Meeting Minutes
ADVANCED SCIENTIFIC COMPUTING ADVISORY COMMITTEE (ASCAC)
September 17-18, 2018
Holiday Inn Capitol – 550 C Street Washington, DC 20024

ASCAC Members Present

Keren Bergman
Martin Berzins
Vinton Cerf
Barbara Chapman
Jacqueline Chen
Silvia Crivelli
John Dolbow
Thom Dunning
Tim Germann
Susan Gregurick
Anthony Hey
Richard Lethin
David Levermore
Satoshi Matsouka
John Negele (online)
Daniel Reed (Chairperson)
Krysta Svore (online)

ASCAC Members Absent

Jack Dongarra
Gwendolyn Huntoon
Linda Petzold
Vivek Sarkar
Dean Williams

Also Participating

Pete Beckman, Argonne National Laboratory (ANL)
Steve Binkley, Deputy Director for Science Programs, Office of Science (SC), Department of Energy (DOE)
Arthur "Buddy" Bland, Oak Ridge National Laboratory (ORNL)
Christine Chalk, ASCAC Designated Federal Officer, Program Manager, Oak Ridge Leadership Computing, ASCR, SC, DOE
T. Reneau Conner, Oak Ridge Institute for Science and Energy (ORISE), Oak Ridge Associated Universities
Emil Constantinescu, ANL
Barbara Helland, Associate Director, Advanced Scientific Computing Research (ASCR), SC, DOE
Mike Heroux, Sandia National Laboratory (SNL)
Travis Humble, ORNL
Doug Kothe, ORNL
Carolyn Lauzon, ASCR, SC, DOE
Paul Messina, ANL
Andrew Schwartz, Basic Energy Sciences (BES), SC, DOE
Fred Streitz, Lawrence Livermore National Laboratory (LLNL)

Attending

Francis Alexander, Brookhaven National Laboratory (BNL)
James Ang, Pacific Northwest National Laboratory (PNNL)
Laura Biven, SC, DOE
Ben Brown, SC, DOE
David Brown, Lawrence Berkeley National Laboratory (LBNL)
Richard Carlson, SC, DOE
Susan Coghlan, ANL
Leland Cogliani, Lewis-Burke
Michael Cooke, SC, DOE
Scott Collis, ANL
Kenneth Clark
Claire Cramer, SC, DOE
Eric Cyr, SNL

Lori Diachen, LLNL
Paul Doucette, Battelle
Richard Gerber, LBNL
Shaun Gleason, ORNL
Pamela Green
Mark Guston, Cray
Navindra Gunawardena, PNNL/ National Nuclear Security Administration (NNSA)
Jim Hack, ORNL
Aric Hagberg, Los Alamos National Laboratory (LANL)
Bill Hanson
Philip Harmon, University of California
Kurt Heckman
Bruce Hendrickson, LLNL
Dan Hitchcock, retired
Jeff Hittinger, LLNL
Paul Hovland, ANL
Mike Ignatowski, LLNL
Alexander Lazelere, US Council on Competitiveness (USCOC)
Steven Lee, SC, DOE
Nicholas Malaya
Michael Martin, LBNL
Sonia McCarthy, SC, DOE
Sandra McLean, SC, DOE
Chris Miller, DOE
Kathryn Mohror, LLNL
Tameka Morgan, SC, DOE
Gary Myers
Jeff Nichols, ORNL
Lucy Nowell, SC, DOE
Claire Oto, DOE
Mike Parks, SNL
Robinson Pino SC, DOE
Ashley Predith, DOE
Carly Robinson, Office of Science and Technical Information (OSTI), DOE
Julia Rowe, University of California
John Russell, Office of Senator Daines
John Sarrao, LANL
Mike Schulte
James Sexton, LLNL
Michelle Sneed, DOE
Julie Stambaugh, SC, DOE
James Stewart, SNL
Ceren Suset, SC, DOE
Peter Swart, LANL
Angie Thevenot, SC, DOE
Gina Tourassi, ORNL
Robert Voigt, Krell Institute
Tim Wildey, SNL
Stefan Wild, ANL
David Womble, ORNL
Carol Woodward, LLNL

Monday, September 17, 2018

OPENING REMARKS FROM THE COMMITTEE CHAIR

Dan Reed, ASCAC, called the meeting to order at 8:30 a.m. Eastern Time (ET) and introduced the first speaker.

VIEW FROM GERMANTOWN

Barbara Helland, Associate Director, ASCR, reviewed budget information and updates on ASCR activities and projects. The ASCR fiscal year (FY) 2019 budget from Congress is $935M; the SC budget is $6.6B. Research is allocated $130M, facilities $572M, and Exascale Computing Project (ECP) $232M; Computational Science Graduate Fellowship (CSGF) $10M.

Summit regained 1st place on the Top500 list and received 1st place on the Green500 list (Level 3). AURORA is expected in 2021 and additional applications have been accepted to run. A change in the procurement strategy enabled the National Energy Research Scientific Computing Center (NERSC) 9 to take advantage of recent growth in machine learning (ML) and data capabilities. ESnet6 upgrade is projected to finish in 2023.

