

Finding One in a Trillion

FastBit and HDF5-FastQuery: Efficient Index and Query Technologies

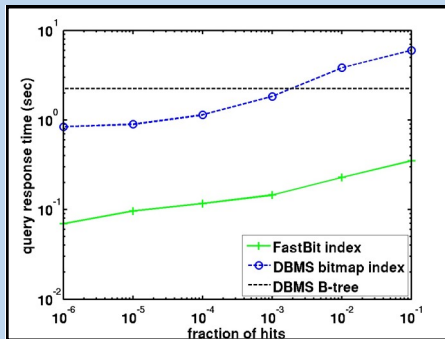
A critical challenge in the trillion-particle project was the ability to **perform queries on the 350 terabytes of information** stored in HDF5. The problem was solved using two novel technologies developed at LBNL called **FastBit** and **HDF5-FastQuery**.

FastBit creates a space-efficient index to multidimensional data and provides a set of functions for querying the data.

HDF5-FastQuery is a hybrid version of FastBit that enabled the team to query the trillion particle data using LBNL's 120,000 processor Hopper computer.

A FastBit compressed index is more than 10x faster than the compressed bitmap index implementation from a popular commercial database management (DBMS) product.

Because of HDF5's ability to mix and match different types of data, HDF5-FastQuery conveniently stores the indexes and the data in the same file.



The figure shows a set of timing measurements using a high-energy physics dataset. **HDF5-FastQuery** took only 10 minutes to index and 3 seconds to query the massive dataset of energetic particles.

For more information about **FastBit** and **FastQuery**, follow these links:

<http://www-vis.lbl.gov/Events/SC05/HDF5FastQuery/>

<http://www.slideshare.net/HDFEOS/hdf-fast-query-shalf>

<https://sdm.lbl.gov/~kewu/fastbit/applications.html>

Flexible technical capabilities—the kind that spur discovery—are possible when powerful I/O is performing for you.



Streams of charged particles called plasma continually boil from the surface of the sun and bombard the magnetic field that surrounds Earth like a protective shield. Most are deflected safely away. But others are pulled in towards Earth's magnetic poles, and when conditions are right, accelerate downwards along magnetic field lines and collide with atoms and molecules in the upper

atmosphere. The energy emitted in each collision bursts across the polar skies as light in brilliant auroras. The same mechanism causes solar flares and can fracture Earth's magnetic shield, wreaking havoc on electronics, power grids, and space satellites.

The question the scientists want to answer is why some particles are accelerated to very high energy and others are not. **It's a trillion-particle question.** To model the process, scientists have to simulate underlying physics on scales that range from the tiny motions of electrons we can't see to 100 times the radii of the Earth, all in three dimensions.

*"What we demonstrated is that **HDF5 can scale to petascale platforms and achieve near peak I/O rates,**" says Prabhat, a researcher at Berkeley Lab's Scientific Visualization Group who helped lead the simulation. "**HDF5 is on the trajectory toward exascale computing.**"*

The first successful simulation of this scale was conducted at the Lawrence Berkeley National Laboratory (LBNL) in 2013 and was **powered by HDF5**.

"We can't save all the data for all the particles over the lifetime of the simulation, so we did the next best thing," says Homa Karimabadi, a physicist at the University of California, San Diego, and one of the lead scientists. "We ran the simulation and stored the particle data at multiple timesteps and then used visualization tools to focus on the time and regions where acceleration was occurring."

The key, says Karimabadi, was finding the small number of particles that mattered—the one hundred, the thousand, or maybe one million in a trillion.

In total, 10 separate trillion-particle datasets, each ranging between 30 and 42 terabytes in size, were written as HDF5 files at rates reaching 90 percent of maximum and a sustained rate of 27 out of a possible 35 gigabytes per second. Larger simulations are in the works. [Learn more.](#)

HDF5 provided the speed and reliability for such first-ever capabilities at this scale. This is the first kinetic plasma physics to demonstrate a power-law distribution; HDF5's analysis capabilities directly facilitated this discovery.

<http://www.mendeley.com/catalog/parallel-io-analysis-visualization-trillion-particle-simulation/>