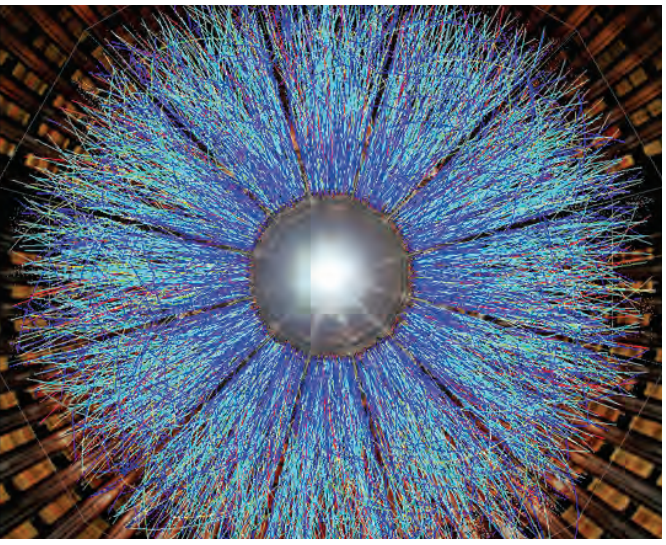
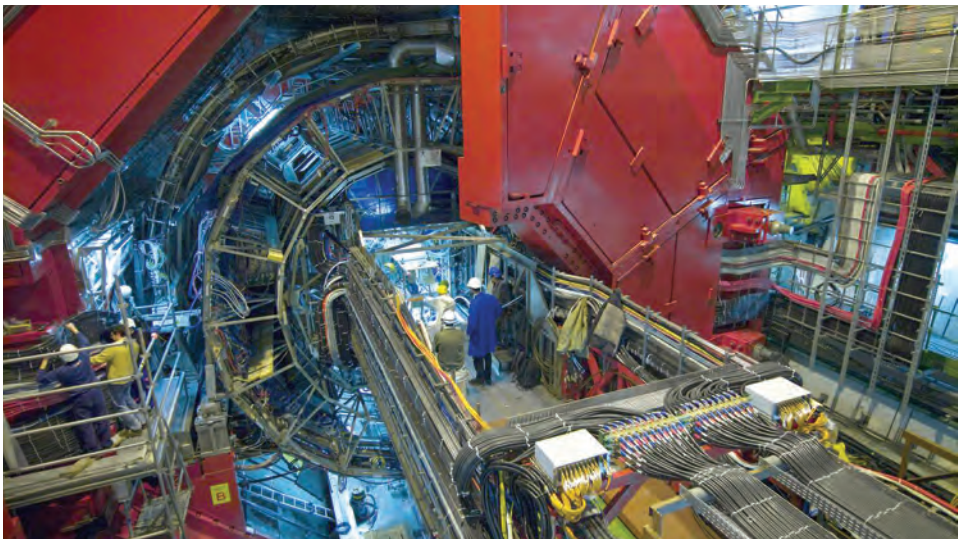
NP Science Network Requirements

Report of the Nuclear Physics
Network Requirements Workshop

Conducted August 25 and 26, 2011

ESnet
Energy Sciences Network





BERKELEY LAB
Lawrence Berkeley
National Laboratory

U.S. DEPARTMENT OF
ENERGY
Office of Science

DISCLAIMER

Nuclear Physics Network Requirements

Office of Nuclear Physics, DOE Office of Science
Energy Sciences Network
Gaithersburg, Maryland — August 25 and 26, 2011

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research. Vince Dattoria is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S. Department of Energy under contract.

DE-AC02-05CH11231.

This work was supported by the Directors of the Office of Science, Office of Advanced Scientific Computing Research, Facilities Division, and the Office of Nuclear Physics.

This is LBNL report LBNL-XXXX

Participants and Contributors

Richard Carlson, DOE/SC/ASCR (Network Research)

Steve Cotter, ESnet (Networking)

Eli Dart, ESnet (Networking)

Vince Dattoria, DOE/SC/ASCR (ESnet Program Manager)

Michael Ernst, BNL (RHIC)

Bill Johnston, ESnet (Networking)

Andy Kowalski, JLAB (Networking)

Jerome Lauret, BNL (STAR at RHIC)

Charles Maguire, Vanderbilt (CMS Heavy Ion at LHC)

Thomas Ndousse-Fetter, DOE/SC/ASCR (Network Research)

Jeff Porter, LBNL (ALICE at LHC)

Martin Purschke, BNL (PHENIX at RHIC)

Gulshan Rai, DOE/SC (NP Program Office)

Lauren Rotman, ESnet (Networking)

Ron Soltz, LLNL (ALICE at LHC)

Brian Tierney, ESnet (Networking)

Domenico Vicinanza, DANTE (Networking)

Chip Watson, JLAB (CEBAF)

Jason Zurawski, Internet2 (Networking)

Editors

Eli Dart, ESnet — dart@es.net

Lauren Rotman, ESnet — lauren@es.net

Brian Tierney, ESnet — bltierney@es.net

Table of Contents

1 Executive Summary

The Energy Sciences Network (ESnet) is the primary provider of network connectivity for the U.S. Department of Energy (DOE) Office of Science (SC), the single largest supporter of basic research in the physical sciences in the United States. To support SC programs, ESnet regularly updates and refreshes its understanding of the networking requirements of the instruments, facilities, scientists, and science programs it serves. This focus has helped ESnet to be a highly successful enabler of scientific discovery for over 20 years.

In August 2011, ESnet and the Office of Nuclear Physics (NP), of the DOE SC, organized a workshop to characterize the networking requirements of the programs funded by NP.

The requirements identified at the workshop are summarized in the Findings section, and are described in more detail in the body of the report.

2 Findings

2.1 General Findings

- Several NP experiments are international in scope, requiring coordinated support by networks in the United States, Europe, and Asia. A desire was expressed at the workshop for a coordinated strategic planning effort among all the networks that support an experiment in order to coordinate capacity and services planning. At a minimum, this would include ESnet, GEANT, and Internet2, though including other networks (e.g., in Asia) would be beneficial. In addition to network capacity and service planning, networks and the experiments need to work with the application programmers for the experiment software stack so that network awareness can be more effectively integrated into data mobility, job scheduling, and other tools.
- Several NP collaborations use ESnet's audio and videoconferencing services (ESnet Collaboration Services, or ECS), and additional collaborations expressed interest at the workshop. ESnet is expanding the service offerings and documentation to facilitate increased use of the service.
- A number of NP experiments are dependent on the Open Science Grid (OSG) software stack, and there is significant concern about OSG's long-term plans and support.
- A desire was expressed for a structure within DOE/SC that encourages strategic planning between ESnet and the major science experiments and research programs.
- Based on the discussions at the workshop and on the case studies presented, the networking requirements of the NP Program Office that are in-scope for ESnet's mission appear to be covered under ESnet's current budget and service projections.

2.2 Findings Specific to Particular Sites or Experiments

- CMS Heavy Ion runs occur one month per year, typically in November. The data are transferred first from CERN to FNAL, and then from FNAL to Vanderbilt. This results in heavy use of the network between FNAL and Vanderbilt for a few weeks in the fall. The path from FNAL to Vanderbilt is currently via ESnet's connection to the Southern Crossroads (SoX) R&E exchange point in Atlanta. The SoX exchange interfaces should be monitored for congestion.
- perfSONAR has been very useful to Vanderbilt in tracking down and solving performance problems.
- Jefferson Lab networking needs may increase in 2015 after an upgrade to its experimental facilities.
- There is an opportunity to explore diverse connectivity to JLAB.
- The ALICE experiment would like to improve its understanding of network topology, performance analysis, and how these interact. Improved WAN performance predictability/stability would make the network a more productive tool for the experiment. ALICE currently uses MonALISA for network performance monitoring, rather than perfSONAR. ALICE attendees were interested in perfSONAR.

- The ALICE experiment has experienced some performance issues when exchanging data with collaborators at UNAM in Mexico.
- The STAR collaboration has implemented a workflow that replaces DSTs with micro-DSTs, reducing data-set size by a factor of five, thus reducing their WAN bandwidth requirements.
- The STAR collaboration is looking into the Cloud Services model for data processing. The collaboration ran some experiments on the DOE Magellan cluster as well as commercial and university based clouds that showed the benefits of this approach. The use of commercial cloud services could have an impact on ESnet commercial peerings.
- Both the DOE/SC HEP and NP program offices fund LHC experiments. Given that the LHC experiments rely heavily on international (e.g. trans-Atlantic) networking, it might be beneficial to integrate the HEP and NP network requirements planning efforts. This would allow LHC experiments funded by the HEP and NP program offices to engage in more closely coordinated planning for international networking capabilities, capacity, and services.

3 Action Items

The following action items have been identified as a result of the workshop.

- The findings of the workshop will influence the capacity and service planning for current and future ESnet infrastructure.
- ESnet, along with ASCR and NP program management, will explore closer integration of the HEP and NP network requirements workshops.
- ESnet will explore options for diverse connectivity to JLAB.

Based on the discussions at the workshop and on the case studies presented, the networking requirements of the NP Program Office that are in-scope for ESnet's mission appear to be covered under ESnet's budget and service projections. ESnet will continue its interactions with the NP program, NP-funded experiments, and NP facilities such that ESnet's understanding of NP networking requirements stays current.

4 Workshop Background and Structure

The strategic approach of the Office of Advanced Scientific Computing Research (ASCR — ESnet is funded by the ASCR Facilities Division) and ESnet to define and accomplish ESnet's mission involves three areas:

1) Working with the Office of Science (SC) community to identify the networking implication of the instruments, supercomputers, and the evolving process of how science is done
2) Developing an approach to building a network environment to enable the distributed aspects of SC science and to continuously reassess and update the approach as new requirements become clear
3) Anticipating future network capabilities to meet future science requirements with an active program of R&D and advanced development

Addressing point (1), the requirements of the SC science programs are determined by:

a) Exploring the plans and processes of major stakeholders, including data characteristics of scientific instruments and facilities; anticipating what data will be generated by instruments and supercomputers coming online over the next 5-10 years; and examining the future process of science: how and where will the new data be analyzed and used, and how the process of doing science will change over the next 5-10 years
b) Observing current and historical network traffic patterns and determining how trends in network patterns predict future network needs

The primary mechanism for accomplishing (a) is the SC Network Requirements Workshops, sponsored by ASCR and organized by the SC Program Offices. SC conducts two requirements workshops per year, in a cycle that repeats every three years:

- Basic Energy Sciences (materials sciences, chemistry, geosciences) (2007, 2010)
- Biological and Environmental Research (2007, 2010)
- Nuclear Physics (2008, 2011)
- Fusion Energy Science (2008)
- Advanced Scientific Computing Research (2009)
- High Energy Physics (2009)

The workshop reports are published at http://www.es.net/requirements/.

The requirements workshops also ensure that ESnet and ASCR have a common understanding of the issues that face ESnet and the solutions that ESnet undertakes.

In August 2011, ESnet and the DOE SC Office of Nuclear Physics (NP) held a workshop to characterize the networking requirements of NP-funded programs.

Workshop participants codified their requirements in a case-study format that included a network-centric narrative describing the science, the instruments and facilities currently used or anticipated for future programs, the network services needed, and the way the network is used. Participants considered three timescales in their case studies — the near term (immediately and up to 12 months in the future), the medium term (two to five years in the future), and the long term (more than five years in the future). The information in each

narrative was distilled into a summary table, with rows for each timescale and columns for network bandwidth and services requirements. The case-study documents are included in this report.

5 Office of Nuclear Physics

5.1 Introduction

Nuclear science began by studying the structure and properties of atomic nuclei as assemblages of protons and neutrons. Research focused on nuclear reactions, the nature of radioactivity, and the synthesis of new isotopes and new elements heavier than uranium. Today, the reach of nuclear science extends from the quarks and gluons that form the substructure of protons and neutrons, once viewed as elementary particles, to the most dramatic of cosmic events—supernovae. At its heart, nuclear physics attempts to understand the composition, structure, properties of atomic nuclei, discover new forms of nuclear matter, including that of the early universe, measure the quark structure of the proton and neutron, and study the mysterious and important neutrino. Rapid advances in large-scale integration electronics, computing, and superconducting technologies have enabled the construction of powerful accelerator, detector, and computing facilities. These provide the experimental and theoretical means to investigate nuclear systems ranging from tiny nucleons to stars and supernovae. Nuclear physics also supports the production, distribution, and development of production techniques for radioactive and stable isotopes that are in short supply and critical to the Nation.

The DOE Nuclear Physics program provides most of the Federal support for nuclear physics research in the U.S. About 1,595 scientists, including 880 graduate students and postdoctoral research associates, receive support from NP. In addition, the program supports three national scientific user facilities.

Other agencies use nuclear physics facilities for their own research. Notable is the use by semiconductor manufacturers to develop and test radiation hardened components for earth satellites to be able to withstand cosmic ray bombardment and by NASA's Space Radiation Laboratory (NSRL) established at Brookhaven Laboratory's Relativistic Heavy Ion Collider (RHIC) Facility to study the radiobiological effects using beams that simulate the cosmic rays found in space.

The DOE Nuclear Physics program helps the U.S. maintain a leading role in nuclear physics research, which has been central to the development of various technologies, including nuclear energy, nuclear medicine, space exploration and the nuclear stockpile. The produces highly trained scientists who contribute to the effort aimed at ensuring that DOE and the Nation have a sustained pipeline of highly skilled and diverse science, technology, engineering, and mathematics (STEM) workers which is knowledgeable in nuclear science.

5.2 Major Facilities

At the largest scale, the NP program supports two unique facilities. The Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory is a world-class scientific research facility used by almost 1,200 physicists from around the world to study what the universe may have looked like in the first few moments after its creation. By colliding heavy nuclei together at nearly the speed of light, RHIC will, for a fleeting instant, heat the matter in collision to more

than a billion times the temperature of the sun. In so doing, scientists are able to study the fundamental properties of the basic building blocks of matter, as well as learn how they behaved collectively some 15 to 20 billion years ago, when the universe was barely a split-second old. What physicists learn from these sub-atomic collisions may help us understand more about why the physical world works the way it does, from the smallest subatomic particles to the largest stars.

The Thomas Jefferson National Accelerator Facility (TJNAF), commonly known as JLAB, is devoted to nuclear physics research. Approximately 1,390 scientists from around the world use TJNAF's Continuous Electron Beam Accelerator Facility (CEBAF) — the first large-scale application of superconducting electron-accelerating technology — to conduct unique world-class nuclear physics experiments. Using high-energy electron beams from the accelerator, experimenters probe the sub-nuclear realm, revealing how quarks make up protons, neutrons and the nucleus itself. Partnering with industry, universities and defense agencies, Jefferson Laboratory also pursues applied research with its free-electron laser and medical imaging programs. TJNAF is in the process of upgrading CEBAF, which will double the electron beam energy.

5.3 **Other Facilities**

What is the origin of the elements, how do stars evolve, and what is the source of high- energy cosmic rays and cosmic gamma rays? The NP program's Low Energy subprogram studies nuclei at the limits of stability, nuclear astrophysics reactions, the nature of neutrinos, and fundamental symmetry properties in nuclear systems. Measurements of nuclear structure and nuclear reactions are carried out primarily at the Argonne Tandem Linac Accelerator System (ATLAS) at Argonne National Laboratory (ANL). Measurements of symmetry properties, particularly of the neutron, are being developed by nuclear physicists at the Spallation Neutron Source (SNS) at ORNL. The Lawrence Berkeley National Laboratory's 88-Inch Cyclotron is being supported to test electronic circuit components for radiation "hardness" to cosmic rays by the National Reconnaissance Office (NRO) and U.S. Air Force (USAF), and for a small in-house nuclear physics research program by the NP program.

University-based research is an important component of the NP Low Energy subprogram. Accelerator operations are supported at two university Centers of Excellence - the Cyclotron Institute at Texas A&M University (TAMU) and the HIGS facility at the Triangle Universities Nuclear Laboratory (TUNL) at Duke University. At the University of Washington, infrastructure is supported to develop scientific instrumentation projects and provide technical and engineering training opportunities.

The future Facility for Rare Isotope Beams (FRIB) at Michigan State University (MSU) is a next-generation machine under development that will advance the understanding of rare nuclear isotopes and the evolution of the cosmos by providing beams of rare isotopes with neutron and proton numbers far from those of stable nuclei in order to test the limits of nuclear existence and models of stellar evolution.

5.4 International Collaborations

The NP program's RHIC and CEBAF facilities attract significant experimental and theory research collaborations from all over the world. Over half the users of NP facilities are from abroad. Scientists from the United States also participate in leading edge scientific experiments abroad. A U.S. national laboratory and university collaboration is participating in the Italian-lead Cryogenic Underground Observatory for Rare Events (CUORE) experiment at the Gran Sasso Laboratory, contributing to the fabrication of the detector which is planned to take data in approximately FY 2014. This experiment will search for evidence that the neutrino is its own antiparticle. In FY 2007, a U.S. university collaboration began limited but crucial participation in the German-lead KArlsruhe TRitium Neutrino (KATRIN) experiment to determine kinematically the mass of the electron neutrino by measuring the beta decay spectrum of tritium. This experiment is expected to become operational in approximately 2014. Building upon the discoveries at the RHIC, a modest U.S. nuclear physics research effort is underway in the ALICE (A Large Ion Collider Experiment) and CMS (Compact Muon Solenoid) experiments at the Large Hadron Collider (LHC) at CERN in Switzerland.

6 Thomas Jefferson National Accelerator Facility

6.1 Background

Thomas Jefferson National Accelerator Facility (Jefferson Lab) is funded by the Office of Science (SC) for the U.S. Department of Energy (DOE). As a user facility for scientists worldwide, its primary mission is to conduct basic research on the atom's nucleus at the quark level.

With industry and university partners, Jefferson Lab has a derivative mission as well: applied research in free-electron lasers (FELs) based on accelerator technology developed at the Laboratory.

As a center for both basic and applied research, Jefferson Lab also reaches out to help educate the next generation in science and technology. The Laboratory is managed and operated for DOE by Jefferson Science Associates, LLC (JSA). JSA is a Southeastern Universities Research Association (SURA)/Computer Sciences Corporation limited liability corporation created specifically to manage and operate Jefferson Lab.

Jefferson Lab is a user facility offering capabilities that are unique worldwide for an international user community of 1,390 active users. One-third of all PhDs granted in nuclear physics in the United States are based on Jefferson Lab research (419 granted, 204 in progress).

6.2 Key Local Science Drivers

6.2.1 Instruments and Facilities

The Continuous Electron Beam Accelerator Facility (CEBAF) at Jefferson Lab provides a high-luminosity electron beam of up to 6 GeV to three halls. Hall B holds the CLAS (CEBAF Large Acceptance Spectrometer) detector, and Halls A and C hold a variety of spectrometers that can be configured to the needs of a particular experiment.

The superconducting radiofrequency (SRF) technology used in CEBAF has also enabled the development of the world's highest-average-power FEL. The FEL has achieved 10, 6.7, 14.2, and 2.2 kW at 10, 2.8, 1.6, and 1.0 microns, respectively, and will, after hardware upgrades, produce 1,000 watts in the ultraviolet range and >100 watts in the terahertz range. This instrument is being further developed, both to extend its capabilities and to exploit it for science.

Jefferson Lab is one of three sites (with Brookhaven National Laboratory [BNL] and Fermi National Accelerator Laboratory [FNAL]) hosting a distributed Lattice QCD (quantum chromodynamics) Computing Facility consisting of 10-100 teraflop/s class clusters tuned to the computing requirements of Lattice QCD (LQCD).

6.2.2 Process of Science

For the Experimental Nuclear Physics Program in the three halls, data is acquired in the counting house, monitored live, and transferred to the computer center to be written to tape in files of size up to 2 GB, typically up to 1 TB/day. Data analysis proceeds by staging a data file to cache disk to be analyzed in the batch farm. The batch system allows submission of meta-jobs that analyze large numbers of files corresponding to a single experiment and configuration.

Pass 1 analysis / reconstruction files of a size comparable to the raw files are written back to disk and to tape, and subsequent batch jobs produce smaller summary files. Most experiments only transfer the smaller files off site, although some experiments have copied all their data out for analysis at their home institutions.

Detector simulation is more distributed, with some work carried out at remote institutions and a larger fraction done at lower priority on the batch farm. Most simulation data is stored in the Jefferson Lab tape library.

The FEL program does not currently produce large amounts of data or networking traffic.

