# The *S*olenoidal *Tr*acker *A*t *R*HIC (STAR) Computing Resource Plan


**January 14, 2009**

# The *S*olenoidal *Tr*acker *A*t *R*HIC (STAR) Computing Resource Plan - 2009

## Executive Summary

The STAR detector and the RHIC accelerator are well on the way to constructing—and in some cases completing—a suite of strategically targeted upgrades of moderate scope which promise to enter in an entirely new era of fundamental heavy ion and spin studies of extended scientific reach. These studies will build on the discoveries of the first phase of RHIC experimentation by utilizing the increased luminosity provided by the RHIC II accelerator upgrade and by implementing new detector instrumentation strategically targeted to enhance STAR's acceptance, particle identification capability, and effective sampling of luminosity. To capitalize on these investments, it is essential that the computing capability of the STAR experiment, now and in the future, also be strategically positioned to receive and analyze the flood of data which the upgraded STAR detector will produce. The plan to meet STAR's future needs for storage, CPU, network capacity, and software and computing workforce in order to accomplish this are summarized here and documented in the following report.

**Structure of the present and future STAR Software and Computing Effort**

Similar to other major international Software and Computing (S&C) projects, the STAR S&C enterprise is organized in a Tier structure according to the availability of capacity and services (e.g., storage, CPU) and dedicated workforce at collaborating institutions.

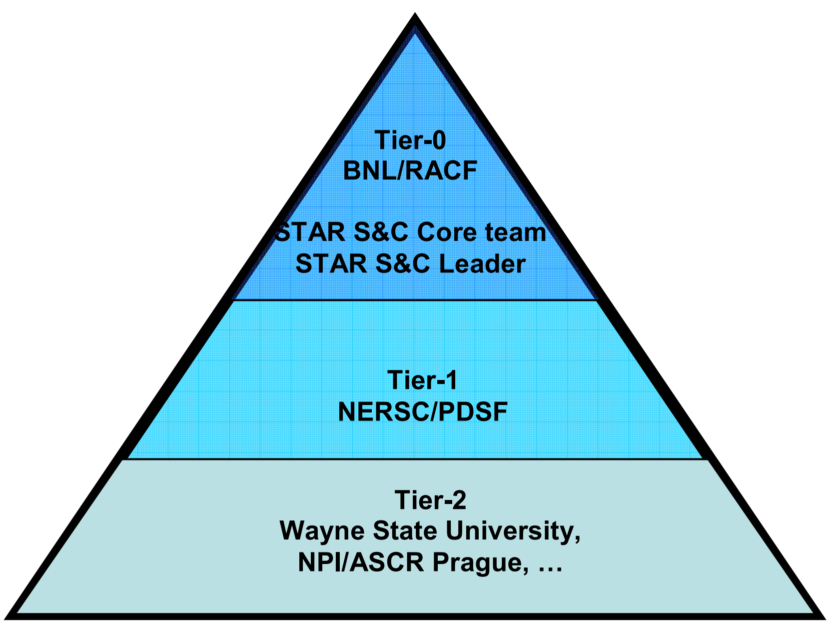


Fig1. Present Structure of the STAR Software and Computing (S&C) effort.
The future addition of an additional Tier-1 center in South Korea is planned.

As shown in Fig. 1, the main tiers in this structure include a Tier-0 center at Brookhaven National Laboratory (BNL), a Tier-1 center at PDSF/NERSC, and Tier-2 centers at collaborating STAR universities.

The STAR S&C resource plan relies heavily on the assumption that the existing Tier centers will continue to play the same basic role they have to this point with a level of effort which (with one exception discussed below in relation to continued evolution toward distributed disk storage) is approximately constant. However, the model in which these activities will be carried out will need to evolve to meet the Collaboration's needs. Thus far, the STAR production and analysis models have mainly relied on centralized user analysis facilities (BNL and PDSF) to provide the bulk of the analysis power for STAR collaborators and scientists. The STAR computing model has, however, been steadily evolving toward a data-grid model in which processed data is made immediately available to remote sites where computing resources may be available. As one step in this direction, in 2008, both event generation (Monte-Carlo) and simulated event reconstruction passes have been centrally managed using standard Grid interfaces for job submission. Ultimately, STAR intends to become fully grid capable and all Tier centers in STAR are now required to provide Grid based access for opportunistic use of resources and all sites are federated within the Open Science-Grid (OSG) infrastructure and project (wherever possible). For reasons which will be clear below in the discussion of future storage and CPU needs, this evolution is viewed as central to the future success of STAR computing.

An additional very important development for the STAR computing structure will be the creation of a second Tier-1 center at the Korean Institute of Science and Technology (KISTI) in Daejon, South Korea. The addition of this new Tier-1 center, as well as the evolution of the computing model for STAR are essential to continue to meet the future software and computing needs of the Collaboration. The activities carried out at the existing STAR Tier centers, as well as those planned for KISTI in the future are outlined in Table 1.

Table 1. Overview of S&C activities ongoing and planned for the future at STAR Tier 0 to 2 sites; "XX" denotes an activity for which the indicated institution is the one primarily responsible within the Collaboration

| | BNL | PDSF/NERSC | KISTI | Tier-2 |
|---|---|---|---|---|
| Prod of Raw Data | XX | | X | |
| Monte Carlo Sim | X | X | X | X |
| Embedding | | XX | | |
| User Analysis | X | X | X | X |
| Calibration | XX | | (X) | |

It is important to note, that although the resources available at STAR Tier-2 sites are modest, these sites strongly leverage the capabilities at the STAR Tier-0 and STAR Tier-1 sites with respect to scientific productivity due to the involvement of scientist end-users in data manipulation and handling. The enhanced scientific productivity provided by these Tier-2 centers is well documented within STAR, which strongly supports their continued participation in the S&C effort.

## Strategy to meet future CPU, storage, and network capacity needs

The increased volume of data (up to an order of magnitude) provided by a new data acquisition system (DAQ1000) combined with the needs of increasingly complex, resource hungry analyses and simulations ongoing and planned for the future, lead to the projected storage, CPU, and network needs for STAR shown in Table 2.

Table 2. Projected CPU, disk storage, and network capacity required by STAR software and computing by year. Row 3 indicates the network capacity required if a safety margin is included to allow for continued streaming of to HPSS without loss of data even without local buffering . Row 4 indicates the network capacity utilized in normal operation if local buffering is in operation and 20% of the raw, high priority triggered heavy ion data is streamed directly to a remote site in "near-real-time" during data acquisition.

| | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| CPU (KSI2K) | 3133 | 11635 | 15895 | 27404 | 20726 | 77372 | 117605 |
| Disk Storage (TB) | 892 | 1729 | 2641 | 3619 | 4021 | 5252 | 6482 |
| Network (Gb/s) | 0.6 | 2.4 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 |
| Network (Gb/s) | 0.4 | 1.3 | 1.6 | 1.5 | 1.2 | 1.4 | 1.4 |

To meet these needs within the funding guidance of the BNL mid-term plan (Table 3), careful optimization of the use of resources is required. Specifically, within the guidance provided by Table 3, a key question which arises is how to optimize the resources devoted to disk storage versus the resources devoted to additional CPU as these are strongly coupled.

Careful consideration of how best to diagonalize this matrix of competing needs in the limit of finite resources leads to several strategies central to STAR's future plan for software and computing.

Table 3 Funding guidance from the BNL mid-term plan ($K). The rows indicate (first) the total capital equipment funding planned for the RACF by year, (second) the amount of capital funding planned for non-experiment-specific RACF infrastructure, (third), the subtotal of RACF funding planned to meet needs (e.g., CPU and storage) specific to the RHIC experiments (STAR, PHENIX), and (fourth), the amount of the subtotal in row 3 projected to be available for STAR-specific needs.

| | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| RHIC Computing | 2000 | 2500 | 3000 | 3000 | 3000 | 3000 | 3000 |
| RACF Facility | 685 | 1594 | 1295 | 1104 | 709 | 1750 | 1017 |
| RACF RHIC Expts | 1315 | 906 | 1705 | 1896 | 2291 | 1250 | 1983 |
| RACF for STAR | 658 | 453 | 853 | 948 | 1146 | 625 | 992 |

The first strategy relates to disk storage. Specifically, in order to move towards a storage model which provides a scalable IO and data access solution as well as to reserve sufficient resources to address the CPU needs of the collaboration, STAR will follow the strategic plan already made in 2005 and will continue to evolve away from centralized storage (e.g. SAN or NAS) to the solution identified involving distributed commodity-based farm storage. This distributed storage solution has been shown to provide a factor ~30 improvement in price for the STAR requirement. Following this strategy, the use of centralized storage will be further limited in the future: only that required for verification when acquired data is being written to the mass storage system (MSS) and a small amount in support of user analysis will be retained. In particular, no simulation data will ever follow a path which leads to or through centralized storage and no produced data set will be stored on centralized storage. In this scenario, the approximate percentage of centralized storage to total storage including distributed commodity-based farm storage is shown in Table 4.

Table 4: Projected percentage of centralized storage as a function of year

| Year | Cent | Distr. | % Cent |
|---|---|---|---|
| **2009** | 242 | 650 | 27.10% |
| **2010** | 289 | 1440 | 16.72% |
| **2011** | 327 | 2314 | 12.38% |
| **2012** | 350 | 3269 | 9.67% |
| **2013** | 350 | 3671 | 8.70% |
| **2014** | 350 | 4902 | 6.66% |
| **2015** | 350 | 6132 | 5.40% |

The viability of this solution has already been demonstrated. Specifically, since 2006 for user analysis, Scalla/Xrootd has been used to aggregate and access data in a scalable way. However, this approach relies on non-scalable components (a catalog and indexing of datasets) as well as partially on manual work. Therefore, before this approach is ready for the type of large scale automated application required for the anticipated order of magnitude increase in data storage, a significant effort will be required to re-write the current data handling mechanism. This effort will require a modest increase in workforce at the level of ~ 0.5 FTE in 2009 and 1.0 FTE in 2010. It will be carried out by the core S&C team at the BNL Tier-0 center.

This storage solution places priority on reserving resources to meet STAR's future CPU needs. However, even in this optimized scenario, a temporary shortfall in CPU capacity is projected in the period 2010-2011 as shown in Table 5

Table 5 Projection of the CPU which can be acquired within the present guidance versus that needed to carry out the intended STAR beam-use plan (KSI2K)

| | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| CPU required | 3133 | 11635 | 15895 | 27404 | 20726 | 77372 | 117605 |
| CPU projected | 3644 | 6630 | 15597 | 30566 | 59531 | 78392 | 121249 |

The profile of CPU required by year, as shown in Table 5, is driven primarily by details of the intended physics program of the Collaboration (Table 6). Specifically, the shortfall in 2010 and excess in 2013 result from the resource intensive full energy Au+Au run and resource-light low energy Au+Au beam energy scan planned in 2010 and 2013 respectively. Presuming the STAR run plan remains the same, some mitigation of this shortfall may still be possible if adjustment of the RHIC computing funding profile is possible. A development which will also help address this issue is the planned near-real-time streaming of ~20% of high priority triggered raw data to the new Tier-1 facility at KISTI.

STAR's future network needs are discussed in detail in the sections which follow. With regard to network capacity, it is assumed as part of the STAR computing plan that the RACF facility will provide the necessary capacity as it is needed.

## Protocol for selection of tape recording technology and use of MSS

A third strategy central to the future STAR computing plan is the adoption of a revised protocol for the number and types of files that will be archived to MSS. In the 2010-2011 timeframe, STAR plans to migrate to LT05 tape recording technology which will afford media capable of recording a factor of ~4 times more dense than at present. However, if STAR continued the protocol followed in the past of archiving the raw data as well as all

Table 6 Planned STAR datasets from 2009 -2014

| Run | Species | Purpose | Dataset |
|---|---|---|---|
| 2009 | p+p 200 GeV | $\Delta G(x)$ with dijets | 900 M events, 50 $pb^{-1}$ sampled |
| 2010 | Au+Au 200 GeV | Precision Au | 600 M events, 2 $nb^{-1}$ sampled |
| | p+p 500 GeV | First W measurements | 250 M events, 10 $pb^{-1}$ sampled |
| 2011 | Au+Au 5-40 GeV | Energy Scan | 50 M events |
| | U+U 200 GeV | Highly elliptical zone at high density | 650 M events |
| 2012 | p+p 500 GeV | Precision W | 550 M events, 150 $pb^{-1}$ sampled |
| | Au+Au 200 GeV | Heavy Flavor, RHIC II | 750 M events, 5 $nb^{-1}$ sampled |
| 2013 | p+p 500 GeV | Precision W | 550 M events, 150 $pb^{-1}$ sampled |
| | Au+Au low E | Energy Scan | 50 M events |
| 2014 | p+p 200 GeV | Au reference with HFT | 2100 M events |
| | Au+Au 200 GeV | Heavy Flavor, RHIC II | 1200 M events, 10 $nb^{-1}$ sampled |

files (DST, micro-DST, calibration, etc.) from all production passes, the cost of media in the out years would be prohibitive ( >> \$0.5M per year). After carefully considering STAR's needs therefore, the STAR computing plan calls for a revised intermediate/economic solution in which a more limited number of essential files will be archived to MSS (raw data, micro-DST's, a fraction of DST's, quality assurance files) for 1 production pass. The relative cost of this solution versus the standard solution used to date is shown in Table 7.

Table 7 Cost of tape media using the standard protocol used to date for archiving files for a run versus the planned revised conservative intermediate/economical model for archiving data to MSS (\$K)

| | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| Standard | 118 | 456 | 503 | 488 | 197 | 694 | 694 |
| Int./Economical | 30 | 115 | 134 | 128 | 46 | 173 | 173 |

## Continued participation by STAR Tier-1 and Tier-2 centers

It bears emphasis that a cornerstone of the future plan for STAR software and computing is maintaining, and if possible slightly enhancing the existing capacity and dedicated effort at the existing STAR Tier-1 and Tier-2 centers.

As noted at the beginning of this summary, Tier-2 centers play a role in the scientific productivity of the collaboration well beyond the relative amount of resources available at these sites due to the fact that scientists who ultimately perform analyses and publish results constitute the main source of workforce at these STAR Tier-2 centers and the scientist end-users there are therefore integrally involved with the handling of processed real and simulated data. Thus, even though the resources invested at STAR Tier-2 sites might be duplicated at other higher tier sites with only a modest increase of funding, the STAR computing plan calls for continued—and if possible expanded—participation by Tier-2 centers by established by collaborating university groups.

Equally important to this plan is continued full participation and robust capability of the PDSF/NERSC center. Thus far, the *P*arallel *D*istributed *S*ystems *F*acility (PDSF), located at the *N*ational *E*nergy *R*esearch *S*upercomputing *C*enter (NERSC) at LBNL has, almost exclusively, provided STAR with supplemental user analysis cycles and embedding cycles and the STAR computing plans calls for PDSF to continue this role. As discussed above in the section on the strategy to meet STAR's future CPU and storage requirements, several effects lead to a dramatic growth in the resources required in the future as shown in Table 2. This growth has important implications for the continued robustness of PDSF/NERSC's contribution to STAR computing. Specifically, unlike the Tier-0 center at BNL which has an operating budget which includes resources to refresh from a quarter to a third of the RACF hardware each year (and address obsolescence), the PDSF/NERSC facility has no identified funding to be used for this purpose. Thus, while the details of any plan to upgrade PDSF/NERSC should be provided by the management of that facility, it is unequivocal that a key element of the STAR computing plan necessary for its success is continued full participation by PDSF/NERSC as a STAR Tier-1 center. Further, to insure the full utilization of this resource, it will be important to sustain dedicated, workforce targeted to facilitate STAR's collaboration-wide use of PDSF at the level of 0.5-1.0 FTE.

Finally, as noted above, an additional very important development for the STAR computing structure will be the creation of a second Tier-1 center at KISTI. This Tier-1 center will not only serve as a regional STAR S&C hub for STAR collaborators in Asia, but it will also provide significant supplemental resources to address STAR CPU and storage needs related to data production and analysis in the collaboration in general.

## Summary

The flood of data acquired by the upgraded STAR detector in the era of high luminosity at RHIC presents a formidable challenge for the STAR software and computing effort, which projects a dramatic increase in required resources from 2009-2015. Within the guidance provided by the BNL mid-term plan, a strategic plan capable in principle of successfully meeting this challenge has been developed and is presented in this report. The key elements of this strategy include:

- Continued evolution along the path identified in 2005 towards distributed commodity based disk storage
- Continued growth of available CPU both through effective growth of RACF capacity resulting from improved price performance over time and the addition of supplemental CPU availability from a new Tier 1 center at KISTI
- Continued full participation by the existing STAR Tier I and Tier II centers including continued resources external to BNL for user (physics) analysis roughly equivalent to those for one data production pass at the BNL Tier 0 center
- A revised protocol for prioritizing the archiving of data to tape
- Increased network capacity which reaches a value of 3 Gb/sec by or before 2011

This success of this plan is crucially dependent on two things:

- a modest 1.5 FTE increase in the "core" STAR Software and Computing workforce at BNL beginning in 2009 to re-write the existing data handling mechanism to accommodate for a scalable distributed commodity based storage solution
- Continued capacity and services at existing or increased levels at STAR Tier 1 and Tier2 centers.

