1 The Solenoidal Tracker at RHIC (STAR) experiment

1.1 The Physics case of STAR and evolution toward eSTAR

The Solenoidal Tracker At RHIC (STAR) Experiment is one of the two large Nuclear Physics (NP) US based experiments at the Relativistic Heavy Ion Collider (RHIC). Located at the Brookhaven National Laboratory (BNL) in New York, Long Island, the facility has been one of the greatest successes of the U.S. Nuclear Physics research program and the first to observe convincing evidence of a new state of quark-gluon matter and in addition, is the world's only polarized proton collider. RHIC has been extremely productive in delivering and accomplishing its scientific mission and the first decade of physics deliverables produced in STAR alone 165 new PhD students, 145 refereed papers (151 cited) with near 16,000 citations.

The most important discovery made in this area over the past decade is that the QGP acts as a strongly interacting system with unique and previously unexpected properties (sQGP). While early expectations and predictions from the community foresaw a QGP behaving like an ideal gas, the matter produced in near-central RHIC collisions was shown to flow as a nearly viscosity-free fluid (a.k.a. "perfect liquid"). Further, yields and flow of mesons compared to those of baryons have establish a scaling behavior that points to collective flow established at the quark level, with hadrons subsequently formed by coalescence of already flowing quarks. Through its unique and versatile polarized proton beam, the RHIC spin program has made great strides towards unraveling the decades old question about the partonic origin of the proton's spin. Longitudinally polarized proton collisions are currently the world's best source of information about the gluon helicity distribution, with the most recent measurements indicating gluons may contribute as much as quarks (~20-30%) to the total spin of the proton. Collisions of 250 GeV beams permit studies of W production, providing direct and theoretically clean access to the flavor separated sea quark helicity distributions. Transversely polarized collisions have allowed STAR to show that the unexplained large asymmetries present in previous fixed target experiments persist even in the collider regime. The origin of these asymmetries is still not understood and has led to a vibrant transverse spin program designed to study the parton spin distributions in transversely polarized protons. RHIC has also engaged and started a Beam Energy Scan program (BES) and is the only machine that can systematically probe the plasma in the vicinity of the transition by varying both temperature and baryon density. In other words, RHIC/STAR can explore a region of the QCD phase diagram (critical point, phase structure, baryon density) much wider than any other facility is able to.

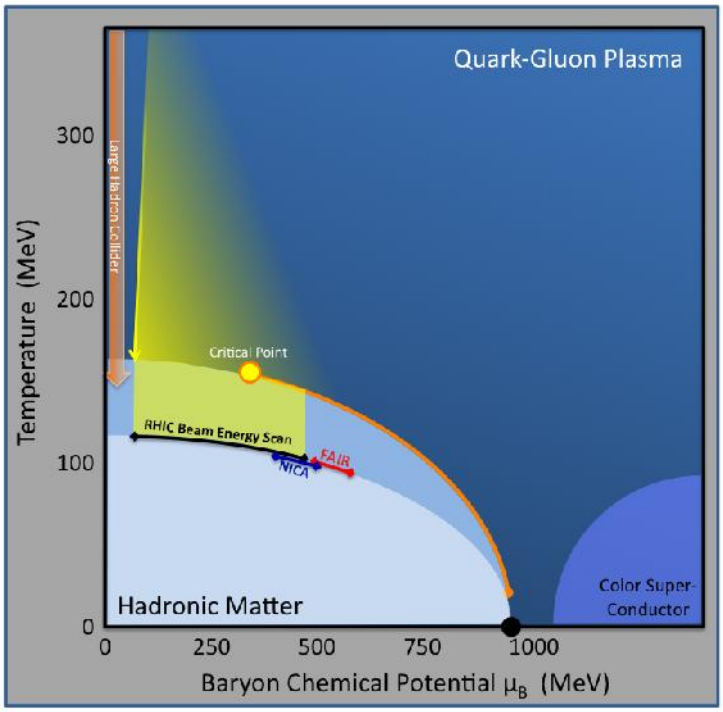

Figure 1: Illustrated phase diagram; in yellow the range RHIC may cover. We also show the LHC, NICA and FAIR coverage of the diagram to better illustrate the unique opportunities of the RHIC program.

RHIC/STAR has now essentially completed a set of major upgrades facilitating the next decade of science. The physics program could be summarized as two major campaign of studies: the first one, from 2014-2016, is focused on Heavy Flavor and Di-leptons measurement to study the properties of the sQGP produced in the high-energy nuclear collisions at μ_B close to 0. The second phase of studies, from 2018 to 2019, will refocus on the RHIC Beam Energy Scan Phase-II. The physics will then be focused on the search for the QCD critical point and study the QCD phase structure at the high baryon-density region $\mu_B > 250$MeV.

To achieve this ambitious program, the first wave of upgrades will provide unique insights on the sQGP properties and focus on the charm and di-lepton measurements. STAR is already equipped with enhanced Particle Identification Detector (PID) systems and is hence able to study a wide variety of secondary decays (including the study of Hyper-nuclei as an offshoot of STAR's physics program). With its new Muon Telescope Detector (MTD), STAR enhances the muon to hadron ratio by orders of magnitudes and will be able to separate upsilon states and study the heavy flavor collectivity and color screening. Combined with the Heavy Flavor Tracker (HFT), STAR will be able to study the prompt J/ψ and non-prompt J/ψ (from B decay) as well as perform detailed studies of the D^0 meson (Run 14 objectives) and later study the charmed lambda or Λ_c. In 2017, the RHIC facility will be equipped with electron cooling capabilities while the STAR sub-system and

central tracking detector will have its inner sector upgraded, allowing for higher tracking precisions (iTPC upgrade). By 2018, STAR will be ready to engage into the deep study of the QCD phase structure and the critical point to gain knowledge as per the characteristics of the phase boundary and the dynamical evolution from cold nuclear matter to hot QGP. The beam-energy scan program has potentials for unparalleled discovery to establish the properties and location of the QCD critical point and to chart out the transition region from hadronic to deconfined matter.

Beyond those time ranges, and past 2020, STAR will have morphed into a superb machine, fully equipped to study the heavy quark, jet and gamma physics and complete its understanding of QCD degrees of freedom as well as covering for a wide range of p+A programs (with a second wave of upgrade including Hadronic calorimetry). The path toward a future eSTAR program will also provide a cost-realizable path to the next QCD frontier with an Electron-Ion Collider (EIC).

Overall, the STAR Beam User Request (BUR) is summarized in Table 1. This BUR is the start of all of our requirements and shall the run plan change or be altered, the numbers reported herein shall change accordingly.

Table 1: STAR Beam User Request from 2014 to 2019.

RHIC run Year	**Species**	**Number of events (B=Billion, M=Million)**
2014	Au+Au 200 GeV Au+Au 15 GeV	2 B (minbias, central) + ~ 0.78 B misc 20 M

2015	p+p 200 GeV p+Au 200 GeV	2.2 B (2 B minbias + trigger mix) 600 M
2016	Au+Au 200 GeV	4.2 B (4 B minbias, …) – large sample
2017	Collider upgrade (eCooling) and STAR/iTPC upgrade	N/A
2018	BES-II p+p 200 GeV longitudinal	400 M (mix of 19.6, 15, 11.5, 7.7 GeV) 1.4 B
2019	P+p 510 GeV, transverse	2 B

We would like to note that the extreme data sample quoted in 2016 is accurately reporting the numbers from the official STAR BUR. However, nota bene is in order: shall STAR be equipped of better vertex constraint capabilities, this data sample will be reduced by a factor of x2. We will re-address this point later in our narrative.

1.2 Data flow background

The RHIC/STAR experiment data taking is initiated from BNL where its data workflow begins. The STAR detector system is currently composed of eight major detector sub-systems (BEMC, EEMC, TPC, HFT, FGT, TOF, GEM, MTD) and numerous triggering systems making the whole data flow composed of ten main area of software coordination.

The Data Acquisition system of STAR itself is currently capable of sustained rates as high as 1.1 GB/sec with peaks at 1.6 KHz event rates. The theoretical limits of the throughput of the DAQ system (based on disk IO for data buffering and local network performance) is 2.5 GB/sec though at a modest cost (about 1k$ / additional 60 MB/sec), the system could be upgraded by adding more hardware online on the STAR side.

STAR is organized in a classic structure of “Tier” centers where BNL is the Tier-0, center of real data collection and the repository of generated simulated data (a copy of the embedding data is brought back to BNL for safe keeping). Tier-1 centers are defined as centers providing a significant resource or service (CPU cycles for data analysis or simulations, archival storage for long term preservation of STAR data, …). Network traffic between BNL and STAR’s Tier-1 centers is the primary object of our requirements. STAR Analysis Centers (SAC, a.k.a. Tier-2 centers) are defined as local compute farms or apportion of main facilities providing analysis cycles to local scientific teams. In this document, we will use the terms of SAC or Tier-2 centers interchangeably. Usually, such center has limited storage resources, hence, network

traffic and load is minimal. However, SAC may move data from anywhere available as STAR has no restriction of strict Tier center hierarchy (and do not see the need for it).

1.3 Collaborators

The STAR institutions' demography and its evolution across the past and present ESnet workshops are represented on Figure 2. As per 2013, STAR remains a strong collaboration composed of 56 active groups and institutions spanning over 3 continents, 5 main geographical groupings (networking wise), 12 countries and 550 scientists.

