WORKSHOP NOTES

10th Annual Workshop on Interconnections Within High Speed Digital Systems

9-12 May 1999

Hilton of Santa Fe
Santa Fe, New Mexico

Sponsored by the IEEE Lasers & Electro-Optics Society and
in cooperation with the IEEE Computer Society and the
IEEE Communications Society

REPORT DOCUMENTATION PAGE

AFRL-SR-BL-TR-99-

0299

1. AGENCY USE ONLY (*Leave Blank*)	2. REPORT DATE	3. REPORT TYPE AND [illegible] COVERED
	12 May 1999	Technical

4. TITLE AND SUBTITLE
10th Annual Workshop on Interconnection within High Speed Digital Systems

5. FUNDING NUMBERS
G 6271 2E
H832/01

6. AUTHORS
Multiple

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
IEEE-LEOS
445 Hoes Lane
Piscataway, NJ 08855-1331

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR
801 North Randolph Street, Room 732
Arlington, VA 22203-1977

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
F49620-99-1-0300

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

10th Annual Workshop on Interconnections within High Speed Digital Systems.

14. SUBJECT TERMS
Interconnections, systems architectures, electronic, optelectronic, and optical interconnections

15. NUMBER OF PAGES
116

16. PRICE CODE

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

Program Committee

Workshop Chair
Howard Davidson
Sun Microsystems
Sunnyvale, CA

Program Chair
Philippe Marchand
UCSD
La Jolla, CA

Tutorials Chair
George Papen
University of Illinois
Urbana, IL

Working Group Chair
Ashley Saulsbury
Sun Microsystems
Sunnyvale, CA

International Liasons
Peter DeDobbelaere
Akzo Nobel
Sunnyvale, CA

Henk Neefs
University of Gent
Gent, BELGIUM

Osamu Wada
Fujitsu Laboratories
Atsugi, JAPAN

Technical Program Committee

Marc Christensen
George Mason University, Fairfax, VA
Kirk Giboney
Hewlett Packard Laboratories, Palo Alto, CA
Anthony Lentine
Lucent Technologies Bell Laboratories, Holmdel, NJ
John Levy
Cisco, San Jose, CA
Tulin Mangir
TM Associates, Santa Monica, CA
John Poulton
University of North Carolina, Chappell Hill, NC
Harold Stone
NEC Research Center, Princeton, IL

Working Group Committee

Lew Aronson
Hewlett Packard Laboratories, Palo Alto, CA
Giorgio Giaretta
Lucent Technologies Bell Laboratories, Holmdel, NJ
Charles Kuznia
University of Southern California, Los Angeles, CA
Rick Lytel
Sun Microsystems, Sunnyvale, CA
Henk Neefs
University of Gent, Gent, BELGIUM
Steve Tam
Cisco, San Jose, CA

Workshop Scope

The continuing rapid increase in the performance of high speed electronics and communications technologies has led to dramatic improvements in advanced computing and communications systems. The rapid growth of computer internetworking and the rise of new applications such as multimedia and virtual reality are driving the requirements for still higher levels of computing and communications. Interconnections within digital computing and switching systems today are often perceived as a performance bottleneck. The purpose of this Workshop is to determine the interconnection requirements of emerging and future computer and communications systems and to disseminate information about state-of-the-art optical and electrical interconnection technologies at the component, packaging, and systems level.

Because of the multi-disciplinary nature of these problems, this Workshop brings together researchers and engineers with expertise in a variety of fields including electronic, optoelectronic, and optical interconnection technologies, advanced systems architectures as well as the systems level perspective of algorithms and applications. The Workshop is comprised of tutorials and invited talks of the highest caliber as well as a few contributed papers. In addition, all attendees participate in smaller working groups to discuss and address a central focus design problem. Working groups are diverse and multi-disciplinary. In the past, problems ranging from high-performance workstation design to tele-medicine applications have been considered. Historically, this workshop has provided a stimulating, highly interactive environment conducive to thought-provoking discussions. Take advantage of this opportunity to contribute to a great Santa Fe experience! More information can be found on the web at **http://soliton.ucsd.edu/ihsds/santafe99**

SUNDAY, 9 MAY 1999

3:00pm - 6:15pm Tutorial Session
Session Chair: George Papen, University of Illinois, Urbana, IL

3:00pm - 3:45pm Tutorial - 1
Device and Interconnect Technologies for ~100 GHz Mixed-Signal ICs, ***Mark Rodwell, UC Santa Barbara, Santa Barbara, CA***
160 Gb/s TDM optical links will require ICs with > 150 GHz analog bandwidth and a 80 or 160 GHz clock. Mixed-signal ICs (DACs/DDSs/ADCs) for digital processing of 2-20 GHz radar signals will have 2000-transistor complexity and ~100 GHz clock rates. To permit clock rates exceeding 100 GHz, transistor current-gain (ft) and power-gain (fmax) cutoff frequencies must be several hundred GHz. The interconnects must have small capacitance per unit length, and wire lengths, hence transistor spacings, must be small. Given that fast transistors operate at high current densities, effective heatsinking is essential. To prevent circuit-circuit interaction through common-lead inductance ("ground bounce"), low wiring ground-return inductance is required within the IC and between IC and package. We report a transferred substrate heterojunction bipolar transistor (HBT) IC technology providing scalable submicron HBTs with record 250 GHz ft and 820 GHz fmax. The interconnects, microstrip on a low-epsilon dielectric, have low capacitance and high velocity and a ground plane for low ground-return inductance. An electroplated Au/Ni/Cu metal substrate with Au thermal vias provides effective heatsinking. Demonstrated ICs include 85 GHz amplifiers and 60 GHz M/S latches. To manage power-delay products in larger circuits, low-voltage-swing (nkT/q) circuits are being investigated.

3:45pm - 4:30pm Tutorial - 2
Overview of Nonlinear Optics for High Speed Communication, *Bahaa Saleh, Boston University, Boston, MA*

4:30pm - 4:45pm Break

4:45pm - 5:30pm Tutorial - 3
Advances in Chip Level Packaging, *John Carson, Irvine Sensors Corporation, Costa Mesa, CA*
Two major directions in chip level packaging will be observed during the next decade: thinner packages (and therefore thinner chips) and more direct chip attach techniques. Package thickness will be pushed to as low as 0.5 mm for various applications enabled by agressive chip thinning techniques. In direct chip attach, peripheral leads in a footprint smaller than the IC carrier will appear in mainstream applications limited only by printed circuit board constraints. Combined, these two trends will drive toward increased use of three dimensional stacking techniques. Examples of thinned chips on flexible substrates and three dimensional assemblies of multi-chip packages are shown to portend these coming events.

5:30pm - 6:15pm Tutorial - 4
Modeling, Analysis and Simulation of Data Networks, *Yusuf Ozturk, San Diego State University, San Diego, CA*
The first topic will be more network modeling and analysis oriented. I can demonstrate some traffic collection tools and later incorporating the data collected into commercial simulation tools. We can work around practical problems for capacity planning and projections to the feature. This fits very good into a workshop program. This talk will reflect my experiences and common mistakes network managers and analysis specialists are doing during the data collection process , analysis and simulation of their network. This tutorial will be mostly a demonstration of network design process starting from data source characteristics, network topology selection, modeling and analysis.

6:30pm - 7:30pm Welcome Reception

8:00pm - 8:30pm
Kickoff Meeting for Working Group Leaders
Session Chair: *Ashley Saulsbury, SUN Microsystems, Mountain View, CA*

MONDAY, 10 MAY 1998

8:00am - 8:15am **Workshop Welcome**
Workshop Chair: *Howard Davidson, Sun Microsystems, Mountain View, CA*

Session: **Short Haul Interconnects**
Session Chair: *Kirk Giboney, Hewlett Packard Laboratories*

8:15am - 8:45am
1.1 Overview of 10Gbit Ethernet, *Peter Wang, 3COM Technology Development Ctr. Santa Clara, CA*
Internet traffic is exploding. Intranet, extranet, E-commerce and Voice-over-IP are all contributing to the growth of data networks. Gigabit Ethernet deployment is ramping up, as are broadband access networks. Carriers are planning for multi-gigabit backbone deployment. Dense Wavelength Division Multiplexing is the talk of the town. Are there alternatives? Is the world ready for 10 Gb/s networking?

This talk will explore the key enabling technologies and the various interconnect options for constructing 10 Gb/s links for the next generation backbone. We will also touch on the challenges of building switching infrastructure for the 10 Gb/s data networks.

8:45am - 9:15am

1.2 PAROLI a Synchronous Optical Interconnection Link with a Through Put of 13 Gbit/s, *Karsten Droegemueller, Siemens AG, GERMANY*

Data communication and telecom switching systems require interconnections with high density, high data throughput and low power consumption. The design, realization, and characterization of a multichannel parallel optical interconnection with a 12 fiber ribbon and with an optical data rate of 1,25 Gbit/s per channel is reported. Two versions will be presented. First, a bit synchronous link with an electrical interface consisting of 22 differential data channels operating at 500 Mbit/s each plus one clock channel. Second, an asynchronous link with 12 electrical differential data channels at 1,25 Gbit/s each. On the transmitter side a vertical-cavity surface-emitting laser (VCSEL) array is employed as light source. Results of reliability test of the VCSEL's are given in the presentation

9:15am – 11:30am

Session: Intra-System Interconnects

Session Chair: *Rick Lytel, Sun Microsystems, Mountain View, CA*

9:15am - 9:45am

1.3 Tb/s Chip I/O – How Close are we to Practical Reality?, *Rick Walker, Hewlett Packard Laboratories, Palo Alto, CA*

Computer and Router designers are counting on Tb/s chip-to-chip data transmission capability to continue expanding their system performances to meet the global demand.

Several prototype serial links, with clock and data recovery, have been published at 2-10 Gb/s data rates per pin. Much work is focussed on lowering the power and size of these links to allow hundreds of links to be integrated onto a single chip.

Even with these advances, some scary system issues still remain. Power supply noise can have disastrous effects on PLL and DLL performance. Signal crosstalk can close up an otherwise open eye. Each advance in CMOS scaling reduces the analog circuit options available to the link designer.

The copper signal-transmission infrastructure is not improving at anywhere near a "Moore's law" rate. FR4 has been the standard dielectric for high-density PCBs for over 20 years, and coax cables are an extremely mature art. Dielectric and skin loss limit data rates to approximately 10Gb/s, and further advances may be slow coming.

This talk will explore these trends and attempt to forecast the future of high-speed serial interconnects.

9:45am - 10:00am Coffee Break

10:00am - 10:30am

1.4 **Interconnect Requirements for Digital Cross-connect Systems,** *Roger Holmstrom and Robert Ward, Tellabs Operations Inc., Lisle, IL*

The requirements for high-speed and high-density board-to-board interconnects in digital cross-connect systems are such that new and emerging technologies are sought. These requirements are discussed in terms of physical, performance, reliability, and cost metrics. Some alternatives are evaluated. For long reach interconnects, parallel optics are favored. For short reach interconnects, electrical interconnects are chosen.

10:30am - 11:00am

1.5 Moore's Law: The Intra-system I/O Challenge, *Craig Theorin, W. L. Gore & Associates, Lompoc, CA*

The modularity of recent high speed digital system designs have created the need for intra-system I/O bandwidth in excess of 10 Gbit/sec. Most system architects anticipate this bandwidth requirement to scale with Mooreís law for the foreseeable future, creating a substantial signal integrity challenge for future data links. We will describe the chip-to-chip signal integrity concerns and likely solutions for intra-system I/O in the early part of the next millennium as aggregate bandwidths scale beyond 100 Gbit/sec.

11:00am - 11:30am

1.6 DDR and RAMBUS (High Speed Bus) DRAMD, *Mian Quddus, Samsung, KOREA*

11:00am - 11:30am

Workshop Problem Statement for the 1998 Workshop:

Ashley Saulsbury, Sun Microsystems, Mountain View, CA

12:00pm - 1:30pm Luncheon & Working Group Session I

1:30pm - 3:30pm Working Group Session II

3:30pm - 6:30pm Free Afternoon

6:30pm - 7:30pm Reception

8:30pm - 9:30pm Special Event 10 Years of Santa Fe Experience

Speaker: *Harold Stone, NEC, Princeton, NJ*

TUESDAY, 11 MAY 1998

8:15am - 9:45am
Session: **Optical Interconnects for High-performance Computing Systems**
Session Chair: ***Harold Stone, NEC, Princeton, NJ***

8:15am - 8:45am
2.1 Interconnects in Scalable, Distributed Mulitprocessor Systems, *Jeffrey Kuskin, Silicon Graphics, Inc, Mountain View, CA*
Communication among processing nodes (that is, CPUs, memories, and I/O devices) is perhaps the key component in the design of a multiprocessor system. Traditionally, multiprocessors have been constructed by connecting a small number of processing nodes to a common, shared bus. The shared bus provides not only a mechanism for the processing nodes to communicate, but also allows all communication to be broadcast to all nodes on the bus.

The broadcast capability of a shared bus greatly simplifies the overall system design. Unfortunately, electrical and mechanical constraints severely limit the number of nodes that a single shared bus can support. For this reason, multiprocessors that scale to large numbers of processing nodes do away with the single shared bus and instead employ a distributed system design in which processing nodes are interconnected via a high-bandwidth, low-latency, switched routing fabric.

The use of a routing fabric overcomes the scalability limitations of a shared bus, but introduces a number of complications of its own. This talk will explore the use of high-performance interconnects in a distributed multiprocessor system. We begin with a short discussion of the basic distributed multiprocessor node architecture and interconnection fabric, and the difficulties that such an architecture creates. We then describe how these problems are solved in practice, with an emphasis on the role of the interconnection network. We conclude with some thoughts on the increasing importance of communication in multiprocessor system designs and the demands that will be placed on future multiprocessor interconnection networks.

8:45am - 9:15am
2.2 The Role of Optics in Balanced Computer System Design, *Mike Chastain, Hewlett-Packard Company, Richardson, TX*
Computer system architects have been waiting and watching the development of parallel optics for five years or more, hoping that breakthroughs in the producibility, pack-aging, and resultant costs would finally make optical links cost competitive with their cop-per counterpoints. The inherent advantages of optical interconnects are well known. The inherent reduction in physical size of both connectors and cables, increased usable com-munication distance, and reduced susceptibility to EMI and EMC have always been appealing; but the costs have always forced designers to do it in copper just one more time.

In the last few years we have all witnessed the almost exponential climb in CPU clock rates, soon to break the gigahertz barrier, and we have seen virtually every performance feature ever implemented in the fastest supercomputers migrate to single chip CPUs. Along with these improvements has come an equally impressive increase in the data band-widths required to keep these CPUs in execution. Soon we may see single chip CPUs capable of consuming 10 Giga-bytes per second or more.

These enormous bandwidths are forcing CPU and ASIC designers alike to push every integrated circuit connection to the maximum frequency in order to maintain pin counts at manufacturable levels. Intelís recent switch to Rambus DRAM is a clear indication that all vendors, even PC vendors, are faced with this problem.

As we increase the frequency of these products we will test the limits of printed circuit technology. Skin effect losses are already a problem, and within a few years these loses will be replaced in priority by dielectric losses, perhaps limiting the usable connection dis-tance to a single backplane or PC planar. Cables with very good dielectrics will, of course, allow longer distance; but there will always be copper trace in the path. As frequencies increase, every 4 to 5 inches of copper trace will reduce the usable cable length by about three feet in addition to decreases due to the cable dielectric losses. We may do well to connect adjacent racks with copper cables, let alone cross machine rooms.

So it appears, within a few years, that copper interconnects of more than a few meters could easily become very expensive, while VCSEL technology and creative packaging may finally yield cost effective parallel optic interfaces.

The inevitable shift to parallel optical technology may occur because of this juncture of over stressed copper and mass produced optics, but there is still a major disconnect between the future needs of computer systems and the roadmap of optical components. Today the optical roadmap is driven by the telecommunications industry, which seems to be increasing communication frequencies in a fixed ì4xî pattern from 622Mhz, to 2.5Ghz, to 10.0Ghz while the computer industry tends to take smaller ì2xî increments. The pre-ferred frequency pattern for the computer industry must include 2.5Ghz and 5Ghz. These frequencies are especially important for computer / optic integration because many technologist now believe the useful limit of printed circuit technology is around 5Ghz. Beyond this frequency, i.e. at 10Ghz, it may be impossible to build a reasonable size backplane.

9:15am – 9:45am
2.3 In Pursuit of a Petaflop: Overcoming the Latency/Bandwidth Wall, *Peter Kogge, Notre Dame University*
The fastest machines on the planet today peak at around a teraflop (10^12, floating point operations per second), with plans over the next few years to approach 10-30 TF. This performance, however, is still insufficient for several important classes of applications. Performance levels of a Petaflop (10^15 flops) thus become a valuable target to aim for. Unfortunately, achieving this with current conventional technology and architecture seems to be difficult, and destined to wait for the 2010-2015 timeframe.

The twin demons in this wall appear to be latency and bandwidth: getting enough data to the right processing logic in a timely enough fashion that the logic can be kept profitably busy, and doing so in a fashion that the the amount of parallelism that must be found in an application is acceptable.

This talk will address one approach to breaking this wall: the HTMT project (Hybrid Technology MultiThreaded acrhitecture). This multi-institution collaboration is in the middle design phase of a long-term effort started in 1994 to find alternatives to conventional architectures and relevant technologies, and if successful, will result in a petaflops level machine by around 2006.

The solutions used by HTMT encompass both technology and architecture. In technology, superconducting logic, very fast WDM all optical networks, Processing-In-Memory (PIM), and 3D holographic storage form the basic underpinnings for a radically different machine. In architecture, multithreading, active memory, and automatic percolation of data throughout a very deep memory hierarchy all are central players.

This talk will overview the inherent problems associated with achieving a petaflops, and discuss the architecture of the current HTMT design. Although all aspects of the machine will be discussed, emphasis will be placed on the active memories, where PIM technology coupled with the concepts of percolation, allow massive parallelism in the memory system to execute large portions of an application in ways that defeat the bandwidth/latency barriers formed by conventional approaches.

9:45am – 10:15am
2.4 Ultra-High Speed Optical Interconnection Network for Supercomputing, *Keren Bergman, Princeton University, Princeton, NJ*
In an attempt to effectively utilize the immense bandwidth of optical fiber interconnects, we designed a completely novel network architecture specifically for optical implementation. This work is part of an aggressive multidisciplinary architecture study of the next generation high performance computing based on hybrid technologies and multi-threaded (HTMT) latency management. The optical network employs multiple node levels with a routing topology that is based on a minimum logic at the node scheme. Our architecture features radically new traffic control logic, having the property that all routing decisions for the self-routing data packets, are based on a single logic operation at each node. The optical network, named the Data Vortex, can scale to interconnect an ultra-high performance computing system in a massively parallel form. Within the framework of the Data Vortex network we are investigating enabling fiber optic technologies and an implementation that consists of fiber interconnects with wavelength division multiplexed and time division multiplexed (WDM/TDM) payload and header. The development and incorporation into the network of fiber optic modules including high speed fiber lasers, amplifiers, and switching nodes will be discussed in this talk.

10:15am – 10:30am Coffee Break

10:30am - 12:00noon
Session: Optical Networks
Session Chair: *Tulin Mangir, TM Associates, Santa Monica, CA*

10:30am - 11:00am
2.5 Ultrafast Optical Interconnect Based on Routing by "Clockwork" in Regular Mesh Networks, *David Cotter, British Telecom, UK, F. Chevalier, and D. Harle University of Strathclyde, UK*
The effectiveness of multi-processor systems (such as future massive-capacity routers and servers) is critically dependent on the speed and efficiency of interconnection. Full connectivity is required with large message throughput and minimal delay. An option under consideration is an ultra-high speed multi-stage packet-switched network, using fixed-length packets at serial bit rates of 0.1-1 Tbit/s. The packets are routed through the network on optical pipes, with >>routing and digital header processing (such as destination address recognition) performed 'on the fly' in the optical domain.

A key requirement for high performance is that the routing mechanisms and processing at network nodes should be as simple as possible. Here we describe a new strategy for routing in regular mesh interconnection networks, based on a method of automatic global scheduled ('clockwork') switching in the optical domain. Using this strategy, the intermediate routing nodes are merely needed to perform an extraordinarily simple function ('for-me-or-not-for-me' header-address recognition), otherwise traffic is routed onwards automatically in the optical domain with absolutely no further intelligent action performed by the node. The throughput is comparable with conventional store-and-forward packet switching, yet the simplicity of this strategy makes it suitable for implementation in digital optical logic. The clockwork approach enables some special capabilities-such as ultra-low latency signalling, bandwidth reservation, ultra-low response delay, and process scheduling.

11:00am - 11:30am
2.6 Large-scale photonic packet switch using wavelength routing techniques, *Koji Sasayama, NTT Network Innovation Laboratories, Kanagawa, JAPAN*
This talk describes the large-scale photonic packet switching system being developed in NTT Laboratories. It uses wavelength-division-multiplexing (WDM) techniques to attack Tbit/s-class throughput. The architecture is a simple star with modular structure and effectively combines optical WDM techniques and electronic control circuits. Recent achievements in important key technologies leading to the realization of large-scale photonic packet switches based on the architecture are described. It is confirmed that a 320-Gbit/s system can tolerate the polarization and wavelength dependencies of optical devices. Experiments using rack-mounted prototypes demonstrate the feasibility of the architecture. The experiments showed stable system operation and high-speed WDM switching capability up to the total optical bandwidth of 12.8 nm, as well as successful 10-Gbit/s 4 x 4 broadcast-and-select and 2.5-Gbit/s 16 x 16 wavelength-routing switch operations.

11:30am - 12:00noon
2.7 Latency and Scaling Issues in High-Speed Optical TDM Networks, *Paul R. Prucnal, Princeton University, Princeton, NJ*
An overview of optical TDM devices and techniques for ultra-high bit rate data communications is given as well as a discussion of the latency and scaling issues present in these systems.

12:00pm - 1:30pm Luncheon and Working Group Session III

1:30pm - 3:30pm Working Group Session IV

3:30pm - 7:00pm Free Afternoon

7:00pm - 8:00pm
Session: Optoelectronic and Optical Technologies
Session Chair: *Marc Christensen, George Mason University, Fairfax, VA*

7:00pm - 7:30pm
2.8 The Commercial Applications of Optoelectronics, A View from the Optoelectronics Industry Development Association (OIDA), *Arpad Bergh, OIDA, Washington, DC*
The Optoelectronics Industry Development Association (OIDA) was formed in 1991 to advance the worldwide competitiveness of the North American optoelectonics industry and to promote the application of optoelectronics technology. The OE industry is a collection of six ore more distinct industries that all depend on OE technology. This fragmentation represents major challenges and opportunities.

It is difficult to draw a technology roadmap that serves all applications. On the other hand, there are great opportunities to share a common infrastructure that can advance a number of non-competing industries. Over the past eight years OIDA had carried out over thirty market survey and technology roadmap activities to identify emerging markets and shortcomings in domestic technology. Industry wide consensus was developed through informal interactions and recommendations were presented to industry and government for action.

The most prevalent impediments identified in these studies are the exploration of new markets for OE enabled applications and the ability to conduct high volume, low cost manufacturing. This talk will describe some of the initiatives that OIDA has undertaken to overcome these deficiencies.

7:30pm - 8:00pm
2.9 Board and Back-plane Level Optical Circuits Using Integrated Thin-cladding Polymer Fibers, *Yao Li, Jan Popelek, and Jun Ai, NEC Research Institute, Princeton, NJ*
This talk summarizes recent research activities at NEC Research Institute on optical interconnections using integrated polymer fibers. The objective of the research is to study large-bandwidth, short-distance, packageable optical solutions to address future interconnection needs at circuit board and back-plane levels. We have studied possibility of using embedded polymer fibers to form a 10 GHz board-level optical clock distribution circuit and demonstrated the feasibility of highly efficient and uniform delivery scheme for up to 128 optical termination's. Specialty thin-cladding polymer fiber bundles were integrated into convention PCB's. Various performance data will be presented. We also extended this embedding concept to include polymer fiber image guides (PFIG's), a cost-effective 2D image transmission components. We have fabricated some packaged and connectorized board-level optical circuits to perform point-to-point 2D parallel optical interconnects for future 2D vertical-cavity surface-emitting laser (VCSEL) and optical detector array based optical interconnects. Among demonstrated are some 16 node (6x6 bits/node) optical shuffle and butterfly interconnect circuits using three-layers of FIG embedding. Low insertion loss (< 2 dB) ad moderate resolution (11 lp/mm) were obtained. To further extend the capability of these 2D parallel optical circuits, we are experimenting a hybrid integration of these PFIG's and free-space micro-optic components so that branching and add/drop capabilities at different optical nodes can be incorporated.

8:00pm – 8:30pm
2.10 Development of Monolithically Integrated Transceivers for Single-and Multi-Channel Fiber-Based Optical Interconnects, *Clifton G. Fonstad, Jr., and Joseph F. Ahadian, Massachuestts Institute of Technology, Cambridge, MA*
The Epitaxy-on-Electronics (EoE) optoelectronic integration technology, in which optoelectronic device heterostructures are grown epitaxially on fully processed GaAs MESFET electronic circuits, has produced uniquely complex monolithic OEICs combining optical emitters and detectors with high-speed VLSI electronics. This paper describes the development of transceivers for fiber-based optical interconnnects using the EoE technology. Recent progress toward implementing the EoE technology with silicon CMOS electronics and more advanced GaAs technologies will also be reviewed.

WEDNESDAY, 12 MAY 1998

8:00am - 9:30am
Session: **Working Group Solution Presentations**
Session Chair: *Ashley Saulsbury, Sun Microsystems, Mountain View. CA*

9:30am - 9:45am **Coffee Break**

9:45am - 11:15am
Session: **Plenary**
Session Chair: *Philippe Marchand, UCSD, La Jolla, CA*

9:45am - 10:30am
3.1 **Java, Jini and High Speed Systems of the Future,** *Bill Joy, SUN Microsystems, Aspen, CO*
Until roughly 1980, high performance systems were built of multiple boards, and organized around a disk operating system. Sun's Solaris and SPARC based products have this form, and applications like Oracle and SAP are focused around management of the information on the disk.

With the emergence of the internet in the last 20 years, systems are more and more often built on networking as the interconnect, often now with TCP/IP playing the role that the disk operating system did, providing the basic interconnect primitives. Sun's work with AOL and Netscape is an example of a major activity for us in this area, defining e-commerce as services relative to the interconnect.

In the future we believe that there will be a third organization for computing systems, those organized around objects and agents. We have built the Java and Jini technologies to support these new kinds of systems.

This talk will discuss these three organizing principles for computer systems (disks, internetworks and objects) and the implications for systems design.

11:15am - 11:30am
Workshop Summary
Matthew Goodman, Bellcore, Red Bank, NJ

Sunday,
9 May 1999

3:00pm - 3:45pm
Sun, 9 May - Tutorial 1

1999 IEEE Workshop
Interconnections within High-Speed Digital Systems

Device and Interconnect Technologies for ~100 GHz mixed-signal ICs

Mark Rodwell
University of California, Santa Barbara

rodwell@ece.ucsb.edu 805-893-3244, 805-893-3262 fax

Device and Interconnect Technologies for ~100 GHz mixed-signal ICs

Two topics:

*ICs *for* high-frequency interconnects*
RF/wireless, optical fiber
*ICs *needing* high-frequency interconnects*
100 GHz digital logic, GHz ADCs/DACs

The organization:

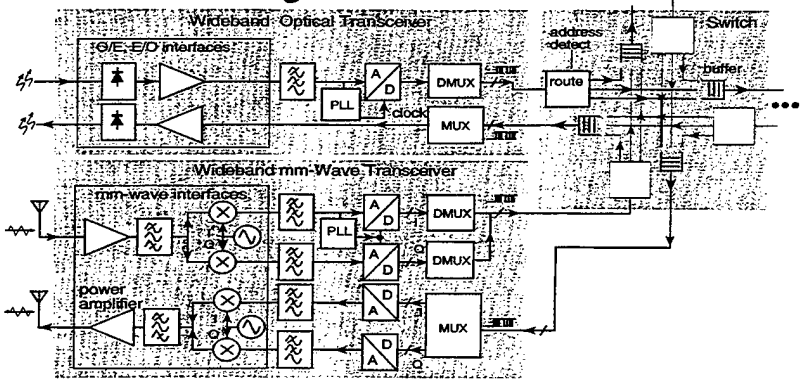
what are the future applications ?
what are the requirements ?
what is the state of the art ?
challenges for future high speed ICs
...and how my group is attacking them

Applications

*ICs "**for**" high-frequency interconnects*

*ICs "**needing**" high-frequency interconnects*

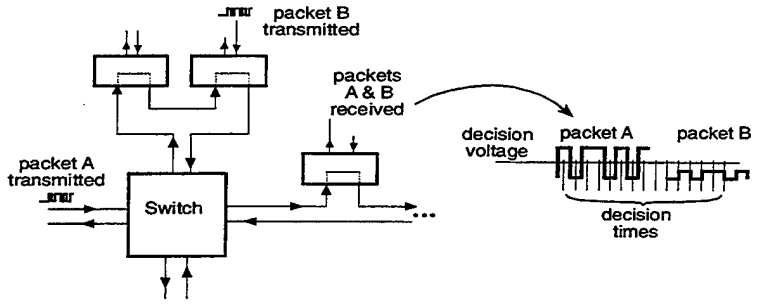
Electronics for GigaHertz Communication



Transceivers: *very fast digital & mixed-signal ICs*
Interfaces: *very wideband analog circuits, optoelectronics, mm-wave power*
Switches: *~10 GHz fast complex digital ICs*

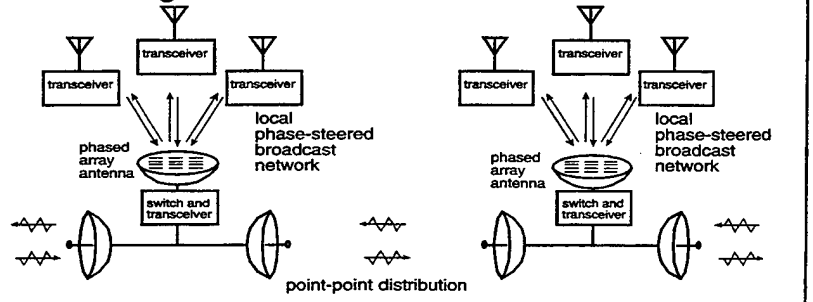
Why electronic switching remains important

Packet switching in a transparent (optical) switch



Burst errors will arise due to timing and amplitude glitches
Fix with digital (electronic) regeneration at switch -> all digital network

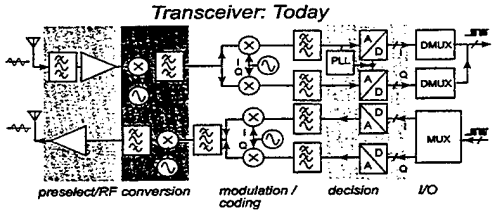
Wireless Digital Transmission: Networks & Distribution



Point-point links: 80 GHz, ~200 GHz line-of-sight
capacities of 10s of Gb/s
Broadcast links: 60 GHz, 120 GHz, ...
phased-array beamsteering
CDMA spread-spectrum coding for multipath

RF/Microwave ADCs/DACs/DDS: Towards the "Software Radio"

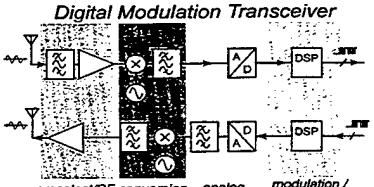
Transceiver: Today



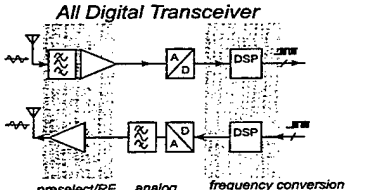
Advantages
digital robustness, frequency agility
complex modulation (spread spectrum...), simultaneous operation on many bands

Challenges
*ADC needs **enormous** dynamic range*
DAC also must have very high SNR
Enormous data reduction in DSP: very complex, very fast logic

Digital Modulation Transceiver

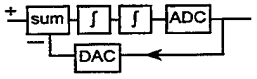


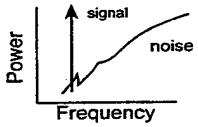
All Digital Transceiver



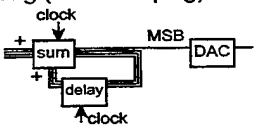
High resolution ADCs/DACs need high IC speed

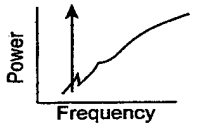
Delta-Sigma (noise shaping) ADC





Interpolating (noise shaping) DAC





ADCs/DACs for radio:
high dynamic range required (10-18 bits)
Oversampling ADCs/DACs:
high resolution obtained through high oversampling

