# MONARC – lessons for ATLAS and implications for ATLAS world-wide computing model

**Krzysztof Sliwa/ Tufts University**
**ATLAS Software Week, May 12, 2000**

i. THERE EXIST A GOOD SIMULATION TOOL THAT SHOULD BE VERY HELPFUL FOR ATLAS-SPECIFIC, MORE DETAILED DESIGN WORK ON DISTRIBUTED SYSTEMS

ii. THREE MAIN RESULTS FROM SIMULATIONS PERFORMED SO FAR:

- ONE NEEDS 622 Mbps BANDWIDTH BETWEEN ANY OF THE TIER-1 CENTERS AND CERN; (possibly the same between TIER-2 and Tier-1) TO REALIZE A DISTRIBUTED COMPUTING MODEL

- IT IS VERY IMPORTANT TO CAREFULLY BALANCE THE LOAD ON DATA SERVERS, I.E. DISTRIBUTE DATA IN AN INTELLIGENT WAY ON THE DISK SERVERS (failing to do so may easily result in deterioration of performance of the system by a factor ~2-4)

- IT IS IMPORTANT TO CAREFULLY DESIGN THE JOB SUBMISSION SCHEME (BATCH JOBS FOR COLLABORATION-WIDE ACTIVITIES, LIKE RAW DATA RECONSTUCTION) TO BALANCE THE LOAD ON CPU BOXES (failing to do so may result in deterioration of performance of the system by a factor of ~2-4)

COMMENTS ON IMPLICATIONS ON THE ATLAS COMPUTING MODEL FROM THE MONARC SIMULATIONS:

i. The simulation ASSUMES an Object Oriented Database capable of maintaining associations between data objects, and maintaining a uniform logical view of all the data, regardless of its physical location. It means that, if someone would decide to built a system NOT based on Objectivity (or Espresso), the simulation results would not be valid (model would not work). I think that it would be very difficult to build an efficient system without an Object Oriented database.

ii. If the network bandwidth to CERN were UNLIMITED, one does not gain anything by having Regional Centers, to the contrary, costs of running the multiple centers would exceed that of running a single center. Ignoring politics, the Regional Center concept is useful only in the case of LIMITED bandwidth from CERN to physicists outside of CERN. One gains in such a case by parallelizing access to the data and by providing faster and more efficient data transfers by using national networks of higher bandwidth than the transatlantic links. Similarly, if one ignores politics, the correct balance between Tier-1 and Tier-2 centers is only a function of the network connectivity on the national level between

the users and a Tier-1 center. If high bandwidth connections are available, one may NOT NEED Tier-2 centers. This point seems to elude many.

iii. Enclosed are selected pages from the MONARC Phase 2 Report from winter/spring 2000.

**CERN/LCB 2000-001**

# Models of Networked Analysis at Regional Centres for LHC Experiments

## (MONARC)

## PHASE 2 REPORT

24th March 2000

**MONARC Members**

M. Aderholz (MPI), K. Amako (KEK), E. Auge (L.A.L/Orsay), G. Bagliesi (Pisa/INFN), L. Barone (Roma1/INFN), G. Battistoni (Milano/INFN), M. Bernardi (CINECA), M. Boschini (CILEA), A. Brunengo (Genova/INFN) J.J. Bunn (Caltech/CERN), J. Butler (FNAL), M. Campanella (Milano/INFN), P. Capiluppi (Bologna/INFN), F, Carminati (CERN), M. D'Amato (Bari/INFN), M. Dameri (Genova/INFN), A. di Mattia (Roma1/INFN), A. Dorokhov (CERN), 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), 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),
A. Putzer (Heidelberg), 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), H. Stockinger (CERN),
D. Ugolotti (Bologna/INFN), E. Valente (INFN), C. Vistoli (CNAF/INFN), I. Willers (CERN),
R. Wilkinson (Caltech), D.O. Williams (CERN).

# Executive Summary

Since Autumn 1998, the MONARC project [1] has provided key information on the design and operation of the worldwide-distributed Computing Models for the LHC experiments. This document summarises the status of MONARC and the results of the project's first two Phases. A third Phase, summarised at the end of this report, is now underway.

The LHC experiments have envisaged Computing Models (CM) involving many hundreds of physicists engaged in analysis at institutions around the world. These Models encompass a complex set of wide-area, regional and local-area networks, a heterogeneous set of compute- and data-servers, and a yet-to-be determined set of priorities for group-oriented and individuals' demands for remote data and compute resources. Each of the experiments foresees storing and partially distributing data volumes of Petabytes per year, and to have to provide rapid access to the data over regional, continental and transoceanic networks. Distributed systems of this size and complexity do not exist yet, although systems of a similar size to those foreseen for the LHC experiments are predicted to come into operation by around 2005.

MONARC has successfully met its major milestones, and has fulfilled its basic goals, including:

- identifying first-round baseline Computing Models that could provide viable (and cost-effective) solutions to meet the simulation, reconstruction and analysis needs of the LHC experiments
- providing a powerful (CPU and time efficient) simulation toolset that will enable further studies and optimisation of the Models,
- providing guidelines for the configuration and services of Regional Centres, and
- providing an effective forum where representatives of actual and candidate Regional Centres may meet and develop common strategies for LHC Computing.

The MONARC work, and discussions between MONARC and (actual and candidate) Regional Centre organisations, has led to the concept of a Regional Centre hierarchy as the best candidate for a cost-effective and efficient means of facilitating access to the data and processing resources. The hierarchical layout is also well-adapted to meet the local needs for support in developing and running the software, and carrying out the data analysis with an emphasis on the responsibilities and physics interests of the groups in each world region. In the Summer and Fall of 1999, it was realised that Computational Grid [2] technology, extended to the data-intensive tasks and worldwide scale appropriate to the LHC, could be used to develop the workflow and resource management tools needed to effectively manage such a worldwide-distributed "Data Grid" system.

The earlier progress of MONARC is documented in its Mid-Project Progress Report [3] (June 1999), and the talks by H. Newman and I. Legrand at the LCB Computing Workshop in Marseilles [4,5] (October 1999). The MONARC Technical Notes [6] cover the specifications for possible CERN and regional centre site architectures, regional centre facilities and services, and the testbed studies used to validate and help develop the MONARC Distributed System Simulation, and to determine the key parameters in the candidate baseline Computing Models. A series of papers on: the structure and operational experience with the Simulation system (using the results of the Analysis Process Working Group); the work of the Architectures Working Group; and the testbed studies and simulation validation in local and wide-area network environments, have been submitted to the CHEP 2000 conference [7,8,9,10,11,12].

## 2.3 Regional Centre Model

The “Regional Centre” is a complex, composite object containing a number of data servers and processing nodes, all connected to a LAN. Optionally, it may contain a Mass Storage unit and can be connected to other Regional Centres. Any regional centre can dynamically instantiate 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 be used to distribute jobs to processing nodes.

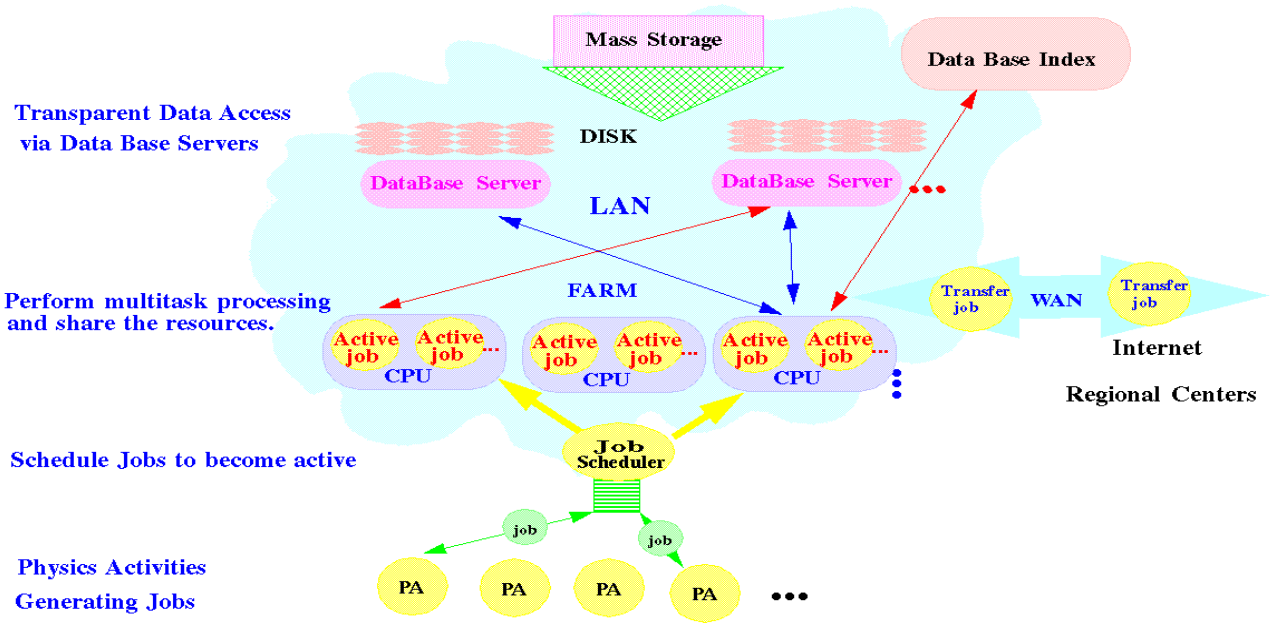


Fig. 2-3 Schematic view of a regional Centre Model.

