Final Minutes
Advanced Scientific Computing Advisory Committee
August 14–15, 2012
American Geophysical Union, Washington, D.C.

ASCAC members present:

Marsha Berger
Marjory S. Blumenthal
Vinton G. Cerf
Vincent Chan
Jackie Chen (via telephone; Tuesday only)
Jack J. Dongarra
Roscoe C. Giles (Chair)
Sharon C. Glotzer
Susan L. Graham
Anthony Hey
Gwendolyn L. Huntoon (via telephone; Tuesday only)
Juan Meza
John Negele (via telephone)
Linda R. Petzold
Vivek Sarkar (via telephone)
Victoria White
Dean N. Williams

ASCAC members absent:

Barbara M.P. Chapman

Also participating:

William Brinkman, Director, Office of Science, USDOE
Christine Chalk, ASCAC Designated Federal Officer, Office of Advanced Scientific Computing Research, Office of Science, USDOE
Daniel Hitchcock, Acting Associate Director, Office of Advanced Scientific Computing Research, Office of Science, USDOE
Paul Messina, Director of Science, Argonne Leadership Computing Facility, Argonne National Laboratory
Lucy Nowell, program Manager, Office of Advanced Scientific Computing Research, Office of Science, USDOE
Frederick O'Hara, ASCAC Recording Secretary
Jeannie Robinson, Oak Ridge Institute for Science and Education
Christiane Jablonowski, Department of Atmospheric, Oceanic, and Space Sciences University of Michigan
Michael Heroux, Computation, Computers, Information, and Mathematics Center, Sandia National Laboratories
Randall Laviolette, Office of Advanced Scientific Computing Research, Office of Science, USDOE
Steven Lee, Office of Advanced Scientific Computing Research, Office of Science, USDOE
Bronson Messer, National Center for Computational Sciences, Oak Ridge National Laboratory
William Harrod, Office of Advanced Scientific Computing Research, Office of Science, USDOE
Grigory Bronevetsky, Computer Scientist, Lawrence Livermore National Laboratory
Thomas Lehman, Information Sciences Institute, University of Southern California
Teresa Quinn, Head, Integrated Computing and Communications, Lawrence Livermore National Laboratory

About 45 others were in attendance in the course of the two-day meeting.

Tuesday, August 14, 2012
Morning Session

Before the meeting, the new members of the Committee were sworn in by Jevecca Romero of the DOE Human Resources Office.

The meeting was called to order by the chairman, **Roscoe Giles**, at 8:59 am. Jeannie Robinson made safety and convenience announcements. Giles welcomed Vinton Cerf and Juan Meza as new members. He introduced **William Brickman** to present an update on the activities off the Office of Science (SC).

There will probably be a continuing resolution based on the FY12 budget and lasting through March 2013. The extension of tax cuts, the FY13 budget and its associated sequestration of funds, and the debt ceiling that will be reached in February will be the focus of Congress and will lend uncertainty to DOE's budget planning.

The President's budget request for Advanced Scientific Computing Research (ASCR) was $455.6 million. The House mark came back as $442 million, which is 0.3% more than the FY12 appropriation. It was also less than the $455.6 million that came back in the Senate mark. The Office of Basic Energy Sciences (BES) lost a lot; the President's request was $1799.6 million, and the Senate mark was $1712.1 million. The Office of Biological and Environmental Research (BER) experienced a cut in climate funds; the President's request was $625.3 million, and although the Senate mark gave it $625.3 million, the House mark gave it $542 million. The Office of High Energy Physics (HEP) seems to have the sympathy of Congress; the President's request was for $776.5 million, and the House mark was for that same amount, and the Senate mark was for $781.5 million. The HEP budget was stretched by construction funding for the Thomas Jefferson National Accelerator Laboratory (JLab), the Relativistic Heavy Ion Collider (RHIC), and the Radioactive Ion Beam (RIB) facility. In addition, the SC budget includes increases for laboratory office buildings. However, there was trouble getting funding for the SC Fellowship Program; this is the last year of the funding for the 3-year program. A continuing resolution to operate at the FY12 budget level is expected; sequestration of funds would be very problematic.

The world is getting hotter. There is a little good news, however. For the first time, the Energy Information Administration (EIA) has said that the amount of carbon dioxide being produced in the United States is going down a little bit. The Arctic is really heating up. Old ice is melting before new ice is made each winter; this is a new cycle. There is a belt in North America that is getting hotter, but warming is not uniform. Sea level is rising. Munich Re has graphed the number of natural catastrophes from 1980 to 2011; it shows a significant rise. Its figures indicate that, in 2011, insured losses topped $100 billion and uninsured losses were about $275 billion. The $^{13}C/^{14}C$ ratio shows that the carbon dioxide in the atmosphere is almost all new carbon dioxide.

New technology on the market to shift energy use from petroleum to electricity includes the Tesla automobile, which can travel 300 miles on a single charge. There are four models of the Tesla. The smallest (least expensive) has 40 kWh of power, has a range of 160 miles, can go

from 0 to 60 in 6.5 seconds, and has a top speed of 110 mph. It recharges at 62 mph with a supercharger. Hybrid-car sales have taken off in terms of percent of all vehicles sold: from 0% in 1999 to 3% in 2012. They offer a factor-of-two increase in mileage (40 MPG versus 20).

Solar and wind are growing slowly. Nuclear could provide a large amount of the United States' energy needs. The natural gas industry has been revolutionized in the past 5 years because fracking has unlocked a plethora of domestic natural gas. A natural gas power plant produces twice the electricity than a similar coal-fired plant with the same carbon dioxide emissions. What is being talked about is burning natural gas with pure oxygen so the only emissions are carbon dioxide and water, easing the need for sequestration.

Normal drilling extracts only 15% of oil reservoirs. Pressurizing old oil wells with carbon dioxide makes petroleum less viscous and forces it up the wellhead. Extracting carbon dioxide from coal-fired flue gas costs $80/ton; the economic value of the carbon dioxide is $10 a ton.

Computer modeling plays a huge role in seeing what we are doing and can do. ASCR's job is very important in the nation's response to climate change. One of the challenges is how to push to the exascale (in an affordable manner). The original $4.5 billion price tag is not in the cards. We need to figure out how to nationalize and popularize this program.

Giles noted that part of the exascale report said that a leadership position was possible for the United States, but that opportunity seems to be slipping away. Brinkman replied that the Office of Science and Technology Policy (OSTP) and the President's Council of Advisers on Science and Technology (PCAST) are not convinced of the need for the extra scale.

Giles pointed out that the travel restrictions put on the scientific community hits the supercomputing community hard. Brinkman said that the Office has run into this problem but has resolved itself to go for the exemptions for several conferences (e.g., Supercomputing 2012).

Berger asked who had zeroed out the fellowship program. Brinkman said that Congress canceled the SC Graduate Fellowship Program. The ASCR fellowship program is still on track.

Hey ask what the earliest date might be that there will be clarity on an exascale plan. Brinkman replied, this fall. Hey noted that China and Japan have shifted capacity to support the exascale. Brickman said that the Office was aware of that.

Daniel Hitchcock was introduced to describe the activities of the Office of Advanced Scientific Computing Research (ASCR). He alluded to a potential barrier between Germantown and Washington, DC. ASCR contributes to DOE's goal to maintain a vibrant U.S. effort in science and engineering as a cornerstone of the nation's economic prosperity with clear leadership in strategic areas. The targeted outcome of the Lead Computational Sciences and High-Performance Computing is to develop and deploy high-performance computing hardware and software systems through exascale platforms. There are data-intensive and computation-intensive challenges. ASCR provides the basic research for the Department. Action needs to be started now so that computing resources are available when needed. President Obama announced the Big Data Initiative on March 29, 2012.

All of the exascale hardware trends affect data-intensive science (in many cases, more than compute-intensive applications). The Square Kilometer Array in Australia needs 100 MW for its computing infrastructure. That produces a significant power bill. Universities and industries will face the same problem. Computing demand doubles every year. Supercomputers now consume 20 MW. Given growth trends, that would lead to a huge, impossible-to-meet power usage by the end of the decade. All science experiments are pushing toward higher data-production rates, so

high that one cannot save the data. One needs to move the computer closer to the experiment. Such data transfer is, in essence, a denial-of-service attack.

The Office got a mark from the House. It said facilities are important; keep upgrades on track; do not spend money on data-intensive computing. The Senate asked for a plan on data-intensive computing. Such a plan has to go through the Office of Management and Budget (OMB), which takes a long time. Funding for the Office is $13 million higher in the Senate mark than in the House mark, but no one knows how it will turn out. The Senate has budgeted funding for high-end computing but has granted no authority to use that funding.