ASCR awarded six new Early Career Awards in 2018 (3 in computer science, 3 in applied mathematics). Two projects have been awarded for Mathematical Multifaceted Integrated Capability Centers (MMICCs). Four projects were awarded by ASCR Computer Science in FY18. Two SciDAC accomplishments noted were for FastMath and Quantum Computing Applications Teams. There were three awards in FY18 for Quantum Testbeds Pathfinder, and two awards for Quantum Testbeds for Science. Two accomplishments in quantum were the first simulation of an atomic nucleus on Quantum Cloud, and Accelerating Applications of High Performance Computing (HPC) with Quantum Processing Units (QPU).

William Vanderlinde will join ASCR on October 1, 2018; SC approved backfills for Applied Math and NERSC Program Managers. Barb Helland highlighted awards given to Lucy Nowell, Daniela Ushizima, Jacqueline Chen, and Buddy Bland.

Discussion

Cerf asked about ESnet6's network implementation, CenturyLink using dark fiber, and special purpose hardware (HW) for ML. **Helland** said that ESnet 6's network plan is to start small, and then build the dark part of the network. The NVIDIA graphic processing units (GPU) are being used heavily on Summit for ML. However, no special purpose HW has been purchased but ASCR is looking at it. **Cerf** stated that Google's tensor processing units (TPU) are showing extraordinary success, and they are using cloud ML to address instability.

ARTIFICIAL INTELLIGENCE (AI) AT THE EDGE R&D

Pete Beckman, ANL, stated they stumbled into Edge when seeking a solution to the extremely high data yield (5.1GB image, 1 sample every 5 minutes, 1TB in a 12 hour period) from a field-placed sensor. The solution was a parallel computer for each sensor and to process the data in the field. Edge computing means a new method of computing coordinating inference and ML together. Edge computing is necessary because of issues with bandwidth, latency, privacy and security, resilience, and energy efficiency.

New ML codes that run in parallel on the Edge but are deployed remotely are desired. A mini rack controller is used to power up and power down every device remotely. The Edge computing platform, Waggle, allows experimental HW at the Edge that can be manipulated and managed remotely. Examples of using Edge computing are in the areas of transportation, hydrology/flooding, disaster planning, earth modeling, national security, energy, manufacturing, atmospheric science, and facilities – light sources.

Continually improving the Edge to HPC system means lightweight learning has to happen at the Edge and the server has to put together all of the learning. Scheduling has to occur at the Edge meaning there is a need for an Edge operating system (OS). Other research questions focus on the programming model, the ML optimized HW, theoretical foundations for failures, dynamic resource management, and fluid HPC.

Discussion

Cerf reinforced the potential brittleness of ML algorithms. He asked if real-world simulated data can be provided for the Edge operations and about security at the Edge in the control system. **Beckman** mentioned collaborations with Daniel Work at Vanderbilt University who is doing transportation modeling with data verified by analyzing actual traffic; and in Detroit, where feedback is needed to make a decision and have a local actuation. All Edge devices and computation have only one path to talk, and it only talks out; there are no open ports. **Cerf** confirmed communication is two-way but only initiated one way.

Chapman was concerned about the amount of data to be transferred between the device and the system. **Beckman** indicated it is still a problem in both directions. Taking sample data from the Edge to improve training data must be streamed at all times.

Levermore asked how the data processed at the Edge interacts with the mainframe. **Beckman** stated that one of the first challenges was time series data. Foundational work must occur in that space; combined and updatable heterogeneous and homogeneous systems, where small perturbations will make a big change in what is computed in the center, are needed. **Levermore** clarified that a deeper understanding of sensitivity issues is needed throughout the system. **Beckman** suggested it was a change in perspective. The notion is ML and computation must be run in both places and they have to be fused.

Bergman asked what technologies are used for communication to happen along the whole path from Edge to HPC. **Beckman** noted all available channels are used. A change in the model is required because moving to the Edge will put public networks in the middle. All things will be pushed together and make the data movement and security more problematic.

VIEW FROM WASHINGTON

Steve Binkley, Deputy Director for Science Programs, SC focused on three topics: DOE leadership, appointee status, and the FY19 budget. Christopher Fall was nominated for the Director of the Office of Science in May 2018. The Senate hearing for Dr. Fall was held June 26, 2018 and SC is awaiting Senate confirmation. Dr. Fall is currently the Principal Deputy Director of Advanced Research Project Agency-Energy (ARPA-E); he served with the Office of Naval Research and the White House Office of Science and Technology Policy.

SC's FY19 budget request was $5.391B; the House-Senate Conference budget is $6.6B. FY19 priorities include continuing operations of the national laboratories, continue exascale computing research, expand quantum computing (QC) and quantum information science (QIS) efforts, focus on robust cybersecurity program, cutting-edge, early-stage research and development (R&D), and maintaining interagency and international partnerships. SC's FY19 budget provides ASCR $935M.

Discussion

Cerf asked about DOE's interest in analog qubits and digital QC. **Binkley** clarified that DOE describes analog as quantum simulations and digital as QC. ASCR is interested in both, but more progress will be made sooner in quantum simulations. The most significant challenge in the general purpose, QC approach, is getting techniques developed for error correction.

