LLNL-TR-660223

LAWRENCE LIVERMORE NATIONAL LABORATORY

Where Big Data and Prediction Meet

J. Ahrens, J. M. Brase, B. Hart, D. Kusnezov, J. Shalf

September 11, 2014

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Where Big Data and Prediction Meet

The powerful combination of big data acquisition and high-performance computing (HPC) will foster innovation across the economy.

James Ahrens[1] , Jim Brase[2] , Bill Hart[3] , Dimitri Kusnezov[4] , John Shalf[5]

[1]Los Alamos National Laboratory, Los Alamos, NM

[2]Lawrence Livermore National Laboratory, Livermore, CA

[3]Sandia National Laboratories, Albuquerque, NM

[4]US Department of Energy, 1000 Independence Ave SW, Washington DC

[5]Lawrence Berkeley National Laboratory, Berkeley, CA

Our ability to assemble and analyze massive data sets, often referred to under the title of “big data”, is an increasingly important tool for shaping national policy. This in turn has introduced issues from privacy concerns to cyber security. But as IBM’s John Kelly emphasized in the last Innovation, making sense of the vast arrays of data will require radically new computing tools. In the past, technologies and tools for analysis of big data were viewed as quite different from the traditional realm of high performance computing (HPC) with its huge models of phenomena such as global climate or supporting the nuclear test moratorium. Looking ahead, this will change with very positive benefits for both worlds. Societal issues such as global security, economic planning and genetic analysis demand increased understanding that goes beyond existing data analysis and reduction.

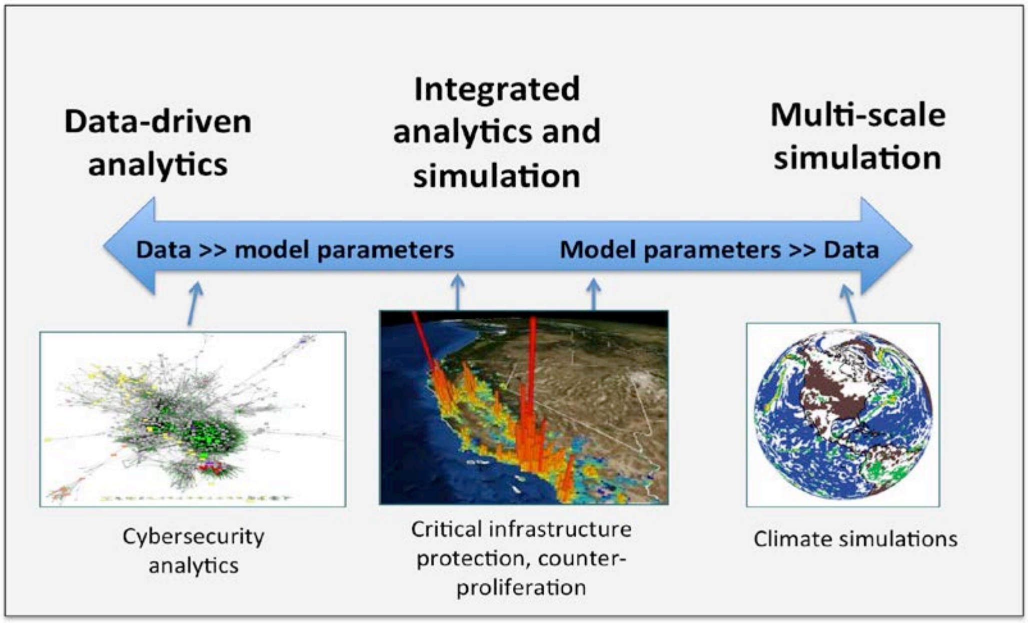
The modeling world often produces simulations that are complex compositions of mathematical models and experimental data. This has resulted in outstanding successes such as the annual assessment of the state of the US nuclear weapons stockpile without underground nuclear testing. Ironically, while there were historically many test conducted, this body of data provides only modest insight into the underlying physics of the system. A great deal of emphasis was thus placed on the level of confidence we can develop for the predictions.

As data analytics and simulation come together, there is a growing need to assess the confidence levels in both data being gathered and the complex models used to make predictions. An example of this is assuring the security or optimizing the performance of critical infrastructure systems such as the power grid. If one wants to understand the vulnerabilities of the system or impacts of predicted threats, full scales tests of the grid against threat scenarios are unlikely. Preventive measures would need to be predicated on well-defined margins of confidence in order to take mitigating actions that could have wide ranging impacts.

There is a rich opportunity for interaction and exchange between the HPC simulation and data analytics communities.

The Spectrum of National Missions – Data-Driven vs. Model-Driven

The classes of problems tackled through HPC range from extremely data rich to extremely data poor. This is illustrated in the figure. The problems on the right end of the spectrum are those where predictions are extremely valuable but the data needed to validate model assumptions is sparse or non-existant. Here collections of models, over many length and time scales, such as transport or chemical reactions in the case of climate modeling, are combined together to make an overall prediction. Only limited sets of data can be brought to bear to help validate the veracity of the overall simulations. At the left end of the spectrum we put problems, which originate from rich sets of data. These could include large sets of experimental data from celestial imaging, particle-physics experiments, genetic sequencers, or infrastructure sensor arrays. It could also include large sets of data of mixed type and origin, lacking understanding of their reliability or content, often in random formats . Data analytics live in this domain where we must work backwards to solve the inverse problem of understanding what is causative in terms of relationships discovered in the data versus what is just a correlation.



