Sifting through a trillion electrons

Eurekalert!

Modern research tools like supercomputers, particle colliders, and telescopes are generating so much data, so quickly, many scientists fear that soon they will not be able to keep up with the deluge.

"These instruments are capable of answering some of our most fundamental scientific questions, but it is all for nothing if we can't get a handle on the data and make sense of it," says Surendra Byna of the Lawrence Berkeley National Laboratory's (Berkeley Lab's) Scientific Data Management Group.

That's why Byna and several of his colleagues from the Berkeley Lab's Computational Research Division teamed up with researchers from the University of California, San Diego (UCSD), Los Alamos National Laboratory, Tsinghua University, and Brown University to develop novel software strategies for storing, mining, and analyzing massive datasets—more specifically, for data generated by a state-of-the-art plasma physics code called VPIC.

When the team ran VPIC on the Department of Energy's National Energy Research Scientific Computing Center's (NERSC's) Cray XE6 "Hopper" supercomputer, they generated a three-dimensional (3D) magnetic reconnection dataset of a trillion particles. VPIC simulated the process in thousands of time-steps, periodically writing a massive 32 terabyte (TB) file to disk at specified times.

Using their tools, the researchers wrote each 32 TB file to disk in about 20 minutes, at a sustained rate of 27 gigabytes per second (GB/s). By applying an enhanced version of the FastQuery tool, the team indexed this massive dataset in about 10 minutes, then queried the dataset in three seconds for interesting features to visualize.

"This is the first time anyone has ever queried and visualized 3D particle datasets of this size," says Homa Karimabadi, who leads the space physics group at UCSD.

The Problem with Trillion Particle Datasets

Magnetic reconnection is a process where the magnetic topology in a plasma (a gas made up of charged particles) is rearranged, leading to an explosive release of energy in form of plasma jets, heated plasma, and energetic particles. Reconnection is the mechanism behind the aurora borealis (a.k.a. northern lights) and solar flares, as well as fractures in Earth's protective magnetic field—fractures that allow energetic solar particles to seep into our planet's magnetosphere and wreak havoc in electronics, power grids and space satellites.

According to Karimabadi, one of the major unsolved mysteries in magnetic reconnection is the conditions and details of how energetic particles are generated.

But until recently, the closest that any researcher has come to studying this is by looking at 2D simulations. Although these datasets are much more manageable, containing at most only billions of particles, Karimabadi notes that lingering magnetic reconnection questions cannot be answered with 2D particle simulations alone. In fact, these datasets leave out a lot of critical information.

"To answer these questions we need to take into full account additional effects such as flux rope interactions and resulting turbulence that occur in 3D simulations," says Karimabadi. "But as we add another dimension, the number of particles in our simulations grows from billions to trillions. And it is impossible to pull up a trillion-particle dataset on your computer screen; it would just fill up your screen with black dots."

To address the challenges of analyzing 3D particle data, Karimabadi and a team of astrophysicists joined forces with the ExaHDF5 team, a Department of Energy funded collaboration to develop high performance I/O and analysis strategies for future exascale computers. Prabhat, of Berkeley Lab's Visualization Group, leads the ExaHDF5 team.

A Scalable Storage Approach Sets Foundation for a Successful Search

According to Byna, VPIC accurately models the complexities of magnetic reconnection at this scale by breaking down the "big picture" into distinct pieces, each of which are assigned, using Message Passing Interface (MPI), to a group of processors to compute. These groups, or MPI domains, work independently to solve their piece of the problem. By subdividing the work, researchers can simultaneously employ hundreds of thousands of processors to simulate a massive and complex phenomenon like magnetic reconnection.

In the original implementation of VPIC, each MPI domain generates a binary file once it finishes processing its assigned piece of the problem. This ensures that the data is written efficiently.

But according to Byna, this approach, called file-per-process, has a number of limitations. One major limitation is that the number of files generated for large-scale scientific simulations, like magnetic reconnection, can become unwieldy. In fact, his team's largest VPIC run on Hopper contained about 20,000 MPI domains—that's 20,000 binary files per time-step. And because most analysis tools cannot easily read binary files, another post-processing step would have been required to re-factor the data into a format that these tools can open.

