

THE AGE OF EXASCALE

Everything Must Change

The advances required to move from one generation of high-performance computer system to the next can often be described as ‘more of the same, but bigger and faster.’ The move to exascale systems is different, however – everything has to change.

ICE | INFRASTRUCTURE COMPUTING FOR THE ENTERPRISE

4 FINDINGS

- The high power consumption of current high-end systems cannot extend to exascale systems. **PAGE 7**
- Exascale systems will have millions of processor cores. **PAGE 4**
- The use of heterogeneous processors will help exascale systems achieve high performance while keeping power consumption low. **PAGE 14**
- Programming systems with millions of processors will need new languages, new tools and new skills. **PAGE 27**

5 IMPLICATIONS

- Tomorrow’s high-end systems will use different memory technologies to deliver low power, including 3D stacking. **PAGE 16**
- Exascale systems will use multicore and many-core versions of x86 processors, augmented by heterogeneous accelerators. **PAGE 12**
- Resilience at large scale is fundamentally different. Instead of trying to deliver resilience, you need to learn to live with the lack of it. **PAGE 10**
- As components change, the balance between their relative capacity and performance will change – and applications must be re-architected to cope. **PAGE 7**
- There is a serious skills shortage. Even as systems are becoming more complex, the level of training for computer scientists is less sophisticated. **PAGE 11**

1 BOTTOM LINE

- The huge changes required to build exascale systems are such that all of the HPC community must work together to design affordable, usable next-generation systems that deliver fantastic levels of peak performance. There will be massive disruption for vendors and users – all must tread carefully.

SEPTEMBER 2011

ABOUT THE 451 GROUP

The 451 Group is a technology analyst company. We publish market analysis focused on innovation in enterprise IT, and support our clients through a range of syndicated research and advisory services. Clients of the company — at vendor, investor, service-provider and end-user organizations — rely on 451 insights to do business better.

ABOUT TIER1 RESEARCH

Tier1 Research covers consumer, enterprise and carrier IT services, particularly hosting, colocation, content delivery, Internet services, software-as-a-service and enterprise services. Tier1's focus is on the movement of services to the Internet — what they are, how they are delivered and where they are going.

© 2011 The 451 Group, Tier1 Research and/or its Affiliates. All Rights Reserved. Reproduction and distribution of this publication, in whole or in part, in any form without prior written permission is forbidden. The terms of use regarding distribution, both internally and externally, shall be governed by the terms laid out in your Service Agreement with The 451 Group, Tier1 Research and/or its Affiliates. The information contained herein has been obtained from sources believed to be reliable. The 451 Group and Tier1 Research disclaim all warranties as to the accuracy, completeness or adequacy of such information. Although The 451 Group and Tier1 Research may discuss legal issues related to the information technology business, The 451 Group and Tier1 Research do not provide legal advice or services and their research should not be construed or used as such. The 451 Group and Tier1 Research shall have no liability for errors, omissions or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve its intended results. The opinions expressed herein are subject to change without notice.



New York

20 West 37th Street, 6th Floor
New York, NY 10018
Phone: 212.505.3030
Fax: 212.505.2630

San Francisco

140 Geary Street, 9th Floor
San Francisco, CA 94108
Phone: 415.989.1555
Fax: 415.989.1558

London

37-41 Gower Street
London, UK WC1E 6HH
Phone: +44 (0)20.7299.7765
Fax: +44 (0)20.7299.7799

Boston

125 Broad Street, 4th Floor
Boston, MA 02109
Phone: 617.275.8818
Fax: 617.261.0688

TABLE OF CONTENTS

SECTION 1: EXECUTIVE SUMMARY	1
1.1 KEY FINDINGS	2
1.2 METHODOLOGY	3
SECTION 2: INTRODUCTION	4
2.1 WHAT'S IN A FLOP/S?	4
<i>FIGURE 1: Floating-Point Operations</i>	4
2.2 DISRUPTIVE CHANGE	5
2.3 WHAT'S THE POINT OF EXASCALE?	5
SECTION 3: EXASCALE ISSUES	7
<i>FIGURE 2: Exascale Requirements</i>	7
3.1 POWER CONSUMPTION	7
3.2 PARALLELISM AND SCALABILITY	9
3.3 RESILIENCY	10
3.4 LACK OF SKILLS	11
SECTION 4: PROCESSORS AND ACCELERATORS	12
<i>FIGURE 3: Convergence of Processor Technologies</i>	12
<i>FIGURE 4: Divergence of Processor Technologies</i>	13
4.1 HETEROGENEOUS ARCHITECTURES	14
4.2 CRAY	14
4.3 IBM	15
4.4 INTEL	15
SECTION 5: MEMORY TECHNOLOGY	16
<i>FIGURE 5: DDR Memory</i>	16
<i>FIGURE 6: Memory Hierarchy</i>	17

SECTION 6: NETWORK TECHNOLOGY	18
<i>FIGURE 7: InfiniBand Roadmap.</i>19
<i>FIGURE 8: InfiniBand Performance.</i>19
SECTION 7: DATA MANAGEMENT	20
7.1 HPC STORAGE SYSTEMS	20
7.2 PARALLEL FILE SYSTEMS.	21
7.3 THE EXASCALE CHALLENGES	22
SECTION 8: SOFTWARE ISSUES	24
8.1 PROGRAMMING METHODOLOGY	25
8.2 SOFTWARE DEVELOPMENT TOOLS	26
8.3 NEW APPROACHES TO PROGRAMMING TOOLS	27
8.4 ALGORITHMS FOR EXASCALE	27
SECTION 9: EXASCALE INITIATIVES	29
9.1 EUROPEAN EXASCALE INITIATIVES.	29
9.1.1 <i>European Technology Platforms</i>29
9.1.2 <i>PlanetHPC</i>30
9.1.3 <i>e-Infrastructure Reflection Group</i>30
9.1.4 <i>Partnership for Advanced Computing in Europe</i>31
<i>FIGURE 9: PRACE Systems</i>32
9.1.5 <i>European Exascale Software Initiative.</i>32
9.2 US EXASCALE INITIATIVES	32
9.2.1 <i>DARPA</i>32
9.2.2 <i>US Department of Energy</i>33
9.2.3 <i>The International Exascale Software Project</i>33
9.3 OTHER INTERNATIONAL INITIATIVES	35

SECTION 10: THE 451 TAKE	36
10.1 IMPLICATIONS FOR USERS, STARTUPS, INCUMBENT VENDORS AND INVESTORS	37
10.1.1 <i>Implications for Users.</i>37
10.1.2 <i>Implications for Startups.</i>37
10.1.3 <i>Implications for Incumbent Vendors</i>37
10.1.4 <i>Implications for Investors</i>38
 INDEX OF COMPANIES AND ORGANIZATIONS	 39

SECTION 1

Executive Summary

There is a great deal of excitement and debate within the high-performance computing (HPC) community about the technical challenges that need to be resolved in order to deliver one-exaflop-per-second performance. However, the most important issue is not the technology from which the computer system is built, but rather the applications that are run on such a powerful system. Exascale systems will be able to run simulations of scientific theories or industrial product designs at a far higher resolution, while various components that could previously only be simulated separately will be integrated in a single, more realistic simulation. Exascale systems may be able to solve many problems that are beyond the reach of today's systems, such as understanding climate change, nuclear fission and how the human brain works.

A side effect of building an exascale system will likely be the availability of more affordable, efficient petascale systems. A petascale system today is housed in hundreds of cabinets and consumes several MWs of power. A petascale system based on exascale technology would likely fit in a single cabinet and consume less than 100kW.

If power consumption in electronic components was not related to clock frequency, this report would not have been written. Power consumption is a function of the voltage squared times the clock frequency. Otherwise, the industry would just wind up the clock speed of processors, and little else would need to change. But because power consumption (and therefore clock speed) is constrained, future large-scale systems must be built from a high number of relatively low-power-consuming multicore processors, with the added option of computational accelerators such as GPUs.

Which brings us to the next problem: How do we connect all of these components to each other, and to memory and storage subsystems? In large-scale parallel systems, the latency of any operation (e.g., compute, memory or storage access) can inhibit the scalability of a program. Today, many HPC applications only scale to a few tens or hundreds of threads. To reach exascale, we are talking about millions or even billions of threads, so extreme scalability – and therefore very low latency – is crucial.

The relationship between processors and memory will change, as will the way memory is manufactured. The way systems are built to house these many-core processors, heterogeneous accelerators and new styles of memory must also change. And as all of the components change, so do the relationships between them. The amount of cache and memory available per processor core will be different, as will the time taken to access these components. So the system balance will shift, and the algorithms used must reflect this shift. These algorithms are translated into programs, which are built using paradigms such as shared memory and message passing, possibly with accelerator components. The development tools we use today were not designed for extreme scalability, and so they, too, must change. As the subtitle of this report says – everything must change.

To build applications that can run efficiently on an exascale system requires both advanced tools and a level of skill that is not prevalent among programmers today. Systems built on such a large scale will suffer regular component failures. The approach to resilience in both hardware and software must adapt to cope with regular failures without taking the system down or stopping an application run. Designing and building all of the hardware and software components to make an exascale system a reality will take an unreasonably high level of investment – one that is beyond any single company and that has to be managed at a national or even international scale. However, while the exascale funding initiatives in place today drip-feed significant funding to a wide variety of research projects, they lack the coordination to make an affordable, programmable exascale system a reality by 2018.

1.1 KEY FINDINGS

- Exascale systems are required in order to meet many of the technological and societal challenges of the 21st Century.
- Power consumption is the driver for change in system design. If we did nothing different, an exascale system would consume more than one gigawatt of power.
- The level of parallelism in an exascale system will be on the order of millions of threads. Both systems and applications must be built very differently in order to scale to this level of parallelism.
- The normal evolution of mainstream processors will not drive exascale systems alone; these processors will be assisted by high-performance, low-power options such as GPUs or FPGAs.
- In order to support faster processing, memory technology and architectures must also keep pace.
- Standard file systems and rotating media have neither the scalability nor the performance to support exascale computation, so they must be replaced or evolve significantly.
- As all of the components change, the system balance will be very different. Applications must cope with different relative capacities and performance of processors, memory and storage systems.
- The scale of these systems will be a challenge for all software components, but particularly for applications. The programming paradigms and languages we use today will no longer work.
- Exascale systems will always include failed components, so resilience must be baked in at the system and application level.
- Programming exascale systems will be very complex, but the typical programmer today is trained in IT, not computer science, and parallel programming is not generally taught at the 'first degree' level.