Reed asked for the biggest areas of promise in SC. **Binkley** said microelectronics technology and lithography were two area of promise. Research is occurring at the national labs on architectural developments, using existing or future lithography's to develop novel architectures. Microelectronics play a key part in the U.S. economy and there are many national security issues associated with microelectronics fabrication.

Dunning asked for clarification on how the offices of High Energy Physics (HEP), Biological and Environmental Research (BER), and Nuclear Physics (NP) contribute to QIS. **Binkley** said HEP's contributions include quantum gravity and black holes (theoretical activities), and new sensors and detectors based on quantum principles of squeezing and entanglement. BER's investment is focused on QIS's impact on biological systems. NP's largest amount of funding is for development of isotopes to be used in devices and there are nano and materials developments in Basic Energy Sciences.

National Cancer Institute (NCI) UPDATE

Carolyn Lauzon, ASCR, explained the partnership between NCI and DOE beginning in 2016 with three calls for *whole of government* support to address challenge problems (National Strategic Computing Initiative (2015), Precision Medicine Initiative (2015), 21st Century Cures Act (2016)). Two current activities within the partnership are the Joint Design of Advanced Computing for Cancer (JDACS4C) and Accelerating Therapeutics for Opportunities in Medicine (ATOM). JDACS4C is composed of pilot projects at three different scales. Cutting across the three scales are CANcer Distributed Learning Environment (CANDLE) and Uncertainty Quantification (UQ). JDACS4C is completing its second year, the broader NCI community is more engaged, and workshops at Supercomputing conferences have been held.

Martin Berzins, ASCAC, provided an overview of JDACS4C and discussed Pilot 1-Pre-clinical Model Development and Pilot 3-Precision Oncology Surveillance, cross-cutting themes, and UQ. Pilot 1's aim is to predict how drugs will influence cancer and requires matching drug performance with the biomarkers of the patient. Pilot 3's challenge is with unstructured data, such as medical reports, which need deep learning (DL) algorithms; the idea is to determine patient health trajectories, see how well drugs will work, and decide what works and what does not by looking at population-level samples. CANDLE is looking at very large models and integrating UQ into each pilot. The pilot studies have contributed benchmarks for the A21 exascale architecture, influenced system requirements for new machines, been selected as one of the early science projects, and run Convolutional Neural Nets on Summit. DL is powerful in terms of pattern making, why it finds certain answers is not understood, and it can return unexpected correlations. Precision medicine needs trusted tools. In computer science a lot of emphasis is on making DL more reliable and more robust.

Fred Streitz, LLNL, reviewed the DOE-NCI Pilot 2-RAS Biology on Membranes which will build a large multi-disciplinary team to address computer system needs to model a biological system. ML is being used to simulate a micron-by-micron membrane and target where molecular dynamics calculations should be done. On Sierra, the molecular dynamics calculations run primarily on the GPUs and the phase-field calculation on the central processing units (CPU). ML, inference, and data management are all separate applications that must run and talk to each other. At any one time several hundred nodes are doing the phase-field calculations and then 1,000s of GPUs are used to do the other calculations, but these run across nodes. Scientists worked with IBM and Frederick National Laboratory to write code to allow that set of transitions to occur. When Sierra is accepted the team will look at the dependence of RAS mobility, aggregation of RAS, and effect of RAS concentration. ML and predictive simulation is envisioned as identifying bottlenecks and optimizing simulation runs; ML is used to guide and choose which phase-field simulations are done and to develop insights faster.

Discussion

Levermore asked if the RAS proteins undergo conformational changes and do those changes affect the way they interact with each other and the lipids. **Streitz** said that is one of the team's courses of study. Two states of RAS, one that admits the binding to the kinase (bindable) and one where that is shielded (non-bindable), are in the model now. Proteins are put into one of the states and everything equilibrates; the ability of doing very long-scale simulations in molecular dynamics allows us to see RAS change into different states. The team is trying to capture the conformations on the membrane. **Levermore** asked about the prediction objectives Pilot 2 is seeking. **Streitz** said reduced-order space has been done. Using a smaller number of parameters (smaller dimension) could recapitulate the conditions of the simulation.

Chen asked about orchestrating the multiscale aspects on a heterogeneous machine. **Streitz** noted two things that were developed in-house by IBM, the Flux Scheduler and the Data Broker. Using the Flux Scheduler allowed tasks to be split across components. The Data Broker is the mechanism by which all the pieces allow data to be shared across the applications.

Crivelli asked about the interaction of the drugs and molecules and about creating synthetic data. **Streitz** said there are two very different directions to go after the simulations, to understand the cascade and to develop the therapeutic; neither are inside the scope of this program. **Gina Tourassi** explained Pilot 3 synthesizes text according to data available from the surveillance community and uses the synthetic data with the experiments run on Summit to optimize the hyperparameter space for different models. A set of hyperparameters is created and brought back; experiments are run to further fine-tune the model with actual clinical data. There is a back and forth between the clinical and synthetic data.

Matsouka mentioned that the approach to using ML is anomaly detection, but in Pilot 2 what is going to happen is unknown. **Streitz** clarified that ML is being used to steer predictive sampling rather than to find anomalous events. **Matsouka** asked what had been enhanced on A21. **Berzins** said no details were given because of non-disclosure agreements.

