IDC
Analyze the Future

UPDATE

October 2009 HPC User Forum Meeting Notes: Lausanne, Switzerland

Steve Conway
Jie Wu
Charles Hayes, CHS
Earl C. Joseph, Ph.D.
Lloyd Cohen

IN THIS UPDATE

This IDC update covers the 35th High-Performance Computing (HPC) User Forum meeting, which took place at the Ecole Polytechnique Fédérale de Lausanne (EPFL) in Lausanne, Switzerland on October 8–9, 2009. Local hosts for the meeting were Dr. Henry Markram, project director of the Blue Brain Project, director of the Center for Neuroscience and Technology, and codirector of EPFL's Brain Mind Institute, along with Dr. Felix Schűrmann, general project manager of the Blue Brain Project and a scientist at the Brain Mind Institute. The principal tasks of the meeting were to:

- Continue the HPC User Forum's dialogue between North American and European HPC users and vendors.
- Share information among HPC users, vendors, and IDC for improving the health of the worldwide HPC industry.
- Continue showcasing examples of HPC leadership and partnerships in government, industry, and academia.
- Explore the use of HPC in bioscience.
- Survey the activities and achievements at leading university HPC centers.

Thursday, October 8

HPC User Forum Steering Committee Chairman Steve Finn, BAE Systems, welcomed attendees and thanked EPFL for hosting the meeting. HPC User Forum Executive Director Earl Joseph, IDC, thanked sponsors Altair, Bull, IBM, and Microsoft, and then presented an IDC update on the worldwide HPC server market.

IDC HPC Market Update

The HPC market saw a 17% reduction in the first half of 2009. The high-end supercomputers segment for systems priced at $3 million and above has really changed. It has moved into a high-growth mode. We thought the workgroup would grow quickly, and it is showing slower growth. Many users enter the market here but quickly move upmarket.

2010 IDC HPC research areas: IDC is doing a lot of end-user research and power and cooling research, along with developing a market model for middleware and

Filing Information: December 2009, IDC #221355, Volume: 1
Technical Computing: Update

Global Headquarters: 5 Speen Street Framingham, MA 01701 USA P.508.872.8200 F.508.935.4015 www.idc.com

management software. We're also closely tracking extreme computing, datacenter assessment, and benchmarking and tracking petascale and exascale initiatives around the world.

EPFL Welcome

Henry Markram welcomed attendees on behalf of EPFL and noted that EPFL has a strong HPC tradition. Most recently, EPFL has been using an IBM Blue Gene system.

Giorgio Margaritondo, EPFL vice president for Academic Affairs, explained that EPFL is a young school that was started 40 years ago. Today, EPFL is the number 1 engineering school in continental Europe. EPFL strongly stresses interdisciplinary study and launched the first Swiss satellite two weeks ago. HPC is extremely important for EPFL, representing the third branch of science, and the Blue Brain Project is EPFL's most visible initiative using HPC.

Neil Stringfellow, CSCS: "The Swiss Initiative for High-Performance Computing"

CSCS was established 20 years ago as an autonomous unit of ETH-Z. CSCS serves the Swiss research community and supports the Swiss national weather center, Meteo Swiss, with 2km high-resolution forecasts. The center supports a varied set of other scientific and engineering applications.

Switzerland's National HPCN Strategy

International competition is increasing with the United States, Japan, Germany, France, the United Kingdom, China, and India. CSCS plans to install a petascale (5+ peak PF) by 2011/2012 and will need a new building for this. Plans include creating a Swiss competency network to connect existing application areas and reach out to new ones. These plans have been approved by the federal government and need to be passed by parliament.

The Swiss stimulus package totals 700 million Swiss francs. 2% of this was allocated for CSCS: 3.5 million Swiss francs for CSCS building and planning, 3 million for HPC education, and 10 million for the new machine. We need to invest in both new algorithms and computer hardware.

CSCS's Cray XT5 is nicknamed "Monte Rosa." It will have 22,000 cores when the upgrade now in progress is completed and will be the fourth most powerful HPC system in Europe. One-third of the jobs require more than half of machine.

Simulations are necessary for science. Those who come first get the scientific credit. HPC provides a competitive edge. With it, you can do simulations faster.

The Role of Science in Switzerland

Switzerland places a high value on scientific research and education. We have a high density of recognized computational scientists, even when compared to the United States. The CSCS user community comes from ETH Lausanne; EPFL; the universities of Zurich, Basel, Bern, and Geneva; and EMPA and the Paul Scherrer Institute. This is a very demanding user community.

Current research examples include:

- 60% of Swiss electricity comes from hydroelectric power, which can get damaged by storms. High-end tourism is also important. Understanding the weather and climate is very important for these things. For predicting the frequency of severe weather events in a changing climate, high-resolution simulations are critical. Resolution is very important for Switzerland. At 2.2km and 0.55km, you can see microclimates. You get similar insights with the regional model. Only at 2.2km do you begin to see terrain heights fairly well.

- We also have avalanches and earthquakes, both of which have been modeled on CSCS systems. In 1356, Basel was destroyed by an earthquake. Simulations are also important for planning nuclear power generation. Climate accounts for 25% of our usage.

Strategic Goals

- Develop HPC leadership, for which Switzerland needs a sustainable ecosystem

- Establish relationships with other leading institutions around the world

- Algorithm development

EPFL, ETH-Z, and the University of Zurich are Switzerland's three leading HPC sites. We want the world's most powerful machines to be only about 5x larger than ours in Switzerland. In 3–4 years, that baseline will be about 800TF.

