Models of Networked Analysis at Regional Centres for LHC Experiments

(MONARC)

MID-PROJECT PROGRESS REPORT

MONARC Members

M. Aderholz (MPI), K. Amako (KEK), E. Arderiu Ribera (CERN), E. Auge (L.A.L/Orsay), G. Bagliesi (Pisa/INFN), L. Barone (Roma1/INFN), G. Battistoni (Milano/INFN), M. Bernardi (CINECA), G. Boschini (CILEA), A. Brunengo (Genova/INFN) J. Bunn (Caltech/CERN), J. Butler (FNAL), M. Campanella (Milano/INFN), P. Capiluppi (Bologna/INFN), M. D'Amato (Bari/INFN), M. Dameri (Genova/INFN), A. di Mattia (Roma1/INFN), G. Erbacci (CINECA), U. Gasparini (Padova/INFN), F. Gagliardi (CERN), I. Gaines (FNAL), P. Galvez (Caltech), A. Ghiselli (CNAF/INFN), J. Gordon (RAL), C. Grandi (Bologna/INFN), F. Harris (Oxford/CERN), K. Holtman (CERN), V. Karimäki (Helsinki), Y. Karita (KEK), J. Klem (Helsinki), I. Legrand (Caltech/CERN), M. Leltchouk (Columbia), D. Linglin (IN2P3/Lyon Computing Centre), P. Lubrano (Perugia/INFN), L. Luminari (Roma1/INFN), A. Maslennicov (CASPUR), A. Mattasoglio (CILEA), M. Michelotto (Padova/INFN), I. McArthur (Oxford), Y. Morita (KEK), A. Nazarenko (Tufts), H. Newman (Caltech), V. O'Dell (FNAL), S.W. O'Neale (Birmingham/CERN), B. Osculati (Genova/INFN), M. Pepe (Perugia/INFN), L. Perini (Milano/INFN), J. Pinfold (Alberta), R. Pordes (FNAL), F. Prelz (Milano/INFN), S. Resconi (Milano/INFN and CILEA), L. Robertson (CERN), S. Rolli (Tufts), T. Sasaki (KEK), H. Sato (KEK), L. Servoli (Perugia/INFN), R.D. Schaffer (Orsay), T. Schalk (BaBar), M. Sgaravatto (Padova/INFN), J. Shiers (CERN), L. Silvestris (Bari/INFN), G.P. Siroli (Bologna/INFN), K. Sliwa (Tufts), T. Smith (CERN), R. Somigliana (Tufts), C. Stanescu (Roma3), D. Ugolotti (Bologna/INFN), E. Valente (INFN), C. Vistoli (CNAF/INFN), I. Willers (CERN), R. Wilkinson (Caltech), D.O. Williams (CERN).

14th June 1999

Executive Summary

The MONARC Project is well on the way towards its primary goals of identifying baseline Computing Models that could provide viable (and cost-effective) solutions to meet the data analysis needs of the LHC experiments, providing a simulation toolset that will enable further Model studies, and providing guidelines for the configuration and services of Regional Centres. The criteria governing the MONARC work are:

- the network bandwidth, computing and data handling resources likely to be available at the start of and during LHC running,
- the computing power and data transport speeds needed for an effective data analysis,
- the features and performance of the distributed database system and
- an overall strategy for data processing, distribution and analysis that meets the needs while using the resources efficiently, with acceptable turnaround times.

The main deliverable from the project is a set of example "baseline" Models. The project aims at helping

to define regional centre architectures and functionality, the physics analysis process for the LHC experiments, and guidelines for retaining feasibility over the course of running. The results will be made available in time for the LHC Computing Progress Reports, and could be refined for use in the Experiments' Computing Technical Design Reports by 2002.

The approach taken in the Project is to develop and execute discrete event simulations of the various candidate distributed computing systems. The granularity of the simulations is adjusted according to the detail required from the results. The models are iteratively tuned in the light of experience. The model building procedure, which is now underway, relies on simulations of the diverse tasks that are part of the spectrum of computing in HEP. A simulation and modelling tool kit is being developed, to enable studies of the impact of network, computing and data handling limitations on the models, and to test strategies for an efficient data analysis in the presence of these limitations.

Chapter 1: Introduction

The scale, complexity and worldwide geographical spread of the LHC computing and data analysis problems are unprecedented in scientific research. Each LHC experiment foresees a recorded raw data rate of 1 PetaByte/year (or 100 MBytes/sec during running) at the start of LHC operation. This rate of data to storage follows online filtering by a factor of several hundred thousand, and online processing and data compaction, so that the information content of the LHC data stores will far exceed that of the largest PetaByte-scale digital libraries foreseen for the next 10-15 years. As the LHC program progresses, it is expected that the combined raw and processed data of the experiments will approach 100 PetaBytes by approximately 2010. The complexity of processing and accessing this data is increased substantially by the size and global span of each of the major experiments, combined with the limited wide area network bandwidths that are likely to be available by the start of LHC data taking.

The general concept developed by the two largest experiments, CMS and ATLAS, is a hierarchy of distributed Regional Centres working in close coordination with the main centre at CERN. The regional centre concept is deemed to best satisfy the multifaceted balance needed between

1. proximity of the data to centralised compute and data handling resources,
2. proximity to the end-users for frequently accessed data,
3. efficient use of limited network bandwidth,
4. appropriate exploitation of regional and local computing and data handling resources,
5. effective involvement of scientists in each country and each world region in the data analysis and the realisation of the experimental physics discoveries.

The use of regional centres is well matched to the worldwide-distributed structure of the collaboration, and will facilitate access to the data through the use of national and regional networks of greater capacity than may be available on intercontinental links.

The MONARC project is the means by which the experiments have banded together to meet the technical challenges posed by the storage, access and computing requirements of LHC data analysis. The baseline resource requirements for the facilities and components of the networked hierarchy of centres, and the means and ways of working by which the experiments may best use these facilities to meet their data-processing and physics-analysis needs, are the focus of study by MONARC.

The primary goals of MONARC are to:

- determine which classes of models, and modes of distributed analysis, are feasible according to the network capacity and data-handling resources available at the collaborating sites
- specify the main parameters that characterise these classes of models

- produce example baseline models which fall into the "feasible" category
- deliver a set of tools for simulating candidate computing models of the experiments
- formulate a set of common guidelines to allow the experiments to formulate their final Models
- formulate a set of guidelines for Regional Centre architecture and functionality, as well as the interactions among the Centres

In order to achieve these goals MONARC has organised itself into four working groups, and is led by a Steering Group responsible for directing the project and coordinating the Working Group activities. Members of the Steering Group are given below:

Table 1: Members of MONARC Steering Group

Steering Group Member	Principal Activity
Harvey Newman (Caltech)	Spokesperson
Laura Perini (INFN Milano)	Project Leader
Krzysztof Sliwa (Tufts)	Simulation and Modelling WG Leader
Joel Butler (Fermilab)	Site and Network Architectures WG Leader
Paolo Capiluppi (INFN Bologna)	Analysis Process Design WG Leader
Lamberto Luminari (INFN Roma)	Testbeds WG Leader
Les Robertson (CERN IT)	CERN Centre Representative
David O. Williams (CERN IT)	Network Evolution and Costs
Frank Harris (Oxford/CERN)	LHCb Representative
Luciano Barone (INFN Roma)	Distributed Regional Centres
Jamie Shiers (CERN IT)	RD45 Contact
Denis Linglin (CCIN2P3 Lyon)	France RC Representative
John Gordon (RAL)	United Kingdom RC Representative
Youhei Morita (KEK)	Objectivity WAN (KEK)

A Regional Centres Committee has been formed, composed of representatives of actual and potential regional centres; which acts as an extended MONARC Steering Group.

The progress of each of the Working Groups is summarised in the following chapters of this report.

As scheduled in the PEP, the **MONARC Simulation WG (Chapter 2)** has developed a flexible and extensible set of common modelling and simulation tools. These tools are based on Java, which allows the process-based simulation system to be modular, easily extensible, efficient (through the use of multi-

threading) and compatible with most computing platforms. The system is implemented with a powerful and intuitive Web-based graphical user interface that will enable MONARC, and later the LHC experiments themselves, to realistically evaluate and optimise their physics analysis procedures.

The **Site and Networks Architectures WG (Chapter 3**) has studied the computing, data handling and I/O requirements for the CERN centre and the main "Tier1" Regional Centres, as well as the functional characteristics and wide range of services required at a Tier1 Centre. A comparison of the LHC needs with those of currently running (or recently completed) major experiments has shown that the LHC requirements are on a new scale, such that worldwide coordination to meet the overall resource needs will be required. Valuable lessons have been learned from a study of early estimates of computing needs during the years leading up to the "LEP era". A study of the requirements and modes of operation for the data analysis of major experiments just coming (or soon to come) into operation has been started by this group. The group is also beginning to develop conceptual designs and drawings for candidate site architectures, in cooperation with the MONARC Regional and CERN Centre representatives.

The **Analysis Process Design WG (Chapter 4)** has studied a range of initial models of the analysis process. This has provided valuable input both to the Architectures and Simulation WG's. As the models and simulations being conducted became more complex, close discussions and joint meetings of the Analysis Process and Simulation WG's began, and will continue. In the future, this group will be responsible for determining some of the key parameter sets (such as priority-profiles and breakpoints for re-computation versus data transport decisions) that will govern some of the large scale behaviour of the overall distributed system.

The **Testbeds WG (Chapter 5)** has defined the scope and a common (minimum) configuration for the testbeds with which key parameters in the Computing Models are being studied. The recommended test environment including support for C++, Java, and Objectivity Version 5 has been deployed on Sun Solaris as well as Windows NT and Linux systems. A variety of tests with 4 sets of applications from ATLAS and CMS (including the GIOD project) have begun. These studies are being used to validate the simulation toolset as well as extracting key information on Objectivity performance.

Distributed databases are a crucial aspect of these studies. Members of MONARC also lead or participate in the RD45 and GIOD projects which have developed considerable expertise in the field of Object Database Management Systems (ODBMS). The understanding and simulation of these systems by MONARC have benefited from the cooperation with these projects.

Chapter 6 of this report summarises the **workplan and schedule**, from now to the end of Phase 2. This chapter also introduces a possible Phase 3 of MONARC which would define and study an optimised integrated distributed system aimed at using the available resources most efficiently, and discusses the relative scope and timing of Phase 2 and 3. The status of the milestones presented in the PEP is reviewed, and more specific milestones are set for the upcoming stage of the project. The status of MONARC's relations to the other projects mentioned in the PEP also is briefly reviewed.

Finally, **Chapter 7** presents an overview of the **system optimisation issues**, and of current or upcoming projects addressing them. Such issues will be the core of the Phase 3 R&D studies. These ideas will be discussed in MONARC within the next few months, in order to formulate a proposal for a PEP extension, to be presented towards the end of 1999.

Chapter 2: Progress Report of the Simulation Working Group

2.1 Introduction

The development of a powerful and flexible simulation and modelling framework for the distributed computing systems was the most important task in the first stage of the project. Some requirements for the framework are listed below:

- allow the study of candidate reconstruction and analysis process architectures for the LHC experiments,
- perform reliable modelling of large computing facilities and the networks connecting them,
- carry out simulations of complex models as fast as possible without jeopardising the correctness of the results
- the model implemented in the framework's simulation program should be equivalent in behaviour to the real system in all important aspects
- understanding of the real system's components in terms of the simulation model should be straightforward.

The distributed nature of the reconstruction and analysis processes for the LHC experiments required the framework's simulation program capable of describing complex patterns of data analysis programs running in a distributed computing system. It was recognised from the very beginning that a process-oriented approach for discrete event simulation is well suited to describe a large number of programs running concurrently, all competing for limited resources (data, CPU, memory, network bandwidth etc.).

2.2 Survey of existing simulation tools

A broad survey of existing tools (SoDA[3], ModNet[4], Ptolemy[5], SES[6], PARASOL[7]), led to a realisation that Java technology provides well developed and certainly adequate tools for developing a flexible and distributed, process-oriented, simulation that would meet the requirements. Java has built-in multi-thread support for concurrent processing, which is an advantage for simulation purposes provided that a dedicated scheduler is developed. Although initially it was thought that the SoDA package developed by C. von Praun at CERN could provide the basis of the tool, it was decided that a considerably more flexible and extensible process-oriented discrete event simulation program could be constructed using readily available Java classes and libraries.

2.3 Development of the Java-based MONARC simulation tools

2.3.1 Modelling the CS2 farm: the first version of the Simulation Program

In February '99 a first version of the simulation program was written quickly (in about one week) by I. Legrand using JDK 1.1. This provided a proof of concept, and a decision was made to continue development in Java. To test the program, a simulation of CERN's CS2 farm was chosen. This parallel computer system is used for data acquisition (DAQ) and processing at CERN. The elements of the simulation program were the variable numbers of DAQ nodes (senders), disk server nodes (writers) and data processing nodes which were connected via a LAN switch. In addition, the DAQ message size, event size, processing time and the sender's time, together with the numbers of various nodes were used as input parameters.

The CS2 farm model was built of passive objects CPU and DISK, which interact via "active objects". The program was a process oriented distributed event simulation based on "active objects" (JOBS and LINKS). Each active object ran concurrently within the simulation program with multi-threading, to "perform" a set of actions which were pre-defined by the user. Each action was defined to have a certain amount of response time, calculated from available shared system resources at the time. If an object was activated or deactivated in the system, an interrupt was signalled to all other active objects, and the response time for the on-going task on each object was re-calculated. The simulation then continued with the re-calculated response time.

The simulation was checked by executing single threads and monitoring virtually every transaction at the most elementary level. Additionally, multi-threaded systems were run in such a way that a comparison of the results could be made with analytical calculations (i.e. by disabling pseudo-random behaviour of the model). The simulation outputs were the total load on the switch (MB/s) and the CPU usage (% of available). These were measured as a function of variable numbers of DAQ, processing and data server nodes, as well as the varying processing time and the message size. The analytically predicted behaviour of the CS2 system (saturation of CPU or the network bandwidth, depending on the chosen set of parameters) was fully reproduced in the simulation. Figure 1 shows the simulation tool in use for the CS2 simulation. Figures 2 and 3 show the simulation results as viewed in the tool GUI.

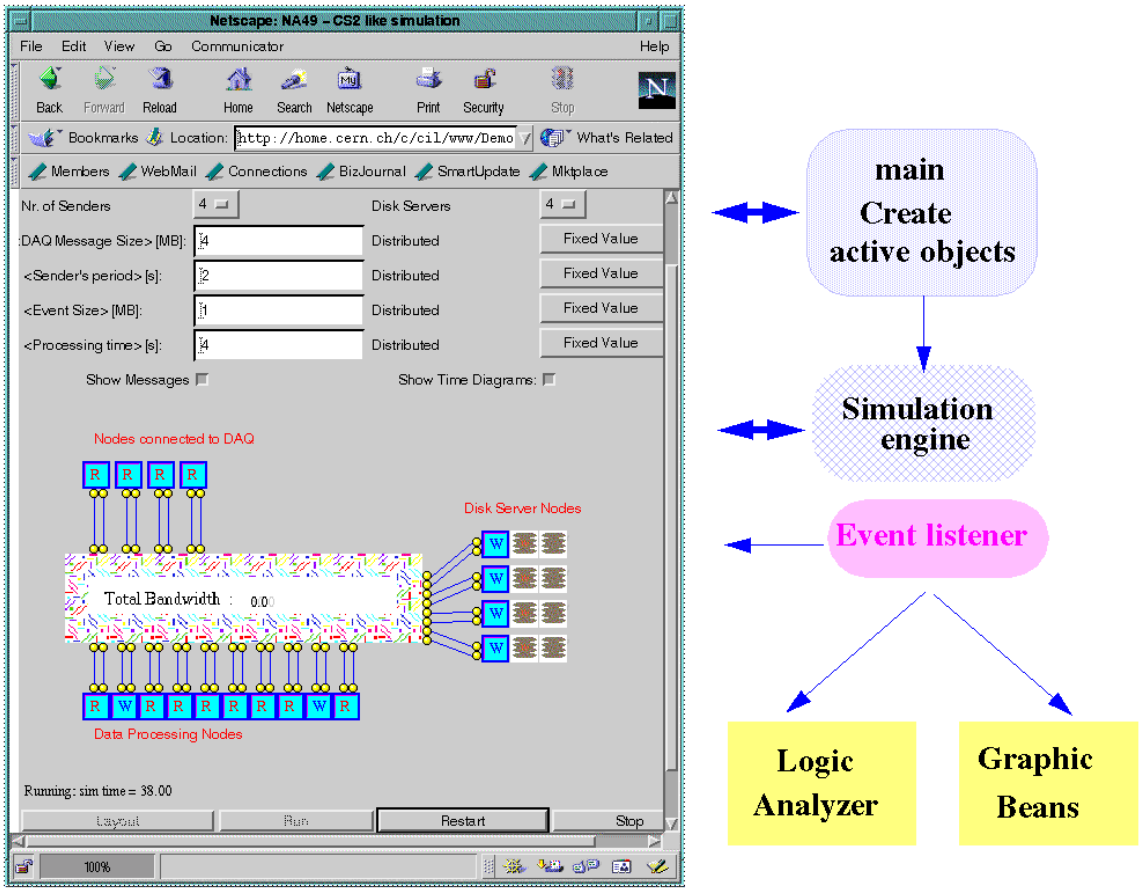


Figure 1: Showing the simulation tool GUI and a schematic of the tool's operation.