The latter bullet has concrete implications for institutions which do not have recurring funds identified to address obsolescence and operational support.

If these elements of the STAR Computing Resource Plan are successfully addressed, STAR will, in the long term, be capable of meeting the future challenge of efficient, timely analysis and publication of science resulting from the vastly increased datasets provided by the DAQ1000 upgrade and increased RHIC luminosity. It is noted however, that within the present outlook, a temporary shortfall in CPU capacity is anticipated in 2010. This shortfall may possibly be addressed by a modified funding profile for the RACF, a modified run plan, a modified production schedule, or some combination of all three.

# 1.0 The STAR physics program

## 1.1 Overview

Over the next decade, the STAR experiment will carry out a forefront program in relativistic heavy ion and spin physics research. These studies will build on the discoveries of the first phase of RHIC experimentation by utilizing the increased luminosity provided by the RHIC II accelerator upgrade and by implementing new detector instrumentation strategically targeted to enhance STAR's acceptance, particle identification capability, and effective sampling of luminosity. A new Time-Of-Flight barrel jointly constructed by the U.S. DOE and the People's Republic of China will provide comprehensive particle identification for more than 95% of charged pions and kaons below a momentum of ~ 2 GeV/c and for electrons of momenta > ~ 200 MeV/c. A new Forward Meson Spectrometer (FMS) will allow a definitive search for mono-jets at forward rapidity, a conjectured signal of gluon saturation at small momentum fraction in relativistic heavy nuclei. Forward tagged proton studies utilizing updated roman pots formerly part of the pp2pp experiment will provide seminal results on the possible existence of GlueBalls and gluon-rich exotica allowed with the framework of QCD. New front end electronics for the STAR *T*ime *P*rojection *C*hamber (TPC) and data acquisition readout electronics will afford an order of magnitude increase in the bandwidth for acquisition of minimum bias data and operation for rare triggers with near zero dead time, increasing the effective luminosity useful for physics research with these probes by 60-70% in heavy ion and proton-proton interactions. A new forward tracker based on Gas Electron Multiplication (GEM) technology,—the Forward GEM Tracker—will improve the tracking precision at forward pseudo-rapidity to allow accurate charge sign determination for high $p_T$ electrons and positrons from W decay. A proposed Heavy Flavor Tracker will afford space point precision of ~ 10 μm near the interaction vertex, allowing efficient detection of secondary decays of particles containing charm and bottom quarks. These improvements signal the beginning of an exciting new era of next generation RHIC experiments focused on precision determination of the properties of the new form of matter discovered at RHIC.

One research avenue which will profit enormously from the second generation capability afforded by the STAR upgrades will be the study of multi-particle correlations. Specifically, the high statistics datasets afforded by the DAQ1000 upgrade combined with the comprehensive particle identification capability provided by the Time of Flight (TOF) upgrade will occasion a sea change in our understanding of how correlations and fluctuations at the quark-gluon stage manifest themselves in final state observables. Uranium beams made possible by the EBIS upgrade to the RHIC accelerator will extend the scientific reach of these studies even further, increasing the energy density of the collision zone and allowing an important test of whether the magnitude of elliptic flow in RHIC Au+Au collisions is already maximal. The DAQ1000 upgrade will also allow triggered datasets (e.g. using the STAR calorimeters) which sample the full RHIC luminosity for precision studies of rare probes such as the Upsilon, J/Psi, and photonhadron correlations. Particles which contain charm and bottom quarks - an especially important probe due to their relatively large mass - will be efficiently detected using a new micro-vertex detector (HFT) with space point accuracy unprecedented in heavy ion interactions.

A second research direction, in addition to discovering the properties of the new form of dense mater created at high temperature and near zero net baryon density, will be to search for a crucial landmark - a critical point - in the QCD phase diagram at finite baryochemical potential. This search will be carried out by performing an energy scan to search for the onset of critical phenomena associated with a first order phase transition in the region of the QCD phase diagram characterized by $\mu_B < \sim$ 400-500 MeV and $T < \sim$ 200 MeV. The discovery of this critical point (if found) would confirm fundamental predictions of lattice QCD and significantly extend our understanding of the equation of state for strongly interacting matter.

In spin physics, RHIC and STAR will measure jets, di-jets and direct $\gamma$ + jet final states to place significant new constraints on the contribution of gluon polarization to the spin of the proton and the magnitude of gluon polarization as a function of momentum fraction ($x_{BJ}$). Inclusive and semi-exclusive forward meson studies using the new STAR Forward Meson Spectrometer will exploit transverse spin phenomena to investigate transversity and the importance of parton orbital angular momentum in accounting for the spin of the proton. Parity-violating W decays studies utilizing the tracking precision afforded by the new STAR Forward GEM Tracker will provide seminal new data concerning the sea quark and anti-quark polarization.

In summary, the STAR detector and the RHIC accelerator are well on the way to constructing—and in some cases completing construction of—a suite of strategically targeted upgrades of moderate scope which promise to enter in an entirely new era of fundamental heavy ion and spin studies of extended scientific reach. To capitalize on these investments, it is essential that the computing capability of the STAR experiment, now and into the future, also be strategically positioned to receive and analyze the flood of data which the upgraded STAR detector will produce. The plan to accomplish this formidable challenge is outlined in the following sections.

## 1.2 Currently available facilities and roles

The STAR Experiment at *B*rookhaven *N*ational *L*aboratory (BNL) is one of the premier particle detectors in the world. Using this device, an international collaboration of 55 institutions from 12 countries, constituting a “team” of more than 590 physicists and skilled specialists, is working diligently to understand the nature of the early universe and the tiniest building blocks of matter through research on nucleus-nucleus collisions at the highest energies achieved in the laboratory. The geographical distribution of the institutions in the STAR Collaboration is given in Table 8

Table 8: Geographical distribution of institutions in the STAR Collaboration as of 2008 - the highest percentage is from the United States/North America followed closely by Europe and Asia

| | | |
|---|---|---|
| USA / North America | 24 | 46% |
| Europe | 12 | 23% |
| Asia (China/Korea) | 8 | 15% |
| India | 6 | 12% |
| South America | 2 | 04% |

### 1.2.1 Tier model and scope, available sites and perspectives

Given the international makeup of the Collaboration, the STAR *S*oftware *and C*omputing (S&C) model has naturally evolved toward a Tier structure, similar to that utilized by other major international S&C efforts. The Tiers (Tier 0, 1, 2) are defined by the services and capabilities available at the institutions within a given classification.

- Hosting the STAR detector and experiment, BNL is at the center of all STAR data taking and it is STAR's unique Tier-0 center by definition. It provides a *M*ass *S*torage *S*ystem (MSS) for archiving all data produced and used for publication, as well as for multiple (as needed) data production passes and a one production pass equivalent level of resources for physics analysis. The BNL STAR S&C team is staffed primarily from RHIC operations support and is referred to as the "core team". Within the STAR S&C model, the Tier-0 is not responsible for providing the required resources for resource intensive simulated data needs but the BNL group's S&C team provides
    - Project management; coordination of calibration, reconstruction, simulation, and embedding
    - Technology development and engineering (e.g., tracking software, framework support)
    - User services such as user help, mailing lists, Web services, accounting, Cyber-security (new activity) and computational run support.

- A Tier-1 center is defined in STAR as a site providing persistent storage (MSS) as a local service and a site providing resources (storage or processing power) to the level of 15% or more required to perform any specific task which would otherwise need to be performed at BNL. A Tier-1 center must have dedicated staff answering to local needs and supporting local operations and maintenance. Such staff would rely on the expertise from the Tier-0 team to help coordinate and guide the local support of STAR software and the related framework. Local staff is entirely responsible however for the deployment of the STAR framework and for its maintenance as well as for keeping up with operating system (OS) upgrades and local database support. Since 2000, NERSC/PDSF has been the only Tier-1 center in STAR. Its main role initially was to provide support for generating and processing simulation data as well as to provide a supplemental resource for user analysis.

- A STAR Tier-2 site is an institution having several tens of TB of storage which can be utilized for local or regional needs for user analysis (supplemental). Tier-2 centers may be provided expert assistance from the core team to carry out their mission. In return, Tier-2 centers are expected to make unused CPU cycles available for general STAR use. Wayne State University and NPI/ASCR in Prague are examples of fully functional STAR Tier-2 centers.

STAR does not have any further definition of Tiers at the moment. With the advent of distributed computing (a.k.a. Grid computing), all Tier centers in STAR are required to provide Grid based access for opportunistic use of resources and all sites are federated with the Open Science-Grid (OSG) project (wherever possible). It is understood that Grid support from STAR's Tier-1/Tier-2 centers requires, at a minimum, additional local support, including attending to Grid operation tickets and participation in the STAR Grid meeting (one hour a week). The added benefit to STAR has been a seamless pool of resources used for Monte-Carlo simulations (not requiring additional storage resources). It is emphasized that the usability of Tier center resources and the scalability of STAR's support model for Tier-1 and Tier-2 depends centrally on the presumed existence of an adequate local Tier center workforce. Without this workforce, resources available at such centers can not be exploited, since the addition of new sites by the core team is not possible in a constant level of effort scenario. Finally, it is noted that the sustainability of the STAR S&C model relies heavily on supporting the STAR software framework on (only) a limited number of operating systems (OS). These are currently all derived from mainstream Linux distributions. Additional resources or sites with specialized hardware or OS architecture is beyond the scope of what can be integrated into STAR by the project's current workforce.

Thus far, the STAR production and analysis models have mainly relied on centralized user analysis facilities (BNL and PDSF) to provide the bulk of the analysis power for STAR collaborators and scientists. The STAR computing model has, however, been steadily evolving toward a data-grid model in which processed data is made immediately available to remote sites where computing resources may be available. As part of the Trilium project (and the Particle Data Grid / PDDG project), STAR developed a data redistribution strategy and data redistribution flow, which included redistribution of full sets of derived data analysis samples (Micro-DST) to the STAR Tier-1 facility (2004). This was extended within the same year to an effort to redistribute user-based format (nano-DST) to a potential Tier-2 facility at the time (USTC). This mode of operation—providing immediate availability of datasets to remote sites—has been shown to increase physics opportunities and shorten turn-around for conference publications. Furthermore, the data transfer pilot project with the STAR institute in Prague in 2007/2008 showed that such a distributed computing "*light weight approach*", concentrating only on the redistribution of data, could provide analysis viability, sustainability, reliability and increased local scientific opportunity by carrying out local analyses, and leveraging local resources (both hardware and human potential). STAR expects to replicate this data redistribution approach to more Tier-2 centers within the next five years. It holds distinct advantages and while the hardware resources from all STAR Tier-2 centers remain modest and could likely be replaced by a small uptick in the funding made available to the STAR Tier-0 center (BNL), such an approach would undercut the potential of the Tier-2 centers as far as the local human potential and scientific achievement is concerned. It is therefore strongly disfavored. The available experience in multiple trials to date consistently shows that, “bringing the data to the scientists”, results in greater productivity than concentrating all computing resources at a single site.

The high concentration of STAR institutions in Asia has led STAR’s S&C leadership to explore the possibility of increasing the scientific productivity in this region further by considering the formation of an additional Tier-1 center with the capability to serve as a dedicated regional center for redistribution of entire data sets to STAR institutions within Asia. One attractive site for such a facility is the *K*orea *I*nstitute of *S*cience and *T*echnology *I*nformation (KISTI) located at Daejon/Korea and its consideration has been folded into this computing resource plan for discussion.

### 1.2.2 The present role of the RACF

Presently, the *R*HIC/*A*TLAS *C*omputing *F*acility (RACF) is the core resource for STAR computing, providing resources at various defined levels in all aspects of STAR computing except for simulation. Using multiple Gigabit connections between the facility and the experimental counting houses, the raw data from the experiments is archived to the *H*igh *P*erformance *S*torage *S*ystem (HPSS) hierarchical *M*ass *S*torage *S*ystem (MSS). Once in HPSS, the data can be retrieved for reconstruction with the reconstructed data being stored back into HPSS and made available on disk for further analysis. A basic principle of STAR’s computing model is that raw data, reconstructed data, and simulated data must be stored permanently on BNL’s MSS. This requirement is in part a direct result of STAR’s bylaws requiring reproducibility of all published analysis. However, a fraction of the raw data and entire sets of quantities useful for physics analysis are regularly redistributed to Tier-1 (or Tier-2) centers. We note that STAR does not have at the moment resources to provide full redundancy (more than one site availability) for the valuable and irreplaceable raw data samples.

The RACF facility has three major components, namely HPSS, the Linux Farm and the centralized disk system. The Linux farm is mainly composed of commodity hardware (hardware acquired at a competitive price following a bidding process). It is used primarily to provide resources for the production of *D*ata *S*ummary *T*ables (DST) and Micro-DSTs for further analysis by the collaborations as a whole. The machinery steering the reconstruction job and providing job management is handled by locally produced software relying on the use of the Condor batch system (Condor is also handling user analysis processes and the farm is co-shared between the two modes of operation). The Data Carousel, also developed locally at BNL and a product of the STAR experiment’s efforts to provide efficient retrieval from mass storage, is coordinating IO request to/from HPSS. To date, the central disk has served as a storage medium for results from the most recent data production pass, making data easily accessible and perusable by users using standard unix commands. However, as discussed in a later section, the projected cost of centralized storage and an increase in the amount of data and processing demands make this approach cost ineffective in the future and this convenient storage model cannot be maintained.

### 1.2.3 Role of NERSC/PDSF in STAR

The success of the STAR effort to date has relied heavily on combining the core resources from BNL with those available at other locations for providing additional supplemental cycles for Monte-Carlo simulations, additional user analysis cycles and simulations for embedding. The latter (embedding) is an absolutely essential activity that has been carried out almost exclusively at remote sites to this point.

The *P*arallel *D*istributed *S*ystems *F*acility (PDSF), located at the *N*ational *E*nergy *R*esearch *S*upercomputing *C*enter (NERSC) at LBNL has, almost exclusively, provided STAR with supplemental user analysis cycles and embedding cycles. PDSF is a large farm of interconnected commercial processors with large disk and archival capacity. PDSF is used for computing by physicists (primarily experimentalists) working in Nuclear and Particle Physics and it has been a valuable resource since its creation in 2000. As a practical matter, users mostly login via ssh and run their jobs locally.

Normally, because there are many different analyses all searching for different signatures in the same data set, multiple requests for embedding data are made simultaneously. The embedding simulations are sensitive to the proper choice of simulation parameters and to the choice of the data into which simulated events are to be embedded. Successful embedding simulations require careful crafting of embedding requests as well as a quality assurance process and a feedback loop with Physics Working Group experts and representatives. This process is time intensive and it is not untypical for the cycle necessary for refining embedding requests to be of month long length to carry out the necessary interactive and iterative discussions. Embedding necessitates careful coordination as well as perfect allocation and balance of resources (human, storage and CPU). It is not uncommon for at least some results to remain "on the shelf" at the time of Quark Matter and other major conferences, precisely because there is more demand than available capacity for handling embedding simulations. To help address this issue, in addition to maintaining adequate hardware capability, the STAR S&C plan calls for a modest workforce (0.2 to 0.3 FTE) to be dedicated to STAR for the handling of embedding requests at sites like PDSF, serving an important production role.

NERSC/PDSF has also served as a key resource for running and processing standard Monte-Carlo simulations for STAR, first as a dedicated site-specific activity and, since 2008, as a Grid-based (STAR dedicated resource) operation integrated into a centralized production management and job handling scheme.

### 1.2.4 Prospect: resources from KISTI

As part of an ongoing effort to realize the full potential of collaborative effort in Asia, STAR approached the *K*orea *I*nstitute of *S*cience and *T*echnology *I*nformation (KISTI) Supercomputing Center in late 2007. Since then, KISTI submitted an application for membership to the STAR collaboration which was accepted by the STAR Council in 2008. A central element of KISTI's proposal to join STAR was the commitment that the KISTI Supercomputing Center would contribute to the STAR experiment in a number of areas (including computing, storage and network resources) and that it would provide a workforce of 5 FTEs, including two particle physicists. In terms of impact, KISTI brings major resources to the STAR experiment's storage (including permanent archival storage) as well as computing resources linked to Pusan National University (a STAR institution) via a fast internal network. KISTI is also directly connected to the GLORIAD network which facilitates bridging with the US.

While the exact level of resources KISTI will contribute long term remains to be finalized, the STAR collaboration expects to use the center for real data processing of the top 15-20% highest priority productions up to 2011. This will require moving entire datasets from BNL to KISTI, possibly relying on the presence of the GLORIAD network for achieving near real-time data transfer. Produced data will ultimately need to be brought back to BNL for permanent archiving, while the derived Micro-DST data sets would be re-distributed to the institutions in Korea and China, making data samples immediately available for physics analysis. A proof of principle for this type of model was a similar exercise carried out in 2004 leveraging the resources at NERSC/PDSF. KISTI has further shown interest in serving as a *S*TAR *A*sian *C*omputing *C*enter (SACC), having a scope appropriate for a Tier-1 center for STAR.

