Update

Perspectives on High-Performance Data Analysis: Government Applications

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

IN THIS UPDATE

This IDC update presents near-verbatim notes from talks on high-performance data analysis (HPDA) use in government, given at IDC-organized HPC User Forum meetings. Together, these talks provide a good feeling for the wide variety of government applications that have begun to exploit high-performance computing (HPC) resources for HPDA. IDC believes that government will become one of the most scientifically and economically important markets for HPDA. IDC coined the term *high-performance data analysis*) to refer to the use of high-performance computing resources to run Big Data workloads, especially advanced analytics and data-intensive modeling and simulation.

Special Keynote: "Scalable Data Mining and Archiving in the Era of the Square Kilometre Array": Chris Mattmann, NASA's Jet Propulsion Laboratory

I'm senior computer scientist at JPL [NASA's Jet Propulsion Laboratory] and teach at USC and am on the board of the Apache Foundation [oversees Apache-Hadoop].

The Square Kilometre Array (SKA) will be built over the next decade. It will take that long to develop the technologies to support 700TBps of streaming data. SKA is an international consortium. Both South Africa and Australia will have components of the SKA onsite. This project will drive a number of Big Data use cases.

SKA will undertake strong field tests of gravity and explore other fundamental astronomical questions of our era. JPL will investigate scalable data archiving, mining, and data management in partnership with the SKA sites and other collaborators.

JPL's Big Data initiative is looking not just at radio arrays but also at climate, energy, life and physical sciences, and other areas. We'll investigate fundamental technologies and techniques in archiving and data science.

Some Big Data Grand Challenges

- How to handle 700TBps of data required by the SKA
- Rapid science algorithm integration (Already-developed algorithms need to be integrated and run on large data sets, as needed by the Western Snow Hydrology Project.)

May 2014, IDC #248349

- How to compare petabytes of climate model output in various formats with petabytes of remote sensing data to improve climate models for the next Intergovernmental Panel on Climate Change (IPCC) assessment
- Cataloguing all of NASA's planetary science data, as required

Technology Thrusts

- Data movement technologies for Big Data systems
- Cloud computing for spaceborne and airborne missions and ground-based sensors
- Extensions to JPL-led Apache Tika and other technologies

The Apache Object Oriented Data Technology (OODT) project is a data management framework for information integration and high throughput for large-scale data processing. It was incubated at the Apache Software Foundation in 2010. NIH, NCI, and others are contributing.

Why Apache OODT? We see open source as a way to collaborate when funding isn't available up front. Apache is the elite open source community for software developers. It provides a governance and management structure, unlike other open source communities. OODT core components include a file manager, a workflow manager, and a resource manager. All are implemented as Web services.

The SKA Project

- NSF does not highly prioritize SKA and is taking a five-year wait-and-see approach, but NASA is using open source to continue working with SKA because we think it's important for the future of Big Data.
- The MeerKAT data repository will take 10PB of data, which isn't all that large. KAT-7 is the precursor to MeerKAT. JPL is establishing a U.S. archive so South African SKA scientists can share data with U.S. scientists. There's an early prototype of KAT-7. We're mirroring this out to South Africa and to the United States. Data movement will be from Cape Town to the U.S. archive.

The National Radio Astronomy Observatory

- This is funded and operated by NSF. JPL activities include leveraging Apache OODT and our architecture experience to define an achievable prototype.
- Our demonstration use case took data from the Expanded Very Large Array (EVLA) summer school project at the National Radio Astronomy Observatory (NRAO) and demonstrated we can execute our code and make the data available anywhere in the world. SKA liked this a lot and has deployed Apache OODT for its data reduction pipeline.
- We're working to implement more parts of Apache OODT.

SKA will be storing more of the raw data than originally intended because they realized the value of it for time-dependent studies.

"Data Analysis and Visualization for the DoD HPCMP": Paul Adams, ERDC

We support the Air Force warfighter.

Data is our problem. We'll add 3PB to our archive this year and expect to reach about 33PB by January 2016. Data locality is our problem. How much disk do we need to buy, and how fast will our network be? How does a researcher with a 100TB file get this moved to his or her home location?

Solutions We Developed

- **Custom visualization.** This is good for researchers who need to show PowerPoint slides to their funders.
- **Community visualization.** Here we teach people how to do the visualization themselves, via a Web site with 200 tutorials.
- **Collaborative visualization.** We teach people one on one how to do the visualization.
- **Remote visualization.** You leave the big files (hundreds of terabytes) where they are. ezViz provides batch-mode visualization on HPC systems. This supports more than 50 input formats and 16 output formats, and it's available on all nonclassified systems. Secure Remote Desktop is a Linux desktop that's been hardened for DoD. It depends on relatively fast network latency and uses less than 150KBps of bandwidth. The future is to provide this service on a Web interface.

