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.

The vast majority of LHC data storage and data access relates to derived or simulated data stored and accessed using ROOT persistency. Efficient and agile data access is already a major issue that is expected to rise in importance by the time of LHC Run 3. Beyond HEP, efficient and agile data access will underpin future successes in data-intensive scientific and commercial applications. It is hard, but not impossible, to believe that HEP will continue to be best served by a persistency solution that appears to be confined to HEP.

Distributed data management is proving to be a major success of the LHC experiments, as illustrated in Figures xx,yy,zz., 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. Not all LSST science will be possible using only the object catalog database. In particular, studies such as those for dark energy effects, of particular interest to the HEP community, are likely to require reprocessing of the LSST image data on HEP analysis facilities. The model for funding and executing these studies is not yet clear.

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 correlators 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. Although SKA currently has no US involvement, it presents a concretely planned example of the technologies and data-related challenges that will certainly be faced by US scientists involved in projects in the same timeframe.

Todays example of the SKA concept is the Murchison Wide-Field array where a raw 15.8 gigabytes/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.

General Considerations for Large-Scale HEP

Large and costly experiments or telescopes, and even some simulations, require international collaboration. Whatever could be done by one nation can be done on a larger scale with more science reach as a collaborative project. In such international projects, the data storage and analysis can take advantage of funding from many nations, access to large shared resources,

opportunistic access to a wide range of resources, and access to and development of distributed expertise. The price to be paid is the complexity of a geographically distributed system.

Thus HEP, and indeed any data-intensive science operating at a comparable international scale, needs to use a combination of wide area networks, distributed storage, distributed computing, and the software technologies to co-manage efficient worldwide work flow and data flow. It seems clear that future HEP does not need many different solutions to this challenge, and may have much to gain by identifying commonalities with other scientific or even commercial fields.

Technology Outlook

A simplistic prediction for the future evolution of technology would to continue to evolve as it has in the past. Figure aa shows a highly selective, but relevant, view of technology evolution, charting how much you can get for $1M in disk storage, CPU power and long-haul network links that have been bought by over three decades for some particular experimental HEP activities[1]. Some notable features of the evolution are:

- Over three decades the line "doubling every 1.3 years" is a good match to the average CPU evolution and is not far from the average disk capacity evolution.
- The data do not exclude a marked slow down in evolution from about 2010.
- The network evolution shows a major discontinuity, corresponding to de-regulation and the end of European PTT monopolies
- Disk accesses per $M – almost totally dominated by the unchanging rotational speed of disks – has changed hardly at all in two decades.
- Comparison of technology evolution with the BaBar and ATLAS raw data rates to persistent storage (just two examples of HEP's "data frontier") shows a similar distance from the technology evolution and hence, perhaps a similar level of technological challenge but separated by more than a decade in time.

Storage Futures

Tape

The death of tape storage has been predicted for more than a decade, but today its future seems to be more assured than at any time in the last decade. Much of this change relates to the problems with disk technology described below, but tape storage does have the intrinsic properties of negligible power usage, a different set of failure modes and often lower failure rate than disks, and persistently lower cost than disks. The scientific and the commercial world has and will continue to have, a need for archival or "just in case we need it" storage with these properties.

Tape does fail, so HEP has traditionally taken advantage of its enforced distributed approach to computing to make a copy or copies elsewhere in the world. A cheaper approach is possible

[1] Specifically, equipment and networking bought with the involvement of R. Mount and/or H. Newman for the Mark J, L3/LEP, BaBar and some aspects of the LHC experiments.

with RAIT (redundant array of individual tapes), analogous to RAID, provided that the data are stored away from areas prone to major natural disasters.

Data volumes in science beyond HEP seem to be doubling every year according to NCSA experience. This makes sense as driven not just be the increasing data hunger of individual sciences, but also by the increasing number of fields that are becoming data intensive. Hence the view from outside HEP is of a growing role for tape in an optimized scientific data management hierarchy.

