Lawrence Berkeley National Laboratory
LBL Publications

Title
Optical Interconnects and Extreme Computing

Permalink
https://escholarship.org/uc/item/4mz8t0xx

Journal
Optics and Photonics News, 27(4)

ISSN
1047-6938

Authors
Bergman, Keren
Shalf, John
Hausken, Tom

Publication Date
2016-04-01

DOI
10.1364/opn.27.4.000032

Peer reviewed

and **Extreme**

The leap to "exaflop" computing will require new optical technologies— and new co-design approaches to developing them.

credit

Keren Bergman, John Shalf and Tom Hausken

Optical Interconnects
Computing

A s they confront ever more complex and data-intensive problems, scientists and researchers increasingly look to the next generation of supercomputing—the high-end segment of high-performance computing (HPC). That next generation will play out in so-called exaflop computers—machines capable of executing at least a quintillion (10^{18}) floating-point operations per second (flops). Such a computer would represent a thousand-fold improvement over the current standard, the petaflop machines that first came on line in 2008. But while exaflop computers already appear on funders' technology roadmaps, making

The combination of small market scale and large system complexity means that the HPC system vendor confronts substantial risks in including a new feature.

the exaflop leap on the short timescales of those roadmaps constitutes a formidable challenge.

A key part of meeting that challenge lies in the system interconnects, where optical technologies could conceivably help bridge some of the current gaps in cost and required throughput. To do so, however, will require much more

than "replacing wires" with optics. It will also require the implementation of solutions—such as wavelength-division multiplexing (WDM), integrated photonics and optical switching—that can bring a significant improvement in performance/cost relationships. Each of those solutions, in turn, will call for new technologies that aren't on current vendor roadmaps. And, on the human side, making the leap to exaflop computing will need new "co-design" approaches, in which project development takes into account both the system and component perspectives from the very beginning.

These findings, and others, emerged from an OSA Incubator Meeting on the use of photonics in extreme-scale, or exascale, computing, held in Washington, D.C., on 10 and 11 August 2015. The meeting was sponsored by the Office of Advanced Scientific Research, Office of Science, U.S. Department of Energy (DOE), a key supporter of supercomputing development to address large-scale problems such as climate change and fusion energy research. In this update, building on the outcomes of the workshop, we talk about some of the market and technology characteristics of HPC in general, and its interconnects requirements in particular, that make the jump to exascale computing so difficult—and what the next steps might be.

The march to exascale

Processing power, in floating-point operations per second (flops), of the most powerful computer (red data), the average of the top 500 (orange data) and sum of the top 500 (teal data). Growth has slowed in recent years, owing to the inability to scale to more nodes.
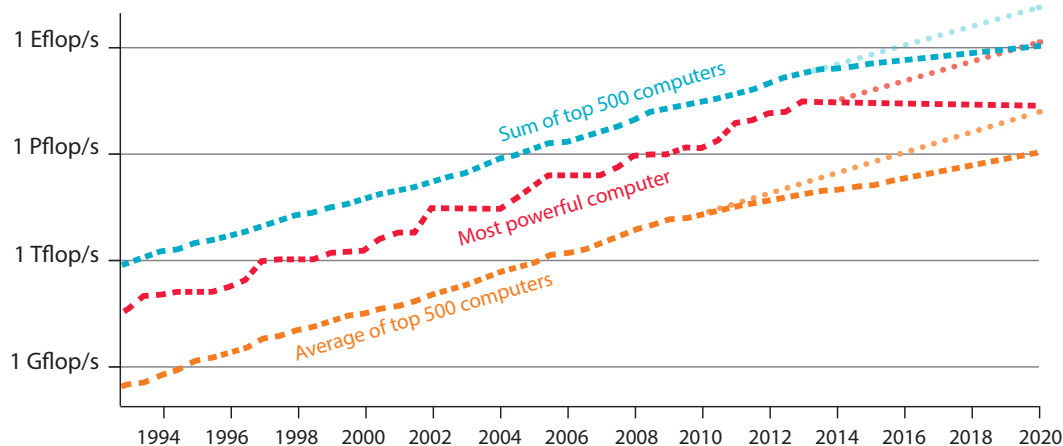
Adapted from E. Strohmier, TOP500 (November 2015).