Microwave ADCs need very fast logic, very fast transistors

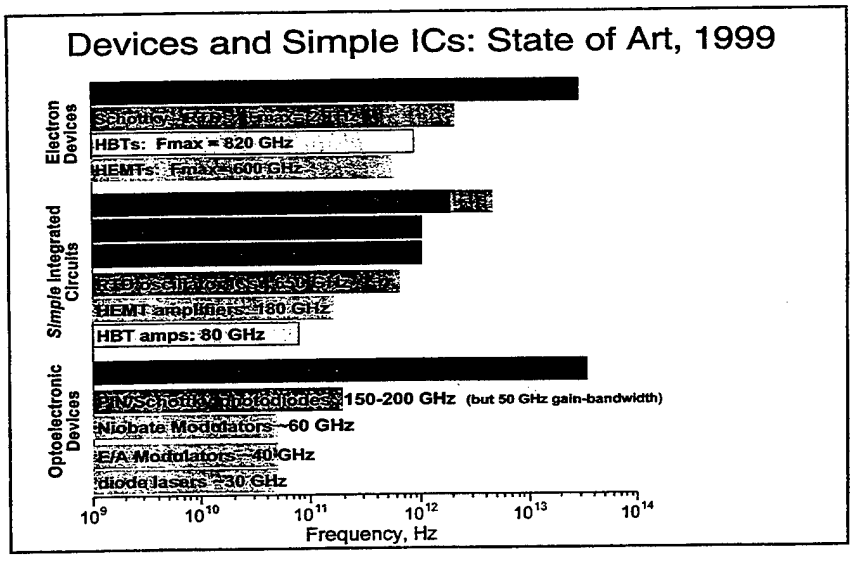
Requirements: 100 GHz clock-rate logic

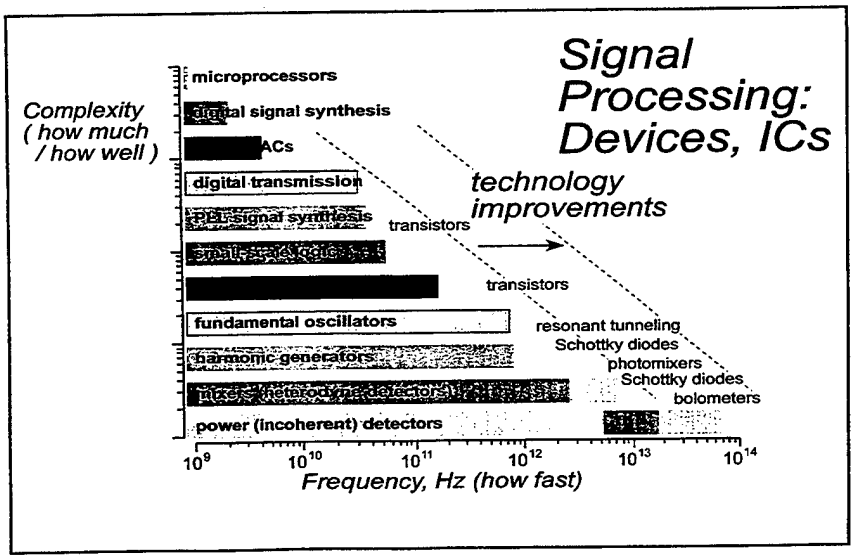
Fast transistors:
ADCs etc need very high ratio of transistor to signal bandwidth

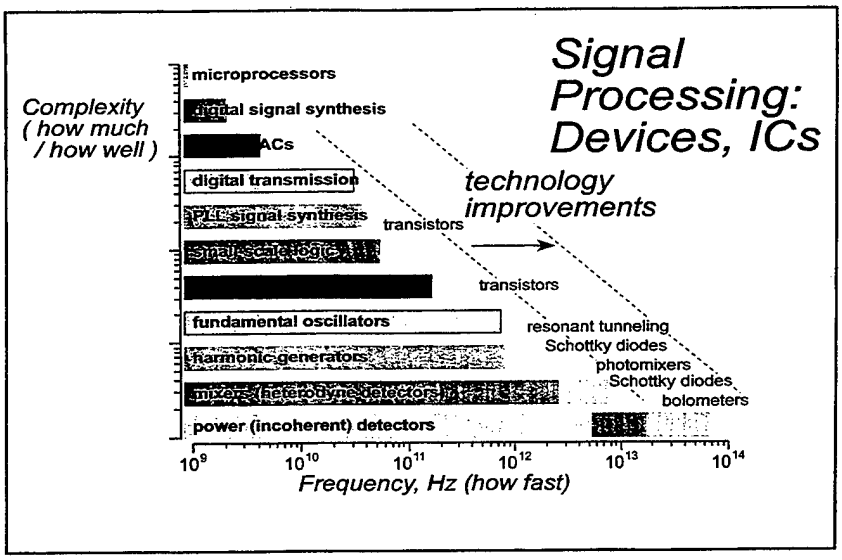
High performance wiring :
millimeter-wave bandwidths with analog & digital signals !
microstrip-lines and ground-planes for signal integrity
power delay products and impact of wiring

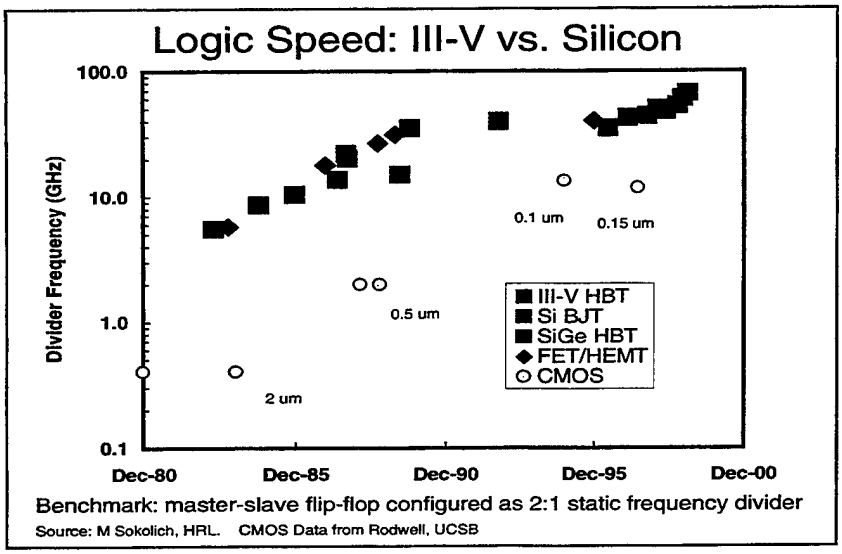
Outstanding heatsinking:
clock rates will be very high, so wiring delays must be small
transistors must be close together !
high performance transistors use high power densities !
power density on die may approach 1 kW/cm^2 !

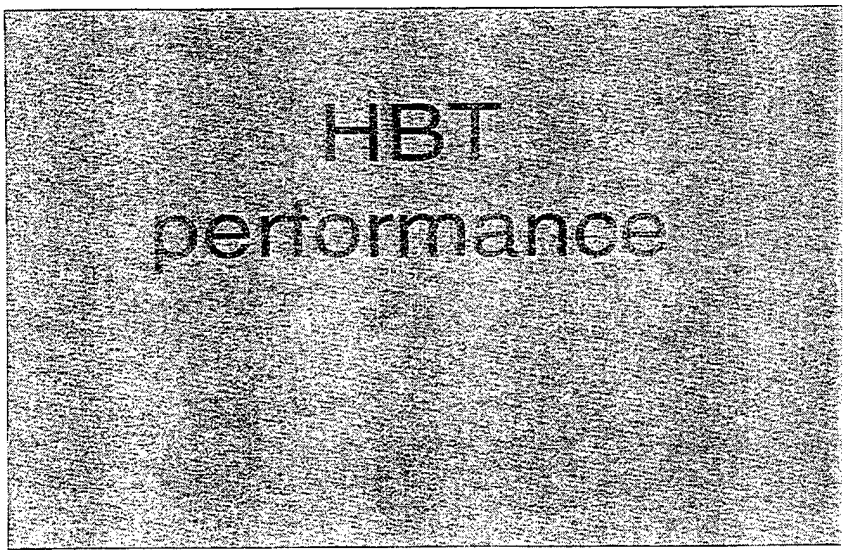
state of the art



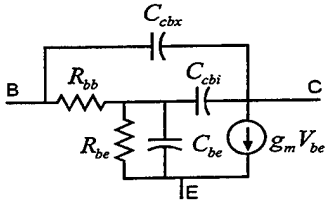



Signal Processing: Devices, ICs
Complexity (how much / how well)
microprocessors
digital signal synthesis
ACs
digital transmission
PLL signal synthesis
small-scale logic
transistors
fundamental oscillators
harmonic generators
mixers (heterodyne detectors)
power (incoherent) detectors
technology improvements
transistors
resonant tunneling
Schottky diodes
photomixers
Schottky diodes
bolometers
10^9 10^{10} 10^{11} 10^{12} 10^{13} 10^{14}
Frequency, Hz (how fast)


Logic Speed: III-V vs. Silicon
Divider Frequency (GHz)
100.0
10.0
1.0
0.1
Dec-80
Dec-85
Dec-90
Dec-95
Dec-00
2 um
0.5 um
0.1 um
0.15 um
III-V HBT
Si BJT
SiGe HBT
FET/HEMT
CMOS
Benchmark: master-slave flip-flop configured as 2:1 static frequency divider
Source: M Sokolich, HRL. CMOS Data from Rodwell, UCSB


HBT
performance

Bandwidth of Bipolar Transistors



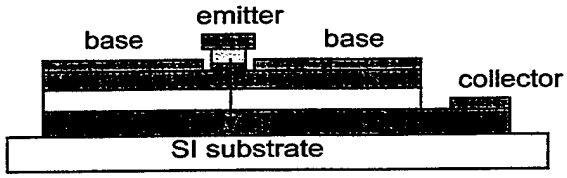
$$f_\tau = \frac{(1/2\pi)}{\tau_{base} + \tau_{collector} + (C_{je} kT/qI_e)}$$

$$f_{max} = \sqrt{\frac{f_\tau}{8\pi R_{bb} C_{cbi}}}$$

f_τ , f_{max} , and C_{cbx} are all important for high - speed circuits

$R_{bb}C_{cbi}$ and $R_{bb}C_{cbx}$ must be reduced

Current-gain cutoff frequency in HBTs

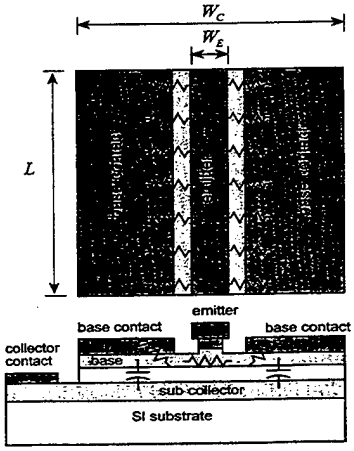


$$\frac{1}{2\pi f_\tau} = \tau_{base} + \tau_{collector} + C_{je}\frac{kT}{qI_E} + C_{bc}\left(\frac{kT}{qI_E} + R_{ex} + R_{coll}\right)$$

$$\tau_{base} \approx T_b^2/2D_n \qquad \tau_{collector} \approx T_c/2v_{sat}$$

Collector velocities can be high: velocity overshoot in InGaAs
Base bandgap grading reduces transit time substantially
RC terms quite important for > 200 GHz ft devices

Fmax in Double-Mesa HBTs



$$R_{bb} = \frac{1}{2L}\rho_{contact,horizontal} + \frac{\rho_{sheet}}{12L}W_E$$

$$C_{cb} = \frac{\varepsilon L}{T_c}W_c$$

Scaling emitter width does reduce base spreading resistance.

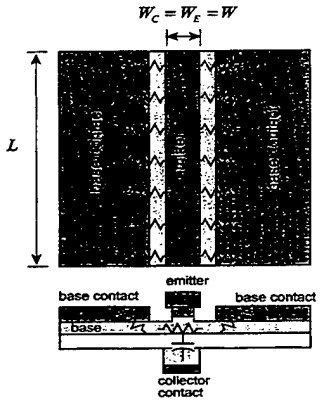
– but –

Minimum base resitance set by base contact resistance.

Minimum collector capacitance set by minimum base contact size

transferred-substrate HBTs

Fmax in Transferred-Substrate HBTs



$$R_{bb} = \frac{1}{2L}\rho_{contact,horizontal} + \frac{\rho_{sheet}}{12L}W$$

$$C_{cb} = \frac{\varepsilon L}{T_c}W$$

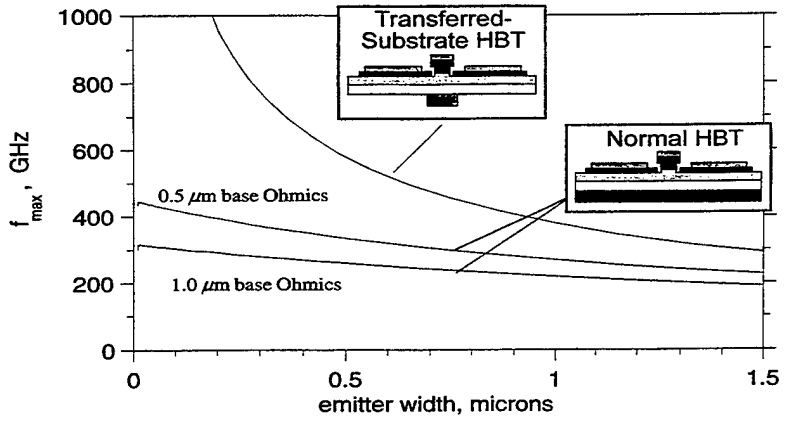
$R_{bb}C_{cb}$ reduces rapidly with deep submicron scaling

Component due to contacts scales as W^1

Base spreading component scales as W^2

F_{max} increases rapidly with deep submicron scaling

Transferred-Substrate HBTs: A ***Scalable*** HBT Technology



- Collector capacitance reduces with scaling: $C_{cb} \propto W_e$
- Bandwidth increases rapidly with scaling: $f_{max} \propto \sqrt{1/W_e}$

Transferred Substrate HBT Process

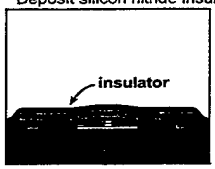
Objectives:

- 1000 GHz transistor bandwidth
- Thermal management for high power density
- Low wiring & packaging parasitics at 100+ GHz

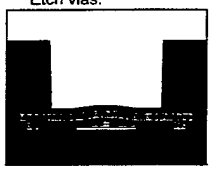
Approach:

- BCB process: standard IC materials
- Metal substrate, thermal vias
- Microstrip wiring: ground vias backside ground plane ε_r=2.7: low capacitance

1) Normal emitter, base processes. Deposit silicon nitride insulator.

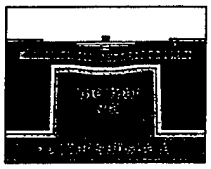


2) Coat with BCB polymer. Etch vias.

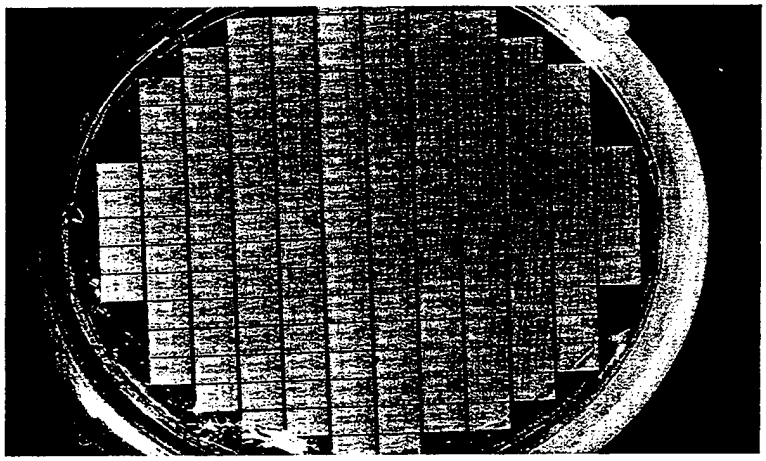


3) Electroplate with Au/Ni/Cu. Wax to carrier substrate.

4) Invert wafer. Remove InP substrate. Deposit collector. Demount from wax.

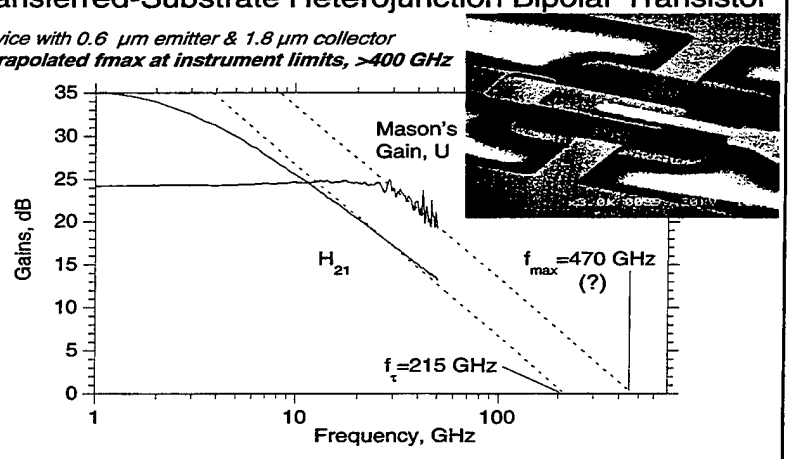


Transfer of Entire 2" HBT MMIC Wafer



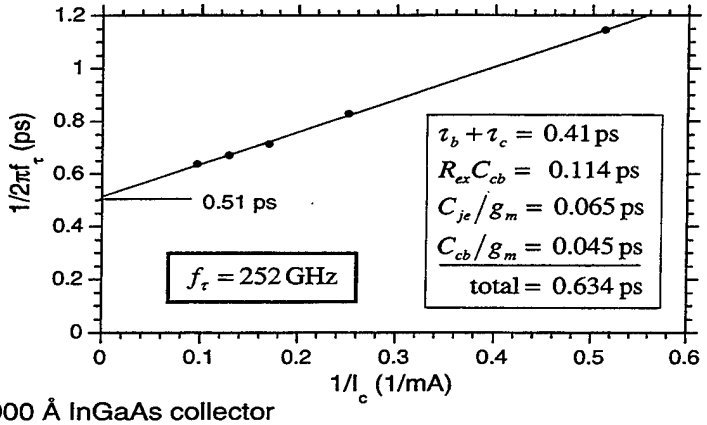
Transferred-Substrate Heterojunction Bipolar Transistor

Device with 0.6 μm emitter & 1.8 μm collector
extrapolated fmax at instrument limits, >400 GHz



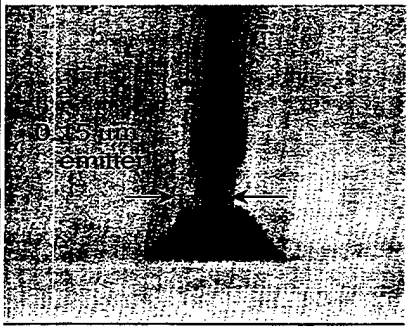
0.25 μm devices should obtain ~1000 GHz fmax

Transit times: HBT with 2kT base grading

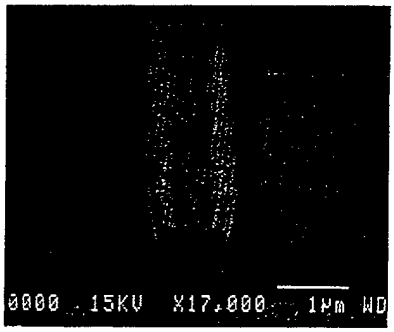


2000 Å InGaAs collector
400 Å InGaAs base, 2kT bandgap grading

SEM Photomicrographs of Deep-submicron HBTs

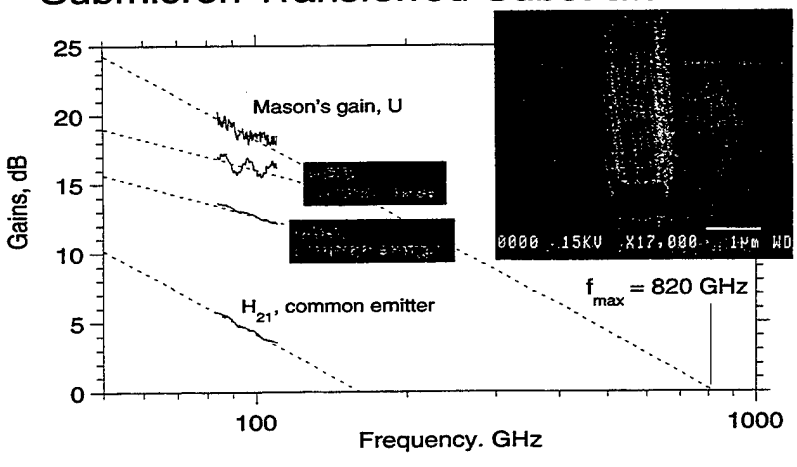


0.15 μm emitter base junction



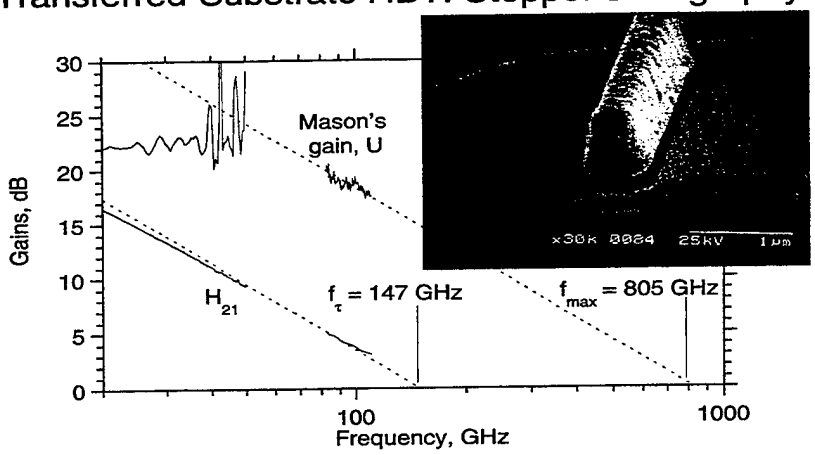
0.4 μm collector

Submicron Transferred-Substrate HBT



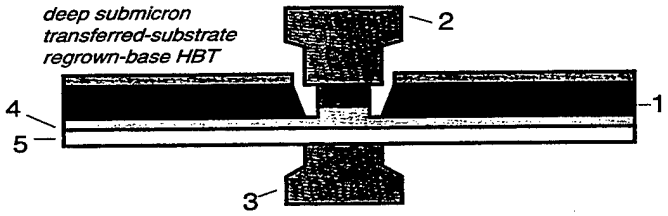
0.4 μm x 6 μm emitter, 0.4 μm x 10 μm collector

Transferred-Substrate HBT: Stepper Lithography



0.4 µm emitter, ~0.7 µm collector

Proposed THz-Bandwidth HBT ?



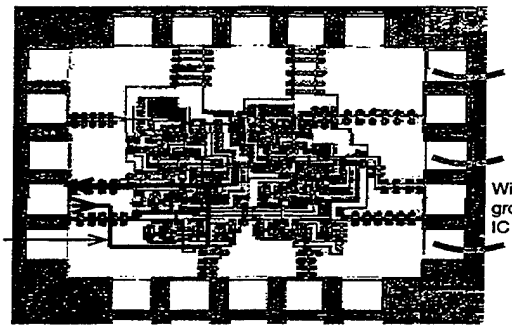
1) regrown P+++ InGaAs extrinsic base --> ultra-low-resistance
2) 0.05 µm wide emitter --> ultra low base spreading resistance
3) 0.05 µm wide collector --> ultra low collector capacitance
4) 100 Å, carbon-doped graded base --> 0.05 ps transit time
5) 1kÅ thick InP collector --> 0.1 ps transit time.

Projected Performance:

Transistor with 500 GHz ft, 1500 GHz fmax

The wiring environment for 100 GHz ICs

Why is Improved Wiring Essential?

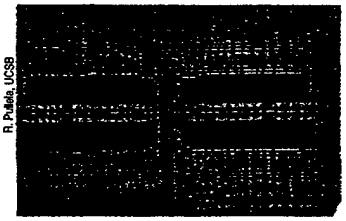


30 GHz M/S D-FF in UCSB - mesa HBT technology

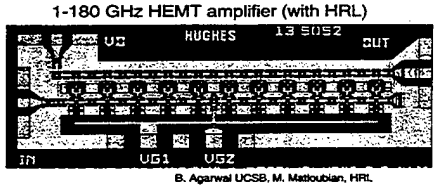
Ground loops & wire bonds:
degrade circuit & packaged IC performance

> 100 GHz CPW ICs: *severe crosstalk & ground bounce*

4-channel 100 Gb/s diode-based DMUX

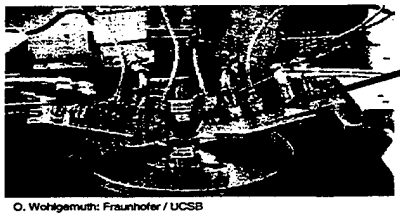


1-180 GHz HEMT amplifier (with HRL)



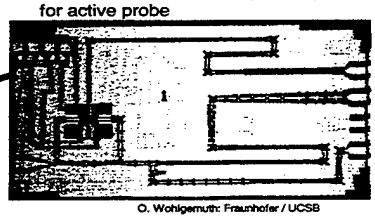
B. Agarwal UCSB, M. Matloubian, HRL

active probes for 70-220 GHz network analysis



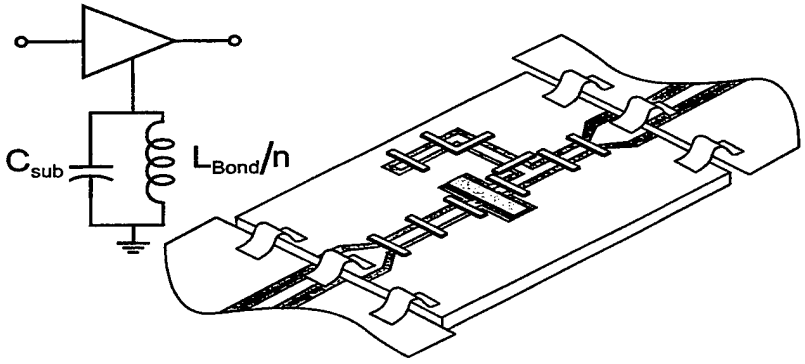
O. Wohlgemuth: Fraunhofer / UCSB

70-220 GHz network analyzer chip for active probe



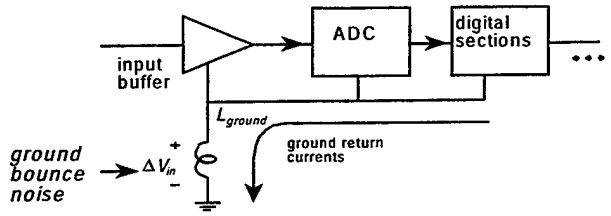
O. Wohlgemuth: Fraunhofer / UCSB

Coplanar Waveguide and Ground-Bounce



Bond wire inductance resonates with through-wafer capacitance at 5-20 GHz

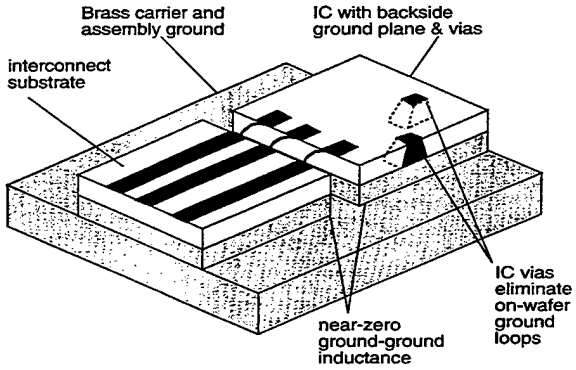
Ground Bound Noise in ADCs



Ground bounce noise must be ~100 dB below full-scale input
Differential input will partly suppress ground noise coupling
~ 30 to 40 dB common-mode rejection feasible
CMRR insufficient to obtain 100 dB SNR

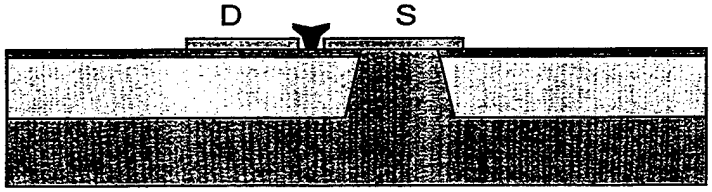
Eliminate ground bounce noise by good IC grounding

Microstrip IC wiring to Eliminate Ground Bounce Noise



Transferred-substrate HBT process provides vias & ground plane.

The microstrip via inductance problem



12 pH via inductance for 100 micron MIMIC substrate

j7.5 Ohms at 100 GHz, j15 Ohms at 200 GHz

A formidable difficulty for > 100 GHz IC design

At 100 um substrate thickness, via spacing must be > 100 um

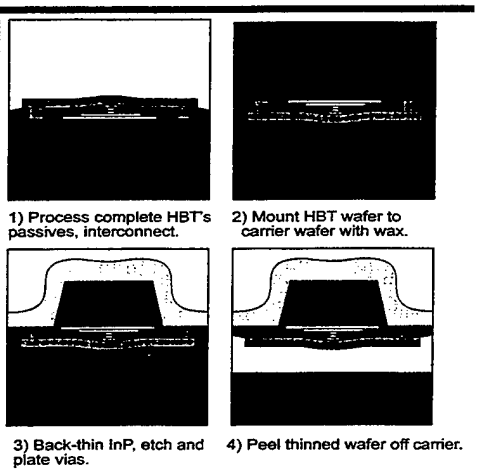
Solutions include "masterslice", flip-chip, substrate transfer

Standard MMIC Microstrip / Via Process

Objectives:
ground plane
microstrip wiring
low via inductance
avoid substrate modes

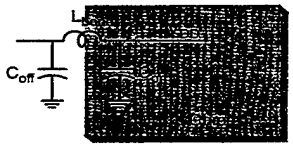
Approach:
the industry standard
frontside processing
wax mounting
wafer lapping
backside metal
wafer release

Limitations:
for 180 GHz must lap to 35 μm

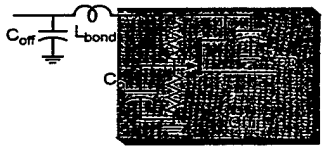


Power Supply Resonance

Resonates at $f = 1/2\pi\sqrt{L_{bond}C_{off}}$
gain peak / suckout, oscillation, etc.



Active (AC) supply regulation

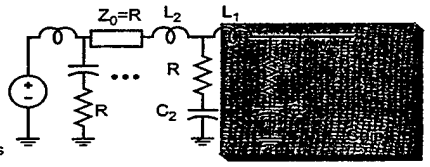


Passive filter synthesis

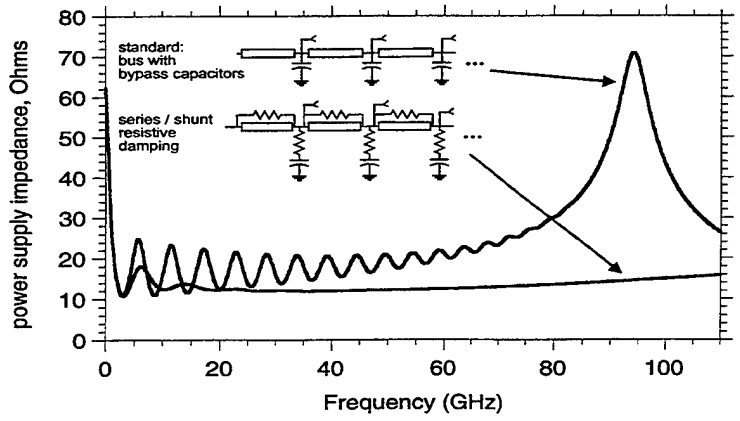
$R = \sqrt{L_1/C_1}$

$\sqrt{L_1/C_1} = \sqrt{L_2/C_2} = \bullet\bullet\bullet$

supply impedance is R at all frequencies

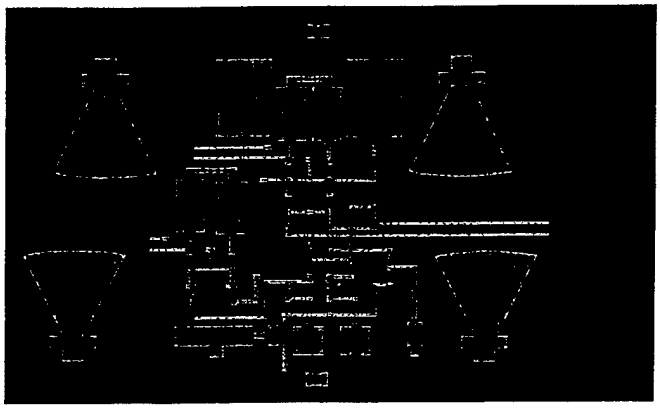


On-wafer power distribution for 100 GHz logic

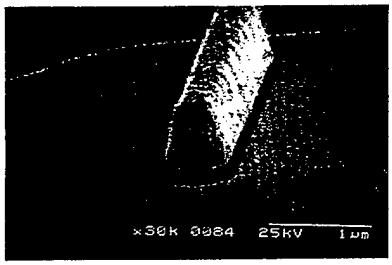


*supply **will** resonate: must prevent during design*

Standard cell showing power busses



Deep submicron HBT logic: ***low power ?***



Device sized for 100 Ω load:
(200 mV ECL logic swing)
0.15 μm x 6 μm emitter
peak speed at 2 mA bias

Shorter stripe length device:
0.15 μm x 0.5 μm emitter
peak speed at 150 μA bias

Small device is low-power but cannot drive 100 Ω line.
drives line with mismatched impedance: capacitance
lower power at higher (wiring-limited) delay

fast low-power logic requires low-voltage-swing-logic

Power-delay product in interconnect limit

$$P_{gate} T_{prop} = (1/2) C_{wire} V_{cc} \Delta V_{\text{logic}}$$

bipolar logic (static power)

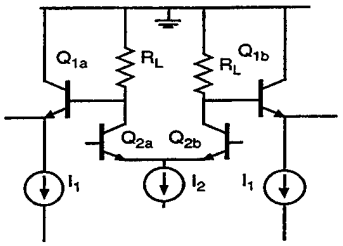
$$P_{gate} / f_{clock} = (1/2) C_{wire} V_{cc} \Delta V_{\text{logic}}$$

CMOS logic (dynamic power)

$(T_{prop} f_{clock})^{-1}$ ~ number gates between latches

For fast, low-power logic: reduce $V_{cc} \Delta V_{\text{logic}}$

The Interconnect Limit and Logic Gate Design



$$A_v = \frac{R_L}{2kT/qI_2} = \frac{I_2 R_L}{2kT/q} = \frac{\Delta V_{\text{logic}}}{2kT/q}$$

Voltage Gain must be >> 1

So: $\Delta V_{\text{logic}} = 10 \bullet (kT/q)$

But: $P_{gate} T_{gate} = (1/2) V_{cc} \Delta V_{\text{logic}} C_{wire}$

$$P_{gate} T_{gate} = (1/2)(1.5 \text{ Volt})(10 \bullet kT/q) C_{wire}$$

(power•delay) is constrained by interconnects

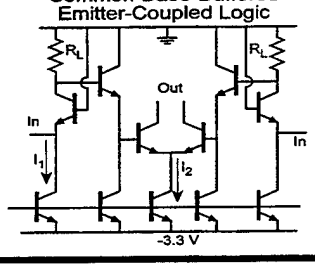
a fast transistor doesn't result in a fast IC

conclusion: a better circuit design is needed

Similar derivation for CMOS (Meindl, Proc IEEE, 1995)

Low-Voltage-Swing Logic Gates

Common-Base-Buffered Emitter-Coupled Logic



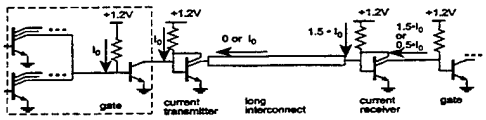
microwave DDS IC effort:
2000 HBTs @ 50-100 GHz clock
dissipation is severe issue

Solution (?):
small HBTs + low-voltage logic

Principle:
low-impedance input current buffer

Challenge:
not increasing transistor count

Current-Mirror-Buffered Integated-Injection Logic



Common Feature:

$$\Delta V_{\text{logic}} = \frac{kT}{q} \ln\left(1 + \frac{I_{switched}}{I_{bias}}\right)$$

Power Density in 100 GHz logic

Transistors tightly packed to minimize wire delays

10^5 W/cm^2 HBT junction power density.

