality of the top layer after MeV implantation. No detrimental degradation in diode leakage current was found. With a dose of $10^{15} cm^{-2}$ and a one-hour annealing at $950^\circ C$, the leakage increase was negligible at 5 volts reverse bias. Even at a dose of $10^{16} cm^{-2}$, the leakage current was less than $10 nA/cm^2$ at 5 volts reverse bias.

Figure Captions:

Figure 1. The effectiveness of gettering can be enhanced by
putting the gettering layer close to the active device region.

Figure 2. (a) SIMS profiles and (b) XTEM micrograph obtained on
a sample implanted with $2 \times 10^{16} cm^{-2}$ 2.4 MeV carbon
and annealed at 1000°C for 1 hour. A large amount
of Au (about $5 \times 10^{13} cm^{-2}$) was trapped in the carbon
implanted region. No extended defects were observed.
In (b) the top $1um$ of silicon was sputtered off during
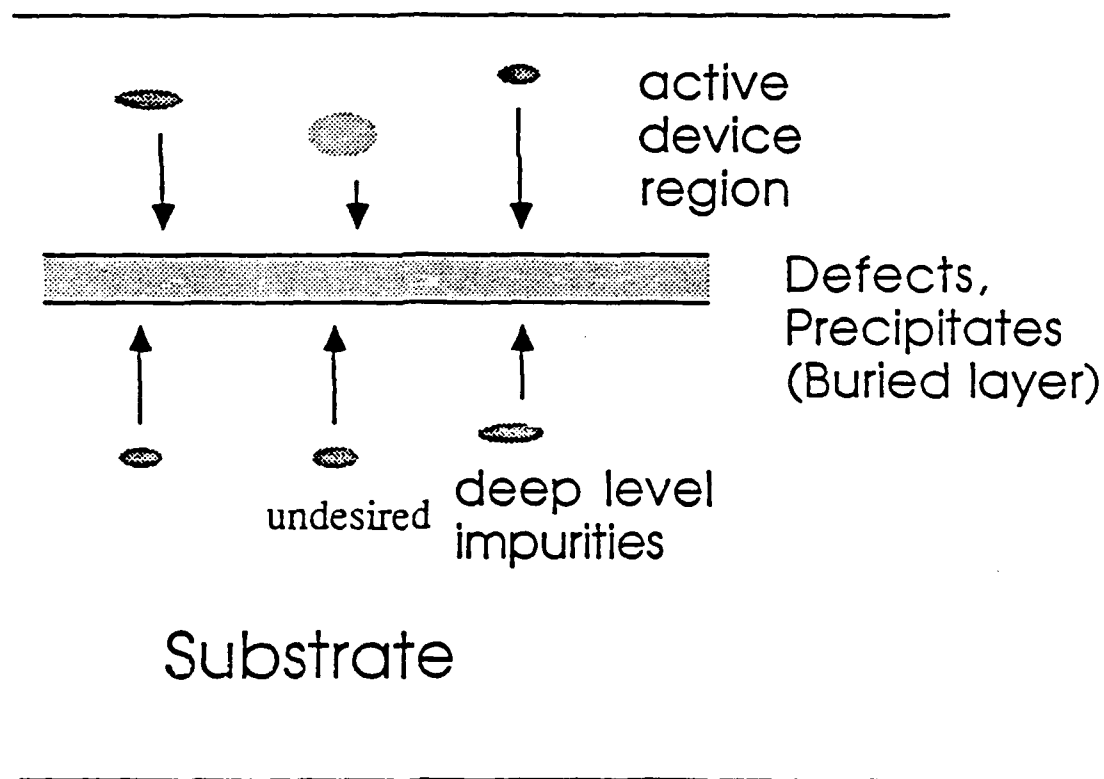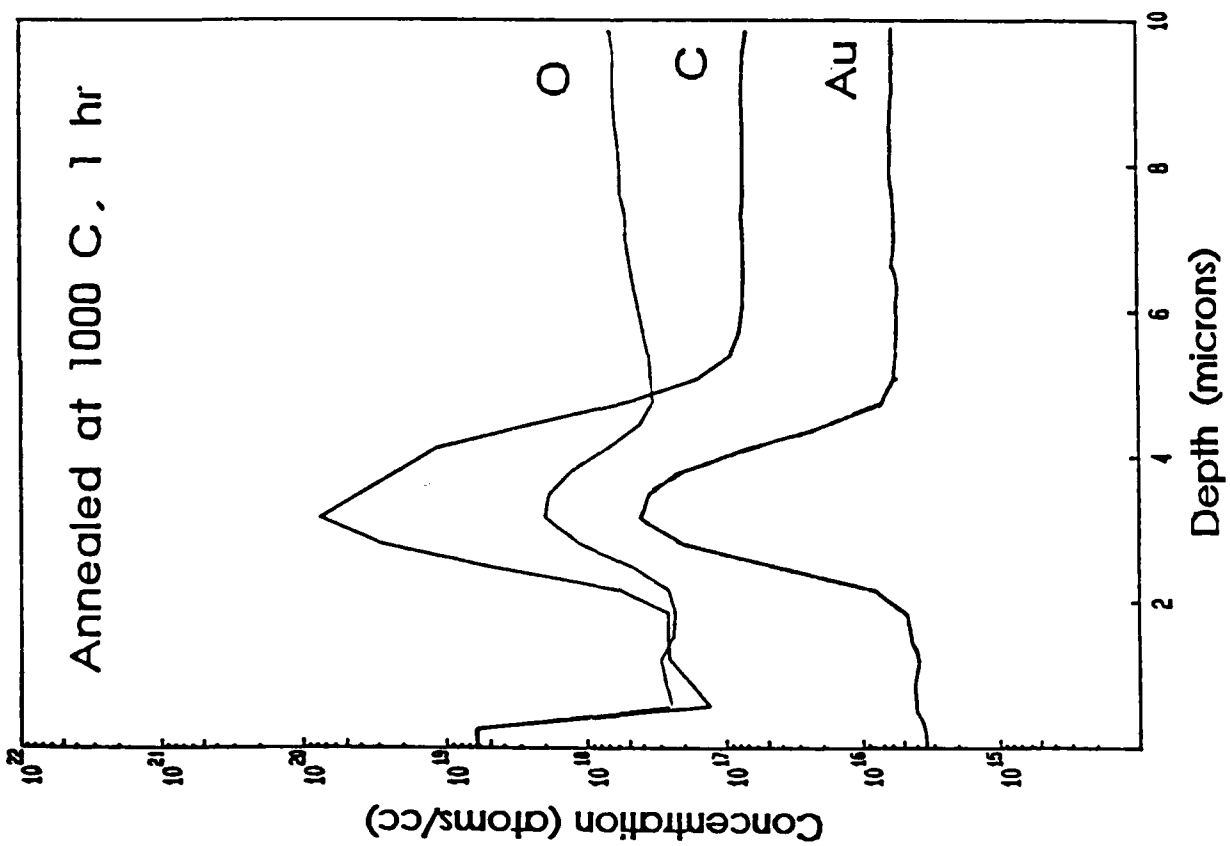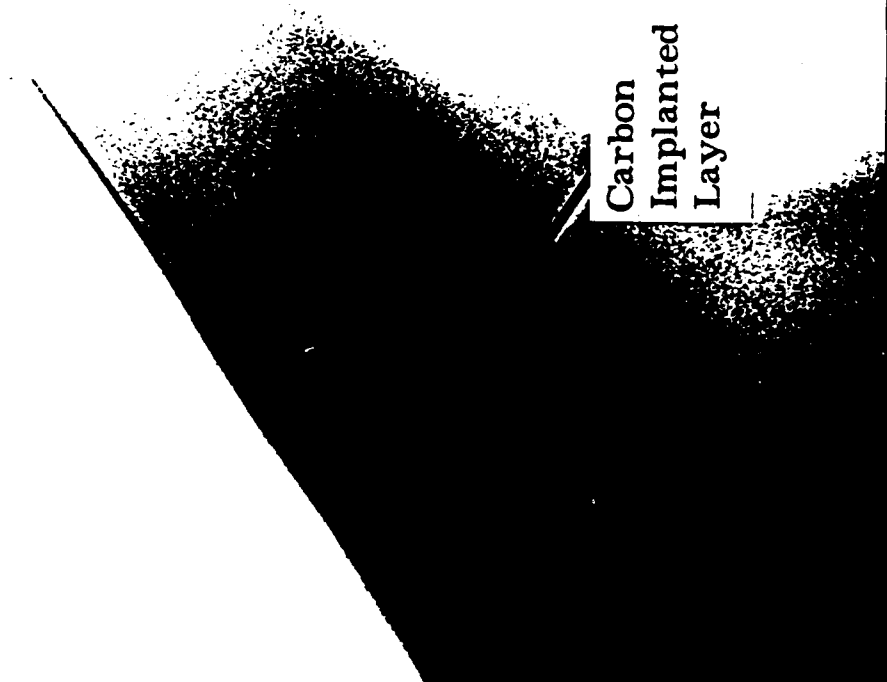TEM sample preparation.

Proximity Gettering With MeV Implantation

Figure 1.

Carbon Implant   2.4 MeV $2x10^{16}$ cm^{-2}

Annealed at 1000 C, 1 hr

0.5 μm

Carbon Implanted Layer

(a)

(b)

Figure 2.

-7-

GaAs probing: Surface Properties to 3-D Field Mapping

Yu-Hwa Lo and Professor S. Wang

Description

Electro-optic probing, pioneered and demonstrated first in $LiNbO_3$ by Mouron and later extended to GaAs by Bloom, has the potential of becoming a new powerful tool for materials and device research. We are exploring the use of the technique to study materials and device properties which are either inaccessible to, or difficult for, other probing techniques. We have carried on experiments in three hitherto unexplored areas: (1) probing surface properties of a semiconductor, 2) measuring standing wave ratio along a coplanar waveguide, and (3) three dimensional mapping of internal fields in devices.

1. Probing Surface Properties:

The surface property plays an important role in device performance. One problem often encountered in processing semi-insulating (SI) GaAs substrates is thermal conversion of the conductivity type. Backgating in GaAs integrated circuits has been attributed to the change in EL_2 defect center density. However, it is almost impossible to detect the change in surface conductivity. Figure 1 shows (a) the configuration of a semi-insulating sample in relation to the probing beam, (b) the measured $d^2V_{e \to o}/dx^2$ signal for an unannealed sample, and (c) the measured $d^2V_{e \to o}/dx^2$ signal for an annealed (at 700° C) sample, where $V_{e \to o}$ is the electro-optic signal. From the direction of the signal movement, the surface type can be ascertained. The surface has changed from semi-insulating in (b) to slightly p type in (c).

2. Standing-wave-ratio Measurement:

For mm-save integrated circuits, it is important to be able to measure the standing wave ration along waveguides to yield information about circuit elements which we design. Figure 2 shows (a) the experimental set up, and (b) the electro-optic signal observed along a coplanar waveguide. The E-O signal shows clearly a standing-wave pattern. The experiment was first carried out at 8.21 GHz and is now extended to 20 GHz. The electro-optic measurement of SWR should complement the S-parameter measurement in characterizing mm-save integrated-circuit elements.

3. 3-D Field Mapping:

To understand the physics of complicated III-V compound devices and to set up physical models for prediction of device performance, it is essential to know the internal potential, field and charge distribution. There is no technique at present available to provide such crucial information in either bulk semiconductors or devices. Our work is to establish the feasibility of providing the information by electro-optic probing. The electro-optic signal senses the field component in the direction of light propagation. By changing the angle of incidence, information is obtained about all three field components. An example is shown in Figure 3 for a simple structure consisting of two electrodes on a GaAs surface. The electro-optic signal expected may be simulated as shown in Figure 4. The more difficult deconvolution problem (for obtaining the orthogonal fields from the E-O signal) is presently under investigation.

SMALL A.C. FROM
LOCX-IN AMPLIFIER

ΔV

D.C. BIAS

TO DETECTOR

n+   n+

S.I. GaAs

x

y

50-100Å Au
(EQUI-POTENTIAL PLANE)

<001> AXIS

PROBING LASER BEAM
(SCANNED ALONG x-AXIS)

fig. 1(a)

$\frac{d^2 V_{e=0}}{dx^2}$ (ARBITRARY UNIT)

— : V_{dc1} = 10 VOLT
--- : V_{dc2} = 30 VOLT
ΔV = V_{rms} = 2 VOLT

0   5   10   15   20   x(µm)
CATHODE   ANODE
(a)

$SCB(V_{dc1} \cdot \Delta V)$ $SCB(V_{dc2})$
$SCB(V_{dc1})$ $SCB(V_{dc2} \cdot \Delta V)$

n+   S.I.   n+   V

fig. 1(b)

$\frac{d^2 V_{e-0I\tau I}}{dx^2}$ (ARBITRARY UNIT)

— : V_{dc1} = 22.5 VOLT
--- : V_{dc2} = 45 VOLT
ΔV = V_{rms} = 2 VOLT

0   5   10   15   20   x(µm)
CATHODE   ANODE
(b)

$SCB(V_{dc2})$ $SCB(V_{dc1} \cdot \Delta V)$
$SCB(V_{dc2} \cdot \Delta V)$ $SCB(V_{dc1})$

n+   p-   n+   V

fig. 2(a)

fig. 2(b)

FIGURE 3   APL vo. 50, p. 1791, June 1987

fig. 3(a)

(a)

(b)

(c)

fig. 3(b)

-12-

Miniaturized of Josephson Logic Circuits

Herbert Ko and Professor T. VanDuzer

As part of the study of the impact of high-T_c, high-energy-gap superconductors on digital circuit performance, the effect of scaling on the performance of Josephson logic circuits to smaller sizes has been studied. Theoretical work has shown the importance of the ratio of the maximum zero-voltage current of a Josephson junction to to its capacitance as the determining factor in setting the speed. This ratio is set in tunnel junctions by the tunneling current density (which has a maximum practical value of about 10 A/cm^2 in niobium and lead-alloy unctions). Detailed simulations of two logic families were carried out. The minimum cell OR gate delay is about 6 ps and the cell delay is about 20 ps in both cases, including the delay of propagation to the next gate.

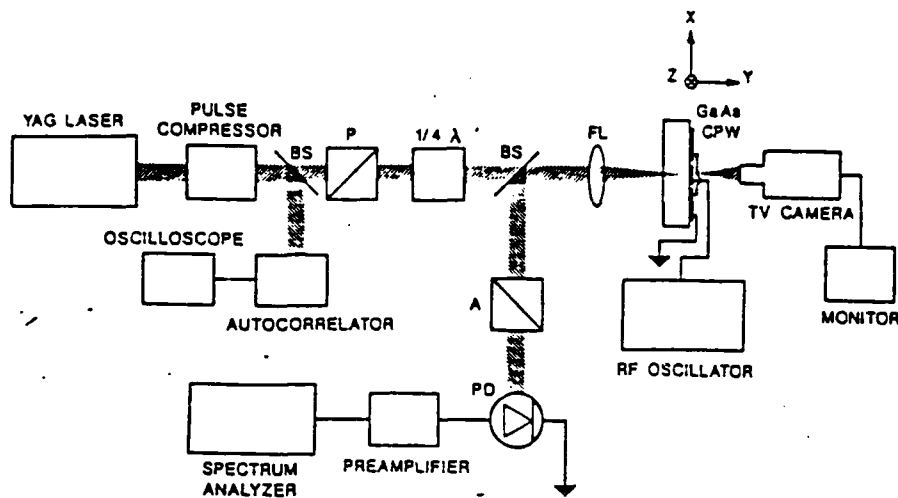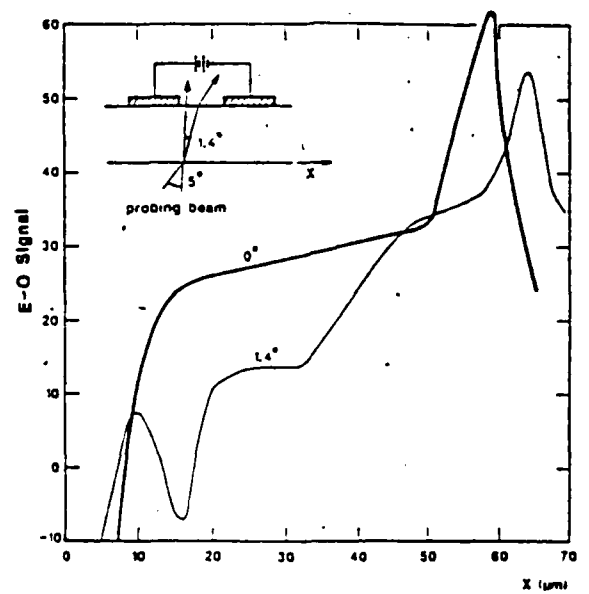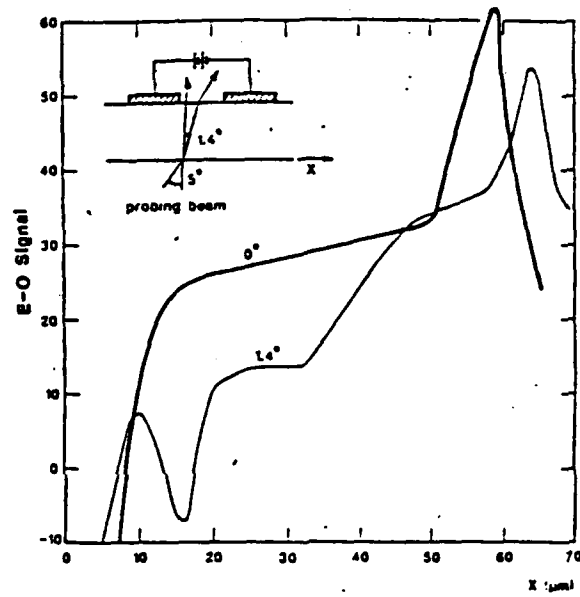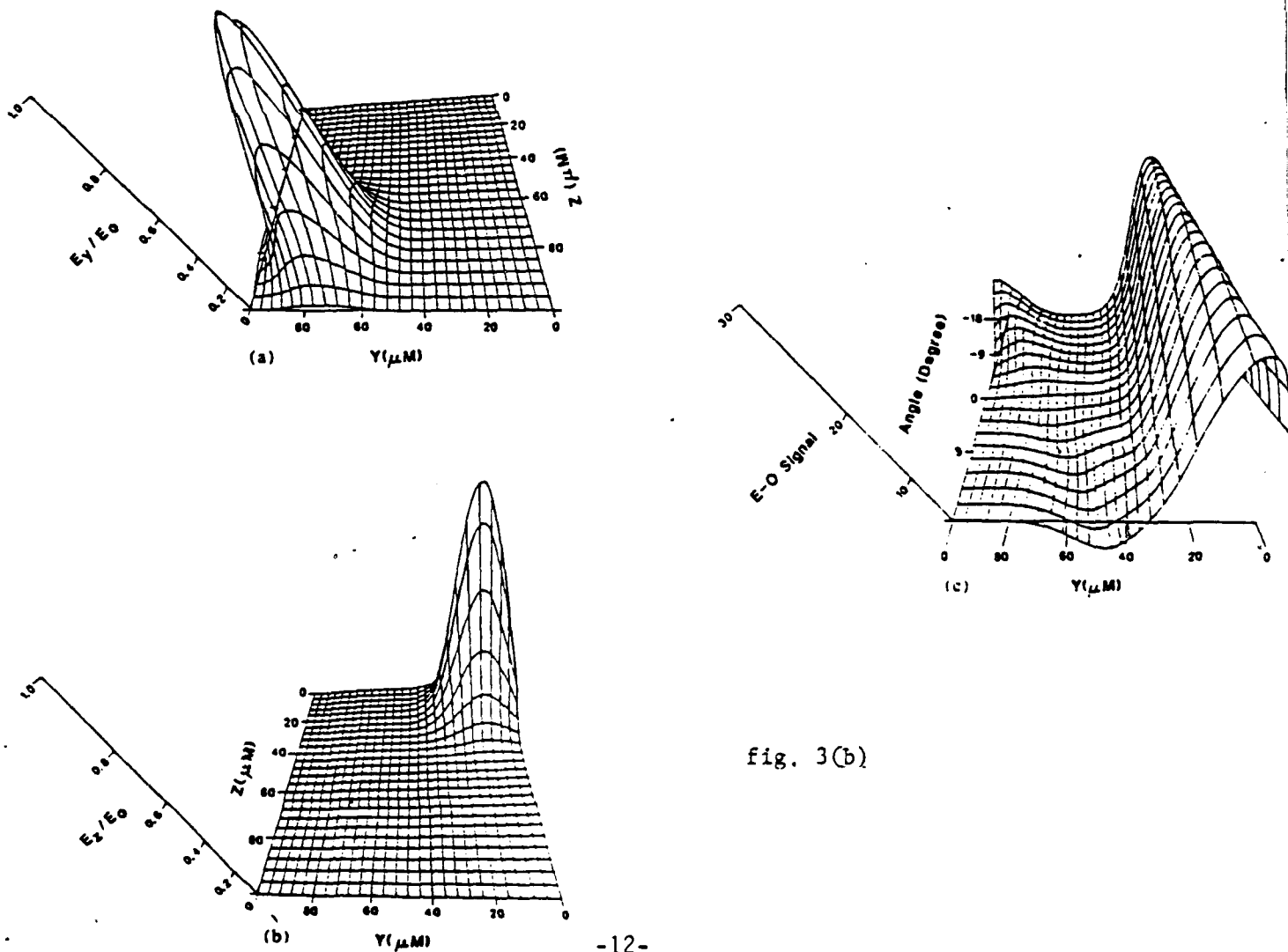Calculations also were made on a model of a logic chip in which the chip is divided into cells, each of which contains two OR gates driving an AND gate. Each cell also has space for ten east-west transmission lines and ten running north-south. The optimization of speed and circuit density can be performed, taking into account practical constraints on the parameters such as maximum tunneling, current density, minimum junction current (noise limitation), and minimum practical dielectric thickness and permittivity. The optimum junction size is about 1 um^2. Thus, new circuit fabrication processes were developed to achieve such a small dimension. A technique was developed that provided good control of the junction dimensions and, therefore, tunneling currents.

Experiments were done to confirm the simulation results for single OR and AND gates. To perform the experiments, an on-chip superconductor sampler circuit was fabricated along with the associated external electronics required to make measurements with picosecond resolution. The measurements made with the sampler have confirmed the scaling, with closest agreement on the faster circuits. Previous measurements of gate speeds have been done on chains of gates so the results were averages and the bias of the individual gates could not be separately controlled. The present work is the first in which the delay of a single gate was studied.

ELECTRONICS RESEARCH LABORATORY
University of California, Berkeley

Superconductive Integrated Circuits

Research Effort (JSEP)

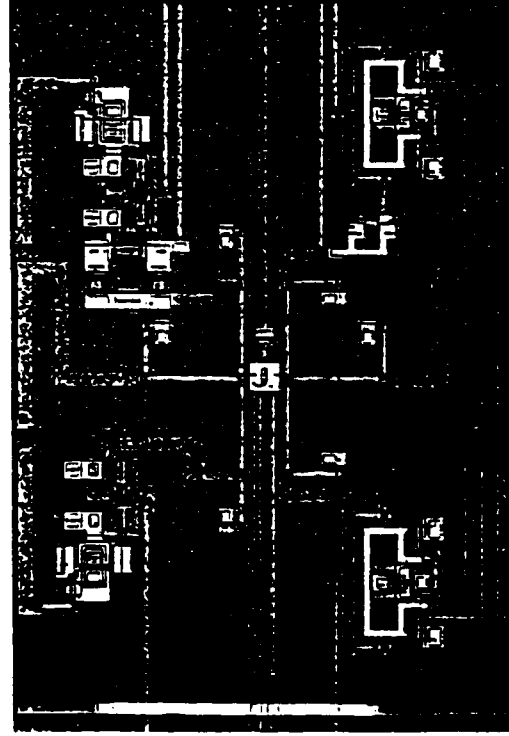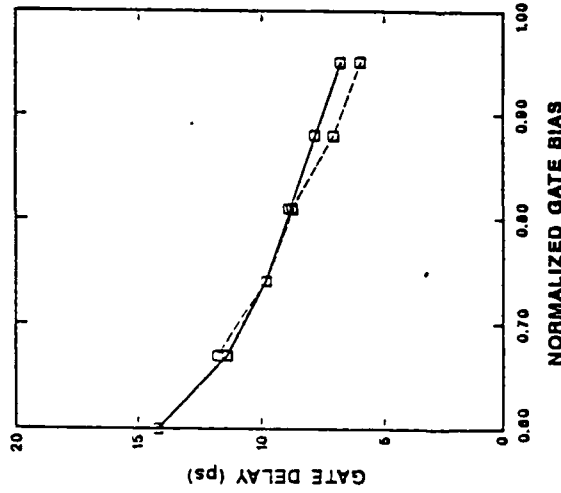- Miniaturization Calculations
- Process Studies for $1\mu m^2$ Josephson Tunnel Junctions
- Picosecond Sampler
- Gate-delay Measurement

Benefits

- Determine Limits on Speed and Density
- New Technique for Junction Fabrication

Accomplishments

- Developed $1\mu m^2$ Junction Process
- Measured Delay of 6ps on Single Gate with on-chip sampler

Superconductive picosecond sampling circuit

Statistical Design of Integrated Circuits

David Riley and Professor Alberto Sangiovanni-Vincentelli

Research in statistical design of integrated circuits in the last year has been aimed at demonstrating the feasibility of a key element of a new problem formulation for statistical design that we have developed.

Statistical design of ICs is concerned with finding optimal circuit parameter values, taking into consideration random disturbances inherent in wafer fabrication that can cause the performance of some of the circuits produced to be significantly degraded.

In the beginning stage of the project, it was decided that several basic assumptions made in conventional formulations for statistical IC design were restrictive when applied to real industrial design, particularly for the design of ICs to be sold on the open market. With particular attention to this case, we have developed a new problem formulation which generalizes conventional formulations in a number of major respects, and which forms the basis for a new statistical design methodology. Each item of the following list compares a significant restrictive assumption of conventional formulations with the related relaxed assumption of the new formulation.

1. The number of electrical performance categories of the circuits produced is two (good/bad); the number of performance categories is arbitrary.

2. The design parameters optimally determined in the formulation are device dimensions; they include all parameters having a first-order effect on the economic value of the circuits.

3. The limits defining the performance categories are prescribed; they are among the designable parameters of the formulation.

4. In the most common problem formulations, parametric yield maximization, the criterion quantity is, in equivalent economic terms, the ratio of revenue to cost; it is a form of profit (revenue minus cost) appropriate to the problem.

5. The only type of random effect degrading the economic value of the circuits produced is parametric fluctuations through deviations in device parameter values; defect phenomena can also degrade the economic value.

Work on the new formulation has entailed development of a combined technical and economic model which specified the following:

1. The criterion quantity (the profit to be expected from the design), the required constraint quantities, the design parameters and, the random variables required in the statistical modeling.

2. The functional dependence of the first two in Item 1 on the latter two.

Development of a realistic model has been made possible through extensive collaboration with Harris Semiconductor. The model has been documented in Ref [1].

Note that the design parameters include fabrication test limits, mask dimensions, back-end fabrication assignments (e.g. to package types or reliability treatments), final product performance specifications and, selling prices.

If the new model is to be useful in industrial practice, it must be feasible to obtain all the data required by the model. Demonstration of this feasibility should precede development of the software implementation of the methodology.

The raw cost data needed for the cost model is readily available to the producer. However, the revenue model requires the modeling of the demand function of the IC being designed - - the expected quantity demanded as a function of the products specs and price. The demand depends on the behavior of potential users in choosing among commercially available realizations of the type of the IC being designed. Data to estimate this choice behavior is not readily available to the producer. It is demonstrating the feasibility of estimating the demand function that has been the aim of the work in the last year.

The demand model is based on the postulation of an appropriate functional form, and on the statistical estimation of parameters of the function. Data for this estimation is gathered in a survey of engineers throughout American industry who select the type of IC of interest for incorporation into electronic systems they design. Additional data obtained in the survey allows evaluation of alternative functional forms. The model has been developed in a collaboration with the Berkeley Economics Department.

A survey aimed at estimating the demand function for a particular type of IC (linear buffer amplifiers) was begun in September of last year to demonstrate the feasibility of our approach. A mixed phone and mail survey has been carried out with the assistance of part-time student employees. About 175 responses have been obtained to date, and the final number of responses is expected to exceed 200. Statistical estimation of parameters for the demand function based on responses obtained to date has been carried out successfully and, using software specially developed for the project, preliminary demand estimation computations have been completed and appropriate graphical displays produced.

Note that elements of our demand modeling approach have application to two very significant, non-open-market scenarios: 1) the procurement of an IC by a systems design division of a firm from an IC design division of the same firm and, 2) the procurement of an IC by the government agencies through competitive bidding. This is because the approach includes a technique for modeling the performance (and price, if applicable), and tradeoff judgments of individuals, or of a collection of individuals, in an electronic system design group.

Reference:

[1] D.C. Riley and A. Sangiovanni-Vincentelli, "Models for a New, Profit-based Methodology for Statistical Design of Integrated Circuits", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. CAD-5, No. 1, pp. 131-169, January 1986.

Timing-Driven Layout for Custom Super Chips

Michael A.B. Jackson, Ren-Song Tsay and Professor E.S. Kuh

This research addresses the topic of timing-driven layout for synchronous digital MOS systems. Here we attempt to constructively influence the timing of the chip by manipulating the topology of the cell placement and the interconnect routes. Our method consists of two main states: The first phase involves a pre-layout timing analysis reminiscent of timing verification, a "fast" placement, and a series of optimization problems formulated as linear programming problems, while the second phase requires layout algorithms that use the results of the optimization step.

The pre-layout timing analysis identifies critical paths through blocks of combinational logic, and the slack times to each output node. In parallel with this, the "fast" replacement yields an estimate of interconnect lengths based on the connectivity of the design. The results of these two analyses are then formulated as linear programming problems. Solutions to the linear programming problems yield an interconnect length (spatial constraint) for every net that guarantees the desired performance, once each spatial constraint is satisfied during layout.

The second phase of the methods consists of placement and routing algorithms that accept spatial constraints and generate layouts satisfying them. This new methodology has been implemented in the routing tools of BEAR, a new building-block layout system under development at U.C. Berkeley. In the event that all constraints are not satisfied, sophisticated interactive techniques are used to resolve unfulfilled constraints. Implementation of these ideas within the floor-planning/placement algorithms is currently under development and preliminary experimental results with the routing tools are very encouraging.

In a separate but related project, we have developed a method to place cells for superchips. The method is based on the solution of linear sparse equations and successive partitioning. It is extremely fast and the quality of the result is excellent. A triple-layer 100K gate array with 26K instances was placed in 50 minutes using VAX 8650. Still to be implemented is timing-driven algorithm described above.

PART C - INDIVIDUAL WORK UNITS

THEME I - High-speed Wide-band Elements for High-frequency Electronics

HFD.1. High Speed/Frequencies GaAs and Si Device Research
C. Hu

The current resolution limit of conventional optical lithography has required that deep-submicron MOS devices be fabricated exclusively by electron beam and X-ray lithography. Both techniques are, however, expensive and complicated. We have developed a photoresist-thinning process that, when used in conjunction with conventional optical lithography, permits the controlled definition of deep-submicron features.

N channel MOS transistors with 0.15-μm effective channel length and showing excellent device characteristics have been successfully fabricated using conventional g-line optical lithography and this thinning technique.

We have found the IV characteristics, gain, and short channel effects in these 0.15 to 0.5 μm MOSFET are well described by the models we have developed for 1 μm devices. The basic scaling guidelines developed for the 1-μm device are also applicable to deep-submicron devices. However, lowering the power supply from 5V to 3.3V fails to solve the hot-electron reliability problems; some form of hot-electron resistant structure will be needed.

We have discovered that when a thin-gate oxide MOSET is biased in the off-state, a large drain leakage current can be observed between drain and substrate. Gate current is also observed at higher drain voltages. These currents are important to short-channel thin-oxide devices and some high-voltage devices, such as EPROM and EEPROM cells. A proposed model based on the band-to-band tunneling theory can explain the experimental results and suggest guidelines for the design of thin-oxide VLSI MOSFETs. We attribute the gate-induced drain leakage current to the band-to-band tunneling in silicon in the gate/drain overlap region, where a deep depletion region is formed. Electrons and holes are generated by the tunneling of valence band electrons into the conduction band and collected by the drain and substrate separately. This effect sets a new limit on submicron device scaling.

As part of the continuing effort to formulate GaAs device models for circuit simulation, we are undertaking improvements to our models and examining fresh approaches. The principal goals are to develop a robust extraction system based on a PC, and to implement the model in SPICE3. The current model displays excellent agreement with both HEMT and MESFET characteristics and provides a one-region description incorporating threshold shift, subthreshold slope variation, transconductance degradation, and breakdown. It also includes a single-region Schottky diode and capacitance descriptions.

To simulate numerically the phenomenon of negative conductivity in GaAs, we have developed a program based on an implicit difference scheme. The simulator can analyze the field and carrier concentration behavior in a device for a variety of doping profiles and boundary conditions, and provides information about the device impedance, microwave power, and efficiency through a Fourier analysis of resulting current and voltage waveforms. The program has been tested by comparison with known results. It is being used to examine the feasibility of a proposed high-frequency oscillator based on the simultaneous propagation of multiple domains through GaAs.

Publications

T. C. Ong, P. K. Ko, and C. Hu, "Modeling of substrate current in P-MOSFETs," *IEEE Electron Device Letters*, vol. EDL-8, no. 9, September 1987, pp. 413-416.

J. Chen, T. Y Chan, I. C. Chen, P. K. Ko, and C. Hu, "Sub-breakdown drain leakage current in MOS-FETs." *IEEE Electron Dev. Lett.*, vol. EDL-8, no. 11, November 1987.

T. Y. Chan, J. Chen, P. K. Ko, and C. Hu, "The impact of gate-induced drain leakage current on MOS-FET scaling," *Technical Digest of International Electron Devices Meeting (IEDM)*, Washington, DC, December 1987.

S. Holland, I. C. Chen, and C. Hu, "Ultra-thin silicon-dioxide breakdown characteristics of MOS devices with n^+ and p^+ polysilicon gates," *IEEE Electron Device Letters*, vol. EDL-8, no. 12, December 1987, pp. 572-575.

M. C. Jeng, J. Chang, A. T . Wu, T. Y. Chan, J. Moon, G. May, P. K. Ko, and C. Hu, "Performance and hot-electron reliability of deep-submicron MOSFET's," *Technical Digest of international Electron Devices Meeting (IEDM)*, Washington, DC, December 1987.

HFD.2. GaAs/Si Devices and Materials Studies
S. Wang

Two important issueds facing GaAs/Si heteroepitaxy are (1) dislocation density and (2) internal stress. We have designed a series of experiments, including selected-area growth, rapid-thermal annealing, and plasma hydrogenation, to test their effects on dislocation density and used TEM to discern the changes in threading dislocations and stacking faults. We have also performed Raman and photoluminescence (PL) experiments to determine the internal stress as a function of distance from the interface in the former and as a result of differential thermal expansion in the latter.

