MAUI: Enabling Fiber-to-the-Processor With Parallel Multiwavelength Optical Interconnects

Brian E. Lemoff, *Senior Member, IEEE, Member, OSA*, Mohammed E. Ali, *Member, IEEE*, George Panotopoulos, Graham M. Flower, *Member, IEEE*, B. Madhavan, *Member, IEEE*, A. F. J. Levi, and David W. Dolfi, *Senior Member, IEEE, Member, OSA*

Invited Paper

***Abstract*—The Multiwavelength Assemblies for Ubiquitous Interconnects (MAUI) program is a collaboration between Agilent Laboratories, Palo Alto, CA, and the University of Southern California, Los Angeles, with the goal of enabling fiber-to-the-processor by developing very-high-density optical interconnects and complementary metal–oxide–semiconductor (CMOS) interface electronics. This paper focuses on the parallel wavelength-division-multiplexed optical interconnect technologies and their potential impact on computer systems.**

***Index Terms*—Computer architecture, multiprocessor interconnection, optical fiber applications, optical fiber communication, optical interconnections, optical receivers, optical transmitters, wavelength-division multiplexing (WDM).**

I. Introduction

BY THE end of this decade, optical interconnects will be expected to carry much of the board-to-board and processor-to-processor bandwidth burden in high-end computer systems. As processors move to higher pin counts and higher bus speeds, the purely electrical backplanes used in today's high-end systems will not be able to provide the necessary distance, density, and power dissipation. While optical interconnects offer a potential solution, the technologies available today, which are tailored to the needs of the networking industry, are far too large, costly, and power-hungry to satisfy the backplane needs of computer manufacturers. New optical technologies that optimize bandwidth per unit area, cost, and power dissipation must be developed. The Multiwavelength Assemblies for Ubiquitous Interconnects (MAUI) program is a Defense Advanced Research Projects Agency (DARPA)-funded collaboration between Agilent Laboratories, Palo Alto, CA, and the University of Southern California (USC), Los Angeles, that aims to develop many of the technologies needed to enable fiber-to-the-processor (FTTP).

Manuscript received December 31, 2003; revised June 8, 2004. This work was supported in part by the U.S. Defense Advanced Research Projects Agency (DARPA).

B. E. Lemoff, M. E. Ali, G. Panotopoulos, G. M. Flower, and D. W. Dolfi are with Agilent Laboratories, Palo Alto, CA 94304 USA (e-mail: brian_lemoff@agilent.com).

B. Madhavan and A. F. J. Levi are with the Department of Electrical Engineering at the University of Southern California, Los Angeles, CA 90089-2533 USA (e-mail: alevi@usc.edu).

Digital Object Identifier 10.1109/JLT.2004.833251

This paper begins with a short overview of the MAUI program and its motivation, followed by a more detailed discussion of those limitations of electrical interconnects that are leading system architects to begin to consider optical alternatives. The parallel wavelength-division-multiplexing (PWDM) approach, being developed at Agilent Laboratories, is then introduced, including a description of the parallel multiwavelength optical subassembly (PMOSA) and comparison with other mutlichannel fiber-optic technologies. The paper then includes a discussion of PWDM system optimization, followed by details of the integrated circuit and PWDM optical multiplexer (mux) and demultiplexer (demux) technologies. Following this is a discussion of a possible MAUI-enabled computer architecture being studied at USC called the *encapsulated processor*. The final section of the paper is devoted to manufacturability and cost issues, which will ultimately determine the success or failure of the technology.

II. Program Overview and Motivation

Today's large-scale symmetric multiprocessor (SMP) computer systems include hundreds of CPUs and terabytes of memory and integrated storage devices, requiring board-to-board bandwidths approaching 1 Tb/s [1]. Without major changes in system architecture, server interconnect systems could be required to carry as much as 40 Tb/s between processor boards by the end of this decade [2, pp. 3–4]. Limitations in printed circuit board and electrical connector technologies make it unlikely that a purely electrical solution could provide the necessary bandwidth density to accommodate this need. Radical changes in architecture, such as the use of directories, novel coherence protocols, much larger caches, application parallelization, and clustering have the potential to reduce the required bandwidth significantly but would be unnecessary if a suitable optical approach were available.

While it is difficult to specify exact requirements that optical solutions will need to meet in order to be suitable for computer backplanes, some goals often mentioned by computer architects are linear bandwidth density (i.e., the duplex bandwidth per unit length along the edge of a circuit board, where it connects to the backplane) in excess of 100 Gb/s/cm, areal bandwdth density (i.e., duplex bandwidth per unit occupied board area) in excess

of 100 Gb/s/cm^2, duplex bandwidth per unit power consumption in excess of 100 Gb/s/W, and a price target of less than \$1/Gb/s for a transmitter/receiver (Tx/Rx) pair, including the required cabling. Today's commercially available optical link technologies fall short of each of these goals by one to two orders of magnitude.

As part of the MAUI program, Agilent Laboratories is developing optical interconnect technology that will optimize bandwidth density, power consumption, and cost through the use of PWDM, which combines parallel optics [3]–[5] with integrated coarse wavelength-division multiplexing (CWDM) [6]–[8]. The PMOSA that Agilent Laboratories is developing will transmit or receive hundreds of gigabits per second through a multimode fiber ribbon, with each fiber carrying multiple wavelengths. The PMOSA is a self-contained assembly consisting of an integrated circuit, optoelectronics, a PWDM optical mux or demux, and mechanical features for aligning to a fiber-ribbon ferrule, which will have a compact footprint (< 50 mm^2) and low-power consumption (<1 W). Particular attention is being paid to developing technologies that will ultimately minimize component and manufacturing cost. As will be discussed in Section III, the cost and performance tradeoffs between a PWDM approach and a single-wavelength two-dimensional (2-D) parallel approach will also be studied.

The University of Southern California is investigating the system impact of FTTP and is developing complementary metal–oxide–semiconductor (CMOS) circuitry that could provide a suitable interface between the PMOSA and a processor or application-specific integrated circuit (ASIC). Experiments that combine the Agilent PMOSA with the USC interface electronics are planned.

III. Limitations of Electrical Backplanes

Electrical backplane interconnects have a number of performance limitations, particularly at bit-rate-distance products in excess of 5 Gb/s-cm, where a controlled impedance becomes necessary on electrical lines. This impedance is necessarily low (25–75 Ω) due to power consumption and line density considerations [9]. Frequency-dependent line loss, both in the conductor and the dielectric substrate, is the primary factor in determining the line-rate-distance product of electrical interconnects. Newer dielectric materials provide lower loss than a traditional FR-4 PC board, but at longer line lengths, the conductor loss alone becomes significant. This is illustrated in Fig. 1, which uses a simple loss model to calculate the maximum bit rate for a given line length. Briefly, the model uses simple analytic expressions for the conductor and dielectric losses [10] for microstrip at a given line length and combines them to yield the frequency response of the line. This response is then combined with an input consisting of the square pulse of a given bit rate (corresponding to a single "1" bit) to determine the net response at the center of the eye, from which eye closure is then determined. For this calculation, the maximum allowable bit rate is chosen to be that which results in a certain percentage of eye closure at the given line length.

Curves are shown for microstrip on FR-4 for two different choices (50% and 75%) of eye closure to indicate how the bit rate improves as more eye closure is allowed. In addition, the 75% calculation is also shown in the case of a lossless dielectric to indicate the effect of conductor loss alone. As can be seen, performance is improved as dielectric loss is removed, but even the conductor loss alone limits the length to less than 0.7 m at 5 Gb/s and less than 0.5 m at 10 Gb/s. By comparison, losses in optical fiber at these distances are completely negligible. This calculation assumes a line impedance of 50 Ω, copper conductors, a line trace width of 101.6 μm (4 mil), and a substrate relative dielectric constant and loss tangent of 4.56 and 0.02, respectively.

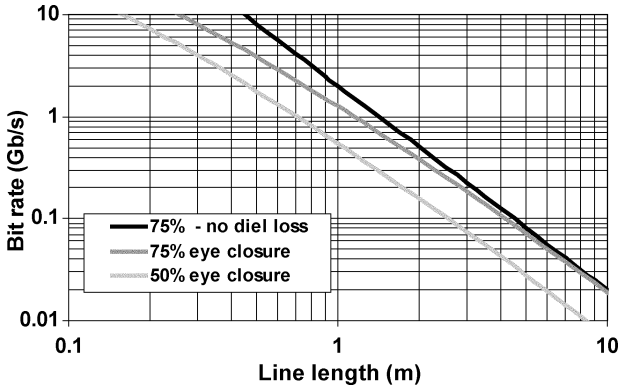


Fig. 1. Maximum bit rate versus line length for a microstrip on an FR-4 board.