## 2.4 The Graphical User Interface and Auxiliary Tools

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

The number of regional centres considered can be changed through the main window of the simulation program. The “Global Parameters” frame allows the (mean) values and their statistical distributions for quantities which are common in all Regional Centres to be changed. The hardware cost estimates for the components of the system can also be obtained. For each Regional Centre in the simulation, the user may interactively select the parameters which are graphically presented (CPU usage, memory load, load on the network, efficiency, Database servers' load etc). Basic mathematical tools are available to examine all simulation results: computation of integrated values, mean values and integrated mean values.

To publish or store the simulation results and all the relevant files used in the simulation, an automatic procedure has been developed. This allows publishing locally, or to a MONARC Web server. The Web

Page thus offers a repository for the MONARC Collaboration [16]. There can be found the configuration files, the Java source code used to certain modules and the results (tables and graphic output) for any given simulation runs. The aim of this facility is to provide an easy way to share ideas and results. The publishing procedure is implemented in Java using the Remote Method Invocation mechanism. The schematic view of how this works is shown in Fig. 2-4. A users guide is in preparation.

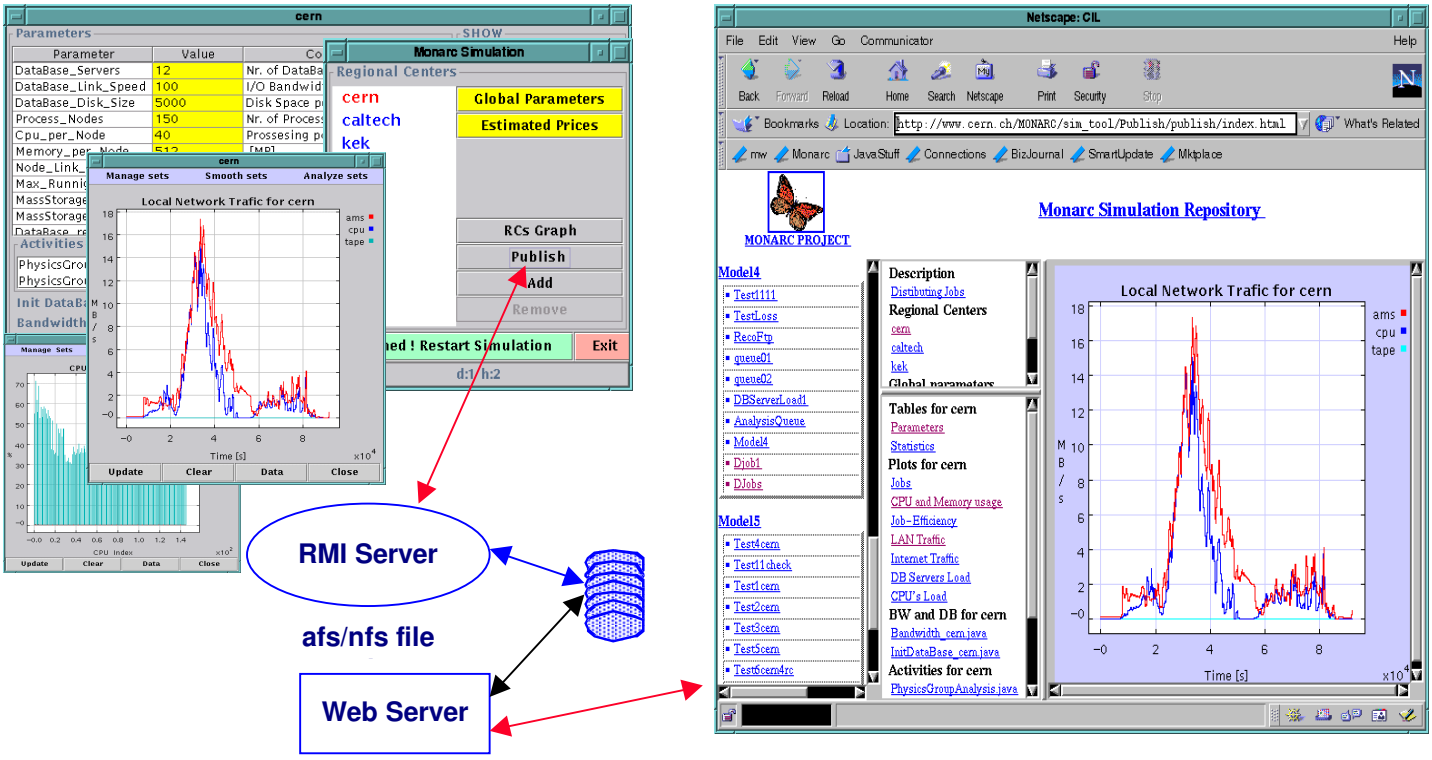


Fig. 2-4 Publishing the simulation results into the web pages

## 4.3 Characteristics of Regional Centres

The various levels of the hierarchy are characterised by services and capabilities provided, constituency served, data profile, and communications profile.

The offline software of each experiment performs the following tasks:

***initial data reconstruction*** (which may include several steps such as preprocessing, reduction and streaming; some steps might be done online); ***Monte Carlo production*** (including event generation, detector simulation and reconstruction); ***offline (re)calibration***; ***successive data reconstruction***; and ***physics analysis***.

To execute the above tasks completely and successfully, both data and technical services are required.

***Data services*** include: (re)processing of data through the official reconstruction program; generation of events; detector response simulation; reconstruction of Monte Carlo events; insertion of data into the database; creation of the official ESD/AOD/tags; updating of the official ESD/AOD/tags under new conditions; ESD/AOD/tag access (possibly with added layers of functionality); data archival/retrieval for all formats; data import and export between different tiers of regional centres (including media replication, tape copying); and bookkeeping (includes format/content definition, relation with Data Base).

***Technical services*** include: database maintenance (including backup, recovery, installation of new versions, monitoring and policing); basic and experiment-specific software maintenance (backup, updating, installation); support for experiment-specific software development; production of tools for data

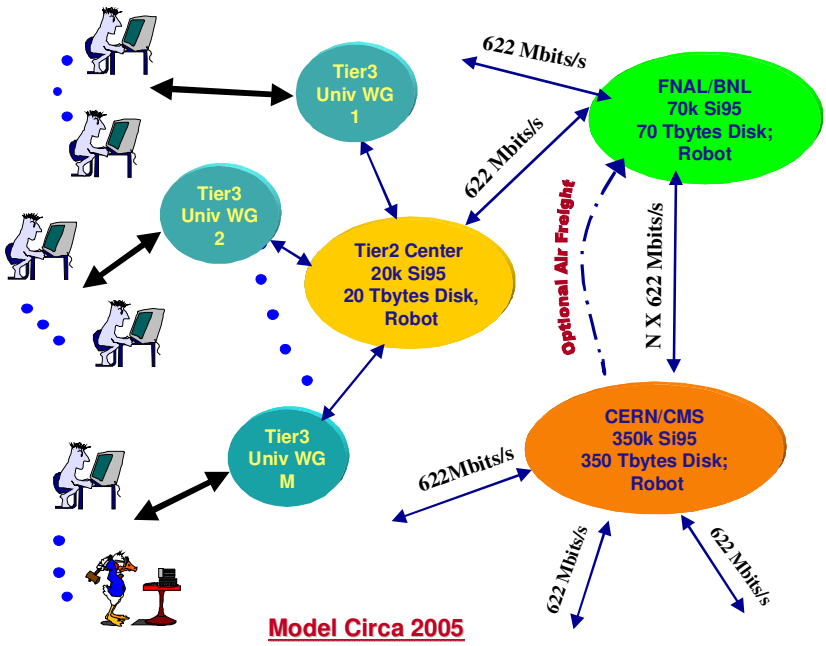


Fig. 4-1 Computing for an LHC Experiment Based on a Hierarchy of Computing Centers. Capacities for CPU and disk are representative and are provided to give an approximate scale).

services; production and maintenance of documentation (including Web pages); storage management (disks, tapes, distributed file systems if applicable); CPU usage monitoring and policing; database access monitoring and policing; I/O usage monitoring and policing; network maintenance (as appropriate); and support of large bandwidth.

## 4.4 Functions of CERN -- the Central Site

The following steps happen at the central site only: online data acquisition and storage; possible data pre-processing before first reconstruction; and first data reconstruction.

Other production steps (calibration data storage, creation of ESD/AOD/tags) are shared between CERN and the regional centres.

