Notes on EDG Tutorial presentation 'Applications and the Grid'

F Harris 1 Nov 2002

Introduction (slides 1-3)

The presentation discusses the current models for the use of Grid services by the 3 application areas within EDG, namely High Energy Physics (HEP), Biomedicine and Earth Observation.

All application areas share a common vision of the Grid in that their application software sits on top of high level services, such as data management services, which in turn interface to the Grid middleware services. The high level services typically have 2 elements, firstly those which are coupled to details of their specific software frameworks, and secondly those which are of a generic nature, and independent of the frameworks. For this latter category one has the possibility to define high level services which could be shared between different applications.

The fundamental attraction of Grid technology is to offer a homogenous way of defining a 'virtual computing laboratory' composed of distributed heterogeneous resources. In addition one can define a virtual organization(VO) which allows a distributed community to work together, and which manages the allocation of resources to authenticated and authorized users. The fundamental functions offered by Grid technology are the 'single log-on' to the grid, together with the capabilities of job submission, data management and monitoring for 'jobs' which obtain resources from the Grid according to their requirements for data, CPU, software, turnaround time etc.

LHC Computing (the 'old' hierachical view) (Slide 4)

The MONARC project between 1998-2000 developed a hierachical model for LHC computing in terms of Tiers 0-4, with Tier-0 being CERN, Tier-1 being a national centre, Tier-2 a regional centre, Tier-3 a university/lab centre and Tier-4 an individual workstation. In fact one can also add the mobile user with a laptop as Tier-5. This classification was in terms both of geographical and financial realities, and the functions associated with the different levels. This remains a useful classification in terms of thinking about the services offered at each level in terms of functions and quality, but the concepts of the management and the geography have advanced with the development of a so-called 'cloud model' for tiers outside CERN. This will be discussed in more detail when we outline the thinking for the development of the first LHC Computing prototype.

Typically (e.g. Atlas/CMS) events of a size of ~ 1 MB will be produced at the LHC at a rate of ~100 Hz into offline processing. Assuming the experiment runs for 2*10**7 sec/year this produces ~2 PB of raw data/year, which has to be stored, processed and distributed according to the computing model of the experiment. In addition this latter processing produces yet more data to be stored and distributed .

For more detail of the computing models for each LHC experiment see

http://lhc-computing-review-public.web.cern.ch/lhc-computing-review-public/Public/Report_final.PDF

Data Handling and Computation for Physics Analysis (Slide 5-6)

The raw data flows first into processing for event filtering and reconstruction producing so-called 'event summary data'(ESD). This takes place in quasi real-time. In addition this can be repeated a few times a year as software and detector calibrations improve. The ESD data then flows into 'batch physics analysis' . It may have been reduced from the original 2*10**9 events depending on whether the 1st stage has filtered further or not. This batch analysis will typically take place ~ once a month on the current ESD sample producing a further reduced samples of ~ 10** 7 events, extracted by physics topic. These are the so-called AODs(Analysis Objects), which are a basic input to individual physics analyses.

A crucial element of processing is the generation and analysis of simulated data. Until the start of LHC in 2007 this is the only source of data for physicists and engineers developing detectors and the associated physics analysis software. It is also vital during experiment running to check the data and software performance with real data. Indeed for LHCb this will be the dominant processing requirement. The majority of this MC generation and processing will be accomplished outside CERN for all the LHC experiments.

The RAW data comprises the basic digital information for detector hits and pulse heights. The picture shows a visualization of hits in the Atlas detector for a Higgs to b,b-bar event, with the b particles having decayed into jets of high energy particles. The reconstruction software, which produces the ESD data, links the basic sub-detector data into track segments and energy clusters., and then in turn into tracks and particles by combining sub-detectors. These particles have energy and momentum associated with them according to provisional particle identification associations. Later processing which produces the AOD data, attempts to group tracks into vertices, and also combine vertices/tracks into events according to physics group processing. A still further level of processing produces essentially a compact AOD with global signatures for the event, such as the total number of tracks, total energy, and probabilities for particle classifications. And crucially the TAG data includes links to the relevant AOD/ESD/RAW data, which will enable steering to all levels of event data as required.

HEP Data Analysis – processing patterns(slides 7-8)

HEP data processing is fundamentally parallel due to the independent nature of events. So processing includes the elements of splitting a 'job' into 'sub-jobs', the execution of the sub-jobs, and the subsequent merging of the sub-job outputs. For example a simulation job, required to produce say 10**6 events, is typically defined as 2000 sub-

jobs, each taking ~ 1 day to execute on a modern PC, and producing ~ 1 GB data. This data being subsequently merged into a dataset of 2 TB.

