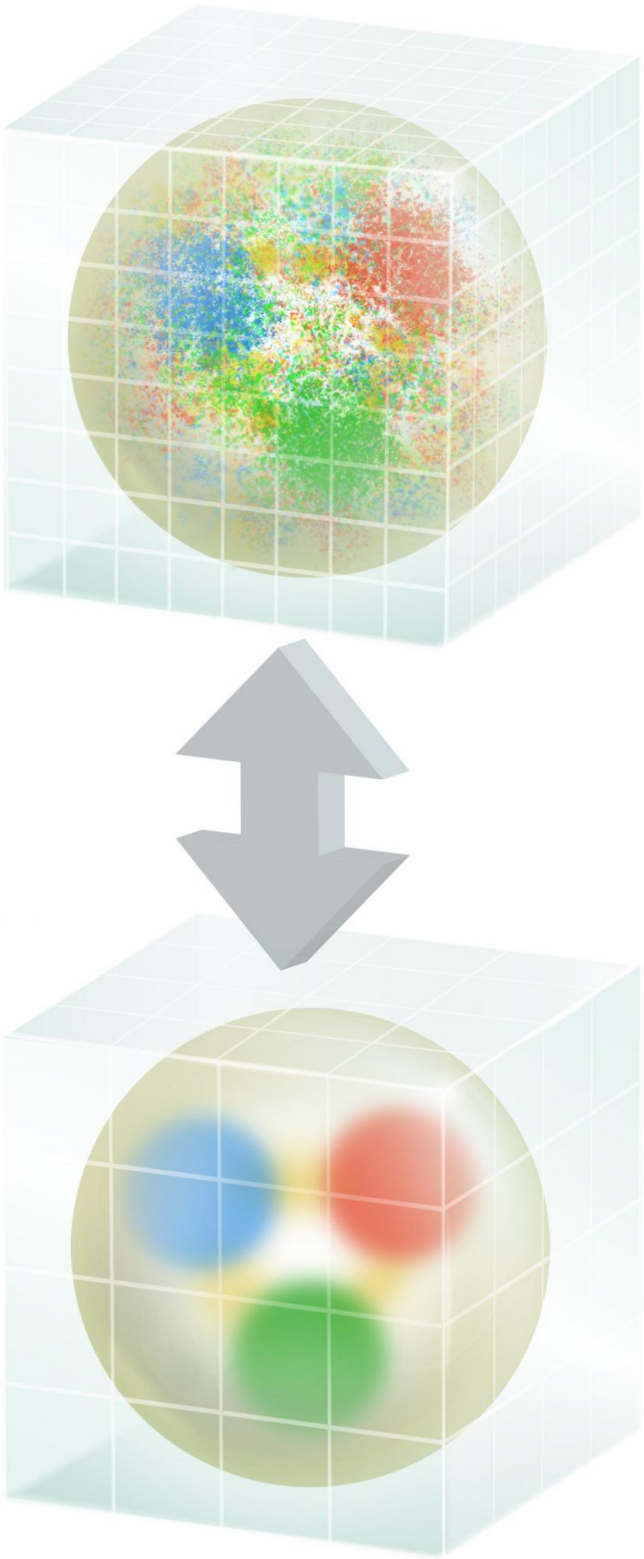
Jefferson Lab-NVIDIA collaboration uses Titan's to boost subatomic particle research

December 8 2016



This image shows an artist's depiction of the team's QCD multigrid method. The top image is fine-grid, which allows the team to simulate high-frequency noise of a proton simulation. The bottom image represents the coarse grid, which supports low-energy, long-distance modes that slow down solvers. Credit: Joanna Griffin, Jefferson Lab Public Affairs

Scientists are only beginning to understand the laws that govern the atomic world.

Before the 1950s the electrons, neutrons, and protons comprising atoms were the smallest confirmed units of matter. With advancements in experimental and theoretical techniques in recent decades, though, researchers now try to understand particles a step smaller and more fundamental.

In recent years large-scale experimental facilities, such as the Large Hadron Collider in Switzerland, have allowed researchers to begin testing theories about how subatomic particles behave under different conditions.