~10^3 W/cm^2 power density on-chip

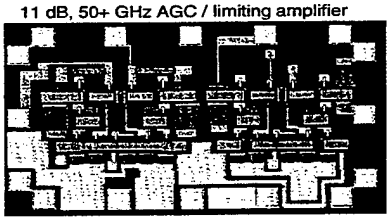
--> 75 C temperature rise in 500 μm substrate.

Solutions:
thin substrate to < 100 μm
replace semiconductor with metal

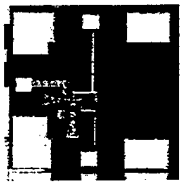
circuit results: transferred-substrate technology

Transferred-Substrate HBT Integrated Circuits

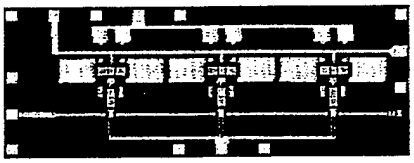
11 dB, 50+ GHz AGC / limiting amplifier



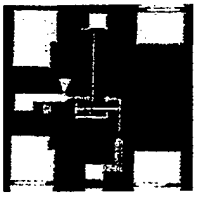
16 dB, DC-60 GHz amplifier



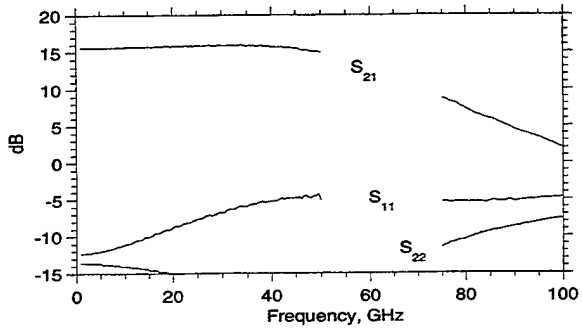
7 dB, 5-80 GHz distributed amplifier

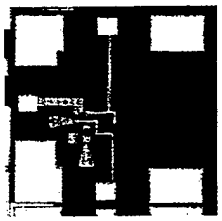


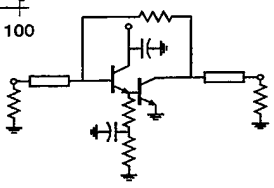
6.7 dB, DC-85 GHz amplifier



Darlington Amplifier - 360 GHz GBW

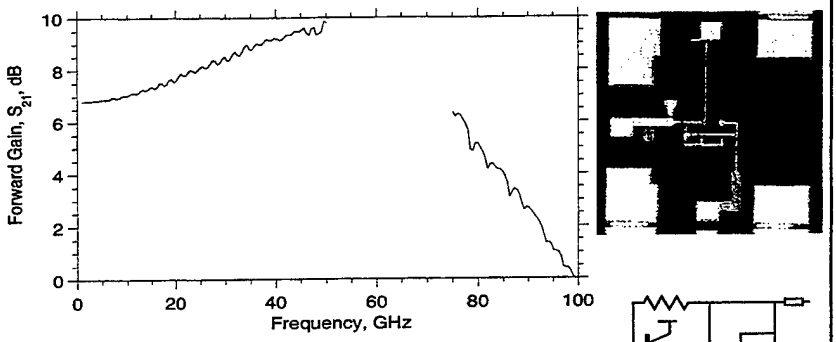


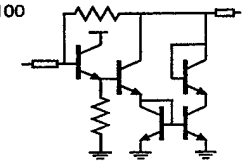




- 15.6 dB DC gain
- Interpolated 3dB bandwidth of 60 GHz
- 360 GHz gain-bandwidth product

6.7 dB, 85 GHz Mirror Darlington Amplifier





- 6.7 dB DC gain
- 3 dB bandwidth of 85 GHz
- f_τ-doubler (mirror Darlington) configuration

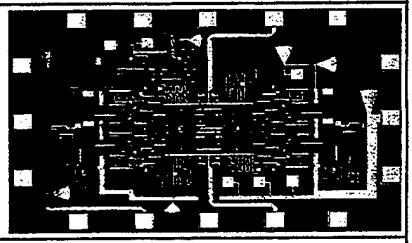
> 66 GHz HBT master-slave flip-flop

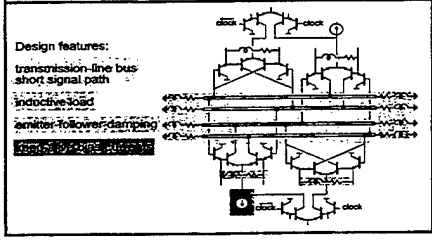
Objectives: 100 + GHz logic

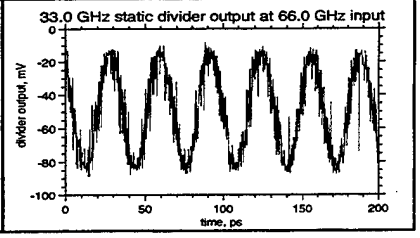
Approach:
transferred-substrate HBTs
efficient circuit design

Simulations:
95 GHz clock rate in SPICE

Measurements:
operation to 66 GHz limit of test setup
now building 75-110 GHz test setup







> 66 GHz HBT master-slave flip-flop

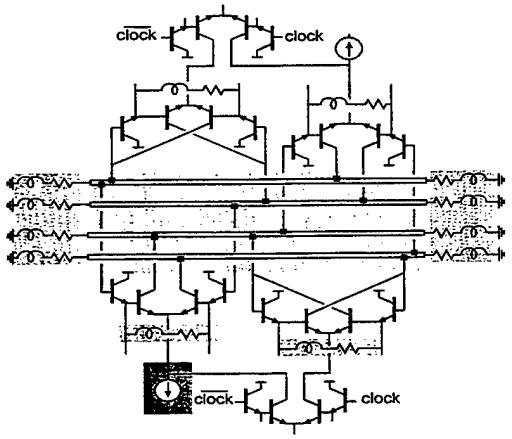
Design features:

transmission-line bus
short signal path

inductive load

emitter-follower damping

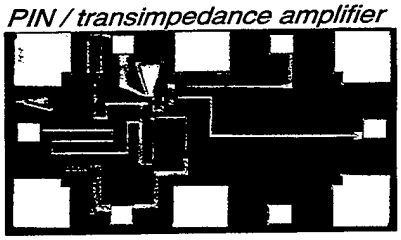
keep-alive bias currents



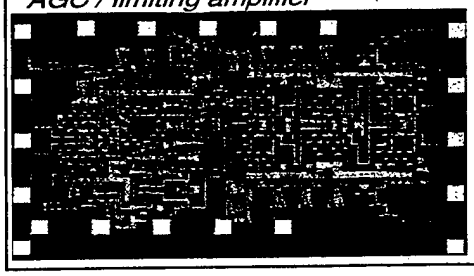
Fiber Optic ICs

not yet tested
(design 40 Gb/s)

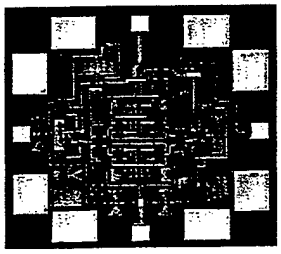
PIN / transimpedance amplifier



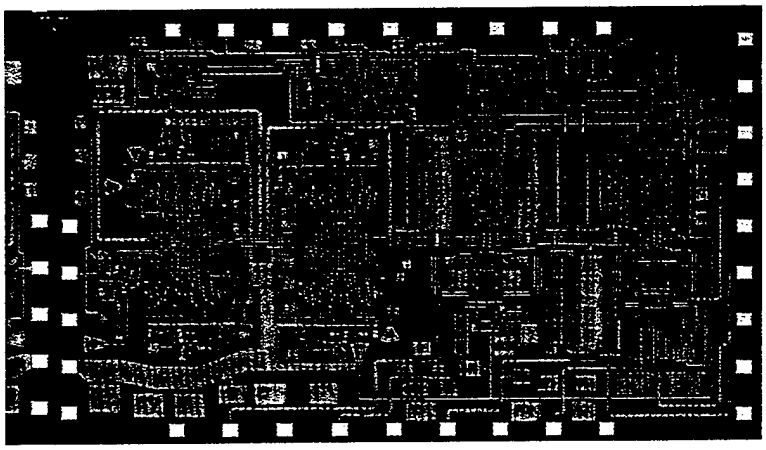
AGC / limiting amplifier



CML decision circuit



Delta-Sigma ADC (300 HBTs)



*Fast ICs **for** fast interconnects*
*Fast ICs **needing** fast interconnects*

ICs for GHz communications:
- Optical fiber transmission to, beyond 40 Gb/s
  - with electronic data switching
- millimeter-wave (60/90/180 GHz) wireless networks
  - at mm-wave, bandwidth is cheap & plentiful
  - ...but the hardware must become cheap
- ADCs, DACs for digital processing of RF signals

Challenges for fast ICs
- Fast transistors: scaling is key
- Wiring environment: signal, ground and power integrity
- Interconnect-limited power-delay products
- Managing high dissipated power densities

4:45pm - 5:30pm
Sun, 9 May - Tutorial 3

IRVINE SENSORS CORPORATION

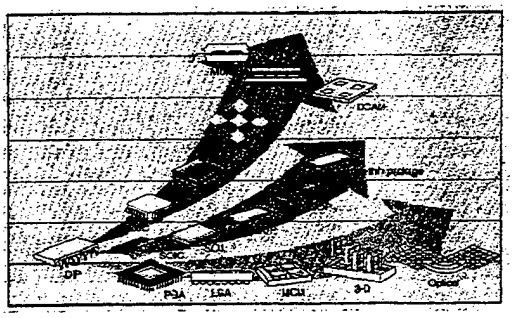
ADVANCES IN CHIP LEVEL PACKAGING

John. C. Carson

Irvine Sensors Corporation
3001 Redhill Avenue Bldg 3
Costa Mesa CA 92626

IRVINE SENSORS CORPORATION

DIFFERENT TRENDS IN PACKAGING



(from R. Heitmann, Universal Instruments Inc. - EPP May 1996)

IRVINE SENSORS CORPORATION

Excerpts from SIA Roadmap for Cost/Performance Category

	1995-2000	2000-2005	2005-2010
feature size (μm)	0.35-0.25	0.18-0.13	0.10-0.07
transistors/cm^2	4M-7M	13M- 25M	50M-120M
pin count	300-1000	1200-2000	2400-3600
package thickness (mm)	1.0- 2.0	1.0	0.5 - 1.0
package cost (cents/pin)	1.4 - 4	1-2	0.6-1.3
package size (mm)	23-45	29-50	35-50
lead pitch- peripheral (mm)	0.3- 1	0.3-0.65	0.3-0.5
lead pitch - array (mm)	1.0 - 1.5	1.0	0.5- 0.65
power (W)	2-18	2-28	2-55
Performance (MHz)	100-200	200-400	400-1000

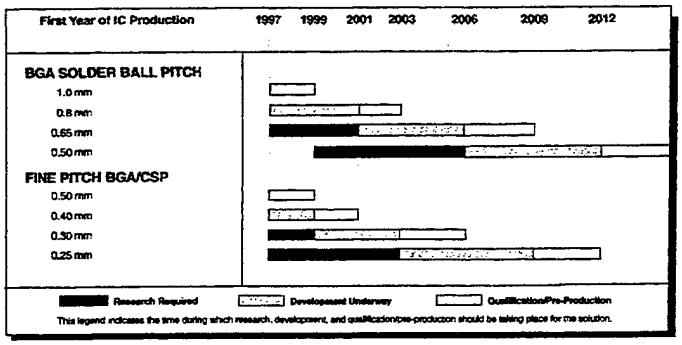
TRENDS OBSERVED

- Packages are getting thinner
- Number of leads is getting larger
- Package footprint decreasing to approach chip size
- Direct Chip attach techniques are emerging
- Package pin pitch is decreasing
- Package pins are being distributed in array format

HOWEVER, SUPPORTING SUBSTRATES ARE BECOMING MORE AND MORE LIMITING

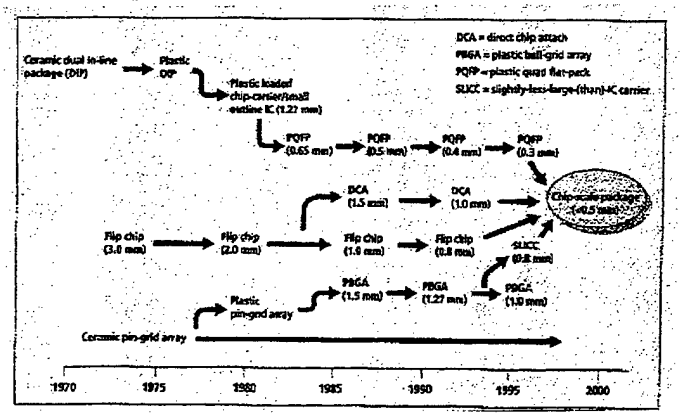
SYSTEM DESIGNERS ARE SEEKING SOLUTIONS IN MULTI-CHIP AND 3D PACKAGING TECHNIQUES ALONG WITH SYSTEM-IN-A-CHIP APPROACHES

FINE PITCH BGA AND CHIP SCALE PACKAGE POTENTIAL SOLUTIONS



SIA Roadmap 98"

EVOLUTION OF CHIP PACKAGES



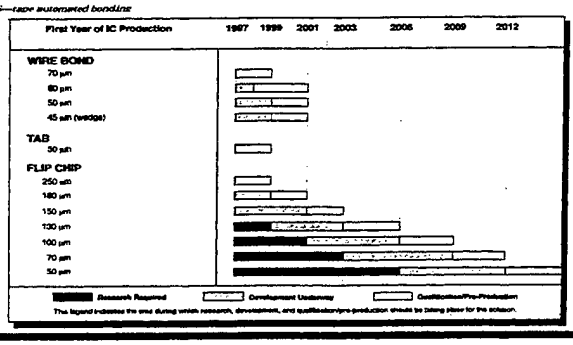
(P. Thompson, Motorola; IEEE Spectrum August 1977)

IRVINE SENSORS CORPORATION

CHIP TO NEXT LEVEL INTERCONNECT NEEDS AND SOLUTIONS

Year of First Product Shipment / Technology Generation	1997 250 nm	1999 180 nm	2001 150 nm	2003 130 nm	2006 100 nm	2009 70 nm	2012 50 nm
Chip Interconnect Pitch (um)							
Wire bond—ball	70	50	50	50	50	50	50
Wire bond—wedge	60	45	45	45	45	45	45
TAB*	50	50	50	50	50	50	50
Flip chip (area array)	250	180	150	130	100	70	50

*TAB—tape automated bonding



SIA Roadmap 98*

Irvine Sensors Corporation

IRVINE SENSORS CORPORATION

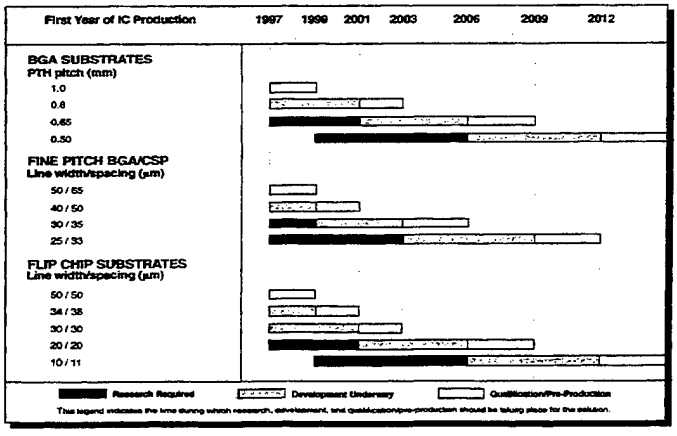
DIFFERENT DIRECT CHIP ATTACH TECHNIQUES

	Surface Mount	Chip on Flex	Flip-chip on Flex
Footprint	Large	Medium	Small
IC pad pitch (µm)	150	150	100
Cost factor	1.0	1.1	0.8-2.8
Pretesting of chips	Yes	No	No
Added wafer process	No	No	Yes
Thermal Resistance	60 C/W	45 C/W	75 C/W
Inductance (nH)	1.0 - 3.0	1.0 - 2.0	0.1 - 0.2
Capacitance (pF)	0.2	0.2	0.03
Advantages	• readily available • high reliability • easy rework • established infrastructure	• small footprint • good electrical performance • compatible with most ICs • excellent thermal resistance	• good electrical performance • lowest cost in high volume • no post cleaning for flux
Disadvantages	• low electrical performance • high processing temperatures • requires flux cleaning • largest outline	• lower yields • no pretesting • no rework after glob-top	• poor thermal performance on flex • wafer bumping required • no pretesting

Irvine Sensors Corporation

IRVINE SENSORS CORPORATION

HIGH DENSITY SUBTRATE POTENTIAL SOLUTIONS



Irvine Sensors Corporation

IRVINE SENSORS CORPORATION

FOOTPRINTS OF CHIP SCALE AND CONVENTIONAL PACKAGES

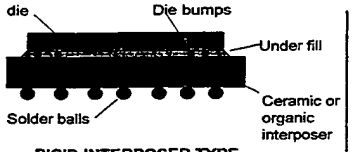
Lead count	Conventional TSOP	Fine Pitch QFP	CSP
28	152	56	11
32	168	63	19
52	169	81	31
64	169	99	36

Areas in mm2,

Adapted from P. Thompson (IEEE Spectrum August 1997)

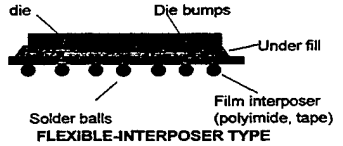
IRVINE SENSORS CORPORATION

TYPES OF CHIP SCALE PACKAGES



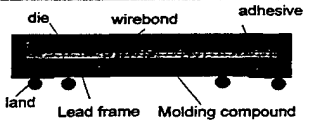
RIGID-INTERPOSER TYPE

NUCSP, Ceramic fine pitch BGA, mini BGA SFFP, Stud bump bond, SLIC, Flip-chip BGA Chip size thin package



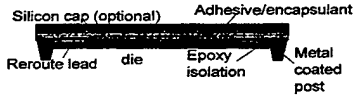
FLEXIBLE-INTERPOSER TYPE

Chip-on-flex, Flip-chip BGA, JACS-PAK, Fine pitch BGA, Resin Molded CSP, FBGA, µBGA, µstar BGA



CUSTOM LEAD FRAME

Small Outline Nolead, Lead on Chip, µstud BGA, Bottom Lead Package, Molded bump, Very Small Peripheral Array, Flip-Tape Carrier



WAFER LEVEL (MICRO SURFACE MOUNT)

Micro Surface Mount, SlimCase, Mini BGA

IRVINE SENSORS CORPORATION

PACKAGES ARE GETTING THINNER

1.2 mm TSOP

0.5 mm Dual In Line Tape Carrier

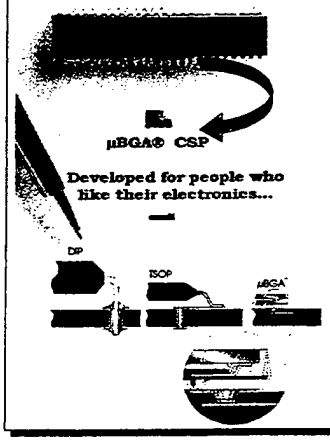
0.4 mm Ultra Thin Chip Package

0.3 mm Slim Case (from ShellCase)

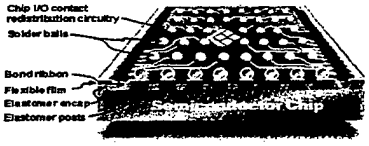
- **Thin packages can be made only 100 microns larger than the die in length, width and height**
- **Note that the incoming wafer thickness is 0.6 - 0.7 mm**
- **The die in thinner packages is in the range of 100 microns (0.1 mm) thin**
- **Advanced die thinning techniques required to support packaging development**

EXAMPLE OF CHIP SCALE PACKAGES

Micro Ball Grid Array (µBGA) from Tessera Inc.

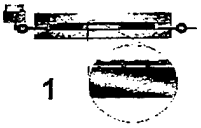


- Direct path from the IC to the PCB provided by a flexible interposer
- Fast electrical interface (e.g. Rambus 0. 8 - 1.6 GHz leads)
- Compatible with "chip shooter"s
- 0.5 - 1.0 mm pitch BGA

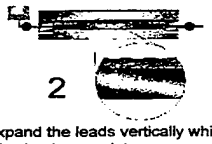


EXAMPLE OF CHIP SCALE PACKAGES

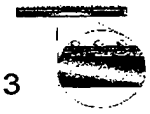
WAVE ™ (Wide Area Vertical Expansion) wafer level packaging technology from Tessera Inc



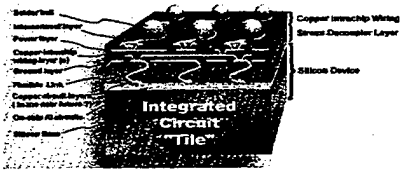
Align interposer and wafer



Expand the leads vertically while injecting low modulus encapsulant

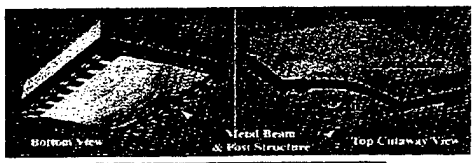


Attach solder balls and dice into single chips



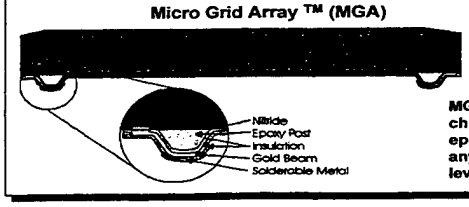
Source: J. Fjelstad, Tessera Inc.

EXAMPLE OF CHIP SCALE PACKAGES



Parameter	MSMT	0.3 QFP	Flip Chip
Height (mm)	.5-.8	1.4	.5-.7
Inductance (nH)	0.1 to 0.2	1-7	0.1 to 0.2
Capacitance (pF)	0.02 to 0.03	0.5 to 1	0.3
Attachment	Solder	Solder	Solder and Underfill

Micro Grid Array ™ (MGA)



Micro Surface Mount technology (MSMT) package from ChipScale Inc.
The package is formed in wafer form

MSMT yields active surface of the chip away from PCB

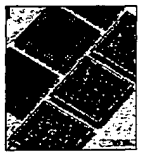
MGA technology produces packages with active face down to PCB

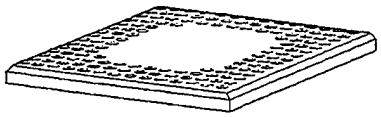
MGA provides a standoff from chip surface by using a compliant epoxy and can be placed anywhere on the chip using wafer level processing

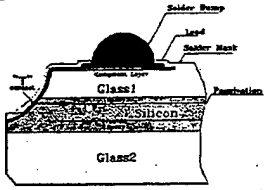
IRVINE SENSORS CORPORATION

EXAMPLE OF CHIP SCALE PACKAGES

ShellCase BGA from ShellCase with pitches down to 500 microns





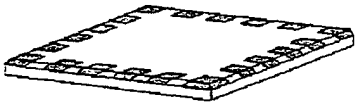


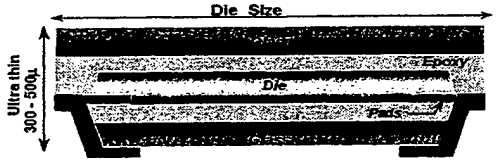
Compliant ShellCase CSP (crosssection)

IRVINE SENSORS CORPORATION

EXAMPLE OF CHIP SCALE PACKAGES

Ultrathin CSP from ShellCase





IRVINE SENSORS CORPORATION

COMPARISON OF CONVENTIONAL AND WAFER LEVEL CHIP PACKAGING

Traditional IC Packaging	Wafer Level Packaging
Wafer is probed, diced and sorted	Wafer moved directly to packaging
ICs packaged away from fab	ICs packaged in fab
ICs are packaged one at a time	ICs are packaged en masse
Burn in performed in sockets	Burn in performed on wafer
Power and ground taken from PCB	Power and ground distributed in assembled structure
Device tested 2-3 times	Device tested once
High pin counts required	Lower external I/O possible
Higher power required	Reduced power requirements
All function in the chip	Function shared between package and chip
More complex substrate required	Simpler substrates possible (lower I/O)
Lead inductance concerns	Lead inductance nearly eliminated

Source: J. Fjelstad, Tessera Inc.

SUPPORTING TECHNOLOGIES FOR ADVANCED PACKAGING

Advanced Packaging requires the utilization of the following techniques extensively :

- thinning of silicon wafers containing circuits
- bump bonding for high I/O density interface
- handling of KGD in die form
- handling of die of different sizes and origins, non-electronic chips (e.g. MEMS, Lasers, Detectors, Fluidic Devices)

Therefore, advances in these techniques will help to increase the density and the functionality of advanced packages

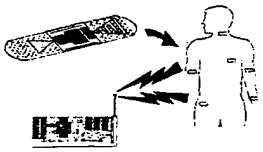
ULTRA-THIN SILICON CIRCUITS

- A Kapton (50 micron thick) based flexible test vehicle has been used to test ultra-thin flash die
- 25 micron thin 16 Mb Flash die has been successfully tested after mounting on the test vehicle
- 25 Micron thin memory die mounted on the flexible substrate is bendable with the substrate. A bending radius of 1 mm can be obtained for each micron of silicon thickness

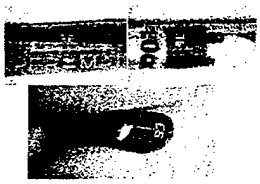
TECHNOLOGIES AND PRODUCTS BASED ON THIN SILICON CIRCUITS

- thin integrated circuit stacks with higher capacity (10 times more)
- flexible circuitry in conformal packaging
  - medical applications (shape conforming sensors)
  - space applications (SOI-like advantages)
  - wearable products
  - smart cards

18 layers of 20 micron thin Si stack compared to ISC's standard short stack: within same height, the new thin stack can accommodate 10 times more layers



"Smart Band Aid" for body function monitoring

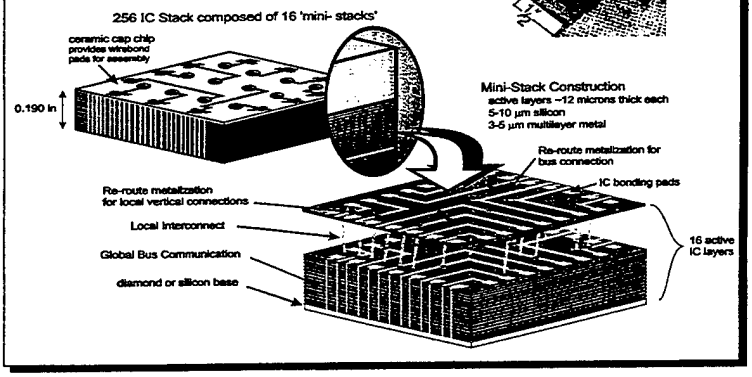


Bendable circuits for space microprobes, medical microprobes and smart projectiles

IRVINE SENSORS CORPORATION

VERY HIGH DENSITY 3D STACKING BASED ON ULTRA THIN SILICON CIRCUITS

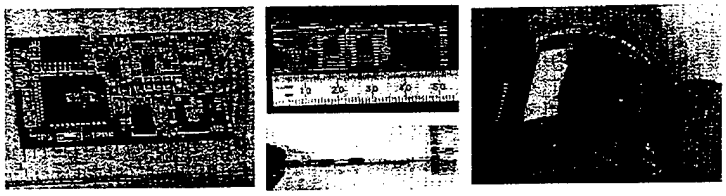
3D-VLSI?



Irvine Sensors Corporation

IRVINE SENSORS CORPORATION

ULTRA THIN CIRCUITS - EVOLUTION OF A BIOMEDICAL SENSOR IMPLEMENTATION



Prototype sensor input and RF transceiver circuit from off-the-shelf components

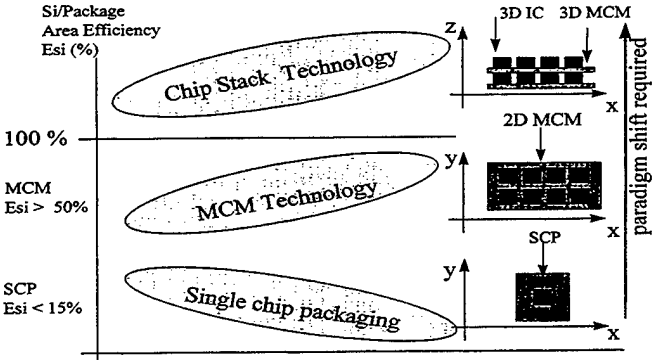
Custom ASIC replaces many ICs and a flex board replaces the rigid PC board

ISC's unique wafer thinning technology results in the silicon becoming flexible without damaging the circuit. Ultra-thin unpackaged ICs mounted on flexible boards

Irvine Sensors Corporation

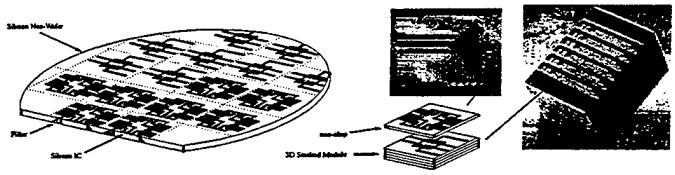
IRVINE SENSORS CORPORATION

PARADIGM SHIFT IN PACKAGING



Irvine Sensors Corporation

NEO-DIE NEO-WAFER AND NEO-STACK CONCEPT



A revolutionary chip-level layering and stacking concept to eliminate the same-size restriction and wafer level inventory requirements of existing processes

- The process is designed to re-create a wafer from individual and heterogenous chips for batch processing by embedding them into an epoxy frame
- After lithography and metalization, the wafer will be diced into neo-die of identical sizes that contain each layer to be stacked
- Mature stacking technology and tools will be used to stack many layers and interconnect layers

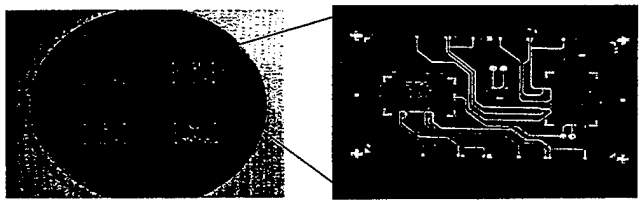
NEO STACKING APPROACH

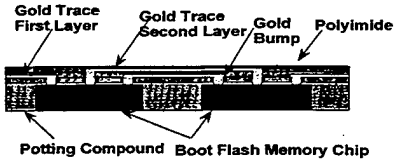
- Starting with KGD, construct a new, or neo-wafer with many dice in a molding compound matrix
- Use a standard neo-die size, just slightly larger than the largest die in the stack
- Add blank silicon to open areas on layers where smaller die are used to enhance thermal conduction between layers if needed
- Perform metalization and thinning in neo-wafer form
- Dice into individual layers
- Laminate into a stack

<u>Neo-stacking is a breakthrough in high density packaging technology</u>

- It allows complete systems in a cube
- It allows the combination of massive electronic functions with extreme miniaturization and integral logic and control functions
- dense layer-to-layer interconnects through the epoxy molding layer
- The process is highly manufacturable through industry standard automated tooling and batch processing

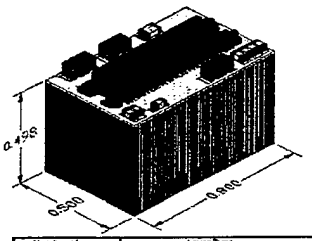
NEO STACKING FABRICATION EXAMPLES

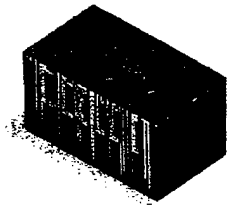




IRVINE SENSORS CORPORATION

NEO STACK FABRICATION EXAMPLE



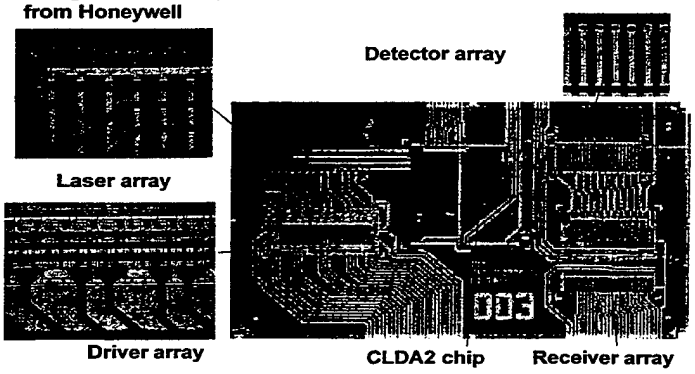


Number of Layers	Layer Type	Routing Layers	Total Chips	Chip Types
1	Cap Substrate	2-sided	-	-
4	Capacitor	1	1	1
32	Flash	1	1	1
1	Flash Driver	2	4	1
1	Microprocessor	1	1	1
1	FPGA	2	1	1
1	Bus Driver	2	4	2
1	Boot Flash	2	2	1
1	IEEE 1394 Interface	2	3	2
4	DRAM	1	1	1
1	Bottom Ceramic	-	-	-
48 Total Layers			52 Total Chips	10 Chip Types

IRVINE SENSORS CORPORATION

NEO-STACKING TECHNOLOGY FOR IMPLEMENTING DIVERSE COMPONENTS IN ONE LAYER

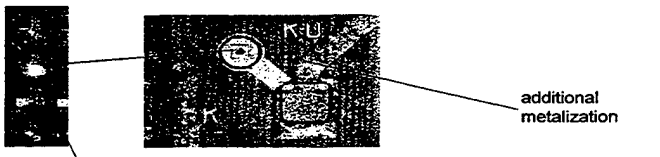
Neo layers with lasers, detectors and electronic chips were built using vertical cavity surface emitting lasers and MSM detectors from Honeywell

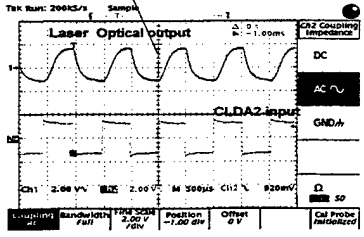


IRVINE SENSORS CORPORATION

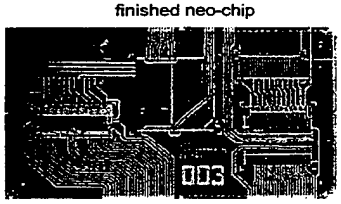
OPTOELECTRONIC NEO-LAYERS

The characterization of the optical links is ongoing

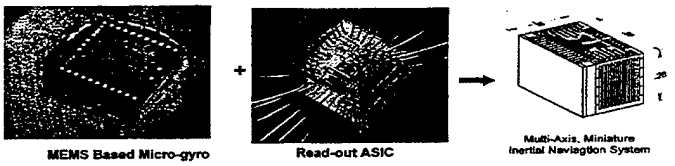




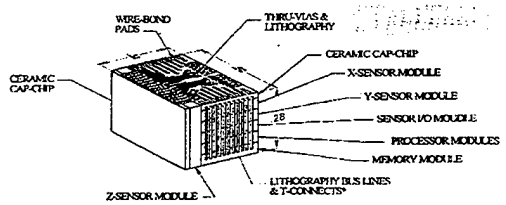
finished neo-chip



INTEGRATION OF DIFFERENT TECHNOLOGIES IN THE SAME STACK
Example: Multi-Axis Miniature Inertial navigation System



Six-axis inertial measurement and navigation unit



HIGH SPEED OPERATION IN CHIP STACKS

- **ALCATEL-Espace demonstrated the basic operation of stacked microwave circuits in Ku-band (10.7 -12. 7 GHz) range**
- **Insertion losses were about -0.5 dB/mm**
- **Results indicate potential operation up to 30 GHz.**

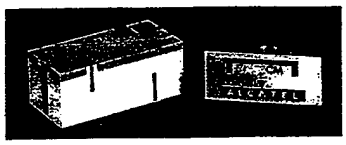


Figure 4 : 3D microwave module by ALCATEL ESPACE (dimension : 20 mm * 10 mm * 6 mm).

