Technologies for exascale systems

To satisfy the economic drive for ever more powerful computers to handle scientific and business applications, new technologies are needed to overcome the limitations of current approaches. New memory technologies will address the need for greater amounts of data in close proximity to the processors. Three-dimensional silicon integration will allow more cache and function to be integrated with the processor while allowing more than 1,000 times higher bandwidth communications at low power per channel using local interconnects between Si die layers and between die stacks. Integrated silicon nanophotonics will provide low-power and high-bandwidth optical interconnections between different parts of the system on a chip, board, and rack levels. Highly efficient power delivery and advanced liquid cooling will reduce the electrical demand and facility costs. A combination of these technologies will likely be required to build exascale systems that meet the combined challenges of a practical power constraint on the order of 20 MW with sufficient reliability and at a reasonable cost.

Introduction

High-performance computing (HPC) is currently experiencing very strong growth in all computing sectors, driven by an exponentially improving performance/cost ratio for HPC machines. As shown in Figure 1, both the average system performance and the performance of the fastest computers as measured by the Top500** benchmark have been increasing 1.8 times every year. Because systems have not dramatically increased in cost, the rate of improvement in performance/cost has also followed this trend. It is inevitable, therefore, that physical experiments will continue to be augmented and, in some cases, replaced by HPC simulations. The current worldwide investment in research and development is approximately $1 trillion, with HPC hardware accounting for merely $10 billion. (Software and services account for an additional $10 billion per year.) This leaves tremendous room for growth in HPC, which will likely continue as long as we can build more cost-effective machines.

The race to build an exaflop computer ($10^{18}$ floating-point operations/second) in the next 8–10 years is representative of the great potential that HPC computers hold. Nevertheless, there are enormous challenges in continuing to improve performance at the near-doubling-every-year rate [1]. Some of these challenges are highlighted in Table 1. This table shows how the underlying technologies are not improving at a rate that is even close to what would be needed to achieve 1.8 times per year through evolutionary means. Thus, there must be innovations in architecture while exploiting existing technologies in new ways, as well as developing and integrating new technologies.

There are primarily three different aspects to the challenges in achieving exascale computing. These are cost efficiency, energy efficiency, and reliability.

To continue improving the performance/cost ratio of computing, we must continue to higher levels of integration; simply leveraging future silicon by putting more performance in a compute chip is not sufficient.
explored by nonexperts in HPC, we should focus on solving these challenges in a manner that results in highly usable balanced systems.

The challenges in processor memory are twofold. We need to have memory devices that support the bandwidth requirements that future processors will need. Additionally, we need to continue improving the cost per bit of memory. As this later metric for dynamic random access memory (DRAM) is not anticipated to increase at a rate that is comparable to the rate of processor improvement, we need to aggressively research and develop new memory technologies. In the near term, this means that we need to change the bandwidth-per-device rule of thumb that relates the bandwidth needed per byte of memory. In the near future, we need much more bandwidth per device of memory. In the long term, we need new technologies to provide significantly better cost per bit over DRAM.

To exploit memory and processors with high bandwidth capabilities, we need to have packaging technology that will allow one to connect these two system components in a cost-effective and power-efficient manner. Technologies such as 3-D packaging are crucial in achieving this goal, as suggested by the illustration in Figure 3. Improvements in the efficiency of power delivery and cooling are also critically needed.

**Figure 1**
Performance of supercomputers: (top curve) total compute power of the highest performing 500 systems, (second curve) performance of the number one computer in the list, (third curve) number ten computer, and (lowest curve) number 500.

**Table 1** Exascale challenges (current annual improvement rates).

<table>
<thead>
<tr>
<th>Component</th>
<th>Improvement Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silicon</td>
<td>Circuit density</td>
</tr>
<tr>
<td></td>
<td>Wiring channel density</td>
</tr>
<tr>
<td></td>
<td>Energy density</td>
</tr>
<tr>
<td>Chip package</td>
<td>I/O density</td>
</tr>
<tr>
<td>Electrical card wiring</td>
<td>Number of layers</td>
</tr>
<tr>
<td></td>
<td>Wiring channel pitch</td>
</tr>
</tbody>
</table>

**Figure 2**
Exascale technologies and their interrelation. Power and cooling are implicitly assumed.
In large systems, the processor-to-processor interconnect plays a critical role. The cost of these interconnects will become prohibitive without developing new technologies. Interconnects of the future will be dominated by optics, as this offers the potential for a far better cost solution for all distances. As we look to exascale, even the connections between processors on a common circuit card needs to be optical because of the amount of bandwidth needed. Silicon photonics represents a nearly ideal solution to the interconnect problem. With silicon photonics, the cost is driven almost entirely by the packaging and optical connector cost, as the core photonics structures can be made in silicon technology using relatively standard tool flow.