1.2 METHODOLOGY

This report on exascale technologies and issues in high-performance computing is based on a series of in-depth interviews with a variety of stakeholders in the industry, including IT managers at end-user organizations across multiple sectors, technology vendors, research institutions and VCs. This research was supplemented by additional primary research, including attendance at a number of trade shows and industry events.

Reports such as this one represent a holistic perspective on key emerging markets in the enterprise IT space. These markets evolve quickly, though, so The 451 Group offers additional services that provide critical marketplace updates. These updated reports and perspectives are presented on a daily basis via the company's core intelligence service – the 451 Market Insight Service. Perspectives on strategic acquisitions and the liquidity environment for technology companies are updated regularly via the company's forward-looking M&A analysis service – 451 TechDealmaker – which is backed by the industry-leading 451 M&A KnowledgeBase.

Emerging technologies and markets are also covered in additional 451 practices, including our Enterprise Security, Eco-Efficient IT, Information Management, Commercial Adoption of Open Source (CAOS), Infrastructure Computing for the Enterprise (ICE), Datacenter Technologies (DCT) and 451 Market Monitor services, as well as CloudScape, an interdisciplinary program from The 451 Group and subsidiary Tier1 Research. All of these 451 services, which are accessible via the Web, provide critical and timely analysis specifically focused on the business of enterprise IT innovation.

This report was written by John Barr, Research Director for HPC, who spent 25 years working at the leading edge of HPC, participating directly in the evolution of parallel compilers and cluster and grid computing. He regularly supports the European Commission's HPC activities as an invited expert, and was the Chairman of the British Computer Society's Parallel Processing Specialist Group before emigrating to Austria.

Any questions about the methodology should be addressed to John Barr at:
john.barr@the451group.com

For more information about The 451 Group, please go to the company's website:
www.the451group.com

SECTION 2

Introduction

The performance advances in HPC seen over the past three decades can be attributed to a combination of higher clock speeds, more sophisticated processor architectures and increases in the exploitation of parallelism, which has been impacted by larger clusters, more sockets per system and, more recently, more cores per processor. Although a bit of a generalization, it is not too wide of the mark to say that many of these performance transitions were made by doing more of the same – just bigger and faster. However, because of the prohibitive power consumption that this approach will lead to, it is no longer an option.

The next performance target is one exaflop per second (exaflop/s), and the HPC industry must rewrite the rule book because several major technology changes are required if this target is to be attained. Why? There are two main drivers here, the first of which is power consumption. It now costs more to power and cool many systems for three years than to buy the systems in the first place, and if the industry continues to build systems the same way, they will be bigger, faster and even more power-hungry. An exascale system built with today's technology would consume more than one gigawatt of power – a situation that is orders of magnitude away from being acceptable. The second driver for a major rethink of system design is parallelism. A handful of the fastest systems today have more than 100,000 processor cores. But exascale systems will have *millions* of cores.

2.1 WHAT'S IN A FLOP/S?

In the past 30 years, the peak performance of HPC systems has advanced from megaflop/s through gigaflop/s and teraflop/s up to petaflop/s. A way to understand the phenomenal performance advances achieved by high-end computing is to appreciate that if the same advances had been made in aircraft performance, you would now be able to fly from London to New York in less than one-ten-thousandth of a second.

FIGURE 1: FLOATING-POINT OPERATIONS

TERM	OPERATIONS PER SECOND	SCIENTIFIC NOTATION
Megaflop/s	One million	10 ⁶
Gigaflop/s	One billion	10 ⁹
Teraflop/s	One trillion	10 ¹²
Petaflop/s	One quadrillion	10 ¹⁵
Exaflop/s	One quintillion	10 ¹⁸

A 'flop' is shorthand for a floating-point operation – that is, an addition, subtraction or multiplication task (division usually takes longer) – and these terms relate to the number of floating-point operations per second. In Europe, the terms above one

million originally used a different scale, with, for example, 'one billion' being a million million. The American names for large numbers (known as the short scale) are now used by most countries, and that is the terminology generally used in computing.

2.2 DISRUPTIVE CHANGE

When disruptive change happens in any industry, those that quickly comprehend it, respond to the change and adapt their behavior accordingly can exploit the new opportunity and profit. Those that try to use yesterday's methods in today's environment will be left behind. This affects both vendors and users. As HPC systems move toward exascale, there will be disruptive change at many levels – including processor, component and system design, as well as programming paradigms and languages.

We are concerned that many experienced HPC staff seem to think that the industry can achieve affordable, programmable exascale systems by doing more of the same but bigger and faster, and by programming these systems the way we program high-end systems today. We are convinced that this is not the case, and we will explain our reasons throughout this report.

2.3 WHAT'S THE POINT OF EXASCALE?

Exascale systems will have a major positive impact on a range of economic, commercial and societal issues. For most of human history, science had two major 'branches' of work – theory and experimentation. But with the advent of high-performance computers in the second half of the 20th century, a third branch was added – that of simulation. It can be extremely difficult, or expensive – and in some cases, it is actually impossible – to perform certain experiments. Good examples of areas where simulation adds significant value, or enables scientific discovery that would be otherwise impossible, include global warming and climate change, cosmology, and nuclear stockpile analysis. Simulations run on exascale systems will provide a more realistic modeling of complex natural processes, and offer better insight into the underlying science of these processes.

The Human Genome Project started in 1989, and had fully mapped the human genome by 2003. Advances in DNA sequencing technology and HPC systems now enable an individual's genome to be sequenced in weeks, at a cost of tens of thousands of dollars. The next major advance in computational ability should reduce the analysis time to just a few hours, with a cost of around \$1,000. This would open new possibilities of predictive, personalized medical care. While an exascale system is not required to sequence the human genome, the technology advances required to deliver an exascale system will also enable the building of efficient, cost-effective systems with perhaps one-tenth or one-hundredth of the compute power of an exascale system.

Increased compute performance enables improved simulation and analysis during the design of complex products such as cars and aircraft. (The Boeing 777 was the first commercial aircraft to be designed entirely with computer-aided engineering tools.) The resolution of models can be greatly increased, uncovering issues that were not apparent at lower resolution. The interaction of structural, aerodynamic, electrical and other systems can be simulated in unison, avoiding problems that may fall between the cracks of separate simulations. Finally, instead of identifying a design that is 'good enough,' many designs can be compared to optimize the final design for functionality and energy efficiency.

Exascale systems can also help reduce the carbon footprint in transportation and design renewable energy sources such as batteries, catalysts and biofuels. They may also aid greatly in enabling a 'first principles' understanding of the properties of fission and fusion reactions, reverse-engineering the human brain, and tackling the major threats to civilization today, such as the healthcare needs of an aging population, climate change, natural disasters, disease epidemics, financial crises, pollution and terrorism.

One of the significant impacts of building an exascale system is that smaller systems will no longer be exotic and unique. So a side effect of an exascale project will likely be small, affordable, programmable and energy-efficient petascale systems.

SECTION 3

Exascale Issues

Figure 2 illustrates that the rates of change for many system components and attributes will be different for exascale systems than they are for the fastest system available today. There are a number of styles of potential exascale systems, so node-level performance and concurrency, as well as the number of nodes, may vary, but the message remains the same. Everything will change, and this change will not always be constant – so the HPC community must cope not only with disruption, but also with the different relative performance levels of system components.

FIGURE 2: EXASCALE REQUIREMENTS

ISSUE	2011	2018	FACTOR
Peak Performance	8.162 petaflop/s	1 exaflop/s	123
Power Consumption	10MW	20MW	2
System Memory	1.6 petabytes	64 petabytes	40
Node Performance	128 gigaflop/s	10 teraflop/s	78
Node Memory Bandwidth	64 GB/s	4 TB/s	62
Node Concurrency	8	10,000	1,250
Interconnect Bandwidth	20 GB/s	400 GB/s	20
Number of Nodes	68,544	100,000	1 1/2
Total Concurrency	548,352	1 billion	2,000
Total Node Interconnect	20 GB/s	2 TB/s	50
MTTI	2 days	1 day	2

Whenever advances are made in HPC, the balance of next-generation systems changes. For example, computer disk capacity increases faster than disk interfaces accelerate, so, in effect, it actually takes longer today to read a whole disk than it did 10 years ago. As new processor technologies are developed for exascale systems, we will see cache size, performance and sharing characteristics all change, and perhaps the amount of memory per processor core will significantly decrease. As the balance between various components changes, the programs and algorithms that were effective yesterday will no longer work.

3.1 POWER CONSUMPTION

If an exascale system were built with today’s technology – even allowing for extrapolating normal advances up to 2018 – the processors, the network and the memory would blow the available power budget on their own. The largest HPC systems today consume up to 8MW. The target for an exascale system is around 20MW, but extrapolating today’s technology to 2018 with no major changes would mean an exascale system consuming more than one gigawatt.

To manage power consumption, we can no longer ramp up processor clock speeds to deliver more performance, but must instead rely on multiple processor cores per chip. Instead of putting more and more standard x86 cores on a chip, we need to consider low-power chip technologies such as the Intel Atom or ARM processors that are now widely used in mobile devices. Special-purpose devices or accelerators such as FPGAs and GPUs can also add considerable performance within a lower power budget, so future HPC systems will undoubtedly be heterogeneous in this regard.

Memory access and data movement within the system consume far more power than the computations themselves. In order to reduce both latency and the power consumed in passing data around a system, much tighter integration is required. Stacked memory will be built into processor chips, changing the amount of memory per processor and also the latencies with which memory is accessed – some will be faster, some slower.

The shared-memory component of future systems will be CC-NUMA (cache-coherent non-uniform memory access). CC-NUMA systems have tried to bridge the gap between easy-to-program shared-memory systems and easy-to-scale distributed memory clusters. The Convex Computer Exemplar (which provided the technology base for the HP Superdome) and SGI's Origin 2000 were two early examples of CC-NUMA systems. While the theory of extending the shared-memory programming model to larger systems is attractive due to the complexity of programming for different memory latencies, the difficulty in scaling CC-NUMA systems to extreme size, and the high cost of CC-NUMA systems compared with clusters, means this approach has seen limited adoption. Now, because of the need to limit power consumption, the industry must find better ways to address the programming issues, since future systems will use a hierarchy of memory to decrease power consumption.

The TOP500 supercomputer list is published twice a year, at the International Supercomputer Conference (ISC) in Europe in June, and at the US version of the conference in November. It is the 'pop chart' for supercomputers, and lists the 500 fastest systems on the planet, as measured by the Linpack benchmark.

