RHIC Computing Facility

June 24, 1996

Abstract

A scalable computing facility capable of meeting the on-site off-line computing requirements of the RHIC experiments is proposed.

Contents

1 Introduction

1.1 Overview

This is a proposal to establish a computing facility at Brookhaven National Laboratory (BNL) to be used for the off-line handling and analysis of data produced by experiments at the Relativistic Heavy Ion Collider (RHIC). The computing and data handling capacities required for RHIC are large on the scale of previous detector systems in either High Energy or Nuclear Physics. The first serious estimate of these needs was made in 1992[1] with a somewhat more detailed follow-up a year later[2]. A 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[3] (ROCC). The performance objectives of the computing facility proposed here are based largely on the needs described in the ROCC report.

Certain aspects of the RHIC computing requirements are most 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 as 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 is explicitly recommended in the ROCC report.

The RHIC Computing Facility (RCF) being proposed here will have primary responsibility for handling and processing the data produced by the experiments but will operate in conjunction with computing facilities at remote locations so network access will be of paramount importance. In particular it will be responsible for the reconstruction of all collider data. It will also be responsible 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, especially for collaborating institutions which have limited local compute resources. 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 modest levels of simulation work during periods of non-peak processing demand for 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 computing environment at the RCF and remote computing sites engaged in RHIC computing.

1.2 Functional Context

It is important that the RCF be part of the complete solution satisfying all of the requirements described in the ROCC report[3]. As mentioned above there are a number of types of computing which are important to RHIC scientists which are not addressed by the RCF. One of these is desk top computing. Desk top computing 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. Collaborating groups must in general expect to continue supplying basic desk top systems for their members at approximately the level at which they have in the past.

The areas in which the ROCC report describes very significant RHIC specific computing which is not being addressed in the RCF are those associated with simulation and modeling. The CPU requirements in these areas, as will be discussed in the next section, are on the same scale as are those 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 their experiments currently expect to satisfy those computing needs which are not addressed by the facility proposed here. 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 resouces 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 Nation Energy Research Scientific Computing Center (NERSC) which recently moved to Lawerence Berkely 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 system 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, the submission of a Grand Challenge proposal, 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

There are two important management issues relating the RCF proposed here to the RHIC community. The first is the manner in which the resources of the RCF will be allocated among the various experiments. The second is the mechanism by which the experiments will have substantive input into technical decision made regarding the facility. Decisions regarding resource allocation include judgements of physics priority as well as an understanding of technical requirements and capabilities. An objective evaluation of the relative physics priorities is expect to be arrived at by the appropriated BNL scientific management with the advice of the laboratories Program Advisory Committee. Approximately once per year, the RCF management will be given general guidance regarding the scientific priorities of the various experiments and perhaps various projects within experiments. This guidance is important to planning both within the facility and by the experiments.

In terms of implementing this guidance, a RHIC Computing Advisory Board (RCAB), comprised primarily of repesentatives from the RHIC experiments will be formed. This board, convened by the head of the RCF, will advise the RCF head, BNL management and the DOE on matters relating to the appropriateness of specific deployments of resources. The RCAB will also serve as the mechanism by which experiments can officially express their views on technical and operational decisions being made regarding the RCF. The RCAB will meet several times a year and will be consulted prior to any major decision regarding technical direction, resource acquisition or operational policy.

Once each year there should be a formal review of the RCF by outside experts. Such a review is a way to assure that the facility is functioning effectively.

2 Requirements

Estimates of the computing needs of major detectors produced years before their turn on will, as a matter of course, 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 most recently in February of 1996. Each of these estimates has been more refined and more 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 most recent and most realistic 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[3]. 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.

Table 1: Total estimated CPU needs of the four RHIC experiments in units of kSPECint92.

	Brahms	Phenix	Phobos	Star	Total
Event Reconstruction	18	175	120	84	397
Models					165
Simulation + Reconstruction	10	75	30	87	202
Physics (Simulation)	4	10	6	180	200
Physics (Data)	5	80	25	180	290
Total	40	415	184	615	1254

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 cost are between $2 and $5 per GByte (see Figure 6). The annual cost of media for RHIC data storage,

Table 2: Estimated storage needs for different stages of analysis. The numbers are in Terabytes per year.

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

depending on which media is 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 recording media. There are some new technologies, such as optial tape, which promise substantial reductions in storage costs and these will 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 use is necessary.

Raw data is typically read into the reconstruction system, hopefully 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 it will never be read into the computing system again. Exceptions are early data, reconstructed before the reconstruction program is fully perfected, which are likely to be read in a second or even more times for reprocessing with improved versions of the reconstruction program 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 in 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 amount of the more expensive storage. 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 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 are today commonly reflected in Hierarchical Storage Management (HSM) systems and in databases, relational and object oriented. 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 premium cost storage capacity required.

While not contained in the tables, a number of additional requirements are expressed in the ROCC report. 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, 50 MBytes/sec of which will be required to handle the recording of data being taken when the detector runs. 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 (75 MBytes/sec) in 1999. Any decision to dramatically upgrade this capacity would involve the entire ESnet community and would involve major additional expenditures. OC12 has thus been defined as the expected level of WAN connectivity in 1999.

2.2 First year requirements

Table 3, also taken from the report of the RHIC off-line computing committee, shows growth in capacity as a function of time for a dedicated computing facility located at RHIC intended primarily to handle and process collider data. The ramp-up shown here was thought by the experiments to be acceptable based on recognition of the fact that in 1999, as a result of the need to commission both the accelerator and the detectors, the data sets actually recorded were likely to be small compared to the nominal year running which was used in the estimates appearing in Tables 1 and 2. Since the ROCC report was published, funding reductions in the RHIC project have further reduced the amount of data which is likely to be taken in 1999. This reduction in need coupled with considerations discussed above have resulted in a set of 1999 design goals for the RCF which is shown in Table 4. These somewhat reduced goals have been discussed with the computing representatives of the various RHIC experiments and found to be acceptable.

Table 3: Accumulated Capacity of RHIC computing facility.

	1997	1998	1999	2000	2001
CPU - kSPECint92	20.7	104.0	520.0	764.0	1065.0
Disk - TBytes	1.2	6.1	30.6	40.1	68.9
Robotic - TBytes	4.0	20.0	100.0	172.9	313.6

Table 4: Goals for the RHIC Computing Facility in 1999

Resource Type	Target Capacity in 1999
Reconstruction CPU	250 kSPECint92
Data mining CPU	75 kSPECint92
Analysis CPU	75 kSPECint92
General Computing CPU	15 kSPECint92
Robotic Storage	100 TBytes
Disk Storage	25 TBytes
I/O Detectors $\rightarrow$ RCF	2 x 50 MBytes/sec
I/O MDS Tape Robots	200 MBytes/sec
I/O MDS Disk $\rightarrow$ CAS	700 MBytes/sec
I/O MDS $\rightarrow$ WAN	ESnet speed (75 MBytes/sec)

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 elements of the facility being proposed here and which functions they will serve.

