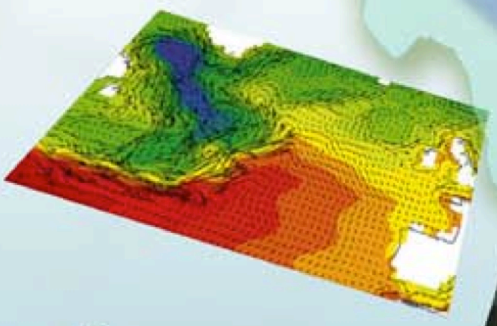
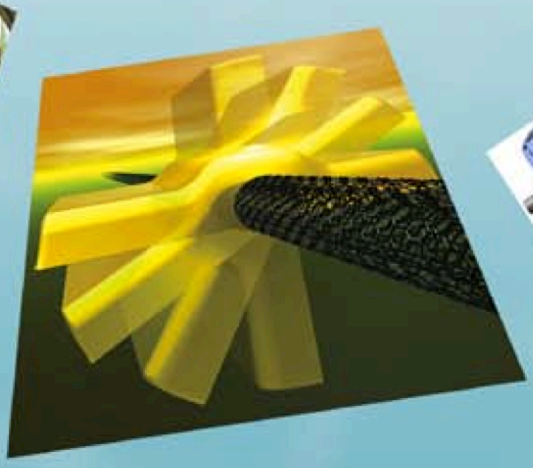
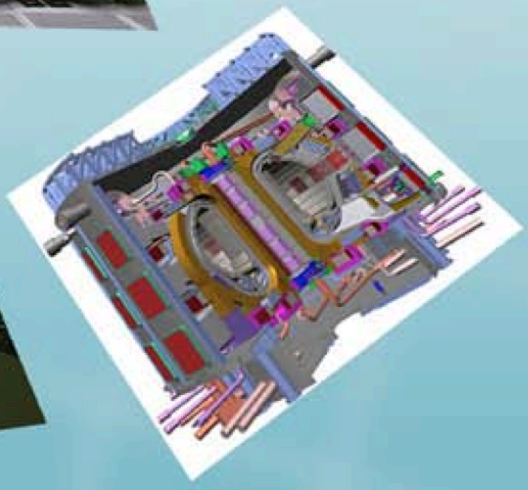
NP Science Network Requirements

ESnet
Energy Sciences Network

Report of the Nuclear Physics Network Requirements Workshop Conducted May 6 and 7, 2008







Berkeley Lab

Office of Science
U.S. Department of Energy

Nuclear Physics Network Requirements Workshop

Nuclear Physics Program Office, DOE Office of Science
Energy Sciences Network
Bethesda, MD – May 6 and 7, 2008

ESnet is funded by the US Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program. 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 US 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.

Participants and Contributors

Rich Carlson, Internet2 (Networking)

Eli Dart, ESnet (Networking)

Vince Dattoria, DOE/SC/ASCR (ASCR Program Office)

Michael Ernst, BNL (RHIC)

Daniel Hitchcock, DOE/SC/ASCR (ASCR Program Office)

William Johnston, ESnet (Networking)

Andy Kowalski, JLAB (Networking)

Jerome Lauret, BNL (STAR at RHIC)

Charles Maguire, Vanderbilt (LHC CMS Heavy Ion)

Douglas Olson, LBNL (STAR at RHIC and ALICE at LHC)

Martin Purschke, BNL (PHENIX at RHIC)

Gulshan Rai, DOE/SC (NP Program Office)

Brian Tierney, ESnet (Networking)

Chip Watson, JLAB (CEBAF)

Carla Vale, BNL (PHENIX at RHIC)

Editors

Eli Dart, ESnet – dart@es.net
Brian Tierney, ESnet – bltierney@es.net

Contents

1 Executive Summary

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

In May 2008, ESnet and the Nuclear Physics (NP) Program Office of the DOE Office of Science organized a workshop to characterize the networking requirements of the science programs funded by the NP Program Office.

Most of the key DOE sites for NP related work will require significant increases in network bandwidth in the 5 year time frame. This includes roughly 40 Gbps for BNL, and 20 Gbps for NERSC. Total transatlantic requirements are on the order of 40 Gbps, and transpacific requirements are on the order of 30 Gbps. Other key sites are Vanderbilt University and MIT, which will need on the order of 20 Gbps bandwidth to support data transfers for the CMS Heavy Ion program.

In addition to bandwidth requirements, the workshop emphasized several points in regard to science process and collaboration. One key point is the heavy reliance on Grid tools and infrastructure (both PKI and tools such as GridFTP) by the NP community. The reliance on Grid software is expected to increase in the future. Therefore, continued development and support of Grid software is very important to the NP science community. Another key finding is that scientific productivity is greatly enhanced by easy researcher-local access to instrument data. This is driving the creation of distributed repositories for instrument data at collaborating institutions, along with a corresponding increase in demand for network-based data transfers and the tools to manage those transfers effectively. Network reliability is also becoming more important as there is often a narrow window between data collection and data archiving when transfer and analysis can be done. The instruments do not stop producing data, so extended network outages can result in data loss due to analysis pipeline stalls. Finally, as the scope of collaboration continues to increase, collaboration tools such as audio and video conferencing are becoming ever more critical to the productivity of scientific collaborations.

2 Workshop Background and Structure

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

In May 2008, ESnet and the Nuclear Physics (NP) Program Office of the DOE Office of Science organized a workshop to characterize the networking requirements of the science programs funded by the NP Program Office. The most network demanding facilities include the Relativistic Heavy Ion Collider (RHIC) facility at Brookhaven National Laboratory (BNL) and the Thomas Jefferson National Accelerator Facility (TJNAF). NP data will also be generated by the CMS and ALICE experiments at the Large Hadron Collider (LHC) facility at the European Organization for Nuclear Research (CERN) in Switzerland.

Workshop participants were asked to codify their requirements in a "case study" format which included a network-centric narrative describing the science, the instruments and facilities used or anticipated for future programs and the network services needed and the way in which the network is used. Participants were asked to consider three time scales in their case studies – the near term (immediately and up to 12 months in the future), the medium term (3-5 years in the future), and the long term (greater than 5 years in the future). The information in the narrative was distilled into a summary table, with rows for each time scale and columns for network bandwidth and services requirements.

3 DOE Nuclear Physics Programs

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 electronic, 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.

The DOE Nuclear Physics program provides most of the Federal support for nuclear physics research in the U.S. About 1,500 scientists, including 800 graduate students and postdoctoral research associates, receive support from NP. In addition, the program supports four 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, and the nuclear stockpile. A group of highly trained scientists expert in fundamental nuclear physics is another important result of the program. This valuable human resource is essential for many applied fields such as nuclear medicine, space exploration, and national security.

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 1000 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 1,200 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 preparing to upgrade the CEBAF, which will double the electron beam energy.

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) and the Holifield Radioactive Ion Beam Facility (HRIBF) at Oak Ridge National Laboratory (ORNL). 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 Texas A&M University (TAMU), the Triangle Universities Nuclear Laboratory (TUNL) at Duke University, and at Yale University; infrastructure is supported at the University of Washington to enable scientific instrumentation projects to be undertaken.

Research and development activities are underway, aimed at the development of a U.S. facility for rare isotope beams. This facility will enable world-leading research opportunities in nuclear structure, nuclear astrophysics, and fundamental studies, and complement the programs of high capability radioactive ion beam facilities elsewhere in the world. Experimental results from such a facility could have a profound impact on our basic knowledge of nuclear structure and the origin of the elements in stars and stellar explosions.

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 has joined the Italian-lead Cryogenic Underground Observatory for Rare Events (CUORE) experiment at the Gran Sasso Laboratory, contributing to the fabrication of the detector that is planned to become operational in approximately FY 2012. 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 will become operational in approximately 2011. 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.

3.1 Nuclear Physics Network Requirements at RHIC

Background

Relativistic Heavy Ion Collider

The RHIC Facility is a National User Facility at BNL that delivers the world's highest energy heavy ion collisions and colliding polarized proton beams for experimental research in nuclear physics. This facility has two large state-of-the-art detectors operated by the Pioneering High Energy Nuclear Interaction eXperiment (PHENIX) and Solenoidal Tracker at RHIC (STAR) research collaborations. Together, these experiments investigate collisions of heavy ions with the primary goal to discover and study a new state of matter called the Quark-Gluon Plasma, and collisions of polarized protons to study the internal structure of the nucleon.

The two RHIC experiments, PHENIX and STAR, are collaborations spanning many countries and are composed of hundreds of collaborators each. The geographically dispersed nature of the collaborations points directly to a requirement for high-quality, flexible and high-performance connectivity between collaborators to enable effective communication and data analysis. The scale of data movement required by the RHIC experiments relies on the availability of high-bandwidth, production quality network infrastructure.

Coming from an initially BNL-centric computing model, several changes in the analysis model and the addition of some new collaborative tools have helped to streamline and expedite the analysis of RHIC data in recent years. These changes have several implications for network bandwidth requirements:

- The computing models of the experiments, formerly relying on few central facilities, have been steadily evolving to a more distributed, data-grid (i.e. STAR) model where the unprocessed and processed data is immediately made available to remote sites where computing resources are available to the experiments
- Entire raw data sets (hundreds of Terabytes per year) are transferred to remote sites for processing (i.e. PHENIX)
- The availability of cost-effective commodity market storage hardware, coupled with improved network bandwidth between BNL and remote institutions, has resulted in a shift towards scientist-local data analysis which places increased demands on the network for data distribution from BNL to remote collaborators
- Because of the vast geographic distribution of the physicists participating in the research program, collaborative tools are critical for maximizing scientific output by fostering the communication across the members of the collaborations

While current annual data collection levels are at Petabyte (10^{12} bytes) scale, the aggregate raw data rate envisioned by the RHIC experiment's program will more than double (but will actually grow by a factor of 10 for STAR) during Run 9 that is scheduled to start in early 2009, reaching an online data acquisition rate of 1 GB/sec. The scale and growth rate of the collected data make data management and distribution an ever-growing challenge. To face the challenges caused by the size of those datasets, while preserving

the Physics quality and turn-around time, the RHIC experiments have adopted a distributed computing model based on the allocation of dedicated or opportunistic remote resources.

As early as 2005, the PHENIX experiment has moved an entire dataset of 6.8 billion events (or 270 TB of data) to their Japanese CC-J facility over the 11 weeks, where it was further reconstructed and analyzed. Over the past few years the STAR computing model has evolved to a data-grid based model, utilizing middleware components developed by Grid projects (e.g. PPDG, OSG and others).

