Scientific discovery and engineering innovation requires unifying traditionally separated high-performance computing and big data analytics.

BY DANIEL A. REED AND JACK DONGARRA

Exascale Computing and Big Data

NEARLY TWO CENTURIES ago, the English chemist Humphrey Davy wrote "Nothing tends so much to the advancement of knowledge as the application of a new instrument. The native intellectual powers of men in different times are not so much the causes of the different success of their labors, as the peculiar nature of the means and artificial resources in their possession." Davy's observation that advantage accrues to those who have the most powerful scientific tools is no less true today. In 2013, Martin Karplus, Michael Levitt, and Arieh Warshel received the Nobel Prize in chemistry for their work in computational modeling. The Nobel committee said, "Computer models mirroring real life have become crucial for most advances made in chemistry today,"[17] and "Computers unveil chemical processes, such as a catalyst's purification of exhaust fumes or the photosynthesis in green leaves."

Whether describing the advantages of high-energy particle accelerators (such as the Large Hadron Collider

and the 2013 discovery of the Higgs boson), powerful astronomy instruments (such as the Hubble Space Telescope, which yielded insights into the universe's expansion and dark energy), or high-throughput DNA sequencers and exploration of metagenomics ecology, ever-more powerful scientific instruments continually advance knowledge. Each such scientific instrument, as well as a host of others, is critically dependent on computing for sensor control, data processing, international collaboration, and access.

However, computing is much more than an augmenter of science. Unlike other tools, which are limited to particular scientific domains, computational modeling and data analytics are applicable to all areas of science and engineering, as they breathe life into the underlying mathematics of scientific models. They enable researchers to understand nuanced predictions, as well as shape experiments more efficiently. They also help capture and analyze the torrent of experimental data being produced by a new generation of scientific instruments made possible by advances in computing and microelectronics.

Computational modeling can illuminate the subtleties of complex mathematical models and advance science and engineering where time, cost, or safety precludes experimental assessment alone. Computational models of astrophysical phenomena, on temporal and spatial scales as di-

» key insights

- The tools and cultures of high-performance computing and big data analytics have diverged, to the detriment of both; unification is essential to address a spectrum of major research domains.

- The challenges of scale tax our ability to transmit data, compute complicated functions on that data, or store a substantial part of it; new approaches are required to meet these challenges.

- The international nature of science demands further development of advanced computer architectures and global standards for processing data, even as international competition complicates the openness of the scientific process.
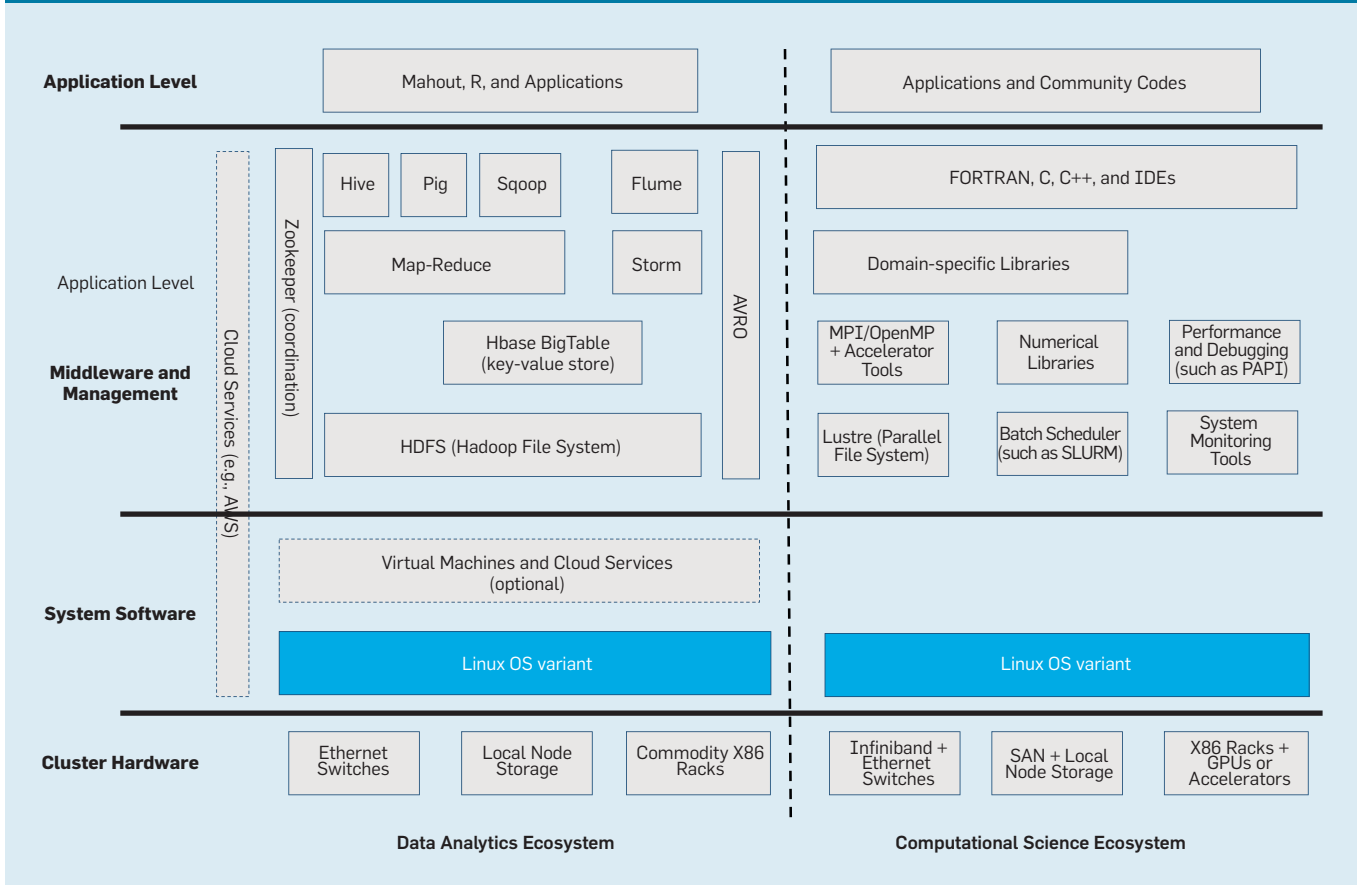
Figure 1. Data analytics and computing ecosystem compared.

Level	Data Analytics Ecosystem	Computational Science Ecosystem
Application Level	Mahout, R, and Applications	Applications and Community Codes
Middleware and Management	Zookeeper (coordination); Hive, Pig, Sqoop, Flume, Map-Reduce, Storm, AVRO, Hbase BigTable (key-value store), HDFS (Hadoop File System); Cloud Services (e.g., AWS)	FORTRAN, C, C++, and IDEs; Domain-specific Libraries; MPI/OpenMP + Accelerator Tools; Numerical Libraries; Performance and Debugging (such as PAPI); Lustre (Parallel File System); Batch Scheduler (such as SLURM); System Monitoring Tools
System Software	Virtual Machines and Cloud Services (optional); Linux OS variant	Linux OS variant
Cluster Hardware	Ethernet Switches; Local Node Storage; Commodity X86 Racks	Infiniband + Ethernet Switches; SAN + Local Node Storage; X86 Racks + GPUs or Accelerators

verse as planetary system formation, stellar dynamics, black hole behavior, galactic formation, and the interplay of baryonic and putative dark matter, have provided new insights into theories and complemented experimental data. Sophisticated climate models that capture the effects of greenhouse gases, deforestation, and other planetary changes have been key to understanding the effects of human behavior on the weather and climate change.

