Yale 40609-1179

Offline Computing at RHIC

(Report of the RHIC Off-line Computing Committee)

B.S. Kumar[1], Chairman
M.D. Baker[2], H.J. Crawford[3], B.G. Gibbard[4], K. Hagel[5], D.L. Olson[6],
R.L. Ray[7], R. Seto[8], S.P. Sorensen[9], T.G. Throwe[4], G.R. Young[10]

1. Yale University, New Haven, CT
2. Massachussetts Institute of Technology, Cambridge, MA
3. University of California Space Sciences Laboratory, Berkeley, CA
4. Brookhaven National Laboratory, Upton, NY
5. Texas A&M University, College Station, TX
6. Lawrence Berkeley Laboratory, Berkeley, CA
7. University of Texas, Austin, TX
8. University of California, Riverside, CA
9. University of Tennessee, Knoxville, TN
10. Oak Ridge National Laboratory, Oak Ridge, TN

(February 14, 1996)

Abstract

We describe the challenges associated with the storage, handling, and analysis of data related to experiments at RHIC. We provide recommendations on how these challenges can be met in a timely manner.

Contents

1 Introduction

The Relativistic Heavy Ion Collider (RHIC) under construction at the Brookhaven National Laboratory is a colliding-beams machine scheduled for completion in 1999. It will accelerate a variety of particle species ranging in mass from protons to Au nuclei at energies in excess of 100 GeV per nucleon to produce the highest energy-density nuclear collisions ever studied in the laboratory. The strongly-interacting matter created in these collisions will be produced at sufficiently high energy densities that it may undergo a phase-change into a novel state of matter known as a quark-gluon plasma. In such a plasma the constituents of the protons and neutrons, namely the quarks and gluons, become deconfined from their parent nucleons to move about freely throughout the entire volume of the plasma. This state of matter is conjectured to have existed briefly at a time some 10^{-6} seconds after the Big Bang and perhaps to exist today in the cores of neutron stars. RHIC includes an accelerator/storage-ring complex designed to produce collisions that might re-create this novel state of matter, and a set of detectors to measure the final-state of these collisions and deduce the conditions during the collision.

There are presently four heavy ion experiments approved for taking data at RHIC. Each experimental group has designed and is building one of the detectors to be placed around the circumference of the RHIC ring. These experiments complement each other in their degree of emphasis on the measurement of hadrons, leptons, photons and jets. There are a large number of proposed signals for deconfinement and many methods proposed for diagnosing the properties of a quark-gluon plasma. This is in large part driven by the exploratory nature of the physics program at RHIC. The quark-gluon plasma would be a new observation. This necessitates pursuing a broad range of experimental approaches as evidenced by the programs put forth by the four experimental groups.

Heavy ion collisions at RHIC energies are expected to produce in excess of 10000 secondary particles in a given collision. The detectors must accordingly be highly segmented in order to record and identify without undue confusion all the particles entering their aperture. Many of the proposed signals require observing final state particles which are created in as few as 10^{-2} to 10^{-4} of the collisions. Hence, in these cases, many collisions must be examined for each "interesting" one found. Finally, the behavior of any proposed signal must be studied over a large enough statistical sample to establish deviations from "normal" strong-interaction physics. This results in the need to obtain and analyze samples of many millions of events each.

Consequently, the experiments studying heavy ion collisions at RHIC will produce large volumes of data which have to be acquired, stored, analyzed, compared to results of model calculations and reference event classes, and compared to results of simulated data. Accomplishing this will require access to large amounts of computing power, data storage, and data-handling capability. The estimated computing resources vastly exceed those presently available to the relativistic heavy-ion nuclear physics community. Recognizing the magnitude of this problem, RHIC management convened our committee to advise them on how these computing and data-handling resources should be provided and/or obtained. Our charge, an overview of the magnitude of the computing resources needed, and an outline of our proposed solution are described in sections 2-4. Subsequent sections provide more details and supplementary information and set forth a plan for obtaining the needed computing resources.

2 The Charge

The committee was given the following charge:

- To provide an updated estimate for the computing resources that will be needed to reduce and analyze data from RHIC experiments, beginning in the year 1999 when the machine becomes operational.
- To reassess the implementation plan for a RHIC computing facility. The new assessment should specifically include necessary equipment at collaborating institutions, and take account of advances in networking capability. If possible, you should provide an updated model for the RHIC computing facility that can serve as the basis of a technical review in the coming year.

3 Scale of RHIC computational and data storage needs

3.1 History

There have been two previous attempts in 1992 and 1993 [1, 2] to assess the computational needs at RHIC. Our report extends substantially on these efforts, and benefits immensely from being able to work with more mature and stable detector designs, and improved data analysis and simulation codes. Also, as will become evident to a reader of all three reports, we have taken into account several items that were not considered in the previous studies. Hence, our overall estimate for computing needs differs from those mentioned in previous reports. As an example, there is a trivial factor of 2 difference because our estimates are based on 4000 hrs per year of RHIC operation rather than the 2000 hrs per year assumed in the 1992 report [1]. A brief chronology of RHIC computing estimates is given in Appendix G. It is important to keep in mind that our estimates reflect our present best understanding of what is needed. As analysis and simulation codes undergo modifications, it is natural to expect that the overall estimates of computation and data storage needs will change. With luck they could even come down!

3.2 Computational functions at RHIC

The following computational and data storage functions have to be performed at RHIC. Of the items listed below, only the acquisition and analysis of raw data was discussed in the ROCOCO-1 report [1].

- Acquisition of data.

  This includes raw physics-event data, which is expected to be produced at the rate of 20 MB/s or more by each of at least two of the RHIC experiments, as well as calibration data, geometrical and survey data, run parameters, and storage ring conditions.
- Storage of data.

This includes storage of all the data noted above plus results of the model and simulation calculations listed below plus storage of results of the intermediate and final stages of reduction and analysis of all these various data.

- Theoretical model calculations of nucleus-nucleus collisions.

  Various theoretical models of different aspects of RHIC collisions exist. Nearly all of these involve a Monte-Carlo step to simulate the observed final state of a RHIC collision. Several thousand to several million such model events must be calculated, depending upon the issues under study, to permit statistically significant comparisons to experimental data. These models will undoubtedly get revised as data become available, and continual comparisons between experiments and the models will be necessary.

- Simulations to assist in detector design.

  Simulations have been ongoing for 4 years in support of the present detector designs and are an essential part of evaluating design tradeoffs in the detectors. These simulations involve Monte Carlo calculations, usually utilizing the CERN GEANT toolkit. They use outputs from the above theoretical models as input in some cases and simple particle spectrum generators in others. They are regularly updated as the design and construction of the detectors progresses.

- Simulations to assist in predicting and determining detector performance.

  These are an extension of the previous item and also involve extensive use of Monte Carlo calculations. Measured and simulated single-particle detector responses have to be mixed with those for full events and the resulting "data" analyzed to determine detector geometrical acceptance and efficiencies of particle detection and identification. Other calculations are needed to analyze event reconstruction efficiency and error rates, to study correct association of "hits" from one particle as it traverses the several layers of a detector, and to determine triggering efficiencies for rare and or specialized classes of events.

- Monitoring of detector performance.

  This includes a realtime component, which is partly handled by the online computers used by each detector, plus longer-timescale components as detector performance is followed over weeks and months of operation. The needed calculations must follow trends, update calibration coefficients and databases, and provide output summaries to alert both groups taking live data as well as those analyzing previously recorded data.

- Reconstruction of events.

  This includes reconstruction of both real physics events as well as reconstruction of simulated events. Both track reconstruction and particle identification steps are included here. The input to this stage is raw data measured by the detector and the output is usually 4-momenta for identified particles. This is expected to be one of the largest single consumers of offline computing power at RHIC.

- Detailed analysis of detector performance.

  This requires access to reconstructed events, both real and simulated, plus test beam measurements of detector performance. Iterative calculations are often required. The time dependence of detailed detector performance over a period of months or years must be calculated as a part of this analysis.

Table 1: Total yearly storage requirements for the RHIC detectors.

Data Type	TBytes/yr.
Raw Data	680
Calibrated Data	518
Theoretical models	50
Simulated Data	154
Data Summary Tapes	193
Micro Data Summary Tapes	55
Database Storage	10
Total	1660

- Analysis of physics and simulation data.

  This analysis is the "final" stage in extracting physics information from data. It requires testing of competing algorithms for signal extraction and background suppression, usually involving iterative analysis of a class of events until acceptable behavior is obtained and systematic errors can be quantified. In contrast to the event reconstruction step, this stage involves a large number of independent analyses carried out by small teams of physicists from the experiment working at their home institutions. Event reconstruction usually involves one computational pass through the data set in a near production-mode environment at a large data-processing center which has fast access to the bulk of the experiment's archival storage. The analyses of physics and simulation data are typically the most geographically dispersed of computational activities.

3.3 Computational and Storage needs

The datasets which must be stored to support all RHIC computing activities will increase in size at the rate of ~ 1.5 Petabytes (10^{15} bytes) per year. A summary by type of data is shown in table 1. [1] A breakdown by experiment is shown in Section 6.

The computing needs to support the RHIC program during steady-state collider operation (allowing for the RHIC overall duty factor) are summarized in table 2. These estimates were derived for each of the approved experiments using their current simulation and reconstruction codes. Detailed estimates for each experiment are given in Section 6 and the Appendices. There are several units in which computing power required may be presented. The tracking code of the PHENIX collaboration was tested on several platforms, and CPU performance appeared to scale better with kSPECint92 ($1000 \times$ SPECint92) than GigaFlops (1 Billion floating point operations per second) (see appendix C). Hence our data are presented here in kSPECint92. However, for those choosing to convert between these units, we suggest a conversion factor of 3 kSPECint92 per GFlop.

[1] The numbers in this document are quoted to 2, sometimes 3 significant figures to keep the arithmetic simple and to be consistent with the numbers across the document. However, it bears emphasis that our numbers, overall, are probably correct to a factor of ~ 2.

Table 2: Total sustained CPU needs of the RHIC experiments. Suggested unit conversions are 3 kSPECint92 = 1 GFlop.

Computation Type	kSPECint92
Physics Event Reconstruction	397
Theoretical Models	165
Simulation Reconstruction	202
Analysis of Simulated Data	200
Analysis of Physics Data	290
Total	1254

Addressing the computing and data-storage needs for the RHIC experimental program will require a multipronged approach.

- A dedicated set of computers with about 400 kSPECint92 of processing power will be required to keep up with the event reconstruction needs of the experiments. These must be reconfigurable into groups of varying size as the needs of individual experiments change with time. This is likely the largest single-purpose allocation of computing power required for RHIC.
- A data storage center capable of archiving and serving several Petabytes of data of all types will be needed. The majority of this storage must be available via high-speed links to the reconstruction computers noted above. Smaller centers handling data particularly for the final analysis projects and modeling computations will be needed and might best be located in regional centers. The large data center must provide a graceful method of moving data to progressively more cost-effective (and therefore slower access) storage as it is requested less and less frequently. A modern hierarchical data storage system appears to be the best architecture for this. This data center must also provide high-speed network links to remote sites where both additional results which must be stored are generated as well as analyses requiring access to results stored here are performed. This particular aspect of the RHIC computing center presents the greatest challenges to the current state of the art and accordingly needs the longest lead time.
- Computing power to handle the theoretical model calculations, event simulation in the actual detectors and reconstruction of these simulated events is required. This computing power does not need access to the live data streams from the detectors and thus could be done at a number of computing centers. Much of the initial theoretical modeling and detector simulation can be done in advance of actual data-taking, although these activities necessarily will continue after real data-taking commences and feedback becomes available on the accuracy of model predictions as well as actual detector performance.
- Computing power, local storage and network access for the analysis tasks carried out by physicists working both at BNL and at their home institutions are required. Although it is expected that computing and storage requirements for any given analysis project will be comparatively modest, it is anticipated that

there will be many tens of such projects proceeding in parallel, each requiring its dedicated resources. These analysis stations are also expected to provide much of the interface for individual physicists to the larger facilities noted in the previous items.

4 Proposed solutions

We propose the following specific solutions for RHIC computing. A RHIC computer center should be designed to meet the bulk of the needs listed under the first two items above. Various existing computing centers should be used in additon to provide part of the needed computing power, particularly for the modeling and simulation tasks. Finally, the provision of a complement of desktop workstations should be continued in order to provide partial support of final analysis tasks as well as to provide physicist interfaces to the RHIC Computing Center.

A central facility, the RHIC Computing Center, should be set up at Brookhaven to provide:

- a large amount of computing power (initially 520 kSPECint92) to handle event reconstruction, simulation and analysis tasks.
- a large storage system (initially with 30 Tb on disk and 100 Tb of robotic mass storage) to archive the raw data streams from the experiments as well as the other necessary large data sets such as those for simulated events, reconstructed events, calibration and geometry databases, and to maintain the necessary relationships among them.
- high-speed network connectivity to remote sites as well as to the experiments in the RHIC ring.
- a unified computing environment, software library, and centralized software management.
- a governing structure to ensure computing and data storage resources were allocated to address continuing computing needs of the RHIC experiments.
- a focal point and management structure to pursue long-term evolution of the facility to keep pace with improvements and upgrades to RHIC and its experiments.

This computing power would be purchased in increments of 4%, 16% and 80% during FY97, FY98 and FY99. This provides a compromise between the need to begin installing and gaining experience operating such a farm and the need to delay the bulk purchase as long as consistent with meeting RHIC computing needs in order to take advantage of ever-improving cost/performance relationships in the computer industry. The most demanding aspect of this facility will be the provision of a storage management system capable of serving the very large data sets needed by RHIC. The aggregate data volume added each year will exceed 1 Petabyte, which is significantly larger than the multi-Terabyte datasets presently handled, thus requiring state-of-the-art solutions. It may be prudent to join a consortium working on development of such large storage systems, such as the High Performance Storage System group. The need to have a workable solution to this storage issue provides a key argument in favor of beginning work on the RHIC Computing Center as soon as is practical.

The RHIC Computing Center will need to provide robust Wide Area Network links to all the other laboratories and universities involved in the RHIC program. All users at remote sites will require, at minimum, X-terminal connectivity in order to interact with (control and receive results from) their computing processes. Local area networks at Brookhaven itself will be needed to handle the anticipated groups of user-supplied workstations which are physically located at Brookhaven. Users needing to receive intermediate and final datasets at their own institutions for intensive local analysis projects will require additional WAN bandwidth. It is anticipated that the needed WAN network bandwidth can be supplied by the ESnet for the ESnet sites. A study needs to be undertaken to assess the needs and recommend improvements for many university sites. Dedicated LANs will be constructed at Brookhaven to handle the links from the experiments to the RHIC Computing Center and from the Center to on-site groups of workstations and X-terminals.

The Center would have a Director appointed by RHIC and BNL management and reporting to the head of RHIC. This Director would be assisted by a standing Board drawn from the RHIC user community which would advise on the allocation of the Center's resources among the various RHIC users and which would also act as an advocate for the continued health of and improvements to the capacity of the Center. We believe that the staffing for the Center would need to reach a level of some 34 FTEs by 1999. It would be drawn from RHIC construction, RHIC operation, the BNL Computing and Communications division, and from experimental groups. The personnel from experiments would be experts whose on-shift round-the-clock duties would include ensuring continued operation of data archiving and event reconstruction tasks as well as handling requests from users. A suggested break-down of the personnel by funding source is discussed in the request for additional experimental equipment [3] and in section 9 of this report.

The computing, storage, networking, software, support and management needs of this facility could be met with an initial total capital investment of $8M by the end of FY99. A continuing annual capital investment of 25% of this sum would then ensure a steady increase in computing power and storage capacity of this facility such that over 1000 kSPECint92 (333 GFlops) of computing power, 65 TBytes of disk storage and 300 TBytes of robotic mass storage would be on hand within 2 years of the RHIC turn-on for experiments. A detailed description of the RHIC Computing Center is given in sections 7 and 8 of this Report.

In addition to facilities at the RHIC Computing Center, the committee anticipates a possible need for the usage of existing supercomputer centers to the extent of about 300 - 450 kSPECint92. Theoretical models and event simulations are prime examples of calculations which can be run at such centers, requiring as they do little access to the large raw-data and calibration databases which must be maintained at the RHIC Center. This need would continue throughout the RHIC experimental program. The needed computing power could be obtained partially at each of several centers, with results being moved via ESnet to the RHIC Computing Center for long-term archiving. The NERSC facility at LBNL has expressed a definite desire to support RHIC computing.

We anticipate continued usage of existing desktop workstations by physicists presently carrying out specialized analysis projects at their home institutions. Some fraction of these workstations are expected to migrate to the RHIC site in anticipation of the beginning of experimentation and to remain at the RHIC site in support of on-site users working on both detector monitoring and data analysis. These workstations necessarily must have network connectivity to the RHIC Computing Center plus the supercomputing centers as noted above; it is anticipated that the DOE ESnet will serve the wide-area networking needs. In order that these workstations keep pace with the rest of the RHIC computing effort, regular, incremental upgrades will be required. An aggregate computing power needed from such desktop workstations of some 120 kSPECint92 (40 GFlops) has been identified, spread out over perhaps some 200 workstations. Assuming the regular upgrading of such workstations over a 4-5 year period can continue, as indeed happens presently supported by the various groups' operating contracts, we estimate that the needed workstation computing power will be provided in this manner and would not require any extra-ordinary capital funds.

Finally, we note that collaboration with CEBAF and other nuclear physics centers, supercomputer centers, groups engaged in construction of new high-energy physics detectors such as Babar, and the like, will be a necessary and fruitful part of the computing effort for RHIC. Such collaboration should be encouraged in that it leads to common solutions for mutual problems in the areas of data storage and file-handling, software maintenance and certification, networking, software proper, and management of distributed computing resources. Collaboration between experiments at RHIC in areas of databases, AFS/DFS, model and simulation codes, and other software tools should similarly be encouraged.

5 Recommendations of Report

1. We recommend the creation of a state-of-the-art computing facility at BNL to facilitate the storage, handling, and analysis of data associated with RHIC experiments. We feel strongly that BNL has a responsibility to play a leadership role in the organization and implementation of this facility. We therefore urge BNL and RHIC management to provide manpower now to address technical issues related to RHIC computing needs.

2. We have formulated a computing model which includes a RHIC computing center with three classes of compute servers, and a large high performance hierarchical storage system. Additional resources would be obtained from local computer centers, supercomputer centers, and desktop workstations. We recommend that the storage of raw data and production of Data Summary Tapes be accomplished at the RHIC computer center. The next stages of data analyses (micro data summary tapes, data mining, etc.) should also take place there. A significant amount of computing resources distributed amongst the collaborating institutions should be used for the final (physics) stages of the data analysis and simulations. Simulation, modeling, and unique analysis tasks which are better suited for a super computer facility should be done there.

3. We recognize that data storage and access will be the most challenging aspect of RHIC computing. Hence, the RHIC computing group needs to immediately allocate manpower and develop expertise and implementation plans in the following areas: High Speed Networking, Large Scale Hierachical Data Storage Systems, Scalable CPU servers, and Data Base technologies.

4. We recommend that RHIC computing collaborate with the HPSS or similar projects on data storage issues. The possible cooperation and sharing of manpower with CEBAF on the development of computer farms would be mutually beneficial.

5. We urge BNL and RHIC management to move swiftly to appoint a recognized leader in the computing field as a director for a new RHIC computing facility.