**Memory technologies**

The complex memory hierarchy shown in Figure 4 is designed to alleviate the gap between processor and memory performance. While individual processor operating frequency peaked at approximately 4 GHz in mid-2000, multicore processors continue to widen the performance gap. In massively parallel high-performance computers, concurrence and locality are taxing the capacity, bandwidth, and latency of the entire memory. In order to meet the prescribed exascale performance target, the capacity of main memory needs to be increased more than 200 times the capacities employed in petascale systems. This increase in memory capacity with limited cost and energy resources places significant burdens on the design of the exascale memory system. The cost and energy challenge is pervasive, affecting every component of the exascale system. Here, we focus our attention on the memory system. The scaling trend of the incumbent memory technologies is studied. Emerging memory technologies potentially capable of providing cost- and power-effective solutions for the exascale memory system are also examined.

Since its invention in 1967 [2], the one-device DRAM has been the workhorse of the memory system in all computing systems. Riding the advances in semiconductor technology prescribed by Moore’s Law, the density of a DRAM chip has doubled every 18 months, and more recently, every 3 years, from the first commercially available DRAM chip of 1 Kb (kilobits) in the 1970s to 4 Gb (gigabits) today. However, the effort for further scaling the DRAM critical dimension (CD) has met with some unprecedented challenges. Current DRAM technology is approaching the 30-nm size limit enabled by 193-nm wavelength immersion lithography [3]. Complexity in DRAM cell layout and tight tolerances...
required of the sense circuits forbid the use of double patterning techniques to half the CD as employed in NAND Flash technology, as depicted in Figure 5. Next-generation extreme ultraviolet lithography with 13.5-nm wavelength for scaling the CD below 20 nm is still under development. The most challenging process in the integration of the DRAM cell is the storage capacitor. The storage capacitor in a typical modern DRAM is a cone-shaped structure with an aspect ratio of the height-to-base area greater than 50:1. In the cone structure, the storage node is sandwiched between the common plates on both sides to increase storage area. Without a major breakthrough for the dielectric material used in the storage capacitor, scaling the capacitor beyond the 20-nm node would require a two-time increase in the aspect ratio of the capacitor structure, whereas the cone-structured capacitor has to be changed to a rod structure. The constant storage capacitor refresh time of 64 ms dictates a stringent off-leakage requirement for the cell access transistor in DRAM to keep the signal charge. This leakage requirement practically forbids scaling of the channel length of the cell access transistor while keeping the memory cell size to a constant number of feature squares, six in state-of-art DRAM. This stringent leakage requirement necessitates the use of 3-D access devices, which increase the process complexity, thus increasing the production cost of DRAM. The slowdown in scaling progress and intrinsic limits to CD scaling mean that DRAM is probably not going to provide a cost-effective solution for the exascale memory system.

Hard disk drives (HDDs) are mechanical devices that are power hungry since they involve constantly spinning the magnetic platters at rotary speeds that rival modern jet engines. With mechanical moving parts, data integrity and reliability have always been major issues confronting the
HDD industry. Extensive and complex error-correction systems such as Redundant Array of Independent Disks (RAID) systems have been devised to protect data stored in HDDs. However, one indispensable merit of HDD is its low cost. State-of-the-art HDDs are capable of storing terabits per square inch, making it the most inexpensive storage device other than tape drives. Today, a gigabyte of storage on an HDD costs less than 10 cents. However, HDD capacity is rapidly approaching its superparamagnetic limit where thermal instability can randomly flip the magnetization direction in nanoscale magnets.

In recent years, solid-state drive (SSD) employing NAND Flash memory as the storage media has been inserted in the memory hierarchy to bridge the huge latency gap between the main memory and the HDD. NAND Flash memory with a cell size of four feature squares (4F²), the theoretical minimum for a planar memory cell, was introduced as high-density nonvolatile memory in the early 1990s [4]. With aggressive scaling and multibit storage capability, NAND Flash has become pervasive in many electronic devices. While NAND Flash is still approximately 10 times more expensive than HDD, it does provide faster access file storage, as well as lower power consumption and cost of ownership because of its smaller footprint. However, NAND Flash based on electrons stored in floating gates is also approaching physical limits in scaling. Because NAND Flash is scaled toward the nanoscale, the number of electrons that can be stored diminishes. For a CD equal to and smaller than 20 nm, there are merely a few hundred electrons stored in the floating gate. Data integrity is becoming a great concern, and extensive error-correction codes are incorporated in SSD to alleviate this problem. In addition, since electrons tunnel in and out through the tunnel oxide of the floating gate in high electric fields, the wear-out mechanism of the tunnel oxide limits the number of reliable programming cycles of the memory cell. Current NAND Flash multilevel cell devices are limited to less than 10,000 programming cycles. The wear-out mechanism in NAND Flash also deteriorates with scaling.