Network transfer and data processing feasibility will be the first test and demonstration of the viability of this plan, which STAR hopes to implement as early as 2009. It is noted that the level of significance KISTI may play in the future of STAR/RHIC software and computing is sufficiently high that a long term commitment secured through more formal, high level channels than the present memorandum of understanding between STAR and KISTI may be necessary.

## 1.3 Foundation of the STAR computing plan

STAR is well positioned in terms of enhanced luminosity, detector instrumentation and data acquisition capability to carry out a forefront program of nuclear science research into the next decade and beyond. A concomitant necessary to insure the success of this plan is a robust plan for software and computing which affords timely production and analysis of data leading to publication of new scientific results. For the purpose of this document, "timely" will be understood to mean that scientific results can begin to be presented at conferences and published roughly after year has been devoted to data analysis and quality assurance.

Within that year, there are several activities requiring software and computing resources which must take place:

- Detector calibration
- Validation of data reconstruction software
- Data reconstruction and *D*ata *S*ummary *T*ables (DST) production
- Physics analysis
- Production of event generator or embedding simulations data samples

The first two activities are labor intensive (as opposed to computing intensive) activities which require an active, aggressive cooperation between the core software team at *B*rookhaven *N*ational *L*aboratory (BNL) and STAR collaborators engaged in detector characterization and calibration. Data reconstruction and data summary tape production are computing intensive activities carried out, thus far exclusively, at the *R*HIC/*A*TLAS *C*omputing *F*acility (RACF) on the RHIC/STAR dedicated resources. Traditionally, the majority of STAR physics analysis jobs have also been spread equally between the RACF resources (providing resources sustaining one pass analysis equivalent) and the resources available at the PDSF facility operated at *L*awrence *B*erkeley *N*ational *L*aboratory (LBNL) by NERSC, with embedding simulation work carried out at NERSC/PDSF.

As discussed later however, particularly for latter two activities, the increased data load resulting from several factors has generated the need for resources for these tasks to become available at multiple sites if timely production of scientific results is to be continued. For the past five years especially, the emergence of smaller facilities has provided local groups resources for additional analysis passes which become more complex as the scientific program evolves. Developments in GRID computing have allowed the aggregation of such sparse resources, with the aggregate capable of being leveraged for event generator based simulation production. However, while already an important and integral part of the STAR research process, GRID-based operation and resources at smaller facilities do not yet constitute a sufficiently large resource to significantly impact the massive resource needs driven by the full spectrum of STAR computing from data production to analysis derived datasets. Hence the need for continued and even increased support at sites remote from the RACF.

There are a number of factors which potentially influence the future resource requirements for the STAR Computing plan, including.

- The STAR physics program and projected event samples
- STAR process of science (from raw to physics data)
- The addition of new detector channels and event sizes and event reconstruction times
- Number of production and analysis passes required prior to obtaining publishable scientific results
- Tape recording technologies

- Choice of storage cost effective solutions

Each factor will be reviewed in the next section.

## 2.0 Factors which drive resource requirements

A useful projection related to STAR's needs can be derived from an estimate of the volume of data that the experiments expect to take each year and actual knowledge of several well known key factors which influence these projections. Projections over a period of seven years doubtless involve substantial uncertainty however. As a simple example, STAR has developed a five year plan for use of RHIC beams, but this plan will be strongly influenced by the amount of beam time actually available during this period, as well as decisions by Brookhaven Management concerning priorities for its use. An attempt has been made below to document factors which influence the present projection concerning future resources required by the STAR software and computing plan. Where substantial uncertainty exists, the impact on this plan will be noted.

### 2.1 The STAR physics program and projected event samples

STAR has a developed a strategic five year plan to optimize the impact of the RHIC scientific program in the near future. Table 9 shows the datasets STAR plans to acquire in the next five years.

STAR is in the process of completing an upgrade for Run 9 of the TPC front-end electronics and DAQ readout chain. This upgrade will increase the bandwidth for acquisition of triggered events from its present limitation of 100 Hz to an upgraded design value of 1 kHz. This increased bandwidth is adequate to complete the scientific program discussed above for the next 5-10 years.

With the installation of the DAQ1000 upgrade in 2009, the size of the datasets that can be taken will outstrip the present downstream computing capability. This means that decisions on how to trigger effectively when acquiring such datasets will be paramount. While DAQ1000 allows for the sampling by the trigger system of nearly all of the collisions provided by the accelerator, it does not allow them all to be recorded for later offline study. Thus, great care must be taken to insure - for physics observables for which it is possible - that the STAR trigger is effective in selecting only the events most interesting for physics. This requires both adequate hardware and software capability in the STAR trigger system.

For perspective, given the resources presently available for producing STAR data at the RACF, a data set of approximately 80 M min-bias Au+Au events takes of order 1 year to produce at the RACF (calibration passes included). Scientific analysis and first publication of results follow within a few months. Clearly from the above table, without additional capability, the effort to maintain timely production and analysis STAR data will become problematic by run 10. Much but not all of the additional capability required will occur through planned regular updates of the RACF hardware, the improvement in price performance effectively growing the RACF capability. Resources beyond those available at RACF however must come from other sources as discussed below.

Table 9: STAR data sets planned to be acquired in the next five years; numbers of events (in Millions) and the beam particle cross sections (in pico-barns) are shown.

| Run | Species | Purpose | Dataset |
|---|---|---|---|
| 2009 | p+p 200 GeV | $\Delta G(x)$ with dijets | 900 M events, 50 $pb^{-1}$ sampled |
| 2010 | Au+Au 200 GeV | Precision Au | 600 M events, 2 $nb^{-1}$ sampled |
| | p+p 500 GeV | First W measurements | 250 M events, 10 $pb^{-1}$ sampled |
| 2011 | Au+Au 5-40 GeV | Energy Scan | 50 M events |
| | U+U 200 GeV | Highly elliptical zone at high density | 650 M events |
| 2012 | p+p 500 GeV | Precision W | 550 M events, 150 $pb^{-1}$ sampled |
| | Au+Au 200 GeV | Heavy Flavor, RHIC II | 750 M events, 5 $nb^{-1}$ sampled |
| 2013 | p+p 500 GeV | Precision W | 550 M events, 150 $pb^{-1}$ sampled |
| | Au+Au low E | Energy Scan | 50 M events |
| 2014 | p+p 200 GeV | Au reference with HFT | 2100 M events |
| | Au+Au 200 GeV | Heavy Flavor, RHIC II | 1200 M events, 10 $nb^{-1}$ sampled |

For projection purposes, we assumed a run would be possible in 2015 with a similar requirement profile to that in 2014.

## 2.2 From raw data to physics results, details and quantification

The process of transforming STAR's raw data to publishable physics requires several steps that will be briefly enumerated below (all are relevant for the following discussions). Specific mention of the storage requirements at each stage within the context of the overall model will be given in each section.

### 2.2.1 Real data handling

The STAR data acquisition system streams raw event data which then needs to be "reconstructed" into physics usable quantities. This "reconstruction pass" is handled by a single reconstruction framework (a.k.a. root4star), a software designed to handle event reconstruction, simulation and user analysis. The reconstruction pass transforms the DAQ file format into *D*ata *S*ummary *T*ables (DST) and other products (quality assurance histograms, Micro-DSTs, event tags, etc...).

The process of event reconstruction is by nature an iterative process as it requires several external sources of information such as detector calibrations as well as bootstrapping procedures for quality control and consistency checks. Calibrations rest on the knowledge of monitored quantities such as the run conditions (beam energy, species, collider parameters such as luminosity which is relevant for the level of background or beam-pipe events or the rate at which events "pile-up") as well as detector environmental conditions (temperature, gas composition, gas pressure, voltage). The second type of parameter is recorded by the STAR online computing infrastructure database; the former is monitored and recorded by *C*ollider *A*ccelerator *D*epartment (CAD) and collected by the experiment in an experiment-specific database. The calibration and quality assurance process takes several steps:

- During data taking, a fraction of the events are analyzed online for a quick signal analysis for quality assurance and quality control (a.k.a. online QA). This stage does not influence the outcome of this plan although the early detection of problems is absolutely crucial to avoid wasting unnecessary effort later to correct problem data (if it can be corrected at all).

- As the data is streamed to mass storage, a controllable fraction of the data is used for immediate event reconstruction purposes (a.k.a. FastOffline). This process is an important step for calibration convergence since the result of the reconstruction is immediately available (in an integrated all-detector sub-system format). As the run evolves, calibrations are carried out progressively and re-injected as part of the continuous FastOffline process allowing verification and quick convergence of the calibration procedure. This is useful for detector sub-systems which are not too dependent on the *T*ime *P*rojection *C*hamber (TPC) tracks; detector sub-systems such as the *T*ime *O*f *F*light (TOF) and the *E*lectro*m*agnetic *C*alorimeters (EMC) take advantage of this process. However, while FastOffline allows for immediate and "as-we-go" calibration of quantities such as the drift velocity (determined by dedicated laser calibration runs), the processing and analysis of several distortions are not put in place by this process and fine grain off-line calibration is typically necessary.

- By mid-run, the process of fine grain detector specific calibration starts. This step allows for processing the fully convolved set of known TPC related distortions. It also affords final passes for verification of other calibrations (based on sampling the entire run period, with emphasis on covering all RHIC fills) and allows for determining the energy loss (dE/dx) which, by itself, is an important step toward particle identification.

Pre-2007, the verifications of the dE/dx modeling alone required producing and analyzing 10% of the entire data sample. However, longer fills and more stable run operation with longer runs have relaxed this need. Today, the overall level of resources required for the final and specific calibrations (beyond just the dE/dx pass alone) is of the order of 10% of what would be needed for full production of a year's worth of data from a typical run.

FastOffline, a round-the-clock continuous process during the run staffed with Quality Assurance personnel as part of a regular shift, is an adjustable resource consumption process (a portion of the farm is given to FastOffline as high priority).

Table10: Summary of CPU resource usage for calibrations, fast processing and associated quality assurance processes. The resources are given as a percentage of what would be needed to process one pass of the year's dataset; the relatively larger usage in 2007 and the decrease in 2008 are explained are exlained in the text.

| Year | Calibration | FastOffline |
|---|---|---|
| 2008 | 11% | 6% |
| 2007 | 17% | 11% |
| 2006 | 12% | 10% |
| 2005 | 12% | 10% |

The overall usage for the past four years is summarized in Table 10. The apparent rise in calibration needs in 2007 is due to additional processing and studies in relation to the STAR inner tracking detectors (*S*ilicon *V*ertex *T*racker a.k.a. SVT and *S*ilicon *S*trip *D*etector a.k.a. SSD). The drop in resource needs for the FastOffline process in 2008 is explained by data growth with two major contributing effects: (a) at constant staffing, the need for quality assurance cannot grow in proportion to the amount of data (b) larger data samples utilizing better organized and more stable trigger parameters produce more uniform datasets and decrease the need for a wider sampling of the data. STAR has started the integration of automated calibration procedures (for the TPC) and auto-QA and in the future, the steady state of FastOffline need is projected to remain between 7% to 10%.

Based on the numbers and the considerations above, it is estimated that asymptotically, the resources required for calibration and quality assurance will require an absolute minimum of 20% of the resources taken for one data reconstruction pass. This minimum amount will be used in the resource estimates below.

Finally, it is noted that Table 10 does not include the resources required for additional studies such as the ones needed whenever a new detector is integrated in the STAR setup (or the exploitation of its physics). Beyond the alignment and calibration passes, the Silicon efforts made between 2006 and 2007 for example required additional partial, but fully calibrated, reconstruction passes (up to 1/3$^{rd}$ sampling of the full dataset) for fine grain analysis and study of subtle effects (hence the need for enhanced statistics). This needs to be taken into account in our resource planning and whenever new detectors will come into play in the STAR detector upgrade planning and timeline.

## Event size and reconstruction times

Table 11 provides an estimate of the event size for the DAQ files by beam species; Au+Au and p+p values were derived from an empirical three year average of recorded event sizes assuming the parameters for triggered data samples in the future would be similar.

Table 11: DAQ event size (MB) as a function of species

| Species | DAQ size / events (MB) |
|---|---|
| U+U & Au+Au central | 1.02 |
| Au+Au minbias | 0.61 |
| p+p | 0.17 |
| p+p (500 GeV) | 0.25 |

The size and time taken to process raw data and produce derived quantities (DST files) is given by Table 12. The reconstructions times are indicative of:

1. Average time per production based on empirical experience producing the year 7 and year 8 data sets in 2008. The facility and farm at this time was based on Intel(R) Xeon(TM) 3GHz 8 core CPU nodes.
2. Averages extracted from empirical data from 2004 and rescaled according to Moore's law to estimate times spent to reconstruct Au+Au central (High Tower trigger).
3. An estimated value for the event size for a future low energy beam energy scan (no empirical data available)
4. Scaling by energy to extrapolate to p+p events at 500 GeV

Table 12: STAR event size (in MB) and reconstruction times (in seconds) per events for expected species at RHIC

| | 200 GeV | | 500 GeV | | low energy | |
|---|---|---|---|---|---|---|
| Species | Event size (MB) | Reco time (sec) | Event size (MB) | Reco time (sec) | Event size (MB) | Reco time (sec) |
| U+U minbias | 2.92 | 17.73 | | | | |
| U+U high tower | 6.46 | 38.32 | | | | |
| U+U central | 6.38 | 34.82 | | | | |
| Au+Au central | 4.25 | 23.21 | | | | |
| Au+Au minbias | 1.95 | 11.82 | 4.87 | 29.55 | 0.77 | 16.32 |
| Au+Au HighTower | 4.31 | 25.55 | 10.77 | 63.87 | | |
| p+p, L2 trigger | 1.03 | 9.33 | 2.58 | 23.33 | | |
| p+p, High Tower | 1.03 | 9.33 | | | | |
| p+p, minbias | 0.77 | 7.62 | | | | |

All data in this category is stored on BNL's MSS and those estimates will hence be used in our resource estimates and planning for STAR's storage and CPU cycles needs.

### 2.2.2 Simulated data handling

Simulated data production, a process by which event generators produce realistic simulated collision events to determine expectations for the values of physical observables, is a two step process involving a pure event generation step and a detector response simulation process. While the former requires self-contained well-known Monte-Carlo based event generator programs (Pythia, Hijing, Mevsim and similar models) which can run almost anywhere, the second step requires more sophisticated detector response simulators and hence it relies on the STAR standard framework.

We estimate that the resources (both storage and processing) needed for handling the Monte-Carlo simulations are of the order of 15% of the disk space and 10% (-0/+ 5%) of the total processing resources required for completing a one pass data reconstruction run.

For the resource planning which will follow, we will add 15% additional storage to the Tier-0 center (BNL) to account for the projected need for simulation but will not account for the resources (CPU) required for running the simulations. The CPU would need to be found from non Tier-0 facilities and hence, represents an immediate non-BNL contribution to the overall resource plan. It is noteworthy to mention that starting in 2008, both event generation (Monte-Carlo) and simulated event reconstruction passes have been centrally managed using standard Grid interfaces for job submission. This makes resources available to STAR at various sites seamless and interchangeable as far as simulated data handling is concerned.

A 15% storage impact has been added to BNL's MSS storage needs to account for storage related to simulations.

### 2.2.3 Embedding process

The embedding process is a process by which simulated events are injected (or embedded) within a background event for the purpose of accurately estimating track reconstruction efficiencies (from merging and splitting of tracks for example) and the efficiency of various algorithms (vertex finding, secondary vertices reconstruction as examples) within a realistic environmental background (incorporating pile-up, hit density, energy, species or trigger bias).

In the past, these have been scheduled sequentially by the Physics Analysis Coordinator in consultation with the S&C Leader according to Collaboration priorities. The management of this activity has however been reshaped in the summer 2006 to incorporate (a) an embedding coordinator with decision making authority (b) a closer relation to the STAR *P*hysics *W*orking *G*roups (PWG), including designated working-group-based embedding helpers (c) embedding deputies responsible for coordinating resource allocations at specific sites and helping to interact with the PWG and (d) a distributed model and structure allowing for the incorporation of additional sites if they materialize. Birmingham (UK) as example, started to provide additional resources for embedding but this institution had to drop out of STAR due to national research strategies in the UK.

Embedding has been handled to date primarily at the NERSC/PDSF facility. It requires that 5 - 15% of the raw data from each run to be transferred from BNL to PDSF as input for this process. The processing power needed for this phase is estimated to be 15% of the resources needed for performing a single pass raw data event reconstruction. Historically, neither the storage nor the processing power required for STAR embedding has been available from the Tier-0 center at BNL and the resources for this activity have had to come from external sources (almost exclusively on PDSF at LBNL).

It should be noted that storage wise, as a practical matter it has been possible to “relax” the basic computing model principle that “all produced data must be archived on BNL’s MSS” through assurances that similar permanent and resilient mass storage could be used at NERSC. If for any reasons this storage would became unavailable, this would immediately negatively impact STAR’s BNL mass storage resource allocation in a retroactive manner (all past data would need to be brought back to BNL). The assumption that NERSC/MSS will continue to be a resource with long term sustainability for STAR is therefore central to the future STAR software and computing plan as no storage impact from embedding is inferred at present inn the Tier-0/BNL storage requirement calculation.

