Hewlett Packard Enterprise

Exascale: A race to the future of HPC

The next generation of computing

Technical white paper

Contents

Back in summer 2008, the first HPC system capable of sustaining more than 1 petaflop (defined as 10^{15} floating-point operations per second [FLOPS] with the HPL benchmark[1]) was launched. It appeared as the world's fastest supercomputer ranked number one on the TOP500 list with a power consumption of about 2.4 MW. It was considered a breakthrough achievement and it marked the end of a row of evolutionary steps.

But shortly afterwards, while first researchers were adapting their codes to petascale systems, the HPC community began a serious discussion about scientific benefits and design challenges of systems faster by a factor of 1,000 or capable of reaching 1 exaflop (10^{18} FLOPS). A paper published in fall 2009, listed the unprecedented opportunities for science as well as critical advances for the U.S. energy needs and security.[2] One year later, in fall 2010, the Department of Energy's (DoE) Office of Science issued an in-detail report on exascale computing.[3]

Similar discussions took place also in China,[4] Europe,[5] and Japan[6] resulting in multiple independent roadmaps toward exascale.

More recently, the United States' National Strategic Computing Initiative (NSCI)[7] aims to drive a path to exascale through cooperation among the nation's technical leadership agencies, including the DoE and the National Science Foundation (NSF). In parallel, an initiative of the Horizon 2020 (the EU Framework Programme for Research and Innovation) is to deliver a broad spectrum of extreme scale HPC systems and develop a sustainable European HPC ecosystem.[8]

It should be noted that exascale computing has to be seen in conjunction with Big Data as a recent paper "Exascale Computing and Big Data" outlines in detail.[9]

While the discussion continued, new number one systems have emerged on the TOP500 in the form of a step function. The current number one system has a peak performance of 1/8 exaflop and sustains about 1/10 exaflop with HPL at a power consumption of about 16 MW.[10]

And here is the challenge: How to increase performance by about a magnitude while staying in the same power envelope? It is obvious that such a goal cannot be reached through additional evolutionary steps—a technological transformation is needed across multiple aspects of a system architecture. We will address this step-by-step.

With the end of Moore's Law,[11] there are some major difficulties to overcome. Compared to the current number one in the TOP500 we want to achieve a ten-fold improvement in computing performance with only a small increase in power. Thus, we have to find a way to deliver FLOPS with less energy consumed. This involves a number of changes in the way we design and provision large-scale computing systems, placing less emphasis on reaching peak FLOPS and more focus on holistic system design through an enhanced memory subsystem and more energy-efficient data motion.

Is Moore's Law still adequate to project growth of computational capability?

In 1965, Gordon Moore, then director of research and development for Fairchild Semiconductor stated, "the complexity for minimum component costs has increased at a rate of roughly a factor of two per year." This observation, later dubbed "Moore's Law," has become synonymous with the exponential growth in processing speed and efficiency over time.

Moore's Law has brought the computing industry to a place where transistors in silicon are now essentially free. In fact, over the last 23 years, the number of transistors per processor has increased by a factor of 2,300, as Figure 1 shows.

Getting the data in and out of the processor is another story. The graphic shows that from 1993 to 2016, pins per package have only increased by a factor of 7.4X. Although the signaling rate has also grown by about 40X, in terms of input/output (I/O) we are struggling to keep pace with the growth of computational capability. This means that while we can pack a lot more compute into a square centimeter of silicon, without new approaches for moving greater amounts of data in and out of the processor package, we won't see a comparable improvement in application performance.

[1] Top 500 List, June 2008
[2] White paper on the Major Computer Science Challenges at Exascale (Al Geist, ORNL, and Robert Lucas), ISI, February 2009
[3] Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee Report on Exascale Computing (Steve Ashby), 2010
[4] China's Exascale Supercomputer Operational by 2020, Chinese Academy of Sciences, June 2016
[5] European Exascale Projects initiative
[6] Fujitsu picks 64-bit ARM® for Japan's monster 1,000-PFLOPS super, The Register, June 2016
[7] Executive order—Creating a National Strategic Computing Initiative, the White House, July 2015
[8] Horizon 2020 High-Performance Computing (HPC) Programme Page
[9] Exascale Computing and Big Data (Daniel A. Reed and Jack Dongarra), Communications of the ACM article, July 2015
[10] Top 500 List, June 2016
[11] The Long Good-bye, IEEE Spectrum, April 2015