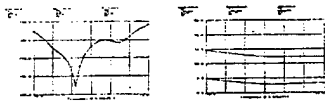


Figure 5 : Measurement of the 3D module.

From: 3D Microwave Modules for Space Applications , P.Monfraix et al

CHIP LEVEL 3D-PACKAGING ROADMAP

YEARS	1998	2000	2002	2004
In-plane line density (lines/cm)	500	1000	1500	2000
In-plane total number of metalization layers	2	3	4	5
Side-face line density (lines/cm)	200	400	800	1000
Side face total number of layers	1	2	2	3
Areal line density (new technology) (lines/cm^2)	900	1600	2500	5000
Maximum Operating Frequency -coplanar lines (GHz)	1	2	4	8
Maximum Operating Frequency –Microstrip lines (GHz)	10	20	30	50

ELECTRICAL CHARACTERISTICS OF TYPICAL PACKAGE INTERFACES

100 mm^2 chip	Bare die with 75 μm Wire bond	Flip-chip 0.5 mm bump	Quad Flat Pack with 75 μm wire bond	Micro Ball Grid array 1 mm bump
pitch (mm)	0.15	0.25	0.30	0.50
Footprint (mm^2)	125	125	785	150
Package/chip area	1.25	1.25	7.85	1.5
Height (mm)	0.4-0.6	0.5-0.7	1.4	0.84
Inductance (nH)	1-2	0.05-0.2	1-7	0.5-2.1
Capacitance (pF)	0.2	0.05- 0.1	0.5-1	0.05-0.2

COMPARISON OF ADVANCED PACKAGING APPROACHES

TECHNOLOGY	ADVANTAGES	DISADVANTAGES	APPLICATIONS
Flip-Chip	– Covers least area – Has excellent electrical performance	– Lacks die availability – Hard to assemble due to planarity – Die shrinks results in board redesign	– Low lead count used (watches, vehicle modules, displays – High reliability systems – Vertically integrated companies
Chip-Scale Package	– System size reduction with standard technology	– New technology lacks reliability and production infrastructure	– Memories – Portable computing and communications – Under 100 leads
Multi Chip	– Early system integration – Best electrical performance	– Needs KGD – Lacks die level availability – Difficult test	– Large systems (e.g. avionics) – Some automotive and communication systems

Adapted from P. Thompson, IEEE Spectrum 1997

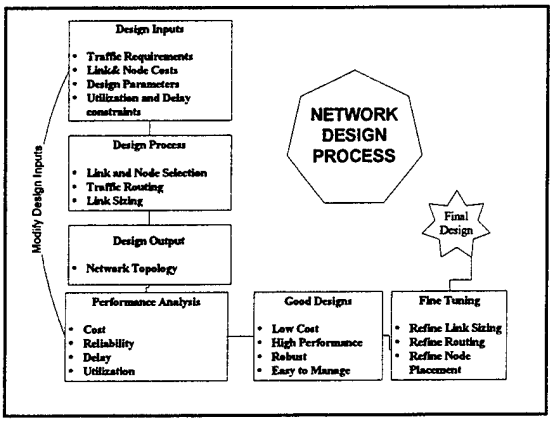
EVOLUTION OF CHIP PACKAGING TRENDS

From	To	Impact
Al pads and metallurgy	Copper pads and metallurgy	bond wires bump materials passivation
Wire bond	flip-chip	low cost wafer bumping underfill materials
leaded packages	area array packages	low cost dense substrates encapsulants
single chip packages	direct chip attach Chip scale package	low cost wafer bumping low cost dense substrates low cost known good die
200 mm wafers	300 mm wafers very thin chips	chip thinning material handling equipment configuration

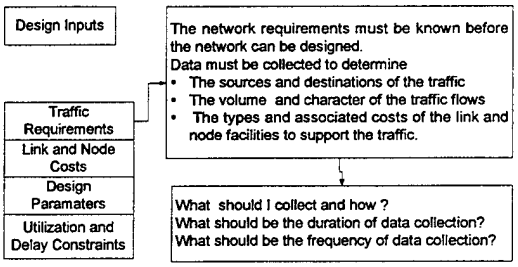
5:30pm - 6:15pm
Sun, 9 May - Tutorial 4

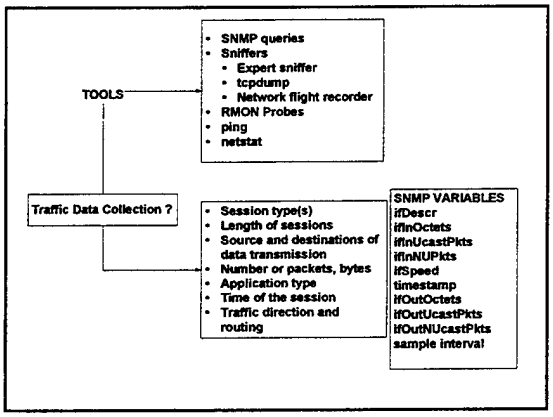
MODELING, ANALYSIS and SIMULATION of DATA NETWORKS

Yusuf Ozturk
San Diego State University
Department of Electrical and Computer Engineering
San Diego, CA 92182-1309
Tel : (619) 594-4550
Fax : (619) 594-3654
mailto:ozturk@kahuna.sdsu.edu



NETWORK DESIGN PROCESS





USEFUL SNMP VARIABLES

- **IfInOctets**
  The total number of octets received on the interface, including framing characters.
- **IfInUcastPkts**
  The number of subnetwork-unicast packets delivered to a higher-layer protocol.
- **IfInNUcastPkts**
  The number of non-unicast (i.e., subnetwork-broadcast or subnetwork-multicast) packets delivered to a higher-layer protocol.
- **ifOutOctets**
  The total number of octets transmitted out of the interface, including framing characters.
- **ifOutUcastPkts**
  The total number of packets that higher-level protocols requested be transmitted to a subnetwork-unicast address, including those that were discarded or not sent.
- **ifOutNUcastPkts**
  The total number of packets that higher-level protocols requested be transmitted to a non-unicast (i.e., a subnetwork-broadcast or subnetwork-multicast) address, including those that were discarded or not sent.
- **ifDescr**
  Describes the interface. It should include – identification information for the physical line and a description of the network.

TRAFFIC GENERATOR PARAMETERS for SIMULATIONS

The data collected must then be related, culled and summarized if they are to be useful.
Ultimately the traffic data must be consolidated to a single estimate of the source traffic destination for each node.

PacketIAT = SampleInterval/(ifInUcastPkts + ifInNUcastPkts)

MeanPacket Size = ifInOctets /(ifInUcas tPkts + ifInNUcast Pkts)

AverageDat aRate = ifInOctets × 8/SampleIn terval

SAMPLING FREQUENCY

- The sampling frequency determines the resolution of the data and its storage requirements.
- Larger sampling intervals result in smoother data summaries and hide the variation between samples.
- Network performance statistics may have a periodic component. If so, the data sampling period should be less than half of that. The data period should not be divisible by the sampling period, otherwise the samples will consistently be taken at peaks, midpoints or low points of the data.
- When using SNMP to measure Internet/Intranet performance, do not let network management traffic swamp regular traffic. (Heisenberg Uncertainty Principle : We disturb the object we measure -- the more precisely we try to measure it the more we disturb the object)

How frequent should we sample for SNMP?

- NumberOfSamplingPoints = 1000
- PacketsPerSample = 2
- TimeBetweenSamples = 60 seconds
- SizeOfPacket = 85 bytes
- MediaSpeed = 256,000 bits/second

$$\text{SampleRate} = \frac{\text{NumberOfSamplingPoints} \times \text{PacketsPerSample}}{\text{TimeBetweenSamples}}$$

$$\text{SampleRate} = 33.33 \text{ Samples/Seconds}$$

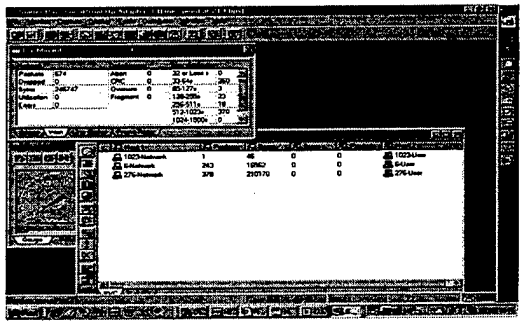
$$\text{Utilization} = \frac{\text{SampleRate} \times \text{SizeOfPacket} \times 8}{\text{MediaSpeed}} \times 100$$

Utilization = 8.854 Less than 10% of link is used for SNMP

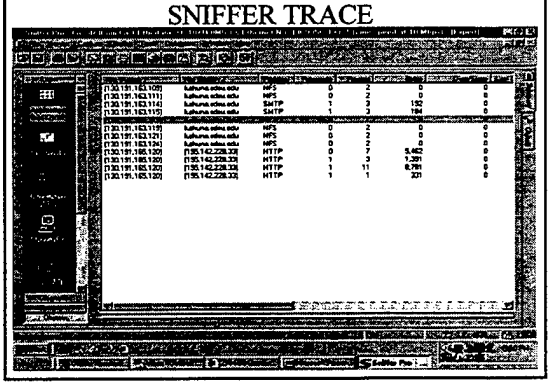
TRAFFIC CHARACTERISTICS

- LAN Analyzers ,Expert Sniffer , RMON MIB , TCPDump will provide information about packet size distributions. When in doubt about the probability distribution function assume exponential.
- Packet inter-arrival times can also be obtained using LAN Analyzers or sniffers. You may use exponential distribution when you do not have a better model for the inter-arrival. Exponential distribution will not fit to all data. For example : RIP (routing information protocol) sends its routing tables on all interfaces every 30 seconds. While the client requests may be exponentially distributed the responses may be fixed sized closely spaced packets.
- Business cycle defines how the average packet rate fluctuates and the peak value should be used for performance analysis.
- The average packet arrival rate and the average packet size should be multiplied for link bandwidth selection.

SNIFFER TRACE



SNIFFER TRACE



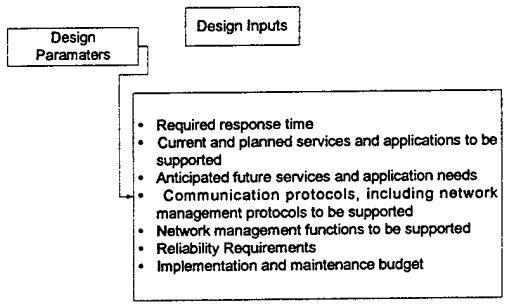
BENCHMARKING

- Before a new application is deployed on a network benchmark data should be collected to predict the impact of the new application on the network resources.
- Questions :
  - Is enough data being collected to provide a statistically valid sample?
  - Will the real network actually experience the type of the traffic being measured?
- While collecting benchmark traffic data for an application the application should run on an isolated segment. Otherwise contamination will occur by the data not participating in the benchmarking study.
- Measurements taken on one type of network (Ethernet for example) does not apply to other networks (Token ring).

BENCHMARKING
Sources of contamination

- **Target and measurement media are different. (Differences such as MTU , framing characters etc. should be compensated)**
- **Network operating systems are dissimilar.**
- **Applications are similar but not identical.**
- **Queues existed at the servers and network links at the time of measurement.**
- **Disk I/O delays are lumped in with CPU instruction delays.**
- **The LAN carried other traffic not related to benchmarking.**
- **Measurements were taken remote from the the server system.**
- **The server is doing other work in addition to our application.**
- **The benchmark network and systems are already highly utilized.**
- *Even if the application will be deployed over a wide area network with a number of remote clients, benchmarking data still should be collected over a LAN system with ONE client and server on the same LAN.*

NETWORK DESIGN PROCESS



DESIGN PARAMETERS

- **What type of traffic will be carried ?**
- **Is the data to be carried time sensitive? If yes , what are the delay and delay jitter requirements?**
- **Is the data bursty in nature ? Will smoothing affect application ?**
- **What are the acceptable bit error rates and packet loss rates?**
- **Is the traffic symmetric (such as in videoconferencing) ?**
- **Are there any applications that will benefit from multicasting or broadcasting (such as digital video broadcasting) ?**
- **What are the reliability requirements ? What are the tolerable times for recovery in case of failures?**
- **What are the security requirements ?**
- **What protocols will be supported ?**
- **Total network budget and the percentage reserved for network management, analysis and data collection tools.**
- **Self similarity of network sites for reducing operating costs versus initial deployment costs.**

WAN DESIGN

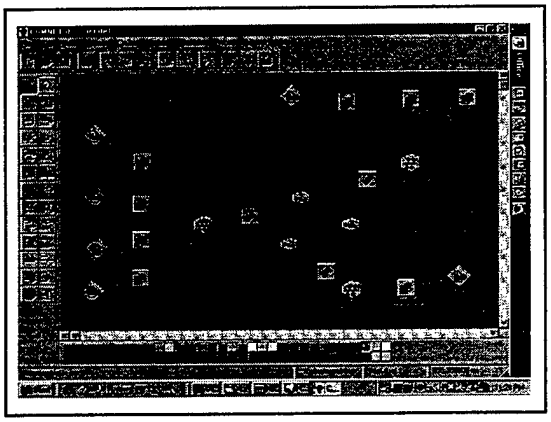
- **Type of service decisions**
  - **Frame Relay**
  - **ATM**
  - **DDS Lines**
  - **SMDS**
  - **Private versus public leased lines**
  - **etc.**
- **Topology decisions**
  - **Star**
  - **Backbone**
  - **Tree**
  - **Mesh**
- **Link and Node selection and sizing**

LAN DESIGN

- **Topology decision (Bus, Ring, Tree , Star etc.)**
- **Access Protocol (CSMA/CD, Token Ring, etc.)**
- **Frame/ Packet Size**
- **Transmission Capacity**
- **Signal propagation delay**
- **Buffer size**
- **Processing delays**
- **Throughput**
- **User traffic profile**
- **Data collision and retransmission**
- **Bridging/Routing decisions**
- **Security**
- **Availability**

Video Network Design

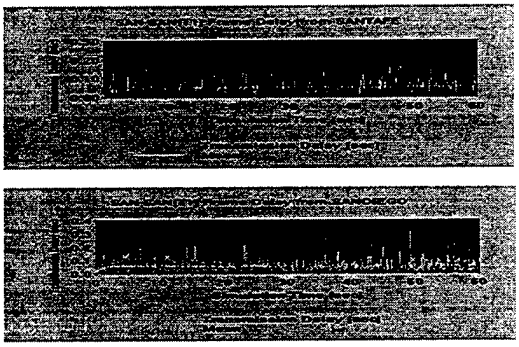
- **Design a network to support video conferencing between three remote locations (Santa Fe, San Diego and Mexico). Each of these locations one or more active participants introducing compressed video at a rate of 384 kbps into the network (Using h.261).**
- **There may be more than one passive participant (Listening only -- destination to the data) at each site.**
- **Employ multicasting to reduce the duplicate traffic on LAN and WAN links where possible.**
- **Simulate the proposed model and provide performance measures on the LAN and WAN Links.**
- **Based on the simulations refine the design (Topology, link and nodes, WAN service type etc) .**



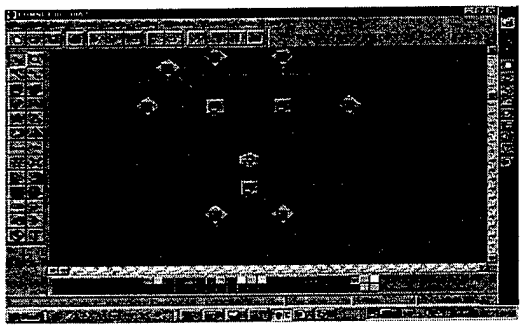
PERFORMANCE METRICS

- Link Utilization
  - Utilization by application
  - Utilization by protocol
- Collision Statistics (CSMA/CD)
- Token Ring Statistics
- Message Delay (maximum , minimum, min)
- Delay Variation
- Node utilization (Router - Bridge - Switch etc.)
- Processing delays
- Throughput
- Buffer size (by node and by port)
- Resistance to node and link failures (Availability)
- Scalability (How easy it will be to increase number of sites or number of users at each site)

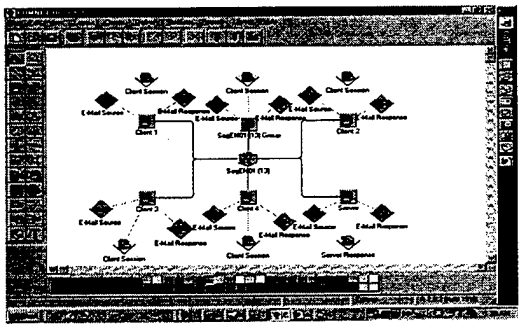
Frame Delay on WAN LINK



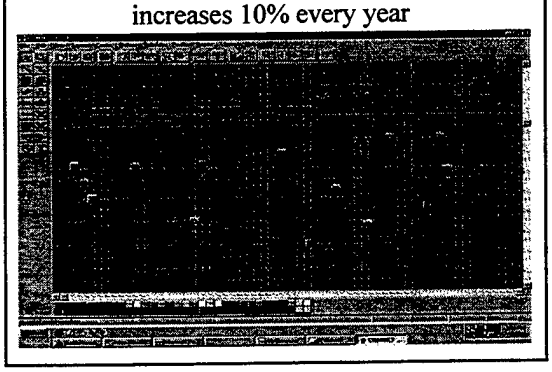
MODELING and SIMULATION of APPLICATIONS



E-MAIL SYSTEM MODEL



How will vBNS perform if the traffic rate increases 10% every year



NOTES

Monday,
10 May 1999

8:15am - 8:45am
Mon, 10 May - 1.1

10 Gigabit Ethernet

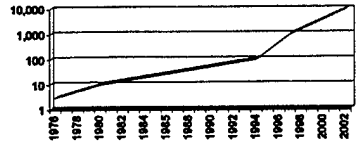
Peter Wang
Technology Development Center
May 10, 1999

Outline

- Drive Towards 10 Gb Ethernet
  - Motivations
  - Applications of 10 GbE
  - Requirements
- Technical Issues & Technology Enablers
  - Transmission medium
  - PMD
  - PMA
  - PCS
  - MAC
- Summary

Drive Towards 10 Gb Ethernet

Moore's Law for Ethernet

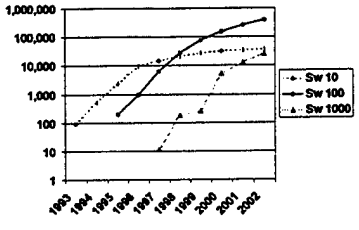


Customer problems to be solved with 10 GbE

- Traditional LAN applications and private enterprise applications
  - GbE to 10 GbE aggregating switches
  - Linking multi-port 10 GbE Switches
  - Linking multi-Gbps Routers inside a LAN
- LAN and Non-LAN private enterprise applications
  - High speed clustered computing interconnects.
    - » NGIO & FutureIO
  - Provide point-to-point backplane connection
- access, MAN, "RAN", WAN(?)

Switched Ports: Cumulative Shipments

2002: 30 M GbE ports

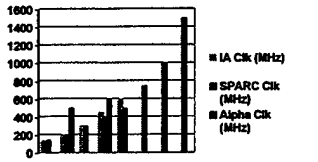


DellOro Group 2/98

Application Traffic Drivers

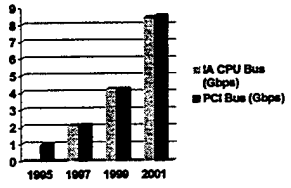
- Digital multimedia production & distribution
  - graphic/image rich business presentations
  - digital post-production
  - medical/scientific imagery
- Data mining & database Synchronization
- The video factor
  - Tele-presence
  - When will HDTV move onto the data network?

Processor Trends



Source: Microprocessor Report

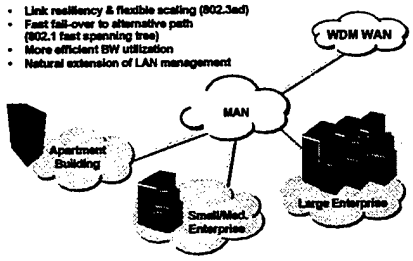
Bus Speed Trend



* NGIO considering 8 Gbps link
* FutureIO considering 10GBps

MAN

- Link resiliency & flexible scaling (802.3ad)
- Fast fail-over to alternative path (802.1 fast spanning tree)
- More efficient BW utilization
- Natural extension of LAN management



Distance Assumptions

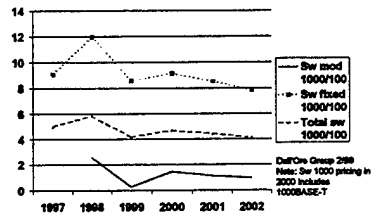
- No building will move because of 10 GE
- Customers will expect to run 10 GE over the same links as GE
- Customers will expect to use 10 GE with new applications that are emerging or on the horizon
  - Access to RAN/MAN
  - cluster computing interconnects

Desired Requirements

- Distance
  - data center, cluster interconnect: 50 m
  - building risers: 500 m
  - campus backbone: 2-10 km
  - MAN access: 10-30 km
  - MAN/RAN backbone: 30+ km
- Media
  - single mode fiber except for shortest reach
- Topology
  - pt-pt, full-duplex, w/ congestion control

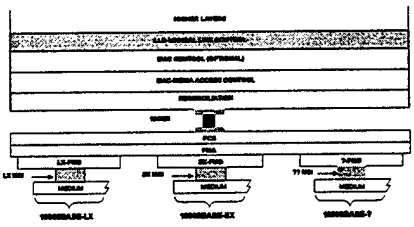
Cost Ratios: Sw 1000 to Sw 100

10 GE cost expectation: < 10x GE



Technical Issues & Technology Enablers

Protocol Stack



Medium

Standard Single-mode Fiber
- Campus backbone and metro area network.
- Suitable for long wavelength lasers
  - » 1.3 um: fiber attenuation limited
  - » 1.5 um: dispersion limited

Multi-mode Fiber
- Server-switch connections; horizontal & vertical risers
- Large core size, low cost transceivers
- ISI from intermodal and intramodel dispersion limits the bit rate and distance
- CWDM on existing or serial on adv. MMF

Copper
- Server-switch in data center
- Extremely limited distance
- Complex PMA if multilevel signal

PMD Components - Lasers

Requirements
- High modulation speed (> 10 G) & output power
- Low driving current & temperature dependence
- High linearity and narrow spectral width

Issues
- Serial vs. CWDM?
  - » Distance vs. complexity/cost trade-off
  - » Support existing MMF?

Technologies
- Distributed-feedback Laser (DFB)
- Fabry-Perot
- Vertical Cavity Lasers (VCSELs)

Laser Technologies

Distributed-feedback Laser (DFB)
- Distributed resonators suppresses multi-modes
- Buried heterostructure:low threshold current (~10mA) and high output power
- Thermal cooling
  - » threshold current is temperature dependent
  - » control mode hopping
- Isolator
  - » eliminate reflection noise
- External modulator
  - » reduce frequency chirping
- Eliminate cooler, isolator and/or external modulator for low-cost shorter reach solutions?

Laser Technologies

Fabry-Perot
- 1.3 um; SMF/MMF
- Simple structure and low cost
- Short reach, multi-mode laser
  - » distance limited by dispersion and mode-partition noise

Vertical Cavity Lasers (VCSELs)
- 0.85 um/1.3 um; SMF/ MMF
- Low cost and easier packaging
- Parallel optics CWDM

PMD Components - Modulator

- **Used for extended reach**

Requirements
- High modulation speed & linearity
- Low driving voltage & cost
- Packaging size

Technologies
- $LiNbO_3$
- Electro-Absorption
- Hybrid integration with DFB lasers

PMD Components - Detector

Requirements
- High receiver sensitivity (responsivity)
- High bandwidth
- Low noise

Technologies
- PIN:
  - » -18 dBm sensitivity
  - » Up to 50 G demonstrated.
- APD
  - » -31 dBm sensitivity
  - » slower than PIN

PMA

Functions
- Mux/Demux circuitry
- Clock data recovery
- Byte alignment (Optional)
- Laser driver & post-amp

Technologies
- **Bipolar**
  - » Mature technology w/ limited BW
- **GaAs**
  - » Mature technology w/ high BW & high power dissipation
- **SiGe**
  - » Emerging next generation Si-bipolar technology
  - » High BW, lower power dissipation & cost
  - » ULSI/BiCMOS compatible, allow system-on-chip integration

PHY Components Status

Component	Technology	Availability	Rate Capabil	Distance
Laser	EML	Now	10 Gbps	80 km
	CW ILM/LiNbO3	Now	10 Gbps	600 km
	Uncooled FP	Jun-99	12.5 Gbps	1 km
	Uncooled DFB	Jun-99	12.5 Gbps	10 km
Photodetector	PIN	Now	12.5 Gbps	N/A
	APD	Now	12.5 Gbps	N/A
Laser Driver	GaAs	Now	10 Gbps	N/A
	SiGe	Dec-99	12.5 Gbps	N/A
TIA	GaAs	Now	12.5 Gbps	N/A
	SiGe	Dec-99	12.5 Gbps	N/A
Limiting Amp	GaAs	Now	12.5 Gbps	N/A
	SiGe	Dec-99	12.5 Gbps	N/A
CDR	GaAs	Now	12.5 Gbps	N/A
	SiGe	Dec-99	12.5 Gbps	N/A
Mux/Demux	GaAs or Bipolar	Now	10 Gbps	N/A
	SiGe	Dec-99	12.5 Gbps	N/A

Source: Lucent presentation @ 802.3 CFI, 3/99

PCS Sublayer

- **Encoding/decoding of data from/to MAC**
- **Need for 10 GMII**
  - Single/Multi-channel operation
- **Auto-negotiation**
  - 1000/10000?

MAC

Ethernet without CSMA/CD
- Connectionless packet transmissions
- Consistant addressing & frame format
- No translation, nor segmentation

Functions
- Full-duplex only, speed/distance independent
  - Inter-frame gap (IFG) & preamble size in bit times
- Flow control
  - Pause operation resonpse time in bit times
- Auto negotiations

MAC Challenges and Solutions

Issues & Challenges
- Bus interface (Backplane/MAC and MAC/PCS)
  - Parallel data width (10/16/20/32/40/etc.; single vs. differential)
  - clock rate (1.25G/625M/622M/312.5M/etc.)
  - chip-chip skew
- Processing Ethernet frames with very short lookup time
- Buffering

Technologies & Solutions
- Interfaces
  - SSTL, PECL, LVDS, CML
- Advances in CMOS
  - 0.25mm process now, migrating to .18mm
  - embedded DRAM for buffer
- Larger min frame size and architectural improvements

Summary

Standard Scope - TBD

Distance/Media
- 802.3z link model?
- Existing cabling standard adequate?
  - New advanced MMF?
  - Is copper (i.e. CX-like) worth considering?

Wavelengths
- 850/1300/1550 nm?

Channels
- Serial only? Or include CWDM?

Coding
- 8B/10B, 14B/15B, 16B/18B, scrambling, multilevel analog?

Quality
- Laser safety?
- BER? Jitter?
- EMI

Potential Starting Point for 10GE Discussions

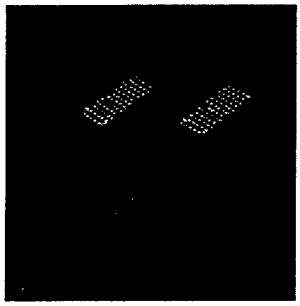
- MAC
  - full-duplex only
  - min. packet size (256 vs 64 bytes)
- MAC/PCS/PMA standard interfaces
- 10 Gbps (data)
- Encoding (8B/10B)
- Single wavelength in 1300 nm window
- Standard single mode fiber
- Up to 30 km

3Com

More connected.™

8:45am - 9:15am
Mon, 10 May - 1.2

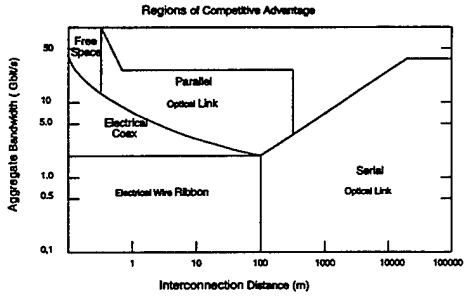
PAROLI a Synchronous Interconnection Link with a Through Put of 13 Gbit/s



Karsten Drögemüller

karsten.droegemueller@infineon.com

Competing Data Link Technologies



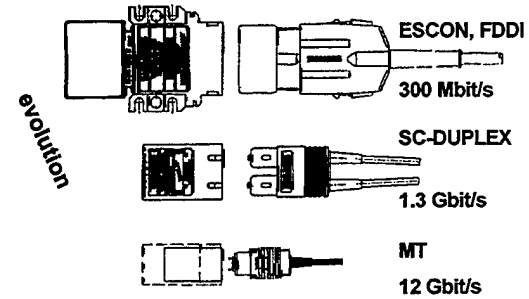
Why Optical Interconnect ?

- Copper: BW x distance product limited
- BW demand drastically increasing
- Cable size is 1/50th of copper
- Optics is the solution to escalating EMI problems
- Cost is not much higher than high performance copper

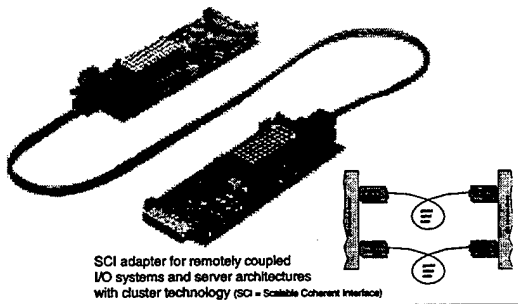
Applications of Optical Interconnect

Applications	Distance	Through Put	DC Couple	Status
Box-Box	5 - 100 m	10 - 30 Gbit/s	Yes and No	Products available
Backplane Extend	1 - 5 m	50 Gbit/s	Yes	R&D
Backplane	25-100 cm	50 - 150 Gbit/s	Yes	R&D
On Board	5 - 25 cm	50 - 100 Gbit/s	Yes	R&D

Evolution of Modul dimension

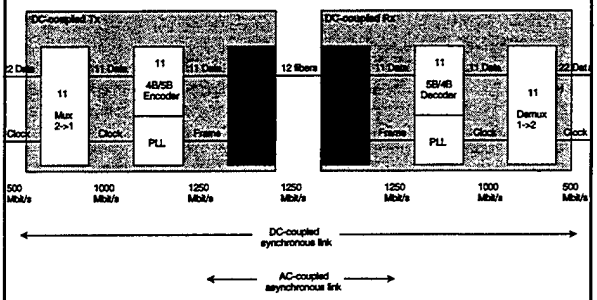


SCI application of SIEMENS (High End Server Group)

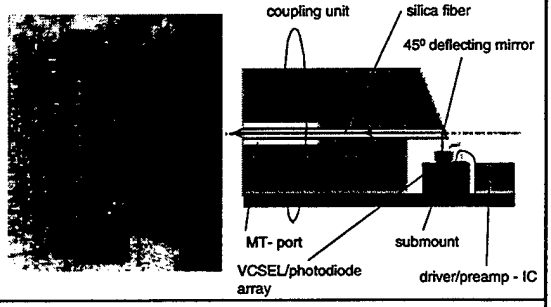


SCI adapter for remotely coupled
I/O systems and server architectures
with cluster technology (SCI = Scalable Coherent Interface)

System Architecture

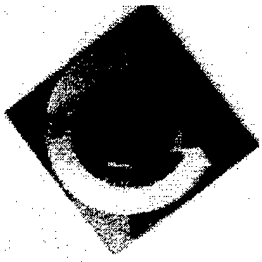


PAROLI Coupling Scheme



Eye Safety

IEC 825-1 Amendment Laser Classification



Maximum optical power in fiber

	1 channel	12 channels
NA = 0.2	443µW	118µW
NA = 0.275	537µW	199µW

9:15am - 9:45am
Mon, 10 May - 1.3

Tb/s Chip I/O - how close are we to practical reality?

