IDC
Analyze the Future

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

## UPDATE

# October 2009 HPC User Forum Meeting Notes: Stuttgart, Germany

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

## IN THIS UPDATE

This IDC update covers the 34th High-Performance Computing (HPC) User Forum meeting, which took place at the University of Stuttgart's High-Performance Computing Center (HLRS) on October 5–6, 2009. Local hosts for the meeting included HLRS Director Dr. Michael Resch, HLRS Head Uwe Kuester, and HLRS Administration Manager Agnes Lampke. The principal tasks of the meeting were to:

- 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 Germany's automotive industry
- Survey the activities and achievements at leading university HPC centers

## Monday, October 5

### *HPC User Forum Welcome/Introductions: Steve Finn (Chair, HPC User Forum) and Earl Joseph (IDC)*

Steve Finn welcomed attendees.

Earl Joseph thanked Altair, Bull, and Microsoft for sponsoring the meeting. He explained the history and purpose of the HPC User Forum (mission statement and goals).

### 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 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 management software. We're also closely tracking extreme computing, datacenter

Filing Information: December 2009, IDC #220795, Volume: 1
HPC User Forum: Update

assessment, and benchmarking and tracking petascale and exascale initiatives around the world.

### *Michael Resch, HLRS: A European View of HPC and Other Remarks*

This is our third time hosting an HPC User Forum meeting. At these meetings, we try to have a U.S. view and a European view. We plan to hear a number of updates over the next two days on how HPC is evolving in Europe. Thomas Eickermann will be here from the Jülich Research Center. We're also happy to have an automotive industry speaker here, Dr. Schelke from Porsche and the Automotive Simulation Center.

Very few of us are interested in HPC only for its own sake, nor are most of our funders. The larger interest is in solutions to important problems like curing cancer. Most money goes into research and some into computational science and engineering. To grow HPC these days, you have to show how it can help grow the economy. Everyone in Germany believes building cars is important because Germany makes a lot of money doing this, but also that the environment is important.

There are a number of interesting European projects in HPC:

- In Stuttgart, we focus on HPC. A few years ago, we set up the Simulation Technology activity, working with industry partners, and we launched the Automotive Simulation Center.
- **PRACE.** The idea is to create a center that can represent HPC in Europe.
  - ❑ The good thing is that HPC is on the European agenda.
  - ❑ The real challenge is to show some meaningful results in the form of achievements for society or research or industry that are due to the existence of PRACE.
  - ❑ Within PRACE are the "principle partners": the United Kingdom, France, Spain, Germany, and the Netherlands.
  - ❑ In Germany, the Gauss Center for Supercomputing is an outcome of PRACE. The Gauss Alliance includes the three national centers plus the regional centers collaborating through this. This is a very coherent German strategy.
  - ❑ PRACE will implement an initial solution by the end of this year. Currently, the idea is to have the headquarters at Lisbon.
  - ❑ If we can do these things, the future could be very interesting.
- **DEISA** began in the Grid era. The idea originally came from IBM. It's an operational grid of HPC systems that includes common standards (e.g., file system). It took three years to become operational.
  - ❑ Now, we have DEISA II. It may be useful if at some point in the future, PRACE and DEISA merge to become even stronger.

- DEISA has a lot of centers, both big ones and very big ones. DEISA could create a link between the very high-end European centers and midrange and smaller centers.

- **HPC-Europa II.** They take a pan-European view but with a different approach. The goal is to have these 27 countries be able to access all of the systems.

  - It's not only about network access, but it's also about getting the full support of the centers, for example, by bringing researchers to the centers for stays of four weeks to a few months, especially for users that don't normally have access to larger HPC systems. They then have established long-term relationships with these centers that are useful when the researchers return to their countries.

We need to find a European position within the framework of the HPC market. One way is to build a European supercomputer system, but that takes a lot of money.

- The hardware architectures of today are basically un-programmable. At the midrange, you find a lot of industry HPC users. According to IDC, few ISV applications scale beyond 128 cores.

- A question is whether Europe should take the lead in addressing this problem and spend a lot of money on software, not for just the high end but for the larger middle layer.

**Comment:** Will a high-end program like the HPCS program in the United States set the direction for the mainstream HPC industry or go off on its own?

**Speaker:** It will be hard for it to be successful because if I'm Intel I don't care about more horsepower for DARPA but about lowering the cost of a laptop. HPC today is like auto industry a few decades ago, where it was concerned about horsepower.

