FERMILAB-CONF-04-531-CD

Computing in High Energy Physics

Ian M. Fisk
Computing Division, Fermilab, PO Box 500
Batavia, Illinois 60510, USA

Received (7 12 2004)
Revised (Day Month Year)

In this review, the computing challenges facing the current and next generation of high energy physics experiments will be discussed. High energy physics computing represents an interesting infrastructure challenge as the use of large-scale commodity computing clusters has increased. The causes and ramifications of these infrastructure challenges will be outlined. Increasing requirements, limited physical infrastructure at computing facilities, and limited budgets have driven many experiments to deploy distributed computing solutions to meet the growing computing needs for analysis reconstruction, and simulation. The current generation of experiments have developed and integrated a number of solutions to facilitate distributed computing. The current work of the running experiments gives an insight into the challenges that will be faced by the next generation of experiments and the infrastructure that will be needed.

Keywords: computing

1. Introduction

When starting to write a review of computing in High Energy Physics (HEP), I intended to address the challenges that are facing the next generation of experiments under construction at the Large Hadron Collider (LHC) and only summarize the progress made by running experiments of the Tevatron and the B-Factories. When we typically discuss the next generation of experiments we emphasize the uniqueness and the new aspects of the computing challenge: the scale, the complexity, the level of distribution. As I researched the computing achievements and plans of the current generation of experiments, I realized the challenges and scale facing the LHC experiments are similar to the those being faced by running experiments.

When the first LHC experiment computing models were proposed in the middle of the 1990's they had many revolutionary aspects. The level of distribution in hierarchical computing tiers, the processing and storage capacity required, and the network connectivity were all proposed at scales that had never been attempted in the field before. After eight years of improvements in technology and increasing requirements from existing experiments, the revolutionary aspects appear to be more common place.

The reconstruction and analysis of the data are the last of many steps involved in realizing an experiment's physics potential and require ample computing resources. Computing is a costly endeavor, but without sufficient computing resources it is not possible to fully capitalize on, the much more costly, accelerator and detector investments.

2. Common Infrastructure Challenges

In preparing this summary I spoke to computing representatives from CDF, DZero, Babar, and STAR. In each case the experiments were facing similar infrastructure challenges to delivering computing resources at the facility level. The accelerating increase of data volumes and the corresponding increase in computing requirements are taxing. The three primary elements of a computing center's physical infrastructure are processing, storage and networking. The budgets for these have become a sizable fraction of the total operations budget of an experiment.

2.1. *Processing Resources*

The need to deliver substantial computing resources at a moderate cost has driven nearly all experiments toward the use of commodity computing solutions. These modular cluster-computing solutions have allowed computing centers to smoothly grow to resource capacities that would have been impossible in the past. While these computing solutions have permitted explosive growth, recently they have not improved as quickly as expected. In addition, operating these large quantities of systems has dramatically increased the computing center requirements for power and cooling.

Moore's law [1] continues roughly to hold as predicted. The number of transistors in each successive generation of central processing units (CPU) has doubled on approximately an eighteen to twenty-four month time scale. Unfortunately the number of transistors does not translate directly into affective computing resources for HEP. One benchmark that has generally accurately predicted the performance of a CPU on high energy physics application code is the integer component of the CPU benchmark from the Standard Performance Evaluation Corporation (SpecInt). [3]. The CPU benchmark is updated and released every few years, and the current version used for measuring resource requirements in high energy physics is the the 2000 version (SpecInt2k).

Using the SpecInt2k benchmark to compare CPUs commonly used in HEP, the performance has not doubled recently on an eighteen month time scale. Every three years, CERN publishes a document called the PASTA report [4], which provides predictions of major computing component evolution. The most recent release was the 2002 report. After two previous versions of very conservative predictions, the processor performance available in 2004 lags significantly behind the 2002 predictions for 2004. The actual performance of processors has not been doubling every eighteen months.

The decrease in CPU performance as compared to predictions has increased the total number of CPUs needed to satisfy the computing requirements. The power and cooling requirements of the CPUs have steadily increased. In 2001 the Silicon Industry Association (SIA) published predictions for silicon devices [2]. The SIA predicted that by 2004 the maximum power per CPU would be greater than 160W. Fortunately for physics, the power required by commodity chips has generally been lower than the prediction. The commonly used systems from Intel and AMD are below 100W and the less commonly used 64bit Itanium systems are below the prediction from SIA.

