Update

Emergent Use Cases in High-Performance Data Analysis: HPC User Forum, September 15–17, 2014, Seattle, Washington

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

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 Arno Kolster and Ryan Quick from PayPal, Shane Corder from Children's Mercy Hospital, and Jack Collins from National Cancer Institute are captured in this document. Arno Kolster and Ryan Quick from PayPal shared insight on using DSPs to develop real-time analytics platform at PayPal. Shane Corder shared insights on using genomic sequencing and HPC to help save the lives of critically ill children. Jack Collins presented details on the use of HPC in cancer research.

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.

Real-Time Analytics — Arno Kolster and Ryan Quick, PayPal

This is our third time here. We're changing things by going right to real-time analytics. We won an HPCwire award at SC13. We also cosponsored a BOF with ORNL last year and will do one again at SC14. We have deployed an SGI UV 2000.

Systems Intelligence is what we call our latest analytics engine and it performs real-time anomaly detection before any negative user experience. This requires correlating events across multiple incoming data streams and understanding what correlations need to occur. We're processing 3 million events per second and this volume keeps increasing. This amounts to 25TB of ingestion per hour, or 20MBps. We're doing metadata around 50,000 relationships. We correlate across all this data and all these streams that include Twitter feeds and other sources.

Systems Intelligence uses multiple sources: applications logs, machine data, environmental data from the datacenter, and social media. All this gets ingested in real time with lots of parsing and semantic analysis. We have a huge shared memory event processor in our SGI UV 2000 that feeds into both graph and relationship databases.

Why Are We Leveraging HPC?

We are tasked with finding "the next big thing" two to five years from now, which means we need to make accurate predictions about which technologies will be useful to PayPal; so we shopped around HPC for tools to meet our real-time requirements. On a whim I went to SC one year to go shopping for things that could do what our enterprise technology couldn't handle. With more expertise, we've noticed that this is a grey area – and we're not the only ones with these types of problems. So at SC we've been talking with others for whom HPC is a little too much but enterprise technology is not good enough.

What We Consider Real Time

In classical concert experiences such as opera, you can hear a bad note in real time. When this happens, you have used predictive analytics because you knew the aria in advance and with real-time analytics you found the note was bad.

eBay has very green datacenters. We all need to be faster, greener, and cheaper.

Last year we were in a meeting where HP talked about Moonshot and a node for standard Web hosting and another one that is very energy efficient, etc. I asked them to go back to the node with DSPs (digital signal processors), which has four ARM 815 cores. We wondered about piping all the data into this. We asked HP more about it, and we found TI and they showed us a diagram called multicore navigator [he showed a node with four of these]. We concluded they built an HPC cluster and put it on a system on a chip. That's what we're here to talk about.

Complex Event Processing as Digital Signals

We're making this official announcement today. There's familiar Linux integration on the ARM side to interface with MySQL, Java tools, etc. in the datacenter. The goal is to take data and encode it as a digital signal. This lets you exploit 40 years of mature algorithms in the digital signal area. We started working with folks on how to take advantage of this with a new app written in OpenCL. You can test it on a laptop and then run it on a cluster. The idea is to turn the data into a signal, something like a sine wave. We have eight of these DSPs and can look at many signals at once and do pattern recognition and take output from one and feed it into others.

We can have multiple filters on an atomic event stream, multiple event streams, and pattern recognition. This solution finds outliers and frequency of occurrence. We've done IP around this. System Intelligence has four sockets per cartridge running at 55W, for 11.2GF/W, which we consider outstanding. This is our first foray into writing our own HPC applications. We are meeting all our design goals. The first use case for Systems Intelligence is flow analysis/graphing to create a graph of services and methods. We're working on this with HP and TI.

Using Genomic Sequencing and HPC to Help Save the Lives of Critically Ill Children — Shane Corder, Center for Pediatric Genomic Medicine, Children's Mercy Hospital, Kansas City, Missouri

The center was established in 2007 by Dr. Stephen Kingsmore. CMH was then working to start a world-class genome center. We are one of the best-known genome sequencing centers in the world. In 2010, our work was in *Time Magazine*'s top 10 medical breakthroughs.

A human being has about 22,000 genes. We don't make designer babies. We run a 40-node cluster, with 600–700 cores. We have 1PB of Isilon storage plus a second tier with SGI.

To date, people go to a doctor and have many tests run at average $23,000 cost, and many times there is still no certain diagnosis. There are roughly 4,100 genetic diseases known to affect humans and these are the main cause of infant deaths.

Infant liver failure example: We did a 25-hour sequencing on 120 billion nucleotide sequences. We narrowed the problem down to two variants. Treatment was corticosteroids and immunoglobulin. The baby is alive and well today.

Our goal is to use this method to change treatment for disease symptoms. In 48% of the cases we worked on, our diagnosis changed the treatment. In the other 52%, at least the parents know all that could be done had been done and they did not need to subject their child to additional tests and treatments.

What's next? Out five-year goal is to provide a diagnosis for every child that comes to us, within one month. We also want to do RNA analysis.

Genomes to Structured Function (and Movies): The Role of HPC — Jack Collins, National Cancer Institute

The more we understand something in science, the more we resort to calculation – whether in genomics or other cases I'll talk about. In science, data is king. All scientists are data scientists because scientists without data are philosophers.

How do we move from the genomics to the chemistry of the DNA down to the physics of what's going on? This involves quantum chemistry.

Biology is not so simple. There are lots of random mutations happening all the time. Which ones cause cancer? Closer to reality is that we're organisms that move in and react with our environment.

Cancerous cells inserted into an animal typically don't develop cancers – this argues that what causes cancer is the interaction between cancerous cells and response systems. Systems biology is where things are focusing.

We've been playing with graph and other types of databases to speed things up because we want to know about all the data and information we can get about the body and the pathways, etc. I need to know this all at the same time and with some kind of time locality.

For therapeutics, functional understanding and drug development require 3D structures. Advances in structural biology are being driven by advances in physics, such as x-ray lasers that can reveal detailed structures; also better detectors/cameras developed for consumers and for astronomy (increases in megapixel resolution). These tools allow us to see crystalline structures we couldn't see before. There are also new algorithms needed. We're starting to see molecular motion for the first time.

By combing quantum mechanics with experiments, we can also learn much more about molecular structures. People are now engineering nanoparticles. It takes simulation to be able to visualize the structures of these particles in detail. (We need to advance risk assessment for the FDA approval process.)

Nanotoxicity

Nanoparticles do serious damage at the cellular level. With HPC, we're pushing toward being able to impact the workflow/work process of the people who are trying to solve the problems.

Role/challenges for HPC:

- Integrating big data into the enterprise workflow
- Integrating heterogeneous computational technologies: CPU, GPU, Phi, etc.
- Efficient software
- Finding enough skilled people
- Preferred programming models today are R, MATLAB, and Python (maybe Java)
- Codes may not be flops dependent
- Macs work well
- People in my field generally prefer open source software
- People are willing to use a cloud because you don't have to wait for IT to provision a system
- We want to be able to work with the latest science
- People are merging enterprise and HPC and optimizing workflows, such as for metastasis analysis

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 #IcUS24694314, 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 #IcUS24555413, 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 #lcUS24300913, 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 #lcUS24190613, 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