HPC TRENDS
FOR FEDERAL GOVERNMENT

DECEMBER 2019

SAIC®
Redefining Ingenuity

HPC TRENDS

40

ATTENDEES

6+

GOVERNMENT ORGANIZATIONS SUPPORTED

4

DEMOS

QUANTUM
CONTAINERS
EXTREME SCALING
DATA ANALYTICS

20

YEARS EXHIBITING

Not familiar with SAIC's role in HPC?
Check out www.saic.com/hpc

As a service to our clients, SAIC offers a few notes from the Supercomputing Conference 2019 (SC19).

Each year, the Supercomputing Conference hosts thousands of the world's leading experts in high performance computing (HPC). The week-long conference offers an in-depth and diverse technical program that includes training, tutorials, and workshops not offered in other places. Researchers and industry experts packed the exhibit area, presenting their latest technology and showing a glimpse of things to come.

This year, SAIC deployed 40 of our HPC professionals in support of programs for the U.S. Department of Defense (DOD), U.S. Department of Energy (DOE), National Oceanic and Atmospheric Administration (NOAA), National Aeronautics and Space Administration (NASA), the Food and Drug Administration (FDA), and other U.S. government agencies. SAIC's computational scientists and HPC experts attended technical sessions, workshops, and tutorials; visited vendors; participated in technology roadmap briefings; and met with government decision-makers to gather — and pass on — valuable insights into the state of the HPC industry.

This document provides observations and insights SAIC gained into current HPC trends at SC19. With an eye toward the application of HPC, HPC experts and non-computational experts like systems engineers, medical professionals, and domain scientists informed our insights. Our discussion remains at a fairly high level, but we invite you to contact our team if you have other topics of interest or would like more details around any of these trends at HPC@saic.com.

State of the Industry

On the show floor and in vendor discussions, several themes emerged: high performance data analytics (HPDA), and artificial intelligence, machine learning, and deep learning (AI/ML/DL) are currently hot topics. Exascale (along with new architectures), quantum, and HPC in the cloud also headlined a lot of talks. Our team tracked the status of these technologies and how our customers might realistically apply them to today's challenges and prepare for tomorrow's hurdles. Figuring out how to get HPC resources to perform as desired includes challenges like application scalability, performance engineering, and industry experts. For anyone new to these terminologies, we offer a quick primer before our more detailed explanation.

HIGH PERFORMANCE DATA ANALYTICS (HPDA):
A lot of organizations and users are finding that they now have access to massive sets of data a traditional desktop computer cannot analyze, like a cancer database. Data is only as good as the insights we can extract from it, and HPDA provides a path to process that volume of data with the massive parallelism offered by HPC to gain insights more rapidly.

AI/ML/DL:
Some attendees spoke of AI, machine learning (ML), and deep learning (DL) in the same breath. We'll use AI as the umbrella term for programs where computers apply logic. Machine learning is AI that can modify itself when presented with new information, or "learn." Deep learning goes one step further from ML, layering artificial neural networks to "think." The exploitation of HPC to enable AI is myriad and exciting. Training AI is very computationally intensive.

NEURAL NETWORKS:
These are computer systems designed to represent and process information like the network of neurons in the human brain. These are the core of deep learning algorithms.

10^{18}

EXASCALE:
Exascale means the ability to perform a billion billion calculations per second! That's 1 followed by 18 zeroes or a thousand-fold increase over the 2008 achievement of petascale calculations.

QUANTUM COMPUTING:
While traditional computing relies on binary or two-state logic, or bits, quantum computing takes advantage of a structure's natural tendency to exist in simultaneous multiple states, or superposition of states that are neither one nor zero. This approach, though in its infancy, has led to the construction of several quantum computers in research production use, with sizes ranging from 1 to 2048 "Qbits," or quantum bits. The size of these quantum computers is increasing with time and promises radical possibilities in future computing.

HYPERION RESEARCH REPORTS THE SIZE OF THE HPC SERVER MARKET GREW FROM $12.2 BILLION IN 2017 TO $13.7 BILLION IN 2018, FORECASTED TO INCREASE TO NEARLY

$20 BILLION BY 2023

Want More?
Throughout this summary, we will offer hyperlinks to more detailed resources like blogs, case studies and white papers.

Read more on exascale' s impact on design decisions.

Watch Dr. Rajiv Bendale's explanation of code optimization on emerging HPC architectures.

Read how code optimization for genomics and bioinformatics applications is helping the FDA.

To Exascale and Beyond

Exascale machines are on the near horizon. China and the U.S. expect to deploy their first instances in 2022. By the middle of the next decade, there will probably be about 10 exascale machines, with each costing hundreds of millions of dollars. We might see 10 times the computing power of exascale by the end of the next decade, though those units will run in the billions of dollars. With that kind of cost, the traditional replace approach is going to shift to systems that can be repaired and maintained.

