MONARC

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. Tufts University (K.Sliwa) on the ATLAS side and Caltech (H. Newman) on the CMS side led the joint effort of creating this common CERN project.

The scale, complexity and worldwide geographical spread of the LHC computing anddata 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

- proximity of the data to centralised compute and data handling resources,
- proximity to the end-users for frequently accessed data,
- efficient use of limited network bandwidth,
- appropriate exploitation of regional and local computing and data handling resources,
- 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 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:

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 MONARC Project has accomplished 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 has been developed and validated against the measurements performed with test-beds. Also, the simulation results were shown to exactly reproduce analytical predictions based on queuing theory.

As scheduled in the PEP, the **MONARC Simulation WG** has developed a flexible and extensible set of common modelling and simulation tools. These tools are based on Java2, 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. Iosif Legrand was the primary developer of the simulation tool. Alex Nazarenko of Tufts University has joined him in the Summer of 1999, and helped develop the improved gui and the built-in set of statistical data analysis tools. The physics activities steps which can be simulated at present are:

- **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.

- **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.

- **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.

- **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.

Tufts University Group (Nazarenko and Sliwa) has developed a complete model of all physics activities foreseen by the LHC experiments, and then proceeded to simulate the two possible computer system architectures – fully centralized (one center at CERN) and a distributed system with multiple Tier-1 and possibly Tier-2 centers. With the completion of this task in December 1999, the primary MONARC goals for its first two stages have been fulfilled. In the last LCB meeting it was agreed that a detailed user’s manual would be the preferred way to summarize the results. Tufts University group is leading this project.

The **Site and Networks Architectures WG** 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** 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** 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 have 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.

STATUS OF MONARC SIMULATION TOOL

Requirements

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.).

Implementing the Data Model

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.

JAVA2–based MONARC simulation tool

The scheme, developed using Java2 tools, 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 (in which the reconstruction and most analyses are performed 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. In table below all parameters currently in use by the MONARC simulation tool are listed.

Table 1. Parameters used 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.

Validation of MONARC simulation tool

A set of measurements of the key parameters of the distributed database, such as AMS read/write speeds with a single user and also with stress tests have been measured. A close discussion between the Analysis WG and Testbed WG helped to identify the key parameters and the dependencies of the parameters needed in the simulation program. The correctness of the scaling behaviour, in both the local and wide-area network environment, which is vital in making any predictions on a large scale distributed system, has been validated successfully in Summer 1999. Later, a series of simulation results were compared against analytical calculations based on queueing theory. Perfect agreement was found.

Milestones

The MONARC project has successfully met all its major milestones, and is well on the way to meeting its primary goals, including

- Identifying first-round baseline Computing Models that could provide viable (and cost-effective) solutions to meet the data analysis needs of the LHC experiments
- Providing a powerful (CPU and time efficient) simulation tools that will enable further computing model studies
- Providing guidelines for the configuration and services of Regional Centers, and
- Providing an effective forum where representatives of actual and candidate Regional Centers may meet and develop common strategies for LHC computing.

A series of short papers on: MONARC simulation tool; the structure and operational experience with the simulation system (using the results of the Analysis Working Group); the work of the Architecture Working Group; and the testbed studies and simulation validaton in local and wide-area network environments will be presented at CHEP2000 Conference.

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). All results from various simulations runs, together with all parameter sets as well as Java classes used in those runs, are available from the WWW pages. A detailed user manual to MONARC simulation is under development, and should be available in the Spring 2000.

MONARC Phase 3

It is believed 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 letter of intent to continue with MONARC Phase 3 has been sent to H. Hoffman and M. Delfino in December 1999. 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.

Tufts University Group will continue its participation in MONARC Phase 3.

ATLAS-specific and US_ATLAS specific simulations with MONARC toolset

Tufts University group will lead both tasks. ATLAS Computing Co-ordinator, N. McCubbin, has asked K.Sliwa to lead a still to-be-formed group which will conduct a systematic study and optimization of possible architectures of ATLAS computing system. Tufts group has all the needed expertise and experience with both the MONARC simulation tools and the problem of optimizing data access to facilitate physics analyses. Analogous optimization studies will be performed for US-ATLAS needs, in which case emphasis should be put on optimizing the balance between Tier-1 and Tier-2 centers. Continuing support for a computing expert/analyst at Tufts University is requested. It is anticipated that results of the first-round studies will be available in the end of 2000.

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

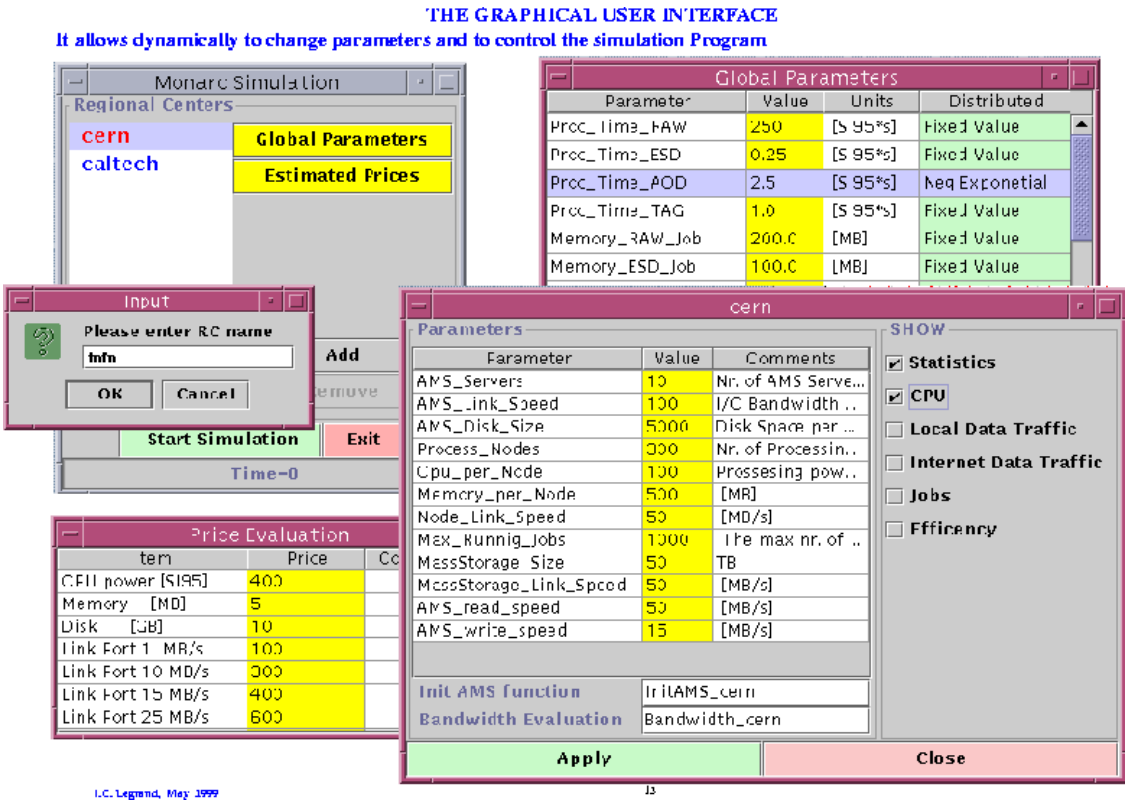

Figure 1. An example of a GUI-based setup of a MONARC simulation run.