Computational science and engineering also enable multidisciplinary design and optimization, reducing prototyping time and costs. Advanced simulation has enabled Cummins to build better diesel engines faster and less expensively, Goodyear to design safer tires much more quickly, Boeing to build more fuel-efficient aircraft, and Procter & Gamble to create better materials for home products.

Similarly, "big data," machine learning, and predictive data analytics have been hailed as the fourth paradigm of science,[12] allowing researchers to extract insights from both scientific instruments and computational simulations. Machine learning has yielded new insights into health risks and the spread of disease via analysis of social networks, Web-search queries, and hospital data. It is also key to event identification and correlation in domains as diverse as high-energy physics and molecular biology.

As with successive generations of other large-scale scientific instruments, each new generation of advanced computing brings new capabilities, along with technical design challenges and economic trade-offs. Broadly speaking, data-generation capabilities in most science domains are growing more rapidly than compute capabilities, causing these domains to become data-intensive.[23] High-performance computers and big-data systems are tied inextricably to the broader computing ecosystem and its designs and markets. They also support national-security needs and economic competitiveness in ways that distinguish them from most other scientific instruments.

This "dual use" model, together with the rising cost of ever-larger computing and data-analysis systems, along with a host of new design challenges at massive scale, are raising new questions about advanced computing research investment priorities, design, and procurement models, as well as global collaboration and competition. This article examines some of these technical challenges, the interdependence of computational modeling and data analytics, and the global ecosystem and competition for leadership in advanced computing. We begin with a primer on the history of advanced computing.

Advanced Computing Ecosystems
By definition, an advanced computing system embodies the hardware, software, and algorithms needed to deliver the very highest capability at any given time. As in Figure 1, the computing and data analytics ecosystems share some attributes, notably reliance on open source software and the x86 hardware ecosystem. However, they differ markedly in their foci and technical approaches. As scientific research increasingly depends on both high-speed

computing and data analytics, the potential interoperability and scaling convergence of these two ecosystems is crucial to the future.

Scientific computing. In the 1980s, vector supercomputing dominated high-performance computing, as embodied in the eponymously named systems designed by the late Seymour Cray. The 1990s saw the rise of massively parallel processing (MPPs) and shared memory multiprocessors (SMPs) built by Thinking Machines, Silicon Graphics, and others. In turn, clusters of commodity (Intel/AMD x86) and purpose-built processors (such as IBM's BlueGene), dominated the previous decade.

Today, these clusters are augmented with computational accelerators in the form of coprocessors from Intel and graphical processing units (GPUs) from Nvidia; they also include high-speed, low-latency interconnects (such as Infiniband). Storage area networks (SANs) are used for persistent data storage, with local disks on each node used only for temporary files. This hardware ecosystem is optimized for performance first, rather than for minimal cost.

Atop the cluster hardware, Linux provides system services, augmented with parallel file systems (such as Lustre) and batch schedulers (such as PBS and SLURM) for parallel job management. MPI and OpenMP are used for internode and intranode parallelism, augmented with libraries and tools (such as CUDA and OpenCL) for coprocessor use. Numerical libraries (such as LAPACK and PETSc) and domain-specific libraries complete the software stack. Applications are typically developed in FORTRAN, C, or C++.

Data analytics. Just a few years ago, the very largest data storage systems contained only a few terabytes of secondary disk storage, backed by automated tape libraries. Today, commercial and research cloud-computing systems each contain many petabytes of secondary storage, and individual research laboratories routinely process terabytes of data produced by their own scientific instruments.

As with high-performance computing, a rich ecosystem of hardware and software has emerged for big-data analytics. Unlike scientific-computing clusters, data-analytics clusters are typically based on commodity Ethernet networks and local storage, with cost

and capacity the primary optimization criteria. However, industry is now turning to FPGAs and improved network designs to optimize performance.

Atop this hardware, the Apache Hadoop[25] system implements a MapReduce model for data analytics. Hadoop includes a distributed file system (HDFS) for managing large numbers of large files, distributed (with block replication) across the local storage of the cluster. HDFS and HBase, an open-source implementation of Google's BigTable key-value store,[3] are the big-data analogs of Lustre for computational science, albeit optimized for different hardware and access patterns.

Atop the Hadoop storage system, tools (such as Pig[18]) provide a high-level programming model for the two-phase MapReduce model. Coupled with streaming data (Storm and Flume), graph (Giraph), and relational data (Sqoop) support, the Hadoop ecosystem is designed for data analysis. Moreover, tools (such as Mahout) enable classification, recommendation, and prediction via supervised and unsupervised learning. Unlike scientific computing, application development for data analytics often relies on Java and Web services tools (such as Ruby on Rails).

Scaling Challenges

Given the rapid pace of technological change, leading-edge capability is a moving target. Today's smartphone computes as fast as yesterday's supercomputer, and today's personal music

collection is as large as yesterday's enterprise-scale storage.

Lest this seem an exaggeration, the measured performance of an Apple iPhone 6 or Samsung Galaxy S5 on standard linear algebra benchmarks now substantially exceeds that of a Cray-1, which is widely viewed as the first successful supercomputer. That same smartphone has storage capacity rivaling the text-based content of a major research library.

Just a few years ago, teraflops (10^{12} floating point operations/second) and terabytes (10^{12} bytes of secondary storage) defined state-of-the-art advanced computing. Today, those same values represent a desk-side PC with Nvidia or Intel Xeon Phi accelerator and local storage. Advanced computing is now defined by multiple petaflops (10^{15} floating operations/second) supercomputing systems and cloud data centers with many petabytes of secondary storage.

