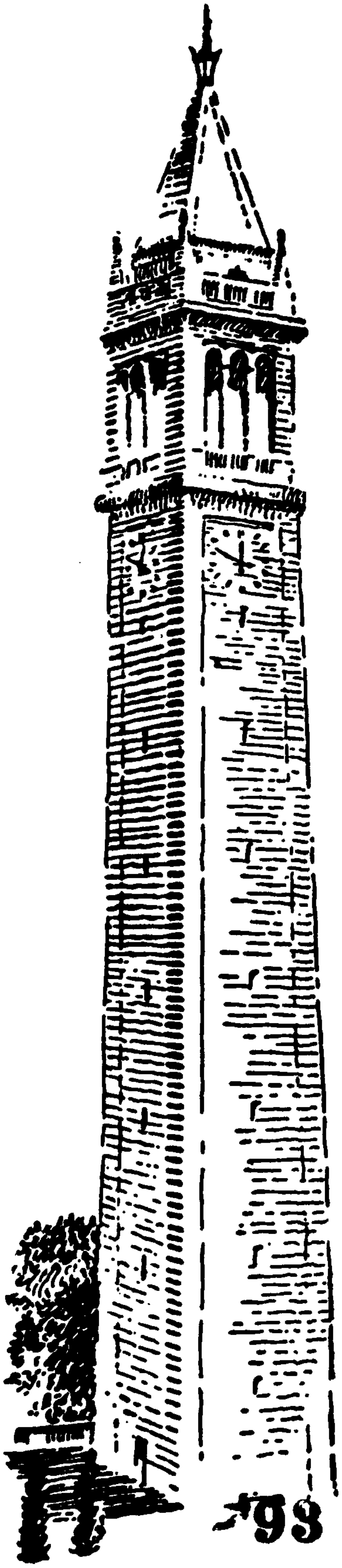


Report No. UCB/ERL 90/3

JOINT SERVICES ELECTRONICS PROGRAM

ANNUAL PROGRESS REPORT
(Contract F49620-90-C-0029)
(1 March 1992 — 28 February 1993)

Chenming Hu and Michael A. Lieberman

5 March 1993

Prepared for:

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

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
March 5, 1993

3. REPORT TYPE AND DATES COVERED
Annual 1 MAR 92 - 2 FEB 93

4. TITLE AND SUBTITLE
Annual Report No. UCB/ERL 90/3

5. FUNDING NUMBERS
F49620-90-C-0029
Project/Task: 2305/A9
Program Element: 61102F

6. AUTHOR(S)
Chenming Hu and Michael Lieberman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Electronics Research Laboratory
253 Cory Hall
University of California at Berekeley
Berkeley, California 94720

8. PERFORMING ORGANIZATION REPORT NUMBER
UCB/ERL 90/3

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
Building 410
Bolling AFB DC 20332-6448

Program Manager: Major Billy R. Smith, Jr.

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE,
DISTRIBUTION UNLIMITED

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This document is the 1992 annual report of research conducted in the Electronics Research Laboratory at University of California at Berkeley under the sponsorship of the Joint Services Electronics Program. The research covers the following project areas: Quantum Electronic Devices: (1-A)Nonlinear Optics in Compound Semiconductors(1-B)Ultrafast Optical Techniques(1-C)Optical Probing of Semiconductor Devices and Interfaces by Electro-Optic and Photo-Elastic Effects; Electronic Devices: (2-A).1 Micron BiCMOS Devices in Bulk and SOI Substrates(2-B)Conductive Oxides and Ferroelectrics for Programmable Devices and(2-C)Insulated-Gate GaAs Field Effect Transistors; Neural Networks and Parallel Computation:(3-A)Stochastic Neural Networks and Application to Signal Processing (3-B) Learning and Generalization by Neural Networks (3-C) Reconfigurable Analog Elements for Neural Nets and (3-D) Architectural Issues in Parallel Computation. There is also a section concerning significant accomplishments. Two supplemental reports are included.

An Appendix of all publications, papers and presentations forms a second part of this report. It is bound under separate cover.

14. SUBJECT TERMS

15. NUMBER OF PAGES

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**

TABLE OF CONTENTS

Page

PART A -- DIRECTOR'S OVERVIEW. 1

PART B -- SIGNIFICANT ACCOMPLISHMENTS 3

AlN$_x$ as Both Insulated-Gate and Semiconductor Material. 3
The CNN Universal Machine 7
Fastest Silicon Transistors 10
Flash EPROM for Analog Computation 14

PART C -- INDIVIDUAL WORK UNITS 17

THEME I -- Quantum Electronic Devices

I-A. Nonlinear Optics in Compound Semiconductors. 17

I-B. Ultrafast Optical Techniques. 18

I-C. Optical Probing of Semiconductor Devices and Interfaces by Electro-Optic and Photo-Elastic Effects 19

SPL.1. Ultrahigh Speed Semiconductor Lasers 20

THEME II -- Electronic Devices

II-A. 0.1 µm BiCMOS Devices in Bulk and SOI Substrates. 22

II-B. Conductive Oxides and Ferroelectrics for Programmable Devices 25

II-C. Insulated-Gate GaAs Field Effect Transistors 27

THEME III-- Neural and Parallel Computation of Neural Networks

III-A. Stochastic Neural Networks and Application to Signal Processing 34

III-B. Learning and Generalization by Neural Networks 42

III-C. Reconfigurable Analog Elements for Neural Nets. 45

III-D. Architectural Issues in Parallel Computation 47

Heterogeneous Architectures for Artificial Neural Networks 47
Fault Tolerance in Layered Artificial Neural Networks 48
Interconnection Network Design Based on Packaging Considerations. 49

SPL-II. Analog Neural Networks for Vision Tasks 52

PART A - DIRECTOR'S OVERVIEW

JSEP continues to play an important and unique role in electronics research at the University of California, Berkeley. Its emphasis on science and relatively stable funding provides an increasingly rare environment for conducting the more basic research and exploring promising new areas. It also provides a unique opportunity for encouraging collaborative research involving multiple principal investigators or new faculty members. Currently, JSEP at U.C. Berkeley partially supports the research of 11 faculty and 22 graduate students.

A long time JSEP investigator, Professor Shyh Wang passed away in April, 1992. Professor Wang was widely recognized as a leader in the field of semiconductor lasers, guided-wave optics and non-linear optics. His two authoritative textbooks on semiconductor physics and devices are widely used and many of his students have become leaders in their fields. The Berkeley JSEP program will miss Shyh for his research contributions and his leadership.

Professor Steve Smith and Professor Ken Gustafson have agreed to supervise the two JSEP students who formerly worked with Professor Wang. Professor Gustafson has also stepped up to serve as the faculty investigator of Work Unit 1-A, Nonlinear Optics in Compound Semiconductor. Professor Kam Lau became a JSEP faculty investigator in April 1992. His work unit is SPL. 1, Ultra High Speed Semiconductor Lasers. Similarly, Professor Leon Chua's work unit, SPL-II Analog Neural Networks for Vision Tasks, also started in April 1992. The annual reports for these two work units, therefore, cover the performance from April through February 1993.

Using the Director's limited discretion, JSEP helped to attract our newest faculty member, Jeff Bokor, by inviting him to become a JSEP investigator immediately on an interim basis and participate in the renewal proposal for continued funding from 1993 onward. This is an example of the unique nature and responsiveness of JSEP. Jeff has been a manager of research at AT&T Bell Laboratory in the areas of quantum and optical electronics and, more recently, VLSI technology and systems prior to joining U.C. B. in November as a full professor. His broad experience in basic science research and research management skills makes him a valuable investigator and a leader of the Berkeley JSEP program.

This annual report covers 12 work units in three theme areas. Theme I, Quantum Electronics Devices, involves research in enhanced nonlinear optical conversion in quantum wells, surface emitting lasers, selective contact technology using shadowed MBE, optical probing of misfit and thermal stress in heterostructures, and a fundamental understanding of the role of carrier transport in the high speed modulation of quantum well lasers. In the surface emitting laser project (Work Unit I-B), uniform MBE growth over the entire wafer was achieved by phase lock epitaxy with periodic rotation. Surface emitting lasers had unparalleled wafer to wafer yield and achieved 13mW continuous output. Their turn-on voltage is 1.7 volt, the lowest reported.

Theme II, Electronic Devices, includes research into 0.1 μm BiCMOS devices, conductive oxides and ferroelectrics, and GaAs-insulator interfaces. We have reported the world's fastest silicon transistor (Work Unit II-A). The inverter delay is 13.5 ps at room temperature with only 1.5 V power supply. (In 13.5 ps, light travels only about 4 mm). This room temperature speed exceeds all previous MOS and bipolar transistor speeds and matches the previous world record set by IBM *at 77 K*. The state-of-the-art industry technology produces a transistor speed about 10 times slower than our record speed. Our result is the culmination of our research into silicon-on-insulator (SOI) devices. The transistors were fabricated in 800Å silicon film and have both higher current and lower capacitance than transistors of identical dimensions constructed in bulk substrate.

Theme III, Neural and Parallel Networks and Applications, addresses issues of signal processing, learning algorithms, reconfigurable devices for neural network, architectures, performance, and a cellular neural network (CNN) universal machine. For the last task (Work Unit SPL-2), the concept of a CNN universal machine was developed. The CNN universal machine is an analog array computer on which programs can be stored and executed, making it the analog equivalent to the digital microprocessor. Of course, the CNN universal machine performs operations on a whole array of data at once. Distributed analog memory and control hardware is used to store intermediate results, direct signal flow, and specify the CNN operation to be performed according to a global program. Many of the vision tasks performed by the retina have been converted to operations that can be represented by a program. A patent has been applied for this invention.

The Appendix, in a separate volume, contains copies of 27 published articles, 10 conference papers, 1 thesis, and 7 papers submitted for publication.

PART B - SIGNIFICANT ACCOMPLISHMENTS AND TECHNOLOGY TRANSITION

AlN_x as Both Insulated-Gate and Semiconductor Material

Professors N. Cheung, C. Hu, and W.G. Oldham with J. Chan

Aluminum nitride has found widespread application due to its direct wide bandgap (6.2eV), good thermal conductivity, and high temperature stability. In particular, it has been investigated as a surface passivation material for GaAs-based IC's, as an insulated-gate material for GaAs-based metal-insulator-semiconductor FETs (MISFET), and as a potential semiconductor material for UV light sources and detectors. In this work, we have extensively studied various aspects of sputtered AlN_x thin films via numerous characterization techniques.

A ceramic resistive heater was employed to heat the substrates to temperatures up to 700°C during growth. X-ray diffraction measurements showed that a preferred orientation along the basal plane was achieved at a minimum temperature of 450°C for AlN_x films on both types of Si substrates and sapphire (Figure 1).

Optical absorption measurements were carried out on the AlN_x on sapphire samples using mercury and deuterium lamps with a Perkin-Elmer monochromater. The absorption spectrum (in the form of absorption versus energy gap) was calculated from which the energy bandgap was extrapolated. The bandgap values ranged from 5.2 to 5.9eV (Figure 2), which agreed with literature.

Current-voltage tests performed on AlN_x/Si MIS capacitors confirmed the insulating properties of the prepared AlN_x films. Electric breakdown fields ranged from $1.3x10^5$ to $7.9x10^5$ V/cm for all heated samples (450-900°C) with leakage current less than $100nA/cm^2$ (Figure 3).

Measurements carried out on MIS capacitors consisting of AlN_x thin film on silicon as well as MBE-grown GaAlAs substrates showed an electric breakdown field approaching $1x10^6$ V/cm; this indicated the potential application of AlN_x thin film as an insulated-gate material in III-V electronic devices. X-ray diffraction showed that even at a substrate temperature of only 450°C, AlN_x film oriented in the basal direction (0001) was achieved, which indicated the feasibility of low-temperature processing of this semiconductor material for future applications in the area of wide bandgap light sources and detectors.

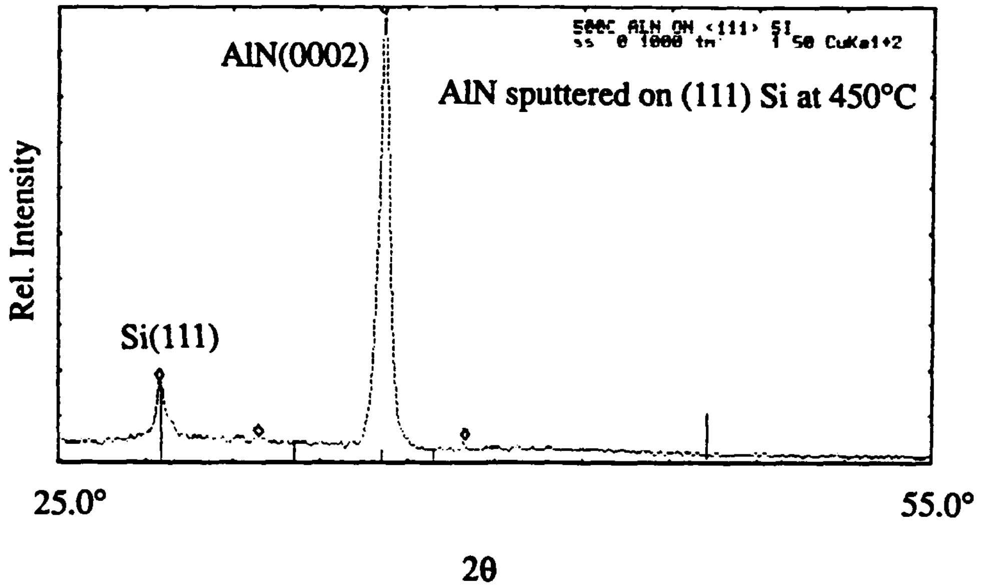


Figure 1: (0001) basal orientation of AlN_x on (111) Si sputtered at 450°C.

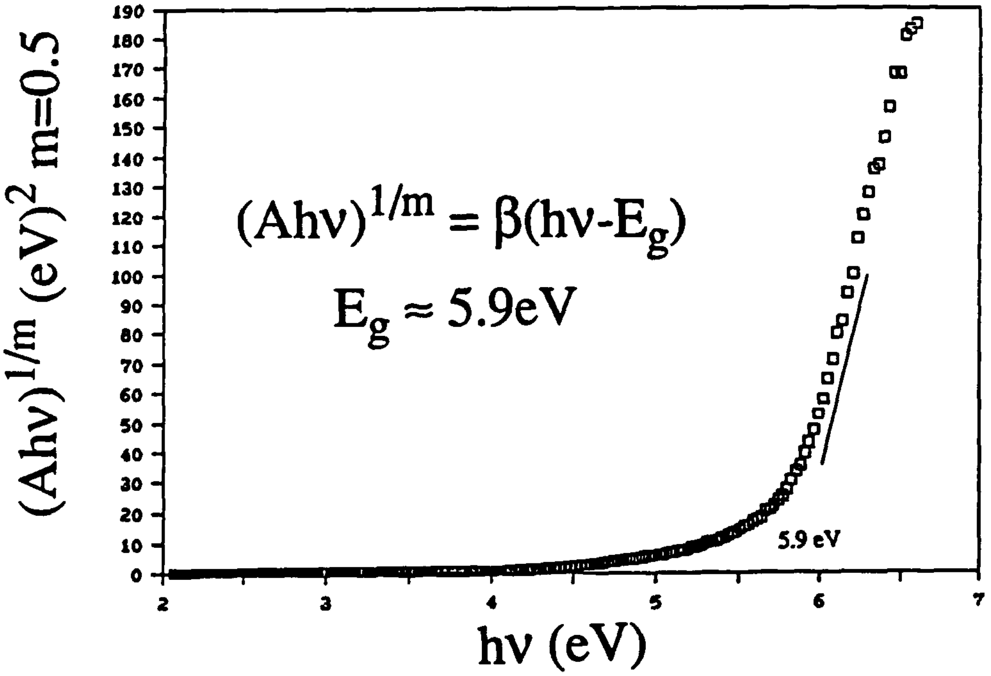


Figure 2: $(Ah\nu)^{1/m}$ vs. hν plot of AlN_x on sapphire, with extrapolated E_g = 5.9eV.