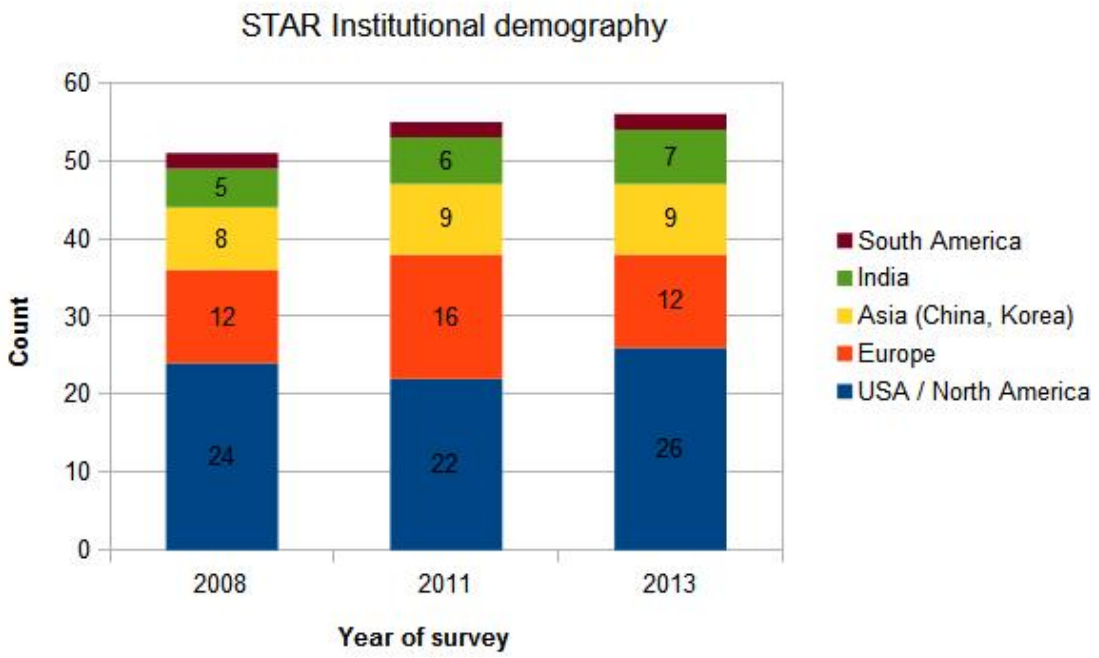

Figure 2: STAR institution evolution over the three ESnet workshops

For the sake of completeness, we would like to note that in Figure 2, we did not include the counting of institutions which are solely composed of emeritus members or institutions phasing out (finishing students) which do not generate network requirements of any kind or are likely to be removed within a year. More importantly, while the demographics remain stable, not all institutions are equal network consumers and it is important to focus on our typical data path and core activities.

Our collaborators remain strongly focused on remotely login-in to either the core facility at BNL *RHIC Computing Facility* (or RCF) located at BNL as STAR's Tier-0 center and the NERSC/PDSF center as Tier-1 center. Both facilities are heavily used for user analysis although that analysis at NERSC/PDSF had been at times

challenged by the need to run aggressive data simulation campaigns (a.k.a. embedding production) sharing the same "rigid" resources (rigid as opposed to "elastic" as a Cloud approach may imply). Our past plan and intents were to ramp up the resources at the *Korea Institute of Science and Technology Information* (KISTI) to palliate for the gap in resource needs to support either the embedding and analysis requirements or create a shift of resources in the data-processing requirements, restoring resources for other processes: at times of conference pressure, the NERSC/PDSF resources have been insufficient and embedding often run at BNL (squeezing users and data production workflows).

In 2013, KISTI previously a Tier-2 center has acquired the status of Tier-1 center – the site not only provides the core processing of the embedding data production but, at 1,000 CPU slots and expanding to another +1,500 slots by the end of 2013, has surpassed the capacity available for STAR at NERSC/PDSF (595 slots usage average in 2012). With its MoU extended to 2017 (included and renewable), the growth at KISTI has opened the possibility to consider the resources as a supplemental for real data production. The network requirements from/to BNL/KISTI would change (as indicated later) but have been already planned in previous workshops (and should not come as a surprise). In addition, a handful of STAR institutions located in China (amongst which, Tsinghua University where our Embedding Coordinator is located) have considered switching some of their analysis workflows to KISTI. It is unclear if this trend will continue as KISTI, as a Tier-1 center, has limited user support possibilities (opening to a large number of users would be counter-productive and STAR is best served by focusing on large scale data productions with limited amount of users).

In our past planning document (SN0548), we envisioned the onset of more *STAR Analysis Centers* (SAC) as the physics program matures and demands for more analysis powers to appear. We also noted that their inventory has been hard to assess but constitute pools of local resources dedicated (not necessarily shared with all STAR users) to local group's physics program needs. We planned on developing strategies to help integrate those centers into a global data analysis and data distribution pool. The status remain the same – as there are no mechanisms to help or encourage SACs to share resources across the collaboration and no clear mechanisms to help supply them with workforce able to maintain/upgrade their local setup, it is not possible to clearly assess their number. In fact, trying to bring those centers in a more consistent picture of the STAR resources by attempting to include them as part of the STAR VO (via the OpenScience-Grid (OSG) software stack and services) has been deterred by the lack of local workforce able to ensure the sustainability of those resources on the OSG/Grid. Support is on a "best effort" basis. Monitoring the number of remote databases (slave servers of BNL master Meta-Data repository), we infer we would have at this stage four active centers which is lower than our past projections by one unit. Those four centers are: Prague (our most stable active center), UIC, Wayne State U. and USNA (MIT has become inactive due to workforce shift). Our new projected number is showed in Table 2. We predict the loss of WSU in 2014 but the regain of one more institution and a plateau in outer years to 3 SACs at most.

Table 2: Projected number of Star Analysis Centers (SAC) from 2013 to 2019. The 2013 estimate represents the number previously projected, the actual number is 4.

	WAN needed for MuDST @ SACs & Tier X						
	2013	2014	2015	2016	2017	2018	2019
Typical number of SACs (STAR Analysis Centers including non-US Tier 2)	5	4	3	3	3	3	3

The STAR computing model continues to rely on a data-grid model and 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 File, Replica and Meta-Data Catalog (we will refer as the STAR FileCatalog), able to make differential inventories between sites within minutes, and the development of in-house tools for reliable data transfer and redistribution. The resources from the OSG are leveraged in a seldom manner and only sites dedicated to STAR's use have been integrated in a Grid based workflow (except the Tier-2 centers as noted above).

1.4 Data size projections –setting the basis for our network requirements

Whenever associated to a file type or file family, the terminology of DAQ or RAW will indicate the files produced by the event collection coming out of the STAR system or the STAR Counting House. The data is essentially composed of raw (not-physics ready) signals coming out of the diverse detector sub-systems packed into binary files. We will use the terminology of DST (Data-Summary-Tape, a rather historical nomenclature) the products of the data reconstruction process where the RAW data is processed and summarized into Physics ready quantities. MuDST or Micro_DST indicates a data sample dominated by the so-called MuDST (but could include a fraction of full event files, histogram based QA and/or tag files the addition of which are not significant). We will refer briefly in our text the pico-DST, a user based slew of derived formats sharing one characteristic across their diversities: their reduced size comparing to the MuDST.

Table 3: Projected event size for RAW and DST files for STAR up to 2019 as a function of species. 2012 and 2013 are showed here as the basics for the extrapolation and projections. The numbers are in units of MB/events.

MuDST size/evts = f(Species)	2012	2013	2014	2015	2016	2017	2018	2019
p+p 500 GeV	0.35	0.42	0.59	1.11	0.92		1.13	1.26
p+p 200 GeV	0.12	0.14	0.29	0.76	0.57		0.64	0.78
U+U 193 GeV	0.45	0.54	0.72	1.26	1.07		1.34	1.47
Au+Au 200 GeV	0.46	0.55	0.73	1.27	1.08		1.36	1.49
Notes		FGT partially added	HFT added (no FGT)	FGT back in STAR + HFT	HFT, no FGT		HFT, iTPC effect	As before + HCAl or similar
p+p 500 GeV	0.59	0.77	0.98	1.62	1.45		1.84	1.99
p+p 200 GeV	0.21	0.27	0.43	0.99	0.83		0.96	1.11
U+U 193 GeV	0.55	0.72	0.92	1.55	1.39		1.75	1.90
Au+Au 200 GeV	0.60	0.78	0.98	1.62	1.46		1.85	2.00
DAQ size/evts = f(Species)	**2012**	2013	**2014**	**2015**	**2016**	2017	**2018**	**2019**

Based on the analysis of past event size, we projected the evolution up to 2019 and summarized the results in **Error! Reference source not found.**. The upper part of the tables shows the size of the MuDST while the lower parts predicts the size per event of the DAQ files as a function of a single species and year. To reach those numbers, projections of the effect of luminosity on the event size have been folded in as well as the phasing-in (and out) of new detectors. The iTPC upgrade alone will cause a data size increase of TPC data by 40% and create a jump is event size.

While imperfect (not all data for the species planned for future runs are available), the expectations of data size growth can be inferred by folding the values from **Error! Reference source not found.** and the STAR run plan alone in Table 1. This would lead to the resulting estimates of Table 4.

Table 4: Event size projections considering the species mixed foreseen by the STAR BUR.

	2013	2014	2015	2016	2017	2018	2019
Species		*Au+Au 200 GeV BES 15 GeV*	*p+p 200 GeV, p+Au 200 GeV*	*Au+Au 200 GeV*	*N/A Machine Upgrade (eCooling) & iTPC*	*BES-II (multiple energies) p+p 200 GeV long*	*p+p 510 GeV trans*
Expected Total number of events	N/A	2.80	2.80	4.20		1.80	2.00
Estimated DAQ event size	*0.77*	0.98	1.06	1.46		0.80	1.99
Estimated MuDST event size	*0.42*	0.72	0.81	1.08		0.54	1.26

From the expected dataset mix (species, trigger) and their respective event size average, we can them make projections as per the yearly dataset size we will encounter for the period of 2014-2019 – while the 2019 is beyond the required timeline of this workshop (up to 2018), it seems judicious to include it for two reasons: (a) 2017 marks a machine and detector upgrade period during which the data requirements for RAW will be null hence, going up to 2019 maintains the same amount of years for the RAW data and (b) the RHIC/STAR BUR sets two clear physics program objectives one of which is past the 2017 machine upgrade. We summarize those projections in Table 5.

Table 5: Projected data set size for the 2014-2019 period. The two first years are showed as basis for the projection and verifications.