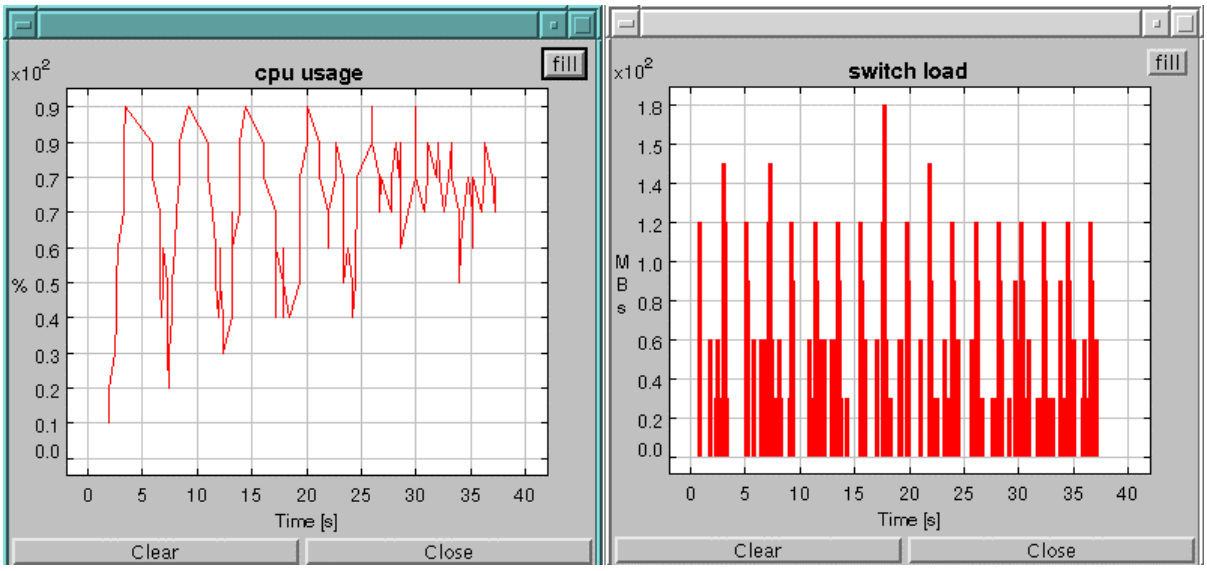


Figure 2: Showing the CPU and I/O usage predicted by the simulation tool for the operation of the CS2

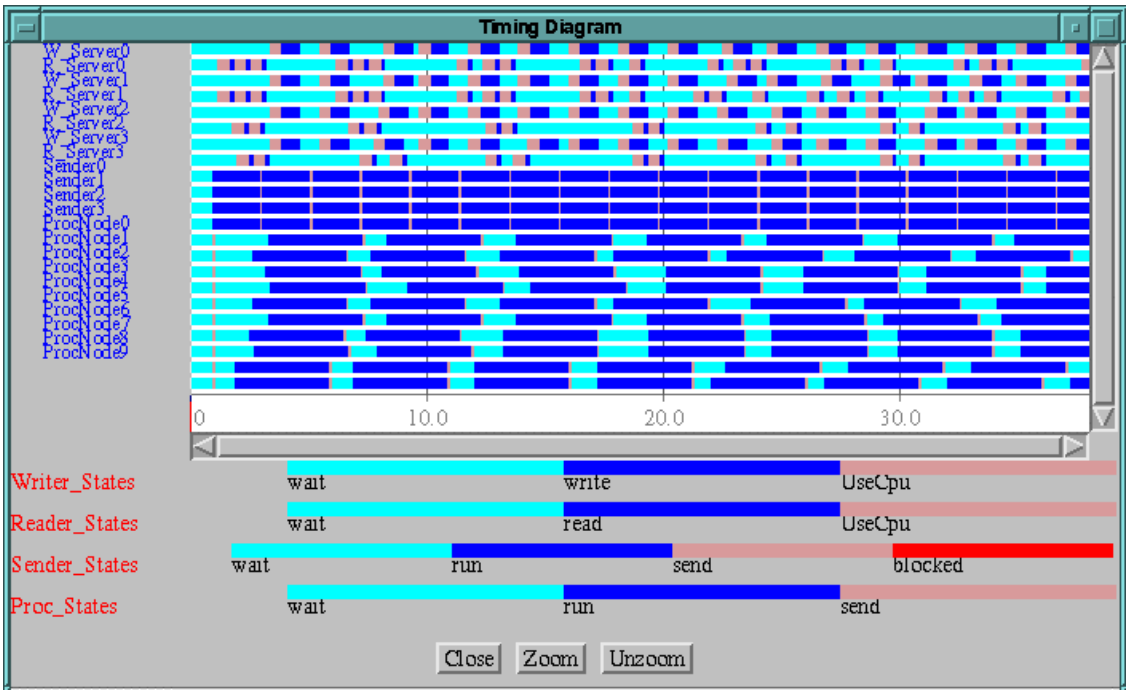


Figure 3: Showing a breakdown of the simulated CS2 task activities as a function of time

2.3.2 Implementing the Data Model

After work on implementation of the Regional Centre simulation had begun, it was soon realised that the Data Model had to be realistically described in the simulation program to allow for a realistic mapping of the behaviour of an object database. As envisaged in the Computing Proposals of the LHC experiments, all data is organised in objects and managed within the framework of an object database. In our case we consider specifically an Objectivity/DB federated database, which allows to distribute sets of objects onto different media, geographically and physically, media (tape, disk...) and data servers (Objectivity AMS servers), while maintaining a coherent and logically uniform view of the entire distributed database. Objects can contain pointers to each other (associations) which enable navigation across the entire

database. The data model implemented in the simulation consists of 4 functionally different groups of objects:

- RAW data; about 1MB/event, most likely to be stored on tape only at CERN
- ESD data (Event Summary Data) - objects with reconstructed information; about 0.1 MB/event
- AOD data (Analysis Object Data) - a subset of ESD (possibly non-overlapping, connected via AOD->ESD associations); about 0.01 MB/event
- TAG - a small set of essential information describing a physics event (jet and lepton multiplicity, trigger masks, values of the transverse energy of the most energetic jets and leptons...) which allows initial selections of which AOD data to process

Data of these four different types are organised in unique containers (files). The simulation has a software equivalent of a real Objectivity/DB database catalogue, which allows each job to identify which containers are needed for processing the data requested by that JOB. The locking mechanism has been implemented on the container level, as in Objectivity federated databases. Different types of operation on the data are modelled by different JOBS; for example RAW->ESD, ESD->AOD and AOD->TAG processing involves different input and output data, and different processing time. For example, if the initial FARM configuration has all data on TAPE, if RAW->ESD jobs are submitted to the queues, they invoke the TAPE->DISK copy process.

The simulation was checked by executing single threads and monitoring virtually every transaction at the most elementary level. Additionally, multi-threaded systems were run in such a way that a comparison of the results could be made with analytical calculations (i.e. by disabling pseudo-random behaviour of the model).

The first version of the Regional Centre simulation program had the passive objects, TAPE, DISK, CPU as generic entities. No realistic configuration was provided, i.e. all DISK was being accessed as if they were part of the same, single, disk server. Discussion on how to implement realistic descriptions of TAPE, DISK, CPU and NETWORK, led to a conceptual design of a second version of the simulation program at the end of February 1999. By implementing the interactions between the different AMS servers (Objectivity disk servers) one could easily built a model with multiple regional centres. Figure 4 shows the simulation components of a model of a Regional Centre.

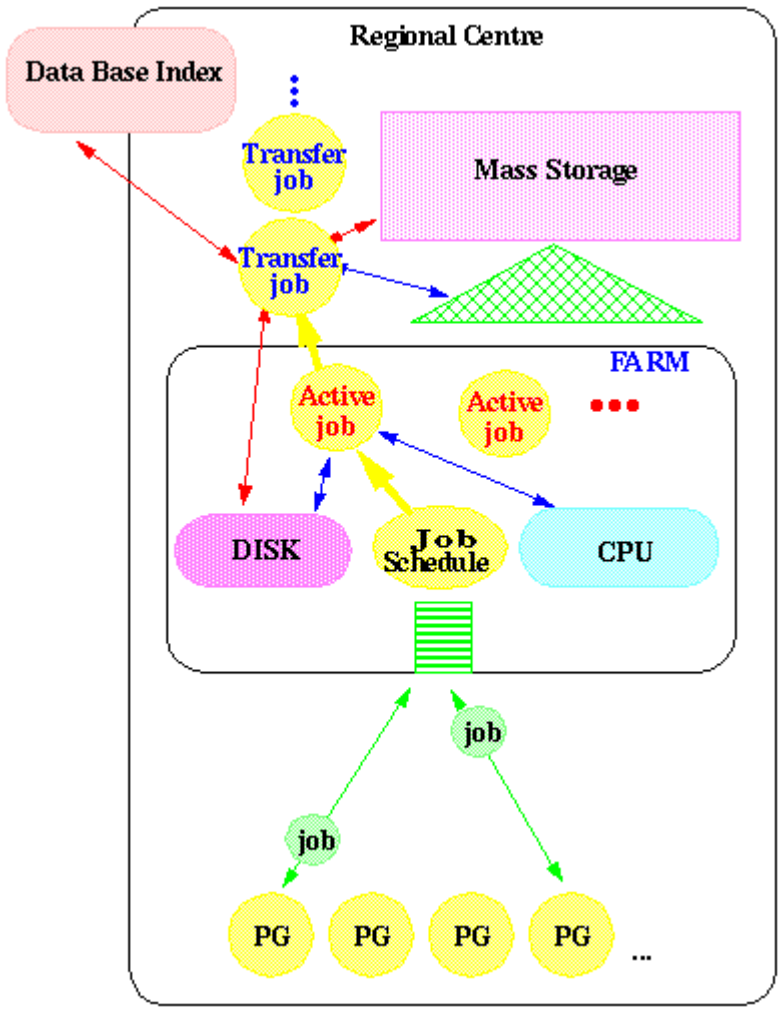


Figure 4: Showing the components of the simulated Regional Centre

2.4 Status of Java-based MONARC simulation tool

2.4.1 Second version of a MONARC simulation tool

The second version of the Legrand simulation program was available for testing at the end of April 1999. It constitutes a major revision of the previous tool, providing an improved description of the Database Model, including multiple AMS servers, each with a finite amount of DISK connected to them. Figure 5 shows the structure of the model used to simulate the Objectivity database system.

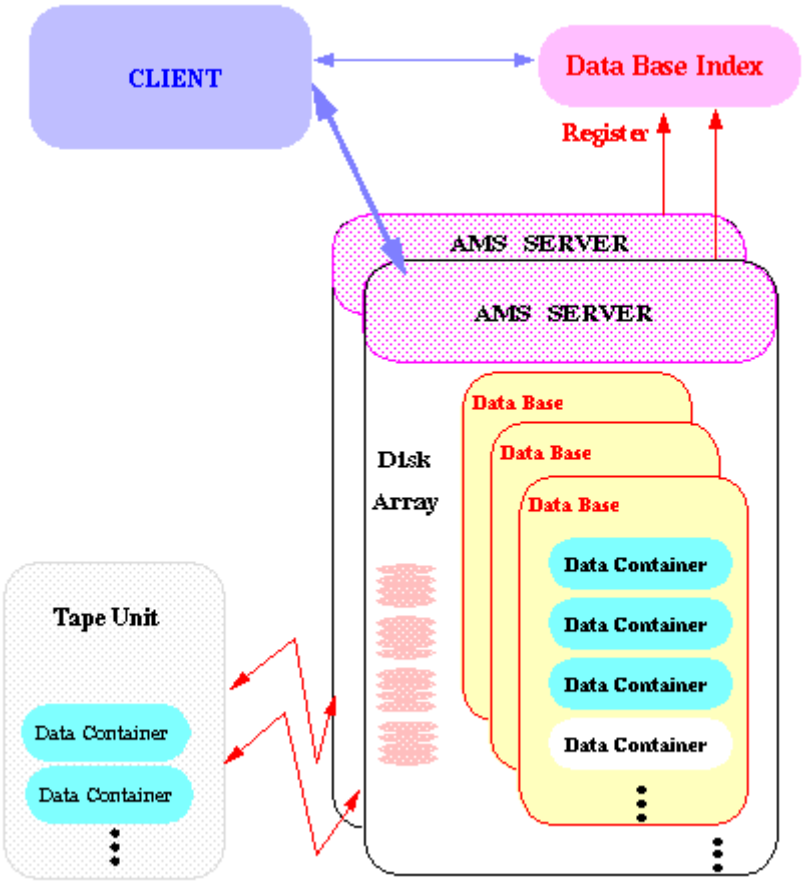


Figure 5: Showing how the Objectivity database is modelled using the simulation tool.

The new scheme provides for an efficient way to handle a very large number of objects and automatic storage management, allows one to emulate different clustering schemes of the data for different types of data access patterns as well as to simulate the order of access following the associations between the data objects, even if the objects reside in databases in different AMS servers.The NETWORK model has been modified as well. It is, at present, an "interrupt" driven simulation. For each new message an interrupt is created, which triggers a re-calculation of the transfer speed and the estimated time to complete a transfer for all the active objects. Such a scheme provides an efficient and realistic way to describe (simulate) concurrent transfers using very different object sizes and protocols. Logically, there is no difference in the way LANs and WANs are simulated. A multi-tasking processing model for shared resources (CPU, Memory, I/O channels) has been implemented. It provides an efficient mechanism to simulate multitasking and I/O sharing. It offers a simple mechanism to apply different load balancing schemes. With the new program it is now possible to build a wide range of computing models, from the very centralised (with reconstruction and most analyses at CERN) to the distributed systems, with an almost arbitrary level of complication (CERN and multiple regional centres, each with different hardware configuration and possibly different sets of data replicated). A much improved GUI, enhanced graphical functions and built-in tools to analyse results of the simulations are also provided. Table 2 shows a list of parameters currently in use by the MONARC simulation tool.

Table 2: A list of parameters currently in use by the MONARC simulation tool

federated database and data model parameters (global)	**regional centre configuration parameters (local)**
database page size	number of AMS_servers
TAG object size/event	AMS link speed
AOD object size/event	AMS disk size
ESD object size/event	number of processing nodes
RAW object size/event	CPU/node
processing time RAW->ESD	memory/node
processing time ESD->AOD	node link speed
processing time AOD->TAG	mass storage size (in HSM)
analysis time TAG	link speed to HSM
analysis time AOD	AMS write speed
analysis time ESD	AMS read speed
memory for RAW->ESD processing job	(maximum disk read/write speed)
memory for ESD->AOD processing job	
memory for AOD->TAG processing job	**data access pattern parameters (local)**
memory for TAG analysis job	fraction of events for which TAG->AOD associations are followed
memory for AOD analysis job	
memory for ESD analysis job	fraction of events for which AOD->ESD associations are followed
container size RAW	
container size ESD	fraction of events for which ESD->RAW associations are followed
container size AOD	
container size TAG	clustering density parameter

A number of parameters can be modified easily using the GUI menus, they include most of the **global** parameters describing the analysis (CPU needed by various JOBS, as well as memory required for processing) and most of **local** parameters defining the hardware and network configuration of each of the regional centres which are part of the model (an arbitrary number of regional centres can be simulated, each with different configuration and with different data residing on it). Also, the basic hardware costs can be input via GUI, which allows simple estimates of the overall cost of a system. This part of the simulation program will certainly evolve to include the price for items which are more difficult to quantify, like inconvenience and discomfort, travel costs et cetera. For each regional centre, one can define a different set of jobs to be run. In particular, one could define different data access patterns in physics analyses performed in each of the centres, with different frequencies of following the TAG->AOD, AOD->ESD and ESD->RAW associations. Figure 6 shows the simulation tool GUI for building a model.

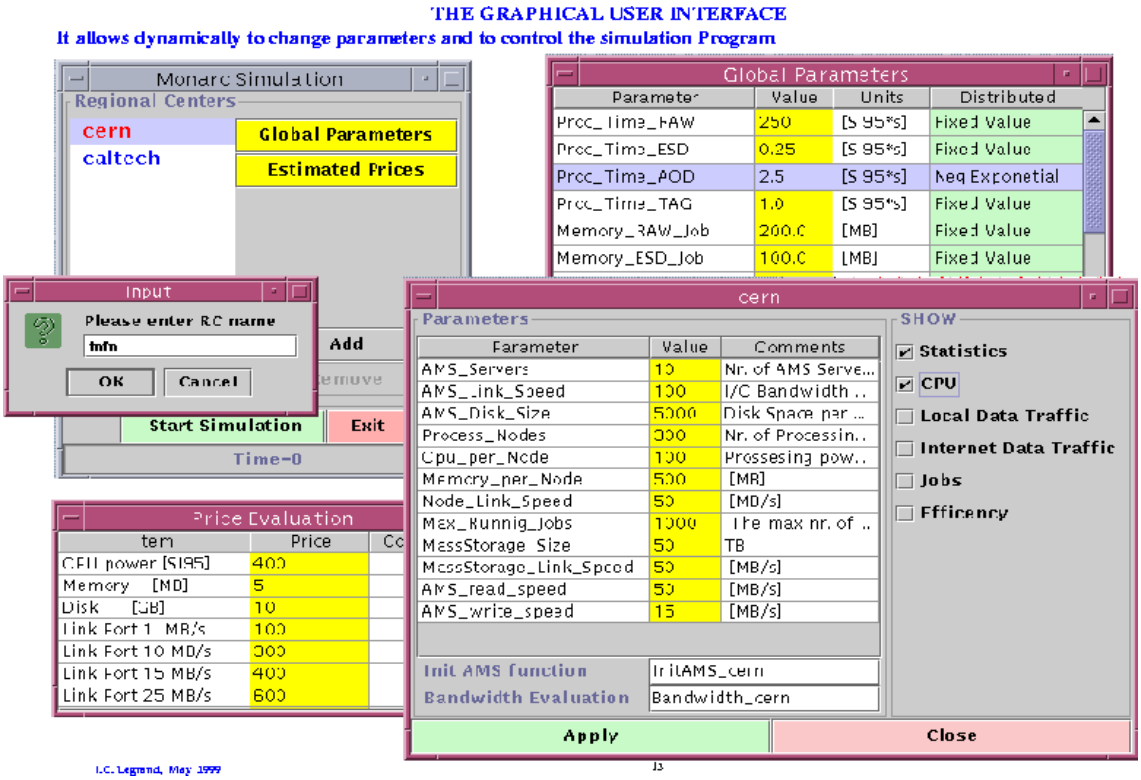


Figure 6: Showing the simulation tool GUI for building a model

Appendix A describes other example models built with the simulation tool.

2.5 Validation of MONARC simulation tool

The next step, which has already started, is to perform validation of the current version of the simulation program and its logical model. Behaviours of the simple CS2 model, and that of a single FARM, have been verified as described in Section 2.3. Similar tests are being performed, with the second version of the MONARC simulation program.

It was realised early in January, 1999, that it will be difficult to compare the simulation results with many of the existing experiments because our knowledge of their computing systems, however detailed, is inadequate to extract a proper "modelling" of the key parameters required for the simulation. The validation of the simulation program should be done by actually measuring the performance of the system with varying job stress and data access patterns. The precision of the simulation can only be increased by iteratively of refining the model of the system and its parameters with the actual measurements obtained from dedicated test-sites.

We require measurements of the key parameters of the distributed database, such as AMS read/write speeds with a single user and also with stress tests. A close discussion between the Analysis WG and Testbed WG has begun, to identify the key parameters and the dependencies of the parameters needed in the simulation program.

