Optical Interconnects: Out of the Box Forever?

Dawei Huang, *Member, IEEE*, Theresa Sze, *Member, IEEE*, Anders Landin, Rick Lytel, *Member, IEEE*, and Howard L. Davidson, *Senior Member, IEEE*

Invited Paper

***Abstract*—Based on a variety of optimization criteria, recent research has suggested that optical interconnects are a viable alternative to electrical interconnects for board-to-board, chip-to-chip, and on-chip applications. However, the design of modern high-performance computing systems must account for a variety of performance scaling factors that are not included in these analyses.**

We will give an overview of the performance scaling that has driven current computer design, with a focus on architectural design and the effects of these designs on interconnect implementation. We then discuss the potential of optics at each of these interconnect levels, in the context of extant electrical technology.

***Index Terms*—Cache memories, data busses, high-speed electronics, optical interconnections.**

I. Introduction

OPTICAL INTERCONNECTS outperform electrical interconnects for long-distance applications. Even for distances as short as 300 meters (m) at bandwidths over 1 Gb/s, fiber is now the default interconnect choice. Recent research [1]–[7] has also suggested that optics has clear advantages even at very short distances, yet optical interconnects are not common in computers.

The reason optical interconnects are not widely implemented in modern computing systems is due to computer architecture design and how interconnects are balanced within those constraints. We will discuss these issues in this paper and explain that despite the fact that optical interconnects have clear advantages in time of flight and bandwidth density, these are not the only design constraints that must be considered. As a consequence, the overall optimization of modern computers has generally precluded the use of highly integrated optical interconnects. We will discuss the historical scaling of logic, memory, and wires, analyze access time and density of dynamic random access memory (DRAM), walk through the steps and timing relations for memory access, and estimate the capabilities of electrical interconnects at several levels. With this background, we shall characterize what would be required for optical interconnects to displace wires at the backplane, board, and chip level.

Manuscript received November 22, 2002; revised February 7, 2003.

D. Huang, T. Sze, and R. Lytel are with Sun Microsystems, San Diego, CA 92121 USA.

A. Landin and H. L. Davidson are with Sun Microsystems, Menlo Park, CA 94025 USA (e-mail: hld@sun.com).

Digital Object Identifier 10.1109/JSTQE.2003.812506

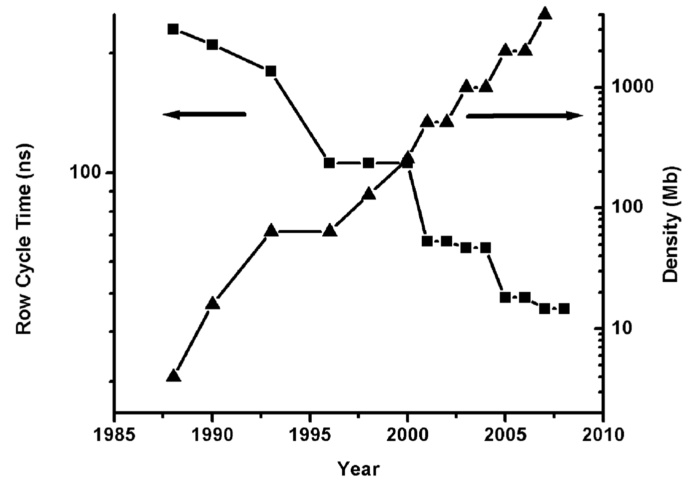


Fig. 1. Access time (nanoseconds) and density (megabit/chip) for commodity DRAM chips by year of introduction. (Note: T_Row_Cycle is the row cycle time for DRAM, which determines the memory access time and latency).

II. A Short Tour of Computer Design

A. Scaling Overview

Power laws compactly express important properties of many natural and engineered systems [8]–[10]. We assert that there are three key empirical-power law scaling relationships for computer systems:

1) You can have twice as many transistors for your next design (Moore's law).
2) A processor should have at least enough memory that it takes one second to touch every word (Amdahl's law).
3) There are many more short wires than long wires (Rent's rule).

These laws guide the design of successful computers.

Logic, memory, and wires respond very differently to scaling down minimum feature size. Approximately, logic speed increases exponentially, memory density increases exponentially, and wires are sensitive to ratios of physical dimensions. The clock cycle for microprocessors has decreased from 1 μs for early 8-bit chips, to 333 ps for recent designs. DRAM density has increased from 1 kb/chip to 1 Gb/chip, while raw DRAM access time has decreased from about 1.2 μs to 50 ns. Fig. 1 illustrates the scaling of DRAM access time and chip density over time. When normalized to processor clock cycles, access time to a main memory built with DRAM chips is becoming exponentially worse with lithographic scaling, and the number of

1077-260X/03$17.00 © 2003 IEEE

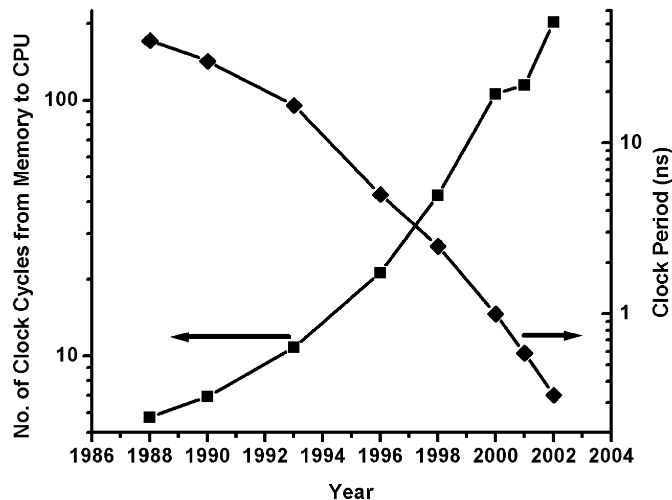


Fig. 2. Processor clock period (nanoseconds) and main memory latency (number of clock cycles from memory to CPU) of a typical computer plotted by year of introduction.

clock cycles required to fetch from main memory has increased from a few clock cycles to between 20 and 400. Fig. 2 illustrates DRAM access time as a function of processor clocks.

B. The Reason DRAM Is Slow

DRAM is slow because of the physics of the DRAM storage mechanism, the commodity nature of DRAM technology, and the complexity of the memory hierarchy. The potential advantages that optical interconnects may have with respect to bandwidth density and time of flight do not address these fundamental issues.

The basis of DRAM is the single 1-T cell, an analog sample and hold, where a bit is represented by charge on a capacitor [11], [12]. The size of this charge is compared with a reference charge by a clocked, latched, differential sense amplifier. The storage capacitor is either buried under the access transistor, or built into the interconnect stack over the transistor. The charge is gated from the storage capacitor onto a bit line by a control signal. Many storage cells and a reference cell share each bit line. The capacitor, bit line, and input capacitance of the sense amplifier form a capacitive divider. The maximum number of bits per bit line is limited by the ratio between the capacitance of the storage node and the capacitance of the bit line and input stage of the sense amplifier. This ratio, and the voltage stored for a 1, set the amplitude of the signal. When the charge in the storage capacitor is gated onto the bit line, the charge spreads by resistance–capacitance (*RC*)-limited diffusion, which results in access time depending quadratically on the number of bits per bit line.

