Exploitation of optical interconnects in future server architectures

A. F. Benner
M. Ignatowski
J. A. Kash
D. M. Kuchta
M. B. Ritter

Optical fiber links have become ubiquitous for links at the metropolitan and wide area distance scales, and have become common alternatives to electrical links in local area networks and cluster networks. As optical technology improves and link frequencies continue to increase, optical links will be increasingly considered for shorter, higher-bandwidth links such as I/O, memory, and system bus links. For these links closer to processors, issues such as packaging, power dissipation, and components cost assume increasing importance along with link bandwidth and link distance. Also, as optical links move steadily closer to the processors, we may see significant differences in how servers, particularly high-end servers, are designed and packaged to exploit the unique characteristics of optical interconnects. This paper reviews the various levels of a server interconnect hierarchy and the current status of optical interconnect technology for these different levels. The potential impacts of optical interconnect technology on future server designs are also reviewed.

1. Introduction

Over the last several decades there has been steady improvement in the performance of all aspects of computer systems, as base computing technologies have evolved from relays and vacuum tubes through several generations of CMOS silicon transistor technology. However, various aspects of the overall system performance have improved at different rates, as exemplified by the slow rate of improvement in memory access time (~5–10% per year) in contrast to, for example, processor performance per chip (in excess of 40% per year, particularly on applications that make good use of caches). A similar, although less drastic, disparity exists in the improvement rate for off-chip I/O bandwidth relative to per-chip performance. The most recent International Technology Roadmap for Semiconductors (ITRS) projects that while per-chip performance will continue to improve at a rate of approximately four times every three to four years, the number of signal pins per module will increase by only two times over the same period, and the maximum bit rate per signal pin will increase by only 35% [1]. Thus, the total off-chip I/O bandwidth (BW) (pin count times bit rate per pin) will increase by roughly 2.7 times, while the internal chip performance improves by four times. Over a decade, this difference in improvement rates would result in a difference in ratio of nearly three times, which would dramatically affect balanced system design. The actual improvement of off-chip bandwidth may in many cases be even lower, since the maximum off-chip bit rate will often be limited to a small number of signal pins, with many pins operating at a lower bit rate.

Other trends highlighted by the ITRS also point to the increasing importance of chip packaging and off-chip interconnect. The cost of packaging as a fraction of the overall packaged chip cost has been steadily increasing. Chip packages have increased in pin count at 10% per year while decreasing per-pin cost only 5% per year, yielding a per-chip increase in package cost of roughly 5% per year, whereas silicon has provided a performance improvement of four times every three to four years, at a nearly constant cost.

These trends illustrate the expectation that many high-performance chips will be increasingly limited by off-chip bandwidth, and there will be increasing need for technologies that provide substantially improved chip-to-chip interconnect capabilities.

At the same time, there has been dramatic improvement in the cost–performance of optical interconnect technologies, at rates which by some

0018-8646/05/$5.00 © 2005 IBM

measures exceed the rate of improvement for silicon chips. These trends are due partly to the natural improvement in cost–performance with greater maturity of optical interconnect technologies and partly to the use of optics in interconnect scenarios where they have not previously been used, with consequent engineering optimizations.

To a large extent, the use of optical interconnect technologies in commercial products has been limited to direct replacement of electrical cables by optical cables to support longer link lengths. As link bit rates have steadily increased, optical cables have replaced electrical cables for shorter cable lengths [2]. As optical interconnect technology becomes less costly and better integrated, there is increasing opportunity for optics to move "inside the box," causing substantial impacts on system packaging, interconnect topology, communication traffic patterns, and other aspects of server architecture [3].

As optical interconnects are considered in servers for progressively shorter links such as I/O buses, memory buses, and particularly symmetric multiprocessor (SMP) or system buses, they will also have to accommodate the demands of these shorter links. This paper considers the effects that these demands, in areas such as bandwidth–distance product, power usage, packaging, bit error ratio, and acceptable cost, will have on the adoption of optical technology for the various server interconnect links. This paper provides an overview of current status and trends in optical technologies at various levels of the server interconnection hierarchy, and projects the impact that new optical interconnect capabilities may have on future server architectures.

Section 2 describes the server interconnection hierarchy from physical, logical, and topological perspectives. Section 3 surveys many of the different types of servers currently being built, with an emphasis on how they differ in the mapping of physical to logical hierarchies for different system scales. Section 4 describes some current research and development efforts in optical interconnection technologies that are likely to have an impact on server architecture and design. Section 5 discusses some of the most likely impacts on server architecture resulting from new developments in optical interconnection, and Section 6 provides a summary.

2. Interconnects

Interconnect links in servers can be described in terms of their logical function and physical implementation, and by the topology of the interconnection network. We provide a short overview of these three different views of server interconnects to help facilitate understanding of areas in which optical interconnects can provide advantages in future server designs.

Logical functions of the interconnect hierarchy for servers

The following descriptions of links are based on their logical functions. They are listed roughly in order of decreasing bandwidth, increasing link distance, and decreasing latency sensitivity. In practice, some of these logical buses share the same physical hardware in many server designs. **Figure 1** shows the logical hierarchy of interconnects used in server systems. Each individual server typically includes a subset of these buses, although some high-end systems may include them all.

SMP and SMP expansion links

An SMP system is a computer system which has two or more processors closely connected, managed by one operating system or hypervisor,[1] sharing the same memory and input/output devices between processors. The SMP bus (or host bus) links processors together, allowing them to share resources and maintain coherence among copies of data in memory and in processor caches. Typically, operations executed by a processor generate bus operations which are transmitted to any or all of the other components (processors, memory controllers, or I/O devices) which may be affected by the operation. This transmission may occur either through system-wide broadcast or through targeted transmission to the affected components. Broadcast buses are typically less complex but require more bandwidth [e.g., link widths up to 16 bytes plus overhead, with aggregate bandwidth up to several hundreds of gigabits per second (Gb/s)], whereas directory-based bus designs can reduce bandwidth requirements by limiting transmission of a bus operation to only the set of components that are affected by it. SMP buses are typically the highest-performance and most performance-sensitive links of a system, particularly for large SMP systems. Along with bandwidth, end-to-end bus latency, including control of latency components at transmitter and receiver, is extremely important, since processors must usually wait until SMP bus operations complete before continuing processing. SMP expansion links (or scalability links) can provide extra distance or fan-out for larger systems, using extra links and multiple chip crossings.

Memory and memory expansion links

A memory bus is used to interconnect dynamic random access memory (DRAM) chips to the memory controller. This is typically an industry-standard bus, having many memory chips connected to a single multidrop bus except in the highest-speed implementations. Bit rates for multidrop memory buses range up to hundreds of MHz. High-performance point-to-point memory buses can

[1] A *hypervisor* is a scheme which allows multiple operating systems to run on the same computer at the same time.

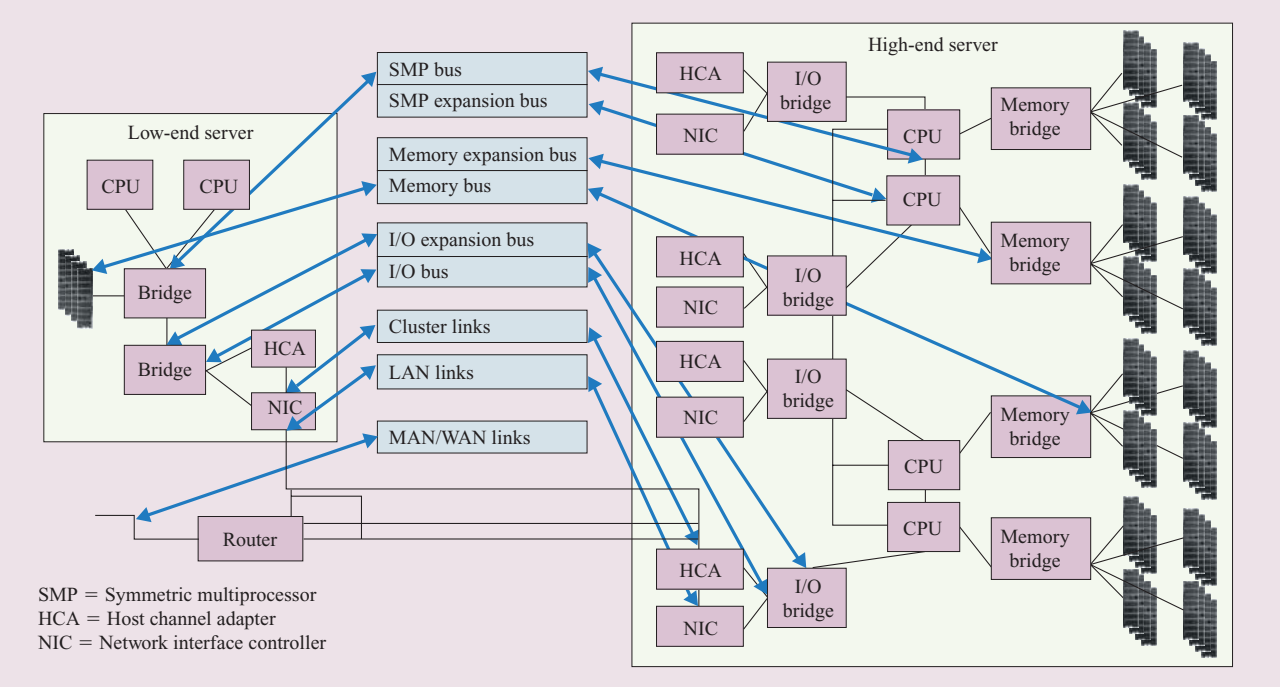


Figure 1

Server logical interconnection hierarchy.

support a range of bit rates of 3–6 Gb/s per pin pair. Low-end servers may support as few as eight memory chips, while high-end SMP servers may support hundreds of memory chips, using multiple buffer/hub chips connected to the processor through memory expansion buses. The memory hub chips then provide fan-out, multiplexing, and packet switching to support connectivity to many more DRAM chips than can be supported by direct connection to the memory controller.

I/O and I/O expansion links

An input/output bus is used for attaching I/O devices such as disks and network cards to the server. Typical widths are 32–100 bits. Typical bit rates range from 33 MHz (PCI[2]) to more than 2 Gb/s (PCI-Express**), with multidrop busing used at the lower speeds and point-to-point links at the higher speeds. Similarly, I/O expansion links with hub and bridge chips provide connectivity (fan-out and distance) to the I/O cards supporting a high-end system.

Cluster links

Cluster networks are used to interconnect multiple SMPs, each with its own memory and operating system, to form a larger cluster server. The individual SMPs do not share memory with one another and communicate only by passing software-initiated messages back and forth. Because of this, the amount of communication traffic between SMPs within a cluster may be roughly an order of magnitude less than the amount of communication traffic within an SMP. There are also clusters that use more tightly integrated communications mechanisms—for example, global memory addressability, which allows processors to directly reference memory attached to other processors using explicit remote load and remote store operations. Compared with SMP bus links, cluster network links usually have reduced bandwidth and latency requirements, but they must typically span much longer distances (10–100 m). Compared with local area network/storage area network (LAN/SAN) links, cluster links more typically have multiple lines per link to obtain higher aggregate bandwidth, whereas LAN/SAN links will usually have only one line per link to simplify cabling through building walls. The difference between cluster and LAN/SAN links is not precise, however, since clusters are often interconnected using LAN network technology.