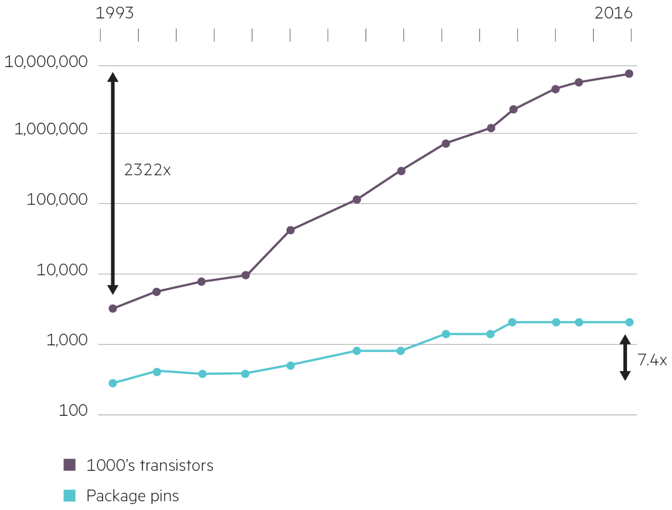

Figure 1. Number of transistors and pins per processor (1993–2016)

The energy consumption challenge

The recently released PathForward Technical Requirements established a power envelope of 20 to 30 megawatts for a capable exascale system. In 2010, the ASCAC report estimated that scaling current systems to an exaflop would consume more than a gigawatt of power. Despite recent advances in this respect—also visible in the TOP500 and the Green500—significantly increasing the energy efficiency of exaflop computing systems is the quintessential issue. This requires new system designs and computing technologies that will reduce the energy requirement to a manageable and economically feasible level.

Troubled memory

On today's platforms, memory capacity and bandwidth struggle to keep pace with the increase in FLOPS. Future HPC platforms will demand both increased memory capacity and very high bandwidth to operate efficiently; for these reasons, exascale supercomputer design must begin by addressing the memory subsystem.

The ratios of memory capacity to FLOPS and memory bandwidth to FLOPS are compelling indicators when evaluating system balance. A quick examination of the specifications for some of the world's most powerful supercomputers demonstrates that while the aggregate figures may seem impressive at first glance, the trends for memory capacity, memory bandwidth, and interconnect bandwidth vs. FLOPS are all headed down.

To boost memory performance, co-packaging is emerging. It refers to memory stacked on the processor or stacks of memory that sits on the same substrate as the processor. High-bandwidth memory stacks are reaching the market today. As we enter the era of co-packaged memory, we will confront a challenging conundrum: with high-bandwidth memory comes smaller capacity and with a non-volatile memory (NVM) comes large capacity but less bandwidth.

For some time now, we have been blessed with the simplicity of (double data rate) DDR-DRAM as a single memory tier. While the next iterations of the DDR standard will prevail in the general-purpose server market for quite some time, DDR is beginning to reveal some serious shortcomings as the last, antiquated, parallel bus in our systems. Consider this example depicting bandwidth per pin density on a modern CPU.

Table 1. Bandwidth per pin density on a modern CPU

	Socket pins	Percentage of socket pins	Data bits	Transfer rate (Gbps)	Bandwidth per pin (Gbps)
Memory	660	33%	384	2.4	1.396
PCIe	192	10%	96	8	4.000
Proprietary	175	9%	40	9.6	2.194
PCH	16	1%	8	5	2.500
Power/Ground/Misc.	968	48%			

The share of pins allocated to DDR adds up to 33 percent on a modern processor (see Table 1), meaning that excluding power and ground, more than half the data pins are tied to a parallel memory bus that keeps lagging behind. Now imagine if those pins could be generalized and instead tied to Serializer/Deserializer (SerDes) interfaces, allowing the system to move away from DDR memory and send and receive data serially at high speed. For example, 128 lanes of next-generation SerDes running at 56 Gbps would result in 1.792 TB/s of bidirectional bandwidth. Adding up to 512 pins, accounting for differential pairs and input and output lanes, results in 28 Gbps of bandwidth per pin on average, which is a 20-fold improvement over DDR2-400.

