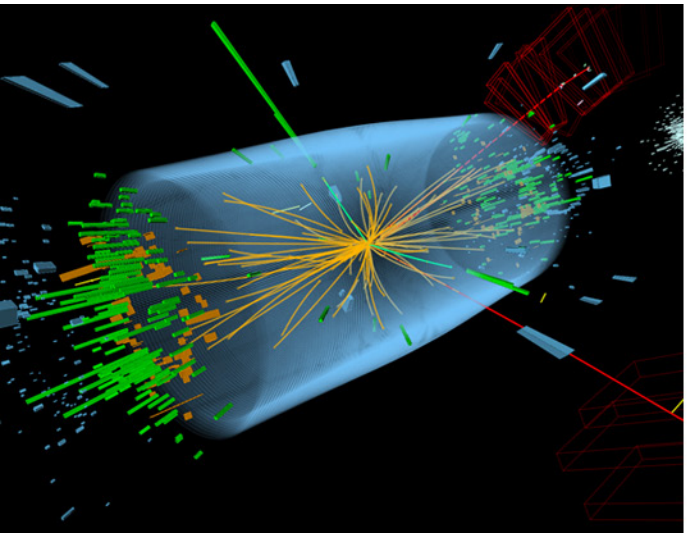
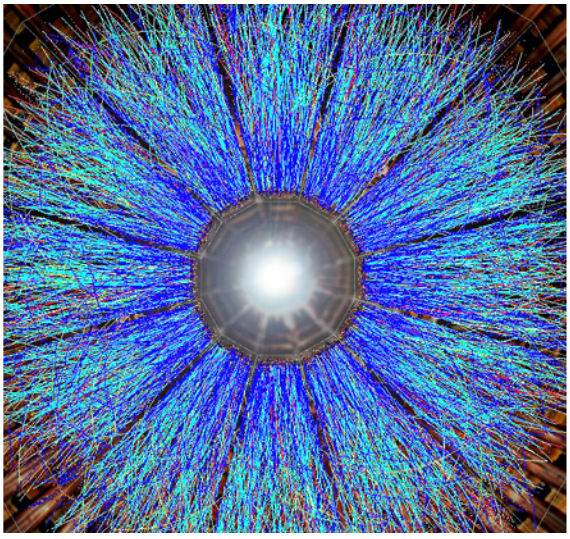
High Energy Physics and Nuclear Physics Network Requirements

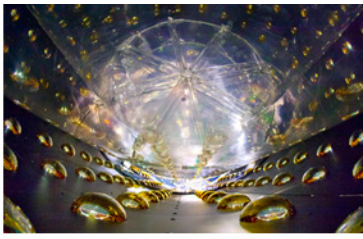
HEP and NP Network Requirements Review
Final Report

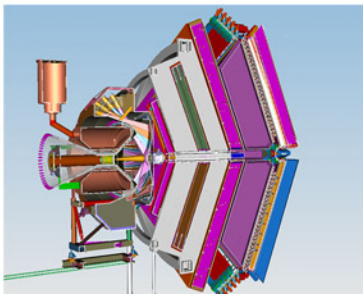
Conducted August 20-22, 2013

ESnet
Energy Sciences Network









BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

DISCLAIMER

High Energy Physics and Nuclear Physics Network Requirements

Offices of High Energy Physics and Nuclear Physics, DOE Office of Science
Energy Sciences Network
Gaithersburg, Maryland — August 20–22, 2013

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 Offices of High Energy Physics and Nuclear Physics.

This is LBNL report LBNL-6642E

Participants and Contributors

Lothar Bauerdick, FNAL (LHC/CMS)

Greg Bell, ESnet (Networking)

Leandro Ciuffo, RNP (Networking)

Eli Dart, ESnet (Networking)

Sridhara Dasu, University of Wisconsin (LHC/CMS)

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

Kaushik De, University of Texas at Arlington (LHC/ATLAS)

Michael Ernst, BNL (LHC/ATLAS, RACF)

Dale Finkelson, Internet2 (Networking)

Steven Gottlieb, Indiana University (Lattice QCD)

Oliver Gutsche, FNAL (LHC/CMS)

Salman Habib, ANL (Cosmic Frontier Simulations)

Stefan Hoeche, SLAC (Non-Lattice QCD)

Richard Hughes–Jones, DANTE (Networking)

Julio Ibarra, FIU (LSST)

Bill Johnston, ESnet (Networking)

Theodore Kisner, LBNL (DESI)

Andy Kowalski, JLab (JLab Experiments)

Jerome Lauret, BNL (RHIC/STAR)

Steffen Luitz, SLAC (SLAC Programs)

Paul Mackenzie, FNAL (Lattice and Non-Lattice QCD)

Charles Maguire, Vanderbilt University (LHC/CMS-HI)

Joe Metzger, ESnet (Networking)

Inder Monga, ESnet (Networking)

Cho-Kuen Ng, SLAC (Accelerator Modeling)

Jason Nielsen, UC Santa Cruz (LHC/ATLAS)

Larry Price, DOE/SC/HEP (HEP Program)

Jeff Porter, LBNL (LHC/ALICE)

Martin Purschke, BNL (RHIC/PHENIX)

Gulshan Rai, DOE/SC/NP (NP Programs)

Rob Roser, FNAL (Intensity Frontier Experiments)

Malachi Schram, PNNL (Belle II)

Craig Tull, LBNL (Daya Bay)

Chip Watson, JLab (JLab Experiments)

Jason Zurawski, ESnet (Networking)

Editors

Eli Dart, ESnet — dart@es.net

Mary Hester, ESnet — mchester@es.net

Jason Zurawski, ESnet — zurawski@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. In support of SC programs, ESnet regularly updates and refreshes its understanding of the networking requirements needed by 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 25 years.

In August 2013, ESnet and the DOE SC Offices of High Energy Physics (HEP) and Nuclear Physics (NP) organized a review to characterize the networking requirements of the programs funded by the HEP and NP program offices.

Several key findings resulted from the review. Among them:

1. The Large Hadron Collider's ATLAS (A Toroidal LHC Apparatus) and CMS (Compact Muon Solenoid) experiments are adopting remote input/output (I/O) as a core component of their data analysis infrastructure. This will significantly increase their demands on the network from both a reliability perspective and a performance perspective.
2. The Large Hadron Collider (LHC) experiments (particularly ATLAS and CMS) are working to integrate network awareness into the workflow systems that manage the large number of daily analysis jobs (1 million analysis jobs per day for ATLAS), which are an integral part of the experiments. Collaboration with networking organizations such as ESnet, and the consumption of performance data (e.g., from perfSONAR [PERformance Service Oriented Network monitoring Architecture]) are critical to the success of these efforts.
3. The international aspects of HEP and NP collaborations continue to expand. This includes the LHC experiments, the Relativistic Heavy Ion Collider (RHIC) experiments, the Belle II Collaboration, the Large Synoptic Survey Telescope (LSST), and others. The international nature of these collaborations makes them heavily reliant on transoceanic connectivity, which is subject to longer term service disruptions than terrestrial connectivity. The network engineering aspects of undersea connectivity will continue to be a significant part of the planning, deployment, and operation of the data analysis infrastructure for HEP and NP experiments for the foreseeable future. Given their critical dependency on networking services, the experiments have expressed the need for tight integration (both technically and operationally) of the domestic and the transoceanic parts of the network infrastructure that supports the experiments.
4. The datasets associated with simulations continue to increase in size, and the need to move these datasets between analysis centers is placing ever-increasing demands on networks and on data management systems at the supercomputing centers. In addition, there is a need to harmonize cybersecurity practice with the data transfer performance requirements of the science.

This report expands on these points, and addresses others as well. The report contains a findings section in addition to the text of the case studies discussed during the review.

2 Findings

The data staging model for the LHC experiments has gone through several evolutionary phases. The original model was based on programmatic replication of datasets from a Tier-1 center to Tier-2 sites. The second phase of the model used on-demand replication, driven by workflow management systems. The model currently being developed is based both on replication and on remote I/O, built on XrootD. Over time, as the remote I/O paradigm matures, more emphasis will be put on remote I/O. The ATLAS experiment calls this the Federated ATLAS XrootD system (FAX), and the CMS experiment calls it AAA (Any data, Any time, Anywhere). Each change has increased the experiments' reliance on network infrastructure stability and performance. Remote I/O using XrootD is likely to increase network utilization at the Tier-2 sites, especially if both Tier-2 and Tier-3 sites rely heavily on remote I/O to access data stored at the larger Tier-2 sites.

The LHC experiments need improved network infrastructure at the Tier-2 sites, which are located primarily on university campuses. The Tier-2 sites play an increasingly important role, in many cases serving primary datasets to other sites (including Tier-2 and Tier-3 sites). That data service role, combined with the adoption of remote I/O technologies, will place significantly higher performance and reliability demands on network infrastructure at the Tier-2 sites. It is expected that several Tier-2 sites will have 100 GbE network connections within the next two years.

The ATLAS experiment wants greater integration between the ATLAS workflow manager PanDA (Production and Distributed Analysis) and the network infrastructure. A project called BigPanDA is under way to build network awareness into PanDA, but more integration with network performance data sources (e.g., with data from perfSONAR or other network monitoring systems) is desired.

The CMS and ATLAS LHC experiments plan to have all Tier-2 sites up and running with XrootD remote I/O by the spring of 2014. This will allow time for testing and hardening of the new software infrastructure before the LHC starts Run 2 in 2015.

The CMS Heavy Ion data volume will increase by a factor of 2 for LHC Run 2. This experiment is expected to produce 1–2 PB/yr but this data will be taken in one month and transferred from Fermi National Accelerator Laboratory (Fermilab) to Vanderbilt University. This may put stress on the path between Fermilab and Vanderbilt (particularly the shared 10 GbE interfaces at the SoX exchange). This will need to be monitored.

The reliability, predictability, and maintainability of the network is becoming increasingly important. Network design for these attributes has become a critical aspect of infrastructure provisioning for data-intensive workflows. This arose in discussions surrounding multiple case studies, including the LHC experiments, Daya Bay, and cosmological simulation case studies.

It is likely that the STAR (Solenoidal Tracker At RHIC) experiment will increase its data exchange with Asian sites, primarily the Korean Institute of Science and Technology Information (KISTI) in South Korea.

The Thomas Jefferson National Accelerator Facility (JLab) wants to increase the reliability of its ESnet connectivity by adding redundant or load-sharing connections. This will be affected by the re-bid of the E-LITE (Eastern Lightwave Internetworking Technology Exchange) regional network in the coming months.

JLab is a heavy user of ESnet Collaboration Services (ECS), which provides both ReadyTalk and videoconferencing services. JLab is generally happy with these services, and would like to see them continue.

The Intensity Frontier HEP experiments do not have the human scale of the Energy Frontier HEP experiments (e.g., LHC/ATLAS, LHC/CMS). Because of this, in many cases the Intensity Frontier experiments will make use of tools and infrastructure developed by the larger collaborations of the Energy Frontier experiments.

Many smaller HEP efforts are organized around a principal investigator (PI) rather than a single experiment or facility. There was a consensus at the review that common frameworks, documentation of best practices, and code/tool reuse would significantly benefit the smaller collaborations.

Based on the experience of the Daya Bay Neutrino Experiment (for trans-Pacific connectivity) and the LHC experiments (for trans-Atlantic connectivity), significant effort is required to engineer the undersea network paths such that experiment operations are not interrupted by cable outages. This affects the LHC experiments, Belle II, RHIC/STAR connectivity to Korea, the LSST, and other efforts with collaborations that span multiple continents.

The Belle II experiment is conducting a series of data and service challenges over the next two years in preparation for the experiment's operations. This will require coordination among ESnet, Pacific Northwest National Laboratory (PNNL), the High Energy Accelerator Research Organization in Japan (KEK), the Science Information Network (SINET), Karisruhe Insitute of Technology (KIT), GEANT (Gigabit European Advanced Network Technology), and other involved parties. The data challenge milestones are 100 MB/sec for 24 hours in summer 2013 (completed successfully), 400 MB/sec for 48 hours in summer 2014, and 1000 MB/sec for 72 hours in summer 2015.

Coordination and strategic planning for data workflows will be needed between the Belle II Collaboration members in Japan, Europe, and the United States.

Cosmology simulations are generating large data volumes, and these are expected to increase by a factor of 10 over the next two to five years. Current site-to-site transfers are in the tens of terabytes, and 100 TB transfers will soon be required. Most of these are currently between the Argonne National Laboratory (ANL) and the National Energy Research Scientific Computing Center (NERSC), with data movement expected to expand to include Brookhaven National Laboratory (BNL), Fermilab, Oak Ridge National Laboratory (ORNL), and SLAC over time. It is expected that these data will be used by multiple observational collaborations (e.g., Dark Energy Survey [DES], Dark Energy Spectroscopic Instrument [DESI], LSST). Exascale systems will have a significant impact here, though it is too early to predict exactly how.

Many large datasets associated with cosmological simulations must be hosted near large-scale computing resources because, in many cases, the analysis of the data does not map well onto the Grid computing model. A few major archive/analysis centers are expected to emerge as central facilities for melding cosmological simulation and data analysis.

Some sites have security policies that significantly hinder the data transfers necessary for data-intensive science. Deployment of the Science DMZ model is often a viable solution for overcoming these issues. We need to ensure that data movement requirements and site security policies are harmonized. Documentation and sharing of architectures and security policies among the DOE facilities would be helpful.

There is a need for regular data transfers between the National Science Foundation (NSF) Extreme Science and Engineering Discovery Environment (XSEDE) computing centers and DOE sites (especially SLAC and NERSC) in support of computational cosmology.

There is a significant reliance on the perfSONAR infrastructure and code base, both by experiments and by the networks that support those experiments. The perfSONAR project does not currently have sustainable programmatic funding. The major perfSONAR stakeholders need to address this issue.

3 Action Items

Several action items for ESnet came out of this review. These include:

- ESnet should host email lists for the coordination of wide area network engineering activities for the Belle II experiment (these lists were deployed between the time of the review and the finalization of this report).
- ESnet should engage with the LHC experiments and their Tier-2 sites to assist with network design to support LHC Run 2.
- ESnet should work with the STAR collaboration to tune data transfer performance between BNL and KISTI.
- ESnet should work with the ACE3D modeling group at SLAC on data transfer performance improvements, e.g., the Science DMZ model.
- ESnet should work with JLab on 10 G network diversity.
- ESnet should work with the DESI collaboration on data transfer performance.
- ESnet must continue to develop and update the fasterdata.es.net site as a resource for the community.
- ESnet should track the progress of the LSST collaboration, and assisting with data movement challenges encountered by the instrument development groups.
- ESnet should continue to assist sites with perfSONAR deployments and continue to assist sites with network and system performance tuning.

In addition, ESnet will continue to develop and deploy the ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS) to support virtual circuit services on ESnet and collaborating networks.

4 Review Background and Structure

Funded by the Office of Advanced Scientific Computing Research (ASCR) Facilities Division, ESnet's mission is to operate and maintain a network dedicated to accelerating science discovery. ESnet's mission covers three areas:

1. Working with the DOE SC-funded science community to identify the networking implications of instruments and 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 the SC mission and to continuously reassess and update the approach as new requirements become clear.
3. Continuing to anticipate future network capabilities to meet new science requirements with an active program of R&D and advanced development.

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

a. A review of major stakeholders' plans and processes, including the data characteristics of scientific instruments and facilities, in order to investigate what data will be generated by instruments and supercomputers coming online over the next 5–10 years. In addition, the future process of science must be examined: How and where will the new data be analyzed and used? How will the process of doing science change over the next 5–10 years?
b. Observing current and historical network traffic patterns to determine how trends in network patterns predict future network needs.

The primary mechanism to accomplish (a) is through SC Network Requirements Reviews, which are organized by ASCR in collaboration with the SC Program Offices. SC conducts two requirements reviews per year, in a cycle that assesses requirements for each of the six program offices every three years.

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

The other role of the requirements reviews is to ensure that ESnet and ASCR have a common understanding of the issues that face ESnet and the solutions that ESnet undertakes.

In August 2013, ESnet organized a review in collaboration with the HEP and NP Program Offices to characterize the networking requirements of science programs funded by HEP and NP.

Participants were asked to codify their requirements in a case study format that included a network-centric narrative describing the science, instruments, and facilities currently used or anticipated for future programs; the network services needed; and how the network is used. Participants considered three timescales in their case studies: the near-term (immediately and up to two years in the future); the medium-term (two to five years in the future); and the long-term (greater 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 Program Perspectives

5.1 High Energy Physics

High energy physics explores the most fundamental questions about the nature of the universe. The DOE HEP Office supports a program focused on three frontiers of scientific discovery. At the *energy frontier*, powerful accelerators investigate the constituents and architecture of the universe. At the *intensity frontier*, astronomically large amounts of particles and highly sensitive detectors offer a second, unique pathway to investigate rare events in nature. At the *cosmic frontier*, natural sources of particles from space reveal the nature of the universe. Together these three interrelated discovery frontiers create a complete picture, advancing DOE missions through the development of key cutting-edge technologies and the training of future generations of scientists.

5.2 Nuclear Physics

Nuclear science began by studying the structure and properties of atomic nuclei as assemblages of protons and neutrons. At first, 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, and 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 NP Office provides most of the federal support for nuclear physics research in the United States. About 1,620 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 these NP facilities for their own research. Notable is the use by semiconductor manufacturers that develop and test radiation-hardened components for Earth satellites to be able to withstand cosmic-ray bombardment and by the National Aeronautic and Space Administration's (NASA's) Space Radiation Laboratory (NSRL) established at BNL's RHIC facility to study the radiobiological effects using beams that simulate the cosmic rays found in space.

The NP program helps the United States 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

program produces highly trained scientists who help to ensure that DOE and the United States have a sustained pipeline of highly skilled and diverse science, technology, engineering, and mathematics (STEM) workers who are knowledgeable in nuclear science.

6 The ATLAS Experiment at the Large Hadron Collider

6.1 Background

The ATLAS experiment at CERN's Large Hadron Collider (LHC) is one of two large, general-purpose LHC experiments to investigate high-energy proton-proton collisions. ATLAS recorded proton-proton interactions at center-of-mass energies of 7 and 8 TeV in 2011 and 2012, respectively. The studies of these interactions have led to more than 250 scientific publications that reported measurements of particle properties and searches for new particles. The chief ATLAS science highlight so far has been the discovery of a Higgs boson with a mass of 125 GeV. In fact, this result was named the "Science Breakthrough of the Year 2012" by Science magazine.

The discovery of the Higgs boson opens a new area of scientific study for the ATLAS experiment. Studies are under way to measure the Higgs boson's properties, including its mass, spin, parity, and coupling to other particles. The Higgs boson solves only part of the mystery of electroweak symmetry breaking in the Standard Model of particle physics — other particles are needed to complete the description. The boson observed by ATLAS may indeed be the Standard Model Higgs boson, or it may be one of several types of Higgs bosons. Therefore, the discovery of the Higgs opens new possibilities beyond our current knowledge and understanding of physics. One theoretical outcome of this discover includes supersymmetry theory, which is a strong candidate to explain some fundamental unanswered questions associated with the Higgs boson: How does the Higgs boson get the observed mass, and what is the mechanism behind electroweak symmetry breaking? In addition, supersymmetry provides a candidate for the dark matter that we know makes up most of the matter in the universe, yet has not been observed. The search for dark matter started many decades before the search for the Higgs. However, the discovery of the Higgs at the LHC could also be the first in a series of breakthroughs that herald a renaissance for particle physics.

The chief science challenge in these studies is the production rate of new particles relative to the rate of proton-proton collisions. For example, only one in a billion inelastic proton-proton collisions at 14 TeV is expected to produce a Higgs boson. Ensuring a statistically significant sample of Higgs bosons requires an enormous dataset of collected proton-proton collision events, which in turn implies a high-luminosity collider. The rate of dark matter production at the LHC could be much smaller, requiring much more difficult data mining.

The LHC accelerator and the ATLAS experiment are currently being prepared for collisions at 13–14 TeV, scheduled to begin in late 2015. The LHC will then operate at the original design luminosity of 1×10^{34} $cm^{-2}s^{-1}$, until a one-year shutdown for further luminosity upgrades in 2018. This expected schedule sets the timeline for the science drivers and needs described below. A proposal to increase the LHC luminosity to 5×10^{34} $cm^{-2}s^{-1}$, beginning in 2023, is currently under consideration.

Since the time of the last HEP ESnet requirements review in 2009, the ATLAS experiment has revamped its computing model to put less emphasis on a strict hierarchical model

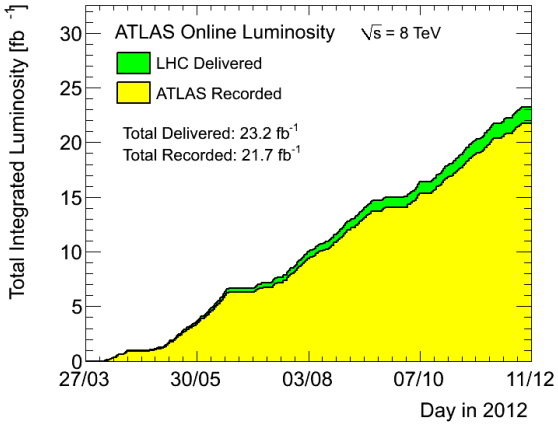


Figure 1. ATLAS Online Luminosity in 2012.

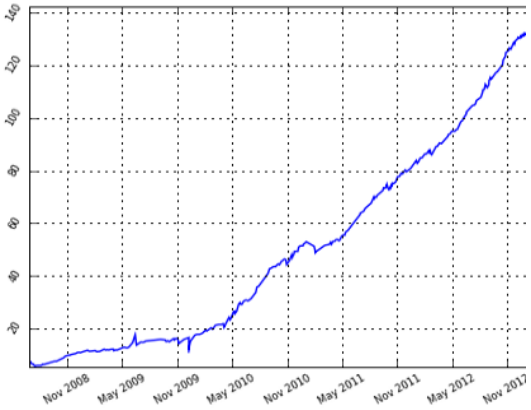


Figure 2. ATLAS data volume on the Grid.

among the computing sites, and more emphasis on a peer-to-peer mesh model. This re-evaluation has been driven by the needs of physics analyzers and the emergence of high-performance research networks. Nevertheless, we maintain the distinctions between Tiers-1/2/3 in our case study to indicate the physical scale of each facility as well as its support of middleware services and its role in physics analysis.

Once the data is acquired and reconstructed, the nature of the LHC physics environment makes its analysis a complex challenge as well. New discoveries in physics expected at the LHC, such as the Higgs boson and supersymmetry, are predicted to occur at very low rates. A typical light Higgs signal, for example, may involve on the order of 1,000 signal events distilled from 100 trillion events occurring in the detector in a year of data taking. The ATLAS trigger provides a rejection factor of 10^5, but the further selection of one event in a million must be performed in offline processing. This presents one of the central computing infrastructure challenges in LHC computing: *the rapid and efficient extraction of sparse physics samples from extremely large datasets*.

Experience gained during the first three years of ATLAS data taking gives us confidence that the distributed computing model developed by ATLAS has sufficient flexibility to process, reprocess, distill, disseminate, and analyze data in a way that utilizes both computing and manpower resources efficiently. New advances in computing, such as cloud computing, can be integrated easily into the current model.

The data management, processing, and analysis tasks required by ATLAS must be conducted in the context of a very large, world wide collaboration. For its distributed analysis system alone, ATLAS has more than 2,000 users. The ATLAS computing system must anticipate more than 1,000 simultaneous users distributed globally who need transparent access to all resources available.

The unique challenges of LHC computing led us to the tiered hierarchy of centers, networked in a worldwide data-intensive grid. The early success of this system has been tremendous. The ATLAS distributed computing systems truly act as an enabler for timely and effective analysis. The LHC experience to date has shown that Tier-2 centers play an

unexpectedly vital role for analysis and ATLAS overall, and that the current and still-evolving future of LHC computing planning is toward a “flatter” architecture that discriminates less between Tier-1 centers and Tier-2 centers in how they interconnect in the distributed facility. This also means, and is agreed upon by ATLAS management, that a cost/benefit analysis is the basis for deciding whether compute and storage capacities are deployed at Tier-1 versus Tier-2 centers. However, this arrangement can only work if the Tier-2 centers have, in addition to compute and storage capacities, the necessary network infrastructure to exploit the resources. This larger Tier-2 role requires the centers to provide high-performance, very reliable network capabilities at a level unforeseen when the hierarchical model was originally designed. In addition to the original role, Tier-2 centers have increasingly become a repository for primary data that is supposed to be served to sites domestically and internationally either through programmatic replication or eventually by letting remote processes access the data directly through the federated storage service (FAX).

Whether hierarchical or flattened, the distributed computing model depends on linking all computational and storage resources within a center as well as all sites through high-speed local and wide area networks (WANs) into a highly functional distributed fabric. This fabric must distribute and manage data and workloads among its massive resources to present a tractable operational load.

Given the modified and added responsibilities and the massive critical resources Tier-2 centers contribute to ATLAS’s centrally managed production and user analysis, it is imperative to deploy and maintain a reliable communication infrastructure. This infrastructure must match the performance requirements of applications running on the computational resources at those centers as well as support programmatic replication of ATLAS data and remote data access.

6.2 The ATLAS Collaboration

The ATLAS collaboration (atlas.ch) consists of 174 institutes from 38 countries. After construction was completed on the ATLAS detector at the LHC, the first collisions were recorded in late 2009. The 44 U.S. ATLAS institutions made major and unique contributions to the construction of the ATLAS detector; provided critical support for the collaboration’s computing and software program and detector operations; and contributed significantly to physics analysis, results, and papers published. The physics results from such a large collaboration rely on efficient networking for transparent access to data and processing across all computing sites, irrespective of the hierarchical or mesh configuration of the Tier-1/2/3 sites.

6.3 Key Local Science Drivers

6.3.1 ATLAS Computing Facilities in the United States

To keep the focus on enabling science, local instruments and facilities are defined as those used directly by physicists analyzing data from the LHC. These are typically large, local computing clusters at universities and laboratories.

While the ATLAS Tier-1 center in the United States runs up to 13,000 concurrent jobs, the five distributed Tier-2 centers currently provide from 4,000 to 7,000 job slots each. The workload is a mix of production and analysis. The majority of the centers give priority to analysis jobs over production jobs such that, to use CPU resources efficiently, network infrastructure must be able to accommodate the bandwidth needs of jobs running on up to 66% of total CPU resources. While the number of user analysis jobs submitted to a Tier-1/2 facility varies widely, the software system keeps all available CPUs busy with centrally submitted jobs. In addition, local computing is done at the Tier-3 sites for the end stages of user analysis.

Facts related to ATLAS analysis jobs:

1. Most, if not all, CPU cycles are spent unzipping ROOT branches and creating unzipped structures in memory.
2. ROOT (gzip) can unzip at a rate of up to 40 MB/sec on a single core. In case of ATLAS D3PDs, the rate is lower, actually 20 MB/sec due to a huge compression ratio and inefficient file structures.

Assuming sparse reading, network latencies, etc., bandwidth requirements are observed between 3 MB/sec and 15 MB/sec per job. That translates to being able to accommodate anything between 80 and 420 jobs on a 10 Gbps network link in the path between CPU and disk storage. This path could be local to a site where both CPU and disk storage are installed, or between the CPU at a site (e.g., University of Illinois at Urbana–Champaign [UIUC]) and storage at a different site (e.g., the University of Chicago). Even at a moderate average rate of 6 MB/sec per job, a 10 Gbps link would be saturated by only 200 jobs, meaning a site should have the ability to serve data at a rate of at least 100 Gbps. At this point, none of the Tier-2 centers are close to this available bandwidth between their compute and storage servers (note that at the Tier-1 site, we have 160 Gbps and we observe full utilization of the path on a regular basis when running about 4,000 analysis jobs). Only large aggregation switches, as recently proposed, and/or significantly higher bandwidth between switches local to sites and across the WAN are necessary to efficiently use current resources and will be even more so in the future.

In the United States, BNL provides Tier-1 computing for ATLAS. The facility is large in absolute size, and in relative size when compared with other Tier-1 computing centers for the LHC. The Tier-1 center is connected to CERN and receives data from the ATLAS Tier-0 facility via an optical private network (OPN). The expectations of the OPN are described in the case study on LHC trans-Atlantic networking.

Currently BNL has more than 10 PB of disk in production and 90 kHS02 of processing. (Processing resources in the LHC are measured in thousands of HEPSpec 2006 [kHS06], which is based on SpecInt 2006.) Data is archived on magnetic tape stored in and managed by automated libraries. The ATLAS inventory of archived data is currently about 8 PB. The BNL Tier-1 facility utilizes dCache to virtualize the large number of physical devices into a storage system.

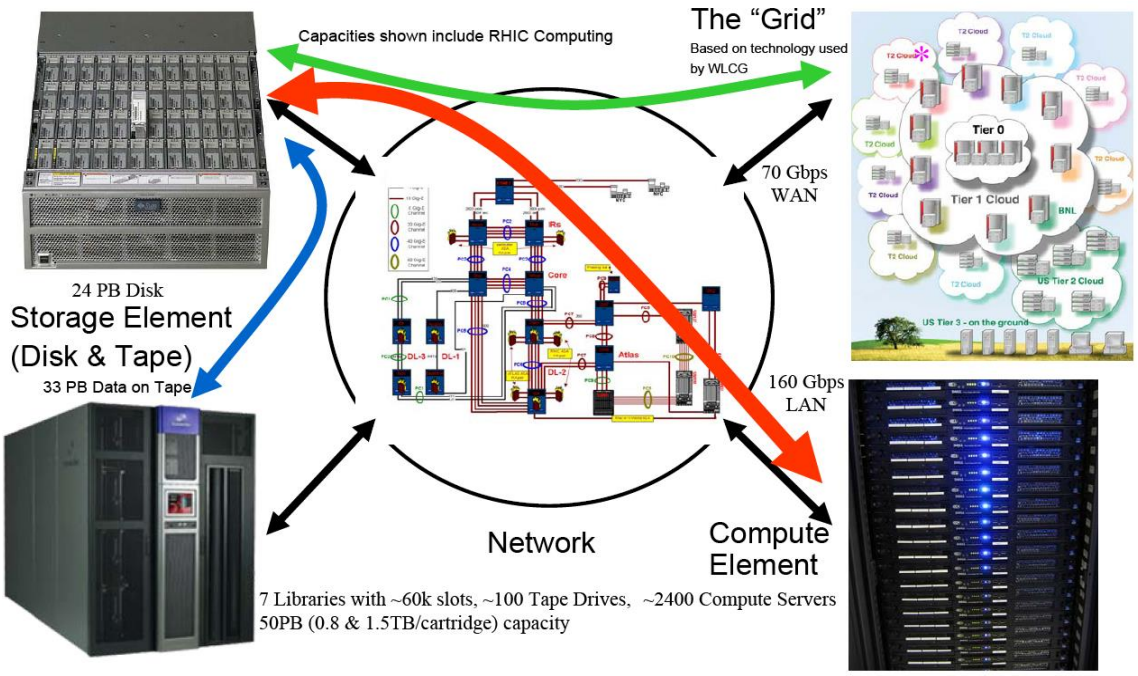


Figure 3. Tier-1 resources and architecture.

Besides the Tier-1 center, there are five Tier-2 computing facilities for ATLAS. Four of the Tier-2 centers are distributed facilities with hardware and operations support located at two or three campuses.

In 2013, a nominal Tier-2 for ATLAS is about 35 kHEPSpec06 of processing, which is roughly 3,000 processor cores, and 3500 TB of usable disk. The available storage space is spread over many physical storage devices and several technologies are used to make them a coherent storage system. In the United States, dCache and XRootD are currently in production.

Compute Nodes

The compute nodes in U.S. ATLAS facilities are multicore, stateless compute servers with a relatively simple data flow profile. To first order, data is read from and written to the storage servers, with reads dominating writes by a factor of 10 or higher. A superficial analysis of interswitch link utilization suggests that 40 Gbps of network bandwidth is required for every 240 compute nodes. However, this assumes that current bandwidth utilization is not limited by storage server bandwidth.

High-bandwidth Data Servers at Tier-1

The high bandwidth data servers at the U.S. ATLAS Tier-1 facility are the dCache storage nodes. These storage nodes are 10 GbE or 20 GbE attached servers that are able to drive their network connections at full line rate. There are currently about 80 storage servers in four flavors. A superficial analysis of interswitch link utilization between compute nodes and storage nodes suggests the following:

- 40 Gbps of network bandwidth is required for every twenty 10 GbE attached storage nodes.

Data flows are observed between disk-storage nodes and the following facility resources:

- Compute nodes (overwhelmingly read-only)
- Data transfer servers
- Custodial storage (high-performance storage system [HPSS])

Read-access to data from compute nodes and data transfer servers runs the gamut, from single “hot” file to uniformly “hot” data storage servers. Write-access from the compute nodes and data transfer servers is assumed to be uniformly spread across storage nodes. Read-and-write access to data from custodial storage (HPSS) to the data-storage nodes is assumed to be uniformly spread over storage nodes.

A rough estimate of the required 10 GbE port density is as follows. Assuming 120 disk drives per 10 GbE connectivity, at 9720 disk drives, the number of required 10 GbE host ports is on the order of 65. In the data-flow section of this document, a minimum of 40 Gbps of network connectivity per twenty 10 GbE attached server is the current rule of thumb. This yields a maximum 5:1 oversubscription ratio between aggregate server connectivity and uplink bandwidth. For sixty-five 10 GbE host ports, a minimum of fourteen 10 GbE uplink ports is required. This translates to a need for at least seventy-nine 10 GbE ports. Even more are required when crosslinks to the disk storage switch in another room and multiple 10 GbE connectivity to HPSS are factored in.

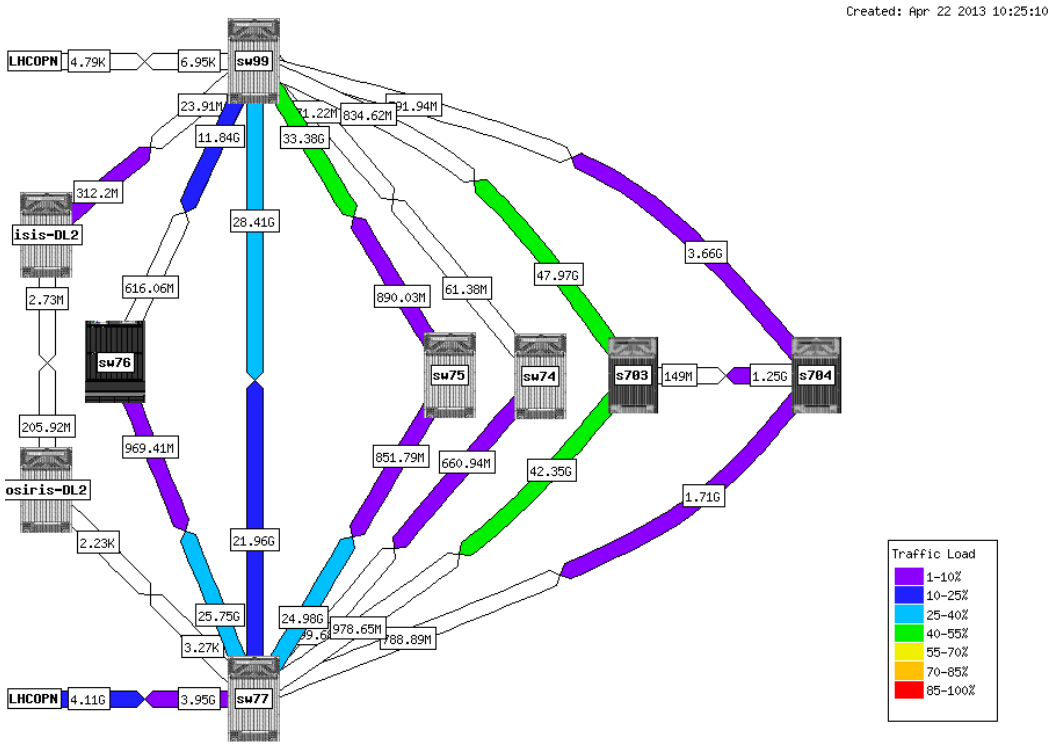


Figure 4. Current U.S. ATLAS Tier-1 network configuration.

In conclusion, based on the view of the set up, it can be assumed that there is no control over data placement, thus eliminating the possibility of partitioning the facility into smaller clusters with minimal data movement between clusters.

This assumption has drastic effects on the design of the network.

Data Transfer Servers at the Tier-1 center
Data transfers are currently performed by dedicated servers. These nodes are used to send/receive data over the LHCOPN, LHC open network environment (LHCONE), and the Great Plains Network (GPN) to/from the dCache storage nodes. These data transfer servers are dual-attached to the network, with one physical connection to the LHCOPN and the other physical connection on the same switch as the dCache storage nodes.

Data Flows in the Tier-1 Network
Large data flows in the Tier-1 network are split into two categories: internal and external. Within the U.S. ATLAS network, the predominant data flow is between the dCache storage subnet and the linux farm subnets. Outside of the Tier-1 network, the predominant data flow is between the dCache transfer server and the WANs. Note that these external data flows from the transfer servers are exactly mirrored by data flows between those and the dCache storage servers. For various reasons, the transfer servers are dual-homed (i.e., have two network interface cards).

A superficial analysis of the interswitch link utilization suggests the following :

1. 40 Gbps of bandwidth is required for every 250 compute nodes.
2. 40 Gbps of bandwidth is required for every twenty 10 GbE attached high-density storage nodes.

However, it should be noted that these ratios may not be independent. It is possible that increasing the number of storage nodes will increase the bandwidth used per 250 compute nodes. It is also possible that increasing the number of compute nodes will increase the bandwidth used per 20 high-density storage nodes. Data flows between dCache and the linux farm have been tuned to utilize both available paths between core switches in the network.

“Network-aware” Applications
A new paradigm will likely change the way applications interact with the underlying network infrastructure. Rather than looking at an opaque piece of infrastructure, as applications do today, they will soon be able to take control of the topology and quality of service parameters — capabilities that make the performance of the communication path in our highly distributed applications highly predictable. Software-defined networking (SDN) is the foundation for the interaction of “network-aware” applications with the infrastructure. Therefore, the Tier-1 and Tier-2 sites are required and are acquiring new equipment to support the up-and-coming SDN technology.

6.3.2 Software Infrastructure

While the bulk of data processing in ATLAS is done at Tier-1 and Tier-2 resources, the end-stage analysis is usually done by users at a local Tier-3 facility. The scale of

computing resources at Tier-3 sites ranges from workstations to small clusters. ROOT is the most common software stack used to analyze the Derived Physics Data (DPD) generated on distributed computing resources. Data transfer is done using ATLAS distributed data management (DDM) tools, which mostly rely on GridFTP middleware. XRootD-based direct data access is also gaining importance wherever high network bandwidth is available. For a small number of users (primarily for detector development and calibration studies), ATHENA software (the ATLAS application framework) needs to be supported in the local environment.

6.3.3 Process of Science

Scientific discovery at the LHC is a massive data-mining problem. The design rate of collisions at the experiment is 40 MHz, with a data collection rate of a few hundred hertz. Only a tiny fraction of the events contains interesting physics and a smaller fraction contains evidence of new discoveries. The experiment's data acquisition systems must preferentially select the one event in a hundred thousand events we can afford to keep. The problem of event selection continues with Tier-1 centers, which are responsible for updating the data samples by reprocessing with improved calibration, and for creating analysis samples for users. In the new mesh model, Tier-2 centers also participate in reprocessing. The events are skimmed in an attempt to make smaller samples focusing on particular physics processes and thinned to concentrate objects relevant to a particular analysis. The ATLAS experiment collects a few billion events per year and, except in the most fortuitous cases, a new discovery is or will be based on a few hundred, or less, very carefully selected events. With future higher luminosities planned, data processing becomes much more challenging due to pile-up effects (multiple collisions in the recorded event), putting more demands on the network due to the flatter distributed computing model involving Tier-2 sites.

With the LHC in its Long Shutdown 1, the ATLAS collaboration is currently focused on analyzing the results of the 7 and 8 TeV data taken between 2011 and 2012. The most popular data format for physics analysis is a DPD format that is essentially a flat ROOT ntuple. These DPD ntuples can be read efficiently with highly optimized tools and can be used in distributed analysis jobs, in local batch queues, or in a PROOF farm. Even though the event size in the DPDs is reduced to about half of the 250 KB/event size in the analysis object data (AOD) files, there are multiple distinct versions of the DPD. A recent estimate found a total of 3.2 PB of existing DPD datasets in ATLAS, including collision data and Monte Carlo (MC) data. A recent initiative to develop a common DPD ntuple format promises to reduce this data volume to 1 PB by eliminating duplication in the DPD definitions.

Data analysis with the local Tier-3 computing systems depends at the moment on specific locally accessible datasets. A particular case study is the analysis of Higgs decays to W boson pairs. In addition to the collision data sample, the analysis processes a total of about 530 MC samples. These samples are skimmed with suitable event filters at Tier-1/2 sites, and approximately 4 TB of input files are downloaded to a local compute farm using ATLAS data management tools, with typical transfer rates of 50 MB/sec. This local

farm is used to produce DPD ntuples in a process that takes about 200–300 CPU hours, and the resulting ntuples are transferred to CERN. Additional ntuples used to study systematic uncertainties in the measurements multiply the requirements by a factor of 50, resulting in a CPU requirement of 10,000 hours and a total ntuple production output of 2.1 TB. This is a fraction of the locally accessible disk at the Tier-3 sites, where the median disk resource is 100 TB; nevertheless, data transfers from the remote sites take a significant amount of time in the data analysis cycle. These datasets are reproduced between 1 and 10 times per month, depending on the studies being performed.

Many analyses do not require local ntuple production, but rather use DPD ntuples produced on the Grid, either centrally or with user analysis jobs. This model offers less flexibility for local analyzers, but it requires fewer local CPU resources. The typical dataset used for analysis is approximately 1 TB and is refreshed once per month with newly produced ntuples. The goal of the Tier-3 systems, as developed and supported by ATLAS, is to allow physicists to analyze their entire reduced dataset in one to two hours. Each local computing installation supports from one to five physics analyses with accompanying datasets. The analyzers at some active Tier-3 clusters submit 2 million Condor jobs each year.

6.4 Key Remote Science Drivers

6.4.1 Instruments and Facilities

Because of the data-intensive nature of the ATLAS scientific program, the ATLAS collaboration implicitly relies on a ubiquitous, high-performing, global network to enable its distributed Grid-computing infrastructure. Providing effective access to petabytes of data for thousands of physicists all over the world would be impossible without the corresponding set of research and education networks that provide 1, 10, and/or 100 Gbps of bandwidth to enable ATLAS data to flow where it is needed. Typical network paths that ATLAS data traverses consist of multiple administrative domains (local area networks at each end and possibly multiple campus, regional, national, and international networks along the path). The Internet's ability to allow these separate domains to transparently interoperate is one of its greatest strengths. However, when a problem involving the network arises, that same transparency can make it very difficult to find the cause and location of the problem. Because of both the necessity of the network for normal ATLAS operations and the difficulty in identifying and locating the source of network problems when they occured, the U.S. ATLAS facilities began an intense collaboration with ESnet and Internet2 to develop and deploy perfSONAR-PS in 2008. The goal was to provide the sites with a set of tools and measurements that would allow them to differentiate network issues from end-site issues and to help localize and identify network-specific problems to expedite their resolution.

As U.S. ATLAS began to deploy perfSONAR-PS monitoring instances, the federated design architecture was found to have shortcomings for the intended use-case. Each site was independently installed, configured, and controlled and it was difficult to see the status of the sites or the intersite measurements without visiting each site and viewing multiple

graphs. To provide a high-level summary of the site test status and to visualize the results of the perfSONAR measurements, a dedicated monitoring system was proposed and developed. The architecture of the system consists of multiple functional components. The first is a set of collectors that gather monitoring information. The results from the collectors are stored in a data store component. Information from the data store is presented to users via a data presentation Web interface. Finally, new monitoring jobs and alerting are defined and configured through a Web interface. Because of the globally distributed nature of the computing facility on which the ATLAS data analysis depends, the perfSONAR tools and services were adopted by the OSG in the United States and the Worldwide LHC Computing Grid (WLCG) to help manage performance at approximately 160 sites worldwide. As dependencies on network performance data increase (ATLAS is in the process of integrating network performance into task brokerage and data placement decisions), it is tremendously important for the collaboration that ESnet contributes to the perfSONAR development and maintenance at least at the present level.

ATLAS is adding direct access to data not available locally. Rather than requiring a process to wait until a programmatic replication of a dataset is completed, the process uses a mechanism that allows transparent discovery of the needed data and access to it over the WAN. U.S. ATLAS is currently using XrootD at both the individual site level and the U.S. ATLAS computing facility level. Tier-2 sites at SLAC and the University of Texas at Arlington use XrootD as their baseline storage system, while the Tier-1 site at BNL, University of Chicago, and University of Michigan are using XrootD as an interface system on top of their dCache-managed storage to serve user analysis activities. The sites each have between 3 PB and 10 PB of usable disk storage installed and serve heavy user analysis activities.

ATLAS recently deployed a FAX aimed at providing direct data access over the WAN. The system allows users to access any data file in the federation via its global unique file name using the XrootD protocol. FAX is implemented via a global XrootD redirector at BNL and some regional redirectors deployed at Tier-2 sites. In addition to Tier-1 and Tier-2 sites, Tier-3 sites are important members of this federation because quite often, "hot" user analysis data are initially produced at Tier-3 institutions.

ATLAS intends to implement and evaluate an even more fine-grained approach to caching below the file level. The approach takes advantage of a ROOT-based caching mechanism as well as recent efficiency gains in ROOT I/O implemented by the ROOT team that minimizes the number of transactions with storage during data-read operations, which, particularly over the WAN, are very expensive in terms of latency. It also utilizes development work performed by CERN-IT on a custom XrootD server that operates on the client side to direct ROOT I/O requests to remote XrootD storage, transparently caching at the block level data retrieved over the WAN and passed on to the application. Subsequent local use of the data hits the cache rather than the WAN. This benefits not only the latency seen by a client utilizing cached data, but also the source site, freed from the need to serve already delivered data. In addition, caching obviously saves network capacities.

Figure 5. Federated data stores and data access with FAX in ATLAS.

Deriving benefit from fine-grained caching depends upon reuse of the cache. As one approach to maximizing reuse, PanDA's (Production and Distributed Analysis; the ATLAS workload-management system) existing mechanism for brokering jobs to worker nodes on the basis of data affinity will be applied to this case, such that jobs are preferentially brokered to sites that have run jobs utilizing the same input files.

Non-PanDA-based applications using data at the cache site will also automatically benefit from the cache. The approach will integrate well with the federated XrootD system; it adds an automatic local caching capability to the federation. It may also be of interest in the context of serving data to applications running in commercial clouds, where the expense of data import and in-cloud storage could make fine-grained caching efficiencies valuable.

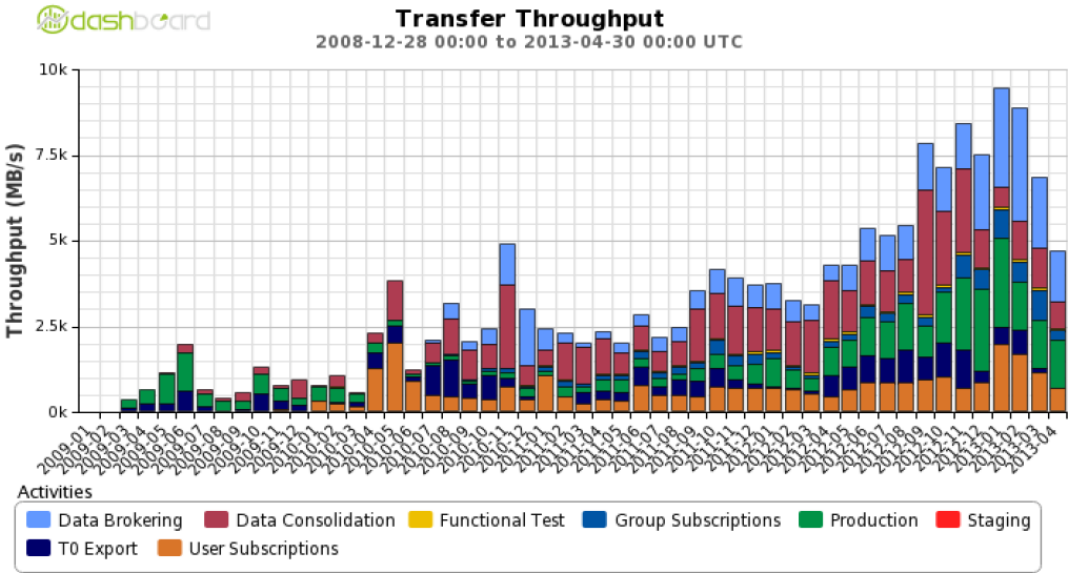


Figure 6. ATLAS worldwide data transfer activities by data category (2009–2013).

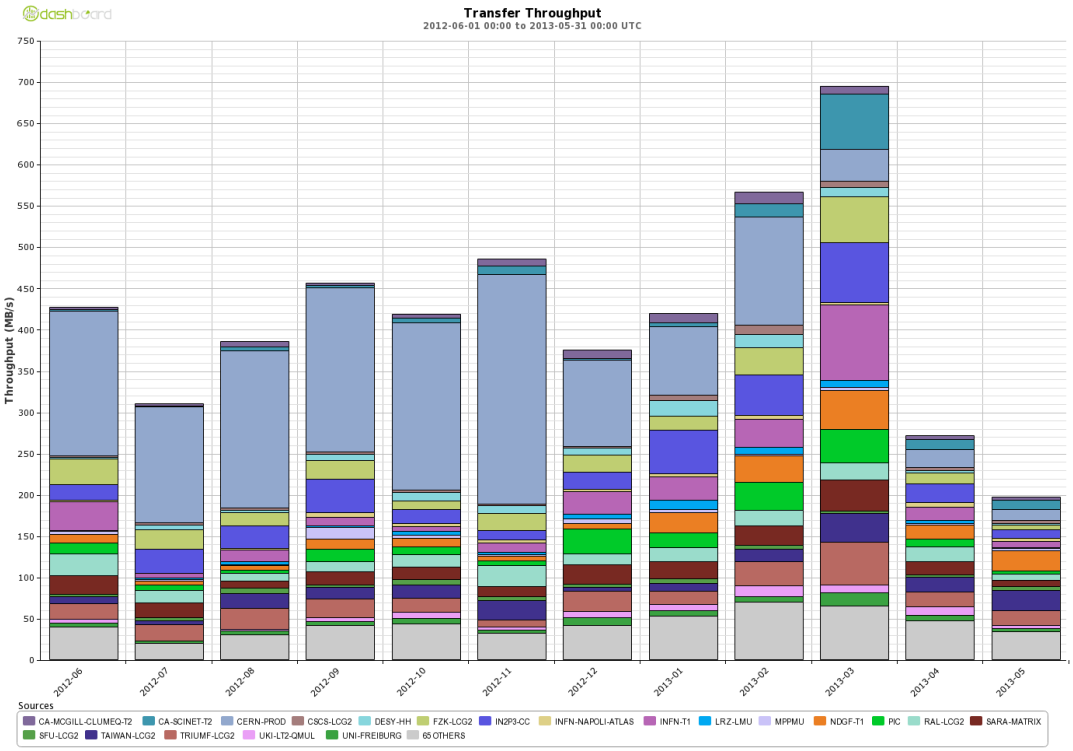


Figure 7. Traffic from non-U.S. sites to the U.S. Tier-1 center at BNL from 6/2012–5/2013.

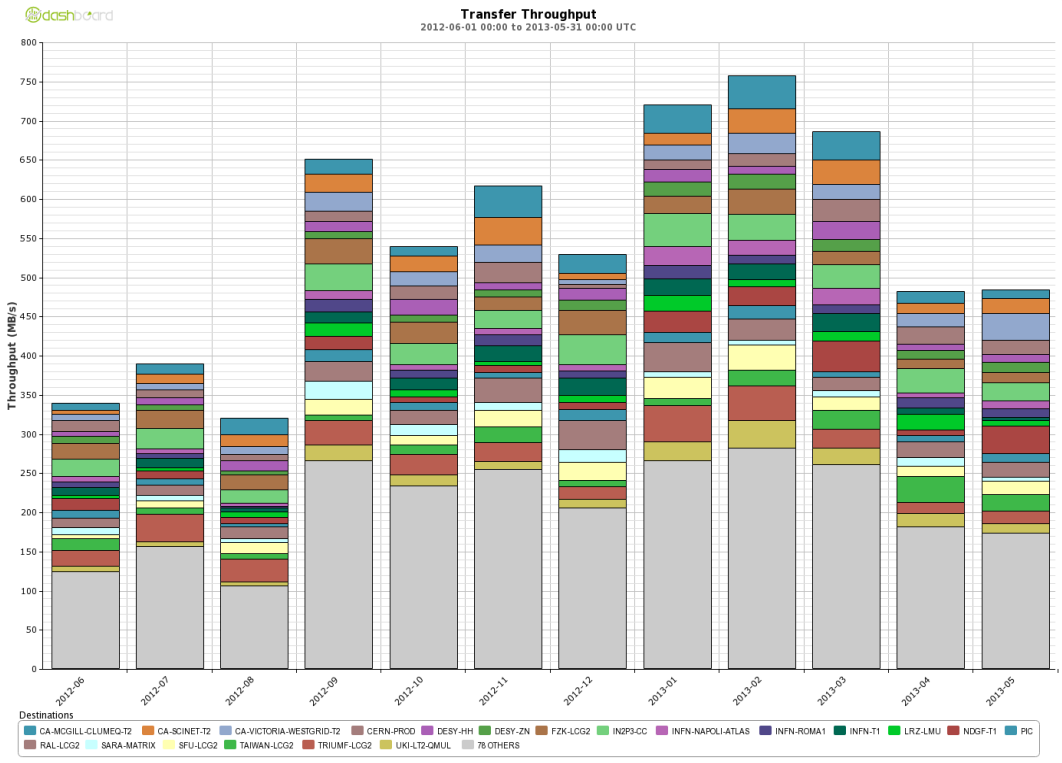


Figure 8. Traffic from the U.S. Tier-1 center to non-U.S. sites from 6/2012–5/2013.