We must establish HPC and computational science programs at Swiss universities. We need a new building to house larger systems. We are at maximum capacity in our current building. A new building is planned in at the new University of Lugano, with 1,500 sq m of space and 10MW of power capacity.

The Swiss plan for high-performance and high-productivity computing includes developing simulation capabilities for 2012–14 platforms, implementing networking strates, and developing scientific curricula for universities.

Issues include huge parallelism, changes in memory, slow improvements in interprocessor communications, stagnant I/O subsystems, and resilience and fault tolerance. For accelerators, the programming languages have to be changed. Algorithms also need to be reengineered.

Henry Markram and Felix Schuermann, EPFL: "Blue Brain Project Update"

Henry Markram: "Blue Brain Project Update"

The Blue Brain Project was aimed at building a facility for building brain models with great detail and accuracy. We were in the proof of concept stage over the last three years with the Blue Gene/L system and performed 15,000 experiments and constructed a first piece of the brain as a proof of concept. The facility today allows us to build any neuron microcircuit with Blue Gene/P. As we increase the power of the facility we can support the creation of more and more brain models. Brain ailments

affect 2 billion humans each year. You need a massive global strategy to bring all knowledge about the human brain together on one platform, and a simulation capability based on powerful supercomputers. Blue Gene allowed us to simulate a cortical column. Now the challenge is scaling up. By 2020, our goal is to simulate a human brain and that will take an exascale computer, not only flops but bits. Without this, there is no global strategy for dealing with neurological diseases.

Felix Schuermann: "Blue Brain Project: Past, Present, and Future Leverage of HPC"

In biology, our experiments surprise us every day. We have to first understand the models, the pieces we bring together, before we can understand at a higher level of integration. This is hard, multidisciplinary work. We have to reverse-engineer the brain. Mammalian brains are very similar to each other. Henry Markram has been doing lots of experiments to identify what things are inside the brain and to quantify these so we can begin to model.

The Blue Brain Approach

- Databasing heterogeneous, multimodel data
- Building 10,000 morphologically complex neurons
- Constructing a circuit with 30,000 dynamic synapses
- Simulating 3.5 million compartments

Note: The supercomputer is involved in all the above steps.

- Validation: expert in the loop

Building Cell Models

- Data-driven constraints
- Genetic algorithm implementation in NEURON
- Inherent generation of electrical cores
- Model management to deal with morpho-electric classes

When we try to connect the cells we move to needing almost the full HPC system. For a second of biological time on one parameter (e.g., electrical potential), the simulation generates 150GB of data.

- Fully parallel setup and simulation on BG/L using NEURON
- Load balance through a new fully implicit solver to parallelize multi-compartment neurons
- The I/O challenge is addressed through a dedicated output library using MPIO and a new framework, Neuodamus, to abstract the compute engine and do the online analysis.

Per neuron, we run ~20,000 differential equations. We can go to 1 million cells on Blue Gene/L. We want to explain and integrate more detailed models of Monte Carlo molecular diffusion and reaction.

CADMOS Blue Gene/P has 4 racks with 16,384 processors, 16TB of distributed memory, 56TF, a 1PB file system, and 10GBps of I/O bandwidth. It enables cellular neuron simulations up to a size of 15 million neurons (larger areas of brain tissue). It enables molecular-level simulations of detailed single neurons. This is a test bed for a multiscale simulation framework and for online analysis.

Moving toward whole brain models, the detailed challenges include:

- Massive data management
- Massive simulation
- Massive visualization
- Massive analysis

Dave Turek, IBM: "Motivation for HPC Innovation in the Coming Decade"

The high-performance computing trends are 1PF in 2008 (achieved), 10PF in 2011, and 1EF sometime in the next decade. Most advances in the past have been due to CMOS improvements. Today, it's more about growth in parallelism. The current world as we know it will end in terms of computing architecture. We have to worry about more challenges going forward. I have a list of eight things with no solution to them today.

No single application is driving the pursuit of exascale computing. There are exascale problems in many domains, and many of these are multiscale problems. There are also exascale problems in business, especially streaming applications for real-time analytics. The initial motivations (for exascale computing) came out of homeland security concerns involving unstructured data and the need to intercept in real time.

Computer design challenges include:

- **Core frequencies.** It will take 100 million cores to deliver an exaflop. You'll see the frequent failures of cores. Also, how do you manage a system like this?
- **Power.** In 2005, Blue Gene/L had 400TF with 2MW of power. Not just the microprocessor is generating power, but I/O and also the memory subsystem.
- **Memory.** In 2011, I'm on the hook to deliver a system where memory cost alone is $100 million. Innovation in processor speeds has outpaced innovation in memory and memory subsystems. Memory per core will decrease because maintaining the same ratio is too expensive.
- **Network bandwidth**
- **Reliability**

- **I/O bandwidth.** Production of data versus ingestion of data will reach a point where you simply can't checkpoint and restart a system at this scale.

Our prototype design for exascale is to do it for 25MW and be highly reliable and manageable, all within nine years. We expect 2014 systems with 200–300PF. If other manufacturers don't make big changes in their architectures, they'll hit walls before this date.

Paolo Masera, Altair Engineering: "PBS Works 10.1"

Altair is based in Detroit. PBS was born in the 1990s at NASA and was introduced in 2003. Altair has had a 22% CAGR over 15 years.