Operating these new systems comes with challenges in system infrastructure, system environment, code, and personnel.

MOORE'S LAW AND THE IMPORTANCE OF CODE

It looks like Moore's Law is grinding to a halt by the mid-2020s. Advancements will still come in the short term, down to 5-7 nanometers, but longer-term advancements will come from the improvements in architecture, such as increased vectorization, greater use of GPUs, and specialized libraries for Field-programmable gate arrays (FPGAs), with the corresponding requirements to code for those architectures. Large cluster users struggle to get legacy codes to support larger environments effectively. Monte Carlo techniques and Computational Fluid Dynamics (CFD) codes are probably the first focus of the National Labs.

Sophisticated coding techniques to take advantage of improvements in architecture are slower but steady and will increasingly dominate as the method to provide more performance. The pendulum is swinging toward major efforts on the software side and away from shrinking processors, which previously provided performance improvements of older code.

HPC Talent

Demand will significantly increase over the next decade for experienced programmers to port code to take advantage of these improved architectures. It will no longer be possible to just run the same code on a faster processor. Instead, it will be necessary to modify the various simulation codes to work effectively, whether that means by vectorization, hooking it to faster custom libraries, optimizing to run on SIMD GPUs, or rewriting it completely in new languages to take advantage of optimized architectures. In-situ programming and other data transference reduction techniques will also increase in use. All of these techniques will require more sophisticated programmers in computer science and domain areas, which means encouraging and incentivizing them to acquire the knowledge and experience to implement these changes.

Academia is not currently churning out the volume of HPC experts needed to directly fill all the needs of government and industry. Instead, industry and government will need to assist in training and incentivizing the next generation of HPC experts. At SAIC, we invest in internships and tailoring young talent to unique customer demands. Internships expose students to the type of work done by HPC experts. Other approaches include scholarships and better training incentives to encourage more students to enter the field and stay current in it. We also need to allow such students remote access to our HPC systems. Universities don't always have the HPC systems on campus that students need to interact with to gain experience in using and designing them.

Read about the SAIC HPC internship experience from one of our 2018 interns

A RECENT STUDY FOUND THAT GOVERNMENT HPC INVESTMENTS RESULTED IN

$112

IN PROFIT OR COST SAVINGS

FOR EVERY DOLLAR SPENT
WWW.HPCUSERFORUM.COM/ROI/

Watch Dr. Chris Powell explain how SAIC evaluates new technology.

New Architectures

As mentioned, GPUs are a significant part of the technology roadmap for the HPC industry. At SC18, there was strong general-purpose GPU (GPGPU) architecture and programming support from NVIDIA, but limited or sparse support from AMD or INTEL. At the International Supercomputing Conference (ISC) in June 2019, NVIDIA announced CUDA support for emerging ARM CPUs. At SC19, NVIDIA announced a reference design platform for building GPU-accelerated ARM HPC systems. In other words, NVIDIA is releasing a beta version of its ARM-compatible software development kit, as well as working on accelerating HPC applications such as LAMMPS and NAMD. NVIDIA's announcement is a sign of progress from last year when there was no support for NVIDIA GPUs within the ARM ecosystem.

Many hardware vendors for the HPC market are starting to push GPGPU as a performance index in their new architectures. This move will heavily benefit machine learning workloads, but will create new challenges to researchers who rely on code with a lot of branch — or warp — divergences. In the past, input/output (IO) constraints could throttle GPU performance when parallelized across multiple GPU devices on a single node.

However, newer architectures — such as Cray's AMD powered Frontier system — are designed to incorporate unified connectivity between the CPU, memory, and GPUs. These architectural changes are paving the way to a transition from CPU-heavy HPC architectures to a stronger focus on GPGPU configuration. Neural networks frameworks, such as the Tensor-Flow library, drive large growth in the GPGPU market. This is also being accommodated in the reduced precision needed for ML incorporated into the newer CPU architectures to enable their use for AI/ML/DL.

FPGAs are becoming more popular and pairing them with high-speed networks and high bandwidth memory may eventually lead to something interesting.

Quantum seems to be the next big architecture with a few different approaches. While it's critical to some specialized challenges like cryptography, it's not widely supportable or applicable today. IBM plans to use QC for AI for better model training, pattern matching, fraud detection, Monte Carlo Computations for portfolio optimization, and risk analysis. More work is needed for broader application of Quantum Computing to classical algorithms, using both gate based (IBM Q) as well as Quantum Annealing based quantum computing (DWave) solutions. A big influx of cooling solutions at this year's event, like fluorinert, was closely tied to these challenges.

