1 The Solenoidal Tracker at RHIC (STAR) experiment

1.1 Background

The **S**olenoidal **T**racker **A**t **R**HIC (STAR) Experiment at Brookhaven National Laboratory is one of the premier particle detectors in the world. While the first six years of the RHIC program have seen STAR very active and tackling several key fundamental research problems (jet quenching, evidence for the existence of a perfect liquid, number of constituent quark scaling, heavy quark suppression, Large transverse spin asymmetries in the sQGP regime and the possible indications of gluon saturation at small x), more exciting topics have unfolded in the past couple years: STAR has made the first ever observations of an anti-hyper-nucleus and anti-alpha, showed possible evidence for an azimuthal charged-particle correlations that may arise from local strong parity violation and leaded the field in producing the most precise constraints to date on the polarization of the gluons.

New area of research are constantly opening (property of the sQGP, existence of a phase diagram critical point, …) and an international collaboration composed of 55 institutions spread over 12 countries and a "team" of more than 560 physicists and skilled specialists is working hard to understand the nature of the early universe and the tiniest building blocks of matter through the study of nuclear collisions at the highest energies achieved in the laboratory.

The STAR institutions are geographically distributed as follow

Regional group	**N**	**%tage**	**2008 census**
USA / North America	22	40%	46%
Europe	16	29%	23%
Asia (China/Korea)	09	16%	15%
India	06	11%	12%
South America	02	04%	04%

Comparing the 2008 census (previous EsNet workshop, 52 institutions @ 590 collaborators), staffing remains near constant with a solid collaboration but the distribution, while remaining stable, indicates a shift toward more foreign institutions joining than US-based institutions. It is however not yet at a level influencing the network requirements as our foreign based colleagues typically uses central facilities in the US (see discussion in the next sections).

The STAR data production and analysis models have been mainly relying on centralized user analysis facilities and namely the *RHIC Computing Facility* (or RCF) located at BNL as STAR's Tier 0 center and the NERSC/PDSF center as Tier 1 center. STAR is also supplied with many STAR Analysis Centers (SACs, similar in scope to Tier 2 centers) and their inventory has been hard to assess but constitute pools of local resources dedicated (not necessarily shared with all STAR users) to local group physics program needs. Their numbers and size have been dramatically fluctuating (from up to 10 to a few

sites, from 15 nodes at 8 cores each to several 100 nodes at 16 cores each) but we noted in 2008 already that shall they be part of a global resource pool (bound by Grid infrastructure for example), the resources they represents would largely cover for the simulation needs of the collaboration.

Based on past trends and future perspective, we estimate the number of SACs to be as described in Table 1. In 2011, our active SACs centers remain Prague (Tier 2), Wayne State University and the MIT sites (Yale, Birmingham and Sao-Paulo phased out or slowed down over the past years and UIC never finalized due to non-stable local computing professional staffing).

Table 1: Projected number of stable STAR Analysis Centers (SAC)

	2010	2011	2012	2013	2014	2015	2016
Typical number of SAC	*4*	3	4	5	4	4	3

The STAR computing model continues to rely on a data-grid model whereas the processed data is made near immediately available to remote sites where computing resources are available. Data distributions tools have been consolidated by the addition of a global Replica-Catalog, able to make differential inventories between sites within minutes, and the development of in-house tools for reliable data transfer and redistribution.

1.2 Key Science Drivers

1.2.1 Instruments and Facilities – RCF

The RHIC Computing Facility located in New York at the Brookhaven National Laboratories (BNL) is the Tier 0 center for the STAR experiment. The STAR detector is located at BNL and the accumulated data is stored on mass storage (HPSS) at the facility.

BNL hosting all RHIC experiments and the core operation and role of the facility is to provide the core CPU computing cycles for a ½ of our user analysis needs, the whole of data reconstructions, support for data calibration, data reduction, database and some local need for simulations. The facility currently provides CPU powers of the order 6,900 kSi2k delivered by over 4,500 CPUs and is projected to reach CPU powers of 100,000 kSi2K by the beginning of the RHIC-II era (2015 onward). The total storage capacity has reached its projected limit at about 500 TB of central storage, served over NFS and usable for data production (and space reserved for dedicated tasks such as calibration, user analysis space, simulation and space for support of STAR's distributed computing program).

Under optimal conditions, the Data Acquisition system (DAQ) is capable of 500-600 MB/sec data streaming to disk cache located online. The reconstruction of our events do not change the size significantly as STAR had already planned to reduce its data demand by removing the need to save the event format files – since 2010, we save the totality of our Micro-DST files, a format ready for physics use and a factor of 5 smaller than our

event files, and 1/10th of our event files for physics verifications (this aggressive space saving was explained in "The STAR Computing Resource Plan, 2009" (CSN0474)). During Run 10 and Run 11, we showed that 600 MB/sec to mass storage was possible, the facility being largely able to absorb our data rate.

1.2.2 Process of Science – RCF

STAR's typical workflow consists of the origin of our Data, acquired by the Data AcQuisition system (DAQ) online at BNL. The DAQ is capable of a rate of 1 kHz at the moment, with event size spanning from ½ of MB to neat a MB depending on luminosity, energy and violence of the collision, colliding species and data acquisition triggers (condition to accept an event on tape). Typically, we assumed in 2011 that a 600 MB/sec store to a Mass Storage System (HPSS), directly from online event buffers, were possible and needed (we will explore the estimated requirements in section 1.2.2.1). Offline, a quality assurance process (a.k.a. Fast-Offline) pulls out of HPSS (while the data is still on HPSS cache) a fraction of its data and process it for quality control purposes on the RCF resources. Up to 15% of the data is pulled for inspection and the results of event processing is placed onto live storage (NFS mounted disk) for the quality assurance team and other detector sub-system experts to mine and verify its quality. Its lifetime is short (two weeks), old data is deleted and replaced by new one. The data is also use to forgo incrementally more precise calibration passes.

Data sets collected for a given year run are typically processed for final production at the end of the run (during the run, previous year data or large simulation requests are run at the Tier 0). When data productions are started, typically at the Tier 0 center at the RCF, two + 1 copies are handled by each job, submitted to a locally-engineered queue system resting on the principle of (mainly) one job equal one input DAQ file (a few output are created). Each production job places one copy of the output(s) in HPSS and one copy on central disk (NFS) and STAR's data management system verifies the presence of the HPSS copy, validates it (expected size checked, no MD5) and removes the NFS resident copy to make space for more files – the NFS storage acts as a "safety net" buffer only. The validated files are indexed (Catalogued in our Replica-Catalog) making it possible for other data management tools to take the files for grab. STAR/BNL's data management also places a copy of the physics ready files (a.k.a. Micro-DST or MuDST) in a virtual name space aggregator system known as Scalla/Xrootd. This system, initially maintained by STAR personnel, is now under the care of the RCF personnel (several groups make use of Scalla/Xrootd at the BNL/RCF). The replicas are also indexed in our Replica-Catalog and we will refer to this storage pool as *distributed storage*. BNL holds as much as 1.5 PB of distributed storage for STAR. Any missed files from distributed storage may be retrieved from mass storage by an inventory differential search. SACs and Tier 1 centers have relied on the BNL data management system to also make differential inventories of data they may need.