Altair's HPC vision is based on three principles:

- **Ease of use.** Vertical portals and an SaaS gateway with well-defined interfaces. Engineers should focus on engineering, not computer science.
- **Reduce risk to ensure business continuity**
- **Resource optimization.** Our product is enabled for cloud computing, business policies, and optimized use.

The PBS Works suite is built around PBS Professional. There are two PBS Professional portals, e-BioChem and e-Render. We also provide green provisioning, the ability to shut off the machine or parts of the machine when you're not using it. PBS Calatyst and PBS Analytics are also available. PBS Professional is a workload manager.

PBS Works 10.1 features include a turnable scheduling formula, green computing support, submission filtering hooks, standard reservations, standards-based meta-scheduling, PBS Application Service, individual user and group limits, and high availability for advance reservations. This ensures that reservations succeed even if there are failures.

PBS Catalyst is application-aware. It features drag-and-drop input for submission and increased user productivity. You can monitor, manage, and prioritize jobs; create profiles for common runs; and connect to multiple PBS servers.

PBS Analytics generates reports automatically out of the box that you can then customize. You can understand usage trends for capacity planning, verify project planning assumptions, and extract accounting data for billing. PBS Analytics License Tracker is for data collection, license monitoring, and data storage.

Jack Collins, National Cancer Institute: "Applying HPC to Biology: The Digital Age"

One problem in biology is a translational problem. When you write something in computer science terms, most biologists can't understand it, and vice versa.

NCI's Advanced Biomedical Computing Center (ABCC) provides the HPC for the NCI and other institutes and groups in the federal government. We provide a lot of the

computational infrastructure and domain experts to enable people to use the computational tools.

Our largest driving goal not is not gigaflops or processors but how many lives we can save. There is a paradigm shift in biology that's generating terabytes and petabytes of data. We should be able to drive mathematical models that can start to impact the experiment and save significant money. Biologists don't need to know how to use the latest programming language; they need new algorithms. If we can have better algorithms that are much more powerful and efficient and let us work smarter, we won't need a million processors. We'll need only a fraction of a big number like that.

NCI's vision for translational research is based on data-driven computation, where integration and understanding are key. To get to regulatory networks, protein pathways, and systems biology, you need to do a lot of integration of data. The real goal is clinical outcomes.

Examples include:

- Next-generation sequencing technologies use high-throughput sequencers, where the output from one illumine paired-end run generates 7TB of raw data from one machine. But what you get is a whole farm of machines, so that creates a real data problem. Today, we generate 20MB of data per hour, and that will go to 2,500MB per hour by 2015. This creates multi-petabytes of data to store. Most of this data doesn't map to a genome. We'll need to map all the data to all the genomes to find out where it goes, but we don't know how everything works. Then we have to find out where all the differences (SNPs) are, and then we have to find out what all the polymorphisms do. We need all your data, including time series (historical data on the person), and we need to get this into a doctor's head. We need to map all this data so researchers can use it to do their experiments in the most effective way.

- The Cancer Genome Atlas: On October 1, it was announced that $275 million of stimulus money is being allocated to start sequencing all the cancer genomes. This is a great idea scientifically. With 600GB/patient/disease, 500 patients/disease, 300TB of data/disease, and 20 cancer types, that generates 6PB of primary data that we also need to annotate, integrate, and analyze for patterns. No one will hand-enter this data. There needs to be text processing, etc., and no one's solved this artificial intelligence problem yet.

- I don't just want results. I want to see the relationships between my results based on ontologies and other metrics. We need to move beyond just getting information into and out of databases. We need to understand it so we can impact the lives of people.

- ABCC has been about analyzing high-dimensional data for a number of years. Things are getting a lot better, but the problems are getting very complex and hard to define. This requires very good people on the computing side. We need people in the large database field and very good computers. We need appropriate computing platforms (memory, multi-core, Cell, FPGA, GPGPU, and maybe other things). Also, we also need to be able to verify we have correct code.

- NCI started funding purely *in silico* centers. They are funding five centers so we can mine and analyze data without being tied to a specific experimental group. We are looking at parts of the genome that are somewhat perplexing. In the genome, structure is very important. We need to know what SNPs do and why. The ABCC does a lot of confocal and other imaging.

- What we really want is to know *a priori* what I'm creating before I go into the lab and create it. I want to be able to do this in the full protein.

- In the imaging world, there's digital pathology, where for example I take a tumor slice and generate the image I need. As the camera resolution goes up, it's along a log2 scale. With all the angles and cameras out there, it's 12TB per image. That lets me see a lot of protein states, but now I need to know which are the high-energy states. I need massively parallel systems to compute problems like these.

- In structure, non-intuitive results explain toxicity.

- My view of the HPC "compute cloud": NIH is starting to gather requirements. I think virtualization will have a bigger impact near term than the cloud. I don't care where people run my problem. I just want to run it and have the results returned to me. I think my chances are better with virtualization.

- I need improvements in compute, storage (I need a lot here!), the network, and turning information into knowledge. I need to distribute data securely, and I also need to access national resources.

In 2008, someone died of cancer every 56 seconds. When I say virtualization, I mean I put together a system and if I ship it somewhere else, it unpacks itself and runs well.

Frederic Hemmer: "CERN and High-Throughput Computing"

We are trying to better understand our universe, which has been expanding and cooling down since the Big Bang. More than 95% is unknown stuff out there that doesn't interact with matter in the ways we understand.