Rick Walker
Hewlett-Packard Company
Palo Alto, California
walker@opus.hpl.hp.com

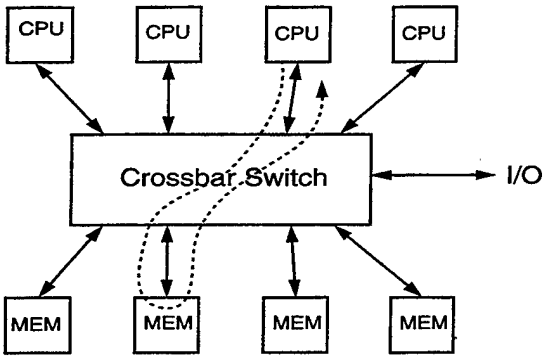
Agenda

- Applications and Key Specifications
- General Architecture for inter-chip communication
- Limitations
  - Skin-Loss
  - Delay Matching for Multi-phase sampling
  - CMOS Scaling
- Industry Trends
- Conclusions

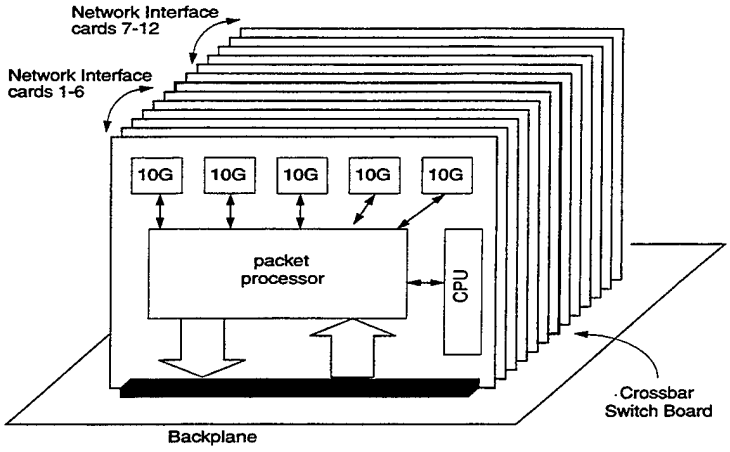
Current Practice

- Current high-performance systems are skew limited using parallel data clocked at 250-500Mb/s.
- Using clock and data recovery on Gb/s links eliminates the skew problem and improves system BW by factor of 8-16X.
  - What are the limits for advanced systems?

CPU-CPU/Memory Application



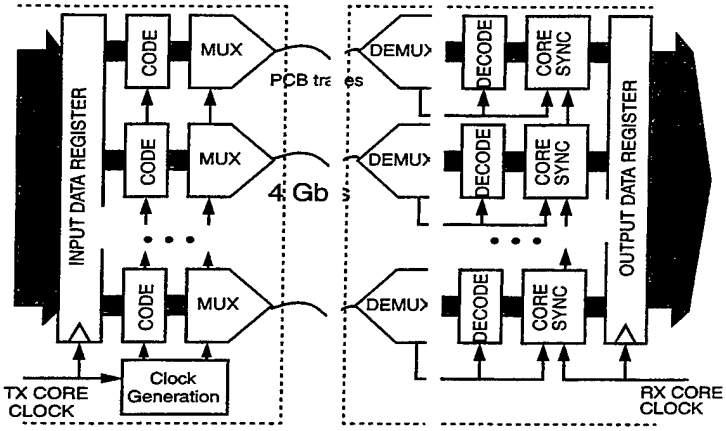
Router Application



Key Specifications

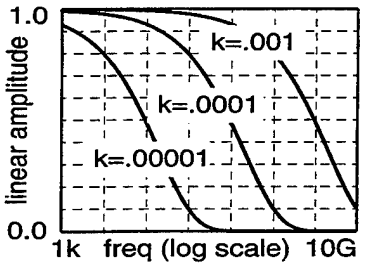
- Speed: As high as possible - at least 1Tb/s I/O per chip
- Latency: critical - less than 10ns plus time of flight
- BW/link: limited to 4-5 Gb/s by PCB loss
- Power: for a 100W chip, all 250 links should dissipate less than 40W -> 160 mW per link
- Size: a typical processor may be $9cm^2$, if links use 20% of the total area, then each 4Gb/s link cell should be less than $720000um^2$ in size.

Genera Arch tecture

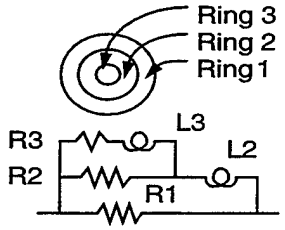


Skin Loss and Dielectric Loss

Nearly all cables are well modeled by a product of Skin Loss $S(f) = e^{-k_s(1+j)l\sqrt{f}}$, and Dielectric Loss $D(f) = e^{-k_d l f}$ with appropriate k_s,k_d factors. Dielectric Loss dominates in the multi-GHz range. Both plot as straight lines on log(dB) vs log(f) graph.

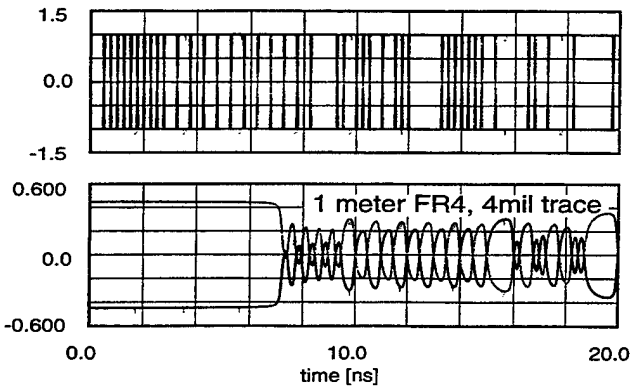


[YFW82]



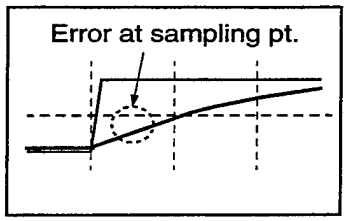
Three-element equivalent circuit of a conductor with skin loss

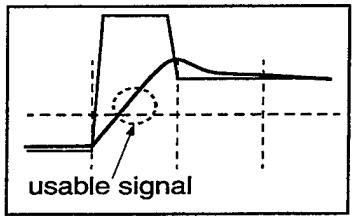
Non-equalized NRZ data



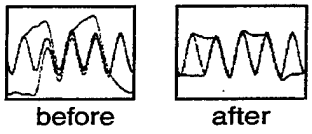
Skin Loss Equalization at Transmitter

boost the first pulse after every transition

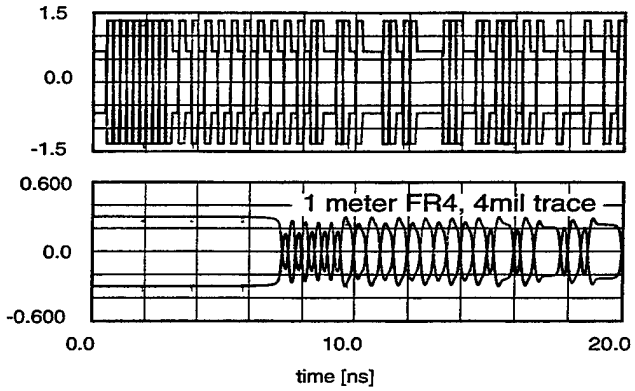




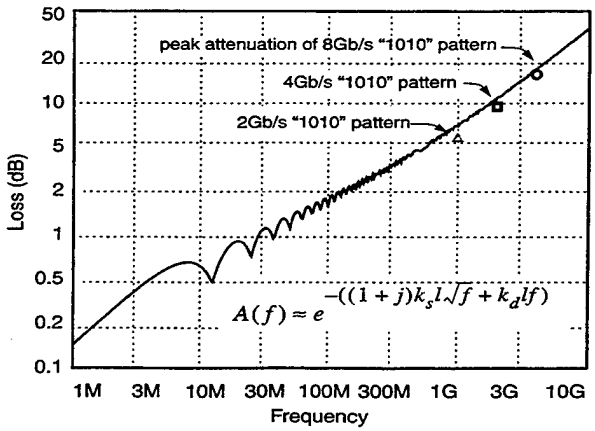
[FMW97]



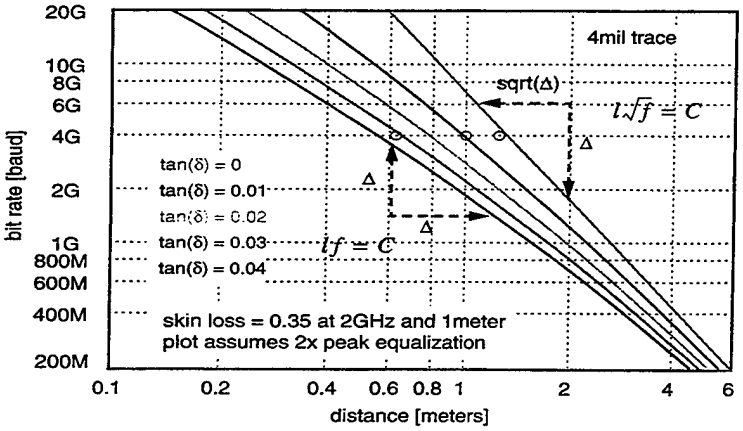
6dB Equalized Data



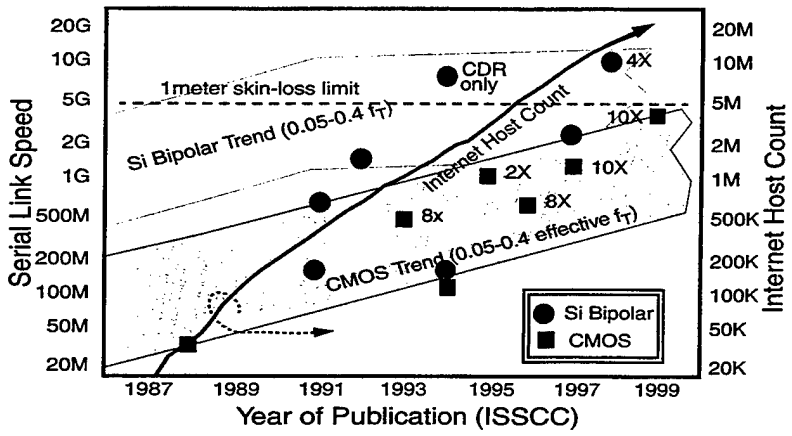
Skin Loss



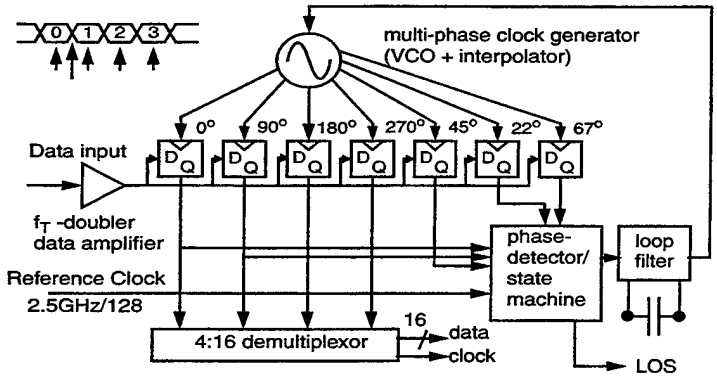
Data rate vs distance and tan(δ)



Communication Trends

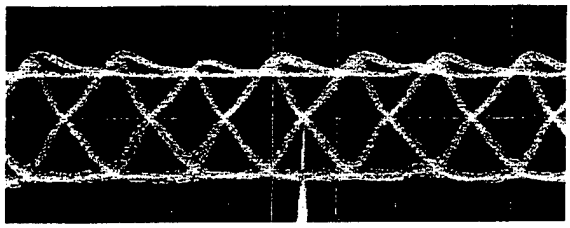


Example Multiphase RX Block Diagram



[WHK98]

Measurement of a Multi-phase System

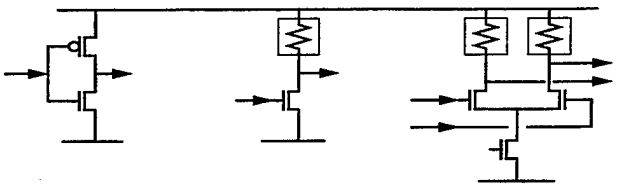


Reported Jitter: 8ps rms, 44ps pk-pk at 3.5Gb/s.

Measurement of photo shows 26ps difference between widest and narrowest eye, so true eye margin for end-end system is $44ps\sqrt{2} + 2 \bullet 26ps = 118ps$, or a total eye closure of 41%.

Attention to delay *matching* is critical!

Techniques to Improve Delay Matching and Power Supply Noise Immunity



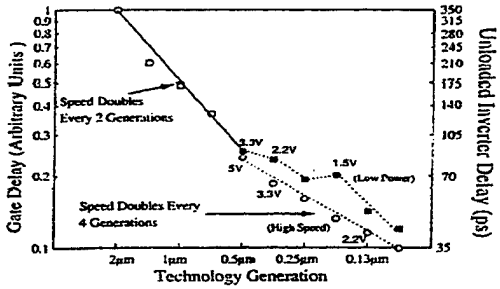
- Rising edge delay dependent on *absolute* V_T
- RC time-constant changes with supply voltage

- Rising edge delay dependent on *absolute* V_T

- Delay depends on VT *matching*
- Current source absorbs supply voltage changes

CMOS Scaling Issues

- Gate delay no longer scales with process

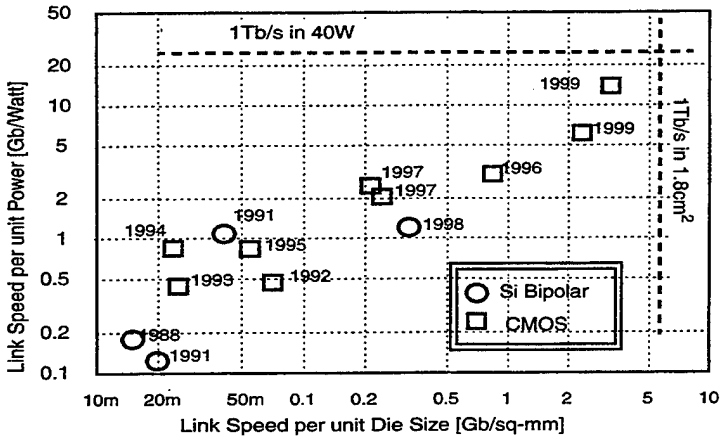


See: Chenming Hu, "Low-Voltage CMOS Device Scaling" 1994 ISSCC Digest, pp 86-87.

CMOS Scaling Issues (continued)

- V_t doesn't track with power supply - so we gradually lose ability to make ECL-like differential circuits.
- Full-swing circuits show worse delay matching than ECL-like topologies.
- Full-swing circuits show worse power-supply delay modulation than differential circuits.
- V_t matching gets worse due to statistical dopant variations in channel.
    - All of these trends make power supply noise rejection and multi-phase alignment more difficult with each process scaling.

Power and die size vs target



Industry Trends

- 50% of all U.S. Families now have home computers
- Computer performance has surpassed needs of most users: witness the drop of P.C. prices in the last 3 years from a stable $2K down to $500 levels.
- Internet host count was doubling every 6 months in 1988, is now doubling every 24 months - we are clearly past the 50% adoption point in the growth curve.
- What applications will continue to drive expensive and exotic improvements in interconnect technology?
    - Without a new "killer app" to drive development, we may by stuck with the limitations of FR4/CMOS for quite some time.

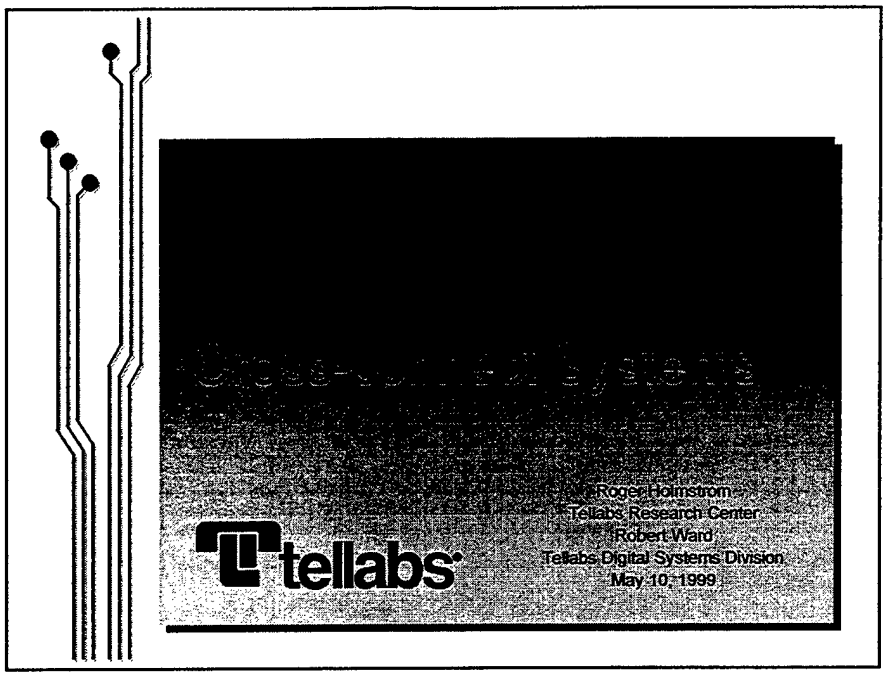
Viability of "exotic" technologies

- Yielded CMOS parts come in at $\$10/cm^2$
- Tb/s chip-chip links are probably feasible in the next few years.
- This performance can be achieved with existing BGA packages across commodity FR-4 PC Backplanes.
- The incremental cost of a Tb/s link in these applications will be about $18 + connector cost.
    - For optical solutions to take hold in these applications, they must provide either significantly higher performance (>10Tb/s) or cheaper system cost (not likely).

Conclusions

- Still much work to be done, but 1 Tb/s chip I/O seems an attainable target.
- 5Gb/s on 1meter PCB is the fastest that can be feasibly supported for the foreseeable future with *low latency.*
- Fiber seems to be progressing along either a 1-10-100-1000-10,000MHz or a 622-2488-10,000MHz evolutionary path. There may be an economically important need for 5Gb/s links.
- 10 Tb/s chip I/O is probably out of the question for current high-volume technologies (CMOS, FR-4 PCB). Computer designs and programs may have to give up cache coherency, and move towards cooperative computing architectures to break out of this limitation.

10:00am - 10:30am
Mon, 10 May - 1.4



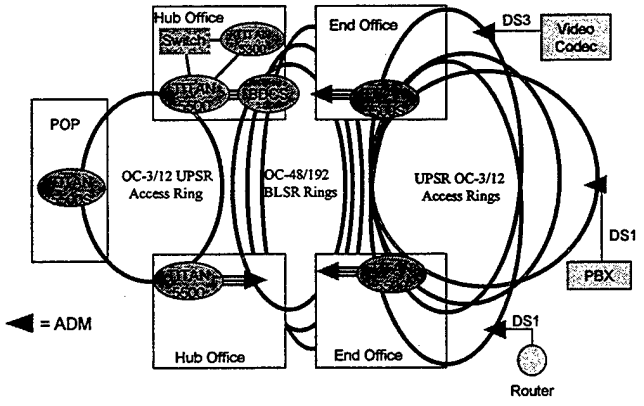
Outline

- Introduction to Digital Cross-connects
  - Applications, Requirements, etc.
- Interconnect Requirements
  - Performance, Physical, Reliability, Cost
- Interconnect Choices
- Conclusions

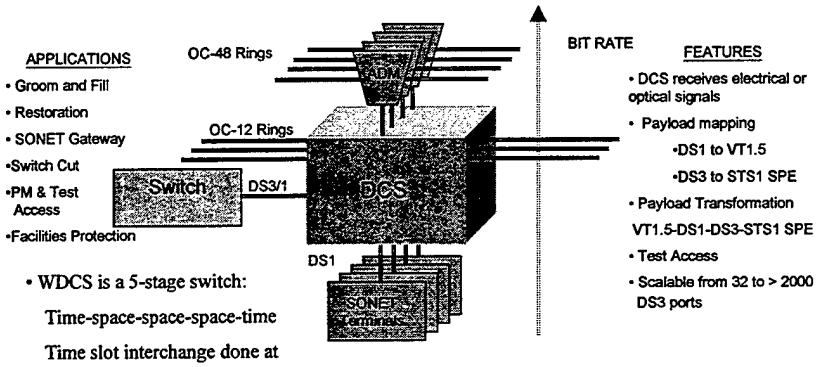
Standard Transmission Rates

Digital	Rate	Equiv. Voice Channels	Optical	Rate
DS0	64 kbps	1	OC-1	51.84 Mbps
DS1 (T1)	1.544 Mbps	24	OC-3	155.52 Mbps
RTD (Digital Radio)	1.544 Mbps		OC-12	622.08 Mbps
DS1C	3.152 Mbps	48	OC-24	1.244 Gbps
DS3 (T3)	44.763 Mbps	672	OC-48	2.488 Gbps
DS3C	89.472 Mbps	1344	OC-192	9.953 Gbps
DS4	274.176 Mbps	4032	OC-768	39.813 Gbps
DS5	470 Mbps			

DCS in a Network



Wideband Cross-Connect



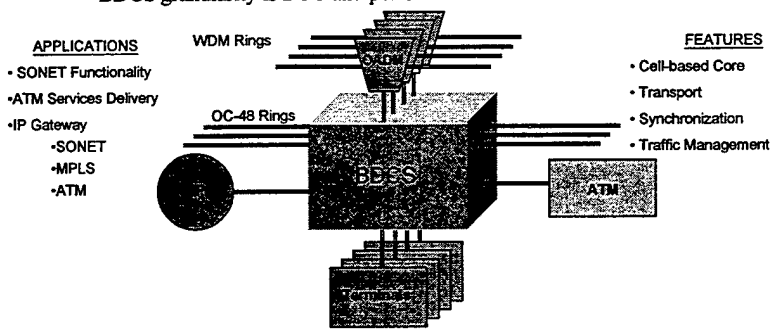
• WDCS is a 5-stage switch:

Time-space-space-space-time

Time slot interchange done at the ports

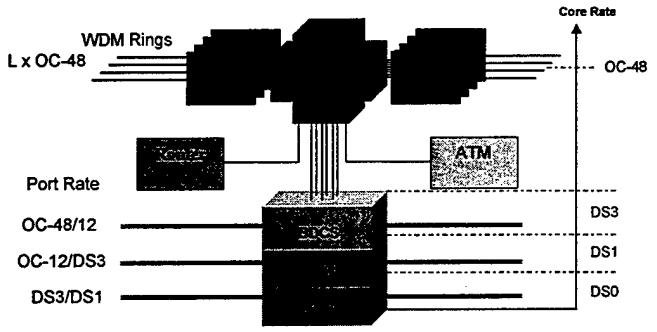
Broadband Cross-Connect

• BDCS granularity is DS 3 and ports are from OC-12 to OC-48



Optical Cross-Connect

Total Bandwidth Management



Optical Channel Cross-Connect

- Motivation
  - Traffic management demand at OC-48
    - Deployment of WDM
    - Growth of data traffic
- OXC Applications
  - Provides interface to optical transport network
    - Maps client signals into Optical Transport Layer
    - Co-existence of several client signals
    - Enables 1:N equipment redundancy
  - Cross-connect in the core network (OTL)
    - manages end-to-end connections in OTL
    - Interfaces between different network segments (equipment)
  - Protection Switching
    - UPSR and BLSR-like protection on Och basis
  - Restoration

Requirements

- Performance All Interconnects
  - Interconnect rate up to 4 Gbps
  - BER < 10^{-15}
  - Jitter - < 0.1 dB peak (complicated issue)
- Port-to Core Interconnect
  - Physical Distance between Ports and Core up to 300 meters
  - Core module connected to four ports
- Within Core
  - Distances < 10 meters
  - Module I/O up to 128 Gbps (one card 12 x 16)
- Space
  - Core Modules
    - Connected to switch end stage modules
      - 200 pins of connector (~8" of connector)
  - End Stage Modules
    - Core Connection
      - 2 x 200 pin connector
    - 4 Port connections (duplex) at 4 Gbps each (optical)
  - Port Modules
    - Two 4 Gbps Connections (duplex) to Core

Requirements, cont.

Reliability

- Definition of Failure
  - *BER spec is < 10^{-15}. If BER > 10^{-15}, has link failed?*
  - End-to-End connectivity: From Bellcore GR-499-CORE - *"A period of unavailable time begins when the bit-error rate in each second is worse than 10^{-3} for a period of 10 consecutive seconds. These ten seconds are considered to be unavailable time".*
- DCS port-to-port DT < 0.1 min/connection/year
  - FIT < 380 for the connection
  - ⇨ 1 + 1 connection redundancy
- Module Failure Rate < 0.25/yr
  - Transceiver MTBF > 16 years (FIT < 7000)

Cost

- Number of interconnects
  - DCS size: 128 to 1056 ports
  - There are ~ 1600 interconnects in a 128 x 128 DCS
- Cost
  - Systems interconnect cost could exceed $ 1M
  - Allowable cost depends upon functionality
    - FEC and CDR included?
  - Top-Down View

$$\frac{\text{(System per-port price)}}{\text{margin factor}} = \text{BOM}$$

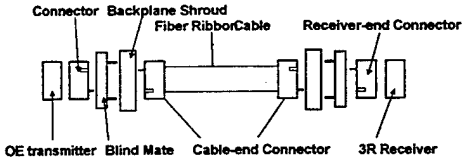
BOM - $core&mech&control = $ interconnect

Optical Interconnect Options

- Small form factor single channel transceiver modules
  - TO Can based modules - semi-cost-effective (~$ 85/Gbps), reliable
    - Bit rate - 1.25 Gbps (2.5 Gbps coming)
    - Conclusion: Still too much board space required
  - Other high-speed Serial link Options
    - 5 Gbps serial link
    - Not there yet (10 Gbps ethernet)
- Parallel Optical Interconnects
  - Advantages
    - Size, performance
  - Disadvantages
    - Availability - single source
    - Immature Product
    - Need for deskewing
    - Costs > $ 200/Gbps
- Muxed (WDM) Interconnect
  - Advantages
    - Multi-Gbps through single connector
  - Disadvantages
    - Cost and availability
    - Etc.
- Issues
  - VCSEL vs edge-emitter
  - Short vs long wavelength
  - Single vs multimode fiber
  - Discrete vs integrated transmit and receive modules
    - Duplex links - integrated
    - Performance - discrete

Structure of the POI

Simplex POI



- POI Consists of FIVE Elements:
  - OE Transmitter Module
  - OE Receiver Module
  - Connector
  - Blind Mate / Backplane Shroud
  - Connectorized Cable

4 Gbps internal transport unit Interconnect Options (#channels x bit rate):

1 x 4 Gbps		$$$
2 x 2 Gbps	maybe	
4 x 1 Gbps	just right	
5 x .8, 8 x .5, 10 x .4		less optimum

DCS Optical Interconnect Specifications

Port-to-Core distances dictate the use of Optics

- Physical
  - Size: L x W x H <= 2" x 1" x 0.5"
  - 0 - 85 °C (ambient) Operating Temperature Range
- Suggested Format: 8X connector (MT)
  - 8X ribbon cable, Blind Mate
- Suggested Bit Rate - 1 Gbps/fiber
  - 4 x 1 Gbps format
  - 8X transmitter module with 2 channels OR
  - 4X transmit + 4X receive (transceiver module)

Electrical Interconnects

Costs dictate the use of electrical interconnects (over optics)

Approximately $ 35/Gbps (dominated by ICs) = 1/10th cost of Optics

- Cables
  - Twin-ax - diff. pair over 10 meters at > 1Gbps
    - Advantages - $, availability, ruggedness,etc,
    - Disadvantages - cable management
  - Microstrip Ribbon Cable
    - Advantages - management,
    - Disadvantages - $, connectors,
- Connectors
  - 2 mm modular PGA systems 6 - 8 rows
    - ADV: mult. Vendors, $, modularity, reliable
  - Card-edge connectors
    - DIS: availability, cost, perform., assy. procedure modifications
- Boards - 2.5 Gbps signals for short distances on FR4.
  - Gtech or cyanate ester better board material
  - Teflon or polymer laminate with μstrip lines
- Line Drivers and Recovery Chips
  - Required at Gbps rates

Summary and Conclusions

- *"The whole system is an interconnect!"*
- Electrical interconnects
  - Speed and density demands are there: 10 Gbps ports in Terabit systems
  - System design must be cognizant of interconnects from the beginning
  - Board I/O and cable density are issues
  - Uniformity and stability of media is increasingly important
  - Driver and recovery chips are vital
- Optical
  - Fast Serial and parallel are needed
  - Connector systems need improvements
  - Cost HAS to come down
  - Opportunity for more functionality in optics - lower system cost

10:30am - 11:00am
Mon, 10 May - 1.5

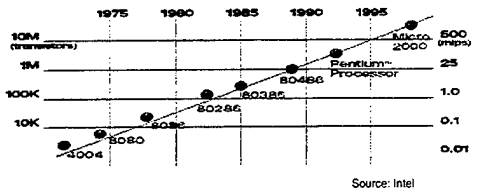
Moore's Law: The Intra-System I/O Challenge

Craig Theorin
May 10, 1999

Overview

- *Introduction: The implications of Moore + Amdahl.*
- *Link Architecture Options*
- *Copper Media Scalability*
- *Fiber Optics Scalability*
- *Conclusion*

Moore + Amdahl = Bandwidth Growth



Source: Intel

- *Moore Observes Exponential MIPs Growth.*
- *Amdahl Necessitates Proportional BW Growth to leverage Moore.*

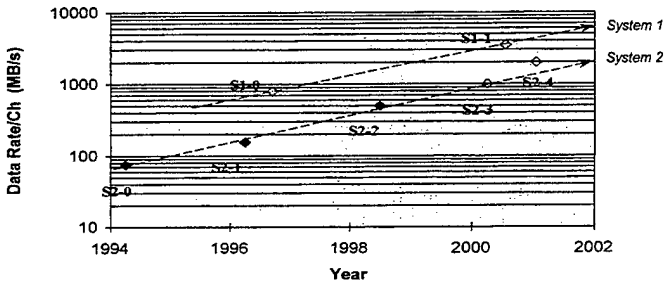
Amdahl's Law: Processor, I/O, Memory Balance

Processor	*I/O Bandwidth*	*Memory*
(MIPs)	*(Mbit/sec)*	*(Mbytes)*

System performance is optimized when MIPs=Mbit/sec=Mbytes

If processors scale with Moore's Law, so must I/O and Memory

Intra-System Data Bandwidth Trends



Keeping Pace with Moore

Year	Bandwidth (Gb/s)	Bit Width (psec)	Rise Time (psec)	Spectrum (GHz)	Zo Discont A.U.	Ch-Ch Sk (psec)
1999	1.0	1000	250	1.4	1.00	250
2000	1.6	630	157	2.2	0.63	157
2001	2.5	397	99	3.5	0.40	99
2002	4.0	250	63	5.6	0.25	63
2003	6.3	157	39	8.9	0.16	39
2004	10.1	99	25	14.1	0.10	25
2005	16.0	63	16	22.4	0.06	16
2006	25.4	39	10	35.6	0.04	10
2007	40.3	25	6	56.4	0.02	6
2008	64.0	16	4	89.6	0.02	4
2009	101.6	10	2	142.2	0.01	2

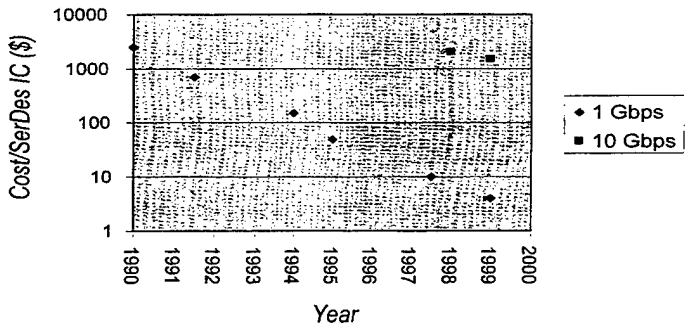
Serial vs. Parallel Data Streams

- *Serialization converts media cost to launch cost.*
  - *long reach applications.*
- *Parallel = N*Serial.*
  - *Maximum I/O BW (ie. 10-100 X serial) & I/O BW*Density.*
  - *BW Reduction for Serial Stream Processing.*

Parallel will be continually obsoleted by increasing SerDes performance/cost ratio.

A decline in serialization performance/cost growth may require parallel.