Historical trend lines   Recent trends

A relatively small market

Understanding the technical challenges of making the exaflop leap begins with an appreciation that, fundamentally, HPC is not a large market—particularly compared with the capital spending on new data centers every year. The International Data Corporation (IDC), a global market-research firm, estimates that spending on supercomputer-scale HPC came in at US$3.2 billion in 2014, a year when spending on large-scale cloud data center infrastructure was about US$26 billion. Total data center infrastructure spending will reach about US$100 billion in 2015, according to IDC—an order of magnitude larger than total HPC spending—and it continues to grow rapidly.

The U.S. government, especially DOE, is a major customer of HPC systems. But as an end-user DOE controls only the application, which has become increasingly driven by software. DOE is not vertically integrated and is neither sufficiently large nor sufficiently resource-rich as a customer to lead hardware development. While DOE plans to buy new bleeding-edge systems on a regular schedule, that amounts to only four highest-end systems spread over two alternating planning cycles—not enough to drive innovation without substantial co-investment. Meanwhile, DOE's small-business grants are not large enough to commercialize many optical technologies that might accelerate exascale development.

Consequently, DOE must rely on an assemblage of HPC vendors to deliver a "box" with the necessary features—a key component of which i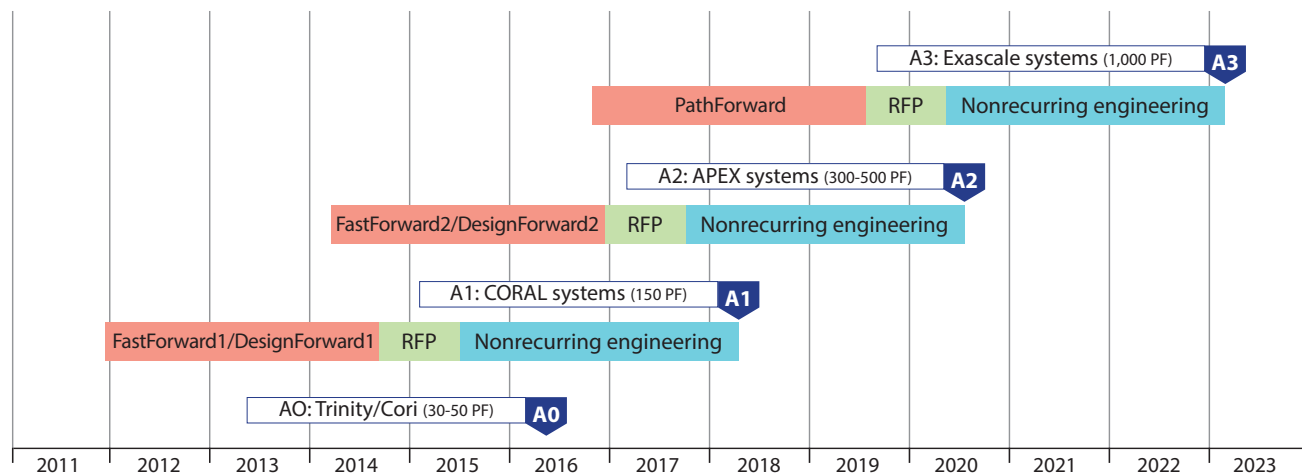s the interconnects. That, in turn, has led to a fragmented HPC supply chain that makes commercialization even more challenging for supercomputing than for data center or telecom networks. Within the data center arena, giant, vertically integrated end-users like Google and Facebook have enough scale and resources to drive innovation for their own cost-driven needs. Even so, they have needed to make partnerships with suppliers, such as by relaxing specifications to allow more competitors and lower the cost. And even doing business in the strong data center segment has not always proved profitable for component suppliers.