6. We feel that BNL and RHIC management should also form a team to design and implement the facility in consultation with the major RHIC experiments and later, should appoint a board with membership from the RHIC experiments and the heavy-ion community to advise on the allocation of the facility's resources. The director and board should serve as advocates for RHIC computing to ensure its viability and financial health.

6 RHIC Computing Requirements

Our charge consists of two parts:

1. To tabulate the requirements for data storage and physics analysis of the four RHIC experiments.
2. To provide a possible implementation model meeting these requirements.

The requirements themselves are derived with a model in mind, one in which each experiment transports data over a network to a central location and onto a large data buffer volume. From there it is simultaneously stored on non-volatile media and analyzed for elementary physics quantities such as single-particle momenta and particle identification, an analysis which distills the raw data down into data-summary-tape (DST) format. These DSTs are then interrogated many times to abstract various physics- or technology- enriched data sets, a process referred to as data mining which results in the production of micro-DSTs (μDST). Final physics analysis is performed on these μDSTs. Given this basic model, we can define the requirements in terms of various technologies : processing power, storage technologies, network technologies, and software.

6.1 Processing Power

We first investigated how to measure processing power requirements, and then each experiment made estimates of its needs as detailed in Appendices B through E. These estimates were based on their currently available analysis and simulation codes, and on extrapolations from similar data sets taken by other experiments. The processing problem was broken down into three pieces:

- The processing required to keep up with the incoming data rate in reconstructing events to produce physics quantities from raw data streams. We expect on average to look at raw events only once and to produce DSTs in the process.
- The processing or data mining required to filter the data into physics topic subsets. We expect to look through DSTs many times to find events having particular trigger requirements or other software selectable characteristics, producing μDSTs in the process.
- The processing power required in final physics analysis. We assume results are obtained by perfecting code on some subset of the data and then processing a complete event set μDST, a process entailing many passes through the data and many comparisons with simulation results.
- The processing estimates take into account our expectation that the computer systems operate continuously year round while the data taking operates fewer hours per year (4000 hours/year).

We assumed a similar processing path for simulation data, with the added requirements of "event generation" and, for a limited subset, "hit generation", functions fulfilled by the detectors for "raw data" sets. We expect to have to analyze at least as many simulated events as real data events, because our physics conclusions will be based on detailed model calculations. However, many of the simulations can forgo

Table 3: Total CPU needs of the four RHIC experiments in units of kSPECint92. The numbers in brackets have larger uncertainties associated with them than the other numbers. Suggested conversion factor 3 kSPECint92 = 1 GFlop.

	Brahms	Phenix	Phobos	Star	Total
	kSPECint92				
Event Reconstruction	18	175	120	84	397
Models	Shared by experiments				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

detector response simulation, since these can be obtained and integrated into physics analysis as acceptance corrections.

Processing power requirements for the event reconstruction (DST production), event generation (models) and the analysis of simulation and physics data are shown in Table 3. The units chosen are kSPECint92, since these were shown to correlate best the advertised CPU power and the measured event analysis done on a variety of platforms. Processing power was considered (though not fully studied) in many forms including MFLOPS, SPECINT, SPECFP, SPECBASE, CERN performance units, and clock speed.

This began as a simple exercise - run code on two platforms and then compare their advertized chracteristics. However, it remains unclear how best to quantify our real needs, given dependence on memory, clock speed, cache size and code. This is an area that requires continuing study, with the goal being to maximize compute power without undue constraints on the software. It bears emphasis that the optimal solution may vary by experiment. We note the many vagaries associated with this comparison and claim the numbers are known to perhaps a factor of two uncertainty. This has some unpleasant consequences in costing algorithms shown later. Nonetheless, this gives us some framework in which to assess needs.

The DST production is viewed as the end of a fire hose - raw data is spooling into the DST production farm at a constant rate of 10-20 MB/s per RHIC detector, 10 months of the year, and DST production must essentially keep up in an equilibrium situation. In our model, this reconstruction "farm" must involve dedicated processing power, relatively free of the vagaries of load variation since the detectors will be producing their data stream at a relatively constant rate whenever RHIC is operating. Note that for most experiments, this rate is presently throttled by final level trigger requirements. DST production is expected to be a real "farm" environment, with trivially parallel processing (one CPU per event). Code is expected to be stable once it is developed on local workstations, primarily by physicists from each of the collaborations. Farm CPUs may be tailored in memory configuration to a particular task, but we expect the individual farm CPUs to look sufficiently like work stations that code porting is nearly transparent.

- Requirement : The DST production farm must maintain equilibrium with an incoming data stream of 20 MB/s per RHIC detector. This is expected to take 400 kSPECint92 of computing power to reduce the raw data to DST format in equilibrium.
- Requirement: The DST farm must accept workstation-developed code in a transparent fashion.
- Requirement: DST production code must be relatively stable in time.
- Requirement: DST output must be available to online processes for diagnostic purposes

A different kind of computation is required for μDST production. This is seen primarily as a data filtering task ("data mining") and we view this as an area requiring active research right now, perhaps in concert with our colleagues at CEBAF and FNAL. Individual users or groups interested in a specific physics topic create lists of selection criteria that require minimal computation but specify which events from a data set are to be culled for further analysis. Such sets may be requested infrequently by each user, but each experiment has many users, and we anticipate many μDSTs generated each year. The frequency with which the DST sets are interrogated sets stringent network and parallelism requirements for μDST production. The μDST event sets may be distributed to local analysis centers for intensive analysis while developing code or concepts, and the mining farm must keep track of what data sets have been generated and where they have gone. Filter algorithms may be simple or complex, but we expect this stage of the process to be very I/O intensive compared to the CPU intensive DST production and mixed physics analysis.

- Requirement: Filter must be able to cull μDST from 100 TB DST data set in $\approx$ 24 hours.
- Requirement: Filter must simultaneously process μDST generation requests from many users.
- Requirement: Filter must keep track of μDST contents and distribution.
- Requirement: μDST output must be available to online processors for diagnostics.

Physics analysis is expected to be performed on μDST data sets which may range in size from a few GB to a few TB. This may be done in part locally on workstations, and in part at centers on farms. Certain analysis tasks may be well suited to super computer centers, while others are suited to individual workstations or local clusters. Almost all will benefit from availability of a dedicated farm where tasks can be run for final analysis, but we do not overlook the real resource of local workstations at participating institutions, especially for code development and physics analysis.

- Requirement: Physics analysis machines must be able to analyze 10TB size μDST data sets more than once.
- Requirement: Physics analysis development must be distributed among the participating physics institutions.
- Corollary: The community must command sufficient processing power to perform analysis code development, beginning now and continuing through the life of RHIC.

We note that a large fraction of the RHIC compute load is in simulations. This is not just for detector response, but also for physics interpretation. Our analysis model involves detailed comparison between data and model calculations in which the models and parameters are subject to large variations until we better understand our physics signatures. The simulation tasks do not require the same dedication that the DST production does, and this affords us more flexibility in resource selection. All groups have real simulation needs right now, to help tune detector designs. We expect to be producing massive simulations starting in about 1998 when detector testing and calibration begin. We are presently working with supercomputer centers at LBNL, ORNL, and LANL to see how their hardware might be better tailored to meet our computing needs and to see how our software might be tailored to fit their hardware architectures. We encourage this activity as well as further exploration of supercomputer time allocation algorithms better suited to our needs.

- Requirement: The RHIC community requires sufficient processing power beginning in 1996 to finalize detector designs and begin physics analysis.
- Requirement: The RHIC community requires sufficient compute power in 1998 to begin model comparison in detail and to establish methods for handling and cataloging these simulated data sets. This is estimated to be $\geq$ 120 kSPECint92.
- Requirement: Provide $\approx$ 600 kSPECint92 of processing power for simulation, data generation, and analysis by 1999.

6.2 Data Storage and Retrieval

These represent real challenges for RHIC computing. RHIC data sets take us into the relatively unexplored areas of PB (peta-byte or 10^{15} bytes) data sets and the problem of mining such sets for information. We have quantified the data sets in Table 4. These are justified in detail in the relevant detector appendices. Note that the raw data sets are expected to be analyzed only once, on average, with a small subset being used many times in developing DST code. DST data sets are expected to be formed from the event reconstruction performed on raw data, and are expected to be stored in an easily retrieved form. The DST sets will be read many times as different physics topics are selected from them to form μDSTs. In the detector appendices there are specific lists of topics we expect to explore, with each topic leading to a number of μDSTs being generated. The challenge here is not only the storage and organization of PB size data sets, but the rapid retrieval of selected portions of the data into 0.1-10 TB size data sets.

- Requirement: RHIC must provide for long-term storage and organization for data sets totalling 1.5 PB per year.
- Requirement: Provide long-term storage at an average rate of 20 MB/s for each RHIC experiment. Long-term is expected to mean 5 years for the shelf life of the raw data.
- Requirement: Provide robotic access to DST data sets, expected to be up to 100 TB in size, in less than 24 hours. Integrated over all RHIC experiments, this means a robot capable of $\geq$ 200 TB storage and access.

Table 4: Storage needs for different stages of analysis. The numbers are in Terabytes per year.

	Brahms	Phenix	Phobos	Star	Total
Raw Data	40	350	60	230	680
Calibrated Data	40	(175)	300	3	518
Models					50
Simulated Data	1	150	2	1	154
Data Summary Tape	10	100	60	23	193
μ Data Summary Tape	1	(15)	(13)	26	55
Database					10
Total	103	809	446	302	1660

The committee recognizes that data storage and mining the greatest challenge to be faced by the RHIC experiments. This requirement exceeds the limits of today's storage and network technologies, and we urge RHIC management to hire at least one person to pursue this problem immediately.

We explored several options like HPSS, OSM, etc. We encourage RHIC to establish direct links to these organizations immediately, with at least one expert dedicated to the high volume storage problem.

6.3 Networks

The picture we use has a direct link between each experiment and an intermediate storage buffer, thought to be a 1 TB disk, a multiply-parallel link between the DST robot and an output storage buffer, typically another 1 TB disk, and a multiply-parallel link between physics analysis sites and the μDST storage area. This leads to three types of networks used in RHIC computing.

- Requirement: Provide optical network connection, with independent backup, capable of 20 MB/s data flow to send raw data from each RHIC experimental area to a 1 TB elasticity data buffer which is simultaneously mounted on the DST production farm and on the data taping station.
- Requirement: Provide network capable of 1 GB/s throughput in a data mining operation (100 TB in 24 hours). This is assumed to be a set of processors acting in parallel on the DST data set being mined. This is an item that requires extensive research and development starting now.
- Requirement: Provide network capable of 10 MB/s operation for each of 10 or more workstations requesting access to μDST data sets. This network is necessary whether wide area access is via X-windows or through direct data copy to local storage. If the access is via X-windows, then provision must be made for perhaps 100 simultaneous users, each occupying 0.1 MB/s in a typical graphical analysis application.

In this model, data flows from each experiment onto a high-speed disk which acts as a buffer storage area. One process connected via a second port is the taping station. Equilibrium rates of 20 MB/s require complex solutions with todays technology, but are expected to be relatively routine by 1999, whether through parallelism or new technologies. A third process connected to the buffer is the DST production farm, presumed to be a farm of sufficient processing power linked on an ATM network. This also requires immediate attention to develop a realistic solution to the I/O requirements of DST production.

The DST files are stored on some medium, presumed to be magnetic tape in this model, whose elements are robotically stored and retrieved so that whole data sets can be interogated and subsets selected in less than one day.

- Requirement: Provide network that can stage 10 TB of data in less than 24 hours, making it available to an array of processors which select data and send it to another storage area to form μDST files.

The μDST data sets are expected to range in size from a few GB to a few TB. Because of the expected trigger selectivity of the different experiments, each will run a variety of simultaneous triggers leading to DST data sets typically 1-100 TB in size. The DST-to-μDST process should take less than a day, since there will be many scientists working on different problems simultaneously and we don't want to wait weeks to obtain data sets for physics analysis. Once the μDST files are available, they are presumably shipped to local storage for access by physics analysis machines. Such machines must have access at a reasoable rate, downloading a μDST in $\leq$ 24 hours.

- Requirement: Provide distribution of μDST data to physics analysis machines at rates to 1 TB/day or 10 MB/s. for each RHIC experiment.
- Requirement: Provide storage for μDST data sets up to 1 TB in size accessible to physics analysis machines. Note that this storage may be at distributed physics analysis centers in our model.
- Requirement: Information from physics analysis must be available to online processes for diagnostics.
- Requirement: Provide 10 MB/s network access to Europe and to Japan assuming each of these will have a physics analysis center. This should be funded through the foreign contributions.

Note that the present network situation at BNL consists of a T3 line (45Mb/s) for all BNL I/O. We have been assured that there will be at least an OC3 line (155 Mb/s or $\approx$ 20 MB/s) for RHIC communication to the outside world by 1998. We will require more than this and would like BNL to install a full OC48 capability from RHIC to the outside world. Present ESnet is T3 but there are clear indications that OC3, OC12 or even OC48 (2.4 Gb/s = 300 MB/s) may be available as a backbone by 1999. Thus there is no technological problem to distribute data at the rates we require, although there may be political ones. If ESnet goes to full OC48 operation, RHIC would be asking for less than 1/10 of the full bandwidth to distribute even the raw data to offsite locations, a scenario we have considered as a backup to a central RHIC computing center. It is clear that AT&T, MCI, and SPRINT are all willing to work with us to develop 10 MB/s distribution as a viable option. Pricing on leased lines is unclear because of government regulations and anticipated changes in them, but there are no technological bottlenecks here.

The RHIC experiments have collaborators dispersed across the world. At this time, we do not know what is needed to move desired data from BNL to X, where "X" will be Japan, Russia, Europe, China, India, S. America, Asian Rim, etc., based just on looking at who has signed on already to RHIC experiments. There are several agencies involved, and it is hard to extrapolate into the future. We note however that their ability to access RHIC data will be an important ingredient in the success of the RHIC venture, and this issue will need further scrutiny once a RHIC center is in place.

A related concern is the network connection to Universities. While National Laboratories have high quality network connections, users in Universities typically do not. The funding agencies might need to make some investment in such infrastructure needed to assure such RHIC users access to RHIC data.

6.4 Software

We imagine many classes of software : taping processes, bookkeeping processes, event reconstruction, data mining, network control, physics analysis. Some of these require development, while some require investigation of market solutions.

6.4.1 Operating Systems

- Requirement: The processors analyzing DST data must run operating systems compatible with workstations where code is developed with no recoding.
- Requirement: The processors used for data mining must accept communication from workstations through scripts describing logical options.

6.4.2 Languages

We anticipate that the RHIC data storage and analysis plans will include a mix of computer languages including C, C++, F77, F90, and SQL. Consequently, plans must include compilers for these, but not necessarily local support.

- Requirement: Each experiment will be responsible for defining their own coding standards and language requirements.

6.4.3 Code Development and Management

- Requirement: Each experiment will define its own code develoment and management system.

RHIC is now supporting the Andrew File System as a means of sharing disk-based data files. We expect the RHIC experiments to agree on a single code management system which RHIC will then support for all.

6.4.4 Databases

- Requirement: Each experiment will define its own database system.

There is an existing agreement, with \$40k expended, to use the Oracle software, and it is presently licensed to operate only on the RHIC IBM cluster. We expect each experiment to define requirements for accessing the data stored in this database, and these requirements will drive, in part, the network technologies employed to connect experiments to the database.

6.4.5 Data Visualization

- Requirement: Each experiment will define its own data visualization requirements and solutions.

We expect each experiment to define requirements for data visualization. These can have significant impact on the technologies employed for network between online and offline processes and between physics analysis and the data. If a standard for visualization can be found and agreed to by the various experiments, then RHIC will explore the feasibility of providing support for such a package.

6.4.6 User Interfaces

- Requirement: Each experiment will define its own code interfaces.

We expect the experiments to define requirements for these interfaces. If a single interface can be found, common to all experiments, then RHIC will investigate the feasibility of providing support for such a package.

7 Computing Model

7.1 Overview

Computing systems today generally consist of a collection of network connected servers and desk top client systems. Server systems support a community of users, either individually, such as serving a file to a particular WorkStation, or collectively, such as performing production reconstruction on a data set which will later be of use by an entire collaboration. A desk top system, which may be a WorkStation or a Personal Computer or an X-terminal, is an individual users interface into the computing environment. The model for RHIC computing includes central server facilities located at BNL, user desk top systems distributed both across the BNL site and at remote institutions and server facilities located at remote institutions. The server facilities located at BNL will be operated directly under RHIC management and will be dedicated to meeting the needs of the experiments. The server facilities at remote sites may consist of either small dedicated facilities established by local RHIC groups to facilite their analysis of RHIC data, or a share of a larger facility such as a departmental server at a University or some fraction of a supercomputer facility. Such remote facilities might be used in support of the local RHIC collaborators or to perform functions of general use to a collaboration. The distribution of capacity between the central facility at BNL and remote facilities will depend on relative costs and on the relative ease of management, operation and use.

We show in Figure 1 the schematic of a RHIC computing model. This model includes a central facility at BNL, example remote facilities, and a distribution of desk top systems both at BNL and remote institutions. The facility at BNL includes three logical components, 1) Central Reconstruction Server (CRS), 2) Managed Data Server (MDS), and 3) Central Analysis Server (CAS). These components reside in a single computer room and are connected via a very high performance network. The basic flow of data in this system is as follows. Raw data is shipped directly via network from the experiments to a staging disk associated with the CRS and is then recorded in the MDS. From the MDS, the data is reconstructed and sent back to the MDS for recording. Data in the MDS is accessed by the CAS where higher level analyses are performed producing results which are displayed on, or otherwise accessed by, desk top systems. A likely function of remote facilities might be the generation and reconstruction of simulated events with the resultant data recorded in the central MDS as shown in the schematic. Another activity which might naturally be performed at a remote facility would be compute intensive analyses, such as two and multi-particle correlation analyses, or the analysis of simulated data. Such remote analysis systems would likely periodically draw data samples from the central MDS as indicated in the schematic.

While Figure 1 shows a single system undifferentiated by collaboration, it is anticipated that portions of the system will be partitioned by experiment. The Reconstruction Server will be partitioned with each experiment having a fraction of the system assigned to it for a period probably measured in months. The various partitions of the system may be tuned differently to satisfy particular characteristics of the needs of different experiments. For example, one may require extra memory while another may require higher bandwidth paths to its data sources. Partitioning of the Managed Data Server and Central Analysis Server would also be done where appropriate to reduce

Figure 1: Schematic diagram of the RHIC computing model.

contention, guarantee shares and streamline day-to-day management. Reallocation of resources between experiments and between functions within the system where possible (perhaps analysis CPU's are interchangeable with reconstruction CPU's) will be done periodically based on an evaluation of programmatic need.

7.2 Reconstruction Server

The CRS is sized to keep up with the reconstruction of events as they are taken. There is a direct network connection from the experiments to disk associated with this system, and there is a fully redundant back-up for this connection. This direct transmission of data reduces the need for data tape handling by approximately a factor of two and centralizes data recording activities allowing more efficient use of tape systems and associated expertise and reducing the requirements for spares. The required bandwidth for the sum of all experiments is 50 MBytes/sec. So including the redundant path the actual installed bandwidth will be $\geq$ 100 MBytes/sec. The amount of staging disk assigned is sufficient to hold 24 hours worth of data, allowing for the possibility that the CRS may be unavailable for short periods. Provision is also made for recording data directly into the MDS at full speed to cover the situation where the CRS is down for an extended period. The raw data can then be brought back into the CRS for processing when it is again available. After reconstruction, the results and as much of the raw data as is deemed necessary will be stored in the MDS.