LAN/SAN links

Local area network (LAN) links are used to interconnect machines, which may be quite heterogeneous, across

[2] Peripheral Component Interconnect.

distances of tens to hundreds of meters, i.e., within a building. Storage area network (SAN) links are used to connect servers with storage systems, and are typically useful when the stored data are shared among multiple servers. Typically Ethernet and Fibre Channel are used for LAN and SAN networks, respectively, and the link bandwidths are currently of the order of 1 to 10 Gb/s.

MAN/WAN links

Metro area network/wide area network (MAN/WAN) links extend server communication outside the machine room or building, typically with standards-based networks. These networks can span distances of many kilometers. WAN links are terminated in dedicated switch/router boxes, connected to the servers by LAN links. Perhaps as 10 Gb/s Ethernet (10GEnet) MAN and WAN links using single-mode fiber and long wavelength lasers are more commonly used, they may be directly terminated in server frames, but they will be packaged using normal LAN/SAN/Cluster I/O adapter cards, and will not place new requirements on server packaging.

Packaging and physical interconnect hierarchy

Current microprocessor chips are placed on a variety of physical packages with varying capabilities for interconnect wiring density and cooling. Low-cost packaging is typically done with organic (plastic) packaging. Ceramic packages offer superior performance and pin density and are often used in high-performance systems. Multichip modules provide significant increases in bandwidth between chips on the module and increased off-module signal I/O count. Typically multichip modules are used to package together processor and cache chips, or to package multiple processor chips in an SMP configuration. In the future, multichip modules could also be used to package together optical components and processor chips. Single-chip modules or multichip modules are mounted on circuit boards, which are packaged in shelves or racks. Low-end systems typically contain one main circuit board, with I/O cards and memory cards plugged onto it with card-edge connectors. Higher-end systems will have multiple circuit boards, typically connected to both sides of a mid-plane circuit board. Cables, implemented with either copper or optical fiber, then connect the shelves or racks to create larger clusters.

Modern servers employ a hierarchy of interconnect technologies based on packaging, performance, cost, and expandability requirements for different models. Interconnects used in servers range from WAN links to on-chip data interconnect wires. These interconnects can be grouped in several categories characterized by link distance. Each category has different power and area requirements, bit error ratio requirements, traffic patterns, and flow control methods, based on the need(s) of the link(s) it serves. Wireless networks, which are important for laptop computers and mobile devices, are not shown here, since servers typically use wired communications.

Figure 2 shows the physical interconnect hierarchy described in the following paragraphs. Also shown in the figure is the use of optics vs. copper cables for the various types of interconnects. As shown in the figure, optics has been used for the longer-distance links for several decades, with progressively less use of optics for shorter links. Currently, for bit rates in the range of 1–10 Gb/s, optical technologies are in common use for links longer than a few meters.

Intra-chip wiring is the densest wiring found interconnecting components within a chip. This can include buses many bytes wide running at the processor frequency. They are used to interconnect processors to on-chip caches, memory ports, and I/O ports, and for processor-to-processor communication for multicore chips. Because the resistivity of wires increases as their cross-section area shrinks, the longer on-chip buses tend to be fabricated with wires wider and thicker than minimum geometry and sometimes run at less than the processor speed. Because of materials and processing challenges and the capability of on-chip electrical wiring, optical links are not likely to be used for intra-chip communication for the foreseeable future.

Intra-module wiring: Multichip modules can contain from two to more than a dozen chips packaged directly on a single substrate. These modules provide significantly higher bandwidth between chips than can be obtained from standard card packaging. While multichip modules are typically currently implemented with ceramic or laminate substrates, silicon carriers are also being studied for future systems (see the paper by Knickerbocker et al. [4] in this issue). Multichip modules are typically used to package together processor and cache chips or to package multiple processor chips together in an SMP configuration. In the future, multichip modules could also be used to package together optical components and processor chips.

Intra-card wiring is used to connect chips on the same card. Typically one to four processor chips (or a single multichip module) are mounted on a card, along with pluggable memory modules and I/O support chips. Optical modules used in current servers are most often mounted on cards and interface with the processor chips through card wiring.

Card-to-card wiring is used to connect chips on different cards. For larger systems, 2 to 16 cards are often plugged into a backplane containing the interconnect links between the cards. Backplanes can be completely passive or can contain active switching components.

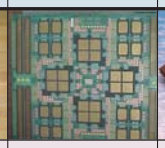
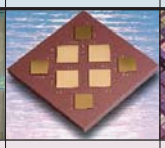
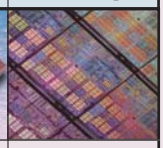
	MAN/WAN	Cables–long	Cables–short	Card-to-card	Intra-card	Intra-module	Intra-chip
							
Length	Multi-km	10–300 m	1–10 m	0.3–1 m	0.1–0.3 m	5–100 mm	0–20 mm
No. of lines per link	One	One to tens	One to tens	One to hundreds	One to hundreds	One to hundreds	One to hundreds
No. of lines per system	Tens	Tens to thousands	Tens to thousands	Tens to thousands	Thousands	Approximately ten thousand	Hundreds of thousands
Standards	Internet Protocol, SONET, ATM	LAN/SAN (Ethernet, InfiniBand, Fibre Channel)	Design-specific, LAN/SAN (Ethernet, InfiniBand)	Design-specific and standards (PCI, backplane InfiniBand and Ethernet)	Design-specific, generally	Design-specific	Design-specific
Use of optics	Since the 1980s	Since the 1990s	Present time, or very soon	2005–2010 with effort	2010–2015	Probably after 2015	Later

Figure 2

Server physical interconnection hierarchy.

For the highest-performance SMPs, the pluggable interconnects between the cards and the backplane are pushing the limits of connector technology, and in some systems the total bandwidth through a card edge can be a bottleneck.

Cables—short (1–10 m): Two or more backplanes can be interconnected with pluggable cables to extend the server size or to interconnect the server to an I/O expansion rack. High-bandwidth versions of this are typically limited to interconnecting backplanes within a single rack, or interconnecting between physically adjacent racks. Cables of this length generally use electrical transmission.

Cables—long (10–300 m): Longer cables are used within a computer machine room to implement cluster or LAN/SAN networks. These cables are distinguished from shorter cables by the fact that they may, depending on transmission bit rate and cable length, be implemented with either electrical transmission or optical transmission using multimode optical fibers. Newer multimode optical fibers allow links with a bit rate–distance product of up to more than 2000 MHz·km (i.e., 4 Gb/s over 500 meters), whereas copper cables have a bit rate–distance product more than ten times lower.

MAN/WAN: Communications outside a machine room can extend to tens or even hundreds of kilometers at multi-Gb/s speeds, using single-mode fibers and long-wavelength (1,300 or 1,550 nm) transceivers. In server applications, these links are commonly used for remote data replication, e.g., for disaster recovery. These links may directly exit a server I/O card or, more typically, may be generated in a router connected to the server through a LAN link.

Interconnect topology

Full mesh interconnect

In this simplest of logical structures, every source has a point-to-point interconnect link with every destination with which it may have to communicate. This structure eliminates the need for arbitration and provides the lowest possible latency. However, the high signal count required for all-to-all connectivity restricts this topology to very small networks, e.g., with four nodes or fewer.

Shared bus

For most of the early SMP server designs, it was common to interconnect multiple processors, memory, and I/O ports using a single shared multidropped bus. This is a "many-to-many" configuration, with multiple senders and multiple receivers on each common electrical line. An arbitration mechanism is used to select one sender to transmit on the bus at a time. This approach has the advantages of simplicity of design and ease of system

expansion. However, the electrical characteristics of a multidropped bus limit its useful frequency to the 200–400-MHz range, and limit its length to 10–20 cm.

As system frequencies move into the GHz range, shared buses are generally being replaced by point-to-point interconnects, with impedance-matched transmission lines. This migration improves the bus frequency, reducing the wire count per chip, and eliminates the need for lengthy arbitration among multiple transmitters. Several options for point-to-point topologies are discussed below. A similar migration from multidropped bus to point-to-point link has occurred with Ethernet, as speeds increased from 10 and 100 Mb/s to 1 and 10 Gb/s. Similarly, the multidrop PCI/PCI-X bus architecture is being supplanted by the InfiniBand** and PCI-Express architectures, which use point-to-point signaling with switching router/hub chips to accomplish fan-out.

Switched

The most generally useful networks currently are multistage packet-switched networks, in which modules are connected with point-to-point links, and switching elements within the modules route data and control packets to their proper destination on the basis of explicit routing headers or address-based routing. There are a wide variety of switched network implementations, depending on topology and integration of switching functionality in other components. The simplest switched networks are loops, with each node on the loop containing a three-port switch that allows packet insertion into and removal from the loop and forwarding along the loop. More complex topologies, such as hypercubes, 2D and 3D tori, and fat-tree or other topologies, provide varying benefits at various network sizes and can be characterized by performance parameters such as chip count, scaling characteristics, network diameter, and bisection bandwidth.

Direct-switched

Switched networks can also be distinguished by the location of the switching capability: Switching can be integrated into end-nodes, into specialized switch-only chips, or into both. End-nodes containing switching features can be connected directly (to form "direct" networks) without external switch chips. However, implementing switching within a chip that also incorporates other functions limits the port count that is practically achievable—typically to six or eight ports. By contrast, a dedicated switch chip can be built with tens of ports, allowing construction of larger-scale networks with lower diameters (fewer chip hops across the farthest-separated nodes). A further design parameter arises in balancing the number of ports per switch chip with the per-port bandwidth: A chip with higher port bandwidth will typically allow fewer ports per chip.

3. Survey of current server designs and packaging

This section describes the variety of packaging styles and technologies that are used to construct servers of various types. Servers vary in size over a wide range of configurations, from single-board systems up to systems occupying multiple frames. Server design and packaging are determined by multiple factors, including interconnect, cooling, modularity, and operating system extent. **Figure 3** shows a classification of various scales of servers, along with pictures of 2005-era systems to illustrate the relative physical sizes of the various systems.

Server designs

Low-end servers

Low-end servers typically consist of one to two microprocessor chips mounted on a card together with dual in-line memory modules (DIMMs) and I/O controllers, along with power, cooling, and cabling infrastructure. They can be packaged as a deskside unit or mounted in a 1U or 2U (1.75-in. or 3.5-in.-high) shelf that fits in a standard rack. They generally have good price–performance because of cost-optimized designs, limited memory and I/O capacity, minimal support for redundancy and concurrent repair, and no support for upgrading to larger systems.

Mid-range SMPs

Mid-range SMPs are typically built from smaller building blocks (boards) containing from one to four microprocessor chip, memory, and I/O ports. Up to four of these building blocks can be interconnected with a cost-optimized SMP bus, commonly implemented as pluggable copper cables. Future bandwidth density requirements may justify the use of optical fibers for these pluggable cables. Mid-range servers typically have a larger memory capacity per microprocessor chip than low-end servers, and they typically support more extensive attachments for I/O devices. Mid-range servers may support concurrent maintenance and improved redundancy.