Even with the better outlook for CPU power usage, the power used by a rack of computing systems can easily require 6 to 8kW of power and cooling. Many HEP computing facilities were designed for large Symmetric Multi Processor (SMP) systems that had a much lower power density. The large HEP computing Facilities have approached the problem of how handle the increase in power density in a variety of innovative approaches. CERN has increased the power available in the computing center and switched to using lower density tower cases for worker nodes. Fermilab has converted and experiment hall to computing space with power and cooling designed for high density, rack mounted systems. GridKa has moved to water-cooled racks that increase the initial cost, but cool much more efficiently. SLAC has purchased some blade servers for Babar. The cost per node in this type of installation is higher, but the power consumption and cooling efficiency is improved.

The increasing requirements of the experiments, combined with a slower than expected improvement in CPU performance, place a substantial burden on computing facilities to afford and house the computing required. Purchasing additional infrastructure requires the availability of scarce financial resources. Hosting computing systems requires physical infrastructure, which is difficult to increase on short time scales even at substantial cost.

2.2. *Storage Resources*

Storage providers are continually asked to increase the scale of the data archiving and data serving services in response to the growing data sets collected by collaborations. The two primary responsibilities of the storage community are to archive and safeguard the data and to serve the data to the collaboration efficiently.

Several sites are now successfully running large mass storage systems to archive data at the several petabyte scale. From a technical standpoint the problem is reasonably well understood, tape drive and robot capacity continues to improve largely as has been predicted. Operating facilities of this size requires substantial effort and commitment, but there are no technical hurdles that should jeopardize archiving data for the next generation of experiments at the LHC. In preparation for the new experiments, CERN has successfully ingested 1GB/s of data during the Alice data challenge [5].

Allowing the data stored on tape to be accessible for applications is much more

difficult. Enabling large quantities of data stored on tape media to be efficiently accessed for analysis is a persistent computing challenge. The increasing scale of the data vaults and the input output needs of the applications have ensured that data serving solutions are evolving. The field currently appears to be consolidating on two solutions: CASTOR [6] from CERN and dCache [7] from Fermilab and DESY.

2.3. *Networking Resources*

The final basic building block of computing infrastructure is networking. The storage and processing resources need to be connected both locally at computing centers, and more recently over wide areas to facilitate distributed computing. On both the local and wide area networks, the network infrastructure is an area that has seen improvement recently. Newer network connections for individual local resources are commonly gigabit Ethernet and the infrastructure to enable this has steadily decreased in cost. Large scale switches that allow hundreds of systems to be connected at wire speed simultaneously over high performance backplanes are available. The trunks that connect clusters of local resources are migrating to 10Gb/s fiber links and the cost has decreased substantially since their introduction.

In wide area networking, the over capacity of wide area networking infrastructure in the United States and elsewhere has lead to availability of high performance network connections at more modest costs. Many sites have single or multi-gigabit production up-links and single and multiple ten gigabit research links are becoming more common.

3. The Push Toward Distributed Computing

Enabling sufficient computing requires commitment, sufficient financial resources, and sufficient infrastructure resources (power, cooling, and space). Not all of these can always be found at the same physical location. The need to provide for increasing requirements of experimental collaborations with larger datasets and fixed budgets, combined with the infrastructure challenges associated with building large processing clusters made of power hungry commodity computers, and the improvements in wide area networking have pushed collaborations and computing providers to examine distributed computing solutions. The benefits of distributed computing are that it allows the support and infrastructure for computing resources to be shared over sites and facilities resources can be leveraged. The cost is paid in complexity.

Once the computing is spread over multiple locations, data management and data accessibility become more complicated. Experiments need to efficiently replicate and synchronize data across multiple sites. The existence of multiple official copies of the data requires good version managing solutions and solutions for distributed meta-data. The data, the processing resources and the users are distributed, so the problem of appropriately matching resources becomes more interesting.