Research institutions funded by the US Department of Energy (DOE) have also made major investments in experimental test facilities. The newest of these facilities lies in Hall D at the Thomas Jefferson National Accelerator Facility (Jefferson Lab). The experiment, known as GlueX, aims to give researchers unprecedented insight into subatomic particle interactions.

"We believe there is a theory that describes how elementary particles interact, quarks and gluons that make up the matter around us," said

Robert Edwards, senior staff scientist at Jefferson Lab. "If so, the theory of QCD suggests that there are some exotic forms of matter that exist, and that's what we're looking for in our Hall D experiment."

Edwards serves as the principal investigator on a project that uses computation to inform the GlueX experiment as well as corroborate experimental findings. To that end the team has been using the Titan supercomputer at DOE's Oak Ridge National Laboratory (ORNL). Titan is the flagship supercomputer of the Oak Ridge Leadership Computing Facility (OLCF), a DOE Office of Science User Facility located at ORNL.

The team wants to make computer codes for quantum chromodynamics (QCD) applications run more efficiently and effectively, and with access to world-class computing resources, the researchers' computational innovations were able to achieve speedups ranging from seven- to tenfold for QCD calculations compared with those achieved in earlier work.

Mathematical mesons

The field of QCD is the study of forces between two major categories of subatomic particles—quarks and gluons.

Quarks serve as the primary force-carrying particles in an atom's nucleus and make up hadrons, a class of subatomic particles that includes protons and neutrons. Gluons, much like their name implies, allow quarks to interact with forces and serve as the "glue" that holds hadrons together.

Quarks can also bind with their inverse, antiquarks, to form mesons. Mesons are among the most mysterious of all subatomic particles because their resonances are in existence for only fractions of a microsecond. Through experiments researchers hope to use GlueX to

confirm the existence of "exotic" mesons that would help advance QCD theory.

When simulating a quantum system of quarks, gluons, and mesons, the number of calculations needed to compute the interactions of the subatomic particle fields explodes in a hurry. Researchers represent quarks and gluons by using a lattice or grid. In fact, researchers using this method call it lattice QCD (LQCD).

Once the theories are expressed in terms of the lattice, the overall simulation becomes similar to a high-school-level model of a crystal—plastic spheres at the lattice points connected by springs between them. One can think of the spheres at the lattice points as representing the quark field with the springs between them representing the quark-gluon interactions. When given energy by pushing or nudging, the model will vibrate. At any given instant, a snapshot of the model would show a particular arrangement of stretched and compressed springs. If one looked at the statistical distribution of these snapshots, he or she could deduce information about the crystal.

QCD works in a similar way. The team's lattices act as snapshots of the states of the gluon fields. By generating a statistical sampling of these QCD field snapshots and analyzing them, the team can compute the properties of the subatomic particles of interest.

Although that process might sound simple, it really isn't. Each snapshot requires a lot of computation. To compute the quark-gluon interactions, the team must repeatedly carry out the complex computation of solving the Dirac equation—a complex wave equation.

Solving the equation is complicated enough, but Jefferson Lab researcher Balint Joo noted that the team's simulations must do it many times. "Our algorithm is one that requires solving the Dirac equation

hundreds of thousands of times for each of the 300 to 500 snapshots that we take," he said.

Such computational demands push even the world's fastest supercomputers to their performance limits, and Joo and NVIDIA high-performance computing researcher Kate Clark have teamed with other researchers from the USQCD collaboration to search for new ways to improve code performance on the Jefferson Lab team's CHROMA code, among other QCD applications. They shared their results in a paper presentation at the SC16 conference, which took place November 13-18.

GPUs as the glue

Since 2005 Clark has focused on methods to improve code performance for the LQCD community. Before moving to NVIDIA, she worked in LQCD algorithms at Boston University with professor Richard Brower, where the team developed a multigrid algorithm. Essentially, computer chips have become so much faster than memory systems that memory can't feed chips the data fast enough, meaning the bottleneck for LQCD calculations comes from the speed of the memory system. Clark has been developing the QUDA library, which takes advantage of a GPU system's computational strength, including its very fast built-in memory, to improve calculation speed.

When developing its new algorithm, the Edwards team began by adding a multigrid algorithm into its code. Multigrid algorithms take the large, fine-grained lattice grid for LQCD calculations; average the various grid points; and create multiple smaller, coarser grids.

