Update

# Technologies and Applications in High-Performance Data Analysis: HPC User Forum, September 15–17, 2014, Seattle, Washington

Steve Conway
Robert Sorensen
Earl C. Joseph, Ph.D.
Chirag Dekate, Ph.D.

## IN THIS UPDATE

This IDC update captures part of the proceedings at the 54th HPC User Forum held in Seattle, Washington. Organizations worldwide are taking decisive action to develop high-performance computing capability through regional, national, and international initiatives. Key leaders in the high-performance computing industry spoke at the HPC User Forum in Seattle, Washington. Presentations made by John Feo of Pacific Northwest National Laboratory, Nima Negahban and Amit Vij of GIS Federal, and Charlotte Crain of SAS Institute are captured in this update. Feo shared experiences on building scalable technologies for semantic analysis. Vij and Negahban presented details about their real-time geospatial-rendering project. Crain shared insights on how commercial firms are exploiting multicluster grids for data-intensive computing.

High-performance computing innovation will be driven by disruptive developments in several of the enabling technologies. Many of the HPC industries will be key in driving the innovation to the market and packaging it for broader consumption. IDC believes that innovations and investments in high-performance computing will be needed on a sustained basis to catalyze regional and national high-performance computing efforts.

## Building Scalable Technologies for Semantic Analysis — John Feo, Pacific Northwest National Laboratory

I'll talk about the semantic wrap engine we've been developing for the past three-and-a-half years. Data tends to be unstructured and very heterogeneous, so fixed data structures are no longer adequate (table format with fixed columns). Data coming in often is not owned or created by the user but comes in from the Internet or other outside sources. You don't know what data will come in, so graphs are the right way to represent the data. NoSQL distributes data over multiple servers, so it's complex to track.

We've developed GEMS. This is a database that supports structured and unstructured data as a graph. It has in-memory data storage and runs on everything from desktops to clouds. There are no special system requirements. We're using a cluster (32 nodes) to process data. Use graph algorithms to process data. Because data is stored as a graph, we don't need to use SELECTs and JOINs, and this saves lots of storage. It works with OpenMP, MPI, Linux, and x86. You just need a C++ compiler

and an MPI. We generate tens of thousands of threads per node. We're using the SPARQL query language. We've developed a multithreaded runtime system that scales on commodity hardware.

Why graphs? Because the data that we operate on is often unstructured. Graph infrastructures don't have much data locality, so they can't use traditional memory structures. We use clusters to scale memory by adding nodes, and we use threads to hide latency.

GEMS is running on 64- and 128-processor Linux systems today. The method can scale up to large clusters. We also have a 128-processor uRiKA system. We're beating uRiKA on the Berlin benchmark by up to 6x. We typically also beat Neo4j by 5–7x.

Our solution is intended as a semantic graph engine. We support a variety of graph algorithms.

## Real-Time Geospatial-Rendering Project — Nima Negahban and Amit Vij, GIS Federal

I'm CTO of GIS Federal. My colleague here is Amit Vij. GPUdb is our database. It's centered mainly around GPUs but could use Phi or other coprocessors. We were on an army project needing hundreds of sources to stream data at runtime without indexing because that would have taken too long. This led to the idea to develop a geospatial-rendering engine that would be easy for a user to use. We've been incubated by the U.S. Army INSCOM and are in the process of going into production for this client in Afghanistan.

### *GPU Supercomputing Cloud Overview*

Accelerator GPU-based supercomputers will be a big trend because of their low power use. We tie the many GPU threads to the data through a cluster. The user doesn't need to be a GPU or a cluster expert, just a normal developer and can still do powerful analytics. An analyst can take a huge query on the fly and do analytics. We're not customized with FPGAs – GPUs are commodity hardware (e.g., Bloomberg bond pricing). Our solution had much lower price, power use, and space requirements than the others we compared it with.

GPUdb Gandhi Cluster has 5 nodes, 7GB RAM, 3TB HDD, 2 Xeon HEX-CORE, and 1 NVIDIA 750Ti (640 cores, 2GB VRAM). Total cost is $5,040. Right now, we're exploring which markets to aim this at.