We also need to validate the correctness of the scaling behaviour, which is vital in making any predictions on a large scale distributed system. Another set of the required measurements concerns the local and the wide area network parametrisation functions. With those parameters in hand, and assuming that no significant changes to the logical model of the program will be required, the present simulation program provides a tool with which one can perform complex and meaningful studies of the architectures of the LHC-era computing systems.

2.6 Key parameters to be measured by the Testbed WG

In the process-oriented approach, response time function of the passive objects such as TAPE, DISK, CPU, NETWORK and AMS will define the precision and granularity of the simulation. The behaviour of AMS and NETWORK objects are of particular interest for MONARC. The response time function of AMS in fact depends on the various internal Objectivity/DB parameters such as the object size, object clustering and page size, as well as "use-case" parameters such as data modelling, mirroring, caching, and data access patterns. A stress test of multiple user jobs with respect to a single user job performance is also important in predicting the scalability of the system. Preliminary measurements have begun in the Testbed WG with a local federated database with a single AMS server (Chapter 5). A function, or functions, characterising the network performance should also be measured.

Different analysis access patterns, using data organised with different size objects and with different frequencies of following the TAG->AOD, AOD->ESD and ESD->RAW associations, will be used to test the more complex behaviours of the distributed systems, and compared with the predictions of the simulation program. Also, measurements of identical access patterns, but using multiple AMS servers connected with different network bandwidths, are foreseen.

Finding the key parameters and the key dependencies, and the scaling between those parameters in various models of distributed computing systems, will constitute a significant step towards validation of the second-round models of the LHC-era experiments. It has been planned that this stage will take place in the summer of 1999, and the project is well placed to meet the associated milestome.

2.7 Workplan

Some minor additions to the program are foreseen in the short term. A database replication protocol has to be implemented, and adding a possibility of physicists analysing data using the CPU available at their workstations rather than CPU at the Regional Centres. An important area of work in the next stage will be the development of an analysis framework, a software system which would allow systematic exploring of the multi-dimensional space of input parameters describing the LHC experiments computing systems architecture and evaluation of its performance.

It is important to mention some key elements that go into the system design, apart from the performance parameters of the components.

- Capacity limitations of disks
- Disk-tape migration
- Caching behaviours in the disk and tape systems (not associated with Objectivity itself)
- Data replication versus recomputation versus "refusal" (e.g. dump to tape and ship) mechanisms

- Prioritisation of tasks; quotas and relegation down to lower priorities once quotas are exceeded
- Queues and queue management: task aging, priority alterations and special operations (such as draining a queue of tasks at higher priority) on a daily or weekly basis.

Obviously not all of the above need be implemented at once, but a realistic model used optimally will include "stress" (full utilisation and some over-subscription) of the components. Hence some of the above elements will have to be taken into account, and a rough time-schedule for the implementation of these and perhaps other key aspects of a "realistic model" should be given, so that we can demonstrate that there are certain "baseline" (minimum) resource requirements.

2.8 Milestones

All milestones have been met, except for that which calls for "validation of the chosen tools with the Model taken from an existing experiment" in January 1999. It was realised that it will be difficult to compare simulation results with many of the existing experiments because our detailed knowledge of their computing systems, and more importantly the measurement of performance and throughput, is inadequate. The basic elements of the logical model and the applicability of Java tools have been verified with the CS2 model, although that system was certainly much simpler than any of the two example models (Reconstruction and Physics Analysis) which have been built with the second version of the simulation tools. We are currently of the opinion that the validation of the MONARC simulation program would be most reliably done by verifying the results of complicated access patterns with the measurements performed on dedicated test-sites, as described in section 2.5.

2.9 Deliverables

All existing information, including various presentations in which the logical model of the MONARC simulation tool has been has presented, some documentation, simple examples and demos are available on from MONARC WWW pages (MONARC->Simulation and Modelling->Status of the Simulation Software):

- CS2 and Objectivity/DB data server ->Demo
- single FARM built with first version of the MONARC simulation tool ->Monarc
- documentation (under development) -> Documents

The two example models built with the second version of the MONARC simulation tool are available from sunitp01.cern.ch/Examples. There exists a group account on that machine, and any MONARC member can either copy the files and run the programs on a local workstation with JDK1.2 installed, or one can run the program on sunitp01.cern.ch using an X-window server. A MONARC collaboration-wide working environment will be prepared shortly on a SUN workstation at CERN (monarc01.cern.ch) to allow participation of more people in developing and validating the program.

It is anticipated that significant improvements to the program documentation will be made during the Summer of 1999.

Chapter 3: Progress Report of the Architecture Working Group

3.1 Introduction

The task of the Architecture Working Group is to develop distributed computing system architectures for LHC which can be modelled to verify their performance and viability. To carry out this task, the group considered the LHC analysis problem in the "large". We started with the general parameters of an LHC experiment, such as

- CPU, disk, and mass (archival) storage requirements;
- the number and geographic distribution of collaborating institutions and individual physicists;
- the nature and size of the various analysis tasks;
- the availability of affordable network bandwidth;
- the cost evolution of basic technologies; and
- support requirements of the various activities.

From there we conducted detailed discussions about how the analysis task will be divided up between the computing facility at CERN and computing facilities located outside of CERN. We considered what kind of facilities will be viable given different analysis approaches and networking scenarios, what kind of issues each type of facility will face, and what kind of support will be required to sustain the facility and make it an effective contributor to LHC computing.

The viability of a given site architecture will ultimately be judged according to its ability to deliver a cost-effective solution to the LHC computing problem. Factors contributing to the viability (and the relative effectiveness) of a given architecture include

- overall equipment cost per unit of computing and data handling capacity
- person-power required to keep the system operational
- the system's operational efficiency, expressed in terms of
    - turnaround time for various user-tasks
    - "availability", meaning its ability to remain "up" for long periods
    - maximum workload (computing and I/O) sustainable per unit time
    - flexibility to respond to varying workloads at different times of the day and year

The judgement of the viability of a given architecture must be done in combination with a well-chosen "analysis strategy" that specifies the profile of, and limitations on, the users' analysis tasks, the partitioning of resources among the production-oriented, group-oriented and individuals' activities in the data analysis process, and the parameters controlling such decisions as recomputation or inter-site transport of portions of the data.

Once a set of viable architectures has been determined, the relative effectiveness of different implementations will need to be determined according to the minimum requirements for turnaround,

system MTBF, and the maximum allowable cost, as determined by the LHC experiments and the host-organisations at the sites. As indicated above, the evaluation of a system's effectiveness must be performed in combination with an intelligent strategy that aims at optimal use.

The general picture that has emerged from these discussions is:

- LHC computing is at such a scale in terms of computing capacity, data handling, and speed of inter-site communications that a worldwide effort to accumulate the necessary technical and financial resources will be required.

- There is considerable uncertainty with respect to the maximum affordable network bandwidth, especially where an ocean must be crossed. This means that it is necessary to develop several scenarios for how computing resources will be distributed and used.

- The optimal use of the financial and manpower resources of CERN and the collaborations and nations involved, and thus the best use by physicists and students of the LHC's opportunities for new physics, will be achieved, if we adopt a model based on a hierarchy of computing centres rather than a highly centralised model based at CERN. Centres are characterised by the range of services and the types of facilities they provide.

- At the top of this hierarchy is a large centre at CERN which has the capability to perform all analysis-related functions but not the capacity to do them completely.

- Below this is a collection of large, multi-service centres with capacities that are a significant fraction -- 10-20% -- of the part of the CERN Centre devoted to ATLAS or CMS. We call these Tier1 Regional Centres (RCs) and will summarise their characteristics below.

- Below the Tier1 RCs, there **may** be smaller centres, called Tier2 Centres, with capabilities and capacities which are more limited than the Tier1 centres but are nevertheless significant.

- These Tier1 and Tier2 centres **may** be further augmented by 'special purpose' centres or 'service centres' whose task may be to provide one or a few specific services to (typically) a single collaboration.

The primary motivation for a hierarchical collection of computing resources, called Regional Centres, is to maximise the intellectual contribution of physicists all over the world, without requiring their physical presence at CERN. An architecture based on RCs allows an organisation of computing tasks which may take advantage of physicists no matter where they are located. Next, the computing architecture based on RCs is an acknowledgement of the facts of life about network bandwidths and costs. Short distance networks will always be cheaper and higher bandwidth than long distance (especially intercontinental) networks. A hierarchy of centres with associated data storage ensures that network realities will not interfere with physics analysis. Finally, RCs provide a way to utilise the expertise and resources residing in computing centres throughout the world. For a variety of reasons it is difficult to concentrate resources (not only hardware but, more importantly, personnel and support resources) in a single location. A RC architecture will provide greater total computing resources for the experiments by allowing flexibility in how these resources are configured and located. A corollary of these motivations is that the RC model allows one to optimise the efficiency of data delivery/access by making appropriate decisions on processing the data. One important motivation for having such 'large' Tier1 RCs is to have centres with a critical mass of support people while not proliferating centres which would then create an enormous coordination problem for CERN and the collaborations.

There are many issues with regard to this approach. Perhaps the most important involves the coordination of the various Tiers. While the group has a rough understanding of the scale and role of the CERN centre and the Tier1 RCs, whether we need Tier2 centres and special purpose centres and what their roles should be has been worked on a little and is much less clear. Which types of centres should be created in addition to Tier1 centres and what their relationship to CERN, the Tier1 centres, and to each other should be will be a major subject of investigation over the next few months. Also, there are a variety

of approaches to actually implementing a Tier1 centre. Regional centres may serve one or more than one collaboration and each arrangement has its advantages and disadvantages.

There are also a number of higher-level issues that are complex, and heavily dependent on the evolution of system and system-software concepts, in addition to the technology evolution of components. It is likely that by LHC startup, efficient use of the hierarchy of centres will involve their use, to some extent, as if they were a single networked system serving a widely distributed set of users. From the individual site's point of view, this "one-distributed-system" concept will have to be integrated with, or traded off against, the fact that the site will be serving more than the LHC program, and often more than one LHC experiment.

To keep its discussions well grounded in reality, the group has undertaken the following tasks, which are described in the MONARC Project Execution Plan (PEP):

1. A survey of the computing architectures of selected existing HEP experiments;
2. A survey of the computing architectures of experiments that are starting data-taking now or in the next year or so;
3. Discussions and meetings with representatives of proposed Regional Centre candidate sites concerning their proposed level of services and support, architecture, and management;
4. Technology evaluation and cost tracking; and
5. Network performance and cost tracking.

Items 1 and 2 help us develop models to input to the Simulation and Testbed Working groups. Item 3 is essential to ensure that the proposed models of distributed computing are "real" in the sense that they are compatible with the views of likely Tier1 RC sites. Items 4 and 5 keep model building within the boundaries of available technology and funding.

3.2 Results from the Last Year

This year, the Architecture Working Group has produced three documents that have been submitted to the full collaboration and are summarised below. The plans for a fourth document are presented.

- Rough Sizing Estimates for a Computing Facility for a Large LHC Experiment, Les Robertson [8];
- Report on Computing Architectures of Existing Experiments, V.O'Dell *et al.* [9];
- Regional Centers for LHC Computing, Luciano Barone *et al.*[10]; and
- Report on Computing Architectures of Future Experiments (in progress).

3.2.1 Report on Computing Architectures of Existing Experiments [9]

This survey included:

- all four LEP experiments at CERN;
- CDF and D0 in 'Run I' (1992-1995) at Fermilab;
- ZEUS at DESY;
- the CERN Fixed Target Experiments NA48 and NA45; and
- the Fermilab Fixed Target Experiments KTeV and FOCUS.

The main conclusion from this report is that the LHC experiments are at such a different scale from the surveyed experiments and that technology has changed so much since some of them ran, that LHC experiments will need a new model of computing. We can, however, derive valuable lessons on individual topics and themes.

Some of the most important lessons on the computing architectures were:

- ***Scale***: LHC experiments will require 60 times more CPU and will generate 10 times more data than CDF anticipates for 'Run II' (2000-2003).
- ***Distribution***:
    - Implementing effective distributed computing systems is not simple and depends critically on good advanced planning. If experiments do not plan from the start to distribute their computing, it does not happen. Event simulation, where the data transfer requirements are relatively modest, is the area which has been most widely and successfully distributed.
    - Support and continuity at remote sites was identified as a major problem for distributed computing.
    - Maintaining the code base and calibration constants at remote sites was a major challenge.
    - Hardware and operating system differences between the central facility at CERN and the remote sites were also sources of problems.
- ***Planning***: An extensive analysis of the early planning for LEP computing indicated a definite tendency to underestimate the resource requirements by a large factor, in some part due to political and budgetary considerations.

3.2.2 Rough Sizing Estimates for a Computing Facility for a Large LHC experiment [8]

This document was prepared by Les Robertson of CERN IT Division. It attempts to summarise a rough estimate of the capacities needed for the analysis of an LHC experiment and to derive from them the size of the CERN central facility and a Tier1 Regional Centre. The information has been obtained from estimates by CMS and cross checked with ATLAS and with the MONARC Analysis Working group. Some adjustments have been made to the numbers obtained from the experiments to account for overheads that are now measured but were not when the original estimates were made. While the result has not yet been reviewed by CERN management, it currently serves as our best indication of thinking on this topic at CERN so we are using it as the basis for proceeding.

Current studies of the full simulation and reconstruction of events at full LHC luminosity tend to indicate that the requirements estimates in this report are not overestimates, and additional work may be required to reduce the computing time per event to these target levels. The report also does not take into account the needs for full simulation and reconstruction of simulated events, which must be processed, stored and accessed at Regional Centres or at local institutes, if not at CERN.

It is assumed that CERN will NOT be able to provide more than about about 1/2 of the aggregate computing need for data recording, reconstruction, and analysis of LHC experiments. This is exclusive of Monte Carlo event simulation and reconstruction of simulated events. The remainder must come from elsewhere. The view expressed by the author is that it must come from a 'small' number of Tier1 Regional Centres so that the problems of maintaining coherence and coordinating all the activities is not overwhelming. This sets the size of Tier1 RCs at 10-20% of the CERN centre in capacity.

Table 3 summarises the total CPU, disk, LAN throughput, tapes, tape I/O, and the number of 'boxes' that will have to be operated to support the data analysis of a large LHC experiment as the LHC moves from turn on around 2005 to full luminosity operation a few years later.

Table 3: Summary of required installed capacity

year	2004	2005	2006	2007
total cpu (SI95)	70'000	350'000	520'000	700'000
disks (TB)	40	340	540	740
LAN thr-put (GB/sec)	6	31	46	61
tapes (PB)	0.2	1	3	5
tape I/O (GB/sec)	0.2	0.3	0.5	0.5
approx box count	250	900	1400	1900

3.2.3 Regional Centers for LHC Computing [10]

Based on Les Robertson's estimates and the issues raised about the problems with distributed computing in the past by the survey Computing Architectures of Existing Experiments, we developed a framework for discussing Regional Centres and produced a document which gives a profile of a Tier1 Regional Centre.

This profile is based on facilities (and the corresponding capacities) and services (capabilities) which need to be provided to users. There is a clear emphasis on data access by users since this is seen as one of the largest challenges for LHC computing, especially where parts of the data may be located at remote sites, and/or resident in a tape-storage system.

It is important to recognise that MONARC cannot and does not want to try to dictate the implementations of the Regional Centre architecture. That is best left to the collaborations, the candidate sites, and to CERN to work out on a case by case basis. MONARC wants to provide a forum for the discussion of how these centres will get started and develop and can assist in the effort to locate candidate centres and bring them into the discussion.

The report describes the services that we believe that CERN will supply to LHC data analysis, based on the physics requirements. These include:

- Online data acquisition and storage
- Possible data preprocessing before reconstruction
- First data reconstruction
- Support for data analysis on-site by a group of a few hundred physicists per experiment

CERN will have the original or master copy of the following data:

- the raw data;
- the master copy of the calibration data; and
- a complete copy of all ESD (reconstructed), AOD (DST), and TAG (thumbnails, nanoDST) data.

The regional centres will provide:

- all technical services and data services required to do physics analysis;
- all of the AOD, TAG and calibration data;
- a significant fraction of the raw and ESD data;
- caching or mirroring of all calibration constants;
- excellent network connectivity to CERN and to the users in the region principally served by the centre;
- human resources to develop or share in the development of common maintenance, validation, and production software with CERN and the collaboration;
- a fair share of the post-reconstruction or re-reconstruction processing;
- human resources to work on common projects with CERN and the collaborations;
- services to members of other regions on a best effort basis;
- excellent support services, training, documentation, and trouble shooting at the regional centre and at remote sites served by it.

Support is called out as a key element in achieving the smooth functioning of this distributed architecture. It is essential for the regional centre to provide a critical mass of user support. It is also noted that since this is a commitment that extends over a long period of time, long term staffing, a budget for hardware evolution, and support for R&D into new technologies must be provided.

3.2.4 Report on Computing Architectures of Future Experiments

Work on this report is just beginning. It will include a study of BaBar at SLAC, CDF and D0 'Run II' at Fermilab, COMPASS at CERN, and the STAR experiment at RHIC. The approach will be to survey the available public literature on these experiments and to abstract information that is particularly relevant to LHC computing. This can be supplemented where required by discussions with leaders of the computing and analysis efforts.There will not be an attempt to create complete, self-contained expositions of how each experiment does all its tasks. We will have a 'contact-person' for each experiment who will be responsible for gathering the material and summarising it for the report. Most of these 'contact-persons' are now in place. There will be an overall editor for the final report.

3.2.5 First meeting of Regional Centre Representatives

On April 13, there was a meeting of representatives of potential Regional Centre sites. It was felt at this point that we had made good progress in understanding the issues of how Regional Centres could contribute to LHC computing and it was now time to share this with possible candidates, to hear their plans for the future, and to get their feedback on our discussions. The three documents discussed above, which had been made available in advance of the meeting, were summarised briefly. We then heard presentations [11] from IN2P3/France, INFN/Italy, LBNL/US(ATLAS), FNAL/US(CMS), RAL/UK, Germany, KEK/Japan(ATLAS), Russia/Moscow.

The general tone of the meeting was very positive. Some organisations, such as IN2P3, expressed

confidence that their current plans and expected funding levels would permit them to serve as Tier1 Regional Centres. Others are involved in developing specific proposals that can be put before their national funding agencies within the next few months or a year. Still others have recently begun discussions within their High Energy Physics community as a first step in formulating their plans. In general, the representatives indicated that their funding agencies understood the scale of the LHC analysis problem and accepted the idea that significant resources outside of CERN would need to be provided. We can conclude from the meeting that there are several candidates for Regional Centres that have a good chance to get support to proceed and will be at a scale roughly equivalent to MONARC's profile of a Tier1 RC. It was also clear that there would be several styles of implementation of the Regional Centre concept. One variation is that several centres saw themselves serving all four major LHC experiments but others, especially in the US and Japan, will serve only single experiments. Another variation is that some Tier1 Regional Centres will be located at a single site while others may be somewhat distributed themselves although presumably quite highly integrated.