Fundamental Physics Questions

- Why do particles have mass?

- What is 96% of the universe made of?

- Why is there no antimatter left in the universe?

- What was matter like during the first second right after the Big Bang?

The Large Hadron Collider (LHC) is like a microscope that goes back in time, close to the Big Bang. CERN was started in 1954. In 2008, more than 100 countries were involved in the LHC.

CERN has 2,300 staff, 700 fellows and associates, 9,500 users, and a budget of 887 million Swiss francs (595 million euros). The budget comes from the member states. Other nations are "Observers to the Council." The largest single user community is an observer, the United States.

CERN's Tools

- The world's most powerful accelerator, LHC
- Very large sophisticated detectors
- Significant computing to distribute and analyze the data
- A grid linking together about 200 computer centers around the world
- Enough power and storage to handle 15PB of data per year, making this data available to thousands of physicists for analysis

The detectors, CMS and Atlas, are discovery machines for new physics. Atlas when running produces 1PB of data per second. Our computing problem is that at each beam crossing, several protons collide.

The LHC computing grid in Europe is the world's largest grid. EGEE is its name. It has 17,000 users, 136,000 CPUS, and 38,000 disks.

Future Direction

- Grid will turn into something that coordinates clouds, using virtualization to provision services.
- Commercial cloud offerings can be integrated for several but not all types of work. They aren't good for simulations or compute-bound applications.
- Workflows will be multidisciplinary and complex.
- Today's middleware is complex, with a large support effort needed.

Sustainability

- We need a permanent, common grid infrastructure.
- We need EGEE to ensure a common infrastructure to be used by research teams, institutions, and others.

Robert Singleterry, NASA

The NASA Ames system has 51,200 cores and uses Lustre. There's a duplicate system at Langley, except with 3,000 cores, so I can solve smaller problems on Langley and, if I need to, run larger versions at Ames. Goddard has 4,000+ Nehalem cores but is running GPFS, so we can't move problems to here. There are smaller resources at other centers.

We can solve science applications and engineering applications, such as CFD for the Ares9I and Ares-V, aircraft, Orion reentry, space radiation, structures, and materials. A lot has to be done numerically, no long physically. We don't have all the wind tunnels we used to have.

NASA Directions from My Perspective

These views are my own and don't represent NASA or NASA Langley.

In 2004, Columbia had 10,240 cores. In 2008, it went to 51,200 cores. So by extrapolation in 2012, we should have 256,000 cores and by 2015, it should be 1,280,000 cores. That's 5 times more cores every four years.

Assuming that power and cooling were not issues, then what will a core be like in next seven years? Will it be powerful like Nehalem and not so many in a system, or many, less powerful as in Blue Gene or Cell? Will it be like a CPU, or something completely new?

Each of the four NASA Mission Directorates owns a part of each large system. Each center and branch resource controls their own machines as they see fit. Someone from Goddard can't use our machines without special permission. Queues limit the number of jobs we can access and also the time any one job can have. So, this looks like a time share machine of old. So in 2016, how many cores can my job get? Do I have an algorithm that can exploit more cores in 2016? An algorithm that uses 2, 000 cores in 2008 would have to use 50,000 cores in 2016. I'd have to recode for this. Are we spending money on the right things? Why spend on better hardware when I can't use it? It's probably better to spend more on software development. Another question is whether we as researchers understand the perspective of the NASA funders.

Case Study: Space Radiation

Particles impinge on Earth's atmosphere and interact with airplanes or a spacecraft wall. They hit and break apart. They don't change speed, they change energy, and we can do things about this, such as add more shielding or develop better absorption with new materials.

Electronics are now intermediaries for steering spacecraft, etc. Astronauts don't do this directly anymore, they hit a pedal that tells the electronic system to do it. So, we have to be careful how we shield electronics.

With previous space radiation algorithm development, you'd design and build spacecraft, and then bring in radiation experts to analyze the vehicle by hand (not parallel) and fix shielding problems. Excess or the wrong type of shielding around a scientific instrument can throw off the science (e.g., a CCD camera in spacecraft).

Today's algorithm development is different. You do a ray trace of the spacecraft/human geometry. You use a transport database, which is mostly serial. A 1,000-point interpolation is parallel. Integration of data at point is parallel. At most, it exploits hundreds of cores and does not scale beyond this.

So now we're going to run the transport algorithm along each ray, each independent of the others. At 1,000 rays per point, 1,000 points per body, a 1 million–element transport runs at 1 second to 3 minutes per ray and point. The integration of data at the points is the bottleneck. The process is parallel if the communications bottleneck were not an issue.

Future space radiation algorithms:

- **Monte Carlo methods:** Data communication is the bottleneck.
- **Forward/adjoining finite element methods:** Phase space decomposition is key.
- **Hybrid methods (the best of first two methods):** This holds promise for better using future HPC systems (on paper, anyway).
- **Variational methods:** It's unknown today how well this would exploit future HPC.

Summary:

- Current space radiation models are not very HPC friendly and don't scale well. Is this good enough? No, because we need the scalability. Funders want new bells and whistles all the time, not just moving it to the HPC world.
- Future methods show better scalability on paper, but we need resources to investigate and implement.
- NASA is committed to HPC, but my concerns are whether we buy machines that can run our algorithms well or do we need to rewrite the algorithms?
- We have no HPC help desk to work with users to achieve better results for NASA work, such as the HECToR model in the United Kingdom.