### 2.2.4 User analysis

The resources needed to support a one pass user analysis are estimated to be of the same order as those required for a one pass data reconstruction. At BNL, resources from both pools (user analysis and any data production whether simulated or real) coexist and share the computing farm resources. In the STAR model, it is also assumed that an equivalent level of resources for user analysis is available from other sites to provide the supplemental resources necessary for one additional analysis pass. To date, the resources from NERSC/PDSF have been used to the extent possible to provide a portion of the supplemental analysis pass.

Recently, due to resource constraints at BNL, the emergence of analysis “*squeeze out*” has occurred at the main Tier-0 facility: high priority data production has taken priority in the instance of finite resources, encroaching on the share of the facility reserved for user analysis. This has caused collaborators to independently seek additional external (and local) resources outside those counted on and accounted for in the initial STAR planning for computing. In November 2006, through a survey of information from a diverse group of collaborating STAR institutions, it was estimated that the total external resources being utilized for analysis (beyond those from BNL and PDSF) was at the level of 40% of those at necessary for one analysis pass, consistent with the supplemental need for one more pass outside BNL (PDSF then provided 0.5 pass additional). We note that:

- The estimated level of need for user analysis based on the STAR model (one normal and one supplemental analysis pass) is very likely an underestimate which is below the real requirement for analyses which are becoming increasingly complex. One reason for this is that over the past years as the RHIC program has matured, topics based on two particle correlations and single inclusive spectra have been decreasing while the number of research studies involving multi-body decays and high rank correlation studies have increased. The increased need due to this cause is difficult to quantify and therefore no attempt has been made to account for additional need beyond the basic assumption of two passes accounted for in the basic STAR model.
- This discussion also underscores that resources allocated to user analysis need to be preserved to maintain the physics competitiveness of the STAR scientific program. An attempt to carry out a similar survey of external resource contributions to STAR for user analysis in April 2008 indicated at present only marginal resources from Wayne State University and a STAR Tier-2 site in Prague are available, constituting a total additional external resource level of only 14% (down by 26% from the previous survey) of one analysis pass. As the need for resources for analysis has not decreased within this period, there is concern that the present STAR plan under-estimates the needs of end users and additional efforts will be needed to identify external resources.

Apart from the storage resources for Micro-DSTs, a bi-product of data production (hence accounted for in the estimated storage needed for support of production output), there is no specific plan to provide space for user analysis at the STAR Tier-0 or Tier-1 sites. Basic codes for user analysis are archived on the Andrew File System (AFS); the RACF hardware and storage planning incorporate growth based on empirical observations and those are marginal comparing to other storage considerations.

Additionally, in 2005, as part of a standard support model, BNL started to offer to remote institutions with resources to invest the possibility to "piggy-back" in purchasing high end storage in support of their institutional research program and R&D. In 2008, the total institutional space (all centralized and aggregated using NAS/NFS) was estimated to be of a total of 32 TB (16%, two-thirds of which was used for R&D support) compared to 153 TB of total production space (78%) and 12 TB of generic user space storage (6%).

## 2.3 Effect of the addition of new detector channels, event sizes and event reconstruction times

As far as it is presently known within STAR, the addition of new detector channels is not expected to significantly impact the data size which must be handled by the computing system. This is a result of several factors:

- As new detectors such as the FMS, FGT, BSMD and future HFT are implemented, old detectors such as the SVT (and possibly the FTPCs) are being de-commissioned.
- For new detectors which are added, zero suppression will be utilized to minimize the number of bits readout for downstream processing.
- A new DAQ file format is planned to provide a more compact and hierarchical structure.
- Derived data formats (TOF, trigger structures) have been under development and include redundant information which will be eliminated. The data structures will be put under review with the objective to remove redundancies and reduce size without hindering physics capabilities.
- The size ratio between DST and Micro-DST is equal to five.

The conclusion, within modest error bars (- 0 + 15%) is that the addition of new detector channels is not expected to significantly impact the resources required for efficient and timely production of data ready for scientific analysis.

We hence believe an overall resource estimate based on a constant event size over the period considered is a reasonable assumption. Sizes for DAQ and DST (plus the derived format) were summarized in Table 11 and Table 12 respectively.

## 2.4 Number of production and analysis passes required prior to obtaining publishable scientific results

The number of passes through data sets in the future will be limited by the available resources. This will place a high premium on insuring that calibrations and updated reconstruction code are fully verified prior to the start of production, requiring an aggressive effort by STAR Collaboration members on these tasks immediately following (perhaps even during) the cessation of data taking in order to maintain timely production of STAR data. As hinted in section 2.2.1, the understanding of the response of newly integrated detector sub-systems may require additional data reconstruction passes we estimate to be at the level of 30% of the samples available. This amount of additional data production is not atypical for the delivery of reasonable quality physics datasets, and it is not inherent to the example given for Silicon detectors. Calorimetric physics requires large samples for high energy calibrations (the information could be re-injected to global reconstruction after statistics has been acquired); luminosity distortion studies, time dependent effects or even occupancy effects due to event "pile-up" and their associated R&D require systematic sampling of the data acquisition timeline. The absolute minimum number of passes expected to be necessary in any scenario is hence ~1.5 (one pass reconstruction only, 20% calibration and quality assurance and 30% partial sampling and iterative convergence for delivery of reasonable physics).

In general it is expected that, based on experience, the number of passes required to insure forefront science-quality reconstruction of the data is closer to 2. On a normal and known run scenario (no surprises, no additional studies, no new detectors), we would hence request a minimal of 2.2 passes as a baseline. In fact, the number of passes also depends on the details of the running configuration, a new configuration or new detectors requiring more passes initially to insure quality of the final production. The projected number of passes required as a function of run in order to insure quality reconstruction in the context of Table 12 is shown in 13 below.

**Table 13: Projected number of data passes versus run configuration**

| 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|
| 2.2 | 2.2 | 2.5 | 2.2 | 2.2 | 2.5 | 2.5 |

The increase in the number of passes change from 2.2 to 2.5 in 2011 and 2014 results from two different assumptions. In 2011, U+U is expected to generate higher occupancy events requiring additional resources for calibration and passes to better study vertex efficiencies and event reconstructions. The increased number of passes occurring in 2014 is attributed to the coming of the STAR Heavy Flavor Tracker (HFT) in the STAR setup in year 2012. Based on our past experience with integration of new detectors in the STAR setup, we infer a full availability of usable data for an integrated tracking on year Y+2 at which point, iterative partial productions will be needed for full understanding and convergence of alignment and calibration constants within an integrated tracking perspective. In particular, procedures for alignment have been developed and tested within high precision detector context, relying on our experience in tracking with Silicon detectors (SVT and SSD). We further assume the increase of 2.5 (2.2+0.3) passes will be needed for a period of two years, allowing for a consistent re-production of the HFT datasets.

## 2.5 Tape Recording Technology

STAR transitioned from 9940B tapes to LTO-3 in 2007 and is presently using LTO-3 tape and drive technology for all MSS storage. LTO-3 technology will likely continue to be used until LTO-5 technology becomes available (skipping the LTO-4 generation) since, given the expected tape density for LTO-5 (a factor of 2 more dense for each generation) LTO-5 will afford a factor of x4 in storage per tape at the same cost that would be required for a transition to LTO-4.

Based on the technology roadmap shown in Figure 2 and the market pricing trend shown in Figure 3, we infer that the LTO-5 technology, available as soon as 2008/2009, will become economically beneficial and reach price stabilization by 2010/2011 (depending on vendor’s availability), justifying the migration of previous storage to a new storage technology (drive and media) by the same FY10/FY11 time frame.

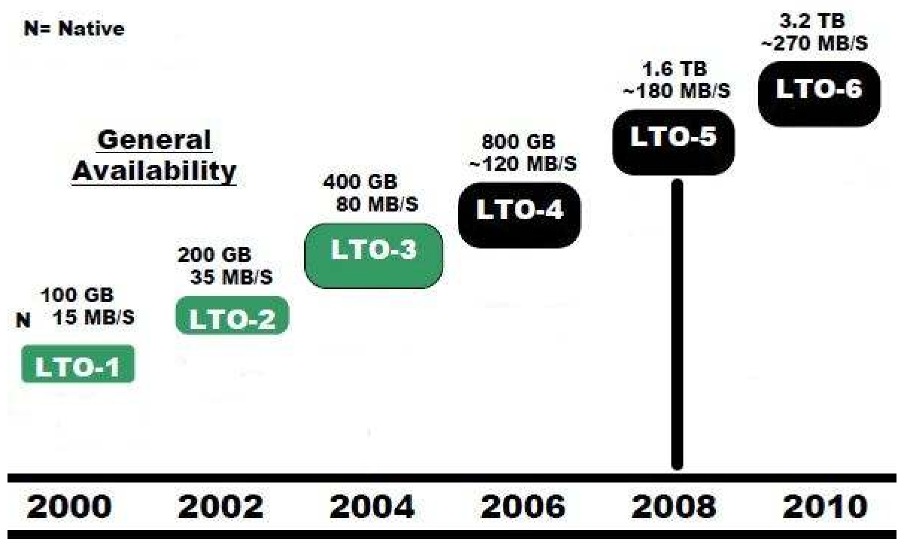


Figure 2: 2006 LTO tape drive technology roadmap taken from *Exabyte Inc*. - Drive speed and capacity are indicated for native (uncompressible) data. The black markers generally indicate technology under development. At the time of this writing, LTO-4 drives and technology are commercially available at market price.

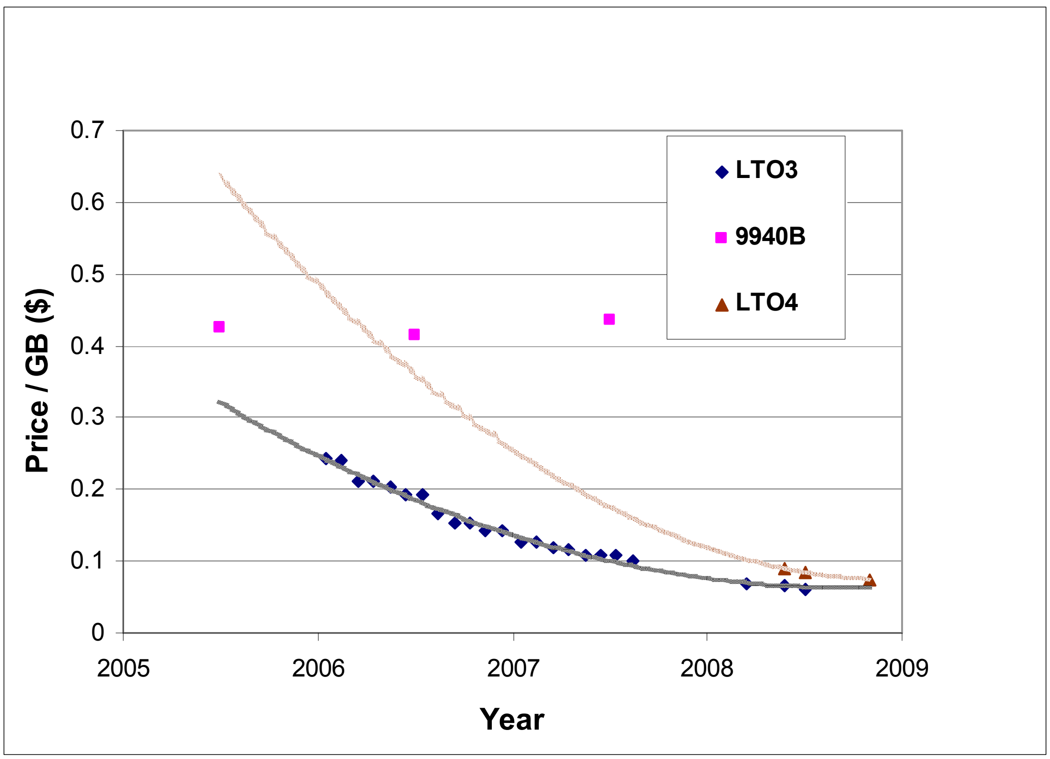


Figure 3: Empirical tape media pricing (actual cost incurred at BNL) is shown for 9940B (magenta), LTO-3 (blue) and LTO-4 generations (red). Pricing trends are represented as shaded curves for the LTO-3 and LTO-4 tape generations. Initial projections for LTO-4 (in red) are based on current market pricing studies.

At the moment, the pricing for media is at an average plateau around $0.065 per GB. This plateau is steady and is assumed to reach a minimum at $0.06 / GB at our next purchase cycle (current market price). The trend of the LTO-5 media is expected to follow a similar pattern to the one seen with the LTO-3 and LTO-4 technology with a pricing plateau at similar value supporting the conjecture of flat pricing for the timeline covered by this plan. The convergence point and value are chosen to be at $0.06 / GB.

It is noted however that unlike the transition from 9940B media to LTO media, there will be no market value for older LTO drives and media and hence, a one to one replacement with vendors (as done in 2007 and 2008 respectively) will not be possible.

## 2.6 Choice of cost effective storage solutions

Current and projected pricing for high availability centralized disk storage (network based RAID storage) compared to pricing of commodity hardware and storage such as cheap disk attached to farm nodes has shown large differences (by one order of magnitude) in the ratio of cost to benefit for the latest approach. The STAR S&C project has therefore invested heavily in the use of a model known as a distributed disk storage model since 2005. This strategy has relied to date on the strengthening and development of scalable components leveraging the Scalla/Xrootd project for its use in a physics production mode environment. No change of this model is envisioned, and STAR—which has pioneered the use of a distributed data model at RHIC—will continue in this direction for further cost savings.

## 3.0 Cost model and projections

An Excel model for facility capacity as a function of year was developed and has been used to project costs out to 2015. The model is a two step model based upon:

- Step 1: the use of the physics requirements as described in section 2.1 (Table 9). From this step, a DAQ rate per species by year is determined which drives the bandwidth requirements from the counting house to the HPSS system. An estimate of required tape storage capacities and the size of the derived data set are also determined.
- Step 2: DAQ rate, number of reconstruction passes (from section 2.4) and estimated derived data size are used in second model which calculates the CPU capacity needs, the final storage requirements and the cost derived from (a) prices realized in recent major facility procurements (b) assumptions regarding improvements in price/performance with time based on observation over the past few years (Moore's law) (c) anticipated changes in technology where they can be anticipated and (d) storage requirements driven by external (from the Tier-0) requirements (e.g., Monte-Carlo simulation output as outlined in section 2.2.2 and user analysis CPU requirements as discussed in section 2.2.4).

This two pass model provides a double blind verification and self-consistent check (the DAQ rate should lead to storage capacity requirements for the raw data similar to what is estimated from the projected number of events). All projections are believed to have a systematic error of approximately 15%.

In doing the modeling for both steps, it was assumed that the STAR Physics requirements as anticipated and described in Table 9 will be satisfied based on a run scenario of 10 physics weeks for each species that is run in a given year except in 2009, when it has been assumed there will be a 12 week, one-species run (p+p). In all estimates, the overall duty factor includes an assumed machine efficiency of 40% and an experimental duty factor of 85%.

## 3.1 RHIC Mid-Term Strategic Plan, RCF funding and availability for STAR

The choice by STAR to move toward a distributed storage model as described in section 2.6 was highly motivated by two factors: (a) the need for a scalable IO and data access solution and (b) funding constraints and the need to concentrate resources on buying the necessary CPUs.

The funding profile proposed in the RHIC mid-term strategic plan (February 2008) was taken as guidance to further evaluate STAR's flexibility and refine the model for the period extending to 2015. The funding guidance contained within the mid-term plan is summarized in Table 14. The funding in 2014 and 2015 was assumed to be constant at the level of the guidance indicated for 2013.

Table 14. Funding guidance from the Feb 2008 BNL mid-term plan ($K). The rows indicate (first) the total capital equipment funding planned for the RACF by year, (second) the amount of capital funding planned for non-experiment-specific RACF infrastructure, (third), the subtotal of RACF funding planned to meet needs (e.g., CPU and storage) specific to the RHIC experiments (STAR, PHENIX), and (fourth), the amount of the subtotal in row 3 projected to be available for STAR-specific needs.

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| Funds (k$) | 2000 | 2500 | 3000 | 3000 | 3000 | 3000 | 3000 |
| Total facility cost | 685 | 1594 | 1295 | 1104 | 709 | 1750 | 1017 |
| Avail. to the exp. | 1315 | 906 | 1705 | 1896 | 2291 | 1250 | 1983 |
| Avail. to STAR | 657.5 | 453 | 852.5 | 948 | 1145.5 | 625 | 991.5 |

The above funding levels do not include the cost of the tapes for HPSS storage since to this date, the purchase of tapes has been funded from the experiment's operation budget.

## 3.2 DAQ rates and bandwidth availability from the counting house to the RACF

Based on the requirements summarized in Table 9, we derived a raw data size for both species proposed for each run in each year. Folding in the overall number of physics weeks, we derive the DAQ data rate requirement by year shown in Figure 4.