Large LQCD jobs are run at one of the DOE or National Science Foundation (NSF) supercomputing centers, producing space-time (quantum vacuum) configuration files. Typical configuration-generation job sizes are in the tens of thousands of cores. These files are then used as input into large numbers of analysis jobs at BNL, FNAL, and Jefferson Lab, with typical sizes up to 1,024 cores or up to 16 graphics processing units (GPUs). In aggregate, these analysis jobs consume even more computing power than the first stage (configuration generation). Propagator files generated from the configuration files at Jefferson Lab are currently in the few hundred MB to 20 GB range, and will grow larger as access to larger supercomputers allows for generating finer lattices.

6.3 Key Remote Science Drivers

6.3.1 Instruments and Facilities

Most of the experimental physics data is acquired and analyzed at Jefferson Lab and therefore the data-related wide area network (WAN) requirements are rather modest. Similarly, the FEL and LQCD programs do not yield significant WAN traffic other than bursts to move a small number of large files. Bursts of inbound traffic are probably correlated with transfers of LQCD files from supercomputing centers. (See network traffic graphs below.)

6.3.2 Process of Science

With a staff of about 750 and a user base of more than 1,300 researchers, there is considerable conventional use of networking at Jefferson Lab (i.e., other than for bulk data transfer), including e-mail, Web, and a growing use of videoconferencing. These tools are essential components in the many experiment collaborations at Jefferson Lab.

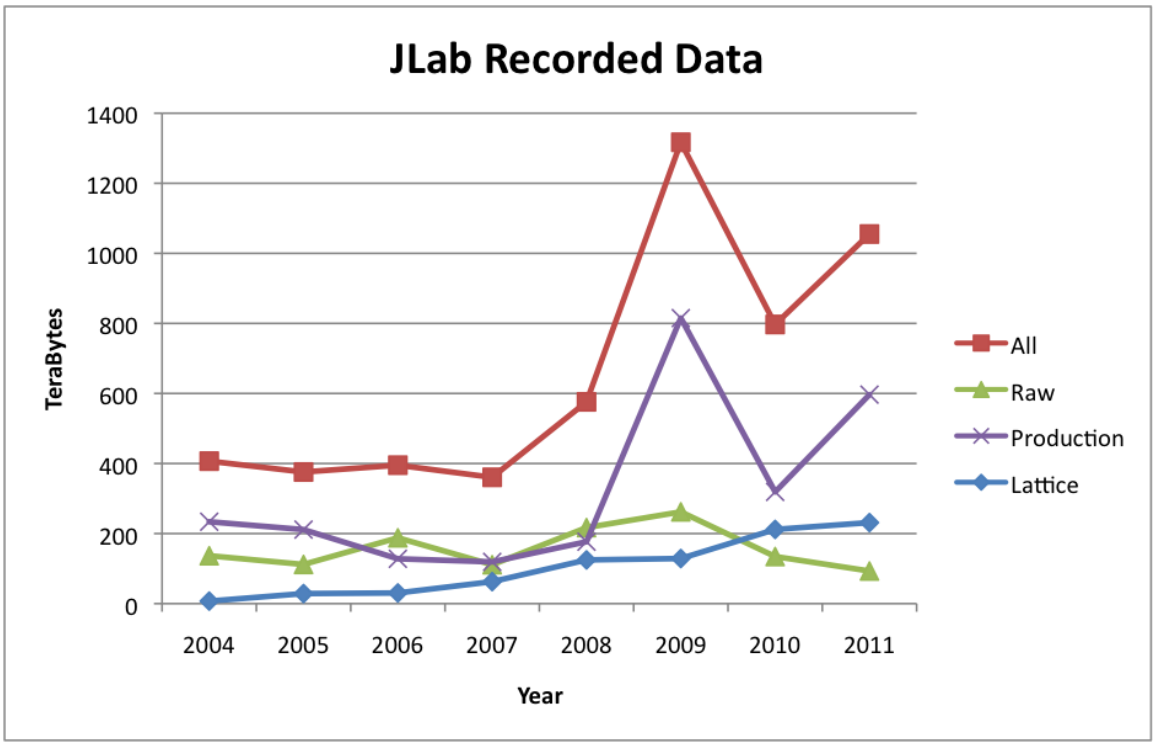


Figure 6-1. Data volume into the tape library; production refers to first-pass analysis.

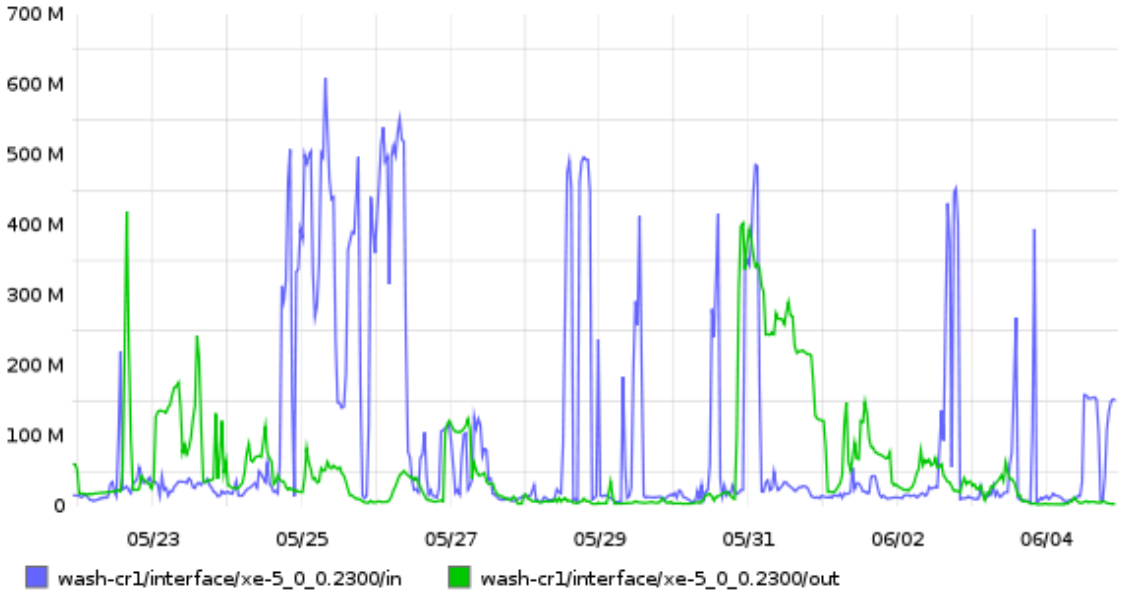


Figure 6-2. WAN traffic for a two-week period starting on May 22, 2011, 5 min average; max from JLAB — 600 Mbps (blue), max to JLAB — 425 Mbps (green).

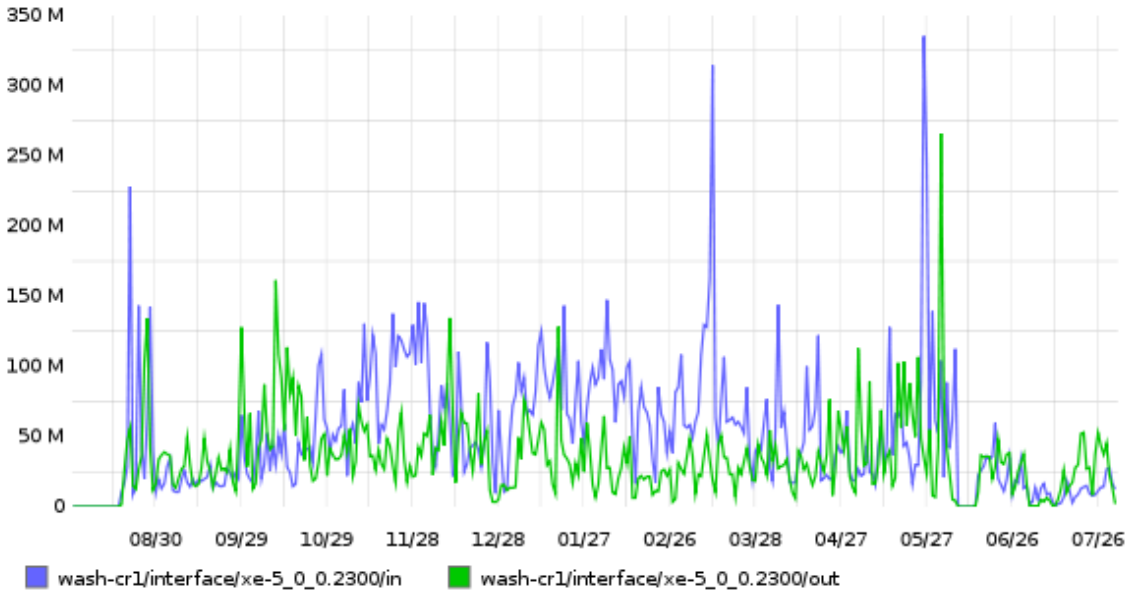


Figure 6-3. WAN traffic from Aug 1, 2010, 1 day average; max from JLAB — 325 Mbps; max to JLAB — 275 Mbps.

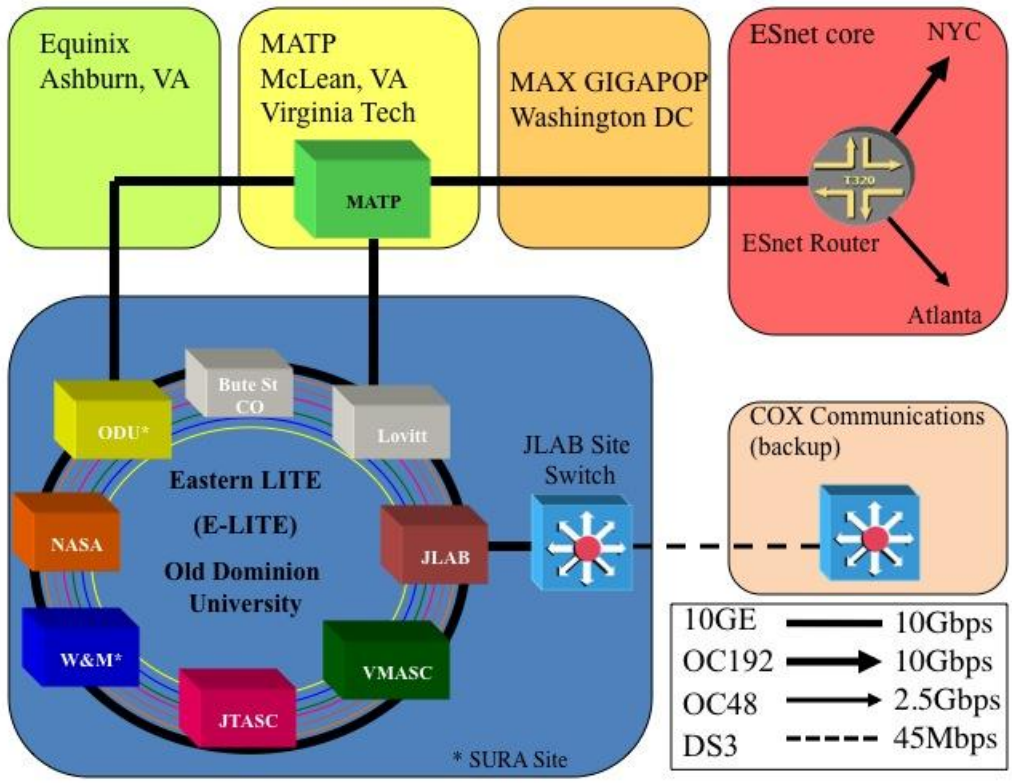


Figure 6-4. Jefferson Lab's current WAN connection via the E-LITE MAN.

Jefferson Lab has benefited from excellent partnerships and collaborations with ESnet, SURA, JSA, and local universities and research centers. With ESnet's knowledge and experience, these local partnerships made possible the Eastern LITE (Lightwave Internetworking Technology Enterprise) or E-LITE metropolitan area network (MAN). The multiwave 10 Gbps E-LITE network (Jefferson Lab's costs paid for by ESnet) provides access to the Virginia Optical Research Technology Exchange (VORTEX) gigaPOP sponsored by Old Dominion University (ODU) and located in Norfolk, Virginia. VORTEX provides access to MATP (Mid Atlantic Terascale Partnership), where ESnet has a presence. Jefferson Lab's membership in MATP was funded by SURA.

In 2011, E-LITE added an alternate 10 Gbps link to MATP services from ODU via Equinix in Ashburn, Virginia. The ESnet VLANs (virtual local area networks) to Jefferson Lab fail over to this alternate path if the primary VORTEX link goes down. Since ESnet has a presence at Equinix, alternate connectivity to ESnet should be provided at the Equinix collocation facility in the future. This would remove MATP as a single point of failure.

6.4 Local Science Drivers — The Next 2-5 Years

6.4.1 Instruments and Facilities

Jefferson Lab has embarked on a doubling of the CEBAF energy from 6 GeV to 12 GeV. Using space already available in the accelerator tunnels, 10 newer, high-performance cryomodules will be installed, and an additional magnet arc will be added to recirculate the beam for one final pass through the north linear accelerator (LINAC) to Hall D. The new experiment in Hall D will use the electron beam to produce a coherent bremsstrahlung beam. Hall D will house a solenoidal detector to carry out a program in gluonic spectroscopy to experimentally test current understanding of quark confinement. All three existing halls will be upgraded to receive the new five pass, 11 GeV beam. The additional experimental equipment proposed for Halls A, B, and C takes advantage of currently installed apparatus. The 12 GeV CEBAF Upgrade project is anticipated to be completed in FY 2015.

6.4.2 Process of Science

Trends in the 6 GeV program show Moore's law outpacing requirements for data analysis. Constant investments have yielded an increasing capacity for simulation.

Requirements for 12 GeV (2012+) will likewise be greater than for 6 GeV, but in terms of box count, the analysis cluster will be smaller than the current experimental physics cluster. Annual data volume for the 12 GeV program will be about 20 times the 6 GeV program, but still considerably less demanding than when the 6 GeV program began. Moore's law thus allows Jefferson Lab to continue a simple, cost-effective, lab-centric computing model.

Current 12 GeV computing plans show that Hall B (CLAS) will continue to be the largest simulation and data-generating hall, with Hall D fairly close, and Halls A and C much lower. The following spreadsheet contains summary numbers for computational and data volume requirements for each hall.

Table 6-1.Storage and computing requirements for Halls A, B, C, and D. Projects assume full running in 2015. Hall D requirements pre-operation (simulation) are being developed.

Year		2011	2012	2013	2014	2015	2016	2017
	Units							
Tape	TB/yr	(Actual)						
A		260	360	250	190	260	890	2770
B		720	800	??	??	5497	5500	5500
C		340	675	675	??	1120	1950	1950
D		??	??	??	??	8000	8000	8000
Total		1320	1835	925	190	14877	16340	18220
Work disk	TB/yr							
A		17	25	25	23	26	102	215
B		79	80	80	80	805	800	800
C		10	60	60	60	90	175	175
D		6	20	60	60	200	200	200
Total		112	185	225	223	1121	1277	1390
Cores	2011							
A		68	12	12	12	42	60	84
B		1228	1200	1200	??	1811	2000	2000
C		27	30	30	??	17	34	34
D		16	??	??	??	9000	9000	9000
Total		1339	1242	1242	12	10870	11094	11118

6.5 Remote Science Drivers — The Next 2-5 Years

6.5.1 Instruments and Facilities

A good estimate of Jefferson Lab WAN requirements would be that the requirements would scale like data volume. However, with a mainly central computing model with fairly modest requirements, this overestimates the networking requirements.

Since data rates will remain constant or decrease between now and the shutdown (2012), the current 10 Gbps WAN will be more than adequate for the next few years. In 2015, as the 12 GeV machine turns on, requirements might grow beyond 10 Gbps.

The LQCD Computing Facility should also grow only modestly in the next five years in terms of server count, and by roughly 10 times in performance by following Moore's law with nearly constant investments. However, LQCD will remain a modest contributor to WAN networking for the foreseeable future.

6.5.2 Process of Science

Use of distributed computing models (Web 2.0, grid, cloud, etc.) will continue to grow, even though the core of the computing model remains lab-centric. Conventional WAN usage, including videoconferencing, will steadily increase as these technologies become ever more widespread. It is difficult to quantify this growth in terms of network bandwidth and other capabilities.

6.6 Beyond Five Years — Future Needs and Scientific Direction

In addition to the 12 GeV program described above, Jefferson Lab is exploring other uses of its leadership SRF (superconducting RF) technology, which will likely lead to support for a number of SC accelerator projects at multiple locations (FRIB, ILC, Project X, SNS II, eRHIC, etc.) and could potentially lead to additional facilities on the campus such as an Electron Ion Collider (ELIC at Jefferson Lab) or a new 4^{th} Generation Light Source based on an FEL.

A light source at Jefferson Lab would necessitate much greater WAN bandwidth, as most light source users take their data home, and will expect to be able to do that over the network.

6.7 Middleware Tools and Services

Jefferson Lab currently participates in the International Lattice Data Grid (ILDG), hosting the U.S. ILDG metadata server, and a share of the U.S. LQCD files. ILDG uses Virtual Organization Membership Service (VOMS) tools for membership, hosted in Europe.

The Laboratory offers some limited use of GridFTP (Grid File Transfer Protocol) and other data-transfer tools. This is expected to grow to ease the transfer of files to and from Jefferson Lab. No computational grid is currently planned.

Videoconferencing continues to grow, and support for robust, easy-to-use tools is essential. Jefferson Lab currently makes use of ESnet's collaboration services for audio- and videoconferencing. Experimental collaborations associated with the 12 GeV program have adopted these services for weekly meetings. Usage is expected to increase through 2015 as the 12 GeV program ramps up.

Jefferson Lab is also expected to make use of federated identity services and InCommon services to authenticate collaborators in the 12 GeV era.

6.8 Outstanding Issues

None at this time.

6.9 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)				
• GeV program • LQCD computing	• Detector simulation, data analysis, mostly lab-centric batch analysis • QCD simulation	• 2 GB * N • 100 MB • 400 MB	• < 1 minute • <10 seconds • few seconds	• < 1 minute • few minutes
2-5 years				
• 12 GeV program	• (as above)	• (as above, N 10x larger)	• 10x higher bandwidth	• 10x higher bandwidth
5+ years				
(tbd)	(tbd)	(tbd)	(tbd)	(tbd)

7 The ALICE Experiment

7.1 Background

The ALICE (A Large Ion Collider Experiment) collaboration constructed and operates a heavy-ion detector to exploit the unique physics potential of proton-proton and nucleus-nucleus interactions at Large Hadron Collider (LHC) energies. The principal goal of the experiment is to study the physics of a new phase of strongly interacting matter at extreme energy densities known as the quark-gluon plasma (QGP). The study is carried out with measurements of the properties of Pb+Pb and p+p collisions produced at the LHC.

The ALICE detector operates year-round in conjunction with the running schedule of the LHC at CERN, taking data during both the p+p and Pb+Pb collision periods each year. Data from the experiment is collected per detected collision (event). Consequently, relevant quantities for network, storage, and computing requirements reduce to per-event quantities, such as event size and processing time, multiplied by the event collection rate or total number of events collected. For ALICE, the overall event rate and subsequent amount of data generated is quite large, with annual event collections of about 1.5 billion p+p events and 150 million Pb+Pb events, corresponding to 3.0 PB and 2.0 PB of raw data, respectively.

The scientific workflow is a sequence of processing over the collected (or simulated) data based on detector and event characteristics. At each step in the process, reduced data sets are created and stored for further analysis. The workflow includes the reconstruction of raw data (detector signals) into interpretable physics quantities such as particle tracks or energy deposition in a detector. The resulting processed data, referred to as event summary data (ESD), are used directly in analysis tasks but also processed further, using standard sets of pattern recognition and filtering algorithms to produce a refined set of quantities known as analysis object data (AOD), used in most end-user analyses. Details about ALICE software and data definitions can be found on the ALICE Offline Computing pages.[1] Individual scientists or subgroups of physicists working on common analyses use these refined data for specific analysis tasks. Throughout the processing steps, the data retain their event-based granularity until the information is eventually reduced to a few sets of numbers or graphs that can be directly interpreted as general physical properties of the colliding system. The event-based granularity allows event processing to be distributed over a large number of independent compute facilities.