Dealing with Data

There is an increased need for persistent services, such as databases, as well as statistical analysis and visualization in High Performance Data Analytics (HPDA), with in-situ analysis becoming increasingly important as data sizes grow and data transfers take progressively longer. Data transfer speeds and networking speeds are increasing but at rates that fall short of fully supporting processing performance increases. Processing performance increases with Exascale in the next decade will still be limited by I/O bottlenecks resulting from potentially smaller data transfer and networking speed increases, creating an effective slow down. Analyzing data where it is produced and stored will continue to become more critical on newer and larger systems, and is sure to be critical on exascale and above systems.

Performing in-situ analysis on a node as data is produced will be a critical component, allowing for effective data reduction. Storing the produced data for further analysis and use in training neural net systems, with a history of the operations performed on the data, will allow performance analysis on the stored data to improve when improved analytical techniques become available. This will allow one to make judgments on the suitability of stored data for particular tasks. On-demand access to this data for use by web services or visualization will improve the usefulness of produced data and enhance its exploration. Many of the large vendors are releasing tools to scale data analytics computations at a higher performance.

Everyone is serious about HPDA and exploring how to combine it with HPC in the cloud. Many new companies are coming out with a front-end API to use the cloud more effectively and economically. The cloud companies are developing new tools to keep themselves relevant and competitive. Cloud offerings will continue to advance, aided by the ongoing transition to container technologies like Docker and Singularity. Google, Microsoft, and Amazon are all leading the way here due to the possibility of no queue time and flexibility.

Containers are increasing in relevance as a method of packaging and deploying software effectively on HPC systems and in the cloud. Cloud services are offering new offerings such as spanning Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS). Services are offered in: compute, networking, storage, web + mobile, databases, data + analytics, AI & cognitive services, Internet of Things (IoT), enterprise integration, security + identity, monitoring + management, and developer tools. Along with this, cloud computing providers, such as Amazon's AWS, are heavily marketing deep learning containers (AWS DL Containers). The cloud services market is capitalizing on simple-to-use containers with pre-built DL frameworks, like PyTorch and TensorFlow, for quick customer deployment of machine learning code.

Watch Dr. Wes Brewer share findings on combining HPC with data analytics challenges.

HPC TRENDS

Check out blogs from SAIC data scientists.

AI/ML/DL

AI and ML seem to be part of almost every discussion, from how to analyze data to how to automate cluster support. It was really interesting to see how seemingly everyone tried to develop ways to use it. Hardware manufacturers continue to make design pushes to suit AI/ML applications, which have a place in climate modeling, though they are not yet a part of our climate models. Neural networks were widely discussed at SC19, with the biggest national labs dedicating a huge portion of their next-gen systems for ML and DL. Many of the HPC vendors demonstrated new AI/ML specific chip architectures integrated with their systems, including architecture from Graphcore IPUs now available in the Cirrascale cloud services. The training requirements of neural networks are asymmetric with the deployment of the trained nets. Training processing power continues to grow and by orders of magnitude larger than the required processing power for the deployment of trained neural nets. We will soon take advantage of exascale-level systems for training our largest and most complex neural nets. However, with larger systems becoming increasingly more expensive and limited in number, it will be necessary to spend more effort on improving neural nets' training efficiency. Otherwise, we will reach the limits on the scale and complexity of the neural nets that we can train and use.

SC19 Takeaways

Ever-increasing HPC usage is a major economic force across the world, with rapid adoption as a game-changing decision aid and technology enabler. HPC now permeates industrial design, manufacturing, defense, healthcare, and almost all fields, and its rapid implementation is an economic driving force among first-world nations. To move forward, HPC needs not only computer science and domain science expertise, but practical technologists — system and network administrators, and everything in between — to ensure the full realization of the potential of HPC. For example, our HPC practitioners at SAIC create better weather models to enhance overall weather prediction in advance of severe weather situations. They are improving the materials in warfighter systems for increased safety. And they create entirely new ways of looking at information and extracting usable, and verifiable, information for mission-critical decisions.

After attending SC19, our experts return to their customer programs with advanced skills garnered from the technical program and new insights from the technology roadmap briefings to better assist government organizations achieve their mission-critical goals, goals that are multi-disciplinary and enhanced by HPC. As evidenced by the SC19 keynote speaker, Dr. Steven Squyres, the principal scientist for the Mars Exploration Rover Project, HPC does not exist in isolation. It interconnects with domains like space and sciences like engineering and helps us achieve greatness.

SAIC experts provide scientific, computational, operational and strategic mission knowledge and know-how that is unencumbered by the need to promote a particular hardware or software solution.

WE HELP REAL PEOPLE USE HPC TO DO REAL WORK.

Please reach out to SAIC for more information on any of these topics or to discuss your HPC challenges.

20-0961

SAIC® Redefining Ingenuity