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 this CPU 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 of the history of the data being used by virtue of there being only a single copy of each

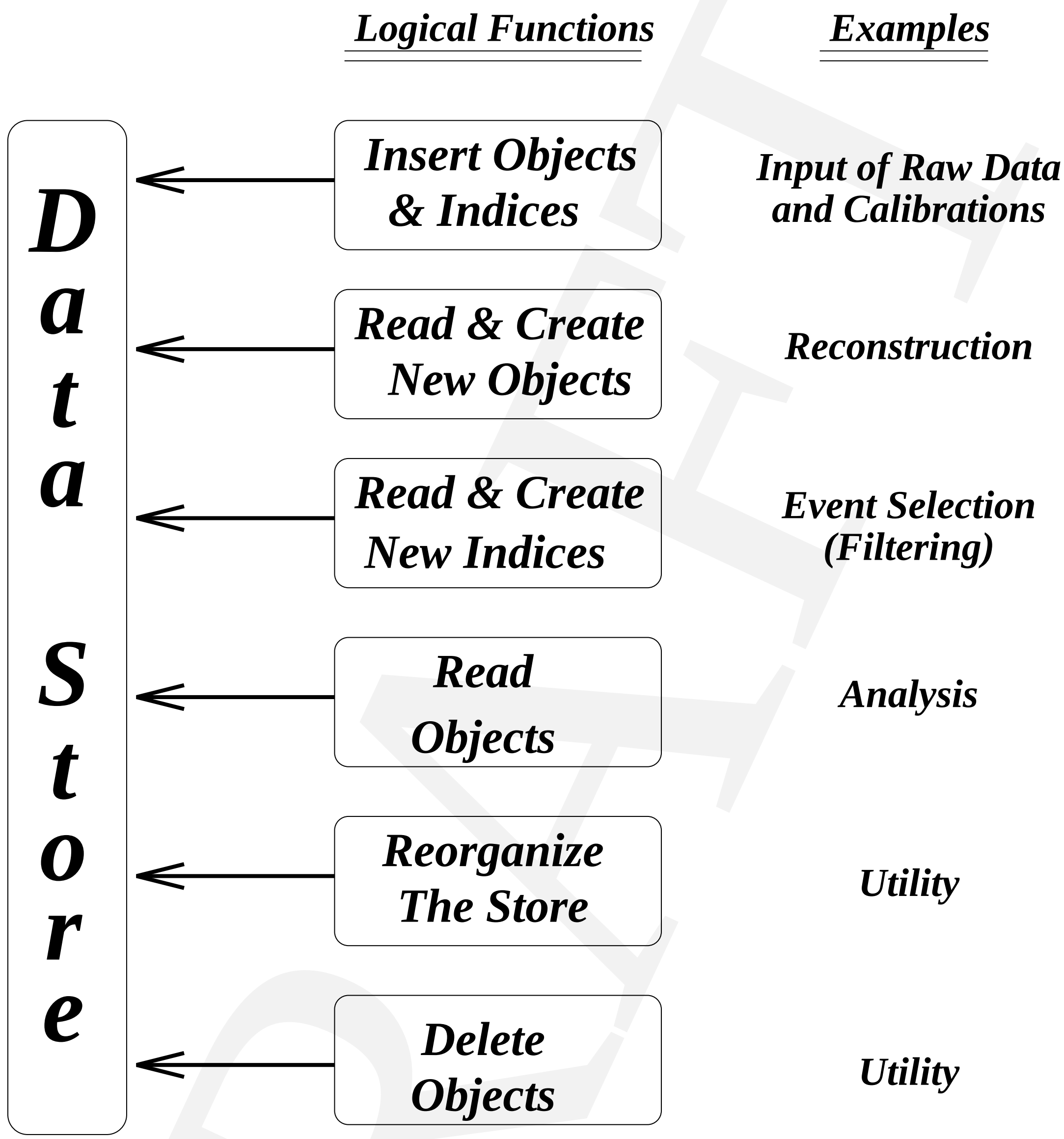

Figure 1: Conceptual model of RHIC Off-line computing.

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, it will usually be physically found on shelves. DST level data, requiring more immediate access, will usually be found physically on robotic tape. Relatively small data sets such as μDST's or n-tuples which result from selection passes through the DST's, a process referred to as "Data Mining", 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, would frequently be physically found on disk.

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 address 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 other 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 proposed here. There are basically four distinct components. 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 including DST's, the data produced as μDST's, and the data being used either locally or remotely for analysis.

The CAS is responsible for μDST creation by accessing the data from the MDS and writing results back to it. The CAS is also responsible for performing analysis and is especially optimized for performing very large scale analyses which are not practical on smaller scale systems. It is planned that μDST production passes 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. This system is the natural extension of the existing RHIC computing cluster.

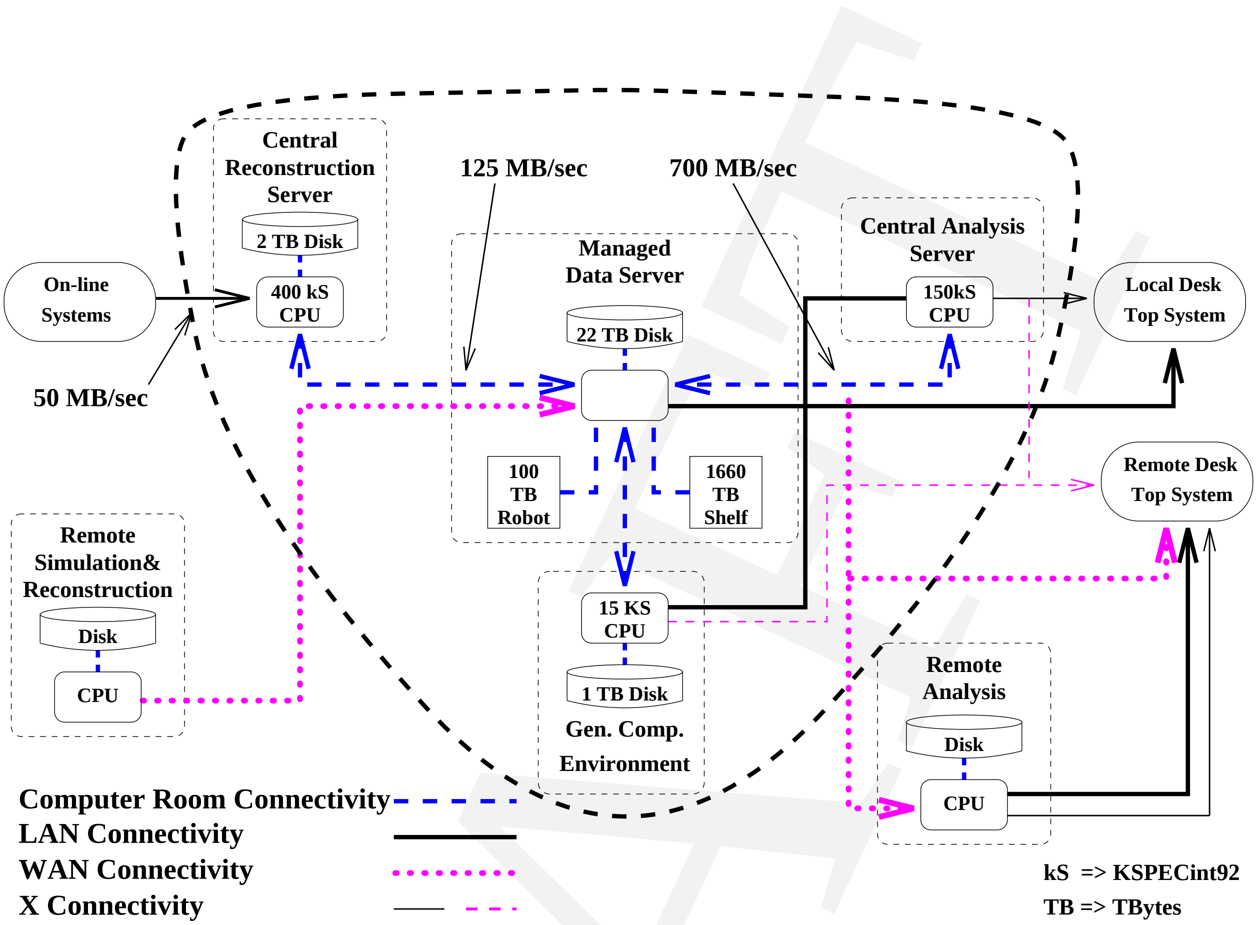

Figure 2: Physical model of RHIC Off-line computing.

4 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 with 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.

4.1 Strategic Issues

Since this proposal is for equipment which will be purchased over a three 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 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 the initial purchase price. However, it is recognized that there are other factors which must also be considered.

It is clearly most cost-effective to purchase computing equipment as late as possible. However, if one is to estabish 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 strategies must be reconciled. It is proposed here that the acquisition process be approximately geometric over the three years. In the first year a system of approximately 4% 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 16% of the capacity will be acquired bring the total to 20%. This will verify the scaling of the system by a factor of four and allow a substantial increase in the the statistics on individual components. This factor of four increase in scale and complexity will also serve to validate simulations and models of the system which will have been developed. In the third and final year the remaining 80% will be purchased, another scaling by a factor of four, bring the system to full size. In this way, while the bulk of the purchases actually occur in the last year, progressively more complete information and relevant experience is acquired in each of the preceding years.