Since the results on Raman and PL experiments have already been published (see publications 1-4, they are discussed first. A bevel was etched in a GaAs epitaxial film grown on a Si substrate so that the Raman spectrum of the GaAs layers can be measured as a function of distance from the GaAs/Si interface. The amount of strain and disorder has been estimated from the GaAs longitudinal optical phonon line shape and frequency to decrease from a value of 5.5×10^{-3} at a distance 500 Å from the interface to a value around 1×10^{-3} at a distance 2000 Å from the interface. Therefore, the misfit strain caused by lattice mismatch is largely relaxed in a film of 5000 Å thick by the generation of edge dislocations.

Even though misfit strain is relaxed, we still have to contend with stress caused by difference in the thermal expansion coefficients of GaAs and Si. For the PL experiments, GaAs films were grown on Si substrates with patterned SiN masks by molecular beam epitaxy. The pattern consists of bare Si stripes of width ranging from 10 μm to 100 μm surrounded by SiN on both sides and a reference area of bare Si. The PL experiments were carried out at 77K to yield information about the internal stress and the optical quality of the hetero-epitaxial film. Under stress, the degeneracy of the valence band is removed, and the heavy-hole and light-hole gap-energies shift with stress. From the PL peak shift and separation, the stress can be determined and is found to be 3.5×10^9 dynes/cm^2 at 77K corresponding to a thermal stress of 2×10^9 dynes/cm^2 at 300K. For 1.5 μm and 3 μm thick films, PL intensity from the 10 μm stripes showed a 140% and 75% increase over the reference area, respectively. This remarkable increase in the PL intensity is believed due to a reduction of dislocations inside the window area.

TEM pictures of selective-area grown samples show that stacking faults are confined to the boundary of single crystal and polycrystalline regions. TEM pictures of the rapid-thermal annealed (RTA) samples show a significant reduction of threading dislocations and an almost elimination of stacking faults in the single crystal (window) region. Work is in progress to obtain TEM pictures for selective-area grown samples after RTA. Preliminary experiments on hydrogenated samples show noticeable improvements in photo-response.

In summary, we have tried a number of approaches to reduce the defect density in and to improve the optical quality of hetero-epitaxial GaAs/Si films. Selective-area epitaxy appears effective in moving the defects to the area near the single crystal and polycrystalline boundary. Rapid thermal annealing also seems effective in defect reduction. Once the density is reduced to below a certain level, other techniques such as doped superlattices may become more effective in further reducing the defect density. We believe that a combined use of selective-area growth, incorporation of doped superlattices, post-growth RTA, and post-growth hydrogenation can bring the defect density to a level suitable for minority-carrier devices.

Publications

Y. H. Huang, P. Y. Yu, M.-N. Charasse, Y. H. Lo, and S. Wang, "A Raman Study of Epitaxial GaAs Layer on Si (100) Substrate," *Appl. Phys. Lett.*, Vol. 51, p. 192, 1987.

Y. H. Lo, M.-N. Charasse, H. Lee, D. Vakhshoori, Y. Huang, P. Y. Yu, Z. Liliental-Weber, M. Werner, and S. Wang, "Raman Scattering Spectrum Along a Bevel-Etched GaAs on Si, TEM Study, and GaAs PIN Photodetector on SI," *Materials Research Society Symposium Proceedings*, Vol. 91, *Heteroepitaxy on Silicon Technology*, p. 149, 1987.

H. P. Lee, S. Wang, Y.-H. Huang, and P. Yu, "Photoluminescence Studies of Selective-Area Molecular Beam Epitaxy of GaAs Film on Si Substrate," *Appl. Phys. Lett.*, Vol. 52, p. 215, 1988.

Y.-H. Huang, P. Y. Yu, H. P. Lee, and S. Wang, "Characterization of GaAs Film Grown on Si Substrate by Photoluminescence at 77K," *Appl. Phys. Lett.*, Vol. 52, p. 579, 1988.

HFD.3. Basic Techniques for Electromagnetic Scattering and Radiation:

Transmission Line Systems for Millimeter/Submillimeter Propagation
D. J. Angelakos and K. K. Mei

Coplanar transmission lines are being used in monolithic submillimeter devices. Because of the geometry, a number of problems arise related to coupling between lines and sections of lines. In addition, some topological arrangements between balanced and unbalanced lines lead to problems of matching and radiation. The time domain finite difference method described above is very suitable for such problem as well. Some preliminary results are already obtained. A full investigation for the coupling problem will require the access of a Cray 2 computer.

Pulse Radiating Antennas
Jiayuan Fang
(Professor K. K. Mei)

There are many types of broadband antennas, but most of them are phase dispersive. So, not all broadband antennas can radiate short pulses. This project intends to identify and investigate antennas which have minimal phase dispersion so that they are high fidelity both in phase and amplitude.

Computations of High-Frequency Circuit Components in Superconductivity Devices
Guo Chun Liang
(Professors K. K. Mei and T. Van Duzer)

Superconducting devices involve numerous circuit components that need to be analyzed and measured. Both computation and direct measurement are important, because neither approach along can provide reliable data. Time-domain techniques will be used to calculate passive components used in superconducting devices. These include microstrips, coplanar waveguides, and slot lines. Recently, we also find that it may be easier to use slow wave structure in the measurement and computations.

HFD.4. Submillimeter-Wave Mixing with Superconductive Tunnel Junctions

Michael A. Bruns
(Professor T. Van Duzer)

The goal of this work is to study the application of small ($\approx 1 \ \mu m^2$) superconductor-insulator-superconductor (SIS) tunnel junctions in an integrated antenna-mixer combination for submillimeter waves at frequencies up to 600 GHz. Recent research at UC Berkeley and elsewhere has already demonstrated that single SIS junctions and junction arrays can operate as efficient heterodyne mixers for millimeter waves with input noise approaching the quantum limit. The mixing properties of the SIS tunnel junction result from the extremely nonlinear dc current-voltage characteristic, the major feature of which is a sharp rise in tunneling current at the gap voltage. Quantum mixer theory predicts very low noise and even conversion *gain* (an impossibility in classical mixer theory) if the current rise of the dc characteristic is sharp on a voltage range of hf/e, where f is the signal frequency, h is Planck's constant, and e is the electron charge.

Another important factor is the junction shunt capacitance, which should be large enough to short out undesired higher harmonics of the local-oscillator frequency, but not so large that the wasted displacement current constitutes a very large fraction of the total junction current. As the local oscillator frequency increases, the junction capacitance must decrease to maintain good mixer performance. This constraint is not as severe when using series arrays of tunnel junctions, because the overall capacitance is reduced with the junction capacitances in series. Inductive tuning structures may also be used to cancel out junction capacitance.

The 600-GHz mixer experiment also differs from similar experiments at lower frequencies in that the signal and local oscillator are not coupled and guided by classical waveguide techniques. Instead, a quasi-optical approach is used, coupling and guiding the signal with mirrors and lenses to a focus on the SIS junction itself.

The first phase of the quasi-optical receiver study employs 90-GHz RF signals, for which the design and testing is easier and more accurate than for the 600-GHz system. A complete test system has been developed in the Physics Department and studies are in progress using junctions from another source.

The first part of the research on this project has been directed towards the fabrication of SIS tunnel junctions with low capacitance and dc I-V characteristics with a sharp current rise at the gap voltage. The junctions will subsequently be incorporated in a thin-film antenna-mixer with integrated impedance-matching structures. It is expected that experiment will deviate from theory, especially for the 600-GHz structures, and iterated design adjustments will be required for the thin-film devices.

We have devised a fabrication procedure for making, with accurate control, 1 μm^2 Nb-AlO$_x$-Nb tunnel junctions using a crossed-strip technique. The various steps are being developed.

SPL.1. Electro-Optic Probing of Semiconductor Materials and Devices
S. Wang

The work on electro-optic probing has been carried out in two fronts: (1) exploring the possibility for three-dimensional mapping of internal field and (2) demonstrating the capability of standing-wave measurements in coplanar microwave waveguides (see publications 1 and 2).

The electro-optic probing technique is for the first time proposed to detect three-dimensional field distribution in linear electro-optic material like GaAs. By changing the incident angles and positions of the probing beam, sufficient information of the field distribution should be included in the phase retardation of the probing beam. We have shown that if certain conditions on the probing beam are satisfied, a very simple linear relation between phase retardation and each field component can be found. A three-dimensional computer simulation was undertaken to illustrate the relation between a given field distribution and the predicted electro-optic signal. A simple experimental demonstrating the sensitivity of the electro-optic signal with the incidence angle was also performed. These results show that three-dimensional field mapping by electro-optic probing is not only mathematical solvable but also experimentally feasible.

We have also successfully measured, for the first time, standing waves in a GaAs coplanar waveguides at frequencies of 8.2107 and 12.310 GHz by using harmonic-mixing electro-optic probing technique. The technique consists of mixing the microwave signal at f_m with the n-th harmonic nf_o of the mode-locked laser pulse of repetition frequency f_o to produce an electro-optic probing signal at $f_m \pm nf_o$. This technique imporives greatly the signal to noise ratio, making possible SWR measurements at high microwave frequencies. The experiment has since been extended to 20 GHz. We believe that the technique can provide valuable information about impedance matching in the design of mon-wave circuits.

In addition to the work on electro-optic probing, we have also carried out exploratory work in several areas: (1) resonant tunneling diodes, (2) second-harmonic generation in GaAs waveguides, and (3) lasing action in GaAs/Si films. These studies are undertaken to explore potential high-speed electronic and opto-electronic devices for GaAs/Si hetero-epitaxy. Our theoretical study shows that a high peak to valley ratio can be achieved in resonant tunneling diodes if the barrier is made high and the

barrier layer is made thin. The study provides guidance for an optimal design of the device, (see publi-cation 3). The second-harmonic experiment originally was intended to detect stress-induced changes in electro-optic effect. During the investigation, we observed a coherent, second-harmonic emission normal to the surface (see publication 4) of a GaAs/AlGaAs waveguide. The observation of SH signal may open up novel possibilities in a number of applications, such as optical interconnect, auto-correlator and spectrum analyzer. Finally, we have fabricated gain-guided laser structures on GaAs/Si films and observed lasing action at 77K. The polarization is definitely TM. The emission spectrum will be analyzed, for the first time, to yield information on the internal stress in GaAs/Si films.

In summary, we have demonstrated the feasibility of electro-optic probing for three-dimensional field mapping and for SWR measurements. Earlier, we have used electro-optic probing to ascertain the cause for the back-gating effect (see Y. H. Lo, Z. H. Zhu, C. L. Pan, S. Y. Wang, and S. Wang, "New Technique to Detect the GaAs Semi-Insulating Surface Property --- CW Electro-Optic Probing," *Appl. Phys. Lett.*, vol. 50, p. 1125, April 27, 1987). One fundamental limitation of probing an electronic cir-cuit by an optical beam is the spatial resolution due to the finite spot size. The spatial resolution is not a problem for measuring SWR and for studying back-gating effect, but is a problem for probing integrated circuits. Ways to overcome this problem is under study. In the mean time, exploratory studies will be initiated in areas related to GaAs/Si hetero-epitaxy to enhance the chip functionality including opto-electronic integration.

Publications

Y. H. Lo, M. C. Wu, Z. H. Zhu, S. Y. Wang, and S. Wang, "Proposal for Three-Dimensional Internal Field Mapping by CW Electro-Optic Probing," *Appl. Phys. Lett.*, 50(25), pp. 1791-1793, 22 June 1987.

Z. H. Zhu, C. L. Pan, Y. H. Lo, M. C. Wu, S. Wang, B. H. Kolner, and S. Y. Wang, "Electro-Optic Measurement of Standing Waves in GaAs Coplanar Waveguides," *Appl. Phys. Lett.*, Vol. 50, p. 1228, May 4, 1987.

D. Vakhshoori and S. Wang, "Resonant Tunneling Diodes with AlAs Barriers: Guides for Improving Room Temperature Operation," *Jour. Appl. Phys.*, Vol. 62, p. 3474, October 15, 1987.

D. Vakhshoori, M. C. Wu, and S. Wang, "Surface-Emitting Second-Harmonic Generator for Waveguide Study," *Appl. Phys. Lett.*, Vol. 52, p. 422, February 8, 1988.

Y. H. Lo, "Topics on GaAs Integrated Circuits: GaAs Grown on Si Substrates, Field-Effect Transistors, and Electro-Optic Probing Techniques," Ph.D. Thesis, October 1987.

THEME II - New Architectures for Parallel Computation

MPC.1. CAD and Architectural Problems in Realizing Massively Concurrent Computation:

Acceleration Techniques for Analog Circuit Simulation

G. Casinovi and A. Sangiovanni-Vincentelli

We have been investigating techniques to accelerate circuit simulation of large scale integrated circuits. Our interest has focused on two ares: aggregation methods, which can be used to speed up the solution of large systems of algebraic equations, and macromodeling, which reduces the computational complexity of the circuit to be simulated.

Aggregation methods are used to improve the efficiency of relaxation-based circuit simulation, which has proved very effective for accurately simulating large-scale digital circuits. Programs such as RELAX and SPLICE are based on relaxation algorithms that have guaranteed convergence properties. These programs can analyze large circuits with the same accuracy and using the same models as SPICE2, however they are over one order of magnitude faster in terms of CPU time used. For circuits with tight feedback loops, the convergence of relaxation algorithms can be slow. Aggregation methods have been proposed to speed up relaxation algorithms used in the solution of large systems of algebraic equations.

A survey of aggregation methods was compiled by F. Chatelin and W. L. Miranker [1], who also showed that all the existing aggregation methods are nothing else but different implementations of the same mathematical idea; more precisely, aggregating a system corresponds to performing a projection on a suitable linear subspace: different aggregation methods correspond to different choices of subspaces.

In that report, the authors show that aggregation methods can be very effective in improving the performance of relaxation processes. At the same time, they prove that the choice of the subspace plays a major role in the gain that can be obtained from an aggregation step; it follows that one should choose carefully the subspace which to perform the projection onto.

We developed a new algorithm that uses information obtained from the relaxation process to build the aggregated system; theoretical analysis and numerical experiments performed on several

matrices show that this algorithm is more effective than the other aggregation methods proposed so far. The results of this research were presented at the 1985 International Symposium on Circuits and Systems [2]. We also showed that our method can be applied successfully to relaxation based circuit simulation. Our conclusions are summarized in a paper that will appear to the *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* [3].

The CPU time needed for cicruit simulation can also be decreased by using models for whole functional blocks rather than for individual devices. Most circuits can be logically divided into a certain number of interconnected functional blocks, and in many instances the designer is not interested in the detailed waveforms of the internal nodes. For instance, modularly designed systems contain several building blocks which are used repeatedly. As a consequence, a certain amount of CPU time can be saved by representing each block with a macromodel that retains approximately the same input-output behavior of the original block, while discarding all redundant information.

A natural way of obtaining a macromodel is to reduce the number of differential equations describing the circuit, trying at the same time to maintain the dynamics of the original system. This approach requires us to be able to quantify the intuitive notion that "two dynamical systems have approximately the same input-output behavior". In addition to that, we also need an algorithm that can tell us how to modify a given macromodel in order to make its dynamical behavior as close as possible to the one of the original system.

The first problem can be solved by taking the difference of the outputs of the original system and the macromodel for the same input, and computing its L^2 norm over a fixed time interval. The supremum of that norm, as the input varies over a given class of "admissible inputs", can be taken as a numerical indication of how much the two systems differ in their dynamical behavior. From a mathematical point of view, this is a classical optimal control problem, and in the literature there are numerous algorithms that can be used for its numerical solution.

In sharp contrast to this, only few and limited results have been published answering the question of how the dynamics of a system are affected by changes in the set of differential equations that describe it. In fact, we can show that this problem can be dealt with in a very general setting by a suit-

able generalization of the classical optimal control problem. Moreover, we already have a mathematical tool available that can be used to solve it, both from a theoretical and computational point of view. This tool is the theory of hamiltonian systems, which has been studied extensively because of its applications to many problems of classical and relativistic physics.

Applying those theoretical results to our case, we have developed an algorithm that can be used to modify a macromodel and make its dynamics fit those of the original system as closely as possible. We are now implementing the algorithm in a form suitable for modeling actual circuit. The program MODEL will use NUTMEG as input and output interface, giving it full compatibility with SPICE 3 and other simulators using the same language.

References

[1] F. Chatelin and W.L. Miranker: Acceleration by aggregation of successive approximation methods, IBM Thomas J. Watson Research Center Report No. RC 8508, 1980.

[2] G. Casinovi and A. Sangiovanni-Vincentelli: Aggregation Methods for the Solution of Large Scale Systems of Linear Algebraic Equations. Proceedings of the 1985 International Symposium on Circuits and Systems. Kyoto, Japan, June 1985.

[3] G. Casinovi and A. Sangiovanni-Vincentelli: A New Aggregation Technique for the Solution of Large Systems of Algebraic Equations. To appear on *IEEE* Transactions on Computer-Aided Design of Integrated Circuits and Systems.

Our research has focussed on three projects: Massive Parallel Computation with particular emphasis on neural networks, and macro-modeling for large-scale analog circuit simulation and placement of modules for large-scale chips.

Scalability in Feedforward Neural Networks
A. Kramer and A. Sangiovanni-Vincentelli

Neural networks have received a great deal of attention recently as a promising new model of parallel computation. While there is still a great deal of variation in definitions of what comprises a neural network, in general these all involve highly interconnected networks of simple analog computing nodes. The output of each of these nodes is a non-linear function of the weighted sum of its inputs. Typically these nonlinear functions are either step functions or sigmoids. The computational function of a neural network is fully described by the nonlinear function used by its nodes, the topology of the interconnections between nodes, and the "weights" on these interconnections.

While, in general, neural networks may have feedback paths (a directed cycle on the graph representing its topology), this results in a dynamical system which is very complex to analyze. Feed-forward networks are a restricted class of network architectures in which nodes are organized into sequential "layers" and connections are made only from a node to nodes on higher layers. The first layer is the "input" layer, the last is the "output" layer, and any layer in between are "hidden" layers (typically only 1). This class of networks is interesting because the dynamics of the network do not affect the function it is computing (and hence dynamics may be ignored), and the class of functions computable by feedforward networks appears to be quite rich.

The usefulness of FFNN's really involves three separate, though related questions:

1) What functions are representable by FFNN's

2) How easy is it to find these representations (learning)

3) How efficient are these representations

While there are many groups investigating representation and learning in FFNN's, the question of efficiency is usually ignored. This is unfortunate, because it is unlikely that the technology of neural networks will be developed until the computations being performed in them are known to be more

efficient than solutions that use existing computational technology (von Neuman, systolic array, etc...).

There are many ways to pose the question of efficiency in FFNN's, including power consumption, physical size, accuracy, robustness, etc... More important than any of these, however, is the question of scalability: how does the network representation of a function "scale" with function "size" (dimensionality). This metric is known for conventional models of computation, and it is a relevant and well-defined measure by which to compare the efficiency of computation in FFNN's.

For most interesting classes of functions, the question of scalability cannot be answered analytically, and for this reason, any investigation of scalability must address issue of learnability. For these functions, the only way to determine whether a given network is capable of representing a particular function is to try to find such a representation via a "learning algorithm," typically gradient descent or backpropogation. These algorithms search the "weightspace" of a network for a point (a set of weight values) at which the "error" between the target function and the network function is equal to zero.

A question related to that of scalability, which is of interest for the reasons just given, is that of the relationship between scalability and learnability: for a function of a given size, does the network representation of the function become easier to find as the network size is increased? (Does learning become easier as the size of the hidden layer is increased?). Clearly any measure of "learnability" depends on the learning algorithm being used. Despite this, because most iterative improvement algorithms have problems searching spaces where "bad" (unacceptable) local minima abound, we have chosen "abundance/badness of local minima" as a useful and semi-universal measure of the "learnability." "Abundance" we measure probabilistically, while "badness" is a description of a local minima which includes depth, basin of attraction, etc...

It should be clear that our investigation involves a great deal of simulation of FFNN's. In particular, while most existing research has concentrated on simulation of small "toy" networks, we are interested in scalability and thus need to be able to simulate large networks in reasonable times. Prohibitively long simulation times are the major reason past work has not addressed scalability. On a conventional sequential computer, FFNN simulation is O(#nodes squared) and this is unacceptable for large simulations. For this reason, we have developed a massively parallel FFNN simulator which runs on the

Connection Machine. This SIMD computer has 64,000 1-bit processors connected in a hypercube. Our simulator on the CM runs in O(log(#nodes)) and can simulate networks of up to 250 nodes per layer. Currently, the simulator is in the final stages of testing.

We also need function classes which are interesting and scalable. Currently, we have several classes in mind:

1) random boolean functions

2) structured boolean functions (pla, etc...)

3) real-valued functions

4) random network functions (a random set of weight values)

5) decoding of error-correcting codes

6) real-world problems (ala expert systems)

7) recognition of geometric patterns

For the moment, we are concentrating on boolean functions.

State - Analysis of Simple Neural Nets
C.H. Séquin and C. Chinrungrueng

1. Motivation

Neural networks — here understood to mean highly interconnected networks of some kind of summing, thresholding, or switching elements — provide an intriguing alternative to other computing circuitry such as the classical von Neumann computer or systolic signal processing pipelines. The true potential of neural networks and their most fruitful areas of application are still a topic of hot debate. This debate will continue until one has developed a quantitative understanding of their properties such as knowledge storage capacity, information retrieval speed, and learning performance.

One goal of our research is to fully map out the error surfaces of a few very simple networks and to make them somehow visible to the reader. These error surfaces are given as some measure of the overall deviation of the actual outputs of the network from the desired outputs (summed over all input patterns) as a function of all possible states of the network as given by the values of all its (synaptic) weights. It is our belief that from these simple examples we can gain some insight into the structure of the error surfaces of more complicated neural networks. A parallel goal is to develop new visualization techniques for these high-dimensional error surfaces.

2. The Networks

The problem we have concentrated on so far is the Boolean exclusive-or (XOR) function of two inputs. This is one of the simplest problems that cannot be solved with a simple Perceptron with a single threshold element (Fig. 1a). If we disallow feedback within the same layer, then we need a feed-forward network with at least two layers of switching elements to construct a network that can perform the XOR function.

The particular networks considered in this study are shown in Fig. 1. The elements in the input layer simply provide an impedance match to the input signals, limit the signals to the amplitude range considered at all the other switching elements, and provide unlimited fanout. The switching elements, or 'neurons' in the middle layer, also called the 'hidden' elements, are connected to some or all of the input elements through adjustable weights. Their outputs are not directly observable, but provide — again through

adjustable weights — inputs for the neurons in the output layer (just one in our case). To feed an input directly to an output neuron, one of the hidden neurons can be replaced with a dummy element with a single input with a fixed weight of one (see h3 in Fig. 1c).
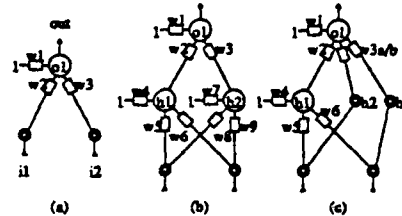

Figure 1. *Three networks used in our analysis and their labeling of the weights:*
(a) Perceptron, (b) regular 3-layer net, (c) net with missing hidden units.

The neurons themselves can take several different implementations. Common to all of them is the fact that they form a weighted sum of all their inputs multiplied with the corresponding weights on those connections. In addition, a bias term that does not depend on any input is added to the sum; in effect this permits one to adjust the switching threshold. This sum is subjected to a 'switching' function that typically takes the form of a sharp thresholding function or, in our case, a smooth monotonic sigmoid function. We have assumed these functions to be symmetrical with respect to the origin in order not to camouflage any inherent symmetries that arise from the symmetry of the network or of the specified problem. We also normalize the amplitudes, so our network inputs and outputs from any neuron will be in the range [-1 . . +1].

3. Error Surfaces

Let us consider the network shown in Fig. 1b. The state of this network is fully defined by the nine adjustable weights w1 through w9; thus the state space of this network is 9-dimensional. If these weights are known, we can compute the network output for each pattern of inputs assuming a smooth sigmoid switching function *swf* inside the neurons:

$$h1(i1,i2) = swf\ (w4 + w5 \times i1 + w6 \times i2)$$
$$h2(i1,i2) = swf\ (w7 + w8 \times i1 + w9 \times i2)$$
$$o1(i1,i2) = swf\ (w1 + w2 \times h1 + w3 \times h2).$$

If we further assume that for each pair of inputs the goal function goal(i1,i2) is known, then we can also compute an error for each input pattern:

$$err(i1,i2) = goal(i1, i2) - o1(i1,i2).$$

Finally we can combine all these error terms into a single term that reflects how well the network overall performs its designated function. We are really interested in the maximum error made for any of the specified input patterns, so a function like *MaxErr* might be most appropriate:

$$MaxErr = max(err(i1, i2)) \text{ for } all\ legal\ input\ patterns\ (i1, i2).$$

However, computationally it is often preferable to have a smooth function to work with during the analysis. Thus often a total square error, *TsqErr*, is computed by summing the squares of the individual errors:

$$TsqErr = \sum_{i1,i2} (err(i1,i2))^2.$$

For a given network and a given problem statement, such an error surface $Err(w1 \cdots w9)$, defined over the domain of the state space, fully describes the possible states of the net and how well it performs for any combination of values of its weights. If we could globally examine this surface, we could readily answer the questions how well a network could perform a given task. Unfortunately these surfaces can be of very high dimension — for this simple example it is already a 9-dimensional hyper-surface embedded in 10-dimensional space.

For the simplest cases we can actually model the error surface. This is illustrated with the trivial network given in Fig. 1a. While the system basically has three weights and thus three degrees of freedom, the symmetry of the problem, i.e. the equivalence of the two inputs allows us to assume that w2 and w3 are identical and then restrict us to exploring the 2-dimensional manifold produced as a function of w1 and of w2=w3.
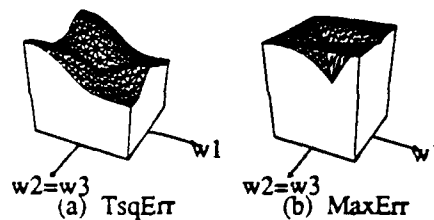
Figure 2. *Error surfaces for XOR function on net 1a.*

Figure 2 shows two surfaces for this network if the desired function is the Boolean XOR operation between the two inputs. Fig. 2a shows the normalized total error $E(x,y) = \sqrt{(TsqErr/4)}$ summed over all four different input patterns. The range of the two weight variables is from -3.5 to +3.5 each. In Fig. 2b we

have plotted instead the maximum error for any of the four patterns. Notice that no part of this surface comes down near the ground plane as would be indicative of a possible solution. This is just another way of realizing that a single neuron (a Perceptron) cannot perform the XOR function.
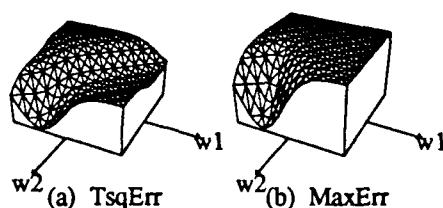

Figure 3. *Error surfaces for XOR function on net 1c.*

If we add a second neuron in the middle (hidden) level (Fig. 1c) then we can find a set of weights that can produce a solution to the XOR problem. In Figure 3 the error function is displayed as a function of the threshold w1 and the weight w2. The diagonal valley goes indeed deep enough to constitute a solution if the weight range is increased to ± 6.0. The other weights are set to: w4=3.0 and w5=w6=3.2.

The network in Fig. 1b is already complicated enough that a single surface in 3-dimensional space is inadequate to represent its solution space. In this case we need to rely on more mathematical techniques to obtain information about its error surface and possible solution states.

4. Symmetries

Often the state space of a given network for a given problem is rather symmetric. This can reduce the scope of the visualization task. The net in Figure 1b has two equivalent hidden neurons that are connected in the same manner to the rest of the network. Thus the role of these two neurons could be interchanged without affecting the overall i/o behavior of the net as a whole. Specifically that means the function of the network will be unchanged if we make the following four substitutions:

opA: w2 ←→ w3, w4 ←→ w7, w5 ←→ w8, and w6 ←→ w9.

Similarly, we have identified five more symmetry operations:

opB: w5 ←→ w6 and w8 ←→ w9.

opC: w2 → -w2, w4 → -w4, w5 → -w5, and w6 → -w6.

opD: w3 → -w3, w7 → -w7, w8 → -w8, and w9 → -w9.

opE: w5 → -w5, w8 → -w8; and w1 → -w1, w2 → -w2, w3 → -w3.

opF: w6 → -w6, w9 → -w9; and w1 → -w1, w2 → -w2, w3 → -w3.

These six operations result in a 64-fold symmetry for the state space of net 1b computing the XOR function.

Such symmetries are very important to discover because they reduce the fraction of the state space that needs to be explored in order to gain a full understanding of the behavior of the neural net. We thus only need to sample a suitably defined region in the 9-dimensional space that amounts to 1/64th of the total domain.

5. Sampling Results

Within one representative domain, we now analyze the function by sampling it at random with several hundred arbitrary starting points, from where we start a steepest descent algorithm to find the corresponding local minimum. It turns out that, for a representative domain of 1/64th of that total state space, we find only three separate local minima. The error *TsqErr* and the corresponding values of the synaptic weights are listed in Table 1.

Table 1. Minima in XOR Function on Net 3b.										
Mult.	Error	w1	w2	w3	w4	w5	w6	w7	w8	w9
4	0.080	2.86	-3.00	-3.00	3.00	-3.00	3.00	3.00	3.00	-3.00
2	2.097	0.00	-3.00	3.00	3.00	-1.06	3.00	3.00	1.06	3.00
2	2.884	0.82	-0.88	-0.74	3.00	-3.00	3.00	3.00	-3.00	3.00

There is only one 'good' minimum that has an error value that is low enough so that the state of this minimum can be considered a solution of the network for the XOR problem.

Because of the 64-fold symmetry of the state space for this problem, each such minimum will be replicated in 64 locations. However some of these replicas may fall on top of one another. The first column in the above table indicates the multiplicity of the observed minima. The good minimum (line 1) has a mul-

-38-

tiplicity of 4, and thus there will be a total of 16 such locations in the whole state space.

6. Ongoing Work

We are now in the process of determining the shape of the various local minima. In particular we are trying to determine:

- The maximum extension of the collection zone, i.e., what is the farthest known source point from the final local minima point where it comes to rest ?

- The maximum radius of complete collection, i.e., what is the largest 9-dimensional sphere around a local minima, inside which all starting points will converge to this minimum ?

- The minimum barrier height for escape from any minimum, i.e., what are the lowest barrier heights between any pair of adjacent minima ?

Once this static analysis of the error surface is completed, we will then study the dynamic aspects of the learning behavior of these networks. This will involve an analysis of the geodesic trajectories that the system takes when it converges to one of the local minima under the gradient descent or the back-propagation algorithm.

PROUD: An Efficient Sea-of-Gates Placement Algorithm for Dense Chip

Ren-Song Tsay and E. S. Kuh

As Application Specific Integrated Circuits (ASIC) become more popular and the gate count on a chip gets increasingly larger, the quality and speed of automatic layout algorithms need to be read-dressed. It is well-known that the problem of module placement is especially crucial to the final outcome of the design. For a sea-of-gates chip with 100K or more gates, conventional deterministic methods are no longer suitable because sparsity of the connectivity specification ought to be considered. Random algorithms such as simulated annealing could take a forbidden amount of time in reaching an acceptable solution.

We introduce a hierarchical method of placement which takes full advantage of the sparsity inherent in the placement specification. The method starts from a quadratic placement formulation and takes the I/O pad specification as input and solves successively *linear* sparse equations. The slot constraints are considered gradually by a simple partitioning on the global placement results solved from linear equations. After partitioning, we use an interesting global optimization technique to handle the inter-influence among separated regions. From the result of several real chips, the quality of the placement is excellent. It is superior to that obtained by TimberWolf3.2 and is comparable with that of TimberWolf4.2. The run time complexity of our method is $O(n \log^2 n)$ and the memory space complexity is linear. For a 100K sea-of-gates chip with 26,000 modules, the run time is about 50 minutes on a VAX 8650 (6 MIPs machine) and the memory requirement is about 11 megabytes.

The algorithm has been extended to do placement on designs with mixed big macros and standard cells. We also find a solution for I/O pad assignment from the general quadratic formulation but using eigenvector approach instead of solving linear equations.

Parallel implementation of the PROUD placement algorithm is now under study.

MPC.2. Systems Problems in Parallel Distributed Computation

Shobana Venkatraman
(Professor Shankar Sastry)

The convergence and adaptation properties of neural networks, their memory capacities and programmability have not been studied completely thus far. We are investigating these properties for a general class of neural networks of which the Hopfield and Grossberg networks are special cases. We have found that in such networks it may be possible to establish very clearly defined basins of attraction around each of the equilibria (the nodes we wish to "remember"). We have also been studying the effect of the hamming distance between equilibria on the location of the spurious equilibria, if any, that are generated. We are working on making the algorithm for the synthesis of this class of neural networks more programmable, in the sense of being able to update a network built to remember 'm' nodes to store 'm+1' equilibria without having to restructure the entire network. We are studying the characteristics of these artificial neural networks by simulating the nonlinear differential equations that describe them.

MPC.3. Interconnection of Dense Computational and Memory Elements:

Implementation of Learning-Capable Neural Nets in an Analog MOS VLSI Environment
Victor Hu and Alan Kramer
(Professors Ping K. Ko and Chenming Hu)

The goal of this project is to investigate the feasibility of implementing learn-capable neural nets using analog MOS VLSI technology. While there is a great deal of variety in approaches to neural-net computation, most proposed architectures are characterized by the weighted interconnection of simple nonlinear threshold computing elements. These connection weights, which define the computational contents and functions of the network, are considered one of the most critical elements in its practical implementation.

The computational capabilities of these networks have been demonstrated using pre-programmed resistors or digitally weighted resistors in the interconnections. However, the real potential of neural computation lies with a fully analog and learn-capable network. We are exploring the possibility of using field alterable resistive elements in this network.

In the past few months, we have focused on the use of the floating-gate EEPROM devices. We have evaluated the analog programming characteristics of commercial EEPROM devices as well as test devices we obtained through our industrial contacts. The results so far have been encouraging and suggest that analog programming of a EEPROM device is indeed feasible and controllable.

In order to build a MOS neural net chip, however, a integrated MOS/EEPROM technology is necessary. We are currently designing an inhouse technology with that capability. A NMOS/EEPROM test run will be started once the test mask design is completed. Several designs of a simple demonstration neural-net chip using the back-propagation algorithm are being investigated.