7.3 Managed Data Server

The MDS consists of three levels of storage. The first is an array of disks, the second a robotic tape system, and the third shelf storage for tape. A Hierarchical Storage Management (HSM) System will be used to manage the data in this system. The MDS must be capable, at minimum, of simultaneously accepting data at 50 MBytes/sec from the CRS while sending data at 50 MBytes/sec to the CRS to satisfy the requirement that it be capable of serving as both source and sink of data for CRS operations. The bandwidth for transfer of data from the MDS to the CAS is required to be substantially greater since higher level analyses tend to be I/O limited as compared to reconstruction which is CPU limited. This high transfer rate, perhaps 1-2 GBytes/sec, would be primarily unidirectional and would be accomplished by heavily parallelizing disk reads and network transfers. The I/O rate between the robotic tape system and the disk array, which is required to handle CRS serving operations in parallel with up-dating the analysis cache function of the disk array, will be in the range of 200 MBytes/sec. It has long been recognized that the organization of the data being used in analysis can critically effect the performance of the analysis system. Innovations addressing this problem include Column-wise N-tuples from CERN and the PASS project, originally LBNL based, but now being realized as the CAP project at Fermilab. Innovations of this type will be used as a level of data organization within the MDS beyond that of the HSM system.

7.4 Central Analysis Server

The CAS is required to have a high I/O bandwidth to CPU ratio since many high level analyses read very large volumes of data while doing only short simple computations on any individual piece. The system must support some form of parallel analysis, perhaps one in which many processors with independent paths to many corresponding partitions of the data set can operate independently and combine partial results into a single final result, in the style of Piaf developed at CERN. Results from CAS processing might take the form of displays which can be viewed via X on a desk top system or sent to a printer, neither of which requires particularly high bandwidth connections and so are relatively insensitive to the quality of networking. Alternately, CAS processing might produce reduced size data sets which could be moved in a reasonable amount of time across networks of modest to high quality to remote servers or desk top systems for further analyses.

7.5 Desk Top Systems

There will be a desk top system for each member of a RHIC collaboration. Desk top systems support individual users, at minimum serving as their interface into the RHIC computing system, but frequently also performing a range of additional support functions including Email, document presentation, web browsing, code development, and modest levels of data analysis. They serves not only as their points of access to RHIC computing but as their points of access into the general community of modern science. The time to obsolescence for a desk top system is approximately four years so it is to be expected that a desk top system will, on average, be purchased for each RHIC scientist between now and first results from RHIC. A desk top system may range from an X-terminal or low end PC up to an extremely powerful WorkStation extensively equipped with peripherals. Even for an inexpensive desk top system infrastructure in the form of a network connection and print capability is required. The costs range from ~\$5K for a minimal system to ~\$25k for a moderately well equipped workstation and on up. There will be a spectrum of desk top solutions depending on the needs and interests of the individuals. Individuals involved in software development or highly interactive display work are apt to need moderately powerful desk top systems. The use of the CAS as described above is intended, not only to make data analyses practical which would otherwise take an unacceptably long time for single workstation, but to make analysis opportunities equally available to scientists independent of the power, and therefore price, of their desk top system or the quality of its connectivity, beyond some modest threshold.

7.6 Remote Departmental Servers

It is often the case that there are server systems at remote institutions to which RHIC collaborators have access. These systems are frequently associated with the department or group of which the collaborator is a member and are specifically intended to support the research of the department's faculty and staff. Such systems are usually more powerful than a single workstation and are better endowed with peripherals; disk in particular. These facilities can contribute significantly to the effectiveness of the individuals at that institution and in some cases may be of general use to the RHIC collaboration of which the individual is a member. The most obvious activity for such a system is high level analysis where the data volume has been pruned down to a manageable level and the decisions about what to do next are made by a single individual or small group. The size of the data set which can be so handled is determined by the amount of disk and CPU which is available and the bandwidth of the connection to the MDS by which it can be refreshed. Local facilities with very substantial capacity may be used to do production level CPU intensive analyses such as may be required for two-particle correlation studies. Just how much usage will be made of such remote facilities depends on the capacity of the systems, the ease of use, and the associated costs.

7.7 Supercomputer Centers

Supercomputer centers are clearly remote facilities which need to be considered as possible contributors to the RHIC computing effort. As the definition of supercomputer evolves to include Symmetric Multi-Processors the hardware at such centers can be expected to include systems reasonably efficient in dealing with the codes run by large detectors such as those at RHIC. There are three areas in which one might imagine making large scale use of such centers, if the costing algorithms are attractive. These are areas which make smaller demands in terms of database needs (calibration tables, response and field maps, etc.) than does real event data, and thus could be handled at a center where monster-data-set handling is not necessarily a specialty. The areas are:

1. Large scale production runs of event generator codes (Fritiof, HIJET, VENUS, HIJING, RQMD, and the other usual suspects). Such codes could run on such machines and don't produce nearly the massive data volume/event that a detector does. It would help to quantify the latter to address data storage and transport issues of the output of such work.
2. Large scale production runs of GEANT where one is making a first pass using Event Generator output as the input vectors and calculating hits into detector elements. Whether this includes the detector response step is not clear at this stage, as that depends on how detailed a database is in use to model that. Such use needs to be investigated.
3. Large scale efficiency studies where test particles are mixed back into real events to see if they can be recovered, or, similarly, large scale mapping of detector response, resolution, etc. Again, the volume of data output and database needed to support the job would have to be addressed.

Table 5: Projected costs of hardware components of RHIC computing facility.

	1995 Price	Projected 1999 price
CPU	\$50/SPECint92	\$8/SPECint92
Disk	\$210/GByte	\$34/GByte
Robotic Mass Storage	\$44,000/TByte	\$7,000/TByte

Table 6: Commodity based costing of RHIC computing facility.

	Project Phase				Operating Phase	
	1997	1998	1999	Sum	2000	2001
CPU	\$0.4M	\$1.0M	\$3.3M	\$4.7M	\$1.2M	\$1.0M
Disk	\$0.1M	\$0.3M	\$0.8M	\$1.2	\$0.2M	\$0.4M
Robotic Storage	\$0.07M	\$0.18M	\$0.56M	\$0.81	\$0.3M	\$0.4M
Network				\$0.3M	\$0.08M	\$0.05M
Software, etc.				\$0.9M	\$0.22M	\$0.15M
Totals	\$0.6M	\$1.5M	\$4.7M	\$7.9M	\$2.0M	\$2.0M

8 Costs

The hardware costs associated with the RHIC computer center part of our computing model are discussed here. In order to arrive at a cost estimate for the RHIC computing facility, we assume that over a period of 18 months the cost per SPECint92 of CPU power drops by a factor of two, the cost per GByte of disk storage drops by a factor of two, and the cost per TByte of robotic tape storage drops by a factor of two. These trends have been observed over the past several years, and they are expected to continue. With these assumptions, given the present cost of CPU, disk and robotic storage, we can estimate a cost for these components in 1999. These projected costs are summarized in Table 5.

We assume a purchase profile such that the following levels of computing capacity are achieved: 4% in 1997, 20% in 1998 and 100% in 1999. Our 100% solution for 1999 would have 500 kSPECint92 of CPU power, 30 TBytes of disk storage, and 100 TBytes of robotic storage. The costs by components of the RHIC computing facility are summarized in Table 6 and the accumulated capacity at the end of each of several years is summarized in Table 7. The column labeled "Sum" indicates the total cost over the "Project" phase of the RHIC computing facility. The remaining columns show estimated hardware upgrade/replacement costs during the "Operational" phase and resulting accumulated capacity.

The lines labeled "Network" and "Software, etc." in the "Sum" column of Table 6 are estimates for the networking costs for interconnecting the various components of the RHIC computing facility, and for software and other miscellaneous costs. We estimate that on the order of 100 devices will be networked together at a cost of \$3000 per

Table 7: Accumulated Capacity of RHIC computing facility at the end of year indicated.

	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

device, for a total of \$0.3M. The software and other miscellaneous costs are estimated at \$0.9M. The total estimated project cost of the RHIC computing facility is \$7.9M.

Operating costs for the facility will include maintenance costs, which are typically 10 to 20% of the purchase price, and replacement/upgrade costs. Since the average lifetime of computer equipment is 3 to 4 years, it is expected that 25% of the facility will be replaced or upgraded each year. The last two columns of Tables 6 and 7 then reflect a \$2M per year replacement/upgrade cost for the facility and the associated additional capacity.

Costs for facility infrastructure are assumed to be provided by Brookhaven.

9 Manpower

As we indicated in the interim report, we believe that the present level of staffing and funding for RHIC computing falls far short of what is needed to have a viable computing facility by the time RHIC experiments start running. In this section we present the requirements for staffing a computing center. The following estimates of the technical effort required to design, acquire, configure, and operate the RHIC central computing facilities were developed by listing the various required activities and estimating the amount of effort required for each, based on previous experience. These initial estimates were then adjusted based on information obtained regarding the technical effort required to operate computing centers at other similar facilities, specifically Fermilab and CEBAF.

The technical effort should be provided by a core group of people, employed by the RHIC computer center, augmented by personnel from the Brookhaven Computing and Communication Division (CCD) and from the RHIC experiments. An effort should be made, where possible, to organize these contributions so that they conform to the expertise and interests of the contributing organization, but the director of the RHIC computing facility should coordinate their activities. The tables below outline one plausible scenario for the distribution of technical effort from various sources.

Table 8 outlines, as a function of year, a possible distribution of this central core component of the computing effort across various activities. We recognize in Table 8 that in 1995 four FTE's were needed in order to start building the expertise need to forge a running computer center by 1999, whereas the core group was actually 2.25 FTE. This reinforces the statement that BNL management must provide additional manpower now to address technical issues related to RHIC computing needs. Failure to do so bears the risk that the computing center will be grossly inadequate when RHIC experiments begin in 1999, thus delaying the analysis of physics data.

Table 8: FTE's in RHIC core computing effort.

	1995	1996	1997	1998	1999	2000
System Support	1	1	2	3	6	6
Network Support	0	0	0	0	0	0
Code Development Supp.	1	1	1	1	1	1
Application Supp.	0	0	1	2	2	2
Technical Devel.	1	3	3	4	4	4
Hardware Support	1	1	1	2	2	2
Admin. & Manag.	0	1	1	2	3	3
TOTAL	4	7	9	14	18	18

It is very important for personnel to be hired as early as feasible so that they may gain in technical experience and become experts by the time the majority of the equipment is delivered. Unlike with the hardware purchases, we do not benefit from backloading the funding profile for the FTE's.

We believe that the technical effort within the core component of the RHIC computing group needs to be augmented. Under the current scenario, RHIC construction

Table 9: RHIC core computing FTE's by funding source.

	1995	1996	1997	1998	1999	2000
TOTAL	4	7	9	14	18	18
RHIC Construction	3	4	5	3	0	0
RHIC Operations	0	0	0	5	8	8
Additional Personnel	1	3	4	6	10	10

Table 10: Non-core FTE's contributed by BNL Computing & Communication Division.

	1995	1996	1997	1998	1999	2000
System Support	0	0	1	1	1	1
Network Support	1	1	1	2	3	3
Code Development Supp.	0	0	0	0	0	0
Application Supp.	0	0	1	1	1	1
Technical Devel.	0	1	1	1	1	1
Hardware Support	0	1	1	2	2	2
Admin. & Manag.	0	0	0	0	0	0
TOTAL	1	3	5	7	8	8

would be the primary funding source for most of the core computing effort in the beginning years. RHIC operations would fund the core computing effort beginning in 1998. It is essential to find funds for additional personnel, either from existing budgets or from additional sources. The number of new personnel needs to ramp up to a steady state value of 10, as indicated in Table 9.

We recognize that there already exists in the BNL-CCD some expertise that could be used effectively by RHIC computing. We propose that the CCD manpower be allocated as shown in Table 10.

Personnel contributed by the various RHIC experiments would make up the remainder of the personnel who would staff the RHIC computing center as shown in Table 11. It is reasonable to expect that these individuals would provide support that would overlap with the needs of their particular experiment. For example we note in Table 11 that these individuals would contribute support in code development, applications support and technical development. These items are particularly well suited to the interests of the experiments. These individuals would also form a valuable liaison between the computer center and the experiments. About half of the personnel contributed by the experiments would also provide system support in the later years leading to RHIC turnon.

It should be emphasized that the details of the preceding discussion were meant to be used as an example of how the RHIC computer center might be staffed. The non-core contributions will have to be organized according to the expertise and interests of the contributing organizations and personnel involved as well as the needs of the

Table 11: Additional FTE's contributed by RHIC experiments.

	1995	1996	1997	1998	1999	2000
System Support	0	0	0	2	4	4
Network Support	0	0	0	0	0	0
Code Development Supp.	1	1	1	1	1	1
Application Supp.	0	1	1	1	2	2
Technical Devel.	0	1	1	1	1	1
Hardware Support	0	0	0	0	0	0
Admin. & Manag.	0	0	0	0	0	0
TOTAL	1	3	3	5	8	8

overall project. An overriding concern of our committee is the immediate need to get manpower committed to the RHIC computing center.

10 The RHIC computing center

10.1 Overview

The RHIC Computer Center is envisioned as a central facility located at BNL to serve the computing needs of RHIC users. Its primary purpose will be to provide the necessary computing power and services to do the reconstruction and storage of raw data coming from the various RHIC experiments. It must be emphasized that, in contrast to other large computing facilities, one of the major tasks of the RHICCC will be to handle the large volume of data ($\sim$ 1 PByte/year) being produced and processed by the RHIC experiments. Some fraction, perhaps one-half of the community will use the RHICCC as their primary computing tool in doing data analysis as well.

The development of the RHIC computing capability involves a number of distinct but overlapping activities. One set of activities has to do with the computing facility proper. For each of the major subsystems of this facility there will be three phases. Phase 1 includes information gathering and evaluation, acquisition and evaluation of models and prototypes and the selection of a specific technology. Phase 2 includes the development and implementation of a plan to do a phased acquisition and installation of the subsystem based on the chosen technology. Phase 3 is the operation of the subsystem. Another major activity is the development of the technical group responsible for establishing and operating the facility. This involves identifying the expertise required as a function of time and then recruiting, borrowing, and training people so that the needs are met.

10.2 Schedule

10.2.1 Facility Development

In those cases where it is possible, the facility should be ramped up over a period of three years, with the installed capacity increasing from about 4% in the first year, to about 20% in the second and up to 100% in the third year, the year of initial RHIC operation. This ramp-up allows adjustments to be made in the details of the configuration and style of its operation while the system is still of a manageable size and the financial investment does not preclude serious reevaluation. It also serves as a compromise between the need to make increasingly realistic projections of the performance of the final system and the desire to exploit technology advances by purchasing much of the system as late as possible. Technical considerations, for example the fact that some subsystems such as the robotic portion of the data server, may be composed of only a few very large components may make this approach impractical in some instances. Funding considerations may also distort this strategy of geometric ramp-up, but where possible it is clearly desirable. In Figure 2 a possible schedule is shown for some of the principle components of the facility. Note that, while there will be significant prototyping done as part of the technology choice phase for all items, the only prototype operation specifically shown is that of a robotic hardware system. It is anticipated that the buy-in costs for this system may be large and therefore not practical on an early 4% scale. If this is the case, the need for operational experience with the final software combined with the needs of the experiments for substantial storage well before turn on is likely to require a significant prototype system in the indicated time frame.

Figure 2: Proposed schedule for the implementation of the RHIC computing facility.

10.2.2 Technical Group Development

The other major aspect of the schedule has to do with developing the technical group to acquire, integrate and operate the facility. As discussed elsewhere, the size of this group is significant and its ramp-up must be keyed to detailed technical needs as a function of time. Early on when the dominant activities are planning, designing, and evaluation, there will be the need for creative individuals with the highest levels of experience and technical expertise. Later, as there begins to be a substantial installed system to operate, the focus will move to individuals who can be relied on to maintain a highly stable operation for extended periods. While it is likely that a good deal of expertise will be developed "on the job" and that in a technology which is advancing very rapidly this is often the only available course, efforts should be made to hire key people with appropriate experience in areas of particular importance.

10.3 Management

It is imperative that the RHIC computing center be established and an appropriate head be appointed in the near future. The initial purchase of systems must be made in early 1997. Many decisions must be made prior to this which will affect the direction of the facility. This position must be filled by someone with appropriate standing in the physics community since his/her task will be to work together with the users, the RHIC and BNL management, and the DOE to assemble the necessary manpower and resources to carry out the mission of the center. He/She must be capable of attracting a dedicated, talented and responsible staff. He/She will be responsible for working with his/her staff, vendors, other super-computer centers, and the RHIC user community to make the decisions for the best purchases of a great deal of equipment that must work for the RHIC experiments to function.

Although the exact management scheme of the RHICCC is not specified in this report, it is envisioned that the head of the RHICCC would report to the management of the RHIC project. In addition, since it is the RHIC users who must set the priorities of the center, the RHIC management should also appoint an advisory board for the RHICCC with members drawn from the RHIC experiments. This group may also include members from the heavy-ion community at large, and experts from outside the heavy-ion community.

This board would have several responsibilities:

- to advise the head of the RHICCC on policies so that resources are allocated equitably among the experiments.
- to advise the head of the RHICCC on short term needs and problems.
- to advise the head of the RHICCC on the long term computing needs of the experiments to plan for the needed upgrades.
- to act as a liaison to the various experiments in negotiating the contributions of the groups to RHIC off-line computing.

In addition, there should be reviews of the RHIC computing effort by an external committee which includes members of the heavy-ion community and outside experts. This committee could review policy, resource allocation, and other issues, and provide a report and recommendations to BNL and RHIC management.

10.4 User Services

An additional task falling on the RHICCC will be the support of users stationed at BNL. We envision that the large experiments will have about one-third of the collaboration in residence at BNL. This means that about 200 additional physicists will be at BNL during data taking all needing access to desk space, terminals, printers etc. We envision that the RHIC and BNL management will provide space for these physicists through each of the experiments. Each of the home institutions will provide terminals and workstations for their on-site physicists. This will provide some of the compute power for DST analysis, however the RHICCC should coordinate with BNL management to ensure the necessary support for networking, software and hardware maintenance for all of the equipment used by these people. This is a similar arrangement to the one used by Fermilab and CERN. The RHICCC also should be ready to provide documentation and software assistance to users in the manner of a help desk for standard packages supported by the center, e.g. the progeny of HBOOK, PAW, TeX, etc.

The RHICCC must be a professionally run operation. It must be available 24 hours a day, and have enough fault tolerance so that the experiments which are dependent on the center for data taking will not experience significant down time because of malfunctioning equipment or software. A major portion of the nuclear physics community will be dependent on this center for access to data and its ability to respond to the needs of its users will be critical to a successful outcome of the RHIC physics program.

10.5 Interaction with BNL Computing and Communications Division

The interaction of the RHICCC with the CCD must be clearly delineated. We presume that several of the CCD staff will become RHICCC members, and that the RHICCC will be located in the CCD building unless another large area is built or assigned to it. RHIC users will be the major user of computing facilities at BNL and it may be cost effective to have a close relationship between CCD and the RHICCC. It must be emphasized however, that the RHIC user community will need to set the priorities for the RHICCC.

