The science of visual analysis at extreme scale

Lucy T. Nowell*
Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy
19901 Germantown Road, Germantown, MD 20874

ABSTRACT

Driven by market forces and spanning the full spectrum of computational devices, computer architectures are changing in ways that present tremendous opportunities and challenges for data analysis and visual analytic technologies. Leadership-class high performance computing system will have as many as a million cores by 2020 and support 10 billion-way concurrency, while laptop computers are expected to have as many as 1,000 cores by 2015. At the same time, data of all types are increasing exponentially and automated analytic methods are essential for all disciplines. Many existing analytic technologies do not scale to make full use of current platforms and fewer still are likely to scale to the systems that will be operational by the end of this decade. Furthermore, on the new architectures and for data at extreme scales, validating the accuracy and effectiveness of analytic methods, including visual analysis, will be increasingly important.

Keywords: data management, visual analysis, scientific visualization, computer architecture, high performance computing

1. INTRODUCTION

Conducting computational science at the Exascale poses such a wide range of challenges that a single scientist or a small team will no longer suffice, especially for scientific endeavors that have potentially vast societal impacts. Solutions for many important problems will require integrative, cross-disciplinary engagement and deep thinking that goes beyond traditional disciplinary boundaries. Furthermore, radically different heterogeneous multi-core machine architecture, a drastically reduced memory capacity per processor, limited input-output (IO) bandwidth, and stringent power management requirements will necessitate rethinking the way application software and user scientists interact with the machines and data. These challenges are not limited to computational science at scale, extending to other disciplines that require analysis of abundant data.

2. CHANGING COMPUTER ARCHITECTURES

Scientific challenges such as understanding the causes and potential impacts of climate change, improving the efficiency of combustion, and unraveling the mysteries of dark energy and dark matter, as well as a variety of national security challenges, require computational capabilities at extreme scale. [1, 2] At the same time, industry reports make it clear that the exponential growth in processor clock speeds that sustained increases in computational speed for more than 15 years has ended. It is likely that Exascale computer systems will be comprised of as many as a billion cores and that such systems will be capable of 10 billion-way concurrency in simultaneous operations. Industry reports indicate that data movement will be the limiting factor for Exascale systems, rather than processors and computational

*lucy.nowell@science.doe.gov; phone 301-903-3191; http://science.doe.gov/ascr/

Visualization and Data Analysis 2011, edited by Pak Chung Wong, Jinah Park, Ming C. Hao, Chaomei Chen, Katy Börner, David L. Kao, Jonathan C. Roberts, Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7868, 786802 · © 2011 SPIE-IS&T · CCC code: 0277-786X/11/$18 · doi: 10.1117/12.881434

operations, especially when power constraints are considered. At the same time, memory per core is expected to decline sharply for Exaflop systems and the performance of storage systems continues to lag far behind. Multi-level storage architectures that span multiple types of hardware are anticipated and will require new approaches to run-time data management and analysis.

Current HPC systems are already sufficiently complex that computational scientists are forced to learn many details about computer hardware, operating systems, storage architectures and file management systems in order to create codes that are essential for their research. This situation is likely to worsen. Measured in floating point operations (FLOPS), system peak performance will increase by a factor of 500. However, system memory is expected to increase by a factor of only 33, node memory bandwidth by a factor of 16, storage by a factor of 20, and input/output bandwidth by a factor of 100. At the same time, total concurrency must increase by a factor of over 4,000. Driven by market forces, these architectural changes and challenges will extend across the full range of computational devices, including laptop computers, which may have as many as 1,000 cores by 2015.

Table 1 summarizes the expected changes in the architecture of leadership class supercomputers by the end of this decade.

Table 1: Potential Exascale computer design for 2018 and its relationship to current high performance computer designs.[3]

	2010	2018	Factor Change
System peak	2 Pf/s	1 Ef/s	500
Power	6 MW	20 MW	3
System Memory	0.3 PB	10 PB	33
Node Performance	0.125 Gf/s	10 Tf/s	80
Node Memory BW	25 GB/s	400 GB/s	16
Node Concurrency	12 cpus	1,000 cpus	83
Interconnect BW	1.5 GB/s	50 GB/s	33
System Size (nodes)	20 K nodes	1 M nodes	50
Total Concurrency	225 K	1 B	4,444
Storage	15 PB	300 PB	20
Input/Output bandwidth	0.2 TB/s	20 TB/s	100

The new system architecture will require fundamental rethinking of the way science is done on our largest computing platforms. The dominant approach to data analysis by the high performance computational science community has been to store simulation data and analyze it at a later time, away from leadership-class supercomputers. However, because the new leadership-class machine architectures will be driven by power and hardware constraints, especially with regard to the memory architecture, this approach will be infeasible. Minimizing data movement will be essential to meeting the power constraints and it will be possible to store only a very small percentage of the data generated by

simulations at scale. These limitations will drive science towards increased reliance on in-situ data reduction and analysis.