SerDes Cost/Performance



Ganged Serial vs. Clock Forwarding

- *Jitter budget has limited 1 Gbps optical clock forwarding.*
  - *Clock and Data Jitter accumulate in budget.*
  - *HiPPI Budget TBD.*
  - *FO centric designs will end clock forwarding.*
- *Future High BW Links will look like hybrid of parallel and serial or "ganged serial".*
  - *Allow deskew of parallel data streams.*
  - *Scalable for future systems.*

Data Coding

- *Typical Code Functions:*
  - *Limit low frequency content (ie. run length) for AC coupling.*
  - *"DC-Balance" the signal to keep duty cycle close to 50%.*
- *Scrambling: Muxing a PRBS w/ Data*
  - *Statistical max run length and DC balance.*
- *Multi-Level Coding*
  - *Lowers max frequency by increasing # of bits/symbol.*
- *Forward Error Correction (FEC)*
  - *Using Error Correction Bits to improve BER*

Copper Scalability

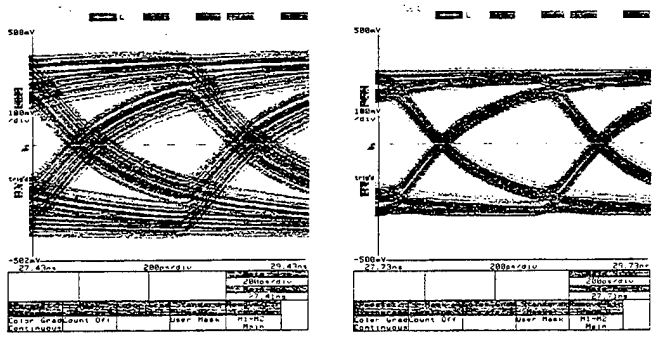
SI Concern	Cause	Solutions
Loss, Eye Pattern	Skin Effect & Loss Tan	Larger Cable, EQ, EOP, Peaking
EMI	Poor Shielding vs Tr, Imbalance	More Shielding, RF Chokes,
Return Loss	Zo Discontinuity	Control, Signal Shielding
Next/Fext	Poor Shielding vs. Tr	More Shielding between Signals
Skew (pair2pair)	Signal Routing, Er Variation	Control, Deskew Circuits.
Imbalance	EMI, Jitter,	Control, More Shielding?

- *A common solution is to increase Tr*
- *"Control", Larger Cable, More Shielding = Cost*

Bandwidth Scalability

Standard Cable | **Eye Opener Plus**

850 Mbps
20 Meters
100 Ohm
26 AWG



Gb Ethernet Copper Modems

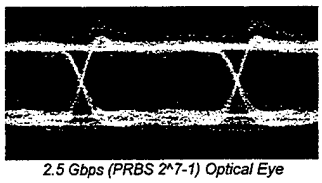
- *1 Gbps data transmitted over 4 pairs*
  - *5 Level Code (PAM5) used to send 2 bits/symbol*
  - *Extra bits for forward error correction (FEC)*
- *Hybrids are used to bi-directionally couple data into cable.*
- *Waveform "shaping" for EMI compliance*
- *Noise Reduction Through DSP*
  - *NEXT and FEXT*
  - *Digital Echo (reflection)*

Optics Scalability

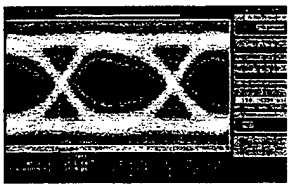
SI Concerns	Cause	Solutions
Optical Eye	Optical Noise from Source	Improved VCSELs and Launch
BER vs. BW	Higher BW Challenges Budget	Longer Wavelengths
Jitter Budget	Source Jitter & Rx BW & Noise	Improved VCSELs and Receiver
Clock/Data Jitter	Sum of clock+data Jitter	Control, Ganged Serial X-mission

- *10 GbE is currently addressing this issue for Serial Optics to 10 Gbps.*
- *For greater BW Parallel FO is required.*
- *At a cost.*

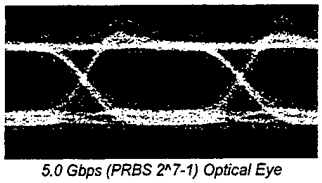
VCSEL Scaling



2.5 Gbps (PRBS 2^7-1) Optical Eye



12.5 Gbps (PRBS 2^7-1) Optical Eye

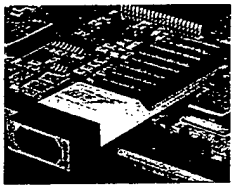


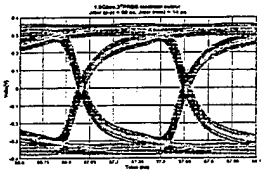
5.0 Gbps (PRBS 2^7-1) Optical Eye

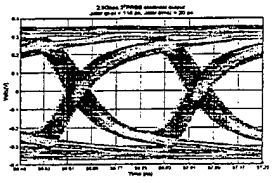
- *VCSEL BW > 20 GHz.*

Parallel Optic Packaging

- *18 Gbps Data Link*
- *High Density*
- *Modular Architecture*
- *Class 1 Eye Safe*

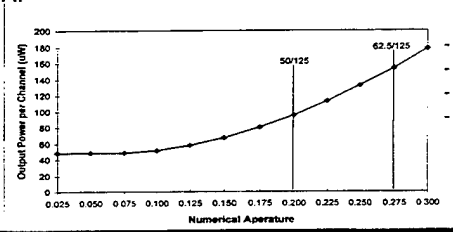






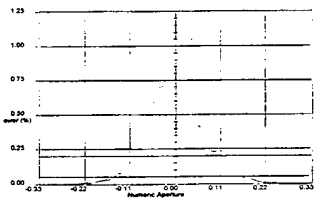
Parallel Optics Eye Safety

- *Tx optical power limited by FDA & IEC.*
- *Standard 250 um pitch 12 wide arrays force potential 6 dB power reduction for safety compliance.*
- *Amount of power allowed is proportional to divergence angle or N.A.*

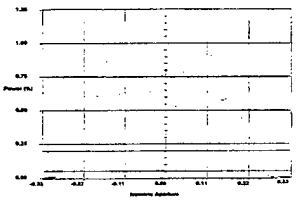


- IEC Class 1
- 12 Wide Array Output
- 850 nm lasers
- 250 um pitch

Increasing Beam Divergence

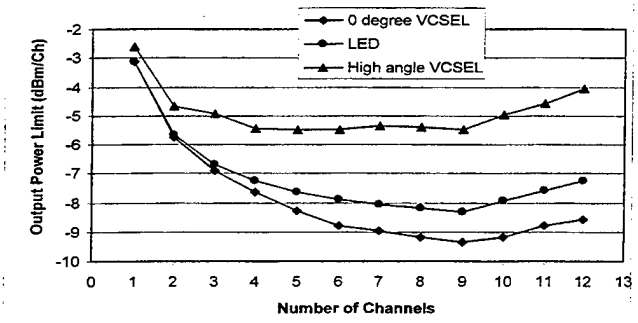


Far field distribution of a 0° VCSEL launch.



Far field distribution of a high angle VCSEL launch.

IEC Class 1 Allowable Output Power



Summary

- *Moore & Amdahl Require 10 Gbps in 5 years and 100 Gbps in 10 years.*
- *Leverage IC Functionality to solve analog transmission problems.*
  - *DSP?*
  - *Encoding, Error Recovery, Peaking, Deskew, etc.*
  - *Copper EMI challenges will be extreme.*
- *Parallel Optic data links will address BW needs.*
  - *Ganged Serial Likely for future Intra-System I/O*

NOTES

Tuesday,
11 May 1999

8:15am - 8:45am
Tues, 11 May - 2.1

Interconnects in Scalable, Distributed Multiprocessor Systems

Jeffrey Kuskin

Silicon Graphics, Inc.

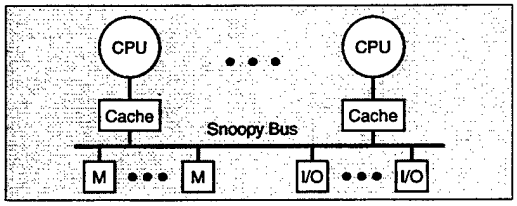
sgi 1

Outline

- Overview of scalable multiprocessor system architecture and issues
- Cache coherence in action
- Origin 2000 network details
- Multiprocessor interconnects in the future

sgi 2

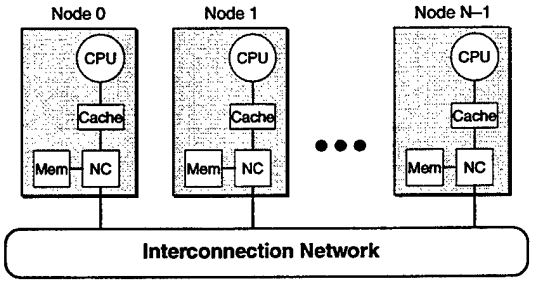
Bus-Based Multiprocessor



- Example: SGI Challenge
- Non-scalable
  - Requires broadcast
  - Limited bus bandwidth
  - Limits on snoop response time, bus transition/settling time, power, etc.

sgi 3

Scalable Multiprocessor



- Computation divided among processing nodes
- Nodes communicate via point-to-point interconnection network

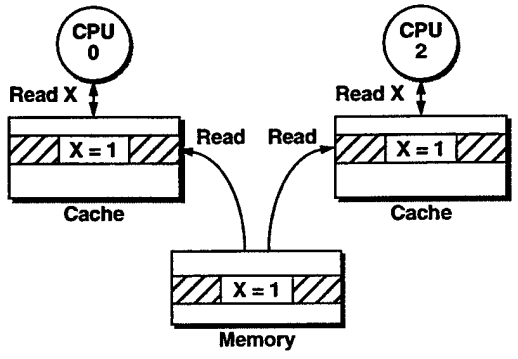
sgi 4

Communication Mechanisms

- Defines the convention used to communicate among nodes
- Message passing
  - Each node has direct access only to its local memory
  - Communication between nodes is requested explicitly
  - Examples: Intel Paragon, Thinking Machines CM-5, IBM SP2
- Shared memory
  - Physically separate memories appear as a single, unified memory
  - Each node may access any memory location using normal loads/stores
  - Examples: HP/Convex Exemplar, SGI Origin 2000, Stanford DASH

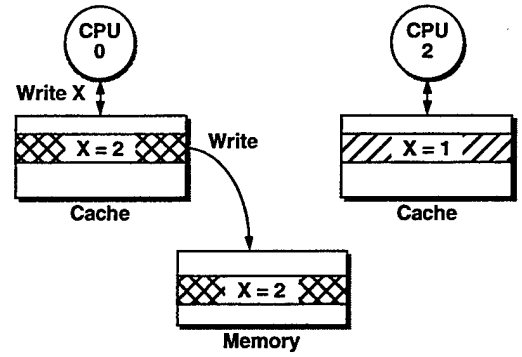
sgi 5

Cache Coherence Problem 1



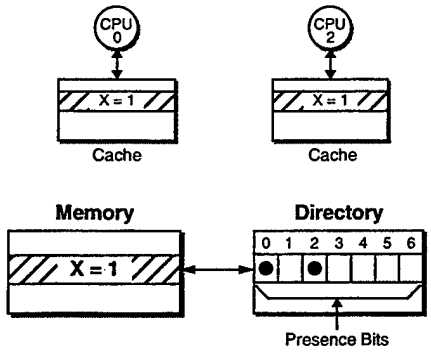
sgi 6

Cache Coherence Problem 2



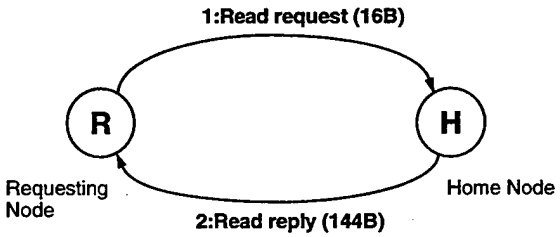
sgi

Cache Coherence Solution: Directories



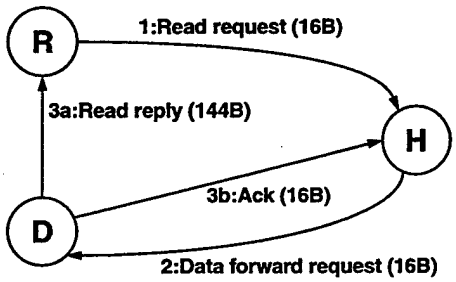
sgi

Distributed Cache Coherence 1



sgi

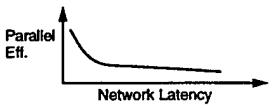
Distributed Cache Coherence 2



sgi

Cache Coherence Implications

- Protocol messages are small
- Network *latency* is crucial
  - CPUs do not tolerate large memory read latencies (getting better)
  - Impact tends to be non-linear

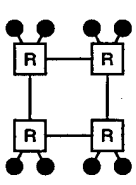


- Network *bandwidth* also important, but less so
  - Needed at basic level to match memory and CPU bus bandwidth
  - Needed to cope with communication bursts (typical, especially in scientific codes)

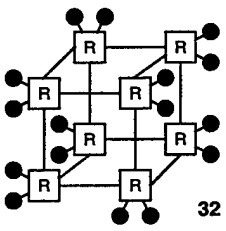
sgi

Cache Coherence in Action: SGI Origin 2000

- Processing nodes contain 2 R10K CPUs, memory, and network interface
- Network implemented via 6-port routers (details to follow)
- Hypercube system topology; often multiple routers per network traversal



16 CPUs



32 CPUs

sgi

Origin 2000 Network Detail 1

- Router characteristics
  - 6 ports connected via a crossbar
  - Each port bidirectional at 800 MB/s in each direction
  - Provides 4 virtual channels
  - Input-buffered with pipelined crossbar arbitration
  - Best-case (fall-through) input-to-output latency of 50 ns
- Links (per direction)
  - 20 data bits, 2 clock bits (differential), 1 data framing bit
  - Clock rate of 200 MHz, sampled on both edges (400 MHz data rate)
  - Credit-based flow control
  - Sliding-window, CRC-based error detection/retransmission

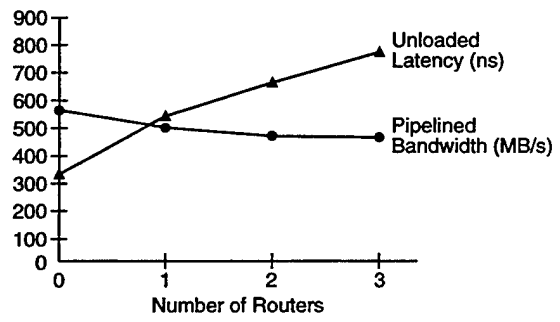
sgi 13

Origin 2000 Network Detail 2

- Implementation
  - 850K-gate ASIC
  - IBM CMOS 5L (0.5µ drawn), 5 metal layers
  - 160 mm^2 die area
  - Core operates at 3.3V, 100 MHz
  - 29 W worst-case power dissipation
- Cables
  - Shielded, electrically-matched wires, 1–5 m
  - Expensive :–(

sgi 14

Origin 2000 Latency/Bandwidth Characteristics



sgi 15

Future Trends

- Communication ever-more critical to overall system performance
- Bandwidth demands growing
  - CPU bandwidth growing, both of system bus and functional units
  - Memory system bandwidth growing: SDR, DDR, DRDRAM
- Network latency becoming more of a problem
  - Decreasing in absolute time
  - But increasing when measured in CPU instruction issue slots
  - Latency impact on overall performance is non-linear
- Will interconnection network become primary limit on overall system performance?

sgi 16

Trend: Merging of Network Interface and CPU

- Desire to move network interface "closer" to the CPU
  - Architecturally
    - User-level, protected access ("OS bypass")
    - Tied more closely to memory system (address translation, etc.)
  - Physically
    - Place on same die as CPU
    - Direct datapaths between CPU internals and network interface
- Challenges
  - Development of reasonable interface to user jobs
  - Electrical, mechanical, physical integration of CPU logic and network interface

sgi 17

Trend: Active Networks

- Current multiprocessor networks are "passive"
  - Message unchanged as it flows through network
  - Network does not interpret message contents
  - Result: network acts mainly as a delay element (though a useful one!)
- Idea: perform computation *in the network* as well as on CPUs
- Benefits
  - Moves computation closer to the data on which it operates
  - Offloads CPUs
- Challenges
  - Programming model, compiler and OS support, protection, etc.
  - Details of computational resources, integration into network fabric, etc.

sgi 18

Conclusions

- Interconnect is a key component of multiprocessor system performance
- Interconnect latency and bandwidth are both important
  - Low latency especially critical for cache coherence
  - Bandwidth for message passing, clustering, traffic bursts
- Future interconnects must continue to improve latency and bandwidth
  - By coupling the network more closely to the CPU
  - By (eventually) making the networks "active"

8:45am - 9:15am
Tues, 11 May - 2.2

The Role of Optics in Balanced Computer System Design

Mike Chastain
Hewlett-Packard
chastain@rsn.hp.com

Parallel Fiber Optic Development

The advantages of parallel fiber versus copper interconnect are well known
- Physical size reduction at high frequency
  - Connectors and cables
- Greater communication distance
- Reduced susceptibility to EMI and EMC

Computer industry has watched parallel fiber development for five years
- Costs have always prevented wide spread system insertion

Computer industry is also slow to adopt new interconnect technologies
- Waits for technology cost crossover; or some "external" forcing function

Industry investment is making parallel fiber more viable
- Real products now appearing from multiple vendors
- Costs are starting to come down
  - Breakthroughs in manufacturing and packaging
  - Optimistic projections of high volume insertion
- Costs are still high relative to copper for short (<10m) links

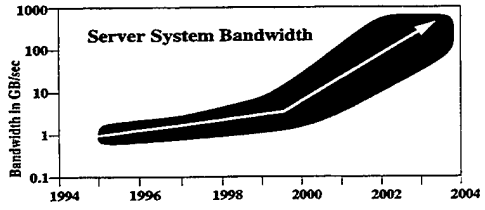
Are there other forces, outside optical development, that may hasten insertion?

Consider CPU Performance

The industry is now increasing CPU performance at an exponential rate
- Single chip CPUs breaking the Gigahertz barrier, and beyond
- Single chip CPUs incorporating "super-computer architecture tricks"

Increasing CPU performance driving corresponding increase in bandwidths
- Soon CPUs may require 8 GB/sec, or more, to sustain performance

Increasing bandwidths forcing maximum frequency at all CPU and ASIC pins
- Designers struggling to maintain reasonable pin counts for manufacturing
- Intel's endorsement of Rambus is an indication of a pervasive problem



Consider Copper Interconnect Limits

HEWLETT PACKARD

At today's interconnect frequencies (up to ~1 Ghz)
- Primary frequency dependent loss mechanism is skin effect
  - Proportional to $\sqrt{f}$

At interconnect frequencies beyond 1 Ghz
- Dielectric loss starts to dominate
  - Impact greater; Dielectric loss increases linearly with f

At interconnect frequencies approaching 2.5 Ghz
- Interconnect distance may be limited to a single backplane or PC planer
  - New low loss PCB materials will be required

At interconnect frequencies approaching 5.0 Ghz
- PCB interconnects may no longer practical

Copper cables are still an option; for now
- Designers will trade copper trace for cable to increase interconnect distance
  - 4-5" of PCB trace is roughly equivalent in loss to 3 feet of copper cable
- Parallel copper cables will still be limited to adjacent racks
  - Six to ten meters at 622 Mhz, dropping linearly with frequency

Machine room level interconnects are already in jeopardy without parallel fiber!

Consider Server Packaging Density

HEWLETT PACKARD

Increased interconnect frequencies coupled with greater interconnect losses
- Driving system designers to reduce interconnect distances
  - Driving system designers to increase system packaging density

To achieve increased packaging density
- More ASIC integration to minimize component count
- More CPUs, ASICs, and RAM per PCB area

(Frequency * Density) is driving power density to the limits
- More gates at greater frequency ==> more power density!
- More high speed I/O ==> more power density!
- More CPUs, ASICs, and RAM per PCB area ==> more power density!

System power density will soon exceed machine room limitations
- By 2002-3, (4 CPUs + ASICs + 16 GB DRAM + Power) => ~850 watts
- Existing rooms are designed for 40-70 W/sq.ft. with an 18" raised floor
  - Floor area + service area => 19" rack occupies ~14 sq.ft. => 980 watts max
- New standards are still inadequate (125 W/sq.ft. 36" raised floor suggested)

Result: Packaging density limited by machine room for foreseeable future

Future Server Designs ?

HEWLETT PACKARD

May therefore consist of medium (CPU count) SMPs
- High frequency signaling on all interconnects
- High integration (and high power) silicon
- High density packaging, power input, and power dissipation

Tightly coupled electrically, but not physically
- Tightly integrated coherent cable interconnects
  - Utilize copper until frequency-versus-distance becomes prohibitive
  - Shift to optical as frequency increases and/or costs come down
    - Frequency may cause a shift in spite of costs!
- Shift to optical now for machine room level interconnects
  - Such as the emerging Future I/O standard

To balance system performance versus machine room constraints
- Spread system across multiple racks to distribute thermal load
  - Match machine room capabilities (power/sq.ft. electrical and thermal)
- Perhaps integrated with storage and I/O components
  - Good volume utilization without adding significant power/sq.ft.

The Role of Optics in Server Evolution

HEWLETT PACKARD

Optical interconnects within servers
- Evolutionary "copper replacement" strategy as frequency increases
  - System architects must work closely with the optical link community
    - Copper link designs must be compatible with optical link limitations
    - Optical components must be compatible with server manufacturing
    - Optical packaging must be consistent with server connector requirements
- Evolutionary "EMI/EMC management" strategy as frequency increases

Optical interconnects between servers and/or I/O within a machine room
- Addressed by the emerging Future I/O standard
  - Parallel optical links at 1/2/4 GB/sec data delivery; up to 300m (@1GB)
  - Network like protocols optimized for both cluster and I/O communication
  - Designed for highly reliable, fault tolerant communication
  - Designed to enable sharing of storage, networks, and other I/O

Optical interconnects between machine rooms
- Still addressed by existing and evolving LAN/WAN infrastructures
  - Tightly bridged to the emerging Future I/O standard

Server Architects must Design for Optics

HEWLETT PACKARD

Server architects are starting to design within limits of available optics

- Accepting 12 bit wide links as a cost effective limit
  - Leveraging telecom volumes for connectors and cables
- Accepting per-bit encoding and self-clocking for AC coupled links
  - Designing in clock recovery and link training sequences
- Accepting multiple bit time skews between end points
  - Performing parallel word re-assembly in end points
- Accepting a "non-zero BER" at high frequencies
  - Designing in transparent link retry and ECC recovery mechanisms

Optical Vendors must Design for Servers

HEWLETT PACKARD

Optical link frequency has been driven by the telecom industry
- Telecom road map is 4x per generation; 622Mhz, 2.5Ghz, 10Ghz
- Server road map is 2x per generation; 1.25Ghz, 2.5Ghz, 5Ghz, 10Ghz(?!)

Optical link packaging is not consistent with server environments
- Server power is generally noisy; Optical links want clean power
- Servers (most) rely on forced air convection for thermal management
  - Optical interfaces cannot assume heat conduction to PCB
- Server manufacturing relies on robotic assembly and test
  - Optical interfaces should support standard pick&place BGA processes
- Servers need blind-mate optical connectors; with EMI containment!
  - 2nd level assemblies to accomplish blind-mate/containment are expensive

Servers need "transparent" optical links
- Server silicon must be re-used; copper links may become optic links
  - Different "products" must make different distance-cost trade-off
- Electrical interfaces consistent with (same as) copper cable interfaces
  - Same frequency, encoded self-clocked, low voltage differential interfaces
- Few (if any) special system considerations beyond equivalent copper cable
  - Example: special system requirements to handle "eye safety"

Summary

HEWLETT PACKARD

Parallel optical links are finally close to reality, but costs are still high

Parallel optical links (and Future I/O) will address machine room interconnects

CPU frequency and associated dielectric losses will drive server density upward
- **But, existing machine room capability will limit the power density per sq.ft.**
- **Future servers may trade PCB trace for cables and distribute the power density**
- **But dielectric loss also reduces copper cable length proportional to frequency**

Therefore server designers may have (non-cost) reason to use internal optical links

Server and optical link designers must work together to enable a smooth transition

9:15am - 9:45am
Tues, 11 May - 2.3

In Pursuit of a Petaflop: Overcoming the Bandwidth/Latency Wall *with PIM Technology In the HTMT Project*

Dr. Peter M. Kogge
McCourtney Prof. Of CS & Engineering
IBM Fellow, IEEE Fellow
CSE Dept., Univ. of Notre Dame

Thesis

- Modern technology: The *Memory Wall*
  - Latency: cannot access data fast enough
  - Bandwidth: cannot get data to logic fast enough
- Next level of supercomputing- *Petaflops*:
  - Impossible without radical change
- A direct assault on the problem - *HTMT*
  - Hybrid Technology, MultiThreaded
  - Mix memory & logic, interconnect optically

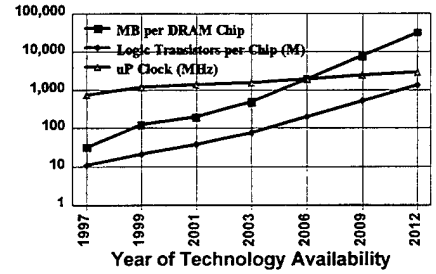
Computer Performance

- 1 *Teraflop* = 10^12 Flops/second
  - = 1,000 "peak" 1999 workstations
- Fastest Computers in world today:
  - ASCI Red: 1 Teraflop peak
  - ASCI Blue: 3 Teraflops peak
- 1 *Petaflop* = 1,000 Teraflops
  - What a 1 GF machine can do in 30 years - takes a Petaflop Machine *15 minutes*

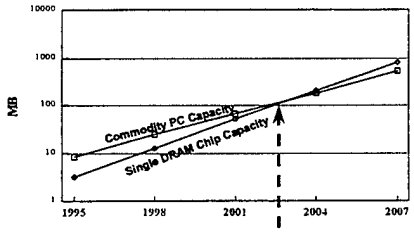
What are *Principal Barriers* to Petaflops?

- Physics Driven:
  - Memory Density
  - Memory Bandwidth
  - Memory Latency
- Semantics Driven (Programming Model)
  - Expressing million way parallelism
  - Without global synchronization

The SIA CMOS Roadmap

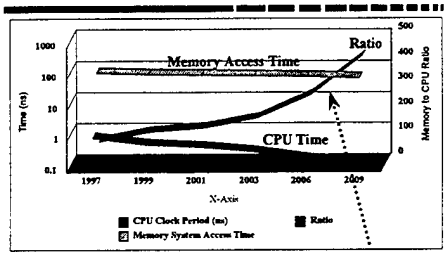


Memory Chips/PC



ONE DRAM chip satisfies *ENTIRE* PC demand by 2003;

Latency in a Single System



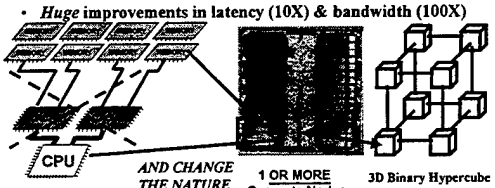
THE *WALL*

Bandwidth

- Today's CPUs: need upwards of 3 billion bytes of data per second
  - *And almost doubling each year!*
- Today's cheap memory: at most 1/30 of this
  - *And increase at only perhaps 7% per year*
- Problem! - Need either
  - Lots of chips
  - Lots of pins on each chip
  - Both

Solution: *Processing-In-Memory*

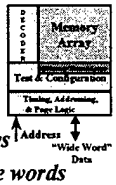
- Mixing Significant *Logic* and *Memory* on same chip
- *Huge* improvements in latency (10X) & bandwidth (100X)



=> Opportunities for *New Scalable Architectures*

PIM Key: Memory Macro

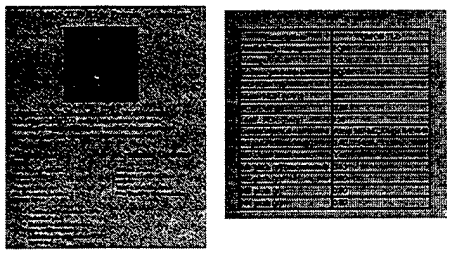
- Multi MB independent units
  - Separate address decoding
  - Separate refresh
  - Separate test and redundancy
- Sub 25 ns access to 1K-4K *full rows*
- Sub 10 ns access to 128-1024b *wide words*
- 2.4 - 4+ GBps bandwidth potential *per macro*



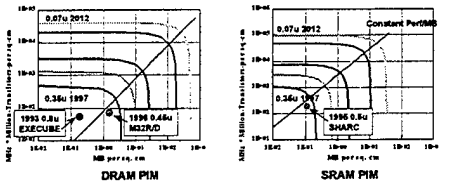
Recent PIM Technology Offerings

- IBM 7LD, 10/97
  - 0.25u DRAM-based process with 5 LM
  - 2 MB/macro in less than 30 mm²
- Silicon Access, 3/98 (www.siliconaccess.com)
  - IP ownership of highly variable embeddable DRAM macro
  - TSMC & UMC alliances for 0.25u
  - 2 MB/macro in less than 23 mm²
- Samsung, 7/98 (www.samsungsemi.com/products/asic/embedded-dram.htm)
  - 0.25u ASIC, stacked capacitor configurable DRAM, 5 LM
  - Up to 128 Mb/ chip with mix of SRAM, Flash, other macros
- IBM SA-27E, 3/99 (www.chips.ibm.com/news/1999/sa27e)
  - 0.18u *logic process* with embedded trench DRAM array, 6 LM *copper*
  - 2 MB is 20 mm², < *13ns access, 33 ps* gates
  - ASIC design flow, rich library of macros

A Real DRAM PIM Technology



Design Space: Memory "+" Logic

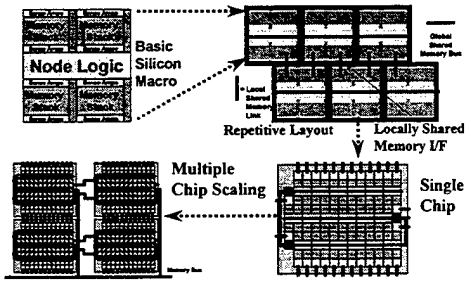


PIM VLSI constraints (as of today):
- Approx. 1 year lag between SIA announce & volume PIM parts
- DRAM logic: 1/2 speed of conventional, but approaching same levels of metal!

Key PIM Subsystems

- *Stacks* of on-chip memory
- Key Subsystems that make up logic
  - *At the Sense Amps Processing* Logic (ASAP)
  - "CPU-like" local GP processing (Core)
- *Node*: Memory stack + ASAP + Core + I/F
  - Potentially many nodes per PIM chip
- Interfaces
  - Intra chip links between nodes (floorplanning)
  - Inter multiple PIM chip systems
  - Between PIM and non-PIM subsystems

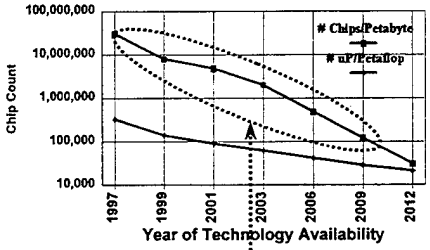
SHAMROCK: Floorplanning for Hugh Inter Node Bandwidth



Current ND PIM Projects

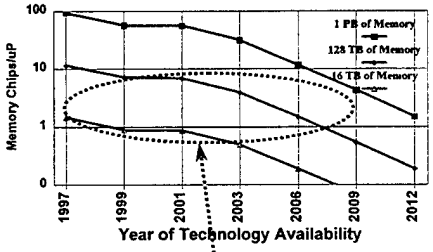
Project	Funded By	Partners	Goal
PIM Fast	JPL	LM	Build Space Computer
DIVA	DARPA	USC, CalTech	Build Smart Memory
HTMT	NASA, NSA	Cal Tech, Princeton, SUNY, Dela. ...	Design 1st Petaflop Computer

A Peta Machine: The SIA Way



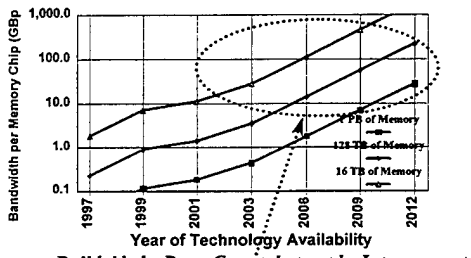
It's The Memory, Stupid!

