THE LHC COMPUTING MODEL

M. Mazzucato
INFN, Italy - Via Marzolo 8 - 35131 Padova

Abstract

The needs of the new experiments in preparation for the Large Hadron Collider (LHC) have imposed a true revolution to the High Energy Physics Community (HEP) in the way the software is produced and in the way the computing is organized. The HEP software community has now fully entered in a transition era from the Fortran to the OO and C++ paradigm and many lectures have presented in detail in this school the key aspects of these new software methodologies aiming to improve the quality of the HEP software system. The computational and data access issues are extremely challenging and are addressed by this lecture.

The next generation of LHC experiments will produce 1 Petabyte of raw data per year for each experiment, several orders of magnitude more than the ones presently running, and these must be analysed and made available to hundreds of physicists spread all over the world. This is certainly one of the key problems that the LHC experiment have to address to fulfil their task. The new computing models based on flexible architectures able to respond to the new technological developments in the processing power, storage and network and to the availability of resources will be reviewed.

1 INTRODUCTION

At present three collaborations, ATLAS, CMS and ALICE, have an approved technical proposals to build an experiment for LHC, while LHC-B is expected to submit a fourth one at the beginning of 1998.

Recently the two collaborations, most advanced in the preparation, ATLAS and CMS, have issued a Computing Technical Proposal CTP) [1] [2] where they present their computing requirements and introduce the main principles of the Computing Models for the LHC era and for the transition phase. These describe, for the moment mainly at architectural level, how they intend to manage the resources that will become available to perform the LHC computing. Both hardware and software aspects are concerned.

The CTP's are intended to be evolutionary documents that will be regularly updated until the presentation of the final Technical Design Reports (TDR), few years before the turn-on of LHC, which will contain the final design choices of each collaboration and the specific implementation details, on the light of the technological advances available at that time.

The LHC computing model can be subdivided into two large components, one being the online Event Filters necessary for a successful acquisition of the interesting physics events and the other concerning the offline event reconstruction, with the related activities of detector calibration and alignment, the Monte Carlo simulation and the physics analysis of both real and simulated data. It is believed that the Event Filters will not put strong constraint on the LHC computing models so the DAQ requirements will be only shortly mentioned in this report dealing mainly with the offline computing.

2 LHC requirements

Contrary to LEP, the present electron-positron collider in operation at CERN, where each event represents a basic interaction of two point like constituents, at LHC the interesting physics interactions have to be selected by an enormous amount of softer background reactions. At the design luminosity L =

$10^{34}\ cm^{-2}sec^{-1}$ the two proton beams collide in correspondence of each experiment with a frequency of 40 MHz and produce 8 x 10^8 interactions/second.

To cope with the intrinsic limitations of the commercial data writing systems and with the constraints deriving from the total data volume and the offline analysis the maximum output event rate for the DAQ has been limited to no more than 100 Hz. In this section the computing requirements needed for this task and the subsequent offline analysis will be reviewed.

2.1 Trigger requirements

Powerful trigger schemes have been designed to provide the necessary overall online event rejection factor of about 10^6. This is achieved in three steps.

The Level-1 Trigger system is completely hardware based and provide a reduction factor of about 500 within 2 μsec.

The Level-2 and Level-3 are based on computing farms, providing Event Filters of increasing complexity in the software algorithms, with rejection factors of 50–100 and 1000–2000 respectively. Level-2 takes a decision in few milliseconds and Level-3 can use up to a tenth of a second per event and assigns already each event to the different physics streams using possibly the offline algorithms. Since errors in the code used in the trigger will cause not retrievable losses, the software for LHC must be of unprecedented high quality. This is felt as a very critical point for the success of the experiments and the collaborations plan to use the best engineering tools available and modern programming languages and methodologies to achieve the necessary robustness and reliability.

2.2 CPU and Storage requirements

The basics assumptions used to derive the requirements for the computing models are:

- Event rate out of Level-3 trigger : 100 Hz
- Event raw data size : 1MB
- Total number of events per year : $10^{\ 9}$
- MC event size : 2MB

Both collaborations assume for the output of the reconstruction an event size of 100 KByte/event, for the physics analysis data of 10 Kbyte/event and for the the event tags of 0.2 KByte/event. The storage volume is dominated by raw data.