	Initial projections						*Outer years projections*	
	2012	2013	2014	2015	2016	2017	2018	2019
Species	U+U 193 GeV, p+p500 and 200 GeV BES 5 GeV Cu+Au	p+p 500 GeV	*Au+Au 200 GeV BES 15 GeV*	*p+p 200 GeV, p+Au 200 GeV*	*Au+Au 200 GeV*	*N/A Machine Upgrade (eCooling) & HLT*	*BES-II (multiple energies) p+p 200 GeV long*	*p+p 510 GeV trans*
Projected N events (B)	2.20	2.50	2.80	2.80	4.20		1.80	2.00
Projected size RAW (TB)	1321.05	1801.44	2800.04	3025.05	5280.19		1466.39	1066.13
		All data	*Trend projections (upper bound considering hsitorical deviations to plans)*					
N events (B)	6.1	2.7	*3.1*	*3.0*	*4.4*		*2.0*	*2.2*
Final size RAW (TB)	2147.24	1985.43	3080.05	3267.05	5594.20		1613.03	4472.75
Deviation to projected	93.18%	8.00%	10.00%	8.00%	5.00%		10.00%	10.00%
	4249958673							
	5.3	2.54						
RAW Data with tracking detector (candidate for data production)							*Projected based on possible excess*	
sum(events) tpx	5586466071	2728446873	*3080000000*	*3024000000*	*4410000000*		*1980000000*	*2200000000*
sum(size) tpx (TB)	2140.1	1980.06	*2868.52*	*3042.68*	*5141.32*		*1502.25*	*4165.57*
Size / events (MB)	0.39	0.76	*0.98*	*1.06*	*1.46*		*0.80*	*1.99*
Initially projected in 2011	*0.59*	*0.70*						
Real up to 2012, complete up to 2011, 2013 onward are projections							*Projected for derived data (MuDST)*	
Total events MuDST	3753824889	*2526128181*	2851613083	2799765572	4082991459		1833179839	2036866488
Fraction of events to RAW	68.33%	92.58%	92.58%	92.58%	92.58%		92.58%	92.58%
Total size MuDST (TB)	931.11	*1060.01*	1967.08	2173.92	4211.19	2784.06	937.13	2450.47
Size / events (MB)	0.26	*0.44*	*0.72*	*0.81*	*1.08*		*0.54*	*1.26*

A few notes are required.

As in the past requirements estimates, we note that STAR has often exceeded its goals in terms of number of events to be taken. For the purpose of science, the more events the better but for the purpose of resource estimates, this has introduced an uncertainty in planning for computing we cope by adding a factor showed in the "Deviation to projected" row. The 2012 and 2013 values are factual numbers while beyond, the values are projected. To better understand how much of the data is usable for data production (hence Physics), the row "Fraction of events to RAW" (last block at the bottom, second row) is a good indicator of data usability – this number can never be 100% for many reasons: early problem detections in the run (detector trips, questionable data quality based on QA plots, …) would account for a measured 3% drop alone. Other reductions include data taken for specialized studies but not including the main tracking detector and data marked as of "no physics quality" as problems may have been uncovered at analysis levels. The 2012 value of an excess of 93.18% is however an artifact – the Cu+Au data sample was not part of the initial STAR's BUR and it is to be noted that on this year, the calculation of "Fraction of events to RAW" does not include this dataset.

The second note is that while our past projections (ESnet report from 2011) expected a RAW event size average of 0.70 MB/events in 2013, the run plan was modified for the benefit of one species (the mix is different, the average event size is impacted). **Error! Reference source not found.** would however indicate an event size of 0.77 MB/events for 2013 and our final number is remarkably accurate at 0.76 MB/events. The MuDST size per events is speculated to be slightly larger due to a few detectors added to the data stream, the information of which will need to be propagated with redundant information so the detector response can be better understood.

We noted in section 1.2 that STAR has for plan an extremely large dataset in 2016. This impressive data sample is driving the requirements but may be reduced by a factor of 2 depending on STAR's ability to select the primary event vertex with a cut of less than 5 cm accuracy. This deliverable is not formally a computing deliverable (and hence not immediately under our control). No detector setup can though achieve this vertex selection at this stage. Nevertheless, this selection can be achieved by ensuring that High Level Trigger (or HLT) vertexing capabilities, in addition of tracking, are in place by 2016. At the time of this workshop, the same team of computer scientists from Germany (FIAS), now having full membership of the STAR collaboration and with whom we collaborated on the HLT tracking before (along with CBM, ALICE and other experiments) are visiting BNL. Along with physicists from several STAR institutions, a focused effort was organized by computing to tackle this problem. For the sake of projecting and making sure STAR does not fall behind network resources, we did not fold this (yet unproven) possibility but did align with the BUR requests for consistency. Though, our conclusions will repeat this fact as words of caution, we will systematically consider the reduction of this dataset by a factor x2 wherever applies and proceed with gross approximations to the lower value. Our confident in the steering of this deliverable in time for the 2016 run is very high. In other words, it would be extremely premature to draw conclusions as per the implied storage requirements and strain on the facility such datasets may imply for the facilities hosting STAR data.

Finally, while there is no run foreseen in 2017 (for the benefit of major machine and detector upgrades), we made calculations of network requirements on this year based on an average data sample size from the previous three years average. In 2017, high priority data re-production will be scheduled as the current CPU resources at our facilities no longer allows for 2 passes of data production (but one).

Our science case being lay down and the rationales behind the derived data sets size being explained, we can now focus on a purely network centric aspect.

1.5 Key Local Science Drivers (e.g. Local Network aspects)

In this section, we will essentially focus on the Tier-0 aspects and LAN requirements and will treat all other facility requirements in section 1.6 and related sub-sections.

1.5.1 Instruments and Facilities

Describe compute, storage, and network capabilities, any connections to any major scientific instruments (e.g.: supercomputers, particle accelerators, tokamaks, genome sequencers, satellite data, computational clusters, storage systems, etc.)

The BNL RHIC Computing Facility (RCF) is hosting all RHIC experiments and the core operation and role of the facility is to provide the core CPU computing cycles

for ½ of our user analysis' needs, the whole of data reconstructions, support for data calibration, data reduction, database and some local need for simulations.

During data taking, the STAR DAQ system streams data to a cache space spread over 10 buffer box nodes (nodes collecting and aggregating the data into streams and files) for a total of 96 TB disk space. In this configuration, and depending on the DAQ rate, but assuming 600 MB/sec rate, STAR would be able to hold its ground for ~ 46 hours without network connectivity before suffering any data loss. At observed peak rates of 1.1 GB/sec, STAR would still maintain operational viability for 24 hours. The data is though pushed to the RCF via 2x10 Gb lines onto a disk cache of 54 TB space (near a 2:1 space match) located in front of the High Performance Storage System (HPSS) tape archiving system. STAR has accumulated about 12 PBytes of storage space in HPSS to date (~ 7.6 of which are RAW data). Datasets from the year e012 onward have been multiple PBytes size large and driving the bulk of this size.

Based on average run time and hours of physics running suitable for data taking observed in previous years as well as the input from Table 5, we infer the LAN requirements from the DAQ to the HPSS systems as showed in Table 6. The maximum line speed (sustained) needed for the entire period exceeds a 1x10Gb line but remains below 2x10 Gb lines. As far as the LAN connectivity is concerned, STAR is currently covered for both sustained and peak rates (peak rates at the 13.6 Gb will exceed the 2x10 Gb line capacity but the data caching will make it possible without additional resources).

Table 6: Network LAN requirements from the DAQ to the HPSS systems for the period covering 2014 to 2019. 2013 is showed for historical purpose. A margin was folded in the calculation to account for possible protocol overheads.

	LAN need from DAQ to HPSS						
	2013	2014	2015	2016	2017	2018	2019
LAN, DAQ to HPSS gross average [+20%] – Minimal (MB/sec)	488.41	625.64	663.63	1339.46	0.00	327.65	908.54
<Peak> DAQ → HPSS LAN [+20%] (MB/sec)	463.74	816.97	866.58	1749.09	0.00	427.85	1186.38
All times LAN rate needed (MB/sec)	566.40	816.97	866.58	1749.09	1749.09	1749.09	1749.09
LAN (Gb/sec)	4.43	6.38	6.77	13.66	13.66	13.66	13.66

The facility currently provides CPU powers of the order 76 k HSPEC delivered by over 9,192 CPU cores. The total storage capacity has reached about 560 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). The CPUs are standard off-the-shelf commodity hardware and nodes hosting local storage (cheap disks) for a total of ~ 3.3 PBytes of distributed storage space holding a portion of our DST files. Distributed storage have come to be the main resources of storage for analysis files since 2010 or so.

1.5.2 Software Infrastructure

Describe the software used to manage the daily activities of the scientific process in the local environment. Please include tools that are used to locally manage data resources, facilitate the transfer of data sets from or to remote collaborators, or process the raw results into final and intermediate formats.

The DAQ data rate and data flow was described in the previous section. The data from the experimental data taking area (DAQ network) to the HPSS storage system is moved using a home-grown version of pftp. This version is more suitable for data streaming and has some intelligence triggering data transfers (round robin mechanism selecting multiple drives attached to each buffer box and avoiding simultaneous read/write if possible, read when disks are not too busy for write). When the data has reached HPSS, we consider the data within the RCF realm (where the CPUs and storage are located). A fraction of the data is analyzed online (online Quality Assurance or onlineQA) for identifying gross problems with detector responses.

The data is retrieved for processing out of HPSS via a data batch system (the ERADAT system) deeply embedded into the data production software (both are home developed systems). Essentially, data production campaigns restore one DAQ file per job and produce many files as output (the essential of which are our DSTs). The optimization done by the production system is that the DAQ files are restored in an optimal manner and as they are located on tape (publication doi:10.1088/1742-6596/331/4/042045 better describes the process of optimization). During data taking, a fraction of the data is sampled and reconstructed via the standard track reconstruction software for additional QA and calibration support – this process is known as the "FastOffline" processing and typically samples ~ 8-10% of the data but limits the processing to a 1,000 events per DAQ file (75% of the runs were QA-ed this way in the run 2013 and improvement to previous years at 50% coverage).