Processing can be classified into 2 types, production and user analysis. The former is planned by the data processing and physics managers of the experiment. It is very large scale and take substantial CPU and storage resources. Consequently rather few people are authorized to launch such productions. The results of productions are subsequently replicated in the experiment according to the policy of the experiment.

User analysis, operating on the TAG and AOD sets of their physics group, is by definition 'chaotic' depending on the current planning of the physics group and the individual. There may be many passes over the group AOD and TAG sets before a particular individual is happy that he has a data subset on which he can concentrate. The organization of such processing will vary from experiment to experiment, but there will need to be facilities to allow very efficient searching of datasets, and protections against 'rogue' jobs via the authorization scheme within the VO.

A logical view of Event Data for physics analysis (slides 9-10)

The individual physics user will interface to the system via requests based on collections of data. The user will, for example, say ' please give me access to the latest set of B-> pi,pi events '. The system will then scan the TAG dataset for events satisfying the selection criteria derived from the high level request. The TAG data will include links to the basic high level data (RAW,ESD and AOD) to enable its retrieval for detailed individual analysis. Although the majority of analyses will be done on TAG+AOD data some access to ESD and RAW data will be allowed for developing reconstruction algorithms and associated data visualization. In these cases it is very likely that the most efficient way will be to organize staging of such data by efficient selection procedures rather than to allow analysis jobs to do it on the fly.

The LCG project has spawned a persistency project to develop common tools for use by all the experiments interfacing their frameworks to experiment data. The software will will be independent of the technology used to store the data. This software will interface to Grid tools to find the 'best' data on the Grid. It will be capable of running without the Grid.

More details of the LCG project and POOL can be found at
http://lcg.web.cern.ch/LCG and http://lcgapp.cern.ch/project/persist/

In the following sections are shown examples of development in the LHC experiments which are relevant to Grid based computing. More detail can be found of such developments in the WP8 long term requirements document which can be found at
http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=332409

Analysis and Grid developments in ALICE (slides 11-12)

The ALICE experiment are developing the use of PROOF(Parallel ROOt Facility) for interactive physics analysis. The user opens a ROOT session and defines a set of criteria on which to select events for data analysis. This criteria are applied to the local or remote TAG database, and the location of the files in which selected data reside is provide by an RDB(Relational Database). Sets of parallel tasks are set up and distributed to where the data is located. The parallel tasks can be either complex analysis sending results back to the user, or be a simple process sending event data back to the local station. This will depend on several factors, including the amount of data and the complexity of the processing to be accomplished.

http://alien.cern.ch

ALICE are developing Alien , a system for managing distributed processing. It includes a File Catalogue based on an RDB and an associated TAG catalogue. It has facilities for job queue management, using either push or pull techniques. Interfaces are being developed from Alien to both EDG and iVDGL(US) testbeds. An interface exists for the EDG Resource Broker.

ATLAS/LHCb software framework and Grid interfacing (slides 13-14)

Both Atlas and LHCb share the same software framework. It is based on a service architecture with the user application seeing a set of services including, for example, data and job options services. It is these services that will be interfaced to the Grid in a manner transparent to the user program. Thus for example it is the persistent data service that will be responsible for accessing data whether it be stored remotely and accessed via the Grid or stored locally on a laptop disconnected from the outside world.

A joint development project GANGA(Gaudi And Grid Alliance) is developing a GUI based application which will handle for the user job preparation, resource booking and submission together with associated monitoring and control. GANGA will interface both to Grid services and to the user program operating in the Athena/Gaudi framework, which in turn may interface to the Grid for data access and monitoring services.

A CMS Grid Job – the vision for 2003 (Slide 15)

CMS are developing a system based on the concept of decomposing a job into sub-jobs. Some jobs , for example a production simulation, may be decomposed into many sub-jobs. Others, such as a simple user analysis, may have few sub-jobs. The idea is that sub-jobs will communicate by passing data via files . A data flow graph can be defined showing the relationship between sub-jobs and files.

A job is created by a user using an analysis tool which passes it to high level Grid services. A high level component, provided by CMS, the grid job decomposition service, defines the sub-jobs and the data flow graph. This interacts with several catalogues and

other grid services to determine the way in which a job is decomposed. The grid scheduler will then map the sub-jobs to the sites, and generates appropriate file replication actions.