Several investigators have recently shown [11] that these limits can be improved by the use of signal processing. Both multilevel signaling, which allows a higher data rate for the same bandwidth, and equalization, which compensates for line loss, have been employed to varying degrees to achieve 10-Gb/s signaling over ~1 m of copper on FR-4. However, this improvement comes at a high price in terms of power dissipation. For example, the MAX3804, available from Maxim [12], which extends the transmission distance for a 10-Gb/s nonreturn-to-zero (NRZ) signal over FR-4 to 30 inches through receive equalization, consumes 115 mW. By comparison, a single MAUI channel running at 10 Gb/s is expected to consume less than 40 mW per Tx/Rx pair, representing nearly a three-fold reduction in power consumption. Since the MAUI Tx can be located very close to the output of the previous stage of electronics, additional power savings can be realized by reducing the transmitted voltage level of this stage relative to that required to drive 30 inches of FR-4 backplane.

Electrical backplanes have other limitations that are harder to quantify than the loss but that are just as important, particularly as line density increases. Electrical interconnects are susceptible to electromagnetic interference (EMI) from the board and rack environment, including (and especially) crosstalk from adjacent channels. These issues are particularly serious at high-density connectors which interface plug-in boards to the backplane (see, e.g., [13] and [14]). These problems become more serious as signal speeds increase, since the requirement of low crosstalk and impedance control conflicts with the desire to make the connector small and the bandwidth density large. The use of dielectric rather than metal-based waveguides makes optical interconnects immune to EMI. The very tight confinement of light in optical waveguides (of the order of micrometers) also makes both

the waveguides and corresponding connectors essentially free of crosstalk.

In spite of these limitations, however, electrical backplanes remain the only commercial solution to high-density interconnects in systems today. This is partly because they can be made to work at the current required line rates, despite their technical limitations. More important, the cost of electrical solutions remains lower than their optical counterparts. As long as electrical solutions remain adequate for the required bit rates *and* provide a cost advantage, they will continue to dominate in low-cost, high-volume markets. The challenge for optical interconnects, as bit rates continue to increase, therefore, is to become cost competitive with electrical solutions while maintaining the attractive technical advantages they currently enjoy over their electrical counterparts. In Section IX, we will describe efforts within the MAUI program to reduce the cost of backplane optical interconnects.

IV. Parallel WDM Optical Interconnects

The technologies used to implement PWDM in the MAUI program leverage heavily from previous work in both parallel optics and integrated CWDM. In a typical parallel optical interconnect, one optical signal is transmitted through each fiber of a 12-fiber ribbon. Typical fiber ribbon has a fiber-to-fiber pitch of 250 μm and a multimode fiber core diameter of either 62.5 or 50 μm, although single-mode fiber has also been used. The Tx typically contains a 12-channel driver integrated circuit (IC) and a 12-element vertical-cavity surface-emitting laser (VCSEL) array, while the Rx contains a 12-channel Rx IC and a 12-element photodiode array. Both Tx and Rx also typically include a 12-element lens array for coupling between the optoelectronics and the fiber.

To improve density and reduce electrical parasitics while allowing for wafer-scale assembly of optical subassemblies, a chip-mounted enclosure (CME) scheme was developed by Agilent Laboratories for use in a 12 $\times$ 10-Gb/s parallel optical interconnect [3], [5]. In this scheme, the mechanical base of the subassembly is the IC. Bottom-emitting VCSEL or bottom-illuminated photodiode arrays are flip-chip bonded to pads on the IC. A silicon ring surrounding the optoelectronics is also soldered to the IC, and a microoptic lens array is soldered to the ring, creating a hermetic enclosure for the optoelectronics. Finally, a mechanical lid, containing alignment pins, is aligned and attached to the lens array, creating a prealigned subassembly that will mate to an MT-ferrule (see, e.g., [15]), which is the ferrule most commonly used in fiber-ribbon connectors.

Integrated CWDM technology was initially developed to increase the bandwidth-distance limits of multimode fiber in local area networks (LANs) and to provide a lower cost alternative to high-speed serial links by using lower speed optoelectronic and electronic components [6]–[8]. Most recently, this has been applied to the 10 GBase-LX4 implementation of 10-Gigabit Ethernet [16]. In an integrated CWDM Tx or Rx, a compact optical mux or demux is directly coupled to an array of photodiodes or a multiwavelength array of lasers, which, as in parallel optics, is driven by a multichannel driver or Rx IC.

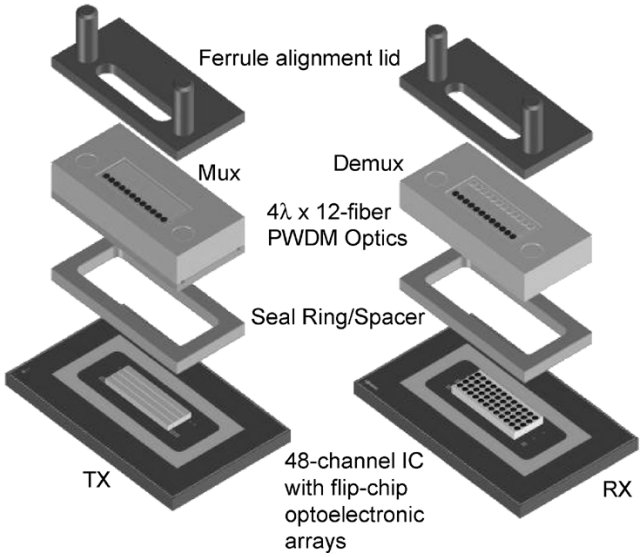


Fig. 2. Exploded view of MAUI PMOSA Tx (left) and Rx (right). The base of the PMOSA is a 48-channel IC with wire-bond I/O pads around the perimeter. A 4 $\times$ 12 optoelectronics array is flip-chip bonded in the center, surrounded by a silicon seal ring/spacer. PWDM multiplexing and demultiplexing optics (4λ $\times$ 12 fiber) are attached over the seal ring, and a mechanical lid that mates to an MT-ferrule sits atop the subassembly. Each PMOSA has a 5 $\times$ 8 mm footprint and is expected to consume less than 1 W while transmitting or receiving between 240 and 720 Gb/s over a 12-fiber ribbon.

PWDM combines parallel optics with CWDM by transmitting multiple wavelengths through each fiber in a ribbon. Some previous PWDM efforts used wavelength for channel selection purposes but did not involve simultaneous links operating on all wavelengths with independent data [17]. Other efforts that did implement simultaneous links on all channels used aggregated single-fiber CWDM subassemblies and did not involve a compact package [18].

The MAUI PMOSA leverages the CME approach, originally developed for parallel optics, while extending CWDM mux and demux technology to include arrays of many 1 $\times$ n muxes or n $\times$ 1 demuxes on a single chip. Fig. 2 shows the first-generation MAUI Tx and Rx PMOSAs that have been built. The driver and Rx ICs, which form the mechanical base of the PMOSA, each have flip-chip solder bumps for a 4 $\times$ 12 array of optoelectronics, with wire-bond pads arranged around the outer perimeter of the IC for electrical input/output (I/O) to the 48 channels. The initial version of the ICs operate at 5 Gb/s per channel, but versions operating at 10 and 15 Gb/s per channel are planned.

The first-generation Tx PMOSA has four 1 $\times$ 12 bottom-emitting VCSEL arrays, each a different wavelength, flip-chip bonded to the IC. The wavelengths used are 990, 1020, 1050, and 1080 nm. The first-generation Rx PMOSA has a 4 $\times$ 12 substrate-illuminated lensed photodiode array flip-chip bonded to the IC.

The PWDM mux is a passive optical component with a 4 $\times$ 12 array of lenses patterned on the underside and a 1 $\times$ 12 array of lenses on the top side, which combines the four wavelengths of the 48 VCSELs into the 12 output fibers. The PWDM demux has a 1 $\times$ 12 array of lenses on the top side that receives input from the 12 fibers and separates them into 48 output beams that

exit the bottom side, where they are focused onto the photodiodes by a 4 × 12 array of lenses integrated into the substrate side of the photodiodes. Both mux and demux are photolithographically patterned in wafer form using a process commonly used to fabricate microlens arrays. Dielectric interference filters are attached to the mux and demux to provide wavelength selectivity.

A mechanical lid is stacked on top of the optics. This lid has an opening in it for light to pass through and alignment pins to mate to an MT-ferrule. Once assembled, the PMOSA is ready to be wire-bonded to an electrical circuit board or other package and optically connected to a fiber ribbon with MT-ferrule termination. First-generation PMOSAs have been built and operated at 5.21 Gb/s per channel to realize a 250-Gb/s link. These results will be published shortly.

Due to the large number of high-speed electrical I/O ports, the incorporation of a PMOSA into a system typically involves two levels of electrical interconnection. As an example, consider the first-generation MAUI PMOSA, where the I/O pads are distributed along the periphery of the IC. The first level of interconnection can be realized using gold wire bonds to connect from the IC to a fan-out package with very high trace density, built with a ceramic or flex circuit board technology. The second level of interconnection could be a ball-grid array (BGA) interface between the fan-out package and the system board.