Friday, October 9

Welcome and logistics: Earl Joseph and Steve Finn, BAE Systems, summarizing the September 2009 User Forum

Victor Reis, U.S. Department of Energy: "HPC, the Department of Energy and the Obama Administration — A Strategic Perspective"

We now call it extreme computing rather than HPC.

Strategic planning: You start with a vision of where you want to be, and when. You look at the assets and tie them to the vision with a strategy. In a successful organization, you line everyone up behind the strategy.

President Obama gave a key speech on April 5 on the nuclear aspects. The Department of Energy (DOE) consists of large science and large computing. I've been trying to get people to see these are part and parcel of the same thing. There are two organizations within DOE: Advanced Simulation and Computing is part of the National Nuclear Security Administration (NNSA), and the Advanced Science and Computing Research group is part of the Office of Science.

President Obama's nuclear vision involves going to a world with zero nuclear weapons and until then maintaining a safe, effective arsenal. We are not going to go back and test. We will also use nuclear energy to combat climate change and advance peace.

On April 27, President Obama doubled funding for variety of things, including supercomputers. On June 27, he talked about energy independence and climate change. On September 22, he gave a speech at the United Nations on the urgent need to deal with climate change.

DOE has done a lot in the past 15 years. It was a mess when I got there in 1993. The Cold War was over, the superconducting supercollider was canceled, and the Advanced Neutron Source was canceled. In 1994, the Republicans won the election and promised to get rid of the DOE. After that, the DOE limped along.

In the past 15 years, there have been many new major science facilities, such as the Spallation Neutron Source, the Advanced Photon Source, and the National Ignition Facility. For nuclear weapons, the idea was to test virtually, without live underground testing. We cleaned up major assets such as Rocky Flats.

DOE has 12,440 Ph.D.s and 25,000 visiting scientists. The largest group is in the NNSA. Today, 8 of the top 10 HPC systems are in the United States and five are at DOE sites.

Accelerated Strategic Computing Initiative (ASCI)

In 1993, President Clinton established the nuclear weapons stewardship program. Simulation became part of this. Robust, balanced, and long-term funding was an important goal. Funding for advanced computing was $10 million. It got to be about $600 million a year. ASCI was built on partnerships among government, academia and commercial firms.

There are three models for HPC development:

- Science model: The Office of Science does this and their program is SciDac.
- Applications model: DOE NNSA-ASCI/ASC. Here the focus is on a time-urgent national problem. It's a closed community. DOE invests in the development of the computers. This requires much bigger investment and much tighter management. How applications-driven with the next generation of HPC architectures be? What are the important DOE applications? Who cares? How much performance, when? What should the scale of the investment be? What about the breadth of applications? What is the model of the DOE HPC program? What level of DOE/industry partnerships should there be? What's the commercial spin-off?
- The commercial model is used by the private sector.

In response to the 2008 presidential election, DOE held a series of workshops, most sponsored by the Office of Science, on grand challenges and a variety of other things. Bill Brinkman's presentation was on HPC focal areas: nanoscience, energy science, and others, plus national security.

The overarching ASC goal is to provide the SSP [Strategic Systems Program] with a sufficient science and technology base to make decisions.

Summary

- Solving current DOE problems could provide a major opportunity for HPC.
- DOE problems are in climate, national security, energy, and science.
- Historically, computer architecture drives the rate of growth.
- The DOE model of development is a balance between science and applications.
- The opportunity for Obama Administration and DOE involves the role of DOE labs, academia, and industry.

Alan Gray, EPCC: "HPC in the United Kingdom — An Update"

The Edinburgh Parallel Computing Centre (EPCC) was founded in 1990 by the University of Edinburgh as the focus for its interest in simulation. Today, EPCC is a leader center for computation science in Europe and manages both U.K. national services, HECToR (Crary XT5) and HPCx (IBM Power 5 server).

HECToR (which stands for High End Computing Terascale Resource) was started in 2007 and supports a wide variety of university applications and some industrial ones. The HECToR service is located at the University of Edinburgh and includes a Cray XT5h system, which is a hybrid of a Cray XT4 and a Cray X2 vector system. The Cray XT5h uses quad-core Opteron processors. The next upgrade will be in 1Q10 to a Cray Baker system of 20 cabinets with a new AMD chip. On a per-core basis, the application performance will decrease, so the apps will need to scale. In 4Q10, we plan to upgrade the Cray Baker system to the Gemini network, and it will jump to 24 cores per node. For Phase 3, the vendor and system are unknown.

HPCx was the previous main U.K. service. U.K. policy has been to have overlapping HPC services. HPCx is due to end in January 2010. It's located at Daresbury and uses 160 IBM e-Server p575 nodes.

For the past two years, we've run both services simultaneously. People who have long jobs, interactive jobs, and advanced reservations go to HPCx. We treat HECToR as the main HPC facility and HPCx as our "national supercomputer."

Case Studies

- Environmental modeling using HIGEM involves U.K. Met and seven U.K. academic groups. Its goal is to develop an advanced Earth System model able to perform multi-century climate simulations.
- Computational materials chemistry is the highest user of the HPCx system and might be highest for HECToR, too. The work is on catalysts and biomaterials:
  - Biomaterials: Fundamental factors related to bone structure, useful for future disease prevention
  - Nanomaterials and nucleation

- Fractal-generated turbulent flows: New industrial fluid flow solutions are urgently needed in the automotive and aerospace industries for fuel savings and reduced environmental impact. Using fractal grids is a very effective approach and is impossible without HPC. The first-ever successful simulations of turbulence from fractal grids were performed on HECToR in 2008–2009.