Figure 2 outlines this exponential increase in advanced computing capability, based on the widely used High-Performance LINPACK (HPL) benchmark[6] and Top500 list of the world's fastest computers.[16] Although solution of dense linear systems of equations is no longer the best measure of delivered performance on complex scientific and engineering applications, this historical data illustrates how rapidly high-performance computing has evolved. Though high-performance computing has benefited from the same semiconductor advances as commodity com-

Figure 2. Advanced computing performance measured by the HPL benchmark.
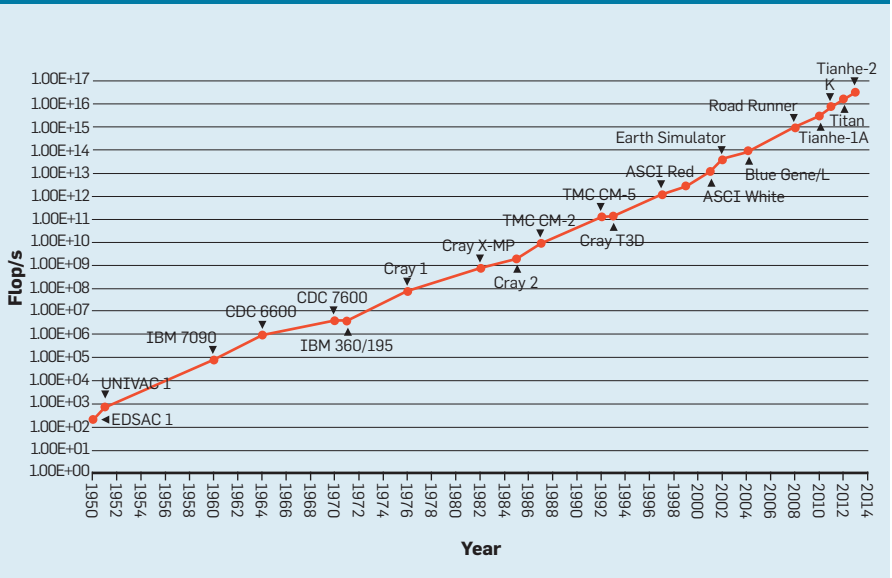
puting, sustained system performance has improved even more rapidly due to increasing system size and parallelism.

The growth of personal, business, government, and scientific data has been even more dramatic and well documented. Commercial cloud providers are building worldwide networks of data centers, each costing hundreds of millions of dollars, to support Web search engines, social networks, and cloud services. Concurrently, the volume of scientific data produced annually now challenges the budgets of national research agencies.

As an example, Figure 3 outlines the exponential growth in the number of objects stored in Amazon's Simple Storage Service (S3). Atop such low-level services, companies (such as Netflix) implement advanced recommender systems to suggest movies to subscribers and then stream selections. Scientific researchers also increasingly explore these same cloud services and machine-learning techniques for extracting insight from scientific images, graphs, and text data. There are natural technical and economic synergies among the challenges facing data-intensive science and exascale computing, and advances in both are necessary for future scientific breakthroughs. Data-intensive science relies on the collection, analysis, and management of massive volumes of data, whether obtained from scientific simulations or experimental facilities. In each case, national and international investment in "extreme scale" systems will be necessary to analyze the massive volumes of data that are now commonplace in science and engineering.

Race to the future. For scientific and engineering computing, exascale (10^{18} operations per second) is the next proxy in the long trajectory of exponential performance increases that has continued for more than half a century. Likewise, large-scale data preservation and sustainability within and across disciplines, metadata creation and multidisciplinary fusion, and digital privacy and security define the frontiers of big data. This multifaceted definition of advanced computing encompasses more than simply quantitative measures of sustained arithmetic operation rates or storage capacity and analysis rates; it is also a relative term encompassing qualitative improvements in the usable capabilities of advanced computing systems at all scales. As such, it is intended to suggest a new frontier of practical, delivered capability to scientific and engineering researchers across all disciplines.

However, there are many challenges on the road to ever more advanced computing, including, but not limited to, system power consumption and environmentally friendly cooling, massive parallelism, and component failures, data and transaction consistency, metadata and ontology management, precision and recall at scale, and multidisciplinary data fusion and preservation.

Above all, advanced computing systems must not become so arcane and complex that they and their services are unusable by all but a handful of experts. Open source toolkits (such as Hadoop, Mahout, and Giraph), along with a growing set of domain-specific tools and languages, have allowed many research groups to apply machine learning to large-scale scientific data without deep knowledge of machine-learning algorithms. The same is true of community codes for computational science modeling.

Hardware, software, data, and politics. Historically, high-performance computing advances have been largely dependent on concurrent advances in algorithms, software, architecture, and hardware that enable higher levels of floating-point performance for computational models. Advances today are also shaped by data-analysis pipelines, data architectures, and machine-learning tools that manage large volumes of scientific and engineering data.

However, just as changes in scientific instrumentation scale bring new opportunities, they also bring new challenges, some technical but others organizational, cultural, and economic, and they are not self-similar across scales. Today, exascale computing systems cannot be produced in an affordable and reliable way or be subject to realistic engineering constraints on capital and operating costs, usability, and reliability. As the costs of advanced computing and data-analysis systems, whether commercial or scientific, have moved from millions to billions of dollars, design and decision processes have necessarily become more complex and fraught with controversy. This is a familiar lesson to those in high-energy physics and astronomy, where particle accelerators and telescopes have become planetary-scale instrumentation and the province of international consortia and global politics. Advanced computing is no exception.

The research-and-development costs to create an exascale computing system have been estimated by many experts to exceed one billion U.S. dollars, with an annual operating cost of tens of millions of dollars. Concurrently, there is growing recognition that governments and research agencies have substantially underinvested in data retention and management, as evinced by multi-billion-dollar private-sector investments in big data and cloud computing. The largest commercial cloud data centers each cost more than $500 million to construct, and Google, Amazon, Microsoft, Facebook, and other companies operate global networks of such centers.

Figure 3. Growth of Amazon S3 objects.
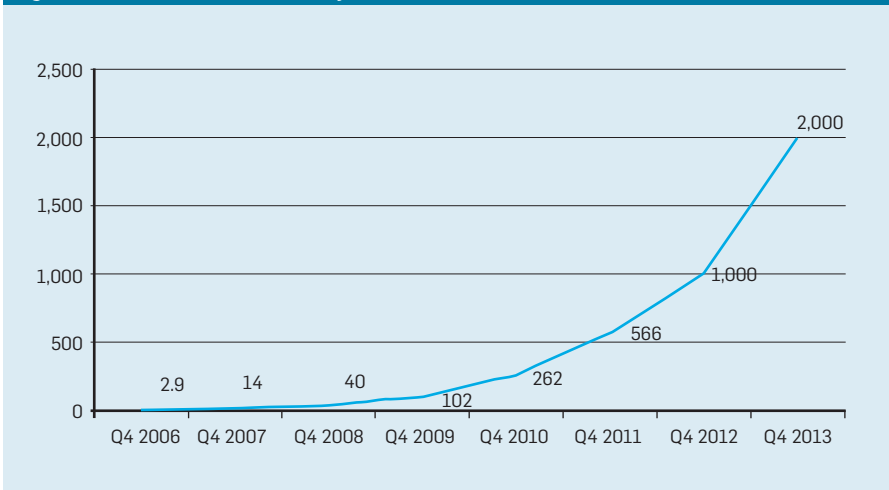
Against this backdrop, U.S. support for basic research is at a decadal low, when adjusted for inflation,[2] and both the U.S. and the European Union continue to experience weak recoveries from the economic downturn of 2008. Further exacerbating the challenges, the global race for advanced computing hegemony is convolved with national-security desires, economic competitiveness, and the future of the mainstream computing ecosystem.