1.2.2.1 Projected Data rate and Data size

Since our network requirements are all derived by our program data demands, we attempted to build a data model projection based on past usage, past known long term planning ("The STAR Computing Resource Plan, 2009" (CSN0474)) and reassessed data

requirements as well as the state of our R&D. Two important factors will be treated separately and folded in our estimates: the RHIC luminosity increase and the STAR strawman Beam User Request (BUR).

1.2.2.1.1 Event size discussion – effect of luminosity

The assessment of the event size in outer years is driven by two factors: the addition of new detectors (we will discuss in the next section) in the STAR system and the RHIC luminosity increase, potentially causing or data acquisition to take more of pile-up events (events happening before/after the triggered event but recorded within the same time window as the speed of recording and the electronic would not allow separating them out). We recently studied the effect of pile-up in the RHIC-II era.

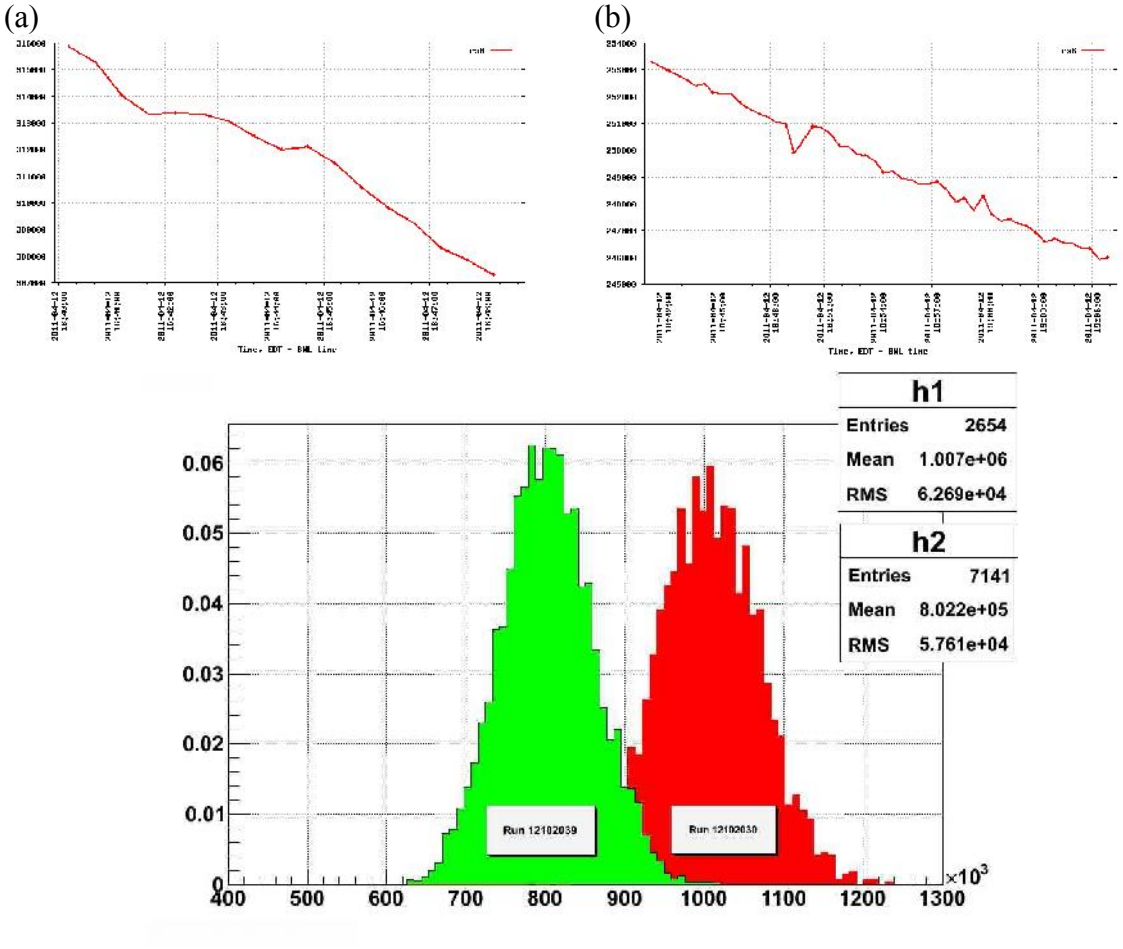

Figure 1: Distribution of event size for two run numbers, 12102030 (Tue Apr 12 18:40:46 EDT to 19:07:06 EDT) and run 12102039 (Tue Apr 12 16:39:44 EDT to 16:48:38 EDT). Panels (a) and (b) show the related ZDC rate proportional to the RHIC beam luminosity, both runs were taken under the same trigger conditions, the X axis is in Bytes. This figure shows a direct correlation between the average event size and the beam luminosity.

On Figure 1, we illustrate the correlation between the average even size and the beam luminosity – while Run 10 and Run 11 luminosities have allowed STAR to keep its event

size to a conservative ~ 0.6 MB/events both runs represented on Figure 1 have event sizes exceeding this average estimates – run 12102030 approaches a peak estimate of 1 MB/events. In Run 11, the lower left tail end of the distribution was selected via pile-up rejection methods but it is to be noted that we expect increase in peak luminosity to a 4 to 5x of those seen during the time run 12102030 was taken (3x average), implying a possible net event size up to 4-5 MB/evts (3 MB/evts average). Considering the RMS of the event size distribution, the probability of being able to select only the 1 MB/evts events during a p+p run is negligible without additional innovative event selection and mechanism suppressing the effect of event pile-up.

It is also noteworthy to mention that shall we successfully reach a point where High Level Trigger (HLT) algorithms, such as the Cellular Automaton seed finding (a multi-core, GPU or CPU, aware fast algorithm developed in-house with help from GSI/CBM colleagues), could be ported to online, there may be a chance to further reduce the size of our DAQ files by saving track seeds instead of hits or eliminating hits not used for tracking. This possibly ambitious path to data reduction would require a full physics evaluation pass as STAR carried for its online clustering algorithm to make sure Physics is not compromised. Potential size reductions are of the order of 2 for the TPC detector may then fold in our estimates (although it is likely the gain will be used for increasing the data samples in number of events, several physics analysis such as the di-leptons and Ds and any rare probes two particle analysis have their statistical precision directly tight to the number of available events).

