HPC TRENDS FOR FEDERAL GOVERNMENT
Engility recently attended SC16, the world’s premier High Performance Computing (HPC) conference, which attracted 349 exhibitors and over 11,000 attendees. With many federal organizations constrained in attendance by budget, Engility attended the show to not only enhance our own capabilities, but also promote the work of our clients and collect data to help identify the best possible HPC solutions.

In a study funded by the Department of Energy, IDC Research found that government HPC investments resulted in $141 in profit or cost savings for every dollar spent!

Engility sent more than 20 computational scientists and HPC experts to attend technical sessions, workshops, and tutorials; visit vendors; participate in technology roadmap briefings; and meet with government decision makers to gather—and pass on—valuable insights into the state of the HPC industry. Several teammates presented posters and workshop papers during the conference, and Engility was privileged to sponsor the SC16 Poster Session, facilitating a deeper technical dialog on a broad range of industry, government, and academic HPC research.

Following, please find a few observations from SC16. We invite you to contact our team if you have other topics of interest or would like more details around any of these trends. Gay Porter, who leads our Technology Solutions Group, can field your questions and connect you with our subject matter experts.

Not familiar with Engility’s role in HPC? Visit www.engilitycorp.com/services/hpc/ for details on our HPC experience and capabilities. You will also find all the papers and posters referenced in this document.

STATE OF THE INDUSTRY

Applied HPC: Exascale has transitioned from buzzword to long-term strategy while cognitive computing has moved to the forefront with many sessions focusing on how HPC is being applied in new and interesting ways. High Performance Data Analytics, health sciences, and cyber were also frequent topics of interest—appearing in multiple conference components and many vendor presentations.

Katherine Frase, who leads the strategy and business development for IBM’s Watson Education unit, delivered the keynote address. Her presentation, “Cognitive Computing: How Can We Accelerate Human Decision Making, Creativity, and Innovation Using Techniques from Watson and Beyond?” explored how we deal with the vast amounts of unstructured data we generate as a society. Data without the ability to derive actionable insight is a major challenge.

Her talk and the plenary session, “HPC Impacts on Precision Medicine: Life’s Future–The Next Frontier in Healthcare” explored how HPC is helping and will continue to help humans do their jobs better. For instance, an oncologist anywhere in the world might access a system that could reference all the specialty-cancer research available and enable the local doctor to ask questions that might result in a better diagnosis and treatment plan for the patient’s unique form of the disease.

Frase stated, “Systems need to fit into the workflow of the real person doing the real work.” Fitting HPC into the real work of users was a constant and vital theme encountered at SC16 and one encountered on the federal programs Engility supports on a daily basis.
TRENDS

Big Data, High Performance Data Analytics and Life Sciences

Issues surrounding big data and the application of HPC were a prevalent and pervasive theme. Show organizers created SCinet on the show floor—a temporary network showcasing 100 Gbit/s technology and demonstrating over 1.2 Terabytes of network traffic on the show floor, serving all 11,000 attendees.

Big data led to conversations of High Performance Data Analytics (HPDA) and life sciences. HPC and the analysis of large data sets are becoming inextricably entwined. Show organizers and industry analysts discussed HPC and HPDA in the same breath—pointing out the HPC-driven computation needed for HPDA with new applications to fields like fraud prevention and detection, business intelligence, and personalized medicine.

Nearly 20% of workloads on government HPC systems are dedicated to big data analytics.

The conference’s HPC Matters panel, which highlighted the Cancer Moonshot and new funding initiatives in the fight against cancer, illustrated the prominence of life sciences as an HPC focus area at SC16. As life sciences have become more data driven, the opportunity to solve important problems in biology has become evident—for instance the analysis of treatment and the study of genomes. Patient data paired with data analytics and HPC could lead to major medical advances.

Engility scientists Ross Smith, Rhonda Vickery, and Jack Harris presented a conference poster on HPC-enabled data analytics for high-throughput, high-content cellular analysis—showing how to transition biologists into power HPC users.

Cyber

Data analytics and how it affects cyber security cropped up in several conversations. HPC has a role to play in this emerging field, but the technology, and how organizations use it, continues to mature.

Engility developed and applied a comprehensive vulnerability assessment and cybersecurity framework for IV&V, Information Assurance, cyber ranges and training for NASA spacecraft missions, USAF weapon systems, Defense cybersecurity assurance, and Intelligence Community programs.