Ratio of Memory Chips to Microprocessors



Memory Light but Decent Ratio & Parts Count

The Bandwidth Problem for Peta Machines

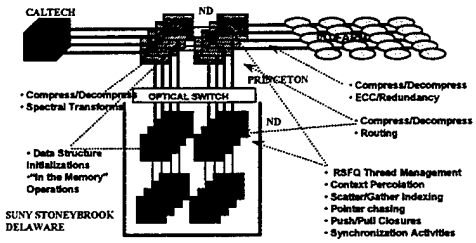


Buildable by Parts Count, but not by Interconnect

HTMT: A Petaflop in 2004-2006 Timeframe

- Multi-Institution program dating back to 1994
- Harness emerging technologies
  - RSFQ for 100GHz "CPUs"
  - PIM for smart memory hierarchy
  - WDM all optical network for interconnect
  - Holographic memory for dense storage
- With latency-tolerant architectures
  - MultiThreading
  - Parcels for "in the memory" function execution

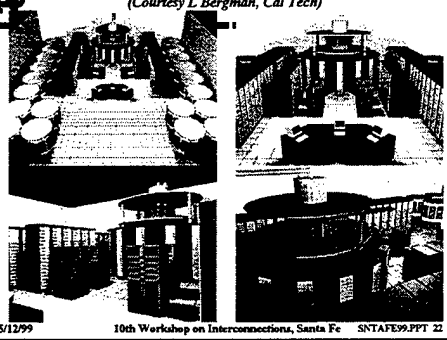
The HTMT System Architecture



PIM: Linchpin of System Latency/Parallelism Reduction

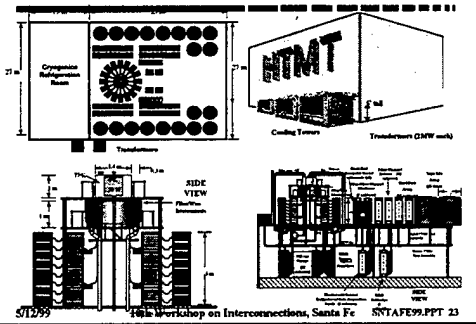
The HTMT Machine Room

(Courtesy L Bergman, Cal Tech)

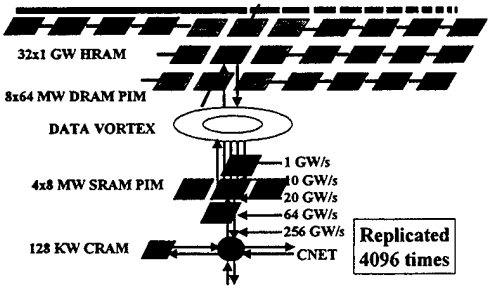


The HTMT Facility

(Courtesy L Bergman, Cal Tech)

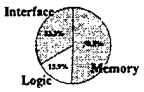


A Replicable HTMT Unit

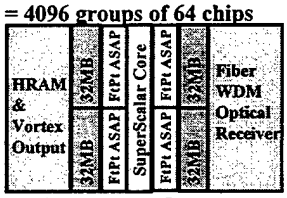


Silicon Budget for HTMT DRAM PIM

- Designed to provide proper balance of memory & support for fiber bandwidth
  - Different Vortex configurations => different #s
- In 2004, 16 TB = 4096 groups of 64 chips
- Each Chip:



By Area

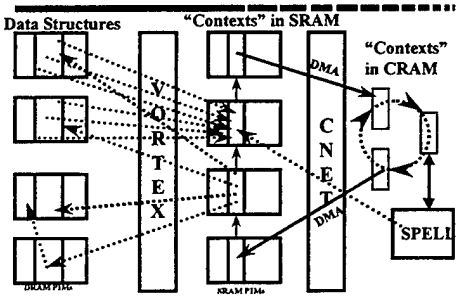


The View from a SPELL

	Capacity Words	Capacity W/Flop	2 Way Latency	Read BW W/Flop	Write BW W/Flop
Local CRAM	128K	0.000005	70	1	1
Local SRAM	32M	0.000125	240+	0.25	0.25
Remote CRAM	512M	0.002	270	0.08	0.08
Remote SRAM	128G	0.5	420+	0.08	0.08
Single DRAM Cluster	512M	0.002	16,000	0.04	0.04
HRAM from 1 Cluster	32G	0.125	67,000	0.04	0.04
4 DRAM Clusters	2G	0.008	16,000	0.16	0.04
HRAM from 4 Clusters	128G	0.5	67,000	0.15	0.04

SPELL can never "miss"; PIM must "guess" first!

HTMT Execution Model



SRAM PIM Functions

- Initiate Gather/Scatter to/from DRAM
- Recognize when sufficient operands arrive in SRAM context block
- Enqueue/Dequeue SRAM block addresses
- Initiate DMA transfers to/from CRAM context block
- Signal SPELL re task initiation
- Prefix operations like Flt Pt Sum

DRAM PIM Functions

- Initialize data structures
- Stride thru regular data structures, transferring to/from SRAM
- Pointer chase thru linked data structures
- "Join-like" operations
- Reorderings
- Prefix operations
- I/O transfer management
  - DMA, compress/decompress, ...

Conclusions

- The Twin Demons: Latency & Bandwidth
- PIM Technology: Solves the local problem
- Petaflops: Global problem still present
- HTMT: Attach global problem by:
  - Making memory smart so many transfers only "one way"
  - Utilizing best of emerging optical technology for bulk of remaining

9:45am - 10:15am
Tues, 11 May - 2.4

Ultra-High Speed Interconnections Network for Supercomputing

Keren Bergman
Princeton University

Sponsors: DOD, JPL/NASA, Caltech, NSF, ONR

---

technologies research group:
Mark Arend
Nathan Kutz (Univ. Wash.)

Brandon Collings (now Lucent)
Jeff Roth
Qimin Yang
Suzanne Sears
Ricky Leng
Dmitriy Krylov
Keir Neuman

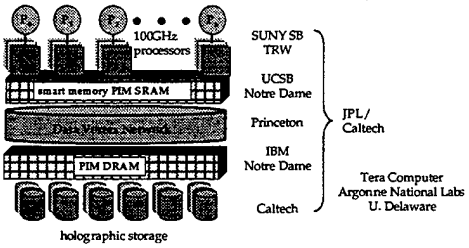
architecture and system design:
Coke Reed (PU Math/IDA)
Charles Fefferman (PU Math)
John Hesse (Interactic)

Outline

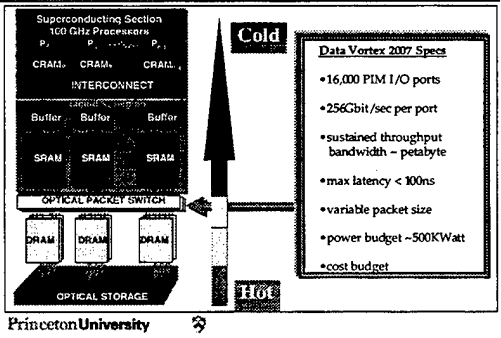
- HTMT Petaflops project overview
- Data Vortex optical network
- Experimental test bed
- Modular packaging and interfaces
- Summary & future directions

Hybrid Technology Multi-Threaded Project Overview

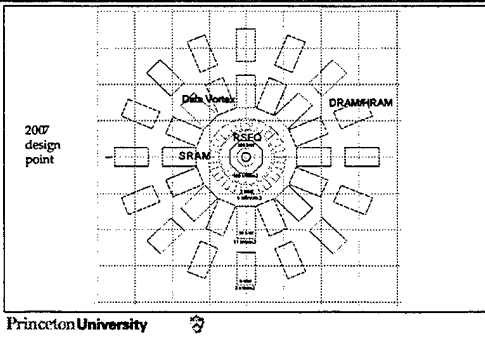
- achieve Petaflop computing in 10 years with hybrid mix of exotic technologies
- very high speed processors (100GHz), deep memory hierarchy
- employ multi-threading techniques for aggressive latency management



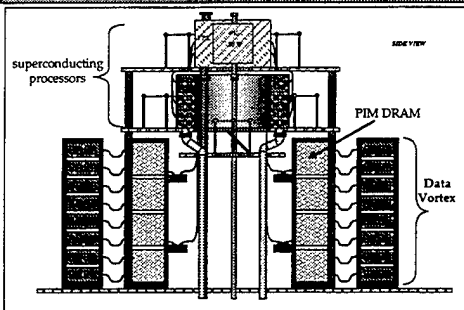
HTMT Machine Architecture



Top Down View of HTMT Machine



2007 HTMT Machine Hardware



PrincetonUniversity

Advantages & limitations of optical interconnect

- **ADVANTAGES:**
  - very high throughput bandwidths
  - potential transparency to data format
  - can scale I/O capacity by TDM/WDM
- **LIMITATIONS:**
  - cannot perform sophisticated routing logic
    - leads to long latencies or resort to scheduler
  - difficult buffering with random access
    - forces one data format
  - expensive components, low levels of integration

unless demand for bandwidth per I/O and throughput is high
➡ use electronic switching networks

PrincetonUniversity

Optical network design requirements for computing

- Low latency (<100ns) between any PEs
- Flat structure, no differentiation between local and remote
- Narrow latency distribution, small variance, smooth structure
- Transparency to payload modulation (TDM/WDM)
- Fine grained access, high efficiency for short packets
- Scalability to >10k PEs with minimal injection penalty
- Feasible implementation, commercially available technologies

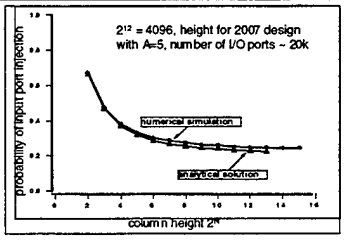
PrincetonUniversity

Unique features of the Data Vortex topology

- Network designed specifically for optics:
  - absolute minimal logic at the nodes
  - no contention resolution at the nodes
  - bufferless
  - transparent to payload data structure
  - implemented with commercially available tech
- Topology and data flow:
  - scales without performance sacrifice
  - simplifies node function (no all-optical switching)
  - hierarchy accomplishes fairness
  - deflection recovered in two node hops
  - timing used to reduce logic
  - low latency and latency variance

PrincetonUniversity

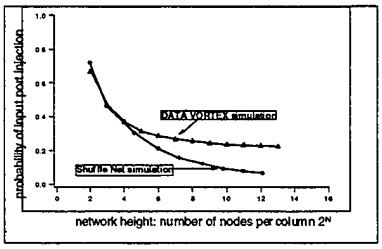
Scalability of Data Vortex



analytical solution - assumes statistically independent packet injection
numerical solution - Monte Carlo

PrincetonUniversity

Scaling comparison with Shuffle network



2^{12} = 4096, height for 2007 design

PrincetonUniversity

Latency Distribution

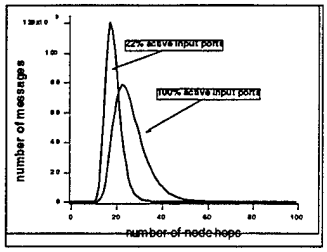
network size = 4096

Narrow latency distribution:

packet arrival nearly in sequence

important for data, voice, and video communications

best operation at 20% injection, knee of performance curve



Princeton University

Data Vortex Network Performance

NETWORK SIZE ⇐ SUSTAINED THROUGHPUT

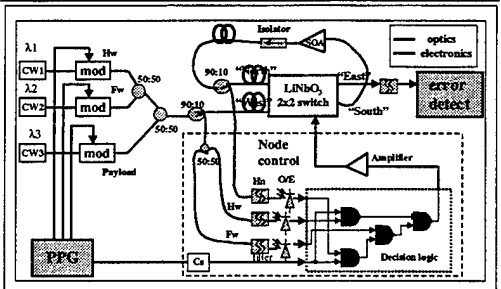
- Sustained (incoming) cross-sectional bandwidth under fully loaded conditions
- Capacity per input port ⇒ number of ports required
- Network dimensions: height (N), angle(A), cylinder (C)

  Number of nodes: $(A)(N)(\text{Log}_2(N)+1)$
  Number of cylinders: $\text{Log}_2(N)+1$

- For nearly no deflection at injection ports, use 22% of available ports for input number of inputs: $(A)(N)/5$

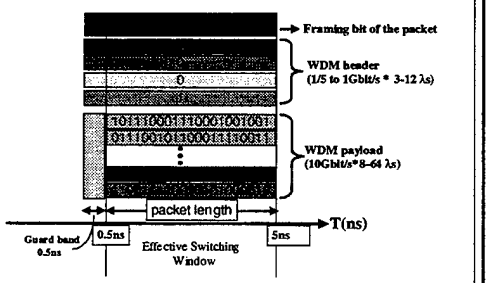
Princeton University

Node Test Bed



Princeton University

Packet modulation format



Princeton University

Payload (8Gbit) eye diagrams with SOA

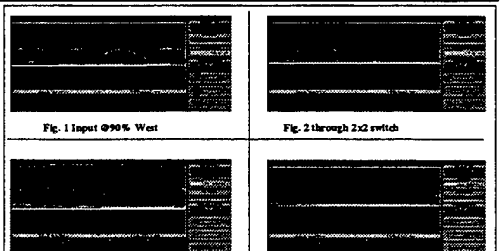


Fig. 1 Input @90% West

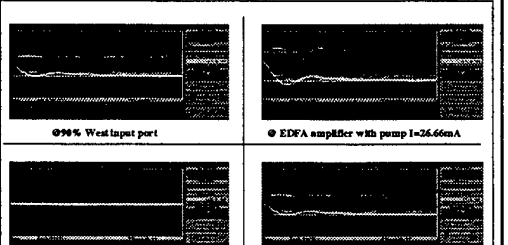
Fig. 2 through 2x2 switch

Fig. 3 Amplified by SOA

Fig. 4 @ 90% North

Princeton University

Payload (8Gbit) eye diagrams with EDFA



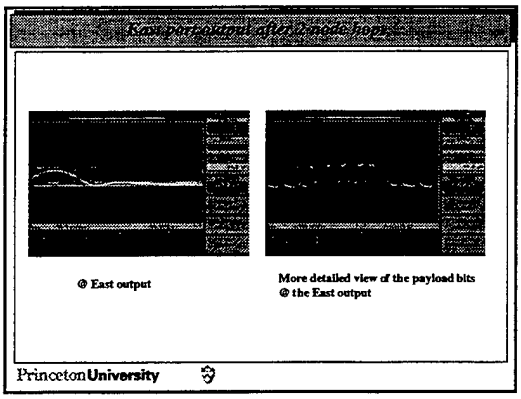
@90% West input port

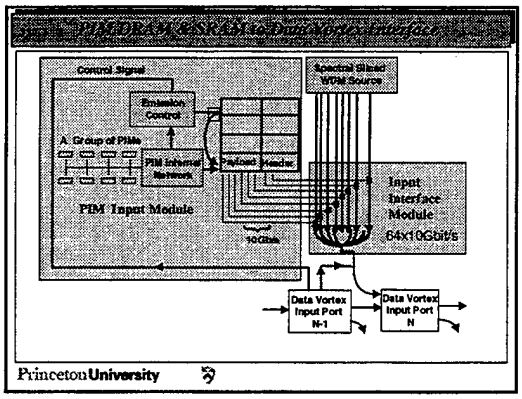
@ EDFA amplifier with pump I=26.66mA

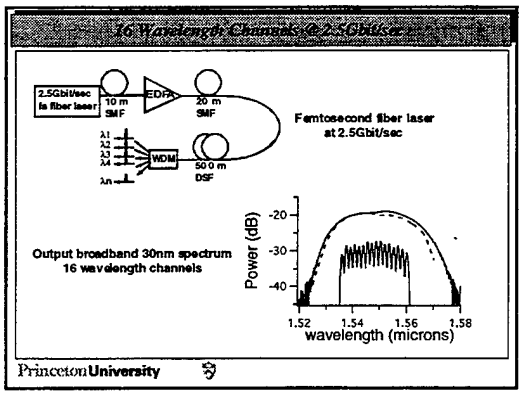
@ South port

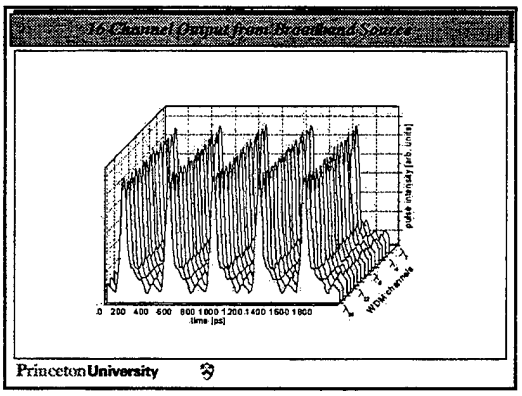
@90% North loop-back input

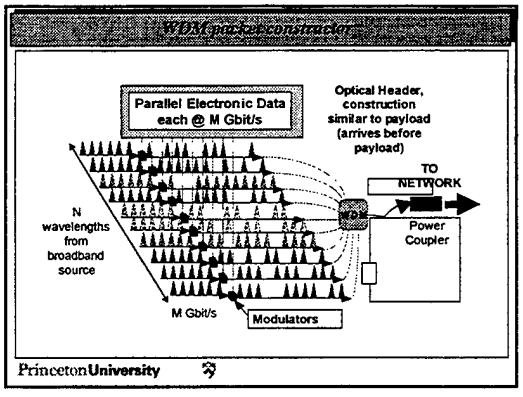
Princeton University

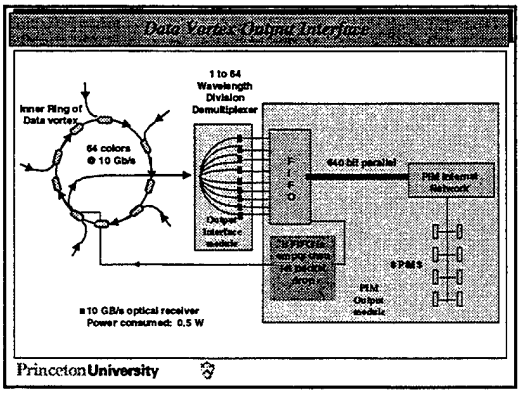

@ East output
More detailed view of the payload bits
@ the East output
PrincetonUniversity


Control Signal
Emulation Control
A Group of PIMs
PIM Internal Network
PIM Input Module
Input Interface Module
64x10Gbit/s
Data Vortex Input Port N-1
Data Vortex Input Port N
PrincetonUniversity

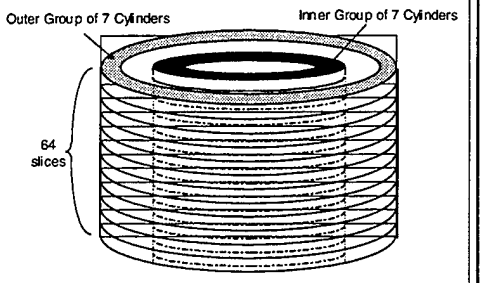

2.5Gbit/sec fs fiber laser
10 m SMF
EDFA
20 m SMF
WDM
50 0 m DSF
Femtosecond fiber laser at 2.5Gbit/sec
Output broadband 30nm spectrum
16 wavelength channels
Power (dB)
wavelength (microns)
PrincetonUniversity


PrincetonUniversity


Parallel Electronic Data each @ M Gbit/s
Optical Header, construction similar to payload (arrives before payload)
TO NETWORK
N wavelengths from broadband source
Power Coupler
M Gbit/s
Modulators
PrincetonUniversity

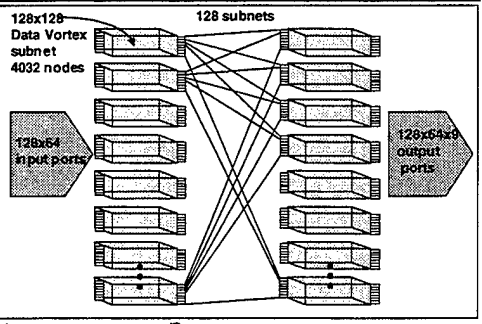

1 to 64 Wavelength Division Demultiplexer
Inner Ring of Data vortex
64 colors @ 10 Gb/s
640 bit parallel
PIM Ethernet Network
@ PIMS
a 10 GB/s optical receiver
Power consumed: 0.5 W
PrincetonUniversity

Modular Packaging Design: New Topology



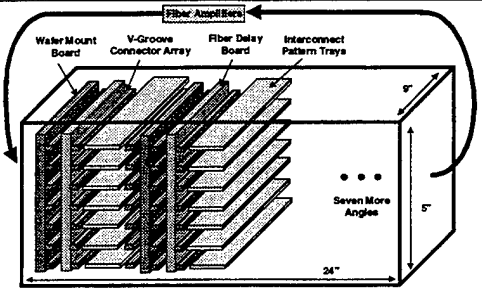
Princeton University

Sub-networks Arrangement



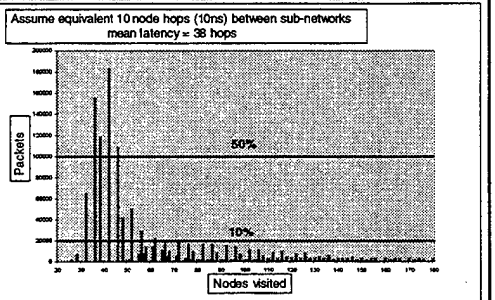
Princeton University

Sub-Network Design



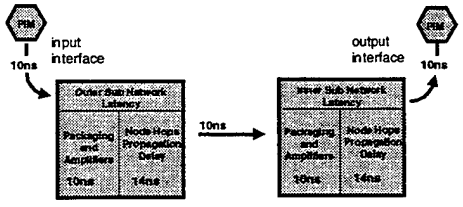
Princeton University

Latency Distribution Modular Topology



Princeton University

Latency through Modular Network



MEAN LATENCY: 78ns

Princeton University

Summary and future directions

- Data Vortex network topology for optical TDM/WDM packet switching implementation
- Simple node logic, low latency, narrow latency distribution, scalable ⟹ promising for computing
- Sustained Petabyte throughput
- Experimental test bed demonstrates WDM routing feasibility
- Modular packaging for HTMT machine
- Implement 16x16 switching fabric as network test-bed

Princeton University

10:30am - 11:00am
Tues, 11 May - 2.5

Ultrafast Optical Interconnect Based on Routing by "Clockwork" in Regular Mesh Networks

D Cotter[1], F Chevalier[2] and D Harle[2]
[1]BT Laboratories, UK
[2]University of Strathclyde, UK

BT

1

Introduction

- Ultrafast interconnection network for multi-processor systems (future massive-capacity routers and servers)
- Multi-stage packet-switched network
  - fixed-length packets, serial bit rates 0.1–1 Tbit/s
  - routing and header processing 'on the fly' in the optical domain
  - no buffering in the optical domain
  - contention-free in the optical domain
- Routing and processing mechanisms as simple as possible
  - regular topology
  - 'clockwork' routing

2

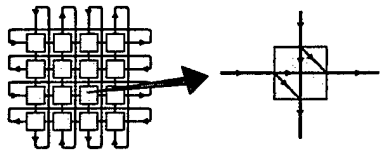
Outline

- 'Clockwork' routing in the Manhattan street network
- Node architecture
- Performance
- Special properties significant for future high-performance multi-processor systems
  - ultra-low latency signalling (e.g. acknowledgement)
  - bandwidth reservation
  - process scheduling

3

Suitable topologies [animated slide]

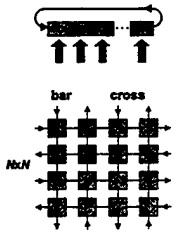
- Eulerian network digraph
  - decomposition into a set of distinct closed directed trails
- Manhattan street network
  - every node is topologically equivalent



4

Global states and transitions

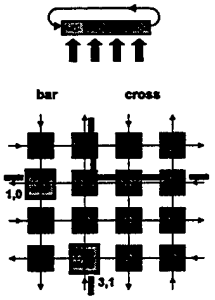
[animated slide]



- Link length = (qN+1) time slots, q=0,1,2,...
  - a packet that leaves a node in time slot j, arrives at the next node in time slot (j+1)mod$N$
- Packet is routed automatically by 'clockwork'
  - no routing decisions needed at intermediate nodes

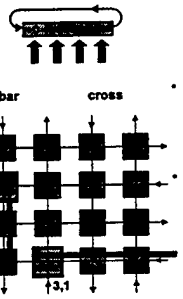
5

Routing by 'clockwork' [animated slide]



6

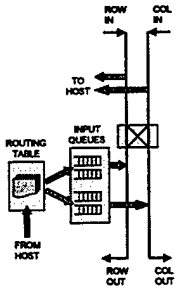
Routing by 'clockwork' [animated slide]



- Single routing decision at source node only
  - placed onto appropriate trail (chosen by time slot and outbound link)
- Intermediate nodes only perform 'me-or-not-for-me' selection

7

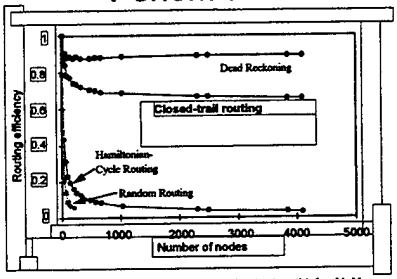
Node architecture



- Contention resolved using peripheral queues (in electronics)
- No optical buffers anywhere inside network
- Optical layer is contention-free
- Free of deadlock and livelock
- Packets delivered in correct sequence
- No 'hidden' queues
- Source has direct visibility of the state of the local queues
  - localised access and flow control mechanisms
  - reduces delay
  - improved handling of bursty traffic

8

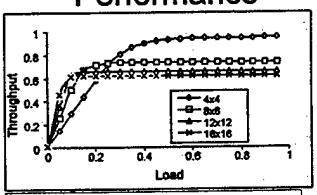
Performance



Routing efficiency (relative to shortest-path) for *NxN* Manhattan street network

9

Performance

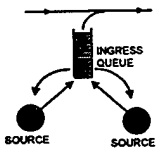


Throughput (relative to store-and-forward) for *NxN* Manhattan street network

- Throughput comparable to store-and-forward
- Greater absolute throughput (packets delivered per second)
- Optimal delay performance with small peripheral queues (depth=5)

10

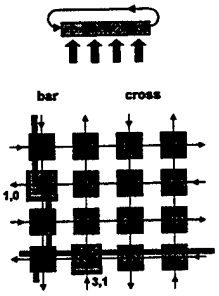
Bursty (self-similar) traffic



- No hidden queues inside network
- Hosts have visibility of ingress buffer state
  - local, fast rate-control mechanisms

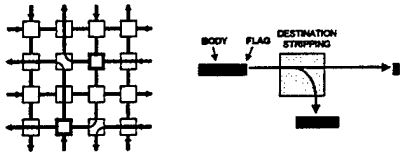
11

Automatic return trail [animated slide]



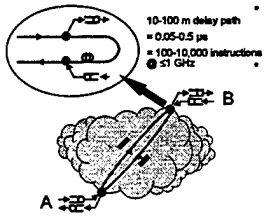
12

Signalling



- Acknowledgement with ultra-low delay
- Sojourn time of flag is precisely fixed (no optical buffers or deflections)
- Identification by return arrival time

Bandwidth reservation and scheduling



- Closed directed trails used for fast simple bandwidth reservation
- Deterministic time of arrival
  - direct transfer into system memory for pre-scheduled process
  - possible close integration of 'clockwork' routing and processor scheduling strategy in multi-processor systems

Summary

- Strategy for optical packet routing in high-speed mesh interconnect
  - throughput comparable with store-and-forward
  - trivial processing at nodes
  - no hidden queues inside the network
    - effective access control for bursty (self-similar) traffic
  - automatic return-path routing
    - ultra-low latency signalling
    - process scheduling

LARGE-SCALE PHOTONIC PACKET SWITCH USING WAVELENGTH ROUTING TECHNIQUES

Koji Sasayama

NTT Network Innovation Laboratories

Outline

WDM star-based switch architecture

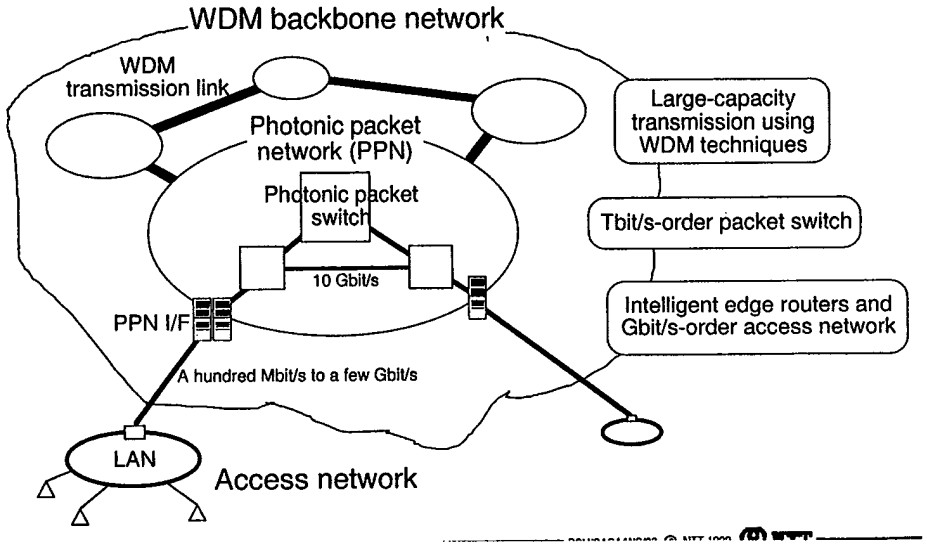
Photonic WDM highway layer

Electrical control layer

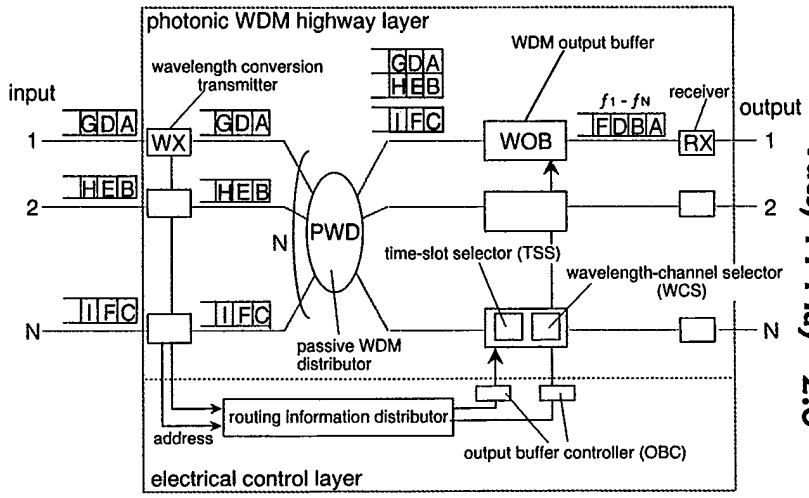
Rack-mounted prototypes

Conclusion

Photonic packet switch in WDM network

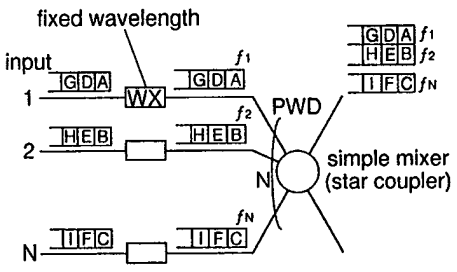


WDM star architecture for large-scale packet switch

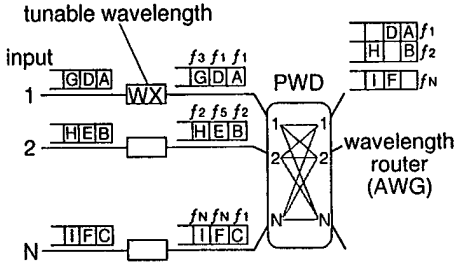


11:00am - 11:30am
Tues, 11 May - 2.6

Two alternative configurations of WX/PWD

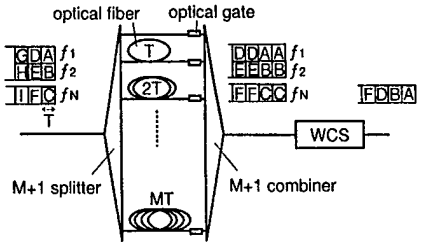


Broadcast-and-select type switch

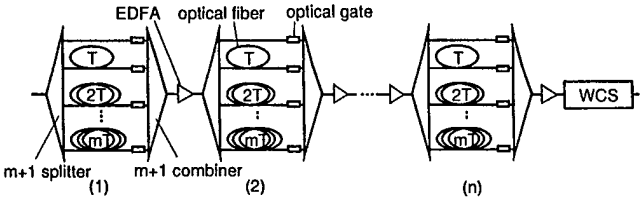


Wavelength-routing type switch

Configuration of time-slot selector (TSS)

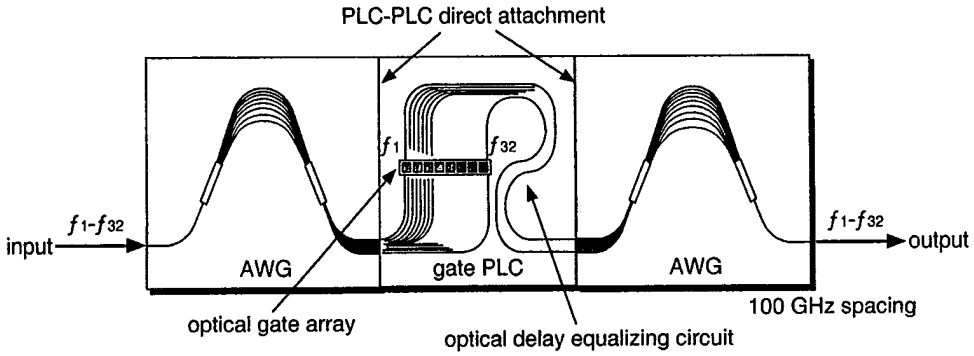


Large TSS

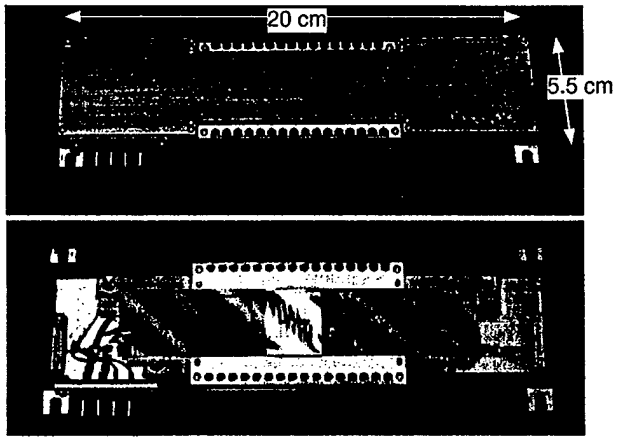


Cascaded small TSS units (nm = M)

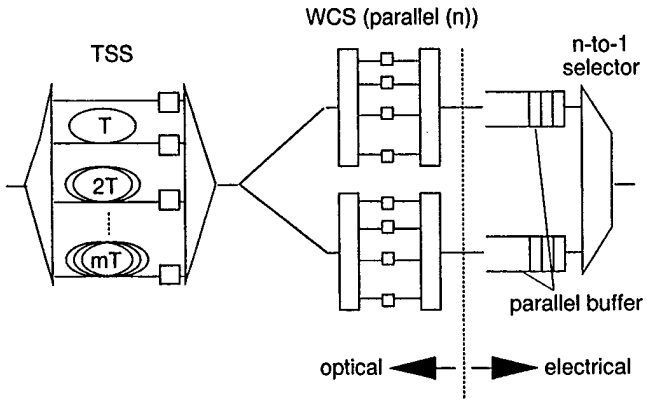
Configuration of 32-ch wavelength channel selector (WCS) module