To estimate the resources needed by the RHIC mid-term program and in the RHIC-II luminosity upgrade era, algorithms were developed to make projections for anticipated RHIC running scenarios. The network estimates for the runs beyond 2008 are based on the acceptable fractional data transfer rates relative to the experiment's data acquisition capabilities. The result and summary of these projections, derived primarily from the expected amount of raw data collected, are shown in the table below.

Table 1: RHIC and RHIC-II planning derived data sets and estimated Wide Area Network (WAN) needs

	FY05	**FY06**	**FY07**	**FY08**	**FY09**	**FY10**	**FY11**	**FY12**
STAR Data (TB/year)	145	54	145	172	1360	1915	2610	2610
PHENIX Data (TB/year)	488	337	650	590	1200	1600	2500	2500
Total Annual Raw Data (TB/year)	633	391	795	762	2560	3515	5110	5110
Required WAN bandwidth (Mbps)	276	1500	2737	3020	6066	8719	13047	13047

ESnet, peering with other domestic (e.g. Internet2) and international network service providers (e.g. GEANT) provides all of BNL's network connectivity. The RHIC experiments assume the ability to move large quantities of data, collaborate efficiently, and effectively distribute computation to serve the needs of the experiments. The RHIC scientists have estimated that the combined network requirements of the RHIC experiments will exceed 3 Gbps in 2008, and will reach more than 13 Gbps in 2011 and beyond. Of particular note are the transition between FY08 and FY09 (projected WAN bandwidth usage doubles) and the transition between FY10 and FY11 (projected WAN bandwidth usage increases by a further 50%).

Current Network Requirements

The PHENIX Collaboration

The PHENIX collaboration currently has about 550 members from 67 institutions, 13 different countries, and 4 continents. Historically, the analysis model has been for collaboration members to perform their analysis at the RHIC Computing Facility (RCF), where the various datasets have been available centrally. Until about 2004, all of the raw data reconstruction took place at the RCF.

Several changes in the analysis model, and also some new collaborative tools, have helped streamline and expedite the analysis of the PHENIX data, as well as change the network bandwidth requirements in several ways:

- In 2005, PHENIX began to transfer entire raw data sets to remote facilities for remote processing;
- With the availability of inexpensive large data storage and increased network bandwidth to remote institutions, there is a shift towards researcher-local data analysis;
- Collaborative tools, especially desktop video conferencing, have network requirements different from those for data transfers (low latency, high availability).

The data acquisition system of the PHENIX experiment currently has a throughput of up to 600 MB/sec. During a RHIC Run, raw data get sent to remote facilities, in part to relieve the RCF, and in part to process the data at the facility where the experts for the given dataset are physically located. This is especially true for the polarized proton (spin) data of RHIC. Many of the “spin” experts are located at Japanese institutions, and PHENIX routinely transfer the entire proton raw dataset to the CCJ computing center at RIKEN close to Tokyo. In this way, the RCF is able to process the previously acquired heavy-ion dataset, while the CCJ and the local experts run the reconstruction of the proton data.

In addition to CCJ, PHENIX transfers portions of the heavy-ion datasets to institutions interested in facilitating a rapid analysis turnaround. In the past heavy-ion datasets have been transferred to ORNL, IN2P3 (CCF) in France, and the ACCRE facility at Vanderbilt University.

The reconstruction of the raw data is the first step in the analysis chain resulting in more compact files called "Data Summary Tapes" (DSTs), which can more readily be analyzed. Once produced off-site, the DSTs are sent back to the RCF for long-term archival storage and analysis at the RCF.

For proton-proton datasets, the size of the DSTs is a fraction (~ 20%) of the original raw data size. For heavy-ion datasets, there is less reduction in size due to the high number of produced particles per collision.

Data transfers from the PHENIX counting house need to happen in near-real-time. The PHENIX data acquisition system has the ability to store a given raw data file for about 20 hours. During this period, the file is copied to the RCF HPSS tape storage system, and is available for transfer to remote facilities. The file then needs to be deleted to make room

for new data. This requirement leads to significant fluctuations in the instantaneous data throughput of up to 300 MB/sec (2.4 Gbps), with averages slightly below 80 MB/sec (640 Mbps).

The RHIC Run in 2008 had a proton-proton component of only 23 days overall, resulting in a small dataset of about 118 TB of raw data which was sent to CCJ. Future running at RHIC will most likely have a larger proton-proton component.

The transfers of reconstructed data back to RCF have fewer fluctuations, because they have no such stringent near-line requirements where the cached data must be transferred before more data arrives.

The STAR Collaboration

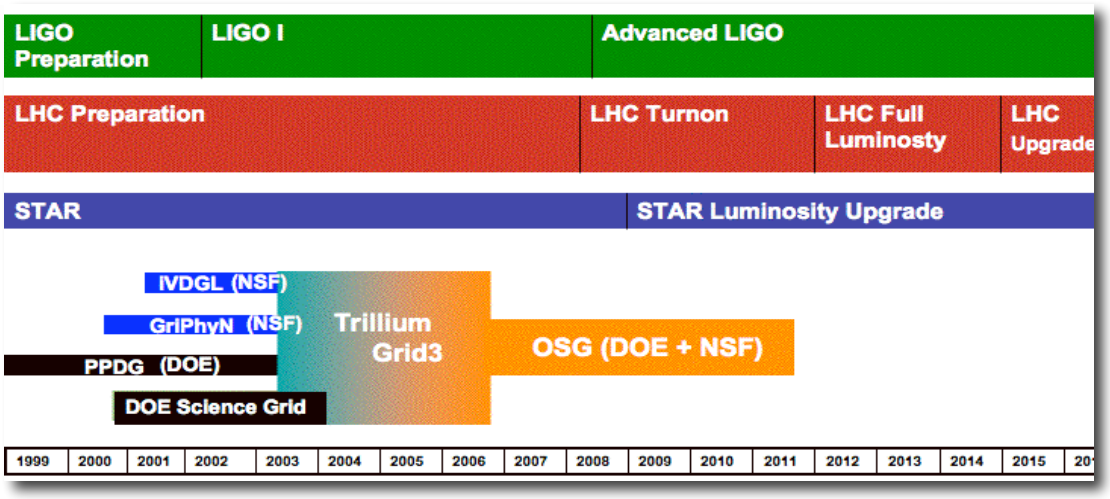
The STAR Experiment is an international collaboration composed of 52 institutions spread over 12 countries, with more than 590 physicists and skilled technicians. Geographically, STAR institutions are distributed as follows

USA / North America	24	46%
Europe	12	23%
Asia (China/Korea)	8	15%
India	6	12%
South America	2	04%

STAR Analysis and Data Flow Model

The STAR production and analysis models have been mainly relying on centralized user analysis facilities (RCF at BNL and the National Energy Research Scientific Computing Center/Parallel Distributed Systems Facility NERSC/PDSF at LBNL). The STAR computing model has however been steadily evolving around a data-grid model where the processed data is made immediately available to remote sites where computing resources may be available.

Data redistribution strategies have been developed through involvement in the Particle Physics Data Grid project, the Trillium project and now the Open Science Grid. The figure below shows the timeline and evolution of these projects.



The STAR distributed computing model has evolved from a centralized processing paradigm as follows:

- With the planned upgrade of the Data Acquisition System, the experiment's data volume is expected grow by an order of magnitude reaching a data acquisition rate of 1 GB/sec by 2009 and making data management and distribution an ever growing challenge. To face this challenge, the STAR experiment will rely on additional Tier-1 facilities and envisions the processing of entire samples at dedicated remote sites. One of those facilities is the Korea Institute of Science and Technology Information (KISTI) located at Daejon/Korea.
- The present off-site Tier-1 facility is NERSC/PDSF, which supports large-scale simulations and "signal" embedded data production required for efficiency studies. Recently more institutions have expressed an interest in contributing resources to this effort. Datasets have been transferred to the University of Birmingham, England and the University of Illinois at Chicago.
- The data redistribution includes the full set of derived data analysis samples (Micro-DSTs or MuDSTs) sent to Tier-1 facilities and while further reduced data samples (nano-DSTs) are sent to Tier-2 centers (e.g. in China). The immediate availability of datasets at remote sites has been shown to increase physics productivity and therefore STAR is planning to engage more Tier-2 centers in the next five years. However, some aspiring Tier-2 institutions are lacking efficient access to the large data sets because of two factors: high network latencies and the absence of adequate network backbone infrastructure
- STAR is a full member of Open Science Grid and as such, makes use of its resources for Monte-Carlo simulation. Recent use of Virtual Machine technology on Amazon EC2 and Nimbus Clusters at the University of Chicago tend to indicate that support for Grid based operation could easily grow from a Monte-Carlo based operation to a full data production based paradigm on opportunistic resources. To date, data production has been confined to the RHIC Computing Facility at Brookhaven National Laboratory.

Network Requirements – the next 5 years

PHENIX

PHENIX currently generates about 50TB of raw data per week while running, resulting in about 600TB of data in a typical 12-week running period. In the future, PHENIX will add several new detector components, which will increase the per-event data sizes by a factor of 2 or more for Gold beam collisions systems, and a factor of 3 or more for proton beam collisions, with the corresponding increases in network bandwidth. The two silicon-based detectors, the Vertex Detector and Forward Vertex Detector, scheduled to be commissioned in 2010, will add significantly to the per-event data size.

The following table summarizes the projected rates, assuming that each future RHIC Run has an equal heavy ion and a proton beam component.

Run/Year	Acquired dataset	Transferred dataset
2008 (d+Au)	460TB	-
2008 (p+p)	118 TB	118TB
2009 (25 weeks)	1200TB	600TB
2010 (25 weeks, partial VTX)	1600TB	800TB
2011 (25 weeks, VTX+FVTX)	2500TB	1200TB

STAR

The planned accumulation of dataset size is shown in column 2 of the table below, and is consistent with the RHIC mid-term strategic planning document for the period of 2006-2011. Column 3 shows the amount of "raw" data amount (mix of light+heavy ion beam Runs), column 4 the derived data size and column 5th, the MuDST dataset size.

	Projected TB on tape (all)	Projected Raw TB	Acquired/re-scoped Raw (TB)	Expected Raw for embedding (15% level) (TB)	Expected derived MuDST TB (1 pass, MuDST only) Tier-2 scale (TB)
2008	870	115+320	165 (37%)	16	33
2009	1720	220+640	650 (proton beam only)	65	130
2010	3000	500+1000	1500	150	300
2011	4160	680+1400	2080	208	416
2012	4160	680+1400	2080	208	416

The difference between projected and "rescoped" data set size in 2008 is due to experiment operating factors.