The actual module footprint and, hence, the areal density of interconnection will be, in this case, determined by the minimum pitch of the BGA interface. However, the chip-like nature of the PMOSA lends itself to its direct integration on to a system chip in a piggyback fashion. In this case, only one level of wirebond interconnection is needed, essentially reducing the module footprint to that of the PMOSA itself. As PMOSA technology evolves toward higher channel counts, I/O ports will likely be distributed into an areal array on the IC surface, in which case the first level of interconnection will be a flip-chip-type interconnection, rather than wire bonds.

With a power dissipation of less than 1 W per end and a footprint of 5 × 8 mm, a PMOSA operating between 5 and 15 Gb/s per channel could transmit or receive between 240 and 720 Gb/s through a 12-fiber ribbon. The complete PWDM link would thus have an areal bandwidth density of between 300 and 900 Gb/s/cm^2 (counting only the chip footprint) and a bandwidth per unit power of between 120 and 360 Gb/s/W.

The linear bandwidth density at the board edge would depend upon the type of backplane connector being used. Most optical backplane connector systems that are commercially available today are based on the MT-ferrule. Teradyne's HD-Optyx connector, for example, can accommodate 3.58 MT-ferrules per inch (data taken from [19]). Using this system, together with a 6 × 12 (72-fiber) version of the MT-ferrule [20], results in a backplane connection with 258 fibers per inch. Assuming that each 72-fiber ferrule has six short 12-fiber ribbons that are each routed to a different PMOSA on the board, the linear bandwidth density of the backplane connection would be between 2580 and 7740 Gb/s/in or 1016 and 3047 Gb/s/cm. This can be contrasted with Teradyne's state-of-the-art five-pair GbX electrical backplane connector that has a connection density of 69 differential pairs per linear inch (data taken from [21]), yielding a linear bandwidth density of 216 Gb/s per linear inch, assuming that each differential line can sustain a 6.25-Gb/s signal in one direction.

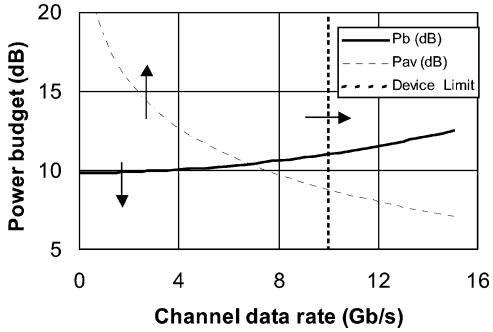


Fig. 3. Available and required optical power budget of a typical PWDM interconnect as a function of channel data rate. Limit of 10 Gb/s imposed by devices is also shown. Arrows indicate future trends, as component technologies improve. With increasing power consumption, the P_{av} curve moves upward in the direction of the arrow. An interconnect length of 10 m on a 50-μm core multimode fiber is assumed.

Ultimately, the choice between the PWDM approach described previously and a purely parallel approach using a single wavelength, such as 2-D parallel optics, [20] will depend on cost and performance tradeoffs. A PWDM approach has the advantages that it will greatly reduce the number of fiber ribbons and optical backplane connectors required in a system and increase the linear bandwidth density of the backplane connection by a factor equal to the number of wavelengths. In addition, single-row parallel connectors are quite a bit less costly than 2-D connectors. The 2-D parallel approach has several advantages: 1) a 2-D lens array should be lower cost; 2) it has less optical loss than a PWDM mux and demux; and 3) only a single wavelength of VCSEL is needed. Which approach will be better suited for FTTP applications will depend upon the cost evolution of parallel optical connectors and fiber ribbons, the cost and loss evolution of the PWDM mux and demux, and the cost evolution of VCSEL arrays. These tradeoffs will be investigated under the MAUI program, and a 2-D parallel version of the PMOSA that incorporates a 4 × 12 lens array in place of the mux and demux is planned. Ultimately, it may be possible to construct a PMOSA that combines these approaches by transmitting multiple wavelengths over a 2-D array of fibers.

V. PWDM System Optimization

The most important factors driving component requirements for a PWDM interconnect are available optical power budget, required optical power budget, total electrical power consumption, and line rate per channel. These factors are very much interrelated, and the tradeoffs are illustrated in Fig. 3. The available optical power budget $P_{\rm av}$, is obtained by subtracting the Rx sensitivity S from the optical signal output from the Tx. $P_{\rm av}$ is a function of data rate and total electrical power consumption, as shown in Fig. 3. The maximum achievable $P_{\rm av}$ is determined by the properties of the VCSELs and photodiodes being used. Higher output power from VCSELs and higher responsivity and lower capacitance of photodiodes can greatly improve $P_{\rm av}$. However, to what extent this improvement can be re-

alized depends on other considerations, such as VCSEL speed and threshold current, and constraints on detector size imposed by coupling optics.

The required optical power budget P_b is the total of all the optical losses, including the mux/demux insertion loss, optical coupling loss, fiber attenuation and link power penalties, plus a margin to account for temperature variation and the aging of devices. For fiber lengths and data rates typical for a PWDM interconnect, fiber attenuation and link penalties are small, and P_b is dominated by the mux/demux insertion loss and optical coupling loss, although at data rates beyond 10 Gb/s, the link penalty due to intersymbol interference (ISI) begins to make a significant contribution to P_b. The condition for an interconnect to operate properly is $P_b <= P_{\rm av}$.

For PWDM system optimization, it is important to tailor the properties of VCSELs and detectors under the given constraints to maximize $P_{\rm av}$ so that it far exceeds P_b. As shown in Fig. 3, the excess power budget can then be traded off to reduce total power consumption of the system or to increase the data rate per channel. As component technologies improve, each of the curves or lines in Fig. 3 moves in the direction of the arrow, making it possible for the PWDM interconnect to consume less power and/or operate at a higher data rate per channel.

Independent of $P_{\rm av}$, there are several other parameters that directly affect electrical power consumption. In the Rx, a significant portion of power is dissipated in the output stage. Thus, reducing the output voltage swing and increasing the system impedance will greatly decrease power consumption. In a proprietary system designed with PWDM Tx and Rx in mind, these parameters can be optimized. In other cases, where PWDM modules have to be designed for a 50-Ω environment with standard logic levels, voltage swing requirements can still be reduced by a careful examination of electrical signal loss in the system.

For the Tx, the characteristics of the VCSELs being used play an important role in reducing power consumption. Lower forward-bias voltage for the VCSEL helps lower the supply rail of the driver IC, leading to a direct reduction in power consumption. In addition, for a given optical modulation power from the VCSEL, the lower the logic-1 level current required, the lower the power dissipation of the driver IC. It should be noted that in order to lower the required logic-1 current, simultaneous optimization of the VCSEL threshold current and slope efficiency is needed.

An optimized PWDM interconnect system is the result of a careful tradeoff analysis of many different parameters across the system producing optimal specifications for each component. For the purpose of illustration, the first-generation MAUI specifications for a 48-ch PWDM system with an aggregate capacity of 240 Gb/s are listed in Table I.

The aggregate capacity of a PWDM interconnect system can be increased by increasing the number of channels and/or increasing the data rate per channel. For a given set of component technologies, the channel data rate can only be increased by increasing the output power from the VCSEL, which, referring to Fig. 3, will move $P_{\rm av}$ upward, shifting the intersection of $P_{\rm av}$ and $P_{\rm b}$ to a higher data rate. Due to saturation of VCSEL output power, the increase in channel data rate is limited, often leaving a channel count increase as the only option for achieving greater aggregate capacity. Even when permissible, increasing the data rate may result in more power consumption compared with increasing channel count. To illustrate the point, let us consider power consumed by two systems, system 1 having N channels each with data rate B and system 2 having 2 N channels each with data rate B/2 so that the aggregate capacity is the same for both cases. For system 1, the VCSEL will need to produce more optical power, requiring its driver IC to dissipate more electrical power. It can be seen from Fig. 4 that if the optical power has to be increased by more than 3 dB, which could easily be the case, then the Tx power consumption will be more than double. In addition, to accommodate twice the bandwidth, the system 1 Rx has to dissipate more power, often by more than a factor of 2. It is then highly possible that total power consumption of system 1 would be larger than system 2. It should be noted that the optimum choice of channel count and channel data rate depends on other considerations as well, including complexity of optics, electrical I/O port density and speed, and crosstalk, among others.

