

# Paving the Roadmap to EXASCALE

Several full science applications are now running at more than a sustained petaflop/s on Jaguar, the U. S. Department of Energy's Cray XT5 supercomputer at ORNL. This capability has kindled a new excitement in the science community to use this resource to make the next science breakthroughs. With Jaguar enabling petascale science today, computer scientists are starting to look ahead to the next decade and ask, what will it take to enable sustained exascale science?

The path from terascale to petascale was driven by the growth of multi-core processors.

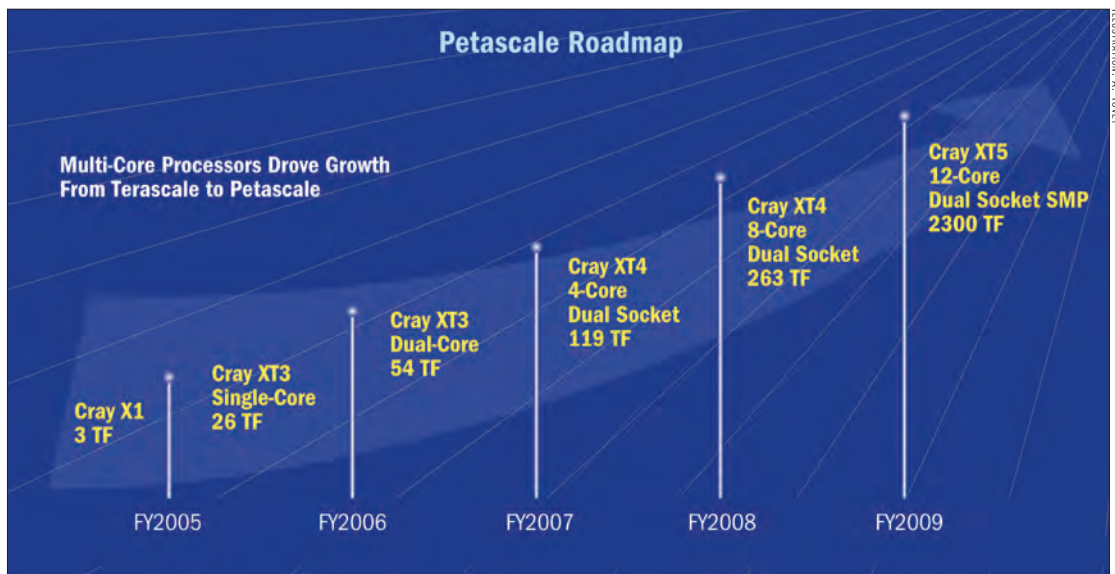
## Petascale Roadmap

The DOE Office of Science Petascale Roadmap started in 2004 with the creation of its Leadership Computing Facilities at Argonne and Oak Ridge national laboratories. Four years later in November 2008, Jaguar, the Cray XT5 computer at Oak Ridge National Laboratory (ORNL) ran the first sustained petaflop/s application in the world, marking the beginning of the era of petascale science. Jaguar was upgraded in 2009 to a peak rate of 2.3 petaflop/s (PF), which is 1,000 times more powerful than ORNL's Cray X1e that was established at the beginning of the Office of Science's Petascale Roadmap. Before describing the Exascale Roadmap to achieve the next 1,000× jump in computational capability, it is important to review the challenges and drivers that allowed the Office of Advanced Scientific and Computing Research (ASCR) to successfully execute its Petascale Roadmap, shown in figure 1.

The path from terascale to petascale was driven by the growth of multi-core processors. In 2005 the 26 teraflop/s Cray XT had only one CPU core on each compute node. Over the next three years, the computational power of this system doubled three times as the number of cores doubled to two, then four, and finally eight cores per node, giving the Cray XT system more than a quarter of a petaflop/s capability. In late 2008, a 1.3 PF Cray XT5 system was installed at ORNL and a year later upgraded to faster processors to achieve its present 2.3 PF computational power using 12 cores per node. The path

to petascale was not without its bumps. The market trend to multi-core processors allowed a petascale system to be built, but it did not enable petascale science. For that to happen, the Leadership Computing Facilities had to overcome four major challenges: operating system scalability, file system performance, message passing scalability, and application readiness to use multi-core.