4.2 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 final year purchases (which will be more than 50% of the total) can take full advantage of any price or performance gains during the previous year. 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).

4.3 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 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.

4.4 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.

4.5 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 satifies particular well define performance requirements, there remainds a significant probability that subtle aspects of their interactions may produce bottlenecks which limit performance substantially below what one would nievely 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.

4.5.1 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.

4.5.2 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.

4.5.3 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.

5 Facility Components

5.1 Central Reconstruction Server

5.1.1 Requirement

As discussed in the introduction, during a nominal running year the four currently approved experiments will acquire on the order of 700 TBytes of data. The amount of CPU power required for the first stage reconstruction has been estimated in the ROCC report[3], and a target value of 250 kSPECint92[1] has been agreed upon for turn on in 1999. 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 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 bench marks 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.

5.1.2 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 (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 1999. 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 4 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 possible configuration of the Central Reconstruction Server (CRS) is indicated in Figure 4. The CRS will consist of four such units, one for each of the four experiments, with the number of components scaling to the required bandwidth for each experiment. Each of the units will consist of one or more *data logger* machines which will contain either a fiber connection(s) from the experimental hall or tape drive(s) (operated, and perhaps provided, by the experiment) to read the raw data tapes produced by the experiment. The ability to read tapes into the *data logger* machines is provided since

[1] 1 kSPECint92 = 1000 SPECint92 or approx. 0.33 GFLOPS

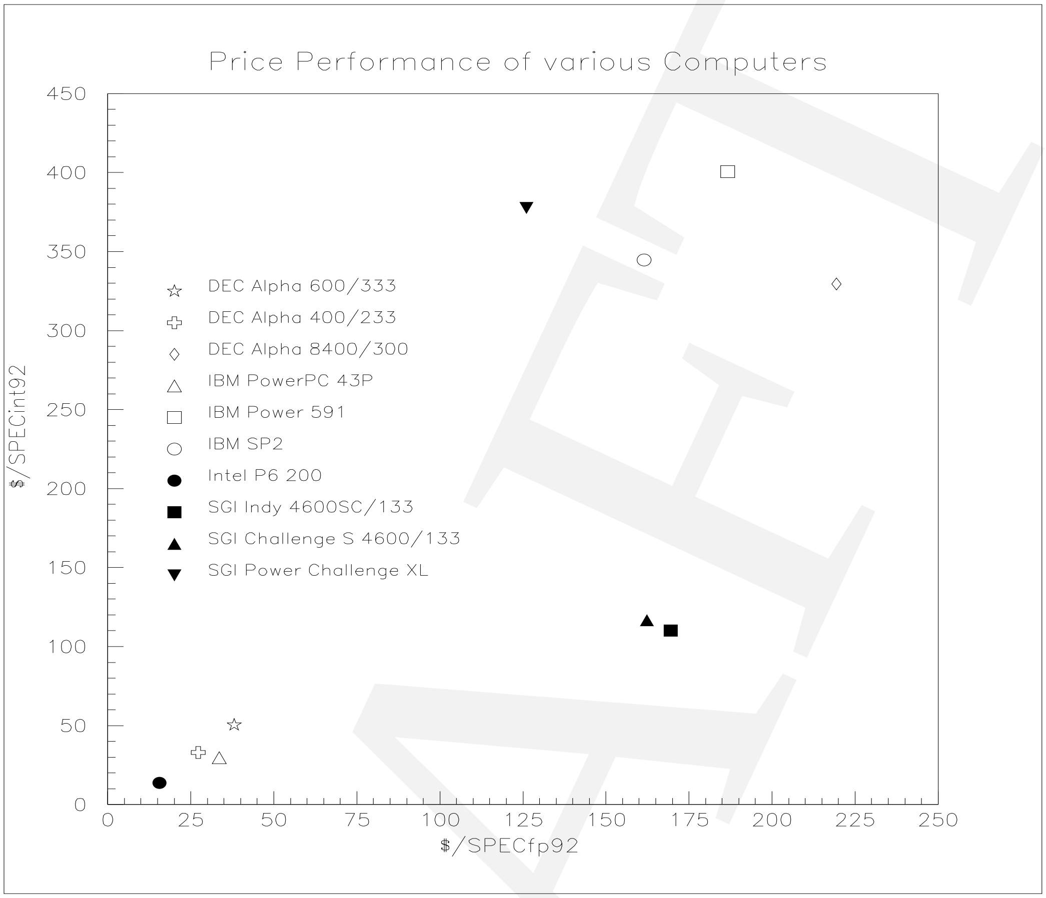

Figure 3: The price/performance ratios of various computers. The list price of the base configuration of each of the models indicated was used.

it is possible that in the early stages of the machine and detector commissioning it may be desirable to write to tape for a variety of reasons. Once all of the systems and reconstruction software matures, it is assumed that the experiments will send all data directly to the RCF via a fiber link and retain the costly (in dollars and in manpower) taping capability for the presumably very rare periods when fiber connectivity to the RCF is lost. The disks on the *data logger* machine will be part of the Hierarchical Storage Management (HSM) system so that files loaded onto the machine will immediately be managed by the HSM. If there is a delay before the data can be reconstructed (e.g., calibrations are not yet available), then the data will naturally be swapped out to tape by the HSM (the HSM will be configured to keep all such swapped out data together to optimize its retrieval during a subsequent reconstruction pass).

Each unit 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 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 and the collection of the results of the reconstruction from the worker

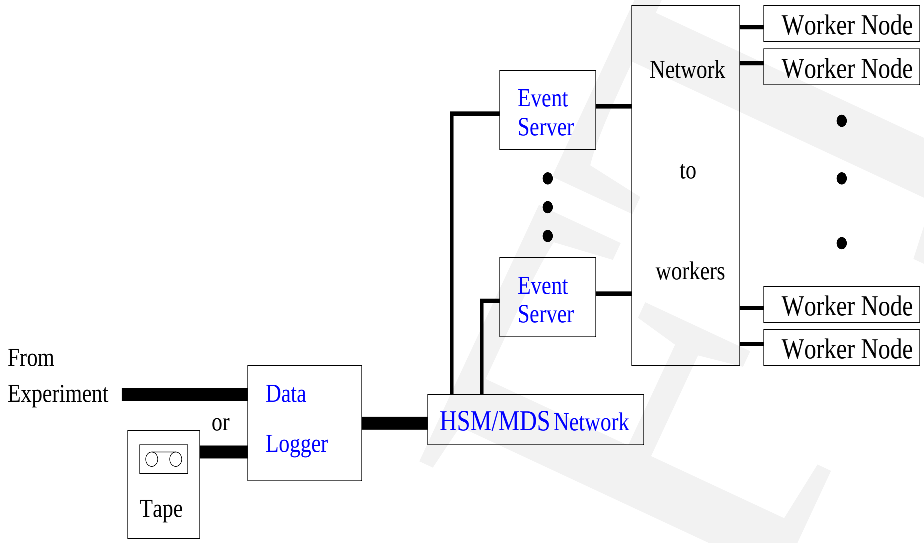

Figure 4: Schematic Diagram of Central Reconstruction Server

nodes. Finally, the *event server* will present the reconstructed data to the HSM in the form appropriate for further analysis on the Central Analysis Server (CAS).

Each unit will also have a number of *worker nodes* (mentioned above) which will each run independent copies of the reconstruction code defined by the collaboration and will completely reconstruct an entire event.

The CRS is set up to be scalable. As mentioned above there will be four of the units described in Figure 4, one for each of the experiments. The number of elements (*data logger*, *event server* and *worker nodes* in each of the units will depend on the requirements for each of the experiments and is expected to grow over time. 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.

5.1.3 Relevant commercial evolution

CPU costs, as measured in $/SPECint92, 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 CPU cost. A recent event of this type was the introduction of the 200 MHz Pentium Pro® (P6) which is rated at 320 SPECint92. This put the cost of CPU at $22/SPECint92 in May of 1996 when six months earlier the best cost was about $50/SPECint92, 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.

5.1.4 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 the HSM and then send an acknowledgment of receipt of the data back to the experiment. For the tape option, the experiment's Data Acquisition System (DAQ) will produce tapes which will be periodically collected and brought to the RCF. The tapes will then be read into the *data logger* machine which will write the data to the HSM as in the fiber option.

