STAR Offline Computing Requirements for the First Year of RHIC Running

Y. Fisyak (BNL), P. Jacobs (LBNL), T. Ullrich (Yale), T. Wenaus (BNL)

STAR Note 382

November 30, 1998

1 Introduction

This document should be seen as an addendum to the *STAR Offline Computing Requirements Report* from March 1998 [1]. While the former report focuses on estimates of computing needs for a fully instrumented STAR detector and nominal RHIC running, the aim of this document is to present estimates for the first year of RHIC operation.

First experiences from MDC1 show no significant deviations from the requirements listed in the first report. We conclude that the numbers derived in that document are still valid and that the requirements for nominal running conditions are up-to-date. In this report we therefore give only a brief description of the differences between nominal and first year running and refer the reader to the more detailed document for further information.

2 Assumption about the first year

In the following we describe the assumptions made in order to evaluate our computing requirements. While estimates on the instrumentation status of the STAR detector can be easily made, the performance of the RHIC machine, i.e., luminosity and availability and their time-dependence is less predictable.

2.1 STAR configuration

The STAR detector will include all baseline subsystems and in addition partially installed detector components where completion is scheduled several years after RHIC turn-on. The setup will consist of the baseline Time Projection Chamber (TPC) and trigger, Central Trigger Barrel (CTB), and 10% of the Electromagnetic Calorimeter (EMC). It is very likely that both Forward Time Projection Chambers (FTPC) and the recently proposed RICH detector will be added to the setup either at startup or during the first year.
The basic assumptions relevant for the data volume are:

Writing speed: 20 MB/s to tape

Event size: one central event will be 16 ± 4 MB

The event size is somewhat higher than the one reported previously because full data compression will not be enabled during the first year of operation.

2.2 RHIC running

The RHIC luminosity at startup is predicted to be around 1% and should increase exponentially to 10% towards the end of the first year. However, it should be noted that STAR's requirements do *not* depend on the actual luminosity profile but rather on the actual *duty factor profile*. While predictions of the first are fairly reliable it appears that assumptions about the latter are more uncertain.
Integrating the official estimated performance and folding with an estimated performance factor of STAR itself yields a data volume equivalent to $4 \cdot 10^6$ central Au+Au events at 100 GeVA/beam which represents approximately 25% of STARs data taken during a nominal year.

3 Raw Data Volume

The data volume recorded by STAR does not suffer from the small luminosity during startup. Because of the large event size, we will saturate our data acquisition rate even at lowest luminosities. Already above 2% of the nominal RHIC luminosity STAR is, in principle, able to saturate its rate with central events (5% of $\sigma_{\rm geom}$) alone. The overall data volume is therefore solely influenced by the event recording rate, the integrated performance of STAR and the availability of the RHIC machine.

The total raw data volume for nominal running was estimated to be 200 TB. Taking into account the somewhat larger event size and the smaller RHIC efficiency that we assume here, we expect a raw data volume of ~ 60 TB for the first year.

4 DST Production

In the STAR computing requirement report the CPU time per event was estimated to be 2.5 kSi95 $\cdot$ sec/event which translates to 70 kbyte/sec for typical Intel Pentium-II processors with 300–400 MHz (15 Si95). During MDC1 values of 50-100 kbyte/(sec CPU) were actually observed, which confirms nicely our previous evaluation. This also supports the usage of Si95 units which appear to allow a reasonable scaling for our purposes.
The total year sample of $4 \cdot 10^6$ events requires consequently $1 \cdot 10^7$ kSi95 $\cdot$ sec. We assume further that each raw event in year one will be processed 1.5 times[1], allowing for reprocessing of a fraction of the data which gives us the final number of $1.5 \cdot 10^7$ kSi95 $\cdot$ sec.
Given a combined STAR/RCF computing duty factor of 0.75, i.e. $2.37 \cdot 10^7$ sec/year, this number translates into 630 Si95 units for the year average. This however has to be seen as not sufficient since the coarse average does not well describe the actual RHIC performance profile. The machine duty factor is expected to improve with time, which will result in an unacceptable delay of months at the end of the first year. The CPU power that would be required to cope in real time with STAR's data taking rate at peak times during the first year is about a factor of 5–6 higher than the year average. In order to arrive at a reasonable number we simulated various optimistic and pessimistic RHIC performance profiles, and obtained numbers between 1.5 – 2 times the average value (~ 1200 Si95) to avoid backlogs of more than 10% of the data sample at the end of the year.
We estimate the DST data volume to be on the order of 10 TB which corresponds to a data reduction factor of 6 rather then 10 as for nominal running. This is mainly due to additional information on the DST which is required for initial systematic studies of the detector responses and the reconstruction algorithms.