In lieu of DRAM, NAND, and HDD, researchers and semiconductor developers are actively working on various emerging memory technologies as replacements or supplements. Resistive random access memory (RRAM), spin-torque transfer magnetic random access memory (STT MRAM), and phase-change memory are among the most actively pursued emerging memory technologies. Figure 6 shows typical structures and brief operating principles of these emerging memory technologies. All three are nonvolatile memory technologies with demonstrated performance rivaling that of DRAM. The target for any emerging memory technologies to be a viable replacement must satisfy the specifications set forth in Table 2.

The memory element of RRAM is a metal–insulator–metal capacitor-like structure in which the insulator is typically a metal oxide. The reversible resistive change phenomenon was first published in 1962 [5]. RRAM has attracted much attention since enormous varieties of insulators not restricted to metal oxides exhibit similar resistive change phenomenon. Although the basic physics underlying RRAM is not completely understood, it is generally agreed that the phenomenon is associated with oxygen vacancies in the oxide film, whereas the physical conduction path is filamentary [6–8]. Recently, tremendous progress has been made in RRAM in emerging memory research from academia to memory developers. On the other hand, STT MRAM is built on the theory revealed in the 1990s [9, 10]. The memory element of an STT MRAM cell is the magnetic tunneling junction (MTJ). Each MTJ consists of two ferromagnetic layers separated by a very thin tunneling dielectric film. Magnetization in one of the layers is pinned or fixed in one direction by coupling to an antiferromagnetic layer. The other ferromagnetic layer is a free layer and is used for information storage. By controlling the direction of magnetization of the free layer with respect to the pinned layer, the MTJ can be configured to a high-resistance antiparallel state or a low-resistance parallel state. The direction of magnetization of the free layer is controlled by the spin of electrons determined by the polarity of the current driven through the MTJ during a write operation. The prospects of STT MRAM for replacing high-performance memory are promising [11]. Of all the emerging memory technologies, parameter random access memory (PRAM) is the most mature with limited production from major memory manufacturers. PRAM is based on reversible transformation of a phase-change material from a polycrystalline to an amorphous state by Joule heating creating a resistance change in the material. The phase-change characteristic in chalcogenide materials was observed in the 1960s [12, 13]. Current PRAM products are designed to replace NOR Flash in mobile applications. Considerably, more progress is required in phase-change materials, memory cell, and array designs in order for PRAM to replace DRAM or NAND Flash.

The three memory technologies discussed in the preceding paragraph represent a short list of emerging memory technologies in the research and development pipeline [14]. The tasks for these memory technologies in meeting the energy challenge for exascale computing are tremendous yet similar. All solid-state memory devices are governed by the operating voltage V and the two basic passive components, i.e., resistance R and capacitance C. With comparable lithographic CDs, these parameters are similar for the four emerging technologies; thus, the bit-cell read and write power differences are very much the same as and not much better than those of incumbent DRAM and...
NAND Flash technologies. To meet the energy requirement for exascale computing, a new memory architecture must be designed to take full advantage of the nonvolatility property of these emerging technologies. In addition, exascale memory systems must rely heavily on advances in 3-D packaging to reduce the input and output distances, since much of the energy spent is in the movement of data. The cost factor is a different story; RRAM and PRAM would have an advantage over STT MRAM since both RRAM and PRAM are capable of storing multiple bits in a single memory cell. This multiple-bit storage capability is the most effective way to reduce cost per bit with little cost added in sensing circuits. The penalty of multiple-bit storage is degradation in latency, which can be hidden by an innovative memory system design. Since RRAM, STT MRAM, and PRAM are emerging technologies, there is not much field data on the reliability aspect of the technologies. Field failure data generally provides a wealth of knowledge for
memories to develop and design next-generation products. SST MRAM and PRAM products in the field today could provide some valuable data, but there may not be enough data to envision a reliable memory system for exascale computing without extensive use of redundancies and error corrections. Reliability is one of the much studied topics in RRAM research. Considering power, cost, reliability, and maturity of development, PRAM is the most probable candidate to be integrated into the exascale computing system.