### *On-the-Fly Time-Referenced Heatmap Video Generation*

We got feeds from many UAVs, so the data is from hundreds of sources. We have no fixed schema or indexing. Another example is tracking all flights originating from LAX, with the ability to add parameters such as showing all flights in a specific altitude range.

Next example is 112 million data points for ship paths. You can plug in parameters such as vessel type = tanker, and you can then zero in on this by geographic locations. Full-text search capability is built in, such as "oil."

Next example is Twitter data, which is a type of problem we encountered in our work for the army. (He tried a search on ISIS.) Many users employ GPUdb as an input to their high-order algorithms. We're graphlike but not a graph database. Our database would be much faster on a bigger 64-bit system. At army, things run on an SGI UV 2000 server.

Comment: What about security?

Speaker: This is all public data (e.g., Twitter).

## Commercial Firms Exploiting Multicluster Grids for HPC — Charlotte Crain, SAS Institute

Data historically has been thrown all over the place; so today, we have long-standing databases of various kinds, and customers are telling us what they want to be able to do with it for auditing/compliance, better financial performance, and various degrees of analysis. Big data in business depends on what the customer considers big. In general, more and more complex analyses are needed to stay ahead of the competition.

So IT sees growing data problems, shrinking budgets, and the need to do more with less, including more sophisticated analytical models. Business requirements such as healthcare regulations are also increasing. The cost of maintaining legacy systems is also growing. So IT has lots of challenges.

### *SAS Grid Types and What We're Doing for Customers*

Customers want to be the best and outcompete rivals. They are moving to these technologies to do that:

- Grid computing
- In-database analytics
- In-memory server

With HPC grid computing, you take what you have and scale out rather than scaling up because scaling out is much less expensive. You use commodity hardware, then distribute and manage the workload across the servers. Many customers are going to this type of grid computing and it's a cost reduction – because you use existing hardware, maybe add some more, and throw Linux on it and you're ready to go.

*In-database analytics* is where you push the analytics to the data and don't move the data. You move the functions into the database. This enables faster time to results. You can run existing SAS codes without modification and run them on all the data. *In-memory analytics* has been around for a few years. Move the analytics into the in-memory server. All the business users can use the same data tables: Banking, retail, and others are looking for this type of solution.

We've been focusing more on Hadoop this year. We can quickly move data stored in Hadoop into memory. We have been doing data for decades and have very accurate data. Our customers would like to get where PayPal is today but aren't doing that. It's too costly for most customers.

SAS has SAS Grid, which includes commodity hardware and a Hadoop cluster. We can move data into SAS Grid or SAS Rack.

### *Trends*

Big data has really taken hold in business in the past few years. Hadoop is driving this and the many data sources such as Twitter feeds and others. IoT is also growing exponentially. We call this the Era of Abundance. In business, there are two mindsets, the scarcity mindset and the abundance mindset, where people think – "I've been doing this 20 years, and this is how I think about my problem" versus thinking from scratch about the best way to do the problem:

- **Era of Scarcity.** This is what we hear a lot from IT. These people spend a lot on storage and limit access to data to a few people. They must stay within a budget. They spend a lot of time doing workarounds. IT provides data but does not meet all the data users' needs. This approach can take months to get the data users really need. Analysts have good ideas they can't implement quickly because they're stuck in this "data prison." Some people do amazing things with the tools and budgets they have, but for most, this is an issue. Because of all these scarcity delays, companies' products are late to market, and this turns into a vicious cycle.
- **Era of Abundance.** Why is it good to change your thinking and adapt? You can know a lot about consumers' habits and what they like and maybe their risky behaviors (e.g., Is someone a motorcycle driver?). What if I could look at all of this data together – would that change how I look at this customer? The old way was siloed data, and one risky behavior could characterize a customer, period. SAS is huge on fraud analytics. Today, we have a huge retailer and we can tell what you like and can offer you a discount when you're in a store, real time. A question is, "How can we do this and scale it?"

