INVITED PAPER

Nanoelectronic and Nanophotonic Interconnect

The emerging integrated circuit interconnection bottleneck may be overcome if tiny, efficient photon-transport waveguides and small transceivers for photon-electron information exchange can be developed.

By Raymond G. Beausoleil, *Senior Member IEEE*, Philip J. Kuekes, *Member IEEE*, Gregory S. Snider, Shih-Yuan Wang, *Fellow IEEE*, and R. Stanley Williams

ABSTRACT | A significant performance limitation in integrated circuits has become the metal interconnect, which is responsible for depressing the on-chip data bandwidth while consuming an increasing percentage of power. These problems will grow as wire diameters scale down and the resistance-capacitance product of the interconnect wires increases hyperbolically, which threatens to choke off the computational performance increases of chips that we have come to expect over time. We examine some of the quantitative implications of these trends by analyzing the International Technology Roadmap for Semiconductors. We compare the potential of replacing the global electronic interconnect of future chips with a photonic interconnect and see that there is in principle a four order of magnitude bandwidth-to-power ratio advantage for the latter. This indicates that it could be possible to dramatically improve chip performance without scaling transistors but rather utilize the capability of existing transistors much more efficiently. However, at this time it is not clear if these advantages can be realized. We discuss various issues related to the architecture and components necessary to implement on-chip photonic interconnect.

KEYWORDS | Architecture; bandwidth; detector; global interconnect; modulator; partition length; photonic bandgap crystal; power; resonator; transceiver; waveguide

I. INTRODUCTION

In 1965, Gordon Moore, then director of Fairchild Semiconductor's Research and Development Laboratories, first noted that the complexity of semiconductor chips had doubled every year since the first prototype integrated circuit was produced in 1959. This exponential increase in the number of components on a chip later became known as Moore's law, which has undergone several revisions over the decades. The time constant is now 18 months, but Moore's law has gone from being the doubling of the number of transistors on a chip to the doubling of microprocessor power to the doubling of computing power at a fixed cost. After nearly 50 years, the information technology industry realizes that the end of Moore's law, however formulated, is on the horizon because of several physical limits. Meindl has identified a five-level hierarchy of limits [1]:

- fundamental;
- material;
- device;
- circuit;
- system.

Fundamental limits are imposed by the laws of physics and are thus absolute and independent of material properties, device structure, circuit configuration, or system architecture. In the same work, Meindl rigorously derived the minimum energy that must be transferred in a binary logic circuit's switching transition and the minimum energy that must be transferred in a single interconnect's binary transition using two different physical models, and found that both have *precisely* the value $E_s = k_B T \ln 2$, where k_B is Boltzmann's constant and T is the absolute temperature [1].

Known material and semiconductor device limits fall orders of magnitude short of this ultimate theoretical boundary. However, based on projections derived from both Meindl's hierarchy of limits listed above and the International Technology Roadmap for Semiconductors

Manuscript received January 29, 2007; revised September 25, 2007. This work was supported in part by the Defense Advanced Research Projects Agency.
The authors are with Hewlett-Packard Laboratories, Palo Alto, CA 94304 USA (e-mail: ray.beausoleil@hp.com; philip.kuekes@hp.com; greg.snyder@hp.com; sy.wang@hp.com; stan.williams@hp.com).
Digital Object Identifier: 10.1109/JPROC.2007.911057

(ITRS),[1] a 10-nm minimum feature size could support a chip with more than 100 billion transistors sometime beyond 2020, if we can:

- develop 10-nm-scale fabrication technologies that will circumvent the expected exorbitant manufacturing costs arising from optical lithographic technologies;
- devise effective methods to handle the necessarily large number of defective components that will be present in such circuits;
- handle the heat dissipation from the projected power densities;
- invent the necessary global interconnect technology to effectively complement 10-nm transistors [2].

Although it has received comparatively little attention to date, perhaps the most problematic of these advances is the fourth: even present day integrated circuits have evolved to the point where the global interconnects of the sub-90-nm feature size circuits have become stringent limitations to performance. As a result—and because applied scientists and engineers have been so successful at continually pushing transistor fabrication technology to the next node on the ITRS—a major uncertainty (e.g., red table entries) in the roadmap has become the global interconnect. In fact, global interconnect is the primary constraint on CPU clock speed and already consumes a significant fraction of the total electrical power dissipated by a chip.

We believe that the global interconnect problem will not be solved by copper, carbon nanotubes, or code-division multiple-access radio frequency (RF) because of bandwidth to power and signal integrity limitations. Instead, it is possible that knowledge systems in the future will be based on monolithically integrated nanoscale electronic-photonic circuits, with the information processing primarily relying on electrons and the majority of the information transfer (above a particular architecture-dependent length scale measured in tens of micrometers) accomplished using photons. Such systems will be dramatically more capable and energy efficient than solely electronic systems. Similar visions have been expounded over the past two decades, but recent advances in the area of nanophotonics—such as photonic bandgap and negative index materials—have made possible the design and manufacture of integrated electronic-photonic systems that utilize existing fabrication plants for silicon integrated circuits, augmented with technology that those facilities must adopt to remain on the ITRS roadmap. If such systems can be realized, they have the potential to extend the 18-month doubling of computing capability at a fixed cost for many decades into the future.

II. THE ITRS INTERCONNECT ROADMAP

For the past decade, the ITRS and its predecessor has been the blueprint that the entire world semiconductor industry uses for introducing new products (DRAM, general-purpose processors, and application-specific integrated circuits) into the market. The ITRS provides direction for all companies that participate in the process and, most importantly, points out the areas where research is urgently needed in order to overcome the biggest obstacles to a particular generation of product. Nearly 100 organizations and companies worldwide participate in formulating the roadmap, which is now completely reformulated every other year (odd years) and revised in the intervening (even) years. In the following discussion, of roadmap issues, all data are taken from the 2005 edition of the ITRS and/or the 2006 revision.

In these most recent versions of the ITRS, global interconnect is seen as a major challenge to the industry. This issue is highlighted throughout the roadmap document, with frequent comments such as: "This dramatic reversal from performance limited by transistor delay to performance limited by interconnect delay shows clearly the inadequacy of continuing to scale the conventional metal/dielectric system to meet future interconnect requirements." Moreover, the ITRS Interconnect projections have become progressively more pessimistic with each new revision (e.g., comparing performance metrics in the 2003 and 2005 editions).

The interconnect stack of a CMOS integrated circuit is divided into three regions: *metal 1* is the direct connection to the semiconductor level, *intermediate* is the next four to eight levels of interconnect (depending on the technology year of introduction), and *global* is the top two to five levels in the hierarchy. The *global interconnect* is responsible for conducting information over "long distances" on the chip, and to and from the chip edges. The global interconnect is becoming a major bottleneck for chip design and operation. The relatively slow propagation of electrical signals down the global interconnect and the relatively long distances that need to be covered are the primary limitations to the clock speed of a processor chip, and this problem is getting worse as feature sizes (e.g., wire widths) are getting smaller. The global interconnect can also consume a significant fraction of the power used on a chip built using *current* architectures, and this issue will worsen as the wire widths sizes shrink.

The problem is manyfold. First, the RC time constant of an electronic interconnect line is proportional to the square of the length of that line, since both R and C are proportional to the wire length. As the clock frequency increases, this places a severe constraint on how long global interconnect lines can be, especially for any synchronous system. Secondly, as the widths of the interconnect wires decrease, the RC time constant per unit length is beginning to increase hyperbolically because the *resistivity* of the copper interconnect wires rises with decreasing wire cross-section. (This increase in the effective wire resistivity noted by the ITRS is the result of the increasing importance of surface and grain boundary

[1] http://www.itrs.net/.

scattering as the surface to volume ratio of the wires increases.) The only known way to partially offset this increase in RC is to use an insulator with a lower dielectric constant, but the industry has been stuck with a dielectric in the range of 3.3–3.6 and has found that "the slower than projected pace of low-κ dielectric introduction" was one of the central issues for the ITRS. Even if the dielectric constant of the insulating layers can be reduced to 2.4, this would still only help the situation by 33%, while the problems with RC are essentially diverging with reduced feature size. This inability to decrease the capacitance is also a major problem for dynamic power dissipation on the chip, which obeys the relation $P_{\mathrm{dyn}} \approx CV^2 f$, where C is the capacitance of the system being charged, V is the voltage to which it is charged, and f is the charging frequency.

Table 1(a) and (b) contains data collected from the 2006 ITRS revision (highlighted in blue) and a set of quantities derived from the roadmap data (highlighted in green) to provide some guidelines for considering the desirability of a hybrid electronic-photonic global interconnect. One must recognize that, for any particular year, the ITRS presents a nominally consistent set of requirements for a hypothetical technology that acts as a benchmark for the industry—these are not predictions for any given product or company but requirements based on reasonable guesses about what will be possible. It is also a compromise among a large number of experts in the different technology areas and institutions that contribute to designing and manufacturing chips. At this time, there are no known manufacturing solutions that will provide all the ITRS requirements for global interconnect in 2010 and beyond.

There are several issues to note about the roadmap. First, it calls for the on-chip clock frequency of a high-end processor to increase from 9.3 GHz in 2007 to 73.1 GHz in 2020, during which time it is anticipated that the effective *resistivity* of Cu wires will *double* and the upper allowable power dissipation of the chip must remain constant (at essentially 200 W). These are physical constraints that will be extremely difficult if not impossible to satisfy simultaneously, and will require at the least a number of major architectural changes. Note also that both the physical area of the chip and its number of input/output ports (1024) are to remain constant, which means that there will be a tremendous bottleneck for getting information onto and off the chip. The intrinsic switching delay of the transistors (called τ in the ITRS, and considered to be the most fundamental physical property of a transistor) is already in