Gregurick encouraged Streitz to consider electro-statics and the electronic nature of the protein, and asked if the pilots are using FDA-approved or pre-clinical trial drugs. **Streitz** said the team is looking at electro-statics and variations of the charge on the lipids and the proteins. **Lauzon** stated Pilot 1 is using pre-clinical drugs. **Streitz** added no new compounds are being considered, rather the data for a set of drugs used against the NCI-60 cancer cell line.

Hey asked about the problems IBM Watson ran into and how these projects are avoiding them. **Berzins** said IBM Watson promised too much and reached too far ahead. These projects are carefully structured with very definite goals, focused on well-defined scientific questions. The idea is to show people what could be done and to open the door to other areas of research.

Lethin asked Streitz about the bottleneck and if investment is needed to speed up any one part of the process. **Streitz** said the bottleneck is the quantity and speed of the molecular dynamics code. A specially written code to run on the GPUs was needed for scalability. Introducing topology to the membrane substantially complicates the mathematics and will end up being a larger part of the calculation.

Levermore asked if there has been any interaction with the National Library of Medicine (NLM) to make data available to academic researchers. **Lauzon** indicated that CANDLE and Pilot 3 have had a series of hack-a-thons to engage the community as well as workshops at Supercomputing conferences. They also have an Intellectual Property (IP) and Data Committee to ensure software (SW) and data products are available. **Carly Robinson** noted there are specific requirements under the Cancer Moon Shot Initiative about open data and open publications, those are all being met and are part of the IP management plan. There is not a direct representative from NLM, but the NIH program managers are very well versed in the policies.

Matsouka asked Streitz which aspect of molecular dynamics was the bottleneck. **Streitz** said even though it is the fastest part of the calculation it is a bottleneck because of the amount.

EXASCALE UPDATE – DOUG KOTHE, MIKE HEROUX

Doug Kothe, ORNL, provided an update on the ECP which consists of three technical focus areas, Hardware and Integration (HI), Software Technology (ST), and Application Development (AD). There are national security projects and open science projects; ECP is negotiating with Oak Ridge Leadership Computing Facility (OLCF), Argonne Leadership Computing Facility (ALCF), and NERSC on which applications will be prioritized and targeted for performance optimization. Kothe mentioned priority 1 and priority 2 targets for each set of applications, AD's challenge problems for each project, and a new co-design center, ExaLearn, for Machine Learning Co-Design. A Science & Technology Council is being formed. Kothe closed with a summary of ECP technical accomplishments, project management, and stakeholder relations.

Mike Heroux, SNL, shared the scope, goals, challenges, and highlights of the ST technical area. An ECP Capabilities Assessment Report was released in July 2018, and will be updated every 6 months. ECP ST contributes to 89 unique projects, and ECP ST staff contribute to standards groups. A SW Development Kit (SDK) is a collection of related SW products put together to improve usability and practices and fosters community growth. The motivation of SDKs began with the discovery that four math libraries could not be built together into a single executable. A side benefit of SDK is community policies which list important team building, quality improvement, and membership criteria. The SDK leadership is well-versed in specific SW and understands the SW ecosystem in their area. ST is creating a horizontal coupling by testing two similar products. ST's next steps include a review in early October 2018 and collecting data from two polls on SW practices.

Discussion

Dunning asked about the long term sustainability of the SW technologies being developed. **Heroux** mentioned that while sustainability is outside the scope of ECP the project can improve the quality of the SW to make it easier to sustain, can engage in the standards committees, and can put as much as possible into the hands of the facilities, vendors, and application users. There is also a quarterly meeting with ASCR program managers to exchange notes for awareness and consistency. **Kothe** added a document for transitioning off of ECP will be created. The document will include a description of the ECP projects and resources needed to maintain and support them. **Heroux** said ECP is also concerned about staff leaving before the project ends due to a lack of long term sustainability plans. **Dunning** stated a plan for ramping down and into the facilities' operating budgets would increase comfort and willingness to adopt technologies. **Heroux** noted that the products are not solely ECP funded; when ECP finishes it does not mean that all funding stops. **Helland** said that ASCR fully understands this problem.

Lethin asked if there are sufficient resources to ensure the efforts are funded to the extent needed. **Heroux** explained the teams get money from other sources, not just ECP and the ECP is often less than half of the teams' total funding. ST's CD-2 will have more expectations and funding per person or team to ensure success. Overall there is adequate funding, if not directly through ECP then through individual portfolios. A bigger problem is the lack of people in the DOE complex with the skill set needed for ECP.

Berzins asked what confidence Heroux has that the approaches being used will carry things forward. **Heroux** said ST is still at the beginning of the transition to the newer, higher concurrent execution models. He did not know if there is enough awareness, skill sets, or portable expression of parallelism to procure all of the SW capabilities past exascale. **Berzins** suggested two methods to carry applications forward: adopting very generic flexibility in SW design, and insuring anything that is available can be run. **Heroux** added the future technical assessment effort is one way ST is paying attention to the rest of the community. For example, a task-based execution model where allocation, initialization, and use of data within the same task is a best practice will be beneficial. While not required for using OpenMP, the code can be written that way to better prepare for tasking environments.

Levermore encouraged caution noting a balance in funding must be driven by meeting scientific needs and providing fertile ground for new ideas and new codes.