At present, the STAR data grid computing infrastructure consists of six sites: NERSC/PDSF, Wayne State University, University of Illinois at Chicago, Nuclear Physics Institute of the Academy of Sciences of the Czech Republic (NPI ASCR) in Prague, University of Birmingham, England and University of Sao Paulo, Brazil. These sites rely on the stability of grid-based operations and efficient large distance data transfer.

The number of Tier-2 computing sites is expected to double by 2010. The U.S. STAR Tier-1 center could be complemented by a secured Tier-1 center at the Korea Institute of Science and Technology Information (KISTI). The growing participation of foreign institutions will require robust international networking to move data from BNL to Tier-1 and Tier-2 centers. In the future, most computational workflows will occur via grid interfaces to facilities and all STAR collaborators will be organized as a grid Virtual Organization.

Additional Network-related Needs

STAR expects significantly more on-demand WAN data transfers to support individual analyses, in addition to the managed bulk placement of datasets. For the ESnet PKI, this implies a 10-20 fold increase in the number of grid credentials due to the increase in user and host credentials that will be needed (from 10s to 100s).

The STAR collaboration currently makes extensive use of services such as teleconferencing and Web publishing, with an increasing amount of IP-based teleconferencing (Skype). Collaboration technologies that are integrated with grid authorization services so that STAR collaborators already registered as members of the VO could use these services without additional registration steps would be helpful.

International Partners for STAR

The following section describes the network requirements between BNL and KISTI in Korea, University of Birmingham, England, and NPI ASCR in Prague, Czech Republic.

KISTI/Korea (Tier-1)

KISTI has joined the STAR collaboration as a full member institution and their proposal includes contributions to computing, storage and network resources, and operations workforce. KISTI's resources are linked to computing resources at the Pusan National University (Korea) and made accessible via a connection to the "Global Ring Network for Advanced Application Development" network (GLORIAD). GLORIAD is a high-speed computer network connecting scientific organizations in Russia, China, United States, the Netherlands, Korea and Canada.

Scope & Usage

While the exact resource level has yet to be finalized, the STAR collaboration plans to use the center for high priority data processing. This involves moving entire datasets from BNL to KISTI. Data rates to the KISTI facility would need to scale to the STAR

program's raw data production. An optimum goal would be to accommodate half of the STAR raw data sets accumulated at BNL. Produced data would need to be brought back to BNL for permanent archiving, while the derived (MuDST) data sets would be re-distributed to the institutions in Korea and China, making data samples immediately available (as was done in the 2004 exercise which leveraged the resources at NERSC/PDSF). The data transfer rate needs to be fast enough to fully utilize the compute resources at KISTI and to allow near real-time transfer of data sets to remote Tier-1 centers. 100 MB/sec (800 Mbps) is the bare minimum required, and will only allow for a subset of the data to be transferred back to BNL (this would also require negotiation with KISTI to deploy additional storage at KISTI to mitigate the lack of network bandwidth). The ability to transfer at all data generated at KISTI back to BNL will require at least 300 MB/sec (2.4 Gbps).

The GLORIAD network, which currently provides a 10 Gbps network backbone (6 Gbps for routed traffic, and 3 x 1 Gbps links for circuit services) all the way to the U.S. east coast, seems to be the natural path.

KISTI could eventually become the gateway for Asian High Performance Computing Collaboration to STAR. KISTI has expressed an interest in establishing connectivity between Representative Supercomputing Centers in the STAR Collaboration (i.e., Shanghai Supercomputing Center, KISTI Supercomputing Center, NERSC, and possibly others).

University of Birmingham, England (Tier-2)

The University of Birmingham site has the potential of migrating from a Tier-2 to a Tier-1 center. In terms of network planning, transatlantic bandwidth will need to be monitored to ensure sufficient data transfer capability exists between the US and Europe particularly in light of the high network demand expected from the ramp-up of the CERN LHC experiments.

NPI ASCR/Prague (Tier-2 pilot)

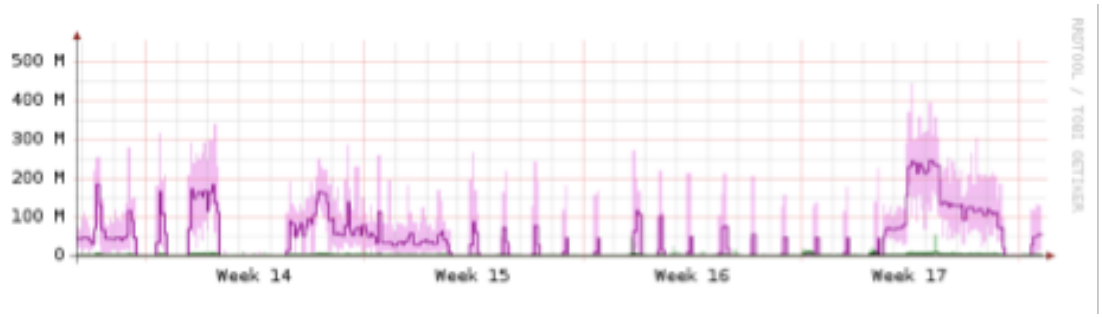
The Ultra-Relativistic Heavy Ion Group of the Nuclear Physics Institute ASCR (NPI ASCR) has been an active STAR participant since 2001. At present, NPI ASCR has two dedicated computer scientists, a supplement of 18 TB of storage space, and increased computing power provided by the "Golias" compute farm to handle the increasing demand of the growing physics group.

The creation of local opportunities in Prague for scientific analysis (without the need for remote connection to BNL) has since attracted additional Physicists, and the group has doubled in the past year. The Prague data processing capability was limited mostly by the ability to transfer the data sets from BNL to the local storage and vice versa. However, in 2008 with help from ESnet and CZnet personnel as well as BNL's IT division of networking, a dedicated 1 Gigabit network path was established between BNL and NPI ASCR. Once the BNL datasets arrive at the local site in Prague, researchers can access the data and launch analysis jobs that run at the Regional Computing Center for Particle Physics "Golias" compute farm, the biggest site in the Czech Republic.

STAR's site at NPI ASCR in Prague is an example of science empowered by the deployment of distributed computing technologies. Typically, a Tier-2 center would

transfer data sets of interest to the extent useful to their local research efforts. A full micro-DST set for a given year or portions (triggered events, species) of multiple years' datasets would be used to sustain their research efforts. At the Prague prototype site, data transfers are handled using the SRM (BestMan) client interoperating with DPM (Disk Pool Manager) SRM at Prague/Bulovka. Local data access leverages the STAR Meta-Scheduler and framework to access a Logical File Name namespace, but the CPU power resides at Golias. These two facilities are interconnected by a high-speed fiber link. The facility also has a full copy of all database information. Therefore, once the data has been transferred to Prague, full analysis and full event reconstruction is possible without the need for continued transatlantic connections back to BNL.

As shown in the figure below, data transfer rates to the Tier-2s are bursty, and need sustained rates of at least 300 Mbps. The requirements for the next five years are to have sufficient link capacity to exploit the full potential of researcher-local investments. As the amount of data taken by the STAR collaboration rapidly increases in near future, the spread of compute power via the Grid will become a necessity, and the mode of operation of the Prague testbed will become a standard mode of operation. Hence, the current peak rate should at be maintained and scale with the derived data rate increase. As noted above, STAR anticipates an increase of Tier-2 centers by a factor of 2 by 2010 driving a network bandwidth requirement to/from Tier-0 or Tier-1 center depending on the geographical location of the Tier-2 center.



WAN Data transfer speeds from BNL to Prague for March, 2008

Summary Table

Time Frame	Science Instruments and Facilities	Process of Science	Anticipated Requirements	
			Network	Network Services and Middleware
Near-term	• RHIC at BNL with currently PHENIX and STAR experiments taking and analyzing data	• PHENIX large scale data transfers to Japan and France for reconstruction and analysis • STAR data transfers to Tier-1 sites (PDSF and KISTI/Korea (planned)) • STAR analysis uses Data Grid model (OSG) • Support of 6-8 STAR Tier-2 sites	• 6 Gbps Wide Area bandwidth	• Grid middleware infrastructure: Computing Element, Storage Element, w/ X.509 authentication and authorization, managed transfers (SRM) • Reliable transfer of large datasets, domestically and internationally (Japan, Korea, France, Czech Republic) • End-to-end monitoring • Grid Security Infrastructure
5 years	• RHIC (STAR, PHENIX) and RHIC-II support	• Enable data analysis for collaborators spanning over 60 institutions >15 countries with support for less developed regions in South America, Russia, India and China • Dynamic replication of frequently accessed datasets (data on demand) • All simulation needs for RHIC-II done on Grid infrastructure	• 20 Gbps Wide Area bandwidth	• Bandwidth and Service guarantees, quality of service • Network bandwidth predictions, guaranteed high bandwidth (accounting) • Secure data access • Grid infrastructure, schedulers, brokers, planners, Processing/data co-scheduling • Object level access • Distributed database
5+ years		• Real-time data analysis, event visualization • Underlying object-on-demand oriented analysis (rather than file) stabilizing network needs • Extensive use of data caching • Interoperability with other grids	• 40 Gbps Wide Area bandwidth	• Resource discovery • Database access on Grid • Dynamic data grid, computational grid relies on co-scheduling of CPU, data and network resources

3.2 CMS-HI Research Program

Background

The Compact Muon Solenoid (CMS) experiment collaboration at the Large Hadron Collider (LHC) consists of over 2000 scientists from 35 countries. The CMS detector will begin taking data from proton-proton collisions for high-energy physics research in a few months time, and will start taking data from heavy ion collisions of Pb+Pb later next year. As with the other large experiments at the LHC, the High Energy Physics (HEP) data processing is designed around a multi-tiered system centered at the CERN Tier-0 site and then fanning out to several Tier-1 sites hosted in some of the member states. For the United States, the Tier-1 site for the CMS-HEP computing is at Fermi National Accelerator Lab. In turn, the FNAL Tier-1 site provides reconstructed data to seven Tier-2 computing sites in the U.S., including one at MIT, and to other international sites for CMS.

The program of heavy ion nuclear physics research is a distinct, recognized subset of the entire CMS collaboration known as CMS-HI. At the LHC there are expected to be heavy ion collisions during one month per year. From the US there are 10 institutions participating in the heavy ion research program at CMS. All of the CMS-HI institutions in the US have CMS-HEP groups as well. MIT is the lead US institution for CMS-HI.