Similar to sound waves, which are really composed of many waves, each with a different pitch or frequency, the team's problem is composed of many modes with different energies. High-energy modes need a fine lattice to represent them accurately, but low-energy modes—which

usually slow down when seeking a solution—can be represented on coarser lattices with fewer points, ultimately reducing the computational cost. By using multiple grids and separating the modes in the problem onto the various grids most efficiently, the researchers can get through their long line of calculations quicker and easier.

"GPUs provide a lot of memory bandwidth," Clark said. "Solving LQCD problems computationally is almost always memory-bound, so if you can describe your problem in such a way that GPUs can get maximum use of their memory bandwidth, QCD calculations will go a lot quicker." In other words memory bandwidth is like a roadway in that having more lanes helps keep vehicles moving and lessens the potential for traffic backups.

However, the more GPUs working on a problem, the more they must communicate with one another. If too many GPUs get involved and the problem size doesn't keep up with the computational resources being used, the calculation becomes very inefficient.

"One aspect of GPUs is that they bring a lot of parallelism to the problem, and so to get maximum performance, you may need to restructure your calculation to exploit more parallelism," Clark said.

Pouncing on parallelism

Essentially, as computing technology has evolved, processing speed has improved faster than the ability of interconnects to move increasingly larger amounts of data across supercomputers' nodes. For simulations in which researchers divide their calculations across many computer nodes, this imbalance can lead to performance bottlenecks.

"With QCD the computational cost doesn't scale linearly; it scales super-linearly," Clark said. "If you double the problem size, the computational

cost goes up by more than a factor of two. I can't keep the same size of computation per node and just put it on a bigger system."

Despite performance gains through implementing the multigrid algorithm, Clark, Joo, and their collaborators noted that for maximum performance impacts, they would need to exploit sources of parallelism other than those that had typically been used in existing LQCD calculations.

Each one of Titan's 18,688 GPUs has 2,688 processing cores. To return to the roadway analogy, each one of a GPU's individual processors is a "lane" on a road, and if only one lane is open, cars back up quickly.

With that in mind, Clark, Joo, and their collaborators worked on opening up as many processing "lanes" as possible for LQCD calculations. The team recognized that in addition to exploiting parallelism by calculating multiple grids rather than a single, large grid, they could also exploit more parallelism out of each grid point.

To create multiple grids from one large, fine-grained grid, each GPU calculates a set of grid points (which appear as mathematical vectors), averages the results, and sends the averages to the middle grid point. Rather than just having one processing "lane" doing all of these calculations, researchers can use four processing cores to calculate the points above, below, and to the left and right of the original grid point.

Much like going from a one-lane road to a four-lane highway, the data throughput moves much faster. This concept works for a two-dimensional calculation, and a four-dimensional calculation can use this same concept to achieve eight-way parallelism.

In addition, the researchers noted that each grid point is not just a number but also a vector of data. By splitting up the vector calculations

to run on multiple processors, the team further increased code parallelism.

Because of these innovations, the Edwards team saw hundredfold speedups on the coarsest grids and a tenfold speedup for finer grids when comparing simulations with those that took place before the QUDA implementation. Clark and Joo pointed out that this approach affects more than the team's CHROMA code. These methods are already being applied to other QCD applications.

Clark noted that as computers continue to get more powerful by using accelerators—such as the OLCF's next-generation machine, Summit, set to begin delivering science in 2018—researchers will have to focus on getting as much parallelism as possible.

"Going forward, supercomputers like Summit will have many more processing cores, so to get high efficiency as a whole, researchers in many fields are going to have to work on how to exploit all the levels of parallelism in a problem," Clark said. "At some point exploiting all levels of parallelism is something that all researchers will have to do, and I think our work is a good example of that."

More information: M.A. Clark, Balint Joo, Alexei Strelchenko, Michael Cheng, Arjun Gambhir, and Richard C. Brower, "Accelerating Lattice QCD Multigrid on GPUs Using Fine-Grained Parallelization." SC16 Proceedings of the International Conference for High Performance Computing, Storage and Analysis, Salt Lake City, UT, November 13–18, 2016, dl.acm.org/citation.cfm?id=301 … 881&CFTOKEN=57960029.

Provided by Oak Ridge National Laboratory