STAR Offline Computing Requirements Task Force Report
(Star Note SN0327)

G. Eppley (Rice), P. Jacobs (LBNL), P. Jones (Birmingham), S. Klein (LBNL), T. LeCompte (ANL), W. Llope (Rice), S. Margetis (Kent State), S. Pandey (Wayne State), L. Ray (Texas), I. Sakrejda (LBNL), H. Spinka (ANL), T. Trainor (Washington), T. Ullrich (Yale), T. Wenaus (BNL), K. Wilson (WSU) and N. Xu (LBNL)

March 2, 1998

Contents

1 Executive Summary

Assumptions Basic assumptions, parameters and units are given in section 2.1 and Table 1.

Raw Data Discussion of raw data size and recording rate is given in section 6.1. Raw event size for central Au-Au collisions is estimated to be 12 ± 4 MB. Data recording rate is 20 MB/s. We assume an effective RHIC/STAR year of 10^7 seconds (i.e. a duty factor of 2/3), giving an annual raw data volume of 200 TB. STAR will record $1.7 \cdot 10^7$ central Au-Au events per year at $\sqrt{s} = 200$ GeV, or the equivalent data volume (i.e. more events) for lower energies or lighter systems.

DST Production The total CPU time and resultant data volume are given in section 6.1. Analysis of a year's data set of central Au-Au events requires $4.3 \cdot 10^7$ kSi95-sec, generating a total data volume of 20 TB.

Data Mining and Analysis Annual CPU needs per general physics category for data mining and analysis are given in Table 9, and annual μDST data volumes are given in Table 10. By both measures, the soft hadronic probes, peripheral physics and spin programs have small impact on the computing requirements. Largest CPU needs are for event-by-event and prospective D meson and lepton physics, with significant needs also for hyperon, correlation and high p_T physics. Total μDST data volume is estimated to be 12 TB per year.

Simulations The requirements for the various types of simulations needed to correct data for instrumental effects and for studies of theoretical models are discussed in section 8 and the the total CPU needs and data volumes are summarized in Table 14. Total annual CPU requirement is 10^8 kSi95-sec, generating 24 TB of data.

- **Corrections for instrumental effects:** By far the largest resources are required for Geant-based calculations for estimate of background due to instrumental effects. About 1.7M full Au-Au central events passed through Geant (GSTAR), requiring about $7 \cdot 10^7$ kSi95-sec CPU and generating about 20 TB of raw data.

- **Event generators:** About $2 \cdot 10^6$ events will be generated, requiring $2 \cdot 10^7$ kSi95-sec CPU and generating about 200 GB of data. Half of this effort is common to the RHIC community and the load can be shared, whereas the other half is STAR-specific.

I/O vs CPU Tables 7 and 8 show a variation among physics categories in the ratio of CPU to input data volume of two to three orders of magnitude for both tasks. For data mining, the greatest amount of computation must be performed for event-by-event physics. We judge a task that requires greater than about 30 Si95-sec/MB to be CPU-intensive (i.e. < 1 MB/sec into a 30 Si95 processor); by this criterion, many STAR analysis tasks are classified as either CPU–intensive or extremely CPU-intensive. Taking into account common I/O-intensive tasks such as histogramming, we conclude that roughly the same number of STAR analysis tasks can be classified as CPU-intensive as I/O-intensive.

High Luminosity pp Raw data volume and DST production time for p-p analysis can be a significant fraction of that for the analysis of heavy ion data, especially in the early years of STAR when online filtering of pileup in high luminosity p-p events is not fully developed.

2 Introduction

The aim of this document is to present estimates of STAR computing needs as explicitly as possible. While many assumptions have been made, our intent is to present them and their consequences clearly enough that the result of a different set of assumptions (or better knowledge of the relevant parameters) can be followed in order to assess their overall affect on the STAR computing resources.

2.1 Assumptions about STAR Configuration

In this report, unless otherwise stated, estimates are for the fully instrumented STAR detector that will be achieved several years after RHIC turn–on, including all subsystems that currently have plans for implementation. These include the baseline Time Projection Chamber (TPC) and trigger, plus trigger upgrades, Silicon Vertex Tracker (SVT), Electromagnetic Calorimeter (EMC), and Forward Time Projection Chamber (FTPC), and a patch of Time of Flight (TOF) which is 10% of the full implementation.

The estimates therefore represent the full range of STAR physics analysis projects and the scale of STAR computing needs several years after RHIC turn–on. We do not assess the time dependence of these quantities following turn–on.

The basic assumptions for annual STAR data volume are:

- **Writing speed:** STAR will write to tape at 20 MB/s
- **Event volume:** One central Au-Au event will be 12±4 MB
- **Live time:** The RHIC year corresponds to 4000 hours. We assume a STAR/RHIC duty factor of 2/3, giving an effective annual live time of 10^7 seconds.

Under these assumptions, STAR will write $1.7 \cdot 10^7$ central Au-Au events to tape per year. We assume that for nucleus-nucleus collisions this data volume is only weakly coupled to the luminosity of RHIC for the dataset. In other words, given a data recording bandwidth, we assume that the trigger will be run loosely enough at a given luminosity to saturate the recording bandwidth.

For p-p running, $6 \cdot 10^7$ pp events will be written per 10 week running period. In this case, the data recording rate is luminosity–limited.

2.2 Methodology

The Task Force was convened by the STAR Spokesman in the summer of 1997, with instructions to carry out a reassessment of overall STAR computing needs and produce a final report by November 1, 1997 (see section A). Membership of the Task Force was drawn from the STAR membership, with members having prior experience in each of the major STAR physics analysis categories.

Each subgroup undertook a detailed analysis of a few representative analyses within its general physics category, in order to identify all significant data analysis steps and estimate the computing resources needed at each step. These consisted of both common projects, with moderate demand on computing resources, and difficult projects, with extreme demand on computing resources. Such a mix was established for some of the physics categories; for others, a single representative or all–inclusive analysis was studied.

For the current iteration of this document, a simple "run plan" was assumed, that of a full year's data taking of central Au-Au events. This run plan will never be realized in practice, of course, but it serves as a useful reference point to understand the main issues for STAR computing. Appropriate scaling of the CPU times and data volumes for a given observable to smaller systems or lower energies can then be carried out and the consequences for computing resources derived. Estimates in this document for the p-p and peripheral A-A physics programs were scaled to be comparable to the nuclear A-A programs.

Since the RCF is designed in such a way that DST production and further analysis (including data mining) are performed on different sets of processors, it was necessary to consider separately the needs for DST production and for further analysis, including data mining. It was also necessary to assess the needed simulations and calibrations.

These estimates were used to investigate the variations in requirements among the various topics, especially variation of the ratio of CPU/IO. Under some assumptions, they were scaled in order to establish the overall magnitude of the computing needs for STAR. This document presents all of these estimates.

All discussion of CPU benchmarks are based on the report "A Short Note on Benchmarks" by Thomas Ullrich (STAR Note SN0314). We use the unit of SPECint95 (abbreviated Si95) throughout.

3 Physics Motivation and Experimental Approach

3.1 Soft Hadronic Probes

Single particle spectra of hadrons up to about p_T=2 GeV/c are among the first observables that will be available when RHIC turns on. Analysis of these spectra yield understanding of freezeout conditions of the reaction, including chemical composition and radial flow. They may also yield insights into more exotic mechanisms such as Disoriented Chiral Condensates.

Because low p_T hadrons are copiously produced in high energy nuclear collisions, high statistics inclusive spectra are relatively undemanding compared to rarer probes in terms of numbers of events required and consequently the needs for computing. However, these analyses will be carried out for many different event classes, based upon global selection via total multiplicity, E_T, etc., or more detailed event-by-event selection, and many parallel such analyses may be carried out.

3.2 Hyperons

If a quark-gluon plasma (QGP) is formed in nucleus-nucleus collisions at RHIC, copious strangeness production is expected to result due to the lowering of the kinematic threshold for strangeness production in a QGP relative to the thresholds for producing strange hadrons within a hadronic gas (HG), and the different cross sections for the microscopic processes which give rise to strangeness production in the QGP and HG.

The combined observation of strangeness saturation and strangeness chemical equilibrium has been suggested as a possible signature of QGP formation. A potential problem for this simple picture is that strangeness can also be produced in rescattering after hadronization of the QGP. This is particularly true of singly strange mesons (kaons) and hyperons (Λs) since their quark content nearly reflects that of protons, neutrons, and pions. However, it is more difficult to produce $\bar{\Lambda}$ and multiply strange baryons such as the Ξ and Ω by hadronic rescattering since their production would require several collisions to build up their quark content. Since these particles are expected to preserve more information about the QGP through the hadronization phase they are of particular interest in STAR.

While the most common singly strange particles (Λ, kaons) are copiously produced and reconstructed, multistrange hadrons, especially Ω, are expected to require millions of events in order to generate spectra with reasonable statistics, and we expect that this phyics program will be one of the major users of computing resources in STAR.

3.3 Correlations

Two-particle correlations provide both geometrical and dynamical information about the source which produced the emerging particles, in addition to that provided by one particle distributions. This information involves the space-time evolution of the system (identical-particle correlations) and the space-time asymmetry of the event (unlike-particle correlations). Due to the dynamics of the expansion there is a strong correlation between space-time point of creation and the momentum of the final state particles, and studying two-particle correlation functions as a function of rapidity, p_T and particle species for a variety of event centralities will shed light on this issue, especially when used in conjunction with single-particle distributions. In addition, higher-order Bose-Einstein correlations may provide information about source coherence and shadowing that is otherwise unavailable. Such information might produce a signal should DCCs or a phase transition occur during the collision.

Kaon correlation functions have much smaller contribution from resonance decays than do pion correlation functions. The study of correlation functions for K_s^0 has additional advantages due to the absence of Coulomb effects, immunity to two track resolution effects, and the possible sensitivity to strangeness distillation effects in baryon-rich matter.

Due to the need to study two particle correlation functions in weakly populated regions of phase space for a variety of systems, correlation studies require moderate to large event statistics and corresponding computing requirements.

3.4 Event–by–Event

Event–by–Event analysis searches for nonstatistical fluctuations in particle distributions which may be attributable to unusual dynamics in the event. Possibly only a small fraction of the event population selected with the same global features, such as total multiplicity, will manifest such anomalous fluctuations. The analysis separates the unusual events into a number of anomalous event classes, with the remaining events serving as a reference pool. Each event class must then be analyzed by more traditional inclusive methods to determine the nature of the anomalous content. This includes both a high-statistics characterization of the phenomenon that resulted in the selection (in order to confirm and better understand the selection process) and inclusive studies of other event properties that do not have enough statistical power to serve directly as event selection criteria. Careful study of such a result for systematic and computational error is of course imperative.

Event selection proceeds through comparison of distributions from real events to reference distributions. The reference should contain no variation from event to event other than that due to finite number statistics, in order to compare the correlation content of each event with that of the event ensemble average having identical statistical variance structure. Techniques to form the reference sample include:

- Formation of a reference "partner" to each event having the same multiplicity and with points sampled randomly from a phase space distribution given by the mean-value distribution of the real-data ensemble;

- Event mixing, analogous to that used in HBT analysis, in which the reference sample consists of actual tracks drawn from different events having the same global characteristics.

Further physics interpretation of the event classes involves simulations based on event generators. The utility of event generators for event-by-event physics is in the study of event class *calibration*, not event *selection*. By calibration we mean the introduction of known, unusual dynamics into a generator containing only "conventional" physics, and the study of signatures of this anomaly in the observed final state, within the conventional event environment. Event selection might be studied by generating a number of simulated events that approaches the number of real–data events. However, this is not possible both because of the prohibitive CPU time needed and the lack of justification based upon the physics content of common event generators.

Event-by-event physics in STAR can be thought of in part as an offline trigger to select events based upon the details of the (primarily low p_T) particle distributions. Because of the need to study each event in detail, this is expected to be the most demanding task computationally for STAR analysis.

3.5 High p_T

Roughly 50% of the transverse energy of a central Au-Au collision at RHIC is in the form of jets or mini-jets. One simply cannot understand a heavy ion collision without understanding jet and minijet production. Additionally, hard probes such as jets and direct photons probe the collision at very small distances (and therefore very early times). To study how the nuclear matter evolves during the collision, it is necessary to understand the early stages of the collision.

Many of the signatures for a quark-gluon plasma rely on the enhancement in effective gluon flux in such a collision. To separate phase transitions from changes in structure functions in cold nuclei (which go by such names as "shadowing" and "the EMC effect"), it is critical that we know the structure functions and parton distributions in nuclei. This is anticipated as being done by looking at p-Au collisions in addition to Au-Au.

We are looking for several signatures in this program:

- Direct photons: produced by gluon Compton scattering, they are sensitive to the gluon densities. Because they are colorless, they do not interact strongly with nuclear matter.

- Jets: produced in g-g and q-g scattering, they are also sensitive to gluon densities. Because they are produced from colored partons, they do interact with nuclear matter, and the nature of this interaction can be studied by comparing photons and jets.

- High p_T electrons: produced by W's and Z's (only in asymmetric p-Au collisions) and in decays of heavy flavored quarks. The relationship between quark momenta and electron momenta in heavy flavor decays is known, so studying this subset of jets gives us different systematics and different quark-gluon ratios than generic jets. The ability to select heavy flavor may make extracting the signal in AuAu collisions easier than the generic jet signal.

We expect that analysis of high p_T phenomena will require moderate levels of computing.

3.6 Leptons and D–mesons

D mesons: Until now the question of open charm production in nucleus-nucleus collision remains unanswered, whereas sufficient data in nucleon-nucleon and nucleon-nucleus interactions are available, primarily from experiments at Fermilab. If there is an enhancement of *total* charm production in nuclear collisions, which reflects the presence of a production mechanism more effective than initially anticipated, the suppression factor for J/ψ formation must be even larger than presently estimated to explain the NA50 results. The obvious strategy for the detection of open charm is the study of the decays $D^+/D^- \rightarrow K\pi\pi$ and $D^0/\bar{D^0} \rightarrow K\pi$ with branching ratios of 9.1% and 3.8%, respectively. Although the STAR experiment was not designed to study D-meson production, it offers two unique features which may allow such a measurement: large acceptance and particle identification capability for the decay products. Previous studies have shown that the current Silicon Vertex Tracker (SVT) setup with 3 layers of silicon drift detectors is not sufficient to clearly resolve such small secondary vertex impact parameters. A micro-vertex detector upgrade may be necessary to obtain a clean signal. The reconstruction of D mesons is highly CPU intensive and we expect that it will be a major consumer of computing resources.

Lepton pairs: There is general agreement that information about the hot, early stages of the collision, and thus about the initial conditions of a quark-gluon plasma, can only be seen by employing probes that do not interact strongly and thus are not influenced by final state interactions. Such probes are real and virtual photons, the latter being observed as lepton pairs (muons or electron-positron pairs).