Table 1(a) Photonic Interconnect Comparison for 2006 ITRS Goals—Near Term

Estimated Year of Production	2007	2008	2009	2010	2011	2012	2013
High Performance MPU properties							
MPU/ASIC metal 1 1/2 pitch (nm)	68	59	52	45	40	36	32
V_{dd} (high performance) (V)	1.1	1	1	1	1	0.9	0.9
On chip local clock (MHz)	9,285	10,972	12,369	15,079	17,658	20,065	22,980
MTransistors per cm^2	357	449	566	714	899	1,133	1,427
NMOS intrinsic delay τ (ps)	0.64	0.54	0.46	0.4	0.34	0.29	0.25
Clock Period (ps)	108	91	81	66	66	50	44
Global Electronic Interconnect							
Minimum global pitch (nm)	210	177	156	135	120	108	96
Conductor resistivity (μΩ-cm)	2.73	2.87	3	3.1	3.22	3.39	3.52
RC for 1mm global wire (ps)	209	316	410	523	687	787	977
bit hop length: RC = clock period (μ)	719	537	444	355	310	252	211
ave electron bit hops to cross chip	24	33	39	49	56	69	83
ave delay or latency (ns)	2.6	3.0	3.2	3.3	3.7	3.5	3.6
MTransistors within bit hop radius	6	4	4	3	3	2	2
Power for bits at clock speed (mW)	1.36	1.02	0.91	0.83	0.72	0.55	0.50
Energy per bit per hop (pJ)	0.147	0.093	0.074	0.055	0.048	0.028	0.022
Bit flux/Watt (**Tbit**/sec/cm/W)	0.140	0.166	0.172	0.186	0.186	0.261	0.276
Global Photonic Interconnect							
Meindl partition length (μ), N=1	1,287	1,184	1,115	1,010	933	876	818
Light speed * clock period/3 (μ)	10,800	9,100	8,100	6,600	6,600	5,000	4,352
ave hops to chip edge for photons	2	2	2	3	3	4	4
Mtransistors within photon hop radius	1,308	1,168	1,166	977	1,230	889	849
ave delay or latency (ns)	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Energy per bit to cross chip (aJ)	90	90	90	90	90	90	90
Potential Bit flux/Watt (**Pbit**/sec/cm/W)	11	11	11	11	11	11	11

(a)

Table 1(b) Photonic Interconnect Comparison for 2006 ITRS Goals—Long Term

Estimated Year of Production	2014	2015	2016	2017	2018	2019	2020
High Performance MPU properties							
MPU/ASIC metal 1 1/2 pitch (nm)	28	25	22	20	18	16	14
V_{dd} (high performance) (V)	0.8	0.8	0.8	0.7	0.7	0.7	0.7
On chip local clock (MHz)	22,980	33,403	39,683	39,683	53,207	62,443	73,122
MTransistors per cm^2	1,798	2,265	2,854	3,596	4,537	5338	7193
NMOS intrinsic delay τ (ps)	0.21	0.18	0.15	0.13	0.11	0.1	0.08
Clock Period (ps)	44	30	25	25	19	16	14
Global Electronic Interconnect							
Minimum global pitch (nm)	84	75	66	60	54	48	42
Conductor resistivity ($\mu\Omega$-cm)	3.73	3.93	4.2	4.39	4.58	4.93	5.38
RC for 1mm global wire (ps)	1353	1601	2210	2794	2983	4064	5795
bit hop length: RC = clock period (μ)	179	137	107	95	79	63	49
ave electron bit hops to cross chip	98	128	164	184	220	279	360
ave delay or latency (ns)	4.2	3.8	4.1	4.6	4.1	4.5	4.9
MTransistors within bit hop radius	2	1	1.0	1.0	0.9	0.7	0.5
Power for bits at clock speed (mW)	0.34	0.33	0.31	0.21	0.20	0.18	0.17
Energy per bit per hop (pJ)	0.015	0.010	0.008	0.005	0.004	0.003	0.002
Bit flux/Watt (**Tbit**/sec/cm/W)	0.349	0.390	0.390	0.509	0.614	0.614	0.614
Global Photonic Interconnect							
Meindl partition length (μ), N=1	818	679	623	623	538	496	459
Light speed * clock period/3 (μ)	4,352	2,994	2,520	2,520	1,879	1,601	1,368
ave hops to chip edge for photons	4	6	7	7	9	11	13
Mtransistors within photon hop radius	1,069	637	569	717	503	430	422
ave delay or latency (ns)	0.09	0.09	0.09	0.09	0.09	0.09	0.09
Energy per bit to cross chip (aJ)	90	90	90	90	90	90	90
Potential Bit flux/Watt (**Pbit**/sec/cm/W)	11	11	11	11	11	11	11

(b)

the subpicosecond time domain, or nearly two orders of magnitude faster than the 2020 clock period for a chip (which is limited by the global interconnect). Given that actual clock speeds are nowhere near the ITRS projection, in large part because of power dissipation and consequent heating, it is apparent that the operating speed of chips has already hit a significant roadblock that is only partially compensated at present by using multicore architectures.

Using information available from the roadmap, it is possible to estimate the order of magnitude of several relevant circuit properties. For example, in 2007, the length of a global interconnect line with an RC time constant equal to the clock period is 719 μm, but this distance decreases to only 49 μm in 2020 when the pitch of the global interconnect is projected to be 42 nm. The major impact of this decrease is that the number of transistors accessible within a single bit "hop" through the global interconnect would *decrease* from 6 million in 2007 to only 0.5 million in 2020. Therefore, the transistor density will not increase rapidly enough to keep the number of addressable transistors within one clock period. For all these systems, a series of repeater units must be incorporated into the global interconnect at approximately the bit hop spacing in order to transmit a signal from an arbitrary point to or from the edge of the chip. For a synchronous system, the average number of clock periods required to transmit a bit from one edge of the chip to the opposite side increases from approximately 24 in 2007 to 360 in 2020. This also has a major impact on the chip architecture, since a substantial portion of the area of the chip will be required for the active circuitry and the vias for the repeaters, and their existence will create major routing and avoidance issues for the global and intermediate levels of interconnect. One of the potential architectural consequences is that the number of cores, or actual processing units on a single chip, could increase exponentially with time from two to four now at traditional Moore's law rates to 500–1000 less capable cores in 2020. Thus, each chip will be a "multiprocessor unit" (MPU) and will require exquisitely timed intra- and interchip communication to pass information among processors and shared memory.

The power and energy requirements of the global interconnect and the repeaters are also major issues and, depending on the architecture, could well consume more than half the power of the chip. The estimated maximum power use in a single global interconnect line in 2007 is 1.4 mW—given that there are approximately 60 000 such lines in a chip, it is easy to see that without careful management, the global interconnect and their

repeaters could easily dominate the power dissipation in a chip. Although the power for bit transmission in the global interconnect in 2020 will only be 0.17 mW per line, the sheer number of such lines (many millions) would require kilowatts of power for the global interconnect, which is beyond any current cooling technology to handle. Of course, the ITRS assumes that it will be possible to decrease the dielectric constant of the insulating layers in the chip to 2.0—if that has not happened, the theoretical power dissipation would be much worse. Another issue to note is that the energy required to charge a global interconnect line up to V_{dd} is currently about 0.15 pJ (or 0.8 MeV), which is roughly a factor of 10 000 times larger than the switching energy of a transistor. To meet the goals of the ITRS, this energy will have to decrease to 2 aJ by 2020, which will be very difficult if not impossible to achieve with known electronic materials.

III. THE NANOPHOTONIC PARTITION LENGTH

Could replacing some or all of the electronic global interconnect improve this situation? Naeemi *et al.* [3] have proposed a "partition length" above which information is more efficiently transported by photons rather than electrons. (Naeemi *et al.* use a different set of assumptions about the wire conduction mechanism and physical constants of materials than the ITRS committee, so calculations based on these assumption are not necessarily quantitatively consistent with the properties calculated from the ITRS tables, but the trends are still the same.) By assuming that board-level global interconnect wires operate in the RLC regime, where bandwidth is limited by the skin depth, they obtain an expression for the partition length L_{part} given by

$$L_{part} = W_{min}\sqrt{\frac{K_0}{f_{max}}} \tag{1}$$

where W_{min} is the minimum wire width (and, by assumption, waveguide width) available at the board level $K_0 = 6.152 \times 10^{16}$ Hz and depends on the electronic conductor material and assumed conductivity mechanism, and f_{max} is the maximum modulation frequency for either the electrical or optical signal. We have plotted (1) as a function of frequency for a variety of wire/waveguide widths (where the wire and waveguide widths are assumed to be equal) in Fig. 1. Note that above 20 GHz, the partition length for a wire/waveguide width of 1 μm is less than 2 mm, a distance roughly equal to the "bit hop length" calculated from the ITRS for the maximum transmission distance of a bit in the global interconnect. Given that chips have dimensions of centimeters, utilizing photons should improve communications bandwidth even on current chips, and depending on the architecture can dramatically shorten latency by decreasing the number of clock cycles needed to transport bits to their intended destination. Global optical interconnect could deliver or accept bits to within one electronic "bit-hop" length from any position on a circuit. However, since the intrinsic delay of the field-effect transistors is less than a picosecond, the clock speed of a chip could be much faster in any given chip generation if the longest

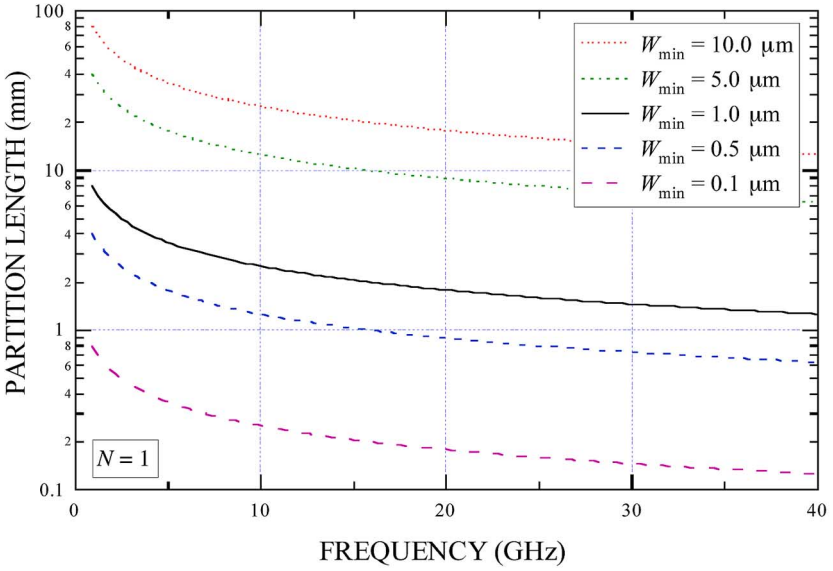

Fig. 1. *The partition length defined in [3] and given by (1) for $N = 1$ channel per optical waveguide. Note that above 20 GHz, the partition length for a wire/waveguide width of 1 μm is less than 2 mm.*

electronic bit-hop length is shortened (especially since the delay is proportional to the square of the length). This statement is self-consistent, since the Meindl partition length is inversely proportional to the square root of the clock speed. Note that in Table 1, both the wire width and the operating frequency change with the estimated year of technology introduction, e.g., they reflect the technology projection for the global interconnect wiring pitch and desired clock speed, so that Fig. 1 and Table 1 address different issues. In Table 1, we want to know the Meindl partition length for each annual technology projection compared to a single optical carrier frequency in a waveguide with a width of one micrometer.