The disks of the *data logger* machines and the *event server* machines will be part of the HSM, so that any data moved to or from the disks by the HSM will travel over the network shared by the HSM and Managed Data Server (MDS). Assuming that the CRS can keep up with reconstructing 50% of the data, it will place a load on the order of 100 MBytes/sec (depending on the size of the reconstruction output relative to the raw data input) on the network during periods when the accelerator is running and a load on the order of 50 MBytes/sec during periods when the accelerator is shut down.

5.1.5 Expected performance

5.1.6 Acquisition and installation

The CRS, as with all components of the RHIC Computing Facility, will be phased in over three years with the bulk of the purchases being done in the last year. The proposed acquisition breakdown is 4% of capacity acquired in 1997, 16% of capacity acquired in 1998 and the final 80% of capacity acquired in 1999. An example of such a ramp-up follows. It is assumed that all of the *worker nodes* are four CPU Symmetric Multi-Processor (SMP) 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).

1996:	10 CPU Prototype	$ 80000
1997:	1 Data Logger	$ 4800
	1 Event Server	$ 4800
	4 23 GB disks	$ 8300
	5 SMP machines	$ 162200
		$ 180100
1998:	3 Data Logger	$ 10200
	3 Event Server	$ 10200
	16 23 GB disks	$ 21000
	14 SMP machines	$ 453900
		$ 495300
1999:	7 Data Logger	$ 16800
	7 Event Server	$ 16800
	40 56 GB disks	$ 66000
	48 SMP machines	$ 1556000
		$ 1655600

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 is being assembled in 1996. This prototype is intended to explore the feasibility of using consumer market computers in the CRS, and serve as a functioning version of what appears at this time to be serious candidate for the final form of the CRS.

5.1.7 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 will be 1 FTE in 1997 at the 4% level of hardware, 2 FTEs in 1998 at the 20% level of hardware, and 4 FTEs in 1999 at the 100% level.

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.

5.1.8 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.

5.2 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.

5.2.1 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. 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, 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 then 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 new μDSTs. 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. Although there have been no final decisions on the details of the data storage within the MDS it is likely that the data mining operation will rely heavily on a standard database query facility. This will require that the MDS provide a relatively high performance database server system with robust commercial database software.

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 22 TBs of online storage via magnetic disk, 100 TBs of nearline storage via magnetic tape with robotic access and, shelf storage for up to a few PB's 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.

5.2.2 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 here is between low cost single disk drives and high-performance RAID devices. Although there are many advantages to RAID based storage, 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 roughly \$1/MB while the cheapest non-RAID disks are only \$0.20/MB. In order to achieve a volume of 22TB 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.

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 are only \$1/GB and 8mm drives are only \$4000/MB/s, the number of cartridges and readers required to reach 100TBs of robotic storage with 200MB/s access are 15000

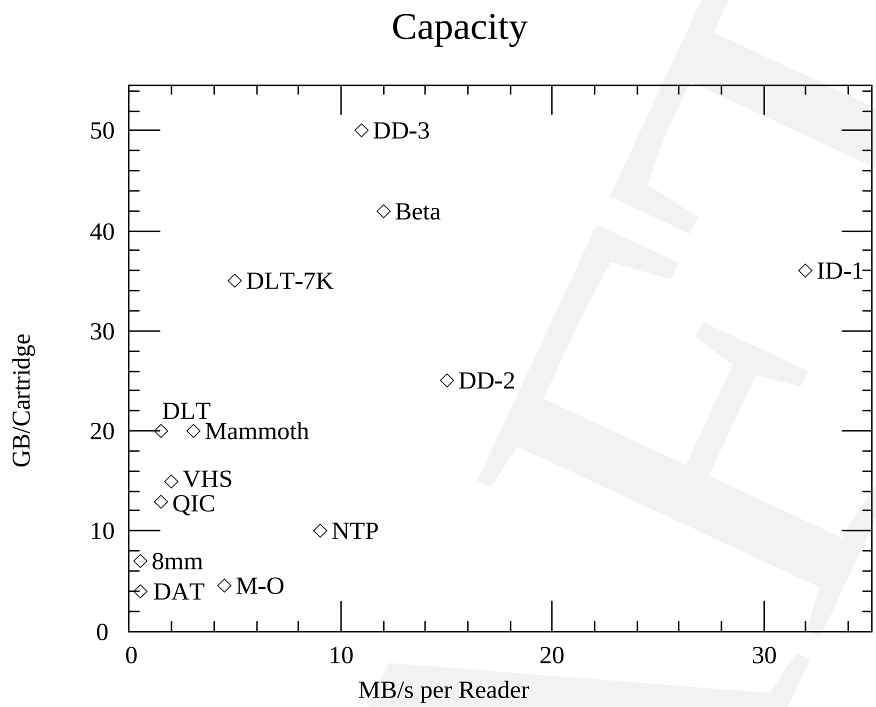

Figure 5: Cartridge Capacities and Access Rates for common tape formats.

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 5 and 6. In Figure 5, 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 6, 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 5 are disfavored in Figure 6 due to their higher cost.

The current best choice under these constraints is Digital Linear Tape (labeled as DLT-7K in Figures 5 and 6. With a capacity of 35GBs/cartridge and 5MB/s/drive, the MDS requirements can be met with a robot containing 2800 cartridges and 40 drives (easily attainable via products from a number of companies including Mountaingate and EMASS). Although DLT may not be the correct final solution for the MDS, it is a good working solution based on products available in today's market. 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 a little 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

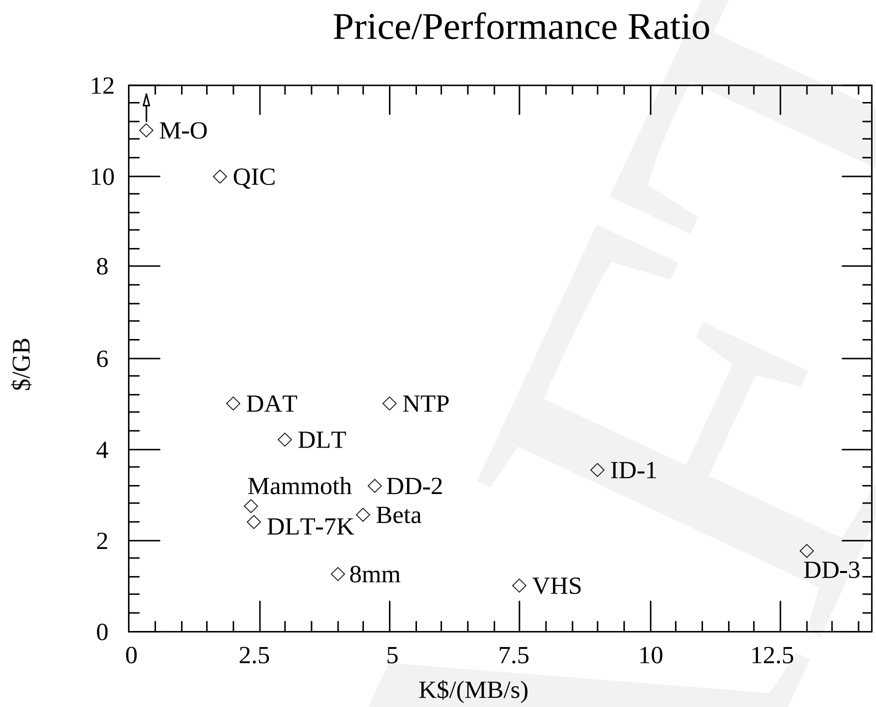

Figure 6: Price/performance ratios for common tape formats in terms of cartridge capacity and data access rate.

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 will initially use an interim low-cost HSM solution, currently being acquired as part of a prototype system, during which 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.

Database: This element of the MDS is the most difficult to predict. Since the experiments are still designing the analysis software, the details of the data storage model are still uncertain. It seems clear that the raw data volume is too large to be included in a standard database, but it is also clear that the data mining operation will at least require some searchable indexes into the stored raw data. Whether this is best accomplished through a relational database, an object oriented database, or a specialized system (like that envisioned in the PASS project) is still to be determined. In any case a substantial database server system will be required, so for the purposes of cost estimates, it is assumed that the required system will be similar to currently available high performance (as measured by the TPC benchmarks) RDBMS systems.