An internal CMS-HI study group has identified Vanderbilt University to be the location of the main CMS-HI compute center. There is a large (1500 CPUs) facility already present at Vanderbilt that has been used successfully in 2006 and 2007 to process raw data arriving soon after it was acquired by the PHENIX experiment at RHIC. Additionally, the CMS-HI study group recommended that the CMS-HI institutions continue to take advantage of the already existing MIT Tier-2. The MIT-HI group has been serving the simulation needs for CMS-HI during the past several years.

A draft proposal for CMS heavy ion computing has been submitted to the DOE. This proposal details the required resources for the main CMS-HI compute center, including numbers of CPUs, disk and tape storage, and the compute model of use for all the US institutions in CMS-HI. This proposal recommends assigning some of the CPU and disk analysis and simulation resources to the MIT Tier-2 site, but all of the first pass data reconstruction will be done at the Vanderbilt site.

Current Local Area Network Requirements

Instruments and Facilities

Both the MIT Tier-2 and the existing ACCRE compute facility possess high-speed local area networks in the range 1 Gbps to several Gbps. Upgrades to 10 Gbps are planned. The LANs at these two sites are proven capabilities for both heavy ion simulation and heavy ion real data processing efforts mounted in the past and continuing at present.

Process of Science

Real data reconstruction jobs for PHENIX in 2006 and 2007 ran automatically upon receipt of the data and database information from RHIC, using in-house developed software. For example, in two months of 2007 some 30 TBytes of raw heavy ion collision data were sent from RHIC to Vanderbilt and processed automatically, with minimal manual intervention. Reconstructed outputs were then shipped back to the RHIC Computer Facility in Brookhaven for immediate inspection and use by PHENIX collaborators while the run was still in progress. Similarly, batch jobs are run on the MIT Tier-2 for CMS-HI using CMS collaboration developed job submission and management software.

Current Wide Area Network Requirements

Instruments and Facilities

The MIT Tier-2 facility is already at 10 Gbps through a connection in New York. As part of the CMS-HEP tiers, the MIT Tier-2 undergoes regular network bandwidth certification of the link to FNAL.

Vanderbilt's ACCRE compute center has been designated as the main site for processing the raw data from the LHC for the CMS-HI research program. On that basis the Provost's office at Vanderbilt has signed contracts to obtain a 10 Gbps line between ACCRE and the Southern Crossroads (SoX) Gigapop in Atlanta. This link will be in operation at the end of calendar 2008. At this time, the precise network topology of the raw data transport from the CERN Tier-0 to the SoX Gigapop in Atlanta has not been decided. The LHC standard for raw data transfer from the Tier-0 is 10 Gbps. It will be important to establish the precise topology in the next few months so the process of certifying the transfer reliability from CERN to Vanderbilt can begin in early 2009. Discussions among representatives from the Internet2 organization, SoX, Vanderbilt, and CMS management have begun to address this issue.

At nominal year operations, which may occur as early as 2011, CMS-HI is expected to generate 300 TBytes of raw data per year, plus an additional amount of calibration data, to be transferred from the Tier-0 center at CERN to the main CMS-HI compute center at Vanderbilt. Processing of that data may generate annually an additional 300 TBytes of output for analysis. For the ramp up to nominal year operations, a conservative estimate is that in the first year 2009 there may be 50-100 TBytes of raw data, and in the second year 150-200 TBytes of raw data. By extension, the plans for the implementation of the network capabilities should be matched at least to these expected volumes of data.

Process of Science

Procedures for the analysis of the CMS-HI data at Vanderbilt will closely match those for the CMS-HEP data at FNAL. The CMS collaboration’s grid based tools for transferring and verifying the data will be put into operation. Two first pass reconstruction and two second pass analysis runs over the data will be done annually. Some of the reconstructed data can be shipped to the MIT Tier-2 center for subsequent analysis, if it is decided to assign some of the CMS-HI compute hardware at that location.

Once the analyzed data output is produced, the other US institutions will extract physics information by submitting analysis/processing jobs using special CMS collaboration software that takes advantages of the Open Science Grid (OSG) architecture. The Vanderbilt University's HEP group is the leader of the NSF REDDnet project. That project is deploying 500 TBytes of storage, over the next several years, for use by a range of data-intensive application communities, including: elementary particle physics (CMS-HEP), structural biology, supernova modeling, as well as medical and geological imaging. Vanderbilt's HEP group will possess the local expertise to allow CMS-HI to take advantage of REDDnet. Data depots can be placed at selected CMS-HI institutions for use by the all the US collaborators in CMS-HI.

Network Requirements – the next 5 years and beyond

After 5 years time, it is prudent to expect that the raw data volume from CMS-HI may double from 300 to 600 TBytes. At 600 TBytes in 2014, the CMS-HI output in one month would be comparable to that of the PHENIX experiment during several months in 2007.

With a doubling of the raw data volume, then the LAN and WAN network requirements should scale appropriately. Similarly, in 10 years time CMS-HI output is expected to exceed the one Petabyte milestone. Hence, WAN speeds for Tier-0 (CERN) to Tier-1 (Vanderbilt) data flows could easily reach 30 Gbps in 10 years

It should be noted that these scaled up requirements for CMS-HI are correlated with those of the CMS-HEP program in the US. Hence, it is expected that the necessary network infrastructure will exist.

Summary Table

Feature	Science Instruments and Facilities	Process of Science	Anticipated Network Requirements	
Time Frame			Local Area Network Bandwidth and Services	Wide Area Network Bandwidth and Services
Near-term (1-4 years)	• CMS at LHC	• Transfer of ~350 TBytes of all raw HI data during ~1 month from CERN Tier-0 to US (likely main CMS-HI compute site at Vanderbilt U. in Nashville, TN) for reconstruction passes and some analysis • Transfer of ~100 TBytes during ~12 months of baseline reconstructed pp data from FNAL to main CMS-HI compute site for special analysis • Transfers of ~200 TBytes of reconstructed HI data from CMS-HI compute site to MIT Tier-2 site for remaining analysis • Transfers of ~50 TBytes of analyzed data to other CMS-HI sites in US, Europe, and Asia	• Local area network at main CMS-HI compute site already capable of 10 Gbps • Tools for local job submission and monitoring already available	• 10 Gbps (from CERN Tier-0 to main CMS-HI compute site (Vanderbilt) • 10 Gbps from main CMS-HI compute site to MIT Tier-2 site; the MIT Tier-2 site already has 10 Gbps to New York • 2.5 Gbps (duplex) between main CMS-HI compute site and other CMS-HI institutions • Collaboration and OSG software for remote analysis job submission from other CMS-HI institution to main CMS-HI compute site • CMSSW collaboration tools • Deadline scheduling • Use of remote site depots (e.g. REDDNet) for distributed computational analysis
5 years	• CMS at LHC	• Continued data acquisition and processing at twice the volume levels indicated above	• Appropriate upgrades of local infrastructure to match pace of data volumes	• Doubling of WAN speeds quoted above • Continued reliance on collaboration software and OSG
5+ years	• CMS at LHC	• Continued data acquisition and processing at multi-Petabyte levels	• Appropriate upgrades of local infrastructure to match pace of data volumes	• Doubling of WAN speeds quoted above • Continued reliance on collaboration software and OSG

3.3 *Thomas Jefferson National Accelerator Facility*

Background

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

With industry and university partners, it has a derivative mission as well: applied research for using the Free-Electron Lasers 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. Jefferson Lab is managed and operated for the DOE by the 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.

JLab is a user facility offering capabilities that are unique worldwide for an international user community of 1,200 active users. One quarter of all Ph.D.s granted in Nuclear Physics in the U.S. are based on Jefferson Lab research (248 granted, 192 more in progress).

Current Local Area Network Requirements

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 SRF technology used in CEBAF has also enabled the development of the world's highest-average-power Free-Electron Laser (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 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 BNL and FNAL) hosting a distributed Lattice QCD Computing Facility consisting of teraflop/s class supercomputers and clusters tuned to the computing requirements of Lattice QCD (LQCD).

Process of Science

For the Experimental Nuclear Physics Program in the 3 halls, data is acquired in the counting house, monitored live, and transferred to the computer center to be written to tape in files of less than 2 GBytes, typically up to 1 TByte/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 results are written back to disk and to tape, and subsequent batch jobs produce smaller summary files. Most experiments

only transfer the smaller files offsite, although there have been instances of experiments copying all of their data out for analysis at their home institutions.

Detector simulation is more distributed, with some work being carried out at remote institutions, and a larger fraction being done at lower priority on the batch farm. Simulation results are stored in the JLab tape library.

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

For LQCD, large jobs are run at one of the DOE or NSF supercomputing centers, producing space-time (quantum vacuum) configuration files. Typical job sizes are in the thousands of cores. These files are then used as input into large numbers of analysis jobs at BNL, FNAL, and JLab, with typical sizes ranging from 8 to 512 cores but in the aggregate an equally large amount of computing power. Generated propagator files are currently in the range of a few hundred megabytes, growing soon to 20 GBytes as access to larger supercomputers allows for using finer lattices.

Current Wide Area Network Requirements

Instruments and Facilities

Most of the experimental physics data is acquired and analyzed at Jefferson Lab and so the data related 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 on the next page.)

Process of Science

With a staff of about 650 and a user base of over 1,200, there is considerable conventional use of networking (i.e. other than for bulk data transfer), including email, web, and a growing use of videoconferencing. These tools are essential components in the operation of the many collaborations at JLab.

