Big Data Meets HPC

Suzanne Tracy, Editor-in-Chief, Scientific Computing and HPC Source

Addressing the major opportunities and challenges of the 21st century

High performance computing (HPC) has already contributed enormously to scientific innovation, industrial and economic competitiveness, national and regional security, and the quality of human life. This crucial role has been emphasized in recent years by U.S. and Russian presidents, as well as by senior officials in Europe and Asia.

Steve Conway, IDC VP HPC explains that, to date, most data-intensive HPC jobs in the government, academic and industrial sectors have involved the modeling and simulation of complex physical and quasi-physical systems. These systems range from product designs for cars, planes, golf clubs and pharmaceuticals, to subatomic particles, global weather and climate patterns, and the cosmos itself. However, he notes that from the start of the supercomputer era in the 1960s — and even earlier — an important subset of HPC jobs has involved analytics, attempts to uncover useful information and patterns in the data itself. For example, cryptography, one of the original scientific-technical computing applications, falls predominantly into this category.

The financial services industry was the first commercial market to adopt supercomputers for advanced data analytics. According to Conway, large investment banks began hiring particle physicists from Los Alamos National Laboratory (LANL) and the Santa Fe Institute in the 1980s in order to employ HPC systems for daunting analytics tasks, such as optimizing portfolios of mortgage-backed securities, pricing exotic financial instruments, and managing firm-wide, global risk. This practice has continued: in 2013, Goldman Sachs lured a particle physicist away from the Large Hadron Collider at CERN.

What Is Driving Demand?

High-performance data analysis — what Conway refers to as "HPDA" — is both an evolutionary and a revolutionary story. He believes that the data explosion fueling the growth of high performance data analysis stems from a mix of long-standing and newer factors:

- the ability of increasingly powerful HPC systems to run data-intensive modeling and simulation problems at larger scale, at higher resolution, and with more elements (e.g., inclusion of the carbon cycle in climate ensemble models)
- proliferation of larger, more complex scientific instruments and sensor networks, from "smart" power grids to the Large Hadron Collider and Square Kilometer Array
- the increasing transformation of certain disciplines into data-driven sciences — biology is a notable example — but this transformation extends even to humanities disciplines such as archeology and linguistics
- growth of stochastic modeling (financial services), parametric modeling (manufacturing) and other iterative problem-solving methods, whose cumulative results produce large data volumes
- availability of newer advanced analytics methods and tools: MapReduce/Hadoop, graph analytics, semantic analysis, knowledge discovery algorithms and others
- the escalating need to perform advanced analytics in near-real-time — a need that is causing a new wave of commercial firms to adopt HPC for the first time

Existing HPC Disciplines Expand Analytics

Conway notes that "some members of the climate research community have begun to augment existing methods with analytics-based knowledge discovery algorithms to promote new insights, and perhaps no field has stronger potential for benefiting from HPC-based analytics than bioscience. He adds that data-intensive applications already in motion in the varied bioscience field "range from advanced research — notably in genomics, proteomics, epidemiology and systems biology — to commercial initiatives to develop new drugs and medical treatments, agricultural pesticides and other bio-products."

Conway believes that one of the world's most socially and economically important HPDA thrusts will almost surely be the multi-year transition from today's procedures-based medicine to personalized, outcomes-based healthcare. Identifying highly effective treatments in near-real-time by comparing an individual's genetic makeup, health history and symptomology against tens of millions of archived patient records poses enormous HPDA challenges that may take another decade to master. When this capability matures, he believes that it will likely serve as a decision-support tool of unprecedented utility for the global healthcare community.

In yet another bioscience example, German-based Schrödinger is using HPC public cloud resources to identify promising candidates for new drugs to combat cancer and other diseases. IDC believes that at least half a dozen pharmaceutical firms are following in Schrödinger's footsteps.

Newer analytics methods and tools are likely to benefit all existing HPC vertical segments at least to some extent. These segments also include computer-aided engineering, chemical engineering, digital content creation and distribution, electronic data automation, financial services, geosciences and geo-engineering (oil and gas), defense, government labs and academia. But the story doesn't end there.

Conway points out that high-potential horizontal analytics applications also are starting to make an important impact in the world of high performance computing.

"Fraud detection, cyber security and insider threats are increasingly crucial challenges for established HPC users in government, academia and industry to meet, and they are causing a new wave of commercial organizations to move up to HPC for the first time. Prominent examples range from PayPal to Italy's Istituto Nazionale della Previdenza Sociale and the U.S. Postal Service," he says.

Tackling these problems often requires moving beyond today's needle-in-a-haystack, static searches for items already known to exist in a database. The challenge presented by these problems is to discover hidden patterns and relationships — things you didn't know were there — and then to track patterns dynamically as they form and evolve.

As the HPDA vendor scene is becoming increasingly heterogeneous and vibrant, the analytics side of the formative HPDA market is where traditional HPC users and first-time commercial adopters are converging most rapidly. "Established vendors that have served each of these customer groups are exploiting this convergence by following their buyers into the new HPDA analytics territory," Conway explains.