Complexity and end-user needs

All of this means that the HPC community cannot rely on market scale to solve the hardware challenges of moving to the next generation. And those challenges are formidable. While current customer roadmaps boldly include exascale computing within a few years, vendors are still far from building a real system—especially one that meets the usability targets set by DOE for delivered application performance on top science applications.

An exascale computer has to manage about 100 to 1,000 times more parallelism than today's leading-edge supercomputers, with as many as one million interconnect endpoints to get to each socket in the system. The difficulty of getting to this level of parallelism within cost constraints is not fully appreciated. Porting existing applications to these new platforms to run efficiently across the vastly more parallel system constitutes another huge challenge.

Exascale platform deployment timeline

Based on data from DOE Exascale by Steve Binkley; http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/201512/2015-1209-ascac-03Binkley.pdf

iStock

Clearing the hurdles

Next-generation, extreme computing creates some unique customer priorities. Optical interconnects could help fulfill them—but only with some new technology and thinking. Here's a capsule summary of some near-term hurdles—and some possible ways to clear them.

COST

The hurdle: Given the huge number of interconnects in exascale systems, the cost of optical links would need to drop from a current average of roughly US$1 per Gb/s to only around US$0.05 per Gb/s to be competitive with short-link copper interconnects.

Clearing the hurdle: Leveraging volume manufacturing, spurred by demand from the much larger data center market, could help drive costs down.

POWER

The hurdle: While in principle reducing power requirements is a key argument for optical interconnects, the energy consumed for current-gen optical links is estimated to be more than twice the energy per bit for copper.

Clearing the hurdle: Suppliers are working to reduce energy requirements to a few pJ/bit by early in the next decade, a level at which optics could work within DOE's stringent exascale computing budgets.

BANDWIDTH DENSITY

The hurdle: Optics could help expand bandwidth density in exascale systems, but it is hard for optics to add value in short links where electronic interconnects are repeaterless.

Clearing the hurdle: Wavelength-division multiplexing and integrated photonics could provide significant scaling, and a nonlinear boost to performance well beyond what's possible with electronics, though these solutions bring trade-offs of their own.

PACKAGING

The hurdle: Exascale computing requirements could increase component count by a hundred times, requiring innovative packaging to meet DOE's reliability standard of less than one fault per day.

Clearing the hurdle: Engineers are experimenting with packaging that places the optics closer to the processor, and other board design schemes to reduce failure rates.

The combination of small market scale and large system complexity means that the HPC system vendor confronts substantial risks in including a new feature. DOE cannot mandate novel uses of optics unless the system vendor is fully behind it, because the cost to correct subsequent issues with performance or functionality might have to be borne by the vendor. For example, a vendor once delivered a computer with connector issues, and was forced to replace the connectors—18,000 of them. Consequently, DOE has programs for "precompetitive" investments (such as "Fast Forward" and "Design Forward") to share the risk of introducing new technologies to the product line.

DOE's timeline for procuring new computers used to be two years from start to finish, which is too short to have substantial influence on system design. DOE has extended the procurement timeline to four to six years, and as the buyer it has sufficient lead-time to influence the design during this period. DOE includes funding for NRE (nonrecurring engineering) expense to directly fund new technologies and features during this extended lead time for system acquisitions once the acquisition contract is signed.

Whatever the timeline, the exascale computer customer has some clearly visible priorities. There are upper limits on the acquisition cost and operating cost—including, importantly, the cost of electrical power. The customer also requires high data throughput, low latency, low raw bit error rates and minimal error correction, and high reliability (including failure detection and early warning on component failure). Despite its importance, an HPC system end-user commonly doesn't want to deal with the communication network—it only wants to see simple communication abstractions. In the next few sections, we explore all of these priorities in the context of the current state of the art for optical interconnects.