TABLE I
EXAMPLE PARAMETERS FOR 240-GBd PWDM SYSTEM

Parameter	Symbol	Value	Unit
Min. data rate	B_{min}	5.0	Gb/s
Max. TX power consumption	$P_{TX,max}$	1.0	W
Min. logic 1 current	$I_{1,min}$	4.0	mA
Min. logic 0 current	$I_{0,min}$	1.2	mA
Min. VCSEL slope efficiency	η_{min}	0.30	W/A
Max. VCSEL threshold current	$I_{th,max}$	0.40	mA
Max. VCSEL forward voltage	$V_{F,max}$	2.0	V
Min. optical extinction ratio	ER_{min}	6.0	dB
Max. link power budget	$P_{L,max}$	11.5	dB
Min. Bit-error-ratio	10^{-12}		
Min. RX sensitivity	S_{min}	-13.5	dBm
Min. RX BW	$f_{RX,min}$	3.75	GHz
Max. RX power consumption	$P_{RX,max}$	1.0	W
Min. photodetector BW	$f_{3dB,min}$	5.0	GHz
Max. detector capacitance	C_{max}	200	fF

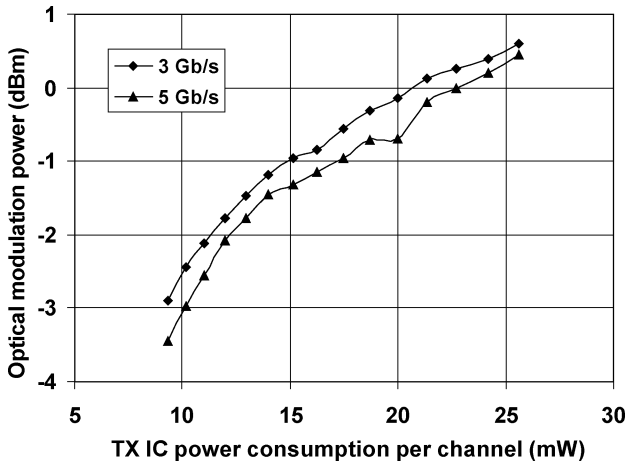


Fig. 4. Optical modulation power from VCSELs as a function of Tx IC power consumption at 3 and 5 Gb/s. These empirical plots are typical for VCSELs and drivers used in a first-generation PWDM system.

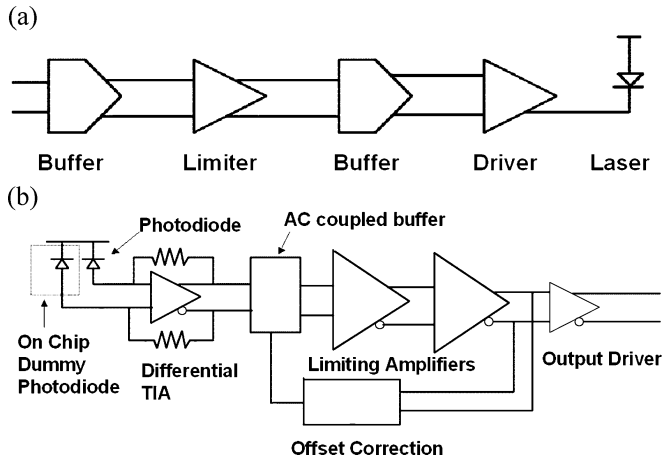


Fig. 5. Block diagram for single channel of (a) Tx and (b) Rx ICs. These diagrams do not include bypassing and filtering electronics but include the essential circuit blocks used in the core electronics. In the Tx, a dummy laser diode (not shown) is included on chip to prevent common-mode feedback effects. Similarly, a dummy photodiode is included on chip on each receive channel.

VI. Integrated Circuits

The key goal in the design of the PMOSA Tx and Rx circuits has been the achievement of maximum data transfer per unit of power dissipated. This is achieved by using low supplies and currents and designing with this goal in mind. Because the fiber losses and link penalties are negligible for the very short distances that MAUI aims to address, and since both ends of the link, including the lasers and photodiodes can be optimized together, power efficiencies in excess of 250 Gb/s/W can be achieved.

The first-generation MAUI ICs are implemented in a 45-GHz f_T SiGe BiCMOS process (i.e., a process that includes both CMOS and SiGe bipolar transistors). A block diagram of the Rx and Tx circuits used for each channel is shown in Fig. 5. A full CMOS implementation is possible but is somewhat restricted by the bias and voltage swing requirements of the VCSELs and photodiodes. In addition, short-channel CMOS is much more expensive to prototype due to mask costs. Since the cost of the ICs is expected to account for only a small fraction of the total PMOSA cost, short-channel CMOS may not be compelling for this application. Eventually, short-channel CMOS may be required if it becomes necessary to integrate higher level functionality or even processing power into the PMOSA.

To illustrate the type of performance that can be expected, Rx and Tx circuits designed in a 45-GHz f_T BiCMOS process can be designed to operate from 2.5-V supplies. By choosing small devices and using currents in the microamp range in the front end and in the receive chain, sensitivities below -17 dBm at 5 Gb/s with a total power consumption of 15 mW are achievable. The bandwidth of the Rx under worst-case conditions is set to about 70% of the data rate to prevent undue losses due to ISI. Transmitter power consumption of 15 mW per channel at 5 Gb/s can also be achieved.

In order to minimize power in the MAUI Rx, it is necessary to reduce the normal postamplifier limiting stages to one and to simplify the offset correction circuitry. In addition, since the MAUI application is short distance and does not require a high dynamic range, further power savings can be gained by compromising the high-power end of the range through reduction of currents in the buffer circuits and amplifier stages and by incorporating as much gain as possible in the transimpedance stage. Reducing the gain of the Rx is enabled by the reduced link budget.

It is of course possible to design a single-ended Rx, and this has the potential to be even more power efficient; however, this approach compromises the power supply rejection ratio, which is very critical in this application [22]. Supply noise is one of the major mechanisms for creating crosstalk, and the dense packing of 48 channels with multigigahertz bandwidths makes this a key consideration. In the differential configuration used, on-chip diodes, which emulate the photodiodes, are used to preserve symmetry to the maximum extent possible and avoid common-mode feedback effects.

The MAUI Tx design requires a careful choice of both the extinction ratio and the bias current to prevent excessive jitter. This requires that the bias current be substantially higher than the laser threshold. The VCSEL should be optimized for threshold rather than speed [23].

In order to minimize crosstalk, all of the MOS current sources are placed into isolated p-wells to reduce substrate injection effects, and all bias references have filter circuitry. Local bypassing is also used whenever possible to minimize supply noise. To maintain isolation from the substrate, metal-to-metal capacitors are used, rather than capacitors using diffused layers. Finally, high-speed I/Os are routed using a differential coplanar waveguide configuration, where the signals maintain a high degree of isolation due to the presence of ground planes between signals.

Because these circuits are running at very low currents, parasitic effects associated with having large-value passive elements have to be carefully evaluated. It is very important to design the passive components with this in mind. Adequate resistor models that include statistical mismatch data for process variations of length and width are essential. Substantial effort must be placed into simulations that include layout parasitics to minimize the circuit area and ensure high yield. Passives must be sized with the tradeoff of increased area versus increased parasitics considered for every device and circuit node.

Because so many minimum-size devices are used in these circuits, designing for the effects of mismatch in transistor base-emitter voltage $\mathrm{V_{be}}$ and MOS threshold voltage $\mathrm{V_t}$ is also a critical consideration. This can be addressed by using Monte Carlo simulations that vary both the process variations and the environmental conditions. The results, both before and after design optimization, of one of the Rx simulations is included in Fig. 6. This identified multiple robustness problems in the circuitry, which were solved by some configurational changes and some changes in device choices. For this circuitry, the Monte Carlo analysis was very valuable in identifying the minimum practical size for various devices and in identifying topology problems.

As IC processes become faster, PMOSA performance can continue to improve. Throughout the duration of the MAUI program, efforts will be made to further increase power efficiency by lowering the supply voltage and decreasing currents, while

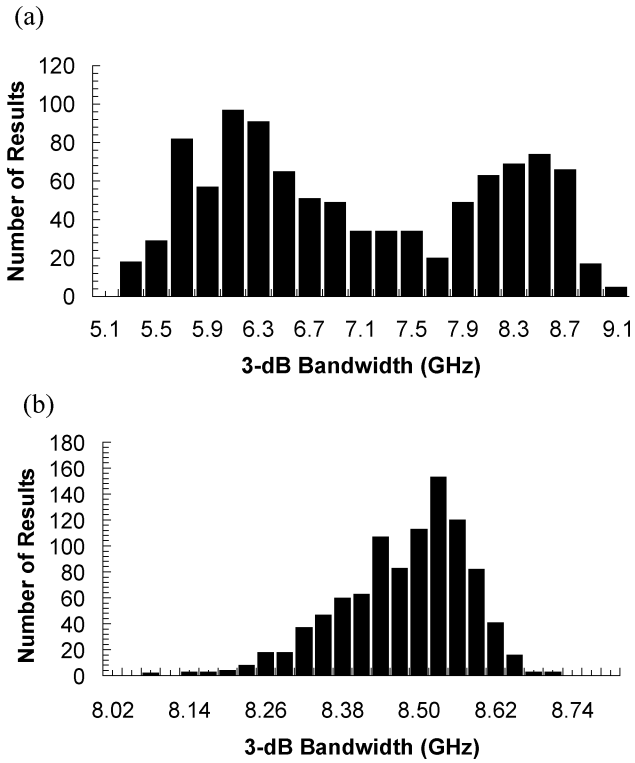


Fig. 6. Example Monte Carlo simulations, each showing 1000 simulation results. In this instance, the bandwidth of the Rx channel is simulated over process, mismatch effects, and temperature to test the robustness of the design. In (a), the wide bimodal distribution and large standard deviation (1.08 GHz) indicate a problem. This problem was traced primarily to a resistor tolerancing issue. A rerun of the Monte Carlo analysis after the layout of the cell and schematic are improved is shown in (b). After the improvement, the standard deviation has been reduced to 97 MHz.

increasing line rates. About 40% of the power is now consumed in the I/Os. This is dictated by both the optical power required by the link budget and by the minimum signals required by downstream CMOS circuits. Reductions in the voltage and current requirements of the VCSELs and photodetectors, as well as reductions in the insertion loss of the PWDM optics will have a significant impact on improving the overall power efficiency of future Tx and Rx ICs.