- **ϕ mesons:** As early as 1985 it was suggested by A. Shor to use the ϕ meson as an indicator of the quark-gluon plasma. In such a plasma one would expect a marked enhancement of $s\bar{s}$ pairs, resulting in an increase in the production of ϕ mesons. Due to the small ϕN cross section, these could escape the reaction volume without rescattering, bearing information about the initial conditions of the reaction. In addition to the question of the pure production cross sections there is an increasing interest in the study of the width and position of the ϕ peak. It is sensitive to medium–induced changes of the hadronic mass spectrum, especially to the possible drop of vector meson mass as a precursor to chiral symmetry restoration.

In the absence of high baryon densities (contrast RHIC and the SPS), modifications of the peak positions are predicted to be small except in the immediate vicinity of the phase transition, whereas the increase in the width of the ϕ due to collision broadening could be substantial.

- J/ψ **mesons:** The suppression of J/ψ production is one of the most promising signatures for a QGP created in nuclear collisions. Quarkonium ground states (J/ψ, Υ) are small and tightly bound resonances of heavy quarks. The J/ψ has a radius much smaller than the normal hadronic scale and its binding energy of 0.6 GeV is much larger than $\Lambda_{\mathrm{QCD}} \approx 0.2$ GeV, requiring hard gluons to resolve and dissociate. Confined matter for temperatures up to about 600 MeV does not contain sufficiently hard gluons to resolve the quark structure of the small J/ψ, whereas deconfined matter for T > 200 MeV easily can. Many theoretical studies therefore propose to use J/ψ suppression as a test for color deconfinement. Extrapolation from present data to RHIC energies is rather difficult since recent measurements suggest a substantial supression already in 160 GeV/n Pb-Pb collisions. By neglecting any supression mechanism, a rough estimate yields 0.03 J/ψ decays in the e^+e^- channel per event in central Au-Au collisions (assuming a 100% efficient trigger scheme).

We expect that dilepton analysis will not be as demanding as that for D-mesons, but still require a significant amount of computing resources.

3.7 Peripheral Collisions

The STAR peripheral collisions program will study two-photon, photon-Pomeron and double-Pomeron collisions, where the nuclei interact coherently via long(er) ranged forces. A major focus of the group is meson spectroscopy, studying scalar and tensor mesons produced in two-photon interactions which decay to a small number of final state particles. Rates are also large for photon-pomeron interactions, where part of the photon wave function is absorbed by the other nucleus, producing ρ, ϕ and J/ψ. These interactions test how the produced vector mesons interact with the other nucleus; by changing nuclear beams and hence the nuclear radius, different thickness targets can be probed. Double-Pomeron interactions are also expected.

These interactions are characterized by final states consisting of a few (typically two or four charged) particles with a well balanced perpendicular momentum, with nothing else visible in the detector.

As the program evolves and the STAR EM calorimeter is completed, peripheral collisions will begin a more general study of large impact parameter events. The focus of this will be studies of the spatial distribution of the nuclear structure functions. By selecting events with relatively large impact parameters, it will be possible to study the structure function near the nuclear surface; as the distance from the center of the nucleus increases, nucleon structure functions should evolve from showing a strong EMC effect toward the free nucleon structure function. We are currently studying measurements of charm production as a function of impact parameter.

Because the events are so small, we expect that the peripheral physics computing needs will be negligible compared to those for the nucleus-nucleus collision programs.

3.8 pp and Spin

It is well established experimentally that less than 30% of the proton spin is carried by the spin of the constituent quarks. It is therefore of great interest to measure the polarization of the gluon field, ΔG. STAR at polarized RHIC is uniquely positioned to make this measurement directly in a regime that is clearly perturbative. Processes of interest are dijet production and direct photon plus jet production. STAR will also measure the polarized u, d, $\bar{u}$, and $\bar{d}$ distributions through W and Z boson production.

The measurement of the parity violating asymmetry A_{PV} in high p_T inclusive jet production will be an interesting test of the standard model and a possible window to new physics. The third fundamental partonic structure function $h_1(x)$ may be extracted from the transverse spin asymmetry A_{TT} in high p_T dijet production or in Z boson production.

The polarization of hyperons produced in pp, pA, and AA collisions can be determined by their decays. The extra CPU time to compute these polarizations from reconstructed hyperons is negligible, and it is assumed that this will be done as a standard part of the hyperon analyses.

Unpolarized pp and pA measurements at STAR will form an essential set of reference data to calibrate what is seen in AA collisions. As an example, the quenching of jets may change the measured yield of jets or single particles as a function of p_T depending upon the medium through which the high p_T parton travels prior to hadronization. This can be observed only by a differential comparison of p_T spectra from AA and simpler colliding systems. The effect of shadowing on the structure functions, which will also contribute to the change, can be determined through the comparison of pp and pA data.

4 General Comments on Offline Analysis

4.1 Offline Analysis Chain

We do not make any assumptions here about the implementation of an Object Oriented Data Base Management (OODBM) scheme for the offline analysis. There are a number of open questions within the collaboration on the use of OODBM (this is not a statement about the desirability of such an implementation), so for clarity at present we define the various data analysis stages without reference to it.

The offline analysis chain consists of the following components:

DST Production All raw data on tape are passed first through a Data Summary Tape (DST) production process, which is the same for all data of a given physics class (e.g. Au-Au collisions). The DST contains physics-related quantities such as track 3-momenta and PID quantities and candidate secondary vertices. No rejection of tracks or events occurs at this step[1]. This step requires a large volume of input data and is very expensive in CPU time. The baseline assumption is that there will be slightly more than one DST production pass for each raw event (including calibration, code development and debugging, etc.), though this remains an important point for discussion. The DST production is carried out using the CRS (Central Reconstruction Server) at the RCF.

miniDSTs The DSTs may also contain objects that are not essential for further physics analysis, such as correlation matrices for track and vertex fits. In applications for which these objects are not essential, it may be desirable to compress the DSTs to "miniDSTs" (mDSTs) containing only essential physics quantities, again without rejection of tracks or events. See further discussion in section 6.4.

Data Mining Data mining is the first step that is specific to given physics analysis project(s). The DST or mDST is read and event selection is made, with the output written to a project-specific μDST. Typically this is a filter making decisions based upon calculations performed during DST production, and therefore requires low cpu per event and high input bandwidth. Both reduction in number of events and reduction in the data volume per event may occur, so that the output μDST data volume can be considerably less than the input DST volume. In an OODBM framework, the output μDST may in fact consist simply of set of pointers to data objects, which are not duplicated from the DST. Though carried out more frequently than DST production, it is not expected that data mining will occur on a daily basis. It seems likely to us that different projects within the same general class of physics analysis will find it desirable to share data mining processes, thereby reducing significantly the number of times the raw data have to be read. Data mining at the RCF occurs in the CAS (Central Analysis Server).

Analysis Analysis specific to a given project is carried out on the μDSTs. This consists both of calculations based upon the information in the μDST and histogramming. It is expected that this step will be performed very often. Though each analysis project requires a modest volume of data, the multiplicity of projects and the frequency of data access can generate a significant part of the computing requirements.

The general pattern of development at all steps of the data analysis chain is one of pilot studies followed by production runs. This will be true throughout

[1] A minor "conceptual" exception to this rule is the formation of secondary vertex candidates using very loose cuts, which nevertheless implies the rejection of tracks from consideration as daughters of vertices already at this level. See section 7.2.

the lifetime of STAR. Some QA tasks will have to proceed with substantial data volumes in pilot studies in order to achieve sufficient statistical accuracy.

4.2 Latency

For the purpose of data quality monitoring, we require about 1% of the raw data to be reconstructed within 12 hours of its recording. Unlike current collider experiments such as CDF, where all data undergo reconstruction within a few days after they are recorded, reconstruction of the full dataset at STAR is expected to occur over a much longer time period. The running period at RHIC is about 4000 hours per calendar year, or about half the year. Because of the very cpu-intensive nature of event reconstruction, this process will be spread out through the full year, thereby significantly reducing the computing needed relative to the CDF model, but at the expense of greatly increased latency.

4.3 Upstream Data Access

This refers to access to event information from previous steps in the analysis chain, μDST$\rightarrow$DST$\rightarrow$raw data, best performed within an OODBM framework.

A general need is for a broad QA program on a very limited subset of events. From NA49 experience this is typically 1 event (space-point clusters in the TPC) or 1000 events (tracks), very rarely $\approx$10,000 events (spectrum backgrounds) for some sort of calibration task, in total probably not more than 1% of total volume.

A second major lookback activity comes with Event–by–Event analysis. The scope of this will depend on what we find. Given an anomalous event class, what does it take to understand the anomaly? From the point of view of skepticism some fraction of anomalous events will need to be examined, beginning with the raw data. This is most probably a very small fraction, say a few thousand selected events spread over several event classes. More importantly, for selected event samples it may be desirable to redo decaying particle searches or PID, or improved algorithms may be developed in response to better understanding of some anomalies as artifacts, leading to better quality cuts.

The need for upstream data access dictates the use of an OODBM framework, but we cannot at this point make any reasonable estimate of the magnitude of the computing task due to this activity, other than to say that it will occur only on a small fraction of the events.

4.4 Calibration

Calibrations include the determination of alignment, gain, and (for tracking devices) spatial distortion correction factors. Experience from TPC experiments such as NA49 shows that determining calibrations for tracking devices in a high energy heavy ion environment, while a subtle and difficult problem which may take considerable time to understand fully, is not computationally intensive. Typically a small amount of data, both from collisons and from other sources such as cosmic rays and laser tracks, is studied continually over a long period.

The EMC can be calibrated using a number of different processes in pp, pA, light AA and peripheral Au-Au collisions. These include the decays $\pi^0 \rightarrow \gamma\gamma, \eta^0 \rightarrow \gamma\gamma, J/\psi \rightarrow e^+e^-$, and $Z^0 \rightarrow e^+e^-$. We calculate the number of events needed to calibrate the EMC as follows. At installation, modules will have been calibrated to $\approx 2\%$. We limit our discussion to π^0s with $p_T > 3$ GeV. These are quite pure, even in central Au-Au collisions, and we get of order 1000 per day (if only central triggers are taken). The energy resolution on each pion is about 10%. So for 2% overall calibration, we need 25 pions per bin. The detector has 4800 towers. Since the physics is ϕ–symmetric, we can normalize all towers at the same η but different ϕ to have the same average photon energy in order to achieve relative calibration. That leaves us with 40 η rings to calibrate. We need 40 x 25 = 1000 π^0s to calibrate each ring to 2%. Since we want both photons to be in the same η ring, this costs us a factor of 2-3 in acceptance, so it would take us 2-3 days to do this sort of calibration using central events. However, because about half the towers contain some energy for a central Au-Au collision, it is much more desirable to perform this calibration on more peripheral events, taken using (e.g.) 10% of the bandwidth. The calibration to 2% is then achievable after 1 month of running, and to 1% after six months. Since the π^0s are triggered, the data volume will be identical to the above estimate for central collisions, i.e. a few thousand central event equivalents. The dataset needed to calibrate the EMC in situ is data that STAR plans to collect anyway, to study AA collisions over a wide range of impact parameters. We conclude that calibration of the EMC requires negligible computation compared to other data analysis.

5 Data Sets

The following sections define the data sets needed for analysis of typical projects within each of the STAR physics categories for central AuAu at $\sqrt{s} = 200$ GeV. These data sets are the input to the data mining process, and are summarized in Tables 2 and 3.

5.1 Soft Hadronic Probes

We estimate the number of events needed for a variety of single particle inclusive spectra for central Au-Au collisions for common primary charged hadrons: π^+, π^-, K^+, K^-, p, $\bar{p}$. We assume the availability of STAR-TOF information from a patch containing 12 trays, or 10% of the full outer area of the TPC. The estimate of numbers of events will therefore include a factor of 10 for the geometrical acceptance of this TOF patch.

dN/dy for specific particle species: We divide the rapidity region $|y| < 1.8$ into 18 bins ($\delta y = 0.2$), and we require maximum statistical error in any of these bins is 1%. Comparing the predictions from the Hijing (v. 1.3.1) and Venus (v. 4.12) models for central Au+Au collisions, the smallest multiplicity of an identified particle in bins of this size is predicted by Hijing for anti-protons at y=0. Taking this value, 2 antiprotons/bin/central Hijing event, implies that 5000 events per central event sample are necessary. Including the factor 10 for TOF acceptance, we estimate that dN/dy analysis requires $5 \cdot 10^4$ events.

dN/d$\mathbf{p}_T$ and dN/$\mathbf{m}_T$d$\mathbf{m}_T$: We require these spectra to span two decades with 3% statistical error in the least populous bin that has width x MeV/c, where x = 0.015*(p_T of bin), set by the momentum resolution. For central Au+Au, both Hijing and Venus predict these bins to occur at 1.5:2.0:2.5 GeV/c for π:K:p at mid-rapidity. It should be noted that the need for TOF information is particularly obvious for this observable - these p_T values are well above those for which STAR can expect reasonable PID from dE/dx alone. We find the following multiplicities per event:

20 MeV bin at p_T=1.5 GeV/c:	N($\pi^{\pm}$) = 1/event (Hijing or Venus)
30 MeV bin at p_T=2.0 GeV/c:	N($K^{\pm}$) = 0.1/event (Hijing)
	N($K^{\pm}$) = 0.2/event (Venus)
40 MeV bin at p_T=2.5 GeV/c:	N($p, \bar{p}$) = 0.1/event (Hijing or Venus)

For perfect PID and tracking, this implies a minimum ≈11000 events. Including the TOF-patch geometrical acceptance of 10%, we estimate that analysis of spectra in (y,p_T) bins require $1 \cdot 10^5$ events.

Scaling to other systems: The above estimate was made for central Au-Au collisions. Study of hadronic spectra are of great interest for all possible variations of experimental configurations (including the addition of new detectors, mentioned abover), trigger conditions (variations in total multiplicity, E_T, etc.), event-by-event selections, variation in beam energy and colliding systems (lighter ions, pA, pp). While the sample size for each of these configurations is modest, the total may add up to a considerable size. We do not attempt to estimate the number of different configurations that will be measured by STAR each year, but make the following observation about data volume: for inclusive spectra, the data volume is driven by the number of tracks in a sparse bin. While the total multiplicity may change enormously with the above variations, the shape of the p_T spectra will change much less and the total required *track statistics*, as distinct from required *number of events*, will be roughly constant. Thus the required *data volume* will be roughly equivalent to that estimated for central Au-Au, independent of the above variations in configurations. We expect that single particle spectra analysis computing requirements will scale linearly with the data volume, and therefore will be the same for any system as that deduced for central Au-Au collisions.

5.2 Hyperons

We discuss both hyperons and the K^0_s meson, since they are reconstructed using the same "V0" technique, differentiating in addition between $|s| = 1$ (Λ, $\bar{\Lambda}$) and the rarer $|s| = 2, 3$ (Ξ, $\bar{\Xi}$, Ω, $\bar{\Omega}$). Table 4 shows the expected yield per event of various "V0" type strange particles[2], including the effect of cuts that reduce the signal by a factor 10 in order to achieve signal to background of at least 2 to 1.