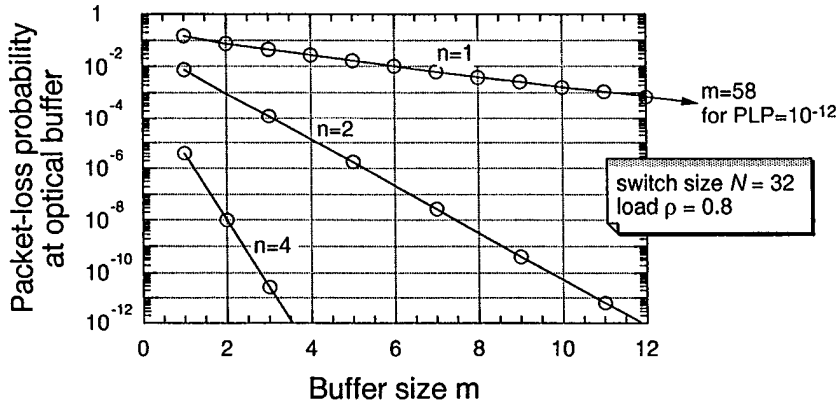
Photograph of fabricated WCS module



Composite optical/electrical buffer configuration

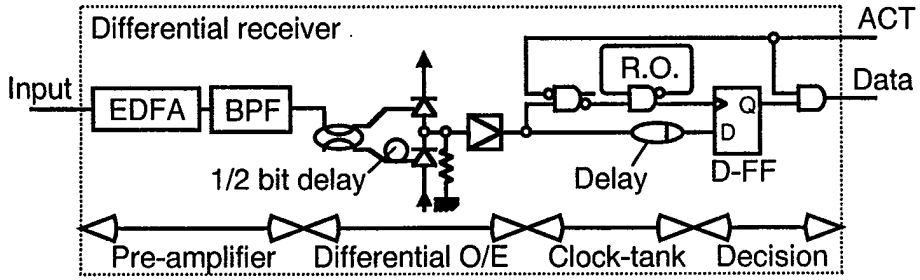


Packet loss in composite output buffer

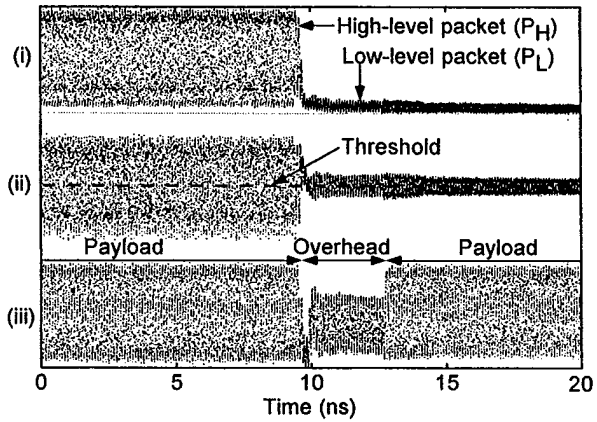


n: number of outputs from optical buffer

Configuration of burst-packet receiver with clock-tank circuit

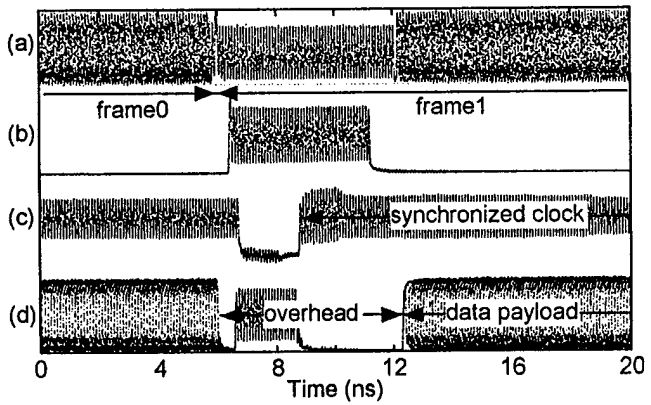


Packet reception for large power fluctuation



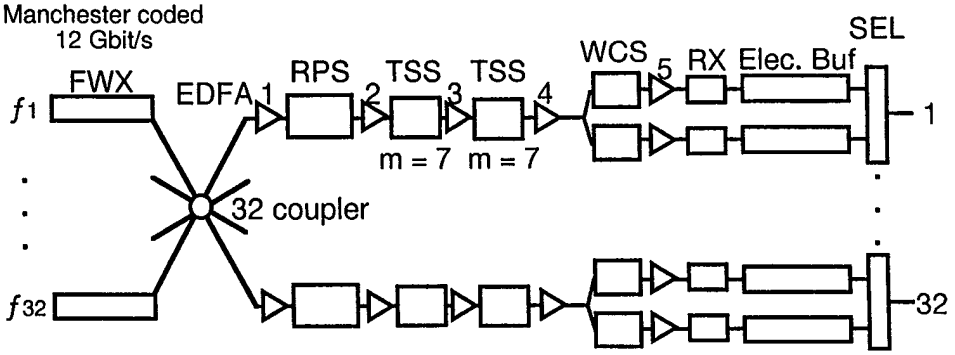
(i) Optical input signal
(ii) Differential
(iii) Regenerated signal

Packet reception for phase fluctuation



(a) Optical input signal
(b) Digital-ring oscillator input signal
(c) Regenerated clock
(d) Regenerated data

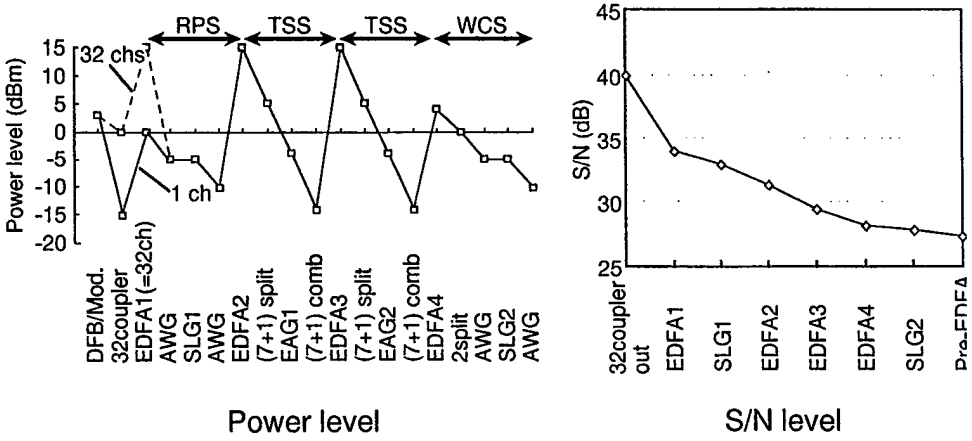
320-Gbit/s system configuration



320-Gbit/s system specifications

- architecture: broadcast-and-select type with WDM output buffers
- wavelength: 1500 nm band
- optical frequency channel span: 100 GHz (0.8 nm)
- number of inputs / outputs: 32
- highway speed: 10 Gbit/s (12 Gbit/s internal speed)
- buffer size / outputs: 12 optical buffers (M=12)+64 electrical buffers
- packet format: 64 bytes (including 4-byte guard band)
- transmission code: IM / Manchester

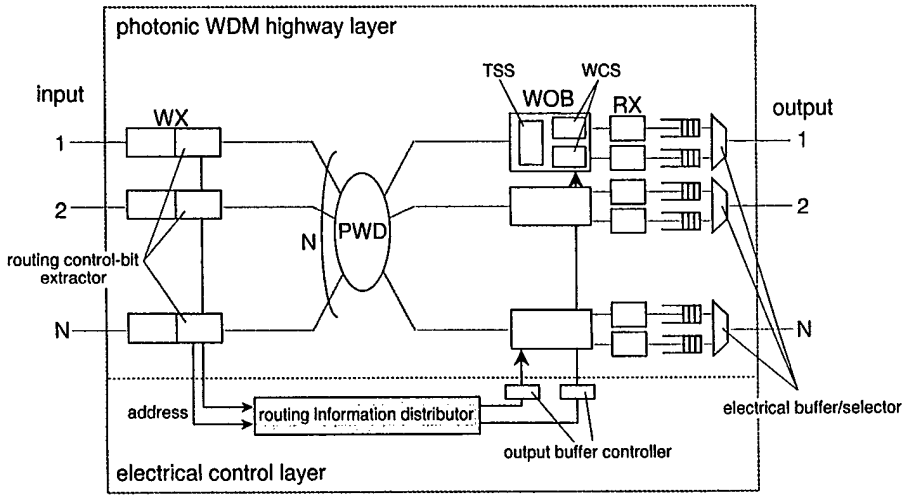
320-Gbit/s system design



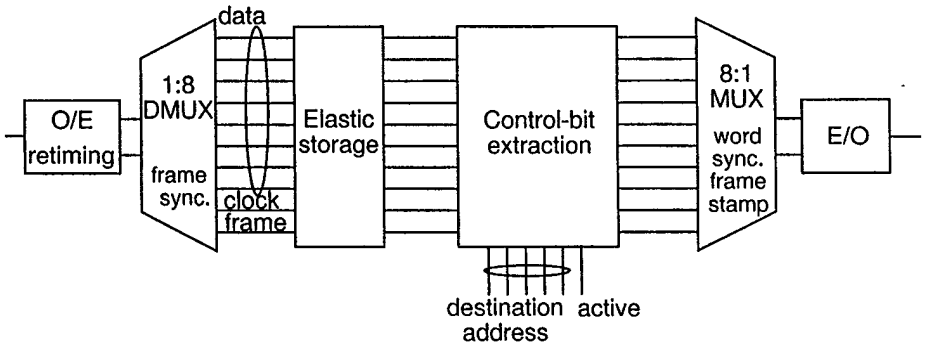
Power level

S/N level

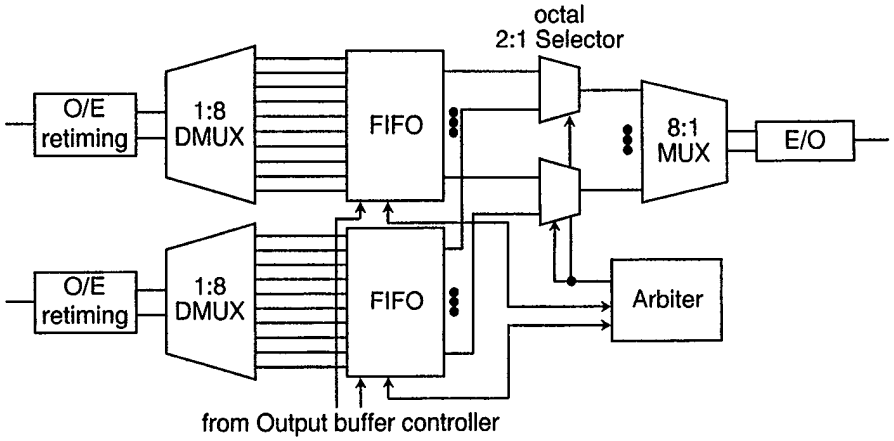
Role of electronics in photonic packet switch



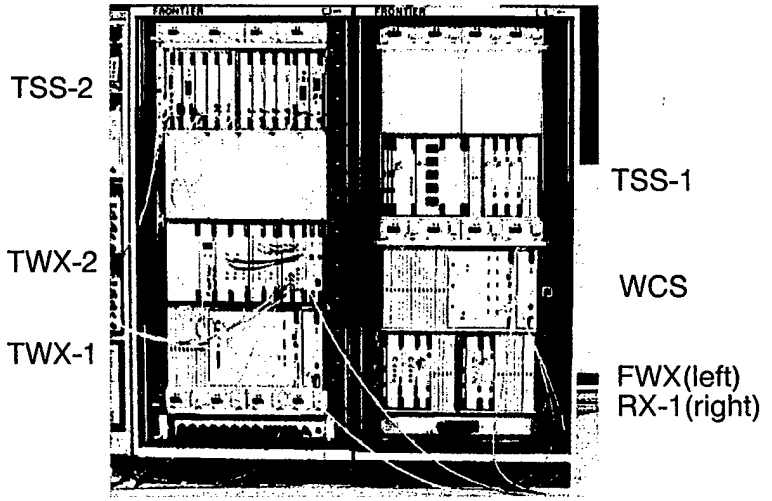
Block diagram of control-bit extractor



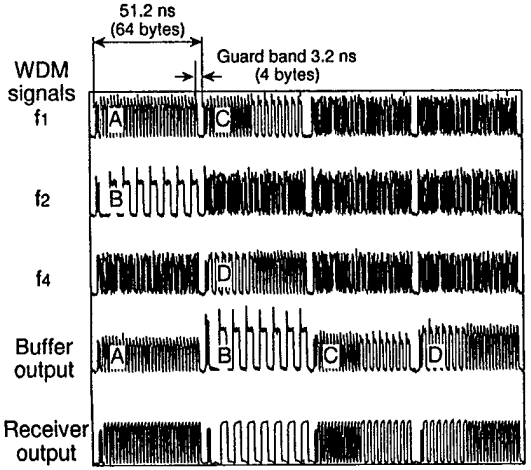
Block diagram of electrical buffer and selector



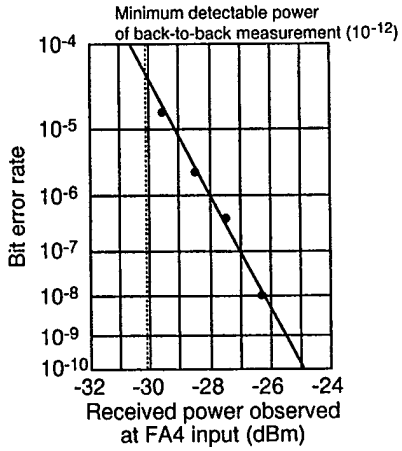
Rack-mounted prototype system



Experiments using broadcast-and-select switch

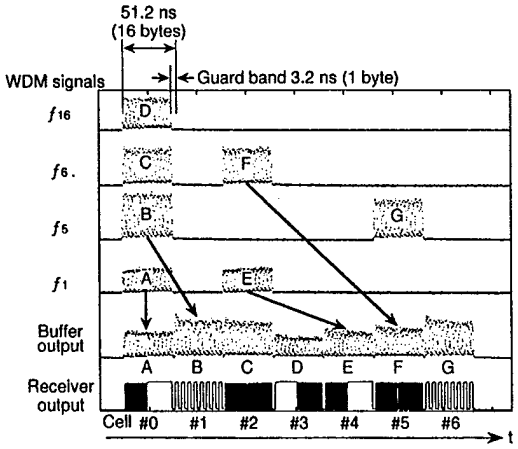


Operational example (10 Gbit/s)

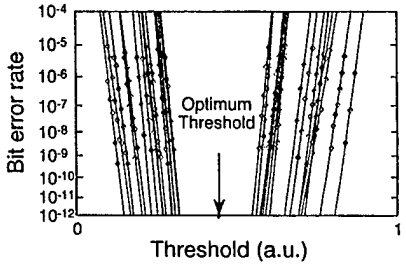


Bit-error-rate

Experiments using wavelength-routing switch



Operational example (2.5 Gbit/s)



Bit error rate vs. threshold

Conclusion

Large-scale photonic packet switches
- simple star architecture with modular structure
- combination of broadband WDM techniques and electrical control circuits

Key technologies needed for sub-Tbit/s switch
- hybrid-integrated 32-ch wavelength channel selector
- 10-Gbit/s burst-packet receiver
- level and S/N design

Rack-mounted photonic packet switch prototypes
- 10 Gbit/s x 4 broadcast-and-select type
- 2.5 Gbit/s x 16 wavelength-routing type

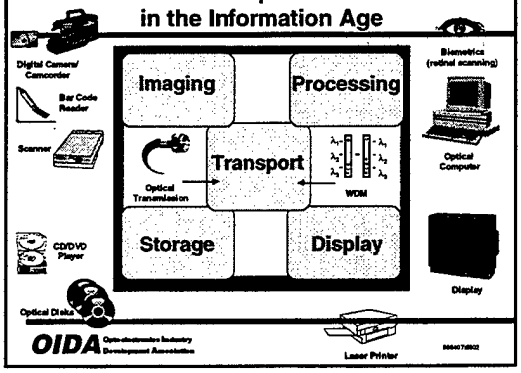
7:00pm - 7:30pm
Tues, 11 May - 2.8

State of the U.S. Optoelectronics Industry

A.A. Bergh
OIDA

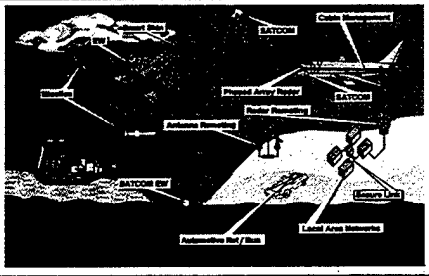
OIDA

The Role of Optoelectronics in the Information Age



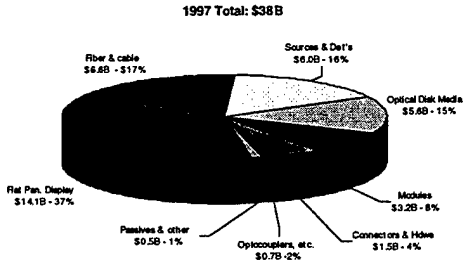
OIDA

Lightwave Technology at Hughes Aircraft



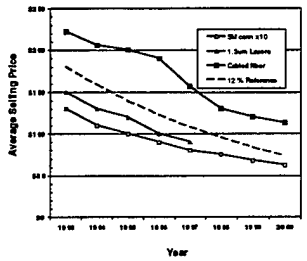
OIDA

Worldwide Optoelectronic Component Production



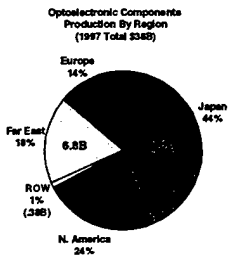
OIDA

Historical Price Behavior for Selected Optoelectronic Products



OIDA

Status of Optoelectronics in Japan Compared to the United States

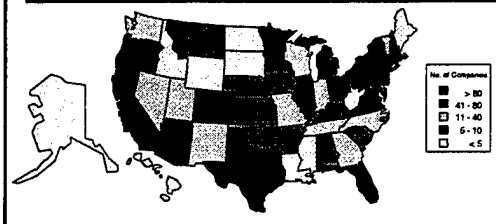


Technologies	R&D Leader*
High Capacity Networks	
Custom Optoelectronics	
Storage	
Sensors	
Emerging Device Technology	
Waveguide Devices	
Display	

* Source: JTEC Panel Report on Optoelectronics in Japan and the United States, February 1996

OIDA

Distribution of OE Companies by State

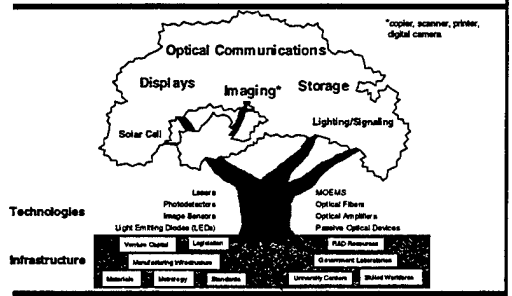


Optoelectronics equipment represents a $20B U.S. industry
- Distributed across 30 states
- With over 150,000 direct employees

Source: Photonics Spectra Buyer's Guide, 1997

OIDA Optoelectronics Industry Development Association

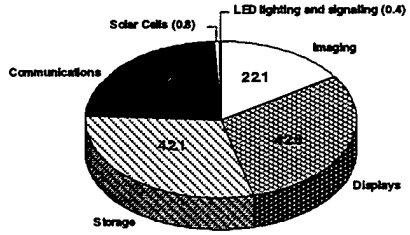
Optoelectronic Industries



OIDA Optoelectronics Industry Development Association

World OE Production ($B) 1998

(Total: $140.6 B, North America Estimate: $42.8 B)



OIDA Optoelectronics Industry Development Association

OIDA Programs

New Business Development
- OE Market Trends
- OE Future Vision
- OE Application Trials

Infrastructure Improvement
- Manufacturing Infrastructure
- Industrial and Measurement Standards

Technology Advancement
- Technology Roadmaps
- R&D Resources
  e.g. University Centers

Unified Voice of OE Industry
- Forum for OE Users and Suppliers
- Advise the U.S. Government
- Interact with Industry Associations Worldwide
- Serve as the U.S. Broker for JOP

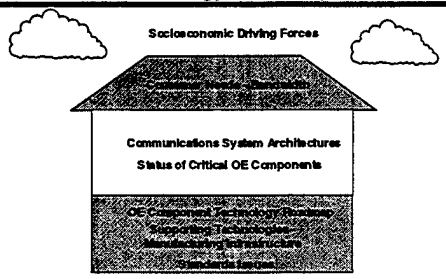
OIDA Optoelectronics Industry Development Association

OIDA Workshops

1998	
Metrology for OE	Feb 98
Optical Communications Roadmap	May 98
Technology Roadmap for Image Sensors	Jun 98
Annual Forum	Oct 98
1999	
Broadband Communications & Switching Components Technology	Apr 99
Advanced Imaging - "Electronic Eye"	Jun 99
International Standards	Aug 99
Annual Forum	Oct 99

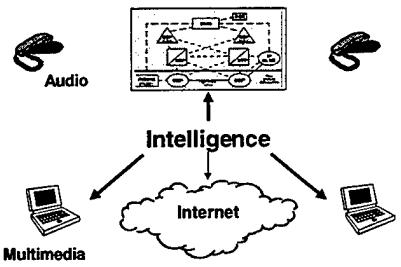
OIDA Optoelectronics Industry Development Association

Layers of Communication Technology Roadmap



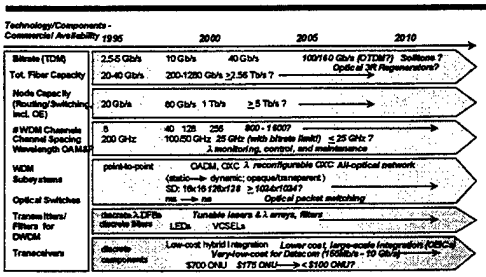
OIDA Optoelectronics Industry Development Association

Paradigm Shift in Communication



OIDA

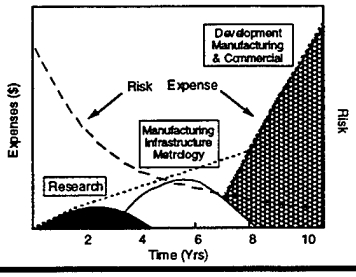
OIDA Communications Roadmap (2/3)



Slanted Font: Major industry efforts are required for commercialization

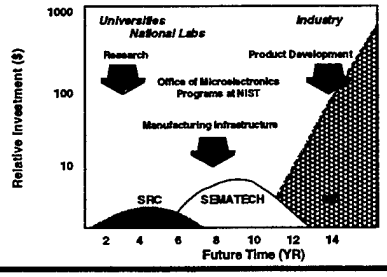
OIDA

Three Phases of the Innovation Process



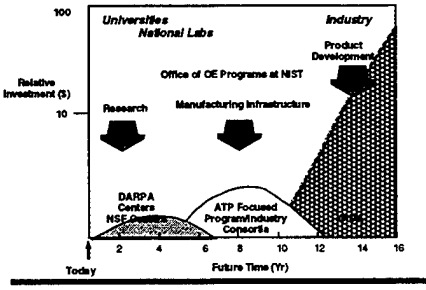
OIDA

Semiconductor Industry



OIDA

Optoelectronic Industry



OIDA

OE DARPA Centers History

Time Period	No. of Centers	Funding
1990 - 1993	3	15 million
1994 - 1997	4	25 Million
1997 - 2000	2	12 Million
2000 - 2003	4?	25 Million?

OIDA

Number of Students Placed in Industry from OE DARPA Centers 1990-1996

NCIPT	71
OTC	48
OMC	60
COST	47
	226

OIDA Opto-electronics Industry Development Association

Market Strengths of Japan and the US

- North American Strengths
  - Communications
  - Industrial uses
  - Military applications
- Japanese Strengths
  - Consumer applications
  - Displays
  - Storage
- Emerging Opportunities!
  - Imaging
  - New Information Age applications
  - Medical technology
  - Transportation

OIDA Opto-electronics Industry Development Association

Board & Back-plane Level Optical Interconnections Using Integrated Thin-cladding Polymer Fibers

7:30pm - 8:00pm
Tues, 11 May - 2.9

Yao Li

NEC Research Institute,
4 Independence Way, Princeton, NJ 08540.
e-mail: yao@research.nj.nec.com

Other Contributors & Collaborators

Jun Ai	NECI,	USA
Jan Popelek	NECI,	USA
K. Kasahara	NEC CRL,	Japan
Y. Takiguchi	Hamamatsu, KK	Japan

@ IEEE-Santa Fe, 05/11/99

Talk Outline

* **Introductions,**
* **POF's as Short-distance Optical Channels,**
* **POF's for Intra-computer Interconnections,**
  Project I: multi-Gb/s on-board clock distributions,
  Project II: 2D parallel optical circuits on PCB.
* **Some experiments,**
* **Summary and Conclusions.**

Introduction

* **Bandwidth bottleneck at PCB level,**
  (>500 MHz on-chip & <200 MHz off-chip)
* **Problems of conventional waveguides,**
  (high cost for glass waveguides, large loss for polymer ones)
* **Space parallelism can be spplied by VCSEL's,**
  (1D & 2D arrays with low fabrication cost, low threshold current)
* **POF offers low-cost, high-rigidity, low-loss,**
  (1/4 of glass fiber cost, breakage-free, < 3 dB/m)
* **Low-cost Polymer fiber-image-guides (PFIG's) are also becoming commecially available**

Basics of Polymer Optical Fibers

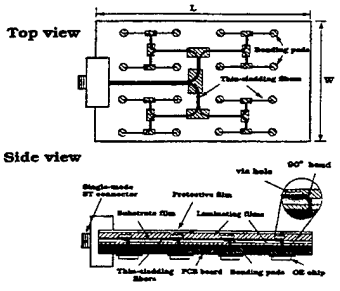
* 1st. POF in 1970's but progress was slow,
* Main applications now in display & lighting,
* Low material & production cost,
  (1/4 of cost of silica fibers)
* High attenuation, (PMMA:12 db/100 m @ 650 nm),
* Thin-cladding (90% core) & Multimodes,
* Low operating temperature (-20 to 80 °C),
* High flexibility and rigidity against breakage.

Two Interconnect Projects at NECI for Board-level POF Circuits

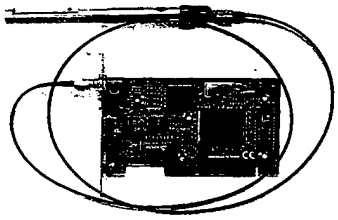
* Multi-Gb/s Optical Clock Distribution Circuit,
  (10 Gb/s, 128 port, connectorized integrated optics)

* 2D Parallel Optical Circuits for VCSEL Arrays,
  (both point-to-point and multi point capability)

Board-level Optical Clock Distribution Using End-tapered Fiber Bundles



A POF Circuit-embedded PCB



Power-loss distribution eye diagrams

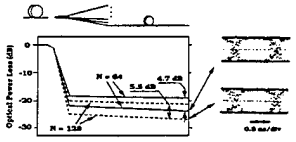
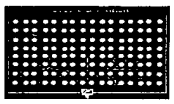
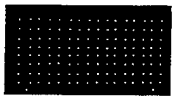


Photo of the fabricated board



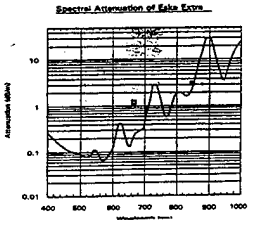
Same board when in operation



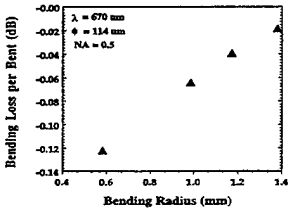
Main Characteristics of Mitsubishi Thin-cladding PMMA Fibers

Fiber Specifications

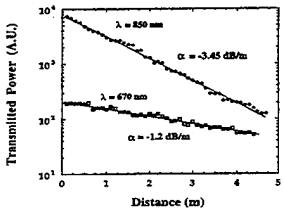
Trademark	**Eska**
Product Code	**CK**
Core Index	**1.492**
Clad Index	**1.402**
NA	**0.51**
Core Diameter	**107 um**
Fiber Diameter	**114 um**
Application	**sign, illumination**



Bending Loss Measurement



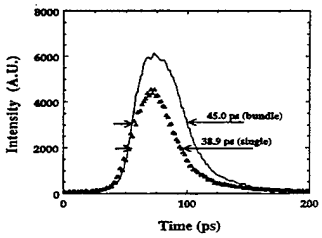
Attenuation Measurement



Temporal & Frequency Domain Measurements

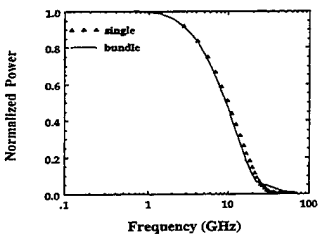
30 fs Ti: Sapphire laser at λ = 850 nm,
10 ps Synchroscan Streak Camera,
1.4 ps Maximum Readout Accuracy

Pulse width & Skew



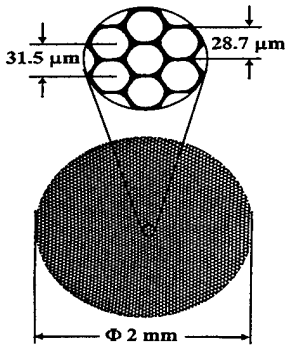
$$\text{skew} = \sqrt{45^2 - 39^2} = 22.5 \text{ ps}$$

Bandwidth

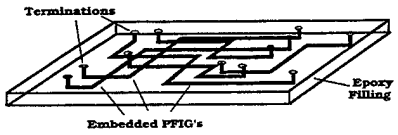


Embedded Optical Circuit Board for VCSEL Arrays

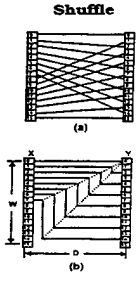
Fiber Image Guide
(PMMA, 3,500 pixel/ϕ=2mm)



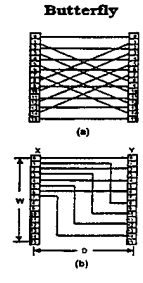
3D Circuit Layout



Shuffle



Butterfly

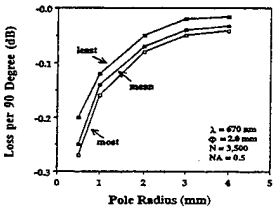


Main Characteristics of a Prototype PFIG

Product Specifications

Company	Asahi Chemicals
No. of pixels	3,500
Materials	PMMA
OD range	Φ/1.0 to Φ/3.0 mm
NA	0.50
Fill factor	> 0.85
Applications	disposable endoscope

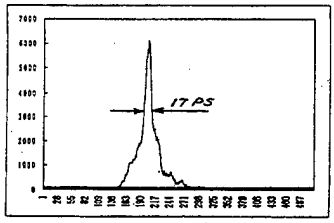
Bending Loss Measurement



Dispersion Measurement

Light Source: Ti-Sapphire laser with λ = 850 nm, Δ = 30 fs.

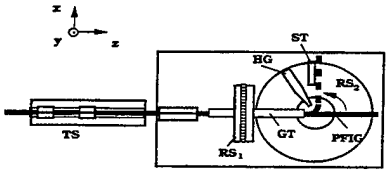
Detector: Hamamatsu Streak camera with τ < 10 ps, sensitivity : < 1 nW



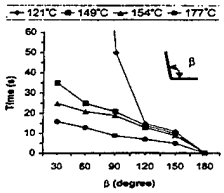
one meter in length

Fabrication of Circuit Preforms Using Thermo Bending

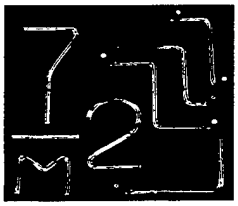
Setup for Thermo Treatment



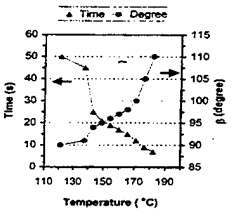
Bending Speed



Preformed Elements



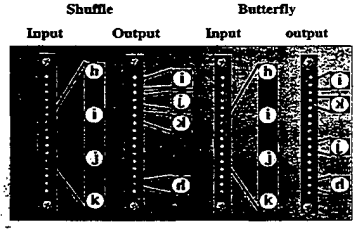
Bending Response at 90°



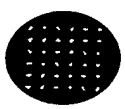
Prototype Board & VCSEL Transmissions

Prototype Board

Material: G10
Dimension: 20 x 12 x 0.6 cm³

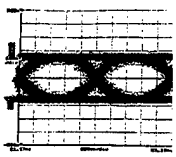


VCSEL Image



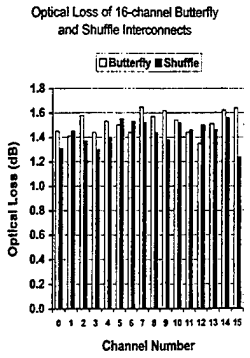
(a)

Eye Diagram

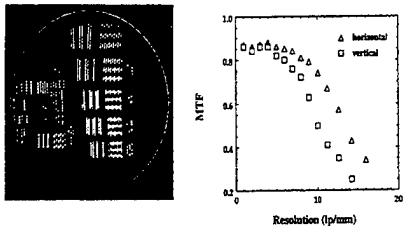


(b)

Power & Resolution Performance Measures



Imaging Result **Transfer Function**



(a) (b)

Summary and Conclusions

* **POF suits better for short-distance applications,**
* **POF offers better packaging capabilities,**
* **Multi Gb/s bandwidth is sustainable,**
* **Free-space optics can add value to POF circuits,**
* **Packaging capability determines practicality.**

NOTES

Wednesday,
12 May 1999

NOTES