Estimating tape costs is challenging. The smallest component is usually the tape media, followed in ascending order by the acquisition of the drive-plus-robot system, the maintenance of the drives and robots, and the highly skilled labor needed to operate arcane tape-data-management systems such as HPSS. It is reasonable to expect that as tape consolidates its position in scientific computing, less labor intensive approaches to its integration in the storage hierarchy will appear, not least due to development efforts at NCSA and other major scientific computing centers that see a clear need for this integration.

Tape technology has, for many decades, been limited by marketing considerations more than by technology. Tape is very far from its fundamental physical limitations and suffers mainly from the existence of very few – approximately two today – major drive manufacturers who can see no reason to compete with themselves. The technology could deliver twice the capacity per dollar every 18 months, but it seems reasonable to expect a doubling time of around three years for the marketed products.

Rotating Disks

The 30-year run of exponential growth in capacity per dollar is almost certainly over. The factors responsible are both market related and technology related.

The consumer market for rotating disks is now declining. An increasing number of consumer devices – for example most HEP laptops – are now being bought with flash memory taking over the historical role of disk. One consequence is that an appealing and for-a-short-time successful HEP storage strategy – buy the cheapest, slowest, largest consumer-market disks and hide them behind a layer of faster enterprise disks or solid-state-disks – no longer works as well as it did.

In the enterprise-disk world, devices are loosely classified by their interfaces. SATA disks with low rotational speed and maximum capacity are used for less demanding “nearline” applications and much more expensive SAS disks are used for the more demanding “online” applications. The cost differences are not really related to the SATA or SAS interfaces, but to the differing rotational speeds, qualities of the mechanical engineering, device monitoring probes and firmware, automatic recovery firmware etc. incorporated into each class of drive. The market for the most expensive disk capacity, currently provided by 2.5 inch 15k rpm drives used primarily for database applications, is already threatened by solid-state storage with much lower latency.

The technological problems faced by magnetic disk are easy to understand. The area of the bits written to current 4 terabyte drives is close to the limit of magnetic stability. The density could be raised if more atoms could be involved by writing bits that used more of the magnetic material below the surface of the disk platter, and/or by using a higher coercivity material that would allow smaller bit sizes. Technological developments along these directions have been in the works for some years, but according to industry experts, none will be advanced enough to bring the next jump in disk capacity to market much before the end of the decade. The prime "use-a-larger-volume-of-magnetic material" approach is shingled recording, conceptually illustrated in Figure bb. In this approach a specially shaped write head lays down tracks with a triangular cross section, which become a set of "shingles" as successive tracks are written. Each shingle goes deep into the material, so can be stable even if it is much narrower than current tracks. It is impossible to re-write sectors or single tracks with this approach. To change any information on the disk, a large block of tracks must be erased and re-written. HEP would have no problem with such a device, but its success in the wider marketplace could be problematic. The "high-coercivity" approaches include Heat-Assisted Magnetic Recording (HAMR) using a uniform surface with higher coercivity and "bit patterned recording" using lithography to lay down a circular pattern of bits that are close together but physically isolated from each other.

The next ten years of disk evolution is thus highly uncertain. It would be prudent and probably correct to assume that the doubling time is now around four years. The impact of this will be considerable. The logic that made it cost-effective to replace storage after as few as three years (approximately twice the doubling time) might now argue for keeping disks eight years or more, if the disks were sufficiently reliable. Disks may also last longer if most of the access traffic can be fielded by a substantial solid-state storage cache as part of the overall storage hierarchy. There will be strong market pressures favoring, long lasting, low power, physically dense, and therefore very heavy disk systems.

Solid-State Storage

The certainty of an increasing role for solid-state storage was suggested by the twenty-year failure of rotating disk to provide more accesses per second per dollar. In terms of this metric, today's solid-state storage provides vastly better value than rotating disk. The solid state storage being used in (exploratory) production at HEP computing centers costs about ten times as much per unit capacity as rotating disk. Thus reducing a rotating disk purchase by 10% and spending the money on solid-state storage is cost neutral and can improve physics analysis throughput.

