The age of exascale and the future of supercomputing

November 16 2021, by John Spizzirri

AVIDAC boasted remarkable computing power for the time, performing 1,000 multiplications per second. Today's smart phones can store around 100 million times more data and do in a single second what would have taken AVIDAC two months. Pictured with AVIDAC is pioneer Argonne computer scientist Jean F. Hall. Credit: Argonne National Laboratory

Argonne looks to exascale and beyond, sorting out the relationship between computing and experimental facilities, the need for speed and

AI's role in making it all work.

In 1949, physicists at the U.S. Department of Energy's (DOE) newly minted Argonne National Laboratory ordered the construction of the Argonne Version of the Institute's Digital Automatic Computer, or AVIDAC. A modified version of the first electronic computer built at the Institute for Advanced Study in Princeton, New Jersey, it was intended to help solve complex problems in the design of nuclear reactors.

With a floor area of 500 square feet and power consumption of 20 kilowatts, AVIDAC boasted remarkable computing power for the time. It possessed a memory of 1,024 words (about 5.1 kilobytes in total), could perform 1,000 multiplications per second, and had a programming capability that allowed it to solve problems consistently and accurately.

Today, your smart phone can store around 100 million times more data, and can do in a single second what would have taken AVIDAC two months.

Since AVIDAC, Argonne has housed an impressive inventory of ever more powerful machines and, for most of its 75-year history, taken a leadership role in advancing supercomputing efforts and strengthening the foundations for pivotal discoveries in cosmology, climate, energy research and much, much more.

Through decades of research, innovation and collaboration, Argonne has helped shape machines of discovery, each more complex than its predecessor. And with a unique perspective as both host and user of supercomputers, the lab continues to steer larger conversations around the future of supercomputing.

Katherine Riley has witnessed the advent and passing of several

generations of Argonne supercomputers and will have input into future designs—a future that seems to arrive faster all the time.

When she joined Argonne in the early 2000s, computer chip designer Intel had achieved, only a few years earlier, a teraflop of computational power—a trillion operations per second. At that time, the lab was discussing the potential for petaflop machines that would eventually perform quadrillions of calculations per second.

"Even when we were looking at petascale systems, there were conversations about what exascale (a billion billion calculations per second) would look like," recalled Riley, now director of science for the Argonne Leadership Computing Facility (ALCF), a DOE Office of Science user facility. "I remember somebody projecting how hot the chips would be in an exascale system, and they thought it would be as hot as the surface of the sun.

"Obviously, we couldn't see the path yet, but we've learned a lot and so much has advanced since then."

Argonne has recently retired its third generation petascale machine, Mira, and is on the cusp of receiving its first exascale machine. The new supercomputer, Aurora, is set to open doors on more complex questions, provide higher resolution simulations and deliver faster, more accurate data analyses.

Fast as it is, Aurora will be augmented by the use of artificial intelligence (AI). And that combination will help drive autonomous, self-driving laboratories that will lead to more efficient experiments and faster solutions to and discoveries aimed at issues important to society.

It is the age of exascale, the new future of supercomputing. At least for now, as Riley and colleagues across the laboratory begin thinking about

and discussing the next chapter in the evolution of supercomputers.

Speed's the thing

"The big difference between the supercomputers of the past and today's supercomputers is that the early supercomputers could solve certain types of problems very quickly, but you couldn't solve very complex problems," said Salman Habib, director of Argonne's Computational Science division. "The supercomputers of today can do that, which makes a lot of sense, because if you're a national laboratory like Argonne, you're tasked with tackling those kinds of problems."

"Those kinds of problems" affect all of us and shape the world around us. They involve complex systems that have a lot of moving parts. They range from the biological relationships within ecosystems to the complexities of climate to the formation of mutations in a virus.

Habib's complex problem is the universe; where it's been, where it's going and what's driving it there. It is a problem that involves developing numerous models of the early universe to understand the fundamental constituents of matter and how their populations and distributions evolve. That takes an awful lot of computational power and time.

Even with Aurora on the doorstep, Habib wonders if an exaflop of calculating power will prove enough to answer his or other big questions. He is not alone. Scientists have struggled with some of these questions for decades and longer, and still not come close to fully solving them, even with giant leaps in computing.

For him, it's about getting an answer on time scales that replicate how we think or as fast as it takes Alexa or Siri to respond to your query.

"We think in time scales that are in the tens of seconds or minutes. If

you have to wait two weeks for a computer to give you an answer, your chain of thought is already gone," said Habib. "In order to aid me in thinking, the machine should answer me literally in seconds, just like when you Google something. If it took two weeks to get an answer, the whole purpose of Google would be lost, right?"

Getting to that kind of speed requires not only changes in computer architecture, but the development of techniques to accelerate the way we ask questions to better understand systems under study.

"We actually don't know how to write down, as equations, the questions we're asking about some of the more dynamical problems," said Rick Stevens, Argonne's associate laboratory director for Computing, Environment and Life Sciences. "Even if we have the world's fastest supercomputer, I actually couldn't write down the equation of cancer, for example."