High-end SMP machines

High-end SMPs support larger numbers of processors, larger amounts of memory, extensive I/O expansion networks which can span multiple frames, and higher levels of reliability, redundant components, and concurrent maintenance. High-end SMPs are typically assembled from basic building blocks (boards) that tightly integrate four microprocessor chips, memory,

Figure 3

Classification of server systems.

and I/O ports. Four to 16 of these building blocks can be plugged into a high-performance SMP network that is typically implemented as a large custom backplane. SMP bus bandwidth requirements push connector technology to the extreme because of the large amount of traffic that must be handled—already hundreds of Gb/s in existing servers and growing.

These basic machine types can be combined into arrays, using several different form factors, as described below.

Clusters and parallel-processing machines

A computer cluster typically consists of two to thousands of SMP servers connected through a cluster network or LAN and used as a single computing resource. Generally, two-node and three-node clusters are employed for redundancy-based reliability. Larger clusters are used to provide significantly higher performance than that from a single high-end SMP. Typically, each server within a cluster has its own memory, communicating with the other servers only by passing software-initiated messages back and forth.

Blades

Blade-style packaging is a particular type of cluster configuration based on modularity and dense packaging of multiple small SMP blades in a rack. Each blade is typically built as a board containing only the components of a low-end server—one or two processor chips, memory, and standard I/O ports. Current systems restrict SMP interconnects to processors packaged within the same blade. The blades are plugged into a backplane or midplane that provides the LAN/SAN or cluster level of interconnect between multiple blades and external systems, as well as shared infrastructure (power supplies, fans/blowers, network and power cables, and integrated system management). This sharing of network, power, packaging, and cooling infrastructure increases density, simplifies heterogeneous system management, and reduces cost for installations requiring more than a few blades.

Today, the backplane interconnect between blades is typically implemented with 1-Gb/s Ethernet or 2-Gb/s Fibre Channel, which provides adequate performance for the systems which can fit on a blade. In the near future this is likely move up to InfiniBand or 10GEnet.

Server physical interconnect hierarchy mapping to logical hierarchy

A summary of the uses of the various interconnect technologies in the various scales of systems is shown in **Table 1**. Several points are evident from the table. SMP links have, to date, been implemented only in on-card or backplane links, since the required widths for a

Table 1 Mapping logical and physical interconnect hierarchies for various system scales.

Interconnect–logical levels	*System scale*			
	Low-end	*Blade*	*Mid-range SMP*	*High-end SMP*
SMP bus	On-card	On-card	On-card	On-card or backplane
SMP expansion	—	—	Cable, if present	Cable, if present
Memory expansion	—	—	On-card	On-card
Memory	On-card	On-card	On-card	On-card
I/O expansion	—	—	Cable, if present	Cable
I/O	On-card	On-card	On-card or backplane	On-card or backplane
Cluster	Cable	Backplane	Cable	Cable
LAN/SAN	Cable	Backplane	Cable	Cable

competitive bus—typically 8 or 16 bytes wide, plus overhead—are impractical to implement with copper cables at competitive cost. SMP expansion links and I/O expansion links are present only in larger systems that extend over multiple boards and shelves. All systems have I/O buses and cluster/LAN/SAN connectivity, but typically only the blade-style systems integrate both ends of a cluster or LAN/SAN link on the same backplane.

Clusters may be, and commonly are, built using all of these base server types, by incorporating cluster adapters and switch network infrastructure. Also, specialized clusters (particularly large-scale or ultra-scale clusters) are built as specialized machines with cluster interconnect network components tightly integrated into the node structures as described later, in Section 5.

4. Optical interconnect technology and trends

This section describes the use of optical technologies in various aspects of server interconnection; it describes recent demonstrations of significantly improved optical interconnect technologies that may be integrated into future server systems.

In the first subsection, existing single-channel bidirectional links based on the optical Ethernet standard transceivers are considered and found to be unsuitable for server interconnections. The second subsection describes existing parallel link modules which contain 12 unidirectional links operating at up to roughly 3.5 Gb/s per channel. The technologies used in these modules, while not yet at the level required for server optical interconnects in terms of density and per-channel bit rate, are a significant step in the right direction. Since the existing commercial modules do not meet the needs of server interconnects, the remaining sections examine current research into higher-performance modules. The third subsection shows how commercial 3-Gb/s parallel optical modules can be made to operate at 10 Gb/s per channel without changing the standard packaging. For the first uses of server optical interconnects, 10 Gb/s is fast enough. The fourth subsection examines still wider parallel links, either with more fibers or with multiple wavelengths on a single fiber. The fifth subsection describes a multiple-fiber, multiple-wavelength module that has been reported by Agilent Technologies. This module, while not yet commercialized, has the density and speed required for use in servers. Finally, the sixth subsection looks briefly at the emerging technology of optical waveguides, which have the possibility of eventually replacing copper traces or optical fibers for on-card and backplane interconnects.

Historically, the use of optical technologies vs. copper link technologies has been governed by tradeoffs between link cost (including development cost, component cost, power penalty, and system complexity cost, in various ratios) for the two technologies at various link lengths and bit rates. In general, optical technologies operating at a particular bit rate have been used for links longer than a particular crossover length, with electrical interconnects used for shorter links. The crossover length is typically shorter at higher bit rates. The optical interconnect research and development projects described in this section are aimed at decreasing the crossover lengths for links at various bit rates.

Ethernet technology comparison with SMP link requirements

The 10GEnet family of standards specifies a number of media-dependent transmission technologies for LAN and WAN applications. The physical media described in various ancillary standards, some still in progress as of this writing, include backplanes, copper cable, multimode fibers, and single-mode fibers. Ethernet networks are well established at network speeds of 1 Gb/s and lower, and (barring potential competition from InfiniBand networks

in tightly coupled and high-performance cluster applications) there is little risk in expecting 10GEnet to be widely deployed through this decade for LAN and MAN applications. To assess the further impact of 10GEnet technology on server architectures, however, it is instructive to consider whether 10GEnet links could be used for SMP bus links.

There are several design factors that affect SMP bus links, dictated by their use in the system. The first is the uncorrected bit error ratio (BER), which should be in the range of 1×10^{-20} or better. Uncorrected bit errors on an SMP link cannot be allowed, since they may affect any piece of data or code handled by the system. The required bit error ratio is determined by system lifetime. In a system having an aggregate SMP bus bandwidth of 100 Gb/s (10^{11} bits per second), a 1×10^{-20} BER would give an average of one error every 10^9 seconds (31.7 years), which is sufficient to ensure essentially error-free operation during the expected lifetime of a system. The 10GEnet link definitions are rated at either 1×10^{-12} or 1×10^{-15} BER (i.e., many errors per day), and do not include error-correction coding. Very substantial additional error-correction circuitry, with attendant latency and complexity penalties, would be needed to overcome this difference between raw and corrected BER.

A second factor is an adequate bit rate–distance product. While the optical versions of Ethernet are acceptable here, most of the electrical versions of Ethernet (XGMII, XAUI, and XFI [5])[3] consist of multiple parallel data lines at a bit rate well below 10 Gb/s. The 10GEnet XGMII interface (37 bits wide, at less than 300 Mb/s) is too wide and low-speed to be competitive in SMP links. The XAUI interface is a four-lane-wide, striped, 8B/10B-encoded, 3.125-Gb/s differential interface with a reach of 70 cm over standard circuit boards. The XFI interface carries full-rate (10.3125-Gb/s) coded data a distance of 20 to 30 cm from the framer to the electronic-to-optical (E/O) and optical-to-electronic (O/E) converter module. Either XAUI or XFI could provide an adequate bit rate–distance product to be used for SMP links, but have other deficiencies in power usage, latency, and form factor, as described below.

A third factor is low power consumption, well below 1 W for a 10-Gb/s line, more typically less than 0.2 W. This requirement arises because the combination of interconnects and processors must not exceed the system cooling capabilities. The electrical Ethernet signaling described in the previous paragraph has a typical power consumption of roughly 1 W for 10 Gb/s, several times higher than typical SMP links. For optical Ethernet, the short-reach, 850-nm multimode fiber implementation of 10GEnet typically requires 2 W for just the E/O and O/E portions, and 3–5 W for a full transceiver. The optical sources for this implementation are 850-nm vertical-cavity surface-emitting lasers (VCSELs) [6]. This option provides the lowest power of the various optical Ethernet standards, yet it is still too power-hungry for SMP applications.

Although 10GEnet technologies are well optimized for LAN (long cable) and MAN/WAN links, neither copper nor optical 10GEnet technologies are well suited for SMP bus applications. Along with the primary considerations discussed above—power usage per Gb/s, bit error ratio, and link length—other considerations, such as latency through the Ethernet framer and mechanical form factors, preclude the use of 10GEnet for high-bandwidth SMP buses. We conclude that the 850-nm fiber transceivers work well for clusters and LAN links; however, today's optical Ethernet transceivers are not suited for use in SMP or SMP bus expansion links in present and future systems.

Multimode 12× fiber ribbon InfiniBand-compatible links

As described above, the 850-nm multimode fiber is the medium that provides characteristics most closely suited to the requirements of SMP links. However, other 850-nm multimode optical module technologies share the advantages of the 850-nm 10GEnet implementation while also providing greater parallelism, better power efficiency, and better opportunity for correcting bit errors that occur during optical transmission. The IBM High Performance Switch (HPS) cluster network [7] utilizes a parallel optical module that is available commercially from multiple manufacturers. Known as the "SNAP12" module (**Figure 4**), this device may also be used for InfiniBand optical links at 12× width, operating at 2.5 Gb/s per channel [8]. A narrower InfiniBand link definition, with four fibers per direction, uses the same technology.

The transmitter version of a SNAP12 module contains a linear array of 12 VCSELs [6] operating at 850 nm; these are co-mounted on an internal electrical flexible printed circuit (flex) with the necessary electronics to drive the lasers. The other end of the flex contains an electrical connector (a 100-pin MEG-Array** connector). This electrical connector and the industry-standard multifiber ferrule for the 1×12 fiber ribbon array largely determine the form factor for the module. The electrical inputs to the module are relatively simple, using differential current mode logic for the high-speed data inputs and outputs. The receiver version of the module consists of an array of surface-illuminated photodiodes (PDs), with amplifiers to drive the output electrical signals. Although VCSELs can be produced to operate at

[3] XGMII: 10-Gigabit Medium-Independent Interface; XAUI: X (for "ten") Attachment Unit Interface; XFI: 10-Gigabit Small-Form-Factor Pluggable Electrical Interface.