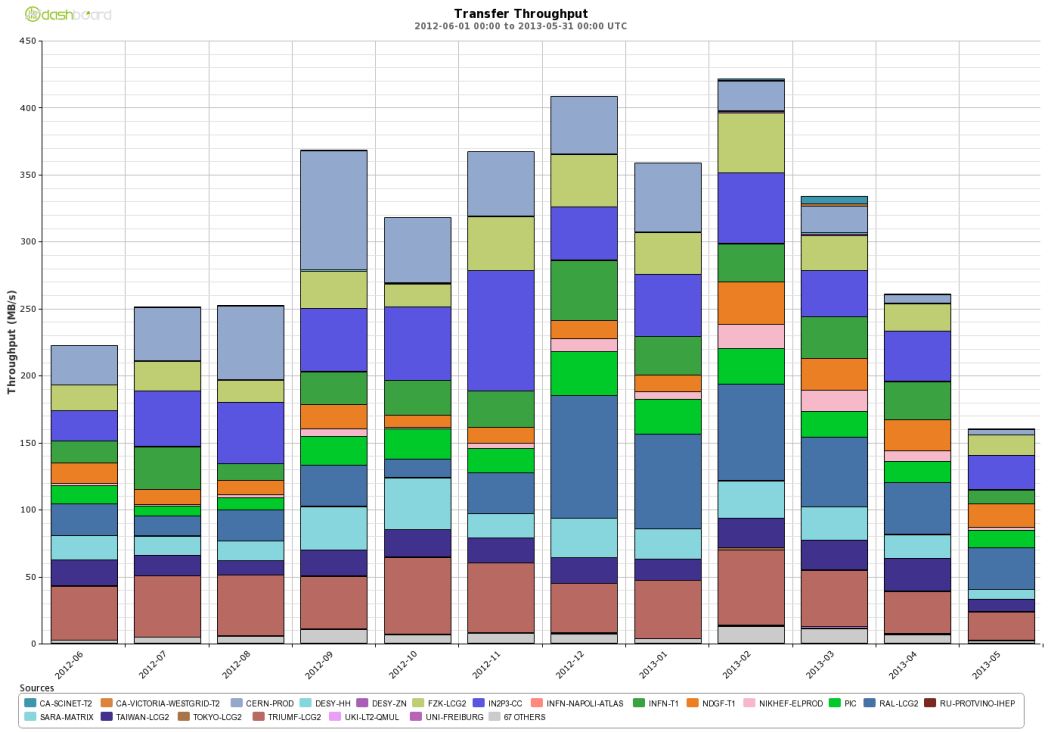


Figure 9. Traffic from Non-U.S. sites to U.S. Tier-2 centers from 6/2012–5/2013.

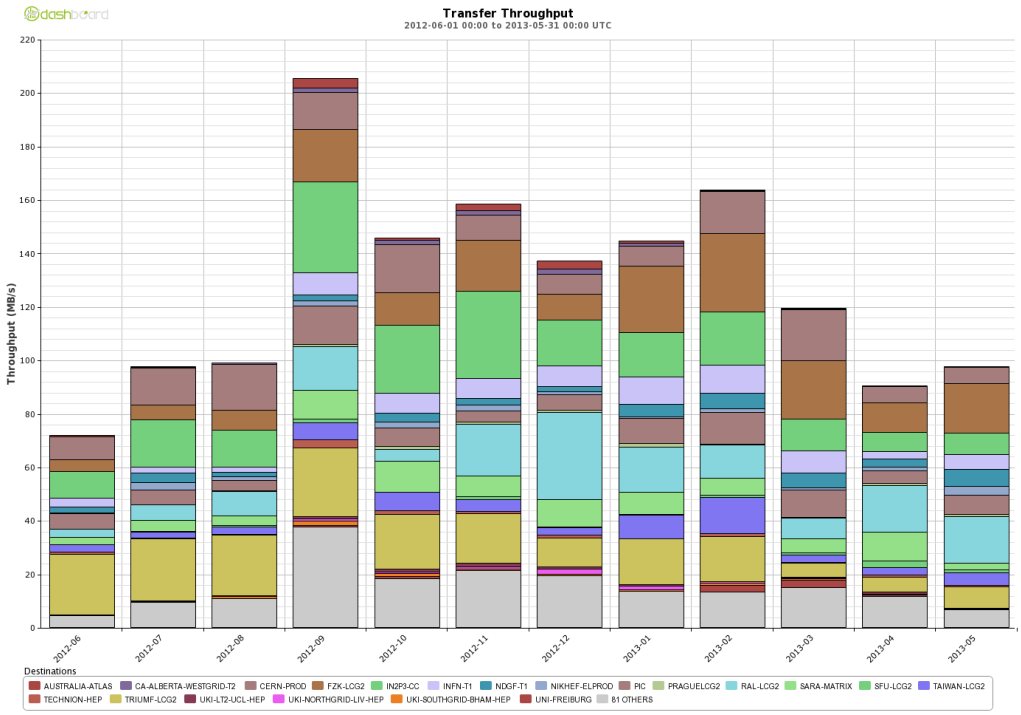


Figure 10. Traffic from U.S. Tier-2 centers to non-U.S. sites from 6/2012–5/2013.

6.4.2 Software Infrastructure

The PanDA workload management system (WMS) is used to manage the distributed workflow in ATLAS. The following section provides a general introduction to PanDA, followed by specific examples relevant to this case study.

The LHC's computational challenge is not limited to the unprecedented size of the data generated. LHC data is highly distributed and accessed by large number of users. A sophisticated WMS is needed to manage the distribution and processing of such data. One of the most successful WMSs developed in the United States for the ATLAS experiment is PanDA. PanDA is also actively being considered for wider use among other big data sciences. The AMS (Alpha Magnetic Spectrometer) is a satellite-based experiment that has dedicated three programmers to adapting PanDA for their use. The CMS and ALICE (A Large Ion Collider Experiment) experiments at the LHC are evaluating PanDA for their distributed analysis system. PanDA is becoming the enabling technology for many scientific discoveries that require access to exascale data.

PanDA delivers transparency of data and processing in a distributed computing environment to ATLAS physicists. It provides execution environments for a wide range of experimental applications, automates centralized data production and processing, enables analysis activity of physics groups, supports custom workflow of individual physicists, provides a unified view of distributed worldwide resources, presents status and history of workflow through a integrated monitoring system, archives and curates all workflow, manages distribution of data as needed for processing or physicist access, and provides other features. The rich menu of options provided, coupled with support for

heterogeneous computing environments, make PanDA ideally suited for data-intensive sciences.

PanDA has a highly scalable and flexible architecture. Scalability has been demonstrated in ATLAS through the rapid increase in usage over the past three years, and is expected to easily meet the expected growth needs over the next decade. PanDA was designed to flexibly adapt to emerging computing technologies in processing, storage, and networking, as well as the underlying software stack (middleware). This flexibility has also been successfully demonstrated over the past five years of evolving technologies adapted by computing centers in ATLAS, which span many continents yet are seamlessly integrated into PanDA.

The PanDA project began in 2005 as part of the U.S. ATLAS program and was managed jointly by Prof. K. De from the University of Texas at Arlington (UTA) and Dr. T. Wenaus from BNL. At the time, a variety of WMS were deployed in ATLAS, based on deployed grid systems, separately for different applications. There were also separate systems for physicists and central production. PanDA emerged as the best system and was adopted as the default and single WMS for ATLAS before the LHC started operating in 2009. PanDA continues to be primarily supported by DOE and NSF, and is managed by the original team of Wenaus and De, while enjoying expanded contributions from many countries in ATLAS. Today, PanDA has grown to support all distributed workflows in ATLAS, and enjoys a huge user- and support-base worldwide.

Through PanDA, ATLAS physicists see a single computing facility that is used to run all data processing for the experiment, even though the data centers are physically

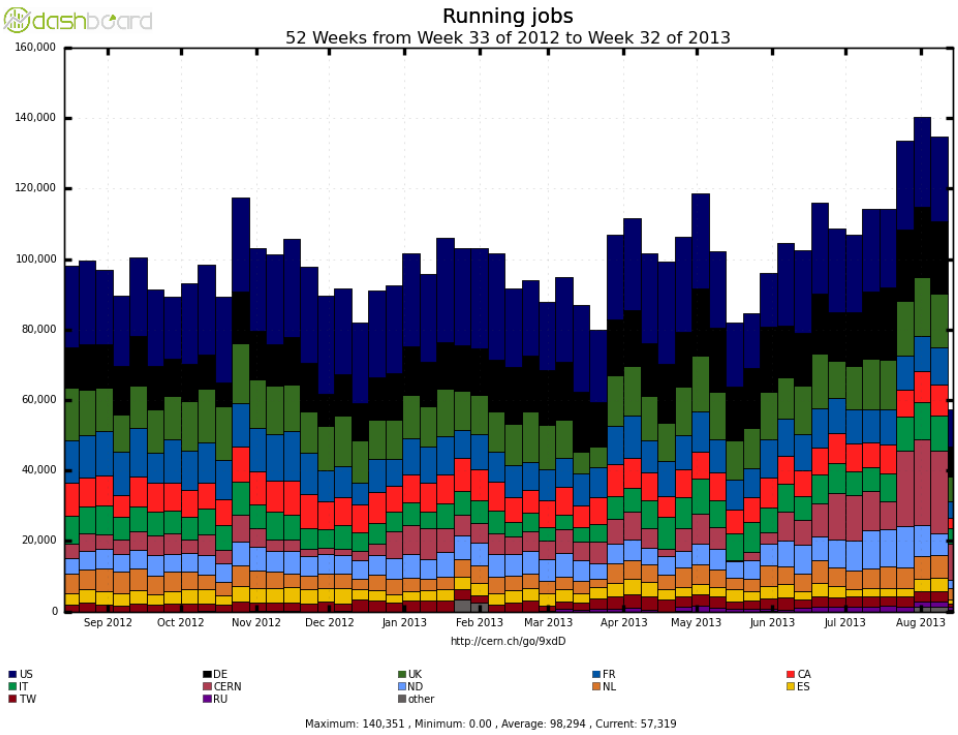


Figure 11. Average number of concurrently running production jobs.

scattered all over the world. Central computing tasks (such as MC simulations, processing and reprocessing of LHC data, reprocessing of MC simulations, mixing and merging of data, and other tasks) are automatically scheduled and executed. Group production tasks, carried out by groups of physicists, are also processed by PanDA. User analysis tasks, which often lead to scientific publications, are seamlessly managed. PanDA is deployed at all ATLAS Tier-1 and Tier-2 centers. Figure 11 shows the number of concurrently running jobs managed by PanDA over the 10 regional groups of tiered centers during the past year.

Each PanDA site provides a Grid-accessible Compute Element (CE) and a Storage Element (SE). Pilot jobs are continuously and automatically scheduled at the CE of each site. When the pilot jobs start execution, they contact PanDA Apache servers, which then dispatch the execution workload. PanDA maintains a central database of all activities, and consequently a central queue of all workflows. This architecture provides an integrated view of all resources managed by PanDA. The pilot-based system also enables integration of non-Grid-based resources. Local resources at universities are integrated using local pilot submission factories. New cloud-based resources are also added to PanDA using the CE model.

The SE plays a central role in the PanDA workflow. For central computing tasks, input data is asynchronously staged in and output data is staged out to the SE. The Tier-1 hierarchy is maintained for all workflow of these types of tasks. However, the user analysis workflow is different. Processing always goes to the location of the data. Both workflows are automatically managed by PanDA. The ATLAS data management system DQ2 is used by the PanDA system for all data registration, data discovery, and data movement. PanDA supports a large variety of SEs across hundreds of computing sites. Recent work has focused on supporting FAX, specifically through XrootD.

Recent developments in PanDA introduce a revolutionary step in WMS design: the concept of a Network Element (NE). Networking services provide an essential infrastructure for all distributed WMS, but they are seldom integrated into the workflow. PanDA is undergoing an evolution as network services are completely integrated into it. The NE will become as ubiquitous as the CE or SE in the PanDA design.

PanDA's capability for large-scale data-intensive distributed processing has been thoroughly demonstrated in one of the most demanding computing environments in large-scale science. PanDA processes a diverse range of workloads — more than 200 million jobs/year on over 100,000 job slots worldwide. Thousands of physicists use it for their personal processing needs. At current scale, PanDA is managing about 1 million jobs daily in ATLAS.

6.4.3 Process of Science

Sites and their complementary roles are essential to the ATLAS analysis model. The process of science at remote locations has a variety of forms. At the remote Tier-1 centers, the synchronized reconstructed data and more summarized analysis formats are served to local Tier-2 sites in the same way they are served to local Tier-2 sites from the

U.S. Tier-1 sites. Data location is managed transparently through automated systems like PanDA and the ATLAS DDM system to provide seamless access to hundreds of PB of data to ATLAS physicists.

The scientific process primarily resides at the remote Tier-2 centers, which are the bulk of the analysis resources for ATLAS. However, with the new mesh configuration of sites, Tier-1 and Tier-2 sites are almost equivalent in their usage by the thousands of ATLAS physicists. Smaller event samples are processed remotely by physicists at Tier-1/2 sites 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.

The 40 ATLAS Tier-3 computing systems in the United States are designed to provide computing resources each for roughly 10–20 physicists at a single institution. As such, they have modest storage and CPU resources. Most of them do not have the middleware required to participate as Grid computing sites, which is in contrast with the Tier-1 and Tier-2 sites. The remote Tier-1 and Tier-2 sites provide Grid-enabled analysis queues, data-storage elements, and MC production capability.

Physicists at home institutions transfer the stored output of MC production (or the converted DPD ntuples) to their local computing clusters for analysis. They may also submit event-skimming jobs to Grid queues at the remote facilities and transfer the output of those analysis jobs to their local cluster for analysis. Strategic data placement makes it possible to fully utilize the U.S. resources for analysis, and large shared storage elements, as part of the ATLAS DDM scheme, make it possible for physicists to collaborate in analyzing a specific dataset.

6.5 Local Science Drivers — the Next 2–5 Years

6.5.1 Instruments and Facilities

During the next 2–5 years, the LHC will go from startup to operating at design luminosity. The complexity of events, trigger rate, event processing times, and average event sizes will increase, but the operating models of the experiments that will be exercised in the next year will be recognizable in the next 2–5 years. Most of the increases in facility capacity for processing, disk storage, and archival storage will come from technology improvements, while maintaining a similar complexity for the facility. Processing and storage nodes will be replaced with faster and larger nodes, though the number of nodes should remain roughly constant. Usage of GPUs may become an integrated part of the hardware. Cloud computing and leadership class high performance computing (HPC) sites may be used transparently through PanDA as they become available to HEP.

6.5.2 Software Infrastructure

The local software infrastructure is not expected to change in the next 2–5 years. ROOT will continue to be used widely, perhaps increasing in importance with the consolidation of data formats expected for the next LHC run. A new ATLAS data management system will be deployed in 2014, which should meet all requirements during the next 5 years.

6.5.3 Process of Science

The scientific process for the LHC will run in cycles over the next 2–5 years. The start of the new Energy Frontier offers the opportunity for rapid discovery as thresholds for production are crossed. Some of these processes, like some supersymmetry channels, turn on extremely fast and can be observed very early, if there is a good understanding of the detector and the background. As more data is analyzed, the process of discovery turns to signals that occur less frequently, an endeavor that requires the analysis of larger quantities of data. In 2015, the LHC experiments will have the opportunity to cross the Energy Frontier at 13 TeV, which will require rapid assessment of the data as we look for new physics knowledge. With the increase in data volume, a very careful and detailed analysis of large datasets will look for more subtle physics.

The ATLAS collaboration will continue to analyze the data from the 7 and 8 TeV runs during the next two years. The high-energy Run 2 is expected to begin in 2015 and continue with proton-proton collisions at a center-of-mass energy of 13 TeV until 2018, by which time approximately 100 fb^{-1} of data will have been collected.

Several recent computing developments are changing the process of science in ATLAS. The first is the effort to define a new data format that combines the flexibility of the object-based AOD format and the convenience of the flat DPD ntuple. This will allow ATLAS to store datasets in a single analysis format instead of both AOD and DPD, thereby reducing the size of the expected 2012 dataset by a factor of 3 relative to today's size. The second is the FAX effort on federated storage access via the XrootD protocol. This will enable direct access of remote data from local analysis jobs. The third is the opportunity to flock analysis jobs submitted on the local computing clusters to high-volume queues in remote facilities, outside of the existing Grid paradigm.

At the moment, an MC production volume of 10 times the number of data events is foreseen for Run 2 physics analysis. We have estimated that the new datasets will occupy about 3 times the current storage space, even taking into account the data format improvements outlined above. A centralized skimming production service will decrease the data volume on the local computing centers, and new tools are being developed to use both parts of the merged data format effectively.

An example analysis in 2015 may begin with a centrally produced skimming algorithm defined by a physics group and using the merged AOD/DPD format dataset as input at the remote facility. The skimmed ntuples (on order of 5 TB) could be transferred to the local computing clusters, or local analysis jobs could access the skimmed dataset directly using the FAX mechanism. The additional possibility of analysis at remote facilities is discussed below.

6.6 Remote Science Drivers — the Next 2–5 Years

6.6.1 Instruments and Facilities

The Tier-1 centers will produce large samples when the whole collected data is reprocessed. These larger products must be synchronized to other Tier-1 centers. The

samples selected by physics groups to be served to Tier-2 centers will increase in size as the integrated luminosity increases, but the time the physics groups are willing to wait for data is probably roughly constant, so the network bandwidth required from both Tier-1-to-Tier-1 and Tier-1-to-Tier-2 will increase. We expect a larger fraction of the derived data to be placed automatically at the Tier-1 and Tier-2 centers through automated caching systems like PD2P (PanDA Dynamic Data Placement), which will increase demands on networking.

One area in which complexity is increasing is in the number of batch slots of processing. The batch-slot count is steadily increasing, as most performance improvements are achieved by increasing the number of processor cores, with more modest improvements in the speed of each individual core. At the Tier-1 and Tier-2 centers, this increases the number of applications operating and increases the overall bandwidth from the local storage. It is reasonably safe to predict that the LHC experiments will see a two- to three-fold increase in the required rate from local storage to accommodate the growing number of cores.

6.6.2 Software Infrastructure

PanDA is evolving continuously to provide physicists with the same common interface regardless of the facilities available for data processing, storage, and networking. Through the ASCR-funded BigPanDA project, various innovations will be introduced. While the Grid middleware may change, and as new cloud computing and Leadership Computing Facilities (LCFs) become available, the user data analysis model will remain the same in the next 2–5 years. There will be increased demands on networking as higher energy and luminosity at the LHC drives the distributed computing model to higher complexity and more direct access to data.

6.6.3 Process of Science

Over the next 5 years, the science goals of the ATLAS experiment will shift from the detailed precision measurements of 7 and 8 TeV collisions to the search for new physics in the early 13 TeV data. Networking performance will play a significant role in enabling physicists to access large datasets repeatedly as new detector calibrations and corrections are developed.

The new ATLAS analysis model, to be commissioned during an MC data challenge in 2014, is intended to reduce the computing resources required for analysis and consolidate the physics data formats. At the beginning of the 13 TeV run, some detector performance studies will access unskimmed data in formats that are as close as possible to raw detector output. These studies will commission the new detector systems for science.

One possible development concerns analysis at remote facilities. Because of concerns about the future of local computing investment at universities, ATLAS has prototyped a remote facilities queue that accepts batch jobs from local clusters. These jobs would run on the remote SEs at the facilities or on data accessible through FAX (as described in Section 6.4.1). ATLAS is also investigating the feasibility of providing a central analysis

cluster that would be accessed remotely by physicists from their institutions. Such a

Table 1. Event sizes, samples, and processing times for resource calculations.

LHC and data taking parameters		2012 pp actual	2013 pp	2014 pp	2015 pp
Rate [Hz]	Hz	400 + 150 (delayed)	0	0	1000
Time [sec]	MSeconds	6.6	0	0	5.0
Real data	B Events	3.0 + 0.9 (delayed)	0	0	5.0
Full Simulation	B Events	2.6 (8 TeV) + 0.8 (7 TeV)	0.4 (2010 MC) + 0.5 (2011 MC) +2.6 (2012 MC)	1.0 (2012 MC) + 2.0 (13 TeV MC)	1.0 (50 ns MC) + 2.0 (25 ns MC)
Fast Simulation	B Events	1.9 (8TeV) + 1.0 (7 TeV)	0.6 (2010 MC) +1.0 (2011 MC) +4.4 (2012 MC)	2.0 (2012 MC) + 2.0 (13 TeV MC)	2.0 (50 ns MC) + 3.0 (25 ns MC)
Event sizes					
Real RAW	MB	0.8	0.8	0.8	1.1 (50 ns) 1. (25 ns)
Real ESD	MB	2.4	1.1 (2010) 1.1 (2011) 2.4 (2012)	1.1 (2010) 1.1 (2011) 2.4 (2012)	2.5 (50 ns) 2.5 (25 ns)
Real AOD	MB	0.24	0.16 (2010) 0.16 (2011) 0.24 (2012)	0.16 (2010) 0.16 (2011) 0.24 (2012)	0.35 (50 ns) 0.25 (25 ns)
Sim HITS	MB	0.9	0.8 (2010 MC) 0.8 (2011 MC) 0.9 (2012 MC)	0.9 (2012 MC) 1.2 (13 TeV MC)	1.2 (50 ns MC) 1.2 (25 ns MC)
Sim ESD	MB	3.3	1.9 (2010 MC) 1.9 (2011 MC) 3.3 (2012 MC)	3.3 (2012 MC) 3.5 (13 TeV MC)	3.7 (50 ns MC) 3.5 (25 ns MC)
Sim AOD	MB	0.4	0.26 (2010 MC) 0.26 (2011 MC) 0.4 (2012 MC)	0.4 (2012 MC) 0.5 (13 TeV MC)	0.55 (50 ns MC) 0.5 (25 ns MC)
CPU times per event					
Full sim	HS06 sec	3100	2700 (2010 MC) 2700 (2011 MC) 2790 (2012 MC)	2511 (2012 MC) 3500 (13 TeV MC)	3500 (50 ns MC) 3500 (25 ns MC)
Fast sim	HS06 sec	260	250 (2010 MC) 250 (2011 MC) 234 (2012 MC)	211 (2012 MC) 250 (13 TeV MC)	250 (50 ns MC) 250 (25 ns MC)
Real recon	HS06 sec	190	108 (2010) 108 (2011) 190 (2012)	108 (2010) 108 (2011) 150 (2012)	230 (50 ns) 180 (25 ns)
Sim recon	HS06 sec	770	200 (2010 MC) 300 (2011 MC) 616 (2012 MC)	493 (2012 MC) 500 (13 TeV MC)	560 (50 ns MC) 500 (25 ns MC)

cluster would be feasible only if the collaboration members have low-latency network connection to the cluster.

The parameters that drive computing and networking resource requirements — LHC operation plans and ATLAS trigger rates, amount of simulated data, event sizes, and processing times — are input parameters to the resource model. Such parameters are shown in Table 1.

Data Distribution Plan for 2013–2015

In the updated model foreseen for 2013–2015, the number of preplaced AOD replicas kept in Tier-1 and Tier-2 disks for both real and simulated data will remain the same as in 2012. AODs from the most recent (real data) reprocessing and the corresponding simulation will be preplaced in two copies, and the corresponding real data DESDs in one copy in Tier-1 disks. All AODs from data (re)processings and simulation production that remain relevant for analysis will be kept in two preplaced disk copies in the Tier-2 sites,

as will the real data DESDs in one copy. The number of copies of AODs corresponding to different (re)processings and simulation productions in Tier-1 sites will be decreased from two to zero, according to their relevance and popularity for physics analysis. While in 2011 and 2012 this was done via a human decision process within the ATLAS Computing Resources Management, in 2013 and thereafter auxiliary automated mechanisms based on popularity will be introduced to provide an additional dynamic component in reducing the number of pre-placed AOD replicas in Tier-1 disks.

In addition, the simulated event summary data (ESD) and raw data object (RDO) (where RDO is the simulation equivalent of RAW), which are produced only upon explicit request for specific simulated samples and which have proved essential to the combined performance and trigger groups, will be kept at Tier-2 sites in one copy in 2013–2015. The total simulated ESD volume is estimated to correspond to 5% of the total simulation

Table 2. Input parameters for resource calculations.

Tier-1 disk policy (sum Tier-1s)	**2012**	**2013**	**2014**	**2015**
Sim RDO disk copies	0	0	0	0
Sim ESD disk copies	0	0	0	0
Sim AOD disk copies	2	2	2	2
Real RAW disk copies	1	0.05	0.05	1
Real ESD disk copies	0.13	0.05	0.05	0.2
Real AOD disk copies	2	2	2	2
Real DESD disk copies	1	1	1	1
Tier-1 processing				
Number of reprocessing/year	1	0.22	1.06	1
Tier-2 disk policy (sum of T2's)				
Sim RDO disk copies	0.05	0.05	0.05	0.05
Sim AOD disk copies	2	2	2	2
Sim ESD disk copies	0.2	0.05	0.05	0.1
Real RAW disk copies	0	0	0	0
Real AOD disk copies	2	2	2	2
Real DESD disk copies	1	1	1	1

statistics in 2012 (20% was assumed in the 2012 Resource Request), and is expected to remain at this level in 2013 and 2014, after which it is expected to increase to 10% in 2015. The RDO fraction is assumed to remain constant at 5% of the total statistics.

6.7 Beyond 5 Years — Future Needs and Scientific Direction

The current CERN schedule for the LHC program shows a high-energy, high-luminosity run beginning in 2019 (Run 3). The ATLAS experiment will be upgraded in a long 1-year shutdown in 2018 to replace calorimeter front-end electronics and trigger hardware, so that full readout information can be used in the calorimeter and muon triggers. A fast-track trigger system will be installed at the same time.

A proposal to increase further the LHC luminosity to 5 x 10^{34} $cm^{-2}s^{-1}$ beginning in 2023 has been presented to the CERN Council. A long run at this increased luminosity would yield 3000 fb^{-1} of 14 TeV collisions, and this data sample would open up the possibility for precision measurements of the Higgs boson couplings and extend the reach for new physics searches.

6.8 Network and Data Architecture

The DOE ASCR-funded Big Data PanDA (BigPanDA) and the NSF-funded Advanced Network Services for Experiments (ANSE) projects propose new enhancements to the networking model that integrates the NE with PanDA WMS. Successful completion of these two projects will enable the science of the future LHC program. A dynamically configurable network architecture will be an essential component of future capabilities.

High-speed Campus Networking

Campuses have a fiber-rich environment, with fiber optic connectivity among buildings. As vendors increase the throughput available with their network equipment, the existing or planned fiber footprint ensures that the basic physical infrastructure can keep pace with technology.

The following is an example of how campus networking will be implemented at the University of Chicago (UC). Similar plans exist at several other institutions hosting ATLAS Tier-2 centers.

To optimize data flows, UC will acquire two items: (1) a dedicated Core Research Switch capable of providing Layer-3 routing and Layer-2 switching functionality with both 10 Gbps and 100 Gbps optics and enough fabric/backplane bandwidth per slot to support multiple 100 Gbps connections and (2) a Science DMZ aggregation switch. The first switch will be a common, high-speed confluence of external and intracampus large data flows, as distinct from the general-purpose network. This will provide the ability to extend high-speed connectivity among multiple laboratories and compute clusters on campus. It will also serve as the main Science DMZ switch. In the future, we will establish a more diverse, distributed Science DMZ/high-performance research network by the acquisition of a second switch via future NSF funding or as campus networking budget allows.

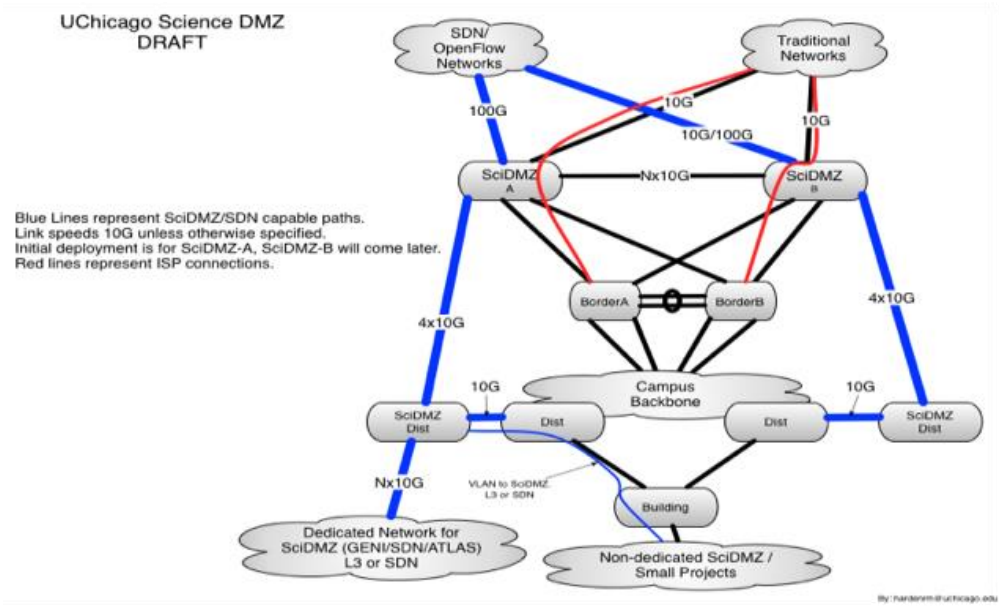


Figure 12. Draft architecture for the UC campus.

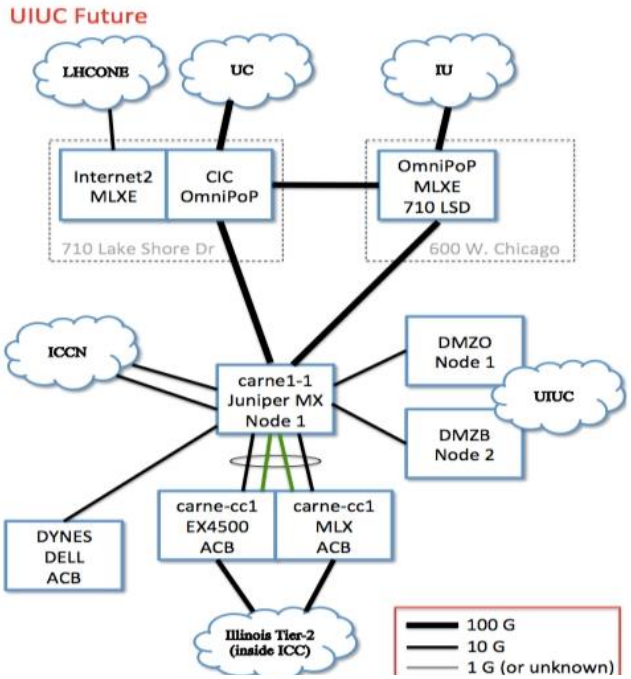


Figure 13. Draft network configuration for site in the Chicago area.

Taking the example of the Midwest Tier-2 Center (MWT2), Figure 13 shows how sites associated with the distributed ATLAS Tier-2 center in the Chicago area plan to arrange their network connectivity in fall 2013. Note: Also planned is to add the Great Lakes Tier-2 (UMichigan and Michigan State University) to this configuration.

6.9 Collaboration Tools

The ATLAS collaboration's main collaboration and communication tool is the Vidyo service, which is capable of voice and video communications through desktop, mobile, and H.323/SIP clients. CERN provides registered users access to this service for CERN-related meetings, but there are equal numbers of distinct guest users and registered users in 2013. Currently approximately 2,000 Vidyo meetings take place per month within ATLAS, and the Vidyo service is integrated with the CERN Indico agenda server. The vast majority of meetings have fewer than five connected clients, but a small number of meetings can have more than 200 clients. In nearly every meeting, only the speaker or main meeting room is shown on video; the other participants transmit audio only.

There is also some need for robust telephone/audioconferencing for meetings with up to 20 participants, especially for colleagues who do not have Vidyo accounts through CERN or who desire a simpler teleconferencing system. Most of these colleagues use the ReadyTalk service.

6.10 Data, Workflow, Middleware Tools, and Services

The primary high-level tools used by the ATLAS experiment are PanDA for workflow management, and the distributed data management system DQ2 (Don Quijote2). These tools work together to provide a uniform interface to the distributed computing and storage resources available to the collaboration. DQ2 will be replaced by the Rucio system in 2014. PanDA is evolving to BigPanDA in 2014–15, which will integrate network elements as a resource in workload management. These high-level applications insulate physicists from the diverse middleware systems and heterogeneous computing systems used in the distributed computing resources available to ATLAS.

PanDA has demonstrated the capability to scale well as the number of users, data volume, and available resources have almost exponentially increased over the past 2–3 years. We expect PanDA and Rucio to evolve and manage the growing needs of ATLAS for the next decade. This will put additional demands on networking as described in this document. Computing clouds are already integrated into PanDA and have provided additional resources to ATLAS as needed for physics publications. We expect LCF facilities to become critical contributors to future physics goals at the LHC.

For the next few years, distributed computing facilities will probably still be accessed primarily through WLCG components, but with an increasing use of Infrastructure as a Service (IaaS) and other technologies. WLCG is expected to concentrate its efforts on a few topics specific to HEP or where it can have a leading role, while integrating industry-standard products in other areas. These will then also be incorporated into the ADC toolkit. Cloud computing is one example.

With the increased load expected after the upgrade in 2018, ATLAS will have to optimize the usage of computing resources within the LHC environment (CPU, storage, network, maintenance manpower), and must evolve to adapt to the changing ATLAS workflows, the changing hardware, and in particular to the increased data flows and processing volumes implied by the upgrades. Changes in technology will imply development that far

exceeds the normal M&O (maintenance and operation), for instance work on clouds and virtualization, or new storage techniques. Networking is a major area needing development to cope with future data flows and volume. Next-generation advanced networked applications (i.e., cloud-based services) will require a set of network capabilities and services far beyond what is available from networks today. A new class of intelligent network services is needed in order to satisfy additional application-specific requirements and to feed the co-scheduling algorithms that will search for real-time and scheduled resources that span the network and application spaces associated with large-volume, worldwide distributed data analyses.

6.11 Summary Table

Based on the parameters shown in Tables 1 and 2 (number of real and simulated events, event sizes, and dataset replication factors, etc.), the following data volumes need to be accommodated on disk and tape worldwide, as the U.S. sites currently host and will continue to host 23% of the total data volume.

The table below summarizes the expected daily data transfer volumes (showing the average and peak) to and from BNL's Tier-1 site and the five U.S. Tier-2 sites. Data are exchanged between sites in the United States domestically and sites in Europe and Asia via trans-oceanic links.

CPU [kHS06]	2012 requested	2012 actual	2013	2014	2015
CERN	*111 (111)*	*111*	*111 [111] (111)*	*111 [111] (111)*	*111 -> 240*
Tier-1	*295 (259)*	*420*	*316 [319] (333)*	*385 [373] (326)*	*478 [502]*
Tier-2	*319 (332)*	*634*	*360 [355] (395)*	*412 [408] (398)*	*522 [540]*
Disk [PB]	WLCG factor 0.7		WLCG factor 1.0		
CERN	*11 (9)*	*10*	*10 [11] (10)*	*12 [11] (10)*	*15*
Tier-1	*29 (30)*	*47*	*35 [35] (36)*	*35 [36] (33)*	*47 [51]*
Tier-2	*48 (45)*	*52*	*51 [52] (49)*	*56 [56] (46)*	*65 [69]*
Tape [PB]					
CERN	*21 (18)*	*21(and 9 ESD)*	*25 [27] (27)*	*29 [31] (31)*	*38*
Tier-1	*31 (38)*	*31*	*42 [43] (41)*	*55 [53] (53)*	*74 [77]*

Table 3. Daily data transfer volumes to/from the Tier-1 and the Tier-2 sites.

Process of Science	To Tier-1		From Tier-1		To Tier-2		From Tier-2	
2014	TB/day/site	TB/day/site peak	TB/day/site	TB/day/site peak	TB/d/site	TB/d/site pk	TB/d/site	TB/d/site pk
Tier-0 Export	0	0						
Data Consolidation	26	70	43	86	8	17	8	17
Group Subscriptions	8	17	4	8	2	4	1	3
User Subscriptions	7	13	8	17	2	4	3	4
Production	8	8	3	8	2	3	4	6
PD2P/Data Brokering	4	8	17	26	4	8	3	8
Analysis, Merging, DQ2 (LAN)	1100	2200			400	800		
2015								
Tier-0 Export	15	40						
Data Consolidation	26	70	43	86	8	17	8	17
Group Subscriptions	8	17	4	8	2	4	1	3
User Subscriptions	7	13	8	17	2	4	3	4
Production	8	8	3	8	2	3	4	6
PD2P/Data Brokering	4	8	17	26	4	8	3	8
Analysis, Merging, DQ2 (LAN)	1100	2200			600	1200		
2016								
Tier-0 Export	20	50						
Data Consolidation	35	70	50	100	12	22	12	22
Group Subscriptions	10	17	8	12	4	8	2	6
User Subscriptions	10	18	12	20	5	10	6	8
Production	12	15	6	12	4	6	8	12
PD2P/Data Brokering	8	16	25	35	10	20	6	16
Analysis, Merging, DQ2 (LAN)	1600	3200			1000	2000		
2017								
Tier-0 Export	20	50						
Data Consolidation	35	70	50	100	12	22	12	22
Group Subscriptions	10	17	8	12	4	8	2	6
User Subscriptions	10	18	12	20	5	10	6	8
Production	12	15	6	12	4	6	8	12
PD2P/Data Brokering	8	16	25	35	10	20	6	16
Analysis, Merging, DQ2 (LAN)	1600	3200			1000	2000		

Note that no LHC data taking will take place in 2018. The data rates are expected to stay at the level of 2017 or be slightly lower. The following tables show a breakdown of the total data volumes into the various data categories. The United States Tier-1 and Tier-2 sites will host 23% of the total data.

Table 4. Tier-1 disk occupation by data category.

Tier-1 Disk ***(PB)***	***2012 requested***	***2012 actual***	***2013***	***2014***	***2015***
Current RAW data	3.5	1	0.5	0.5	5.2
Real ESD+AOD+DPD data	4.5	14	9	5	11
Simulated RAW+ESD+AOD+DPD data	8.0	12	9	13	14
Calibration and alignment outputs	0.4	0.4	0	0	0.3
Group data	6	15	12	12	12
User data (scratch)	2	2	1.4	1.4	1.4
Cosmics	0.2	0.2	0.2	0.2	0.2
Processing and I/O buffers	4.3	3	3	3	3
Total	***29***	***47***	***35 [35]***	***35 [36]***	***47 [51]***

Table 5. Tier-1 tape occupation by data category.

Tier-1 Tape (PB)	*2012 requested*	*2012 actual*	*2013*	*2014*	*2015*
Real RAW+AOD+DPD data	15	10	11	13	21
Cosmics and other data	4	4	4	4	4
Group + User	3	0.3	4	5	6
Simulated HITS+AOD data	12	17	24	33	43
Total	***35***	***31***	***42 [43]***	***55 [53]***	***74 [77]***

Table 6. Tier-2 disk occupation by data category.

Tier-2 Disk (PB)	*2012 requested*	*2012 actual*	*2013*	*2014*	*2015*
Real AOD+DPD data	13	9	7	11	15
Simulated HITS+RDO+ESD+AOD	22	18	21	21	26
Calibration and alignment outputs	0.3	0.3	0.2	0.2	0.2
Group data	10	19	20	20	20
User data	2	2	1.4	1.4	1.4
Processing and I/O buffers	1	3	2	2	2
Total	***48***	***52***	***51 [52]***	***56 [56]***	***65 [69]***

7 CMS Physics Analysis Case Study

7.1 Background

Networks are the backbone of CMS physics analyses. The raw data (1 PB/year) comes from the CMS experiment in CERN, Geneva (Tier-0) and the simulated MC data are generated at its fifty or so Tier-2 centers around the world (generating about 2 PB/year) and transferred to its seven Tier-1 centers where they are reconstructed (3 PB/year). The summary of the reconstructed data (about 0.5 PB/year) is used by physicists at Tier-2 and Tier-3 centers to obtain results for publication; thus enabling a reliable free flow of these summary data to centers around the world, in response to physicists' requests, is key to our success.

Between 2010 and 2012, CMS physics analyses were primarily based at its Tier-2 centers worldwide. Typical analysis jobs access both real data from the LHC experiment and several MC datasets. The amount of data accessed for a typical analysis is 50 TB of real data and 100–200 TB of MC data. Typically, an analysis uses serious amounts of computing for a few months, requiring several tens of thousands of CPU hours. The CMS has more than a hundred such analyses, pursued in parallel by collaborators around the world. In this document, we study one such analysis and provide aggregate usage of resources for physics analyses performed in the past year at the U.S. CMS Tier-2 computing centers.

7.2 Collaborators

The CMS is the virtual organization participating in this case study. The contributing facilities are the CMS Tier-0 computing center at CERN, Geneva; CMS Tier-1 computing centers at Fermilab, KIT (Germany), RAL (U.K.), CNAF (Italy), IN2P3 (France), PIC (Spain), and ASGC (Taiwan); and CMS Tier-2 U.S. computing centers at the Univeristy of Wisconsin–Madison, University of California at San Diego, University of Nebraska Lincoln, Purdue, MIT, Univeristy of Florida, and Caltech. Note that Tier-0 and Tier-1 are primarily involved in providing access to their custodial data. Tier-2 centers provide both computing resources and ephemeral storage for physics analyses. Roughly 1,000 users are involved in running jobs, whereas a dozen users are involved in the detailed studies reported.

7.3 Key Local Science Drivers

7.3.1 Instruments and Facilities

The OSG infrastructure is the key science driver for CMS physics analyses in the United States. The U.S. CMS Tier-2 computing centers, each with about 3,000 cores and 1 PB of usable storage, are interconnected via multiple 10 Gbps network links. Peer-to-peer connections are made among themselves and to the Fermilab-based Tier-1 computing center. All U.S. Tier-2 sites are also connected to international sites (both Tier-1 and Tier-2) and, of course, to the Tier-0 center at CERN, which originates the bulk raw data from the detector.

7.3.2 Software and Data Infrastructure

The primary software used by the analysts is based on CMSSW C++ framework (the CMS Offline Experiment Software) that supports simulation, reconstruction (both real and MC), and analysis workflows. CMSSW is based on ROOT I/O for object streaming to/from files. The CMS Computing group does the bulk processing of simulation and reconstruction jobs centrally to produce result files containing AOD, including simulation information (AODSIM) for MC. Table 7 provides a description of the CMS data tiers and total data volumes for 2012. These AOD/AODSIM ROOT files with custom data formats are transferred around the network using Grid tools under the auspices of the PhEDEx (Physics Event Data Export) system. Table 8 shows volumes of data at Tier-1 sources and integral data volume hosted at U.S. Tier-2 sites. Figure 14 shows the data transfer rate peaking at 10 Gbps.

Analysis workflow begins with the processing of AOD data (50–100 TB of AOD and 100–200 TB of AODSIM). Over more than a few weeks, about 10–100,000 are submitted using either the CMS Remote Analysis Builder (CRAB) system, which automatically verifies existence of data files at any CMS center and queues jobs to those resources; or local resources directly (using Condor scripts) when data is known to be available locally.

The first set of jobs, typically processed by organized groups pursuing a set of similar analyses, consists of making reduced datasets (10–50% of data processed) shared by several analysts. These reduced data files, possibly also in CMSSW data format, are placed in the local Tier-2 storage systems to which the users have access, even when the jobs run on remote machines using CRAB. These jobs are usually run once or twice in an analysis cycle, as they are time-consuming. Because this stage of analysis is based on fairly reliable centrally written software, it is stable and usable by many individuals.

Table 7. CMS Data tiers and total data volumes for 2012. A factor-of-5 increase is expected for the 2015–2017 LHC run.

Type	Description	Location
RAW (2.24 PB)	Compressed data from detector	Tier-0/Tier-1 Tape — used for organized (re)reconstruction by production operations team
GEN-SIM (3.32 PB)	MC simulated data including simulation detail	Tier-1 Tape — used for organized (re)reconstruction by production operations team
AOD (0.77 PB)	Subset of reconstructed data with analysis objects	Multiple Tier-2/Tier-3 disks — used for analysis using Grid or local computational resources (especially Tier-3)
AODSIM (4.55 PB)	Subset of reconstructed data with analysis and simulation (truth) objects	Multiple Tier-2/Tier-3 disks — used for analysis using Grid or local computational resources (especially Tier-3)

Table 8. Data hosted at CMS Tier-0 and Tier-1 data centers, which are moved to Tier-2 sites on an as-needed basis. Also listed are the total volumes of data hosted at Tier-2 data centers.

Location	Data Volume	Data Type
T0_CERN	11.9 PB	RAW tape
T1_US_FNAL	20.4 PB	RECO tape/cache
T1_ES_CNAF	6.26 PB	RECO tape/cache
T1_DE_KIT	4.26 PB	RECO tape/cache
T1_UK_RAL	3.81 PB	RECO tape/cache
T1_FR_CCIN2P3	3.38 PB	RECO tape/cache
T1_ES_PIC	1.94 PB	RECO tape/cache
T1_TW_ASGC	1.49 PB	RECO tape/cache
T2_US_* (7+Vanderbilt)	6.10 PB	Disk (noncustodial)
T3_US_LPC	1.04 PB	Disk (shared T1 cache

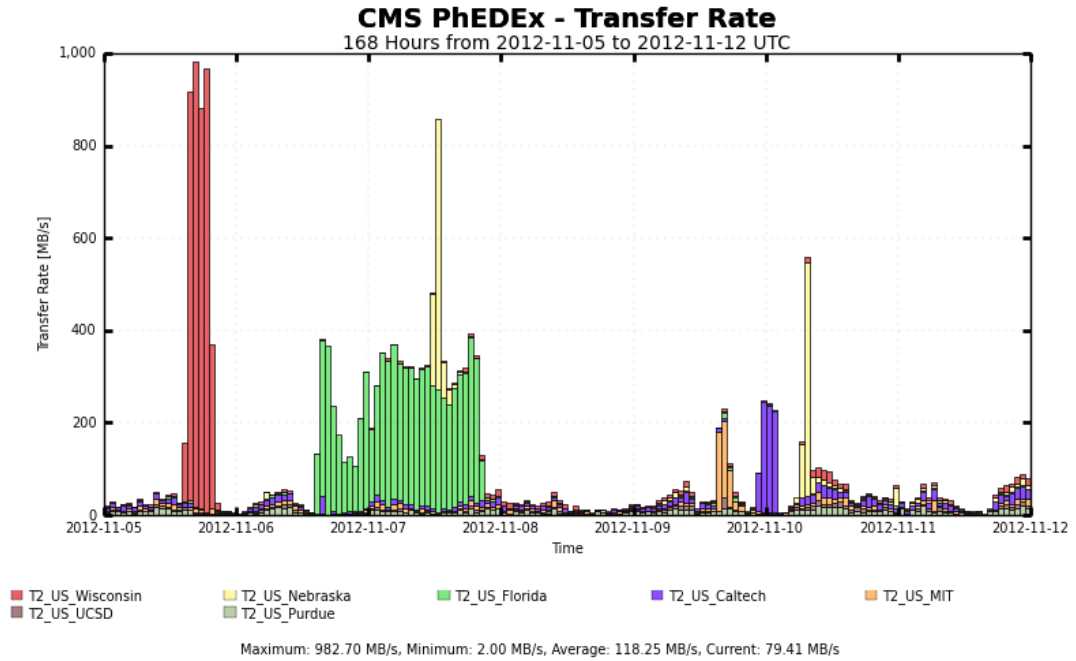


Figure 14. Data transfer rate from Fermilab Tier-1 to U.S. Tier-2 centers, showing saturation of a 10 Gbps link. The transfer rates show bursts when datasets are requested.

The latter sets of jobs typically rely on analyst-written code and are subject to frequent changes as the analysis progresses. This analysis process is iterative and takes several weeks to produce final histograms and tables. These final root files are rather small, amounting to less than few gigabytes, and are often shared using Web access. The analysis often culminates with complicated fits to extract physically meaningful quantities that are plotted and described in publications.

7.3.3 Process of Science

An individual CMS analysis project that leads to publication is usually produced by a collaboration of 10–30 physicists across the world. Before using real data, the analysis is attempted on a simulated dataset to ensure that the signal being sought or measured is cleanly separated from backgrounds. Often the analysts request MC signal samples for the project, which are produced by the central MC production service. The much larger background samples are identified and transferred to the physicist's favorite Tier-2 site, or processed using CRAB where they are located. The collaboration develops analysis software and makes several passes on the MC datasets to fully define the analysis. Often multivariate analysis techniques will require the training of neural networks or decision trees. The primary dataset for the analysis is processed. To extract the physical qualities of interest, special software — which may use maximum likelihood techniques — is

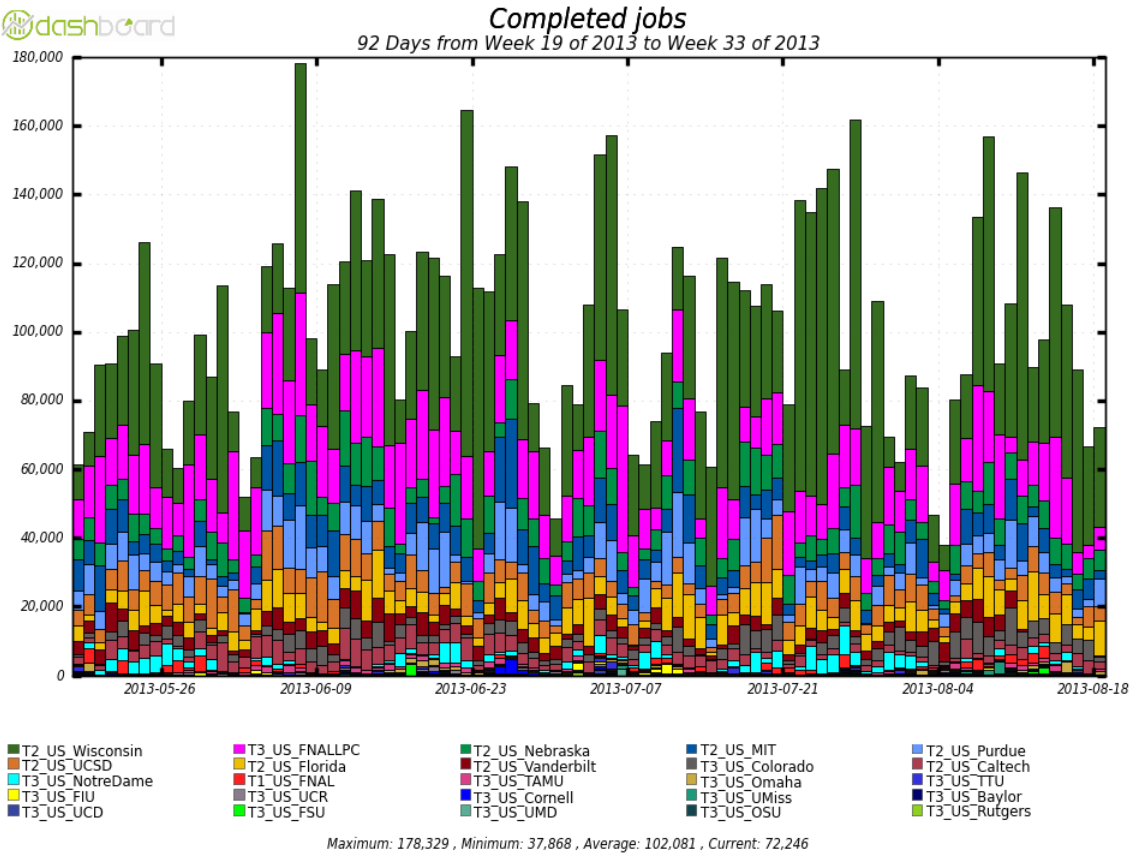


Figure 15. Completed daily analysis job count at U.S. Tier-2 and Tier-3 centers, showing a peak of 178,000 and sustained level of 102,000 per day.