11 Summary and Conclusions

The experiments at RHIC will probe uncharted territory in the studies of nuclear matter under extreme conditions of temperature and pressure, and its expected transition into a quark gluon plasma. The detectors will generate 700 TBytes of raw data per year. An additional 1000 TBytes of data per year will be generated as processed data, and simulations data. The timely analysis of this information will require the implementation of 1250 kSPECint92 of processing power, and the availability of high bandwidth networks to connect between the experiments and the data storage and processing facilities, and with users scattered across the globe.

The storage, handling, and analysis of RHIC data presents a formidable challenge to the RHIC scientific community. Our committee has taken a critical and careful look at the important issues related to the computing challenge, and has provided recommendations for how these can be met. Our thinking has necessarily been influenced by the present best estimates of computing need, and our understanding of technology. Undoubtedly, there will have to be detailed proposals prepared for funding based on the ideas we have presented, the current needs of the experiments, and the technologies that are available at that time. The simultaneous purchase of equipment from more than one competitive vendor is highly recommended. We are of the unanimous opinion that there exists a means by which the RHIC computing needs can be met. The ability of RHIC management to provide the resources needed to analyze RHIC data will depend crucially on the speedy establishment of a RHIC computing center, and on the swift and judicious deployment of able personnel.

12 Acknowledgements

We thank the many persons who provided valuable input to our committee. In particular, we are pleased to acknowledge invaluable discussions with Mike O' Connor, Ken Kliewer, Stu Loken, David Quarrie, Tom Trainor, and Roy Whitney, all of whom made presentations during our meetings. We are also indebted for the hospitality provided by BNL, ORNL, and LBNL, where our committee met. This work was supported in part by the U.S. Department of Energy.

A Text of Interim Report

(Submitted on October 16, 1995)

A.1 Introduction

A RHIC offline computing committee was convened in August, 1995 by Tom Ludlam and charged with providing recommendations on how computing should be done in the RHIC era. The committee has thus far held one phone conference, and three meetings (two at BNL, and one at LBNL). During these meetings, we heard presentations about Data Storage, Networks, and Data Handling. We also heard short reports on how other large experiments are handling their computing problems. The committee has reviewed the present status of RHIC computing, and discussed the creation, management and implementation of a RHIC computing center. Foremost in our concerns is that there needs to be a concerted effort to get the project on the road. We recognize that our computing problems are, as an example, two to three orders of magnitude larger than the problems presently being tackled at CERN. Furhther, our present investment in computing resources and manpower is about an order of magnitude smaller than CERNs! In order to get from where we are to where we want to be in four years, RHIC computing needs people, money, and leadership. Our present recommendations are summarized below.

A.2 The Charge

The committee was charged with the following

- To provide an updated estimate for the computing resources that will be needed to reduce and analyze data from RHIC experiments, beginning in the year 1999 when the machine becomes operational.
- To reassess the implementation plan for a RHIC computing facility. The new assessment should specifically include necessary equipment at collaborating institutions, and take account of advances in networking capability. If possible, you should provide an updated model for the RHIC computing facility that can serve as the basis of a technical review in the coming year.

A.3 Recommendations of Interim Report

1. The committee feels strongly that Brookhaven National Laboratory has a responsibility to play a leadership role in the organization and implementation of a computing facility at Brookhaven to support the storage and analysis of data acquired by the RHIC experiments.
2. The committee recognizes that the present level of staffing for RHIC computing falls far short of what is needed to have even a partial implementation of a RHIC computing facility by the time RHIC experiments start running. It therefore urges BNL and RHIC management to provide manpower now to address technical issues related to RHIC computing needs, as discussed below.
3. The committee urges BNL and RHIC management to move swiftly to to appoint a recognized leader in the computing field as a director for a new RHIC computing facility.

4. BNL and RHIC management should also form a team to design and implement the facility in consultation with the major RHIC experiments and later, should appoint a board with membership from the RHIC experiments and the heavy-ion community to decide on the division and allocation of the facility's resources. The director and board should serve as advocates for RHIC computing to ensure its viability and financial health.
5. The committee has discussed various models for how RHIC data should be stored and analyzed. It is our present opinion that the first stage of data handling (production of Data Summary Tapes) be accomplished at a computing center located in Brookhaven. For most experiments, a large fraction of the next stages of data analyses (micro data summary tapes, data mining, etc.) would also take place at Brookhaven. However, with the current trends in wide area networking, a significant amount of computing resources distributed amongst the collaborating institutions will be used for the final (physics) stages of the data analysis and simulations. Unique analysis tasks which we feel would be better suited for a super computer facility would be discussed and enumerated in the final report.
6. The committee recognizes that data storage and access will be the most challenging aspect of RHIC computing. Hence, the RHIC computing group needs to immediately allocate manpower and develop expertise and implementation plans in the following areas: High Speed Networking, Large Scale Hierachical Data Storage Systems, Scalable CPU servers, and Data Base technologies.

A.4 Items to be discussed in the Final Report

The final report will elaborate on the recommendations of this interim report, and pay particular attention to the following issues.

- What computing resources will be needed to analyze the data?
- When will the resources be needed?
- Short-term implementation strategy for the resources.
- How much will the resources cost?
- How much manpower will be needed, and when?
- Computing models.
- Specific needs of the major RHIC collaborations.
- Long term implementation strategy for the resources.
- Structure and management of a RHIC computing center.
- Computing resources in collaborating institutions and supercomputing centers.

B BRAHMS

B.1 Introduction

BRAHMS is one of four experiments approved to run at RHIC starting in 1999. It will focus on the measurement of particle spectra over a wide rapidity range using small acceptance spectrometers. Its CPU and data storage requirements are the smallest among the four experiments.

B.2 CPU Need for Event Reconstruction

Based on the present non-optimized codes in simulations on an ALPHAstation 400 4/233 a 122 MFLOPS machine, for central events at the 2 degree and 34 degree setting, local tracking requires about 10 sec/event, global tracking requires 2 sec/event and particle identification requires less than 1 sec/event. This leads to 122 MFLOPS * 12 sec = 1.4 GFLOPS-sec. Peripheral events are estimated to take 20% of the CPU time to analyze as the central events. For a ratio of 1:5 central:peripheral events, we have 1.4 GFLOPS-sec*50 + 1.4*.2GFLOPS-sec * 200 where 50 and 200 represent the number of central and peripheral events, respectively, per second. This leads to 126 GFLOPS. We estimate that the code will run a factor of 10 faster when optimization is turned on which leads to 13 GFLOPS. If we agree to be able to analyze one RHIC year of data (4000 hours) in one annual year, we arrive at 6 GFLOPS. For event reconstruction we therefore estimate that we will need 6 GFLOPS (or $\sim$ 20 kSPECint92).

B.3 Data Storage

Given the extensive experience this group has had with similar types of experiments at the AGS, we extrapolate to obtain the requirements we expect for the BRAHMS experiment at RHIC. Experiments E802/E866 had running times of 640 hours with 2000 events being recorded per minute, each of size 10kB/event, giving 1.2 GB/hr which led to a total data set of .8 TB.

The average rates for BRAHMS are expected to be 250 events/sec with average event lengths of about 10 kB. The event lengths will range from 3kB for the most peripheral events to 40kB for the most central events.

If we scale from our previous experience to what we expect for the BRAHMS experiment, we arrive at a factor of 50 which would imply a data set of 40 TB when we assume a RHIC year of 4000 hrs. This is consistent with the above rates for 4000 hours of running time. Our previous experience has been that the data expands by a factor of around 2.5 in the first pass of data "reduction". Given the nature of the TPC's in this experiment, that factor might be somewhat less than 2.5, perhaps even as low as .8 or .9. We should plan for an extra 40 TB of reduced data for a total requirement of 80 TB.

B.3.1 Nearline Storage Media

The state of the art of slow storage media currently appears to be DLT which can store 50GB of data. The experiment would then lead to 800 DLT tapes per year and first pass data reduction would lead to an extra 800-1000 tapes per year. We would envision a robotic system to have that data accessible at a central computing center in the order of an hour. For the later pass data in the form of DST tapes, we anticipate a reduction of around a factor of 10 leading to about 4TB. This data should be kept in nearline storage so that as many physicists as possible can analyze it. As this would reduce the number of DLT tapes to around 80, one could envision that these could be sent to home institutions for analysis there.

B.3.2 Online Storage Media

Once the data were reduced to "Ntuples" in our previous experiments, it occupied around 800 MB of disk space. Scaling this by a factor of 50 for BRAHMS we anticipate an Ntuple data base to be around 40-45 GB. We feel that we therefore need about 75 GB of disk space for the analysis of this experiment.

B.4 Networks

The distribution of data to collaborators (Raw data and DST) is an important concern. As noted above, the bulk of the raw data is expected to stay at RHIC/BNL using the robotics envisioned. The DST/Ntuples will probably be distributed to participating institutions, preferrably through the network. We must note that a good network connection needs to extend to Europe in addition to the U. S.

B.5 Software

We plan to concentrate our programming efforts on the essential physics programming relying on existing packages (or newer if developed by others) for the bulk of the analysis software. We intend to extensively use the Cern library including Paw and Geant. For online bank handling we will examine the following:

- E866 packages YBOS, Analysis_Control,... to be updated for a distributed environment.
- ADAMO and the distributed packages DAD and PINK, interfaces based on TCL/TK. This set of packages were developed for ALEPH and are now in use at HERA.

B.5.1 Operating Systems

UNIX

B.5.2 Languages

Fortran 77, C, C++, Fortran 90

B.5.3 Database Needs

Three calibrations per channel translate into about 60 kBytes/run which scales to about 100-200 MByte/run period. A one GByte database is therefore deemed sufficient. We will use a RHIC supplied database which is presently Oracle.

B.6 Schedule

The computing facilities currently available are very close to adequate for the work we anticipate doing before RHIC turnon in October, 1999, assuming, of course, that the facilities remain available and do not get dominated by increased activities of the other experiments. Also for the first 6 months after the experiments begin, we anticipate being very occupied with the execution of the experiment and the full capability to analyze the data will not be necessary until around 6 months after RHIC turn-on. From our point of view, the equipment should be purchased as late as possible to take advantage of lower prices and increases in technology. We would think, however, that it would be productive to have at least a working prototype of the robotics system in place so that its operation could be understood and that it could be used in a production mode as soon as enough data is taken for that to be necessary. Databases and software should also be available at some level. We feel that mid 1998 would be a good time for these things to occur.

C PHENIX

C.1 Assumptions

The estimates of PHENIX's computing requirements are based on the assumptions listed in the following table:

Assumptions	
Beams	100 GeV Au + 100 GeV Au
Luminosity	$2 \times 10^{26} \mathrm{cm}^{-2}\mathrm{s}^{-1}$
Au+Au reaction cross section	6 barn
Accelerator duty factor	46% (=4000 hours)
Trigger	10% most central collisions
Charged particle multiplicity	Based on "single" HIJET
CPU Hardware utilization factor	90%
Interaction rate	250 Hz
Event size	300 kB
Data recording rate	20 MB/s
Derived Quantities	
Accepted event rate	20 MB/s / 300 kB = 67 event/s

PHENIX is currently assuming a maximal data recording capacity of 20 MBytes/s. Our data rate at 100% trigger efficiency would, however, be 75 MBytes/s. At the nominal luminosity we are therefore not BEAM limited, but instead limited by our data recording capacity. It is therefore likely that the self imposed constraint of 20 MB/s will be increased by a factor of 3-4 in order to match the nominal luminosity. This increase will probably not occur from day one, but over a period of several years as the real beam luminosity approaches the nominal value.

The conversion factor 1 GFlop = 3 kSPECint92 has been used in this appendix unless explicit tests have shown another conversion factor to be more realistic for the particular application.

C.2 PHENIX Computing Model

In order to better understand the discussion below of PHENIX's global computing requirements it is necessary to discuss briefly our computing model shown in Figure 3.

The raw data from the DAQ will be sent through a routing layer (most likely a powerful workstation) directly via a high-speed network to the managed data storage system (MDS) at the RHIC Computing Center (RHICCC). One or several back-up tape drives will be located in the counting house, but will only be used for emergencies or for special needs. The routing layer will also send data to a multitude of monitoring tasks running on workstations in the counting house. These monitoring tasks will use the same software, the PHENIX Event Processor (PEP), as will be used for all off-line tasks. The online system will also utilize the central PHENIX database.

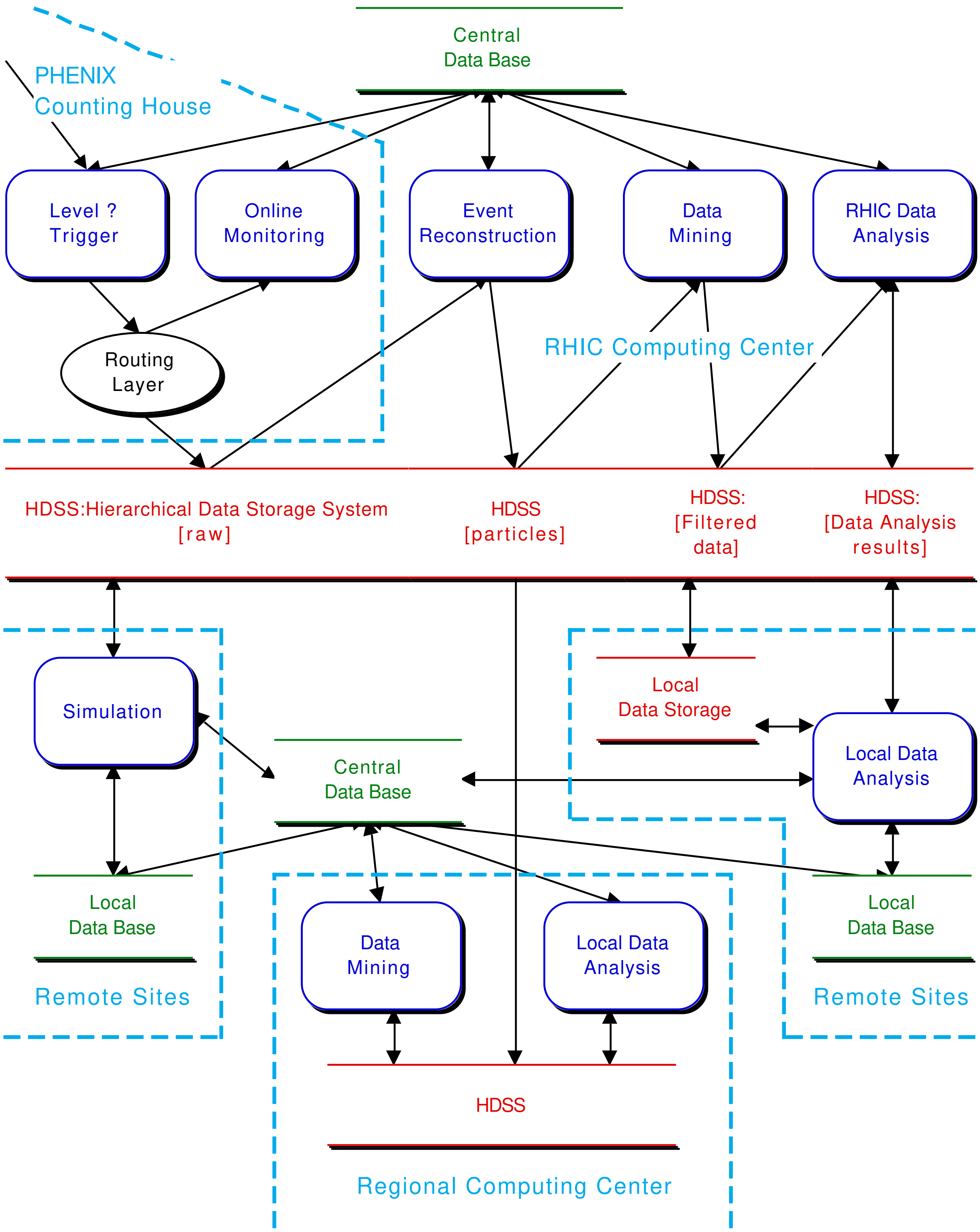

Figure 3: A pictorial view of the PHENIX Computing Model. Please refer to the text for details.

The event reconstruction will be performed on the main PHENIX Event Reconstruction Cluster at RHICCC by reading events from the MDS after calibrations have been performed. It is assumed that after an initial start-up phase the event reconstruction will be done with only a few hours time delay after the data collection. The event reconstruction layer will output calibrated data and higher order objects like hits, track-stubs, tracks, calorimeter clusters, RICH rings, identified particles etc. to the MDS in a format suitable for subsequent data mining.

The data mining stage will permit any user to search through all the data and create datasets suitable for their use. This stage corresponds to the production of μ-DSTs in other collaborations, but we find the data mining idea interesting since it allows each user to basically create their own μ-DSTs. The datasets from the data mining will either be stored in the central MDS or be transferred to remote sites for further analysis.

The final data analysis will be performed both at RHICCC and at remote sites. It is currently assumed that 50-75% of the PHENIX users will use the central I/O intensive cluster for this analysis and will only send X-window traffic over the network.

We assume that the creation of simulated data, both model calculations and detector hit generation, will initially be performed off-site at local institutions or at super-computer centers. At a later stage we hope to move these calculations to the central RHICCC site.

We do not see any need for regional computing centers for PHENIX within the USA, but our Japanese collaborators have expressed a strong interest to create a regional center in Japan. This center will be financed entirely by Japan.

In the context of the computing model discussed in section 7, we envision the hardware behind the functional blocks to vary from block to block and as a function of time. Currently we assume that the event reconstruction cluster and the data analysis cluster will consist of a large number of loosely-coupled workstations and/or PC's, whereas the data mining and routing layer probably will consist of high-end workstations or shared memory machines with high I/O capabilities.

The MDS will consist of one or several large scale robots for the entire RHICCC with several fast tape drives allocated to each collaboration. The size and configuration of the disk farms in front of the robot will be determined by the needs and performance of the MDS for on-line buffering.

C.3 CPU Requirements

In the following paragraphs we discuss the CPU requirements for PHENIX. Table 12 summarizes these requirements.

C.3.1 CPU Requirements for Event Reconstruction

The CPU time for event reconstruction seems to be dominated by the pattern recognition in the central arms. The muon arm reconstruction is less time consuming. Currently three different algorithms have been tested for the central tracking and they differ in CPU consumption by a factor of 2-3 depending on platform. Table 13 shows the platform dependence of the CPU consumption for the code.

The platform dependence seems to scale much better with SPECint92 than with MFlops. The average CPU consumption for the DC & PC tracking is 424 MFlops or

Table 12: Overview of PHENIX estimated requirements for CPU.

	CPU		Location
	GFlops	kSPECint92	
Event Reconstruction	70	175	RHICCC
Data Mining	10	30	RHICCC
Data Analysis	20	60	Primarily RHICCC
Simulation	25	75	RHICCC & Off-site
Models	25	75	Off-site
Total On-site	115	310	RHICCC
Total Off-site	35	105	Off-site
Total	150	415	

Table 13: Platform dependence of CPU consumption for the current PHENIX central tracking algorithm for one central Au+Au event. Platform dependence only shown for tracking through drift- and pad-chambers.