Deployment of the LHC Global Grid service (Slide 16)

The LHC Computing Grid (LCG) project is aiming to set up a first prototype LHC production prototype, based on Grid technology, in the latter part of 2003. The planning for this has resulted in the development of the hierachical model of tiers by mixing in an element of a 'cloud' model. As before one has a service oriented model for tier functionalities, with lower levels of functionality and quality of service as one goes from Tier-0, Tier-1 down to regional and institute levels. For example Tier-0 and Tier-1 will have to offer substantial CPU and storage facilities, and staffing to support a 24 hour, 7 day service at a national level. They will have to support substantial levels of production processing and the replication of data for all the participating experiments. By the time one gets to Tier-3, the institute centres, one will expect just local user support, and operation according to local agreements.

It should be emphasized that the experiment and LHC computing models are developing, and particularly the analysis models. Physics analysis will involve the cooperation of groups of physicists who will be distributed throughout the collaboration. How this will be managed will be experiment dependent. It will have to take into account the chaotic nature of individual physics analyses and the sharing of data and results within a physics group.

With the cloud model services will be available through an Internet portal which channels requests to appropriate sites. Thus a Tier-1 would be replaced by a 'Grid Service Exchange'(GSX) which coordinates the use of resources in a 'Grid Service Zone'. The GSX is responsible for the service to the other regional centres in the LHC Computing Grid. Similarly Tier-2 can have a distributed definition, with its resources being managed by a 'Grid Zone Centre'. The cloud model gets away from the concept of monolithic centres, and takes into account the distribution of resources connected by very high performance networking. An excellent example of this is in Dutchgrid which makes extensive use of a high bandwidth network SURFnet connecting the academic institues in Holland. All partners have a t least a Gbit connection to a backbone which has 20 Gbit bandwidth. It is easy for a university group to have access to central mass storage at the Academic Computer Centre SARA. Consequently the Dutch Cluster can be regarded as a single regional centre.

A review of current thinking for the LHC computing model can be found at
http://lcg.web.cern.ch/LCG/SC2/RTAG6

It is also useful to consider the status of development of international research networks in Europe. Details can be found at
http://www.dante.net/geant/

The application of the Grid in Biomedicine (EDG workpackage 10) - challenges and requirements (slides 17-19)

Unlike HEP, which has well defined international centres for computing all over the world, the field of biomedicine has not had such established centres , nor a general awareness of the technical problems and possible solutions for distributed computing. However computation and the management of data is an increasingly important issue and grid technology opens up a possibility for viable technical solutions. A first biomedical grid is being deployed by the Datagrid IST project.

Currently there is a growing awareness of common needs such as agreed standards for formats in storing data, and the necessity of common long-term investments. It should be remembered that there is a very large potential user community based in hospitals, medical centres and biology research laboratories throughout the world. This growing community numbers tens of thousands.

The basic requirements are similar to those in HEP. For example a single hospital can gather Terabytes of medical image data in a year, which need indexing, versioning and perhaps replica management. Security is a key issue in medicine, and also in biology. There is a growing need for large scale computation involving both parallelism and pipelining.

An overview of the basic requirements for Grid services in biomedicine can be found at
http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=332412

Grid Biomedical Projects (slide 20)

Biomedical projects typically involve 3 classes of software running over Grid middleware. Firstly there is the development of the basic framework for sharing resources and algorithms in a cooperative manner according to the demands of a particular experiment. One may draw an analogy with a HEP analysis program being configured according to its requirements for data and algorithms.

A second crucial aspect is the development of 'grid-aware' algorithms for bio-informatics, data mining etc. These algorithms should be made available to the community via the Grid Services.

A third, very important aspect is the development of Grid portals which interface the user to the Grid by accepting his requests for resources according to his job definition, and attempts to satisfy the request by the Grid.

Grid impact on data handling (slide 21)

Grid replication services will greatly facilitate the mirroring of important databse information to research centres. For example the major biomedical centres at Cambridge,UK and Geneva can publish the existence of new database information to the Replica Catalogue. A centre subscribing to this Catalogue at Paris can then, using the replica management handling, obtain a copy of this new data, and, in turn, copy it to participating satellite nodes.

Web Portals for biologists (Slides 22-23)

A biologist will enter a genomic sequence through an interface , and trigger the piplelined and parallel execution of comparative algorithms acting on genome files. Such processing is very CPU intensive, and is currently growing in scale with the size of the databases. Currently processing can take days on small institute clusters. With the sharing of resources by the Grid this can become much more efficient.

A particular development has been with 'Visual Datagrid Blast' which offers a graphical interface to enter the query sequences and select the reference database. A graphical interface presents the result of the analysis.

This development work has been based on the use of the portal GENIUS which was developed in EDG/INFN in the context of HEP, but which has general purpose functionality which can be tailored to the needs of the application. Details of this portal software can be found at http://genius.ct.infn.it

Summary of 'added value' provided by the Grid for Bio-medicine(slides 23-24)

The development of a Virtual Organisation gives the benefits which are common to all application areas. The particular technical benefits resulting from the setting up of a coherent, resource sharing community are in several areas.