Neural Networks

This year seemed to be the year of the neural network at SC16. Heavily pushed by NVIDIA and many of their partners, deep learning, convolutional neural networks, and other algorithms that are based on the premise of a neural network were on wide display on the show floor. The concept even factored heavily in the Plenary Talk on data-driven/personalized medicine. However, presented work did not reflect this emphasis; only three of the technical papers included the words “neural” or “convolutional.” The gap suggests adjudicated research is still catching up with the anticipated roadmaps of vendors.

Use of neural networks, including its use in the realm of machine learning and data analytics, is an area of interest to several of Engility’s federal clients. Neural networks are not a magic potion though, and codes and architectures designed for direct physics calculations will likely remain the bread and butter of HPC solutions for quite some time.
Software discussions at SC16 were very much driven by hardware concerns. Just looking at the exascale systems under development at DOE, the entire community recognizes that a sizeable portion of the time needed to reach exascale won’t be the hardware—it will be creating application codes that can run at scale on the new systems. Uncertainty in next-generation HPC hardware architectures is causing further uncertainty in software approaches.

New software approaches (e.g., MetaMorph) are emerging for “performance portability,” which means obtaining reasonably good performance (e.g., MetaMorph has the goal of achieving 90% of native code performance) on all emerging hardware platforms while only having to code the algorithms once. Examples of portable and native software solutions being considered to meet uncertainty in hardware include Kokkos, RAJA, C++118, MPI, OpenMP, OpenACC, CUDA, and many more. Compatibility with these solutions, I/O libraries, and in situ libraries may be a challenge.

OpenStack is emerging as a helpful virtualization platform for HPC cloud computing. One panel speaker noted that Linpack may get 99% of bare metal performance on OpenStack; however, real applications such as image deconvolution slow down by a factor of 4-5x. Cloud computing for HPC is evolving and worth following.

Engility’s Ross Smith explored code optimization at the PyHPC workshop at SC16.

The high performance Linpack (HPL) measurement is the test for the Top 500 list and is commonly used to establish the fastest computers in the world. For many years, however, HPC practitioners have encouraged the use of a broader range of benchmarks to truly measure performance for systems. This is not new, and one of the full-day workshops presented at SC16 was the 7th International Workshop on Performance Modeling, Benchmarking and Simulation of HPC Systems that covered alternatives and replacements to Linpack. The most notable alternative to HPL is HPCG (high-performance conjugate gradient). Conjugate gradient is one of the most prevalent math algorithms used in most applications that solve large linear algebra problems.

Systems designers are also dealing with the inability to increase processor frequency—with each putting forth a different approach. There are attempts at trying to abstract underlying hardware disarray and divergence, but users can’t escape the fact that they may need to change the mathematical representation of a physics problem in order to match the underlying hardware better. Many application software codes will need to be rewritten to get the performance out of future machines. This may involve serious rethinking and reimplementation of algorithms—an expensive prospect. This state indicates the best path forward is an integrated approach for software development. For example, a computational fluid dynamics researcher does not want to worry about whether a certain kind of memory is configured as a cache or how much power is applied to some portion of the code. Engility’s HPC experts navigate these issues every day, helping federal partners to find the best software solutions.

On one Engility-managed federal program, researchers optimized computational fluid dynamics applications for a modern multi-core and accelerator HPC system.

Under another DoD program, Engility and its partners explored parallel and scalable algorithms to reduce the time complexity of extreme-scale simulations.
Visualization and Analysis

In situ libraries are not fulfilling user needs, given the number of custom in situ implementations published at SC16. The SENSEI generic in situ interface may help solve some of this challenge. The primary difficulty with in situ and visualization in general is the ability of the average scientist to use it.

Engility’s Sean Ziegeler was invited to speak in the conference’s In Situ Analysis and Visualization workshop where he presented a paper and served as a panelist.

Another topic around visualization was the need for systems that make use of emerging HPC technologies. For instance, how do we ensure the next generation of weather models accounts for vertical resolution? The variable is negligible at today’s current 13-km grid sizes but will be a large consideration as we reach resolutions of less than 2 km.

HPC applied to weather modeling was explored in depth by Engility’s Shannon Rees in a conference poster dealing with NOAA’s Next Generation Global Prediction System.