Packaging, interconnection, and energy management technologies

Packaging, interconnection, and energy management present formidable challenges to achieve exascale computing within both system power and cost objectives for the program. Complementary metal-oxide semiconductor (CMOS) feature size will continue to scale, but dissipated power will not. If power per operation would remain constant, an exaflop machine will have 50 times the power of the anticipated 20-Gflop/s (giga-floating-point operations per second) IBM Blue Gene/Q supercomputer now under construction [15]. If instead we are to realize a 20-MW machine, we must find a way to dramatically reduce power losses in data movement, power conversion, and cooling while also meeting the increased packaging and interconnection densities to support the performance objectives. We discuss this subject in inverse order, where each of these technology challenges can be described in more detail.

Table 2  Preliminary memory specifications for an exascale memory system.

<table>
<thead>
<tr>
<th>Type</th>
<th>File storage</th>
<th>Main memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cost</td>
<td>Competitive to NAND Flash</td>
<td>3–4× cheaper than DRAM</td>
</tr>
<tr>
<td>Capacity per chip</td>
<td>128 Gb at 22 nm (or 1× nm)</td>
<td>32 Gb at 22 nm</td>
</tr>
<tr>
<td>Nonvolatility</td>
<td>Essential</td>
<td>Important; enables memory-based checkpoint/restart</td>
</tr>
<tr>
<td>Raw write endurance</td>
<td>$10^5$ (10^7 in future generations, for enterprise solid-state drives)</td>
<td>$10^8$ (10^9 in future generations)</td>
</tr>
<tr>
<td>Read latency LR</td>
<td>2 μs</td>
<td>100 ns</td>
</tr>
<tr>
<td>Write latency LW at SET</td>
<td>10 μs</td>
<td>500 ns</td>
</tr>
<tr>
<td>Bandwidth per chip: READ</td>
<td>0.4 Gb/s</td>
<td>4.0 Gb/s</td>
</tr>
<tr>
<td>(SET) WRITE</td>
<td>0.2 Gb/s</td>
<td>2.0 Gb/s</td>
</tr>
<tr>
<td>Chip interface</td>
<td>Modified double data rate</td>
<td>Modified double data rate</td>
</tr>
<tr>
<td>Retention time</td>
<td>2-year (minimum) at 85°C</td>
<td>1 hour at 110°C 1 year at 25°C</td>
</tr>
</tbody>
</table>

Cooling

Energy-efficient cooling may be realized through the use of water cooling. If the water can be maintained at an acceptable inlet temperature by equilibration with the outside air, a practice known as free cooling [16], then cooling may become particularly inexpensive as water chillers are not required, and insulation to prevent condensation can be eliminated. Water at 35°C or cooler may be easily obtained anywhere on the planet through evaporative cooling towers and the like [17]; thus, above-ambient-temperature water cooling is an emerging technology. The challenge is to cool the processors, memory devices, switches, and power supplies of this exascale machine with such relatively warm water while containing the device temperatures at acceptable levels. This will require continued innovation in thermal interface materials to meet the needs of mechanical compliance. Additionally, thermal interface materials within soldered components must withstand the increased Pb-free solder reflow temperatures. Field replacement of serviceable parts will demand affordable separable interfaces, and we may move from “quick connects” where we break the water flow to “thermal connectors” where we can break a thermal conduit without introducing high cost or high thermal resistance.

Power delivery

Efficient power conversion requires a thoughtful look at conversion from high-voltage transmission lines down to the final transistor power rails. Today, there can be as many as four power conversion steps, from rectification of, for
example, 240-V alternating current (ac) to direct current (dc), then power factor correcting for the largely capacitive load of the switching electronics back onto the grid, then dc–dc down conversion in two or more steps, to finally the sub-1-V levels expected for exascale processor chip cores. Certainly, reducing the number of conversion steps will be an important first step, and making each conversion step as efficient as possible is a logical next step. We anticipate and expect continued improvements in the efficiency of both ac–dc and external dc–dc converters. Making the last conversion step as close to the processor as possible, either in the processor chip or immediately adjacent to it, will reduce resistance losses from the necessarily large currents. However, as more and more functions are integrated onto the processor chip, a proliferation of different precision voltages in the form of references, static random access memory (SRAM) supply voltages, I/O driver voltages, and receiver threshold voltages requires additional power supplies of modest current, which compete for the precious near-die real estate. This points to an opportunity for on-chip dc–dc voltage conversion, with the mix of on- and off-chip conversion depending on the required area and subsequent efficiency. Resistive series regulators are fundamentally limited to low (~50%) conversion efficiency values because of the inherent resistive divider network; these are suitable for very low-current applications or for voltage trimming. Buck converters are more efficient but require on-chip inductors, which are difficult to produce with high-quality factors. Switched-capacitor circuits may be an effective solution for on-chip voltage conversion. Such circuits have been built with limited efficiency [18], but recent designs using trench capacitors can potentially enable on-chip conversion efficiency values of more than 90% [19]. Indeed, the Second International Workshop on Power Supply On Chip, i.e., PwrSOC10 [20], has garnered considerable attention. Workshop papers cite current densities of up to 10 A/mm²; however, these come with rather high losses on the order of 15% or more. Therefore, as a hedge against on-chip dc–dc power conversion, we must also consider what could be done with near-chip power conversion. For a few amperes per square millimeter, 95% efficient near-chip regulator operating off of a 3.3 V of supply or higher may be interesting. IBM is active in such endeavors.