Notice that $|s| = 2, 3$ particles and anti-particles are produced in about the same abundance, and the acceptance for particle and anti-particle is the same. The number of Λs found at mid-rapidity depends especially strongly upon the model. Hijing predicts about 30% more Λs at mid-rapidity than Fritiof.

We estimate the total number of events required for various measurements, using experience from NA49 as a general guide,

- 10^3 reconstructed particles will measure yields
- 10^4 reconstructed particles will measure 1-d distributions
- 10^5 reconstructed particles will measure double-differential c.s.

Table 5 gives estimates of the required number of events for various K^0_s and hyperon measurements, based upon NA49 experience and the rates from Table 4 for the TPC+SVT case (where there are reasonable estimates for the reconstruction efficiency). For the TPC alone, the rates may be down by a factor 3-5 for $K^0_s$and Λ, probably 10 for Ξ (work in progress).

5.3 Correlations

We consider six data sets for correlation studies: charged pions (both signs), charged kaons, K^0_s, multi-particle correlations, protons, and non-identical particles. The following estimates, appropriate for central Au-Au collisions, are based upon a pion rapidity density dN/dy=300 per charge sign at mid–rapidity with a p_T distribution similar to that observed by NA49 at the SPS at $\sqrt{s} \approx 20$ GeV. Of course, analysis for impact parameters other than central will also be carried out; however, the resource requirements for these analyses will be accounted for in the scaling from central Au-Au analysis to overall STAR needs in section 9.

Identical charged pion pairs: We consider the phase space $-1.5 \leq y \leq 1.5$ and $0 \leq p_T \leq 1.5$ GeV/c, resulting in six rapidity bins with Δy=0.5 and five p_T bins: 0-200, 200-400, 400-700, 700-1200, >1200 MeV/c, giving 30 bins for each charge sign. In order to have 50K useful pairs (defined as Q_{inv}<100 MeV/c) in the least populous bin (p_T>1200 MeV/c), we require $2 \cdot 10^6$ events.

Identical charged Kaon pairs: We divide phase space into $0 \leq |y| \leq 0.5, 0.5 \leq |y| \leq 1.0, 1.0 \leq |y| \leq 1.5$ and p_T=0-400, 400-1500MeV/c, giving 6 bins per charge sign. We assume zero net baryon number and therefore the

[2] Result of Hijing simulations by Ken Wilson, Spiros Margetis *et al.*

same population of K^+ and K^-. We estimate that the number of events required for sufficient statistics in the low p_T bin is comparable to that of high p_T pion data set, based on the ratio of K^+/π ~15% usually seen in heavy ion collisions. We therefore require $2 \cdot 10^6$ events.

Proton pairs: Per event statistics and therefore binning is the same as kaons.

Non-identical particles: We consider π^+-π^-, π-K, π-p, and K-p correlations in three rapidity bins from $0 \leq |y| \leq 1.5$ and two p_T bins with $p_T \leq 400$ and $p_T > 400$ MeV/c (boundary determined by PID capability) for a total of 24 bins. We can extract good statistics from the 2-pion data set.

Multiparticle correlations: We consider only positives. We can extract good statistics from the two–pion data sets, allowing us to bin as $0 < |y| < 0.5, 0.5 < |y| < 1.0, 1.0 < |y| < 1.5$ and p_T=0-200, 200-400MeV/c, and 400-700MeV/c, for a total of 9 bins.

K_s^0 pairs: The yield is reduced relative to charged kaon pairs by two factors: the branching ratio and reconstruction efficiency. The branching ratio is about 65%. We assume that the combination of acceptance and efficiency cuts the yield to about 0.2%. The required numbers of events is in the range of 4M for 80K useful pairs.

Summary: For the two particle correlation functions studies, the statistics are driven by the K_s^0 data set:

- 500k events/per setting is required for most 3-D pion analyses;
- 2M events/per setting is required for most 3-D pion and kaon analyses and possibly 1-D $C_2(Q_{inv})$ for K_s^0;
- 4M events/per setting is required for 2-D K_s^0 analysis.

5.4 Event–by–Event

The event–by–event program in STAR will look at every full-energy Au-Au central event. Events from p-A, lighter A-A and Au-Au at lower energies will most likely be used for calibration purposes. While unusual dynamical fluctuations may occur in any colliding system, those with lighter projectiles or lower bombarding energies have poorer statistical power due to lower multiplicities, making the detection of possibly subtle event-by-event signatures more difficult.

5.5 High p_T

Online triggering on high p_T π^0s and γs in central Au-Au collisions will be highly efficient and pure at p_Ts where there is still reasonable rate. Jet triggering for these collisions will not be as clean, and we assume that every event that STAR writes will need to be studied for high p_T signals. Because statistics on the highest p_T phenomena are always limited by cross-section, this statement will be true for the duration of STAR.

We anticipate every event being examined for a direct photon candidate and a jet candidate (possibly by more than one algorithm). We expect that fewer than 1% of the events will be deemed interesting enough for further analysis. The fraction 1% is somewhat arbitrary, but that is how high p_T is defined: "higher than some fraction of the events". Initially that fraction is 1%. Over time, that 1% will fall to .1%.

A year's data set of $1.7 \cdot 10^7$ central Au-Au collisions will yield 5000 γs with $p_T > 10$ GeV and 5000 jets with $p_T > 20$ GeV.

For lighter systems, the purity and efficiency of the triggers will improve quickly as multiplicity decreases, so for lighter systems the data volume will be much smaller than that calculated simply by scaling from Au-Au with multiplicity.

5.6 Leptons and D–mesons

The following estimates for dileptons are based upon preliminary investigations for STAR (not using the full analysis chain) and experience with the NA45 experiment. The D meson estimates are extrapolated to STAR multiplicities from a more extensive MC study for NA49. All numbers are given for central Au-Au collisions at $\sqrt{s} = 200$ GeV. Note that not all requirements scale linearly with multiplicity.

Studies of J/ψ and D mesons cannot be performed by the baseline STAR detector but require some components of the proposed STAR upgrades. An efficient trigger scheme based on a fully equipped EMC is required to extract a statistically significant J/ψ signal. For the D meson, most likely an upgrade of the current SVT (additional layers) is necessary to enable reconstruction of secondary vertices at the required precision.

$\phi \to e^+e^-$: We aim for an effective signal (i.e. background–free equivalent) of about 10^4 ϕs distributed in several bins of (e.g.) E_T, in order to study the p_T of the ϕ. Considering central collisions only, this sample requires 10^7 events. For other E_T bins more events are required due to lower multiplicity, but the computing requirements per event are correspondingly less. We ignore this variation and conclude that the equivalent of about 10^7 central Au-Au events are needed. Scaling from Au-Au to p-Au is approximately linear.

$J/\psi \to e^+e^-$: In order to obtain an effective signal on the order of 10^3, again roughly 10^7 events must be scanned. Note that although the cross section for the J/ψ is much smaller than for the ϕ, the signal/background is much more favourable and hence the effective signal is reasonable. The data volume scales approximately linearly with multiplicity from Au-Au to p-Au. The CPU time scales more strongly than linear but not quadratically; we propose to use a linear scaling also for CPU.

D mesons: Most numbers are taken from a NA49 D–meson study scaled to STAR multiplicities. We conservatively assume 2 D mesons/event. Taking the branching ratios ($D^0 \rightarrow K\pi$, $D^\pm \rightarrow K\pi\pi$) into account, assuming an overall efficiency of 5% (optimistic) and assuming an S/B ratio of 1:10, we need $3.5 \cdot 10^6$ events for an effective signal of 1000 Ds (D^0, $\bar{D}^0$, $D^\pm$). Both the data volume and the CPU requirements do not scale linearly with multiplicity. A good approximation to scale from Au-Au to p-A is to scale with multiplicity to the power 1.6.

5.7 Peripheral Collisions

We estimate the computing requirements for the peripheral collisions program, for a 10^7 second/year run, at design luminosity. We will assume that the peripheral collisions trigger writes to tape at 2 Hz.

Because peripheral collisions events are small (2 or 4 tracks), this uses considerably less than 1% of the data volume. Even if there are 20 background tracks in the TPC, this is only about 1% of the data bandwidth. In any case, it is hoped that, as we gain experience with the trigger, most of the background tracks can be eliminated at level 3, and only clusters from real tracks recorded on tape.

This leaves us with $2 \cdot 10^7$ events per year. In more useful terms, this is $6 \cdot 10^7$ tracks per year, or the equivalent of $1.5 \cdot 10^4$ central events per year. Before and during DST production, background tracks (from other beam crossings) in the TPC will increase the load by a factor of 10, to $1.5 \cdot 10^5$ equivalent events. During DST production, these tracks and clusters will be removed, reducing the sample to $1.5 \cdot 10^4$ equivalent events.

On a longer time scale, after the barrel calorimeter is complete, peripheral collisions will explore some topics in photon-gluon and gluon-gluon fusion in large impact parameter collisions. This will probably involve events with somewhat higher multiplicities (6≤M≤12?), so may double or triple our data and consequent requirements, even if the same 2 Hz trigger rate is retained (we would give up some channels like low invariant mass 2-track events to maintain the trigger rate).

Summary: Computing Requirements for Peripheral Collisions are minimal, because the events are so small.

5.8 pp and Spin

In this section we include a discussion of *pp* running for Au-Au reference as well as spin physics. The current run plan calls for 10 weeks of polarized *pp* running annually, beginning in year 2. At 66% up time, this is $4 \cdot 10^6$ seconds.

The spin physics program needs many more events than, for example, cross section measurements, since effects are small and difficult to detect. A typical trigger for interesting processes consists of a threshold in energy and cuts on topological features.

For example, consider one physics process that will be a large component of the spin physics data sample: dijet production, used to extract the gluon polarization as a function of the parton momentum fraction, x. The cross section for two back-to-back cone R=0.7 jets in $|\eta| < 0.7$ with $E_T > 15$ GeV is 200 nb. At $2 \cdot 10^{31}/cm^2/s$ (design luminosity), this is an event rate of 4 Hz. If most of the extraneous TPC hits from pileup can be dropped by Level 3 track finding, the event size can be held to 0.5 MB, including the readout of a fully instrumented EMC.

The other physics processes of interest in the spin program (except possibly the inclusive photon sample) are rarer than dijet events. We therefore tentatively set a DAQ budget of 15 Hz for 0.5 MB events or $6 \cdot 10^7$ events/year, where a year is $4 \cdot 10^6$ seconds, yielding a raw data volume of 30 TB/year. This is approximately 15% of the raw data volume from a year of Au-Au running.

In the years before the EMC is fully implemented, the event rate will still be high since triggering will be much less efficient. A DAQ output of 15 Hz may still produce too many events. If Level 3 does not have sufficient time or resources to perform the desired triggering, then one could consider a Level 4 trigger (as envisioned by some LHC experiments) that would simply drop events at the event reconstruction level.

During the initial year of design luminosity $(2 \cdot 10^{31})/cm^2/s$, it may be desirable to write out the full TPC including the pileup (i.e. without Level 3 tracking and hit rejection) so that a large reproducible data set will exist for testing Level 3. This may be required a second time when the luminosity increases substantially $(2 \cdot 10^{32})$. This would produce a data set of about 40% of the Au-Au data set in those years.

There is also the possibility that for some classes of events, the raw event size could be greatly reduced by, for example, writing out tracks instead of hits.

6 DST Production

6.1 DST Data Volume and Processing Time

Raw data volume for a central Au-Au collision is currently estimated to be 12 MB/event, based upon Hijing events run through GSTAR with all processes active. (The version of Hijing used to obtain this number generates rapidity density dN/dy(charged)=1200 at midrapidity.) This figure for the data volume has considerable uncertainty, of course. We hope that nature holds surprises for us, but even within the scope of current understanding this number is not well established, as is exemplified by the variation of its estimate among common models: simulations using Venus events as input give a data voume approximately 100% larger than those using Hijing. On the other hand, the TPC may be operated with a clock speed and drift velocity such that many fewer time buckets will contain data than was assumed to obtain the above figure. Bearing these points in mind, we estimate a raw data volume of 12 ± 4 MB/event for central Au-Au collisions.

The DST production is estimated to reduce the data volume by a factor of 10, consistent with existing TPC-based experiments. We are mindful of the fact that this factor is a function not only of the physics of the event but also machine backgrounds and spurious data such as electronic noise, but in the absence of any additional information, we must use the factor 10 and consider the backgrounds simply to add to the uncertainty of the raw event size.

All events written to tape by STAR will undergo event reconstruction. Both CPU time and data volume should scale linearly with event multiplicity, with the exception of secondary vertex finding (this is discussed separately in section 7.2). We assume that "on the fly" DST production will be performed for only a small fraction of the events during data taking, of order 1%.

For the reduction of raw data to DSTs, the original ROCOCO-2 report (February '96, table 20) reported 33 GFlop-sec/evt, whereas the update to the ROCOCO-2 report (July '97, table 13) reported 150 kSi92-sec/event. Using the rough conversion factors from SN0314 of Si95/MFlop≈13 and Si95/Si92≈40, we derive 2.5 kSi95-sec/event and 3.75 kSi95-sec/event for ROCOCO-2 and its update, respectively. Given the uncertainty in these conversions, and bearing in mind the recent progress in optimizing STAR TPC tracking, we choose to use the lower number of **2.5 kSi95-sec/event**, or about 600 seconds/event on an HP735/125. This will probably improve with time but is not out of line with what has been experienced by NA49, and we will continue to use this number for this report.

Consequently, a sample of 1M central Au-Au events, useful for the bulk of STAR's "moderate" analysis tasks, (see Table 2) requires a reconstruction time of $2.5 \cdot 10^6$ kSi95-sec. The total yearly sample of **17M events** requires $4.3 \cdot 10^7$ kSi95 seconds for one pass through the full raw data set.

We assume that each raw event in STAR will be processed 1.5 times, allowing for reprocessing of a fraction of the data. In addition, we assume a combined RCF/STAR computing duty factor of .75.

Event reconstruction times for p-p should roughly scale to Au-Au by the number of tracks or by event size, which is nearly the same scaling. Since optimized global track fitting and excellent calorimeter calibration are essential to good jet resolution, we anticipate several passes of the data to get the DST's correct. Hence we estimate that 30% of the Au-Au event reconstruction time will be sufficient for several passes of a data set 15% of the Au-Au data set in size. This should also cover the required time for calibration. For the years in which the full TPC is written out, the data set will approach 40% of the Au-Au data set and only one pass will be possible. For those years, event reconstruction must be treated as a trigger and be fully tested before use.

6.2 DST Content

We assume that the DST contains (at least) information about all global tracks. We can speculate whether some space point information will be saved and what the consequences of this might be later.