It is assumed that the number of MC events to be generated is approximately 10 % of the real events one.

The resulting requirements for CPU and storage are resumed in the following table.

	ATLAS	CMS
LEVEL 3 RATE (TO STORAGE)	100 MB/sec	100 MB/sec
LEVEL 2+3 CPU POWER *	4 X 10^4 SPECint95	12 X 10^4 SPECint95
EVENT RECONSTRUCTION *	7 X 10^4 SPECint95	5 x 10^4 SPECint95
PHYSICS ANALYSIS *	15 X 10^4 SPECint95	20 X 10^4 SPECint95
MC PRODUCTION *	5 X 10^4 SPECint95	largely in background
DATA PRODUCTION	1 TB/DAY	1 TB/DAY
TOTAL DATA VOLUME	1 PB/YEAR	1 PB/YEAR
MC DATA VOLUME	200 TB/YEAR	200 TB/YEAR

1 SPECint95 = 40 SPECint92 = 40 MIPS

$*$ uncertainty factors are large !!

2.3 Networking requirements

One of the key requirements for the physics analysis is that any physicist from a collaboration institute should be able to query the data set relevant for his analysis and retrieve the results on his home desktop.

He should also be allowed to effectively collaborate with co-workers dispersed around the world to reach his physics goal.

ATLAS has made an initial attempt to estimate the need of network bandwidth for each individual user for this task, including the usage of collaborative tools such as video-conferencing. Here are the results :

- Bandwidth/user in year 2005 : 2 Mbit/Sec
- Total bandwidth for ATLAS : 1000 Mbit/Sec (500 simultaneous users)

3 Technology trends

The above requirements for the LHC experiments are at least two order of magnitude higher than for the current collider experiments . On the other hand the evolution of the technology has continuously provided in the past a systematic increase in the computing resources available for a fixed cost.

Technology tracking teams were set up by CERN/IT to predict the state of the technology at the turn-on of LHC and the main conclusions [3] [4] [5], summarized here, were used for the definition of the computing model. In the year 2005 the current chip fabrication facilities, built for a feature size of 35 μm, should be able to handle 0.13 μm. This will allow :

- individual memory chips of 4 Gbit (500 Mbyte) for a cost of 200 \$
- Processor CPU of$\geq$ 100 SPECint95 at the cost of today processors
- SMP parallel architectures (4-8 CPU/unit) as “standard”
- Magnetic disks at a cost 10 \$/GByte $\rightarrow$ 10M \$/PB (not counting new technologies)

- Tape storage stagnant at 2 M$ /PB
- Fast Ethernet (100 Mbit/sec) and Gbit Ethernet as candidate technology for LAN
- ATM, now slowly starting for WAN. Not clear if market cost for LAN desktops will be actractive

In conclusion it is reasonable to expect that the necessary CPU and storage for LHC will be obtainable within the foreseen budgets. Powerful computing farms based on commodity processor are producing very encouraging results now [6].

A key point of concern is the affordable networking bandwidth. Problem is cost,not technology : ATM links at 622 Mbit/sec are already available now. But the WAN cost has been only slowly decreasing in the past, due to absence of any market forces in Europe, and without the cut on prices, expected by the very recent opening of the commercial network market, could put severe constraints on the computing model. This has to be kept flexible to adapt to the reality of the year 2005.

4 Computing Model

The basic element of the computing models of ATLAS and CMS is constituted by an object store which will contain, following the OO paragdim, all the event data needed to perform the physics analysis or detector studies as persistent objects.

The commercial solution for the management of data objects capable to scale up to the PB level in a fully-distributed and heterogenous environment is provided by the Object DataBase Management Systems (ODBMS) that are under investigation by RD45 [7]. These are expected to provide transparent access to any stored object without knowing any details of its specific location. Many other interesting features of ODBMS are described in the lectures dedicated to them. The RD45 collaboration has shown that this solution can potentially scale to the requirements of LHC and has chosen one commercial product (Objectivity/DB) as the most suitable for LHC. RD45 is currently carrying on the work on ODBMS to prove full scalability,vendor independence and effectiveness for the physics analysis work.