- Interactive biomolecular modeling aims at further understanding of "ion channel" proteins in the nervous systems. They control nerve signals. The work is done on HPCx and could lead to better drugs and treatments for people suffering from certain diseases of muscle, kidney, heart, or bones. The method is computational steering.

- FireGrid is a next-generation emergency response system. Better information leads to more effective responses (e.g., World Trade Center). The idea is to have an intelligent network of sensors in your building that sends signals to a command-and-control center. It must be faster than real time so you can use the simulation results. You would use the information and simulations for decision making leading to more-effective responses.

HPC systems are getting more complicated to use. Scientists and engineers don't want to worry about HPC systems. We're collaborating with scientists and engineers to make HPC easier to use.

Jim Kasdorf, Pittsburgh Supercomputer Center: "National Science Foundation Directions"

NSF 2005 announced Track 1 and Track 2 procurements. Track 1 provides for a petaflop system in 2011 at $200 million. Tracks 2A, 2B, 2C, and 2D each are for one year each at $30 million per year.

Track 2A, in 2006, awarded $59 million to TACC for the "Ranger" Sun Opteron InfiniBand capacity system with a final peak of 529TF. Track 2B went to the University of Tennessee.

The National Institute of Computational Science at ORNL is going from Jaguar today to Phase 2, a 1PF NSF Cray system in the second half of 2009. Phase 3 might be a >1PF NSF Cray system in 2010.

NSF is investing in a number of projects:

- **SDSC:** An Appro Intel ScaleMP with flash memory and 200TF peak.

- **Georgia Tech:** Initially an HP + NVIDIA Tesla system, then in 2012 new technology with 2PF peak performance

- **An award for a future grid to Indiana University:** For a test bed for a network allowing isolatable, secure experiments

Track 1: There was an award to NCSA with rumored specs being an IBM Power 7 with 38,900 eight-core chips, 10PF peak performance, and 620TB of memory.

TeraGrid Phase III includes designing an XD grid architecture; two awards were made, to TACC and to the University of Tennessee.

What's next: NSF Advisory Committee for Cyber Infrastructure (ACCI) has formed task forces on various topics. The task forces also includes members from other U.S. government agencies.

Thomas Eickermann, Juelich Supercomputing Centre, PRACE Project Update

PRACE stands for Partnership for Advanced Computing in Europe. Supercomputing drives science through simulation and is needed in all areas of science and engineering. It addresses top societal issues as defined by the EU.

This began with meetings of European scientists under HET (2006) to establish the HPC needs of science and engineering. The report went to the European Commission. This was a key impetus for starting PRACE.

The United States is far ahead of Europe in HPC based on Top500 rankings. 91% of European HPC power is represented in the PRACE countries. The vision is to provide world-class HPC systems to support world-class science in Europe to attain global leadership in public and private R&D.

The mission is to create a world-leading, persistent high-end HPC infrastructure, and to deploy 3–5 systems of the highest order. The first PF system in Europe was installed at the Julich Research Center. The goal is to ensure a diversity of architectures. 2008 was the PRACE preparatory phase. The PRACE implementation phase is 2010–2012, with the operational phase starting in 2013.

ESFRI is the European Roadmap for Research Infrastructure, an advisory group for the EC.

Fourteen member states joined the PRACE preparatory phase, which runs from January 2008 through December 2009. The objective is to perform all the legal, administrative, and technical work to create a legal entity and start providing Tier-0 HPC services in 2010.

We've procured some prototype systems that we're currently evaluating. Software for petascale is one of the biggest challenges. We're also preparing a package for joint procurements in the future and developing a procurement strategy for 2010.

Applications should be representative of European HPC usage. We surveyed PRACE partners' systems and applications (24 systems, 69 applications). We developed a quantitative basis for selecting representative applications. We disseminated this in our technical report.

The representative benchmark suite has 12 core applications plus 8 additional ones. This is now finalized. There are applications and synthetic benchmarks. This is called the Juelich Benchmark. We looked at how well they would run on different architectures. Most run well on MPPs and clusters. Some still run best on vector architectures. There is also serious interest in porting some codes to Cell and other

accelerators. We looked at IBM Blue Gene, IBM Power 6, Cray XT5, NEC SX9, and Intel clusters.

We analyzed recent European procurements with the goal of developing a general procurement process for future use.

What comes next? We're nearing the end of the preparatory projects. Contracts for the legal entity are in final negotiation, with signatures planned for 2009. The Tier-0 infrastructure will become operable in the first half of 2010. The access model will be based on peer review: We are targeting "the best systems for the best science."

The vast majority of funding (90%) is from national money, and the rest from the EC. So we need to monitor the number of projects from each country and watch for any major imbalances.

Panel on Using HPC to Advance Science-Based Simulation

Panel Moderators: Henry Markram and Steve Finn

Panel members: Jack Collins, Thomas Eickermann, Victor Reis, Markus Schulz (CERN), Neil Stringfellow, Henry Markram, Felix Schűrmann, Paul Muzio, and Charles Hayes (CHS)

Markram: I want to pose a challenge to the panel. I'm a biologist trying to simulate the human brain. As a user, I realized a while ago you can't just say, give me the HPC and let me do the job. The biggest problem we'll face in simulating the human brain is that it's not just a hardware problem. We'll get the hardware. You can't just expect the software to catch up or you'll get exascale systems that are very inefficient. We need to develop software in tandem with hardware. Can we bring all the disciplines together, from biology to astrophysics, to help develop the exascale hardware? Can we form an international consortium around the software agenda?