The distributed computing model is characterized by a Tier system, composed of a single Tier 0 (T0) center at the experiment site, several Tier 1 (T1) centers providing additional processing and both tape- and disk-storage capacities, and many smaller Tier 2 (T2) centers for processing and disk-storage capacities. The roles of the Tiers are noted briefly here. Raw event data is stored at the single T0 computing facility at CERN, where detector calibrations and initial event reconstruction passes are run. The rest of the computing workflow is done on the ALICE Grid consisting of about 80 additional facilities, seven T1 and about 70 T2 centers distributed around the world. The T1 facilities are relied upon for: (1) long- term custodial storage of a copy of the

[1] http://aliweb.cern.ch/Offline/

raw and reconstructed data, (2) additional reconstruction passes over the raw data, (3) further processing and analysis of the reconstructed data, (4) disk resident storage of and access to ESD and AOD data, (5) processing and storage of simulation data in quantities comparable to the real event data, and (6) running end-user analysis tasks. The T2 facilities provide the same functions as the T1 facilities except for (1) and (2) above. Since 90% of the processing on T1 and T2 sites is devoted to analysis or simulation tasks, there is little distinction between T1 and T2 facilities for the general work carried out on the ALICE Grid facility. In practice, however, the T1 and larger T2 sites tend to have a larger amount of disk storage relative to the number of CPU-cores and as such perform a larger number of tasks that require significant input data.

At each step in the process, data is moved out from the T0 and T1 centers to the T2 centers while Monte Carlo results are moved from T2 to T1 centers. Multiple copies of ESD and AOD files are generated automatically at processing time and written to grid-enabled Storage Elements (SEs) in the ALICE Grid facility. One copy remains on the site where the job ran (assuming an SE is available at the site) while other copies are distributed onto remote SEs. The distribution process (run at the end of each job) uses information on storage capacity and network proximity of potential destinations to decide where to send copies but also includes a random selection for one destination site. A record of each copy is stored in a global file catalog. As a result of this process, data are distributed to a variety of remote sites, to be available for further analysis. These data transfers are characterized by a nearly steady-state migration of data out to the distributed resources.

The ALICE Grid is designed to allow all users to analyze data directly on the distributed facility. An ALICE scientist submits a task to a central task queue located at CERN. Submission can be done from any facility or personal computer with the appropriate client software and the ability to authenticate with a personal grid certificate to connect to the AliEn (ALICE Environment) grid infrastructure.[2] The tasks are broken up into many identical jobs, each accessing a small subset of the data. The individual jobs are executed through a process in which the participating grid sites pull work from the central task queue. Specifically, a middleware component at each site monitors both its local resources and the pending jobs on the task queue, taking jobs from the queue when the local resources meet the needs of the job. One criterion for a site running a job is the availability of input data needed by the job on the site. That is, the infrastructure favors running jobs on sites where the needed input data exist. However, when priority dictates, jobs are run on sites on which data are accessed dynamically over the WAN.

To minimize contention for accessing data, ALICE analysis jobs are typically organized into "trains." An analysis train is a collection of many analysis tasks, run on the grid together over a predetermined set of data. Thus, instead of each analysis independently reading the same input data from disk, the data is read once for the entire train, reducing the cost, for example, of reading dynamically over the WAN. Data produced by an individual's analysis task within a larger job are written to an ALICE SE and logged in the global AliEn File Catalog, but limited to single copies. Those data files can then be accessed on the ALICE Grid or copied to a scientist's local site by AliEn client tools for more direct access.

[2] http://alien2.cern.ch/

Although all data processing and analysis can be run on the ALICE Grid, scientists often need to run an analysis task repeatedly over a fixed subset of data with a very fast turnaround time. The short turnaround time allows a scientist to make modest changes and refine analyses. For this type of work, ALICE supports independent analysis facilities (AFs) where subsets of data are staged to disk and accessed by many users and analyzed in parallel via PROOF (Parallel ROOT Facility).[3] The staging process requires pulling data from the distributed grid-based SEs to the AF, and can be characterized by relatively large data transfers (10-100 TB) over short time intervals from distributed sources to an individual facility.

7.2 Key Local Science Drivers

7.2.1 Instruments and Facilities

U.S. participation in ALICE from a compute facility perspective is concentrated at two new T2 centers at the National Energy Research Scientific Computing Center/Parallel Distributed Systems Facility (NERSC/PDSF) at Lawrence Berkeley National Laboratory (LBNL) and the Livermore Computing Center (LC) facility at Lawrence Livermore National Laboratory (LLNL). The two facilities are comparable in size: 1,000 CPU-cores and 0.5-1.0 PB of disk space. Together, they are of similar size to other ALICE T1 sites. Both sites are integrated into the ALICE Grid and accessed via the AliEn client framework. The NERSC/PDSF site also allows direct login for registered users and supports the client software for job submission or data access. The NERSC facility also includes tape storage via an allocation on the NERSC high-performance storage system (HPSS), for which several hundred TB of storage are planned for each year. That tape storage capacity allows the NERSC facility to become a T1 facility for ALICE and should be considered as such for understanding its network requirements in the coming year and beyond.

The two U.S. facilities represent about 7% of all of ALICE Grid computing resources in terms of both CPU-cores and disk space. The grid-enabled SE at each site is composed of several modest-size (50-70 TB) file servers, each with at 10 GigE connection to the facility core router and then to ESnet directly (NERSC/PDSF) or via additional routers (LLNL/LC) also with 10 GigE connection. The compute nodes at each facility are 1 GigE attached. The multiple file servers are integrated into a single facility-wide SE using XRootD. Each facility has an XRootD manager (redirector) to which each server connects. Each redirector, with its ALICE SE name, is registered with AliEn and with the ALICE global XRootD redirector. The site SE supports processing done on the site, accepts data copied from remote processing, and, as needed, supports I/O for remote processing.

7.2.2 Process of Science

As mentioned above, the ALICE grid allows all users to perform their analysis tasks on the grid facility. Thus it is common for a scientist to submit a task to the grid as if it were a cluster, retrieving the output to a local facility for final processing. However, often the work of refining an analysis requires turnaround times of hours rather than days, run over smaller but repeatable subsets of the entire data set. This type of workflow is more optimally run on an

[3] http://root.cern.ch/drupal/content/proof

ALICE AF. In general, these facilities match a reasonable number of processors with dedicated disk space for staging data sets for common use. Jobs are submitted directly to and processed on the AF. In particular, ALICE computing only supports AFs based on PROOF clusters for implementing job-parallelism, which by design include XRootD installations for data staging and I/O.

In terms of the science process, both U.S. T2 facilities support jobs run on the ALICE Grid. The NERSC/PDSF facility also allows direct logins by ALICE scientists and supports use of AliEn client tools. That is, ALICE scientists can submit jobs to the ALICE Grid from PDSF and copy data locally from the grid for more direct access. While NERSC/PDSF resources are not currently configured as an ALICE AF, users have begun using the resources opportunistically for this type of work. It is reasonable to assume AF-type use will grow in the future, perhaps even with dedicated resources.

7.3 Key Remote Science Drivers

7.3.1 Instruments and Facilities

The ALICE experiment and the Grid's T0 facility are located at CERN and, in fact, most of the ALICE Grid resources are located in Europe. This includes all seven current ALICE T1 sites, with each T1 site supporting a number of T2 sites in their respective countries and nearby regions. The ALICE grid facility and operations were developed in this environment.

A shift in the concentration of ALICE resources from Europe has begun in the past year with new projects in the United States, Korea, and Mexico. The two U.S. facilities became operational at the end of 2010, with a steady ramp-up of resources going into 2012. For the purpose of this case study, the NERSC facility should be considered a T1 center, with its tape storage fully integrated into the ALICE Grid framework. A T1 facility is being constructed at the Korean Institute of Science and Technology (KISTI) center with a planned 1,500 cores and 1 PB of storage in 2012. A new ALICE T1 center of approximately 1,000 cores and several hundred TB of storage has been approved at the National Autonomous University of Mexico (UNAM) in Mexico City, with deployment scheduled in early 2012. For each of these new T1 facilities, the data-transfer path from CERN and other European centers goes through the United States, directly to the U.S. facilities, or to Korea or Mexico via international links.

Current processing and data transfers at the U.S. facilities are like those to any ALICE T2 site. The storage element at each site supports jobs run on the site, providing input data and retaining output data. Data is also copied into and out of the site as part of the normal distribution process. Since these U.S. facilities have been operating in that mode for about a year, we can provide initial estimates of the data transfers. Data movement into each of the two U.S. sites averages about 40 MB/sec while data transferred out is somewhat less, on the order of 25 MB/sec. These data transfers, done as automatic operations in the ALICE Grid, are nearly continuous and spread uniformly to several countries, primarily in Europe.

For this case study, we assume that the NERSC facility becomes an ALICE T1 center and thus will be responsible for custodial storage of a fraction of all raw and primary reconstruction (ESD) data for ALICE sent directly from the T0 center at CERN. The percentage, currently about 7%, is

fixed by the size of the U.S. participation measured relative to all of ALICE. As mentioned earlier, ALICE takes about 3 PB of p+p raw data and 2 PB of Pb+Pb raw data each year. The p+p data are acquired over about a nine-month period while the Pb+Pb data are acquired over a 20-day stretch each fall before the winter shutdown. The goal for these data transfers is to achieve a steady-state transfer model where the Pb+Pb data is copied over three months and the p+p data over the following nine months. This translated to 350 TB per year or a steady transfer rate of about 12 MB/sec. In addition to the raw data, the site will accept a steady fraction of the ESD data. Several copies of ESD data are distributed, so that the fraction sent to the U.S. T1 become closer to 25%. The ESD data are about one-tenth the size of the raw data, adding about 100 TB to the steady transfers from CERN. The combined raw and ESD data transferred are then about 500 TB per year, or a steady 16 MB/sec.

As a consequence of providing custodial storage of raw and ESD data files, the U.S. facilities will participate more directly in the data-distribution process. Copies of new ESD files, produced by additional reconstruction passes over the raw data and new AOD files from additional ESD processing, will be distributed from the U.S. T1 site[4] to other ALICE Grid sites. In addition, another ALICE T2 site is operated at the Ohio Supercomputer Center (OSC), providing a few hundred cores but limited local storage. The network proximity to the T2 sites at LBNL and LLNL will allow the OSC site to participate in more data-intensive tasks, using WAN for direct I/O. From these considerations, it is expected that the outgoing traffic, now at 25 MB/sec, will grow to 40 MB/sec.

7.3.2 Process of Science

In the previous *Process of Science* section, two modes for how scientists typically do their analysis work were noted: They run grid submissions that produce output that is retrieved for further study and they run repeatedly on a fixed subset of data that is pre-staged locally. In the first model, the stress on the network is limited, as the input of the submission is largely metadata while the output is small. In the second model, though local running has little effect on the WAN, the initial staging of the data set can be significant, not to mention that thanks to the XRootD software, data can also be accessed directly on WAN from the AF. These data subsets have total sizes of tens of TB and need to be refreshed as more data are taken or new calibrations upstream of the analyses happen. Thus, to meet the scientists' needs, periodic staging of 5-10 TB/day over several days will not be uncommon, but limited to the resources that can be devoted to such local end-user processing. As to the effect on the network, while such data are pulled from multiple distributed sources on the ALICE Grid, this traffic does add short-term bursts of 50-100 MB/sec to the WAN, perhaps a few times per year.

7.4 Local Science Drivers — The Next 2-5 Years

7.4.1 Instruments and Facilities

The U.S. facilities are in production and, after the remaining ramp-up in the NERSC/PDSF site this next year, are expected to grow at a rate consistent with replenishing retired hardware

[4] We expect LLNL/LC to operate much like a T1 due to its network proximity to NERSC/HPSS.

with compute and data servers that have higher resource densities and are expected from the commodity hardware market. Such a growth pattern over the next five years would increase the two U.S. facilities to a few thousand cores and several PB of disk space each. As a result, the rate of data migration within the normal ALICE Grid operations should keep pace, likely doubling before the end of the five-year period.

7.4.2 Process of Science

The most likely change in the two-to-five-year period that could significantly affect network requirements is the possibility of deploying a dedicated U.S. ALICE AF at NERSC. Elsewhere, such facilities have become popular with scientists. Compared with the current opportunistic use of resources for this type of work, a dedicate AF would require more managed transfers of specific data sets into the facility. As a result, the need for 100 MB/sec bursts over a several-day period would occur more frequently, perhaps monthly.

7.5 Remote Science Drivers — The Next 2-5 Years

7.5.1 Instruments and Facilities

The biggest change in the two-to-five-year period is the long shutdown of the LHC after the 2012 running period. The shutdown, which will significantly reduce data transfers from CERN in early 2013, will last until the LHC resumes operation in 2014. The ALICE Grid facility will continue to operate, reprocessing data from the first three years and running new simulation and analysis tasks. Estimates for changes in ALICE operations for the period after the LHC resumes are best guesses at this point. While no dramatic changes are expected, a number of smaller additions are planned that will increase the overall resource demands by about 50% over current requirements, starting in 2014 and increasing modestly for the following couple of years. Thus, we expect that the steady-state 50 MB/sec rates for data moving in and out of each facility will increase to perhaps 100 MB/sec at the end of that period.

7.5.2 Process of Science

One possible change in the two-to-five-year period is the development of a truly grid-based AF. The goal would be to give the scientist the same fast turnaround times with repeated access to specific subsets of data as is achieved with a single-site AF. It is unclear at this time how this might affect use of the WAN. It may mean replacing the pre-staging process with more processing on the sites where the data exists but it may also mean more dynamic I/O, reading data directly from the WAN. The goal is not to minimize WAN traffic, but to provide fast turnaround on the analysis of specific data sets. There may be multiple ways to build such a workflow.

7.6 Beyond 5 years — Future Needs and Scientific Direction

Increase in both CPU and storage may be anticipated, depending upon future plans of the ALICE collaboration and DOE Nuclear Physics priorities. Specific information is not available at this time to determine the impact on network requirements.

7.7 Middleware Tools and Services

ALICE relies on XRootD for data storage. ALICE-XRootD includes a Grid Security Infrastructure (GSI) authentication plug-in and is configured with a global redirector to provide a functional global file system. Personal grid certificates for ALICE's U.S. scientists are signed by the DOEGrids CA through the Open Science Grid (OSG) Registration Authority as part of ALICE Virtual Organization (VO) participation in OSG.[5] Client requests for data made directly to the global redirector will return direct access to the requested data, with data served by the XRootD protocol. ALICE data, however, are ultimately managed using the AliEn File Catalog, where information about each file is stored. Client requests for data from the File Catalog will be given site-level SE information, allowing the client code to connect directly to the XRootD redirector at the site instead of the global redirector. For WAN access to data, XRootD allows multiple streams and even has a "BitTorrent-like" option for pulling data off multiple sources.

ALICE Computing has adopted MonALISA (MONitoring Agents using a Large Integrated Services Architecture) for monitoring all sites and services on its grid infrastructure. The data are archived at the central ALICE MonALISA collector at CERN and presented on the Web at http://alimonitor.cern.ch/map.jsp. In particular, ALICE continuously monitors network connections between all sites with MonALISA by periodically performing an automated memory-to-memory file transfer between each of the ALICE Grid sites (VO boxes) using the FDT tool.[6] These measurements are done every few days, providing single-stream bandwidth and round-trip time measures between every ALICE site. Tables for the two U.S. T2 sites are provided on the MonALISA Web display:

- http://alimonitor.cern.ch/speed/index.jsp?site=LBL
- http://alimonitor.cern.ch/speed/index.jsp?site=LLNL

7.8 Outstanding Issues

Understanding the topology of network connections and learning how to measure and respond to network issues is perhaps the most confusing part of the operation. This is important to ALICE, as the ALICE Grid contains ~80 sites that are all used dynamically to access data. For example, in the table at the above links, it is expected that LBNL and LLNL are often the closest sites to each other - but this is not always the case. On occasion, the measured bandwidth between the U.S. sites is poor, and they appear to have better network connection to individual sites in Europe or Asia. Understanding these measurements and whether they are good proxies for network proximity between different sites could have a significant impact on how ALICE uses the WAN.

[5] Unrelated to networks, ALICE-USA relies on OSG middleware for job submission (OSG client) and site interface (OSG CE). This allows direct site monitoring and job accounting by OSG, forwarded to the Worldwide LHC Computing Grid (WLCG).

[6] http://monalisa.cern.ch/FDT/

7.9 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)				
• ALICE Detector at LHC generates about 4 PB of event data per year. U.S. facilities are part of the ALICE Grid and represent about 7% of ALICE's resources.	• A large fraction of data analysis is done directly on the ALICE Grid; however, a significant subset is done separately on dedicated analysis facilities.	• Data volume in normal grid operations ~2-3 TB/day. • Local data sets of 10-100 TB refreshed a few times/ year. • Typical file size = 500 MB.	• Jobs run ~few hours and process tens of GB. Large per-job I/O range: 0.1 – 10 MB/sec. Approx. 1,000 concurrent jobs/site.	• Steady transfers into each site of 50-60 MB/sec growing to >80 MB/sec. Transfers out per site from ~25 MB/sec to ~40 MB/sec. • Added peak transfers into NERSC/PDSF at 100 MB/sec a few times per year. • Some targeted transfers from CERN, but routine transfers between all ALICE Grid sites.
2-5 years				
• Short-term reduction due to LHC shutdown in 2013. Steady increases afterward.	• Potential greater reliance on and deployment of a dedicated AF at NERSC/PDSF. • Potential for more dynamic grid-based AF could alter how the scientific workflow is implemented.	• Steady increase to 5 TB/day. • Local data sets of 10-100 TB refreshed more frequently, ~monthly. • Files remain 500 MB.	• Little change. Jobs run ~few hours and process tens of GB. Large per-job I/O range: 0.1 – 10 MB/sec. Approx. 2,000 concurrent jobs/site.	• Steady transfers into the facility reaching ~100 MB/sec. Transfers out to keep pace at ~50-100 MB/sec. • Peak transfers of 200 MB/sec lasting a few days occurring more frequently. • Targeted transfers from CERN + routine transfers between all ALICE Grid sites.
5+ years				
• Detectors, DAQ, and new analysis channels will continue to evolve.	• New analysis tools could again alter how the scientific workflow is implemented.	• Steady increase in overall volumes.	• Per-node I/O may become important with many cores per node and independent jobs per core.	• Steady transfers into the facility should increase as the facility grows.

8 CMS-HI Research Program

8.1 Background

The Compact Muon Solenoid (CMS) experiment collaboration at the Large Hadron Collider (LHC) consists of more than 2,000 scientists from 35 countries. The Heavy Ion (HI) Analysis group, a subset of this collaboration, includes about 120 scientists from nine countries. The primary physics interest of this group is the study of the quark-gluon plasma (QGP) phase of matter produced in the collisions of Pb+Pb nuclei accelerated by the LHC. The heavy ion research effort in CMS is fully integrated into the management and operations of the entire CMS physics program. The heavy ion research programs conducted by U.S. institutions in the CMS are supported by the Nuclear Physics (NP) Program Office of the DOE Office of Science.

The first heavy ion collisions at the LHC took place in November 2010, at a center-of-mass energy of 2.76 TeV per colliding nucleon pair. This run was very successful. Shortly after, a major international conference in the relativistic heavy ion field was held in France, where the ALICE and CMS heavy ion groups presented an enormous amount of heavy ion data analyses. These first analyses appeared to confirm that the model of a strongly interacting liquid for the QGP was persisting in the LHC collisions, at more than an order of magnitude increase in beam energy compared with the Relativistic Heavy Ion Collider (RHIC) experiments.

The computing model for the CMS heavy ion program is tightly integrated into the general computing model for the CMS high-energy program. There is a central Tier 0 (T0) facility at CERN that initially collects the raw data from the experimental apparatus and performs a prompt reconstruction. When those initial steps are completed, the raw data and prompt reconstruction production are shipped to Tier 1 (T1) facilities for archival to tape storage and further data re-reconstruction. In the case of the heavy ion data, this means a transfer of all these data files to the T1 site at Fermi National Accelerator Laboratory (FNAL) for the tape archival storage. The new feature for heavy ion computing is that these FNAL data files are then subscribed to immediately by a new, specially designed Tier 2 (T2) facility at Vanderbilt University. Of course, since the heavy ion beams run only for about four weeks of the year, typically mid-November to mid-December, the intense network traffic for its data takes place for one month's duration.

The Vanderbilt T2 then supports user physics analyses of the prompt reconstruction production, and undertakes new re-reconstruction steps as the physics analysis demands. Other CMS T2 sites in France, Russia, Brazil, and Turkey can transfer reconstruction production from the Vanderbilt T2 site to their own, smaller disk systems to diversify the analysis venues. This took place for the 2010 data set.