Another key point under investigation is a Mass Storage System (MSS) capable of taking care, in a user transparent way, of the migration of the data from a rapid-access medium (magnetic disk) to low cost sequential media (robotic tapes)according to the usage. This solution is dictated by the expectation that the costs will not allow multi-PB's of data to be stored on disk even in the year 2005. Unfortunately,contrary to ODBMS, there seems not to be a commercial market for MSS systems up to the PB's size .At present only one product is available, HPSS,developed through a joint effort of IBM,US government and major US research laboratory, which is now on the way to be installed at CERN.

A major future task for RD45 is to address the issue of interoperability of ODBMS and HPSS and this is considered the major factor of risk for the ODBMS solution so it is followed with the necessary attention.

4.1 The traditional data flow

In this approach, used in present colliders as LEP, all events pass through a chain of subsequent processing :

- Reconstruction : produces all particle 4-vectors with relative detector points and shower hits
  - Output is DST, at Lep $\sim$ 30 Kbyte/event
- Short DST: produces all particle 4-vectors, particle identification, track elements and complete showers
  - Output is SHORT DST, at Lep $\sim$ 10 Kbyte/event

- MINI/N-tuples generation : produces selected events and physics quantities relevant for a specific analysis
  - Output are MINI or Ntuples, at Lep $\leq$ 1 Kbyte/event

At each step the data are subject to the reconstruction of more elaborated physics quantities and to the selection of the minimal information relevant for the physics analysis. No links are maintained between the different data sets.

The main drawback for this solution is that the general users access freely the data only at the SHORT DST level since at this stage the size is reduced enough to allow the storage on direct access magnetic disks. Raw data and other intermediate large data set are stored, as sequential files, on tapes. In this way, if analysis indicated the need for a more refined treatment for an event sample, e.g. events with high momentum electrons, requiring access to raw data, this was possible only passing through all the tapes where the interesting events were dispersed. The common practice was then to collect all the improvements for all detectors few times per year and then launch a general reprocessing of all events in order to minimize the very heavy tape mounting. In this way several months could pass before improvements were made available to users.

Moreover, due to the lack of pointers between data sets, if analysis selected events for data visualization on SHORT DST's or ntuples, the visual scan could only start after the corresponding raw data were extracted from tapes and the graphic information generated reprocessing the event.

It was clear that this was not an optimal solution but current experiments were forced to choose this strategy by the limitation of the available hardware and software technology.

4.2 The traditional data processing and data access

Real events are generally centrally processed and reduced data set, such as SHORT DST's, are stored in a central analysis facility, e.g. SHIFT, and made available to all users. Laboratories and Institutes create MINI DST's or ntuples via batch job on the central facility and import them for final analysis on local workstations.

Access to raw data is generally possible only through the central facility.

Only few major regional centres import copies of DST's, normally via shipment of physical tapes, in order to increase the number of queries they can satisfy using local resources. They are completely independent processing units and used only by regional users. The replication and management of the data sets in these centres and the experiment library update is entirely under the outside teams responsibility.

4.3 A new "extreme" computing model for LHC

At LHC the experiments are much larger and complex and a general reprocessing is much a bigger effort. Data distribution and access is much more difficult taking into account that,compared to the present ones, the number of physicists is multiplied by 4 and the number of events, the data volume and the offline CPU by 1000. For LHC four types of general data objects for all events have been proposed :

- Raw :$\sim$ 1 MB
- Reconstructed : $\sim$ 100 KB
- Analysis : $\sim$ 10 KB
- Event tag : $\sim$ 0.2 KB

The "extreme" model envisages that, at the limit, only the raw data and event tag will be permanently resident on the high performance event object store, at least for the whole period during which

the code and calibration files will be in evolution. All higher level objects (reconstructed and analysis or user objects) are created on demand, using the most updated version of the reconstruction code and calibration files.

At the same time the high performance object store contains all these higher level data objects corresponding to old raw data already migrated to the lower performance store.

Again all user objects are created on demand starting from existing general objects.

General (re)-processing are foreseen but in the sense that they create the objects needed by further classification and analysis of all events. Subsequent partial processing will be in general started on user demand.