Today's HPC systems have reached a point where DDR cannot keep pace. Furthermore, pins dedicated for specific protocols are pins that can't be used in a flexible way. So while everybody claims to want more efficient computing systems, the fact is that many still deploy systems that achieve high peak FLOPS—the recent TaihuLight system[12] being a great example of such—but that may not be the best choice for real application workloads.

Because DDR is quickly becoming outdated, next-generation systems will need more capable memory subsystems with co-packaged DRAM and pools of NVM to enable better data motion in and out of the compute engine. These systems will not only achieve enhanced floating-point performance but also greater overall computing and energy efficiencies, thanks to adequate aggregate memory bandwidth and capacity.

Memory-centric computing and the Machine

In order to achieve exascale by the early 2020s, a transition to a faster and more general memory interface will be necessary to access both close and distant memory devices using a unified protocol. This memory protocol will have a notion of near vs. far, large vs. small, and fast vs. slow. It loads or stores domain that can be used to map addresses at large scale, without necessarily trying to maintain coherence using a hardware directory protocol.

While co-packaged memory will provide the critical growth in bandwidth, greater capacity will have to be supplied through other means. The good news is that several technologies are lining up as a persistent replacement to DIMMs, with greater capacity, comparable latency, and bandwidth. Hewlett Packard Enterprise has been hard at work on memristor[13] for several years now, and the fruits of those efforts are nearly ready to go to market. With such a device, we can now envision a world where the memory becomes the center of the system architecture or, in other words, memory-centric computing. Once attached to a serial memory controller, the persistent memory now lives on the fabric and can be accessed from anywhere. Such memory can then be co-located closer to the processors in a more distributed way, or aggregated in larger pools allowing for the re-unification of storage tiers such as the burst buffer and the parallel file system.

The Machine is our strategic vehicle to advance memory-centric computing and other enabling technologies, which aims to integrate standard cores, application-specific cores, memory, management, and fabric all in a single package. As Hewlett Packard Enterprise builds early prototypes, higher-level research around distributed metadata, disaggregated namespace, and other fundamental I/O and filesystem capabilities will be driven forward.

[12] Top 500 List, June 2016

[13] X-ray Experiments Show Hewlett Packard Team* How Memristors Work, SLAC National Accelerator Laboratory, June 2016
*= With reference to the Hewlett Packard Labs team

The evolution of fabrics and topologies

We need energy-optimized and commoditized optical communication in order to realize the promise of a capable and affordable exascale system. Data motion is the most energy-intensive activity in supercomputers and as we are quickly reaching the limits of electrical SerDes, next-generation optical technologies will offer the data transmission capabilities that will be required to achieve the next level in HPC.

Using today's HPC interconnects, which operate in the range of 50 to 100 picojoules per bit, energizing an exascale system's fabric alone would consume more than the DoE's 20-megawatt target. So we have to build a physical signaling layer that is an order of magnitude more efficient. At Hewlett Packard Enterprise and other companies and labs, a very "energetic" optical technology research and development program on vertical cavity surface-emitting lasers (VCSELs) and silicon photonics is shooting to deliver the highly desired target of 1 picojoule/bit in the next decade.

As we scale to hundreds of thousands of endpoints on the fabric, common knowledge and widely accepted topologies should be revisited. In systems of this scale, the interconnects will have a significant impact on cost, power, complexity, and performance. Multi-dimensional all-to-all topologies, in particular, the HyperX,[14] might prove far more usable than a fat-tree, 3D torus, or even a dragonfly, providing for reduced network latency through fewer switch hops, high bisection bandwidth, better path diversity, and improved scheduling flexibility. The HPE vision of the HPC fabric of the future co-locates the switches very close to the processing elements, resulting in a simpler structure to deliver the high levels of performance required for exascale, all within the narrow energy envelope.

System resiliency despite hardware failure

Reliability, availability, and serviceability (RAS) is another key issue surrounding exascale system design. In order to offer the required computing capability, exascale machines will contain tens to hundreds of thousands of nodes and will shatter the expectation that it is possible for every component to be online at all times. Supercomputing on such a large scale must be able to tolerate some degree of hardware failure without inhibiting workload execution.

As the ultimate goal is to provide exascale computing on a continual basis without interruption, significant improvements to existing RAS systems must be made. Advanced analytics and modeling capabilities can be integrated to predict hardware failures and monitor the reliability of individual components as well as the system as a whole on an on-going basis. Predictive fault tolerance measures can also be put in place to migrate the compute away from the components on the verge of failure, therefore, increasing the overall system resiliency.