MONARC expects to follow up this first meeting with another meeting towards the end of 1999. We will have a draft of the final document on Regional Centres available before the meeting for comment by the Regional Centres' representatives. In addition to hearing plans, status reports and updates, we hope to have discussion on the interaction between the Regional Centres and CERN and between the Centres and their constituents. We also plan to be able to present to the Regional Centres representatives MONARC results which may help them develop their strategies.

3.2.6 Technology Tracking

The main initiative in technology tracking was to take advantage of CERN IT efforts in this area. We heard a report on the evolution of CPU costs by Sverre Jarp of CERN who serves on a group called PASTA which is tracking processor and storage technologies [12]. We look forward to additional such presentations in the future.

3.3 Goals and Milestones for the July-December period

mid-July	**Complete the Report on Computing Architectures of Future Experiments**
end-'99	**Produce the final document on the Regional Centres**
end-'99	**Consider the strategic objectives of MONARC modelling[1]**
end-'99	**Develop cost evolution model for networking**
end-'99	**Develop cost evolution model for CPU, disk and mass storage systems**

[1] *In a first phase it is important that the Simulation and Analysis WGs develop confidence in the detailed validity of the MONARC simulation tools on small systems with all activities under our control. This means convincing ourselves, and others, that we really have produced working models of the distributed computing process. In parallel, or as soon as possible afterwards, the Architecture WG should consider the "in the large" issues that MONARC needs to model. For example, how will priorities be determined between large-scale production jobs, group analysis and work by individuals? What commitments will CERN and the Tier 1 Regional Centres need to make with each other?*

Chapter 4: Progress Reports of the Analysis Process Design Working Group

The task of the Analysis Process Design Working Group was to develop a preliminary, but nevertheless feasible, design of the Analysis Process in the LHC era.

4.1 Results from Phase 1 of the project

The principal results obtained are presented below, following the organisation of the PEP subtasks. Further details may be found on the Analysis Process Design Working Group's Web page[13].

The "user-requirements" approach has contributed most to the generation of the first Analysis Process scenarios for LHC experiments to go into the MONARC simulation in Phase 1.

Limited studies of scenarios heavily influenced by available resources have been performed. More detailed studies will be undertaken as we receive feedback from Simulation and Architecture WGs.

The first approximation to parameters of the Analysis Process scenarios, their values, ranges and later distributions will be refined through successive iterations of simulation and progressively more detailed configurations of resources.

4.1.1 Analysis of contemporary production and analysis procedures.

A survey of the Analysis Processes of experiments taking data now and in the next three years was performed (Phase 1B, subtask 4.4.1). Inspection of experiments at LEP and at FNAL (including RUN-II) revealed methodologies highly tuned to their physics channels, backgrounds, and detector performances which employ mature technologies for most of their installed computing resources [14].

The dimension of the computing resources needed, the dispersion and number of the analysing physicists, and mainly the distributed approach to the analysis, set a scale for technology and architectures which requires a distributed and coherent design from the beginning of LHC era.

Although we may be guided by past and present experience, particularly for the way an individual physicist user needs access to the relevant data during analysis, there is evidence that the techniques used cannot easily scale to LHC.

Our survey showed that a new approach to the Analysis Process for LHC is needed in order to cope with the size, constraints and distributed requirements of the future experiments.

Following the "user-requirements" approach, we considered the specific physics goals at LHC, with the anticipated trigger, signal and background rates and the data volume to be recorded and analysed. Thus there is a firm basis in the anticipated LHC physics for the initial parameters and distributions used to design our Analysis Processes.

We concluded that some hierarchy has to be built into the Analysis Process from the beginning. Our model is that the experiment(s) define "official" Physics Analysis Groups (PAG), developing algorithms, refining calibrations and studying particular channels. We start with each PAG requiring access to a subset of the order of a few percent of the accumulated experimental data (10^9 events per year at full Luminosity). The Analysis Process follows the hierarchy: Experiment-> Analysis Groups-> Individuals. Coordination between the PAGs and between the Individual physicists is needed; the logical and physical overlap in data sample storage, event selection and trigger specification is most relevant for our studies.

A typical Analysis Group may have about 25 active physicists, spread in different (and perhaps overlapping) World Regions. Table 4 gives a summary of the "Group approach" to the Analysis Process.

Table 4: Summary of the "Group Approach" to the Analysis Process

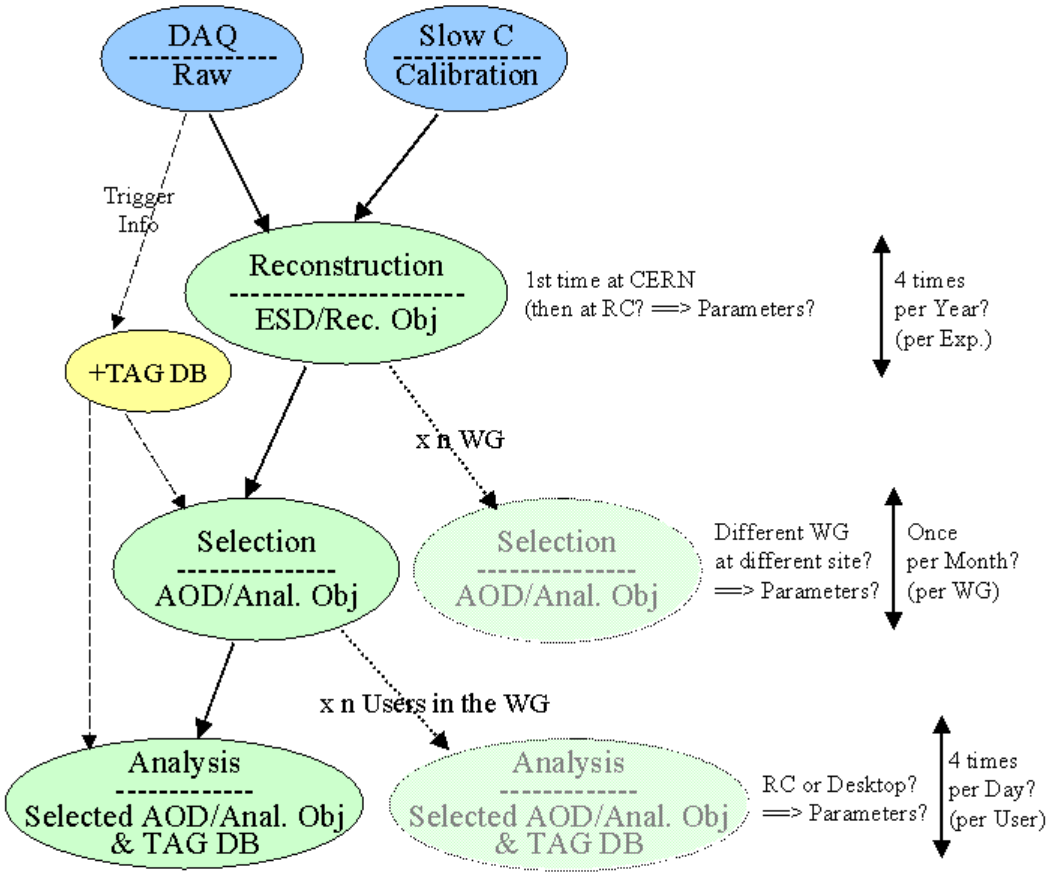
LHC Experiments	Value	Range
No. of analysis WGs	20/Exp.	10-25/Exp.
No. of Members of WG	25	15-35
No. of RCs (including CERN)	8/Exp.	4-12/Exp.
No. of Analyses per RC	5	3-7
Active time of Members	8 Hour/Day	2-14 Hour/Day
Activity of Members	Single RC	More than one RC

The above considerations lead to a Group approach for the reduction of the data-sample (using a common facility) and to a local (Regional Centre) approach for individual activities[15][16].

4.1.2 Identify user requirements.

The possible initial phases of the Analysis process were investigated and some preliminary data sets accessed during the various steps were defined (Phase 1B, subtask 4.4.2). Given the Group/Individual Model above described, the analysis process can be represented as in the following scheme:

Analysis Process



Raw Data : On Tape, at CERN and at RC
ESD/Rec.Obj : On Tape at CERN, on Disk at RC (Including CERN RC)
for the samples needed by analysis at a given RC
AOD/Anal. Obj : On Disk
Selected AOD/Anal. Obj : On Disk
TAG DB : On Disk

The Analysis steps therefore are:

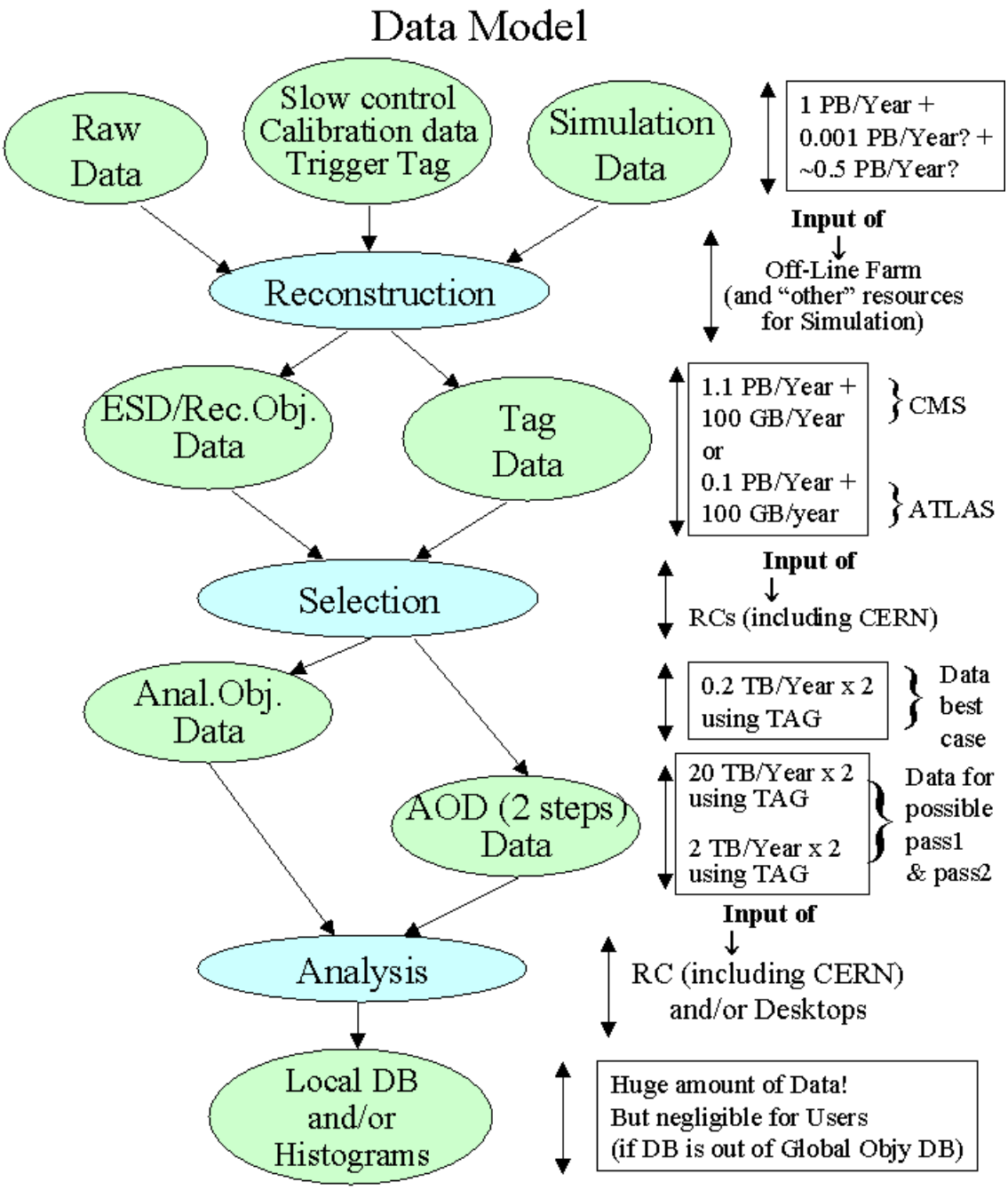
1. **Reconstruction**. The process has to be performed at the Off-line Farm at CERN for all the WGs. This in fact means the filling process of the Objects in the Object Database. Possible re-reconstructions are one of the parameters of the Model, including their possible location (either at CERN or partially at Regional Centres). The so called ESD are produced during these processes. Data produced are of the order of 100 TBytes/year and they reside also in the Regional Centres for the part needed by the "regional" activities. Disk storage media are foreseen for this type of (output) data sample. Tapes may be also needed, depending on cost and technology evolution.

2. **Selection**. The data-sample is selected and reduced in size and number of events, eventually in two subsequent Passes triggered by individual Groups, in order to provide the database information relevant for the analysis. This is the more relevant and delicate process, producing the so called AOD. Data produced are evaluated for different selections. The results are strongly dependent on the number of "passes" and designed activities, ranging from final 2TB/year to 0.2 TB/year for the whole experiment. Disk storage at the Regional Centres should be the choice for these data samples.

3. **Analysis**. The group-produced data sample is inspected by individual components so as to obtain physics results. Simulated data will also be used during this process. Data samples will certainly be stored on disks and the jobs will run at the Regional Centres. The possibility of undertaking part or all of this activity on Institute resources (Desktops) is under evaluation.

4. **Simulation.** The model includes the distributed production of Monte Carlo event simulation, and the reconstruction. The current practise in HEP experiments of distributing and coordinating simulation is well established: this fact led us to retain distributed simulation in the LHC computing model. Group simulations may use dedicated (Tier2 Regional Centres) and/or distributed resources available to the Collaboration.

The following diagram shows one of the possible implementations of the Analysis Model. The initial CTP differences between ATLAS and CMS are here expressed as an example of how the "Selection Pass" can lead to quite different Models and therefore to a spread of architectures for Analysis Design.



4.1.3 Identify feasible models to be simulated.

The identification of a first Analysis Model for an LHC experiment was performed in order to provide input for simulation (Phase 1B, subtask 4.4.3). The architecture has been designed taking into account the many parameters and the constraints for the steps of the analysis, some of them being reported in the following table:

Table 5: Parameters for the Reconstruction Step (per Experiment)

Parameter	**Value**	**Range**
Frequency	4/ Year	2-6/ Year
Input Data	1 PB	0.5-2 PB
CPU/event	350 SI95.Sec	250-500 SI95.Sec
Data Input Storage	Tape (HPSS?)	Disk-Tape
Output Data	0.1 PB	0.05-0.2 PB
Data Output Storage	Disk	Disk-Tape
Triggered by	Collaboration	1/2 Collab. - 1/2 Anal. WGs
Data Input Residence	CERN	CERN + some RC
Data Output Residence	CERN + some RC	CERN + RCs
Time response	1 Month	10 Days - 2 Months
Priority (if possible)	High	-

Table 6: Parameters for Selection Pass 1 (per Physics Analysis WG)

Parameter	**Value**	**Range**
Frequency	1/ Month	0.5 - 4.0 / Month
Input Data	100 TB	20-500 TB
CPU/event	0.25 SI95.Sec	0.10 - 0.50 SI95.Sec
Triggered by	WGs	-
Data Input Storage	Disk	1/2 Disk - 1/2 Tape

Output Data	1 TB	0.5 - 10 TB
Data Output Storage	Disk	Disk and Tape
Data Output Residence	RCs	Specific RC + Other RCs
Time response	3 Days	1 Day - 7 Days

Table 7: Parameters for Selection Pass 2 (per Physics Analysis WG)

Parameter	**Value**	**Range**
Frequency	1/Month	0.5 - 4.0 / Month
Input Data	Output of Pass 1	
CPU/event	2.5 SI95.Sec	1.0 -5.0 SI95.Sec
CPU Residence	RC	RC + Desktops?
Triggered by	WG	-
Output Data	0.1 TB	0.05 -1.0 TB
Data Output Storage	Disk	Disk
Time response	1 Day	0.5 - 3 Days
Note: Desktop = Institute resources		

Table 8: Parameters for Analysis Activities (per Physicist)

Parameter	**Value**	**Range**
Frequency	4/Day	2 - 8 /Day
CPU/event	3.0 SI95.Sec	1.5 -5.0 SI95.Sec
Triggered by	User	1/3 WG - 2/3 Users
Time response	4 Hours	2 - 8 Hours

The most relevant parameters are:

- CPU power for each of the analysis phases/steps.
- Frequency of each of the operations started by the Groups or by Individuals. Major reconstructions have been limited to a few units per year, selection by the groups is in the range of once per month and the job submission by individuals must be coherent with working hours and time zones.
- Data set residence and accesses; this is a first approach to the "data model", which is one of the major issues of the Computing model for LHC. The Federated DB is at the basis of the process and gives constraints to the distributed access to the data, imposing a careful study of where the different data sets should be. The Analysis Process has to take care of the behaviour of data replication and retrieval.
- Distribution of analysis tasks onto the different resources, ranging from the CERN Regional Centre to the other Regional Centres and eventually to the "desktops".
- Parameters and spread of values for the quoted items (and all of the others not quoted here) have been proposed, often with possible distribution functions [17].

Some of the ranges and eventually some of the possible combination of them can lead to unfeasible models, either in terms of required resources or in terms of turnaround responsiveness. Studies were performed to establish constraints on the parameters to avoid unfeasible approaches [18]. For example some of the proposed time responses for a given Analysis may lead to required resources (either CPU power or data storage) that remains "idle" for the most of the time. This is a clear indication of an unfeasible Model. Another example might be a Model that meets the requests of the expected Analysis needs, but cannot be afforded because of the associated networking resources (latency, round trip time for database transactions and bandwidth).

4.1.4 Elaborate policies, priorities and schedules for different models.

The task aimed to establish how the different schemes of access to the collaboration resources could be mapped into the analysis jobs needs. (Phase 1B-1C, subtask 4.4.4).
Implementing priorities, schedules and policies in a distributed Analysis approach should include them directly into the architecture. Performing the Analysis at LHC in an hierarchical Experiment -> Group -> Individual implementation is a starting point.
As already said in other parts of the Progress Report, there is also the need for a definition of rates and percentages of accesses to the hierarchy of data (TAG -> AOD -> ESD -> RAW) for any of the Analysis steps. Having understood that criteria and priorities (or having *a priori* defined them because of resources constraints), can be explicitly implemented into different Models in order to evaluate performances and costs.
The task is till under development and in particular what is needed for the first delivery is the identification of possible resource architectures and the mapping of the data and job analysis into them.