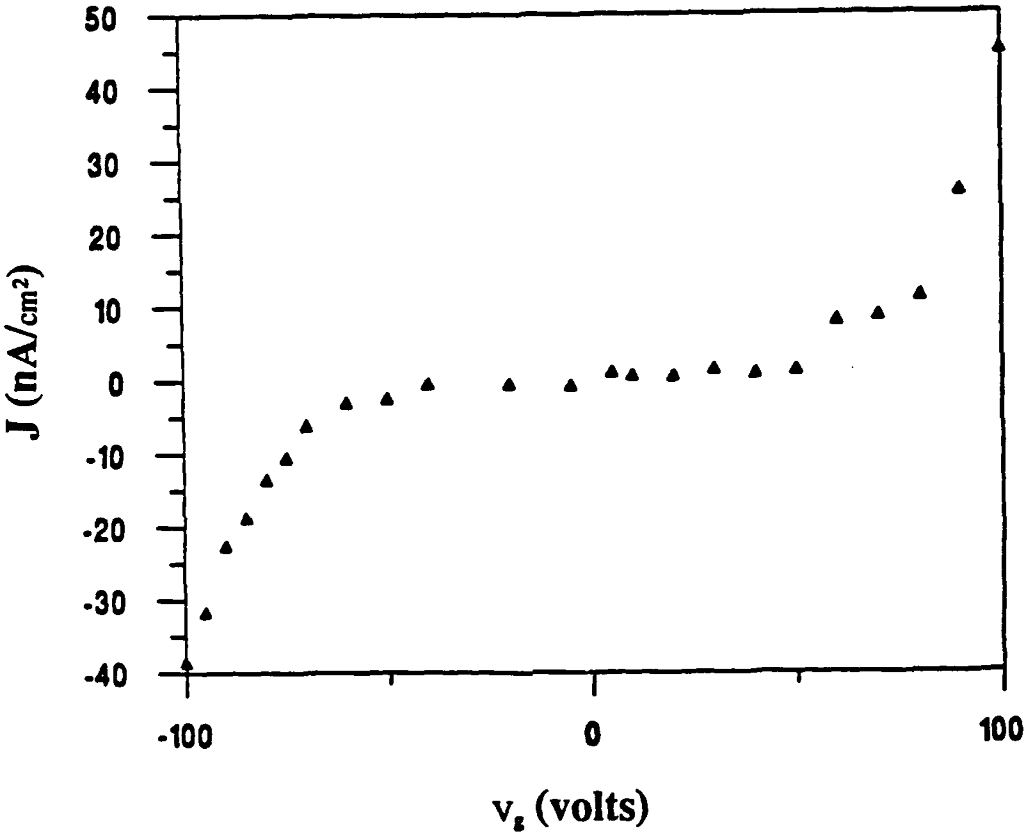
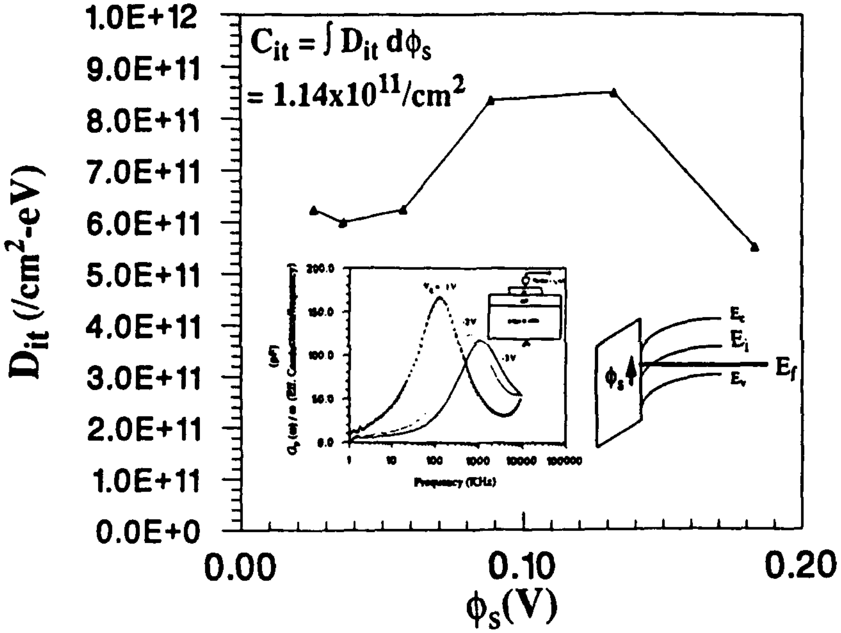
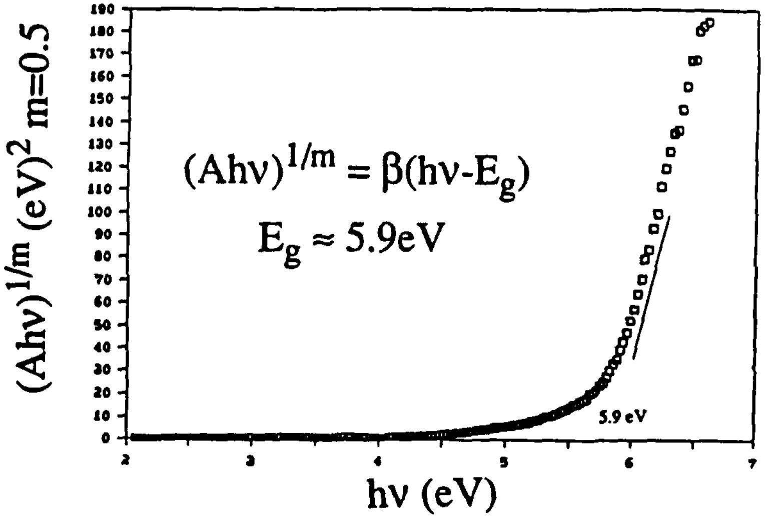


Figure 3: Current density vs. voltage characteristics of AlN_x/Si structure, with breakdown field approaching 10^6 V/cm, and leakage current level on the order of 100nA/cm^2.

AlN_x as both insulated-gate and semiconductor material





Properties of AlN_x:

- Insulator-semiconductor interface trap density on the order of 10^{11} cm^{-2}
- Extrapolated semiconductor bandgap of 5.9eV
- Preferred basal plane texture when grown at a substrate temperature of 450°C

The CNN Universal Machine

Professor L. O. Chua

Introduction

The development of the concept of a CNN Universal Machine[1] is a major breakthrough in our research on Cellular Neural Networks (CNN) during the past year. Elementary parts of this new architecture have already been designed and some critical implementation issues have been examined. In addition, some key applications have been identified and investigated.

The CNN Universal Machine is an analog array computer on which programs can be stored and executed, making it the analog neuromorphic equivalent to the digital microprocessor. As a dual to the Arithmetic Logic Unit of microprocessors, the CNN array is used to perform operations on a whole array of data at once. The architecture of the Universal Machine allows it to perform a sequence of different CNN operations on the same array. Distributed analog memory and control hardware is used to store intermediate results, direct signal flow, and specify the CNN operation to be performed in accordance with a global program.

The global stored-program concept is the key element of the CNN Universal Machine architecture. Many CNN templates have already been developed to perform interesting processing tasks. The programmability of the CNN Universal Machine allows these templates to be time multiplexed to implement complex and useful algorithms. This breakthrough replaces hundreds of dedicated CNN chips with a single programmable, real-time, VLSI chip with comparable significance to the invention of the microprocessor.

The CNN Universal Machine will have an enormous number of applications as an intelligent sensor or display driver as information acquisition, transmission, and presentation systems become widespread with the advent of high speed digital communication networks. In addition, the CNN Universal Machine will be an important tool in studying the behavior of biological and phvsical systems through the modeling of partial differential equations. A CNN Universal Machine Supercomputer is proposed for such research-oriented activities. We have applied for a patent for the CNN Universal Machine and Supercomputer.

Hardware Description of the CNN Universal Machine

The original Cellular Neural Network architecture introduced in 1988 is composed of an rectangular array of analog circuit processors called "cells"[2]. By changing the interconnections between these cells, different types of computations can be performed on the data stored on the array. The pattern of interconnections, referred to as a "template", is restricted to be purely local, enabling the CNN to be compactly implemented on a single chip using current analog VLSI technology.

This analog array architecture, along with some additional elementary logic and storage elements located at each cell, is the nucleus of the new CNN Universal Machine. In addition to the CNN array, the Universal Machine also contains global processing and control units which control the CNN array and reconfigure it to execute each of the instructions in the stored program. Because of the unique combination of both digital and analog processing provided by the CNN Universal Machine, we refer to it as an "analogic" processor, combining the two terms "analog" and "logical".

Due to the large amcunt of data which can be processed at one time, reading data in and out of the CNN Universal Machine could be a significant bottleneck in the processing. However, there are several ways to overcome this problem. High input rates can be achieved by fabricating sensors directly on the CNN Universal Machine chip. This allows simultaneous parallel input to all of the cells. In addition, it is also feasible to build CNN Universal Machine chips which can be surface mounted directly onto large scale sensor arrays. In some cases, the data from the CNN Universal Machine can also be output in parallel. For instance, multiple CNN Universal Machine chips could be directly mounted onto flat panel displays to perform image processing or decoding tasks. In other cases, the data will be read out serially. In this case, a sequence of CNN templates can be used to perform analog-to-digital conversion inside the CNN Universal Machine.

Applications

The CNN Universal Machine is universal in a very broad sense. In fact, we have shown it to be as universal as the Turing machine. Thus, practically any conceivable analogic model can be solved on the CNN Universal Machine. In particular, many useful image and array processing algorithms have been identified and implemented as CNN templates[3,4].

The invention of this new architecture was heavily inspired by biological systems. As a result, the CNN Universal Machine can serve as prototype modeling tool for a broad class of biological and cognitive processes. In fact, the CNN Universal Machine architecture has a profound correspondence with the biological retina. Many of the early vision tasks performed by the retina have already been converted into a sequence of CNN templates which can be represented by a program.

Many CNN Universal Machine chips can be used in an array to form the CNN array supercomputer. This supercomputer can be used for solving a very broad class of nonlinear partial differential equations, including reaction-diffusion equations, autowaves, spiral waves, and solitons. These simulations can be performed at a fraction of the time or cost required by existing supercomputers. Therefore, computationally intensive problems such as calculating stress distribution in mechanical structures can be solved efficiently.

The algorithms implementable by sequences of CNN templates represent a new world of analogic software. The CNN algorithms can be described by an analogic CNN language. This language is translated by a compiler to the analogic CNN machine code. Indeed, some elementary "subroutines" of simple CNN templates have already emerged. Thus, the software base is already developing for applications of the CNN Universal Machine.

References

[1] T. Roska and L. O. Chua, "The CNN Universal Machine: An Analogic Array Computer," *IEEE Transactions on Circuits and Systems--II: Analog and Digital Signal Processing,* Mar. 1993, to appear.

[2] L. O. Chua and L. Yang, "Cellular Neural Networks: Theory and Applications," *IEEE Transactions on Circuits and Systems,* Vol. 35, Oct. 1988.

[3] *Proceedings of the 1990 IEEE International Workshop on Cellular Neural Networks and Their Applications,* Budapest, Hungary, Dec. 16-19, 1990.

[4] *Proceedings of the 1992 IEEE Second International Workshop on Cellular Neural Networks and Their Applications,* Munich, Germany, Oct. 14-16, 1992.

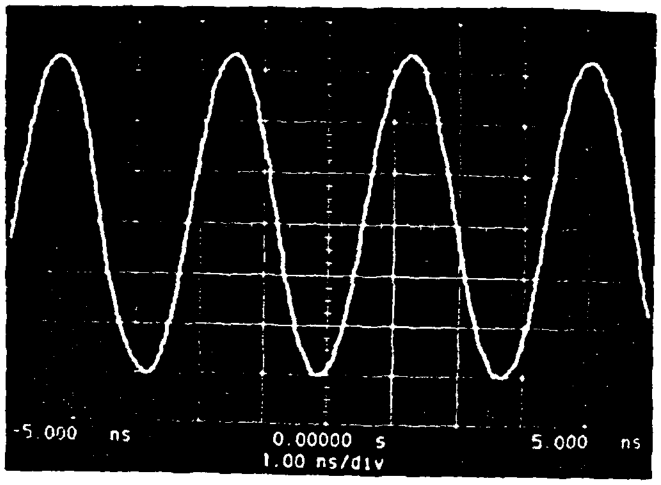
Fastest Silicon Transistors

Professors C. Hu and P.K. Ko with F. Assaderaghi, J. Chen, M.S. Chan, S. Parke and Z.H. Liu

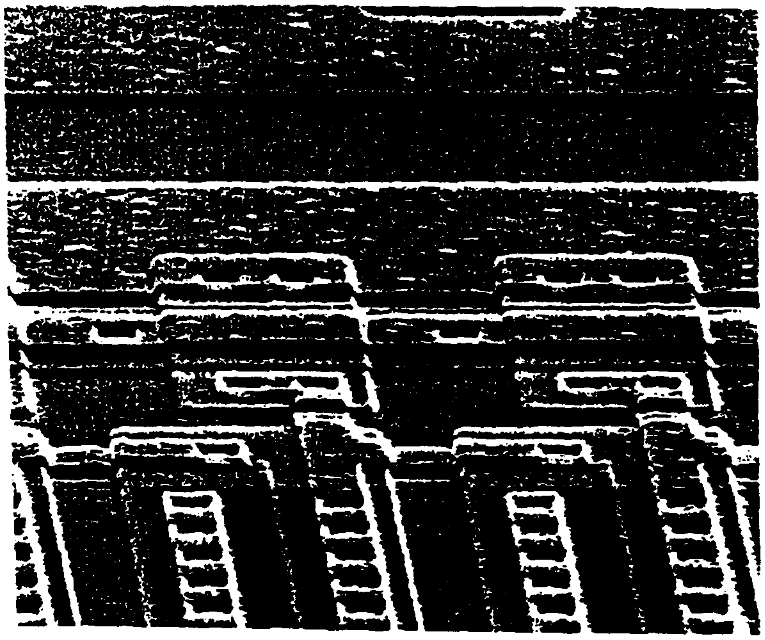
We have reported the world's fastest silicon transistors.[8] The speed was measured with 101 stage inverter ring oscillators.

The delay per stage is 13.5 ps at room temperature at only 1.5V power supply. (Figure 1(a),(b)). This MOSFET speed is the fastest silicon transistor speed for both MOSFET and bipolar transistors. The previous speed record was set in 1988 by IBM with 0.1 μm MOSFET's, which had 13.5 ps speed at 77 °K. Our transistor can achieve this speed at room temperature and at lower voltage. State of the art 0.5 μm MOSFET has 100 ps speed at 5V and 500 ps speed at 1.5V.

This result is the culmination of our research into silicon-on-insulator (SOI) devices. SOI MOSFET's were fabricated in 800 Å silicon film over 4,000 Å buries oxide produced by oxygen implantation. Junction capacitances are all but eliminated, leading to faster transistor speed. We have also optimized the SOI MOSFET's through a thorough understanding of the device physics and therefore achieved much faster speed than previous SOI research did. (Until now, SOI has not produced faster transistors than bulk technology.) We realized that high performance depletion mode n-channel transistors can be made in SOI while they cannot be made in bulk substrate because of punchthrough. We used a photoresist-ashing technique to make 0.2 μm gates. The 75 Å thick gate oxide is conservative for the 1.5V operation.



(a) Ring oscillator output waveform at VDD=1.5V



(b) SEM photo of ring oscillator

Figure 1. Propagation delay of a 101 stage unloaded ring oscillator is 14 ps/stage at VDD=1.5V. The SOI film thickness is 500A. Tox=70A and Leff=0.17 um.