There is a question of how much processing to do at the DST production level, and how much to do at the μDST production level. Event–by–event analysis needs to look at *all* the events at the μDST level and can delay some calculation to that level, but analysis of rare processes might benefit from significant event selection based upon DST information only and would benefit from more CPU time and resources used earlier in the chain.

For the case of V0's, we should seriously think about limiting the analysis at the DST level so that we get through one pass of the data quicker. This would mean that there must be sufficient CPU capacity on the Central Analysis Server to complete the reconstruction. We estimate that from 10-30% of the analysis may be performed after reconstructing global tracks. The argument against this comes from the search for rare decays, such as Ωs. To avoid reading every DST event during the rare decay reconstruction, it makes sense to do some finding to flag 1-5% of events (say) which may have a candidate. This means running V0 finding and Ξ and Ω finding on all events. The V0 finder must be made as efficient as possible and it absolutely must make some cuts.

6.3 pp and Spin

In addition to the calibrations, the pp running may require more than one pass to produce DSTs because of the pileup effects in the TPC and perhaps SVT. The data analysis software will most likely be optimized for heavy ion interactions, with a unique interaction vertex, while problems with high luminosity running with pp give many separate events within a TPC or SVT readout time. It is our guess that the handling of the pileup may not be optimized on the first pass of the data, especially when high luminosity first occurs. Perhaps after this first high luminosity running this will be a nonissue.

6.4 miniDST

DST production is a data compression process in which some fraction of the tracking information is abandoned, and ideally none of the particle (kinematic) information is lost. Because of the uncertainty in this compression process and the great size of the computational investment in the compression step (implying the possibility of only one pass of DST production), there is inevitably a substantial amount of redundant tracking information retained. This implies that a further compression step without event or track selection may be appropriate prior to the data mining process.

One further compression step results in what can be termed a miniDST. This entity contains, for all events and all particles, the minimum data volume required to represent all particle kinematic information and negligible tracking information. The specific data volume for a miniDST is typically 100-500 bytes per particle. This implies a yearly miniDST volume of about 8 Tbyte. The DST volume is already sufficiently reduced from the raw data (200 TB$\rightarrow$20 TB) that miniDST volumes may be produced several times in the history of STAR if refinements to this compression step are developed.

The processing time for miniDST production should be small. There is only minimal computation, mainly extracting certain data structures from the general DST format. It is possible that one or a few miniDST processors could keep up with the DST production rate.

7 Data mining and analysis

In this section we estimate the input and output data volumes for the data mining and subsequent analysis steps. The data sets that are input to the data mining process are defined in the leftmost column in Table 2.

In the following, estimates are made only for one experimental configuration (beam species, beam energy, magnetic field, ...). We will attempt to account for multiple configurations when scaling to the full STAR requirements. The results are summarized in Tables 7 and 8.

7.1 Soft Hadronic Probes

Data Mining From Table 2, input data set is $2 \cdot 10^5$ events. Input volume is therefore 240 GB, output volume is 24 GB. This process is a simple filter and data compression and should require negligible cpu time. We expect this process to occur a few times per DST event.

Analysis This step involves particle ID determination, possibly through fitting and deconvolution of PID spectra (e.g. dE/dx from TPC), as well as other histogramming and fitting of spectra. Input volume is 24 GB. Output volume and cpu time are negligible. This process will be perfomed very frequently.

7.2 Hyperons

In NA49 a V0–specific μDST is created, with about 40 words of information for each candidate. More stringent cuts are made during the μDST production process. If we assume that in STAR we filter at most 100 times more candidates than signal, then to obtain 100K particles signal, we would have a 1.6 GB μDST.

For the rare particle searches, it is essential to have a way of selecting events on the DSTs. That is, some small subset of information from the DST should be saved either in a separate file, or in a more sophisticated database. Events would be tagged if they have a likely Ξ or Ω candidate. For these rare particle searches it is likely that we will need to sample about 10% of the DST data for Ξs and about 1% for Ωs. This assumes that V0 and Ξ finding have been run during DST processing.

As a rule of thumb, the V0 finding and fitting should take 10-20% of the CPU time needed to reconstruct the event. We have measured the V0 finding for STAR to take about 40 Si95-sec/event. We note that the event reconstruction time given in section 6.1 is 2.5 kSi95-sec/event, so that these measurements currently correspond to only 3% of the total. However, recent studies have shown that the event reconstruction can benefit from significant improvent in efficiency.

In NA49, the V0 fitting takes the same amount of time as the V0 finding, so we double the total figure for STAR to estimate 80 Si95-sec/event for V0–finding and fitting. However, the V0 candidate finding can be performed during the DST production, whereas the V0 fitting is carried out on the tracking information using the error matrices. We therefore assign only the V0 fitting time, 40 Si95-sec/event, to the data mining step.

For a "moderate" analysis to measure dN/dydp_T of Λ (see Table 2), about 1M events are needed, corresponding to $4 \cdot 10^4$ kSi95-sec. A hard analysis, requiring an event sample of 20M events, will require $8 \cdot 10^5$ kSi95-sec.

Data Mining For the data mining step we assume that some V0 finding using loose cuts has been performed at the DST production step. The data mining then consists of selecting those events with V0 candidates. However, for central Au-Au collisions, every event will have many such candidates, so that the effect of data mining is only to reduce the data volume per event, not the number of events. Table 4 shows that there will be of order 4 real V0's per event. Using the above figure of 100 more candidates than signal and 40 words (160 B) per candidate, this yields an output from data mining of about 60 kB/event.

- For a "moderate" analysis (dN/dydp_T of Λs), 1M input events are needed (1.2 TB DST data), resulting in an output μDST of 60 GB. CPU time required is $4 \cdot 10^4$ kSi95-sec.
- For a "hard" analysis (yield of Ω), 20M input events are needed (24 TB DST, resulting in an output μDST of 1.2 TB. CPU time required is $8 \cdot 10^5$ kSi95-sec.

Analysis This step consists of histogramming, peak fitting, etc., with input data volumes of 60 GB for a moderate analysis and 1.2 TB for a hard analysis. CPU time per event will be negligible. We expect this step to be performed very frequently.

The estimate of CPU usage would change substantially if, for instance, points in the vicinity of V0 candidate tracks are saved in order to perform refitting at this step.

7.3 Correlations

Given the data sets from section 5.3, we have the following correlation functions calculated for e.g. central Au-Au collisions:

$\pi^{\pm}$	30 bins $\cdot$ 2 = 60
$K^{\pm}$	6 bins $\cdot$ 2 = 12
p	6 bins $\cdot$ 1 = 6
multiparticle	9 bins $\cdot$ 1 = 9
K_s^0	6 bins $\cdot$ 1 = 6
non-identical	6 bins $\cdot$ 4 = 24

for a total of 117 bins (correlation functions) for one experimental configuration.

Data Mining Most of the bins can be populated from the tracks of 1M events. Assuming 40 bytes/track and a simple filter at low CPU cost on good track candidates, we have a μDST volume of
1M events $\cdot$ 2000 good tracks/event $\cdot$ 40 bytes = 80 GB

The rest of the bins can be filled by selecting tracks from 3M more events via p_T and PID selection cuts, also at low CPU cost. We estimate an additional 20 GB of additional μDST volume for this, making a total of 100 GB. Input is 4M events (section 5.3), corresponding to 5 TB of DST data. Output is 100 GB. This selection will be performed a few times per DST production cycle.

Analysis This step involves PID, p_T, event multiplicity and rapidity cuts, all made at low CPU cost, followed by the creation of the real and mixed-pair background distributions and iteration of corrections, at significant CPU cost. We estimate this cost to be about 1 day on an HP735/125 (4 Si95), or about 350 kSi95-sec. For all data sets, this corresponds to a total of $4 \cdot 10^4$ kSi95-sec, with input volume of 100 GB. This process will occur a few times per DST production cycle. Further analysis, primarily fitting of correlation functions, will occur very often but with negligible input data volume (correlation histograms) and required cpu time.

7.4 Event–by–Event

Beyond the miniDST, which may serve as a common data base for many different physics analysis programs, a further compression step is desirable for EbyE physics. This is the formation of the event spectrum. By various correlation analysis procedures, the multiparticle distribution in an event is represented by a minimal set of parameters or a vector. The specific data volume for an event spectrum is typically 100-500 bytes per event, very similar to the volume for a particle in a particle spectrum. It is the event spectrum that serves as the basis for event selection. For Scaled Correlation Analysis (SCA), event spectrum production from the miniDSTs reduces 8 Tb to 8 Gb. The computing cost for this process is significant, corresponding to up to $7 \cdot 10^6$ kSi95-sec for 2-D SCA of momentum spectrum (which should dominate analysis time). The resulting data volume is negligible. Analysis of the resulting event spectra can be carried out at home institutions. Data volume and CPU requirements are trivially small.

The analysis task required to form the event spectrum may be eased (with regard to I/O demands) by prior formation of a 'horizontal' μDST, defined as containing *partial* kinematic information for *every* event. The information to be included would have sufficient statistical power to provide a basis for nontrivial event selection. This means that some particle classes (e.g., hyperons) and some kinematic information would be omitted. One might choose to retain only momentum information for pions and kaons as an example. This may reduce the 5 Tb mDST volume to a horizontal μDST volume less than 1 Tb. In the context of an OODBMS this μDST (and others?) may be irrelevant, but the details of what a realizable OODBMS implies in the context of μDST formation and usage must be better defined.

Given that several event classes have been selected based upon differential density distributions in real data and reference event spectra, there is need for a "lookback" process. One wants to collect the full dynamical information for selected events and particles in order to form inclusive spectra for further physics analysis. This data volume may be called a 'vertical' μDST, defined as containing *full* kinematic (and some tracking?) information for *only some* events. The vertical μDST data volume in terms of events would be no more than 10% of the total data, with this estimate driven not only by physics estimates but by a lack of justification for further statistical power in an inclusive analysis. Thus, several event classes of order 100k events each should be anticipated, or 1M events total. This would correspond to about 3 Tb of vertical μDST volume if one wants this degree of information (including some tracking information). There is little justification at this time to want to access 10% of the raw data. Raw data access should be well under 1% (i.e., less than 1 Tb), motivated mainly by studies of possible systematic errors or suspected spurious tracking effects.

The selected events will be distributed randomly throughout the data population. To retrieve 10% of the DST volume would require 4 days at 10 Mb/s, assuming immediate availability from a tape robot. With multiple tape drives there should be little conflict among these numbers. The need to retrieve such an event population in a shorter time period is not apparent. Similar numbers would apply to retrieval of 1% of the raw data complement. Since this would be mainly on shelves there might be a greater time required for manual intervention. However, the real bandwidth bottleneck at this stage of analysis will be human ability to understand the event population at the event spectrum and inclusive analysis level.

To estimate CPU requirements, we refer to analysis currently carried out on HP C180 processors (12 Si95). A 1-D SCA analysis (e.g. m_T, y) takes 1 second for each event or 2 seconds for each real-data/reference pair, or 24 Si95–sec. For a 2-D analysis (e.g. m_T,y) the cost would be 10-30 times this or about 0.5 kSi95–sec. For $1.7 \cdot 10^7$ events this is $8.5 \cdot 10^6$ kSi95–sec. This task may need to be carried out several times as refinements are made. We estimate a total requirement of $4 \cdot 10^7$ kSi95–sec for analysis.

The resulting event spectrum data volume, at 150 bytes/event, is 2-4 Gbyte. This volume is trivially transportable to home institutions. This serves as a major basis for event selection.

Data Mining All events are used as input. As discussed above, initial reduction of the DST to a mDST results in a data volume of 8 TB for 17M events. 2-D SCA analysis to form an event spectrum requires 0.5 kSi95-sec/event, for a total of $8.5 \cdot 10^6$ kSi95-sec for 17M events. Output data volume of the event spectrum is about 8 GB. It is expected that this will be performed about 5 times for each DST production. Analysis of the event spectrum for anomalous events will result in selection of several event classes for subsequent inclusive analyses, which we estimate will each correspond to between 1% and 10% of the total set of events. Thus, the output of data mining and event spectrum analysis is a set of several μDSTs consisting of selected events, each of which may comprise 1% and 10% of the total data set, and whose size is on average about 1 TB. We include the event spectrum analysis in Data Mining, both because it is a true data mining process and because its output is a set of μDSTs, similar to the Data Mining Process in all other sections.

Analysis Further inclusive analysis on the selected events will procede as in other sections (hadronic spectra, hyperon production, HBT, etc.) We do not attempt to estimate the CPU times required for these processes, but they should be of the order or less than those estimated in the other sections (due to smaller data sets), and certainly smaller than the CPU times required for the event-by-event data mining.

7.5 High p_T

We anticipate that every event will be examined for direct photon, jet and electron candidates, and that these will be available as "objects" for further analysis. The jet algorithms are not yet specified, and may be subject to change as soon as real data is available. Experience at CDF suggests that the CPU necessary to do this is roughly 2% of that needed to reconstruct and track the event, or 50 Si95-sec/event for STAR.

The output objects from the jet finding are typically a few percent of the total event size. We anticipate most analysis being done on the "Jet", "Photon" and "Electron" datasets, with return to lower levels of data abstractions (e.g. tracks and hits) only for a handful of extremely unusual events.

Data Mining Processing time is 50 Si95-sec/event. Input data is complete annual event sample of 17M events. Total CPU is therefore $8 \cdot 10^5$ kSi95-sec. Input data volume is 20 TB, output data volume per analysis project is about 80 GB. We expect this process to occur about 5 times per DST event.

Analysis Analysis consists of low–cpu computations and histogramming. Input volume for one year's data is 80 GB.

7.6 Leptons and D–mesons

CPU estimates are based on analysis carried on on an HP C160 (10 Si95). Estimated numbers of events needed for the following physics projects are given in section 5.6.

7.6.1 $\phi \rightarrow e^+e^-$

1. μDST-level1: all electron tracks with $0.2 < \mathrm{p}_T < 1$ GeV/c which have 3 hits in the SVT and 40 hits in the TPC. The upper p_T limit is defined by the PID capabilities of the TPC and/or the kinematics of the ϕ. According to VENUS, an upper $\mathrm{p}_T$cut of $\approx$1 GeV/c should not affect the signal significantly. (Simulations show that by comparing dE/dx in the TPC and the SVT one might even have sufficient resolution for electron PID up to p = 2 GeV/c.)

   This leads to an average number of e^+e^- tracks of $\approx$100/evt. The size of the μDST is then roughly 32 kb/event; 10^7 events$\Rightarrow$320 Gb μDST. Estimated CPU time is 2 sec/event (PID, loose cuts, monitoring, conversion rejection etc.). So for 10^7 events this is $2 \cdot 10^7$ seconds CPU.

2. μDST-level2: Apply further cuts and combine all tracks to form like–sign and unlike–sign pairs. Only tracks from pairs with an invariant mass of $0.8 < m < 1.2$ GeV/c^2 are stored. Assuming a signal to background of 1:5 (conservative, we see 1:3), the μDST-level2 contains 5 pairs/event. The size should be approximately 1.6kb/event or 16 Gb for 10^7 events. The CPU needed is $\approx$1 sec/event, for a total 10^7 sec for this step. The μDST-level2 can easily be further analyzed at home institutions.