During a full data production campaign, all files (and all events) flagged for data production goes through the data production process. As the data is distilled into DST, the result are then double copied: one copy goes to the HPSS storage for permanent archiving and a second copy is randomly placed on one of the 80 file system partitions available as data production space (the random placement is done for load balancing purposes). Indexer daemons picks newly created files as they appear and immediately Catalogs them in the STAR's FileCatalog. During this process, the file's checksum and size (queried or computed during production time) are verified– if either do not validate, the file is not Catalogued and flagged as "bad". At the end of the production campaign, they may be re-produced. This paranoid check, mainly implemented in case of network communication oddities, has not detected a single occurrence of such event for the past two years (below a 2% loss due to this effect or other core common problems, we do not re-submit). As the Cataloguing occurs, the presence of an HPSS copy is checked – if present, the NFS file may be removed immediately, if not present in HPSS, the NFS copy is pushed again

to HPSS (a few percent failures in the production workflow in moving the data to HPSS occurs). Typically, the NFS files are NOT removed to allow the next stage to take place.

The workflow above was envisioned to be altered to avoid the extra step of a copy in HPSS (xrdcp or a direct copy into "a" local disk space was planned). However, STAR distributed storage capacity is at ~ 70% of its requirements therefore, the deployment of a streamlined workflow was delayed (in other words, we did not see as a benefit to add the handling of errors and delays due to failure of a copy via an xrdcp method tight to the production flow but leave an external process to handle it). We projected this obstacle will be removed by the 2014 purchase cycle (within the past established funding profile and projections, distributed storage will be sufficient to automate the production workflow in the outlined manner).

Datasets of interests are registered in the STAR Data-Management system as candidate for distributed disk population. Individual daemons from ~ 500 nodes consult the STAR FileCatalog and evaluate the missing dataset portion from distributed disk. If the missing dataset is found from NFS, the files are copied over the network unto the "a" node's local storage using a standard 'cp' command. If not, a centralized process issues a full differential list and schedules the missing datasets for restore to the DataCarousel. The data management system knows of disk space availability at all times. Apart from its coordination, built-in faire-share and optimization mechanisms, the DataCarousel relies on a connection to HPSS via pftp but could rely on any other tools. The central storage data is either removed on demand or automatically and bulk removed (for example, logic such as "*if the data is on distributed storage, remove from NFS*" or "*make sure at least two copies exists on storage element XX*" or "*remove all data from NFS from the 2010 campaign*" are trivially possible actions in the current STAR's data management system).

At the end, the data is evenly spread over the massive 3.3 PBytes virtual storage aggregated using Scalla/Xrootd and hence, access to "a" dataset over the facility likely involves the whole set of nodes (there is no special or logical portioning done at this stage). However, and providing all daemons are maintained active and in good standing, the temporary loss of a fraction of the dataset will be detected (within 20 mnts) and the missing data restored.

Typically, a Gb/sec interface to each node is sufficient to restore the occasional data loss from each node. A massive restore of data (1/2 PBytes) to 500 nodes with this network bandwidth can be done within 2.3 hours assuming no constraint of throughput from HPSS.

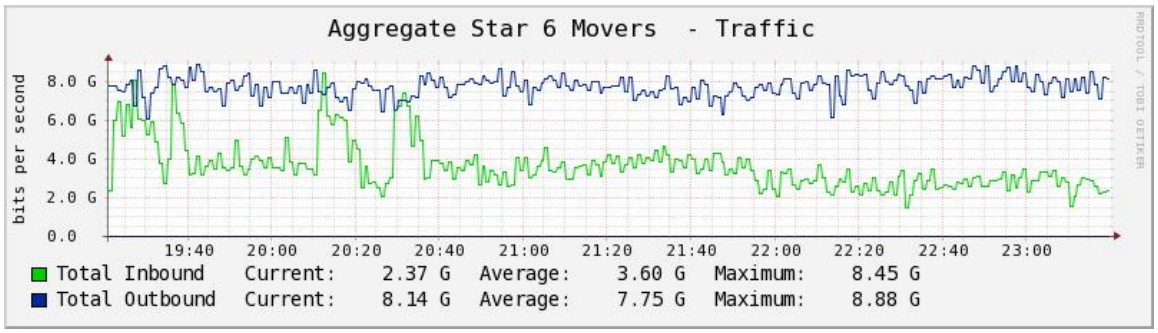

Figure 3: Input/Output graph during a Mock-IO Challenge to/from HPSS. Both data move from the experiment (blue) and data read for production purposes (green) were simulated in a realistic load environment.

Figure 3 represents the expected data throughput to/from HPSS in simultaneous read/write mode. The HPSS system has been showed to provide an aggregate of 4 Gb/sec (peaks at 8.45 Gb/sec are rare in this case) and thus, the restore of such dataset loss (½ PBytes) would actually map to a 12 days restore of the data. To reduce this intrinsic limitation, capacities to the HPSS itself would need to be expanded (the network is not the limiting factor in this process).

One consequence of those lengthy restore of our large datasets (and becoming larger) is that the dynamic "*on the fly*" (or on-demand) disk population of datasets is a rather conceptual ideal of no practical use unless jobs submitted to a batch system could be delayed for as much as weeks long. Therefore, STAR data are pre-staged on distributed disk based on feedback and observations. There are two sources for such feedback and input: (a) The Physics Working Group (PWG) are regularly polled for their dataset usage intents (ordered by priority) – those input are summarized across all PWG and, depending on space availability, the datasets of highest cardinality in the number of request dimension are replicated across the virtual storage pool while the lowest priority (and lowest occurrence) have a single copy available. (b) The second input is the usage from STAR users themselves – STAR users submit their jobs via a job submission interface allowing them to specify datasets based on Meta-Data declaration. Their usage is recorded and monitored. The monitoring includes aggregate information related to the currently accessed datasets and most accessed datasets and data production campaign as a function of time range (past days, weeks, months, year). An example of such a graph is showed in Figure 4. Evolution of analysis pattern as well as indicators of hot datasets (datasets most used) can be inferred from those graphs and datasets replicated accordingly.

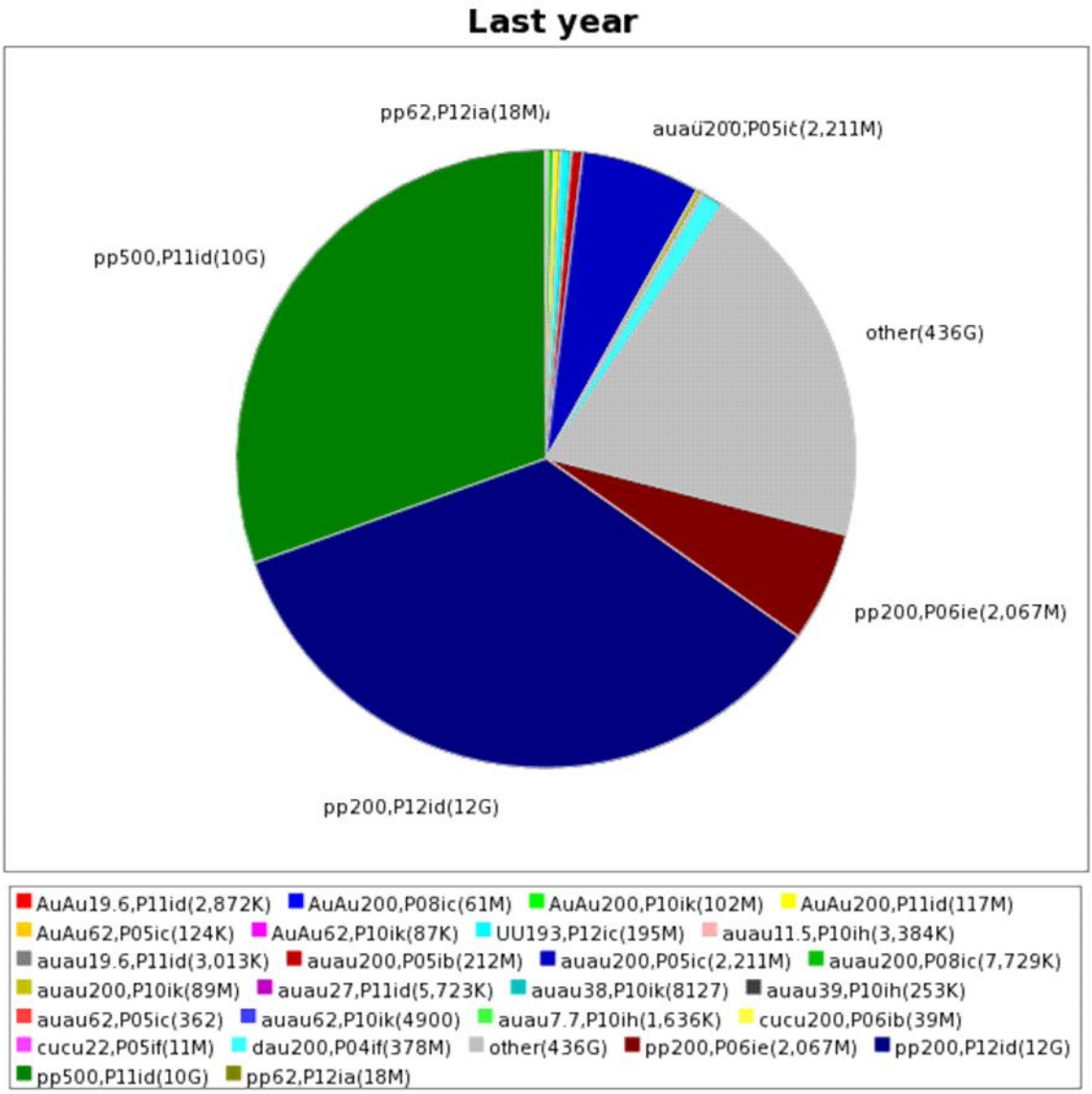

Figure 4: Usage statistics example of data access pattern from users for the past year, in this case the statistics gives an idea of dataset access pattern by production and collision species. This information is used to determine the actual "hot" datasets.

Finally, the lengthy cycle for data restores in case of a data loss points to the need to secure the distributed storage (resilience and redundancy). The generalized use of RAID-5 based local storage will reduce such data loss scenario. This will be in place in all future storage and the space loss for going toward a local RAID solution will be folded in our storage requirements calculations.

1.5.3 Process of Science

Describe the process by which scientists use the instruments and facilities for knowledge discovery, emphasizing the role of networking in enabling the science.

In STAR, Scalla/Xrootd has been used since its very early days and is still in use in STAR. All science processes from data production, calibration, user analysis or simulation are handled by a single framework a.k.a. root4star. This single framework relies on the ROOT package and Scalla/Xrootd plugin is a de-facto component installed along the STAR software.