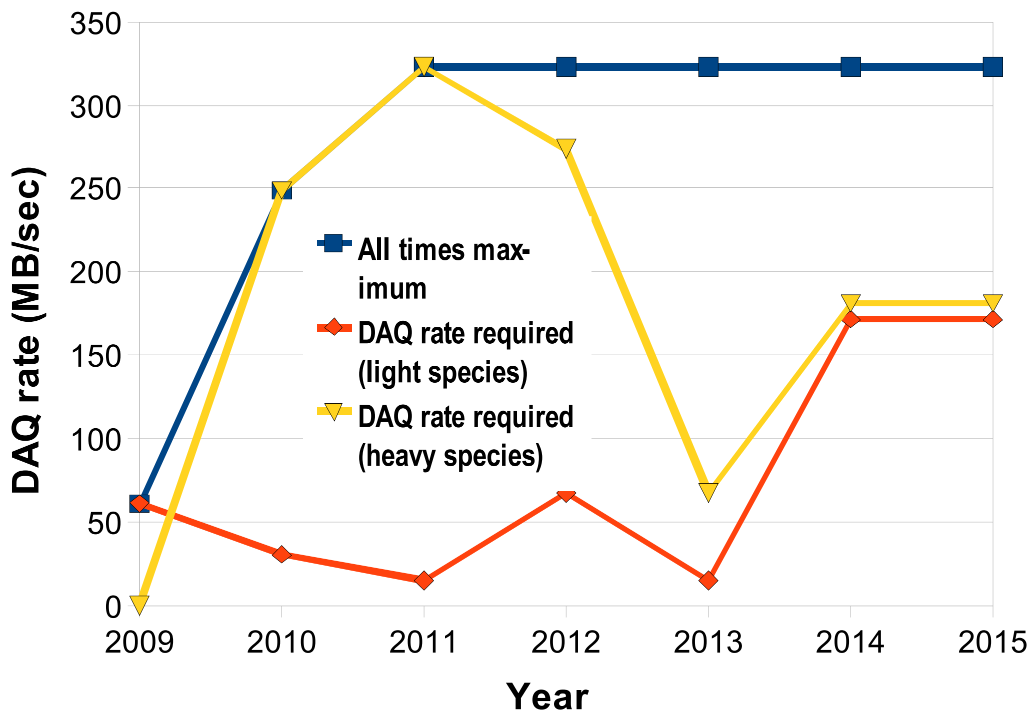


Figure 4: DAQ rate needed to accomplish STAR's Physics program and run plan as compiled in Table 9. The yellow curve represents the rate needed to sustain STAR's plan for acquiring the samples for the heavier species (e.g. Au+Au or U+U) while the red is the rate needed to accommodate the lighter ones (e.g. p+p and light ions). The blue line shown is as an all-time maximum (data streaming with no local buffering) used as a basis to later estimate the required LAN capacity planning (Figure 5).

Figure 5 shows the required data transfer bandwidth for several scenari. All curves accounts for a 20% TCP protocol overhead as a safety margin and are derived from the required DAQ rates by appropriate conversion to Gb (for a better mapping to network requirements).

Unless otherwise specified, the use of data buffering online is also considered in the calculation of the Local Area Network (LAN) transfer rate to MSS. We assume an even transfer rate taking advantage of the downtimes (due to duty factors) reducing the overall instantaneous rate which is needed. In blue, the minimal network bandwidth needed to sustain the data transfer rate to HPSS is represented without any additional transfers between the counting house and the Local or Wide area network (LAN /WAN). The rate reaches a maximum of 1.03 Gb/sec in this scenario, corresponding to the blue curve from Figure 4 (~ 320 MB/sec spread over a 34% overall efficiency and adding 20% TCP overhead).

In green and yellow, the network bandwidth needed to sustain a 20% data transfer to an offsite facility in near-real-time is indicated for the lighter and heavier species respectively within a two-species per year configuration. This bandwidth represents the transfer to HPSS (LAN) as well as the transfer off BNL (WAN) compounded to yield an overall capacity needed for STAR's counting house networking. It is noted in the context of a possible contribution from KISTI, only heavier nuclear species seem to be of interest to Asian institutions in the collaboration at this stage (hence, the yellow curve is the one of interest). For the additional WAN transfer, a near real-time assumption is used rather than a buffered assumption since the transfer would need to avoid retrieving files from HPSS for later transfer. A mode of transfer based on restoring files from MSS would be very inefficient in the first place and would require the purchase of additional MSS drives which is beyond the scope of this proposal and present facility integrated planning at BNL. Hence, a direct stream of data from the counting house to the remote facility is foreseen.

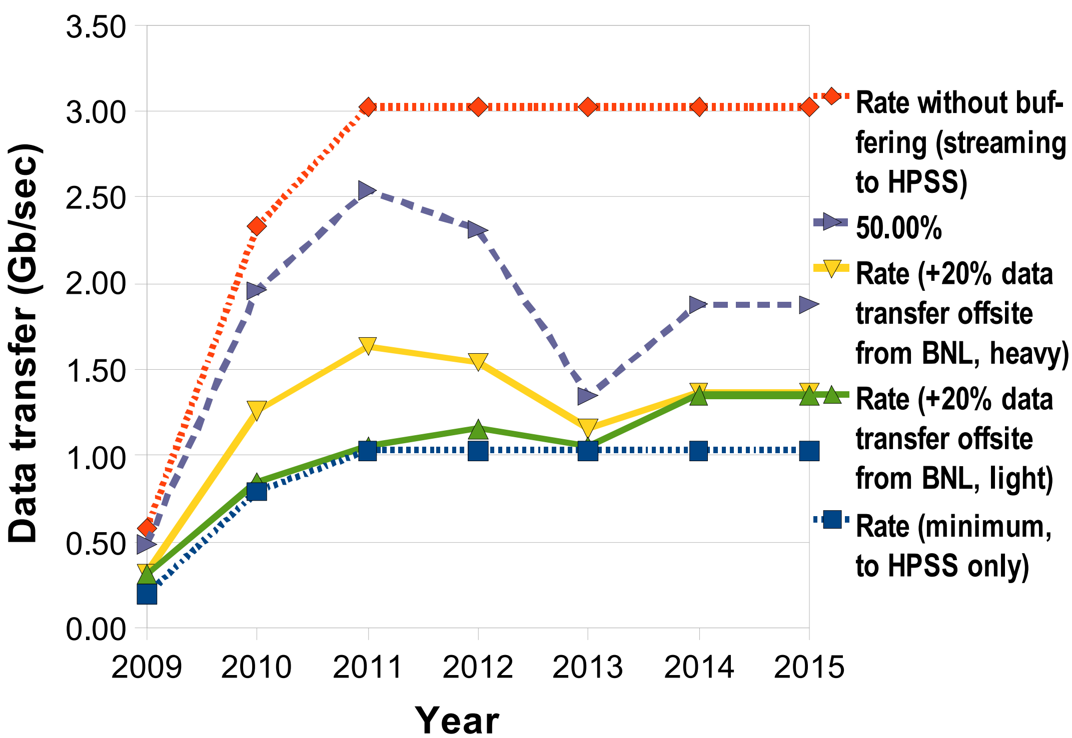


Figure 5: Network capacity needs for (respectively from bottom to top) [dot / blue] a minimal data saving capacity (using online buffering), [solid / green] a rate needed to sustain 20% light species data transfer offsite from BNL, [solid / yellow] a rate needed to transfer 20% heavy species, [dash / purple] the same with a 50% data transfer and [dot / red] the ultimate maximum (streaming to HPSS with no buffering).

We would also like to point at the fact that behind the hidden assumption underlying the scaling of the average DAQ rate per period/species according to a 20% fraction, is the possibility to sample 20% of the total acquisition stream (at any point in time) in order to select high priority data to be transferred offsite. This could be achieved by using triggered samples (event selection based on triggered particles such as the J/Psi or upsilon, or high momentum particles). If by chance no sub-fraction of the events within a given time frame can be identified for transfer but instead, the full data stream at that time is flagged as high priority, the network transfer rate capacity would need to be in excess of 4 Gb/sec for transferring the U+U samples in 2011 (this rate is considered unlikely), 3.5 Gb/sec in 2010 and 2012 and below 2 Gb/sec otherwise. The purple curve indicates the same profile with 50% sampling of the data to be transferred outside BNL to a remote facility in near-real-time and the red curve, a scenario in which the online data buffering scheme would fail, requiring instantaneous streaming of data from the data collector box at the STAR hall to HPSS. Within the current run plan, the red dotted curve represents an ultimate upper limit for the worst case (no buffering) which would, in that instance, be handled without any data loss.

STAR's position is that it should plan for this upper limit which also encompasses the requirements for off-site data transfers. For transfer offsite, streaming and sub-sampling of the data acquired will need to be done to an extent sufficient to keep the overall transfer rate below the one shown by the red dotted curve.

At present, STAR utilizes two 1 Gb/sec fiber lines to transmit data for storage at the RACF facility. As illustrated in Figure 5, an additional fiber may be needed by 2010. It is assumed this will be provided by the RACF and ITD networking teams at the appropriate time. A shared (STAR and PHENIX) 10 Gb/sec backbone and trunk is planned to be operational for Run 9. Its full availability depends on the installation of a new network switch and proper connectors at both ends. Such a line would provide for and sustain all network needs for STAR for the full period covered by this plan.

## 3.3 Projections and operations cost for tapes

Taking the projected cost for tapes discussed in section 2.5, the cost impact of storing all STAR data on BNL's MSS was evaluated. For this, four scenari were considered:

- One traditional approach, used prior to the 2007 data productions, was to save all raw data and the all files from all production passes (All raw data, all passes production DST, Micro-DST and other byproducts of production such as QA histograms and event level tags as well as output from FastOffline production and other calibrations for traceability purposes)
- A recent change based on economic considerations is to save all the raw data but only one production pass. This model impacts the flexibility of physics analyses which can be performed, since the STAR bylaws require any published analysis to be reproducible upon request. Saving only one production pass (which implies

deleting the older ones) adds a layer of management (strict enforcement of the use of a unique approved production pass). It has however been possible to achieve this model for the 2007 Silicon detector based sample.

- A more aggressive model is one in which all the raw data is saved, all of the Micro-DSTs are archived but only a faction of the DST is saved ($1/10^{th}$ or less) for quality assurance, global calibration (dE/dx) and verification purposes. In this scenario, we also consider that all products of calibration and quality assurance (a.k.a. FastOffline) should also be permanently archived on tape. This approach differs only slightly from the previous one as historically, STAR had to re-use the DST only once in eight years of data taking (and only a partial use was necessary).
- Finally, an intermediate model is one in which all passes are considered (including the calibration passes) but as for the previous aggressive model, only the Micro-DST are archived as well as a small fraction of the DST for the explained purpose.

Figure 6 illustrates the result of the four models. In this projection, we did not consider the impact of saving the data produced by simulations since it results in only a small perturbation (7% on the total storage for a two-pass processing cycle, 15% for one-pass processing) to the total cost which is not relevant for the argument which will follow.

Over the course of the 2009-2015 period:

- the traditional model would impact the experimental operations funds at the level of 3.2M$ (integral cost for the entire period with close to 700 k$ per year in the years beyond 2013)
- The economic model would impact the operations cost at the 1.5 M$ level (already a reduction factor of ~ 2.2).
- The aggressive model would reduce the overall cost to an integrated total of 700 k$. This represents a cost impact ~ 5 times less than the traditional, safe model and 2.3 less than the aggressive model.
- Finally, the intermediate model would impact the operations budget by roughly 1M$, representing a savings of a factor of 3.3 compared to the traditional model, and ~ 1.5 compared to the economic model but an increase in cost of a factor of 1.5 compared to the aggressive model.

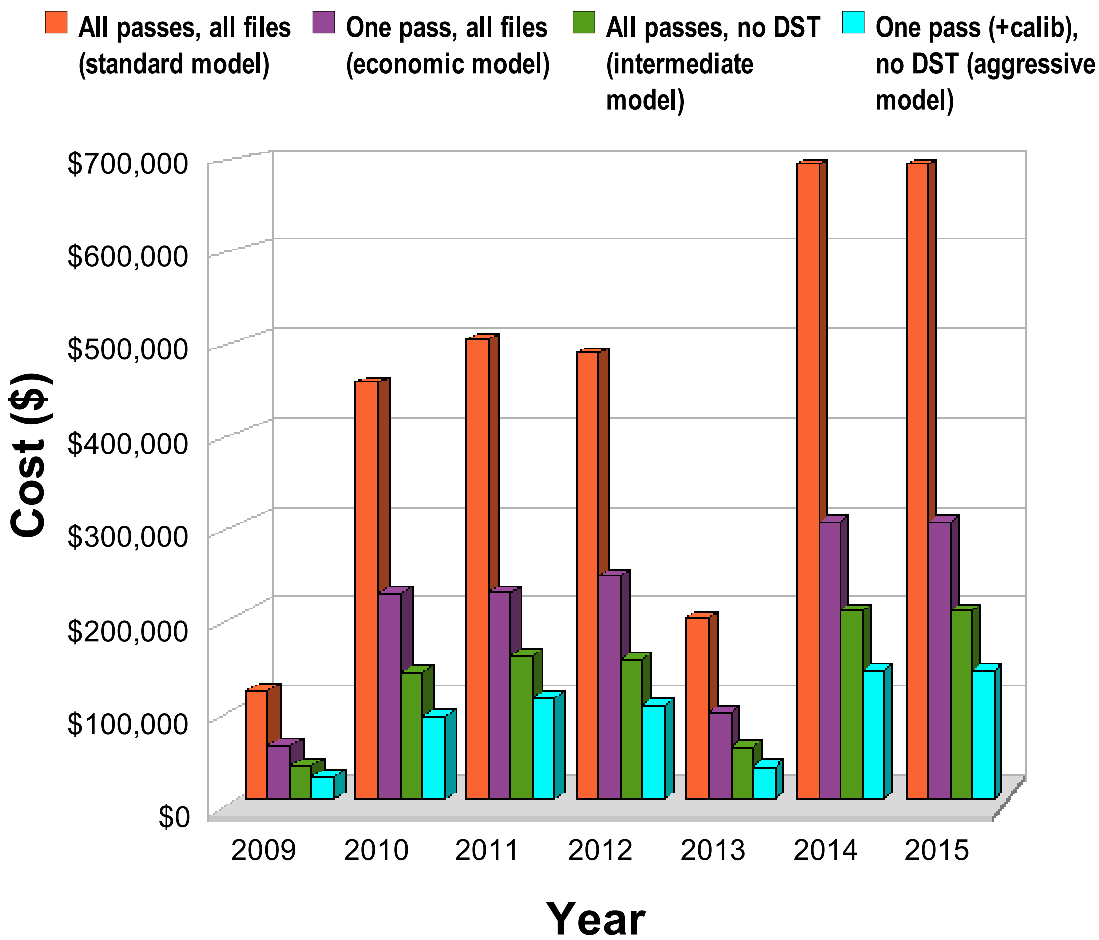


Figure 6: Tape cost impact to STAR within several data saving model assumptions.

STAR's preferred strategy is to take a middle path which combines two models: we would use the aggressive approach whenever possible and otherwise, would use the intermediate model. This strategy is viewed to be stronger because within this approach, the cost savings could potentially be devoted to add an IT professional for data handling and data management which, in the end, would provide more benefit to the Collaboration.

The final funding profile, including 15% additional storage resulting from the need to store Monte-Carlo simulation output (as described in section 2.2.2) is shown in the expanded Figure 7 in the instance of the intermediate space saving model for the cost of storage.

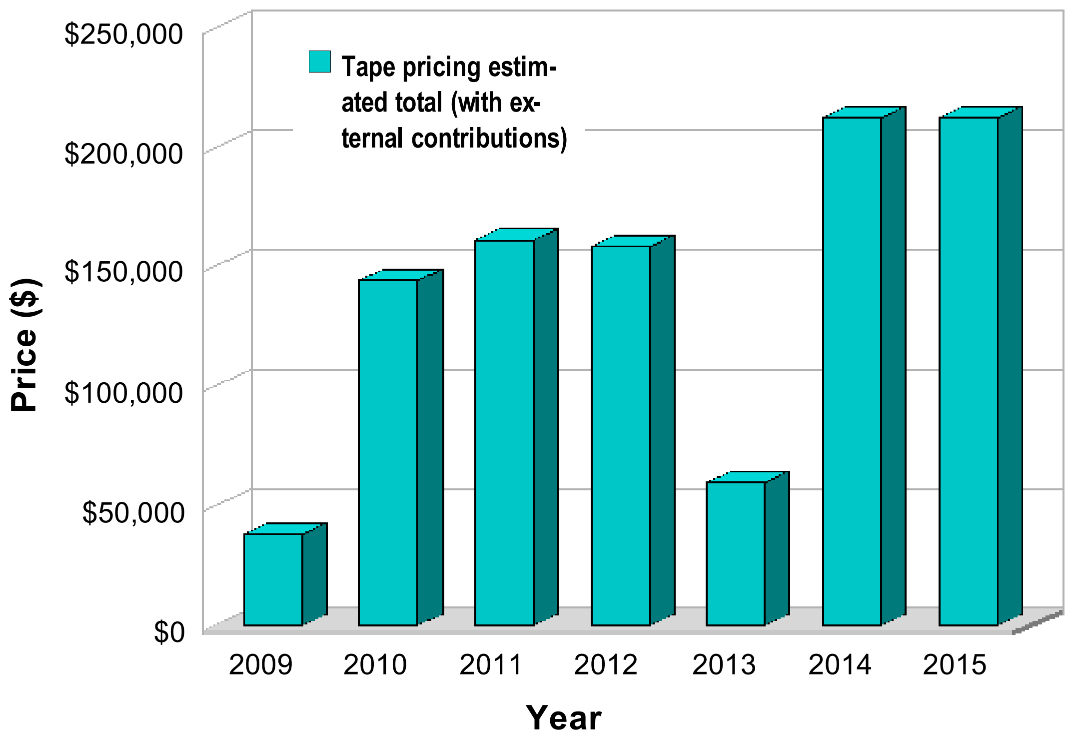


Figure 7: Cost impact on STAR operation budget, considering an "intermediate space saving model" and contributions from external sources.