In our calculations, we will fold factors of x1.41, x1.73 and x2 event size increase for p+p events in respectively 2013, 2014 and 2015 to account for the luminosity effect (and assuming we will cope for most of the event size increase). We will assume this increase does not affect the heavy ion runs (where pileup effects are minimal and tail selection will likely be possible). Outer years will remain at the same size increase estimates. Finally, we will also assume a 1:1 ratio of light versus heavy species, reducing the estimated factors to a rounded down x1.2, x1.4 and x1.5.

1.2.2.1.2 BUR & Projections up to 2015

Although a year by year run plan and BUR over the period requested is hard to predict with accuracy (a two years exercise is done and re-assessed every year at BNL through a formal process of case presentation to a Program Advisory Committee or PAC), Table 2 summarizes the state of our current knowledge.

The overall profile of 2012-2013 is driven in one hand by the proton program and the study of the gluon polarization at low X and the W boson program. The measurement is not believed to be luminosity limited but bandwidth limited. Each event frame may contain several collision or “pileup” as discussed in section 1.2.2.1.1, hence creating larger events, with a lower number of events at the highest p+p energy (considered in 2013 estimates).

In 2014, STAR is planning to take a large enough Au+Au sample so we can study the Heavy Flavor Tracker (a.k.a. HFT) detector sub-system response and event reconstruction (the detector is assumed to become physics usable in 2014 but a prototype will be installed in 2013). The physics goals would include high precision measurement

of the charm cross-section and possibly get a first look at charm flow for preparing for the high precision measurements of the later years. The 2014-2016 will be driven by the HFT program and the charm program as well as in 2015 a large p+p run for reference sample. The higher number (comparing to our past planning) in 2013 is due to the fact that our older plans assumed the low energy scan program would take place mostly in that year while our Run 10 & 11 BUR allowed for some of the beam energy scan to already take place (consistent with the suggestion we made in section 3.5.2 of CSN0474). Beyond 2016, STAR is likely to begin the onset toward the eSTAR program, adding detector in the forward region (this period is not part of this document).

Table 2: Projected data for the period 2012 to 2016. The two first years are showed as a cross check and trend projections purpose for outer years.

	Initial projections		*Outer years projections*				
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Projected N events (B)	0.85	2.4	2.20	2.50	2.00	2.00	2.00
Projected size RAW (TB)	550	1552.94	1321.05	1801.44	2125.78	2314.58	2314.58
Candidate for data production and transfer			*Trend projections (upper bound)*				
Final N events considered	1.5	2.6	2.38	2.75	2.2	2.2	2.2
Final size RAW (TB)	970.59	1682.35	1431.14	1981.58	2338.36	2546.04	2546.04
Deviation to projected	76.47%	8.33%	8.33%	10.00%	10.00%	5.00%	5.00%
Verification Catalog			*Projected based on possible excess*				
sum(events) tpx	1657150926	2660748136	2383333333	2750000000	2200000000	2100000000	2100000000
sum(size) tpx	861.99	1084.47	1332.86	1845.49	2177.76	2263.4	2263.4
Size / events (MB) – before luminosity effect	0.55	0.43	0.59	0.59	0.8	0.75	0.75
Size / events (MB)	0.55	0.43	0.59	0.70	1.04	1.13	1.13
2010 = real, 2011 = projected							*Projected*
Total events MuDST	1596593951	2563516884	2296239602	2649507232.8	2119605786	2023260069	2023260069
Fraction of events to RAW	96.35%	96.35%	96.35%	96.35%	96.35%	96.35%	96.35%
Total size MuDST (TB)	619.14	778.94	794.09	1099.51	1598.32	1630.15	1630.15
Size / events (MB)	0.41	0.32	0.36	0.44	0.79	0.84	0.84

For network bandwidth projections, we allowed uncertainties on the number of events as showed by the "Deviation to projected" row. A higher margin is set of 2013 and 2014 where/when the HFT detector system for STAR is believed to be integrated (currently target for full installation is October 2013, prototype begins in 2013 as hinted above. This fuzz-factor provides a safety margin for our projections. The size of the raw data is increased proportionally to the new detector phased in the STAR system (+100kByte for the FGT and +300kByte for the HFT) in respectively 2012 and 2014 while the following years considers a size decrease due to the phasing out of non-zero suppressed data in a few sub-systems (the FGT included in 2012 is considered to zero suppress as soon as 2013). For clarity, we separated the detector inclusion size effect and show the number in the "before luminosity effect" row of our Table 2. Similar trend affects the size of the derived data (MuDST) on the last line and historical data (trend) for accumulated data, usable data and data passing the Physics selection criterion to be saved in the final MuDST are considered.

1.2.2.2 LAN requirements, transfer from online to Mass Storage (HPSS)

From Table 2, we derive the LAN need as showed below in Table 3. A 20% margin was added to account for possible TCP protocol overheads and miscellaneous transfer problems causing lags and summarized in the first row as gross average.

Table 3: LAN need at BNL to sustain the projected data rates.

	LAN need (no consideration of species granularity)						
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
LAN, DAQ to HPSS gross average [+20%] – Minimal (MB/sec)	*196.03*	*339.78*	289.05	400.22	472.28	514.22	514.22
<Peak> DAQ → HPSS LAN [+20%] (MB/sec)	*139.65*	*394.32*	335.44	457.41	539.77	550.00	550.00
All times LAN rate needed (MB/sec)	*139.65*	*394.32*	394.32	457.41	539.77	550.00	550.00
LAN (Gb/sec)	1.09	3.08	3.08	3.57	4.22	4.30	4.30

However, those rates representing averages are based on the total projected data rates (in size) and the associated operational hours recorded by our online accounting tool (RunLog) for the whole run. The <Peak> numbers on the second row are actually average rates over the time we took good data rather than an all-time upper limit.