The economics of integrated circuit production encourage designers to place as many bits on a line as possible, thereby increasing the ratio of memory array to support circuit area, and increasing the number of bits of memory that can be placed on a chip of a given size.

The personal computer (PC) and the Internet now control the economics of DRAM manufacturing. More memory is sold for use in PCs than any other single application. This vast market has produced tremendous competitive pressure on the price and nature of DRAM. Parts produced in the highest volume generally have the lowest unit cost. Because the PC provides a universally accepted target specification for memory chips, the lowest cost per bit for a system can be obtained by designing with such commodity parts.

Memory manufacturers periodically test the waters by introducing a part with half the number of bits per bit line as their mainstream component. These new devices have generally failed in the marketplace. Furthermore, there is little motivation for large increases of DRAM bandwidth when the devices themselves are slow, and latency ultimately limits performance.

In summary, DRAM latency is a significant design constraint, limited by the device design optimization. Optical interconnects cannot solve this DRAM access problem.

C. Simple Machines

Much of the analysis of the potential of optical interconnects has focused on tradeoffs between bandwidth, power, circuit area, and signal integrity. Let us consider some of the design tradeoffs for a single-processor machine.

A large fraction of the work done by a computer merely produces memory addresses and moves bits back and forth. For example, in a single-accumulator computer, it takes four memory accesses to perform a single operation on two inputs: one to fetch the instruction, two to fetch the operands, and one to write back the result.

Fortunately, locality in programs can be exploited to maximize the return for this effort. A small section of a program, perhaps an inner loop and some operands, might recirculate in the processor's registers for many clock cycles. Instructions that were executed frequently do not need to be repeatedly fetched. Intermediate results of sequential sums do not have to be written to memory, only to be immediately returned to the processor. The number of instruction fetches and writes required to execute a particular piece of code can be greatly reduced by taking advantage of such locality. This idea can be applied more generally.

From the beginning, computer engineers have faced the tradeoff between memory speed, memory density, and memory cost. Faster memory tends to be less dense and more expensive. One response has been to place a small fast memory, called a cache, in the path between the processor and main memory. Contemporary designs employ up to three levels of cache, each larger and slower than the previous one. A cache retains copies of recently fetched memory locations. If the processor should request a value already present in the cache, the local copy is quickly delivered and latency is reduced. If the processor should request a value that is not in the cache (cache miss), a memory access is required which results in increased latency. For well-behaved programs, these methods result in a notion of "distance from the processor." The control complexity, access time, and cost required increase rapidly with size and number of levels of cache memory.

Notwithstanding, this hierarchy of memory system successfully exploits the locality of the program in single-processor machines. With multiprocessors, things are different.

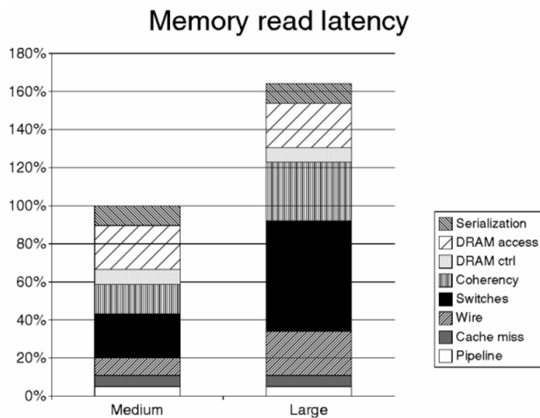


Fig. 3. Time needed for the steps of a main memory access in two different size multiprocessor servers [13].

D. Architectural Issues in Multiprocessor Systems

The 1990s saw a dramatic increase in the popularity and commercial success of the symmetric multiprocessor, or multiprocessor server, notwithstanding that in the mid 1980s, most of the academic architecture community asserted that this architecture was doomed by scalability limitations of the shared electrical bus—the traditional backbone of both single-processor and multiprocessor systems. What has allowed this success is an evolution of the interconnect away from the shared bus to more elaborate topologies, point-to-point links and switches that allow far more efficient signaling. These point-to-point links maintain the required logical functionality and provide increased bandwidth. Memory operations still logically appear to happen in the same sequential, atomic fashion like a shared bus connecting all processors to a single memory.

Fig. 3 shows how the memory-read latency is spent in a medium size and a large size contemporary commercial multiprocessor. DRAM access and the wire latencies are only a small fraction of the total read latency, while transfers through switches and actions to maintain access ordering and cache coherence, are a significant fraction. In a large server, these two categories represent over half the access latency. There are several reasons for significant switching delays. They are caused both by a desire to locate switching elements close to the devices that source the data, and from wanting to use common part in both large and small systems.

Fig. 3 also illustrates that switching delays are related to the size of the system. Using a standard chip package, there are a limited number of data ports available per switch chip. Therefore, to increase the size of a system (i.e., number of processors), switches are inserted in the path to accommodate additional point-to-point connections. A drawback is that significant latency is spent in additional switch stages.

A potential advantage of optical interconnects at the backplane level is that if higher density optical input/output (I/O) can be provided to a switch chip, a reduction in the number of switch stages would be possible. This could significantly reduce latency in large servers. The switches would still be electrical, and their bandwidth and port count would limit possible topologies.

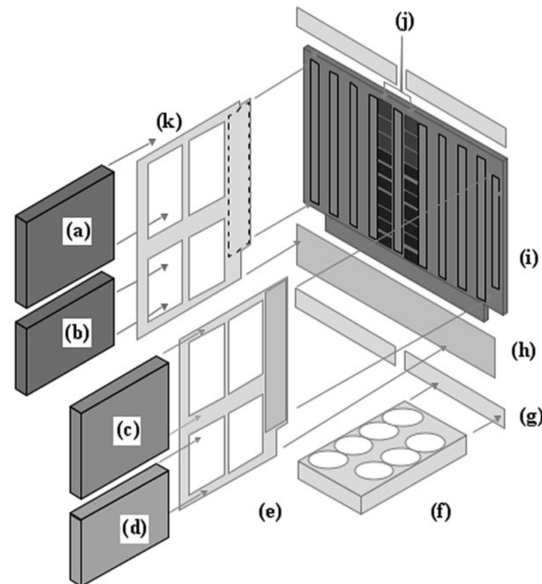


Fig. 4. Mechanical arrangement of the front half of the cabinet of a Sun Microsystems SF 12 K–15 K. (a) Uniboard (also shown in Fig. 4) the main CPU board. Each system can contain up to 18 of these boards. (b) I/O adapter board or MaxCPU board, a two-CPU board alternative to I/O. Each system can have up to 18 of these boards. (c) and (d) System controller board and the system controller peripheral board. (e) and (k) Frames and expander boards that allow boards designed for smaller cabinets to be used in the SF15K chassis. (f), (g), (h), (i), and (j) Fan trays, fan centerplane, power centerplane, logic centerplane, and centerplane ASICs.

