International Journal of Modern Physics A

WHAT YOUR NEXT EXPERIMENT'S DATA WILL LOOK LIKE: EVENT STORES IN THE LARGE HADRON COLLIDER ERA

DAVID MALON*

Decision and Information Sciences Division, Argonne National Laboratory
9700 South Cass Avenue, Argonne, Illinois 60439, United States of America

Received 31 October 2004
Revised 31 October 2004

Each new generation of collider experiments confronts the challenge of delivering an event store having at least the performance and functionality of current-generation stores, in the presence of an order of magnitude more data and new computing paradigms (object orientation just a few years ago; grid and service-based computing today). The ATLAS experiment at the Large Hadron Collider, for example, will produce 1.6-megabyte events at 200 Hz–an annual raw data volume of 3.2 petabytes. With derived and simulated data, the total volume may approach 10 petabytes per year.

Scale, however, is not the only challenge. In the Large Hadron Collider (LHC) experiments, the preponderance of computing power will come from outside the host laboratory. More significantly, no single site will host a complete copy of the event store–data will be distributed, not simply replicated for convenience, and many physics analyses will routinely require distributed (grid) computing.

This paper uses the emerging ATLAS computing model to provide a glimpse of how next-generation event stores are taking shape, touching on key issues in navigation, distribution, scale, coherence, data models and representation, metadata infrastructure, and the role(s) of databases in event store management.

Keywords: Data management; databases; event stores

1. LHC Event Store Characteristics

LHC experiments such as ATLAS have approximately an order of magnitude more readout channels than current Fermilab Run II experiments. ATLAS, for example, expects more than 3 petabytes of raw data per year. Event Summary Data (ESD—the principal output of reconstruction) currently amounts to over 1.0 MB/event, but this does not include a great deal of calorimetry data that may also need to be retained. Analysis Object Data (AOD), the primary input to analysis, is currently about 25 KB/event. With simulation, ATLAS expects more than 10 petabytes of event data annually.

*Work supported in part by the US Department of Energy, Division of High Energy Physics, under Contract W-31-109-ENG-38

The raw data distribution model is to place a fraction 1/N at each of N Tier 1 sites (N = 6), with an archival copy at CERN. ESD from first-pass reconstruction will be distributed so that a fraction 2/N is available at each of the N Tier 1 sites. The full AOD is expected to be available at all Tier 1s, and propagated to Tier 2 sites and beyond. For first-pass AOD production at CERN, the current (highly contentious) proposal is that all streams share same the first-pass AOD schema, to support potential analyses that cross stream boundaries. Customized AOD may be produced by physics working groups running production jobs at Tier 1 sites.

An analysis job that requires access only to AOD may run at many large LHC sites, or be distributed across several sites. Jobs that require more than AOD content (e.g., access to ESD) must run in a distributed environment. Re-reconstruction will always be distributed, not simply to take advantage of remote CPUs, but because remote sites are where the raw data live. The role of the host laboratory is diminished compared to previous-generation experiments: at least 2/3 of computing power will reside outside of CERN, and this fraction can only increase. Remote sites provide essentially all simulation cycles, and simulation data storage. Distributed processing (the grid) must work well, and not just for production managers—end-user physics analyses will rely upon it.

2. Event Store Technologies

Three of four LHC experiments (ATLAS, CMS, LHCb) will use an LCG-developed object streaming technology called POOL for event storage. The fourth, ALICE, will use ROOT directly. POOL is designed to support hybrid stores built from file-based and relational database components. Its object streaming layer, while designed for technology neutrality, is aimed principally at file-based storage for most data, with a primary implementation based upon ROOT. Relational databases (and grid replica catalogs) host the associated file catalogs. POOL also supports relational-database-hosted event collections and event-level metadata services.

Cross-file pointers are integral to this approach. LHC experiments incur the usual consequences of a persistence strategy: the reader's view is essentially the writer's view. This limitation is one of the reasons people like ntuples, which allow direct access to the data members they need and no others, and from which they can, in principle, build their own objects. (Such an approach may not be very object-oriented, but should one care?)

Objects do not have inherent identity, and hence are not readily relocatable. A persistent representation of a pointer to a data object records the id of the logical file in which it was stored. The object can be retrieved from any replica of the original file, but if the object is copied into a file with different contents, old pointers will not know how to find that object copy.

Since objects in one file can point to objects in another, it is natural to wonder whether HEP experiments will finally have a transparently globally interconnected object storage system. ATLAS will support such interconnectedness architecturally, but limitations are likely in deployment and as a matter of policy. In ATLAS, there are navigable pointers from AOD to ESD to RAW, for example, but whether one will be able to follow them may depend upon such factors as whether files that contain the pointed-to objects are local, or whether jobs are allowed to trigger remote data transfers. It is likely, for example, that AOD-to-ESD navigation will work well at a Tier 1 site and not at all somewhere else.

3. Metadata

With petabyte-scale event stores, a metadata system's ability to support efficient discovery, identification, selection, location, and access to data of interest is crucial. LHC common metadata infrastructure is still immature, and several experiment-specific approaches are in use in LHC data challenges. Common metadata tools work is ongoing, e.g., in POOL, and in metadata service components identified in gLite (EGEE). It is not yet clear how useful generic tools will be, and even generic middleware providers such as the Globus team are undecided about the role and value of non-domain-specific tools. Because metadata encompass many things and come from many sources, and integration is a major challenge.

An expanded role for event-level metadata is one potential difference in LHC event stores. Event tags contain metadata about an event that allow physicists to make first-pass decisions about whether the event is of interest to their analyses. Tags also contain pointers to events in persistent storage, to allow efficient retrieval. The model is this: one queries the tag database, gets a list of (pointers to) events that satisfy the query predicate, and iterates over exactly those events and no others. Middleware determines which files are needed to provide selected events as input, and marshals the data and computational resources needed to process the selection.

In principle, ATLAS physicists can select whatever events satisfy their selection criteria, independent both of stream boundaries and of standard "skims," and extract them into their own files. This democratization of analysis, allowing selections orthogonal to predefined samples while maintaining appropriate provenance information, is a desideratum of ATLAS event store design.

4. Summary

Views of LHC event stores are coalescing. The basic structure and data products in LHC event stores will be familiar–no major paradigm shifts are expected before LHC turn-on. Stores are shaping up to be essentially file-based for bulk data storage, with relational databases for catalogs and metadata. The role of metadata is increasing, as is the functionality of metadata systems. Event data will be widely and routinely distributed: no single site may host sufficient data to allow many standard jobs to run there alone. Event store deployment is strongly dependent upon still-emerging distributed computing infrastructure and middleware. Physicists will be able to run jobs all over the world—and may need to.