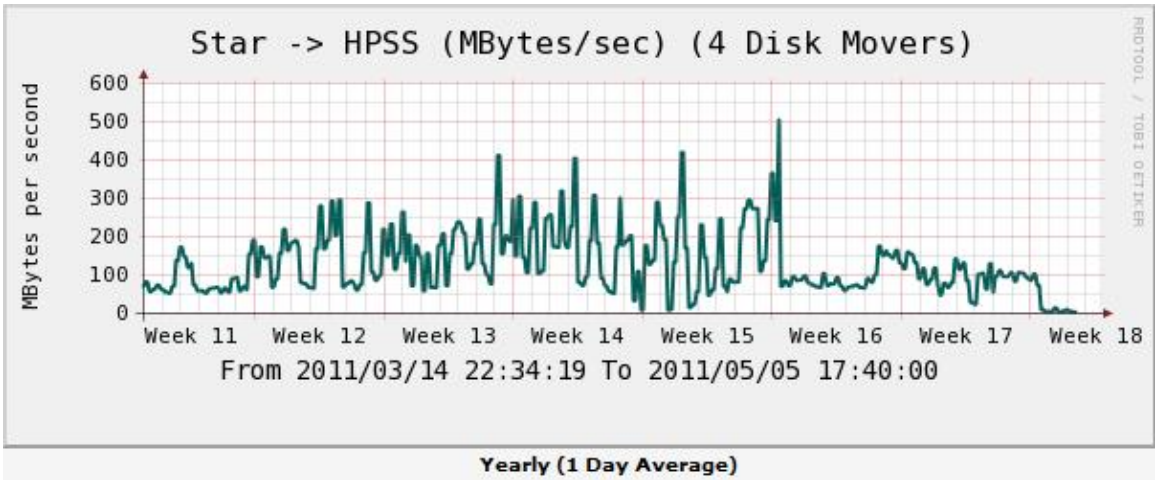

Figure 2: Data mover statistics from STAR online to Mass Storage (HPSS) at the RCF. Note the peaks at 500 MB/sec. Over the run period, we had averages of 250 MB/sec, consistent with our estimate (modulo the 20% overhead) of ~ 300 MB/sec of Table 3.

While this estimate tends to stand the usage verification for the past two years, previous estimates considering fluctuations across species, run weeks and run efficiencies showed needed rates as high as 550 MB/sec needed at real peak times. Modulo appropriate use of online disk caching, an ultimate 550 MB/sec would account for "surge" of data rates and showed to be feasible. For 2011, an example of DAQ rates is showed on Figure 2 - real peaks are at ~ 500 MB/sec. The last row of Table 3 shows the LAN line rate in Gb/sec.

1.2.2.3 WAN requirements, transfers of Micro-DST to other Tier

Table 4 summarizes the network needs for transferring the physics ready Micro-DST to other Tier centers. While Tier 1 should have a full copy of the MuDST which could be moved over large period of times (as produced), SACs tend to transfer data a-posteriori, by bulk and depending on local analysis demands. When they are in need of a dataset, their expectations are typically a fast turnaround for data movement (within days) and a "*we will take all you give*" approach to network bandwidth. Hence, we assumed full dataset transfer over a period of 3 months for a Tier 1 center and a one week delay for a SAC or a Tier 2 (the demand was typical of the observed pattern for our Prague site; dataset rotation i.e. the replacement of the local data by a brand new set, would happen on the order of 4 times a year).

Table 4: Data transfer rates for sustaining redistribution of Micro-DSTs to SACs or Tier centers.

	WAN needed for MuDST @ SACs & Tier X						
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
Typical number of SACs (STAR Analysis Centers including non-US Tier 2)	*4*	3	4	5	4	4	3
Tier 1 center [100%, 3 months] (Gb/sec)	*0.65*	0.82	0.84	1.16	1.68	1.72	1.72
Individual SAC/Tier 2 bdwdth need [rotation at 10% datasets, week] (Gb/sec)	*0.84*	1.06	1.08	1.49	2.16	2.21	2.21
Total SACs bdwdth out of BNL [assume 2/3, 1/3] (Gb/sec)	*2.24*	2.11	2.87	4.96	5.77	5.89	4.42
Total SACs bdwdth out of NERSC [assumes 1/3, 2/3] (Gb/sec)	*1.12*	1.06	1.43	2.48	2.89	2.94	2.21

Finally, for total network estimates from NERSC and/or BNL, we assume 2/3rd of our institutions would acquire the data from BNL while 1/3rd would do this from NERSC. This is a target goal (but not representative of today's habits – near all institutions in need of data take it from the BNL/RCF's mass storage as sole "trusted and complete" source for datasets).

1.2.2.4 WAN requirements, Monte-Carlo simulations

While extremely CPU intensive (full simulation or "slow" simulator can take as long as 30-45 mnts on modern CPU for a Au+Au 200 GeV collision event), our Monte-Carlo production on the grid does not generate significant bandwidth requirements. A typical 24 hours process would generate of the order on 100 GB output, a small perturbation to the overall requirements. Concurrency of jobs and stability of transfer services are however key to Grid usability – STAR has made heavy use of the data transfer capabilities of SRMs to de-couple CPU slot usage and output data transfers back to the BNL/RCF.

1.2.2.5 WAN requirements, embedding support

STAR's only stable Tier 1 center to date has been the NERSC/PDSF resource facility. STAR principle uses for PDSF are to sustain some of its user's analysis, access the resources from BNL via Grid for Monte-Carlo simulation productions and use a large

fraction (~ ½) of PDSF's STAR allocated resources for the embedding simulation production. We summarize WAN requirements in Table 5 with the justifications as follows.

Table 5: Network requirements in Gb/sec for sustaining the embedding process at one Tier 1 center.

	WAN speed for embedding support at Tier 1						
	2010	2011	2012	2013	2014	2015	2016
%tage N needed [only 10% of the data can be used for embedding]	10.00%	10.00%	10.00%	15.00%	15.00%	12.00%	10.00%
Minimal WAN for raw offsite [embedding] (Gb/sec)	0.01	0.03	0.03	0.05	0.06	0.05	0.04
Size raw to move TB	8.62	10.84	13.33	18.45	21.78	22.63	22.63
WAN desired, raw offsite within 2 days (Gb/sec)	0.41	0.51	0.63	0.87	1.03	1.07	1.07
Data size produced by embedding (TB)	60.34	75.91	93.30	129.18	152.44	158.44	158.44
WAN needed for medium priority move of the data back to Tier0 [within 1 week] (Gb/sec)	0.82	1.03	1.26	1.75	2.06	2.15	2.15
Total needed for embedding [ideal] (Gb/sec)	1.23	1.54	1.90	2.62	3.10	3.22	3.22

While 10% to 15% of our data was needed in past planning to sustain the embedding simulations, the increase of the data sets in size has not caused a proportional increase of the need to transfer raw DAQ files to NERSC. Only 10% (or less) of our data contains information necessary for handling this kind of production: those files contain "raw" signals while the rest of our data contains already formed track hits or clusters (online clustering is performed to reduce the DAQ output size). On the first line of Table 5, we show an estimate of the percentage of data needed for our embedding operations: higher percentages in 2013 to 2015 accounts for the introduction of new detector sub-systems which may require enhanced simulations to understand them fully.