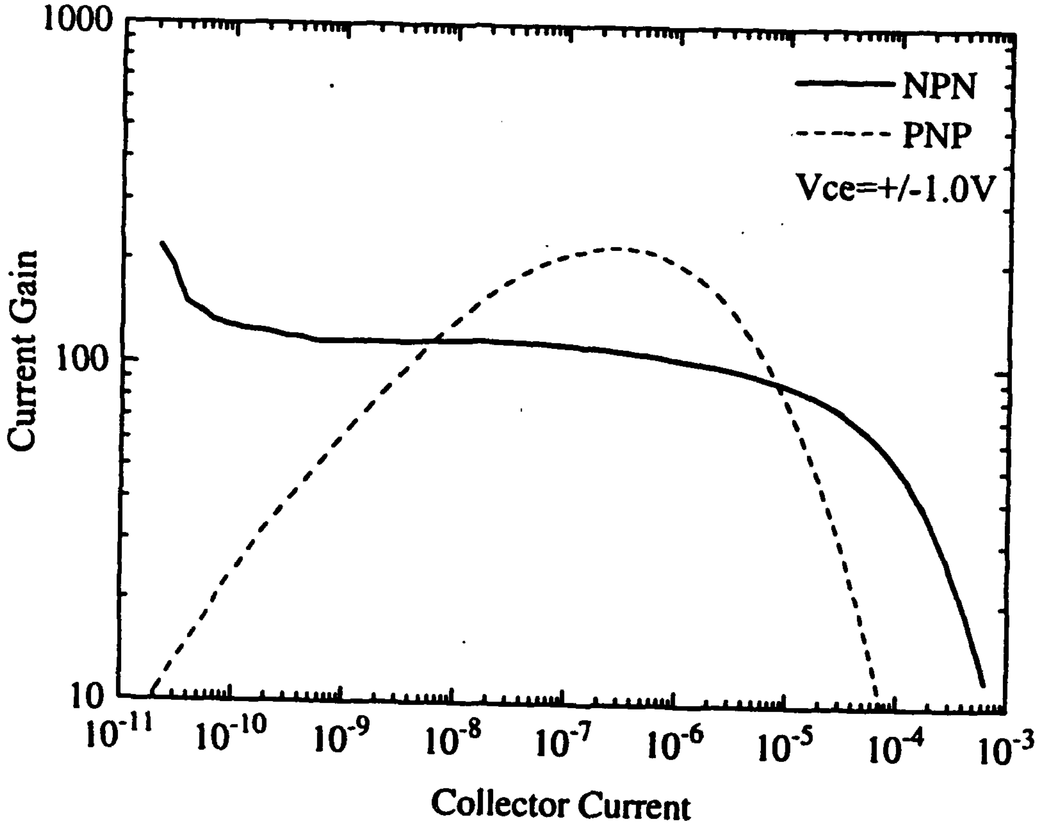
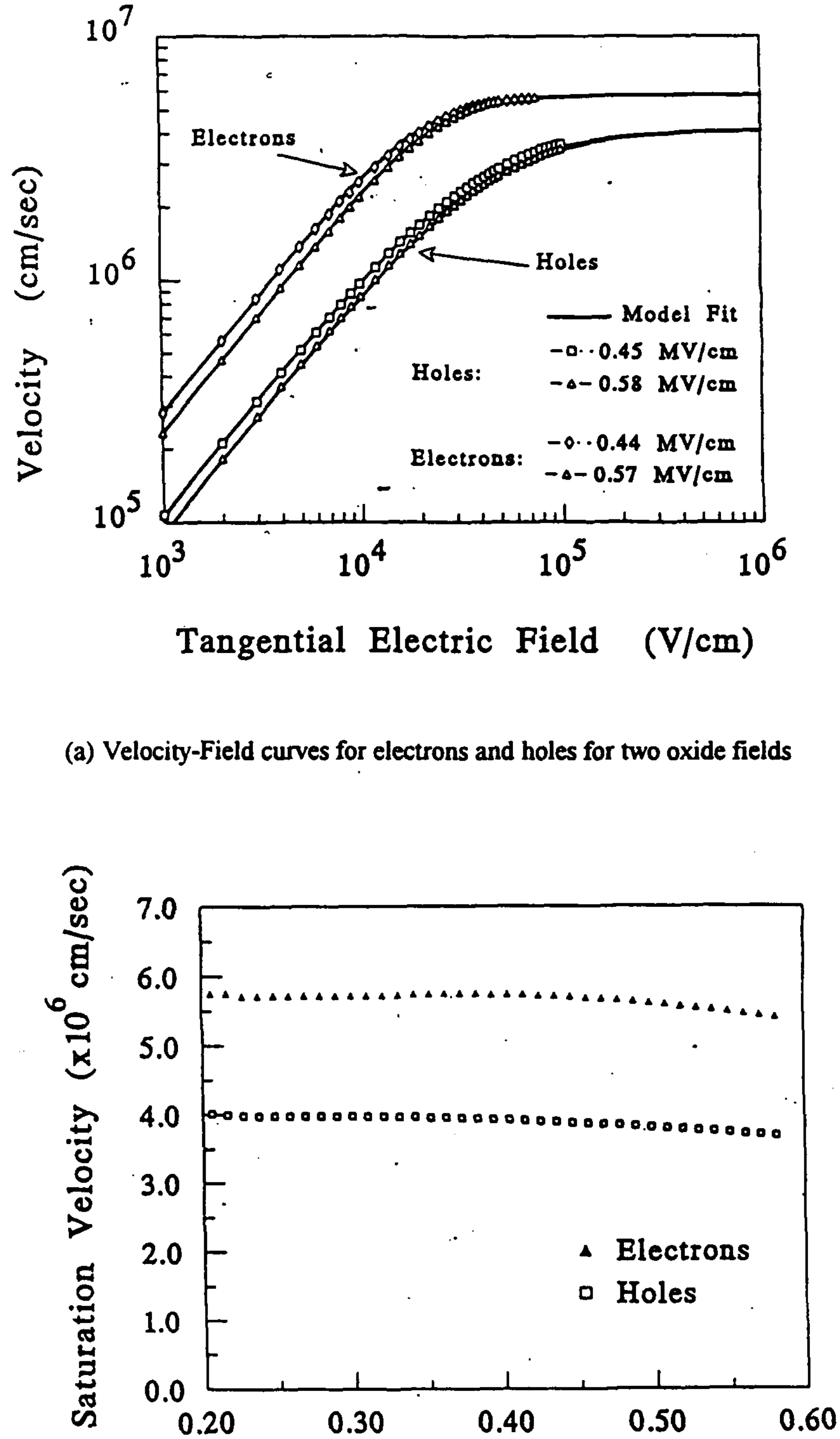


Fig.2 Current gain of NPN and PNP lateral bipolar junction transistors on SOI. Maximum current gains of 120 and 225 for the NPN and PNP respectively, are the highest values that have been reported.



(a) Velocity-Field curves for electrons and holes for two oxide fields

b) Saturation velocity as a function of oxide field for electrons and holes showing very weak dependence

Fig. 3 Velocity-Field curves for electrons and holes in the Si-SiO2 inversion layer as measured from the back-gate MOSFET that functioned as a uniform-channel-field device.

Flash EPROM for Analog Computation

Professor Ping K. Ko with Alan Kramer, Larry Hung and C.K. Sin

Flash devices present an interesting problem for analog programming in comparison to EEPROM devices because, since flash arrays are typically bulk-erase, the cost of overshooting the desired threshold during programming is very high, as to lower the threshold will necessitate erasing the entire array. For the reason of this asymmetry, the analog programming algorithm that was developed earlier in this project for analog programming of EEPROMs (which program in a symmetrical manner) is unusable for Flash-EEPROMs. We have developed and successfully tested (on the before-mentioned wafers) a new programming algorithm for analog threshold adjustment of Flash EEPROMs. This new algorithm utilizes the fact that, based on the physics of EPROM cell programming, the rate of threshold voltage change with programming time is monotonically decreasing. The threshold voltage of the cell is sampled after each programming step and the programming time(and/or voltage) for the next step is determined by a linear approximation to the calibrated programming curve of the Flash device in a way that allows the device to be programmed to an analog value without overshoot(Figure 1). Eight-bit values have been successfully programmed.

In addition, a novel architecture for massively parallel analog computation is invented which uses an EEPROM as the essential computational building block. The proposed architecture is similar in appearance to a memory array, in that the chip stores a set of d-dimensional memories, one per row(Figure 2). In contrast to standard digital memory, each row in the proposed chip is capable of storing in analog either a point, a hypersphere, or a hyperrectangle in d-dimensional space. The other important difference between this architecture and a standard memory chip is that a substantial amount of computation can be performed in parallel directly on all the stored memories. In particular the chip computes geometric relationships based on Euclidean distance between the stored memories and a new d- dimensional query point which allow it to access the stored memories in order of proximity to the query point. The speed [ms settling time] and density [1E6 elements per chip, i.e. 1E6 d-dimensional vectors] of the proposed architecture promise to make it a powerful engine for real-world computational tasks such as associative memory, pattern classification and function approximation.

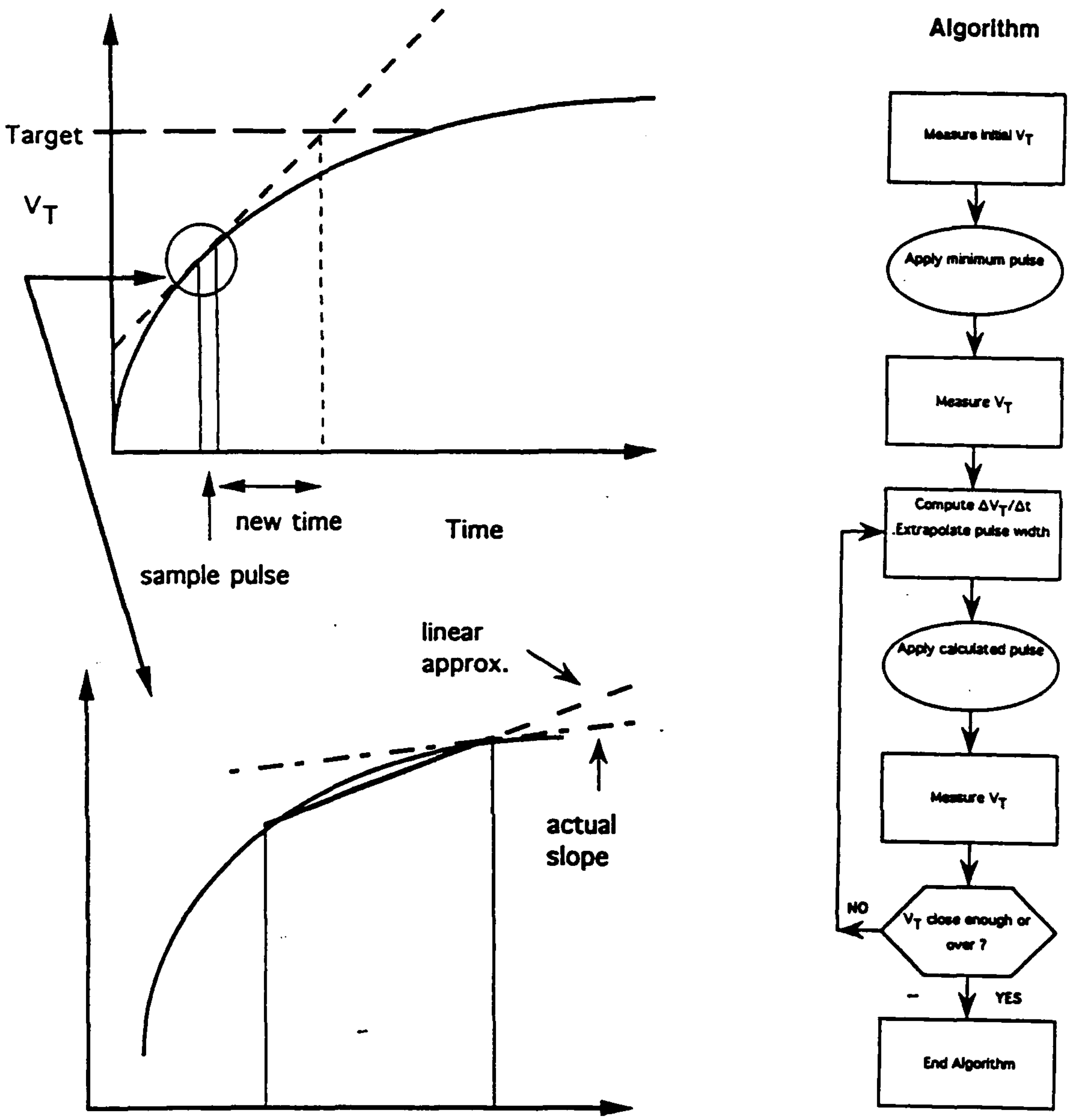
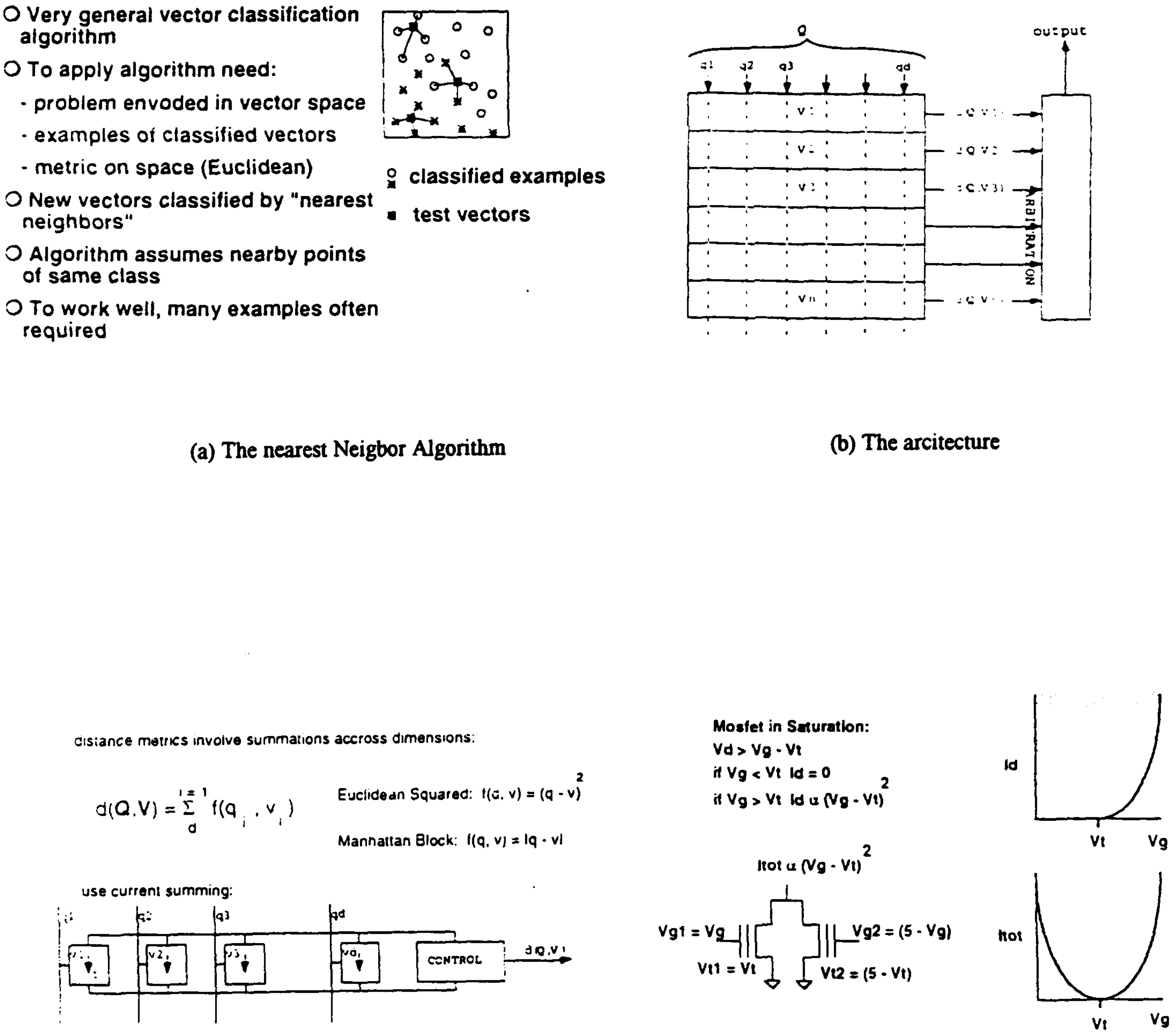


Figure 1. Analog programming algorithm for Flash EPROM.



- ○ Very general vector classification algorithm
- ○ To apply algorithm need:
  - problem envoded in vector space
  - examples of classified vectors
  - metric on space (Euclidean)
- ○ New vectors classified by "nearest neighbors"
- ○ Algorithm assumes nearby points of same class
- ○ To work well, many examples often required

(a) The nearest Neigbor Algorithm

(b) The arcitecture

distance metrics involve summations accross dimensions:

$$d(Q,V) = \sum_{d}^{i=1} f(q_i, v_i)$$

Euclidean Squared: $f(q, v) = (q - v)^2$

Mannattan Block: $f(q, v) = |q - v|$

use current summing:

Mosfet in Saturation:
Vd > Vg - Vt
if Vg < Vt Id = 0
if Vg > Vt Id α $(Vg - Vt)^2$

Itot α $(Vg - Vt)^2$

(c) Distance computation

Figure 2. A new architecture for massively parallel analog compuation based on a Flash EPROM memory array.

PART C - INDIVIDUAL WORK UNITS

I-A. Nonlinear Optics in Compound Semiconductors

Professor T.K. Gustafson with Patrick Harshman

Vakhshoori at AT&T, has recently pointed out that periodic heterostructures can be used to quasi-phase-match the second harmonic generation in the perpendicular direction, thereby potentially enhancing the second harmonic intensity by orders of magnitude. In addition he has proposed the utilization of active wave-guides to circumvent the loss problem.

We (Harshman and Wang) [1] have investigated, as another possible approach to overcoming both the problem of phase-matching the second-harmonic generation and the small magnitude of the nonlinearity, the use of quantum transitions between confined states in asymmetric quantum wells. Theoretical calculations [1] indicate that the magnitude of the resonantly-enhanced nonlinearity of an appropriately designed quantum well structure using an interband transition is only comparable to that of bulk GaAs. However, the nonlinearity can also exhibit a sign reversal when the asymmetry of the well is reversed. This ability to change the sign of the quantum well nonlinearity implies that asymmetric quantum wells in the guiding region can also be used to quasi-phase-match the nonlinear conversion. The second-harmonic power radiated from the surface of the multiquantum well waveguide has been calculated to be over a factor of 100 times greater than the power radiated from a bulk AlGaAs guide.

We have also begun to consider the use of wide band-gap materials such as GaN as a nonlinear material. Such wide band-gap materials offer the possibility of using intersub-band transitions for the nonlinearity, which can be much higher than the bulk nonlinearity.

JSEP Publications

[1] P.J. Harshman and S. Wang, "Asymmetric AlGaAs Quantum Wells for Second-Harmonic Generation and Quasi-Phase Matching of Visible Light in Surface Emitting Waveguides," *Appl. Phys. Lett.*, Vol. 60, pp. 1277, 1992.

[2] P.L. Kelley, P.J. Harshman, O. Blum, T.K. Gustafson, "Radiative Renormalization Analysis of Optical Double Resonance," *Technical Digest of the Nonlinear Optics*, Materials, Fundamentals and Applications Conference, Mauii, Hawaii, 1992 (August), paper, Tues 4. (Partially supported by JSEP).