The central site holds: a complete archive of all raw data; a master copy of the calibration data (including geometry, gains etc.); and a complete copy of all ESD, AOD, tags possibly online.

The estimate for the amount of data taken is:

- 1 PB raw data per year per experiment
- $10^9$ events (1 MB each) per year per experiment
- 100 days of data taking (i.e. $10^7$ events per day per experiment)

Current estimates for a single LHC experiment capacity to be installed by 2006 at CERN are given in [25].

In the following, resources for the regional centre will be expressed in terms of percentage of the resources available at CERN as specified in the above document.

## 4.5 Configuration of Tier-1 Regional Centres

Architectural diagrams of a typical regional centre are shown in Figure 4-2, 4-3, and 4-4. These are not meant to be physical layouts, but rather logical layouts showing the various work-flows and data-flows performed at the centre[1]. In particular services, work-flows and data-flows could be implemented at a single location or distributed over several different physical locations connected by a high performance network.

The overall architecture is shown in Fig. 4-2. Production services are shown in the upper 80% of the diagram and consist of data import and export, disk, mass storage and database servers, processing resources, and desktops. Support services are arrayed along the bottom of the chart, and include physics software development, R&D systems and test-beds, information and code servers, web and tele-presence servers, and training, consulting, and helpdesk services.

[1] Numbers in these figures are based on our current understanding of ATLAS and CMS requirements. The requirements are still being studied, and these numbers will certainly change.

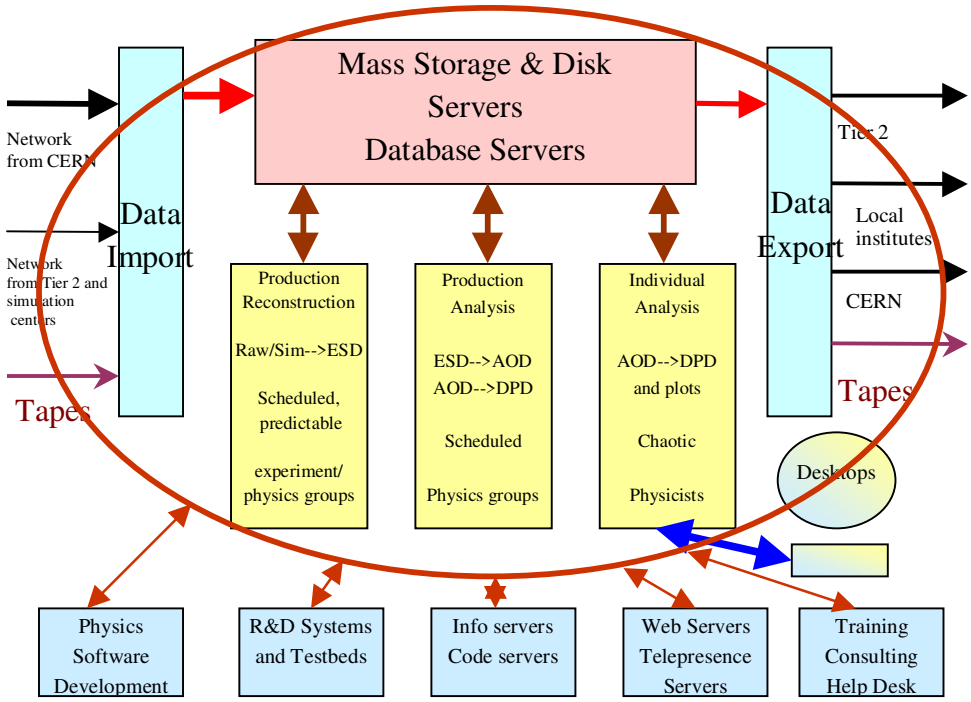


Fig. 4-2: Overall Architecture of a possible Regional Centre

Fig. 4-3 charts the workflow at the centre, with individual physicists, physics groups and the experiment as a whole submitting different categories of reconstruction and analysis jobs, on both a scheduled and spontaneous basis. Shown also are the characteristics of these jobs and an indication of the scale of resources required to carry them out. Fig. 4-4 shows an overview of the data-flow at the centre, where data flows into the central robotic mass storage from the data import facility (and out to the data export facility), and moves through a central disk cache to local disk caches on processing elements and desktops.

## 4.6 Tier-2 Centres

A Tier-2 regional centre is similar to a Tier-1 centre, but on a smaller scale; its services will be more focused on data analysis. Tier-2 centres could be seen as "satellites" of a Tier-1 with which they exchange data. A Tier-2 regional centre should have resources in the range 5 to 25 % of a Tier-1 regional centre.

# Chapter 5: Main Modelling Results

## 5.1 Scope of Modelling

The most important goal of the MONARC project was to develop a set of viable baseline models for the LHC experiments' computing systems. A set of data reconstruction jobs, physics analysis jobs and data transfers needed to satisfy the analysis jobs' database queries, and the data replications required to maintain coherence of the continuously updated federated database, has been defined. Each set satisfies the user requirements defined by the MONARC Analysis Working Group, and allows physicists to access the required amount of data in the desired time.

Tape handling and its I/O capability under a multi-user environment could be one of the most crucial

aspects in the LHC experiments. Although a model of tape robotics has been implemented in the MONARC simulation tools, one needs detailed use cases of data access patterns and the realistic time response of tape drives and robotics to perform reliable and viable modelling. In the phase 2 of this project, we modelled other hardware components such as CPU farm, disk, and the bandwidth of wide area networks. The modelling effort of the tape robotics will continue in the next phase of the project, based on the real use cases of the analysis program developed by each LHC experiment. Detailed plans for evaluating use cases for each experiment are summarised in Chapter 7.

Some sets of the defined jobs have been executed both in a centralised computing system with just one centre (CERN), and in a distributed system with a number of Regional Centres. Having fixed the set of activities to be performed, one can evaluate with the existing models the hardware resources and the network bandwidth needed to finish all jobs in the required time. Both central and distributed classes of models have been shown to be feasible with the CPU, disk and network resources that are within those expected to be available in 2005. The models, together with all the results obtained in the simulation runs, are available on the MONARC Simulation and Modelling Working Group Web pages [16].

## 5.2 Data Model

A hierarchical data model, similar to those developed within the ATLAS and CMS collaborations, has been incorporated in the MONARC simulations. The experiment's data events are written at the central site, CERN, as RAW objects of the size of 1 MBytes/event. After the full reconstruction, the event summary data (ESD) objects are created of the size of 100 kBytes/event, as well as the analysis object data (AOD) of the size of about 10 kBytes/event, and the TAG objects of about 100 Bytes/event[2]. The full reconstruction is expected to take place twice a year. Redefinition of the AOD and TAG objects, based on re-analysis of the ESD data, is expected to take place once per month.

The baseline models developed by MONARC are all based on a hierarchical set of computing centres. The CERN computing centre will store all data types: RAW, ESD, AOD and TAG. The Tier-1 Regional Centre (RC) will have replicas of the ESD, AOD and TAG; the Tier-2 RC only AOD and TAG. The individual physicists may have just TAG at their desktops and possibly private collections of events in various data formats. It would be possible to introduce variations on the above model, for example by allowing subsets of ESD data at the Tier-2 RC's, or subsets of RAW data at Tier-1 RC's, etc.

The smallest unit of the simulated federated database is a container, or a file. A single integer, an event number, is the basis of the simulated event catalogue. It allows a unique mapping of objects of various types to data containers (files) and to distribute them among numerous data servers (AMS servers). The system is capable of identifying the files and the data servers that contain an event, or a range of events, as defined by their event numbers. The current implementation of the data model allows simulating even very complicated data queries that follow associations between TAG->AOD, AOD->ESD and ESD->RAW. The user-defined factors describing the frequency of such traversals across different data types are parameters of the data model.

[2] Object sizes are current estimates, and are subject of modifications by each experiment. Study of CMS ORCA, for example, has recently shown that ESD is more likely close to 500 kBytes/event for CMS.

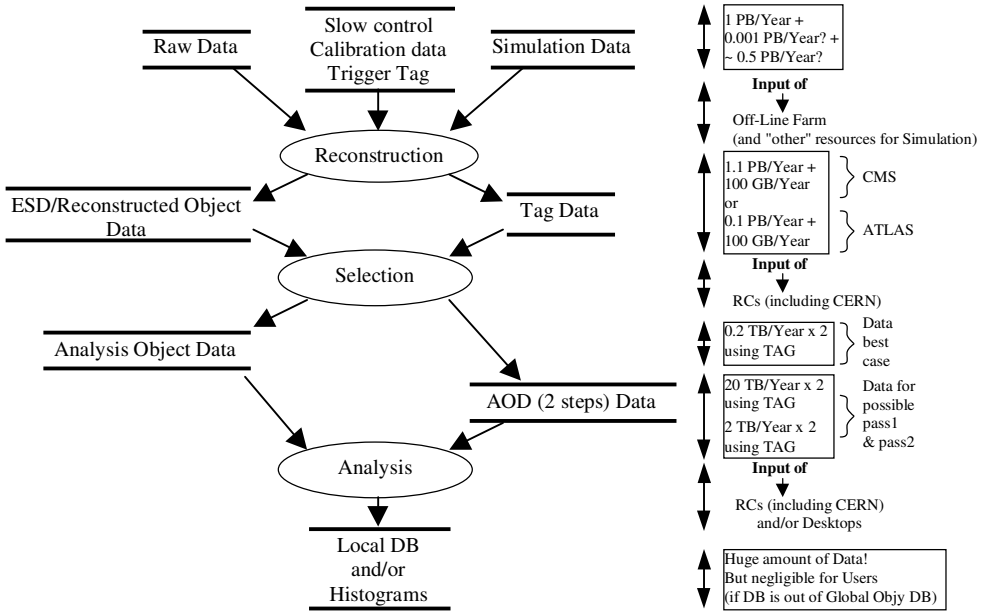


Fig. 5-1 Main tasks of the analysis process and a data flow model.