**Packaging and interconnection**

Power losses due to data communication between multiple die in the system is another significant challenge. An approach to reduce power significantly is to bring the communicating components into close proximity where short-reach data communications are possible. A balanced architecture that combines lowest power short-reach links and optimized long-reach communications channels can be achieved with advancements in packaging and interconnection. For short-distance channels, electrical communication links can provide significant power savings in transmit circuits, short-length wires, and low-power receiver circuits. These channels may consist of ultrashort-reach interconnection structures such as within thinned die stacks or other short-reach channels for data communications such as with high-density packaging between multiple die or die stacks. Power savings on the order of between 10 and 1,000 times reduction compared to traditional off-chip power levels are possible within these types of communications channels. In addition to power savings, the use of these short-reach high-density interconnectivity channels can provide the necessary bandwidth, low latency, and cumulative data rates of communications to scale the performance of the HPC system while also meeting cost objectives. The packaging and interconnection technology requirements drive both higher density interconnection and higher rates of data communication that support efficient utilization of die area for I/O, low power per bit of data transferred, and signal integrity with acceptable noise margins at the targeted data rates needed to meet the performance objectives of the system. Specially developed I/O circuits that can transmit and receive data at low power can provide the power savings while also maintaining a small area footprint compatible with high interconnection density. For long-reach data communications, enhancements in optical communications can be leveraged, as was previously discussed.

Another power-saving approach is to power down circuits when not in use. The ability to support these power savings necessitates the ability to quickly power down and power up on-demand. Highly evolved I/O cells that minimize the energy exchanged per bit of transferred data and that maintain a high data rate but also low latency are critical. Optimization of cells that meet these challenges can provide energy savings and high-bandwidth data communications, as well as meet the objectives for the architecture and design of the system within the cost objectives.

The increased interconnection with I/O density and wiring and resulting power savings can be achieved through the use of 3-D die stacking, 3-D high-density packaging, and high-density interconnection. The exascale system optimization can leverage stacked memory, high I/O processor die, and high interconnectivity packaging beyond traditional solutions. Multi-high-die stack technology with through-silicon vias (TSVs) provides the highest level of integration at the lowest power levels for data communication. Use of fine-pitch, off-chip I/O and off-chip-stack I/O combined with high-density packaging provides another level of power savings while supporting module integration. Optimization of power efficiency, component cost, wiring length, number of channels, operational data rate, and wire length for module integration are ongoing studies toward an efficient exascale computing solution. Examples of technical
advancements of 3-D memory die stacking, 3-D high-density interconnection, and 3-D packaging using lead-free solder interconnection that may be applicable to exascale computing have been reported [21–25].

**Optical communication technology**

Optical interconnects are already a widely accepted solution in today’s HPC systems. Starting with the first petaflops machines, short-reach active optical cables (AOCs) based on vertical-cavity surface-emitting lasers and multimode (VCSEL/MM) fibers have been deployed in large volumes, with more than 100,000 optical ports per system [26] to provide high aggregate bandwidth communications between racks and the central switch. In order to maintain reasonable compute-to-bandwidth ratio, parallel optical links have been further developed to provide broadband connectivity to multichip modules in the IBM P775 system [27]. Small form-factor VCSEL/MM-based optical engines have been developed [28] to provide up to 1.4 TB/s (terabytes per second) of aggregate switching capacity on a switching socket while significantly decreasing power dissipation. The number of parallel optical links in the whole system increased to more than two million, demonstrating an impressive 120% compound annual growth rate [29]. Utilization of optical interconnects is necessitated by the ever-increasing dramatic mismatch between computational operations and memory bandwidth. With this tendency likely to extend to the next decade, massive numbers of parallel optical links on the order of 100 million, which will be used to connect racks, boards, modules, and chips together, are expected in the HPC exascale systems. In order to be massively deployed in an exascale system, optical links should provide reliable communication while maintaining extremely low power dissipation on the order of just a few picojoules per transmitted bit. It is worth mentioning that besides accounting for just electrical-to-optical and optical-to-electrical conversion in transceivers, this power budget should also include all electrical circuitry required to provide high-integrity electrical signaling on a card or on a module for an I/O link. These numbers are approximately 50–100 times lower than what is available with today’s technology. Analogous aggressive scaling is expected for cost reduction for optical interconnects from today’s tens of dollars for transmitted gigabits per second of data to below tens of cents. Although it is expected that further development of the next-generation VCSEL/MM-based transceivers might provide some of this scaling, serious difficulties are envisioned to meet aggressive power and cost savings required for exascale systems while simultaneously maintaining the reliability.