Reed asked if there will be a *must have* and a *desirable* extended set of SW. **Heroux** stated it is important to allow a larger portfolio of SW. In the heterogeneous computing world a runtime for every device on the node is needed. Programming models and application programming interfaces (API) will need at least one runtime. There is a special API for writing code for TPUs. Heroux expected an explosion in programming models and runtimes because of heterogeneity in the node architectures.

Matsouka asked to what extent ST engages the commercial SW developers to set up development disciplines. **Heroux** indicated where there are opportunities for collaboration or at least awareness, communication was key; specific engagements come from conversations. **Matsouka** asked to what extent classical SW engineering practices will be adopted. **Heroux** stated pull/ merge request codes are thoroughly tested against the entire regression suite and classical SW engineering practices are being used. For example, no piece of code gets into the repository if the entire test suite does not pass and if the coding policy is not followed. **Matsouka** emphasized the need for sustainability, on a systematic basis, to ensure the specifications are written down and the procedures laid out. **Heroux** said the SDKs promote that and the community policies statements are best practices. **Kothe** added that Altair Engineering, an Industry Council member, is scrutinizing what is done and offering advice. ECP is reaching out to industry to adopt best practices.

James Ang suggested looking at the System on a Chip (SoC) community because the OS can deal with a diversity of architectural IP blocks. **Heroux** said he hoped there is only one OS, but there has to be more than one runtime. Different devices exist because each presents a different execution model that is particularly well-adapted to a type of computation. Whether in the same or different languages, different programming models will be needed. **Ang** speculated that there will be an execution model that deals with heterogeneity. **Heroux** indicated industry can afford to hire programmers, who specialize in a given device, to work with a domain expert to produce a SW that runs across all devices. **Ang** asked about the vision for extreme heterogeneity. **Heroux** stated ECP is committed to using MPI + X where X is the execution and programming model of the node. ECP has multiple runtimes and multiple programming models.

Chen asked about optimizing mappers to map extreme heterogeneity and architecture to the runtime. **Heroux** considered mappers an effective approach for scientific computing, but not convinced of the same for portability.

Berzins commented that rewriting runtime versus rewriting every application that uses the runtime is a much more portable solution to many of these problems and involves a lot less work.

QUANTUM ACCELERATED HIGH-PERFORMANCE COMPUTING – EARLY CAREER AWARDEE

Travis Humble, ORNL, explained his early career research on QC. QC is quantum mechanical computation which manipulates the wave function to perform calculations. QC is epitomized by the Schrödinger equation and manipulated by changing the Hamiltonian.

Algorithms in a QC model take fewer steps to solve problems. This project is specifically asking are QPUs compatible with modern scientific computing. Through design (using modeling and simulation), can quantum-accelerated HPC demonstrate some performance advantages that makes the pursuit worthwhile?

Humble's approach is to begin with a quantum accelerator node component model which consists of a CPU, GPU, memory system, and a network interface card. The QPU is its own device, it is part of the node but can be decomposed further into a control unit, an execution unit, a register, and perhaps a quantum network interface card. On the node the program flow is controlled by the CPU, using a directive-based language that off-loads kernels onto the QPU. The framework is eXtreme-scale ACCelerator (XACC) based on Low Level Virtual Machine (LLVM). The node runtime environment is through a client-server interaction. The OS has to be an important part of the runtime; the execution model needs a scheduler inserting itself into how code is sent to the QPU and how quickly results are processed. A HW abstraction layer is also required to facilitate interaction with the system.

QPUs are known to have multiplied power if they interact making a quantum interconnect essential in the long-term for developing quantum-accelerated HPC. The quantum interconnect moves quantum information between the QPUs using quantum network interfaces. The simulation component is accomplished by using a Structural Simulation Toolkit (SST). QC in the future may provide an enormous gain in energy efficiency during computation. Humble closed with the goal for the quantum accelerator research project and broader community engagement in SW, benchmarking, and green QC.

Discussion

Bergman asked about implementing models in the quantum processes within the SST environment. **Humble** is following the component model (QCU, QEU, and register parts). Once a component receives information logic is implemented. Quantum physics simulations are done at the register level.

Gregurick asked if libraries of Hamiltonians will be available. **Humble** envisioned a hierarchy of users and that libraries will be available. However, developing the quantum mechanics reductions will require a new quantum SW engineer with a detailed understanding of the quantum algorithms themselves and the capabilities of the HW system.

Hey asked if QC could simulate large quantum systems. **Humble** illustrated this ability. The Hubbard model, in condensed matter physics, is an extremely important test case. With QC the Hubbard model can be specified through the Hamiltonian operator and the quantum computer could solve it by involving the quantum state under the Hamiltonian and observing it. **Hey** asked about experiments on the prototype systems and trapped-ion computing. **Humble** indicated he has access to several devices through ASCR and can program up to 20 qubits. Superconducting based qubits and trapped-ion qubits are the two leading commercial technologies and they have demonstrated sufficient control of the quantum mechanics to become programmable.

Crivelli asked what type of programs his team is currently running. **Humble** said the primary example is the variational problem Eigensolver method. A program asks the quantum processor to prepare a particular quantum state, similar to asking it to prepare a distribution, then a sample is taken from that distribution. If the right state has been prepared the samples drawn will yield a very good characterization of the minimum energy required for the problem.