High-Speed MOSFET/BJT Technology

James Moon
(Professors Ping K. Ko and Chenming Hu)

We are developing a high-speed BJT/CMOS technology for the implementation of analog and digital circuits.

For the MOS transistors, this technology features Elevated Drain and Source (EDS) regions that are grown epitaxially from the substrate, and the Contact-Over-Oxide (COO) structure developed earlier in our laboratory. The COO contact scheme occupies a much smaller area than conventional junction contacts, has significantly lower junction-to-substrate capacitances, and does not require a diffusion barrier layer for shallow junctions. In addition, this structure also promise greater radiation hardness to SEU (Single Event Upset) than conventional techniques. With the EDS scheme, an LDD (Lightly Doped Drain) structure for hot-electron protection can be realized simply by controlling the implant energies of the n- and n+ regions, instead of using the more difficult oxide-spacer technology. One of the goals in our device design is to make the devices operational even at channel lengths down to 0.1 to 0.2 μm. For the bipolar transistor, we are investigating the designs of both lateral and vertical injection structures. Simplifying the device structure and making it MOS technology compatible are our main goals.

Since the project started six months ago, we've developed a deep-submicron lithographic process which is based on conventional optical lithographic in conjunction with a novel photoresist ashing technique. Using this process, we are able to fabricate FET's with controlled channel lengths in the 1/4 μm regime routinely. Some of the important achievements so far are: (1) An n-channel NMOS transistor with channel length of 0.15 μm and transconductance of 650 mS/mm at room temperature - the highest reported for the MOS transistor. (2) A propagation delay of 21 ps per stage has been achieved for a NMOS depletion-load logic gate chain at room temperature - the fastest reported for a MOS logic circuit.

We have fabricated n-channel MOSFETs with large variations in device dimensions and process parameters. A comprehensive study was performed on the characteristics of these devices. Important scaling guidelines for device performance and reliability have been established.

At present, we are focusing on various technology integration issues. We have also begun designing test devices and circuits for evaluating the capabilities of the technology.

Process Simulation for Extraction of Topography Dependent Electrical Characteristics:

Application to VLSI Neural Networks
Edward Scheckler
(Professor A. R. Neureuther)

The implementation of highly interconnected computational structures, such as neural networks, is critically dependent on minimizing performance degradation from parasitic loading effects. The layout and topographical features in processing determine the magnitude of parasitic loading. Efforts to link profiles generated by the SIMPL and SAMPLE process simulators to programs for device and RC-parasitic modeling are being pursued in order to investigate questions of VLSI neural network performance.

Simulations of prototype networks have yielded basic information on parasitic effects. Some general trends regarding speed limitations and scaling effects with larger, higher density networks have been identified for Hopfield networks based on the implementation developed at AT&T Bell Labs. Investigation of other architectures, including prototypes being developed here by Professor P. K. Ko, is being pursued. General trends in neural network research are being followed to identify which areas are most significantly affected by VLSI issues.

APPENDIX

413

Modeling of Substrate Current in p-MOSFET's

T.-C. ONG, P. K. KO, AND CHENMING HU, SENIOR MEMBER, IEEE

Abstract—It is shown that the substrate current characterization method and modeling approach used for n-MOSFET's is also applicable to p-MOSFET's. The impact ionization rate extracted for holes is found to be $8 \times 10^6 \exp(-3.7 \times 10^6/E)$, where E is the electric field. Based on our measurement and modeling result, roughly twice the channel electric field is required for p-MOSFET's to generate the same amount of substrate current as n-MOSFET's. The hot-carrier-induced breakdown voltage is therefore also about two times larger.

I. INTRODUCTION

IT HAS LONG been recognized that hot-carrier effects impose a limit upon future device scaling. Hot-carrier effects in n-MOSFET's, which can be correlated with the substrate current [1], are more serious than in p-MOSFET's. Models of the substrate current in n-MOSFET's have been well developed [2], [3]. While there have been studies of hot-carrier effects in p-MOSFET's [4]–[6], it is not clear that the substrate current model for n-MOSFET's is also applicable to p-MOSFET's. In this paper, we show that the channel electric fields and the substrate currents in n-MOSFET's and p-MOSFET's can be modeled by the same physical approach, although the values of the semiempirical parameters are different. The test devices used in this study are n^+ polysilicon-gate MOSFET's without threshold adjust implant. Three kinds of gate oxide thicknesses were fabricated for both n- and p-MOSFET's: 432, 152, and 90 Å. The substrate concentration is 2×10^{16} cm^{-3} for all the n-MOSFET's and 4×10^{16} cm^{-3} for all the p-MOSFET's. The source/drain junctions are 0.3-μm arsenic junctions in the n-MOSFET's and 0.4-μm boron junctions in the p-MOSFET's.

II. CHANNEL FIELD AND IMPACT IONIZATION COEFFICIENTS

It has been shown that for conventional n-MOSFET's the maximum lateral channel field at the drain end E_m can be expressed in terms of the drain voltage V_D as [7], [8]

$$E_m = \sqrt{\left(\frac{V_D - V_{D\,SAT}}{l}\right)^2 + E_{SAT}^2} \approx \frac{V_D - V_{D\,SAT}}{l} \quad (1)$$

$$l = 0.22\, T_{ox}^{0.33}\, X_j^{0.5} \quad (2)$$

where E_{SAT} is the channel field at which the carriers reach

saturation velocity. The corresponding drain voltage is $V_{D\,SAT}$. T_{ox} is the gate oxide thickness and X_j is the source/drain junction depth. E_{SAT} is about 4×10^4 V/cm for n-MOSFET's. Using this relationship, the substrate current I_{SUB} can be expressed as

$$I_{SUB} = \frac{I_D A_i E_m l}{B_i} \exp(-B_i/E_m) \quad (3)$$

where A_i and B_i are the constants in the expression for the impact ionization rate, i.e., $\alpha = A_i \exp(-B_i/E)$, and I_D is the drain current. Equations (1)–(3) were derived and have been verified by experiments and simulations for n-MOSFET's [8]. However, there is nothing in the derivations to suggest that they are not equally applicable to p-MOSFET's.

$V_{D\,SAT}$ can be determined experimentally for n-MOSFET's [9]. The same method is applied to p- and n-MOSFET's as shown in Fig. 1. For p-MOSFET's, parallel shift of the constant I_{SUB}/I_D locus is observed and $V_{D\,SAT}$ can be experimentally determined as well. From (1)–(3), a plot of $\ln(I_{SUB}/(I_D(V_D - V_{D\,SAT})))$ versus $1/(V_D - V_{D\,SAT})$ will yield one single straight line for all the devices and bias voltages. The vertical intercept and slope of this line give A_i/B_i and B_i, respectively. In Fig. 2, such a plot is shown for both the n-MOSFET's and the p-MOSFET's. Using $E_{SAT} = 1.2 \times 10^5$ V/cm, a single straight line also is obtained for different sizes of p-MOSFET's at different biases. This proves that (1)–(3) are applicable to p-MOSFET's, too. Using (2) for l, A_i and B_i deduced from Fig. 2 are

n-MOSFET: $B_i = 1.7 \times 10^6$ V · cm^{-1}, $A_i = 2 \times 10^6$ cm^{-1}

p-MOSFET: $B_i = 3.7 \times 10^6$ V · cm^{-1}, $A_i = 8 \times 10^6$ cm^{-1}.

A_i of the p-MOSFET is within the range of published data [10]–[12] ($\approx 4 \times 10^7$ cm^{-1} in [10], 2.25×10^6 cm^{-1} in [11], and 2×10^6 cm^{-1} in [12]). The values of B_i are close to those obtained from p-n junction experiments ($\approx 3.45 \times 10^6$ V·cm^{-1} in [10] and 3.26×10^6 V·cm^{-1} in [11]), suggesting that (2) is applicable to p-MOSFET's as well.

After A_i and B_i for holes are extracted from measurements, I_{SUB} of the p-MOSFET's can be calculated. A physical expression for $V_{D\,SAT}$ [3]

$$V_{D\,SAT} = -\frac{|V_G - V_t| E_{SAT} L_{eff}}{|V_G - V_t| + E_{SAT} L_{eff}} \quad (4)$$

is found to fit the experimental $V_{D\,SAT}$ very well and is used in the calculation of E_m. V_t is the extrapolated threshold voltage.
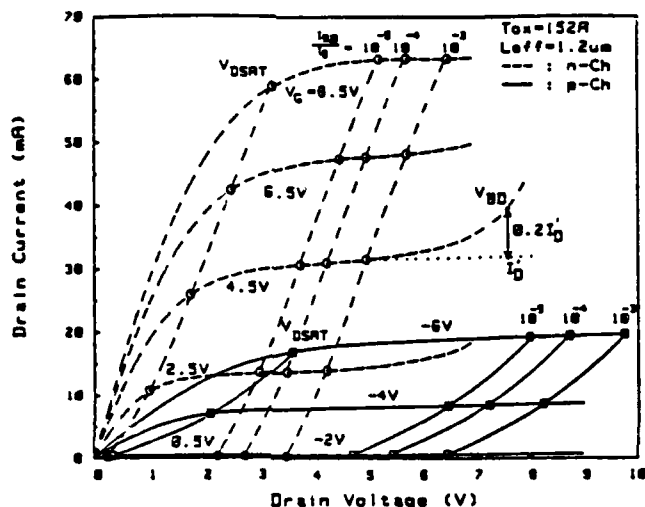
Fig. 1.   Constant I_{SUB}/I_D loci and $V_{D\,SAT}$ loci are superimposed on the I_D-V_D characteristics of an n- and a p-MOSFET. V_{DSAT} loci are determined with the method described in [9]. T_{ox} = 152 Å.
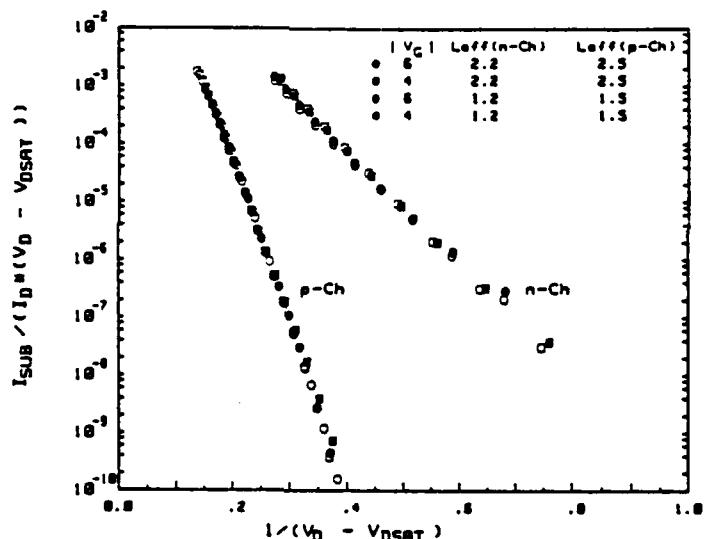

Fig. 2.   Plot of log $[I_{SUB}/(I_D(V_D-V_{DSAT})]$ versus $1/(V_D-V_{DSAT})$ to show the different impact ionization rates between n- and p-MOSFET's. Circle: $|V_G|$ = 6 V; square $|V_G|$ = 4 V. T_{ox} = 152 Å.

I_D in the saturation region is

$$I_D = v_{SAT}\,WC_{ox}(V_G - V_t - V_{D\,SAT}) \qquad (5)$$

where v_{SAT} (= 5.6×10^6 cm/s) is the saturation velocity of holes obtained by fitting the measured I_D with (5) for different channel lengths. W is the device width and C_{ox} is the gate capacitance. The effect of drain voltage on V_t can be incorporated into (5) to improve the accuracy of I_D. I_{SUB} calculated by this simple model is shown in Fig. 3. For buried-channel p-MOSFET's, modification of this model is expected since E_m also depends on the depth and the doping concentration of the buried layer.

In (3), B_i may be interpreted as the ratio of the impact ionization energy Φ_i to the carrier mean free path λ. Φ_i is expected to be very roughly equal for electrons and holes (if

one considers the different effective masses of holes and electrons, holes can have a larger Φ_i than electrons [13], [14]). Given this assumption and λ = 92 Å for electrons [15], we find λ = 43 Å for holes. Notice that $B_i\lambda$ of the p-MOSFET's is higher than that of the n-MOSFET's by a factor of 2.2. This means the p-MOSFET's can take twice the V_D-$V_{D\,SAT}$ to generate the same I_{SUB} (see (1)–(3)) as n-MOSFET's. As a result, the hot-carrier-induced breakdown voltage of the p-MOSFET is also about two times larger as shown in Fig. 4. The hot-carrier-induced breakdown voltage V_{BD} is defined as the drain voltage at which the drain current is 20 percent larger than in the saturation region as illustrated in Fig. 1. In Fig. 4, notice that for the 152- and the 90-Å devices, the V_{BD}-V_G curves of both the n- and p-MOSFET's show the C-shape characteristics of hot-carrier-induced breakdown [16], while
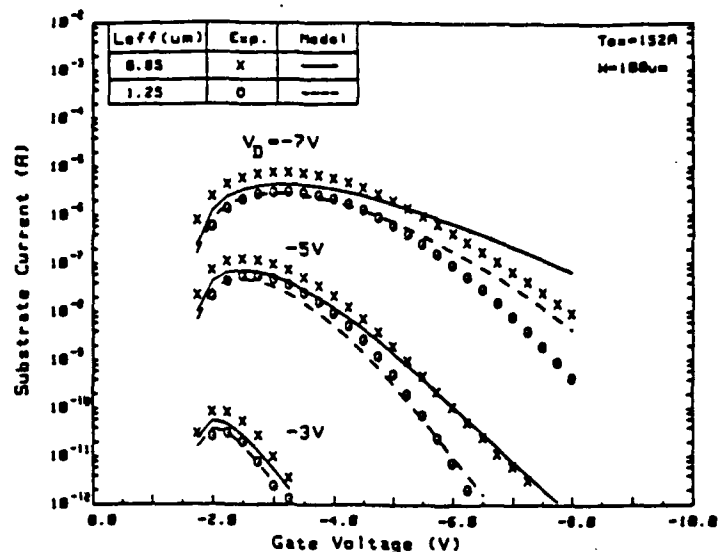
Fig. 3. I_{SUB}–V_G characteristics of p-MOSFET's. L_{eff} = 0.85 and 1.25 μm. T_{ox} = 152 Å.
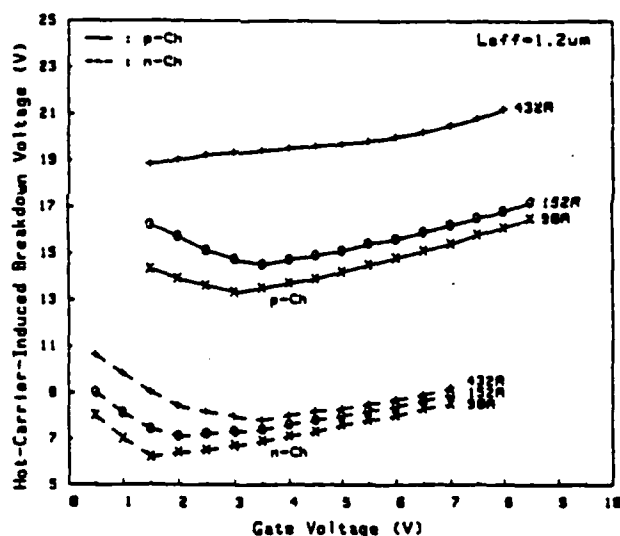
Fig. 4. Breakdown voltage V_{BD} versus V_G for n- and p-MOSFET's with T_{ox} = 432, 152, and 90 Å.

for the 432-Å devices the C-shape is observed only in the n-MOSFET's. The breakdown voltage of the 432-Å p-MOSFET's is about 20 V, limited by the p^+-n junction breakdown.

III. CONCLUSIONS

The dependence of the maximum channel field on terminal voltages is found to be the same in conventional n- and p-MOSFET's. The substrate currents of p- and n-MOSFET's thus can be characterized with the same method. Substrate current measurement and modeling reveal that p-MOSFET's can withstand twice the channel field of n-MOSFET's. For small-geometry p-MOSFET's, the breakdown voltage also is dominated by hot-carrier-induced breakdown as in n-MOS-FET's, but the value is about twice as large.

REFERENCES

[1] C. Hu, S. Tam, F.-C. Hsu, P. K. Ko, T. Y. Chan, and K. W. Kyle, "Hot-electron-induced MOSFET degradation—Model, monitor, and improvement," IEEE, Trans. Electron Devices, vol. ED-32, pp. 375–385, Feb. 1985.
[2] C. Hu, "Lucky electron model of channel hot electron emission," in IEDM Tech. Dig., 1979, pp. 22–25.
[3] C. Hu, "Hot-electron effects in MOSFETs," in IEDM Tech. Dig., 1983, pp. 176–181.
[4] E. Takeda, Y. Nakagome, H. Kume, N. Suzuki, and H. Asai, "Comparison of characteristics of n-channel and p-channel MOSFETs for VLSI's," IEEE Trans. Electron Devices, vol. ED-30, pp. 675–680, June 1983.
[5] R. Radojcic, "Hot-electron aging in p-channel MOSFETs for VLSI

CMOS." *IEEE Trans. Electron Devices*, vol. ED-31, p. 1896, Dec. 1984.

[6] T. Tsuchiya and J. Frey, "Relationship between hot-electron/holes and degradation of p- and n-channel MOSFETs," *IEEE Electron Device Lett.*, vol. EDL-6, pp. 8–11, Jan. 1985.

[7] Y. A. Elmansy and D. M. Caughey, "Modeling weak avalanche multiplication currents in MOSFETs in IGFETs and SOS transistors for CAD," in *IEDM Tech. Dig.*, 1975, pp. 31–34.

[8] T. Y. Chan, P. K. Ko, and C. Hu, "Dependence of channel electric field in device scaling," *IEEE Electron Device Lett.*, vol. EDL-6, pp. 551–553, Oct. 1985.

[9] T. Y. Chan, P. K. Ko, and C. Hu, "A simple method to characterize substrate current in MOSFETs," *IEEE Electron Device Lett.*, vol. EDL-5, pp. 505–507, Dec. 1984.

[10] C. A. Lee, R. A. Logan, R. L. Bardorf, J. J. Kleimack, and W. Wiegmann, *Phys. Rev.*, vol. 134, p. 761, 1964.

[11] S. M. Sze, *Physics of Semiconductor Devices*, 1st ed. New York: Wiley, p. 62.

[12] W. N. Grant, "Electron and hole ionization rates in epitaxial silicon at high electric field," *Solid-State Electron.*, vol. 16, p. 1189, 1973.

[13] C. T. Sah, J. Y. C. Sun, and J. J. Tzou, "Generation-annealing kinetics and atomic models of a compensating donor in the surface charge layer of oxidized silicon," *J. Appl. Phys.*, vol. 54, p. 944, 1983.

[14] J. L. Moll and R. Van Overstraeten, "Charge multiplication in silicon p-n junctions," *Solid-State Electron.*, vol. 6, pp. 147–157, 1963.

[15] S. Tam, P. K. Ko, and C. Hu, "Lucky-electron model of channel hot-electron injection in MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-31, no. 9, pp. 1116–1125, Sept. 1984.

[16] F.-C. Hsu, P. K. Ko, S. Tam, C. Hu, and R. S. Muller, "An analytical breakdown model for short-Channel MOSFETs," *IEEE Trans. Electron Devices*, vol. ED-29, no. 11, pp. 1735–1740, Nov. 1982.

Subbreakdown Drain Leakage Current in MOSFET

J. CHEN, T. Y. CHAN, I. C. CHEN, P. K. KO, AND CHENMING HU, SENIOR MEMBER, IEEE

Abstract—Significant drain leakage current can be detected at drain voltages much lower than the breakdown voltage. This subbreakdown leakage can dominate the drain leakage current at zero V_G in thin-oxide MOSFET's. The mechanism is shown to be band-to-band tunneling in Si in the drain/gate overlap region. In order to limit the leakage current to 0.1 pA/μm, the oxide field in the gate-to-drain overlap region must be limited to 2.2 MV/cm. This may set another constraint for oxide thickness or power supply voltage.

I. INTRODUCTION

IN A MOSFET with a thin gate oxide and with the gate grounded, significant drain leakage current can be observed at drain voltages much below what is usually considered the "breakdown voltage" [1]. We shall refer to this as the subbreakdown leakage current. This subbreakdown leakage current is also observable in the subthreshold I_D-V_G characteristics of thin-oxide devices. Fig. 1 shows the typical subthreshold characteristics for a 88-Å gate oxide n-MOSFET. The large leakage current near zero V_G is another manifestation of the subbreakdown leakage. It increases with increasing V_D and decreasing V_G. A good understanding and control of this MOSFET subbreakdown leakage current is very important to thin-oxide VLSI MOSFET's and some high-voltage devices such as nonvolatile memory cells or programming circuits.

Deep-depletion breakdown in MOS capacitors has been studied in detail [2]. In [1], Feng *et al.* have studied the effect of scaling oxide thickness and channel doping concentration on the drain avalanche breakdown voltage. This paper explores the mechanisms responsible for the MOSFET drain subbreakdown leakage current.

II. EXPERIMENT

Measurements were performed on n-channel MOSFET's fabricated with conventional n-type polysilicon gate technology. The p-bulk was doped at 10^{16} cm^{-3} and the source/drain diffusions were 0.4-μm As junctions. The channel was 20 μm long and 20 μm wide. Gate oxide thicknesses ranging from 55 to 350 Å were examined.

Fig. 2 shows the measured drain current at $V_G = 0$ V for several oxide thicknesses. These currents flow from the drain to the substrate. For the 155-Å gate oxide device, data are presented for three different temperatures. 25, 100, and 150°C. As temperature increased, a V_D-independent leakage
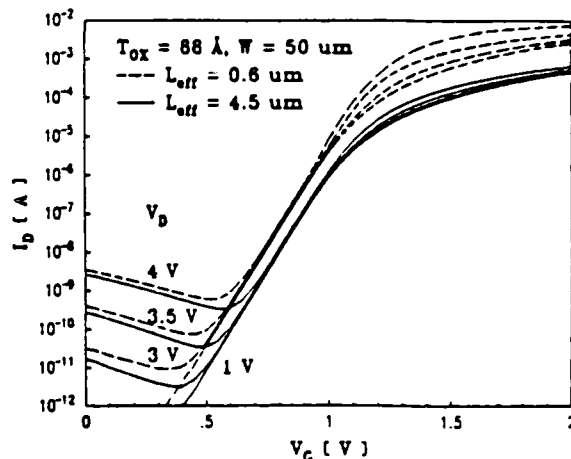
Fig. 1. Subthreshold characteristics for two n-MOSFET's with T_{ox} = 88 Å, W = 50 μm. and L = 4.5 and 0.6 μm. Significant drain leakage current can be observed when V_{DG} is high.
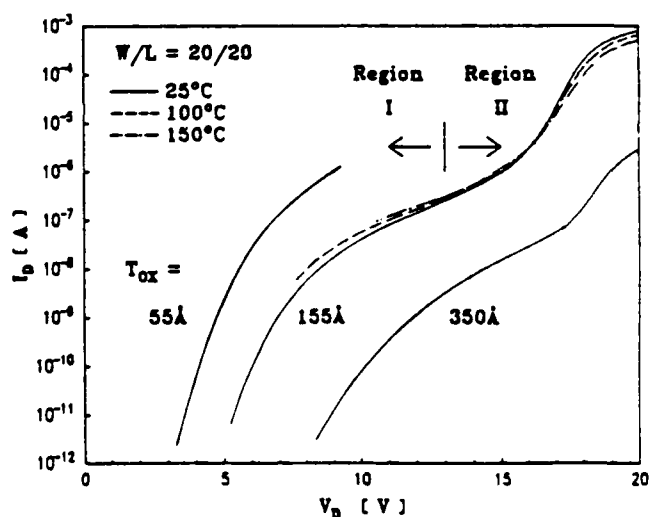

Fig. 2. Drain current characteristics for n-MOSFET's at different temperatures with gate grounded. A V_D-independent leakage current. believed to be the thermal generation current, has been subtracted for clarity. The drain current characteristics for devices with T_{ox} = 55 and 350 Å at room temperature are also included.

current (2.28 × 10^{-9} at 100°C and 1.17 × 10^{-7} at 150°C). believed to be the thermal generation currents collected by the drain junction, appeared as expected. The data shown in Fig. 2 have this current component subtracted for clarity.

III. MODEL AND DISCUSSION

One may attempt to attribute the subbreakdown leakage current to thermal current amplified by impact ionization. The
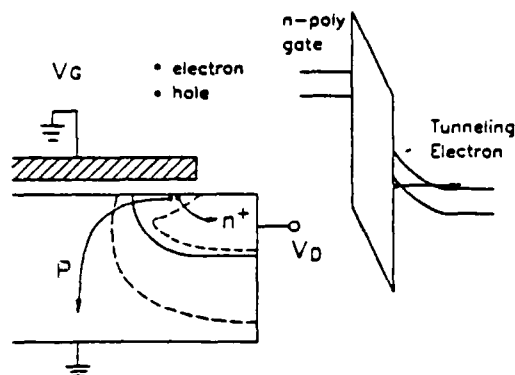
Fig. 3. A deep-depletion region is formed in the gate-to-drain overlap region. The energy-band diagram illustrates the band-to-band tunneling process and the flow of carriers. Valence-band electrons tunnel into the conduction band and are collected at the drain. The holes created flow to the substrate.
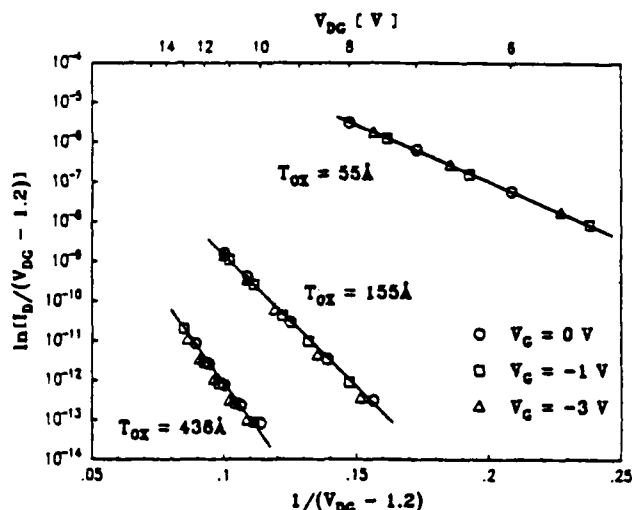

Fig. 4. Plot of ln $(I_D/(V_{DG} - 1.2))$ versus $1/(V_{DG} - 1.2)$, where V_{DG} is the voltage difference between the drain and gate. All measured data with the same oxide thickness fall on a straight line, which is in agreement with the band-to-band tunneling model.

data at different temperatures show that this leakage current is insensitive to temperature while the thermal leakage current varies by more than 10^5 times over the temperature range. This indicates that the subbreakdown drain current is not amplified thermal leakage current. Another possible origin of this subbreakdown current is the amplification of the electron tunneling current from the gate. However, the drain current is more than three orders of magnitude larger than the gate current which rules out this possibility.

We interpret the subbreakdown current as due to the band-to-band tunneling process in the gate-to-drain overlap region as illustrated in Fig. 3. Drain current is due to the tunneling of valence-band electrons into the conduction band. The holes created by the tunneling of electrons flow to the substrate. Note that tunneling is only possible in the presence of a high electric field. The field in silicon at the Si–SiO$_2$ interface depends on the doping concentration in the diffusion region and the difference between V_D and V_G, i.e., V_{DG}.

Band-to-band tunneling current density is the highest where the electric field is the largest and the band bending is larger than the energy bandgap E_g. A simple expression for the surface electric field at the tunneling point in the gate-to-drain overlap region can be obtained as follows:

$$E_s = \frac{V_{DG} - 1.2}{3 T_{ox}} \qquad (1)$$

where E_s is the vertical electric field at silicon surface, 3 is the ratio of silicon permittivity to oxide permittivity, and T_{ox} is the oxide thickness in the overlap region. A band bending of 1.2 eV is the minimum necessary for band-to-band tunneling to occur. The theory of tunneling current predicts [3]

$$I_D = AE_s \exp\left[-\frac{\pi m^{*1/2} E_g^{3/2}}{2\sqrt{2}q\hbar E_s}\right] = AE_s \exp\left(-B/E_s\right) \qquad (2)$$

where A is a preexponential constant and $B = 21.3$ MV/cm with $m^* = 0.2m_o$ [4]. According to (1), E_s is proportional to

$V_{DG} - 1.2$. Fig. 4 shows that measured ln $(I_D/(V_{DG} - 1.2))$ plotted against $1/(V_{DG} - 1.2)$ is linear in agreement with (2). Measured data yield an experimental value of $B = 18$ MV/cm. This is in reasonable agreement with the theoretical value of $B = 21.3$ MV/cm, which has never been verified to this accuracy. Fig. 4 also verifies that the tunneling current only depends on the vertical field, namely, the difference between drain and gate voltages as in [1]. These facts strongly support our interpretation that the currents are due to band-to-band tunneling. Using measured A and B, a drain leakage current of $I_D = 0.1$ pA/μm can be created when E_s is 0.75 MV/cm and the oxide field is about 2.2 MV/cm. Empirically, we have found $I_D = 0.1$ pA/μm when $V_{DG} \cong 1.2 + T_{tox} \times E_{crit}$ with E_{crit} ranging from 2.2 to 3 MV/cm for a wide range of MOSFET technologies, where T_{tox} is the measured gate (channel) oxide thickness. The fact that E_{crit} is larger than 2.2 MV/cm suggests that the oxide near the gate edge is thicker than T_{tox} due to polysilicon gate reoxidation. This sets another limit to the minimum oxide thickness or the maximum power supply voltage.

As seen in Fig. 2, there is a noticeable kink in the drain I-V characteristics. So far we have only addressed Region I, to the left of the kink. In Region I. I_D increases slightly with increasing temperature probably due to a decrease in E_g in (2). In Region II, I_D rises above the predicted tunneling current. We attribute the rise to the amplification of the tunneling current by avalanche impact ionization. I_D in Region II is characterized by a negative temperature coefficient due to the decrease of the impact ionization coefficient at higher temperature and by the sensitivity to the substrate bias (not presented here), of which I_D in Region I is independent.

IV. SUMMARY

Subbreakdown leakage is the significant drain leakage current observable at below-breakdown voltage. The domi-

nant mechanism for this current is band-to-band tunneling at the Si–SiO$_2$ interface in the gate-to-drain overlap region. At higher drain voltages, impact ionization amplifies this current and results in even larger leakage currents until breakdown occurs.

In thin oxide devices, the OFF-state drain current often increases with decreasing V_G measurably due to the same mechanism (Fig. 1). The voltage required to cause band-to-band tunneling will decrease when oxide thickness decreases. This may be another barrier to VLSI MOSFET scaling. Possible solutions to the problem are the use of lightly-doped drain devices which have lower E_s, or graded gate oxide devices [5] in which the oxide is thicker in the gate-to-drain overlap region than that in the channel.

REFERENCES

[1] W. S. Feng, T. Y. Chan, and C. Hu, "MOSFET drain breakdown voltage," IEEE Electron Device Lett., vol. EDL-7, p. 449, 1986.
[2] E. H. Nicollian and J. R. Brews, MOS Physics and Technology. New York: Wiley, 1982, p. 378.
[3] J. L. Moll, Physics of Semiconductor. New York: McGraw-Hill, 1964, p. 253.
[4] S. M. Sze, Physics of Semiconductor Devices. 2nd ed. New York: Wiley, 1981, p. 525.
[5] P. K. Ko, S. Tam, and C. Hu, "Enhancement of hot-electron currents in graded-gate-oxide (GGO)-MOSFETS," in 1984 IEDM Tech. Dig., p. 88.

TECHNICAL DIGEST

iedm

international
ELECTRON
DEVICES
meeting

1987

WASHINGTON, D.C.
DECEMBER 6-9, 1987

The Impact of Gate-Induced Drain Leakage Current on MOSFET Scaling

T.Y. Chan*, J. Chen, P.K. Ko and C. Hu

Electronics Research Laboratory
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley CA 94720

Abstract

Significant gate-induced drain leakage current can be detected in thin gate oxide MOSFETs at drain voltages much lower than the junction breakdown voltage. This current is found to be due to the band-to-band tunneling occurring in the deep-depletion layer in the gate-to-drain overlap region. In order to limit the leakage current to 0.1pA/μm, the oxide field in the gate-to-drain overlap region must be limited to 1.9MV/cm. This may set another constraint for the power supply voltage and/or oxide thickness in VLSI MOSFET scaling. Device design considerations for minimizing the gate-induced drain leakage current are discussed.

I. Introduction

As MOS technology is scaled down to the sub-half micron regime, it is anticipated that the power supply voltage *will need to be reduced to below 5 volts: Hot-electron reliability*, oxide reliability, and punch-through are widely discussed reasons[1]. This paper reports a new mechanism - Gate-Induced Drain Leakage Current - that imposes an additional constraint on the acceptable supply voltage and/or oxide thickness. Significant drain leakage current can be detected at drain voltage much lower than the breakdown voltage. This drain leakage current is caused by the gate-induced high electric field in the gate-to-drain overlap region and can dominate the drain leakage current at zero gate bias in thin oxide MOS-FETs. A model based on band-to-band tunneling is proposed and can satisfactorily explain the experimental results.

Its dependence on fabrication process and device parameters has been carefully studied. The exact oxide thickness and doping profiles in the gate-to-drain overlap region are found to play important roles in the gate-induced drain leakage current. The guidelines for designing thin oxide VLSI MOSFETs and high voltage devices such as those used in non-volatile memory circuits are discussed.

II. Gate-Induced Drain Leakage Current Characteristics

A typical subthreshold I_D–V_G characteristics for 88-$\overset{\circ}{A}$ gate oxide n-MOSFETs with two different channel lengths are shown in Fig. 1. Significant amount of drain leakage current near zero V_G is observed. This drain leakage current increases

---

* Present address : Intel Corporation, 3065 Bowers Avenue, Santa Clara, CA. 95051.

with increasing V_D and decreasing V_G, and shows no apparent dependence on channel length. Fig. 2 shows the off-state (V_G = 0V) drain leakage currents for several oxide thicknesses. These currents flow from the drain to the substrate. For the 155-$\overset{\circ}{A}$ gate oxide device, data are presented for three different temperatures - 25, 100, and 150°C. A V_D-independent leakage current, believed to be the thermal generation current, has been subtracted from the high temperature curves for clarity.

III. Leakage Mechanism

One may attempt to attribute the off-state drain leakage current to thermal current amplified by impact ionization. The insensitivity of the leakage current to temperature shown in Fig. 2 indicates that the drain leakage current is not amplified thermal leakage current. Another possible origin of this drain leakage current is the amplification of the electron tunneling current from the gate. However, the drain current is more than three orders of magnitude higher than the gate current which rules out this possibility.