**Comment:** In the United States, it's hard to find students prepared to write new algorithms and software. What about Europe?

**Speaker:** European students are generally better prepared in that respect.

**Comment:** I'm worried about the future of HPC if the memory and system connections can't keep up with increases in core and node speeds.

**Speaker:** The true rule is, "don't communicate" except when you have to. We have to integrate more or find ways of reducing communications, which means we have to become more asynchronous and do much more structured communications.

### *Robert Singleterry: NASA HPC Directions, Issues and Concerns: A User's Perspective*

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 five 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 and so forth. 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 one second to three 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.

**Comment:** You didn't talk about the need to certify rewritten algorithms.

**Speaker:** That's a major issue I don't have time to go into.

### *Tom Sterling, Louisiana State University: Trends and New Directions in HPC*

We've been on an S-curve of evolutionary change. My take-home message is that we're about to experience a change.

A year ago, I was waiting for the power in my house to come on for two weeks because Hurricane Gustav was around. Simulating a hurricane storm surge using HPC is a critical thing. [He showed a slide with projections four days before Gustav hit, showing very different potential paths.] We need better computation. We need to be able to simulate problems like this at the mesoscale levels. We need exascale computing to integrate this into the climate modeling problem.

HPC is in a phase change — every 10 years or so, the technology pushes through the punctuated equilibrium. We've had five of these phase changes so far: sequential instruction execution, sequential instruction issue, vector (pipelined arithmetic, registers, memory access, Cray), SIMD (MasPar, CM-2), and CSP (Communicating Sequential Processes = MPI), which means killer microprocessors (MPP, clusters, and MPI).

### The Four Principle Challenges in Computer Architecture

- Starvation: Insufficient parallelism
- Latency (access to remote resources)
- Overhead
- Waiting (contention for shared resources)

We're in midst of a paradigm shift in organizing computing architecture, programming models, OSs, and runtimes.

The news from the DARPA ExaScale Technology Study is challenging to the user. [He showed a concurrency chart.] In 2020, we will have between 100 million–way (heavyweight cores) and 1 billion–way (lightweight cores) concurrency. Well, billion-way may turn out to be light by one order of magnitude.

Strategic requirements include scalability, efficiency, storage capacity, power consumption, reliability, and programmability.

### Exascale Design Points

We can't get to exascale with linear extrapolations of the current architecture, including for power consumption reasons. What exascale will likely look like:

- 22–11 nanometers
- 10 billion–way parallelism
- 100 million to 1 billion cores
- 1–2GHz
- 10- to 100-way concurrency per core
- 100s of cores per die
- 3D packaging
- Global address space without cache coherence
- Optical bandwidth of 1Tbps

If we follow history and have a paradigm shift, then what's the role of the new model of computation? This is not just an intellectual exercise. It's a critical tool for effective exascale applications to address the challenges imposed by technology changes. A new model of computation is the system, the machine.

MPI will continue to evolve and be valued, but neither it nor UPC will bring us to the level to exploit these future architectures. Newer models will overlap with MPI, and each will be used for certain things and certain parts of codes.

### *Wolfgang Dreyer: New Software Technology Directions at Microsoft*

ISC09 was a great success, with 119 exhibitors and 1,670 attendees versus 1,375 in 2008: 17% of attendees were senior management, 27% management, 29% professional experts, and 27% staff. The move to Dresden was absolutely necessary. For 2010, we expect similar year-over-year growth. The 25th anniversary will be in 2010. There will be a half-day session on HPC in the life sciences; sessions on evolving markets (China, the Middle East, and Russia); fault tolerance for many-core systems; future architectures; and high-performance computing and networking, storage, and flash technology; Linpack — reducing runtime; innovative applications; and exascale computing.

### *Jim Kasdorf, Pittsburgh Supercomputer Center: National Science Foundation Office of Cyberinfrastructure*

NSF 2005 announced Tracks 1 and 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 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 529TFLOPS. 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 1PFLOPS NSF Cray system in the second half of 2009. Phase 3 might be a >1PFLOPS NSF Cray system in 2010.

NSF is investing in a number of projects:

- SDSC: An Appro Intel ScaleMP with flash memory and 200TFLOPS peak
- Georgia Tech: Initially an HP + NVIDIA Tesla system, then in 2012 new technology with 2PFLOPS peak performance
- An award for a future grid to Indiana University (This is 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 POWER7 with 38,900 eight-core chips, 10PFLOPS peak performance, and 620TB of memory.