The shift from personal computers to mobile devices has also further raised competition between the U.S.-dominated x86 architectural ecosystem and the globally licensed ARM ecosystem. Concurrently, concerns about national sovereignty, data security, and Internet governance have triggered new competition and political concerns around data services and cloud-computing operations.

Despite these challenges, there is reason for cautious optimism. Every advance in computing technology has driven industry innovation and economic growth, spanning the entire spectrum of computing, from the emerging Internet of Things to ubiquitous mobile devices to the world's most powerful computing systems and largest data archives. These advances have also spurred basic and applied research in every domain of science.

Solving the myriad technical, political, and economic challenges will be neither easy nor even possible by tackling them in isolation. Rather, it will require coordinated planning across government, industry, and academia, commitment to data sharing and sustainability, collaborative research and development, and recognition that both competition and collaboration will be necessary for success. The future of big data and analytics should not be pitted against exascale computing; both are critical to the future of advanced computing and scientific discovery.

Scientific and Engineering Opportunities

Researchers in the physical sciences and engineering have long been major users of advanced computing and computational models. The more recent adoption by the biological, environmental, and social sciences has been driven in part by the rise of big-data analytics. In addition,

Computing technology is poised at important inflection points, at the very largest scale, or leading-edge high-performance computing, and the very smallest scale, or semiconductor processes.

advanced computing is now widely used in engineering and advanced manufacturing. From understanding the subtleties of airflow in turbomachinery to chemical molecular dynamics for consumer products to biomass feedstock modeling for fuel cells, advanced computing has become synonymous with multidisciplinary design and optimization and advanced manufacturing.

Looking forward, only a few examples are needed to illustrate the deep and diverse scientific and engineering benefits from advanced computing:

Biology and biomedicine. Biology and biomedicine have been transformed through access to large volumes of genetic data. Inexpensive, high-throughput genetic sequencers have enabled capture of organism DNA sequences and made possible genome-wide association studies for human disease and human microbiome investigations, as well as metagenomics environmental studies. More generally, biological and biomedical challenges span sequence annotation and comparison, protein-structure prediction; molecular simulations and protein machines; metabolic pathways and regulatory networks; whole-cell models and organs; and organisms, environments, and ecologies;

High-energy physics. High-energy physics is both computational- and data-intensive. First-principles computational models of quantum chromodynamics provide numerical estimates and validations of the Standard Model. Similarly, particle detectors require the measurement of probabilities of "interesting" events in large numbers of observations (such as in 10^{16} or more particle collisions observed in a year). The Large Hadron Collider and its experiments necessitated creating a worldwide computing grid for data sharing and reduction, driving deployment of advanced networks and protocols, as well as a hierarchy of data repositories. All were necessary to identify the long-sought Higgs boson;

Climate science. Climate science is also critically dependent on the availability of a reliable infrastructure for managing and accessing large heterogeneous quantities of data on a global scale. It is inherently a collaborative and multidisciplinary effort requiring sophisticated modeling of the physical processes and exchange mechanisms

among multiple Earth realms—atmosphere, land, ocean, and sea ice—and comparing and validating these simulations with observational data from various sources, all collected over long periods. To encourage exploration, NASA has made climate and Earth science satellite data available through Amazon Web Services;

Cosmology and astrophysics. Cosmology and astrophysics are now critically dependent on advanced computational models to understand stellar structure, planetary formation, galactic evolution, and other interactions. These models combine fluid processes, radiation transfer, Newtonian gravity, nuclear physics, and general relativity (among other processes). Underlying them is a rich set of computational techniques based on adaptive mesh refinement and particle-in-cell, multipole algorithms, Monte Carlo methods, and smoothed-particle hydrodynamics;

Astronomy. Complementing computation, whole-sky surveys, and a new generation of automated telescopes are providing new insights. Rather than capture observational data to answer a known question, astronomers now frequently query extant datasets to discover previously unknown patterns and trends. Big-data reduction and unsupervised learning systems are an essential part of this exploratory image analysis;

Cancer treatment. Effective cancer treatment depends on early detection and targeted treatments via surgery, radiation, and chemotherapy. In turn, tumor identification and treatment planning are dependent on image enhancement, feature extraction and classification, segmentation, registration, 3D reconstruction, and quantification. These and other machine-learning techniques provide not only diagnostic validation, they are increasingly used to conduct comparative and longitudinal analysis of treatment regimes;

Experimental and computational materials science. Experimental and computational materials science is key to understanding materials properties and engineering options; for example, neutron scattering allows researchers to understand the structure and properties of materials, macromolecular and biological systems, and the fundamental physics of the

It is important for all of computer science to design algorithms that communicate as little as possible, ideally attaining lower bounds on the amount of communication required.

neutron by providing data on the internal structure of materials from the atomic scale (atomic positions and excitations) up to the mesoscale (such as the effects of stress);

Steel production. Steel production via continuous casting accounts for an important fraction of global energy consumption and greenhouse gases production. Even small improvements to this process would have profound societal benefits and save hundreds of millions of dollars. High-performance computers are used to improve understanding of this complex process via comprehensive computational models, as well as to apply those models to find operating conditions to improve the process; and

Text and data mining. The explosive growth of research publications has made finding and tracking relevant research increasingly difficult. Beyond the volume of text, principles have different or similar names across domains. Text classification, semantic graph visualization tools, and recommender systems are increasingly being used to identify relevant topics and suggest relevant papers for study.

There are two common themes across these science and engineering challenges. The first is an extremely wide range of temporal and spatial scales and complex, nonlinear interactions across multiple biological and physical processes. These are the most demanding of computational simulations, requiring collaborative research teams, along with the very largest and most capable computing systems. In each case, the goal is predictive simulation, or gleaning insight that tests theories, identifies subtle interactions, and guides new research.

The second theme is the enormous scale and diversity of scientific data and the unprecedented opportunities for data assimilation, multidisciplinary correlation, and statistical analysis. Whether in biological or physical sciences, engineering or business, big data is creating new research needs and opportunities.

Technical Challenges in Advanced Computing

The scientific and engineering opportunities made possible through advanced computing and data analytics

are deep, but the technical challenges in designing, constructing, and operating advanced computing and data-analysis systems of unprecedented scale are just as daunting. Although cloud-computing centers and exascale computational platforms are seemingly quite different, as discussed earlier, the underlying technical challenges of scale are similar, and many of the same companies and researchers are exploring dual-use technologies applicable to both.

In a series of studies over the past five years, the U.S. Department of Energy identified 10 research challenges[10,15,24] in developing a new generation of advanced computing systems, including the following, augmented with our own comparisons with cloud computing:

Energy-efficient circuit, power, and cooling technologies. With current semiconductor technologies, all proposed exascale designs would consume hundreds of megawatts of power. New designs and technologies are needed to reduce this energy requirement to a more manageable and economically feasible level (such as 20MW–40MW) comparable to that used by commercial cloud data centers;

High-performance interconnect technologies. In the exascale-computing regime, the energy cost to move a datum will exceed the cost of a floating-point operation, necessitating very energy efficient, low-latency, high-bandwidth interconnects for fine-grain data exchanges among hundreds of thousands of processors. Even with such designs, locality-aware algorithms and software will be needed to maximize computation performance and reduce energy needs;

Driven by cost necessity, commercial cloud computing systems have been built with commodity Ethernet interconnects and adopted a bulk synchronous parallel computation model. Although this approach has proven effective, as evidenced by widespread adoption of MapReduce toolkits (such as Hadoop), a lower-cost, convergence interconnect would benefit both computation and data-intensive platforms and open new possibilities for fine-grain data analysis.