In this mode, solid-state storage is certain to play an increasing but not dominant role in HEP, and perhaps an even larger role in the commercial world. The woes of rotating disk described above are likely to allow solid-state – specifically flash-memory-based – storage to lower its cost relative to disk and gain market share. It seems plausible that the cost differential would go down from around a factor ten now to around a factor 3 within ten years. Industry experts do not expect current solid-state technologies to "kill" disk in the foreseeable future, not least because it is almost inconceivable to create the large number of multi-billion dollar "fabs" – chip fabrication facilities – that would be needed to displace the world's many exabytes of disk.

Storage Middleware and Data Management

With the notable exception of the adoption by several HEP sites of the semi-commercial HPSS (High Performance Storage System) to manage tape data, single-site storage middleware in HEP has been entirely home made. In the late 1990s, it seemed that it would be an achievement to limit HEP's invention of tape data management systems to only one per lab.

HEP's middleware stacks have proved quite successful, for example the seamless integration for BaBar at SLAC of petabyte (HPSS) tape storage and tens of terabytes of disks distributed over 50 servers, but these successes have not translated into wide HEP adoption, let alone interest by other sciences or industry.

In the wider scientific world of middleware such as Globus Online (GO), which is a data transfer mechanism that has nice retry and a graphical display, is currently the norm. These tools continue to grow and change, but are still based on the gridftp protocol. There are other tools for data transfer, but it still has to be managed by the application. There are alsotools that are emerging that transfer data at the file system level.

There are scalable file systems such as Lustre and GPFS that are connecting the file system with the archive or nearline environment so there is seamless access, from the users prospective, to all files. The data-managed file system means that the inode remains in the file system with a stub of the first blocks of the data, or maybe nothing of the file at all, and the I/O is captured until the data returns to the local disk of the file system server. The data can be recalled from any ftp archive as long as permissions and security have been prearranged. This builds basically an endless file system. Connecting these large endless file systems is then next on the horizon.

There are some test sites in the next few years that will be using GO-Storage to connect these large managed file systems together, so the meta-data is separate and in a replicated environment and the data could be anywhere in the world. As data is retrieved from locations, it moves closer and locally depending on the applications requirements. These methods all have small latencies added and depending on the data actually location, the latency could be great if it's on tape somewhere. If these efforts yield systems and services that are widely adopted by science they must become serious candidates for meeting HEP needs in the future.

Even so, at the current time there are no commercial or widely used open-source offerings meeting the HEP needs for worldwide data and workflow management. The absence of commercial offerings is probably largely due to HEP's need to integrate tens to hundreds of autonomous computer centers, having many different funding sources, many different technologies, and a varying level of affiliation, from loose to none at all, with any particular HEP project. Multinational commercial entities are generally orders of magnitude more coherent in their technologies and management. The wider world of collaborative science does face many of the HEP challenges, but HEP is still at the bleeding/leading edge in terms of overall complexity.

The commercial world of distributed data-intensive applications should not be ignored – individual technologies and de-facto standards relevant to HEP are almost certain to appear. The wide scientific world must also not be ignored. HEP will be able to bring benefit, reap credit and attract non-HEP funding by helping adapt some of its most successful distributed computing technologies for other sciences and smaller HEP activities. Beyond this, HEP can benefit from participating in the development of widely useful distributed computing tools, and should studiously avoid the arrogance that could postpone the adoption of tools developed outside HEP to serve the exponential growth of data-intensive science.

A major stimulus to obey the exhortations above must be the labor intensive aspects, in both ongoing development and operations, of the distributed computing systems used in HEP today.

Data, Software and Physics Preservation

HEP Outlook and Recommendations

Impact of the Technology Outlook on HEP

The coming decade and beyond will see experiments and observations at several frontiers straining against the limits of data management and storage technology. Storage technology is likely to evolve in capacity/cost more slowly than in the last decades, making it ever more important that the role of storage is carefully optimized with respect to other costly elements of the scientific programs. Figure dd gives attempts to illustrate this challenge by comparing the data rates expected from the most challenging activities with the disk technology evolution predicted earlier in this report.