The real promise of a photonic global interconnect becomes clear when we realize that Naeemi *et al.* have implicitly assumed that each optical waveguide carries only one channel of information [3]. Instead, if we explicitly assume that each chip incorporates N distinct optical wavelength-division multiplexing (OWDM) transceivers, so that each optical waveguide can carry N signal channels, then we have

$$L_{\text{part}} = \frac{W_{\min}}{N}\sqrt{\frac{K_0}{f_{\max}}}. \tag{2}$$

We have plotted (2) as a function of frequency for a wire/waveguide width of one and a variety of values of N in Fig. 2. Now we note that above 20 GHz, the partition length for a waveguide width of 1 μm carrying 30 channels is less than 60 μm, a distance roughly equal to the electronic "bit hop length" in 2020. We conclude that the use of OWDM can allow the global interconnect system to connect on-chip cache to the CPU at speeds much higher than purely electrical connections would allow. Since there are no capacitive charging and discharging losses for photonic bits, the energy required to transmit a bit over a long distance (on or off chip) is much lower for the photonic interconnect. The major energy requirement is to convert the photonic signal to an electronic signal to carry the bit over the last bit-hop to its eventual destination on chip (or conversely, to transmit the bit electronically on its first hop to a modulator that impresses the bit onto an existing photon stream). Thus, if photonic gateways are present at the same density as repeaters on a purely electronic circuit, and the operating energy of a photonic transceiver is approximately the same as for a repeater, the relative energy dissipation will be approximately the inverse of the average number of bit hops for the purely electronic case, e.g., ~1/360 for the technology anticipated in 2020. This represents a significant saving in power for long-distance communication on the chip and enables an engineering tradeoff in which increasing the density of photonic gateways can both decrease the power dissipation and increase the clock speed for the chip—a level of flexibility not available for all electronic systems.

We can also see by multiplying the speed of light in a dielectric medium with $\kappa = 3$ by the ITRS clock period that the clock frequency is so fast that even light cannot make it all the way across the chip within one clock period. Photonic solutions for global interconnect need to account for that issue. Thus, the repeaters required for electronic global interconnect are replaced by photonic transceivers.

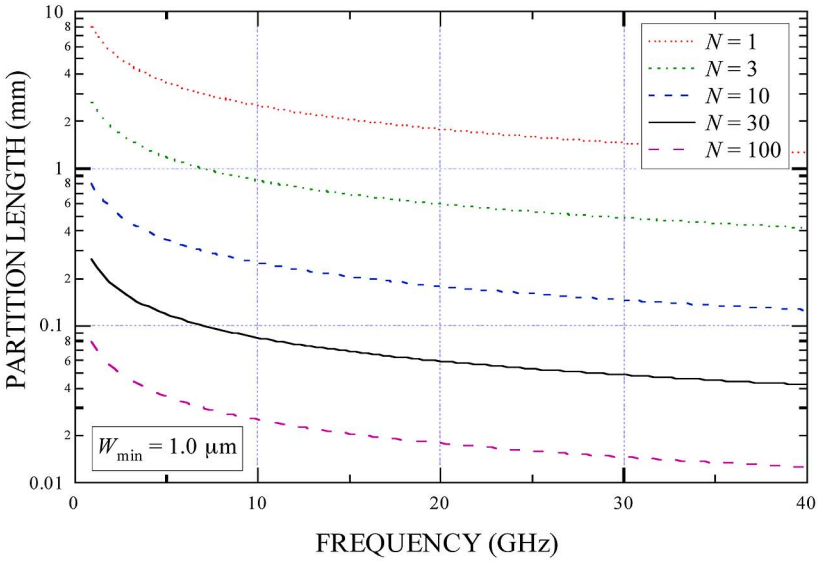

Fig. 2. *The partition length given by (2) for a wire/waveguide width of 1 μm. Note that above 20 GHz, the partition length for a system using 30 wavelength channels is less than 60 μm.*

The power and chip-area requirements of those transceivers are an important issue that must be carefully considered, since they could completely outweigh any benefits from photonic interconnect. Generating the photons on the chip will require a huge amount of power, which would make optical photonics less attractive than an RF communications strategy, which is being seriously monitored by the ITRS. However, if the photons are generated off chip, then a series of modulators and receivers, or transceivers, can be used to impress information onto the photonic bitstream or to convert photons to electrons. This puts significant constraints on the operating power and size of the transceivers.

The higher the density of the photonic transceivers, the faster the on-chip clock can run. There is also a significant energy improvement within the global interconnect itself. Assuming that a bit of information is carried by 500 photons with an energy of 0.83 eV per photon (infrared), the total photonic energy of each transmitted bit (66 aJ) is orders of magnitude lower than for electronically transmitted bits (Table 1). Thus, most of the energy required to transmit a bit is determined by the first or last electronic bit hop. Photonic interconnect at the global level can drastically cut the power dissipation in the global interconnect and perhaps also in the intermediate interconnect, depending on how far down the interconnect stack the optical circuitry can extend.

Using optical interconnect can greatly increase the data bandwidth over an all-electrical interconnect. In any particular generation of chip, this can both substantially increase the on-chip clock speed of the chip while at the same time decreasing its total power dissipation, depending on the engineering tradeoffs desired. This can provide a substantial improvement over the performance of chips as represented in the ITRS, as well as enable future generations of chips that are completely out of the realm predicted by the roadmap.

The electronics industry has been driving the scale of device features to the de Broglie wavelength of the electron for 50 years, which has enabled the exponential improvement in device performance with time. Photonic technologies have not yet seen a similar acceleration of performance improvements; in current optical information technology, the term "high level of integration" implies that functionality has been added during postprocessing or packaging. Although optical IT is ultimately limited by the wavelength of light and the energy required to generate photons, photonic crystals [4] may allow researchers to reach optoelectronic feature sizes as small as $\lambda/10$ and devices capable of operating at light levels of only a few photons. The ultimate goal of a photonic interconnect research effort is the demonstration of giant nonlinear optical effects at extremely low light levels in structures that have been manufactured using economically promising methods, such as nanoimprint lithography. In this sense, a "high level of integration" of photonic functional devices could occur in nanoscale monolithic fabrication rather than in postprocessing, ushering in an era of Moore's law for optics as discussed in various nanophotonic roadmaps [5], [6].

IV. GLOBAL PHOTONIC INTERCONNECT ARCHITECTURES

The desirability of both off- and on-chip photonic interconnect has been discussed for more than 20 years, but to date it has not been possible to actually implement it because the necessary components were much too large to integrate on Si chips and required far too much power to be practical. However, recent developments now make photonic interconnect look feasible, even as the feature sizes of integrated circuits move into the few tens of nanometer range.

A. Logical-to-Frequency Addressing

The purely electronic architecture of existing microcircuits is the origin of the long-term roadblocks caused by global interconnect. For example, in the typical memory multiplexer/demultiplexer architecture shown in Fig. 3, an access request by a CPU for a particular logical memory address must ultimately be converted into a physical memory address: the multiplexer must "know" where every available byte of RAM is physically located. Ultimately, then, electrical wires must be connected from a high-level multiplexer to every byte of physical RAM. In principle, stages of submultiplexers can be designed (each of which communicates with one multiplexer above and many below) but the buffering and switching that must be incorporated into any such architecture limits the complexity and performance of the system. The problem only gets worse as the components shrink deep into the nanoscale, and the intrinsic capacitance of the multiplexing electronics greatly exceeds that of the RAM: the effective RC time constant of the multiplexed

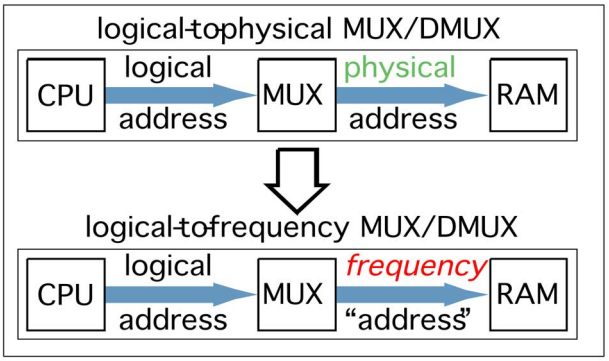

Fig. 3. *Schematic diagram of two different memory multiplexing architectures. The first (logical-to-physical) requires that each subunit of RAM be electrically connected to the CPU through a physically deterministic multiplexer. A possible approach removes this determinism while improving performance: the multiplexer associates a unique frequency "fingerprint" with each subunit of RAM and can operate without any specific information regarding the physical location of a particular subunit of RAM.*

circuit path becomes too large to permit high-speed operation. Furthermore, the information carrying capacity of each nanowire in the circuit—proportional to the wire area/length^2—decreases as the system is brought down to the nanoscale, and closely spaced wires cannot be accessed at high speeds without inducing currents in the adjacent wires.

An alternate photonic scheme is to use the "logical-to-frequency" multiplexing concept shown in Fig. 3, where:

1) the total RAM is subdivided into many identical RAM subunits, each with its own local multiplexing system;
2) each RAM subunit is assigned a unique frequency "fingerprint";
3) a two-way optical data transfer bus is provided between a central signaling system controlled by the CPU and each RAM subunit.

It is important to note that these optical data pathways can be driven in parallel and operate at the group velocity of light in the optical substrate, providing high performance with low system complexity. This approach removes the physical determinism implied by Rent's rule while improving performance: the multiplexer associates each frequency "fingerprint" with a particular subunit of RAM and can operate without any specific details regarding the physical location of that subunit. As long as each subunit has access to the full data stream, the relevant information can be extracted and processed. Even if individual nanocircuits cannot operate at high speed, the high degree of parallelism offered by this nanophotonic OWDM allows the circuit—composed of many such tiles—to operate at extraordinarily high speeds. (All of these ideas—described here in the context of nanoscale RAM circuits—can be extended to nanoscale logic and sensors with trivial modifications [7], [8].)