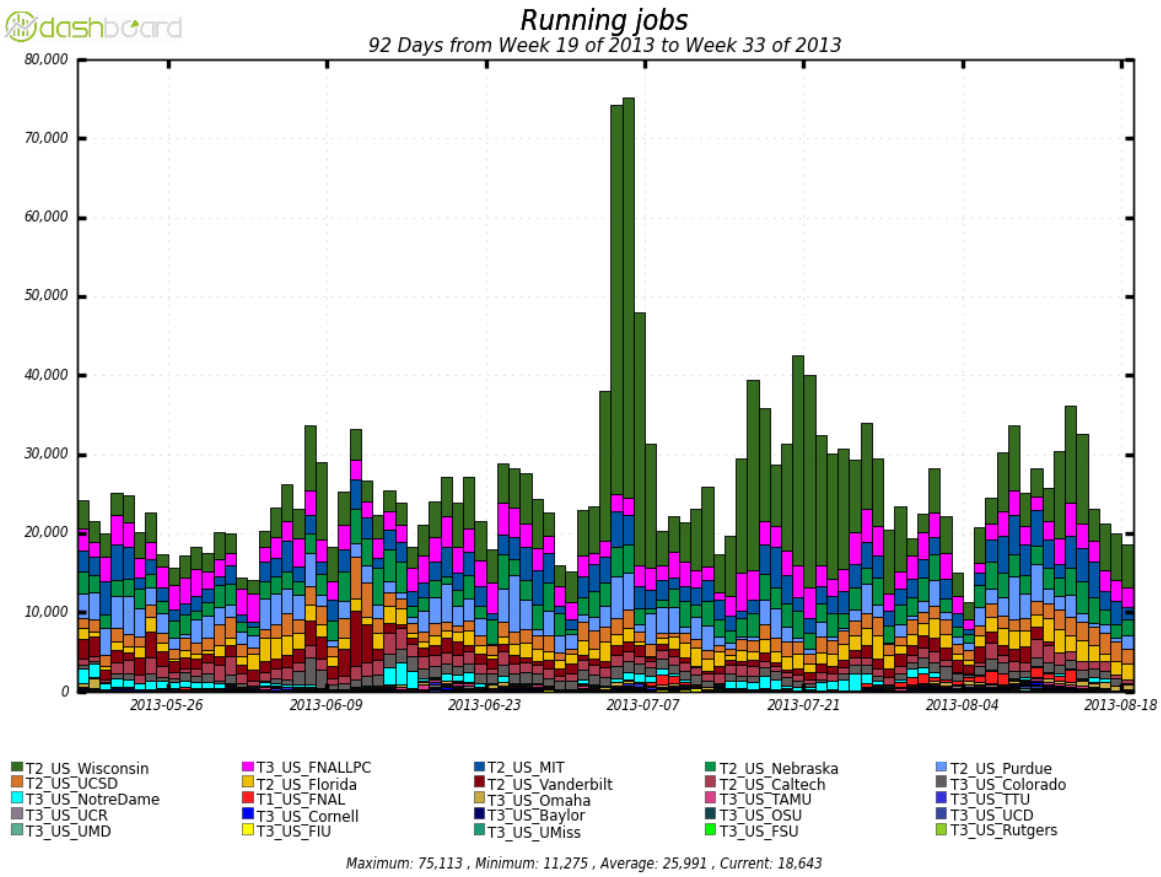


Figure 16. Running analysis job count — the large spike was due to multiple very short jobs, an oddity that should be ignored. The mean number of analysis jobs running at U.S. Tier-2/Tier-3s is 25,000.

implemented. From the inception, deriving final results takes several weeks. Each project requires network and storage services for hundreds of terabytes of data transfer, computing services for processing tens of thousands of jobs, and video collaboration services. Hundreds of such projects are in operation at any given time within a CMS experiment.

The number of analysis jobs completed each day at U.S. Tier-2 centers is shown in Figure 15, indicating sustained processing of 102,000 jobs per day and a peak of 178,000 jobs on the best day chosen. Since the analysis job time varies significantly, depending on the level of work done, the running job rate is of relevance, and is shown in Figure 16. The mean number of analysis jobs running at U.S. Tier-2/Tier-3 sites is 25,000.

Recently, ubiquitous object-level data access to the CMSSW jobs is being provided through XRootD mechanisms over the WAN using products from the AAA project. This allows even local jobs to read files from remote locations, avoiding file-transfer requests. The use of AAA is low but is likely to grow significantly. The system is especially useful for processing the shared reduced dataset files on the WAN with fellow analysts. Figure 17 shows the 200 MB/sec data usage on the WAN by 800 AAA-served jobs primarily running on Notre Dame Tier-3 compute resources.

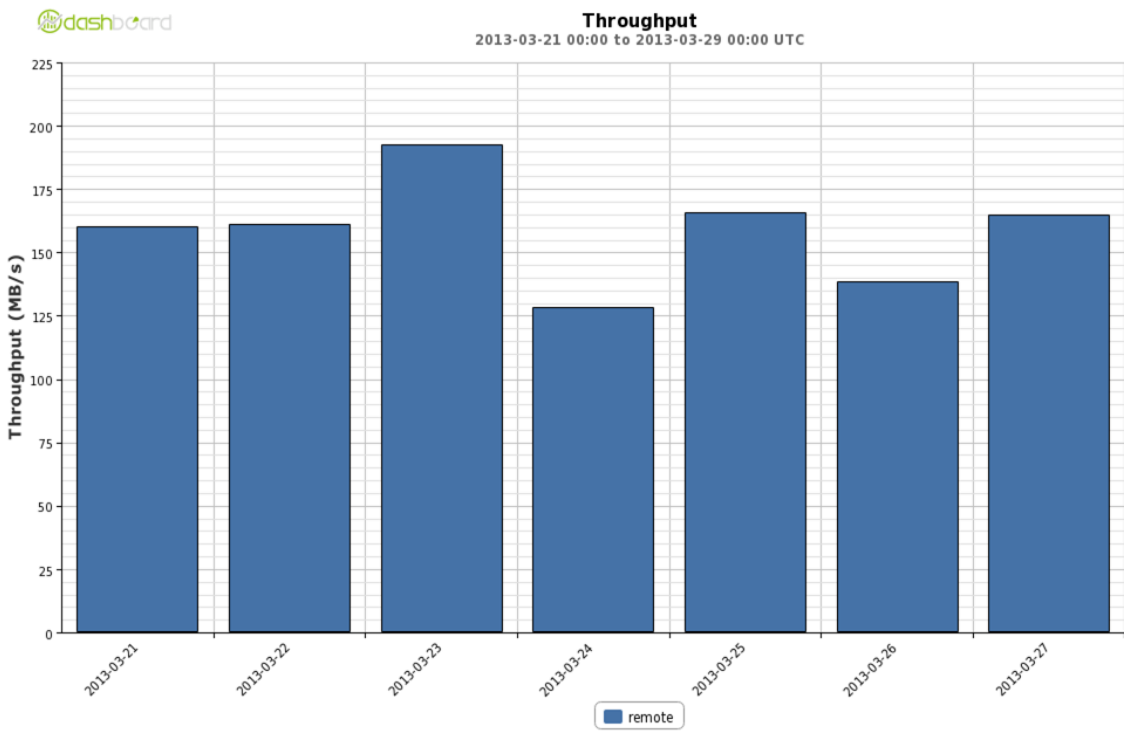


Figure 17. Data throughput to AAA-served jobs, showing a maximum of 200 MB/sec when 800 jobs were operating at Notre Dame, all reading from the WAN.

7.4 Key Remote Science Drivers

7.4.1 Instruments and Facilities

The CMS analysis workflow requires access to real LHC data totaling about 0.5 PB for the AOD data storage. The custodial storage for the AOD data is at one of six Tier-1 computing centers. Tier-1 centers also have custodial MC data in AODSIM form, amounting to several petabytes. Datasets are typically transferred to Tier-2 centers worldwide at the request of analysis operations/physics groups/individual physicists, and is available in multiple locations. Data transfer or remote job submission using CRAB are available options for processing. The physicist users on the Grid have write-access to one or two Tier-2 sites, where they transfer their processed data. Data analysis requires access to WLCG services, storage privilege at remote centers, and high network bandwidth among various tiers of CMS computing.

7.4.2 Software Infrastructure

Grid services for both storage access (e.g., storage resource manager [SRM]) and compute servers (e.g., gLite or Condor-G), PhEDEx servers, CRAB servers, and XRootD servers are used for file transfer and job management.

7.4.3 Process of Science

PhEDEx, CRAB, and XRootD/AAA services are the mainstay of the software infrastructure needed to access remote data for transfers, remote processing, or processing on the WAN.

7.5 Local Science Drivers — the Next 2–5 Years

7.5.1 Instruments and Facilities

We expect that computing centers will double or triple in capacity in the next 2–5 years. Network needs will increase proportionately.

Local resources at Tier-3 facilities will be much easier to put together, as they will not need to be capable of storage services using AAA technologies. Increased use and numbers of Tier-3 centers on campuses around the country are likely. Ease of use of non-owned or opportunistic resources or rented resources, (e.g., Amazon cloud) will result in the use of new, highly distributed resources.

7.5.2 Software Infrastructure

We anticipate that the staged transfers using PhEDEx will be reduced in favor of WAN-based analysis using AAA. Location independence of data is our primary goal; we anticipate that reliance on network services will increase. Network reliability and throughput are important criteria for the future.

7.5.3 Process of Science

The scientific process is expected to only evolve adiabatically. However, the computing services needed to achieve scientific goals can and will evolve quite quickly. Currently, physicists must pay attention to the location of data and reliability of remote hosts, often having to specify in “white lists” or “black lists” which remote hosts to use. Location independence enabled by AAA features will be well-received and adapted quickly.

7.6 Remote Science Drivers — the Next 2–5 Years

7.6.1 Instruments and Facilities

The CMS tiered computing infrastructure will grow two- to three-fold through renewed investments and (reduced) Moore’s law scaling, assuming continued support from funding agencies. Additional opportunistic resources will be brought in using temporary opportunistic resources, e.g., UCSD Supercomputer Center resources recently, and possibly commercial vendors. Flexibly provisioned resources accessed over the WAN will be the way of the future.

LHC luminosity is likely to rise by a factor of 2, accompanied by a center-of-mass energy increase of approximately a factor of 2, resulting in significantly more busy events. As a result, data growth will be substantial. It is anticipated that five times the 2010–2012 data will be collected in the period of 2015–2018. Event complexity will increase significantly. The MC simulated data volume will also rise proportionately.

7.6.2 Software Infrastructure

Analysis software infrastructure is expected to only evolve adiabatically. We do anticipate potential use of multi-coprocessor architecture, which requires modified software infrastructure, but adaptation will not have large impact on network resources yet.

7.6.3 Process of Science

The increased use of WAN-based analyses means remote centers will need to carefully and skillfully monitor and provision their resources. Nimble methods to transfer peak loads to partnering computer centers could make the process of science more efficient. Essentially, content delivery becomes the responsibility of the experiment's network services, thereby allowing the user to focus on the physics problem they are solving.

7.7 Beyond 5 Years — Future Needs and Scientific Direction

7.8 Network and Data Architecture

7.9 Collaboration Tools

The CMS collaboration primarily uses Vidyo services for videoconferencing. All weekly meetings, typically about a dozen, take place concurrently from 7 a.m.–2 p.m. EST. Often Skype is used by smaller groups for informal chats when needed.

7.10 Data, Workflow, Middleware Tools, and Services

The factor-of-5 increase in data volume over the next 5 years will result in substantial changes in analysis projects. Data placement in multiple locations, followed by jobs seeking sites with compute resources co-located with data, will not be easy. Rather, a single or small number of copies of data will be placed at Tier-2 centers, which will be accessed over the WAN as envisioned in AAA products. With sufficient provisioning of WAN bandwidth, non-owned and opportunistic compute resources will play a big role.

It is expected that groups at Tier-3 sites and on university campus grids will use their resources to process data located remotely. It is also anticipated that commercial cloud resources will be leased temporarily to satisfy peak usage.

Assuming all data access is done remotely using AAA, which maximizes the use of the WAN, we scale the current 25,000 jobs (see Figure 16) to 125,000 concurrently running jobs to process five times the data expected with increased LHC luminosity. Scaling using the Notre Dame experience of 200 MB/sec for 800 jobs (see Figure 17), we obtain a total of 313 Gbps for physics analysis. Distributing these jobs evenly to all seven Tier-2 sites, we expect to see a data read rate of 45 Gbps at each Tier-2 site. Because the data is also distributed from each Tier-2, we expect an equal 45 Gbps data serving rate from each Tier-2 center.

Data hosting at Tier-2 sites will continue, albeit at a slightly higher level (perhaps twice as high, as five times as high is unaffordable). The burst transfer rates of at 10 Gbps (See Figure 14) will probably continue at that level.

Realistically, we anticipate a mixed use of data resources, with some jobs using remote access and others accessing locally transferred data. While it is difficult to envision the exact division at this time, we can take equal division as a lower bound. In this case, the data-access rate will be 23 Gbps, whereas the increased rate of transfers for populating local caches at five times the rate would probably need additional bandwidth.

Therefore, our estimate for Tier-2 WAN connection needs ranges from 33 to 55 Gbps. Factoring in the operational efficiencies, the CMS Tier-2 sites should anticipate and provision a 100-Gbps-level connectivity. Multi-10 Gbps network access is likely necessary at the user Tier-3 sites if they are provisioning a few thousand cores.

7.11 Outstanding Issues

N/A

7.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• CMS 2010–2012 AOD datasets amount to 0.5 PB, with raw data of about 1.0 PB and reconstructed data at 1.5 PB. • Current MC simulated dataset AODSIM accessed by users is over 2.0 PB. • 2013–2014 period will see significant high center-of-mass energy MC production, which is likely to amount to 2 PB of data. • Primary mode of data access now is to use storage co-located with the compute servers.	• AOD and AODSIM is processed first using CMSSW programs to reduce to group level tuples – this analysis is carried out on the Grid. • Group tuples are processed to obtain distributions of interesting quantities – this process is also carried out on the Grid. • Distributions are statistically analyzed, for example using ML fits to extract physically meaningful quantities – this process is usually performed on user computers.	• Typical size of datasets is 20 TB. • MC signal datasets are a few GB whereas large single-lepton datasets are 100 TB. • Dataset is composed of large 2 TB files.	• LAN activity is primarily for direct object access by thousands of concurrently running jobs at a typical Tier-2. • Depending on the activity, the jobs can be I/O limited, in which case the peaks seen currently are ~10 Gbps.	• WAN transfer-time need is primarily set by user experience. Currently overnight delivery of TB-size datasets is acceptable. • Typical time to transfer datasets is about 3 hours for 10 TB datasets across the WAN. • About a dozen simultaneous requests from users are still deemed acceptable. • Data transfers to U.S. Tier-2 sites amounted to 140 TB/week. • Data transfer out of U.S. Tier-2 sites amounted to 65 TB/week. • Current provisioning of bandwidth at Tier-2 sites, 3-5 x 10 Gbps at each institute, is sufficient to already deploy object-level reading of data over WAN using AAA technologies.

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
2–5 years				
• CMS 2015–2018 AOD dataset will amount to about 5 PB, with 10 PB of raw data volume. • MC AODSIM for 2015–2018 will likely be of order 25 PB. • Future data access mode will most likely be over the WAN using AAA technologies.	Same as above	Same as above	Number of concurrently running jobs is likely to go up by a factor of 5–10, resulting in modified data-access pattern, resulting in tenfold increase in the LAN bandwidth needed — 100 Gbps being the target for Tier-2 sites.	• Factor of 5–10 growth in dataset sizes can lead to unacceptable delays in data transfer time — therefore, our anticipation that Tier-2 sites will have 100 Gbps WAN service. • Anticipated change to reading analysis objects over the WAN is likely to alleviate the storage space burden, but increase the sustained WAN activity. • Peer-to-peer transfers across all Tier-1 and Tier-2 sites and AAA analysis patterns imply necessity of 100 Gbps connectivity at all Tier-2 sites.
5+ years				
Nominally the LHC operations after 2018 will be at twice the instantaneous luminosity compared to 2015–2018 period, resulting in twice the data volume.	• Multipurpose machines with co-processors will enable more complex parallel processing jobs. • Workflow is likely to change substantially but the patterns are as yet unknown.	Currently assumed to be about the same as above.	Yet another factor-of-2 increase.	Factor of 2–3 more than above.

8 Production Transfers to Support CMS Physics

8.1 Background

The CMS is one of the four experiments recording proton-proton collisions at the LHC in Geneva, Switzerland. All detector signals of a collision are called an event. In 2012, the LHC produced 20 MHz of collisions in the center of the CMS detector. In 2015, this will increase to 40 MHz. A powerful multistage trigger system reduces the data taking rate and selects only collisions interesting for physics studies. The data taking rates in 2012 reached 1 kHz after trigger, while 2015 will start with a trigger rate of at least 1 kHz.

The CMS uses a tiered setup of distributed computing sites, shown in Figure 18, to process, store, and analyze the recorded and triggered proton-proton collisions. It relies on networks between the centers to move data around.

Many collisions are stored in files of 2–8 GB size, which are optimized for tape storage. Files are grouped in blocks smaller or equal to a typical tape cartridge size, optimized for tape writing and recall. Blocks are grouped in datasets of the same physics content.

The CMS primarily knows about two kinds of types of datasets/blocks/files: data recorded by the detector from real proton-proton collisions, and MC simulations that use the mathematical framework of the Standard Model to simulate events.

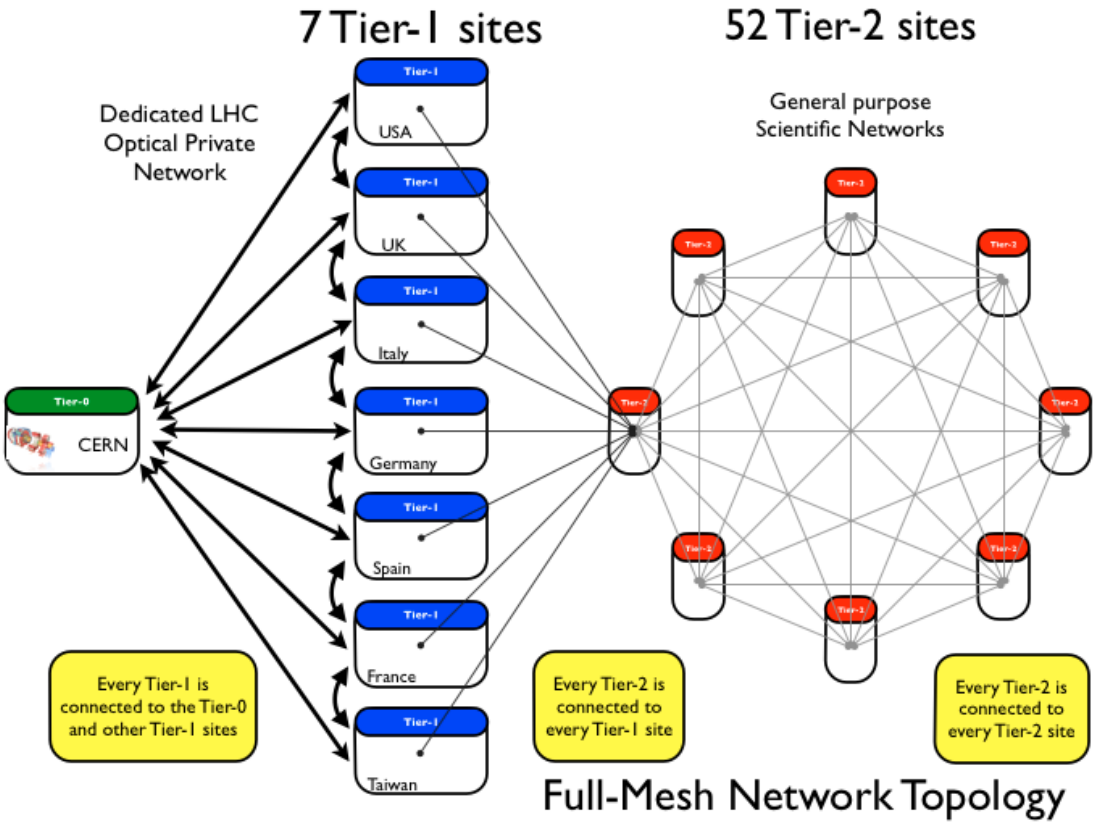


Figure 18. Tier structure and network connections of distributed CMS computing infrastructure.

The CMS records billions of data events during an LHC data-taking year and also simulations billions of events.

8.2 Collaborators

The CMS collaboration is its own virtual organization. The collaboration consists of approximately 2,500 physicists. The primary distinction of use cases on the distributed computing infrastructure consists of central tasks like the reconstruction of all data, the simulation of MC, and analysis. Analysis is chaotic, with all users submitting workflows or tasks in parallel; production is organized and uses a special role and gets special priority on all CMS computing resources.

Table 9 summarizes the distributed CMS computing infrastructure and the sites participating in it. In the United States, Fermilab is the CMS Tier-1 site. The following sites are U.S. CMS Tier-2 sites: Wisconsin, Nebraska, Caltech, MIT, Florida, Purdue, and UCSD.

Table 9. Summary of sites in distributed CMS computing infrastructure.

Tier-Level	U.S. Sites	Non-U.S. Sites	Total Sites
T0		1	1
T1	1	6	7
T2	7	45	52
T3	30	33	63

8.3 Key Local Science Drivers

8.3.1 Instruments and Facilities

A standard CMS site provides the following Grid-accessible services:

- Worker nodes organized though a local batch system
  - Access through GRID CE used by central production and analysis users (sometimes additional local access is granted to special users)
- Mass storage system (MSS) managing the available disk (dCache, Hadoop, DPM, etc.)
  - Jobs access the MSS through optimized local-access protocols
- Files placed on MSS at sites through GridFTP

Worker nodes at sites are interconnected through a low-latency local area network (LAN) infrastructure based on single gigabit infrastuctures or already emerging 10-gigabit infrastructures.

8.3.2 Software Infrastructure

The CMS uses a C++ software framework called CMSSW, which fulfills all needs of central processing and MC production. Software releases are either installed at the sites or distributed through read-only file systems based on http caches (Squids) like CVMFS (CERN Virtual Machine Filesystem). Access to calibration and alignment constants is

provided through a system that translates database access into http calls, which are also cached through Squid http caches.

8.3.3 Process of Science

Central production workflows process and prepare data and MC for analyses. In the time evolution of a physics analysis of LHC proton-proton collisions, which needs input from both data and MC simulations, central production workflows are coming before physicists analyze the data. The CMS distinguishes four primary types of production workflows, summarized in Table 10, which also gives the average event sizes from 2012.

Recorded data is processed for the first time at Tier-0 at CERN and then transferred via the network to the Tier-1 sites for safekeeping on tape (custodial storage), further processing and event selection (skimming), and further transfer to the Tier-2 sites for analysis. Each Tier-1 site has only a subset of the total amount of recorded data; CERN has a complete backup copy of all RAW data.

MC production is CPU-intensive and primarily performed at Tier-2 sites. The output is the simulation counterpart of the RAW detector data. After production, it is transferred to the Tier-1 sites for custodial storage.

Processing of the MC simulation is performed at the Tier-1 sites because of its I/O intensive nature while simulating different PileUp conditions. (Every LHC proton-proton collision consists of a primary collision and several parasitic secondary collisions called PileUp. In 2012, up to 30 PileUp collisions were produced and subsequently simulated.)

Data reprocessing is also performed at the Tier-1 sites. Both these reprocessing workflows need to stage large amounts of data from tape to disk.

Table 11 summarizes the main network workflows for CMS for the four main production workflow types.

Table 10. CMS production workflows with their input and output data formats.

Workflow	Location	Input	Output
Prompt Reconstruction	T0	Detector RAW data 2012: 0.8 MB/event	Analysis Object Data (AOD) 2012: 0.25 MB/event
Data Rereconstruction	T1	Detector RAW data 2012: 0.8 MB/event	Analysis Object Data (AOD) 2012: 0.25 MB/event
MC Reconstruction	T1	Simulated collisions 2012: 1.5 MB/event	Simulation Analysis Object Data (AODSIM) 2012: 0.3 MB/event
MC production	T2	None	Simulated collisions 2012: 1.5 MB/event

Table 11. Main production network workflows for CMS.

Network Workflow	Main Transfer Route	Transfer Pattern
Archive prompt processing output (RAW and products)	T0・T1	Sustained
Archive MC production output	T2・T1	Sustained with small bursts
Replicate analysis samples	T1・T2	Bursts

8.4 Key Remote Science Drivers

8.4.1 Instruments and Facilities

The CMS uses its distributed computing infrastructure described in Section 8.1.

8.4.2 Software Infrastructure

The CMS uses the PhEDEx system for organized transfers. It handles destination-based transfer requests (transfer dataset X to site A). The system picks out sources of sites, depending on link quality and load (a single dataset can be transferred from several source sites optimized by the system), and transfers the files of a dataset to their destination. In the end, PhEDEx schedules GridFTP transfers between sites through several layers of software to protect sites, enabling multiple streams and taking care of authentication.

8.4.3 Process of Science

In 2012, the CMS produced and reprocessed 13.5 PB of MC files (see Figure 19) and recorded and processed 9.4 PB of proton-proton collisions (see Figure 20).

The total volume transferred in 2012 amounts to more than 40 PB. Figure 21 shows the transfer rate averaged over one week for 2012.

To illustrate the transfer load worldwide, we pick a good week with many transfers in the last months of 2012 in which data taking activity was the highest.

Figure 21 shows the main transfer workflows for this week, with the sustained transfers from the Tier-0 to the Tier-1 sites and the burst transfer modes to the Tier-2 sites, which peak around 10 Gbps.

Figures 23 and 24 show the total data volumes for the different main transfer streams for this particular week. The same information is shown in Figures 25 and 26, only for all of 2012.

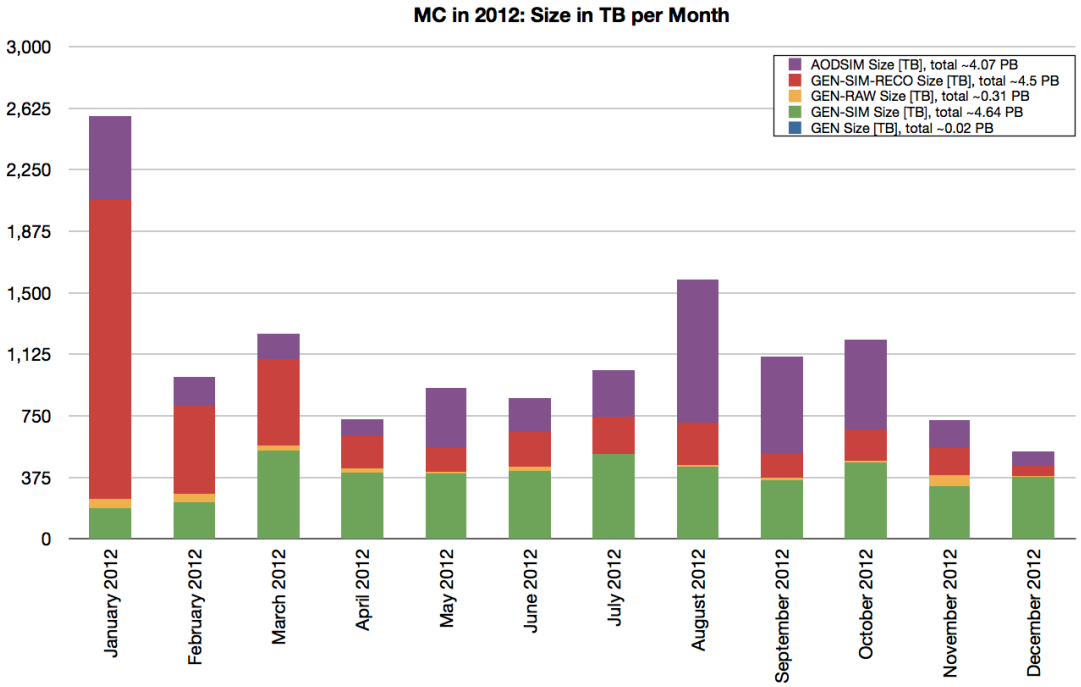


Figure 19. MC produced and reprocessed in 2012.

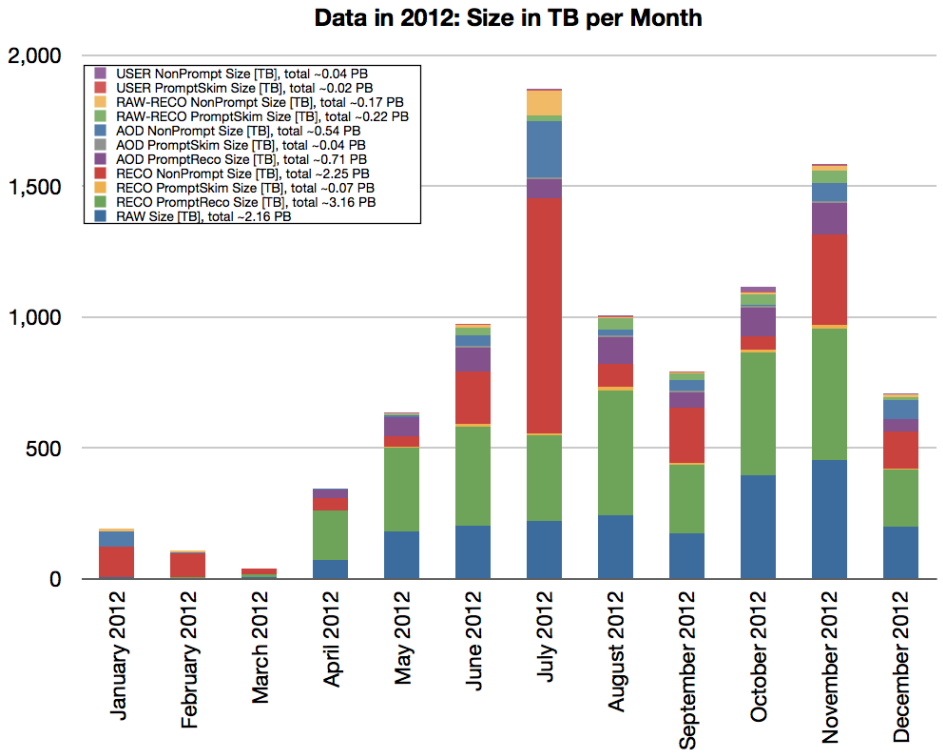


Figure 20. Volume of data recorded and processed in 2012.

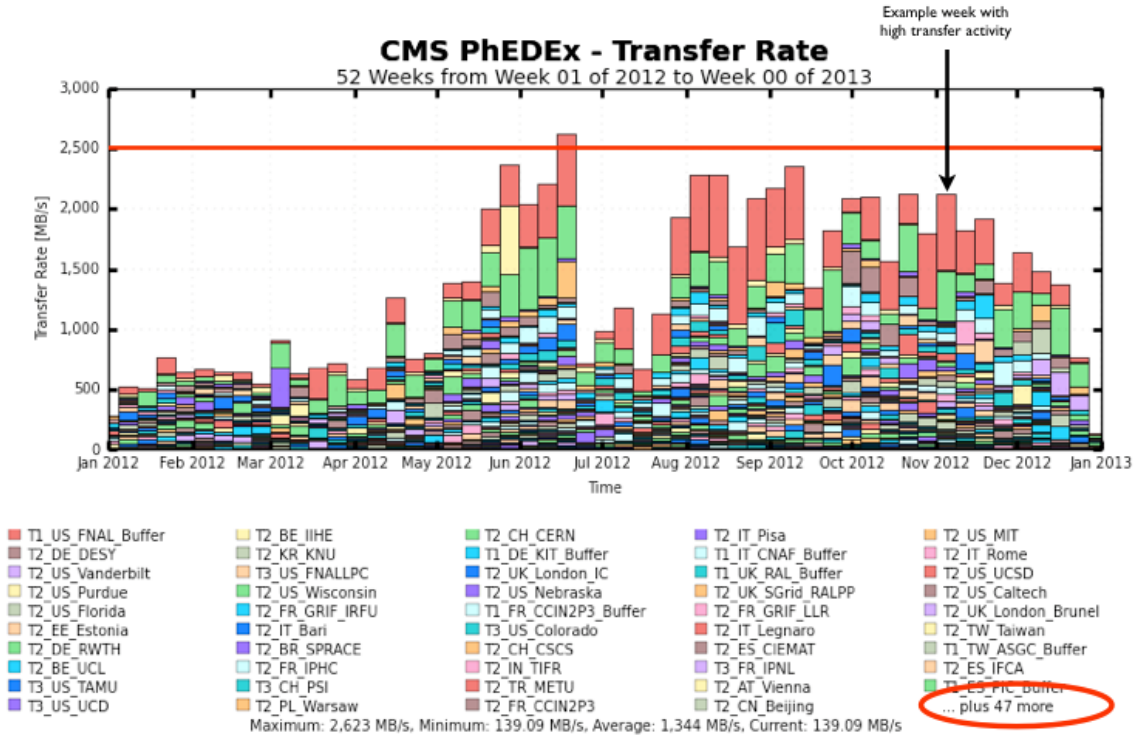


Figure 21. Transfer rate averaged over one week for 2012.

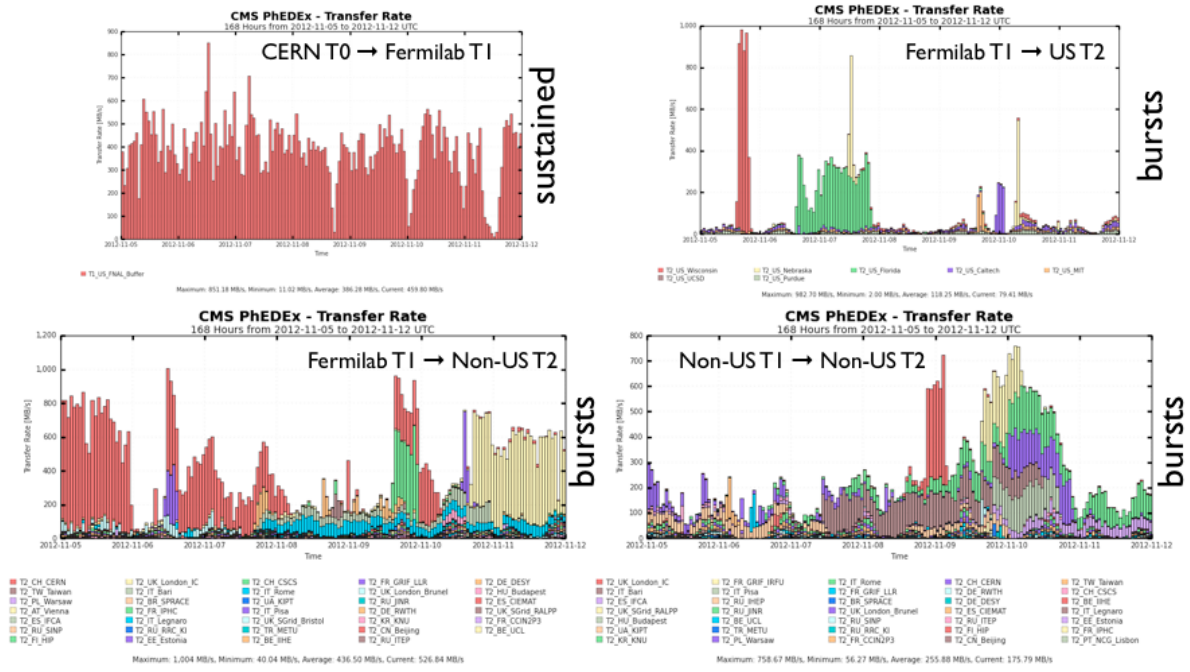


Figure 22. Main transfer workflows for one week in 2012.

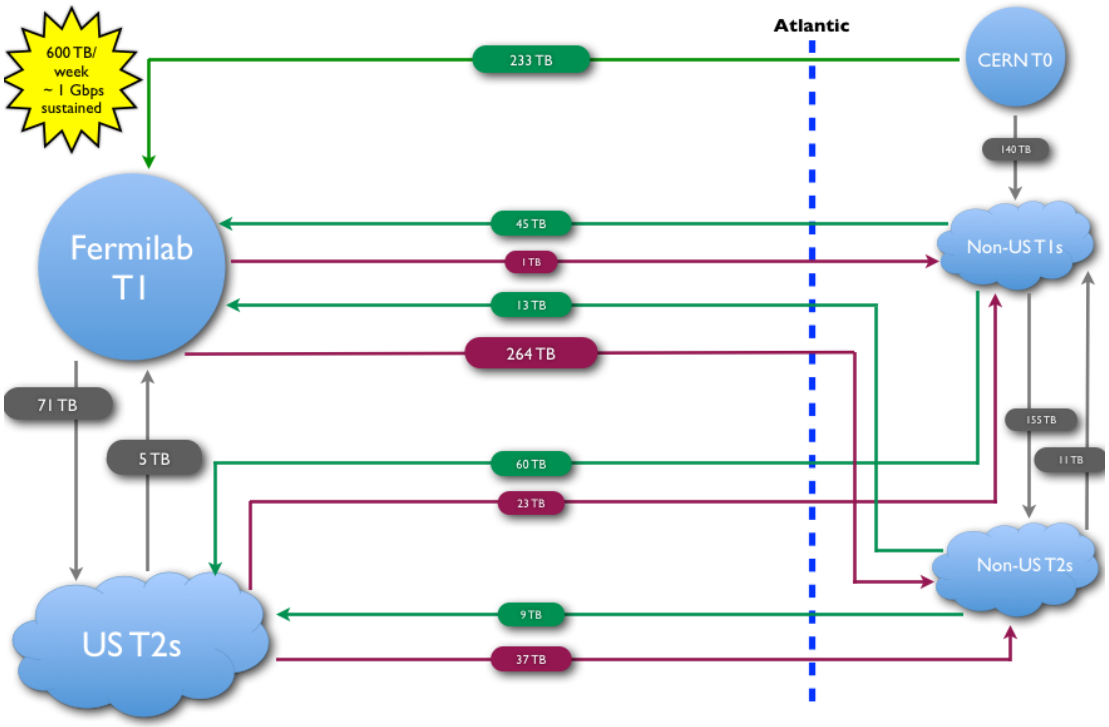


Figure 23. Total transfer volume for main transfer streams for one week in 2012.

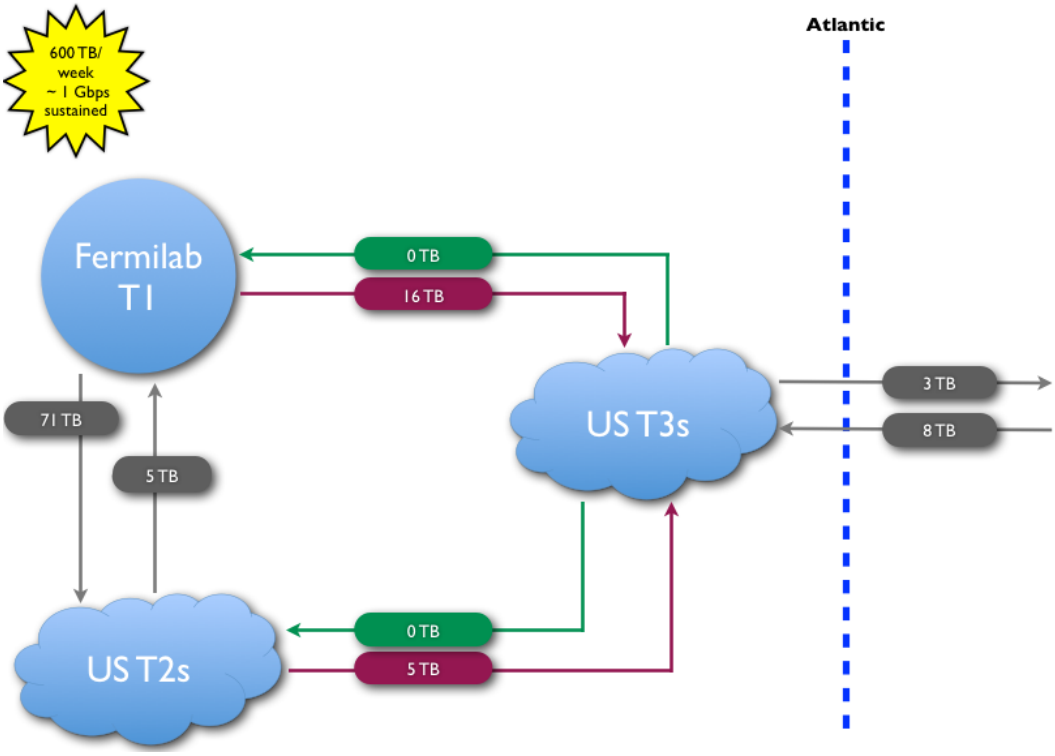


Figure 24. Total transfer volume for transfers to and from U.S. Tier-3 sites for one week in 2012.

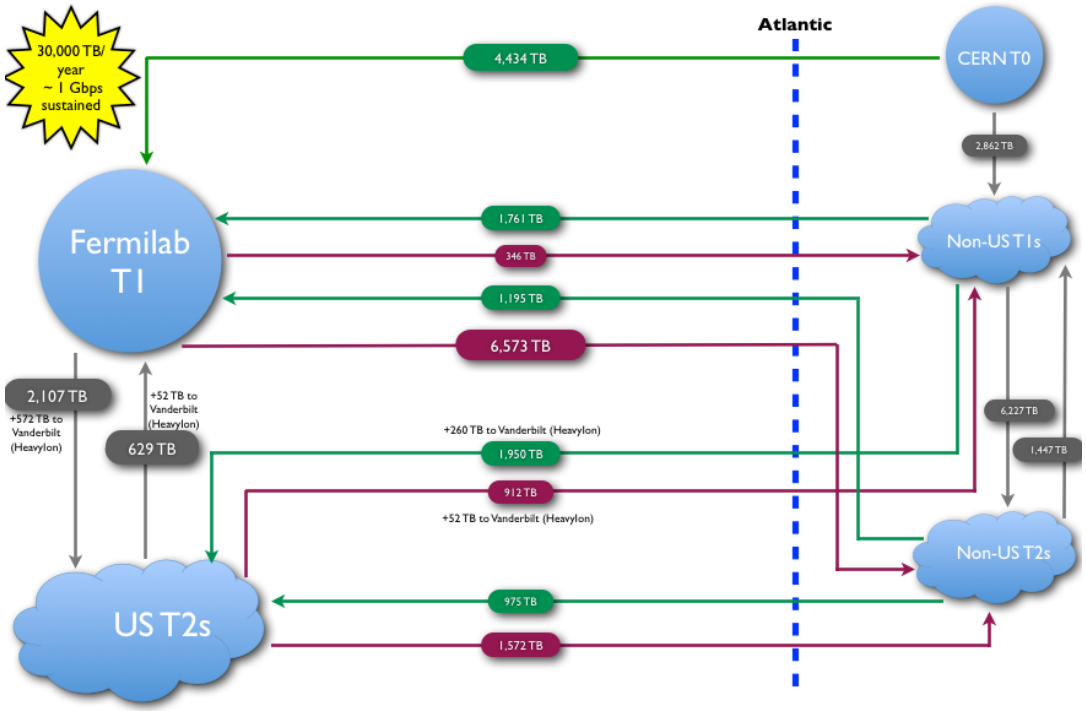


Figure 25. Total transfer volume for main transfer streams in 2012.

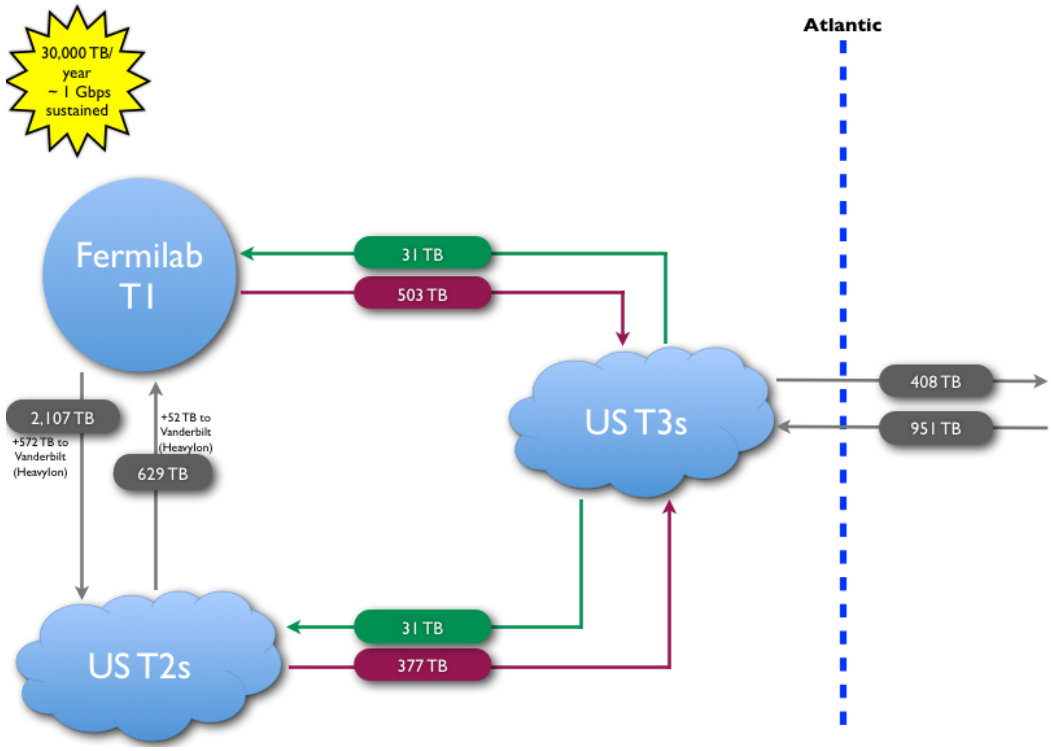


Figure 26. Total transfer volume to and from U.S. Tier-3 sites in 2012.

8.5 Local Science Drivers — the Next 2–5 Years

For LHC Run 2, several key parameters will change:

- Energy will go up from 8 TeV to about 13 TeV, leading to increased event complexity, larger event sizes and longer processing times
  - Smaller effect: bunch spacing will be reduced from 50 ns to 25 ns; lower PileUp per event; smaller event sizes and shorter processing times
- Trigger rate will go up from 300–400 Hz to 1 kHz
  - Although the last two months in 2012 already saw such high trigger rates

Summary:

- Expect event processing times to increase → more resources needed to process and analyze
- Expect trigger to select more relevant events for analysis → analysis datasets grow, which have to be transferred to the Tier-2 sites for analysis

8.5.1 Instruments and Facilities

Apart from sites changing their access interface from Grid to cloud-based technologies, we expect that more and more Tier-3 sites will be set up without any disk organized in an MSS. These diskless Tier-3 sites access data and MC files transparently through the WAN. The technology used is based on the XrootD protocol and was implemented for CMS by the AAA project. A first example is the Tier-3 at the University of Notre Dame. It has a capacity of about 800 job slots and during ongoing analysis sustains about 200 MB/sec input rate sustained over 24 hours while reading files through the WAN (see Figures 27 and 28).

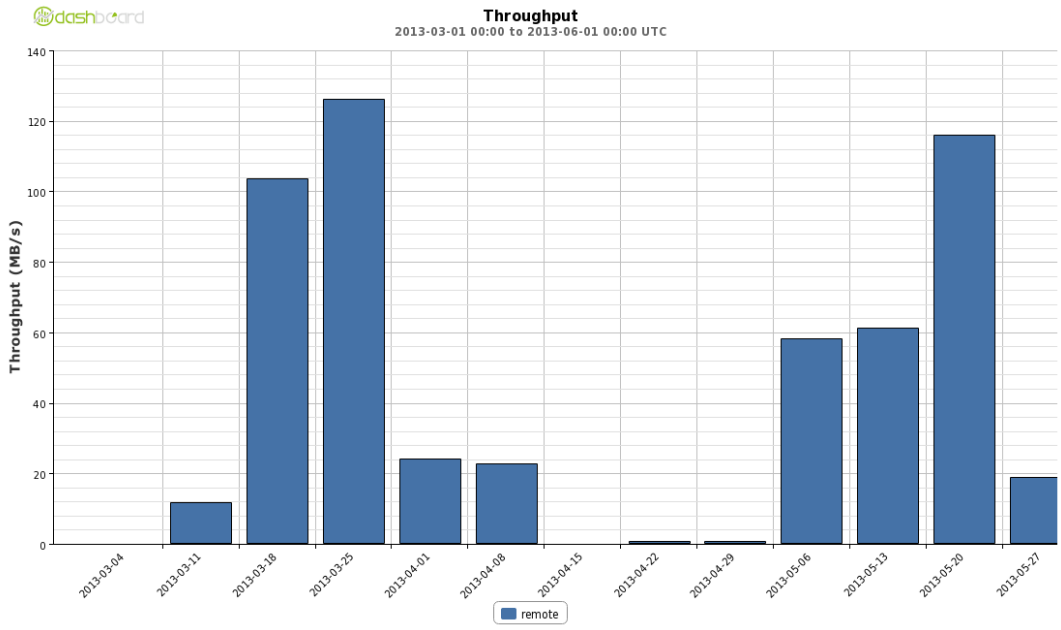


Figure 27. Network throughput to the Tier-3 at the University of Notre Dame during three months in 2013.

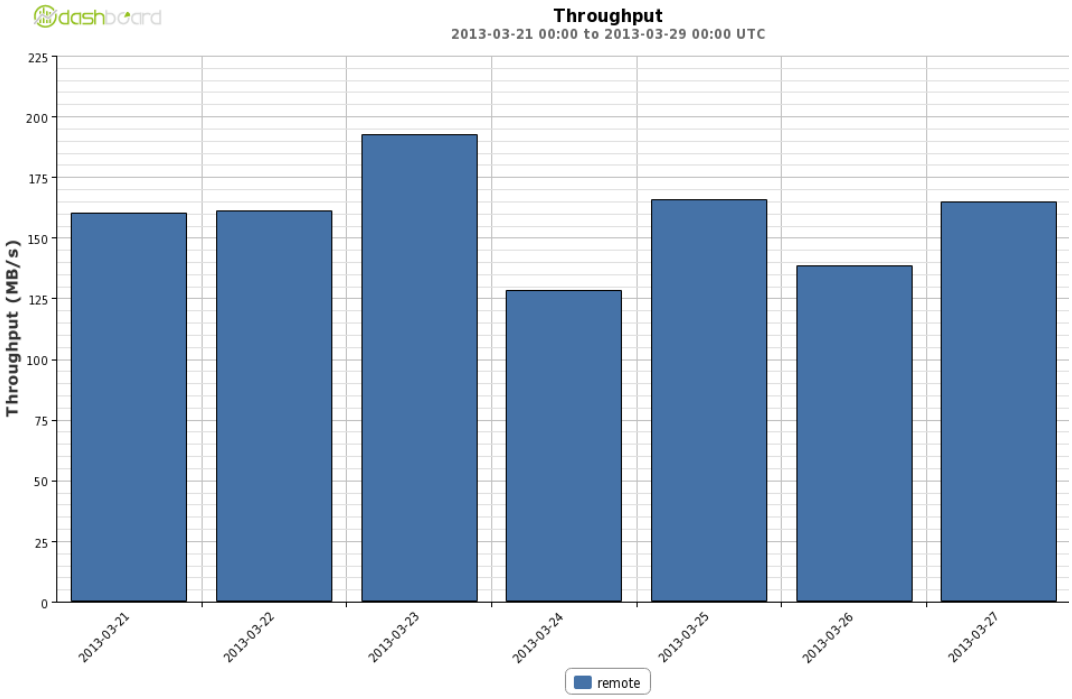


Figure 28. Network throughput to the Tier-3 at the University of Notre Fame during one week in 2013.

The campus research traffic is normally 500 MB/sec but when the Tier-3 site is analyzing data, it increases to 2–3 GB/sec.

8.5.2 Software Infrastructure

Every CMS site will publish its stored files in the CMS data federation. The data federation allows access to all published files through the WAN transparently — independent of the location. The federation is based on XrootD; every site will bring up XrootD servers that provide access and stream files. The site can restrict the total outbound bandwidth of the XrootD servers to protect the local storage. This requires good network connectivity down to the smallest sites.

8.5.3 Process of Science

There is no fundamental change planned for the described science process. Only the CMS data federation will be used to optimize resource usage and use resources also independent of data locality. This optimization is planned for all major workflows outside Tier-0.

8.6 Remote Science Drivers — the Next 2–5 Years

8.6.1 Instruments and Facilities

The CPU resources needed to process all data and simulate all MC will be significantly higher in LHC Run 2. We expect that requested resource increases would have to be

augmented with access to high-performance supercomputing centers and other types of opportunistic resources like commercial clouds from Amazon and Google when they become financially viable options. The mode of operation will also change, as allocations on these resources most probably cannot be made for the whole year but rather for a limited time period. This requires either fast stage-in of datasets or access through the CMS data federation. Also, other Grid sites not primarily for CMS will have to be used opportunistically. In any case, good network connectivity to these resources is required.

A good example is the usage of the San Diego Supercomputer Center (SDSC) in 2013, where we processed large datasets for specific SuperSymmetry (SUSY) analyses. About 1.7 million core hours were used in four weeks to reprocess 400 million proton-proton collisions. 150 TB of input RAW data was transferred to SDSC and access was provided through the local Data Oasis storage system. This collaboration enabled the CMS to complete the processing of these datasets needed for SUSY analyses several months ahead of the original schedule (http://www.sdsc.edu/News%20Items/PR040413_lhc.html).

8.6.2 Software Infrastructure

For LHC Run 2, the CMS will optimize the use of the available disk space at Tier-1 and Tier-2 sites. The strict request-based dataset distribution model will be augmented to take popularity information for datasets into account. The system will automatically create more replicas of a dataset and analysis jobs will be rerouted based on a dataset's popularity with regard to analysis jobs in the queue wanting to access that dataset. At the same time, the system will be able to release the cache of less popular or unpopular samples automatically to make room for more popular samples. This enhanced mode of operation is not expected to have a big effect on network load. In the first place, unpopular samples will not be distributed at all, which saves bandwidth. This is compensated by samples that are more often replicated, creating more bandwidth usage. An optimal operations point will have to be found.