Embedding productions require long preparation and typically, the target goal is for the samples to be transferred within a short time period – we assumed samples need to be transferred within two days. However, the results need (in principle) to be brought back to the BNL/RCF (low priority, within a week) driving an additional network requirement on transferring the results from NERSC to BNL. This has not been done consistently to date but we expect a change in the coming year (the storage allocation at the BNL/RCF accounts for this transfer). A note that the output tends to be larger by a factor of 7 than the input as many files are generated: event files, Micro-DST files and Geant association files are all needed for efficiency corrections.

Finally, while Table 5 tends to suggest a one-time embedding transfer process, several Run years' worth of data is handled simultaneously (and often, the datasets needed do not overlap with previously transferred DAQ files). Instead of making a fine grain estimate over all embedding series within a year, averaging the total bandwidth over the course of the year, we chose to use one number representative of a burst transfer operation.

1.2.2.6 WAN requirements, Cloud and data preservation scenari

As part of the EsNet workshop, we would like to present and entertain the idea of the possibility to process on a National Laboratories Cloud facility up to 20% of our data.

Such percentage would allow out-sourcing the cycles needed for our Fast-Offline QA process, quoted as of the order of 15% of our data in section 1.2.2.1 Table 2, or an equivalent "emergency" production of data in a fast turn-around manner. While not yet considered as a regular workflow (Cloud facilities not being guaranteed), such operation would dramatically enhance our Physics capabilities and possibly allow a better use of remote facilities which otherwise, would not be accessible to STAR. For example, none of our software was installed at ANL but we could deliver a stable production stream on ANL's Magellan Cloud within a virtual machine facility.

Table 6 summarizes the additional bandwidth requirements for this idea to become concrete and feasible at levels of 20% (Fast-Offline) and 50% of the data.

Table 6: WAN requirements for handling a Cloud operation at a level of 20% of our data.

	WAN needs, N% processed offsite						
	2010	2011	2012	2013	2014	2015	2016
WAN need for 20% RAW moved offsite [Cloud] (Gb/sec)	N/A	0.62	0.52	0.71	0.84	0.86	0.86
WAN need for 20% MuDST back to BNL [Cloud]	N/A	0.26	0.26	0.36	0.53	0.54	0.54
Total WAN for 20% offsite processing [Cloud] model (Gb/sec)	N/A	0.87	0.79	1.08	1.37	1.4	1.4
Total WAN for 50% offsite processing [1/2 pass "as we go"] (Gb/sec)	N/A	2.18	1.97	2.7	3.43	3.5	3.5
Total WAN a one time copy of all raw offsite (Gb/sec)	N/A	3.08	2.62	3.57	4.22	4.3	4.3

It is also worth noting that the STAR collaboration is exploring the possibility to leverage Hadoop file system and the Google Map-Reduce paradigm for data processing on distributed resources. While at its infancy, if such exploratory work would reveal a path to better exploit and tight storage and computational resources, further Cloud-like approach may appear for the sake of efficient use of resources.

Within the same idea of exploring additional network paths, we do not consider at the moment the full transfer of all our raw datasets to a remote storage facility. This precluded data safety at the RCF and STAR is subject to raw data loss as tapes decay (frequent access for reprocessing the data tends to wear them out). Furthermore, the incoming of new mass storage technologies, such as the HPSS T10k-C cartridges (5 TB at first generation up to possibly projected 40 TB storage per cartridges), will put STAR at risk of losing all early year's data or a large fraction of recent targeted datasets (a low energy point sample for example with the loss of a single cartridge. From this observation, two scenari offers itself as natural solutions: (a) double the storage at the BNL/RCF center (replicate each tape to another, HPSS allows dual copy) or (b) move an entire dataset of raw files to an alternate facility. The former would be immediately possible and under full control of BNL/RCF and RHIC/STAR operations while the second would provide an additional geographical data safety (two disconnected centers are unlikely to accidently lose data at the same time). The combination of both model are not orthogonal (geographical safety and local dual copies could be both done, further enhancing DOE's ability to safely preserve invaluable data for the long term). Modulo the logistic of economics and the understanding of how to securely provide long term

archival capacity at NERSC, the last row of Table 6 shows the bandwidth needed to achieve this plan (this would allow streaming data from online to offsite over the run period).

1.2.2.7 WAN requirements, summary from a BNL/Tier 0 perspective

Table 7: Requirements totals – greyed cells indicates passed or unlikely achievable scenari.

	WAN totals, by Tier						
	2010	2011	2012	2013	2014	2015	2016
SACs and Tier 2 centers (need for any / each)	0.84	1.06	1.08	1.49	2.16	2.21	2.21
Tier 1 center, MuDST and embedding support	1.23	1.54	1.90	2.62	3.10	3.22	3.22
[A] Tier 0 center, general support (Gb/sec)	2.24	2.11	2.87	4.96	5.77	5.89	4.42
[B] Tier 0 center, general support + 1/2 pass offsite (Gb/sec)	2.24	2.18	2.87	4.96	5.77	5.89	4.42
[C] Tier 0 center, general support (Gb/sec) + 1/2 pass offsite + complementary 1/2 saving at Tier 1	2.24	2.18	3.52	5.86	6.83	6.96	5.49

Table 7 shows the total requirements considering all the factors previously explained and roughly decomposed by Tier center levels. The three last rows are Tier 0 centric and consider respectively [A] the basic network needs for a "standard' STAR workflow [B] the same adding a 50% of one pass data production level on a cloud based operation and [C] a similar workflow as in the previous line, adding as traffic to accommodate for another half of our data to transferred to a Tier 1. A note that summing the numbers in a linear manner would not be adequate (peak requirements represents only a worst-case scenario). Instead, we set the required bandwidth as the maximum bandwidth for all the data from previous numbers except for the last row where a max is made but the bandwidth necessary for transferring ½ of our data over a period of time twice of the length of a run (lower priority transfer) is added linearly.

The assumption behind scenario [C] is that if we already produce ½ of our data to a Cloud based operation located at (for example) at NERSC, then we could store the data on mass storage as part of the same workflow and only have the other 50% to be transferred to achieve full dataset safety and preservation (but without processing). We do not show the requirements this would impose on the Tier 1 center (it would follow a similar arithmetic guided by our numbers from Table 6 and Table 7). Note as well that STAR have modeled its processing needs based on a minimum of 2.2 to 2.4 passes per year of data – this estimate may have been far too conservative as precision Physics may require more iteration – the scenario above only represents ~ +0.5 pass additional for science convergence (coupled to the prospect of a full dataset saving at a remote facility).

The maximum requirement in all scenari are (rounded up) a 7 Gb/sec for BNL/RCF connectivity to the world, a 4 Gb/sec for NERSC/PDSF (5 Gb/sec in data preservation, scenario [C] mode) and a maximum of 2.5 Gb/sec per SAC.

