Computing Frontier Group I5: Data Management and Storage

Michelle Butler/NCSA, Richard Mount/SLAC; Mike Hildreth/Notre Dame

Input from the Science Frontiers

Energy Frontier

Experiments at energy frontier hadron colliders already generate over a petabyte per second of data at the detector device level. Triggering and real-time event filtering is used to reduce this by six orders of magnitude for a final rate to persistent storage of around one gigabyte per second in the case of LHC experiments at the start of Run 2. The main requirement dictating the rate to storage is keeping the storage cost, and the cost of the computing to analyze the stored data, at a tolerable level.

Science at energy frontier lepton colliders is unlikely to be constrained by data management and storage issues.

The current practice of ATLAS and CMS is to treat the all the data written to persistent storage equally through the production phases of reconstruction, re-reconstruction and worldwide distribution of a complete set of data ready for physics analysis. Flagging a large fraction of the persistent data for storage on tape only, with no further reconstruction or distribution unless a physics case arose, would cut costs or allow the rate to persistent storage to be raised.

Distributed data management is proving to be a major success of the LHC experiments and is an area where the US brings much of the intellectual leadership. But this success is costly. For example the US-ATLAS M&O-funded effort that contributes to developing and operating the ATLAS distributed workflow and data management system amounts to around $4.5M per year. This is in addition to the effort required to operate the US Tier 1 and Tier 2 facilities. It is difficult to imagine that the LHC-focused effort could continue at this level for two decades, and even more difficult to imagine that other HEP or wider science activities could take advantage of the LHC experiments current achievements given their operational cost and complexity.

With respect to data preservation and open availability, the LHC experiments are actively developing appropriate policies.

Intensity Frontier

Lepton colliders in intensity frontier “factory mode” also run up against the cost of storage, but the physics of lepton collisions is relatively clean and recording all events relevant to the targeted physics has proved possible in the past and is a realistic expectation for the future. The Belle II TDR estimates a data rate to persistent storage of 0.4 to 1.8 gigabytes/s, which is comparable to LHC Run-2 rates but without the need to discard data with significant physics content.

Most of the many other intensity frontier experiments do not individually challenge storage capabilities, but there is a recognition that data management (and workflow management) is

often inefficient and burdensome. Most experiments find it hard to escape from the comfort and constriction of limiting all their data-intensive work to a single site – normally Fermilab. The statement “all international efforts would benefit from an ATLAS-like model” was made and should probably be interpreted as a need for ATLAS-like data management functions at a much lower cost and complexity than the current ATLAS system.

CTA has a rather specialized real-time data challenge where some 30 gigabytes/s of data must be gathered and processed in real time from about 100 telescopes spread over a square kilometer.

With respect to data preservation and open availability, the intensity frontier community does not have a plan, but recognizes that the issue exists across the frontiers.

Cosmic Frontier

The cosmic frontier presents several faces, each presenting its own challenges for data management and storage: terrestrial sky survey telescopes, terrestrial radio telescopes, HEP-detector-in-space telescopes, and large-scale simulations.

Sky Surveys

The Sloan Digital Sky Survey pioneered the use of innovative database technology to make its data maximally useful to scientists. This approach continues with LSST, notably the development of a multi-petabyte scalable object catalog database that is capable of rapid response to complex queries. The data management needs of the sky surveys – handling image catalogs and object catalogs – appear very different from those of experimental HEP, but nevertheless, the baseline LSST object catalog employs HEP’s xrootd technology in the key role of providing a switchyard between MYSQL front ends and thousands of MYSQL backend servers. LSST’s 3.2 gigapixel camera will produce 15 terabytes per night, building up to over 100 petabytes of images and 20 petabytes of catalog database during the first ten years.

Although the basic data-access technology to make LSST science achievable has already been demonstrated, it is certain that a vigorous LSST science community will want to attempt many scientific studies that will be poorly served without major additional developments.