IDC forecasts that revenue for HPDA-focused servers will grow robustly (13.3% CAGR), increasing from $743.8 million in 2012 to approach $1.4 billion in 2017. HPDA storage revenue will approach $800 million in the latter year. The most serious technical challenge to liberating HPDA growth is data movement and management, although the HPDA market should be seen more fundamentally as a war among clever algorithms.

Meeting Data-intensive Challenges

The growing market for high-performance data analysis — using HPC for data-intensive challenges — is already enlarging HPC's contributions to science, commerce and society, and HPDA promises to play a major role in helping to address the major opportunities and challenges of the 21st century. Barry Bolding, VP, Storage & Data Management at Cray reports that "Cray continues to see an increasing trend in the HPC marketplace that we are calling 'data-intensive' supercomputing. The dramatic growth in scientific, commercial and social data is resulting in an expanded customer base that is asking for much more complex analysis and simulation. There is a feedback loop between more and bigger datasets and more complex simulation modeling, and Cray is seeing this across our customer base and market segments. "

As the term big data continued to dominate as a source of technology challenges, experimentation and innovation in 2013, Jorge Titinger, SGI's CEO, observes that "It's no surprise then that many business and IT executives are suffering from big data exhaustion, causing Gartner to deem 2013 as the year the technology entered the 'Trough of Disillusionment.'"

However, Titinger notes that 2013 also saw early signs of the re-entry of high performance computing (HPC) within the enterprise. "The technology has a legacy within the scientific and research community, but big data served as a catalyst for HPC adoption within traditional enterprises. While the term big data might have dipped into disillusionment, HPC holds the promise of guiding big data towards business growth and productivity," he says.

Jason Stowe, Chief Executive Officer at Cycle Computing, notes that "We're seeing several large trends as it relates to big data and analytics. We started talking about this concept of Big Compute back in October 2012 with a blog post titled 'BigData, meet BigCompute: 1 Million Hours, 78 TB of genomic data analysis, in 1 week.' In many ways, it's the collision of where HPC is meeting the challenges of big data. As our technical capabilities continue to expand in the ways we can collect and store data, the problem of how we access and use data is only growing. We're seeing several themes around this issue, and feel strongly that the ability to easily orchestrate and access big data on the cloud provides clear solutions."

He adds that “The cloud allows us the ability to access extremely large amounts of data — some of which is being collected to in near-real-time — and allows us to tap into virtually unlimited computing power.” He describes a few themes that highlight this:

- **Ask the right question:** There is a shift underway where researchers, engineers and analysts can change the very way they think about problems. Previously, we have been limited by the computing resources we have — the clusters we have on premise. Today, we can change the very way we ask our questions. Ask the right questions — and use the cloud to create the size of system needed to answer your questions.
- **Unprecedented collaboration:** Today, scientific breakthroughs come from teams of people, instead of the lone scientist. Often the teams are, in fact, collaborating on different continents. It’s amazing that technology has enabled this type of worldwide collaboration. But it’s even more exciting to consider how HPC in the cloud is taking this ability to a completely new level.
- **Enabling technologies for streaming analytics:** A few technologies come to mind as innovative and allowing for these themes to come alive. Some of these include NoSql databases, RabbitMQ, and something we at Cycle Computing are calling Jupiter. Jupiter was designed to enable low overhead, streaming computations and analytics on hundreds to hundreds of thousands of cores. Highly resilient, and with extremely low overhead, Jupiter has already proven itself when we used it to conduct a record-breaking 156,000+ core cloud computing run over all eight AWS Regions, running Schrödinger Materials Sciences tools — called The MegaRun. Jupiter was critical in making this happen — and will be a key tool for cloud computing runs of all sizes in the future.

A Comprehensive Approach

How can organizations embrace — instead of brace for — the rapidly intensifying collision of public and private clouds, HPC environments and big data? The current go-to solution for many organizations is to run these technology assets in siloed, specialized environments. However, according to Robert Clyde, CEO of Adaptive

Computing, “This approach falls short, typically taxing one datacenter area while others remain underutilized, functioning as little more than expensive storage space.” Clyde explains that “As larger and more complex data sets emerge, it becomes increasingly more difficult to process big data using on-hand database management tools or traditional data processing applications. To maximize their significant investments in these datacenter resources, companies must tackle big data with ‘Big Workflow’ — a term we’ve coined at Adaptive Computing to describe a comprehensive approach that maximizes datacenter resources and streamlines the simulation and data analysis process.”

Adaptive’s Big Workflow approach utilizes all available resources within the datacenter, including HPC environments, as well as other datacenter resources like private and public cloud, big data, virtual machines and bare metal. Under the Big Workflow umbrella, all datacenter resources are optimized, eliminating the logjam and turning it into an organized workflow that greatly increases throughput and productivity.