Another issue is how to create reliable working environments at all locations. In the absence of a completely common working environment, the experiment needs

tools to ensure results created at distributed locations are reliable and reproducible. In addition to the software and processing environment, the experiment needs to replicate databases for catalogs, conditions, and triggers. This opens the interesting and complicated field of high performance distributed databases.

Once the data, the resources, and the environment are prepared, the computing must be made accessible to users in a scalable way. Interfaces to the system are needed that allow generic grid type submission or individual interactive logins. Infrastructure needs to be created to manage large organizations of users so that access can be granted at multiple sites, while enforcing and accommodating the local access and usage rules.

3.1. *Progress in Distributed Computing*

In preparing this summary of computing in high energy physics I performed an unscientific survey of running experiments to assess why they had considered distributed computing solutions, if they had, and how much progress they had made implementing those solutions. The most common reason quoted for desiring to implement a distributed computing solution was the need to deploy resources locally to obtain funding. All the collaborations I contacted had international participation; in order to obtain computing resources the facilities had to be hosted in the country providing the resources. Another response was that the host lab priorities for expending valuable infrastructure resources did not completely align with the experiment requests, or that the host lab lacked additional infrastructure resources.

Considerable progress on adopting distributed computing solutions has been made in many running collaborations. In the Fermilab Run2 experiments, the CDF collaboration now has 25% of their computing resources located off-site, they expect to ramp to 50% in the coming year. They have deployed two satellite computing facilities in the US as well as centers in Italy, Taiwan, Korea, and Canada. These resources are predominantly used for Monte Carlo generation, but analysis use is increasing. The other Run2 experiment DZero is already at 50% use of off-site computing resources. Monte Carlo simulation is produced off-site, and the DZero data management system, SAM [8,9], allows reasonable remote access to data from 40 installed SAM stations.

The B-factories have also made substantial progress. Babar currently has approximately 50% of the computing resources located off-site at 4 large centers in Europe. Babar has gone as far as now performing all primary event reconstruction away from the host lab. Raw events are shipped to Padova, where they are reconstructed and replicated to remote centers. Analysis is performed at all remote facilities. The Belle experiment currently performs 25% of the simulated event production off-site. Belle also benefits from excellent networking between the experiment and regional universities, which allows more transparent distributed analysis.

The STAR experiment located at Brookhaven (BNL) has moved to a evenly distributed computing system with 50% resources at Brookhaven and 50% of the

resources located at Laurence Berkeley National Lab (LBNL). STAR was unique among the running experiments I spoke that they had moved to a model where computing was provided by two equal peers. Even in the other experiments that had moved to 50% off-site resources, the host lab was still the single largest center.

4. Basic Work Flow for Distributed Computing

The technical challenges facing collaborations deploying distributed computing infrastructure have been approached in a variety of ways from very basic and functional to complicated and visionary. To describe some of the progress in these technical challenges, one needs the basic elements of a distributed computing work-flow. The basic elements are managing data, replicating data to remote sites, serving data to applications, and running applications.

4.1. *Manage Data*

The process of versioning, selecting, replicating and tracking the data distributed at remote locations has been approached in a variety of ways. Some experiments have chosen to replicate all data between centers, or to partition the data statically so describing the data available on a site can be done with web pages or catalogs implemented in databases. The most advanced data management system current in production is arguably the SAM [8,9] systems from the DZero collaboration.

SAM stands for Sequential Access via Meta Data. It is built around a large central Oracle database and it manages both the data and the meta-data in addition to managing the application access to the data. SAM manages file storage and replica catalogs. Data files are stored in tape systems at Fermilab and elsewhere and files are cached around the world for fast access. SAM also manages the file delivery, users at all sites retrieve files out of file storage and SAM handles caching for efficiency, so users do not need to know about physical file location. SAM manages file meta-data cataloging, so there is meta-data held for each file and users can make queries to determine files needed for analysis. Finally, SAM manages analysis bookkeeping; it remembers what files were run over and which were processed successfully.