The Linpack benchmark solves a dense system of linear equations. It is something of a blunt instrument, since no one number can accurately describe the performance of a system on a range of applications with different characteristics. However, as the problem size scales with system size, it provides a metric that captures the compute performance of a system, and the benchmark has been widely used to describe the performance of HPC systems since 1993. The Linpack performance is affected by the size of the system, the speed of the processors, the optional use of accelerators such as GPUs and the network performance.

The June 2011 TOP500 list had a surprise new entry in the number-one position. Hosted at the RIKEN Research Institute in Kobe, Japan, the K system (manufactured by Fujitsu) was not only at the top of the list, but it was more than three times faster than the previous top system. The water-cooled K system is not yet fully installed, so we antic-

ipate that its performance will increase by the time the November list is published. Indeed, K (or Kei) means a quadrillion – 10 to the power 16, or 10 petaflop/s – which is its ultimate performance goal. For now, its 672 cabinets (with 548,352 processor cores) peaks at ‘only’ 8.162 petaflop/s. When completed, its power budget for 10 petaflop/s performance will be 9.9MW. An exascale machine will be at least 100 times more powerful than the K system, but it is generally agreed that a power requirement in excess of 20MW is unacceptable, so power consumption must be improved by a factor of 50 if we are to see a useable, affordable exaflop/s system.

The most power-efficient x86 processors today deliver around 400 megaflop/s per watt. Extrapolating that to an exascale system corresponds to 2.5 gigawatts, and that’s just for the processors. The most power-efficient processors today (which may not be the easiest to use) achieve less than 2 gigaflop/s per watt. If the 20MW target is to be achieved for an exascale system, that equates to 50 gigaflop/s per watt, or an improvement between a factor of 25 and a factor of 125, depending on whether x86 or special accelerators are considered.

Meanwhile, the cost of electricity varies enormously by location, but an average figure is between \$500,000 and \$1m per year for each MW of average power draw, or up to \$20m per year if the target of 20MW is achieved. However, some industry luminaries suggest that initial exascale systems might consume up to 200MW, with production systems consuming ‘only’ 100MW by 2022. We don’t believe this would be affordable or acceptable with current concerns over climate change, and we are confident that a system consuming close to 20MW will be possible, although almost certainly not within the 2018 time frame.

3.2 PARALLELISM AND SCALABILITY

Since we can no longer just wind the processor clock up to deliver more speed, future performance gains must predominantly come from a massive increase in the levels of parallelism exploited. High-end systems today have hundreds of thousands of processors. Exascale systems will have millions of processors and – with additional parallelism used to hide data-access latency – may require billions of parallel threads.

It is a significant problem in its own right to build a system with millions of processors and to decompose an application into millions or billions of parallel threads, but unless synchronization times and inter-process latencies are so tiny that they are effectively zero, application performance will not scale beyond even a few thousand processors.

Strong Scaling

If the time to solution for an application decreases as the number of processors applied to a problem increases, it is said to display the property of strong scaling.

Weak Scaling

If an application can only deliver increased performance if the problem size is increased in line with additional resources, it is said to exhibit weak scaling.

For applications to be able to efficiently use an exascale system, they cannot rely on today's implementations and strong scaling. Exascale applications must handle massive data sets to exploit weak scaling, use new algorithmic approaches, integrate application components that today are processed separately (e.g., full aircraft simulation, instead of separate simulations of fuselage, wings and engines) and optimize the solution of a problem instead of finding a single solution that is 'good enough.'

More energy is consumed, and more time taken, moving data around a system than is used in executing computation on that data. If the goal is to build an exascale system with the caveat that it cannot consume vast amounts of power, then we need to do far more than just optimize the computation. It is important to understand locality of data reference and to ensure that data is processed where it is (whenever possible) rather than moving it to be processed elsewhere. The motivation for optimization will switch from maximizing compute performance to minimizing data movement. Indeed, it has long been the case that when a compute-bound problem is optimized, it turns into a data-access-bound problem.

3.3 RESILIENCY

When systems have millions of cores, avoiding component failure can be very expensive, and is more often completely impractical. In 2009, the fastest system in the world, the Cray XT-5 Jaguar system at Oak Ridge National Laboratory, averaged 52 hours between failures. IBM Blue Gene systems with a similar number of cores fare slightly better. If today's technologies are extrapolated to exascale, the bottom line is that there will always be failed components in a system. Perhaps, in order to save power, some components of exascale systems won't be able to afford luxuries like error correction – so we should anticipate even more component failures, and learn to cope with them.

New approaches and algorithms are required that can handle both massive scalability and resilience. Perhaps algorithms that were rejected in the past because they were not efficient on a single processor may find a new lease on life. In the past, systems provided resiliency. In the future, applications must take on this role.

The most common way that resiliency is built into large HPC applications today is with 'checkpoint restart.' Key data that represents the program's state is stored at regular intervals. If there is a system outage, the application can recover from the time of the last checkpoint. But, when a system has many millions of components, the mean time between incidents will be very short, and the amount of data to be checkpointed very large – so, as the system size grows, you quickly reach the point where the system spends more time dumping checkpoints than doing useful work. Furthermore, how do we cope with component failure that occurs during a checkpoint or restart? The bottom line is that we need to develop a different class of algorithms that can be used to build resilient applications on components that are expected to fail.

The recent RFI published by the US Department of Energy talks of the system overhead to handle automatic fault recovery reducing application performance by up to half. For an exascale system, that is equivalent to five times the total performance of all systems on today's TOP500 list being wasted – just to cope with resiliency issues. We think it is worth significant investment to find a better way of delivering resilient applications on systems where regular component failure is to be expected.

3.4 LACK OF SKILLS

Programming today's multicore GPU-accelerated clusters is becoming more and more complex. At the same time, the level of knowledge of a typical computer science graduate is less deep than it was a decade ago. There are impressive exceptions of course, but many computer science courses teach 'IT' rather than 'computer science,' and many graduates do not have a firm grasp of processor or systems architecture. Where will we find the staff to efficiently program midrange HPC systems, let alone exascale systems requiring billions of parallel threads? This is not just a problem for high-end HPC systems. By the end of the decade, a single rack will hold around 100,000 processor cores. Who is going to write business applications to run on that?

A major IT company told us that of the 1,000 staff that worked on its database product, only 10 programmers were involved in handling parallel programming issues. This is illustrative of where the industry is in terms of coping with parallelism – separating the parallel issues from the majority of the programming and bringing a team of experts to bear on the parallel issues. But if you are one of the 990 serial programmers, do you not need to understand parallel processing? And as parallelism becomes part of the natural landscape of computing, processors will have hundreds of cores, and high-end HPC systems will have millions of cores. It will no longer be satisfactory for a small team to deal with parallelism – in the future, everyone must be a parallel programmer.

A senior manager at a major investment bank recently told us that “programmers today don't understand how computers work” – and his bank pays high salaries for some of the best programmers around. The commoditization of HPC has brought it to a much larger market, while at the high end, systems are becoming far more complex – and we just don't have the overall skills base to cope with it.

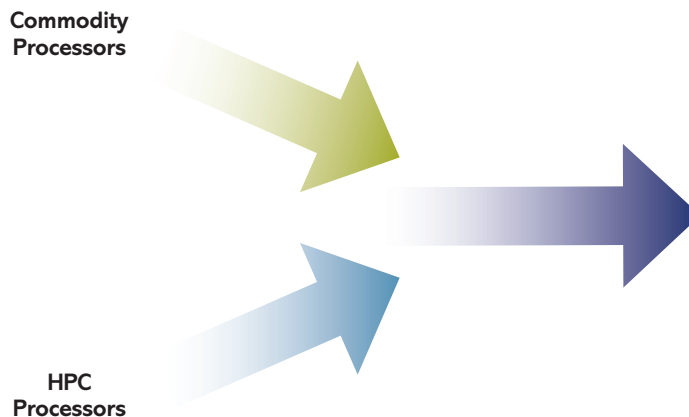
SECTION 4

Processors and Accelerators

The biggest challenge the industry faces in building an exascale system is to do so within an affordable power budget, generally agreed to be around 20MW. This equates to 50 gigaflop/s of performance for every watt consumed. Today, mainstream processors deliver about 400 megaflop/s per watt, while GPUs deliver up to 5 gigaflop/s per watt. NVIDIA's roadmap to 2016 shows that it anticipates reaching the exascale target performance-per-watt level with its extreme-scale Echelon project. Indeed, it claims that a 400-cabinet exaflop/s system would consume 'just' 15MW. One of the conclusions drawn by NVIDIA is that the next challenge is software – which is absolutely right.

In the early days of HPC, the processors used in HPC systems were very different from those deployed in mainstream systems. HPC systems used exotic vector processors built out of gallium arsenide, and even when the first microprocessors were deployed in HPC, it was the obscure Intel i860 that was associated with the 'attack of the killer micros' in the 1990s, and not the mainstream Intel Pentium. All that has changed during the first decade of the new millennium. Almost without exception, from entry level to the fastest supercomputers – including 90% of the systems on the TOP500 list – HPC systems are now based on x86 processors from Intel and AMD, with the optional GPUs from mainstream graphics firms NVIDIA and AMD (through its ATI acquisition).

FIGURE 3: CONVERGENCE OF PROCESSOR TECHNOLOGIES



However, we think that the apparent convergence here may be an illusion. Sure, things are changing. But the major long-term goal for datacenter or cloud processors is the efficient support of multiple virtual machines, while HPC's needs are to handle as many floating-point operations per watt as possible. Can these two goals be met by a single architecture? Possibly, but it is also possible that the use of x86 processors in HPC is not the endgame, but was merely the intersection of mainstream and HPC technologies before they diverge again, driven by very different requirements.

FIGURE 4: DIVERGENCE OF PROCESSOR TECHNOLOGIES



The pressure for very power-efficient processors that exascale requirements bring will create opportunity for a new breed of low-power-consuming, many-core processors.

The option that Intel is pushing is the Knights family of ‘many integrated core’ (MIC) processors, following its experiments with the 80-core Polaris chip, the 48-core ‘cloud on a chip’ and the failed graphics chip known as Larrabee. The first MIC product, code-named ‘Knights Corner,’ will use Intel’s 22-nano-meter manufacturing process and provide more than 50 x86 cores. Development kits (code-named ‘Knights Ferry’), based on pre-production components with 32 cores running at 1.2GHz, were available in 2010. A later variant known as ‘Knights Landing’ is in development, but details have not been announced. At the International Supercomputer Conference in Hamburg in June 2011, Intel claimed that the MIC family would be able to drive exascale systems at just twice the power budget of today’s petaflop/s systems, and that it would be supported by the programming languages and tools delivered by Intel today. However, details of this design and exactly how it will be programmed efficiently have not yet been described.