The resources for STAR at the RCF are separated into two sections: an analysis farm (a.k.a. "CAS") and a production farm (a.k.a. "CRS"). While data movement through the CRS nodes are hard to interpret, at full farm occupancy, the jobs on the CAS are essentially user jobs reading data from Scalla/Xrootd and reducing the data to picoDST or histogram files (the IO of which is negligible). A few typical IO profiles of our nodes are showed in Figure 5 and Figure 6. Both nodes have similar storage space and show an IO rate in the node around 12 MB/sec and out of the node at about 5 MB/sec. To first order, those rates do not concern us considering the 1 Gb/sec network interface.

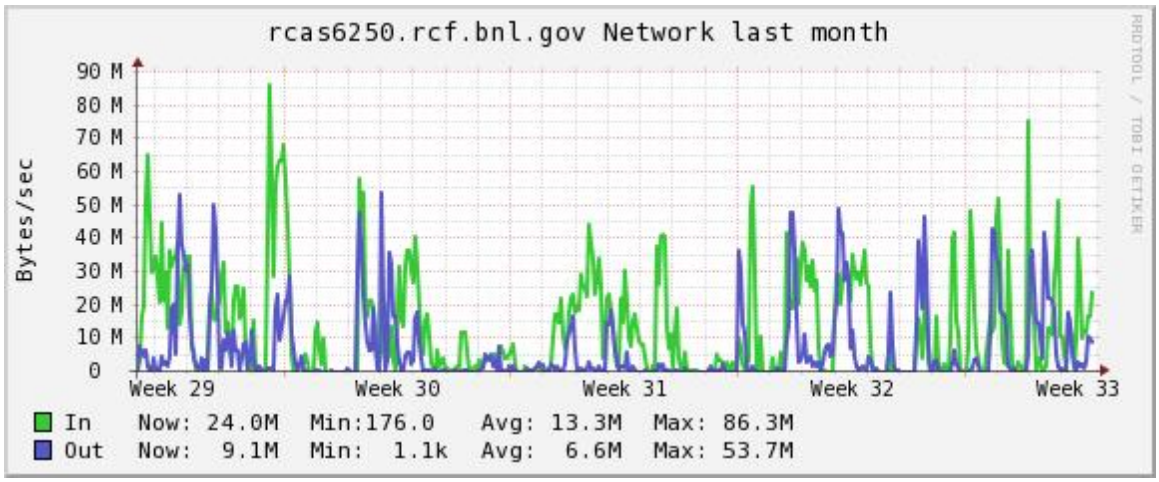

Figure 5: Typical IO in and out of a node on an Analysis node.

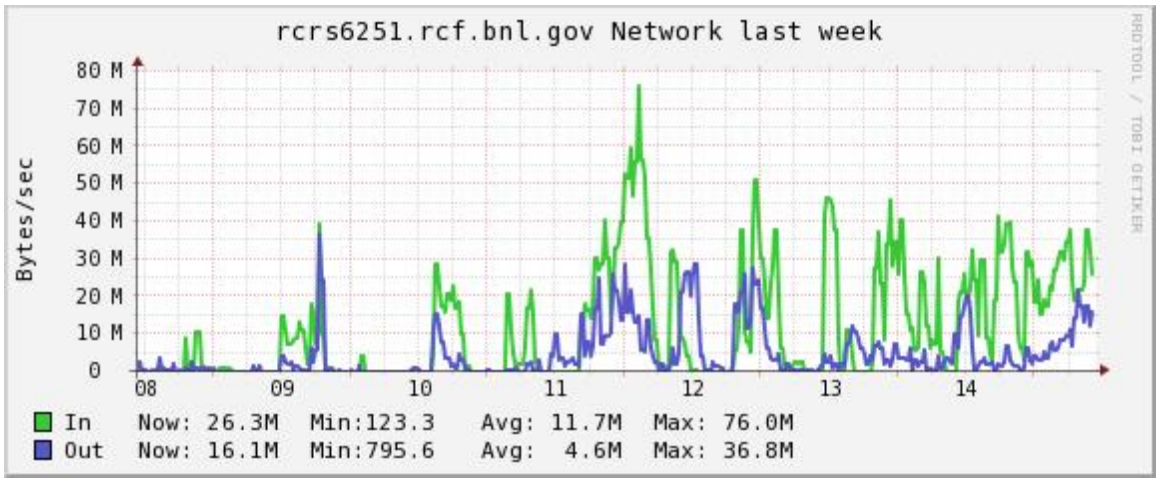

Figure 6: Typical IO profile in a period of no data production campaign.

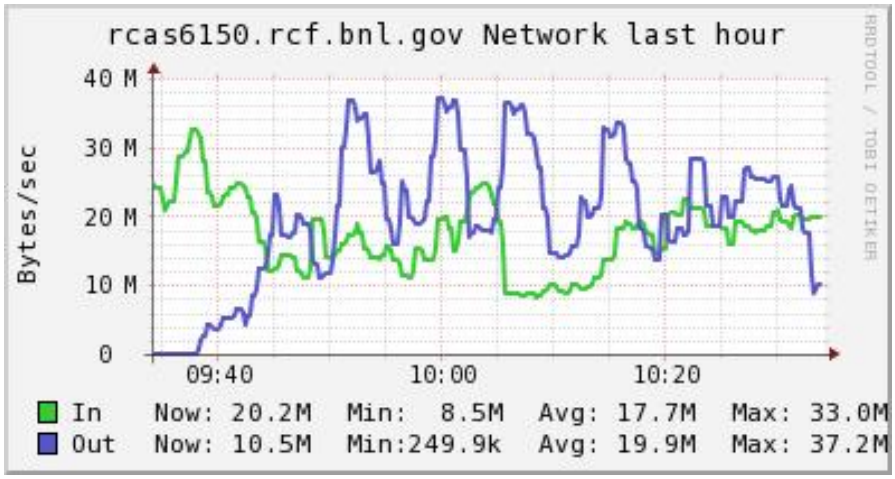

Figure 7: IO rate for an Analysis node narrowed to a peak load

However, we are concerned of the IO rates as showed on Figure 7, a "zoom-in" presentation of peak load of activities. The main component of the IO "Out" (in blue) could only be explained by access of the node's local data via Scalla/Xrootd access, data going out of the node on the LAN serving other nodes / jobs on the farm. In this example, we see peaks at 37 MB/sec (and have observed flat IO rates at 40 MB/sec during analysis intensive periods so this example, taken before this workshop, is not uncommon).

We know the IO "out" will be proportional to both the amount of data located on a given node and the number of batch slots across the facility. With a new incoming farm node purchase with x4 more data attached to each node, the risk to exceed the capacity of a 1 Gb line (hence having potentials for lengthy IO saturations, causing job efficiency issues via IO starvations) seem to be an imminent and palpable reality we will have to face in the coming year. The need for capacities > 1 Gb is, in our

view, a tangible and an immediate LAN requirement within our distributed data and data flow model. Perhaps the enabling of ROOT/Scalla IO re-ahead (not done to date) may alleviate this issue (for the most IO challenging jobs, it is likely to make it worst). Compute nodes with x2 the number of cores will not create this demand as far the IO "in" is concerned but the evolution of core density and storage space will certainly need to be watched and considered on this widely distributed data model as impacting LAN requirements.

1.6 Key Remote Science Drivers (e.g. Wide Area Network aspects, remote collaborators, data transfers)

1.6.1 Instruments and Facilities

Describe remote access to or transfer of data from remote compute, storage, and network capabilities, any connections to any major scientific instruments (e.g.: supercomputers, particle accelerators, tokamaks, genome sequencers, satellite data, computational clusters, storage systems, etc.)

The NERSC/PDSF and KISTI facilities are primary consumer and producer of data from the STAR/BNL Tier-0 center.

The resources at NERSC/PDSF are focused on providing CPU cycles for the embedding process, a process where real data and simulation signals are fused into the same data stream and thereafter reconstructed as real data would. The analysis of how efficiently the simulated data could be reconstructed gives a measure of the geometrical, reconstruction and environmental effects on detection efficiencies. Efficiency corrections are needed for all STAR published papers if any quantitative comparisons are to be made – this represents most of our papers making the embedding production and particularly important step of our scientific deliverables. The resources at NERSC/PDSF are also used for providing a number of users (a few groups in the "*region*" constitute the most common users, including the local scientific group at LBNL, UC-Davis and their visiting scientists) a pool of resources for user analysis. Effectively, any STAR user many request an account at PDSF.

The resources at NERSC/PDSF are shared between many projects and apportioned based on resource allocation cycles. In 2012, STAR had 300 slots of official allocation and 595 slots of actual average usage. The excess in resource usage can be easily understood as the site, of very modest size, always tend to have more jobs than what our allocation may digest. Hence, at a low down of other experiment's usage, the additional CPUs are taken. NERSC/PDSF as a Tier-1 center also provides permanent archival storage. In our planning, we consistently aim at providing space for a full copy of our DST in the NERSC/HPSS system. Practically, we lack a dedicated person at that Tier-1 center for data handling (the embedding deputies try at best efforts and transfer the DAQ files needed for embedding, the DST transfers tend to lag far behind) and only a small fraction of the DST are moved.

Table 7 shows the network bandwidth needed for the diverse categories of transfers. The first row, presented in Table 2, is used for evaluating the compounded network load on the facilities holding the data. The second row indicates the network resource needed to a Tier-1 center for being able to transfer all MuDST within a 6 months period. This minimal bandwidth is needed for NERSC/PDSF. The last row assumes that 1/3rd of all SACs takes the data from PDSF while 2/3rd would from BNL (4th row). Those network requirements allowing SACs to transfer data from our Tier-1 and Tier-0 respectively are indicated for completeness. In the case of PDSF, those requirements are not additive (the time frame for transferring the MuDST is quoted as 3 months while the data transfers are estimated as burst transfers over year period). Typically, the larger of the two numbers is needed as a connection speed from PDSF.

Table 7: Network bandwidth needed by SAC or Tier-1 centers depending on activities.