As outlined earlier, the rate of writing raw events to persistent storage for the LHC experiments is substantially influenced by the cost of storage and can be expected to evolve at a similar rate. It is likely that other activities with petabyte/s rates for raw data from front-end devices will be compelled to employ real-time data processing and selective rejection in order to reach persistent rates that are affordable in the context of each activity.

At a more detailed level, the roles of tape, rotating disk and solid-state storage will evolve:

- Tape will fall slightly in relative cost should and play a more important role than, for example, in the LHC experiments today where it is arguably underused.
- Rotating Disk, will not improve in capacity/cost as rapidly as tape or solid-state storage and is for the next decade may take around four years to double capacity/cost. As a result, it will be cost-effective to wait about eight years before replacing disks, provided the space, power and cooling are available to make this possible.
- Solid-state storage will not kill rotating disk in the enterprise and data-intensive science markets, but it will become relatively cheaper, perhaps by a factor of three, and will merit full consideration in the overall optimization of HEP computing environments, where it offers the promise of efficient sparse or random access to HEP data. To make effective

use of the still small affordable quantities of solid state storage, application-aware approaches to caching data on such storage appear essential.

The relative rates of cost evolution of tape, rotating disk, solid-state storage, networks and CPU are clearly uncertain. However, it is certain that HEP computing models will need to adapt to make the most efficient use of all these elements. The best way to adapt would be to make the implementation of the models intrinsically adjustable to adapt to a wide range of situations. The now old, but never really implemented, concept of "virtual data" would go a long way to making HEP computing dynamically adaptable to wide range of relative storage and CPU costs. In a virtual data system, all derived (or simulated) data products begin existence solely as the rigorously complete recipes for creating them. These virtual data products are then instantiated or replicated based on real or algorithmically anticipated demand. The physical instances are retained based on anticipated future use and the relative cost of storage versus re-creation. HEP has often appeared to be close to developing the rigorous provenance databases needed to implement virtual data and a full virtual data implementation may be in reach of the LHC experiments for Run 3.

Findings and Recommendations

CpF15 Finding 1a: The largest HEP experiments have developed, and are improving functional distributed data and workflow management systems meeting their needs. These systems are expensive to develop and operate and are thus rarely appropriate for smaller experiments.

CpF15 Finding 1b: HEP currently benefits from, but can also be constrained by, the highly successful ROOT features supporting reading and writing of persistent data. No other major scientific field uses ROOT or appears interested in it. Major developments in persistency technology will be required to take advantage of storage hardware on the timescale of LHC Run 3.

CpF15 Recommendation 1: HEP should maintain and promote a vision of the future in which fully functional and low-operational-cost distributed computing and persistency management is supported by software that is widely used in data-intensive science. To this end, developments in industry and the wider science community should be monitored actively, HEP should work with the wider science and computer science community to export and adapt HEP technologies and vice-versa. In distributed computing, HEP should organize itself to significantly reduce the number of diverse approaches and provide the benefits of ideas and software developed in the largest experiments to other activities where they are needed.

CpF15 Finding 2a: Rotating disk storage will suffer a marked slowdown in the evolution of capacity/cost. This may be the largest perturbation of HEP computing models that must attempt to optimize the roles of tape, rotating disk, solid-state storage, networking and CPU.

CpF15: Finding 2b: Many of the components required to support virtual data already exist in the data and workflow management software of the largest experiments. The rigorous provenance recording required to support the virtual data concept would also benefit data preservation.

CpF15 Recommendation 2: Computing model implementations should be flexible enough to adapt to a wide range of relative costs of the key elements of HEP computing. In preparing for Run 3, the LHC program should seriously consider virtual data as a way to accommodate scenarios where storage for derived and simulated data becomes relatively very costly.