Yukiko Sekine is retiring. Dave Goodwin will be the new National Energy Research Scientific Computing Center (NERSC) program manager. The Office hopes to add an applied mathematician position. A computer-science program manager position is posted.

Cerf stated that the exascale will be expensive and asked if there were an opportunity to involve other agencies in this endeavor. Hitchcock replied that there is a lot of collaboration with the National Nuclear Security Administration (NNSA) and the international community. However, most concurrency is now on the nodes, and the need for memory is overwhelming. Low-power memory systems that operate at high bandwidth are needed. The memory and CPU need to cooperate, requiring a new technology that industry will not build for just one customer. There is a need to collaborate with industry to bring the cost down. Cerf said that more memory on a chip is needed, even with commodity computers. There are opportunities for reinventing computing for both the exascale and commodity computing. Hitchcock replied that that was an interesting challenge.

The Early Career Program operates across SC. ASCR has six Early Career awardees at two national laboratories and three universities. This is the fourth year of the program. This funding opportunity was posted on July 20, 2012; pre-applications are due September 6; applications will be encouraged by October 4; and full applications are due November 26 on two topics: development of (1) mathematical descriptions, models, methods, and algorithms to accurately describe and understand the behavior of complex systems involving processes that span vastly different time and/or length scales and (2) the underlying understanding and software to make effective use of computers at extreme scales to transform extreme-scale data from experiments and simulations into scientific insight.

ASCR builds and operates world-class facilities: NERSC with Hopper and NERSC-7, the Argonne Leadership Computing Facility (ALCF) with Mira, the Oak Ridge Leadership Computing Facility (OLCF) with Titan, and the Energy Sciences Network (ESnet) that operates at 100 Gb/sec nationwide.

The Innovative and Novel Computational Impact on Theory and Experiment (INCITE) Program awards time on the ALCF and OLCF systems for researchers to pursue transformational advances in science and technology. The 2013 Call for Proposals closed June 27, 2012. Total requests amounted to about 14 billion core-hours, nearly 3 times as much as requested last year; 143 proposals were submitted, an increase of nearly 20% over last year. Awards of about 5 billion core-hours will be announced in November for CY 2013. It involves a lot of university people (40%).

Argonne National Laboratory (ANL) has the Blue Gene installed. The floor had to be reinforced to hold the machine. They will need a power upgrade.

At Oak Ridge National Laboratory (ORNL) in September, they will add new 20-PFLOPS graphic processing units (GPUs) to the Jaguar to make the Titan. Cold water may be pumped out of the bottom of Melton Hill Dam to cool the system (returning the warm water to the river).

The $53 million NERSC-7, a 2-PFLOPS Cray Cascade had its Critical Decision 3 (CD-3) approved in June 2012; the contract with Cray was signed in June; and delivery is expected in 2014.

ESnet is doing 2 years of planning, design, and development will culminate in deployment of production 100G network to support DOE science missions, November 2012. A 20-step build process is now under way. A testbed is now operating. Cerf noted that Google just implemented OpenFlow with great success. Hitchcock said that DOE is implementing similar technology to stitch together an optical circuit all the way from CERN {Conseil Européen pour la Recherche Nucléaire [now European Organization for Nuclear Research (Organisation Européenne pour la Recherche Nucléaire)]} to NERSC. It has a different need: to move very large data loads to a few places. ESnet has inaugurated a Policy Board; has a new home at the Scientific Networking Division of Lawrence Berkeley National Laboratory; connected DOE supercomputing sites to more than 20 active international research projects, supporting major scientific discoveries at the large Hadron Collider (LHC), Daya Bay Neutrino Experiment, and Palomar Transient Factory.

ASCR is conducting research on applied mathematics, computer science, partnerships, and next-generation networks for science. It is trying to integrate discretization with solvers; it has solicited proposals on this topic. Ten Projects are being funded on next-generation networking for science. There are Scientific Discovery Through Advanced Computing (SciDAC) partnerships. All of the SC offices were polled to learn their future computational needs and what opportunities should be pursued. Workshops were held to define the needs. Outreach to industry was conducted. An Small Business Innovative Research (SBIR) topic of high-performance computing (HPC) has been established.

A workshop was held with BES to understand problems emerging in the light sources. The workshop report is now out. It recommends:

- Integrating theory and analysis components seamlessly within the experimental workflow,
- Moving analysis closer to the experiment, and
- Matching data management access and capabilities with advancements in detectors and sources.

A Geant4 [GEometry ANd Tracking, v.4] workshop was held to identify applied-mathematics and computer-science challenges in the effective transformation of Geant4 and to examine opportunities for discovery to meet these challenges. Cerf noted that there is an issue (bitrot) in not collecting enough data to understand what the data means and what should be saved. Hitchcock admitted that ASCR does not have a lot of investment in that area. Other programs have been looking at what metadata should be saved with the data. ASCR is trying to figure out what could be done to have the maximum impact. Cerf said that many institutions have this problem; collaboration and co-investment are needed.

In the FY12 budget, there is $73.4 million for the exascale:

- $5 million in Applied Mathematics for Uncertainty Quantification
- $20 million in Computer Science for Software Environments
- $ 5 million in Computational Partnerships for Software Environments

- $30 million in Research and Evaluation Prototypes for Industry Partnerships
- $13.4 million in Computational Partnerships for Co-design

A funding opportunity announcement (FOA) was focused on runtime systems. Awards have been selected and are in negotiation.

ASCR and NNSA's Advanced Simulation and Computing exascale principle investigators (PIs) held a meeting on April 19-20, 2012, in Portland, Oregon, to contribute to the exascale software research plan. There also has been a JASON study on June 27-29, 2012, on the technical challenges associated with developing scientific and national-security applications for exascale computing. It looked at the applications and what they need to do in the future.

Graham asked if October would be too soon to get an update on the co-design project. Hitchcock replied that uncertainty quantification has already been used in nuclear reactor design. There should be a lot to talk about.

Berger asked if the study on using air cooling was done at NERSC. Hitchcock responded, yes; they have monitoring air temperature throughout the facility to qualify the concept. However, most of the easy things have already been done to minimize power usage.

Cerf pointed out that, in the discussions at the Association for Computing Machinery's Turing celebration in June, it was obvious that synchronous computing is on the way out and that asynchronous computing is very attractive; in the same vein, the exascale propositions should be challenged, also. Hitchcock said that asynchronous input/output (I/O) has been used by the Office's facilities. A lot of things, including the need for coherence, are being questioned. Cerf asserted that the cooling problem is most effectively addressed by finding ways to use warmer water for cooling. Hitchcock pointed out that the new NERSC computer will use water at 75°F. The Titan will use water from the Clinch River. Cerf noted that Google is looking at the possibility of locating facilities in the Arctic.

White asked if there were any research initiatives for building tools for the exascale that will benefit a broad range of uses. Hitchcock replied that ASCR does high-performance computing (HPC) research for the users at DOE facilities, and it works on some sustainability plans. It is hoped that small businesses develop (and maintain) and support ASCR's software for the community market. Cerf noted that nobody wants to pay for infrastructure. Google decided to produce basic software that others would build on top of. The more generalized the software is, the more sustainability one gets.

Giles noted that one needs to have room to fail; one needs room for research; one has to take risks and look to the long term. Bringing people from the national laboratories into the workshops is getting harder. Hitchcock pointed out that de Lorenzo always said that this year is worse than last year but better than next year. One has to accept limitations and move forward. Brinkman added that the budget does not change much from year to year, and things seem to work out. We do not have long-range commitments. Congress will not allow that. It is an enormous problem in this country. SC gets budget increases during recessions. SC did not send the exascale request to OMB because it knew it would not get the funding. Dongarra asked about the data-intensity issue. Brinkman responded that this is an issue that will be straightened out at the international level. However, DOE has been getting negative comments from OSTP and PCAST.

Chan asked if moving analysis closer to the experiment is already happening in high-energy physics. Hitchcock responded that most analysis is done in the facility (doing data triage). How

this is done will be critical. One cannot save all the data. One will have to figure out what is not worth saving. One needs to look at the metadata to see what needs to be recomputed.

Hey stated that ASCR needs to act on the exascale, the base computer science program, and data-intensive computing. Brinkman responded that the Office takes the committee of visitors (COV) recommendations very seriously, although it may not act on all of them. Hitchcock added that it is a continual balancing act to fund the best proposals. To face the coming challenges at a time of revolutionary change means that one has to do things that make a lot of people uncomfortable.