	WAN needed for MuDST @ SACs & Tier X						
	2013	2014	2015	2016	2017	2018	2019
Typical number of SACs (STAR Analysis Centers including non-US Tier 2)	5	4	3	3	3	3	3
Tier 1 center [100%, 3 months] (Gb/sec)	1.12	2.07	2.29	4.44	2.93	0.99	2.58
Individual SAC/Tier 2 bdwdth need [rotation at 10% datasets, 3 weeks] (Gb/sec)	0.48	0.89	0.98	1.90	1.26	0.42	1.11
Total SACs bdwdth out of BNL [assume 2/3, 1/3] (Gb/sec)	1.60	2.37	1.96	3.80	2.51	0.85	2.21
Total SACs bdwdth out of NERSC [assumes 1/3, 2/3] (Gb/sec)	0.80	1.18	0.98	1.90	1.26	0.42	1.11

The KISTI Tier-1 center is a center equipped of a 1,000 CPU slots and 150 TB of centralized storage space. With a steady growth planned for the period of our extended MoU (up to 2017, renewable). Another installment of 1,500 CPUs is planned by the end of fall. The CPU growth is foreseen as of the order of 500 to 1,000 CPUs / year for the period covered by this report (exact number need to be confirmed by mutual agreement – the final resource plan evolution for KISTI was not yet crystalized at the time of this report). The facility is rather heavily used and all slots allocated to STAR are typically busy as showed on Figure 8.

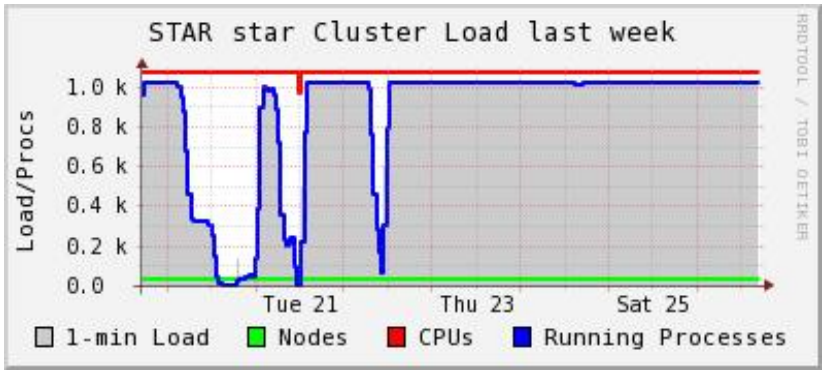

Figure 8: CPU load typical profile at KISTI. The offset running/maximum is a monitoring artifact: all nodes, including our databases and Grid gatekeepers are part of the same graph but do not run jobs.

The site currently and essentially supports the embedding process and its operation has exceeded those of NERSC/PDSF. A small amount of (local) users (from Tsinghua and Macau) also use the resources for user analysis. As noted in our introduction section 1.3, our KISTI Tier-1 center is supplied with minimal user support (we have one point of contact from the facility handling all user's requests) and hence, we do not envision the growth of user analysis activities beyond the opportunistic use from those who supports the embedding data productions at KISTI. KISTI does not have permanent archiving storage and hence, the data produced are either brought back to NERSC or to BNL.

The requirements for the transfers of DAQ files from to BNL/NERSC and/or KISTI for embedding support are not indicated nor considered in any of our calculations. This is due to an extreme streamlining of our embedding process at this stage of experiment maturity. The embedding productions now require only a very small fraction of the RAW data for processing. The streamlining has been effectively achieved by an enhanced coordination and planning of the process and workflow. The Physics Working Groups are polled far in advance, the requests for embedding filed in a request system, similar requests are identified and often DAQ files usable for multiple requests are located and tagged for transfers, reducing the demand for large sampling. An order of magnitude is that KISTI has held about 5 TB worth of DAQ files for the past 6 months of constant operation while NERSC/PDSF has seen of the order of 50 TB of DAQ files at most.

Finally, all data produced by the embedding workflow are to be brought back to BNL. At a ratio of input/output of 1:7 to 1:10, the amount of data to be transferred is still below the threshold to create even a second order effect on network requirements.

Table 8: Network bandwidth requirements necessary for moving DAQ/RAW data from BNL to "a" remote site for remote data processing. Since the result of production must be brought back to BNL, we also indicated the bandwidth needed on the BNL side for this process to occur.

	WAN needs, N% processed offsite						
	2013	2014	2015	2016	2017	2018	2019
WAN need for 20% RAW moved offsite [Cloud / Tier1] (Gb/sec)	0.76	0.98	1.04	2.09	1.37	0.51	1.42
WAN need for 20% MuDST back to BNL [Cloud / Tier1]	0.28	0.52	0.57	1.11	0.74	0.25	0.65
Total WAN for 20% offsite processing [Cloud / Tier1] model (Gb/sec)	1.04	1.50	1.61	3.21	2.11	0.76	2.07
Total WAN for 50% offsite processing [1/2 pass "as we go"] (Gb/sec)	2.61	3.74	4.03	8.02	5.26	1.9	5.17
Total WAN for a one time copy of all raw offsite (Gb/sec)	3.82	4.89	5.18	10.46	6.85	2.56	7.1

However, and due to the rapid growth of KISTI, STAR computing is considering its use for real data production. Constrained to essentially one pass of data reconstruction per year at BNL (far below acceptable Physics objectives and below our planning), the resources at KISTI cannot be under-considered. The rapid CPU growth is in fact essentially planned within that objective in mind. Table 8 gives estimates of the network bandwidth needed to allow data production to occur to a remote site (or Cloud processing). The first row indicates the bandwidth needed for a 20% data transfer occurring right away during and along data taking (while a copy is done to HPSS, another would be pushed through to the remote site – in collaboration with ESnet, this has been exercised in STAR and showed to be possible in 2009). The second row indicates the additional bandwidth required for bringing the data back to BNL. The third is the sum of the first two indicating the bandwidth needed in total to/from KISTI. The 4th row is the same global calculation pushing data production of ½ of the data at KISTI (this would allow restoring at least 2 production passes within one year – it is our actual target).

Other facilities and activities worth noting are:

1. the support of SAC centers, summarized in Table 7, indicates on the 3rd row the bandwidth needed for each SAC for being able to use their limited storage and copy datasets (at a 10% level replacement or transfer every 3 weeks) for sustaining local science. The bandwidth indicated there are marginal but need to be considered by each SAC.
2. The possibility of a full copy of ALL RAW data to a secondary facility for the long term preservation and safety of STAR data has been long discussed and considered. The bandwidth required for this process is indicated in the last row of Table 8. The possibility of leveraging our current partial data copy away from our Tier-0 center will need to be decided within the next 2 years.

In both Table 7 and Table 8, we would like to remind our reader of an uncertainty for year 2016 which will likely see a factor of x2 drop in the network bandwidth requirements.

1.6.2 Software Infrastructure

Describe the software used to manage the daily activities of the scientific process in the wide area environment. Please include tools that are used to manage data resources for the collaboration as a whole, facilitate the transfer of data sets to or from remote collaborators, or process the raw results into final and intermediate formats. The objective is to facilitate discussion of the software tools that move data over the network.

All STAR sites use the root4star framework for their scientific process.

The use of Scalla/Xrootd is at a test level at NERSC/PDSF and access of data is essentially done via centralized storage at both PDSF and KISTI through NFS/GPFS storage. Our Prague site continues its use of a mix of DPM (historical use) and direct NFS access of the data. Typically, no other tools than our STAR unique framework (relying on ROOT and its adequate site specific plugins) are needed.

Most sites use the STAR Unified Meta Scheduler (SUMS) for submitting jobs. This tool monitors and records user's requests as we already noted in section 1.5.2 though, at remote sites, the monitoring capability is often not enabled. The benefit of using SUMS is that similar (to identical) job description can be seamlessly moved between sites for achieving the same results (providing the same datasets are available) regardless of the site's choice of batch system. Most workflows are local (that is, not based on distributed computing, Grid or Cloud processing).

It is a noteworthy notice to mention that the user's general pattern has also included the use of so-called picoDST. Of no specific designed format (but based on ROOT trees), their size are a fraction of those of the MuDST and from a $1/5^{th}$ to a $1/10^{th}$. The data transfers are handled in a non-organized way in some instances (BNL to PDSF transfers are using grid tools but transfers are also ongoing between PDSF to China with no clear prescription).

Simulation production and library regression tests suites are steered from BNL also using SUMS but in "Grid" mode. The jobs are in this case distributed. Library validation and regression test suites of software installed at our remote sites constitute a marginal operation comparing to the massive need for data production. But those operations allow maintaining thin support teams at remote sites (as the libraries and codes are centrally validated by a single "librarian") and hence of high value. We would like to note however that in the case of a KISTI based data production, the workflow being tested as this report is being written will be relying on a distributed computing paradigm (leveraging grid tools for data transfers to first order) – KISTI being interested in Cloud computing, the infrastructure is opened to questions but the 2013 exercise will leverage the in-place grid gatekeepers from both sides. Our KISTI site is already part of the OSG infrastructure (registration as a STAR resource need to be verified).

1.6.3 Process of Science

Describe the process by which scientists use remote access to instruments and facilities or data from remote instruments and facilities for knowledge discovery, emphasizing the role of networking in enabling the science.

Please also describe the data workflow used – what tools are used to move the data, what sites are involved, how the analysis tools interact with data movement, what performance is currently achieved (and what performance is needed, if different), and so forth.

Data transfer flows will be described essential from a NERSC/PDSF, KISTI and Prague viewpoint.

Between NERSC/PDSF and BNL, grid based data transfers are used. Typically, Globus Online (GO) and globus-url-copy (guc) are used for transfers. Data may be grabbed from Xrootd onto an export cache using xrdcp (this load is not significant to impact user access to the distributed data at BNL). STAR is equipped of 4 Grid gatekeepers (2 are shared with the OSG general VOs, 2 are dedicated to STAR specific use). On the NERSC side, two end points may be used for the transfers. Rates of 200 MB/sec would be typical for transfers using guc while 100 MB/sec using GO but those transfer rates are limited by the end point capacity. Those rates are sufficient for the 2013 data transfers at low priority but will likely not suffice at the onset of larger datasets as seen from 2014 onward.