When ASCR started down the Petascale Roadmap, there were serious concerns that the Linux operating system would not work at 100,000 processors. At that time, Linux had never been scaled to more than 5,000 processors. The concerns were two-fold. If the algorithms used inside the operating system were not scalable, then the computer may be unusable – much like a person getting a blue screen of death on his laptop every time he tries to use it. The second concern was that even if the algorithms were scalable, the “noise” in the operating system would make the applications run slowly. Operating system noise is a term used to describe frequent interruptions that an operating system makes to a running application (for example, checking the keyboard several times a second to see if a key has been pressed, checking for incoming messages, and so forth). Such disruptions on one processor can cause a science application to slow down on several processors as they wait for the delayed task to catch up. While waiting, operating system disruptions occur on other processors and the delays build up across the entire system. Both of these concerns were realized.



**Figure 1.** ORNL's Petascale Roadmap was driven by the growing number of cores on a processor chip.

Argonne quantified these issues as part of the ZepetoOS project and was successful at getting Linux on the IBM Blue Gene/P system at the Argonne Leadership Computing Facility. This new capability has enabled new classes of applications to run on the Blue Gene, including many-task applications such as molecular docking and MRI analysis. Similarly, ORNL and Cray collaborated to deploy Linux at scale on Cray systems by replacing internal algorithms and turning off all unnecessary interruptions. After a year of research and development, the Linux operating system was able to scale to the full size of Jaguar, and the noise was low enough to allow science applications to run at more than a sustained petaflop/s. Scalability will be an even bigger technical challenge to address on the Exascale Roadmap, as discussed later.

Every large computer system established at the beginning of the Petascale Roadmap had problems with file system scaling. A simplistic example illustrates the problem. Imagine a 100,000 processor petascale computer running a science application. If each processor tries to read an input file, then the file system can be overwhelmed by 100,000 requests to read the same file. Similarly, the file system can be overwhelmed if all the processors try to write out their results at the same time. Parallel file systems are much more complicated than this simple example. They have thousands of disk drives, and a single file will be spread across many different disks in order to have the required volume and read/write performance. A major challenge on the Petascale Roadmap was to make sure that a petascale file system would have the scalability and performance needed for petascale science. The ORNL Leadership Computing Facility uses the Lustre parallel file system. To address the file system challenge, ORNL created a Lustre Center of Excellence at the start of the Petas-

cale Roadmap. This Center, staffed by Lustre engineers, worked with ORNL staff to locate and fix scaling and performance problems in Lustre. By the time Jaguar was installed, the ORNL Leadership Computing Facility had in place the largest and fastest Lustre file system in the world, able to handle the needs of petascale science. Similarly, the Argonne Leadership Computing Facility has deployed the PVFS, a parallel file system, on their Blue Gene/P and has proven the scalability of this approach. Recent work by the SciDAC Scientific Data Management Center on MPI-IO, Parallel netCDF, and the ADIOS I/O library has dramatically improved application I/O performance on the Cray and Blue Gene systems. Although satisfying the petascale science needs, these advances are still a long way from being able to handle the file system and I/O challenges at the exascale.

The message passing paradigm is at the heart of most petascale applications, and an efficient implementation of the Message Passing Interface (MPI) standard is a critical component of any Petascale system. For example, if one processor needs to send its result to everyone else a naïve implementation might have that processor make 100,000 sends – one to each of the other processors. But an efficient implementation would send the result to one processor, then the two processors that know the result send it to two others, then the four that have the result send to four others and so on. Using this logarithmic algorithm all 100,000 processors can get the result in the time it takes the naïve approach to send the first 17 messages! ASCR computer scientists are leaders in the development of the popular MPICH and Open MPI implementations. One of the petascale challenges they solved was how to efficiently have not just one, but every processor wanting to send its result to every other processor at the same time. In MPI,

Scalability will be an even bigger technical challenge to address on the Exascale Roadmap.

The Exascale Roadmap takes us from the petascale science of today to tomorrow's exascale science, where the nation can tackle some of its most important problems in energy, climate change, health, and security.