The cost challenge

The first of the customer priorities, cost, is the greatest challenge for achieving exascale computing. The capital acquisition cost for a leadership-class computing system is estimated today at about US$100-US$150M, and will likely rise in the future to US$200 million (a figure we will use here for illustrative purposes). The total cost of ownership, including both acquisition and operating costs, over an expected five-year life is about twice the acquisition cost.

The cost of processing and memory squeezes the cost of the interconnects to about 10 to 20 percent of the overall capital cost, with 15 percent as a realistic upper limit. At US$200 million for the largest systems, that leaves at most US$30 million for interconnects.

If power consumption were to scale linearly with performance, the computer's power needs would soon exceed its operating cost budget.

At this point, with current technology, the math starts to look unpromising for optical interconnects. Assuming 100,000 to one million endpoints per exascale system, and a best-case average of 1.5 optical links per endpoint, there would be as many as 1.5 million optical links, each at 400 Gbps. To meet the US$30 million interconnect budget, the optical links would have to cost at most US$0.05 per gigabit per second (Gbps), versus about US$1/Gbps today—already a very aggressive baseline. And that cost estimate doesn't even include the electrical and other optical interconnects that are needed.

Based on those numbers, optics clearly has a long way to go to compete with copper interconnects for short links. The current estimate is that photonics is 10 to 30 times more expensive than copper for links less than a meter long. For the very shortest links, copper is almost free.

The key to overcoming cost challenges will be to leverage volume manufacturing, where possible, from a larger-volume market segment, and use that to reduce the cost for HPC end-users. Suppliers must invest many millions of dollars for each redesign of an electronic interface product, so they have to make choices over which designs to revise and which to reuse. Working together as an industry reduces the risk and cost of these choices. There is already consolidation of physical interfaces for HPC connectors, but how far up the supply chain vendors can leverage synergies between the HPC and data center markets remains an open question.

Power: The elephant in the room

Typically, end-users are not primarily concerned with the electrical power requirement. Yet if power consumption were to scale linearly with performance, the computer's power needs would soon exceed its operating cost budget. The power dissipation is also a factor in the cooling requirements on the board and within the rack.

A common rule of thumb for the cost of electrical power is about US$1 million per megawatt per year. Current computers operate in the 5- to 7-MW range, and DOE is prepared to go to as much as 20 MW for exascale. Over a nominal five-year life for a new HPC system, therefore, the cost for electrical power could amount to as much as US$100 million.

Memory and processing squeeze the available electrical power for the communication network to about 20 percent of overall system power, or about 4 MW. Electronic switches

consume as much as half of this, leaving about 2 MW for the interconnects. Using the previous example, this works out to less than 3.3 pJ/bit per optical link end-to-end across the system, including switching and intermediate connections, compared with about 35 pJ/bit today. Or, viewed another way, an exaflop computer using 20 MW of electrical power amounts to 50 Gflop/J (20 pJ/flop) across the system. Today's systems commonly operate at 2 Gflop/J, with top systems at 5 Gflop/J, a factor of 10 from the goal.

Not every flop results in a byte sent across the network, however. Nonideal "verbosity" constrains the power consumption further. Verbosity measures the available bytes of communication bandwidth per computation, measured in flops. A verbosity of one byte/flop may be the ideal needed for a high-bandwidth interconnect, but a system-wide goal of 0.1 byte/flop may be more realistic. Systems today are more likely 0.001 byte/flop or worse, however, meaning that communications-intensive applications suffer in performance.

Low utilization of the links also reduces the power efficiency. Unlike copper interconnects, optical links use some power even when the channel is not transmitting or receiving. Assuming no penalties due to WDM, the laser might require about 0.1 pJ/bit—but 10 percent utilization drives this to 1 pJ/bit. Utilization can be improved, but not to 100 percent, because high utilization would lead to contention issues and queueing.