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

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

## *Erich Schelkle, ASCS/Porsche: HPC Applications for the Automotive Industry*

### The Automotive Simulation Center Stuttgart

The Automotive Simulation Center Stuttgart (ASCS) will be run as a nonprofit association and is the second of three planned competence centers in Germany. Its role is to develop numerical simulation methods for the automotive industry, with attention to benefiting the environment. It will combine competencies in engineering science, computer science, and business.

It includes OEMs and suppliers, HPC providers, universities and research institutes, and ISVs. It differentiates between capacity and capability and commodity production versus computational research. HLRS is responsible for the IT for HPC, SimTech for numerical fundamental research, and ASCS for transforming this into automotive science. The focus will be multidisciplinary optimization, optimization of the mixture formation, and hybrid and pure electric vehicles.

### History of HPC in CAE, 1970–2000

- 1970s: Mainframes, academic studies, and failure analysis
- 1980s: Supercomputers, usage grows, and simulations get much more complex
- 1990s: CAE becomes a mainstream tool on workstations/PCs; better efficiency and accuracy

### History of CAE Safety Research

- Crash simulations are very important. The first full car crash at Porsche was performed in 1966 using a Porsche 904. Starting in 2005, in-house and consumer requirements, along with legislation, have been considered in crash simulations.
- Increasing demand on forecast quality: turnaround time, model size, and data and model handling. New "Panamera" simulation involves 3.5 million elements.

The "magic triangle" of product development includes:

- Increasing product complexity
- Increasing the number of derivatives [derivative products]
- Increasing number of car lines

To cut the development time of a new vehicle in half, the auto industry will have to replace hardware prototyping with virtual prototyping. In 1990, development time was 40–50 months. In 2005, it was 25–30 months. The goal in 2015 is 18–24 months.

### The Current Status of CAE

- Concept development is almost entirely dependent on digital prototypes to avoid the cost/time of physical prototypes, which come later.

- Four stages: sourcing and assembling, setup, execution, and analysis. Real and virtual testing both occur at all four stages.

- Porsche uses 14 digital models, 5 for physical layout (e.g., full-car thermal model), and the rest for geometrical layout. We are planning 5–7 new digital models.

- Our hardware concept requires three computer platforms with different demands for each: the supercomputer (highly scalable, long runtimes), big workstations (I/O intensive, large memory, and limited scalability), and desktops.

- CAE has contributed significantly to shortening development times. The new Panamera took 16% less time than the 1997 G2.

- The FEM crash model during Panamera development increases at each phase of development.

- The number of simulation runs varies from 100 to 3,200, depending on the phase of development and which type of analysis is being carried out. For the Panamera, we did 4,300 test runs in total.

- Current CAE status (2,000s). We use high-power desktops and servers for simulation-driven product development and do parametric studies using multiphysics.

#### Directions and Challenges

- Front-load CAE methods

- Integrated data and analysis

- Extend the life cycle through the whole process

- Expand the use of HPC to suppliers

- Expand simulation to new methods (e.g., digital production and digital repair)

To simulate the experience of the future product, new visualization and real-time simulation technologies are being introduced. In 2010, concept design will be driven by analysis. We'll use Cell and multicore systems as our standard architecture for testing designs and ideas in the computer.

There are new demands on our CAE road map from new regulations, design/technology trends, tools for virtual development, and CAE software and hardware trends.

Challenges for future include mobility and transport, energy/environment, safety/security, and affordability/competitiveness.

### *Vijay K. Agarwala, Pennsylvania State University: Developing a Coherent Cyberinfrastructure from Local Campus to National Facilities: Challenges and Strategies*

We had a workshop last summer on how to build the computing pyramid and fund it from federal agencies, universities, industry, and what the feedback mechanism should be. A report was issued in April 2009.

The top layer consists of $200 million PFLOPS systems, with prices of $200 million for 100TFLOPS+ and $40 million for 5–100TFLOPS. So, there has not been much investment in the sector where 90% of the high-performance computing takes place. We need a balanced investment strategy, the committee concluded.

The goal is to enable individual researchers to use campus and national resources. This needs full support at each level. We need to do a better job of integrating national and campus resources. Funding should encourage resource sharing, and we need to develop standards for data and for data life-cycle processes. We need to make data a societal asset. Very little is invested today in the preservation of data.