The CMS-HI computing resource base also has a special heavy ion simulation and analysis facility at the MIT T2. Before the Vanderbilt T2 facility came into existence in 2011, the MIT T2 heavy ion resource was the main computing site during the several years of simulation investigation into the capabilities of the CMS detector in this program. The MIT heavy ion facility is about one-fourth the size of the Vanderbilt T2 facility.

In the United States, other heavy ion institutions are the University of Maryland, the University of Illinois at Chicago, Kansas University, the University of California at Riverside, and a newly starting program at Rutgers University. Collectively, these other sites are called Tier 3 (T3) institutions, indicating that they have the minimum set of local hardware and software needed to interact efficiently with the T2 facilities at MIT or at Vanderbilt, as described below.

8.2 Key Local Science Drivers

8.2.1 Instruments and Facilities — CERN Tier 0

The resources of the CERN T0 facility are allocated among four experiments - ALICE, A Toroidal LHC Apparatus (ATLAS), CMS, and Large Hadron Collider beauty (LHCb) - according to a prescription set by the LHC Computing Resources Scrutiny Group. During the approximately one month of heavy ion running at the LHC, the entire CMS T0 allocation is in principle available to process the heavy ion data. However, the computing power of the CMS T0 allocation vastly exceeds the needs of heavy ion prompt reconstruction, at least for the 2010 and anticipated 2011 data taking. It is estimated that the prompt reconstruction step for the expected 2011 HI data set will be about one week compared with the four weeks nominally available, even with only half the total allocated T0 computing power being used. Projections for future years, when the heavy ion event set will grow more complex because of the use of more sophisticated high-level triggering, are that the prompt reconstruction step will ultimately consume 30 days with half the available computing power.

Concerning disk storage, the 2010 data taking was a special circumstance in which the CMS experiment provided 1 PB of disk space for the heavy ion data set for several months into 2011. This special provision was made because the 2010 data were taken in non-zero suppressed (NZS) mode, resulting in about 800 TB of raw data being initially written. The NZS mode was necessary owing to great uncertainty about how the normal CMS zero suppression software would perform in heavy ion events as compared with p+p events. Some hypotheses stated that the tracking systems in particular would be adversely affected by highly ionizing tracks, which are more prolific in heavy ion events than in p+p events. Therefore, the first two months in 2011 were dedicated to studying the zero suppression effects on the NZS heavy ion raw data. In the end it was found that the adverse effects of zero suppression in heavy ion events could be easily overcome with slight improvements to the zero suppression software. In February 2011, the 800 TB of NZS raw data were reduced to 150 TB of ZS, which then produced 190 TB of reconstructed data at the T0. These amounts were then transferred to the FNAL T1, according to the computing plan. It would not have been feasible to transfer the initial 800 TB of NZS raw data to FNAL.

For the 2011 data taking, the plan is to do the zero suppression in the high-level trigger system, with the raw data then reconstructed immediately at the T0, just as in the case of p+p data taking. The maximum possible amount of data to be transferred to FNAL, both raw and prompt reconstruction, has been set at 500 TB for 2011.

8.2.2 Instruments and Facilities — Tier 1

The only T1 facility for the CMS-HI program is at FNAL, which has the highest capabilities among all other CMS T1s. The initial plan for the CMS-HI computing model was to have the Vanderbilt site perform as a full-fledged T1, with its own tape-archiving capability. However, in a CMS computing workshop held in Bologna in September 2009, it was determined that a separate tape-archiving capability at Vanderbilt would unnecessarily duplicate resources and operations at the FNAL T1. For CMS computing, the single month of heavy ion running is just one more month of "normal" data operations, the only difference being that all file transfers go to FNAL, not just the largest fraction. There was, therefore, no need for Vanderbilt to meet the stringent standards of T1 data transfer capability, which FNAL was already meeting.

The only real modification of the general CMS computing model is to allow the Vanderbilt T2 site to subscribe to, i.e., transfer, raw data files from the FNAL T1 site. Normally T2 sites have no need for raw data files and thus cannot access these.

In March 2011, the initial 190 TB of prompt reco generated at the T0 were transferred to FNAL and then to Vanderbilt. The transfer to Vanderbilt took 19 days, i.e., 10 TB/day. For reasons explained in the next subsection, the transfer of the 150 TB of zero-suppressed raw data did not take place until August 2011.

For the November 2011 heavy ion data set, and for future data taking, the plan is to do immediate data reconstruction at the T0, and subsequent transfers to FNAL and then to Vanderbilt.

8.2.3 Instruments and Facilities — Tier 2

The T2 facilities for the CMS-HI program consist of the new site at Vanderbilt, a special heavy ion component at the MIT T2 site, and four non-U.S. T2 sites in France, Brazil, Russia, and Turkey. The largest of these is the Vanderbilt T2 site, which had 480 cores (4.1 kHS06) and 620 TB of disk space for processing the 2010 heavy ion data set. The MIT heavy ion facility has about one-quarter of this size, and likewise the non-U.S. T2 heavy ion resource commitments are each comparable to the MIT heavy ion facility.

The Vanderbilt T2 facility did not exist when the first heavy ion data were taken at the LHC in November 2010. The internal and external review processes that resulted in a strengthened proposal lasted into the summer of 2010, with final spending authority granted on November 1, 2010. First hardware was not commissioned until January-February 2011, and the system went "live" in March 2011 to receive the first prompt reco data sets from FNAL. This initial transfer of 190 TB took 19 days to accomplish, a much lower average speed of ~1 Gbps than the several Gbps that is typical of a CMS T2 transfer rate in the United States.

Part of the low rate of the initial transfers was due to the Phase I disk I/O software, which had just been developed for this T2, and which uses its own mass disk storage system not present elsewhere in CMS. This I/O software was extensively upgraded between March and August 2011 into a Phase II version, which was used to transfer the 150 TB of raw data files from FNAL. The result with the upgraded software was a sustained rate of 2.4 Gbps, with the expectation that 3.6 Gbps would be easily within reach after a known faulty network switch was repaired.

Since even the lower rate of 2.4 Gbps corresponds to more than 25 TB/day, the Vanderbilt T2 is judged as ready to accept as much as 500 TB over four to five weeks for the 2011 heavy ion data taking.

8.2.4 Process of Science — Tier 0

The function of the T0 in the CMS-HI computing model is to provide the initial prompt reconstruction data set for immediate use by physicists. There is additionally a data-quality monitoring effort to assure the detector, including the trigger systems, works properly for the heavy ion collision events. Depending on disk availability, the prompt reco data may remain accessible to users to do analyses at CERN for perhaps one month after the heavy ion run is concluded.

The 2010 data taking was a special case, wherein only 10% of the data were initially reconstructed at the T0 immediately after being acquired. This “core” reconstruction data set formed the basis of the CMS-HI physics analyses until the completely reconstructed data set became available in late March.

In 2011 and future years, 100% of the heavy ion data will be promptly reconstructed at the T0.

8.2.5 Process of Science — Tier 1

The FNAL T1 has no direct science role for the HI program in CMS. There is specifically no user access to the HI data on disks at FNAL, nor can any other CMS T2 except Vanderbilt transfer those files from FNAL. The role of the FNAL T1, after receiving the heavy ion data from the T0, is to provide the secondary tape archive storage, with a first, read-only storage at the T0. This secondary storage is mandated at a T1 in the general CMS computing model.

8.2.6 Process of Science — Tier 2

In the CMS computing model, the T2s are the main sites on the computing grid where users do physics analyses on the reconstructed data. The T2 sites also provide simulation computing resources for the collaboration. This same model is applied for the heavy ion program. Both the Vanderbilt T2 system and the MIT heavy ion analysis facility are accessible to users via the CMS Remote Analysis Builder (CRAB) software tool. With the CRAB tool, a user can compose an analysis task to process the entire subset of reconstructed data at Vanderbilt, or copies of these data files that have been transferred to MIT or other non-U.S. T2 heavy ion facilities in CMS. These analysis jobs can be submitted from any T3 or T2 site in CMS. The internal network speeds of the Vanderbilt system are such that the entire 190 TB of the 2010 reconstruction data set can be scanned in only a few days. Typical read and write rates of 2-3 GB/sec are observed.

The Vanderbilt T2 system went “live” for CRAB job submissions in early April 2011, just after the reco data were received and several weeks before the May 2011 physics conference at Annecy, where the first CMS heavy ion physics results were presented. Several of those analyses benefitted from the ability to scan the entire data set at Vanderbilt. Previously, only the 10%-size data set, initially produced at CERN in November 2010, had been available for physics analyses.

One of the particular advantages of the Vanderbilt T2 is that it is embedded into a much larger (>3,000 cores) largely homogeneous computing facility. This allows the possibility of opportunistic computing, meaning the ability to run on otherwise idle cores that were not bought as part of the CMS-HI computing base. In fact, the number of CMS job slots is set currently at 950, almost twice the number of purchased cores. During especially busy times, such as before a major conference, the number of running CMS jobs is high. This opportunism is possible because the CMS jobs run on a 24/7 basis, while much of the local non-CMS use at the Vanderbilt Advanced Computing Center for Research and Education (ACCRE) facility is pegged to an eight-hour weekday basis. The internal network speeds of this T2 are designed with this "bursting" or opportunistic computing capability in mind.

8.3 Key Remote Science Drivers

8.3.1 Instruments and Facilities

In regard to the networking capabilities for the CERN T0 and the FNAL T1, the CMS-HI research program benefits immensely from the idea that the heavy ion data taking is regarded as just one more month of normal CMS data taking. Therefore the entire set of local network resources at the T0, and the transfer operations to the FNAL T1, are available and the CMS-HI data transfers are simply regarded as standard work in CMS. Since the heavy ion data volumes with zero suppression are roughly comparable with those of the p+p program on a monthly basis, the existing infrastructure is under no appreciable additional strain.

8.3.2 Process of Science

Although all CMS personnel are available for the heavy ion running, the process of science for the heavy ion running is driven by the much smaller heavy ion group in CMS. This group decides which events to select, and whether to partition the available data acquisition bandwidth among minimum bias events and specially selected triggers to probe the QGP. The group also must certify modifications to the standard CMS reconstruction software to process the heavy ion events. The heavy ion events are intrinsically more complex to handle than isolated p+p events. On the other hand, the much-increased luminosity of the p+p collisions leads to pileup complications (multiple collisions in the same event trigger), which fortunately are not a difficulty for the heavy ion events.

The heavy ion T2 facilities in CMS operate more or less independently of the decisions made for the non-heavy ion Tier 2 in CMS. The choice of which skim files to generate for physics analysis, which simulation sets are necessary, and when to do a new reconstruction pass are taken by the heavy ion group alone. The external network speeds between the T2s for the heavy ion program are specially tested to ensure prompt file transfers when necessary. In 2011, the network transfer speeds between the T2s have not been a problem for the physics analyses.

8.4 Local Science Drivers — The Next 2-5 Years

8.4.1 Instruments and Facilities - CERN

The LHC is expected to run Pb+Pb beams in November 2011 and again in 2012, with the possibility of changing to p+Pb collisions in 2012. During a long shutdown planned for 2013, the superconducting magnets will undergo modifications to enable the LHC to reach its design energy of 14 TeV. Heavy ion operations may resume in 2014 or 2015, likely at twice the current collision energy. At a minimum, this will mean larger event sizes and more complicated events than are presently being recorded. The natural trend would be to allow for at least 50% more data volume than the current level. The actual data volumes for the 2010 data set at 2.76 TeV were somewhat underpredicted, at ~20-30%, but enough reserve capacity was built into the computing assumptions so that the extra complexity was not a serious problem for the available resources.

8.4.2 Process of Science

The science process at the T0 is expected to change significantly over the course of the next few years. 2011 will see more significant use of the high-level trigger system due to the high luminosities of the Pb+Pb collisions. This trend will expand in 2012 if Pb+Pb beams are run again. On the other hand, if it is decided that running p+Pb collisions is feasible at the LHC in 2012, a major study will be needed on how the CMS detector will operate in this mode. A strong motivating factor for running p+Pb collisions is that they probe the “cold nuclear matter” effect in heavy ion collisions. The p+Pb collisions at the LHC, like the d+Au collisions at RHIC, will constrain models of the heavy ion reactions that allow non-QGP scenarios to mimic QGP scenarios. There has not been much simulation study of p+Pb collisions at the LHC to date. If it is decided that such collisions are possible - and the decision may not be reached until February 2012 - then an intense effort will be necessary at that time to maximize the potential of this physics program.

Similarly, if in 2014 or 2015 there are to be Pb+Pb collisions at over 5 TeV, a major investigation will be needed during 2013 to be ready for such higher-energy data. Fortunately, this will be only a factor of two jumps in beam energy, compared with the factor of 15 from RHIC.

8.5 Remote Science Drivers — The Next 2-5 Years

8.5.1 Instruments and Facilities

As currently planned, the T1 and T2 facilities for CMS-HI should largely keep pace with the expected changes in 2011 and 2012. The present plan for the Vanderbilt T2 is to grow to more near 20 kHS06 in the next four years, with some 1.3 PB or more of disk space. The CMS data acquisition system could easily deliver more data than either the p+p or the heavy ion downstream computing resources could handle. The amount of data being taken is limited to what can be processed and stored in a given year; it is expected that data will be sufficient to produce significant new physics results. That expectation of new results was certainly borne out with the 2010 data set, and should continue to be true for the 2011 and 2012 data sets.

In 2014 or 2015, with collisions at twice the current energy possible, more disk space and higher throughput capacities will be required. If 2011 is limited to 500 TB of raw plus prompt reco data in four weeks, 2014 or 2015 requirements will likely expand to 750 TB or even 1 PB during the month of heavy ion running. This will impact both the FNAL T1 resources and the heavy ion T2 resources in CMS. Since the p+p program is looking forward to collisions at 14 TeV, the likely growth in FNAL capability for the high energy physics will match what is expected for the heavy ion physics.

8.5.2 Process of Science

The impact on the users will scale with the anticipated volumes of data. Intensive users at present may generate several hundred GB of output for their analyses. In five years, this could easily be a few TB. At present, end-user T3 facilities may able to host a few tens of TB of local storage while having 100 MB/sec of external network capability. In five years, 100 TB of local user storage with several hundred MB/sec may be necessary to meet the physics demands.

8.6 Beyond 5 years — Future Needs and Scientific Direction

LHC plans for the next 10 years inevitably entail much higher luminosities for both the p+p and the heavy ion programs. The major push will be to improve detector technologies and the trigger system to cope with much higher instantaneous data rates. Data volumes and transfer rates may need to increase by factors of 3 from what is now required. For the heavy ion field, the way forward will depend on what is seen at the higher energy. Is there a significant transition away from the liquid behavior that is now apparent? Which probes are especially useful to discriminate among models of the QGP? Can events with these problems be enriched in the final data volume by more sophisticated trigger strategies?

8.7 Middleware Tools and Services

Middleware tools, such as those provided by the Open Science Grid (OSG), are vital to the success of the CMS computing model. For the commissioning of the Vanderbilt T2, with its open-source (noncommercial) Portable Batch System (PBS) operating system, OSG experts were consulted, and much time was spent discussing the correct hooks needed to make the CMS CRAB system work well at that site.

As a result of the case-study discussion during the 2011 ESnet/DOE-NP workshop, a consensus was reached to increase focus on the OSG component of the networking system. For the CMS-HI case study, this meant soliciting post-workshop comments from the local facility staff on the OSG system, especially given the numerous interactions with the OSG group during the prior five months of bringing the Vanderbilt T2 into routine operations.

The overriding impression is that while the OSG group is performing a valuable and irreplaceable service, major improvements are urgently needed. The following is summary of the comments from the local staff:

1) Documentation is at times poor or contradictory. Different sets of instructions for installing a Compute Element were found at three OSG sites, but each set lacked a crucial piece of information. Eventually a user's T3 site provided the best set of instructions.

2) The update process has improved. Previously a major reinstallation was necessary even for minor updates.
3) The OSG ticket site GOC is ineffective, as ticket response is glacial. It is more effective to post questions on OSG user mailing lists and hope that other T2 system administrators have already solved the specific problem.
4) The wiki-based documentation system leads to corrupted documentation, depending on the expertise of the last author posting instructions.
5) No clear direction is provided on whether to use Pacman or RPM for installations. The RPM installs are completely alpha/unsupported at present, yet RPM appears to be pushed as the approach.
6) No good "best practices" information on monitoring or security certificates is in place. It took a huge hit-or-miss effort to get to the current stage of self-monitoring, and each new site should not have to repeat the same learning curve.
7) The Physics Experiment Data Export (PhEDEx, a CMS transfer software tool) certificate renewal method is unclear, and the way it is currently done may not be approved. There is the impression that this casual approach does not matter.
8) There is no good U.S. liaison for the SAM/CERN Nagios systems. Again, it is a trial-and-error process to discover the meaning of some monitoring test results.
9) The atmosphere at the OSG conferences can be hostile. One example is an antagonistic attitude to attendees who are not completely expert in public key infrastructure (PKI).
10) The OSG Resource and Service Validation (RSV) metric is considered a pointless exercise, as it is trivial to fool, and more comprehensive metric tools are already in place.
11) No good database of sites in terms of which versions of software are being run, no timeline for when RPM will become the standard, and insufficient workshops for migrating to newer versions of software components such as PhEDEx 4 or BeStMan 2.

To realize optimum performance in the network links between the FNAL T1 and the heavy ion T2, sophisticated network diagnostic tools and good communications with network experts along the transfer path are essential. There is a question about what role, if any, the ESnet point of presence in Nashville should play in the transfer of data between FNAL and Vanderbilt.

Coordination is needed between DOE-High Energy Physics (HEP) and DOE-NP on the outbound network capacity of FNAL to Vanderbilt during the intense one-month period of data traffic when the LHC is running heavy ion beams. During that time, FNAL must continue to meet its HEP responsibilities in CMS for outbound traffic. The FNAL peak networking needs for the CMS-HI program will not conveniently be addressed in the ESnet/DOE-HEP workshop. Similarly, it is not the normal role of the ESnet/DOE-NP workshop to study the networking needs of the FNAL facility. The precise technical specification - whether FNAL should be able to deliver 3 or 5 or 7 Gbps data flow to Vanderbilt, if needed - is not particularly arduous to contemplate. However, this item should be recognized as overlapping two DOE research branches.

8.8 Outstanding Issues

The heavy ion program at the LHC is operating in an unexplored energy domain. Whereas the Higgs boson searchers at the LHC can employ the standard model and existing phenomenology to model exquisite constraints on background contributions, the predictions for the QGP

properties are not nearly as precise. When a switch to p+Pb data taking at the LHC is contemplated, or a change to a factor-of-2 higher in energy is foreseen, the resulting data characteristics are at best an educated guess. As already implied, there were built-in factors-of-2 uncertainties in the data event sizes for the 2.76 TeV running, as extrapolated from the RHIC running. That the actual particle multiplicities were only 20-30% higher than planned is considered a good result. So this intrinsic higher “systematic error” in planning for the heavy ion computing infrastructure requirements must be always kept in mind.

8.9 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)				
• Pb+Pb collisions from the LHC into CMS detector, with use of high-level trigger. • Growth of Vanderbilt T2 facility for CMS-HI. • Growth of MIT heavy ion facility.	• Choice of trigger bandwidth to maximize different probes of the QGP. • Accommodation of many external users to do physics analyses at the heavy ion T2.	• 100 TB/week for 4-5 weeks from CERN, Nov. — Dec. • File sizes of a few GB each, i.e., ~100K files • User analysis production at tens of TB.	• Local CERN networking 800 MB/sec to 1,000 MB/sec from experiment to local storage. • Several GB/sec at T2 analysis. • 4-5-week time span, typically Nov-Dec.	• WAN transfers of 3 Gbps from CERN to FNAL, and same from FNAL to Vanderbilt, Nov.-Dec. • 1 Gbps from Vanderbilt to other HI T2, all year (MIT, France, Brazil, Russia, Turkey). • Few hundred Mbps from Vanderbilt to HI T3, all year (UMD, UIC, KU, UCR, Rutgers).
2-5 years				
• Possible change to p+Pb running in 2012. • Probable change to twice the Pb+Pb collision energy by 2015. • Expected increases in beam intensities. • 2013 is a no LHC data year.	• Change to p+Pb running may have a late decision date, early 2012. This will limit the time for sufficient modeling of the data stream. • Increased energy and/or intensities may easily double the data volumes.	• For the p+Pb or Pb+Pb in 2012, the same data rates as above. • For 2015 running, can expect a factor of 2 from above numbers.	• Same as above for 2012-2015, with 2013 a down year.	• Same as above for 2012. • A factor of two more in 2015.
5+ years				
• Major increase in luminosity of LHC and detector capability, perhaps a factor of 10.	• More sophisticated trigger and trigger computing power to have only a factor of 3 more data volume.	• Assume a factor of 3 from above estimates.	• 3 GB/sec to 5 GB/sec from experiment to T0.	• 30 Gbps from CERN to U.S., FNAL to Vanderbilt, Nov.-Dec. • 1-2 Gbps from T2 to T3.