As an example, parallel file systems have now been proven inadequate to handle supercomputers checkpointing at scale. While burst-buffer implementations may provide an interim solution, the long-term goal should be to provide a fast storage tier that allows checkpointing in seconds, not minutes, such that the workload can go back to execution mode faster while also allowing for more frequent checkpointing. With such a capability, one can now envision allowing for standby servers across the system that would dynamically take over for a crashed server, or even better, pick up the latest checkpoint from a server about to crash, in a form of predictive resilience, where failures are handled before they actually happen.

Inevitably, increasing the resilience of exascale systems will require a holistic approach incorporating multiple hardware and software technologies geared toward both predicting crashes and finding ways to keep the system stable despite failures.

Programming at exascale

We are reaching a point where Big Data is becoming too big to move and computing systems simply cannot analyze data if they cannot ingest it. We need to allow data storage in fast, in NVM tiers and make the fast storage tier work for more than checkpoints: for post processing, for common data sets, for global shared data structures, and in some cases for fast swap files that soften the limit of in-node fast memory. Programming to a new storage hierarchy becomes a key software need.

Exascale system architectures will break records for parallelism—in fact, a billion-way concurrency—and this substantial increase in system concurrency will challenge software, hardware, and applications. Programming paradigms are critical in order to make effective use of these never before seen levels of concurrency and equally new, giant data sets.

[14] HyperX: topology, routing, and packaging of efficient large-scale networks, Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis, 2009

We see growing investment in a variety of possible solutions, under the MPI+X umbrella. We now have distributed systems with shared-memory nodes and where MPI ranks on the same node can share memory. New interconnect mechanisms can allow access to data across the Machine, creating new possibilities for controlled memory sharing in bigger contexts. Thus, there are +X alternatives that will not displace message passing, but can coexist and complement it.

Node complexity can make programming harder. That node complexity has been driven by the demand for peak performance but achieved performance across a range of applications has been harder to accelerate. Our common goal should be to try to make a one-level memory on the node work, with tremendous simplification to the programming model compared with explicit management of a two-level memory system.

The HPE contribution

Hewlett Packard Enterprise's leadership position in the HPC market provides the unique set of capabilities needed to drive innovation in the future of computing. Leveraging the largest server revenue in the IT industry,[15] the targeted R&D spending will then fuel invention to the benefit of exascale while also bringing it all to a point of affordability and availability for every enterprise—large or small.

To help address the many challenges on the path to exascale, Hewlett Packard Enterprise is also leading an industry-wide approach that will revolutionize system architecture. The development of a new and open protocol, temporarily dubbed the next-generation memory interface (NGMI), will increase flexibility when connecting memory devices, processors, accelerators, FPGAs, and switches. It allows the system architecture to better adapt to any given workload. Our advances in silicon photonics will drive increased input and output to the computing elements without exploding the energy budget. Then, emerging NVM technologies such as the memristor will provide the throughput to compete with DDR, but in a persistent and more energy-efficient way.

The HPE approach to exascale is geared to breaking the dependencies that come with outdated protocols. An open architecture will help us foster a vibrant innovation ecosystem and drive the industry to rethink how next-generation computing systems will now be built.

Conclusion

As exponential data growth reshapes the industry, engineering, and scientific discovery, success has come to depend on the ability to analyze and extract insight from incredibly large data sets. Exascale computing will allow us to process data, run systems, and solve problems at a totally new scale and this will become vitally important as problems grow ever larger, ever more difficult. Our unmatched ability to bring new technology to the mainstream will provide systems that are markedly more affordable, usable, and efficient at handling growing workloads.

At Hewlett Packard Enterprise, innovation is our legacy and our future. Driving an NGMI, core innovation in fabrics and topologies, commercialization of silicon photonics, and revolutionary new memory devices such as the memristor, Hewlett Packard Enterprise stands at the forefront of the next wave of computing, all the way to exascale.

Learn more at
labs.hpe.com/research/themachine/

[15] IDC Worldwide Quarterly Server Tracker, March 2016

Sign up for updates

Hewlett Packard
Enterprise

4AA6-7771ENW, September 2016