The new technology of silicon photonics is emerging and has the promise of revolutionizing short-reach optical interconnects. This promise is connected to the idea of integrating optical and electrical circuitry on a single silicon die utilizing mature CMOS technology. If successful, such integration might result in significant cost reduction, highly increased density of optical interconnects based on very low power, and highly reliable all-silicon optical transceivers. Various approaches have been explored from building optical circuits such as modulators and photodetectors in the front end of CMOS line (FEOL) [30] to adding additional low-temperature processing steps at the back end of the line (BEOL) [31]. Several products utilizing some of these approaches mostly target the AOC market [32]. However, in order to meet very aggressive requirements for the exascale era in reliability, cost, and power dissipation, significant development is needed.

Over the last several years, IBM has developed a variant of silicon photonics technology called CMOS-integrated silicon nanophotonics [33, 34]. This technology allows monolithic integration of deeply scaled optical circuits into the FEOL of a standard CMOS process. The light signals produced by an external DC infrared laser propagated on a silicon-on-insulator die inside single-mode silicon waveguides with submicrometer dimensions. These optical circuits share the same silicon device layer with the bodies of nearby metal-oxide semiconductor transistors. Several processing modules have been added into a standard CMOS FEOL processing flow. These modules require a minimal number of additional unique masks and processing steps while sharing most mask levels and processing steps with the rest of the conventional CMOS flow. For example, passive optical waveguides and thermooptic or electrooptic modulators require addition of only a single additional mask to the flow [33, 34], since they share the same silicon device layer with the CMOS p- and n-channel field-effect transistors. To build a high-performance Ge photodetector, a “Ge-first” integration approach was developed using a rapid melt growth technique concurrent with the source-drain anneal step [35].

Utilization of advanced scaled CMOS technology allows one to control dimensions of optical nanophotonic waveguides within just a few nanometers, opening the way to build very dense optical circuitry. Indeed, the demonstrated optical devices such as wavelength division multiplexers (WDMs) [36], high-speed electrooptical modulators [37], germanium photodetectors [35–38], and fiber edge couplers [39] are working close to the optical diffraction limit. With such integration density, the area occupied by optical circuitry is becoming comparable to or sometimes even smaller than the area occupied by surrounding analog and digital CMOS circuits that provide signal amplification for robust I/O links.

Demonstrated integration density of just 0.5 mm^2 per channel in IBM CMOS-integrated silicon nanophotonics technology allows the design of massively parallel terabit-per-second-class optical transceivers fabricated on a single CMOS die occupying an area less than 5 × 5 mm^2.
Such integration density achieved in a standard CMOS foundry allows one to expect not only dramatic cost reduction and power savings for optical interconnects but also significant improvement of failure rates potentially approaching the reliability of CMOS circuitry. With such advances, it is possible to expect that Si photonics transceivers will form the interconnect backbone at all levels from AOC connecting racks to standalone transceiver chips at the edge of the card or the module. To meet these expectations, significant additional effort is needed, along with advances in compact, low-power, and reliable lasers, for full realization of all the advantages promised by silicon nanophotonics, as compared to traditional VCSEL/MM parallel links. Both have promise to scale power per transmitted bit to just a few picowatts per bit and even lower. In VCSEL/MM transceivers, this is expected to come from further increase of the line data rate to 25 Gb/s and beyond and utilization of more power-efficient electrical analog circuits designed in advanced CMOS or SiGe technologies. In silicon nanophotonics, analogous or even better power efficiency can be envisioned coming from efficient utilization of a single external laser whose power is shared between multiple parallel optical channels. Further power savings can come as a result of intimate integration of electrical and optical circuitry on the same silicon die, thus minimizing parasitic effects in the more complex hybrid package typical for VCSEL/MM transceivers. With the high level of integration of almost all transceiver components in a single silicon die, the cost structure of transceivers based on Si nanophotonics technology will be dominated mostly by the cost of packaging and testing. Significant breakthroughs are needed in these areas to make this technology more attractive than more traditional VCSEL/MM optical links. Thus, it is most likely that these two technologies will coexist with each other in the next decade for building the most cost- and energy-efficient parallel optical links.