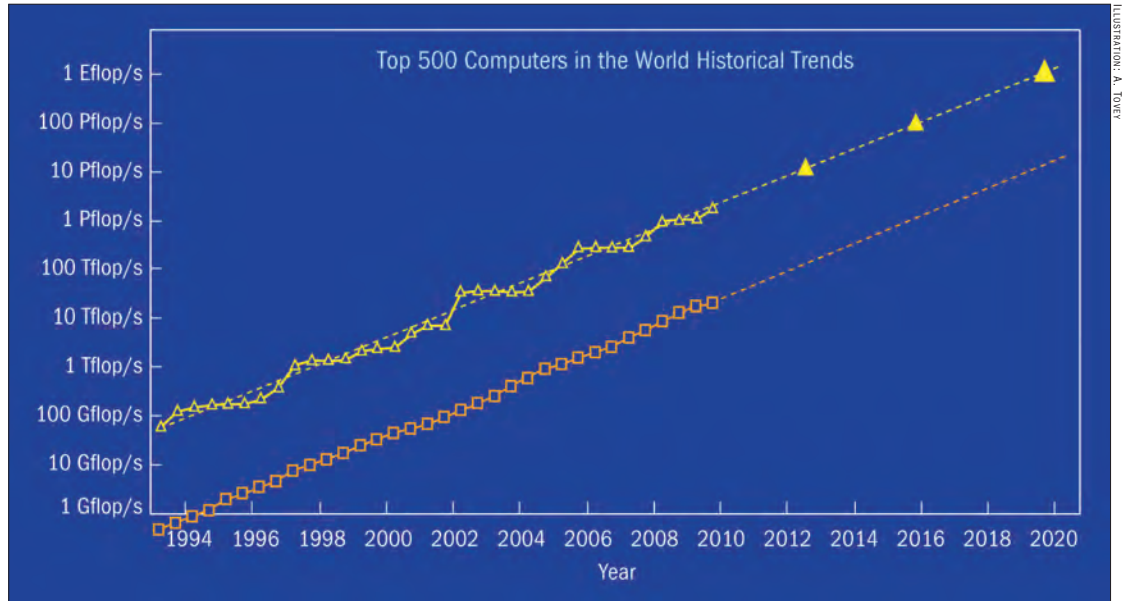


Figure 2. Projecting the TOP500 performance curves out to determine when exascale will be reached.

this is called ALLTOALL and is a difficult algorithm to scale to huge numbers of processors. A second petascale challenge solved in the MPICH and Open MPI implementation was how to quickly get communication established when the application first starts up on 100,000 processors. Initially, MPI implementations did not scale and took several minutes to start up; now, startup on petascale systems takes around 3 seconds. The MPICH and Open MPI implementations form the basis of vendor implementations at both Leadership Computing Facilities, and ongoing collaborations work to ensure that the algorithms in these implementations are best matched to the architectures of these and future systems.

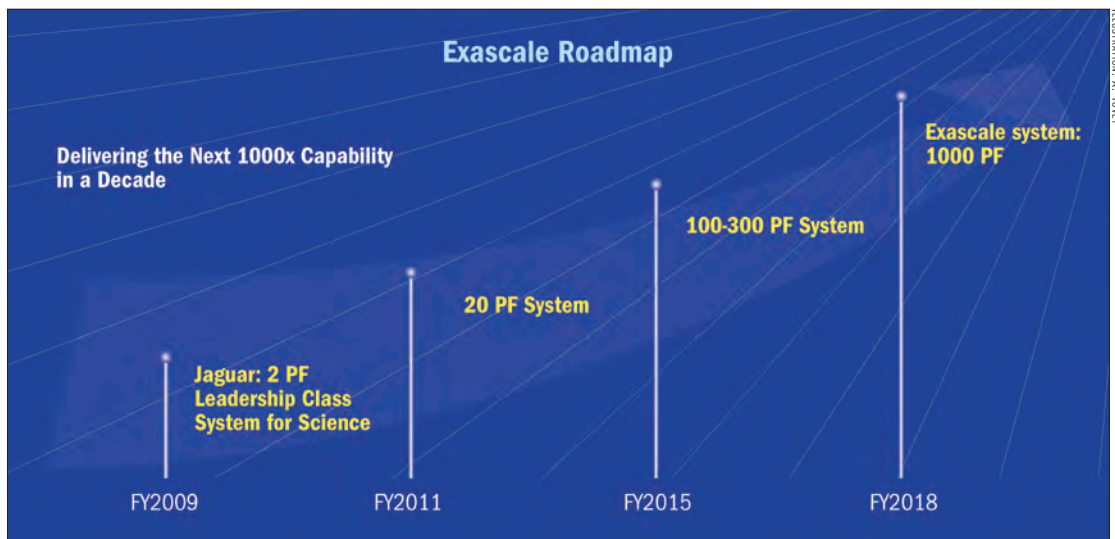
The key to petascale science, as well as exascale science, is making sure that the science application codes can effectively use the computational resource. These science codes are huge and complex, often involving millions of lines of code and coupling multiple complex phenomena happening at different time scales. For example, modeling the combustion in a high efficiency engine involves fluid flow, air mixing, combustion chemistry, heat transfer, and moving interfaces – all having to be solved at the same time. These science application codes take decades to create and build up confidence in their accuracy. It is a major undertaking to modify these applications to utilize the 100,000 processors in today's leadership computers. Two lessons were learned from the petascale success. First, have intermediate systems for application teams to use as stepping stones for adapting their application codes to the scale and to the new technologies being developed to reach exascale. Second, place liaisons from the Leadership Computing Facilities on the application teams. These liaisons have expertise in the new