We attribute the gate-induced drain leakage current to the band-to-band tunneling process in silicon in the gate-to-drain overlap region as illustrated in Fig. 3. The cross-section shown in Fig. 3 is simply a gated-diode configuration. When high voltage is applied to the drain with the gate grounded, a deep-depletion region is formed underneath the gate-to-drain overlap region. Electron-hole pairs are generated by the tunneling of valence band electrons into the conduction band and collected by the drain and substrate, separately. Since all the minority carriers generated thermally or by band-to-band tunneling in the drain region (holes in n-MOSFETs and electrons in p-MOSFETs) flow to the substrate due to the lateral field, the deep-depletion region is always present and the band-to-band tunneling process can continue without creating an inversion layer.

Band-to-band tunneling is only possible in the presence of a high electric field and when the band bending is larger than the energy band gap, E_g. The field in silicon at the Si-SiO$_2$ interface also depends on the doping concentration in the diffusion region and the difference between V_D and V_G, i.e. V_{DG}. A simple expression for the surface field at the dominant tunneling point can be expressed as

$$E_s = \frac{V_{DG} - 1.2}{3T_{ox}} \qquad (1)$$

where E_s is the vertical electrical field at silicon surface, 3 is the ratio of silicon permittivity to oxide permittivity, and T_{ox}

is the oxide thickness in the overlap region. A band bending of 1.2 V is the minimum necessary for tunneling process to occur. The theory of tunneling current predicts [2]

$$I_D = AE_s\exp(-B/E_s) \tag{2}$$

where A is a pre-exponential constant and $B = 21.3$ MV/cm [3]. Fig. 4 shows the measured $\ln[I_D/(V_{DG}-1.2)]$ plotted against $1/(V_{DG}-1.2)$. All measured data with the same oxide thickness fall on straight lines, in agreement with the tunneling theory (Eq. (2)). Fig. 4 also verifies that the tunneling current only depends on the vertical field, namely, the voltage difference between the drain and gate. The slope of the lines in Fig. 4 is $3T_{ox}B$ (Eqs. (1) and (2)) and is plotted in Fig. 5. An experimental value of $B = 18$ MV/cm was obtained, in reasonable agreement with the theoretical value of $B = 21.3$ MV/cm [4]. These facts strongly support our interpretation that the currents are due to the band-to-band tunneling.

At high V_{DG}, I_D was found to rise above the tunneling current predicted by Eq. (2) as shown in Fig. 6. We attribute the rise to the amplification of the tunneling current by avalanche impact ionization which accounts for the negative temperature coefficient due to the decrease of the impact ionization coefficient at higher temperature (Region II as indicated in Fig. 2) and the sensitivity to the substrate bias (Fig. 7), of which I_D at low V_{DG} (Region I as indicated in Figs. 2 and 7) is independent. The kink observed in n-MOSFETs shown in Fig. 6 is due to hole trapping in the oxide which affects the increase of the surface field.

IV. Impact on MOSFET Scaling and Design

To limit the undesirable gate-induced drain leakage current to 0.1pA per μm channel width, the oxide field in the gate-to-drain overlap region must be limited to 1.9MV/cm as can be obtained in Fig. 8. This sets another limit to the minimum oxide thickness or the maximum power supply voltage in MOSFET scaling,

$$V_{CC} = 1.2 + T_{ox} \times 1.9 MV/cm \tag{3}$$

With this constraint, a 100-$\mathring{A}$ device can only tolerate a maximum supply voltage of 3.1V. The measured V_{DG} for $I_D = 0.1pA/\mu m$ is usually larger than predicted by Eq. (3) depending on the process technologies as demonstrated in Fig. 8. This is due to the larger oxide thickness near the gate edge resulted from source-drain re-oxidation. Since the gate-induced drain leakage current depends on the oxide field, the doping type of the polysilicon gate used in thin oxide MOSFETs strongly affects the acceptable supply voltage. The acceptable V_{CC} for 0.1pA/μm is expected to be smaller by about 1V if the poly-gate and the drain region have opposite doping types, e.g. p-MOSFET with n-type poly-gate, because of the work function difference. This gives another reason to use p$^+$ poly-gate in sub-half micron p-MOSFETs, in addition to the concern for making surface-channel devices.

Possible means to reduce this gate-induced drain leakage current are the use of lightly-doped (LDD) devices or graded gate oxide devices [5]. However, the effectiveness of using LDD device strongly depends on the doping concentration in the n$^-$ region. Fig. 9 shows the calculated V_{DG} needed to create an oxide field of 1.9MV/cm in the gate-drain overlap region and induce 1.2V band bending as a function of the doping concentration. A minimum is obtained around 10^{-5}cm^{-3}. This suggests that LDD MOSFETs must have a n$^-$ doping lower than 10^{18}cm^{-3} to raise the acceptable V_{CC} over that of

non-LDD MOSFETs. LDD MOSFETs usually have a moderately doped n$^-$ region in order to reduce the series resistance and minimize the hot-electron degradation. If the doping concentration is higher than 10^{18}cm^{-3} at the gate edge, a point with doping concentration of 10^{18}cm^{-3} always exists in the gate/drain overlap region, which makes LDD MOSFETs no better than conventional MOSFETs for the gate-induced drain leakage current. Buried LDD MOSFETs with the peak n$^-$ doping concentration several hundred angstroms underneath the Si-SiO$_2$ interface seems to be the more favorable device structure for both hot-electron reliability [6] and gate-induced drain leakage current.

Since the T_{ox} to be used in Eq. (3) is that present over the point where the drain doping concentration is 10^{18}cm^{-3}, graded gate oxide device structure could effectively reduce the leakage current. However, the weakened gate control near the gate edge might become a serious concern in device reliability [7]. Judicious design of the gate edge profile to compromise device reliability and gate-induced drain leakage current is necessary if the graded gate oxide device structure is adopted. The gate-induced drain leakage current in n-MOSFETs can also be significantly enhanced after hot-electron stressing due to electron trapping in the oxide and at the interface as shown in Fig. 10, while p-MOSFETs show continuous decrease in the drain leakage current because the dominant trapped charges are still electrons. This may pose a long-term device drift issue in n-channel devices.

V. Conclusion

The gate-induced drain leakage current will inevitably become an important issue in device design and fabrication when thinner gate oxides are used in sub-micron and deep sub-micron MOSFETs. The voltage required to cause band-to-band tunneling decreases when oxide thickness decreases. Since the band-to-band tunneling process is caused by the high oxide field in the gate-to-drain overlap region, the exact oxide thickness and doping profiles in the gate/drain overlap region require great care in current and future device design and fabrication.

References

[1] Masakazu Kakumu, Masaaki Kinugawa, Kazuhiko Hashimoto and Jun'ichi Matsunaga, 1986 IEDM Technical Digest, pp. 399-402.

[2] J.L. Moll, "Physics of Semiconductor," McGraw Hill, P. 253, 1964.

[3] S.M. Sze, "Physcis of Semiconductor Devices," 2nd edition, Wiley New York, p. 525, 1981.

[4] J. Chen, T.Y. Chan, I.C. Chen, P.K. Ko, and C. Hu, IEEE Electron Device Lett., November 1987.

[5] P.K. Ko, S. Tam and C. Hu, 1984 IEDM Technical Digest, p. 88.

[6] Hugh R. Grinolds, Masaaki Kinugawa, and Masakazu Kakumu, 1985 IEDM Technical Digest, p. 246.

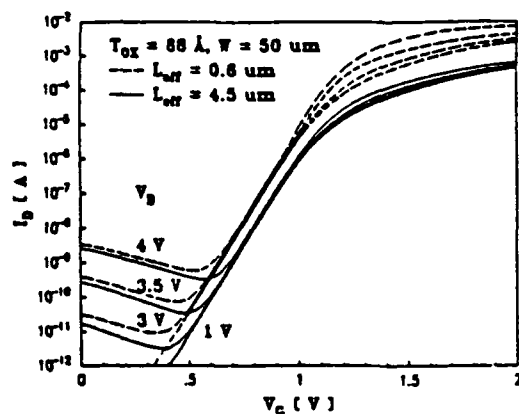[7] T.Y. Chan, A.T. Wu, P.K. Ko, and C. Hu, IEEE Electron Device Lett., vol. EDL-8, p. 326, July 1987.

31.3

Fig. 1 Subthreshold characteristics of thin oxide n-MOSFETs with $T_{ox} = 88\mathring{A}$, and $L_{eff} = 0.6$ and 4.5 μm. Significant off-state ($V_G = 0V$) drain leakage current can be observed when V_D is high.
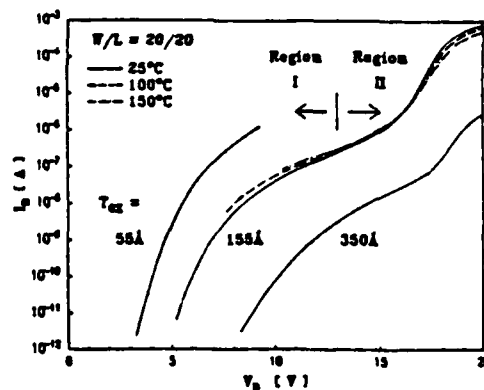

Fig. 2 Drain current characteristics for n-MOSFETs with gate grounded. The drain leakage currents show strong dependence on the oxide thickness, but are insensitive to temperature. A V_D-independent leakage current, believed to be the thermal generation current, has been subtracted for clarity.
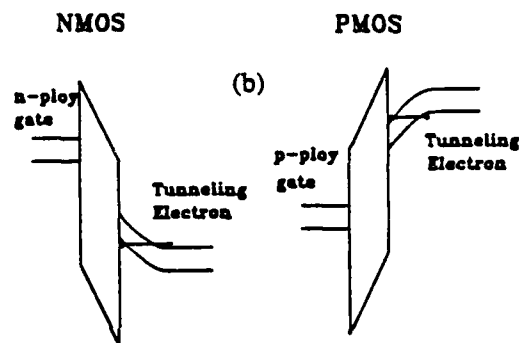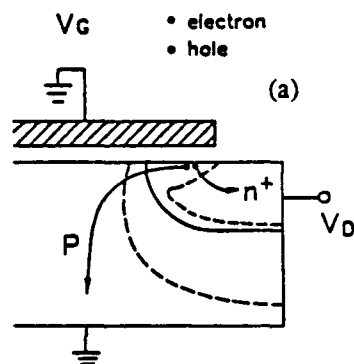

Fig. 3 (a) A deep-depletion region is formed in the gate/drain overlap region. (b) The energy band diagram illustrates the band-to-band tunneling process in silicon in the gate/drain overlap region. Electron-hole pairs are generated by the tunneling of valence band electrons into the conduction band and collected by the drain and substrate, separately.
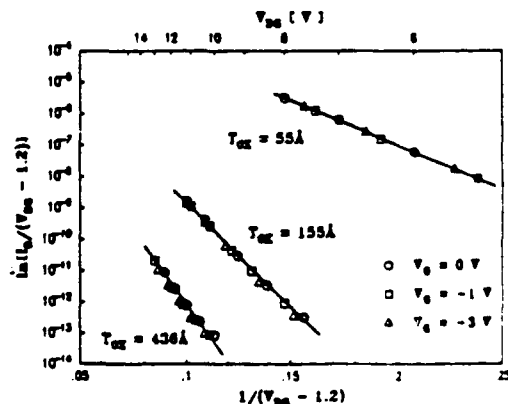
Fig. 4 Plot of $\ln[I_D/(V_D-1.2)]$ versus ⌈⋯⌉, where V_{DG} is the voltage difference between the drain and gate. All measured data with the same oxide thickness fall on straight lines, which is in agreement with the band-to-band tunneling theory.
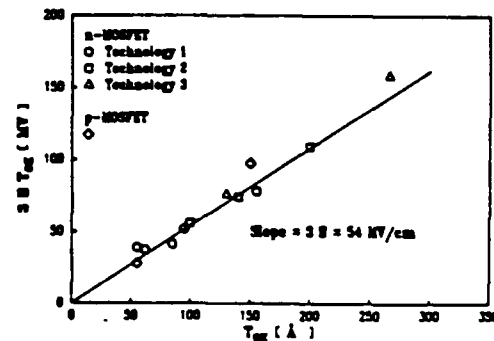
Fig. 5 The slope, $3T_{ox}$ B, of the lines in Fig. 4 as a function of the oxide thickness for both n- and p-MOSFETs. An experimental value of B = 18 MV/cm was obtained, in agreement with the theoretical value of B = 21.3 MV/cm.

31.3
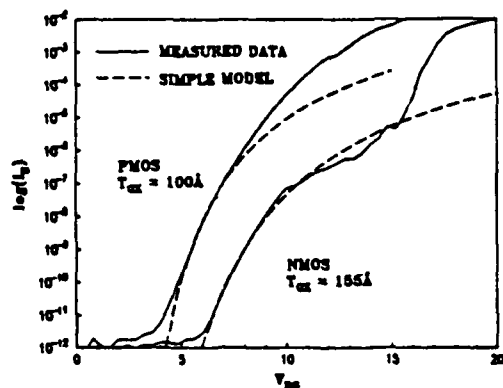
Fig. 6 I_D–V_{DG} characteristics from measurement and predicted by the simple model. The rise of measured I_D above the predicted tunneling current is attributed to the amplification of the tunneling current by avalanche impact ionization. The kink observed in n-MOSFETs is due to hole trapping in the oxide.
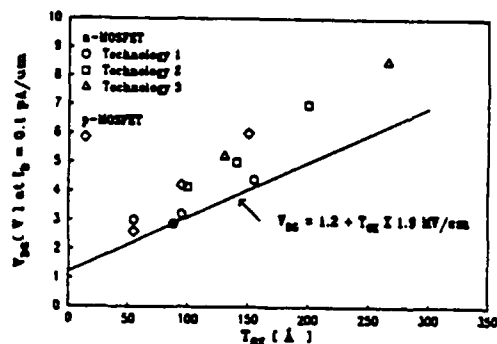

Fig. 8 Measured V_{DG} for 0.1pA/µm drain leakage current as a function of oxide thickness for both n- and p-MOSFETs. It is found that $I_D = 0.1$pA/µm when $V_{DG} = 1.2 + T_{ox} \times$ 1.9MV/cm. The higher V_{DG} for some process technologies is due to the larger oxide thickness near the gate edge resulted from source-drain re-oxidation.

(a)
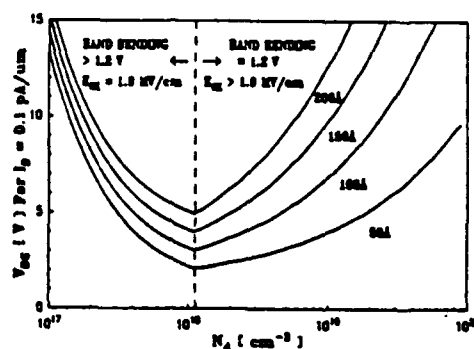
Fig. 9 V_{DG} needed to create an oxide field of 1.9MV/cm in the gate/drain overlap region and to induce 1.2V band bending as a function of the doping concentration. A minimum is observed around $10^{18} cm^{-3}$.

(b)

Fig. 7 I_D–V_{DG} characteristics at different substrate bias for (a) n-MOSFET and (b) p-MOSFET. The sensitivity of the drain leakage current to the substrate bias at high V_{DG} indicates the amplification of the tunneling current by avalanche impact ionization.
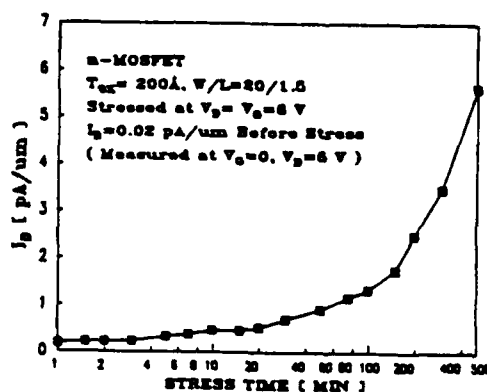
Fig. 10 The gate-induced drain leakage current for a n-MOSFET measured at $V_G = 0$V and $V_D = 5$V as a function of the stressing time. The leakage current is significantly enhanced after hot-electron stressing due to electron trapping in the oxide.

31.3

Ultra-Thin Silicon-Dioxide Breakdown Characteristics of MOS Devices with n$^+$ and p$^+$ Polysilicon Gates

S. HOLLAND, I. C. CHEN, AND CHENMING HU, SENIOR MEMBER, IEEE

Abstract—In this work we investigate the effect of the gate material on the breakdown characteristics of ultra-thin silicon dioxide films at low voltages (<6 V). When MOS capacitors are stressed with a positive gate voltage, the charge to breakdown and time to breakdown at a fixed oxide-voltage drop are significantly smaller in p$^+$ polysilicon-gate capacitors than in n$^+$ polysilicon-gate capacitors. The results are interpreted in terms of a simple model of hole tunneling resulting from hot-hole generation in the anode by hot electrons entering from the silicon dioxide. Extrapolation of high-voltage-breakdown lifetime measurements for relatively thick-oxide devices to low voltages may be complicated by this mechanism.

I. INTRODUCTION

BREAKDOWN in thin films of silicon dioxide has received considerable attention in the recent technical literature [1]-[5]. In these previous studies the voltage applied during the breakdown measurement is usually greater than 10 V, which is much larger than the actual power-supply voltage. Hence the study of silicon-dioxide breakdown in the low-voltage regime (≈ 5 V) is of much interest.

The investigation of low-voltage breakdown in a conveniently short time requires the use of utlra-thin oxides (thinner than ≈ 5 nm). For oxide thicknesses in this range it has been reported that the charge to breakdown Q_{BD} (the total charge injected into the oxide necessary to cause breakdown) and time to breakdown t_{BD} are strongly dependent on the stressing electric field [3], [6], more so than in thicker oxides. In addition, ultra-thin silicon dioxide films are being considered for subhalf-micrometer MOS transistors [7], [8] and therefore the study of breakdown in these films is of practical importance.

II. SAMPLE PREPARATION

Polysilicon-gate MOS capacitors were fabricated for this study. The starting substrates were 4-in-diameter 6–12-$\Omega\cdot$cm n-type and 15–30-$\Omega\cdot$cm p-type. Gate oxides were grown at 800°C after patterning and etching of a steam-grown field oxide. Before gate oxidation the furnace tube was cleaned for 8 h at 1050°C in an O$_2$/trichlorethane (TCA) ambient. Polysilicon was deposited undoped at 650°C immediately after

gate oxidation. For each wafer the polysilicon was doped p$^+$ on one half and n$^+$ on the other half by using a CVD oxide as a diffusion mask. The n$^+$ polysilicon was doped from a POCl$_3$ source at 950°C and the p$^+$ polysilicon was doped from a solid source containing B$_2$O$_3$ at 900°C.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Fig. 1 shows the measured t_{BD} and Q_{BD} versus the oxide-voltage drop V_{ox} during stress for MOS capacitors of area 1 × 10^{-6} cm^2 and thickness 3.3 nm. Data are shown for both p$^+$ and n$^+$ polysilicon-gate devices. The solid curve is a theoretical plot that will be discussed later. The gate voltage during stress was held constant (positive with respect to the substrate). The tunneling current during stress remained essentially constant until breakdown occurred. In some cases the tunneling current was observed to increase slightly during the measurement indicating the possibility of positive charge trapping. Breakdown was detected as a sudden large increase in the current through the device. V_{ox} is the actual voltage drop across the silicon dioxide, i.e., the applied voltage minus the work-function difference between the polysilicon and n-type substrate minus any band bending. For the p$^+$ polysilicon-gate devices a work-function difference plus band bending of 1.1 V was assumed. We found that for the n$^+$ polysilicon-gate devices a work-function difference plus band bending of 0.4 V was necessary in order to have equal tunneling currents on a gate current versus V_{ox} plot (equal current implies equal V_{ox}). Similar experiments on magnesium- and aluminum-gate MOS capacitors yielded Q_{BD} values on the order of 0.1 C/cm^2 [9], and hence metal-gate devices are not suitable for this type of study.

As can be seen in Fig. 1, there is a reduction in t_{BD} and Q_{BD} for the p$^+$ polysilicon-gate devices of approximately three orders of magnitude when compared to the n$^+$ devices. For this experiment the cathode (the negatively biased electrode) during stress is the n-type substrate and hence is the same for both types of devices. The anode during stress (the polysilicon gate) is not the same, however.

Fig. 2 shows the results of a similar experiment except that in this case the gate is biased negative with respect to the substrate. The substrate is p-type and the capacitor area is 1 × 10^{-4} cm^2. The solid curve is a theoretical plot. A constant gate voltage was applied and as before the tunneling current during stress was essentially constant. For this case. the t_{BD} at a fixed current density (hence fixed oxide-voltage drop) is indepen-
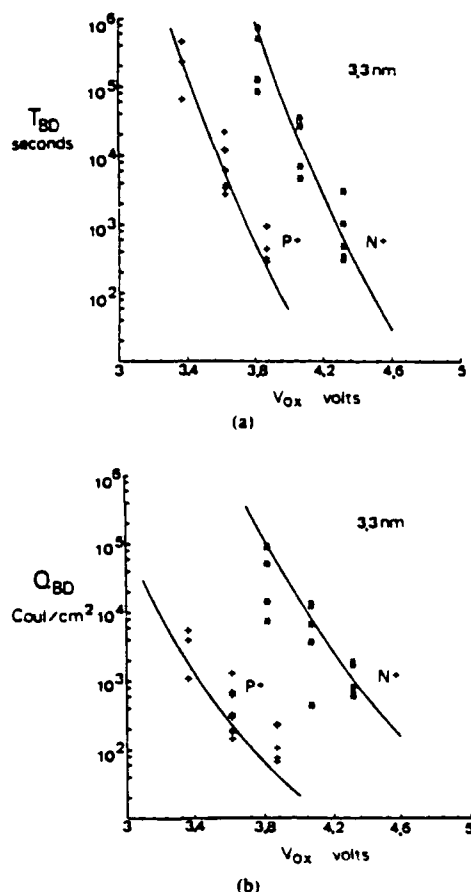
(a)

(b)

Fig. 1. (a) Time to breakdown t_{BD} and (b) charge to breakdown Q_{BD} versus oxide-voltage drop comparing n$^+$ and p$^+$ polysilicon gate MOS capacitors (oxide thickness = 3.3 nm, capacitor area = 1 × 10^{-4} cm^2). The gate voltage was constant and positive with respect to the substrate. All devices are from the same wafer. The solid curve is the theoretical model.
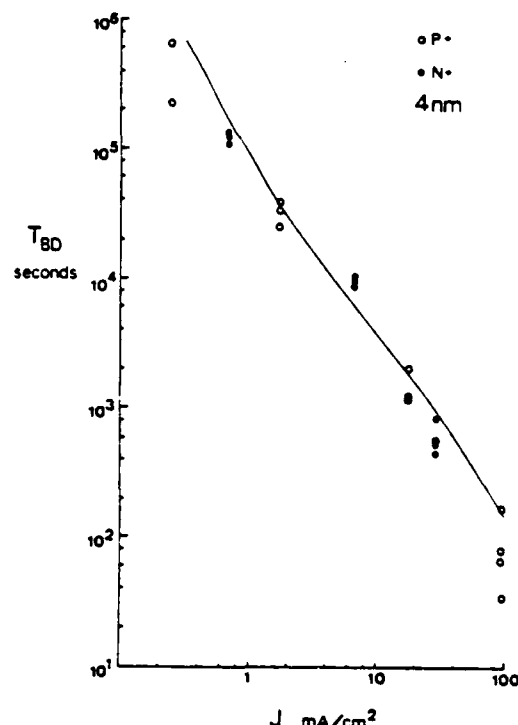

Fig. 2. Time to breakdown t_{BD} versus stressing current density J comparing n$^+$ and p$^+$ polysilicon gate MOS capacitors (oxide thickness = 4 nm, capacitor area = 1 × 10^{-4} cm^2). The gate voltage was constant and negative with respect to the substrate. All devices are from the same wafer. The solid curve is the theoretical model.

dent of gate type. In this experiment the anode during stress (the p-type substrate) is the same for both types of devices.

Since the breakdown characteristics of Figs. 1 and 2 are compared at equal oxide-voltage drops, the energy gained by the tunneling electrons traversing the silicon dioxide should be the same for both types of devices. Hence it is unlikely that the breakdown is a result of hot-electron effects in the silicon dioxide itself. Also, the maximum potential energy that can be gained by the tunneling electrons in the silicon dioxide is $q(V_{ox} - \phi_n)$, where ϕ_n is the barrier height between the silicon and silicon dioxide. For this set of experiments the maximum energy gained in the silicon dioxide is approximately 1–2 V, which is probably too small to cause impact ionization in the silicon dioxide.

However, the electrons will enter the anode electrode with qV_{ox} electronvolts of energy relative to the conduction band in the anode if scattering in the silicon dioxide is negligible as shown in Fig. 3. Since this energy can be substantially greater than the bandgap of silicon, it is possible to generate hole-electron pairs in the anode electrode [10]. It is thus possible to generate hot holes, a fraction of which can tunnel into the silicon dioxide as originally proposed by Weinberg et al. [11]. These holes will tend to increase the cathode electric field which in turn could lead to breakdown. In the n$^+$-gate case the

valence-band electrons must be excited to the bottom of the conduction band. However, in the p$^-$ case the valence-band electrons only need to excited to the top of the valence band since there are available states there. Thus, for the p$^-$ case the barrier height for hole tunneling will be 1.1 eV (the silicon bandgap) lower than in the n$^+$ case as shown in Fig. 3.

If the fraction of tunneling electrons which give up all of their energy by creating hole–electron pairs is α_H, then the hole-tunneling current arising from hole generation in the anode will be given by $J_p = \alpha_H J_n \Theta_H$ where J_n is the electron tunneling current from the cathode and Θ_H is the hole-tunneling probability. The charge associated with the trapped holes at time t is thus $Q_p(t) = \eta J_p t$ where η is the hole trapping efficiency. Assuming that breakdown occurs when a critical number of holes have tunneled into the SiO$_2$ and have been trapped [5], then Q_{BD} will be given by

$$Q_{BD} = J_n t_{BD} = C\Theta^{-1} = C \exp\left(\frac{B'X_{ox}}{V_{ox}}[\phi_p(V_{ox})]^{3/2}\right) \quad (1)$$

$$\phi_p(V_{ox}) = E_{g,\text{SiO}_2} - q\phi_n - qV_{ox} - E_{g,\text{Si}} \quad \text{p}^+ \text{ polysilicon}$$

$$\phi_p(V_{ox}) = E_{g,\text{SiO}_2} - q\phi_n - qV_{ox} \quad \text{n}^+ \text{ polysilicon}$$

where C is a constant, X_{ox} is the oxide thickness, $B' = B/\phi_n^{3/2}$ is the Fowler–Nordheim slope parameter, E_{g,SiO_2} and $E_{g,\text{Si}}$ are the bandgap energies of silicon dioxide and silicon, respectively, and ϕ_n and $\phi_p(V_{ox})$ are the barrier heights for electrons and holes, respectively. (Equation (1) is derived assuming Fowler–Nordheim tunneling and equal electron and
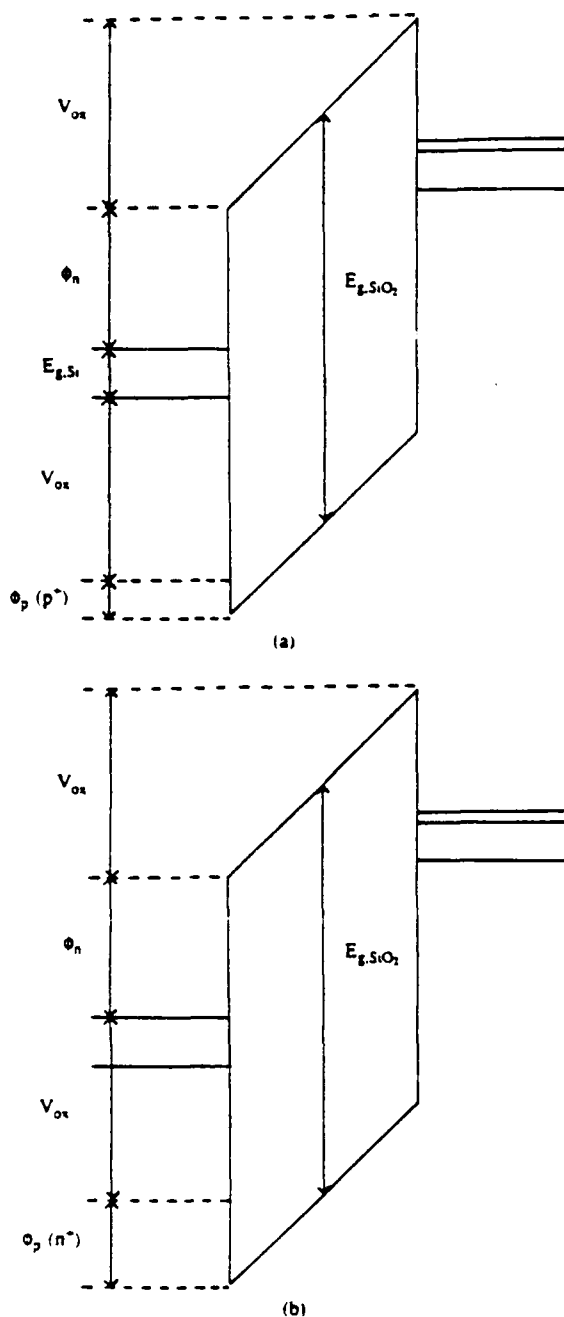
(a)

(b)

Fig. 3. Energy-band diagram depicting the mechanism of hole tunneling due to hole generation in the polysilicon gate (the anode for the case shown). The gate is biased positive with respect to the n-type substrate. ϕ_p is the barrier height for hole tunneling. (a) p⁻ polysilicon gate. (b) n⁻ polysilicon gate.

hole effective masses. In addition, the field dependence in the pre-exponential factor of the Fowler–Nordheim expression has been neglected for simplicity, i.e., $J_n = A \exp [-BX_{ox}/V_{ox}]$ where A is a material constant.) The dependence of t_{BD} on V_{ox} can be determined by dividing (1) by J_n.

The solid curves shown in Figs. 1 and 2 are the theoretical predictions. For Fig. 1 the pre-exponential factor C was arbitrarily fit at one point for the n⁻ data and the same pre-exponential factor was used to generate the theoretical curve for the p⁻ data. The model parameters used were 9 eV for the
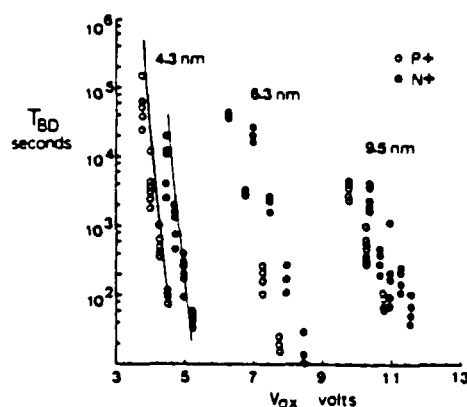
Fig. 4. Time to breakdown t_{BD} versus the oxide voltage drop V_{ox} for oxide thicknesses 4.3, 6.3, and 9.5 nm. The solid curve is the theoretical plot for the 4.3-nm devices. The gate voltage was constant and positive with respect to the substrate.

silicon-dioxide bandgap, 3.2 eV for the barrier height for electron tunneling, and $B'\phi_n = 250$ MV/cm [12]. As is evident from Fig. 1, both the dependence on V_{ox} and the difference in magnitude resulting from p⁺ and n⁻ gate doping are well represented by the simple model. In Fig. 2 V_{ox} was determined from the measured $J_n - V_{ox}$ characteristics.

Fig. 4 compares the effect of both the gate type and oxide thickness on the time to breakdown. Theoretical plots are shown for the 4.3-nm oxide. However, the 6.3- and 9.5-nm oxides will have significant electron scattering in the oxide and hence the simple model presented does not apply directly. In addition, for the thicker oxides the difference in t_{BD} due to gate type is not as large as in the 4.3-nm case, and the difference becomes smaller at higher voltages suggesting that a different mechanism of hole generation may be operable at relatively high voltages. If so, a simple extrapolation of high-voltage t_{BD} measurements may not give a valid estimation of the device lifetime at low voltages.

IV. SUMMARY

In conclusion, we have presented a simple model for the breakdown of ultra-thin silicon-dioxide films at low voltages. The difference between p⁻ and n⁻ polysilicon-gate t_{BD} and Q_{BD} can be explained if one assumes that at low voltages holes are generated in the anode and that some fraction of these holes then tunnel into the oxide. In this paper the discussion has been in terms of V_{ox}. When one compares the results at the same applied voltage the differences due to gate material are not as significant. Hence we do not believe that this effect should determine the choice of n⁻ and p⁻ gate technology.

REFERENCES

[1] S. Lassig and M. S. Liang. "Time-dependent degradation of thin gate oxide under post-oxidation high-temperature anneal." IEEE Electron Device Lett., vol. EDL-8. no. 4. p. 160. Apr. 1987.
[2] Z. A. Weinberg and T. N. Nguyen. "The relation between positive charge and breakdown in metal-oxide-silicon structures." J. Appl. Phys., vol. 61. no. 5. p. 1947. Mar. 1987.
[3] M. S. Liang and J. Y. Choi. "Thickness dependence of oxide breakdown under high field and current stress." Appl. Phys. Lett., vol. 50. no. 2. p. 104. Jan. 1987.
[4] S. Holland and C. Hu. "Correlation between breakdown and process-

induced positive charge trapping in thin thermal SiO_2," *J. Electrochem. Soc.*, vol. 133, no. 8, p. 1705, Aug. 1986.

[5] I. C. Chen, S. Holland, and C. Hu, "Electrical breakdown in thin gate and tunneling oxides," *IEEE Trans. Electron Devices*, vol. ED-32, no. 2, p. 413, Feb. 1985.

[6] I. C. Chen, S. Holland, and C. Hu, "Hole trapping and breakdown in thin SiO_2," *IEEE Electron Device Lett.*, vol. EDL-7, no. 3, p. 164, Mar. 1986.

[7] M. Miyake et al., "Subhalf-micrometer p-channel MOSFET's with 3.5 nm gate oxide fabricated using X-ray lithography," *IEEE Electron Device Lett.*, vol. EDL-8, no. 6, p. 266, June 1987.

[8] T. Kobayashi, S. Horiguchi, and K. Kiuchi, "Deep-submicron MOSFET characteristics with 5 nm gate oxide," in *IEDM Tech. Dig.*, Dec. 1984, p. 414.

[9] S. Holland and C. Hu, unpublished results.

[10] C. Chang, C. Hu, and R. W. Broderson, "Quantum yield of electron impact ionization in silicon," *J. Appl. Phys.*, vol. 57, no. 2, p. 302, Jan. 1985.

[11] Z. A. Weinberg, W. C. Johnson, and M. A. Lampert, "High-field transport in SiO_2 on silicon induced by corona charging of the unmetallized surface," *J. Appl. Phys.*, vol. 47, no. 1, p. 248, Jan. 1976.