Meza said that it seems that $5 million for applied mathematics is on the short side. Hitchcock answered that there is the exascale crosscut that identifies something Congress does not agree with and takes all the money away. In the risk calculus, it was agreed that it was not wise to ask for more applied mathematics funding.

John Negele said that, on the leadership-class machines, there is a learning curve. IBM limited access to the code. He asked what DOE can do to allow early users to use the machine at turn-on. Hitchcock said that IBM's decisions are IBM' s decisions, and they have more lawyers than DOE does. Negele pointed out that it took 1.5 years to negotiate a non-disclosure agreement with IBM's lawyers with no access during that time. DOE should pressure IBM to cut the legal nonsense. Hitchcock stated that this is determined by what one can negotiate with IBM. He noted that Paul Messina will talk about early access on the following day of the meeting.

Williams asked if the 100-Gb/sec network would go into production this year. Hitchcock replied, yes, to do experiments on using the bandwidth. In experiments, 97% usage of the bandwidth was achieved.

The floor was opened to public comment. There being none, a break was declared at 11:08 a.m. The meeting was called back into session at 11:24 a.m.

Christiane Jablonowski was asked to describe her Early-Career Award research on climate and weather modeling. Her group is looking at

- High-order, finite-volume, nonhydrostatic, dynamical-core modeling on cubed-sphere grids
- Adaptive mesh refinement (AMR) and variable-resolution grids
- Evaluations of dynamical cores.

To develop a new model, one performs 2-D shallow-water test cases, 3-D dry dynamical core test cases, 3-D dynamical core and moist simple physics, the 3-D Aqua-Planet Experiment, and the 3-D Atmospheric Model Intercomparison Project assessment with the full model.

One problem is how to accurately remap data from a cubed-sphere grid to a latitude-longitude grid and vice versa. The group used high-order (third or fourth order nonlinear) mapping. The problem occurs at edges of the cube. Information is extended at the edges with high-order calculated "ghost cubes." On a Cartesian mesh, the group is developing a non-hydrostatic wobble at a one-kilometer scale. It deals with issues like high-speed sound waves (that produce friction). Everything is then put on cubed-sphere grids that

- Are fourth-order in the horizontal with explicit time stepping,
- Are second-order in the vertical, implicit,
- Can be configured for shallow and deep atmospheres, and
- Prepare for adaptive mesh refinement applications on cubed-sphere grids.

Finite-volume methods allow physical consistency with built-in conservation and monotonicity. The use of high-order data hides grid imprinting or spectral ringing and increases accuracy. Equations are conservatively written in flux terms and source terms. The equation of state relates temperature, volume, and pressure. Integrating the system of equations leads to compact notation of the state vector and fluxes, which can be split into horizontal and vertical components. Then cell-centered components of the state are computed, a fourth-order edge value is reconstructed, fluxes are computed with Riemann solvers, and the cell-averaged flux is recovered. Several tests have been run, and the results have been published.

Variable-resolution modeling cannot be done with current computing resources. Several types of meshes are used as a result.

The group is using an adaptive mesh for a cubed sphere that tracks the effect under study across the sphere. Conforming meshes (with soft edges) are being developed to produce high-resolution representations of, say, transitioning cyclones.

Refined meshes look very much like uniform meshes (i.e., the refined mesh moves with the phenomenon under study). The adaptive measures mimic the uniform meshes very well but at much higher resolutions; and they are dynamical rather than being static.

A workshop was held at the National Center for Atmospheric Research (NCAR) in August 2012 to compare dynamical-core models. The goals of the workshop were to explore new test cases designed for hydrostatic and non-hydrostatic dynamical cores on the sphere, for both shallow and deep atmosphere models with a special focus on non-hydrostatic models and high resolutions and to provide standard diagnostics for model evaluations.

In summary, the Michigan research group is pushing the frontiers of

- Dynamical core modeling for weather and climate applications by developing physically consistent fluid dynamics solvers based on high-order finite-volume methods,
- Variable-resolution modeling by exploring dynamic and static mesh adaptations, and
- Objective dynamical core evaluations via new test cases.

Giles asked what the range of results was in the model-workshop "bake-off." Jablonowski replied that the models do not agree. There are a lot of differences. A lot of choices have to be made in any dynamical model.

Cerf asked about ground truth. Jablonowski responded that it is difficult to compare results to ground truth because the models do not have observational input. Cerf noted that the more diffusers there are, the more drifting from truth occurs. The initial conditions have a great effect.

Berger asked if the mesh were refined in time. Jablonowski answered, yes, but the implicit solver for two directions has not been written, yet. Finer timestamps are performed at smaller mesh sizes.

The floor was opened to the public. There being no public comment, a break for lunch was declared at 12:14 p.m.

Tuesday, August 14, 2012
Afternoon Session

The meeting was called back into session at 1:30 p.m.

Michael Heroux was asked to review parallel applications for the "year of the exascale." The old commodity computing trends are failing, but new trends are being established (e.g., in

energy efficiency through threadcount occupancy and state per thread; vectorization; heterogenicity through performance variability and core specialization; and memory per node). Parallelism is essential. One needs to take advantage of these trends.

What major role will the message-passing interface (MPI) play? It will likely be MPI + X (MPI enhanced) with resilience and programmability.

Single program, multiple data (SPMD) and MPI have been successful because of portability, separation of parallel and algorithmic concerns, and preserving code investments. MPI was disruptive but not revolutionary. New parallel applications should preserve these dynamics.

We live in an expanded ecosystem that includes terascale laptops, petascale desktops, and exascale mainframes. A broad community effort is being supplemented by HPC value added.

Almost all DOE scalable applications use MPI. It provides a portability layer. Application developers access MPI via a conceptual layer, although they could swap in another SPMD approach. Even dynamic SPMD is possible. However, adoption is expensive. The entire computing community is focused on X. MPI and X interactions are well understood because they are a straightforward extension of the existing MPI+Serial. New MPI features will address specific threading needs.

Effective node-level parallelism is the first priority. Future performance is mainly from node improvements; the number of nodes is not increasing dramatically. Application refactoring efforts on the node are disruptive: almost every line of code will be displaced. A successful strategy similar to the SPMD migration of the nineties is needed. If there is no node parallelism, failure will ensue at all computing levels.

The single-program, multiple-data approach fits most programming needs. However, if one has fourth-order Laplace, one needs to redo the programming approach.

The first step of parallel-application design is to identify parallel patterns. Every parallel programming environment [Open MP, Intel Threading Building Blocks (TBB), and NVIDIA's Compute Unified Device Architecture (CUDA)] supports basic patterns like parallel-for and parallel-reduce. They all do the same type of things and are not new.

Why patterns? Because they are essential expressions of concurrency, they describe constraints, they map to many execution models, and there are lots of ways to classify them.

On our palette, we will want a parallel framework for no loop-carried dependence, rich loops, and the use of shared memory for temporal reuse and efficient device data transfers. We want a parallel-reduce framework to couple with other computations and out of a concern for reproducibility.

Pipelines establish filters that are sequential and parallel. Filters executed in sequence. The programmer’s concern is to determine whether a filter can execute in parallel, to write the filter, and to register it with the pipeline.

Another approach is a thread team with multiple threads; a fast barrier; and a shared, fast-access memory pool. A qualitatively better algorithm results because of threaded triangular solve scales, fewer MPI ranks (with fewer iterations and better robustness), and data-driven parallelism.

It is desirable to allow a domain scientist to remain a domain scientist. The languages needed are already here.

MPI-X preserves programmability. MPI applications preserve sequential programmability via abstractions. Most X applications do this also via patterns. There are critical issues in

migrating to X: identifying latent node-level parallelism; identifying and replacing current, essential node-level sequentiality; isolating computation to stateless kernels; and abstracting physics i,j,k from data structure i,j,k. Any beyond-MPI platform must also preserve programmability.

Scientific applications tend to be written in Fortran. C++ gives more opportunities and advantages. To develop resilient applications, one must embrace performance variability, localize failure, and handle soft errors. The keys to addressing resilience are algorithms and codesign.

It would be good to remove correctness if it meant improving performance. If one slows down one process, all parallel processes are slowed down, also. Algorithms can address problems (like latency-tolerant algorithms) by using pipelines to hide latency to produce multiple levels of look-ahead. Another resilience scenario is when an answer does not show up because of a local failure, and the job gets killed. A persistence process needs to be developed to squirrel away some data to be used during recovery when additional hardware is brought in. It would enable a fundamental algorithm to aid fault recovery. A third scenario is selective (un)reliability. A single bit clip out of billions of operations can give wrong results. Selective reliability enables reasoning at the approximation-theory level for implicit applications. It enables "running through" faults.