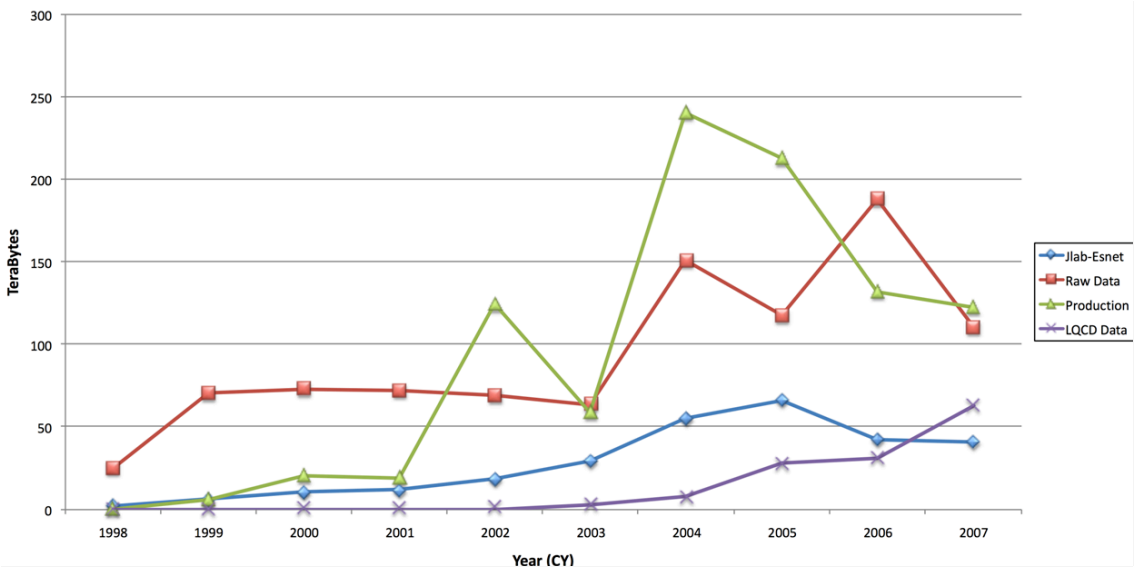


Figure 1: Data volume into the tape library, and from JLab to ESnet.

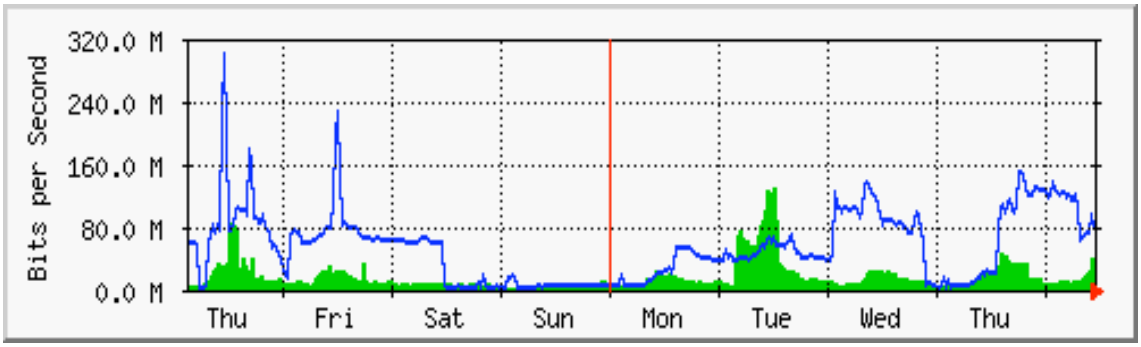


Figure 2: Looking back from 2008 May 2, weekly – 30 min average; max from JLAB – 130.4Mbps (filled green), max to JLAB – 300.4Mbps (blue)

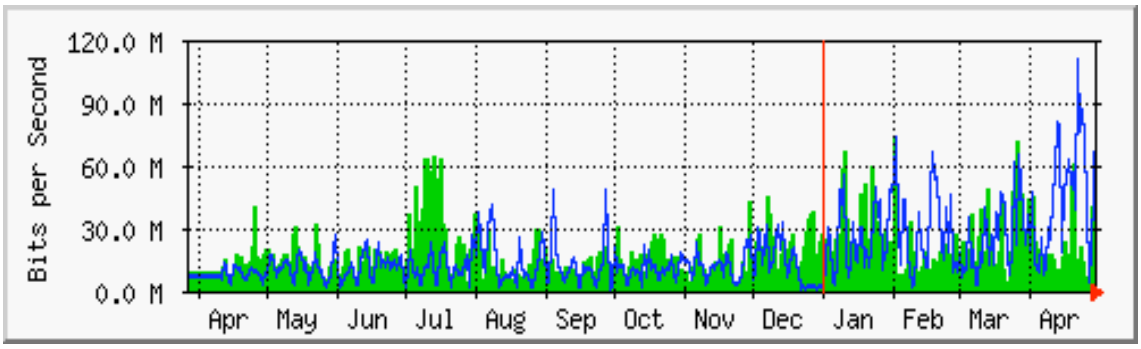


Figure 3: Looking back from 2008 May 2, yearly – 1 day average; max from JLAB – 73Mbps; max to JLAB – 109.8Mbps

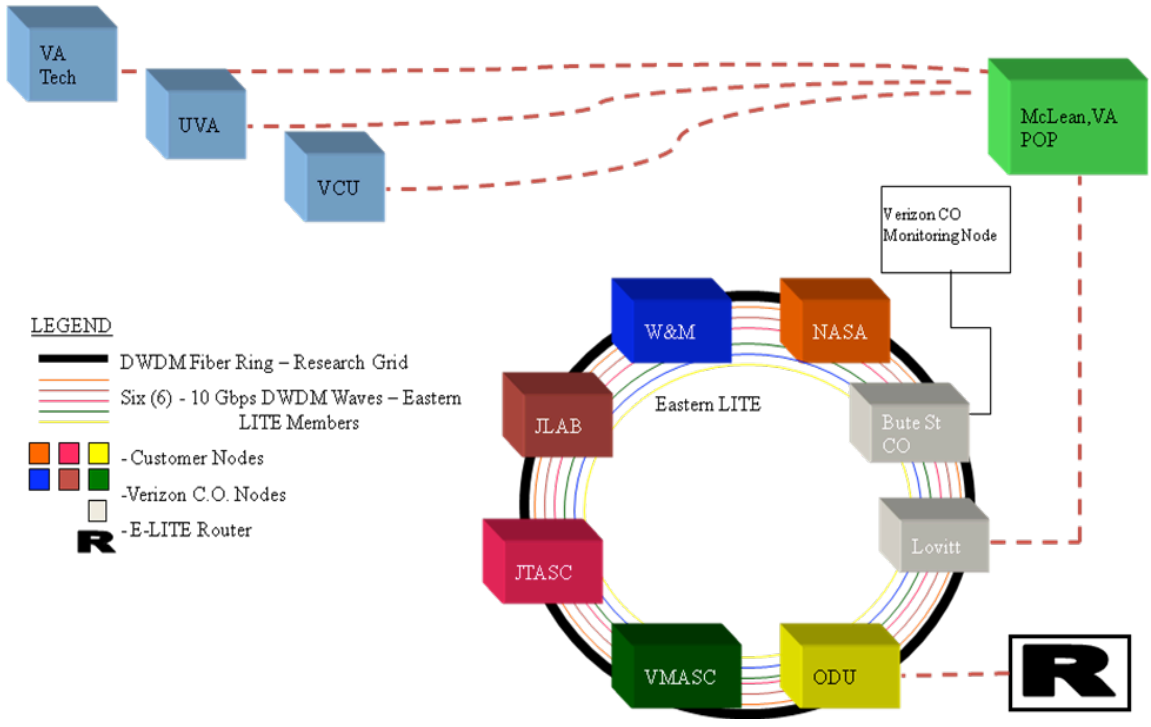


Figure 4: Jefferson Lab's current WAN connection via the E-LITE MAN

JLab has benefited from excellent partnerships and collaborations with ESnet, SURA, JSA, and local Universities and research centers. Local partnerships made possible the Eastern LITE (Lightwave Internetworking Technology Enterprise) or E-LITE metropolitan area network (MAN). ESnet's knowledge and expertise made possible the use of E-LITE and other partnerships to provide JLab with 10Gbps connectivity to ESnet. The multi-wave 10Gbps E-LITE network (JLab's costs paid for by ESnet) provides access to a VORTEX POP located in Norfolk (POP in Norfolk sponsored by ODU). Virginia Optical Research Technology Exchange (VORTEX) provides access to MATP and the ESnet network. JLab's costs in MATP (Mid Atlantic Terascale Partnership) are funded by SURA (Southeastern Universities Research Association).

Local Area Network Requirements – the next 5 years

Instruments and Facilities

The laboratory has embarked on a doubling of the CEBAF energy from 6 GeV to 12 GeV. Using space already available in the accelerator tunnels, ten newer, high performance cryomodules will be installed, and an additional magnet arc will be added to re-circulate the beam for one final pass through the north linac to Hall D. The new experimental Hall D will use the electron beam to produce a coherent bremsstrahlung beam and house a solenoid 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 5 pass, 11 GeV beam. The additional experimental equipment proposed for Halls A, B and C take advantage of currently installed apparatus.

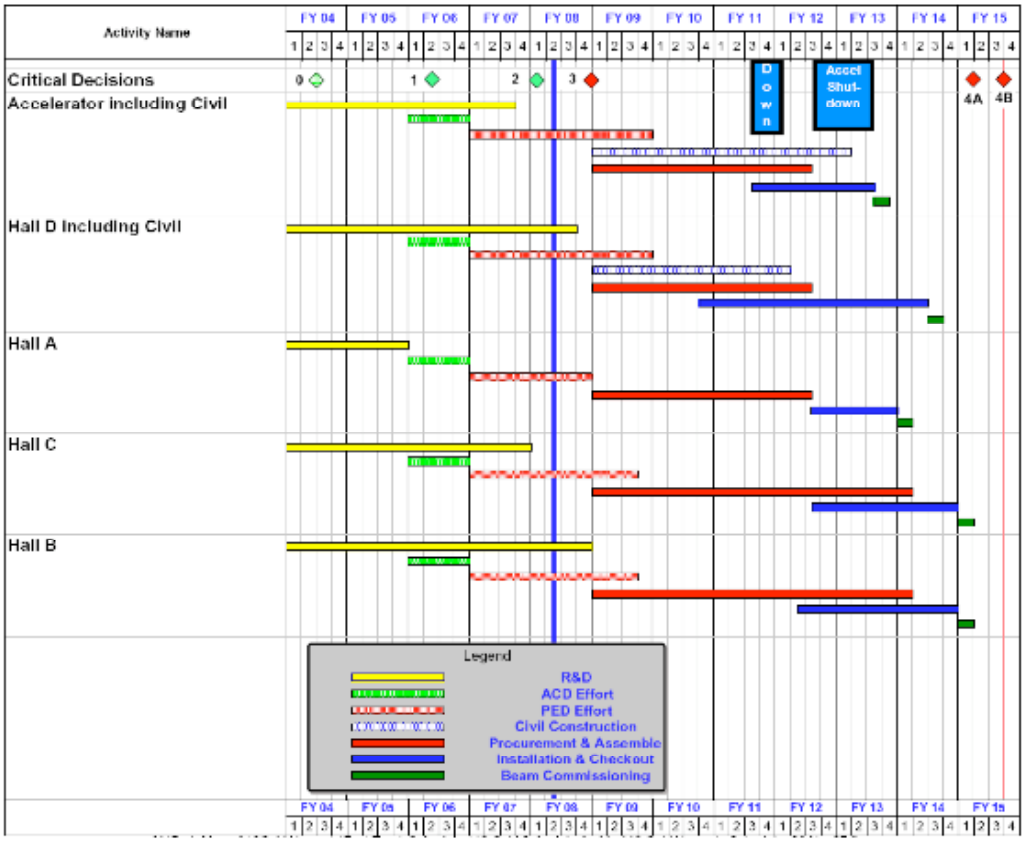


Figure 5: JLab 12 GeV Upgrade Schedule

Running time in the next several years will be budget limited, with the 12 GeV upgrade being the highest priority task. Physics ends for the upgrade shutdown in mid 2011, and resumes in 2015.