The full range of software challenges presented by emerging computer architectures is beyond the scope of this paper. Further information can be found in reports commissioned by the Defense Advanced Research Projects Agency (DARPA)[4,5,6,7] and the Department of Energy[1,8], as well as the International Exascale Project[9].

3. DATA ABUNDANCE

The challenges presented by computer architecture come in an era when the nature of science itself is changing. In a relatively short time, science has shifted from data scarcity to an overwhelming abundance of data, as simulations and experiments generate many Petabytes of data, with some sciences facing Exabytes of data near term. Exponential growth in data generation rates result from a combination of improved sensors, refinements in scale, and improved availability of and access to high performance computing systems. For example, the Large Hadron Collider (LHC) is expected to produce roughly 15 Petabytes of data annually over its estimated 15 year lifespan[10]. A recent report states that climate model data are growing faster than the data set size for any other scientific discipline, with collections of hundreds of exabytes expected by 2020 [11] and the planning process for major scientific projects now includes planning for Exabyte scale data sets.

The impact on science as a result of such collections is discussed in *The Fourth Paradigm: Data Intensive Scientific Discovery* [12] and a host of other reports, including a special issue of the *Nature* [13]. The explosive growth of data worldwide is documented in a variety of reports, including two white papers by the IDC, *The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010*, which notes that the information added annually to the digital universe will reach 988 exabytes per year by 2010 [14], and *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011* [15].

4. DATA MANAGEMENT AND ANALYSIS CHALLENGES

The value of scientific data is realized only when data are effectively analyzed and results are presented to the science community, policy makers, and the public in an understandable way. With extreme-scale data, increased reliance on automated data analysis is essential. Humans are capable of reading only a few gigabytes of text in a lifetime. We cannot even skim an Exabyte of numeric or textual data and our ability to detect patterns in numeric data is not strong.

As discussed above, industry reports suggest that the high power cost of data movement will constrain computation, not the availability of computer cycles. Furthermore, exascale systems are likely to have much deeper memory hierarchy than current systems, with resulting need for intelligent resource management. These expected machine characteristics underscore the need to re-think data analysis so that as much analytic processing as possible is done at the time of data generation, despite the small memory footprint of the machines. Because of power costs, message passing across nodes must also be minimized, so new approaches to integrative analysis are required. This is especially true for data from multiple simulations, whether running as an ensemble with varying parameters or a suite of models required to simulate complex phenomena.

The challenges of analyzing massive scientific data sets are compounded by data complexity that results from heterogeneous methods and devices for data generation and capture and the inherently multi-scale, multi-physics nature of many sciences, resulting in data with hundreds of attributes or dimensions and spanning multiple spatial and temporal scales. The combination of massive scale and complexity is such that high performance computers will be needed to analyze data, as well as to generate it through modeling and simulation. As the *International Exascale Project Roadmap* notes, “the potential impact of Exascale computing will be measured not just in the power it can provide for simulations but also in the capabilities it provides for managing and making sense of the data produced.

…Individual simulations would potentially produce Petabytes+ of data due to scaling, and when combined with multiple executions the data could approach Exabyte scale. Thus, managing scientific data has been identified by the scientific community as one of the most important emerging needs because of the sheer volume and increasing complexity of data." [9]

Sharing, re-use, and re-purposing of scientific data and integration of data from multiple simulations and multiple disciplines are required to address mission-critical challenges in complex systems. Analysis of massive heterogeneous data sets are required, for example, for understanding the impact of stockpile decay on containment materials over decades or understanding the causes and potential impacts of climate change. Such analyses may engage hundreds to thousands of scientists at multiple locations and from multiple disciplines. Integration and/or comparison of data from simulations and observations are necessary for model validation, as well as requiring analysis in their own right. The challenges of analyzing massive scientific data sets are compounded by data complexity that results from heterogeneous methods and devices for data generation and capture and the inherently multi-scale, multi-physics nature of many sciences, resulting in data with hundreds of attributes or dimensions and spanning multiple spatial and temporal scales.