I-B. Ultrafast Optical Techniques

Professor John Stephen Smith

Our work on surface emitting lasers will now continue under an AFOSR grant. Recent results include phase lock epitaxy growths with periodic rotation, resulting in growths which are uniform over the entire wafer. We have also obtained results on relative intensity noise in vertical emitters, and have demonstrated mode control using external cavities. Our vertical cavity surface emitting lasers include devices which produced maximum output powers of up to 13 mW cw. They also exhibit low series resistance, with turn on voltage of 1.7 volts, the lowest reported, and unparalleled wafer to wafer yield.

Our theoretical work on compensating group velocity dispersion in quantum well lasers has been completed and published in Applied Physics Letters. The conclusion was basically that the mode control necessary to achieve zero dispersion was possible, but that the parameters involved would be challenging to satisfy experimentally.

Our work on selectively contacted NIPI structures has resulted in many applications, including nipi based high-speed detectors and bistable switches with gain, nipi bandfilling modulators, and tunable electroabsorption modulators,

We have demonstrated AlGaAs/GaAs optical hetero-nipi modulators based on bandfilling effect with inter-digital selective contacts with five orders of magnitude change in I-V characteristics from forward to reverse junction bias. This indicates the high quality of the inter-digital contacts. The measured reflectance spectra show a change of the absorption constant of about $7850 cm^{-1}$ with applied bias voltages as low as 1.5V.

In a laterally contacted AlGaAs/GaAs hetero-nipi doping superlattice the internal electric fields can be tuned to control the depletion or the confinement of the carriers in the quantum wells. The difficulty to obtain high quality inter-digital lateral contacts had previously prohibited a realization of this bandfilling modulator.

JSEP Publications

[1] S.P. Dijaili, J.M. Wiesenfeld, G. Raybon, C.A. Burrus, A. Dienes, J.S. Smith, and J.R. Whinnery, "Cross-Phase Modulation in a Semiconductor Laser Amplifier Determined by a Dispersive Technique," *IEEE Journal of Quantum Electronics*, Vol. 28, No. 1, pp. 141-150, January 1992.

[2] D.M. Kuchta, J. Gamelin, J.D. Walker, J. Lin, K.Y. Lau, and J.S. Smith, "Relative Intensity Noise of Vertical Cavity Surface Emitting Lasers," submitted to *Applied Physics Letters*.

[3] G.C. Wilson, D.M. Kuchta, J.D. Walker, and J.S. Smith, "Transverse Modes and Spatial Hole Burning in Vertical-Cavity Surface-Emitting Laser Diodes," submitted to *Applied Physics Letters*.

I-C. Optical Probing of Semiconductor Devices and Interfaces by Electro-Optic and Photo-Elastic Effects

Professor J.S. Smith (S.Wang) with P.J. Harshman

We have used optical probing experiments to study heterostructures in which misfit and thermal stresses are important. In particular, we have studied (111) oriented zinc-blende strained-layer heterostructures. These structures are of interest because they possess large built-in electric fields through the piezoelectric effect, an effect that vanishes in (100) oriented strained layers due to symmetry constraints. The built-in electric field is produced by strain in the quantum well and is related to the strain tensor elements through the piezoelectric tensor. Thus, optically probing the built-in electric fields has allowed us to detect the presence of biaxial strain. Specifically, we have reported on both the growth and characterization of a strained (111)B AlAs/AlInAs multiquantum well structure which exhibits smooth surface morphology and narrow photoluminescence spectra. The excitonic transition energy for this structure has been calculated, taking into account both the effects of strain on the band structure and also the effects of quantum confinement. The result of the calculation differs from the experimentally obtained fundamental excitonic transition energy by 140meV. This discrepancy may be explained either by uncertainties in quantum well width and composition or, possibly, by spontaneous ordering of the AlInAs alloy. In addition to the fundamental c1-hh1 photoluminescence peak, a second peak which is believed to be related to the c1-hh2 exciton has been observed. Observation of the c1-hh2 exciton can be attributed to the existence of an internal electric field which mixes the parities of the quantum well states. Also, the c1-hh1 photoluminescence peak has been found to undergo a gradual blue-shift with increasing optical probe intensity. As the optical probe power is increased from 0.5mW to 145mW, the c1-hh1 peak undergoes a blue shift of approximately 7meV, roughly equal to the strain-induced built-in stark (red) shift estimated from lattice mismatch data. This result can be attributed to the fact that carriers excited by the above-gap optical probe screen the strain-generated internal electric fields.

JSEP Publications

[1] P. J. Harshman and S. Wang, "Investigation of (111) Strained- Layers: Growth, Photoluminescence, and Internal Electric Fields," *J. Appl. Phys.*, Vol. 71, pp. 5531, 1992.

SPL-I. Ultra-high Speed Semiconductor Lasers

Professor K. Lau with M. Daneman and M. Kiang

In the past year, we have concentrated our activities on obtaining a fundamental understanding of the role of carrier transport in quantum well lasers in their high speed modulation characteristics. For the past decade, it has been assumed that such transport effects are too fast to be of substantial consequence. Not until the past year did we begin to understand that intrinsic quantum capture of carriers into the quantum well, which occurs on a sub-picosecond time scale, has a severe effect on the modulation of quantum well lasers in the tens of gigahertz range. The effects are particularly severe in lower dimensional quantum materials. Through a combination of experimental and theoretical studies we are uncovering the physics behind these effects.

We investigated the quantum capture limited modulation bandwidths of various lower-dimensional semiconductor lasers. We showed that, for buried quantum well, wire, and dot lasers, the maximum bandwidth is proportional to the "packing density" of the active region. For the quantum wire lasers grown on V-grooved substrates, the maximum bandwidth is enhanced by the pre-capture of carriers from 3-dimensional states to 2-dimensional states before the capture into the 1-dimensional states.

A simple, physical explanation of why the packing density of the quantum well, wire and dots should affect the quantum-capture speed is based on the fact that the capture cross-section is proportional (to first order) to the number of 3-D carriers at the location of the quantum well(s). A low packing density implies that the carriers are spread-out across the barrier region and hence the carrier density at any given location is low. The capture cross-section is thus small and the corresponding capture time is long. This adversely affects the modulation speed limit in these lasers. The problem is more severe in quantum wires, and even more so in quantum dots, since the packing density (volume ratio) increases exponentially with decreasing dimensionality. This poses a fundamental question: are all the recent predictions and excitements about the superiority of quantum wire and dot lasers in terms of high speed properties all misguided, since they do not take into account the quantum capture effects?

Perhaps a more appropriate question to ask is: Are there any ways to negate the adverse effects of quantum capture and hence realize the full potential of these lower-dimensionality materials? The answer, it turns out, lies in how the bandstructure of the barrier (SCH) region is configured. We have analyzed the influence of bandstructure design of the SCH region on the transport-limited modulation bandwidth in quantum well lasers. By properly grading the SCH region, limitations due to physical-space transport can largely be removed. The limitation due to intrinsic quantum capture (state-space transport) then become the dominant one, though this, too, can be alleviated (but not completely removed) by proper bandstructure design.

To summarize our conclusions, a graded-bandgap barrier region assists tremendously in reducing transport effects, both in the physical-space (diffusion) and in the state-space (quantum capture). A simple, physical explanation of these results is as follows: the graded-bandgap barrier creates an electric field which accelerates the carriers towards the quantum wells (wires or dots) and hence, reduces significantly the diffusion effect. Moreover, the graded barrier tends to accrue carriers in the vicinity of the bottom of the barrier-well, where the quantum wells (wires or dots) are typically located. This increases the capture cross-section of the carriers and hence reduces the effect of quantum capture.

An experimental setup has been completed where the intrinsic modulation response of quantum well lasers can be measured in a parasitic-free manner, and most importantly, the carrier can be injected directly into the quantum well or into the barrier layers by a wavelength-selective optical injection-modulation method as described in the original proposal. The experimental system is now under calibration and quantitative results will be obtained shortly after calibration has been completed.

For a detailed description of the results described above, please consult the following two publications.

JSEP Publications

[1] S.C. Kan, D. Vassilovski, T.C. Wu, and K.Y. Lau, "Quantum Capture Limited Modulation Bandwidth of Quantum Well, Wire and Dot Lasers," accepted for publication in *Appl. Phys. Lett.*, May 1993.

[2] T.C. Wu, S.C. Kan, D. Vassilovski, and K.Y. Lau, "Influence of Separate-confinement Layer Bandstructure on the Transport-limited Modulation Bandwidth in Quantum Well Lasers," submitted for publication in *IEEE Photon. Tech. Lett.*

II-A. 0.1 μm BiCMOS Devices in Bulk and SOI Substrates

Professors Chenming Hu and Ping K. Ko with F. Assaderaghi, Jian Chen, ManSan Chan, Steve Parke and Z.H. Liu

We continued to investigate the physics of and technology for BiCMOS devices in bulk and SOI substrates with dimensions in the vicinity of 0.1 um.

We explored the physics of gate induced band-to-band tunneling current in silicon MOSFET. We proposed an analytical physical model to explain the dependence of this current on the drain profile in conventional and LDD MOSFET's[1][13]. The model has been applied to examine such issues as low leakage drain engineering and oxide scaling limits, which are of primary importance to future scaled MOS technologies. The band-to-band tunneling phenomenon was also investigated for its impact on SOI MOSFET's and was applied to measure the current gain of the MOSFET's parasitic bipolar junction transistor[5][8]. Through this study, we were able to quantify how this parasitic device limits the breakdown voltage of the MOSFET.

We invented a new technique to measure v-E(drift velocity versus electric field) curves for electrons and holes in the inversion layer. By applying a very high gate bias (>30 V) at the back gate of a SOI MOS device, we can establish a nearly uniform channel field in the backside inversion channel. Velocities for electrons and holes have been measured up to field strengths in excess of 1E5 V/cm[6]. Saturation velocities for electrons and holes are found to be 6.5E6 cm/s and 3.5E6 cm/s respectively at room temperature, and 9E6 cm/s and 7E6 cm/s respectively at 85K. Our results provide the first experimental verification that saturation velocity of inversion-layer electrons and holes are insensitive to the gate field.

Fabrication of many new device structures (and the associated technologies) conceived in 91/92 was completed in mid to late 1992.

Among these devices is a PMOS transistor on SOI substrate, with channel length of 0.1 um and gate oxide thickness of 54 Å, that displays a transconductance of 250 mS/mm - believed to be the highest achieved so far. We produced a depletion-load NMOS ring oscillator on SOI substrate that exhibits a minimum propagation delay of 14 ps at a supply voltage of only 1.5 V [12] at room temperature. The 14 ps result is the fastest reported for a silicon technology at any temperature.[14]

We have also successfully demonstrated lateral NPN and PNP transistor with current gains larger than 100 on a SOI substrate[3][14] and explored a BJT-MOSFET hybrid that promises high application potential in low-voltage, high speed circuits. With only 10 mask levels, the lateral BJT's can be integrated into a versatile complementary BiCMOS technology on SOI substrate[11].

We continued the investigation of nitrided thermal oxide as an alternative to pure thermal oxide as the thin gate dielectric [2][9] of choice for future MOS technologies. We demonstrated that nitrided oxides grown using a simple two-step process, with post-oxidation annealing of thermal oxide performed in an N_2O environment, are as good as those produced using much more complex processes[11].

We have also reported the most quantitative model for the ever troublesome short channel effect[15]. This will become the dominant challenge for future MOSFET scaling. For gate oxide, we continued the research into stacked thermal/CVD oxide. We reported a high quality stacked oxide technology [16] that is suitable for 60 Å oxide. We found that only 20 Å CVD oxide is sufficient to "repair the defects" and significantly reduce the oxide defect density.

JSEP Publications

[1] S. A. Parke, J. E. Moon, H.C. Wann, P. K. Ko, and C. Hu, "Design for Suppression of Gate-Induced Drain Leakage in LDD MOSFET's Using a Quasi-Two-Dimensional Analytical Model," *Trans. Electron Devices*, Vol. 39, No. 7, p. 1694, July 1992.

[2] Z. Liu, H.J. Wann, P. K. Ko, C. Hu, and Y.C. Cheng, "Effects of N_2O Anneal and Reoxidation on Thermal Oxide Characteristics," *EDL-13*, No. 8, p. 402, Aug. 1992.

[3] S. Parke, C. Hu, and P. K. Ko, "Deep Sub-micron Bipolar-MOS Hybrid Transistors Fabricated on SIMOX," *Proceedings of the IEEE SOI Conference*, p. 82, Oct. 1992.

[4] S. Parke, C. Hu, and P.K. Ko "Complementary, High-performance Lateral BJT's in a SIMOX C- BiCMOS Technology," *Proceedings of the IEEE SOI Conference*, p. 142, Oct. 1992.

[5] J. Chen, F. Assaderaghi, P. K. Ko, and C. Hu, "The Enhancement of Gate-Induced Drain Leakage(GIDL) Current in SOI MOSFET's as a Function of Temperature, *"Proceedings of the IEEE SOI Conference*, p. 84, Oct. 1992.

[6] F. Assaderaghi, J. Chen, P.K. Ko and C. Hu, "Measurement of Electron and Hole Saturation Velocities in Silicon Inversion Layers Using SOI MOSFET's, *"Proceedings of the IEEE SOI Conference*, p. 112, Oct. 1992.

[7] J. Chen, R. Solomon, T.Y. Chan, P.K. Ko, and C. Hu, "Threshold Voltage and C-V Characteristics of SOI MOSFET's Related to Si Film Thickness Variation on SIMOX Wafers," *Trans. Electron Devices*, Vol. 39, No. 10, p. 2346, Oct. 1992.

[8] J. Chen, F. Assaderaghi, P.K. Ko, and C. Hu, "The Enhancement of GIDL Current in Short-Channel SOI MOSFET and its Application in Measuring Lateral Bipolar Current Gain β," *Electron Device Letters*, Vol. 13, No. 11, p. 572, Nov. 1992.

[9] Z.H. Liu, H.J. Wann, P.K. Ko, C. Hu, and Y.C. Cheng, "Improvement of Charge Trapping Characteristics of N_2O-Annealed and Reoxidized N_2O-Annealed Thin Oxides," *EDL-13*, No. 10, p. 519, Oct. 1992.

[10] S. Parke, F. Assaderaghi, J. Chen, C. King, C. Hu, and P.K. Ko, " A Versatile, SOI BiCMOS Technology with Complementary Lateral BJT's," *Technical Digest of IEDM*, p. 453, Dec. 1992.

[11] Z. Liu, J. Krick, H. Wann, P.K. Ko, C. Hu, and Y.C. Cheng, "The Effects of Furnace N_2O Anneal on MOSFET's," *Technical Digest of IEDM*, p. 625, Dec. 1992.

[12] J. Chen, S. Parke, J. King, F. Assaderaghi, P.K. Ko, and C. Hu, "A High Speed SOI Technology with 12/18ps Gate Delay Operating at 5V/1.5V," *Technical Digest of IEDM*, p. 35, Dec. 1992.

[13] H. Wann, P.K. Ko, and C. Hu, "Gate-Induced Band-to-Band Tunneling Leakage Current in LDD MOSFET's," *Technical Digest of IEDM*, p. 147, Dec. 1992.

[14] S. Parke, C. Hu, and P.K. Ko, "A High-Performance Lateral Bipolar Transistor Fabricated on SIMOX," *EDL-14*, No. 1, p. 33, Jan. 1993.

[15] Z.H. Liu, C. Hu, J-H. Huang, T-Y. Chan, M-C. Jeng, P.K. Ko, Y.C./ Cheng, "Threshold Voltage Model for Deep-Submicrometer MOSFET's," *IEEE Transactions on Electron Devices*, Vol. 40, No. 1, pp. 86-95, January 1993.

[16] R. Moazzami, C. Hu, "A High-Quality Stacked Thermal/LPCVD Gate Oxide Technology for ULSI," *IEEE Electron Device Letters*, Vol. 14, No. 2, pp. 72-73, February 1993.

II-B. Conductive Oxides and Ferroelectrics for Programmable Devices

Professor Chenming Hu with Hyungcheol Shin

The future viability of DRAM devices has been called into question because of the need for ever larger capacitance per micron square. We have investigated ferroelectric PZT films for this application[3]. Ferroelectric lead zirconate titanate (PZT) films with as much as 2.5 times the storage capacity of the best reported silicon oxide/nitride/oxide (ONO) stacked dielectrics have been fabricated. A 2000-Å $PbZr_{0.5}Ti_{0.5}O_3$ film with an effective SiO_2 thickness of 10 Å is demonstrated. Because of the extremely high dielectric constant ($\varepsilon_r \geq 1000$), even larger storage capacities can be obtained by scaling the ferroelectric film thickness whereas the thickness of ONO films is limited by direct tunneling through the film.

The films exhibit ohmic behavior at low fields (with a resistivity of $3.5\times10^{10}\Omega\cdot$cm and an activation energy of 0.33 eV) and exponential field dependence at high fields. Electrical conduction is primarily attributed to electronic hopping. At the same charge storage capacity, the leakage and time-dependent dielectric breakdown characteristics are superior to other dielectric systems. However, lifetime extrapolation to the desired charge storage capacity equivalent to 10 Å of SiO_2 with 2.5-V operation show that time-dependent dielectric breakdown. Leakage current as low as 9×10^{-8}A/cm^2 at 2.5 V for a 4,000 Å film is obtained with the addition of lanthanum and and iron to compensate for lead and oxygen vacancies in the film. Further improvement in both leakage current and time-dependent dielectric breakdown characteristics are necessary to ensure reliable DRAM operation.