3. Iterations: Most of the data mining procedures can be checked on a much smaller sample ($\approx$10%) before the full production is submitted. CPU $= 0.1 * 2 \cdot 10^7 + 0.1 \cdot 10^7 = 3 \cdot 10^6$ sec.

Data Mining Equivalent to μDST-level1 production. Input data volume is 12 TB, output is 320 GB. Total CPU is $2 \cdot 10^5$ kSi95-sec. This will be performed a few times per DST event.

Analysis Equivalent to μDST-level2 production. Input data volume is 320 GB, output data volume is 16 GB. Total CPU is $1 \cdot 10^5$ kSi95-sec. This will be performed several times per DST event. Further analysis and data volumes are negligible.

7.6.2 $J/\psi \rightarrow e^+e^-$

Unlike the case of ϕs, hard pions contribute significantly to the J/ψ background. It is therefore most favourable to combine candidate tracks into pairs already at an early stage and accept only pairs within a mass window around the J/ψ mass. Only tracks from those pairs are stored in the μDST. Track candidates are defined as being identified as an electron with $p_T > 0.8$ GeV/c and having at least 3 SVT and 40 TPC hits. This yields about 10 candidate tracks/event, which are mostly pions misidentified as electrons. Size of μDSTs for 10^7 events is about 32 Gb. We estimate 3 CPU sec/event yielding $3 \cdot 10^7$ sec in total. CPU time is about $3 \cdot 10^7$ seconds per pass through the data.

Data Mining Input data volume is 12 TB, output is 32 GB. Total CPU is $3 \cdot 10^5$ kSi95-sec. This will be performed several times per DST event.

Analysis Further analysis and data volumes are negligible.

7.6.3 D–meson production

1. μDST-level1: Find the primary vertex, calculate the impact parameter for Kaon and pion tracks, apply cuts and reconstruct the secondary vertices. The μDST-level1 contains all tracks from secondary vertices (loose cuts); in total 400 tracks/event. Accounting for the combinatorics, required CPU time is roughly 100 sec/event, in total $3.5 \cdot 10^6$ events$\cdot$100 sec = $3.5 \cdot 10^8$ sec. Data volume is $\approx$130kB/event or 455 GB for the $3.5 * 10^6$ events.

2. μDST-level2: Scan μDST-level1 and apply stronger cuts, accept only tracks from pairs in a mass window around the D mass. Required CPU time is 1 sec/iteration/event. Probably at least two iterations are needed, hence total CPU time is $2 * 1sec * 3.5 \cdot 10^6 = 7 * 10^6$ sec. Resultant data volume is negligibly small.

Data Mining Equivalent to μDST-level1 production. Input data volume is 4 TB, output is 450 GB. Total CPU is $3.5 \cdot 10^6$ kSi95-sec. This will be performed slightly more than once per DST event.

Analysis Equivalent to μDST-level2 production. Input data volume is 450 GB, output data volume is negigible. Total CPU is $7 \cdot 10^4$ kSi95-sec. This will be performed several times per DST event.

7.7 Peripheral Collisions

For peripheral collisions, DST (μDST) production can proceed by selecting events on the basis of multiplicity, with tracks selected by requiring that they pass somewhere near the origin, with the cut chosen to eliminate background tracks but preserve those from K_s^0 and Λ decays.

Because of the low multiplicity, tracking confusion should not be a problem, so at the μDST level only trackfitting output is required, ($\sim$ 100 bytes/track), or roughly 10 Gbytes/year. This is a disk (or at the most Exabyte) sized sample. At this stage, the major computing requirement is a partial wave analysis (PWA), something that is easily undertaken on even today's workstations.

Due to the very small data volume, data mining and analysis requirements are negligible.

7.8 pp and Spin

There will be at least five different data streams:

- exclusive dijets
- inclusive photons
- inclusive electrons
- high Pt inclusive jets
- Au-Au reference data, zero bias, min bias, calibration events

If full tracking is performed by the Level 3 trigger, the average DST event will be slightly larger than the raw data event, 180kB/event, but by using Level 4 triggering it will contain only about 25% of the events. However, for the present estimate we do not make either of these assumptions, but rather assume that the DST size is 10% of that for the raw data, or 50 kB/event, for an annual DST volume of 3 TB for 60M events. The μDST should be reduced to 2.5kB/event.

CPU time: We scale the cpu time needed for p-p data mining and analysis as follows. Annual data volume for p-p is 15% of that for central Au-Au. Assuming that the cpu time is dominated by cluster and track finding, it should scale as multiplicity or data volume, so that DST production for 60M events should require $6.5 \cdot 10^6$ kSi95-sec (compare section 6.1), or about 100 Si95-sec/event. By the rule of thumb from CDF cited in section 7.5, cpu for data should be 2% of this or about 2 Si95-sec/event.

Data Mining For a data set of 60M events, processing time is $6 \cdot 10^4$ kSi95-sec. Input volume is 3 TB, ouput volume for each μDST is 150 GB or less, of which there are at least 5. Data mining on the full data set will occur a few times per DST event.

Analysis Analysis and data fitting should only be a few percent of the above CPU time, or about 100-200 kSi95-sec.

8 Simulations and Corrections

8.1 General Remarks on Simulations

Corrections for instrumental effects: In general, simulations may be used to derive corrections to the data due to the effects of finite acceptance and efficiency, and background to a given signal due to the physics of the event or the instrumental response. While much work has gone into the development of a detailed model of the STAR detector, both within the Geant simulation package for propagation of tracks through the detector and specific detector responses to generate simulated "raw data", these tools can be computationally very expensive and must be used with care. It is not possible to approach the statistics of the real data set with Geant–based simulations in which most physical effects of the detector are explicitly modelled. This is because of both the very large CPU requirements of such a model and the difficulty of verifying the detector model to that level of precision. For analyses for which very high statistics detector simulations are appropriate, parametrizations or "fast simulators" must be developed. It is important to note that such flexibility in detail of modelling is being built into the GEANT4 package, scheduled for public release in 1998.

Acceptances and efficiencies are calculated by embedding individual simulated tracks into real events, running the standard tracking, and evaluating the results. Such tracks can be chosen according to any convenient input distribution that maximizes the statistical power of the cpu needed for the reconstruction (see Appendix B). Detailed considerations of the required number of embedded tracks and reconstructed events is given for various physics categories in the subsections below.

Backgrounds are calculated by passing events from a reasonably physical event generator ("reasonable" to be defined) through a detailed, Geant-based model of the detector response (GSTAR). Experience with TPC-based detectors has shown that simulations computing requirements are dominated by the background calculations. These typically require 10 times as much CPU time per event as the reconstruction. However, typically only a tenth the number of events need to be simulated for accurate modelling of the background. Discussions of this point relevant to various physics categories are given below, but reasonable estimates of background rates cannot be made at this point and the overall requirements for GSTAR-based simulations will be established using this 10% rule of thumb. For a year's sample of 17M central Au-Au events, we estimate that about 1.7M full Au-Au central events passed through GSTAR will be needed for background calculations. GSTAR is currently benchmarked at 36 kSi95-sec/event. With the addition of the response simulation and reconstruction, we round this number to 40 kSi95-sec/event. Then 1.7M events will require about $7 \cdot 10^7$ kSi95-sec CPU and generate about 20 TB of "raw" data.

Event generators: We consider here the need for event generator simulations in addition to Geant-based simulations, in order to investigate the physics of the generators themselves and to compare them to STAR data. (We refer here to event generators in the "public domain". Additional event generator studies targeted at Event-by-Event physics will be discussed in section 8.5).Following the calculations proposed by PHENIX, we assume that a typical generator requires 10 kSi95-sec per central Au-Au event, that it is appropriate to run a given generator with a given set of switches to about $3 \cdot 10^4$ before exhausting the physics content of the generator, and that about 30 such combinations of generator and switches will be required. The output volume of an event generator is typically 100 kB per central Au-Au event, so that 10^6 events will be generated, requiring 10^7 kSi95-sec CPU and generating about 100 GB of data. As argued in section 8.5, an equal amount of event generator simulation targeted at Event-by-Event investigations will be required, giving a total of 200 GB of data and $2 \cdot 10^7$ kSi95-sec CPU reuqired.

8.2 Soft Hadronic Probes

In this section we build upon the discussion in the introduction and the Appendix B. Any "raw" spectrum, e.g. an m_T distribution, has to be corrected for geometrical acceptance and detector efficiency (usually in a single step), as well as for contamination (background). This needs to be done with high systematic accuracy, since systematic errors start dominating the spectra very early. It is therefore necessary to have a very precise modeling and estimation of all three factors (acceptance, efficiency and background).

The acceptance plus efficiency corrections are best estimated using the embedding technique. Some alignment and distortion residual uncertainties, which are basically unknown and therefore impossible to simulate, are usually not contributing much to the overall systematic error and can be partially recovered at the evaluation step. Embedding is very efficient and quick, only a small fraction of events from the analysed sample have to be used. Since the samples needed for hadronic spectra are not large anyway, embedding CPU and disk space needs can be neglected. As an illustration, we modify the example given in Appendix B.1 with more pessimistic assumptions. We divide our phase-space in 20 bins in rapidity (e.g. $y = \pm 1$ and $\Delta y = 0.1$) and 40 bins in p_T, so that $n_{bin} = 800$. Let's us also take the conservative number of 30 tracks embedded per event, which is about half the width of the fluctuations in the number of tracks in the TPC. Then for a 1% relative error we would need about 30K events. The rarer the particle species is (e.g. $\bar{p}$) the fewer the number of bins and thus the fewer the number of embedded events needed.

The background contamination correction factors are calculated using a detailed, GSTAR based description of the detector using input from an event generator which reasonably describes the data. The sources of background are mainly secondary (non-vertex) hadronic interactions of the produced particles with the detector material, gamma convertions, and weak decays. The accurate modeling of the background is the most difficult task and it dominates the systematic error of the measurements. Some iterative feed-back to the input from the data will be needed for tuning the event input as well as the whole chain in general. For our purposes, the CPU time needed per event is about 10 times the amount needed for the reconstruction of a real event. If we exercise the formula in Appendix B.2 on the data samples in section 5.1 (pion case), i.e. 10K events in the sample, one pion per bin (1% statistical accuracy) and 10% contamination, then the number of fully simulated events is about the same, 10K. For Kaons and protons, especially at higher p_T, this correction scheme breaks down, mainly because the errors will be strongly dominated by PID unfolding issues rather than background. We therefore conclude that several tens of thousand of events will probably be sufficient to model the background. Again the CPU and disk needs are not exceedingly high.

8.3 Hyperons

Corrections are usually divided into three parts: background subtraction and correction for losses due to geometrical acceptance and reconstruction efficiency. The division between the latter two is somewhat arbitrary, but generally geometrical losses are classified as those which are independent of the detailed detector response. For a Λ hyperon for example, acceptance may be defined by the number of padrows crossed by the two charged daughters. The acceptance can be calculated easily by running a large number of Λs through GSTAR with a flat rapidity-transverse momentum distribution.

Calculating the reconstruction efficiency entails finding whether an accepted Λ was properly reconstructed. As such it relies not only on the performance of the analysis code itself, but also on the detailed response of the detector and the environment in which the tracks are found. The best way to calculate the reconstruction efficiency is to embed Monte-Carlo particles, already filtered by the acceptance, at the raw data level. The events into which the Monte-Carlo particles are embedded must be representative of the entire data sample (be of the same average multiplicity, for example). Since the entire event must be reconstructed in order to determine whether the embedded particles were found, it is vital to embed particles which have already been determined to be within the acceptance of the detector.

For rare particles, for instance those which are found in less than 10% of all events, there exists the possibility of embedding only into those events in which a candidate has not already been found. This has the desirable advantage that no new criteria must be used to settle the question whether the particle found is the particle embedded. This does not, however, obviate the need for background subtraction.

Background subtraction is performed by plotting the invariant mass distribution for the candidate particle. The cuts used to extract the signal are usually adjusted so that the background around the mass peak is reasonably flat. In this case, the background can be subtracted by choosing an invariant mass window above and below the mass peak, such that the background under the peak itself can be determined by interpolation. This is undertaken as a function of rapidity and transverse momentum. The limitation is that there must be a sizeable mass peak in each bin for this method of background subtraction to be of practical use.

The implication of this for embedding rare particles is that an event can only be used if no candidates were found within the mass interval defined by the lower and upper limit of the two background regions. Embedding one or more strange particle decays into the selected events inevitably results in new background candidates. This new background must be subtracted after embedding.

An alternative approach would be to follow the Monte-Carlo hits through the embedding process and hence label those points resulting from the Monte-

Carlo. The potential advantage of this microscopic approach is that you can tell uniquely whether the tracks found were the ones added to the event. The problem with this simple picture is that an additional criterion must be used to decide whether the embedded track was found or not. An extreme example would be to imagine a track with 50% hits from a Monte-Carlo track and 50% from hits associated with the event. Was the embedded track found? This uncertainty does not invalidate the approach, but illustrates that work must be done to study the potential bias this sort of decision introduces. The benefit of following this approach is that needless reconstruction losses (due to program bugs for example) can be identified and fixed along the way. Additionally, it is likely for the more common hyperons (lambdas for example) a candidate will be found in every event. This may mean that only the point-matching approach will be applicable for these particles.

Estimating now the number of embedded events which are needed, we aim to arrive at the same total number of reconstructed particles in the embedded sample as in the data themselves. The total number that need to be embedded is then approximately (# in data)/(efficiency). The efficiency is not 10%, as might be concluded from table 4. The estimates in that table include losses due to cuts, so that the apparent reduction in rate is partly due to acceptance. Arguing backwards, a reconstruction efficiency of 50% for Λ, $\bar{\Lambda}$ and K_s^0 implies a single track reconstruction efficiency of 70%, which is probably not too far off the mark, remembering that these are secondaries.

Given a reconstruction efficiency of 50% and a total data set of 100K particles at most (for double differential cross sections), then 200K embedded particles need to be reconstructed. If we embed 1 particle per event (most pessimistic) then we would need 200k embedded events per particle. We would most likely embed more than one particle per event, the final number being dependent upon the background formed between the embedded tracks themselves. If this looks very different from the random contamination between tracks in the event and the embedded tracks then this will ultimately determine how many particles can be embedded in practice. We make a conservative estimate that up to 10 particles (20 tracks) could be embedded into each event. This would lower the number of embedded events to 20k per particle. For Ξ and Ω the efficiency will be $(0.7)^3 = 34\%$. In this case we would need 300K embedded tracks. However, we will probably not see 100k Ω in STAR!