## 5.3 Analysis Activities and Data Access Patterns

There are several phases in the analysis of an experiment's data. The first is to reconstruct the RAW data and to create the first version of Event Summary Data (ESD), Analysis Object Data (AOD) and TAG objects ("pass-1" analysis). Each Physics Analysis Group will then define its standard data-set, and finally physicists will run their physics analysis jobs. AOD and TAG data are expected to be re-defined more often than ESD ("pass-2" analysis). The frequency of each of the operations, the volume of input and output data, and the amount of computing hardware resources needed to accomplish the task are the most important parameters of a LHC experiment computing model. In Figure 5.1 we present the main tasks of the analysis process, and a sketch of the resulting data flow model.

### 5.3.1 Reconstruction of RAW data.

These jobs create the ESD (Event Summary Data objects), the AOD and the TAG data-sets based on the information obtained from a complete reconstruction of RAW data that has been already recorded. The newly created ESD, AOD and TAG are then distributed (by network transfers, or other means) to the participating Regional Centres. This is an Experiment Activity. It is assumed that experiments should be able to perform a full reconstruction of the RAW data and distribution of the ESD, AOD and TAG data, 2-4 times a year.

### 5.3.2 Re-definition of AOD and TAG data.

This job re-defines the AOD and the TAG objects based on the information contained in the ESD data. The new versions of the AOD and TAG objects are then replicated to the participating Regional Centres by network transfers. This is an Experiment Activity that is expected to take place with a frequency of about once per month.

### 5.3.3 Selection of standard samples within Physics Analysis Groups

This class of jobs performs a selection of a standard analysis group sample, a subset of data that satisfies a set of cuts specific to an analysis group. Event collections (subsets of the TAG database or the AOD database with only the selected events, or just pointers to the selected events) are created. Re-clustering of the objects in the federated database might be included in this Analysis Group activity.

### 5.3.4 Generation (Monte Carlo) of "RAW" data set.

This job creates the RAW-like data to be compared with real data. These jobs can be driven by a specific analysis channel (single signal) or by the entire Collaboration (background or common signals). This is an Analysis Group or an Experiment Activity, and can take place both at CERN and at Regional Centres.

### 5.3.5 Reconstruction of "RAWmc" events to create ESDmc, AODmc and TAGmc.

This job is very similar to the real data processing. Since RAWmc may be created not only at CERN the reconstruction may take place at the Regional Centres where the data had been created. The time requirements of the reconstruction of these events are less stringent than for the real RAW data.

### 5.3.6 Re-definition of the Monte Carlo AOD and TAG data.

This job has the same characteristic of the ones at 5.3.2 and 5.3.5. The difference may be in the need for the final analysis to access the original simulated data (the "Monte Carlo truth") at the level of the kinematics or the hits for the purpose of comparison.

### 5.3.7 Analysis of data sets to produce physical results.

These jobs start from data-sets prepared for the respective analysis groups, accessing Event Collections (subsets of TAG or AOD data-sets), and follow associations (pointers to objects in the hierarchical data model – TAG->AOD, AOD->ESD, ESD->RAW) for a fraction of all events. Individual physicists, members of Analysis Groups submit these analysis jobs. In some cases, co-ordination within the Analysis Group may become necessary. Analysis jobs are examples of Individual Activities or Group Activities (in the case of enforced co-ordination).

### 5.3.8 Analysis of data sets to produce private working selections.

This job is a pre-analysis activity, with a goal to isolate physical signals and define cuts or algorithms (Derived Physics Data). These jobs are submitted by individuals physicist, and may access higher data hierarchy following the associations, although (as test jobs) they require perhaps a smaller number of events than Analysis jobs described in 5.3.7. These jobs are examples of Individual Activities.

The main characteristics of the major analysis tasks, such as the frequency with which the tasks will be performed, the number of tasks run simultaneously, the CPU/event requirements, the I/O needs, the needed time response et cetera, are summarised in Table 5-1.

### 5.3.9 Regional Centres and the Group Approach to the Analysis Process

The analysis process of experiments data follows a hierarchy: Experiment->Analysis Groups->Individual Physicists. A typical Analysis Group may have about 25 active physicists. Table 5-2 gives a summary of the “Group Approach” to the Analysis Process.

| | Full reconstruction | | Re-Define AOD/TAG | | Define Group data-sets | | Physics Analysis Job | |
|---|---|---|---|---|---|---|---|---|
| | **Value used** | **range** | **Value used** | **Range** | **Value used** | **Range** | **Value used** | **Range** |
| Frequency | 2/year | 2-6/year | 1/month | 0.5-4/month | 1/month | 0.5-4/month | 1/day | 1-8/day |
| CPU/event (SI95*s) | 250 | 250-1000 | 0.25 | 0.1-0.5 | 25 | 10-50 | 2.5 | 1-5 |
| Input data | RAW | RAW | ESD | ESD | DB query | DB query | DB query | DB query |
| Input size | 1 PB | 0.5-2 PB | 0.1 PB | 0.02-0.5 PB | 0.1 PB | 0.02-0.5 PB | 0.1-1TB (AOD) | 0.001-1TB (AOD) |
| Input medium | DISK | TAPE/DISK | DISK | DISK | DISK | DISK | DISK | DISK |
| Output data | ESD | ESD | AOD | AOD | Collection | Collection | – | Variable |
| Output size | 0.1 PB | 0.05-2 PB | 10TB(aod) 0.1TB(tag) | 10 TB(aod) 0.1-1TB(tag) | 0.1-1TB (AOD) | 0.1-1TB (AOD) | – | Variable |
| Output medium | DISK | DISK | DISK | DISK | DISK | DISK | – | DISK |
| Time response (T) | 4 months | 2-6 months | 10 days | 5-15 days | 1 day | 0.5-3 days | 12 hours | 2-24 hours |
| Number of jobs in T | 1 /experiment | 1 /experiment | 1 /experiment | 1 /experiment | 1/Group | 1/Group | 20 / Group | 10-100 / Group |

Table 5-1 Characteristics of the main analysis tasks

| **LHC Experiments** | Value USED | Range |
|---|---|---|
| Number of analysis groups (WG) | 20/experiment | 10-25/experiment |
| Number of members per group | 25 | 15-35 |
| Number of Tier-1 Regional Centres (including CERN) | 5/experiment | 4-12/experiment |
| Number of Analyses per Regional Centre | 4 | 3-7 |
| Active time of Members | 8 Hour/Day | 2-14 Hour/Day |
| Activity of Members | Single regional centre | More than one regional centre |

Table 5-2 Summary of the "Group Approach" to the Analysis Process.

The concept of a distributed computing system, with a number of Regional Centres distributed in the world, each with replicas of the AOD, TAG and a partial or complete (depending on the needs) of the ESD, maps very well to the Analysis Group approach to Analysis Process. Physicists working on the same analysis tend to work together as sophisticated analyses require joint effort of faculty, post-doctoral research associates and students. It is difficult to imagine that all physicists involved in physics analyses could move to CERN! It is highly probable that Tier-1 Regional Centres will become focal points for different analysis efforts. The original motivation to create Tier-1 Regional Centres, namely, to provide faster and more efficient access to the experiments' data by exploiting the anticipated better WAN network bandwidth in a given region, as compared to the WAN connection to CERN, gains importance if the physics analyses are distributed world-wide as well.

## 5.4 Parameters of the Model

A complete list of global and local parameters that characterise the federated database, regional centre configuration and the data model used in MONARC simulation is presented in Appendix A.