A new type of dielectric programmable device is the antifuse. In the simplest case, a dielectric film such as 90 Å oxide-nitride-oxide film is sandwiched between a polysilicon electrode and heavily doped substrate. To program the antifuse into a conducting state, a short voltage pulse breaks down the dielectric film and forms a conductive link. This device is very compact in size and low in resistance in the programmed state. Therefore, it is the preferred device for field programmable gate arrays.

With Actel[2], we examined SEM and cross-sectional TEM of the breakdown spots. They revealed a dome-shaped loose network of the dielectric over the breakdown spot. The many channels through the loose network are apparently the conductive paths of the breakdown dielectric. The diameter of the dome is roughly 3,000 Å for 16mA programming current. Apparently, during breakdown (programming), programming current produces sufficiently high temperature to melt the silicon and ONO film over a small volume centered around the point of breakdown. ONO film breaks up into a loose network and protrudes toward anode, probably due to the drift momentum of the electrons in the molten silicon. Upon cooling at the end of programming, nitride (melting point 1,900 °C) and oxide (melting point 1,700 °C) solidify into the dome first, then followed by silicon (melting point 1,400 °C) solidification. High quality epitaxial growth of silicon takes place when the melt volume is large, e.g., at 16mA current. When the melt volume is small, such as ata 3.5mA current, the temperature gradient and cooling rate are large -- resulting in defects in the breakdown spot below ONO film. This prevents formation of epitaxial single crystal inside the channel. In addition, a small-grain polysilicon crest is formed over the dome. It shows that the molten silicon region has a larger diameter than the molten ONO because of silicon's lower melting point. We presented an invited paper on this type of device [5] at the 1992 IEDM.

Modifications are made to Fowler-Nordheim tunneling current analysis to accurately model the measured conduction characteristics of insulator layers thinner than 6 nm. The most significant is direct tunneling for which a closed form expression is introduced. Polysilicon depletion and electron wave interference are also considered. Four nanometers is found to be a practical limit for SiO_2 scaling due to direct tunneling leakage almost independent of power supply voltage[1].

Low-field current following Fowler-Nordheim stress of thin gate oxides is studied[14]. The conduction mechanism is attributed to trap-assisted tunneling of electrons. For oxides thicker than 100 Å, this stress-induced current is observed to decay as traps are filled without significant tunneling out of traps. In thinner oxides, steady-state current flows when there is an equilibrium between trap filling and emptying processes. This model is observed to be consistent with stress-induced current behavior in a wide range of oxide thicknesses (60 Å to 130 Å) and process technologies. Stress-induced current may cause oxide failure prior to dielectric breakdown and consequently cannot be neglected especially in critical applications such as EEPROM's and DRAM's.

JSEP Publications/Presentations

[1] K. Schuegraf, C. King, and C. Hu, "Ultra-Thin Silicon Dioxide Leakage Current and Scaling Limit," *1992 Symposium on VLSI Technology Digest,* pp. 18-19, June 1992.

[2] S. Chiang, R. Wang, T. Speers, J. McCollum, E. Hamdy, and C. Hu, "Conductive Channel in ONO Formed by Controlled Dielectric Breakdown," *1992 Symposium on VLSI Technology Digest,* pp. 20-21, June 1992.

[3] R. Moazzami, C. Hu, and W.H. Shepherd, "Electrical Characteristics of Ferroelectric PZT Thin Films for DRAM Applications," *IEEE Trans. Electron Devices,* Vol. 39, No. 9, pp. 2044-2049, Sept. 1992.

[4] R. Moazzami and C. Hu, "Stress-Induced Current in Thin Silicon Dioxide Films," *Technical Digest International Electron Devices Meeting,* pp. 139-142, San Francisco, Dec. 1992.

[5] **Invited Paper,** C. Hu, "Interconnect Devices for Field Programmable Gate Array," *Technical Digest International Electron Devices Meeting,* pp. 591-594, San Francisco, Dec. 1992.

II-C. Insulated-Gate GaAs Field Effect Transistors

Professors N. Cheung, C. Hu, and W.G. Oldham with J. Chan

Introduction

Aluminum nitride has found widespread application due to its direct wide bandgap (6.2eV), good thermal conductivity, and high temperature stability. In particular, it has been investigated as a surface passivation material for GaAs-based IC's, as an insulated-gate material for GaAs-based metal-insulator-semiconductor FETs (MISFET), and as a potential semiconductor material for UV light sources and detectors [1]. In this work, we have extensively studied various aspects of sputtered AlN_x thin films via several characterization techniques. Material properties of AlN_x thin films were analyzed by Rutherford Backscattering Analysis (RBS), X-ray diffraction, and optical absorption methods. Insulating properties of AlN_x were deduced from current-voltage (I-V) electrical measurements. Finally, defect state characteristics of AlN_x were studied via capacitance-voltage (C-V) and conductance-frequency (G-ω) techniques.

Summary of Experimental Results

Material Properties

AlN_x thin films were reactively sputtered using an RF magnetron sputterer equipped with a 2" Al target (99.999% purity). Various RF power settings, N_2/Ar gas ratios, and chamber pressures were employed and compared. Increasing the RF power improved the deposition rate of AlN_x films, whereas high N_2/Ar gas ratios improved the stoichiometry of the films. Nitrogen-rich nitride films were grown on silicon (both 111 and 110 orientations), R-cut sapphire and carbon substrates. RBS analysis showed that a high N_2/Ar gas ratio (7:1) produced stoichiometric AlN_x (x=1.0) (Figure 1). However, there was a 10% oxygen concentration present in the film which can be attributed to the large free energy of formation of Al_2O_3 over AlN_x (Figure 2).

A ceramic resistive heater was employed to heat the substrates to temperatures up to 700°C during growth. Dependence of preferred texture on deposition temperature was investigated. X-ray diffraction measurements showed that a preferred orientation along the basal plane (0001) was achieved at a minimum temperature of 450°C for AlN_x films on both types of Si substrates and sapphire (Figure 3).

Optical absorption measurements were carried out on the AlN_x on sapphire samples using mercury and deuterium lamps with a Perkin-Elmer monochrometer. The absorption spectrum (in the form of absorption versus energy gap) was calculated from which the energy bandgap was extrapolated. The bandgap values ranged from 5.2 to 5.9eV (Figure 4), which agreed with literature [2]. Furthermore, the average index of refraction values of the AlN_x thin films was measured by an ellipsometer as 2.064, which again compared favorably with data reported elsewhere [3].

Implanting nitrogen into nitrogen deficient AlN_x films was also investigated. A method called plasma immersion ion implantation (PIII) was employed to perform the nitrogen implants. The implanted films were studied for changes in film composition using Rutherford Backscattering (RBS). Optical absorption behavior of the films was analyzed using a Perkin-Elmer monochrometer, and electrical characteristics were measured using both high frequency capacitance-voltage (CV) and current-voltage (IV) measurements. The implanted films were subsequently rapid thermal annealed in a nitrogen atmosphere from 550 to 1060°C. Experimental results indicated that we have successfully increased the N/Al ratio of these nitrogen-deficient films through the implantations. In general, the implanted AlO_xN_y films suffered 10 times more leakage current than non-implanted samples. However, we found that ion implantation stabilized the leakage current of the AlO_xN_y film and minimized mobile charge generation during high temperature annealing (1060°C). Furthermore, optical absorption measurements showed that both the implanted and unimplanted samples showed comparable improvements in energy bandgap values (~6 eV) with annealing temperature [4].

Insulator Properties

Current-voltage tests performed on AlN_x/Si MIS capacitors confirmed that AlN_x is indeed an insulator. Electric breakdown fields ranging from $1.3x10^5$ to $7.9x10^5$ V/cm were measured for all heated samples (450-900°C) with leakage current on the order of at most $100nA/cm^2$ (Figure 5). In addition, a survey of past attempts to fabricate insulated gates for III-V FETs led us to the concept of a gate structure in the form of insulator/$Al_yGa_{1-y}As$/GaAs. This structure will preserve the high interface quality of the $Al_yGa_{1-y}As$/GaAs interface. The "insulator" was a high bandgap semiconductor which prevents thermionic or tunneling current through the stacked gate structure. The insulated-gate structure being studied was a capacitor consisting of Al contact on reactive-sputtered AlN_x on top of a linearly graded n-type $Al_xGa_{1-x}As$ layer which varied from x=0.33 at the GaAs substrate surface to the top surface (Figure 6). The motivation for using this structure was to improve upon the AlN/$Al_xGa_{1-x}As$ interface condition through a solid-phase epitaxial reaction. Current-voltage tests performed on AlN_x capacitors sputtered on the graded-AlGaAs substrate measured leakage current density on the order of $10nA/cm^2$ (Figure 7).

Defect State Characterization

C-V measurements revealed a hysteresis behavior which could be attributed to interface traps at the AlN_x/Si interface (Figure 8). G-ω measurements showed that interface trap density was on the order of $10^{11}/cm^2$ for both the Si and MBE MIS structures (Figure 9).

Conclusion & Future Work

We have confirmed that AlN_x is an insulator which has a direct bandgap value up to 5.9eV. Furthermore, electrical measurements carried out on MIS capacitors consisting of AlN_x thin film on silicon as well as MBE showed an electric breakdown field approaching $1x10^6$ V/cm; in addition, the interface trap density of the AlN_x/substrate interface was on the order of $1x10^{12}/cm^2$. X-ray diffraction showed that even at a substrate temperature of only 450°C, AlN_x film oriented in the basal direction was achieved, which is important for low temperature processing of crystalline direct wide bandgap material for applications in the area of UV light sources and detectors. Future work will include investigating the optical properties of AlN_x specifically for use as a crystalline semiconductor material rather than as an insulating film. The completely miscible and direct-gap Al-Ga-N system [5] will also be studied, and the resulting films characterized via X-ray diffraction, optical absorption, Hall measurements as well as other electrical methods mentioned earlier.

References

[1] J.H. Edgar,"Prospects of Device Implementation of Wide Band Gap Semiconductors," *J. Mat. Res.*, Vol.7, p. 235, 1992.

[2] K.L. Ho, K.F. Jensen, S.A. Hanson, J.F. Evans, D.C. Boyd, and W.L. Gladfelter,"MOCVD of Wide Bandgap III-V Semiconductors by Using Novel Precursors," *Mat. Res. Soc. Symp. Proc.*, Vol.162, p. 605, 1990.

[3] D.H. Wang and L. Guo, "Optical Properties of Sputtered AlN Films and Coated GaAs," *Thin Solid Films*, Vol.158, p. L39, 1988.

[4] J.S. Chan, N.W. Cheung, and K.M. Yu, "Low Energy Ion Beam Modification of AlNO Thin Film for Insulated Gate Field Effect Transistors," *Mat. Res. Soc. Symp. Proc.*, Vol. 268, p. 377, 1992.

[5] M. Asif Khan, J.M. Van Hove, J.N. Kuznia, and D.T. Olson, "High Electron Mobility GaN/AlGaN Heterostructures Grown by Low-Pressure Metalorganic Chemical Vapor Deposition," *Appl. Phys. Lett.*, Vol. 58, p. 2408, 1991.

JSEP Publications

[1] J.S. Chan, N.W. Cheung, and K.M. Yu, "Low Energy Ion Beam Modification of AlN_xO_y Thin Film for Insulated Gate Field Effect Transistors," *Mat. Res. Soc. Symp. Proc.*, Vol. 268, p. 377, 1992.

[2] J.S. Chan, T.C. Fu, and N.W. Cheung, "Comparison of AlN Thin Films Deposited by RF Magnetron Sputtering and Ion-Assisted Molecular Beam Epitaxy," to be presented at the Spring MRS Meeting, San Francisco, April 12-16, 1993.

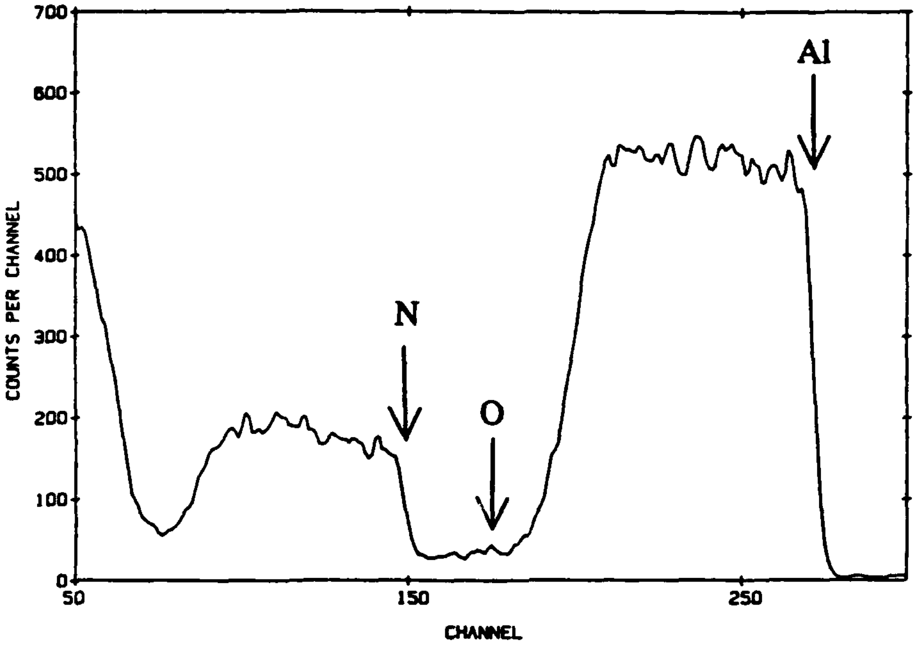
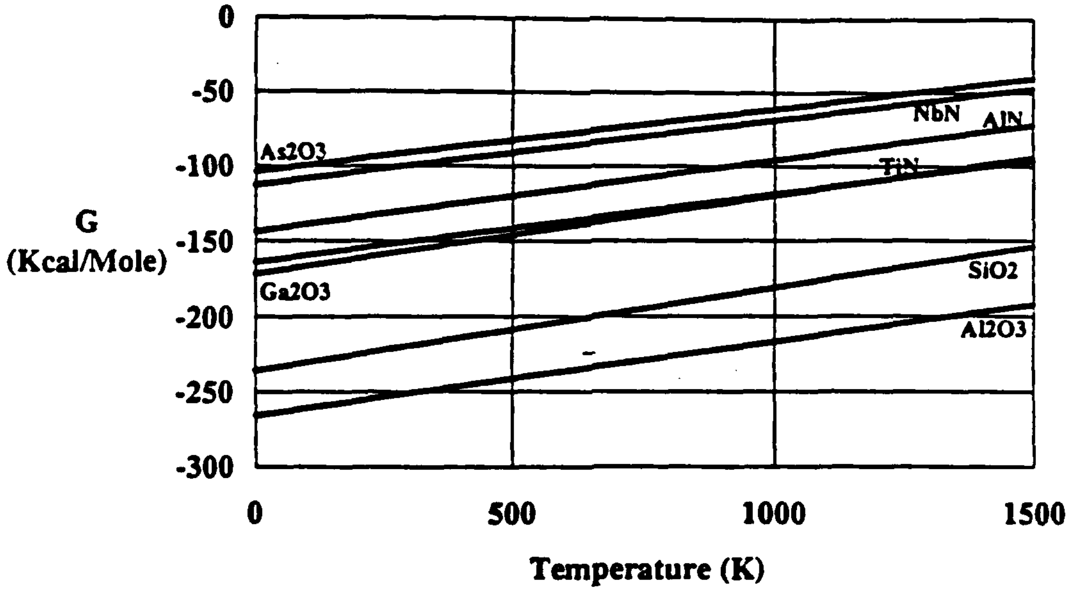


Figure 1: RBS analysis of AlN_x film sputtered on carbon substrate at 200Watts, 10mtorr, and a N_2:Ar ratio of 7:1.



From T. Reed, "Atlas of Charts for High Temperature Chemical Calculations"

Figure 2: Free energies of formation of selected oxides and nitrides.

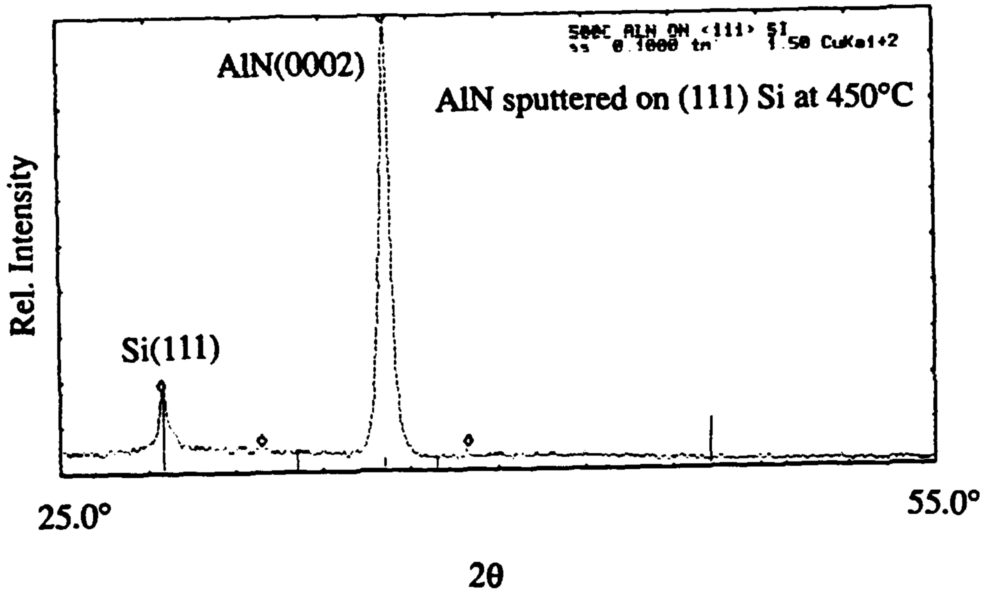


Figure 3: (0001) basal orientation of AlN_x on (111) Si sputtered at 450°C.