8.4 Correlations

We use the embedding technique to assess the efficiency and resolution of measuring various components of the momentum difference $\vec{q}$ of two tracks close to each other in phase space. For the efficiency, assume that at low p_T for the pair and $|\vec{q}| > 20$ MeV/c, the efficiency is 100% and the resolution is simply $\sqrt{2}$ times the single track momentum resolution (the limit of 20 MeV will increase for stiffer tracks). We need to study the window $0 < |\vec{q}| < 20$ MeV/c with 5 bins in each direction, leading to 125 $\vec{q}$ bins for each (y, p_T) bin. To reach 1% statistical error in each bin, 10K entries per bin are needed. In section 5.3 it was concluded that 30 bins in (y, p_T) are needed for one pion species, leading to 125 q-bins·10K pairs·30 (y,p_T) bins=$4{\cdot}10^7$ embedded pairs, or $4 \cdot 10^5$ reconstructed events with 100 embedded pairs per event (see Appendix B.1), for each pion species. Assuming that the net baryon number is small in the phase space where these measurements are made, particle ID efficiency is the same for π^+ and π^- and the same calculation can be used for both. However, separate calculations are needed for K^+, K^-, and protons, but with a factor 5 fewer (y, p_T) bins for each. Thus, a total of about $6 \cdot 10^5$ events will need to be reconstructed for the calculation of efficiency and resolution at low $|\vec{q}|$.

For multiparticle correlations we will have to revisit the reconstruction issues. However, we will not perform 3D analysis here (at least not initially). We expect 15 bins at low $|\vec{q}| \cdot$ 10K pairs·30 (y,p_T) bins=$5{\cdot}10^6$ embedded sets of pairs or $\approx 5 \cdot 10^4$ events to be reconstructed.

For the K_s^0 correlation function, because the members of the pair will decay at different points we assume that reconstruction of one member of the pair will not interfere measurably with reconstruction of the other, in other words that the K_s^0 pair reconstruction efficiency equals the square of that quantity for a single K_s^0 down to $|\vec{q}|$=0. In section 8.3 it is argued that the bulk of the losses in V0 reconstruction are due simply to lifetime (i.e. acceptance). For efficiency studies via embedding of MC K_s^0 pairs in real events one can then choose singles or pairs whose decay topology generates daughters within the fiducial volume. The reconstruction efficiency for such V0s is therefore of order 50% for singles (see section 8.3) or 25% for pairs. From Table 4 it is seen that of order 1-3 K_s^0 will be reconstructed per event. To study reconstruction efficiency we embed MC pairs in real events at the rate of about 10/event. In order to achieve the same track statistics for efficiency studies via MC pair embedding as for the data (conservative upper limit), the number of embedded events to be reconstructed is approximately (# real pairs/# embedded pairs)·(1/reconstr. eff.)≈0.4·number of real events, or 1M events for 1-D and 2M for 2-D K_s^0 correlation analysis (section 5.3).

An additional simulations task is to generate about 10^4 events which contain proper Bose-Einstein and Coulomb correlations. These must be imposed by running events from an event generator into a "correlator", which either modifies the momenta of the generated particles or rejects some subset of them, requiring the combining of several uncorrelated events to obtain one correlated event. It is desirable that at least a subset of these correlated events from a generator be treated to the full simulation chain including GSTAR. In total this task represents a minor addition to the overall computing load and we do not tabulate its requirements separately.

8.5 Event–by–Event

Remarks in section 8.1 refer to the application of MC simulations to studies of acceptance, efficiencies and backgrounds. These concepts are applicable to conventional mean-value particle distributions. The use of MC generators in event-by-event analysis is somewhat different. We can identify three main areas in which simulations can be used for EbyE analysis (this list is not exhaustive):

- Event spectrum calibration - *a priori* examination of an event spectrum space using EGs containing anomalous features put in 'by hand.' This requires of order 1M events (see below), but without GSTAR. GSTAR is not needed because for this analysis one can analyze the EG momentum space output directly. The detector response remains a separate issue. One is trying to minimize CPU usage while still properly characterizing the nature of the event spectrum space.
- Null test - given that one or more anomaly classes have been identified one must demonstrate that an anomaly cannot be reproduced by hadronic or other *known* physics and/or detector effects. This should require $\leq$ 1M events (e.g., 0.1M events per anomaly class) using several EGs plus GSTAR.
- Hypothesis test - a specific model of anomalous physics (e.g., DCC) intended to account for an anomaly class is inserted into one or more EGs and put through a full simulation and reconstruction chain. This should require $\leq$ 1M events (e.g., 0.1M events per model) with several specific EGs, GSTAR and reconstruction.

What does this imply for Offline simulations? The number of events required for event spectrum calibration will be of order 1M. This simulation should provide a survey of different anomaly types and amplitudes to map an event spectrum. This implies of order 10-30K events for each of 10-30 type/amplitude combinations. However, this does not mean GSTAR/tracking processing of these events. It is possible that the EG output can be directly used for event-spectrum calibration purposes. We are not studying detector response in this process, we are calibrating a spectrum. This should significantly reduce the computation requirement for this task. This will require 1M events with special catalogued correlation content not covered in STAR 'standard' EG runs.

Anomalous events used for calibrations or hypothesis tests must be described and catalogued in detail to enable interpretation of results. This is a sizable program. The EG(s) used to carry out this study must be carefully chosen on the basis of a tradeoff between runtime and realism. One of the more important aspects of this simulation is to determine the degree of attenuation of early- and intermediate-stage correlation development (due to "plasma" formation or other symmetry change) by later hadronic rescattering.

The detector response enters into the event-spectrum calibration process at some point because the correlation content of particle spectra will be altered by detector effects (e.g. two-hit resolution, distortions, efficiencies, background due to interactions). This "sensitivity reduction" may be represented at some point by a transfer-function approach, determined with a reduced event sample by standard full simulation techniques and an elementary EG.

Simulations are carried out for mean-value distributions in part for background corrections. If the background in a particle spectrum exhibits no unusual EbyE variations then it is 'invisible' to EbyE analysis (assuming it does not overwhelm the desired particle species - a form of attentuation). However, if there are sources of *nonstatistical* background with trivial origin (e.g., fluctuations due to hadronic processes in material) then EG-GSTAR background studies will be quite important for EbyE. Because of the computational expense of detailed, GSTAR-based simulations, care must be taken to maximize the background discrimination power of the event analysis. In general, the approach to backgrounds, if any, for EbyE will be different than for spectra. It should be a bottom-up process rather than a top-down process. That is, study of specific effects found in anomalous events should justify EG studies rather than routine large-volume EG studies searching for possible noise sources for EbyE analysis.

In summary, we expect that the dominant CPU usage for EbyE analysis will consist of GSTAR-based simulations of 1-2M events per year. Some of this may be carried out in common with the rest of the STAR physics program but a significant portion (up to 50%) is expected to be of an exceptional nature.

8.6 High $\mathbf{p}_T$

The "purity" of finding jets or γs is high in the sense that the background to 20 GeV jets is 15 GeV jets and the background to 10 GeV photons is 12 GeV jets. The detector doesn't create jets or photons out of whole cloth, but it can mismeasure or we can misinterpret what we see. To assess these backgrounds, simulation of high p_T datasets by a fast parametric simulator is appropriate. Although simulated datasets that are larger than real data by some reasonable factor (e.g. 5-10 times) are desired, these do not have to be simulated down to the hit level. Typically, other experiments use parametric simulations, smearing the detector response to some known parameterization. That being the case, CPU usage will be minimal, as will disk space (as only the highest level of objects need to be saved.)

In the technique we propose, we don't treat jets like single particles - embed them in an event and see if we can find it again. What we do is we take single particles, embed them, and determine the reconstruction efficiency and resolution as a function of p_T, η, and track density. Once we have those, we can build up jets and do all the studies mentioned, just using parameterizations of the resolution and efficiency functions. This is expected to require minor computing resources.

It will be necessary to overlap ("embed") photons and jets and reconstruct the equivalent about 50K real events. Jet overlapping is dsirable because: a) there are global event variables that have to be considered (e.g. energy balance), b) we don't reconstruct all of a jet and overlapping only part of the jet may be misleading, c) by overlapping MC tracks and showers on real events, we can build up the effective jet overlapping by overlapping constituents.

In addition, simulations of high p_T particles, γs and jets are needed for the following studies:

- The effects of the $\eta = 1$ gap in the calorimeter (how does the reconstructed

p_T compare to the true p_T when the particle is near $\eta = 1$, etc.)

- The effects on jets in AA collisions: how are the p_T and η smeared in the presence of many other tracks and particles
- The reconstruction efficiency of jets and photons with pileup at high lumminosities, especially for polarized pp (see section 8.9).

8.7 Leptons and D–mesons

Simulations for the lepton pairs analysis are needed for: (a) tuning the cuts and selection mechanism before the different data mining steps, (b) determining the efficiency and acceptance for the normalization, and (c) studying the background to evaluate the systematics on the signal. Experience shows that a small sample of Monte Carlo events with an significantly enhanced signal (e.g. $\sim 50\ \phi \to e^+e^-$ per event) is already sufficient for task (a). For STAR on the order of $\sim$ 2K central events need to be generated with a fast detector response simulator only. Pure event generator studies (PYTHIA, Genesis) are imperative and a reasonable sample of $\sim$ 50K events can be easily generated at low cost. To fulfil the requirements of task (b), a sample of $\sim$ 5K events with a considerable number (50-100) of embedded signal pairs per event is needed. Many other studies for (b) can be performed using the data themselves. The largest sample ($\sim$ 10K) is needed for the background studies (c), which, should consist of standard MC events (i.e. no enhanced embedded signal). Since many other physics analysis studies need this type of generated events, these datasets do not need to be specially produced but can be shared with other projects. We conclude that the overall simulation requirements for lepton-pair studies are minimal.

An estimate for the D-mesons simulations is currently rather difficult. Since the technology for additional vertex detectors is not determined yet (SDD, pixel, CCD) we can only speculate about the actual requirements for simulating these devices[3]. The simulations are needed to determine the efficiency and acceptance of the D decay products as well as the background. This involves studies of TPC data and the vertex detection devices and their reconstruction power. One can assume that the understanding of the TPC data will already be very much advanced when the first D-meson measurements are carried out, which reduces the simulations requirements significantly. One can envision a scheme where a 100% efficient TPC is assumed and its actual efficiency is later folded into the results from the simulations of secondary vertex reconstruction. Again, events with an enhanced D-meson signal of $\sim$ 50/event can be used. Assuming a reconstruction efficiency of 1%, an effective signal of $S_{\text{eff}} = 1000$ could be obtained in $\sim$ 20K events, sufficient to calculate corrections for a much smaller signal. This effort is small and the resources needed are negligible compared to the actual data analysis of several millions of events.

[3]For example, a detailed study of a Silicon Drift Detector is much more demanding than for a pixel based device.

8.8 Peripheral Collisions

Because multi-track confusion is not a problem, our simulation requirements are similarly modest. We can use the fast simulator (or data) to parameterize resolution (including dE/dx) and tracking efficiency as a function of p, p_T, ϕ and η, and then use this parameterization in our Monte Carlo. This parameterization should be a standard part of the STAR analysis software.

One area of special concern is particle identification via dE/dx. While this is primarily studied by looking at data, we also need well tuned simulations to fill in the gaps and cover the entire solid angle and momentum space. This should be part of the overall STAR PID effort; in the worst case, we would need to simulate at most 10M tracks, equivalent to 4000 central events.

We also need to simulate several types of backgrounds. Most of these can be evaluated at the 4-vector or fast simulator level, with correspondingly limited computer requirements; we have already generated over 500,000 Fritiof and Venus peripheral collision and beam gas events without straining our resources.

However, we will also need to study triggering efficiency at low multiplicity, both for tracks coming from the interaction diamond and from outside it. Some of these studies will require detailed simulations, at least of the trigger detectors. With pre-selection of appropriate events, a complete simulation would probably be required for no more than 100,000 events (1,000 events for 3% accuracy times 10 event origins, spread along the beampipe, times 10 models), with an average of 4 tracks each. This corresponds to the equivalent of about 100 central events.

8.9 pp and Spin

Simulations will be mostly carried out with a fast parameterized simulator as proposed for high p_T in section 8.6, with one possible exception described below. The number of simulated events should not exceed the number in the data set. These simulations are very rapid compared to Au-Au events, and can be handled on a physicist's workstation. The computing load is negligible.

For unpolarized pp events, these studies will be most important in the first few years, when sizeable polarized pp data sets will not exist. Perhaps on the order of a few$\cdot 10^7$ events will be required. One particular problem will be the comparison of the relative triggers for pp and pA data. There is a fair probability that there will be no particles in the forward calorimeters in most pp events, while this is much less likely for pA and especially for AA interactions. The effect on the determination of the interaction point must also be considered by simulations.

For polarized pp interactions, one need for simulations will be to determine the constants in the equations linking the observed asymmetries to the related polarized structure function as a function of x. For this purpose a few$\cdot 10^8$ events will be required. A few workstations can generate these events in a timely manner. Simulations of π^0 decays that appear as direct photons will also be needed, and the number of such events should again not exceed the number in the data set.

The main concerns in the analysis of the polarized pp data are for stability of the beam properties and the detector efficiency, acceptance, and resolution in order to cancel systematic errors. Knowledge of absolute efficiencies or acceptances is not as important. If the beam properties (polarization, intensity, bunch length or size, etc.) differ from bunch to bunch, or if there are efficiency changes (perhaps due to pileup in the SVT and TPC) in STAR, then simulations may be necessary to correct for systematic errors. To be conservative, it will be assumed this is needed every year as the luminosity of RHIC improves and the pileup effects become more severe. Assuming the smallest asymmetry to be measured is 0.001 with an error of ± 0.001, then the efficiency, acceptance, and resolution must be known to a precision on the order of ± 0.001. Taking pileup in combination with differences in beam bunch properties as the most likely problem, then there will be efficiency changes for individual tracks or resolution changes for reconstructed jets. The associated systematic errors can be estimated by embedding tracks in pp events and parameterizing the results in terms of η, track momentum, and localized track density or luminosity. An upper limit to the number of such embedded tracks is $2.5 \cdot 10^7$ (250K tracks for 4 values of eta, 5 of momentum, and 5 of track density). Assuming 10 tracks embedded per reconstructed event, this gives a total of $2.5 \cdot 10^6$ events with embedded tracks to be analyzed per year.

9 Scaling to total STAR requirements

In previous sections we have restricted discussion to the computing requirements for a few specific analysis topics within each STAR physics category. In this section we address the question of how to scale these estimates to correspond to the computing requirements of the STAR collaboration as a whole. We attempt to guess at the future behaviour of the STAR Collaboration from evidence supplied to us by the collaboration itself: membership in the Physics Working Groups, as indicated by the number of collaborators subscribing to the email distribution list for each PWG. Table 6 gives the number of email subscribers for each of the PWG groups where such a list exists, and an arbitrary number of 20 (10% of total) for those groups whose lists we are not aware of. No attempt has been made to account for double counting of members.