## 5.5 Description of the simulated activities

The Baseline Models that have been built by MONARC simulate the following activities that will be performed at CERN, and/or at the Regional Centres:

i. individual physicist's analysis jobs (at all participating centres)

ii. analysis groups' selection jobs, which define the standard analysis samples (at all participating centres)

iii. reconstruction of RAW data at CERN, which leads to creation of the new ESD, AOD and TAG data

iv. re-processing of ESD data at CERN, which leads to creation of the new AOD and TAG data

v. replication of new ESD, AOD and TAG data from CERN to all Regional Centres, using an ftp-like transfer protocol

vi. generation and reconstruction of RAW Monte Carlo events at Tier-1 RC's, and of ESD Monte Carlo events at Tier-1 (or Tier-2) RC's

vii. reconstruction of Monte Carlo events generated at Tier-1 RC's

viii. generation of the "fast Monte Carlo" events at Tier-2 RC's

The number of events to be processed by various jobs, and the elapsed time in which the jobs should be finished were defined by the MONARC Analysis Working Group. For example, the full reconstruction of RAW data should be done twice a year, and the re-definition of AOD once a month, with the task itself taking no more than 10 days, etc.

| | **RAW** | **ESD** | **AOD** | **TAG** | **Monte Carlo** |
|---|---|---|---|---|---|
| **Number of events in RAW, ESD, AOD, TAG and Monte Carlo data types and their location** | | | | | |
| #events, location | 1,000,000,000 CERN | 1,000,000,000 each Tier1:locally each Tier2:at Tier1 | 1,000,000,000 each RC: locally | 1,000,000,000 each RC: locally | 100,000,000 each Tier1:locally |
| **Volume of replicated data (ftp); number of events and data volume accessed by analysis activities** | | | | | |
| **Reconstruction** input events Accessed per day | 6,000,000 | – | – | – | 1,000,000 |
| **Reconstruction output** | No | Yes | Yes | Yes | Yes |
| **FTP transfers (replication)** | | 0.6 TB to each Tier1 centre | 60 GB to each Tier1 and Tier2 RC | 600 MB to each Tier1/Tier2 RC | 100 GB from each Tier1 RC to CERN |
| **Definition of AOD / input** | – | 100,000,000 events/day | – | – | |
| **Definition of AOD / output** | – | – | Yes | Yes | |
| **FTP transfers (replication)** | | | 1 TB to each Tier1 and Tier2 RC | 10 GB to each Tier1/Tier2 RC | |
| **Number of events and data volumes of different type to be accessed per day by different analysis activities** | | | | | |
| **Physics Group Selection job** (data accessed per single job); 20 jobs running, 1 per analysis group. | 0.001% of 1,000,000,000 (per job) (0.01 TB/job) | 0.1% of 1,000,000,000 (per job) (0.1 TB/job) | 10% of 1,000,000,000 (per job) (1 TB/job) | 100% of 1,000,000,000 (per job) (100 GB/job) | |
| **Physics Analysis** (data accessed per single job); 200 jobs running, 10 jobs per analysis group. | 0.01%of AOD data (per job) (on average 0.045 TB/job) | 1% of AOD data (per job) (on average 0.45 TB/job) | Follow 100% of the group set (per job) (on average 0.45 TB/job) | Group Data-set: 1-10 % of all TAG objects (per job) (on average 4.5 GB/job) | |

Table 5-3 Model of Daily Activities of the Regional Centres

Here we consider a specific model called the Distributed Daily Activity Model (DDAM). In this model, activities in a typical 24-hour period in an LHC experiment were considered, such as reconstruction, analysis and data replication. There are 20 different analysis groups in this model. Each analysis group could have different standard data samples. For 10 analysis groups their standard data-set contained 1% of the number of events, for 5 groups the data-set contained 5% of the total number of events, and for the remaining 5 the data-sets contained 10% of all events (on average 4.5% per analysis group). Details of the data access involved in the tasks to be performed in the model of Daily Activities of the Regional Centres is presented in Table 5-3, together with the number of events that are processed (per day) to satisfy the experiment and user requirements.

In a model that describes a fully centralised scenario, all jobs are run at one centre (CERN). In a model of a distributed computing system architecture (in the case of DDAM there are 5 Tier-1 Regional Centres and a single Tier-2 Centre), the analysis jobs are distributed among all participating Regional centres, while the reconstruction jobs are run at CERN only.

## 5.6 Results and conclusions

### 5.6.1 Results and group repository

All the results obtained in the baseline models developed by the MONARC Collaborations have been made available ("published") in the MONARC Simulation and Modelling Group repository [16]. The files needed to construct the models, run the simulation jobs, and verify the results are available from the Web pages. A detailed presentation of results of MONARC simulations can be found in a paper presented at CHEP2000 [12].

A fully centralised model (with all the activities taking place at CERN), and partially distributed models (with a number of Tier-1 and Tier-2 Regional Centres), were simulated. Their performance was evaluated with a fixed set of tasks, with the requirement that all simulated activities had to be finished in a desired time interval (one or two days, depending on the models). Various levels of optimisation of load balancing of the CPU and database access speed were tried and evaluated. The resources necessary to complete the specified set of tasks (CPU, memory, network bandwidth, distribution of RAW, ESD, AOD and TAG data among the multiple data servers, etc.) were adjusted until the system was capable of finishing all jobs in the desired time. The optimisation was performed "by hand", i.e. the parameters of a particular model were changed, the new simulation was run and the results examined. The final parameters used for the DDAM are reported in Appendix A.

With the set of tasks to be performed and the elapsed time in which all tasks should finish fixed, the cost of the system is the variable that reflects the quality of solution. Also, the amount of resources necessary to accomplish the required tasks should be within the expected limits. To first order, the difference in cost of hardware between the fully centralised and the partially distributed scenarios is the additional price for storage media and data servers for replicated ESD, AOD and TAG data. However, a distributed computing system with replicas of parts of data may be more flexible and in the sense that the load of analysis jobs is also distributed. Data I/O are also distributed to different servers and therefore more robust against bottleneck operations. It should be emphasised that in both classes of models we found that the required resources did not exceed the planned CPU, memory, data-server I/O and network bandwidth of the computing systems for the LHC experiments.

At present, no serious price versus performance comparison between the centralised and distributed computing models is available, as only the hardware costs and the network connection costs are included in the cost function. However, with a more complete cost function that will include travel costs and, more importantly, quantify differences in the human aspects of different architectures of computing systems, finding an optimal solution should be possible.

All the results should be treated as preliminary. For example, tape handling is not covered yet. The baseline models describe mature experiments, in which all the data has been already reconstructed at least once, and with ESD, AOD and TAG data available at all Regional Centres. The models will evolve in the direction of automatic load balancing and resource optimisation. However, the three main

conclusions that emerge from the simulations performed with the current baseline models are unlikely to change. These are summarised in the following sub-sections.

### 5.6.2 Network implications

For a distributed computing system to function properly (i.e. to support the data transfers requested by the analysis jobs, and, simultaneously, the data transfers necessary to replicate the ESD, AOD and TAG objects) a network bandwidth of 30 MBytes/s between CERN and each of the Tier-1 Regional Centre is required. Of course one must take into account that this bandwidth requirement may well be competing with other demands for the total available bandwidth. On the other hand, one can envisage replicating ESD, or AOD data by means other than network transfers such as shipping tapes or CDs, as has been assumed in the baseline models. Such a hybrid (network and non-network) replication scheme would reduce the demand for the network bandwidth.

However, the current results suggest that it should be possible to built a useful distributed architecture computing system provided the availability of CERN->Tier-1 Regional Centre network bandwidth is of the order of 622 Mbps per Regional Centre. This is an important result, as all the projections for the future indicate that such connections should be commonplace in 2005. This means that distributed computing systems will be technologically viable at the time when the LHC experiments will need them. A preliminary but similar result was also obtained for a minimum bandwidth requirement of a Tier-2 to Tier-1 connection. To answer the question of how the Tier-2 centres will function requires a further study. In Figure 5-2 we present the plots obtained with the DDAM showing the WAN traffic as a function of time.

In this figure, the plot to the left presents our simulation of the WAN traffic between CERN and any of the 5 Regional Centres that are part of the partially distributed computing system. The assumed 30 MBytes/s bandwidth is close to being fully saturated for all connections. ("Caltech2" is a Tier-2 regional centre, while all others are Tier-1 regional centres.) In the plot on the right, the WAN traffic for one of participating centres to all other centres is shown. Here, only the connection to CERN is active (if data is unavailable at a Regional Centre, then the database associations point to data at CERN). One can clearly see that the assumed bandwidth of 30 MB/s is almost fully saturated.