As discussed previously, the CMS data federation plays a prominent role in CMS plans to optimize resource usage and increase flexibility for production and processing. The CMS data federation is based on the AAA project: "Any data, Any time, Anywhere" with XrootD at its core (see http://xrootd.slac.stanford.edu/). All CMS files on disk at CMS sites will be accessible through XrootD. Applications can open files directly via the WAN; the CMS event contents have been optimized for WAN access. The WAN access itself introduces minimal additional latency. As an example, processing RAW data through AAA produces 50 KB/sec per application and higher data rates for MC and other workflows if more data is read through the WAN. AAA is decoupling the application from the location of the data. A redirector setup resolves the request to open a file (see Figures 29 and 30) and redirects the file open request to the XrootD servers of a site providing access to the requested file, and at the same time also provides load balancing.

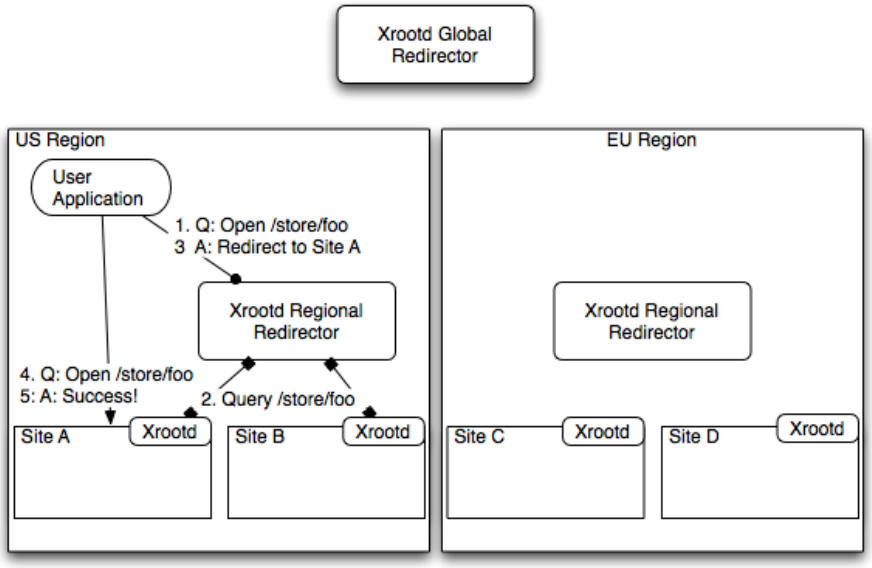


Figure 29. Local redirector setup for AAA.

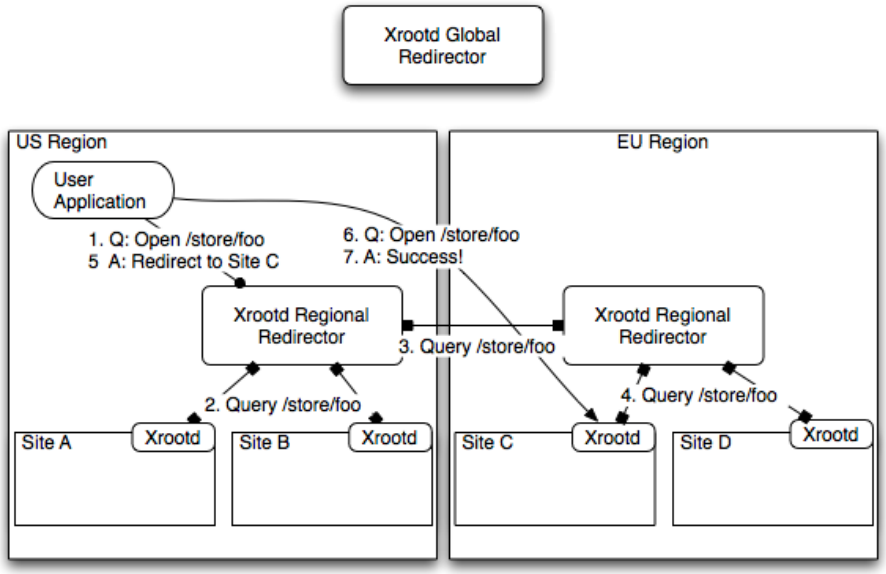


Figure 30. Cross-region redirector for AAA.

All CMS sites will be configured for “fallback” mode. Every application running locally at the site that tries to read a file that cannot be found locally at the site is automatically falling back through XrootD and WAN access. AAA has a different data access pattern than organized PhEDEx sample placements, which are done in burst mode. AAA distributes the network load much more evenly and is expected to fill the gaps in the network usage shown by example in Section 8.4.3.

8.6.3 Process of Science

The change in LHC Run 2 parameters will cause analysis dataset sizes to grow as more interesting events are selected among all the background events. Larger analysis datasets will cause the data transfer volume to go up. The increase is not expected to be exponential. Figure 31 shows the increase in data volume during LHC Run 1.

There was a small increase in total transfer volume between 2010 and 2011 but the transfer volume had the same center-of-mass energy. In 2011, data replaced MC on the analysis level, which contributed to the small increase. From 2011 to 2012, the center-of-mass energy increased (larger cross section of physics processes) and the trigger rate increased, resulting in larger datasets to be analyzed and transferred to the Tier-2 level. In summary, LHC Run 1 saw an increase in transfers but the effect is not exponential. For LHC Run 2, we estimate a conservative increase in the total transfer volume by a factor of 2–5, including the effect of the CMS federation filling the gaps in between the bursts of organized PhEDEx transfers.

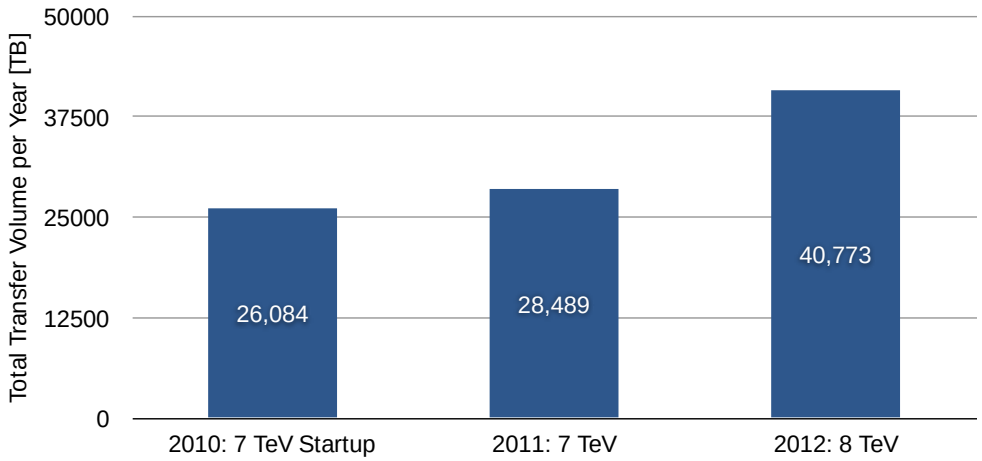


Figure 31. Transfer volume per year for LHC Run 1.

8.7 Beyond 5 Years — Future Needs and Scientific Direction

8.8 Network and Data Architecture

Apart from the already mentioned good network connectivity of all CMS sites down to the Tier-3 level — which is required for the use of the CMS data federation and good connectivity to supercomputing centers, non-CMS Grid resources, and cloud providers — the reliability and performance of the whole network infrastructure is extremely important. Monitoring of the infrastructure plays a key role. The U.S. CMS network monitoring strategy is based on using network monitoring tools on different levels of the network infrastructure:

- Application level: PhEDEx
- Site/fabric level: FTS monitoring DashBoard
- Network level: perfSONAR

Here perfSONAR is a very good tool and its support should be maintained.

If CMS detects low-level network problems or performance reductions, we normally contact networking experts at Fermilab who also uses perfSONAR to determine how well the network infrastructure is performing by looking at information on throughput/ latency for network paths between perfSONAR hosts. CMS is working with the OSG networking group to establish a complete perfSONAR mesh that helps to track and visualize the network throughput/latency metrics for all sites. By doing so, many network problems can be detected before application-level end users become aware. Thus network infrastructure monitoring is necessary to supplying high throughput and reliable network connections especially at the LAN. Experience shows, however, that sometimes this is not always sufficient for solving the majority of WAN performance problems.

PerfSONAR monitoring does not extend to the end system(s), and often stops at the border between the site and the WAN providers (last-mile perfSONAR monitoring gaps remain). To solve these problems, the Fermilab team adopted an end-to-end focus, from the network up through the application level, looking at packet traces. This gives a true end-to-end picture of what's happening with a particular application. The ESnet Fasterdata Knowledge Base could help to optimize WAN data movement, but the knowledgebase serves moreso as guidance than an integrated debugging/troubleshooting tool.

8.9 Collaboration tools

CMS uses Vidyo for videoconferencing collaboration-wide. All meetings have Vidyo conferences attached in their Indico agendas.

8.10 Data, Workflow, Middleware Tools, and Services

Summary of LHC Run 2 expectations:

- Bigger analysis datasets (higher cross section means a higher number of selected physics events)
  - Analysts have to look at more events that are more complicated
  - More selected events result in more transfers to Tier-2 sites for analysis
  - Burst mode in data placement will remain or even increase slightly
- WAN access through the CMS data federation will fill in gaps between transfer bursts
- Resource requirements for processing will increase (complexity of events will increase, which means longer processing times)
  - An increase in resource diversity will play a big role (the CMS needs to use HPC installations like NERSC; clouds like Amazon, Google, etc.; university campus computing centers) to augment CMS-owned resources
  - Example: Resources on university clusters get allocated for two months to CMS; CMS must be able to use it effectively; also if no disk space is available on site, data needs to be accessed/transferred through the WAN

Trends:

- No huge demand in more peak performance around 10 Gbps
- Factor of 2–5 increase in sustained rates
- In general, more reliance on network reliability and robustness to more and diverse sites, which will result in more monitoring, troubleshooting, etc.
- Trans-Atlantic component of traffic remains very important

8.11 Outstanding Issues

N/A

8.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• The LHC brings protons to collisions inside the CMS detector. • The CMS detector records the events, filters them in the trigger system, and stores the RAW information. • The CMS distributed computing infrastructure of more than 100 sites stores, processes, and analyzes the collisions to extract physics results.	• Proton-proton collision events are recorded with the detector and reconstructed and stored on the distributed computing infrastructure. • MC events are simulated, reconstructed, and stored on the distributed computing infrastructure. • Analysis data formats are transferred to the Tier-2 level where they are accessed by analysis applications.	• In 2012, 13.5 PB MC and 9.4 PB data was stored. • In 2012, a total of more than 40 PB was transferred between the CMS sites.	• LAN is used within individual sites to access files on the MSS using local optimized file access protocols. • Worker nodes have Gigabit connectivity, which is slowly upgrading to 10 GbE.	• CMS is placing datasets centrally or per requests of physics groups at Tier-2 sites for analysis. • These organized transfers are showing a burst-like behavior.

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
2–5 years				
• LHC Run 2 will see an increase in center-of-mass energy and trigger rate, which will primarily result in larger datasets to be analyzed. • The increase in center-of-mass energy will also result in larger CPU resource demands. CMS will have to augment its resources by using supercomputer centers, non-CMS Grid resources, and cloud providers.	• New developments and systems will help increase the efficiency of the CMS operation. • Dynamic data placement and automatic cache release will optimize disk usage but will have no net effect on the transfer volume. • CMS data federation will allow location-independent access to CMS files on disk at one of the sites. • New access mode will produce a more even usage of the network and is expected to fill in the gaps between the bursts of organized transfers.	• Conservative estimates are an increase of the total transfer volume by a factor of 2–4. • No huge demand in more peak performance.	Excellent connectivity down to the smallest Tier-3 sites, super computer centers, and cloud providers is required.	Use of CMS data federation will fill in gaps between organized transfer bursts around 10 Gpbs.
5+ years				

9 CMS-HI Research Program

9.1 Background

The CMS is one of four experiments taking data produced by the LHC accelerator located at the CERN laboratory near Geneva, Switzerland. For the most part, the LHC operates in the proton-proton (pp) collision mode designed to explore the frontiers of high energy physics. The announcement on July 4, 2012, that a Higgs-like boson particle had been discovered is one of the spectacular successes of this HEP research at the LHC. While several months of the year have been devoted to pp physics, the LHC is also capable of colliding heavy ion (HI) nuclei such as lead-on-lead (PbPb), or even asymmetric collisions such as protons-on-lead (pPb). In each of the past three years, the HI running has taken place for a period of about 24 days after the proton running has completed.

The goal of the HI research program is to create and study the properties of a novel state of matter called the quark gluon plasma (QGP). This is a state of matter predicted to be created in HI collisions where the produced temperatures and densities are so large that the normal nuclear matter constituents of protons and neutrons melt into their composite quarks and gluons. This phase is thought to be the state of matter persisting in the early universe until a few microseconds after the Big Bang. One of the early surprises in this field of high energy nuclear physics was the discovery that the QGP behaves like a strongly interacting liquid with accompanying hydrodynamic behavior, such as flow, and not like a weakly interacting partonic gas, as had been expected. This discovery was made by colliding gold nuclei (AuAu collisions) at the Relativistic Heavy Ion Collider (RHIC at BNL).

The LHC HI research program also analyzes pp and pPb collisions as so-called reference data for the more complex PbPb collisions. Originally it had been assumed that the QGP would not be formed in the simpler systems, and that QGP effects in the PbPb data could be disentangled from normal nuclear matter effects by scaling up from the pp and pPb measurements. However, close analyses of the pp and the pPb data taken in 2012 and 2013 lead to a further discovery that there is a highly correlated pattern in some of the produced particles (“the ridge”) reminiscent of the flow-like behavior seen in the RHIC AuAu and the LHC PbPb collisions. There are recent speculations that QGP “droplets” may even be found in the pp or pPb collisions observed at the LHC. Clearly, the field is still in its early stages of understanding these surprising collision phenomena.

After the completion of the special HI run of reference pp collisions in February 2013, the LHC went into its scheduled Long Shutdown 1 (LS1) upgrade phase. During the LS1, the accelerator’s dipole magnets are to be modified to handle a doubling of the collision energy. Simultaneously, the experiment’s collaborations are to upgrade their detector systems to handle anticipated large increases in beam luminosities and data volumes expected when collisions resume in early 2015. These increases will be in both the pp and the HI programs. In particular, the CMS-HI detector physics group is proposing a series of upgrades during LS1 to be able to accommodate the enhanced capabilities of the CMS detector as of 2015. The three-year history of the CMS-HI data taking at the LHC

(see Table 12) serves as a basis for extrapolating the computing and network needs in the next few years corresponding to the scope of this case study.

Table 12 lists the collision systems with their respective center-of-mass collision energies per nucleon. The raw collision rates for producing data events are given, followed by the event rates passing a first-level set of trigger cuts, and then the event rates passing a high-level trigger (HLT) set of cuts. Essentially, the volume output event stream from the HLT, which is calculated as the HLT rate multiplied by the event data size for the particular collision system, must be constrained to fit within the various data-handling capacities downstream of the detector. These constraints include the Tier-0 prompt reconstruction limit, the file transfer limits to the remote site national Tier-1 facilities, and their tape archiving quotas.

Table 12. HI collision systems and data rates for the CMS during 2010-2013.

Dates	System	Maximum Collision Rate	Max Level 1 (L1)	HLT Output	Integrated Luminosity
Nov/Dec 2010	PbPb 2.76 TeV	200 Hz	200 Hz	120 Hz	7 μb^{-1}
Nov/Dec 2011	PbPb 2.76 TeV	4.5 kHz	2.7 kHz	200 Hz	150 μb^{-1}
Jan/Feb 2013	pPb, Pbp 5.02 TeV	260 kHz	60 kHz	1 kHz	31 nb^{-1}
Feb 2013	pp 2.76 TeV	3 MHz	90 kHz	1.2 kHz	5.4 pb^{-1}

In 2010, all data were taken in so-called minimum bias mode, meaning the collision events on average were less complex than a smaller fraction of more interesting, complex events. The volume of Raw data generated from the detector was 150 TB, for which there were 190 TB of prompt reconstruction (Reco) files subsequently produced at the Tier-0 computer facility at CERN. For the 2011 data taking, when there was a huge jump in the heavy collision luminosity — as shown in the collision rate column of Table 12 — there was a total of 915 TB of Raw and Reco files produced, with the majority of those files (628 TB) in selected trigger mode. The intrinsically much less voluminous pPb and pp running in 2013 produced 313 TB.

In 2015 after LS1, the LHC will run PbPb collisions at 5 TeV with at least an 8 kHz collision rate rising perhaps to a maximum rate of 20–30 kHz. The HI event sizes will naturally be larger at the higher energies for the two to three PbPb runs contemplated in 2015-2017. In 2018 there will be a scheduled Long Shutdown 2 (LS2). After LS2, the beam collision rate could be at 50 kHz. There is an approved HI upgrade project for the L1 to handle at least the post-LS1 rates. The HLT will be configured in 2015 and 2016 to produce on the order of 50% more HI data than was produced in 2011.

In addition to the Raw and Reco data files, the HI program requires simulation file production. Unlike the HEP program, in which the number of simulated events may be

comparable with the number of data events, the HI physics analyses generally require a significantly smaller fraction of events, perhaps a few million simulations compared with tens of millions of data events. The HI program is not a “golden” events program to discover new types of particles, for which the possible background production magnitudes must be precisely known. Instead, the HI analyses typically concentrate on well-known particles, often looking at spatial correlations among them (jets, flow), or detecting rare probes (J/Psi, Upsilon, Z-boson) whose interaction or lack thereof with the QGP phase can provide special physics insights.

9.2 Collaborators

Of the approximately 2,500 physicists in the CMS collaboration, about 120 physicists from nine countries have a primary focus in the HI program. The majority of these scientists are based in U.S. institutions: the University of Kansas, University of Maryland, MIT, Purdue, Rice, Rutgers, University of California Riverside, University of Illinois at Chicago, and Vanderbilt. There are also important overseas CMS HI groups in France, Korea, Hungary, New Zealand, and Russia.

The computing system for the HI group in CMS conforms as much as possible to the overall computing system in CMS. The HI group in CMS is far too small to support any nonstandard capabilities in the transport, storage, or processing of the data. The HI Reco files for user data analysis are primarily stored at the Vanderbilt Tier-2 facility. Users access these files via a Grid-based computing interface called the CMS Remote Analysis Builder (CRAB). The CRAB system recognizes, via the CMS database component of the data files transfer system (PhEDEx, see Section 9.4.2), where particular data files are located. Users’ job configurations are typically assembled at their local computing facilities or at special gateway facilities at Fermilab or at CERN. There is no special need for a user to have an account at any of the Tier-2 sites. These job configurations are then directed to the remote Tier-2 data hosting facilities for processing, where the accompanying user’s Grid certificate is recognized for authorized data access. Upon completion of the jobs, their analysis output files can be returned to the user’s home computer facility or transferred to some Tier-2 facility where the user has an allowed storage quota. All CMS-HI users have an allowed storage quota at the Vanderbilt Tier-2, and the quota can be managed with remote access tools.

Besides the main Vanderbilt Tier-2 sites, there are other CMS-HI Tier-2 facilities or components at MIT and in France and Russia. These sites can accommodate smaller subsets of the data, called prompt skims. These are files containing the most important events for an analysis and are produced at the Vanderbilt Tier-2 soon after the main Reco files are available during data acquisition.

9.3 Key Local Science Drivers

9.3.1 Instruments and Facilities

In addition to the obvious LHC accelerator facility and the CMS detector instrument, the key components for this research program are the computational systems. These start

with data acquisition (DAQ), which feeds the central Tier-0 computational resource where the data are initially reconstructed; then the high-speed network systems by which these data files are transported to national Tier-1 facilities for tape storage and eventual re-reconstruction; and finally the Tier-2 satellites, which are associated with a particular Tier-1 by their own high-speed networks. Originally the system was quite hierarchical, with the Tier-2 generally supporting a particular set of physics files. More recently, as network speeds have increased between the Tier-2 sites or even between a Tier-2 and certain Tier-3 sites, the AAA model has arisen. Files are opened and streamed over the WAN to the user's job application via the XrootD servers that are implementing this AAA model.

The components, specifications, and missions for a Tier-1 or a Tier-2 are well defined in CMS. There will be a CE consisting of a set of worker nodes. There will be an SE managing the data files on disk, which are accessed by the worker nodes via a high-speed local network infrastructure (10 Gbps in the case of the Vanderbilt Tier-2). In addition, the tier-site must be visible to the rest of CMS for GridFTP activity, including for file transfers and CRAB job submissions.

The HI Vanderbilt Tier-2 is somewhat unique in CMS computing in that it also does some Tier-1-like activities although not the archival tape storage of the data. The CMS-HI group relies on the tape facilities of the Fermilab Tier-1 for that service. On the other hand, the Vanderbilt Tier-2 does perform re-reconstruction of the data, a mission normally done by the Tier-1 facilities in CMS. Similarly, during the HI data-acquisition periods, the Vanderbilt Tier-2 is involved in the rigorously regulated prompt reconstruction transfer cycle (see Section 9.3.3), acquiring data files at high speed from both the Fermilab Tier-1 and the CERN Tier-0. When the prompt reconstruction files arrive at Vanderbilt, they are processed for prompt skim production and output back to tape at Fermilab and to disk storage at other CMS Tier-2 sites.

The MIT Tier-2 also plays a special role in the CMS-HI program. Although this facility is largely supported by the NSF-HEP program, it also receives support from DOE-NP. The particular role of the HI component at the MIT Tier-2 is to provide simulation production support to the HI group, as well as to provide another significant CRAB jobs analysis resource.

9.3.2 Software Infrastructure

The CMS software environment is completely enveloped in a framework called CMSSW, for which a prescribed schedule of updates and releases is well established. These releases are typically associated with the chronology of the data taking. Sites like the Vanderbilt Tier-2 automatically subscribe to these releases by the CVMFS read-only file mechanism. Database, calibration, and alignment information are also centrally supported and transparently accessed from the user's viewpoint.

9.3.3 Process of Science

Raw data originate at the detector and are transported by the DAQ system to the central Tier-0 facility. A 48-hour delay is allowed for alignment and calibration information to be

developed from special-purpose files taken at the same time. After the alignment and calibration information is available, another 48 hours are allocated to do the prompt reconstruction of the Raw files. When the prompt reconstruction completes, the Raw and Reco files are transferred to the particular national Tier-1 sites that have previously subscribed to certain subsets of the data. No one Tier-1 site obtains all the datasets for the pp program. There are some 12 hours assigned for the transport of the files to the Tier-1 sites, where these files are placed on tape storage for archiving. Of course there are multiday capacity data storage buffers at the Tier-0 that are intended to copy with transient outages in the external network systems, as described below. The central Tier-0 facility also contains a copy of the Raw files on tape for emergency use. After the Reco files are at the Tier-1 site, they are processed for prompt skimming and distribution to Tier-2 sites. When the archival step at the Tier-1 site is completed, then these same files may be released from the disk buffer areas at the Tier-0 to make room for newly arriving Raw data.

The workflow for the steps of calibration/alignment, prompt reconstruction, transfer, tape archival, and prompt skim is naturally automated and tightly controlled. The steps must proceed in a continuous assembly-line fashion to prevent saturation of the disk buffers at the Tier-0 site. Such saturation could result in a retarding of the DAQ. In this respect, the transfers to the Vanderbilt Tier-2 during periods of HI DAQ must be carefully monitored.

An example of this monitoring is shown in Figure 32, which depicts the file-transfer rates over a 5.5-day period ending on February 13, 2013. As it happens, there was a transient transfer software problem at the CERN Tier-0 site during the week-end of February 9-10,

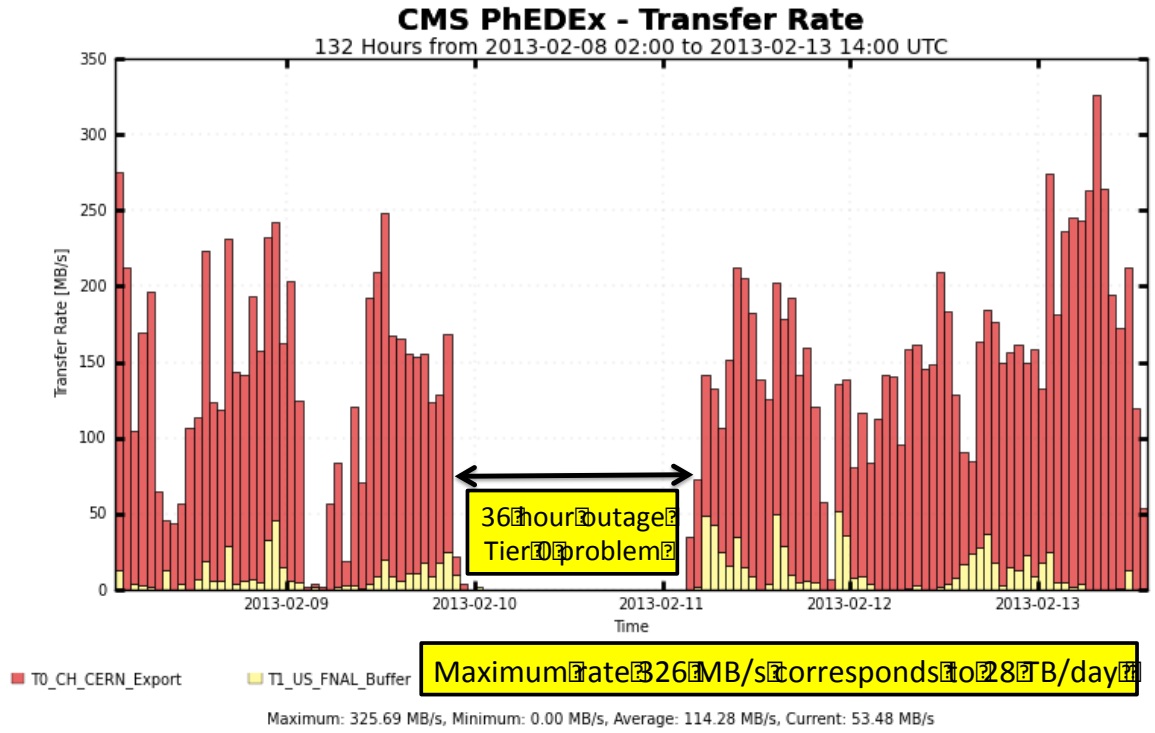


Figure 32. Monitoring of the data-transfer rate to the Vanderbilt Tier-2, Feb. 13, 2013.

which was not resolved until the following Monday. This problem caused a shutdown of the file transfers during a 36-hour period. Fortunately, this outage (the only one during the 2013 data taking, and none happened in 2010 or 2011) had absolutely no impact on the data production schedule, as the weekend also corresponded to a time when the LHC was changing from pPb running to pp running. Two days after the interruption, the backlog of data transfers was completely eliminated as the transfer speeds surged to 28 TB/day, i.e., almost 10% of the total data volume produced during the 24 days of the 2013 HI run in CMS.

9.4 Key Remote Science Drivers

9.4.1 Instruments and Facilities

The CMS experiment has a widespread distribution of collaborators and computing resources that are tightly coupled and highly uniform from a data processing perspective. These features are largely true also for the CMS-HI research program. The variations specific to the CMS-HI operations, such as in some of the extra missions of the Vanderbilt Tier-2 or the tape archival storage of the data, can be attributed to the relatively small sizes of the HI group itself and the HI data volumes as compared with the corresponding HEP sizes.

9.4.2 Software Infrastructure

The global file transfer system used throughout the CMS experiment is called PhEDEx (Physics Event Data Export). This is an extremely sophisticated system using the GridFTP mechanism that automatically optimizes the transfer routes. It is a critically important component during the periods of data acquisition as noted in Section 9.3.3.

The PhEDEx system has excellent monitoring and historical retrieval services. Figure 33 shows an example of the data transfers into the Vanderbilt Tier-2. The figure emphasizes the episodic nature of data transfers related to that Tier-2. Prior to the HI data taking period in early 2013, there were two periods of intense activity in the fall of 2012. In October 2012, some 120 TB of muon physics data files were retrieved from their tape storage location at the French national Tier-1 site. This set of data had to be re-reconstructed urgently at the Vanderbilt Tier-2. Similarly, in December 2012 another set of file transfers came from Fermilab tape storage to accommodate a set of three different physics re-reconstructions at Vanderbilt. In early 2013, a spike in data activity was associated with the pPb and pp new file transfers, most of which were chosen by PhEDEx to come directly from CERN. The next six months in 2013 were relatively quiet.

The outbound activity from the Vanderbilt Tier-2 is even more episodic, as show in Figure 34. There was a modest amount of activity in November 2012, which was the output of the muon re-reconstruction going back to the HI Tier-2 in France (T2_FR_GRIF_LLR). In January–February 2013, modest outbound transfers were associated with the prompt skim data production workflow at Vanderbilt during the LHC data taking. A more intense spike followed in March–April 2013, related to yet another physics dataset re-reconstruction.

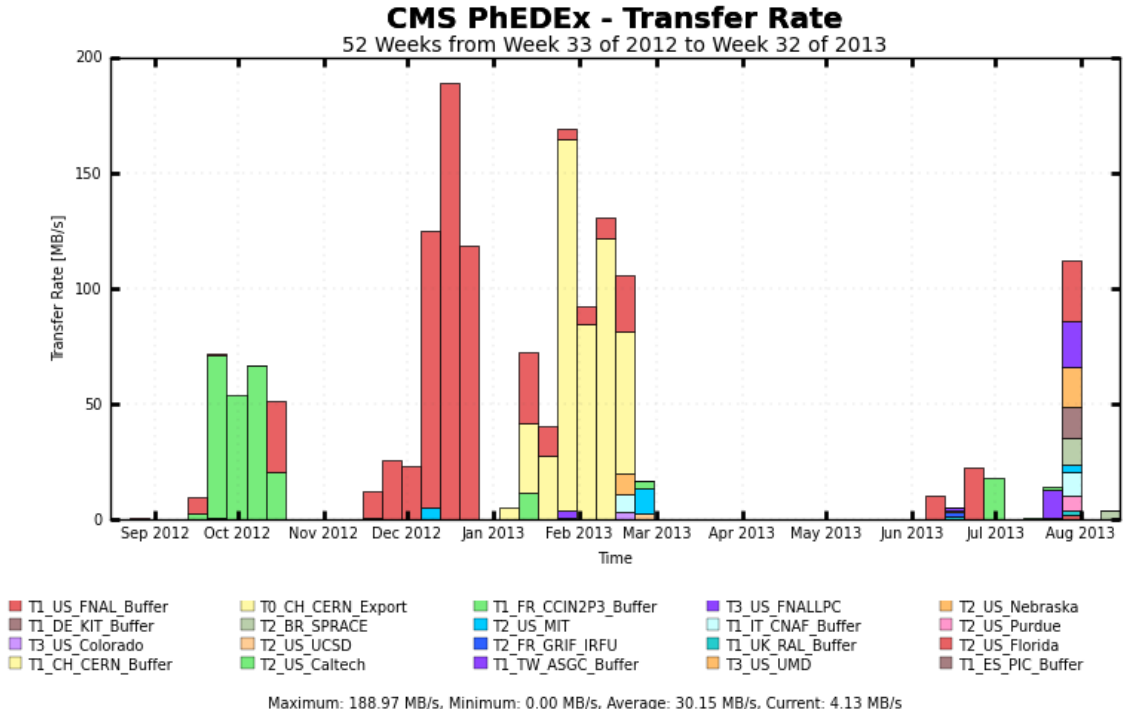


Figure 33. One-year record of PhEDEx transfers into the Vanderbilt Tier-2, as of August 10, 2013.

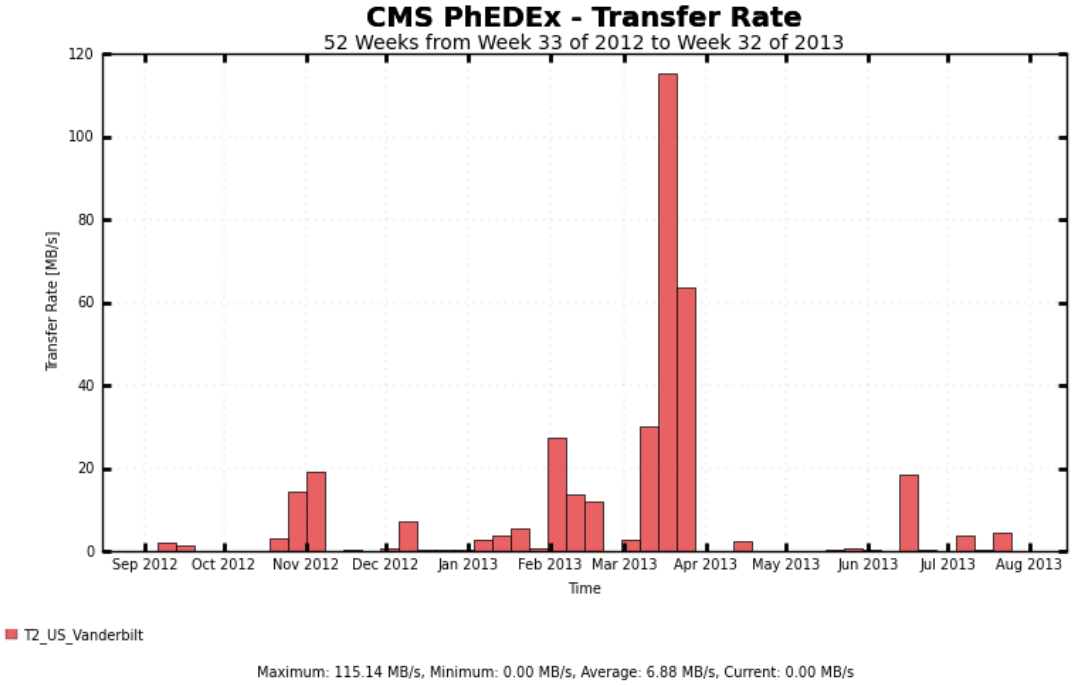
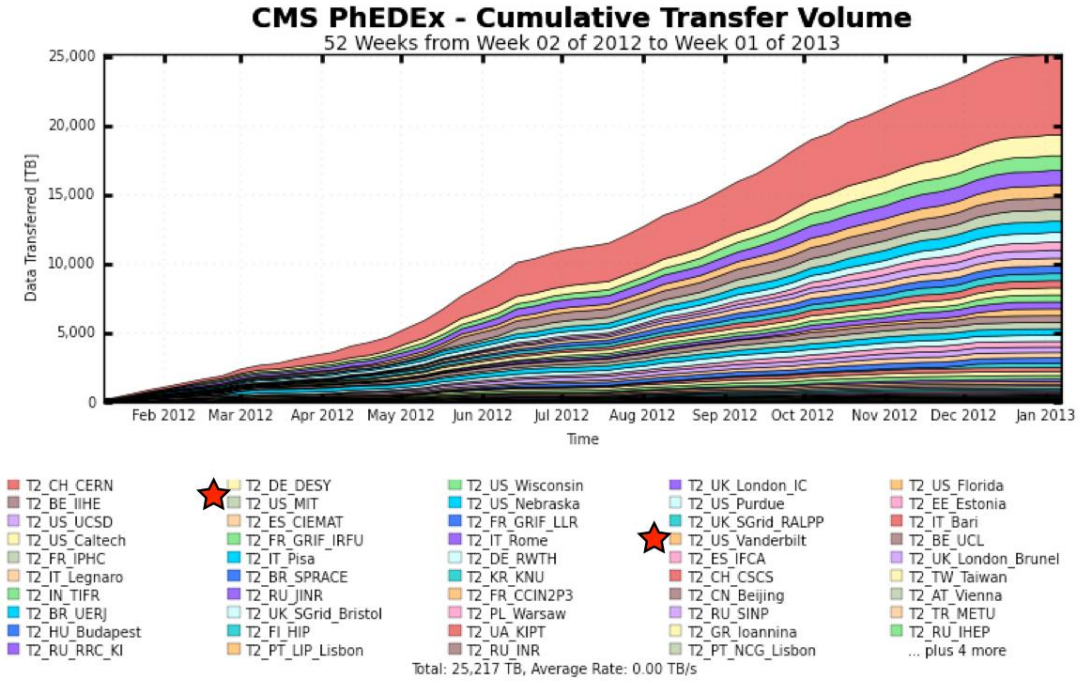


Figure 34. One-year record of PhEDEx transfers into the Vanderbilt Tier-2, as of August 10, 2013.

These last two sets of figures confirm the highly variable nature of the network activity into and out of the Vanderbilt HI site. This cycle of activity would seem well suited to dynamic network capacity allocation systems.

For the MIT Tier-2 site, equivalent figures could be generated but these would be dominated by HEP activity rather than HI activity. Some measure of the HI network activity at MIT compared to Vanderbilt can be inferred from the relative sizes of the mass storage systems at the two sites. The Vanderbilt site currently has about 2 PB of HI data files, whereas the MIT site has about 500 TB of HI files.

The relative comparison of inbound traffic among all CMS Tier-2 sites during calendar 2012 is shown in Figure 35. The total amount for the Vanderbilt Tier-2 is given as 491 TB. As it happens, calendar 2012 was a "down" year for the HI data taking. All the data from the 2011 run had been transferred as of December 2011. The data taking for the pPb and pp 2013 runs did not start until January 2013. Hence, what is shown for calendar 2012 is a baseline of non-LHC running transfer volume, such as for re-reconstruction purposes or dedicated network performance testing.



850 TB into MIT, 491 TB into Vanderbilt

Figure 35. Record of cumulative transfers into all CMS Tier-2 facilities during 2012.

9.4.3 Process of Science

As noted above, the individual CMS-HI users employ the CRAB system to submit data analysis jobs to remote sites. The output from these jobs can be elevated to global use ("published"), effectively creating secondary datasets. These published outputs can become accessible to both the PhEDEx and the CRAB systems upon being validated for

registration with the CMS StoreResults service. This registration will enable global access to the new files by other users and at other Tier-2 sites.

9.5 Local Science Drivers — the Next 2–5 Years

9.5.1 Instruments and Facilities

Coming out of the LS1 in 2015, the LHC is expected to provide nearly double the collision energy to 5 TeV for PbPb collisions, compared with the 2010–2011 runs. The beam intensity will also go up, perhaps giving a factor of 4–5 in collision rate. More powerful L1 and HLT capabilities will be essential to cope with the increased data-acquisition capacity of the detector. The event volumes will go up at least 50%, and it will be computationally more demanding to process these higher energy collision events.

The central Tier-0 computing resource must be substantially scaled up in any case to handle the more complicated pp datasets as of 2015. In fact, some of the prompt Reco may have to be offloaded to Tier-1 sites, as in some scenarios even the expanded Tier-0 resource will lack enough processing power. The scaling factors for the pp data processing should be compatible with the increased HI prompt reconstruction processing demands. The main HI computing facilities at MIT and Vanderbilt will have their worker-node and disk storage capacities increased to keep pace with this increased data load. The Vanderbilt Tier-2 facility will double its worker node size to 2,000 by 2016 and increase its usable storage amount to 3 PB.

It is also expected that the Tier-2/Tier-3 distinction will increasingly blur as the AAA model gains prominence. Because all the U.S. HI members are already partners with their HEP groups in the same institution, the HI people at these Tier-3 sites will be able to process the Tier-2 stored HI data files over the WAN just as effectively as their HEP colleagues. It is a matter of upgrading the network capabilities into the Tier-3 sites in order to make this “diskless” mode of data analysis feasible.

9.5.2 Software Infrastructure

In the CMS, the XrootD implementation of the AAA model of data analysis jobs is coming into being as of 2013. This new model is integrated into the complete CMS data federation to optimize the matching between more distributed new computing resources and the existing more centralized data locations. Experience with this model will enable the Tier-2 sites to estimate their new outbound bandwidth demands and the impact on their local storage systems. Obviously, limits to local storage access must to be respected.

9.5.3 Process of Science

The process of science for the CMS-HI should remain fundamentally the same for the next 2–5 years, apart from an anticipated more decentralized system of computing resources to process analysis jobs.

9.6 Remote Science Drivers — the Next 2–5 Years

9.6.1 Instruments and Facilities

The CMS-HI group will track with the rest of the CMS collaboration in its possible use of commercial cloud computing resources such as Amazon or Google. Because this science is so data intensive, some means of getting those commercial resources to access the large quantities of physics data files should be implemented. Since the annual HI datasets are naturally about an order of magnitude smaller than the pp datasets, with a correspondingly smaller number of HI analyzers, the CMS-HI research subset might be a useful test case for this new mode of analysis.

9.6.2 Software Infrastructure

Besides the XrootD-AAA model of analysis job processing, the most significant change in software infrastructure for the CMS-HI group may come in the "popularity" monitoring tool for the datasets analyzed by various users. Even with the anticipated 3 PB of storage capacity for the HI program at Vanderbilt in 2016 and several hundreds of terabytes at other HI Tier-2 sites, some triage decisions of what datasets to retain on disk and which to purge must be made by 2015. Some of the older datasets from 2010 or 2011 should be replaced to allow for storage of the 2015 and 2016 datasets. This will be a physics-driven process in the next 18 months, to decide which datasets have become less valuable.

9.6.3 Process of Science

As mentioned above, after the LS1 completes in early 2015, the LHC is expected to deliver more intense HI beams at higher collision energies, resulting in a significant increase in data volumes compared with what they were in 2011. On the other hand, clear financial constraints limit how much storage space can be funded in the near term. A conservative estimate is that the HI data volumes will grow by a factor of 2 for the 2015 and 2106 running periods, after which the LS2 starts and no new data acquisition may be expected for another 2 years at least. The 2011 PbPb run saw 685 TB transferred to Fermilab in a one-month period of time. A factor-of-2 increase would mean about 1.4 PB of HI data transferred from the Tier-0 to the Fermilab Tier-1 in one month, meaning an average transfer rate of 4.3 Gbps or daily average 47 TB. The same rate, or half at least, would have to be achieved for transferring the data files to Vanderbilt, at least the Reco files for local prompt skim production. The transfer of the Raw files could be deferred until they were needed for a re-reconstruction pass.

As such, these estimates foresee a need for an incremental increase in the inbound network performance in the next 2–5 years, as far as HI data transport is concerned. On the other hand, the outbound network capacity at the Vanderbilt Tier-2, which so far has not been stressed, is likely to increase as the AAA-XrootD model takes hold and more Tier-3-like computing resources come online for the HI program.

The network outlook can be summarized by saying that the best present sustained performance during dedicated testing periods is something over 4 Gbps inbound, with a

nominal 10 Gbps capacity. In 2 years, that nominal capacity should grow to 20 Gbps for the Vanderbilt Tier-2 to keep pace with the influx of new HI data from CERN. Five years from now, the capacity will have to grow to 100 MB/sec. This was the conclusion of a recently completed external review of the Vanderbilt Tier-2 conducted by DOE-NP. The outbound forecast is less well-determined since that will depend on the use patterns of the HI group in the next 2–5 years, but planning for 100 Gbps network capacity is the prudent course. The local jobs capacity at the Vanderbilt or MIT Tier-2 facilities is constrained by the number of installed worker nodes. If the HI group adopts a wider use of Tier-3 facilities for data analysis, taking advantage of the AAA/XrootD model, the outbound network traffic will grow to the limits of the local disk storage system. Those limits are still to be refined for the Vanderbilt Tier-2 site, as the current load of running CRAB jobs has not yet tested them.

9.7 Beyond 5 Years — Future Needs and Scientific Direction

The range beyond 5 years comes after the LS2 completion, and becomes more speculative regarding the HI program. At a minimum, the LHC will be able to deliver an order-of-magnitude more collision rate compared with the 2011 experience. The L1 and HLT systems would have to be upgraded in major ways in order to select the most interesting events to record in a reasonable amount of data volume, say a few PB during the typical 3–4 week HI running period

9.8 Network and Data Architecture

As described above, the inbound network requirements for the HI Tier-2 at Vanderbilt are highly episodic and predictable. During the periods of HI data collisions at the LHC, the Vanderbilt Tier-2 acts almost like a national Tier-1 in its mission to acquire the prompt reconstruction data from CERN as rapidly as possible. This high-performance period will generally last for 4 weeks at most in a calendar year. During other months of the year, the inbound demands will be less stringent unless a re-reconstruction pass is needed and Raw data files must to be procured from a Tier-1 site such as at Fermilab. Again, these re-reconstruction periods of time are predictable several weeks at least in advance, and typically will last only a few weeks themselves.

These predictable, episodic periods of intense bandwidth use may be compatible with ESNet's On-Demand Secure Circuits and Advance Reservation System (OSCARS), assuming that a multiweek (or even a multiday for re-reconstruction tasks) duration of intense bandwidth use is feasible. As it happens, Vanderbilt University is a member of the NSF-funded ANSE (Advanced Network Services for Experiments) project for the HEP community. This project is actively looking at the use of dynamically allocated circuits for the needs of experiments like the CMS. According to one of the Vanderbilt members, the underlying software to access dynamic circuit technology is not yet reliable enough and more effort is needed to establish this functionality as a production tool.

9.9 Collaboration Tools

In 2012–2013, the LHC experiments changed to the Vidyo videoconferencing system, having used the EVO (Enabling Virtual Organizations) videoconferencing system until then. All major CMS meetings among collaborating institutions are conducted now with the Vidyo system.

9.10 Data, Workflow, Middleware Tools, and Services

OSG middleware provides a crucial interface for the operations of Tier-2 facilities such as Vanderbilt. The Vanderbilt technical staff relies on this package and the response to the staff's questions, especially during upgrades of the OSG components.

9.11 Outstanding Issues

For the Vanderbilt site, the network connection with the Fermilab Tier-1 site is the most important. There have been dedicated time periods when the achievable network bandwidth has been actively investigated, especially in the month or two prior to a major run at the LHC, to confirm that the performance during the run will be successful. Generally, these tests have gone well, although the maximum sustained rate has never been above 4 Gbps even though there should be 7 Gbps allowed into the Tier-2 site. Getting the "last mile" of performance out of this Fermilab–Vanderbilt connection has proved elusive. Such future dedicated performance tests are planned during the LS1 and these tests will need to use whatever tools are available to make a full end-to-end scan to find network bottlenecks.

9.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• The LHC and CMS are the principal instruments, along with a large ensemble of tiered computing facilities. • The CMSSW is a self-contained software framework that is centrally supported and is used by all collaborating institutions.	• Data are acquired by the detector and a first-pass analysis is made at the central Tier-0 compute site. These analysis Reco files are transported to Tier-1 and Tier-2 sites for storage and for user analysis. The prompt-Reco files are re-passed with new reconstruction software leading to improved analyses. • The HI program takes data for about 3-4 weeks per year with the majority of the Reco files stored at a single Tier-2 at Vanderbilt. • HI MC events are generated but on a reduced scale compared with the HEP program. The MIT HI computing site has the main MC responsibility.	• In the 2011 PbPb LHC run, some 915 TB of data were initially produced. • In the 2013 pPb run, 300 TB of LHC data were produced. • Individual file sizes range from a few to 10 GB. • Tens of thousands of files in different datasets must be managed.	LAN storage systems see multi-Gbps performance.	• During LHC running periods, as much as 1 PB/month should be moved from CERN to the U.S. • Collaborating sites can be transferring tens of terabytes of datasets or more over one or a few weeks.

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
2–5 years				
• Both the LHC and the CMS detector are expected to nearly double the respective capacities as of 2015, when LHC operations resume. • The increased collision energies for the HI events will produce more complex events at larger event sizes. In turn, analysis of these more complex events will place increased demands on the computing time and memory requirements of the analysis jobs.	With the use of the AAA/XrootD model of data processing, the science analysis will make more optimum use of the available computing resources. This mode of operation will rely on a more robust WAN among all the U.S. CMS institutions, and not just a selected set of Tier-2 sites that are hosting the data.	Annual dataset sizes may increase by a factor of two for the HI program in CMS.	The future data analysis model has more Tier-3 participation, contingent on having fast enough WAN speeds at those sites. A minimum of 20 Gbps is seen for 2 years hence, and up to 100 Gbps has been recommended at the 5-year limit.	• The future data analysis model assumes that the data can be streamed rapidly to non-Tier-2 sites that do not have massive storage systems. • The number of existing Tier-2 sites will not be expanded, but their internal resources will grow as the annual data volumes increase.
5+ years				
After two years of new operations in 2015 and 2016, the LHC will go into its LS2 phase and not resume until past 2018.	Details of the HI science program at the LHC post-LS2 are just coming under discussion.			

10 The ALICE Experiment

10.1 Background

The ALICE (A Large Ion Collider Experiment) collaboration has constructed and operates a heavy ion detector to exploit the unique physics potential of proton-proton and nucleus-nucleus interactions at LHC energies. The principal goal of the experiment is to study the physics of strongly interacting quark gluon plasma (QGP), a novel phase of matter produced at extreme energy densities. The study is carried out with measurements from PbPb, pPb, and pp collisions at the LHC.

To extract the most physics information from the measurements, ALICE, like all of the LHC experiments, requires the collection and processing of an unprecedented amount of experiment data. The LHC experiments have adopted a distributed computing model for the processing, analysis, and archiving of data organized within the Worldwide LHC Computing Grid (WLCG) collaboration. For ALICE, all participating countries are expected to contribute CPU, disk, and mass storage within the sponsoring country in proportion to its Ph.D. participation. These resources are made available to ALICE by their connection to the ALICE Grid facility discussed below. The initial ALICE-USA obligations corresponded to about 6% of all ALICE computing resource needs and are currently about 7% of those requirements. The ALICE-USA computing project was proposed in 2008 to deliver those required resources by deploying two U.S. ALICE Grid sites located within two computing centers: Livermore Computing (Lawrence Livermore National Laboratory [LLNL/LC]) and NERSC/PDSF (Parallel Distributed Systems Facility). The proposal was formally approved by DOE-SC/NP in early 2010 and the two facilities have been fully operating within the ALICE Grid facility since summer of 2010.

The ALICE detector operates in conjunction with the running schedule of the LHC at CERN, taking data during pp and PbPb (or pPb) collision periods each year. The broad LHC schedule consists of multiyear periods of operation separated by long shutdown periods for maintenance and upgrades to the collider and experiments. After three consecutive years of operation from 2010 into early 2013, the LHC is currently in its first shutdown period, LS1. LS1 is expected to last for about two years, with the LHC due to resume operations in 2015 for the Run 2 period. As such, current “steady-state” operations of the ALICE Grid facility as they exist now during LS1 will be used to provide network requirements for the coming 2 years, while changes expected from Run 2 will be reflected in the network requirements in the next 2–5 years. The LHC has scheduled a second shutdown, LS2, starting in about 2018, in which both the collider and ALICE detector are expected to make some dramatic changes that affect ALICE data taking capabilities. These changes will be discussed broadly in the final section covering future needs beyond 5 years.

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, to be 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. In 2011, for example, event collections of about 1.2 billion pp events and 150 million PbPb events corresponded to about 3 PB and 2 PB of raw data, respectively. In addition to the raw data, a comparable volume of MC simulation data, used to evaluate measurement efficiencies and systematic uncertainties, is required with each dataset produced and stored on the ALICE Grid facility.

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 datasets 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), is not only used directly in some analysis tasks but is 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 scientist or subgroups of physicists working on common analyses make use of these refined data for their specific analysis tasks. Throughout the processing steps, the data retains its 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 for event processing to be distributed over a large number of independent compute resources.

ALICE's distributed computing is carried out on the ALICE Grid facility that is characterized by a tiered set of sites composed of a single Tier-0 center at CERN for primary data storage and initial processing, several Tier-1 centers providing additional processing and both tape and disk storage capacities, and many smaller Tier-2 centers with CPU and disk storage capacities. Raw event data is stored at the single Tier-0 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, 8 Tier-1 and about 70 Tier-2 centers distributed about the world. The Tier-1 facilities are relied upon for: (1) long-term custodial storage of a copy of the 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 MC simulation data in quantities comparable to the real event data, and (6) running end-user analysis tasks. The Tier-2 facilities provide the same functions as the Tier-1 facilities except for (1) and (2) above: long-term custodial storage of data, and additional reconstruction passes. About 90% of the processing on Tier-1 and Tier-2 sites is devoted to analysis or MC simulation and, as such there is little distinction between Tier-1 and Tier-2 facilities for the general work carried out on the ALICE Grid facility. In practice, however, sites with

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

large storage, like all Tier-1 and many larger Tier-2 sites, are likely to accommodate more data-intensive tasks like analysis, while sites with more CPU than disk are more likely to do MC simulations. All of this is predicated on the amount and priority of the different tasks that are queued.

At each step in the process, data is replicated. Processed real data is copied from the Tier-0 and Tier-1 centers to the Tier-2 centers while MC results are replicated and copied to and from Tier-1 and Tier-2 centers. Multiple copies of ESD and AOD files are generated automatically at processing time and distributed to Grid-enabled Storage Elements (SEs) in the ALICE Grid facility. The data 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 replicas of the reduced or produced data. A record of each copy is stored in the ALICE File Catalog, a global single-instance catalog of all ALICE data files. As a result, all data is available at multiple sites for further analysis.

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] A task is broken up into many identical sub-jobs, each requiring a subset of the data. The individual sub-jobs are executed through a process in which the participating Grid sites pull work from the central task queue. That is, AliEn services at each site monitor both the local resources and the pending jobs on the central task queue, pulling jobs from the queue when local resources are available to meet the job requirements. The infrastructure strongly favors running jobs on sites where required input data exists locally. However, when priority dictates and CPU resources exist without local access to data, jobs will be run on sites on which data is accessed dynamically over the WAN.