Meanwhile, the chip designer that has been unexpectedly thrust onto the HPC stage is ARM Holdings. The main requirement of chips in HPC systems to date has been high performance, but for exascale systems, it is

Maxeler Technologies

Today’s fastest computer is Japan’s K system, which will fill some 800 cabinets in its final configuration. Using today’s technology, 100 of these systems would be required to hit the exaflop/s performance level, and it would consume around 1GW of power. Maxeler claims that it can deploy what it calls an ‘exascale equivalent’ system in a very small footprint – perhaps just a handful of cabinets. For some classes of applications, it can build a relatively small machine that delivers the same application performance as a true exascale system by exploiting the unique features of FPGAs.

The future of HPC is heterogeneous systems – with the heterogeneity required to deliver both low power consumption and high performance. The real smarts in future HPC platforms will be compilers and development tools that hide the complexity and help developers access the high theoretical performance levels. We wouldn’t bet against Maxeler (or similar technology) driving future exascale systems.

power efficiency. The ARM family of chips today drives smartphones, tablets and other mobile devices, which require components to consume a very small amount of power – not because of the massive scale of these devices, as will be the case for exascale systems, but to maximize battery life. So ARM may have the right technology for exascale systems, even if it was designed for another purpose entirely.

4.1 HETEROGENEOUS ARCHITECTURES

We are seeing increasing use of application accelerators in HPC systems – mainly GPUs, but we have also seen the use of IBM Cell processors and FPGAs for specialist tasks. This is just the tip of the iceberg, since we cannot build an exascale system by simply throwing more and more commodity processors at the problem due to cost, complexity and power consumption. NVIDIA already uses ARM cores in its Tegra chipset, which also includes audio, video and image processing capabilities. According to NVIDIA, it chose to work with ARM because of the “groundbreaking mix of performance, power consumption and form factor” – exactly the requirements of exascale systems. Project Denver, a collaboration between ARM and NVIDIA, goes beyond the use of ARM cores in the Tegra family, and is explicitly aimed at delivering processors for the supercomputer market, although details of this initiative have not yet been made available.

FPGAs have been on the edge of supercomputing for some time. Convey Computer and Maxeler Technologies both use FPGAs combined with optimized libraries, clever compiler technology and consulting skills to deliver a high level of performance in a very small form factor. The complexity of programming FPGAs has maintained their position as niche products, but all exascale systems are going to be very complex to program, and the significant performance, power-consumption and form-factor benefits that FPGA-accelerated systems deliver may bring opportunities for FPGAs in exascale systems.

4.2 CRAY

Cray is one of the leading vendors of very-high-end HPC systems, and it anticipates delivering an exascale system in the 2018-20 time frame, with budget being the limiting factor, not technology. It plans to use stacked memory delivered with TB/s bandwidth, but exascale systems will have lower network bandwidth, memory bandwidth and memory capacity per floating-point operation than is available today. Cray understand that software tools are fundamentally important in being able to deliver exascale applications, and it is working on auto-tuning compilers and programming models.

4.3 IBM

Many of IBM's laboratories across the US, Europe, Asia and South America tackle issues important to exascale systems. For example, its Zurich facility works in the areas of nano-technology (work that has generated several Nobel prizes), photonics, chip design, and power and cooling issues, while its Dublin facility focuses on application issues for exascale systems.

IBM has been a major supplier of HPC system for many years, and has worked closely with the National Laboratories in the US and major research facilities in Europe to develop several generations of Power-architecture-based supercomputers and Blue Gene systems.

Its most recent project is the \$200m collaboration with the University of Illinois and the National Center for Supercomputing Applications to develop the 10 petaflop/s Blue Waters system. However, as this report was being prepared, it was announced that this contract was terminated because “the innovative technology that IBM ultimately developed was more complex and required significantly increased financial and technical support by IBM beyond its original expectations.” We anticipate that this will not be the only multi-petaflop/s or exaflop/s project that will come unglued.

4.4 INTEL

Intel supports 23 labs across Europe under the banner 'Intel Labs Europe' that employ several thousand people. Three of these facilities – at the University of Versailles, France; the University of Leuven, Belgium; and the Jülich Research Centre, Germany – are focussed on developments looking toward exascale systems, with application scalability, resilience and power management being at the core of the research agenda.

SECTION 5

Memory Technology

There is no point in building a system with millions of processors if the memory is too slow to feed them, or if the memory consumes the total power budget of the system. So new memory technologies are required to improve both connectivity performance and power consumption.

Progress in DRAM manufacture has generally been measured in capacity, not performance. This might be fine for mainstream computing, but it is a major problem for HPC applications. In the 30 years from 1980 to 2010, the clock cycle time of a processor improved by almost three orders of magnitude (more if you factor in the effect of multicores), while memory access time improved by less than a single order of magnitude. So we are more than 100 times worse off than we were in 1980 in terms of delivering data to processors from memory. And this situation will only get worse unless the way that we build memory changes.

A number of techniques are now used to mitigate the problem of slow memory access – we have several levels of cache supported by data pre-fetching – but it’s essentially a case of papering over the cracks, and the cracks are getting bigger. By the time exascale systems arrive, you will be able to drive a bus through the cracks.

Current memory DIMMs (dual in-line memory modules, the current standard in memory chips) are not power-efficient. In order to read 32 bits of data, 8K cells are activated. If this technology were used in an exascale system today, the memory alone would consume more than 100MW. In short, there are many problems that need to be overcome in order to provide the high-performance, high-density, low-power, low-cost memory required for exascale systems.

FIGURE 5: DDR MEMORY

FAMILY	LAUNCHED	I/O BUS CLOCK	VOLTAGE
DDR	2001	100-200 MHz	2.5
DDR2	2004	100-266 MHz	1.9
DDR3	2007	400-1067 MHz	1.5
DDR4	2012	1067-2133 MHz	1.2

Note that DDR stands for ‘double data rate,’ with the ‘double’ signifying that data is transferred on the rising and falling edges of the clock signal, so the effective clock rate is double that shown in Figure 5. DDR4 is an evolutionary technology that won’t bridge the power-consumption gap, while DRAM manufacturer Micron Technology’s proposed hybrid memory cube (HMC) technology takes a different approach that gets much closer to the needs of exascale systems. HMC stacks a number of layers of DRAM on top of a logic layer and the package substrate to deliver 20 times the performance of a DDR3 module, while using one-tenth of the energy in one-tenth of the space. One-tenth of the energy reduces the power requirement of memory in an exascale system to around 10MW, which is in the right ballpark but still high – so while HMC is the right direction, it is not yet the end of the story. Micron expects HMC to be available in volume in 2013.

Another memory technology that may address some of the needs of exascale systems is Memristor, developed by HP Labs. The name is a combination of memory and resistor, a reference to the way the device operates – its resistance ‘remembers’ the level of the last current passed through the device. It is able to store multiple values (not just one or zero like other digital devices), and could be used to build very compact nonvolatile, solid-state devices (SSDs), possibly using three-dimensional design techniques.

The memory hierarchy in future generations of HPC systems will undoubtedly be complex, always aimed at balancing the needs of performance, power consumption and price.

FIGURE 6: MEMORY HIERARCHY

PROCESSOR
Registers
Level 1 Cache
Level 2 Cache
Level 3 Cache
‘Near’ DRAM (e.g., HMC)
‘Far’ DRAM
Flash Memory
Local SSD
Backing Store

Companies that manufacture flash memory include Intel, Micron, Samsung, SanDisk and Toshiba, while Fusion-io, Nimbus Data Systems, Texas Memory Systems, Violin Memory and Virident Systems ship SSD products. Another, slightly different take on SSD is provided by Viking Modular, which populates DRAM slots with SSDs called SATADIMMs. It may appear counterintuitive to use DRAM slots for storage – why not just use more DRAM and cache data? Actually, there are several good reasons for using SATADIMMs. First, the maximum DRAM module today is 32GB, while SATADIMM can grow to 512GB. DRAM is volatile, so data is lost if power fails, while SATADIMM is nonvolatile (like disk storage).

In older generations of systems, it was necessary to populate all the memory slots in order to maximize bandwidth, but with the Intel Nehalem and later processors, this is no longer true. Indeed, higher bandwidth is delivered if fewer DIMM slots are populated, which means that HPC systems tuned for performance often have empty DIMM slots, which can be populated by SATADIMMs to deliver high-performance, low-power storage in a form factor that is effectively zero. HPC system designers can remove drives, drive bays and midplane; run fans slower; and improve air flow for better thermal behavior – all in a smaller chassis that delivers data to applications faster. A 1U Supermicro server can be configured with 12 SATADIMMs for a total of 6TB of storage that can deliver 1.4 million IOPS. SATADIMM looks just like a disk to the system or user – but it draws 2 watts instead of 15.

SECTION 6

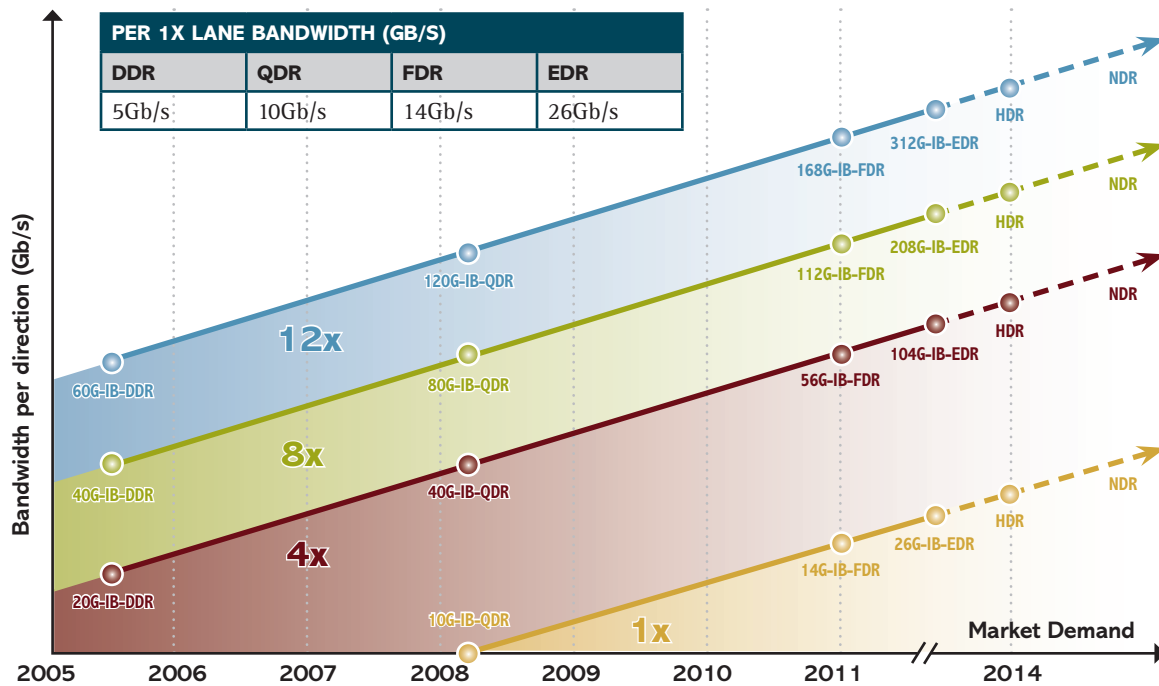
Network Technology

The network for an exascale system must support up to one million nodes with a bandwidth of hundreds of GB/s per link. Bandwidth, latency, resiliency and cost are four crucial dimensions of the networking requirement for an exascale system. The leading networking technology for high-end systems today is InfiniBand. Gigabit Ethernet actually leads the way in the TOP500 list, with 232 deployments against InfiniBand's 206 (in third position is 'proprietary' technology with 23 systems – mainly Cray). But if you look higher up the list, Ethernet's influence drops off – indeed, it only supports one system in the top 50.