Parallel I/O and Non-Volatile Memory

Burst buffers were a hot topic in discussing I/O performance. Burst buffers sit in between memory and hard drive storage—faster than hard drives and slower than memory. They are similar to non-volatile memory like that found in your phone. Burst buffers are helpful, but middleware infrastructure is needed to use them effectively. Real applications are seeing a 3-5x speedup in I/O performance with burst buffers, but they could do more with some tuning. Experimentation will determine optimal burst buffer configurations for each type of simulation.

Engility’s I/O experts are currently working on HDF5, a parallel I/O library that stores data in a standard self-describing format. This allows for easy and fast data interchange. We are also working with the developers of HDF5 on an auto tuner that enhances the performance of HDF5 on HPC systems.

Amdahl’s Law makes post-processing an increasingly significant issue. The law predicts the theoretical speedup of a fixed task when a system’s resources are improved. Satellite-observed and archived-simulation data are growing at an exponential rate due to the much higher image resolution of new satellites and higher-resolution simulations. Burst buffers may help with the simulation I/O but do not necessarily help with reading data back in for post-process. During simulations, in situ analysis or workflows done entirely in memory (or at least burst buffer space) may be necessary. Treating the entire simulation and post hoc workflow as a task parallel graph may be the solution. POSIX (a family of standards for compatibility between operating systems) enforces an unnecessary level of coherence that significantly penalizes performance. Other options include object stores and database-like storage.

Post-Moore’s Law and Dennard Scaling

Moore’s Law and Dennard Scaling were again a major theme. Moore’s Law predicted the number of transistors on a circuit would double every 18 months, and Dennard Scaling states power use stays in proportion with area. These held true for nearly 40 years, but as of 2010, chips are nearing saturation. Among attendees, there was an incredibly varied view of what Moore’s Law meant, whether we were already past Moore’s Law, and the ability for Moore’s Law to extend 5, 10, or even 20 years into the future. One thing that did seem to have consensus was the eventual limit to shrinking silicon feature sizes. One workshop was completely dedicated to the successor-technology conversation on future computing-chip architectures.
Several speakers talked about different implications of quantum computing and quantum materials. This complex topic is growing in two major areas: quantum computers (which are receiving major investments from several big and small players) and the software to exploit the computers. A frequent observation is that if you think you understand quantum physics, you likely are wrong. The competition is to produce a functional quantum computer, and there are several candidate technologies, each with its own strengths and weaknesses. The end goal is to produce a machine that exploits quantum characteristics (a bit can have a value of 0 and 1 at the same time) to allow simultaneous computations (e.g., 5 qubits can perform 32 computations at once; traditional computers have to perform them successively). The uses for this kind of computing will include intractable problems like efficient derivative trading, quantum chemistry and chromodynamics, and condensed matter physics, as well as cryptography (especially code breaking). Other challenges include workforce development, industrial base production, and stable funding. Researchers project different dates for a commercially viable functional computer, from 10 to 20 years.

Efficiency

Flop/s are floating-point operations per second. Flop/s measure computer performance.

SCI16 saw a continued push to not just increase the computational power of computers, but also a push toward increasing efficiency, as measured by flop/s per watt. There were two separate workshops on energy efficiency. Additionally, the efficiency of the inter-node communication infrastructure, and not just that of the computational engine (CPU), became a significant talking point. There are also continued trends toward increased power management in CPUs and accelerators—some parts of silicon may turn off when not in use or may use lower frequencies to reduce heat dissipation requirements. Users may increasingly encounter the need to run jobs with constraints on power consumption, which will impact performance.

Containers

The software containers concept received a lot of attention and buzz at SCI16. Particularly useful for executing modern data analytics tools in traditional HPC environments, containers are similar to virtual machines, but they do not run a separate operating system. Containers are extremely useful in making code more portable because an executable and the libraries it depends on are distributed, ensuring smooth execution across multiple hardware architectures. Modern tools are being updated extremely rapidly and make use of cutting-edge software. Containers mitigate the challenge of integrating these tools into environments with much slower update cycles (e.g., HPC systems).