Advanced memory technologies to improve capacity. Minimizing data movement and minimizing energy use are also dependent on new memory technologies, including processor-in-memory, stacked memory (Micron's HMC is an early example), and non-volatile memory approaches. Although the particulars differ for computation and data analysis, algorithmic determinants of memory capacity will be a significant driver of overall system cost, as the memory per core for very large systems will necessarily be smaller than in current designs;

Scalable system software that is power and failure aware. Traditional high-performance computing software has been predicated on the assumption that failures are infrequent; as we approach exascale levels, systemic resilience in the face of regular component failures will be essential. Similarly, dynamic, adaptive energy management must become an integral part of system software, for both economic and technical reasons.

Cloud services for data analytics, given their commercial quality-of-service agreements, embody large numbers of resilience techniques, including geo-distribution, automatic restart and failover, failure injection, and introspective monitoring; the Netflix "Simian Army"[a] is illustrative of these techniques.

Data management software that can handle the volume, velocity, and diversity of data. Whether computationally generated or captured from scientific instruments, efficient in situ data analysis requires restructuring of scientific workflows and applications, building on lessons gleaned from commercial data-analysis pipelines, as well as new techniques for data coordinating, learning, and mining. Without them, I/O bottlenecks will limit system utility and applicability;

Programming models to express massive parallelism, data locality, and resilience. The widely used communicating sequential process model, or MPI programming, places the burden of locality and parallelization on application developers. Exascale computing systems will have billion-way parallelism and frequent faults. Needed are more expressive programming models able to deal with this behavior and simplify

the developer's efforts while supporting dynamic, fine-grain parallelism.

Much can be learned from Web and cloud services where abstraction layers and domain-specific toolkits allow developers to deploy custom execution environments (virtual machines) and leverage high-level services for reduction of complex data. The scientific computing challenge is retaining expressivity and productivity while also delivering high performance.

Reformulation of science problems and refactoring solution algorithms. Many thousands of person-years have been invested in current scientific and engineering codes and in data mining and learning software. Adapting scientific codes to billion-way parallelism will require redesigning, or even reinventing, the algorithms and potentially reformulating the science problems. Integrating data-analytics software and tools with computation is equally daunting; programming languages and models differ, as do the communities and cultures. Understanding how to do these things efficiently and effectively will be key to solving mission-critical science problems;

Ensuring correctness in the face of faults, reproducibility, and algorithm verification. With frequent transient and permanent faults, lack of reproducibility in collective communication, and new mathematical algorithms with limited verification, computation validation and correctness assurance will be much more important for the next generation of massively parallel systems, whether optimized for scientific computing, data analysis, or both;

Mathematical optimization and uncertainty quantification for discovery, design, and decision. Large-scale computations are themselves experiments that probe the sample space of numerical models. Understanding the sensitivity of computational predictions to model inputs and assumptions, particularly when they involve complex, multidisciplinary applications requires new tools and techniques for application validation and assessment. The equally important analogs in large-scale data analytics and machine learning are precision (the fraction of retrieved data that is relevant) and recall (the fraction of relevant data retrieved); and

---

a http://techblog.netflix.com/2011/07/netflix-simian-army.html

Software engineering and supporting structures to enable productivity. Although programming tools, compilers, debuggers, and performance-enhancement tools shape research productivity for all computing systems, at scale, application design and management for reliable, efficient, and correct computation is especially daunting. Unless researcher productivity increases, the time to solution may be dominated by application development, not computation.

Similar hardware and software studies[1,14] chartered by the U.S. Defense Advanced Research Projects Agency identified the following challenges, most similar to those cited by the Department of Energy studies:

Energy-efficient operation. Energy-efficient operation to achieve desired computation rates subject to overall power dissipation;

Memory capacity. Primary and secondary memory capacity and access rates, subject to power constraints;

Concurrency and locality. Concurrency and locality to meet performance targets while allowing some threads to stall during long-latency operations;

Resilience. Resilience, given large component counts, shrinking silicon feature sizes, low-power operation, and transient and permanent component failures;

Application scaling. Application scaling subject to memory capacity and communication latency constraints;

Managing parallelism. Expressing and managing parallelism and locality in system software and portable programming models, including runtime systems, schedulers, and libraries; and

Software tools. Software tools for performance tuning, correctness assessment, and energy management.

Moreover, a 2011 study by the U.S. National Academy of Sciences (NAS)[9] suggested that, barring a breakthrough, the exponential increases in performance derived from shrinking semiconductor feature size and architectural innovation are nearing an end. This study, along with others, suggests computing technology is poised at important inflection points, at the very largest scale, or leading-edge high-performance computing, and the very smallest scale, or semiconductor processes. The computing community remains divided on possible approaches, with strong believers that technical obstacles limiting extension of current approaches will be overcome and others who believe more radical technology and design approaches (such as quantum and superconducting devices) may be required.

Hardware and architecture challenges. Although a complete description of the hardware and software technical challenges just outlined is beyond the scope of this article, review of a selected subset is useful to illuminate the depth and breadth of the problems and their implications for the future of both advanced computing and the broader deployment of next-generation consumer- and business-computing technologies.

Post-Dennard scaling. For decades, Moore's "law" has held true due to the hard work and creativity of a great many people, as well as many billions of dollars of investment in process technology and silicon foundries. It has also rested on the principle of Dennard scaling,[5,13] providing a recipe for shrinking transistors and yielding smaller circuits with the same power density. Decreasing a transistor's linear size by a factor of two thus reduced power by a factor of four, or with both voltage and current halving.

Although transistor size continues to shrink, with 22-nanometer feature size now common, transistor power consumption no longer decreases accordingly. This has led to limits on chip clock rates and power consumption, along with design of multicore chips and the rise of dark silicon—chips with more transistors than can be active simultaneously due to thermal and power constraints.[8]

These semiconductor challenges have stimulated a rethinking of chip design, where the potential performance advantage of architectural tricks—superpipelining, scoreboarding, vectorization, and parallelization—must be balanced against their energy consumption. Simpler designs and function-specific accelerators often yield a better balance of power consumption and performance. This architectural shift will be especially true if the balance of integer, branch, and floating-point operations shifts to support in situ data analysis and computing.

Chip power limits have stimulated great interest in the ARM processor ecosystem. Because ARM designs were optimized for embedded and mobile devices, where limited power consumption has long been a design driver, they have simpler pipelines and instruction decoders than x86 designs.