The InfiniBand Trade Association (IBTA) roadmap shows the planned development of InfiniBand for the next few years, but does not yet include capabilities that would support exascale systems. A single basic InfiniBand link – or lane – delivers 2.5Gbps. The total bandwidth delivered is a function of both the data rate of each link and the number of links. For example, eight lanes at quad data rate will deliver $2.5 \times 8 \times 4 = 80\text{Gbps}$, represented by the green dot in 2008 on the IBTA roadmap in Figure 7.

InfiniBand silicon, switches and adapters are built by both Mellanox and QLogic. Even in the short term (i.e., by 2014), the InfiniBand roadmap reaches EDR, which has a bandwidth of 26Gbps per lane, and supports up to 12 lanes for a peak node bandwidth of 312Gbps, or almost 40GB/s. Various commentators have proposed that an exascale system requires an inter-node bandwidth of between 50 and 100GB/s. As the InfiniBand roadmap reaches 40GB/s in 2014, it appears that InfiniBand will be able to match the requirements of exascale systems by 2018, at least as far as bandwidth is concerned. Latency is another important issue, and projected latency figures have not yet been divulged.

FIGURE 7: INFINIBAND ROADMAP



Source: InfiniBand Trade Association

FIGURE 8: INFINIBAND PERFORMANCE

NAME	DESCRIPTION	RATE (GBPS)
Serial		2.5
DDR	Dual	5
QDR	Quad	10
FDR	Fourteen	14
EDR	Enhanced	26
HDR	High	Undefined
NDR	Next	Undefined

Alternatives to InfiniBand for driving exascale interconnects include offerings from Cray, EXTOLL and Gnodal. Cray’s Gemini interconnect, used for its XK6 and XE6 supercomputers, is built in a three-dimensional torus (which harks back to the T3D architecture) with a bandwidth of 20GB/s per node, and scaling to support hundreds of thousands of cores.

Meanwhile, two new kids on the block are Gnodal and EXTOLL. Gnodal (whose staff have many years of experience in this market at Meiko and Quadrics) has built a high-performance, low-latency switch for 10/40 Gigabit Ethernet, initially targeting the storage market, but it will have a stronger offering for the HPC market when 10-Gigabit Ethernet is commonly available on motherboards (to keep the cost down).

EXTOLL (a startup out of Heidelberg University) claims to deliver higher bandwidth than QDR InfiniBand, with a latency of one-third that of InfiniBand. It also uses a switch-less 3D torus architecture, and claims lower power consumption than its rivals.

SECTION 7

Data Management

Petascale systems today handle hundreds of thousands of concurrent I/O operations simultaneously, supporting an aggregate bandwidth of 200GB/s. The target aggregate I/O bandwidth for exascale systems was estimated to be roughly 20TB/s in the US Department of Energy Exascale Initiative Roadmap.

While processing speed has gone through the roof in recent decades, disk performance has not kept pace. This was the initial motivation for parallel file systems. As supercomputers evolve and run faster and faster, access to the data becomes more of a bottleneck – so a highly scalable parallel file system, along with appropriately architected storage servers, is required to complement the parallel processing capabilities of modern supercomputers. As the industry moves toward exascale systems, the data-access problem will only get worse.

7.1 HPC STORAGE SYSTEMS

In terms of HPC storage systems, BlueArc develops high-performance NAS systems that are accelerated by the use of optimized code running on FPGA processors. Many file systems can be tuned to be blindingly fast at one type of I/O operation, at the cost of poor performance elsewhere. But BlueArc has demonstrated predictable and consistent performance across a wide range of read/write mixes, with small and large block sizes, where performance also scales as more disks are added. The addition of parallel NFS (pNFS) support will give BlueArc a much stronger story in the scale-out storage market segments showcased by cloud and compute clusters.

DataDirect Networks does not have its own scale-out file system, but because it supports IBM General Parallel File System (GPFS), Lustre and Quantum's StorNext, the company has the flexibility to fit into a wide range of markets. To simplify things for customers, it has developed pre-integrated offerings such as its ExaScaler HPC file storage system, which comes with SFA or S2A storage systems and parallel file storage gateways running Lustre. For customers requiring HPC storage with advanced capabilities such as snapshots and data migration, DataDirect has a GridScaler offering that integrates its storage with IBM GPFS.

Meanwhile, the PAS 12 is Panasas' flagship system. In a standard 40U rack, the PAS 12 can hold up to 400TB of storage while delivering 15GB/s of throughput performance. Panasas claims the PAS 12 can scale up to 4PB of capacity, in 40TB increments, with maximum throughput performance of 150GB/s. A PAS 12 can deliver 1,500MB/s of throughput in benchmarks. For single-client NFS workloads, the PAS 12 can handle up to 7,000 IOPS, compared with 3,500 IOPS for the PAS 8. In addition to the new hardware, Panasas had added new software functions, including user quotas. It also now has active/active failover support for network connectivity (in place of the old active/passive design) to maximize bandwidth utilization.

Xyratex is a major OEM supplier of storage systems, with 25% of its products being deployed in the HPC market. All of the components of its HPC strategy are brought together in the ClusterStor CS-3000, a Lustre storage appliance that delivers reliability, scalability and performance. The CS-3000 is built from modules that comprise a pair of servers, 84 3.5-inch disks and network connectivity, and it can scale to 672 drives – that’s more than 2PB – in a single rack. The base unit supports 5GB/s bandwidth; the full rack supports over 20GB/s; and the system can expand to over 30PB and 100GB/s over QDR InfiniBand or 10-Gigabit Ethernet fabrics. The third dimension to this product – complementing the hardware platform and Lustre – is the management stack that enables rapid deployment and provisioning, performance tuning, monitoring and management with non-disruptive rolling upgrades. Health monitoring covers both the physical and virtual layers, as well as the Lustre file system, and supports root-cause diagnostics and proactive preventative maintenance.

7.2 PARALLEL FILE SYSTEMS

IBM’s GPFS provides concurrent, high-speed file access to applications running on multiple nodes of AIX, Linux, Windows and heterogeneous clusters. It also provides storage management and information lifecycle management tools. The file system separates metadata and user data, and uses byte-level locking to maintain consistency and coherence during concurrent access from multiple nodes. Data is striped across all disks in a storage pool for performance and scalability. GPFS provides high-performance metadata access, while data availability is supported by high-availability clustering, file-system snapshots and replication for metadata and user data.

Cluster configuration options include shared disk (direct-attach to all machines in a cluster by a block-level protocol such as SCSI using a SAN, InfiniBand or iSCSI); network-based block I/O over a LAN when access to multiple SANs is required, which enables sharing data between clusters at higher performance than CIFS or NFS; and sharing data between clusters. GPFS is the second most popular parallel file system for high-end supercomputers behind Lustre.

Lustre is a scalable, object-based parallel file system, which originated in the Linux cluster market – hence its name. A Lustre file system (which is available under the GNU General Public License) is composed of a metadata server, one or more object storage servers, and clients that use the data. LNET, the Lustre Network, has been implemented on a range of technologies, including InfiniBand, and can use RDMA to improve performance. Lustre uses RAID and file striping to deliver resiliency and performance, and a distributed lock manager to manage concurrent file access. Further resiliency features include high-availability clustering for both metadata servers and object storage servers, making server failures, reboots and software upgrades transparent to the user.

Lustre supports POSIX file-system semantics, industry-standard servers, and native InfiniBand and Ethernet. It scales to petabytes of data delivered at hundreds of Gbps aggregate performance.

Following Oracle’s 2010 announcement that future versions of Lustre would only be supported on Oracle/Sun hardware platforms, a number of community groups were formed guarantee the

future of an open version of Lustre. These were the European Open File System Cooperative (EOFS) and two US-based initiatives – the Open Scalable File System (OpenSFS) group and the High-Performance Cluster File System (HPCFS) group. OpenSFS raised funds from national laboratories and interested vendors to support the ongoing development and support of Lustre. In 2011 the grassroots HPCFS and OpenSFS merged, and also signed a memorandum of understanding with EOFS to provide a single community effort supporting the future of Lustre.

Both Xyratex and Whamcloud are committed to supporting Lustre for the HPC market. One part of that strategy is to make Lustre easier to use and manage, in order to take it to a much wider audience than its high-end users today. Whamcloud has a multiyear contract with OpenSFS covering performance and namespace enhancements, and an online file-system checker that will maintain distributed coherency across the file system.

NFS is ubiquitous in mainstream computing, while the parallel file-system market, which is dominated by Lustre and GPFS, is a niche market, partly due to the lack of a standard parallel file system. The arrival of parallel NFS (which is included in the NFS 4.1 standard, published early in 2010) expands the market for parallel file systems, and may challenge the incumbents. One attraction of pNFS is that a generation of systems administrators is used to managing NFS environments, and the step to pNFS is a relatively small one compared to the move to GPFS or Lustre. Development of pNFS has been supported by BlueArc and Panasas, as well as many other vendors for which HPC storage is not a large part of their business. Volume deployments, as part of standard Linux distributions, are expected in 2012. pNFS copies the behavior of parallel processing by using multiple NFS servers operating in parallel to overcome the serialization bottleneck of access to metadata.