In summary, node-level parallelism is the new commodity curve. Domain experts need to "think" in parallel. Most future programmers won't need to write parallel code. Fortran can be used for future parallel applications, but complex parallel patterns are very challenging, parallel features lag, and the lack of compile-time polymorphism hurts. Resilience is a major front in extreme-scale computing. MPI+X is and will be a dominant platform for the tera and peta scales. MPI+X will be a dominant platform for the exascale. Ongoing efforts are needed in MPI to address emerging needs. Migrating to emerging industry X platforms is critical and urgent.

Graham asked whether, if he were not worried about migrating codes, this would still be the best approach. Heroux answered, yes.

Cerf said that he would not want to foreclose any possibilities or new ideas (e.g., new programming models). Heroux responded that DOE does not want to close out any new ideas. Cerf asked if there were any tools for checking whether a calculation is correct. Heroux said that strongly analyzable code is the best tool.

Negele stated that MPI is likely to be viewed in the same way as Fortran and asked how that plays into the big computations needed for big data. Heroux said that libraries will be needed. An assessment will have to be made whether the data or analytics are bigger.

Giles asked what other elements besides big data and graphs do not fit into the MPI-X model. Do error control and resiliency and energy usage for data storage and operations enter in? Will those problems all be on the X level? Heroux replied that is where the greatest gains are to be had. Adaptable address space is one approach that is appearing in some programs, but it reduces the convergence rate of the solution of partial differential equations (PDEs), and the two compete. Maybe one does not do PDEs. Clever techniques will likely evolve.

Berger noted that resilience is being put on the algorithm, but an algorithm cannot do much on resilience. Heroux responded that one wants to demonstrate a payoff and then get the program and runtime people to come along in concert. A recent resilience workshop was hardware centric. Maybe that will be the answer.

Hey asked whether any HPC languages might be adapted at the exascale. Heroux answered that Chapel has the best chance. The challenge is to identify a market that will drive the development.

The floor was opened to the public; there was no public comment.

Randall Laviolette was asked to present an update on SciDAC.

SciDAC started in 2001 to pull together scientists, computer scientists, and programmers to exploit HPC. SciDAC-3 has four large software projects ("SciDAC institutes"). SciDAC-3 solicited proposals last year. The FOA offered $15 million for three years for up to five institutes. The institutes had to be architecturally aware.

Three awards were made, but "data" was missing, so a supplemental FOA was issued for a data institute. The four institutes are FASTMath (Frameworks, Algorithms and Scalable Technologies for Mathematics), QUEST (Quantification of Uncertainty in Extreme Scale Computations), SDAV (Scalable Data Management, Analysis, and Visualization), and SUPER (Institute for Sustained Performance, Energy, and Resilience). All of them are in California. They will address the issues associated with debugging, load balance, fault tolerance, multicore, vector floating point (VFP) units/accelerators, and power.

FASTMath helps application scientists overcome two fundamental challenges. It has three broad topical areas: tools for problem discretization, solution of algebraic systems, and high-level integrated capabilities. All FASTMath technologies will focus on performance engineering for multi-/many-core architectures.

At SDAV, tools are being enhanced in several ways. SDAV's goals are to actively work with application teams to assist them in achieving breakthrough science; to provide technical solutions in the data management, analysis, and visualization regimes that are broadly used by the computational science community running on Leadership Class machines; and to use existing robust tools to the extent possible and develop/adapt tools on as as-needed basis.

The objectives of QUEST are to deliver expertise, advice, and state-of-the-art uncertainty-quantification tools on advanced computational architectures and to shepherd forward the QUEST repertoire of uncertainty-quantification theory, algorithms, and software while enhancing their effectiveness for relevant benchmark problems.

The goal of SUPER is to ensure that DOE's computational scientists can successfully exploit the emerging generation of HPC systems.

During the past month, there was a workshop for Secretary Chu. It started with an educational colloquium and had panels, Q&A with Chu, and breakout sessions. A workshop report is being written.

The road ahead: SciDAC seeks to develop application partnerships with SC offices and industry, broadening its base. There will be a SciDAC-3 PI meeting on September 10–12, 2012. The overall portfolio and management of Institute awards is expected to cover a significant portion of DOE computational science needs on current and emerging computational systems. Basic research programs prepare the way for SciDAC-4 and extreme-scale institutes.

Steven Lee was asked to present an update on SciDAC Scientific Computation Application Partnerships. SciDAC wants to partner with all SC programs on projects of strategic importance. SciDAC puts out joint solicitations with those partner programs. Collaboration between scientists and computer scientists must be substantial. The first project selected was in Fusion Energy Sciences (FES; up to $33 million) on plasma-edge physics, multi-scale integrated modeling, and

materials science. The partnership in High Energy Physics (HEP $12 million) focuses on (1) the search for dark matter with *N*-body simulations of the cosmos; (2) lattice gauge theory; and (3) accelerator-science modeling and simulation. The partnership in Nuclear Physics (NP; up to $20 million) focuses on low-energy experiments probing the properties of nuclei at the Facility for Rare Isotope Beams (FRIB) and at A Toroidal LHC ApparatuS (ATLAS); medium-energy experiments at JLab to probe the properties of hadrons; and heavy-ion-collision experiments at RHIC and the LHC to probe quark-gluon-plasma properties and the quantum chromodynamic equation of state. The partnership in Earth Systems Science (maximum $32.5 million over 5 years) focuses on climate: to develop (1) physics and dynamics for atmosphere, ocean, or ice sheets to run efficiently and accurately with high-resolution or unstructured grids; (2) efficient and accurate schemes for simulating atmospheric or oceanic chemical or biogeochemical tracers; and (3) methods to validate and to characterize uncertainty in climate simulations. The partnership in the Materials and Chemical Sciences (up to $30 million over 5 years) focuses on development of first-principles treatments of excited states and excited-state processes and electron correlation in finite and extended systems. Other priority research directions include solar energy, chemical reactions, magnetism and superconductivity, materials in extreme environments, separations, and energy storage.

One hundred twenty six pre-proposals were received, of which 65 were encouraged to submit (and *did* submit). These proposals were reviewed by 134 reviewers, and 17 were deemed successful. Each project involves many (up to 14) institutions.

Lucy Nowell noted that the solicitation for the data institute came in after the other institutes were selected. Supplemental funding was provided so the projects could make use of the data institute.

Lee continued: Projects target strategic needs of SC program partners. Three-fifths of the ASCR-funded effort supports institute personnel. All projects are science-led. ASCR supports more than 80 named faculty or staff in 25 institutions in multi-disciplinary collaborations with domain scientists.

Chan asked how SciDAC-institute scientists get paid. Lee replied that they had two charge numbers: one for partnership work and one for institute work.

Blumenthal asked how well the approach to reviewing proposals with domain scientists and computer science reviewers worked. Lee answered that he was surprised how well it worked. All but one of the proposals was processed by snail mail; the physical panel reviewers interacted well. The mathematics and computer science people were more credible than the domain scientists. Things that were hated or loved were hated or loved by all.

Meza asked if the mismatch in the time frames could be reduced. Lee responded that there is a problem in that not everyone chose a 5-year project. That problem was not solved, but it is not as bad as before. Laviolette added that there was a staggered start: the institutes first, and the projects 9 months later.

Berger asked what was useful from the workshop. Lee replied that some barriers were legal and administrative, and they are now getting resolved.

Dongarra noted that the SciDAC institutes get additional funding. Lee agreed; they initially got funding and then they got more for each partnership that they participated in.

White asked what will happen to the software and algorithms produced by SciDAC and how will it be sustained as usable. Lee answered that some software was cycled over to SciDAC-3.

SciDAC has $15.5 million for the next 5 years. It leverages what is out there. It is hard to say what will happen. Dean Williams asked if they had asked the PIs what happens with the codes. Lee said that the PIs will be asked that at the next meeting.

David Branson (via telephone) asked what happened to the SciDAC Executive Directors Council. Lee responded that the program is still in the thinking stage about that possibility.

A break was declared at 3:10 p.m. The meeting reconvened at 3:23 p.m.

Tony Hey was asked to present the report of the COV to ASCR. Its charge was to assess the efficacy and quality of the award processes used during the past three years and, within the boundaries defined by DOE missions and available funding, to comment on how the award process has affected the breadth and depth of ASCR's research portfolio, the degree to which the program is addressing the challenges of multicore hybrid computing and petascale-to-exascale scientific data management, and the national and international standing of the program.