In this new world, hardware/software co-design becomes de rigeur, with devices and software systems interdependent. The implications are far fewer general-purpose performance increases, more hardware diversity, elevation of multivariate optimization (such as power, performance, and reliability) in programming models, and new system-software-resource-management challenges.

Resilience and energy efficiency at scale. As advanced computing and data analysis systems grow ever larger, the assumption of fully reliable operation becomes much less credible. Although the mean time before failure for individual components continues to increase incrementally, the large overall component count for these systems means the systems themselves will fail more frequently. To date, experience has shown failures can be managed but only with improved techniques for detecting and understanding component failures.

Data from commercial cloud data centers suggests some long-held assumptions about component failures and lifetimes are incorrect.[11,20–22] A 2009 Google study[22] showed DRAM error rates were orders-of-magnitude higher than previously reported, with over 8% of DIMMs affected by errors in a year. Equally surprising, these were hard errors, rather than soft, correctable (via error-correcting code) errors.

In addition to resilience, scale also brings new challenges in energy management and thermal dissipation. Today's advanced computing and data-analysis systems consume megawatts of power, and cooling capability and peak power loads limit where many systems can be placed geographically. As commercial cloud operators have learned, energy infrastructure and power are a substantial fraction of total system cost at scale, necessitating new infrastructure approaches and operating models, including low-power designs,

cooling approaches, energy accountability, and operational efficiencies.

Software and algorithmic challenges. Many of the software and algorithmic challenges for advanced computing and big-data analytics are themselves consequences of extreme system scale. As noted earlier, advanced scientific computing shares many of the scaling problems of Web and cloud services but differs in its price-performance optimization balance, emphasizing high levels of performance, whether for computation or for data analysis. This distinction is central to the design choices and optimization criteria.

Given the scale and expected error rates of exascale computing systems, design and implementation of algorithms must be rethought from first principles, including exploration of global synchronization-free (or at least minimal) algorithms, fault-oblivious and error-tolerant algorithms, architecture-aware algorithms suitable for heterogeneous and hierarchical organized hardware, support for mixed-precision arithmetic, and software for energy-efficient computing.

Locality and scale. As noted earlier, putative designs for extreme-scale computing systems are projected to require billion-way computational concurrency, with aggressive parallelism at all system levels. Maintaining load balance on all levels of a hierarchy of algorithms and platforms will be the key to efficient execution. This will likely require dynamic, adaptive runtime mechanisms[4] and self-aware resource allocation to tolerate not only algorithmic imbalances but also variability in hardware performance and reliability.

In turn, the energy costs and latencies for communication will place an even greater premium on computation locality than today. Inverting long-held models, arithmetic operations will be far less energy intensive and more efficient than communication. Algorithmic complexity is usually expressed in terms of number of operations performed rather than quantity of data movement to memory. This is directly opposed to the expected costs of computation at large scale, where memory movement will be very expensive and operations will be nearly free, an issue of importance to both floating-point-intensive and data-analysis algorithms.

Programming models and tools are perhaps the biggest point of divergence between the scientific-computing and big-data ecosystems.

The temporal cost of data movement will challenge traditional algorithmic design approaches and comparative optimizations, making redundant computation sometimes preferable to data sharing and elevating communication complexity to parity with computation. It is therefore important for all of computer science to design algorithms that communicate as little as possible, ideally attaining lower bounds on the amount of communication required. It will also require models and methods to minimize and tolerate (hide) latency, optimize data motion, and remove global synchronization.

Adaptive system software. Resource management for today's high-performance computing systems remains rooted in a deus ex machina model, with coordinated scheduling and tightly synchronized communication. However, extreme scale, hardware heterogeneity, system power, and heat-dissipation constraints and increased component failure rates influence not only the design and implementation of applications, they also influence the design of system software in areas as diverse as energy management and I/O. Similarly, as the volume of scientific data grows, it is unclear if the traditional file abstractions and parallel file systems used by technical computing will scale to trans-petascale data analysis.

Instead, new system-software and operating-system designs will need to support management of heterogeneous resources and non-cache-coherent memory hierarchies, provide applications and runtime with more control of task scheduling policies, and manage global namespaces. They must also expose mechanisms for finer measurement, prediction, and control of power management, allowing schedulers to map computations to function-specific accelerators and manage thermal envelopes and application energy profiles. Commercial cloud providers have already faced many of these problems, and their experience in large-scale resource management has much to offer the scientific computing ecosystem.

Parallel programming support. As the diversity, complexity, and scale of advanced computing hardware has increased, the complexity and difficulty of

developing applications has increased as well, with many operating functions now subsumed by applications. Application complexity has been further exacerbated by the increasingly multidisciplinary nature of applications that combine algorithms and models spanning a range of spatiotemporal scales and algorithmic approaches.

Consider the typical single-program multiple-data parallel-programming or bulk-synchronous parallel model, where application data is partitioned and distributed across the individual memories or disks of the computation nodes, and the nodes share data via network message passing. In turn, the application code on each node manages the local, multilevel computation hierarchy—typically multiple, multithreaded, possibly heterogeneous cores, and (often) a GPU accelerator—and coordinates I/O, manages application checkpointing, and oversees power budgets and thermal dissipation. This daunting level of complexity and detailed configuration and tuning makes developing robust applications an arcane art accessible to only a dedicated and capable few.

Ideally, future software design, development, and deployment will raise the abstraction level and include performance and correctness in mind at the outset rather than ex situ. Beyond more performance-aware design and development of applications based on integrated performance and correctness models, these tools must be integrated with compilers and runtime systems, provide more support for heterogeneous hardware and mixed programming models, and provide more sophisticated data processing and analysis.

Programming models and tools are perhaps the biggest point of divergence between the scientific-computing and big-data ecosystems. The latter emphasizes simple abstractions (such as key-value stores and MapReduce), along with semantics-rich data formats and high-level specifications. This has allowed many developers to create complex machine-learning applications with little knowledge of the underlying hardware or system software. In contrast, scientific computing has continued to rely largely on traditional languages and libraries.

New programming languages and models, beyond C and FORTRAN, will help. Given the applications software already in place for technical computing, a radical departure is not realistic. Programming features found in new languages (such as Chapel and X10) have already had an indirect effect on existing program models. Existing programming models (such as OpenMP)

have already benefited through recent extensions for, say, task parallelism, accelerators, and thread affinity.

Domain-specific languages (DSLs) are languages that specialize to a particular application domain, representing a means of extending the existing base-language by hosting DSL extensions. Embedded DSLs are a pragmatic way to exploit the sophisticated analysis and transformation capabilities of the compilers for standard languages. The developer writes the application using high-level primitives a compiler will transform into efficient low-level code to optimize the performance on the underlying platform.

Algorithmic and mathematics challenges. Exascale computing will put greater demand on algorithms in at least two areas: the need for increasing amounts of data locality to perform computations efficiently and the need to obtain much higher levels of fine-grain parallelism, as high-end systems support increasing numbers of compute threads. As a consequence, parallel algorithms must adapt to this environment, and new algorithms and implementations must be developed to exploit the computational capabilities of the new hardware.