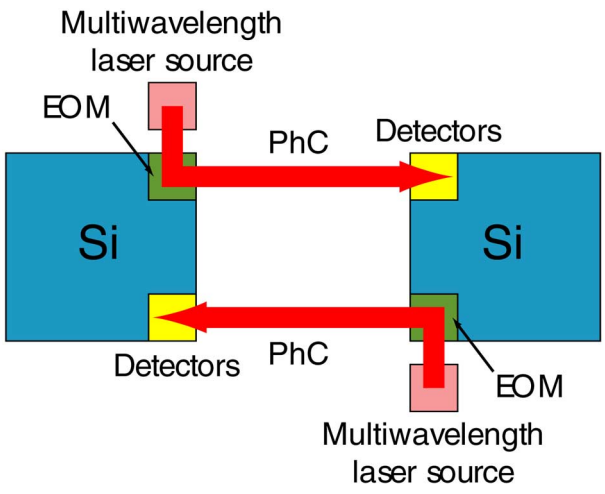

Fig. 4. *Conceptual architecture of a nanophotonic data transfer system for chip-to-chip communication. Lasers are typically difficult to fabricate on silicon; instead one can use off-chip multiwavelength laser sources and on-chip electrooptic modulators (EOMs) and photodetectors.*

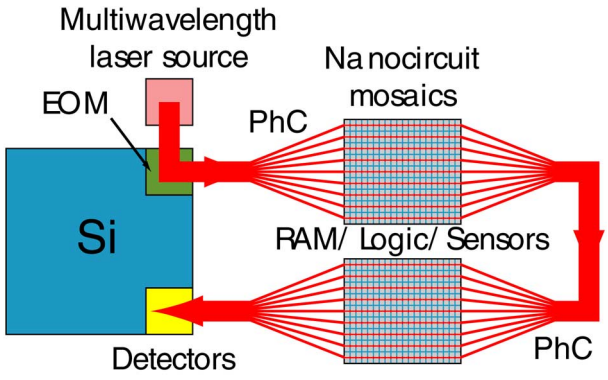

Fig. 5. *Conceptual architecture of a nanophotonic data transfer system for chip-to-RAM/sensor/logic communication. The signal can be spatially multiplexed using 3 dB couplers.*

This concept can be implemented directly using nanophotonic structures, as shown in Figs. 4 and 5. In Fig. 4, we show a conceptual architecture of a nanophotonic data transfer system for chip-to-chip communication. Laser sources are typically difficult to fabricate on silicon, so one can use off-chip multiwavelength laser sources (such as those available for telecommunication applications at $\sim$1.5 μm), and on-device electrooptic modulators and photodetectors, to imprint a signal onto an optical data stream propagating through one or more photonic bandgap crystal (PhC) waveguides from one chip to another. The receiving chip decodes these data using an appropriate number of photodetectors. Similarly, Fig. 5 shows a nanophotonic data transfer assembly that allows a silicon chip to communicate with memory, sensor, and/or logic arrays composed of nanocircuits. The modulated signal emerging from the chip can be split using 3-dB couplers into a number of waveguides, each of which will provide encoded data in parallel for all nanocircuits. Each collection of nanocircuits will sample the optical signal, disambiguate the contents, extract any instructions or data intended for that particular set, and then—through a local electronic multiplexer—either read or write bits to/from that set. The high degree of parallelism available through this technique is best applied to large nanoscale memory, sensor, and/or logic arrays generating or accessing data at rates of 100 Gb/s or more; the power expense of creating photons is the price paid for this high performance (see the comparisons for electrical and photonic bit flux/watt in Table 1). This expense can be amortized over an entire server for cost effectiveness, using the off-chip (or off-board) laser sources.

B. Data Transfer Among Tiles and Mosaics

Fig. 6 shows a schematic diagram of nanophotonic data transfer components and their interfaces with arrays of (for example) nanoscale RAM, logic or sensors. The fundamental architectural building block of this system is a "tile," which is a small region incorporating nanocircuits of

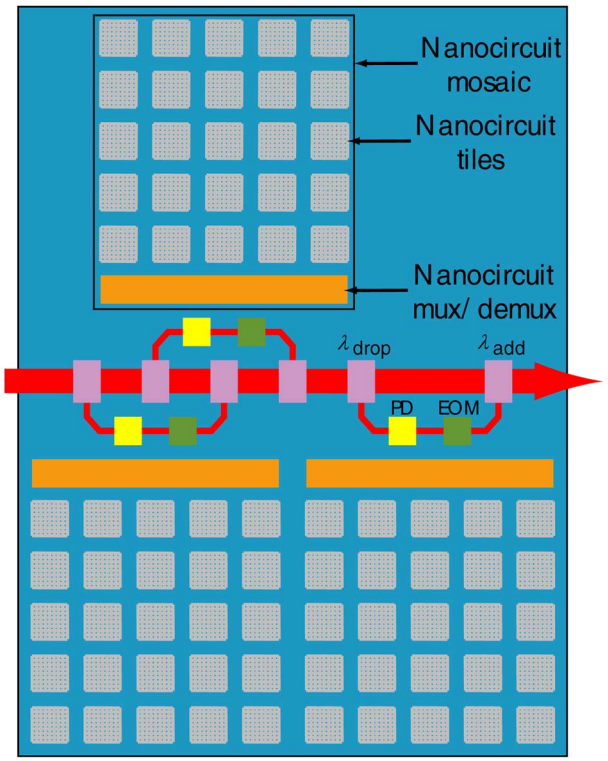

Fig. 6. ***Schematic diagram of WDM nanophotonic interconnect components and their interfaces with mosaics of molecular RAM/logic. Data for each mosaic are encoded on a specific wavelength and then extracted and (possibly) demodulated using a wavelength drop and a photodiode. Data is encoded onto a wavelength using an electooptic modulator.***

a single device type [7], [8]. These tiles are collected into larger structures called "mosaics," which perform some useful function and are connected to even larger structures via an intermediate-layer multiplexer circuit [7]. We adopt an OWDM architecture in this example, although optical time-domain multiplexing (OTDM) could be used instead [9]. OWDM divides an optical signal into "virtual fibers" that may be separately encoded and decoded by photodetectors and modulators that are connected to the waveguides via wavelength-specific directional couplers (drops and adds). As implied by Figs. 4 and 5, we use one or more multifrequency lasers to provide many narrow-band coherent channels. Data for each mosaic are encoded onto a specific wavelength by modulating the light at that wavelength over a bandwidth B that is smaller than the frequency separation between adjacent channels. This modulation can be accomplished at a central processor (for writing to the mosaic) or at the mosaic itself (for reading by the CPU). The detector and modulator for a particular channel are segregated from the primary waveguide by a "wavelength drop," a coupler that shunts a large (approaching 100%, if possible) fraction of the power at that channel's wavelength onto a local waveguide. The corresponding signal can be demodulated by a photodetector and then written to the memory mosaic by the local intermediate demultiplexer.

V. COMPONENTS FOR LOCAL PHOTONIC INTERCONNECT

Fig. 7 illustrates a template for the two-dimensional PhC components. A p-i-n material (e.g., SOI or InGaAs) can be fabricated onto a glass substrate, with a lattice of holes providing an optical bandgap that shapes and directs propagating electromagnetic radiation [10], [11]. Traditionally, these structures are fabricated using charged-beam (e.g., electrons or focused ions) lithography, but more economical processing methods, such as nanoimprint lithography, are very promising [12]. Intentional defects can be introduced into the lattice to produce particular optical components: a point defect is a resonator, and a line defect is a waveguide. The transverse mode diameter of an optical field propagating along a PhC waveguide can be as small as $\lambda/3n$, while the mode volume of a PhC nanoresonator can be as small as $2(\lambda/n)^3$.

Typically, PhC structures rely on periodic spatial variations in refractive index to provide both confinement and coupling. Fig. 8 shows a possible implementation of a nanophotonic data transfer junction using PhC technology. The waveguides and wavelength drops are coupled through the evanescent fields surrounding these features; the coupling can be strongly enhanced by fabricating the drops as point defects (nanoresonators) with high Q factors. The resonant frequency of a nanoresonator at a particular location in the integrated structure can be statically tuned by adjusting either the refractive index of the p-i-n material or the spacing and/or size of the lattice of holes during the fabrication process. The Q of the resonator can be modeled using finite-difference time domain numerical methods and has been designed with an unloaded $Q > 10^6$ [13], which makes loaded Q's in the 10^3–10^4 range seem

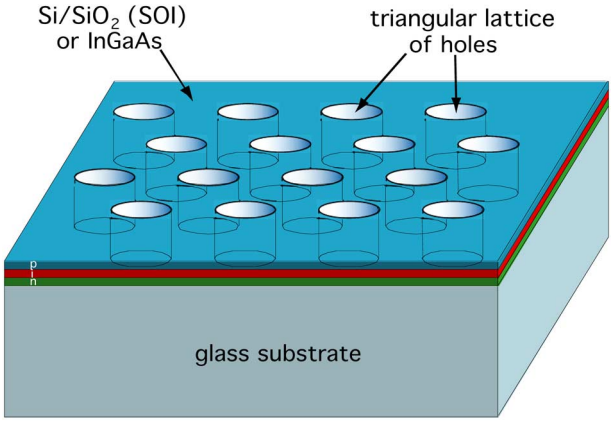

Fig. 7. ***Triangular-lattice photonic bandgap crystal fabricated in a $p-i-n$ layer on a glass substrate.***

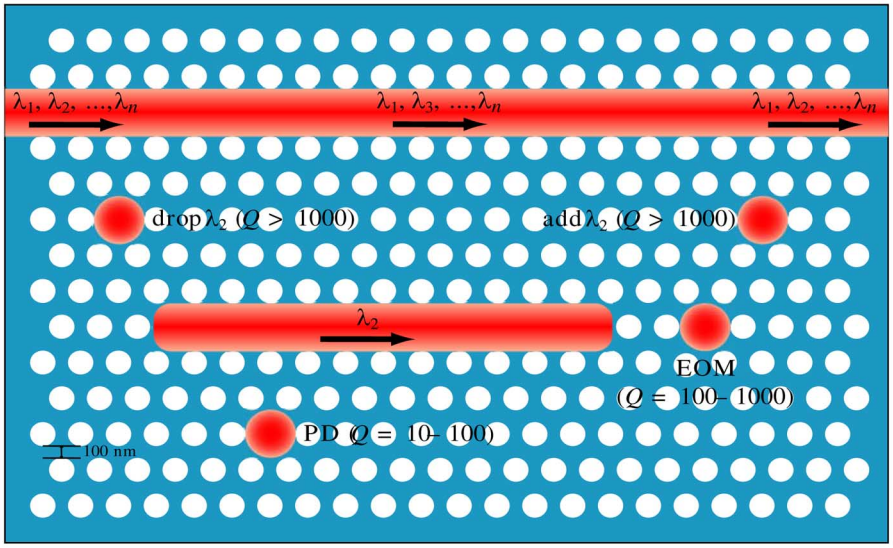

Fig. 8. *Possible implementation of a nanophotonic data transfer junction using PhC technology. The waveguides and drops are coupled through the evanescent fields surrounding these features; the coupling can be strongly enhanced by fabricating the drops as point defects (nanoresonators) with high Q factors.*

reasonable. The resonant transmission bandwidth of a resonator with quality factor Q and resonant frequency $\nu_0 = c/\lambda_0$ is $\pi\nu_0/Q \approx 500\ \text{THz}/Q$; the resonant power transmission fraction of this resonator is $1-1/Q$, and the nonresonant insertion loss is $1/Q$ (assuming a configuration similar to that shown in Fig. 8, where nonresonant light can propagate freely along the waveguide). In other words, if we require that $\pi\nu_0/Q > B$, a fraction $1-1/Q$ of the light in the channel at wavelength λ_0 will be redirected and transmitted through the wavelength drop while an identical fraction of the light at other wavelengths will continue to propagate along the waveguide. Therefore, as a design choice, we constrain the number of nanophotonic junctions (i.e., drop/add pairs) to less than $Q/2$ so that the net nonresonant loss per waveguide will be less than $1-(1-1/Q)^Q \approx e^{-1} = 63\%$. This choice guarantees that the last drop/add pair on the waveguide can extract at least 37% of the light originally entering the waveguide at the corresponding wavelength. Of course, other choices are possible given different design goals.