We propose to use the percentages given in the last column of Table 6 to scale the computing estimates for each physics category determined in sections 5, 7, and 8.

Tables 7 and 8 summarize the annual requirements for data mining and analysis for the selected analyses discussed in the previous sections. The ratio of CPU rquired over total input data volume is given in the last column of both tables, and serves as the measure of whether a process is CPU or I/O intensive. By various measurements, the lower bound for an I/O intensive task is found to be about 1 Si95-sec/MB.

Table 9 summarizes the total CPU needed by STAR for data mining and analysis, assuming 100 analysis projects. A mean CPU requirement for a given physics category is calculated by weighting moderate and hard projects in the ratio .8/.2. We assume that about 5 data mining passes are carried out on each DST-level event per physics category (i.e. this specifies how many times the DST data set must be read). Note that there is also a hidden assumption here that different analysis projects within the same physics category will share some data mining passes. Likewise, we assume that each separate project will carry out about 20 analysis passes per DST production, in this case independent of other projects within the same physics category. A large variation in the number of data mining passes would have significant consequenses for the required data access bandwidths.

Table 10 summarizes the total data volume of μDST for STAR, assuming 100 analysis projects. These are the data for which frequent and low latency access is required, and the data volume should be compared to the available disk space at RCF. Where projects within a physics category will share a common μDST the number of projects is set to 1, as indicated.

Table 14 summarizes the simulations requirements appropriate to a dataset of 17M central Au-Au events.

10 Tables

The following Tables 1 to 14 are discussed in the foregoing text. Tables 15 to 19 have been specified by the RCF in order to summarize STAR's computing needs, and represent the same information in a different format. In particular, Tables 17 to 19 present an overall summary of STAR's computing requirements.

Unless explicitly stated otherwise, the tables represent STAR computing requirements for a simple run plan consisting of 17M central Au-Au events. Raw data volume and DST production time for p-p analysis can be a significant fraction of that for the analysis of heavy ion data, especially in the early years of STAR when online filtering of pileup in high luminosity p-p events is not fully developed. Note in particular that Tables 15 to 19 are appropriate to this simple run plan and do not include requirements for p-p running.

Quantity	Value	Comments
RHIC year	$1.4 \cdot 10^7$ sec	
Combined RHIC/STAR duty factor	.67	
Data recording rate	20 MB/s	
effective STAR year for HI	$1.0 \cdot 10^7$ sec	
Data volume (central Au-Au)	12 ± 4 MB	12 MB from Hijing
Central Au-Au events per year	$1.7 \cdot 10^7$	
effective STAR year for p-p	$4 \cdot 10^6$ sec	2/3·10 weeks
p-p events recorded per year	$6 \cdot 10^7$	recorded at 15 Hz
Combined RCF/STAR duty factor for reconstruction	.75	
Avg # of times each raw event processed	1.5	
dN/dy(charged) at $y_{cm} = 0$	1200	primaries, Hijing
	2200	primaries, Venus
Units		
SPECint92	SPECint95/40	abbreviated Si92, Si95
MFlop	SPECint95/13	
Example: HP 735/125	4 SPECint95	

Table 1: Basic assumptions and units.

Physics Category (section)	Observable	Particle	# central Au-Au events
SOFT HADRONIC PROBES (5.1)	dN/dy	$\bar{p}$	50K
	<p_T >	$\bar{p}$	20K
	dN/dydp_T	K^0_s(5.2)	50K
		$\bar{p}$	200K
HYPERONS (5.2)	dN/dydp_T	Λ	1M
		Ξ^-	20M
	dN/dy	Ξ^-	2M
	total yield	Ω^-	20M
CORRELATIONS (5.3)	R_{out}, R_{side}, R_{long}vs. y, k_T	$\pi\pi$, p_T>1 GeV/c	2M
		$K^{\pm}K^{\pm}$	2M
		pp	2M
	2-D	$K^0_sK^0_s$	4M
	multiparticle	pions	2M
EVENT–BY–EVENT (5.4)			17M (all events)
HIGH p_T (5.5)	p_T> 10 GeV	γ	>17M central + γ trigger
	p_T> 20 GeV	Jets	>17M central + jet trigger
LEPTONS AND D–MESONS (5.6)	dN/dp_T	$\phi \to e^+e^-$	10M
		$J/\psi \to e^+e^-$	10M
		D mesons	3.5M

Table 2: **Data sets:** Required number of events for central Au–Au collisions at $\sqrt{s} = 200$ GeV for a variety of STAR observables, for processes with both moderate rates and low rates (the latter determine the size of the dataset). These are to be considered the input data sets to the data mining process.

Physics Topic (section)	Observable	# Events	# equivalent central Au-Au events
PERIPHERAL COLLISIONS (5.7)	all, 1 year's running	20M	15K
PP AND SPIN	all, 10 week's running	60M	2.5M
(5.8)	A_{LL}	260M	11M
EMC Calibration	π^0, noncentral Au-Au		10K

Table 3: **Data sets:** Number of events written to tape per year for non–nuclear collisions in STAR, and largest required data set (multi–year accumulation).

Particle	4π (Hijing)	$>$ 9 TPC Hits	Reconstructed in TPC+SVT	Reconstructed in TPC
K_s^0	250	30	3	1
Λ	80	5	0.5	0.1
antiΛ	50	5	0.5	0.1
Ξ	5	0.3	0.003	
antiΞ	5	0.3	0.003	
Ω	0.05	0.003	$3 \cdot 10^{-5}$	
antiΩ	0.05	0.003	$3 \cdot 10^{-5}$	

Table 4: Estimated rate per event from Hijing of K_s^0 and various hyperons predicted by the Hijing model and rates of reconstruction in STAR for the TPC alone or for the TPC+SVT.

Particle	Yields	dN/dy	dN/dydp$_T$
K_s^0	.5K	5K	50K
Λ	10K	100K	1M
Ξ	200K	2M	20M
Ω	20M		

Table 5: Estimated numbers of events needed for various measurements of K_s^0 and hyperon spectra, based upon rates in Table 4 for TPC+SVT as well as NA49 experience.

Physics Category	# subscribers	% of total
Soft Hadronic Probes	30	12
Hyperons	42	16
Correlations	20	8
Event–by–Event	50	19
High p_T	38	15
Leptons and D–mesons	20	8
Peripheral Collisions	19	8
pp and Spin	39	15
Total	258	

Table 6: Estimate of fraction of total number of STAR analysis projects within each physics category, determined by fraction of total number of subscribers to each email distribution list. Where we are not aware of a list, an arbitrary number of 20 subscribers has been assigned.

Physics Category (section)	Observable	Input Volume (GB)	Output Volume (GB)	CPU (kSi95-sec)	CPU per Input Vol (Si95-sec/MB)
SOFT HADRONIC PROBES (7.1)	dN/dp$_T(\bar{p})$	240	24	(240)	(1)
HYPERONS	dN/dydp$_T$ (Λ)	$1.2 \cdot 10^3$	60	$4 \cdot 10^4$	30
(7.2)	total yield (Ω)	$2.4 \cdot 10^4$	$1.2 \cdot 10^3$	$8 \cdot 10^5$	30
CORRELATIONS (7.3)	all	$5 \cdot 10^3$	100	$(5 \cdot 10^3)$	(1)
EVENT–BY–EVENT (7.4)	2D SCA	$8 \cdot 10^3$	$1 \cdot 10^3$	$8.5 \cdot 10^6$	1000
HIGH P$_T$ (7.5)	Jets, γs	$2 \cdot 10^4$	80	$8 \cdot 10^5$	40
LEPTONS AND D–MESONS	$\phi \rightarrow e^+e^-$	$1.2 \cdot 10^4$	320	$2 \cdot 10^5$	20
(7.6)	$J/\psi \rightarrow e^+e^-$	$1.2 \cdot 10^4$	32	$3 \cdot 10^5$	30
	D mesons	$4 \cdot 10^3$	450	$3 \cdot 10^6$	750
PERIPHERAL COLLISIONS (7.7)	all	20	small	(20)	(1)
PP AND SPIN (7.8)	all	$3 \cdot 10^3$	150	$1.2 \cdot 10^5$	40

Table 7: **Data Mining:** Annual data volume, CPU, and their ratio for data mining for typical and intensive projects, for a single experimental configuration. CPU times in parentheses are derived from Input Volume assuming the minimum required to read data is 1 Si95-sec/MB, and are for one pass of the input data.

Physics Category (section)	process	Input Volume (GB)	CPU (kSi95-sec)	CPU/IO (Si95-sec/MB)
SOFT HADRONIC PROBES (7.1)	dN/d$p_T(\bar{p})$	24	(24)	(1)
HYPERONS	dN/dydp_T (Λ)	60	(60)	(1)
(7.2)	total yield (Ω)	$1.2 \cdot 10^3$	$(1.2 \cdot 10^3)$	(1)
CORRELATIONS (7.3)	make corr fn	100	$4 \cdot 10^4$	75
EVENT–BY–EVENT (7.4)	see section 7.4			
HIGH P_T (7.5)	Jets, γs	80	(80)	(1)
LEPTONS AND D–MESONS (7.6)	$\phi \rightarrow e^+e^-$	320	$1 \cdot 10^5$	300
	$J/\psi \rightarrow e^+e^-$	32	(32)	(1)
	D mesons	450	$7 \cdot 10^4$	200
PERIPHERAL COLLISIONS (7.7)	all	small	small	(1)
PP AND SPIN (7.8)	all	150	100	1

Table 8: **Analysis:** Annual data volumes, CPU, and their ratio, for data analysis. Typical and intensive projects for a single experimental configuration are shown. CPU times in parentheses are derived from Input Volume assuming the minimum required to read data is 1 Si95-sec/MB, as indicated.

Physics Category	Data Mining mean CPU	# passes per DST prod	Analysis mean CPU	projects * passes per DST prod	Total (kSi95-sec)
SOFT HADRONIC PROBES	240	5	24	12 * 20	$7 \cdot 10^3$
HYPERONS	$2 \cdot 10^5$	5	300	16 * 20	$1 \cdot 10^6$
CORRELATIONS	$5 \cdot 10^3$	5	$4 \cdot 10^4$	20 * 5	$4 \cdot 10^6$
EVENT–BY–EVENT	$8.5 \cdot 10^6$	5	8	19 * 20	$4 \cdot 10^7$
HIGH P_T	$8 \cdot 10^5$	5	80	15 * 20	$4 \cdot 10^6$
LEPTONS AND D–MESONS	$8 \cdot 10^5$	5	$1 \cdot 10^4$	8 * 20	$6 \cdot 10^6$
PERIPHERAL COLLISIONS	20	5	small	8 * 20	100
PP AND SPIN	$1.2 \cdot 10^5$	5	100	15 * 20	$6 \cdot 10^5$
Total (kSi95-sec)	$5 \cdot 10^7$		$5 \cdot 10^6$		$6 \cdot 10^7$

Table 9: **Total CPU for Data Mining and Analysis** Summary of annual CPU requirements **in units of kSi95-sec** extracted from Tables 7 and 8. Where appropriate, mean CPU per project is weighted between moderate and hard projects in the ratio .8/.2. For Data Mining, number of passes over full data set per DST production run is specified. For Analysis, same quantity specified but multiplied in addition by number of projects, from Table 6. Total is product of 2nd and 3rd columns plus product of 4th and 5th columns.

Physics Category	# projects	μDST vol, moderate (GB)	μDST vol, hard (GB)	wghtd total (TB)
Soft Hadronic Probes	12	24		.3
Hyperons	16	60	$1.2 \cdot 10^3$	5
Correlations	8	100		1
Event–by–Event	19→3	1000		3
High p_T	15→1	80		.1
Leptons and D–mesons	8	320	450	3
Peripheral Collisions	8	small	small	small
pp and Spin	15→1	150		.2
Total				13

Table 10: **Total Data Volume for μDST:** Estimated annual μDST volume in GB per project and TB per total, by physics category. Where appropriate, moderate and hard projects weighted in ratio .8/.2. Data volume scaled by number of projects, assuming 100 projects in total (Table 6), except for categories where a single μDST serves all analyses, as indicated.

Physics Category	Total Data Mining CPU (Si95)	Input Rate per Process (MB/Si95-sec)	Total Input Rate (MB/sec)	Ratio of Ouput to Input Volume	Total Output Rate (MB/sec)
SOFT HADRONIC PROBES	.08	1	.08	.1	.008
HYPERONS	67	.03	2.0	.1	.2
CORRELATIONS	1.7	1	1.7	.1	.2
EVENT–BY–EVENT	2800	.001	2.8	.1	.3
HIGH P_T	270	.03	8.1	.004	.03
LEPTONS AND D–MESONS	270	.01	2.7	.02	.05
PERIPHERAL COLLISIONS	.007	1	.007	small	small
PP AND SPIN	40	.03	1.2	.05	.06
Total	3400		18		.8

Table 11: **Data Rates for Data Mining:** Summary of bandwidth requirements for Data Mining. Total CPU in units of Si95 are derived from Table 9 via: CPU (kSi95-sec) * # DM passes per DST * # Reconstruction passes(=1.5) / (duty factor (=.75) * $3 \cdot 10^7$ sec). Input bandwidth per process and ratio of input to output volumes are from Table 7. Input total bandwidth is product of columns 2 and 3; output total bandwidth is product of columns 4 and 5.

Physics Category	Total Analysis CPU (Si95)	Input Rate per Process (MB/Si95-sec)	Total Input Rate (MB/sec)
Soft Hadronic Probes	.4	1	.4
Hyperons	6.4	1	6.4
Correlations	200	.01	2
Event–by–Event	see section 7.4		
High p_T	1.6	1	1.6
Leptons and D–mesons	100	.01	1
Peripheral Collisions	small	1	small
pp and Spin	2	1	2
Total	310		13

Table 12: **Data Rates for Analysis:** Summary of bandwidth requirements for Analysis. Total CPU in units of Si95 are derived from Table 9 via: CPU (Si95-sec) * # projects * # passes per DST * # Reconstruction passes(=1.5) / (duty factor (=.75) * $3 \cdot 10^7$ sec). Input rate per process and ratio of input to output volumes are from Table 8. Total input rate is product of columns 2 and 3.

Physics Category (section)	Observable	# Embedded Events per analysis	# Number Background Events
SOFT HADRONIC PROBES (8.2)	most	10^4	$5 \cdot 10^4$
HYPERONS (8.3)	Λ, Ω	$2 \cdot 10^4$	
CORRELATIONS (8.4)	pions, kaons, proton K_s^0(2D)	$6 \cdot 10^5$ $2 \cdot 10^6$	$1 \cdot 10^5$
EVENT–BY–EVENT (8.5)			$1 \cdot 10^6$
HIGH P_T (8.6)		50K	0
LEPTONS AND D–MESONS (8.7)	$\phi \rightarrow e^+e^-$	$1 \cdot 10^4$	0
	D mesons	50K	?
PERIPHERAL COLLISIONS (8.8)	all	0	4K Au-Au central equivalent
PP AND SPIN (8.9)		$2.5 \cdot 10^6$	

Table 13: **Simulations for Corrections to Data:** Estimated numbers of events for correction of data based upon Geant simulation: "Embedded" refers to superposition of a small number of simulated tracks on a real event followed by reconstruction of those events; "Background" refers to full simulation and reconstruction of physics events from a representative event generator.