VII. PWDM Optics

Both the PWDM mux and demux being developed for the MAUI program are based on the same basic optical zig-zag architecture that has been used in single-fiber CWDM demuxes [6]–[8]. In the PMOSA, however, we have extended this architecture to apply to a 12-fiber ribbon with a 250-μm fiber-to-fiber pitch. Thus, each optical chip has 12 four-wavelength muxes or demuxes integrated side by side on a 250-μm pitch, as can be seen in Fig. 2. Fig. 7 shows ray-tracing models of a single-fiber cross section for the first-generation MAUI mux and demux.

Unlike earlier CWDM optics, which were individually fabricated injection-molded plastic parts, the MAUI optics are patterned photolithographically in wafer form. This approach is taken because it is more suitable for devices containing large numbers of lenses, it allows for wafer-scale optical alignment and assembly, and it allows lenses to be fabricated in high-refractive-index materials.

The basic optical elements used in this architecture are lenses, mirrors, and bandpass filters. Lenses are used at the inputs and outputs to angle light beams and to either collimate or focus. Mirrors are used to bounce the beam internally along a zig-zag path. Bandpass filters are used to separate or combine wavelengths through spectrally selective transmission. Bandpass filters transmit all wavelengths within a particular band and reflect all others.

Since the MAUI mux and demux are meant to always be used together as part of a link, tradeoffs in their losses relative to one another must be considered. One such tradeoff is the choice of fiber type. The use of large-core multimode fiber facilitates the mux design, since coupling is always easier to a larger fiber. On the other hand, the use of smaller core fiber would simplify the design of the demux, since the output of a smaller core fiber is easier to collimate and can be focused to a smaller spot. The fiber type should be chosen to minimize the combined loss of the mux and demux.

It should be noted that independent of fiber type, minimizing the VCSEL emitting area and numerical aperture (NA) and maximizing the photodiode area will tend to minimize the overall loss of the optics. It is also worth noting that by ordering the wavelengths such that the channel with the most loss in the mux has the least loss in the demux, and vice versa, the combined loss of the optics can be minimized.

The mux can be thought of as an imaging system that projects the image of the VCSEL aperture onto the fiber input facet. Since the NA of the VCSEL is usually larger than the NA of the fiber, magnification is required. This magnification imposes a tighter alignment tolerance of the VCSEL to the mux than would be required in a one-to-one imaging system.

In the demux, the light is directed to the photodiodes, which have a very high collection angle. Consequently, the goal is to achieve the smallest possible focused spot on the detector. This allows for improved alignment tolerances and smaller detector active areas, which in turn have improved high-speed performance. To minimize the area of the focused spots, the focusing lenses are integrated into the photodiode substrate, allowing them to be closer to the focal plane than had lenses on the bottom side of the demux been used.

In the design of the PWDM optics, either refractive or diffractive lenses can be used. Diffractive lenses are easier and less expensive to manufacture but impose an inherent loss penalty due to finite diffraction efficiency. One advantage is that a vortex design can be used to minimize back reflections to the laser sources, which can cause instabilities at high modulation speeds [24]. Refractive lenses do not have an inherent loss penalty, but they require sophisticated and costly manufacturing techniques, such as gray-scale lithography, and are more prone to deviations from design parameters.

Concave relay mirrors are used to maintain collimation over long optical paths and become necessary as the number of wavelengths increases. In addition, relay mirrors can be used to improve alignment tolerances in the mux and demux. The use of relay mirrors, however, induces higher complexity and increases manufacturing cost.

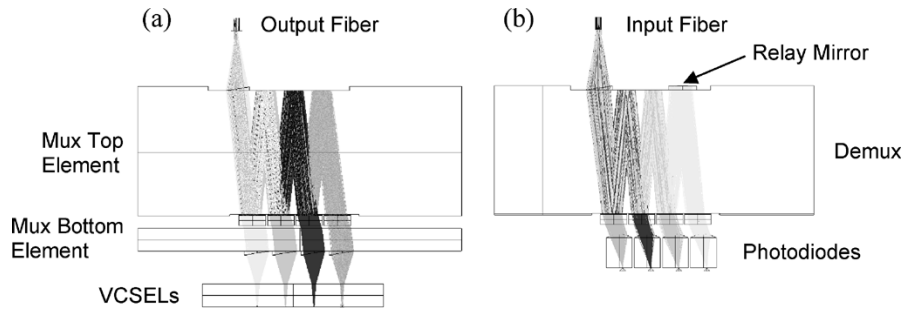


Fig. 7. Single-fiber ray-tracing models of the first-generation PWDM optical elements. (a) Multiplexer: Light is emitted upward from the four VCSELs at the bottom. Four lenses on the bottom mux element collimate and angle the light, which then passes through bandpass filters matching each VCSEL wavelength. Through consecutive bounces off the top reflective surface of the top element and the filters, the four beams converge to a single beam. The output lens at the top left of the top mux element angles and focuses that beam onto the fiber input facet. (b) Demultiplexer: Light is emitted downward from the fiber. The lens at the top of the demux collimates and angles the light, which then bounces between the bandpass filters at the bottom and the reflective top surface. A concave relay mirror helps collimate the light on the last bounce. Each filter allows a specific wavelength to pass. Each output beam is collected by a lens on the top side of the photodiode and is focused onto the active area on the bottom side of the photodiode. The actual PWDM optical elements include 12 replications of the models shown, one for each fiber, on a 250-μm pitch.

Refractive and reflective optics do not impose any loss due to geometrical limitations. Nevertheless, there are several secondary loss sources. These include scattering due to surface imperfections, filter and reflector losses, alignment losses, imperfect anti-reflection coatings, and losses from lens and relay mirror fabrication errors. Should diffractive lenses be used, an additional loss is imposed to the ones previously mentioned, due to the finite diffraction efficiency of these elements.

Of the loss sources mentioned, only the filter and reflector losses are unique to PWDM, the rest being present in standard parallel optics. While some of the other loss sources might be accentuated due to the increased complexity of the PWDM optics, it should still be possible to implement PWDM with a minimal loss penalty relative to 2-D parallel optics.

PWDM optical interconnects are inherently 2-D, with one axis extending along the fiber array and the second along wavelength. The scaling of PWDM interconnects can be achieved by extending either or both of these axes.

To scale along the wavelength axis, one has to use more VCSEL wavelengths. This involves reducing the wavelength spacing between channels, extending the overall wavelength range, or both. As one reduces the wavelength spacing, tolerances on the VCSELs and optics become tighter, the mux and demux dimensions must become larger, and the PWDM optics design becomes more difficult. On the other hand, extending the wavelength range requires new VCSEL wavelengths to be developed. To accommodate more channels, the PWDM optics will need to incorporate additional relay mirrors to maintain collimation over a longer optical path. The alignment tolerance of the devices will be decreased, since some misalignments are amplified with additional bounces. Finally, material absorption could become an issue, depending upon the wavelengths and the optical material used.

In the first-generation MAUI PMOSA, four wavelengths are used with a 30-nm spacing. The shortest wavelength, 990 nm, was chosen by the requirement that the indium phosphide (InP) photodiode substrate be transparent, and the longest wavelength, 1080 nm, was chosen by the desire to use four VCSELs that could be easily fabricated in the same indium gallium arsenide (InGaAs) multi-quantum-well system. By changing the filters, the same MAUI mux and demux will work with any set of wavelengths that are spaced by at least 20 nm and are not absorbed by the material in which the microoptics are fabricated.

Scaling to eight wavelengths may be practical without a major change in the mux and demux architecture by adding more VCSEL wavelengths. As tolerances on both the microoptic lens fabrication and on alignment improve with time, scaling beyond eight wavelengths may be possible.

Scaling along the fiber axis is straightforward as far as the optics are concerned. To accommodate more than 12 fibers, a PWDM mux or demux must simply integrate the required number of replications of the single-fiber mux or demux in an array matched to the fiber connector. Limits are set by the manufacturability and cost of the fiber connectors. Ultimately, scaling in either dimension may be limited by the yield of the optoelectronics.