4.1.5 Identify key parameters to evaluate simulated models.

Establishing a preliminary set of parameters for the evaluation of the models simulated is the first goal of this task (Phase 1B-1C, subtask 4.4.5). The process is under way, the major issue being the identification of clear, even if preliminary, parameters that can classify the models into the planned resources. Some propositions have been advanced, like obvious parameters such as occupancy of CPUs, of storage, of network, etc., and less obvious parameters like number of Regional Centres, network use, management of the system, coordination etc. [19]. The global cost of a given Analysis Model is one of the major elements for the evaluation and it requires a careful inspection of the technologies trends by the Technology Tracking Group in order to produce prices scales. Another very important key parameter is the isolation, via the simulation, of possible bottlenecks of the architecture/infrastructure of the RC

models. Moreover there are also some parameters that can only be evaluated taking into account the whole Computing System Design, like the ability to respond in due time to an "urgent", medium complexity, analysis.
More informations about this important deliverable can be found also in Chapter 6.

4.2 Workplan for the second part of the project.

During the current Phase 1C and the next Phase 2 of the PEP there will be activity on a large number of issues, some covering both phases and some only foreseen for the last phase. Below is a list of these issues.

- Understand better the Data Model. Access to the data and their availability in the database is one of the major issues, as well as the architectural behaviour of the object database. Phase 2 at the beginning will need some preliminary results, and further results will be requires before the end of the project.

- Monte Carlo production and data use while doing analysis must be better quantified for incorporation into the analysis process.

- Prioritisation and coordination of the analysis jobs have to be modelled in order to get complete simulations of the Computing Models for LHC. This task is an important issue but it is also difficult, being experiment dependent, and therefore is foreseen for the end of Phase 2.

- Definition of evaluation criteria for the simulated results on different analysis approaches are being addressed now, but evaluations will be available near the end of the project (phase 2).

- Different strategies of analysis, not only in term of parameter range, have to be designed. This process in under way (during phase 1C) and will need a resumed study during Phase 2.

- The role of the Desktops, or more generally of the Institutions' resources must be incorporated into the Model. This issue is already under study and first results are due for Phase 1C, while more refined roles have to be planned for Phase 2.

- Mapping the Analysis Process into the distributed architecture of the Regional Centres is already being addressed and will require iteration with the Architecture WG and with the Simulation WG during phase 2.

Chapter 5: Progress Reports of the Testbed Working Group

The aim of the Testbed Working Group is to provide the measurements of some of the key system parameters which govern the behaviour and the scalability of the various models of the distributed computing system. The measurements have to implement the "use-cases" of the data analysis models and the data distribution models which are defined by the Architecture Working Group and the Analysis Process Design Working Group. The result of the measurements will then be fed back to the Simulation Working Group to check the validity of the simulation models and the selection of the key parameters.

A simple computing and network architecture was implemented by the end of January '99 (a month later than foreseen), then many more sites followed to set up the machines devoted to the measurements in February. Resources devoted to these measurements at CERN as a "central facility" were delivered at the beginning of April.

In parallel, a suitable set of test applications was identified. Actual measurements were started in April '99 and first preliminary results were obtained at the beginning of May '99. Implications of the results are now being discussed within the group, in conjunction with the Simulation Working Group.

In the meantime, up-to-date information about the performance of Objectivity/DB has been collected, mainly from RD45, BaBar and GIOD. Recommendations for our test environments have been defined accordingly:

5.1 Configuration of the testbed environments

To study the distributed aspect of object database such as database replication, the following list of basic software combinations has been selected as a reference environment of the testbed:

- SUN Solaris 2.6, with compiler C++ v. 4.2;
- Objectivity/DB v. 5.1 with /DDL, /c++, /stl, /FTO and /java options.

In addition to the above list, other platforms used in the study includes Intel PC's running Windows NT v. 4, with Visual C++ v. 5.0, and Linux.

For the "use-cases" software and the data modelling, the following set of applications have been identified as suitable for studying various system parameters. In fact, some of them are already used in other Objectivity/DB benchmark measurements. These applications have been tailored and tuned to reflect the various data models of our interests.

- **ATLFAST++** [20]: ATLAS fast simulation and analysis program has been ported into the LHC++ environment. The program can populate an Objectivity database with events, ~40 KB each, following the Tag/Event data model, as well as access data and fill histograms;
- **GIOD** project [21]:a target of ~1,000,000 CMS fully-simulated events, ~250 kB each for a total of ~250GB of data, are being produced and stored in an Objectivity database. C++ and Java code for populating the database from CMSIM FZ files, accessing the data, reconstructing physics objects (tracks, clusters, etc...), performing di-jet analysis is also available, as is an event viewer in Java 3D;
- **ATLAS** 1 TB milestone program [22]: fully-simulated ATLAS events with GEANT3 have been ported into the LHC++ environments to populate an Objectivity database. A fast read-access program has been developed and tested to study the clustering of the raw data structure of the ATLAS data;
- **CMS** test beam runs [23].

5.2 Sites and resources available for the tests

The group plans to involve all the participating sites with the above environments to test the performance of globally distributed databases. A dedicated facility at CERN has been set up with the required software. In addition, the following facilities are now available as the testing environments (Table 9).

Table 9: Showing the facilities available at the MONARC collaborating institutes

CERN	SUN Enterprise 450 (4*400MHz CPUs, 512MB memory, 4 UltraSCSI channels, 10*18G disks) Use of mass storage management (HPSS) facility is being planned.
Caltech	HP Exemplar SPP 2000 (256 CPUs, 64 GByte memory) HPSS (600 TB tape + 500 GB disk cache) HP Kayak PC (450 MHz, 128 MB memory, 20 GB disk, ATM) HP C200 (200 MHz CPU, 128 MB memory, 10 GB disk) Sun SparcStation 20 (80 GB disk) Sun Enterprise 250 (dual 450Mhz CPUs, 256 MB memory[1]) Micron Millennia PC (450 MHz CPU, 128 MB memory, 20 GBytes disk) ~1 TB RAID FibreChannel disk (to be attached to the Enterprise 250[1]) *[1] shortly to be ordered*
CNAF	SUN UltraSparc 5, 18 GB disk
FNAL	ES450 Sun Server (dual CPUs), 100 GB disk + access to a STK Silo
Genova	SUN UltraSparc 5, 18 GB disk
KEK	SUN UltraSparc, 100 GB disk
Milano	SUN UltraSparc 5, 18 GB disk Access to non dedicated facilities is available at CILEA: to a SUN system similar to the dedicated one and to the HP Exemplar SPP 2000 of the Centre, for agreed tests.
Padova	SUN UltraSparc 5, 117 GB disk + SUN Sparc 20, 20 GB disk
Roma	SUN UltraSparc 5, 27 GB disk
Tufts	Pentium II 300 MHz PC, 12 GB disk (+ Pentium-II 400 MHz PC, 22 GB disk, in July)

A network test topology, giving access to the network advanced services, is being set up on the layout provided by the Italian project Garr-it2. For the network connectivity at Caltech, NTON (OC12->OC48), CalREN-2 (OC12), CalREN-2 ATM (OC12) and ESnet (T1) will be utilised. A link between KEK and CERN will be a public link of NACSIS 2 Mbps line as well as a dedicated 2 Mbps satellite ATM virtual link of Japan-Europe Gamma project.

5.3 Measurement of key parameters

The set of measurements to be carried out in the Testbed WG has been defined in agreement with the other working groups, and particularly with the Simulation WG, as described in Section 2.5. The behaviour of the database server needed to be defined with a response time function for read and write

transactions to the database.

The response time function is a combination of various transaction overheads, which are internal to the database software, and the CPU speed and the data transfer speed, which will vary from system to system. It also depends on the job load as number of jobs increase on the system.

A set of measurements have been performed using the ported ATLFAST++ program with a local federation [24]. A total number of 100000 events (~4 GB) are stored on a SUN UltraSparc 5 workstation . The measurements are made on:

- stress tests consisting of submitting an increasing number of concurrent jobs;
- timing tests of changing the access pattern from sequential access to non-sequential access;

Preliminary results are obtained from these tests and the the group is now trying to understand the results to give feedback to the Simulation Working Group.

In the stress test, the program read both Tag and Event attributes from the same Objectivity containers with a small amount of CPU cycle used in the analysis program. A linear dependence of the execution time on the number of concurrent jobs is shown until a divergent behaviour starts. By tuning the cache size of the Objectivity client, the divergence disappears and the system behaves linearly up to 60 or so concurrent jobs on SUN Ultra 5 with 128 MBytes of memory. The cause of the divergent behaviour in the initial cache size is now being investigated, but the system is proven to behave linearly for a reasonable number of concurrent jobs.

In the timing tests, CPU time and the wall-clock execution time for a single job have been measured. The CPU time per MBytes read is the same for sequential and non-sequential data access, which is consistent with the previous study of the CPU requirements of database I/O transaction[25]. The wall-clock execution time of the job for reading selected event sample is slower than the sequential reading, which suggests an additional overhead in data I/O or in database transaction. For reading a half of the event sample in the database, the wall-clock time difference is about 80 seconds per 50,000 events. This study will give us the knowledge of the impact on system performance due to efficient or inefficient use of data modelling and access patterns. The exact cause of the wall-clock time overhead is now being investigated.

In another set of tests using the ATLAS 1TB milestone data model, a performance of the client-server configuration of Objectivity has been measured[26]. A preliminary study of the result suggests that the number of concurrent jobs and the use of client-server configuration changes the behaviour of the system from CPU-bound state to IO-bound state. A set of studies for different hardware configurations and the networks, including the test over WAN is planned.

In parallel to these measurements, previous results (see [25] and references quoted therein) have been analysed and plans of tests are being defined to evaluate how the system performances depend on:

- size of objects and clustering,
- data access patterns,
- the data replication overhead in the federated database,
- the different configurations of CPU and storage distribution.

Specific measurements are intended to cross-check simulation results or to parameterise complex system components [27].

In particular, regarding system performances over WAN, advanced network services (like QoS, multicast, etc...) have been investigated [28], mainly with respect to:

- **reliability**: how to guarantee access to the data;
- **efficiency**: how to obtain adequate allocation of resources;
- **network monitoring and control:** how to adapt in a dynamic way the computing model and the available network resources.

5.4 Workplan

Further studies are planned to better understand how the system performance depends on global parameters, the data server configuration and the data model (see 2.4.1). Local tests will be repeated using AMS, to the extent it could be useful with the present version of AMS: indeed the next version will be multi-threaded and thus some relevant performance figures are expected to change, especially for concurrent access.

We also plan to identify and list the key parameters which are internal to the Objectivity performance tuning, which will give a universal guideline for all testbed measurements on Objectivity.

A thorough comparison of access to local and remote data is also needed, in order to suitably parameterise the network effect on system performances.

Tests on data replication over several distributed federations world-wide are being planned to measure the feasibility of two major tasks: distribution of data produced centrally (like calibrations) and centralisation of data generated remotely (like MC events).

Furthermore, we plan to set up a "use-case" where a number of different "virtual" users will access the same event sample (i.e. 100000 events), and will perform concurrently their own analysis on personal collections of events, or on Generic Tags in which they can save the main attributes used in their analysis. They will also have the possibility of re-clustering the events according to specific criteria and performing their analysis on the re-clustered samples. This test will give some hints on the most effective ways of working for the end-user (less time consuming, more performing).

Regarding the quality of the network services, the aim is to define a set of minimal/critical requirements and extend the test network to the European Quantum experimental layout, to the NACIS and Japan-Europe Gamma networks layout, and to the ESnet layout.

Chapter 6: Workplan and Schedule

We review in this chapter the planning, resources and schedule of the project as detailed in the PEP. The status described is as of the end of May'99.

The working groups have met regularly, with the Architecture WG meeting every 2 weeks, and the others according to demand. In the last 2 months the Simulation and Analysis WGs have had joint meetings. In addition general meetings have been held, with good participation, at a frequency of about 2 every 3 months. For all meetings video-conferencing has been used. Thus overall we have managed to overcome the problem of widely distributed human resources. Before mid-January 2 meetings with RD45 were held, and we continue a close collaboration.

Overall there has been a broad participation from the MONARC members, and indeed the collaboration has grown in the past months.

6.1 Progress Achieved and time-scales

A good amount of work was accomplished in the 2 months prior to the official LCB approval in December

1998. However various delays have resulted in a month shift in the time-scales set out in the PEP.

6.1.1 Review of Milestones

The Main Milestones set in the PEP till May were:

December '98	**Choose modelling tools**
December '98	**Hardware setup for testbed systems**
January '99	**Validate the chosen tools with a Model taken from an existing experiment**
March '99	**Complete the first technical run of Simulations of a well defined LHC-type Model[1]**
March '99	**Start measurements on testbed systems**
April '99	**Choose the range of models to be simulated**

[1]Here the meaning of the word "technical" is that the simulation is required to run with all the main ingredients needed for simulating an LHC Model; but the first realistic models were not scheduled for this time.

Allowing for the overall delay of a month all of these milestones have been fulfilled, except for that of 'model validation'. This has been due both to the limitations of manpower for modelling, and the fact that we have revised our ideas on the most effective way to validate the modelling tool. We intend now to perform the validation, which we regard as being extremely important, in close collaboration with the MONARC Testbeds WG.

Also for the Architecture WG the internal milestones were revised following the advice of the LCB. Consequently the group has given priority to the task of developing architectures for LHC, and developing guidelines for Regional Centres. Thus the 'Survey of existing experiments' has been completed just recently, and the 'Report on near future experiments' will be restricted in scope, and is due in mid-July.

The next main milestone, **which will complete Phase 1,** is in July 99.

July '99	**Completion of the first cycle of simulation of possible Model for LHC experiments. First classification of the Models and first evaluation criteria.**

We believe we will be in time for completion of this milestone.

6.1.2 Completion of Phase 1

The completion of Phase 1 requires:

- Runs of the Simulation to be performed by the various groups involved: at least CERN, Tufts, Bologna, Caltech.
- A first assessment of the results on the basis of a preliminary resources vs performance

evaluation

- A first identification of the sensitive areas of the Model (bottlenecks, resource consuming procedures etc.) as basis for further model generation.

It is clear that already at this stage the project has started requiring a much closer interaction between the WG's; the above summary about Phase 1 completion already makes this point quite evident.
The steering group mandate encompasses the coordination of the work between the WG's and this task is going to be crucial in the next months. Some reorganisation of the WG structure and individual mandates in MONARC could be decided for the next stage, if it will be deemed useful.

The Workplan for Phase 2 will be discussed in Section 6.3 and new more specific Milestones will be set there.

6.1.3 Review of the relations with other projects

The external projects with whom we have working relations were listed in Section 3.2 of the PEP. We repeat this list here with supporting comments.

- **Pre-LHC Experiments:**Contacts have been established in view of the related Report of the Architecture WG (see Chapter 3).

- **RD45:** The collaboration is going on well. MONARC has adopted Objectivity as the baseline Database system, on the basis on the choice made by the LHC experiments involved in the project. The experiment choice in turn was prompted by the work and advice of RD45.
  The measurements and the tests performed in the contest of the Testbed WG should however be well suited also for assessing different fall-back solutions if they are of interest for any of the experiments participating in MONARC.

- **HPSS:** The contacts are established, but the issue of HPSS is for Phase 2 and 3 and not for Phase 1.

- **GIOD:** The project is active and well advanced. Installations of the GIOD database have been made at CERN, Padova and Fermilab. Wide area tests of database clients are being carried out or planned between Caltech, CERN, Fermilab, SDSC, Padova. Several of the CMS and RD45 milestones have been met by the project, including the demonstration of >100 MBytes/second writing to Objectivity and successful tests with up to 240 database clients. Current work with MONARC is focusing on the wide area deployment of the GIOD database, which contains fully simulated, realistic LHC event data.

- **Technology Tracking:** The team resumed his work and first results from PASTA were presented in MONARC meetings and taken into account in the documents of the Architecture WG. Final results should be available in time for Phase 2 simulations.

- **Existing Computing Centres:** Good contacts are established and a first organised information exchange has taken place with the meeting of April 13th, MONARC with Regional Centres representatives. This line of work will be pursued in the next months.

- **ICFA Network Task Force:** The contacts exist, but not much happened in the network sector of MONARC till now. The network issue will be part of Phase 2 studies, as mentioned in the next section 6.2 and in Chapter 5.

6.1.4 Review of resources

The hardware resources promised for the setting up of the Testbeds have been granted and are now in use (see Chapter 5). As for manpower, the hiring of people has taken place not only in CERN, but also in Milan (INFN/CILEA, from January 99), and in Tufts (in May 99). People with previous experience in

related computing matters have joined the project (from computing teams, services and centres), as well as new young people (some having previously worked on physics analysis) willing to acquire an experience in such matters (e.g. in Bologna). The CERN team fully devoted to MONARC, that consisted only of Iosif Legrand, has recently acquired the valuable contribution of Youhei Morita. The MONARC project is seen as strategic and the prospects of getting new people for working in it seem good in various countries (e.g. in Italy and in the US).

The next section sketches some ideas of importance for Phase 2. The following 6.3 will give the lines for a workplan till end-99 with the relevant main milestones.

6.2 Ideas on the Evolution of Models: Phase 2 and possibly beyond

Starting this Summer, based on the experience gained from the study and simulation of the Models developed up until that time, a systematic top-down specification of the overall system should be performed. This will involve detailed choices of a range of parameters characterising the site architectures (computing capacity, internal channel I/Os, LAN speeds), the Analysis Processes, and the network loads resulting from users' activities other than event data analysis. This design specification should possibly include

- A complete set of user, intra-site and inter-site "tasks" occurring during the data analysis.
- A definition of key limits in storage capacities, quotas for users and groups, relative priorities, and maximum acceptable turnaround times for various tasks.
- A definition of "cost metrics" that permit a determination of which responses to a given set of requests is most "cost effective", in terms of the resources used and the length of time required to satisfy the requests. A key example here is decisions to recompute certain quantities, transport data over the network, or dump the data to tape and ship it (by Express Mail).
- A definition of key parameters setting the scale of the analysis: how much data is accessed, processed or transmitted; by how many people and how often.
- A specification of the key "events" that occur within a site or between sites; examples include queueing up requests for similar data or data on the same data volumes, for staging in or migration out from tape, so that many requests are satisfied by one or a few jobs.
- A specification of priority changes, "refusal of service" or other responses to exceeding quotas or overflowing the maximum capacities, since these will tend to govern the behaviour of the system when it is fully loaded.
- A specification of the high water marks (such as system utilisation, or network occupancy level) that will trigger a change in some of the key system behaviours. This means that some of the key decisions (such when to migrate data to tape; when to recompute or transport data over the network) will depend on the overall system state. Apart from abnormal conditions, this aspect is important for evaluating the ability of the overall system to keep up with the total workload by using nights, weekend, or certain times of the year to catch up on accumulated medium or low priority tasks. Another variation of this concept is that certain parts of the workload that are time-critical can be completed by redirection of resources (sometimes in an extreme way) from lower priority tasks.