Muzio: The hardware we're using is not well suited to HPC. We're starting with the wrong building blocks. The HPC community is not large enough to get the chipmakers to change directions. In the past, my colleagues did work with PGAS models that are a lot easier to use in parallel apps, but we can't get anyone to support this in hardware. The last system to do this was a Cray X1. I think Cray had too limited resources to accomplish the design.

Comment: Aren't the Gemini chips going to support PGAS?

Muzio: Yes, but that's only one vendor.

Schűrmann: HPC systems today leverage DRAM or flash memory, neither of which fits HPC requirements well.

Collins: We definitely should develop hardware and software together. If you think of the computer as a tool, I can't hire a team of Ph.D. scientists to make the machine run. Can we get people together to do this? Only if users form almost a union to lobby the funder, or we change the funding process. It's almost impossible to pool money for different things under federal rules.

Markram: I think industry's ready to do this because it will be $1 billion to get the software to work on future HPC systems. The vendors need to work with the biological community.

Reis: $1 billion is not a lot of money. We [the U.S. Government] just spent $3 billion on "cash for clunkers." I'm encouraged that if you give the vendors the money and the goal, they generally say, "We think we can get there." There aren't many vendors you could spend it on. The same people come to these meetings and have been around for years. It's a matter of pulling them all together. At a recent workshop, I asked people what they think are the top 3 problems for exaflop computing. [He showed a slide with survey results. The top results were general modeling/simulation, climate, fission nuclear energy, and nuclear weapons.]

Markram: Do I trust that the IBM exascale system will solve my problem in 10 years, or should I begin to work with others toward the right solution?

Reis: You need an urgent mission to generate excitement and funding. My approach has been to pick two or three goals. I think it's important for this effort to be international.

Markram: But you really want a machine designed to perform very well on your own applications.

Collins: If I have $1 billion to spend, will I spend it on application software or an interface that lets me optimize my code for whatever hardware is out there. I don't see a lot of tools that help me map my apps so they run on, say, 70% of the computer.

Singleterry: IBM and other vendors won't be able to supply this software. The effort has to begin with the end users.

Markus: That already failed once with the vector machines. Today's systems are parallel, with different hierarchies of memory and different latencies. You could invest $50 billion and not solve the problem. You need to train people better and form closer collaboration between scientists and engineers and the people who develop the hardware and software systems.

Schűrmann: No one sees a way past MPI at the moment. The hardware folks are leading the way forward without addressing the software problem.

Muzio: [He showed a hummingbird flight simulation.] This is a problem you can't do with MPI. The cost for Intel to develop the chip for a Cray system to do this is $30 million, but Intel will say there's no market for this. We can address things like this between now and exascale.

Stringfellow: There's a chicken and egg effect. At a Cray conference a few years ago, they had a long list of things to do, including Chapel. Unless we give up some of these things we won't get the basic problem solved.

Muzio: There are many applications that would benefit from moving-mesh technology, such as accurate heart modeling.

Markram: The brain is a very plastic thing that keeps changing its structure, so a structural model needs to have full detail plus a simulation model and these need to remain accurate almost in real time.

Singleterry: Unless you want to spend money on changing the hardware to match the algorithm, you'll need to change the algorithm to match the hardware.

Stringfellow: Commodity components will always be cheaper for some things, and beyond that you can add specific features to do certain things such as global address space. The very large machines are created in the United States between the vendor-partner and the site.

Hayes: It would be wonderful if you could get IBM to design a machine to simulate the human brain, but the market needs to be there.

Finn: U.S. government agencies publish a glossy book with the gaps in application performance and what it would take to close the gaps. Another issue is, How do you know your model is right?

Schűrmann: The user community has to be able to communicate its vision clearly.

Singleterry: How is this different from what NAG and others are doing now?

Schűrmann: We need to train our students to think more in these libraries and to abstract from them. If you have a big enough user base, maybe you can convince hardware vendors to help with the software.

Collins: We've tried to work with many of the new technologies, such as FPGAs and GPGPUs. Most people don't have the ability to play with these things for five years the way we do. Most places have a small cluster. If we don't get software designed better, we'll be using 5% of the new hardware systems.

Markram: Everyone seems aware of the massive explosion of memory. How will we manage exascale data volumes?

Singleterry: You'll have to do analysis on the fly, without a core dump or restart. Your software will need to weather hardware failures. You're making your apps so complex, I'm not sure it will be workable. Without restart, if your job fails before completion you have a real problem.

Collins: The amount of data we have to analyze is growing so fast we're creating cross-disciplinary teams to look at the algorithm designs and the potential need to redesign the algorithms.

Markram: Within a few years, the data explosion will be a nightmare. Within five years, scientists will have sequenced the genomes of 500–1,000 species.

Stringfellow: Is there any plan for non-lossless compression, where you could retain the main elements but not everything?

Singleterry: It depends on what you need. In NASA, we have to maintain all the real data.

Collins: In medicine, we can't compress it because if a doctor misdiagnoses because of missing information, this can result in a lawsuit. For some data, it's cheaper to run the experiment again.

Schűrmann: The amount of data you need to regenerate can be very large, though.

Collins: We already have the best compression algorithm for DNA. It's called DNA.