Process of Science

Trends in the 6 GeV program show Moore's law outpacing requirements for data analysis. Constant investments (either direct or re-cycling old LQCD nodes) will yield an increasing capacity for simulation.

Requirements for 12 GeV (2012+) will likewise be smaller, in terms of box count, than the current farm. Annual data volume for the 12 GeV program will be about 10x-20x of the 6 GeV program. Thus, Moore's Law allows JLab to continue a simple, mostly 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, followed by Hall D and then Halls A & C. The following spreadsheet contains summary numbers for computational and data volume requirements for CLAS.

Event Simulation	**2012**	**2013**	**2014**	**2015**	**2016**
SPECint_rate2006 sec/event	1.8	1.8	1.8	1.8	1.8
Number of events	1.0E+12	1.0E+12	1.0E+12	1.0E+12	1.0E+12
Event size (KB)	20	20	20	20	20
% Stored Long Term	10%	25%	25%	25%	25%
Total CPU (SPECint_rate2006)	5.7E+04	5.7E+04	5.7E+04	5.7E+04	5.7E+04
Petabytes / year (PB)	2	5	5	5	5
Data Acquisition					
Average event size (KB)	20	20	20	20	20
Max sustained event rate (kHz)	0	0	10	10	20
Average event rate (kHz)	0	0	10	10	10
Average 24-hour duty factor (%)	0%	0%	50%	60%	65%
Weeks of operation / year	0	0	0	30	30
Network (n*10 GigE)	1	1	1	1	1
Petabytes / year	0.0	0.0	0.0	2.2	2.4
1st Pass Analysis	**2012**	**2013**	**2014**	**2015**	**2016**
SPECint_rate2006 sec/event	1.5	1.5	1.5	1.5	1.5
Number of analysis passes	0	0	1.5	1.5	1.5
Event size out / event size in	2	2	2	2	2
Total CPU (SPECint_rate2006)	0.0E+00	0.0E+00	0.0E+00	7.8E-03	8.4E-03
Silo Bandwidth (MB/sec)	0	0	900	900	1800
Petabytes / year	0.0	0.0	0.0	4.4	4.7
Total SPECint_rate2006	5.7E+04	5.7E+04	5.7E+04	5.7E+04	5.7E+04
SPECint_rate2006 / node	600	900	1350	2025	3038
# nodes needed (that year)	**95**	**63**	**42**	**28**	**19**
Petabytes / year	**2**	**5**	**5**	**12**	**12**

Wide Area Network Requirements – the next 5 years

Instruments and Facilities

A good estimate of WAN requirements would be that it would scale like data volume, but with a mostly central computing model (with modest requirements), this would overestimate the networking requirements.

Since data rates will remain constant or decrease between now and the shutdown (2011), the current 10 Gbps WAN will remain more than adequate in the next five years. In 2015, as the 12 GeV machine turns on, requirements might grow beyond 10 Gbps.

The LQCD Computing Facility should also grow significantly in the next 5 years, roughly 10x by following Moore's Law with nearly constant investments. With a corresponding increase in off-site supercomputing cycles, LQCD could grow to be a larger contributor to WAN networking than experimental physics in the 6 GeV era.

Process of Science

Use of distributed computing models (web 2.0, grid, cloud, etc.) will continue to grow. Conventional use, including videoconferencing, will also steadily increase as these technologies become ever more widespread. It is difficult to quantify this growth in terms of network bandwidth and other capabilities.

Beyond 5 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 Office of Science accelerator projects at multiple locations (FRIB, ILC, Project X, SNS II, etc.) and could potentially lead to additional facilities on the campus such as an Electron Ion Collider (ELIC at Jefferson Lab and/or eRHIC at BNL) and a new 4th Generation Light Source.

Summary Table

Time Frame	Science Instruments and Facilities	Process of Science	Anticipated Requirements	
			Local Area Network Bandwidth and Services	Wide Area Network Bandwidth and Services
Near-term	• 6 GeV accelerator and related halls • LQCD Computing	• Detector simulation, data analysis, mostly lab-centric batch	• 1 GigE / 10 GigE hybrid infrastructure being upgraded to all 10 GigE	• 10 Gbps
5 years	• Larger LQCD facility	• No change	• Expanding use of 10 GigE	• Slowly fill 10 Gbps
5+ years	• 12 GeV program	• Possibility of another science instrument	• (tbd)	• (tbd)

3.4 LBNL/NERSC NP Heavy Ion Program

Background

The Relativistic Heavy Ion physics program at LBNL includes participation in the STAR experiment at RHIC/BNL and the ALICE experiment at LHC/CERN. The NERSC facility serves as a major computational facility for this program providing resources to local researchers as well as collaborators nationally and internationally.

The main physics thrust of the program is the study of matter under the most extreme conditions of energy density available in the laboratory caused by the collisions of atomic nuclei at relativistic energies. This encompasses studies over a broad range of beam energies and masses; from a few GeV/c to 100 GeV/c at RHIC (250 GeV/c for protons) and up to 2.5 TeV/c at the LHC (7 TeV/c for protons); spanning protons to gold and lead nuclei.

The experimental program at RHIC began in 2000 and recently concluded run 8 while the LHC (ALICE) expects first beam collisions in 2009/10. STAR is upgrading the data acquisition system in the next year which will cause increased dataset sizes in the coming years.

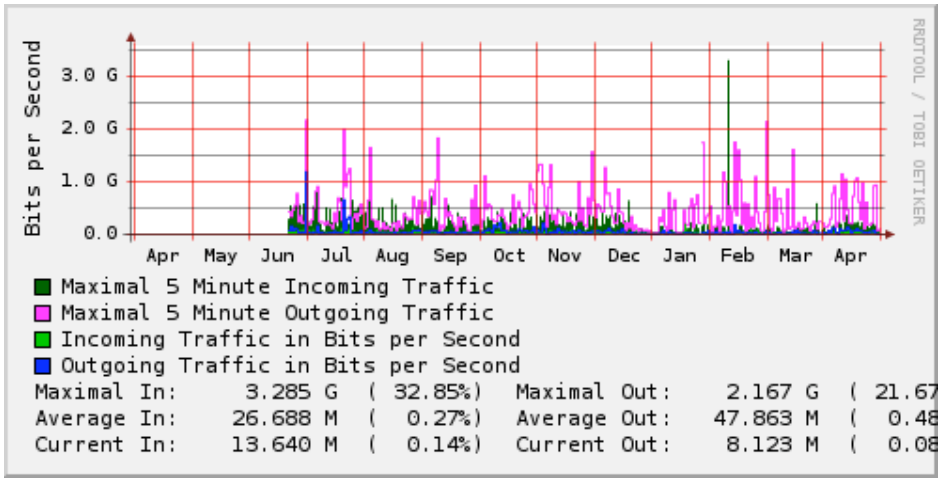
Current Local Area Network Requirements

Instruments and Facilities

STAR:

The local network at NERSC/PDSF shows current rates of utilization of ~ 50 Mbps average (~ 3 Gbps peak).

PDSF Network June 2007 – April 2008



ALICE:

The PDSF facility at NERSC is used for analysis and simulations for ALICE.

Process of Science

STAR:

Currently supports managed bulk data transfer to/from BNL, data analysis and simulations at PDSF/NERSC. Users mostly login via ssh and run locally. Simulation workload has migrated to grid interfaces (GRAM, GridFTP) doing data transfer on WAN as part of workflow using srm/gridftp. Most of workload is done via jobs locally submitted to batch system with minimal random data transfer on WAN. Data analysis is entirely based on locally available datasets at this point, but it is anticipated that this will change in the future with more on-demand WAN data access.

ALICE:

Analysis of simulated data for ALICE currently consists of manual procedures using the aliroot code and manual data handling and processing in the AliEn framework, which is a distributed processing data management and workflow environment for ALICE. AliEn uses X509 grid credentials for authentication and the U.S. participants get their X509 certificates from the DOEGrids CA operated by ESnet.

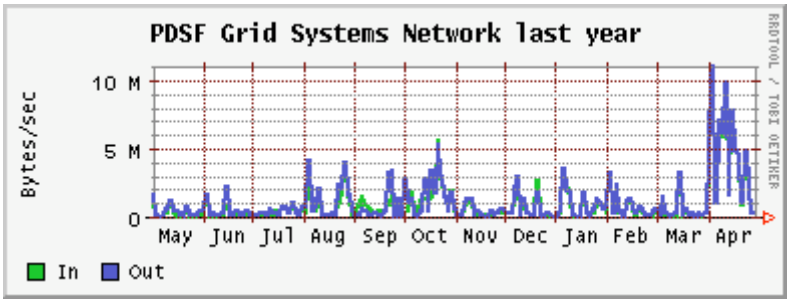
Current Wide Area Network Requirements

Instruments and Facilities

STAR:

Currently the main WAN traffic is between BNL and NERSC.