The review process began with an overview of ASCR and its operations by Office's staff. The budget for the Computer Science program had risen from $30.7 million in FY09 to $47.3 million in FY11; however, although Lucy Nowell and Sonia Sachs were now full-time federal program managers, it was clear that the CS program was still under-resourced. The ASCR Computer Science Research Program falls into five general categories:

- Operating and file systems,
- Performance and productivity tools,
- Programming models,
- Data management and visualization, and
- Extreme-scale architectures.

It had convened 14 workshops on the scientific challenges and technology issues posed by extreme-scale computing. It issued four FOAs (in 2008, 2010, 2010, and 2010). Success rates were considered reasonable. There were renewals for 3 years after the respective awards. Discussions were held on international collaborations and the exascale initiative. Peer-reviewed applications from FOAs were reviewed, and nothing untoward was found.

The program is well-managed and fairly administered. The COV made five specific recommendations in this area:

- Continue to improve the online information management capabilities of the program (and related ASCR programs that incorporate computer science research), informed by an overall plan, and by best practices from other funding organizations, such as the National Science Foundation (NSF) and the National Institutes of Health (NIH).
- Expand the information management capabilities to incorporate a reviewer database that records areas of expertise, quality of past reviews, responsiveness, and conflicts of interest, and a PI database that identifies previous successful and unsuccessful DOE proposals, links to research and project websites, and all currently active DOE-funded projects.
- Introduce mechanisms to provide balanced and knowledgeable reviewers by using a less crude, more refined approach to conflicts of interest.
- Provide a longer-term, more coherent schedule of planned solicitations, adapted as necessary to budget contingencies and ongoing research advances.
- Incorporate some mechanism for funding the exploration of promising new ideas that might not conform to the planned research programs.

The efficiency and quality monitoring of active awards is well done, but budget restrictions have put burdens on program managers, and more-efficient methods will be needed. The COV made three specific recommendations in this area:

- Computer science program managers should be encouraged to consider how new technologies and new media, including social environments and hubs, could be used to provide more efficient oversight.
- Better metrics should be developed for evaluating the impact and future needs for workshops and other conferences used as oversight mechanisms.
- A team approach needs to be developed to utilize the staff of ASCR and the Computer Science program managers most efficiently while maintaining adequate oversight of current research activities.

Concerning the breadth and depth of the portfolio, the awards were appropriate. There is a lot of high-quality research being supported. There is a tension between doing new work and developing libraries and maintaining expertise. The COV made two specific recommendations in this area:

- It is important that ASCR's CS program maintain a balance between its focus on exascale research and the traditional research strengths of the CS research groups at the DOE labs.
- The CS program should consider the importance of research into energy-efficiency, machine learning and data analytics for exascale systems within the context of its overall planning for the exascale computing, and more prominence should be given to these topics in future solicitations.

The programs on multi-core hybrid computing and peta-to-exa scientific data management have been progressing. Machine learning is a major lack. The COV made three specific recommendations in this area:

- The review panel should ideally contain a mix of university and DOE Laboratory researchers.
- The CS program should work with the BES and BER experimental data communities as well as ASCR's traditional simulation and modeling community in its scientific data management and analysis program.
- ASCR should consider setting up a research program to build expertise in Machine Learning and Data Mining technologies in support of the Office of Science's data mission.

On the international standing of the portfolio, China and Japan are committing resources to develop chip-design expertise. The COV made two specific recommendations in this area:

- ASCR should do all that it can to ensure that it receives sufficient investment in exascale for the United States to remain internationally competitive.
- The program should maintain its leadership role in high-end computing by continuing to engage with the international community.

The COV also made some general comments: The exascale is an important part of the ASCR portfolio. Resolving its uncertainties will be very helpful. There are three unfilled program-manager positions.

Graham added that the documentation of data has improved, and the staff needs better tools to work with.

Giles noted that (1) the recommendation on conflict of interest needs to be rephrased to emphasize that the current rules are too strict; (2) there is not a finding that the schedule has been pushed back to a certain date; and (3) ASCAC cannot work with ASCR to secure funding.

Chan asked if there were any directions given for developing a workshop metric. Hey replied that one could have a discussion about outcomes or counts of proposals from workshop participants. Graham added that some people feel put upon being asked to attend workshops. There may be a better mechanism to gather information.

Blumenthal noted that one may get only a partial response when one has a recommendation about two things. Data analytics and data mining should be split.

Giles called for a vote to accept the report of the COV subject to editorial changes. The vote was unanimously in favor of the motion. Giles thanked the Subcommittee and its leadership.

Bronson Messer reported on early science on the Titan machine at the Oak Ridge Leadership Class Facility (LCF). It is an upgrade of the Jaguar Cray from XT5 to XK6. The first phase has been completed: the replacement of the cores. Fermi GPUs have been put on 960 nodes. Kepler GPUs will replace the Fermis and will be put on 14,592 minus 960 other nodes to produce a 20+ PFLOPS peak performance.

From the outset, six projects were accelerated: Wang-Landau–Locally Self-Consitent Multiple Scattering (WL-LSMS), S3D, Non-Equilibrium Radiation Diffusion (NRDF), Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS), Community Atmospheric Model–Spectral Element (CAM-SE), and Denovo. NRDF is used as a testbed for a lot of advanced mathematics.

A big component in how these projects are selected is the Center for Accelerated Application Readiness (CAAR) algorithmic coverage (Phil Collela's seven dwarfs: fast-Fourier transform, dense linear algebra, sparse linear algebra, particles, Monte Carlo, structured grids, and unstructured grids). Also, the programming languages, libraries, algorithms, grid type, etc. used are noted to ensure that as many as possible such aspects are covered.

These applications will have a lot of time on the machine immediately upon startup, using one or more of the access programs. Acceleration by a factor of 3 to 12 is expected. Some community codes have been run on the machine and have shown a speed-up of a factor of 1.5 to 3.0.

There will be a short transition period after acceptance of the machine. The acceptance test will include mockups of these codes. The transition period will ramp up the machine over a dozen weeks, with "friendly" codes being run first and INCITE codes being slowly introduced. The runs selected will use from 10,000,000 to 150,000,000 core hours on the Titan.

The science problems that will be run will include:

- S3D in a compressible Navier–Stokes equation solver; the re-factored code is 2 times faster on the Cray XT5
- Denovo is a 3-D radiation transport code written in C++ and using CUDA; the re-factored code is 3.2 times faster on the Cray XT5
- LAMMPS, written with CUDA, addresses the classical N-body-problem solver of atomistic modeling; the refactored code is 3.2 times faster on the Cray XT5
- The Wang-Landau LSMS calculates the statistical mechanics of magnetic materials by combining classical statistical mechanics with first-principles calculations; it is 1.6 times faster on the Cray XT5

- CAM-SE is a spectral elements community atmosphere model that employs equal-angle, cubed-sphere grids and terrain-following coordinate systems; it is 1.7 times faster on the Cray XT5

Results are expected to be better with the Kepler processors. Hyper-Q will be obtained with 32 simultaneous work queues, allowing multiple CPU threads and GPUDirect enables GPUs to directly exchange data without needing to go to a CPU/system memory.

Hybrid computing is now mainstream with Titan, Blue Waters, Stampede, etc. Therefore, a lot of community codes are making the leap to GPUs (e.g., Chroma and QMPACK and CP2K). However, we are not going to wait for the communities; rather, we will conduct field tests and educational training events (conferences, workshops, tutorials, etc.) so users will be able to expose hierarchical parallelism; use compiler directive-based tools; analyze, optimize, and debug codes; and use low-level programming techniques, if required.

In summary, via CAAR and work by colleagues at NVIDIA, the Swiss National Supercomputing Centre (CSCS), National Center for Supercomputing Applications (NCSA), and other organizations has produced a set of vanguard GPU-capable applications. More than a dozen codes are poised to make immediate use of Titan. Strong overlap with INCITE and the ASCR Leadership Computing Challenge (ALCC) awardees, combined with the flexibility afforded by our Director's Discretionary (DD) Project program, will allow us to facilitate first-in-class simulations immediately following Titan acceptance, producing significant scientific results in short wall-clock times.

Giles asked, in terms of publicizing results, whether a lot of visibility will be evident at a certain conference in November. Messer replied, yes; and a lot of press had already been received.

Meza asked if there had been any lessons learned. Messer said that it was planned to do that in the training sessions. Many broad-based things have been learned:

1. The code selection is important.
2. One wants to maximize what one gets down the line.
3. One can get help from people who have been working in the field for years.

Meza said that this is a good way to sell the exascale to industry and users.

Chan asked which ones solve questions that could not be solved before. Messer responded, all of them, and all have scientific significance.