Significant model development, algorithm redesign, and science-application reimplementation, supported by (an) exascale-appropriate programming model(s), will be required to exploit the power of exascale architectures. The transition from current sub-petascale and petascale computing to exascale computing will be at least as disruptive as the transition from vector to parallel computing in the 1990s.

Economic and Political Challenges
The technical challenges of advanced computing and big-data analytics are shaped by other elements of the broader computing landscape. In particular, powerful smartphones and cloud computing services are rapidly displacing the PC and local servers as the computing standard. This shift has also triggered international competition for industrial and business advantage, with countries and regions investing in new technologies and system deployments.

Computing ecosystem shifts. The Internet and Web-services revolution is global, and U.S. influence, though

substantial, is waning. Notwithstanding Apple's phenomenal success, most smartphones and tablets are now designed, built, and purchased globally, and the annual sales volume of smartphones and tablets exceeds that of PCs and servers.

This ongoing shift in consumer preferences and markets is accompanied by another technology shift. Smartphones and tablets are based on energy-efficient microprocessors—a key component of proposed exascale computing designs—and systems-on-a-chip (SoCs) using the ARM architecture. Unlike Intel and AMD, which design and manufacture the x86 chips found in today's PCs and most leading-edge servers and HPC systems, ARM does not manufacture its own chips. Rather, it licenses its designs to others, who incorporate the ARM architecture into custom SoCs that are manufactured by global semiconductor foundries like Taiwan's TSMC.

International exascale projects. The international competition surrounding advanced computing mixes concern about economic competitiveness, shifting technology ecosystems (such as ARM and x86), business and technical computing (such as cloud computing services and data centers), and scientific and engineering research. The European Union, Japan, China, and U.S. have all launched exascale computing projects, each with differing emphasis on hardware technologies, system software, algorithms, and applications.

European Union. The European Union (EU) announced the start of its exascale research program in October 2011 with €25 million in funding for three complementary research projects in its Framework 7 effort. The Collaborative Research into Exascale Systemware, Tools and Applications (CRESTA), Dynamical Exascale Entry Platform (DEEP), and Mont-Blanc projects will each investigate different exascale challenges using a co-design model spanning hardware, system software, and software applications. This initiative represents Europe's first sustained investment in exascale research.

CRESTA brings together four European high-performance computing centers: Edinburgh Parallel Computing Centre (project lead), the High Per-

formance Computing Center Stuttgart, Finland's IT Center for Science Ltd., and Partner Development Center Sweden, as well as the Dresden University of Technology, which will lend expertise in performance optimization. In addition, the CRESTA team also includes application professionals from European science and industry, as well as HPC vendors, including HPC tool developer Allinea and HPC vendor Cray. CRESTA focuses on the use of applications as co-design drivers for software development environments, algorithms and libraries, user tools, and underpinning and crosscutting technologies.

The Mont-Blanc project, led by the Barcelona Supercomputing Center, brings together European technology providers ARM, Bull, Gnodal, and major supercomputing organizations involved with the Partnership for Advanced Computing in Europe (PRACE) project, including Juelich, Leibniz-Rechenzentrum, or LRZ, GENCI, and CINECA. The project intends to deploy a first-generation HPC system built from energy-efficient embedded technologies and conduct the research necessary to achieve exascale performance with energy-efficient designs.

DEEP, led by Forschungszentrum Juelich, seeks to develop an exascale-enabling platform and optimization of a set of grand-challenge codes. The system is based on a commodity cluster and accelerator design—Cluster Booster Architecture—as a proof-of-concept for a 100 petaflop/s PRACE production system. In addition to the lead partner, Juelich, project partners include Intel, ParTec, LRZ, Universität Heidelberg, German Research School for Simulation Sciences, Eurotech, Barcelona Supercomputing Center, Mellanox, École Polytechnique Fédérale de Lausanne, Katholieke Universiteit Leuven, Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique, the Cyprus Institute, Universität Regensburg, CINECA, a consortium of 70 universities in Italy, and Compagnie Générale de Géophysique-Veritas.

Japan. In December 2013, the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) selected RIKEN to develop and deploy an exascale system by 2020. Selection was based on its expe-

rience developing and operating the K computer, which, at 10 petaflop/s, was ranked the fastest supercomputer in the world in 2011. Estimated to cost ¥140 billion ($1.38 billion), the exascale system design will be based on a combination of general-purpose processors and accelerators and involve three key Japanese computer vendors—Fujitsu, Hitachi, and NEC—as well as technical support from the University of Tokyo, University of Tsukuba, Tokyo Institute of Technology, Tohoku University, and RIKEN.

China. China's Tianhe-2 system is the world's fastest supercomputer today. It contains 16,000 nodes, each with two Intel Xeon processors and three Intel Xeon Phi coprocessors. It also contains a proprietary high-speed interconnect, called TH Express-2, designed by the National University for Defense Technology (NUDT). NUDT conducts research on processors, compilers, parallel algorithms, and systems. Based on this work, China is expected to produce a 100-petaflop/s systems in 2016 built entirely from Chinese-made chips, specifically the ShenWei processor, and interconnects. Tianhe-2 was to be upgraded from a peak of 55 petaflop/s to 100 petaflop/s in 2015, but the U.S. Department of Commerce has restricted exports of Intel processors to NUDT, the National Supercomputing Center in Changsha, National Supercomputing Center in Guangzhou, and the National Supercomputing Center in Tianjin due to national-security concerns.

U.S. Historically, the U.S. Networking and Information Technology Research and Development program has spanned several research missions and agencies, with primary leadership by the Department of Energy (DOE), Department of Defense (DoD), and National Science Foundation (NSF). DOE is today the most active deployer of high-performance computing systems and developer of plans for exascale computing. In contrast, NSF and DoD have focused more on broad cyberinfrastructure and enabling-technologies research, including research cloud services and big-data analytics. Although planning continues, the U.S. has not yet mounted an advanced computing initiative similar to those under way in Europe and Japan.

International collaboration. Although global competition for advanced computing and data-analytics leadership continues, there is active international collaboration. The International Exascale Software Project (IESP) is one such example in advanced computing. With seed funding from governments in Japan, the E.U., and the U.S., as well as supplemental contributions from industry stakeholders, IESP was formed to empower ultra-high-resolution and data-intensive science and engineering research through 2020.

In a series of meetings, the international IESP team developed a plan for a common, high-quality computational environment for petascale/exascale systems. The associated roadmap for software development would take the community from its current position to exascale computing.[7]

Conclusion

Computing is at a profound inflection point, economically and technically. The end of Dennard scaling and its implications for continuing semiconductor-design advances, the shift to mobile and cloud computing, the explosive growth of scientific, business, government, and consumer data and opportunities for data analytics and machine learning, and the continuing need for more-powerful computing systems to advance science and engineering are the context for the debate over the future of exascale computing and big data analysis. However, certain things are clear:

Big data and exascale. High-end data analytics (big data) and high-end computing (exascale) are both essential elements of an integrated computing research-and-development agenda; neither should be sacrificed or minimized to advance the other;

Algorithms, software, applications. Research and development of next-generation algorithms, software, and applications is as crucial as investment in semiconductor devices and hardware; historically the research community has underinvested in these areas;

Information technology ecosystem. The global information technology ecosystem is in flux, with the transition to a new generation of low-power mobile devices, cloud services, and rich data analytics; and

Private and global research. Private-sector competition and global-research collaboration are both necessary to address design, test, and deploy exascale-class computing and data-analysis capabilities.