How much this will be completely free and how much this should be coordinated remains to be studied.

A key issue is to understand how the different event objects should be clustered to optimize the response time to the most frequent user queries.

The ODBMS should manage dynamically the data migration from disks to tapes to keep in the most performing central store the most frequently accessed data. It should also take care of the dynamical replication of data to outside institutes according to the access frequency.

4.4 System architecture

A fundamental requirement of LHC experiments is that all the data reside in what appears as a single object database to the users.

It is expected that, thanks to ODBMS , this may be implemented by geographically distributed servers.

The architecture should be flexible and able to dynamically evolve with time from a CERN-tral model where all the data are stored at CERN and coherently centrally managed and accessed to a partially decentralized model where the database (or part) is replicated in regional centres having the necessary resources to analyse them .

This replication process should be transparent to the users and managed by the ODBMS.

The ideal architecture is the one where the physicist recognize the presence of a regional centre only by the reduced response time to their queries and is able to continuously exploit new resources (at least of sufficient scale), made available worldwide at any time to improve the overall analysis performance.

At the same time a regional centre could offer general services to the whole collaboration such as MC event generation and analysis.

This represents a drastic change from the past attitude and will require large efforts to integrate regional centres and central facility in a unique experiment computing system with optimized usage of all the resources.

In the partially decentralized model physicists outside CERN normally access data in their regional store cached from the CERN store. Analysis teams use collections of events cached from the regional store to a team store located in the physics departments. Individual stores contain data from queries to the team or regional store.

The model should allow maximal flexibility being capable of continuous decisions on object retrieval, recalculation and storage according to the available resources, taking in particular into account the network performance.

5 The generation of MC events

The generation of MC events has been traditionally provided for a large extent by outside institutes.

This is expected to be true also for the LHC experiments.

A clever and efficient way to manage the MC production is Funnel, the system developed originally by the Zeus experiment and subsequently adopted by other experiments,L3....

The idea behind is simple. The I/O request for Zeus MC generation is not very large,typically $\sim$ 1 KByte/sec/workstation (2.5 SPECint95). Normally each physicist has access to a personal workstation which is in general idle outside the normal working hours. Adding together all these workstation one can easily obtain all the CPU necessary for the experiment simulation.

Funnel allows to build in each site a farm of workstations connected in a LAN and, taking profit that each institutes has a network connection with ftp/telnet/rsh, establishes a remote and centralized control to minimize the management manpower.

The physics teams submit jobs with event generation request to a central manager in DESY (MCC). This forwards each input job to one of the site for processing. Here a para-manager program distributes events in parallel between the different processing units, handle exceptions and output the data via ftp or tapes via PTT or air freight to the physics team.

The system has proven to work very efficiently and minimize the management request: it is claimed that few-man hours/day are sufficient to follow the whole Zeus MC production ($\sim 10^{10}$ event/year).

It is certainly interesting to evaluate the possibility to incorporate the Funnel functionality in Geant4 [8], the new simulation framework being developed for LHC and for HEP in general following the OO paradigm.

6 A general event processing for LHC

In view of the success of Funnel it was proposed already at CHEP95 [9] to extend the underlying ideas to a general event processing system for HEP. In fact all steps of data analysis, MC production, event reconstruction and physics analysis, present large similarities. They consist of loops :

- Read input events
- Process them
- Write output

with additional access to calibration database or control files.

What is different are the I/O requirements and I/O streams may be very complicated and spread out over a large database.

Even more complicated is to handle exceptions, decide on how to proceed and report back the errors, clean all the effects of errors on archiving, bookkeeping, output streams etc ., a key point for the success !

A similar system is certainly in the direction of what is needed for LHC.

The LHC physicist dream is to submit a query with selected input objects, may be via a collective name, and the processing software, possibly easily built from the software computing environment, and have a system which take care of the rest :

- Create a processing system
- Manage computing resources (geographically dispersed world-wide)
- Take care of the security issues
- Process the data where they are in parallel

- Take care of disk space archiving and bookkeeping
- Handle and report back exceptions
- Follow and give back informations on the query status

The system should free physicists time from the details of the event processing, leave them more time for physics and minimize the management of computing resources.