9 The PHENIX Experiment at RHIC (BNL)

9.1 Background

The Pioneering High Energy Nuclear Interaction eXperiment (PHENIX) is one of two large detector systems at the Relativistic Heavy Ion Collider (RHIC). The experiment has about 1 million readout channels in about 15 detector systems. The experiment's data rates are driven by a high sampled and recorded data rate (5-8 KHz to disk), leading to peak data rates of about 1.2 GB/sec to disk. That number is likely to increase with upcoming detector upgrades.

9.2 Key Local Science Drivers

9.2.1 Instruments and Facilities

The experiment is located in the Building 1008 complex at the RHIC ring. The data are recorded to disk locally in the Counting House, and sent to the RHIC Computing Facility for long-term storage in the high-performance storage system (HPSS).

The local buffering on so-called "Buffer Boxes" at the experimental site helps level the ebb and flow of data, which varies with the RHIC beam intensity, the fill cycle, and other parameters. The local buffer capacity is about 70-100 hours, depending on the beam species and current RHIC luminosity. With the buffer setup, we can ride out short service interruptions of the HPSS service or the LAN. In addition, the most recent data are available locally for monitoring processes and calibration procedures.

The PHENIX experiment relies mainly on the resources of the RHIC/ATLAS Computing Facility (RACF) for long-term data storage, retrieval, and analysis. In the past we sent whole raw data sets through the network to remote facilities for processing (to the Computing Center in Japan [CCJ] and Vanderbilt). Moving large-scale data sets to use remote processing capability proved inefficient. We currently process the data locally at the RACF.

9.2.2 Process of Science

The raw data are reconstructed and converted into data summary tapes (DSTs). Most calibrations, corrections, clustering, track reconstruction, and similar CPU-intensive processing is done once in this process. The resulting DSTs contain higher-level information such as cluster and track parameters, positions, and particle energies, and can be analyzed with relatively modest CPU requirements. The DSTs can be further refined into micro-, pico-, or nano-DSTs, which contain information relevant for a specific analysis project and can be analyzed in a very short time.

The PHENIX collaboration has adopted a centralized processing model, where the vast majority of computing is performed at the RACF. The retrieval and transfer of data is by far the most expensive component in terms of cost and time, and we try to maximize the return on a given file retrieval. Sending the data off site for end-user analysis typically happens at the pico- or nano-DST level, if at all, while an analysis requiring a substantial amount of processed data is virtually always performed at the RACF. The only remote computing center that might request a substantial amount of DST-level data is CCJ, where a large number of PHENIX collaborators are

involved with the spin program. Data sets transferred to Japan tend to be larger in runs that have a major spin (pp) component.

The key technology used at the RACF is the concept of an “analysis train.” The train is, at its core, a file-retrieval management system that has boosted our data throughput at the analysis stage by at least a factor of 15. In the analysis of the early RHIC runs, we tried the well-known data-staging model, where a user or process requests a file that then resides on disk for a given period of time and is then deleted to make room for new requests. This model of unmanaged retrievals is very inefficient, as it typically leads to multiple retrievals of the same file in a short period of time.

The analysis train is run typically once a week (or more frequently in times of high demand) and collects all analysis projects from users interested in a particular data set. The train retrieves the files of the data set, and all registered analysis modules then process the files. In this way, a retrieved file is used by a large number of analysis projects, maximizing the return on the investment of the file retrieval. The waiting period for the train — at most a week - for the desired data set to start is much shorter than the time it took in the past until a given analysis project had gotten hold of the complete data set in an unmanaged fashion.

9.3 Key Remote Science Drivers

9.3.1 Instruments and Facilities

The PHENIX Collaboration currently consists of 70 institutions in 13 countries. About 250 scientists routinely access the RACF for analysis, processing, and presentation of data. The relatively small component of larger-scale data transfers, as opposed to interactive access, shifts the focus from the highest bandwidths to low latency requirements.

9.3.2 Process of Science

Analysis work in PHENIX must be associated with a Physics Working Group (PWG). The PWG has local resources at the RACF for its members, which it manages largely autonomously within the different analysis projects. Most local and remote collaborators draw on the resources of the PWG.

9.4 Local Science Drivers — The Next 2-5 Years

9.4.1 Instruments and Facilities

The PHENIX Collaboration is undergoing an ambitious detector upgrade. The first part, the Silicon Vertex Detector (VTX), has been commissioned in Run 11. The second part, the Forward Silicon Vertex Detector (FVTX) will be installed and commissioned in Run 12. Each component will increase the overall data rate by increasing the per-event size, boosting the expected peak data rates from 700 MB/sec pre-Run 11 to 1200 MB/sec in Run 11, to an estimated 1500 MB/sec in Run 12. The increases in data volume will proportionally affect the WAN transfer rates, which are typically a fraction of the original raw data size. Despite fluctuations, that fraction has been holding more or less steady for several years.

The upgrades beyond Run 12, working name sPHENIX, are currently being designed. This radical upgrade will most likely feature a new magnet, new calorimetry, a new tracking system, and a new data acquisition (DAQ) system. It is too early for data-rate estimates, but the working assumption is the doubling of data rates as compared with Run 12.

9.4.2 Process of Science

We assume we will continue the current concept with PWGs, and remote collaborators using RACF resources, although no specific decision has been made.

9.5 Remote Science Drivers — The Next 2-5 Years

9.5.1 Instruments and Facilities

The immediate impact of the sPHENIX design is a large simulation effort of various detector aspects. We expect a significant fraction of the simulation effort to take place remotely, with simulated data flowing back to Brookhaven National Laboratory (BNL).

Other than that, we assume that the RACF-centric computing model will most likely persist.

9.5.2 Process of Science

9.6 Beyond 5 years — Future Needs and Scientific Direction

The commissioning of the future sPHENIX will almost certainly bring higher data rates, and most likely new data-processing paradigms. We believe currently available technology will sustain the envisioned data rates, even without a progression of the current rate of improvements in data storage and processing technologies.

9.7 Middleware Tools and Services

To the extent that we perform large-scale data transfers off site, we will continue to use the grid middleware tools that have served us well in the past.

Minor network impacts are expected from the current shift from phone-based meetings to videoconferencing and Voice over Internet Protocol (VoIP) services, which impose modest latency requirements on the networks.

9.8 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)				
• PHENIX upgrades with the VTX (commissioned) and FVTX (Run 12) detectors.	• Centralized processing to DSTs. • Analysis trains. • Modest-size off-site transfers.	• ~3 PB raw data. • Reconstructed to ~2.5 PB DSTs. • 800 TB (estimated) transferred.	• Near-line. • 300,000 raw data files. • 10 GB each. • Network in place.	• Virtually no near-line requirements. • 800 TB volume estimated.
2-5 years				
• Data taking with VTX+FVTX. • Simulations for the sPHENIX upgrade.	• Centralized processing as above. • Distributed simulations.	• Estimated data rates according to beam species and energies (estimated in Appendix A).	• Near-line.	• No near-line requirements.
5+ years				
• sPHENIX commissioning and operation.	• No change in computing paradigm envisioned.	• 3 GB/sec peak (weak estimate). • Move to larger file sizes (100 G?).	• Near-line.	• No near-line requirements.

10 The Solenoidal Tracker at RHIC (STAR) Experiment

10.1 Background

The Solenoidal Tracker At RHIC (STAR) is one of two large detector systems at the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory (BNL). STAR is very active international collaboration of 55 institutions over 12 countries. Its goal is to understand the nature of the early universe and the tiniest building blocks of matter through the study of nuclear collisions up to the highest beam energies available at RHIC.

The STAR institutions are geographically distributed as follows:

Table 10-1.

Regional group	N	Percentage	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%

Compared with the previous ESnet workshop in 2008, which included 52 institutions and 590 collaborators, staffing has remained near constant with a solid collaboration, but the distribution, though stable, indicates a shift toward more non-U.S.-based institutions. This shift is not yet at a level to influence network requirements, as our non-U.S.-based colleagues typically use central facilities in the United States (see discussion in the next sections).

The STAR data production and analysis models mainly rely on centralized user analysis facilities, with the RHIC Computing Facility (RCF) at BNL functioning as STAR's Tier 0 (T0) center and the National Energy Research Scientific Computing Center's Parallel Distributed Systems Facility (NERSC/PDSF) center as the Tier 1 (T1) center. STAR is also supplied with many STAR Analysis Centers (SACs, similar in scope to Tier 2 [T2] centers) and although their inventory has been hard to assess, they do constitute pools of local resources dedicated (not necessarily shared with all STAR users) to local group physics programs. Their numbers and size have dramatically fluctuated (from up to 10 to a just few sites, from 15 nodes at 8 cores each to several 100 nodes at 16 cores each) but, as we noted in 2008, they will be part of a global resource pool (bound by grid infrastructure, for example), and their resources would largely cover the simulation needs of the collaboration.

Based on past trends and future perspective, we estimate the number of SACs in Table 10-1. In 2011, our active SACs are Prague (T2), Wayne State University, and the MIT sites (Yale, Birmingham, and Sao-Paulo have phased out or slowed over the past years).

Table 10-2. Projected number of stable STAR Analysis Centers (SACs).

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

The STAR computing model relies on a data-grid model, in which the processed data are made nearly immediately available to remote sites where computing resources are available. Data-distribution tools have been consolidated by means of a global replica catalog that makes differential inventories between sites within minutes, and the development of in-house tools for reliable data transfer and redistribution.

10.2 Key Science Drivers

10.2.1 Instruments and Facilities — RCF

The RCF at BNL is the T0 center for the STAR experiment. The accumulated data is stored on mass storage (HPSS) at the facility.

BNL hosts all RHIC experiments. The facility's main operation and role is to provide the core CPU computing cycles for half 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 more than 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 (with 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 (DAQ) system is capable of streaming 500-600 MB/sec of data to an online disk cache. The reconstruction of our events does not change the size significantly, as STAR had planned to reduce its data demand by removing the need to save the event format files - since 2010, we save all of our microDST files, a format ready for physics use and a factor of 5 smaller than our event files, and one-tenth of our event files for physics verifications. During Runs 10 and 11, we showed that 600 MB/sec to mass storage was possible, the facility being largely able to absorb our data rate.

10.2.2 Process of Science — RCF

STAR's typical workflow consists of the origin of our data, acquired by the 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 almost one MB depending on luminosity, energy and violence of the collision, colliding species, and data-acquisition triggers (condition to accept an event on tape). In 2011, we assumed that a 600 MB/sec store to a mass-storage system (HPSS), directly from online event buffers, was desirable (see estimated requirements in Section 10.2.2.1). Offline, a quality-assurance process (a.k.a. Fast Offline) pulls a fraction of its data out of HPSS (while the data are still on HPSS cache

disk) and processes it for quality-control purposes on the RCF resources. Up to 15% of the data are pulled for inspection and the results of event processing are placed onto live storage (NFS mounted disk) for the quality-assurance team and other detector subsystem experts to mine and verify its quality. Its lifetime is short (two weeks); old data are deleted and replaced by new data. The data are also used to forgo incrementally more precise calibration passes.

Data sets collected for a given year 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 T0). When data productions are started, typically at the T0 center at the RCF, two + 1 copies are handled by each job, submitted to a locally engineered queue system based (mainly) on the principle of one job equals one input DAQ file (a few outputs are created). Each production job places one copy of the output(s) in HPSS and one copy on the central disk (NFS). 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 (cataloged in our replica catalog), allowing other data-management tools to access the files. STAR/BNL's data management also places a copy of the physics-ready files (a.k.a. microDST 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 staff (several groups make use of Scalla/XRootD at the BNL/RCF). The replicas are also indexed in our replica catalog and we 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 T1 centers also have relied on the BNL data-management system to make differential inventories of data they may need.

10.2.2.1 Projected Data Rate and Data Size

Because our network requirements are all derived from 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 and the state of our R&D. Two important factors will be treated separately and folded into our estimates: the RHIC luminosity increase and the STAR strawman Beam User Request (BUR).

10.2.2.2 Event Size Discussion — Effect of Luminosity

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

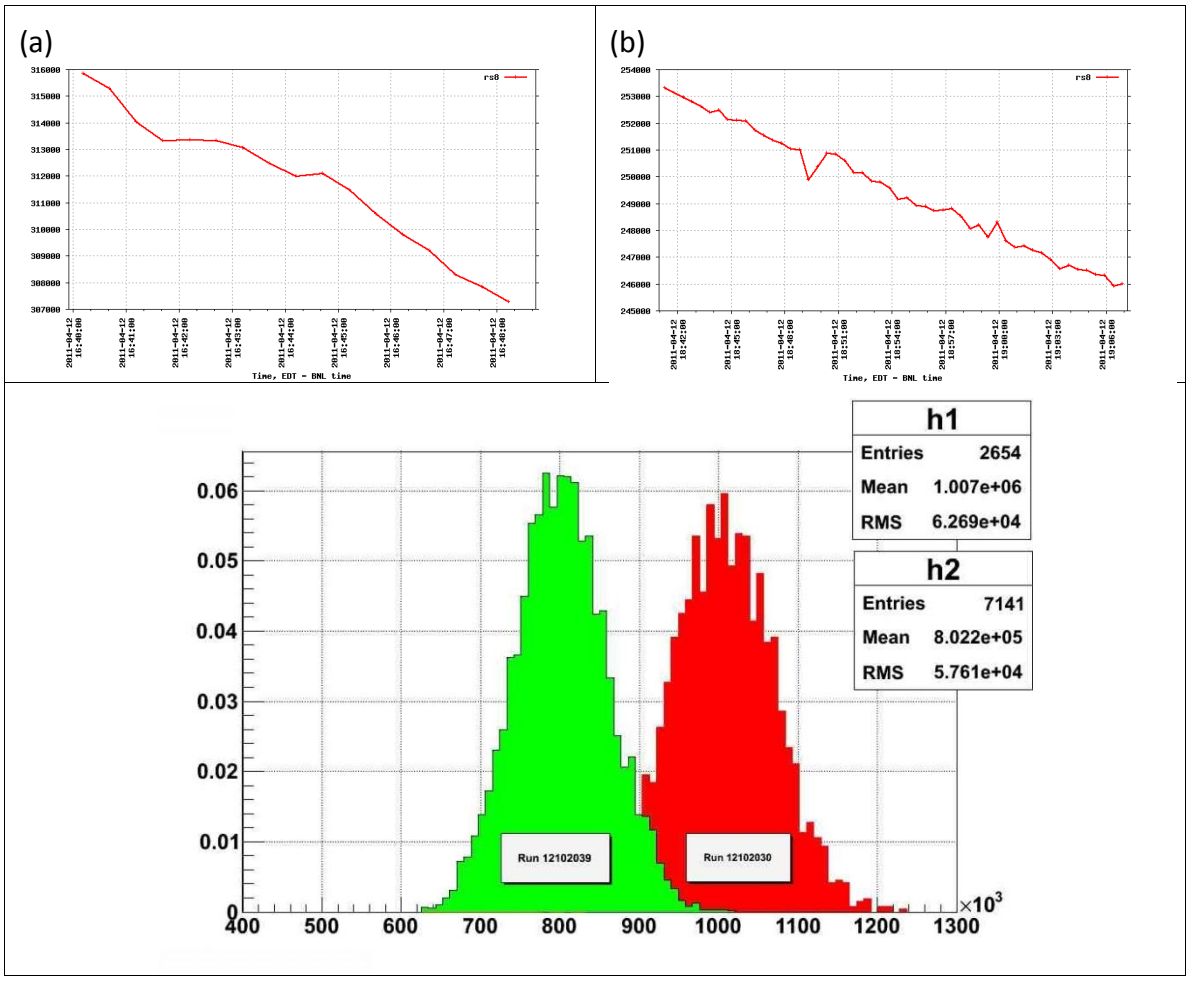


Figure 10-1. Distribution of event size for two run numbers, 12102030 (Tue Apr 12 18:40:46 EDT to 19:07:06 EDT) and 12102039 (Tue Apr 12 16:39:44 EDT to 16:48:38 EDT). Panels (a) and (b) show the related ZDC rate proportional to the RHIC beam luminosity; both runs were taken under the same trigger conditions; the X axis is in bytes. This figure shows a direct correlation between the average event size and the beam luminosity.

Figure 10-1 illustrates the correlation between the average event size and beam luminosity — while Runs 10 and 11 luminosities have allowed STAR to keep its event size to a conservative ~0.6 MB/event, both runs represented have event sizes exceeding this average estimate. Run 12102030 approaches a peak estimate of 1 MB/event. In Run 11, the lower-left tail end of the distribution was selected via pileup rejection methods but we expect an increase in peak luminosity to 4-5x those seen during the time Run 12102030 was taken (3x average), implying a possible net event size up to 4-5 MB/event (3 MB/event average). Considering the RMS of the event size distribution, the probability of being able to select only the 1 MB/event events during a p+p run is negligible without additional innovative event selections and mechanisms suppressing the effect of event pileup.

It is noteworthy that if we successfully reach a point where High Level Trigger (HLT) algorithms, such as the Cellular Automaton seed finding (a multicore 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 similar to what STAR performed for its online clustering algorithm to make sure the physics is not compromised. Potential size reductions on the order of 2 for the TPC detector may then fold into our estimates (although it is likely the gain will be used for increasing the data samples in number of events, several physics analyses such as the dileptons and Ds and any rare probes two particle analysis have their statistical precision directly tied to the number of available events).

In our calculations, we will fold in factors of x1.41, x1.73, and x2 event-size increases for p+p events in 2013, 2014, and 2015, respectively, 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). In the out-years we will retain 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.

10.2.2.3 Projections up to 2015

Although a year-by-year run plan and Beam Use Request (BUR) over the period requested is hard to predict with accuracy (a two-year exercise is done and reassessed every year at BNL through a formal process of case presentation to a Program Advisory Committee or PAC), Table 10-3 summarizes our current knowledge.

In 2014, STAR plans to take a large-enough Au+Au sample to study the Heavy Flavor Tracker (HFT) detector subsystem response and event reconstruction (the detector is assumed to become physics usable at the latest in 2015 but a prototype will be installed in 2013 and an early installation targeted for 2014). The physics goals would include high-precision measurement of the charm cross section and a possible first look at charm flow for preparing for the high-precision measurements of the later years. The span of 2014-2016 will be driven by the HFT program and the charm program - as well as a large p+p run for reference samples in 2015. The higher number in 2013 (compared with our past planning) is due to the fact that our older plans assumed the low energy scan program would take place mostly in that year while our Runs 10 and 11 BUR allowed for some of the beam energy scan to already take place. Beyond 2016, STAR is likely to begin the onset toward the eSTAR program, adding a detector in the forward region (this period is not part of this document).

Table 10-3. Projected data for 2012 to 2016. The two first years are shown as a cross check and trend projections purpose for out-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 but focus on the likelihood of an early HFT installation with the 2013-2014 time frame. A higher deviation to plan factor is hence set for 2013 and 2014 (where/when the HFT detector system for STAR is estimated to be integrated) to provide a safety margin to our projections. The size of the raw data is increased proportionally to the new detector to be phased into the STAR system (+100 kByte for the FGT and +300 kByte for the HFT) in 2012 and 2014, respectively, while the following years anticipate a size decrease due to the phasing out of non-zero suppressed data in a few subsystems (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 Table 10-3. Similar trends affect 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.

10.2.2.4 LAN Requirements, Transfer from Online to Mass Storage (HPSS)

From Table 10-3, we derive the LAN need, shown in Table 10-4. A 20% margin was added to account for possible TCP protocol overheads and miscellaneous transfer problems that could cause lags. This is summarized in the first row as gross average.

Table 10-4. 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, the 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.