The study of climate requires integration of data from a wide variety of models, including models of oceans, atmosphere, clouds, land surfaces, sea ice, solar radiation, etc. These data have varying formats and characteristics that already present major challenges for analysis using current practices. For disciplines such as those involved in climate modeling, both software codes and data are shared community property and changes to the models require community consensus and validation. Data format standards have taken many years to develop and adopt across an international community and changes to them will require extensive investigation and discussion. Recording data provenance – the data production/generation, transformations and reduction processes so that the data support discovery – will be increasingly important.

These challenges are exacerbated by the data reduction process that will be essential for scientific simulations running on Exascale systems. As mentioned above, for simulations on Exascale computing platforms, the combination of restrictions on data movement, limited storage and IO bandwidth mean that it will be possible to store only a very small percentage of the data generated by simulations at scale. For scientists who are accustomed to saving all of their data for subsequent analysis, this change presents profound challenges. Automating data triage so that the most valuable data are preserved and important potential discoveries are not aborted because the data are no longer available for analysis is chief among them.

Different knowledge representation schemes enable different types of automated analysis, so knowledge representation choices constrain subsequent analytic methods. Integration of data from multiple sources and/or disciplines is also facilitated or hindered by knowledge representation choices. Thus, critical scientific advances depend on engagement of knowledge representation experts with the science community. Knowledge representation and machine reasoning are needed for a variety of problems, especially including representation of and reasoning about uncertainty and sources of uncertainty in the capture/generation and analysis of scientific data. Also needed are knowledge representation methods that support automated analysis of large scientific data sets that include tensor flow fields, including vectors, such as electromagnetic fields, elastic and plastic strain in materials, viscosity, and velocity fields.

Understanding, quantifying and managing uncertainty also requires research. "Uncertainty quantification" refers to the broad range of activities aimed at assessing and improving confidence in simulation. There are many different sources of uncertainty and error that arise in the modeling and simulation of complex systems. For increasing the confidence of simulations, it is important to accurately characterize and quantify the effects of uncertainties and errors on mathematical models and computational algorithms. Data may also be a source of uncertainty because of variations in

the quality of data that are captured on diverse instruments, a variety of data reduction and summarization methods, and errors resulting from storage system faults, to mention but a few factors[16].

Among the data management and analysis challenges we face are the following:

- How can data be represented in the system so as to maximize its analytic value while also minimizing the power and memory cost of the analytic process?
- How can data provenance, which is essential for validation and later reuse/repurposing, be captured and stored without overburdening a system that is IO bound?
- For complex scientific problems that require integrated analysis of data from multiple simulations, observatories, and/or disciplines, how can the expected IO and memory constraints be overcome to support re-use and repurposing of data?
- In the context of these memory and IO constrained systems, how can simulation data be compared to or integrated with observational/experimental data, both to validate the simulations and to support new types of analysis?
- Are new abstractions needed for long-term data storage that move beyond the concept of files to more richly represent the scientific semantics of experiments, simulations, and data points?
- How can data analysis contribute to generating the ten to one hundred billion way concurrency that future machines will support and need to mask latency?
- How can data management and analysis applications help to mitigate the impact of frequent hardware failures and silent faults?

5. VISUAL ANALYSIS

Visual analysis systems that enable interaction between the scientist users, the data analysis system, and the data are critical for supporting scientific discovery and understanding, as well as enhancing communication about science outcomes with the science community, policy makers, and the public. By visual analysis systems, we mean scientific visualization software that is deeply coupled to both an underlying data analysis engine and the data itself, making it possible for scientists to interact with the data analysis system and the data. Such systems should support evolving understanding of both the data and the analytic process.

The sheer quantity and complexity of data call for new thinking about visualization as a tool for data exploration and interactive analysis, as data will far surpass human analytic capacity. Estimating based on the standard three kilobytes per page of text, typical human will read on the order of one gigabyte of text in a lifetime. A really avid reader might get to five to ten gigabytes, but not more. The only way we have any prayer of keeping up with the proliferation of information is through visual analysis, and that is just for text. Furthermore, the pace of creating digital information and data is increasing exponentially.