VIII. MAUI-Enabled Computer Architecture

The PWDM optical interconnects being developed under the MAUI program can be used in many applications in which very high bandwidth density is required over relatively short distances. In this section, we focus on one possible MAUI-enabled architecture being studied at USC for a system-area network (SAN) based on a new building block called the encapsulated processor. The encapsulated processor is a single CMOS chip with fiber-optic ports as the only means of off-chip high-speed data communication. In addition to a theoretical study of the encapsulated processor architecture, USC is developing circuitry that can be used in this architecture to interface between a PMOSA and a processor.

By incorporating the advantages of fiber optics with the integration capability of scaled CMOS electronics, the encapsulated processor leads to a new microprocessor design point. Fiber optics brings the advantages of reduced power dissipation from high-speed chip I/O, improved edge-connection density bandwidth, low crosstalk, and zero EMI to systems, enabling high-bandwidth access to key resources such as main memory and the SAN, scaling to larger multiprocessor systems,

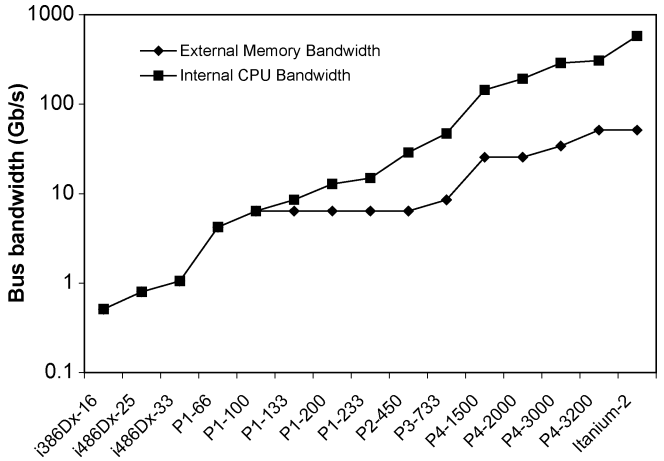


Fig. 8. Imbalance between microprocessor internal data path bandwidth and memory-bus bandwidth continues to grow with each microprocessor generation, already in excess of 0.5 Tb/s [18].

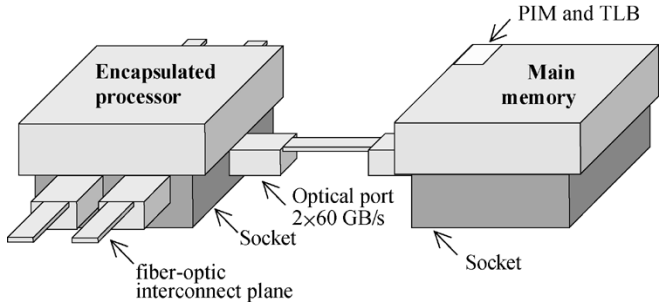


Fig. 9. Encapsulated processor shown is a single CMOS chip with fiber-optic ports as the only means of external high-speed data communication. In the implementation shown, a socket supplies power and ground and includes optical ports, each of which provide 2 × 60 GB/s (480 Gb/s) of external data bandwidth.

and reduced overall system power density in distributed multiprocessor systems.

Increasing the per-port bandwidth has a particularly large impact in multiprocessor systems where it is a simple way to dramatically reduce SAN latency. Under equal loading conditions in a multiprocessor SAN, an interconnection network with a bandwidth of 32 GB/s/port has a six times lower latency compared with an interconnection network with a bandwidth of 3.2 GB/s/port [25]. Data transfer latency from memory will also be dramatically reduced through the increased bandwidth available with the encapsulated processor.

The microprocessor as we know it continues to be implemented as a distinct computing engine with small, fast on-chip memory and a significantly larger off-chip external memory. This fundamental system partition demands low-latency and high-bandwidth access to memory for any high-performance processor. While the advantages of high-memory bandwidth are understood, conventional bus-based electrical solutions are performance limited due to difficulties with electrical signaling at multigigahertz rates over low-cost packaging and printed circuit board (PCB) interconnects as discussed in Section III. This has resulted in an increasing mismatch between internal microprocessor bandwidth and external memory-bus bandwidth, as shown in Fig. 8 (data from [26]). Fig. 8 accounts for a superscalar microprocessor architecture by multiplying internal data path width by the number of instructions that can be issued simultaneously. The internal CPU bandwidth of the Itanium-2 processor is more than an order of magnitude greater than the external memory-bus bandwidth. Attempts to increase external memory-bus bandwidth using electrical signaling result in increased power dissipation, reduced noise tolerance, and latency.

In the coming years, the imbalance between microprocessor performance and memory access will be driven to a crisis point [25], [27]. As illustrated by Fig. 8, with successive generations, the difference between internal microprocessor bandwidth and memory-bus bandwidth will continue to grow beyond an order of magnitude. Without a new approach, the microprocessor will lose hundreds of process cycles while waiting for a single read from main memory.

Data transfer latency from memory cannot be hidden by pre-fetch when a user application is memory-bandwidth bound. High-bandwidth memory interfaces improve microprocessor performance even for cache-friendly applications [25] by reducing the L2 miss penalty at the cost of a higher L2 miss rate, allowing reduced on-chip memory, better yields, and lower power dissipation.

As shown in Fig. 9, the encapsulated processor is a single CMOS IC with fiber-optic ports as the only means of external high-speed data communication. This can be implemented, for example, with a socket that, in addition to supplying power and ground, includes the PMOSAs and the interface ICs required for encoding and multiplexing the raw data into a form suitable for optical transmission. There are low-power, noncontrolled impedance, short-distance electrical links embedded in the socket for interconnecting the processor, the interface ICs, and the PMOSA. Eventually, the interface circuitry may be integrated into the PMOSA. The processor IC can have separate thermal management from the PMOSAs and interface ICs. Optical ports are available for memory, I/O, multiprocessor, and scalable SAN interconnection.

The encapsulated processor includes a single CMOS chip consisting of two or more CPUs with an L1 and L2 cache connected by a crossbar switch, as shown in Fig. 10. The crossbar connects to an on-chip L3 cache and multiple high-speed fiber-optic ports. Main memory could be connected directly to an optical port or via an external crossbar. Main memory may have its own processors (PIM) and pipelined translation look-aside buffer (TLB) whose purpose is to efficiently feed the encapsulated processor.

The bisection bandwidth of the eight-port crossbar switch integrated into the encapsulated processor shown in Fig. 10 would be 960 GB/s (7.68 Tb/s). Innovative circuit design using 90-nm CMOS technology predicts that the crossbar switch core could dissipate a total power of less than 6 W.

The direct replacement of an electrical link with optics introduces an electrical-to-optical and optical-to-electrical conversion delay. Typical values for this delay are less than 0.5 ns for a complete link. In practical applications, this is compensated for by the reduced time of flight of an optical signal traveling in glass fiber compared with an electrical signal propagating in an FR-4 dielectric. Such signal delays are insignificant compared with other latencies in the system.

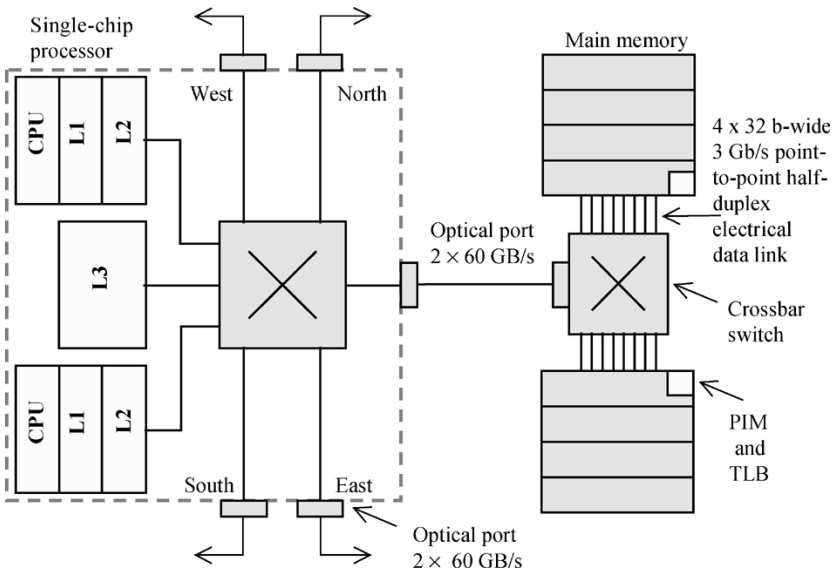


Fig. 10. Encapsulated processor consists of two or more CPUs with L1 and L2 cache connected by a crossbar switch. Main memory could be connected directly to an optical port or via external crossbar using dual-simplex point-to-point electrical links signaling at 3 Gb/s.