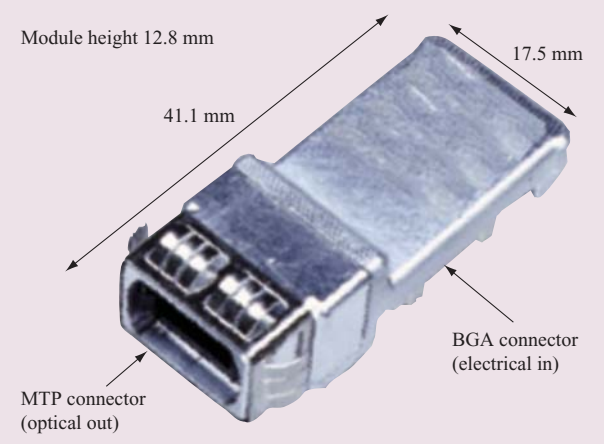


Figure 4

SNAP12 module showing form factor and dimensions.

wavelengths from approximately 650 nm to 1,550 nm, the 850-nm wavelength is chosen primarily because it is where the lowest-cost, highest-reliability devices from multiple vendors can be found. Commercial SNAP12 modules commonly operate at up to approximately 3.3 Gb/s per channel, for an aggregate throughput of approximately 40 Gb/s for a module pair.

The choice of VCSELs for the lasers in parallel optical modules such as SNAP12 and the Infineon Paroli** arises from the need to fabricate a low-cost 1×12 array of lasers for the parallel optics. Other communications lasers, such as Fabry–Perot and distributed feedback lasers, are non-surface-emitting devices in which a cleaved facet acts as the output mirror. Such lasers naturally lend themselves to linear array configuration but are generally more expensive than VCSELs because of yield and testing issues.

The fibers that connect the receiver and transmitter SNAP12 modules consist of a 1×12 linear array of fibers on a 250-μm pitch. The link uses multimode fibers (MMFs) with a core diameter of either 50 or 62.5 μm. The choice between multimode and single-mode fibers is based on transceiver cost and the short-distance requirement of optical interconnects. The use of MMFs lowers the transceiver cost by providing a mechanical alignment tolerance of 5–10 μm between the optoelectronic device and the fiber array, whereas single-mode fiber alignment requires submicron mechanical alignment precision. The extra cost of this precision mechanical alignment is justified in single-mode telecommunication systems, in which the transmission distance is measured in tens to hundreds of kilometers and the fiber costs dominate the link cost. For data communication systems, on the other hand, and for the coming optical interconnects in servers, the achievable distance with MMFs of up to 1 km [9] is more than adequate.

12 × 10-Gb/s VCSEL-based multimode fiber link

The SNAP12 technology described in the previous section offers a 40-Gb/s aggregate fiber optic link bandwidth when a 3.33-Gb/s per channel bit rate is used. For many anticipated server connections, this link rate is not fast enough. Since package development can dominate the cost and timeline for an optical transceiver, reuse of the SNAP12 package at 10 Gb/s per channel could speed the development of higher-throughput parallel optical modules. Although a faster commercial version of SNAP12 has yet to be announced, research into 12-channel parallel optics modules has been pursued at several corporate research laboratories [10, 11]. This section describes the results of a 120-Gb/s link based on the SNAP12 package [12]. This work explored the performance of the existing SNAP12 flex-based technology and BGA (ball grid array) connector at 10 Gb/s per channel. The important issues are signal integrity, crosstalk, power dissipation, and link bit error ratio.

To make the SNAP12 package ready for 10 Gb/s per channel, the flex layout was reengineered to minimize skew and improve shielding of the differential transmission lines. Power-supply decoupling was also improved. Characterization of the MEG-Array electrical connector showed that it could support 10 Gb/s, with reassignment of the pins (again to minimize skew and improve shielding). The optical coupling and alignment schemes remained unchanged, and wire bonds continued to be used for both the integrated circuits (ICs) and the optoelectronic devices. The 3-Gb/s commercial SNAP12 ICs were replaced with 10-Gb/s SiGe ICs. Since the power dissipation of the 10-Gb/s SiGe ICs was nearly the same as for the 3-Gb/s ICs, the thermal properties of the SNAP12 package were sufficient for heat-sinking. The transmitter and receiver each dissipated about 110 mW per channel. It should be noted that these ICs were not optimized for low power dissipation.

The optical signal integrity of the transmit eye was quite good, with rise and fall times of 38 ps [see **Figure 5(a)**]. The received electrical signal is also shown [**Figure 5(b)**], with rise and fall times of 36 ps following several connectors and cables. Electrical package crosstalk of the MEG-Array plus flex circuit was measured to be less than 30 dB up to 20 GHz. To achieve a good overall level of crosstalk in the transmitter, it was found that a common cathode VCSEL array should be avoided. Transmitters that were built with isolated VCSEL devices (on a semi-insulating substrate) were measured to have a crosstalk penalty of less than 0.1 dB.

This measurement indicates that the crosstalk within the IC is very small, and that wire bonds can be used at these data rates. The worst-case crosstalk penalty in the receiver was about 3 dB. The worst-case receiver crosstalk occurs when the total link attenuation is at a minimum, so that the signals reaching the receiver are at their largest. With nominal amounts of link attenuation, the receiver crosstalk penalty was about 2 dB.

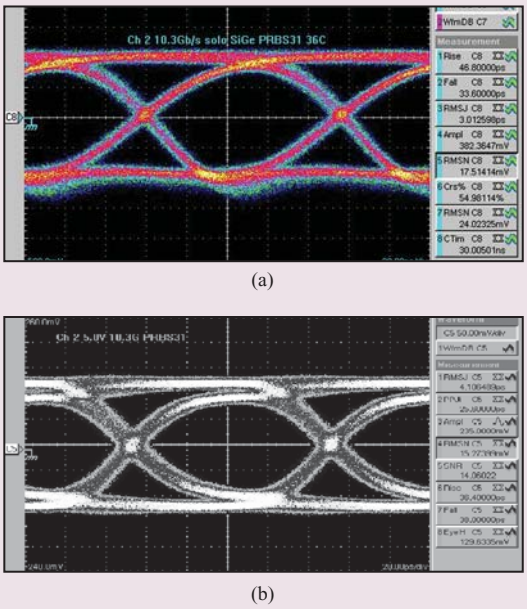
Link experiments were conducted on a range of different fiber lengths from 10 m to 300 m. For the 300-m experiments, next-generation multimode (OM-3) 50-μm fiber was used. During one long-term experiment in which the link was operating at 120 Gb/s, nine errors were observed in 68 hours for a BER better than 1×10^{-14}.

This characterization of the signal integrity, power dissipation, crosstalk, and link error rate showed that the existing SNAP12 form factor and technology are suitable for the higher data rates demanded by server interconnects. A similar investigation using 10-Gb/s ICs on a flex but without a pluggable connector led to a similar conclusion [11].

Increasing density beyond 1 × 12 optical modules

The preceding subsection demonstrated the ability of existing 1×12 parallel optical modules to migrate to higher bit rates by optimizing the O/E devices, Si driver electronics, and electrical flex packaging. However, as noted in Section 2, a substantial amount of communication is required between nodes, particularly in the SMP and memory links, and electrical pin-count densities on the first-level package (e.g., an MCM or SCM) limit the amount of information that can be moved on or off the package. As a result, it would be very desirable to have the electrical interface to the optical interconnect be very close to the processors. In addition to overcoming pin-count limitations, placing the optics close to the processors avoids the need for transferring the electrical signals to optical modules located at the card edge. At high bit rates, there is a substantial power penalty for driving the electrical signal to the card edge. Because of the pin-count limitations and because of the power required to drive electrical signals to the card edge, 1×12 modules with a SNAP12 footprint, even if operating at high bit rates of 10 Gb/s or beyond, are not dense enough.

Increased density can be achieved by decreasing the module size, and smaller footprints help to enable operation at higher bit rates. However, given the high number of signals and congestion at card edges, managing fibers bundled into 1×12 cables would probably require aggregation of the 1×12 optical ribbon cables into 2×12, 4×12, or 6×12 cables at the board edge.



(a)

(b)

Figure 5

Measured 10.3-Gb/s eye diagrams of one channel of the reengineered 1×12 Snap12 optical modules: (a) transmitted; (b) received.

A more attractive approach to increasing density is to increase the module parallelism. Rather than continuing to go to wider linear arrays, however, there is likely to be a trend toward more use of two-dimensional arrays such as 2×12, 4×12, or 6×12 arrays of fibers and VCSELs or photodiodes. Although such modules may be bigger than a 1×12 SNAP12 module, such modules can provide better aggregate density per channel. There are several examples of modules, either under development or shipping as commercial products, that achieve this improved density [11, 13].

One key issue to be addressed for two-dimensional parallel modules is elimination of the wire bond that is used in the 1×12 parallel modules to electrically contact the VCSEL or PD arrays. It is very difficult to contact an array with more than two rows at the typical 250×250-μm pitch of typical MMF ribbon and connectors using wire bonds. This problem has been addressed by a flip-chip attach between the drive electronics and the O/E array [14], as illustrated in **Figure 6**. This flip-chip attach presents some thermal management issues for the temperature-sensitive VCSELs, since they are attached electrically to the IC on one side and optically coupled to the fiber on the other side, leaving no direct surface of the VCSEL available for a heat sink. However, the thermal issues have been found to be manageable by optimizing the VCSEL for reliable operation at slightly elevated

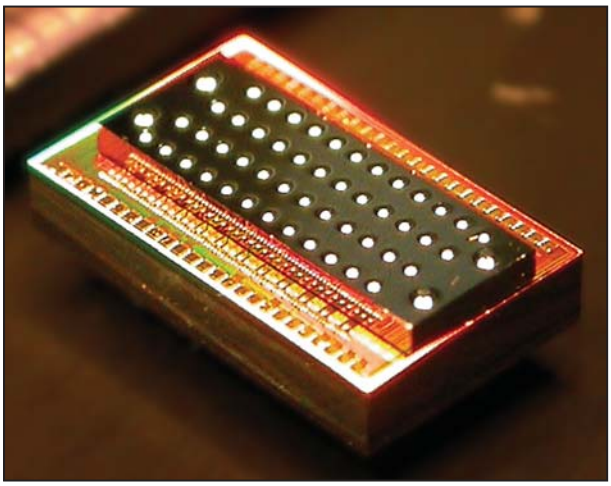


Figure 6

Flip-chip attachment of optical-to-electronic devices on ICs (48 VCSELs on a 48-channel driver IC, after [15]). Structure illustrated in Figure 7.

temperatures. A virtue of the flip-chip attach is that it can present less parasitic series inductance than wire bonding, which is helpful for a high-frequency interconnect. Flip-chip optical devices use their substrate as the optically active surface and thus operate at wavelengths much greater than 850 nm (e.g., 980 nm or 1,310 nm), where the substrate materials (GaAs and InP) are optically transparent.

It is also possible that parallel optics could evolve in more complex directions than simply increasing the channel bit rate and the width of the link. One of the most straightforward ideas is to use multiple wavelengths on a single fiber (wavelength division multiplexing, or WDM). A version of a parallel WDM transmitter and receiver has been demonstrated and is described in the next subsection [11, 16]. This system is considered "coarse" WDM (CWDM), since the wavelength spacing between channels is typically 10–15 nm, as opposed to the 0.5–1.0 nm channel spacing typical of the dense WDM (DWDM) systems used in long-haul telecommunications.