Suppliers may reduce the energy requirement to a few pJ/bit by 2023 and eventually to hundreds of fJ/bit by 2030, or perhaps much sooner. The industry has been talking about this goal for many years (if only as a way to justify R&D funding), but compared with many challenges, it is "just" engineering, while the cost of ownership is the greater challenge. Meanwhile, the current estimate is that the energy requirement for optical links is greater than two times the energy per bit for copper.

Bandwidth density and WDM

Given the cost and power situations today, where can optics make a difference in enabling exaflop systems? The answer could lie in improving the bandwidth density, where integrators may have a large opportunity to improve performance. Data rates have improved about 10 times in 10 years, but the power consumption has not scaled at the same rate. Another

iStock / kentoh

10 times could be achieved if Ethernet-compliant products advanced from 100 Gbps to 1 Tbps; with fiber ribbons, the industry might even get to 10 Tbps in less than 10 years.

The repeaterless reach of electronic links is becoming very short, even at 25Gbps. Copper interconnects currently operate with differential signaling at 8-16 Gbps per lane. Greater lane rates are possible, but are less energy efficient. Copper can support 20 Gbps/mm at the periphery, and 2 Gbps/mm through the connectors.

In light of that, some believe that interconnects inevitably will need to employ WDM and integrated photonics, because of the scale promised by these technologies. As previously with long-haul fiber links, WDM could provide a big "non-linear" opportunity to improve performance relative to cost, power and complexity. The initial pull of WDM-compatible single-mode fiber is expensive, but scaling performance by adding more wavelengths is relatively inexpensive. Likewise, coherent solutions based on integrated photonics are not as challenging at the shorter distances inside exascale computers as they are at longer distances.

Nonetheless, some vendors are skeptical that highly integrated photonic solutions are necessary so soon for interconnects. And solutions like WDM bring their own trade-offs. The total bandwidth is the product of the number of lanes, the number of wavelengths per lane, the data rate per wavelength and the signaling rate within the nominal data rate. There are opportunities to extend bandwidth with each of these, but each poses challenges. For example, multilevel signaling (PAM4) extended data rates, but reduced margins and increased error rates, leading to the need for strong forward error correction (FEC), which in turn adds latency and power consumption. This trend is acceptable for data centers, but not for HPC.

Packaging and reliability

Finally, components for exascale computing will require innovative packaging, bringing on "the revenge of the mechanical engineers." With pinout for advanced packages already up to 6,000 pins, no opportunity remains to scale performance simply by increasing the pinout. Moreover, DOE views reliability as a critical issue. As computers reach a million cores or more, the greatest source of potential failure becomes the sheer number of solder joints and cables required. DOE's target rate for exascale system is less than one fault per day. Meeting that requirement will be a challenge given the potential hundred-fold increase in component count and the desire for substantially greater bit rates.

Panel mounting of optical I/O is becoming less practical. Instead, the optical-to-electrical (O-E) and electrical-to-optical (E-O) conversion must be closer to the processor than in the past. The conversions cannot go as far as being on the electronic die itself, however, suggesting an interposer design that places the optics alongside the memory and the processor, or allowing replacement of parts after assembly. However, there continues to be disagreement whether the laser can be placed on the chip or must be placed apart from the electronics.

The system vendors also have some highly specific preferences with regard to packaging to minimize costs. And boards with optical traces are considered preferable relative to optical cables, for the simpler design and lower failure rate. Such boards have only been demonstrated, however, and are still immature for a large design.

From interconnects to optical switching

Beyond interconnects, an important role for optics might lie in switching (see "Optical Networks Come of Age," OPN, September 2014, p. 50). Electronic packet switches require a large share of the network power budget. That suggests that switching might be "the real gem"—an opportunity where optics can make a substantial difference.

Addressing switching requires an understanding of the traffic patterns inside supercomputers. The size of DOE's supercomputing jobs scales with the size of the computer, but the distribution of jobs by size across different sizes of computers follows a power law, from very large jobs to small ones that have to be triaged for efficiency. With the rise of large-scale parallel data analytics applications, data center operators are now starting to see more of a power law distribution as well.