The spectrum of data rich and sparse data problems

Historically, HPC has focused on the right end of the spectrum. A driver and exemplar of this is the annual assessment of the state of the US nuclear weapons stockpile. Many of the largest HPC systems have been dedicated over the years to high-fidelity physics-based simulation of the performance of these complex devices.

The right side of the figure with its national security heritage has driven an emphasis on 'uncertainty quantification' and 'verification and validation' to affirm the veracity of predictions in the modeling world. These classes of simulations drive the technology of high-performance computing, requiring large amounts of memory on the processors and large bandwidth between processors.

A large segment of the federal investment in HPC has been based on this particular class of problems for decades but this has also had a very positive effect on other business and academic fields. This investment has also driven the amazing progression in parallel processing that has progressed from ten-thousand processors in the 90s, to millions today, and billions in the near future. This parallel processing capability has been applied to a wide variety of problems. Industry innovations have been an essential part of the development of parallel processing through significant clock speed improvements as well as multi-core on-chip parallelism. Early investments in storage led to high performance parallel storage systems, such as Lustre. At one time, a high percentage of the world's fastest computers have used the Lustre file storage system. Investments in communication technologies, including high-speed networks such as Infiniband, led to high bandwidth networking solutions that are in use in data center facilities today. The history of this development was outlined in several articles in the 3rd quarter issue of Innovation devoted to HPC.

Convergence of Big Data Analytics and HPC

Today, many of our national priorities require computing in parts of the spectrum that are driven by both data analytics and modeling. These crucial applications that span the space from data to modeling will define requirements for new classes of computing technology.

Advanced national security applications as well as new classes of business and social media analytics, will require a large amount of data movement within the analysis process. The data centers at companies like Google and Facebook are now beginning to ask for high-performance interconnects in their analytic systems that have typically been the hallmark of HPC systems. New scalable search methods for database analysis and dense communication patterns in emerging machine learning methods will push available memory bandwidth to the limit ,and stress the performance of large-scale interconnects.. Extending algorithms and computing methods to the large-scale naturally leads to HPC architectures. Data analytics and HPC for simulation are beginning to look very similar in their computing and data movement patterns. They face the same set of challenges in taking advantage of next-generation computing technology.

The High Performance Computing Crisis

HPC is at a transition point with technology reaching important limits in data movement capability and power consumption. Federal mission needs are working to drive HPC performance into new domains: one of low power, more memory on processor, billion level parallelism, smarter algorithms and data management to foster efficient use of resources. These directions are not driven by today's market forces, but can have tremendous derivative value. The robust portable electronics, cloud services, and healthcare markets, for instance, can typically accept lower resource efficiencies and have not been shaping the conversation on the transformational technologies for analytics and large scale modeling that define Fig. 1.

Historically, computation was the biggest cost in systems and the biggest source of energy consumption. This is no longer the case. The largest source of power consumption is the cost of data movement, while relatively speaking, computations are nearly free. But as data movement becomes the main challenge for both analytics and simulation, HPC faces a growing crisis in its ability to have data in the right place at the right time. . Today US data centers consume roughly 2% of US electricity production and while most data is 'dark', in that it sits idle, the opportunities presented by Big Data could increase computing and storage demands. HPC has typically distinguished itself through the interconnect and memory bandwidth and innovations in efficient computer memory and supercomputing architectures could enable more effective use of Big Data while reducing the carbon footprint of this sector.

Addressing the Challenges

The HPC simulation and analytics communities must face these challenges together. Progress on addressing data movement, energy needs, and programming models can have broad impact over the full spectrum of critical computing applications. Government–industry-university partnerships can support these co-design activities through public software and data sets that enable broad participation in the needed R&D and ultimately create an open community to address these challenges. Targeted investments to enable the convergence of the needed R&D will directly support the emerging national priorities for predictive analytics and simulation. The convergence of both driving mission applications requirements and computing technology will profoundly change the high-performance computing and Big Data worlds.

Big Data and HPC Link to address Crucial National Problems

Nuclear counter-proliferation

Our national security priorities are increasingly focused on the security, resilience, and activities of the complex information and transportation networks that we depend on. The discovery of illicit nuclear smuggling requires the integration of data from many sources including financial networks, shipping transactions, and communication networks. Continuing rapid advances in computing and algorithm technologies are greatly improving our capability to discover and ultimately prevent these threats. At the same time, new data protection technologies can work together with clear policy guidance to allow these systems to balance threat detection performance with privacy protection.

Energy Critical Infrastructure

The Nation's energy grid and critical infrastructure are growing increasingly networked providing the ability for smartmeters to help regulate and monitor energy delivery to homes. This emerging data will continue to grow in its diversity and content in the next years as systems become more instrumented. Big data will serve as a foundation in the identification of threats and vulnerabilities in our large networks as well as improve delivery and efficiency. It will help in dealing with the impacts of natural disasters. Developing the policies and technologies needed to balance the protection of critical infrastructure with protecting the privacy of the broad set of customers will be a crucial component in assuring the resilience of our critical infrastructure.