The data of computational science, on the other hand, are mostly numeric, with some non-numeric metadata, and it is even more difficult for humans to extract meaning from streams of numbers than from text. Faced with many Petabytes or Exabytes of numeric data, scientists must rely on automated analysis to comprehend their data and to detect patterns within it.

Current scientific visualization systems are typically designed for use by a single person and are characterized by limited by display resolutions that do not allow even one pixel per datum (a challenge that is likely to persist), limited interaction with the analytic engine and data, difficulty of use such that scientists seldom create their own visualizations, and inadequate attention to human perceptual and cognitive characteristics that influence how information is extracted from the visualization. Research is needed to facilitate visual analysis of extreme scale,

heterogeneous, high-dimensional scientific data, including support for multiple users who may not be co-located. In particular, research is needed to develop methods for visual representation of uncertainty, visual comparison of multiple data sets or outcomes, and visual representations for tensor flow fields and vectors. Interactive visual representations of system performance that support fault management and/or user intervention in long-running simulations and/or analytic processes on extreme-scale systems with heterogeneous architectures are also needed.

We collect data to help us understand some phenomenon. The data are simply a signal – evidence about the phenomenon. It is the phenomenon itself that is important to us, not the data per se. If we visualize data to recognize outliers or detect errors in processing, that is valuable but it is not sufficient. What we really need to do is use the data to create visual representations of the phenomena that motivated us to capture the data. Fortunately, many developers of scientific visualizations have understood this and have developed powerful representations that advance scientific understanding.

A key challenge will be to develop and validate visual analysis systems that support scientific discovery, in addition to confirmatory analysis and communication, while maintaining user scientists' ability to control the course of analysis and ensuring their faith (and that of policy makers and the public) in the integrity of the process of scientific analysis. It is not enough to produce an answer; scientists must be able to understand, trust and explain the results of analytic processes, communicating them effectively to other scientists and to policy makers. Providing deep integration of visualization processes with data analytic engines and the data will be especially challenging on the IO-constrained platforms of the future. Such integration is essential to support scientists' interaction with their data, and such interaction is essential to supporting human insight, intuition, and learning. Visualization alone is not enough.

Cognitive psychologists tell us that humans learn best when multiple senses are engaged simultaneously. From personal experience, I know that I listen and learn most effectively when I take detailed notes. It is not the notes themselves that are important, since I may never look at them again. But the process of writing focuses my attention more effectively, so that I hear more completely and I make a mental record of what I write.

Interaction may play a more important role in cognition. In a 1963 study, Held and Hein[17] worked with kittens, keeping them in total darkness until the experiment began. For a given pair of littermates, one kitten was confined in a basket that was on wheels. The rolling basket was tethered to the other kitten, which was free to roam. At the end of the experiment, the kitten that had been free to move about was normal. The kitten that had been confined to the basket and pulled about was functionally blind, though its eyes functioned normally. Though it had been able to see all of the same things as the kitten pulling the basket, the inability to control the experience seemed to short-circuit visual information processing. This is reminiscent of what it is like to be merely a passenger in a vehicle, admiring the scenery but unable to describe the route taken from one location to another. Like the kitten in the rolling basket, we don't retain much information from an experience when we are just passengers. Insight and learning depend upon manipulation of our environment. Thus, our visual representations need to be interfaces to the systems that create them – more like steering mechanisms and exploratory toolkits than passive displays to be observed without interaction. How to create this kind of visual analytic environment in the face of emerging changes in computational platforms is not clear, but surely reaching the point where processor power is no longer a limitation offers abundant opportunities to support the human faces of science.

Visual analysis systems are fundamentally about communication. They enable the "bits to brains" transfer of information from a computer system to a human user. While much attention is paid to validating algorithms for transforming data into visual representations, rarely is the effort made to validate that the information put into the visual representation has been or can be accurately extracted and interpreted by human users. Indeed it is not clear that effective methods for validation of visual analysis systems have been developed, but such validation is critical to building trust in the systems and to advancing the field of visual analysis from art to science. Issues that must be

addressed go far beyond usability evaluation and utility evaluation to assessing the perceptual accuracy and efficacy of the visual representations, in addition to validating the underlying analytic methods for data sets so large that only automated methods are possible. The visual analysis research community must develop ways of measuring the impact of visualization and visual interaction and routinely evaluate the systems using the new metrics. It is only when reliable and broadly recognized methods of validation exist and are rigorously applied that visualization and visual analysis will move beyond the realm of one-off art to be accepted as scientific disciplines.