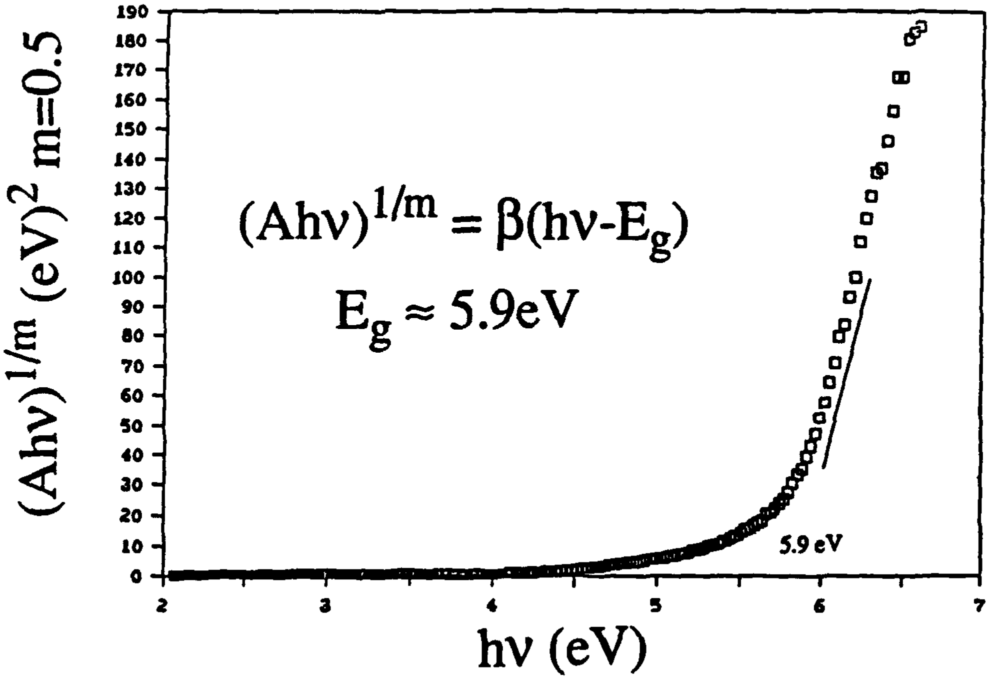


Figure 4: $(Ah\nu)^{1/m}$ vs. hν plot of AlN_x on sapphire, with extrapolated E_g = 5.9eV.

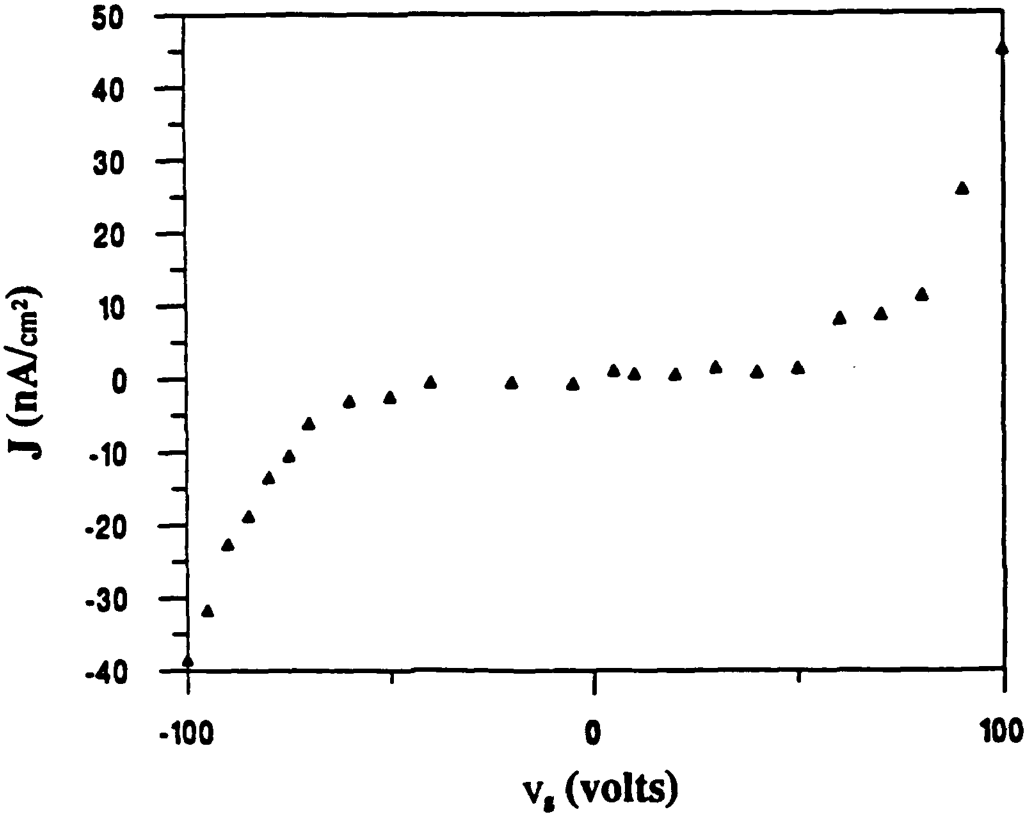


Figure 5: Current density vs. voltage characteristics of AlN_x/Si structure, with breakdown field approaching 10^6 V/cm, and leakage current level on the order of $100nA/cm^2$.

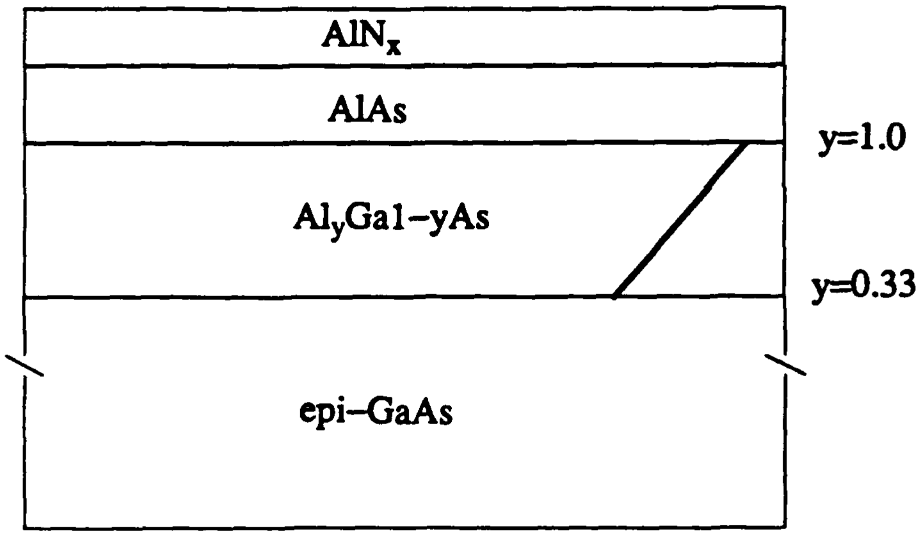


Figure 6: MISFET Structure with graded $Al_yGa_{1-y}As$ substrate grown by molecular beam epitaxy and reactively sputtered AlN_x. AlN_x sputtered on MBE substrate.

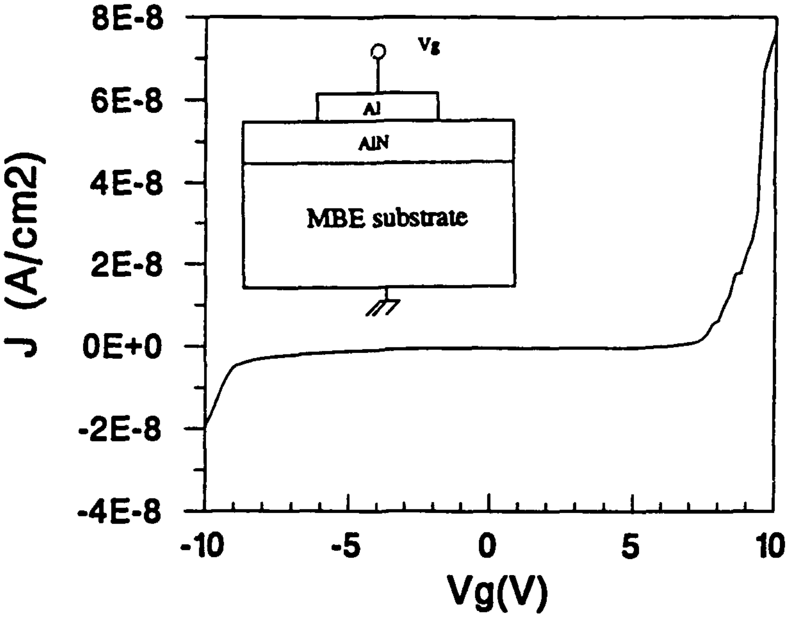


Figure 7: Current density vs. gate voltage of AlN_x/MBE structure, with leakage current on the order of $10nA/cm^2$.

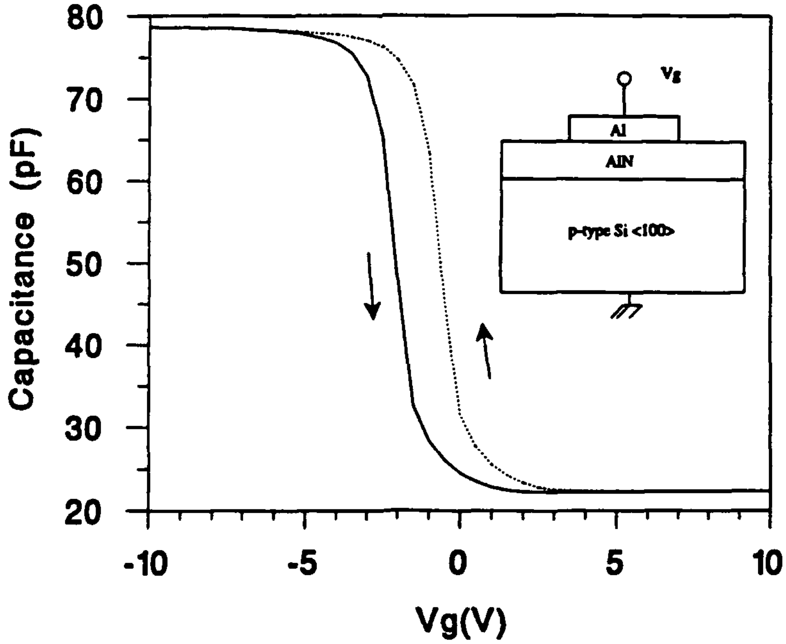


Figure 8: Capacitance-voltage characteristics for Al/AlN/Si MIS test structure, showing hysteresis behavior due to interface traps.

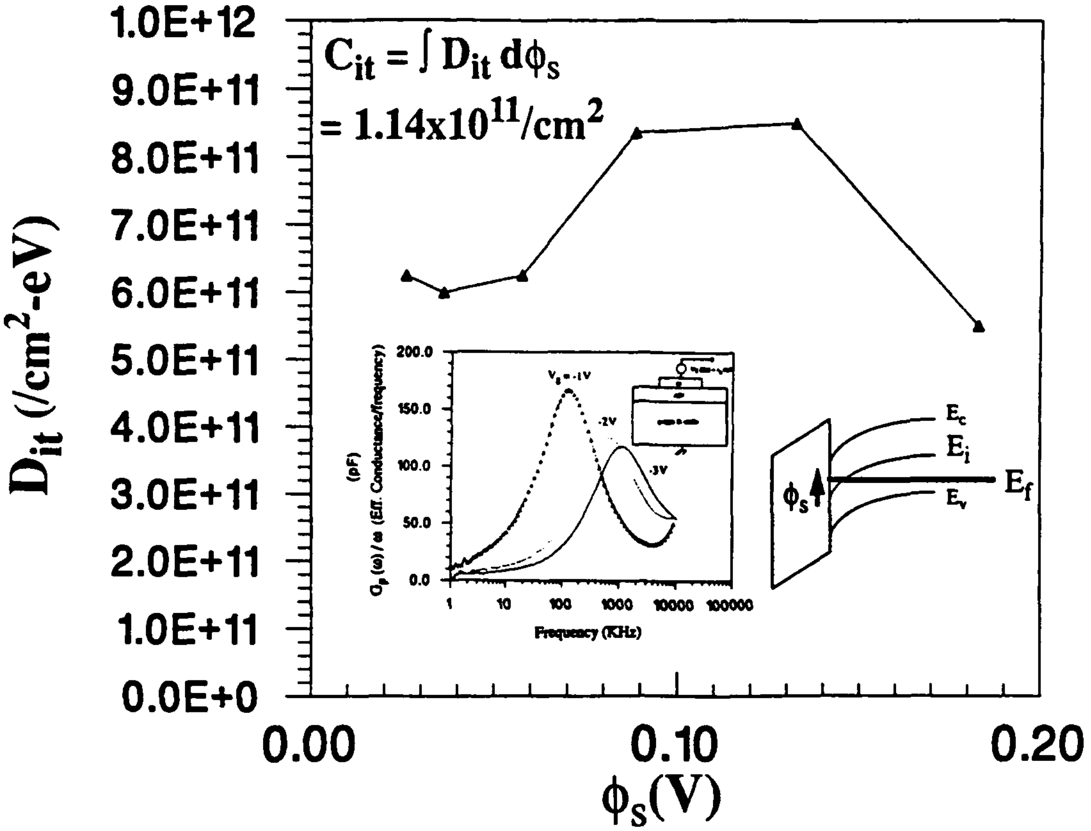


Figure 9: Conductance-frequency characteristics of Al/AlN/Si MIS test structure, showing D_{it} on the order of 10^{11}/cm^2.

III-A. Stochastic Neural Networks and Application to Signal Processing

Professor A. Zakhor with Nick Cobb

During the past year, we continued our efforts in classification of active sonar data using time frequency transforms. We have chosen the Wigner-Ville Transform (WVT), the Wavelet Transform (WT) and the scalogram for our active sonar classification problem. The WVT has been chosen because of its desirable mathematical properties that permit us to match mathematical optimality with physical interpretation. We choose not to use the windowed versions of the WVT, in order to avoid adaptive choice of the window. To reduce the cross terms, we take the Hillbert transform of the real input data to obtain an analytical signal. In addition to the WT, we also consider the square of the modulo of the WT, which is referred to as scalogram, for feature extraction and classification purposes. Similar to the WVT, the scalogram is a bilinear transform, and therefore suffers from cross terms.

As an example, the WVT and the WT of a 5:1 finite steel cylinder with hemispherical endcaps for a 30° off axis incidence and of an empty steel shell are shown in Figures 1 and 2. To use the pictorial time-frequency and time-scale information in Figures 1 and 2 in classification, a finite dimensional feature vector should be selected and fed into the classifier. Since feeding the entire TF transform into the classifier would undermine its generalization capability, we have chosen to use the integral of the TF distribution over square regions as our features. Figures 3 and 4 show our choice of the integration areas for the WVT and the WT. The areas of integration for the scalograms are identical to those of the WT. The numbers inside the squares in Figures 3 and 4 represent a feature that is obtained from the integration of the points in that region. For the WVT, features 1 through 5 represent the sum over the frequency axis. In both cases, the locations and the size of the integration areas have been chosen to maximize the classifier performance while keeping the number of features small.

Our experimental data consists of synthetic returns from two solid elastic cylinders with hemispherical endcaps and a length to diameter ratio of 10 and 5 respectively, along with a solid sphere, an aluminum spherical shell and two steel shell of different thicknesses. This results in 6 different classes; moreover the data from the cylinders are obtained for various angles of incidence, resulting in a total of 7 different aspect ratios for each of the 2 cylinder classes. We add random white Gaussian noise of up to -4dB of SNR to the original data. The SNR is calculated by computing the power of the signal over the entire burst of 512 time samples. This way, we generate 12,000 returns, of which 8,000 are used to train the decision tree and a polynomial classifier, and 4,000 are used to optimally prune the same decision tree. Another set of 1,000 independent returns for each of the SNR levels is used for testing the performance of the classifier.

The classification rates for the decision tree and second order maximum margin polynomial classifier are shown in Tables 1 and 2 respectively. As seen, at high SNR, WVT, scalogram and WT perform more or less in the same way. At low SNR however, the wavelet transform outperforms both Wigner-Ville and scalogram for both the decision trees and the maximum margin polynomial classifier. This can be attributed to the lack of cross terms in the wavelet transform as compared to the scalogram and the Wigner-Ville transform.