Tools going in the direction mentioned above are being developed by many groups: Condor [10],Nile [11],Us teams [12].

It is clear however that this is a very challenging task and a coordinations of the efforts is highly desirable.

Simulation tools should be developed to evaluate the possible different implementation choices and optimize the model taking care of institutes and regional centres resources, LAN and WAN bandwidth and task specifications together with real life effects such as unavailabity of WAN links for some time or crashes of some computing units.

7 Collaboration at distance

Collaboration in a project has always implied very frequent meetings at CERN or other major laboratories, but with the growing of the size of the LHC collaborations to about 1500 physicists it is believed that an effective coordination of the efforts of all these people sparse in the world would require a much larger usage of new emerging tools such as video-conferencing or other computer-based systems for work at distance. This is the other crucial issue, together with the computing model that need to be solved to allow in practice to physicists of outside institutes to take effectively part in the experiment efforts.

Recently a very positive and fast progress was happening in this field.

A project aimed to introduce and integrate modern video-conference methods into the daily working environment of the collaborators of the LHC experiments was recommended by the LHC Computing Board (LCB) [13], the CERN advisory committee set up to foster common projects in computing for the LHC experiments, and approved by the CERN management. The project [14] aims to overcome the limitations of the present systems.

In particular the CODEC system proposed by the Telecom companies, shows to be quite inflexible in use and booking and not spontaneous or interactive.

The Packet system (Mbone) based on internet has unpredictable and often inadequate performance and there is a lack of integration between CODEC and Packet world.

The first phase of the project, ending at the end of 1998, will be dedicated to investigate available solutions,develop the necessary improvements and set up a demonstrator installation at CERN with the aim to have final recommendations for a full deployement of the videoconference service that should be relized starting from 1999.

8 Conclusions

The complexity of the LHC experiments has forced the HEP community to enter into a new revolutionary phase. New powerful tools and methodologies are being adopted to solve the very challenging tasks in the computing domain. The first results are very encouraging :

- Geant4 has shown the validity of the new paradigm to build complex LHC software
- RD45 has shown that ODBMS can provide new very powerful tools for access to very large data volumes

- RD47 is showing the effectiveness of computing farms based on commodity processors
- A new project on video-conferencing will provide much better tools for collaborative work
- The technology evolution is helping as expected

but the process to build an efficient analysis environment for the LHC era is very challenging and continuous research and development efforts from the LHC collaboration in the computing field will be needed to arrive to a new computing model.

References

[1] ATLAS Computing Technical Proposal,
$http://atlasinfo.cern.ch/Atlas/GROUPS/SOFTWARE/ctp/ctp-work.html$.

[2] CMS Computing Technical Proposal,
$http://cmsdoc.cern.ch/ftp/CMG/CTP/index.html$.

[3] "PASTA-The LHC technology Tracking team for processors, Memory,Architectures,Storages and Tapes", Status Report, August 1996.

[4] "The LHC Networking technology Tracking Team,Status Report", October 1996.

[5] "Object Database and Mass Storage Systems : The prognosis", CERN/LHCC 96-17,1996.

[6] RD47 collaboration,
LCB status report in preparation.

[7] RD45 collaboration,
$http://wwwcn1.cern.ch/asd/cernlib/rd45/index.html$.

[8] Geant4 Collaboration,
$http://wwwinfo.cern.ch/asd/geant/geant4.html$.

[9] D. Weiss, "Funnell: Towards comfortable event processing",
CHEP 95 Proceedings, Rio de Janeiro, Brasil, september 1995,pp. 59-63.

[10] Condor,
$http://www.cs.wisc.edu/condor/$.

[11] Nile,
$Untitled, URLhttp://w4.lns.cornell.edu/$.

[12] W.E. Johnston, "Network-based Remote Instrument and Experiment Control", CSC 97.

[13] LCB,
$http://www.cern.ch/Committees/LCB/$.

[14] Video-conference project,
$http://sunmed2.cern.ch:8000/www_lcb/html/pep/pep.ps$,
$http://sunmed2.cern.ch:8000/www_lcb/html/pep/pep.html$.