Challenges to be addressed in scientific visual analysis include but are not limited to the following:

- How can a visual analysis system represent data sets for which there is not even one pixel per data point without falsely conveying a sense of uniformity in data? How can small anomalies in petascale to exascale data be made perceptible?
- How can data from multi-scale, multi-physics simulations and experiments be represented so that the complexity of the science is accessible while also maintaining comprehensibility of the displays?
- Can visual analytic systems support collaborative data analysis and communication across distributed multi-disciplinary teams of scientists, without loss of disciplinary semantic context in the visual representation?
- Can visual analysis systems help scientists understand and manage the uncertainty that results from exascale system characteristics?
- Can visual analysis systems provide insight to developers and application scientists about the behavior of applications on exascale computation platforms and support application optimization?
- How can the accuracy of communication with the human users of visual analysis systems be validated and maintained across multiple data sets and potentially vast user communities?

CONCLUSION

Significant changes in computer architectures will occur over the next few years, with impact across the full spectrum of computational devices. Providing support for unprecedented levels of parallelism in application software, the new architectures will also drive dramatic changes in the way data are analyzed in all disciplines, whether the data are textual, multimedia, or the numeric data of science. Research is needed across a broad range of topics and issues to ensure our ability to make sense of data and continue to make scientific discoveries in support of national priorities.

REFERENCES

[1] Ashby, S. et al, The Opportunities and Challenges of Exascale Computing: Summary Report of the Advanced Scientific Computing Advisory Committee (ASCAC) Subcommittee (2010), http://science.doe.gov/ascr/ASCAC/Reports/Exascale-Subcommittee-Report.pdf.
[2] ASCR Scientific Grand Challenges Workshop series, http://extremecomputing.labworks.org/index.stm
[3] DOE Exascale Initiative Roadmap, Architecture and Technology Workshop, San Diego (2009).
[4] Kogge, Peter, et al, ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems: The International Exascale Software Project Roadmap, report for the DARPA Information Processing Techniques Office (IPTO) (2008), http://www.cse.nd.edu/Reports/2008/TR-2008-13.pdf.
[5] Amarsinghe, S., et al, ExaScale Software Study: Software Challenges in Extreme Scale System, report for the DARPA Information Processing Techniques Office (IPTO) (2009), http://users.ece.gatech.edu/mrichard/ExascaleComputingStudyReports/ECSS%20report%20101909.pdf
[6] Binachini, R, et al, System Resilience at Extreme Scale, report for the DARPA Information Processing Techniques Office (IPTO), http://institutes.lanl.gov/resilience/docs/IBM%20Mootaz%20White%20Paper%20System%20Resilience.pdf
[7] Sarkar, V., Harrod, W.,and Snavely, A.E."Software Challenges in Extreme Scale Systems,"in SciDAC 2009, http://www.cs.rice.edu/~vs3/PDF/Sarkar-Harrod-Snavely-SciDAC-2009.pdf

[8] Modeling and Simulation at the Exascale for Energy and the Environment, http://science.doe.gov/ascr/ProgramDocuments/Docs/TownHall.pdf.
[9] The International Exascale Software Project Roadmap, http://www.exascale.org/iesp/Main_Page
[10] http://public.web.cern.ch/Public/en/LHC/Computing-en.html
[11] Challenges in Climate Change Science and the Role of Computing at the Extreme Scale, http://extremecomputing.labworks.org/climate/report.stm.
[12] Hey, T., Tansley, S., and Tolle, K. (ed), [The Fourth Paradigm: Data Intensive Scientific Discovery] Microsoft Research, Redmond, WA (2009) http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx.
[13] Nature (462), 722-723 (10 December 2009), http://www.nature.com/nature/journal/v462/n7274/full/462722a.html).
[14]The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010, 9, (http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf),
[15]The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011, http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf.
[16] http://science.doe.gov/ascr/Funding/Notices/DE-FOA-0000315.pdf
[17] Held, R. and Hein, "A Movement produced stimulation in the development of visually guided behavior," Jou. Comp. and Phys/ Psych (56) 872-876 (1963).

ACKNOWLEDGEMENTS

The ideas articulated in this paper result from intensive discussion with personnel in the Office of Advanced Scientific Computing Research (ASCR), participants at ASCR-supported workshops, and researchers whose work we support.