Finally, comparing our results to existing results in the literature by Gorman and Sejnowski [1], we find that our approach results in superior classification rates. In particular, they report detection rates of 84% for two objects for SNR between 4 and 15dB, while the classification rate among 6 objects for our wavelet based technique at -4 dB SNR is around 92%.

In the coming year, we will complete our work on sonar and move on to another application of neural networks, namely face recognition.

References

[1] Gorman and Sejnowski, "Learned Classification of Sonar Targets Using a Massively Parallel Network," *IEEE Transactions on Acoustics, Speech and Signal Processing,* Vol. 36, No. 7, July 1988.

JSEP Publications/Presentations

[1] F. Lari and A. Zakhor, "Automatic Classification of Active Sonar Data Using Time Frequency Transforms," SPIE International Symposium on Applied Science and Engineering, San Diego, California, July 1992. Also presented at the IEEE Signal Processing Workshops on Time-Frequency and Time Scale Transforms, Vancouver, Canada, October 1992; also submitted for publication to *IEEE Transactions on Signal Processing.*

SNR	% correctly classified					
	DETECTION			CLASSIFICATION		
	WVT	scal	WV	WVT	scal	WV
10	100	100	99	99	100	98
8	100	98	100	99	98	100
6	100	98	99	97	97	98
4	99	96	98	95	95	94
2	97	94	97	94	90	89
0	95	88	97	90	86	90
-2	86	78	94	73	71	86
-4	71	65	92	47	56	76

Table 1: Detection and classification of decision trees for wavelets, scalograms and the Wigner-Ville Transform

SNR	DETECTION		CLASSIFICATION	
	WV	scal	WV	scal
6	99	100	98	99
4	98	99	98	99
2	99	99	98	97
0	97	95	95	90
-2	92	90	88	82
-4	85	74	77	62

Table 2: Detection and classification of second order maximum margin polynomial classifiers for wavelets and scalograms.

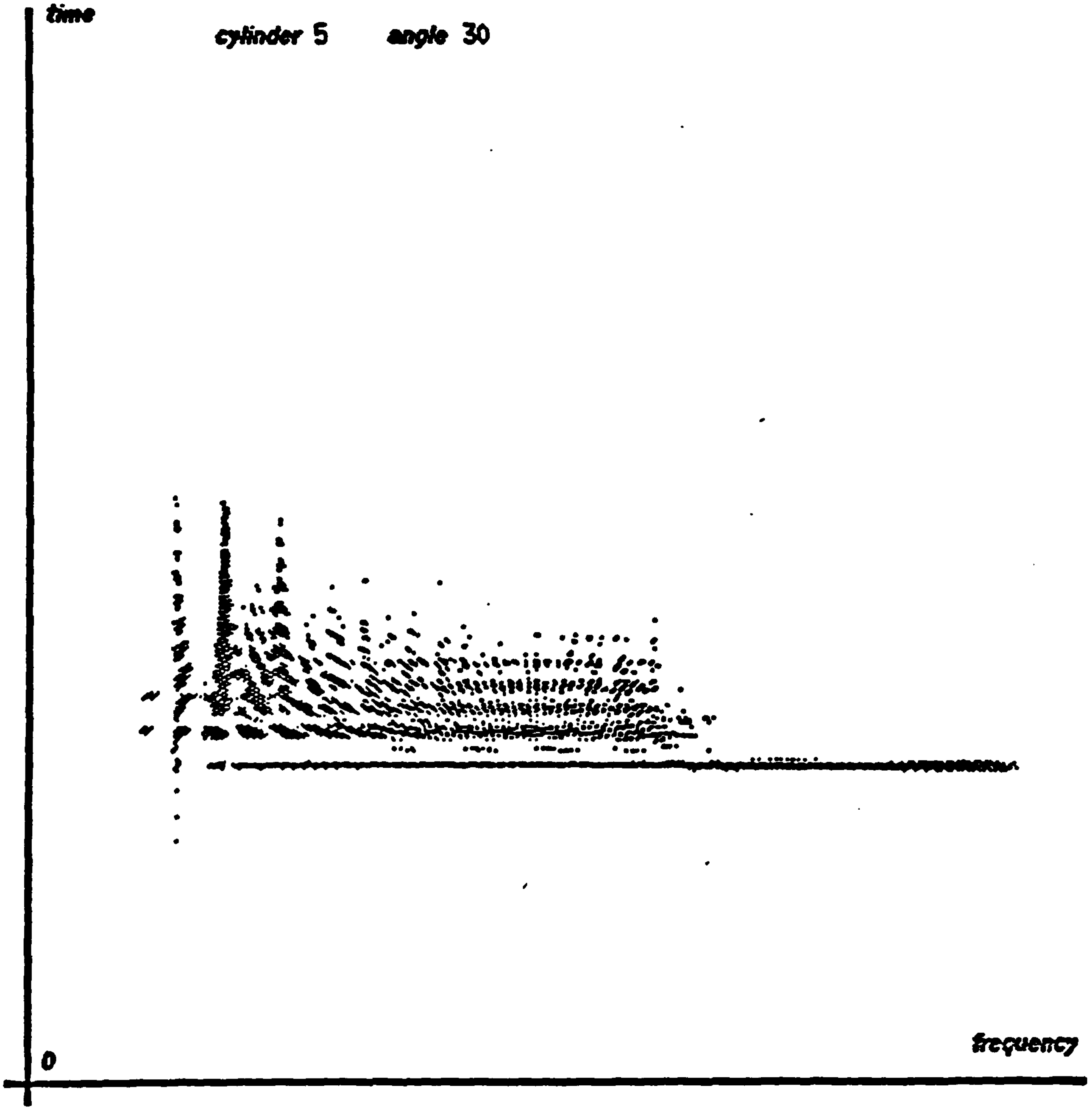


Fig. 1: An example of Wigner-Ville Transform of the return of a cylinder of 5:1 ratio at angle 30°.

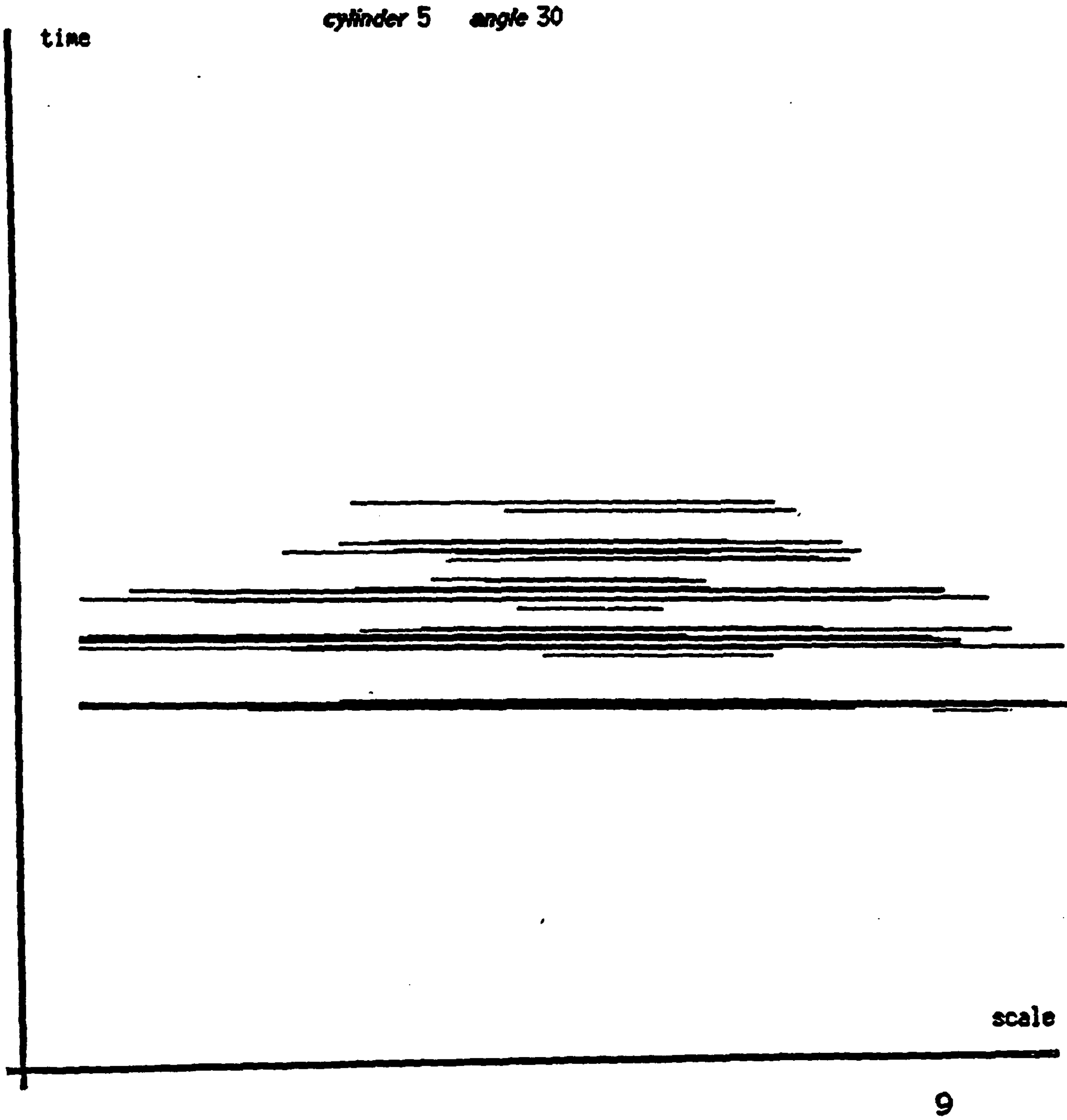


Fig. 2: An example of the Wavelet Transform of the return of a cylinder of 5:1 ratio at angle 30°.

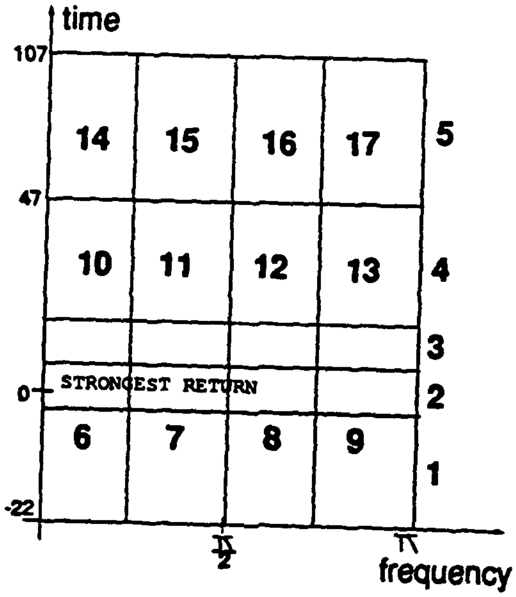


Fig. 3: Feature extraction for
the Wigner-Ville Transform.

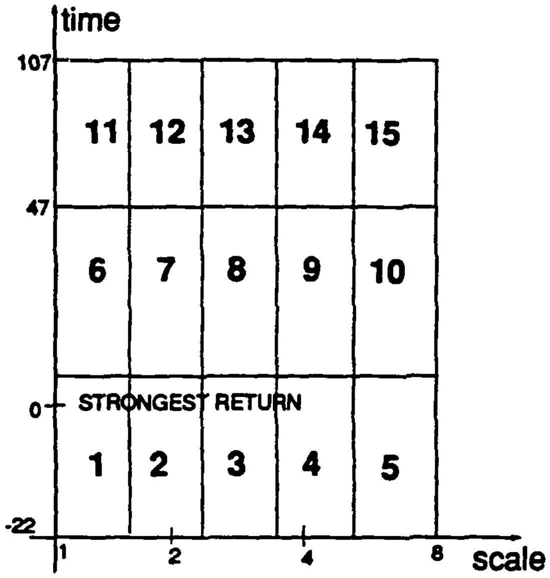


Fig. 4: Feature extraction for wavelets

III-B. Learning and Generalization by Neural Networks

Alberto Sangiovanni-Vincentelli with Arlindo Oliveira

During this period, we actively pursued research related to the design of networks capable of being implemented using standard digital technology.

Since conventional neural network training algorithms are not appropriate for the generation of networks that map naturally into this technology, the major thrust was on the development of new learning algorithms.

In previous work, algorithms for the design on two-layer networks (one hidden and one output layer) were proposed and shown to outperform alternative approaches in problems that accept a compact two-level description. However, for more complex problems, the full power of multi-layer networks is required to achieve good generalization.

Multi-layer networks are more powerful because the first layers extract simple features that are combined into more and more complex ones in the upper layers. The final layers take as input a new set of variables that represent a simpler problem than the original. In principle, the training algorithm should be powerful enough to derive the appropriate functions for the nodes in the first layers. However, many standard training algorithms for neural networks degrade rapidly when the depth of the network increases. In many cases, the features that are to be extracted by the first layers have to be hand-designed.

An important result of this research is an algorithm that selects, level by level, a useful set of features. From the many possible features that can be computed, the algorithm selects a minimal set that is appropriate to perform the classification task at hand. Theoretical and empirical results show that such an approach leads to good generalization accuracy. This approach is also less likely to degrade when the network that is required to learn the problem is very deep. We have shown that the algorithm has a better generalization accuracy than either decision trees or constructive algorithms found in the neural networks literature for a variety of problems. The following table of results compares the performance of this algorithm (MIFES) to the performance of a decision tree algorithm, ID3, and the performance of a constructive neural network algorithm, Cascade-Correlation. The problems selected are a mix of artificially generated data and real-world data extracted from the publicly available database at UC Irvine.

Generalization Error Averaged Over 5 Runs

Problem	No. inputs	MIFES	ID3	Cascade-Correlation
carry4	8	3.7	3.5	4.8
xor8	8	0.0	9.6	38.9
sm12	12	0.0	5.3	11.1
sm18	18	21.1	21.1	15.9
str18	18	5.1	13.5	12.1
tictactoe	27	1.7	6.9	12.8
krkp	38	2.1	4.1	1.8
mushroom	122	0.2	9.3	0.2
Average over all problems	-	4.2	9.3	12.2

Ongoing research is concentrated on the problem of learning efficiently using digital networks even when the problem is noisy (misclassified examples or wrong value of some attributes). The approach we are currently following relaxes the requirement of exact learning of the training set that was present in the previous work. The algorithm currently in development is based on the design of a network that maximizes the mutual information between the output and the class of the example. It has been shown that mutual information maximization is a powerful paradigm in both artificial and natural neural networks. In particular, it has been shown that this paradigm can explain certain forms of self-organization found in the visual cortex, like the appearance of ocular-dominance columns. This is important from a practical point of view because the appearance of such characteristics seems to require from the system a high level of ability to extract useful features from images without direct supervision.

This algorithm is expected to improve the results obtained so far in real-world problems like handwritten character recognition and feature extraction in images. Other applications like the evaluation of positions in games are also being actively investigated.

JSEP Publications/Presentations

[1] Arlindo L. Oliveira and Alberto Sangiovanni-Vincentelli, "What Can Boolean Networks Learn?" 1992 Computational Learning Theory and Natural Learning Systems Workshop, Madison, Wisconsin, August 1992.

[2] Arlindo L. Oliveira and Alberto Sangiovanni-Vincentelli, "Constructive Induction Using a Non-Greedy Strategy for Feature Selection," Ninth International Conference in Machine Learning, Scotland, United Kingdom, July 1992.

[3] Arlindo L. Oliveira and Alberto Sangiovanni-Vincentelli, "Synthesis of Minimal Multi-Level Networks," presented at the Neural Networks for Computing Workshop, Snowbird, Utah, April 1992.

III-C. Reconfigurable Analog Elements for Neural Nets

Professor Ping K. Ko with Alan Kramer, Larry Hung and C.K. Sin

Investigation of the use of Flash EEPROM devices as reconfigurable analog weights continued. An industry test wafer with a small flash array was used to verify the feasibility of programming analog threshold to a precision of eight bits in flash devices. Flash devices present an interesting problem for analog programming in comparison to EEPROM devices because, since flash arrays are typically bulk-erase, the cost of overshooting the desired threshold during programming is very high, because lowering the threshold will necessitate erasure of the entire array. Because of this asymmetry, the analog symmetric programming algorithm that was developed earlier in this project for analog programming of EEPROMs is unusable for Flash-EEPROMs. We have developed and successfully tested (on the before-mentioned wafers) a new programming algorithm for analog threshold adjustment of Flash EEPROMs. This new algorithm utilizes the fact that, based on the physics of EPROM cell programming, the rate of threshold voltage change with programming time is monotonically decreasing. The threshold voltage of the cell is sampled after each programming step and the programming time(and/or voltage) for the next step is determined by a linear approximation to the calibrated programming curve of the Flash device in a way that allows the device to be programmed to an analog value without overshoot. Eight-bit values have been successfully programmed.

In addition, a small test layout has been completed for fabrication in an industrial flash EEPROM process. The main intent of these test circuits is to test an analog flash array for device mismatch when used as analog transistors or capacitors as well as to test the ability of these devices to be used for a novel neural network architecture based on distance computing neurons. This architecture, which uses flash EEPROM's as variable-threshold capacitors in a way that allows it to compute sums of absolute values very efficiently, operates in the charge domain and suffers severely from the problem of parasitic source and drain capacitance. A scheme to compensate for this has been devised and is also the subject of an experiment on the newly fabricated test structure. The silicon has just come back from fabrication and is currently under testing. Initial results have been promising.