There are great opportunities and great challenges in advanced computing, in both computation and data analysis. Scientific discovery via computational science and data analytics is truly the "endless frontier" about which Vannevar Bush spoke so eloquently in 1945. The challenges are for all of computer science to sustain the research, development, and deployment of the high-performance computing infrastructure needed to enable those discoveries.

Acknowledgments

[C]

References

1. Amarasinghe, S. et al. *Exascale Software Study: Software Challenges in Extreme-Scale Systems.* Defense Advanced Research Projects Agency, Arlington, VA, 2009; http://www.cs.rice.edu/~vs3/PDF/Sarkar-ACS-July-2011-v2.pdf
2. American Association for the Advancement of Science. *Guide to R&D Funding - Historical Data.* AAAS, Washington, D.C., 2015; http://www.aaas.org/page/historical-trends-federal-rd
3. Chang, F. et al. Bigtable: A distributed storage system for structured data. *ACM Transactions on Computer Systems 26*, 2 (June 2008), 4:1–4:26.
4. Datta, K. et al. Stencil computation optimization and auto-tuning on state-of-the-art multicore architectures. In *Proceedings of the 2008 ACM/IEEE Conference on Supercomputing* (Austin, TX, Nov. 15–21). IEEE Press, Piscataway, NJ, 2008, 1–12.
5. Dennard, R.H., Gaensslen, F.H., Yu, H.-n., Rideout, V.L., Bassous, E., and LeBlanc, A.R. Design of ion-implanted MOSFETs with very small physical dimensions. *IEEE Journal of Solid State Circuits 9*, 5 (Jan. 1974), 256–268.
6. Dongarra, J.J. The LINPACK benchmark: An explanation. In *Proceedings of the First International Conference on Supercomputing* (Athens, Greece, June 8–12). Springer-Verlag, New York, 1988, 456–474.
7. Dongarra, J.J. et al. The international exascale software project roadmap. *International Journal of High Performance Computing Applications 25*, 1 (Feb. 2011), 3–60.
8. Esmaeilzadeh, H., Blem, E., Amant, R.S., Sankaralingam, K., and Burger, D. Dark silicon and the end of multicore scaling. In *Proceedings of the 38th Annual International Symposium on Computer Architecture* (San Jose, CA, June 4–8). ACM, New York, 2011, 365–376.
9. Fuller, S.H. and Millett, L.I. Computing performance: Game over or next level? *Computer 44*, 1 (Jan. 2011), 31–38.
10. Geist, A. and Lucas, R. Major computer science challenges at exascale. *International Journal of High Performance Applications 23*, 4 (Nov. 2009), 427–436.
11. Gill, P., Jain, N., and Nagappan, N. Understanding network failures in data centers: Measurement, analysis, and implications. *Proceedings of ACM SIGCOMM 41*, 4 (Aug, 2011), 350–361.
12. Hey, T., Tansley, S., and Tolle, K. *The Fourth Paradigm: Data-Intensive Scientific Discovery.* Microsoft Research, Redmond, WA, 2009; http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf
13. Kamil, S., Shalf, J., and Strohmaier, E. Power efficiency in high-performance computing. In *Proceedings of the Fourth Workshop on High-Performance, Power-Aware Computing* (Miami, FL, Apr.). IEEE Press, 2008.
14. Kogge, P., Bergman, K., Borkar, S. et al. *Exascale Computing Study: Technology Challenges in Achieving Exascale Systems.* U.S. Defense Advanced Research Projects Agency, Arlington, VA, 2008; http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf
15. Lucas, R., Ang, J., Bergman, K., Borkar, S. et al. *Top Ten Exascale Research Challenges.* Office of Science, U.S. Department of Energy, Washington, D.C., Feb. 2014; http://science.energy.gov/~/media/ascr/ascac/pdf/meetings/20140210/Top10reportFEB14.pdf
16. Meuer, H., Strohmaier, E., Dongarra, J. and Simon, H. *Top 500 Supercomputer Sites, 2015*; http://www.top500.org
17. Nobelprize.org. Nobel Prize in Chemistry 2013; http://www.nobelprize.org/nobel_prizes/chemistry/laureates/2013/press.html
18. Olston, C., Reed, B., Strivastava, U., Kumar, R., and Tomkins, A. Pig Latin: A not-so-foreign language for data processing. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (Vancouver, BC, Canada, June 9–12). ACM Press, New York, 2008, 1099–1110.
19. Partnership for Advanced Computing in Europe (PRACE). 2014; http://www.prace-ri.eu/
20. Pinheiro, E., Weber, W.-D., and Barroso, L.A. Failure trends in a large disk drive population. In *Proceedings of the Fifth USENIX Conference on File and Storage Technologies* (San Jose, CA, Feb. 13–16). USENIX Association, Berkeley, CA, 2007.
21. Schroeder, B. and Gibson, G.A. Understanding disk failure rates: What does an MTTF of 1,000,000 hours mean to you? *ACM Transactions on Storage 3*, 3 (Oct. 2007), 8.
22. Schroeder, B., Pinheiro, E., and Weber, W.-D. DRAM errors in the wild: A large-scale field study. In *Proceedings of the 11th International Joint Conference on Measurement and Modeling of Computer Systems* (Seattle, WA, June). ACM Press, New York, 2009, 193–204.
23. U.S. Department of Energy. *Synergistic Challenges in Data-Intensive Science and Exascale Computing.* Report of the Advanced Scientific Computing Advisory Committee Subcommittee, Mar. 30, 2013; http://science.energy.gov/~/media/ascr/ascac/pdf/reports/2013/ASCAC_Data_Intensive_Computing_report_final.pdf
24. U.S. Department of Energy. *The Opportunities and Challenges of Exascale Computing.* Office of Science, Washington, D.C., 2010; http://science.energy.gov/~/media/ascr/ascac/pdf/reports/Exascale_subcommittee_report.pdf
25. White, T. *Hadoop: The Definitive Guide.* O'Reilly Media, May 2012.

Daniel A. Reed (dan-reed@uiowa.edu) is Vice President for Research and Economic Development, and Professor of Computer Science, Electrical, and Computer Engineering and Medicine, at the University of Iowa.

Jack Dongarra (dongarra@icl.utk.edu) holds an appointment at the University of Tennessee, Oak Ridge National Laboratory, and the University of Manchester.

Watch the authors discuss their work in this exclusive *Communications* video. http://cacm.acm.org/videos/exascale-computing-and-big-data