[12] Z. A. Weinberg, "On tunneling in metal-oxide-silicon structures," *J. Appl. Phys.*, vol. 53, no. 7, p. 5052, July 1982.

international
ELECTRON DEVICES
meeting

1987

WASHINGTON, D.C.
DECEMBER 6-9, 1987

TECHNICAL DIGEST

PERFORMANCE AND HOT-ELECTRON RELIABILITY
OF DEEP-SUBMICRON MOSFET'S

M.-C. Jeng, J. Chung, A.T. Wu, T.Y. Chan, J. Moon, G. May, P.K. Ko, and C. Hu

Department of Electrical Engineering and Computer Sciences
University of California, Berkeley, CA 94720

ABSTRACT

A study of the performance and hot-electron reliability of submicron n-channel MOSFET's is presented. Well-established hot-electron-based physical models are adequate in explaining the general behaviors of the drain, substrate, and gate currents of these devices. These results suggest that the basic physics is rather well-understood and the design criteria developed for micron-size devices can be extended to cover their deep-submicron counterparts. Hot-electron studies reveal a channel-length dependence in device degradation. This phenomenon together with gate-induced drain leakage current [1] will impose an upper limit on the supply voltage and a lower limit on the gate oxide thickness. Based on device degradation results alone, the power supply voltage for a quarter-micron device with oxide thickness of 86 Å should be limited to 2.5 V if no degradation-resistant structure is used.

INTRODUCTION

Performance and hot-electron reliability of submicron MOSFET's have been the two major concerns for their feasibility in circuit applications. Recent years have seen increased activities in the studies of deep-submicron MOSFET's due to their potential applications in high-performance VLSI circuits and systems [2]-[4]. However, the emphases of most previous reports have been to demonstrate the feasibility of fabricating these devices. This paper attempts to construct a unified understanding of their characteristics, thereby establishing some basic device design guidelines.

The threshold voltage, drain-induced-barrier-lowering (DIBL), transconductance, current driving capability, substrate current, and degradation of submicron devices were characterized for various channel dimensions, doping concentrations, and gate oxide thicknesses. Discussion of the experimental results are presented in later sections.

DEVICE FABRICATION

The devices used in this study were fabricated using an NMOS technology for which the temperature cycle after source/drain implantation (As, $3 \times 10^{15} cm^{-2}$, 50 Kev, 0° inclination) was limited to 900 °C or below to reduce the junction depth. Conventional g-line optical lithography was used to define all levels including that for the gate. The ultra-fine gates were obtained by calibrated thinning of the optically defined photoresist features in oxygen plasma. An SEM photograph of a transistor cross-section with $L_{eff} = 0.22$ μm is shown in Fig. 1. With this technique, we have been able to fabricate deep-submicron MOSFET's routinely with excellent device characteristics using standard semiconductor processing equipment. Transistors of four oxide thicknesses (36 Å, 56 Å, 86 Å, and 155 Å) with substrate dopings ranging from $6 \times 10^{16} cm^{-3}$ to $8 \times 10^{17} cm^{-3}$ were fabricated. The junction depth is about 0.15 μm for all devices. In Fig. 2(a), the strong-inversion drain current of a device with $L_{eff} = 0.15$ μm and $T_{ox} = 36$ Å is shown which exhibits very good characteristics. This device also exhibits very well-behaved subthreshold characteristics (Fig. 2(b)) with insignificant punchthrough problem. The practical channel length appears to be limited by the technology controllability than by any inherent physical mechanism.

DEVICE CHARACTERISTICS
AND DISCUSSION

(A) Drain current and short-channel effects

A good indicator of the extent of short-channel effects for a given process technology is the threshold voltage shift due to charge-sharing and DIBL effects. In Fig. 3(a), the linear-region threshold voltage is plotted versus effective channel length for four oxide thicknesses. The substrate doping for each oxide thickness is chosen so that the threshold voltages for long-channel devices are in the range of 0.5 - 0.7 V. The threshold voltage shift at V_{DS}

31.1

= 3V versus effective channel length is shown in Fig. 3(b). Both figures show that although thicker oxide devices present worse threshold shift as expected, the threshold voltages do not suffer severe change until the effective channel length is shorter than 0.25 μm for all oxide thicknesses.

The transconductance and the saturation drain current of an array of devices are plotted in Fig. 4. Very-well-behaved trends agree with a simple physical model [5]-[6], indicating that the basic physics is well-understood. The model equations are summarized below.

$$I_{DSAT} = W_{eff} v_{sat} C_{ox} (V_{GS} - V_{th} - V_{DSAT}) \quad (1)$$

and

$$g_m = \frac{\partial I_{DSAT}}{\partial V_{GS}} \quad (2)$$

where

$$V_{DSAT} = \frac{E_c L_{eff} (V_{GS} - V_{th})}{E_c L_{eff} + (V_{GS} - V_{th})} \quad (3)$$

$$E_c = \frac{v_{sat}}{2 \mu_{eff}} \quad (4)$$

$$\mu_{eff} = \frac{\mu_0}{1 + (\frac{E_S}{E_0})^n} \quad (5)$$

E_S is the effective vertical field in the channel and can be approximated by $E_S = [2Q_B + C_{ox}(V_{GS}-V_{th})]/2\varepsilon_{si}$. Q_B is the depletion bulk charge, $E_0 = 0.67$ MV/cm, $n = 1.6$, $\mu_0 = 670$ cm^2/V sec, and $v_{sat} = 8 \times 10^6$ cm/sec. The same model parameters were used for all device dimensions and oxide thicknesses. The data shown in Fig. 4 have been corrected for the source/drain and contact resistance ($= 30 \Omega$/side). The inversion-layer capacitance which is more important for thin-oxide devices [7] at low gate bias, has not been included in these equations. The device with 36 Å gate oxide and $L_{eff} = 0.15$ μm exhibits a maximum transconductance of 640 mS/mm, the highest ever reported for a MOSFET at room temperature. While simplistic in formulation, these model equations do provide good physical as well as quantitative understanding of the current performance of MOS devices down to submicron regime. They would also serve well as first order calibrators for process control and diagnosis.

(B) Hot-electron effects

The substrate current peak is higher for a device with shorter channel length and thinner oxide as expected. The peak substrate current at $V_{DS} = 3$ V versus channel length for different oxide thicknesses is shown in Fig. 5. More informative

plots of I_{SUB}/I_{DS} versus $1/(V_{DS} - V_{DSAT})$ [8] for devices with $L_{eff} = 0.3$ μm and different gate oxide thicknesses are shown in Fig. 6. The straight lines suggest that the same basic physical mechanism prevail independent of channel length. However, the slopes for the lines seem to deviate from the $T_{ox}^{-1/3}$ dependence predicted by the model.

In the degradation studies, the device lifetimes show a dependence on the channel length. Results for devices with 86 Å gate oxide are presented in Fig. 7. Similar results are observed for other oxide thicknesses. This channel-length dependent degradation can partly be explained by the increase of the ratio of damaged interface area to the total channel area as the channel length decreases. A useful variation of Fig. 7 which provides direct device design guidelines is shown in Fig. 8, where the maximum supply voltage to ensure a 10-year device lifetime (corresponding to a 3% change in drain current [9]) is plotted against channel length. This maximum supply voltage appears to decrease as the channel length decreases. For a quarter-micron device, the maximum supply voltage is 2.5 V suggesting that some kinds of hot-electron-resistant structures are still desirable even if the power supply is lowered to 3.3 V.

CONCLUSION

Performance and hot-electron reliability of deep-submicron MOSFET's are characterized for various oxide thicknesses and substrate dopings. It is found that the basic physics involved in deep-submicron devices is similar to that of micron-size devices. Existing physical models are applicable to these devices with the need for only minor modifications. The device degradation results suggest that the maximum supply voltage for a conventional quarter-micron device with 86 Å oxide thickness should be limited to 2.5 V based on a 10-year lifetime extrapolation.

REFERENCES

[1] T.Y. Chan, et al, IEDM Tech. Dig., this issue.

[2] W. Fichtner, et al, IEDM Tech. Dig., p.272, 1982.

[3] W. Fichtner, et al, IEDM Tech. Dig., p.384, 1983.

31.1

[4] S. Horiguchi, et al, IEDM Tech. Dig., p.761, 1985.

[5] C.G. Sodini, et al, IEEE Trans. Electron Devices, ED-31, p.1386, 1984.

[6] M.-S. Liang, et al, IEEE Trans. Electron Devices, ED-33, p.409, 1986.

[7] G. Baccarani, el al, IEEE Trans. Electron Devices, ED-31, p.452, 1984.

[8] T.Y. Chan, et al, IEEE Elec. Dev. Lett., EDL-5, pp.505, 1984.

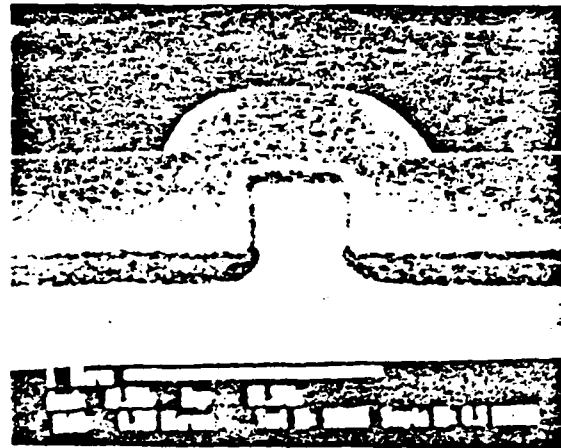[9] C. Hu, et al, IEEE Trans. Electron Devices, ED-32, pp.375, 1985.

Fig. 1 SEM photograph of a transistor cross-section with L_{eff} = 0.22 μm. The junction depth is about 0.15 μm.
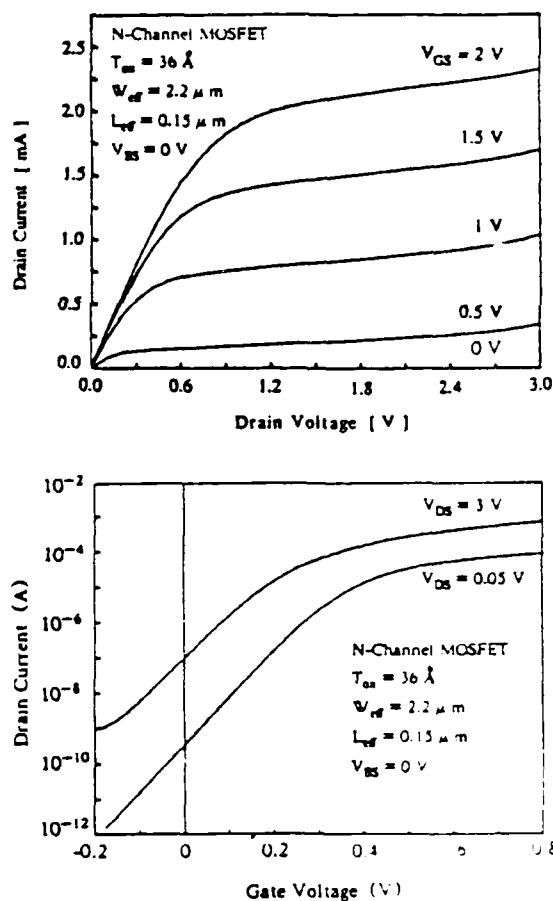
Fig. 2 Drain current characteristics of a deep-submicron MOSFET fabricated with the photoresist-thinning technique. (a) strong inversion characteristics, (b) subthreshold characteristics
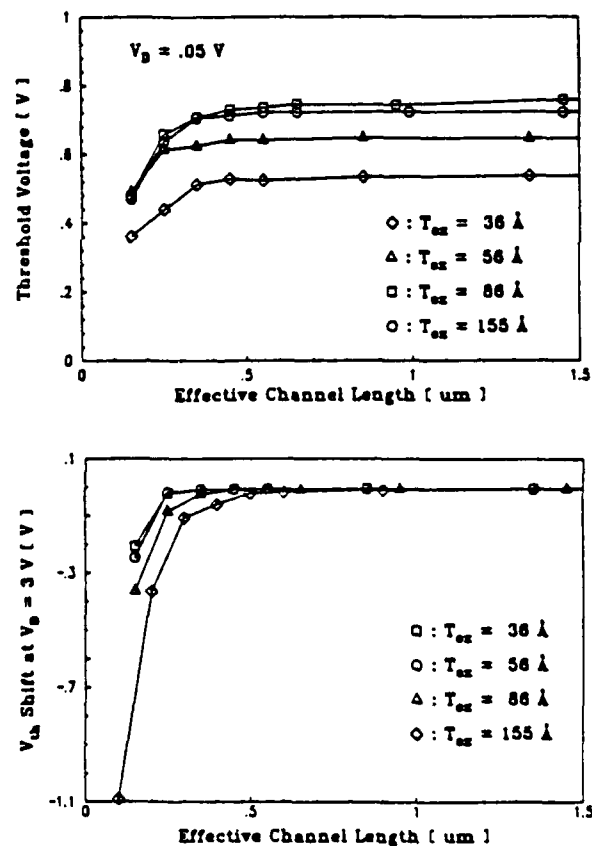
Fig. 3 (a) linear-region threshold voltage, (b) threshold voltage shift at V_{DS} = 3 V, as a function of effective channel length for various oxide thicknesses.

31.1
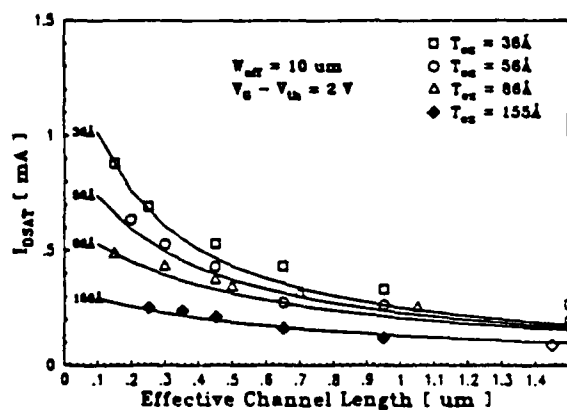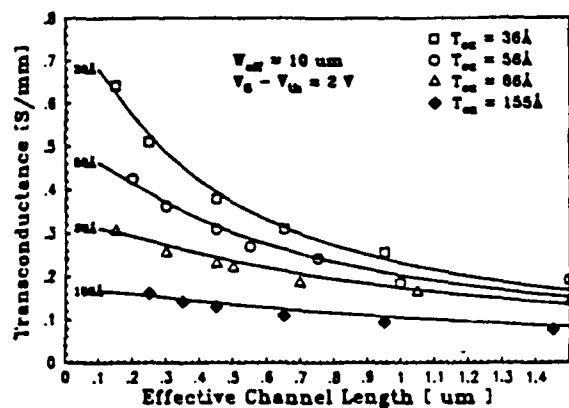
Fig. 4 Saturation characteristics as a function of effective channel length. Symbols represents measured data. Lines are calculated from a simple physical model. (a) transconductance, (b) drain saturation current.
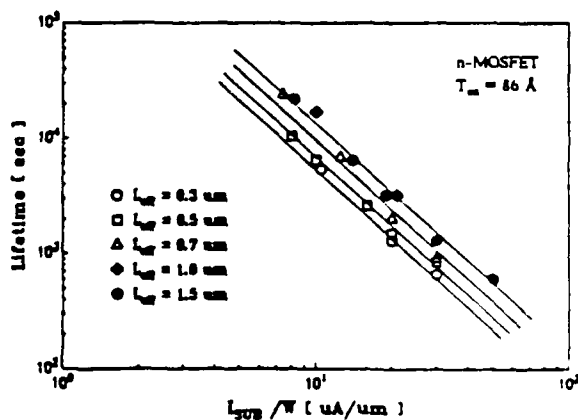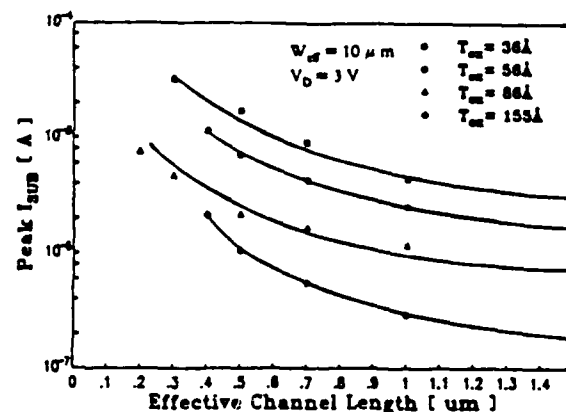

Fig. 5 Peak substrate current at $V_{DS} = 3$ V as a function of effective channel length for various oxide thicknesses.

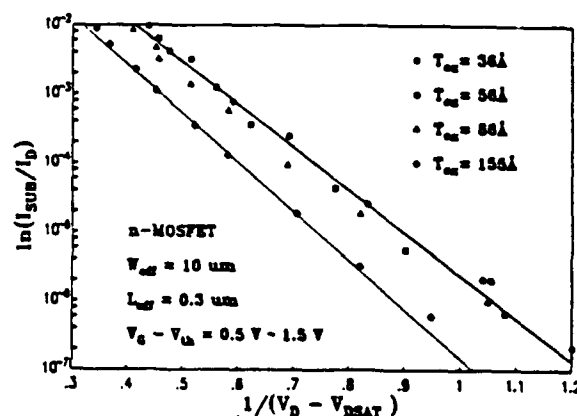Fig. 6 Normalized substrate current versus $1/(V_{DS} - V_{DSAT})$ for transistors with $L_{eff} = 0.3$ μm and various oxide thicknesses.
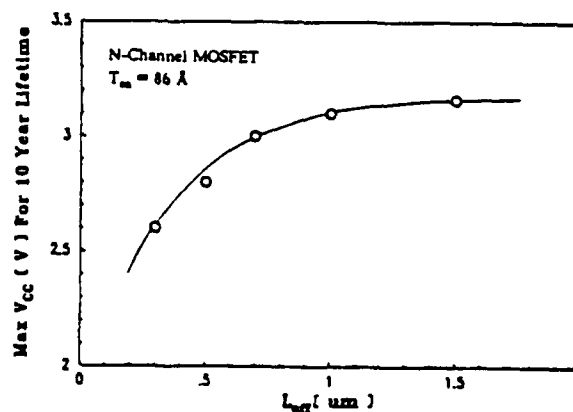

Fig. 7 Device lifetime versus substrate current per unit channel width for $T_{ox} = 86$ Å and various channel lengths. Lifetime is defined at the current degradation.

Fig. 8 Maximum operating voltage versus effective channel length for 86 Å devices based on a 10-year lifetime extrapolation.

31.1

Raman study of an epitaxial GaAs layer on a Si [100] substrate

Yihe Huang and Peter Y. Yu

Department of Physics, University of California and Materials and Chemical Sciences Division, Lawrence Berkeley Laboratory, Berkeley, California 94720

Marie-Noelle Charasse,[a] Yuhua Lo, and Shyh Wang

Department of Electrical Engineering and Computer Science, and Electronics Research Laboratory, University of California, Berkeley, California 94720

A bevel has been etched in a GaAs epitaxial film grown on a Si substrate, so that the Raman spectrum of the GaAs layers can be measured as a function of distance from the GaAs/Si interface. The amount of strain and disorder in the GaAs film has been estimated from the GaAs longitudinal optical phonon line shape and frequency. Both the strain and the amount of disorder were found to decrease with increase in the distance from the interface.

Recently there has been much interest in GaAs epitaxial films grown on Si substrates. This is due to the many potential applications which can result from combining the fast speed and excellent electro-optic properties of GaAs with the highly developed Si technology.[1-14] Two major problems have been encountered in the growth of high quality GaAs films on Si substrates. Firstly there is a 4.1% lattice mismatch between the two materials, and secondly the random occupation of the Ga and As sublattices leads to the presence of antiphase domains. The first effect results in high densities of dislocations and possibly also strain in the GaAs films. The antiphase domains contain large amounts of antisite defects at the domain boundaries. Raman scattering has been useful in the study of defects in GaAs caused by ion implantation, polishing, and high-energy neutron and electron irradiations.[5-10] We report here the first Raman study of epitaxial films of GaAs grown on Si. From this study we determined the crystallinity and strain in the GaAs film as a function of its distance from the interface.

Our sample was a 2.3-μm-thick GaAs film grown by molecular beam epitaxy (MBE) on a Si substrate oriented 4° off the [100] direction towards the [110] axis. To avoid the formation of islands, the first few hundred angstroms of GaAs was grown at $T < 350$ °C. This thin layer of GaAs was amorphous. After the Si surface was completely covered by GaAs, the substrate temperature was increased to 600 °C. Two layers of strained InGaAs-GaAs superlattices, each of five periods, are included in the GaAs film as shown in the inset of Fig. 1. The strained superlattices bend the dislocations and drastically reduce the density of the dislocations which can propagate through the superlattice. A small angle bevel was etched by gradually dipping the sample into $NH_4OH:H_2O_2:H_2O$ (1:1:20) to expose the GaAs layers near the interface as shown in the inset of Fig. 1. To maintain a uniform etching rate, the sample passes over de-ionized water before touching the etchant.

The Raman experiment was performed at room temperature in the $x(yz)\bar{x}$ backscattering geometry, where x,y,z refer to the [100], [010], and [001] crystallographic axes,

respectively. A SPEX double spectrometer and a multichannel detector, which are similar to those described in Ref. 11, were used to analyze the Raman signal. Typically 300 mW of Ar^+ laser power (5145 Å) was focused with a cylindrical lens to a spot of about 10×40 $(\mu m)^2$ on the sample with the longer side parallel to the edge of the bevel. The GaAs lattice temperature at the focus was estimated to be 360 K.

The Raman spectra were obtained successively by translating the sample in a direction perpendicular to the bevel edge with a micrometer. From the known bevel angle and distance of laser focus to the bevel edge, the thickness of the GaAs film probed by the laser can be calculated. The accuracy of this thickness measurement is about 300 Å. When the focus was scanned towards the bevel edge from the GaAs side, the GaAs Raman signal decreased while the Si optical phonon (at 519.5 cm^{-1}) intensity increased (see Fig. 1). Since the laser beam has to penetrate the GaAs film to reach the Si substrate, we can determine the absorption coefficient of the GaAs film from the variation of the Si phonon intensity provided that this absorption coefficient is not varying
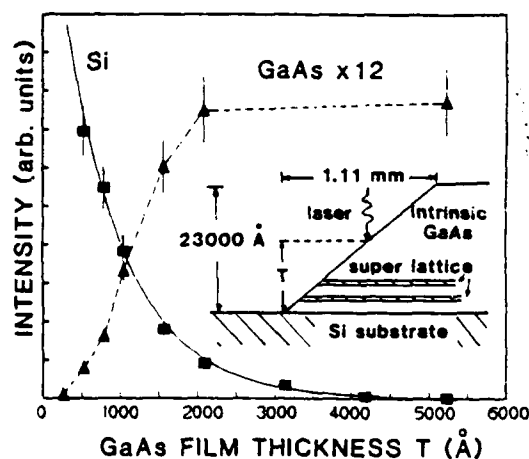

FIG. 1. Variation of the Si optical phonon and GaAs LO phonon intensities with the thickness of the GaAs film. The solid curve is a fit to the Si phonon intensity calculated by assuming that the GaAs film has an absorption length of 1580 Å. The broken curve guides the eyes. The GaAs LO phonon intensities have been multiplied by a factor of 12 in order to be plotted on the same scale as the Si phonons. The inset shows the sample geometry.

[a] Present address: Thomson–CSF, Laboratoire de Recherches. Domaine de Corbeville, Orsay, France.

with the GaAs film thickness. The solid curve in Fig. 1 shows a fit to the experimental points with a constant GaAs penetration depth of 1580 ± 20 Å. This value is consistent with that of bulk GaAs.[12]

The GaAs longitudinal optical (LO) phonon intensity in Fig. 1 was actually the sum of the Raman scattering from GaAs layers at different depths from the interface up to the surface. To determine the Raman spectra of the GaAs layers as a function of its distance to the interface, we have adopted the following procedure. First the Raman spectrum of GaAs was measured with the laser focus at a spot as close as possible to the edge of the bevel. This is taken to be the Raman spectrum of the GaAs layer at the interface. The laser focus is then translated by a few hundreds of angstroms away from the interface. The resultant Raman spectrum is then the sum of the Raman spectrum from the GaAs nearest to the interface plus that from an additional overlayer. By subtracting one Raman spectrum from the other, we obtain the Raman spectrum of any additional layer. The attenuation of light due to absorption in the GaAs film has to be corrected for in this subtraction process. This process can be continued until the spot reaches the surface of the GaAs film. Figure 2 shows the normalized GaAs LO phonon line shape of GaAs layers at different distances to the interface obtained in this way. When the layers are far away from the interface, their LO phonon spectra are narrow, symmetric, and Lorentzian in shape. After correction for the instrumental response, the intrinsic GaAs LO phonon linewidth is about 2 cm^{-1} for layers farther than half-micron from the interface. The GaAs LO phonon peaks become broader and more asymmetric when the layers are closer to the interface. Figure 3(a) shows the asymmetry and broadening of the GaAs LO phonon as a function of the layer distance to the interface.

In a perfect crystal only the zone center ($q = 0$) phonon modes are observed in one-phonon Raman scattering due to quasi-momentum conservation. The phonon line shapes are

FIG. 3. (a) Linewidth, frequency shift, and asymmetry of the GaAs LO phonon as a function of distance to the interface. These results have been obtained from Raman spectra of the GaAs LO phonon measured with better resolution than those shown in Fig. 2. The inset shows quantitatively the definition of the parameters Γ_A and Γ_B and the asymmetry parameter: Γ_A/Γ_B. (b) Strain and spatial correlation lengths of the GaAs film as a function of distance to the interface deduced from (a).
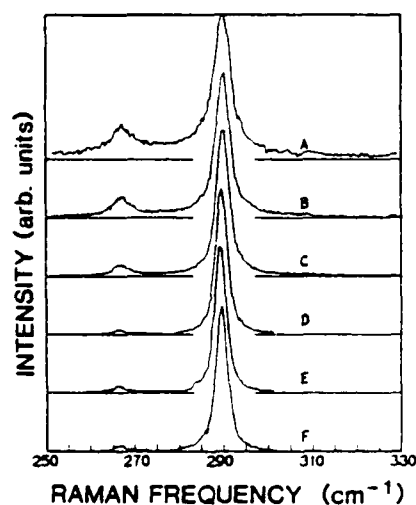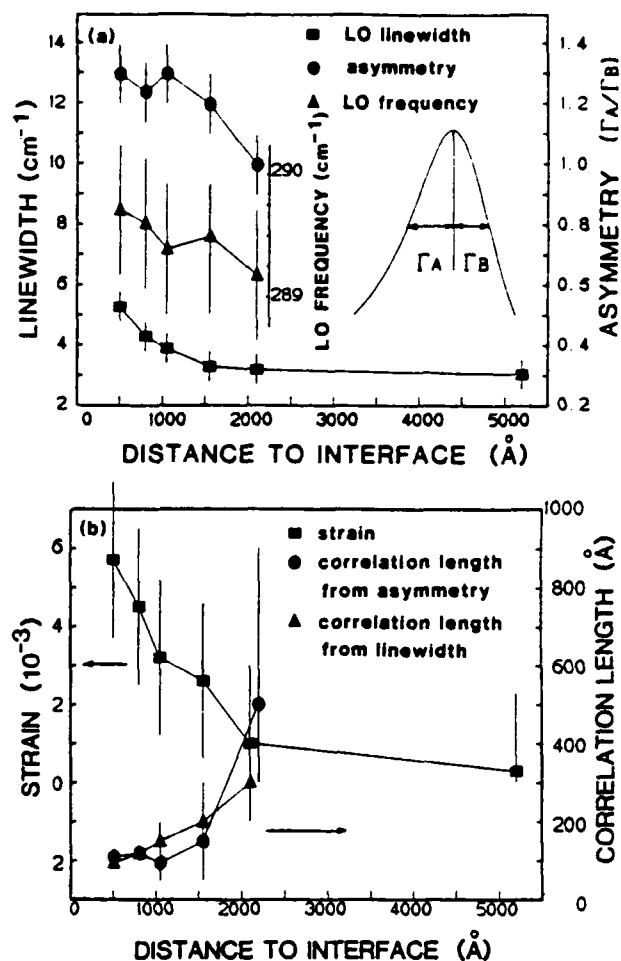

FIG. 2. Raman spectra of GaAs layers at different distances to the interface obtained by taking the difference of two Raman spectra measured from two adjacent spots along the bevel. A: 0–500 Å; B: 500–800 Å. C: 800–1000 Å; D: 1000–1500 Å; E: 1500–2100 Å; F: 2100–5200 Å.

usually Lorentzian with the phonon linewidths determined by the phonon lifetime. In a disordered crystal, however phonons can be localized to regions of extent L. As a result quasi-momentum conservation is relaxed,[7–9] and the LO phonon Raman peak becomes asymmetric and shifted towards lower frequency. On the other hand, compressive stress can shift the LO phonon towards higher frequency.[8,12,13,15] Thus the LO phonon line shape and peak positions contain information on both the crystal quality and on the amount of strain in the GaAs film. We adopt this spatial correlation model to interpret our experimental results.[7,9] In this model, L can be deduced from one of these quantities: shift in LO phonon frequency, LO phonon width, or the asymmetry of LO phonon line shape. Figure 3(b) shows the correlation lengths of the GaAs film as a function of distance to the interface deduced from the linewidth and the asymmetry. Both show the same increase in the correlation length as the GaAs film becomes farther from the GaAs/Si interface.

We found that the GaAs LO phonon frequency is blue

shifted for GaAs layers nearer to the interface in contrast to the prediction of the spatial correlation theory. We attribute this difference to the presence of compressive strain in the GaAs layers near the interface due to the lattice mismatch between GaAs and Si. To determine the magnitude of this strain we assume a uniform and isotropic two-dimensional strain for simplicity. We also assume that the GaAs film free surface is unstrained and use its LO phonon frequency as reference ω_0. From the phenomenological theory in Refs. 8,13–15 the strain-induced frequency shift $\Delta\omega$ is given by $\Delta\omega/\omega_0 = -1.3X$, where X is the strain. Figure 3(b) shows the variation in the compressive strain as a function of distance from the interface. The strain measured in our sample is much smaller than the 4% strain expected from the lattice mismatch. This suggests that the strain in GaAs films on Si is relieved by dislocations. For GaAs layers more than ~2000 Å away from the interface, the strain is below our measurement sensitivity.

A very small TO phonon Raman peak was also observed in our sample. Since the TO phonon is forbidden by a selection rule in our scattering geometry, we have interpreted this peak as due to disorder-induced scattering. The intensity of the TO phonon relative to the LO phonon increases towards the interface. This is consistent with our result that disorder increases for GaAs layers closer to the interface.

In conclusion, we have used the Raman scattering from a bevel etched on the GaAs film grown on a Si substrate by MBE to examine the GaAs quality as a function of its distance to the interface. The lattice spatial correlation and strain were deduced from the LO phonon line shapes and frequency shifts. We found that both the strain and the disorder in the GaAs film decreased in the first 2000 Å to a level below the sensitivity of the Raman measurement.

[1]R. Fischer, N. Chand, W. Kopp, C. K. Peng, and H. Morkoç, IEEE Trans. Electron Devices ED-33, 206 (1986).

[2]R. Fischer, N. Chand, W. Koop, H. Morkoç, L. P. Erickson, and R. Youngman, Appl. Phys. Lett. 47, 397 (1985).

[3]J. P. van der Ziel, R. D. Dupuis, and J. C. Bean, Appl. Phys. Lett. 48, 1713 (1986).

[4]R. Fischer, H. Morkoç, D. A. Neumann, H. Zabel, C. Choi, N. Otsuka, M. Longerbone, and L. P. Erickson, J. Appl. Phys. 60, 1640 (1986).

[5]D. J. Evans and S. Ushioda, Phys. Rev. B 9, 1638 (1974).

[6]C. S. Rama Rao and S. Sundaram, J. Appl. Phys. 54, 1808 (1983).

[7]H. Richter, Z. P. Wang, and L. Ley, Solid State Commun. 39, 625 (1981).

[8]H. Shen and F. H. Pollak, Appl. Phys. Lett. 45, 692 (1984).

[9]K. K. Tiong, P. M. Amirtharaj, F. H. Pollak, and D. E. Aspnes, Appl. Phys. Lett. 44, 122 (1984).

[10]R. S. Berg, P. Y. Yu, and E. R. Weber, Appl. Phys. Lett. 47, 515 (1985).

[11]J. C. Tsang, G. S. Oehrlein, I. Haller, and J. S. Custer, Appl. Phys. Lett. 46, 589 (1985).

[12]B. O. Seraphin and H. E. Bennett, "Optical Properties of III-V Compounds," in *Semiconductors and Semimetals*, edited by R. K. Willardson and A. C. Beer (Academic, New York, 1967), Vol. 3, p. 513.

[13]B. A. Weinstein and M. Cardona, Phys. Rev. B 5, 3120 (1972).

[14]R. Trommer, H. Muller, and M. Cardona, Phys. Rev. B. 21, 4869 (1980).

[15]F. Cerdeira, C. J. Buchenauer, F. H. Polak, and M. Cardona, Phys. Rev. B 5, 580 (1972).

194   Appl. Phys. Lett., Vol 51, No. 3, 20 July 1987

Huang et al.   194

RAMAN SCATTERING SPECTRUM ALONG A BEVEL ETCHED GaAs ON Si, TEM STUDY AND GaAs P-I-N PHOTODETECTOR ON Si

Y.H. Lo, M.-N. Charasse, H. Lee, D. Vakhshoori, Y. Huang and Peter Yu
University of California, Berkeley, CA
Z. Liliental-Weber and M. Werner
Lawrence Berkeley Laboratory, Berkeley, CA
and S. Wang
University of California at Berkeley and Lawrence Berkeley Laboratory

Abstract

Raman scattering is measured along a bevel etched GaAs epitaxial film grown on Si by molecular beam epitaxial (MBE). From the correlation length profile of Raman scattering, most dislocation lines are confined in the 2000Å regions close to the interface. The strain profile calculated from the Raman peak shift shows that about 0.6% compressive strain exists near the interface because of lattice mismatch. However, as one moves away from the interface, the compressive strain is gradually counterbalanced by thermal expansion. Transmission electron microscope (TEM) studies of the local dislocation image and properties show that an ultra clean Si surface is essential for dislocation confinement. From high resolution TEM, we find that the distance between dislocations at the interface is nonuniform, varying from 50Å to 125Å with an average distance at 81Å. Finally, a GaAs p-i-n photodetector on Si substrate is fabricated. Even though a normal photoresponse curve is obtained, the high dark current (50nA) and relatively low responsivity (0.01A/W) show that the material quality needs to be further improved to make a minority carrier vertical transition device.