In STAR's preferred model, the cost projection leads to a total integral cost of the order of 984 k$ for the entire period 2009-2015. An "aggressive cost model" would lead instead to a cost of 700k$ but as we already hinted, this model may be overly optimistic and not achievable, as deletion of all previous production passes for a given year's worth of data may not be possible. Even within the intermediate model, an aggressive effort to delete previous (unused for physics analysis) data production passes will need to be made.

We note that the drop in price in 2013 is due to the projected run configuration. This year has low size datasets for p+p and is influenced as well by the low size expected for the 50 M event Au+Au beam energy scan sample (total expected size will be of the order of barely 40 TB—DAQ + reconstruction total space).

As pointed out in section 2.5, the additional cost impact for full replacement of all media (migration of tape technologies) is, within the aggressive space saving model, estimated to be close to ¼ million dollars in 2010. This impact will be closer to ½ a million if the upgrade is delayed to mid-2012.

The cost details for the three scenari are summarized in Table 15.

Table 15: Tape cost summary. The last two columns represent the saving of the average of the aggressive model and the intermediate model comparing to respectively, the standard and economic data set archiving scheme (see text for more information).

| Year | All passes, all files (standard model) | One pass, all files (economic model) | One pass (+calib), no DST (aggressive model) | All passes, no DST (intermediate model) | Savings comparing to standard | Saving comparing to economic |
|---|---|---|---|---|---|---|
| **2009** | $118,000 | $58,000 | $23,000 | $36,000 | $88,000 | $28,000 |
| **2010** | $455,000 | $225,000 | $91,000 | $139,000 | $340,000 | $110,000 |
| **2011** | $503,000 | $225,000 | $110,000 | $156,000 | $369,000 | $92,000 |
| **2012** | $488,000 | $244,000 | $102,000 | $152,000 | $360,000 | $116,000 |
| **2013** | $197,000 | $95,000 | $35,000 | $56,000 | $151,000 | $49,000 |
| **2014** | $694,000 | $303,000 | $140,000 | $205,000 | $521,000 | $130,000 |
| **2015** | $694,000 | $303,000 | $140,000 | $205,000 | $521,000 | $130,000 |
| **Total cost / saving** | $3,152,000 | $1,456,000 | $643,000 | $952,000 | $2,509,000 | $812,000 |

## 3.4 Storage and CPU capacities

Based on the funding profile given in Table 14 and the parameters defined from the first stage model as well as the considerations enumerated in section 3.0 for the second stage model, both storage and CPU capacities were evaluated. The model allowed for varying parameters such as the portion of storage allocated to centralized disk (Network File System, network based RAID systems) compared to the amount of standard commodity Linux based systems. The cost for CPU was based on facility provided costs for 2U, 8 core, 16 GB memory systems. The combination of memory and local storage growth was presumed to follow (at constant pricing) Moore’s law and the overall power of such a box following a SpecSI2k growth per core is shown in Figure 8. The onset of the multi-core (or many-core) era is not addressed in this resource computing plan, but it is noted that STAR, based on an embarrassingly parallel processing approach, is not yet equipped with a purely parallel framework. To take advantage of the best price/performance equipment available on the market (all trends indicates this is moving toward a many-core architecture), the STAR analysis and reconstruction framework will have to undergo a software re-engineering process for optimal use of modern computer architectures. The evaluation of the best path forward for this transition started in 2007 (multi-core era task force) with the expectation of a technically sound approach by mid-2009 (and in production by the end of 2010).

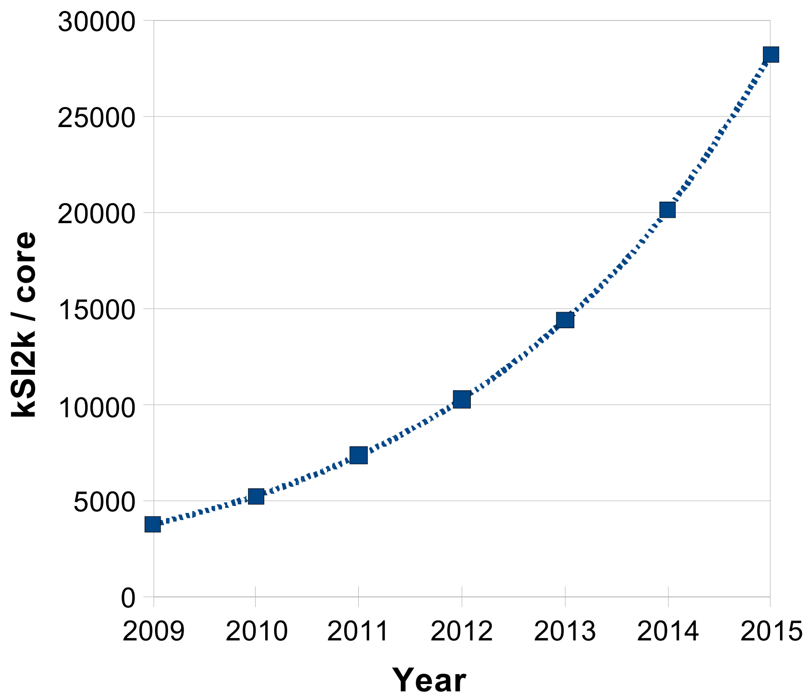


Figure 8: CPU power growth profile in kSi2k as a function of year (facility provided)

Nonetheless, within such a balance model, allocation of funds for storage reduces the availability of funds for CPU and over-allocation of CPU reduces storage capacity. Indicators and reality checks indicate the amount of CPU or storage headroom/shortfall for each year. Wherever resource constraints appeared and STAR requirements were not satisfied, storage capacity was given priority. The result of the findings is summarized below.

### 3.4.1 Central and distributed disk model capacities

Facility provided central storage (BlueArc solution or similar) is (as per 2008) of the order of 4.096 $/GB compared to market pricing for standard SATA drives (500 GB capacity) of the order of 0.13 $/ GB. For the same cost, distributed storage capacity would be 30 times greater than in the centralized storage solution. To achieve the storage capacity needed to sustain availability of derived data (Micro-DST) for analysis while leaving room for CPU capacity, the use of distributed disk space continues to appear to be the most cost effective solution.

Balancing CPU and storage, a possible profile of respective allocations in central and distributed storage is given in Table 16.

Table 16: Storage capacity (TB) proposed by year for central (Network file system, network based RAID systems) storage solution and distributed (cheap internal storage to farm nodes). Ratios are provided for guiding the argument.

| Year | Central | Distributed | %tage central |
|---|---|---|---|
| 2009 | 242 | 650 | 27.10% |
| 2010 | 289 | 1440 | 16.72% |
| 2011 | 327 | 2314 | 12.38% |
| 2012 | 350 | 3269 | 9.67% |
| 2013 | 350 | 3671 | 8.70% |
| 2014 | 350 | 4902 | 6.66% |
| 2015 | 350 | 6132 | 5.40% |

#### 3.4.1.1 Data access model: usage of central versus distributed

We noted in section 2.6 the introduction in 2005 and use in production user-analysis mode in 2006 of Scalla/Xrootd as a cost effective solution for storing and accessing data. During our feasibility study, the model started with a redundant copy of the derived data (also available on centralized storage) for the current and past year data subject to active analysis. In addition to cost effectiveness, this feasibility study showed the solution to be scalable (growth of storage as nodes would be added), providing an order of magnitude better IO than the central solution used at the time with little to no change to the user code needed to access the data in such virtual storage.

The downsides were rooted in several factors such as;

- The virtualization of storage: to date, not all physicists have adopted the notion that files could be accessed but a standard file listing may not be possible other than consulting an external catalog
- In early 2007, STAR attempted to enable dynamic disk population, a model by which Scalla/Xrootd would restore requested (but missing files) from MSS onto its storage space. While attractive in theory (as space would re-configure itself upon user demand) this approach had severe practical issues:
    - Users would request large datasets exceeding the storage space available, causing excessive deletion and replacement of data by MSS-based files. This constant streaming reduced performance to a stall – jobs would wait infinitely for files waiting in a long queue of requests and eventually would timeout. Processing efficiency showed 15% to 20% job loss.
    - The file size of the Micro-DST was shown to be too small to take full advantage of the tape system's peak performance, further reducing the ability to provide a fast and responsive file retrieval mechanism from MSS

to live storage. Work is underway to resolve this issue (at its source – the DAQ file size).

- For the data access model to work, since Scalla/Xrootd file access would prevent access to data when there is a high load on the farm nodes hosting the data, the data replication (or redundancy) needed for a fully functional system is estimated to be an additional 25% of the dataset which will need to be replicated at least twice and 10%, three or more times. This estimate is based upon monitoring the dynamic file restoration model. The actual storage needs for a dataset of size N is a storage space of N x1.35 at a minimum. Since Scalla/Xrootd is based on accessing the least loaded node having a copy of the requested file, this factor is a minimum; a larger number would help improve the data access performance.
- Since not enough storage was available as distributed storage, by mid-2007 STAR's storage model required strict control of which datasets would be approved as available on live storage. The impact of this policy was additional workforce required for data management drawn from a constant level of effort core software team. This approach has remained in effect and led to regular polling of the physics working groups to identify the "hot" datasets accompanied by semi-automated bulk dataset restoration of files from MSS to live storage.

Up to 2007, the central storage was used to provide access to at least one pass of data production, and to provide 15% additional storage space needed to buffer the product of Monte-Carlo based simulations as well as 15% for calibration and fast reconstruction studies.

This model changed in 2008 with the centralized disk being used strictly to buffer the result of ongoing data productions, Monte-Carlo simulations (SRM/gridftp buffer space of entire datasets before moving into MSS), FastOffline (15 days retention time with automated deletion and MSS saving), and test data production for calibration purposes or ongoing local real data production (with finite life times and contained within a few TB of dedicated and "recyclable" storage space):

For real-data handling, a temporary copy is made on central disk space to verify that the result of data production has been safely stored to MSS. The comparison process is automated. It relies heavily on file registration in STAR's File and Replica Catalog with space being released as files are checksum-ed and their presence in MSS is confirmed.

Upon writing this document and studying the resource model and accounting for the buffering of the Monte-Carlo outputs (a Grid based operation) it was realized that the cost impact of allowing those full Monte-Carlo datasets to be saved on centralized disk before Cataloging and saving into MSS was significant enough that it could potentially cause a serious CPU shortfall over the 2010, 2011 period with a bare minimum available in 2014. Therefore, the optimistic approach was taken of planning for simulated data output to be handled similar to real-data handling in a fully automated manner. The projections and target goals are summarized as follow:

Table 17: Percentage of Monte-Carlo outputs stored on centralized storage within a year - the numbers represent a target goal.

| Year | *2008* | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|
| Fraction buffered | *100%* | 67% | 47% | 33% | 0% | 0% | 0% | 0% |

In other words: currently, no single entire real data sets are available on centralized storage anymore and by 2012, none of the simulated data should transit through central storage.

### 3.4.1.2 Consequence on the computing model

Within the storage profile showed in Table 17, the following assumptions were used:

- Up to 2009, there will be no change in the centralized storage usage and the growth from 2008 is assumed to scale as a function of the CPU power (more CPU, faster IO). Modest changes in the model for the Monte-Carlo data retention (1/3$^{rd}$ would need to be streamed to MSS and deleted from live storage) do not require large changes in the infrastructure; rather only a solid cataloguing scheme for simulated files.
- By 2010, the central storage model will change. Growth includes only the increased need due to the remaining Monte-Carlo simulation buffering. The hidden model assumption and related change is an assumed smaller retention time for datasets from the real data production stream; that is, faster cataloguing, comparison with MSS, and handling of additional data transfer of missing files from MSS comparing to central storage. The required overall gain in speed for those operations comparing to the 2008 performance (a factor 2) will require some of the core software group’s activities to focus on database and catalogue performance in 2009.
- In 2011, the storage model will evolve further. Leveraging distributed disk space, the buffering will need to use the farm’s local storage. To make this achievable, the core software group’s activities will need to include a complete re-write of the current data handing mechanism. For one thing, the files will need to be deposited directly into the data aggregator (Scalla/Xrootd) namespace. On the other hand, since the cataloguing paradigm will need to include access from many nodes (comparison to MSS is still needed hence a catalog will still be needed), scalability of the number of clients (3O) will need to be addressed and fully functional for this mode of operation to occur.
- By 2012, the centralized storage space is expected to remain constant for the rest of the period. The last hidden implication is that all handling of transient data would utilize the distributed space. No simulated data will ever reach the central storage.

Provided STAR can address (with proper staffing and priorities) the development and scalability studies needed to achieve this plan, there are no known technology road blocks foreseen to achieve STAR's roadmap for the storage model change. It is noted that STAR does not currently have the dedicated support personnel required to fully benefit from a distributed storage model and a-fortiori for the success of this transition, some modest, targeted increase in staffing will be necessary.

### 3.4.1.3 Distributed storage adequacy

Within the proposed storage profile as given in Table 17, it is possible to consider relaxing one of the potentially most constraining impacts on physics production from the past distributed disk model: imposition of strict control of approved datasets on live storage rather than allowing, in any year, any data to be analyzed at any point in time.

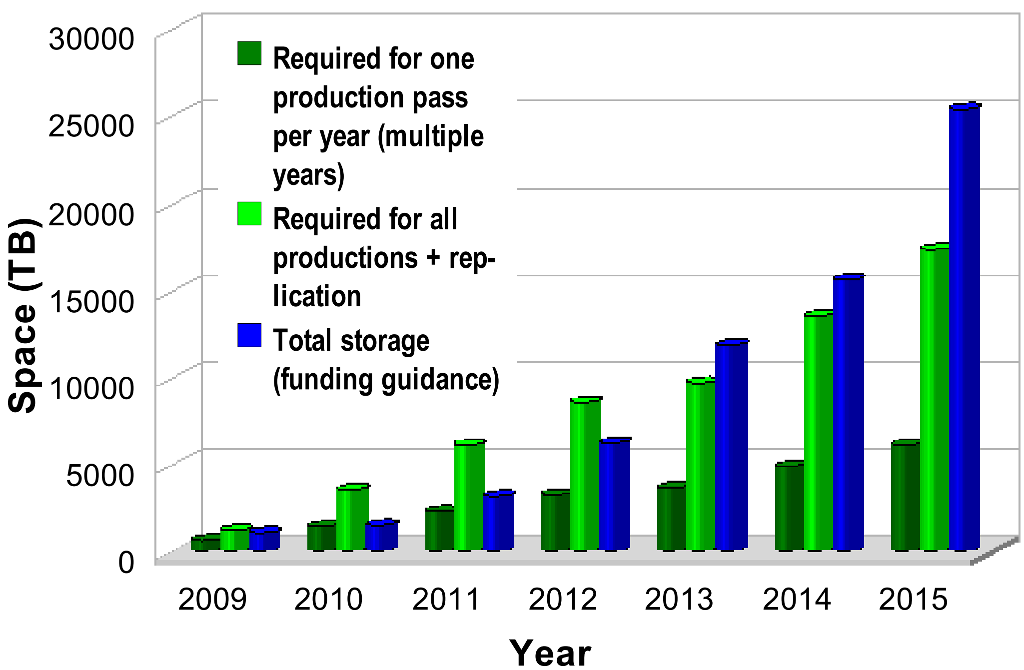


Figure 9: Storage capacity needed for two different storage models compared with acquired distributed storage.

Figure 9 illustrates the feasibility of using distributed storage in this manner. In blue is shown the storage STAR would acquire within the current funding guidance. The way in which CPU allocation drives the distributed storage allocation will be discussed in section 3.4.2.

To better estimate if the storage projected in this model will be sufficient and whether latitude in providing flexible data access will be possible, two additional bar charts are shown. In all estimates, Micro-DSTs and the need to store transient simulated data are taken into account. In dark green, we represent the storage needed to hold on disk one full production pass for that year as well as one pass production from the previous year's data. In this model, need for transient storage to hold simulated data starts as early as 2011. In light green, we represent the space needed to store the current data production pass and all previous real data production passes. Simulated data storage remains transient and starts from 2013 onward to limit the shortfall this additional space would cause. More importantly, this estimate includes the minimal space required to provide efficient distributed space management that is, an overhead of 1.35 to accommodate for replication across the virtual storage.