We need a single, open, standards-based system for identification management, authentication, and authorization. We need a single identity that can move across different systems instead of having to have many user names and passwords.

We need to develop technologies and tools for the emerging cyberinfrastructure for education and scholarship. We need to make IP protection no longer a stumbling block.

Cloud computing: Will it disrupt the cyberinfrastructure landscape? My definition is a virtualized/remote computing resource from which users can purchase what they need, when they need it. To build a datacenter costs at least $15 million today, so the question arises whether it's always necessary to build one. Cloud computing could be an alternative. Our faculty and students want more computing resources and services delivered on our campus, because things on campus are tailored to their needs.

Through the Google-IBM-NSF partnership, 14 projects are being funded around the United States at $300,000 to $400,000 each to explore roles of cloud computing in HPC.

The HP-Yahoo!-Intel partnership aims to develop a global open source test bed, with Open Cirrus. This also involves the University of Illinois, KIT (Germany), IDA Singapore, and others.

Our conclusion at Penn State is that a public cloud like Amazon is cheaper than what you can do yourself for only a narrow range of computing tasks, but it's widely believed that Amazon is cheaper for many things. I would argue for an expanded role for university-based centers, including supporting the business use of advanced modeling and for educating future workers in computational science.

### *Thomas Eickermann, Jülich Research Center: PRACE Program 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 European Union (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 three to five systems of the highest order. The first PFLOPS system in Europe was installed at the Jülich 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). Western Europe 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 Jülich 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 POWER6, Cray XT5, NEC SX-9, 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.

**Comment:** What drove the process, the applications or having large resources available that will drive science?

**Speaker:** The applications came first.

## Tuesday, October 6

Steve Finn welcomed people to day two of the meeting and summarized the September 2009 HPC User Forum meeting in Broomfield, Colorado. Those presentations will be available soon on the Web site: **www.hpcuserforum.com**.

### *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.

The ABCC [NCI's Advanced Biomedical Computing Center] 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 driving goal is not gigaflops or processors, but how many lives can we save. The paradigm shift in biology is 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.

The NCI vision for translational research focuses on data-driven computation, and for this, 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:

- Next-generation sequencing technologies using high-throughput sequencers. 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 2500MB/hour by 2015. This creates multiple 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 this project is getting $275 million of stimulus money to start sequencing all the cancer genomes. This is a great idea scientifically. It entails 600GB per patient per disease, and with 500 patients/disease, 300TB data-points/disease, and 20 cancer types, that amounts to 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, and so forth, and no one's solved this artificial intelligence problem yet.

- Google in my mind is an HPC model. The company deals with huge volumes of data and allows you to access it. But 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.

- Analyzing high-dimensional data. The ABCC has been about this task 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, multicore, Cell, FPGA, GPGPU, and maybe others things). We also need to be able to verify we have correct code.

- NCI started funding purely *in silico* centers. It is 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 [single-nucleotide polymorphisms] 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 is also digital pathology. You take tumor slice and generate the image you 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.

- Structure: Non-intuitive results explain toxicity.

My view of the HPC "compute cloud": I think virtualization will have a bigger impact in the near term than the cloud. I don't care where people run problems. I just want it to run and be returned to them. I think my chances are better with virtualization.

- I need improvements in compute, storage (need a lot), networking, and turning information into knowledge.

- I need to distribute data securely and to access national resources.

### *Marie-Christine Sawley, ETH Zurich, CERN Group: Data Taking and Analysis at Unprecedented Scale: The Example of CMS*

In the past 70 years, discoveries have ranged from the electron to the quark. In each time, we've reduced the scale 300x. Particle physics wants to find a model that can reconcile the infinitely small with the infinitely large. Humans can intuitively understand in the range of objects from 10^4 to 10^-4, without tools.

CERN was inaugurated in 1954 and pursues nuclear research — its resources are used by 20 EU member states and associated states, including the United States. We have 8,000 researchers.

The Large Hadron Collider (LHC) is a tube in a tunnel with a 27-km circumference. In an extreme vacuum, it produces hydrogen atoms that collide with a detector present. CMS is one of these detectors. Protein collisions are rare because the densities of the materials are so slow. 1 boson appears per year on average. The detectors are Atlas and CMS. Atlas is very big, 7,000 tons. CMS is 12,500 tons. Both look at the same physical phenomenon, but using different methods with different teams and funding. CMS is 14x denser than Atlas. It has a huge coil like a solenoid with a powerful magnetic field. It detects different particles depending on their energy and mass. In September 2008, the first beam injection sent two beams in different directions so they would collide.