computer technologies and the particular science area and can help guide the evolution of the applications to solve the national challenges made possible by leadership computers. Further support is provided by the computer scientists and mathematicians at the SciDAC centers and institutes, who bring specific expertise such as I/O performance, numerical techniques, and more.

While ASCR computer scientists and mathematicians are playing active roles in the development of scalable system software for upcoming machines, to achieve exascale science it will be necessary to go even further and co-design the applications and architecture together. Some of the early examples of such co-design are IBM's double precision floating point units in the Cell architecture and Nvidia's memory protection in their latest accelerators. Both these architecture changes were made specifically for the requirements of science applications and would not have happened otherwise. DOE also has an ongoing collaboration with IBM to develop the successor to the Blue Gene/P system, an example of co-design applied at the system level, with the needs of DOE science applications guiding the design of all aspects of the system.

**Exascale Roadmap**

The Exascale Roadmap takes us from the petascale science of today to tomorrow's exascale science, where the nation can tackle some of its most important problems in energy, climate change, health, and security. Using the TOP500 trends (figure 2), this 1,000x increase in computational capability is projected to take a decade. Figure 2 shows the number one system and the 500<sup>th</sup> most powerful computer in the world as measured by the performance on the



**Figure 3.** The Exascale Roadmap calls for a series of systems of increasing computational power to enable applications and system software to be scaled to the exascale.

Linpack benchmark. The trends have been remarkably consistent for the past 15 years. Projecting these trends out, it is interesting to note that by the time an exascale system exists, there will be more than five hundred 10 PF systems in the world. The time frame for the first 10 PF, 100 PF, and 1 exaflop/s systems are highlighted with yellow triangles. This information was used to construct the multiple stages in an Exascale Roadmap, shown in figures 3 and 4 (p56).

Learning from the success of the Petascale Roadmap, the Exascale Roadmap has multiple stages to allow the applications and system software to adapt to the ever-increasing scale. Beginning today with petascale systems such as Jaguar, the roadmap shows that systems with a few tens of petaflop/s will be available in the 2011–2012 timeframe. The first example may be the National Science Foundation’s Blue Waters system, which is expected to be delivered to the National Center for Scientific Applications in 2011. By 2015, systems with a few hundred petaflop/s are expected to exist. These systems will provide scientists with the capabilities needed to make scientific discoveries in mid-decade and, just as importantly, they will also test new technologies needed to reach exascale. In the 2018–2020 timeframe, the first exascale system is expected to appear. The images used in figure 3 are placeholders to illustrate that the systems could be quite different at each stage of the roadmap.

Multi-core chips were the technology driver that enabled petascale science. The question arises: what will be the technology driver that will allow the Exascale Roadmap to be successfully executed? All the paper studies over the past couple of years have shown that just continuing to ride the multi-core trend will not be sufficient to reach exascale. Instead, the driving technology to build an exascale system

appears to be the emerging heterogeneous, many-core processors. While multi-core processors have a few to tens of cores, many-core processors will have hundreds of cores on each chip. The heterogeneity arises from analysis of the most effective way to use all these cores. Rather than making them the same, it is much more efficient to establish a few general purpose cores and designate the rest as specialized cores. One example of such heterogeneity is the Advanced Micro Devices (AMD) Fusion project. A year ago, the AMD team announced plans to put a graphics processor unit (GPU) on the same chip with their standard multi-core processor. GPUs can have hundreds of graphics specialty cores. Thus, when Fusion appears in the next few years, it will be a heterogeneous mix of many cores. Similar heterogeneous many-core chips have been announced by Intel and other chip manufacturers. Heterogeneous, many-core processors have the potential to overcome a major exascale challenge—power consumption—by providing two orders of magnitude more flops per watt than multi-core processors. But this driver increases the programming challenge.