The more complex decisions implied by the above set of design specifications and concepts could lead to long and complex (multi-step) decision processes that should be the subject of careful, and potentially protracted study (see Chapter 7). In order to keep to the defined scope and schedule of MONARC as approved (for Phases 1 and 2 of the project), the overall system capacity and/or network speeds should be allowed to vary over a considerable range, so that the majority of the workload may be satisfied with

acceptable turnaround times. These may be minutes for interactive queries, hours for short jobs, a day for most long jobs, and few days for the entire workload. In this way, by the end of Phase 2, critical baseline resource requirements may be determined (in first approximation) and peculiarities or flaws in certain types of "Analysis Process" may be isolated.

Some of the tools to be designed and/or simulated, that would enable the above goals during phase 2, or eventually phase 3, are

- Query estimators that estimate the time it will take to satisfy a given request for data
- "Affinity evaluators" that will determine the concurrence of multiple requests that are proximate in space (file location) or which occur close to one another in time
- Strategies for caching, re-clustering, mirroring, or pre-emptively moving data, so that the requests can be satisfied in an acceptable time-interval

6.3 Workplan for Phase 2 till end-99

The main, general milestones for Phase2, as set out in the PEP were:

August '99	**Completion of the coding and validating phase for second-round Models**
November '99	**Completion of the second cycle of simulations of refined Models for LHC experiments**
December '99	**Completion of the project and delivery of the deliverables.**

The goals to be reached by end-99 (set of "baseline" models, guidelines both for model building and for Regional Centres) were set with a timing based on the need for MONARC to provide a useful contribution to the Computing Progress Reports of ATLAS and CMS (expected end-99 too). As it appears also from the previous section 6.2, a high level of detail will have finally to be taken into account in a realistic implementation oriented model.

The minimal scope of Phase 2 is the iteration of at least a couple of simulation cycles after the first one, in order to acquire expertise on the sensitivity of the models to the different features and parameters; at the same time incorporating a first definition of the "cost metrics".
Issues like the study of detailed priority schemes, of data caching and tape "staging", as well as job-migration vs network data transfer will enhance the value of the MONARC contribution to the CPR's but are not required for this contribution being useful and significant.

The planning made here assumes Phase 2 simulations to end on November 99. If the CPR's are delayed MONARC will surely be able to take advantage of the added time for addressing the issues of section 6.2 that cannot be considered in the shorter schedule.

The Phase 3 as it was foreseen in the PEP was centred on prototyping and implementation issues; it is now clear that in the first stage of Phase 3, the core of the work will be devoted to system optimisation studies (see Chapter 7). The boundary Phase 2-3 is thus for MONARC largely a matter of external opportunity.

It is proposed to link the end of Phase 2 to the CPR date. If a Phase 3 is approved, the end of Phase 2 will coincide with the CPR completion date, and Phase 3 will address spreading the modelling knowledge

(code, guidelines, etc.) into the experiments. In absence of Phase 3, Phase 2 should include an organised knowledge transfer to the experiments and should therefore end some two months after the CPR completion.

6.3.1 Goals for Phase 2 in 1999

As said above, this section deals with a time span for simulation ending in November 99.
The following results need to be achieved with such timing:

- Perform at least two full simulation cycles after the first one
- Elaborate a complete set of "tasks" during the data analysis
- Define a reliable "cost metrics". Technology tracking is crucial here.
- Define the key parameters that set the scale of the analysis and of the required resources
- Clarify the role of MonteCarlo: resources, localisation, interference with the rest of the computing tasks
- Provide a top-down description of possible site and network architecture, granting no relevant aspects are missed in the modelling
- Provide convincing validation of the Simulation performing tests with fully understood and real implementations (testbeds are the primary resource to be used in this role)
- Perform model evaluation on the basis of criteria taking into account some or all of the following items as far as possible:
    - Adaptability of the system to architectures likely to exist in different Centres
    - Responsiveness as ability to respond to peak load for "urgent" analysis
    - Scalability as level of performance as the data volume and component performances increase with time
    - Flexibility as ability to adapt or migrate to a different model over time.
      This is a crucial point as we cannot expect to be able to forecast technologies,costs and resources on the full time span of LHC life; actually even 5 years are a difficult time span.
    - Overall performance vs cost. Last but not least.
- Perform comparison between a suitable CERNtralised model and a suitable Regional Centre model.

The timing of Phase 2 after end-99 will depend both on the timing of the CPRs and on the decisions about Phase 3, as stated above. The planning for a possible Phase 2 extension and operational proposals for a Phase 3 extension will be presented to the December LCB. The final status report of MONARC Phase 2 could also be presented at this time, or a few months later, according to the timing of the CPRs.

6.4 Main Milestones from now on

The main MONARC milestones until end-99 are:

July '99	**Completion of the first cycle of simulation of possible Model for LHC experiments. First classification of the Models and first evaluation criteria.**
September '99	**Reliable figures on Technologies and Costs from Technology Tracking work to be inserted in the Modelling.**
September '99	**First results on Model Validation available.**
September '99	**First results on Model Comparison available.**
November '99	**Completion of a simulation cycle achieving the goals described in 6.3.1**
November '99	**Document on Guidelines for Regional Centres available**
December '99	**Presentation to LCB of a proposal for the continuation of MONARC**

Chapter 7: Ideas for a Possible MONARC Phase 3

We believe that from 2000 onwards, a significant amount of work will be necessary to model, prototype and optimise the design of the overall distributed computing and data handling systems for the LHC experiments. This work, much of which should be done in common for the experiments, would be aimed at providing "cost effective" means of doing data analysis in the various world regions, as well as at CERN. Finding common solutions would save some of the resources devoted to determining the solutions, and would ensure that the solutions found were mutually compatible. The importance of compatibility based on common solutions applies as much to cases where multiple Regional Centres in a country intercommunicate across a common network infrastructure, as it does to sites (including CERN) that serve more than one LHC experiment.

A MONARC Phase 3 could have a useful impact in several areas, including:

- facilitation of contacts, discussions, interchanges, for the planning and mutually compatible design of centre and network architecture and services (among the experiments, the CERN Centre and the Regional Centres)
- modelling consultancy and "service" to the experiments and Centres
- providing a core of advanced R&D activities encompassing system optimisation, and pre-production prototyping
- taking advantage of the work on distributed data-intensive computing systems beginning this year in other "next generation" R&D projects

Details on the synergy between a MONARC Phase 3 and R&D projects such as the recently approved Next Generation Internet "Particle Physics Data Grid" (PPDG) may be found in [29]. The PPDG project (involving ANL, BNL, Caltech, FNAL, JLAB, LBNL, SDSC, SLAC, and the University of Wisconsin) shares MONARC's aim of finding common solutions to meet the large-scale data management needs of high energy (as well as nuclear) physics. Some of the concepts of a possible Phase 3 study are briefly

summarised below.

The Phase 3 study could be aimed at maximising the workload sustainable by a given set of networks and site facilities, or at reducing the long turnaround times for certain data analysis tasks, or a combination of both. Unlike Phase 2, the optimization of the system in Phase 3 would no longer exclude long and involved decision processes, as the potential gains in terms of work accomplished or resources saved could be large. Some examples of the complex elements of the Computing Model that might determine the (realistic) behaviour of the overall system, and which could be studied in Phase 3 are

- **Resilience**, resulting from flexible management of each data transaction, especially over wide area networks
- **Fault tolerance**, resulting from robust fall-back strategies and procedures (automatic and manual, if necessary) to recover from abnormal conditions (such as irrecoverable error conditions due to data corruption, system thrashing, or a subsystem falling offline).
- **System state tracking**, so that the capability of the system to respond to requests is known (approximately) at any given time, and the time to satisfy requests for data and/or processing power may be, on average, reliably estimated, or abnormal conditions may be detected and in some cases predicted.

MONARC in Phase 3 could exploit the studies, system software developments, and prototype system tests completed by early 2000, to develop more sophisticated and efficient Models than were possible in Phase 2. The Simulation and Modelling work of MONARC on data-intensive distributed systems is likely to be more advanced than in PPDG or other NGI projects in 2000, so that MONARC Phase 3 could have a central role in the further study and advancement of the design of distributed systems capable of PetaByte-scale data processing and analysis. As mentioned in the PEP, this activity would potentially be of great importance not only for the LHC experiments, but for scientific research on a broader front, and eventually for industry.

Appendix A: The MONARC Simulation Tool

A-1. Introduction

The aim of this note is to describe a simulation program, being developed as a design and optimization tool for large scale distributed computing system for future LHC experiments. The goals are to provide a realistic simulation of distributed computing systems, customised for specific physics data processing and to offer a flexible and dynamic environment to evaluate the performance of a range of possible data processing architectures.

A discrete event, process oriented simulation approach, developed in Java$^{(TM)}$ was used for this modelling project. A Graphical User Interface (GUI) to the simulation engine, which allows to change dynamically parameters, and to monitor and analyse on-line results, provides a powerful development tool for evaluating and designing large scale distributed processing systems.

A-2. Design Considerations of the simulation program

The simulation and modelling task for MONARC requires the description of complex data processing programs, running on large scale distributed systems and exchanging very large amounts of data. Building the logical simulation model requires the abstraction from the real system all the components and their time dependent interaction. This logical model has to be equivalent to the simulated system in all important respects. An Object Oriented design, which allows an easy and direct mapping of the logical components into the simulation program and provides the interaction mechanism, offers the best solution for such a large scale system and also copes with systems which may change dynamically.

A process oriented approach for discrete event simulation is well suited to describe concurrent running programs as well as all the stochastic arrival patterns, specific for such type of simulation. Threaded objects or "Active Objects" (having an execution thread, program counter, stack...) allow a natural way to map the specific behaviour of distributed data processing into the simulation program.

This simulation project is based on Java technology which provides adequate tools for developing a flexible and distributed process oriented simulation. Java has build-in multi-thread support for concurrent processing, which can be used for simulation purposes by providing a dedicated scheduling mechanism. Java also offers good support for graphics and it is easy to interface graphics with the simulation code. Proper graphics tools, and ways to analyse data interactively, are essential in any simulation project.

Currently, many groups involved in Computer Simulation are moving towards Java. Perhaps the best known project, Ptolemy II, is a complete new redesign of the Ptolemy simulation environment. The reasons for which we decided to write a new "simulation engine" for process oriented, discrete event simulation were, first, at the time we started this project, Ptolemy II was not available, and second, a dedicated core for the simulation engine can be more efficiently implemented. However the modular structure of this simulation package does not exclude the possibility to be interfaced with the engines of other simulation tools like Ptolemy II.

A-3. The components models

A-3.1 Data Model

It is foreseen that all HEP experiments will use an Object Database Management System (ODBMS) to handle the large amounts of data in the LHC era. Our data model follows closely the Objectivity architecture and the basic object data design used in HEP. The model should provide a realistic mapping of an ODBMS, and at the same time allow an efficient way to describe very large database systems with a huge number of objects.

The atomic unit object is the "Data Container", which emulates a database file containing a set of objects

of a certain type. Objects assumed to be stored in such data files are considered in the simulation to be in a sequential order. In this way the number of objects used in the simulation to model large number of real objects is dramatically reduced, and the searching algorithms are simple and fast. Random access patterns, which are necessary for realistic modelling of data access are simulated by creating sequence of indices Clustering factors for certain types of objects, when accessed from different programs, are simulated using practically the same scheme to generate a vector of intervals.

A Data Base unit is a collection of containers and performs an efficient search for type and object index range. The AMS server simulation provides the client server mechanism to access objects from the database. It implements response time functions based on data parameters (page size, object size, access is from a new container...), and hardware load (how many other requests are in process at the same time). In this model it is also assumed that the AMS servers control the data transfers from/to mass storage system. Different policies for storage management may be used in the simulation. AMS servers register with a database catalogue (Index) used by any client (user program) to address the proper server for each particular request.

This modelling scheme provides for an efficient way to handle a very large number of objects and automatic storage management, allows one to emulate different clustering schemes of the data for different types of data access patterns, as well as to simulate the order of access following the associations between the data objects, even if the objects reside in databases in different AMS servers.

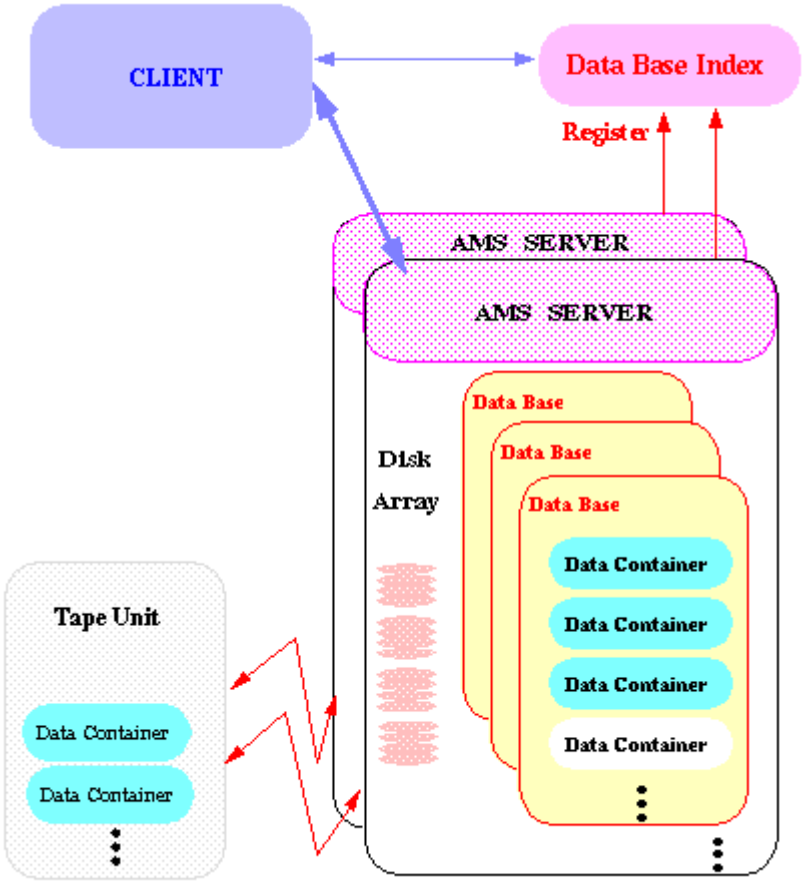


Figure A-1: A schematic diagram of data model based on ODBMS architecture

A-3.2 Multitasking Data Processing Model

Multitasking operating systems share resouces such as CPU, memory and I/O between concurrently running tasks by scheduling their use for very short time intervals. However, simulating the detail of how tasks are scheduled in the real system would be too complex and time consuming, and thus it is not suitable for our purpose. Therefore we need to model the multitasking data processing.

Our model for multitasking processing is based on an "interrupt" driven mechanism implemented in the simulation engine. An interrupt method implemented in the "Active" object which is the base class for all running jobs, is a key part for the multitasking model. The way it works is shown schematically in Figure A-2.

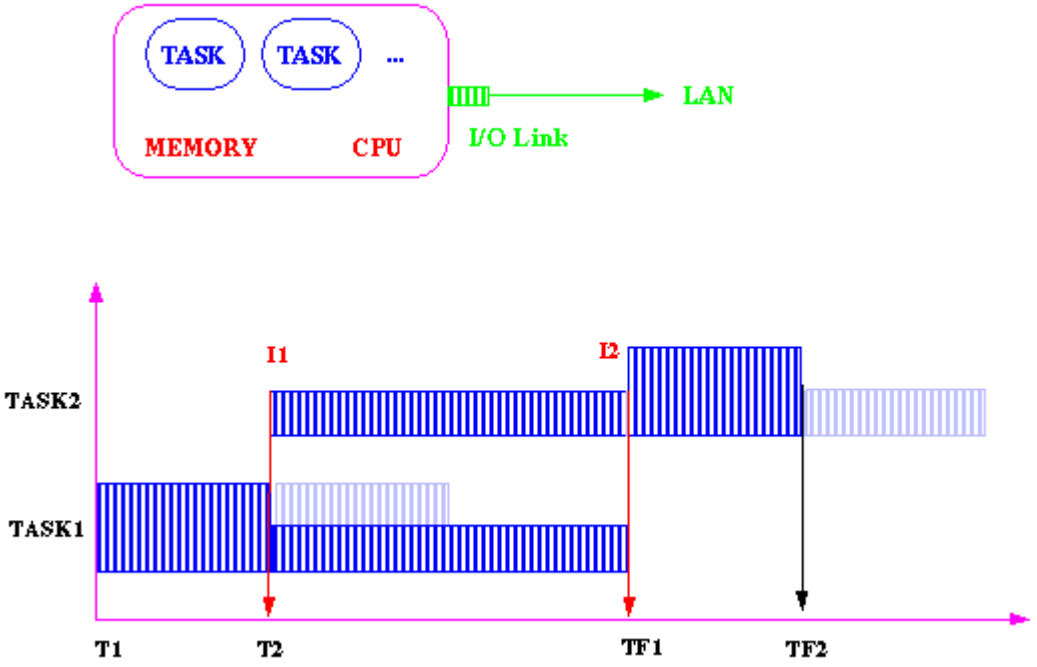


Figure A-2: Modelling multitasking processing based on an "interrupt" scheme

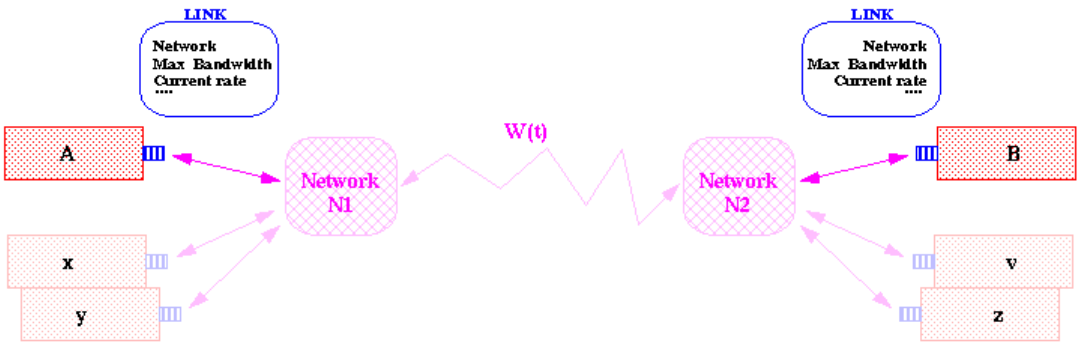
When a first job starts, the time it takes is evaluated and this "Active" object enters into a wait state for this amount of time and allows to be interrupted. If a new job starts on the same hardware it will interrupt the first one. Both will share the same CPU power and the time to complete for both of them is computed assuming that they share the CPU equally. Both active jobs will enter into a wait state and are listeners to interrupts. When a job is finished it also creates an interrupt to re-distribute the resources for the remaining ones. This model is in fact assuming that resource sharing is done continuously between any discrete events in the simulation time (e.g. new job submission, job completion) while on real machines it is done in a discrete way but with a very small time interval. This provides an accurate and efficient model for multiprocessing tasks.