The floor was open to public comment, but there was none.

Daniel Hitchcock was asked to present the new charge to the Committee. A data explosion is occurring everywhere in DOE: in simulations, genomics, HEP event-style data, light sources, and climate (which produces hundreds of exabytes). All of the exascale hardware trends impact data-intensive science. The data problem leverages investments in exascale computing to maximize the impact on the science missions. Data from instruments are still on an 18- to 24-month doubling path because of the huge data production by complementary metal oxide semiconductor (CMOS) detectors. Significant hardware infrastructure is needed to support this data gathering, which probably will not be replicated at users' home institution.

The Office's current activities in this area include SciDAC SDAV, mathematics of large data, data intensive codesign, the data portfolio in Computer Science, and workshops. A charge letter asks what part of the big data apple DOE should bite off and what is the intersection of

research needed for exascale computing and exabytes data. What would a machine-learning program look like in DOE?

Meza asked if he were asking what ASCR should bite off or what DOE should bite off. Hitchcock replied, ASCR, mostly for the SC applications, all in the open, not classified.

White asked what was included in "big data." Hitchcock responded that the committee gets to define what "big data" is. There has been a lot of work done on federal investments in data. This is not a charge where the Office knows the answer and wants validation. The Office needs people outside the Office to tell us what they think. The parts dealing with scale are ASCR's because no one else will do them.

Glotzer noted that DOE defined and developed HPC and pushed the architecture. In big data, there are two issues: (1) big data for the Department and (2) infrastructure everyone needs. If DOE does not take on that role, who will? Hitchcock answered that there are parts of the big data that DOE cannot afford to take on (e.g., automatic text translation of streaming Farsi). There are other things outside the DOE mission. Having a scope that increases ASCR's budget by a factor of 10 would not be helpful. Giles pointed out that it is part of the charge to figure this out.

Petzold said that the areas chosen were very good, but the analysis of Twitter feeds could be considered part of the DOE mission. Hitchcock responded that trying to change behavior by Twitter feeds would get us in trouble. Petzold said that DOE could be a testbed source on climate information. Hitchcock agreed that that is important but pointed out that it is hard to see how ASCR could ask OMB and Congress for money to do that. It is important to limit the scope of discussion to the doable.

Dongarra said that there may be scope here to work with the NSF. Hitchcock said that that is part of the question. Who is doing what [e.g., NNSA, Defense Advanced Research Projects Agency (DARPA), and NSF] is important here.

Hey put forward three thoughts: (1) BER, BES, and Fusion Energy Sciences (FES) have needs to be met. (2) Others (e.g., NCAR) already do a lot of this research, and we do not want to replicate it. (3) There needs to be a core of expertise and competence.

Giles pointed out that the report for this charge is due in March 2013. He asked if that were a critical date. Hitchcock replied that to delay the deadline and to expand the scope is the death spiral for information-technology (IT) projects. If it is developed by March, it becomes part of the FY14 discussions in Congress and part of the FY15 planning.

White asked how much definition of the intersection had been done. Hitchcock answered that there is a lot of information out there, especially from workshops and the writings about the Systems Biology Knowledgebase (KBase) in BER. There has been a lot written about "big data" that will help the Committee to decide what to look at.

The floor was opened to the public. Greg Bronevetsky said that, in modeling of computer systems, ASCR should recognize that (1) computer systems themselves are sources of the data and (2) CERN runs giant clusters distributed around the world, and ASCR can do better than one physics laboratory.

There being no other public comment, the meeting was adjourned for the day at 5:21 p.m.

Wednesday, August 15, 2012

The meeting was reconvened at 8:29 a.m.

William Harrod was asked to present an update on the exascale effort. ASCR does not have a budget, yet. The world has flattened out: performance, speed, and frequency are not increasing. Multiple cores are being used to increase speed, but those cores need to be kept busy. Parallelism is imperative in increasing performance. The Top 500 list indicates that an exaflop computer could be made by 2020.

A million processors are possible, but the leading computers today use 10 MW of electricity. An exaflop machine might use 20 MW. A lot of work has gone into achieving a 20-megawatt exaflop machine. High energy efficiency is the name of the game. Radically reducing overhead, both in moving data and setting up problems, is an important strategy. Optics show promise. There is a lot of focus on software. Building an exaflop computer is a grand challenge. Changing the technology to field an exaflop computer is a major effort. Programmers and system software engineers need to control data movement. The system needs to be highly programmable. System-wide checkpointing will not be possible.

There is a plan, but it is still under development. The strategy is to conduct critical R&D efforts, develop exascale software stacks, fund computer technology vendors to move required technology from research to product space, fund the design and development of exascale computer systems, conduct a joint effort with NNSA, and collaborate with other U.S. Government agencies and other countries. The next 5 years need to be spent influencing software vendors to develop exascale-effective software; the technology needs to be matured.

Cerf noted that Harrod had not mentioned the fundamental nature of computation. Instead, the current vectors of technology were emphasized. He asked if something revolutionary were needed. Harrod replied, no. Many issues are being looked at for much more refined computing. The operating system is being looked at, and a solution will be reached in FY13 if there is a budget for FY13. New architectures and abstract machine models are being considered to give people something to focus on. Performance modeling and analysis will have to be highly dynamic. Vectorized computing will have to be faced whether an exaflop machine is built or not. Solicitations are being issued on various programming environments, extreme-scale algorithms, and cross-cutting technologies. One needs a plan to deploy the technology into the vendor space. A workshop is being held on data-intensive computing.

A Council has been established to investigate issues, challenges, and solutions for the DOE software plan for developing exascale operating-system and runtime software. The plan is due in November 2012.

An FOA was issued on June 8, 2012, for resilient extreme-scale solvers (RX-solvers) to establish a foundation for research on extreme-scale scientific computing. Proposals are being received, and initial reviews are under way.

An earlier FOA had solicited proposals for the X-stack, focusing on programming models, languages, runtime systems, and new energy-efficient and resilient programming techniques. It was issued Nov. 22, 2011; peer review was conducted in April 2012; and funding recommendations were completed on June 8, 2012.

FastForward is an effort to influence vendors to develop advances in processes, memory, and other technologies that they otherwise would not undertake. The codesign centers are considered to be the center of the universe of exaflop computing. A prototype-build phase is planned, but no acquisition is now planned.

Cerf asked if there were any problems of export control. Harrod replied that there are some discussions with the Europeans. Export control on computers has largely failed. That does not mean that such a policy would not be adopted by some government or administration.

Software will be open source. Vendors will not do anything that will hurt their products and revenues. The real challenges are avoiding mediocre solutions, practicing codesign, developing a new software stack for exascale systems, exploring radical concepts rather than developing practical solutions, and designing new computers based on a new execution model.

White asked where modeling and simulation were in the plan. Harrod said that they are spread across the whole plan.

Blumenthal asked if DOE had any allies. Harrod replied that there are a lot of allies and interest in exascale computing, but no one else is interested in building an exascale computer.

Negele stated that there is a tremendous number of scientific needs that can only be met with the exascale. The urgency of moving down this path must be emphasized. Harrod agreed that a lot of people in this government recognized the importance of this effort.

Giles noted that DOE's opportunity for leadership in this arena is slipping away as funding is awaited. Harrod said that the United States has led HPC for a long time. It is now trying to change how computers operate. Other countries are looking at this problem, also. It would not be desirable to have some other country or vendor develop an exascale computer that does not meet DOE's needs.

Giles asked if a 100-PFLOPS computer dilutes the exascale effort. Harrod said that the main concern is pushing technological change into the vendor space.

Greg Bronevetsky was asked to review his research on application-behavior analysis.

Future HPC systems will face significant constraints in terms of performance, usability, power, and accuracy. This will make systems more complex, including deep, heterogeneous memory hierarchies, heterogeneous processors, and caps on power. As a result, applications will need to become more complex and adaptable to run productively on these systems. Optimizing system and application productivity is very challenging and must take into account many fine details of hardware capabilities and application needs. Significant analysis of the software's behaviors and resource needs is required to achieve such optimization.

The approach pursued here is based on four components:

- *Measurement* of application and system behavior,
- *Analysis* of behavior based on these measurements,
- *Deployment* of application-level adaptivity into real applications, and
- *Action* to configure the application to improve its efficiency and performance.

This approach has been validated in the context of sparse linear algebra computations, which are representative of complex HPC applications because they are scalable, they are a key subroutine in many scientific applications, and their behavior is highly dependent on the properties of the input data.