5.2.3 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. Two new generations (the Elite 3 and Elite 9) have been introduced and each of these represented a capacity improvement of nearly a factor of 3. In each of these cases the introductory price was an increase of roughly 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 nearly two years have passed since the introduction of the Elite 9, this analysis suggests that a new device should be available soon, and Seagate has announced a new Elite 23 for fall 1996 availability which will increase capacity by a factor of about 2.5 (no pricing is available at this time). It is reasonable to expect that one more new generation will be become available within the time frame of this proposal, so for the purpose of cost estimates, it is assumed that a 56GB drive will be available in early 1999.

DLT tape technology is also evolving, although somewhat less predictably. 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 2 years. For the purpose of cost estimates, it is assumed that this device will be available in late 1998 and that the cost will again be 50% more than the current technology. Subsequent generations of DLT will probably come too late for this project. An additional factor which may improve the price/performance is that DLT is beginning to replace 8mm as media of choice for general purpose applications. If this trend continues there should be some volume based reductions in the cost of DLT drives and media. This potential improvement has not been included in the cost estimates.

Robotic systems are not commodity items and have been developing at a much slower pace. It is not expected that the costs of the robotic devices will change substantially during the duration of this proposal.

5.2.4 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 1 GB/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.

5.2.5 Acquisition and installation

The RHIC Computing Facility will be phased in over three 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 4% in 1997, 16% in 1998 and 80% in 1999. Here is one 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 as an example of how the proposed system could be achieved, not as a literal blueprint of expected purchases). The purchases described in 1996 are for a prototype system being established to allow both users and operators to acquire general experience with HSM systems, to study patterns of usage and to serve as a development platform for specialize data access software. Only purchases described for 1997 and beyond are part of the proposal being made here.

1996:	1 Mountaingate D-360 robot	$ 85000
	2 Quantum DLT-7000 drives	$ 30000
	50 35 GB cartridges	$ 4200
	Metior HSM License	$ 30000
		$ 149200
1997:	1 Sparc 5 server	$ 11500
	10 23GB disk drives	$ 27400
	50 35 GB cartridges	$ 4200
	HPSS HSM software	$ 115000
		$ 158100
1998:	HPSS HSM software	$ 135000
	2 Mountaingate D-480 robots	$ 170000
	3 Sparc 5 servers	$ 36000
	4 Quantum DLT-9000 drives	$ 68000
	36 23GB disk drives	$ 82800
	600 50 GB cartridges	$ 72000
	Database Server	$ 100000
	Database Software	$ 50000
		$ 713800

1999:	1 Mountaingate D-360 robot	$ 85000
	1 Mountaingate D-480 robot	$ 85000
	18 Sparc 5 servers	$ 216000
	16 Quantum DLT-9000 drives	$ 240000
	1330 50GB cartridges	$ 146300
	374 56 GB disk drives	$ 860200
	Database Server Expansion	$ 200000
	Database Software Upgrade	$ 75000
		$ 1907500

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. 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. 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).

5.2.6 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 1996 (existing items), the maintenance costs will begin in 1997 and achieve a stable plateau in 2000.

	1997	1998	1999	2000
Disk Maintenance	$ 0	$ 0	$ 0	$ 0
CPU Maintenance	$ 0	$ 1700	$ 22100	$ 84500
Robot Maintenance	$ 12900	$ 12800	$ 38200	$ 63800
Tape Drive Maintenance	$ 4500	$ 4500	$ 14700	$ 48400
Software Maintenance	$ 5000	$ 8000	$ 35000	$ 42500
	$ 22300	$ 27000	$ 110000	$ 239200

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.

5.2.7 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 late 1997 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, if an integrated data management package like the concept being pursued in the CAP project is available it might be used as a complete replacement for a standard HSM with additional custom software provided by the RHIC experiments. However, unless a clearly better solution is just over the horizon, this choice should not be delayed beyond 1997 - 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 (including DLT) 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.

5.3 Central Analysis Server

5.3.1 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 analyze 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 CAS is expected to provide 150 kSPECint92 of computing power and a network bandwidth to the MDS of 700 MB/sec, as stated in the Requirements section above. An additional 15 kSPECint92 of computing power located in the CAS will 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 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.

5.3.2 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 expect to be base 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 320 SPECint92 per processor) to provide an aggregate of 150 kSPECint92 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 and administrative advantages 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 700 MB/sec during data mining operations. To meet this need each node would require 100 Mbit/sec switched ethernet at a minimum, for an aggregate theoretical bandwidth of 1.5 GB/sec. More likely, these nodes will be configured with ATM OC3 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

5.3.3 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 interface cards in the hope of further increases in MDS/CAS connectivity.

5.3.4 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 a 100 Mbit/sec 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.

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.

5.3.5 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 planned acquisition breakdown is 4% in 1997, 16% in 1998 and the final 80% in 1999. 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:	3 SMP machines	\$ 97300
1998:	9 SMP machines	\$ 291800
1999:	29 SMP machines	\$ 940080
	3 (GCSE) SMP machines	\$ 97300
		\$ 1426480

5.3.6 Operation and maintenance

There will also 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.

5.4 General Computing Environment

5.4.1 Requirements

In addition to the various specialized servers previously describe 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 asside form those specialize 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. This includes providing platforms corresponding to those which make up the servers and 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 application and utilities, commerical and otherwise, which are being employed at the RCF.

5.4.2 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 two or three 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, and two multi-CPU cycle-servers (an eight CPU SGI SMP and a 16 node IBM integrated farm).

The present hardware configurations will be reorangized utilizing load-sharing strategies and augmented to include platforms corresponsing to the various specialized server systems contained in the RCF. In particular, the GCE will be configured so that a user will be automaticly logged into the less heavily loaded nodes and jobs which are not host-specific will also automatically execute on less heavily loaded nodes. The system will be configured so that activities such as compilation and linking will take place on appropriate platforms. The GCE will also serve as an access point for initiating tasks on the CAS.

Most of the fileserver capability, including user files, collaborative software efforts, library code, and most files other than system executables and scratch disk space, will be migrated to disk managed by the HSM system. Appropriate migration schemes will be employed to ensure that active user files are not displaced by the incoming experimental data.

5.4.3 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.

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

5.4.5 Spending Plan

5.5 Data Access Software

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 infrastructure level software. At a level above this there will be commercial data handling software including a

Table 5: Funding profile for General Computing Support and Environment in dollars.

	1997	1998	1999	Total
Hardware	70K	140K	210K	420K
Software	75K	150K	100K	325K
Total	145K	290K	310K	745K

Hierarchical Storage Manager (HSM) and a Database Management System (DBMS), most likely an Object Oriented DBMS. At a level above this, but possibly supplanting portions of the commercial layer, will be a layer of software which supplies a customized API for data access by RHIC, or perhaps more generically by HEP/NP, experiments allowing access to very specific data objects located in a hierarchical storage system. The PASS and CAP projects are examples of development work being done along these lines. Once this API exists there will be a data definition layer which is customized to the needs of each experiment, reflecting the specific way in which their data is best organized.

Realization of the above described Data Access Software (DAS) represents a potentially major R&D effort. At minimum, adapting existing software to satisfy RHIC needs will be required and very likely the development of substantial new software will be necessary. In the event that there is no package adequately developed to adopt or adapt, the goal would be to integrate established, probably commercial, software to perform the required data access functions. An OODBMS system might be used to store meta-data regarding the location of objects in a file/directory structure and then an HSM would be used to store and manage the meta-data describing the physical location of file/directory structure within the hierarchy of storage media. The development effort would consist of defining and writing the API and the interfacing layers between the API, the OODBMS, and the HSM. Proceeding in this way one would hope to minimize the actual quantity of code which must be written. If appropriate standards are used one might be able to produce a system which would allow the transparent replacement of the OODBMS and and HSM when appropriate. The project will in any case represent a major effort in which members of the RHIC computing group will need significant contributions from the experiments. The first step of this process will be to specify the requirements for such an access system. This will be done in parallel with reviewing the goals and status of existing projects of this type. With the requirements and reviews in hand a decision will be made as to which of the existing projects is sufficiently promising to adopt or whether a ground up development is more appropriate.