In summary, about 10PB of data per year is sent to CERN by the high-level trigger. We have Dell cluster with 33 racks of Harpertown and 200TB of disk.

I work in core computing. Our group takes the data and distributes it to all the scientists working around the world. We also do all the visualization. We have our own simulation software called Monte Carlo Production (MCPROD), which simulates the phases of the event.

The CMS computing project provides the resources and services to store/serve 10PB of data/year. We provide access to most interesting physics events. CMS collaborators are located in 200 institutions around the world. There are 120 people within CMS working specifically on the computing side of things (there are many more people than this at CMS).

Flow in the computing grid goes to a large storage system, then one copy is put on archival tapes and the other is used for calibration and express stream analysis. Then, each of our seven tier 1–associated centers around the world gets sent the data (e.g., Fermilab in Chicago). We expect to do five full reconstructions that will take 70,000 computing cores, 26PB of disk, and 38PB of tape in 2009–2010. The tier 1s need very good links to us. It's a dedicated, private optical network.

The data transfer challenge. In a 2009 test, we sent the data to all tier 1s with no problem. In August, we ran 17,000 concurrent jobs. Tier 1s plus tier 2s equals 40 sites. We run site readiness tests in the background at all times.

Conclusions for CMS from the past five years:

- We built a large expertise in day-to-day grid computing operations.
- Data challenges and cosmic data are letting us run in near-real conditions.

Do we have things to learn from other communities?

- From HPC, the need for balance between collecting, filtering, simulating, distributing, and interpreting large data volumes that comes in bursts.
- From data driven science, online filtering of a deluge of experimental and observational data.
- We're at the crossing point between experiments and simulations. The value is in the data and the knowledge it supports.
- We will increasingly see data from multiple disciplines.

### *Paul Muzio: HPC at the City University of New York*

CUNY was founded in 1847 for the children of immigrants and the working class, with no tuition. In 1870, Hunter College was established for women entering teaching. There are 23 CUNY campuses today. CUNY itself was formed in 1961. In 1975, a financial crisis in New York led to the end of free tuition and open enrollment.

This is the decade of science for CUNY. The goal is to upgrade CUNY as a research university, including hiring 1,200 new faculty and upgrading admission standards and research facilities. We have a cyberinfrastructure/high-performance computing initiative, with a charge to prove the value of the investment. CUNY enrolls 460,000 students and has had 12 Nobel Prize winners and many Pulitzer Prize winners. $6.5 million was allocated for a new, 90,000 sq ft building in the next five years or so. The expected total cost is $75 million.

CUNY's HPC philosophy: The university wants to see more research done, more research dollars coming in, and more publications. We can't compete with large NSF dollars. Cheap flops aren't necessarily the answer. MPI has learning curve problems and limitations. Will there be better alternatives in the future? We focus on certain key applications and want to accomplish things in those areas, plus bring in new architectures with external funding.

NSF recently funded us to acquire an MPI GPU cluster. We expect to get a PGAS system in 2010. We're working on other unique systems for the 2011 time frame.

The MPI GPU system adds 384 Nehalem cores and 24 NVIDIA Fermi GPUs to our existing 384-core Nehalem SGI ICE system. Some of our users have codes that fit GPUs. We expect to see x86 plus vectors on a chip in the future. Applications for this system include environmental science (WRF, WRF-chem, urban traffic, and air pollution), Monte Carlo (finance, condensed matter, photonics), creative media, and genomics. Wall Street is interested in GPUs.

### *Jean-Marc Denis: Bull Technology Update — Extreme Computing*

Bull can cover HPC needs from small departmental computers to large, world-class computers and also storage. We have a GPU-based product and use FPGAs in special products with the oil and gas industry and other industries. We work with InfiniBand technology. Bull Cluster Suite is a complete software layer based on Red Hat. We have plans to work with SUSE as well.

We provide expertise and project management. In Europe, we have more than 500 people dedicated to HPC. Bull believes that as a midsize company, we should not try to do everything. Our model for developing technology, especially software, is to work with partners. With CEA, we developed the concept of computer room optimization with a goal of PUE close of 1.05. With another partner, we developed our B505 Accelerator Blade. We have many software collaborations, including GPU compilers and Linux/Microsoft hybrid clusters.