As with the petascale roadmap, having the drivers and the ability to build an exascale system will not guarantee exascale science. The Leadership Computing Facilities must also provide the expertise and tools to enable science teams to productively utilize the exascale systems. To help prepare the science community for heterogeneous, many-core systems, the Hybrid Multicore Consortium was formed. This consortium, announced at Supercomputing09, is a multi-organizational partnership to support the effective development (productivity) and execution (performance) of high-end scientific codes on large-scale, accelerator based systems. The Hybrid Multicore Consortium is a part of the Exascale Roadmap

Learning from the success of the Petascale Roadmap, the Exascale Roadmap has multiple stages to allow the applications and system software to adapt to the ever-increasing scale.

ILLUSTRATION: A. TOWER

Systems	2009	2011	2015	2018
System Peak Flops/s	2 Peta	20 Peta	100-200 Peta	1 Exa
System Memory	0.3 PB	1 PB	5 PB	10 PB
Node Performance	125 GF	200 GF	400 GF	1-10 TF
Node Memory BW	25 GB/s	40 GB/s	100 GB/s	200-400 GB/s
Node Concurrency	12	32	0(100)	0(1000)
Interconnect BW	1.5 GB/s	10 GB/s	25 GB/s	50 GB/s
System Size (Nodes)	18,700	100,000	500,000	0(Million)
Total Concurrency	225,000	3 Million	50 Million	0(Billion)
Storage	15 PB	30 PB	150 PB	300 PB
I/O	0.2 TB/s	2 TB/s	10 TB/s	20 TB/s
MTTI	Days	Days	Days	0(1Day)
Power	6 MW	~10 MW	~10 MW	~20 MW

**Figure 4.** Estimated specifications for Exascale Roadmap systems based on expected technologies in the respective time frames.

in a similar way that the Lustre Center of Excellence was a part of the Petascale Roadmap. The programming environment challenge involves the programming models and tools needed to scale and tune science applications on exascale systems.

In addition to the programming environment challenge, the other major technical challenges for the Exascale Roadmap are exponentially increasing scale; resilience in the face of constant faults; fast, efficient data movement both within a heterogeneous node and between nodes; and a power consumption in flops per watt that is a thousand times better than today’s computers. Significant research and development efforts are required in each of these areas over the next decade in order to achieve productive exascale science.

**Scalability Challenge**

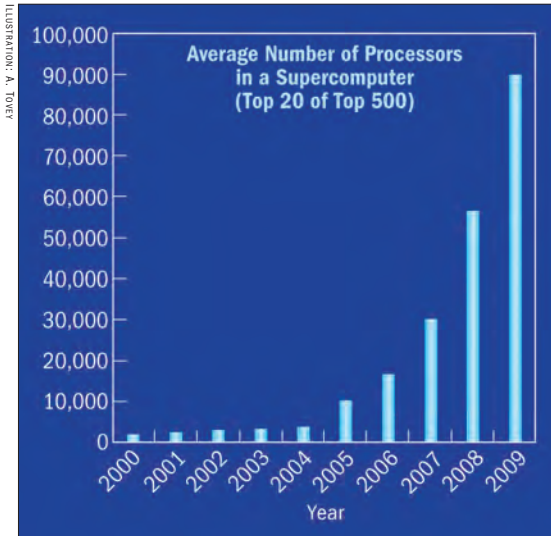
As shown in figure 5, the largest systems are growing exponentially in the number of processors. Projecting this growth into the future, an exascale system in 2018 will have more than 100 million processors. This rapid increase in scaling is the biggest challenge because it makes existing challenges, such as programming, system management, and file system performance, harder. It also creates new challenges in areas such as resilience and power consumption. What is driving the increase in processors and will it continue to 2018?

