
CERN

AMASSING DATA TO FIND A NEW KIND OF PARTICLE

A COMBINATION OF HARDWARE AND SOFTWARE QUICKLY DECIDES WHICH 100 OF THE 40 MILLION COLLISIONS OCCURRING EACH SECOND CONTAIN INTERESTING FEATURES THAT COULD BE THE SIGNATURE OF A NEW KIND OF PARTICLE.

What's a Petabyte of Data?

A stack of books tall enough to reach halfway from the Earth to the moon or enough DVDs to keep you watching movies 24 hours a day, seven days a week, for 45 years (that's about 43 billion frames). The information content of these books or movies in digital compressed form is approximately one petabyte (PB) of data: what each of the two general purpose high energy physics experiments—CMS (Compact Muon Solenoid) and ATLAS (A Toroidal LHC Apparatus)—expects to collect during 2008, the first year of operation of the Large Hadron Collider (LHC) for physics studies.

Snapshots of Collisions

The data from the CMS experiment will consist of an electronic "snapshot" for each of about two billion proton-proton (pp) collisions that were routed to storage by the experiment's trigger. A combination of hardware and software quickly decides which 100 of the 40 million collisions occurring each second contain interesting features that could be the signature of a new kind of particle. Each snapshot records digitized information from the subset of the 15 million individual detector elements that registered signals as one of the particles ejected.

Sorting through the Snapshots

To allow the 2,000 physicists collaborating on CMS to analyze these data efficiently, each snapshot of detector information must be transformed into a different representation: the direction, speed, and type of particles emanating from the pp collision. In order to accomplish this transformation, sophisticated pattern recognition software—involving millions of lines of programming code—combs through each snapshot, grouping the signals that are consistent with the passage of each individual particle. For example, particles with an electric charge will leave a "footprint" in each of a series of detector layers. These footprints will form a track along the trajectory of a particle, which the tracking pattern recognition can reconstruct, just like a good guide can reconstruct the path of an animal. And


CERN

just like the guide can judge the weight of an animal from the depth of its footprint, the reconstruction software can judge the mass of a particle, and hence its type, from the strength of each footprint signal. Sorting out all of the information from several hundred particles that are produced in a collision takes some time—about 30 seconds per collision. This would take a laptop about two millennia to process and would double the data size. Each collaboration therefore needs several tens of thousands of computers.

Massive Data Collection Meets the Grid

The computing environment that enables the analysis of this data faces further challenges. The approximately 2,000 CMS physicists and their computing and storage resources are distributed over every continent. Each analysis undertaken by these physicists requires a detailed simulation of these very complex detectors to understand how the detector, never perfect, might distort the physical quantities being measured and how other physical processes might mimic their signal. The aggregate of the simulated data sets can easily outstrip the real data in both size and required computing power.

LHC physicists are turning to a new type of computing model—Grid computing. ... You plug in your application, and it accesses computing resources from a wide variety of sources without needing to know where the resources are located.

PARTICLES WITH AN ELECTRIC CHARGE WILL LEAVE A "FOOTPRINT" IN EACH OF A SERIES OF DETECTOR LAYERS. THESE FOOTPRINTS WILL FORM A TRACK ALONG THE TRAJECTORY OF A PARTICLE, WHICH THE TRACKING PATTERN RECOGNITION CAN RECONSTRUCT, JUST LIKE A GOOD GUIDE CAN RECONSTRUCT THE PATH OF AN ANIMAL.

To face these computing demands and take advantage of the computing resources distributed throughout the world, the LHC physicists are turning to a new type of computing model—Grid computing. Grid developers liken the computing model to the power grid: you plug in your application, and it accesses computing resources from a wide variety of sources without needing to know where the resources are located—national labs, universities or university groups, or even home computers. Advances in network capabilities over the last few years have made such a distributed computing model feasible. A software suite known as middleware sits on top of the network resources and provides grid control and resource management, so that the system's complexity can be hidden from the grid user. Various versions of this middleware currently exist, developed, for example, in Europe for the LHC Computing Grid and in the United States separately for the Open Science Grid and TeraGrid. As physicists gain experience with these grids, they are learning how to standardize the interfaces to this controlling middleware, so that user applications can eventually interoperate seamlessly with multiple grid infrastructures.

Via the computing grid, each country and university can share in the resource "wealth," maintaining control of its own computing power, yet contributing to and benefiting from a larger global grid facility. The Grid will not only address the computing challenge, but also help resolve a perhaps even stickier issue: the political and sociological prickles associated with amassing these resources from around the world.

GRIDS STAND ... TO CHANGE THE WAY SCIENTISTS CAN AND WILL APPROACH COMPUTING IN THE FUTURE. AND A GOOD THING FOR CMS ... IT WOULD TAKE A LOT OF 747s TO HAUL THAT STACK OF BOOKS AROUND!

Grids stand, once again, to change the way scientists can and will approach computing in the future. And a good thing for CMS, because it would take a lot of 747s to haul that stack of books around!

Lawrence Gibbons
Physics

Frank DiMeo

Lawrence Gibbons

The 2,000 CMS physicists and their computing and storage resources are distributed on every continent.

For more information:

E-mail: lkg@mail.lns.cornell.edu