## LEARN MORE

## Related Research

Additional research from IDC in the technical computing hardware program includes the following documents:

- *Worldwide Broader HPC 2014-2018 Forecast: Servers, Storage, Software, Middleware, and Services* (IDC #248835, June 2014)
- *When Massive Data Never Becomes Big Data* (IDC #IcUS24922014, June 2014)
- *Worldwide Technical Computing Server 2014-2018 Forecast* (IDC #248779, May 2014)
- *Perspectives on High-Performance Data Analysis: The Life Sciences* (IDC #248348, May 2014)
- *Global HPC Market Dynamics in 2013* (IDC #248137, April 2014)
- *Industrial Partnership Programs and High-Performance Computing: HPC User Forum, April 7-9, 2014, Santa Fe, New Mexico* (IDC #248113, April 2014)
- *Disruptive Technologies in High-Performance Computing: HPC User Forum, April 7-9, 2014, Santa Fe, New Mexico* (IDC #248112, April 2014)

- *Advances in Processors, Coprocessors, and Accelerators in High-Performance Computing: HPC User Forum, April 7-9, 2014, Santa Fe, New Mexico* (IDC #248111, April 2014)
- *International Perspectives on Industrial High-Performance Computing Partnerships: HPC User Forum, April 7-9, 2014, Santa Fe, New Mexico* (IDC #248122, April 2014)
- *Worldwide HPC Public Cloud Computing 2014-2017 Forecast* (IDC #247846, April 2014)
- *Summary of IDC's 2014 Research in the Use of HPC by Oil and Gas Organizations* (IDC #247704, March 2014)
- *IBM Sale to Lenovo Opens Opportunity for Other HPC Vendors* (IDC #lcUS24694314, February 2014)
- *IDC's Worldwide High-Performance Computing Predictions 2014* (IDC #WC20140211, February 2014)
- *Seagate Looking for the X Factor in Its Acquisition of Xyratex* (IDC #lcUS24555413, December 2013)
- *Micron Demonstrates Technologies to Address Emerging Challenges in Big Data Applications* (IDC #244843, December 2013)
- *Market Analysis Perspective: Worldwide HPC, 2013 – Directions, Trends, and Customer Requirements* (IDC #244742, December 2013)
- *HPDA Pulse: 2013 Software and Consulting Market Analysis* (IDC #244513, November 2013)
- *HPDA Pulse Results: 2013 Hardware and Storage Market Analysis* (IDC #244493, November 2013)
- *HP FY13: Revenue Declines Abate on Stronger Core Business* (IDC #lcUS24466413, November 2013)
- *Catalyst Supercomputer Heralds Shift to More Balanced Architectures* (IDC #lcUS24437513, November 2013)
- *China Eyes 10,000-Fold Data Reduction for Internet of Things* (IDC #lcUS24392513, October 2013)
- *HPC User Forum, October 2013, Seoul, Korea* (IDC #243786, October 2013)
- *Tools and Techniques for Technical Computing in Life Sciences: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243778, October 2013)
- *Perspectives on Quantum Computing: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243777, October 2013)
- *National and International Initiatives: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243776, October 2013)
- *Issues in High-Performance Computing: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243775, October 2013)
- *High-Performance Data Analysis in the Life Sciences: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243774, October 2013)
- *Chinese Research in Processor Designs for High-Performance Computing and Other Uses* (IDC #243502, October 2013)

- *World's Fastest Supercomputer Set to Reach Customer in October 2013* (IDC #IcUS24300913, September 2013)
- *The Broader HPC Market 2012-2017 Forecast: Servers, Storage, Software, Middleware, and Services* (IDC #242742, August 2013)
- *IDC's Worldwide Technical Server Taxonomy, 2013* (IDC #242725, August 2013)
- *China Regains Top Supercomputer Title* (IDC #IcUS24190613, June 2013)
- *10 Things CIOs Should Know About High-Performance Computing* (IDC #241565, June 2013)
- *Worldwide High-Performance Data Analysis 2013-2017 Forecast* (IDC #241315, June 2013)
- *Top Issues for HPC Sites: HPC User Forum, April 29-May 1, 2013, Tucson, Arizona* (IDC #241463, June 2013)

## About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

## Global Headquarters

5 Speen Street
Framingham, MA 01701
USA
508.872.8200
Twitter: @IDC
idc-insights-community.com
www.idc.com

---