Stevens is also a principal investigator on a multi-institution cancer research project. For him, part of the answer to deeper understanding of complex problems, like developing treatments for cancer, lies in the integration of AI and computer design and function. The combination could help fill gaps in information, which might then lead to faster, innovative discoveries.

One AI technique gaining significant traction in the scientific community is called machine learning, which is used to find patterns in data. In a complex area of study such as cancer research, machine learning can, for example, uncover and fill in information on the behavior of tumor cells or the relationships between tumor cells and drug molecules.

"That means we can actually understand things by combining what we know from theory and what we learn from data into a coherent system.

Then we use the supercomputer to predict how that system is going to behave," Stevens explained. "I think future applications are going to be hybrids like this."

Solving for supercomputer challenges

The story of supercomputing is based in part on exponentials, said Stevens. Of particular note is Moore's Law, which states that the size and speed of computing processors doubles every two years.

Though not achieved entirely on the strength of processors alone, the operating power of supercomputers has increased something like a trillion times over the last 40 years, exceeding Moore's Law.

"Now, name anything else in your entire understanding of the world that's become a trillion times faster in the last 40 years," Stevens said. "I'll wait for you."

But one thing that doesn't go very well with exponentials is power, he added. And as supercomputers become faster, they will require much more power, or energy, to operate them.

Aurora will require about 50 megawatts of power, 2,000 times more than was consumed by AVIDAC and more than that used by a typical small town. As speeds increase, future machines will require double Aurora's power and more. The budgetary implications and space requirements associated with that increase have the potential to slow the forward momentum of supercomputing.

But by pushing the envelope of what's possible, even by small increments, researchers can begin to minimize the problem.

As director of the Mathematics and Computer Science division, Valerie

Taylor's research focuses on performance and power analysis. Work conducted by her group could make current and future machines less power hungry and more efficient.

"Energy is the amount of power you use over time," she explained. "So, we take an application, for example, and look at the average power it consumes over the execution time of a program and ask if there are efficient ways to reduce power requirements."

Turns out, there are a number of different strategies to achieve better energy efficiency. These could include adjusting the power requirements throughout execution or reducing execution time itself, added Taylor. Modifications can be made to software codes, and machine learning methods can offer insights into how applications can be improved.

"Those are just a few ways to reduce power requirements," she said. "And sometimes we may achieve a 10 to 15 percent increase in energy efficiency, and other times it might be as much as 40 to 45 percent. So, no matter what your power budget is, there are ways to utilize it more efficiently."

Thinking outside the lab

When it comes to the next big thing, innovation doesn't necessarily have to originate from within the lab. Today, the DOE is working with large hardware developers, such as Hewlett Packard and Intel, to deliver the next generation of supercomputers. But tomorrow, innovation may come from any number of players.

"We're interested in getting the most innovative technologies into our computing resources, whatever their source," said ALCF Director Michael Papka. "The faster we can do that, the faster our users can benefit from them."

New generations of high performance computing resources are introduced every five or six years, he added. But what if, instead of replacing the entire machine each time, components could be updated as innovations and ideas become available, making the resource itself more of an experiment?

"The challenge," Papka noted, "is making these upgrades transparent for the users."

The idea is really no different than the way many of us upgrade our home computers. You run out of space, you buy a bigger hard drive; your gaming runs a little slow, you buy a new graphics card.

"The capability to push software updates overnight or at designated times is already enabled in the software space today, where you can wake up the next morning with a new capability on your phone or in your car," Papka said. "We are now having the conversation about how to perform such upgrades for hardware."

Additionally, by making the systems easier to use, a wider range of users might start conducting more diverse research.

Already, the integration of simulation, AI techniques and data analysis on one machine, for example, is opening the doors to new sciences for which the thought of using supercomputers wasn't in the cards. And this integrated approach to addressing complexities is allowing scientists to explore different ways of asking their questions.

"Look at the biological research community. They were doing smaller experiments. But now, with the explosion of genomic capabilities, they are tackling problems they never thought would be viable," suggested Riley.

Just recently, she added, researchers studying the complex SARS-CoV-2 virus, responsible for the COVID-19 pandemic, were added to the roll of supercomputer users.

A collaboration between Argonne, academic and commercial research partners, for example, pushed the computational envelope to achieve near real-time feedback between simulation and AI approaches.

By coupling two distinct types of AI-enabled hardware, data from simulations was streamed from one platform to another to simultaneously analyze interactions in the SARS-CoV-2 virus that help it elude the host's immune system.

As supercomputers get faster and data continues to increase, real-time data analysis is becoming a tool not only for current and future supercomputing, but for new machine-to-machine interactions.

Technology integration to drive real-time analysis

It wasn't that long ago that computers were mainly stand-alone entities, being fed and spitting out data and answers. But as new machines are built, developers are acknowledging the growing relationship between supercomputers and external machines, like particle colliders and telescopes, scientific tools of discovery that produce ever-greater data loads.