A more radical version of optical interconnects could use DWDM. At present, the costs of such systems appear to be prohibitive compared with those of the parallel optics technologies discussed above. However, since DWDM systems can multiplex many wavelengths into a single fiber, there is a tremendous possibility for a reduction in fiber count and cost of distributing the optical signals. Against this reduction in fiber count, however, is the fact that DWDM is inherently a single-mode system with comparatively expensive components. Today, each wavelength of a DWDM system typically consists of a distributed-feedback laser, which produces unmodulated light, and a separate modulator (sometimes on the same substrate). Each laser must be temperature-controlled to within ~1°C. The separate wavelengths are multiplexed, e.g., with an arrayed waveguide grating (AWG), again temperature-controlled. At the receive end, there is typically an additional AWG, and more single-mode optics to convey the signal to the PDs. At present, the cost of a DWDM link is ~100 times more expensive than a VCSEL/MMF data communication link on the basis of cost per Gb/s. This cost and the bulk of the optical transceivers in such systems must be overcome before DWDM can be considered for intrasystem interconnects; however, there are promising developments in this direction. One technology option is to use highly integrated "photonic integrated circuits (PICs)" with laser modulators, AWGs, waveguides, and other optical functions on relatively large InP substrates [17, 18]. If such an InP PIC can be produced reliably and at low cost, it could have a major impact on both long-haul WDM systems and optical interconnects. There is also significant ongoing work in silicon photonics, using a combination of III–V laser and III–V or Ge-based detectors and active optical circuits such as modulators and AWGs built using CMOS and silicon-on-insulator (SOI) technology [19, 20].

Ultimately, the decision on whether to use simple parallel optics or some form of WDM will revolve primarily around the cost of the various interconnects. The cost difference between CWDM and single-wavelength parallel modules depends on the tradeoff between the costs associated with the multiplex/demultiplex (MUX/DEMUX) and tighter laser wavelength control compared with handling a larger number of fibers. DWDM systems can use more wavelengths and even fewer fibers than CWDM. However, CWDM lasers do not require the precise temperature regulation of DWDM lasers, and inexpensive VCSELs may be used as the laser source. The enormous cost disparity of DWDM compared with single-wavelength or CWDM parallel solutions will have to be eliminated before it can be considered a serious contender for optical interconnects.

Wider VCSEL multimode links using CWDM

Reference [11] describes the demonstration of a 48-channel high-density optical transmitter and receiver pair, termed the Multiwavelength Assemblies for Ubiquitous Interconnects (MAUI) module. One feature of the MAUI modules is that it reduces the fiber count by a factor of 4 by using CWDM, as explained in the previous subsection. The target overall bandwidth is more than 10 Gb/s in each of 48 channels. This bit rate should match or exceed the clock speed of microprocessors and so be adequate

for the initial use of SMP bus optical interconnects in servers.

The four 1×12 VCSEL arrays used in MAUI are flip-chip-attached to the IC, as explained in the previous section. After attachment of the O/E array to the IC, the optical coupling is added as shown in **Figure 7(a)**. Since the fiber ferrule has spacing between the pins of 4.6 mm, the electronic/optical assembly is quite compact. The complete assembly can be wire-bond-attached to an electrical flex similar to that used for the SNAP12 modules. However, the higher electrical density makes it difficult to package the module with an electrical connector. Instead, contacts to the electrical flex are accomplished by a direct BGA solder attach. This allows the contact pitch to be as small as 0.5 mm, as opposed to 1.27 mm for an electrical connector such as the MEG-Array.

Since this module uses a four-wavelength CWDM design to reduce the fiber count, an optical MUX and DEMUX are needed. These components are based on sequential reflection from dichroic optical wavelength filters, with a ray-trace diagram as shown in **Figure 7(b)**. The loss associated with the wavelength MUX is about 5 dB, which, since there is a companion DEMUX at the receiver, is a substantial loss of optical power that must be overcome by higher laser powers or better receiver sensitivity.

Denser optical links

Properly packaged, the fiber-based parallel modules described above will probably be suitable for the initial use of optical interconnects in servers. However, as serial rates continue to increase beyond 10 Gb/s and the required density of interconnects on a card or a backplane continues to increase, it would be desirable to replace the on-card and backplane electrical interconnects with high-density, low-power optical interconnects. This replacement will require even lower optical interconnect cost and higher density and bit rate than those delivered by the modules described in the preceding subsection. In addition, management of the optical "wires" will likely be impractical if the "wires" continue to be optical fiber ribbons which are not integral to the circuit card.

As a result of these considerations, much research is currently focused on polymer optical waveguides, which have the possibility of being fabricated as part of the organic electrical circuit card. A number of challenges must be overcome in order to make these waveguides a practical technology, including the following:

- Optical loss of the waveguides (less than 0.1 dB/cm for ~1-m links).
- Stability of the waveguides during the board manufacturing process, chip attachment, and system operation (less than 250°C for manufacturing, 80°C during operation).
- Scalability of waveguide manufacturing to board dimensions (e.g., $\sim 60 \times 80$ cm).
- Low-loss optical interface between O/E module and waveguides, and between board and backplane (less than 2 dB per interface).

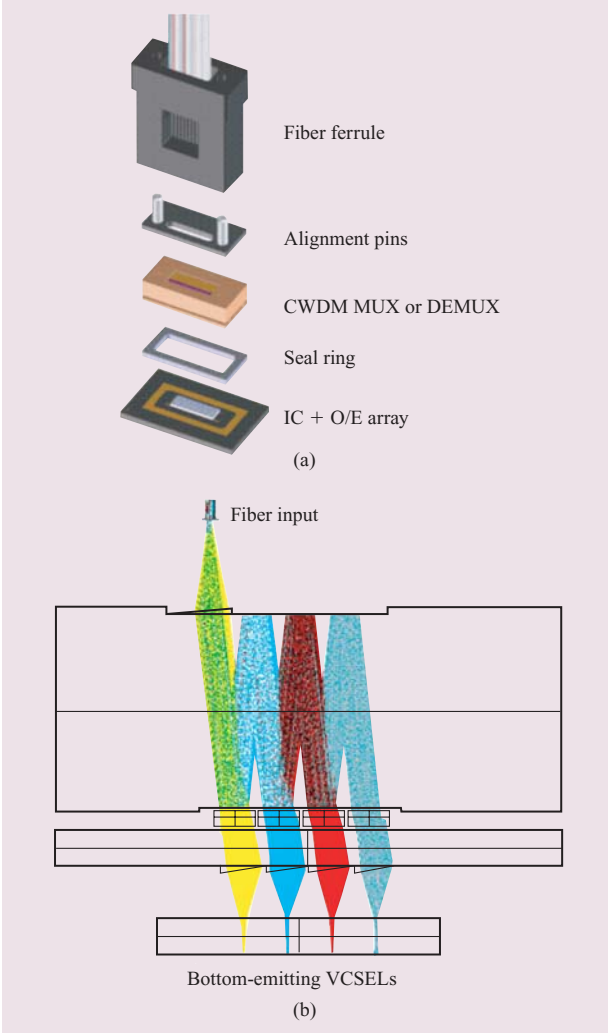


Figure 7

(a) MAUI module, showing (b) parallel wavelength division multiplexing optical signal paths.

The work on waveguide-based optical interconnects is at a preliminary stage; there is much additional work to be done in order to determine a complete commercial technology. For example, no consensus exists as to how to manufacture the waveguides, and whether they should be attached to the surface of the circuit card or laminated

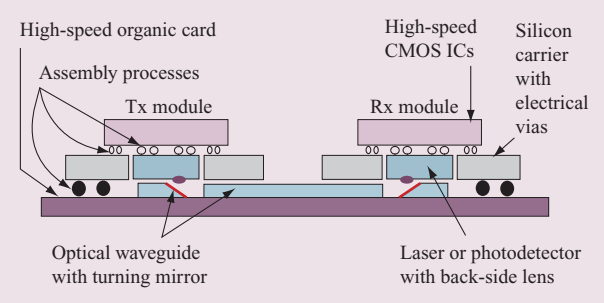


Figure 8

Cross section of Terabus module structure. (Tx: transmitter; Rx: receiver.)

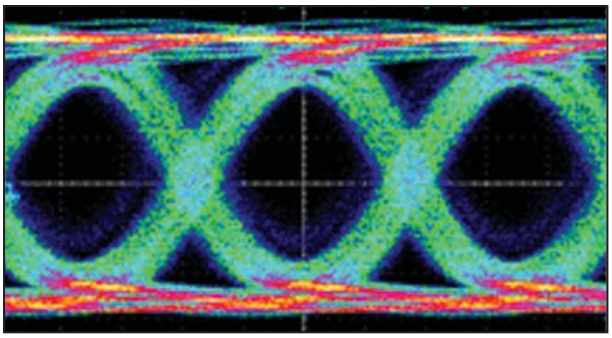


Figure 9

14-Gb/s electrical eye diagram at the output of a Terabus receiver. The optical input is from a Terabus transmitter.

inside. Waveguides could be fabricated as a single layer or in multiple layers. Such fundamental choices will influence other aspects of the overall technology, such as the optical coupling schemes that are possible. Similarly, appropriate board-to-backplane and board-to-fiber optical connectors are required if this sort of technology is to achieve acceptance, and high-density, low-loss connectors of this sort are at a very preliminary stage.

Nonetheless, there has been significant progress. For example, a number of low-loss acrylate and polysiloxane materials have been developed [21] which are photo-definable by lithography or direct laser writing and appear to have adequate thermal stability. The materials may be deposited by a spin-on process, spray-coating, or doctor-blading (spreading with a flat blade), and structures such as mirrors can be created by laser ablation. Both deposition on top of and lamination inside organic circuit cards have been demonstrated. The materials have also shown the ability to support high-bit-rate communications [22].

Efforts are currently underway to demonstrate complete operation of board-level optical waveguide interconnects. Such demonstrations will go a long way toward determining the appropriate packaging and materials choices for waveguide technology. One such recent effort is the Terabus program [15], which is a joint effort between the IBM Research Division and Agilent Laboratories. The overall approach of the program is to create a "chiplike" silicon-carrier-based optical module that is coupled electrically and optically to an organic circuit card, as illustrated in **Figure 8**.

Some of the most critical components for Terabus have been demonstrated. For example, 4×12 CMOS IC and O/E arrays have been fabricated and joined as shown in Figure 6, and a complete fiber-based link at 14 Gb/s has been demonstrated. An electrical eye diagram from a Terabus laser diode driver and VCSEL that is fiber-coupled to a PD and an amplifier is shown in **Figure 9**. The link ran at a bit error ratio of less than 1×10^{-12}. Analysis and simulation of the Terabus packaging shows that it will support future bit rates of at least 20 Gb/s.

5. Impact of optical technologies on server architectures

Optical links for SMP buses

Present SMP links and SMP expansion links between server nodes are served by dense, impedance-controlled connectors and backplanes or cables. Typical link distances are in the range of one meter of standard circuit board material and up to two meters of cable. Future higher-performance systems will require higher-bandwidth buses between processor complexes, or between system racks in the case of very large clusters.