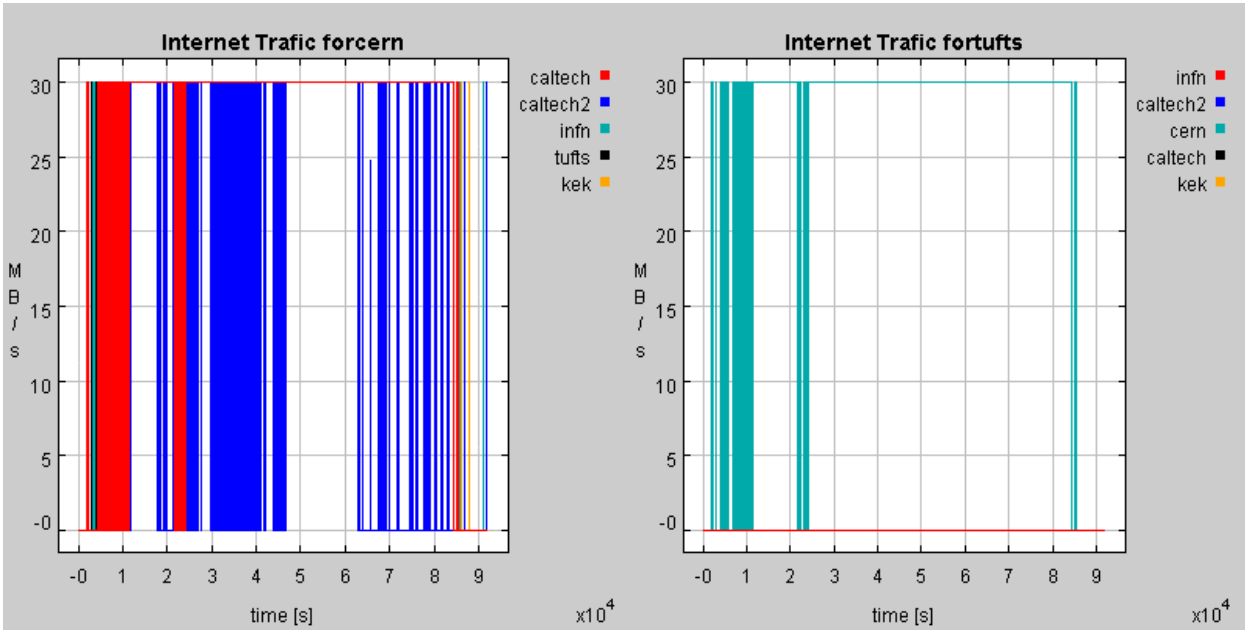


Fig. 5-2 Wide area network traffic activities in DDAM (Distributed Daily Activity Model)

### 5.6.3 Optimisation of job submission scheme

The results of simulations with the baseline models indicate clearly that load balancing through optimising job submission is a very important factor in tuning the performance and cost of the system. For the systems with large number of CPUs, it is much better to submit many jobs, each processing a smaller number of events, rather than submit a few jobs each processing an enormous amount of events. Such optimisation of the job submission process exploits the stochastic nature of the problem, and leads to a much better utilisation of the distributed resources. This result is easy to understand intuitively as one can much easier keep all the CPU's active with many small jobs. We have found that one could reduce the overall cost of the system by a large factor (2-4) simply by load balancing the CPU by optimisation of job submission. Without optimisation, one would have to provide more, or faster, CPU power in order to finish the jobs in required time. At present, optimisation of job submission was performed in each Regional Centre independently, with jobs being submitted at a local Regional Centre. However, one could consider system-wide load balancing schemes, which could lead to still greater gains in optimised utilisation of resources.

### 5.6.4 Load-balancing of database servers (AMS servers)

It was found that it is also very important to balance the load on the data servers. Failing to distribute containers (files) of different types of data among the data servers uniformly may lead to significant bottlenecks, which in turn may lead to increases in the time it takes to finish the assumed set of tasks (easily by a factor of 2). Jobs sit idle in memory waiting for data to arrive from the data servers (AMS servers), as can be seen in Figure 5-3. This points to a need for careful design of the federated database layout, and a need for dedicated simulations of the future CERN, Tier-1 and Tier-2 data management systems in order to maximise the effective I/O throughput.

In this figure, the CPU/memory utilisation plot as a function of time for the CERN centre with non-optimised AMS servers is shown in the upper-left plot and better optimised AMS servers in the upper-right plot. In either case, the same set of jobs was submitted. Also shown are: AMS read load for non-optimised case (lower-left plot); and better optimised case, in which data was more evenly distributed among servers (lower-right plot).

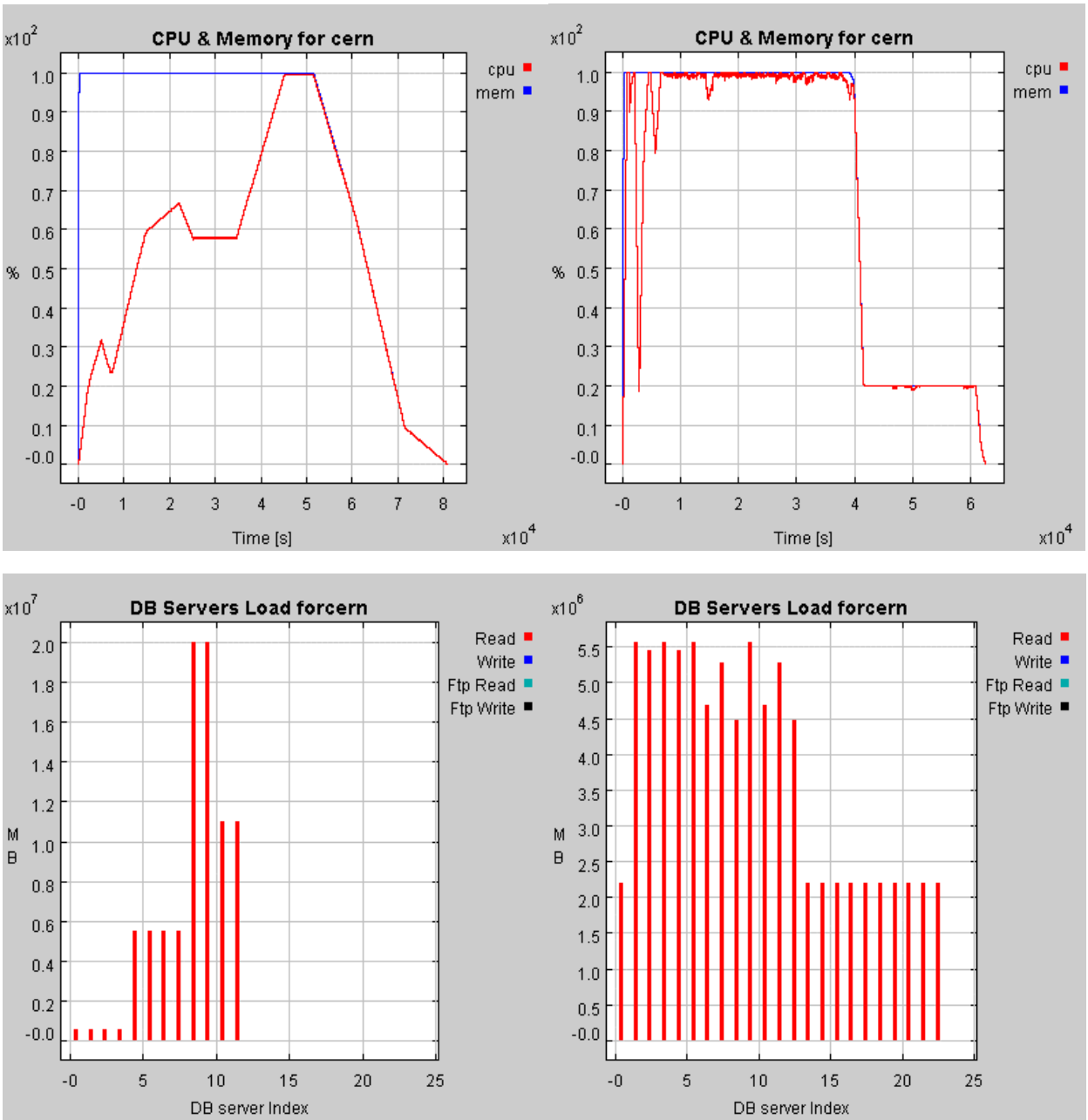


Fig.5-3 Effect of load-balancing in the database servers

# Chapter 6: Phase 2 Conclusions

The main conclusions of the MONARC work can be summarised as follows:

a) MONARC has demonstrated that a hierarchical model of computing resource distribution, based on Regional Centres, is feasible. There exists willingness in many national communities to participate in the development of a computing infrastructure such as the one defined by the MONARC model, and the infrastructures needed to develop and deploy the MONARC computing model are potentially available. The MONARC model seems capable of dealing with the computing needs of the experiments.

b) The MONARC simulation tool has proven to be an excellent instrument for computing model studies. The basic elements of LHC computing, object database and wide-area network performances, jobs activities and resource utilisation have been implemented, with flexible facilities for varying key model parameters. The simulation results have permitted assessment of the matching between a given set of resources and a static load of activities. The modularity of the simulation tool permits the easy addition of new modelling blocks for iterative validation steps.

c) The further development of realistic models, suitable for implementation and optimised for resources and performance, including the large-scale mass storage system, will require the following steps:

    1. Development of middleware for farm management, allocation of resources, query estimation and priority setting, network monitoring, etc. The GRID projects are expected to be able to provide several of these fundamental tools.

    2. Set up of use-cases and realistic prototypes by the experiments. Feedback from the these test-benches will allow the iterative development and refinement of the model

    3. Simulation tool use (and further development), for helping in the optimisation of the system and in the identification and solution of possible bottlenecks.