In the example shown in Fig. 10, the total latency to request data from main memory and return it to the microprocessor of about 37 ns is dominated by the 25-ns core row-access latency of the dynamic random access memory (DRAM) itself. As future memory designs improve on this value, the advantages of using a PWDM fiber-optic port in combination with crossbar switches become more evident. The high-bandwidth port consumes less power and less board area compared with any all-electrical alternative. In addition, memory can be scaled while incurring minimal additional latency by adding crossbar switches. In general, a high-bandwidth SAN also allows intelligence (and the associated power dissipation) to be distributed throughout the system.

The encapsulated processor, serviced by high-bandwidth fiber-optic interconnects, can become a basic general-purpose building block for systems because the high-bandwidth fiber-optic ports enable a wide range of specialized system configurations. FTTP is a natural technology convergence point that should first find application in bandwidth-intensive mainframe and server applications, followed by rapid adoption in the PC, workstation, and router market segments [28].

IX. Manufacturability and Cost

One of the primary goals of the MAUI program is to develop technologies that will minimize the final cost of the interconnect solution. In order to achieve a $1/Gb/s price target per link, a 48 $\times$ 10.4-Gb/s PMOSA Tx/Rx pair plus the needed connectors and cabling would need to sell for $500.00. In addition to very low-cost components, this will require very high-throughput assembly and test and very high optoelectronic yields.

One technology that has the potential to dramatically reduce the overall cost of the PMOSA is wafer-scale optical alignment and assembly. Because the mechanical base of the PMOSA is an integrated circuit, it is possible to assemble a large array of PMOSAs with the ICs in wafer form [3], [5]. Since the PWDM mux and demux are constructed out of photolithographically defined microlens arrays, they can be manufactured as wafers and then optically aligned and attached to a wafer or portion of a wafer of ICs. It may also be possible to align and attach the lid in wafer form.

By replacing many precision optical alignment and attachment steps with a single wafer-scale step, significant cost savings are possible. This potential cost savings must be weighed against the improved performance and yield that can be obtained by individually actively aligning and assembling each PMOSA. These tradeoffs will be investigated as part of the MAUI program.

For many practical reasons, it does not make sense to attach optoelectronic components in wafer form; nevertheless, to allow for subsequent wafer-scale assembly of the PWDM optics, it may make sense to attach singulated optoelectronic arrays to ICs that are still in wafer form. Traditional flip-chip die-attach systems solder a single die at a time, making the VCSEL and photodiode attachment step a potential bottleneck. As part of the MAUI program, an automated die-attach system is being developed that will be capable of simultaneously aligning and soldering multiple die. This technology has the potential of dramatically reducing the time required to populate the ICs with optoelectronics.

A large part of the cost associated with any multichannel optical interconnect technology is in the testing. With 48 high-speed differential channels, tests will have to be performed on 96 separate electrical lines at data rates as high as 15 Gb/s, potentially requiring a great deal of time and expensive equipment. To address this problem, multiplexed testing technologies are being developed that will allow any number of channels to be quickly and easily tested with standard single-channel test equipment. These multiplexed testing technologies include inexpensive high-bandwidth selector switches and multichannel pattern-generation chips. Optical testing of the PWDM mux and demux is also being addressed through the development of an automated system for quickly testing insertion loss, optical crosstalk, and spectral properties of all 48 optical channels of each device.

A final issue that must be resolved before MAUI PWDM technology can be made practical for computer interconnects is VCSEL yield. In order for a PMOSA containing 48 VCSELs to be low-cost, a very high VCSEL yield must be achieved. While improving VCSEL manufacturing processes is beyond the scope of the MAUI program, great improvements can be made in VCSEL yield, resulting in dramatic cost reductions, simply by relaxing electrical and optical performance specifications as well as operating temperature requirements. These can be addressed through proper optimization of the laser driver circuitry and PWDM optics, as well as by optimizing the thermal environment in the end application. The impact of optimizing these system parameters on VCSEL cost will be studied, and every effort will be made to optimize these parameters during the program.

ACKNOWLEDGMENT

A great many talented individuals are contributing to the MAUI program. The authors would like to thank G. Rankin, E. de Groot, A. Schmit, B. Law, H. Xia, D. Lande, J. Simon, P. Mendoza, L.-K. Chia, R. Tella, W. Gong, H. Martins, K. Djordjev, D. Lin, A. Tandon, C. Kanemura, S. Close, M. Trieu, E. Luiz, M. Amistoso, L. Windover, M. Tan, and R. Ritter for their contributions.

REFERENCES

[1] J. M. Tendler, J. S. Dodson, J. S. Fields Jr., H. Le, and B. Sinharoy, "POWER4 system microarchitecture," *IBM J. Res. Develop.*, vol. 46, no. 1, pp. 5–26, 2002.

[2] A. F. Benner. (2003, Sept.). IBM Systems Group, Poughkeepsie, NY, private communication. Available: http://www.stanford.edu/group/SPRC/Report/talks/lemoff.pdf

[3] K. Giboney, J. Simon, L. Mirkarimi, B. Law, G. Flower, S. Corzine, M. Leary, A. Tandon, C. Kocot, S. Rana, A. Grot, K. J. Lee, L. Buckman, and D. Dolfi, "Next-generation parallel-optical data links," in *Proc. 14th Annu. Meeting IEEE Lasers & Electro-Optics Society*, vol. 2, San Diego, CA, Nov. 2001, pp. 859–860.

[4] P. Rosenberg, K. Giboney, A. Yuen, J. Straznicky, D. Haritos, L. Buckman, R. Schneider, S. Corzine, F. Kiamilev, and D. Dolfi, "The PONI-1 parallel-optical link," in *Proc. 49th Electronic Components Technological Conf.*, 1999, pp. 763–769.

[5] 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, H. Xia, G. Rankin, and D. W. Dolfi, "Parallel Optical Interconnects Beyond > 100 Gb/s," *J. Lightwave Technol.*, vol. 22, pp. 2055–2063, Sept. 2004.

[6] B. E. Lemoff, "Coarse WDM transceivers," *Optics Photonics News*, vol. 13, no. 3, pp. S8–S14, Mar. 2002.

[7] B. E. Lemoff, L. A. Buckman, A. J. Schmit, and D. W. Dolfi, "A compact, low-cost WDM transceiver for the LAN," in *Proc. 50th Electronic Components Technological Conf.*, 2000, pp. 711–716.

[8] L. B. Aronson, B. E. Lemoff, L. A. Buckman, and D. W. Dolfi, "Low-cost multimode WDM for local area networks up to 10 Gb/s," *IEEE Photon. Technol. Lett.*, vol. 10, pp. 1489–1491, Oct. 1998.

[9] D. W. Dolfi, "Multi-channel optical interconnects for short-reach applications," in *Proc. 53rd Electronic Components Technological Conf.*, New Orleans, LA, 2003, pp. 1032–1039.

[10] K. C. Gupta, R. Garg, and I. J. Bahl, *Microstrip Lines and Slotlines*. Norwood, MA: Artech House, 1979, ch. 2.

[11] J. Zerbe, C. Werner, V. Stojanovic, F. Chen, J. Wei, G. Tsang, D. Kim, W. Stonecypher, A. Ho, T. Thrush, R. Kollipara, G. J. Yeh, and M. Horowitz, "Equalization and clock recovery for a 2.5–10 Gb/s 2-PAM/4-PAM backplane transceiver cell," in *IEEE Int. Solid State Circuit Conf. Dig.*, vol. 1, 2003, pp. 479–480.

[12] Maxim corporate website. [Online] . Available: http://pdfserv.maxim-ic.com/en/ds/MAX3804.pdf

[13] B. Rothermel and D. Helster. 2 mm HM Connector Noise Analysis Differential Pair Signal Placement Comparison. Tyco Electronics Circuits and Designs. [Online]. Available: http://www.tycoelectronics.com/products/simulation/files/papers/hmcna_2.pdf

[14] B. Rothermel and D. Helster. Z-Pack HM-Zd Connector Noise Analysis for Xaui Applications. Tyco Electronics Circuits and Designs. [Online]. Available: http://hmzd.tycoelectronics.com/documents/connector_noise_analysis.pdf

[15] US Conec corporate website.. [Online]. Available: http://www.us-conec.com/pages/product/ferrule/mtpfer/mainfrm.html

[16] *IEEE Standard for Carrier Sense Multiple Access With Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications-Media Access Control (MAC) Parameters, Physical Layer and Management Parameters for 10 Gb/s Operation*, IEEE Standard 802.3ae, 2002.

[17] R. R. Patel, S. W. Bond, M. D. Pocha, M. C. Larson, H. E. Garrett, R. F. Drayton, H. E. Petersen, D. M. Krol, R. J. Deri, and M. E. Lowry, "Multiwavelength parallel optical interconnects for massively parallel processing," *IEEE J. Select. Topics Quantum Electron.*, vol. 9, pp. 657–666, Mar.–Apr. 2003.