A. Integrated Optoelectronic Components

These considerations also lead to an improved design for an integrated PhC photodetector and demodulator. The intrinsic capacitance of a silicon photodetector with an area of A square micrometers is approximately $100A$ aF. A typical transverse dimension of the photodetector shown in Fig. 9 is 100–150 nm, so that the corresponding intrinsic capacitance of the doped region is 2 aF. This capacitance is low enough that the current fluctuations due to Johnson noise should be insignificant. Hence, as shown below, we expect the bit error rate (BER) of an integrated device to be dominated by the statistics of the laser source. The small size of the detector implies that the fraction of the light absorbed by the active area of the detector will be quite small. We can compensate for this reduced absorption by incorporating the doped region into a resonant cavity with a Q of 10 to 100. It has been shown that such a resonant cavity enhancement method can dramatically increase the efficiency of broadband silicon photodetectors [14], and we can also alloy the cavity with Ge to increase the intrinsic absorption of the material. With an appropriate choice of Q to impedance-match the optical input losses of the cavity to the internal absorption loss of the detector, it should be possible to increase the detection efficiency to 50%. Similar considerations can be applied to the design of a resonant cavity enhanced (RCE) modulator; using electrooptic techniques, modulation depths as high as 90% can be obtained if $Q > 1000$.

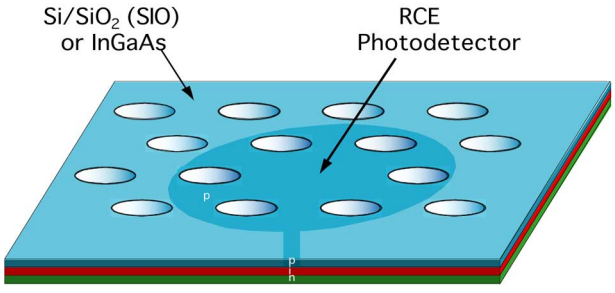

Fig. 9. *Example of an RCE photodetector for demodulation of an encoded wavelength channel. The weak absorption of the small doped region is compensated by allowing the incident radiation to pass through the active area multiple times. Similar considerations can be applied to the design of a PhC modulator.*

B. Integration of Optical and Electronic Components

How far can this massively parallel architecture be scaled? Let us assume that we have K identical waveguides, each comprising a total of $N \leq Q/2$ drop/add pairs, with a nominal maximum value of N of about 100. Now, the nth mosaic (supported by the nth drop/add pair of nanoresonators) on every waveguide has the same resonant frequency ν_n. Therefore, to distinguish the signal intended for the nth mosaic on waveguide k from those of the other $K-1$ waveguides, we must encode the signals for all K mosaics onto the single channel with wavelength λ_n. Hence, the total number of memory and/or logic mosaics supported by this system of interconnects is NK. Clearly, mosaic k on each waveguide must be assigned a unique encoding (or "fingerprint") so that the signal intended for mosaic k on channel n can be distinguished from the other $K-1$ mosaics encoded on channel n. Suppose that the maximum data I/O rate that can be supported by a given mosaic is B. Then, assuming that the modulation bandwidth that can be applied to each channel is at least KB (and that the transmission window of the drop/add nanoresonators satisfies $\pi\nu_0/Q > KB$), the total throughput of the system is NKB. This approach allows us to access all mosaics in parallel without any foreknowledge of the physical location of *any* mosaic. Additional multiplexing is achieved simply by adding mosaics.

If we desire a BER of 10^{-22} and assume RCE detection efficiency of 50%, then a total of 500 photons/bit (or 250 detected photons/bit) are required to transmit bits in photonic interconnect. If we wish to operate at a full throughput of 1 Tb/s at a wavelength of 1.55 μm, then the minimum power needed (incident on all photodetectors) is 65 μW. If we also assume that we are using 1024 waveguides (supplied by a ten-stage binary splitter system with a total insertion loss of $\log_2 1024 \times 0.1$ dB $= 1$ dB), each with a nonresonant insertion loss of 63%, we find that we need only 200 mW of optical power to supply the entire system. Even at this power, cross-phase modulation is weak enough that it can be neglected over the centimeter-scale distances that would be used in this system.

Single tiles of nanocircuits may not be able to operate at high speeds because of the capacitance of metal interconnects connecting the tile to other distant electronic components. In this case, the minimal implementation of a nanophotonic interconnect using PhC technology is shown in Fig. 10. There are only two wavelengths used in this OWDM scheme, corresponding to a single memory/sensor/logic mosaic to be accessed. One wavelength (λ_1) is used for writing/input and another (λ_2) for reading/output. Note that we are not concerned about insertion loss for wavelength λ_1 immediately after the extraction for photodetection, so the EOM can be placed directly in the path of the input waveguide. It is also possible that a high data rate is not required for data/transfer to/from a single mosaic, so that it is convenient to divide the spectrum previously available for a single wavelength into a larger number of slowly modulated wavelengths. As long as the aggregate bandwidth of the sub-OWDM wavelengths is less than the resonant transmission bandwidth of the resonator, the intracavity detector can be partitioned into a number of insulated semiconductor "slices," each sensitive to a specific wavelength. The charge produced in the detectors can be accumulated separately.

C. Plasmonic Waveguides and Couplers

The use of plasmonic waveguides formed by arrays of nontouching metallic nanoparticles has been discussed in a number of papers, with a review of original experiments and ideas given by [15]. A plasmon is well confined to the metallic nanoparticle waveguide [16]. This work has demonstrated a propagation loss of excited *resonant*

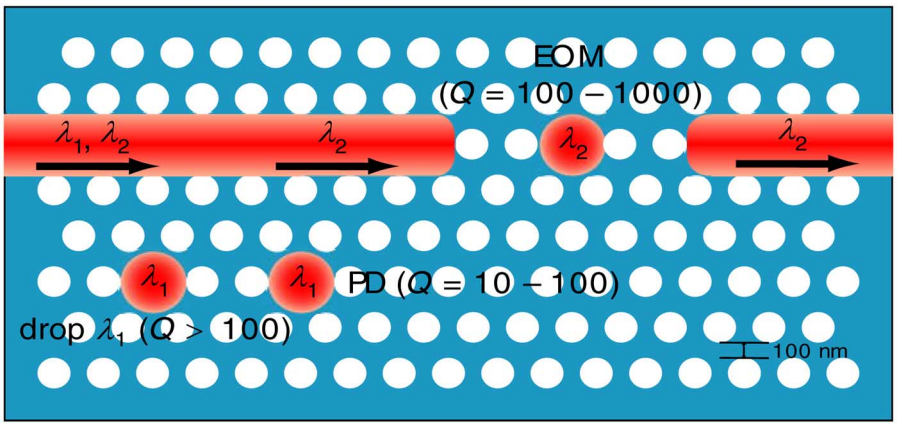

Fig. 10. *Minimal implementation of a nanophotonic data transfer junction using photonic crystal (PhC) technology. There are only two wavelengths used in this simple WDM scheme, corresponding to a single memory mosaic to be accessed. One wavelength* (λ_1) *is used for writing to a drop filter and photodetector (PD) and another* (λ_2) *for reading via an EOM.*

(plasmon) mode of 3 dB/15 nm, which is very large. The estimated coupling efficiency of exciting the subwavelength scale modes in one-dimensional (1-D) nanoparticle waveguides with near-field fiber probes with light throughputs is below 0.1%, and it is not mode-selective.

Using a recently developed design concept for a very low-capacitance and relatively low-loss metal nanoparticle plasmon coupler based on silicon-on-insulator (SOI) technology, a mode-selective energy transfer in the 1.5-μm wavelength band may be realized from a conventional fiber taper to a an electronic device (e.g., the gate of a transistor) via plasmons with efficiencies up to 75% [17]. This concept is extendable to higher frequencies and promises applications in energy guiding and optical sensing with high efficiencies. The waveguide consists of a hybrid structure of SOI and a lithographically defined square lattice of metal nanoparticles on an optically thin, undercut silicon membrane. In order to allow for nonresonant excitation of the metal nanoparticles to reduce the absorptive heating losses without a concomitant increase in radiative loss, the authors employed a lateral grading in nanoparticle size to confine the mode to the center of the waveguide. Vertically, the confinement was ensured both by bound metal/air surface plasmons and the undercut geometry of the silicon membrane.

Importantly, this design concept can also be scaled to higher frequencies towards the visible regime of the spectrum by an appropriate change in lattice constant. The higher absorptive losses for near-resonant excitations at lower wavelengths can then be partially counteracted by a change of the materials system to silver. The high efficiency of power transfer into a plasmon waveguide should thus allow applications at visible and near-infrared frequencies. For example, the use as a coupling structure to other planar plasmonic devices such as cavities and resonantly excited 1-D particle waveguides can be envisioned. The plasmonic route towards optical manipulation on the chip is very intriguing, and the very large coupling efficiency demonstrated by Painter *et al.* [17] suggests that it may become a viable approach in some cases. This technology certainly deserves thorough experimental and theoretical investigation, particularly over relatively short distance scales bridged by the intermediate interconnect layer.

D. Nanoscale Resonant-Cavity Modulators and Photodetectors

In order to implement photonic data transfer, nanoscale low-power optical modulators are needed. The most efficient way to realize this is to use resonant cavity modulators, each constructed to operate on a single OWDM channel. Resonant modulators operate through two mechanisms: a shift of the resonant peak away from the optical channel center and a change in the resonant cavity loss. In practice, the first technique is implemented by changing the cavity center frequency through variations of the refractive index. The shift of a resonant cavity solely due to a change of the refractive index is given approximately by

$$\Delta\nu_c = f\frac{c}{\lambda}\frac{\Delta n}{n} \tag{3}$$

where $\Delta\nu_c$ denotes the shift in the original cavity resonant frequency $\nu = c/\lambda$, c is the speed of light, λ is the original center wavelength, n denotes the original cavity refractive index, Δn is the change in refractive index, and f denotes the fraction of the optical mode which experiences the refractive index change. The modulation depth M (defined as the on–off power ratio) depends on both the original signal bandwidth $\Delta\nu_s$ and the shift of the cavity resonant frequency through the relation

$$M = 1 - \frac{1}{1 + \left(\frac{\Delta\nu_c}{\Delta\nu_s}\right)^2}. \tag{4}$$

This relation assumes the cavity bandwidth matches the signal bandwidth. Table 2 shows the required ratio of the shifted center frequency to the channel signal bandwidth to obtain the corresponding modulation depth. In practice, the desired modulation depth is determined by the acceptable bit error rate and power requirements for the specific application, as dictated by the specifications of the photodetector and the overall losses in each channel.