Assuming that a significant effort is required the basic time line for the development of Data Access Software (DAS) will be approximately as follows:

- Jul 1997: Joint RHIC Computing Group/Experiments working group formed.
- Oct 1997: Requirements document & reviews of existing projects completed.
- Dec 1997: Optimum course of action decided including essential planning.
- Jul 1997: First prototype test of DAS.
- Dec 1997: Version 1.0 release of DAS - basic function.

- Jul 1998: Version 2.0 release of DAS - full function.
- Jul 1999: Version 3.0 release of DAS - performance optimized.

The profile of technical effort for this project, which would include contribution in approximately equal proportion from the each of the two major experiments and BNL in the form of the RHIC computing group and CCD would be approximately as indicated in Table 6.

Table 6: Technical effort profile for the development of Data Organization and Access Software. Numbers are in FTE's including contribution from PHENIX, STAR and RHIC/CCD.

	FY 1996	FY 1997	FY 1998	FY 1999	FY 2000
Technical Effort	0.3	3.0	4.0	3	3.0

5.6 Local and Wide Area Networking

- Using the network configuration presented here, the required performance can be achieved with currently available technology.
- Considerable price reduction and/or performance gains can be expected for networking products in the next two years. Despite this favorable trend, the projected cost of the required solution is approximately $644k. 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.

5.6.1 Requirement

1. High port count and switch density

   In this model, the greatest challenge is accommodating the large number of systems to be connected to the overall switch fabric as opposed to very high bandwidth. (Bandwidth to any individual system does not exceed 20 MB/s, which is well within the capacity of OC-12 ATM.) Currently available switches typically have capacities up to 16 OC-12 modules each, either as single OC-12 interfaces or groups of 4 OC-3 ports. At the required port densities (approximately 44 OC-12 and 153 OC-3 ports) this necessitates a group of switches acting as a single switch fabric. The disadvantage of such a configuration is the cost of ports used to interconnect the switches and the potential physical level load balancing that this implies.

2. Performance over the Wide Area Network (WAN)

   The WAN requirements specified in the ROCC report[3] present a significant technical challenge in terms of bandwidth and the implied end to end quality of service. The report concludes that there are "no technological bottlenecks here", but does not acknowledge the very formidable financial issues involved. It is also not generally appreciated that the overall WAN path between BNL and any RHIC collaborator location typically contains several links that are not within BNL's sphere of influence. Improving these links would require a substantial investment of resources.

3. Dynamic requirements, technologies, resources

   As the cutover date for the RHIC project nears, frequent changes can be expected in data handing, system capabilities, and performance modeling which may mandate revision of the network configuration. This trend is already evident. During the same time period, evolving products in the networking arena provide us with alternate means of meeting these requirements, compounding the problem of arriving at a stable design so that work can proceed with some degree of certainty.

5.6.2 Definitions and Underlying Assumptions

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

1. Definitions

   LAN: (Local Area Network) The portion of the network that resides at BNL. Within this campus area it is assumed that single and multimode fiber optic cable are available for the RHIC experiments to allow direct connection to the switch fabric. In some cases this may include parts of the existing BNL campus network.

   WAN: (Wide Area Network) The remainder of the network connections between off-site collaborators and/or remote computing facilities. This is almost entirely assumed to be via ESnet.

   OC-3: ATM providing 155 Mb/s raw throughput; most of these connections will run on Category 5 UTP (unshielded twisted pair) for cost reasons.

   OC-12: ATM providing 622 Mb/s raw throughput; these connections are all fiber optic connections, using multimode fiber optic cable within the RHIC Computing Center, and single mode fiber optic cable for connection to the data acquisition systems approximately 3 kilometers distant.

   Switched Ethernet: Ethernet connections providing a dedicated 10 Mb/s path to each system. These connections run on Category 5 UTP cabling.

2. Assumptions

   (a) Total bandwidth coming from the experiments' data acquisition systems is based on the ROCC report[3].

(b) 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
- 11 data loggers
- 11 event reconstruction servers
- 67 Reconstruction systems
- 45 Analysis systems, 1 analysis server
- 22 Managed data store / HSM servers
- 1 Calibration database, 1 event index database server
- 10 General computing systems

(c) Flow control is assumed to exist at the application, transport, and data link levels. The ATM data link level acts as the last resort for flow control and enforcement of agreed traffic levels on any given logical connection.

(d) ATM switch configurations and prices are based largely on the Fore ASX-1000; no vendor selection or recommendation is implied by this choice, but rather a representative "real" product to serve as an accurate baseline for the 1996 time frame.

(e) The price of the network interface cards are included in the cost estimates for the individual systems, not part of the price estimate shown in Section 5.6.3 below.

5.6.3 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 at OC-12 or OC-3 speeds, depending on the experiment. (See Section 5.6.2 above.)

   The required fiber optic infrastructure is largely already in place, and will be completed with the upgrade of the AGS to CCD link in the second half of 1996. The resulting connection will provide at least 8 strands of single mode fiber from any given experimental hall to the RCF to allow for primary, backup, and future connection requirements.

   The ROCC report stipulates an "independent backup" for each connection. Investigation of 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 the full 2 Km between locations. The solution was to provide spare fiber capacity over the existing path and include additional backup ATM switch interfaces in the design. The data acquisition systems will then be equipped with a "standby" ATM interface card to take advantage of this feature.

The bulk of the RHIC Off-line computing network will reside in an approximately 10,000 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. This close proximity allows for the use of OC-3 ATM over Category 5 copper cabling, which represents a significant cost savings over multimode fiber - primarily in the cost of installation and patching facilities, and to a lesser extent, the switch port and interface cards.

The core of the proposed network is an ATM switch fabric composed of a group of separate switches; in 1996, this would consist of 6 switches, each holding up to 16 modules. Each module provides 622 Mbps of non-blocking switch capacity. The cost projection assumes the use of 18 OC-12 links to interconnect the switches. In the 1998 time frame, it is reasonable to assume that the same connectivity will be available in 2 - 4 larger switches, with some nominal decrease in price.

The choice of ATM as the central switching technology was driven by a wide range of factors which are listed below:

- Network interface cards are available in the appropriate bandwidth ranges and for a wide variety of platforms;
- Availability of relatively large switches with non-blocking performance;
- Provision for traffic shaping within the switches as a means of fairly distributing capacity between experiments and sub-functions (reconstruction / analysis / storage);
- Seamless integration of ATM-based WAN services expected to be part of the ESnet offering by the time the experiment goes on-line;
- Support for multiple simultaneous connections with the ability to control fractional segments of bandwidth within them;
- Fairly well documented, if less than stellar, performance;
- A great deal of commercial development of products likely to increase the performance and / or lower the cost of the network;

In order to reduce cost, 10 Mb/s switched ethernet interfaces will be used on the 67 reconstruction systems. Two switches will each support 34 ports via two OC-3 (155 Mb/s) ATM connections, with a resulting raw bandwidth of approximately 9 Mb/s to each of the systems.

The use of an ATM network in the discussions in this proposal is not intended to exclude other network fabrics (such as FibreChannel) from consideration for the purchased system. It is simply the best currently available technology for estimating needs and costs.

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. In order to count on end to end ATM-based services between the RCF and remote sites, ESnet will have to offer access in the form of a switch instead of a router. This has many implications elsewhere, and once again should not be assumed to be a trivial exercise. (ESnet ATM switch interfaces are being introduced at some sites in 1996.)

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.

A small project is underway at this time to help collaborators document "weak links" in their path to BNL in hopes of facilitating upgrades in bandwidth and connection quality to each remote location before 1999. BNL is also participating in a larger ESnet effort to work on connectivity issues and tools with SLAC, FNAL, and ORNL.

3. Cost Projection

   Table 7 below shows the cost estimate breakdown for the network design. The primary underlying premise was to provide the required connectivity using technology available now and then to use that 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 two years before purchase.