JSEP Publications

[1] C.K. Sin, A. Kramer, V. Hu, R.R. Chu, and P. K. Ko, "EEPROM as an Analog Storage Device, with Particular Applications in Neural Networks," *Trans. Electron Devices*, Vol. 39, No. 6, pp. 1410-1419, June 1992.

[2] L. Hung, "A Programming Algorithm for Flash EPROM Analog Storage," M.S. Thesis, University of California, Berkeley, Dec. 1992.

[3] A. Kramer, P. K. Ko and A. Sangiovanni-Vincentelli, "Massively Parallel Analog Geometric Computation Using EEPROM's," abstract for Neural Networks for Computing Conference, Snowbird, Utah, submitted November 1992.

III-D. Architectural Issues in Parallel Computation

Heterogeneous Architectures for Artificial Neural Networks
Professor C. H. Séquin with Chedsada Chinrungrueng

Existing *monolithics* artificial neural network (ANN) architectures, are not sufficient to cope with complex problems, such as processing of speech or vision. To solve such complex tasks, a classical ANN would have to be of exorbitant size; to build it in parallel VLSI hardware would be too costly and have impractical connectivity requirements. Furthermore, training a large-scale network as a monolithic system is impractical because the dimension of the adjustable parameter space is very large and this would result in unacceptably slow convergence rates.

The existence of *heterogeneous* organization in mammalian visual systems suggests that large scale networks should be composed of a variety of *heterogeneous* modules. Several researchers have proposed a class of heterogeneous architectures that are composed of both supervised and unsupervised learning architectures. In these architectures, an unsupervised learning algorithm, such as the *k-means algorithm*, is used by a gating network for decomposing assigned tasks, and a supervised learning algorithm, such as those based on *gradient descent*, is used by each expert module to solve its assigned subtask.

We have developed an enhancement of the heterogeneous architectures that are based on k-means partitioning. The enhanced architecture is characterized by a novel k-means algorithm that integrates into its partitioning process the information about the input distribution as well as about the structure of the goal function and of the expert modules. This new k-means algorithm allows each individual region in the partition to adjust its size so that the representation resources in all the region are optimally used. In order to enable the new k-means algorithm to achieve its optimal performance and be usable for both stationary and non-stationary situations, we also have introduced into the new k-means algorithm two new mechanisms: The first mechanism is for biasing the partitioning process so that it can avoid being trapped in a bad local minimum. The second mechanism is for adjusting the learning rate dynamically to match the instantaneous characteristics of a problem, permitting the algorithm to converge first very rapidly and later very closely towards an optimal solution.

We have evaluated the *performance* and *complexity* of this enhanced heterogeneous architecture compared to that of a homogeneous radial basis function architecture and to a multilayer perceptrons trained by the error back-propagation algorithm on the Mackey-Glass time series prediction and hand-written capital letter recognition. The evaluation results indicate that the heterogeneous architectures are more efficient in solving these two test problems on both performance and complexity.

JSEP Publications

[1] C. Chinrungrueng and C.H. Séquin, "Adaptive K-Means Algorithm with Error-Weighted Deviation Measure," to be presented at the 1993 International Conference on Neural Networks, San Francisco, California, March 1993.

Fault Tolerance in Layered Artificial Neural Networks
Professor C. H. Séquin with Reed Clay

We have been continuing our investigation into the fault tolerance of feed-forward artificial neural networks. In particular, during 1992, we have documented the additional benefits of our training technique for fault tolerance as it relates to generalization and to the problem of overfitting to training data. Also, through variations of our basic method and through more elaborate studies, we have gained new insights and have refined our understanding of several issues related to fault tolerance training. This includes: how and why fault tolerance can arise naturally in a network, how to use this knowledge to achieve a desired level of fault tolerance faster, and how our technique compares to other work in this area.

The basic idea behind our fault tolerance training technique is to randomly introduce - during training - the types of failures that one might expect to occur during the actual operation of the network. We have shown that this can make the network tolerant not only against the specific faults trained for, but also against faults that have never been presented explicitly and which may also be considerably more severe (i.e., double or triple faults) than any faults ever presented to the network during training.

We have also noted several beneficial side effects that arise from this particular method of training for fault tolerance [1]. Specifically, we have examined the power of generalization of networks trained for fault tolerance. Some applications show a tendency to overfit the network to a noisy set of training data and thus to generalize to new test data more poorly than is warranted. Our fault tolerance training method was able to reduce this amount of overfitting noticeably.

Some other applications can achieve a significant degree of fault tolerance simply by prolonged training with a standard backpropagation method. This phenomenon is probably what has led to the popular misconception that neural nets are inherently fault tolerant. The mechanism behind this phenomenon is the fact that the prolonged training will continue to push the weights of the network in the desired direction, even after the desired classification can be achieved by the network. This is because the output "error" is measured with respect to a perfect saturated output value - not just a value that lies on the proper side of some threshold. As the weights take on more distinct values in this prolonged training, the margins with respect to the separating thresholds of the classification become larger, leading to some fault tolerance. Unfortunately, this method does not work in general, and we have constructed some very simple demonstration cases where this standard training method fails to provide fault tolerance.

The larger magnitude weights can mean that fewer units need to add up to produce the desired output. In this situation, fault tolerance can be achieved simply by longer training since this gradually leads to larger weights. However, larger weights only lead to fault tolerance if the relative settings of the weights have already reached appropriate ratios. For example, using standard backpropagation training (and assuming that there are more than the minimal number of hidden units required to ''learn'' a task), there is usually a significant number of hidden units which do not contribute to the desired solution in any useful way. To effectively recruit this redundant hardware to be used for achieving fault tolerance, different training techniques (such as ours) are required to explicitly force these units to develop useful settings of weight ratios.

Since fault tolerance typically depends on both the relative settings of the weights as well as the increased magnitude of the weights, this implies another possible variation for achieving fault tolerance even more speedily: The network is trained with our technique of randomly introducing typical faults but it is tested by using hard threshold units (which is mathematically equivalent to scaling up the magnitude of the weights). This combination can produce the desired level of fault tolerance faster than using just our fault tolerance training technique.

We also compared our technique to the method of adding random noise to the weights of the units during training. For a fault model where hidden units are simply ''lost'' (i.e. where faulty units output a ''0'', or neutral value), we have seen that this method can be as good as our technique described above. However, for the type of faults where a hidden unit outputs an extreme value (i.e. ''+1'' or ''-1''), the technique of training with sporadic errors of exactly this failure type is still significantly better; obviously, it is advantageous to train the network with the specific types of faults that might actually occur during useful operation.

JSEP Publications

[1] Reed D. Clay and Carlo H. Séquin, "Fault Tolerance Training Improves Generalization and Robustness," *Proc. Int. Joint Conf. on Neural Networks*, pp. I-769-774, Baltimore, Maryland, June 1992.

Interconnection Network Design Based on Packaging Considerations
Professor Abhiram Ranade with M. T. Raghunath

A central problem in building large scale parallel computers is the design of the interconnection network. The design of the interconnection network significantly affects the performance, cost and availability of the parallel computer. Most parallel programs require a large amount of communication bandwidth between the processors and the performance of such programs depends more on the performance of the interconnection network than the processor speed. Building high performance interconnects is expensive and usually accounts for a large fraction of the total cost of a machine. Fault-tolerance is an important issue in parallel computers since they consist of thousands of components. It is important to ensure that faults in the interconnection network do not drastically reduce the performance of the machine.

We examine in detail the problem of interconnection network design for a 1024 processor machine. Our objective is to design a network with high performance, low cost and high availability.

We develop a few abstract models of packaging technology to characterize network cost. In general, it is difficult to model packaging costs precisely, because costs change as technology evolves. Further, large scale parallel machines employ several levels of hierarchy, e.g., racks, boards and chips, and each level has different cost functions. These costs functions, or even the levels of hierarchy themselves may change with advances in technology. Instead of precisely modeling costs for a specific technology, we characterize the generic constraints and costs that are likely to be valid even with changes in technology. Each of our abstract packaging models deals with the costs at progressively increasing levels of detail.

For each packaging model, we evaluate several network organizations that have equal cost under that model. Each network is defined by specifying the interconnections at the different levels of the packaging hierarchy. From among the various network organizations that are possible, we select promising ones based on theoretical analysis. We measure the performance of each of these networks using detailed simulations of random communication traffic. Evaluations are based on both the commonly used *open-network* work-load model and on a more realistic work-load model based on multi-threaded processors. In the latter model, the processors stall for outstanding communication operations unlike the *open-network* model. The presence of multiple threads provides the processors with some ability to tolerate latency.

We also devise fault-tolerance schemes for a few of the interesting networks identified on the basis of the above-mentioned performance evaluations. Most of the networks falling in this category turn out to be variants of the butterfly network. The best networks, however, are different from the butterfly in small but important ways. Most of these networks have a single path between any pair of processors. This property makes the networks vulnerable to faults. We present a simple scheme to tolerate faults and evaluate the performance of this fault-tolerance scheme.

Our results indicate the following general principles: 1) Making the networks denser at the lower levels of the packaging hierarchy has a significant positive impact on network performance, even when the higher levels of packaging use a sparse interconnect, 2) It is better to organize a fixed amount of communication bandwidth as a smaller number of high bandwidth channels rather than a larger number of low bandwidth channels, 3) Providing the processors with the ability to tolerate latencies (by using multi-threading) is very useful in improving performance, 4) For a machine where communication is based on shared-memory primitives, it is better to use fewer high bandwidth memory units. 5) Fault-tolerance schemes that are inexpensive and easy to implement provide high reliability with small degradation in performance.

Although interconnection network design has been a popular research topic for a number of years, there is still no consensus on what constitutes a best interconnection network. Parallel computers announced in the past 4--5 years have used different network topologies such as hypercubes, meshes, fat-trees, etc., indicating that the network design problem is still open. We present a framework for reasoning about the problem. We also provide several network designs that are optimal under different sets of costs and constraints as characterized by our abstract models of packaging technology.

JSEP Publications

[1] M.T. Raghunath and Abhiram Ranade, "Customizing Interconnection Networks to Suit Packaging Hierarchies," *Technical Report UCB//CSD-93-725*, Computer Science Division, University of California, Berkeley, California 94720, January 1993.

[2] M.T. Raghunath and Abhiram Ranade, "Fault-Tolerant Routing in Partitioned Butterfly Networks," submitted to the 1993 International Conference on Parallel Processing.

SPL-II. Analog Neural Networks for Vision Tasks

Professor Leon O. Chua with J.M. Cruz, K.R. Crounse and B.E. Shi

For the period starting March 1992, we have obtained many promising results in image halftoning, spatio-temporal filtering, and VLSI design for Cellular Neural Networks.

Image Halftoning

Halftoning is the process for coding gray-scale images by using binary (black and white) values at each picture element. Upon display it is hoped that, by the blurring of the eye, the halftone image will appear similar to the original continuous toned image. Digital image halftones are required in many present day electronic applications such as facsimile (FAX), electronic scanner/copying, laser printing, and low bandwidth remote sensing. Many algorithms have been devised for halftoning digital images. These algorithms all suffer from defects which are readily apparent when the image is displayed at the marginally sampled resolution and viewed at the critical pixel merge distance.

We have investigated the use of the CNN to perform image halftoning. A class of CNN templates was shown to accomplish the same type of minimization as the well known error diffusion algorithm, a halftoning standard often considered to give good performance [1]. However, unlike error diffusion where the filter coefficients are chosen arbitrarily, the CNN uses an internal model for the filtering done by the human visual system. The algorithm continuously uses this model to evaluate the current output and direct the transient towards the best solution. In addition, the directionality and microstructure apparent in error diffusion is reduced by the circular symmetry of the template and careful choice of coefficients.

Our solution provides excellent quality halftones along with the possibility of an implementation operating at very high speeds. Moreover, some additional image processing functions, such as noise removal, can be incorporated with no extra hardware complexity when implemented on the CNN Universal Machine. The applications which could take advantage of high-speed CNN halftoning include electronic document scanning/copying and real time image compression and binarization [2].

Spatio-temporal Filtering

During the past year, we have obtained several promising results in the analysis of stability and performance prediction for linear space-invariant CNNs. It turns out that despite the high dimensionality of the system, we can characterize the input/output behavior of a linear CNN using a well defined spatio-temporal transfer function. This work has resulted in simple graphical stability criteria. These give necessary and sufficient conditions for the stability of linear space-invariant CNN arrays, which, unlike previously developed stability criteria, fully exploit the reduction in complexity of the system which results from the assumption of spatial invariance. These criteria also have the advantage that they give some indication of the stability of the network in the presence of parameter variations and can be adapted to take into account parasitic elements.

The practical benefit of this research has been in the development of image motion sensitive CNNs based upon the spatio-temporal filtering behavior of motion sensitive cells found in the primary visual cortex[3]. In the computer vision literature, these types of filters have been used in the computation of the optical flow[4], the two dimensional velocity vector field on the image plane resulting from relative motion between a camera and its environment. The CNN approach to spatio-temporal filtering is much more computationally efficient than previous approaches, which relied upon digitizing and storing many image frames and explicitly computing a convolution in space-time. In contrast, the CNN approach uses the continuous time dynamics of the CNN array to perform the convolution in real-time with no need for storage of any image frames.

We are also currently investigating the use of these CNN arrays in phased array radar systems to perform beamforming. Beamforming is the process of combining the signals from an array of sensors in such a way that the output contains only the signal incident upon the array from a given direction. In contrast to current approaches to beamforming, which rely only upon spatial filtering of the data from the array, the CNN approach combines filtering in both space and time. We expect that the CNN approach will be more robust in the presence of sensor noise and element failure than current approaches and that it will have significant advantages for wideband beamforming.

VLSI Design

As we have mentioned in the section describing the CNN Universal Machine, one of the possible analog processing units of the Universal Machine is the generalized Chua circuit. The Chua circuit is one of the very few physical systems in which a formal proof of the existence of chaos has been accomplished and in which the theoretical, simulation and experimental results match precisely. It is also the simplest autonomous circuit which can exhibit bifurcation and chaos. These factors have made the Chua circuit a standard paradigm for studying chaotic phenomena.

In the CNN, arrays of Chua circuits have been successfully designed to generate autowaves. Autowaves, or autonomous waves, are nonlinear wave phenomena which can be excited in an active medium. The famous Zhabotinsky gas reaction demonstrated the presence of autowaves in physical medium. Autowaves are different from classical waves in that they do not obey reflection and interference properties. Instead, autowaves annihilate upon collision with each other and obstacles. However, they do still obey diffraction as predicted by Huygens principle. Recently, interest has been developing in using autowaves for certain image processing tasks[5]. If an image is imposed on an active medium, autowaves can be excited in such a way as to perform interesting distributed parallel computations such as edge detection, contour restoration, and closed-curve detection. In addition, optimization tasks such as finding the least energy path between two points along a nonconservative hilly landscape have also been considered[6].

As a first step toward developing the powerful capabilities of the CNN Universal Machine with the Chua circuit as the basic analog processor, we have recently designed and tested an analog integrated circuit implementation of the only non-linear element in the Chua circuit, the Chua diode. This is an important first step in the eventual design and fabrication of a complete two dimensional array of Chua circuits.

References

[1] R. Ulichney, *Digital Halftoning*, The MIT Press, Cambridge, Massachusetts, 1990.

[2] M. Tanaka, K. R. Crounse and T. Roska, "Template Synthesis of Cellular Neural Networks for Information Coding and Decoding," *Proceedings Second IEEE International Workshop in Cellular Neural Networks and their Applications*, pp. 29-35, 1992.

[3] C. L. Baker, Jr., "Spatial- and Temporal-Frequency Selectivity as a Basis for Velocity Preference in Cat Striate Cortex Neurons," *Visual Neuroscience*, Volume 4, 1990, pp. 101-113.

[4] D.J. Heeger, "Optical Flow from Spatiotemporal Filters," *Proc. Intl. Conf. Computer Vision*, pp. 181-190, 1987.

[5] V. I. Krinsky, V. N. Biktashev, and I. R. Efimov, "Autowave Principles for Parallel Image Processing," *Physica D*, Vol. 49, pp. 247-253, 1991.

[6] V. Perez-Munuzuri, V. Perez-Villar, and L. O. Chua, "Autowaves for Image Processing on a Two-Dimensional CNN Array of Excitable Nonlinear Circuits: Flat and Wrinkled Labyrinths," *IEEE Transactions on Circuits and Systems--I: Fundamental Theory and Applications*, Mar. 1993, to appear.

JSEP Publications

[1] K. Crounse, T. Roska, and L. Chua, "Image Halftoning with Cellular Neural Networks," *IEEE Transactions on Circuits and Systems--II: Analog and Digital Signal Processing, Special Issue on Cellular Neural Networks*, Mar. 1993, to appear.

[2] B. Shi, T. Roska, and L. Chua, "Design of Linear Cellular Neural Networks for Motion Sensitive Filtering," *IEEE Transactions on Circuits and Systems--II: Analog and Digital Signal Processing, Special Issue on Cellular Neural Networks*, Mar. 1993, to appear.

[3] J. Cruz and L. Chua, "An IC Diode for Chua's Circuit," *International Journal of Circuit Theory and Applications*, to appear.