The data flows to/from KISTI consist of two paths. DAQ files are transferred from BNL using the Fast Data Transfer (FDT) tool and the product of embedding production for permanent archiving are also brought back to BNL using FDT. The current data rates are 40 MB/sec, not an impressive data transfer rate but sufficient for the current need. Shall raw data transfers occur, the network connectivity and expected speed would need to be revisited – as previously discussed, a 2013 operation would require a ~ 1 Gb connectivity while a 2014 operation would require 1.5 Gb capability. Typically, these bandwidths are in place but end-to-end tuning is needed to take the full capacity. Embedding results are also copied from KISTI to NERSC/PDSF using guc. Using multiple threads for the transfer (after studying the saturation point), rates of 300 MB/sec has been showed to be possible between those two sites.

Data transfers from NERSC/PDSF and/or BNL to Prague are handled using FDT as the underline transport. Data is also grabbed from BNL/Xrootd using xrdcp. Prague has continued onward to consolidate 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 already showed, reported and published that the use of such techniques has the potential of reducing data transfer makespan by 30%. Recent work focused on the use of local data caches and best space reclamation strategies (based on user's access and data demand pattern). All work has been carried through thesis students (master or PhD in Computer Science). We feel that within a year or two, a fully optimized system will be complete for STAR use, factoring in multiple sources for dataset provenance, network bandwidth and availability and cache optimization.

1.7 Local Science Drivers – the next 2-5 years

1.7.1 Instruments and Facilities

Describe the instruments and facilities as they will be in the next 2-5 years (e.g. beyond the current fiscal year's budget cycle and out to 5 years).

With the next 2-5 years, STAR's focus will be on the Phase-I of the program i.e. the Heavy Flavor and Di-leptons measurement (and the study of sQPG properties). The detector upgrades and making the challenging datasets (especially those taken by the HFT) a success with certainly be our very first priority.

1.7.2 Software Infrastructure

Describe the proposed software infrastructure and tools as they will be in the next 2-5 years (e.g. beyond the current fiscal year's budget cycle and out to 5 years). Note any products you are test driving, or major revisions expected for the current generation of tools.

No major change of the software infrastructure is seen for what concerns network requirements. STAR computing will however go through dramatic changes and upgrades including (a) the onset of a new track reconstruction software (b) a new Meta-Data collection facility online (based on the *Advanced Message Queuing Protocol* or AMQP) which will completely replace the old system (direct MySQL access) will be in effect in 2014 (c) a strong push toward moving computational resources closer to the experimental device (HLT track reconstruction and vertexing).

Enhancement of our STAR FileCatalog will be needed to support increase operations as well as data accumulation – spanning over more than a decade of data taking, advanced queries for comparative identification of dataset will be needed. We have also not consistently catalogued the embedding datasets, essentially relying on the records of our simulation and embedding request tracker. This has caused some issues related to the fast identification and location of possible viable past embedding processing. This is an organizational item only and in the past year, the workflows have more consistently brought the data samples back to BNL where

they are Catalogued by the local workforce (automation should be in place by next year).

We have concerns as per the rapid evolution of the computing landscape especially in the many-core dimension. The mix of architecture is inevitable and the use of Xeon/Phi' like architecture of general interest for STAR's online HLT program. We have been grateful for the help of Intel in this matter, providing free resources and expertise for evaluating the possible usage of the Xeon/Phi in STAR.

1.7.3 Process of Science

Describe how the process of science will change over the next 2-5 years.

We do not see a dramatic change in our process of science within this timeframe. There will be hidden changes not directly relevant to network requirements apart from HLT based vertexing, needed to reduce the size of the massive dataset forecasted in 2016 by better selecting the events of interests.

We have not dared to proceed (yet) with application of data reduction algorithm at the source – not recording hits which would not be considered for tracking has its data size advantages (and may impact data set sizes by reduction factors of ~ 40%) but are not done without risks: the drop of hits is irreversible and more studies would be needed before considering such high risk path. This may come to a natural development however as more and more computing power is moved online for High Level Trigger purposes and early event transformation will be possible.

Focus on real-time decision making filters (HLT, pattern recognition) as well as data reduction and repacking methods (fast online tracking, pile-up rejection at the source for data reduction) and even moving detector calibration processes closer to the data taking so real-time first pass track reconstruction in HLT and collision vertex reconstruction could lead to better decision making in regards of the selection of the collision event holding the highest potentials for key physics measurements are all likely activities and development for the next 2-5 years.

1.8 Remote Science Drivers – the next 2-5 years

1.8.1 Instruments and Facilities

Describe the use of remote instruments and facilities as they will be in the next 2-5 years (e.g. beyond the current fiscal year's budget cycle and out to 5 years).

We have already described our upgrade plans and schedule as well as our main facilities. We compile the network requirements in Table 9. Most network bandwidths summarized are calculated as the maximum of diverse previous requirements showed in Table 7 and Table 8 as the diverse transfers are not

continuous along the year (some are burst transfers, some for 6 months length and other second order effects).

Table 9: Summary tables for all network requirements

	WAN totals, by Tier						
	2013	2014	2015	2016	2017	2018	2019
SACs and Tier 2 centers (need for any / each)	0.48	0.89	0.98	1.90	1.26	0.42	1.11
Tier 1 center, MuDST (and embedding support)	1.12	2.07	2.29	4.44	2.93	0.99	2.58
Total WAN for 50% offsite processing [1/2 pass "as we go"] (Gb/sec)	2.61	3.74	4.03	8.02	5.26	1.9	5.17
[A] Tier 0 center, general support (Gb/sec)	1.60	2.37	2.29	4.44	2.93	0.99	2.58
[B] Tier 0 center, general support + 1/2 pass offsite (Gb/sec)	2.61	3.74	4.03	8.02	5.26	1.90	5.17
[C] Tier 0 center, general support (Gb/sec) + 1/2 pass offsite + complementary 1/2 saving at Tier 1 a year later	3.59	4.70	5.25	9.31	7.88	3.61	5.81

The key essential components will be:

- Each SAC will need networking at a capacity < 2 Gb/sec for the time period envisioned as showed on row 1.
- To sustain operation of NERSC/PDSF, network rates of ~ 3 Gb/sec will be needed for this period on the NERSC side for STAR usage – this is showed in row 2 (we again purposely ignore the 2016 estimate with caution; 4 Gb/sec would though be workable in any scenari).
- A faster pace use of the KISTI facility and strong push toward real-data processing to the extent possible – we will plan for a facility growth able to consume data transfers up to a 50% level; the required bandwidth as a function of years has been showed in Table 8. Those rates are showed on row 3 and will unlikely exceed the 5-6 Gb/sec rate.
- To sustain both operations, BNL connectivity will need to be provided at levels consistent with scenario [B] (row 5).
- Depending on how critical data preservation to another site is (and to the extent possible), the required bandwidth from BNL would be as showed on row 6, scenario [C].

1.8.2 Software Infrastructure

Describe the proposed software infrastructure and tools as they will be in the next 2-5 years (e.g. beyond the current fiscal year's budget cycle and out to 5 years). Note any products you are test driving, or major revisions expected for the current generation of tools.

The only fundamental changes we see within this period are the possible exploitation of hybrid Cloud/Grid infrastructure on two fronts:

- The online computational resources especially are in the process of being "Cloudified" and would be used for additional processing on-site)
- With a 2 years' timeframe, it is highly probably that operations at KISTI will be carried on a Cloud basis

Those changes will not alter our current network requirements.

1.8.3 Process of Science

Describe how the process of science will change over the next 2-5 years, including remote data sources, remote access, remote facility operation, expanded collaboration, etc.

We do not see an expansion of our collaboration at this stage but have strong evidence of a sustainable body of workforce over this period. Keeping our collaborators active in the STAR operations is another story as there has been no official check-and-balance of dues/consequences but in one process: shift dues. At the time of the LHC where team tends to over-commit and spread thin, policies are needed in STAR to address this issue and get a clear picture of commitments. This is to be addressed in the coming year.

The decrease of SAC to a minimal 3 will be derived from the building of massive compute resources at major centers compounded with the extreme size of datasets use in STAR.

To date, STAR has not yet leveraged the use of a global Xrootd namespace (and global redirector), the networking available as well as the prioritization of where to access the data appearing to be insufficient to us. However, the maturing of our understanding in global data movement planning and scheduling may change this view within this timeframe.

Another change may be the move of our NERSC/PDSF operation to a mainframe machine such as the "carver" system (an IBM iDataPlex system). Early tests by our users have showed this path is feasible. The phasing out of facilities such as the one as NERSC/PDSF for the benefit of a "Carver-like" mainframe operation is likely (from an experimental standpoint, performance, support and reliability are the only relevant factors). Possibly, Cloud based approach could also be used for sharing resources (a "Virtual PDSF") in a more elastic manner.

1.9 Beyond 5 years – future needs and scientific direction

Describe future plans for compute, storage, software, and network capabilities, any connections to any major scientific instruments (new or existing) that are coming in the next 5+ years, facility upgrades, or other changes.

Our upgrades will continue to move forward with, by 2018, the incoming of the iTPC. We noted in section 1.4 the event size growth this upgrade will cause. Though, the run plan envisioned a less stressful data demand as the species planned in 2018

drives the smallest event sizes amongst all species followed will compensate for the initial increase. The 2019 runs, which may be altered depending on capabilities available by then, will see the full impact of this event size impact by setting the program on more challenging (size wise) species. The order is logically considered with resource growth and detector capabilities in mind. We expect KISTI to remain part of STAR success and the MoU extended to the beginning of the eSTAR era (2020).

As the transition from RHIC to eRHIC (eSTAR) will occur, there will be a shift in the collaboration demography if EIC ought to be approved as the next generation of experiments. We cannot fully forecast the nature of this shift at this time (the groups interested in the EIC science and the HI science are not all overlapping). The data driven by the future experiments will be highly dependent on the final detector designs (in progress but not sealed with a stamp of approval).

Before those times, the frameworks will need to be drastically refactored (or new frameworks designed) leveraging the reality of vectorization, parallelism at the compute core level, asynchronous IO operations and MQ-like communications expanded. Agile data structures and representations will need to be folded into those frameworks and today's capabilities of systems such as STAR's AMQP system perhaps extended to real data.