Introduction

Lately the research on epitaxial growth of GaAs on Si substrate is developing quickly because it not only invokes a number of interesting scientific challenges but also provides practical benefits in applications such as GaAs VLSI digital circuit and opto-electronic integration.[1-5] However, several fundamental problems exist in the epitaxial growth of GaAs on Si. First, the growth of a polar crystal on a non-polar crystal tends to form anti-phase domains (APD) which induce static charge sheets in the GaAs thin film. Second, the relatively large lattice mismatch (≈4%) between GaAs and Si will induce many dislocations near the interface and some of them may propagate towards the crystal surface. Third, the large difference in thermal expansion coefficients between GaAs and Si causes a very high stress which often makes the sample crack during wet etch processing.

To investigate the strain and dislocation properties as mentioned above, we perform a Raman scattering experiment which can measure the strain and dislocation density profile from the GaAs and Si interface to the GaAs surface. On the other hand, cross section TEM and high resolution TEM provide us with information about dislocation image and properties. A GaAs on Si p-i-n photodetector is also fabricated for the first time and the results are compared with a normal GaAs p-i-n photodetector.

Raman Scattering

GaAs is grown on a Si substrate oriented 4 degrees off the [100] direction towards the [110] axis by MBE. As shown in the insert of Fig-1, a bevel of 0.1 degree has been etched in the GaAs epitaxial film, so that the Raman spectrum of GaAs layers can be measured as a function of distance from the GaAs/Si interface. The 5145Å line from an Ar^+ laser is used to trigger the Raman signal. The Raman experiment is performed at room temperature in the $x(yz)\bar{x}$ back-scattering geometry, where x,y,z refer to the [100], [010] and [001] crystallographic axes respectively. In this scattering geometry, the longitudinal optical (LO) phonon of GaAs is allowed by the selection rules, while the transverse optical (TO) phonon is forbidden. The LO phonon Raman scattering intensity variation with distance from the interface is shown in Fig-1. We can determine the transmission coefficient of the GaAs film form the variation of the Si phonon intensity provided that this absorption coefficient is not varying with the GaAs film thickness. The solid curve in Fig-1 shows a fit to the experimental points with a constant GaAs penetration depth of 1580±20Å.

The normalized Raman spectrum at different depths from the interface is shown in Fig-2. The GaAs LO phonon peaks become broader and more asymmetric when the layers are closer to the interface. This phenomenon can be explained as follows.

In a perfect crystal, only zone center (q=0) phonon modes are excited for one-phonon Raman scattering because of quasi-momentum conservation. For a disordered lattice, however, phonons can be localized to a region of space L, and as a result quasi-momentum conservation will be relaxed so that off-center phonons (q≠0) can be excited.[6-8] The participation of off-center phonon scattering causes the asymmetric non-Lorentzian lineshape near the interface. On the other hand, the phonon linewidth near the interface is broader than the phonon linewidth for layers farther than half micron from the interface because the phonon lifetime is reduced by disordered material. Both the LO phonon lineshape and linewidth, therefore, contain information about the crystal quality. We adopt the spatial correlation model developed by Richter[6] and Tiong et al.,[8] to calculate the spatial correlation length L from either the phonon width or the lineshape asymmetry. Fig-3 shows the correlation lengths of the GaAs film as a function of distance from the interface. Those calculated correlation lengths are related to the average distance between two dislocation lines. The abrupt increase of correlation length around 2000Å from the interface shows that most dislocations can not propagate further.

In addition, the strain profile can also be calculated from the Raman peak frequency shift if a uniform and isotropic two-dimensional strain is assumed.[7] As shown in Fig-3, a compressive strain is present in the GaAs layers near the interface is present. This may be due to the lattice mismatch between GaAs and Si. This compressive strain from lattice mismatch is counterbalanced by the tensile strain from thermal expansion, so the GaAs surface becomes almost strain free for a 2.3μm film.

For the region thicker than 5000Å, the dislocation density is too low to be detected by Raman scattering techniques. The discrete feature and local image of dislocation lines will be studied by TEM.

TEM Study

Figs-4 and 5 show the TEM cross sections of two GaAs on Si samples grown by MBE. Fig-4 shows the growth of GaAs on an oxygen and carbon contaminated Si surface. Very characteristic triangular protrusions are observed near the Si surface giving rise to a high dislocation density. Many dislocations and stacking faults generated at the interface propagate to the surface layer of GaAs. The density of dislocations estimated in the near surface area is in the range of 5 to $9 \times 10^9 cm^{-2}$. If the Si surface is clean (Fig-5), triangular features are not formed at the interface and the density of stacking faults formed at the interface is lower as well. Many of the misfit dislocations are tangled and confined in the near Si/GaAs area within 0.3μm from the Si substrate. This film quality is quite acceptable even without any strain layer superlattice or thermal annealing. This observation indicates that an ultra clean Si substrate is necessary for high quality GaAs on Si growth.

As was found before,[5] two types of dislocations are formed at the interface. Most of the dislocations (≈85%) have their Burgers vector parallel to the interface (Fig-6) and only about 15% have their Burgers vector inclined to the interface. This small fraction of dislocations with inclined Burgers vector is probably the reason for the much lower density of dislocations ≈5×10^7 cm^{-2}) in the near surface area of the GaAs. The average distance between the dislocations at the interface is one dislocation per 81Å. However, the distribution of the dislocations at the interface in not uniform. There are areas where the distance between the dislocations (estimated from high resolution TEM pictures) is only 50Å apart but there are other areas where dislocations are separated as much as 125Å. Higher stress (≈2.5%) is built up in the areas where dislocations are closer to each other. Stress is almost completely released (≈0.1%) in the areas where dislocations are far from each other. One would expect that dislocations which are close together interact differently from dislocations which are further apart. This nonuniformity will influence the final dislocation distribution in the top surface layer of GaAs, where one can find both areas which are dislocation free and areas which have higher dislocation densities.

P-I-N Photodetector

In terms of device characteristics, most majority carrier devices like FETs have been reported by several groups[1,2] showing comparable performance with those devices made on a GaAs substrate. LEDs and lasers have also been recently reported even though reliability and lifetimes are much worse than those on a GaAs substrate.[9,10] Photodetectors, however, are almost not reported yet besides the lateral conductive photoconductors which is most insensitive to dislocations and other defects. We fabricated a p-i-n homojunction GaAs photodiode having the device structure shown in Fig-7. The commonly used two-step growth were adopted and three InGaAs/GaAs strain layer superlattices, ten 50Å/100Å periods each superlattice, are grown as dislocation filtering layers. The high dark current (50nA at -0.2volt) restricts this device from working at higher bias. The normalized photoresponse with respect to wavelength is shown in Fig-8. The peak responsivity at room temperature occurs at 0.78μm (1.59ev) and cuts off at 0.88μm (1.41ev). The lack of abnormal peaks at longer wavelength (up to 1.5μm) means that there is no significant deep level absorption. High dark current can be explained by higher dislocation density (approximately 10^3 to 10^4 times) in the device active region than in the GaAs substrate. The responsivity without an anti-reflection coating is about 0.01A/W, which is also relatively low.

152

Conclusion

In summary, we use Raman scattering to scan across a beveled edge of GaAs on Si to find out the strain as well as the average dislocation density profiles. Compressive strain is present at the interface because of the GaAs and Si lattice mismatch but this strain is gradually counterbalanced by the tension from thermal expansion. The dislocation image and its propagation are studied by TEM, which indicates that an extremely clean Si surface is essential. The distance between two dislocations near the interface is nonuniform. Finally, the first GaAs on Si p-i-n photodetector is demonstrated. Despite its high dark current and relatively low responsivity, a normal photoresponse curve without any deep level peak is observed.

Reference

[1] H. K. Choi, B.-Y. Tsaur, G. M. Metze, G. W. Turner, and J. C. C. Fan, IEEE Electron Device Lett. EDL-5, 207 (1984).

[2] R. Fischer, N. Chand, W. Kopp, C. K. Peng, and H. Morkoc. IEEE Trans. Electron Devices, ED-33, 206 (1986).

[3] R. Fischer, N. Chand, W. Kopp, and H. Morkoc, Appl. Phys. Lett. 47, 397 (1985).

[4] J. P. van der Ziel, R. D. Dupuis, and J. C. Bean, Appl. Phys. Lett. 48, 1713 (1986).

[5] R. Fischer, H. Morkoc, D. A. Neumann, H. Zabel, C. Choi, N. Otsuka, M. Loqngerbo, and L. P. Erickson, J. Appl. Phys. 60, 1640 (1986).

[6] H. Richter, Z. P. Wang, and L.Ley, Solid State Commun. 39, 625 (1981).

[7] H. Shen and F. H. Pollak, Appl. Phys. Lett. 45, 692 (1984).

[8] K. K. Tiong, P. M. Amirtharaj, F. H. Pollak, and D. E. Aspesns, Appl. Phys. Lett. 44, 122 (1984).

[9] T. H. Windhorn, G. M. Metze, B. Y. Tsaur, and J. C. C. Fan, Appl. Phys. Lett. 45, 309 (1984).

[10] J. P. van der Ziel, R. D. Dupuis, R. A. Logan, R. M. Mikulyak, C. J. Pinzone, and A. Savage, Appl. Phys. Lett. 50, 454 (1987).
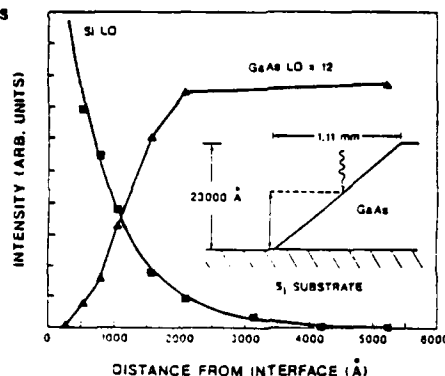
Figures

Figure 1. The LO phonon Raman scattering intensity profile. The epitaxial structure of the sample with bevel angle 0.1 degree is also shown.
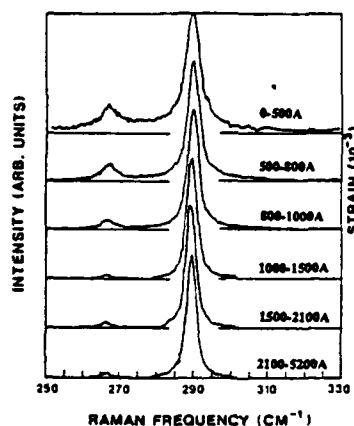
Figure 2. Normalized Raman scattering at different distances from the interface. Both LO phonon and TO phonon peaks can be identified.
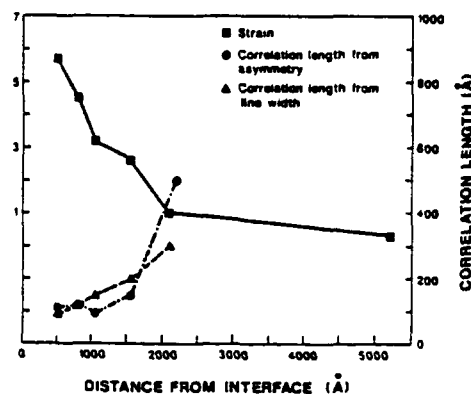

Figure 3. Calculated strain and correlation length profiles from Raman scattering. The correlation length obtained from either Raman linewidth or lineshape asymmetry corresponds to the average dislocation distance.
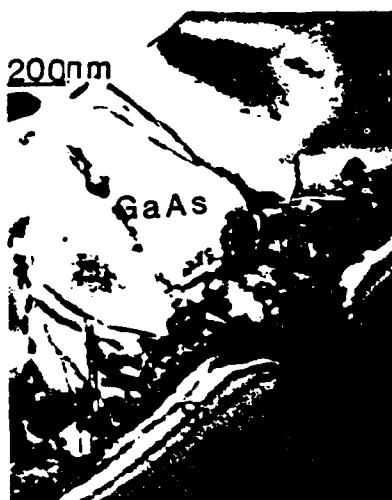

Figure 4. TEM picture of GaAs grown on an oxygen and carbon contaminated Si surface. A number of dislocations and stacking faults are observed and many of them propagate to the GaAs surface layer.
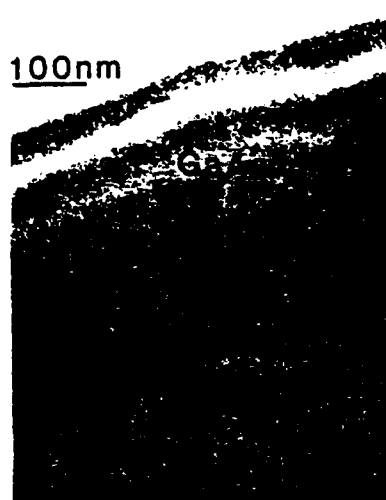

Figure 5. A TEM picture of GaAs grown on a clean Si surface. In contrast to Fig-4, most of the dislocations are confined near the interface or are bent in the middle so that the GaAs surface has very low dislocation density.
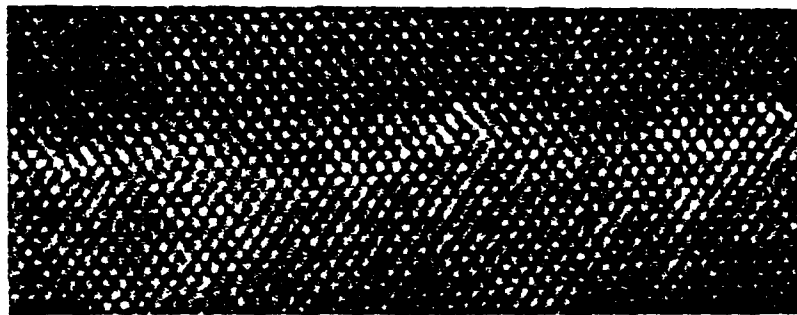
154

Figure 6. High resolution TEM picture of GaAs and Si interface. In this picture, We look at (110) plane where dislocations at the interface are found on the (100) plane with Burgers vector parallel to the interface. All of these dislocations in this picture are not inclined.
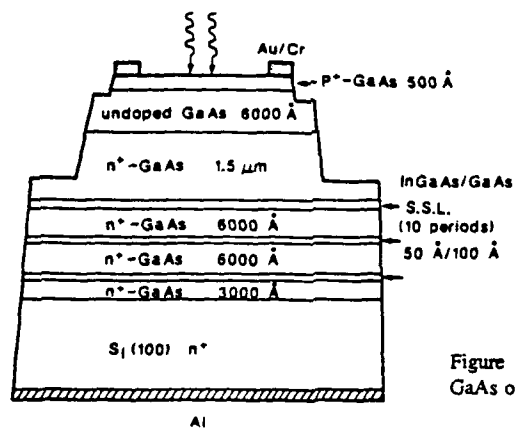

Figure 7. Schematic diagram of the GaAs on Si p-i-n photodetector.
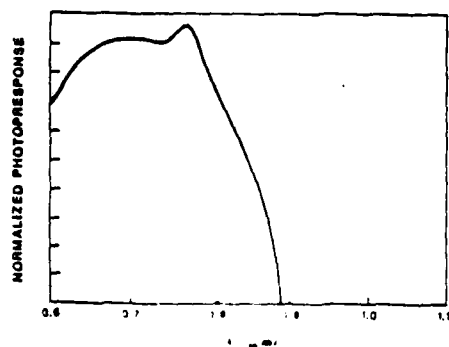

Figure 8. The normalized photoresponse curve of the GaAs on Si p-i-n diode.

Photoluminescence studies of selective-area molecular beam epitaxy of GaAs film on Si substrate

Henry P. Lee and Shyh Wang[a]

Department of Electrical Engineering and Computer Sciences and Electronics Research Laboratory,
University of California, Berkeley, California 94720

Yi-He Huang and Peter Yu

Department of Physics and Materials and Chemical Sciences Division, Lawrence Berkeley Laboratory,
University of California, Berkeley, California 94720

GaAs films have been grown on Si substrates patterned with SiN by molecular beam epitaxy. The pattern consists of bare Si stripe of width ranging from 10 to 100 μm surrounded by SiN on both sides and a reference area of bare Si. 77 K photoluminescence (PL) spectrum and intensity are measured on the single crystalline GaAs films grown on these stripes and the reference area. For 1.5- and 3-μm-thick films, PL intensity from the 10 μm stripe shows 140% and 75% increase over the reference area, respectively. This remarkable increase in the PL intensity is believed due to the reduction of dislocations inside the window area. The improvement in the optical quality makes selective-area molecular beam epitaxy a very attractive technique for the fabrication of optical devices on Si substrate.

Heteroepitaxy of GaAs thin films on Si substrate offers a unique opportunity for monolithic integration of GaAs based optoelectronics and high-speed devices with Si integrated circuits (IC's). Since 1984, GaAs on Si metal-semiconductor field-effect transistors IC's,[1] high electron mobility transistors, heterojunction bipolar transistors, light-emitting diodes and lasers, have been successfully demonstrated by various groups[2] using molecular beam epitaxy (MBE) and metalorganic chemical deposition (MOCVD). Despite these impressive achievements, epitaxial GaAs films on Si substrates have one inherent problem independent of the epitaxy techniques used; namely, the high threading dislocation density (10^7–10^8/cm^2) arising from the 4% lattice mismatch between the GaAs and Si lattice constants. High dislocation density and large tensile stress (low 10^9 dynes/cm^2) are two main obstacles to be overcome before GaAs on Si technology can realize its full potentials.

To reduce dislocation density away from the heterointerface, strained-layer superlattices (SLS's) such as InGaAs/GaAs[3] and GaP/GaAs/GaPAs[4] have been incorporated by many groups during the epitaxial growth to bend upward propagating dislocations. In this letter, we study the possibility of laterally terminating dislocations through the use of selective-area molecular beam epitaxy (SAMBE)[5] with SiN as the on-wafer insulating mask. In SAMBE growth, single crystalline epitaxial films are formed only on the exposed area of crystalline substrate. Areas covered by noncrystalline films (such as SiO, SiN) result in the growth of polycrystalline (poly) films. Since poly films contain many grain boundaries, dislocations entering them from the single crystalline areas are effectively absorbed by the poly film. We report in this letter successful SAMBE growth of GaAs films on Si substrates and their 77 K PL characteristics. We found significant increase in PL intensity for both

1.5- and 3-μm-thick films grown by SAMBE over films grown by planar MBE. This finding provides strong evidence that crystalline/poly interface is responsible for the reductions of dislocation density.

Heavily doped n^+-Si (001) substrate tilted 3.5° towards ⟨110⟩ axis is used in this work. The wafer is first cleaned repeatedly in 5:1 H_2SO_4:H_2O_2 before depositing 800 Å of SiN by thermal CVD. The wafer is patterned and etched to form 2-mm-long stripes of exposed Si with widths ranging from 10 to 100 μm along ⟨$\bar{1}$10⟩ orientations. Three identical stripes separated by 20 μm of SiN are grouped together as shown in Fig. 1. Next to the stripe pattern is a 4 mm by 4 mm area of bare Si. GaAs film grown over this area is used as a reference for comparison since its properties are identical to GaAs film grown over a patternless substrate. The etching of SiN is done by reactive ion etching (RIE) using a mixture of CHF$_3$ and O$_2$ and a 20% of overetch time. Etching of Si substrate is estimated to be 100 Å. After RIE, the wafer is thoroughly cleaned and processed by the usual pre-MBE preparation procedures[6] before loading into the MBE chambers. A two-step MBE growth procedure is used where the initial 2000 Å layer grown at a growth rate of 0.6 μm/h and substrate temperature of 400 °C is followed by a normal growth rate of 1 μm/h at substrate temperature of 570 °C. Two undoped samples having a thickness of 1.5 and 3 μm are grown in this way, respectively. The films grown inside the stripe windows show identical color and texture compared with the film grown over the patternless area when viewed under the microscope. Significant lateral growth in ⟨$\bar{1}$10⟩ direction is observed as in the case of SAMBE on GaAs substrate.[7]

The two samples are mounted next to each other in a liquid-nitrogen Dewar sitting on a movable X-Y stage. A 1-μm-thick unintentionally doped GaAs film is also included as a standard sample for comparison. The 488 nm line of an Ar$^+$ laser with a power of 150 mW focused to 1.5 mm by 0.76 mm spot size is used as the excitation source in the

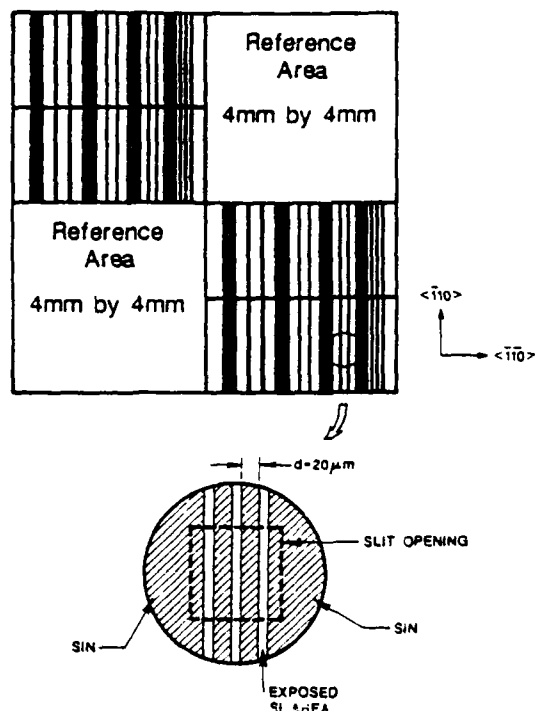[a] Also Center for Advanced Material, Lawrence Berkeley Laboratory.

FIG. 1. Layout of the stripe windows (not in full detail). The darkened region is area covered with SiN film. The rectangular box formed by the broken line represents the slit opening where only luminescence inside the box enters the spectrometer.

FIG. 2. PL spectra of standard GaAs sample and GaAs on Si samples inside a 4 mm by 4 mm reference area. The calculated PL spectra (shown in solid line) using a biaxial stress model[10] are superimposed on the measured data. Excellent agreement between the measured and calculated spectra is achieved here.

photoluminescence (PL) measurement. Relative low power density (28 W/cm^2) is used here to minimize the temperature variations of GaAs films inside different window sizes. The detector consists of a SPEX double grating spectrometer and a scanning photocounting GaAs detector cooled to −20 °C. The spectrometer is scanned from 12 350 cm^{-1} (8097 Å) to 11 750 cm^{-1} (8510 Å) during the measurement. The scanning rate of the spectrometer is set at 2 cm^{-1}/s with a spectral resolution of 4 cm^{-1}. The luminescence signal from the sample is first imaged onto an adjustable slit before entering the spectrometer. By adjusting the width of the slit, only luminescence signal from areas under study is detected. The large size of the laser spot covers illumination over a large area of the sample and makes accurate comparison meaningful. Before each measurement the wafer position is carefully adjusted to maximize detector counts at a wavelength near the PL peak. This ensures that all the emission from a particular area of the wafer under study is collected.

The PL spectra of a standard GaAs film, a GaAs on Si film inside 10 μm and 100 μm stripes, and in the reference area are shown in Fig. 2 for the 1.5- and 3-μm-thick samples, respectively. All GaAs on Si films show two clearly resolved PL peaks (A and B) shifted towards longer wavelength compared to standard GaAs sample. This has been observed by many groups[4,8,9] and is generally attributed to the splitting of valence bands by the biaxial tensile stress inside the film. The calculated spectra taking into account of a uniform biaxial tensile stress, exciton transitions, and temperatures of excited carriers are also plotted as solid lines in Fig. 2.[10] The
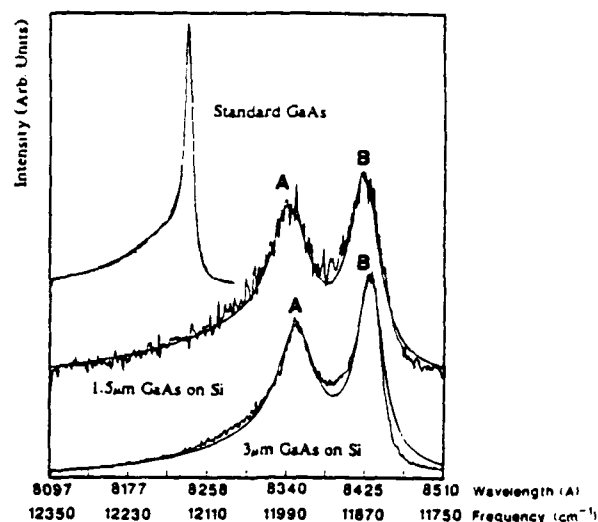
agreement between the calculated and measured spectrum is excellent. From the theoretical calculation, the biaxial tensile stress not only splits the $J = 3/2$ and $J = 1/2$ valence bands but also causes an overall shift of the two splitted bands. The tensile stress calculated from the shift of peak A and B and their splitting agrees well. The tensile stress is found to be 3.2 ± 0.4 kbar for the 1.5-μm film and 3.7 ± 0.5 kbar for the 3-μm film respectively at 77 K. The PL peak wavelengths from the film inside the 10-μm width, the 100-μm width, and the reference area show no appreciable difference, indicating that the tensile stress of the films is not affected by the width of the window stripes in which the films are grown. The linewidth of the peak B is, however, broader for the 10-μm-wide film compared to the 100-μm-wide film probably due to some nonuniform distribution of stress inside the 10-μm film. The total PL intensity is calculated by first subtracting off the dark counts of the detector and then integrating PL counts over all frequency. PL intensity from various stripes is then normalized with respect to their window area, the slit size, and the PL intensity of the standard GaAs sample. The results are shown in Fig. 3 for both samples. PL intensity on the 4 mm by 4 mm reference area is taken as the PL intensity for infinite stripe width. The improvement in PL intensity for the film in the 10 μm stripe over the 100 μm stripe is over 140% for the 1.5-μm-thick sample and 75% for the 3-μm-thick sample. Indirect evidences from high-temperature annealing experiments[11,12] indicate that increase in PL intensity is related to the reduction of dislocation. Although the dislocation reduction mechanism is not yet clear at this stage, we believe that increase in PL intensity resulted from dislocation density reduction inside the selective-area grown film. It is expected that careful transmission electron microscopy (TEM) studies will reveal the mechanism of dislocation termination on the poly/crystalline boundary.
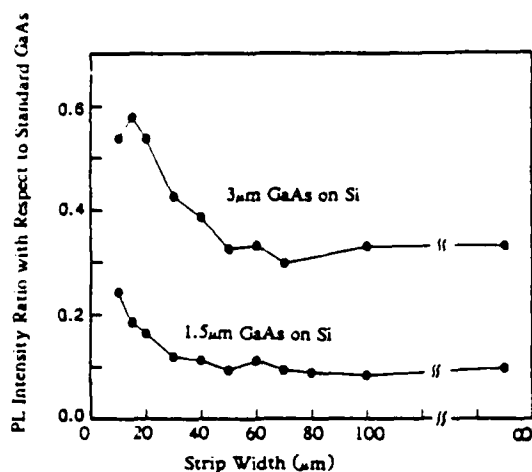
FIG. 3. PL intensity normalized with respect to the stripe area and PL intensity of standard GaAs sample for film of different stripe widths. The 4 mm by 4 mm reference area is taken as having infinite stripe width.

In summary, we have found significant increase in the PL intensity for GaAs films grown on Si substrates by SAMBE. We attribute the increase in PL intensity to the reduction of dislocation density. The PL spectra of GaAs on Si films can be fully explained by the presence of biaxial tensile stress in the film. It is also shown that the presence of poly-GaAs on either side of the strip does not change the stress of the crystalline GaAs film grown in between. The

remarkable improvement in the film quality makes SAMBE a very attractive technique for the fabrication of optical source and detector on Si substrates.

[1] H. Shichijo, J. W. Lee, W. V. Mclevige, and A. H. Taddiken, IEEE Electron Device Lett. EDL-8, 121 (1987).
[2] See, for example, Materials Research Society Symposia Proceedings, edited by J. C. C. Fan (Materials Research Society, Pittsburgh, 1986), Vol. 67.
[3] R. Fisher, D. Neumann, H. Zabel, H. Morkoç, C. Choi, and N. Otsuka, Appl. Phys. Lett. 48, 1223 (1986).
[4] T. Soga, S. Hattori, S. Sakai, and M. Umeno, J. Cryst. Growth 77, 498 (1986).
[5] A. Y. Cho and W. C. Ballamy, J. Appl. Phys. 46, 783 (1975).
[6] A. Ishizaka and Y. Shiraki, J. Electrochem. Soc. 133, 666 (1986).
[7] A. Y. Cho, J. Vac. Sci. Technol. 16, 275 (1979).
[8] W. T. Masselink, T. Henderson, J. Klem, R. Fisher, P. Rearah, H. Morkoç, M. Hafich, P. D. Wang, and G. Y. Robinson, Appl. Phys. Lett. 45, 1309 (1984).
[9] S. Zemon, S. K. Shastry, P. Norris, C. Jagannath, and G. Lambert, Solid State Commun. 58, 457 (1986).
[10] Y. H. Huang, P. Yu, H. P. Lee, and S. Wang (unpublished).
[11] N. Chand, R. People, F. A. Baiocchi, K. W. Wecht, and A. Y. Cho, Appl. Phys. Lett. 49, 815 (1986).
[12] J. W. Lee, H. Shichijo, H. L. Tsai, and R. J. Matyi, Appl. Phys. Lett. 50, 31 (1987).

Characterization of GaAs film grown on Si substrate by photoluminescence at 77 K

Yihe Huang and Peter Y. Yu

Department of Physics, University of California, and Materials and Chemical Sciences Division. Lawrence Berkeley Laboratory. University of California, Berkeley, California 94720

Henry Lee and Shyh Wang

Department of Electrical Engineering and Computer Science, and Electronics Research Laboratory. University of California. Berkeley, California 94720

Photoluminescence of GaAs films grown on Si substrate has been investigated quantitatively at 77 K. The peak shift and splitting of the exciton luminescence are shown to result from tensile stress in the film. Information on carrier lifetime has been deduced from the line shape of the photoluminescence.

GaAs film grown on Si substrate by molecular beam epitaxy has been of much interest recently.[1-4] Many different types of devices have been fabricated successfully on these thin films. But there are still serious problems remaining to be solved. One such problem is the high density of dislocations resulting from the lattice mismatch between GaAs and Si. This high dislocation density results in short minority-carrier lifetimes making it difficult to fabricate high-quality laser diodes. The second problem is thermal stress which results from difference in thermal expansion coefficients between GaAs and Si. Several techniques have been used to measure this thermal stress in GaAs films grown on Si. Measurement of the stress-induced curvature is possible for large wafers such as those grown by metalorganic chemical vapor deposition.[5] The lattice constant of the strained GaAs films can also be measured by x-ray diffraction.[6] But both techniques measure the magnitude of the stress averaged over large areas. The line shape and frequency shift of phonon Raman scattering have been shown to yield information on both the crystal quality and stress.[7] This technique suffers from low sensitivity. Determination of stress in GaAs film on Si using photoluminescence technique has been reported previously.[8] Zemon *et al.* found that the band-edge luminescence peak was shifted and split. They suggested that this shift and splitting were caused by strain and could be used to determine the thermal stress in the GaAs film.[9] In this letter we report a quantitative analysis of the photoluminescence (PL) of several GaAs on Si samples with different film thicknesses and widths. We show that from an analysis of the splitting of the PL peaks, their line shapes, and relative intensities, one can deduce besides the thermal stress other useful information on sample quality such as minority-carrier lifetime.

Our GaAs films were grown by molecular beam epitaxy on either Si or GaAs substrates. The film thicknesses are listed in Table I. Sample *D* is only 10 μ wide while other samples were more than 2 mm wide. The samples were unintentionally *n*-type doped to less than 10^{16} cm^{-3} The 5145-Å line of an Ar$^+$ laser was used to excite the PL. Since the line shape of the PL spectra changed at high excitation power densities, the laser power density was kept at a low level of 29 W/cm^2 which we will show to cause minimal heating of the samples. A Spex 1403 double monochromator and a photon counting system were used to measure the PL spectra. The experiment was performed at liquid-nitrogen temperature with all the samples mounted side by side on the same cold finger.[10]

Sample *A* was assumed to be free of thermal stress since its substrate was also GaAs. Indeed its PL spectra showed one sharp peak (see Fig. 1) with no detectable sign of splitting. The frequency of this peak was used as the reference for determining the stress-induced shift in the PL of other GaAs/Si samples. In contrast the PL spectra of all the GaAs/Si films showed two peaks similar to those reported by Zemon *et al.*[9] Although these spectra were qualitatively similar to each other, the widths of the PL were quite different between spectra *B*, *C*, and *D*.

To interpret these PL spectra quantitatively, we have used the following model. First we assume that each peak in the PL spectrum contains two contributions: transitions due to the lowest (1*s*) exciton level and the corresponding band-to-band transition. These two contributions were not fully resolved in our experiment because of the smallness of the exciton binding energy (E_{ex}) in GaAs and the high temperature ($>$ 77 K) of the experiment. In the absence of strain the heavy hole (hh) and light hole (lh) bands are degenerate at the Brillouin zone center so that these bands contribute only one luminescence peak. The split-off band is 0.34 eV higher

TABLE I. Summary of characterization of GaAs films and sample parameters deduced by fitting the experimental photoluminescence spectra.

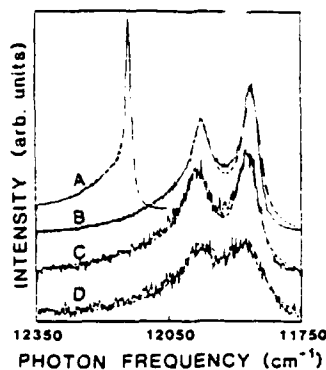| Sample	*A* GaAs/GaAs	*B* GaAs/Si	*C* GaAs/Si	*D* GaAs/Si
Thickness (μ)	1.5	3	1.5	1.5 (10 μ wide)
Exciton linewidth (cm^{-1})				
Δ_h (heavy hole)		44	53	85
Δ_l (light hole)	13	52	70	110
Carrier temperature (K)	81	127	187	216
Stress (kbar)	0	3.7 ± 0.5	3.2 ± 0.4	3.6 ± 0.5
A	1.8	1.9	3.5	4.8
Carrier lifetime (ps)	2000	79	28	18

FIG. 1. Experimental photoluminescence spectra of GaAs films at 77 K (solid curves). A: GaAs on GaAs substrate. B: GaAs on Si substrate (3 μ thick). C: GaAs on Si substrate (1.5 μ thick). D: GaAs on Si substrate (1.5 μ thick and 10 μ wide). The broken curves are theoretical fits to the experimental curves.

in energy so at 77 K its population is negligible. To simplify the calculation, we have neglected the discrete excited states of the excitons and the excitonic enhancement of the continuum band-to-band transition.[11] The error introduced by this simplification was not serious for GaAs because of its very small exciton binding energy. The exciton luminescence is represented by a Lorentzian $L(E)$ centered at the exciton energy ($E_{g_i} - E_{ex_i}$) while the band-to-band transition is described by the function $F(E)$:

$$L_i(E) = |M_i|^2 \mu_i^{3/2} \frac{\Delta_i^2 e^{-(E_{g_i} - E_{ex_i})/kT}}{(E - E_{g_i} + E_{ex_i})^2 + \Delta_i^2}, \quad (1)$$

and

$$F_i(E) = |M_i|^2 \mu_i^{3/2} \begin{cases} (\sqrt{E/E_{g_i} - 1})e^{-E/kT} & E > E_{g_i} \\ 0 & E < E_{g_i} \end{cases} \quad (2)$$

In Eqs. (1) and (2), i refer to either hh or lh bands, $|M_i|^2$ are their optical transition matrix elements, E_{g_i} and μ_i are their band gaps and joint-density-of-states masses, E_{ex_i} and Δ_i are the exciton binding energies and linewidths, T is the carrier temperature, and k is the Boltzmann constant. The function $F(E)$ describes an asymmetric peak whose high-energy side is determined by the temperature T of the carriers. The total PL intensity is obtained by summing the contributions from both the heavy hole and light hole bands:

$$I_{PL} \propto \sum_{i = hh}^{lh} [AL_i(E) + F_i(E)]. \quad (3)$$

A is an adjustable parameter, which represents the relative intensity of the lowest exciton level to the band-to-band transition.