Of great importance to Argonne and researchers around the world is the Advanced Photon Source (APS), a DOE Office of Science user facility. The high-intensity X-ray beams produced at the APS allow researchers from academia, industry and government the opportunity to explore, in great detail, the structure and function of matter and materials.

One of the most powerful sources of high-energy X-rays in the world,

the APS is in the midst of an upgrade that will allow the facility to generate X-rays 500 times brighter than its current output. The increase will generate tremendous amounts of new data that previously could not be captured.

Argonne principal computer scientist and group leader Nicholas Schwarz is leading the effort to deliver the new computing resources the APS Upgrade will require in the coming years.

"The upgraded facility will enable researchers to ask and help solve some of the most challenging and novel scientific questions facing the world today in areas such as materials research, quantum information, energy systems and medicine," said Schwarz. "Doing so requires us to analyze, in real-time, two to three orders of magnitude more data per year."

To handle that increase, future supercomputing and AI technologies will couple and process these increased data volumes captured from experiments and generated from large-scale simulations into a feedback loop. In this scenario, scientific ideas are explored through simulations, the results of which are validated by experiments at the APS. Those results are then used to correct or suggest new simulations. The process iterates, bringing scientists closer to new insights more quickly.

Another advantage of closely coupled computing and experiment facilities is the ability to study events that occur too quickly to understand. Real-time, on-demand supercomputing will allow researchers to analyze large amounts of data as soon as it is collected to help identify rare events.

For example, a materials science experiment might produce a data set that, if analyzed quickly, could help determine where or how a fault might form and propagate in an alloy used in a building or a vehicle.

"Often, interesting phenomena and rare events occur too quickly for humans to recognize and react. They occur, and poof, they're gone before you realize it," noted Schwarz. "Advanced computing resources and scientific instruments need to be ready, on demand to respond to an event when it occurs. We can't tell nature to come back later when we're ready."

Current "future thinking" is already making this kind of rapid analysis possible. Faster machines and the integration of simulation and data analysis are paving the way for automation in experiments and other laboratory functions.

AI and automation

While we often think of automation as the realm of robots, much of the automation that lab leaders have in mind is driven by advanced AI.

Part of the case for AI is its ability to automate repetitive, time-consuming jobs currently done by scientists and to do them faster and more reliably. For example, the goal of automated or self-driving labs is to accelerate simulations, physical experiments and, inevitably, discovery.

"We'll soon have computers big enough and powerful enough to undertake both very powerful simulations and large-scale AI computations," noted Ian Foster, director of Argonne's Data Science and Learning division. "This means that we can create a new class of intelligent, continuously learning simulation tools that repeatedly explore many possible responses to a question or problem—and then learn from those simulations to choose the next best responses to consider."

Using car battery research, Foster points to the design of electrolytes as a case in point. There are literally millions of possible electrolyte

candidates to improve a battery's conductivity. To simulate, construct and test them all would take a lifetime.

But AI techniques like machine learning can be trained to look for electrolytes with very specific characteristics. By simulating these select materials, AI systems can further determine which ones get the researchers closer to their goal, and then use the knowledge gained from those simulations to help choose the next candidates to simulate.

Similarly, researchers are developing catalogs of materials and molecules, even cosmological structures, vetted by AI technologies that sift through millions of pages of scientific journals to extract specified information.

"We end up with an AI-enabled discovery process, which we expect to be far more efficient, perhaps speeding the discovery process by a factor of hundreds of thousands." said Foster. "This is what we want to do with Aurora, and it's going to be really interesting to see how it works."

Faster, bigger, farther

Salman Habib's earlier ask for faster computing seems to resonate with many users. The question remains, how do we get there? Faster might come in the guise of efficiencies or AI acceleration or newer processing units.

Aurora is equipped with both CPUs (central processing units), the more traditional processing hardware, and GPUs (graphic processing units)—which handle many operations by breaking them into thousands of smaller tasks—that will help drive AI training methods and more realistic simulations. But newer options, like quantum chips, are already on the drawing board.

The big buzz word in science today, quantum anything is often complicated to explain and more complicated to address in computing architecture.

"The target is not a singular quantum computer, but this very heterogeneous supercomputer that has little bits of everything," suggested Riley. "We're not betting on that for the next system, but like everything else, it's coming and faster than we think."

Whatever the solution, all agree that, when it comes to the future of supercomputing, business as usual won't cut it. Exploring novel concepts, approaches and collaborations will have to be the norm.

"We are on this eternal journey to build the fastest machines," said Stevens. "Ideally, we'll have access to all of these innovative technologies to make them faster, because there are always going to be more and bigger problems we need to solve. And we can only make progress if we deliver on these frameworks in AI and automation."

Asked if he's happy that the next iteration of supercomputing is right around the corner, Stevens will tell you, "I'm happy while I'm happy. But I also recognize that if we want to accomplish big things, we're going to need a lot bigger computers."

Provided by Argonne National Laboratory