Process	# events	CPU per event (kSi95-sec)	Total CPU (kSi95-sec)	MB/evt	Total Vol (TB)
EvGen (RHIC-wide)	$1 \cdot 10^6$	10	$1 \cdot 10^7$	.1	.1
EvGen (STAR-specific)	$1 \cdot 10^6$	10	$1 \cdot 10^7$	.1	.1
Embedded tracks	$2 \cdot 10^6$	2.5	$5 \cdot 10^6$	2	4
Geant for background	$1.7 \cdot 10^6$	40	$7 \cdot 10^7$	12	20
Total			10^8		24

Table 14: **Total CPU and Data Volume for Simulations:** Summary of requirements from Section 8 for annual requirement for simulations appropriate to a dataset of 17M central Au-Au events. Event generators (EvGen) are divided into generic studies with RHIC-wide applicability, which could be a common task among the experiments, and STAR-specific studies, for instance for Event-by-Event Physics (section 8.5). "Embedded tracks" refers to the full reconstruction of data events in which Monte Carlo-generated tracks have been embedded, for acceptance and efficiency studies. "Geant for background" refers to full GSTAR simulation of physical events, including detector response simulations and full reconstruction.

Data Analysis Requirements		Comments
Raw Data Recording Rate (MB/sec)	20	
Assumed RHIC/STAR seconds per running year	$1 \cdot 10^7$	
Raw event size (MB)	12±4	
Raw Data Volume (TB)	200	
Calibration Data Volume (TB)	negligible	
Required Reconst. Latency (Hours)	12 hours for 1% of data	online monitoring
Raw Data Reconst. CPU, 1 Pass (Si95-sec/evt)	2500	
Average Number of Reconst. Passes	1.5	Also assume RCF/STAR duty factor = .75
Event Summary Data (DST) Volume (TB)	20	
Other Reconst. Output Data Volume (TB)	0	
Number of Analyses Projects	100	
Typical Analysis Data Mining: Passes Required Per Year	5	
Required Input DST Volume (TB)	1→20 (4)	
Natural Density in Greater DST (fraction)	.001-1 (.01)	estimate; impossible to establish global average
CPU-sec Required/MB of DST (Si95-sec/MB)	$1 \rightarrow 10^3(30)$	see table 7
Output μDST Volume (TB)	$.005 \rightarrow 2$ (.1)	see table 7
Number of Concurrent Useful μDSTs	100	
Ratio of CPU to I/O Intensive Analyses	1:4	see table 8
Typical CPU Intensive Analysis:		
Passes Required Per Day	.1	
μDST Volume (TB)	.5	
CPU-sec Required/MB of μDST (Si95-sec/MB)	200	
Typical I/O Intensive Analysis: Passes Required Per Day	.1	
μDST Volume (TB)	.1	
CPU-sec Required/MB of μDST (Si95-sec/MB)	1	lower limit for data access

Table 15: **RCF table:** requirements for data analysis. Numbers in paratheses are approximate averages for quantities with large range.

Simulation Requirements		Comments
Modeled Data Volume (TB)	.2	
Modeling CPU (Si95)	600	
Simulated Data Volume (TB)	20	
Simulation CPU (Si95-sec/evt)	36000	
Simulated Data Reconst. CPU(Si95-sec/evt)	2500	
Simulated Event Summary Data Volume(TB)	2	
Typical Simulated Data Mining: Passes Required Per Year		
Required Input DST Volume (TB)		
Natural Density in Greater DST (fraction)		
CPU-sec Required/MB of DST(Si95)		
Output mDST Volume (TB)		
Number of Concurrent Useful mDST's		
Typical CPU Intensive Simulated Analysis: Passes Required Per Day		
mDST Volume (TB)		
CPU-sec Required/MB of mDST(Si95)		
Typical I/O Intensive Simulated Analysis: Passes Required Per Day or Week		
mDST Volume (TB)		
CPU-sec Required/MB of mDST (Si95)		

Table 16: **RCF table:** Requirements for simulations.

Requirement		Comments
Raw data	200	
Calibration data	.1	
Event Summary (DST) data	30	include reprocessing factor 1.5
Other Reconstruction Output Data	small	
Data Mining Output (μDST) Data	13	
Disk Resident Raw Data	2	
Disk Resident DST Data	2	
Disk Resident μDSTData	15	
Model Data	.2	
Simulated Data	20	
Simulated DST Data	2	
Simulated μDST data	.2	
Disk Resident Simulated Data	.2	
Disk Resident Simulated DST Data	.2	
Disk Resident Simulated μDST Data	.2	
Total	266	

Table 17: **RCF table:** Data Storage Volume Estimate (TB) for nominal year running (2001). Disk resident data calculated by standard RCF formula: 1% of raw data, 10% of current DST data, 100% of current μDST data.

Requirement		Comments
Event Reconstruction	2800	reprocessing factor 1.5; RCF/STAR duty factor .75
CPU Intensive Data Mining	3400	
I/O Intensive Data Mining	1	
CPU Intensive Analysis	250	
I/O Intensive Analysis	20	
Modeling	600	
Simulation	2800	
Reconstruction of Simulation	280	
Data Mining of Simulation	240	
CPU Intensive Analysis of Simulation	25	
I/O Intensive Analysis of Simulation	2	
Total	10100	

Table 18: **RCF table:** CPU Estimates (SPECint95) for nominal year running (2001).

Requirement	Rate (MB/sec)	Comments
Raw Data Recording Rate	20	
Reconstruction Input Rate (MB/sec)	14	reprocessing factor 1.5, duty factor .75
Event Summary (DST) Data Output Rate (MB/sec)	1.4	
Other Reconstruction Ouput Rate	small	
CPU Intensive Data Mining Input Rate	16	Table 11
I/O Intensive Data Mining Input Rate	2	Table 11
Data Mining Output Rate	1	Table 11
CPU Intensive Analysis Input Rate	3	Table 12
I/O Intensive Analysis Input Rate	13	Table 12
Simulated DST Read Rate	1	
CPU Intensive Simulated μDST Read Rate	.3	
I/O Intensive Simulated μDST Read Rate	1	
RCF Located Totals	73	

Table 19: **RCF table:** I/O Rate Estimates (MB/sec) for nominal year running (2001). Data rates are from Tables 11 and 12. CPU-intensive processes are those having input rates much less than 1 MB/Si95-sec; I/O intensive processes are all others. Reconstruction Input Rate is calculated via: 12 MB/evt * 17M events * reprocessing factor (=1.5) / (duty factor (=.75) * $3 \cdot 10^7$ sec).

A Charge to the Task Force from the STAR Spokesman

Objective: Review and update STAR offline computing requirements. Provide this as input to STAR Spokesman and the RHIC Task Force, which is responding to the recommendation of the recent RHIC Offline Computing Review that offline requirements should be updated.

Charge: Develop the offline computing requirements (including compute cycles, data storage on shelf tapes, robotic tape storage and disc, network bandwidth between storage and local and remote processors, etc.) for the analysis and interpretation of STAR data from running years 1 (assumed to begin in October 1999) and future running scenarios. Consider the full STAR physics agenda including nucleus-nucleus physics (pp, pA, and AA), spin physics and peripheral collision (two-photon) physics. Develop a reasonable scenario for event reconstruction, data analysis (including DST and micro-DST stages) and simulations needed to correct and interpret the data and for comparisons of data with appropriate physics models. It is expected that the analysis of STAR offline computing requirements will be based on the scenario that you develop. Indicate analysis topics or specialized analysis approaches that drive particular computing requirements, so that the STAR Collaboration can consider their impact on offline computing requirements and resources. Analyses with unusual cpu to I/O requirements should also be flagged. The precision of this analysis is clearly limited by very significant uncertainties. It seems not to be worthwhile to aim for a precision of better than a factor of two.

Approach: A task force reporting to the STAR Spokesman and consisting of STAR physicists with recent and relevant experience in specific physics analysis approaches will be asked to do this reassessment of STAR offline computing requirements. The task force will be chaired by Peter Jacobs (LBNL), with Lanny Ray (Texas) and Torre Wenaus (BNL) as ex officio members. It will consist of the following STAR Collaborators who will concentrate on the physics topics listed by their names.

Topic Task Force members:

Chair P. Jacobs (LBNL)

Soft hadronic physics W. Llope (Rice), S. Margetis (Kent State)

Hyperon physics K. Wilson (WSU), P. Jones (Birmingham)

Leptons and D-mesons T. Ullrich (Yale)

High Pt, jets and photons T. LeCompte (ANL), W. Christie (BNL)

Event-by-event physics T. Trainor (Washington), I. Sakrejda (LBNL)

Spin physics G. Eppley (Rice), H. Spinka (ANL)

Peripheral collision (two-photon) physics S. Klein (LBNL)

Particle correlations (HBT) S. Pandey (Wayne State), N. Xu (LBNL)

Ex Officio Lanny Ray (Texas), Torre Wenaus (BNL)

Torre Wenaus, Head of STAR Computing and Software, will provide an information channel between the STAR task force and the RHIC Offline Computing management. He will also advise the task force on issues related to the impact of STAR requirements on the RHIC Computing Facility and other STAR computing resources and on the overall STAR software. This task force will be provided with information concerning the current state of planning for STAR data taking and physics analysis by the Chair of the STAR Runtime Committee and the Convenors of the STAR Physics Working Groups, respectively. The task force is asked to submit a preliminary report to the STAR Spokesman by October 1, and a final report to the Spokesman by Nov 1.

B Comments on Simulations

B.1 Efficiency and Acceptance Calculations via MC Embedding

We discuss here techniques to determine the single track acceptance and efficiency correction factors. While the tracking acceptance and efficiency are conceptually different, in practice they are interdependent and need to be estimated together. Indeed, the only meaningful correction factor is the product of the two. In the following we refer to this product as simply the "efficiency".

It has been found by a number of collaborations working at Bevalac through SPS energies that in the very high track density environment of high energy heavy ion collisions, single track reconstruction efficiency cannot be reliably estimated based purely upon Monte Carlo modelling. Evidently, the MC simply cannot reproduce the event environment with sufficient accuracy without the investment of enormous effort and computing time. A practical alternative is to generate "raw data" for single tracks by Monte Carlo means and embed these into real events. These hybrid events are then sent through the track reconstruction procedure, and the probability for the embedded MC track to have been reconstructed by this procedure can be evaluated by a variety of means (phase space matching, point matching,...). The number of MC tracks embedded in a single real event must be much less that the number of tracks in the event itself, in order not to perturb the event environment. The amount of CPU time required to generate these MC tracks is small compared to the time needed to track the complete event. Thus, the CPU time and data volume required for the efficiency estimate are directly proportional to the number of events reconstructed, which for a given statistical precision is inversely proportional to the number of MC tracks embedded per event. These conflicting requirements - minimizing CPU time while not perturbing the environment - require careful optimization.

We give here a rule of thumb for estimating the number of embedded events needed for single particle spectrum analysis at STAR. Let

n_{bin}	$=$ # phase space bins in analysis
$n_{TrackPerBin}$	$=$ # MC tracks embedded per bin
n_{embed}	$=$ # MC tracks embedded per real event
σ_{eff}	$= \sqrt{1-\varepsilon}/\sqrt{\varepsilon \cdot n_{TrackPerBin}}$

where σ_{eff} is the desired relative error in the efficiency estimation and ε is the estimated efficiency (the efficiency appears because the the number of reconstructed embedded tracks will be binomially distributed). Note that σ_{eff} probably doesn't need to be smaller than .01 (except for spin measurements), so that for $\varepsilon = .9$, $n_{TrackPerBin} = 10^3$.

The required number of embedded and tracked events to reach this statistical precision is then:

$$N_{event} = \frac{n_{bin}}{\sigma_{eff}^2 \cdot n_{embed}} \cdot \frac{1-\varepsilon}{\varepsilon}$$

As a simple example, consider negative hadrons for central events binned in 100 MeV/c bins in p_T between 0 and 2 GeV/c and .1 in y between 0 and 1 (use STAR symmetry)$\rightarrow n_{bin} = 200$. Also, $\sigma_{eff} = .01$, and assume $n_{embed} = 100$ and $\varepsilon = .9$. Then $N_{event} = 2 \cdot 10^3$, not a large number.

How reasonable is $n_{embed} = 100$ for STAR? The surface area of the inner surface of the TPC is about $12.5m^2$. If the 100 embedded tracks are confined to the inner half of this surface (rough guess at phase space effects) 100 tracks are spread over an area of about $6.25m^2$, or one track per $625cm**2$. Since the two track resolution is of order 1 cm., these tracks likely do not interfere significantly and the number of embedded tracks per event can be somewhat larger without perturbing the event. This can be tested in practice by varying n_{embed} to check whether the calculated efficiencies are stable against this variation. Additionally, the kinematics (momentum and direction) of the embedded tracks can be chosen to minimize their interference.

B.2 Monte Carlo Estimates of Background

We consider here the need for GEANT–based calculations for estimates of physics backgrounds. Backgrounds are estimated by tracking full events from a "representative" event generator through a detailed GEANT model of the detector at the apporpriate level of detail of physics modelling. The phase space distributions of common particles (pions, kaons, protons, perhaps Λ) need to be reproduced by the event generator, but not the fine details of the event. This process is very CPU–intensive and not always appropriate for phsyics studies. For instance, if a mass peak is reconstructed, techniques such as interpolation from outside the peak, mixed events, or (for dileptons) like–sign pairs may give more reliable estimates. But often there is no substitute for a detailed Geant calculation.

We give here a rule of thumb for estimating the needed number of Geant-simulated events for estimation of background for a single particle spectrum analysis at STAR. Let:

N = # raw events in event sample
n_{tot} = population of bin with poorest S/B
n_{bkgd} = background counts in this bin
b = n_{bkgd}/n_{tot}
σ = σ_{bkgd}/n_{signal}

where σ is the required relative statistical uncertainty on background in this bin. The number of background events needed to achieve the relative error σ due to the statistical precision of the background estimate is:

$$N_{BkgdRequired} = \frac{N}{n_{tot}} \cdot \frac{b}{\sigma^2 \cdot (1-b)^2}$$

As a simple example, consider $n_{tot} = 100, b = 0.3$ (i.e. S/B=2:1) and $\sigma = 0.3$ (i.e. desire 30% error due to background estimate). Then $N_{BkgdRequired} = 0.07 \cdot N$, i.e. $N = 10^6$ requires the generation of 70K MC events. This simply expresses the fact that if the background is low, it doesn't need to be known very well. The number of MC events required may in fact be very large, depending upon the physics signal and the signal to background of the measurement.