An external hydrostatic stress changes the lattice constant and shifts the energy bands, while a uniaxial stress can lift some of band degeneracies. The stress-induced band-gap shifts and splittings in GaAs have been studied extensively by PL and many other experiments.[12,13] We assume that the stress distribution inside the GaAs/Si films is biaxial with the stress perpendicular to the film equal to zero, while the stress in the plane of the GaAs films is isotropic. One can

show easily that for a GaAs film grown on (100) Si substrate this stress is equivalent to the sum of a hydrostatic pressure and a uniaxial stress perpendicular to the film. The effect of such uniaxial stress on the optical transitions of GaAs can be described by the strain Hamiltonian proposed by Bir et al.[14] For GaAs subjected to a tensile stress in the [100] direction, this Hamiltonian gives the following stress-induced band-gap shifts to the lowest order in stress:

$$\frac{\Delta E_h}{X} = \left[-\frac{2}{3}\left(\frac{\partial E_g}{\partial P}\right) + b(S_{11} - S_{12}) \right]$$
$$= -10.4 \text{ meV/kbar,} \quad (4)$$
$$\frac{\Delta E_l}{X} = \left[-\frac{2}{3}\left(\frac{\partial E_g}{\partial P}\right) - b(S_{11} - S_{12}) \right]$$
$$= -4.9 \text{ meV/kbar,}$$

where ΔE_l and ΔE_h are, respectively, the lh and hh band-gap shifts, S_{11} and S_{12} are components of the compliance tensor of GaAs, X is the magnitude of the isotropic stress in the plane of the films, ($\partial E_g/\partial P$) is the hydrostatic pressure coefficient of the band gap, and b is the deformation potential for uniaxial stress along the [100] direction. To fit the PL spectra for GaAs films grown on Si substrate, we use Eq. (4) to calculate the band gaps E_{g_i} under stress and substitute these values into Eq. (3). Figure 1 shows the experimental PL spectra (solid lines) and the corresponding theoretical fits (broken curves). In fitting curve A (GaAs/GaAs sample) the adjustable parameters are A, E_{ex}, and T and their values are listed in Table I. Within our experimental accuracy we found that $E_{ex} = 0$ meV gave the best overall fit to the experimental spectrum. If we use $E_{ex} = 5$ meV, which is the generally accepted exciton binding energy in GaAs, the theoretical curves show a dip between the exciton peak and the continuum band-to-band contribution. The dip would not be present if the higher discrete exciton states plus the exciton enhancement to the continuum density of states were included in our model. Thus the use of the adjustable parameter A and the assumption that $E_{ex} = 0$ compensate for these omissions. We also noted that in sample A the carrier temperature T was very close to the lattice temperature suggesting that laser heating of the samples was negligible. In fitting the spectra for the GaAs/Si samples we again assumed $E_{ex_i} = 0$ for both hh and lh transitions. The adjustable parameters in these cases are X, Δ_l, Δ_h, T, and A. The results are shown in Table I also. From Eq. (4) the stress X causes a shift in the "center of gravity" of the hh and lh transitions in addition to splitting them. The fact that a single value of X can fit both the shift and splitting of the hh and lh transitions is a strong indication that the biaxial stress we have assumed is basically correct. Furthermore, the fact that X deduced from the different GaAs/Si samples are more or less the same is consistent with its thermal origin.

Besides the splitting of the PL peaks, there are also differences in line shapes and relative peak intensities between all the samples. For example the linewidths of the exciton peaks in the GaAs/GaAs sample were considerably narrower than those of the GaAs/Si samples. The broadening of the exciton peaks in the GaAs/Si samples can result from either shorter exciton lifetimes in the GaAs/Si samples due to dis-

locations or from inhomogeneous broadening due to variation of stress in the sample. We will show later that the exciton lifetime in our GaAs/Si samples is longer than tens of picoseconds (ps) so homogeneous lifetime broadening in these samples is negligible. Thus to reconcile this inhomogeneous stress broadening with the biaxial stress distribution, we have to assume that the stress is locally homogeneous and isotropic but there are variations over an area of the size of our focal spots which were typically larger than 300 μm across. Presumably this stress variation is largest near the edge of the film. This will explain the larger exciton linewidth in the narrower sample D. The stress variation over the film can be mapped out by a tighter focusing of the laser spot and this will be the subject of future investigations.

From the relative intensities of the hh and lh exciton peaks and the line shapes of the higher energy tail of the band-to-band contribution, we found that the carrier temperatures in the GaAs/Si samples are considerably higher than the lattice temperature. (See Table I for the carrier temperatures.) We propose to explain this "hot-carrier" effect in the GaAs/Si samples by the shorter lifetime of photoexcited carriers in these samples due to the higher density of dislocations. To correlate the carrier lifetime with its temperature we assume that the photoexcited electron-hole (e-h) pair lifetime is τ. The rate at which energy is transferred to each e-h pair from the laser beam is given by $H = (\hbar\omega - E_g)/\tau$ where $\hbar\omega$ is the laser photon energy. The rate at which each e-h pair will lose energy due to emission of longitudinal optical (LO) phonons is[15]

$$L = \frac{\hbar\omega_{LO}}{\tau_0}\left[\exp\left(-\frac{\hbar\omega_{LO}}{kT}\right) - \exp\left(-\frac{\hbar\omega_{LO}}{kT_L}\right)\right], \quad (5)$$

where $\hbar\omega_{LO}$ is the LO phonon energy, τ_0 is the time for an e-h pair to emit a LO phonon and is ~ 0.1 ps,[16] and T_L is the lattice temperature assumed to be 77.4 K. At steady state we expect $H = L$ and from this equation we can calculate τ from the experimentally deduced carrier temperature. The resultant τ are shown in Table I for samples A–D. The shorter carrier lifetimes in samples C and D as compared with sample B are presumably caused by their smaller thickness so there are more dislocations at the surface of these films. The carrier lifetimes estimated in this way from these films are somewhat shorter than the lifetime of photoexcited free-car-

riers in GaAs containing nonradiative recombination centers such as Cr.[17] They are consistent with the relative quantum efficiencies of these films determined with the same optical setup. The results of these quantum efficiency measurements will be published elsewhere.[18]

In conclusion, we have shown that from the photoluminescence spectra of GaAs films grown on Si, one can determine the thermal stress, carrier lifetime, and stress inhomogeneity in these films.

[1]R. Fischer, N. Chand, W. Kopp, C. K. Peng, and H. Morkoç, IEEE Trans. Electron Devices ED-33, 206 (1986).

[2]R. Fischer, N. Chand, W. Koop, H. Morkoç, L. P. Erickson, and R. Youngman, Appl. Phys. Lett. 47, 397 (1985).

[3]J. P. van der Ziel, R. D. Dupuis, and J. C. Bean, Appl. Phys. Lett. 48, 1713 (1986).

[4]R. Fischer, H. Morkoç, D. A. Neman, H. Zabel, C. Choi, N. Otsuka, M. Longerbone, and L. P. Erickson, J. Appl. Phys. 60, 1640 (1986).

[5]Tetsuo Soga, Shuzo Hattori, Shiro Sakai, and Masayoshi Umeno, J. Cryst. Growth 77, 498 (1986).

[6]D. A. Neumann, X. Zhu, H. Zabel, T. Henderson, R. Fischer, W. T. Masselink, J. Klem, C. K. Peng, and H. Morkoç. J. Vac. Sci. Technol. B 4, 642 (1986).

[7]Y. Huang, P. Y. Yu, M. Charasse, Y. Lo, and S. Wang, Appl. Phys. Lett. 51, 192 (1987).

[8]S. Zemon, S. K. Shastry, P. Norris. C. Jagannath, and G. Lambert, Solid State Commun. 58, 457 (1986).

[9]S. Zemon, C. Jagannath, E. S. Koteles, S. K. Shastry, P. Norris, G. Lanbert, A. N. M. Choudhury, and C. A. Armiento, in Proceedings of the 13th International Symposium on Gallium Arsenide and Related Compounds, Las Vegas, 1986 (Institute of Physics, Bristol, 1987), p. 141.

[10]In the GaAs on Si samples we expect that the film quality varies with distance from the interface. Although the excited carriers can diffuse away from the surface into the bulk, because of the short lifetime of carriers in GaAs on Si films, we estimate that the ambipolar diffusion length is at most comparable to the penetration length.

[11]R. J. Elliott. Phys. Rev. 108, 1384 (1957).

[12]R. N. Bharagava and M. I. Nathan. Phys. Rev. 161, 695 (1967).

[13]P. Y. Yu, M. Cardona, and F. H. Pollak. Phys. Rev. B 3, 340 (1971).

[14]G. L. Bir, E. I. Butikov, and G. E. Picus, J. Phys. Chem. Solids 24, 1467 (1963).

[15]See, for example, S. A. Lyon, J. Lumin. 35, 136 (1986) and reference therein.

[16]J. A. Kash, J. C. Tsang, and J. M. Hvam, Phys. Rev. Lett. 54, 2151 (1985).

[17]J. S. Weiner and P. Y. Yu, J. Appl. Phys. 55, 3889 (1986).

[18]H. Lee, S. Wang, Y. Huang, and P. Y. Yu (unpublished).

Proposal for three-dimensional internal field mapping by cw electro-optic probing

Y. H. Lo, M. C. Wu, Z. H. Zhu,[a] S. Y. Wang,[b] and S. Wang

Department of Electrical Engineering and Computer Sciences, and Electronics Research Laboratory, University of California, Berkeley, California 94720

The cw electro-optic probing technique is for the first time proposed to detect the three-dimensional internal field distribution in linear electro-optic material like GaAs. By changing the incident angles and positions of the probing beam, sufficient information of the electric field distribution is included in the phase retardation of the probing beam. If certain conditions on the probing beam are satisfied, a very simple linear relation between phase retardation and each field component can be found and the whole problem becomes not only mathematically solvable but experimentally feasible. Finally, a three-dimensional computer simulation is undertaken to illustrate the relation between field distribution and detected electro-optic signal.

To understand the physics of more and more complicated III-V compound devices and to find out proper physical models to describe or predict the device performance, it is essential to know the internal potential, field, and charge distributions. However, there is no technique presently available to detect such crucial information in either bulk semiconductor or devices. For those compound materials having linear electro-optic properties such as GaAs, the phase retardation of light is affected by the electric field inside the material, and thus intrinsically contains the information of electric field in the area swept by the probing light. This attractive property has been widely applied to detect the logic state and transient response of a very fast GaAs circuit,[1-5] or in cw operation, to detect the surface potential variation and the boundary of a *p-n* junction.[6,7] However, none of the work has ever explored the possibility of the great importance of extracting three-dimensional internal field distribution from the electro-optic signals. In this letter, we propose for the first time a nondestructive, high-resolution three-dimensional field measurement method by means of the cw electro-optic probing (CWEOP).

Because GaAs is the most extensively used electronic and optic material having linear electro-optic effect, we will consider it as an example in this paper. The index ellipsoid of GaAs under the electric field is described in Eq. (1).

$$(x^2 + y^2 + z^2)/n_0^2 + 2\gamma_{41}(E_x yz + E_y zx + E_z xy) = 1 , \qquad (1)$$

where γ_{41} is the only nontrivial electro-optic coefficient for the cubic crystal GaAs. If the wave vector k of the probing laser beam has directional cosines α, β, γ with respect to the three crystal axes, the plane normal to k and passing through the origin is described as Eq. (2).

$$\alpha x + \beta y + \gamma z = 0 . \qquad (2)$$

Geometrically, those points satisfying both Eqs. (1) and (2) form an ellipse C normal to k as shown in Fig. 1. The direc-

tions of the major and minor axes of the ellipses represent two eigenvectors of the polarization states of the light, and their magnitudes are equal to the refraction indices of the slow and fast optical axes. In spite of the clarity in geometry, the algebra to find those two axes is very involved. However, if the k is in the *y-z* plane and the angle with the z axis is less than 45°, the problem can be greatly simplified. These conditions are satisfied for all the following derivations and proposed measurements. The simplest mathematical approach is to first project the ellipse C normal to k into an ellipse C' in the *x-y* plane as shown in Fig. 1.

After obtaining the directions and magnitudes of the axes of C' in the way of planar analytic geometry, we can easily find the axes of ellipse C by Eq. (2) since each point in C' possesses the same x and y coordinates as the corresponding point in C. The results are summarized in the following set of equations:

$$\cot(2\theta) \equiv \xi = (2a + c)/d , \qquad (3)$$

where

$$a = (\alpha^2)/(\gamma^2) - 2n_0^2(\alpha/\gamma)\gamma_{41}E_y , \qquad (4a)$$

$$c = 2n_0^2\gamma_{41}(\alpha E_y + \beta E_x)/\gamma - (\alpha^2 + \beta^2)/\gamma^2 , \qquad (4b)$$

$$d = 2[(\alpha\beta)/\gamma^2 + n_0^2\gamma_{41}(\gamma E_z - \alpha E_x - \beta E_y)/\gamma] . \qquad (4c)$$

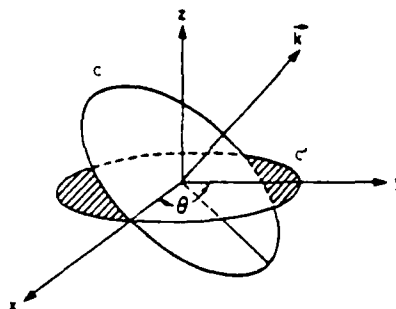The angle θ in Eq. (3) is the angle between the x axis and the

FIG. 1. Configuration of the index ellipse C normal to the wave vector k and its projection C' in the x-y plane. θ is the acute angle between the axis (major or minor) of C' and the x axis.

[a] Permanent address: Department of Radio and Electronic Engineering, Zhejiang University, Hangzhou, China.
[b] Permanent address: Hewlett-Packard Laboratories, Palo Alto, CA.

axis (major or minor) of ellipse C' such that $0 < \theta < (\pi/2)$. The directional cosines α, β, γ are those of k. The two extrema of the ellipse normal to k have the coordinates described as follows:

$$x_\pm = n_0 \cos\theta \sqrt{(1+c') \pm d'}, \tag{5a}$$

$$y_\pm = n_0 \sin\theta \sqrt{(1+c') \pm d'}, \tag{5b}$$

$$z_\pm = -(\alpha x_\pm + \beta y_\pm)/\gamma, \tag{5c}$$

and

$$n_\pm = \sqrt{x_\pm^2 + y_\pm^2 + z_\pm^2}$$
$$= n_0\sqrt{(1+c') \pm d'}\left(1 + \frac{(\alpha\cos\theta - \beta\sin\theta)^2}{\gamma^2}\right), \tag{6}$$

where $c' = c/2$ and $d' = d/[2\sin(2\theta)]$.

After finding the field dependence of the eigenstates for arbitrary incident light, we look for the relation between the detected variation and spatial distribution of electric field by the prescribed electro-optic effect. For simplicity but without loss of generality, let us assume the probing light initially linearly polarized along the x axis, and then its polarization state can be represented by the Jones vector as

$$J = \begin{bmatrix} 1 \\ 0 \end{bmatrix}. \tag{7}$$

To find the field distribution from a finite number of probing beams, we need to set up the mesh points in the tested area. For each small unit of those meshes, the transformation of the polarization state from input beam to output beam is described as follows[8]:

$$\begin{bmatrix} J_{xi} \\ J_{yi} \end{bmatrix} = e^{-i\Delta_i} R_i^{-1} \begin{bmatrix} e^{-i\Gamma/2} & 0 \\ 0 & e^{i\Gamma/2} \end{bmatrix}$$
$$\times R_i \begin{bmatrix} J_{xi-1} \\ J_{yi-1} \end{bmatrix} = W_i \begin{bmatrix} J_{xi-1} \\ J_{yi-1} \end{bmatrix}, \tag{8}$$

$$R_i = \begin{bmatrix} \cos\Phi_i & \sin\Phi_i \\ -\sin\Phi_i & \cos\Phi_i \end{bmatrix}, \tag{9}$$

$$\Gamma_i = (\omega l_i/c)(n_{si} - n_{fi}), \tag{10}$$

$$\Delta_i = (\omega l_i/2c)(n_{si} + n_{fi}). \tag{11}$$

where R_i is the matrix rotating the initial polarization state, along x axis in this case, to the slow axis of the index ellipse. The quantity Γ_i is the phase retardation, and n_{si} and n_{fi} are slow and fast refraction indices of the ith mesh, and l_i is the length of the optical path in mesh i. Because the $\exp(-i\Delta_i)$ term in Eq. (8) does not have any physical significance but a phase shift, we will omit it in the following derivations.

If the y component of the polarization state is detected from the emerging beam, the final J_{yk} is obtained from a series of multiplication of the transformation matrices W_i in Eq. (8) along the light path.[8]

$$J_{yk} = \frac{-i}{2} \sum_i \Gamma_i \sin(2\Phi_i). \tag{12}$$

where the summation is over all the mesh points passed by the probing beam with wave vector k. In Eq. (12), we have approximated $\exp(-i\Gamma_i) \approx 1 - i\Gamma_i$ since the phase retardation in each mesh is always much smaller than one radian.

In general, both Γ_i and Φ_i in Eq. (12) are functions of E_{xi}, E_{yi}, and E_{zi} resulting in a number of cross terms in Eq. (12). To simplify the analysis, we must choose the incident beam to be either in y-z plane or in x-z plane (i.e., $\alpha = 0$ or $\beta = 0$). Assuming k in y-z plane (i.e., $\alpha = 0$) as an example, we simply set Φ_i in Eqs. (9) and (12) equal to θ_i in Eq. (3) if d in Eq. (4c) > 0 or equal to $\theta_i + \pi/2$ if $d < 0$. The $\pi/2$ difference between Φ_i and θ_i as $d < 0$ is due to the previous definition that the angle Φ_i is between x axis and the slow axis but θ_i is confined in $(0, \pi/2)$ to ensure $\sin(2\theta_i)$ in Eq. (6) larger than zero. By combining Eqs. (3), (4), (6), (10), and (12) and the above relation between θ_i and Φ_i with the condition $\alpha = 0$, J_{yk} can be explicitly represented in terms of the field components as follows:

$$J_{yk} = \frac{K}{|\gamma|} \sum_i \left(E_{zi} - \frac{\beta}{\gamma}E_{yi}\right)\left[1 + \left(\frac{\beta}{\gamma}\right)^2 \sin^2\theta_i\right]^{1/2}, \tag{13}$$

$$K = -i(\omega l_0 n_0^3 \gamma_{41})/(2c). \tag{14}$$

The field-dependent nonlinear term in the square root of Eq. (13) is negligible as long as the incident beam has a small angle with the z axis, in other words, $\beta^2/\gamma^2 \ll 1$. Therefore, Eq. (13) can be approximated as

$$J_{yk} \approx \frac{K}{|\gamma|} \sum_i \left(E_{zi} - \frac{\beta}{\gamma}E_{yi}\right). \tag{15}$$

Equation (15) is the equation setting up the linear relation between the field components and the detected probing beam from different angles and positions. The E_{xi} in Eq. (15) is missing because of the condition $\alpha = 0$. To find E_{xi}, we can replace βE_{yi} with αE_{zi} in Eq. (15) assuming that the light is in x-z plane. The signal from a photodetector is proportional to the magnitude of J_{yk} based on the same principle as the amplitude modulation of electro-optic material.[9] Therefore, the three-dimensional electric field distribution measurement becomes not only theoretically possible but practically feasible by means of the recently developed cw electro-optic probing technique.[6,7] From the measured field distribution, the real values of potential profile, field intensity, and charge density can be substantially obtained since the applied voltage supplies the necessary boundary conditions.

Figure 2(c) shows the computer simulation results of the electro-optic signals from Eq. (15) under some given E_y, E_z distributions shown in Figs. 2(a) and 2(b). The detected area in Fig. 2 is $100 \times 100 \ \mu m^2$ in the y-z plane and the incident probing beam scans from $-18°$ to $18°$ with $1°$ difference between adjacent beams. Not only does the beam incident angle vary, the probing beam also moves along the y axis with $0.18 \ \mu m$ per step to supply enough linearly independent equations to solve the field components at each mesh point. A larger mesh size can be used in those regions of slow-varying field to reduce the number of probing beams and save computer memories without sacrificing the accuracy. Finally, a simple experiment is done to justify our proposed method in practice. As shown in Fig. 3, the electro-optic signals measured under normal incidence and $1.4°$ tilted incidence are clearly distinguishable. Those features occurring in the tilted incidence curve are partly due to the different optical path from the normal incident one and partly due to
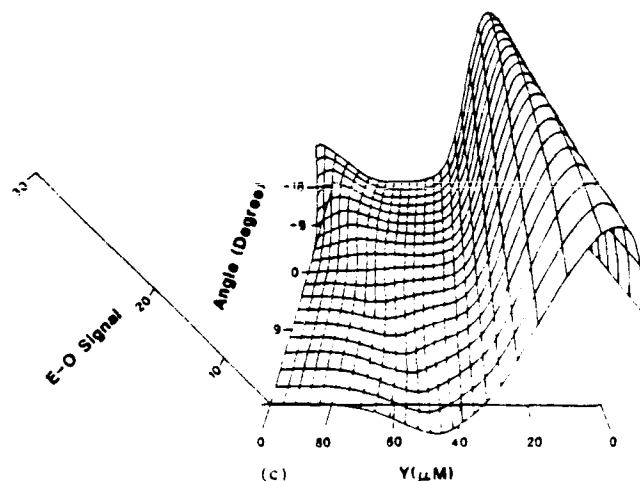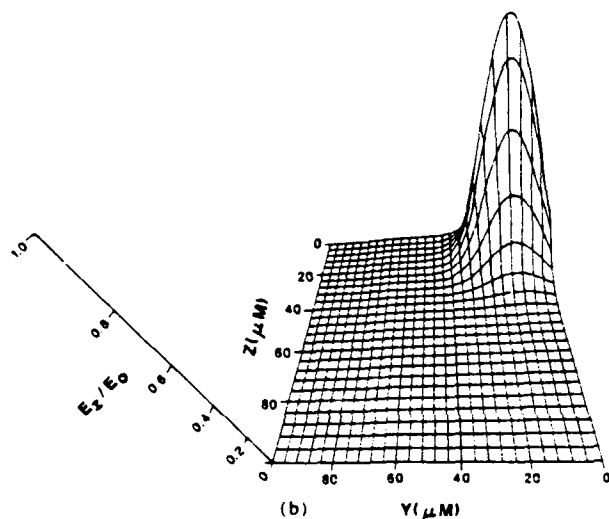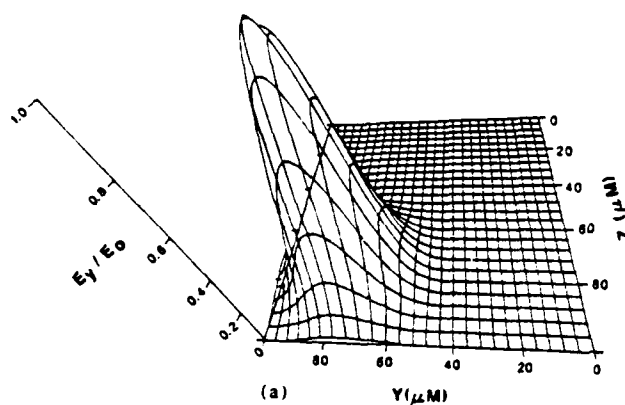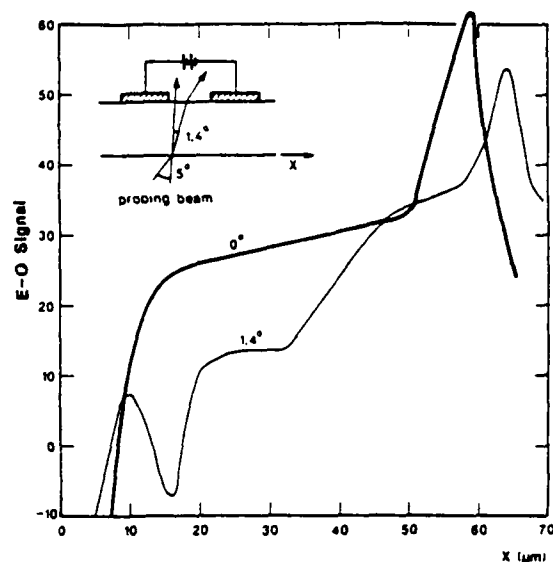
(a)

FIG. 3. Measured electro-optic signal across two electrodes on GaAs substrate. The dark curve is measured under normal incidence and the light one is measured under incidence at an angle. Lock-in technique is used in this experiment.

the influence of lateral field. The more complete experiment and the detailed analysis of the experimental data will be published elsewhere.

In conclusion, we propose a very new method to nondestructively detect the three-dimensional field distribution in linearly electro-optic materials such as GaAs, and mathematically formulate it. If an appropriate incident angle is chosen, the electro-optic signal is a linear combination of the field components in the light path. This property facilitates the data analysis and makes the experiment more feasible. To obtain more accurate results, the finite beam spot size, multiple reflection inside the material, reflection at the interface, as well as the light shape function all need to be taken into account.

(b)

(c)

FIG. 2. Computer simulation of the electro-optic signal under the given field profile in y-z plane. E_y and E_z profiles are shown in (a) and (b) but E_x is not shown here because it does not affect the electro-optic signal if the light has the wave vector in the y-z plane; (c) shows the signal variation as both the incident angle and the incident position along the y axis change.

[1] B. H. Kolner and D. M. Bloom, IEEE J. Quantum Electron. QE-22, 69 (1986).

[2] J. L. Freeman, S. K. Diamond, H. Fong, and D. M. Bloom, Appl. Phys. Lett. 47, 1083 (1985).

[3] K. Stenersen and R. K. Jain, IEEE Electron Device Lett. EDL-5, 422 (1984).

[4] A. J. Taylor, J. M. Wiesenfeld, R. S. Tucker, G. Eisenstein, J. R. Talman, and U. Koren, Conference on Lasers and Electro-Optics Technical Digest 312 (1986).

[5] R. K. Jian, X-C. Zhang, M. G. Ressl, and T. J. Pier, Conference on Lasers and Electro-Optics Technical Digest 312 (1986).

[6] Z. H. Zhu, J. -P. Weber, S. Y. Wang, and S. Wang, Appl. Phys. Lett. 49, 432 (1986).

[7] Y. H. Lo, Z. H. Zhu, C. L. Pan, S. Y. Wang, and S. Wang, Appl. Phys. Lett. 50, 1125 (1987).

[8] A. Yariv and P. Yeh, Optical Waves in Crystals (Wiley, New York, 1984), pp 121-125.

[9] A. Yariv, Quantum Electronics (Wiley, New York, 1975), pp. 339-341.

Electro-optic measurement of standing waves in a GaAs coplanar waveguide

Z. H. Zhu,[a] C. L. Pan,[b] Y. H. Lo, M. C. Wu, and S. Wang
*Department of Electrical Engineering and Computer Sciences and Electronics Research Laboratory,
University of California, Berkeley, California 94720*

B. H. Kolner and S. Y. Wang
Hewlett–Packard Laboratories, Palo Alto, California 94304

We have successfully measured, for the first time, standing waves in a GaAs coplanar
waveguide at frequencies of 8.2107 and 12.310 GHz by using harmonic-mixing electro-optic
probing. The technique is nondestructive and has great potential in expanding the measuring
frequency band to millimeter wave. This letter describes the principle of the technique, the
experimental setup, and the measurement results.

In recent years, many exciting applications of the linear electro-optic effect as a measurement technique have been explored. The electro-optic sampling technique has been applied to detect the transients of ultrafast electronic and optoelectronic devices.[1,2] More recently, cw electro-optic probing (CWEOP) has been used to measure the potential profile, field, and charge distribution in GaAs material and devices.[3–5]

In GaAs microwave integrated circuits, a standing-wave electric field will be established in a waveguide if the impedance of the load is mismatched to the characteristic impedance of the waveguide. Therefore, in principle, the electro-optic probing technique can be applied to detect the distribution of standing-wave fields in GaAs microwave waveguides.

Based on this idea, we have successfully measured, for the first time, the distribution of a standing-wave field formed in a GaAs waveguide using a new technique, harmonic-mixing electro-optic probing. The principle of the technique, the experimental setup, and results are presented in this letter.

Conventional slotted line techniques become impractical at high microwave frequencies in waveguides. No technique exists for monolithic microwave transmission lines. Here we employ the harmonic-mixing electro-optic probing to overcome these difficulties. Briefly, when a laser pulse train from a mode-locked YAG laser, modulated by an electric field, illuminates a photodetector, the photocurrent produced by the photodetector is proportional to the product of the light intensity and the modulating voltage. The modulator (GaAs coplanar waveguide sample in our case)-photodetector combination can be viewed as a mixer. In the frequency domain, any signal at frequency f_m propagating in the waveguide will mix with all harmonics of the fundamental repetition frequency f_0 of the mode-locked laser pulse. Sidebands will appear at frequency $nf_0 - f_m$. The mixing current is given by[6]

$$i_{mix} = i_{avg}\pi\frac{V_m}{V_\pi}\sum_{n=1}^{\infty}\left(\frac{\sin(\pi nf_0\tau)}{\pi nf_0\tau}\right)$$
$$\times\left[\sin 2\pi(nf_0 + f_m)t - \sin 2\pi(nf_0 - f_m)t\right], \quad (1)$$

where i_{avg} is the average photocurrent, V_m is the amplitude of the modulating voltage, V_π is the half-wave voltage of the GaAs modulator, n is the order of harmonic component, and τ is the pulse width of the mode-locked laser. Insofar as the mode-locked laser as a local oscillator is concerned, the only requirement is to select a proper frequency f_m of rf signal, then we can easily measure the mixing electro-optic signal at an intermediate frequency f_i, which is given by

$$V_0 \propto i_{avg}\pi(V_m/V_\pi)R_L, \quad (2)$$

where V_0 is the electro-optic peak voltage developed across a load resistor R_L, and is proportional to the modulating voltage V_m.

The experimental setup is shown in Fig. 1. A laser pulse train of about 10 ps duration at a repetition rate of $f_0 = 82$ MHz is generated by the compressed output of a cw mode-locked Nd:YAG laser ($\lambda = 1.06 \mu m$). The quarter-wave plate introduces a phase shift of $\pi/2$ between the "fast" and "slow" components of light, so that the electro-optic modulator (i.e., the GaAs waveguide sample) operates in the linear region. The laser beam passes through the GaAs sample and is focused to a diameter of about 8 μm on the central electrode of the coplanar waveguide. The widths of the central electrode, side electrodes, and the spacing between them are 135, 590 and 65 μm, respectively. The rf output with frequency f_m of an oscillator is fed to the GaAs coplanar waveguide (CPW) via a connector (type SMA), and the other terminal of the waveguide is open, so a standing-wave field is established along the waveguide. The laser beam modulated by the standing-wave field is reflected from the electrode, then passes through an analyzer to convert the phase change into amplitude change. A low-speed germanium photodetector is used to detect the mixing signal at the intermediate frequency f_i. A preamplifier is inserted between the photodetector and a spectrum analyzer to optimize the S/N ratio. A TV camera is used to monitor the measuring position exactly. By scanning the laser beam from the open-circuit terminal to the other terminal along the lon-

[a] Permanent address: Department of Radio and Electronic Engineering, Zhejiang University, Hangzhou, China.
[b] Permanent address: Institute of Electro-Optical Engineering, Tung University, Hsinchu, Taiwan, China.
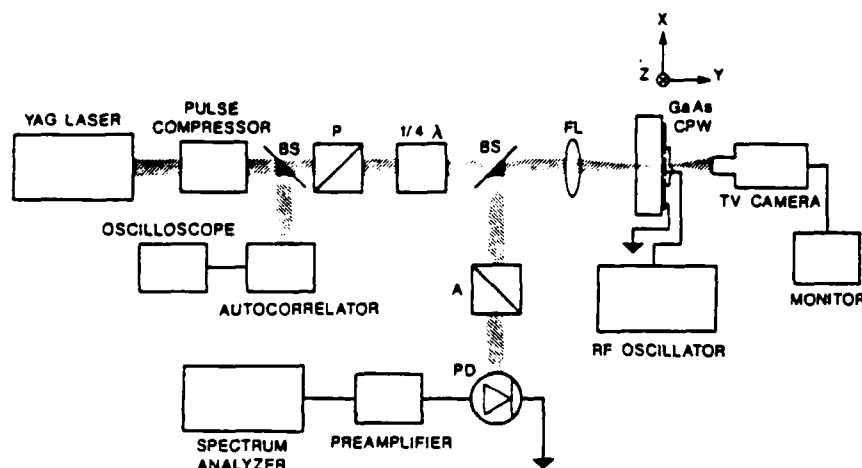
FIG. 1. Schematic diagram of the experimental setup: P—polarizer, A—analyzer, BS—beamsplitter, FL—focusing lens, PD—photodetector.

gitudinal ⟨011⟩ direction (z axis in Fig. 1) of the CPW electrode, we can detect the standing-wave pattern.