5 Data mining and analysis

Because of the only partial instrumentation of the STAR detector during the first year several projects cannot be carried out. This applies to studies of lepton pairs,

[1] A factor of 1.5 is a particularly conservative number for year 1 running.

some fraction of the high-p$_\perp$ physics program and to the studies of rare hyperons. Nevertheless the majority of physics analysis programs can be performed with the installed baseline detectors and subsystems: studies of soft hadronic probes, hyperons (Λ, K_s^0), high-p$_\perp$ hadrons, two-particle correlations and event-by-event analysis. We derive the computing requirements by scaling down the relevant projects by the number of recorded events in year one as compared to nominal running conditions (factor of 1/4). From this we obtain an overall number for data mining and analysis of $1.2 \cdot 10^7$ kSi95 $\cdot$ sec for the total CPU needs and 2.3 TB for the μDST data volume. Assuming a STAR/RCF-CAS computing duty factor of 0.75 the first number translates to 480 Si95 units for the year average. Again, the same issues concerning peak times as described in section 4 apply. A significant fraction of the first months of STAR running will be used for quality assurance, detector calibration and alignment while the dominant load will occur at a later time during year one when larger DST samples are available and studies with calibrated detector systems can begin. The year average does not describe this profile at all. We estimate the need for at least twice the average value, i.e., 1000 Si95, to cope with the increased demand for CAS resources towards the end of year one.

6 Simulations

In STAR, simulations are used to derive corrections to the data due to effects of finite acceptance and efficiency, and background to a given signal due to the physics of the event or the instrumental response. As such they are essentially independent of the data sample size and thus we contemplate no change in the requirements as compared to nominal running, although we expect significant data mining and analysis of simulations to begin late in year one. The requirements derived in [1] are 10^8 kSi95 $\cdot$sec total annual CPU and an output volume of 24 TB. The annual CPU is the sum of $9 \cdot 10^7$ kSi95 $\cdot$sec for the generation of physical events including detector response simulations (Geant) and $1 \cdot 10^7$ kSi95 $\cdot$sec for full reconstruction. A large fraction of the former will be done off-site.

As mentioned in section 4 we see little need for CRS resources beyond the year average shortly after RHIC startup, while at peak times and towards the end of the year we anticipate the need for twice the annual average. Assuming the presence of resources equivalent to 1200 Si95 throughout the year the available overhead before the bulk of real data arrives can be used to reconstruct a considerable fraction of the simulated data. In the following table we therefore do not add reconstruction CPU for simulated data as a separate item but assume that the 1200 Si95 cover all reconstruction tasks during year one.

7 Summary

The following table summarizes the STAR offline computing requirements for year one as discussed in the foregoing text.

Assumption		Comment
Au+Au events recorded	$4 \cdot 10^6$	25% of nominal year
Event size	16 ± 4 MB	
Data recording rate	20 MB/s	
Combined RCF/STAR duty factor for reconstruction and analysis	0.75	
Number of reconstruction passes	1.5	conservative
Storage requirements		
Raw data volume	60 TB	
DST volume	10 TB	
μDST volume	2.3 TB	
Simulated data volume	24 TB	
CPU requirements		
Total for reconstruction	$1.5 \cdot 10^7$ kSi95 $\cdot$ sec	includes 1.5 passes
Average reconstruction CPU	630 Si95	flat RHIC duty cycle
Actual reconstruction CPU	1200 Si95	realistic duty cycle
Total for analysis	$1.2 \cdot 10^7$ kSi95 $\cdot$ sec	
Average analysis CPU	480 Si95	flat RHIC duty cycle
Actual analysis CPU	1000 Si95	realistic duty cycle
Simulations - event generation	$9 \cdot 10^7$ kSi95 $\cdot$ sec	not at RCF
Simulations - reconstruction	$1 \cdot 10^7$ kSi95 $\cdot$ sec	see section 6

References

[1] G. Eppley *et al.*, STAR Offline Computing Requirements Task Force Report, Star Note SN0327 (1998)