In simple terms, there are two ways for a processor to do more operations per second. One is to increase the frequency of the processor allowing it to compute faster. The second way is to replicate the

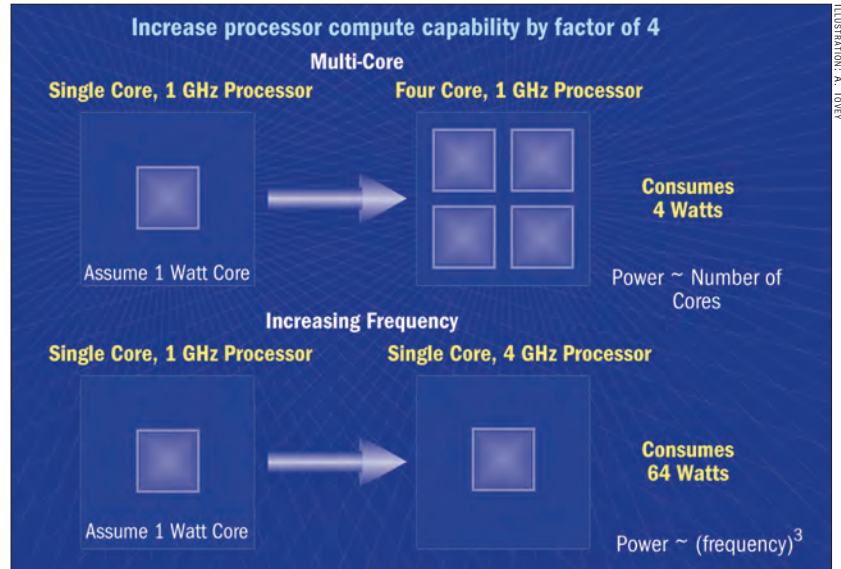
computing units on the chip. Through the 1980s and 1990s, the processors got faster by increasing the frequency. But chip manufacturers realized that if they continued to increase the frequency, the chips would become too hot and use too much electrical power. The power required by a processor is proportional to its frequency cubed. Thus, doubling the speed of a processor by doubling its frequency increases the power it draws by a factor of eight, which would rapidly rundown phone and laptop batteries. By comparison, doubling the speed of a processor by replicating two compute cores on the chip only increases the power by a factor of two. Because multiple cores allow more computing for less power (figure 6) multi- and many-core processors will continue to be a driving force for an exascale system where power consumption is a big challenge.

Petascale applications are using a quarter-million processors today. Because the number of cores is going to continue to increase, exascale applications will need to find a way to exploit an additional 1,000× in concurrency. Looking at the roadmap (figure 4), there is only a 30× increase in system memory size, so applications will not be able to simply solve 1,000× larger problems. Instead, scientists will need to find ways to break the problem into an order of magnitude more pieces that can be solved concurrently. This approach often involves having to rethink the entire solution approach, which is why the intermediate systems in the Exascale Roadmap are critical for the applications software to evolve and adapt to the changing bytes/flop ratios of the future leadership systems.

Significant research and development efforts are required in each of these areas over the next decade in order to achieve productive exascale science.



**Figure 5.** Exponential growth in the number of processors in the largest computing systems has been observed over the past decade.



**Figure 6.** Lower power and cooling requirements are the driving forces behind future many-core processors.

### Data Movement Challenge

Data movement has always been the bottleneck for large-scale systems (sidebar “Data Movement, Not Flops, is the Bottleneck to Performance” p58). Sometimes referred to the memory wall, the bottleneck is expected to grow with scale (which is itself growing exponentially). Increasing scale increases the amount of data movement because of the likelihood that the required data are on or being used by another processor. And as the speed of floating point operations increases, the relative time to fetch memory grows longer. Studies have shown that today’s computers spend most of their time moving data rather than performing mathematical operations. The data storage hierarchy on the node of today’s supercomputers is 5 levels deep: registers, L1 cache, L2 cache, L3 cache, and main memory. Just determining where data are in the hierarchy can take many operations. But it is worth determining because it is several orders of magnitude faster to fetch data from cache versus main memory. Node architecture designs for future systems will be further complicated by being heterogeneous many-core with mixtures of shared (coherent) memory and non-shared memory, and having global memory addressing.

By 2018, the data movement challenge is further complicated because data movement and storage will consume more than 70% of the total system power. In figure 4, most of the 20 MW will go just to power the 10 PB of total system memory.

