# What Your Next Experiment's Data Might Look Like: Events Stores in the LHC Era

**David M. Malon**

**Argonne National Laboratory**

APS Division of Particles and Fields Meeting

Riverside, California

30 August 2004

# Questions

- Will LHC event stores look like current-generation event stores, only larger?
- Will issues of scale be resolved simply by increases in processor speed and decreases in storage costs?
  - Tens of petabytes may have been daunting on SSC time scales, but will they be routine when LHC turns on?
- Will event stores look different, be structured differently, be accessed differently than current-generation stores?

# Illustrative Scale: ATLAS

- Approximately an order of magnitude more readout channels in ATLAS than in current Fermilab Run II experiments
- Raw data:
  - Expect 1.6 MB/event, 200 events/second → >3 petabytes/year
- Event Summary Data (principal output of reconstruction)
  - Size is currently ~0.5 MB/event, but this does not include a great deal of calorimetry data that may also need to be retained
- Analysis Object Data
  - Currently ~100 KB/event
- With simulation, expect ~10 petabytes/year ATLAS event data

# Event data flow

- RAW data: 1/N to each of N "Tier 1" sites (N = ~6); archival copy at CERN
- ESD from first-pass reconstruction: 2/N to each of the N Tier 1 sites (ESD for any given event is available at two Tier 1 sites)
  - Each Tier 1 site assumes principal responsibility for 1/N share, hosts a replica of another 1/N share
  - ~1 petabyte of ESD/year per reconstruction pass
- AOD: Full AOD available at all Tier 1s, and propagated to Tier 2s and beyond
  - 200 TB AOD/year (single pass)

# Event data flow

- First-pass AOD production at CERN
- Current proposal (highly contentious): all streams share same 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

# Consequences

- An analysis job that requires access only to AOD may run in many places (and while many single sites could host such a job, it may in fact be distributed across sites)
- Jobs that require more than AOD content (e.g., access to ESD) must run in a distributed environment
  - U.S. Tier 1 may be an exception here, and host complete ATLAS ESD
- Re-reconstruction will always be distributed, not simply to take advantage of remote CPUs, but because remote sites are where the raw data live
- Diminished role of host laboratory:  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, and not just for production managers

# Storage technology

- Three of four LHC experiments (ATLAS, CMS, LHCb) will use an LCG-developed object streaming technology called POOL for event storage
  - ALICE will use ROOT directly
- Designed to support "hybrid" stores:
  - Object streaming layer, while designed for technology neutrality, is aimed principally at file-based storage for most data: primary implementation based upon ROOT
  - Corresponding file catalogs supported in relational databases (and grid replica catalogs)
  - Relational-database-hosted event collections and event-level metadata services
- (POOL is a bit broader than this, but this is not a POOL talk)
- While there are always good reasons for technology-neutral layers and higher levels of abstraction, the perennial challenge for this kind of strategy is to deliver sufficient added value without making it needlessly difficult for people to take advantage of the underlying technology
  - Does a thin layer over ROOT help you more than it gets in your way?

# Streaming layer notes

- Cross-file pointers are integral to this approach
- Usual consequences of "persistence" approach: a reader's view is essentially the writer's view
  - If I stream a Foo object out, a Foo object is what I read in (or a Foo Version 2, if I have schema evolution)
  - Aside: This is one of the reasons people like ntuples: direct access to the data members they need and no others, from which they can, in principle, build their own objects (…not very object-oriented, but do you care?)
  - Some machinery in place to do more general things; work is ongoing

# Streaming layer notes

- Objects do not have inherent identity, and hence are not readily relocatable
- 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
  - CMS can also find the object in a tar file that contains the original file; still, LHC experiments are largely at the point where objects can only be found in replicas of the original files in which they were stored

# Navigation

- Since objects in one file can point to objects in another, will we finally have a transparently globally interconnected object storage system?
  - No.
  - Yes.
  - Sort of…
- 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 you 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, or …
  - Likely, for example, that AOD→ESD navigation will work well at a Tier 1 site and not at all somewhere else

# Metadata infrastructure

- Metadata will be key: with petabyte-scale event stores, much depends upon how well an experiment's metadata system supports efficient discovery, identification, selection, location, and access to data of interest
- LHC common metadata infrastructure is not very mature yet; several experiment-specific approaches in use in the data challenges
  - Common metadata tools work ongoing, e.g., in POOL, and in metadata service components identified in gLite (EGEE)
  - Not clear yet how useful generic tools will be (e.g., Globus team, while it has a Metadata Catalog Service (MCS), is not including MCS in Globus distributions, and is undecided about the role and value of non-domain-specific tools)
- Metadata come from many sources, and integration is a major challenge

# Event-level metadata

- An expanded role for event-level metadata ("tag" databases) is one potential difference in LHC event stores (in ATLAS, anyway)
- In principle, end-user physicists can select whatever events satisfy their selection criteria—and extract them into their own files—across stream boundaries, and independent of standard "skims"
- Certainly not the first time this has been tried
- It remains an open question how useful this will be at the end-user analyst level
  - Marching to the beat of a different drummer may be difficult, but it should not be impossible

# Event Tags and Event Collections

- Event tags contain event-level metadata: 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
- Event collections are lists of pointers to events in persistent storage, along with (optionally) event-level metadata (event tags)
  - You may hear these things called "explicit collections" (correctly so)
  - ATLAS tag databases are event collections with agreed-upon tags
- Model:
  - Query the resulting tag database
  - Get a list of (pointers to) events that satisfy the query predicate
  - Iterate over exactly those events and no others
  - Middleware handles issues of which files are needed to provide selected events as input
- Lots of middleware (grid and other) needed to build jobs from event-level selections

# The mandatory bow to grids

- Grids are currently used largely as distributed batch queues by production managers (and workflow management above the level of marshalling resources for and running a single job tends to be done with experiment-specific tools), but this will change
- Grid projects are discovering web services, and this is a good thing
- Grid-based distributed analysis infrastructure projects are underway in several venues in which LHC experiments participate
- LHC data challenges are exposing many shortcomings in the still-emerging grid-based distributed computing fabric
  - A daunting but progressively clearer development agenda remains

# Summary

- Basic structure and data products in LHC event stores will be familiar
  - No major paradigm shifts here
- Stores are shaping up to be essentially file-based for bulk data storage, with relational databases for catalogs and metadata
- Role of metadata is increasing (as is 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
  - You will be able to run jobs everywhere—and you may need to
- Event store deployment is strongly dependent upon still-emerging distributed computing infrastructure and middleware
- Views of LHC event stores are coalescing (…but then, we have no data yet)