A-3.3 LAN/WAN Networking Model

Accurate and efficient simulation of networking is also a major requirement for MONARC. The simulation program should offer the possibility to simulate data traffic for different protocols on both LAN and WAN. This has to be done for very large amounts of data and without precise knowledge of the network topology (as in the case of long distance connections). It is practically impossible to simulate the networking part at a packet level.

The approach used to simulate the data traffic is again based on an "interrupt" scheme. When a message transfer starts between two end points in the network, the time to completion is calculated. This is done using the minimum speed value of all the components in between, which can be time dependent, and related with the protocol used. The time to complete is used to generate a wait statement which allows to be interrupted in the simulation. If a new message is initiated during this time an interrupt is generated for the LAN/WAN object. The speed for each transfer affected by the new one is re-computed, assuming that they are running in parallel and share the bandwidth with weights depending on the protocol. With this new speed the time to complete for all the messages affected is re-evaluated and inserted into the priority queue for future events. This approach requires an estimate of the data transfer speed for each component. For a long distance connection an "effective speed" between two points has to be used. This value can be fully time dependent.

This approach for data transfer can provide an effective and accurate way to describe many large and small data transfers occurring in parallel on the same network. This model cannot describe speed variation in the traffic during one transfer if no other transfer starts or finishes. This is a consequence of the fact that we have only discrete events in time. However, by using smaller packages for data transfer or artificially generating additional interrupts for LAN/WAN objects the time interval for which the network speed is considered constant can be reduced. As before, this model assumes that the data transfer between time events is done in a continuous way utilising a certain part of the available bandwidth.



$$Bandwidth_{AB}(t) = F(Protocol, L_A, L_B, N_1(t), N_2(t), W(t))$$

Figure A-3: The Networking simulation model

A-3.4 Arrival Patterns

A flexible mechanism to define the stochastic process of submitting jobs is necessary. This is done using the "dynamic loadable modules" feature in Java which provide the support to include (threaded) objects into running code. These objects are used to describe the behaviour of a "User" or a "Group of Users". It should be able to describe both batch and interactive sessions, and also to use any time dependent distribution describing how jobs are submitted. An "Activity" object is a base class for all these processes for which current experience should be used to estimate the time dependent patterns and correlations.

In order to provide a high flexibility in modelling all these activities, the user can provide very simple sections of Java code, to override the "RUN" method of the "Activity" class. Any number of such "Activities" can be dynamically loaded into the "Regional Centre" object, simulating the "Users" using the computing facilities. A schematic view of such objects is presented in Figure A-4 together with a very simple RUN method in which a new job is submitted to the farm every 1000 seconds. The job is an Analysis job which uses TAG data, and for 1% of events is processing ESD data, and for another 0.5% is accessing RAW data.

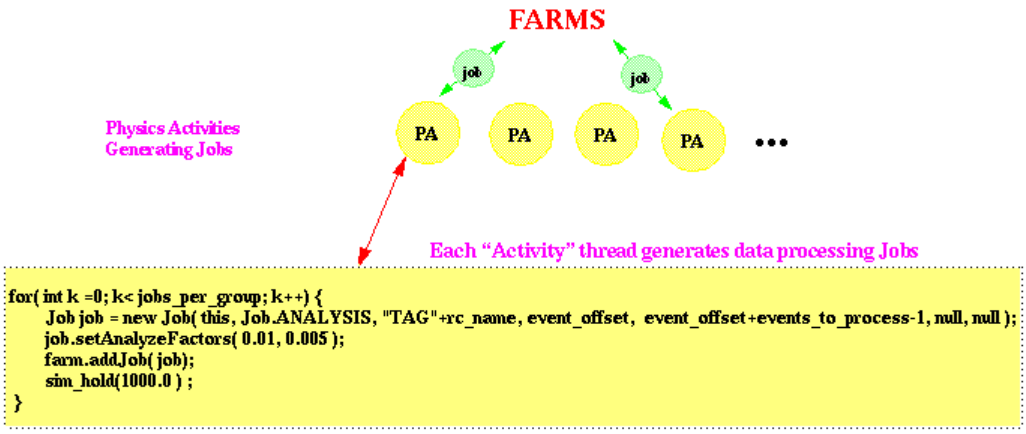


Figure A-4: The "Users" description model

A-3.5 Regional Centre Model

A Regional Centre object is a complex, composite object containing a number of data servers and processing nodes, all connected to a LAN. As an option it may contain a Mass Storage unit and can be connected to other Regional Centres. Any regional centre can instantiate dynamically a set of "Users" or "Activity" Objects which are used to generate data processing jobs based on different scenarios. Inside a Regional Centre different job scheduling policies may used to distribute the jobs to processing nodes. Currently, a simple load balancing mechanism is used, which does not allow swapping on the processing nodes, and queues jobs when no more active memory is available.

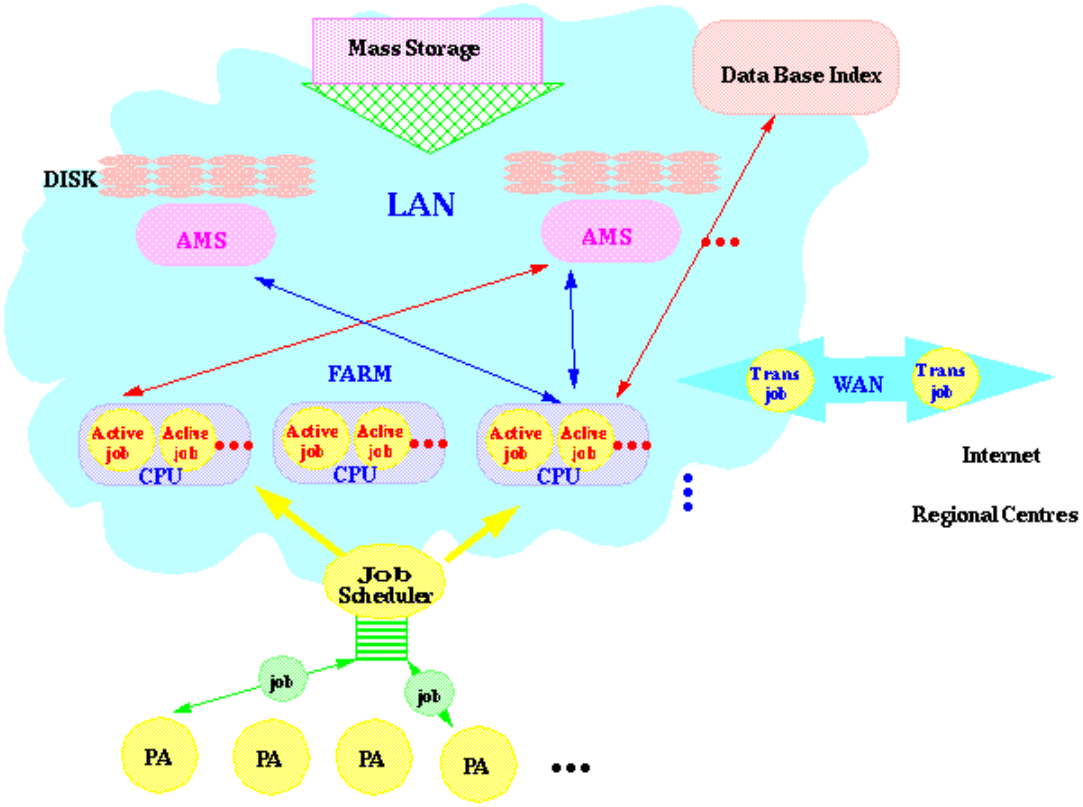


Figure A-5: A schematic view of a Regional Centre object as a composite object

With this structure it is now possible to build a wide range of computing models, from the very centralized (with reconstruction and most analyses at CERN) to the distributed systems, with an almost arbitrary level of complication (CERN and multiple regional centres, each with different hardware configuration and possibly different sets of data replicated)

A-4. The Graphical User Interface

An adequate set of GUIs to define different input scenarios, and to analyse the results, are essential for the simulation tools. The aim in designing these GUIs was to provide a simple but flexble way to define the parameters for simulations and the presentation of results.

In Figure A-6 the frames used to define the system configuration are presented.

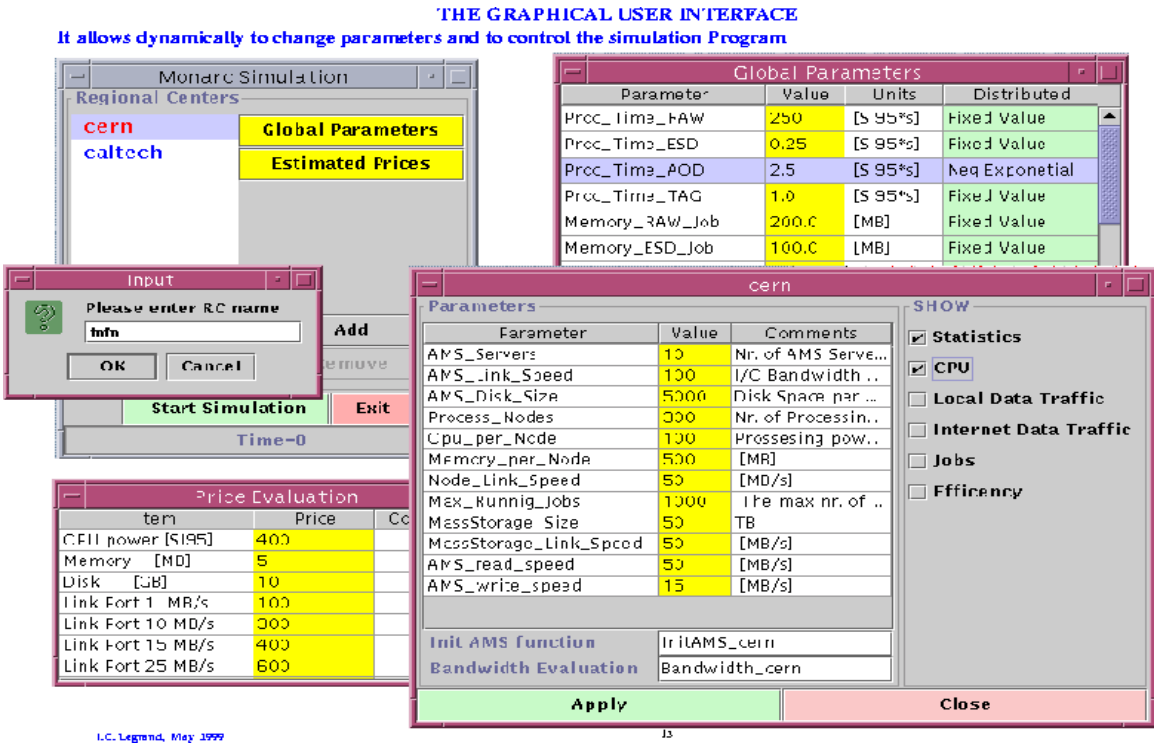


Figure A-6 : The GUI Frames used to define the system configuration

The number of regional centres considered can be easily changed through the main window of the simulation program.

The "Global Parameters" menu allows to change the (mean) values and their statistical distributions for quantities which are common in all Regional Centres.

The Price List table contains basic hardware cost estimates for the components in the system and can be used to evaluate the total cost of the Regional Centre. Currently a very simple scheme is used, but soon we will develop a more elaborate mechanism to evaluate the overall costs for such systems.

For each Regional Centre the configuration is defined at startup in a simple ASCII file, which can be modified at any time through the specific graphical Frame. Parameters currently accessible from the GUI are listed in Table A-1.

Table A-1: A list of parameters currently in use by the MONARC simulation tool

federated database and data model parameters (global)	**regional centre configuration parameters (local)**
database page size	number of AMS_servers
TAG object size/event	AMS link speed
AOD object size/event	AMS disk size
ESD object size/event	number of processing nodes
RAW object size/event	CPU/node
processing time RAW->ESD	memory/node
processing time ESD->AOD	node link speed
processing time AOD->TAG	mass storage size (in HSM)
analysis time TAG	link speed to HSM
analysis time AOD	AMS write speed
analysis time ESD	AMS read speed
memory for RAW->ESD processing job	(maximum disk read/write speed)
memory for ESD->AOD processing job	
memory for AOD->TAG processing job	**data access pattern parameters (local)**
memory for TAG analysis job	fraction of events for which TAG->AOD associations are followed
memory for AOD analysis job	
memory for ESD analysis job	fraction of events for which AOD->ESD associations are followed
container size RAW	
container size ESD	fraction of events for which ESD->RAW associations are followed
container size AOD	
container size TAG	clustering density parameter

A frame of a regional centre appears when the name of the centre is selected in the main window. In this window the user may select which parameters to be graphically presented (CPU usage, memory load, load on the network...). For all these results basic mathematical tools are available to easy compute integrated values, mean, integrated mean value...

Additional statistical tools will be added soon.

A-5. Two simple examples

In what follows, two simple examples are presented. Parameters used in these examples may not be realistic, as our first aim was to describe and test the simulation tool and its ability to describe specific problems related with data processing for LHC experiments. Testing the time diagrams, integrated values, locking problems were done for typical data flow scenarios, using fixed value parameters to check the program.

A-5.1 Event Reconstruction Example

This task is basically foreseen to be done in a Single Centre (at CERN) (Figure A-7), and is a typical CPU intensive activity. It requires access to Mass Storage units, and is a good example for which the a right balance between CPU power and LAN traffic can be studied.

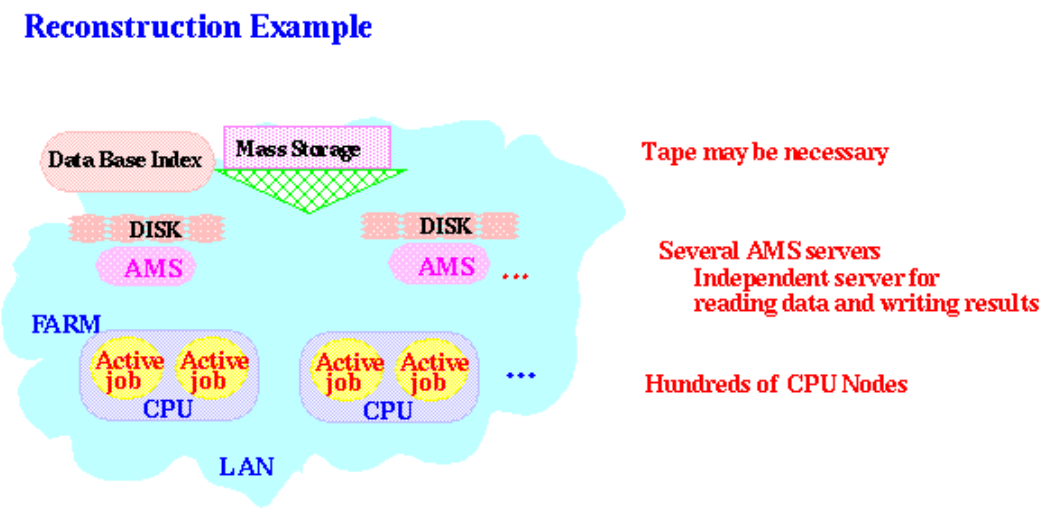


Figure A-7 : Schematic view of the components involved in the Single Centre Reconstruction problem

Parameters used for this problem, and an example of the presentation of the results, are shown in Figure A-8. In this example 250 processing nodes are used and the maximum number of active jobs is 500. The total number of events "processed" is 10^7. It uses a simple scheme for the response time of the AMS servers, i.e. constant values for reading and writing data. As an initial condition all data are stored on the Mass Storage unit and it requires first to move data on the disks. CPU usage is affected by data availability and at the beginning all AMS servers (10 in this case) are starting data transfers to disk. The network traffic to the Mass Storage (green line) is saturated during this time.