The data movement challenge in the Exascale Roadmap is divided into three categories, each with complementary research activities:

- **On node** – node architecture design, new memory management schemes, and improved memory capacity and speed through 3D stacking
- **Between nodes** – interconnect design, optical communication, performance, scalable latency, bandwidth, and resilience
- **File System I/O** – scalability, performance, and metadata

### Resilience Challenge

Users want resilience in the execution of their applications. They want to be able to submit a long-running job and have it run to completion in a timely manner. Exascale systems will have millions of processors in them, and some projections say they will have a billion threads of execution. The major challenge in resilience is that faults in extreme scale systems will be continuous rather than an exceptional event. This requires a major shift from today’s software infrastructure. Every layer of the exascale software ecosystem has to be able to cope with frequent faults; otherwise, applications will not be able to run to completion. The system software must be designed to detect and adapt to frequent failure of hardware and software components. With the potential that exascale systems will have constant failures somewhere across the system, the application software will not be able to rely on check-pointing to cope with faults. For exascale systems, new fault tolerance paradigms will need to be developed and integrated into both existing and new applications.

There are several factors driving up the fault rate on the Exascale Roadmap.

Every layer of the exascale software ecosystem has to be able to cope with frequent faults; otherwise, applications will not be able to run to completion.

## Data Movement, Not Flops, is the Bottleneck to Performance

Data movement, not the amount of floating point operations (flops), will be the performance bottleneck in future supercomputers. Consider a simple computation:  $A = B + C$ . First the processor must go out to memory and get the value of  $B$ , and then it has to go out and get the value of  $C$ . It rapidly adds these two values together and sends the result  $A$  back out to memory. This simple example illustrates how data movements can dominate performance. A delay in any of the three data movements will make the one floating point operation (no matter how fast it is) appear to take a long time.

Making data movement fast and efficient inside a supercomputer is a major challenge. Delays can occur if  $B$  and  $C$  are in local memory (tens of

cycles), in another processor's memory (hundreds of cycles), or stored on a hard drive off the machine (tens of thousands of cycles). Now complicate matters by having thousands of processors making data movement requests at every moment. The hundreds of miles of wire inside a supercomputer and the data packets traveling along these wires can be imagined as a giant highway system. And just like a highway system, it can have traffic jams (data congestion), delays (data arrive late), and even crashes (if some component becomes overwhelmed by incoming traffic). The challenge is ensuring the data arrives on time just like you try to get to work on time.

- The number of components for both memory and processors will increase by a factor of 100, which will increase hard and soft errors (sidebar "Types of Errors") by that same amount due to the commercial error rates in these components.
- Smaller circuit sizes, running at lower voltages to reduce power consumption, increase the probability of switches flipping spontaneously due to thermal and voltage variations, increasing soft errors.
- Chip manufactures realize the reliability for extreme scale systems will require additional detection and recovery logic right on the chips to detect silent errors. They estimate these extra circuits will increase power consumption by 15% and increase the chip costs, so there is resistance to making such changes.
- The thermal and mechanical stresses, caused by power management cycling of chips on and off just when needed, significantly decrease.
- Heterogeneous systems make error detection and recovery even harder. For example, detecting and recovering from an error in a GPU can involve hundreds of threads simultaneously on the GPU and hundreds of cycles in drain pipelines to begin recovery.

Flash (or phase change) memory can be exploited to extend the life of existing checkpointing techniques through 2015 of the Exascale Roadmap by providing a relatively fast, non-volatile memory on neighboring nodes to save checkpoints. In the longer term, increasing system resiliency without excess power or performance costs will require a multi-pronged approach that includes innovations at each level of the system hierarchy and a vertically integrated effort that enables new resiliency strategies across the levels.

It is clear the solution to this challenge requires a coordinated hardware, system software, and application software effort. Research in the reliability and robustness of exascale systems for running large simulations is critical to the effective use of these systems. New paradigms must be developed for handling faults within both the system software and user applications. Equally important are new approaches for integrating detection algorithms in both the hardware and software and new techniques to help simulations adapt to faults.

### Power Consumption Challenge

Projections in the Defense Advanced Research Projects Agency's extreme scale study show that the power consumption for an exascale system in 2018 even under optimistic assumptions would be 100–200 MW. This amount of power is equal to the amount produced by a small power plant and is not tenable from the infrastructure standpoint of trying to safely bring that much power into a building. Even if it were feasible from an infrastructure perspective, the electric bill for that much power would be more than \$100 million per year.