Absorptive resonant modulators (also known as electroabsorption modulators) operate by increasing the cavity loss, with the modulator center frequency staying fixed with respect to the optical channel. These devices will typically operate in a regime where the original cavity is strongly overcoupled to the optical waveguide, such that with no electrical signal applied the optical channel experiences low loss. Upon increasing the optical cavity loss (commonly performed through injection of free carriers), the cavity absorbs a significant fraction of the signal power. For an absorptive cavity modulator with a large modulation depth, the required change in cavity absorption is approximately given by

$$\Delta\alpha \approx \frac{2\pi n}{c}\Delta\nu_s. \tag{5}$$

Table 2 Fractional Cavity Shift of a Dispersive Cavity Modulator Versus Modulation Depth

Modulation depth M	Fractional cavity shift $\Delta\nu_c/\Delta\nu_s$
0.99	10.0
0.94	4.0
0.90	3.0
0.80	2.0
0.50	1.0

E. Candidate Modulator Designs

All-Silicon Modulator: A fast, low-power silicon-based modulator would be a boon to the photonics industry. However, because of the long recombination lifetime in silicon (~500 ps), it is not clear how current-injection based devices can operate at multigigabit data rates. One possibility to increase the operation frequency is to reduce the free carrier lifetime in silicon. For example, by implantation of fluorine ions, the recombination lifetime can be reduced to ~30 ps, which leads to an operating frequency of ~6 GHz (corresponding to a data rate of 3 Gbit/s). Another possibility is to consider dc-biasing the modulator so any carriers generated will be quickly swept out of the modulator, with the bandwidth limited by the carrier transit time (for an electrode spacing of ~3 μm, this gives a bandwidth of ~6 GHz). As the bandwidth is linearly dependent on the electrode separation for strong dc biasing, frequencies approaching 50 GHz may be reached (corresponding to electrode spacings of 300 nm), although there will be a much larger optical loss due to increased optical overlap with the highly absorbing electrodes.

In order to get an initial estimate of the required electrical power for each type of modulator, we assume that each optical channel will operate at a rate of 2 Gbit/s at a wavelength of 1.55 μm. For a dispersive cavity modulator, the required refractive index change for a modulation depth of 90% is 2.4×10^{-4}. As this change in refractive index is much larger than can be obtained through electrooptic effects in pure silicon at reasonable voltages, we will ultimately need to find a new material or use current injection instead.

For a current injection based device, the density of injected carriers required for this modulation depth is $\sim 1.4 \times 10^{17}$ carriers/cm^3. Assuming a carrier velocity of 10^7 cm/s and a photonic crystal defect modulator with an electrical contact area of ~1 μm^2, the current requirement is approximately ~2 mA. In contrast, an absorption-based device operating with the same parameters requires $\sim 5 \times 10^{17}$ carriers/cm^3, resulting in a current consumption of ~8 mA. Although these numbers may sound attractive, the heat dissipated by the modulators incorporated within 1000 transceivers on a chip might be unacceptably large.

Hybrid Silicon Modulators: In order to reach faster modulation speeds, it is necessary to consider materials other than silicon. For example, lithium niobate modulators are commonly used in the optical telecommunication industry to make modulators with bandwidths up to 40 GHz in a Mach–Zehnder configuration. A simple calculation (illustrated below) shows that a microcavity lithium niobate modulator (for example a photonic defect crystal design) operating at 5 Gb/s requires an applied electric field of ~17 V (over a distance of 3 μm) to obtain a modulation depth of 90%. While these voltages are possible in a nanophotonic system (especially since in principle there is no current flow), scaling up to higher data rates is unattractive because of the need for high voltage components. One possibility is to consider other electrooptic materials that have been found to have dramatically larger electrooptic coefficients. For example, some polymers and inorganic crystals have electrooptic coefficients ~20 times that of lithium niobate. In this case, operation at 40 Gb/s requires a voltage of only ~7 V for a 90% modulation depth. These materials must be further studied, however, as some are not stable at high electric field strengths.

Another possibility is to combine silicon with other semiconductor materials more suitable for optical modulation. There has been great progress in heterogeneous

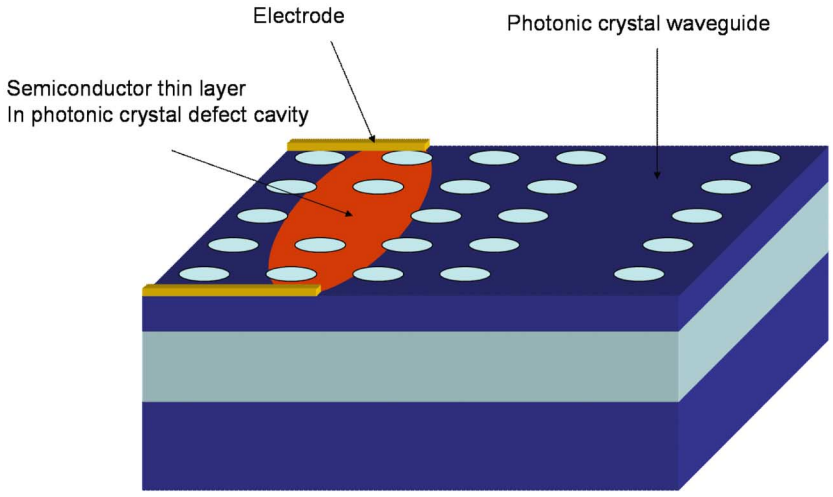

Fig. 11. *Possible design of a photonic crystal modulator based on a thin layer of semiconductor media grown on silicon. The shape of the layer and electrode placement are chosen to minimize excess loss.*

integration of high-quality semiconductor materials on silicon, ranging from thin layers to nanowires, which do not require expensive and time consuming approaches such as flip-chip bonding. Semiconductor materials from the III–V class of the periodic table such as InP and GaAs are often used for both high-speed electronics and photonics in the 1.55-μm telecommunication band. For example, the carrier mobility in GaAs is $\sim$6 times higher than that of silicon, giving a bandwidth of $>$ 30 GHz.

Active Layered Modulator: Another possible design using these compound semiconductors would involve fabricating a thin layer of III–V material directly onto the cavity structure, as illustrated in Fig. 11. Here carriers are injected through lateral metal contacts placed such that there is negligible overlap with both the optical bus waveguide and the optical mode in the modulator cavity. In this case, electrode separations of $\sim$2 μm will result in essentially no added absorption of the optical mode while allowing high-speed operation.

F. Nanoscale Si Photodetectors

Nanoscale photodiodes can be combined with resonant cavities to allow efficient detection at selected wavelengths. When coupled to a waveguide, such a device allows detection of a single OWDM channel while allowing other channels to pass with minimal attenuation. For an RCE photodiode, the absorbing region is placed inside the cavity, allowing for higher efficiencies than those attainable with nonresonant devices of similar size. Alternatively, a conventional photodiode can be placed after a drop filter.

For the design shown in Fig. 12, a thin germanium active layer is integrated into a Si/SiO_2 photonic crystal cavity. Germanium absorbs at $\lambda = 1.5$ μm and is compatible with Si crystal growth. Photogenerated carriers are collected by means of a metal–semiconductor–metal (MSM) structure in which thin metal electrodes are placed directly on the surface of the Ge active layer. An applied dc voltage provides an electric field that draws electrons and holes to the electrodes. MSM-type photodiodes are relatively simple to fabricate and have small capacitances, allowing for large detection bandwidths. An MSM photodiode with a hydrogenated amorphous silicon-germanium absorbing layer, operating at a wavelength of 850 nm, has been reported with a full-width at half-maximum response time of 51 ps [18].

The photodetector can be efficient only if absorption in the Ge active layer dominates other cavity losses. To ensure that the cavity bandwidth exceeds the signal bandwidth $\Delta\nu_s$, we require

$$f \geq \frac{2\pi n}{c\alpha}\Delta\nu_s \tag{6}$$

where f is the effective fraction of the optical mode that exists in the Ge region, n is the refractive index, c is the speed of light in vacuum, and α is the Ge absorption coefficient. For example, for a signal bandwidth of 20 GHz (requiring a $Q \sim 10\,000$ cavity), using $\alpha \approx 4 \times 10^3$ cm^{-1} at $\lambda = 1.55$ μm, we estimate $f \approx 0.004$. Thus the Ge active region is quite small. For example, in a photonic crystal cavity with mode volume $V = (\lambda/n)^3$, the Ge region could be a cube 70 nm on a side or a film 10 nm thick and 200 nm on a side. The exact geometry of the active region must be chosen to maximize the collection efficiency of photogenerated carriers. For the design shown in Fig. 12, a thin Ge layer covers the defect region of a photonic crystal cavity. The photogenerated carriers are confined within the Ge layer

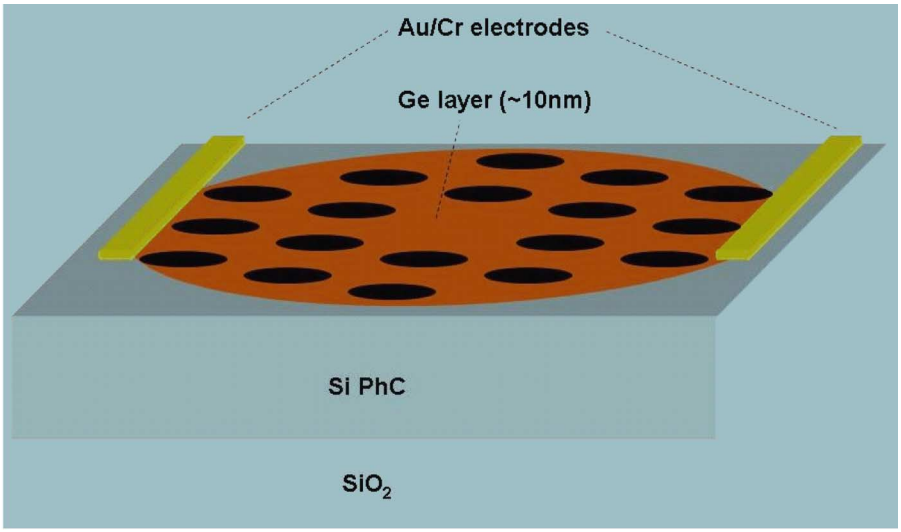

Fig. 12. *A Si-Ge MSM photonic-crystal-cavity-enhanced photodiode.*

because of its smaller bandgap and are pulled toward the electrodes by the applied electric field. If surface recombination becomes a problem, this might be solved by covering the Ge layer with Si to form a quantum well. The metal electrodes contact the Ge layer away from the cavity center. This placement minimizes absorption and diffraction losses.