Svore asked about device benchmarks. **Humble** said he has worked with the IEEE to discuss their role to support QC, the topic of benchmarks immediately came up in conversation and

National Institute of Standards and Technology (NIST) representatives were there. **Svore** asked about SW stack scaling as the machine grows, asynchronous calls, and classical feedback and classical control of operations that run in a state of coherence. **Humble** said to scale up better examples and test cases are needed. The event manager supports asynchronous communication.

PREDICTIVE SCIENTIFIC SIMULATIONS FOR COMPLEX SYSTEMS – EARLY CAREER AWARDEE

Emil Constantinescu, ANL, explained that there are accuracy challenges in complex systems and convergence degrades as physics is added. With added complexity there is a decrease in accuracy. Assessing predictive simulations in complex systems may include simple, complex, and stochastic models. Sources of error include simulation data, modeling and numerical errors, observational model, and measurements. Constantinescu reviewed and provided examples of numerical errors, model form errors, and probabilistic predictions. The project is providing an efficient way to quantify and control numerical errors, it is estimating model form errors by using spatio-temporal stochastic processes, and has developed ways to calibrate models with probabilistic solutions.

Discussion

Berzins asked about the challenges getting the methods into advanced simulation codes. **Constantescu** indicated just taking numerical errors does not provide a universal error estimate for the spatial discretization. There is not a general approach, integration will have to be incremental. He has introduced stochastic modeling by going from classical, well-established to more fringe methods.

Dolbow asked how the selection process is informed by the quantity of interest (QoI). **Constantescu** said the QoI provides a strategy to estimate point-wise error and is a forward estimate approach. The benefit is that the errors for all distinct variables are gathered in one shot, requiring only one quantum pass.

PHOTONIC INTEGRATED SUBSYSTEMS FOR NEXT GENERATION LEADERSHIP CLASS HPC

Keren Bergman, ASCAC, discussed the need for photonic interconnects in HPC and highlighted the collaborative SBIR program supported by ASCR. Trends of the top 10 systems from the Top500 list indicate that computation performance has risen at a much faster rate than communication at the node. Novel interconnect technologies and architectures are necessary to improve energy efficient data movement. The #1 machines increased from 6.1 GFlop/W to 18.4 GFlop/W energy efficiency in two years due to new development in interconnect technologies.

At the chip photonics starts at the same energy consumption level as electronics. Once the signal is modulated it can be sent over much longer distances without having to regenerate the bits and expend more energy like is required in electronics. Optics flattens the energy curve; communication can occur on the chip, on the DRAM, or across the system to a storage place and have the same energy consumption as on the chip bringing complete flexibility to the system.

Rich Carlson created a collaborative SBIR program bringing a small number of companies together to make one prototype. There was significant enough funding to develop something that can impact large-scale HPCs. Bergman shared examples of an optically-connected storage system, and an optically-connected memory.

Once the optical/ electrical input/output (I/O) are put together into a universal connectivity the speed or modulation format does not matter, the I/O propagates transparently through the waveguide. Having an optical fabric to connect the evolving heterogeneous nodes is compelling. Having a transparent, scalable high-bandwidth density type of connectivity in HPCs is both compelling and energy efficient.

Discussion

Berzins asked about limits to the all-to-all interconnects. **Bergman** said the optical technology is a limit, and making a high port count, single optical switch is challenging. To increase a typical optical switch to a larger scale, such as on Summit, requires scaling it to some type of topology which is possible in spatial switching but has optical losses. An optical switch is a broadband switch that provides both spatial switching and wavelength selectivity. **Berzins** inquired if this a technology might be available by the mid-2020s. **Bergman** considered that is very realistic. In electronics the fabrication, the chip, is very expensive but not the packaging. In optics it is the opposite, the chip is not very expensive but the packaging is.

Lethin asked if there is something about optics that make it ideal for an SBIR collaboration. **Bergman** said there is nothing unique about optics, that large companies and the labs could participate. However, small businesses are very agile and are able to collaborate easily. One goal of SBIR is to bring the technology to large vendors and small companies are eager to have the large companies be their future customers.

Bland asked if optics have a place in commercial machines at a smaller scale, or if they only become feasible in large scale systems. **Bergman** stated data centers are the main market for the optical technologies. A potential market is the Internet of Things; when optics are in the chip everything is lighter, energy consumption is reduced, and bandwidth is higher.

Matsouka asked about latency reduction. **Bergman** said that using photonics in other modalities is the subject of research. For example, her team is using the wavelength domain and much shorter latency to do accelerated neural networks. The physical distance will determine the ultimate latency. **Matsouka** added that having multiple wavelengths is inherently advantageous. The electronics determines the energy. Going at a slower rate per wavelength will be more efficient than high rates on a smaller number of channels.

PUBLIC COMMENT

None.

Reed adjourned the meeting at 5:38 p.m. (ET)

Tuesday, September 18, 2018

OPENING REMARKS FROM THE COMMITTEE CHAIR

Dan Reed, ASCAC, called the meeting to order at 8:30 a.m. (ET)

NEW CHARGE – TRANSITIONS

Barbara Helland, ASCR, shared a new charge to ASCAC concerning the identification of key elements of the ECP that need to be transitioned back into ASCR/ SC at the end of the project to address opportunities and challenges for future HPC capabilities.