3D TVs are now common, enabling us in the future to deliver 3D scenes to end users without using expensive CAVEs. We'll also provide augmented reality with tablet visual solutions.

We directly support the warfighter, such as with modular protective systems that can be delivered directly to the war zone. They protect warfighters from RPGs, attacks, and suicide bombers. The system is a kind of container for two men. It's been tested by physical and digital tests and has been deployed.

"Making Sense of 'Big Data' in Biology and Cancer Research": Jack Collins, National Cancer Institute/SAIC-Frederick

People in 2001 talked about the "data tsunami" but kept focused on floating point. What do we mean by "Big Data"? There are big data sets from LHC experiments, etc., and also "lots of data" from tens of thousands of genomes or YouTube videos, etc. These are very different animals.

Storage and access to storage are critical issues. Security and data integrity are huge problems. Users want easy access to everything on the fastest disks, and that's not always possible. It's not easy to move data in hierarchical storage systems. Storage environments are distributed today. In medicine, it's being generated worldwide.

If I run a calculation on your genome to see if you'll get cancer, I give you an answer today. In six months, the software has changed and you might get a different answer. Metadata repositories are also a major issue, as are who pays for the data.

Another question is what is HPC? It's increasingly becoming an appliance and part of the infrastructure needed to analyze data, make decisions, and impact medical science on a daily basis and in an almost unnoticeable way.

Traditional HPC experts often have a narrow view of the new applied user world. The exascale user base is very narrow. HPC should be baked into workflows to optimize human time, our most expensive resource. Hardware is cheap. People write software and are the limiting factor. Pad computers and clouds will become portals to ubiquitous HPC infrastructure. HPC will drive systems biology, personalized medicine, etc.

We need plug-and-play workloads. Visualization will be king for making sense of mountains of data.

Many research methods and tools have not been optimized. It's common for researchers to waste much time and effort trying without success to implement an underdeveloped method or tool.

At NCI, the dominant factor is data management and analysis. Inside the workload, we need to build HPC into the modules.

Sequencers got really, really cheap and this started the whole conversation. It's pennies per gigabyte to generate the data and dollars per gigabyte to analyze it.

HPC can enable us to move from diagnosis to treatment. Time to solution becomes critical when it's a cancer diagnosis. So does security. The interface for clinicians is quite different from the one for researchers.

Oncology workflow: HPC can be involved at all stages. You characterize the tumor by analyzing various images from various types of sources. HPC shortened the time from two days down to 15 minutes at our center, and this got a lot of attention.

Then the pathologist annotates the image – a whole stack of images – and these images need to be registered. We used HPC at NSCA to do this and reduced the time to two weeks.

Then you characterize the tumor at the genome level.

Analyses are becoming much more complex, and the data is analyzed multiple times.

Now I want to know where the mutations are. I do a large-scale analysis and simulation. This took a few hundred thousand hours on a supercomputer. You're looking for the mechanism that leads to the mutation. You need this information to know how to treat the cancer. Then you mine and integrate various data sources using graph-based methods. And then you want to design a drug, which involves chemistry.

Why a high failure rate for new drugs? We might have 90 cases, 90 controls, and 300 features. But the data is completely random data that we just made up. You can get really good fits while doing really bad science.

On the human side, you need to integrate people with math, science, chemistry, etc. skills. You need skilled computational analysts and to visualize the data (critical). We need a better computer interface, more collaborative environments, and very robust search engines.

Comment: At Boeing, data analysis at large scale is a very big issue. We have to keep the data for the life of the air flight program.

"The National Library of Medicine's Visible Human Project and the DARPA Virtual Soldier Project": Brian Athey, University of Michigan

I also want to talk about the more recent "Virtual Soldier Project." I chair the Informatics Department, including bioinformatics.

The Visible Human Project

The National Library of Medicine (NLM) created two data sets to represent in detail the human male and human female anatomies. This was the original Big Data project. The male was 15GB and the female 50GB. That was considered huge at the time. The male was created in 1994. We talked with Pittsburgh Supercomputing Center (PSC) about how to do this, and I got a contract to do it in conjunction with PSC, from NLM. The Visible Human later became the Virtual Soldier Project.