G. Achieving the Required Bit Error Rate

For an optimal photodetection circuit, the variance in the number of detected photoelectrons n in a single pulse is approximately

$$\langle \Delta n^2 \rangle \approx \langle n \rangle + kT/(e^2/2C) \tag{7}$$

where k is Boltzmann's constant, T is the temperature in Kelvin, e is the charge of an electron, and C is the photodiode capacitance. The first term, representing Poisson-distributed shot noise, is equal to the mean number of detected photoelectrons. The second term represents thermal noise, which is expected to follow a Gaussian distribution. The factor $e^2/2C$ is equal to the single-electron charging energy of the device. We do not include some other possible noise sources such as excess noise in the amplifier circuit, leakage current and $1/f$ noise due, for example, to traps. Leakage current and $1/f$ noise are unlikely to be important at the high bit rates considered here. The small capacitance possible in a nanoscale photodetector provides an important advantage over large-area detectors. For integrated micrometer-scale structures, capacitances can be as small as 100 aF, giving a ratio of $kT/(e^2/2C) \approx 32$.

A decision threshold is chosen so that the error probability is the same either for an "on" pulse or an "off" pulse. The mean number of photons required in an "on" pulse so that the error probability P_{err} is less than $\exp(-b)$ is approximately

$$n_{\min} \approx \frac{2b}{\eta M^2}\left(2 - M + 2\sqrt{1 - M + \frac{M^2}{2b}\sigma_{\mathrm{th}}^2}\right). \tag{8}$$

Here M is the modulation depth, so that the mean photon number in an "off" pulse is $(1-M)\,n_{\min}$. The total quantum efficiency of the photodetector is given by η, and σ_{th}^2 is the thermal noise. This formula, based on a Gaussian approximation, is valid for the range of parameters considered below. For an optical channel to be useful with minimal error correction in a reliable computer, we estimate that the error probability must be less than 10^{-22}. Table 3 below was calculated for $P_{\mathrm{err}} = 10^{-22}$ and $\sigma_{\mathrm{th}}^2 = 32$.

The values in Table 3 can be divided by the quantum efficiency to predict the required number of photons per pulse. For example, a 90% modulation depth ($M = 0.9$) and 50% quantum efficiency ($\eta = 0.5$) require ~500 photons per "on" pulse. For a 2-Gb/s communication rate at a wavelength of 1.5 μm, this translates to < 130 nW of optical power dissipated per channel. Even for 40 Gb/s, the optical power per channel is still less than 2.6 μW.

Table 3 Minimum Required Photon Number for an "On" Pulse (Normalized to Unity Quantum Efficiency) for Various Modulation Depths

M	$\eta n_{\min}$
0.99	221
0.94	255
0.9	287
0.8	391
0.7	548
0.5	1225
0.1	38500

VI. NANOPHOTONIC WAVEGUIDES AND RESONATORS FOR GLOBAL INTERCONNECT

There are a wide range of approaches for chip-based photonic information transfer, with the best solution dependent on the desired application. Here we consider high-density low-power approaches appropriate to large-scale integration with high-performance nanoelectronic circuits. This precludes most previous efforts into chip-scale electrical/optical integration, where the focus has been on chip-to-chip data transfer. Correspondingly, relatively large photonic structures can be used, as the chip real-estate is dominated by electronics. Here we focus on implementations where optics and electronics are tightly entwined, in order to solve the global interconnect problem introduced previously.

In the "mosaic" design (Fig. 13), small nanoelectronic circuits are connected with optical busses in an OWDM/OTDM architecture. However, as the optical interconnects are by necessity much larger than any nanoelectronic circuit element, it is desirable to minimize the aggregate optical footprint of the optical waveguides and transceivers. In order to accomplish this, there are two main approaches for high-density nanophotonics: photonic bandgap crystal and ridge waveguide technologies.

A. Photonic Crystal Waveguides and Resonators

Let us consider the physical footprints required for a PhC nanophotonic OWDM system. Typically, each waveguide will require at least three lattice periods of holes on each side of the waveguide to provide the bandgap necessary for high optical confinement [19]. At a lattice spacing of about 300 nm (as commonly used in a slab-type PhC structure designed to operate at an optical wavelength of 1.55 μm), each PhC waveguide has a physical width of 1–2 μm. For each add/drop filter or photodetector/modulator, if we assume a design consisting of a single defect PhC cavity and a modest Q, we need approximately 3–4 μm^2 of physical area. Thus, if we adopt the simple transceiver architecture, each complete transceiver would occupy approximately 30–40 μm^2.

The use of PhCs presents cost and complexity challenges during fabrication. The presence of a periodic high-index

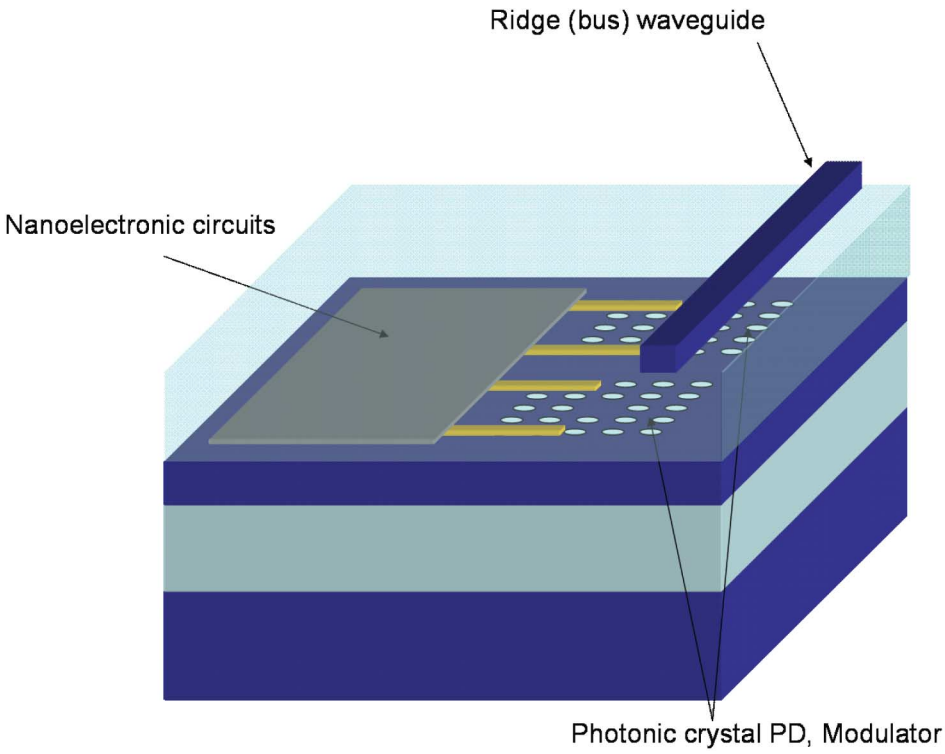

Fig. 13. ***Illustration of a possible nanophotonics implementation. Bottom layer contains nanoelectronic circuitry along with electrically connected photonic crystal-based photodetectors and modulators. These PDs and modulators are evanescently coupled to the ridge waveguide optical bus residing on a top layer.***

contrast lattice results in a large number of possible scattering surfaces. As the optical losses in most planar waveguides arise from scattering from surface roughness introduced during the fabrication process, the larger overlap between the scattering surfaces and the optical mode in a PhC structure suggests scattering losses are larger than those of more conventional waveguides. Experimentally, this is found to be true, where published losses for PhC waveguides are considerably higher than the closest corresponding planar ridge waveguides, when considering similar material systems and fabrication methods: about 6 dB/cm for a PhC waveguide [20] compared to about 3 dB/cm for a ridge waveguide [21].

B. Ridge Waveguides and Microresonators

Conventional ridge waveguide technology is very well known and established, with considerably easier design, fabrication, and understanding than PhCs. However, the modal areas/volumes for these structures are larger than those of PhCs, resulting in less compact devices. When considering a high-index contrast ridge/strip waveguide, the modal area is only a factor of about two larger than the corresponding PhC structure, and the actual physical footprint is about 1 μm. Notice this is actually *smaller* than the PhC structure by about a factor of two, due to the need for a few lattice periods to create a bandgap. While this suggests that conventional ridge waveguides are the best choice for high-density optical waveguides, considerations such as crosstalk (PhCs can be engineered to have very little optical crosstalk for waveguides only 300 nm apart, much less than is the case for ridge waveguides) and waveguide bends make the optimal choice dependent on the actual architecture of the system.

Microcavities based on ridge waveguides generally consist of either Fabry–Pérot cavities using Bragg reflection, such as 1-D gratings or inline PhCs (neither of which is optimal for inline electrooptic OWDM transceivers) or whispering-gallery cavities, such as microring resonators [22]. The use of microring resonators for both active and passive devices has been thoroughly investigated, with numerous demonstrations of low-loss devices in a wide variety of material systems. Furthermore, a reasonable degree of integration with other optical/electrical components has been already achieved. The main drawback for a microring-based transceiver is the much larger physical space required for a whispering-gallery-mode structure. Assuming a high-index contrast structure such as SOI, for low losses the ring must have a diameter larger than 5 μm, resulting in at least 25 μm^2 of real estate *per* add/drop filter, photodiode, or modulator (with a more reasonable estimate in practice of 50–100 μm^2). Considering an entire transceiver, the required real estate is approximately 200 μm^2 (a factor of about ten larger than that of the corresponding PhC design). Nevertheless, the fabrication of ridge waveguides and

microcavities is considerably less complex than PhCs, and exhibit losses that are about a factor of two lower.

C. Hybrid Three-Dimensional Architectures

An elegant way to mitigate the size difference between optical and electrical components is to employ a three-dimensional (3-D) design, where the optical components reside on a different layer than the nanoelectronic circuitry. One possible way to do this is illustrated in Fig. 13. This photonic system consists of an optical bus layer containing ridge waveguides combined with photonic crystal based transceivers on the nanoelectronics layer. This design allows the low losses of ridge waveguides to be used for the main data busses and the compact footprints of PC nanoresonators to be used for the add/drop/PD/modulators. Here the photonics space required on the nanoelectronics layer is minimized, while allowing easy electrical contact between the PD/modulator and the nanoelectronic circuits. Furthermore, the use of PC-based PDs and modulators may allow increased bit rates with respect to ridge technology due to the much smaller physical lengths between electrical connections. In principle, it may also be possible to place the add/drop filters, as well as the photodetectors and modulators, in their own layers.

These ideas motivate the exploration of new structures and architectures that use the flexibility of 3-D integration to potentially ease these limitations. Further, use of 3-D circuits might diminish the need to significantly reduce the feature size of individual components. Nevertheless, expanding circuit design into a third dimension raises many new questions. Careful consideration must be given to all portions of the structure, from circuit design and layout to the construction of new multilayered structures. As feature sizes shrink, 3-D intermediate interconnects and "through-wafer" contacts (possibly including nanowires and carbon nanotubes) are likely to become necessary in any case, even though a number of materials processing have a number of critical and unanswered problems. Critical issues will include thermal management under possibly high heat loads; the thinning and bonding of wafers; multilayer lithography and alignment; patterning, etching, and filling dense inter/intrachip vias; reliability and manufacturing costs; and contacts and interfacial impedances of nanoelectronic circuits.