1.2.3 Instruments and Facilities – NERSC/PDSF

The NERSC facility serves as a major computational facility for the RHIC/STAR experiment, providing resources to local researchers as well as collaborators nationally and internationally. While the LHC/Alice experiment usage is ramping up, as seen on STAR remains the top user at NERSC/PDSF.

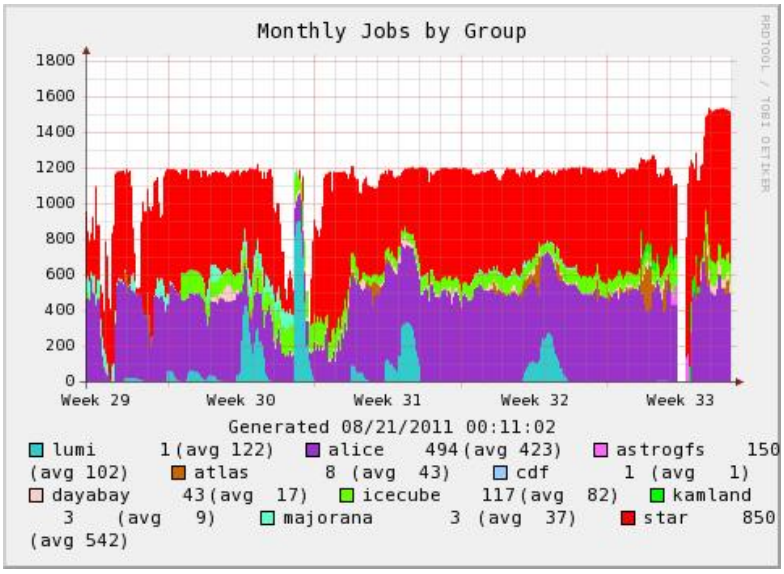

Figure 3: Batch queue usage at NERSC/PDSF by group. STAR remains the largest user and share near equal resources with LHC/Alice.

The main physics thrust of the program is the study of matter under the most extreme conditions of energy density available in the laboratory caused by the collisions of atomic nuclei at relativistic energies. The local group at Berkeley remains active is the support of the STAR central detector, the STAR TPC, and is also heavily involved in the detector upgrade program of STAR, with a focus on the Heavy Flavor Tracker (HFT). The presence of data close to the local research group makes the STAR Berkeley group and vibrant research collaborative institutions.

The Tier 1 Currently supports the deployment of the STAR software library releases and manages bulk data transfer to/from BNL. The single framework STAR provides through its library releases then allows PDSF users to perform data analysis and dedicated STAR workforce to handle the embedding simulation production. The local support team also maintains the Grid infrastructure further allowing the BNL production team to handle remotely steered Grid based Monte-Carlo productions but most of the workload is done via jobs locally submitted to batch system with minimal data transfer on WAN.

WAN data transfer is carried out in bulk managed fashion (using the Berkeley DataMover, SRM with gridftp, gridftp or alternatively, the use in recent times of the FDT[1] tool) with local catalogs showing what datasets are available for local analysis. The

[1] Fast Data Tansfer, FDT: http://monalisa.cern.ch/FDT/

Replica Catalog is now global, allowing any site to query and inventory the full set of replicas.

Data analysis if entirely based on locally available datasets at this point, but it has been discussed to leverage the deployment of remotely accessible Scalla/Xrootd service to share data from the NERSC/PDSF and the BNL/RCF.

1.2.4 Process of Science – NERSC/PDSF

Workflow for embedding simulations, typically carried at the NERSC/PDSF, are complex simulations aimed to combine simulated tracks embedded into raw data signals (serving as a background) – our code ability to reconstruct of the embedded tracks and identify them as close to the original track characteristics is directly link to detector efficiencies (geometrical acceptance, functional coverage i.e. estimate of the effect of "dead" zones), code and data reconstruction artifacts (algorithmic efficiencies) and biases on momentum or even particle identifications. The efficiency corrections allow comparing the data to models, data to results from other experiments also correcting for efficiencies or quoting uncertainties on our physics results. This process requires apportion of DAQ data to be transferred from the BNL/RCF to the NERSC/PDSF and a copy of the resulting outputs supposed to be brought back to the BNL/RCF. The bandwidth requirements for those transfers were explained and presented in section 1.2.2.5. The numbers will drive our network need estimates in the next sections.

1.2.5 Remote science driver – NERSC/PDSF

0-2 years case

The operations at NERSC should remain standard, a balance between local user analysis, embedding productions and Grid based Monte-Carlo. We do not anticipate major changes apart from the possible use of SCalla/Xrootd global redirector, the impact of which for WAN requirements which will need to be studied and understood (we have no practical experience as per how much data may be pulled from one site to the other front at this time). However, user analysis varies from 100 Hz events data consumption to a second event reading and from Table 2, event sizes spanning from 0.36 MB to 0.61 MB lead us to conclude that only a Hybrid model (not a fully shared data exchange scheme) may be possible. In other words, even one event per second at 0.36 MB and 2000 slots at BNL reading data from remote would imply a 7 Gb/sec transfer if no data would exists at BNL – this is not envisioned within our bandwidth request. Within a hybrid model, Scalla/Xrootd may transfer missing data between the two sites via gateways, using whatever bandwidth is available to synchronize the data pools.

Within this two years period, we expect our BNL/RCF Tier 0 network requirements to follow a standard "general support" requirement (scenario [A] from Table 7) at a maximum of ~ 3 Gb/sec WAN bandwidth needs while NERSC/PDSF Tier 1 center will require of the order of 2 Gb/sec connectivity to BNL for sustaining STAR science.

2-5 years case

Large data samples, driven by precision physics topics (with key players at our Berkeley and international colleagues facilities) and the possibility of produce data fast and use all resources will likely force the STAR collaboration to offload some of its processing to remote sites. Depending on resource availabilities, we envision that our "Cloud based" offsite processing scenario [B] (technically possible today and already demonstrated by STAR) would be a path to follow shall opportunistic resources be made available. By then, the lifetime of the PDSF facility will also be questionable (economy of scale) making a virtualized operation even more so likely. Assuming this would be the direction of NERSC (leverage larger more economic clusters and normalize smaller operation through support of their science via Virtualization), the WAN requirements would follow the guidance indicated by the numbers given scenario [B] in Table 7. Those numbers remain at ~ 5-6 Gb/sec maximum for BNL/RCF connection to NERSC/PDSF at ~ 3-4 Gb/sec.

By then, and if this scenario is possible, it is likely that the currently run OSG-based simulation productions may be reshaped to fit within a Cloud-based or Virtualized infrastructure (managed or not by the OSG, depending on NP office's interests).

5+ years scenario

In the long term, we view the copy and preservation of our past data to another center as vital for ensuring data safety and longevity for DOE's scientific data and as STAR will be morphing into eSTAR and possibly, BNL phasing into eRHIC. We view the NERSC/PDSF mass storage as a natural place to place another full integral copy of our data.