As shown in figure 4 (p56), the goal of the Exascale Roadmap is to build a productive exascale system with a power consumption of 20 MW. This goal will require a 300× improvement in flops per watt over the technology available today. Exacerbating the power reduction challenge is the fact that the solutions to the critical challenges of resilience and data movement require an increase in the power consumption rather than a reduction. As such, these other challenges must be addressed first; otherwise, any power improvements could just be negated by the solution approaches for data movement and resilience.

Developments in two technologies – 3D stacked memory and heterogeneous, many-core processors – offer the potential to achieve the specified power-consumption goal. The emerging solution

Research in the reliability and robustness of exascale systems for running large simulations is critical to the effective use of these systems.

is to increase the flops per watt by 300× by having many of the cores specialized in accelerating mathematical operations and a few focused on moving data into and out of the accelerator cores. In a recent *HPCwire* article, Nvidia proposed such an exascale system. Increasing the flops per watt is only part of the problem. As discussed in the data movement challenge, more than 70% of the exascale system power consumption will be memory and data movement. In fact, one reason the amount of system memory in the exascale system is relatively small is to make the power goal feasible. During the Exascale Roadmap, the development and commercialization of 3D stacked memory is a possible solution to getting the bytes/watt and data movement power reduced to the necessary levels. 3D stacked memory involves stacking memory wafers on top of each other and running the circuit and power lines vertically up through the layers. There is also discussion about stacking the memory directly on top of the processor to reduce the distance and power required to move data from memory to the processor.

In addition to the above hardware power efficiency improvements, there also need to be software power efficiency improvements. Current power management software, such as found in laptops, only control a few things such as processor speed, screen brightness, and disk drive. Exascale power management technologies will need to encompass all system components, including processor cores, memories, storage, I/O circuitry, power supplies, and service processors. These technologies include slowing down components, such as dynamic voltage and frequency scaling of processor and memory chips, and speed control of data communication on the interconnect. Software power management will be performed voluntarily under application control as well as involuntarily under operating system or even hardware control. For example, power management can be instigated by programmer control by providing hints to the operating system about the parts of the program where processor speed can be reduced (for example, while waiting for a message). However, the programmer may not have ultimate control. For example, the hardware may decide to slow down the processor to reduce thermal stresses and hot spots under extreme conditions of workload intensity.

Research in power-aware architectures and management policies will allow a system with a fixed power limit to achieve a larger fraction of its peak performance by adapting its power budget to the needs of the application.

### Hardware–Software Co-Design

Historically, huge supercomputers were built and delivered with little or no software on them. The application developers were left with heroic efforts

## Types of Errors

**hard errors** permanent component failure, such as a hardware crash or software failure

**soft errors** transient problems, such as a blip or short-term failure of a hardware device or software program

**silent errors** undetected errors, either hard or soft, due to lack of detectors for a component or inability to detect (such as when a transient effect is too short), and where the real danger is that an answer may be incorrect but the user would not know

to get their simulations to run efficiently on these systems. There is a large gap between the peak theoretical performance of supercomputers and the actual performance realized by today's applications. The Exascale Roadmap will encourage hardware/software co-design in order to achieve the maximum benefit to the science applications. Each of the roadmap challenges presented here has both a hardware and software component in the solution.

The Exascale Roadmap systemically includes the concept of co-design in all layers of the software stack from the applications, down through the programming environment, middleware, system software, to the processor core instructions and all layers of the hardware architecture from the many-core processor design, through the heterogeneous node architecture, the inter-node communication fabric, to the overall system architecture.

Ultimately, the needs of the science applications drive the challenge solutions and potential shortcuts. For example, by characterizing the computational intensity, locality, memory and communications patterns, and resiliency of existing petascale applications and then projecting this information into power and failure-rate models for different exascale architectures, a better understanding emerges of where the energy budget is being expended, what the performance bottlenecks are, and where irrecoverable failures will likely happen given different system design points.

National laboratory–industry–university partnerships will be extremely important to meeting the aggressive goals of the Exascale Roadmap. Fundamentally, this approach represents a shift from simply procuring and operating large scale systems. The organizations involved in the Exascale Roadmap, vendors and researchers alike, will actively engage in the development and co-design of advanced architectures and algorithms that address the nation's critical problems in energy, climate, security, and health. ●

National laboratory–industry–university partnerships will be extremely important to meeting the aggressive goals of the Exascale Roadmap.

**Contributor** Al Geist, ORNL