There is no single best topology that addresses all traffic patterns, however. Data centers have so-called

Some believe that interconnects inevitably will need to employ WDM and integrated photonics, because of the scale promised by these technologies.

elephant flows that are well-suited for optical circuit switching. The value of optical circuit switches in supercomputers is less understood, but a number of research studies of HPC applications over the past decade demonstrated substantial and persistent structure to communication patterns.

Yet there is some doubt about optical switching as a grand solution to the exascale challenge, because optical cross-connects introduce a loss in signal that must be restored. The low average link utilization also limits the efficiency of optical switching. A trade-off exists between improving utilization and minimizing queuing of data. In a strictly optically switched network, the bits cannot flow while the circuit is being reconfigured, yet there is no optical buffer to store the bits during that time. If the packets are in the range of a few nanoseconds in extent, then the reconfiguration time has to be short and rare—in the range of 1 ns—and the optical switch has to have agility on a nanosecond timescale.

The lack of such switches makes all-optical packet-switched networks impractical. With photonics limited to the circuit-switched domain, solutions will involve hybrids of photonic circuit switches and electronic packet switches that minimize the number of conversions to the electrical domain for buffering and packet switching.

MEMS-based optical switches might achieve sufficient port densities to make a difference. The port cost of MEMS switches is currently hundreds of dollars per port, however, because so few are sold. HPC is not a large enough market to reduce the port cost sufficiently, but the use of MEMS switches in data centers may provide needed volume to make them attractive for the HPC market, too.

The case for co-design

Taken as a whole, this broad analysis suggests that optical interconnects and switching have a role to play in enabling exascale computing, but that new technologies and new ways of working will be necessary to make the leap. The industry will not magically achieve a successful exascale goal—especially not by the aggressive DOE target dates—unless hardware, software and application engineers work together to develop new architectures and code bases that work effectively in an integrated environment. This kind of hardware/software co-design has been at the core of recent exascale computing development.

Indeed, bridging the gap between system integrators and equipment and optics suppliers, and providing a freer exchange to enable co-design, was both a key motivator and a key closing message of the 2015 OSA Incubator. That gap remains formidable, however. Optics suppliers ask for specifications, without knowing where novel designs might help to make breakthroughs in overall system performance. In return, integrators commonly ask for product roadmaps with incremental improvements in performance and lower cost. The integrators can't develop novel architectures with parts that don't exist—but the products don't exist because there are no current customers for them.

Yet increasing awareness of the need for co-design may provide reason for guarded optimism. And, last July, the Obama administration announced a new National Strategic Computing Initiative aimed at ensuring U.S. leadership in HPC. That executive order was followed by a government-hosted workshop on October 20, and the U.S. Congress has reportedly approved potential spending of more than a billion dollars over the course of a decade. There is little in the project that's specific to optics. Still, if the funding is forthcoming, the ratcheting-up of U.S. government support could provide an extra impetus toward making these complex new computing systems a reality. OPN

Keren Bergman (kb2028@columbia.edu) is with the Department of Electrical Engineering, Columbia University, New York, N.Y., USA. John Shalf (jshalf@lbl.gov) is with the Lawrence Berkeley National Laboratory, Berkeley, Calif., USA. Tom Hausken (thausken@osa.org) is OSA's senior industry advisor.

References and Resources

▶ J. Shalf et al. "Rethinking hardware-software codesign for exascale systems," IEEE Comp. **44**(11), 22 (2011).

▶ P. Kogge et al. "Computing trends: Adjusting to the 'new normal' for computer architecture," Comp. Sci. Eng. **15**(6), 16 (2013).

▶ S. Rumley et al. "Silicon photonics for exascale systems," J. Lightwave Tech. **33**, 547 (2015).

▶ www.osa.org/en-us/meetings/osa_incubator_meetings/photonics_in_exascale_computing_incubator/

▶ www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative