Off-line Computing for RHIC

Contributors to Orginal Document

M. Baker[1], J. Flanagan[2], B. Gibbard[2], K. Hagel[3], T. Healy[2], S. Kumar[4], W. Love[2], E. Nicolescu[2], D. Olson[5], C. Price[2], L. Ray[6], T. Schlagel[2], S. Sorensen[7], A. Stange[2], T. Throwe[2], F. Videbaek[2]

Contributors to Document Update

J. Flanagan[2], B. Gibbard[2], R. Healy[2], T. Healy[2], R. Hogue[2], E. Nicolescu[2], C. Price[2], G. Rabinowitz[2], D. Stampf[2], T. Throwe[2], M. Strongson[2], G. Tsai[2]

[1] *Massachusetts Institute of Technology, Cambridge, MA*
[2] *Brookhaven National Laboratory, Upton, NY*
[3] *Texas A&M University, College Station, TX*
[4] *Yale University, New Haven, CT*
[5] *Lawrence Berkeley National Laboratory, Berkeley, CA*
[6] *University of Texas, Austin, TX*
[7] *University of Tennessee, Knoxville, TN*

July 20, 1997

Abstract

The plan for Off-line computing for the RHIC experiments is described. Included in this plan is a scalable computing facility located on-site at BNL capable of satisfying the bulk computing needs associated with the data taken by the experiments. This document is an updated version of a previous document, "RHIC Computing Facility"[1]. Appendices A - D, describing the off-site computing plans of the individual experiments remain totally unchanged and Tables 1 and 2 in the requirements section are unchanged except for a transformation of units from SPECint92 to SPECint95 (a division by 40). The primary foci of the updates are the transfer of $1.5M of project funding from FY 1999 to FY 2000, progress in technology evolution over the past year, and a change in the preferred choice in tape storager technology.

Contents

1 Introduction

1.1 Overview

The computing and data handling capacities required for Relativistic Heavy Ion Collider (RHIC) detectors are large when compared to previous detector systems in either High Energy or Nuclear Physics. The first serious estimate of these needs was made in 1992[2] with a somewhat more detailed follow-up a year later[3]. A more recent estimation and discussion of these needs based on a detailed understanding of the physical characteristics of the detectors and the scientific goals of the collaborations is found in the report produced in February of 1996 by the RHIC Off-line Computing Committee[4] (ROCC). The performance objectives of the computing facility in the previous proposal[1] and the plan proposed here are based largely on the needs described in the ROCC report.

Certain aspects of the RHIC computing requirements are appropriately handled by a dedicated facility located at and under the direct management of RHIC. These are the aspects associated with the handling and processing of the actual data produced by the detectors. Other aspects of the RHIC computing requirement, in particular those associated with theoretical models, event simulation and certain compute intensive or low data volume types of analyses are less critically linked to the operation of the detectors themselves and so can be done effectively at locations remote from RHIC. The possibility of satisfying such needs at existing locations such as departmental facilities at collaborating institutions or at regional or supercomputing centers at substantial dollar savings to the RHIC project is explicitly recommended in the ROCC report and is a part of the plan proposed here. In the event that adequate reduced cost computing is not available elsewhere, the computing mission of the computing facility at RHIC, as described below, will be adjusted to address those additional needs.

The dedicated RHIC Computing Facility (RCF) being developed at BNL has primary responsibility for handling and processing the data produced by the experiments and will operate in conjunction with computing facilities at remote locations and so will requiring high levels of wide area network connectivity. The RCF will be specifically responsible for the reconstruction of all collider data and for recording such raw and derived data as the experiments deem necessary. It will serve as a data mining and serving facility for this raw and derived data and will also function as a primary analysis facility. As mentioned, large scale theoretical modeling and event simulation are expected to occur mostly at existing remote sites. The storage of some data sets associated with simulation at BNL and the use of BNL facilities for simulation work prior to RHIC turn-on and during periods of non-peak demand for processing collider data are expected. Similarly the export of various levels of processed data from the RCF to remote facilities for later stages of analysis is also expected. To this end it is important that there be a high level of compatibility between components of the RCF and remote computing sites engaged in RHIC computing.

1.2 Principle Components of Plan

It is important that the RHIC off-line computing plan satisfy all of the requirements described in the ROCC report[4]. As mentioned above there are a number of types of computing which are important to RHIC scientists which are not directly addressed by the RCF or by other remote computing facilities or centers. These are the so called "desk top" requirements which includes such activities as receiving and sending e-mail, accessing information (commonly done today via the World Wide Web, WWW), preparing documents or presentations, and serving generally as an interface into the universe of networked computers. Also within the realm of desk top computing is frequently included the development of software and the performance of very high level analyses on small data sets. Collaborating groups will in general have to continue supplying basic desk top systems for their members at approximately the same level as they have in the past.

The CPU requirements for simulation and modeling, as will be discussed in the next section, are on the same scale as are those required for processing the actual event data. This proposal was generated in consultation with computing representatives of the various RHIC experiments. They have provided descriptions of how each experiments expect to satisfy those computing needs which are not addressed by the RHIC Computing Facility being established at BNL. These descriptions are contained in appendices A through D. The solutions vary substantially from experiment to experiment.

BRAHMS, the smallest of the experiments, expects to be able to satisfy all of it computing needs with a combination of the RCF and desk top workstations.

PHOBOS expects to have access to computing resources at a number of collaborating institutions, including the Massachusetts Institute of Technology, the University of Illinois at Chicago and the University of Maryland. These resources in combination with desk-top systems are expected to be adequate for those PHOBOS computing needs not addressed by the RCF.

STAR anticipates using resources associated with the National Energy Research Scientific Computing Center (NERSC) which recently moved to Lawrence Berkeley National Laboratory (LBNL). LBNL is a major center of STAR activity and since NERSC has made a recent commitment to developing and supplying resources useful to detector systems such as STAR, STAR feels it is in a good position to exploit this resource. There are initiatives including the transfer of the PDSF compute facility from the SSC site to NERSC, development of a Grand Challenge project, and internal initiatives within LBNL, which seek to enhance NERSC's capacity to support computing of the type required by HEP & NP experiments.

PHENIX expects that a facility will be established in and funded by Japan, capable of serving as a regional computing center for collaborators on PHENIX located in eastern Asia. This regional center is expected to support physics analysis for people in that region and to be capable of satisfying virtually all of the modeling and simulation needs of the experiment. There is also the possibility of obtaining substantial amounts of computing from existing facilities at collaborating institutions, most particularly the three participating national laboratories (Ames, LANL, ORNL).

1.3 Management

In order to assure compatibility across RHIC related sites, take advantage of economies of scale in acquisitions, share development efforts, and optimize the use of limited DOE funding for RHIC computing, a substantial level of coordination is required across the whole of RHIC off-line computing. The head of RHIC computing is responsible for overall coordination. He has direct line responsibility for the RHIC Computing Facility at BNL with advisory responsibility for other sites engaged in computing for RHIC. With respect to those sites for which the DOE's Nuclear Physics Division is supplying funding he will also advise the DOE regarding the appropriate distribution of RHIC computing funds.

There are a number of important management issues related to the RHIC off-line computing plan and in particular the RCF. These include the allocation of computing capacities among the various experiments, the allocation of limited funds across the various types of computing capabilities, disk, tape, CPU, etc, and the appropriateness of the technical directions and decisions taken.

In the area of technical directions the head of RHIC off-line computing will be advised by the RHIC Computing Advisory Committee which including technically knowledgeable individuals both from within the RHIC community and from other Nuclear and High Energy Physics computing facilities. This committee, convened by the head of RHIC off-line computing, will meet approximately biannually to review technical progress and plans for RHIC off-line computing in general and the RCF in substantial detail. The result of this semi-annual review will be a report containing evaluations and recommendations.

Decisions regarding resource allocation between experiments involve judgments of physics priority as well as an understanding of technical requirements and capabilities. An objective evaluation of the relative physics priorities is expected to be reach by the appropriate BNL scientific management with the advice of the Program Advisory Committee. Approximately yearly, the head of RHIC off-line computing will be given guidance regarding the scientific priorities of the various experiments and perhaps various projects within experiments. In terms of implementing this guidance, the Head of RHIC off-line computing is advised by the Experiment Computing Representative Board (ECRB), again convened by him, which is comprised of representatives from each of the RHIC experiments. This board will specifically advise on how to effectively allocate existing resources, consistent with the above mentioned scientific priorities. The ECRB will also advice the head of RHIC off-line computing regarding the projected future needs of the experiments. This advice will be use in charting the direction of evolution in facility capabilities. The ECRB will meet several times a year and will be consulted prior to significant action regarding major new equipment procurements, the allocation of exiting resources, or the alteration of operational policy.

2 Requirements

Estimates of the computing needs of major detectors produced years before their turn on will of necessity have large uncertainties associated with them. The first serious effort to quantify the computing needs of the RHIC experiments occurred in 1992. Additional estimates were made in 1993, 1995 and the most recent detailed estimate was made in February of 1996. Estimates have become increasingly detailed and realistic. The earliest estimates were done at a time when the detectors themselves were only conceptually designed and virtually no analysis code had been written. The assumptions about running time have also changed significantly. Early estimates were for 2000 hours of RHIC running per year while the most recent estimates are for 4000 hours of running. Early estimates did not fully consider computing associated with comparisons to theoretical models, extensive event simulations or certain types of very compute intensive high level analyses. As a result of all of the above factors the identified needs have grown in successive estimates. The February 1996 estimates were reviewed by the experiments in June of 1997 and judged to still be good estimates of their anticipated nominal year computing needs. These most recent, realistic and relatively stable estimates are now being used to define the scale of the computing facility to be established. The experiments are aware that once this scale is set, while the architecture may be scalable, from a funding plan perspective, it is rather like setting the size of the experimental halls and from this point on they will have to find a way to fit within the capacities that they have defined as acceptable.

2.1 Nominal RHIC running year

The numbers in Table 1 and Table 2, which describe RHIC computing requirements, are basically taken from the report of the RHIC Off-line Computing Committee[4]. The PHENIX numbers and correspondingly the totals have been corrected for a misunderstanding regarding the definitions of categories at the time the table was originally generated but this does not affect any conclusions. The numbers in these tables are based on a full year of running for the various detector systems in their design configuration. The original units for specifying CPU capacity were SPECint92's, a measure experimentally observed to correlate very well, better than other measures such as MFLOPS, with the performance of computers on PHENIX simulation and reconstruction code. The SPECint92 is no longer commonly quoted so all units have been converted to the commonly quoted SPECint95. This conversion consists, in general, of dividing the SPECint92 number by 40 to produce the SPECint95 number. For the sake of those more familiar with MFLOPS, there are approximately 13 MFLOPS per SPECint95 for some typical computers.

An important observation is that the total annual storage of 1.5 PB is a very large number. A survey of current storage media costs, excluding low-density 8 mm tape, which is unacceptable for a variety of reasons, shows that storage costs are generally in the range $2 and $5 per GByte (see Figure 8). The annual cost of media for RHIC data storage, depending on which media are used, will thus be in the range $3-8M. Investigation of the time dependence of the costs of these media shows no obvious trend. This $3-8M/yr component of RHIC operating cost is a serious enough matter to warrant careful consideration of this requirement and what can be done to reduce it. It is clear that these media costs must significantly influence the choice of

Table 1: Estimated storage needs for different stages of analysis. The numbers are in Terabytes per year.

	Brahms	Phenix	Phobos	Star	Shared	Total
Raw Data	40	230	60	230		560
Calibrated Data	40	120	300	3		463
Models					50	50
Simulated Data	1	150	2	1		154
Data Summary Tape	10	175	60	23		268
μData Summary Tape	1	25	13	26		65
Databases					10	10
Total	92	700	435	283	60	1570

Table 2: Total estimated CPU needs of the four RHIC experiments in units of SPECint95. The RCF is sized to be able to satisfy the CPU requirements specified in the first two lines of this table.

	Brahms	Phenix	Phobos	Star	Shared	Total
Event Reconstruction	450	4,375	3,000	2,100		9,925
Physics (Data)	125	2,000	625	4,500		7,250
Models					4,125	4,125
Simulation + Reconstruction	250	1,875	750	2,175		5,050
Physics (Simulation)	100	250	150	4,500		5,000
Total	925	8,500	4,525	13,275	4,125	31,350

recording media. There are some new technologies, such as optical tape, which promise substantial reductions in storage costs and these will be followed closely but none is currently sufficiently mature to serve as a basis for RCF planning.

In order to translate the above described storage requirements, which are expressed in terms of data set types and sizes, into quantities of different types of storage technology, a brief description of how the various data set types are used is necessary.

Raw data is read into the reconstruction system, when possible directly from the data acquisition system as that would significantly reduce tape handling, but perhaps from pre-recorded tape, after which it is expected to spend most of its time stored on tape on a shelf. The vast amount of of the raw data will never be read into the computing system again. Exceptions are early data, reconstructed before the reconstruction programs are fully perfected, which are likely to be read in multiple times for re-processing with improved versions of the reconstruction programs and individual events which are found to be of special interest and require detailed re-study of the basic raw data to fully evaluate. The frequency with which these exceptions can occur is limited by the available CPU cycles and data access resources. For rarely accessed data, such as the raw data, which require that a tape be manually moved from a shelf to a tape reading system, the latency (time delay required for access) will typically be between several hours and one day.

DST data, the output of the reconstruction process, is expected to be scanned periodically to select out pieces of interest from events of interest which are then used to produce μDST's. Since DST data is likely to be reread on a time scale of a few days or weeks, it is desirable that it be stored on tapes located in a robotic system so that when the data is requested it can become available within a few minutes rather than a few hours.

μDST's are expected to be accessed very frequently by individual physicists as part of their final or near final analysis. Ideally μDST's will reside on disk so that access with a frequency of minutes or hours can be accomplished with a latency of less than a second.

Thus in simple terms one would expect that the various types of data sets could be directly mapped onto the amount of shelf storage, robotic storage and disk storage required. The situation is somewhat more complex however. Since the hierarchies of storage are progressively more expensive as the latency is reduced, strategies are used to minimize the overall storage cost. One strategy is to maintain a history of the usage of a particular piece of data and to select a type of storage for it according to that usage. Another is to partition data sets so that one can keep readily available exactly that data which one needs to access frequently on low latency media without having to store on such expensive media any associated data that is not wanted. The unwanted data may consist of uninteresting events found in the same run or less interesting pieces of a particular event. Depending on patterns of usage and the care with which one employs these strategies, very significant savings in the use of expensive storage can be obtained. These strategies commonly use by Hierarchical Storage Management (HSM) systems and in relational and object oriented databases. While in principle DST's and μDST's are data sets which have been optimized in terms of being compact, experience has shown that, as patterns of usage develop, the strategies described above can still produce significant additional optimizations of access performance relative to storage expenditures. This implies that the data set volumes indicated in the table should be regarded as upper limits on the amount of high cost storage capacity required.

While not contained in the tables, a number of additional requirements are expressed in the ROCC report associated with the bandwidth for moving data to and from various systems. These include a requirement that there be access to DST's and μDST's at bandwidths of order 1000 MBytes/sec. While access to data on disk at 1000 MBytes/sec is in principle practical, the number and cost of tape drives required to produce a tape I/O bandwidth of 1000 MBytes/sec appears excessive and so the design goal taken for this parameter has been reduced to 200 MBytes/sec of I/O to or from tape, up to 75 MBytes/sec of which may be required to handle the recording of raw data when the detectors run. There is further a specification of wide area network access to BNL at OC48 (300 MBytes/sec). This access is controlled by the ESnet backbone bandwidth and must be shared with a variety of users including CEBAF, BaBar, the Fermilab Collider experiments, the LHC experiments, etc. The expectation is that this backbone bandwidth is unlikely to be greater than OC12 (64 MBytes/sec) in 1999. Any decision to dramatically upgrade this capacity would involve the entire ESnet community and would involve major additional expenditures. OC12 is thus the expected level of WAN connectivity in 1999.

2.2 Capacity requirements as a function year

The capacities requirements were estimated for nominal year running. It has been recognized for some time that at turn on in 1999 the actual computing requirements would be substantially lower than those of a nominal year since the first running would occur late in that year and both the accelerator and detectors would be in a commissioning mode. For this reason the strategy was to define an initial year goal for the RCF, which would be achieved in 1999 for RHIC turn-on, at levels approximately 50% below nominal year capacity. This would be achieved with a schedule of initial capital purchases between 1996 to 1999. Beginning in FY 2000 "out-year" replacement/upgrade purchases, would be used to bring the facility to full nominal year capacity over the next two years. Recently the estimates for capacity needs in FY 1999 have been revised further downward. Even though the nominal year needs, in FY 2001, have remained at original levels this does allow one to defer money from FY 1999 to FY 2000. The deferral of $1.5M of RCF initial capital purchases from 1999 to 2000 meets these boundary conditions while relieving funding pressure on detector system construction in 1999. By retaining the original schedule of replacement/upgrade purchases beginning in FY 2000 and combining them with the funds deferred from 1999, one is able to catch up over the next two years, satisfy computing requirements in 2000 and meeting the nominal year capacity requirements in 2001.

While previously there had been two relatively distinct capacity milestones in the development of the RCF, initial year capacities to be achieved in FY 1999 with initial funding and nominal year capacities to be achieved in FY 2001 using replacement/upgrade funding, the situation is now less clearly delineated. In Table 3 the capacity goals for various components of RHIC off-line computing are shown as a function of year with the nominal year requirements achieved in FY 2001. These profiles are extrapolated from input received from the experiments in June 1997. This table displays the target profiles for computing capacities toward which RHIC off-line computing is working. The items above the line are computing capacity specifically to be satisfied by the RCF directly funded by DOE NP division funding while those items below the line are capacities which have been assumed to be satisfied from other sources.

Table 3: Required Capacity as a Function of Fiscal Year.

	1997	1998	1999	2000	2001
RHIC → RCF I/O - MByte/sec	20	50	2x50	2x50	2x50
Reconstruction CPU - SPECint95	0	0	750	5,500	10,000
Data Mining CPU - SPECint95	0	0	625	1,875	3,750
Analysis CPU - SPECint95	0	0	625	1,875	3,750
Disk Storage - TByte	1	1	12	24	40
Disk I/O - MByte/sec	20	80	400	800	1200
Robotic Storage - TByte	2	20	60	130	250
Robotic I/O - MByte/sec	10	40	120	200	200
RCF WAN Bandwidth - MByte/sec	5	16	16	64	64
Modeling CPU - SPECint95	75	250	750	1,750	4,125
Simulation CPU - SPECint95	500	2,125	4,250	6,875	6,875
Special Analysis CPU - SPECint95	175	500	1,125	2,250	5,000
Sim/Model/Spec Analy Disk - TByte	2	3	6	11	25
Sim/Model/Spec Analy Robotics - TByte	3	25	40	120	220

3 Models

In the following section RHIC off-line computing is described in terms of models which deal with the problem at various levels of abstraction. First, there is a high level conceptual model for how the off-line analysis of RHIC data is expected to be performed. Second, there is a functional/geographical model enumerating the actual functions which must be performed and indicating where various functions are expected to be performed. Finally, there is a physical model which describes physical components of the RHIC computing plan and which functions they will perform.

3.1 Conceptual Model

The model which has evolved out of recent experience at operating collider detectors is one in which data access is the most critical computing concern. Advancing technology has resulted in dramatic decreases in the cost of compute cycles. The inherent appropriateness of employing coarse grained parallelism based upon the *Event* character of the data has made possible very effective application of these compute cycles to the computing problem. Farms of inexpensive processors, each working on a single event, are not only highly efficient but, at least conceptually, easy to manage. Data access, on the other hand, has not progressed so rapidly. Even though there have been dramatic decreases in the cost of hardware associated with data access, networks, tape drives, disk drives and robotic systems, they have not been as dramatic as those associated with CPU. In addition the strategies for using hardware in parallel to solve data access problems have been slower in evolving and have proven to be generally more complex conceptually.

At a most abstract level, the model which has evolved is one in which all data resides in a single highly structured data store for which there are a set of methods by which the data can be accessed or otherwise manipulated. The structure of the store can be thought of as a set of indices which allow one to find objects of interest and objects related in some useful way to other objects within the store. One of the primary operations that one needs to perform is the insertion of raw data from the detector into the store along with some appropriate indices. Analysis of the data then consists of accessing objects in the store and from them creating new objects which are also inserted into the store and appropriately indexed. Reconstruction is a primary example of this activity where detector type objects such as hits or cluster are accessed and physics type objects, such as particles or vertices, are calculated and added to the store again with appropriate indices. Further computational passes through the data in the store may result in the addition of more objects and indices. In some cases indices may be added without new data objects. An example of the production of "*indices only*" is a filtering pass which identifies events or particles within events which are of interest. Since the resultant store contains all information, by exploiting different sets of indices, users with different needs are all able to use this same store.

The underlying attraction of such a model is that, if implemented effectively, it allows highly specific access to exactly the objects needed with little overhead associated with the movement of unwanted data and it allows many different users to access a single copy of objects eliminating the need for multiple customized versions of the same data set. Thus one can be more efficient in the use of both I/O bandwidth and storage media while naturally maintaining easier and thus presumably better control

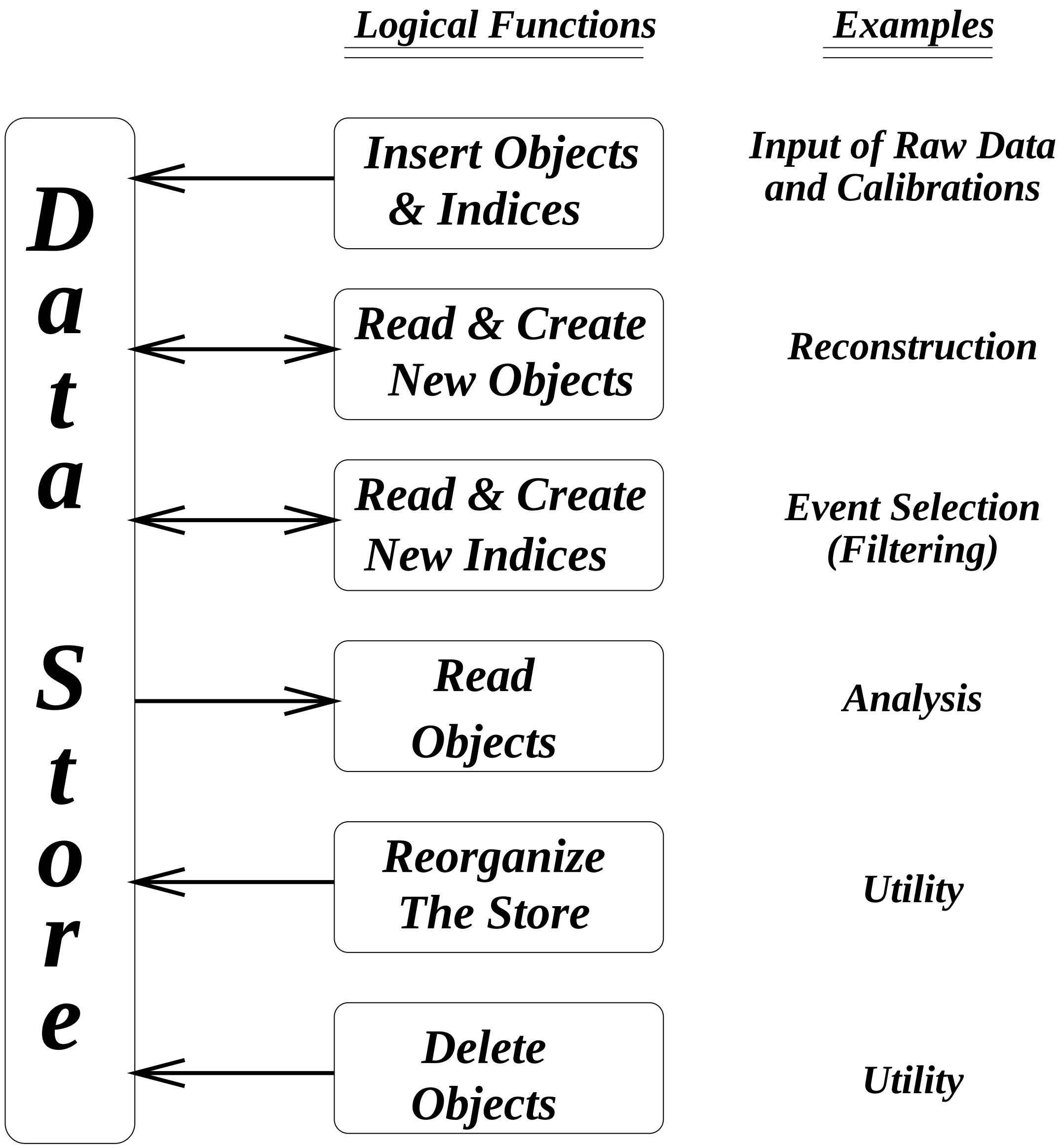


Figure 1: Conceptual model of RHIC Off-line computing.

of the history of the data being used by virtue of there being only a single copy of each data object.

A schematic of this model is shown in Fig. 1.

3.2 Functional/Geographical Model

There are many well defined computing functions required by RHIC. Those functions which are extraordinary in terms of their demand on resources have been the focus of the requirements section, but there are other functions which must be identified as to where and how they will be performed.

Data Recording/Storage: There are a variety of types of data which must be recorded and stored. In some cases the recording is of an archival nature, in the expectation that the data will rarely, if ever, be accessed again. In other cases the data is recorded and stored in the expectation that it will be frequently accessed and that the ease and speed of access is of critical importance. Large scale data sets will be recorded where produced. Thus the raw detector data and data derived from the reconstruction pass, such as DST's, will be recorded at RCF. Similarly major physics modeling, detector simulation and associated reconstruction passes on such simulations will be recorded at the regional or supercomputing center at which it is produced. While in the conceptual model, raw data is logically to be found in the unified data store, since it is rarely accessed, hopefully accessed only once for reconstruction, it will usually be physically found on shelves. The primary output of the reconstruction pass, historically called DST level data, requiring more frequent and immediate access, will usually be found physically on robotic tape. Relatively small highly distilled subsets of the data, historically called μDST's or n-tuples, are produced by selection passes performed on the DST data, a process referred to as "Data Mining". This component of the data will generally be recorded and stored local to their production but will also frequently be replicated and in some instances uniquely stored at remote sites including individual workstations, departmental facilities at collaborating institutions, and regional or supercomputer centers. This type of data, in the same logical store as the raw and DST data, will frequently be physically found on disk because of the need for very frequent and fast access as final analyses are being performed.

Event Reconstruction: Event reconstruction is the process of transforming the raw detector data into physics variables. This is generally the single most compute intensive aspect of the data processing. The primary result of the reconstruction process is usually a DST. The reconstruction of all collider produced data is expected to be performed at the RCF. Reconstruction of simulated events produced to understand detector performance issues are expected to be performed at the site that produces the simulated events. At times when the reconstruction capacity at the RCF is not saturated by reconstruction of collider data, it could be applied to such simulated data as well. However, the RCF as described in this proposal is not sized to perform the reconstruction of simulated events in parallel with the reconstruction of collider data.

Physics Modeling: In order to interpret results it is frequently necessary to compare signals observed in the collider data with the signals which would be produced in the detector by events corresponding to a particular Physics Model. The generation of such events can require large amounts of computing capacity. This type of computation is expected to be performed at departmental facilities at collaborating institutions and at regional and supercomputer centers. Again while the RCF is capable of doing such work when not saturated by collider data, it is not sized to perform this function in general.

Event Simulation: Event simulation refers to the computer simulation of the response of a detector to an event or particle. Such simulations are required to understand the response of the detector. The most common issue being addressed is the acceptance of the detector. This frequently requires the production of numbers of events comparable to the number of actual events of a particular type observed in the detector. Depending on the details of the simulation, the required computer time to perform such a simulation can range from being relatively small to being much greater than the time required to reconstruct an event. Such simulations are expected to be done at remote sites such as regional and supercomputer centers.

μDST Production: The production of a μDST is most generally accomplished by making a pass through a DST data set applying criteria to select events and objects within events. The resultant μDST then consists of the subset of objects of interest from the subset of events of interest and is thus much smaller and more easily accessed during later repetitive stages of analysis. μDST production generally requires a relatively small ratio of CPU to I/O and is thus generally limited by the bandwidth and specificity by which the DST's can be accessed. The RCF is intended to be the primary site for such μDST production and the facility is scaled to meet requirements in this area. Certain regional or supercomputing centers may choose to locally store subsets of the DST's and so may also have μDST production capability for some types of data.

It is also possible to produce additional μDST's from existing μDST's. This is expected to frequently be the case in constructing final very selective data sets. Frequently the final very selective summary of the data will be in the form of an n-tuple. The RCF is explicitly intended to perform such functions but, when the storage and compute cycle needs are relatively small, it is recognized that these functions may be done remotely, for example using departmental resources at collaborating institutions.

Analysis: Once a final highly selected data set has been identified the analysis process of studying the physics significance of the data is typically performed by repetitive passes through the data set. These passes consist of calculating additional objects of physics significance, applying various additional selection criteria, plotting distributions and numerically and visually comparing and correlating signal, background, acceptance and theoretical model distributions. Depending on the size of the data set and the scale of the computations required these needs may range from those which can be satisfied on an inexpensive workstation to those which require a large facility with parallel coordinated operations across many processors operating on large data sets distributed across many disks. The RCF is intended to serve as a facility for such analysis in the expectation that small scale analyses will often be performed on workstations perhaps at remote institutions but that there will be many large scale analyses which require a major facility. The intent is that by having such a central facility, any physicist can pursue an interesting analysis even if it requires computing resources beyond the means of her local institution.

Software Development: This activity is highly labor intensive and involves the use of CASE systems, languages, class and template libraries, debuggers, static and performance analyzers, distributed computing environment utilities, configuration management systems, and more. While this activity takes place at many remote sites, it is the RCF which is the focus and is responsible for supplying many of the required software components. Probably the most common location for software development will be a programmer's or physicist's desk top workstation. However, the RCF must also have platforms available to support such activities for those working at RHIC and for those without appropriate platforms on their own desk or at their home institution.

General Interactive Computing: The modern experimentalist performs a vast number of activities via computer, ranging from e-mail and document preparation to querying databases and displaying visualizations of events or physics distributions. Some of these activities are quite independent of the particular experiment on which he is working while others take on a particular significance as a result of his role in a particular experiment. The RCF will not be supplying the hardware (x-terminal, pc, or workstation) by which physicists interface to the computing world. Such will remain the responsibility of the various collaborating institutions. In general it is expected that the home institution will supply, in addition to the screen and keyboard, the basic level of computing required to perform routine desktop functions. However, the RCF, in so far as it will serve the computing needs of many short term visiting RHIC collaborators, will have some capability for this kind of routine interactive computing support. In addition, there are a variety of overhead services (DNS, NTP etc.) which must also be provided as part of the general computing environment of the RCF.

3.3 Physical Model

The physical model of RHIC computing is shown in Fig. 2. This figure shows, generically, the elements which compose the complete model and specifically those physical elements which will comprise the RCF being discussed in detail here. There are four distinct components to the RCF. They are the Central Reconstruction Server (CRS), the Managed Data Server (MDS), the Central Analysis Server (CAS) and the General Computing Environment (GCE) system.

The CRS is responsible for reconstructing all collider produced event data. It is highly desirable from a data handling perspective that this be done in real time as the data is produced. However provision is made for recording some or all of the data to be later read back in and reconstructed.

The MDS is responsible for storing and making available for access all forms of data. This includes raw data, the output of the CRS, *DST's*, the output of "data mining", *μDST's*, and any other data being used either locally or remotely for analysis.

The CAS is responsible for μDST type data creation, accessing the data from the MDS and writing results back to it. The CAS is also responsible for performing analysis and is specifically intended to perform very large scale analyses which are not practical on smaller scale systems. It is expected that μDST type production passes will run as background activities to analysis work for which more rapid turn around is desired.

The GCE serves as the interface to the other server systems in the RCF. It also supplies general interactive computing at RHIC, including software development and is the base for supporting the general RHIC computing environment.

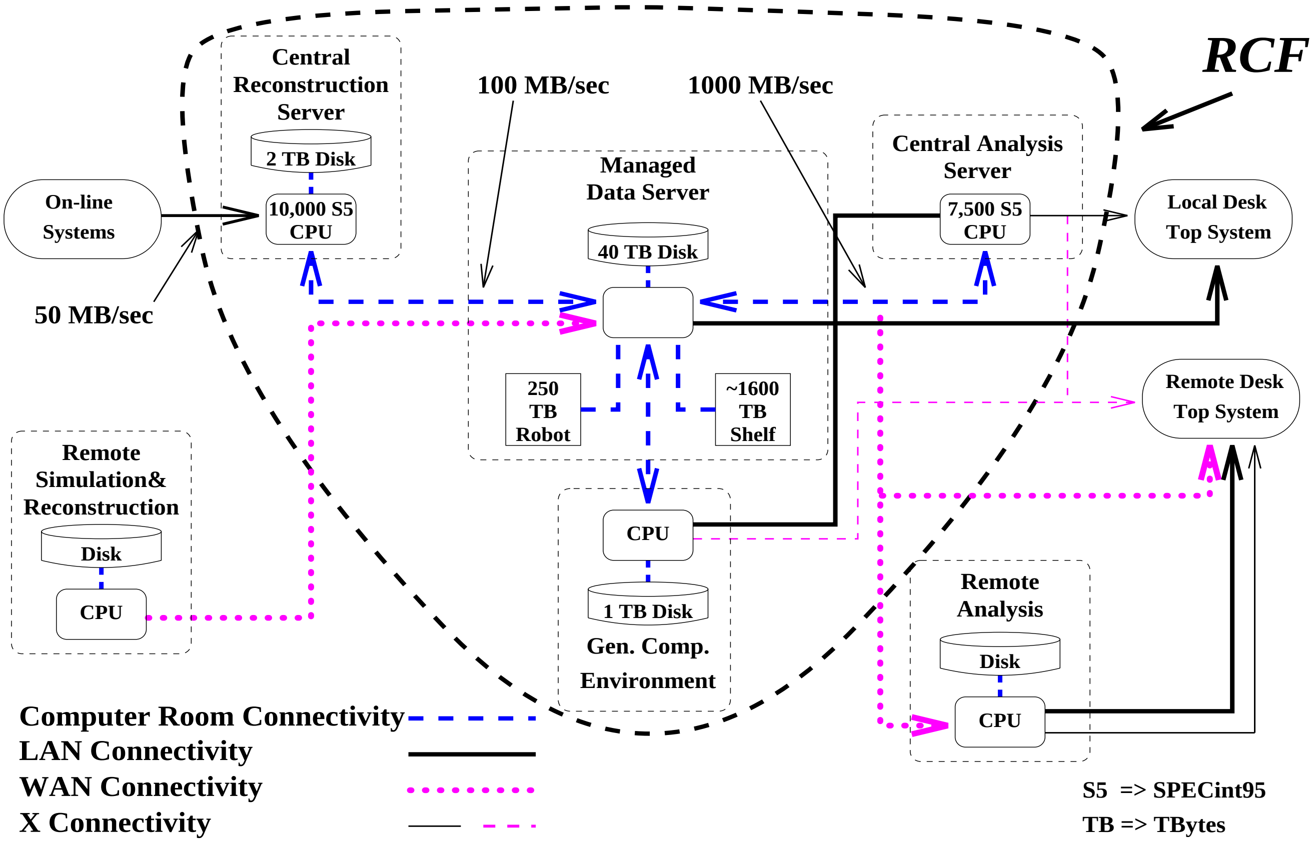


Figure 2: Physical model of RHIC Off-line computing.

4 Computing at Collaborating Institutions

4.1 Function Requirement

The primary missions of the computing equipment purchased at individual collaborating institutions are a) to supply a desk top interactive interface for each of its physicists into the RHIC Wide Area Network computing environment including effective access to RCF resources, b) to supply basic desk top computing services, Email, document preparation, etc. to each physicist, c) to facilitate the local development of software including access to network distributed software libraries and the testing of codes on small data samples, d) to support small scale local analysis projects at what has in the past been referred to as the n-tuple or small μDST data set level. At any given time, even during RHIC operation, the majority of the physicists collaborating on RHIC experiments are expected to be physically located at their home institutions. Therefore the typical physicist's primary desk top interface into the RHIC computing environment will be located at his home institution. This is consistent with resent historic precedent when physicists worked on smaller, shorter term, experiments and returned to their home institution to do the analysis. While in previous experiments such home institution located analyses may have represented a significant fraction of, or perhaps the complete, analysis of an experiment's data, in the RHIC case it will represent analysis of only final, highly pre-processed and distilled, data. In the RHIC model, data analyses which exceed the scale of analyses which have been done on home institution computing systems in the past are expected to be done on computing resources located at the RCF or in some cases other remote computing facilities.

4.2 Estimate of Equipment Needs and Costs

A profile for the equipping of an individual physicist is as follows. Each physicist needs a desk top system; x-terminal, PC, or workstation. Each physicist needs the dedicated compute power of a modern CPU with approximately 64 MB of memory and a system/swapping disk. A disk is required for data storage if (s)he is engaged in analysis, in the near term a 23 GBytes disk will probably be standard. Given the low cost and high performance of modern PC's and low end workstations, in general one such would serve as both the desk top system, source of compute cycles and point of attachment of the disk. It is expect that the effective lifetime of such system is four years and that purchases made over the next four years can be considered, on average, to fall two years from now or mid 1999. Therefore the performance of CPU's will have approximately doubled and the price of recently released 23 GB disks will have fallen by approximately a factor of two when the typical purchase is made. The maintenance costs of an individual's system as described above will be relatively low. Disks now come with lifetime warranties, Intel based computers and some workstations come with three year warranties while other workstations include at least one year warranties. Since maintenance costs are typically 15% of the capital cost but more than half of the equipment is typically still under original warranty, the maintenance costs are estimate to be 7% per year of the capital investment. Software costs for such a system are likely to be dominated by specialize products like AFS licenses and database products and is again expected to be modest, less than $1,000 per year. Staff and graduate students based for extended periods at BNL will bring their desk top system with them. The institutions which have physicists periodically visiting for short periods at BNL will supply x-terminal desk top systems to be time share by their people when at BNL. A typical institution may need one or two such. The cost of these and some other items are more appropriately figured on a per institution basis. Other examples are network costs, printer costs, and the cost of a DLT tape drives to exchange data with RCF or other collaborating institutions. Cost estimates for printers, DLT tape drives and local network equipment such as hubs are relatively easy to make. In the

Table 4: Typical Individual Computing Cost

	Cost
PC or low end workstation	$ 5,000
23 GB disk	$ 2,500
4 year maintenance	$ 2,100
Software	$ 3,000
TOTAL (4 years)	$12,600
Annual Cost	$ 3,150

area of Wide Area Network, estimates are more difficult. It is important that a site's Wide Area Networking at a minimum, support high quality x-traffic between users at a that institution and the RCF. RCF will have a connection to the ESnet backbone at the highest available speed. Sites which currently have direct access to the ESnet backbone are expected to satisfy this requirement. Sites which are participating in NSF's Internet 2 or DOE's Next Generation Internet initiatives are likely to also have connectivity meeting this requirement. Groups at sites which are not so fortunate may have to arrange for leased lines to the nearest hub on the ESnet backbone. It is currently estimated that a T1 connection costs of order $30k per year per site. It is hoped that in the 1999 time frame the majority of the institutions will not required leased lines and that smaller groups which do will be adequately serviced by a frame relay or fractional T1 connection. However, the degree of success of these new network initiatives can not be accurately estimated so allowing for the possibility of perhaps one in three institutions requiring a leased line connection may be prudent. While this is moderately expensive, it is preferable to reproducing computing equipment which already exists at the RCF at multiple individual institutions.

Table 5: Typical Institution Specific Computing Costs

	Cost
Printers	$ 2,000
DLT drive	$ 4,500
Time shared x-terminal(s)	$ 3,000
Network equipment	$ 2,000
4 year maintenance	$ 1,275
TOTAL (4 years)	$12,775
Annual Cost	$ 3,194

4.3 Cost Estimate for DOE Institutions

A serious estimate of the current baseline DOE funding level for the RHIC collaborating groups is needed to accurately determine what incremental funding is needed from the DOE to support the institutions and individuals at the levels described above. This baseline estimate would include all purchases made for computing in the past and might also reasonably include some additional equipment purchases such as scopes, meters, etc. because with the turn on of the RHIC detectors, collaborating institutions will enter into an extended period in which they are primarily involved in data analysis rather than the development of experimental equipment. Since institutional computing is defined to be at a level supported for previous experiments with more compute intensive activities performed at the RCF and/or at other computing facilities, it is expected that the only required incremental funding will be for improved network connectivity, x-terminals for use when visiting at RHIC and to support real growth in the size of the experimental groups. A poll of the collaborations indicates that there are approximately 330 individuals from 50 institutions who are DOE funded and require support in this form. This leads to an estimate of total a DOE Nuclear Physics

Division annual incremental costs of approximately \$37,500 for shared x-terminals and perhaps as much as \$500,000 for leased lines in the absence of highly successful Internet 2 and Next Generation Internet projects.

5 Facility Optimization

In the design of any large Computer Facility there are a number of competing interests which must be balanced. For the facility described in this proposal, finding this balance will be particularly crucial. The RCF requires both high performance and high capacity within a modest budget. For example, the disk cache in the CRS must accept a 50MB/s data stream while also serving data to a large CPU farm, but it must also provide a total storage capacity of several TBs in order to adequately buffer the incoming data during periods in which the CRS is saturated. The first of these needs would argue for high performance (and therefore high cost) RAID hardware with multiple host connections, while the second requirement suggests low cost single disk drives in order to increase the affordable capacity. For each element of the proposed facility these competing needs must be evaluated in order to determine the proper balance between capacity and performance.

5.1 Strategic Issues

Since this proposal is for equipment which will be purchased over a five year period, it is important to consider pricing trends in order to develop a strategy which will exploit any predicted changes. Over any long period, products with large markets experience significant improvements in both price and performance, while products in small, niche markets experience rather stagnant development and prices tend to simply track inflation. This is particularly true in the computer industry. Prices for PCs and PC based software have fallen dramatically due to the competition for the very large corporate customer base, while prices for super computers have remained relatively constant.

In order to take advantage of this market pressure, the RCF should rely heavily on items with commodity pricing. For example, PC based CPU cycles, popular tape formats (like DLT or 8mm), and widely used network components should be the most cost-effective solutions based strictly on the initial purchase price. However, it is recognized that there are other factors which must also be considered, like scalability and maintenance costs.

It is clearly most cost-effective to purchase computing equipment as late as possible. It is also valuable to delay major technology decisions in order to take advantage of new developments and products. However, if one is to establish a highly reliable system of known performance, it is necessary that the system be assembled sufficiently early that substantial testing and debugging can be conducted. These two opposing strategic considerations must be reconciled. It is proposed here that the acquisition process be approximately linear over the five years. In the first year a system of approximately 2% of the capacity of the final system will be estabished. This will permit one to obtain experience with the components of the final system individually and in small scale agregation. In the second year approximately an additional 8% of the capacity will be acquired bring the total to 10%. This will verify the scaling of the system by a factor of five and allow a substantial increase in the the statistics on individual components. This factor of five increase in scale and complexity will also serve to validate simulations and models of the system which will have been developed and will give the experiments an opportunity to gain experience with the system at a substantial scale. In the third year an additional 40% will be purchased, scaling by a factor of four, to bring the system to sufficient size to handle the inital physics runs in 1999 and early in 2000. in the fourth year, an additional 35% using the final $1.5M of “project” funding and the initial $2M of “replacement” capital to bring the system to 75% and in the final year using the replacement capital and taking advantage of the fact that only 2% of the system is ready for replacement (purchased 4 years earlier) an additional (net) 25% to bring the system to its full capacity as described earlier. In this way, while the bulk of the purchases actually occur in the later years, progressively more complete information and relevant experience is acquired in each of the early years. This plan takes advantage of the fact that the initial physics runs will place lower demands on the computing system that will be typical of the steady state operation of the RHIC facilty.

5.2 Market Considerations

Commodity items have price advantages because of extended vendor competition in the general market place. In addition, the RCF should be able to realize significant price advantages if a vigorous competitive bidding process can be used. Highly specialized single vendor solutions should be avoided even if the initial costs seem competitive simply because the long term cost of being locked into a single vendor or product line will ultimately overwhelm any initial advantage. Likewise, since the RCF will purchase significant resources well in advance of the final purchases, a high degree of flexibility must be maintained so that the later year purchases (which will be more than 50% of the total) can take full advantage of any price or performance gains during the previous years. These arguments based on healthy competition suggest that the RCF should rely either on commodity items where many vendors can provide the same product (like PCs) or broad markets where many companies can provide similar performance (like UNIX workstations).

5.3 Operational Considerations

In addition to the initial cost, the RCF must also install, service and maintain a large facility. The amount of manpower required for this task can be quite sensitive to the details of the hardware choices. The best example of this is the number of people required to maintain the CPU farms in the CRS and the CAS. If each of these consists of a small number of SMP machines with many CPUs each, one or two FTEs can keep up with the operating system maintenance tasks. However, if each facility contains a large number of single CPU systems (probably hundreds) all running the same OS, then the number of required FTEs is closer to ten than one. If those same single CPU boxes were running ten different versions of the UNIX operated system, then management would require fifteen or twenty FTEs if a coherent management scheme were even possible. Similar manpower issues arise for most of the components of the RCF, commodity pricing generally argues for a large number of low performance devices, while management and maintenance costs scale roughly with the number of devices.

5.4 Performance Considerations

Commodity items may also fall short in areas where performance is critical. Even though 100 Exabyte 8mm tape drives may provide the same aggregate read/write rate as four high performance SONY ID-1 drives, it may be quite difficult to effectively use the larger number of devices. For example, to stream the raw data from the detectors directly onto tape, the difficulty of striping a 20MB/s data stream onto 40 8mm drives would likely outweigh the benefits of their lower cost. This sort of performance difference is also relevant to the question of single CPUs versus SMPs, RAID devices versus single disks and ATM networks versus parallel ethernets. Each of these questions must be carefully studied before final decisions are made.

5.5 System Modeling

The proposed RCF is a large system which will require the complex interoperation of a variety of components. While it is relatively easy to design a system which is composed of components, each of which satisfies particular well defined performance requirements, there remains a significant probability that subtle aspects of their interactions may produce bottlenecks which limit performance substantially below what one would niaevely expect. Simulation and modeling of integrated systems is a way to locate, in advance, many problems of this type. Since the effectiveness of such modeling is limited by the validity of the models, the models begin developed to design the RCF will be compared at each phase of the project, beginning with the current prototyping phase, to the preformance of installed systems. In so doing iterations of the models based upon these comparisons should assure reasonable levels of validity.

5.5.1 Performance vs utilization

There are two variables of primary interest when discussing system performance: latency and bandwidth. Latency is the measure of time it takes the system to complete one operation. Bandwidth is the number of operations that are completed in some unit of time. The two are not necessarily directly related, since, for example, multiple operations can be performed in parallel.

In the CRS, bandwidth is the important consideration. The CRS needs to process as many events as possible averaged over time in order to keep up with the incoming data streams. It makes little difference how long a particular event waits in the reconstruction queue so long as all events are ultimately being reconstructed. On the other hand, in the CAS there are individual scientists posing queries and waiting for results and their nonproductive time waiting for query results is an overriding consideration. A better quality of physics analysis can be done using a system with low latency, even at the expense of total throughput, simply because it makes more efficient use of the scientists' time.

Simulations are being performed based on a variety of assumptions in order to determine the optimal configurations for either maximizing throughput (bandwidth) or minimizing latency. Preliminary results indicate that having a significant amount of "excess capacity" in the CAS will be crucial for reducing time spent waiting for results. This excess can be easily achieved without compromising total performance by executing long running processes, like the data-mining operations which create new

μDSTs, in the background at a low priority. These jobs then ensure that CPU cycles are not wasted without impacting the higher priority queries.

5.5.2 Configuration issues

There are a number of systems in the RCF in which the precise ratio of the various components must be optimized. The most obvious example is in the MDS where there will be a hierarchy of storage media. Simulations will be performed to determine the appropriate balance of on-line (disk), near-line (robotic tapes), and off-line (shelf tapes) media and the required bandwidths between them. Simulations will also be used to investigate the possibility that some amount of a less traditional media (like rewritable magneto-optical) might fill a specific need. Likewise, in both the RCS and the CAS modeling will be used to estimate the optimal configurations with respect to memory size, swap space, and network connectivity.

5.5.3 Data block size optimization

The packaging of the raw and reconstructed data is also very important. Smaller data block sizes allow for smarter and in principle more economical manipulation of data. For example, costs can be reduced by using smaller staging storage for incoming data on the CRS. On the other hand the use of small data blocks require that one keep track of this larger number of blocks and that the fixed overhead associated with handling an individual block is a larger fraction of the capacity required. The organization of data packages in the reconstructed and DST data can also impact system performance. For example, since an entire multi-event data block must be read from tape even if only a small portion of it is required to complete a user query, it is important to design a data storage model that either utilizes small data packages or organizes data in such a way that there is a high probability that entire packages be of use in single queries. This is essential both to limit the demands on the robotic system and to minimize the required network bandwidth. The details of such a storage model will be investigated in future simulations.

6 Facility Components

6.1 Central Reconstruction Server

6.1.1 Requirement

As discussed in the introduction, during a nominal running year the four currently approved experiments will acquire on the order of 600 TBytes of data. The amount of CPU power required for the first stage reconstruction of this data has been estimated in the ROCC report[4], and this capacity as a function of time is shown in Table 3. Due to the nature of the analysis codes being developed by the experiments, the codes are observed to scale better with a machine's SPECint[1] rating rather than with its MFLOPS rating; however, it is recognized that this scaling is approximate at best. As the reconstruction codes are improved, a more accurate measure of relative performance will be developed in the form of benchmarks from each experiment. This measure will be used in evaluating the performance of various CPU solutions; however, it is not expected that the overall scale of the required computing will be changed.

6.1.2 Rationale for and description of solution

To solve the event reconstruction computing problem posed by the magnitude of the data collected by the RHIC experiments will require the use of the most cost effective, commodity based hardware. For the same reasons that mainframe computer solutions were abandoned in the late 80's in favor of more cost effective RISC based UNIX workstations, premium priced, high performance, many CPU SMP computers and integrated farm systems must be abandoned in favor of a high performance consumer market based solution (presently with limited SMP capability). This paradigm shift is not without cost. Premium priced SMP and integrated farm machines provide significant system management advantages, but the initial cost of these machines is so high as to make it impossible to purchase the target computing power for 2001. Fortunately, network based system administration tools are becoming available which will help to offset the loss of the system management advantage of the SMP and integrated farm machines, and limited (up to 6 way) SMP systems are presently available in the commodity based market.

A comparison of the Price/Performance ratio of a number of computer systems is shown in Figure 3. The figure shows the list price (as of early 1996) of the base configuration of each of the represented computers divided by the manufacturer's published performance rating. As can be seen, in general the cost for large SMP machines is much higher than for single CPU workstations, and the consumer market Intel PC at this time has the best cost.

A number of scenarios for the configuration of the Central Reconstruction Server (CRS) will be examined over the coming year. Two of these scenarios are indicated schematically in Figure 4 and represent a network based system and a dual ported disk based system. We will not limit ourselves to these two scenarios, but will also look at a tape based system (the experiment writes tapes at the experimantal hall and brings them to the RCF to be read), a Serial Storage Architecture (SSA) based system (where

[1] 1 SPECint95 = 40 SPECint92 and is approximately 0.0132 GFLOPS

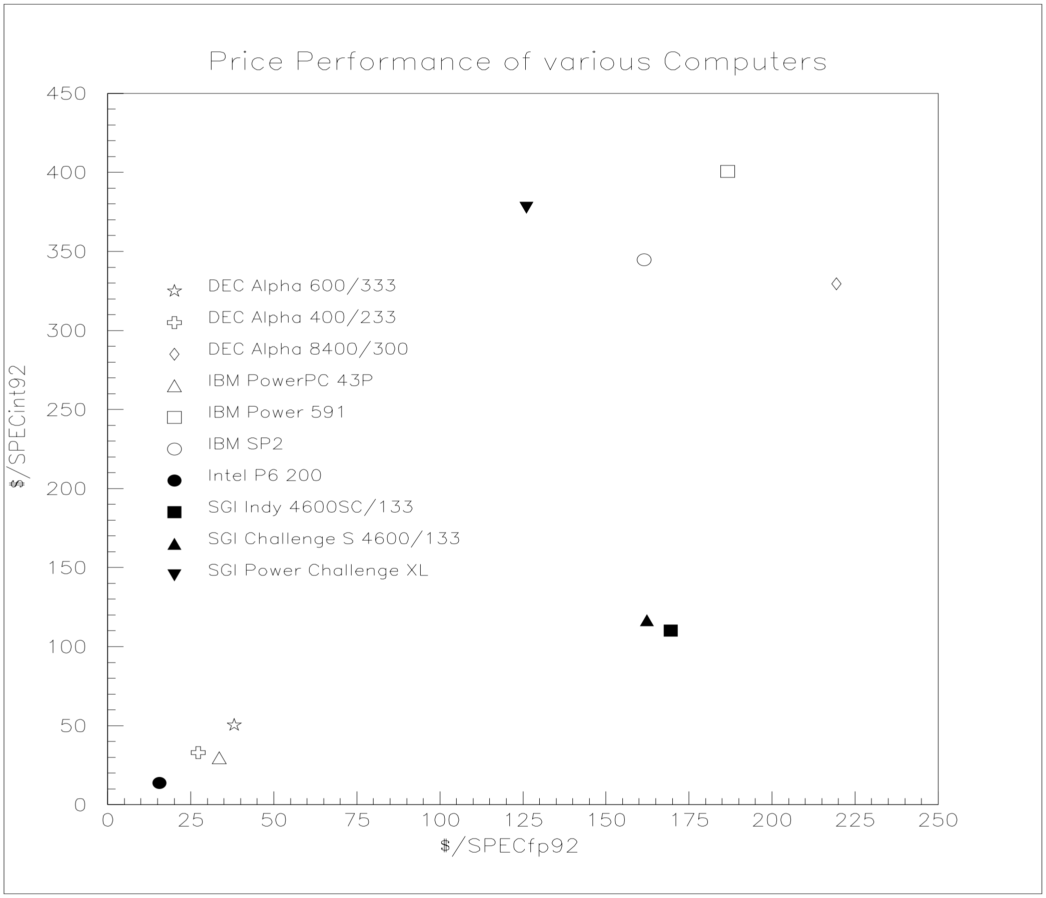


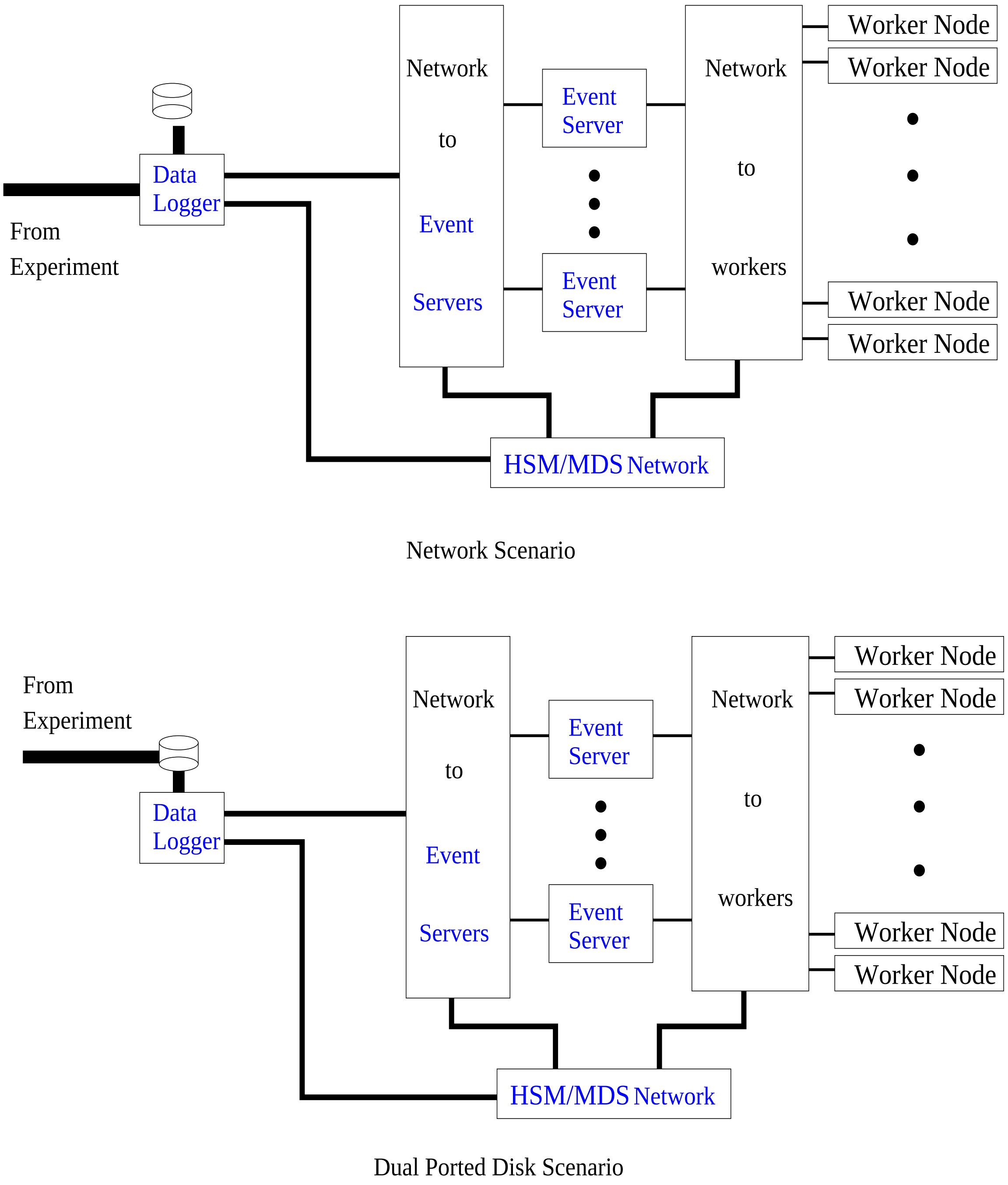
Figure 3: The price/performance ratios of various computers. The list price (as of early 1996) of the base configuration of each of the models indicated was used.

the disk and computers are on an SSA ring rather than a network) and variations of the above.

In the network based system, the data will come to one or more *data logger* machines in the RCF via a fiber optic cable. The *data logger* machine(s) will have two additional network connections, one to the CRS farm and one to the MDS. The data will be written to a disk buffer on the *data logger* machine(s) and then half of it will be sent to the CRS for immediate reconstruction and half to the MDS for reconstruction when the collider is shutdown. This scenario puts a large network load on the *data logger* machine(s).

A possible way to reduce the network load on the *data logger* machine(s) is to use a dual ported disk with one port of the disk being directly connected to a machine in the experimental hall via, for example, Fiber Channel. This would eliminate the data “network” connection from the experimental hall to the RCF and the data would be read from the disk by the *data logger* machine and sent to the CRS and MDS.

For the actual reconstruction of the data, each scenario will also have a number of *event server* machines which will have the task of coordinating and controlling the reconstruction process. These machines will stage the data to be reconstructed, that is, they will initiate the retrieval of the data from the mass storage system or ensure



Network Scenario

Dual Ported Disk Scenario

Figure 4: Schematic Diagram of Two Possible Scenarios for the Central Reconstruction Server

that data to be reconstructed which is currently on the disk does not get swapped out to tape. The *event server* machines will also split the data into "events", if necessary, and coordinate the distribution of the events to the worker nodes. Finally, the *event server* will either collect and present the reconstructed data to the HSM, or it will inform the worker node as to where in the MDS to send its reconstructed data.

The *worker nodes* (mentioned above) will each run an independent copy of the reconstruction code defined by the collaboration and will completely reconstruct an entire event.

Whichever scenario is adopted, the CRS will consist of four such units, one for each of the four experiments, with the number of elements (*data logger*, *event server* and *worker nodes*) in each of the units depending on the requirements for each of the experiments. The number of elements in each of the units is expected to grow over time as the experiments evolve. In theory, the individual elements in the CRS can be logically reconfigured depending on demand, but it is expected that such a reconfiguration would occur only on a time scales of several months.

6.1.3 Relevant commercial evolution

CPU costs, as measured in $/SPECint95, have been seen to drop by a factor of two over a period of 18 months during the last few years, and this trend is expected to continue. Generally speaking, this trend is due to a combination of gradual drops in price for specific CPU's and distinct events, such as changes in architecture, which cause larger drops in CPU cost. A recent event of this type was the introduction of the 200 MHz Pentium Pro® (P6) which is rated at 8 SPECint95. This put the cost of CPU at $880/SPECint95 in May of 1996 when six months earlier the best cost was about $2000/SPECint95, thus giving a factor of two in six months rather than 18 months. Continuing with the P6 example, 300 MHz chips are expected in approximately six months and the P7 is expected in 24 months in keeping with the exponential drop in cost. Historical data of actual purchases is shown in figures 5 and 6. Each figure shows three curves starting at different times with the earliest starting point showing the cost estimate before any purchases of Pentium machines and the next two showing actual purchase prices. The first figure is plotted assuming a 24 month doubling period and the second figure is plotted assuming an 18 month doubling period. In both cases the purchased machines are below the previous curve(s).

6.1.4 Interface to experiments and MDS

A detailed discussion of the network interface between the CRS and the Experiments and the MDS is presented in the networking subsection of this section of the proposal. A general discussion of the interface will be given here.

Each of the experiments has the option of being connected to the CRS either via a dedicated and redundant pair of fiber optic cables capable of transferring data at 20 MBytes/sec, or by way of transported tapes. For the fiber option, the data will be presented to the *data logger* machine which will write the data to a disk buffer and then send an acknowledgment of receipt of the data back to the experiment. In the dual ported disk scenario, the experiment will be in direct control of the writing of the data to disk in the RCF via the fiber. For the tape option, the experiment's Data Acquisition System (DAQ) will produce tapes which will be periodically collected and

Figure 5: Projected cost of reference machine over time assuming a 24 month doubling time.

brought to the RCF. The tapes will then be read into the *data logger* machine which will stream the data to the local disk buffer.

The traffic out of the *data logger* machines will depend on the scenario. In the case of a tape solution, there will be a total of 25 MBytes/sec out of the *data logger* machines and into the *event server* machines. In the remaining scenarios, when the accelerator is running there will be 50 MBytes/sec out of the *data logger* machines with half going to the *event server* machines and half going to the MDS. When the accelerator is shutdown, data will either flow back to the *data logger* machines from the MDS at 25 MBytes/sec and then out to the *event server* machines as when the collider is running, or it will flow directly to the *event server* machines bypassing the *data logger* machines depending on how the *event server* machines are configured.

Traffic from the *worker nodes* to the MDS, when the accelerator is running, will be between 25 and 50 MBytes/sec depending on the size of the reconstructed data and whether or not the raw data is being permanently stored. When the collider is shutdown, the traffic from the *worker nodes* to the MDS will just be the reconstructed

Figure 6: Projected cost of reference machine over time assuming an 18 month doubling time.

data, since the raw data will already have been stored in the MDS, and so up to the order of 25 MBytes/sec.

It is assumed in the above discussion that the CRS can keep up with the immediate reconstruction of 50% of the data so that all of the data is reconstructed by the end of the year with the collider operational for half of the year. In the event that reconstruction cannot take place immediately but is delayed due to, for example, the need to generate calibration constants, the role of the *data logger* mcahines is greatly reduced and, perhaps, eliminated. In addition, all data to be reconstructed will come from the MDS and the data would flow directly to the *event server* machines from the MDS.

6.1.5 Acquisition and installation

The CRS, as with all components of the RHIC Computing Facility, will now be phased in over four years with the bulk of the purchases being done in the last two years. An example of a possible ramp-up of the CRS follows. It is assumed, for costing purposes, that all of the *worker nodes* are four CPU Symmetric Multi-Processor (SMP) Pentium Pro and successor systems and that the doubling time for CPU performance is 24 months (a conservative assumption given that 18 months is the generally accepted doubling time). It is also assumed that in FY1997 and the beginning of FY 1998 the required characteristics of the *data logger* and *event server* machines will be determined either with existing equipment or with loaner machines. The actual machines will be obtained in FY 1998 and FY 1999.

1996:	10 CPU Prototype	$ 80000
1997:	4 23 GB disks	$ 8300
	10 SMP machines	$ 153000
		$ 161300
1998:	1 Data Logger	$ 50000
	2 Event Server	$ 22000
	16 23 GB disks	$ 21000
	19 SMP machines	$ 282000
		$ 375000
1999:	2 Data Logger	$ 75000
	4 Event Server	$ 36000
	40 56 GB disks	$ 66000
	59 SMP machines	$ 846000
		$ 1023000
2000:	26 SMP machines	$ 440000
		$ 440000

The above ramp-up produces a cumulative CPU capacity of 320 SPECint95 in 1997, 1,100 SPECint95 in 1998, 4,100 SPECint95 in 1999 and 6,170 in 2000. The proposed plan permits the managers of the system to gain experience with the details of managing the distributed computing environment outlined above while delaying the bulk of the acquisition until as late as possible. The intermediate step provides a means of learning how to scale the system up.

In addition to the plan proposed above, a prototype system based on 200 MHz Pentium Pro computers was assembled in 1996. This prototype was intended to explore the feasibility of using consumer market computers in the CRS, and to serve as a functioning version of what appears, at this time, to be a serious candidate for the final form of the CRS.

6.1.6 Operation and maintenance

The manpower required to operate the CRS will ramp up over time as does the hardware. The manpower directly associated with the CRS will be responsible for monitoring, configuring, reconfiguring and upgrading the machines which are part of the CRS. It is expected that the manpower needed for the CRS, assuming a ramp-up similar to that discussed in the previous section, will be 1 FTE in 1997 , 2 FTEs in 1998, and 4 FTEs in 1999.

Maintenance costs for the CRS will greatly depend on the type of solution that is ultimately selected. If a commodity based, PC solution is achieved, then the machines will likely come with a three year warranty (PC warranty periods have increased with time, so the warranty may span the replacement time of the machine (assumed to be 4 years) when RHIC turns on). If a PC solution is not achieved, then hardware maintenance is generally 20% of the list price of the machine. A PC solution then would significantly save on hardware maintenance costs, and maintenance would consist of swapping in spare machines when there is a failure and returning the broken machine under warranty for repair or replacement. The repaired or replaced machine would then be returned to the "spare" pool.

As mentioned, it is assumed that the machines will be replaced on a time scale of about 4 years. With an expected budget of $180K in 2000 and $450K in 2001, the CRS would achieve a capacity of 7,060 SPECint95 in the year 2000 and 10,000 SPECint95 in the year 2001.

6.1.7 Significant alternative branch points

At each stage of the ramp-up of the CRS outlined in the previous sections the computer market environment will be reassessed in term of the most cost effective means of obtaining compute cycles. While it may not be feasible at each of these stages to completely change the computing model of the CRS without losing the earlier investment, every effort will be made to exploit the most cost effective computing solution available and to try to incorporate available innovations.

6.2 Managed Data Server

The volume of data collected by the four experiments will present a significant challenge for the proposed facility. The storage, management and retrieval of this data will require state of the art hardware and software solutions.

6.2.1 Requirement

The requirements for data management in the RCF can be defined from three basic functions: data storage, data mining and data analysis. First, the facility must be able to store the data collected by the experiments. Since data will be collected for 4000 hours per year and the CRS will contain only enough CPU to reconstruct all incoming data by running 8000 hours per year, at least half (and possibly all) of the raw data must be recorded for later reconstruction. After completing the first stage reconstruction (either in real time or by rereading stored raw data), the calibrated data and first level data summary tapes (physics quantities) must be stored. Thus the total volume stored is likely to be 1.5 to 2 times larger than would be expected from the 50MB/s collected in the experimental halls, which already amounts to as much as 4TB per day. The proposed facility will need to be able to store this data and automatically move it to progressively more cost effective storage as the need for fast access diminishes. While it may be possible to delete much of this data when subsequent analysis passes have been completed, it is reasonable to expect that access to at least several months of this data will be required. This implies that the RCF must provide several hundred TBs of online, nearline and offline storage.

The second function of this facility is data mining. Once the data summary tapes have been stored, the next step is to produce μDSTs which contain only those events or parts of events which are relevant to a specific physics analysis task. This requires passing through a large volume of DST data and writing the selected parts into a single logical data set which will then be actively accessed by subsequent analysis. Typical DSTs will range in size from 1-100 TBs with μDSTs of between a few GBs and a few TBs. Production of new μDSTs should be possible on a time scale of days rather than weeks. This will require an access rate to the DSTs in the range of 50-200 MB/s along with sufficient nearline storage for the DSTs and sufficient online storage for many concurrently accessed μDST's. It is also clear that careful organization of the data within the DSTs will be crucial in order to minimize the fraction of a DST which must be accessed in order to produce each new μDST. If every byte of stored DST data must be read in order to produce each new μDST then data mining will become the limiting factor in the data analysis chain. The data mining aspect of this facility is now the subject of an active Grand Challenge project which is addressing both data organization within the MDS and method of access. Although there have been no final decisions, the current efforts are focusing on HPSS and Object Oriented storage models.

The third function of this facility is providing access to the μDSTs for the subsequent analyses that will be performed on the Central Analysis Server. If several hundred physicists are actively working on a variety of μDSTs, the Managed Data Server must provide online storage (or very fast access nearline storage) for tens of TBs of μDST data with a very high network access rate. Although technically this is the simplest service provided by the MDS, it is also the least predictable. The demands will be generated by physicists working interactively and will therefore not be subject to batch scheduling policies which can be used to balance demands for various resources. It is therefore crucial to employ storage management software which can automatically reorganize data storage in order to eliminate bottlenecks caused by heavy access to single servers or devices.

The proposed MDS will meet these requirements by providing data storage through a tiered structure which will include 40 TBs of online storage via magnetic disk, 250 TBs of nearline storage via magnetic tape with robotic access and, shelf storage for up to a few PBs of magnetic tapes which can be returned to nearline storage within 24 hours. The MDS will also provide 200MB/s access to nearline storage via tape drives which will stage the data to and from disk. In addition, the MDS storage resources will be managed by Hierarchical Storage Management software which will provide the automatic migration functions necessary to effectively use a multi-tiered storage system.

6.2.2 Available Technology

There are four basic choices to be made within the MDS: network fabric, disk storage architecture, tape format, and HSM software. The first of these, network fabric, is discussed in section 5.6.

Disk storage architecture: The basic choice is between low cost single disk drives and high-performance RAID devices. Although there are many advantages to RAID based storage, for the RCF, the high price of most RAID hardware makes it impractical for an installation of the size required in the MDS. Today, a typical RAID device would cost over $0.50/MB while the cheapest non-RAID disks are $0.12/MB. In order to achieve a volume of 40TB the MDS will have to rely on low cost single drives which benefit from commodity pricing. Some of the RAID advantages will be regained via software RAID implementations on the storage servers. Although these solutions do not currently provide the high-performance of RAID hardware, they can be implemented at a reasonable cost.

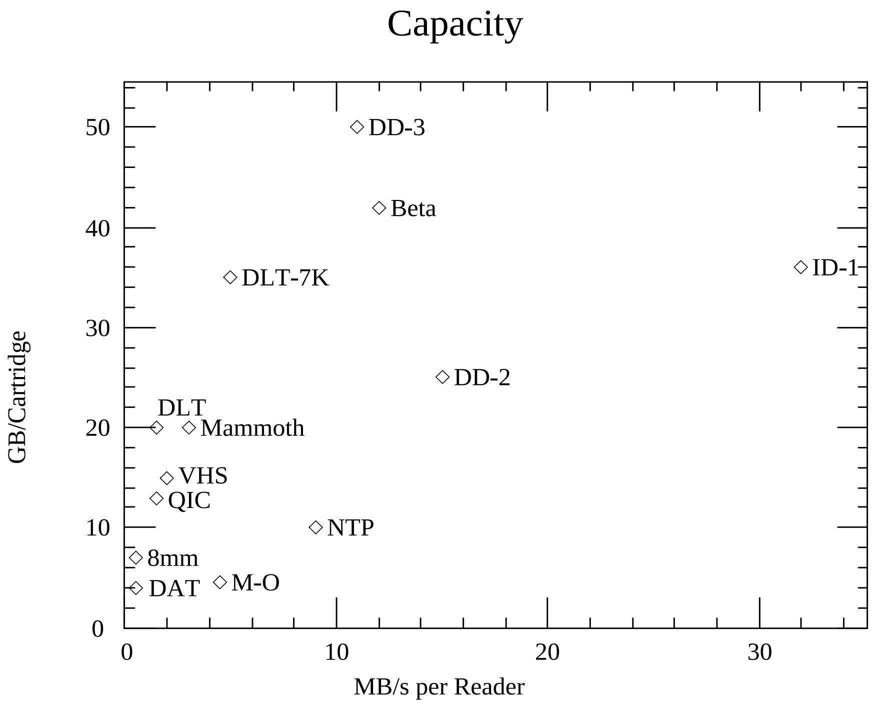


Figure 7: Cartridge Capacities and Access Rates for common tape formats.

Tape format: The current tape storage market offers at least a dozen tape format choices covering a wide range of cost and performance. The choice of format will impact the cost and performance of the MDS in three ways. There is the direct cost of the media itself, the cost and performance of the corresponding tape drives, and the required size and cost of the associated tape robot hardware. Although it is tempting to simply look for the lowest cost per GB of tape media and lowest cost per MB/s read/write rate of tape devices, this does not necessarily lead to the most cost effective solution. A good example of this effect is 8mm Exabyte tape. While 8mm cartridges

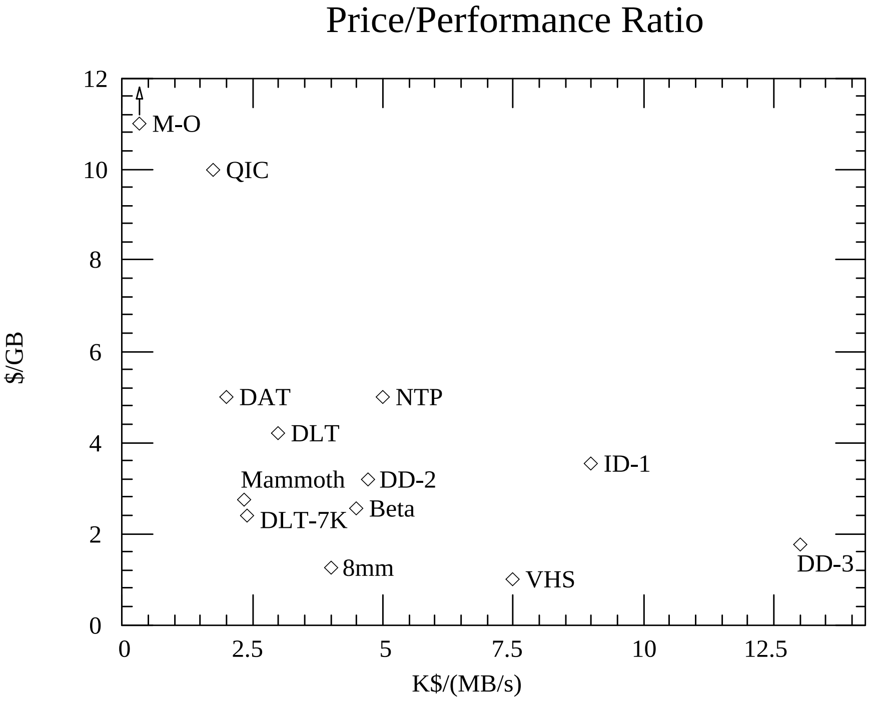


Figure 8: Price/performance ratios for common tape formats in terms of cartridge capacity and data access rate.

are only $1/GB and 8mm drives are only $4000/MB/s, the number of cartridges and readers required to reach 250TBs of robotic storage with 200MB/s access are 37500 and 400 respectively. These numbers imply a very large robotic system (perhaps even larger than can be achieved in today's market) with a cost of well over a million dollars. In order for a solution to be cost effective it must provide a balance between low cost per GB (MB/s) and high capacity per cartridge (reader). This balance is summarized in Figures 7 and 8. In Figure 7, the number of required cartridges (readers) increases toward the bottom (left) of the figure. So, in general, robot costs are very high for formats in the lower left and decrease rapidly for those nearer the upper right. In Figure 8, total costs for media will be lowest for formats near the bottom of the figure and total costs for readers will be lowest near the left edge. Most of the high performance formats which appeared to be favored in Figure 7 are disfavored in Figure 8 due to their higher cost.

The current best choice under these constraints appears to be Digital Linear Tape (labeled as DLT-7K in Figures 7 and 8). With a capacity of 35GBs/cartridge and 5MB/s/drive, the MDS requirements can be met with a robot containing 7200 cartridges and 40 drives (attainable today via products from a number of companies including Mountaingate, EMASS and StorageTek). A more natural choice is StorageTek's Redwood (DD-3) drives and media. StorageTek is the market leader in large scale storage systems, and Redwood drives are already in use by a number of projects similar in scope to the proposed MDS. The primary drawback to Redwood technology is the high cost of the individual drives. While the long term cost saving associated with the Redwood media will probably offset the initial higher cost of the drives it still creates budget problems since the drives must be purchased far in advance of the bulk of the media. Finally, there are several emerging technologies (such as optical tape) with the potential to provide very high capacities and access rates. Although these technologies are very interesting and should be watched closely, they are all too far from the product stage to be seriously considered at this time.

HSM software: The very high access rates between the MDS and CAS will require a very high performance HSM system. The most promising product is HPSS which is currently under development at the National Storage Laboratory in cooperation with a number of National Labs and software vendors. HPSS is being specifically designed for the type of high speed, parallel access that will be required by the RHIC project, so it is quite likely that HPSS will be an integral part of the MDS. However, since HPSS is still in development, availability is a serious concern. The proposed MDS facility is initially using an interim low-cost HSM solution during which time HPSS and other products can be evaluated in more detail before a final choice is made. Since essentially all HSM products adhere to the IEEE OSSI reference model, changing HSM software should not have an unacceptably large impact on RHIC specific software development projects provided that the change is made well in advance of project completion.

6.2.3 Commercial Evolution

In order to estimate the actual costs of constructing the MDS, it is important to evaluate market trends. Most computer products undergo a rapid product development cycle, and significant improvements in both performance and price/performance can be expected during the lifetime of this proposal.

Magnetic disks are probably the most volatile component on the MDS. Over the last three years the most cost effective magnetic disk has been the 5 1/4 inch form factor drives from Seagate. Three new generations (the Elite 3, Elite 9 and Elite 23) have been introduced and each of these represented a capacity improvement of a factor of between 2.5 and 3. In each of these cases the introductory price was an increase of somewhat more than 50% over the prior generation, but that price fell to nearly the old levels within about six months. This trend suggests that a new generation should be expected about every 2 years with price/performance improvements of somewhat more than a factor of 2. Since the Elite 23 has only recently become available, it is reasonable to expect that one more new generation will become available within the time frame of this proposal, so for the purpose of cost estimates, it is assumed that a 46GB drive will be available in the middle of 1999. However, given the increasing popularity of the 3 1/2 inch form factor drives (which is primarily driven by the home PC market), it is quite possible that the minimum of the price/performance curve will cease to follow the largest capacity 5 1/4 inch drives.

Tape technology is also evolving, although somewhat less predictably. For example, the current technology (DLT-7000) represents a 75% improvement in media capacity and a 300% improvement in access rate over the previous generation (DLT-4000); however the prices for the new generation are expected to remain somewhat higher (perhaps 50%) so the price/performance ratio is improving more slowly than is the case for magnetic disk. The next generation drives/media (DLT-9000) are currently in development and are expected to have capacities of 10 MB/s and 50GB/cartridge with availability in about 18 months. For other tape formats, similar evolution is possible but there is substantial variation. For example, no significant advances are expected in StorageTek's Redwood drives although price's are expected to drop slowly over the next two years. StorageTek is striving to increase tape capacity through thinner tape media; however, progress is sufficiently slow that the availability cannot be predicted.

Robotic systems are not commodity items and have been developing at a much slower pace. Also, new robot technology focuses on increased swaps per hour and other measures which are not considered to be limiting factors in the proposed MDS. Therefore, it is not expected that the costs of the robotic devices will change substantially during the duration of this proposal.

6.2.4 Interface to CRS, CAS, experiments, and WAN

The MDS will require extremely high bandwidth network connections. The data rate between the MDS and CRS will be between 50 and 100 MB/s depending on what fraction of the raw data can be processed in real time. The need for access due to data mining operations is virtually unlimited but careful management of resources should allow a bandwidth of 700-1000 MB/s between the MDS and CAS to accommodate a reasonable level of μDST production and data analysis queries. In addition, the MDS should have the highest possible connectivity to the WAN in order to allow offsite users to effectively use their existing computing resources to perform analyses on relatively small subsets of the data stored in the MDS. This WAN connectivity will be limited by BNL's connection to ESnet and the ESnet itself.

6.2.5 Acquisition and installation

The RHIC Computing Facility will be phased in over five years in order to allow sufficient time for installation and evaluation and to provide sufficient resources for software development. The proposed acquisition breakdown is by capacity 2% in 1997, 10% in 1998, 40% in 1999, 75% in 2000 and 100% in 2001. Here is an example of the equipment to be purchased in each of the years of the proposal (including existing items). Costs are estimated from current prices with extrapolations as described above. (This chart is included only as an example of how the proposed system could be achieved, not as a literal blueprint of expected purchases).

1997:	1 StorageTek 9710 robot	$ 45000
	2 Quantum DLT-7000 drives	$ 20000
	100 35 GB cartridges	$ 10000
	1 Sun E3000 Server	$ 50000
	15 9GB disk drives	$ 30000
	OSM HSM License	$ 27000
		$ 182000
1998:	2 servers (E3000 or equiv.)	$ 80000
	1 StorageTek 4400 silo	$ 100000
	4 StorageTek Redwood drives	$ 400000
	46 23GB disk drives	$ 106000
	400 50 GB cartridges	$ 32000
	HPSS HSM software	$ 200000
		$ 918000
1999:	4 servers	$ 160000
	6 Redwood drives	$ 510000
	240 46GB disk drives	$ 720000
	1100 50 GB cartridges	$ 80000
		$ 1470000
2000:	7 servers	$ 280000
	8 Redwood drives	$ 680000
	339 46GB disk drives	$ 678000
		$ 1638000
2001:	2 servers	$ 8000
	225 54GB disk drives	$ 450000
		$ 530000

While the hardware purchases are clearly skewed to the later years in order to take advantage of the expected price/performance improvements, the bulk of the learning curves associated with installing, maintaining, and using this system are skewed to the early years. Specifically, the greatest challenge to providing an effective system will be in selecting, deploying and tuning the HSM software and/or integrating the solution provided by the data storage Grand Challenge project. Therefore, the manpower required to acquire and install the MDS is more evenly distributed than the distribution of hardware purchases might suggest.

An appropriate distribution of manpower is (in FTEs) 2 in 1997, 4 in 1998, and 6 in 1999 and beyond. These would be split between experts in hardware support, software support, system installation, acquisitions, and technology tracking. These FTEs are specifically dedicated to the MDS and do not include those required to provide support to the experiments' software development efforts (see Cost and Schedule section).

6.2.6 Operation and maintenance

Beyond the scope of this proposal, there will be several sources of ongoing costs associated with the operation and maintenance of the MDS. First, since all hardware has a finite lifetime a continuing capital budget will be required to allow for replacement of aging equipment. For the MDS it is appropriate to replace all equipment as it approaches its expected lifetime since a longer replacement cycle would result in an unacceptably high failure rate that could substantially degrade overall performance. This replacement cycle, which is expected to average about four years, also allows the MDS to take advantage of continuing product development and cost/performance improvements to provide additional capacity and performance with fewer physical devices.

A second source of ongoing costs is maintenance for both the hardware and software. Since the MDS must be operational in order for either reconstruction or analysis to proceed, the hardware must be available on a 7 by 24 basis. Maintenance contracts for this type of coverage, typically are about 15% on the purchase price per year. However, this coverage should be provided only for the MDS server systems and the robotics. Disk drives typically have a multiyear warranty, and can be cost effectively maintained using cold spares and deciding whether to repair or replace on an individual basis. Software maintenance is necessary in order to obtain upgrades and some level of telephone support for problem resolution. Although there is no standard rate for software maintenance, 10% of the purchase price is typical. Since the capital purchases begin in 1997 (existing items), the maintenance costs will begin in 1998 and achieve a stable plateau in 2001.

	1998	1999	2000	2001
Disk Maintenance	$ 0	$ 0	$ 0	$ 0
CPU Maintenance	$ 7500	$ 19500	$ 43500	$ 85500
Robot Maintenance	$ 6750	$ 21750	$ 21750	$ 21750
Tape Drive Maintenance	$ 3000	$ 63000	$ 139500	$ 241500
Software Maintenance	$ 5000	$ 20000	$ 20000	$ 20000
	$ 22250	$ 125750	$ 224750	$ 368750

Obviously there are uncertainties in this estimate and it may be possible to maintain some of the hardware using BNL personnel; however, on average this should be a reasonable estimate of the ongoing maintenance costs.

Finally, a staff will be required to handle unplanned failures, perform software upgrades and routine maintenance, manage the data library (shelf storage), assist users, and investigate next generation storage solutions. A continuing staff of 6 FTEs is probably sufficient for these tasks. Additional FTEs would be required if maintenance costs are reduced by handling some hardware maintenance in house. The tradeoffs in manpower costs versus maintenance costs will be investigated before any final decision is reached.

6.2.7 Significant Decision Points

In the discussion of the MDS, there are two significant decision points which could lead to substantially different solutions. First, the final choice of HSM is made in early 1998 and is presumed to be HPSS. If HPSS fails to develop sufficiently or if an attractive alternative becomes available, a different choice could be made. Likewise, an integrated data management package developed through the Grand Challenge effort might used as a complete replacement for a standard HSM or might integrate an HSM package other than HPSS. However, unless a clearly better solution is just over the horizon, this choice should not be delayed beyond the beginning of 1998 - the experiments must have sufficient time to fully integrate the HSM software into their data management schemes.

Also, the final choice of tape format is a significant decision point. Once a choice is locked in, it will be difficult to take advantage of advances in technology in other formats. Making a final decision too early could lead to either a costly switch late in the deployment cycle or a missed opportunity to provide increased performance. The problem with delaying this choice is that the robotic hardware represents a significant investment and at least some of that investment must be made quite early. This difficulty can be reduced substantially if a flexible robotic system can be purchased. For example, EMASS offers a system which can be used for a number of tape formats and can even provide access to multiple formats at the same time. If an appropriately sized EMASS system can be obtained for a competitive cost it would allow the final choice of tape format to be made quite late while incurring only a relatively small cost in refitting the existing robotics. This choice must be tentatively made in 1997 in order to begin the initial installation but can probably be changed up until 1999 if the benefits are sufficiently large. If, on the other hand, a flexible robot is not a reasonable choice, then the final choice of tape format must be made no later than early 1998.

6.3 Central Analysis Server

6.3.1 Requirements

The goal of the Central Analysis Server (CAS) portion of the RHIC Computing Facility (RCF) is to provide a local, dedicated farm of computers with a high I/O bandwidth to the Managed Data Server (MDS) to support the repeated analysis of very large volumes of data. The computing activities expected to occur on the CAS would include production of μDST data sets from DSTs ("data mining"), as well as, final physics analysis of μDSTs. The expected capacity of the CAS as a function of time is shown in Table 3. The network bandwidth between the CAS and the MDS is also available from Table 3. In additional, at the request of the Experiments, approximately 10% of computing power located in the CAS will be made available to serve some of the General Computing Environment needs.

Some fraction of the final physics analysis is expected to be performed on remote facilities including desk top systems, existing collaborating institution facilities, and regional or supercomputer centers. These remote sites will be expected to have a moderate to high bandwidth connection to the RCF in order to transfer the requested μDST data sets (GB to TB). Such remote analyses are expect to favor those with smaller data set sizes which can reasonably be accessed via Wide Area Network and stored on modest facilities. Performing data-mining operations remotely will in general be impractical due to the extremely high bandwidth requirements they impose. The RCF is thus designed to be the primary facility for performing "data-mining" and the facility used for most large scale analyses.

6.3.2 Rationale for and description of solution

The design of the CAS is likely to be a variation of the Central Reconstruction Server (CRS) design. This will provide ease of management and administration, a crucial consideration in a large system where manpower will be at a premium. It will also allow interchangeability of components. The CAS therefore is also expected to be based on processors with low price/performance as discussed in the CRS section. For example, using Intel technology today, the CAS would require 117 4-CPU SMP 200 MHz Pentium Pro systems (rated at 8 SPECint95 per processor) to provide an aggregate of 3750 SPECint95 of compute power.

A currently attractive operating system choice would be to run Solaris for Intel on the CAS nodes; however, all options (Linux, Windows NT, etc.) would be investigated for performance, availability of required software, administrative advantages and trends in the HEP/NP community before a final decision were made.

Network connectivity to the MDS is crucial, as this is where all of the data is stored. Data mining operations will require a high I/O bandwidth to CPU performance ratio. Requirements for connectivity are approximately 1000 MB/sec during data mining operations. To meet this need each node would require switched Fast Ethernet at a minimum, for an aggregate theoretical bandwidth of 1.5 GB/sec. More likely, these nodes will be configured with Gigabit Ethernet cards which is expected to be standard in 1999.

The CAS will not require any local disk space, save for the operating system and swap space for each node (1-2 GB disk). The experimental data will be accessed directly from the hierarchical storage system.

6.3.3 Relevant commercial evolution

Over the past several months, the Pentium Pro has emerged as the processor with the best price/performance ratio for integer performance. Continued improvements in this price/performance ration as well as that for other processors is certain to occur. As the price/performance ratio improves the CAS will require significantly fewer nodes to meet the requested processing power. A dramatic decrease in the number of nodes may require the purchase of multiple network interfaces per processor system or the purchase of higher performance interfaces in order to maintain the desired I/O to CPU ratio.

It is assumed that in data mining operations, the ratio of I/O to computing is quite high. Thus improvements in connectivity between the CAS and the MDS are likely to improve the throughput of the system. Close attention will be paid to the evolution of the performance of ATM and Gigabit Ethernet interface cards in the hope of further increases in MDS/CAS connectivity.

6.3.4 Interface to MDS and WAN

The bulk of the CAS will only be accessed through a set of local interactive hosts. These hosts will be used for code development and generating queries (see the section on RCF General Computing Environment). The work flow for physics analysis will be as follows. A user on an interactive host will submit a query to the CAS. A query is a request to analyze some subset (GB to TB) of μDST data. The query will then be queued with other user queries. When a query starts, it generates requests for data from the MDS along with pointers to the code to analyze it. The MDS will provide this data, either by pointing to where it already resides on MDS disk, or by first reading it from tape to disk. As data becomes available, jobs are started on the CAS to process this data. As more data becomes available, more jobs are started. Finally, when all of the data has been made available and all of the jobs processing it have completed, the results are gathered and returned as the result of the query.

The CAS will retrieve all data from the MDS, either from some lookup into a file system, or through a sockets connection through an MDS API. Each CAS node is currently expected to contain at least one Fast Ethernet card at a minimum. Each node will be connected to a switch which can handle an aggregate of 1.5 GB/sec. This should theoretically allow each node to access the MDS at the full 100 Mbit/sec of Fast Ethernet.

It is expected that the results of CAS analyses will be sent to the requesting user as an X display or as raw histograms. Neither of these operations will require high bandwidth connections. Some CAS processing will generate data sets of reduced size which will be further processed at remote locations. This will require wide area network connectivity of modest to high quality.

6.3.5 Acquisition and installation

The CAS design is similar to that of the CRS, with the component nodes configured in essentially the same way. Therefore, the acquisition and installation of the CAS will closely follow that of the CRS (see the Acquisition and installation subsection of the CRS section of this proposal). The following table shows an example acquisition plan, assuming that all of the nodes are four CPU SMPs, and that the CPU performance doubles every 24 months.

1997:	4 SMP machines	$ 80000
1998:	14 SMP machines	$ 220000
1999:	41 SMP machines	$ 548500
	3 (GCE) SMP machines	$ 46500
2000:	16 SMP machines	$ 300000
		$ 1195000

With the above plan the cumulative CPU capacity of the CAS would be 160 SPECint95 in 1997, 776 SPECint95 in 1998, 2,710 SPECint95 in 1999 and 4,120 SPECint95 in 2000.

6.3.6 Operation and maintenance

There will be continuing hardware maintenance costs, and a need for manpower to operate, maintain and upgrade the CAS hardware and software. The required manpower is roughly 1 FTE beginning in 1997 and increasing to 2 FTEs by 1999.

Also, as with the CRS, it is assumed that the CAS machines will be replaced on a time scale of about 4 years. If the expected budget of $220K in the year 2000 and $330K in the year 2001 is realized, then the total capacity of the CAS would be 5,250 SPECint95 in 2000 and 7,480 SPECint95 in 2001.

6.4 General Computing Environment

6.4.1 Requirements

In addition to the various specialized servers, general purpose computing support is required at the RCF and a system to supply such functionality will be established. This General-purpose Computing Environment (GCE) system will satisfy all basic computing needs aside from those specialized production level activities for which the CRS, MDS and CAS have been designed.

The GCE will serve the basic computing needs of the many visiting scientists who will be located at BNL, as well as the staff which maintains and operates the RCF. It will provide scientists and programmers with a common file space in which to organize and develop code which will run on the specialized servers. To this end, the GCE should include platforms corresponding to those which make up the other servers as well as the required software development tools.

The GCE will be the location of the software and license serving functions for RHIC off-line computing. It will also serve as a central source of information by and about the experiments, making available (via the World Wide Web and wide and local-area distributed filesystems) documentation, results, and other timely information.

It will be through the GCE that most scientists interact routinely with the RCF both in terms of actual computing services and in terms of general technical assistence. An important function of the GCE staff will be to supply technical expertise to the RHIC user community on the systems, hardware and software, and on the applications and utilities, commerical and otherwise, which are being employed at the RCF.

6.4.2 Hardware

The proposed strategy is to build the GCE as a cluster of workstation class computers with sufficient resources (memory, disk, printers etc.) to support a large user community. These CPUs would be supported by three or more dedicated fileservers. This strategy allows incremental performance upgrades by simply adding additional CPU, disk or fileservers to the cluster.

CPU-intensive testing of analysis and reconstruction codes would involve using a subset of the CPU in the Central Analysis Server. Roughly 10% of the CAS CPUs will be reserved for this activity. Cost estimates for these CPUs are included in the discussion of the CAS.

The GCE will be an evolution of the current RHIC Computing Cluster. The current environment consists of 2 IBM LAN file servers with approximately 125 GB of shared disk space, an IBM WAN file server with roughly 10 GB of disk, 2 SGI interactive hosts, two multi-CPU cycle-servers (an eight CPU SGI SMP and a 16 node IBM integrated farm), and two single-CPU PentiumPro UNIX systems.

6.4.3 Distributed computing infrastructure

User files and tools will be available on all hosts in the system via the file servers. Files will be available to remote sites via WAN fileservers, using AFS (currently) and/or DFS (future) as the protocol. The present licenses will need to be augmented or upgraded as client sites are added and the transition to DFS is made.

6.4.4 Database access

The present GCE includes a 16-user license for Oracle for use by the experimental groups. This system may ultimately evolve into the event data database described in connection with the MDS.

At present it is not clear whether that database will be relational (as is the current Oracle system), object oriented or some combination. Developer licences for Objectivity have been purchased with the aim of evaluating the technology and the role it will play in the RCF.

In any case, the database licenses will need to be upgraded and expanded. This cost is included in the MDS estimates.

6.4.5 Spending Plan

Table 6: Funding profile for General Computing Environment in dollars.

	1997	1998	1999	2000	2001
Hardware	60 K	100 K	150 K	150 K	100 K
Software	0 K	100 K	175 K	125 K	100 K
Total	60 K	200 K	325 K	275 K	200 K

6.5 Data Access and Resource Management Software

In recent years the quality, the variety and the timeliness of the physics results obtained from large collider detector systems such as those at CERN and Fermilab have been constrained by limits on access by the physicists to the post reconstruction data. This is primarily a result of the vast quantities of data produced by these experiments and the fact that relatively small and generally different subsets of that data are of interest to each of the tens, perhaps as many as a hundred, analyses being conducted.

Realizing fast and efficient data access requires attention at a number of levels. At the lowest level, appropriate hardware is required to actually store and move the data. Integral to this hardware is the operating system and other system level software required to perform the basic operations. Beyond this it is important to understand in some detail the nature of the various types of access required, where and when bottlenecks arise, and to develop strategies which minimize the effects of these bottlenecks.

There are two distinct aspects to the post reconstruction data access problem. First, selecting the data of interest from the whole of the post reconstruction data set, commonly referred to as the Data Summary Tapes (DST's) but perhaps more appropriately now sometimes called the Event Summary Data (ESD). This selection process involves identifying the events which are of interest from within the total sample and identifying within each event those parts of that event (objects in the event) which are required for the analysis under question. Second, accessing the above selected data subset, commonly referred to as a μDST (now sometimes more appropriately called the Data Analysis Set (DAS)), multiple times, subjecting it to various manipulations, comparisons and forms of visualization. Ideally this repetitive access can be done in real time by a physicist sitting at a workstation or x-terminal tens of times in a single day.

While the division between these two aspects, data selection and repetitive analysis, is sometimes blurred, the two activities are usually sufficiently different that it makes sense to treat them separately. Selection passes take hours or more commonly days, while analysis passes ideally take minutes and hopefully rarely more than a couple of hours. There is an iterative aspect to these activities in that after a period of repetitive analysis on a given Data Analysis Set (DAS), a new selection pass is frequently made through the entire Event Summary Data (ESD) producing a new DAS.

Over the course of the last ten years a number of initiative have been undertaken to implement strategies which facilitate access. The general approach of these projects has been to address the access problem by optimizing the organizational structure of the stored data. These projects have either developed custom software or based their development on commercial software. Examples of such projects are the PASS[5] project, the CAP[6] project and most recently the RD45[7] Objectivity initiative at CERN. Within the last 6 months a new DOE funded Grand Challenge project has been initiated to address the issue of data access for large Nuclear and High Energy Physics experiments. PHENIX, STAR and members of the RHIC Computing Facility are all collaborators in this project, the stated highest priority aim of which is to deliver data access software which is useful to RHIC. This project is a research effort whose success can not be guaranteed and for which serious performance projections have not yet been made.

There are three elements currently represented in the Grand Challenge project. 1) The organization of the data at the event level, most specifically on tape in a robotic system, with indices which support targeting particular event categories when doing selection passes. In so doing one is able to obtain events of interest while avoiding reading events which are of no interest. It is intended most specifically to reduce the number of tapes which must be mounted to accomplish any given selection pass. This is currently being referred to as "Clustering". 2) The organization of the data within an event so that one need not access all of the data of an event if only certain parts of it are necessary. In implementation this might mean that there would be separate physical stores of data containing collections of objects of the same type from different events. This is most relevant to storage of Data Analysis Sets (DAS) which reside on disk to allow fast access during repetitive analysis and which therefore must be of limited size. This type of access is most appropriate to an object oriented data model where the event is navigated by pointers. This is currently referred to "ODMG compliant" access (in reference to the Object Data Management Group defined standard interface for access to object oriented data). 3) The organization of access requests such that those requests can share tape mounts and tape file reads whenever possible. This type of "Coherent Access" makes optimal use of facility resources and can dramatically improve typical and worst case performance without having a significant negative impact on "best case" performance.

It is in the area of Coherent Access that RCF members are most active. An important characteristic of this component of the project is that it will be effective, and perhaps even more important, if it is not practical to employ one of the other components of the project.

The Coherent Access model is imagined to work approximately as follows: A process or set of processes successively mounts and accesses data from tapes found in a "Tape Required" FIFO queue. When a tape is accessed, events on it are tested against sets of criteria defined by all currently active data requesters, some of whom may have just begun accessing data while others may be near completion. The components of interest in the events satisfying complete sets of criteria are preserved on disk and labeled as part of the resultant Analysis Data Set of each requester whose criteria were satisfied. When a new requester becomes active, its request is converted into a list of needed tapes, and files on those tapes, plus a set of criteria. This list of needed tapes is compare to the "Tape Required" queue and those tapes not found in it are added to the queue. The new requresters set of criteria is added to the other sets of active criteria and will be applied as soon as any appropriate event is encountered. Tapes are removed from the "Tape Required" queue as accessed, and when the last tape needed by a given requester is removed, its access pass through the data is complete.

Using very simple models, this coherent access mechanism has been compared to the simple mechanism of processing each requester in isolation. Using a simple representation of the PHENIX and STAR data selection problems, this mechanism was found to produce generally superior performance and in cases of heavy system loading improvements in throughput by nearly a factor of 100 were produced. Within the limits of this simple model, which assumes that data access rather than CPU is the limitation, the Coherent Access mechanism was found to be able to satisfy both the PHENIX and STAR throughput requirements at any level of system loading with hardware capacities which are expected at the RCF. More sophisticated modeling of these mechanisms is now under development and will be used to refine these results and study algorithms which might be used to manage queues in the coherent access model and further optimize throughput.

6.6 Local and Wide Area Networking

6.6.1 Overview

This section presents the evolution of the networking component of the RCF since the October 1996 version of this document.

- Considerable price reduction and/or performance gains have already been realized in the past year and can be expected to continue. Considering this favorable trend, the projected cost is approximately $576K, which represents a $68K reduction from the 1996 estimate. Cost estimates and how they were derived are found below.
- The design presented here and the resulting cost estimates are heavily influenced by a number of factors which are likely to change before the equipment is purchased. Among these are port count, available switch size, and implementation choices affecting data flow patterns within the network. This design is tied directly to the assumptions of the overall off-line computing proposal in terms of system quantities, function, and required bandwidth.
- The availability of large amounts of WAN (Wide Area Network) bandwidth to remote sites is not a certainty and should not be depended on for key elements of the overall RHIC off-line computing model.

6.6.2 Underlying Assumptions

In order to qualify what the sections below describe, it is necessary to detail some of the key assumptions.

1. The components will be logically and, in most cases, physically partitioned by experiment for performance reasons. The exceptions to this general rule involve access to shared components of the RCF. For design purposes the percentage of resource allocation was agreed by the collaborations to be 35/35/10/10 for Phenix/Star/Phobos/Brahms respectively. (The remaining 10% are to be allocated at a later date.)
2. Total bandwidth coming from the experiments' data acquisition systems is based on the ROCC report[4].
3. The numbers of systems/ports connected to the switch fabric are based on the current interpretation of the collective off-line system design as described in the previous sections of this proposal. Some of the key numbers are:
   - 4 Inputs from data acquisition
   - 22 Managed data store (MDS) servers
   - 114 Reconstruction systems
   - 75 Analysis systems
   - 10 General computing systems
   - 4 Data Logging / Caching systems
4. With the exception of the 22 MDS servers, the price of the network interface cards are included in the cost estimates for the individual systems, not as part of the price estimate shown in Section 6.6.3 below.

6.6.3 Proposed Solution

1. Design Overview

   The proposed network is intended to provide the complete interconnection of all RHIC off-line computing systems, beginning with the delivery of raw experimental data gathered at each of four experimental halls approximately 3 km from the RHIC Computing Facility (RCF). The four separate data flows are transported via single-mode fiber optic cable connections. (See Section 6.6.2 above.)

   The required fiber optic infrastructure is largely already in place. The resulting connection will provide at least 6 strands of single mode fiber from any given experimental hall to the RCF to allow for primary, backup, and future connection requirements. It now appears likely that multiplexing substantially will reduce the number of fibers required.

   The ROCC report stipulates an "independent backup" for each connection. Upon investigation the cost to provide a completely independent fiber path from the ring to the RCF proved to be prohibitively expensive - estimated at $897,373. - as this would require trenching a full 2 km between locations. The solution was to provide spare fiber capacity over the existing path and include additional backup switch interfaces in the design.

The bulk of the RHIC Off-line computing network will reside in an approximately 14,400 square foot area within the Computing and Communications Division (CCD) building. This computer room environment will house the network switches as well as the reconstruction and analysis nodes, the storage servers, and the entire disk and tape storage system.

The core of the proposed network is a switch fabric composed of a group of separate switches. In the 1997 proposal, ATM switches were the only choice able to provide the required bandwidth and port count within the budget; in 1998, these are expected to be high capacity switches providing a mix of Gigabit Ethernet (1000BaseSX/LX) and Fast Ethernet (100BaseTX/FX). This aspect of the design, as before, is contingent upon successful qualification testing in the second half of 1997 and the first quarter of 1998.

2. WAN

   Significant obstacles remain to placing heavy dependence on the availability of large amounts of WAN bandwidth. Although other possibilities exist, this aspect of the plan has assumed that all WAN resources will be provided by ESnet. It seems reasonable to assume that the BNL - ESnet link speed will be OC-3, and perhaps even OC-12 by 1999, but the OC-48 value given in the ROCC report is unlikely to be available.

   It is also important to note that while the present BNL ESnet is provided by ATM, this ATM connection terminates in a router rather than an ATM switch. Dates for introduction of OC-3 and OC-12 to BNL from ESnet are not known at this time. The cost estimate allocates a fixed block of $58,500. for a high speed router to connect between ESnet/BNL and RCF if required. This system will provide the anticipated 1000BaseSX to ATM OC-12 adaptation. A future requirement for end to end ATM WAN services would require a different solution.

   Currently, it is not entirely agreed to what extent and in what form remote collaborators will access experimental data. It seems prudent to use a very conservative estimate for available ESnet bandwidth for all calculations while working to improve the situation.

   Similarly, it is important to raise collaborators' awareness of RHIC / BNL's role, or lack thereof, in delivering any degree of end to end service quality beyond the scope of ESnet's influence on the Internet as a whole. This is not a function of BNL or ESnet; it is a function of the worldwide collection of service providers that form one or more segments of the path between ESnet and the collaborator's site. Once this dependence has been acknowledged, we can begin to develop ways to improve the situation where possible in the years before the experiments come on-line.

   Barring objections from ESnet, we will recommend that individual remote institutions contact ESnet and initiate connection negotiations directly.

3. Cost Projection

   Table 7 below shows the cost estimate breakdown for the network design. The estimate is based on using June 1997 price quotes as a base reference to infer the cost when purchased in 1998. The table lists the primary costs and a percentage by which each cost center is expected to decline in the coming year.

Table 7: Networking Cost Estimates

		Unit	1997	97→98	98/99
	Quantity	Cost	Cost	% Disc.	Cost
Ethernet Switch Chassis	7	$12,000	$ 83,700	10%	$ 75,300
Switch modules	1	n/a	279,400	10%	251,500
Management Software	1	27,500	27,500	10%	24,700
RCF Cabling and Patching	1	20,000	20,000	0%	20,000
Gigabit Ethernet Monitor	1	62,000	62,000	5%	58,900
Equipment Racks	3	1,800	5,400	0%	5,400
High Speed Router	1	65,000	65,000	10%	58,500
Modeling, Docu. Software	1	35,000	35,000	10%	31,500
Network Mgmt. Workstation	1	8,500	8,500	10%	7,700
GbE NICs for MDS (SBUS)	24	2,000	48,000	10%	43,200
Total					$576,700

6.6.4 Design Vulnerabilities

The network design outlined above is based on a set of assumptions which have proven to be subject to change. Many of these variables could significantly impact the cost of the overall solution. The most important of these include:

- The number of processors per network interface within the reconstruction and analysis farms defines the bandwidth required per port as well as the total number of switch ports at a given speed.
- The number of ports and internal capacity of the switches affects the total number of switch chassis and the number of high speed ports used to link these switches together.
- The processing power allocated to each experiment and the efficiency of their analysis code likewise determine the bandwidth required to each node.
- Variations in the ratio of raw data input from the detectors to the output of the reconstruction phase determine in part the network bandwidth requirements.
- The chance that the anticipated reduction of hardware prices will not be realized.

6.6.5 Relevant product evolution

Networking hardware stands to realize continued improvement in price / performance in the next year, perhaps more than any other aspect of the RHIC off-line computing system.

- Switch port density probably will continue to increase in this period. While this will help the RCF in terms of system installation, management costs, etc., it is not clear if there will be a significant reduction in the per port purchase cost. The cost estimate above assumes a modest reduction in keeping with recent networking product trends while factoring in the already significant impact of Gigabit ethernet technology.

- The introduction of a fast ethernet switch with an OC-12 ATM or Gigabit ethernet interface now allows the use of fast ethernet interfaces to one or more classes of machines, thereby realizing a significant reduction in cost.
- It is possible that the evolution of switches could result in the availability of a large switch capable of handling a much larger subset of the required connections for a lower cost.
- Alternate disk connection technologies (such as fiber channel) might provide a more cost effective means to meet the required performance goals.

6.6.6 Acquisition

In order for the network to be available for capacity testing and network / application tuning, it will be necessary to select the products in the first half of 1998. The products will have to be released by this date, tested by the RCF (or trusted third parties), and have documented performance in order to be ordered and installed before the system goes on-line in 1999.

An initiative will be started in the beginning of 1998 to draft an RFP for switch vendors to respond to. This effort may have to be repeated prior to the actual purchase if there are significant design or technology changes in the interim.

6.6.7 Installation, Staffing, and Maintenance

- Installation

  Installation of the networking equipment will be in three standard equipment racks in the RCF. For all 100BaseTX connections this effectively means that the switch must be within 90 meters of the systems. The systems connected by fiber optic cable (ATM or Gigabit ethernet) connected by fiber optic cable are not constrained by distance within the RCF.

  Fiber optic and enhanced category 5 cabling will be installed by BNL's telephone contractor and/or CCD.

  Costs for the switches, racks, and patch panels have been included in the estimate. No significant installation labor costs are anticipated as the bulk of the installation will be performed by CCD.

- Staffing

  A significant portion of the operation and monitoring of the RCF network is expected to continue to be provided by CCD. The projected staffing requirements (FTE) for networking, including the CCD component, are shown in Table 8 below. These values are averaged over the year and do not reflect the peak number of individuals who may be required for any given activity.

- Maintenance

  Annual hardware maintenance costs are estimated at 13% of the total cost, or approximately $66k in 1999. Software maintenance is estimated at 20%, or $12.7k.

Table 8: Networking Manpower Estimates

Activity	1997	1998	1999
Design / Project Mgmt.	0.50	0.75	0.50
WAN Testing and Reporting	0.25	0.25	0.25
Configuration and Qualification	0.10	0.50	1.00
Troubleshooting	0.10	0.20	0.25
Network Management	0.10	0.30	0.50
	1.05	2.00	2.50

6.7 Physical Plant and Infrastructure

The RHIC Computing Facility consists of many individual pieces of hardware for each of the main RCF sections. The CRS and CAS alone account for 100+ individual nodes and their related power and networking connections. To accomodate all of the equipment a large room will be required with sufficent power, network access, air conditioning, safety equipment, and secure access.

Fortunately, there is already such an area in the BNL Computing and Communications Division Data Center. A separate area within the data center, which in the past had been used to house an IBM mainframe, is being set aside for the exclusive use of RHIC Computing. The room is 1440 square feet in size, and should have sufficient room to house the robot, tape drives, disk, and computing nodes.

Entry to the data center area is currently restricted by a card reader lock on the access doors. In addition the RCF area has two separate sets of double doors which could enable additional security restrictions if necessary.

The room already has a raised floor for wiring, as well as its own air conditioning unit and fire extinguishing unit. There is also a separate power distribution unit (PDU) in the room which will be used for the electrical connections.

Although the data center currently has an uninterrubtible power supply (UPS) for the data center machines, it is not clear whether it would have the capacity to also handle all of the RCF equipment nor is it clear what fraction of the RCF equipment actually requires such service.

The cost of installing a separate UPS for RCF is currently estimated at $200,000. A final estimate can be made once the power requirements of the RCF are better known. This and other infrastructure costs are expected to be provided for by Brookhaven.

7 Cost and Personnel Summary

There is substantial history behind the plan for RHIC Off-line Computing. The massive computing requirements of the RHIC experiments in combination with serious limitations on funding have result in the decision to follow three significant non-technical strategies in addition to the technical strategies discussed in earlier sections of this document. The first such decision was to attempt to out-source significant components of the computing requirement to remote facilities at which free, in terms of DOE Nuclear Physics Division funding, or at least significantly reduced cost, computing could be obtained. This is an excellent strategy to the extent that it is successful but if the required computing is not forthcoming and must be satisfied at the RCF, it will represent a change of scope and require funding beyond that discussed in this section. The second decision has to do with the sources of personnel who will establish and operate the RCF. This will be discussed later in this section. The third decision was to develop a plan to arrive at capacities in the RCF which satisfy nominal year requirements by a combination of "project" funds associated with but distinct from the RHIC construction project and upgrade/replacement funds in the initial years of operation.

The initial bump of funding associated with the RHIC project, and thus protected from the normal G&A of 32.5% is $7.9M. This funding, taxed at a G&A rate of only 1.6%, would be used to ramp-up to the level required for initial operation during the period from FY 1996 through FY 1999. Computing equipment generally becomes obsolete on a time scale of three to four years. Where by obsolete one refers to the situation in which the annual maintenance cost for a piece of equipment bgins to approach the cost required to purchase new equipment of comparable performance. Another aspect of obsolecence is the appearence of incompatibilities between the obsolete product and new products appearing on the market. It is important that a facility with an ongoing mission replace its equipment on about a four year time scale. This implies a replacement/upgrade budget of about 25% per year of the total installed capital value to keep a facility current. In the case of the RCF this level of facility replacement and upgrade, called by some facility "fresh", will require $2M of purchases per year. It was realized that since a large fraction of the RCF equipment would be purchased in 1998 & 1999, as late as possible in the initial project bump phase, a substantial fraction of refresh funding in 2000 & 2001 could be used to bring the facility to full nominal year capacity, there being little four year old equipment in the facility in those years.

The following description of facility funding is then based on this strategy. A distinction is made between the initial G&A sheltered "project" funds and the "refresh" funds which are subject to full G&A. The facility capacity profile ramps-up to the levels required for nominal year running in 2001 based on significant contributions from these refresh funds as well as the initial project bump.

7.1 Capital Cost & Capacity Profiles

The capital costs associated with the RHIC Computing Facility are summarized here. A detailed discussion of the cost of each of the components of the RCF is included in that component's section of this proposal. The purchase profile is designed to satisfy the goals established in Table 3 in the requirements section of this document. The costs by component of the RHIC computing facility are summarized in Table 9 and

Table 9: RHIC Computing Facility direct capital costs.

	1997	1998	1999	2000α	Project	2000	2001
CRS	$ 163k	$ 375k	$ 1023k	$ 440k	$ 2001k	$ 180k	$ 450k
MDS	196k	918k	1470k	729k	3313k	909k	529k
CAS	80k	220k	595k	300k	1195k	220k	330k
GCE	60k	100k	150k	50k	360k	100k	100k
Network	30k	420k	157k	0k	607k	50k	50k
Software	0k	100k	175k	25k	300k	50k	50k
Totals	$ 529k	$ 2133k	$ 3570k	$ 1535k	$ 7776k	$ 1509k	$ 1509k

Table 10: The capacities of the RCF achieved during the indicated year.

	1997	1998	1999	2000	2001
CRS CPU - SPECint95	320	1100	4100	7100	10000
CAS CPU - SPECint95	160	780	2700	5300	7500
MDS Disk - TBytes	.3	1.3	12	28	40
MDS Disk I/O - MBytes/sec	20	125	375	975	1200
MDS Robotic Tape - TBytes	5	20	75	270	270
MDS Tape I/O - MBytes/sec	10	44	110	200	200

the accumulated capacity achieved each year is summarized in Table 10. The column labeled "Project" indicates the total funding summed over the initial ramp-up phase of the RHIC Computing Facility.

7.2 Technical Support & Operating Cost Profiles

An estimate of the personnel requirements associated with establishing and operating the RCF was conduct both bottoms up and by comparison with other facilities with similar missions. The magnitude of this requirement in combination with ever present funding limits lead to a plan involving contributions to the technical effort required to establish and operate the RCF from the BNL Computing and Communication Division and from the RHIC experiments. The estimated personnel requirements for the RCF are summarized in Table 11. A more detailed discussion of the manpower for the various components of the RCF is given in the corresponding subsections of this document. The personnel for each year is shown in three columns indicating contributions from the RHIC Experiments (Exp) and from the BNL Computing & Communications Division (CCD), as well as the manpower directly funded for the RCF. The categories in the table are not absolute in that there are people counted in the upper five categories who would fit into the lower four categories. The last four categories represent people not explicitly tied to one of the subsystems of the RCF represented by the upper five categories.

Direct labor costs, including appropriate fringes, addressed here are for 18 FTE's in the out year limit. This includes 8 FTE's which are a part of the original RHIC pre-

Table 11: Estimated personnel requirements for the RCF

	1997			1998			1999			1999
	Exp	CCD	RCF[2]	Exp	CCD	RCF	Exp	CCD	RCF	Totals
CRS	0	0	1	1	0	1	2	0	2	4
CAS	0	1	0	0	0	1	1	0	1	2
MDS	0	0	2	1	1	2	1	1	4	6
GCE/User Support	1	0	2	1	0	4	2	1	5	8
DAS	1	0	2	1	1	2	1	1	1	3
Network Support	0	1	0	0	2	0	0	2	0	2
Technical Devel.	1	1	1	1	1	1	1	1	1	3
Hardware Support	0	1	0	0	2	1	0	2	1	3
Admin. & Manag.	0	0	1	0	0	2	0	0	3	3
Totals	3	4	9	5	7	14	8	8	18	34

Table 12: Profile of annual facility operating costs.

	1997	1998	1999	2000	2001
Funded FTE's (Effective)	5.5	10.5	16	18	18
FTE S&W	$ 495k	$ 945k	$ 1440k	$ 1620k	$ 1620k
MSTC	75k	176k	439k	624k	805k
Direct Total	$ 570k	$ 1127k	$ 1879k	$ 2244k	$ 2425k

operations and operations budget. The Materials, Supplies, Transportation and Communication (MSTC) costs listed include maintenance costs on non-warranted equipment and software, which are typically 10 to 20% per year of the purchase price of equipment or software, computing facility supplies, personnel support costs (telephone, travel, etc), sub-capital equipment, and other purchased services. Costs for facility infrastructure are assumed to be provided by Brookhaven. The institutional G&A burden will only be 1.6% during the project phase of RHIC but with turn-on will become 32.5%. These annual facility operating costs are indicated in Table 12.

References

[1] M. Baker *et al.*, RHIC Computing Facility, October 8, 1996.

[2] W. Love *et al.*, Report of the RHIC Offline Computing Committee, Sept 30, 1992.

[3] J. Featherly, *et al.*, RHIC Offline Computing Study Group Interim Report, June 30, 1993.

[4] B. S. Kumar *et al.*, Offline Computing at RHIC, Feb 14, 1996.

[5] D. R. Quarrie *et al.*, Proceeding of Computing in High Energy Physics 1994, pp 229-232.

[6] http://fnhppc.fnal.gov/cap/cap.html

[7] http://wwwinfo.cern.ch/asd/cernlib/rd45/

A BRAHMS Plans

The computing needs for BRAHMS have been outlined in the ROCC report and will not be repeated here. The requirements are relative to the other RHIC experiments modest, approximately 5% of the total required capacity. Even then the amount of data is larger by one or two orders of magnitude than present AGS experiments thus needing specialised facilities.

The computer center as specified in this proposal will fulfill all of the computing needs that the BRAHMS collaboration anticipates at this time. In particular the facilities proposed for the CRS with its Managed Data server is seen as crucial to fullfil the needs of BRAHMS. For the later stages of analysis where datavolumes are smaller we have discussed how one might get around the envisioned lack of cpu resources, and the possibility of being "squezeed out" by the large experiments. The consensus is that the necessary resources could be found using desktop computers. Assuming that the typical workstation purchased in 1999 and beyond is expected to have a speed of around 200 MFLOPS, we feel that piecing together desktops from different parts of the collaboration will make up any difference we might experience due to lack of resources and "competition". This is certainly not the mode we prefer to work in and will without doubt cause the analysis to take longer. We are not currently making plans to use other computer centers beyond the "shared" model calculations that are to be done once for all experiments probably at other computer centers.

B STAR Computing Resources

The offline detector simulation and event reconstruction software for STAR is currently in an advanced prototype phase which has been used as the basis for our estimates of the computing resource requirements for STAR. Physics analysis software, which acts on reduced sets of data summary tapes (DST), or μDSTs, for the purpose of obtaining publishable physics results, is not as well developed. However HBT analyses will most likely dominate STAR physics analysis cpu requirements. Reasonable estimates based on other experiments and actual STAR simulations were used to obtain cpu requirements for STAR HBT analyses. Data volumes associated with the raw data, simulated data and DST production are relatively well understood whereas that associated with physics analysis is less well known. Wherever possible the estimated cpu and data requirements for STAR are based on recent computing experiences for STAR simulations. We have also compared our estimates with actual usages from other TPC-based heavy-ion experiments (LBL-EOS, NA36, NA44 and NA49) as much as possible. The details of these estimates are documented in the RHIC Off-line Computing Committee (ROCC) report. The summary listed below was prepared from that report. In addition, the present state of plans and activities related to meeting these requirements for STAR are described.

B.1 Summary of STAR requirements

The table below lists the total annual estimated requirements for cpu cycles for the various computing activities for STAR. The cpu needs are divided into these various categories because the necessary operating conditions for these categories can be different. For example, event reconstruction, which is a single pass on all the data, is best handled at a dedicated facility sized to just meet that need while capacity needed for physics analysis can vary greatly from week to week based upon the interests of the individual physicists and the timing of conferences and meetings.

Table 13: Annual CPU requirements for STAR.

Process	kSPECint92/ev	Events/yr	kSPECint92-days / year
Event Rec. (real data)	150	14.4M	25,000
Event Gen. Models	75	28.8M	25,000
Simulations (Ev. gen. and Mixed)	150	14.4M	25,000
Physics analysis (data)	195	14.4M	32,500
Physics analysis (sim)	96	28.8M	32,000
Total			**140,000**

The other major computing resource for STAR is the data storage required. This is listed in the table below for various categories as the volume of data generated per year of operation.

The access requirements for this data fits the behavior of an HSM system quite well. Some data needs to be available on a short time scale (minutes) and should exist within the robotic system (tape in robot) while other data only needs to be accessible within

Table 14: Qualitative characteristics of process types.

Process	**Comment** [a]
Event Rec. (real data)	Best suited to dedicated single-purpose facility with modest CPU:I/O ratio.
Event Gen. Models	Suitable for shared facility with high CPU:I/O ratio.
Simulations (Ev. gen. and Mixed)	Best suited to optimized facility with high CPU:I/O ratio. Can be augmented with SCC and LCC.
Physics analysis (data)	Best suited to facility optimized for data access and relatively (compared to other STAR processing) low CPU:I/O ratio. Can be augmented with SCC and LCC.
Physics analysis (sim)	Best suited to facility optimized for data access and relatively (compared to other STAR processing) low CPU:I/O ratio. Can be augmented with SCC and LCC.
General comments	Due to the large event size for STAR it may be that the event reconstruction and GEANT simulation facilities optimized for STAR are not well optimized for experiments with smaller event sizes.

[a] SCC is a Supercomputer Center, LCC is local computer center including user's workstations.

Table 15: Annual data volume summary for STAR

Data Item	MB/ev	Number per yr	Total prod. per yr.(TB)	Total saved per yr.(TB)	Comments
Event Gen.	1	28.8M	28.8	28.8	Same number as real data for sim. plus another set for comparison with data
GEANT+g2t	17 to 29	14.4M	262	0.26	Save 10^{-3}; 14.4M with 10% full, 90% phys. off; 14.4M mixed ev. negligible
Slow sim.	0.05	14.4M	0.7	0.7	10 tracks/event 14.4M mixed ev.
Raw data	16	14.4M	230	230	Tape archive
Calibrated data	32	0.1M	3.2	3.2	during development of DST production
DST	1.6	43.2M	69	23	Assume 10% of raw data size; 1 real + 2 sim. ev., save real data DSTs only
μDST	0.16	72M	11.5	11.5	Assume 10% of DST; 5 per raw event
ntuples	0.2	72M	14.4	14.4	One per μDST
Calibration data	NA	NA	0.002	0.002	calibration database
Total			620	**312**	

24 hours (stored on shelf). The raw data needs to be read only once (on average) and can sit on a shelf while the remaining data is accessed many times and is best stored in the robot. While more data is generated each year some of the data will lose popularity and not be accessed after some point in time. Until more is known about the real access patterns for STAR data as assumption that a one year supply of the non-raw data is the capacity required within a robot is reasonable. This is 82 TB.

B.2 Location of resources

The requirements listed above are large by any computing standards and are larger by orders of magnitude than previous high-energy or nuclear physics experiments. Because of the scale of these requirements, it was recommended in the ROCC report that a serious effort should be made to find resources outside of BNL to help meet this need. In particular, the DOE supercomputer centers were listed as possible sources. Due in large part to the characteristics of the cpu needs, it was recommended that while the majority of the required robotic storage capacity should be installed at BNL, only about half of the cpu should be installed there with the other half being sought at other sites.

At this point the main focus of additional resources for STAR is at the NERSC center which recently moved from LLNL to LBNL. The move of the PDSF computer farm from the SSC lab to NERSC and the close proximity of NERSC to the STAR group at LBNL are the primary motivating factors for interest in NERSC rather than other supercomputer centers at this point in time. The STAR group at LBNL is committed to utilizing PDSF for simulations studies as soon as it is operational at NERSC. Several groups in STAR will be applying for time at NERSC via the normal annual NERSC application process.

There are activities underway related to enhancing the capabilities and usefulness of NERSC for high-energy and nuclear physics (HENP) applications. The nuclear science and physics divisions at LBNL have jointly proposed strengthening PDSF as an internal initiative at LBNL. There is also a Grand Challenge proposal being prepared to address the issue of “Data Access and Data Analysis of Massive Datasets for High-energy and Nuclear Physics”. This proposal is a collaboration of STAR, PHENIX, CLAS, RHIC, NERSC and others and, if approved, will greatly enhance the ability of RHIC experiments to use the supercomputer centers as well as enhance the capabilities for data analysis at the RHIC Computing Facility.

C PHENIX computing, a complete solution

The proposed RHIC Computing Facility (RCF) is going to be the main computing resource for the PHENIX Collaboration. As described in detail in Appendix C of the ROCC report, the computing requirements of the PHENIX experiment are huge and will require that there be a dedicated computing center at the same site as the detector. In the PHENIX computing model practically all of the data will be stored centrally within the storage system at the RCF and all event reconstruction, data mining and data scanning will be done at the RCF. We think it is very important that a very large part of the data analysis and data evaluation be performed on-site at Brookhaven where the students and post-docs doing the analysis will have easy access to the various detector experts and where there would be an envigorating intellectual environment for the exchange of ideas.

However, as also outlined in the ROCC report, the RCF alone will not be able to meet all of the PHENIX computing requirements. PHENIX needs an estimated 150 SPECint92 of CPU power to perform theoretical model calculations and detector simulations of backgrounds, efficiencies and acceptances. Estimates of this additional CPU power done after the ROCC report also indicate that more CPU power is needed.

In order to satisfy this need for CPU power PHENIX is currently exploring the possibility of a regional PHENIX computing center at RIKEN in Japan. The proposed center will perform the following functions:

1. PHENIX Detector Simulation

   Creation and reconstruction of large-scale simulated data for acceptance and efficiency calculations and for background studies.

2. Theoretical Model calculation

   Calculation of large sets of data for comparisons of the models to the PHENIX data. Of special interest will be the creation of a lepton event generator in contrast to the current hadronic models. This model might require large amounts of CPU power in order to perform an unbiased simulation of the rare leptonic signals.

3. A regional data analysis center

   This would allow additional Japanese collaborators (and maybe even many other Asian collaborators) to get access to parts of the data, that would otherwise not be available to them due to bandwidth limitations. Such a regional analysis center would especially allow students and post-docs from the smaller Asian groups to participate more actively in the analysis, since they would otherwise not be able to afford to send people to BNL for long durations.

The scope and functions of this proposed Japanese regional center closely mirrors a similar center that was originally considered as part of Japan's contribution to the SSC in Dallas. Currently we envision, that the center will have a size of 25-33% of the RCF.

In case the timeprofile for the Japanese regional center to become operational should not match PHENIX's needs it is our intention to apply for access to supercomputing resources within USA, in particular from three of the national laboratories participating in PHENIX: ORNL, LANL and Ames.

PHENIX does not plan to create additional regional centers, either in the USA or in Europe. RCF and the proposed regional center in Japan would satisfy the current large-scale needs of PHENIX computing.

We do, however, still envision, that individual institutional groups within PHENIX will continue to update and replace their computing equipment (workstations, local disks etc.) at the same rate as is currently the case. Our goal will be for every active collaborator within the USA to have a small workstation or a PC at his/hers desk and have access to a local file system. This goal should in general be compatible with the current funding level from DOE and NSF to the various institutions through their research grants. In addition we assume, that ESNET will be upgraded so each institution will have very high bandwidth access to RHIC.

Conclusion: RCF in the configuration specified in the current proposal together with the proposed regional center in Japan will fully meet PHENIX's large-scale computing requirements and therefore, PHENIX does not anticipate any additional large-scale computing requests to the nuclear physics office of DOE.

D PHOBOS Off-site Computing Needs

The PHOBOS collaboration plans to perform the bulk of our off-line computing using the RHIC Computing Facility: event reconstruction, DST generation, μDST generation, and some fraction of the analysis of the μDSTs. The off-site computing needs can be divided into model event generation, which should be shared between experiments, and PHOBOS-specific simulation, analysis of simulated data, and some fraction of the analysis of real data.

We estimate the CPU needs for the PHOBOS-specific off-site computing to be in the range of 20–40 kSPECint92 in the year 2000. We can reasonably expect to have access to this level of computing distributed throughout various PHOBOS member institutions. For instance, by the year 2000, we expect to have access to some fraction of a 15–30 kSPECint92 farm at the Massachusetts Institute of Technology as well as access to some fraction of similar facilities at the University of Illinois at Chicago and at the University of Maryland. Assuming reasonable growth and upgrades to these facilities as well as desktop computing and other facilities at other PHOBOS member institutions, we should be able to meet our off-site needs.

It should be noted that high-quality, high-reliability networking is a crucial ingredient in our computing scheme. We are assuming that such networking will be available in the U.S. and most of Europe at reasonable cost in 1999 and beyond, but this is beyond our control. In particular, there is some concern about the quality of the network connection between Krakow, Poland and the rest of the collaboration which has not yet been resolved.

Overall, the PHOBOS computing needs appear to be addressed by the RHIC Computing Facility proposal and other existing and planned resources. Assuming that the RCF is approved and built as planned, the only area of concern is networking to Krakow. This problem should be manageable.

E Glossary of Terms

μDST micro Data Summary Tape produced from a DST.

100BaseTX 100 Megabit Ethernet on UTP cables.

8mm A currently popular magnetic tape format.

AFS Andrew File System, a caching wide area filesystem.

AGS Alternating Gradient Syncrotron, one of the accelerators at BNL which will be used as an injector for the RHIC machine.

API Application Program Interface.

ATM Asyncronous Transfer Mode.

Beta A video tape format used in Sony drives.

BNL Brookhaven National Laboratory.

BRAHMS Broad RAnge Hadron Magnetic Spectometer.

BaBar The B/B-bar detector at SLAC's Collider.

CAP Computing for Analysis Project at Fermilab.

CAS Central Analysis Server.

CASE Computer Aided Software Engineering.

CCD Computing and Communications Division at BNL.

CEBAF Continuous Electron Beam Accelerator Facility recently renamed the Thomas Jefferson National Accelerator Facility.

CISC Complex Instruction Set Computer.

CLAS CEBAF Large Acceptance Spectrometer.

CPU Central Processing Unit.

CRS Central Reconstruction Server.

DAQ Data AcQuisition.

DAS Data Access Software.

DAT Digital Analog Tape, a 4mm magnetic tape format.

DBMS Data Base Management System.

DD-2 A video magnetic tape format.

DD-3 A video magnetic tape format.

DFS Distributed File System.

DLT Digital Linear Tape.

DLT-7K The model 7000 DLT tape drive.

DNS Domain Name Server.

DOE Department of Energy.

DST Data Summary Tape.

EMASS A mass storage company which is a subsidiary of E-Systems, Inc.

ESnet Energy Sciences Network.

FNAL Femi National Accelerator Laboratory.

FTE Full Time Equivalent.

FY Fiscal Year.

GB Gigabyte or 10^9 bytes.

GCE General Computing Environment.

GFLOPS One billion FLoating point Operations Per Second.

HBT Hanbury-Brown-Twiss.

HENP High Energy and Nuclear Physics.

HEP High Energy Physics.

HPSS High Performance Storage System project at the National Storage Laboratory.

HSM Hierarchical Storage Management.

I/O Input/Output.

ID-1 19mm helical scan magnetic tape format defined in the American National Standard Institute (ANSI) X3.175-1990 standard.

IEEE Institute of Electrical and Electronics Engineers.

LAN Local Area Network.

LANL Los Alamos National Laboratory.

LBNL Lawerence Berkely National Laboratory.

LCC Local Computer Center.

LHC Large Hadron Collider.

LLNL Lawrence Livermore National Laboratory.

Mammoth The latest 8mm magnetic tape format.

MDS Managed Data Server.

MFLOPS One Million FLoating point Operations Per Second.

MHz Megahertz or one million cycles per second.

M-O Magneto-Optical, a read-writable optical disk format.

NA44 A CERN based High Energy Physics experiment.

NA49 A CERN based High Energy Physics experiment.

NERSC National Energy Research Scientific Computing.

NFS Network File Server.

NP Nuclear Physics.

NSF National Science Foundation.

NTP Network Time Protocol.

NTP A magnetic tape format used on Sony drives.

OC-3 Optical Carrier - 3, a 155.52 Mbits/sec optical transmission standard.

OC-12 Optical Carrier - 12, a 622.08 Mbits/sec optical transmission standard.

OC-48 Optical Carrier - 48, a 2488.32 Mbits/sec optical transmission standard.

OODBMS Object Oriented Database Management System.

ORNL Oak Ridge National Laboratory.

OS Operating System.

OSSI Open Storage Systems Interconnection.

PAC Program Advisory Committee.

PASS Petabyte Access and Storage Solutions.

PC Personal Computer.

PDSF Particle Detector Simulation Facility.

PDU Power Distribution Unit.

PHENIX Pioneering High Energy Nuclear Interaction eXperiment.

PHOBOS

QIC A half inch magnetic tape format.

RAID Redundant Array of Inexpensive Disks, or, since most are expensive, Redundant Array of Independant Disks.

RAM Random Access Memory.

RCAB RHIC Computing Advisory Board.

RCF RHIC Computing Facility.

RHIC Relativistic Heavy Ion Collider.

RIKEN The Institute of Physical and Chemical Research, Japan.

RISC Reduced Instruction Set Computer.

ROCC RHIC Offline Computing Committee.

SLAC Stanford Linear Accelerator Center.

SMP Symmetric Multi Processor.

SPECint92 Standard Performance Evaluation Corporation integer benchmark from 1992.

SSC Superconducting Super Collider.

STAR Solenoidal Tracker At RHIC.

TB Tera Byte or 10^{12} Bytes.

TByte Tera Byte or 10^{12} Bytes.

TPC Time Projection Chamber.

UPS Uninterrubtible Power Supply.

UTP Unshielded Twisted Pair.

VHS Video Helical Scan, a video magnetic tape format.

WAN Wide Area Network.

WWW World Wide Web.