Containers also greatly minimize I/O load on meta-data servers, and according to Cray’s CTO, they play a critical part in providing exascale capability. Containers have made their way from cloud computing to HPC environments and are playing an important role in bringing cutting-edge data-analytics tools to traditional HPC environments. Widely known containers include:

<table>
<thead>
<tr>
<th>Docker</th>
<th>Shifter</th>
<th>Singularity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Docker is most widely known, but it requires escalated privileges.</td>
<td>Shifter appears more secure and is deployed on Cray systems at DOE, but it does require setuid (root) access.</td>
<td>Singularity, which can be installed in a home directory without root privileges, may allow users to deploy containers to HPC systems without any security implications.</td>
</tr>
</tbody>
</table>
Accelerators

Traditional industry powerhouses such as Intel and NVIDIA continue to push toward the use of accelerators to improve performance.

NVIDIA released their new Pascal architecture earlier this year and proclaimed how it will power the next generation of applications with improved floating-point performance. NVIDIA has also proclaimed that they have adjusted their hardware releases to emphasize the increased interest in deep learning, and the Pascal architecture reflects this with 16 Tflop/s of half-precision performance.

The Naval Research Laboratory booth at SC16 displayed an all-GPU unit that reached 142 Tflop/s in a 7U form factor with 5KW power consumption.

Intel has presented their accelerator solution with their latest Xeon Phi release called Knights Landing (KNL). It is intended to address many of the previous shortcomings of the Knights Corner architecture with faster cores, faster internal communication, use of the Intel Omni-Path fabric, and a fast memory called multi-channel DRAM (MCDRAM). The previous Xeon Phi (Knights Corner) promoted the fact that codes could run on the platform with a simple recompile, and Knights Landing furthers that commitment by being binary compatible with traditional Xeon processors. Intel has also made waves with the announcement of a future processor that incorporates traditional cores with a built-on FPGA alongside the cores.

While there was little in the technical program or workshops that dealt directly with the new Intel Knights Landing processor, the new processor was definitely a considerable topic of discussion given its inclusion in many of the top-10 fastest supercomputers in the world. The prospect of a many-core architecture that can be relatively easily programmed (significantly easier than Knight’s Corner or CUDA) has the attention of many HPC users. Of significant note, codes compiled to run on a standard Xeon processor should run on a KNL without any modifications. KNL is self-hosting, but modifications for performance are still needed.

Time will tell how software developers approach the 16 GB of on-chip MCDRAM. Will they attempt to make use of it directly or ignore it? Ignoring it can be a viable approach, given that the MCDRAM can be configured as an L3 cache and still provide a significant benefit when compared to a traditional L3 cache. It is anticipated that use of fast memory such as MCDRAM and burst buffer-like architectures will influence HPC codes in enhancing data movement within codes for highest code efficiency and scalability.
WE HELP REAL PEOPLE USE HPC TO DO REAL WORK.

SC16 TAKEAWAYS AND 2017

We know that everyone can’t attend SC16 and we hope this summary has provided a few helpful glimpses into the state of HPC and imminent trends presented at the conference. A common theme throughout is the need for strong partnership among government, academia, and industry to address the HPC challenges presented by your critical missions.

• Energy efficiency is becoming more important as computer sizes, sophistication, and data growth increase.
• The combination of advancing machine learning and HPC capacity offer new analytical areas for health, cyber security, and many more federal government missions.
• The uncertainty of future HPC architectures necessitates trusted partnerships with industry and academia to develop the tools and solutions needed to gain actionable insight from data sources.

Although the HPC environment involves technical tradeoffs, budget constraints, legacy issues, and dozens of other obstacles, HPC users understand that the use of HPC resources can provide them with the opportunity to overcome the many challenges they face. Hardware and software vendors strive to craft solutions, but at the end of the day, it is experienced HPC scientists and technologists who understand the underlying issues and mission needs to deliver success.

Engility experts provide scientific, computational, operational, and strategic mission knowledge that is unencumbered by the need to promote a particular hardware or software solution. Echoing the themes of SC16’s keynote address, we help real people use HPC to do real work. Please reach out to Engility for more information on any of these topics or to discuss your HPC challenges.

For more information, please contact:

Gay Porter | VP, Technical Solutions Group | 703.405.6411 | gay.porter@engilitycorp.com

This presentation consists of Engility general capabilities and DOES NOT contain controlled technical data as defined by the International Traffic in Arms (ITAR) Part 120.10 or Export Administration Regulations (EAR) Part 734.7-11. Per EGL-12152016