VII. CONCLUSIONS

We have seen that although the potential is great, there are many challenges to implementing on-chip photonic global interconnect. What is needed are low loss and small area waveguide structures to efficiently transport photons, and small-volume, low-power, and high-efficiency transceivers to exchange information between photons and electrons at lower levels of interconnect. The physics that can enable such components to be built has only recently been understood, and the challenges of actually fabricating and integrating them with silicon electronics are formidable. In general, it appears that they will be based on photonic bandgap structures that will form integrated photonic circuits. These structures can contain the necessary broad band waveguides, add–drop filters, modulators, and resonant detectors needed for photonic interconnect. The challenges going forward are to actually build components with the necessary performance, integrate them together into systems, and then demonstrate that they can be built using standard semiconductor fabrication procedures.

It is important to note that here we have made the "nanophotonic" case for significant changes to modern integrated circuit architecture and fabrication based on rather generic observations about the physics of metal interconnects. However, we have not made the system-level architectural case for nanophotonic interconnects based on specific future requirements for either programmability or computational performance. Nor have we demonstrated that orders-of-magnitude improvements in global interconnect bandwidth will necessarily translate into similar increases in overall computational bandwidth. Our further investigations into these additional areas of investigation are underway, and will be published in the near future. ■

REFERENCES

[1] J. D. Meindl and J. A. Davis, "The fundamental limit on binary switching energy for terascale integration," *IEEE J. Solid-State Circuits*, vol. 35, p. 1515, 2000.

[2] J. D. Meindl, "Beyond Moore's law: The interconnect era," *Comp. Sci. Eng.*, pp. 20–24, Jan./Feb. 2003.

[3] A. Naeemi, J. Xu, A. V. Mule, T. K. Gaylord, and J. D. Meindl, "Optical and electrical interconnect partition length based on chip-to-chip bandwidth maximization," *IEEE Photon. Technol. Lett.*, vol. 16, p. 1221, 2004.

[4] E. Yablanovitch, "Photonic band-gap crystals," *J. Phys. Condens. Matter*, vol. 5, p. 2443, 1993.

[5] M. Haurylau, G. Chen, H. Chen, J. Zhang, N. A. Nelson, D. H. Albonesi, E. G. Friedman, and P. M. Fauchet, *IEEE J. Sel. Topics Quantum Electron.*, vol. 12, p. 1699, 2006.

[6] R. Kirchain and L. Kimerling, "A roadmap for nanophotonics," *Nature Photon.*, vol. 1, p. 303, 2007.

[7] G. Snider, P. Kuekes, T. Hogg, and R. S. Williams, "Nanoelectronic architectures," *Appl. Phys. A*, vol. 80, pp. 1183–1195, 2005.

[8] G. Snider, P. Kuekes, and R. S. Williams, "CMOS-like logic in defective, nanoscale crossbars," *Nanotechnology*, vol. 15, pp. 881–891, 2004.

[9] Ramaswami and K. N. Sivarajan, *Optical Networks: A Practical Perspective*. New York: Academic, 2002, 2/e.

[10] J. D. Joannopoulos, R. D. Meade, and J. N. Winn, *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton Univ. Press, 1995.

[11] K. Sakoda, *Optical Properties of Photonic Crystals*. Berlin, Germany: Springer, 2001.

[12] W. Wu, G.-Y. Jung, D. L. Olynick, J. Strasnicky, Z. Li, X. Li, D. A. A. Ohlberg, Y. Chen, S.-Y. Wang, J. A. Liddle, W. M. Tong, and R. S. Williams, "One-kilobit cross-bar molecular memory circuits at 30-nm half-pitch fabricated by nanoimprint lithography," *Appl. Phys. A*, vol. 80, pp. 1173–1178, 2005.

[13] K. Srinivasan and O. Painter, "Fourier space design of high-Q cavities in standard and compressed hexagonal lattice photonic crystals," *Opt. Express*, vol. 11, p. 579, 2003.

[14] M. K. Emsley, O. Dosunmu, and M. S. Unlu, "High-speed resonant-cavity-enhanced silicon photodetectors on reflecting silicon-on-insulator substrates," *IEEE Photon. Technol. Lett.*, vol. 14, p. 519, 2002.

[15] S. A. Maier *et al.*, "Plasmonics—A route to nanoscale optical devices," *Adv. Mater.*, vol. 13, p. 1501, 2001.

[16] S. A. Maier, P. G. Kik, and H. A. Atwater, "Observation of coupled plasmon-polariton modes in Au nanoparticle chain waveguides of

different lengths: Estimation of waveguide loss," *Appl. Phys. Lett.*, vol. 81, p. 1714, 2002.

[17] S. A. Maier, M. D. Friedman, P. E. Barclay, and O. Painter, "Experimental demonstration of fiber-accessible metal nanoparticle plasmon waveguides for planar energy guiding and sensing," *Appl. Phys. Lett.*, vol. 86, p. 071103, 2005.

[18] L.-H. Laih, J.-C. Wang, Y.-A. Chen, W.-C. Tsay, T.-S. Jen, J.-S. Chen, and J.-W. Hong, "Improving the transient response of a Si metal-semiconductor-metal photodetector with an additional i-a-SiGe:H film," *Jpn. J. Appl. Phys.*, vol. 36, p. 1494, 1997.

[19] J. D. Joannopoulos, R. D. Meade, and J. N. Winn, *Photonic Crystals: Molding the Flow of Light*. Princeton, NJ: Princeton Univ. Press, 1995.

[20] E. Dulkeith, S. J. McNab, and Y. A. Vlasov. (2005, Apr.). Mapping the optical properties of slab-type two-dimensional photonic crystal waveguides. [Online]. Available: http://arxiv.org/abs/physics/0504132

[21] Y. A. Vlasov and S. J. McNab, "Losses in single-mode silicon-on-insulator strip waveguides and bends," *Opt. Express*, vol. 12, p. 1622, 2004.

[22] S. M. Spillane, T. J. Kippenberg, and K. J. Vahala, "Ultralow-threshold Raman laser using a spherical dielectric microcavity," *Nature*, vol. 415, pp. 621–623, 2002.

ABOUT THE AUTHORS

Raymond G. Beausoleil (Senior Member, IEEE) received the bachelor's degree in physics from the California Institute of Technology, Pasadena, in 1980 and the Ph.D. degree in physics from Stanford University, Stanford, CA, in 1986.

He has been a member of Technical Staff with Boeing High Technology Center, Director of Research at Solidlite Corporation, Cofounder and President of Cygnus Laser Corporation (and subsidiaries), and Director of Research at ElseWare Corporation, during which time his research focused on high-power all-solid-state laser and nonlinear optical systems as well as mathematical algorithms for computer firmware. In 1995, he became a member of Technical Staff with HP Laboratories, Palo Alto, CA, where he is now a Principal Scientist performing basic research in the Quantum Science Research Department in nanoscale optics for classical and quantum information processing. He has been a member of the affiliate Faculty of Stanford University, Stanford, CA, since 1995, where he has led the research, development, and implementation of a numerical model of the optical response of the NSF Laser Interferometer Gravitational-Wave Observatory detector to environmental perturbations, thermal focusing and deformations, and gravitational radiation.

Philip J. Kuekes (Member, IEEE) received the B.S. degree in physics from Yale University, New Haven, CT, in 1969.

He is Chief Architect with the Quantum Science Research, Hewlett-Packard Laboratories, Palo Alto, CA. He designed mega-op array processors in the 1970s with Raytheon and giga-op systolic processors in the 1980s with TRW. In 1991, he joined HP Laboratories as Project Manager for Teramac, a trillion operations per second reconfigurable computer. Teramac is the largest defect-tolerant processor ever made. He has developed various architectures for chemically assembled electronic nanocomputers for the HP molecular electronics program. He has received 21 patents in molecular electronics and parallel computer architectures.

Mr. Kuekes received the 2000 Feynman Prize in Nanotechnology (with J. R Heath and R. S. Williams). He was named to the "Scientific American 50" list of technology leaders for 2002 and Researcher of the Year in 2005 by *Small Times* magazine.

Gregory S. Snider is a Researcher with Hewlett-Packard Laboratories, Palo Alto, CA, investigating nanoelectronic architectures, circuits, compilation, and simulation. Previously he has worked in communications, processor design, medical instrumentation and imaging, networking, operating systems, computer security, memory systems, compilers, hardware and software synthesis, e-services, simulation, and programmable hardware. In the early 1990s, he was the Architect and Compiler Designer of the Teramac simulation system, a defect-tolerant computer built from hundreds of custom field-programmable gate arrays.

Shih-Yuan Wang (Fellow, IEEE) received the B.S. degree in engineering physics and the Ph.D. degree in electrical engineering and computer sciences from the University of California, Berkeley, in 1969 and 1977, respectively.

In 1977, he joined the HP Microwave Semiconductor Division, working on GaAs field-effect transistors. In 1980, he joined the Applied Physics Lab/HP Labs, working on high speed photodiodes, III–V traveling-wave electrooptic modulators, vertical surface emitting lasers, and GaN blue light-emitting diodes and lasers. After four years with Agilent Technologies and a startup he cofounded, he rejoined HP in 2003 and is now with the Quantum Science Research group at HP Labs, working on optical frequency metamaterials, nanoimprint lithography, nanoscale electronic devices, nanowire devices on silicon, nanophotonics and photonic interconnect. He has received more than 50 patents. He is on the Editorial Boards of *Applied Physics A* and the *Journal of Nanoengineering and Nanosystems*.

Dr. Wang is a Fellow of OSA.

R. Stanley Williams received the B.A. degree in chemical physics from Rice University in 1974 and the Ph.D. degree in physical chemistry from the University of California, Berkeley, in 1978.

He is HP Senior Fellow with Hewlett-Packard Laboratories and founding Director of the Quantum Science Research group. He was a Member of Technical Staff with AT&T Bell Laboratories from 1978 to 1980 and a Faculty Member (Assistant, Associate, and full Professor) of the Chemistry Department, University of California, Los Angeles, from 1980 to 1995. His primary scientific research during the past 20 years has been in the areas of solid-state chemistry and physics. He has received 62 U.S. patents and published more than 300 papers in reviewed scientific journals.

Dr. Williams has received numerous awards for scientific and academic achievement, including the 2000 Julius Springer Award for Applied Physics, the 2000 Feynman Prize in Nanotechnology, and the 2004 Birnbaum Prize (the last two shared with P. Kuekes). He was named to the inaugural Scientific American 50 Top Technology leaders in 2002 and 2005.