Platform & Rating	CPU time sec	SPECint92 *sec	MFlop *sec
DEC Alpha 200 4/233 158 SPECint92, 45 MFlop	6.45	1,108	291
IBM RS6000 390 114 SPECint92, 55 MFlop	11.7	1,257	605
SGI Indy 4600PC 133 MHz 85 SPECint92, 15 MFlop	13.7	1,163	206
HP 9000/755 99MHz 109 SPECint92, 45 MFlop	13.2	1,440	599
Average		1,242	424
RMS		5%	21%
Including TEC		836	2,100
Including RICH, EMCal, Momentum, PID & Muon Arm		1200	3,000

1,242 SPECint92. When the Time Expansion Chamber is included in the tracking the CPU time increases to 836 MFlops or 2,100 SPECint92. When finally the processing of the Ring-Imaging Cherenkov Counter, the EM calorimeter, the momentum reconstruction, the particle identification is added and the muon arm reconstruction is also added the CPU time increases by approximately 50% to an estimated value of 1,200 MFlops or 3,000 SPECint92 per central Au+Au event.

The estimated uncertainties in these values are a factor of 2-3. If it turns out that the low-mass vector meson physics is very interesting (mass shifts etc.) we will have to do a very thorough job of reconstructing low-p_t tracks. This could easily increase the CPU time by a factor of 2-4. On the other hand, the speeds of the reconstruction algorithms are improving with more efficient coding.

The total annual need for CPU power for event reconstruction can be estimated in two different ways depending on our mode of operation:

Real time operation: Here we will calibrate and reconstruct the data within 2-24 hours after the data is recorded. In this case the duty factor of the RHIC accelerator becomes irrelevant, since the Event Reconstruction Facility should be viewed as part of the detector and should be able continuously to accept the output from the DAQ. The CPU need in this mode is:

event rate * CPU time per event / CPU duty factor = 67 * 1200 / 0.9 = 89 GFlops or 223 kSPECint92.

Data storage mode: Here the raw data is stored on tape (as is usually the case in nuclear physics experiments) and then replayed later during the event reconstruction. In this case we will have to multiply the above estimate with the RHIC duty factor. Annual averaged CPU need: 89 GFlops * 46% = 41 GFlops or 102 kSPECint92. It should, however, be stressed that in this mode the requirements for near-line data storage increases dramatically, since all the raw data needs to get stored in an easily accessible way.

A prudent and realistic estimate of the total CPU requirement for event reconstruction will probably involve an intermediate scenario, where part of the data is processed during accelerator shut-down and another part is reconstructed in real-time. We will also assume that after we gain some more experience in characterizing the events, we might choose not to reconstruct all tracks in each central event, but for many events we will only identify lepton tracks and the nearby hadron tracks. The estimated speed-up from a reduced track reconstruction is probably not more than a factor of 2-3 due to the complexity of the tracks through the PHENIX magnetic field.

PHENIX's requirements for CPU power for event reconstruction is therefore estimated to be approximately 70 GFlops (175 kSPECint92) with an uncertainty of a factor of 2-3.

C.3.2 CPU Requirements for Data Mining and Data Analysis

Following the PHENIX computing model, we will need a substantial amount of CPU power in order to do the data mining and data analysis after the event reconstruction. It is again difficult to estimate the needed CPU power, since we are not in a position now to perform the needed tests of the data mining facility and the data analysis.

For the data mining we will assume that a powerful shared memory machine of ~10 GFlops will be able to serve the 5-10 concurrent users, who are performing data searches. This estimate is based on the experiences with machines currently used by large private corporations doing data mining (Walmart etc.).

The CPU requirements for the actual data analysis, where the results of the data mining is rescanned and investigated numerous times, is not expected to surpass the requirements for event reconstruction. The main emphasis in PHENIX is on physics signals based on leptonic probes and since in general the number of detected leptons per event will be less than 10, the complexity of each event is limited. The most CPU intensive part of the data analysis will probably be the analysis of the hadronic sector, especially HBT and $\phi \rightarrow KK$.

We find, however, that the best way of estimating the requirements for CPU power for data analysis is not to try to estimate the need for each kind of physics individually, but instead to argue as follows: 100 physicists will concurrently perform analyses, each using a state-of-the-art workstation. In year 2000 we estimate such a workstation to be around 200 MFlops. So approximately 20 GFlops should be sufficient to do the data analysis. Furthermore, using in the order of 20% of the total CPU power for data analysis seems to be consistent with the experience of most larger HEP collaborations.

The estimated CPU requirement for data mining and data analysis is around 30 GFlops (90 kSPECint92).

C.3.3 CPU Requirements for Simulation and Model Calculations.

Substantial amounts of CPU power are required for calculation of multi-dimensional acceptances, efficiencies and backgrounds. But since the major thrust of the physics interest in PHENIX is based on event distributions (in contrast to event-by-event analysis) the need for as much simulated data as real data is not obvious. It seems more realistic to assume that the amount of fully simulated events created will be around 20% of the real data. Even with good, fast shower generators the CPU time needed to create an event is similar to the event reconstruction time. The CPU time needed to create and reconstruct the simulated data will therefore be:

$$20\% \times 70 \text{ GFlops} \times 2 \approx 28 \text{ GFlops}$$

Additionally we need to perform a large set of model calculations in order to compare theoretical models to the experimental results. For the hadronic models we can share event libraries with the other RHIC experiments, but for the leptonic signals we need to create our own event libraries. The creation of the hadronic libraries is estimated to require a total of 40 GFlops, so PHENIX's responsibility will be 10-20 GFlops. The leptonic generators are assumed to run much faster than the hadronic, since rescattering is small (this is the main reason for utilizing leptonic probes!), but on the other hand we need a very large sample of events, since their production cross sections are small. An additional 5-10 GFlops should be sufficient to produce the needed simulated leptonic data.

We therefore estimate the total need for CPU power to create and analyze simulated data is around 50 GFlops (150 kSPECint92).

C.3.4 The location of CPU servers.

As indicated in the discussion of PHENIX's computing model it is assumed that the CPU servers for event reconstruction, data mining and the major part of the data

Table 14: Overview of PHENIX estimated annual requirements for data storage and distribution among storage media. It is assumed that a tape vault will be the final storage medium for the data and that data will not accumulate on the disks or in the robots.

Type of data	Annual requirement TBytes	Storage Medium Disk	Robot
Raw data	350	1	12
Reconstructed data	175	1	8
Analysis data sets	100	5	10
Simulated data	150	1	5
Model data	10	1	5
Data base	0.1	0.1	
Total	$\approx$ 800	$\approx$ 10	$\approx$ 40

analysis will be located at RHIC within the RHIC Computing Center. We hope to be able to get access to off-site resources for the simulation needs. We hope to access the large computing resources within the national labs associated with PHENIX.

C.3.5 CPU Server Hardware

We might prefer many CPUs with only 32MBytes or 64MBytes of memory. This might be in contrast to some recent estimates from STAR, where STAR seems to prefer fewer CPU's with larger memory 128-256 MBytes in order to contain the larger event sizes recorded by STAR. We are particular interested in pursuing the use of cheap, commercially available PC type "pizza" boxes, which currently seem to give the best price/performance.

C.4 Data Storage

In the following paragraphs we discuss the data storage requirements for PHENIX. Table 14 summarizes the requirements.

C.4.1 Raw data

The requirement for storage of the raw data is:

$$20 \text{ MBytes/s} \times 3600 \text{ s/hour} \times 4000 \text{ hours} = 288 \text{ TBytes}$$

We would, however, like to store this amount of data in a smarter format that allows easier access (could be an object data base or as DSPACK structures etc). This could result in an increase of 20-25% as additional search indices, headers, meta-file etc, are added.

PHENIX therefore requires $\approx$ 350 TBytes annually for storing raw data.

C.4.2 Data storage after event reconstruction

The output of the event reconstruction pass (in general called DSTs) will contain higher order data structures (hits, track stubs, tracks, clusters, identified particles etc.) The amount of storage needed for these data depends in PHENIX critically on the degree of reduction of the information from the TEC. Assuming, optimistically, that it will be possible to store the calibrated and corrected TEC information in a compact format, we will estimate the DST to be compressed by a factor of 2, resulting in an *additional 175 TBytes annually from the DSTs.*

C.4.3 Data storage for simulation and model calculations

Model calculations will create $\approx 10^8$ events annually. Each event will typically have 10^4 particles and each particle will be characterized by 40 bytes of information. This will require 40 TBytes to store, but this task will be shared by the experiments.

The simulated raw data constitutes 20% of the real raw data. But since the simulated data also contain additional information (track/hit links etc.) the simulated data probably will totally constitute 30% of the real data. Since the real data is around 500 TBytes/year it is estimated that the simulated data will require additional 150 TBytes of data storage annually.

C.4.4 Data storage for data base

We estimate that the on/off-line database will increase by 100-200 GBytes annually. We will require that the database system work together with the data storage system, so all old records, which are not stored on a disk file, will be accessible by the database management system in a robot. Compared to the rest of the data storage needs data base storage requirements are small, but as stressed below, that storage medium for the data base is more critical.

C.4.5 Near-line data storage

For each event we will store of order 10^3 bytes as search objects for the data mining and with a production of $\approx 10^9$ events annually we will need 1 TBytes in disk space annually for these "cardinal" data. These data should always be stored on disk.

The data mining facility will search through the "cardinal" objects and create data sets based on the queries. These data sets will each typically be from 10 GBytes to 10 TBytes, depending of the nature of the query. A full scale HBT analysis or an evaluation of the efficiency of the tracking algorithms will require large data sets, whereas an investigation of the degree of the suppression of the Upsilon might only require a few MBytes. The biggest driver for the storage requirements seems to be low-mass vector meson physics due to the low signal/background ratio. In order to do a reasonable job of analyzing the ϕ or even optimistically the ρ we need in the order of 10^8 events near-line. This will correspond to $10^8 \times 30$ kBytes (hits and higher order objects) = 3 TBytes. We will need this for several projectile-target combinations. Assuming 100 active physicists each working with a average 100 GBytes data set will require 10 TBytes in near-line storage continuously. Each physicist will likely perform 10 data mining operations over a year so the accumulated annual storage could be in the order of 100 TBytes, if we choose to store old data mining results.

In order to have a reasonable buffer between the DAQ and the event reconstruction cluster we need robot-space for one week of raw data equal to $\approx$ 12 TBytes.

The total need for robot space is itemized in table 14. *The requirement of the hierarchal data storage system will be of order 40 TBytes.*

The data in the MDS will, however, not accumulate, but will continuously be updated and the old data will be migrated to shelves.

C.4.6 Data Storage Media

We do not at the moment have any strong preferences concerning the data storage medium for the majority of the event data. Our only requirements are that the robot access time is only in the order of a few minutes to any file located in the robot and that any file can be retrieved within a day from shelves by an operator.

Part of the online database will need RAID discs in order to ensure that critical data, like cold and warm start configurations are always available, so we don't have any DAQ down-time due to disk failure. We don't know how much RAID disk space is needed, but probably not more than 10-20 GBytes.

We also need disk space for the i/o intensive cluster, but the need for this disk space should be determined by the performance requirement for the hierarchical data storage system. It will probably need a disk cache capacity of 10-20% of the total capacity of the robot system.

C.5 Networking

In the ideal world, we would be able to send all of our data over a high speed network to the RHICCC and let the hierarchical data storage system handle the further storage. Since we need to write continuously at 20MBytes/s from day one and the other experiments probably want to do the same, it has to be a network with a very high bandwidth especially since the peak rates might be 2-3 times higher. Since we might also want to record more than 20 MBytes/s as the luminosity increases it seems like we have to require a bandwidth of the network from the counting house to the center of at least 200 MBytes/sec (=1.6 GBaud!).

Between the major functional units in the cluster we probably just need to be able to transfer data at 20-100 MBytes/s.

The connectivity of the CPU's within the RHICCC is another open question. If we are going for a loosely coupled set of workstations we need a very high connectivity. PEP is based on a design philosophy, where many nodes collaborate on the analysis of a single event. We might therefore be interested in more tightly coupled systems, like massive parallel systems or shared memory systems, which have a very high bandwidth built-in.

An ATM network with sustained transfer rate up to 100 MBytes/sec seems to satisfy PHENIX's requirements for the first several years after the start of data taking.

C.6 Functional Requirements

PHENIX has several functional requirements to the RHICCC:

- We need operators, system managers, technicians, etc., to keep all systems and the network operating.

- We need an infrastructure with secretaries, a manager etc.
- We need a building and office space for the PHENIX collaborators, who will be analyzing data on-site.

We need a lot of software and associated expertise in the following areas:

1. Data base system
2. Hierarchical data storage system
3. Advanced graphics systems
4. Freeware packages including CERNLIB
5. Distributed file system (AFS/DFS)
6. Distributed object software (a CORBA implementation)

For these software systems we need on-site and off-site license or discount arrangements (like the recent AFS agreement). We need the associated hardware and we need RHIC people, who can act as consultants and technical experts.

D PHOBOS

D.1 Introduction

This section explains how the numbers reported for PHOBOS computing and data storage needs in Tables 3 and 4 were generated. These estimates are fairly firm for the amount of data storage needed by PHOBOS. The CPU estimates are much rougher because the code is still under development. Furthermore, it is clear that if more CPU power were available, it could easily be put to good use since the following results assume that manpower is put into improving the CPU efficiency of the code which could be used elsewhere.

The numbers are only valid for the nominal beam luminosity and for multiplicities close to HIJET values. Large changes in either of these numbers will yield corresponding changes in the CPU and storage needs, unless the triggering is deliberately changed. PHOBOS data-taking is beam limited at the nominal RHIC luminosity even for a minimum bias trigger.

D.2 Assumptions

Table 15 lists the general set of assumptions that should be common to all experiments at RHIC. Table 16 lists the set of assumptions that are specific to PHOBOS. Two of the these specific assumptions are variable and should be expanded upon. The first is the assumption that interactions occurring farther than 10 cm from the nominal interaction point will be ignored. The motivation for this is that the acceptance of PHOBOS is optimized for vertices near the nominal interaction point and these are the most useful events. The second variable assumption is that by the time that RHIC reaches full luminosity and PHOBOS is fully on line, the experiment will be triggering on the 30% most central events, with only occasional running periods devoted to truly minimum bias data. One of the physics goals of PHOBOS is to measure the centrality dependence of various physics signals. In order to achieve this, PHOBOS is designed to be able to handle the full minimum bias data rate, but the expected limitations in offline computing resources will force a compromise. Events which are 30% central are currently judged to be peripheral enough to use as a baseline for comparison with very central events. This number may change once actual data have been examined.

Table 15: General RHIC assumptions.

Luminosity:	$2 \times 10^{26}/cm^2/s$
Cross-section:	6 barns
Charged dN/dy:	based on HIJET
Running time:	4000 hours / year $(1.4 \times 10^7 s/yr)$
Interaction rate:	$1200 Hz$

Table 16: PHOBOS-specific assumptions.

Vertex position trigger:	$\|Z_{VTX}\| < 10\ cm$ (60%)
Centrality trigger:	30%
PHOBOS event rate:	200 Hz
PHOBOS event size:	18 KBytes/event (for 30% central)

D.3 Needs

Multiplying the event size by the event rate by the running time per year *yields an average raw data recording rate of 60 Tbytes/year.* The amount of storage needed for reconstructed data is a multiple of the raw data. Based on experience it was decided that 5× the raw data was a reasonable amount of space for the reconstructed data. In order to calculate the amount of storage needed for a Data Summary Tape (DST), we assume that the DST consists primarily of particle 4-momenta and charges from the PHOBOS spectrometer and a coarse-grained $dN/d\eta d\phi$ measurement from the multiplicity detector. The size of this DST can easily range from $\frac{1}{2}$ to 2× the original raw data. It is difficult to reduce the DST size much further because the zero-suppressed PHOBOS raw data is a very compact way to store data.

The CPU needs for PHOBOS were determined by measuring the CPU needed to reconstruct a central event using existing PHOBOS code and extrapolating to the full system. The PHOBOS collaboration reports that it takes 60 Alpha × s on average to analyze one central event on an Alpha (3000/700) which is rated at 30 Mflop or roughly 90 SPECint92 units. Equivalent timing results were obtained by using 2 very different track reconstruction codes. One of the codes was also run on the RHIC cluster node rsgi00 and again yielded fairly consistent results. Using this number of 60 Alpha × s, we can calculate the number of Alpha machines that it would take to keep up with the data-taking.

$$\frac{1}{3} \times \frac{2}{3} \times 60 Alpha \cdot s/ev \times 200 ev/s = 2700 Alpha$$

In other words it takes 2700 Alphas to keep up with the data-taking. This is 240 kSPECint92 or roughly 80 Gflops. The factor of 1/3 comes from a projected improvement in the CPU efficiency of the reconstruction code over the next few years and the factor of 2/3 comes from the reduced time for reconstructing collisions which are not quite at $b = 0$ and which are not at $Z_{vtx} = 0$.

We can achieve a further reduction because the average CPU power needed is only 1/2 times the instantaneous value due to the roughly 50% assumed duty factor of the accelerator and experiment. *This yields a total CPU usage need of 120 kSPECint92 or 40 Gflops.*

The above calculation concerned the CPU needs for the production of DSTs. The CPU needs for the analysis of DSTs and for analysis of simulated data were obtained by multiplicative factors, again based on previous experience. We assume that the event-generator modelling will be performed in cooperation with the other RHIC experiments. We assume that the detector simulation and (simulated) event reconstruction will be held to about 25% of the data event reconstruction. We assume that the physics analysis will take about 25% of the CPU of the event reconstruction.

Table 17: PHOBOS computing needs summary

Raw Data Recording rate:	60 TBytes/yr.	(for 30% central)
Reconstructed Data :	300 TBytes/yr.	
DST:	60 TBytes/yr.	
microDST:	12 TBytes/yr.	
Event Reconstruction CPU:	120 kSPECint92	$\sim$ 40 Gflop
Simulation (and Reconstruction):	30 kSPECint92	$\sim$ 10 Gflop
Physics Analysis (Data):	30 kSPECint92	$\sim$ 10 Gflop
Physics Analysis (Sim.):	6 kSPECint92	$\sim$ 2 Gflop

D.4 Summary

The final numbers for PHOBOS CPU and data storage needs for one year of running at nominal luminosity are recorded in table 17. It is clear from this table that PHOBOS, being a high rate experiment, is not "small" when it comes to CPU needs or even storage. The data recording rate discussed here, 200 Hz, corresponds to several thousand raw charged tracks per second. This is comparable to the number of tracks which will be written per second by STAR. We should expect PHOBOS and STAR to require a comparable amount of data storage space for the DSTs and a comparable amount of CPU time for event reconstruction, which is indeed the case, to within a factor of 2–3. Despite this near equality in data taking rate, however, PHOBOS will need fewer total resources than STAR. There are two reasons for this. First, the PHOBOS raw data format is more compact than STAR's, requiring less space/track. Second, STAR will record many more *pairs* of particles than PHOBOS will, which means that post-reconstruction analyses such as HBT or phi-meson reconstruction should require much less total CPU time for PHOBOS.

E STAR

E.1 Introduction

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 cpu 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 (LBNL-EOS, NA36, NA44 and NA49) as much as possible.

The following sections describe the assumptions used in obtaining these estimates, the off-line data processing model, the cpu requirements, the data volume requirements and a summary of the annual computing needs for a robust physics program for STAR. There is also a schedule of requirements for STAR to be able to achieve an efficient operation shortly after the beam comes on. Finally, a description of the expected use of resources at BNL and other sites is described.