[18] M. Laha, "Coarse WDM opens the road beyond very-short-reach markets," *WDM Solutions*, vol. 3, no. 10, pp. 37–43, Oct. 2001.

[19] Teradyne corporate website. [Online]. Available: http://www.teradyne.com/prods/tcs/products/connectors/optical/

[20] J. Trezza, H. Hamster, J. Iamartino, H. Bagheri, and C. DeCusatis, "Parallel optical interconnects for enterprise class server clusters: needs and technology solutions," *IEEE Commun. Mag.*, vol. 41, no. 2, pp. S36–S42, Feb. 2003.

[21] Teradyne corporate website. [Online]. Available: http://www.teradyne.com/prods/tcs/products/connectors/backplane/gbx/

[22] D. V. Plant, M. B. Venditti, E. Laprise, J. Faucher, K. Razavi, M. Chateauneuf, A. G. Kirk, and J. S. Ahearn, "256-channel bidirectional optical interconnect using VCSEL's and photodiodes on CMOS," *J. Lightwave Technol.*, vol. 19, pp. 1093–1103, Aug. 2001.

[23] P. W. Mena, J. J. Morikuni, S. M. Kang, A. V. Harton, and K. W. Wyatt, "A comprehensive circuit-level model of vertical-cavity surface-emitting lasers," *J. Lightwave Technol.*, vol. 17, pp. 2612–2632, Dec. 1999.

[24] C. L. Coleman, Y. Chen, X. Wang, H. Welch, and B. TeKolste, "Diffractive optics in a parallel fiber transmitter module," in *OSA Trends Optics Photonics (TOPS)*, vol. 75, Diffractive Optics and Micro-Optics, OSA Tech. Dig., Washington, DC, 2002, postconference ed., pp. 249–252.

[25] P. Wijetunga and A. F. J. Levi, "The case for fiber-to-the-processor," in *Proc. Electrochemical Society*, vol. 2002-4, 2002, pp. 381–397.

[26] Intel Corp. website. [Online]. Available: http://www.intel.com/pressroom/kits/quickreffam.htm

[27] A. F. J. Levi, "Optical interconnects in systems," *Proc. IEEE*, vol. 88, pp. 750–757, June 2000.

[28] N. Savage, "Linking with light [high-speed optical interconnects]," *IEEE Spectrum*, vol. 39, no. 8, pp. 32–36, Aug. 2002.

Brian E. Lemoff (SM'04) received the B.S. and M.S. degrees in physics from the California Institute of Technology, Pasadena, in 1989 and the Ph.D. degree in physics from Stanford University, Stanford, CA, in 1994.

He is currently Project Manager of the Optical Interconnects group at Agilent Laboratories, Palo Alto, CA, where he is the Principal Investigator for the Multiwavelength Assemblies for Ubiquitous Interconnects (MAUI) program. Since joining Agilent Laboratories (formerly Hewlett-Packard Laboratories) in 1994, he has been working in the area of low-cost high-density optical interconnects, including coarse wavelength-division multiplexing (CWDM) and parallel optics. His pioneering work in CWDM for the local area network was critical in successfully gaining acceptance for CWDM in the IEEE 802.3ae 10-Gigabit Ethernet standard. During his doctoral work at Stanford University, he made key contributions in the areas of ultrashort-pulse lasers, chirped-pulse amplification, and soft X-ray lasers. He has published more than 30 papers and holds six U.S. patents.

Dr. Lemoff is a Member of the Optical Society of America (OSA) and has been a voting member of the IEEE 802.3 Ethernet Standards Committee. For his doctoral work, he received the 1995 American Physical Society Award for Outstanding Doctoral Research in Atomic, Molecular, and Optical Physics.

Mohammed E. Ali (S'96–M'01) was born in Lalmonirhat, Bangladesh. He received the B.Sc. degree in electrical and electronic engineering from the Bangladesh University of Engineering and Technology (BUET), Dhaka, in 1992 and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Los Angeles (UCLA) in 1997 and 2000, respectively.

He joined Agilent Laboratories, Palo Alto, CA, in 2001, where he is currently working as a Research Engineer in the Optical Interconnects group. His research interests include test, evaluation, integration, and system optimization of high-capacity optical links and interconnects.

George Panotopoulos was born in Athens, Greece, in 1974. He received the Diploma degree in electrical and computer engineering from the National Technical University of Athens, Athens, Greece, in 1997 and the M.S. and Ph.D. degrees, both in electrical engineering, from the California Institute of Technology, Pasadena, in 1998 and 2002, respectively.

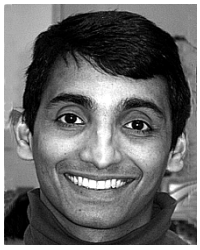
He is currently a Research Engineer with Agilent Laboratories, Palo Alto, CA, conducting research in the field of short-reach optical interconnects. Other research interests include photorefractive materials and the use of strong volume gratings for information processing and communications.

Dr. Panotopoulos is a Member of the The International Society for Optical Engineers (SPIE) and the Technical Chamber of Greece.

Graham M. Flower (S'77–M'83) was born in Cincinnati, OH, in 1957. He received B.S. degrees in mathematics, physics, and electrical engineering from the Georgia Institute of Technology, Atlanta, in 1979 and the M.S. degree in electrical engineering from the University of Florida, Gainesville, in 1984. His thesis work involved the measurement and modeling of high-voltage devices and was supported by a contract with Bell Laboratories, Allentown, PA.

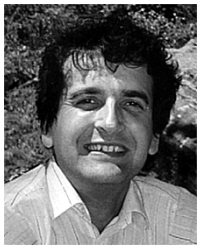
From 1979 to 1982, he was with Harris Semiconductor, Palm Bay, FL, where he was engaged in the design of data acquisition circuits. In 1984, he joined the Microwave Semiconductor Division of Hewlett Packard, San Jose, CA. Since 1984, he has been engaged in the design of communications integrated circuits. In 2000, he joined the Photonics and Electronics Research Laboratory, Agilent Laboratories, Palo Alto, CA, where he is currently a Principal Project Scientist. He has published 14 papers.

Mr. Flower is a Member of the American Physical Society, Tau Beta Pi, and Sigma Pi Sigma. In 1985, he received the Best Paper Award at the IEEE Temperature and Thermal Measurements Symposium.



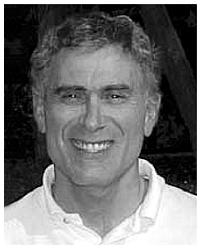
B. Madhavan (S'93–M'01) received the Ph.D. degree in electrical engineering from the University of Southern California (USC), Los Angeles, in 2000.

After working at Archway Digital Solutions, Mountain View, CA, for two years, as the System Interconnect Architect of a multiprocessor system with inter-chassis parallel–optical interconnects, he joined USC in 2002. His research interests include mixed-signal circuit design for multigigabit-per-second-per-channel electrical and optical interfaces for computer interconnects as well as high-frequency packaging.



A. F. J. Levi received the Ph.D. degree in physics from Cambridge University, Cambridge, U.K., in 1983.

After working for ten years at AT&T Bell Laboratories, Murray Hill, NJ, in mid-1993, he joined the faculty of the University of Southern California (USC), Los Angeles, where he is currently a Professor. He invented hot electron spectroscopy and the microdisk laser and carried out pioneering work on parallel fiber-optic interconnect components in computer and switching systems. His current research interests include the scaling of ultrafast electronic and photonic devices and the system-level integration of advanced optoelectronic technologies, manufacturing at the nanoscale, and the subject of adaptive quantum design. To date, he has published more than 200 scientific papers and several book chapters and is author of the book *Applied Quantum Mechanics* (Cambridge, U.K.: Cambridge Univ. Press, 2003) and holds 12 U.S. patents. Additional information about his research group can be found at http://www.usc.edu/alevi



David W. Dolfi (SM'01) received the A.B. and Ph.D. degrees in physics from the University of California, Berkeley, in 1970 and 1979, respectively. His dissertation was on the study of resonant optical solitons in sodium vapor.

He joined Hewlett-Packard (now Agilent) Laboratories, where he has worked on and managed a large number of fiber-optic-related projects (including millimeter-wave optical modulators, waveguide filters, and polarization devices) and multichannel optical subsystems, including parallel optic and coarse-wavelength-division-multiplexing transceivers. He has managed several Defense Advanced Research Projects Agency (DARPA) programs in the optical interconnect area and acts as Technical Coordinator for the Multiwavelength Assemblies for Ubiquitous Interconnects (MAUI) program. He is currently Department Manager for the Communication Technologies Department in the Photonics and Electronics Research Laboratory of Agilent Laboratories, Palo Alto, CA. He holds several patents in the area of millimeter-wave modulator design and has published or presented more than 80 journal and conference papers.

Dr. Dolfi is a Member of the Optical Society of America (OSA). He has served as a Committee Member on the Optical Fiber Communication Conference (OFC) and the European Conference on Optical Communication (ECOC) and as Chairman of the OFC Devices subcommittee.