4.2. *Data Replication*

Once data can be managed, selected, and tracked, tools are required to efficiently replicate the data between computing centers. Independent solutions have been adopted by each experiment and there is a lack of a common standard. Babar implemented a high performance multiple stream version of FTP called Babar FTP [11], which is used by Babar and was used by DZero in the past. DZero primarily now uses GridFTP [12], which is the grid standard for high performance parallel stream transfers. Fermilab developed a tool called FCP that allowed load balanced copying, which is in use at CDF. The experiments at the Large Hadron Collider (LHC) are adopting a common interface called Storage Resource Manager [10] (SRM).

SRM is also used by STAR to replicate data from BNL to LBNL. SRM is a management layer that sits on top of lower level transfer tools and provides a uniform interface to a diverse and growing set of storage systems. The SRM system sets up transfers and performs the authentication. It provides load balancing and traffic shaping to ensure that the resource is not over-burdened by trying to accommodate too many requests. Although many multiple underlying transfer protocols are supported by SRM, most installations are currently configured with GridFTP. Data integrity is ensured with end-to-end check sums and fault tolerance is provided by retries.

4.3. *Data Serving*

Once data has arrived on a site it needs to be made accessible to applications. At small sites this has been successfully done using NFS mounted file systems. DZero has been successful using cache space managed by SAM. The next generation of LHC experiments is currently looking CASTOR and dCache, which are solutions from CERN and Fermilab/DESY respectively. In both cases the data serving solutions provide a way of making a group of physical disk resources appear as a common filesystem with a defined interface to local mass storage and wide area data transfers. The filesystem provides a POSIX-like interface to allow locally running applications to access data, while the interface to mass storage allows archived data to be efficiently accessed and output results to be safely stored.

The dCache solution has been extensively used by the CDF experiment and the throughput rate achieved while serving multiple applications approaches what is expected to be required for LHC experiments. dCache provides load balancing for heavily accessed files by automatically creating replica copies. It also provides fault tolerance by staging files to alternative cache disks if a file server fails.

A relatively new data serving product is Xrootd, which was developed by SLAC for use in Babar. Xrootd has some interesting properties. Xrootd provides high performance file-based access that is scalable and extensible. Xrootd has fault tolerant features similar to what dCache is providing for server failures in a natural way and for dynamic adding and removing of servers. It has a flexible security model and is compatible with rootd.

4.4. *Running the Application*

Once the data has been managed, transferred and can be served to applications, the applications have to be prepared to receive it. A lot of effort has been expended in developing tools to ensure the software environment can be reproduced and information needed to consistently run applications is available at distributed facilities. The goal is to ensure that applications run at distributed centers have results which are as reliable as results run at the host lab. The quality of service and the tools for information flow need to be as good at all centers as they are at the host lab.

In addition, the quality of software environment packaging is critical to efficient distributed computing.

In addition to the basic software environment, experiments need to replicate large quantities of conditions and detector information typically installed in databases. There have been several attempts at using database technologies to perform database replications and synchronizations with varying degrees of success and performance. Recently a new product called FroNtier [13] was developed at Fermilab to use web server technology to replicate databases. This technology and technique is also being looked at by some LHC experiments.

FroNtier was developed within CDF to replicate read-only databases to remote or local facilities. It works by caching complete SQL queries in web server caches and serving the results to applications as needed. The system is capable of providing load balancing and fail-over fault protection. The advantages of the system are that it uses open source software that is well tested using web server technology to provide database information in a distributed way. The caches can be placed at the remote center, close to the running applications. The initial performance tests are very promising. Database queries from a Oracle server at Fermilab were cached at a remote CDF center at the University of California, San Diego.

Providing the tools to allow a distributed group of users to easily and securely run applications on a distributed set of resources is also a computing challenge. It is possible to continue to provide interactive logins for all participants at all sites, as was done in Babar. Although at some point this is operationally intensive and ceases to scale as the number of users and distributed sites grows large. More advanced approaches include the use of distributed batch queues and the use of grid interfaces to local batch queues.

CDF has deployed a particularly interesting solution on the CDF Analysis Facility [14] (CAF). The CAF system provides a consistent interface to local and remote facilities, and uses the Fermilab kerberos infrastructure for authentication and virtual organization management. The CAF usage Model is that users develop and debug on their local desktops, and then the applications are packaged as a sand box to be submitted to a remote cluster. The interface at run time combines elements of a traditional batch queue with elements from an interactive environment. The users can submit and kill the remote process, but they also have the ability to tail log files and attach a debugging session to a running process. The CAF monitoring maintains full information about cluster status and utilization history for hardware and users.

