PUBLIC RELEASE: 8-DEC-2016

What to do with the data?

Physicists and scientific computing experts prepare for an onslaught of petabytes

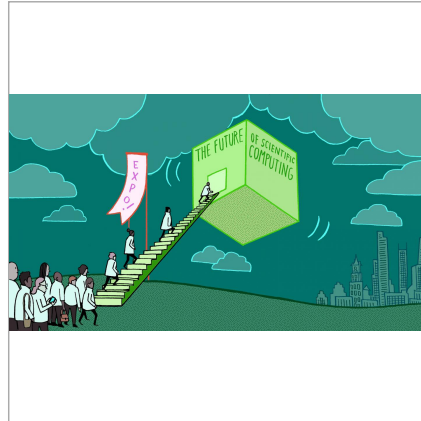DOE/US DEPARTMENT OF ENERGY

IMAGE: THE FUTURE OF SCIENTIFIC COMPUTING IS ILLUSTRATED. **view more ›**

CREDIT: ILLUSTRATION BY SANDBOX STUDIO, CHICAGO WITH CORINNE MUCHA

Rapid advances in computing constantly translate into new technologies in our everyday lives. The same is true for high-energy physics. The field has always been an early adopter of new technologies, applying them in ever more complex experiments that study fine details of nature's most fundamental processes. However, these sophisticated experiments produce floods of complex data that become increasingly challenging to handle and analyze.

Researchers estimate a decade from now, computing resources may have a hard time keeping up with the slew of data produced by state-of-the-art discovery machines. CERN's Large Hadron Collider, for example, already generates tens of petabytes (millions of gigabytes) of data per year today, and it will produce ten times more after a future high-luminosity upgrade.

Big data challenges like these are not limited to high-energy physics. When the Large Synoptic Survey Telescope begins observing the entire southern sky in never-before-seen detail, it will create a stream of 10 million time-dependent events every night and a catalog of 37 billion astronomical objects over 10 years. Another example is the future LCLS-II X-ray laser at the Department of Energy's SLAC National Accelerator Laboratory, which will fire up to a million X-ray pulses per second at materials to provide unprecedented views of atoms in motion. It will also generate tons of scientific data.

To make things more challenging, all big data applications will have to compete for available computing resources, for example when shuttling information around the globe via shared networks.

What are the tools researchers will need to handle future data piles, sift through them and identify interesting science? How will they be able to do it as fast as possible? How will they move and store tremendous data volumes efficiently and reliably? And how can they possibly accomplish all of this while facing budgets that are expected to stay flat?

"Clearly, we're at a point where we need to discuss in what direction scientific computing should be going in order to address increasing computational demands and expected shortfalls," says Richard Mount, head of computing for SLAC's Elementary Particle Physics Division.

The researcher co-chaired the 22nd International Conference on Computing in High-Energy and Nuclear Physics (CHEP 2016), held Oct. 10-14 in San Francisco, where more than 500 physicists and computing experts brainstormed possible solutions.

Here are some of their ideas.

Exascale supercomputers

Scientific computing has greatly benefited from what is known as Moore's law--the observation that the performance of computer chips has doubled every 18 months or so for the past decades. This trend has allowed scientists to handle data from increasingly sophisticated machines and perform ever more complex calculations in reasonable amounts of time.

Moore's law, based on the fact that hardware engineers were able to squeeze more and more transistors into computer chips, has recently reached its limits because transistor densities have begun to cause problems with heat.

Instead, modern hardware architectures involve multiple processor cores that run in parallel to speed up performance. Today's fastest supercomputers, which are used for demanding calculations such as climate modeling and cosmological simulations, have millions of cores and can perform tens of millions of billions of computing operations per second.

"In the US, we have a presidential mandate to further push the limits of this technology," says Debbie Bard, a big-data architect at the National Energy Research Scientific Computing Center. "The goal is to develop computing systems within the next 10 years that will allow calculations on the exascale, corresponding to at least a billion billion operations per second."