"It takes a really long time to perform a simple Linux search of a 20,000-file directory; and since the data is not stored in standard data formats, such as HDF5, existing data management and visualization tools cannot directly work with the binary file," says Byna. "Ultimately, these limitations become a bottleneck to scientific analysis and discovery."

But by incorporating H5Part code into the VPIC codebase, Byna and his colleagues managed to overcome all of these challenges. H5Part is an easy-to-use veneer layer on top of HDF5, that allows for the management and analysis of extremely large particle and block-structured datasets.

According to Prabhat, this easy modification to the code-base creates one shared HDF5 file per time-step, instead of 20,000 independent binary files. Because most visualization and analysis tools can use HDF5 files, this approach eliminates the need to re-format the data. With the latest performance enhancements implemented by the ExaHDF5 team, VPIC was able to write each 32 TB time-step to disk at a sustained rate of 27 GB/s.

"This is quite an achievement when you consider that the theoretical peak I/O for the machine is about 35 GB/s," says Prabhat. "Very few production I/O frameworks and scientific applications can achieve that level of performance."

Mining a Trillion Particle Dataset with FastQuery

Once this torrent of information has been generated and stored, the next challenge that researchers face is how to make sense of it. On this front, ExaHDF5 team members Jerry Chou and John Wu implemented an enhanced version of FastQuery, an indexing and querying tool. Using this technique, they indexed the trillion-particle, 32 TB dataset in about 10 minutes, and queried the massive dataset for particles of interest in approximately three seconds. This was the first time anybody has successfully queried a trillion-particle dataset this quickly.

The team was able to accelerate FastQuery's indexing and query capabilities by implementing a hierarchical load-balancing strategy that involves a hybrid of MPI and Pthreads. At the MPI level, FastQuery breaks up the large dataset into multiple fixed-size sub-arrays. Each sub-array is then assigned to a set of compute nodes, or MPI domains, which is where the indexing and querying occurs.

The load-balancing flexibility happens within these MPI domains, where the work is dynamically pooled among threads—which are the smallest unit of processing that can be scheduled by an operating system. When constructing the indexes, the threads build bitmaps on the sub-arrays and store them into the same HDF5 file. When evaluating a query, the processors apply the query to each sub-array and return results.

Because FastQuery is built on the FastBit bitmap indexing technology, Byna notes that researchers can search their data based on an arbitrary range of conditions that is defined by available data values. This essentially means that a researcher can now feasibly search a trillion particle dataset and sift out electrons by their energy values.

According to Prabhat, this unique querying capability also serves as the basis for successfully visualizing the data. Because a typical computer displays contain on the order of a few million pixels, it is simply impossible to render a dataset with trillions of particles. So to analyze their data, researchers must reduce the number of particles in their dataset before rendering. The scientists can now achieve this by using the FastQuery tool to identify the particles of interest to render.

"Although our VPIC runs typically generate two types of data—grid and particle—we never did a whole lot with the particle data because it was really hard to extract information from a trillion particle dataset, and there was no way to sift out the useful information," says Karimabadi.

But with the new query-based visualization techniques, Karimabadi and his team were finally able to verify the localization behavior of energetic particles, gain insights into the relationship between the structure of the magnetic field and energetic particles, and investigate the agyrotropic distribution of particles near the reconnection hot-spot in a 3D trillion particle dataset.

"We have hypothesized about these phenomena in the past, but it was only the development and application of these new analysis tools that enabled us to unlock the scientific discoveries and insights," says Karimabadi. "With these new tools, we can now go back to our archive of particle datasets and look at the physics questions that we couldn't get at before."

"Most of today's simulation codes generate datasets on the order of tens of millions to a few billon particles, so a trillion-particle dataset—that is, a million-million particles—poses unprecedented data management challenges," says Prabhat. "In this work, we have demonstrated that the HDF5 I/O middleware and the FastBit indexing technology can handle these truly massive datasets and operate at scale on current petascale platforms."

But according to Prabhat, exascale platforms will produce even larger datasets in the near future, and researchers need to come up with novel techniques and usable software that can facilitate scientific discovery going forward. He notes that one of the primary goals of the ExaHDF5 team is to scale the widely used HDF5 I/O middleware to operate on modern petascale and future exascale platforms.

Source URL (retrieved on *12/21/2013 - 12:09pm*):
http://www.ecnmag.com/news/2012/06/sifting-through-trillion-electrons