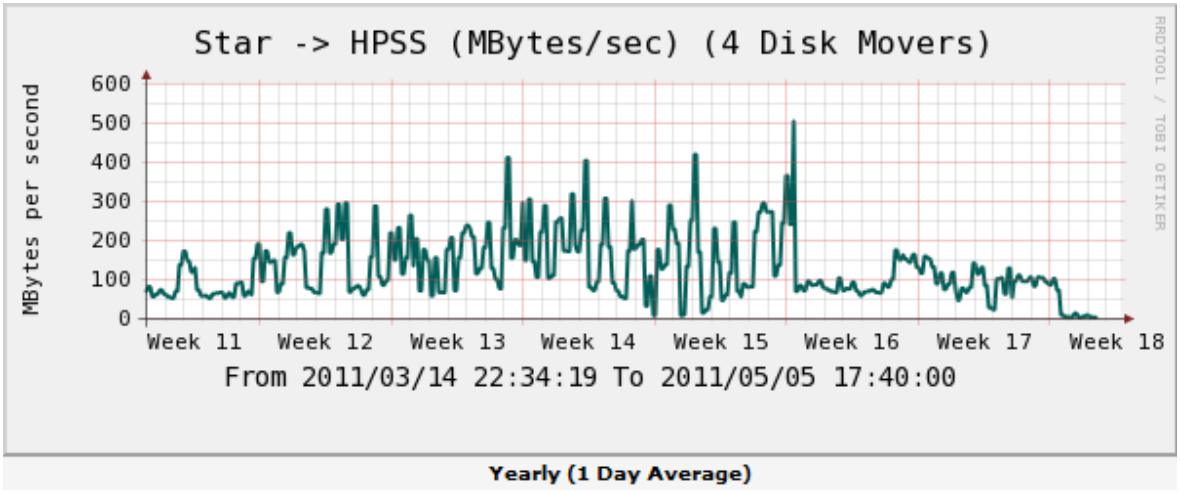


Figure 10-2. Data mover statistics from STAR online to Mass Storage (HPSS) at the RCF. Note the peaks at 500 MB/sec. Over the run period, we had averages of 250 MB/sec, consistent with our estimate (modulo the 20% overhead) of ~ 300 MB/sec of Table 10-4.

While this estimate tends to correlate with the usage verification for the past two years, previous estimates that considered fluctuations across species, run weeks, and run efficiencies showed needed rates as high as 550 MB/sec at real peak times. Modulo appropriate use of online disk caching, an ultimate 550 MB/sec would account for "surge" of data rates and was shown to be feasible. For 2011, an example of DAQ rates is showed on Figure 10-2 - real peaks are at ~500 MB/sec. The last row of Table 10-4 shows the LAN line rate in Gbps.

10.2.2.5 WAN Requirements, Transfers of Micro-DST to Other Tiers

Table 10-5 summarizes the network needs for transferring the physics-ready microDST to other Tier centers. While T1 should have a full copy of the MuDST, which could be moved over large periods of time (as produced), SACs tend to transfer data a-posteriori, by bulk, and depending on local analysis demands. When they need a data set, they typically expect a fast turnaround for data movement (within days) and a "we will take all you give" approach to network bandwidth. Hence, we assumed full data-set transfer over a period of three months for a T1

center and a one-week delay for a SAC or a T2 (the demand was typical of the observed pattern for our Prague site; data-set rotation, i.e., the replacement of the local data by a brand new set, would happen around four times a year).

Table 10-5. Data transfer rates for sustaining redistribution of microDSTs 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 two-thirds of our institutions would acquire the data from BNL while one-third would do this from NERSC. This is a target goal (but not representative of today's habits — nearly all institutions in need of data sets take them from the BNL/RCF's mass storage as sole "trusted and complete" source for data sets).

10.2.2.6 WAN requirements, Monte Carlo simulations

While extremely CPU intensive (full simulation or "slow" simulators can take as long as 30-45 minutes on a 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-hour process would generate an output of about 100 GB, 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 Storage Resource Managers (SRMs) to decouple CPU slot usage and output data transfers back to the BNL/RCF.

10.2.2.7 WAN Requirements, Embedding Support

STAR's only stable T1 center to date has been the NERSC/PDSF resource facility. STAR's principal uses for PDSF are to sustain some of its users' analyses, access 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 10-6 with the justifications as follows.

Table 10-6. Network requirements for sustaining the embedding process at one T1 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 size increase of the data sets has not caused a proportional increase of the need to transfer raw DAQ files to NERSC. Only 10% (or less) of our data contain information necessary for handling this kind of production: Those files contain “raw” signals while the rest of our data contain already formed track hits or clusters (online clustering is performed to reduce the DAQ output size). The first line of Table 10-6 estimates of the percentage of data needed for our embedding operations: Higher percentages in 2013 to 2015 account for the introduction of new detector subsystems, which may require enhanced simulations to understand them fully.

Embedding productions require long preparation; typically, the 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). Note that the output tends to be larger by a factor of 7 than the input, as many files are generated: event files, microDST files, and Geant association files are all needed for efficiency corrections.

Finally, while Table 10-6 tends to suggest a one-time embedding transfer process, several run-years’ worth of data are handled simultaneously (and often, the data sets 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.

10.2.2.8 WAN Requirements, Cloud and Data Preservation Scenarios

As part of the ESnet workshop, we would like to present the idea of processing up to 20% of our data on a National Laboratories cloud facility. This would allow outsourcing the cycles needed for our Fast Offline QA process, quoted as approximately 15% of our data in section 10.2.2.1 or an equivalent “emergency” production of data in a fast turnaround manner. While not yet

considered a regular workflow (cloud facilities not being guaranteed), such an operation would dramatically enhance our physics capabilities and possibly allow a better use of remote facilities that otherwise would not be accessible to STAR. For example, none of our software was installed at Argonne National Laboratory (ANL) but we could deliver a stable production stream on ANL's Magellan cloud within a virtual machine facility.

Table 10-7 summarizes the additional bandwidth requirements for this idea to become feasible at levels of 20% (Fast Offline) and 50% of the data.

Table 10-7. 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

The STAR collaboration is exploring the possibility of leveraging the Hadoop file system and the Google MapReduce paradigm for data processing on distributed resources. While in its infancy, if such exploratory work reveals a means to better exploit and tighten storage and computational resources, more cloud-like approaches may appear for the sake of efficient use of resources.

While exploring additional network paths, we are not considering the full transfer of all our raw data sets 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 T10KC cartridges (5 TB at first generation, up to a projected 40 TB storage per cartridge), will put STAR at risk of losing all early years' data or a large fraction of recently targeted data sets (a low-energy point sample, for example, with the loss of a single cartridge). From this observation, two scenarios are offered as natural solutions: (a) Double the storage at the BNL/RCF center (replicate each tape onto another, HPSS allows dual copy) or (b) move an entire data set 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 additional geographical data safety (two disconnected centers are unlikely to accidentally lose data at the same time). The combination of both models is 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 logistics of economics and the understanding of how to securely provide long-term archival capacity at NERSC, the last row of Table 10-7 shows the bandwidth

needed to achieve this plan (this would allow streaming data from online to off-site over the run period).

10.2.2.9 WAN Requirements, Summary from a BNL/Tier 0 Perspective

Table 10-8. Requirements totals — grayed cells indicate passed or unlikely achievable scenarios.

	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 10-8 shows the total requirements, considering all factors previously explained and roughly broken down by Tier center levels. The three last rows are T0-centric and consider respectively [A] the basic network needs for a "standard' STAR workflow, [B] the same adding 50% of one full production-level pass of data analysis on a cloud-based operation, and [C] a similar workflow as in the previous line, adding additional network traffic to account for the other half of our data to be transferred to a T1. Adding the numbers in a linear manner would not be adequate (peak requirements represent 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 half 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 half of our data to a cloud-based operation located at (for example) NERSC, we could store the data on mass storage as part of the same workflow and only have to transfer the other 50% to achieve full data-set safety and preservation (but without processing). We do not show the requirements this would impose on the T1 center (it would follow a similar arithmetic guided by our numbers from Table 10-7 and Table 10-8). Note as well that STAR has modeled its processing needs 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 data-set saving at a remote facility).

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

10.2.3 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 national and international collaborators. While the LHC/ALICE experiment usage is ramping up, STAR remains the top user at NERSC/PDSF.

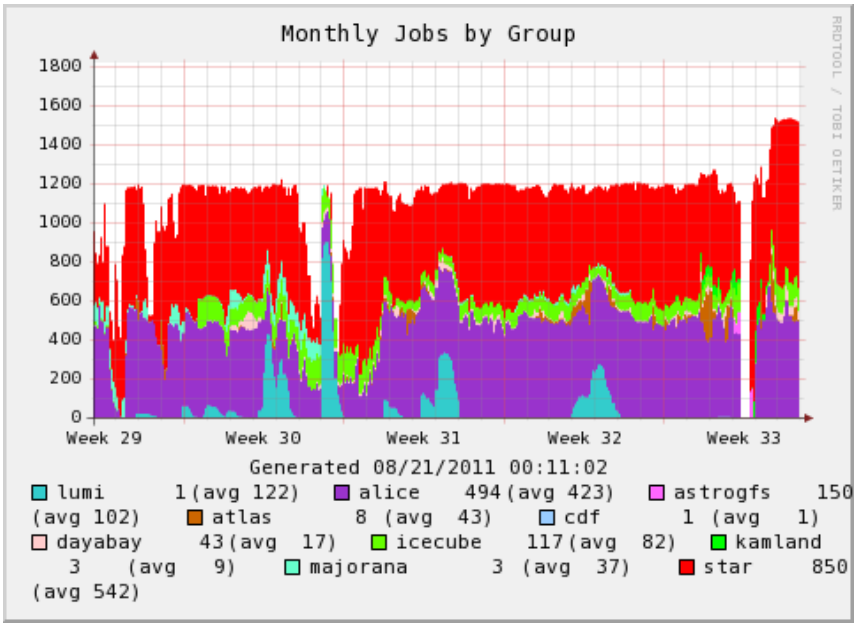


Figure 10-3. Batch queue usage at NERSC/PDSF by group. STAR remains the largest user and shares near equal resources with LHC/ALICE.

The program's main physics thrust 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 actively supports the STAR central detector, the STAR TPC, and is also heavily involved in the STAR detector upgrade program, with a focus on the HFT. The presence of data close to the local research group makes the STAR Berkeley group a vibrant research collaborator.

The T1 currently supports the deployment of the STAR software library releases and manages bulk data transfer to and from BNL. Through its library releases, the single-framework STAR allows PDSF users to perform data analysis and provides 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[7] tool) with local

[7] Fast Data Transfer, FDT: http://monalisa.cern.ch/FDT/

catalogs showing what data sets are available for local analysis. The replica catalog is now global, allowing any site to query and inventory the full set of replicas.

Data analysis is entirely based on locally available data sets at this point, but there is discussion on whether to leverage the deployment of remotely accessible Scalla/XRootD service to share data from the NERSC/PDSF and the BNL/RCF.

10.2.4 Process of Science — NERSC/PDSF

The work flow for embedding simulations, typically carried out at the NERSC/PDSF, consists of complex simulations that combine simulated tracks embedded into raw data signals (serving as a background). Our code ability to reconstruct the embedded tracks and identify them as close to the original track characteristics is directly linked 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, and correcting for efficiencies or quoting uncertainties on our physics results. This process requires a portion of DAQ data to be transferred from the BNL/RCF to the NERSC/PDSF and a copy of the resulting outputs to be brought back to the BNL/RCF. The bandwidth requirements for those transfers were explained and presented previously. The numbers will drive our network need estimates in the next sections.

10.3 Remote Science Drivers — NERSC/PDSF

10.3.1 0-2 Years Case

The operations at NERSC should remain standard, a balance among 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, whose impact on WAN requirements will need to be studied and understood (we have no practical experience for how much data may be pulled from one site to the other front at this time). However, user analysis varies from 100 Hz event data consumption to a second event reading and from Table 10-3, event sizes spanning from 0.36 MB to 0.61 MB, leading 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 2,000 slots at BNL reading data from remote would imply a 7 Gbps transfer if no data 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-year period, we expect our BNL/RCF T0 network requirements to follow a standard "general support" requirement (scenario [A] from Table 10-8) at a maximum of ~3 Gbps WAN bandwidth needs while NERSC/PDSF T1 center will require of the order of 2 Gbps connectivity to BNL for sustaining STAR science.

10.3.1 2-5 Years Case

Large data samples, driven by precision physics topics (with key players at our Berkeley and international colleague facilities) and the possibility to 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" off-site processing scenario [B] (technically possible today and already demonstrated by STAR) would be a path to follow should opportunistic resources become available. By then, the lifetime of the PDSF will also be questionable (economy of scale), making a virtualized operation even more likely. Assuming this direction for 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 in scenario [B] in Table 10-8. Those numbers remain at ~5-6 Gbps maximum for a BNL/RCF connection to NERSC/PDSF at ~3-4 Gbps.

By then, and if this scenario is possible, it is likely the currently run Open Science Grid (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).

10.3.1.1 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 BNL possibly phasing into eRHIC. We view the NERSC/PDSF mass storage as a natural place for another full, integral copy of our data.

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

10.3.2 Instruments and Facilities — Prague / SAC or Tier 2 Case

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

10.3.3 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 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 Gbps dedicated, this line was dropped to about 140 Mbps

throttled as illustrated in Figure 10-4. With the change of network bandwidth, the group has adapted to the new reality and shifted its 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, and publication material. The group ensures both data safety and resilience (a complete collapse of networking would allow it to continue to work locally). Proactive feedback and surveys helped the group to understand needs and the shift of scheme emphasized that connections latency is the show stopper and the tipping point for remote centers.

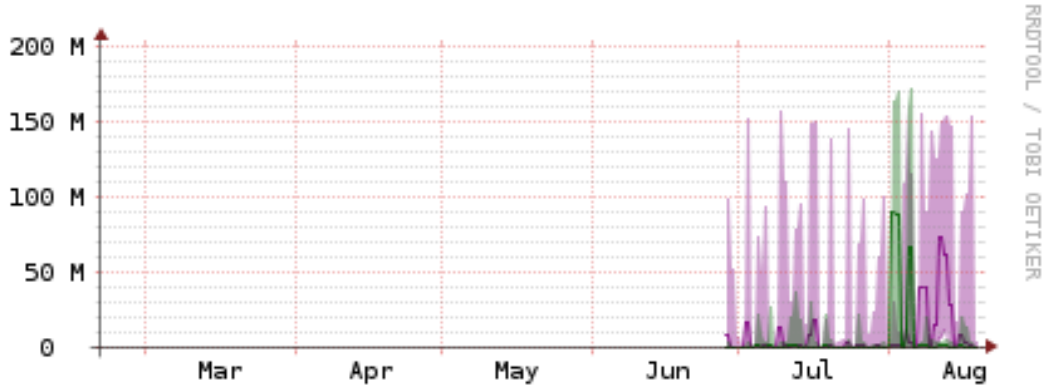


Figure 10-4. Network data transfer to Prague from BNL. The saturation at ~150 Mb/sec is due to a bandwidth throttling made in 2009.

Typically, a SAC or a T2 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 I/O).

Prague has also been heavily involved in the development of theoretical computing models (based on constraint programming or mixed-integer programming) and the development of data planners to enhance data transfers and leverage the presence of data sets from multiple sources (data sources as well as sites) for the most efficient data transfers to a destination. Such a new approach may dramatically change bandwidth requirements. Preliminary studies in STAR on the use of such data planners (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 10-5, the planner 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.

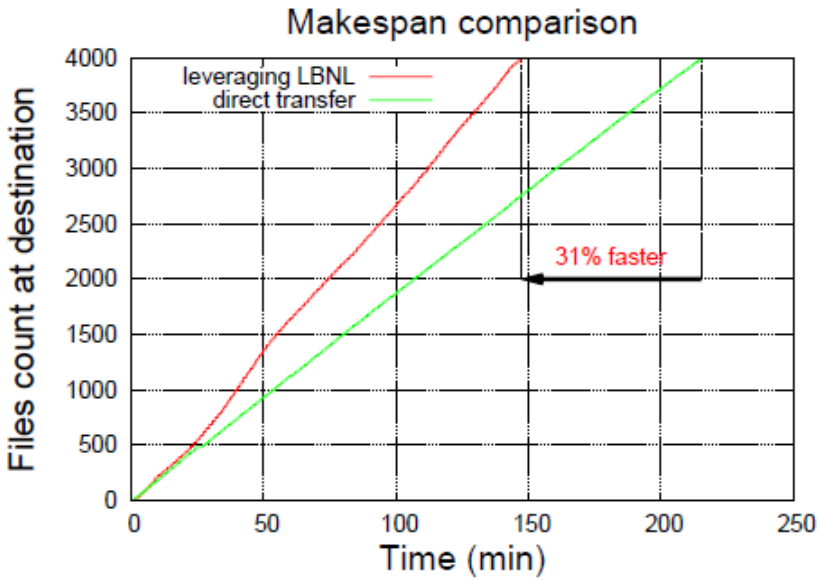


Figure 10-5. Moving the same number of files from a selected data set, makespan comparison between a direct site-site transfer and one leveraging two sites with an independent network path to the destination.

This methodology certainly holds nontrivial consequence for network bandwidth usage. With its dynamic discovery feature of available path to a destination and features such as an advanced network bandwidth reservation and/or predictor, a resulting comprehensive data transfer and management solution may very well allow saturating unused or little-used network segment experiments we would not otherwise discover.

10.4 Remote Science Driver — Prague / SAC or Tier 2

10.4.1 0-2 Years Case

Table 10-2 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 STAR data management system), we believe the profile will remain standard with no surprises.

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

10.4.2 2-5 Years Case

We feel the SACs and T2 centers, with “immediate” need (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 Gbps estimate. This number was given in our summary, but the U.S. funding and science profile may create conditions for short windows of opportunity for researchers to harvest the science of the RHIC-II era. This would ultimately pressure remote sites to stress the T0 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 remains by then such a thing as an unused or little-used network path).

10.4.3 5+ Years Case

The requirement will remain stable at a 2.5 Gbps link speed maximum.

10.5 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 (mostly EVO and Skype-client based communications which are free, easy, and versatile). EVO has had its share of dropped connections and barely audible remote speaker, perhaps related to the use of a common network path, which is 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 that integrates with grid authorization services would 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 recently the FDT[8] tool. STAR has developed data planners in house and deployed them over a few sites in test mode (see above for a quick description) but this tool essentially relies on FDT to actually move the data.

10.6 Issues

- We remain convinced that the distribution of our physics-ready data sets allow for 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 intended to be fully automated, the need to verify the validity of data productions over larger samples and our current inability to invalidate data sets placed at remote sites caused delays in data redistributions. They are typically not done synchronously to data productions but after a physics evaluation period (which may take months). As a result, higher bandwidths are needed over short periods of times and, considering workforce at remote sites, we do not see this modus operandi as changing soon.
- The Birmingham institution’s move away from STAR has to some extent slowed our plan to expand our embedding operation and outreach to other T1 centers. However, STAR can now balance (in emergency situations) the workload between BNL and NERSC for this kind of operation. Potential new SACs from the United States include our institutions in Texas.

[8] Fast Data Transfer, FDT: http://monalisa.cern.ch/FDT/

- STAR is 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 Monte Carlo simulations (not requiring real data as input, that is, unlike embedding) are carried on grid resources (emergency running may be carried at BNL). However, only STAR's already dedicated sites (those for which NP funds a base resource and hardware for STAR) are used, as the heterogeneity of grids has made running complex work flows practically unachievable — the use of pre-installed software packages takes care and attention for full reproducibility and perfection of science (full validations can take several days and troubleshooting remains difficult on the 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 recently made massive use of the Magellan cloud facility for raw data processing[9,10]. 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 very little storage space at BNL, the NERSC cloud facility, and the ANL cloud facility, respectively. 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 presentation of 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 Matter 2011 conference. This truly opportunistic mode of operation showed that (a) STAR is able to and equipped to run the most complex work flows on distributed (virtualized resources) and (b) the use of burst resources (availability of elastic resource) remains fundamental to the ability of an experiment to treat, under heavy load and demand, physics cases that would otherwise be dropped. The consequence on networking is, however, nontrivial: Set at National Laboratories, the Magellan cloud has a predictable network path; true commercial clouds do not (perhaps suggesting a strong case for continuing to sustain National Laboratories cloud-based infrastructures).