One strategy pursued was to explore ways to save power by lowering processor voltage at the cost of some computation errors. Algorithmic resilience techniques were used to reduce the effect of these errors to acceptable limits (power versus accuracy). Errors in sparse linear algebra can be detected by multiplying the computation by a check vector, which projects the error onto the vector. However, the results are highly dependent on the properties of the inputs and the choice of the vector, leading to erratic performance and erratic resilience. A matrix-sensitive

statistical model of detector effectiveness was trained, and the model-guided error detector was found to be consistently efficient and accurate across the input space. The key lesson learned is that effective operation of these resilience techniques requires measurement and analysis of their behavior across their input space. These measurements and analyses are used to drive the configuring of the application to use the most cost-effective technique for each input.

A second thrust of the research has focused on performance of sparse linear algebra applications. Scalable performance is achieved by distributing computation across many processors in a way that balances computation and minimizes communication. To make this possible, a model was developed that predicts the algorithm's computation and communication needs for each matrix row. This procedure makes it possible to allocate work intelligently in a way that is tailored to the needs of each input matrix. This technique results in significantly better performance, with models of both communications and computation cumulatively contributing to the speed-up.

To make model-guided optimization available to real HPC applications, in-deep research has been conducted in each of the four main thrusts. For the *measurement* thrust, improved techniques were developed to quantify application behavior in terms of actual use of system resources. The main insight gained is that, instead of relying on low-level hardware-centric metrics, application behavior must be captured in terms of how resource availability affects its behavior. To this end, Active Measurement was developed; it employs threads that use a known amount of each system resource to make these resources unavailable for use by the application. The effect of this interference on application performance can indicate the application's utilization of each resource and how contention for each resource affects performance; in addition, efficiency can be quantified.

For the *analysis* thrust, research was conducted on detection and localization of system faults, focusing on detecting, localizing, and characterizing complex faults that reduce application performance but do not crash the system. This characterization requires precise measurements and models of application behavior, which is highly irregular, with a few code regions that are strongly affected but that behave normally for the most part. Because a given code region's vulnerability to a given fault is unknown, filtering must be performed to infer this information and to focus the analysis on only the abnormal behaviors resulting from the fault. This inference of hidden influences significantly improves the accuracy of fault characterizations, enabling more effective operation on complex, large-scale systems.

Modeling makes it possible to exploit and manage an application's flexibility. For the *deployment* thrust, modeling is being made more effective by using compiler transformations to create new flexibility. Compiler analyses are being developed that leverage the semantics of libraries to enable library-specific transformations that exploit the libraries' full capabilities. The current focus is on MPI applications. An implementation of MPI has been developed in which ranks are OS threads and can thus communicate via direct copies or by passing pointers. To make these techniques practical, a compiler is needed that can handle the complexity of today's applications. Because real applications feature many complex behaviors that require different types of analyses, a new compositional symbolic analysis framework is being developed that enables analyses to use the results of other analyses without knowledge of application-program interfaces (APIs) or the abstractions that each implementation may use. The composition of independent analyses thus enables complex transformations of real applications.

Incorporating model guidance into existing runtimes is complex. For the *action* thrust, a custom work manager has been developed to prototype model-guided optimizations. This system is significantly simpler than full-scale runtime systems are, but it incorporates their key properties. Further, it significantly simplifies the development and validation of models that can capture behavior by allowing developers to separate applications into tasks clearly and to identify application-specific task inputs explicitly.

Detailed analyses of exascale software are needed. Behavior analysis and modeling are required to use applications and systems productively. The utility of this approach has been demonstrated in representative use-cases. Research is ongoing to increase the capability and generality of the approach.

Graham asked how useful it would be to accumulate analytical data over time. Bronevetsky answered that this is statistical analysis. More data would allow one to refine the models and their predictive power over time.

Williams asked how well this system would benefit Jablonowski's output and analysis. Bronevetsky replied that each module receives errors and produces errors. One should quantify the errors and model the relationships among performance, energy efficiency, and result accuracy to intelligently select application configurations that effectively balance these parameters.

Negele noted that Bronevetsky had correlated application parameters with performance and asked how strong a correlation was seen and whether there were any likelihood of being able to tell the developer which parameters correlated more strongly with, say, energy. Bronevetsky responded that error rates are 40%. They are trying to reduce that rate to 20 to 30%. Guidance can be given in two ways:

- Separating the application into regions that behave consistently (e.g., loops or functions) and
- Identifying properties of application inputs (e.g., matrix sparsity or condition number) that may correlate with its behavior.

Also, because the effectiveness of the approach depends on accurate compiler analyses and transformation, developer-provided constraints on possible application behaviors (e.g., a no-alias annotation or more detailed specification via an embedded domain-specific language) can significantly improve the capabilities of these compiler tools.

Paul Messina was asked to review the early science being done by the 16 early science applications at the Argonne LCF.

The big system Mira is at ANL. There is water flowing through it, but it has not started its acceptance phase. Since March, 1 of the 48 racks has been available for use.

A three-way codesign effort among Lawrence Livermore National Laboratory (LLNL), IBM, and ANL led to the Blue Gene architecture. On the Blue Gene/P (Intrepid), each node has four cores, one hardware (HW) thread per core, 2 GB of memory, double hummer, and a peak of 13.6 GFLOPS. The maximum parallelism is 163,840.

Blue Gene/Q (Mira) has 48,000 nodes, 768,000 cores, 786 TB of memory, a peak of 10 PFLOPS, and 35 PB of disk storage; each node has 16 cores, 4 HW threads per core, 16 GB of memory, and a peak of 205 GFLOPS; the maximum hardware parallelism is 3,145,728. The system is installed as 3 rows of 16 racks. One has to have lots of disk space. It is expected that 100 MB of storage will be needed for visualization and data analytics. In comparison to Blue Gene/P, it has

- 15 times the FLOPS per node,
- 3 times the memory bandwidth, and
- –74% the latency.

It also has 35 petabytes of long-term (months to years) storage. It can be configured for non-homogeneous operation. One can run various MPI ranks/node and threads/node (2 to 64).

Cerf asked what the messaging was used for. Messina replied, for data parsing.

For the 10 science applications benchmark suite, the architectures are similar. They cannot be manually tuned; only compiler optimizations are allowed. Three of the applications are threaded; the remainder are 100% MPI applications. For 100% MPI applications, multiple MPI ranks per core were tested. For MPI + OpenMP applications, 1 MPI rank per core and multiple OpenMP threads per core were tested. For core-to-core comparison, one gets 3.8 times the FLOPS per core and 15.1 times the FLOPS per node. If one plays with the ranks/node, one can get different Blue Gene P/Q ratios (speed-ups ranging from a factor of 5.9 to 9.1) with FLASH. Five models have been imported now. There is a performance-tool project to help people figure out what is going on in performance.

Sixteen projects were chosen for early science use of the system for which they get a postdoc and 2 billion core hours. The early science program was launched to prepare key applications for the architecture and scale of Mira and to solidify libraries and infrastructure. The projects have a running start for delivery of exciting new science.

There is a good spread of applications. Some projects completed 16-rack runs, and one has been given full 48-rack access. The early science runs are expected to be conducted in the second half of calendar year 2012. Mira is committed to go live for INCITE on Oct. 1, 2013, with 768 million core-hours for allocation, although that allocation may be increased. All 16 projects are running on Blue Gene/Q. NWChem was ported to Blue Gene/Q in 3 hours.

Some of the early science projects are

- Cosmic structure probes of the dark universe
- Hardware/hybrid accelerated cosmology code (HACC) framework
- High-speed combustion and detonation
- FLASH (a compilation of codes for astrophysics, cosmology, high-energy-density plasmas, and incompressible fluid dynamics)
- GAMESS (General Atomic and Molecular Electronic Structure System)
- NAMD [Not (just) Another Molecular Dynamics program], the engine for large-scale classical molecular-dynamics simulations of biomolecular systems based on a polarizable force field
- Tokamak plasma microturbulence
- Materials design and discovery: catalysis and energy storage
- Lattice quantum chromodynamics
- Direct numerical simulation of autoignition in a jet in a cross-flow

Early experience confirms that Mira will enable advances in a broad spectrum of applications. The Early Science Program is paying off. All the applications are running. Many valuable insights on tuning and scaling are being obtained. Applications are being enhanced to model more complex phenomena with higher fidelity. The Blue Gene/Q Tools and Libraries

Project has yielded substantial software tools very early in the life of Mira. We look forward to exciting scientific results in the next few months.

Chan asked if a postdoc were being provided for each INCITE project. Messina replied, yes, they are called catalysts; each one has three INCITE projects. These are highly experienced people; not necessarily postdocs. We would like to have permanent employees in those positions. Sometimes, we hire on as full-time a postdoc who was working for an early science project.