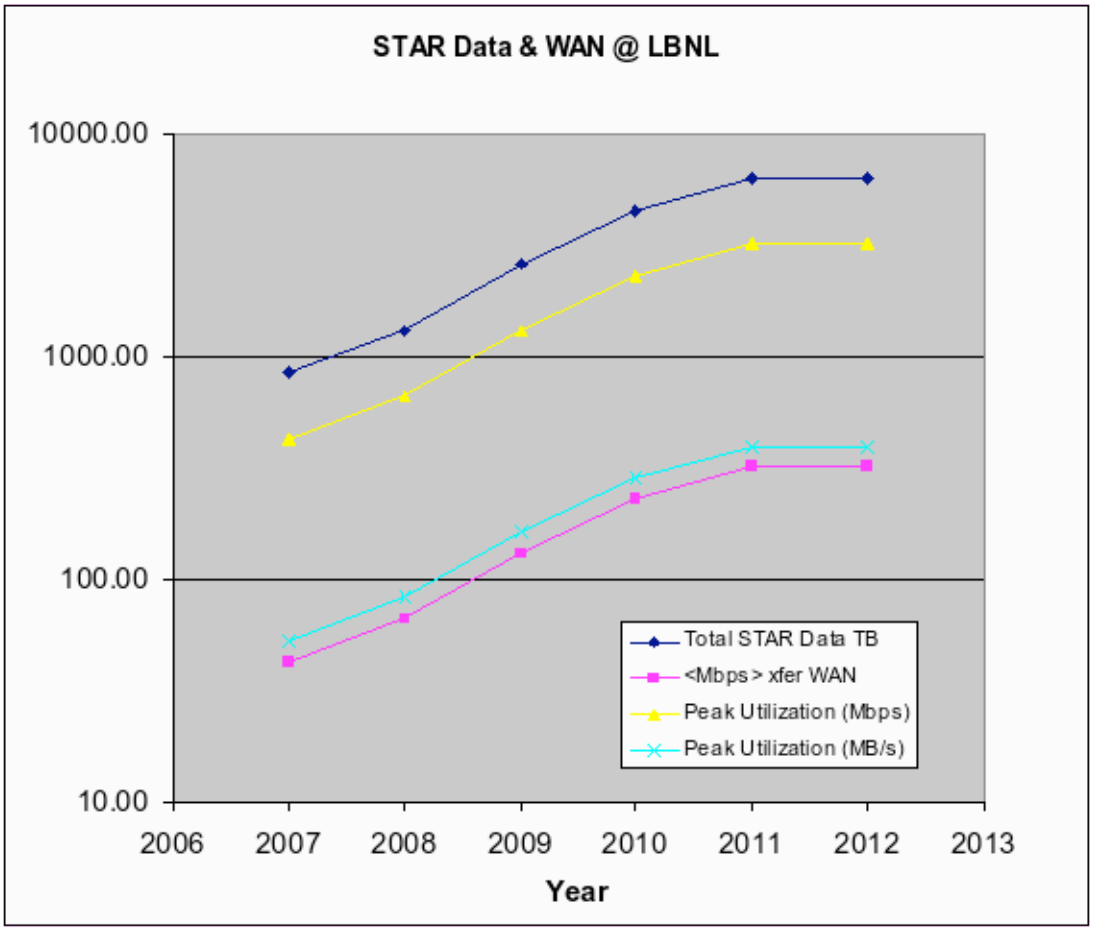
PDSF Ganglia report from Grid interfaces, May 2007 – April 2008.



The graph above shows the average network utilization by the grid interface systems at PDSF, reflecting primarily the WAN data transfer.

The graph below show the projections from current usage into the future reflecting planned upgrades of the experiment and computational facilities assuming a similar distribution of workload as exists today where the primary network utilization is between BNL and NERSC. It is anticipated that the workload distribution may change such that more of the overall STAR workload is distributed to many more sites nationally and internationally, and that NERSC is likely to become a mirror site to BNL for distributing reconstructed data and micro-DST data sets for analysis at universities. This would

increase the network load from NERSC by some factor, probably on the order of 2-5 times.



ALICE:

The LHC computing model has Tier-1 facilities located in countries outside of CERN which are connected to CERN via optical private network (LHC-OPN). In the U.S., FNAL is the Tier-1 facility for CMS and BNL is the Tier-1 facility for ATLAS. For ALICE the data connection to CERN is via LBNL/NERSC which is described as part of the U.S. Tier-2 Federation with Tier-1 capabilities. LBNL/NERSC is expected to store about 8-10% of the raw data for ALICE and it will serve data of several other centers in the Americas. At this point in time, the other computational facilities in the Americas are LLNL, OSU/OSC, U. Houston, UNAM Mexico and the Brazil T2 Federation.

From the LCG Technical Design Report, June 2005.

> The p–p raw data are immediately reconstructed at the CERN Tier-0 facility and exported to the different Tier-1 centers. For heavy-ion data, the quasi real-time processing of the first reconstruction pass, as it will be done for p–p data, would require a prohibitive amount of resources. ALICE therefore requires that these data be reconstructed at the CERN Tier-0 and exported over a four-month period after data taking. During heavy-ion data taking only pilot reconstructions and detector calibration activities will take place. Additional reconstruction passes (on average three passes over the entire set of data are

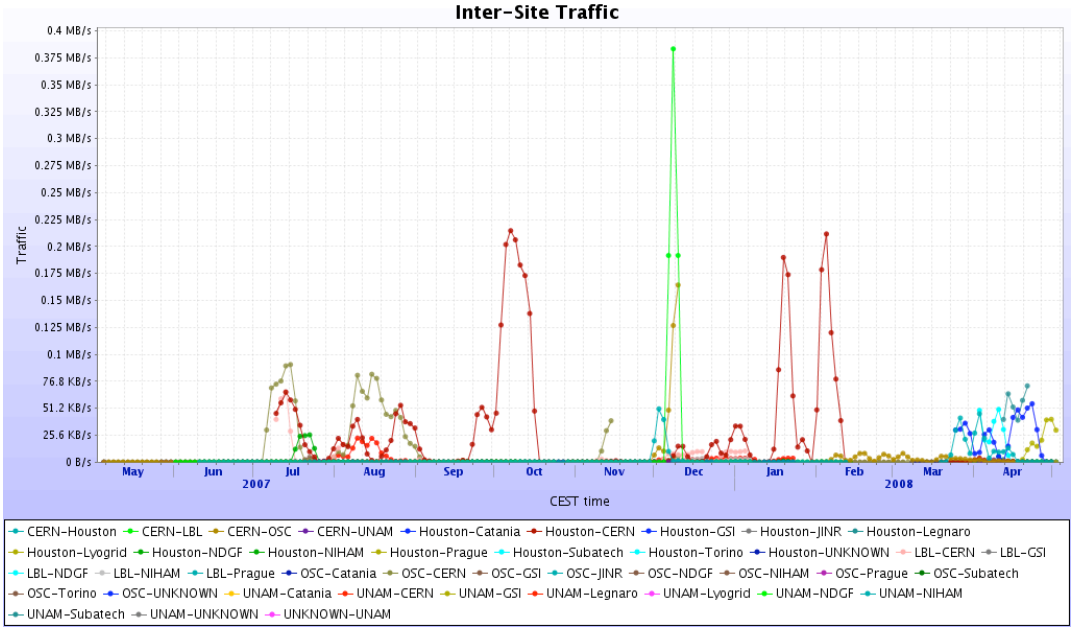
considered in the ALICE computing model) for p–p and heavy-ion data will be performed at Tier-1s, including the CERN Tier-1.

Raw data will be available in two copies, one at the Tier-0 archive and one distributed at the Tier-1 sites external to CERN. The reconstruction process generates Event Summary Data (ESD) with raw data size reduction coefficient of 10 for heavy-ion and 25 for p–p. Two copies of the ESD are distributed in the Tier-1 sites for archive. Multiple copies of active data (a fraction of raw data, the current set of ESD) are available on fast-access storage at Tier-1 and Tier-2 sites.

Analysis is performed directly on ESD or on Analysis Object Data (AOD). Two analysis schemes are considered: scheduled and chaotic. The scheduled analysis tasks are performed mainly at Tier-1 sites on ESD objects and produce various different AOD objects. These tasks are driven by the needs of the ALICE Physics Working Groups. The chaotic analysis tasks (interactive and batch) are launched by single users predominantly on AODs but also on ESD. These tasks are mainly processed at Tier-2 sites. It is the responsibility of the Tier-2 sites to make the data available on disk to the collaboration, the archiving being the responsibility of the Tier-1 sites.

To date, seven sites (including CERN) have pledged Tier-1 services to ALICE and about 14 sites (including CERN) have pledged Tier-2 services. The amount of resources provided by the various Tier-1 and Tier-2 sites is very uneven with a few Tier-1s providing a relatively small contribution compared to others. CERN will host the Tier-0, a comparatively large Tier- 1 with no archiving responsibility for the raw data (this will be done by the associated Tier-0), and a Tier-2. During the first-pass reconstruction, when all the CPU resources installed at CERN are required for the Tier-0, no computing tasks typically allocated to Tier-1 and Tier-2 will be processed at CERN. The first-pass reconstruction for heavy-ion data is scheduled to take place during the four months following the heavy-ion data-taking period.

The ALiEn monitoring graph below shows the inter-site traffic touching the facilities in the Americas for the past year. The bandwidth is not large (no beam data yet) but it shows a significant number of international connections.



Process of Science

STAR:

WAN data transfer is carried out in bulk-managed fashion with local catalogs showing what datasets are available for local analysis.

Current collaboration technologies employed are primarily teleconference and web publishing, with an increasing amount of IP-based teleconferencing (Skype).

Web publishing is becoming problematic with current methods of access control/authorization causing trouble and significant disruptions of service.

ALICE:

Data in ALICE is published into a catalog (AliEn) globally and is available to anyone in the collaboration for analysis. This allows (encourages) inter-site data transfers between participating ALICE sites, as can be seen in the graph above for "Inter-Site Traffic". Depending on the characteristics of the computational workload there are many cases where the input data is not moved but the jobs go to where the data is located and only the output data is transferred on the WAN.

ALICE scientists currently uses the AliEn environment for doing data analysis and simulations use grid credentials for authentication. The U.S. scientists get their certificates from the DOEGrids CA operated by ESnet. It is estimated that there will be 50 or more U.S scientists in ALICE in the next year or two.

Local Area Network Requirements – the next 5 years

Instruments and Facilities

The LAN network and computational requirements are expected to scale approximately linearly with the total data size. It may be that it scales more closely as the integral of all the data rather than the new data each year, but with dataset sizes increasing each year that is not a large discrepancy. The graph above shows the projections.

Process of Science

STAR:

Most computational workflow is expected to occur via grid interfaces to facilities and all STAR collaborators will be organized as a grid Virtual Organization, using grid credentials for authentication, and grid tools (Open Science Grid software) for data and job management. STAR has its own workflow management system (SUMS) providing the user interface to grid resources.

ALICE:

ALICE data analysis and simulations is handled within the AliEn environment and uses the Aliroot framework for software integration and Xrootd as the storage and data access system.

Wide Area Network Requirements – the next 5 years

Instruments and Facilities

STAR:

See description with graph above in the "Current WAN Requirements" for STAR.