To minimize contention for accessing data, a significant amount of ALICE analysis jobs are organized into “trains” — a collection of many different analysis tasks that require the same input data, combined into a single task. Thus, instead of each analysis independently reading the same input data from disk, the data is read once for the entire train, reducing I/O cost and increasing CPU efficiency for the data processing. While users are not required to run their individual analyses within a larger train, they are encouraged to do so by references to ease of operation (train operators run the jobs) and faster turnaround (higher priority) in their analyses.

Although all data processing and analysis can be run on the ALICE Grid, scientists often can make good use of more direct processing in which an analysis task is run repeatedly over a fixed subset of data. Such a processing mode provides fast turnaround times, allowing scientists to efficiently refine their analyses. For this type of work, ALICE

---

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

supports several independent ALICE Analysis Facilities[1] (AAFs) where subsets of data are staged to disk for access by many users and analyzed in parallel via Parallel ROOT Facility, PROOF.[2] The staging process requires pulling data from the distributed Grid-based SEs to the AAF, and can be characterized by relatively large but sporadic data transfers over short time intervals from distributed sources to an individual facility.

The ALICE data processing described above is summarized in Table 13.

Table 13. General types of processing carried out by ALICE scientists.

Processing	Activity	Location	Input & Source	Output & Destination
Raw data reconstruction	Organized	T0 & T1	Raw data experiment or archive	ESD/AOD files ALICE Grid SE
MC simulation + reconstruction	Organized	T1 & T2	Configuration data	Simulated data + ESD/ AOD files ALICE Grid SE
Analysis trains	Organized	T1 & T2	ESD/AOD (usually) local SE	User output ALICE Grid SE
User analysis on the Grid	Chaotic	T1 & T2	ESD/AOD (usually) local SE	User output ALICE Grid SE
Non-Grid user analysis	Chaotic	AAF	ESD/AOD AAF XRootD system	User output AAF Storage

10.2 Collaborators

The worldwide ALICE virtual organization (VO) is for use by ALICE scientists interacting with Grid organizations such as the WLCG and the OSG in the United States. The registry of members, including information on roles with respect to computing activities, is maintained in the Virtual Organization Management and Registration Service (VOMRS) by the WLCG at CERN for ALICE.[3] The VO manager, Latchezar Betev (CERN), is also in charge of Grid operations for the ALICE Grid facility. Several hundred ALICE scientists are registered with the ALICE VO as are many ALICE users of the ALICE Grid facility.

The ALICE computing project is led by Predrag Buncic (CERN), with the Management Board oversight lead by Yves Schutz (CERN, University of Nantes, France). A Computing

[1] http://aaf.cern.ch/

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

[3] https://lcg-voms.cern.ch:8443/vo/alice/vomrs

Board[4] meets monthly to receive updates and provide feedback to the project. Each major detector system has a representative on the computing board as does each participating country (or funding agency). The ALICE-USA computing project is a DOE-funded project with Jeff Porter (LBNL) as the project lead and active member of the ALICE Computing Board. Operations at the ALICE-USA sites are coordinated by a steering committee that also meets monthly and consists of the project leader, ALICE representatives from each of the two sites (R. Soltz, LLNL, J.Porter), a system administrator (J. Cunningham, LLNL, and I. Sakrejda, LBNL), and a support person (L. Gerhardt, LBNL), as well as an at-large ALICE-USA collaborator (B.S. Nilsen from Creighton University).

10.3 Key Local Science Drivers

10.3.1 Instruments and Facilities

ALICE-USA participation from a compute-facility perspective is concentrated at two Tier-2 centers at the NERSC/PDSF facility at LBNL and the LC facility at LLNL. The two facilities are comparable to each other in size: 1,000 CPU cores (>10 kHS06) and 0.7 PB of disk space. Both sites are integrated into the ALICE Grid and accessed via the AliEn Grid framework.

The two U.S. facilities combine to represent about 7% of all of ALICE Grid computing resources in terms of both processing capacity and disk space. A Grid-enabled SE exists at each site and both have several modest size (50–70 TB) file servers, each with 10 GbE connection to the facility core router and then to ESnet directly (NERSC/PDSF) or via additional routers (LLNL/LC [Livermore Computing]). The compute nodes at each facility are primarily 1 GbE attached, though about half the PDSF compute nodes have InfiniBand (IB) interconnects and are accessible via IP over IB protocol.

While the two facilities are of comparable size, there are important differences. The LLNL/LC site is a pure ALICE Grid site in which user log-ins are not allowed and no other groups can access the resources except under specific arrangements. The NERSC/PDSF site does allow direct login for registered users and supports the client software for job submission or data access. NERSC/PDSF is not, however, operated as an ALICE AAF. NERSC/PDSF is also a multigroup facility supporting non-ALICE LBNL HEP/NP research groups, primarily STAR, ATLAS, and Daya Bay. While multigroup access has little impact on resource utilization by ALICE, it does make it difficult to disentangle network use specific to ALICE. Therefore, for the purposes of this report, network use measurements from LLNL/LC will be used as representative of either facility.

Finally, the NERSC facility also includes tape storage via an allocation on the NERSC HPSS system on which an annual growth of several hundred terabytes of storage can be reserved for ALICE. That tape storage capacity could allow NERSC to become a Tier-1 center for ALICE, a goal written in the original ALICE-USA proposal but deferred, perhaps

[4] http://aliceinfo.cern.ch/Management/Boards/Computing/members.html

indefinitely. The prospect of NERSC becoming a Tier-1 center for Run 2 will be considered when evaluating network requirements beyond the next 2 years.

10.3.2 Software Infrastructure

Each facility is configured as an ALICE Grid site with ALICE middleware services run on the site. Specifically, an ALICE VO box is an independent machine with network access to the ALICE central services managed at CERN and direct access to the local compute cluster and SE. Services on the VO box collect monitoring information about the cluster and local SE and are responsible for such tasks as job submission and software installation onto the cluster, performed as needed to support processing on the site.

The multiple file servers that make up an ALICE SE are integrated into a single facility-wide SE using XRootD software. Each facility has an XRootD manager (redirector) to which each server connects and is registered. Each redirector, with its ALICE SE name, is registered with AliEn central services. For the U.S. sites, the two SEs are ALICE::LBL::SE and ALICE::LLNL::SE. Each file copied into an SE is registered with the ALICE File Catalog with the associated SE for later access.

10.3.3 Process of Science

A significant amount of processing carried out within the scientific investigation is done within an organized production model, as listed in Table 13. For example, once the raw data is taken and detector calibrations are determined, a reconstruction pass is done over the data managed by the central team to produced data files that can be used by individual physicists. Similar production processes are carried out for MC simulations. The ALICE Grid facility, however, is constructed to allow all users to perform their analysis tasks directly on the Grid facility. Individual scientists can submit tasks to the Grid or within an analysis train as if the Grid were a monolithic cluster. Those tasks analyze the data generated during the production processing and produce further-refined data that can be accessed directly by individual scientists for final inspection and interpretation.

In addition to running full analysis passes over a dataset, a scientist will typically refine an analysis with runs over smaller but repeatable subsets of the entire dataset, yielding short turnaround times. This type of workflow is more optimally run on an AAF, which pairs a reasonable number of processors with dedicated disk space for staging datasets for common use. Jobs are submitted directly to and processed on the AAF. In particular, ALICE computing only supports AAFs based on PROOF (Parallel ROOT Facility) clusters for implementing job parallelism, which by design includes XRootD installations for data staging and I/O.

As noted in Table 13, ALICE jobs that run on the U.S. Tier-2 sites are either MC simulation jobs that do not require input data, or analysis jobs (either organized trains or individual user tasks) that require input data that exists on the local SE. Thus the bulk of the job I/O is done between the compute cluster and the local SE, either storing MC output or reading analysis input. Network traffic on the local site SE also includes data copied to it by remote processing during data replication as is discussed in the next section.

Although there are no AAFs in the United States, the NERSC/PDSF facility does allow direct log-ins by local 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 from the Grid to local resources for more managed access. In addition, U.S. scientists have organized local analyses using other NERSC resources for this type of work. It is reasonable to assume AAF-type use will grow in the future, perhaps even with dedicated resources.

10.4 Key Remote Science Drivers

10.4.1 Instruments and Facilities

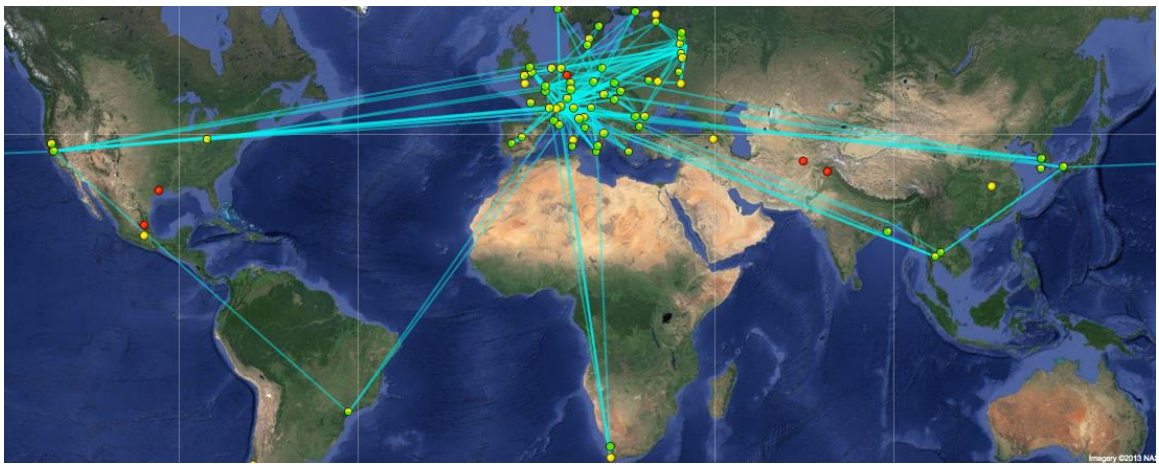
The ALICE experiment and its Tier-0 facility are located at CERN and, in fact, most of the ALICE Grid resources are located in Europe. This includes all current ALICE Tier-1 sites, with each Tier-1 site supporting a number of Tier-2 sites in their respective countries and nearby regions. That cluster of resources in Europe is illustrated in the map shown in Figure 36; the ALICE Grid facility and its operations have been developed in this environment.

The two ALICE-USA sites at LLNL/LC and NERSC/PDSF are operated as Tier-2 centers and, as such, do not participate in processing of raw data. The raw data reconstruction passes at the Tier-0 and Tier-1 sites noted in Table 13 produce ESD and AOD files, which are replicated for distribution on the ALICE Grid. Historically, three copies of each file are distributed in a algorithmic fashion: one file sent to the SE local to the processing, one file to the nearest SE as measured by network tests, and one file randomly to any SE monitored as ready to accept data. This same formula is used for the produced MC simulation files, which are generated at both Tier-1 and Tier-2 sites. Thus, the wide area data distribution mode for the ALICE-USA sites: (1) receive a fraction of ALICE ESD and AOD data files produced in Europe, (2) receive MC simulation files produced at Tier-2 sites, and (3) send copies of MC simulation files produced locally to other sites, including U.S. sites.

As can be seen in the map in Figure 36, a third U.S. ALICE Grid site is operated at the Ohio Supercomputer Center (OSC). That site was established prior to the ALICE-USA computing project and delivers resources in support of ALICE collaborators at Ohio State University. ALICE has access to a CPU capacity of about 200 cores at OSC but no disk storage due to an incompatibility between the disk technologies at OSC and the ALICE XRootD storage infrastructure. Thus, without local input data, OSC runs mostly MC simulation jobs, which produce data typically stored on one of the Tier-2 SEs at LLNL/LC or NERSC/PDSF. A modest shift in the concentration of ALICE resources from Europe began in 2010 with new projects in the United States, South Korea, and Mexico. The two U.S. facilities became operational at the end of 2010, with a steady ramp-up of resources to their present capacities noted above. A new ALICE Tier-1 facility has recently been commissioned at the KISTI center in South Korea and a large ALICE Tier-2 has been approved at UNAM in Mexico City, though its deployment schedule is uncertain. For each of these new facilities, the data transfer path from CERN and other European centers goes through the United States, directly to the U.S. facilities or to South Korea or Mexico

via international links. The impact on U.S. sites from the newly created Tier-1 site at KISTI was already observed when a reconstruction pass done recently at KISTI produced a noticeable increase in rate of data transfers into the U.S. SE by about 40 MB/sec.

Figure 36. A worldwide map of ALICE Grid sites. Dots represent sites and lines represent a snapshot of dynamic data transfers between sites occurring at the time the image was queried. The image illustrates the large density of sites located in Europe.



10.4.2 Software Infrastructure

As noted in Section 10.3.2, each center is configured to be a site in the ALICE Grid facility by running ALICE middleware services that interact with the local cluster and SE, the central services at CERN, and other ALICE Grid sites. Information about each site is maintained in a central LDAP repository, which is accessed by VO box services to navigate site-specific configurations, such as local batch system details or file system definitions for managing log and cache files. The Grid manager at CERN, under guidance from local contacts, maintains that information in the LDAP repository.

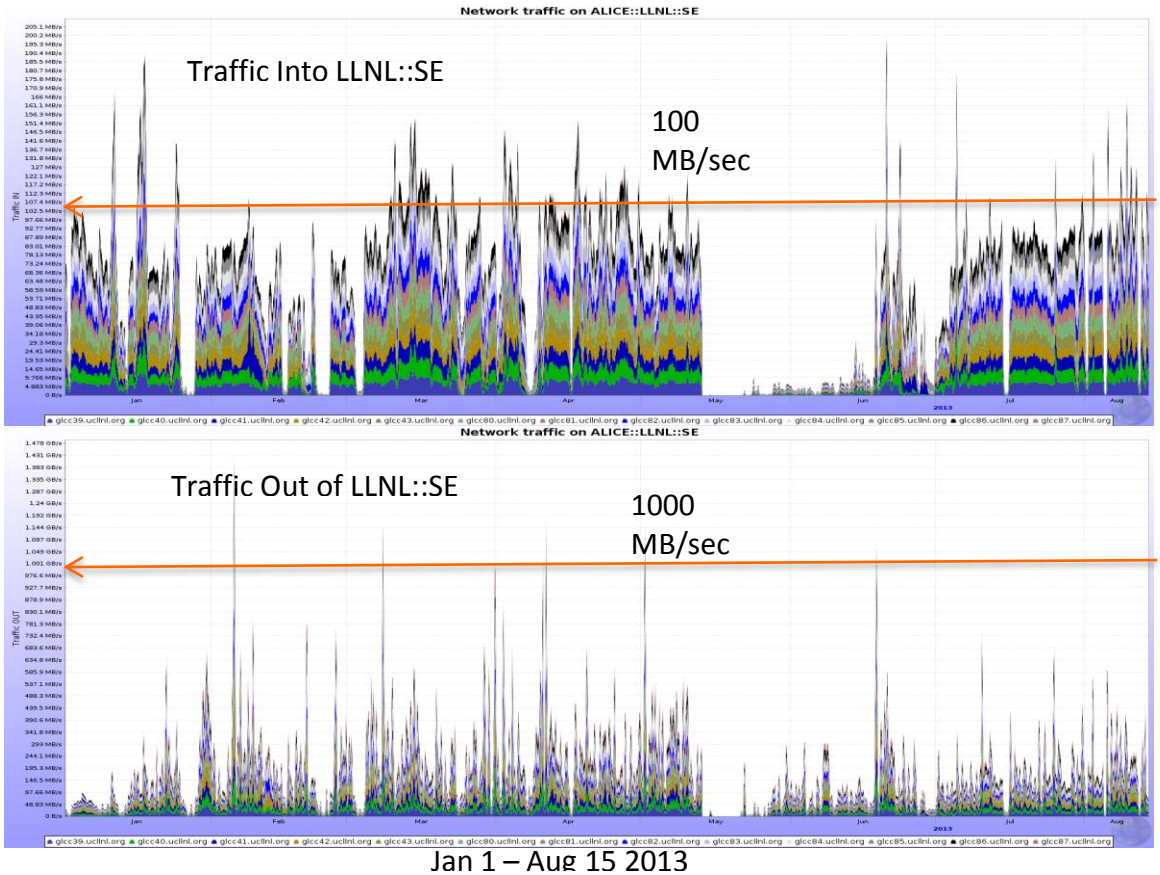
In addition to operational details for active processing on the cluster, the project is required to meet pledged goals for providing resources to ALICE. For this reason, the WLCG records information about each site, such as resource utilization (CPU and disk) and site availability and reliability. For U.S. sites, that information is gathered by site participation in the OSG. OSG middleware is run on each site to gather accounting information for assessing ALICE utilization of site resources in terms of CPU and wall time, weighted by processing capacity (HS06). Additional OSG probes monitor specific critical functions to determine availability and reliability measures at each site. These data are forwarded to the WLCG, which provides ALICE with independent assessment of performance of each site relative to its pledged contributions.

The ALICE computing project expects that the Grid software infrastructure will undergo modest evolution in coming years. For example, application software now deployed either by the ALICE PackMan service or an ALICE BitTorrent implementation will be

replaced by using CVMFS[5] in the coming year. However, the basic services that make up the AliEn Grid infrastructure are expected to function as they exist now for the next couple of years.

10.4.3 Process of Science

Current processing and data transfers at the U.S. facilities are like those at any ALICE Tier-2 site. The SE 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. While some occasional data access is done directly between a running job on the cluster and an external source, the majority of network interactions are done via the local SE. A plot of network traffic into and out of the LLNL::SE is shown in Figure 37 over the period from January 1 to August 15, 2013, which indicates data transfer rates as values averaged over several-hour-long periods. The top panel shows the nominal data rates going into the SE at about 100 MB/sec, while the bottom plot indicates that outgoing traffic averages about 400 MB/sec. This asymmetry is consistent with the expected use of the SE. Most of the data going into the SE is from local or nearby MC simulations, which have relatively modest output rates per CPU hour.



Jan 1 – Aug 15 2013

[5] https://twiki.cern.ch/twiki/bin/view/CvmFS

Figure 37. Traffic monitored into and out of the LLNL::SE since January 2013. Plot automatically averages data over several hours for such a long plot ranges, suppressing short-term large rate spikes. Shorter time interval plots indicate that peak rates can be larger by a factor of 10. Nominal rates of 100 MB/sec into the SE and 500 MB/sec out are consistent with data flow as described in the text.

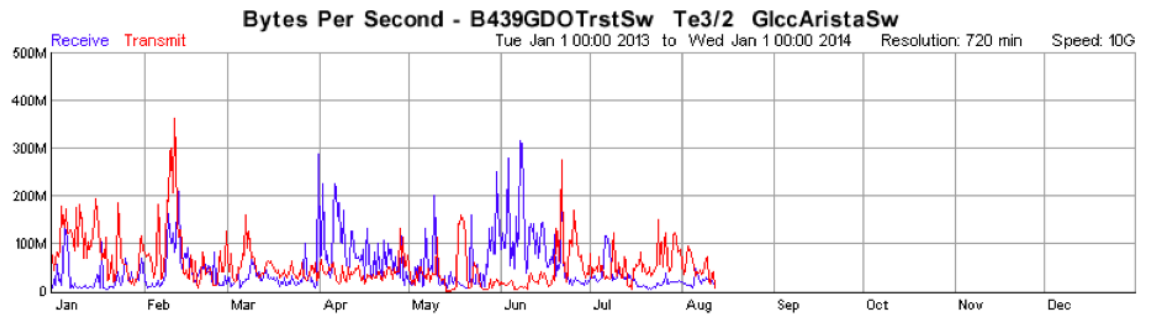


Figure 38. Traffic monitored at the LLNL/LC switch connected to the ALICE cluster showing nominal rates of 50–100 MB/sec averaged over 12-hour periods. The traffic includes data moved between the compute cluster and the SE, as well as transfers to and from the WAN.

Data going out of the SE is largely input data for analysis done on the local cluster, which tends to have relatively high I/O demands.

In addition to the traffic measured at the SE, the network switch to which the LLNL/LC ALICE facility is connected is also monitored. The results over the same period in the SE monitor plots of Figure 37 are shown in Figure 38. The traffic results plotted as 12-hour averages show nominal rates of 50–100 MB/sec, not inconsistent with the values in Figure 37, which have finer grained data rate averages.

10.5 Local Science Drivers — the Next 2–5 Years

10.5.1 Instruments and Facilities

The US facilities are expected to grow in capability at a rate consistent with the growth in capability of commodity compute servers and commodity storage (periodic hardware refresh results in capability growth that follows the commodity computing market). Such a growth pattern over the next 5 years would increase each of the two U.S. facilities to more than 2,000 cores and 1–2 PB of disk space. As a result, the rate of data migration within the normal ALICE Grid operations should keep pace, likely doubling before the end of the 5-year period.

The two most recent CPU purchases at NERSC/PDSF were combined with larger purchases at NERSC and, for that reason, include IB interconnects between the nodes of the cluster and then to the rest of the facility. This pattern is likely to continue as older PDSF hardware is fully replaced within the next 2 years. Thus, we expect that the individual compute nodes will have a much higher bandwidth capacity to the SE and WANs in the next 2–5 year period. The cluster at LLNL/LC will also be fully replaced in 2015 and the project will consider whether IB or other fast local area network (LAN) interfaces will be included on these nodes at the time of purchase.

A potential change in operations would occur if the NERSC/PDSF facility was to transition to a Tier-1 center as originally planned. In that case, an additional amount of new raw data, and then locally reconstructed ESD/AOD data, would end up on the U.S. facilities. This would require a steady bandwidth of about 50–100 MB/sec devoted to raw data transfers from CERN to NERSC and additional bandwidth out to the WAN as ESD and AOD files generated at NERSC are replicated and distributed onto the ALICE Grid. Since the ESD and AOD files combine to be about 20% of the size of raw data, that additional outgoing bandwidth is expected to be no more than about 20 MB/sec.

10.5.2 Software Infrastructure

ALICE site services, referred to as the AliEn Grid infrastructure, will likely begin a larger evolution during the next 2–5 years. Both CERN IT in general and ALICE in particular will more aggressively push for maintaining uniform computing environments as provided by lightweight virtual machine (VM) technologies. The capability for the VO to instantiate a single computing environment at all its sites is very attractive for both the maintenance of the application software and the reliability of the actual science results. In the past (and currently) effort has gone into making sure that the base code is portable to a broad number of modern Linux systems. In the near-term future, code and architecture complexities may require that code portability become unachievable, to which VM technologies can offer a more direct solution. Evolving to using more “cloud-like” solutions will change the AliEn middleware in ways that are not yet very clear.

10.5.3 Process of Science

The process of science will likely see little change in the next 2–5 years. Data taking will resume at CERN, which adds tasks for new calibrations and new data structures, but this is considered more a steady-state operation of the experiment than the hardware installation and testing work going on now during LS1. The project expects that NERSC will continue to be used for local analysis. In fact, the scientific computing community has a growing interest in supporting serial high-throughput workflows that are typified by HEP/NP event-based data analysis. ALICE is well positioned at NERSC to foster that interest with ongoing tasks. As such, it is plausible that ALICE-USA could deploy a dedicated ALICE AAF at NERSC, an arrangement that has been popular with scientists at other institutions. A dedicate AAF would require managed transfers of specific datasets into the facility. As a result, network usage, both WAN and internal, would be subject to frequent bursts on the order of 100-1000 MB/sec as datasets are staged for processing.

10.6 Remote Science Drivers — the Next 2–5 Years

10.6.1 Instruments and Facilities

The next 2–5 year period coincides with the resumption of data taking for the Run 2 period (2015–2017). Not only will new data begin to arrive, but data will accumulate at a larger rate than in Run 1. Current estimates for Run 2 indicate that both data volumes and number of events will increase by about a factor of 2 over the previous period. The

experiments' needs for MC simulations follow directly with the real data event rates and thus will grow by a similar factor.

10.6.2 Software Infrastructure

The ALICE computing project is in the process of building a long-term software development task targeted to meet challenges expected after LS2. That effort includes a significant rewrite of the AliRoot framework, optimized use of GEANT4 and perhaps even new ways to manage MC simulations (use of parameterized MC or embedding). One target is to make use of heterogeneous architectures. As these code rewrites become available and adopted, they will be rolled out during the next 2–5 year period.

10.6.3 Process of Science

Overall, a comparison of network needs by ALICE-USA in the next 2–5 year period with current needs yields an increased bandwidth by a factor of 2 or 3. The factor of 2 should occur directly from the increase in real and simulated data relative to the Run 1 period and the matching increase in CPU and disk resources expected for the two U.S. Tier-2 facilities. The full impact of the new Tier-1 center at KISTI is not well known but one could expect another 25% increase in bandwidth needs due to the proximity of that facility to the U.S. sites. In addition, more aggressive use of NERSC for local analysis work, particularly on non-PDSF hardware, will add to the network reliance on staging datasets locally.

Two additional changes may occur in the next 2–5 years. There is an effort in HEP/NP to make direct use of U.S. HPC facilities for MC simulations being carried out at Argonne National Laboratory (ANL) and ORNL. That work, in which ALICE is participating, may produce a workflow that can harness significant processing resources for HEP/NP projects. As there is no expectation for significant Grid-enabled disk storage at the HPC facilities, any data produced by ALICE from these efforts will be stored at two ALICE-USA SEs, leveraging the good network connectivity between LBNL, LLNL, and other U.S. HPC centers. Finally, if NERSC does transition to a Tier-1 center, network bandwidth needs will increase as raw data is shipped from CERN and processed data is replicated out to other ALICE Grid sites. In Section 10.4.3, we showed today's nominal WAN use to be a steady about 50–100 MB/sec (both input and output), with bursts to about 1 GB/sec. These should increase to nominal values of about 250 MB/sec, with bursts at 2.5 GB/sec in the next 2–5 years. LAN use is currently several times that amount and will likely require nominal rates of 1 GB/sec, with 10 GB/sec bursts.

10.7 Beyond 5 Years — Future Needs and Scientific Direction

The period just beyond 5 years presents a special challenge for ALICE. While LS2 will likely begin in 5 years (2018), the resumption to data taking following LS2 (Run 3) will include a very large data rate increase in ALICE due to increased luminosity of the PbPb beams at the LHC and changes to the ALICE detector. Specifically, the ALICE collaboration has produced an upgrade letter of intent (LOI), approved by the LHC governing committee, that shows a hundredfold increase in event rates over the previous running

period. At these event rates and with the new detectors expected to be in place in ALICE, the raw data production is estimated to be over 1 TB/sec. The network and data storage costs for coping with nearly 100 PB/day are expected to be prohibitive even at the end of this decade. The current software re-engineering project mentioned in Section 10.6.2 is specifically targeted to the ALICE online HLT system, which will be used to do full event reconstruction in real-time. That reconstruction will include real-time calibration and data quality assurance in order to perform hit and event filtering needed to accomplish ALICE data reduction goals, which reduce data rates from the 1 TB/sec coming off the detector to 10–20 GB/sec written to disk and tape storage. The new envisioned computing model, referred to in ALICE as O2 (online/offline), uses the online systems to generate what are essentially today's ESD and AOD files, with additional information that allows for redoing reconstruction passes with refined calibrations (albeit over filtered event data).

Since the two U.S. facilities currently operate as Tier-2 centers that only run MC and data analysis work, the key factor is not the raw data volume acquired but the number of events. That number, for one, determines the amount of MC needed to adequately do efficiency corrections and systematic uncertainty evaluations. It is expected, however, that the data volume per MC event will decrease either due to utilization of the same techniques to reduce the raw (reconstructed) data per event or through streamlining techniques such as MC parameterization and embedding. In addition to MC considerations, the volume of ESD and AOD files (now about 10% of the raw data) will be the "raw" data. As a result, ALICE network needs may be 100x that of today, with wide-area bursts of 200 GB/sec and local area bursts at TB/sec by the end of this decade.

10.8 Network and Data Architectures

Many sites on the ALICE Grid have configured their resources to be on the LHC Open Network Environment (LHCONE) virtual network. The ALICE team has reported improved performance, increased reliability, and fewer interventions to deal with network problems at those sites. As a result, the team has asked all ALICE sites to participate. Initial discussion with ESnet and local networking has begun and the project expects both sites to join the virtual network this year.

CERN IT has recently announced that it is running out of IPv4 addresses and needs sites to move to IPv6. ALICE-USA has discussed this with the two Tier-2 sites and has relayed a request to be IPv6 compatible. At this point, it appears some work is needed to ensure this will happen by the end of 2013 as requested by CERN.

Finally, network monitoring is an important ingredient to dynamic processing, in which decisions are made based on network capacity between facilities. Two examples presented include the ALICE data distribution model, in which a copy of data produced at a site is automatically sent to the nearest SE; and the processing model that augments jobs run locally to data with jobs that rely on remote data access in cases of high-priority processing. The critical information needed is the nominal available bandwidth between

the ALICE resources. Currently, ALICE relies on VO box or SE communication to monitor that bandwidth.

10.9 Collaboration Tools

ALICE is a worldwide collaboration of about 1,000 scientists and relies heavily on collaboration tools. ALICE has followed (and guided) the decisions by CERN as to the set of tools to use. CERN has contracted with Vidyo[6] to provide VC access, and hosts an Indico[7] site with privileged access on which meetings are organized and presentations are archived. The ALICE-USA computing project is centered at LBNL and relies on ReadyTalk for phone conferencing and LBNL-hosted Google site for managing its relevant meetings.

10.10 Data, Workflow, Middleware Tools, and Services

ALICE has been developing its distributed computing model for over a decade and relies heavily on networks for data movement and access. Several years ago, ALICE chose to use XRootD for data storage and transfers. ALICE XRootD includes a GSI (Grid Security Infrastructure) authentication plug-in to manage access to its data. Personal grid certificates for ALICE-USA scientists are obtained through the DigiCert Grid cooperative agreement (CA) using OSG Registration Authority as part of ALICE VO participation in OSG.[8] However, ALICE-USA participants may also use their CERN CA signed certificate that is available to all ALICE users. Client requests for data are made directly to the AliEn File Catalog, where information about each file is stored. This information is used to direct jobs to the site where the data reside. Any ALICE scientist can request data from the File Catalog and will be given site-level SE information allowing the client code to connect directly to the XRootD redirector at the site.

ALICE computing has adopted MonALISA for monitoring all sites and services on its grid infrastructure. The data is accessible via the Web-based repository, http://alimonitor.cern.ch/map.jsp. As noted in Section 10.8, 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 Fast Data Transfer (FDT) tool.[9] 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. Tier-2 sites are provided on the MonALISA Web display:

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

[6] http://www.vidyo.com/

[7] http://indico.cern.ch/

[8] 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 WLCG.

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

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

10.11 Outstanding Issues

Understanding the topology of network connections and learning to measure and respond to network issues is a challenging aspect of the operation. This is important to ALICE, as the ALICE Grid contains about 80 sites that are all used dynamically to access data. For example, in the table at the above links, LBNL and LLNL are often the closest sites to each other as expected — but not always. On occasion, the measured bandwidth between the U.S. sites is poor and individual sites in Europe or Asia appear to have a better network connection to the U.S. sites than the U.S. sites to each other. 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.

10.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0-2 years)				
• LHC is currently in a shutdown (LS1). • ALICE workflow for the next two years includes new reconstruction pass of some data and significant MC simulation running. • U.S. facilities are part of the ALICE Grid and represent about 7% of ALICE resources.	• MC simulation and a majority of data analysis are done directly on the ALICE Grid. • Analysis jobs are sent to sites with the data. • Some high-priority jobs access data remotely.	• Local data volume is ~0.5 PB in support of data analysis done on the cluster. • New data generated locally is ~1 TB/day • Typical file size ~500 MB.	• MC jobs have low transfer time demands. Those Jobs run few hours and produce a few GB. • Analysis jobs can have large per-job I/O range: 0.1–100 MB/sec. • Approx. 1,000 concurrent jobs/site. • 0.25-2.5 GB/sec.	• Steady transfers into and out of each site of 50–100 MB/sec with bursts reaching 1 GB/sec. • Some targeted transfers from CERN, but routine transfers between all ALICE Grid sites.

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
2–5 years				
LHC resumes operation. Data rates to increase by order x2.	• MC simulation and a majority of data analysis done directly on the ALICE Grid. • Analysis jobs are sent to sites with the data. • Some high-priority jobs access data remotely — such access may become more frequent. • Development of dynamic analysis facilities.	• Local data storage to increase to 1-2 PB per U.S. site. • New data generated increased to ~2 TB/day • Files likely remain 500 MB but could grow if processing speed is increased.	• MC jobs have low transfer time demands. Those Jobs run ~few hours and produce tens of GB. • Analysis jobs can have large per-job I/O range: 0.1 – 100 MB/sec. • Approx. 2,500 concurrent jobs/site. • 1-10 GB/sec.	• Steady transfers into and out of each site of ~250 MB/sec with bursts reaching 2.5 GB/sec. • Some targeted transfers from CERN, but routine transfers between all ALICE Grid sites. • Potential modest increases specific to decision on whether NERSC becomes a Tier-1 facility.
5+ years				
Large increase in PbPb luminosity and detector readout yielding 100x increase in event rate.	• Large increase in MC requirements and analysis activities on the ALICE Grid. • Reliance on HPC / heterogeneous architectures may become important and may develop specialized MC facilities.	• Local data storage to increase 10x. • New data generated increased 20 TB/day. • File sizes likely to grow beyond GB/file as processing speed is increased.	• MC jobs may have dramatic speedup with new architectures but may be partially offset by use of filtering or embedding, producing per-job rates reaching 100 MB/sec. • Analysis jobs will retain large per-job I/O range. • >10,000 concurrent jobs/site. • 100-1000 GB/sec.	• Steady transfers into and out of each site of ~10 GB/sec with bursts reaching 100 GB/sec. • Some targeted transfers from CERN, but routine transfers between all ALICE Grid sites. • Potential modest increases specific to decision on whether NERSC becomes a Tier-1 facility.

11 The PHENIX Experiment at RHIC (BNL)

11.1 Background

The PHENIX (Pioneering High Energy Nuclear Interaction eXperiment) Experiment is one of the two large detector systems at the Relativistic Heavy Ion Collider (RHIC). After major detector additions in 2011 and 2012, the experiment has about doubled its channel count to about 2 million readout channels in 17 different detector systems. The data rates of the experiment are driven by a high sampled and recorded data rate (5–8 kHz to disk), leading to peak data rates of about 1.4 GB to disk.

The PHENIX collaboration is actively engaged in an ambitious detector upgrade, code-named "sPHENIX." With the exception of the newest two detector systems, the Vertex Detector (VTX) and Forward Vertex Detector (FVTX), virtually all of the existing detectors will be decommissioned and replaced by a magnetic solenoid, an electromagnetic calorimeter, and a hadronic calorimeter. This detector is also anticipated as a basis for a detector system at the planned Electron-Ion Collider. For the purview of this document, the sPHENIX computing needs mostly focus on simulations.

11.2 Key Local Science Drivers

11.2.1 Instruments and Facilities

The PHENIX experiment is located in the Building 1008 complex at the RHIC ring. The data are recorded to disk locally in the Countinghouse, and sent to the RHIC Computing Facility for long-term storage in the HPSS storage system.

The local buffering on "buffer boxes" at the experimental site helps to 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 whole raw datasets were sent over the network to remote facilities for processing (to the Computing Center in Japan [CCJ] and Vanderbilt). The moving of large-scale datasets to make use of remote processing capabilities has not proved efficient. Therefore the majority of the data processing is now performed locally at RACF.

A number of reconstruction passes over raw data, a slightly modified data analysis setup, and the onset of simulations for the sPHENIX apparatus have led to a very efficient use of the PHENIX resources at the RACF. We have begun the process of running a fraction of the sPHENIX simulations outside of the RACF on the Grid, giving priority to data-intensive tasks to be run at the RACF.

11.2.2 Process of Science

The raw data get 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 the goal is 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. Datasets transferred to Japan tend to be larger in runs that have a major spin (proton-proton) 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 has proved very inefficient, as it typically leads to multiple retrievals of the same file in a short period of time.

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

The recent change involves a more on-demand setup for trains. We now start trains whenever there is sufficient demand for a data pass, where the threshold for a train to start can be adjusted based on the load levels at the RACF. This has helped to eliminate extreme peaks in network load at the start of a train by distributing the start times more randomly.

11.3 Key Remote Science Drivers

11.3.1 Instruments and Facilities

The PHENIX collaboration currently consists of 70 institutions in 13 countries. About 250 scientists who are part of the PHENIX collaboration 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. With the onset of Grid-based simulations for sPHENIX, it is probable that there will be an increase in WAN traffic, although the intention is to offload mostly the peak demands to the Grid.

11.3.2 Process of Science

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

11.3.3 Local Science Drivers — the Next 2–5 Years

11.3.4 Instruments and Facilities

The PHENIX collaboration has recently completed an ambitious detector upgrade with the commissioning of the VTX in 2011 and the FVTX in 2012. Each component has increased the overall data rate by increasing the per-event size, although the impact has been reduced by the lack of an HI component in Run 13. The long AuAu Run in 2014 will likely yield the largest raw dataset ever.

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 sPHENIX efforts will require large-scale simulations. Currently, there are three categories of simulations:

1. Event generators. Those do not simulate any detector responses and are used to establish the capabilities and limits of the detector geometries under investigation. This type of simulation is well suited to be run on the Grid (or at remote institutions) because of the relatively low I/O to CPU ratio. Recent simulation runs, performed at the RACF concurrently, have each produced about 25 TB of data in a time span of about a week. Each such run typically focuses on a special setup with specific impact parameters and other properties such particle flow parameters.
2. Detector design studies. Those take the output from the event generators and, using GEANT4, simulate the energy deposition in a generic detector, that is, without a specific segmentation. For example, the hadronic calorimeter is treated as a homogeneous cylinder, and the energy depositions in the material are recorded at an interaction-by-interaction basis. This allows us to impose different detector segmentation and detector cell arrangements later by summing up the energy

deposited in a given cell volume. This kind of simulation is not well suited to be run remotely because of the large output volumes of 5–40 GB of data per event, and will likely only be run at the RACF.

3. Detector performance studies. As the detector setup is refined and approaches the final design, the detector segmentation is at the GEANT4 simulation stage, which allows us to write out only the final detector response. This will lead to a reduction in output volume compared with the design studies by an estimated factor of 100. We expect this kind of simulation to be the main contributor to the WAN traffic.

11.3.5 Process of Science

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

11.4 Remote Science Drivers — the Next 2–5 Years

11.4.1 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 BNL.

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

11.4.2 Process of Science

No changes are anticipated in the 2–5 year time frame.

11.5 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. PHENIX believes the currently available technology will be able to sustain the envisioned data rates, even without a progression of the current rate of improvements in data storage and processing technologies.

The schedule, funding, and support for the sPHENIX are unknown at this time.

11.6 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, which have worked well in the past.

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

11.7 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Dataset 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 in 2014, 1.5 PB in 2015. • Reconstructed to ~1 PB DSTs.	• Near-line. • 300,000 raw data files. • 10 GB each. • Network in place.	• Virtually no near-line requirements. • 800 TB volume estimated.
• PHENIX detector data.	• Distributed analysis.	• 300 TB (2014). • 150 TB (2015).	• Random transfers of few files. • 2 weeks of successive transfers of 2.2 GB files.	• Random transfers of few files. • 2 weeks of successive transfers of 2.2 GB files.
• sPHENIX detector simulations.	• Distributed simulations.	• 24 TB/project.	•	•
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. • 2 TB/year estimated.	• Near-line. • Network switch upgrades (optical infrastructure in place).	• No near-line requirements. • 1-1.5 PB/year estimated.
5+ years				
• sPHENIX commissioning and operation. • One year without data taking. • Detector response simulations.	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. • 2 PB/year estimated.

11.8 APPENDIX A

PAC Decision about Run 14 and 15

Run 14:

- 14 Weeks of 200 GeV Au+Au running
- 3 Weeks of 15GeV Au+Au

Run 15:

- 200 GeV p+p
- 200 GeV p+Au
- Additional mixed species running (p+Si, p+Cu, d+Au, ^{3}He+Au)

12 The Solenoidal Tracker at RHIC (STAR) Experiment

12.1 The Physics Case of STAR and Evolution Toward eSTAR

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

The most important discovery made in this area over the past decade is that the QGP acts as a strongly interacting system with unique and previously unexpected properties (sQGP). While early expectations and predictions from the NP community foresaw a QGP behaving like an ideal gas, the matter produced in near-central RHIC collisions was shown to flow as a nearly viscosity-free fluid (a.k.a. "perfect liquid"). Further, yields and flow of mesons compared to those of baryons have established a scaling behavior that points to collective flow established at the quark level, with hadrons subsequently formed by coalescence of already flowing quarks. Through its unique and versatile polarized proton beam, the RHIC spin program has made great strides toward unraveling the decades-old question about the partonic origin of the proton's spin.

Longitudinally polarized proton collisions are currently the world's best source of information about the gluon helicity distribution, with recent measurements indicating gluons may contribute as much as quarks (about 20–30%) to the total spin of the proton. Collisions of 250 GeV beams permit studies of W-boson production, providing direct and theoretically clean access to the flavor-separated sea quark helicity distributions. Transversely polarized collisions have allowed STAR to show that the unexplained large asymmetries present in previous fixed-target experiments persist even in the collider regime. The origin of these asymmetries is still not understood and has led to a vibrant transverse spin program designed to study the parton spin distributions in transversely polarized protons. RHIC has also engaged and started a Beam Energy Scan (BES) program and is the only machine that can systematically probe the plasma in the vicinity of the transition by varying both temperature and baryon density. In other words, RHIC/STAR can explore a region of the quantum chromodynamics (QCD) phase diagram (critical point, phase structure, baryon density) much more widely than any other facility is able to do.

RHIC/STAR has now essentially completed a set of major upgrades facilitating the next decade of science. The physics program could be summarized as two major campaigns of studies: the first, from 2014–2016, is focused on Heavy Flavor and Di-leptons measurement to study the properties of the sQGP produced in the high-energy nuclear collisions at μ_B close to 0. The second phase (2018–2019), will refocus on the RHIC Beam

Energy Scan Phase-II. The physics will then focus on the search for the QCD critical point and study the QCD phase structure at the high baryon-density region μ_B > 250MeV.

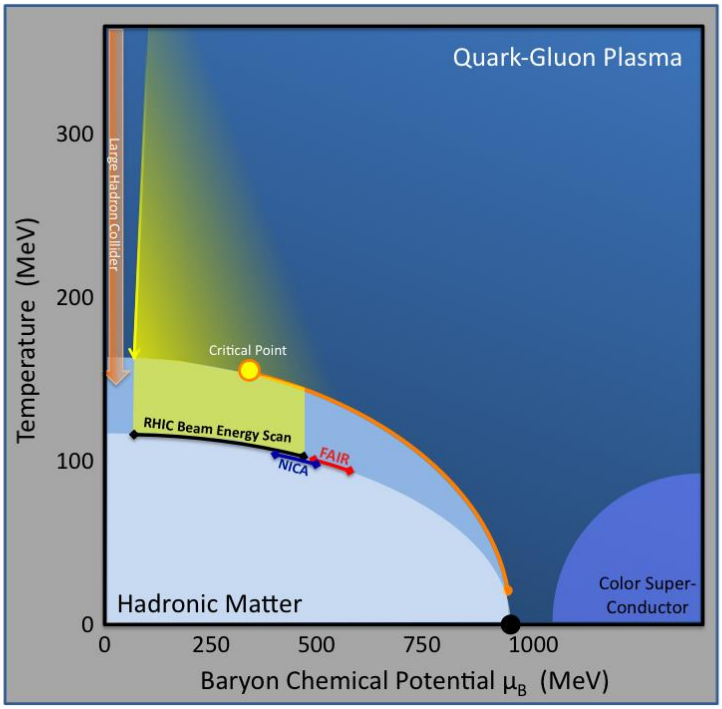


Figure 39. Illustrated phase diagram. The yellow portion shows the specific temperature and baryon chemical potential range that RHIC may cover. Also shown are the LHC, NICA, and FAIR coverage of the diagram to better illustrate the unique opportunities of the RHIC program.

To achieve this ambitious program, the first wave of upgrades will provide unique insights on the sQGP properties and focus on the charm and di-lepton measurements. STAR is already equipped with enhanced Particle Identification Detector (PID) systems and hence able to study a wide variety of secondary decays (including the study of hyper-nuclei as an offshoot of STAR's physics program). With its new Muon Telescope Detector (MTD), STAR enhances the muon-to-hadron ratio by orders of magnitude and will be able to separate upsilon states and study the heavy flavor collectivity and color screening. Combined with the Heavy Flavor Tracker (HFT), STAR will be able to study the prompt J/ψ and nonprompt J/ψ (from B-decay) as well as perform detailed studies of the D^0 meson (Run 14 objectives) and later study the charmed lambda or Λ_c. In 2017, the RHIC facility will be equipped with electron cooling capabilities while the STAR subsystem and central tracking detector expect to have the inner sector upgraded, allowing for higher tracking precisions (iTPC upgrade). By 2018, STAR will be ready to engage into the deep study of the QCD phase structure and the critical point to gain knowledge of the characteristics of the phase boundary and the dynamical evolution from cold nuclear matter to hot QGP. The beam-energy scan program has potential for unparalleled discovery to establish the

properties and location of the QCD critical point and to chart out the transition region from hadronic to deconfined matter.

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

Overall, the STAR Beam User Request (BUR) is summarized in Table 14. This BUR is the start of all of our requirements and should the run plan change or be altered, the numbers reported herein will change accordingly.

Note that the extreme data sample quoted in 2016 is accurately reporting the numbers from the official STAR BUR. However, if STAR is equipped with better vertex constraint capabilities, this data sample will be reduced by a factor of 2. This point is re-addressed later.

Table 14. STAR beam user request from 2014 to 2019.

RHIC run Year	Species	Number of events (B=Billion, M=Million)
2014	Au+Au 200 GeV Au+Au 15 GeV	2 B (minbias, central) + ~ 0.78 B misc 20 M
2015	p+p 200 GeV p+Au 200 GeV	2.2 B (2 B minbias + trigger mix) 600 M
2016	Au+Au 200 GeV	4.2 B (4 B minbias, …) – large sample
2017	Collider upgrade (eCooling) and STAR/iTPC upgrade	N/A
2018	BES-II p+p 200 GeV longitudinal	400 M (mix of 19.6, 15, 11.5, 7.7 GeV) 1.4 B
2019	P+p 510 GeV, transverse	2 B

12.2 Data Flow Background

The RHIC/STAR Experiment's data taking initiates from BNL, where its data workflow begins. The STAR detector system is currently composed of eight major detector sub-systems (BEMC, EEMC, TPC, HFT, FGT, TOF, GEM, MTD) and numerous triggering systems; the whole data flow is thus composed of 10 main areas of software coordination.

The DAQ system of STAR itself is capable of sustained rates as high as 1.1 GB/sec with peaks at 1.6 KHz event rates. The theoretical limits of the throughput of the DAQ system (based on disk I/O for data buffering and local network performance) are 2.5 GB/sec, though at a modest cost (about $1000 per additional 60 MB/sec), the system could be upgraded by adding more hardware online on the STAR side.

STAR is organized in a classic structure of tiered centers, where BNL is the Tier-0 center of real data collection and the repository of generated simulated data (a copy of the embedding data is brought back to BNL for safekeeping). Tier-1 centers provide a significant resource or service (CPU cycles for data analysis or simulations, archival storage for long-term preservation of STAR data, etc). Network traffic between BNL and STAR's Tier-1 centers is the primary object of our requirements. STAR Analysis Centers (SACs, a.k.a. Tier-2 centers) are defined as local compute farms or a portion of main facilities providing analysis cycles to local scientific teams. Usually, such SAC centers have limited storage resources; hence, network traffic and load are minimal. However, a SAC may move data from anywhere available, as STAR does not need or have a strict tier hierarchy.

12.3 Collaborators

The STAR institutions' demography and its evolution across the past and present ESnet reviews are represented in Figure 40. As of 2013, STAR remains a strong collaboration composed of 56 active groups and institutions spanning three continents, five main geographical groupings (networking wise), 12 countries, and 550 scientists.

For the sake of completeness, note that Figure 40 does not include the counting of institutions that are solely composed of emeritus members or institutions phasing out (finishing students), which do not generate network requirements of any kind or are likely to be removed within a year. More importantly, while the demographics remain stable, not all institutions are equal network consumers and it is important to focus on the typical data path and core activities.

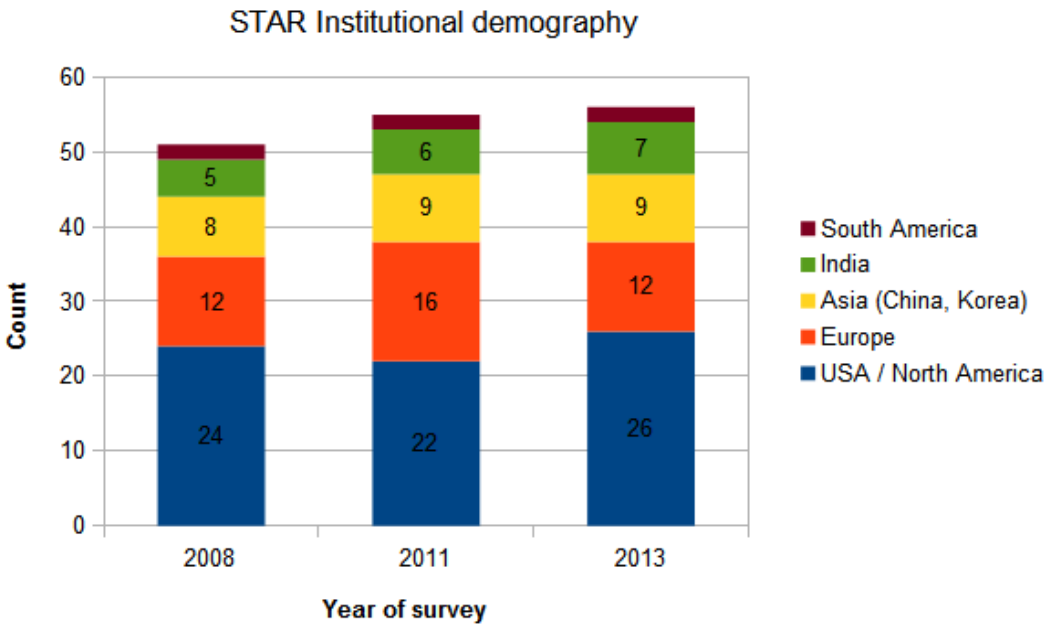


Figure 40. STAR institution evolution over the three ESnet reviews.

Our collaborators remain focused on remote log-in capabilities to either the core facility at the BNL RHIC Computing Facility (RCF), which is STAR's Tier-0 center, or the NERSC/PDSF center as the Tier-1 center. Both facilities are heavily used for analysis although the analysis at NERSC/PDSF has been at times challenged by the need to run aggressive data-simulation campaigns (a.k.a., embedding production) sharing the same "rigid" resources (rigid as opposed to "elastic," as a cloud approach may imply). Our past plan was to ramp up the resources at KISTI to make up for the gap in resource needs to support either the embedding and analysis requirements or to create a shift of resources in the data processing requirements, restoring resources for other processes.

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

In the past planning document (SN0548), we envisioned adding more STAR Analysis Centers (SACs) as the physics program matured and demands for more analysis power grew. We noted that their inventory was hard to assess but constituted pools of local resources dedicated (not necessarily shared with all STAR users) to a local group's physics program needs. We planned to develop strategies to help integrate those centers into a global data analysis and data distribution pool. The status remains the same — there are no mechanisms to help or encourage SACs to share resources across the collaboration and no clear mechanisms to help supply them with a workforce able to maintain/ upgrade their local setup. It still is not possible to clearly assess their number. Trying to include those centers as part of the STAR VO (via the OSG software stack and services) has been deterred by the lack of local workforce able to ensure the sustainability of those resources on the OSG/Grid. Support is on a "best effort" basis. Monitoring the number of remote databases (slave servers of BNL master metadata repository), we infer we would have at this stage four active centers, which is lower than our past projections by one unit. Those four centers are: Prague (our most stable active center), UIC, Wayne State University (WSU), and USNA (MIT has become inactive due to workforce shift). Our new projected number is shown in Table 15. We predict the loss of WSU in 2014 but

Table 15. Projected number of Star Analysis Centers (SACs) from 2013 to 2019. The 2013 estimate represents the number previously projected; the actual number is 4.

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

expect to regain an additional institution and to reach a plateau in future years to three SACs at most.

The STAR computing model continues to rely on a data grid model. The processed data is made almost immediately available to remote sites where computing resources are available. Data distribution tools have been consolidated by the addition of a global file replica and metadata catalog (we will refer to it as the STAR FileCatalog), able to make differential inventories between sites within minutes, and the development of in-house tools for reliable data transfer and redistribution. The resources from the OSG are seldom used and only sites dedicated to STAR's use have been integrated into a Grid-based workflow (except the Tier-2 centers as noted above).

12.4 Data Size Projections — Setting the Basis for Our Network Requirements

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

Based on analysis of past event size, we projected the evolution up to 2019 and summarize the results in Table 16. The upper part of the table shows the size of the

Table 16. Projected event size for RAW and DST files for STAR up to 2019 as a function of species. The basics for the extrapolation and projections are shown for 2012 and 2013. The numbers are in units of MB/events.

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

MuDST while the lower part predicts the size per event of the DAQ files as a function of a single species and year. To reach those numbers, projections of the effect of luminosity on event size have been folded in as well as the phasing in (and out) of new detectors. The iTPC upgrade alone will cause a TPC data size increase of 40% and create a jump in event size.