Cache coherence protocol is another significant factor for memory latency. When large caches are placed close to each processor, a significant fraction of memory accesses can be satisfied with data from other processors' caches, rather than from main memory. To reduce latency for these accesses, it is helpful to broadcast requests directly to all processors, as well as to main memory. For example, if several processors are cooperating on a single program, it is likely that several processors will need to access the same word in memory. If more than one processor has recently accessed a word, there will be copies in more than one cache. One processor may modify the word while a different processor still has the previous value in cache. Preventing incorrect program execution in such cases requires an efficient mechanism for invalidating the word across the whole computer. This mechanism is called broadcast snooping, and its complexity grows nonlinearly with the number of processors. With the improvements in the interconnect topology mentioned above, the limitation on coherence through broadcast snooping is the bandwidth with which caches can be accessed and interrogated for snoop results, and not in the bandwidth of the interconnect medium, which limits the potential of optics at this interconnect level.

An example of a modern multiprocessor server [13] is shown in Fig. 4. This particular system is the SunFire 12 K–15 K family and can accommodate up to 96 processors. Fig. 5 shows a detailed view of the central processing unit (CPU) board, which contains four processors and 32 GB of DRAM. The switch chips used to maintain memory coherency are located on the CPU boards as well as on the centerplane.

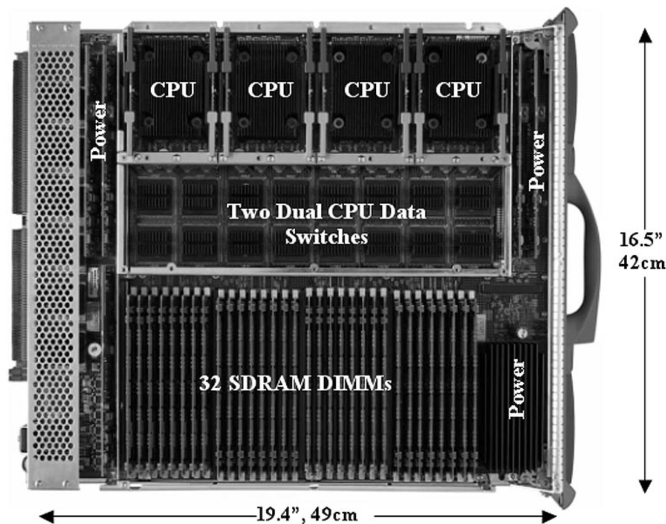


Fig. 5. Sun CPU board, or Uniboard, consists of four CPUs with associated E-cache DIMMS, two dual CPU data switches, 32 SDRAM DIMMS for a total of 32 GB of memory, and address and data switch ASICs for off-board access. This board corresponds to Fig. 4(a).

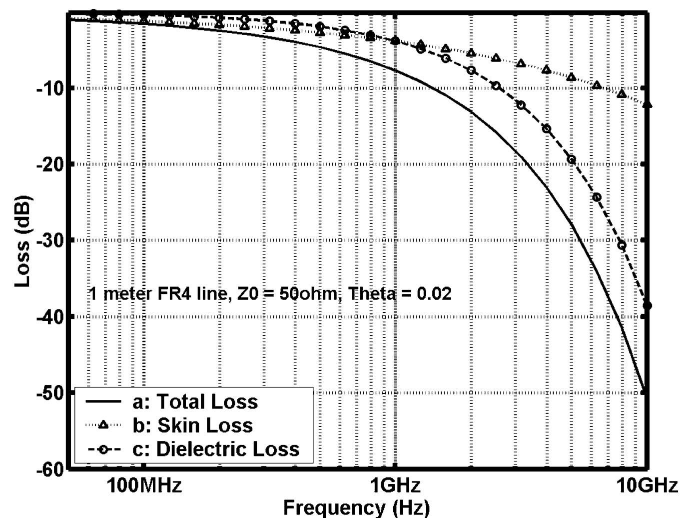


Fig. 6. Contributions to loss versus frequency for copper traces on FR-4 circuit board [16].

III. Electrical and Optical Interconnects

Several justifications can be put forward for the use of optical interconnects in multiprocessor systems. These fall into a number of broad categories.

1) Optics will speed up memory access.
2) It will become impossible to route enough bandwidth through a backplane, circuit board, or module.
3) Unfavorable wire scaling will make clock distribution and signal wiring increasingly difficult.

In this section, we consider these applications at the backplane, board, and chip level.

A. Circuit Board Wiring Density

Two factors control the wiring density on a circuit board, the pitch of the signal lines and via construction [14]. Standard printed circuit boards are constructed from etched copper foil traces on FR-4 laminate with plated through holes. These boards can be manufactured with trace width and spaces between traces of 100 μm.

In order to route signals between layers on a board, the traditional via process requires a hole to be drilled completely through the board. In areas where it is necessary to bring many closely spaced connections to the surface of the printed circuit board (PCB) to connect to a chip, for example under a processor or dual in-line memory module (DIMM) socket, the via holes may block half the available space for wiring on every layer of the board. Additional wiring layers may be required to make routing possible.

A generic PC is built on a six-layer board, with two layers dedicated to power distribution. A server board, with its greater complexity, may need a twenty-layer board.

Alternative methods for forming vias do not block all the layers at each location. Laminate boards may incorporate buried or blind vias. Boards so produced are generally more expensive than conventional boards, because of lower yield. However, a crossover occurs when decreased board-layer count offsets the cost of lost yield. Additionally, as manufacturers gain experience with these new processes the crossover layer count declines.

One alternative to laminated circuit boards is multilayer ceramic circuit boards, which are available with alumina and low temperature glass–ceramic bodies. Traces are formed by screen printing metal inks on sheets of unfired ceramic. Line and space widths are 50 μm. This process inherently supports buried vias. These boards are used for integrated circuits packages, and for multichip modules. This technology has the additional benefit that the copper traces have a lower resistance per unit length than the alumina of standard FR-4.

Ceramic boards are available with up to 100 layers and in sizes up to 15 cm on a side. All the integrated circuits required to build a 32-way parallel processor, except for the main memory chips and I/O adapters, can be mounted on a single board [15].

B. Electrical Wiring Limitations

To maintain signal integrity, the signal speed is limited by circuit board material losses, and wiring density on a circuit board is limited by noise control.

The ideal operational region for circuit boards is within the low dielectric loss region. As shown in Fig. 6, dielectric loss in FR-4 laminate rises rapidly above 1 GHz, where it rapidly becomes larger than skin effect losses. Frequency-dependent losses reduce signal-to-noise ratio and introduce timing errors. At high data rates, where bit time is very small, even a miniscule timing error significantly reduces the timing margin. Since dielectric loss is a fundamental material property caused by dipole rotations in the polymer, it can only be reduced by changing the material not by changing the shape of the conductor. To improve bandwidth density, both lower loss materials and signal processing techniques can be used.