SMP links require cost-competitive, high-bandwidth (more than 100 Gb/s), low-power, low-latency interfaces. In next-generation systems, each point-to-point link in an SMP system, whether cable or backplane, will have an aggregate, bidirectional throughput of 100 to 500 Gb/s. Using present electrical I/O technology would require hundreds of electrical signals and, for short links, several watts of power dissipation. For longer cable links, cable bulk and bend radius become a concern; some multiplexing of the data to higher rates is required in order to maintain bandwidth while maintaining a manageable cable. Electrical solutions for multigigabit cable transmission for distances greater than approximately 2 m would require two to three times the power required by shorter electrical links. This is the area of application in which parallel optical links could first offer advantages.

Optical technologies can be engineered to satisfy these SMP cable link requirements, but the needs of the entire system must be kept in mind. The majority of SMP

cable links will be in the range of 2 to 10 m, allowing technology tradeoffs not possible for longer links. System architects will prefer optical solutions which operate efficiently at a few speed multiples above the standard board-level electrical chip-to-chip communication rates, since custom chips and additional latency would be required to multiplex the data to much higher rates for optical transmission. These system requirements could be handled by two to four SNAP12-like modules operating in the range of 5 to 10 Gb/s per fiber. In addition, SMP systems will benefit from forward error-correction coding information transmitted along with the data to avoid the need for higher-latency retry protocols to recover from transmission errors. These requirements are different from current optical link standards, so it is likely that first implementers of optical SMP buses will use SNAP12 or other standards-driven physical layers in combination with custom bus conversion chips to condition, transmit, receive, and correct the data sent over parallel optical SMP links.

Servers with optical multidrop buses

As described in Section 2, the simplest and lowest-latency structure for connecting small numbers of nodes is a multidrop bus, where multiple transmitters may, after arbitration, take turns broadcasting packets simultaneously to multiple receivers over the same physical transmission medium. This structure is particularly useful for SMP buses, which use a "broadcast-and-snoop" protocol to maintain cache coherence. Such coherence protocols operate as follows. When multiple processors with per-processor local caches share access to the same memory, they may locally cache copies of the memory data, and must read and write data in a way that ensures that other processors can be sure of getting the most current version. For example, when a processor executes a "Load" instruction for a piece of data not in the processor cache, the cache controller can broadcast the corresponding bus operation to all of the cache controllers and memory controllers in the system. The other cache controllers then give a response, indicating whether they have a shared (read-only, unmodified from the copy in memory) or modified version of the relevant data. The responses from all of the cache controllers are aggregated, and, depending on the aggregated response, the data is brought into the requesting cache either from memory or from another cache, and the cache-line status in all processor caches is modified appropriately. Further details on simple broadcast-and-snoop cache coherence protocols can be found, e.g., in [23].

Broadcast-and-snoop coherence protocols are relatively simple to implement; they provide low-latency, high-performance cache coherence, but require that all operations which may affect cache state are broadcast to all cache controllers. Directory-based coherence protocols, which reduce the amount of bus traffic, do exist [24], but they require more complexity and more local state, which can reduce performance for moderate-scale systems.

With SMP buses at moderate bit rates (less than 500 to 800 MHz), the broadcast function can be implemented electrically at the physical layer, over bus lengths of a few tens of centimeters. For buses operating at higher speeds and longer distances, however, including backplane-length (20–1,000-cm) links, transmission-line effects such as reflections and matched impedance become important, and point-to-point links must be used. The logical broadcast function must therefore be implemented as a switched multicast function, with operations packets being replicated as they pass through multiple chips to connect the cache controllers.

In practice, splitting an electrical signal reliably across more than two receivers at Gb/s bit rates is a difficult proposition, especially where the physical environment includes significant attenuation and impedance discontinuities in connectors, vias, and chip module packages. However, splitting an optical signal to multiple receivers can be done without impedance discontinuities that affect the achievable bit rate. The primary limit to achievable bit rate comes from power limitations, as a signal split among N receivers provides $1/N$ or less of the power available to a single receiver.

If an optical broadcast bus could be physically achieved, there would be substantial advantages to system design vs. a multistage interconnect network. These advantages include the following items:

- Latency: No multistage transmission.
- Power: Each transmitter transmits to multiple receivers, replacing multiple retransmissions of each operation with data recovery and latching in each stage.
- Modularity: Adding more nodes on a network does not change the system timing.

There are some significant challenges that must be overcome to enable such a system to operate correctly. These challenges include many-to-one transmission, link power budget, and wide-bus transmission:

- *Many-to-one transmission*. The opposite of a broadcast bus (one-to-many, with one transmitter and many receivers) is a many-to-one bus. The primary challenges with many-to-one transmission are transmitter squelch and timing. The squelch issue arises from the fact that, once an arbitration cycle is

done, the winning transmitter should not be affected by crosstalk from the inactive transmitters. In normal point-to-point optical links, an optical "0" does not correspond to zero optical power—the on/off ratio in normal transmission is typically between 10:1 and 3:1, but may be even smaller. In a many-to-one scheme, however, the transmitters which are "off" must be truly off (squelched) in order not to interfere with the active transmitter. The timing issue arises from the fact that, in switching from one transmitter to a different transmitter over the same shared medium, there is generally some path length difference, which appears at the receiver as a phase difference. There must be some method for compensating for this phase difference, either at the transmitter or at the receiver.

- *Link power budget*. For a one-to-many broadcast bus to operate correctly, the received signal at each receiver, after signal splitting, must be strong enough to allow reliable data reception. For a 1-to-8 broadcast bus, for example, the power budget must allow an extra margin of more than 9 dB for splitting losses. To some extent, this link margin can be compensated for with a lower bit rate (optical transceiver circuitry capable of 10–20-Gb/s transmission can provide extra margin when operated at 2.5–5 Gb/s, which is a more typical speed for SMP buses for processor frequencies of less than 5GHz). However, since reduced link margin affects error ratio, this feature requires detailed engineering analysis to determine performance and cost tradeoffs.
- *Wide bus transmission*. To be useful for an SMP bus, the optical broadcast link must transmit data at aggregate link speeds well above 100 Gb/s. Typically, SMP buses operate with 8-byte or 16-byte widths, which, with overhead, are in the 72-bit to 144-bit range. Splitting a serial line to, e.g., eight receivers can be accomplished with a fiber coupler, but splitting 72 or 144 lines operating in synchrony requires a more practical implementation than 72 or 144 separate individual fiber couplers.

In summary, multidrop optical buses offer great promise for implementation of broadcast-and-snoop low-latency SMP buses as bus frequencies exceed 1 Gb/s. However, there are significant challenges in research and engineering design to be overcome before such structures can be practically considered for commercial product design.

Modular servers

As described in Section 3, there is significant difference between the packaging of mid-range servers (4–16 processors), which fit in shelves, and high-end servers, which use midplanes or backplanes in a packaging suited to rack-sized systems. Significant costs are associated with delivering both types of system, since different boards, components, and cooling structures must be used. Currently, however, there is little alternative, since the bandwidth and reliability required for high-end SMP links cannot be supported across cables, necessitating backplane-style interconnects to construct high-end SMP systems. The commercial systems that are built using aggregations of mid-range server building blocks with electrical interconnects (e.g., [25], [26], or [27]) depend heavily on locality-aware operating systems, hardware, and system software to minimize the traffic over electrical SMP expansion links.

With widely parallel optical links of the type described in the subsection on denser optical links, an optical cable can support enough bandwidth to construct SMP systems with better bandwidth uniformity. Such widely parallel optical links can support the required bandwidth at sufficiently low power only if the optical transceivers are sufficiently close to the processor chips.

One factor that fundamentally affects SMP bus design is bit error ratio. To date, the error detection and correction (EDC) strategies for SMP buses and for optical cables have been substantially different, since SMP buses have used hardware EDC, whereas LAN/SAN and cluster links (and longer links) have allowed software error correction. An uncorrected SMP bus bit error may cause tremendous problems, since there is no mechanism for software correction. There are fundamental physical limits that characterize errors in optical and electrical links. The theoretical limit to the noise floor of an optical link includes both shot noise of the photons and thermal noise in the receiver. For an electrical receiver, the noise floor is only thermal. Both electrical and optical links have a similar thermal noise limit. Because optical receivers are seldom operated substantially below room temperature for communications, at the typical photon fluxes that are used (tens to hundreds of microwatts of optical power), the noise tends to be dominated by thermal noise in the receiver, and there is no substantial difference between electrical and optical links with respect to bit error ratio. For cost and pragmatic reasons (e.g., test time), optical modules are seldom tested to better than 1×10^{-12} BER. However, since these errors have fairly predictable characteristics in properly designed transceivers, they can be compensated for with error-correction coding of appropriate strength, at nominal extra delay.

If such SMP links could be built with adequate bandwidth, reliability, power efficiency, and latency penalty, one could imagine constructing high-end SMP machines from mid-range machines, or even from moderately powerful blades, with much less

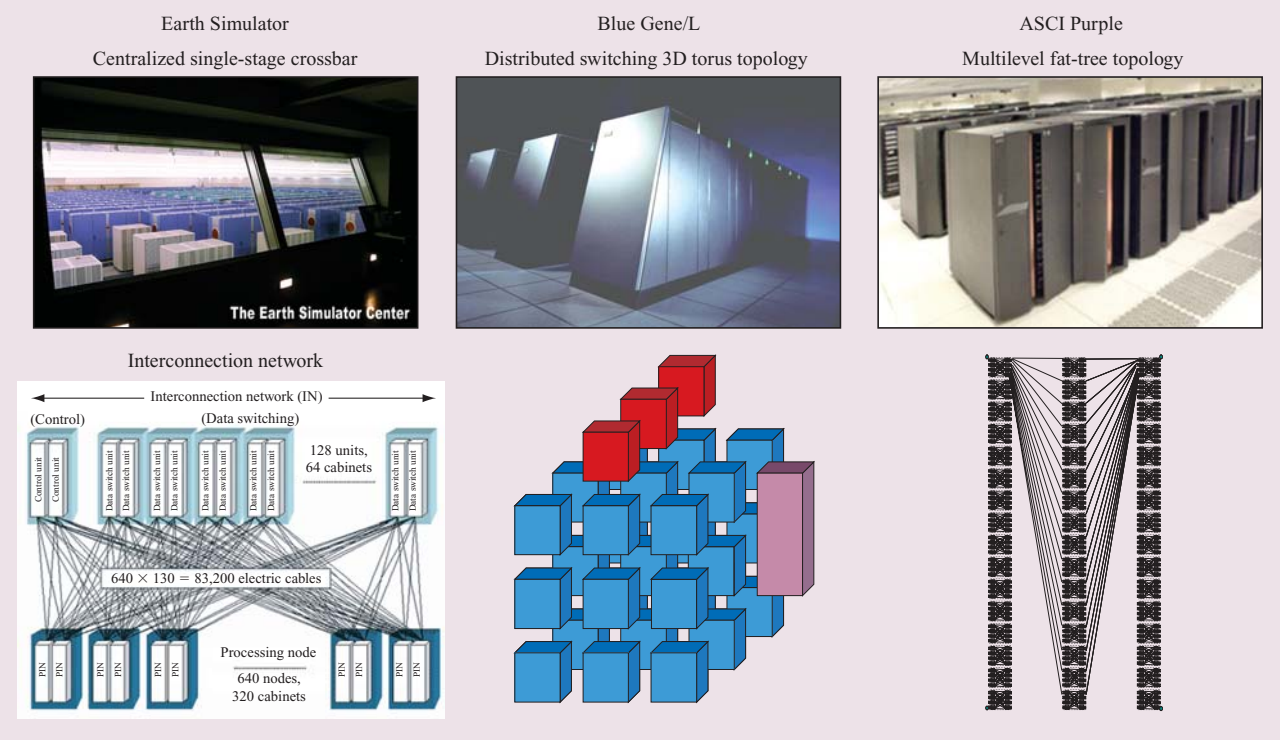


Figure 10

Clusters and massively parallel machines: Earth Simulator, Blue Gene/L, and ASCI Purple.

nonuniformity than is practical with electronic links. It would be only moderately more difficult to imagine incorporating configurability in these links so that they could be used either for maintaining cache coherence in an SMP bus or for message-passing of cluster traffic. Customers could combine operating system software and system management software to configure the same processors either as large-scale cache-coherent shared-memory processing machines or as distributed-memory machines with separate independent caches, depending on application efficiency requirements and time of day. This capability would be extremely valuable for a wide range of large server customers.