7.3 THE EXASCALE CHALLENGES

The HPC community has had to adopt nonstandard storage technologies and parallel file systems in order to support the needs of high-end systems today. But extending today's implementations may not be sufficient to support the needs of exascale systems. The headline performance targets for exascale file systems highlighted in the US Department of Energy Exascale Initiative Roadmap are 20TB/s throughput and 300PB capacity. The challenges faced in building exascale file and storage systems include:

- Addressing the issues of lower storage capacity and bandwidth relative to the peak performance of today's systems
- Achieving extreme scalability
- Ensuring resiliency (both high availability and reliability of data)
- Supporting a mix of cheap components and high-performance devices.

Exascale systems will need a thousand I/O nodes with an aggregate bandwidth of around 20TB/s. How can data be stored at this rate in a fault-tolerant way? Data must be staged off compute nodes as quickly as possible – but, if all of these threads write data simultaneously, how do we avoid data flow congestion? Indeed, the current resiliency model of always expecting a server to recover will no longer be valid at exascale.

The file systems used by the TOP100 supercomputers in November 2010 were Lustre (65), GPFS (21), PanFS (3) and ‘other’ (11). Six months later, Lustre had increased its footprint to 70 systems, including eight of the top 10 and all of the top three.

Of the two companies that collaborate to support the Lustre file system today, Whamcloud believes that Lustre is the file system for exascale, while Xyratex thinks it is time to develop a new file system to meet the needs of exascale. Lustre is far more widely used at the top end of HPC than at the entry level, and exascale is the next ‘top’ for HPC. The two companies agree that there are massive challenges to be faced, whether an evolution of Lustre, or an alternative, is the file system for exascale.

Xyratex’s take on an exascale file system is that Lustre is almost 10 years old, and it may be easier to implement a file system to support exascale from scratch than to augment Lustre, which, while widely used by top systems today, is very complex. Xyratex plans an alternative called Colibri, but has not yet talked about it in detail. Whichever file system is used on the server side – Lustre, Colibri or something else – the scalable object store must support replication, migration and tiered storage, and pNFS will probably feature at the client end.

SECTION 8

Software Issues

Software issues for extreme-scale computing involve more than just parallel programming, and also cover operating systems; runtime systems for scheduling, memory management, communication, performance monitoring, power management and resiliency; computational libraries; compilers; programming languages; and application frameworks. Critical challenges that are common to both small and large HPC systems are multicore processors, with the expectation of hundreds or thousands of cores per chip; massive parallelism; constraining energy consumption; and building resilient applications on top of an infrastructure that will always include failed components. The main issues that impact software for exascale systems can be summarized as energy efficiency, parallelism and resiliency.

The majority of today's HPC systems are high-performance clusters based on commodity components, running more or less standard Windows Server or Linux operating systems. In a system with thousands or millions of nodes, minimizing and hiding latency of I/O, data transfer and inter-node messaging is a must. This may be achieved by optimizing existing operating systems, or it may require a lightweight operating environment to be developed for compute nodes, while a full OS runs on management nodes. The operating environment must minimize not only the latency of operations, but also the variability of that latency (or jitter), as the performance of a parallel application is often influenced by the performance of the slowest node.

OpenMP (Open Multi Processing) is an application programming interface (API) that supports portable shared-memory parallel programming in C/C++ and Fortran.

MPI (Message Passing Interface) is an API that supports communication through message passing for parallel programs running on distributed-memory systems, such as clusters.

CUDA (Compute Unified Device Architecture) is a parallel computing architecture developed by GPU manufacturer NVIDIA, but the term CUDA is commonly used to describe its parallel programming framework, which is used to program GPUs as compute accelerators in HPC systems.

OpenCL (Open Computing Language) is a framework based on C99 for writing programs that execute on heterogeneous platforms consisting of processors, GPUs and other types of accelerators.

PGAS (Partitioned Global Address Space) is a parallel programming model that uses a global memory address space that is logically partitioned across distributed memory nodes.

8.1 PROGRAMMING METHODOLOGY

Early HPC programs targeted a single processor, since that was all that was available. Many codes were then adapted for vector machines or shared-memory parallel systems, initially using proprietary compiler extensions, and later using OpenMP.

As clusters of commodity systems became popular, the HPC industry moved from OpenMP, via proprietary message-passing libraries, to MPI. This process took many years, initially requiring the standardization of MPI, then support from software vendors to make a body of applications available.

The jury is still out as to whether the existing programming paradigms of MPI between nodes and OpenMP within nodes, with CUDA (or something similar) targeting heterogeneous accelerator technologies, will work at exascale. If this mix won't work, then the industry is in a deep hole – the transition to MPI took about a decade, but the first exaflop/s systems are expected to arrive in 7-8 years. There is a great danger that we will have fantastically powerful systems, but no applications to run on them. And don't forget resiliency – our applications must be built to expect and cope with component failures. Can OpenMP and MPI do that today? The answer is no.

Automatic parallel compilation has been the unreached 'holy grail' of compiler writers for several decades. As the order of concurrency in applications for exascale systems reaches the billions, current methods will not scale. PGAS languages show promise, but they work best with regular, static parallelism, while real-world problems are dynamic.

The good news is that the world is a naturally parallel place – we need to unlearn the process of turning parallel phenomena into serial programs. We also need tools that allow us to express problems without worrying about the underlying architecture the application will run on – tools that will exploit parallelism, support portability across platforms and handle resiliency. Without those we won't have any applications to run on our shiny new exascale systems.

There are a number of measures of efficiency that should be considered. Previously, the only measure of real concern was the compute performance of an application, but now we must also consider programmer productivity (as heterogeneous architectures are difficult to program) and power efficiency. Programmer productivity is not just about writing an application; it's about doing so portably. If the future of HPC systems is, as we believe, to be heterogeneous, then porting an application from one system to another is potentially far more complex than moving it from one x86 cluster to another x86 cluster.

8.2 SOFTWARE DEVELOPMENT TOOLS

The programming paradigms, languages and software tools used for exascale systems will almost certainly be different to what is used today, but the companies that currently lead in this field, and some of the emerging approaches and technologies, are likely to be part of the landscape going forward.

Intel has a comprehensive set of parallel development tools, which have been supplemented through its acquisitions of companies with promising technologies including Cilk Arts, Pallas and RapidMind. Intel's toolkit today comprises parallel compilers and libraries (including MPI), the Intel Cluster Studio, and many other tools that support the development, debugging and optimization of parallel codes.

Compilers from the Portland Group (known as PGI, a subsidiary of STMicroelectronics) support multicore processors and NVIDIA GPUs – either in CUDA mode or using PGI's more general Accelerator programming model. PGI has also developed debuggers and profilers that support OpenMP and MPI. NVIDIA supports its GPUs with the CUDA C compiler, mathematical libraries and its participation in a very active HPC community.

Another vendor that supports CUDA is France-based CAPS entreprise, whose heterogeneous multicore parallel programming tool, HMPP, is a high-level source translator for C and Fortran that targets CUDA and OpenCL for NVIDIA GPUs and AMD's FireStream. PathScale's ENZO also supports the HMPP directives, which are being discussed within the OpenMP community as a possible extension to the standard.

OpenCL was originally written by Apple, and its development is now maintained by the nonprofit Khronos Group – initially in collaboration with AMD, IBM, Intel and NVIDIA, but now supported by a much larger group of vendors. OpenCL is a framework for writing parallel programs on heterogeneous systems, including multicore processors and GPUs. It comprises a language specification (based on C99), a platform-layer API that provides a hardware abstraction, and a runtime API to map the execution to the available resources. Its execution model is a host program that calls kernels, which may be task- or data-parallel. A variety of memory types are supported, and the movement of data between them is explicit. OpenCL compilers are available from AMD, Apple, IBM, Intel, NVIDIA and others.

Meanwhile, ET International, a spinoff from the University of Delaware, has developed its Swift Adaptive Runtime Machine (or SWARM) in a collaborative project with IBM, funded by the US Department of Energy. Its innovative way of running applications delivers greatly improved parallel efficiency for C and Fortran codes that use OpenMP or MPI, and it currently works with IBM's experimental Cyclops64 platform, x86 systems and the Epiphany processor from Adapteva.

Compilers are only part of the software development story – debugging and profiling tools are even more important for parallel codes than they are for serial codes – and 'printf' is not an effective way of debugging a million-way parallel code. Allinea Soft-

ware's DDT debugger has been deployed on some of the world's largest systems, and it can monitor and trace information at great scale in the blink of an eye. TotalView (acquired last year by Rogue Wave Software) is the main alternative parallel debugging tool. And Jinx, a software tool from Corensic, discovers concurrency bugs in parallel codes – which can be intermittent and therefore difficult to track down and fix, especially at large scale.

8.3 NEW APPROACHES TO PROGRAMMING TOOLS

Partitioned Global Address Space (PGAS) describes a class of programming languages that use an abstracted address space to exploit the best of both worlds – the ease of programming of shared-memory OpenMP code, and the scalability of distributed-memory MPI applications. Languages that use the PGAS approach include Unified Parallel C, Co-Array Fortran (a small extension to Fortran 95), Titanium (an extension to Java), X-10 (an open source project led by IBM) and Chapel (a new language developed by Cray).

The Global Address Space Programming Interface (GPI), was developed by the Fraunhofer Institute and is sold by Scapos, its commercial arm. GPI provides an interface that can be used directly with C, C++ and Fortran, so the programmer doesn't need to learn a new language.

Meanwhile, The Barcelona Supercomputing Center has been developing its Encore compiler suite for several years. This is an extension to OpenMP that assists with the automatic detection of parallelism, targeting shared-memory multicore systems, GPUs and FPGAs. The goal for Encore is improve the usability, portability and performance of parallel code through the exploitation of nested parallelism and the automatic overlap of communication and computation. Nested parallelism means that parallelism is available at many different levels within a system. A processor core can do more than one thing at a time, there are multiple cores on a chip, there are multiple chips in a server, and there are multiple servers in a system. Rather than treating all cores as being part of a flat landscape, it is more efficient to chop an application into chunks and process each chunk on a server, then use parallelism at a lower level to spread computations across processors and cores.

8.4 ALGORITHMS FOR EXASCALE

Many of the algorithms in use today were selected because they were efficient at relatively low levels of parallelism. There are two attributes that are very important for algorithms to be run on exascale systems. One is extreme scalability, the other is resiliency. The latter issue will be discussed in more detail in the next section. As noted earlier, minimizing data movement is more important than maximizing compute performance. The industry will need to revisit the algorithms used for HPC applications to find ones that fit the requirements of 2018, which will be very different from the requirements now. In some cases, old algorithms that were deemed to be inefficient in the early days of HPC may provide the scalability and resiliency required. In other cases, new methods will have to be developed.