The Dark Energy Survey (DES) can be considered a precursor to LSST, taking data with a 0.6 gigapixel camera for five years from 2012 culminating in a petabyte dataset.

The images, and tabular object catalogs of sky surveys and other image-based astronomy are readily intelligible by other scientists and even the general public. From the experimental HEP perspective, data preservation and open availability is relatively simple to achieve once policies have been decided.

Terrestrial Radio Telescopes

Arrays of radio telescopes can present a data-volume challenge comparable with that posed by energy frontier hadron collider experiments. The most extreme example now being planned is the European-led Square Kilometre Array (SKA) project that expects to complete its Phase I

system in 2020. SKA will feed petabytes/s to correllators that will synthesize images in real time, producing a reduced persistent dataset on the scale of 300 to 1500 petabytes per year. These volumes can only be realized if considerable evolution of computing and storage costs happens by the time SKA data flows.

Todays example of the SKA concept is the Murchison Wide-Field array where a raw 15.8 gigbytes/s is processed to a produce a stored 400 MB/s.

"HEP Detectors" in Space

Examples include the Fermi Gama Space Telescope (FGST) and the Alpha Magnetic Spectrometer (AMS-02). These detectors have front-end data rates far lower than LHC experiments and would not be constrained by terrestrial storage and data analysis capabilities. The choke points determining their trigger rates to persistent storage are the limited bandwidth of the downlinks that bring data back to Earth. The necessary conservatism applied by NASA and other space agencies to placing new technology in space seems guaranteed to keep downlink bandwidths well below rates that would make storage and data distribution a challenge in the future. Nevertheless, these detectors are built and operated by large collaborations and thus require functional distributed data management.

The raw persistent data from these devices contrasts markedly with that from the image-based telescopes. Like data from almost any HEP experiment, they are intelligible only to a few experts until substantial reconstruction and analysis has been performed. They present the same data preservation and open availability challenges as HEP experiments, and may be subject to the higher availability expectations typical of image-based astronomy

Simulations

Simulation provides our only way to perform "experimental cosmology" since only one universe is observable. Simulation also plays a vital role in understanding all aspects of astrophysics, such as supernovae, for which only very limited observation data can be collected for each occurrence. Finally, simulation is needed for the design of observational programs and for their detailed technical elements.

Already today, post-processing of simulation data presents a major data-intensive computing challenge, requiring data management, large-scale databases and tools for data analytics. Some of today's pain relates to the much more ready availability of national resources for computation than those for data management and analysis: "we can easily generate many petabytes from simulations and have [almost] no place to store them and analyze them".

There is some expectation that compute-intensive simulation will be co-located with the data-intensive facilities for analysis of the simulation, but powerful, easy to use analysis tools will still need to be developed.

Lattice Field Theory (LQCD)

Like other simulation-based sciences, LQCD appears to use massive simulations on national supercomputer facilities, followed by intense analysis of the resultant data. However, in the

more conventional view of LQCD, the configuration generation step is performed on massively parallel supercomputers because these are best adapted to the problem, whereas the subsequent, and by no means less compute-intensive, analysis step is performed on HEP-funded throughput-optimized systems because these are the most appropriate for this step.

LQCD has significant, but not problematic data volumes that must be managed and transmitted between the two steps, but does not face major data-related challenges.

Perturbative QCD

Data-related challenges are expected to remain minor in relation to those of other branches of HEP.

Accelerator Science

In broad summary, accelerator science is not a driver in data (or networking), but would certainly welcome access to the easy-to-use data management and analysis tools that are the goal of a wide range of HEP experiments.

Accelerator science has a long held dream of being able to perform predictive simulations in close-to-real time so that feedback can be provided to physicists in the control room as they strive to optimize accelerator performance. A likely scenario involves running a massively parallel simulation for a relatively short time on a remote Leadership Class Facility, followed by the rapid transfer of tens or hundreds of terabytes of simulated data to local facilities for rapid analysis. This scenario is becoming achievable, but will stretch the limits of data transmission bandwidths and of rapid data analysis.