Table 7: Networking Cost Estimates

	Quantity	Unit Cost	Related Cost	1996 Cost	96→98 % Disc.	98/99 Cost
OC-12 ATM Port Cost	45	$12,097	$43	$546,294	40%	$327,777
OC-3 ATM Port Cost	155	1,651	23	259,422	25%	194,566
Switched Ethernet Port	67	1,114	23	76,184	25%	57,138
Spare switch (chassis)	1	58,462	0	58,462	30%	40,923
Equipment racks (2 per)	4	1,750	0	7,000	5%	6,650
Switch Mgmt software	1	21,495	0	21,495	20%	17,196
Total						$644,251

5.6.4 Design Evolution

1. Design Vulnerabilities

   The network design outlined above is based on a wide range of assumptions which have proven to be subject to change. Many of these variables could dramatically 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 per ATM switch 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 / or the efficiency of the analysis code likewise determine the bandwidth required to each node.
   - Variations in the size ratio of raw data input from the detectors to the output (calibrated data + DST) from the reconstruction phase determine network bandwidth, buffering, and storage requirements.
   - The chance that the anticipated reduction of hardware prices will not be realized.

2. Incremental improvements

   In the time period between 1996 and 1999, a number of improvements can be anticipated that will change the design before the equipment is actually purchased:

   - Higher utilization of the available raw bandwidth provided by ATM through more efficient drivers, "stripped down" protocol stacks, or improved hardware;
   - More efficient use of available bandwidth due to improvements in the dispatching of jobs across multiple systems;
   - Feedback from advanced modeling of the applications and the network combined to optimize physical connections, virtual circuit connections, etc.

3. Relevant product evolution

   Networking hardware stands to realize substantial improvement in price / performance in the next two years, perhaps more than any other aspect of the RHIC off-line computing system. Several items have the potential to dramatically impact the cost of the network, though this is not a certainty.

   - Switch port density probably will increase at least two fold 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.
   - The introduction of a fast ethernet switch with an OC-12 ATM or Gigabit ethernet interface might allow the use of fast ethernet interfaces to one or more classes of farm machines, thereby realizing a significant reduction in cost. It is the lack of such a switch interface that currently excludes such products from consideration, despite the fact that the bandwidth required in many cases could be provided by switched fast ethernet (100BaseTX).

- It is possible that the evolution of ATM switches could provide a large switch capable of handling all of the required connections for a lower cost. Switches exist now for the telephone / carrier industry, but typically lack support for large numbers of OC-3 and OC-12 ports.
- Other developing network technologies (FibreChannel, Gigabit Ethernet, etc.) might displace ATM as the most cost effective solution which meets the required performance goals.

5.6.5 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 quarter 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 complete system goes on-line in 1999.

Assuming the approval of the basic networking plan, an initiative will be started to draft an RFP for switch vendors to respond to. This effort will have to be repeated at the end of 1997 or beginning of 1998 just prior to the actual purchase.

5.6.6 Installation, Staffing, and Maintenance

- Installation

  Installation of the networking equipment will be in standard equipment racks strategically located among the approximately 180 systems that it will interconnect in the RCF. For all OC-3 UTP connections this effectively means that the switch must be within 90 meters of the systems it serves. The OC-12 systems are not constrained by distance within the RCF, but will be limited by disk / tape subsystems.

  The mapping of individual system connections to a given switch in order to optimize performance is the subject of upcoming modeling by the RHIC off-line computing group. The same models will be scaled up to determine the number of OC-12 trunks required to interconnect any two switches. The result of these activities will determine much of the physical layout of equipment and cabling within the RCF.

  Installation costs for the switches, racks, and patch panels are not included in the estimate above as too much remains to be defined concerning the layout of the RCF and the types and sizes of all the equipment involved.

- Staffing

  A significant portion of the operation and monitoring of the RCF network is expected to be provided by BNL Computing and Communications Division (CCD). CCD has already invested heavily in network management software which can be expanded to handle the additional systems. The projected staffing requirements (FTE) for networking are shown in Table 8 below. These values are averaged over the year and do not reflect the peak number of individuals which may be required for any given activity.

- Maintenance

Table 8: Networking Manpower Estimates

Activity	1997	1998	1999
Design / Project Mgmt.	0.50	0.75	0.25
WAN Testing and Reporting	0.25	0.25	0.25
Configuration and Qualification	0.10	0.50	1.00
Troubleshooting	0.10	0.20	0.50
Network Management	0.10	0.20	0.50
	1.05	1.90	2.50

Annual hardware maintenance costs for the switching equipment is estimated at 13% of the total cost, or approximately \$84k. in 1999. Software maintenance is estimated at 25%, or \$4.3k. A replacement rate of 25% per year for switching hardware totals \$161k per year beginning in 2000.

5.7 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, tapes, and computing nodes.

Entry to the main data center area is currently restricted by a combination lock on the access doors. In addition the RCF area has two separate sets of double doors which could have additional security restrictions such as a lock or secure card 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 known. This and other infrastructure costs are expected to be provided for by Brookhaven.

Table 9: RHIC Computing Facility cost estimate.

	1997	1998	1999	Sum
CRS CPU	$ 172k	$ 474k	$ 1590k	$ 2236k
CRS Disk	$ 8k	$ 21k	$ 66k	$ 95k
MDS	$ 158k	$ 714k	$ 1908k	$ 2780k
CAS	$ 97k	$ 292k	$ 1037k	$ 1426k
GCE	$ 70k	$ 140k	$ 210k	$ 420k
Network		$ 644k		$ 644k
Software, etc.	$ 75k	$ 150k	$ 100k	$ 325k
Totals	$ 580k	$ 2435k	$ 4911k	$ 7926k

Table 10: Accumulated capacity of RHIC Computing Facility at the end of year indicated.

	1997	1998	1999
CRS CPU - kSPECint92	10.5	51.9	252.6
CRS Disk - TBytes	0.08	0.4	2.0
MDS Disk - TBytes	0.23	1.06	22.0
MDS Robotic - TBytes	3.5	33.5	100.0
CAS CPU - kSPECint92	6.3	32.9	166.7
GCE Disk - TBytes	0.2	0.4	1.0

6 Cost and Schedule Summary

6.1 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.

We assume a purchase profile such that approximately the following levels of computing capacity are achieved: 4% in 1997, 20% in 1998 and 100% in 1999. Our 100% solution for 1999 will have 420 kSPECint92 of CPU power, 25 TBytes of disk storage, and 100 TBytes of robotic storage which meet the goals established in Table 4 in the requirements section of this document. The costs by component of the RHIC computing facility are summarized in Table 9 and the accumulated capacity at the end of each year is summarized in Table 10. The column labeled "Sum" indicates the total capital cost over the three year ramp-up of the RHIC Computing Facility.

The total estimated capital cost of the RHIC Computing Facility is $7.93M.

6.2 Technical Support & Operating Cost Profiles

The estimated manpower 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 each of the appropriate subsections of the proposal. The manpower for each

Table 11: Summary of estimated manpower requirements for the RCF

	1997			1998			1999		
	Exp	CCD	RCF	Exp	CCD	RCF	Exp	CCD	RCF
CRS	0	0	1	1	0	1	2	0	2
CAS	0	1	0	0	0	1	1	0	1
MDS	0	0	2	1	1	2	1	1	4
GCE/User Support	1	0	2	1	0	4	2	0	5
DAS	1	0	2	1	1	2	1	1	1
Network Support	0	1	0	0	2	0	0	3	0
Technical Devel.	1	1	1	1	1	1	1	1	1
Hardware Support	0	1	0	0	2	1	0	2	1
Admin. & Manag.	0	0	1	0	0	2	0	0	3

Table 12: Profile of estimated annual facility operating costs.

	FY 1997	FY 1998	FY 1999	FY 2000
Additional Labor	3 FTE's	6 FTE's	10 FTE's	10 FTE's
Additional Labor Costs	$ 113k	$ 338k	$ 600k	$ 750k
Non-labor Operating Costs	$ 87k	$ 195k	$ 425k	$ 610k
Unburdened Total	$ 200k	$ 533k	$ 1025k	$ 1360k
45 % Overhead Burden	$ 90k	$ 240k	$ 461k	$ 612k
Burdened Total	$ 290k	$ 773k	$ 1486k	$ 1972k

year is shown in three columns indicating the contributions from the RHIC Experiments (Exp) and from the BNL Computing & Communications Division (CCD), as well as the manpower directly connected with the RCF. The number of new personnel requested in this proposal is 10, with the other 8 RCF personnel being funded initially from RHIC construction funds and eventually from RHIC operations funds. 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.