HPC is great for simulating volumetric data for many purposes. The Visible Human data was linked to anatomical nomenclature. This is similar to what we do with sequencing the genome, including annotation.

The roots of the Visible Human Project (VHP) date back to 1986 in the long-range planning of the NLM. The aim was to create a digital image data set of a complete male and female cadaver in MRI, CT, and anatomical mode. Human anatomy with physical cadavers is the most expensive course to offer at medical schools. We also create a miniature version of the VHP at 20MB size. The VHP was named the most successful biological project of the past half century.

The VHP allowed us to look at everything from tissues and organs to full anatomy, but not at the molecular and cellular level. We weren't able to do much with the brain. The whole bioinformatics and genomics thing came later, but the VHP taught us a lot about how to approach these things.

The Virtual Soldier Project

Multiscale human anatomy is described by the foundational model of anatomy (FDA). The DARPA Virtual Soldier Project was to build a VS on an electronic dog tag to diagnose and predict combat injury and battlefield mortality. Accurate, quick diagnosis and treatment saves lives. This project included building a computer model of the generic patient for comparison. You use pre- and post-wounding individual data to highlight differences for diagnostic purposes. The full hierarchical model of the soldier is called "the holomer" and includes a 3D model of the beating heart. This project provides support to medics on the battlefield.

We consumed lots of supercomputer cycles at academic centers all over the country. Highly integrated physiology model outputs (porcine) were expanded to include pressures, volumes, flows of the heart/respiratory system, airways/lungs, and more. We used porcine heart data for a lot of this because the pig heart closely resembles the human heart. The status monitor displays model data and actual data from the soldier.

We did a lot of modeling of ballistic trauma to the heart using FEM on supercomputers, with collaborators including ATK. Oak Ridge was also involved.

HPC should be used much more to help with the causal reasoning model, to guide the simulation, and to help with the what-if predictions about what would happen when bullets or other ballistics hit certain places on the anatomy. GE helped with imaging automatic segmentations (views of affected parts of the targeted anatomy).

The average battlefield medic is 19 years old and has six weeks of training, so it's hard for them to make quick life-and-death decisions. That was one of the main rationales for the Virtual Soldier Project. All soldiers now have electronic dog tags based on this project. The Virtual Soldier data is more relevant in a conventional war than an Iraq 2-type war.

"The New NASA Earth Exchange: A Big Data Project in NASA Earth Science": Tsengdar Lee, NASA

The Landsat mission has been running for 40 years. We have been building this NASA Earth Exchange (NEX) platform for a couple of years.

Landsat satellites take all sorts of measurements of the earth with various technologies. Our job is to turn the observations into knowledge products. Downlink data volumes vary from megabytes to petabytes (advanced sensors). Exploration technologies support the geospatial, communications, and computing infrastructure.

The sequence goes from data acquisition to data processing/mission control to data transport to Distributed Active Archive Centers (DAACs) to science processing/management/archiving/distribution, etc.

Data transfer happens using the 10GbE National Lambda Rail network. Every night, individual researchers were transporting gigabytes of data from NASA Ames to Goddard, so we needed a fast network. The researchers had a day to analyze the data and then it was erased; otherwise, the disks would stay filled up. A typical university researcher (PI) could spend $10,000 or so on equipment. This model does not scale with growing data volumes.

According to Gray's laws, data-centric computing in science is the fourth paradigm. The solution is scale-out architecture. We need to bring the computations to the data. For this reason, we picked a couple of projects to gear up for Big Data analytics.

Traditional datacenters focus on data archives, access, and distribution. Scientists download specific data sets to a local machine for analysis. With more observational data, this is impractical, so datacenters are starting to provide additional data services. NASA is building data analytics platforms such as NEX and the NASA Data for IPCC Earth System Grid Gateway.

NEX

- It's a collaborative platform to engage and enable our community to do discovery and decision making by combining observations, supercomputing, and social networking.
- We moved a lot of data from our centers to NASA Ames. This is an experiment. NEX has 9PB of online storage, 50PB of tape storage, 512 readily accessible CPUs, and 180,000 CPUs in total.
- We need models to analyze the data, so we have many models and modeling tools built into the NEX platform. We have prestaged many terabytes of data to save time for researchers. Many researchers spend 60% of their time just on managing data, so this saves time.
- We can duplicate the workflow of other researchers' experiences to save additional time for users. They can build on top of these existing workflows.
- NEX also has a portal for submitting jobs. Experienced users will use the "sandbox" to configure their data. The most experienced users will go deeply into the system to map their problem onto the Pleiades supercomputer that supports NEX.
- The portal captures who's doing what, where. It has a knowledge management system with search capabilities and reporting capabilities. It includes the workflows, archived seminars, searchable publications, and spatial distribution of funding.
- We also built a virtual institute that offers summer short courses, seminars, conferences, and presentations. We aim to have each funded project host two seminars. Everything can be downloaded. One course is on Big Data analysis.
- NEX downloaded and processed all the Landsat data, the first time this has happened.
- Benefits include efficient use of resources and other advantages.