Berger asked how nested parallelism in open MPI was done. Messina said that he did not know but noted that it was done dynamically.

Williams asked how much support was given to bring models onto the system. Messina replied that it depends on the project and whether an appropriate postdoc can found for a project. If one cannot be found, some of the effort is pushed back on the modeler.

Dongarra asked what constituted the tools and libraries effort. Messina replied that he did not know. Buddy Bland said that there was not an organized effort. There are many codes running at both centers (ORNL and ANL). It is a good idea. It should be done. Hitchcock added that the Office tries to get as much commonality as makes sense.

The floor was opened to the public for comment. There being none, a break was declared at 10:49 a.m. The meeting was called back into session at 11:07 a.m.

Thomas Lehman was asked to review traffic engineering for dynamically provisioned federated networks. A lot of this was initiated by DOE in 2007 with NSF's Dynamic Resource Allocation via GMPLS Optical Networks (DRAGON). ESnet developed the On-Demand Secure Circuits and Advance Reservation System (OSCARS). OSCARS was deployed as a production service in ESnet in mid-2007. About 50% of ESnet's total traffic is now carried via OSCARS circuits. It has been adopted by SciNet since 1999 to manage network bandwidth resources for demonstrations and bandwidth challenges. It was adopted by LHC to support Tier 0–Tier 1 and Tier 1–Tier 2 transfers. It is currently deployed in more than 20 networks worldwide, including wide-area backbones, regional networks, exchange points, local-area networks, and testbeds. It has been adopted by the NSF's Dynamic Networking System (DYNES), which will result in more than 40 more OSCARS deployments.

It is a simple service, sort of like an Ethernet. It has full control of services within a domain. An attempt is being made to give users a scheduling capability. A protocol has been developed to do the multidomain signaling.

Cerf asked whether the routine saw anything besides the Internet-protocol addresses. Lehman replied that it interconnects with VLAN. The real-time reservation protocol is complex and makes it easy for the user.

The driving vision is to treat all the network layers in a holistic and integrated fashion. The network needs to be available to application workflows as a first class resource in this ecosystem.

Today our dynamic provisioning systems see only a single layer. But the networks are really multilayer. We would like to provision services at a lower layer to create a topology element at the higher layer (a link between routers) and to offer services directly at the lower layer on a multidomain basis. These resources should not be left idle. Jobs should be scheduled to be responsive to user needs and to maximize the use of resources. The network needs to be able to respond to "What is Possible?" and "What do you recommend" questions. These are "intelligent network services." Multilayer control and intelligent network services drive the research.

In multilayer network control, routing domains are different between the layers. Vendor-unique functions and capabilities must be understood. The result of multilayer control is dynamic *topologies* instead of dynamic *services*. This situation can create instability in the network if not managed properly.

In intelligent network services, resource computation in response to open-ended questions can be complex and processing intensive. Being limited to "scheduled" services will help. For a single domain, one can have a single state-aware entity; but for a multi-domain system, one will likely need a two-phase commit type of process. A common capability in the form of multi-constraint resource computation is needed to enable both of these capabilities. Multidomain topology sharing and multidomain messaging also present challenges, but not to the degree of computation.

Advanced Resource Computation for Hybrid Service and TOpology NEtworks (ARCHSTONE) components are the Advanced Network Service Plane and Network Service Interface (which provide the "request topology" and "service topology," a Common Network Resource Description schema, and formalization of the application to network interactions) and the Multi-Dimensional Topology Computation Element (MX-TCE, which provides high-performance computation with flexible application of constraints). Multi-constraint topology computation is the main challenge to enable OSCARS to become multilayer network aware and to provide intelligent network services. One should use OSCARS v0.6 as a base infrastructure and as a development environment.

The technical challenges include: Topology computation is an advanced path computation process that is an order of magnitude more complex in the constraint and network graph dimensions. Traffic engineering constraints are categorized for subsequent treatment in the multistage computation process. This is a tractable problem. Different approaches were looked at: the constrained shortest path first, the constrained breadth first search, graph transformation, and heuristic search. Multiple combinations of these approaches were evaluated. The initial conclusion was a multistage K-shortest path (KSP) with ordering criteria for initial implementation. Future services may require other techniques.

The network graph is transformed to a different format and calculation algorithms are run over that. A channel-graph technique is similar. In the end, one can solve the problem and get answers quickly.

MX-TCE architecture has been implemented to provide answers to such questions. To make OSCARS understand the network, features have been added for multilayer topologies, multipoint topologies, requests in the form of a "service-topology," vendor-specific features, technology-specific features, and node-level constraints. The result is a schema superset to what OSCARS now uses.

In summary, to OSCARS was added the ARCHSTONE network "service plane" formalization, extensions to OSCARS topology and provisioning schemas, MX-TCE (multi-dimensional topology computation engine), and a new class of network services referred to as "Intelligent Network Services." There are other advanced network research activities under way: software-defined networking, OpenFlow, clouds, and network as a service. These are all tools. Implementing intelligence that can understand the network underneath it is being looked at. How extensive it can be is unclear. There are a lot of vendor-specific capabilities that could be enlisted. A lot of complexity has to be built in to make things easy for the user. The focus is on

heterogeneous technologies and control planes, multiple control and policy domains, multi-constraint resource computations, the need for flexible interaction with application workflows, and maintenance of service states.

Other ASCR research projects [the Coordinated Multi-Layer Multi-Domain Optical Network (COMMON) project; the (ESCAPES) hybrid network traffic engineering system, and the Virtual Network On Demand (VNOD)] complement the efforts. An effort is being made to integrate all of these capabilities. We are building a prototype network service plane with intelligent network services. This will allow people to lock in resources (e.g., three days in advance) of multiple networks and to choose their optimum usage.

Cerf noted that, in OpenFlow, a decision of flow is done on a packet-by-packet basis. This system maps out a topology and then senses what route is dedicated for any given packet. This is a hard problem. Lehman agreed.

Teresa Quinn was asked to review FastForward.

DOE's investment in the HPC industry has benefited the nation. FastForward is a story about two DOE offices investing in the U.S. HPC computing industry primarily to benefit DOE missions and secondarily to benefit the nation's economic competiveness. It was conducted by two DOE offices, seven national laboratories, and five U.S. companies. Between February 2012 and June 30, 2014, it is to fund $62.5 million of R&D for processors, memory, and storage technologies for a broad market with the objective of influencing critical HPC technologies. Contracts were awarded by June 29, 2012. DOE is now setting up 2-year collaborations with FastForward awardees. Why? The nation will fare poorly if it does not invest in technology for the future. Codes will need a lot memory and memory bandwidth, high peak performance, and efficient performance. One would need to buy 9 or 10 times the computer hardware to get a doubling or tripling of performance. The ratio of memory bandwidth and capacity to computing is actually shrinking. In addition, operating costs are expected to increase by a factor of 2 or 3 because of system power. Going up to 30 MW adds $150 million in electricity costs to annual operating costs.

High-value R&D is being sought to increase performance of DOE simulations, decrease energy usage, benefit the broad market, and be available in large-scale DOE systems in 5 to 10 years. FastForward has made five awards and has gotten most of the big players to participate. One or two have disruptive technologies that could be exciting.

Questions can be raised about why companies would care about serving the HPC market. HPC is a market leader. What HPC needs now, consumers will need later. Apple spends 2% of its revenue on R&D. R&D money is dear to these companies. The companies have to put in 40% of the FastForward costs and may or may not get a product out of their efforts.

DOE learned lessons from earlier programs like PathForward and Advanced Technology Systems:

- Investments are needed in R&D for technology and systems not just components.
- Investments in commodity technology are paid back over a long time.
- Long-term collaborations have the greatest potential for impact.

The next steps entail the possible funding of system R&D (because investing in technologies is not enough) and considering options to increase the likelihood that the R&D will be available in systems that we can buy in 5 to 10 years. Getting these projects awarded took the efforts of a large team of people.

Cerf agreed with the need for long-term investments and asked whether that was what this program was going after. Quinn replied that the program wants to address the need for disruptive technology. The participants are addressing memory, memory bandwidth, and power consumption with some very clever approaches.

Hey noted that ARM Holdings has some very low-power and asynchronous chips and asked whether they were considered in the procurement. Quinn answered that they chose not to participate, but others will likely employ the ARM technology.

The floor was opened for public comment. There being none, the meeting was adjourned at 12:08 p.m.

Respectfully submitted,
F. M. O'Hara, Jr.
Recording Secretary
Oct. 1, 2012