MONARC Phase-3 will address the issues summarised in c) above, with the aim of favouring the best possible synergy with the initial phase of the planned EU-GRID project and with the ongoing US-GRID activities.

# Appendix A : Global and local parameters used in models.

Table A.1: A list of global parameters currently in use by baseline models built with the MONARC simulation tool (Two-Day Activities Model [34] and Daily Activities Model [35])

| **federated database and data model parameters (global)** | | |
|---|---|---|
| **Global parameter name** | **Daily Activities model** | **2-day activities model** |
| Database page size | 64 kB | 64 kB |
| TAG object size/event | 100 B (neg.exp) | 100 B |
| AOD object size/event | 10 kB | 10 kB |
| ESD object size/event | 100 kB | 100 kB |
| RAW object size/event | 1 MB | 1 MB |
| Processing time RAW->ESD | 250-500 SI95*s | 500-1000 SI95*s |
| Processing time ESD->AOD | 25 SI95*s (normal) | 25 SI95*s |
| Processing time AOD->TAG | 2.5 SI95*s (normal) | 5 SI95*s |
| Analysis time TAG | 0.25 SI95*s (normal) | 3 SI95*s |
| Analysis time AOD | 2.5 SI95*s (normal) | 3 SI95*s |
| Analysis time ESD | 25 SI95*s (normal) | 15 SI95*s |
| Generate RAWmc | – | 5000 SI95*s |
| Generate ESD | – | 1000 SI95*s |
| Generate AOD | – | 25 SI95*s |
| Generate TAG | – | 5 SI95*s |
| Memory for RAW->ESD processing job | 200 MB | 100 MB |
| Memory for ESD->AOD processing job | 200 MB | 100 MB |
| Memory for AOD->TAG processing job | 200 MB | 100 MB |
| Memory for TAG analysis job | 200 MB | 100 MB |
| Memory for AOD analysis job | 200 MB | 100 MB |
| Memory for ESD analysis job | 200 MB | 100 MB |
| Container size RAW | ~200 GB | 10 GB |
| Container size ESD | ~5.4 GB | 10 GB |
| Container size AOD | ~ 3 GB | 10 GB |
| Container size TAG | ~30 MB | 10 GB |

Table A.2: A list of local parameters currently in use by baseline models built with the MONARC simulation tool (Daily Activities Model and Two Day Activities Model)

| **Regional centre configuration parameters (local)** | | |
|---|---|---|
| **LOCAL parameter name** | **Daily Activities model** | **2-day activities model** |
| AMS link speed | 200 MB/s | 100 MB/s |
| AMS disk size | 125 TB | 20-100 TB |
| Number of AMS servers | 85 (CERN); 37 Tier1 RC | 10-58 |
| Number of processing nodes | 600 (CERN); 200 at Tier1 RC | 20-1000 |
| CPU/node | 500 SI95 | 500 SI95 |
| Memory/node | 200 MB | 1 MB |
| Node link speed | 50 MB/s | 10 MB/s |
| Mass storage size (in HSM) | 1000 TB (0 for Tier1 RC) | 50-1000 TB |
| Link speed to HSM | 2000 MB/s (0 for Tier1 RC) | 100 MB/s |
| AMS write speed | 200 MB/s | 100 MB/s |
| AMS read speed | 200 MB/s | 100 MB/s |
| Network bandwidth to/from each RC | 30 MB/s | 40 MB/s |

Table A.3: A list of local parameters (that could be defined per activity or even per job) defining database queries (following associations between objects) currently in use by baseline models built with the MONARC simulation tool (Daily Activities Model and Two Day Activities Model.

| **Data access pattern parameters (local)** | |
|---|---|
| Fraction of events for which TAG->AOD associations are followed | 10-100% |
| Fraction of events for which AOD->ESD associations are followed | 1% |
| Fraction of events for which ESD->RAW associations are followed | 1% |
| Clustering density parameter | Unused |

# Appendix B[3] : Motivations for MONARC Phase 3

The motivations for MONARC Phase 3 were spelled out in the Progress Report in June 1999:

***"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:

- facilitating 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)
- providing a modelling consultancy and "service" to the experiments and Centres
- providing a core of advanced development activities aimed at system optimisation, and pre-production prototyping
- taking advantage of MONARC's synergy with (and complementary to) the work on distributed data-intensive computing systems beginning this year in other "next generation" R&D projects[4], such as those on Grid Computing.

The Phase 3 study will be aimed at maximising the workload sustainable by a given set of networks and site facilities, while reducing the long turnaround times for certain data analysis tasks. Unlike Phase 2, the optimisation of the system in Phase 3 would no longer exclude long and involved decision processes, where a momentary lack of resources or "problem" condition could be met with a redirection of the request, or with other fallback strategies. These techniques could result in substantial gains in terms of work accomplished or resources saved.

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.

---

[3] This appendix is extracted from the MONARC Phase 3 Letter of Intent, submitted to H. Hoffmann and M. Delfino January 15, 2000.

[4] Details on the synergy between a MONARC Phase 3 and R&D projects such as the recently approved "Particle Physics Data Grid" (PPDG) project and the proposed "GriPhyN" (Grid Physics Network) project may be found at http://www.cern.ch/MONARC/docs/progress_report/longc7.html. Also see the PPDG and GriPhyN Websites at http://www.cacr.calech.edu/ppdg and http://www.phys.ufl.edu/~mre/.

MONARC in Phase 3 could exploit the studies, system software developments, and prototype system tests scheduled by the LHC experiments during 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 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.

## Goals and Scope of MONARC Phase 3

MONARC Phase 3's central goal is to develop more realistic Computing Models meeting the LHC Computing Requirements than were possible in the Project's first two phases. This goal will be achieved by confronting the Models with realistic large scale "prototypes" at every stage, including the large scale trigger, detector and physics performance studies that will be initiated by some of the experiments in the coming year. By assessing these "Use Cases" involving the full simulation, reconstruction and analysis of multi-Terabyte data samples[5], MONARC will able to better estimate the baseline computing, data handling and network resources needed to handle a given data analysis workload.

During Phase 3, MONARC will participate in the design, setup, operation and operational optimisation of the prototypes. The analysis of the overall system behaviour of the prototypes, at the CERN site and including candidate Regional Centre sites, will drive further validation and development of the MONARC System Simulation. This is expected to result, in turn, in a more accurate evaluation of distributed system performance, and ultimately in improved data distribution and resource allocation strategies. Strategies that will be recommended to the experiments before their next round(s) of event simulation, reconstruction and analysis studies.

As a result of this mutually beneficial "feedback", we also expect to obtain progressively more accurate estimates of the CPU requirements for each stage of the analysis, and of the required data rates in and out of storage and across networks. We also expect to learn, in steps, how to optimise the data layout in storage, how to cluster and re-cluster data as needed, how to configure the data handling systems to provide efficient caching, and how to implement hierarchical storage management spanning networks, in a multi-user environment.

In addition to the large scale studies of simulated events initiated by the LHC experiments, MONARC will develop its own specific studies using its Testbed systems[6] to explore and resolve some of the problems and unexpected behaviours of the distributed system that may occur during operation of the large-scale prototypes. These in-depth studies of specific issues and key parameters may be run on the MONARC testbeds alone, if adequately equipped, or in tandem with other large computing "farms" and "data servers" at CERN and elsewhere.

In the course of studying these issues using testbeds and prototype systems, we expect to identify effective modes of distributed queue management, load balancing at each site and between sites, and the use of "query estimators" along with network "quality of service" mechanisms to drive the resource management decisions.

One technical benefit for the HEP and IT communities that will result from MONARC Phase 3 is the development of a new class of interactive visualisation and analysis tools for the distributed system simulation. This work, based on new concepts developed by MONARC's chief simulation developer I. Legrand, has already begun during MONARC Phase 2. Based on the initial concepts and results, we are confident that by the end of Phase 3 we be able to make available a powerful new set of Web-enabled visual tools for distributed system analysis and optimisation, that will be applicable to a broad range of

[5] One example, related to the CMS High Level Trigger studies, is described briefly below.
[6] At CERN, in Italy, Japan, and the US.

scientific and engineering problems.

## Large-Scale Prototype Examples

Following the ORCA3 software release[7] and the CMS High Level Trigger (HLT) 1999 milestone, it became evident that a co-ordinated set of future ORCA releases and HLT studies of increasing size (in terms of the numbers of events and data volumes) and sophistication would be required. In order to carry out the ORCA4 release of the software and the subsequent HLT study in the first half of 2000, two of CMS' major milestones have been advanced to next Spring:

- Simulation of data access patterns March 2000
- Integration of databases and mass (tape) storage March 2000,

where we will use large volumes of "actual" (fully simulated and reconstructed) data[8]. This has led to a strong and immediate demand for MONARC's help with the design and optimisation of data structures, data access strategies, and resource management, to make good use resources at some Regional Centre sites as well as CERN.

In a similar vein, ATLAS is planning large-scale studies using large samples of GEANT4 data, and ALICE is planning a series of increasingly large "data challenges".