While imperfect (not all data for the species planned for future runs are available), the expectations of data-size growth can be inferred by folding the values from Table 16 and the STAR run plan alone in Table 14. This would lead to the resulting estimates of Table 17.

Table 17. Event size projections considering the species mixed foreseen by the STAR BUR.

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

From the expected dataset mix (species, trigger) and their respective event size average, we can make projections about the yearly dataset size we will encounter for the period of 2014-2019 — while 2019 is beyond the required timeline, it seems judicious to include it for two reasons: (1) 2017 marks a machine and detector upgrade period, during which the data requirements for RAW will be null — hence, going up to 2019 maintains the same amount of years for the RAW data, and (2) the RHIC/STAR BUR sets two clear physics program objectives, one of which is past the 2017 machine upgrade. We summarize those projections in Table 18.

Table 18. Projected dataset size for the 2014-2019 period. The two first years are shown as a basis for the projection and verifications.

	Initial projections						*Outer years projections*	
	2012	2013	2014	2015	2016	2017	2018	2019
Species	U+U 193 GeV, p+p 500 and 200 GeV BES 5 GeV Cu+Au	p+p 500 GeV	*Au+Au 200 GeV BES 15 GeV*	*p+p 200 GeV, p+Au 200 GeV*	*Au+Au 200 GeV*	*N/A Machine Upgrade (eCooling) & iTPC*	*BES-II (multiple energies) p+p 200 GeV long*	*p+p 510 GeV trans*
Projected N events (B)	2.20	2.50	2.80	2.80	4.20		1.80	2.00
Projected size RAW (TB)	1321.05	1801.44	2800.04	3025.05	6280.19		1466.39	4066.13
	All data		*Trend projections (upper bound considering hsitorical deviations to plans)*					
N events (B)	6.1	2.7	*3.1*	*3.0*	*4.4*		*2.0*	*2.2*
Final size RAW (TB)	2147.24	1985.43	**3080.05**	**3267.05**	**6594.20**		**1613.03**	**4472.75**
Deviation to projected	93.18%	8.00%	10.00%	8.00%	5.00%		10.00%	10.00%
	4249958673							
	5.3	2.54						
RAW Data with tracking detector (candidate for data production)							*Projected based on possible excess*	
sum(events) tpx	5686466071	2728446873	*3080000000*	*3024000000*	*4410000000*		*1980000000*	*2200000000*
sum(size) tpx (TB)	2140.1	1980.06	*2868.52*	*3042.68*	*6141.32*		*1502.25*	*4165.57*
Size / events (MB)	0.39	0.76	*0.98*	*1.06*	*1.46*		*0.80*	*1.99*
Initially projected in 2011	*0.59*	*0.70*						
Real up to 2012, complete up to 2011, 2013 onward are projections							*Projected for derived data (MuDST)*	
Total events MuDST	3753824889	*2526128181*	2851613083	2799765572	4082991459		1833179839	2036866488
Fraction of events to RAW	88.33%	92.58%	92.58%	92.58%	92.58%		92.58%	92.58%
Total size MuDST (TB)	931.11	*1060.01*	**1967.08**	**2173.92**	**4211.19**	**2784.06**	**937.13**	**2450.47**
Size / events (MB)	0.26	*0.44*	*0.72*	*0.81*	*1.08*		*0.54*	*1.26*

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

The second note is that while our past projections (ESnet report from 2011) expected a RAW event size average of 0.70 MB/event in 2013, the run plan was modified for the benefit of one species (the mix is different, the average event size is impacted). Table 16 would however indicate an event size of 0.77 MB/events for 2013 and our final number is remarkably accurate at 0.76 MB/events. The MuDST size per events is estimated to be slightly larger due to a few detectors added to the data stream, the information of which will need to be propagated with redundant information so the detector response can be better understood.

We noted in Section 12.2 that STAR plans for an extremely large dataset in 2016. This impressive data sample is driving the requirements but may be reduced by a factor of 2, depending on STAR's ability to select the primary event vertex with a cut of less than 5 cm accuracy. This deliverable is not formally a computing deliverable (and hence not immediately under our control). No detector setup can achieve this vertex selection at this stage. Nevertheless, this selection can be achieved by ensuring that HLT vertexing capabilities, in addition to tracking, are in place by 2016. Currently, the same team of computer scientists from Germany (FIAS), who now have full membership in the STAR collaboration and with whom we collaborated on the HLT tracking before (along with CBM, ALICE, and other experiments), is visiting BNL. With physicists from several STAR institutions, computing organized a focused effort to tackle this problem. For the sake of projecting and making sure STAR does not fall behind network resources, we did not fold in this (yet unproven) possibility but did align with BUR requests for consistency. However, because our confidence in the steering of this deliverable in 2016 is very high, we will systematically consider the reduction of this dataset by a factor of 2 wherever it applies, and we will proceed with gross approximations to the lower value. In other words, it would be extremely premature to draw conclusions about the possible storage requirements and strain on the facility such datasets may impose on the facilities hosting STAR data.

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

12.5 Key Local Science Drivers

In this section, we focus on the Tier-0 aspects and LAN requirements and will treat all other facility requirements in Section 12.6 and related subsections.

12.5.1 Instruments and Facilities

The BNL RHIC Computing Facility (RCF) hosts all RHIC experiments. The facility's core operation and role is to provide the core CPU computing cycles for half of our users' analysis needs and all the data reconstructions; and support for data calibration, data reduction, database, and some local need for simulations.

During data taking, the STAR DAQ system streams data to a cache space spread over 10 buffer-box nodes (nodes collecting and aggregating the data into streams and files) for a total of 96 TB disk space. In this configuration, and depending on the DAQ rate, but assuming 600 MB/sec, STAR would be able to hold its ground for about 46 hours without network connectivity before suffering any data loss. At observed peak rates of 1.1 GB/sec, STAR would still maintain operational viability for 24 hours. The data is, however, pushed to the RCF via 2 x 10 Gb links onto a disk cache of 54 TB space (near a 2:1 space match) located in front of the HPSS tape archiving system. STAR has accumulated about 12 PB of storage space in HPSS to date (about 7.6 of which are RAW data). Datasets from 2012 onward have been on the order of multiple petabytes and are the dominant contributors to the overall data storage volume.

The LAN requirements from the DAQ to the HPSS systems shown in Table 19 are based on average run time and hours of physics running suitable for data taking observed in previous years as well as the input from Table 18. The maximum line speed (sustained) needed for the entire period exceeds a 1 x 10 Gb link but remains below 2 x 10 Gb links. For LAN connectivity, STAR is currently covered for both sustained and peak rates.

The facility currently provides CPU powers of the order 76 k HSPEC delivered by more than 9,192 CPU cores. The total storage capacity has reached about 560 TB of central storage, served over NFS and usable for data production (and space reserved for dedicated tasks such as calibration, user analysis space, simulation, and space for support of STAR's distributed computing program). The CPUs are standard off-the-shelf commodity hardware and nodes hosting local storage (cheap disks) for a total of about 3.3 PB of distributed storage space holding a portion of our DST files. Distributed storage has come to be the main storage resource for analysis files since 2010 or so.

Table 19. Network LAN requirements from the DAQ to the HPSS systems for the period of 2014 to 2019. For historical purposes 2013 is included. A margin was folded into the calculation to account for possible protocol overheads.

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

12.5.2 Software Infrastructure

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

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

During a full data-production campaign, all files (and all events) flagged for data production go through the data production process. As the data is distilled into DST, the results are double copied: one copy goes to the HPSS storage for permanent archiving and a second copy is randomly placed on one of the 80 file-system partitions available as data production space (the random placement is done for load-balancing purposes). Indexer daemons pick up newly created files as they appear and immediately catalog them in the STAR's FileCatalog. During this process, the file's checksum and size (queried or computed during production time) are verified — if either do not validate, the file is not cataloged and is flagged as "bad." At the end of the production campaign, they may be reproduced. This paranoid check, mainly implemented in case of network

communication oddities, has not detected a single occurrence of validation failure for the past two years (below a 2% loss due to this effect or other core common problems, we do not resubmit). As the cataloging occurs, the presence of an HPSS copy is checked — if present, the NFS file may be removed immediately; if not present in HPSS, the NFS copy is pushed again to HPSS (a few percent failures in the production workflow in moving the data to HPSS occur). Typically, the NFS files are NOT removed to allow the next stage to take place.

We had planned to alter the workflow described above to avoid the extra step of a copy in HPSS (xrdcp or a direct copy into a local disk space was planned). However, STAR distributed storage capacity is at approximately 70% of its requirements; therefore, deployment of a streamlined workflow was delayed (also, error handling for the failure of an xrdcp copy might disrupt the production workflow). We planned to remove this obstacle by the 2014 purchase cycle (within the past established funding profile and projections, distributed storage will be sufficient to automate the production workflow in the outlined manner).

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

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

Typically, a 1Gbps interface to each node is sufficient to restore the occasional data loss from each node. Even a massive restore of data to 500 nodes with this network bandwidth can be done within 2.3 hours, assuming no constraint of throughput from HPSS.

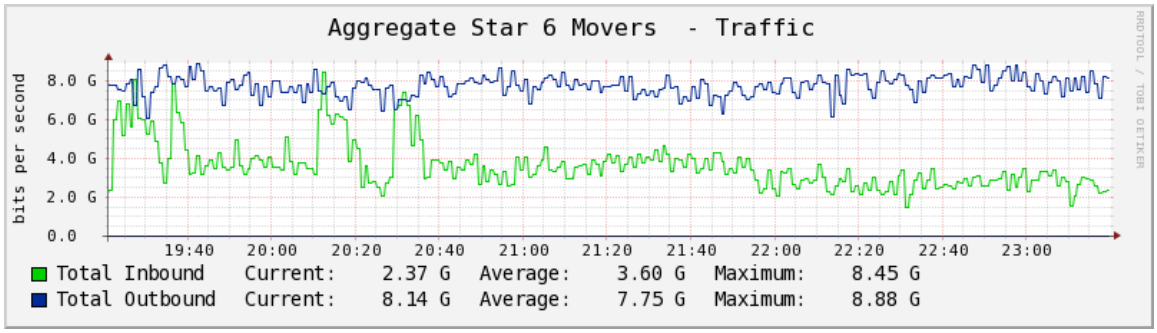


Figure 41. I/O graph during a mock I/O challenge to/from HPSS. Both data move from the experiment (blue). Data read for production purposes (green) were simulated in a realistic load environment.

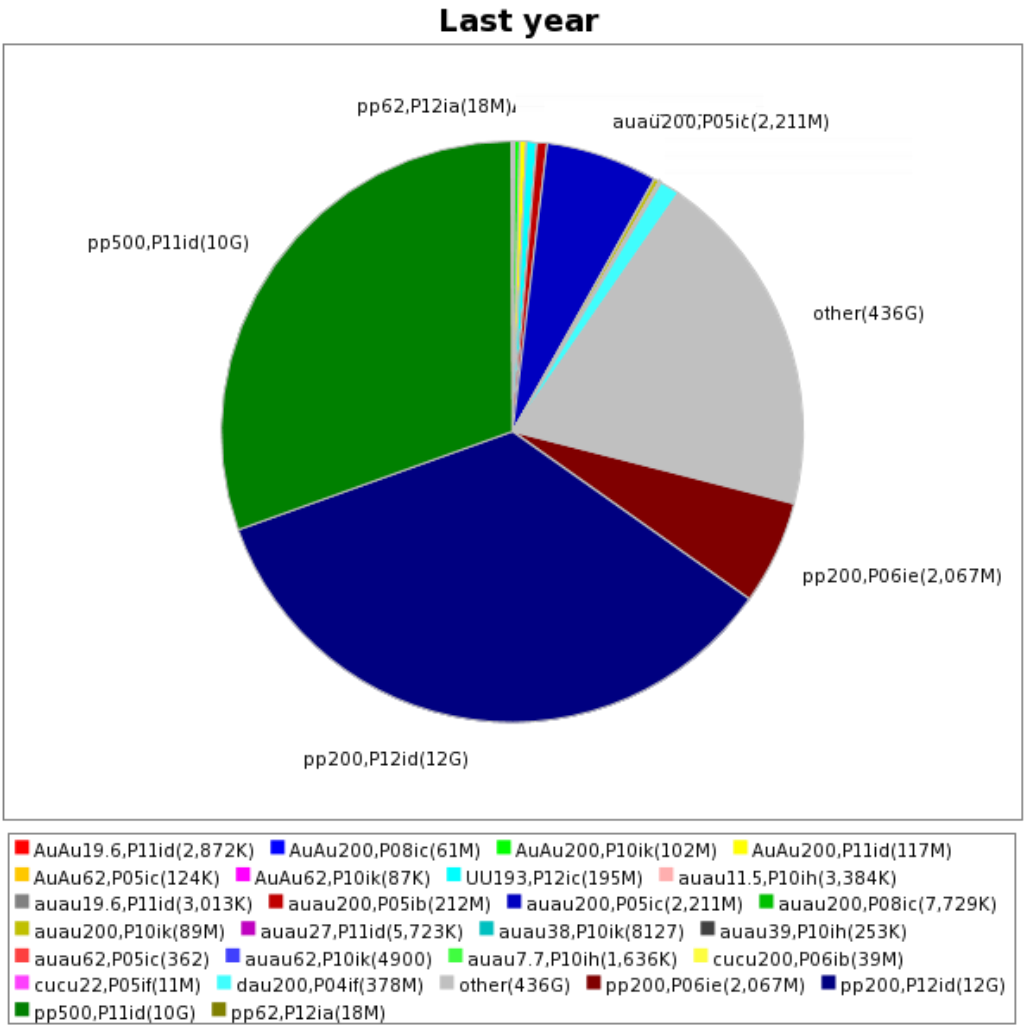


Figure 42. Usage statistics example of data access pattern from users for the past year. In this case, the statistics give an idea of dataset access pattern by production and collision species. This information is used to determine the actual "hot" datasets.

Figure 41 represents the expected data throughput to/from HPSS in simultaneous read/write mode. The HPSS system has been shown to provide an aggregate of 4 Gbps (peaks at 8.45 Gbps are rare in this case) and thus, the restoration of such dataset loss (500TB) would actually map to a 12-day restoration of the data. To reduce this intrinsic limitation, the performance of the HPSS infrastructure itself would need to be expanded (the network is not the limiting factor in this process).

One consequence of those lengthy restorations of our large (and becoming larger) datasets is that the dynamic "on-the-fly" (or on-demand) disk population of datasets is a conceptual ideal of no practical use unless jobs submitted to a batch system can be delayed for up to weeks. Therefore, STAR data are pre-staged on distributed disks based on feedback and observations. There are two sources for such feedback and input: (1) the PWG is regularly polled for its dataset usage intents (ordered by priority) — those inputs are summarized across the PWG and, depending on space availability, the datasets with greatest demand are replicated across the virtual storage pool while the lowest priority (and lowest occurrence) have a single copy available; and (2) the usage from STAR users themselves — they use a job submission interface to specify datasets based on metadata declaration. Their usage is recorded and monitored. The monitoring includes aggregate information related to the currently accessed datasets and most-accessed datasets and data-production campaigns as a function of time range (past days, weeks, months, year). Figure 42 is an example of such a graph. Evolution of the analysis pattern as well as indicators of hot datasets (datasets most used) can be inferred from those graphs and datasets replicated accordingly.

Finally, the lengthy cycle for data restorations in case of data loss points to the need to secure distributed storage for resilience and redundancy. The generalized use of RAID-5-based local storage will reduce the possibility of such data loss. This will be in place in all future storage systems, and the additional disk space used by RAID-5 will be taken into consideration for storage requirement calculations.

12.5.3 Process of Science

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

The resources for STAR at the RCF are separated into two sections: an analysis farm (CAS) and a production farm (CRS). While data movement through the CRS nodes are hard to interpret, at full farm occupancy the jobs on the CAS are essentially user jobs reading data from Scalla/XrootD and reducing the data to picoDST or histogram files (the I/O of which is negligible). A few typical I/O profiles of our nodes are shown in Figures 43 and 44. Both nodes have similar storage space and show an I/O rate in the node of around 12 MB/sec and out of the node at about 5 MB/sec. To first order, those rates do not concern us considering the 1 Gbps network interface.

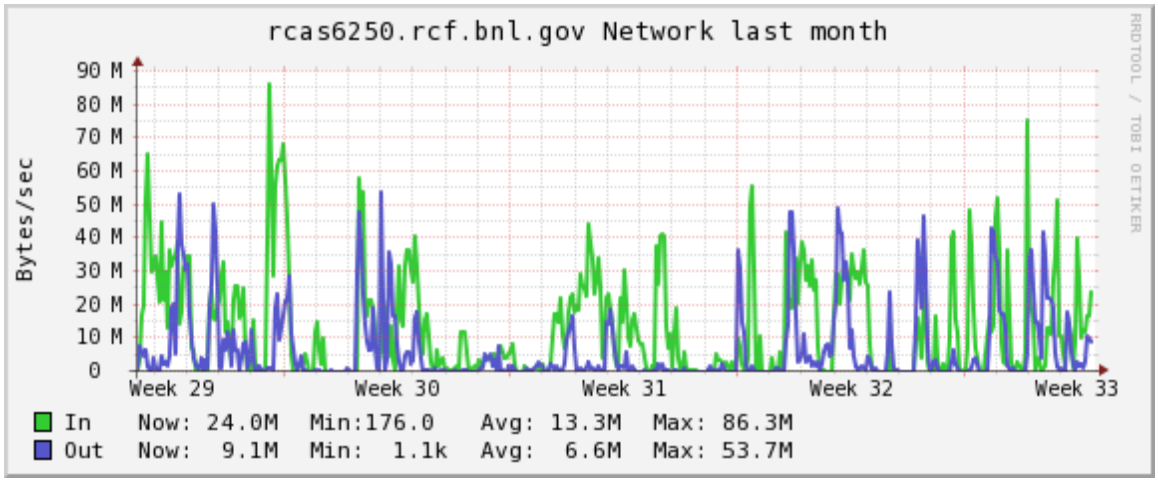


Figure 43. Typical I/O in and out of a node on an analysis node.

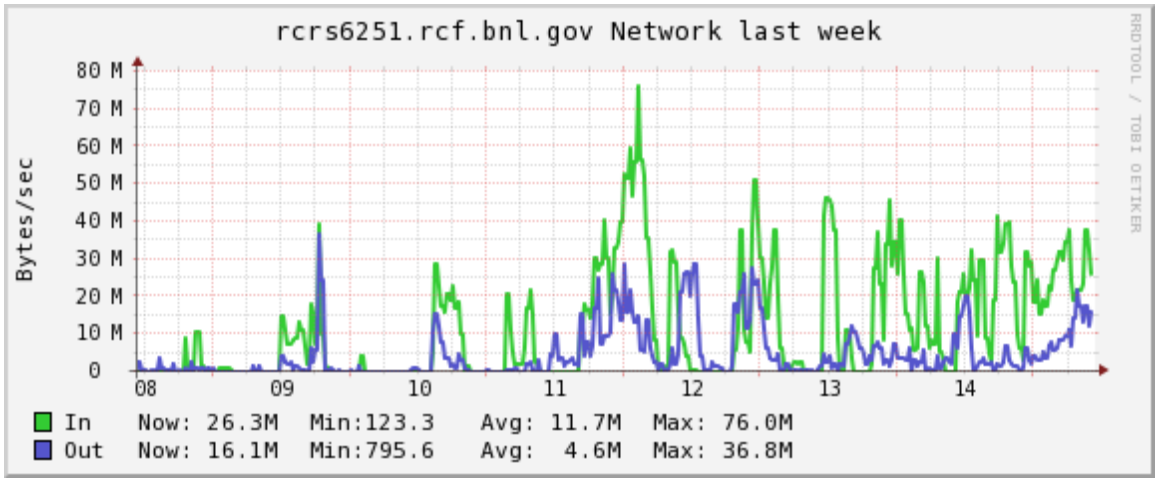


Figure 44. Typical I/O profile in a period of no data production campaign.

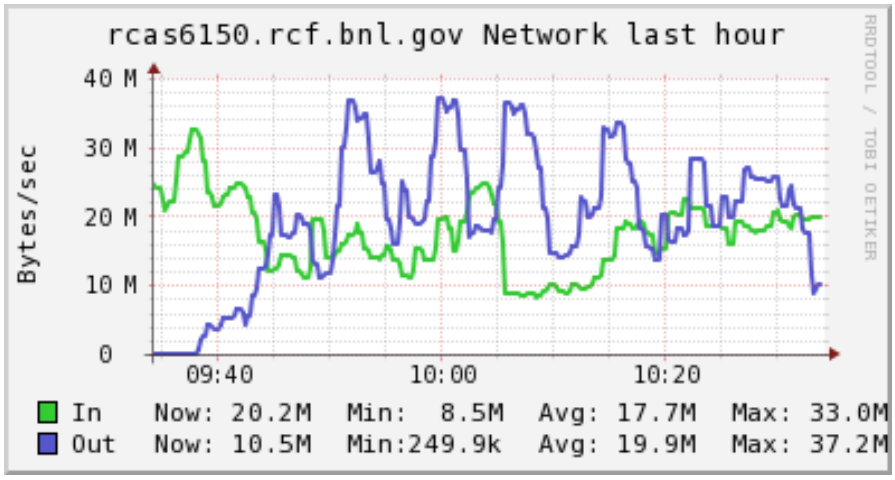


Figure 45. I/O rate for an analysis node narrowed to a peak load.

However, there is concern regarding the growth of I/O rates as shown in Figure 45, a "zoom-in" presentation of peak load activities. The main component of the I/O "out" (in blue) could only be explained by access to the node's local data via Scalla/XrootD access, which shows data going out of the node on the LAN serving other nodes/jobs on the farm. In this example, peaks at about 37 MB/sec are notable (flat I/O rates at about 40 MB/sec during analysis-intensive periods have been observed before this review).

The I/O "out" will be proportional to both the amount of data located on a given node and the number of batch slots across the facility. With a new incoming farm node purchase with four times more data attached to each node, the risk of exceeding the capacity of a 1 GbE line (hence having potentials for lengthy I/O saturations, causing job efficiency issues via I/O starvations) could be possible in the coming year. The need for capacities greater than 1 GbE is an immediate LAN requirement within our distributed data and data flow model. Perhaps the enabling of ROOT/Scalla I/O read-ahead (not done to date) may alleviate this issue (for the most I/O challenging jobs, it is likely to make it worse). Compute nodes with twice the number of cores will not create this demand as far the I/O "in" is concerned but the evolution of core density and storage space will certainly need to be watched and considered on this widely distributed data model as it will impact LAN requirements.

12.6 Key Remote Science Drivers

12.6.1 Instruments and Facilities

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

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

The resources at NERSC/PDSF are shared among many projects and apportioned based on resource allocation cycles. In 2012, STAR had 300 slots of official allocation and 595 slots of actual average usage. This discrepancy is due to the use by STAR of cycles that go unused by other projects.

As a Tier-1 center, NERSC/PDSF also provides permanent archival storage. In our planning, STAR consistently aims to provide space for a full copy of the DSTs in the NERSC/HPSS system. In practice, only a small fraction of the DSTs are moved.

Table 20. Network bandwidth needed by SAC or Tier-1 centers, depending on activities.

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

Table 20 shows the network bandwidth needed for the categories of transfers. The first row, presented in Table 15, is used to evaluate the compounded network load on the facilities holding the data. The second row indicates the network resources needed at a Tier-1 center to be able to transfer all MuDST within a 6-month period. This minimal bandwidth is needed for NERSC/PDSF. The last row assumes that one-third of all SACs take the data from PDSF while two-thirds would be from BNL (fourth row). For completeness, we indicate those network requirements allowing SACs to transfer data from our Tier-1 and Tier-0, respectively. In the case of PDSF, those requirements are not additive (the time frame for transferring the MuDST is quoted as 3 months while the data transfers are estimated as burst transfers over a year period). Typically, the larger of the two numbers is needed as a connection speed from PDSF.

The KISTI Tier-1 center is equipped with 1,000 CPU slots and 150 TB of centralized storage space, with a steady growth planned for the period of our extended MOU (up to 2017, renewable). Another installment of 1,500 CPUs is planned by the end of fall. The CPU growth is foreseen as 500 to 1,000 CPUs/year for the period covered by this report (the exact number will be confirmed by mutual agreement — the final resource plan for KISTI had not been finalized at the time of this review). The facility is rather heavily used and all slots allocated to STAR are typically busy, as showed in Figure 46.

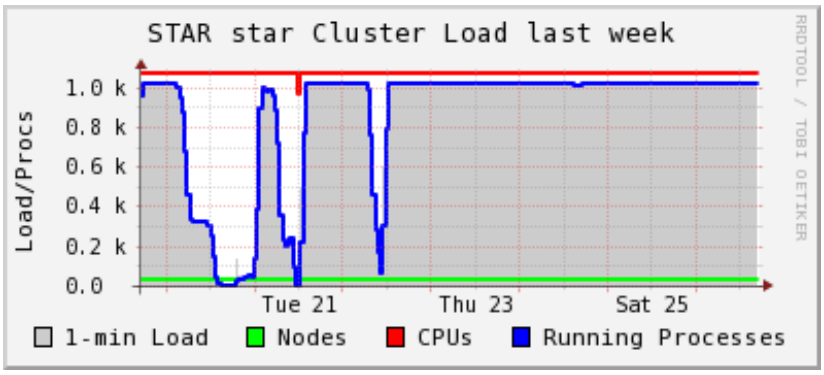


Figure 46. KISTI typical CPU load profile. The offset running/maximum is a monitoring artifact: All nodes, including our databases and Grid gatekeepers, are part of the same graph but do not run jobs.

The site currently and essentially supports the embedding process. A small number of (local) users (from Tsinghua and Macau) also use the resources for user analysis. As noted in our introduction, Section 12.3, the KISTI Tier-1 center is supplied with minimal user support (we have one point of contact from the facility handling all user requests) and hence, we do not envision the growth of user analysis activities beyond the opportunistic use from those who support the embedding data productions at KISTI. KISTI does not have permanent archiving storage and hence, the data produced are either brought back to NERSC or to BNL.

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

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

Due to the rapid growth and available allocation slots at KISTI, STAR computing is considering this site for real data production. Constrained to essentially one pass of data reconstruction per year at BNL (far below acceptable physics objectives and below our planning), the resources at KISTI cannot be overlooked. The site's rapid CPU growth is in fact essentially planned with that objective in mind. Table 21 gives estimates of the network bandwidth needed to allow data production to occur at a remote site (or cloud processing). The first row indicates the bandwidth needed for a 20% data transfer occurring right away during and along with data taking (while a copy is stored in HPSS,

another would be pushed through to the remote site — in collaboration with ESnet, this has been exercised in STAR and shown to be possible in 2009). The second row indicates the additional bandwidth required for bringing the data back to BNL. The third is the sum of the first two, indicating the bandwidth needed in total to/from KISTI. The fourth row is the same global calculation pushing data production of half of the data at KISTI (this would allow restoring at least two production passes within one year — it is our actual target).

Table 21. Network bandwidth requirements necessary for moving DAQ/RAW data from BNL to an arbitrary remote site for remote data processing. Because the result of production must be brought back to BNL, bandwidth is indicated as needed on the BNL side.

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

Other facilities and activities worth noting are:

1. The support of SACs is summarized in Table 20, which indicates on the third row the bandwidth needed for each SAC to be able to use its limited storage and copy datasets (at a 10% level replacement or transfer every 3 weeks) for sustaining local science. The bandwidth numbers indicated there are low when compared to the bandwidth needs at BNL, but they indicate the need for each individual SAC.
2. The possibility of a full copy of all RAW data to a secondary facility for the long-term preservation and safety of STAR data has long been discussed and considered. The bandwidth required for this process is indicated in the last row of Table 21. The possibility of leveraging our current partial data copy away from the Tier-0 center will need to be decided within the next 2 years.

The information in Tables 20 and 21, is a reminder of the uncertainty for year 2016, which will likely see a factor of two times the drop in the network bandwidth requirements.

12.6.2 Software Infrastructure

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

Scalla/XrootD is still in a testing stage at NERSC/PDSF and data access is essentially done via centralized storage at both PDSF and KISTI through NFS/GPFS storage. The Prague site continues its use of a mix of DPM (historical use) and direct NFS access of the data.

Typically, no tools other than our STAR unique framework (relying on ROOT and its adequate site-specific plugins) are needed.

Most sites use the STAR Unified Meta Scheduler (SUMS) for submitting jobs. This tool monitors and records user requests as noted in Section 12.5.2 though, at remote sites, the monitoring capability is often not enabled. The benefit of using SUMS is that similar (or identical) job descriptions can be seamlessly moved between sites for achieving the same results (providing the same datasets are available) regardless of the site's choice of batch system. Most workflows are local (that is, not based on distributed computing, Grid, or cloud processing).

The general user pattern has also included the use of so-called picoDST. Of no specific designed format (but based on ROOT trees), their size is a fraction of those of the MuDST, from one-fifth to one-tenth. The data transfers are handled in a non-organized way in some instances (BNL to PDSF transfers are using Grid tools but transfers between PDSF and China are more ad hoc).

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

12.6.3 Process of Science

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

Grid-based data transfers are used between NERSC/PDSF and BNL. Typically, Globus and globus-url-copy (guc) are used for transfers. Data may be grabbed from XrootD onto an export cache using xrdcp (this load is not significant enough to affect user access to the distributed data at BNL). STAR is equipped with four Grid gatekeepers (two are shared with the OSG general VOs, two are dedicated to STAR-specific use). On the NERSC side, two endpoints may be used for the transfers. Rates of 200 MB/sec would be typical for transfers using guc with 100 MB/sec using Globus but those transfer rates are limited by endpoint capacity. Those rates are sufficient for the 2013 data transfers at low priority, but these rates will likely not suffice for the larger datasets that are expected for 2014 and after.

Data flows to/from KISTI consist of two paths. DAQ files are transferred from BNL using the Fast Data Transfer (FDT) tool and the products of embedding production for

permanent archiving are also brought back to BNL using FDT. The current data rates are 40 MB/sec, not an impressive data transfer rate but sufficient for current need. Should raw data transfers occur, network connectivity and expected speed would need to be revisited — as previously discussed, a 2013 operation would require a 1 Gbps connection while a 2014 operation will require 1.5 Gbps of capacity. Typically, these bandwidths are in place but end-to-end tuning is needed to achieve rates close to full capacity. Embedding results are also copied from KISTI to NERSC/PDSF using guc. Using multiple threads for the transfer (after studying the saturation point), rates of 300 MB/sec are proven to be possible between those two sites.

Data transfers from NERSC/PDSF and/or BNL to Prague are handled using FDT as the underlying transport. Data is also grabbed from BNL/XrootD using xrdcp. Prague has continued onward to consolidate the development of theoretical computing models (based on constraint programming or mixed-integer programming) and the development of data planners to enhance data transfers and leverage the presence of datasets from multiple sources (data sources as well as sites) for the most efficient data transfers to a destination. We already showed, reported, and published that the use of such techniques has the potential to reduce data transfer makespan by 30%. Recent work focused on the use of local data caches and best space reclamation strategies (based on user access and data-demand pattern). All work has been carried out by graduate-level computer science students. We feel that within a year or two, a fully optimized system will be complete for STAR use, factoring in multiple sources for dataset provenance, network bandwidth, and availability and cache optimization.

12.7 Local Science Drivers — the Next 2–5 Years

12.7.1 Instruments and Facilities

With the next 2–5 years, STAR's focus will be on Phase I of the program, i.e., the heavy flavor and di-lepton measurement (and the study of sQPG properties). Detector upgrades and making the challenging datasets (especially those taken by the HFT) a success is a high priority.

12.7.2 Software Infrastructure

No major software infrastructure changes that may affect the network requirements are expected. STAR computing will, however, go through dramatic changes and upgrades, including (1) the onset of new track reconstruction software; (2) a new metadata collection facility online (based on the Advanced Message Queuing Protocol or AMQP), which will completely replace the old system (direct MySQL access) and will be in effect in 2014; and (c) a strong push toward moving computational resources closer to the experimental device (HLT track reconstruction and vertexing).

Enhancement of our STAR FileCatalog will be needed to support increased operations as well as data accumulation. Spanning more than a decade of data taking, advanced queries for comparative identification of datasets will be needed. We have not consistently cataloged the embedding datasets, essentially relying on the records of our

simulation and embedding request tracker. This has caused some issues related to the fast identification and location of possible viable past embedding processing. This is an organizational item only and in the past year, workflows have more consistently brought the data samples back to BNL, where they are cataloged by the local workforce (automation should be in place by next year).

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

12.7.3 Process of Science

STAR does not foresee a dramatic change in the process of science within this time frame. Any changes will indirectly affect the network requirements with the exception of HLT-based vertexing, which will reduce the size of the massive datasets expected in 2016 by better selecting the events of interest.

The STAR collaboration has not yet tested or applied any data reduction algorithms at the source, as this is a high-risk process. Dropping any potential events of interest is irreversible and more studies would need to be performed before this type of workflow is considered. However, one advantage of pursuing this path may be that dataset sizes can be reduced by 40%. A method like this may develop naturally, however – as more and more computing power is moved online for HLT purposes, early event transformation will be possible.

Likely activites, developments, and research over the next 2–5 years include:

- Focus on real-time decision-making filters (HLT, pattern recognition),
- Data reduction and repacking methods (fast online tracking, pile-up rejection at the source for data reduction),
- Moving detector calibration processes closer to the data taking so that real-time first-pass track reconstruction in HLT and collision vertex reconstruction can improve decision-making.

12.8 Remote Science Drivers — the Next 2–5 Years

12.8.1 Instruments and Facilities

We have already described our upgrade plans, schedule, and main facilities. The collaboration's aggregate network requirements are listed in Table 22. From the previous network requirements shown in Tables 20 and 21, most of the network bandwidth calculations in Table 22 represent the maximum bandwidth requirements, as data transfers are not continuous throughout the year (i.e., some transfers are bursty, and others are consistent for a 6-month duration).

Table 22. Summary table for all network requirements.

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

The key components will be:

- Each SAC will need less than 2 Gbps of network bandwidth for the time period envisioned on Table 22, Row 1.
- To sustain operations at NERSC/PDSF, network rates of about 3 Gbps will be needed at NERSC — this is shown on Row 2 (we again purposely ignore the 2016 estimate with caution; 4 Gbps would be workable in any scenario).
- The collaboration is working towards a faster pace use of the KISTI facility with a strong push toward real-data processing — we plan for a facility growth able to consume data transfers up to a 50% level; the required bandwidth as a function of years is shown in Table 21. Those rates are shown on Row 3 and are unlikely to exceed the 5–6 Gbps rate.
- To sustain both operations, BNL connectivity will need to be at levels consistent with Scenario B (Row 5).
- Depending on how critical data preservation is to another site (and to the extent possible), the required bandwidth from BNL would be as shown on Row 6, Scenario C.

12.8.2 Software Infrastructure

The only fundamental changes STAR can foresee within the 2–5 year period are the possible exploitation of hybrid cloud/Grid infrastructure on two fronts:

- The online computational resources are in the process of being “cloudified” and would be used for additional processing on site.
- With a 2-year time frame, it is highly probably that operations at KISTI will be carried on a cloud basis.

Those changes will not alter STAR’s current network requirements.

12.8.3 Process of Science

The decrease to a minimum of three SACs is anticipated as new resources at major centers come online.

To date, STAR has not made use of a global XrootD namespace and global redirector. It may be that XrootD capabilities are leveraged within the next 5 years if we gain a better understading of the data movement scheduling.

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

12.9 Beyond 5 Years — Future Needs and Scientific Direction

STAR upgrades are expected to advance with the possible advent, by 2018, of the iTPC. Section 12.4 describes the growth in event size that this upgrade could cause. The full impact of the event size increase may not manifest until the 2019 runs because of the species currently planned for the 2018 and 2019 runs.

It is likely that KISTI will remain a part of the STAR collaboration, and that connectivity to Asia will continue to be very important for STAR.

It is possible that additional scientists and institutions will join the STAR collaboration in the eSTAR era, depending on the type of experiments run in the eSTAR era.

Before the eSTAR era, STAR's computing frameworks will need to be refactored or refreshed to take better advantage of recent technological innovations, including better multicore support, asynchronous I/O, and MQ-like communications.

12.10 Network and Data Architecture

Better connectivity to Asia is critical to STAR's science productivity. While bandwidth to KISTI has improved (and ESnet has helped greatly in the past), the connection to China remains problematic; the connections are too slow and intermittent to carry out decent remote work.

In the interim, STAR institutions in Asia have reported that the use of remote persistent session and tools such as NX (Desktop Virtualization and Remote Access Management) are helpful and convenient.

12.11 Collaboration tools

For collaborative tools and video services, the STAR collaboration needs standard phone bridge and videoconferencing capabilities (with slide display essentially).

The RHIC collaborations have maintained a paid subscription to the SeeVogh Research Network (SRN), the successor of EVO services. This service has proven to be useful and cost-effective.

Skype is still in use for daily communication among STAR collaborators.

12.12 Data, Workflow, Middleware Tools, and Services

The availability of predictive and/or advanced network reservation capabilities would be of a benefit for planners and data movement schedulers. A joint effort between ESnet and STAR personnel could explore the benefits of sharing bandwidth between multiple consumers.

12.13 Outstanding Issues

The slow adoption of cloud computing (even at the conceptual level) may be the only issue STAR sees in the U.S.-based distributed computing consortiums. There are some positive signs this may change within the next 2 years and a collective program may emerge but planning for cloud-based resources (from the OSG, for example) within a 2–5 year time frame appears uncertain.

STAR does not have other outstanding issues but notes the tremendous benefit of the OSG support center in reporting problems to us (as per our grid infrastructure) and facilitating communications between teams through a much-improved and enhanced ticket system. The transition to the new OSG CA has been very smooth — a great job overall and also a much-improved process for acquiring a certificate via OIM (OSG Information Management).

12.14 Summary Table

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

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
2-5 years				
• End of Phase-I physics program, beginning of Phase II program. • RHIC/STAR data taking of large samples — BES Phase II, QCD critical point and study the QCD phase structure. • Online HLT event vertexing, possible event filtering and reduction. • iTPC in 2018, eCooling. • Onset of "lego-block" processing (workflows seamlessly running online/offline for calibration – adapters, MQ based I/O). • Cloudified cluster online a "standard" + full use of KISTI.	• Data flows remain near identical. • Elastic computing (cloud resources may "join" a pool for embedding + KISTI). • Possible move of PDSF to Carver-like platforms. • Situations for transfer of datasets off Tier-0 clarified.	• Uncertainty in 2016 data size. • Overall similar datasets up to 2019.	• Similar bandwidth needs to/from the same endpoints. • SAC profile unknown (changes certain within 2 years/will need reassessing). • KISTI connectivity @ 5–6 Gbps. • BNL with a 5+ Gbps pipe (data archiving plan influence).	• Similar time frames and peers. • Possible reshape of the SAC landscape. • Possible use of opportunistic cloud resources (at lower levels) – OSG/cloud?
5+ years				
• Heavy flavor program and B-physics + eSTAR by 2024. • Lego-block frameworks with async I/O + filter / repack capabilities (MQ framework-like) likely.	• Similar landscape foreseen. • Predictions beyond 2020 unclear.	• Expecting similar datasets.	• No changes forecasted.	• Peering is unclear but likely the same until 2020.

13 RHIC Computing Facility (RCF)

13.1 Background

Located at BNL, the Relativistic Heavy Ion Collider (RHIC) program is an NP 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 the world leader in the scientific quest to understand how mass and spin combine into a coherent picture of the fundamental building blocks nature uses for atomic nuclei. It is also providing a unique insight into how quarks and gluons behaved collectively at the very first moment our universe was born. The main RHIC experiments, PHENIX (550 physicists from 75 institutions spread over 15 countries) and STAR (580 physicists from 59 institutions spread over 12 countries), are collaborations spanning many countries and more than a thousand collaborators.

Having reached petabyte-scale data recording per year (10^{12} bytes), the RHIC experiment envisions that its aggregate RAW data rate per Run (or year) will more than triple from the current about 1 PB to greater than 3 PB per experiment in 2014 and 2015, reaching an archival data rate of almost 2 GB/sec per experiment (i.e., STAR), making data management and data distribution an ever-increasing challenge. To face the challenges caused by the size of those datasets while preserving 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.

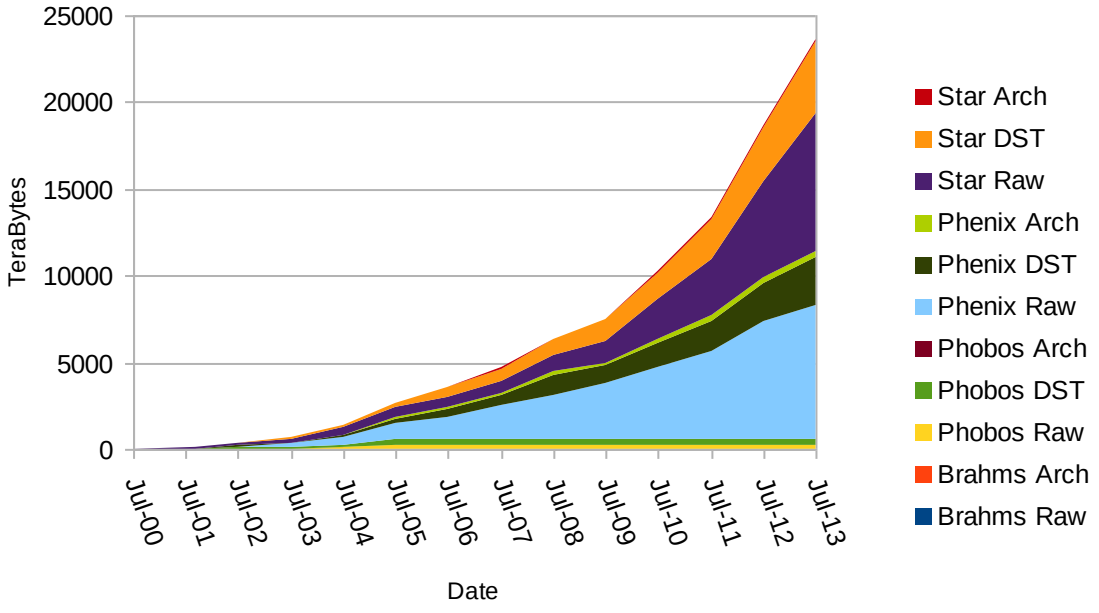


Figure 47. Data accumulated and archived by the RHIC experiments from 2000–2013.

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

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 are the aspects associated with the handling and processing of the actual data produced by the detectors. Other aspects of the RHIC computing requirements, 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 financial 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 the data produced by the experiments, and operates in conjunction with computing facilities at remote locations, therefore requiring decent 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. The RCF serves as a data-mining and -serving facility for the raw and derived data and functions as the primary analysis facility. Remote sites not only manage and process large amounts of data, but also do large-scale theoretical modeling and event simulation. Some datasets from simulation are stored at RCF, and there is some use of the RCF for simulation work during periods of reduced demand for collider data processing. The RCF exports data, which has received various levels of processing, to remote facilities for later-stage analysis as well.

The BNL campus network provides high-performance network connectivity that supports many worldwide scientific disciplines. Main users of the network capabilities are PHENIX and STAR at the RHIC and ATLAS at the 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 seven 10-gigabit circuits divided into two distinct classes of service for the user community. First, general-purpose IP connectivity is provided by two 10-gigabit links to the Internet via ESnet. These links provide the default connectivity between most external scientific facilities and BNL. Secondly, ESnet provides five 10-gigabit links that primarily support the Science Data Network (SDN) bandwidth requirements between BNL and the LHC Tier-0 center at CERN and Tier-1 sites around the globe, and between BNL and four of the five U.S. ATLAS Tier-2 centers at universities and at SLAC. As to NP applications on these links, there is a 1 Gbps 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 10-gigabit 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.

This status update focuses on major upgrades and enhancements since the previous report from 2011.

13.2 Key Local Science Drivers

13.2.1 Instruments and Facilities

The RCF at BNL provides the majority of computing power (90% for PHENIX, 85% for STAR) 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 supporting high energy and nuclear physics experiments. As to network connectivity, the RCF uses ESnet, which is peering with other domestic and international R&E and commercial network

By the end of 2013, the RCF will have more than 13 PB of disk space in production and 220 kHS02 of processing power (we measure processing resources in thousands of HepSpec 2006 [kHS06], which is based on SpecInt 2006). As to the archival storage volume, we expect that to grow to more than 25 PB. Particularly challenging will be the

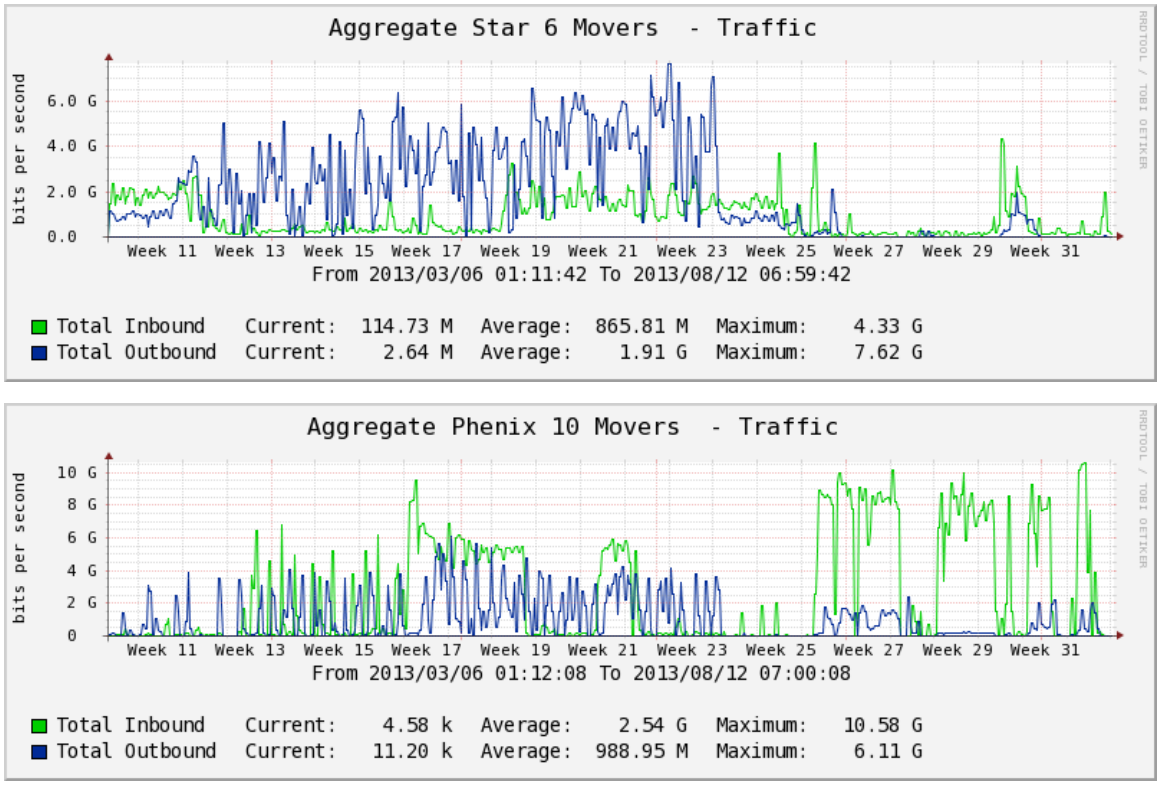


Figure 48. Mass storage (HPSS) I/O for PHENIX and STAR (last 6 months; I/O from the network switch perspective connecting to the HPSS movers).

run in 2016, when STAR plans to accumulate a Raw data volume of more than 6 PB that will be augmented by more than 4 PB of processed/DST data; meaning STAR anticipates adding more than 10 PB of data to the tape archive that year.

At the 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 (40–120 Gbps).

Extensive upgrades to the internal BNL networking infrastructure were conducted over the course of the past few years. Most of the interswitch links were upgraded to 100-gigabit ether channel.

13.2.2 Process of Science:

Many well-defined computing functions are 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 is recorded and stored in the expectation that it will be frequently accessed and that the ease and speed of access is of critical importance. Large-scale datasets are recorded where produced. Thus, the raw detector data and data derived from the reconstruction pass are recorded at the RCF. The primary output of the reconstruction pass (historically called DST-level data), which requires more frequent and immediate access, 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 generally performed at the RCF (STAR sent a fraction of the Run 11 data to cloud resources at NERSC for Fast Offline QA processing). 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, the RCF is not sized to perform the reconstruction of simulated events in parallel with the reconstruction of collider data.

Physics Modeling. In order 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 the 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 being 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.

Micro-DST Production. The production of a micro-DST is most generally accomplished by making a pass through a DST dataset, 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. Micro-DST 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. The RCF is (intended to be) the primary site for such micro-DST 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 micro-DST production capability for some types of data.

It is also possible to produce additional micro-DSTs from existing micro-DSTs. This is frequently the case in constructing final very selective datasets.

Often the final very selective summary of the data is in the form of an ntuple. The 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 dataset has been identified, the analysis process of studying the physics significance of the data is typically performed by repetitive passes through the dataset. 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 dataset and the scale of the computations required, these needs may range from those that can be satisfied on an inexpensive workstation to those that require a large facility with parallel coordinated operations across many processors operating on large datasets distributed across many disks. The 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 terabytes) 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 the RCF and PDSF.

13.3 Key Remote Science Drivers

13.3.1 Instruments and Facilities:

When looking at the various steps involved in the process of getting from RAW data to physics results, there are two that involve resources external to RCF: event simulation and, to some extent, user analysis. Applicable in particular for STAR, we estimate that the resources (both storage and processing) needed for handling the MC 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 (MC) 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 OSG). Using Grid or cloud interfaces makes resources available to STAR at various sites seamlessly and interchangeably.

While the PHENIX experiment is managing and running almost all its user analysis at its RCF share using AnaTrain, STAR's high-priority data production has pushed analysis aside, 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 the RCF and PDSF) was 40% of what is necessary for one analysis pass. To serve the wide area bandwidth needs from the RCF to the three to five STAR Tier-2 centers, between 2 Gbps and 5 Gbps of network capacity is needed, depending on the number of Tier-2 centers and the run scenario in a particular year.

Currently, BNL is serviced by a total of seven 10-gigabit links 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 strung along the South Shore. In the event of a circuit, router, optical component, 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 have been configured with redundant secondary interfaces and multiple Border Gateway Protocol (BGP) peerings that can detect the most common failures and reroute around the defective components almost instantaneously and transparently to the applications.

Both BNL and ESnet staff have completed the deployment of a "dark fiber" solution in 2011 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

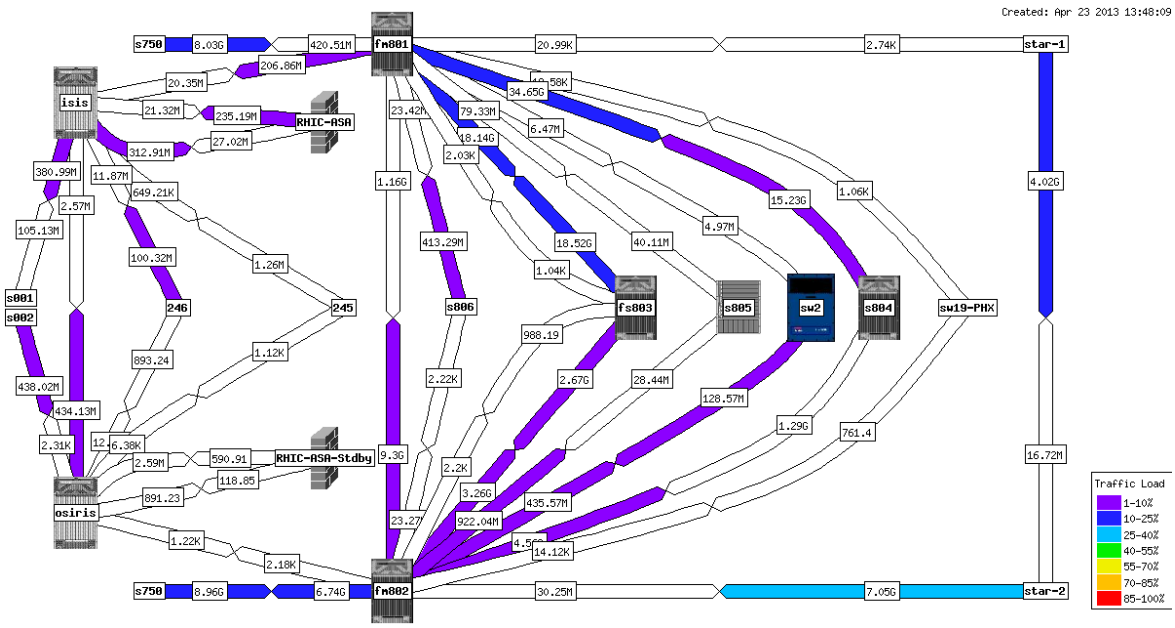


Figure 49. RHIC network current state.

locations in Manhattan. As currently configured, the optical switch gear (Infinera) provisioned for the fiber deployment can support up to 100 gigabits per second.

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 increasing use of ESnet's On-Demand Circuits and Advance Reservation System (OSCARS).

Each of the existing six circuits has been allocated between 10 and 1 gigabit (minimum) bandwidth, the latter with oversubscription capability for the idle bandwidth on the circuit.