E.2 Assumptions

The cpu and data volume requirements for STAR presented in this appendix were based on the following assumptions:

1. The data acquisition rate for full (TPC + SVT + XTPC + TOF + EMC + trigger) Au + Au central events (or equivalent) is 1 per second for 4000 hours of RHIC operation per year for a total of 14.4M events per year. We assume that the CPU and storage requirements for systems other than Au + Au will scale according to the data recording bandwidth limit set at 1 Hz central Au + Au.
2. All raw data events are permanently saved, are read only once, and are analyzed to produce data summary tapes. DST production keeps up with the data event rate on an annual average basis.
3. DSTs are read once for each μDST produced. Several μDSTs are produced for each person doing physics analysis.
4. The raw data event size is 16 MBytes.
5. DSTs are 10% of the raw data size; each μDST is 10% of the DST size.
6. Three analysis passes on the full data (DSTs) are required to produce publishable physics results.

7. Charged particle multiplicity per event is that predicted by FRITIOF for 100 AGeV Au + Au central collisions (or equivalent data size, this is about 3000 charged tracks in STAR's acceptance).
8. Present event reconstruction and detector simulation software (other than event generators and GEANT) will be increased in speed but additional functionality will be added, such that the eventual cpu required will be 1.5 times that of the present software.
9. As many simulated events as real events are reconstructed per year.
10. Physics analysis cpu for simulated events is equal to that for real events. This results primarily from the fact that for relativistic heavy-ion physics studies the connection between experimental observables and the primary physics quantities is available only through theoretical model calculations which are themselves Monte-Carlo computer programs which produce events that must be analyzed in the same manner as the real data.
11. Twice as many event generator and theoretical model events as real events are generated per year. The budgeted CPU for the complex model calculations is taken to be equal to reconstruction of real events.
12. 100 physicists will be doing physics analysis each year.
13. CPU production efficiency is 80% while that for physics analysis is 50%.

Further discussion and justification of these assumptions is included in the following paragraphs.

E.3 Data Processing Model

For the purpose of estimating the computing resources needed for STAR the following model is used (see the dataflow diagrams in Figures 4 and 5). Tables 18 and 19 give the definitions of the processes and data (respectively) for the model.

In brief, the model has two main paths. Raw data is processed through event reconstruction to make DSTs. Some type of query on the DST data is performed in order to generate a micro-DST dataset which is used for physics analysis. The raw data is read once, the DST data is read once per micro-DST and each micro-DST is read many times in order to produce the physics results. There are one to a few micro-DST datasets per physicist doing data analysis. The simulation path starts with event generators, followed by GEANT (called GSTAR) and then either the fast or slow simulators. The output of these simulators is processed through the normal event reconstruction chain. The final-stage physics analysis path for simulation is to read event generator output directly into the physics analysis process.

We envision the following four scenarios for data processing:

1. event reconstruction, DST and μDST production and physics analyses on the real data for 14.4M events per year,
2. event generator - GEANT - fast or slow simulators - reconstruction - physics analyses on many events during the early stages of RHIC operation, before significant amounts of data have been acquired, but relatively few after significant data acquisition,

Table 18: Computational processes in the STAR data flow model.

Process	**Description**
copy on-line to off-line	Move raw data from the on-line cache to the off-line reconstruction farm cache
Copy simulation data to farm	Copy simulation data to reconstruction farm
Event generator	A Monte Carlo process to produce synthetic events of particles suitable for GEANT detector simulation or filtering for physics analysis
event reconstruction	The process of reconstructing the raw (or simulated) data into DST's for physics analysis
Fast sim.	The fast simulator for STAR. Reads GEANT output and generates hits for event reconstruction
generate calibs off-line	Generate calibration data needed to carry out the event reconstruction
generate calibs on-line	This represents the generation of calibration data in the on-line system which is used for off-line data processing
generate final calibs	Generating the final calibrations needed to produce final DST's for physics analysis
GSTAR det. sim.	The STAR GEANT detector simulation program
off-line taper	This copies data from the raw data tapes to the off-line reconstruction cache
physics analysis	The collection of activities required to produce some physics results from DST data
Query to produce micro-DST	The selection of data (events and sub-events) from the DST data to produce a dataset (micro-DST) suitable for a particular physics analysis
Slow sim.	The slow simulator for STAR

3. event generator - GEANT - slow simulator for few (10) tracks mixed with real data, followed by event reconstruction and physics analyses for 14.4M so called "mixed" events per year once significant amounts of data have been taken,
4. event generator - acceptance/efficiency filter - physics analyses on 14.4M events per year to generate statistically significant theoretical model comparisons with data.

Each of these scenarios can be described as following some of the paths in figure 4. For each of these scenarios there is a final stage of analysis that constitutes data mining. This is shown in figure 5. This data mining stage is the most part of the computing problem in high-energy and nuclear physics and is the one which will likely take the most lead time to solve.

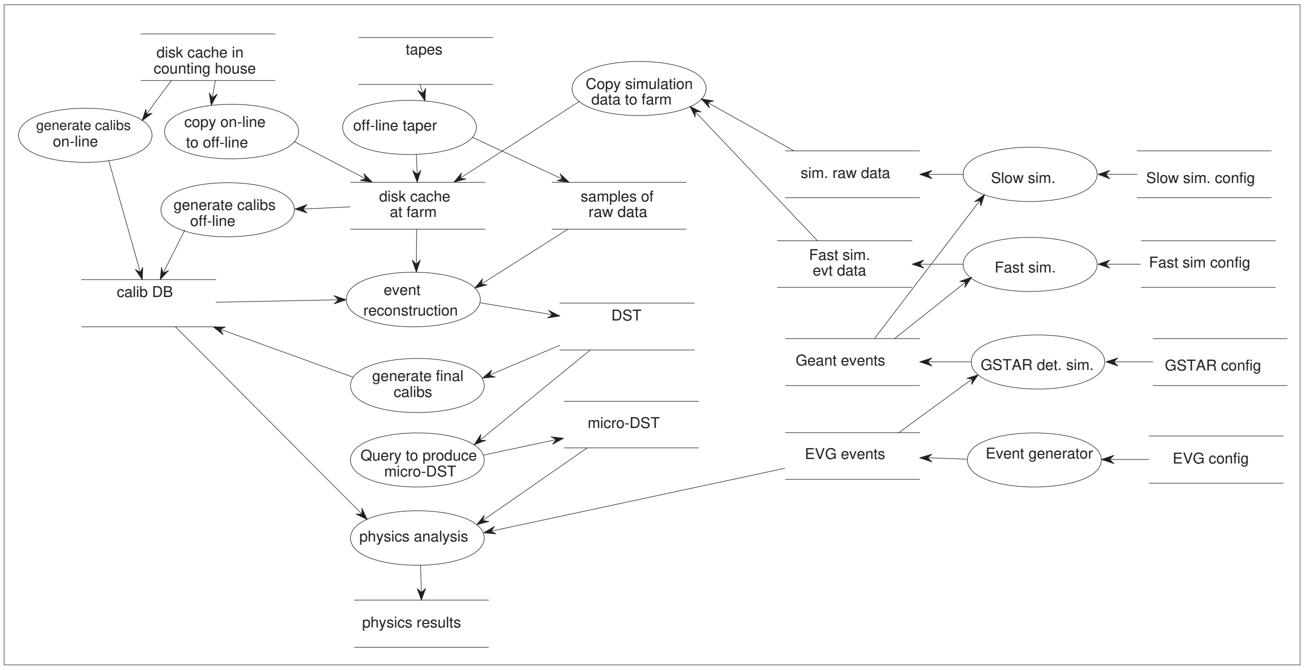

Figure 4: Data flow diagram of STAR Off-line Processing. See itemized descriptions in Tables

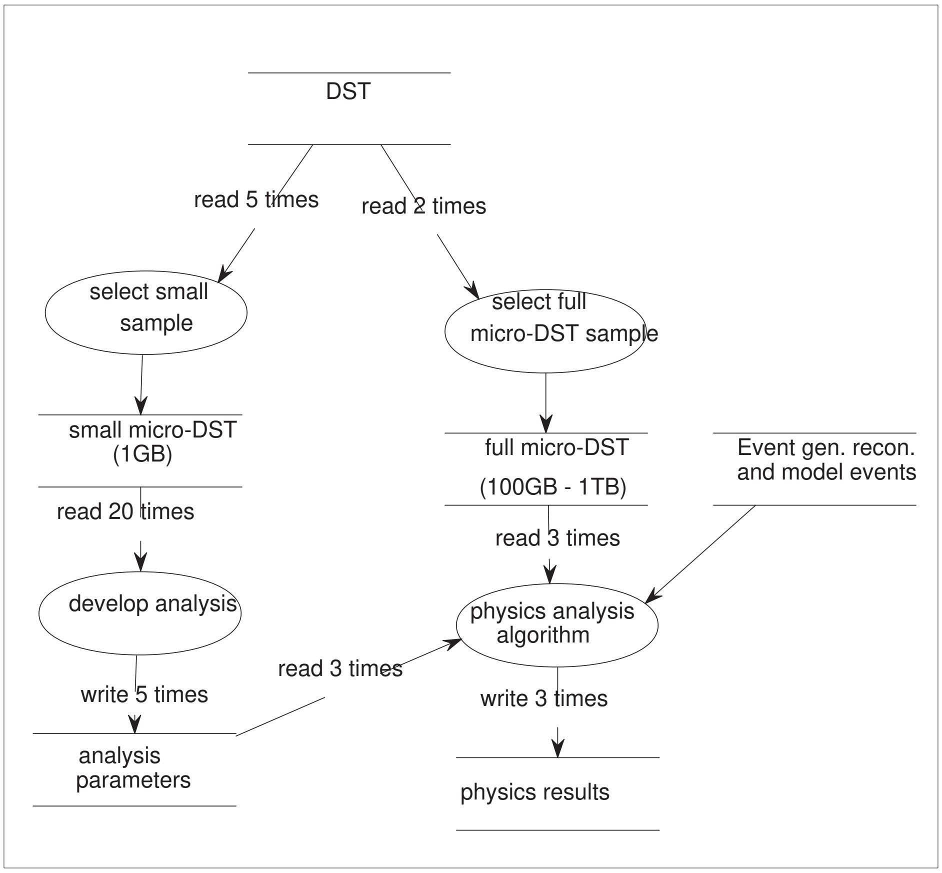

Figure 5: Detailed data flow diagram of STAR Physics Analysis. This activity is commonly called data mining.

Table 19: Data stores in the STAR data flow model.

Data store	Description
calib DB	The calibration database
disk cache at farm	The disk buffer used as input for the event reconstruction farm
disk cache in counting house	The buffer in the on-line system used for de-synchronizing raw data between the on-line and off-line systems. This is part of a possible future scenario with data coming to the reconstruction farm via the network rather than tapes
DST	The primary data summaries produced by the event reconstruction program
EVG config	The configuration data needed to run an event generator
EVG events	Event data from an event generator. Consists primarily of the list of particles in each event.
Fast sim config	The configuration data needed to run the fast simulator
Fast sim. evt data	Data produced by the fast simulator. Consists primarily of space points
GEANT events	Event data generated by GEANT. This consist mostly of energy desposition data in the detectors
GSTAR config	The configuration data needed for running the STAR GEANT detector simulation process
micro-DST	A selected set of DST data used for a particular physics analysis
physics results	The results of a physics analysis, histograms, etc.
sim. raw data	Simulated raw data produced by the slow simulator
Slow sim. config	The configuration data needed to run the slow simulator
tapes	Tapes with raw data

E.4 CPU Requirements

We discuss the cpu requirements for event generation, GEANT, detector simulation, event reconstruction and physics analysis. The event reconstruction process produces DSTs from raw data or from detector simulation data. Detector simulation includes so-called slow simulators which reproduce the physical and electronic response of the detectors and fast simulators which directly produce space points for the tracking detectors. Physics analysis acts on μDSTs producing histograms, cross sections, correlations and other publishable physics results.

E.4.1 Event Reconstruction

The cpu requirements for STAR event reconstruction (DST production) for either raw data or simulated events are shown in Table 20. All values are based on existing software except as noted in the comments. The total amount is 33.2 GFLOP*sec/event, of which 20 is for hit reconstruction in the tracking detectors.

Optimization of the existing software should result in 20 to 50% reduction in cpu usage based on experience with STAR software development over the past few years. However, the addition of further, necessary functionality will probably increase the cpu usage by a factor of 2. We therefore assume an overall increase by a factor of 1.5 from the estimates in Table 20 based on our present software.

In experiment NA36 the average number of tracks per event was 70 and the full tracking, particle ID and V0 reconstruction cpu usage was 270 MFLOP*sec/event. Linearly extrapolating this to STAR (3000 tracks per event) gives 11.6 GFLOP*sec/event, in reasonable aggreement with Table 20.

The vertex TPC tracking code for experiment NA49 takes about 9 to 11 GFLOP*sec per event for about 600 charged tracks per event. This code is still in its early development stage and, so far, only does hit finding and track reconstruction. Scaling this amount linearly to STAR based on the number of clusters per track (typically 25 for STAR and 60 for NA49 TPC) and the number of tracks per event yields an estimate for STAR TPC tracking of (9 to 11)(3000/600) (25/60) = 19 to 23 GFLOP*sec per event. This agrees very well with the comparable STAR values (for TPC hit + TPC track) in Table 20.

The hit finding plus tracking time in EOS takes about 1 GFLOP*sec per event for 100 charged tracks per event. Of this, about 15% is spent hit finding and the rest is spent in tracking and V0 reconstruction. Again, scaling by the number of clusters per track we extrapolate to STAR as (1)(3000/100)(55/25) = 66 GFLOP*sec per event. One known difference between EOS tracking and the current STAR processing is the handling of the inhomogeneous magnetic field and distortion corrections which are part of the EOS processing.

E.4.2 Simulations

The cpu requirements for STAR event generators, GEANT and detector simulation are given in Table 21. For comparison the results from experiment NA49 (main TPC only, about 600 charged tracks) are also shown where available. As with the reconstruction software we also expect the final detector simulation software to require about 50%

Table 20: CPU requirements for STAR event reconstruction based on present software.

Process	STAR (Au + Au) (GFLOP*sec/event)	Comments
TPC hit	18.0	includes deconvolution of merged hits
SVT hit	0.2	no merged hit deconvolution
XTPC hit	1.8	assume 10% of TPC
TOF hit	0.2	
EMC hit	0.1	
TPC track	2.0	includes track fragment joining
SVT track	0.6	
XTPC track	0.2	assume 10% of TPC
Matching	0.4	SVT-TPC track-to-track
V0 rec.	6.0	First pass with loose cuts
Kink rec.	2.4	From NA35 experience
		First pass with loose cuts
Global track	0.3	
Event vertex	0.2	
Track Filter	0.1	
Global PID	0.7	includes SVT, TPC, TOF
Total	**33.2**	

more cpu time than the present version in Table 21. We do not expect any change in the cpu usage for the event generators or GEANT.

Based on experience from NA49 and EOS we will need to reconstruct as many simulated events as real events for determining acceptance corrections, reconstruction efficiencies and systematic errors. In order to compare theoretical models to data with statistical accuracy, as many theoretical events as real events must be analyzed. In addition, many simulated events must be analyzed to determine the acceptances of the various sub-detectors of STAR over varying kinematic ranges and the reconstruction efficiencies for primary particles and rare particles.

Also based on previous experience we have found that the best method of characterizing the detector response is to mix raw data generated with the slow simulator for the few tracks which contain the real physics signal to be studied (high p_T, K, p, K^0_s, Λ, $\bar{\Lambda}$, Ξ, Ω, jets, γ, etc.) with raw data from real events. This produces a so-called "mixed event".

At the physics analysis stage there is a path of taking events directly from the event generator models, passing the particle 4-momenta through a simple filter to characterize the detector and carry out the physics analysis algorithm on these simulated particles.

To account for the considerations above, our model for estimating the amount of CPU and storage requirements for simulations is to generate as many "mixed events" as real events, and in addition to generate as many complete events from theoretical models that go directly to the physics analysis stage.

To characterize generating the mixed events we assume mixing 10 tracks from the GEANT and slow simulator chain into real events and then reconstructing these mixed events. We assume that the cpu time for GEANT and the slow simulator for these mixed events scales by the number of tracks ($\sim 10/3000$). The cpu requirement per mixed event is:

$$\frac{10}{3000}(77 + 1.4 + 1.5 \times 242) + 1.5 \times 33.2 \quad = \quad 51.3 \text{ GFLOP} * \text{sec/event}$$

Pure simulated data will be used in significant amounts before there is real data but afterwards the mixed simulation type is expected to dominate. We expect to run the slow simulators enough on whole events in order to characterize the behavior of the fast simulators. For a given feature being included in the simulator this is expected to require about 1000 events. The fast simulator is expected to be used only for event generator events and not the mixed events.

E.4.3 Model calculations

We expect that a significant amount of cpu is required by all RHIC experiments for the production of event generator output which is required for simulations and for comparing theoretical models with experiment. Relativistic heavy-ion collision models are typically based on Monte Carlo, cascade type approaches which produce output for *single events*, *i.e.* the models do not calculate cross sections or angular distributions directly. Such quantities are obtained from the models, as in the experiment, by accumulating a sufficient number of events for good statistics.

The cpu times for models in Table 21 range from 0.24 to 300 GFLOP*sec/event. Comparison of data with RQMD for example can be extremely cpu time consuming, requiring much more cpu time for the reaction model calculations than for reconstructing and analyzing the data. An accurate estimate of the computing requirements for these model calculations should be done but is beyond the scope of the present study. For now, we assume that as many event generator events as real data events are also needed for comparing the models with data. Additionally we assume that the amount of cpu time for all the model calculations will be the same as that used for event reconstruction of the raw data.

E.4.4 Physics Analysis

The physics analysis cpu requirements for STAR are very preliminary at this point. In order to provide an estimate we have studied the cpu times needed to carry out basic analysis operations such as reading in large, compressed data files, computing histograms and forming two-particle correlations (HBT). The resulting cpu times for histogramming are given in Table 22. Reading 100 events of 1.4MB each (approximate STAR DST size) from an NFS mounted disk across an FDDI network requires 0.02 GFLOP*sec/event.

One of the main goals of the STAR experiment is to calculate relevant physics observables for individual events (*e.g.* pion temperature, mean p_T, K/π ratio, source geometry) and bin the resulting data according to these event-by-event observables, as well as with respect to the mass and energy of the beams, centrality of the collision and other trigger parameters. Table 23 lists a sample of the physics projects that we expect to pursue with STAR along with an estimate of the cpu requirements per event. This table is not a complete listing of all the physics analysis activities which will take place in STAR but we can use it to characterize the range of computing tasks and the typical cpu requirements per event.

Table 21: CPU requirements for STAR simulations based on present software.

Process	STAR (Au+Au) (GFLOP*sec/event)	NA49 (Pb+Pb) (GFLOP*sec/event)
Event Generators:		
HIJING	0.93	–
FRITIOF	0.24	–
VENUS	65	40
RQMD (varied)	300	158
GEANT:		
Physics on	77	36
Physics off	19	13.5
Zebra to data struc. (g2t)	1.4	1.4 (g2ds)
Detector Simulators:		
Slow (SVT+TPC)	242	–
Fast (SVT+TPC)	0.98	–
Intermediate	–	2.7

Table 22: CPU requirements for histogramming from column-wise ntuples with PAW (Au+Au events with 300K ntuple entries).