There are two inherent advantages of Si nanophotonics differentiating it from VCSEL/MM that can potentially make this technology more attractive for building exascale systems. One is based on the possibility to utilize WDM concepts to significantly increase the aggregate bandwidth and to minimize the number of optical fibers in the system. Indeed, as opposed to VCSEL-based links utilizing multimode fibers, Si nanophotonics is based on single-mode optics and single-mode fibers. Moreover, utilization of small chirp external modulation, combined with low dispersion in single-mode fibers, allows the design of optical links that are almost independent of distance. Once in the optical domain, the data can be transmitted as close as the nearest analogous compute socket a few centimeters away or as far as a distant rack located at the furthest side of the building or even several kilometers away. The second and probably the most important advantage is the possibility to integrate Si nanophotonics transceivers into a more complex 3-D package in a way very similar to what is envisioned for integration of memory stacks and compute nodes. While utilization of Si photonics transceivers in HPC systems and high-end data centers will most likely start to happen even earlier than the time horizon for the exascale era, the full promise of the technology will be fully appreciated when 3-D integration will mature enough to stack several silicon dies in a single compute socket, as shown in Figure 3. Since from a packaging standpoint, a CMOS nanophotonics transceiver die is not much different from a typical CMOS die, it is indeed possible to envision that such dies can be stacked together using 3-D integration technology with dense through-silicon vias. The sole function of a photonic layer in this vision is to provide very high-bandwidth (up to several tens of terabytes per second of data) optical communication off the socket. Bringing optical communication to the chip level would allow new opportunities for various new architectural solutions. Apparent independence of the communication signal integrity on the distance allows the connection of the 3-D integrated compute socket to other sockets, switching nodes, racks, or memory banks located very far away.

**Conclusion**

The HPC market is expected to continue its strong growth, if we can continue to find solutions that allow year-over-year growth in computing performance at constant cost and power. Indeed, if the current rate of growth continues, we can expect to reach an exaflop supercomputer, achieving a theoretical peak speed of 10^{18} floating-point operations per second, within the decade. This drive to the exascale will demand advances in 3-D packaging, high-speed electrical and optical signaling and efficient power conversion, and cooling and memory technology. These are all areas of active research both within and outside of IBM. The next decade will be extremely exciting as advances in these areas come to fruition.

**Acknowledgments**

The authors would like to acknowledge the leadership, deep technical insight, and deeply rewarding conversations with Alan Gara. The manuscript benefitted from insightful suggestions of Dale Becker. Many colleagues on IBM’s exascale program, too numerous to name, all contributed through their imagination, insight, and ongoing research.

*Trademark, service mark, or registered trademark of International Business Machines Corporation in the United States, other countries, or both.*

**Trademark, service mark, or registered trademark of Top500.org in the United States, other countries, or both.*


40. IBM Corporation, Made in IBM Labs: Breakthrough Chip patents.

41. 90 papers in the field of electronic packaging and holds 105 U.S. patents. He is honored with the National Medal of Technology and Innovation. He has managed the Systems Packaging Group since 1994, where he directs design and packaging of this family of supercomputers, recently presented at the Conf. Lasers Electro-Optics, Baltimore, MD, 2011, Paper PDPB9.

42. 14 : 12

43. Received May 5, 2011; accepted for publication July 4, 2011

44. Paul W. Coteus IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (coteus@us.ibm.com). Dr. Coteus is a Manager and Research Staff Member in the Systems Research department at the IBM T. J. Watson Research Center. He received his Ph.D. degree in physics from Columbia University in 1981. He continued to design an electron-proton collider and, from 1982 to 1988, was Assistant Professor of Physics at the University of Colorado, Boulder, studying neutron production of charmed baryons. In 1988 he joined the IBM T. J. Watson Research Center as Research Staff Member. He has managed the Systems Packaging Group since 1994, where he directs and designs advanced packaging for high-speed electronics, including I/O circuits, memory system design, and standardization of high-speed DRAM, and high-performance system packaging. He was Chairman in 2001 and Vice-chairman from 1998 to 2000 of the JEDEC (Joint Electron Devices Engineering Council) Future DRAM Task Group. He is the Chief Engineer of Blue Gene systems and leads the system design and packaging of this family of supercomputers, recently honored with the National Medal of Technology and Innovation. He is a senior member of the IEEE, a member of the IBM Academy of Technology, and an IBM Master Inventor. He has authored more than 90 papers in the field of electronic packaging and holds 105 U.S. patents.