6-7 Gb/sec transfers will then be minimally needed on the BNL/RCF side while NERSC/PDSF would require 4.5 Gb/sec (not represented in our summary table).

1.2.6 Instruments and Facilities – Prague / SAC or Tier 2 case

The Ultra-Relativistic Heavy Ion Group of the Nuclear Physics Institute ASCR has been active STAR participant since 2001. From the yearly times they have been pursuing a path of local computing as the most efficient way of data processing and physics analysis. The group has been involved in computationally intensive correlation analysis (HBT) and detector simulations (SVT and now Heavy Flavor Tracker or HFT a key upgrade project for STAR). Realizing that the efficiency of the offline analysis is dependent on available computing power, storage elements and dedicated human resources, the group has heavily invested in all of these areas. At present time, ASCR has dedicated computer scientists to take care of a local farm allowing 25 TB of storage space.

1.2.7 Process of Science – Prague / SAC or Tier 2

The creation of local opportunities for scientific analysis (without the need for remote connection to BNL) was projected to attract more scientists and in fact, a new group joined STAR from Prague (now two institutions and a pool of 20 scientists). The local data processing capacity has been so far limited mostly by the ability to transfer the data sets for analyses from the BNL/RCF to the local storage and vice versa. A breakthrough came in 2008 with the creation of a dedicated routed line. Initially at 1 Gb/sec dedicated, this line was dropped to about 140 Mb/sec throttled as illustrated in Figure 4. The change of network bandwidth has made the group adapt to the new reality and shift their investment to purchase storage at the BNL/RCF, the dedicated routing still allowing decent remote work. The storage local to Prague/Bulovka is then used as backup for analysis results, code, macros, publication material and the group ensures both data safety and resilience (a complete collapse of networking would allow them to continue to work locally). Proactive feedback and survey made to understand their need and shift of scheme emphasized that the connections latency is the show stopper and the tipping point for remote centers.

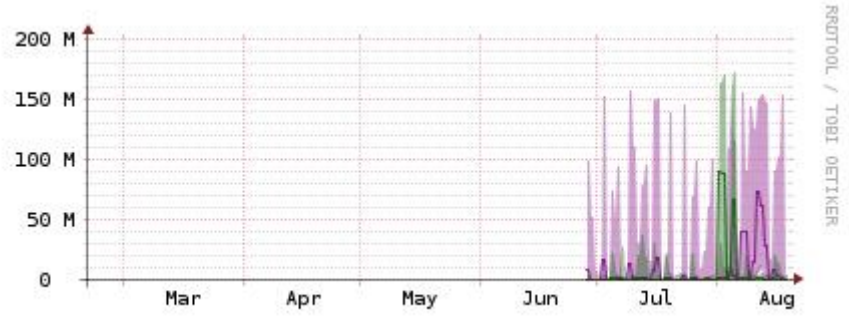

Figure 4: Network data transfer to Prague from BNL. The saturation at ~ 150 Mb/sec is due to a bandwidth throttling which was made in 2009.

Typically, a SAC or a Tier 2 center would transfer data sets of interest to the extent useful to their local research efforts. At Prague, Data transfers are handled using FDT (Fast Data Transfer) a highly portable java-based client optimizing network and local disk end-to-end disk capacity (local IO).

Prague has also been heavily involved in the development of theoretical computing models (based on constraint programing or mixed integer programing) and the development of data planers to enhance data transfers and leverage the presence of datasets from multiple sources (data sources as well as sites) for the most efficient data transfers to a destination. We would also like to note such new approach may change dramatically the bandwidth requirements. Preliminary studies in STAR of the use of such data planers (relying on existing data movers but knowledgeable or reacting to network capacities, links and local storage availability) showed a 30% makespan improvement for data transfers over a direct one network path data transfer from BNL to Prague. Our tests used data movers at NESRC and all relied on FDT to move data across sites. As

illustrated on Figure 5, the planer was able to leverage the data cache at the PDSF to move data "faster" to our center in Prague, allowing maximal use of all network links.

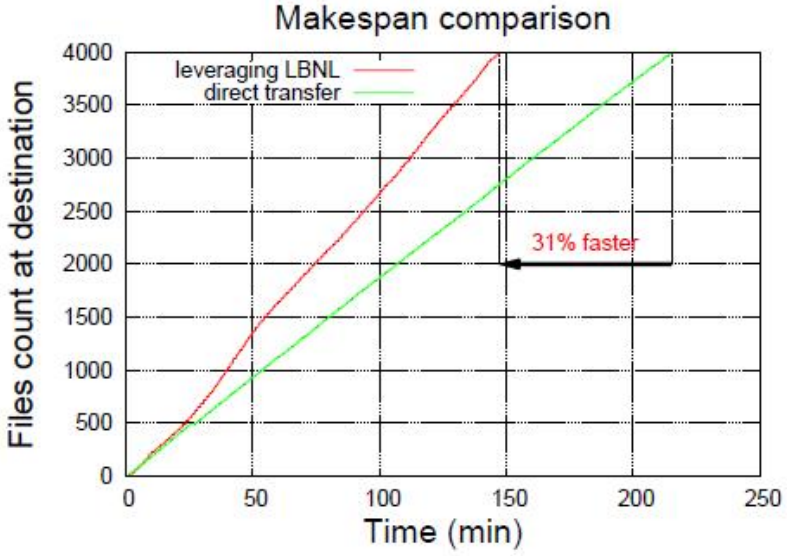

Figure 5: Moving the same number of files from a selected dataset, makespan comparison between a direct site-site transfer and one leveraging two sites with independent network path to the destination.

This methodology certainly holds non trivial consequence for network bandwidth usage. With it dynamic discovery feature of available path to a destination and coupled with features such as advanced network bandwidth reservation and/or predictor, a resulting comprehensive data transfer and management solution may very well allow saturating unused or low-used network segment experiments do not otherwise discover.

1.2.8 Remote science driver – Prague / SAC or Tier 2

0-2 years case

In section 1 Table 1 we gave a projection of a possible number of SACs. While this number is hard to estimate in STAR (sites come and go, some do not declare their presence and do not fully integrate to the data management system of STAR), we believe the profile will remain standard with no surprises.

1.5 Gb/sec connectivity to/from SACs seems sufficient for sites in need to move data closer and make use of their local CPU resources.

2-5 years case

We feel the SACs and Tier 2 centers, with "immediate" (low amount of data but in need of short time spans to acquire them) may drive data demand to a level beyond our ~ 2.5 Gb/sec estimate. This number was given in our summary but funding and science profile in the US may create conditions for short window of opportunity for scientists to harvest the science of the RHIC-II era – this would ultimately pressure remote sites to stress the

Tier 0 to the extreme and can drive bandwidth demand from those centers up to twice the projected needs. However, empowered with tools such as efficient data planners, it is likely that no change or increase in the infrastructure would be needed (if there remain by then such a thing as an unused or low used network path).