Low-loss laminate materials like Rogers 4000 have been developed for printed circuit boards. The raw material is five times more expensive than FR-4, and more difficult to process into high layer count boards. However, it does provide higher bandwidth density. Fig. 7 compares the bandwidth density on circuit

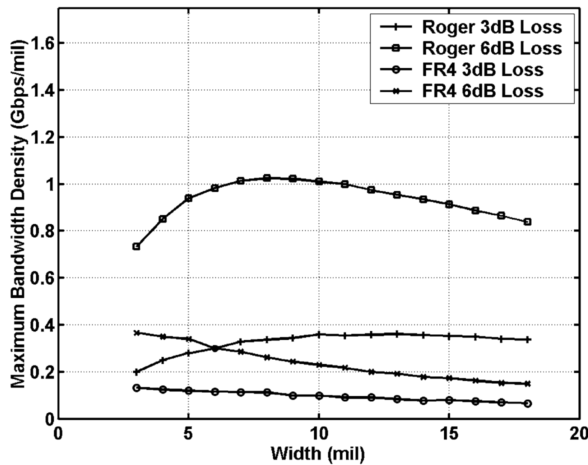


Fig. 7. Bandwidth density as a function of trace width for different circuit board laminates and permissible insertion losses [16].

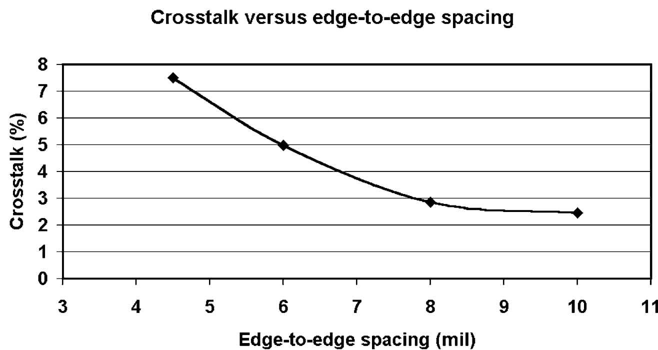


Fig. 8. Crosstalk versus trace spacing for typical FR-4 multilayer circuit board construction [16].

boards for two materials and four acceptable values of attenuation over a range of trace widths. Higher attenuation values require more complex driver and receiver circuits. With Rogers 4000 more than 1-Gb/s/mil bandwidth density can be obtained with modest transceiver technology while only 400 Mb/s/mil is available with FR-4.

Signal processing techniques can be used to increase bandwidth on FR-4 boards. By using driver and receiver cells that employ preemphasis, equalization techniques, and multilevel signaling and coding, deterministic jitter can be reduced, and timing and voltage margins can be increased. Although, this does require more transistors per I/O, the continued drop in transistor cost over time makes these techniques increasingly attractive. These methods will extend the useful lifetime of FR-4 PCB technology.

Crosstalk is a major source of noise that can limit wiring density. Generally it should be limited to 5% of the signal swing, in order to maintain signal integrity. Fig. 8 shows near-end crosstalk as a function of edge-to-edge line spacing for a particular FR-4 multilayer board. In this example, the edge-to-edge spacing between traces should be greater than 150 μm for long runs. For a 100-μm trace width, the center-to-center trace pitch is 250 μm, equal to the center-to-center fiber spacing in a fiber ribbon cable. In this case, fibers have no advantage in wiring density at the same data rate.

C. Bandwidth Density to Memory

While the use of optical interconnects will not help with memory access latency significantly, it may help to relieve wiring congestion to memory.

Modern large computers have more than four DIMM slots per processor and a single cabinet may house more than 100 processors. The board in Fig. 5 is from a computer currently in production. It has 32 DIMM sockets, which produces a difficult routing situation under the sockets. To route each memory connector contact on the surface of the printed circuit board to the memory controller, it must be spread out on multiple layers. The vias that connect the surface of the PCB to lower layers blocks wiring channels on some, or all, of the layers of the board. A similar problem occurs under the application-specified integrated circuits (ASICs), which are visible in the photograph.

It is possible to imagine using optics instead of a circuit board to connect a memory chip to a controller. A free-space optical system would not have wiring congestion problems. It would require adding optical I/O to each DIMM, or to each memory chip. Therefore, to be competitive at the DIMM level, optical I/O would have to cost about the same as the circuit board and connector that are replaced. The motivation would be to relieve wiring congestion, not to reduce time of flight.

D. Backplanes

Backplanes are a possible insertion point for optical interconnects. We will review electrical backplanes, and discuss two possible implementations of optical backplanes.

1) Electrical Backplanes: The circuit boards used in current server designs are up to 0.6 m wide and may contain as many as 40 layers of wiring. Such a board might have 48 connectors, each with 700 signal pins. The insertion force required to seat one of the connectors is almost one hundred pounds.[1]

Currently, there are two broad classes of backplanes. Conventional busses use a traditional multidrop bus topology. The other system is connected point-to-point, and uses electronic switching to manage traffic. In some cases, both topologies are used in the same machine to allow independent optimization of different paths.

Conventional bus topologies can operate at 400 MHz, or a bit more, for 25-cm distances. If extraordinary effort is taken to minimize the effects of connectors and stubs, large backplanes can operate at 1.5 Gb/s. Depending on implementation details it is possible to transport from 5 to 20 GB/s on a bus topology backplane [17].

Point-to-point connections are displacing busses in new designs. With this topology, using off the shelf components, it is possible to realize 3-Gb/s data rates over a 75-cm length of differential trace on FR-4 circuit board [18], [19]. By using this technology, one could obtain a total of 25 TB/s on a backplane. The board would contain 8.4×10^3 nets.

There are several tiers of SERDES integrated circuits available in the marketplace. The lower tier is implemented in silicon complementary metal–oxide–semiconductor (CMOS). SiGe bipolar circuits are popular for the 10-GB/s range. Both

[1]http://www.teradyne.com/prods/tcs/products/connectors/backplane/vhdm/operchar.html#mechanical

silicon and III–V compound semiconductor circuits have been operated at 40 Gb/s. CMOS continues to move up in operating frequency, displacing the more exotic technologies.

Once the SERDES circuits have been chosen, the selection of an optical or electrical medium for a backplane will depend on the relative properties of the transmission medium, availability of connectors, predicted reliability, the confidence of engineers and managers in each technology, and cost.

2) Optical Backplane Technologies and Limitations: Optical "backplanes" that are in use today are more like patch panels than backplanes. They are an assembly of transceivers connected by fiber ribbon "routed" through fiber patch cords.

Many different optical backplane constructions have been suggested and demonstrated, including glass fibers embedded in polymer sheets, polymer light guides [20], lens and mirror systems [21], [22], diffractive optical elements [23], and active pointing with microelectromechanical systems devices. These approaches have focused on demonstrating the viability of an optical platform and not on use in a computer system.