# events	histogram type	cuts or calculation	GF*sec/ev
100	1-D	none	0.0017
100	2-D	none	0.0037
100	2-D	p(2)%sqrt(ptot) de⟨0.1	0.0082

HBT analysis is expected to dominate the physics analysis cpu requirements for STAR. The 10 GFLOP*sec/event (Au+Au) is based on experiences with HBT analyses in NA44 as well as simple benchmark calculations prepared especially for STAR. This is a fairly significant cpu requirement so extra care was taken to check it's validity. This cpu time assumes a modest ratio of 10:1 of background:foreground pairs in the calculation.

Correlation of event-by-event observables with 3-D source dimensions and source duration is one of the prime analysis techniques of STAR. Clearly this requires HBT analysis on all events (and on many sub-ensembles selected on the basis of other observables). Inclusive HBT analyses with high statistics will be needed using three or more momentum space dimensions (three dimensions assuming azimuthal symmetry, five when reaction plane effects are present) for quantitative studies of source geometry and source dynamics. In typical HBT analyses the computation of the uncorrelated background requires most of the computation time. For STAR data it will be necessary to calculate this background for each beam energy, beam particle species, trigger criteria and sub-ensemble selection criteria. The physics program of STAR is therefore best served by two sets of HBT analyses, event-by-event and inclusive 3-D, on all triggered nucleus-nucleus events.

Comparison of theoretical model predictions with experiment requires that a comparable number of event generator events be analyzed in order to achieve similar statistics. We assume that this analysis includes fast acceptance and efficiency filters which account for instrumental effects followed by HBT correlation analysis. Therefore a significant amount of HBT physics analysis cpu is also required for theoretical model comparisons.

We recognize that many passes on a small sample of events from a given run will be required in order to optimize the analysis parameters. But we assume that one or two passes on the *full* data sample, which constitute the bulk of the cpu demand, should be sufficient.

In addition to the two-particle correlation analyses considered in these estimates, three-body and perhaps higher n-body correlation studies (on selected data samples) may prove useful in differentiating effects due to source coherence in the measured correlation functions from instrumental effects due to finite momentum resolution, particle ID contamination and finite two-track resolution. Multiparticle correlations may also provide increased sensitivity to non-Gaussian components of the source distributions. Since the cpu requirements for three-particle HBT analysis will be considerably greater than that for two-particle HBT we limit our consideration of this topic to small, selected data samples and will scale the overall, two-body HBT cpu estimates by three passes, rather than by one or two passes.

Most of the physics projects listed in Table 23 require about 10^6 events or about 7% of one year's worth of data for each bin in beam energy and particle mass and centrality. Thus in one year's worth of data we expect to accumulate data for a number of such bins.

For a collaboration the size of STAR, experience from CDF and D0 at FNAL suggest that of order 100 physicists (graduate students, post-docs, faculty and staff at universities and researchers at national labs) will work on a physics analysis project during the course of a year. If each person analyzes one, typical trigger bin worth of data ($\sim 10^6$ events) this amounts to 1/14.4 of a years worth of data. For 100 people (100

Table 23: CPU requirements for STAR physics analysis.

Physics Project	Number of events to analyze per E-by-E observable bin	Number of event to analyze per $\sqrt{s}$,A trigger bin	CPU per event GFLOP*sec/ev	Comments
Correlation of E-by-E observ.	NA	10^6	0.5	Calculation of T, $\langle p_T \rangle$, K/π; histograms
π,K,p spectra	10^5	10^6	0.2	Histogram plus efficiency corrections
K_s^0,Λ,$\bar{\Lambda}$	2×10^4	2×10^5	0.6	Assume $\frac{1}{10}$ of first pass V0 reconstruction cpu
Ξ, Ω	10^5	10^6	0.6	Assume $\frac{1}{10}$ of first pass V0 reconstruction cpu
$K^{\pm}$ (kinks)	10^4	10^5	0.24	Assume $\frac{1}{10}$ of first pass kink finder cpu
ϕ mesons (K^+K^-)	4×10^4	4×10^5	0.1	Assume $\sim (K/\pi)^2$ of HBT or $\sim \frac{1}{100}$ of HBT
Triple diff. cross sections; reaction plane	10^5	10^6	0.2	Histogram plus efficiency corrections
Pion E-by-E HBT	10^4	10^5	10	From STAR simulations
Inclusive 3D HBT	5×10^4	5×10^5	10	From STAR simulations
$W^{\pm}$, Z $\gamma\gamma$ γ-gluon fusion	NA NA NA	4M 1M 1M	0.025 0.004 0.008	

physics projects) this corresponds to each of the 14.4M events per year being analyzed for 7 different physics projects. We assume these physics projects include HBT event-by-event, inclusive 3-D HBT plus 5 other hadronic physics projects (see Table 23). The W, Z and γ physics projects (last three entries in Table 23) are expected to consume relatively few cpu cycles and will not be used in these estimates. The average cpu per event for these 5 hadronic physics projects are taken from the first seven items in Table 23 and is about 0.35 GFLOP*sec/event. Many passes on small samples of the data (few 1000) are required in order to develope analysis tools and optimize cuts and parameters. We assume 3 passes on the full data are required to produce publishable results. The resulting cpu requirements for STAR physics analysis is then estimated to be:

$$\begin{aligned}(14.4 \times 10^6 \text{events/year})[0.35\ \text{GFLOP} * \text{sec/event} \times 5\ \text{projects} \\ +10\ \text{GFLOP} * \text{sec/event} \times 2\ \text{HBT projects}] \times 3\ \text{passes} \\ = 940\text{M GFLOP} * \text{sec/year}\ .\end{aligned}$$

Physics analysis will also include processing twice as many simulated events (using event generators and mixed events) as real data. This includes determination of ac-

Table 24: CPU requirements summary per event for offline STAR computing (1 Au+Au central event).

Process	GFLOP*sec/event	Comments
Event Generators	0.24 - 300	
GEANT+g2t (phys. on)	78	
GEANT+g2t (phys. off)	20	
Slow simulators	363	
Fast simulators	1.5	
Generate mixed events (10 tracks)	1.5	Assume linear scaling of GEANT and slow sim.
Hit reconstruction	30	
Tracking, PID, V0	19.8	
Produce μDST (minimal selection calculation)	0.08	
Read μDST and 20 hist.	0.51	
HBT per pass	10	
Typical hadronic physics analysis	0.35	
Typical W, Z and γ physics analysis	0.02	

ceptances and reconstruction efficiencies for π, K, p, K_s^0, Λ, $\bar{\Lambda}$, Ξ, Ω, jets, γ, etc. as well as HBT analysis of simulated events. We assume the analysis of simulated events will be more focused than that for real data and will result in cpu requirements equal to that for the real data, rather than twice as much.

A summary of the per event cpu requirements for all off-line STAR computing is given in Table 24. The assumed increase of 50% in the simulation and event reconstruction code from present values was included in Table 24.

E.5 History of Estimates of STAR CPU Requirements

In the first ROCOCO report from September 1992 the cpu estimate for STAR was 10 GFLOP*sec/Au+Au event. This was primarily based on the TPC tracking software available at the time. The estimated steady state cpu rate for STAR was 10 GFLOP assuming 2000 hours per year of RHIC operations. The 10 GFLOP included an estimated factor of 4 to 5 times as much cpu for simulations as for event reconstruction of the real data. These estimates did not account for event reconstruction for the additional STAR detectors, massive running of event generators and theoretical models for comparison with data, and physics analysis of the simulation and real data DSTs.

In 1993 an interim report by the RHIC Off-Line Computing Study Group called for an overall four-fold increase in the total RHIC cpu rate from 40 GFLOP to 160

GFLOP (one fourth for STAR). Again this assumption was based on a 2000 hour running year for RHIC but used more realistic simulation software developed by the STAR collaboration. The study group concluded that such a facility (to be completed by 1997) was too costly and therefore continued to analyze computing models which were based on the 40 GFLOP scale presented in the 1992 ROCOCO report.

An internal STAR estimate of cpu requirements in 1994 arrived at a cpu need of 240 GFLOPS. Considerably more effort has been expended refining this estimate with more benchmarking of the newest STAR codes and more detailed comparisons with other TPC experiments. All of this recent effort is reflected in the present results.

With our current assumption of 4000 hours per year of RHIC operation and assuming the simulation cpu required scales with the number of real data events, the 1993 RHIC Off-Line Computing Study Group's estimate for STAR simulations and event reconstruction would have been 80 GFLOPS. This should be compared with our present estimates in Table 26 for event reconstruction of real data (28 GFLOP) plus simulations (29 GFLOP) or 57 GFLOP *which is actually less than the comparable estimate in 1993.*

In the present report we account for the cpu requirements for large scale production running of event generators and theoretical models (28 GFLOP) and most significantly, as far as cpu requirements, physics analysis of simulations and data DSTs (60 GFLOP each, 120 GFLOP total). Estimates of the physics analysis cpu requirements for STAR have only been done recently. Most (~90%) of this physics analysis cpu requirement is for HBT analyses.

The increased cpu estimate for STAR in the present report, compared to those in '92 and '93, is therefore a result of basing the estimates on much more realistic software, assuming twice as much data per year and accounting for theoretical model running and physics analysis.

E.6 Data Volumes

Table 25 lists the various data objects per event for typical Au+Au equivalent central events. Refer to the computing model diagram in Fig. 4.

We assume that small sample μDSTs (1 GB) will be selected from the master DST data store and will be read several times for tuning the analysis modules and parameters. After tuning the parameters a full μDST (0.1 - 1 TB) will be produced and read 3 times during the course of doing the physics analysis.

E.7 Summary of Annual Total CPU and Data Requirements

This summary applies to a nominal steady state operating scenario after the accelerator and experiment are operating near the design specifications, presumably about the year 2000. The following sections deal with the time dependent issues relating to ramping up to this level of operation.

Tables 26, 27 and 28 summarize the total, annual cpu rate and data volume and storage requirements for STAR offline simulations, event reconstruction and physics

Table 25: Data volume for 1 Au+Au central event, or equivalent, for STAR.

Data Item	Size (MB)	Comments
Event Generator	1	
GEANT+g2t (phys. on)	29	
GEANT+g2t (phys. off)	17.5	
Raw data (from slow sim.)	16	
Raw data (from fast sim.)	16	
Raw data (from DAQ)	16	
Calibrated data	32	Temporary during development of DST production
DST	1.6	1 real, 2 simulated
μDST	0.16	5 for each real event
ntuples	0.2	1 for each μDST

Table 26: Annual CPU rate summary for STAR.

Process	GF*sec/ev	Events/yr	Duty factor	Capacity (GFLOP)
Event Rec. (real data)	49.8	14.4M	0.8	28
Event Gen. Models	25	28.8M	0.8	28
Simulations (Ev. gen. and Mixed)	51.3	14.4M	0.8	29
Physics analysis (data)	65	14.4M	0.5	60
Physics analysis (sim)	32	28.8M	0.5	60
Total				**205**

analysis. Duty factors for production running are assumed to be 0.8; this includes event reconstruction, event generators, and simulations. Physics analysis was (optimistically) assumed to operate at 50% efficiency. *The total, annual cpu rate for STAR is 205 GFLOP; the total, annual data storage volume is 312 TB.*

E.8 CPU Hardware Requirements

The event reconstruction and GEANT simulation processes for STAR require approximately 128MB physical memory and nearly 1 GB virtual memory for full Au+Au events. This will set the requirements for the majority of the farm processors as well as the average desktop computers which get used for data samples and development of every type of computing task. The networking requirements for the farm processors can be determined based upon the task (event reconstruction, model calculation, GEANT simulation, data mining, etc.) and the CPU power of the machines chosen for purchase.

The desktop computers of the physicists in the collaboration will be used for a variety of tasks including code development and every type of processing task for samples of data. Based upon these uses and the requirements mentioned above the typical desktop computer should be a workstation with about 0.5 GFLOPS cpu power and 100 to 200 MB memory.

Table 27: 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 28: 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**	

Due to the large event sizes, each physicists will need many GB (10GB or more) of local disk space in order to have a few samples of data the development and analysis tasks. A typical small sample dataset is about 1 GB and each person will need several.

E.9 STAR schedule

The steady-state scenario described above amounts to quite a significant computing operation. For this to be achieved, from both the computer center point of view and the STAR experiment point of view, there are many things that need to be developed, tested and debugged ahead of time. In this section we describe the development profile required for STAR and the implications of this on the computer facility profile.

We assume that the computer facilities available to STAR provide the basic hardware and system software for cpu cycles, data storage, data access, networking and operations personnel to keep these resources operating. In addition we assume that any general purpose commercial or third party software (commercial databases, GNU software, CERN libraries, etc.) are provided as part of the computer facilities operations as well. In addition to these basic facilities STAR will develop the necessary higher-level software systems to carry out the data processing operations and bookkeeping for the data processing status and data access for physics analysis.

The STAR development schedule is described with two main branches. One is for developing the production system software and tools (sys) and the other is for developing the physics analysis algorithms and detector simulation studies (phy).

- **October 1995**
  (sys) Begin architectural design of off-line production system. Begin implementation of final analysis framework.
  Develop prototype off-line production system to support simulations and test concepts for final system.
  (phy) Initial version of full event reconstruction. Detector simulation of tens of events.

- **April 1996**
  (sys) Begin testing distributed data access concepts. Identify network limitations between DOE labs and university groups.
  (phy) Test prototype off-line production system with simulations.

- **October 1996**
  (sys) Finalize the interface descriptions between the STAR software components and the computer facility hardware and software.
  Begin development of production system components.
  Develop prototype data access and distribution methods.
  (phy) First complete event reconstruction program in final analysis framework.
  Operate prototype off-line system for production simulations.

- **October 1997**
  (sys) First testing and debugging of the individual software components for the entire data processing system.
  Use prototype off-line system for stress testing of final components.
  (phy) Begin physics analysis of simulations at collaborating institutions using the prototype data distribution procedures.

Table 29: Profile for STAR processing before achieving steady-state operations

Feature	**FY96**	**FY97**	**FY98**	**FY99**	**FY00**
GEANT events	10^5	5×10^5	2×10^6	10^7	1.4×10^7
Model events	2×10^5	10^6	4×10^6	2×10^7	2.8×10^7
GFLOPS capacity[a]	1	5	20	100	205
Total data volume (TB)	2.2	11	44	220	312
On-line storage (TB)[b]	2.2	10	20	40	90

[a] The cpu capacity indicated takes into account the low duty factors which will be achieved during the development phase.
[b] On-line storage is the subset of the total data generated that needs to be accessible on a short time scale (minutes).

- **October 1998**
  (sys) First full installation of complete set of software for the entire data processing operation. This marks the start of the full system testing and debugging.
  (phy) Complete full chain of simulation, event reconstruction and physics analysis software. Operate full chain including all simulation and physics analysis activities at collaborating institutions.
- **July 1999**
  Achieve smooth operation of all aspects of data processing using simulated data at full data bandwidth and cpu loads. This constitutes a full load test of the computing facilities.
- **October 1999**
  Start processing real data.
- **March 2000**
  Operating with full efficiency at design level.

Table 29 lists a schedule of number of events generated, number of cpu cycles and storage requirements associated with the descriptive schedule above. The quantities in the table correspond to the usage of the facilities during the year indicated.

E.10 Comments on resources needed beyond the RHICCC

The schedule for the RHIC Computer Center meets part of the STAR computing need and is focused primarily at being available for meeting the need for event reconstruction of real data. There are, however, some important areas where we expect to satisfy STAR's additional requirements with other facilities.

In the time period before the end of FY97, at which point the first phase of the RHIC center becomes operational, there is an increasing simulation and software development effort which we think can be met with a combination of resources from NERSC and individual STAR groups. In particular, beginning the effort to develop a solution to the data mining problem is important to start soon since it has a long lead time. This activity is a combination of two difficult problems, the simultaneous access by many physicists to massive amounts of data, and distributing large quantities of data over a wide area network to the home institutions of these physicists. The expertise at NERSC and LBNL in distributed computing, networking, hierarchical mass storage, and high-bandwidth data access in combination with the STAR computing personnel at LBNL means that a collaboration between LBNL, NERSC and RHIC Computing would be very effective at addressing this problem. Also the mass storage facilities at NERSC can provide a necessary resource for this effort.

In FY98 we need to install the STAR production software on the RHIC computing facilities in order to be able to be ready for the event reconstruction load of real data which starts in FY99. In keeping with the RHIC Computing Model (Fig. 1) we expect to continue to need significant facilities for simulations and data analysis at remote sites after the RHIC Computer Center becomes operational as well.

F Other "Big" experiments

F.1 The D0 experiment at Fermilab

The intellectual center for the analysis of D0 physics is Fermilab.

- D0 farms out a significant fraction of its simulation work, maybe 50%, to collaborating institutions which have substantial computing resources available. The D0 simulation load is of order 10% of the total compute load.
- All other D0 production processing is done at Fermilab.
- The situation is approximately the same for CDF.

The total Fermilab computing is probably about 3 times what D0 has. The other two thirds service CDF and the fixed target program.

At Fermilab, huge farms of trailers house approximately 200 D0 personnel at any given time. Virtually every collaborating institution has several graduate students and postdocs in residence there for extended periods.

Event reconstruction is done in farms. There are presently 75 Indigo SGI machines, and 25 IBM RS6000/320 machines in 4 sub-farms doing D0 reconstruction. The computers are in racks, and do not have monitors. The reconstruction farms have been tuned to match processing power to I/O speed. This tuning has changed with time and has been different for D0 and CDF.

D0 writes data at 2-3Hz. Their event size is 500-600 KB for a recording speed of 2 MB/s, at a duty factor of 2/3. The experiment has been running for about 2 years.

The reconstruction cluster produces Standard output (STA), and DST output. The STA is about the same data volume as the raw data. The DST is about 1/10 this volume. The μDSTs are about 1/100 of the STA. The total data volume is $\sim$ 100 Tb in $\sim$ 50K 8mm tapes. The D0 goal is to have all of the μDST data on disk.

There is a streaming farm that splits data into seperate streams depending on event type. DODAD is based on a CERN product and permits access of individual events from a random access store eliminating the need for separate stores for each stream. The steaming thus becomes "virtual" for even sets which can be stored on disk, such as the DST's.

There is presently a fileserver consisting of 4 alphas, and 20 other VAX stations. The disk storage is $\sim$ 1Tb. Data access is done using the CERN program FATMAN.

There are about 6 Fermilab employeed D0 physicists whose service contribution to the experiment consists of managing the farms and implementing software. There are others (students and postdocs) who take shifts keeping the farms running.

There is an analysis cluster of about 150 Vaxes and alphas with peripherals the majority of which have been brought by the collaborating institutions. These are operated and managed by Fermilab employees.

There is an offline computing policy board to allocate resources within D0.

There are 2 high performance analysis systems in use by D0 at Fermilab. One uses the PIAF system from CERN on an SGI challenge platform. The other is a development effort based on the PASS project started at LBNL a few years ago, which, in this case, uses a 24 processor IBM SP-2, and is refered to as a data mining project. The STK silo system will be connected to one or both systems. The progress in these projects should be monitored.

F.2 CLAS at CEBAF

Experiments running at CEBAF are currently acquiring data at a rate of 1 GB/day. This is expected to increase to 100 GB/day in the coming year and by 1997, when CLAS comes on line, experiments will acquire 1TByte/day at a rate of 10-20MBytes/sec. The data acquisition will consist of crates going through an ATM switch to online CPU farms which will process events. Data will then be shipped back to one CPU to order events and store to tape. There will be virtually no operators, so the system must be highly automated. Experimenters have expressed a strong desire to have data collection separate from the data storage, so the question of how to get the data into a mass storage system with a minimum number of copies is still being addressed.