5+ years case

The requirement will remain stable at a 2.5 Gb/sec link speed maximum.

1.3 Middleware Tools and Services

The STAR collaboration currently makes extensive use of services such as teleconferencing and Web publishing with a net increase of IP-based teleconferencing and mostly, EVO and Skype-client based communications (free, easy and versatile). We note that EVO have had its share of drop connections and barely audible remote speaker perhaps related to the use of common network path, heavily used for sustaining science based on data movements. The extent to which the EsNet collaboration services are useful depends a lot on what happens in the commercial/free services world (such as Skype). A service which integrates with the grid authorization services would though be useful as STAR collaborators already register as members of the VO and could use the collaboration services without additional registration steps.

Use of Grid tools may remain strong as far as our distributed computing program remains sustainable. Data transfers are handled using the Berkeley DataMover, SRM with gridftp, native gridftp or alternatively, the use in recent times of the FDT[2] tool. STAR has developed data planners in house and deployed it over a few sites in test mode (see section 1.2.7 for a quick description) but this tool essential relies on FDT to actually move the data.

1.4 Issues

- We remain convinced that the distribution of our Physics ready data allows for an enhanced productivity where they become available. The effect is often geographical – for example, users from institutions "close to" NERSC/PDSF would typically use that Tier facility for their user analysis (the connectivity and latencies providing the most convenient environment).

- While our process of transfer to NERSC/PDSF was aimed to be fully automated, the need to verify the validity of data productions over larger samples and our current inability to invalidate datasets placed at remote sites have caused delays in data redistributions. They are typically not done synchronously to data

[2] Fast Data Tansfer, FDT: http://monalisa.cern.ch/FDT/

productions but after a physics evaluation period (which may take months). As a result, higher bandwidth are needed over short periods of times and, considering workforce at remote sites, we do not see this modus operandi as changing in the coming years.

- The move away from STAR of the Birmingham institution has to some extent slowed down our past plan to expand our embedding operation and outreach to other Tier 1 centers. However, STAR can now balance (in emergency situations) the workload between BNL and NERSC for this kind of operation. Potential new coming SACs from the US include our institutions in Texas.

- STAR remains a member of the Open Science Grid (OSG) and as such, continues to make use of its resources for Monte-Carlo simulation. Near all STAR direct Mote-Carlo simulations (not requiring real data as input that is, unlike embedding), is carried on Grid resources (emergency running may be carried at BNL). Though, only STAR's already dedicated sites (those for which Nuclear Physics fund a base resource and hardware for STAR) are used as the heterogeneity of Grids has not made running complex workflow practically unachievable in our views – pre-installed software packages takes care and attention for a full reproducibility and perfection of science (full validations can take several days and troubleshooting remain difficult on Grid). More than ever, we are fully confident that the use of Virtualization capabilities on Grids would allow STAR to have access to much more opportunistic resources.

- STAR has recently made massive use of the Magellan Cloud facility for raw data processing[3],[4]. The workflow included the transfer from our Fast-Offline facility of a fraction of the raw DAQ files to the cloud, leveraging a 3 TB, 20 TB and near no space of storage buffers at respectively BNL, NERSC Cloud facility and ANL Cloud facility. STAR made a quick "preview" production pass of the totality of its "W" boson candidate data as well as a pass of the Beam Energy Scan data. The latest allowed presenting results necessary to make a case for an additional lower energy point as part of the Run 11 cycle and this, during a time when all the STAR CPU resources were allocated to satisfy the demand for the Quark Mater 2011 conference. This truly opportunistic mode of operation showed that (a) STAR is able to and equipped to run today the most complex workflow on distributed (virtualized resources) and (b) the use of burst resources (availability of elastic resource) remain fundamental to the ability of an experiment to treat, under heavy load and demand, physics cases which would otherwise be dropped. The consequence on Networking is however non-trivial: set at National Laboratories, Magellan cloud has predictable network path; true commercial clouds do not (perhaps suggesting a strong case for continuing to sustain National Laboratories Cloud base infrastructures).

[3] Magellan Tackles Mysterious Proton Spin, NERSC Science News
[4] The case of the missing proton spin, Science Grid This Week, June 2011

Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• RHIC/STAR at BNL taking data and standard data production support and distribution to Tier centers	• Data taking	• 1.4-1.9 PByte/year	• 400-500 MB/sec - peak at 500 MB/sec	
	• Physics ready Micro-DST (MuDST) transfer to NERSC/PDSF	• 800-1000 TB, 400-500 k files		• 1-1.5 Gps Transfer to NERSC/PDSF over 3 months
	• Partial delivery of MuDST to ~ 3-4 SACS & Tier 2 centers	• 80-100 TB or less in burst		• SAC support @ 1.5 Gbps, data moved within a week
	• OSG use for simulations			
	• Embedding simulation support at NERSC/PDSF	• 10 k files, ~15-20 TB and of the order of 100k files and 90-130 TB results back to BNL		• ~ 2 Gbps in/out of NERSC and BNL
	• Estimated totals			• NERSC 2 Gbps and BNL 3 Gbps
2-5 years				
• RHIC/STAR data taking, Heavy Flavor program, local and distributed data production	• Data taking	• 2.0-2.5 PBytes/year	• 550 MB/sec from online to RCF	
	• Distributed infrastructure based simulations and Fast-Offline (50%) – Cloud-like	• ~ 1.0-1.2 PB during runs		• 3-3.5 Gbps for streaming data from online to remote site ("live")
	• MuDST copy at NERSC/PDSF	• 1-1.6 PB, 400-500 k files		• 2 Gps Transfer to NERSC/PDSF over 3 months
	• MuDST delivery to 3-	• 100-160 TB		• SAC support @ 2.5 Gbps,

	4 SACS & Tier 2 centers • Embedding simulation support at NERSC/PDSF • Estimated totals	or less in burst • 20-22 TB input and ~ 160 TB output		data moved within a week • 2-3 Gbps in/out of NERSC and BNL • NERSC 3-4 Gbps and BNL 5-6 Gbps
5+ years				
• End of RHIC-II era? STAR moving to eSTAR	• Same type of operations as mid-range • Transfer of data set off-site for permanent redundant archival storage • Estimated totals	• Similar datasets • Data size ~ 1.6 PB to move + back years	• Similar rates	• Assume additional bandwidth of 1 Gbps for 50% transfer + use of existing build infrastructure • NERSC 4.5 Gbps and BNL at 6-7 Gbps