Two key constraints mentioned earlier are the size of the cabinet, which bounds the number of processors and amount of memory that fits inside while meeting acceptable thermal and mechanical requirements, and the routing density beneath high bandwidth chips such as DRAM, processors, and ASICs. Glass fibers in polymer sheets need to be laid over a PCB, limiting heat flow and interfering with chip placement. Lens and mirror systems, diffractive optical elements, and active pointing devices either eliminate heatsink area on top of the board or block wiring channels by routing the optical path through the board.

Among the demonstrated technologies, polymer or fiber lightguides embedded in the circuit board have the most potential to be implemented with current design approaches. There is continuing progress in reducing loss in these waveguides, and in improving their ability to survive the mechanical and thermal stress that they would be exposed to in manufacturing and during long-term use. Inexpensive optical launch into the waveguide, and right angle turns within the board, have not been satisfactorily demonstrated.

Another problem with optical backplanes is that they assume reliable optical transmitters and receivers. Most multichannel optical modules are based on vertical-cavity surface-emitting lasers (VCSELs) because of their relatively high reliability, low power consumption, and high modulation speeds. However, the average failures-in-time (FIT) rate (failures per 10^9 device hours) of an oxide or implant VCSEL is about 20 [24]. According to the Semiconductor Industry Association (SIA), the average FIT rate of a mature processor is about 50, whereas the average FIT rate of a high-speed backplane pin is 0.2 [25]. Clearly, if the proposal is to take each of the 500 or so high-speed pins from a bus, and replace the backplane connector pin with a VCSEL, the overall reliability of a computer would drop by a factor of 100 per connector.

One possible solution is to use redundant VCSELs to improve the reliability of the optical interconnect. We have shown that simply adding one redundant channel to a 500-VCSEL array can improve reliability by close to an order of magnitude [26]. This would require circuits that could detect the decay of a VCSEL and automatically switch it to another channel. Instead of SONET's link reestablishment requirement of 1 ms, the optical interconnect link would have to be reestablished in a few processor clock cycles. Clearly, there is a tradeoff between number of redundant channels and latency for reestablishment of service.

An advantage of optical connectors is that they can be denser than electrical ones. Multirow ribbon fiber connectors with a 0.25-mm pitch have 36 times the signal density of current backplane connectors Polymer waveguides have been demonstrated with 0.030-mm pitch.[2]

As an example, a bare board backplane that can accommodate 48 slots, 125 input and output per slot CPU board, costs about \$5000 [27]. The total cost including all the connectors, hardware, mechanical structure, other components, and assembly is about twice as much. If the board material is changed to a fluorocarbon laminate, and uses buried vias, the cost of the board itself might increase by five times. According to the SIA roadmap, the cost per pin of connectors should not be very sensitive to bandwidth. In this case, the total cost would rise to approximately \$30 000. For optics to replace this design competitively, the cost per channel must drop 20 times from current values. This is actually a very encouraging number. It was not long ago that a single channel 1-Gb/s fiber link was \$1000.

A transition to optical backplanes might be encouraged by accumulating difficulties with the electrical ones. Signal integrity, routability, total delivered bandwidth, cost, and mechanical issues all become more difficult as backplane performance increases.

3) Optical Backplane Technology Disruption: There is more motivation to implement a potentially risky and costly new technology when it provides a feature that cannot be obtained in the old one. Optical interconnects could fall in this category, but the tradeoffs are difficult to understand.

Massive fanout, and freedom from wiring congestion, are advantages claimed for optical interconnects. In exchange for the optical components part of the backplane, multiple routing chips, and some protocol overhead could be eliminated.

A well-designed cache can invalidate one entry during each processor clock cycle. When more than one processor performs a write on the same clock cycle, the invalidate operations end up serialized in each cache. In this case, optics provides performance not available from an electrical interconnect, but the other critical elements of the system such as electrical chips used for cache invalidate cannot scale accordingly. The increased bandwidth of optical interconnects are not useful because the rest of the system cannot absorb them.

An example of a unique optical solution is global clock distribution in backplane design. Each slot in the backplane requires at least one copy of the system clock. Buffers and crosstalk add jitter. For historical reasons, the system clock is usually distributed at a low frequency and used as the reference input for on-chip frequency synthesizers that generate the local high-speed clocks. This frequency multiplication introduces additional jitter and circuit complexity. An alternative method would be to provide a moderate power optical signal at the highest

[2]http://www.optical.crosslinks.com/pdf/PitchLinkDataSheet.pdf

clock frequency in the system, and distribute many copies with an optical splitter. The timing relationships of a passive optical splitter are very stable. This would introduce less jitter than the usual digital fanout tree.

E. Chip-to-Chip Connections

The processor complex in a modern machine contains a small group of chips, typically a processor, a switch, and a number of cache memory devices. The processor chip may have as many as 5000 connections. One- to two-thirds are typically for power; the remainder are high-speed signals. There are two common methods for constructing these complexes. Typically, the processor is mounted face down in a package, which transforms the 0.25-mm typical pitch of the solder ball connections on the chip to a 1-mm grid suited for connection to a circuit board. A chip 2 cm on a side may require a package 5 cm on a side. Cache memory chips and switches are mounted in similar packages nearby on the board. In other designs, several processors and all their associated switches and cache memory will be located on a single multilayer ceramic circuit board. These boards are very similar in technology to the packages used for individually mounted processors.

The choice of implementation method depends on many factors beyond performance. One is the cost difference between a single processor and a module containing several processor chips, and the associated cost for an upgrade or in case of a failure. Acceptable manufacturing yield depends on the availability of tested unpackaged integrated circuits, and robust rework technology. There have been many proposals for replacing chip-to-chip connections in this part of a computer with optical interconnects. Most have involved grafting laser diodes onto the silicon, optics to direct light to the correct destinations, and grafted or directly integrated photodetectors.

Alignment of the integrated circuits to the optical system remains an open issue. Passive alignment works for multimode fiber. Single-mode fiber devices are still manufactured with active alignment techniques. The accuracy of machine-vision guided equipment used for automatic wirebonding and mask alignment is adequate for placing transferred optoelectronic devices onto a silicon chip. The equipment used for placing chips in packages is less precise; therefore, the cost of aligning waveguides to a chip with active optoelectronic components is an open issue.

The materials used for supporting an array of chips that are optically interconnected must be selected very carefully. Coefficient of thermal expansion mismatch and residual stress from manufacturing or assembly will cause thermal and temporal drift of beam pointing. The bimetal inherent between the bulk silicon of a chip and its metal system is sufficient to change the warp of a 2-cm square chip, packaged on a polymer substrate, by 100 μm for a 100 °C temperature excursion.

An additional important consideration is the operating temperature of the optical devices. Current generation CMOS is guaranteed to meet lifetime requirements at a junction temperature of 105 °C. For deep submicrometer very-large-scale integration this may be reduced to 90 °C. At the other extreme, a computer may at thermal equilibrium in room at 0 °C when it is turned on. VCSEL performance and lifetime are very temperature sensitive, as is photodetector leakage. We expect that special design and qualification would be required for VCSELs intended to operate while directly attached to a processor.