HPSS seems an ideal choice for the data storage needs, but is still two years away from production. That is not fast enough for the CEBAF time frame, but may very well be good for the RHIC computing time frame. They will use OSM. Several other labs have some potentially attractive software which might be adapted to get raw data tapes from the experiment into the HSM. This problem is also still under consideration.

The computer center will have 10,000 MIPS in offline machines to do the analysis of the raw data. The analysis estimate is that users will make 2-3 passes through the raw data. This will occur onsite using the mass storage system and the computing facilities provided. The "reduced" tapes will be carried to the home institutions for the physics analysis probably in the DLT format. Since there is often no reduction of data quantity in the "reduced" tapes, these tapes may, in fact, stay at CEBAF. A few labs, if any, might access the reduced data over the network if they have access to a connection with a high enough baud rate. It is expected that some groups will purchase analysis stations and park them at CEBAF in order to access the data locally logging in from off site.

The computer center staff consists of 3 technicians, 1 system programmer, 1 data systems analyst which will manage data collection, 4 system administrators, 2 personal computer specialists and the Computer Center Manager. Two MIS programmers, 1 CAD coordinator and one batch systems manager will be added in the near future. This group plus a few students is responsible for handling all LAN/WAN design/upgrade/installation/management/maintenence, manage all UNIX/VMS systems (about 15-20 clusters, total of 150-200 workstations), provide hardware/software support for 400+ PC's, 200+ printers, 300+ terminals, 100+ xterms, software support for 250 Macs, modems, do initial user training, user assistance (over 1000 active users), video conference/other video support, planning/reporting, PLUS support the experimental program which is just beginning. This group faces in the next 12 months to integrate two new operating systems, upgrade the network to ATM, make the decision and implement an automated data storage system. They will have to install and manage all tape devices, silos, HSM or other data management software with more or less the existing staff. It is therefore obvious that the computer center staff is swamped, so the assistance the computer center gives to users is limited to providing a good set of commercial and public domain products for general computing needs and informing users how to access them with virtually no assistance beyond that. All assistance with coding is handled within the workgroups and collaborations.

In conclusion, the CEBAF experiments will acquire about 350 Tbytes/year at 10-20 Mbytes/sec. The offline computing model is that the data flows from the experiment to a raw tape which is then transported in a yet to be determined way into the mass storage system. The details are still under consideration. Event reconstruction is performed onsite at CEBAF using about 10000 MIPS of offline computing which generates DST's which may or may not constitute a reduction in data volume. These DST's will be "carried to the home institutions" for physics analysis although "carried home" may end up to be defined as the home institutions purchasing analysis machines that are located at CEBAF to access the data locally with users logging in from offsite. The computer staff is responsible for keeping the system running, but will provide little assistance to the end users. Experiments will be responsible for that. The computer staff for the computer center is expected to consist of around 15 individuals.

We thank Rita Chambers for providing us with the above information.

F.3 The BaBar Experiment at SLAC

The BaBar experiment at SLAC is in many respects comparable to the two large RHIC experiments, PHENIX and STAR. The BaBar collaboration consists currently of close to 500 members distributed over 77 institutions in 10 countries. The detector is scheduled to begin data taking in 1999. The global computing requirements are somewhat smaller than at RHIC:

- Data recording rate: 2.5 MBytes/sec.
- Annual data storage: 85 TBytes.
- CPU required: 5-10 GFlop.

In the design and construction of the computing system BaBar is building on a long tradition in high energy physics and has chosen a very interesting and more radical strategy than similar, but earlier, experiments. The computing system is being treated at the management level on an equal footing with the mechanical and the electronic systems. The software development scheme is completely object oriented. The use of C++ for all computing tasks is strongly recommended and large parts of the software will be developed with the Rational/Rose CASE tool based on the Booch methodology. BaBar has started an aggressive training program for many of its physicists in the use of C++.

The BaBar computing model is very similar to the one developed for RHIC. The primary data storage, the central database and the CPU-intensive event reconstruction will be on-site at SLAC. Like PHENIX, BaBar will develop overseas regional centers, in England and France where most likely the bulk of the Monte Carlo calculations will be performed.

Many former SSC computing experts are now working within BaBar, which makes it interesting to follow the design of the BaBar computing system. BaBar has in particular experience in large scale data storage based on CORBA from the former PASS project and several contacts between BaBar and RHIC computing have started to develop within the last year with the purpose of trying to generate a common solution to the data storage problem.

F.4 CERN

F.4.1 Computing for NA49

NA49 is a fixed target heavy-ion experiment at CERN consisting of two vertex TPCs, two main TPCs (only one is operational so far), plus a time-of-flight system. The detector is designed for the high multiplicities produced by the SPS Pb beam where a typical Pb+Pb central collision produces about 1000 hadrons. The experiment presently has 167k electronic channels (more than the STAR TPC) and during the first year's run in 1994 achieved an average event rate of 1 - 2 per second. The raw data size is approximately 10MB/event. The data rate during a beam spill was 64MB/sec, which, with the 25% duty factor for the SPS, resulted in a data recording rate of 16MB/sec. The data were recorded with a SONY tape system. This average data recording rate is comparable to that anticipated for either STAR or PHENIX.

The 7 day run in 1994 resulted in 1.5TB of raw data. It is anticipated that the scheduled 1995 run will generate ~10TB of raw data and that a total of ~50TB of raw data will be generated during the lifetime of the experiment. The annual data volume of ~10TB is, however, much less than any of the RHIC experiments whose total annual data volume is estimated to be about 1 PetaByte or 1000TB.

Raw data analysis and DST production for NA49 is done on the CERN CORE (Centrally Operated RISC Environment) processor farm, specifically that portion known as SHIFT, which is a dedicated facility for event reconstruction and DST production. The SHIFT facility is shared among all the CERN experiments, but is primarily used by the four LEP experiments (see section on CERN computing). The present implementation of the NA49 event reconstruction code requires 10.8 GFLOP*sec of cpu per central event. This results in an average, annual cpu requirement of 0.34 GFLOP, which is about 1/100 of that estimated for STAR event reconstruction. Data analysis on the SHIFT facility was supplemented by five HP735 processors which were provided by the experiment for both batch and interactive computing. Data volume reduction to DSTs is presently 10:1 resulting in ~1TB/year of DST production which will be stored on 1000 8mm Exabyte tapes and later on ~100 digital linear tapes (DLT).

Raw data access for DST production is provided by SCSI disks and tape robots. NA49 typically uses 100GB of SCSI disks (out of ~2TB for CORE) and 800GB (soon to be increased to 1.5TB) of tape robot capacity out of the total robotic capacity for CORE of 80TB. NA49 uses the bulk of a single 2TB tape robot mounted from an IBM 3494. The experiment uses the CERN File and Tape Management software (FATMAN) whereby the user refers to the data by a unix-like name and the FATMAN software matches the name to the physical data set and storage medium.

Simulations (event generators, Geant and main TPC fast simulator) require about 1/2 the cpu per event that STAR requires (see section on STAR cpu requirements) and are done on the CORE CSF (Central Simulation Facility) which consists of a farm of 45 HP processors.

Physics analysis of DSTs will be done at five of the NA49 collaborating institutions [Munich (MPI), IKF, U. Birmingham, U. Washington, and LBNL] using local workstations (from 5 to 10 at each institution) and several platforms (HP, Sun, AIX).

Experience from NA49 computing has so far identified a number of problems which are relevant to the RHIC experiments. These include:

1. The widely varying time zones (U.S. and Europe) result in slow turn around

(typically 12 hours) with respect to correcting problems with the software library, tape robots and the processor farms.

2. Unix scripts are not necessarily platform independent as expected.
3. The batch queue system does not account for, nor handle, hardware failures.
4. Users need to have access to the batch processing disks for real-time monitoring of batch processing results.
5. Support of multiple platform types (Sun, HP, AIX) has resulted in the location of event reconstruction software bugs which might not have been detected otherwise.

F.4.2 CORE

The information shown below was extracted from WEB pages.

CORE, the Centrally Operated RISC Environment, is the collective name for the physics data processing services offered by the CERN Computer Centre, and operated on over 200 RISC microprocessors, 2 TeraBytes of SCSI disk and 100 magnetic tape drives. The processors are installed in a variety of computer systems, including Meiko CS2 and IBM SP2 scalable parallel computers, Silicon Graphics Challenge-XL multiprocessor systems, and clusters of Hewlett-Packard, Digital Equipment Corporation and SUN Microsystems workstations.

CORE presents the users with a series of different services, each configured to provide the required performance and capacity using the most cost-effective equipment. For example, the CSF simulation facility is configured for maximum computational capacity at minimum cost, while the SHIFT DST analysis service aims to combine general purpose batch computing power with high bandwidth access to large capacity disk and tape storage.

The aim of CSF is high capacity, low cost computation for physics event simulation. An average CSF job uses the GEANT event simulation program and runs for 12 to 36 hours on a single processor, generating 200 MB of data which it then copies out to tape using the CORE tape service. CSF consists of 25 H-P model 735/99 workstations, shortly to be expanded by the addition of 20 H-P model 712 systems.

SHIFT was developed, starting in 1990 to provide batch computing services on inexpensive RISC processors, with fast access to large amounts of disk data, and good tape support. It had to be vendor independent, cheap and easy to expand, but with the integration, reliability and overall quality of the mainframe services which had long been the workhorses of physics computing at CERN. SHIFT quickly grew to overshadow the mainframes, keeping up with the computing demands of the LEP experiments while enabling the associated budget to be progressively reduced.

SHIFT provides a general purpose batch service, but its main use has been analysis of LEP DSTs (data summary tape - the name given to the experiment's master physics database). The DST of each experiment occupies about 100GB, and is generally held online on disk.

SHIFT is configured as a series of sub-services, each organised for a specific physics collaboration, thereby ensuring that the experiment has guaranteed computational and storage capacity and can therefore schedule its work in a predictable manner.

In order to provide the required performance and reliability in this distributed and heterogeneous environment, it was necessary to develop or acquire a certain amount of basic software:

- the tape daemon, a portable Unix tape subsystem (multi-user, labels, multi-file, operator and robot support);
- rfio - a fast remote file access system;
- rtcopy - a distributed tape-disk file mover; a disk space manager which deals with pools of file systems distributed across many different disk servers, avoiding limits on file system capacity and disk server performance;
- the tape stager which uses the above tools to implement reliable caching of tape files on disk;
- a clustered version of the NASA NQS batch system which provides load levelling across cpu servers;
- integration of the above facilities with standard physics I/O packages such as FATMEN, RZ, FZ, and EPIO;
- tools for network operation and monitoring.

The IBM SP2 Service is a general purpose public service operated on an IBM SP2, a large 64-processor scalable parallel computer. The service is available to all CERN users, but collaborations with major computational requirements are subject to review by the COCOTIME resource allocation committee. A 16-processor partition of the SP2 provides general purpose interactive services run by CN's DCI Group. The remainder of the machine provides a CORE batch service compatible with the interactive service. Some of the SP2 nodes act as CORE disk and tape servers.

PIAF was developed in CN's ASD Group by Rene Brun, Alfred Nathaniel and Fons Rademakers, with support from Hewlett-Packard. It provides a data-parallel environment for the PAW (Physics Analysis Workstation) system for interactive analysis of physics ntuples. The service is offered on eight Hewlett-Packard model 755 workstations, each with fast SCSI disks. The ntuple files are striped across all of the server workstations, and a PAW transaction is split into parts which are executed simultaneously on all of the servers, each analysing the local section of the ntuple file. The PIAF server combines the results for display to the user.

The Meiko CS2 computer is a distributed memory scalable parallel system using SPARC microprocessors and a smart interconnect which enables programs to read and write memory in remote nodes without context switching. The CERN CS2 has 32 nodes, each with one processor and 32 MB of memory. The computer will be upgraded in March 1995 when each node will have dual 100 MHz HyperSPARC processors and 64 MB of memory.

The CS2 has been funded by a European Commission R&D project under the Esprit framework, called GPMIMD2 The project also funds a certain amount of manpower for parallel applications development at CERN. The CS2 service is at present used primarily for the support of these applications, in particular an event-parallel version of the GEANT simulation program, and a fast Monte Carlo developed by the NA48 collaboration and using a distributed pre-computed particle shower library. It also provides services to other partners in the GPMIMD2 project and to Europort, another European Commission funded project.

The Computer Centre tape vault contains some 250,000 cartridge tapes stored in manual racks, 17,000 cartridges in each of two IBM model 3495 automated cartridge loaders (robots), and over 3,000 cartridges in an IBM 3494 and a DEC TL820. This provides a total storage capacity of 80 TeraBytes. About 30,000 tape volumes are mounted each week, over 80% of them for LEP experiments.

The CORE tape service is provided through fourteen tape servers: eleven RS/6000s (using IBM ESCON, Parallel Channel, and SCSI connections), two SUN SPARC stations, and a DEC Alpha server.

Evaluation of new high density tape equipment from DEC and IBM is under way.

F.4.3 CORE Infrastructure

This includes facilities common to the different services, such as registration, the home directory file base, the magnetic tape storage services, and the public disk storage pools. The infrastructure services include operation and support of basic software facilities such as remote file access and batch job scheduling. One of the goals of CORE is the maintenance of a set of standards for service management to ensure that a wide variety of services can be provided with a minimum of manpower.

F.4.4 CORE Networking

High performance, reliable networking is essential to the success of the distributed CORE environment. The CORE network handles 8 TeraBytes of data each week, using four network technologies:

- UltraNet: The original network of CORE, installed in 1990, it has a very high performance (1 Gigabit/second) backbone and supports data transfer rates on individual connections of up to 14 MB/second.
- FDDI: A 100 Mbit/second network standard, which is configured with several segments and a DEC Gigaswitch backbone.
- Ethernet is used for systems which do not have high performance requirements, and is also used as the standard connection to CERN's public local area network.
- HiPPI is currently being installed for very high performance applications.

F.5 URLs for HEP/NP computer centers

- FNAL
  CAP - Computing for Analysis Project http://fnhppc.fnal.gov/cap/cap.html
  CLUBS - Clustered Large Unix Batch System
  http://fnhppc.fnal.gov/clubs/sys_over.html
- SLAC
  SLACVX - SLD VAX cluster for off-line processing
  http://www-sld.slac.stanford.edu/sldwww/slacvx/slacvx.html
- CEBAF
  http://www.cebaf.gov/comp_center/com_center.html
- HPSS
  http://www.llnl.gov/liv_comp/nsl/hpss/hpss.html
- Storage System Standards Working Group
  http://www.arl.mil/IEEE/ssswg.html

- CERN
  CORE - Centrally Operated RISC Environment
  http://wwwcn.cern.ch/pdp/Services.html
- DESY
  Computing info under the User Consulting Office
  http://www.desy.de/
- KEK
  Computing at KEK
  http://www.kek.jp/kek/computing.html
- GSI
  Computing
  http://www.gsi.de/computing/dvee.html
- IN2P3
  Computer Center
  http://www.in2p3.fr/html/ccin2p3/e_systemes.html
- NERSC and NSL
  National Energy Research Supercomputer Center and National Storage Lab.
  http://www.nersc.gov/

G History of RHIC computing estimates

In the following, we present a brief chronology of activities related to the RHIC project. The assessment of computing needs for RHIC has evolved in response to the finalization of the designs of the detectors, and improved understandings of all of the tasks that have to be performed to do RHIC physics.

1984
First ever meeting (2 days) on RHIC detectors at BNL. Rough concepts developed for calorimeters, lepton and hadron spectrometers, 4π devices. No discussion whatsoever of data rates or computing.

1985
Week long workshop at BNL. Only rough channel counts discussed for 4π, dimuon, single-arm, and forward-arm spectrometers. Only vague references to computing needs.

1986
First real data taken with high energy heavy ion beams. Data used to revise event generators.

1987
Week long workshop at LBNL. Detector concepts developed and defended for first time. Electron and photon spectrometers actively discussed. First monte-carlo efforts (still pre-GEANT). Still no discussion of data volumes.

1988
Week long workshop at BNL. Identified physics and detector concepts actively discussed. First real-time computing estimates (Watson + Levine). First GEANT simulations, but no detector response yet. Some rough guesses at data volumes and simulation efforts.

1989
Expressions of interest. Detector R&D program started. RHIC final funding proposal sent to DOE. Funds included for initial computing effort.

1990
Formal expressions of interest. Initial estimates of detector channel counts for "real" detectors that evolved into actual program. GEANT and Monte-Carlo work still concentrating on seeing physics signals - no detector response included. Active experiments at SPS and AGS enter 5^{th} year of datataking, but storage needs still met by 1-2 tape drives and existing laboratory computer centers for playback/reconstruction. Departmental "VAXes" used for physics analysis.

1991
RHIC project gets contruction approval. Formal proposals for detectors backed up by first simulations of real detectors. Earliest computing power and storage estimates all based on scaling up of WA80, NA35, E802, E814, E810 processing needs. Projections tended to lead to a "few" GFlops and "several thousand Exabyte tapes of data per experiment per year." That would mean $<$ 10 TByte per experiment per year of data (Exabyte tapes were 2.3 GB then and still novel). Initial purchases of RHIC computing hardware, plus initial manpower (Tom Throwe) to operate system plus organization of HE-NP computing group in BNL Physics

Department under Bruce Gibbard. First organization of "consulting group" with RHIC users on computing needs.

1992

STAR gets design approval, starts assembling group and building detailed simulation programs. First estimates of actual raw data rates and the first realization that several 100 TBytes/year of storage would be needed driven by need for >150K channels and>40M pixels in detector. PHENIX gets physics approval, first estimates of raw data rates. New detector concept as mandated by RHIC leads to need for > 250K detector channels, raw event sizes of 5 MB (zero-suppressed). Realization that raw data storage needs are similar to STAR. BRAHMS and PHOBOS concepts put forth. Active use of RHIC computing cluster purchased with project money. Report from ROCOCO-1 committee chaired by Bill Love [1]. The estimates in this report are based on a 2000 Hr/year RHIC operation.

1993

STAR gets construction approval, first simulations for actual detector concept, first estimates of computing for event reconstruction. Realization that several 10s of GFlops of computing needed. PHENIX gets design approval, first estimates of reconstruction computing needs and first estimates of analysis needs, all based on timing of early simulation code with first-ever track reconstruction code. First realization that computing power in excess of 100 GFlops might be needed. RHIC study group report released [2]. Recognized need for massive storage, robotic support, centralized facility, massively parallel computing farms of 100 GFlops at least. First staffing estimates.

1994

Approvals for BRAHMS and PHOBOS at various levels. Main development of STAR and PHENIX simulation and pre-offline codes. Firming up of estimates for computing needed. Physics analysis needs still rudimentary and based on "desktop workstation model." Initial development of BRAHMS and PHOBOS codes and estimates of needs. Submission of field work proposal to DoE to request funds for RHIC computing. Submission of proposal for additional experimental equipment to DoE.

1995

Presentations of RHIC-CC to NSAC subcommittee, NSAC, NSAC-LRPWG. Development of estimates for physics analysis computing needs. Realization that 100-200 GFlops analysis power also needed. Development of collaboration computing models (workstations, networks, supercomputer usage).

1996

This report. The computing estimates are based on a 4000 Hr/year RHIC operation.

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] Experimental Equipment for RHIC, T. Ludlam, J. Harris, and S. Nagamiya, September 1994, p. 56.