5. New Challenges at the LHC

The Large Hadron Collider is expected to begin taking in the fall of 2007 and brings new challenges to high energy physics. Scientifically, the enormous increase in energy and luminosity provides the potential for important discoveries early in the program. From a community perspective the size of the collaborations represents

a new challenge. The largest LHC collaborations are already at approximately 2000 people. From the computing perspective, the LHC represents new technical and infrastructure challenges.

On a computing technical level, at the LHC the beams collide more frequently, which leads to more data. The experiments expect to record petabytes of raw and reconstructed data each year. The detectors have more channels and the events are subsequently larger. The luminosity is higher so there are more collisions per beam crossing, which leads to larger and more complicated events. From an infrastructure standpoint, the LHC computing is planned from the beginning to be completely distributed. In addition, most LHC computing centers will support more than one experiment.

5.1. *Comparison of Technical Challenges*

There are technical challenges at the LHC due to the increase and energy and luminosity, but they are not insurmountable. As an example one should compare the technical challenges facing a running experiment like CDF and the planned operations of a large LHC experiment like CMS. In 2004 the CDF data logging rate was about 20MB/s, while the planned rate for CMS is 100MB/s at the start of the experiment. This seems substantial, but CDF has proposed luminosity and trigger upgrades that are expected to boost their logging rate to 60MB/s by 2007. The CMS rate is still larger, but the difference is much smaller. The same holds true for other interesting computing quantities. If we compare the 2004 CDF number for processing resources or for storage capacity, the CMS numbers look like a computing revolution. If we compare the expected CDF capacity at the start of the CMS experiment, the results are much more of an evolution.

One technical area that is a significant change of method is the level of distribution expected in the computing centers. CMS and ATLAS have both proposed a series of computing Tiers.

- The largest single center is the Tier-0 at CERN, where primary reconstruction occurs and there is analysis capacity for a small fraction of the collaboration. This represents approximately one-third of the total computing capacity.
- The next third is located at the sum of 5 or 6 Tier-1 Centers at national labs and universities.l These centers are responsible for re-reconstruction of data and they provide the bulk of capacity for user analysis.
- The final third of the computing capacity it distributed across 25 Tier-2 Centers at universities. These centers provide the Monte Carlo production capacity for the experiment and significant resources for user analysis.

This tiered distribution of resources has some interesting implications. For the first time the host lab is a comparatively small percentage of the total computing resources. There will not be enough analysis resources in any one place to complete

the scientific program of the experiment even at the beginning of data taking. For the experiments to be successful and competitive, the data management, data distribution, environments, and access to distributed resources all has to work properly at the beginning. Previous experiments have used distributed resources as a way of augmenting the program. At the LHC they are critical to the success from the start. Service level at remote centers needs to meet or exceed the service level at the host lab. The computing for the experiments will be provided by a collection of peers.

Another interesting aspect of the LHC computing is that most of the remote computing facilities plan to support multiple LHC experiments. In some cases, LHC experiments, Run 2 experiments, and B factory experiments will be supported simultaneously at the same center. With the current generation of experiments, supporting two experiments involves supporting two complete and often independent computing environments. The services required and techniques used are often created and deployed by the experiments to solve immediate problems. In order to share a computing center resources are frequently partitioned, which makes less efficient use of resources and places a high operational load on the support people.

There is a desire from the computing centers to support a common set of services that support a number of experiments. This should be able to improve the quality of service by reducing the operational load. In addition the facility usage can be governed by dynamic policy and allow for better opportunistic use of computing resources. There is still significant development efforts needed on policy, auditing and accounting. It is important for the facilities to be able to demonstrate that they are meeting their obligations to the supported experiments, when the resources are not partitioned and therefore not as easily identified.