[9] Magellan Tackles Mysterious Proton Spin, NERSC Science News
[10] The Case of the Missing Proton Spin, Science Grid This Week, June 2011

10.7 **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 • Physics-ready microDST (MuDST) transfer to NERSC/PDSF • Partial delivery of MuDST to ~ 3-4 SACS & T2 centers • OSG use for simulations • Embedding simulation support at NERSC/PDSF • Estimated totals	• 1.4-1.9 PB/year • 800-1000 TB, 400-500 k files • 80-100 TB or less in burst • 10 k files, ~15-20 TB and of the order of 100 k files and 90-130 TB results back to BNL	• 400-500 MB/sec - peak at 500 MB/sec	• 1-1.5 Gbps transfer to NERSC/PDSF over 3 months • SAC support @ 1.5 Gbps, data moved within a week • ~ 2 Gbps in/out of NERSC and BNL • NERSC 2 Gbps and BNL 3 Gbps
2-5 years				
• RHIC/STAR data taking, Heavy Flavor program, local and distributed data production	• Data taking • Distributed infrastructure-based simulations and Fast-Offline (50%) — cloud-like • MuDST copy at NERSC/PDSF • MuDST delivery to 3-4 SACS & T2 centers • Embedding simulation support at NERSC/PDSF • Estimated totals	• 2.0-2.5 PB/year • ~ 1.0-1.2 PB during runs • 1-1.6 PB, 400-500 k files • 100-160 TB or less in burst • 20-22 TB input and ~ 160 TB output	• 550 MB/sec from online to RCF	• 3-3.5 Gbps for streaming data from online to remote site (“live”) • 2 Gbps Transfer to NERSC/PDSF over 3 months • SAC support @ 2.5 Gbps, data moved within a week • 2-3 Gbps in/out of NERSC and BNL • NERSC 3-4 Gbps and BNL 5-6 Gbps

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
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 data sets • 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

11 RHIC Computing Facility (RCF)

11.1 Background

Located at Brookhaven National Laboratory (BNL), the Relativistic Heavy Ion Collider (RHIC) program is a nuclear physics program composed of a world-class scientific research facility with complex detectors and an accelerator that drives two intersecting beams of gold ions head-on in a subatomic collision. In terms of luminosity in heavy ion collisions, RHIC is the biggest facility of its kind to date. It is becoming the world leader in the scientific quest toward understanding how mass and spin combine into a coherent picture of the fundamental building blocks nature uses for atomic nuclei. It is also providing unique insight into how quarks and gluons behaved collectively at the very first moment our universe was born. The main RHIC experiments - the Pioneering High Energy Nuclear Interaction eXperiment (PHENIX) (550 physicists from 67 institutions spread over 13 countries) and the Solenoidal Tracker At RHIC (STAR) (560 physicists from 55 institutions spread over 12 countries) - are collaborations spanning many countries and involving more than 1,000 collaborators.

Having reached petabyte-scale data recording per year (10^{12} bytes), the aggregate raw data rate envisioned by the RHIC experiment's program per run (or year) will more than double from currently ~1 PB to >2 PB per experiment in 2015, reaching an archival data rate of 1 GB/sec per experiment, which will make data management and data distribution an ever-increasing challenge. To face these challenges caused by the size of those data sets and the need to preserve the physics quality and turnaround, the RHIC experiments have adopted a distributed computing model or are using a model based on the combination of dedicated and, whenever appropriate and available, opportunistic remote resources.

The computing and data-handling capacities required for the detectors at the RHIC are large when compared with previous detector systems in nuclear physics.

Certain aspects of the RHIC computing requirements are appropriately handled by a dedicated facility located at and under the direct management of the RHIC operations program. These aspects are associated with the handling and processing of the actual data produced by the detectors. Other aspects of the RHIC computing requirement, in particular those associated with theoretical models, event simulation, and certain compute-intensive or low data-volume types of analyses, are less critically linked to the operation of the detectors themselves and so can be done effectively at locations remote from the RHIC facility. The possibility of satisfying such needs at existing locations - such as departmental facilities at collaborating institutions or at regional or supercomputing centers - at substantial dollar savings to the RHIC project was and is explicitly considered by the collaborations. If adequate reduced-cost computing is not available elsewhere, the computing mission of the computing facility at RHIC is adjusted to address those additional needs.

The dedicated RHIC Computing Facility (RCF) at BNL has primary responsibility for handling and processing data produced by the experiments and operates in conjunction with computing facilities at remote locations and so requires high levels of WAN connectivity. The RCF is specifically responsible for the reconstruction of collider data and for recording and archiving

the raw and derived data, as the experiments deem necessary. RCF serves as a data-mining and data-serving facility for the raw and derived data and also functions as the primary analysis facility. In addition to managing and processing of large amounts of data, RCF supports data transfers to enable the large-scale theoretical modeling and event simulation that occurs mostly at remote sites. The storage of some data sets associated with simulation at RCF and the use of RCF resources for simulation work during periods of non-peak demand for processing collider data augments the offsite simulation work. Similarly, the export of various levels of processed data from RCF to remote facilities for later stages of analysis is facilitated by the RCF.

The BNL campus network provides high-bandwidth connectivity that supports many worldwide scientific disciplines. Main users of the network capabilities are PHENIX and STAR at the RHIC and A Toroidal LHC Apparatus (ATLAS) at the Large Hadron Collider (LHC). These two programs account for the majority of the network bandwidth consumed within the BNL computing environment.

For WAN connectivity, BNL is currently provisioned with four 10Gbps circuits, which are divided into two distinct classes of service for the user community. First, classical Internet Protocol (IP) connectivity is provided by a single 10Gbps link to the Internet via the Energy Sciences Network (ESnet). This link provides the default connectivity between most external scientific facilities and BNL. Secondly, ESnet provides three 10Gbps links that primarily support the Science Data Network (SDN) bandwidth requirements between BNL and the LHC Tier 0 (T0) center at the European Organization for Nuclear Research (CERN) and Tier 1 (T1) sites around the globe, and between BNL and four of the five U.S. ATLAS Tier 2 (T2) centers at universities and at the SLAC National Accelerator Laboratory. As to nuclear physics applications on these links, there is a 1Gbps circuit between BNL and the Nuclear Physics Institute (NPI) ASCR in Prague. The SDN circuits are purpose-built, end-to-end connections between dedicated computing resources at BNL and the corresponding peer scientific institutions. The primary link between BNL and CERN is split over two of the 10Gbps links to enhance throughput and reliability. Additionally, a backup link to CERN is provisioned. To support site redundancy, any of the operational links can be reconfigured to transport any or all network traffic types.

Note that there are two additional 10Gbps circuits (provisioned out of a bundle of ~140 new fiber pairs between BNL and ESnet's peering points in Manhattan, and other locations on Long Island) currently in the commissioning phase. Both will be used in the context of LHC computing to improve domestic and international connectivity.

This status update focuses on major upgrades and enhancements since the previous report from 2008.

11.2 Key Local Science Drivers

11.2.1 Instruments and Facilities

The RCF at BNL provides the majority of computing power and storage capacity for the currently active experiments at RHIC (PHENIX and STAR). The facility is large in absolute size, and in relative size when compared with other computing centers that support high energy and

nuclear physics experiments. As to network connectivity, RCF uses ESnet, which peers with other domestic and international R&E and commercial networks.

By 2013, RCF will have more than 9 PB of disk space in production and 100 kHS02 of processing power (we measure processing resources in thousands of HepSpec 2006 [kHS06], which is based on SpecInt 2006). We expect archival storage volume to grow to more than 20 PB.

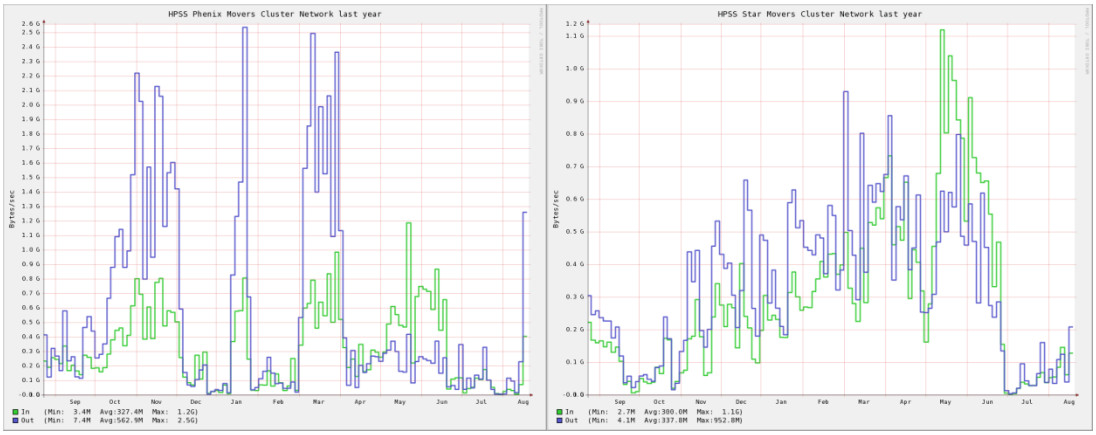


Figure 11-1. Mass Storage (HPSS) I/O for PHENIX and STAR (last 12 months).

At RCF, PHENIX and STAR virtualize the large number of physical storage devices into a storage system by using dCache and XRootD. The LAN traffic between the distributed disk servers (disk-heavy worker nodes) and processes running on worker nodes is on average between 1-4 MB/sec per processor core, which corresponds to 5-15 GB/sec (40 Gbps-120 Gbps).

Extensive upgrades to the internal BNL networking infrastructure were conducted over the course of the last two years. Most of the inter-switch links were upgraded to a 100Gbps ether channel. The multilayer-based distribution system that supports the RHIC and ATLAS enclaves has been upgraded to Force10 switches. New high-performance, multi-petabyte storage appliances with line-rate 10Gbps interfaces have been incorporated into the network. An additional Force10 multilayer switch has been installed to support a high-density compute farm. To increase the number of available compute nodes in these clusters, three additional BNL subnets have been allocated for their use.

Lastly, the internal VLAN connecting data servers to the BNL perimeter routers has been separated from other internal BNL network segments. This cut-through architecture eliminates two internal BNL hops to reach these data servers.

11.2.2 Process of Science

There are many well-defined computing functions associated with RHIC data analysis. A variety of types of data must be recorded and stored. In some cases, the recording is of an archival nature, in the expectation that the data will rarely, if ever, be accessed again. In other cases, the data are recorded and stored in the expectation that the information will be frequently accessed and the ease and speed of access is of critical importance. Large-scale data sets are recorded where produced. Thus, the raw detector data and data derived from the reconstruction pass are recorded at RCF. The primary output of the reconstruction pass,

historically called DST-level data, requires more frequent and immediate access and is usually found physically on robotic tape libraries. Relatively small, highly distilled subsets of the data, historically called DSTs or ntuples, are produced by selection passes performed on the DST data, a process referred to as "data mining." This component of the data is in general recorded and stored local to their production but is frequently replicated and in some instances uniquely stored at remote sites, including individual workstations, departmental facilities at collaborating institutions, and regional or supercomputer centers. This type of data in the same logical store as the raw and DST data is physically found on disk because of the need for very frequent and fast access as final analyses are being performed.

Event Reconstruction. Event reconstruction is the process of transforming the raw detector data into physics variables. This is generally the single most compute-intensive aspect of the data processing. The primary result from the reconstruction process is usually a DST. The reconstruction of all collider-produced data is in general (STAR sent a fraction of the Run 11 data to cloud resources at NERSC for Fast Offline QA processing) performed at the RCF. Reconstruction of simulated events produced to understand detector performance issues are performed at the site that produces the simulated events. When the reconstruction capacity at the RCF is not saturated by reconstruction of collider data, the unused compute cycles can (and actually are) applied to such simulated data as well. However, RCF is not sized to perform the reconstruction of simulated events in parallel with the reconstruction of collider data.

Physics Modeling. To interpret results, it is frequently necessary to compare signals observed in the collider data with the signals produced in the detector by events corresponding to a particular physics model. The generation of such events can require large amounts of computing capacity. This type of computation is typically performed at departmental facilities at collaborating institutions and at regional and other centers. Again, while RCF is capable of doing such work when not saturated by collider data, it is not sized to perform this function in general.

Event Simulation. Event simulation refers to the computer simulation of the response of a detector to an event or particle. Such simulations are required to understand the response of the detector. The most common issue addressed is the acceptance of the detector. This frequently requires the production of numbers of simulated events comparable to the number of actual events of a particular type observed in the detector. Depending on the details of the simulation, the required computer time to perform such a simulation can range from being relatively small to being much greater than the time required to reconstruct an event. Such simulations are done at remote sites such as regional centers.

MicroDST Production. The production of a microDST is most generally accomplished by making a pass through a DST data set applying criteria to select events and objects within events. The resultant micro-DST then consists of the subset of objects of interest from the subset of events of interest and is thus much smaller and more easily accessed during later repetitive stages of analysis. MicroDST production generally requires a relatively small ratio of CPU to I/O and is thus generally limited by the bandwidth and specificity by which the DSTs can be accessed. RCF is (intended to be) the primary site for such microDST production and the facility is scaled to meet requirements in this area. Certain regional or other centers may choose to locally store

subsets of the DSTs and so may also have microDST production capability for some types of data.

It is also possible to produce additional microDSTs from existing microDSTs. This is frequently the case in constructing final very selective data sets.

Often the final very selective summary of the data is in the form of an ntuple. RCF is explicitly intended to perform such functions but, when the storage and compute cycle needs are in a reasonable range, it is recognized that these functions may be done remotely, for example using departmental resources at collaborating institutions.

Analysis. Once a final highly selected data set has been identified, the analysis process of studying the physics significance of the data is typically performed by repetitive passes through the data set. These passes consist of calculating additional objects of physics significance; applying various additional selection criteria; plotting distributions; and numerically and visually comparing and correlating signal, background, acceptance, and theoretical model distributions. Depending on the size of the data set and the scale of the computations required, these needs may range from those that can be satisfied on an inexpensive workstation to those requiring a large facility with parallel coordinated operations across many processors operating on large data sets distributed across many disks. RCF serves as a facility for such analysis (e.g., PHENIX's AnaTrain that aggregates tens to hundreds of analysis tasks running over tens to hundreds of TB) in the expectation that small-scale analyses are often performed on workstations at remote institutions. In addition, there are many large-scale analyses that require a major facility like RCF and the Parallel Distributed Systems Facility (PDSF).

11.3 Key Remote Science Drivers

11.3.1 Instruments and Facilities

Among the steps involved in getting from raw data to physics results, two involve resources external to RCF: event simulation and, to some extent, user analysis. In particular for STAR, we estimate that the resources (both storage and processing) needed for handling the Monte Carlo simulations are of the order of 15% of the disk space and 10% of the total processing resources required for completing a one-pass data reconstruction run. Starting in 2008, both event generation (Monte Carlo) and simulated event-reconstruction passes have been centrally managed using standard grid interfaces for job submission to collaborating sites or sites that offer resources on an opportunistic basis (e.g., via Open Science Grid [OSG]). Using grid or cloud interfaces makes resources available to STAR at various sites in a seamless, interchangeable fashion.

While the PHENIX experiment is managing and running almost all its user analysis at its RCF share using a mechanism called AnaTrain, at STAR high-priority data production has pushed analysis aside, thus reducing the resource share formerly devoted to user analysis. This has caused collaborators to independently seek additional resources outside those counted on and accounted for in the initial STAR resource planning for computing. In November 2006, through a survey of information from a diverse group of collaborating STAR institutions, it was estimated that the total capacities utilized for analysis (beyond those from RCF and PDSF) was at 40% of

what's necessary for one analysis pass. To serve the wide area bandwidth needs from RCF to the three to five STAR T2 centers, between 2 Gbps and 5 Gbps of network capacity is needed, depending on the number of T2 centers and the run scenario in a particular year.

Currently, BNL is serviced by a total of four 10Gbps links, of which three are leased circuits from Keyspan Energy Systems and one from OCG with connectivity provided by ESnet. To support survivability and redundancy, these links provide path diversity, with half of them traversing the North Shore of Long Island and the remaining two circuits strung along the South Shore. In the event of a circuit, router, DWDM system, or other hard failure, any of the remaining circuits can be provisioned to support either IP or SDN network traffic, although at reduced capacity. Finally, both the BNL and ESnet routers are configured with redundant secondary interfaces and multiple Border Gateway Protocol (BGP) peerings, which can detect most common failures and reroute around the defective components almost instantaneously and transparently to the applications.

Both BNL and ESnet staff have just completed the deployment of a "dark fiber" solution to meet both the current and long-term future WAN capacity requirements. As is BNL's standard practice, this project will provide redundant ring topologies along both the North and South Shores of Long Island into the BNL campus from two main hosting locations in Manhattan. As currently configured, the optical switch gear (Infinera) provisioned for the fiber deployment can support up to 100 gigabits per second by using multiple 10Gbps interfaces. Almost operational, this new fiber infrastructure will provide BNL with two additional 10Gbps circuits for the SDN. A third component of the dark fiber project is part of a 100Gbps "test-bed network," initially intended to be used by computer scientists for advanced networking-related research projects. These projects are expected to enter an active state as soon as the test bed becomes available.

As the demand for dependable and interference-free connectivity between BNL and collaborating sites in the United States and abroad is constantly growing, BNL is making increasingly use of ESnet's On-Demand Secure Circuits and Advance Reservation System (OSCARS).

Each of the existing six circuits has been allocated between 10Gbps and 1Gbps (minimum) bandwidth, the latter with oversubscription capability for the idle bandwidth on the circuit. Two more 10Gbps (SDN) circuits have been provisioned by ESnet. The BNL network group is currently in the process of connecting the circuits to the perimeter routers. In case of an outage of circuits, the generic Internet path is used as a backup.

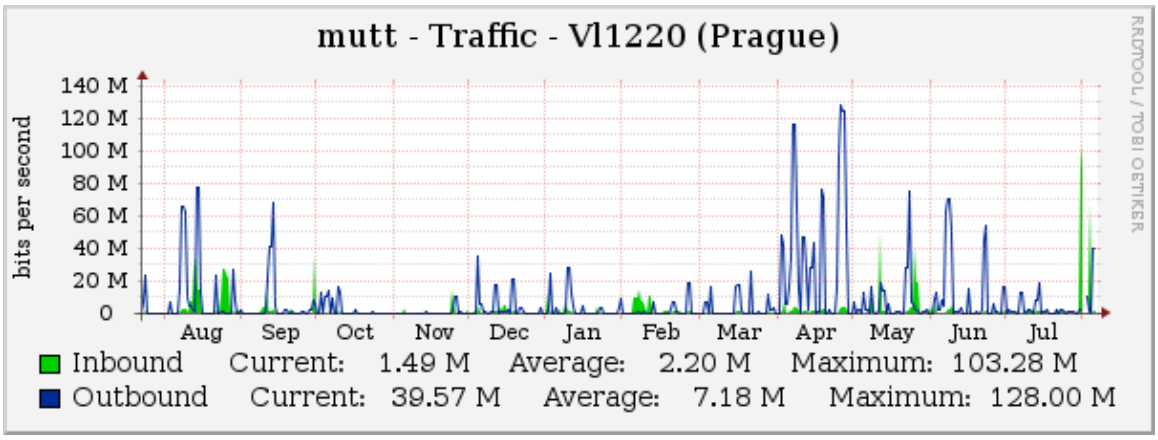


Figure 11-2. BNL to NPI/Prague circuit utilization.

11.3.2 Process of Science

The process of science at remote locations has a variety of forms. At RCF, the reconstructed data or a fraction thereof and more summarized analysis formats (DSTs and microDSTs) are served to PHENIX and STAR analysis sites in the United States and worldwide.

The scientific process mainly resides at the remote analysis centers, which are the bulk of the analysis resources for primarily STAR and to a lesser extent PHENIX. Smaller event samples are processed comparing the expected signal from the predicted background. In this case, the signal can be a source of new physics, or the standard model physics being investigated.

11.4 Local Science Drivers — The Next 2-5 Years

11.4.1 Instruments and Facilities

Table 11-1. Proposed scientific milestones for future RHIC runs.

	Year	#	Milestone
spin	2013	HP8	Measure flavor-identified q and $\bar{q}$ contributions to the spin of the proton via the **longitudinal-spin asymmetry of W production**.
spin	2013	HP12 (update of HP1)	Utilize polarized proton collisions at center of mass energies of 200 and 500 GeV, in combination with global QCD analyses, to determine **if gluons have appreciable polarization over any range of momentum fraction** between 1 and 30% of the momentum of a polarized proton.
spin	2015	HP13 (new)	Test unique QCD predictions for relations between **single-transverse spin phenomena** in p-p scattering and those observed in deep-inelastic lepton scattering
Heavy ion	2014	DM9 (new)	Perform calculations including viscous hydrodynamics to quantify, or place an upper limit on, **the viscosity of the nearly perfect fluid** discovered at RHIC.
Heavy ion	2014	DM10 (new)	Measure **jet and photon production and their correlations** in A≈200 ion+ion collisions from medium RHIC energies to the highest achievable energies at LHC.
Heavy ion	2015	DM11 (new)	Measure bulk properties, particle spectra, correlations and fluctuations in Au + Au collisions at √sNN between 5 and 60 GeV to **search for evidence of a critical point** in the QCD matter phase diagram.
Heavy ion	2016	DM12 (new)	Measure production rates, high pT spectra, and correlations in heavy-ion collisions at √sNN = 200 GeV for identified hadrons with heavy flavor valence quarks to constrain the **mechanism for parton energy loss** in the quark-gluon plasma.
Heavy ion	2018	DM13 (new)	Measure real and virtual **thermal photon production** in p + p, d + Au and Au + Au collisions at energies up to √sNN = 200 GeV.