Discussion

Dunning asked for the level of funding in the last year of the ECP. **Helland** said the last year of funding depends on the outcome of the current review. The highpoint of funding is the 2019 ASCR request of $232M.

Reed asked if Helland envisioned continuing R&D for the technologies that were part of ECP or for something completely different. **Helland** hopes to build Extreme Heterogeneity and Scientific Machine Learning under the current administration's priorities. She is open to where to support the technologies and for ASCAC to indicate what ASCR needs to keep from ECP going forward.

Dunning explained that ECP has assembled excellent teams that understand the challenges associated with new technologies and know how to convert those into science and engineering advances. **Helland** said building up SciDAC is critical and is a negotiation with the other Associate Directors. She is relying on ASCAC to help identify what to argue for. **Reed** added there is not a single way to view this, there are multiple parameters. **Helland** agreed and said the ECP meetings are phenomenal, very strong teams have been built and they communicate well.

Chen echoed Dunning, it takes years to build interdisciplinary teams around a given grand challenge problem. **Helland** explained the teams were carefully chosen; they had big footprints in the programs. Kothe and his team tell every program office of the accomplishments achieved by the ECP projects.

Ang stated that co-design is a multidisciplinary activity. Projects like ECP are needed to pull together all the disciplines and focus them on a concrete objective. He recommended that ASCAC consider how to retain the co-design, multidisciplinary focus as ECP transitions. **Helland** noted that ASCR has SciDAC and the co-design institutes grew out of SciDAC. ECP brought investment from vendors outside the facilities.

Jeff Nichols, ORNL, emphasized how important sharing the ECP projects between the weapons labs and SC labs has been. ECP is an unprecedented combination of successful projects that have not been enabled in the past by the joint combinations of SC and NNSA labs. The SciDAC program was successful in bringing domain-sciences from other areas together. What ECP did that was unprecedented was gave ASCR an opportunity to own domain application development. This is a challenge where ASCAC can help with recommendations on maintaining and supporting the applications going forward because the facilities will not have equal expertise across all of the domain applications. **Dunning** stated that SW needs to be thought of as infrastructure, like a facility. The ECP SW needs to be maintained just the way instruments in the facilities need to be maintained. **Helland** reminded folks that work on exascale began in 2007 with Town Hall meetings. **Matsouka** stated that co-design is a perpetual process, especially when there is a needed boost in capital; SW is an asset.

Messina suggested having an office whose responsibility is SW infrastructure; something with a label, something institutional.

Reed expressed concern over striking the right balance in the ASCAC recommendations. **Helland** reminded ASCAC that the charge is only to identify the key elements. **Reed** said he was wondering if ASCAC would be able to have that conversation ignoring resources. **Helland** indicated she welcomed a report that recommends the amount of funding required.

UPDATE ON SUBCOMMITTEE DOCUMENTING ASCR IMPACTS

Paul Messina, ANL, reminded ASCAC of the charge concerning ASCR impacts. Multiple examples of the types of materials gathered were shared as were the areas to be covered in the report. Although there is not yet a draft of the report the subcommittee is moving forward. Messina requested input from ASCAC.

Discussion

Dolbow mentioned a CSGF longitudinal study in 2017 on Krell Institute's website. **Chalk** said ASCAC completed a CSGF report in 2008.

Hey commented because of continuity, longevity, and enlightened investment in libraries, PETSc, and MPI, DOE has SW used around the world.

Dunning mentioned the national labs have been sources of expertise that have flowed out into industry, universities, and other areas; DOE has boosted the workforce with that expertise.

Reed mentioned the huge impact access to parallel computing machines had on education. **Helland** ASCR's predecessor had a number of education programs in the 1990s such as Adventures in Supercomputing (AIS) and the Undergraduate Computational Science Education (UCSE) group; CSGF wrote the first computational science online manual.

Chapman asked about international stories. **Hey** asked if MATLAB was included. **Messina** said Cleve Moler, MathWorks cofounder, mentioned the influence of his time at ANL on MATLAB. **Matsouka** added there has been significant international impact from the programs, the education, and the SW assets; he agreed to provide a list of SW that is being used in Japan.

Hey recommended including the Supercomputing conference which is a major conference for the whole IT industry. DOE was instrumental in setting up the first conference. Hey suggested contacting George Michael at LLNL.

UPDATE ON SUMMIT

Buddy Bland, ORNL, provided a status report on Summit. It has a broad architecture that can do traditional HPC modeling and simulation, HP data analytics, and AI; powerful CPUs for scalar operations, accelerators that do HP arithmetic at 64, 32, and 16 bit precision, a high speed interconnect with switch-based collective operations, and a HP file system.

The non-recurring engineering included design, packaging, and cooling. Significant time has been spent on training and education and will continue on Summit.