The labor and labor costs, including appropriate fringes, addressed here are those not included in the existing plans for RHIC operation. Non-labor operating costs for the facility 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, 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 overhead burdent is estimated to be 45%. These annual facility operating costs are indicated in Table 12.

6.3 Anticipated Out Year Costs and Capacities

Table 13: Expected upgrade/replacement costs for the RCF in the out years.

	2000	2001
CRS CPU	$ 590k	$ 575k
CRS Disk	$ 30k	$ 60k
MDS	$ 619k	$ 619k
CAS	$ 376k	$ 367k
GCE	$ 111k	$ 108k
Network	$ 170k	$ 166k
Software, etc.	$ 100k	$ 100k
Totals	$ 1996k	$ 1995k

Table 14: Expected accumulated capacity of the RCF at the end of each of the indicated out years.

	2000	2001
CRS CPU - kSPECint92	359	491
MDS Disk - TBytes	28	40
MDS Robotic - TBytes	130	245
CAS CPU - kSPECint92	240	305

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 approaches the cost required to purchase new equipment of comparable performance. While beyond the scope of this proposal, based on this type of analysis, it is expected that capital requests for replacement and/or upgrade of facility equipment will be made in FY 2000 and beyond at about the level of 25% of the initial capital investment. As discussed in the Requirement section of this proposal, the capacity levels achieved in 1999 are sufficient to satisfy expected initial year requirements but are significantly short of "nominal" year requirements. By the judicious use of the 25% annual replacement/upgrade requests, assumed here to be $2M per year, one can, as demonstrated in Tables 13 and 14, bring the facility capacity to that required for nominal year running in 2000 and 2001. This is possible for two reasons. First, the price/performance ratio for computing equipment is expected to continue to improve so that money spent in 2000 and 2001 will be more effective in delivering capacity than money spent early in the project. Second, since purchases during the project were deliberately skewed toward the later years, relatively little of the equipment will actually be obsolete during the first two year of the replacement and upgrade, allowing for substantial net increase in the actual installed hardware base. This proposal is thus directed toward the goals of supply the compute capacity for initial year operations and establishing a facility capable of satisfying "nominal" and likely "out" year requirements based on annual capital requests at a level of 25% of the initial capital investment.

6.4 Time Line & Principal Milestones

Figure 7 indicates a schedule for the implementation of the RHIC Computing Facility. The figure shows the expected 4% - 20% - 100% ramp-up of the facility and important decision points in some of the subsystems. The time scale of the figure is in fiscal years.

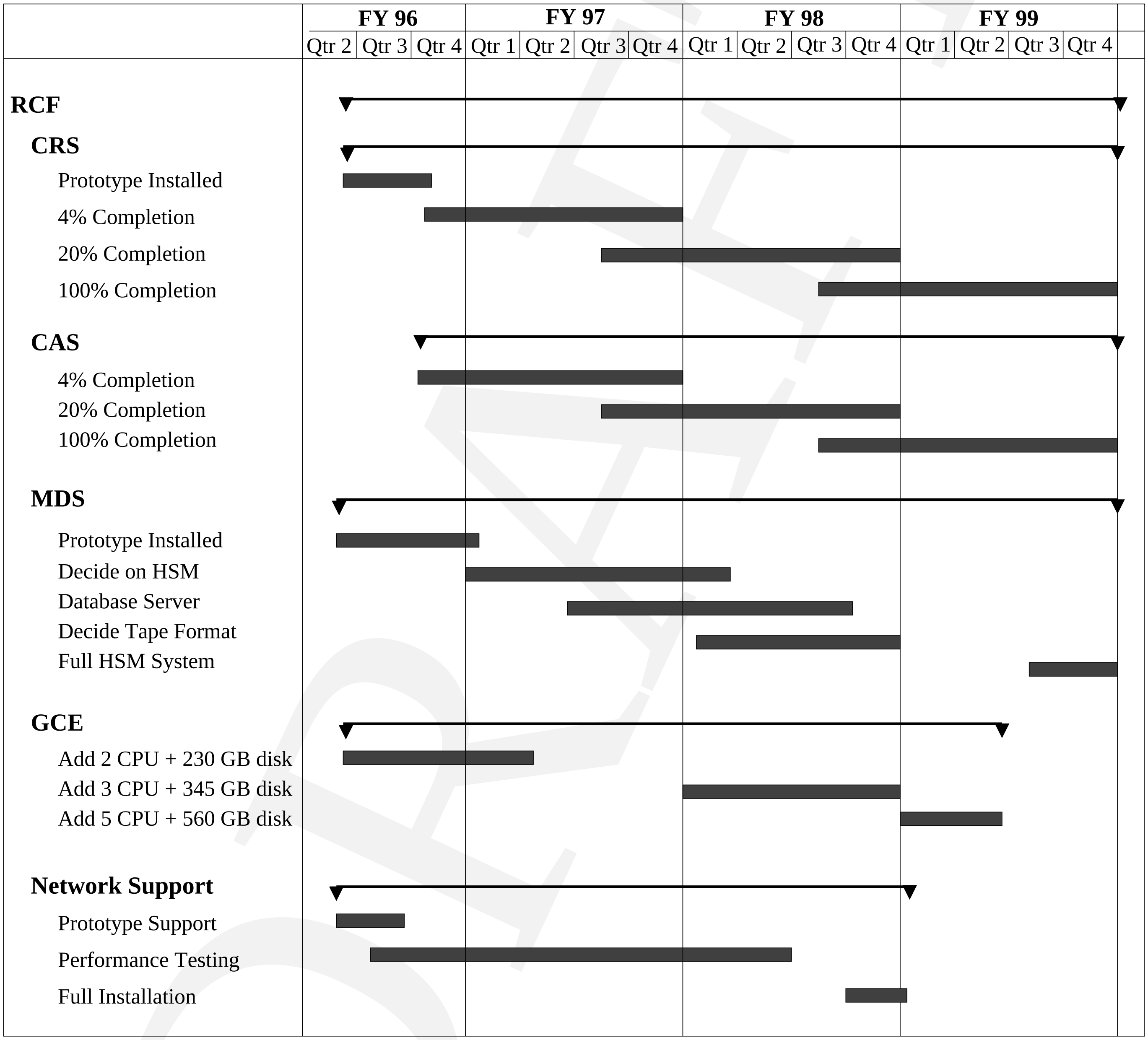

Figure 7: Proposed schedule (in Fiscal Years) for the implementation of the RHIC computing facility.

References

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

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

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

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 15: 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 16: 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 17: 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, the 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 30-40 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 1998, we already expect to have access to some fraction of a 10-15 kSPECint92 farm with about 1Tb of robot space 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.

At this point, there are only two outstanding questions: DST access and networking. If the full beam luminosity is reached in the year 2000 and PHOBOS runs efficiently, we don't have a clear plan for achieving fast access to our full DST sample to generate μDSTs since only a fraction of year's worth is expected to fit in our share of the RCF tape robot, which we assume to be 15-20%. If our share is smaller, it only exacerbates the problem. Also, as mentioned in the main body of the RHIC proposal, 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 areas of any concern are networking to Krakow and access to DSTs at the central facility. These problems should be manageable.

E Glossary of Terms

μDST micro Data Summary Tape produced from a DST.

100BaseTX 100 Megabit Ethernet.

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

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

CLAS CEBAF Large Acceptance Spectrometer

CPU Central Processing Unit

CRS Central Reconstruction Server

DAQ Data AcQuisition.

DAS Data Access Software

DBMS Data Base Management System

DFS Distributed File System

DLT Digital Linear Tape

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

MDS Managed Data Server

MFLOPS One Million FLoating point Operations Per Second

MHz Megahertz

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

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

PAC Program Advisory Committee

PASS Petabyte Access and Storage Solutions

PC Personal Computer

PDSF Particle Detector Simulation Facility

PDU Power Distribution Unit

PHENIX

PHOBOS Son of MARS

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

WAN Wide Area Network

WWW World Wide Web