The storage profile shown by the blue bars (funding guidance profile) follows a constant increasing trend indicating a steady build-up of storage capacity. Compared to a minimal single pass production model (dark green), the acquired storage appears to exceed the required storage in all cases. It is noteworthy that although not fully practical, such a storage strategy would only lightly impact physics deliverables: STAR would still need to manage and impose a single "official production pass" per year (which may exclude the possibility to store test, R&D and other additional passes) but, ignoring the overhead of distributed space management pointed in section 3.4.1.1, at least one pass per year would be present for physics. However, the more practical and workable model which affords STAR the ability to keep multiple productions passes on distributed disk and, more importantly, account for the overhead of data management in Scalla/Xrootd (light green) shows a deficiency up to 2013. This tends to indicate that a careful selection of which datasets should reside on disk will be needed until at least 2013 (i.e., physics deliverables management overhead until 2013 will be required) as the space management overhead is required in all cases for this storage model to work. Both light and dark green storage models are summarized in Table 18 and in a differential graph which is more explicit in Figure 10.

Table 18: Relative excess/shortfall of the storage capacity model as shown in Figure 9. While all years would provide storage flexibility, a fully operational model including distributed space management overhead would not be possible until 2013.

| Year | One production for each year available for Physics | All productions available + replication |
|---|---|---|
| 2009 | 40.10% | -12.22% |
| 2010 | 5.17% | -134.71% |
| 2011 | 27.51% | -92.00% |
| 2012 | 47.76% | -37.37% |
| 2013 | 69.08% | 18.25% |
| 2014 | 68.64% | 13.41% |
| 2015 | 75.88% | 31.71% |

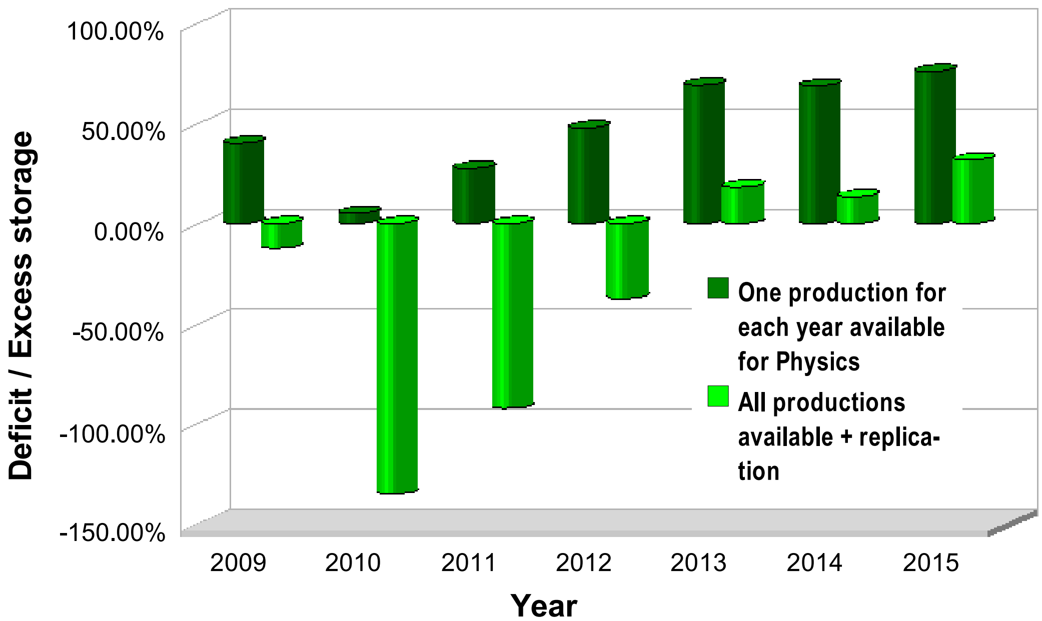


Figure 10: Relative excess and deficit of capacity for the models shown in Table 18 as a percentage of the acquired storage. The light green bars represent STAR's asymptotic storage model which is not achievable until 2013.

Overall, between the two models discussed above there appears to be flexibility to address STAR's storage requirements and while judicious choices may be needed, this higher selectivity could be combined with a strategy STAR has used in the past: utilization of storage at other sites to provide disjoint / non-overlapping dataset access by on a site by site basis. In this instance, this mode of operation would need to be maintained until 2013 (to the extent other sites (Tier-1 and Tier-2) can provide capacity) to support user analysis.

In 2014 and 2015, the excess storage of 14% and 32% respectively will be used to provide additional replication, and re-enabling of dynamic disk population (restoring files from MSS as analysis requires them) would be a safe assumption in 2014.

### 3.4.2 CPU capacity profile

Figure 11 shows the CPU profile within STAR's current plan along with the corresponding funding guidance from the BNL mid-term plan.

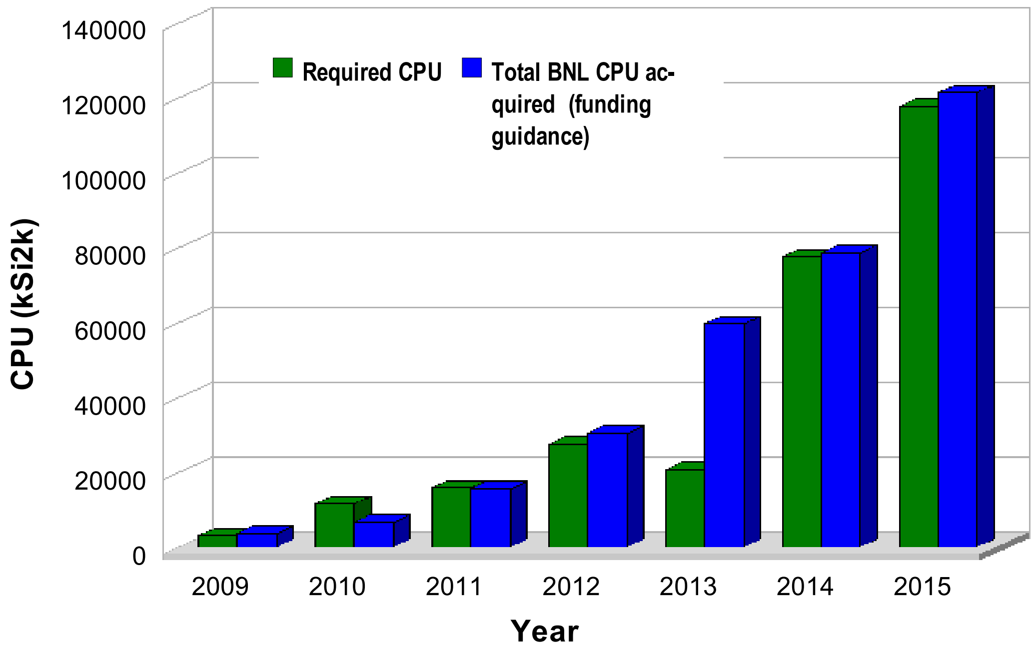


Figure11: CPU profile within proposed funding compared to the capacity required for N passes (as shown in Table 13 of section 2.4).

In blue is shown the CPU (in kSI2k units) expected to be acquired with the present funding guidance. In green, the funding required to allow STAR to perform all data reconstruction passes on datasets taken within the corresponding year is shown. As observed, a shortfall is expected in 2010-2011. During this period, it will not be possible to accomplish all objectives in time and a down-scale of STAR's objectives will be required (partial data processing). By 2012 however, the CPU allocation shows a small excess. The apparent excess of acquired CPU versus that required in 2013 is an effect similar to the one already mentioned regarding tape storage cost in section 3.3. Specifically, STAR's plan currently envisions species and beam energy combinations which result in a low demand on resources in 2013. However, the CPU growth is constant and the increase in 2013 is put to good use by 2014 where the difference is again marginal.

## 3.5 Summary and discussion of expenditures, headroom and deficits

Table 19 summarizes the relative cost allocated to central storage and CPU respectively within this plan. While central storage is a main cost driver up to 2011, as time progresses, its importance is decreased and emphasis is switched to a gradual increase of CPU and distributed storage space, a choice driving greater cost benefit and flexibility to the plan. By 2013, all cost is allocated to CPU and distributed disk.

We note that the proposed plan optimizes effective use of the full funding projected to be available within the present guidance in the mid-term plan.

Table 19: Summary of expenditures for storage and CPU, relative costs, and CPU headroom and deficits in both absolute and relative terms

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| Cost of Central Disk | $457,138 | $169,882 | $136,192 | $70,638 | $0 | $0 | $0 |
| CPU Cost | $199,506 | $282,228 | $715,302 | $875,880 | $1,143,510 | $622,848 | $987,798 |
| Total Cost | $656,644 | $452,110 | $851,494 | $946,518 | $1,143,510 | $622,848 | $987,798 |
| Unspent funds | *$856* | *$890* | *$1,006* | *$1,482* | *$1,990* | *$2,152* | *$3,702* |
| Relative cost %tages | | | | | | | |
| %tage cost central storage | 69.62% | 37.58% | 15.99% | 7.46% | 0.00% | 0.00% | 0.00% |
| %cost CPU | 30.38% | 62.42% | 84.01% | 92.54% | 100.00% | 100.00% | 100.00% |
| %tage unspent | *0.13%* | *0.20%* | *0.12%* | *0.16%* | *0.17%* | *0.34%* | *0.37%* |
| | | | | | | | |
| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
| CPU required (kSi2k) | 3133 | 11635 | 15895 | 27404 | 20726 | 77372 | 117605 |
| CPU acquired, funding guidance (kSI2k) | 3644 | 6630 | 15597 | 30566 | 59531 | 78392 | 121249 |
| CPU Headroom and deficits | | | | | | | |
| CPU (kSI2k) | 511.6 | (5004.3) | (297.1) | 3162.0 | 38804.9 | 1019.8 | 3644.0 |
| %tage acquired to required | 16.33% | (43.01%) | (1.87%) | 11.54% | 187.23% | 1.32% | 3.10% |

As pointed out in section 3.4.2, the current funding profile will cause a shortfall of CPU in 2010-2011. This shortfall is due to a reduction in the budget available for the experiments as shown in Table 14. For the period 2010-2012, the facility incurs a high impact from the cost of storage robotics (HPSS) with the highest cost impact in 2010 (at the level of 1 M$) where, as a consequence, the STAR shortfall in available CPU is the greatest.

### 3.5.1 Moving funds, a hypothetical solution to shortfalls

A possible remediation of this shortfall with no additional integral cost for the interval 2009-2015 would be a re-distribution of funds from year to year:

Table 20: A possible alternative funding profile to mitigate the CPU shortfall in 2010-2011

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| **Funds (k$)** | 1900 | 3750 | 2233 | 2617 | 1700 | 4300 | 3000 |

As shown by the profile in Table 20, this would involve moving funds between fiscal years to reflect the profile shown in Table 21 in order to reach a CPU capacity sufficient to balance STAR's needs as outlined in STAR's baseline plan. This change would still result in the same integral amount of spending for equipment (19.5 M$) until 2015.

This scenario implies:

- Deferring 100k$ of expenditures from 2009 and to 2010
- Bringing forward funds from 2011 and 2012 to augment and meet the needs of the 2010 run (absorbing facility costs without impacting the budget available to the experiment)
- Reducing the cost in 2013 (low energy scan) for the benefit of later years and to take more benefit from the effect of Moore's law.

Table 21: Headroom and shortfall within the alternative funding profile of Table 20

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| CPU required (kSi2k) | 3133 | 11635 | 15895 | 27404 | 20726 | 77372 | 117605 |
| CPU acquired, proposed finds (kSI2k) | 3314 | 11634 | 15956 | 27632 | 41484 | 76631 | 124133 |
| **CPU Headroom and deficits** | | | | | | | |
| CPU (kSI2k) | 181.6 | (0.3) | 61.7 | 228.0 | 20757.6 | (741.3) | 6528.1 |
| CPU %tage relative to acquired | 5.48% | (0.00%) | 0.39% | 0.83% | 50.04% | (0.97%) | 5.26% |
| CPU %tage relative to required | 5.80% | (0.00%) | 0.39% | 0.83% | 100.15% | (0.96%) | 5.55% |

This solution, however, may not be practical or desirable for several reasons:

- "Borrowing" from later years to reach 3.7 M$ and 4.3 M$ in respectively 2010 and 2014 may not be possible within the global context of the RHIC budget
- This scenario may address the availability of CPU, but STAR's storage capacity would suffer since the full distributed disk model would not be possible until 2014 (comparing to 2013 as explained in section 3.4.1.3).

Finally, while such a profile might benefit STAR, this argument would need to be made within the overall context of benefit for RHIC (PHENIX, STAR, and RACF). Regardless of the final outcome, this exercise shows in principle it may be possible to address the 2010 problem for STAR by modest re-shaping of the budget plan in the intervening time.

### 3.5.2 Possible reshaping of the run plan

Another observation is that the current run plan is non-optimal from a resource allocation perspective. If the low energy occurred in 2010, the more modest resource requirements for this program would help offset the deficit in 2010. Alternatively, having a low energy scan at a later stage does not allow STAR to benefit from Moore's law and the growth of capacity from better price performance as a function on increasing time. However, although non-optimal from an S&C's perspective, the present run plan is likely realistic considering CAD, experiment's readiness and other programmatic constraints.

### 3.5.3 Reshaping of the production plan

One other possibility explored was to delay production passes (producing the second high physics quality pass the year after the data was acquired) in order to meet the deficit of resources resulting from a constrained budget. The schedule profile in this scenario is given in Table 22. Within this model, 2009 appears most problematic as the single pass processing capability within 2009 would need to accommodate the 20% requirement quoted calibration needs. Therefore, that year would not see a complete (lower quality) physics quality production pass but rather ~80% of the dataset produced within that year available for physics. Later years would at least make one full pass data processing always available within the same year, with a better quality production pass the year after). Although not optimal, this scenario is a viable strategy.

Table 22: Delayed production scenario. Within this model, STAR would not be able to produce the full cycle of data processing (that is, multiple passes within the same year of data taking) up to 2012.

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| Number of production passes | 2.2 | 2.2 | 2.5 | 2.2 | 2.2 | 2.5 | 2.5 |
| Number of passes to move to year Y+1 | 1.2 | 1 | 1 | 0 | 0 | 0 | 0 |

Within this model, a savings of 500k$ could also be achieved in 2012 with little impact (a shortfall in CPU of the order of 5%) propagating to 2014 and 2015. The conclusions drawn in section 3.4.1.3 about storage capacity would not change.

## 4.0 External contributions

In the past, external contributions have provided supplemental resources which have been necessary to handle Monte-Carlo simulated data, embedding simulations, and additional supplemental analysis passes in support of user analysis.

Such resources have been primarily provided by:

- The Open Science Grid for Monte-Carlo simulations (c.f., section 2.2.2) (both STAR and non-STAR dedicated resources)
- NERSC/PDSF for embedding simulations (c.f., section 2.2.3)
- Additional institutional resources for user analysis from STAR collaborators such as NERSC/PDSF and as noted in 2.2.4 a contribution from other smaller local institutional farms which is difficult to quantify and evolves as a function of time

A potential new prospect for STAR which may be very important is the addition of the *K*orea *I*nstitute of *S*cience and *T*echnology *I*nformation (KISTI). KISTI has been accepted as a full STAR member and has expressed its intent to provide additional processing power to the STAR experiment (section 1.2.4). The following sections elaborate how these diverse contributions impact the science deliverables within STAR.

## 4.1 KISTI

Figure 12 shows the resources required for a one pass real data reconstruction performed at BNL.

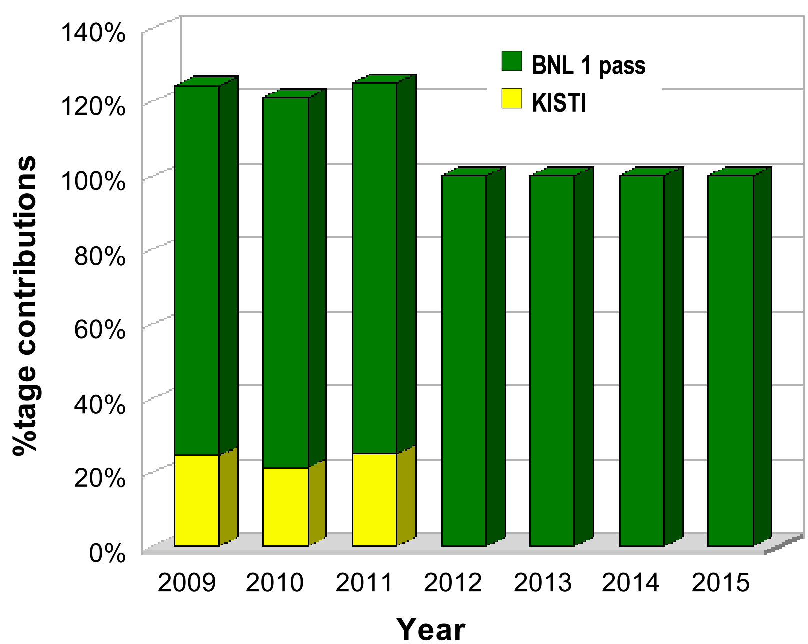


Figure 12: Relative role and contributions of KISTI comparing to a one pass production requirement.