1.10 Network and data architecture

Please describe any specific items of interest in regard to high-performance data transfers, network architecture (e.g. a Science DMZ –http://fasterdata.es.net/science-dmz/) or other site, campus, or facility networking issues. This also includes any interaction with Big Data requirements or initiatives (e.g. the Federal Big Data initiative). Indicate if there are ways in which changes in network architecture or performance could significantly improve your pursuit of science.

Looking at the steady demographic of STAR, better connectivity to Asia is critical to STAR science. While bandwidth to KISTI has improved (and ESnet has helped in the past success story), the connection to China remains a liability to science (the connections are two slow and intermittent to carry a decent remote work). Closer collaborations, needed at the international level, are likely to boost US science for any collaboration having a 1/3rd of its institutions in this region of the globe as it is the case for STAR.

In the interim, the use of remote persistent session and tools such as "NX" (Desktop Virtualization and Remote Access Management) has been reported by our remote colleagues in those regions to be of a dire help and convenience.

1.11 Collaboration tools

Please discuss any current or planned use of videoconferencing services such as Skype, ESnet's ECS service, or other similar services. Also please include any telephone/audio conferencing needs, or other collaboration services that are used or needed.

Our needs for collaborative tools and video services include the need for standard phone bridge and video conferencing capabilities (slide display essentially).

The RHIC collaborations have maintained a paid subscription to the SeeVogh Research Network (SRN), the successor of EVO services somewhat dropped by the HEP community. This service has no real match at this point in time and equivalent services from CISCO have been showed to be cost prohibitive (the cost of the Vidyo service not attractive comparing to SRN, we barely understand the rationales behind the EVO decision).

Other services (for "registered" users, such as ESnet ECS) did not appear to be adequate for our user's community essentially composed of remote collaborators holding no official hire at any national laboratories. While we have been frequently told the service is opened for all of our users as "RHIC users" (employees or guests), when prospected, it appeared great confusions prevented us to even consider this service (any request from a collaborator from a sensitive country would make the registration raise red flag and no go anywhere). A full understanding of the demography of our community is needed before solutions are selected or pushed forward.

Skype is still in use for daily communication amongst members. Skype calls are also frequent while abroad (the cost of a communication plan from any US based telephone service being outrageously out of proportions in comparison to VoIP alternatives).

1.12 Data, Workflow, Middleware Tools and Services

Most scientific disciplines are experiencing significant data growth. Please describe the ways in which data growth, workflow changes, and so forth might impact the science described in the case study, and how networks, workflow tools, and so forth might help. If it would be useful for ESnet personnel to work with you to evolve or enhance your workflow, please let us know.

Tools examples include Globus Online or other data transfer tools, automated data transfer toolkits, distributed data management tools, etc.

Please also include planned use of emerging services such as commercial cloud computing, storage, etc.

There is no doubt in our minds that the availability of predictive and/or advanced network reservation capabilities would be of a benefit for planners and data movement schedulers. While we have tackled all other dimensions (network availability, bandwidth, cache, streams) we have not been able to study (lack of

capabilities and generalization) how those features could help efficiently sharing bandwidth between multiple consumers.

1.13 Outstanding Issues (if any)

Please identify any outstanding issues involving the network or the use of the network.

Please also describe the ways in which the experiment or collaboration workflow is likely to change (or ways in which there is desire to change)

While data preservation as well as data sharing (open access) have been a noticeable push by DOE (multiple surveys were made in 2012-2013 to gather feedback on our plans), we do not understand if additional source of funding would appear to support and/or sustain such efforts. If no additional funds appear, forcing this issue will ultimately have to be addressed with existing workforce and the impact on scientific deliverables direly affected at most sensitive time for RHIC and US science. Making multiple PBytes of data accessible to the world will not be resolved by asking more of the same questions through more surveys – of a sizeable and non-trivial problem, possible further data reduction and scalable data re-distribution would need to be supported via R&D and minimally, time and level of efforts. Support for long term archiving of our data (a copy at NERSC for example) will require additional secured funding we have not identified nor heard off; the problem size however implies starting as soon as possibly achievable (bandwidth required if we start late in the game would make such a plan impossible to achieve). Those are important questions we wish to hear back from DOE.

The slow adoption of Cloud computing (even at the conceptual level) may be the sole issue we see in the US based distributed computing consortiums. It appeared at times that even using the word “Cloud” became taboo for the benefits of securing not only the current production environment but tools and home-grown technologies. While we partly understand the motivations, this approach has been, in our view, detrimental to scientific progress and innovation – a balance needs to be achieved. There are signs this may change within the next 2 years and a collective program may see its life but planning for Cloud based resources (from the OSG for example) within a 2-5 years’ timeframe is now uncertain.

We do not have other outstanding issues at this time but would like to note in passing and acknowledge the tremendous benefit of something as simple as the OSG support center in reporting problems to us (as per our grid infrastructure) and facilitating communications between teams through a much improved and enhanced ticket system. Operational support is often the forgotten child of science but the impact of our ability (at low effort level) to carry forward a Grid program is undeniable. We would also like to note that the transition to the new OSG CA has been more than smooth – a great job overall and also a much improved process for acquiring a certificate via OIM (OSG Information Management).

1.14 Summary Table

Note well – the table asks for several different types of information (e.g. the size of data sets and the time required to transfer them) separately. Please try to avoid saying "I need 10Gbps" and provide real-world data set information if at all possible.

Key Science Drivers			**Anticipated Network Needs**	
Science Instruments, Software, and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time needed**	**WAN Transfer Time needed**
Near Term (0-2 years)				
• RHIC/STAR data taking of large samples with the HFT and MTD upgrades – Heavy Flavor and Di-leptons measurement to study the properties of the sQG • Online/HLT, Xeon/Phi based seed finding and vertexing proof of principle • MQ based Meta-Data collection (online) • New high precision track reconstruction software offline, same framework • IO re-read ahead enabled	• Data flows for data productions moves MuDST to Xrootd (BNL) • Transfer of MuDST to NERSC/PDSF + partial transfers to SAC • Embedding simulations at NERSC/PDSF and KISTI • OSG use for simulations and library validations • Possible ½ pass data reco at KISTI (Grid or Cloud model) • Transfer of datasets off Tier-0 for long germ permanent archival storage a possibility	• 3-3.5 PB RAW & 2-3 PB MuDST • 2 PBytes candidate for transfer • 500-600 k files • Files size average are fixed to 4 GB • Marginal data transfer load from embedding • 1.5 PB to KISTI and 1 PB from	• RAW transferred as produced (during runtime) • Distribute disk population as produced (8-10 month periods) • > 1 Gb connection of farm's compute nodes • SAC need < 2 Gb • NERSC/PDSF ~ 3 Gb • KISTI 3-4 Gb • BNL WAN pipe @ 4-5 Gb as baseline, possibly 5-6 Gb for RAW data transfer to secondary location	• MuDST transfer as we go • Embedding as fast possible • Remote production: provider / consumer • MuDST movement from BNL to NERSC/PDSF (marginal DAQ) • RAW data from BNL to KISTI • Data from KISTI to BNL (embedding and MuDST) • Data from PDSF to BNL (embedding) • Data from NERSC or BNL to SAC (un-identified link)
2-5 years				
• End oh Phase-I	• Data flows remain	• Uncertainty	• Similar	• Similar time

physics program, beginning of phase II program • RHIC/STAR data taking of large samples – BES Phase II, QCD critical point and study the QCD phase structure • Online HLT event vertexing, possible event filtering and reduction • iTPC in 2018, eCooling • Onset of "lego-block" processing (workflows seamlessly running online / offline for calibration – adapters, MQ based IO) • Cloudified cluster online a "standard" + full use of KISTI	near identical, • Elastic computing at the heart (Cloudified online resources may "join" a pool for embedding + KISTI) • Possible move of PDSF to "Carver-like" platforms • Situations for transfer of datasets off Tier-0 clarified	in 2016 data size • Overall similar datasets up to 2019	bandwidth needs to/from the same end-points • SAC profile unknown (changes certain within 2 years / will need re-assessing) • KISTI connectivity @ 5-6 Gb • BNL with a 5+ Gb pipe (data archiving plan influence)	frames and peers • Possible reshape of the SAC landscape • Possible use of opportunistic cloud resources (at lower levels) – OSG/Cloud?
5+ years				
• Heavy Flavor program and B-physics + eSTAR by 2024 • EIC long term vision should crystalize within this timeframe • Lego-block frameworks with aync IO + filter / repack capabilities (MQ framework-like) likely	• Similar landscape foreseen • Predictions beyond 2020 unclear	• Expecting similar datasets	• No changes forecasted	• Peering is unclear but likely the same until 2020

[end of case study – see FAQ pointer and notes below]

1.15 Notes

There is a FAQ and other supporting information on the ESnet web site. The base page for the network requirements information is http://www.es.net/requirements/

Please include "the other end" of network connections as best you can when describing wide area collaborations or data transfers. We realize that in many cases sites have many remote users, and that each of these many hundred or thousand (or more) remote users represents a unique endpoint – that's fine. However, any commonality in user traffic flows that you can describe allows us to do end-to-end planning, performance tuning, etc. This input is very helpful.

Please include a discussion of the data transfer, middleware and workflow tools that you use. Please describe how these tools are used by scientists, any outstanding issues that exist, future plans, etc.

Note – by "process of science" we mean the way in which scientists use the instruments, facilities, supercomputer centers, etc. for knowledge discovery. This is especially important if the way in which the science is conducted produces network usage patterns that are not obvious when looking at the other information available.

We will use Mbps, Gbps, etc. to describe megabits per second and gigabits per second, etc. We will use MB/sec, GB/sec, etc. to describe megabytes per second, gigabytes per second, etc. We have found the difference in notation to be helpful for disambiguation and for catching typographical errors.

If you have needs for ESnet's ECS (ESnet Collaboration Services - audio and video conferencing), please describe those needs, including projections for usage growth or other coming changes.

Also, if you are having difficulties using the network, please describe them. In particular, the following information is helpful:

- Data set sizes that can be moved vs. data set sizes that cannot be moved (this helps indicate the scale of the problem)
- Users that have difficulty vs. users that do not (the same goes for institutions – this might point to specific issues that could be resolved, e.g. packet loss when trying to move data to a particular site)
- Tools, resources, etc. that might help you use the network more efficiently