GPUs are the heart of performance and power efficiency. Summit has 27K+ NVIDIA Tesla V100s, over 5,000 CUDA cores, and 640 Tensor cores on NVIDIA Volta. Tensor cores were not part of the original plan but they gave Summit much greater performance especially in ML, data analytics, and AI. Summit is not yet accepted but is doing great science. With 1/4 of the nodes as Titan, Summit has much higher node performance, more memory per node and higher memory bandwidth, faster interconnect, higher bandwidth between CPUs and GPUs, a larger and faster file system, and 7x the performance.

The HW is stable and performing well, all nodes are functional, Infiniband is performing as expected, and 30K discs are running. The copper SAS cables are being replaced with fiber. Summit is working through the acceptance tests now and will go into production January 19, 2019 with Innovative and Novel Computational Impact on Theory and Experiment (INCITE); ASCR Leadership Computing Challenge (ALCC) allocations will begin in July 2019.

Discussion

Reed asked if the non-volatile memory (NVM) link only comes off of one of the Power 9s. **Bland** said that was correct but it is not a fast link.

Matsouka noted that the real value of Summit is the significant increase in the memory bandwidth. For Summit to have won a signature achievement in this respect is a stellar triumph that should be stressed more. **Bland** agreed; adding bandwidth is very important and expensive.

Gregurick indicated there will be a much bigger impact in biomedical sciences and encouraged continued outreach and training activities. **Bland** said the thought is that more users will come to the table from the AI and data analytics worlds. **Gregurick** asked if people will be taking advantage of data computing that feeds into the model-driven HP traditional simulations. **Bland** thought so and said it was one reason for the investment in the large file system. The NVM on every node will serve as a high speed cache to the file system, but it can also be used in other ways.

Lethin inquired about the cables being replaced with optical fiber. **Bland** clarified those are the disc cables. The SAS cables that link the disc to the controllers and the controllers to the system were already fiber. The network has always been fiber.

Bergman asked about the programmability between Titan and Summit. **Bland** explained Titan has the first generation of NVIDIA Tesla series of GPUs designed for computing. Volta is at least two generations newer, much more efficient, and more programmable. Summit is able to provide more efficiency and parallelism than Titan. The NV link makes the programming much easier because it is fully coherent and directly addressable; the GPU can directly address the DRAM memory, and the CPU can directly address the high bandwidth memory.

Svore asked about hand tuning, investment in the SW stack, and failure modeling. **Bland** said more investment is required when moving a code that has never been put on a GPU. Codes that have been running on Titan typically move to Summit with relative ease and get good performance. Tensor cores are just now being explored. Titan did not have a good OpenMP implementation to use the GPUs effectively; IBMs XL compiler has an exceptionally good Open MP implementation. OLCF will continue to support OpenACC, but also wants to support OpenMP. Various failure recovery proposals exists for MPI but none have been accepted by the standards committee. OLCF tells Titan users to ask for extra nodes.

Chen asked if current computing models are adequate to helping users maximize the compute resources, and about the balance of future users with mixed workflows versus physics-based workflows. **Bland** responded that physics-based modeling and simulation will be the core of what SC and NNSA users do on the machines, but discovering new ways to use the ML and data analytics will improve the physics-based modeling and simulation. Kokkos and RAJA are being developed for using the processors and making it easier to move codes from one type of architecture to another.

BASIC RESEARCH NEEDS (BRN) FOR MICROELECTRONICS WORKSHOP

Andrew Schwartz, ASCR, provided an update on the upcoming BRN on Microelectronics scheduled for October 23-25, 2018. At the May meeting topical areas were identified where fundamental research over 5-10 years could have a significant impact on microelectronic technologies. The output of the meeting was common themes in materials science, device physics, emerging architecture approaches, cross-cutting areas, and a list of questions to be addressed. Schwartz mentioned factors that created the call to action including complementary metal–oxide–semiconductor (CMOS) challenges, Moore's Law, U.S. leadership position, *whole of government* approach, and DOE and SC's role. The BRN will have four break-out panels on four areas, each with two co-leads.

Discussion

Bruce Hendrickson mentioned the National Science Foundation's Big Ideas activities that were focused on power and the need for low power electronics. **Schwartz** stated he was aware of the Big Ideas activities. Thinking about power and energy efficiency and power needs for future electronics is critical.

Dunning asked if there would be discussion of materials beyond CMOS. **Schwartz** indicated that the 2nd session is focused on the devices-to-architectures; the 1st session will be materials.

Hey asked if there is a roadmap for more future plans. **Schwartz** indicated there is not a specific roadmap. A number of facilities are in the process of defining their computing and electronics needs for the future. **Hey** thought it was not the computing needs but the big data collection. **Schwartz** said the challenge is to think about the electronics and human needs required to address what comes after the detectors and sensors have generated the data.

Gregurick asked where a workshop on in-situ processing in sensors and detectors might happen. **Schwartz** envisioned a follow-on workshop to look at detection capabilities, but that is not in the near term planning. **Reed** added that part of the conversation was just about making people aware of the scaling of the data rates to inform thinking on real-time analysis and requirements. **Hey** noted the challenge in Europe is the Square Kilometre Array (SKA) as they have extremely high data rates and computing demands.

PUBLIC COMMENT

None.

Reed adjourned ASCAC at 11:12 a.m. (ET)

Respectfully submitted,
T. Reneau Conner, PhD, MIS, PMP, AHIP
ORISE
October 5, 2018