It is critical to start at the algorithm level. Synchronization will be the main limiting factor – that and data movement are the major challenges, and should be targets for optimization, as opposed to flop/s. The use of algorithms that support ‘mixed precision’ should be considered in order to minimize the amount of data to be moved. Numbers are usually held in a 64-bit word, and most calculations are performed on these – this is known as ‘double precision,’ since many years ago the normal data size was 32 bits, or ‘single precision.’ Of course 64-bit data and calculations are more accurate, and this is normally a good thing. However, sometimes the data itself is not that accurate, and 32 bits is quite sufficient. For applications that only require 32-bit data and calculations, using double precision would normally be seen as mild overkill, and nothing more. But at exascale, doubling the amount of data handled by a system can be very significant, so this issue should be reconsidered. Fault-resilient algorithms are an absolute necessity, since components will fail during application runs.

Many users, particularly groups responsible for porting code to new platforms, are concerned about the reproducibility of results. While this is an important issue, it must be addressed in a reasonable way. It is impossible to guarantee the reproducibility of results from a parallel application at the bit level unless the computations are constrained by guaranteeing the order of computation, and that destroys performance. Floating-point numbers are inexact, and performing the same calculations on floating-point numbers in different orders will result in different answers. Not wrong answers, just different answers – all within certain limits.

SECTION 9

Exascale Initiatives

The research and development costs of an exascale system will be in the billions of dollars, while the purchase price (and the annual maintenance, power and cooling costs) will be measured in hundreds of millions of dollars. No commercial company can afford to invest on this scale, while few outside of national laboratories can afford to buy the resultant system. Having said that, the results of an exascale development will have massive spinoff benefits for midrange and entry-level systems. At any rate, the support of national and international exascale initiatives is of fundamental importance.

There are many organizations, projects and initiatives considering aspects of the technology evolution that is required to reach exascale – probably too many. The hardware, software and user application challenges that exascale systems will bring are many and complex. We do not believe that the approach of many projects tackling different aspects of these challenges will provide a coherent solution. Indeed, in some cases, there are several projects tackling the same issues. To solve such a big problem, perhaps a ‘benevolent dictator’ is required to orchestrate international efforts – ideally, one with very deep pockets. This section briefly discusses a number of the major projects that address exascale issues.

9.1 EUROPEAN EXASCALE INITIATIVES

The European Commission has set several exascale objectives to be targeted with current project funding. They include:

- Projects to develop a small number of advanced computing platforms (100 petaflop/s in 2014, with potential for exascale by 2020)
- The development of optimized application code driven by the computational needs of science and engineering, and of today’s grand challenges
- Addressing the issues of extreme parallelism with millions of cores (programming models, compilers, performance analysis, algorithms, power consumption).

The European Commission has recently agreed to fund three projects – called Mont Blanc, Cresta and Deep – that focus on the software issues that must be addressed if exascale systems are to become a reality, although two projects also include hardware components from Cray and Intel.

9.1.1 EUROPEAN TECHNOLOGY PLATFORMS

European Technology Platforms (ETPs) are bodies supported by industry and academia that align research priorities in industrial areas that are critical to the growth of the European economy. The PROSPECT e.V. organization was formed to promote the ETP

for HPC, whose purpose is to stimulate growth of the HPC supply chain and to help guide European research priorities. The ETP for HPC's vision for 2020 is:

1. Europe will be a global leader in inventions that successfully exploit HPC resources.
2. HPC will provide increasingly accurate and responsive predictions on
3. both short- and long-term phenomena that impact European citizens, such as weather, climate change and epidemics.
4. The use of HPC will enable Europe to achieve world-leading levels of
5. energy efficiency in areas such as electricity generation and transmission, transport, food supply chain, water supply, and product design.
6. HPC technologies designed in Europe will lead in energy efficiency.
7. Europe will have the know-how and skills to master the development of HPC technology components in the areas where it excels, and to master the exploitation of HPC across a wide range of industries and disciplines.

While we applaud the objectives of PROSPECT, we have some misgivings. First, it is not ambitious enough. Second, we wonder how 'European' an organization can be when it includes Cray, Dell, HP, IBM, Intel, NVIDIA and Oracle as members. Finally, where is ARM Holdings, the leading provider of low-power processors – a crucial technology for exascale systems?

9.1.2 PLANETHPC

PlanetHPC is a Support Action funded by the EU's 7th Framework Programme. Launched in November 2009, this two-year initiative provides a forum for European researchers and industrialists to identify the research challenges facing HPC. It brings together major European HPC players from the scientific and business sectors – users, service providers, hardware providers and software providers – with the objective of coordinating activities, strategies and roadmaps for HPC in Europe. A recent workshop of HPC stakeholders (in which The 451 Group participated) concluded that there is a significant shortage of HPC skills, and that the industry is on the brink of significant disruptive change.

9.1.3 E-INFRASTRUCTURE REFLECTION GROUP

The mission of the e-Infrastructure Reflection Group (e-IRG) is to pave the way toward a general-purpose European 'e-infrastructure.' Its vision is an open e-infrastructure enabling flexible cooperation and optimal use of all electronically available resources. The e-IRG is an inter-governmental policy body, recognized as an advisory body by the European Commission, with national delegates appointed by the member-state ministries from more than 30 European countries, as well as representatives from the European Commission. The e-IRG holds a number of workshops and delegate meetings each year to analyze key issues and to form and communicate consensus recommendations. The outcomes of e-IRG recommendations so far have been the establishment of PRACE and the European Grid Initiative (EGI).

The e-IRG has published a white paper that discusses key e-infrastructure issues and topics that require policy action, and forms the basis for proposing formal e-IRG recommendations at the national and EU levels. It has also published a roadmap responding to emerging technologies, paradigm shifts and long-term strategic issues.

The e-IRG white paper identifies a number of challenges in moving to exascale systems, including:

- The need for a new programming model beyond MPI.
- Establishing a performance indicator more sophisticated than the flop/s rate
- Leveraging of heterogeneous computing by operating systems, software libraries, compilers, toolkits, etc.

The recommended approach proposed a tight focus on software issues and development tools to support scientists and engineers who are not computer scientists. The white paper's recommendations regarding exascale were:

- Encourage the development of European hardware technology in order to compete and cooperate with the current leading countries in HPC.
- Dedicate resources to the study of new programming models, algorithms and languages, porting software libraries and software tools to exascale environments, and preferring open source software solutions to leverage existing know-how in a cost-efficient way.
- Identify new grand challenges in science that are able to utilize the exascale platforms.
- Partnership between users of exascale computing, industry and computer scientists must be encouraged, and scientists should be given the opportunity to liaise with programming experts.
- Specialists must create training materials, including robust and easy-to-use 'cook books' for users, especially for those who are not computer scientists.
- Ensure that the value of the scientific case for exascale computing is well understood and appreciated by society at large by means of knowledge dissemination and engagement with the public, policy-makers and industry.

While these recommendations are sound, the cost to implement them effectively would be hundreds of millions of euros.

9.1.4 PARTNERSHIP FOR ADVANCED COMPUTING IN EUROPE

The Partnership for Advanced Computing in Europe (PRACE) supports a small number of Tier 0 facilities across Europe providing services for the scientific and industrial user communities, coordination with HPC service providers on all tiers, and cooperation with the European HPC industry and other e-infrastructures.

PRACE was established in 2010 with around €500m in funding for the period 2010-15, from the countries where the Tier 0 facilities are hosted, the other PRACE members and the European Commission. Figure 9 illustrates current and emerging PRACE facilities. Access to the systems is free of charge, and is based on peer review of project proposals.

FIGURE 9: PRACE SYSTEMS

PRACE CENTER	LOCATION	NAME	SYSTEM	PERFORMANCE
Forschungszentrum Jülich	Jülich, Germany	JUGENE	IBM BlueGene/P	1 petaflop/s
CEA-GENCI	Paris, France	CURIE	Bull Tera 100	1.6 petaflop/s
High Performance Computing Center	Stuttgart, Germany	HERMIT	To be installed at end 2011	Initially 1 petaflop/s
Leibniz Rechenzentrum	Munich, Germany	SuperMUC	IBM, to be installed in mid 2012	3 petaflop/s
N/A	Italy		To be announced in 2012	
N/A	Spain		To be announced in 2013	

9.1.5 EUROPEAN EXASCALE SOFTWARE INITIATIVE

The European Exascale Software Initiative (EESI) is a Coordination and Support Action funded by the European Commission with 17 academic and industrial partners across Europe. It has produced a report describing existing major HPC initiatives worldwide. The motivation of EESI is to coordinate European contribution to the IESP, to enlarge the European community involved in software roadmapping activity, and to build and consolidate a vision and roadmap at the European level to address the challenge of using multi-petaflop/s and exascale systems effectively. It has a strong focus on industrial applications, and is examining the needs of the aeronautics, structures, chemistry, energy, oil and gas, engineering and life sciences markets. One interesting conclusion already reached is that to handle the real-time simulation of an aircraft in flight, exascale systems are not enough – zetascale systems are required!

9.2 US EXASCALE INITIATIVES

The main users and funders of supercomputers in the US are the Defense Advanced Research Projects Agency (DARPA), US Department of Defense (DoD), US Department of Energy (DoE), the National Aeronautics and Space Administration (NASA) and the National Science Foundation (NSF).

9.2.1 DARPA

The DARPA Ubiquitous High-Performance Computing (UHPC) program is developing architectures and technologies that will provide the underpinnings, framework and approaches for the resolution of power consumption, resiliency and productivity problems. The UHPC program aims to develop computer systems, from embedded level to

cabinet level, that have extremely high energy efficiency, and are dependable and easily programmable. These systems will have dramatically reduced power consumption while delivering a thousand-fold increase in processing capabilities.

Dependability technologies developed by the UHPC program will provide adaptable and hardened cyber-resilient systems. Productivity will be significantly improved by developing scalable, highly programmable systems that will not require significant expertise for the development of high-performance applications. UHPC has funded four projects for four years, at a total of \$76.6m. This activity is augmented by the DARPA Omnipresent HPC program, which supports UHPC by developing advances in a wide range of software capabilities.

9.2.2 US DEPARTMENT OF ENERGY

The mission of the DoE Advanced Scientific Computing Research (ASCR) program is to discover, develop and deploy computational and networking capabilities to analyze, model, simulate and predict complex phenomena important to the DoE. The program funds a range of projects that tackle exascale-related challenges, including high-performance networking, exascale co-design centers, advanced architectures and scientific data management. Project funding is typically \$5m-\$10m per year for five years.