A lot more modeling can happen because the data is prestaged. Prestaging the data, the model, and the computer resources lowers the barrier to entry for users.

In the future, we will focus on efforts in data management (e.g., distributed multisite analytics), workflow reuse, workflow discovery, and doing this in a cloud.

"Data-Intensive Research at PNNL": John Feo, Pacific Northwest National Laboratory

How do we specify the problems and give the analysts the tools they need to do their work? We have lots of data, big machines, and lots of problems. So what's the problem? The analysts don't have the tools to specify their problems. How do you do this in SQL? SPARQL? Java? C? So they have to write their own programs in many cases. Do we expect doctors to write their own Java or C programs? It won't happen, so we need to give them an easy interface and way to specify.

Knowledge discovery: The search and optimization problems form the basis for many of these problems.

Complex query workshop a year ago, sponsored by Pacific Northwest National Laboratory (PNNL) and others, to develop a set of abstract graph query patterns that could be adapted to users' specific problems so users could produce compelling standard queries. Workshop report: **hpc.pnl.gov/people/haglin**.

Example: NSF proposal conflicts of interest query. Goal: Among proposals, find a subset of suitable authors without conflicts. This is a difficult query to express because it involves inferencing (geographic, topical), recursion, aggregation, disjunction, directed and undirected links, and units of measure (months).

Another example: Party problem. It's an optimization problem.

Facebook: What happens here will heavily drive what happens with systems and tools in HPC for Big Data.

Steiner tree is an undirected graph with weighted edges.

At PNNL, we're trying to create an entire framework stack for modern searches.

Components of modern searches:

- User interface
- Search/query
- Data storage and manipulation
- Analysis

It's important to use the best type of system for your problem type. MapReduce problems tend to fit well onto shared nothing clusters.

Center for Adaptive Supercomputing Software: We work in many areas of data-intensive computing, including architectural studies (hardware, software, evaluation), compilers, tools, and runtime systems.

Partners: Cray, Georgia Tech, Sandia, Mayo Clinic, UMD, CASL, and others.

We need to take the problem specification layer seriously!

"Data Analysis and Visualization of Very Large Data": David Pugmire, ORNL

We have three petascale machines at ORNL today.

Scientific Analysis and Visualization Today

The purpose of computing is insight, not numbers. We categorize data sets from small to medium, large, and hero size. Small can be done on a laptop or workstation, medium on a fat workstation. Large data sets are painful to move and need distributed parallelism, and hero data sets are impossible to move and can only be analyzed on the supercomputer on which the data was produced.

We have three good tools that can operate on data sets of all these sizes. Apps are pushing data set sizes in spatial resolution, temporal resolution, multivariate, and ensembles. The tools today are

client/server architectures, including a thin client with a GUI and viewer, and communication to the server that does the input/output (I/O), analysis, and visualization before shipping the results back to the client. This model works well if you push the server onto an analysis cluster or supercomputer.

You have a parallel simulation code, and data chunks are written out to disks. The data flow networks filter the data, and you get the result as a plotted visualization.

Experiments with Very Large Data

- We generated a data set of the kind we expect to have in 10 years by taking an existing data set (supernova simulation) and uploading it onto a trillion-cube mesh. We then ran two common algorithms and ran the problem with these algorithms on every supercomputer we could access. We did this in 2010. Basically, only the I/O mattered because it consumed 90-95% of the time. It ran at up to 2.3PF on "Jaguar" (Cray supercomputer at ORNL).
- For climate, we ran a statistical analysis using R as the de facto standard for statistical analysis. We did an extreme value analysis on 100-year data. It scaled up very well.
- Flow analyses are critical for understanding simulations but are hard to parallelize and to understand. Lagrangian methods have proved to be effective for understanding flows. Flows are hard to parallelize because computational requirements are unknown a priori. We developed some algorithms, one of which monitors the load balance in the system and maximizes the use of the resources.