45. John U. Knickerbocker IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (knicker@us.ibm.com). Dr. Knickerbocker is an IBM Distinguished Engineer, Master Inventor, and Manager at the IBM T. J. Watson Research Center. He is the manager of 3-D Silicon Integration including through-silicon-vias (TSVs), wafer finishing, 3-D integrated circuit die and wafer assembly, and fine-pitch test, modeling, characterization, and reliability evaluation. He received his Ph.D. degree in 1982 from the University of Illinois studying materials science and engineering. He joined IBM in 1983 at IBM Microelectronics in East Fishkill, NY, where he held a series of engineering and management positions during his first 21 years with IBM including director of IBM worldwide packaging development. In 2003, he joined the IBM Research Division, where he has led the development of next-generation 3-D silicon integration. His 3-D research advancements include design, wafer fabrication with TSVs, wafer bonding and thinning, wafer finishing, thin die stacking and assembly with fine-pitch interconnections, wafer stacking, fine-pitch wafer test, mechanical, electrical, and thermal modeling, and characterization and reliability characterization. Dr. Knickerbocker has received an IBM Corporate Award and four Division Awards, and 47 Invention Plateau awards. He has authored or coauthored 188 U.S. patents or patent applications and more than 60 technical papers and publications. He serves as a member of the SEMATECH (Semiconductor Manufacturing Technology) 3-D Working Group. He has been a member of IEEE, International Microelectronics and Packaging Society (IMAPS), and a Fellow of the American Ceramic Society. He has more than 28 years of experience at IBM and is a member of IBM’s Academy of Technology. He continues to actively serve as a member of the Advanced Packaging Committee for the Electronic Components and Technology Conference as an active member of the IEEE Technical Society.

46. Chung H. Lam IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (clam@us.ibm.com). Dr. Lam is an IBM Distinguished Engineer, Master Inventor, and Manager at the IBM T. J. Watson Research Center. He received his B.Sc. degree in electrical engineering at Polytechnic University of New York in 1978. He joined IBM General Technology Division in Burlington in 1978 as a memory circuit designer. He worked on 144-Kb CPROS (cost performance read-only storage), 36-Kb EEPROM and led the development of a 64-Kb nonvolatile RAM. He was awarded the IBM Resident Study Fellowship in 1984 with which he received his M.Sc. and Ph.D. in electrical engineering, at Rensselaer Polytechnic Institute in 1987 and 1988 respectively. Upon returning from resident study, he had taken responsibilities in various disciplines of semiconductor research and development including circuit and device designs, as well as process integrations for memory and logic applications in IBM Microelectronics Division. In 2003, he joined the IBM Research Division at T. J. Watson Research Center to lead the research of new nonvolatile memory technologies. Currently, he manages the Phase Change Memory Research Joint Project with partner companies at T. J. Watson Research Center. He is an author or coauthor of 120 patents and 70 technical papers. He is a member of the Technical Committee of the IEEE International Memory Workshop (IMW), International Electron Device Meeting (IEDM) 2011, International Symposium on VLSI Technology, Systems and Applications (VLSI-TSA), China Semiconductor Technology International Conference (CSTIC), and the International Technology Road Map for Semiconductors Committee (ITRS). He is also a member of the IBM Academy of Technology.

47. Yuri A. Vlasov IBM Research Division, Thomas J. Watson Research Center, Yorktown Heights, NY 10598 USA (yvlasov@us.ibm.com). Dr. Vlasov is a Research Staff Member and Manager at the IBM T. J. Watson Research Center. He has led the IBM team on silicon integrated nanophotonics for on-chip optical interconnects since joining IBM in 2001. Prior to IBM, Dr. Vlasov has developed semiconductor photonic crystals at the NEC Research Institute in Princeton, and at the Strasbourg IPCMS Institute, France. He also was, for over a decade, a Research Scientist with the Ioffe Institute of Physics and Technology in St. Petersburg, Russia, working on optics of semiconductors and photonic crystals. He received his M.S. degree from the University of St. Petersburg (1988) and the Ph.D. degree from the Ioffe Institute (1994), both in physics. Dr. Vlasov has published more than 100 journal papers, filed more than 40 U.S. patents, and delivered more than 150 invited and plenary talks in the area of nanophotonic structures. He served on numerous organizing committees of conferences on nanophotonics. Dr. Vlasov is a Fellow of both the Optical Society of America and the American Physical Society, as well as a Senior Member of the IEEE. He was awarded several Outstanding Technical Achievement Awards from IBM and was named a scientist of the year by the Scientific American journal. Dr. Vlasov has also served as an adjunct professor at Columbia University’s Department of Electrical Engineering.