Current State-of-the-Art

“Looking at the industry, we are currently seeing state-of-the-art tools like OpenStack and Hadoop being used for big data processing. Late last year, we joined the OpenStack Community and announced an integration on OpenStack. In addition, we integrated Moab and Intel HPC Distribution for Apache Hadoop software, a milestone in the big data ecosystem that allows Hadoop workloads to run on HPC systems. Now, organizations have the ability to expand beyond a siloed approach and leverage both their HPC and big data investments together,” Clyde notes.

“State-of-the-art for data-intensive supercomputing has to bring productive parallelization to the fingertips of the customer,” Bolding explains. “Whether this is powerful parallel processing hardware and software, blazingly fast parallel file sytems, or parallel graph analytics, parallelism is a key component of state-of-the-art.” He adds that “Open systems, where data is not permanently tied to a particular company or technology are also key. Cray heavily relies on open operating systems, open tiered storage and open analytics tools for many components of its system to insure the longevity of customer data and customer-developed tools.”

At SGI, Titinger notes that “Our customers have quickly realized the benefits of pairing HPC with big data. One prime example is PayPal. The leader in online payments, PayPal processes 13 million financial transactions a day and deployed SGI’s UV platform and HPC technology to ensure fraud was caught early. In the first year, SGI analyzed PayPal’s big data and identified $710 million in fraud that would have otherwise been undetected, identifying suspicious patterns across separate sumptuous transactions.” He explains that “as the company grew, real-time analytics was no longer enough — technology tools have expanded to include predictive analytics as a key competitive advantage for companies such as PayPal. Big data analytics, as a whole, has grown in importance by providing unprecedented and intuitive insights into an array of business functions, such as sales, supply chain and network operations.”

Looking Forward

According to Bolding, “Cray sees this ever-increasing tie between data analytics, data management and computation to be a long-term trend. This tight coupling will lead to discoveries, both scientific and social, which will make their mark in the consciousness of both the high performance computing user and society as a whole. HPC has long had difficulty showing the general public how our products impact their lives, and our industry needs to get better at this. These impacts will become more pervasive and, if we use them wisely, they will make all our lives more productive and more meaningful.”

In 2014, Titinger predicts that key HPC capabilities will enhance the level of intelligence and speed big data analytics can provide for the enterprise, including:

- **Graphing and mapping**: HPC-powered data mapping and graphing will lead to greater accuracy in business forecasting
- **Pattern visualizations:** HPC-powered tools will emerge that can provide an intuitive view of complex data sets, enabling rapid identification of relationships for simple analysis
- **Scaling in-memory databases:** HPC will allow enterprise in-memory systems to handle larger data workloads — allowing closer to complete data sets (over partial sets) to benefit from real-time analytics while in motion

- **Meta-data:** The importance of metadata will jump dramatically — we'll see enterprises realize leveraging meta-data analytics for virtualization and relational mapping can yield enhanced accuracy, new business insights and even reveal security threats

"Overall, 2014 will be the year when HPC elevates its status within the enterprise — becoming a must-have business technology to extract the largest value from the largest amounts of data," Titinger says. "Arguably, the biggest issue today is the failure to recognize the true scope of available relevant data. Big data analytics is most effective when it combines not only internal structured and unstructured data, but pairs this with externally available data from all information sources such as social, market, Web and sensor data. By pairing HPC with big data, companies will maximize the intelligence from all these sources, processing data at high volumes, with the speed and accuracy enterprises need to continue to thrive."

Suzanne Tracy is editor-in-chief of Scientific Computing and HPC Source. She may be reached at editor@ScientificComputing.com.

Related Content

- High Performance Data Analysis: Big Data Meets HPC [1] by Steve Conway
- Big Compute: The Collision of where HPC is Meeting the Challenges of Big Data [2] by Jason Stowe
- Scalable Productivity and the Ever-Increasing Tie between Data Analytics, Data Management and Computation [3] by Barry Bolding
- Big Workflow: The Future of Big Data Computing [4] by Robert Clyde
- Big Data & HPC: The Modern Crystal Ball for 2014 [5] by Jorge Titinger

Source URL (retrieved on *10/25/2014 - 4:51pm*):
http://www.scientificcomputing.com/articles/2014/03/big-data-meets-hpc

Links:
[1] http://www.scientificcomputing.com/blogs/2014/03/high-performance-data-analysis-big-data-meets-hpc
[2] http://www.scientificcomputing.com/blogs/2014/03/big-compute-collision-where-hpc-meeting-challenges-big-data
[3] http://www.scientificcomputing.com/blogs/2014/03/scalable-productivity-and-ever-increasing-tie-between-data-analytics-data-management-and-computation
[4] http://www.scientificcomputing.com/blogs/2014/03/big-workflow-future-big-data-computing
[5] http://www.scientificcomputing.com/blogs/2014/03/big-data-hpc-modern-crystal-ball-2014