The bullx blade system uses general-purpose accelerators for the compute part. We will start shipping the GPU blade in the first half of 2010. The bullx B505 accelerator blade provides 18.9TFLOPS in 7U.

Ultracapacitor module (UCM) provides embedded protection against brownouts of up to 250 microseconds. With this, you can avoid onsite UPS and improve your overall availability to run longer jobs.

The Bull Cool Cabinet Door enables very dense HPC solutions. With 40W in 1.4 sq m, it provides 77% savings over air conditioning. We have large systems at Jülich (300TFLOPS) and the Regionales Rechenzenturm Köln.

### *Lutz Schubert, HLRS: Workflow and HPC?*

Workflows are about centralized communications and are not about great performance. Classically, a workflow is a predefined sequence of tasks exposed in a standard fashion. So, by this definition, any program is a workflow. At the same time, it's the least efficient way to execute a program: centralized communication and communications overhead through standards are just the opposite of what HPC is trying to achieve.

But why should workflows act directly on the process level? What benefit can they have? They can also act as a way to control higher-order processes, such as to initiate an HPC job as part of a larger process or trigger jobs on specific events or coupling jobs according to their status. Examples of higher-order processes with HPC tasks include the "Grid" concept, a collaborative engineering project with British Aerospace, and also a virtual engineering collaboration with HLRS and industry that we're involved in.

Event-based job control entails the "cloud" principle, where you are making resources available as an on-demand utility. This is not a workflow as such, but it's easier to define and adapt by making use of workflow descriptions. An example is financial calculation in the stock market. You do the computation on demand, always involving specified steps each time.

Coupled applications are multiple HPC jobs that depend on each other. The next job (including its nature) is triggered by the completion of the prior job. Intermediary results may trigger evaluation jobs in parallel (example: a material stress test in vehicle design). You model the whole behavior, but it's too complex to model all at once, so you model the elements of the job separately. This is a slow process, so not an attractive workflow model. But you can model the critical points to make it more efficient. Another example is modeling a virtual physiological human. This involves the same base model, but different detailed elements. There is no direct coupling of the elements. New knowledge about one element can impact other behavior. The simulation of specific diseases may lead to coupling, depending on the goal (e.g., how medicine spreads if the heart muscle is affected, or how a muscular disease spreads to affect the heart). Data is only exchanged between elements under certain conditions. The workflow must model these events and conditions.

In summary:

- Workflows can be used for modeling to control behavior, events, and conditions.
- You should understand these tasks not as jobs but as triggers for other jobs.
- Data sets should not be provided via a workflow.
- Recurrent tasks that aren't time critical can be supported by workflows.

Workflows are good for:

- Describing relationships between actors and tasks
- Recurring configurations and processes
- Result-dependent relationships between jobs
- Modeling event–based triggers

Workflows are bad for:

- Tightly coupled process control and execution
- Distribution of large data between jobs
- Fast interactions with external processes

### *Beat Sommerhalder: Parallel Computing at Microsoft*

Microsoft's vision for HPC is based on reduced complexity, making HPC mainstream and creating a broad ecosystem for HPC. Microsoft stepped into the HPC market in 2006. The driver was the multicore 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.

Looking forward, hard problems include the following:

- Scaling distributed systems is hard.
- Data sets are increasing in size.
- Programming models are complex. 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 breadth developers; evolve hybridized and scale-out models).
- Seek opportunities for "automatic" parallelism (e.g., F#, DryadLINQ).

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 is restructuring 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.

### Meeting Wrap Up

Earl Joseph, Steve Finn, and Michael Resch thanked everyone for attending and thanked the presenters. The presentation slides are posted at: **www.hcuserforum.com**.

## LEARN MORE

### Related Research

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

- *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

This IDC research document was published as part of an IDC continuous intelligence service, providing written research, analyst interactions, telebriefings, and conferences. Visit www.idc.com to learn more about IDC subscription and consulting services. To view a list of IDC offices worldwide, visit www.idc.com/offices. Please contact the IDC Hotline at 800.343.4952, ext. 7988 (or +1.508.988.7988) or sales@idc.com for information on applying the price of this document toward the purchase of an IDC service or for information on additional copies or Web rights.