Impact of optical technologies on clusters

The earliest and most profound impact of new optical transmission technologies in the near future may be in the design and construction of large cluster machines and cluster switches. The largest and fastest machines in the world [28] are constructed with thousands of processors, connected through tightly integrated cluster networks fabricated from switching elements and point-to-point links.

Consideration of a few of the fastest supercomputers shows the variation in possible system designs for these machines. **Figure 10** shows the three representative machines considered here: the Earth Simulator system, Blue Gene/L, and ASCI Purple.

Earth Simulator

The Earth Simulator, completed by NEC Corporation in 2001, was the world's fastest machine for several years. The topology of the Earth Simulator is somewhat unusual for such a large system in that it uses a single-stage centralized crossbar switch, with 640 ports operating at 1.25 Gb/s, to interconnect 640 compute nodes. To provide adequate bandwidth for each of the compute nodes, 128 copies of this central switch are provisioned, with traffic flowing in parallel through all of the switches simultaneously. The Earth Simulator uses no optical cables: All of the 83,000 separate cables are electrical.

In some sense, this structure is an ideal topology for a supercomputer. Every compute node is connected to every other compute node by a single switch hop,[4] and each node is equidistant from every other compute node. Since all of the switching is centralized, there are no switch-to-switch links, with their consequent extra cost in

[4] A *switch hop* is a traversal through a single switching element.

cables, connectors, and transceivers. However, the large number of individual cables per compute node (128 per node), with serial transmission on each, is less cost-effective and space-efficient than parallel links, where the overhead of a link can be amortized over all of the lines in the cable. Also, since every cable extends from the central array of switches to the compute nodes, a large proportion of the cables must be relatively long, in comparison with links used for other topologies.

Blue Gene/L

The Blue Gene/L machine, constructed by IBM, has been the world's fastest machine since September of 2004. It uses a combination of interconnect networks, with the primary network being a 3D torus interconnecting more than 64,000 processors in a direct network. The overriding characteristic of Blue Gene/L is the system-on-a-chip (SOC) design of the computer/switching nodes, in which a single chip provides two processing cores, on-chip high-bandwidth memory, a memory controller interface, and multiple network link ports, with port-to-port switch capability, on the same chip.

From a packaging perspective, the Blue Gene/L design point excels in system density, since the integration of switching and processing on the chip and the direct connection of node chips in 3D torus and tree topologies allow dense packing of nodes without separate switch boxes. In the 3D torus topology, the vast majority of links are extremely short—a few inches or less—and can be implemented with on-card interconnect. Since on-card interconnect links are cheaper and require less power than cable links, the overall machine can be very power-efficient and relatively inexpensive. The 3D torus topology is non-optimal for several reasons (the farthest neighbors are up to 64 switch hops farther away than the nearest neighbors, and the bisection bandwidth of the torus topology is quite low compared with the individual node link bandwidth), but the short distance of the majority of links allows them to be built with quite high bandwidth relative to the processing capability of the node. Blue Gene/L also uses no optical cables: The vast majority of links in the machine are short, on-card links, and the short cables between racks use electrical signaling, driven by separate re-drive chips.

ASCI Purple

ASCI Purple represents something of a more commercial mid-point design between the Earth Simulator and Blue Gene/L in terms of performance and topology. The ASCI Purple design makes use of standard POWER5*-based SMP nodes, connected through a very-high-performance cluster network with (20+20)-Gb/s links called the High Performance Switch (HPS) [7]. The network for this cluster uses 32-port switch shelves, with ports that may be connected either to compute nodes or to other switch ports. The 32-port switch shelf allows the construction of clusters with single-cable connection of a moderate number of nodes (a 16-node cluster of 64-way POWER5 SMP nodes provides a very substantial computing resource), and a multistage interconnection network topology allows the same switches to support installations with more than 10,000 processors and more than 100 teraflops per second of aggregate computing power.

One substantial characteristic of the HPS network used for ASCI Purple is that the bandwidth of an individual point-to-point link is roughly an order of magnitude higher than the individual link bandwidth in either the Earth Simulator or Blue Gene/L networks. This allows one or a few cluster links to supply a large set of high-end processors, where more than a half dozen Blue Gene/L links are used to supply two processor cores, and the Earth Simulator compute node is supplied by 128 separate links.

Another substantial difference between the pSeries HPS and the networks in the other two systems is that the HPS switches allow configurability of the links to use either copper cables or optical cables (using the SNAP12 transceivers described in Section 4). The less expensive copper cables are used for short links (up to 5 or 10 m), and optical cables are used for the longer links—up to 40 m—that are necessary to build very large systems.

Future systems

As we look to the next generation of ultrascale systems and examine the impact of the new optical interconnect technologies on them, it is clear that optical interconnect can have substantial impact on the construction of high-end cluster switches and topologies. As link bit rates grow to 10 Gb/s per line and higher, it is clear that electrical transmission will support more limited distances. Current estimates are to limits of less than 20 inches on circuit board and less than 10 meters in cable, without very power-intensive signal-processing circuitry. At the same time, optical links operating at 10 Gb/s have been clearly shown to operate at lengths far greater than the 40 meters required to interconnect a large machine room.

These basic physical characteristics point toward the use of topologies in which electrical transmission is used for short-distance links (e.g., inside-the-box, on-card, and backplane links), and optical transmission is used for links which require going to cable. The other characteristic pointing toward the increased use of optical cable instead of copper cable is the relatively large size of high-bit-rate copper cables and connectors, which increases system weight and decreases potential for cooling airflow. The network topology which most closely matches these technology limitations is a structure with short-distance links (3D torus or hypercube topologies)

inside a box, and optical links of essentially unlimited distance connecting boxes to one another or to centralized switches of very high port count, using multistage interconnection network topologies. The bandwidth of these links can be made very high, using widely parallel on-card links and, e.g., 4×12 optical links, and the bandwidth of the links can be kept constant, whether the links stay on-card or exit through optical cables. This combination of torus or hypercube topologies for inside-the-box networks and multistage topologies for box-to-box cluster interconnects would provide a very attractive structure with the advantages of Blue Gene for inside-the-box density, and of Earth Simulator for global flatness and delay equality across very large systems.

6. Summary

Aggregate interconnect bandwidth, measured per chip, per board, or per rack, is steadily rising. To provide the necessary bandwidth and bandwidth density, optical interconnects are being used in applications which draw closer and closer to the processor chip. On the basis of server requirements for interconnect density and bandwidth–distance product, significant use of optical interconnects inside servers can be expected within the next several years. The most immediately promising type of optical technology is highly parallel, short-wavelength vertical-cavity surface-emitting laser arrays coupled into multimode optical ribbon fiber links. Other optical technologies, including coarse wavelength division multiplexing, polymer waveguide, long wavelength, and single-mode components, may come to play a role in future server interconnects. System and technology evolution will determine which of these optical technologies will be selected to replace electrical links in future systems.

Several challenges must be addressed before optical link technologies gain more widespread use. First, the cost is still too high for configurations that support tens to hundreds of Gb/s in a single link. Research projects in tightly integrated transceivers and widely parallel links are addressing this issue. Another significant challenge is in density, since the optical transceivers should be close enough to the processor chips to simplify the intermediate electrical link. Other issues include the latency in data coding for error detection and correction, thermal issues, and the need for voltages higher than those required for CMOS processors or memory. These and other system packaging issues must be addressed in a holistic manner in order to minimize the cost of designing optical links to a system.

Major aspects of server architecture can change if optical links become available that meet practical implementation challenges. Three major modifications in system packaging and topology can be expected.

The most speculative is the optical multidrop bus. There are significant advantages of multidrop buses in servers, since there are a number of functions (e.g., cache coherence, memory access across large numbers of memory chips, scalable I/O hierarchies) that require distribution of operations across multiple units. Currently, electrical links are undergoing a massive migration from electrical multidrop buses to packet-switched structures with point-to-point electrical links and chips that do switching and replication of operations. An alternative migration to optical multidrop structures should be physically realizable and would improve power, latency, and bandwidth. The more pragmatic limitations of signal splitting across wide buses, signal combination, and link margins for multidrop buses must be addressed before such architectures can be practically considered by system designers.

For clustered systems, the increasing use of optical technologies may affect the topologies used for interconnecting systems. There may be less use of torus and fat-tree topologies built from low-port-count switch units, and more use of centralized high-port-count switches with high bandwidth–distance optical links to the nodes, improving global connectivity and cluster network cost.

If optical link technology is adopted in more server interconnects, system packaging issues will be modified. There will be reduced need for high-performance electrical backplanes, and increased use of drawers or blades that provide better cooling and power delivery, with simpler board cross sections. High-end machines can be increasingly modular, with connected mid-range machines.

To summarize, there is a convincing case for the use of optical interconnects within SMP servers, particularly as a replacement for electrical cabling. Today, optics is heavily used in data communications applications. It must be recognized, however, that it has yet to achieve widespread use inside servers. The bandwidth and density of parallel optics have been shown to exceed those available with copper interconnects. If optics can be packaged close to the processors, it will allow more interconnect bandwidth between processors at lower total power consumption than electrical links. There is no technological barrier to optical interconnects; the final step required for its widespread adoption is the introduction of commercial optical modules at costs competitive with those for copper links. Once the use of optical interconnects becomes common, costs will drop further, resulting in still wider use of optics. As new technologies such as waveguides become feasible, even backplane- and board-level interconnects can include

optical interconnects. At that point, optics will have migrated from a straightforward replacement for copper cables to a technology that enables new and better server architectures.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of PCI-SIG Corporation, InfiniBand Trade Association, FCI Americas Technology, Inc., or Infineon Company.