There are two research directions that may enable directly integrating light emitting devices into CMOS. One is nanoporous silicon; the other uses rare earth element doping. One particularly interesting method for producing well-controlled nanoporous silicon is ion implantation [28]. There is a great deal of ongoing research on methods for producing quantum confinement structures in silicon, including nanowires and nanoparticles. This work, and related work in photonic bandgap devices may result in totally new devices. A very recent result, [29], [30] describes efficient light emission from rare-earth doped oxide layers on silicon. These layers were grown by a chemical-vapor-deposition process modified to introduce excess silicon in the form of dispersed 300-nm particles. These devices are excited by tunnel current through the oxide and operate at 2 V.

Although there is potential of using optoelectronic devices at the chip-to-chip level, currently cost and manufacturing issues prevent current technologies to be directly implemented at this level. Research areas that will allow direct integration of optoelectronic devices with processors will ease the integration challenges.

F. On-Chip Wiring

On-chip wiring conveniently sorts along two dimensions. One is local versus global, the other is *RC* or resistance–inductance–capacitance propagation mode [31], [32].

Local wires connect transistors to gates. Global wires connect functions across the chip. The lengths of local wires scale with the lithography resolution. Global wires scale with the edge length of the chip. The maximum edge length has remained near 2 cm for several generations, a dimension limited by the field of view of popular projection aligners.

The thickness of both the wire and dielectric layers in the interconnect stack on an integrated circuit are controlled by deposition technology. Thickness could be controlled to much less than 1 μm long before lithography reached 1 μm. Until lithography reached about 0.5 μm, almost all the wires and insulating layers in the metal stack on integrated circuits were between about 0.5 and 1 μm in thickness. Meanwhile, the *RC* time constant of wires improved with the square of lithographic scaling.

In practice, it is very difficult to control etching processes well enough to yield wires and via holes more that are twice as tall as they are wide. Local wires are now almost square in cross section, and the width and dielectric thickness track the lithography dimension. This models nicely as a square coaxial cable. Capacitance per unit length depends only on the ratio between conductor dimensions and the dielectric constant, resulting in a linear reduction in capacitance for the short wires that scale with lithography. Wires that scale with chip edge length have fixed capacitance.

Wire resistance follows the ratio of length to cross section. Both the width and thickness of local wires track the lithography dimension. Average length drops directly with scaling,

while area decreases with the square of scaling. Under these constraints, the *RC* product of minimum cross section wires increases with process shrinks.

Semiconductor process engineers responded by changing from aluminum to copper metallization, resetting the curves to less resistance per square. The next step, currently underway, is to use insulators with a lower dielectric constant than silica. For deep submicrometer processes the situation is even worse. The resistance per unit length is increasing even faster than would be expected from the reduction in cross section, because the surface of copper wires must be covered with a diffusion barrier for adequate reliability. The resistivity of the barrier material is much higher than that of copper. The diffusion barrier has a minimum useful thickness, and thus becomes a larger fraction of the cross section with each process shrink.

The evolutionary endpoint of local wires is not obvious. Electron transport in normal metals at room temperature is dominated by scattering mechanisms. Quantum mechanics predicts ballistic transport of electrons in a cold perfect lattice. Single-walled carbon nanotubes (SWNT), rather surprisingly, have shown ballistic transport over distances of several micrometers at room temperature [33]. Quantum effects cause a contact resistance near 12 $\text{k} \cdot \Omega$. They can sustain at least five orders of magnitude higher current density than small metal wires [34]. If it becomes possible to build wires with similar properties in an integrated circuit process, short wires will not limit scaling for many technology generations.

Global wires are harder. SWNT do not help tremendously beyond a few micrometers. Inserting repeaters in the line is the most common approach. The scaling properties are well understood [35], [36].

The distribution of wire lengths on chip is still following Rent's rule. This very naturally leads to placing local wires in minimum width metal in the lower layers of the stack, and using wider lines in the upper levels for global wires. Upper layer metal is usually thicker, but the constant capacitance per unit length scaling is still enforced. Thicker and wider wires do permit longer absolute distance between repeaters.

Optics has been suggested as a replacement for both local and global wires at the extremes of scaling [1]–[7]. Waveguide and free space have been proposed to use for wiring and clock distribution.

If global wires were defined to be those longer than 100 times the lithography resolution, a 100-nm process would have detector pairs with 10-μm center-to-center spacing, which might not be possible in a foreseeable future. A detector with reasonable responsivity at high speed will require a relative large absorption area, approximately 50–100 μm. Additionally, crosstalk control is likely to be very difficult for that geometry. Therefore, it is not clear what can be gained from on-chip optical wiring.

Another possible application for optical interconnects is clock distribution. The accumulated skew through the clock tree on a deep submicrometer CMOS chip can easily become larger than the clock cycle. This problem already exists at the system level, on circuit boards, and on many chips. Although several engineering workarounds are popular, a potential no-skew optical clock distribution is highly desirable because of the synchronous nature of the processor. On the other hand, clever circuit and logic techniques can provide acceptable clock synchronization across a large chip. An asynchronous design style could be extended to cover a whole processor or system. If either of these approaches continues to be adequate, there is little motivation to add optical components to distribute a clock.

For waveguides, the ratio of the index of refraction of a semiconductor material near an absorption edge to that of silica or air is much larger than the ratio between the core and cladding in a glass optical fiber. This requires semiconductor optical waveguide to have a cross section comparable to wide global wires, which suggests that the wiring density will not be improved. The only benefit will be its potential of carrying very high-frequency signal, which is limited by the electronics circuit. Also, the waveguide will have very high loss compared with glass optical fibers, but short lengths, likely on the order of a centimeter. A useful waveguide based on chip interconnect system will also require compact corner turning elements. Micromachined mirrors and photonic bandgap structures have been suggested.

An alternative method is free-space optics, which launches light vertically into an optical system that images each source onto one or more destinations. Both microlens array and bulk optics have been used in previous demonstrations. The issues with this approach are the blocking of thermal path and the tight alignment tolerance requirement for high-density connection.

Another implementation issue is, traditionally, the routing of logic signals and power on a chip bury every point on the surface under multiple layers of metal. Providing windows for photodetectors integrated directly in the silicon at the clock points appears to have overwhelming advantages in terms of stray capacitance and process complexity compared with placing the detectors on the top surface, and wiring them down. Therefore, special design rules and processes would have to be enforced to provide windows for the photodetectors.

On-chip interconnects are facing some significant hurdles. However, at the small dimensions that must be addressed, optics will have difficulty to show clear advantages over electronic routing solutions.