Firstly the coherence in data management will allow the efficient organisation of data mining on genomic databases which have exponential growth.
Secondly the proper management of indexed, distributed medical images will greatly enhance medical studies, and also their utitisation by medical practitioners.

The provision of a collaborative framework for defining experiments will greatly enhance the development of , for example, large scale epidemiological studies.

Lastly the potential of getting hold of large scale distributed computing resources will enhance the use of parallel and pipelined processing in several applications, such as database searches and 3D modelling.

The application of the Grid in Earth Observation (EO) (slides 25-26)

Current collaborative distributed computing is being employed for the processing and validation of Global Ozone (GOME) satellite data processing. This is being performed by KNMI(Holland), IPSL(France) and ESA(Italy). Datagrid is being investigated as providing an enhanced collaborative environment for the project.

A new satellite Envisat was launched in Feb 2002. This sends 200 Mbps to ground, and 400 TB will be archived each year. The data will be processed by more than 10 dedicated sites in Europe,and there are more than 700 approved scientific user projects.

A summary of the requirements for Grid Services in EO can be found at
http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=332411

GOME data processing model (slides 27-30)

Two different satellite data processing techniques are applied. One, a tightly coupled technique using MPI, is used by KNML. The other , using neural networks, is used by ESA. The results from the 2 techniques are validated by IPSL using ground-based LIDAR measurements.

The level-1 raw data is processed to give level-2 data which provides measurements ozone in vertical columns above the earth. Coincident data consists of level-2 data compared with the ground LIDAR measurements.

The total dataset in a year is 267 GB, of which 190 GB is in the form of ~19 million files of ~10KB.

GOME processing and the Grid (slides 31-33)

The first steps involve transferring the level-1 data to a Storage Element(SE) and its subsequent registration with the Replica Manager (RM). Consequently replicas are made to other SEs if necessary.

Jobs are then submitted to process level-1 data producing level-2 data, which must be transferred to an SE. This phase involves use of the Resource Broker, the Replica catalogue and the Information Index. Finally the level-2 data is validated with the LIDAR data, and results are visualized.

This system is currently being commissioned with EDG middleware

Summary Comments(slides 34-36)

All 3 application are currently evaluating EDG middleware on the applications testbed. They have all achieved some measure of success. For example all the experiments have successfully interfaced their simulation production systems to EDG middleware, and hope with the next EDG release to use it in a production environment. The application area share common needs in the area of provision of high level application interfaces to middleware services. There is hope to identify common project work in this area.

All application areas are interested in interactive computing on the Grid, whether it be for interactive HEP physics analysis, or for interactive image processing in medicine or Earth Observation. This will pose special problems in provision of high priority processing power and organization of fast data access.

Within the scope of the LCG project a detailed analysis was made of common use cases for HEP applications. This analysis will continue, and will be used as the basis for the evaluation of grid middleware. The document can be found at
http://lcg.web.cern.ch/LCG/SC2/RTAG4

There are many grid projects in the international community. Those of relevance to HEP are EDG, Crossgrid, Nordugrid, Datatag in Europe, and Griphyn, PPDG and iVDGL in the USA. All are investigating the possibility of inter-working, and the sharing of future developments. This is essential for the provision of a true international grid

Web Sites and Documents

–LCG http://lhc-computing-review-public.web.cern.ch/lhc-computing-review-public/Public/Report_final.PDF (LHC Computing Review)

http://lcg.web.cern.ch/LCG

http://lcg.web.cern.ch/LCG/SC2/RTAG6 (model for regional centres)

http://lcg.web.cern.ch/LCG/SC2/RTAG4 (HEPCAL Grid use cases)

–GEANT http://www.dante.net/geant/ (European Research Networks)

–POOL http://lcgapp.cern.ch/project/persist/

–WP8 http://datagrid-wp8.web.cern.ch/DataGrid-WP8/

http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=332409 (Requirements)

–WP9 http://styx.srin.esa.it/grid

http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=332411 (Requirements)

–WP10 http://marianne.in2p3.fr/datagrid/wp10/

http://www.healthgrid.org
http://www.creatis.insa-lyon.fr/MEDIGRID/

http://edmsoraweb.cern.ch:8001/cedar/doc.info?document_id=332412 (Requirements))

-Some US HEP Grid sites

http://www.ppdg.net
http://www.ivdgl.org
http://www.griphyn.org
http://www.hicb.org
http://www.lsc-group.phys.uwm.edu/vdt/