As seen in Figure 12, KISTI's contribution remains at a constant level for the 2009-2011 (period of the currently defined pledge). It is of the order of 20% of the resources required for one reconstruction pass performed at BNL. In a normal situation, use of this 20% "boost" would be best devoted to taking care of the second highest priority data production pass, achieving first and second priority first pass data processing in half the time otherwise needed. However, in view of the shortfall in resources shown in Table 19, KISTI's resources (apart from scientific/geo-political considerations) would likely be better used to bring relief to the shortfall shown in Table 23.

Table 23: Headroom and shortfall after conclusion of the KISTI contribution

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| KISTI contribution (MOU pledged) | 240 | 800 | 1120 | 0 | 0 | 0 | 0 |
| %tage KISTI contribution to required | 7.66% | 6.88% | 7.05% | 0.00% | 0.00% | 0.00% | 0.00% |
| %tage KISTI to reco, N passes | 11.14% | 10.00% | 9.87% | 0.00% | 0.00% | 0.00% | 0.00% |
| %tage KISTI to reco, 1 pass | 24.51% | 22.00% | 24.66% | 0.00% | 0.00% | 0.00% | 0.00% |
| Total available CPU with KISTI (kSI2000) | 3884.4 | 7430.4 | 16717.4 | 30565.88 | 59530.86 | 78392.26 | 121249.3 |
| CPU Headroom/Deficit (kSI2k) after KISTI | 751.6 | (4204.3) | 822.9 | 3162.0 | 38804.9 | 1019.8 | 3644.0 |
| %tage Headroom/Deficit to total assumption | 23.99% | (36.14%) | 5.18% | 11.54% | 187.23% | 1.32% | 3.10% |

When KISTI's contribution is taken into account, the only year which indicates a shortfall is 2010. Although remote data production has not been attempted in STAR

Table 24: WAN capacity (in Gb/sec) for transferring respectively 20% and 50% of selected data to a remote facility and bringing the results back in parallel. This table differs from Figure 5 as it includes only the WAN requirements and omits the LAN transfer from the counting house to BNL's **MSS.**

| Year | 20%, data transfer only | 50%data transfer level |
|---|---|---|
| 2009 | 0.23 | 0.57 |
| 2010 | 0.93 | 2.33 |
| 2011 | 1.21 | 3.03 |
| 2012 | 1.03 | 2.56 |
| 2013 | 0.25 | 0.63 |
| 2014 | 0.68 | 1.70 |
| 2015 | 0.68 | 1.70 |

before, there is no real technology challenge for success other than ensuring proper data transfer capacity for timely data processing and handling and deploying and maintaining the STAR software stack at the remote site.

Assuming the current 20% level of data handling as described in section 3.2, and taking into account the fact that the produced data would not reside indefinitely on the remote site (the present storage pledge does not allow saving both raw and reconstructed data), the WAN bandwidth required would be as shown in Table 24. In this table, we show the results beyond 2011 for guidance as well as the requirement if the data transfer level is increased to 50%. As indicated, a 3 Gb/sec transfer capability would accommodate STAR's need in both scenari with the minimum requirement for a 20% transfer level being 1.21 Gb/sec.

## 4.2 Grid resources

Resources from the OSG virtual facility as well as resources from Amazon EC2 and the Argonne Nimbus cluster have already been utilized to a limited, but beneficial extent. Beyond storage and CPU resources, the OSG also provides valuable service towards site monitoring, troubleshooting, middleware distribution, and accounting services, all of which are leveraged by STAR to the extent possible. It is stressed that without the ability of resource virtualization, it will be difficult for STAR to make use of its resources and un-used CPU cycles available on remote sites.

STAR grid operations have thus far mainly achieved simulated data production and the STAR grid effort has sustained itself without the need for additional resources. It has been customary to leverage non-STAR resources from the OpenScience Grid (OSG) project (milestone reached in summer 2007) for Monte-Carlo event generation while running the response simulator pass on STAR dedicated resources (resources specifically allocated to and secured for STAR use with local IT staffing for deploying and maintaining the STAR framework and related components). The relative proportions of these contributions are 10-15% (mainly pre-allocated use of resources from Fermi-Grid) and 85-90% respectively from non-dedicated and dedicated STAR sites. Although small compared to the potential of full OSG partnership, OSG resources, like the one harvested on the Fermi-Grid resources, are subject to available cycles in an "on-demand" last minute allocation process. Such resources are useful as request for "emergency" processing cycles are possible through a shared virtual facility. In contrast and within STAR "shares" alone, resources are subject to an internal zero sum game on reserved and dedicated resources.

Table 25 summarizes the resources necessary for simulation as well as (for guidance) the 10 and 15% (respectively) resource levels that would normally be used from OSG STAR non-dedicated resources for event generation purposes.

Table 25: Monte-Carlo simulation needs

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| Simulation needs (kSi2k) | 195.80 | 727.17 | 908.26 | 1712.74 | 1295.37 | 4421.28 | 6189.75 |
| Portion normally estimated to OSG, non STAR site | 20.00 | 73.00 | 91.00 | 171.00 | 130.00 | 442.00 | 619.00 |

Since resources for simulations could be provided transparently by BNL whenever CPU headroom appears, all resources need to be accounted for before making a definite conclusion on the benefits of use of non-STAR pre-allocated resources. However, it is stressed that (a) the balance of resources between sites is possible with minimal staffing precisely due to the existence of Grid based interfaces and (b) while the current funding profile (with a planned increase of 0.5 - 1 M$ per year from 2010 forward) may change current thinking, the presence of Grid based resources continues to provide an important "buffer" to address funding uncertainties and site-by-site funding capacity fluctuations. It allows STAR to be more resilient to unplanned budget issues as far as simulated data is concerned and to pursue other necessary activities such as developing data transfer tools and strategies very much in demand by data intensive scientific collaborations such as those at RHIC.

It is also pointed out that from an OSG facility stand point, all resource usage made by STAR (all of which are available on OSG) is accounted for as activity by the STAR Virtual Organization and the total aggregate is presented as "OSG use" by the OSG project with no fine grain separation or distinction between STAR dedicated and opportunistic use of resources.

## 4.3 NERSC/PDSF

The use of NERSC/PDSF was summarized in section 1.2.3. In general, STAR's request for NERSC/PDSF resources is targeted to cover resources needed for the embedding process and one pass user analysis (additional analysis passes). The latter, under a flat budget scenario, has been limited by the extent of availability of resources at NERSC. Additionally, as noted in the previous section, STAR has moved all simulation production to a Grid based operation, aggregating in a seamless manner resources available at any site. Such resources may come from either NERSC/PDSF or BNL with a small contribution (opportunistic) from other STAR sites (marginal Tier-2 contributions we acknowledge but will ignore to first order).

Table 26 shows a summary of the resources missing within the context of STAR's basic plan. The first and second rows are the embedding and user analysis resource usage levels STAR's plan projects to be needed. Typically, STAR would make a request to NERSC for the resources needed to meet its full embedding needs as well as half of the user analysis needs (4th row), assuming the remaining portion needed for user analysis would be available from sparse local resources (see 2.2.4).

Table 26: External supplemental resources needed to cover for STAR full resource needs

| Year | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|
| Embedding resource needs (kSI2k) | 293.70 | 1090.75 | 1362.39 | 2569.12 | 1943.06 | 6631.92 | 9284.63 |
| User analysis, 1 pass (kSi2k) | 978.99 | 3635.84 | 4541.30 | 8563.72 | 6476.87 | 22106.41 | 30948.76 |
| User analysis, 50% (kSi2k) | 489.50 | 1817.92 | 2270.65 | 4281.86 | 3238.43 | 11053.21 | 15474.38 |
| *Typical PDSF request (kSi2k)* | *783.20* | *2908.67* | *3633.04* | *6850.97* | *5181.49* | *17685.13* | *24759.01* |
| *Monte-Carlo needs* | *195.80* | *727.17* | *908.26* | *1712.74* | *1295.37* | *4421.28* | *6189.75* |
| BNL CPU Headroom/Deficit | 511.62 | -5004.28 | -297.13 | 511.62 | 38804.88 | 1019.82 | 3644.02 |
| | -22.12 | 1817.92 | 2270.65 | 3770.24 | -35566.44 | 10033.39 | 11830.36 |
| NERSC/PDSF supplemental + user analysis (kSI2k) | 489.5 | 3635.84 | 4541.3 | 8052.1 | 3238.43 | 21086.59 | 27304.74 |
| **Summary** | | | | | | | |
| Supplemental (non user analysis) required external (kSi2k) | **0.00** | **1817.92** | **2270.65** | **3770.24** | **0.00** | **10033.39** | **11830.36** |
| NERSC/PDSF possible request profile | 489.50 | 3635.84 | 4541.30 | 8052.10 | 9564.41 | 14760.61 | 27304.74 |

Given the present BNL funding profile, some of the years covered by this plan show a slight excess in the level of Tier-0 resources (c.f., BNL Headroom/deficit row). This does not significantly alter the level of external resources called for in this plan. The Monte-Carlo base operation resource levels, as previously quantified in Table 26 must be accounted for from either BNL or NERSC/PDSF resources. In the interest of scientific opportunity and productivity by members of the collaboration, STAR S&C management believes it to be fundamental to preserve the user analysis portion from the requested NERSC/PDSF allocation and hence, the headroom projected in the current STAR plan does not reduce the external resource needed for the user analysis portion.

The supplemental resource level required (non user analysis based) is shown on the first summary line . On the last line, a smoothed resource allocation level for the period 2009-2015 is proposed which could represent a NERSC/PDSF allocation which would address STAR's need to maintain a level of remote site analysis and provide support for data processing power (embedding or Monte-Carlo). The profile is also represented on Figure 13.

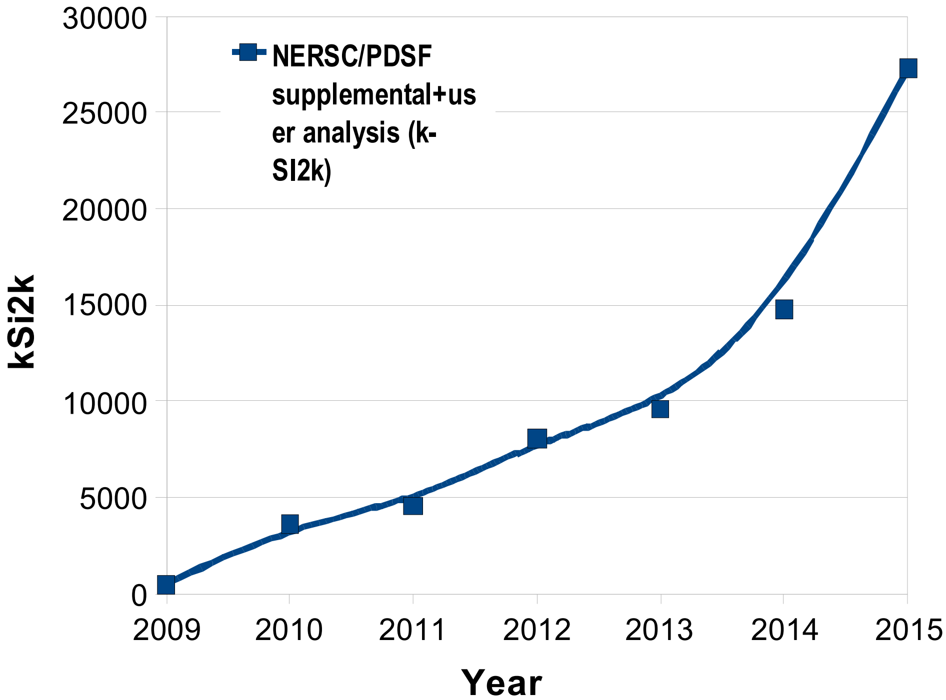


Figure13: Suggested resource profile growth for NERSC/PDSF. The line is to guide the eye and provide an additional smoothing baseline.

## 4.4 Contributions summary

Figure 14 summarizes all contributions except the source of a missing 50% one pass equivalent user analysis needed to supplement the 1.0 pass at BNL and 0.5 at NERSC. This missing portion would equate to 10% of the total resources accounted in this figure, which is well within the margin of error for these projections. It is reiterated that it is beneficial to STAR and its scientific program to have local institutional farms and resources to absorb this type of difference.

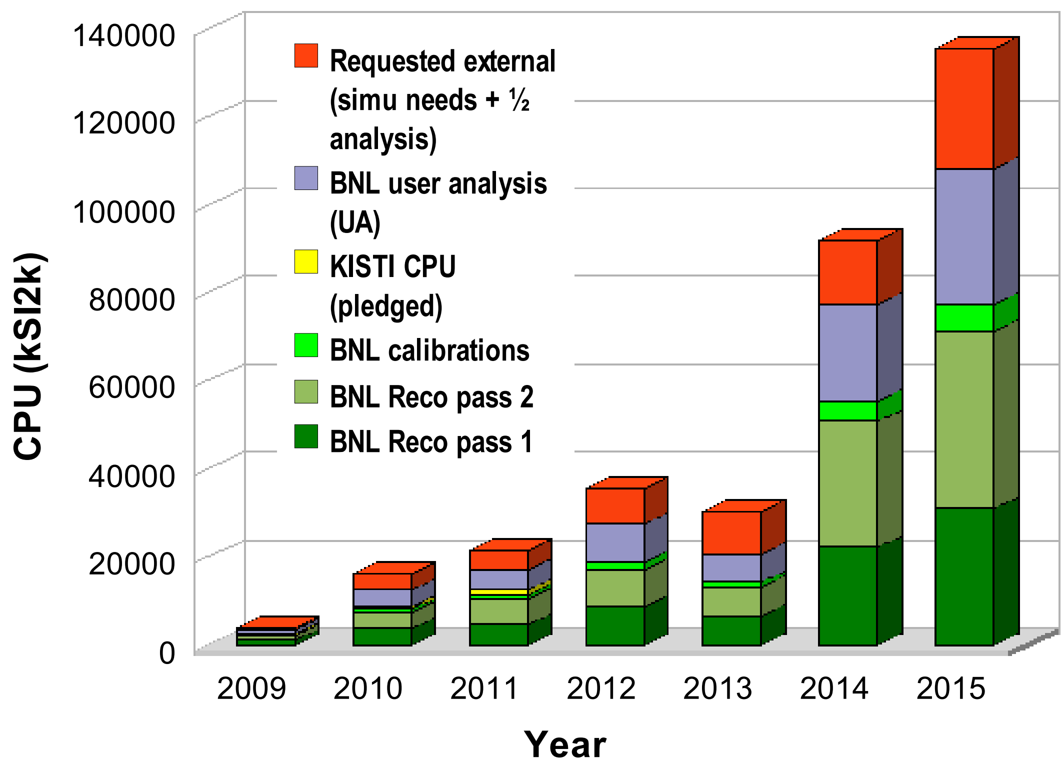


Figure 14: Relative contributions of the diverse CPU needs. The some of all green represents the portion of resources allocated to real data processing, in purple the portion needed to sustain one pass analysis over the data and in red, the resource we would request to sustain this model.

KISTI resources are only pledged to 2011 and are hence not visible from 2012 onward.

The level of external resources required within this plan is coupled to the current funding guidance in the BNL mid-term plan. Within that guidance the integral of overall resources needed in this category is expected to be at the level of 42 MkSi2k additional between 2009 and 2015.

## 5.0 Conclusions

The flood of data acquired by the upgraded STAR detector in the era of high luminosity at RHIC presents a formidable challenge for the STAR software and computing effort, which projects a dramatic increase in required resources from 2009-2015. Within the guidance provided by the BNL mid-term plan, a strategic plan capable in principle of successfully meeting this challenge has been developed and is presented in this report. The key elements of this strategy include:

- Continued evolution along the path identified in 2005 towards distributed commodity based disk storage
- Continued growth of available CPU both through effective growth of RACF capacity resulting from improved price performance over time and the addition of supplemental CPU availability from a new Tier 1 center at KISTI
- Continued full participation by the existing STAR Tier I and Tier II centers including continued resources external to BNL for user (physics) analysis roughly equivalent to those for one data production pass at the BNL Tier 0 center
- A revised protocol for prioritizing the archiving of data to tape
- Increased network capacity which reaches a value of 3 Gb/sec by or before 2011

This success of this plan is crucially dependent on two things:

- a modest 1.5 FTE increase in the "core" STAR Software and Computing workforce at BNL beginning in 2009 to re-write the existing data handling mechanism to accommodate for a scalable distributed commodity based storage solution
- Continued capacity and services at existing or increased levels at STAR Tier 1 and Tier2 centers.

The latter bullet has concrete implications for institutions which do not have recurring funds identified to address obsolescence and operational support.

If these elements of the STAR Computing Resource Plan are successfully addressed, STAR will, in the long term, be capable of meeting the future challenge of efficient, timely analysis and publication of science resulting from the vastly increased datasets provided by the DAQ1000 upgrade and increased RHIC luminosity. It is noted however, that within the present outlook, a temporary shortfall in CPU capacity is anticipated in 2010. This shortfall may possibly be addressed by a modified funding profile for the RACF, a modified run plan, a modified production schedule, or some combination of all three.