IV. Scaling Into the Future

Scaling laws still dominate computer design. Clocks are getting faster, memory is getting larger, and the number of very short wires is exploding. Chip size and backplane length continue to change slowly. According to the International Technology Roadmap for Semiconductors, silicon and copper scaling should continue for at least another ten years, perhaps longer. This implies that the barriers for entry opportunities of optical technologies will remain stable or increase. Optics, then, is most likely to gain a foothold in computer system design in uses where it is already ubiquitous, viz., point–point and multiplexed communications between signaling and receiving nodes at the box, board, or card level. It will greatly ease the transition if the optical technology can enable a new capability.

Given the assumption that the bandwidth demand at the backplane level scales as some fraction of the processor clock, backplane bandwidth demand continues to increase. As the bandwidth signal density on electrical backplanes increases, elec-

trical solutions become more expensive. Each connector has many pins that have to be inserted into the housing, and then the connectors must be assembled to the board. This is not hugely different from the manufacturing process for multichannel optical transceivers. However, in order for the pricing of optical backplanes to become competitive with electrical alternatives, a standard would need to be established.

The difficulty here is defining standard interfaces that are acceptable for a wide range of computer designs. It should be no more difficult for a designer to use an optical backplane than to upgrade to a new version of transceiver chip and electrical connector. This is not overly restrictive since the corollary in the electrical world is that there are only a few types of high-speed electrical transceiver chips that define the state of the art at any one time.

At the memory interface, it has been shown that the latency is due to the speed of the DRAM component itself, and not due to time of flight limitations. Therefore, there is little that optics can do to speed up access to memory.

At the lower length scales, the acceptable cost per connection drops precipitously since there are exponentially more wires at each smaller length scale. For optics to compete at any of them there must be manufacturing processes in place for both the optical devices, and the guidance system, that have the same economics of scale as integrated circuit production. Any technology that requires handling individual components in such large quantities cannot succeed.

Chip-to-chip optical communication is the easiest of the short distance applications. Each module would need only 10 000–100 000 connections. Because chips are remaining about 2 cm on a side, it is possible to contemplate a connection system that only has two types of component, and alignment tolerances compatible with passive alignment. One component would be an optical decal that could be attached to each chip, and the other is an optical block that connects the group of chips. If an optical decal is no more difficult to attach than an ordinary integrated circuit package, mass alignment of the chips to the optical block is tractable, and the whole system does not cost much more than an advanced circuit board; this technology might become common.

The number of wires on deep submicrometer chips are so large that we see no way optical interconnects can be utilized unless all the optical functions are created with the rest of the wafer during processing. We do not know if launching, guidance, and adequate crosstalk control are possible for optical signals at submicrometer pitches.

Recent work on photonic bandgap materials has renewed interest in hybrid optical-electrical interconnects at this scale, but it is our view that this approach will require years of investment for exploration and understanding, during which time silicon and wire scaling will continue to move forward. That this has happened for over two decades with optical interconnection and its relationship to conventional solutions should not be ignored.

Acknowledgment

The authors would like to thank N. Nettleton, of Sun Microsystems, for very useful discussions, and for providing detailed information about the real behavior of source synchronous interconnects in a large system. The authors would also like to thank J. Freeman, of Sun Microsystems, and G. Papen, of the University of California, San Diego, for useful discussions.

References

[1] A. Krishnamoorthy and D. Miller, "Scaling optoelectronic-VLSI circuits into the 21st century: A technology roadmap," *IEEE J. Select. Topics Quantum Electron.*, vol. 2, pp. 55–76, Apr. 1996.

[2] D. A. B. Miller and H. M. Özatkas, "Limit the bit-rate capacity of electrical interconnects from the aspect ratio of the system architecture, Special Issue on Parallel Computing with Optical Interconnects," *J. Parallel Distrib. Comput.*, vol. 41, pp. 42–52, 1997.

[3] D. A. B. Miller, "Physical reasons for optical interconnection," *Int. J. Optoelectron.*, vol. 11, pp. 155–168, 1997.

[4] G. Yayla, P. Marchand, and S. Esener, "Speed and energy analysis of digital interconnections: Comparison of on-chip, off-chip, and free-space technologies," *Appl. Opt.*, vol. 37, pp. 205–227, Jan. 1998.

[5] M. W. Haney and M. P. Christensen, "Performance scaling comparison for free-space optical and electrical interconnection approaches," *Appl. Opt.*, vol. 37, pp. 2886–2894, 1998.

[6] D. A. B. Miller, "Rationale and challenges for optical interconnects to electronic chips," *Proc. IEEE*, vol. 88, pp. 728–749, June 2000.

[7] ——, "Optical interconnects to silicon," *IEEE J. Select. Topics Quantum Electron.*, vol. 6, pp. 1312–1317, Dec. 2000.

[8] D. W. Thompson, *On Growth and Form, Complete Edition*. New York: Dover, 1992.

[9] M. Schroeder, *Fractals, Chaos, Power Laws: Minutes from an Infinite Paradise*. New York: Freeman, 1991.

[10] J. B. S. Haldane, *On Being the Right Size and Other Essays*, J. M. Smith, Ed. London, U.K.: Oxford Univ. Press, 1985.

[11] P. Betty, *Semiconductor Memories*, 2nd ed. New York: Wiley, 1996.

[12] A. Sharma, *Semiconductor Memories: Technology, Testing, and Reliability*. Piscataway, NJ: IEEE Press.

[13] *Sun Microsystem Internal Publication*, Sun Microsystems, San Diego, CA, Nov. 2000.

[14] C. A. Harper, *High Performance Printed Circuit Boards*. New York: McGraw-Hill, 2000.

[15] J. U. Knickerbocker *et al.*, "An advanced multichip module (MCM) for High-performance UNIX servers," *IBM J. Res. Develop.*, vol. 46, no. 6, pp. 779–808, Nov. 2002.

[16] *Sun Microsystem Internal Publication*, Sun Microsystems, San Diego, CA, 2002.

[17] V. A. Barber, K. Lee, and A. H. Obermaier, "A novel high speed multitap bus structure," *IEEE Trans Adv. Packag.*, vol. 24, pp. 54–59, Feb. 2001.

[18] A. Deutch, G. V. Kopacsay, C. W. Surovic, P. W. Coetus, A. P. Lanzetta, and P. W. Bond, "Characterization and performance evaluation of differential shielded cables for multi-Gb/s data-rates," *IEEE Trans. Adv. Packag.*, vol. 25, pp. 102–117, Feb. 2002.

[19] Signal Integrity for PMC-Sierra 3.125/2.488/1.5b PBPS Links. PMC-Sierra, Burnaby, BC, Canada. [Online]. Available: http://www.pmc-sierra.com

[20] R. T. Chen, L. Lei, C. Chulchae, Y. J. Liu, B. Bihari, L. Wu, S. Tang, R. Wickman, B. Picor, M. K. Hibb-Brenner, J. Bristow, and Y. S. Liu, "Fully embedded board-level guided-wave optoelectronic interconnects," *Proc. IEEE*, vol. 88, pp. 780–793, June 2000.