Table 11-2. Proposed RHIC run scenarios and science goals.

Year	Likely Beam Species	Science Goals	New Detector Sub-systems	New Machine Upgrades	Gain from Machine Upgrades	Comments
FY12 (from PAC advice)	Cu+Au and U+U at ~200 GeV; 500 & 200 GeV p+p	Alter initial-state geometry for flow and E loss systematics; pp reference data for VTX; anti-quark and low-x gluon polarizations	PHENIX FVTX and μ trigger; PHENIX DAQ/trig upgrades; STAR FGT partial	Full yellow + blue horiz. stoch. cooling (6 planes in all)	Further heavy-ion luminosity improvements; U beam capability	First run with EBIS as RHIC heavy-ion source; tandems in back-up mode
FY13	200 GeV Au+Au; 500 GeV p+p	RHIC-II HI goals: heavy flavor, γ-jet, quarkonium, multi-particle correlations, etc.; transverse spin asymmetry for Drell-Yan (2015 spin milestone); continue W prod'n	STAR HFT prototype; STAR FGT complet'n; STAR MTD partial	Polarized source upgrade; Electron lenses	improved pp luminosity and beam polarization	Electron lens commissiong; "Proton cannon" increases pol. source current, to allow scraping to improve polarization
FY14	Heavy-ion runs motivated by earlier results	Continue pursuit of γ + jet, energy scan, identified heavy flavor (DM 10-12) milestones + further quantify QGP properties	STAR HFT and MTD completion	RHIC collimator upgrade; 56 MHz SRF; coherent e-cooling install starts in IP2	Full RHIC-II heavy-ion luminosity + improved vertex & store length	Detailed collimator upgrade plans still being developed

During the next two to five years, the RHIC machine and the PHENIX and STAR detectors will undergo significant upgrades, leading to increased luminosity and increased data rates from the detectors. The complexity of events, the event processing times, and the average event sizes will increase (e.g., the introduction of the Silicon Vertex Tracker [VTX] detector at PHENIX in Run 11 doubled the event size from Run 10, which could double again with the introduction of the Forward Silicon Vertex Detector [FVTX] in Run 12), but the operating models of the experiments that will be exercised in the next year will be recognizable in the next two to five years. Most increases in facility capacity for processing, disk storage, and archival storage will come from technology improvements, while maintaining a similar facility complexity. Processing and storage nodes will be replaced with faster and larger nodes, though the number of nodes should remain roughly constant.

RHIC plans to operate during 2012 at 100 GeV/nucleon in Heavy Ion (HI) (Cu+Au) operations mode and 100/250 GeV in polarized p+p operations mode. Several upgrades to the machine are yielding an increase in luminosity in 2014 or later from 30 to 40 10^{26} cm^{-2} s^{-1} for HI operation and from 24/90 to 60/300 10^{30} cm^{-2} s^{-1} for p+p operation. Polarization in p+p mode is planned to grow from presently ~50% to 70%.

As to the experiments, PHENIX has offloaded RCF from p+p reconstruction of 270 TB of raw data in 2005 by replicating it to the computer center of their Japanese collaborators at the Computing Center in Japan (CCJ). Given the number of actual events in recent runs and the expected number of events in future runs, the reconstruction times per event, and the actually available and expected compute capacity at RCF, the collaboration has at this point no plans to ship raw data files offsite. A few collaborating institutes, primarily CCJ, are asking for replicas of the smaller (~70% of the raw) derived DST data sets. As to PHENIX RHIC, runs with a substantial p+p component have an impact on wide area networking.

11.4.2 Process of Science

The PHENIX and STAR collaborations and the RHIC collider accelerator division are completing a suite of strategically targeted upgrades of moderate scope that promise to usher in an entirely new era of fundamental heavy ion and spin studies of extended scientific reach. These studies will build on the discoveries of the first phase of RHIC experimentation by using the increased luminosity provided by the upgraded RHIC II accelerator and implementing new detector instrumentation strategically targeted to enhance the detector's acceptance, particle identification capability, and effective sampling of luminosity. To capitalize on these investments, it is essential that the computing capability of the experiments, now and in the future, also be strategically positioned to receive and analyze the flood of data which the upgraded detectors will produce.

11.5 Remote Science Drivers — The Next 2-5 Years

11.5.1 Instruments and Facilities

At RCF, both collaborations will produce large samples when the data collected with increased RHIC machine luminosity and upgraded detectors are processed. The larger data products will need to be distributed for analysis. The samples selected by physics groups to be served to analysis centers (T2s for STAR) will increase in size as the integrated luminosity increases, but the time the physics groups are willing to wait is probably roughly constant so the network bandwidth requirements both for RCF to T1 and RCF to analysis centers will increase.

11.5.2 Process of Science

The changes in the process of science expected at the remote facilities is the same as the change described above for the local facilities. The centers will be performing similar actions at they do now except with larger data samples as the integrated data collected grows. The data collected in a few years will increase according to particle species and with complexity of the events.

11.6 Beyond 5 Years — Future Needs and Scientific Direction

We expect similar requirements as described for the two-to-five-year year period.

Table 11-3. PHENIX and STAR projected data-set volume and estimated WAN needs.

	FY08	FY09	FY10	FY11	FY12	FY13	FY14	FY15
PHENIX Data (TB/year)	590	595	1660	1420	2380	2210	1700	3400
STAR Data (TB/year)	172	230	1570	1930	2231	2224	4372	4095
Total Annual Data (TB/year)	762	825	3230	3350	4611	4434	6072	7495
Required WAN bandwidth (avg) (Mbps)	276	1500	<2k	<10k	<10k	<15k	<15k	<18k

Note the projections for wide area bandwidth for PHENIX and STAR are very different. From the aggregate WAN bandwidth in Table 11-3, PHENIX requires between 1 and 2 Gbps.

11.7 Grid Middleware and Cloud Computing

The RHIC/ATLAS Computing Facilities (RACF) are heavily involved in the deployment and operations of grid middleware, mainly through the OSG. The ATLAS T1 center at BNL, the largest and most successful of its kind for ATLAS, is hosted in the same facility as RCF and is operated within the same group. A tremendous amount of grid-related experience and expertise has been gathered over the course of the past five years that is fed back into OSG

operations and future evolution. At the facility, about 1,200 compute servers with 7,500 job slots and more than 9 PB of disk and 5 PB of tape storage are used through OSG in the context of Worldwide LHC Computing Grid (WLCG) operations by a variety of virtual organizations.

As to cloud computing, the facilities are building on the virtualization activities that are by now dominating service provisioning at RCF and ACF.

Our activities fall into five areas. The majority overlap with OSG's technology investigation interests.

- Basic infrastructure (hardware and software expertise)
- Running a cloud provider
- Enabling cloud-based clusters/grid sites/worker nodes
- Dynamic expansion of static site to cloud resources
- Complete job virtualization

Particular projects with some dependencies (three-month time horizon)

- Establish a cloud/virtualization test bed at BNL (15 nodes), each capable of running two VMs, including mechanisms for easy rebuilding/reconfiguration. (Done)
- Create an automatically deployable vanilla Condor cluster, configured via Puppet. (Done)
- Install and run an OpenStack-based cloud provider infrastructure on the physical test-bed nodes (begun at Nebraska, will expand at BNL). Test scalability.
- Establish and run Condor VM-universe capability on test-bed Condor cluster.
- Investigate running VM jobs via grid interface. Test scalability.
- Investigate the use of Lawrence Berkeley National Laboratory’s (LBNL's) CloudCRV system for cloud-based infrastructure management.
- Create and deploy VM image able to run jobs
- Deploy VM image with full grid clients, CVMFS-based software distribution mechanism (ATLAS Releases), and cached data stage-in.
- Create and deploy VM image able to act as Condor cluster worker.
- Investigate the use of BaBar's Cloud Scheduler software for on-demand expansion of static batch systems to cloud-based worker nodes.
- Create BNL-based RPM repository that could be sourced by VMs or used to build VM images.

11.8 **Summary Tables**

Table 11-4. RHIC Requirements Summary — STAR.

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 • Physics-ready microDST (MuDST) transfer to NERSC/PDSF • Partial delivery of MuDST to ~ 3-4 SACS & T2 centers • OSG use for simulations • Embedding simulation support at NERSC/PDSF • Estimated totals	• 1.4-1.9 PB/year • 800-1000 TB, 400-500 k files • 80-100 TB or less in burst • 10 k files, ~15-20 TB and of the order of 100 k files and 90-130 TB results back to BNL	• 400-500 MB/sec - peak at 500 MB/sec	• 1-1.5 Gbps transfer to NERSC/PDSF over 3 months • SAC support @ 1.5 Gbps, data moved within a week • ~ 2 Gbps in/out of NERSC and BNL • NERSC 2 Gbps and BNL 3 Gbps
2-5 years				
• RHIC/STAR data taking, Heavy Flavor program, local and distributed data production	• Data taking • Distributed infrastructure-based simulations and Fast-Offline (50%) — cloud-like • MuDST copy at NERSC/PDSF • MuDST delivery to 3-4 SACS & T2 centers • Embedding simulation support at NERSC/PDSF • Estimated totals	• 2.0-2.5 PB/year • ~ 1.0-1.2 PB during runs • 1-1.6 PB, 400-500 k files • 100-160 TB or less in burst • 20-22 TB input and ~ 160 TB output	• 550 MB/sec from online to RCF	• 3-3.5 Gbps for streaming data from online to remote site (“live”) • 2 Gbps Transfer to NERSC/PDSF over 3 months • SAC support @ 2.5 Gbps, data moved within a week • 2-3 Gbps in/out of NERSC and BNL • NERSC 3-4 Gbps and BNL 5-6 Gbps

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Data Set Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
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 data sets • 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

Table 11-5. RHIC Requirements Summary — PHENIX.

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)				
• PHENIX upgrades with the VTX (commissioned) and FVTX (Run 12) detectors	• Centralized processing to DSTs • Analysis trains • Modest-size off-site transfers	• ~3 PB raw data reconstructed to ~2.5 PB DSTs • 800 TB (estimated) transferred	• Near-line 300,000 raw data files • 10 GB each • Network in place	• Virtually no near-line requirements • 800 TB volume estimated
2-5 years				
• Data taking with VTX+FVTX • Simulations for the sPHENIX upgrade	• Centralized processing as above • Distributed simulations	• Estimated data rates according to beam species and energies (estimated in Appendix A)	• Near-line	• No near-line requirements
5+ years				
• sPHENIX commissioning and operation	• No change in computing paradigm envisioned	• 3 GB/sec peak (weak estimate) • Move to larger file sizes (100 G?)	• Near-line	• No near-line requirements

11.9 **Appendix**

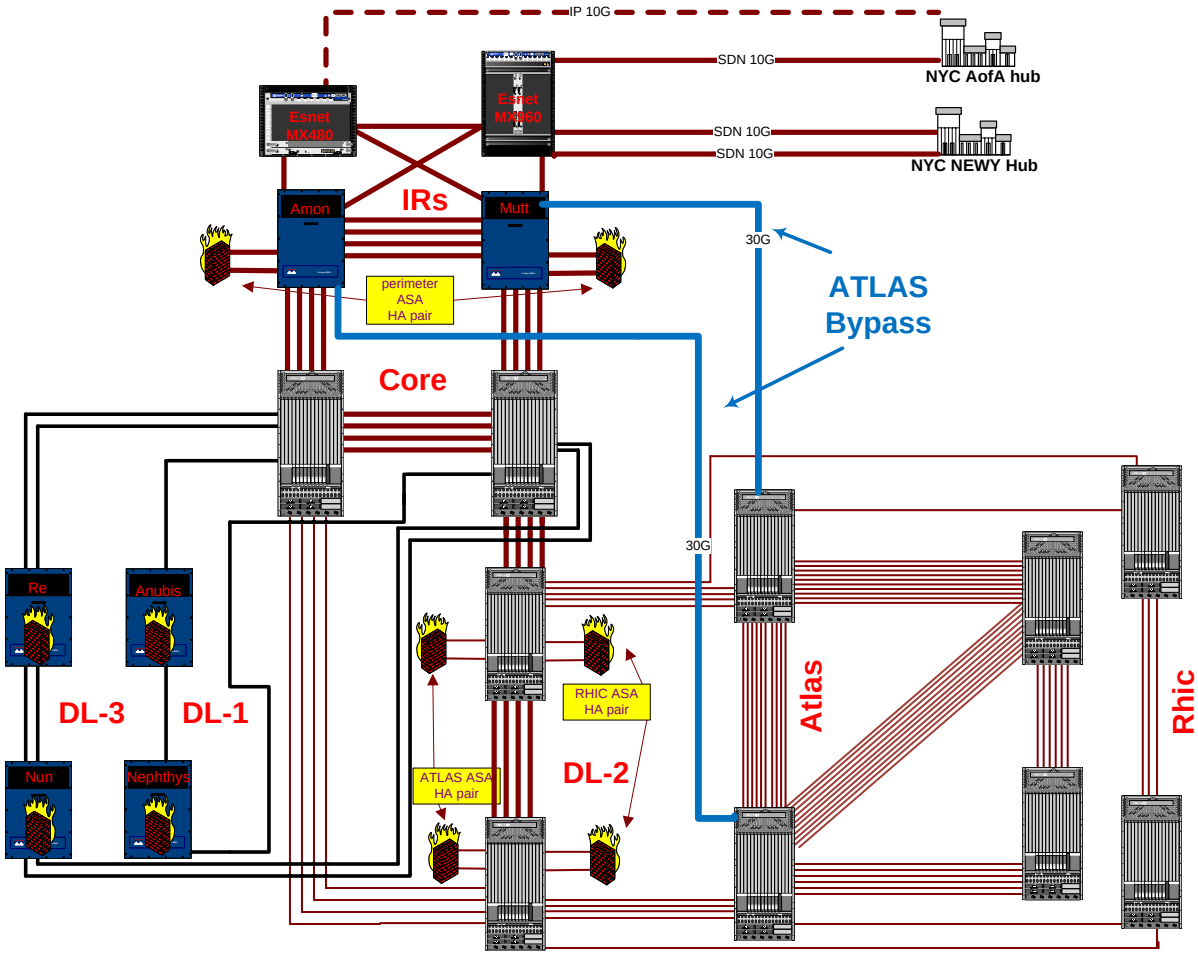


Figure 11-3. BNL Local Area Network as of spring 2011.

12 Glossary

GB/sec: Gigabytes per second – a measure of network bandwidth or data throughput

Gbps: Gigabits per second – a measure of network bandwidth or data throughput

MB/sec: Megabytes per second – a measure of network bandwidth or data throughput

Mbps: Megabits per second – a measure of network bandwidth or data throughput

PB/sec: Petabytes per second – a measure of network bandwidth or data throughput

Pbps: Petabits per second – a measure of network bandwidth or data throughput

TB/sec: Terabytes per second – a measure of network bandwidth or data throughput

Tbps: Terabits per second – a measure of network bandwidth or data throughput

12.1 Acronyms

ACCRE	Advanced Computing Center for Research and Education
AF	Analysis facility
ALICE	A Large Ion Collider Experiment
AliEn	ALICE Environment
ANL	Argonne National Laboratory
AOD	Analysis object data
ASCR	Office of Advanced Scientific Computing Research
ATLAS	A Toroidal LHC Apparatus
BGP	Border Gateway Protocol
BNL	Brookhaven National Laboratory
BUR	Beam User Request
CBM	Compressed Baryonic Matter
CCJ	Computing Center in Japan
CEBAF	Continuous Electron Beam Accelerator Facility
CERN	European Organization for Nuclear Research
CLAS	CEBAF Large Acceptance Spectrometer
CMS	Compact Muon Solenoid
CPU	Central Processing Unit
CRAB	CMS Remote Analysis Builder
CVMFS	CERN Virtual Machine Filesystem
DAQ	Data acquisition
DOE	Department of Energy
DST	Data summary tape
ECS	ESnet Collaboration Services
ELIC	Electron Ion Collider
ESD	Event summary data
ESnet	Energy Sciences Network
EVO	Enabling Virtual Organizations
FDT	Fast Data Transfer
FEL	Free-electron laser
FGT	Forward GEM Tracker
FNAL	Fermi National Accelerator Laboratory
FTD	File Transfer Daemon
FTP	File Transfer Protocol
FVTX	Forward Silicon Vertex Detector

GB	gigabyte
GEANT	Gigabit European Advanced Network Technology
GPU	Graphics processing unit
GSI	Grid Security Infrastructure
HBT	Hanbury Brown-Twiss
HFT	Heavy Flavor Tracker
HLT	High Level Trigger
HPSS	High-performance storage system
ILDG	International Lattice Data Grid
JSA	Jefferson Science Associates
KISTI	Korean Institute of Science and Technology Information
LBNL	Lawrence Berkeley National Laboratory
LHC	Large Hadron Collider
LINAC	Linear accelerator
LITE	Lightwave Internetworking Technology Enterprise
LQCD	Lattice QCD
MAN	Metropolitan area network
MATP	Mid Atlantic Terascale Partnership
MonALISA	MONitoring Agents using a Large Integrated Services Architecture
MuDST	Micro DST
NERSC	National Energy Research Scientific Computing Center
NFS	Network Filesystem
NP	Office of Nuclear Physics
NSF	National Science Foundation
NZS	Non-zero suppressed
ODU	Old Dominion University
ORISE	Oak Ridge Institute for Science and Education
OSC	Ohio Supercomputer Center
OSCARS	On-Demand Secure Circuits and Advance Reservation System
OSG	Open Science Grid
PAC	Program Advisory Committee
PBS	Portable Batch System
PDSF	Parallel Distributed Systems Facility
PerfSONAR	PERformance Service Oriented Network monitoring ARchitecture
PHENIX	Pioneering High Energy Nuclear Interaction eXperiment
PKI	Public key infrastructure
PROOF	Parallel ROOT Facility
PWG	Physics Working Group
QCD	Quantum chromodynamics
QGP	Quark-gluon plasma
RACF	RHIC/ATLAS Computing Facility
R&D	Research and development
RCF	RHIC Computing Facility
RF	Radiofrequency
RHIC	Relativistic Heavy Ion Collider
RMS	Root Mean Square
RPM	RPM Package Manager (historically Red Hat Package Manager)
RSV	Resource and Service Validation
SAC	STAR Analysis Center
SAM	Service Availability Monitoring
SC	Office of Science
SDN	Science Data Network
SE	Storage element

SoX	Southern Crossroads
SRF	Superconducting radiofrequency
SRM	Storage resource manager
STAR	Solenoidal Tracker At RHIC
SURA	Southeastern Universities Research Association
SVT	Silicon Vertex Tracker
TB	terabyte
TPC	Time Projection Chamber
T0, T1, etc.	Tier 0, Tier 1, etc.
UIC	University of Illinois at Chicago
VLAN	Virtual local area network
VM	Virtual Machine
VO	Virtual organization
VoIP	Voice Over Internet Protocol
VOMS	Virtual Organization Membership Service
VORTEX	Virginia Optical Research Technology Exchange
VTX	Silicon Vertex Tracker
WAN	Wide area network
WLCG	Worldwide LHC Computing Grid
ZDC	Zero Degree Calorimeter

13 Acknowledgements

This work would not have been possible without the contributions and participation of those who provided information and attended the workshop. ESnet would also like to thank the Nuclear Physics program office for its help in organizing the workshop and providing insight into the facilities supported by the Nuclear Physics program. In addition, the Oak Ridge Institute for Science and Education (ORISE) conference support and logistics staff was very helpful.

ESnet is funded by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research (ASCR). Vince Dattoria is the ESnet Program Manager.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the U.S .Department of Energy under contract

DE-AC02-05CH11231.

This work was supported by the Directors of the Office of Science, Office of Advanced Scientific Computing Research, Facilities Division, and the Office of Nuclear Physics.

This is LBNL report LBNL-XXXX