In the course of MONARC-assisted studies such as these, working closely with the experiments, MONARC is confident that it will be able to progressive develop more realistic Computing Models, and more effective data access and handling strategies to support LHC data analysis.

## Phase 3 Schedule

The preliminary schedule for Phase 3 covers a period of approximately 12 months, starting when Phase 2 is completed. The completion of Phase 2 will be marked by the submission of the final MONARC Report on Phase 1 and 2, in March 2000.

We foresee that Phase 3 will proceed in several sub-phases:

- **Phase 3A:** Decision on which prototypes to exploit and/or build. Develop general plan for co-operative work with the LHC experiments and CERN/IT. This will require a
  - **Joint MONARC/Experiments/Regional Centres Working Meeting**
- **Phase 3B:** Specification of resources and prototype configurations
  - Setup of simulation and prototype environment
- **Phase 3C:** Operation of prototypes; operation of MONARC Simulation System; analysis of results
- **Phase 3D:** Feedback between prototypes and studies with MONARC Simulation; strategy optimisation.

The MONARC Phase 1 and 2 Report will contain a proposal for a somewhat more detailed set of milestones and schedule.

[7] This is the first release of mainstream OO software by an LHC experiment involving persistent objects, and qualified for initial use in support of trigger and physics performance studies.

[8] A typical data-set for one of these studies, as discussed with CERN/IT, would be $10^6$ events, requiring $10^{10}$ SI95-sec for production processing (simulation and reconstruction) and 1-to-several Terabytes of diskspace.

## Equipment Needs and Network Requirements for Phase 3

The equipment needs for Phase 3 involve access to existing or planned CERN/IT facilities, with some possible moderate upgrades depending on the scale of the prototype simulation/reconstruction/analysis studies to be carried out by the LHC experiments. An disk and memory upgrade to the existing Sun E450 server (MONARC01) purchased by CERN for MONARC will also be needed.

While the equipment requirements will be better specified in Phase 3A, we include a list of preliminary requirements for discussion with CERN/IT, and for planning purposes:

- Access to a substantial computing and data handling system managed by CERN/IT, consisting of a Linux CPU farm, and a Sun data server, linked over Gigabit Ethernet to internal and external networks
- Access to a Multi-Terabyte robotic tape store
- Non-blocking access to wide area network links to the main (potential) Regional Centres[9]
- Temporary use of a large volume of tape media (e.g. for ALICE).

There is a specific need to upgrade the Sun MONARC01 server, to make it a sufficiently capable "client" that will be used together with the larger system indicated above:

- Memory upgrade to at least 1 Gigabyte
- Attachment of RAID disk array of at least 1 Terabyte.

During MONARC Phase 3, we expect to take advantage of the substantially higher bandwidth network connections (in the range of 30 to 155 Mbps) that will become available this year between CERN and Europe, Japan and the US. We will work with CERN/IT to better understand the technical requirements and means to best use these networks to further study and prototype the LHC distributed Computing Models, as well as the requirements for reliable and secure high throughput connections to key points on the CERN site.

## Relationship to Other Projects and Groups

During MONARC Phase 3 we intend to continue our close collaboration with the LHC experiments, and also to work in closely with the CERN/IT groups involved in the development and use of large databases, as well as data handling and processing services. Our role with respect to the LHC experiments will be to seek effective strategies and other common elements that may be used in the experiments' Computing Models. While MONARC will have its own unique role, using distributed system simulations to optimise present as well as future large scale data analysis activities for the LHC experiments, we will also keep close contacts with present (PPDG) and future Grid Computing projects in the US (GriPhyN) and in the European Community.

[9] Examples include Italy, France, Japan and the US. In the latter two cases, at least 10 Mbps of bandwidth is expected to be available for dedicated mission-oriented and distributed system development purposes, starting in the Spring of 2000.

# Appendix C : References


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

2) I. Foster and C. Kesselman, *The GRID: Blueprint for a New Computing Infrastructure,*
Morgan Kaufmann Publishers, San Francisco, 1998.

3) The MONARC Progress Report, June 1999
http://www.cern.ch/MONARC/docs/progress_report/Welcome.html

4) H. Newman, *Distributed Computing and Regional Centres Session*
LCB Marseilles Workshop (1999) http://lcb99.in2p3.fr/HNewman/Slide1.html

5) I. Legrand, *MONARC Distributed System Simulation*
LCB Marseilles Workshop (1999) http://lcb99.in2p3.fr/ILegrand/Slide1.html

6) MONARC Technical Notes: http://www.cern.ch/MONARC/docs/monarc_docs.html

7) H. Newman, *Worldwide Distributed Analysis for the Next Generations of HENP Experiments*,
CHEP2000, Padua, Italy (2000), (http://chep2000.pd.infn.it/, paper number 385).

8) I. Legrand *Multi-threaded, discrete event simulation of distributed computing system*,
CHEP2000, Padua, Italy (2000), (http://chep2000.pd.infn.it/, paper number 148).

9) C. Vistoli et al., *Distributed applications monitoring at system and network level,*
CHEP2000, Padua, Italy (2000), (http://chep2000.pd.infn.it/, paper number 127).

10) H. Sato et al., *Evaluation of Objectivity/AMS on the Wide Area Network,*
CHEP2000, Padua, Italy (2000), (http://chep2000.pd.infn.it/, paper number 235).

11) Y. Morita et al., *Validation of the MONARC Simulation Tools,*
CHEP2000, Padua, Italy (2000), (http://chep2000.pd.infn.it/, paper number 113).

12) I. Gaines et al., *Modeling LHC Regional Computing Centers with the MONARC Simulation Tools*,
CHEP2000, Padua, Italy (2000), (http://chep2000.pd.infn.it/, paper number 169).

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

14) MONARC simulation program
http://www.cern.ch/MONARC/sim_tool/

15) Objectivity: http://www.objy.com/
CERN Objectivity Page: http://wwwinfo.cern.ch/asd/lhc++/Objectivity/index.html

16) MONARC simulation repository
http://www.cern.ch/MONARC/sim_tool/Publish/publish/

17) B. R. Haverkort, *Performance of Computer Communication Systems*,
John Wiley & Sons Ltd.

18) Y. Morita et al., *MONARC testbed and a preliminary measurement on Objectivity AMS server*,
MONARC-99/7, http://www.cern.ch/MONARC/docs/monarc_docs/1999-07.ps.

19) A. Brunengo et al., *LAN and WAN tests with Objectivity 5.1*
MONARC-99/6, http://www.cern.ch/MONARC/docs/monarc_docs/1999-06.pdf

20) K. Holtman, *CPU requirements for 100 MB/s writing with Objectivity*
MONARC-98/2, http://www.cern.ch/MONARC/docs/monarc_docs/1998-02.html

21) A. Dorokhov, *Simulation simple models and comparison with queueing theory*
MONARC-99/8, http://www.cern.ch/MONARC/docs/monarc_docs/1999-08.pdf

22) V.O'Dell et al., *Report on Computing Architectures of Existing Experiments*
MONARC-99/2, http://www.cern.ch/MONARC/docs/monarc_docs/1999-02.html

23) A report of survey of computing architectures of near future experiments will be published in Spring 2000 in the MONARC web page [6].

24) Regional Centres for LHC Computing - Report of the MONARC Architecture Group
http://www.fnal.gov/projects/monarc/task2/rcarchitecture_sty_1.doc

25) L. Robertson, *Rough Sizing Estimates for a Computing Facility for a Large LHC Experiment*
http://nicewww.cern.ch/~les/monarc/capacity_summary.html

26) R. Mount, *Data Analysis for SLAC Physics,*
CHEP2000, Padua, Italy 2000,.(http://chep2000.pd.infn.it/, paper number 391).

27) MONARC Regional Centre Representatives Meeting, 13th April 1999
http://www.cern.ch/MONARC/plenary/1999-04-13/Welcome.html
MONARC Regional Centre Representatives Meeting, 26th August 1999
http://www.cern.ch/MONARC/plenary/1999-08-26/Welcome.html
MONARC plenary meeting 10th December 1999
http://www.cern.ch/MONARC/plenary/1999-12-10/Welcome.html

28) "First Analysis Process" to be simulated
http://www.bo.infn.it/monarc/ADWG/Meetings/15-01-99-Docu/Monarc-AD-WG-0199.html

29) High Energy Physics Data Grid Initiative
http://nicewww.cern.ch/~les/grid/welcome.html

30) Results of the First ALICE Mock Data Challenge
http://root.cern.ch/root/alimdc/alimd_0.htm

31) ALICE Internal Note 99-46

32) The Second ALICE Data Challenge
http://root.cern.ch/root/alimd100/md100_0.htm

33) CMS Workshop on High-Level Trigger (HLT), 4 Nov 1999
http://cmsdoc.cern.ch/cms/TRIDAS/distribution/Meetings/TriDAS.workshops/99.11.04/Agenda.html

34) Two Day Activities Model (Model 1 in the repository [16])

35) Daily Activities Model (Model 3 in the repository [16])