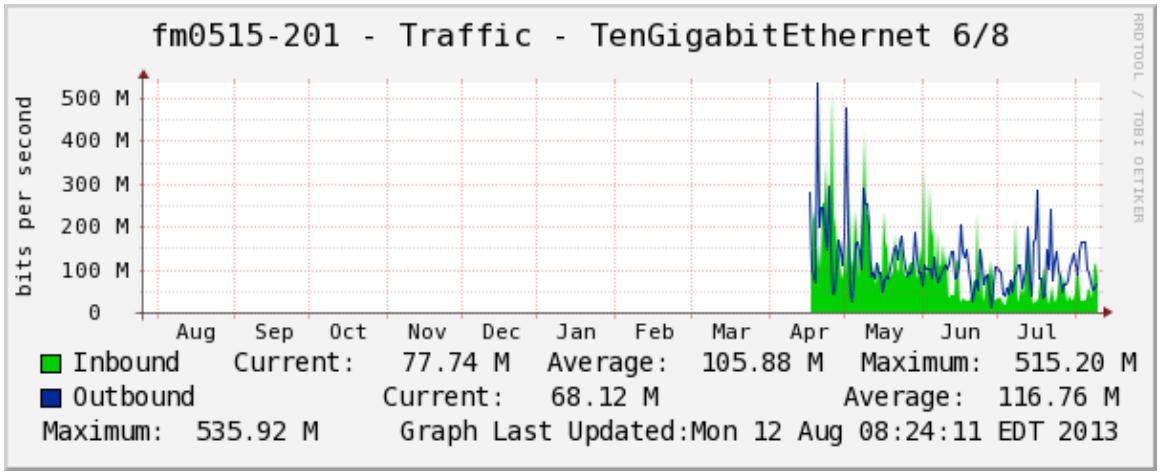


Figure 50. RHIC WAN traffic April–August 2013.

13.3.2 Process of Science

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

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

13.4 Local Science Drivers — the Next 2–5 Years

13.4.1 Instruments and Facilities

Table 23. Proposed scientific milestones for future RHIC Runs

	Year	#	Milestone
spin	2013	HP8	Measure flavor-identified q and anti-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.

During the next 2-5+ 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 VTX detector at PHENIX in Run 11 doubled the event size from Run 10 and increased significantly again with the introduction of the FVTX in Run 12), but the operating models of the experiments that have been exercised in the past year will be recognizable in the next 2-5 years. Most of the increases in facility capacity for processing, disk storage, and archival storage will come from technology improvements, while maintaining a similar facility complexity.

Table 24. Proposed RHIC run scenarios and science goals.

Years	Beam Species and Energies	Science Goals	New Systems Commissioned
2013	• 500 GeV pol p+p	• Sea quark and gluon polarization	• upgraded pol'd source • STAR HFT test
2014	• 200 GeV Au+Au • 15 GeV Au+Au • Fixed Au target test	• Heavy flavor flow, energy loss, thermalization, etc. • Quarkonium studies • QCD critical point search	• Electron lenses • 56 MHz SRF • full STAR HFT • STAR MTD
2015-2016	• p+p at 200 GeV • p+Au, d+Au, ^{3}He+Au at 200 GeV • High statistics Au+Au	• Extract η/s(T) + constrain initial quantum fluctuations • More heavy flavor studies • Sphaleron tests	• PHENIX MPC-EX • Coherent electron cooling test
2017	• No Run		• Electron cooling upgrade
2018-2019	• 5-20 GeV Au+Au (BES-2)	Search for QCD critical point and deconfinement onset	• STAR ITPC upgrade
2020	• No Run		
2021-2022	• Long 200 GeV Au+Au w/ upgraded detectors • p+p/d+Au at 200 GeV	• Jet, di-jet, γ-jet probes of parton transport and energy loss mechanism • Color screening for different QQ states	• sPHENIX
2023-24	• No Runs		Transition to eRHIC

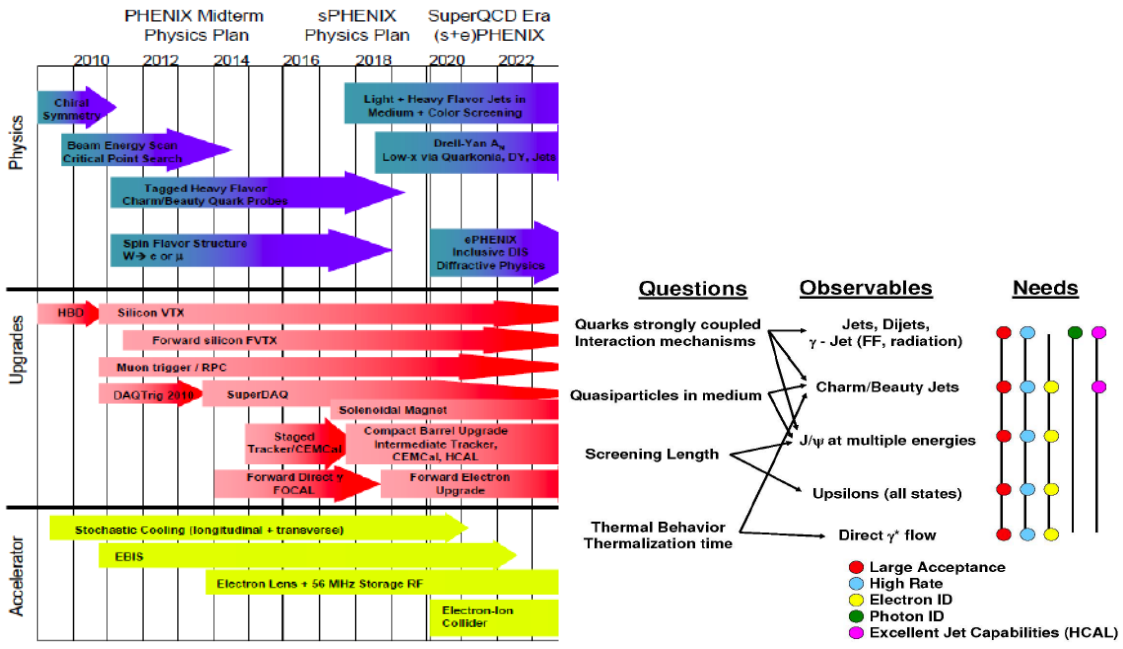


Figure 51. PHENIX decadal plan, physics questions, and needs (STAR has an equivalent plan).

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 2014 at 100 GeV/nucleon in Heavy Ion (HI) (Au-Au) 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.

As to the experiments, PHENIX has offloaded the RCF from p-p reconstruction of 270 TB of RAW data in 2005 by replicating it to the computer center of its Japanese collaborators at CCJ. Given the number of actual events in recent runs, the expected number of events in future runs, the reconstruction times per event, and the actually available and expected compute capacity at the RCF, the collaboration has at this point no plans to ship RAW data files off site. A few collaborating institutes, primarily CCJ, are asking for replicas of the smaller (about 70% of the RAW) derived DST datasets. As to PHENIX, RHIC runs which have a substantial p-p component have a greater impact on wide area networking.

13.4.2 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 begin a new era of fundamental HI and spin studies of extended scientific reach. These studies will build on the discoveries of the first phase of RHIC experimentation by utilizing the increased luminosity provided by the upgraded RHIC II accelerator and by 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 that the upgraded detectors will produce.

13.5 Remote Science Drivers — the Next 2–5 Years

13.5.1 Instruments and Facilities:

At the RCF, both collaborations will produce large samples when the data collected with increased RHIC machine luminosity and upgraded detectors is processed. The larger data products will need to be distributed for analysis. The samples selected by physics groups to be served to analysis centers (Tier-2 centers 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 Tier-1 and RCF to analysis centers will increase.

13.5.2 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 actions similar to what 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.

13.6 Beyond 5 Years — Future Needs and Scientific Direction

We expect similar requirements as described for the 2–5 year period.

Note the projections for wide area bandwidth for PHENIX and STAR are very different. Based on experience gained in previous years, PHENIX users typically transfer about 10% of the RAW data volume taken in a run (300 TB in 2014, 150 TB in 2015) from BNL to several institutions in the United States, Europe, and Japan. And PHENIX is planning to start simulation projects in preparation of a possible sPHENIX detector. The data volume generated by several OSG sites will be a total of 24 TB per project, which is estimated to last 2 weeks.

The PHENIX summary table is in the PHENIX case study (Section 11.7).

	FY11	FY12	FY13	FY14	FY15	FY16	FY17	FY18
STAR Data (TB/year)	1930	3071	3040	4835	5216	10352	2784	2439
PHENIX Data (TB/year)	1271	2192	2216	4000	2000	-	-	-
Total Annual Data (TB/year)	3201	5263	5256	8835	7216	10352	2784	2439
Required WAN Bandwidth (avg) (Mbps)	276	1500	<10 k	<10 k	<10 k	<15 k	<5 k	<10 k

The STAR summary table is in the STAR case study (Section 12.14).

Table 25. PHENIX and STAR projected dataset volume and estimated WAN needs.

14 Thomas Jefferson National Accelerator Facility

14.1 Background

Thomas Jefferson National Accelerator Facility (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 in free-electron lasers (FELs) based on accelerator technology developed at the laboratory.

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

JLab is a user facility offering capabilities that are unique worldwide for an international community of nearly 1,400 active users. One-third of all Ph.D.s granted in nuclear physics in the United States are based on JLab research (444 granted, 186 more in progress).

14.2 Key Local Science Drivers

14.2.1 Instruments and Facilities

The Continuous Electron Beam Accelerator Facility (CEBAF) at JLab is being upgraded to provide a high-luminosity electron beam of up to 12 GeV to four halls. Hall B holds the CLAS (CEBAF Large Acceptance Spectrometer) detector; Halls A and C hold a variety of spectrometers that can be configured to the needs of a particular experiment, and Hall D holds the new GlueX detector. Commissioning of the upgraded accelerator and the detectors will begin in FY 2014.

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 μm, respectively, and will, after hardware upgrades, produce 1,000 watts in the ultraviolet range and more than 100 W in the terahertz range. This instrument is being further developed, both to extend its capabilities and to exploit it for science.

JLab is one of three sites (with BNL and Fermilab) hosting a distributed Lattice QCD Computing Facility consisting of 10–100 teraflop/s class clusters tuned to the computing requirements of Lattice QCD (LQCD).

14.2.2 Process of Science

For the Experimental Nuclear Physics Program in the four halls, data will be acquired in one of the two countinghouses, monitored live, and transferred to the computer center to be written to tape in files of size up to 20 GB, typically up to 30 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, which 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 event summary files. Most experiments only transfer the smaller files off site, 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 carried out at remote institutions and a larger fraction done at lower priority on the batch farm. Most simulation data is produced at JLab, and most is 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 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, Fermilab, and JLab, with typical sizes up to 1,024 cores or up to 16 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 JLab are currently in the few hundred megabytes to the 20 gigabyte range, and will grow larger as access to larger supercomputers allows for generating finer lattices. File transfers among the sites are sporadic, and can be multiple terabytes. LQCD will not be a bandwidth driver for the laboratory.

14.3 Key Remote Science Drivers

14.3.1 Instruments and Facilities

Most of the experimental physics data is acquired and analyzed at JLab and therefore 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 modest number of large files. Bursts of inbound traffic are probably correlated with transfers of LQCD files from supercomputing centers. (See network traffic graphs in Figures 52 and 53.)

14.3.2 Process of Science

With a staff of about 800 and a user base of nearly 1,400 researchers, there is considerable conventional use of networking (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 operation of the many collaborations at JLab.

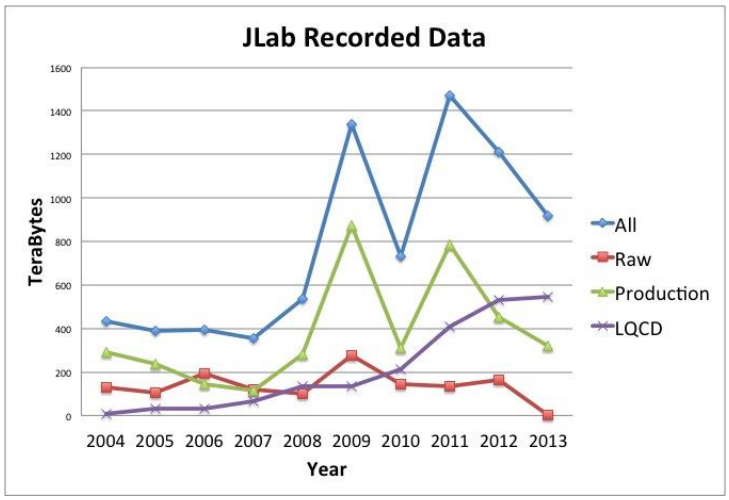
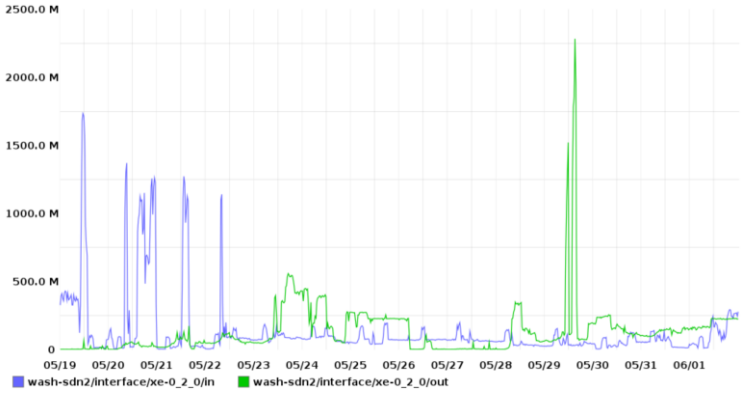
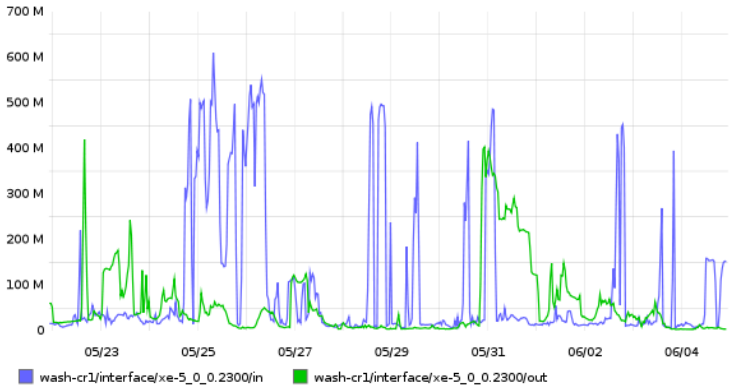


Figure 52. Data volume into the tape library; production refers to first-pass analysis.



(a)



(b)

Figure 53. WAN traffic comparison for a two-week period; (a) 2013 May 19 and (b) 2011 May 22 (last case study).

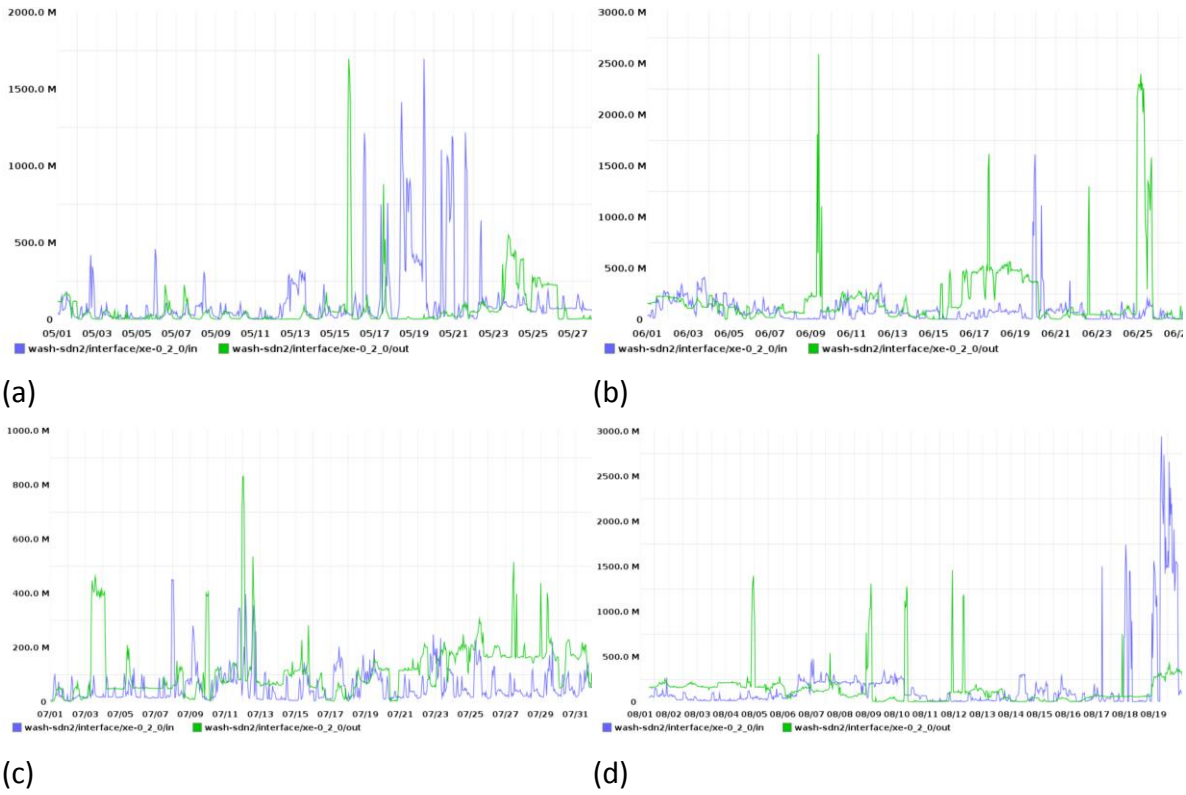


Figure 54. WAN traffic from (a) 2013 May, (b) 2013 June, (c) 2013 July, and (d) 2013 August.

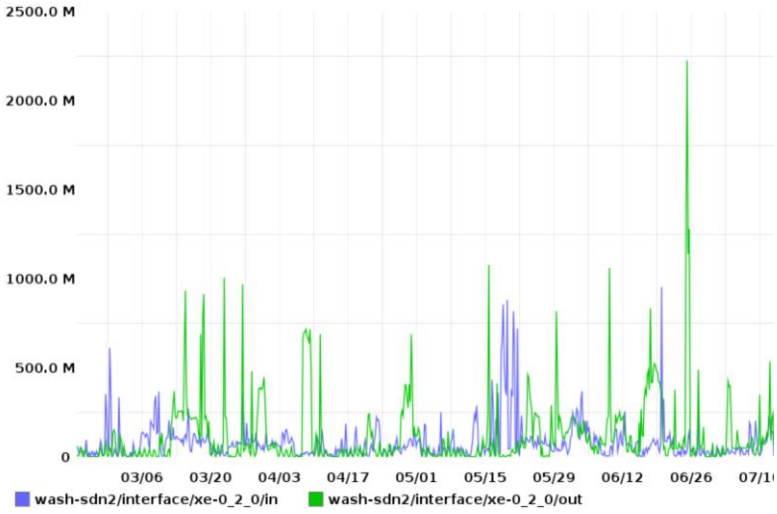


Figure 55. Six months of WAN traffic from 2013 March 1 through 2013 August 19.

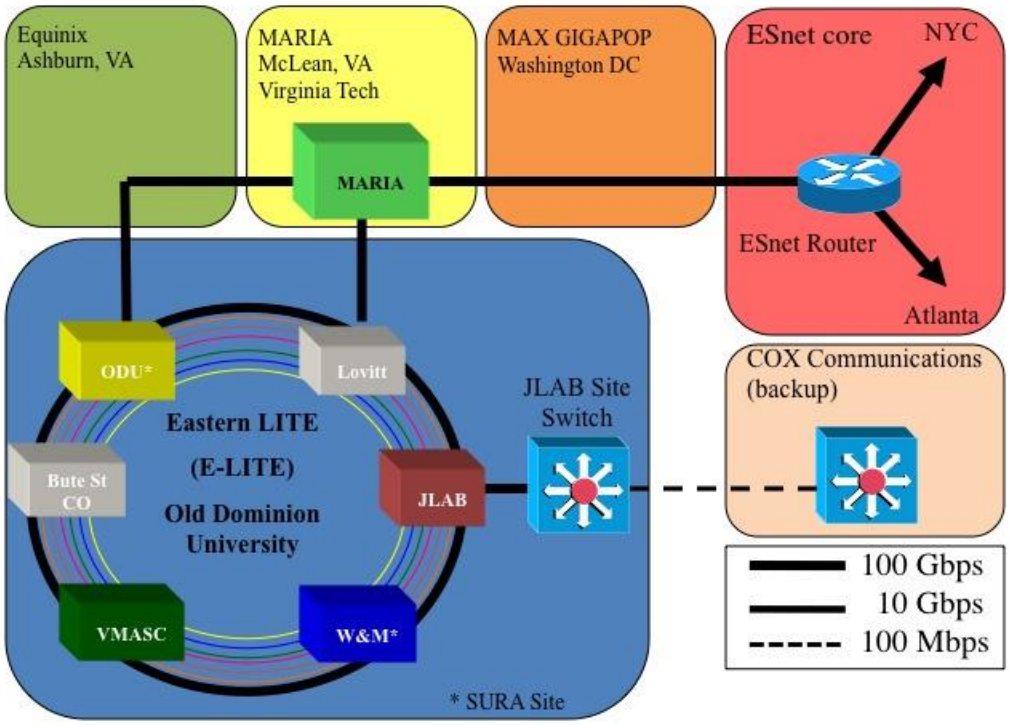


Figure 56. JLab's current WAN connection via the E-LITE MAN.

JLab has benefitted 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 (JLab'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 Mid-Atlantic Research Infrastructure Alliance (MARIA), where ESnet has a presence. MARIA was formerly known as MATP (Mid-Atlantic Terascale Partnership) and JLab's membership in MATP was funded by SURA.

In 2011, E-LITE added an alternate 10 Gbps link to MARIA services from ODU via Equinix in Ashburn, Virginia. The ESnet VLANs to JLab fail over to this alternate path if the primary VORTEX link goes down. However, JLab's connectivity to ESnet at MARIA is a single point of failure.

Both MARIA and E-LITE are developing plans to shut down the VORTEX link and move MARIA's presence to Atlanta. E-LITE plans to maintain its 10 Gbps connection to Ashburn and to add a 10 Gbps connection to MARIA in Atlanta. ESnet has a presence at the Equinix facility in Ashburn and should look into providing JLab connectivity there before MARIA and VORTEX vacate their facility in McLean, Virginia. Additionally, ESnet has a

presence in Atlanta and should connect to the MARIA presence there once it has moved. This would removed a single point of failure provide by providing JLab two divergent paths to ESnet.

14.4 Local Science Drivers — the Next 2–5 Years

14.4.1 Instruments and Facilities

The laboratory is nearing completion of 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 have been installed, and an additional magnet arc added to recirculate the beam for one final pass through the north 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 original halls have been upgraded to receive the new 5-pass, 11 GeV beam.

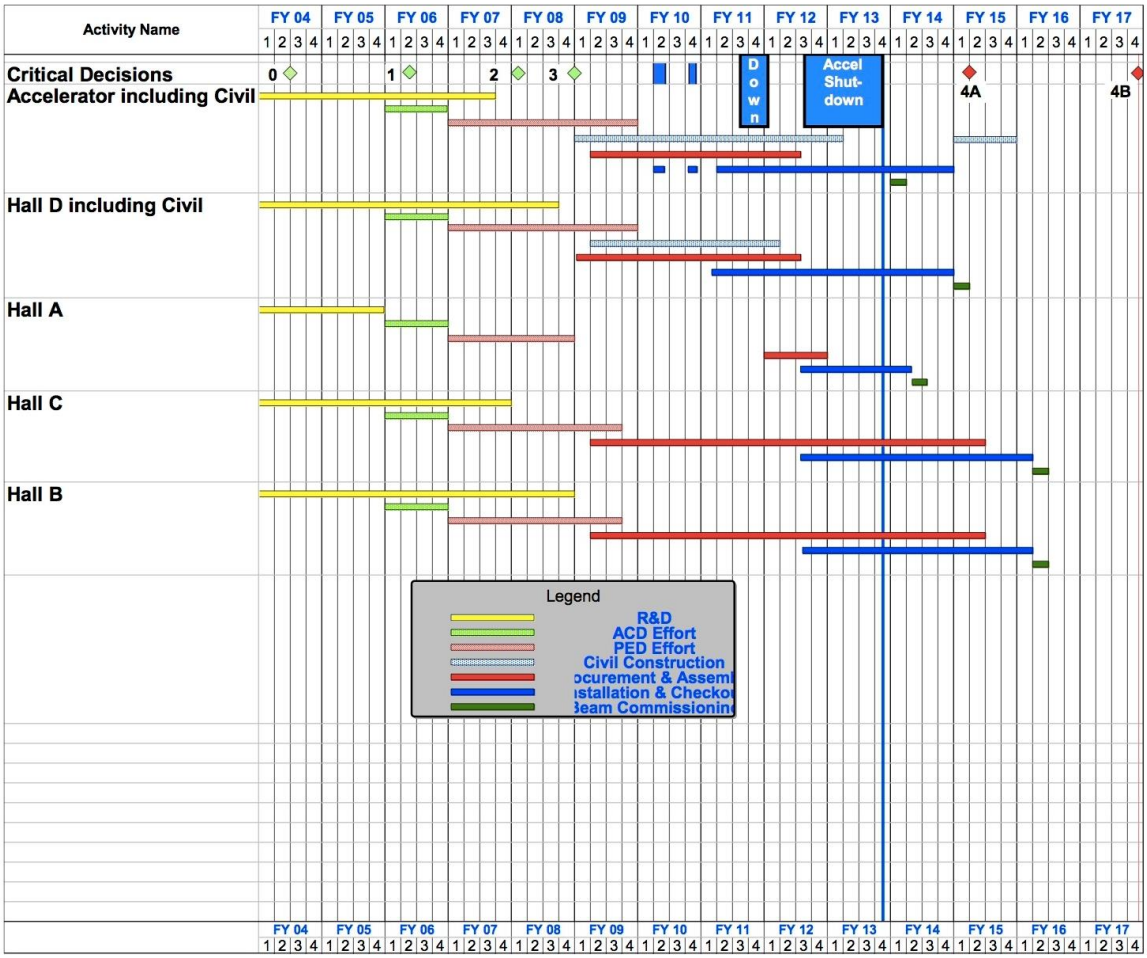


Figure 57. 12 GeV upgrade schedule. Physics resumes in 2015.

14.4.2 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 only comparable to 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 JLab to continue a simple and 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. In the 2018 to 2020 time frame, a new detector in Hall A may bring it up to the same level. The following spreadsheet in Table 26 contains summary numbers for computational and data volume requirements for each hall.

Table 26. Storage and computing requirements for Halls A,B,C,D. Projects assume full running in 2015. Hall D requirements pre-operations (simulation) are still being developed.

Cores	2013	2014	2015	2016	2017
A	65	8	14	17	17
B	975	33	33	2348	4227
C	182	0	10	14	14
D	91	1000	5000	10000	10000
Total	1300	1041	5057	12379	14258
Disk (TB)					
A		12	30	127	
B		6	6	272	
C		0	27	41	
D		150	720	1970	
Total		168	783	2410	
Tape (PB/y)					
A		0.08	0.24	1	
B		0	0	6.7	
C		0	0	1.4	
D		0.8	4	8	
Total		0.88	4.23	17.1	

14.5 Remote Science Drivers — the Next 2–5 Years

14.5.1 Instruments and Facilities

A good estimate of JLab WAN requirements is that it will scale like data volume. However, with a mostly central computing model with somewhat modest requirements, this overestimates the networking requirements.

Data rates will remain constant or decrease between now and the beginning of production running in late 2015, thus the current 10-Gbps WAN will remain more than adequate in that time frame. In 2016, as the 12 GeV science program grows, requirements might grow beyond 10 Gbps.

The LQCD Computing Facility should also grow only modestly in the next 5 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, although burst traffic will grow in volume somewhat slower than Moore's law.

14.5.2 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.

Redundancy in the WAN to ensure the resiliency of JLab's 10-Gpbs connectivity to ESnet is the priority. Figure 53 shows increases to both JLab's business and scientific WAN traffic for the same period of time. Both fit within the existing 10-Gbps connection, but at times the day-to-day business traffic exceeds the laboratory's 100-Mbps backup connection. This makes the resiliency the ESnet connection all the more important.

The increased dependency on remote access and various Internet services, including the use of cloud services, has made the WAN all the more critical to conducting the business operations of the laboratory. While the scientific program at the laboratory can survive with 99.9% availability, business operations require 99.99% availability.

14.6 Beyond 5 Years — Future Needs and Scientific Direction

In addition to the 12 GeV program described above, JLab is exploring other uses of its leadership SRF (superconducting radiofrequency) technology that will likely lead to support for a number of SC accelerator projects at multiple locations (e.g., the Facility for Rare Isotope Beams [FRIB], International Linear Collider [ILC], Project X, Spallation Neutron Source [SNS II], eRHIC, etc.) and could potentially lead to additional facilities on the campus such as an Electron Ion Collider (ELIC at JLab) or a new fourth-generation light source based on an FEL.

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

14.7 Middleware Tools and Services

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

The laboratory offers Globus and other data transfer tools. No computational grid is currently planned.

JLab currently makes use of ESnet's Collaboration Services for audio, Web, and videoconferencing. Videoconferencing continues to grow, and support for robust, easy-to-use tools is essential. 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.

JLab is also expects to expand its use of Federated Identity services and InCommon services to authenticate collaborators in the 12-GeV era.

14.8 Outstanding Issues

None at this time.

14.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• 6-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)

15 Heavy Photon Search

15.1 Background

The Heavy Photon Search Group (http://arxiv.org/abs/1301.1103) at SLAC is collaborating with physicists at JLab, Fermilab, University of New Hampshire (UNH), the University of Orsay, and UC Santa Cruz on the Heavy Photon Search (HPS) experiment, which is aimed at discovering a hidden sector heavy photon. Such a particle would have mass in the range 0.1 to 1.0 GeV/c^2, couple weakly to electrons, and decay into electrons and positrons (e+ e-). It would be produced by electron bremsstrahlung on a heavy target, and be identified as a narrow e+/e- resonance.

15.2 Collaborators

The list of collaborating institutions includes JLab, SLAC, Ohio University, Santa Cruz Institute for Particle Physics (SCIPP) at UC Santa Cruz, SUNY Stonybrook, and UNH in the United states, with international collaborators at University of Orsay in France and the Perimeter Institute in Canada.

15.3 Key Local and Remote Science Drivers

15.3.1 Instruments and Facilities

The HPS detector will be located in Hall B of JLab, where the test detector is now located. We are and will be using the Hall B scientific computing cluster and online system. The vertex detector is being constructed at SLAC and UC Santa Cruz; and the readout system is being developed in collaboration with JLab.

15.3.2 Software Infrastructure

The main HPS data analysis software is built onto the org.lcsim framework, a set of software tools written in Java originally for detector studies for the International Linear Collider (ILC). The detector simulation is done with the Simulator for the Linear Collider (SLIC), a GEANT4-based MC simulation that allows for a very flexible geometry setup that is identical to the geometry used by the analysis software. The MC output or the actual raw data are analyzed for tracks and particle identification using a dedicated reconstruction code written using the org.lcsim framework. The output is in the Linear Collider I/O (LCIO) format and can be further analyzed for physics signals directly or by transforming it to a set of ROOT-based data summary tapes (DSTs).

The processing of the raw data is expected to occur at JLab. Only data summaries of events satisfying preselection criteria for targeted analyses will be exported to remote sites. The simulation will be processed and stored at JLab and only data summaries or small samples of the full data will be exported. Analyses needing access to hit-level information will be run at JLab or run on small samples of exported data unless they can take advantage of the data summaries. Data summaries will be written as ROOT trees. These will be generated and stored on tape at JLab, and mirrored on tape at SLAC.

Disk space at JLab will be needed for code releases and scratch areas. Disk space will also be needed at SLAC for staging, code releases, and scratch areas. Both needs are covered by existing computing infrastructure.

15.3.3 Process of Science

2014 and 2015 data will be collected at JLab using the HPS detector as explained above. It will be processed at JLab and then the DSTs will be exported to SLAC. Much, perhaps most, of the analysis will occur at SLAC on the DSTs stored there. There will be a corresponding set of simulated data that will follow a similar chain from JLab to SLAC.

15.4 Local and Remote Science Drivers — the Next 2–5 Years

Unknown at this time.

15.5 Beyond 5 Years — Future Needs and Scientific Direction

Unknown at this time.

15.6 Network and Data Architecture

The raw data will be in EVIO format, processed data in SLCIO format, DSTs in ROOT format. The data will be transferred by scp/bbcp.

15.7 Collaboration tools

WebEx conferences occur twice weekly. We also use the SLAC Confluence wiki for managing most of the HPS documentation. CVS is use for the code repository. GIT is used for collaborative work on proposals.

15.8 Data, Workflow, Middleware Tools, and Services

For the data storage summary, data (raw,rec,sim) storage is at JLab only, while DST storage is common to JLab and SLAC.

We currently plan to use the SLAC SRM Data Catalog for the book keeping of all of our data.

Microsoft Project is used for planning.

The constants database and data file management database are currently in development. A framework for accessing conditions already exists within the lcsim.org framework and current intentions are to take advantage of this with the actual storage of the metadata in an SQL database.

The HPS storage requirements are summarized in Table 27.

Table 27. HPS storage requirements.

Storage category	2014 (TB)	2015 (TB)
Raw data	140	192
Reconstructed data	304	394
Simulated data (raw and reconstructed)	27	31
Total data	472	618
DST (run data)	62	86
DST (simulated data)	3	3
Total DST	65	89

15.9 Outstanding Issues

We are preparing a data challenge but have not decided on the scope. It will likely start near the end of 2013 or beginning of 2014.

UNH has a postdoc and students who will analyze the HPS data but will they be running their jobs at UNH on DSTs copied there or running jobs on the already available data at JLab and SLAC. If at UNH, will the data be pulled from SLAC or from JLab?

15.10 Summary Table

See table above.

16 Intensity Frontier Experiments at Fermilab

16.1 Background

Particle physics experiments at the Intensity Frontier (IF) explore fundamental particles and forces of nature using intense particle beams and highly sensitive detectors. One of the ways researchers search for signals of new physics is to observe rarely interacting particles such as neutrinos, and their corresponding antimatter particles. Some of these experiments search for evidence of the process that theorists hypothesize allowed our universe full of matter to bloom rather than being annihilated by an equal amount of antimatter created in the Big Bang. Other experiments seek to observe rare processes that can give researchers a glimpse of unknown particles and unobserved interactions. This is the new thrust of Fermilab.

Note that there are IF experiments elsewhere in the world such as Tier-2K, Daya Bay, and SNO, all of which have some U.S. participation. This study will focus on those at Fermilab.

16.1.1 Neutrino Physics

Neutrinos are some of the most fascinating of the known particles. They abound in the universe but interact so little with other particles that trillions of them pass through our bodies each second without leaving a trace.

Neutrinos come in three types, called flavors: muon, electron, and tau. They have no electric charge. Their mass is so small that the heaviest neutrino is at least a million times lighter than the lightest charged particle.

At Fermilab, physicists use a beam of protons from the Main Injector accelerator to create the most intense high-energy neutrino beam in the world. Magnets direct the protons onto a graphite target. When the protons strike this target, they take the form of new particles called pions. A magnetic lens called a horn focuses and collects the positively charged pions and discards the negatively charged ones. The positively charged pions travel through a long, empty space and ultimately decay into antimuons and muon neutrinos. Experimentalists filter the resulting mix of debris, antimuons, undecayed pions, and muon neutrinos through a steel-and-concrete absorber, which stops all but the weakly interacting neutrinos. To make a beam of antineutrinos, they reverse the magnetic field of the horn to collect negatively charged pions that decay to negatively charged muons and muon antineutrinos.

The facility that creates Fermilab's neutrino beam is called NuMI, for Neutrinos at the Main Injector. The neutrinos travel between two detectors for an experiment called MINOS, or Main Injector Neutrino Oscillation Search. One sits at Fermilab; the other is located 450 miles away in the Soudan Underground Laboratory in Minnesota. The NuMI beamline is aimed downward at a 3.3-degree angle toward the underground laboratory. Neutrinos interact so rarely with other particles that they can pass untouched through the entire Earth.

Although the beam starts out at 150 feet below ground at Fermilab, it passes as much as 6 miles beneath the surface as it travels through the earth toward Soudan. Neutrinos travel at the speed of light and make the trip from Illinois to Minnesota in just 2.5 thousandths of a second. Researchers at Fermilab use the NuMI beamline as a source of neutrinos for other IF experiments as well.

16.2 Collaborators

The IF experiments are a mere shadow of what HEP is used to in recent years — namely the LHC experiments with thousands of collaborators. The IF experiments are typically composed of 200–400 members who are typically (at the moment) from U.S.-based institutions. The problems that IF experiments are trying to solve are equally as challenging as in the Energy Frontier, but with fewer people. Thus, experiments will need more laboratory help to achieve their goals.

16.3 Key Local Science Drivers

16.3.1 Instruments and Facilities

Historically, IF experiments have not placed much demand on computing systems — especially when compared with either the Tevatron or LHC experiments. However, the next generation of experiments expect to gather data on the order of petabytes per year and will require significant simulation programs as well. Thus the IF experiments will need to utilize the Grid resources through OSG to be successful. To date, the experimental HEP program has not made significant use of supercomputers. However, that will change. The LHC experiments are working on making use of these platforms — as they are successful, the knowledge will be transferred to the IF experiments as well.

16.3.2 Software Infrastructure

A broad range of experiments inevitably leads to a broad range of frameworks. The Fermilab-based IF experiments (from g-2, NOvA to LAr experiments including MicroBooNE and LBNE) have converged on the Fermilab-developed software “ART” as a framework for job control, I/O operations, and tracking of data provenance. Some highlights and advantages of this framework are listed below.

- It was developed and maintained by the Scientific Computing Division at Fermilab by computing professionals. It has perhaps the largest user base within IF at this time.
- Increased resources for this framework could enable some of the needs experiments such as more accessible parallelization of experiment’s code, for example using standard thread libraries (OpenMP, Threading Building Blocks).
- Experiments outside of Fermilab (or before ART) use LHC-derived frameworks such as Gaudi or homegrown frameworks like MINOS(+), IceTray, and RAT.
- The level of support for development and maintenance of such frameworks varies depending on whether the experiment is a significant stakeholder and/or significant human resources are available.

Software packages

- ROOT and GEANT4 are the bread and butter of all HEP experiments. They are critical to all experiments in IF. Support for these packages is essential.
- GEANT4 has traditionally focused on Energy Frontier experimental support. More ties/stronger support to IF experiments is a requirement.
- As an example, GEANT4 is barely suitable for large scintillation detectors, given a complex geometry and large number of photons to track.
- The community desires improved efficiency for both of these packages. For example better ROOT I/O and GEANT multithreading.
- Neutrino experiments use specialized packages for neutrino interactions: GENIE and Neut. GENIE is a public package that would benefit from continued support as it is heavily used in U.S. experiments.
- LArSoft is a common simulation, reconstruction, and analysis toolkit used by experiments using liquid argon time projection chambers (LArTPCs) that is managed by Fermilab. All U.S. experiments using LArTPCs currently use LArSoft. Similarly, the LAr and NOvA experiments share a simulation toolkit.
- Joint efforts where possible make better use of development and maintenance resources.
- A number of other specialized physics packages are in use by the community, for example: FLUKA for beamline simulations, CRY for simulating cosmic ray particles, NEST for determining ionization and light production in noble liquid detectors, GLOBES for experiment design.

16.3.3 Process of Science

Typically, the data are reconstructed in quasi-real time and made available to the collaboration for analysis. Simulations are normally handled through a central group within each group — and collaborators will either use the library of events available or request specialized production runs. The experiment will want to reprocess its data once a year.

16.4 Key Remote Science Drivers

16.4.1 Instruments and Facilities

The data for IF experiments will be stored centrally at Fermilab on robotic tape systems. The reconstructed data will be cached for faster access. The simulated data will also be stored at Fermilab. Analysis will be done both locally on Fermilab Grid machines as well as at other institutions. Analysis done at remote sites will typically pre-stage the data needed. Simulation will be done typically off site and then transferred to Fermilab, where it is cataloged and stored at its central tape facility.

16.4.2 Software Infrastructure

GridFTP (as part of the OSG stack) will be used for data transfers.

16.4.3 Process of Science

There is a high degree of commonality among the various experiments' computing models despite large differences in type of data analyzed, the scale of processing, or the specific workflows followed.

The model is summarized as a traditional event-driven analysis and MC simulation using centralized data storage distributed to independent analysis jobs running in parallel on-grid computing clusters. Peak usage can be 10 times more than the planned usage.

For large computing facilities such as Fermilab, it is useful to design a set of scalable solutions corresponding to these patterns, with associated toolkits that allow access and monitoring. Provisioning an experiment or changing a computing model would then correspond to adjusting the scales in the appropriate processing units.

Computing should be made transparent to the user, such that non-experts can perform any reasonable portion of the data handling and simulation — IF scientists are not as computer savvy in general as those at the Tevatron or LHC.

16.5 Local Science Drivers — the Next 2-5 Years

16.5.1 Instruments and Facilities

In the next 5 years, several new experiments will come online — g-2 and perhaps mu2e — which will place increasing demands on computational resources.

16.5.2 Software Infrastructure

The evolution of the computing model follows several lines, including taking advantage of new computing paradigms such as storage clouds, different cache schemes, GPU, and multicore processing.

In computing technology, there is a concern that as the number of cores in CPUs increases, RAM capacity and memory bandwidth will not keep pace, causing the single-threaded batch-processing model to be progressively less efficient on future systems unless special care is taken to design clusters with this use case in mind.

There is no current significant use of multithreading, since the main bottlenecks are GEANT4 (single-threaded) and file I/O. However there is interest in real parallelization at the level of ART, for example. Development is well under way with respect to ART for multithreading.

Greater availability of multicore/GPU hardware in grid nodes would provide a motivation to upgrade code in order to use it. For example, currently we can only run GPU-accelerated code on local, custom-built systems. A proposed example for GPU use included "repeated frequent tasks like quick down-going cosmics identification for pre-reconstruction filtering."

16.5.3 Process of Science

The science process is not expected to change over the next 2–5 years. As tools get more sophisticated and make better use of the more modern computing platforms, experimenters will use them — though it should be done such that they are unaware the platforms have/are changing.

16.6 Remote Science Drivers — the Next 2-5 Years

16.6.1 Instruments and Facilities

16.6.2 Software Infrastructure

There will be a demand for software that can make better use of the highly parallelized environments that are expected — and have memory footprints that fit within the next generation of hardware constraints. (See Section 16.5.2.)

16.6.3 Process of Science

The IF will keep a close eye on how the Energy Frontier LHC experiments operate; these experiments have the resources and need to push the envelope of what is possible in order to do their science. IF experiments don't require the state of the art in the same way. The IF experiments will adopt newly developed best practices, but they will not lead the way.

16.7 Beyond 5 Years — Future Needs and Scientific Direction

The Energy Frontier experiments will pave the way with respect to computing – the rest of HEP will learn from them. (See Section 16.6.3.)

16.8 Network and Data Architecture

Networking is critical for the success of HEP. The large "pipes" available have enabled the computing models of the LHC, namely Any data, Any time, Anywhere. While the IF's demands are less, this group would benefit greatly from the ability to move data to computing and back again.

16.9 Collaboration tools

Collaboration tools are very important to IF experiments. Traditionally, IF scientists remain based in their home institutions and do not, as a group, spend significant time at national laboratories. Therefore, tools to enable participation in meetings from remote sites are critical.

16.10 Outstanding Issues

All pertinent issues are discussed above.

16.11 Summary

- Current and future IF experiments have significant computing requirements.
- The quality and impact of the IF effort depends heavily on efficient and transparent access to dedicated computing resources.
- While resources are available for Fermilab-based experiments, all efforts will benefit from dedicated and transparent access to grid resources.
- Dedicated grid resources for the IF (perhaps in the form IF VO) would have the largest impact on international efforts.
- Computing professionals are in demand as support for key software frameworks, software packages, scripting access to grid resources, and data handling.
- Efforts (and problems) are shared across frontiers: significant investments in ROOT and GEANT4 optimizations, HPC for HEP, transparent OSG access, and open data solutions.

17 SLAC — Participation in Current and Future off-site Experiments and Collaborations

17.1 Background

SLAC is currently participating/planning to participate in a number of HEP/NP experiments and collaborations where the experiments are not located at SLAC:

- Fermi Gamma-ray Space Telescope (current)
- ATLAS (non-Tier-2 activity, current)
- Enriched Xenon Observatory (EXO) (current), nEXO (future)
- HPS (current)
- MicroBooNE (current)
- DES (current)
- Long Baseline Neutrino Experiment (LBNE) prototype (future), LBNE (future)
- Cryogenic Dark Matter Search (CDMS) (current), SuperCDMS (future)
- LZ (future)
- DarkSide (future)

Generally, the networking requirements of these activities are relatively modest and do not merit individual case studies. However there is enough commonality to instead suggest a consolidated case study. For many of the future projects, the level of SLAC participation in computing has not yet been determined, so we can only present very rough estimates here.

17.2 Collaborators

A list of collaborators in the above collaborations would be prohibitively long.

17.3 Key Local Science Drivers

17.3.1 Instruments and Facilities

All experimental facilities and major data sources described here are not located at SLAC, however portions of the production and analysis computing for these experiments are or will be performed at SLAC. Dataset sizes range from tens of terabytes to a few petabytes. Facilities at SLAC to perform the analysis are compute clusters and storage systems.

17.3.2 Software Infrastructure

While the specific implementation of software tools varies with the experiment, common tools to manage the scientific process include workflow management systems, batch systems, and databases for file and experiment metadata.

17.3.3 Process of Science

While different in detail, the overall analysis and science workflow is generally very similar. “Raw” data from the experiment is processed into “reconstructed” datasets, which are then made available to scientists and analysis groups within the collaboration

for physics analysis. Since most of these activities are performed using distributed (locally or globally) computing methodologies, networking is essential for moving data.

17.4 Key Remote Science Drivers

17.4.1 Instruments and Facilities

- Fermi Gamma Ray Space Telescope (FGST): Transferring 15 GB/day raw data from NASA Goddard Space Flight Center (GSFC) to SLAC
- ATLAS (non-Tier-2 center): 1.5 Gbps from/to other ATLAS sites
- EXO: Transferring a few terabytes per week from the Waste Isolation Pilot Plant (WIPP) to SLAC
- HPS: Transferring DST data (on the order of 100 TB/year) from JLab to SLAC
- DES: Transferring 200 GB/day from Chile (Cerro Tololo) to SLAC
- MicroBooNE: Amount of computing to be performed at SLAC not known yet
- LBNE prototype / LBNE: No details known yet; amount of computing to be performed at SLAC is not known
- SuperCDMS(Soudan): Very little involvement by SLAC in computing, Fermilab main processing site, Stanford main analysis site

17.4.2 Software Infrastructure

17.4.3 Process of Science

17.5 Local Science Drivers — the Next 2-5 Years

17.5.1 Instruments and Facilities

Except for modernization of local computing hardware at SLAC, very little change is expected.

17.5.2 Software Infrastructure

The main change we see is that even small experiments and collaborations are moving toward Grid-based distributed computing models.

17.5.3 Process of Science

SLAC will be participating in more off-site activities, each possibly at a smaller scale than current collaborations.

17.6 Remote Science Drivers — the Next 2–5 Years

17.6.1 Instruments and Facilities

- FGST: No change expected.
- ATLAS (non Tier-2): Data rates will increase to approximately 10 Gbps over this time period.
- EXO, nEXO: No change expected.

- HPS: No change expected.
- DES: No change expected.
- MicroBooNE: Not known yet.
- LBNE prototype / LBNE: Not known yet.
- SuperCDMS (SNOLAB): No change in data rates or dataset sizes expected.
- LZ: If funded, the experiment will start taking data toward the end of the 5-year period. Data rates and dataset sizes depend heavily on achievable data reduction and compression. For raw data processing at SLAC, a few hundreds of megabytes per second from Sanford Underground Research Facility (SURF) to SLAC would be required.
- DarkSide: If funded, the experiment will start taking data toward the end of the 5-year period. Data rates and dataset sizes depend heavily on achievable data reduction and compression. For raw data processing at SLAC, up to a few hundreds of megabytes per second from the National Laboratory of Gran Sasso (LNGS), Italy, to SLAC would be required.

17.6.2 Software Infrastructure

We expect to move more toward standardized frameworks and software tools, especially in the small collaborations.

17.6.3 Process of Science

17.7 Beyond 5 Years — Future Needs and Scientific Direction

17.8 Network and Data Architecture

Generally, architectures, even of small experiments, will move toward Grid-based computing models and architectures. We also anticipate a tighter interaction with Stanford University (possibly sharing compute facilities and resources). In this model, the Science DMZ approach appears very interesting.

We also expect the emergence of the use of private and public clouds with the associated use of networks to access these computing resources.

17.8.1 Collaboration tools

ReadyTalk phone/audioconferencing is in widespread use in all collaborations. For videoconferencing, the predominant tools appear to be Skype and SeeVogh.

17.9 Data, Workflow, Middleware Tools, and Services

17.10 Outstanding Issues

None.

18 Daya Bay Neutrino Experiment

18.1 Background

The Daya Bay Reactor Neutrino Experiment is a China-based multinational particle physics project studying neutrinos. The multinational collaboration includes researchers from China, the United States, Taiwan, Russia, and the Czech Republic. The U.S. side of the project is funded by DOE HEP.

The experiment studies neutrino oscillations and is designed to measure the mixing angle θ_{13} (theta-one-three) using antineutrinos produced by the reactors of the Daya Bay Nuclear Power Plant and the Ling Ao Nuclear Power Plant. Scientists are also interested in whether neutrinos are CP (charge parity) violators.

On 8 March 2012, the Daya Bay collaboration announced a 5.2σ discovery of $\theta_{13} != 0$, with $\sin^2 (2* \theta_{13}) = 0.092 +/- 0.016(stat) +/- 0.005(syst)$.

Data taking is continuing with a nominal, steady-state rate of about 350 GB/day to improve the precision measurement of θ_{13}, and conduct other studies and research programs (e.g., reactor characteristics).

18.2 Collaborators

Daya Bay is a medium-size HEP collaboration of 230 scientists and 38 institutions in the United States, China, Russia, Czech Republic, and Taiwan. Major computing facilities include:

- On-site Daya Bay:
    - Dedicated networking, computers, and storage are located on site for DAQ and slow controls (DCS) functions, control, data transfer, and on-site real-time data quality monitoring.
- Institute of High Energy Physics (IHEP), Beijing, China:
    - The China Tier-1 Facility at IHEP is an offshoot of the BES-III computer facility. All data are stored and all processing occurs on the IHEP cluster.
- LBNL, NERSC, Berkeley, California:
    - The US Tier-1 facility at LBNL includes NERSC's PDSF cluster, Global File System (GFS), and HPSS tape system. There are more than 250 current and past user accounts on the Daya Bay repo (account). All data are stored and all processing occurs on PDSF.
- BNL and universities:
    - Institutional clusters and compute resources at each institution vary dramatically in scale and usage, from individual desktop machines to large, institutional cluster and shares of major facilities (like RACF).

18.3 Key Local Science Drivers

18.3.1 Instruments and Facilities

The on-site network must support both data transfers and our new interactive Remote Shift capability and collaboration services such as videoconferencing. Daya Bay uses DCS videoconferencing between U.S. and Chinese institutions. We have installed videoconferencing hardware on site at the Daya Bay nuclear power plant. We also routinely use SeeVogh and Skype for one-on-one communications, larger meetings, and as part of our Remote Shift toolkit.

All experimental halls are connected via fiber optic Ethernet to the control room, which is connected over circuits that provide OC3-level speeds to IHEP and CSTNet (this is actually over a 1-Gbps fabric restricted to 150 Mbps rather than a physical OC3 line).

DAQ, Detector Control Systems (DCS), and offline computing have on-site resources at the Daya Bay and Ling Ao power plants sufficient to record data (with a buffer of about 4 weeks' worth of full experiment operation) and transfer data off site. Data are migrated in real-time to a computer facility at IHEP in Beijing, and to NERSC in Berkeley. Data are transferred to disk at both facilities and archived to tape within 30 minutes (nominal) of close of file. Networking out of Daya Bay is a dedicated OC3 to Beijing. From there, CSTNet, GLORIAD, and ESnet are used to migrate data to U.S. scientists.

On-site scientists serve shifts using IBM blades and servers, including machines for control and monitoring of the DAQ, machines for control and monitoring of the DCS, and a small user cluster.

18.3.2 Software Infrastructure

We use SPADE as an orchestration layer to transfer data between Daya Bay, IHEP, and LBNL with an underlying transfer protocol of GridFTP (configurable). On-site scripts take care of managing local user disk space with a high/low-watermark triggered age-based algorithm. SPADE ensures delivery and validity of data to IHEP and LBNL, and archiving onto NERSC's HPSS before the on-site copy is released.

We have a real-time PQM (Physics Quality Monitoring) program running a lightweight NuWa to generate ROOT histograms and plots presented via a Web interface. PQM runs on the on-site user blade cluster.

We have recently developed a Remote Shift capability that allows collaborators full access to the DAQ, DCS, and offline monitoring systems via a Web interface. Collaborators serve 8-hour shifts from their home institutions almost as effectively as on site. End-to-end network bandwidth and latency from a typical U.S. institution is more than sufficient for this purpose.