References

1. International Technology Roadmap for Semiconductors (ITRS), Assembly and Packaging Chapter, Semiconductor Industry Association, 2003.
2. David W. Dolfi, "Multi-Channel Optical Interconnects for Short-Reach Applications," presented at the 53rd Electronic Components and Technology Conference, 2003.
3. A. F. J. Levi, "Optical Interconnects in Systems," *Proc. IEEE* **88**, No. 6, 750–757 (2000).
4. J. U. Knickerbocker, P. S. Andry, L. P. Buchwalter, A. Deutsch, R. R. Horton, K. A. Jenkins, Y. H. Kwark, G. McVicker, C. S. Patel, R. J. Polastre, C. D. Schuster, A. Sharma, S. M. Sri-Jayantha, C. W. Surovic, C. K. Tsang, B. C. Webb, S. L. Wright, S. R. McKnight, E. J. Sprogis, and B. Dang, "Development of Next-Generation System-on-Package (SOP) Technology Based on Silicon Carriers with Fine-Pitch Chip Interconnection," *IBM J. Res. & Dev.* **49**, No. 4/5, 725–753 (2005, this issue).
5. University of New Hampshire Interoperability Laboratory 10 Gigabit Ethernet Knowledgebase; see *http://www.iol.unh.edu/consortiums/10gec/knowledgebase*.
6. H. Iga and K. Li, Eds., *Vertical-Cavity Surface-Emitting Laser Devices*, first edition, Springer-Verlag, Berlin, 2002.
7. O. Lascu, Z. Borgosz, J.-D. S. Davis, P. Pereira, and A. Socoliuc, "An Introduction to the New IBM eServer pSeries High Performance Switch," IBM Redbooks, 2003; see *http://www.redbooks.ibm.com/redbooks/pdfs/sg246978.pdf*.
8. InfiniBand Architecture Specification, Volume 2; see the InfiniBand Trade Association website, *http://www.infinibandta.org*.
9. P. Pepeljugoski, M. Hackert, J. S. Abbott, S. Swanson, S. Golowich, J. Ritger, P. Kolesar, C. Chen, and J. Schlager, "Development of Laser Optimized 50 μm Multimode Fiber for Multi-Gigabit Short Wavelength LANs," *IEEE J. Lightwave Technol.* **21**, 1256–1275 (2003).
10. L. A. Buckman Windover, J. N. Simon, S. A. Rosenau, K. Giboney, G. M. Flower, L. W. Mirkarimi, A. Grot, B. Law, C.-K. Lin, A. Tandon, R. W. Gruhlke, G. Rankin, and D. W. Dolfi, "Parallel-Optical Interconnects Beyond 100Gb/s," *IEEE J. Lightwave Technol.* **22**, 2055–2063 (2004).
11. B. E. Lemoff, M. E. Ali, G. Panotopoulos, G. M. Flower, B. Madhavan, A. F. J. Levi, and D. W. Dolfi, "MAUI: Enabling Fiber-to-the-Processor with Parallel Multiwavelength Optical Interconnects," *IEEE J. Lightwave Technol.* **22**, 2043–2054 (2004).
12. D. Kuchta, Y. Kwark, C. Schuster, C. Baks, C. Haymes, J. Schaub, P. Pepeljugoski, L. Shan, R. John, D. Kucharski, D. Rogers, M. Ritter, J. Jewell, L. Graham, K. Schrodinger, A. Schild, and H.-M. Rein, "120-Gb/s VCSEL-Based Parallel-Optical Interconnect and Custom 120-Gb/s Testing Station," *IEEE J. Lightwave Technol.* **22**, 2200–2212 (2004).
13. See *http://www.tycoelectronics.com/fiberoptics/light/Create_Light_0904/High_Speed_PARA-OPTIX_Transceivers_092004.pdf*.
14. A. V. Krishnamoorthy, L. M. F. Chirovsky, W. S. Hobson, R. E. Leibengath, S. P. Hui, G. J. Zydzik, K. W. Goossen, J. D. Wynn, B. J. Tseng, J. Lopata, J. A. Walker, J. E. Cunningham, and L. A. D'Asaro, "Vertical-Cavity Surface-Emitting Lasers Flip-Chip Bonded to Gigabit-per-Second CMOS Circuits," *IEEE Photonics Technol. Lett.* **11**, 128–130 (1999).
15. D. Kucharski, Y. Kwark, D. Kuchta, K. Kornegay, M. Tan, C.-K. Lynn, and A. Tandon, "A 20Gb/s VCSEL Driver with Pre-Emphasis and Regulated Output Impedance in 0.13μm CMOS," *International Solid-State Circuits Conference Digest of Technical Papers*, ISSCC 2005, February 2005, pp. 222–223.
16. L. B. Aronson, B. E. Lemoff, L. A. Buckman, and D. W. Dolfi, "Low-Cost Multimode WDM for the Local Area Networks up to 10 Gb/s," *IEEE Photonics Technol. Lett.* **10**, 1489–1491 (1998).
17. See *http://www.infinera.com/solutions/technology.html*.
18. See *http://www.35ph.com/download/090401_ThreeFive_Whitepaper.pdf*.
19. DARPA Electronic and Photonic Integrated Circuit (EPIC) Program solicitation; see *http://www.darpa.gov/mto/solicitations/BAA04-15/S/index.html*.
20. See *http://www.intel.com/technology/itj/2004/volume08issue02/art06_siliconphoto/p01_abstract.htm*.
21. R. Dangel, U. Bapst, C. Berger, R. Beyeler, L. Dellmann, F. Horst, B. J. Offrein, and G.-L. Bona, "Development of a Low-Cost Low-Loss Polymer Waveguide Technology for Parallel Optical Interconnect Applications," *Proceedings of the IEEE–LEOS Summer Topical Meetings*, 2004, pp. 29–30.
22. F. E. Doany, P. K. Pepeljugoski, A. C. Lehman, J. A. Kash, and R. Dangel, "Measurement of Optical Dispersion in Multimode Polymer Waveguides," *Proceedings of the IEEE–LEOS Summer Topical Meetings*, 2004, pp. 31–32.
23. T. Shanley, *PowerPC System Architecture*, Addison-Wesley Publishing Co., Inc., Reading, MA, 1995.
24. D. Lenoski and W.-D. Weber, *Scalable Shared-Memory Multiprocessing*, Morgan Kaufmann Publishers, Inc., San Francisco, 1995.
25. See *http://www-1.ibm.com/servers/eserver/xseries/xarchitecture/enterprise/index.html*.
26. See *http://www.newisys.com/products/extendiscale_whitepaper.pdf*.
27. See *http://www.sgi.com/products/servers/origin/3000/overview.html*.
28. See *http://www.top500.org*.

Received October 3, 2004; accepted for publication March 9, 2005; Internet publication September 14, 2005

Alan F. Benner *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (bennera@us.ibm.com).* Dr. Benner is a Senior Technical Staff Member in the IBM Systems and Technology Group, working on system architecture, design, and development of optical and electronic networks for high-performance servers and parallel systems. He received a B.S. degree in physics in 1986 from Harvey Mudd College and joined AT&T Bell Laboratories, doing photonic networks and components research until 1988. He received M.S. and Ph.D. degrees from the University of Colorado at Boulder in 1990 and 1992, researching nonlinear interactions between wavelength-multiplexed optical fiber solitons in the Optoelectronic Computing Systems Center. In 1992, he joined the newly created Power Parallel Supercomputing Systems Laboratory, which developed RS/6000* SP* systems; since then he has been doing research and development on various aspects of high-performance server networks at all levels of the network stack. Dr. Benner has written two editions of a book on Fibre Channel and has co-authored several of the specifications for the InfiniBand architecture. He has more than 15 technical publications and 11 issued patents in the U.S. and other countries.

Michael Ignatowski *IBM Systems and Technology Group, 2455 South Road, Poughkeepsie, New York 12601 (ignatow@us.ibm.com).* After receiving a B.S. degree in physics from Michigan State University in 1979 and an M.S. degree in computer engineering from the University of Michigan in 1982, Mr. Ignatowski joined IBM in Poughkeepsie, New York to work on performance analysis for S/390* servers. In 1991 he received an IBM Outstanding Innovation Award for contributions to the ES/9000* storage hierarchy and MP design. He is the coauthor of eight patents in the area of storage hierarchy and MP design. Mr. Ignatowski is currently a Senior Technical Staff Member working on future server technology and designs.

Jeffrey A. Kash *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (jeffkash@us.ibm.com).* Dr. Kash is a Research Staff Member at the IBM Thomas J. Watson Research Center and manages the Optical Link and Systems Design Group. He joined IBM in 1981 after receiving a Ph.D. degree in physics from the University of California at Berkeley. His group develops and explores advanced multimode optical links and link components for use in the servers of the future. In past work, Dr. Kash has investigated the optical properties of hot electrons in semiconductors, including the development of picosecond imaging circuit analysis, which is used by major semiconductor manufacturers today to debug CMOS ICs. Dr. Kash is a Fellow of both the American Physical Society and the IEEE.

Daniel M. Kuchta *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (kuchta@us.ibm.com).* Dr. Kuchta is a Research Staff Member in the Communication Technology Department at the IBM Thomas J. Watson Research Center. He received B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from the University of California at Berkeley in 1986, 1988, and 1992, respectively. He subsequently joined IBM at the Thomas J. Watson Research Center, where he has worked on high-speed VCSEL characterization, multimode fiber links, and parallel fiber optic link research. Dr. Kuchta is an author or coauthor of nine patents and more than 25 technical papers; he is a Senior Member of the Institute of Electrical and Electronics Engineers.

Mark B. Ritter *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (mritter@us.ibm.com).* Dr. Ritter received a B.S. degree in physics from Montana State University in 1981 and M.S., M.Phil. and Ph.D. degrees from Yale University in 1987. In the fall of 1986, Dr. Ritter took a postdoctoral research position in the Physical Sciences Department of the IBM Thomas J. Watson Research Center in Yorktown Heights, New York. His postdoctoral work with Dr. David Awschalom centered on nonlinear optical properties of fluid-filled porous media and carrier dynamics in diluted magnetic superlattice structures. Dr. Ritter became an IBM Research Staff Member in 1989. Since 1993 he has been manager of a group designing analog and digital circuits for data communications. His group has contributed to the Fibre Channel standard for storage-area networks and to the 802.3ae 10 Gb/s Ethernet standard, as well as to the IR wireless (IrDA) standard. The group's work has also resulted in a number of IBM products, including Fibre Channel transceivers, IrDA transceivers, and, most recently, 10-Gb/s Ethernet transceivers, where his group contributed to the design of the analog front-end circuits at 10 Gb/s. He and his group are exploring silicon–germanium and CMOS circuits for communications links at 40 Gb/s and beyond, as well as silicon circuits and technologies suitable for future terabit communications links within switches and servers. Dr. Ritter received the 1982 American Physical Society Apker Award, three IBM Outstanding Innovation Awards, and several Research Division and Technical Group Awards. He is an author or coauthor of 33 technical publications and holds 16 U.S. patents.