Analysis and Visualization on Next-Generation Architectures

- **Scalable Data Management, Analysis, and Visualization (SDAV).** SDAV is a DOE initiative. Part of this is Adaptable IO System (ADIOS), an I/O abstraction layer aimed at minimizing I/O impact on running applications. Data is a self-describing stream. It's open source and has been very useful in fusion research.
- **Extreme Scale Analysis and Visualization Library (EAVL).** Existing workflows will fail at exascale because of costs, file-centric visualization/analysis, and more. A goal is to create a better data model. De facto data models like VTK can't represent graphs, for example. Running in memory-constrained environments required memory-efficient representations, as well as memory-aware algorithms and memory and computationally efficient algorithms. Another EAVL goal is heterogeneous scalability through implementations for CPU, GPU, and MIC.

Client/server architecture has served us well and will continue to do so, but what goes on inside of it will need to undergo some major, painful changes.

"Processing Large Volumes of Experimental Data": Shane Canon, LBNL

We're in the Big Data era. Much of this is driven by social media and the commercial space. In HPC, instruments are also generating a lot of data: accelerators, sensor networks, and so forth. This is not new in HPC, especially in fields like hydrogen physics. These disciplines think about data management and handling from the start. They also typically know what they are looking for, and this reduced the challenges on the computing side. For the LHC, the raw data is hundreds of megabytes per second, but after filtering, it's hundreds of terabytes per second. They've also built up a cottage industry of technologies in the hydrogen physics community for software technologies, etc.

ORNL has been working with this community since the mid-1990s, with projects such as the collider at Brookhaven National Lab and more recent neutrino projects. This community has been able to control their data growth.

By contrast, the genomics community is being driven by the commoditization of sequencing. Costs have dropped two orders of magnitude for sequencing a genome since 2001, from $100 million to $1,000 (source: National Human Genome Research Institute). Price declines will continue at least for a while longer.

DOE's Joint Genome Institute asked NERSC for access to the HPC infrastructure and to NERSC expertise. DOE wants to know how to turn cellulose into ethanol, for example, and this is a genomics problem. How do you mimic what happens in a termite's stomach?

Synchrotron light sources: DOE operates four of these around the United States. They use these devices for things like building better battery technologies or understanding how photosynthesis works at the cellular level. It used to be that detectors could take two to three pictures per second and now it's hundreds. That's 1GBps of output, which is higher than LHC. Future detectors will look at phenomena at the femtoscale, in maybe four to five years.

Hadoop was inspired by Google's work on MapReduce. An ecosystem has built up around Hadoop. Limitations for science applications:

- Java centric so not for legacy apps
- Text based/line oriented
- Weak for binary data formats
- Not all apps fit MR model
- Maturing support for shared environments

We're looking at how we can run throughput-oriented workloads on our HPC systems. It turns out our systems are fairly cost competitive for this kind of work. We've developed tools for running these workloads, including a "task farmer" and others.

KBase predictive biology project: Berkeley, Argonne, and Brookhaven. Integrate genomic and expression data from other types of experiments and compare them in a Web-based framework. Started seven months ago.

Advanced Networking Initiative, ESnet: 100Gb national network. NERSC, Argonne, and ORNL end points are in place.

Scalable Data Management, Analysis, and Visualization: Obama administration announced this under Big Data initiative with $200 million in funding.

NERSC recently announced a data-intensive big science project offering up to 1PB of storage, up to 20 million core hours, priority access to Hadoop cluster, and more.

Key challenges:

- You have to keep an open mind because the instruments drive much of what needs to happen. You need to meet the user's partway.
- It's not just about the absolute volumes of data but about the velocity and acceleration of data aggregation.
- Big Data puts stress on parallel file systems.
- There's a growing need for supporting services such as Web portals and NOSQL databases.
- You need to work closely with the user community.

"'Gordon,' a Data Monster": Robert Sinkovits, San Diego Supercomputer Center

Gordon's main technology is flash.

Designed for data-intensive and memory-intensive problems that don't run well on traditional distributed memory machines. Traditional systems don't do these things well:

- Large share memory requirements
- Serial or threaded (OpenMP, Pthreads)
- Limited scalability
- High-performance data base applications
- Random I/O combined with very large data sets
- Large scratch files

We're supported by NSF. Gordon is available to all U.S. academic researchers on a competitive basis and on a limited basis for free to most nonacademic users. Appro is the integrator, Mellanox provides the 3D torus.