Figure 2 shows the harmonic-mixing electro-optic signal displayed on the spectrum analyzer by mixing the rf signal ($f_m = 8.2107$ GHz) with the 100th harmonic component ($n = 100$) of the fundamental repetition rate ($f_0 = 82$ MHz) of the mode-locked pulse. The intermediate frequency f_i of the mixer signal is equal to 10.7 MHz. The corresponding standing-wave pattern measured is given in Fig. 3. The zero of the abscissa corresponds to the point where the measurement began, but it should be pointed out that the real zero point corresponding to the edge of the open terminal of CPW locates at about 45 mil (≈ 1.14 mm) to the left of the zero of the abscissa. From the pattern we can see that the distance λ_{sw} between valleys is about 270 mil (≈ 6.9 mm) which is equal to the half-wavelength of the rf signal. Substituting these data into the following well-known formula for evaluating the effective index n_e of GaAs CPW,

$$n_e = c/2\lambda_{sw}f_m, \qquad (3)$$

where c is the speed of light in air, we get $n_e = 2.65$, which is in good agreement with the value used in the design of micro-

wave GaAs integrated circuits. Similar measurement at the rf frequency of 12.310 GHz (by mixing with the 150th harmonic component of the fundamental repetition rate $f_0 = 82$ MHz of mode-locked pulse) yields $\lambda_{sw} = 180$ mil (≈ 4.6 mm). This is in excellent agreement with the value of λ_{sw} as evaluated using $n_e = 2.65$. From the measured standing-wave pattern, we can evaluate the voltage standing-wave ratio (VSWR) ρ and the reflection coefficient Γ of the GaAs CPW, the result is as follows: $f_m = 8.2107$ GHz, $\rho = 5.97$, $\Gamma = 0.71$; $f_m = 12.310$ GHz, $\rho = 7.94$, $\Gamma = 0.78$. The reason for $\rho \neq \infty$ can be attributed to fringing field capacitance at the end of the transmission line and electrode losses.

In summary, we have successfully measured the standing-wave pattern in a GaAs coplanar waveguide at frequencies of 8.2107 and 12.310 GHz by using harmonic-mixing electro-optic probing. The experimental results are in agreement with the theory and the reproducibility of the measurement is excellent. The technique is reliable and noninvasive. Since the input bandwidth of this technique easily exceeds 100 GHz for picosecond mode-locked pulses, we believe in the near future the technique can be developed to become a standard in making noncontact on-chip measurement of standing waves in GaAs millimeter wave integrated circuits.
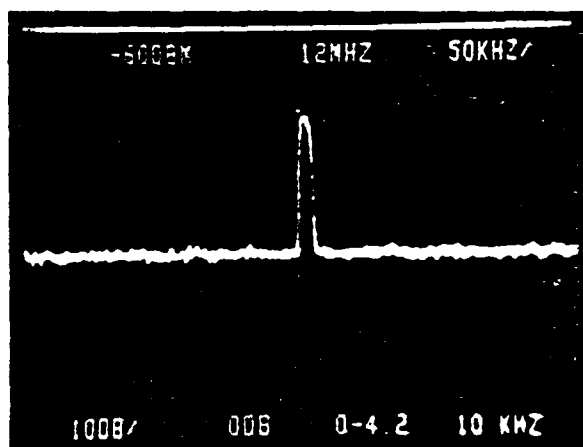
FIG. 2. Harmonic-mixing electro-optic signal $\cdots$ $\cdots$ $= 82$ MHz. $n = 100$, and $f_i = 10.7$ MHz).
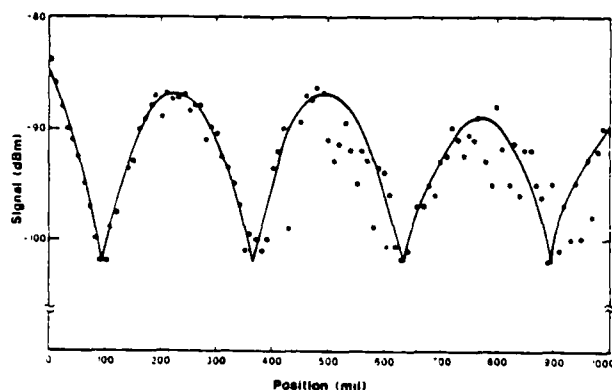

FIG. 3. Standing-wave pattern in GaAs CPW with an open terminal ($f_m = 8.2107$ GHz. $f_0 = 82$ MHz. $n = 100$, and $f_i = 10.7$ MHz).

[1]J. A. Valdmanis and G. A. Mourou, IEEE J. Quantum Electron. QE-22, 69 (1986).

[2]B. H. Kolner and D. M. Bloom, IEEE J. Quantum Electron. QE-22, 79 (1986).

[3]Z. H. Zhu, J-P. Weber, S. Y. Wang, and S. Wang, Appl. Phys. Lett. 49, 432 (1986).

[4]Z. H. Zhu, Y. H. Lo, S. Y. Wang, and S. Wang (unpublished).

[5]Y. H. Lo, Z. H. Zhu, C. L. Pan, S. Y. Wang, and S. Wang, Appl. Phys. Lett. 50, 27 April (1987).

[6]B. H. Kolner, Ph.D thesis, Stanford University, 1985.

Resonant tunneling diodes with AlAs barrier: Guides for improving room-temperature operation

Daryoosh Vakhshoori and Shyh Wang

Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, California 94720

In this communication experimental results of AlAs/GaAs/AlAs resonant tunneling diodes are compared to theory, and the effect of higher resonant states on the room-temperature resonant tunneling current peak-valley ratio is studied. It is shown that for a well thickness of 70 Å, the leakage current through higher resonant states of the well can reduce the peak-to-valley current ratio by more than 400% when compared to a similar diode of 50-Å well thickness. Through these studies a summary of suggestions is listed, which is intended to improve the room-temperature operation of resonant tunneling diodes.

The idea of using the semiquantized states in ultrathin heterostructures for obtaining negative differential resistance was first proposed by Esaki and Tsu[1] and developed by other workers[2] as early as 1970. As the molecular-beam epitaxy and other epitaxial growth techniques developed, renewed interest appeared in the field, and theoretical and experimental papers aiming at obtaining room-temperature observation of negative differential resistance were published.[3-5] The observation of Sollner *et al.*,[6] who have reported 2.5-THz operation of resonant tunneling diodes at 77 K, has proven the potential of these diodes. Also, the recent results of Tsuchiya *et al.*[3,4] show that the material quality of resonant tunneling diodes has improved enough to allow the observation of peak-valley current ratio (P/V) of 3 at 296 K. Comparing the results of Tsuchiya *et al.* to other results of AlAs barrier resonant tunneling diodes[7] shows that design and material quality, and not phonon scattering, are the factors limiting the peak-valley ratio at room temperature. The potential large peak-valley current ratio and ultrahigh-frequency characteristic of these diodes at room temperature make them ideal for high-frequency generation and detection systems. However, to realize the maximum capability of the resonant tunneling diodes, a systematic study of the parameters affecting the current-voltage characteristic of these devices is necessary. These parameters include the effective masses in the barrier and well, the thickness of the well region, and the voltage at which the current resonance peak occurs. In this paper the significance of these parameters is analyzed, and conclusions are drawn which discuss the choice of parameters that can improve the peak-valley current ratio at room temperature.

The structure under consideration is shown in Fig. 1. In the conventional approach to the calculation of resonant tunneling current, the Schrödinger equation is solved in the barrier, and the well and the solutions are matched at the boundary to satisfy the continuity requirements. Then the contribution to the current from each solution is weighted by the Fermi–Dirac function and integrated to obtain the total tunneling current. This procedure is very difficult to be carried out self-consistently; hence approximations must be made to simplify the process. In the approach used in this paper, the following two points have been considered in obtaining the tunneling current. First, in solving the Schrödinger equation, the Wentzel–Kramers–Brillouin (WKB) approximation has been used. This approximation allows us to incorporate the effect of electric field on the solution of Schrödinger wave equation. Second, to obtain a more accu-
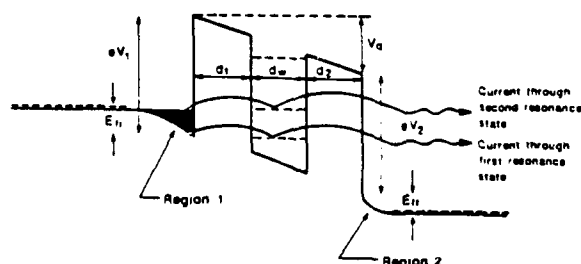

FIG. 1. General resonant tunneling diode under consideration. The quantization of electron energy in region 1 is not expected to influence our analysis substantially.
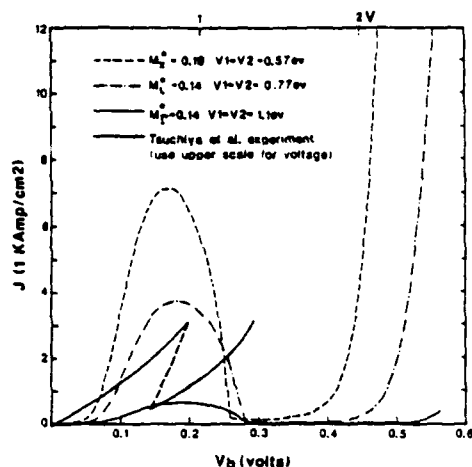
FIG. 2. Low-temperature comparison of Tsuchiya et al.'s (Ref. 4) experiment with current contribution from Γ, L, X valleys. Note that the Γ-valley current alone cannot give agreement with experiment. The dashed line in the experimental curve is due to finite ohmic contact resistance. The device parameters are $d_1 = d_2 = 22.6$ Å, $d_w = 70$ Å, temperature $= 77$ K, and $N = 8 \times 10^{17}$ cm^{-3}.

rate picture of the distribution of electrons in the conduction band, a self-consistent energy-band-bending model, taking into account the voltage drop due to the accumulation of electrons in region 1 and depletion of electrons from region 2 of Fig. 1, is used. The model can be summarized as follows (GaAs base materials only):

$$V_b \ (V) = (E_{fi} + eV_q + E_{ff})/e , \tag{1}$$

$$E_{fi} \ (eV) \simeq 2.65 \times 10^{-4} \left(\frac{V_q^2}{(d_1 + d_2 + d_w)} \right)^{2/3} , \tag{2}$$

$$E_{ff}(eV) \simeq 313 N_d^{-1} \left(\frac{V_q}{(d_1 + d_2 + d_w)} \right)^2 , \tag{3}$$

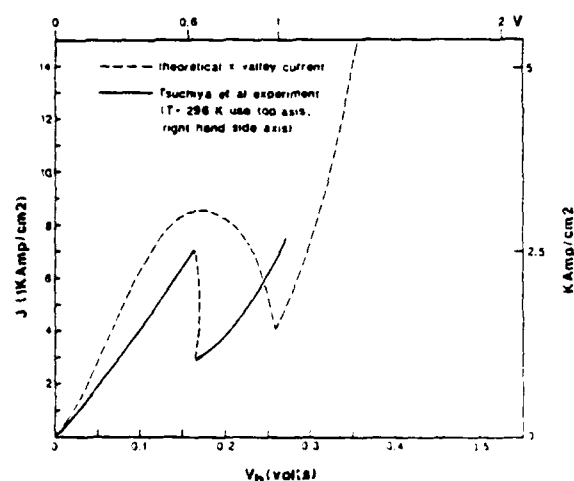where N_d (cm^{-3}) is the doping outside the barrier, d_1, d_2,

FIG. 3. Room-temperature comparison of experimental [Tsuchiya et al. (Ref. 4)] and theoretical resonant tunneling currents. The general behavior between the two curves agree well. The parameters used are those of X valleys, as indicated in Fig. 2.
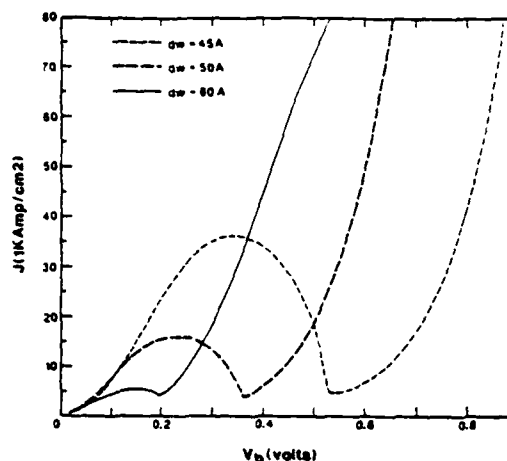

FIG. 4. Current-voltage characteristic of resonant tunneling diode for three well thicknesses. As the well thickness decrease, the room-temperature peak-valley ratio increases, and the voltage at the current peak (V_t) also increases. The other parameters are the same as those of Fig. 3.

and d_w are in meters, and the quantities E_{fi}, E_{ff}, V_q, d_1, d_2, and d_w are defined in Fig. 1. It should be noted that E_{fi} is calculated by linearly approximating the conduction-band energy in region 1, and E_{ff} is obtained by the usual depletion approximations in region 2.

With the use of the above-mentioned consideration and approximations, the effect of parameters on the tunneling current will not be discussed. To analyze the importance of the effective mass of electrons in an AlAs barrier, we have to decide which band minima to use. AlAs is an indirect band-gap material, whereas GaAs is a direct one. Thus, to analyze the tunneling current contribution of each conduction-band minima (Γ, L, X) of AlAs barrier to the overall current, three effective masses (m_Γ^*, m_L^*, m_X^*) are assigned to the corresponding conduction-band minima of the AlAs barrier region. Then each band-minimum state of GaAs is separately matched to the Γ-band minima of the GaAs regions. The corresponding resonant tunneling currents and experimental results of Tsuchiya et al.[4] are plotted in Fig. 2. Different axes are used for theoretical and experimental values to account for series ohmic resistance. As can be seen, it is no longer valid to assign the Γ-band effective mass of GaAs to that of AlAs barrier. By using the effective masses given in Adachi's review article,[8] it is concluded that the Γ valley of AlAs alone cannot account for the observed current-density peak of 2 kA/cm^2 of Tsuchiya et al.'s[4] experiment. A linear combination of all three band minima are needed to account for the resonant tunneling current. Note also that the current is a sensitive function of effective mass and barrier height. Without a detailed knowledge of the real and complex Bloch wave function and energy-crystal momentum-dispersion relation of AlAs barrier, it would be useless to calculate the magnitude of total resonant tunneling current.

However, for studying the room-temperature behavior of resonant tunneling diodes, we do not need to go into this detail, the tunneling current due to the X valley will suffice as our qualitative framework. This choice is not a bad one, since X minima of AlAs represents the lowest energy barrier

(~0.57 eV) to the electrons of the GaAs well. From conventional quantum-mechanical calculations, it is known that the smaller the barriers of quantum well, the less is the separation of the first and second quantized energy levels. This smaller energy separation will in turn contribute to higher unwanted tunneling current through a second resonant state for the X valley compared to the Γ or L valley, especially at room temperature where some electrons are energetic enough to tunnel through the second resonant state. Figure 3 compares the qualitative behavior of calculated X-valley resonant tunneling current to that of Tsuchiya et al.'s experiment at 296 K. As can be seen, the peak-valley current ratio (P/V) of experiment and theory agree well. The thermionic current was also included, but the effect is negligible.

To decrease the unwanted current through the second resonant state and thus to increase the peak-valley current ratio, the well region should be decreased in thickness. Figure 4 shows the current-voltage characteristics of these diodes for different well thickness at room temperature. It can be seen that the peak-to-valley ratio can easily be improved by a factor of 4 by decreasing the well thickness from 70 to 50 Å. The result shown in Fig. 4 is very encouraging for room-temperature operation of resonant tunneling diodes of AlAs/GaAs/AlAs type. However, one should not expect to observe such a large improvement in practice. Not all unwanted leakage currents are due to higher resonance states; the leakage current due to phonon-assisted processes currents due to deep levels and surface states, and conduction through the L valley of the GaAs well should also be considered. The other mechanisms can be considered as an effective shunt conductance. As the thickness of the well region is decreased, the voltage at which current peak occurs (V_r) will increase (Fig. 4). Also the leakage current through the above-mentioned effective shunt conductance increases as the applied voltage is increased. The increase of leakage current as the well thickness is decreased is also observed experimentally. Tsuchiya and Sakaki[9] observed that the valley current, which is a measure of leakage current, increases superlinearly as the well thickness decreases. Thus, as the well thickness decreases, the leakage current at the bias voltage required for resonance (V_r) increases. Therefore, there is an optimum well thickness at which the leakage current due to other phenomena will become more important than the unwanted current due to second resonant state. This thickness is expected to be less than 60 Å.

To summarize the discussion above for enhancing the nonlinear characteristic of resonant tunneling diodes, the following conclusions are made: (a) To build up the resonant state, the electron scattering should be small. This means that the interface quality should be good. To get abrupt, homogeneous interfaces, the method of interrupted growth,[10] which has resulted in one of the best-reported room-temperature peak-valley ratios, is recommended. (b) Use AlAs instead of $Al_x Ga_{1-x} As$ as the barrier. Higher barrier means more quantization of states in the well, which can reduce unwanted current leakages. (c) Use well thickness of less than 70 Å to maximize peak-valley current ratio. Consideration of the above-mentioned points is expected to be of use in designing and growing resonant tunneling diodes of large nonlinear current-voltage characteristics to be operated at room temperature.

[1]L. Esaki and R. Tsu, IBM J. Res. Dev. 14, 61 (1970).
[2]R. F. Kazarinov and R. A. uris, Sov. Phys. Semicond. 6, 120 (1972).
[3]M. Tsuchiya, H. Sakaki, and J. Yoshino, Jpn. J. Appl. Phys. 2, L466 (1985).
[4]M. Tsuchiya and H. Sakaki, Jpn. J. Appl. Phys. 25, L185 (1986).
[5]B. Ricco and Y. Azbel, Phys. Rev. B 29, 1970 (1984).
[6]T. C. L. G. Sollner, W. D. Goodhue, P. E. Tannenwald, C. D. Parker, and D. D. Peck, Appl. Phys. Lett. 43, 588 (1983).
[7]A. R. Bonnefoi, R. T. Collins, T. C. McGill, R. D. Burnham, and F. A. Ponce, Appl. Phys. Lett. 46, 285 (1985).
[8]S. Adachi, Jpn. J. Appl. Phys. 58, R1 (1985).
[9]M. Tsuchiya and H. Sakaki, Appl. Phys. Lett. 49, 88 (1986).
[10]H. Sakaki, M. Tanaka, and J. Yoshino, Jpn. J. Appl. Phys. 24, L417 (1985).

Surface-emitting second-harmonic generator for waveguide study

D. Vakhshoori, M. C. Wu, and S. Wang

Department of Electrical Engineering and Computer Sciences, and the Electronics Research Laboratory, University of California, Berkeley, California 94720

A novel surface emission of coherently generated second-harmonic wave is reported for the first time. The technique is used for the observation of the difference in propagation constant of the TE_0 and TM_0 mode of GaAs/AlGaAs waveguide cavity to a high degree of accuracy. In this technique the second-harmonic signal propagates out from the top surface of the waveguide structure, converting the modal phase difference between TE_0 and TM_0 modes into intensity variation along the waveguide length. The second-harmonic signal is easily observable by the naked eye, and the technique does not require wavelength tuning or mechanical movement for the measurement of birefringence.

In this letter we report phase-matched, coherent, second-harmonic generation of a mode-locked Nd:YAG laser beam propagating pependicular to the direction of flow of the fundamental beam. In conventional second-harmonic-generation techniques, the amount of second harmonic propagating out perpendicular to the direction of a forward propagating fundamental beam is negligible, since the phase-matching condition is violated. However, in the technique presented here, a standing wave of second-harmonic polarization is generated along the waveguide cavity. The second-harmonic signal generated this way is show to be coherent and to satisfy the wave vector conservation. To the best of the authors' knowledge, this is the first time the second-harmonic wave has been generated in this manner and used to study waveguide parameters.

The parameter studied in this experiment is the modal birefringence of a waveguide structure. The difference in propagation constants of the TE_0 and TM_0 mode of a GaAs/AlGaAs waveguide is of paramount importance for an efficient electro-optic waveguide TE↔TM mode conversion[1] and wavelength filtering.[2] Waveguide polarization conversion is a basic signal processing function and will be useful in realizing single-mode communication systems. In addition to its integrability, the TE↔TM mode convertor together with integrated polarizer[3] can be used as a high-speed optical modulator. TE↔TM mode convertor can be realized as a broad-band or narrow-band (full width at half-maximum ~4.5 Å)[2] polarization convertor. Thus TE↔TM convertors can be used as a polarization convertor, an optical modulator, or a wavelength filter in future opto-electronic circuits. However, to realize any of the above devices, an exact knowledge of modal phase difference $\Delta\beta = \beta_{TE} - \beta_{TM}$ is required. The observation of surface emission of the second-harmonic signal gives the modal phase difference accurately (error ≲1%), directly, and without any ambiguity. Furthermore, the requirement that the second-harmonic signal is not absorbed strongly by the top cladding layer of the waveguide is not severe, since some sommon electro-optic polarization modulators on GaAs are made with transparent top cladding layer.[1]

The experimental setup used is shown in Fig. 1. Abut 20 mW average power from a mode-locked Nd:YAG ($\lambda = 1.064\ \mu$m) laser output with a repetition rate of 82 MHz and pulse width of 100 ps was focused by a 40× objective lens on a facet of a waveguide cavity 1 mm long. The waveguide structure is shown in the inset of Fig. 1. The output power from the other facet was measured to be ~1 mW. The second-harmonic generation was monitored by looking into a low-magnification microscope and optimized by changing the polarization of the incident fundamental beam by using a half-wave plate to equalize the power coupled into TE_0 and TM_0 mode of the waveguide. Once the green light (532 nm) was optimized by eye, the image of the near field was recorded on a 1600 ASA Fuji film exposed for 3 min. The picture was taken by using a commercial camera and looking into the eye piece of the microscope. The picture is shown in Fig. 2. There are 16 periods of intensity modulation in 1 mm corresponding to $\beta_{TE_0} - \beta_{TM_0} = 50$ rad/mm. This result is obtained by theoretical consideration of the relation between second-harmonic-intensity modulation and modal phase difference.

From theory we have

$$E_i(z) = \mathrm{Re}\left[e^{i\omega t}(A_i e^{-\gamma z} + B_i e^{\gamma(z - 2L)}) \right],$$

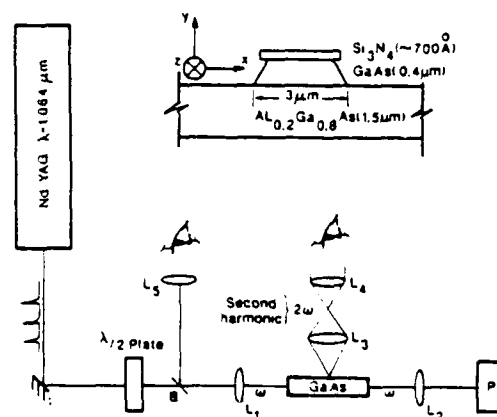$$A_i = \frac{E_{0i} t_i}{1 - r_i^2 e^{-2\gamma L}}, \quad \frac{B_i}{A_i} = r_i, \tag{1}$$

FIG. 1. Experimental setup used. L_1 couples light into the GaAs waveguide. L_2 collimates the light to the power meter P. L_3 and L_4 form a low-magnification microscope. L_5 and beam splitter B are used to adjust coupling through lens L_1 to the waveguide. The $\lambda/2$ plate is used to change the polarization of the incident fundamental beam. The waveguide under consideration is shown in the inset. It was grown by molecular beam epitaxy over [100] oriented GaAs wafer.
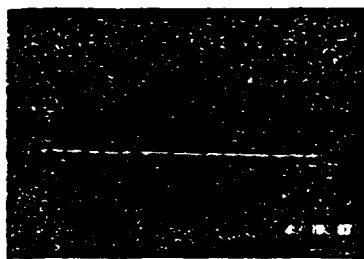
FIG. 2. Picture of the square of the near-field pattern of the second harmonic of Nd:YAG generated across the waveguide. There are 16 periods of variation over 1 mm of waveguide length, giving a period of $\Lambda \sim 63\ \mu m$.

where $i = x$ and y corresponds to the E_x and E_y component of TE_0 and TM_0 modes, respectively (z is the direction of propagation along the waveguide), E_{0i} is the incident electric field on the input facet of the waveguide, t_i and r_i are transmission and reflection coefficients, and $\gamma_i = \alpha_i + i\beta_i$ is the complex propagation constant. The wafer under consideration has [100] surface orientation. This fact combined with the symmetry of the nonlinear susceptibility tensor $\chi^{(2)}$ of GaAs yields second-harmonic polarization of the form $\mathbf{p}^{(2\omega)} \propto (2E_x E_y \hat{x} + E_x^2 \hat{y})$. The first component of polarization is along the polarization of TE_0, and the second term is polarized like the TM_0 mode. The TM_0-like component of second-harmonic polarization has y-directed polarization and hence cannot contribute to the radiation of second harmonic coming out of the top surface of the waveguide. However, the TE_0-like polarization of the second harmonic can radiate out of the top surface. If we keep only the sum frequency term and using Eq. (1), we obtain

$$p^{(2\omega)} \propto Re(E_1 E_2)$$

$$= Re\left[e^{2i\omega t}(A_1 A_2 e^{-2\bar{\gamma}z} + B_1 B_2 e^{+2\bar{\gamma}(z-2L)})\right]$$

$$+ Re\left[e^{2i\omega t}(A_1 B_2 e^{-2\bar{\gamma}L}) \right.$$

$$\left. \times \left(e^{-\Delta\gamma(z-L)} + \frac{B_1 A_2}{A_1 B_2} e^{\Delta\gamma(z-L)}\right)\right], \quad (2)$$

where $\bar{\gamma} = \frac{1}{2}(\gamma_1 + \gamma_2) = \bar{\alpha} + i\bar{\beta}$, $\Delta\gamma = \gamma_1 - \gamma_2 = \Delta\alpha + i\Delta\beta$. The term of the form $e^{2i(\omega t = \bar{\gamma}z)}$ cannot radiate perpendicular to z direction and will be ignored. Also from Eq. (1), $B_1 A_2/A_1 B_2 = r_1/r_2$. Using this in Eq. (2), we obtain

$$p^{(2\omega)} \sim Re\left[e^{2i\omega t}(A_1 B_2 e^{-2\bar{\gamma}L})\left(e^{-\Delta\gamma(z-L)} + \frac{r_1}{r_2} e^{\Delta\gamma(z-L)}\right)\right].$$

$$(3)$$

The time-averaged second-harmonic radiation is then

$$\langle P^{(2\omega)}\rangle_t \propto \langle p^{(2\omega)}p^{*(2\omega)}\rangle$$

$$\propto |A_1 B_2|^2 e^{-4\bar{\alpha}L}$$

$$\times \left[\frac{1}{2}\left(a + \frac{1}{a}\right) + \cos[2\Delta\beta(z-L)]\right], \quad (4)$$

where $a = (r_1/r_2)e^{2\Delta\alpha(z-L)} = 1 - \Delta a$. In terms of Δa, Eq. (4) reduces to

$$\langle P^{(2\omega)}\rangle_t \propto |A_1 B_2|^2 e^{-4\bar{\alpha}L}\left(1 + \frac{(\Delta a)^2}{2} + O(\Delta a)^3\right.$$

$$\left. + \cos[2\Delta\beta(z-L)]\right). \quad (5)$$

For low-loss birefringence ($|2\Delta\alpha L| = |2(\alpha_1 - \alpha_2)L| \ll 1$), Δa becomes $1 - r_1/r_2$, which is less than 0.2 for usual waveguides.[4] Therefore, we can safely ignore Δa in Eq. (5). Figure 3 shows the theoretical variation of $\langle P^{(2\omega)}\rangle_t$ as a function of length across the waveguide. For the purpose of calculation, the actual waveguide is approximated by a slab waveguide shown in the inset of Fig. 3. As can be seen, the theory agrees well with the experimentally observed second-harmonic variation of Fig. 2. The most uncertain parameter of the waveguide is the aluminum mole fraction of the cladding layer, which was approximately 20%. Then according to Eq. (5), the birefringence at $\lambda = 1.0642\ \mu m$ is given by

$$\Delta\beta = \beta_{TE_0} - \beta_{TM_0} = \pi/\Lambda = 50 \text{ rad/mm}, \quad (6)$$

where $\Lambda \approx 63\ \mu m$ is the period of the variation of the second harmonic signal observed. The experimentally observed $\Lambda \approx 63\ \mu m$ agrees well with the theoretical prediction of $\Lambda \sim 60\ \mu m$ (Fig. 3). There are several points that need further discussion. For small Δa (that is $r_1/r_2 \approx 1$), the second-harmonic polarization has a standing-wave profile, varying sinusoidally in z along the direction of the waveguide. This standing-wave polarization acts as a source for the radiation of second-harmonic signal. Assuming that the waveguide is charge free, the polarization profile generates an electric field along its direction, $\mathbf{E} = -4\pi\mathbf{P}$. Since the lens of the camera in our experiment maps the near-field pattern of the electric field across the waveguide to the image plane on the film negative, the resulting picture of Fig. 2 registers the time-averaged square of the near-field pattern or the time-averaged square of polarization profile across the waveguide. This point shows the relation of Eq. (5) to Fig. 2.

Another point to consider is the coherence of the second harmonic signal (SHS). The far-field pattern is approximately proportional to the Fourier transform of the near-field pattern. With the use of the near-field pattern discussed above, the far-field pattern is calculated to have intensity peaks at angles $|\theta| = \lambda_{2\omega}/2\Lambda \ll 1$ rad, where the angle is measured in the y-z plane (Fig. 1) with the y axis referring to $\theta = 0$. This prediction of the theory can account for the fact that when the lenses L_3 and L_4 in Fig. 1 were tilted from the direction perpendicular to the sample surface in the y-z plane, the intensity of SHS would sharply decrease. This observation confirms the coherent nature of the second-harmonic signal.
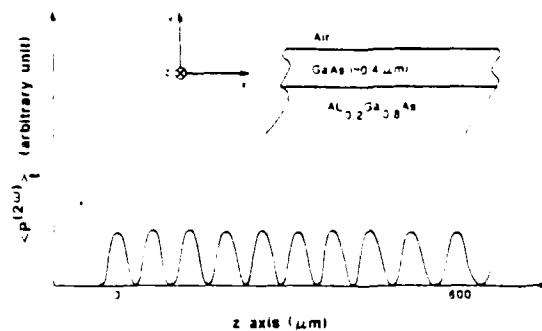

FIG. 3. Theoretical variation of $\langle P^{(2\omega)}\rangle_t$ along the waveguide length. The actual waveguide is approximated by an equivalent slab waveguide shown in the inset of the figure. For the parameters used, we obtain $\Lambda \sim 60\ \mu m$.
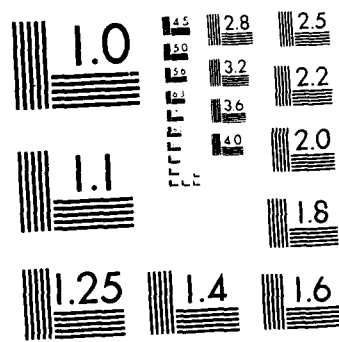
The third point to consider is the phase-matching condition. A standing wave can be decomposed into a forward and a backward propagating wave with wave vector K and $-K$. Therefore, the net wave vector of the standing wave is zero. Furthermore, if we express the fundamental electric field profile [Eq. (1)] in terms of a sum of a standing wave and a propagating wave and following the steps leading to Eq. (5), we find that the second-harmonic radiation perpendicular to the waveguide surface is only the result of interaction of the standing-wave component of TE_0 and the standing-wave component of TM_0 modes. Also, the near-field pattern of the SHS is described by a standing wave. This means that the sum of the initial interacting fundamental wave vectors is equal to the sum of the final second-harmonic wave vectors (which equals zero). It should also be noted that since the SHS is not generated collinear to the fundamental beam, the concept of phase matching is not applicable in its conventional sense. For noncollinear second-harmonic generation, the condition for the highest possible efficiency is the conservation of total wave vectors.

In summary, a novel second-harmonic generation technique is theoretically and experimentally analyzed. The second-harmonic signal obeys wave vector conservation. is coherent. and propagates perpendicular to the fundamental beam. This perpendicular propagation makes this technique useful for the study of waveguide parameters. The experiment reported yielded the modal phase birefringence $\Delta\beta = \beta_{TE_0} - \beta_{TM_0}$ accurately, directly, and unambiguously in the GaAs/AlGaAs waveguide structure under consideration. The knowledge of the modal phase birefringence is crucial in design of $TE \leftrightarrow TM$ mode convertors, wavelength filters, and $TE \leftrightarrow TM$ optical modulators. These devices are important part of wavelength multiplexing systems and other future integrated opto-electronics circuits.

[1] F. K. Reinhart. R. A. Logan. and W.R. Sinclair. IEEE J. Quantum Electron. QE-18. 763 (1982).
[2] R. C. Alferness. Appl. Phys. Lett. 36, 513 (1980).
[3] M. Masuda and J. Koyama, Appl. Opt. 16, 2994 (1977).
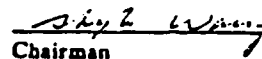[4] T. Ikegami. IEEE J. Quantum Electron. QE-8. 470 (1972).

Topics on GaAs Integrated Circuit:

GaAs Grown on Si Substrates, Field-Effect Transistors,

and Electro-Optic Probing

Yu-hwa Lo

Ph.D.

Department of Electrical
Engineering and
Computer Sciences

Sky L Wang

Chairman

Abstract

This thesis deals with several crucial topics on GaAs integrated circuits. To begin with, we consider the possibility to integrate GaAs devices with Si devices by means of heteroepitaxy growth on Si. The GaAs films are grown on Si substrates by molecular beam epitaxy. The material defects and thermal induced stress in GaAs films are two fundamental problems of particular interests. The defects are mainly due to the inclined interface dislocations and stacking faults. We find that these defects can be reduced if GaAs is grown on a clean and double-stepped Si surface. The stress generated from different thermal expansion coefficients between GaAs and Si can cause film cracking. Our analysis indicates that selective growth is an effective way to reduce the stress and consequently, it could be a solution to the problem of film cracking.

Various kinds of field-effect transistors (FET's) are demonstrated and analyzed. The buried-gate junction FET made by a submicron self-aligned process achieves a transconductance of 180mS/mm. Besides, no back-gating effect is observed in this kind of device. Another novel device named as top-back-gate FET is also reported. One interesting characteristics of this four

terminal device is its tunable transconductance and impedance. This device can be used as a NAND gate logic unit because it can be turned off by either the top-gate or the back-gate.

In order to characterize the GaAs material and device in a non-invasive way, we develop the electro-optic probing technique. Because the refractive indices of GaAs are modified by electric fields, we can obtain the information about internal fields in GaAs device or material from the phase retardation of a probing beam. The experimental results in potential profile probing of various device structures are reported. The problem of three-dimensional internal-field probing is analyzed and a simple experiment is done to demonstrate the feasibility. In addition to probing low frequency characteristics, microwave signal can also be probed if the probing beam is replaced by a short optical pulse train. The standing wave pattern along a coplanar waveguide is measured at 8GHz by this technique. The measured effective index in microwave frequency is very consistent with the value measured by network analyzer. However, the detection bandwidth of electro-optic probing can be over 100GHz if subpicosecond pulsewidth and low timing jitter light source is used.

Approved for public release;
distribution unlimited.

END

DATED

FILM

8—88

DTIC