ALICE:

The table below shows the estimated peak bandwidth utilization between LBNL/NERSC and CERN, and LBNL/NERSC and the other T2 centers in the Americas. In this case, peak bandwidth means largest mean bandwidth utilization in a month during the year, so one can expect the instantaneous bandwidth peaks to be significantly greater.

Type \ Year	2007	2008	2009	2010
LBNL-CERN Mbps	5.2	43	248	248
LBNL-T2s Mbps (sum)	136	680	624	624

Process of Science

STAR:

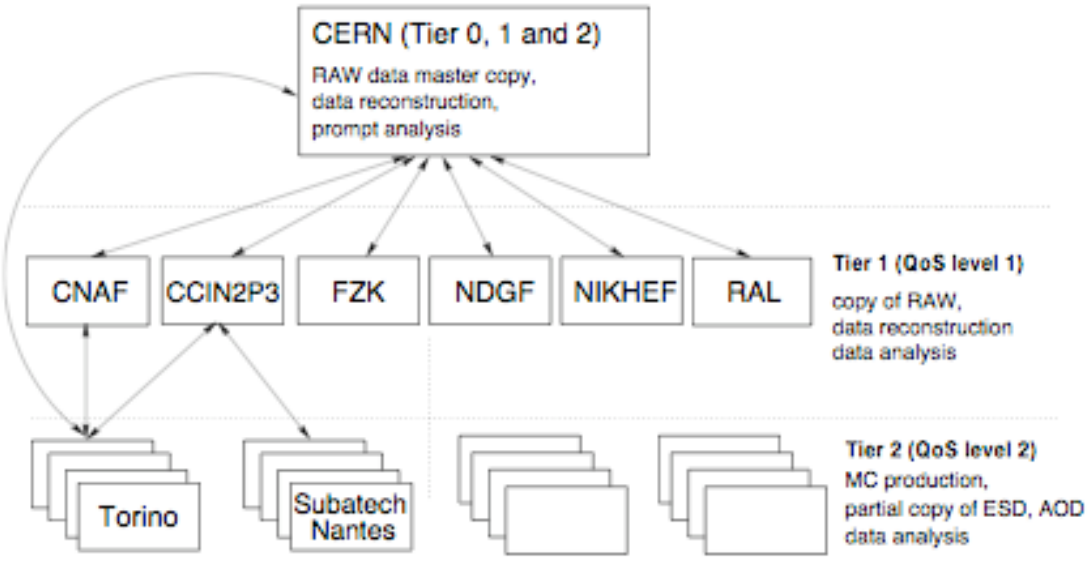
Significantly more on-demand WAN data transfer is expected to support individual analyses, in addition to the managed bulk placement of datasets.

For the ESnet PKI, this implies a 10-20 fold increase in grid credentials (from 10s to 100s).

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). However, a service that integrated with the grid authorization services so that STAR collaborators already registered as members of the VO could use the collaboration services without additional registration steps would be helpful.

ALICE:

From ALICE Computing TDR, June 2005:



Schematic view of the ALICE offline computing tasks in the framework of the Tiered MONARC model.

The ALICE computing model foresees that one copy of the raw data from the experiment will be stored at CERN (Tier-0) and a second copy will be distributed among the external (i.e. not at CERN) Tier-1 centers, thus providing a natural backup. Reconstruction to the Event Summary Data (ESD) level will be shared by the Tier-1 centers, with the CERN Tier-0 responsible for the first reconstruction pass. Subsequent data reduction to the Analysis Object Data (AOD) level, analysis and Monte Carlo production will be a collective operation where all Tiers 1 and 2 will participate. The Tier-1's will perform reconstruction and scheduled analysis, while the Tier-2's will perform Monte Carlo and end user analysis.

Beyond 5 years – future needs and scientific direction

STAR:

There has been extensive R&D and planning for an upgrade to the RHIC complex to increase beam luminosity about 10-fold using electron cooling of the ion beams. This is called RHIC-II and would likely occur in the 5+ year time frame. A consequence of the increased luminosity is an increase in data acquired and subsequent increase in network bandwidth requirements, scaling approximately with the increase in data.

ALICE:

The LHC program will be achieving maturity in the 5+ years time frame and, using history as a guide, ALICE and the other experiments will have achieved ever increasing data rates and computational requirements, following approximately Moore's Law.

Summary Table

Time Frame	Science Instruments and Facilities	Process of Science	Anticipated Requirements	
			Local Area Network Bandwidth and Services	Wide Area Network Bandwidth and Services
Near-term	• NERSC/PDSF for STAR & ALICE:	• STAR - Managed bulk data transfer BNL to NERSC for data analysis. Production computing moving to grid interfaces. End user data analysis by submission to local scheduler with local data catalog. • ALICE – Production computing and data transfer via AliEn grid environment. End user simulation & analysis via submission to local batch system. Analysis of soon to arrive real data will use AliEn grid tools. • Both – X509 grid credentials from ESnet, many teleconferences, some using ESnet	• NERSC LAN usage moving from 3 Gbps peak to 10 Gbps peak.	• Increasing from 1 to 5 Gbps utilization in next few years to/from U.S sites plus international sites listed below. • 1 Gbps to/from CERN. • Expecting 0.1 Gbps to Brazil, 0.01 Gbps to Mexico
5 years	• NERSC/PDSF for STAR & ALICE	• Most computational workload using grid interfaces. Implies 100's of grid credentials. • Significant on-demand data transfer in addition to managed transfer.	• LAN requirements scale with CPU capacity. Aggregate bandwidth should by multi 10Gbps.	• 10 Gbps from NERSC. • 2 Gbps to/from CERN • B/w to Mexico and Brazil should increase by unknown amount.
5+ years	• NERSC/PDSF for STAR & ALICE	• Continuation of previous trends.	• Continuation of previous trends.	• 15-20 Gbps from NERSC. • CERN bandwidth likely to reach 10 Gbps. • Brazil and Mexico likely to increase by unknown amount.

4 Findings

The following issues were reported and discussed at the workshop.

Connectivity and Bandwidth

An important finding is that both STAR and PHENIX will require a significant amount of international bandwidth. Scientific productivity is enhanced through the availability of local data access by researchers. A significant number of scientific publications were observed in case studies where data was made easily accessible to new groups of scientists, thus providing a strong argument favoring bulk data transfer from instrument sites (e.g. RHIC) to repositories housed at collaborating institutions.

For PHENIX, the key international site is the CC-J facility at RIKEN. Bandwidth requirements to RIKEN will be 2-3 Gbps in the near term and 10-20 Gbps in 5-6 years. The PHENIX data cache size at BNL is finite (it can store approximately 20 hours of data), and it is important to empty the cache quickly enough to avoid overwriting data. Therefore PHENIX requires bandwidth guarantees to ensure this does not happen, as well as sufficient network redundancy to ensure that no long-duration outages occur. Also, PHENIX will undergo significant increases in bandwidth in 2009 and 2011.

For STAR, KISTI in Korea is a potential new Tier-1 center which is expected to require 4 Gbps bandwidth from BNL to KISTI by 2010. Other international sites for STAR include Prague, Czech Republic, and possibly Birmingham, England. These sites may have to compete with the LHC Tier-1 to Tier-2 data flow traffic on the shared transatlantic IP links, which could become an issue.

For ALICE, NERSC/PDSF is the main US site, and will effectively act as a Tier-1 site. Other key sites are LLNL, Ohio State University, and the University of Houston. Bandwidth requirements are relatively modest due to the fact that ALICE only collects data 1 month/year, but data transfer then takes place over 4 months.

Another the finding is that dynamic virtual circuits will be important for CMS-HI research program, as they will only be collecting data for 1 month/year. The key US sites for CMS-HI are MIT and Vanderbilt University.

End-to-End Performance

The Nuclear Physics community is using GridFTP for all its underlying bulk data transport, and hence can usually fully utilize the network as long and end hosts are tuned and there are no firewall issues. Currently there is an issue with transfers between PDSF at NERSC and BNL that is being worked on. Part of the problem is very old versions of Linux at NERSC that do not support TCP auto-tuning and hosts that only have 100 Mbps network interfaces.

Collaboration Services

The Nuclear Physics community depends heavily on the certificate infrastructure supported by the Open Science Grid. There was concern about long-term support for OSG, as they provide an essential service for this community.

Video and audio conferencing services are heavily used by the Nuclear Physics community and this will likely increase in the future. In particular, IP telephony applications such as Skype are playing an increasing role.

5 Requirements Summary and Conclusions

Most of the key DOE sites for NP related work will require significant increases in network bandwidth in the 5 year time frame. This includes roughly 40 Gbps for BNL, and 20 Gbps for NERSC. Total transatlantic requirements are on the order of 40 Gbps, and transpacific requirements are on the order of 30 Gbps. Other key sites are Vanderbilt University and MIT, which will need on the order of 20 Gbps bandwidth to support data transfers for the CMS Heavy Ion program.

In addition to bandwidth requirements, the workshop emphasized several points in regard to science process and collaboration. One key point is the heavy reliance on Grid tools and infrastructure (both PKI and tools such as GridFTP) by the NP community. The reliance on Grid software is expected to increase in the future. Therefore, continued development and support of Grid software is very important to the NP science community. Another key finding is that scientific productivity is greatly enhanced by easy researcher-local access to instrument data. This is driving the creation of distributed repositories for instrument data at collaborating institutions, along with a corresponding increase in demand for network-based data transfers and the tools to manage those transfers effectively. Network reliability is also becoming more important as there is often a narrow window between data collection and data archiving when transfer and analysis can be done. The instruments do not stop producing data, so extended network outages can result in data loss due to analysis pipeline stalls. Finally, as the scope of collaboration continues to increase, collaboration tools such as audio and video conferencing are becoming ever more critical to the productivity of scientific collaborations.

Action Items

The action items for ESnet that came out of this workshop include:

- Work with STAR to ensure good bandwidth to Korea. This work as already begun, with active ongoing collaboration between ESnet, BNL, and KISTI.
- There were general concerns expressed at the workshop about redundant connectivity for BNL. This issue is currently being addressed, and BNL will have fully redundant paths soon.
- Work with SINET to ensure bandwidth requirements to CCJ in Japan are met.
- Help ensure that bandwidth requirements are met for CMS-HI from CERN to Vanderbilt University, and from Vanderbilt to MIT.
- Continue development and deployment of the ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS - http://www.es.net/oscars/), as bandwidth management is important to the NP community.

6 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 NP and ASCR program offices for their help in organizing the workshop and providing insight into the science supported by the NP program. In addition, the LBNL conference support and logistics staff was very helpful.

ESnet is funded by the US Department of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program. 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 US 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.