[21] X. Zheng, P. Marchand, D. Huang, and S. Esener, "Free-space parallel multichip interconnection system," *Appl. Opt.*, vol. 39, pp. 3516–3524, July 2000.

[22] P. S. Guilfoyle, J. M. Hessenbruch, and R. V. Stone, "Free-Space optical interconnects for high performance optoelectronic switching," *IEEE Trans. Comput.*, vol. 31, pp. 69–75, Feb. 1998.

[23] D. W. Prather, S. Venkataraman, M. Lecompte, F. Kiamilev, J. N. Mait, and G. J. Simonis, "Optoelectronic multichip module integration for chip level optical interconnects," *IEEE Photon. Technol. Lett.*, vol. 13, pp. 1112–1114, Oct. 2001.

[24] S. Xie, R. Herrick, G. Brabander, W. Widjaja, and U. a. Koelleet, *Reliability and Failure Mechanisms of Oxide VCSEL's in Non-Hermetic Environments*. San Jose, CA: SPIE Photonics West, 2003.

[25] *Bellcore Standard TR332*.

[26] T. Sze, D. Huang, and D. McElfresh.. presented at IEEE Tech. Meeting Computer System Packaging. [Online]. Available: http://www.optical.crosslinks.com/pdf/PitchLinkDataSheet.pdf

[27] Internal Communication.

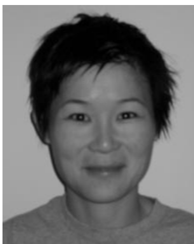
[28] X. Luo, S. B. Zhang, and S. H. Wei, "Chemical design of direct-gap light-emitting silicon," *Phys. Rev. Lett.*, vol. 89, no. 7, p. 076 802-1, Aug. 2002.

[29] A. Polman, “Teaching silicon new tricks,” *Nature Mater.*, vol. 1, pp. 10–11, Sept. 2002.
[30] (2002) ST Microelectronic Sets World Record for Silicon Light Emission. Press Release, Geneva. [Online]. Available: http://www.st.com
[31] K. Banerjee and A. Mehotra, “Analysis of on-chip inductance effects for distributed RLC interconnects,” *IEEE Trans. Computer-Aided Design*, vol. 21, pp. 904–915, Aug. 2002.
[32] B. Kleveland, X. Qi, L. Madden, T. Furusawa, R. W. Dutton, M. Horowitz, and S. S. Wong, “High-Frequency characterization of on-chip digital interconnects,” *IEEE J. Solid-State Circuits*, vol. 37, pp. 716–725, June 2002.
[33] C. T. White and T. N. Todorov, “Nanotubes go ballistic,” *Nature*, vol. 411, pp. 649–650, 7, June 2001.
[34] R. De Picciotto, H. L. Stormer, L. N. Pfeiffer, K. W. Baldwin, and K. W. West, “Four-termina resistance of a ballistic quantum wire,” *Nature*, vol. 411, pp. 51–54, May 2001.
[35] R. Ho, K. W. Mai, and M. A. Horowitz, “The future of wires,” *Proc. IEEE*, vol. 89, pp. 490–504, Apr. 2001.
[36] K. Banerjee and A. Mehrotra, “Accurate analysis of on-chip inductance effects and implications for optimal repeater insertion and technology scaling,” in *Tech. Dig. Papers, 2001 Symp. VLSI Circuits*, pp. 195–198.

Dawei Huang (M’00) received the B.S. degree in precision instruments from Tsinghua University, Beijing, China, in 1997 and the M.S. degree in electrical engineering from the University of California, San Diego, in 1999.

His previous research work at Tsinghua University and the University of California, San Diego were on diffractive optics elements, micro-FP pressure sensor, and free-space optical interconnect. Since 1999, he has been working at Sun Microsystems Chief Technology Office, Physical Sciences Center, San Diego, CA, on various interconnect-related projects. His current research areas include high-speed electrical signaling and circuits, optical interconnect technology, and computer interconnect architecture.

Mr. Huang is a member of OSA.



Theresa Sze (M’00) received the B.S. degree in electrical engineering from the University of Illinois, Urbana-Champaign, in 1991 and the M.S. degree in electrical engineering from the University of New Mexico, Albuquerque, in 1993.

She is a Staff Engineer at Sun Microsystems with the Chief Technology Office, Physical Sciences Center in San Diego, CA. Her interests include interconnect scaling, optical interconnects, and optical device scaling. Since joining Sun Microsystems in 1993, she has worked on silicon device modeling, silicon IC design, hand-held computers, and high-speed system design.

Anders Landin received the M.S. degree in computer science and engineering from Lund University, Lund, Sweden.

He is the Lead Architect at the Advanced Systems Development Center for Enterprise Systems Products at Sun Microsystems, Menlo Park, CA. In 1989, Landin joined the research staff at the Swedish Institute of Computer Science (SICS) where he pioneered the field of cache-only multiprocessor architectures (COMA) and lead the research in parallel computer systems. He has been an architect for various enterprise server products at Sun Microsystems since he joined the company in 1997.

Rick Lytel (M’86) received the Ph.D. degree in physics from Stanford University, Stanford, CA, in 1980.

He joined the Lockheed Palo Alto Research Laboratory, started the Optical Physics Group, and led the development of various integrated optical devices based upon electrooptic polymers. In 1993, he started the Akzo Nobel Photonics company and was its Executive VP and General Manager until 1996, when he left and spent a year with AMP Inc. He joined Sun Microsystems, San Diego, CA, in 1998 as a Distinguished Engineer, and has since created Sun’s first Physical Sciences Center and Advanced System Development Center.

Dr. Lytel has won numerous awards and published nearly 80 papers, presented hundreds of talks, and is Adjunct Professor of Physics at Washington State University, Pullman.

Howard L. Davidson (M’88–SM’94) received the Ph.D. degree in physics from West Virginia Wesleyan College, Buckhannon, WV.

He is currently a Distinguished Engineer at Sun Microsystems RAS Computer Analysis Laboratory, Menlo Park, CA, with responsibility for advanced physical technology. Current research areas include high-power density cooling, optical interconnects, molecular electronics, quantum computing, and SETI. He has previously held product development and research positions at Schlumberger Palo Alto Research, Lawrence Livermore National Laboratory, CRAY Research, and Hewlett-Packard. Previous research areas include integrated circuit testing, supercomputer design, ultralight satellites, high thermal conductivity thermal materials, and electronic instrumentation. He has presented many papers, invited papers, and tutorials on high-speed digital system design, thermal management, and optical interconnects for digital systems. He is the author of the chapter on inspection and testing in *Multichip Module Design, Fabrication, & Testing* (New York: McGraw-Hill, 1995). He holds 24 patents.

Dr. Davidson is a recognized authority on physical system design. He has chaired the IEEE Computer and System Packaging Workshop and the IEEE Workshop on High-Speed Interconnections Within Digital Systems. He has served as an Associate Editor for IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI) SYSTEMS. He is a member of APS and AAAS.