5.2. *The Need for Grid Projects*

The need for a consistent set of services and interfaces to enable distributed computing can driven the need for large scale grid computing projects. These projects have been well funded both in the United States and in Europe and they have resulted in a large number of promising prototypes. Two promising large scale systems are the LHC Computing Grid [15] (LCG) in Europe and Grid2003 [16] and the developing Open Science Grid [17] (OSG) in the U.S. The Grid projects are developing and deploying a consistent and scalable set of services. The services range from low level batch queuing system adapters and interfaces to transfer data, up to high level resource brokers and virtual data systems.

The Grid projects started by deploying fairly low level Globus services to provide consistent interfaces to batch queues through Globus GRAM, a consistent and authenticated file transfer interface with GridFTP, and a basic information publishing tool Globus' Monitoring and Discovery Service (MDS). These were deployed to successfully build systems to handle the computing intensive, but relatively predictable, problem of producing large sample of simulated events.

The Grid prototypes have evolved to deploy higher level functionality. SRM has been chosen as a interface to manage storage transfers. Global file catalogs and replica location services have been deployed. There have been significant improvements in the quantity of data that is monitored and common schema have been agreed to for information providers. Resource brokers now interpret the information provided and can make more intelligent scheduling decisions.

As the scale and complexity increases the experiments and Grid projects are re-examination of architecture and services used. Self describing and discover-able web services are being implemented and new architectural models that improve scaling are being examined.

6. Summary

Computing in High Energy Physics continues to be an interesting and challenging activity. The volume of data collected and complexity of the working environment push the facilities to provide more capacity and more services. Facility limitations and the speed of improvement in underlying technology are complications to progress. By deploying innovative approaches like distributed grid computing, the HEP computing continues to meet the requirements of the experiments.

The next generation of HEP experiments represents new challenges, but they are not unique challenges. The running experiments and the next generation of LHC experiments are both facing similar technical hurdles. Neither group has a complete solution and the common needs of both groups represent an opportunity to work for mutual benefit.

References

1. Moore's Law, 2003; http://www.intel.com/research/silicon/mooreslaw.htm.
2. Semiconductor Industry Association et al., "The International Technology Roadmap for Semiconductors", http://www.itrs.net, 2001.
3. Standard Performance Evaluation Corporation, "SPEC CPU 2000 Results", http://www.spec.org. Accessed 2004.
4. I. Fisk et al, "Working Group A Report", http://david.home.cern.ch/david/pasta/pasta2002Report.htm, Accessed 2004.
5. CERN Bulletin. http://bulletin.cern.ch/eng/articles.php?bullno=13/2003&base=art&artno=BUL-NA-2003-031, number 13/20, 2003.
6. CERN CASTOR Home Page, http://castor.web.cern.ch/castor, Accessed 2004.
7. dCache Home Page, http://www.dcache.org, Accessed 2004.
8. I. Terekov, et al. "Meta Computing at D0"; in Nuclear Instruments and Methods in Physics Research, Section A, NIMA14225, vol. 502/2-3 pp 402-406.
9. I. Terekov, et al. "Distributed Data Access and Resource Management in the D0 SAM System", in Proceedings of the 10th IEEE International Symposium on the High Performance Distributed Computing (HPDC-10), San Francisco, California.
10. A. Shoshani, et al. "Storage Resource Managers: Middleware Components for Grid Storage", Laurence Berkeley National Laboratory Berkeley California, http://www.lbl.gov/ arie/papers/srm.mss02.pdf, Accessed 2004.

11. bbftp Home Page, http://doc.in2p3.fr/bbftp, Accessed 2004.
12. GridFTP Home Page, http://www.globus.org/datagrid/gridftp.html, Accessed 2004.
13. S. Kosyakov, et al. "FroNtier: High Performance Database Access Using Standard Web Components in a Scalable Multi-Tier Architecture", in Proceedings of Computing in High Energy Physics (CHEP) 2004, Interlakken, Switzerland.
14. T.H. Kim et. all., "The CDF Analysis Farm", IEEE NSS 2003 proceedings.
15. LHC Computing Grid Home Page, http://lcg.web.cern.ch/LCG/, Accessed 2004.
16. Grid3 Project Home Page, http://www.ivdgl.org/grid3/, Accessed 2004.
17. Open Science Grid Home Page, http://www.opensciencegrid.org, Accessed 2004.