Seven DoE labs – the Argonne National Laboratory, Lawrence Berkeley National Laboratory, Lawrence Livermore National Laboratory, Los Alamos National Laboratory, Oak Ridge National Laboratory, Pacific Northwest National Laboratory and Sandia National Laboratories – have formed a consortium known as E7 to publish an RFI for the HPC industry to plan for the development of exascale technologies and systems. Its roadmap has prototypes and test beds available in the 2017-18 time frame, with exascale systems and applications being delivered in 2020-21. The purpose of the RFI is to enable the DoE to plan its exascale program, which will include significant design collaboration between the DoE and the HPC industry, supported by significant investment.

The DoE's targets for 2019-20 include a Linpack performance of one exaflop/s with a power consumption of no more than 20MW, 128PB of memory, 4TB/s node memory bandwidth, and 400GB/s node interconnect bandwidth. In terms of resiliency, the RFI says "The system overhead to handle automatic fault recovery must not reduce application efficiency by more than half."

9.2.3 THE INTERNATIONAL EXASCALE SOFTWARE PROJECT

The International Exascale Software Project (IESP) was set up to address the need for a software infrastructure to support exascale calculations – which it does not believe exists today. Hardware has advanced significantly in recent years (and further revolutionary hardware changes are required to reach exascale), but the software ecosystem has stagnated. The IESP's goal is to improve the coordination and development of the HPC software environment. It is building an international plan for developing next-

generation open source software for scientific HPC. It is funded by the DoE and NSF. The roadmap that the IESP is working on covers the following components:

- System Software
 - Operating Systems
 - Runtime Systems
 - I/O Systems
 - Systems Management
 - External Environments
- Development Environments
 - Programming Models
 - Frameworks
 - Compilers
 - Numerical Libraries
 - Debugging Tools
- Applications
 - Algorithms
 - Data Analysis and Visualization
 - Scientific Data Management
- Cross-cutting Dimensions
 - Resiliency
 - Power Management
 - Performance Optimization
 - Programmability

The IESP is currently undergoing a series of transitions. It is moving from deciding what to build toward working out how to build it, translating the roadmap into a project plan. It is developing the appropriate organizational structure and collaborating with funding agencies, laboratories, universities and vendors to promote a co-design approach for future architectures.

9.3 OTHER INTERNATIONAL INITIATIVES

The fastest computer in the world today is at the RIKEN Research Institute in Japan, and it was built by Japanese company Fujitsu. The system it displaced as the fastest is the Tianhe system in China, which was built by NUDT, a Chinese company. The fastest computer in Russia is the Lomonosov system at Moscow State University, which was built by T-Platforms, a Russian company. Although built by indigenous computer companies, the systems from NUDT and T-Platforms use components made by Intel and NVIDIA, two US companies – although the Chinese government plans for a future generation of supercomputers to use processors developed in China.

Meanwhile, India has announced an almost unbelievable plan to develop a super-computer with a peak performance of 132.8 exaflop/s by 2017, according to Ashwini Kumar, the Minister of State for Planning. A budget of around \$2bn has been allocated to this project.

SECTION 10

The 451 Take

The HPC industry's transition from petascale to exascale systems will require disruptive change to many technologies. Vendors will have to change the way they build components and systems, while users must change the way they write applications. We have no doubt that such systems can be built, but we have serious concerns about whether the software industry will be ready with high-performance, scalable, resilient applications to run on these exascale systems when they appear toward the end of this decade. Without software support, an exascale system will be akin to Howard Hughes' H-4 Hercules, nicknamed the Spruce Goose, which is still – by wingspan – the largest aircraft ever built. It was a marvelous feat of engineering, but it only flew once, and was consigned to aviation museums after Hughes' death.

The Gordon Bell prize, awarded to recognize outstanding achievement in HPC applications and encourage development of parallel processing, was awarded for an application that ran at 1 gigaflop/s in 1988, 1 teraflop/s in 1998 and 1 petaflop/s in 2008. If this rate of progress is followed, the 2018 Gordon Bell prize would be won by an application that runs at a sustained exaflop/s rate. Given the new technologies that must be built and integrated in order to deliver an exascale machine, we do not expect an application to have achieved a 1 exaflop/s performance rate in this time frame. Indeed, we would be surprised if the simple Linpack benchmark has exceeded 1 exaflop/s by 2018.

With so many disruptive changes required at the same time, it is unlikely that any single organization can successfully design all of the components and build an exascale system on its own. Indeed, IBM's decision to walk away from the Blue Waters project supports this position. While significant funding is being applied to exascale research, we believe there are too many relatively small pots of money funding independent activities. Coordination of R&D activities supported by a very significant sum of money is required if the HPC industry is to come close to delivering an affordable, usable exascale system by the end of the decade within the power budget of 20MW.

This presents a great opportunity for Europe to challenge the incumbent US-based supercomputer companies. But finding the political will to aggregate and manage sufficient funds to drive a broad exascale program to the benefit of European industry is proving difficult, even if the model demonstrated spectacular success in the past with Airbus, which now outsells Boeing in the commercial airliner market.

10.1 IMPLICATIONS FOR USERS, STARTUPS, INCUMBENT VENDORS AND INVESTORS

10.1.1 IMPLICATIONS FOR USERS

User of high-end supercomputers must give serious thought to rewriting their applications to cater to the capabilities of exascale systems. This is a big task, and is not something to be undertaken lightly. However, million-way-plus parallelism, changes in internal system balance and the fact that resiliency must now be built into applications – and not the system itself – will force the hand of most application owners. While 2018 seems a long way away, it will take years to re-architect applications, test them on intermediate-sized systems, optimize performance for the final target systems and revalidate the new version of the application. Even starting now may already be too late.

Few users will be able to afford exascale systems in 2018 (after all, there are only 13 systems in the world today with a peak performance over one petaflop/s), but leading-edge technology quickly becomes mainstream. Users with smaller budgets need to prepare their applications for affordable petascale systems.

10.1.2 IMPLICATIONS FOR STARTUPS

It is not often that so many disruptive changes happen to an industry at the same time. The old order will be challenged on many fronts, bringing opportunities for new entrants. The hardware challenges will no doubt be very expensive to overcome and, while the software challenges are no less complex, perhaps this is an area where startups can find a niche to make their own. Compiler technology has progressed slowly, steadily and incrementally since the 1980s. If exascale systems are to be programmable, the industry needs compiler technology to take an innovative leap forward, handling complexity and massive parallelism efficiently, while also offering ease of use to the developer.

10.1.3 IMPLICATIONS FOR INCUMBENT VENDORS

The big win for incumbent HPC vendors is not in delivering a small number of exascale systems, which will, even if things go very well, require massive investment to develop and support – and which are unlikely to be profitable on their own. It is the thousands of resulting petascale systems that will provide a strong financial return.

The HPC industry has been transformed during the last decade by the adoption of commodity components. The transition to exascale will almost certainly break that model, with the processors used to drive high-end HPC applications being very different from those that manage Web services in the cloud or databases in enterprise datacenters. There will be great opportunities and serious challenges for all involved in building exascale systems. Building a coherent transition path from today's high-end HPC systems and applications to those of tomorrow will not be easy, but will be necessary to keep customers on board through the transition.

That journey – the transition from one generation of technology to the next – is never-ending, and is at its most extreme in HPC. Vendors need to balance the needs of the lunatic fringe with those of the mainstream – bearing in mind that leading-edge platforms delivered to the former today will become the profitable, high-volume products for the latter in just a few years.

10.1.4 IMPLICATIONS FOR INVESTORS

At the system level, the sums of money involved in developing, building and running a machine with exascale capability will be enormous, as will the risks in funding such a project. At the component level, and for software tools, the costs are more reasonable, and the requirements clear – to handle massive parallelism with low latency and resiliency built in. These requirements apply to processors, networks, memory, middleware and software development tools. Small companies with innovative approaches and unique skills will be able to find a market or – more likely – an acquirer if they can demonstrate that they have effectively solved part of the exascale problem.

INDEX OF COMPANIES AND ORGANIZATIONS

- Adapteva 26
- Allinea Software 26
- Apple 26
- Argonne National Laboratory 33
- ARM Holdings 8, 13, 14, 30
- Barcelona Supercomputing Center 27
- BlueArc 20, 22
- CAPS entreprise 26
- Cilk Arts 26
- Convex Computer 8
- Convey Computer 14
- Corensic 27
- Cray 10, 14, 18, 19, 27, 29, 30
- DataDirect Networks 20
- Defense Advanced Research Projects Agency 32, 33
- Dell 30
- e-Infrastructure Reflection Group 30, 31
- ET International 26
- European Commission 3, 29, 30, 32
- European Exascale Software Initiative 32
- European Grid Initiative 30
- European Open File System Cooperative 22
- EXTOLL 19
- Fraunhofer Institute 27
- Fujitsu 8, 35
- Fusion-io 17
- Gnodal 19
- Heidelberg University 19
- HP 8, 17, 30
- IBM 10, 14, 15, 20, 21, 26, 27, 30, 32, 36
- IBTA 18
- IESP 32, 33, 34, 35
- InfiniBand Trade Association, The 18
- Intel 8, 12, 13, 15, 17, 26, 29, 30, 35
- International Exascale Software Project 33

Jülich Research Centre, Germany 15

Khronos Group 26

Lawrence Berkeley National Laboratory 33

Lawrence Livermore National Laboratory 33

Los Alamos National Laboratory 33

Maxeler Technologies 13, 14

Meiko 19

Mellanox 18

Micron Technology 16, 17

Moscow State University 35

National Aeronautics and Space Administration 32

National Center for Supercomputing Applications 15

National Science Foundation 32, 34

Nimbus Data Systems 17

NUDT 35

NVIDIA 12, 14, 24, 26, 30, 35

Oak Ridge National Laboratory 10, 33

Oracle 21, 30

Pacific Northwest National Laboratory 33

Pallas 26

Panasas 20, 22

Partnership for Advanced Computing in Europe 30, 31, 32

PathScale 26

PlanetHPC 30

Portland Group, The 26

PROSPECT e.V. 29, 30

QLogic 18

Quadrics 19

Quantum 20

RapidMind 26

RIKEN Research Institute 8, 35

Rogue Wave Software 27

Samsung 17

Sandia National Laboratories 33

SanDisk 17

Scapos 27

SGI 8

STMicroelectronics 26

Supermicro 17

Texas Memory Systems 17

Toshiba 17

T-Platforms 35

University of Delaware 26

University of Illinois 15

University of Leuven 15

University of Versailles 15

US Department of Defense 32

US Department of Energy 11, 20, 22,
26, 32, 33, 34

Viking Modular 17

Violin Memory 17

Virident Systems 17

Whamcloud 22, 23

Xyratex 21, 22, 23