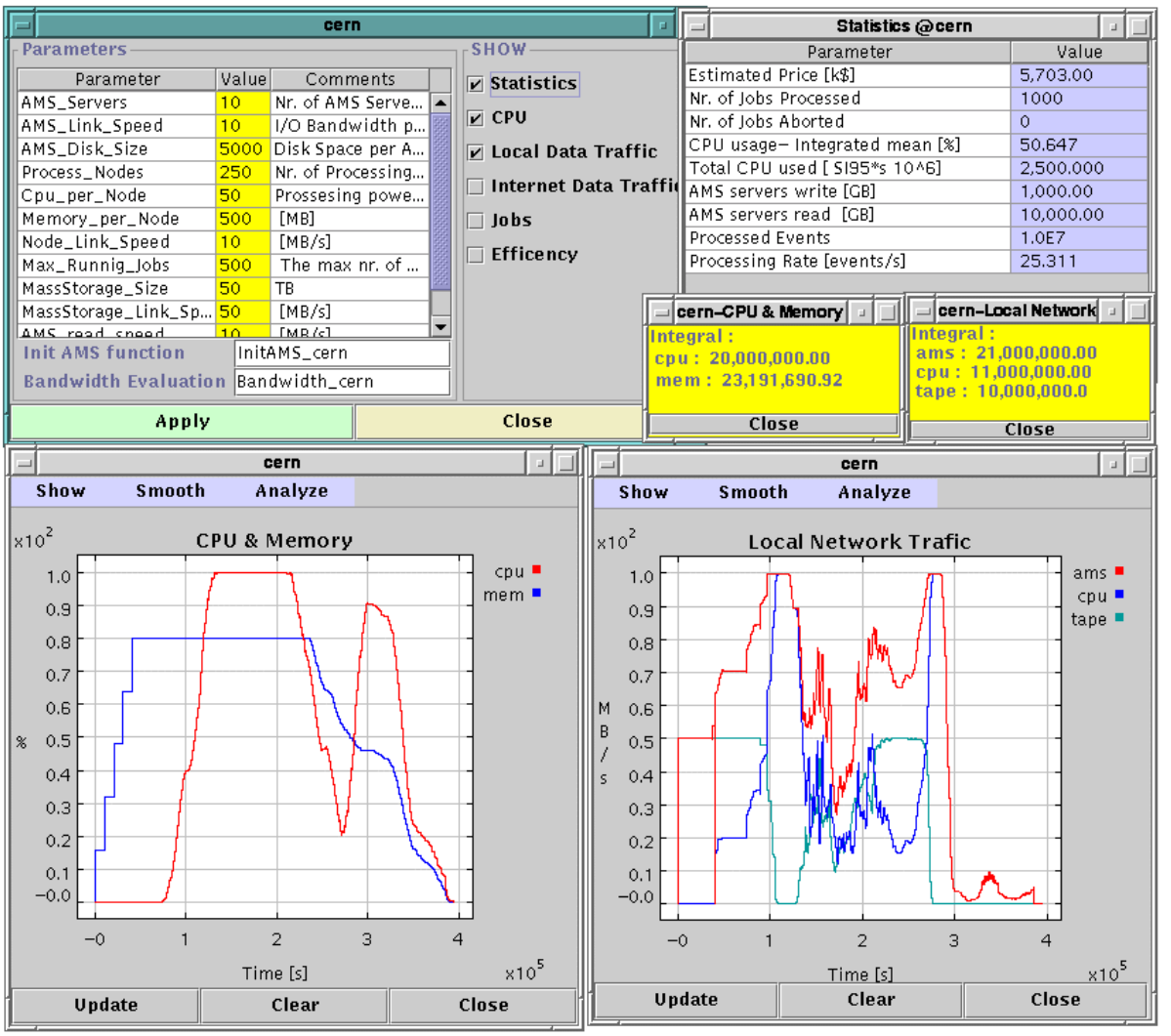


Figure A-8 : A few results from a simple "reconstruction" problem

This simulation uses about 600 threads in parallel, and it takes about 1.5 minutes on an Ultra 5 Sun system with CPU at 333 MHz. It also uses about 50 MB of memory. A significant part of the resources are used for the graphical part of the program.

A-5.2 Physics Analysis Example

The aim of this example is to illustrate data analysis being performed in Regional Centres. The Central site (CERN) stores all the data (RAW, ESD, AOD, TAG) while Regional Centres have copies only for AOD and TAG in this example. A schematic view of the set up is shown in Figure A-9. Jobs will mainly use local data, but for 1% of the events will ask for ESD data and for 0.5% RAW data from the Main Centre.

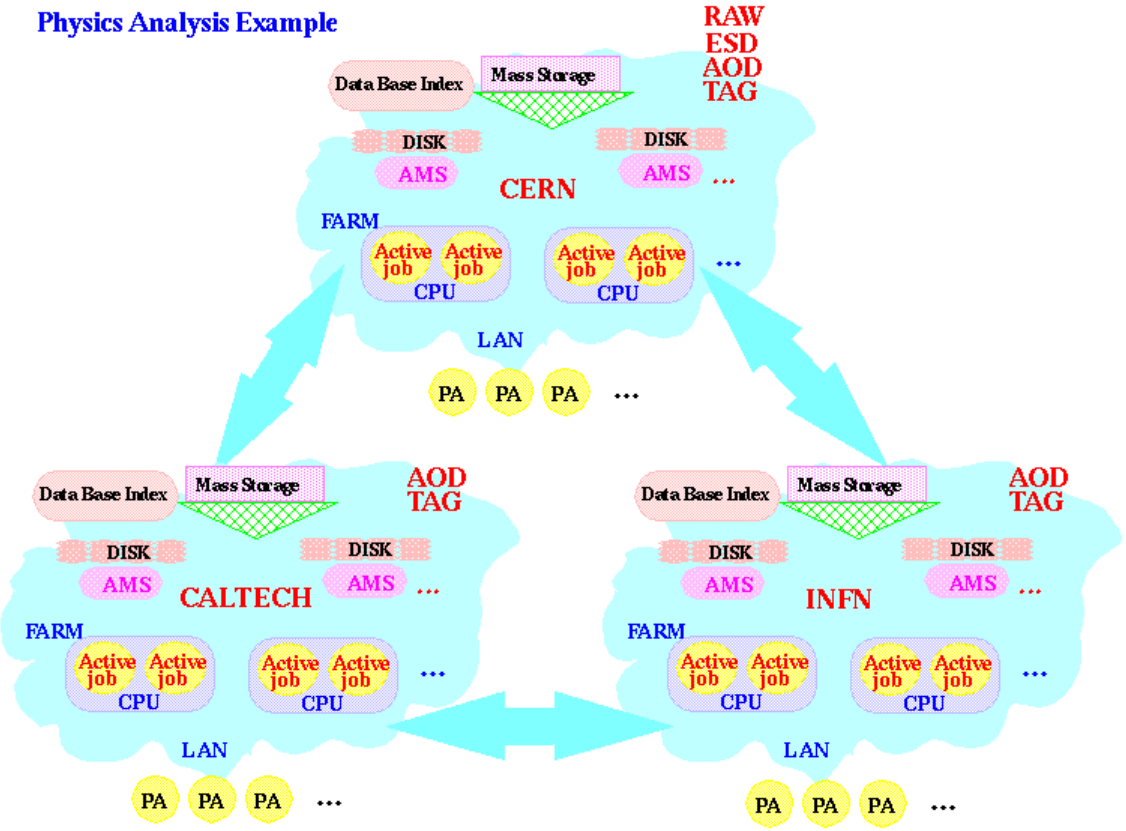


Figure A-9 : Data distribution for the Physics Analysis Example

In all three centres exactly the same jobs are done. Centres have connections of different bandwidth. A sample of the parameters used and the results are presented in Figure A-10a, b. The total internet traffic between the two regional centres and the main one should be the same, as the jobs are identical. This can be seen (Figure A-10a) as the integrated values for the Traffic are equal while the connection bandwidth are different. The CPU usage and LAN traffic for two centres is presented in Figure A-10b.

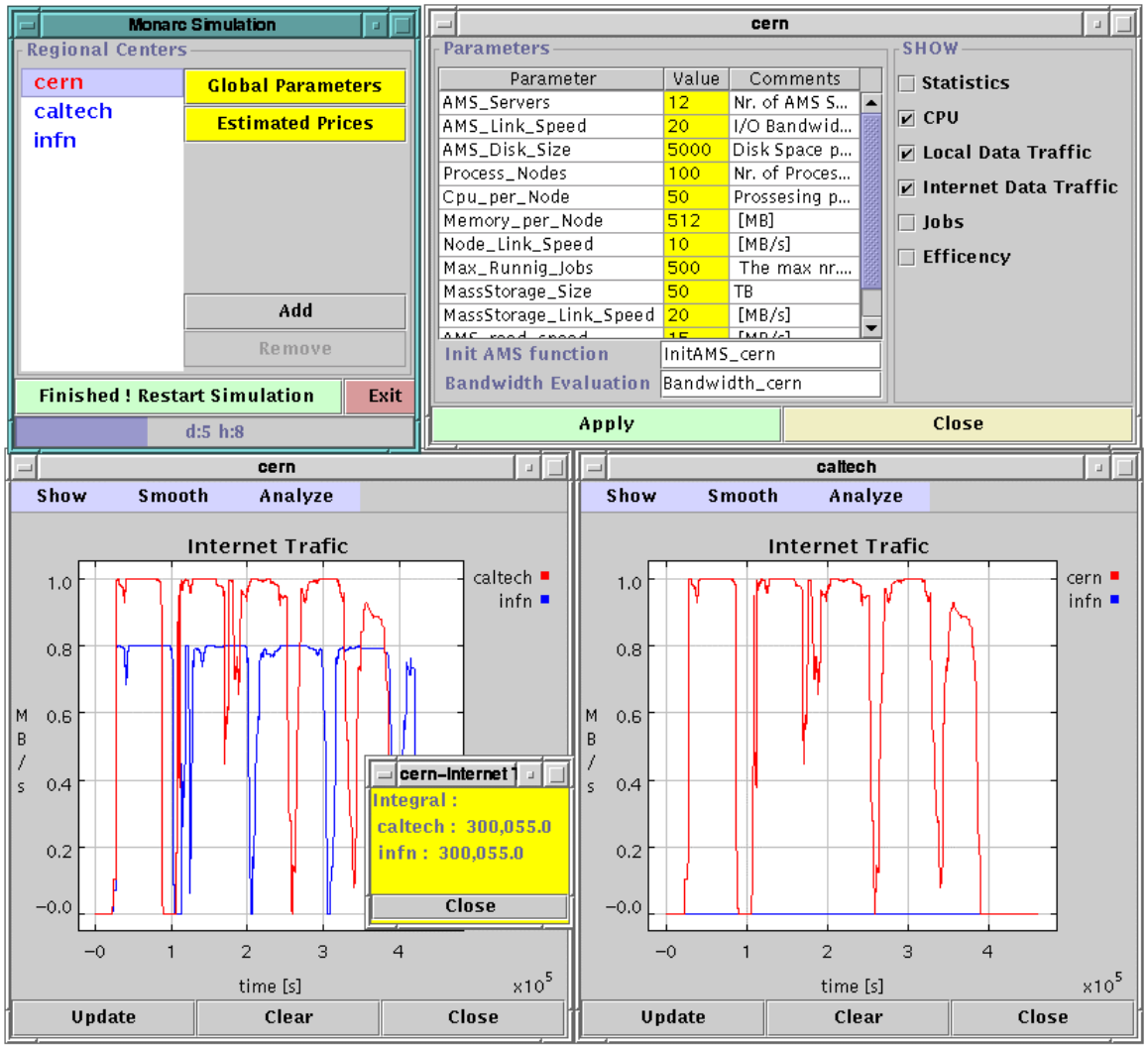


Figure A-10a : Part of the configuration parameters and the internet traffic diagrams for Physics Analysis Example

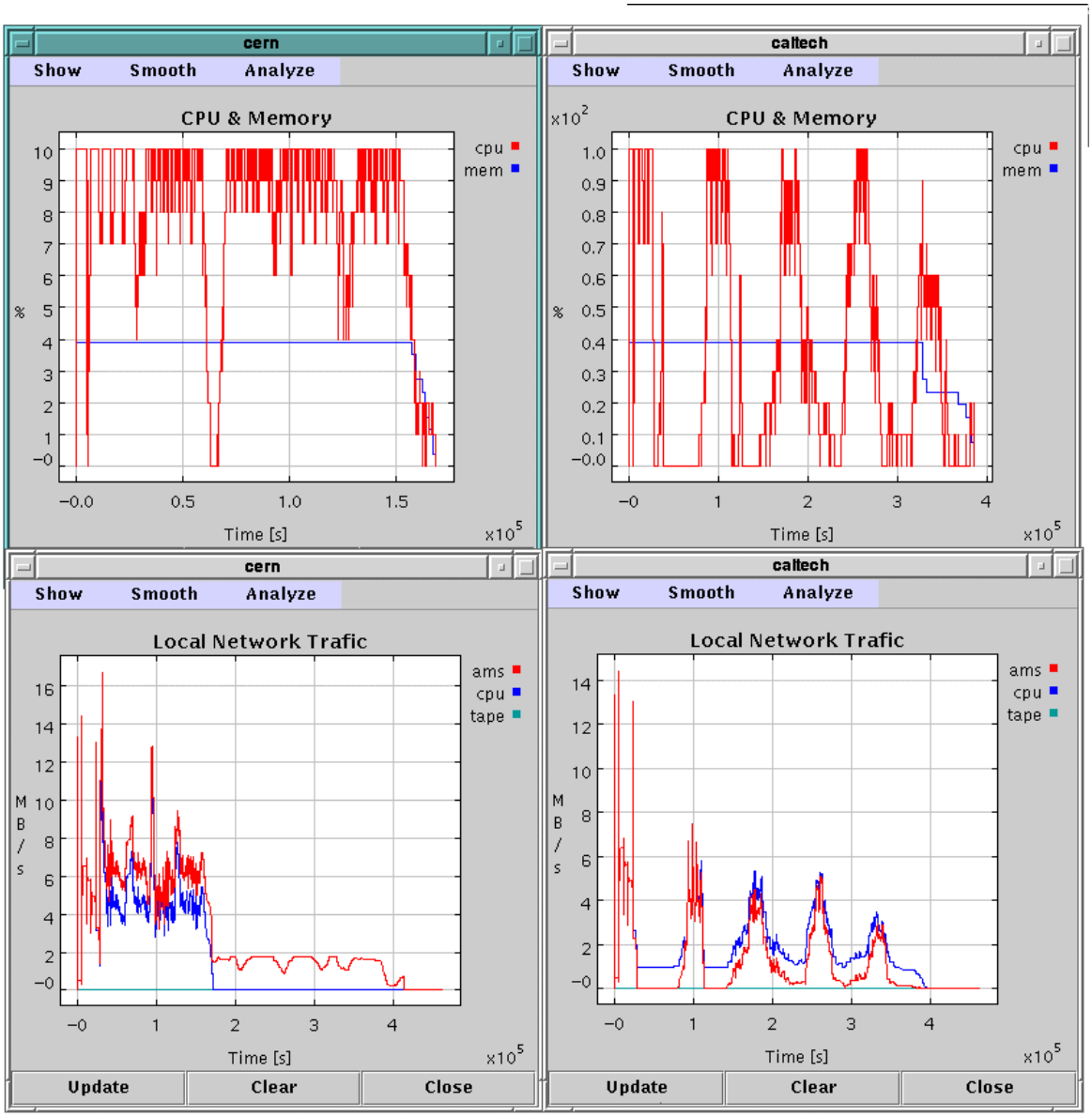


Figure A-10b : CPU usage and Local Data Traffic diagrams for Physics Analysis Example

This simulation job takes less than one minute on the same hardware, and it uses less active threads than in the previous example but has more events in the data traffic part.

Appendix B: References

1) The WWW Home Page for the MONARC Project
http://www.cern.ch/MONARC/

2) The MONARC Project Execution Plan, September 1998
http://www.cern.ch/MONARC/docs/pep.html

3) Christoph von Praun: Modelling and Simulation of Wide Area Data Communications.
A talk given at the CMS Computing Steering Board on 19/06/98.

4) J.J.Bunn: Simple Simulation of the Computing Models:
http://pcbunn.cithep.caltech.edu/results/model/model.html

5) The PTOLEMY Simulation Tool:
http://www-tkn.ee.tu-berlin.de/equipment/sim/ptolemy.html

6) The SES Workbench
http://www.ses.com/Workbench/index.htm

7) **PARASOL** - C/C++ simulation library for dist / parallel systems
http://www.hensa.ac.uk/parallel/simulation/architectures/parasol/index.html

8) Rough Sizing Estimates for a Computing Facility for a Large LHC experiment, Les Robertson.
MONARC-99/1.
http://nicewww.cern.ch/~les/monarc/capacity_summary.html.

9) Report on Computing Architectures of Existing Experiments, V.O'Dell et al. MONARC-99/2.
http://home.fnal.gov/~odell/monarc_report.html

10) Regional Centers for LHC Computing, Luciano Barone et al. MONARC-99/3. (text version)
http://home.cern.ch/~barone/monarc/RCArchitecture.html

11) Presentations and notes from the MONARC meeting with Regional Center Representatives April 23, 1999
http://www.fnal.gov/projects/monarc/task2/rc_mtg_apr_23_99.html

12) PASTA, Technology Tracking Team for Processors, Memory, Storage and Architectures:
http://nicewww.cern.ch/~les/pasta/run2/welcome.html

13) Home page of the Analysis Design Working Group:
http://www.bo.infn.it/monarc/ADWG/AD-WG-Webpage.html

14) Analysis Processes of current and imminent experiments:
http://www.bo.infn.it/monarc/ADWG/Meetings/Docu-15-12-98.html

15) Monarc Note 98/1:
http://www.mi.infn.it/~cmp/rd55/rd55-1-98.html

16) CMS TN-1996/071 The CMS Computing Model

17) Parameters of the initial Analysis Model:
http://www.bo.infn.it/monarc/ADWG/Meetings/15-01-99-Docu/Monarc-AD-WG-0199.html

18) Unfeasable models evaluations:

http://www.bo.infn.it/monarc/ADWG/Meetings/Docu-24-01-99.html (to be released)

19) Preliminary evaluation criteria (slide 8)
http://bo_srv1_nice.bo.infn.it/~capiluppi/monarc-workshop-0599.pdf

20) ATLFAST++ in LHC++:
http://www.cern.ch/Atlas/GROUPS/SOFTWARE/OO/domains/analysis/atlfast++.html

21) GIOD (Globally Interconnected Object Databases) project:
http://pcbunn.cithep.caltech.edu/Default.htm

22) ATLAS 1 TB Milestone:
http://home.cern.ch/s/schaffer/www/slides/db-meeting-170399-new/

23) CMS TestBeam web Page
http://cmsdoc.cern.ch/ftp/afscms/OO/Testbeams/www/Welcome.html

24) MONARC-99/4: M. Boschini, L. Perini, F. Prelz, S. Resconi: Preliminary Objectivity tests for MONARC project on a local federated database:
http://www.cern.ch/MONARC/docs/monarc_docs/99-04.ps

25) K. Holtman: CPU requirements for 100 MB/s writing with Objectivity:
http://home.cern.ch/~kholtman/monarc/cpureqs.html

26) K. Amako, Y. Karita, Y. Morita, T. Sasaki and H. Sato: MONARC testbed and a preliminary measurement on Objectivity AMS server:
http://www-ccint.kek.jp/People/morita/Monarc/amstest.ps

27) K. Sliwa: What measurements are needed now?:
http://www.cern.ch/MONARC/simulation/measurements_may_99.htm

28) C. Vistoli: QoS Tests and relationship with MONARC:
http://www.cnaf.infn.it/~vistoli/monarc/index.htm

29) H. Newman: Ideas for Collaborative work as a Phase 3 of MONARC
http://www.cern.ch/MONARC/docs/progress_report/longc7.html