Software reengineering

Running more data analyses on supercomputers could help address some of the foreseeable computing shortfalls in high-energy physics, but the approach comes with its very own challenges.

"Existing analysis codes have to be reengineered," Bard says. "This is a monumental task, considering that many have been developed over several decades."

Maria Girone, chief technology officer at CERN openlab, a collaboration of public and private partners developing IT solutions for the global LHC community and other scientific research, says, "Computer chip manufacturers keep telling us that our software only uses a small percentage of today's processor capabilities. To catch up with the technology, we need to rewrite software in a way that it can be adapted to future hardware developments."

Part of this effort will be educating members of the high-energy physics community to write more efficient software.

"This was much easier in the past when the hardware was less complicated," says Makoto Asai, who leads SLAC's team for the development of Geant4, a widely used simulation toolkit for high-energy physics and many other applications. "We must learn the new architectures and make them more understandable for physicists, who will have to write software for our experiments."

Smarter networks and cloud computing

Today, LHC computing is accomplished with the Worldwide LHC Computing Grid, or WLCG, a network of more than 170 linked computer centers in 42 countries that provides the necessary resources to store, distribute and analyze the tens of petabytes of data produced by LHC experiments annually.

"The WLCG is working very successfully, but it doesn't always operate in the most cost-efficient way," says Ian Fisk, deputy director for computing at the Simons Foundation and former computing coordinator of the CMS experiment at the LHC.

"We need to move large amounts of data and store many copies so that they can be analyzed in various locations. In fact, two-thirds of the computing-related costs are due to storage, and we need to ask ourselves if computing can evolve so that we don't have to distribute LHC data so widely."

More use of cloud services that offer internet-based, on-demand computing could be a viable solution for remote data processing and analysis without reproducing data.

Commercial clouds have the capacity and capability to take on big data: Google, receives billions of photos per day and hundreds of hours of video every minute, posing technical challenges that have led to the development of powerful computing, storage and networking solutions.

Deep machine learning for data analysis

While conventional computer algorithms perform only operations that they are explicitly programmed to perform, machine learning uses algorithms that learn from the data and successively become better at analyzing them.

In the case of deep learning, data are processed in several computational layers that form a network of algorithms inspired by neural networks. Deep learning methods are particularly good at finding patterns in data. Search engines, text and speech recognition, and computer vision are all examples.

"There are many areas where we can learn from technology developments outside the high-energy physics realm," says Craig Tull, who co-chaired CHEP 2016 and is head of the Science Software Systems Group at Lawrence Berkeley National Laboratory. "Machine learning is a very good example. It could help us find interesting patterns in our data and detect anomalies that could potentially hint at new science."

At present, machine learning in high-energy physics is in its infancy, but researchers have begun implementing it in the analysis of data from a number of experiments, including ATLAS at the LHC, the Daya Bay neutrino experiment in China and multiple experiments at Fermi National Accelerator Laboratory near Chicago.

The most futuristic approach to scientific computing is quantum computing, an idea that goes back to the 1980s when it was first brought up by Richard Feynman and other researchers.

Unlike conventional computers, which encode information as a series of bits that can have only one of two values, quantum computers use a series of quantum bits, or qubits, that can exist in several states at once. This multitude of states at any given time exponentially increases the computing power.

A simple one-qubit system could be an atom that can be in its ground state, excited state or a superposition of both, all at the same time.

"A quantum computer with 300 qubits will have more states than there are atoms in the universe," said Professor John Martinis from the University of California, Santa Barbara, during his presentation at CHEP 2016. "We're at a point where these qubit systems work quite well and can perform simple calculations."

Martinis has teamed up with Google to build a quantum computer. In a year or so, he says, they will have built the first 50-qubit system. Then, it will take days or weeks for the largest supercomputers to validate the calculations done within a second on the quantum computer.

We might soon find out in what directions scientific computing in high-energy physics will develop: The community will give the next update at CHEP 2018 in Bulgaria.

<div align="center">###</div>

Media Contact

Ethan Alpern
news@science.doe.gov
202-586-4307

http://www.science.doe.gov ➦