Markram: Can better compression algorithms be developed? It's amazing how quickly you can put terabytes on your machine today.

Finn: There's the potential for deduping, which is more prevalent in the business world today. For example, if you and I have stored the same data.

Stringfellow: If we dealt with all the data we generate, we'd be swamped. Should there be central or national datacenters to hold all this?

Schulz: In Europe, there is good ability to move the data around, but to store and manage a few petabytes of data is a whole different matter. National or European or science-topic centers for holding and managing data would be very popular.

Collins: NIH has a central repository and is asking if they should keep doing this because it could eat up their entire budget after a while.

Schulz: You have to budget a significant percent of your budgets for tape and other storage.

Victor Reis: Climate keeps coming up as an issue. [He showed a spreadsheet with a model he designed to try to explain climate modeling.]

Beat Sommerhalder, New Software Technology Directions at Microsoft

Microsoft's vision for HPC is based on reducing complexity, making HPC mainstream, and creating a broad ecosystem for HPC. Microsoft stepped into the HPC market in 2006. The driver was the multi-core and many-core strategy. Microsoft recognized that high-end HPC users were exploiting lots of processors. Microsoft created development tools that could work in parallel.

Reducing complexity means easing deployment for larger-scale clusters, simplifying management for clusters of all scales, and integrating with the existing infrastructure.

Hard problems include the following:

- Scaling distributed systems is hard.
- Data sets are increasing in size.
- Programming models are complex, and we need simpler models.

Multithreaded programming is hard today. Customers don't want to get deeply involved in the technical issues. They and their developers want to focus on their own businesses.

Crossing the chasm:

- Embrace existing programming models. MPI is important for Microsoft's target market.
- Increase the reach of existing codes (cluster SOA, .NET/WCF, Excel integration).
- Invest in mainstream parallel development tools (unlock multi-/many-core for the breadth developers; evolve hybridized and scale-out models).
- Seek opportunities for "automatic" parallelism (e.g., F#, DyadLINQ).

PLINQ is a parallel version of LINQ-to-objects. MSFT's Visual Studio 2010 is coming out and will include PLINQ. The goals are tied to developer accessibility, including the ability to express parallelism easily, and to simplify the design and testing of parallel applications.

Microsoft has restructured its whole HPC group to add development and language groups to the HPC group. Vertical targets differ by region. In Germany, they are FSI, manufacturing, and academia.

LEARN MORE

Related Research

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

- *Massive HPC Systems Could Redefine Scientific Research and Shift the Balance of Power Among Nations* (IDC #219948, September 2009)
- *The Second PRACE Industry Seminar* (IDC #220029, September 2009)
- *China HPC Directions and Trends Looking at Evolution of the China TOP100 List* (IDC #219952, September 2009)
- *Living in a Difficult Economy: 2009 and 2010 Growth Areas by Industry/Application Segments* (IDC #219869, September 2009)
- *The Race for the Fastest Computer Is Still On — Fujitsu's Petascale Project Plans* (IDC #lcUS21929009, July 2009)
- *Back-End Compiler Technology: HPC User Forum, April 2009, Roanoke, Virginia* (IDC #219119, June 2009)
- *I/O and Storage: HPC User Forum, April 2009, Roanoke, Virginia* (IDC #219121, June 2009)
- *HPC and Industrial Product Design: HPC User Forum, April 2009, Roanoke, Virginia* (IDC #219120, June 2009)

- *Petascale Computing: HPC User Forum, April 2009, Roanoke, Virginia* (IDC #219117, June 2009)
- *Alternative Processor Technology: HPC User Forum, April 2009, Roanoke, Virginia* (IDC #219118, June 2009)
- *HPC and New Energy Solutions: HPC User Forum, April 2009, Roanoke, Virginia* (IDC #219122, June 2009)
- *IBM and Cray Battle for Having the Fastest Computer in the World* (IDC #lcUS21902709, June 2009)
- *Worldwide Technical Computing Server 2009–2013 Forecast Update* (IDC #217881, May 2009)
- *Rackable Systems Acquires Booster Rocket* (IDC #lcUS21774909, April 2009)
- *Worldwide Technical Computing Server 2009–2013 Forecast* (IDC #217232, March 2009)
- *IDC's Worldwide Technical Computing Server Taxonomy, 2009* (IDC #216847, February 2009)
- *Petascale Supercomputer Sales Continue to Grow: Cray Announces Largest Revenue Year Ever, with 52% Growth* (IDC #lcUS21683709, February 2009)
- *IBM Sells First Petascale Supercomputer in Europe* (IDC #lcUS21683809, February 2009)
- *IBM Advances the Frontiers of Computing Power with 20 Petaflops DOE Order* (IDC #lcUS21656709, February 2009)
- *Economic Crisis Response: Worldwide High-Performance and Technical Computing Server 2008–2012 Forecast Update* (IDC #216022, January 2009)
- *Worldwide Technical Computing 2009 Top 10 Predictions* (IDC #216296, January 2009)
- *Petascale and Exascale Directions in HPC: From the HPC User Forum Meeting, September 2008 — Tucson, Arizona* (IDC #215657, December 2008)
- *IDC Predictions 2009: An Economic Pressure Cooker Will Accelerate the IT Industry Transformation* (IDC #215519, December 2008)
- *The Need for a U.S. and Global Economy Simulator, and Perhaps a New U.S. Government Agency: NEAA* (IDC #215700, December 2008)

Copyright Notice