SDCS and other national labs/centers with academic users have limited control over their user bases, so they need to have general-purpose machines.

Data analytics can involve large amounts of data but small outputs. Intermediate scratch files for graph algorithms, etc., can be very large compared with the input/output files. Simulations can involve modest input sizes and large output sizes, such as CFD, weather/climate simulations. These need to store large 4D data sets.

Gordon hardware:

- 1,024 dual-socket compute nodes: Xeon E5
- 64 dual-socket I/O nodes, including 300TB of total flash memory plus 2x Westmere 2.66 GHz processors
- Dual-rail 3D torus InfiniBand QDR (40Gbps)
- 4PB Lustre-based parallel file system

Number 48 on the November 2011 top 500 system. It would be number 35 today.

Access to data comes with a latency penalty. Gordon latency is two orders of magnitude faster than going out to disk.

16 compute node racks. Based on Appro GreenBlade 8000 series. Compute node: 64GB DRAM, 16 cores, 2.6GHz, 80GB flash.

I/O node: 48GB DRAM, 12 cores, 2.66GHz, 4.8TB flash.

Flash will be made available to compute nodes using iSER. The 3D torus is 4 x 4 x 4, with a 6-hop maximum. The scheduler will be aware of the torus geometry and assign nodes to jobs accordingly.

Gordon is a green system. By removing the local hard drives on the compute nodes we reduced power quite a bit. It's number 30 on the Green500, which is very high for a general-purpose HPC system.

We're using ScaleMP software to create large shared memory nodes that appear to the user as single SMP nodes with 2TB memory and 512 compute cores. We'd eventually like to be able to deploy these vSMP nodes on the fly.

We're also trying a novel allocation process for a few of our users, who can have dedicated use of an I/O node plus 1–4 compute nodes for up to a year.

Traditional users: CFD, high-energy physics, climate, and weather. Plus people making use of Gordon's novel features: genome assembly, computational finance, and more.

"Preparing Applications for Next-Generation IO/Storage": Gary Grider, LANL

Drivers for change include the following:

- Scale (machine and storage sizes)
- Data volumes
- Bursty behavior
- Technology trends
- Economic trends

IO people have a small bag of tricks:

- Burst buffer economics say that at exascale, you have to meet two requirements: 100TBps burst from 1 billion processing elements and 1EB of scratch capacity (30ish memories).
- Buying flash for capacity is inexpensive. Buying disk for capacity is also expensive, but with disk for capacity, you get bandwidth for free.

- The Trinity system will have 2–4PB of DRAM, 5–12PB of flash/burst buffer, and 100PB of disk or thereabouts. There is opportunity for in-transit analysis (before data goes to tape).
- Unlimited archiving will become cost prohibitive. Capacity is no longer the sole cost driver for archive as it was for 25 years. Bandwidth is now a major contributor to the TCO of archives.
- We need to figure out how to allow the storage system to fail. This challenge is in the FastForward project. For reliability, a strategy is to put in checksums all along the way.

Regarding how your apps will change, there will be a need for non-blocking APIs, and more workflows will need to consider the expense of going to disk and especially to archive.

LEARN MORE

Related Research

Additional research from IDC in the high-performance data analysis program includes the following documents:

- *Micron Demonstrates Technologies to Address Emerging Challenges in Big Data Applications* (IDC #244843, 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)
- *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)
- *High-Performance Data Analysis in the Life Sciences: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243774, October 2013)
- *High-Performance Data Analysis – PayPal Breaks New Ground: HPC User Forum, September 2013, Boston, Massachusetts* (IDC #243772, October 2013)
- *European Conference Underscores Movement of Enterprise Big Data Analytics to HPC Capabilities* (IDC #243618, October 2013)
- *HPDA: Importance of Hardware in Effective Big Data Solutions* (IDC #243495, September 2013)
- *eBay Deploys Innovative Green Power-Efficient Datacenter Technologies* (IDC #lcUS24350013, September 2013)
- *Worldwide High-Performance Data Analysis 2013–2017 Forecast* (IDC #241315, June 2013)
- *Changing Market Dynamics: HPC Meeting Big Data and IDC's Projected Evolution of the Market* (IDC #240365, March 2013)

- *High-Performance Data Analysis: HPC Meets Big Data* (IDC #DR2013_LSIS1_SC_CD, March 2013)
- *High-Performance Data Analysis: The Visible Human Project* (IDC #238253, December 2012)
- *High-Performance Data Analysis at NASA JPL* (IDC #238254, December 2012)

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