18.3.3 Process of Science

The ability to routinely and quickly transfer and process raw files in real time (keep up production [KUP]) means that scientists rely upon the KUP output for near real-time feedback and data quality assurance. Our Science Data Gateway is called ODM (Offline

Data Monitor) and is a sophisticated Django framework using NERSC's NEWT (NERSC Web Toolkit) to present an interactive, real-time interface to hundreds of analysis artifacts for each data file, data run, detector, and experimental hall. Artifacts include ROOT histograms and plots, MySQL information and queries, DAQ and DCS configuration and monitoring data, ELog entries, etc. All of this is presented through a Web interface in real time with easy-to-navigate entry points. It is used by all collaborators in the United States, China, and elsewhere. It is used by scientists while on site as it is the most functional and complete presentation of data available.

18.4 Key Remote Science Drivers

18.4.1 Instruments and Facilities

The PDSF Cluster at NERSC is the U.S. Tier-1 center for Daya Bay simulation and data processing. The HPSS mass storage system at NERSC is our main U.S. data archive for all data and information. This includes all raw data, simulated data, derived data, and associated database backups and other files.

Starting in December 2011, we began steady-state run of six antineutrino detectors (ADs) with a raw data rate of approximately 260 GB/day. In July 2012, 8 AD runs began at about 350 GB/day. We are currently generating about 225 TB of storage usage annually.

The datasets consist of 1 GB data files. Additional metadata are transferred, as well as database transactions for support of analysis functions. The data are transferred from the detector site at Daya Bay to IHEP in Beijing, and then from IHEP to NERSC. The path from IHEP to NERSC is via CSTNet from IHEP to Hong Kong, via GLORIAD from Hong Kong to Seattle, and via ESnet from Seattle to NERSC. The data reside on disk at Daya Bay, IHEP, and NERSC. The data at Daya Bay are deleted once they have been transferred successfully to IHEP and NERSC.

The network path from Daya Bay to NERSC and the rest of the United States has changed dramatically over the course of the experiment. CSTNet is the Chinese national network we use for communication and data transfers, and is connected to the United States via 10 Gbps GLORIAD. We have trans-Pacific network outages that can last three to six weeks due to suboceanic cable damage on an irregular basis (e.g., twice in the past 36 months), usually due to ship traffic around Hong Kong. Under such circumstances, we fail over to the 2.5-Gbps link with ASGC through South Korea until repairs are made. CSTNet also connects to Russia via another leg of GLORIAD and to Europe at 10 Gbps over Orient+ via CERNet (the other Chinese national network).

During network outages, we can buffer data on site (about 40 TB of on-site disk cache), or at IHEP, and then transfer data at double speed in recovery mode. We are able to transfer data directly from on site to NERSC, and do so during IHEP machine downtimes (e.g., cluster maintenance and/or problems).

All network traffic from Daya Bay goes through IHEP in Beijing. Indeed, the on-site private subnets are inside the IHEP networking domain, and all external IPs are seen as

IHEP subnet addresses. So, though IHEP and Daya Bay are geographically 2,000 km apart, Daya Bay networking is topologically inside of IHEP.

Daya Bay does not explicitly use any Grid PKI (public key infrastructure) services, though our data migration system (SPADE) uses GridFTP as one of the plug-in transfer protocols.

NERSC's PDSF consists of approximately 2,400 cores of Ethernet and IB-connected Linux machines, of which Daya Bay has about 350 dedicated cores and access on an opportunistic basis to the rest, as well as 5 million CPU hours allocated on the Carver cluster. Daya Bay has about 900 TB of disk, and more than 1500 TB of tape (HPSS) available for raw, processed, and user data. The cluster at IHEP is of the same order of magnitude.

Daya Bay software is a suite of tools including SPADE (data migration), P-Squared (data processing workflow), Offline_DBI (DBI-based offline information), and NuWa (Gaudi-based simulation and analysis framework). All production processing is done on the IHEP and NERSC clusters using NuWa. Real-time processing nominally occurs and is available to U.S. collaborators within two hours of data taking. Full dataset analyses and simulations occur at both IHEP and NERSC, and the resultant datasets are compared between IHEP and NERSC. Individual analysis is done using NuWa and ROOT. Detector simulation is based upon GEANT4 and NuWa.

18.4.2 Software Infrastructure

We use SPADE as an orchestration layer to transfer data between Daya Bay, IHEP, and LBNL with an underlying transfer protocol of GridFTP (configurable). On-site scripts take care of managing local user disk space with a high/low-watermark triggered age-based algorithm. SPADE ensures delivery and validity of data to IHEP and LBNL, and archiving onto NERSC's HPSS before the on-site copy is released.

A data Warehouse Catalog at LBNL keeps track of all raw and processed files (not individual user files). The Warehouse Catalog can be queried directly (using SQL) or through a python module, which is part of NuWa (can be used independently).

Most U.S. collaborators log into PDSF to access and analyze data, but can download data to their home institutions manually, or using SPADE (requires instantiating a server at the receiving end). BNL uses resources associated with the RACF and XRootD to analyze and simulate Daya Bay data.

P-Squared is a job management and submission system used to define, schedule, monitor and control large numbers of batch jobs on PDSF and Carver. PDSF used a Sun Grid Engine and Carver PBS for their batch queues. PDSF uses a fair-share algorithm for access to the queues.

KUP happens automatically as raw data files arrive at PDSF. KUP is triggered by SPADE and managed by P-Squared. Raw data files (1 GB) typically arrive at LBNL within 20 minutes of close-of-file by the DAQ, and are processed within 120 minutes of close-of-file (including queue wait times).

18.4.3 Process of Science

KUP is responsible for real-time processing and ODM for WAN presentation of those results 24/7. Collaborators daily, if not hourly, check ODM for status, physics, and data quality questions.

Large-scale production processing of raw data happens on PDSF (and IHEP) about two to four times per year. All raw data are on spinning disk and processed on PDSF. A full production takes about 4–6 weeks using the full Daya Bay allotment of CPUs on PDSF. However, we can routinely complete a production within one week using either opportunistic CPU resources or through cooperative agreements with other experiments on PDSF (e.g., ATLAS, ALICE, STAR, IceCube, etc).

Scientists who log on to PDSF can run their own analysis against the data using the Warehouse Catalog python module to access data and submit batch jobs. Use of P-Squared by nonproduction managers is rare.

18.5 Local Science Drivers — the Next 2–5 Years

18.5.1 Instruments and Facilities

No change in 2–5 years.

18.5.2 Software Infrastructure

With the development of the Remote Shift system, we have likely seen the last of the major developments of Daya Bay local infrastructure. General maintenance and improvements to the systems are ongoing to optimize performance and stability and respond to user issues and requests.

18.5.3 Process of Science

No change in 2–5 years.

18.6 Remote Science Drivers — the Next 2–5 Years

18.6.1 Instruments and Facilities

No change in 2–5 years.

18.6.2 Software Infrastructure

With the development of the Remote Shift system, we have likely seen the last of the major developments of Daya Bay local infrastructure. General maintenance and improvements to the systems are ongoing to optimize performance and stability and respond to user issues and requests.

ODM improvements and increased use of Carver at NERSC will occur over the 2-year to 4-year time frame.

18.6.3 Process of Science

There are additional science questions being asked and addressed that require the development of new NuWa algorithms and ROOT analyses. But for the foreseeable future, the general process will be stable.

18.7 Beyond 5 Years — Future Needs and Scientific Direction

Network usage will ramp down as the experiment concludes. Juno is a follow-on experiment in China that may lead to future needs and direction.

18.8 Network and Data Architecture

In our experience, trans-Pacific networking and intra-Chinese networking are much better today than they were 5–7 years ago. However, they are still not as stable, performant, and reliable as U.S. national and local networks. We constantly monitor network and data transfer and respond as needed.

Undersea cable outages cause significant interruptions in connectivity. The following is a partial list of undersea cable service interruptions affecting GLORIAD (the primary provider for U.S.-China connectivity for the Daya Bay Neutrino Experiment):

- **February 2012**: GLORIAD down 28 days (ship dragged anchor at Hong Kong). Alternate connectivity via ASGC worked well, but with some reduction in performance (500 Mbps was reduced to 300 Mbps).
- **October 2011**: GLORIAD down 6 weeks (ship dragged anchor at Hong Kong). Alternate connectivity via ASGC worked well.
- **August 2009**: Typhoon Vamco took out GLORIAD for 6–8 weeks. Manual rerouting required to transition traffic to TransPac2 network.
- **December 2006**: Hengchen earthquake brought down GLORIAD for 5 weeks. The alternate route had 1% of the performance of the normal production route.

It is clear from this list that care must be taken when provisioning connectivity on undersea cables — backup connectivity is simply required to ensure continuity of operations.

U.S. collaborators routinely have problems with Chinese content filters when in China at the experiment site or collaborating institutions. The interaction between CSTNet and the “Great Firewall of China” is not clear to even IHEP network engineers. We have not had any problems with data transfers (other than occasional outages) for several years now, but continue to monitor the situation and communicate with IHEP and CSTNet network engineers and managers.

Though U.S. scientists and engineers were largely responsible for the data transfer and network specifications for Daya Bay, all communication, procurement, and interactions with vendors went through IHEP and Chinese channels.

18.9 Collaboration Tools

We use Skype, SeeVogh, ESnet's ECS, and combinations of these on a daily basis. We have a SeeVogh Virtual Control Room (VCR) and a Skype Shifter account up and active 24/7 for communication with on-site shifters and as a drop-in for collaborators and remote shifters.

We experience the same routine problems with audio quality using these services that everyone using such tools faces. The last-mile quality of networking at some Chinese institutions exacerbates the problem, as does the combination of technologies (e.g., when someone uses Skype to call into a DCS phone call). We would like to see improvements in all tools in both stability and quality, and better interoperability.

18.10 Data, Workflow, Middleware Tools, and Services

At this point, Daya Bay will see no significant data growth for the next phase. If the United States participates in other Chinese–U.S. collaborations (e.g., Juno), we expect that tools like Globus and Globus Sharing will be large considerations.

18.11 Outstanding Issues

See above.

18.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• Daya Bay nuclear power plant near Shenzhen, China (as a neutrino source), with 8 antineutrino detectors (4 near and 4 far). • Datasets are 1 GB files. • Derived datasets are 100% of raw dataset size. • Simulated datasets are 10% of raw dataset size. • Database synchronization and other traffic in addition to data traffic. • Remote shift and videoconference traffic. • SPADE and GridFTP for data transfer. • NuWa, ROOT, GEANT for processing, analyzing, and simulation. • P-Squared, SGE, and PBS for job management and execution. • Data Warehouse Catalog (SPADE) and Postgres for data file management. • MySQL and DBI (Database Interface) for calibration and time-dependent parameters. • ODM, PQM, NEST, Django for real-time data presentation.	• Analysis of raw, derived, and simulated data to determine the θ_{13} mixing angle. • Transfer of raw data from detectors to IHEP in Beijing, and from IHEP to NERSC. • Transfer of simulated and derived datasets between IHEP and NERSC.	• 1 GB per file. • 350 GB per day. • 175–350 GB per run. • Calibration runs are much smaller. • Dataset is composed of all similar runs — currently about 70,000 files @ 1 GB each.	• We process datasets in place and do not transfer them. • Global disk I/O is a limiting factor in large-scale analysis. (i.e., we have far more CPU nodes than required to saturate the shared disk resource).	• Raw data files are transferred as they are taken. They are transferred and archived within 20 minutes of DAQ. • KUP processing occurs within 2 hours, and is dominated by execution time.

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
2–5 years				
Same as above	Same as above	Same as above	Same as above	Same as above
5+ years				
None	None	None	None	None

19 Belle II Experiment

19.1 Background

The Belle II computing system has to handle an amount of data eventually corresponding to about 50 PB/year under an operation of SuperKEKB accelerator at design luminosity. To achieve the physics goals within a timely manner, raw data must be processed without any delay to experiment data acquisition. In addition, MC samples corresponding to more than six times the beam data must be produced for physics analyses. Belle II has adopted a distributed computing model based on the Grid. A key component of this model is the establishment of a remote data center at the Pacific Northwest National Laboratory (PNNL), where the raw data can be reprocessed in parallel with KEK within a Belle II distributed computing framework.

Assuming the expected instantaneous luminosity and the raw data event size of 300 KB, the data rate at KEK is estimated to be 1.8 GB/sec. Assuming the maximum file size of the raw data to be 4 GB, a raw data file will be generated roughly every 2 seconds, ultimately amounting to more than 10,000 files created in a typical physics run. Once the raw data file is closed on the online storage disk, the Data Acquisition (DAQ) System will return an acknowledgement to the offline computing system, and then the data transfer to the offline tape storage can start. During this procedure, the file metadata is extracted and registered in AMGA so that it can be accessed from the DST production expert via Grid jobs (this scheme is still under discussion). After the completion of the data transfer, the raw data will be processed on the Grid system, and the resultant mDST file is stored on the offline disk storage at KEK. Because the DAQ network should be separated from the Internet, we will have a special network path between the online storage disk and the offline computing system.

The raw data and metadata are replicated from the offline tape storage at KEK to disk at PNNL. It can be then processed in parallel with the raw data processing at KEK and/or reprocessed later with the updated detector calibration constants. The data files that result from the raw data processing will be kept on disk for distribution to scientists for analysis.

An experiment-specific network requirements report for the Belle-II experiment is available at http://www.es.net/assets/pubs_presos/Belle-II-Experiment-Network-Requirements-Workshop-v18-final.pdf

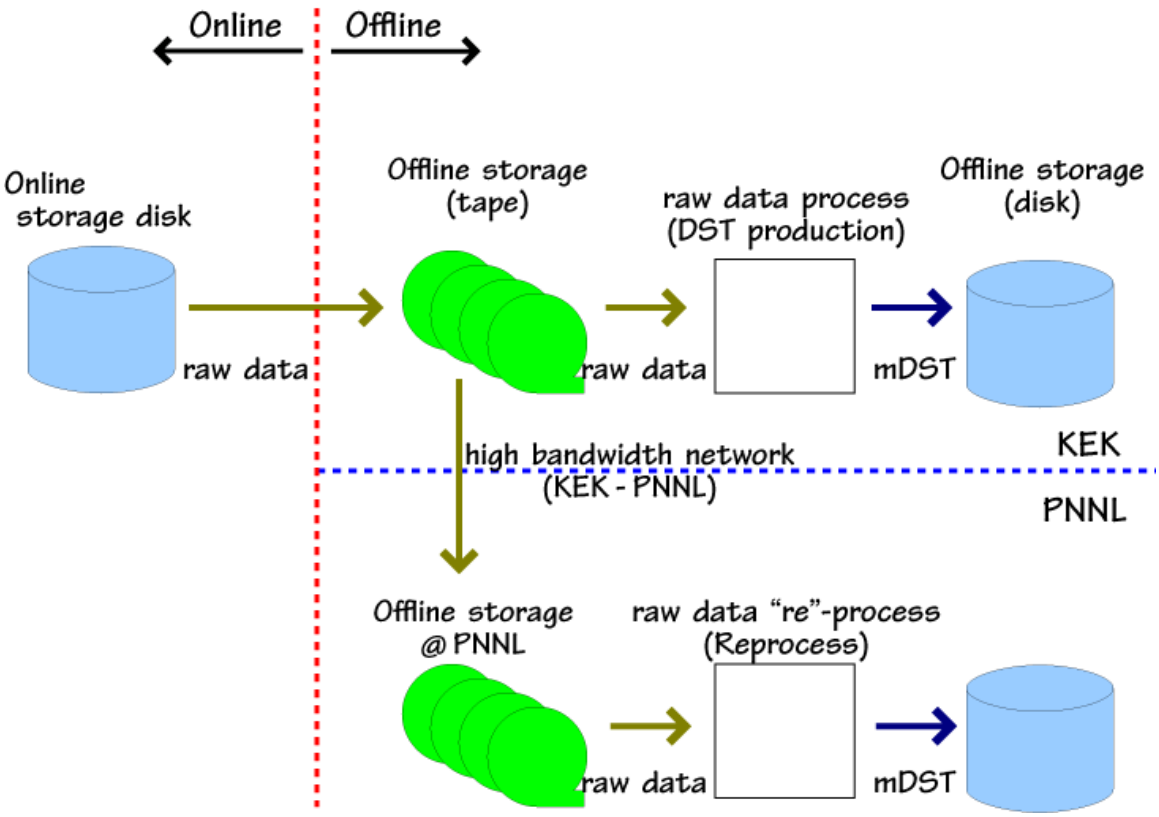


Figure 58. Belle II data flow.

19.2 Collaborators

The U.S. Belle II institutions are PNNL, Carnegie Mellon University, University of Cincinnati, University of Hawaii, Indiana University, Kennesaw State University, Luther College, University of Mississippi, University of Pittsburgh, University of South Alabama, University of South Carolina, Virginia Tech, and Wayne State University. In September 2012, the U.S. Belle II DOE project managed by PNNL achieved the CD-1 milestone.

The Belle II Collaboration includes more than 500 scientists from 22 countries — Japan, United States, Australia, Austria, Canada, China, Czech Republic, Germany, India, Korea, Malaysia, Mexico, Poland, Russia, Saudi Arabia, Slovenia, Spain, Taiwan, Thailand, Turkey, Ukraine, and Vietnam.

19.3 Key Local Science Drivers (e.g., Local Network aspects)

19.3.1 Instruments and Facilities

PNNL Grid Computing:

1. 1 Gatekeeper node using condor (investigating SLURM)
2. 1 storage element using Bestman2
3. 2 GridFTP servers with 10 Gbps connections
4. 768 cores:
    a. AMD Opteron Processor 6272, 2.1 GHz

    b. System memory 64 GB
    c. 2 CPUs/node
    d. 16 cores/CPU
5. DIRAC server: Grid jobs scheduling, monitoring, and data management Grid software stack
6. AMGA server: Metadata catalog
7. 2.75 PB shared Lustre storage
8. 100 TB dedicated Lustre storage for network data challenges

19.3.2 Software Infrastructure

Software Stacks Used:

1. Scientific Linux versions 5 and 6
2. Open Science Grid software stack for:
    a. Compute Element (Condor)
    b. Storage Element (BeStMan)
    c. Worker Nodes
3. Monitorix: Cluster monitoring tool
4. DIRAC Server and Client: Grid software stack for job scheduling, monitoring, and data management
5. Basf2: Belle II common software framework
6. gBasf2: Belle II grid software (basf2 wrapped within DIRAC with supplemental information).

19.3.3 Process of Science

Under the Belle II computing design, every possible Grid site is expected to have some number of output files resulting from raw data processing/reprocessing and MC events in proportion to the number of Ph.D. physicists assigned to that Grid site.

Within the United States, PNNL is expected to have the full raw and mDST datasets and will redistribute the mDST and to participating sites.

19.4 Key Remote Science Drivers

19.4.1 Instruments and Facilities

Over the next 2 years, the infrastructure for replicating the raw data will be developed, deployed, and tested. This will require the development of network configuration, data transfer node configuration, security policy development, and workflow integration.

Several aspects of these tasks were discussed at the Belle II Experiment Requirements workshop held at PNNL November 17–18, 2012. One aspect is whether to use a standard routed network service or a virtual circuit service for data replication. The consensus of the group was to explore a virtual circuit service because of the additional capabilities of traffic isolation and traffic engineering that a virtual circuit service provides — these were seen as advantages over a best-effort routed service.

Currently, Belle II Data Challenges use FTS2 with well-defined channels (endpoints) over a shared network.

19.4.2 Software Infrastructure

The Belle II experiment has adopted the Grid computing model to enable the processing of the very large volume of experimental data and MC samples that the collaboration must analyze. In order to realize this, we also need access to different types of computing resources. The following is an itemized list of software stacks that are used to enable the Belle II grid computing:

1. Middleware
    a. Open Science Grid (U.S.)
2. gLite (Europe/Asia/Canada) DIRAC: Grid jobs scheduling, monitoring, and data management Grid software stack
3. AMGA (ARDA Metadata Grid Application): provides efficient and scalable metadata searching
4. Basf2: Belle II common software framework (composed of several modules to perform various task such as physics analysis, full detector simulation, etc.)
5. gBasf2: Belle II grid software (basf2 wrapped within DIRAC with supplemental information)
6. FTS2: Currently being used for large-scale data transfers (Data Challenges)

19.4.3 Process of Science

Over the next two years, the data replication workflow must be developed and tested.

The Belle II Experiment Requirements workshop attendees discussed the use of data challenges, wherein the data replication workflow is run for a period of time with simulated data. Each data challenge would have a performance target, with each successive challenge having a higher performance target until the final challenge, which would run at the peak performance level expected for the first year or two of production physics runs on the Belle II experiment.

The workshop reached consensus that the first data challenge would be held by the summer of 2013. A table containing the ideal goals of the first challenge and two additional data challenges are below:

Table 28. Goals of the Belle II data challenges.

Date	Summer 2013	Summer 2014	Summer 2015	Production
Rate	100 MB/sec	400 MB/sec	1000 MB/sec	1000 MB/sec
Duration	24 hours	48 hours	72 hours	24 hours/day

It is likely that some portion of the data transfer nodes, storage, and network equipment that will be used when the experiment begins production operation will be purchased sometime in 2015. The data challenge in the summer of 2015 should be conducted using the equipment that will be used in production operation of the experiment.

Since the raw data replication and data analysis workflows will run concurrently when the experiment is running in production, it is expected that at least the 2014 and 2015 data challenges will run concurrently with the data challenges for the data analysis workflow.

19.5 Local Science Drivers — the Next 2–5 Years

19.5.1 Instruments and Facilities

PNNL, as a raw data storage center, plays an important role in data reprocessing. We assume that the reprocessing will be repeated most frequently in the first year of the data collection, then the number of reprocessings per year is expected to decrease as the reconstruction software matures. Finally, after four years of operation, the collaboration must stop reprocessing activities, except in the case that a more sophisticated reconstruction algorithm is invented. On the other hand, the amount of beam data will increase as the instantaneous luminosity increases. PNNL will mainly handle the reprocessing in the early stage of the experiment and evolve into a data storage role in the latter stage. PNNL will store the latest and second-latest versions of the mDST. As with the reprocessing of the raw data, the corresponding MC samples will also be produced in proportion to the number of Ph.D. physicists in each Grid site (15% for PNNL). Another role of PNNL will be to distribute the reprocessed mDST to the Belle II Grid sites. Table 29 shows the required computing resources for PNNL.

Table 29. Required Belle II resources at PNNL.

PNNL Resources									
Year	2013	2014	2015	2016	2017	2018	2019	2020	2021
Tape [PB]	0.00	0.00	0.00	0.00	9.62	27.22	51.77	76.94	102.3
Disk [PB]	1.00	1.00	2.00	5.00	12.00	17.00	22.00	27.00	32.00
CPU [kHepSPEC]	5.00	5.00	10.00	15.00	59.11	95.81	76.58	82.65	87.63
WAN [Gbit/s]	0.50	1.00	2.50	4.00	8.65	15.75	18.82	19.29	19.44

19.5.2 Software Infrastructure

The U.S. Belle II computing relies on the OSG software stack and anticipates doing so for the duration of the experiment. Currently, PNNL is relying on a remote FTS2 server for scheduling large-scale data transfers. However, the PNNL site plans to deploy a FTS3 server and run several Data Challenges to evaluate this new technology. In addition, other technologies will be investigated.

19.5.3 Process of Science

As the performance requirements for the raw data replication workflow increase, there will be a need for development, test, and measurement of additional systems and software capabilities. It is expected that these activities will be conducted during accelerator downtime.

19.6 Remote Science Drivers — the Next 2–5 Years

19.6.1 Instruments and Facilities

Production operation of the Belle II experiment is scheduled to begin in 2016. Once production operation begins, the raw data replication workflow is expected to run for two-thirds of the year, increasing in data volume as the capabilities of the detector increase. The expected data production volume and data rate of the raw data replication workflow for the years 2016, 2017, and 2018 is contained in the following table.

Table 30. Expected Belle II raw data production and replication volume.

Year	2016	2017	2018
Tape (PB)	0.82	9.62	27.22
Disk (PB)	0.39	4.57	12.94
WAN (Gbit/s)	0.84	9.71	18.83

19.6.2 Software Infrastructure

The U.S. Belle II computing will continue to rely on the software infrastructure described in Sections 19.4.2 and 19.5.2.

19.6.3 Process of Science

As the performance requirements for the raw data replication workflow increase, there will be a need for development, testing, and measurement of additional systems and software capabilities. It is expected that these activities will be conducted during accelerator downtime. This will require coordination among the operational groups responsible for the different parts of the infrastructure, including KEK, SINET, ESnet, and PNNL.

19.7 Beyond 5 Years — Future Needs and Scientific Direction

Very little process change from the 2–5 year case is expected.

19.8 Network and Data Architecture

Trans-Pacific data transfers will be continuous and increasing over the operational lifetime of the Belle II detector (2015–2021). Aggregate data transfer from KEK to PNNL will exceed 100 PB, including raw data transfers and data challenges; refer to Sections 19.4.3, 19.5.1, and 19.6.1 in this document.

19.9 Collaboration tools

The Belle II collaboration is currently using SeeVogh for video/audio conference calls.

19.10 Data, Workflow, Middleware Tools, and Services

Data growth is planned for and will drive increases in data transfer rate requirements as described in Sections 19.4.3, 19.5.1, and 19.6.1 in this document. The collaboration plans

to rely on Grid computing middleware (particularly the OSG software stack within the United States) through 2021.

19.11 Outstanding Issues

PNNL is responsible for receiving the raw data from KEK and redistributing the mDSTs to Europe/Canada. As such, we would greatly benefit by having a KEK–PNNL virtual circuit for our ongoing Data Challenges.

In addition, we should determine how to proceed to enable a PNNL–Europe virtual circuit.

19.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
Development of the Belle II raw data replication workflow system and infrastructure.	• Development, test, verification, and commissioning. • Periodic data challenges to ensure data replication workflow is ready.		• 2 PB data to be copied from online disk to offline disk to test workflow. • 2.5 Gbps bandwidth to computational analysis for testing prompt reconstruction workflow.	• Virtual circuit configuration from raw data replication workflow. • 100 MB/sec for 24 hrs in first data challenge. • 400 MB/sec for 48 hrs in second data challenge. • 1000 MB/sec for 2 hrs in third data challenge. • Periodic test flows for debugging and performance analysis of workflow
2–5 years				
• First few years of physics using Belle II. • 300 KB event size. • Raw data files of 4 GB, more than 10,000 files from each run. • Increasing data production as experiment is refined.	• Replication of raw data from KEK to PNNL. • Processing of raw data into mDSTs at PNNL and KEK. • Data challenges for increased replication rates during experiment shutdown periods.		• 5 PB data to be copied from online disk to offline disk. • 10 Gbps LAN bandwidth to computational analysis for prompt reconstruction.	• 80 MB/sec from KEK to PNNL for raw data replication in first year. • Growth to 1500 MB/sec for raw data replication by 2018.
5+ years				
Progression to full luminosity at Belle II	No change		• 110 PB data to be copied from online disk to offline disk. • 25 Gbps LAN bandwidth to computational analysis for prompt reconstruction.	• 1500 MB/sec from KEK to PNNL for raw data replication in 2018. • Growth to 1900 MB/sec for raw data replication by 2022.

20 Dark Energy Spectroscopic Instrument

20.1 Background

The Dark Energy Spectroscopic Instrument (DESI) will measure the optical spectra of millions of galaxies and quasars over a large fraction of the sky. This information will be used to build a model of our locally observable universe out to a distance of approximately 10 billion light-years. Measurements of this large-scale structure will help improve our insight into the nature of dark energy by determining its impact on the expansion history of the universe. It will also allow us to put tighter constraints on the neutrino mass hierarchy when combined with other datasets.

DOE has selected DESI as the first stage-IV dark energy experiment, which will fill the time gap in between the Baryon Oscillation Spectroscopic Survey (BOSS) and the Large Synoptic Survey Telescope (LSST). DESI is a large endeavor that requires significant expertise from a variety of people within DOE and the larger cosmology community. The DESI collaboration is still being formed, but already many universities and government entities have expressed interest in joining the project. Below is a list of such institutions.

AAO	KASI	Univ. College London
Argonne	LAM/CPPM	UC Berkeley
Brazil	Mexico	UC Irvine
Brookhaven	NOAO	UC Santa Cruz
Carnegie Mellon Univ.	New York Univ.	U. Edinburgh
Durham	Portsmouth	U. Michigan
ETH Zurich	Saclay	U. Pittsburgh
FNAL	SJTU	U. Utah
Harvard	Spain	USTC
IAA Spain	Texas A&M	Yale
Kansas	The Ohio State Univ	

The DESI team is involved in many aspects of project execution. In particular, a huge effort is being directed at designing and building the focal plane, spectrographs, and supporting physical infrastructure. In this document, we focus only on the areas of DESI that we expect will require significant use of network resources. The four broad categories are:

1. Transfer of raw data from the telescope to NERSC and mirroring of this data to a secondary site (TBD).
2. Transfer of targeting data from a variety of locations to NERSC.
3. Movement of subsets and/or partially processed outputs of simulations.
4. Serving processed data to the collaboration and the public — both syncing of data to academic institutions and Web-based download to individuals.

These areas come in to play to varying degrees during the three time periods considered in this document (next 2 years, 2–5 years, 5+ years).

In order to conduct a spectroscopic survey, we must first decide which objects are potential candidates, based on images gathered by other instruments in several color bands. This "targeting" process builds up a catalog of potential objects that can then be selected for spectroscopy. The raw images used in the targeting process are actually a larger data volume than the raw DESI data itself.

DESI will have 5,000 optical fibers positioned mechanically on the focal plane of the 4-meter Mayall telescope at Kitt Peak National Observatory near Tucson, Arizona. For each exposure of 15–20 minutes, the fibers are positioned to point at "objects of interest" selected from the targeting data, as well as empty sky (used for a reference in the data processing). The light travels down these fibers to 10 spectrographs, each fed by 500 fibers. Each spectrograph splits the light from a fiber into three color bands, and each color band passes through a grating where the resulting spectrum is projected onto a narrow strip of a charge-coupled device (CCD). These 30 CCD images are compressed at the telescope before transfer off site.

Simulation activities supporting DESI range from large N-body and hydrodynamic modeling of a portion of the universe to detailed simulations of telescope data acquisition and processing, to high-level simulations that optimize targeting and observing strategies for extracting cosmological information. Network usage of large-scale astrophysical modeling is covered in Section 22 (*Cosmic Frontier Simulations*). Here we reference that document where needed. After the start of observations, serving of processed data products to the collaboration and the public will become an increasing use of network resources.

The tentative project road map for DESI is listed in Table 31. This fairly aggressive schedule has data acquisition beginning in 2018 and running through 2022.

Table 31. Current estimated dates for project milestones.

Critical Decision (CD)	Fiscal Year
CD-0, Approve Mission Need	2012
CD-1, Approve Alternative Selection	2013
CD-2, Approve Performance Baseline	2014
CD-3, Approve Start of Construction	2015
CD-4, Approve Project Completion	2018

20.2 Science Drivers — Next 2 Years

The current and near-term work of the collaboration is focused on construction and testing of the instrument, performing several types of simulations, and conducting targeting observations in order to plan out the survey.

Several types of relevant simulations are ongoing for DESI. Current work on large astrophysical simulations focuses on running software tools at scale and improving the

types of physics being modeled. See the simulation case study (Section 22) for more details on the challenges of data movement for such runs. We are conducting a second type of simulation that involves the generation of artificial raw data, with realistic noise and other detailed systematics as it would look after passing through our spectrographs. The purpose of these simulated data files is to verify that the instrument design created by the hardware team produces data that can be processed to meet our science goals. These simulations also serve as a test bed for new data reduction algorithms. A third type of simulation we are planning is a series of higher-level end-to-end simulations that go from artificial catalogs of objects all the way through the survey optimization, data acquisition and processing, and extraction of cosmological data. Each step of this end-to-end simulation will exclude enough detail to make it tractable to run many times, and this type of simulation will give us insight into the broad impacts of survey design and targeting choices.

The targeting effort for DESI consists of using data from a variety of sources to build up a catalog of objects that are candidates for our spectroscopic survey. Some of this data is archival, and some will be collected specifically for this purpose. In all cases, the targeting data is transferred to NERSC for processing.

20.2.1 Instruments and Facilities

All three types of near-term DESI simulations are primarily conducted at NERSC and the DOE Leadership Class Facilities. As much as possible, the outputs of simulations are analyzed on the same machine where the simulation was generated in order to minimize data movement. In the case of large astrophysics simulations, see the DOE HEP Cosmic Frontier Simulations case study (Section 22). Both the detailed instrument simulation/ processing tests and the high-level end-to-end simulations can be run and analyzed at NERSC. These two simulation types require only local network use between the machines at NERSC and the HPSS storage system.

Work on targeting during the next two years of operations involves processing large archival datasets, continued observations from the Palomar Transient Factory (PTF), a dedicated observing campaign using the Dark Energy Camera (DECam) installed at the Blanco telescope in Chile, and additional observations of the North Galactic Cap (NGC) from another instrument (TBD). Archival data from the Wide-Field Infrared Survey Explorer (WISE) satellite is stored at NASA's Infrared Processing and Analysis Center (IPAC) located on the campus of the California Institute of Technology. The primary mission of this satellite has ended, but an extension to the mission will likely be approved and this new data would begin flowing to IPAC in a few months. Data from PTF is acquired at Palomar Observatory and sent over a wireless relay to the San Diego Supercomputer Center (SDSC). Data from DECam is transferred from the Cerro Tololo Inter-American Observatory in Chile and made available from the National Optical Astronomy Observatory (NOAO) in Tucson, Arizona.

20.2.2 Process of Science

20.2.2.1 Simulations

Astrophysical simulations are performed at NERSC and the Leadership Class Facilities. These large runs usually dump their full state information for purposes of checkpoint/restart. For hydrodynamic simulations of the universe, these dumps can be many terabytes. Typically these full dumps are processed into a reduced set of parameters at each Grid point, and these reduced datasets can be transferred to NERSC and other locations (see Section 22 for more details).

Instrument simulations require only a modest amount of data, stored locally at NERSC. Typical inputs are simulated spectra of the different object types targeted by the survey. These input spectra are projected through a model of our spectrograph onto a CCD; noise and other systematics are added; and this simulation is calibrated and processed as if it were "real" data. Input data is transferred from HPSS tape storage to scratch disk at the beginning of such runs and the outputs are saved to HPSS afterward.

High-level simulations will make use of data products (such as object catalogs) that are also generated at NERSC. Intermediate data products are currently written to an HDF5 file on local disk, but this format may be changing.

20.2.2.2 Targeting

The WISE satellite acquired approximately 50 TB of data from its primary mission and this data is stored at IPAC. We have made a mirror of this data on the NERSC global filesystem. There is a significant chance that NASA will re-activate the WISE satellite for another three years of operation with half of its detectors. If this mission extension is approved, there will be an additional 75 TB (only part of this will be during the near-term time range) acquired that must be transferred to NERSC. After the data is transferred to NERSC, all files are scanned and metadata is written out to a simple flat-file format for later reading and querying. This metadata scheme will likely eventually move to a locally hosted database of some type.

The PTF transfers approximately 100 GB/day from the Palomar Observatory to SDSC over the High Performance Wireless Research and Education Network (HPWREN), that has a maximum throughput of 155 Mbps. From there, the data is sent both to NERSC and to IPAC. At NERSC, the data is indexed in a PostgreSQL database and is passed through a fast "real-time" processing pipeline to look for supernovae. At IPAC, the data is processed by a slower but more accurate pipeline and archived. For DESI targeting purposes, we will need to copy this archive (300 TB) to NERSC and index it for searching. We will also need to begin doing daily synchronization of this data from IPAC to NERSC.

There is currently a DESI proposal to use DECam to survey 9,000 square degrees of the sky for targeting purposes. This will consist of approximately 25 TB of uncompressed raw data. Previous experience with similar data indicates that a factor of at least 2 for lossless compression is likely possible. This data is mirrored from Chile to NOAO (Tucson). We will need to transfer this data to NERSC. Note that NOAO is currently mirroring 200 GB of

data daily between Chile, NOAO, and the National Center for Supercomputing Applications (NCSA) using their custom Data Transport System (http://www.aspbooks.org/publications/434/260.pdf).

The data from all of these instruments consists of compressed images, typically written as FITS data files (http://heasarc.gsfc.nasa.gov/docs/software/fitsio/fitsio.html). For a given patch of sky, we have images from multiple instruments and different color bands. We take all of these images and solve simultaneously for the maximum likelihood catalog of objects for that sky patch.

20.3 Science Drivers — Next 2–5 Years

During this time period, the collaboration will be continuing the simulation and targeting effort of the previous chapter, as well as beginning DESI observations on the Mayall telescope at Kitt Peak National Observatory.

Assuming the WISE mission extension is approved, the third and final year of data will be arriving during this time. In 2015, PTF will be shut down in order to upgrade the camera. Observations will then restart and be known as the Zwicky Transient Factory (ZTF). The daily data volume that is copied from SDSC will increase by a factor of 2.

DESI is scheduled to begin collecting data in 2018. In the months leading up to that, all of the optics, spectrographs, and supporting infrastructure will be installed at Kitt Peak. In addition to transferring data to NERSC, the data will likely be mirrored to another U.S. institution (TBD). After processing of raw data at NERSC, the output data products will be distributed to other institutions within the collaboration, as well as to the public.

20.3.1 Instruments and Facilities

Simulation activities will continue to use NERSC and the DOE Leadership Class Facilities. There will likely be a slight increase in the size of output data products from the large astrophysical simulations transferred between these facilities. Instrument simulations and end-to-end simulations will continue to operate using only local resources at NERSC. The ZTF instrument at the Palomar Observatory will be a substantial upgrade over PTF in terms of the speed in which it can survey the sky. Data from ZTF will continue to be sent to SDSC before being copied to NERSC. It is likely that the HPWREN wireless network will need to be upgraded to handle this increase in data volume. If not already completed, additional targeting observations of the NGC (instrument TBD) will be carried out to complement the DECam observations. DESI science data will begin to be transferred from Kitt Peak. Kitt Peak National Observatory is located in the mountains 55 miles southwest of Tucson, Arizona. It is connected to NOAO in Tucson by a 1 Gbps optical fiber connection. NOAO is in turn connected to the University of Arizona by a 10 Gbps connection.

20.3.2 Process of Science

The usage of simulations will be the same as in the last chapter.

20.3.2.1 Targeting

Continued observations from the WISE satellite are archived at IPAC and transferred daily to NERSC. Metadata from these images would likely be appended to whatever database is used to search targeting images for purposes of object extraction. This final year of WISE data would be about 25 TB. ZTF will generate approximately 200 GB of compressed raw data per day, which is transferred to SDSC and then copied to NERSC where its metadata is appended to the targeting database. For ZTF, we anticipate running a version of the full processing pipeline at NERSC, so there will no longer be a need to transfer processed data from IPAC. The additional data transferred from the NGC targeting observations will be approximately 14 TB of uncompressed data, but the instrument that will be used for this is not yet decided. By the start of the mission, the accumulated targeting data from all sources stored at NERSC will be slightly more than a petabyte.

20.3.2.2 Raw Data

Raw data transferred from the Kitt Peak Observatory consists mainly of CCD images capturing the output of the spectrographs. During the afternoon, calibration images are taken with the spectrographs looking at laboratory sources with known spectral properties. Through the night, images of projected spectra from astrophysical sources are acquired. Calibration data generally does not compress as well as the science data. Even with conservative estimates for the maximum number of observing hours, the maximum number of exposures per hour, and the compression rates for the astrophysical and calibration images, this amounts to less than 100 GB of (compressed) data per night. Typically this number will be even smaller (say 30 GB per night), but it is good to consider an upper limit on the daily data volume. If we assume 8 hours of observing, then this is approximately 30 Mbps, which is much less than the capacity of the slowest link (1 Gbps from Kitt Peak to NOAO). There should be no issue with near real-time transfer of the data to NOAO.

After the data arrives at NOAO, it is transferred to NERSC. From NERSC, it will likely be mirrored to another institution. At NERSC, the raw data is backed up to HPSS and is also uncompressed and processed. In addition to daily processing, we will likely perform periodic reprocessing of all data (e.g., after improvements to pipeline software).

20.3.2.3 Data Serving

After processing at NERSC, the output data products (spectra and redshifts) are available to other members of the collaboration. There will also be periodic data releases available to the public for download. It is challenging to estimate the amount of data that will be served through this mechanism. Our best estimates can be obtained by looking at the statistics for a current spectroscopic survey (BOSS). Over the course of 3 months, a dataset consisting of 58 TB of raw and processed data was available to the BOSS collaboration and the public. Over this time period, 32 TB of data were served to more than 300 unique IP addresses. Of this 32 TB, 40% of the traffic was to collaborators and the rest was to the public. If we assume an approximate DESI data size on disk of 200 TB

(compressed raw data plus processed, for a single year of data), we might expect to serve approximately 400–500 TB/year.

20.4 Beyond 5 Years — Future Needs and Scientific Direction

DESI observations are planned to run from 2018–2022. During this time period, there will be the ongoing transfer of raw data from NOAO to NERSC and mirroring to another institution. Most targeting data should be completed by the start of observations. Large astrophysical simulations will continue to be necessary for interpreting the physical conditions in the universe that gave rise to our observations (see Section 22). There will be an approximate 50x increase in the data volume of these simulations during this time.

Serving of processed data will continue to be a major use of the network for DESI. In the summary table for this time period, we assume a model similar to BOSS. In this scenario, each (yearly) data release includes the new data and a reprocessing of the previous years' data. For this situation, we might assume that external institutions and users will want to transfer the new version of the reprocessed archival data as well as the newly acquired data. This is likely an upper limit on the data volume to be served. One can easily imagine the annual data volume served to exceed a petabyte.

20.5 Network Tools and Services

Although the raw bandwidth demands of DESI (outside of large astrophysical simulations) are not too onerous, there are several areas in which the project would benefit from stronger coordination with ESnet. Currently we do make use of Globus for data transfer needs within our collaboration. Distribution of data to the public is still typically performed with rsync, wget, etc. We are interested in improving this situation and learning about new tools that may make our data management tasks easier.

Another area that has been historically challenging is “the last mile,” which deals with debugging firewalls and network bottlenecks at the “endpoints” of ESnet. Although these endpoint problems are outside of ESnet’s network domain, it would be useful to have contacts within ESnet for assistance with such issues.

20.6 Summary Table

Key Science Drivers			Anticipated Network Needs	
Instruments and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
0–2 years				
• NERSC and the LCFs are used for simulations. • Targeting data transferred from IPAC (Caltech) to NERSC. • Targeting data transferred from NOAO to NERSC.	• Generate N-body and hydrodynamical simulations. • Instrument and end-to-end simulations. • Targeting data processed at NERSC.	• Workhorse Level 2 simulation datasets are 20 TB (100/year), and heroic runs are 1 PB (2/year). • Other simulations are few tens of TB total. • WISE data: 50 TB one-shot plus 70 GB daily. • PTF data: 300 TB one-shot plus 100 GB daily. • DECam 13 TB total in daily chunks.	• Tens of minutes per small astrophysical Level 2 dataset. • Order minutes for other smaller simulation data. • Targeting data spinning at NERSC (no LAN requirements).	• Astrophysical simulations need few TB / hour for transfers. • Daily targeting transfers must complete in less than a day.
2–5 years				
• NERSC and the LCFs are used for simulations. • Targeting data transferred from SDSC to NERSC. • Targeting data transferred from IPAC (Caltech) to NERSC. • DESI raw data transferred from NOAO to NERSC. • NERSC serves data to the world.	• Generate N-body and hydrodynamical simulations. • Targeting data processed at NERSC. • Raw DESI data processed at NERSC. • Processed and raw data served to public.	• Workhorse Level 2 simulation datasets are 20 TB (100/year), and heroic runs are 1 PB (2/year). • ZTF data: 200 GB daily. • WISE data: 70 GB daily. • DESI data: less than 100 GB daily. • 400–500 TB served to public per year.	Tens of minutes per small astrophysical Level 2 dataset.	• Astrophysical simulations need few TB/hour for transfers. • Daily targeting transfers must complete in less than a day. • Daily raw data transfers must be limited only by connection to mountain.

Key Science Drivers			Anticipated Network Needs	
Instruments and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
5+ years				
• NERSC and the LCFs are used for simulations. • DESI raw data transferred from NOAO to NERSC. • NERSC serves data to the world.	• Generate N-body and hydrodynamical simulations. • Raw DESI data processed at NERSC. • Processed and raw data served to public.	• Workhorse Level 2 simulation datasets are 20 TB (100/year), and heroic runs are 1 PB (2/year). • DESI data: less than 100 GB daily. • 0.5–1 PB served to public per year.	• Tens of minutes per small astrophysical Level 2 dataset.	• Astrophysical simulations need few TB/hr for transfers. • Daily raw data transfers must be limited only by connection to mountain.

21 Large Synoptic Survey Telescope (LSST)

The Large Synoptic Survey Telescope (LSST) is the most ambitious survey currently planned in optical astronomy. LSST will have unique survey capability in the faint time domain. The LSST design is driven by four main science themes: probing dark energy and dark matter, taking an inventory of the solar system, exploring the transient optical sky, and mapping the Milky Way.

The telescope and site infrastructure, together with the data management system, are a proposed Major Research Equipment and Facilities Construction (MREFC) project of NSF. The 3.2-gigapixel camera is a proposed Major Item of Equipment (MIE) project of the DOE Office of High Energy Physics (HEP), managed by the SLAC National Accelerator Laboratory. The LSST Project (hereby referred to as "the Project") is jointly governed by the two funding agencies. Construction is currently planned to begin in FY 2014, commissioning in FY 2020, and regular operations in FY 2022.

Within HEP, the LSST is a response to the CD-0 declaration of mission need for a Stage 4 dark energy experiment, and as such, HEP is planning to fund the dark energy data analysis and research (which is not itself part of the NSF MREFC and DOE MIE projects) and the required computing resources to carry it out. An LSST Dark Energy Science Collaboration (DESC) has been formed to plan for and coordinate this research, incorporating collaborators with HEP funding (at national laboratories and universities) and others supported by other sources, notably NSF Astronomy (AST).

The Project and DESC are described separately in this document.

The Project covers the construction and operation of the telescope and camera, and the generation and service to users of a calibrated image archive and a catalog database of observations and measured properties of detected objects. The networking needs of the Project are defined to be covered as part of the NSF-funded construction and operating costs of the Project, and as planned do not involve any operational role for ESnet.

The DESC covers the analysis of the catalog and image archives to produce a wide variety of measurements of the phenomenon of dark energy. Since this work will involve substantial transfers of data to and among DOE national laboratories and HEP-funded university groups, we currently envision a significant role for ESnet.

An immense variety of other science investigations will be enabled by the LSST dataset as well, but these will not be covered here.

21.1 Background

The LSST telescope and camera will be a large, wide-field, ground-based system designed to obtain multiple images covering the sky that is visible from the summit of Cerro Pachón in Northern Chile. The current baseline design, with an 8.4 m (6.7 m effective) primary mirror, a 9.6 deg^2 field of view, and a 3.2-gigapixel camera, will cover the sky using pairs of 15-second exposures twice per night, enabling the seasonally visible sky to

be covered completely every 3.5 nights on average, with typical 5σ depth for point sources of $r \sim 24.5$ (AB). The survey will continue for 10 years.

The system is designed to yield high image quality as well as superb astrometric and photometric accuracy. The total survey area will include 30,000 deg^2 with $\delta < +34.5°$, and will be imaged multiple times in six bands, *ugrizy*, covering the wavelength range 320–1050 nm. In the core survey area of 18,000 deg^2 coaddition of 825 or more visits to each location on the sky will enable the creation of a map and database reaching $r \sim 27.5$ (AB).

Image data (about 15 TB/night) will be transferred in real time from Cerro Pachón to a base site in La Serena, Chile, for archiving, and on to the National Center for Supercomputing Applications (NCSA) in Urbana, Illinois, for archiving and immediate processing. Within 60 seconds of acquisition, image data will be analyzed for transient events and alerts sent out to enable follow-up at other observatories. Annually the full cumulative dataset will be processed to create co-added image maps and detailed catalogs of detected astronomical objects. The image and catalog data will be served to users at two Project-funded Data Access Centers at NCSA and at the base site in La Serena.

By the tenth year of the survey, a single copy of the full raw image data will amount to 22 PB. The final processing of the full survey will yield stored co-added image products amounting to at least 11 PB, and a catalog database of 3 PB. In addition, 63 PB of individual calibrated images will be available for re-creation on demand.

21.2 Collaborators

The Project, supported by both DOE HEP MIE and NSF MREFC funds, is a collaboration of laboratory and university groups, with significant elements of its construction contracted out. Operations will also be funded jointly by NSF and DOE, with additional contributions from non-U.S. sources anticipated.

The Project Office in Tucson, Arizona, is organized as a Center under the Association of Universities for Research in Astronomy (AURA). It provides overall management and systems engineering for the Project, as well as receiving and managing the NSF MREFC construction funds for the telescope and site, and data management subsystems.

The NSF project involves the National Optical Astronomical Observatory (NOAO), NCSA, the Caltech/JPL Infrared Processing and Analysis Center (IPAC), SLAC (in a work-for-others role), and several universities (especially in the data management area). We are evaluating the possibility of the French CC-IN2P3 computing center playing a major role in providing additional computing infrastructure, and performing a major fraction of the annual data release processing.

The camera construction project is managed by SLAC, with major contributions from BNL and other DOE labs and DOE-funded university groups.

The Project's wide-area networking components are currently assigned to the NSF-funded part of the construction project, with no baseline role for ESnet. We are

nevertheless investigating the possibility of ESnet being able to provide cost-effective path diversity for the Project's WAN connections.

21.3 Key Science Drivers, Intra-Project (Local and Remote Combined)

We combine local and remote aspects of the core Project here because it is inherently a distributed effort.

21.3.1 Instruments and Facilities

The primary operational facilities will be the Cerro Pachón Summit Facility, the base facility and Data Access Center in La Serena, Chile; the archive facility (and principal processing site) and Data Access Center at the NCSA in Urbana, Illinois; and a yet-to-be sited headquarters facility in the continental United States. The CC-IN2P3 computing center is planned to provide a substantial fraction of the annual data release processing computing requirements, with data transferred by network.

The installation at NCSA will be housed in the National Petascale Computing Facility.

Principal data flows will be from the summit to the base, over a Project-controlled dedicated optical fiber link; from the base to the archive and back over international fiber links and domestic connections in Chile and the United States; from the archive to CC-IN2P3 and back;, and from the Data Access Centers to public and research Internets as needed to support access by the user community.

All computing and storage required to meet the requirements of the Project will be funded by the planned construction and operations funding, and are planned to be provided as Project-owned dedicated facilities (with the exception of the CC-IN2P3 collaboration). The use of commercial cloud computing or storage is not in the baseline plan, but is regularly evaluated as a cost or performance optimization.

21.3.2 Software Infrastructure

The software required to perform local processing and data transfers and support the necessary high-performance database will primarily be LSST-developed, released under an open-source license. Significant portions of it are expected to be used in other optical imaging astronomy projects, notably Hyper Suprime-Cam at the Japanese Subaru telescope.

A variety of components of the software are expected to be obtained from other open-source efforts, particularly in the areas of messaging middleware (e.g., ActiveMQ), workflow management (e.g., HTCondor), and global file collection and data transfer management (e.g., iRODS).

The specific software to be used to perform high-throughput WAN transfers has not yet been identified.

The Project's software development effort is organized using tools such as *git* to support a distributed team working from a number of remote sites.

21.3.3 Process of Science

Raw image data from the telescope and camera are analyzed on the Project-provided infrastructure described above, with the above software, to produce a variety of data products, ranging from alerts for detected transient phenomena to the calibrated images, image co-additions, and catalogs mentioned above. This all requires corresponding levels of network bandwidth within the NCSA-based LSST computing facilities.

21.4 Key Remote Access Science Drivers

21.4.1 Instruments and Facilities

The LSST Project expects to make use of a variety of external datasets to augment its data analyses, especially for astrometric and photometric standards to support calibrations. The largest such dataset will probably be the astrometric catalog produced by the European Space Agency GAIA mission.

Beyond the Project-provided user computing allocations at the LSST Data Access Centers, we expect major user-driven analysis efforts based on LSST data to occur at many remote locations, requiring the transfer of large quantities of data from the Data Access Centers to these sites. Even when analyses occur entirely on Project-provided resources, we expect these to involve large numbers of simultaneous connections from remote sites. These interactions with the Data Access Centers will arise from U.S. universities, laboratories, and national computing facilities, and comparable non-U.S. institutions.

21.4.2 Software Infrastructure

These interactions will be carried out with a mix of LSST-provided software tools and other community-provided tools, especially those supporting Virtual Observatory (VO) protocols. LSST is committed to providing VO-based access to its data to the greatest extent consistent with the availability, performance, and maturity of the VO standards and associated community-developed tools.

21.4.3 Process of Science

Remote scientists will perform queries, request data downloads, request data analyses at the Data Access Centers, create visualizations on both Data Access Center and user facilities, and the like.

Astronomers and physicists will access the data from the Project-provided Data Access Centers using Project-developed graphical and programmatic user interfaces to perform queries and retrieve data of relevance to their scientific investigations. The Project provides an allotment of computing and storage capacity within the Data Access Centers for further user-directed analyses and creation of derived data products.

21.5 Science Drivers, Intra-Project (Local and Remote Combined) — the Next 2–5 Years

The Project is currently in its final design phase. The proposed start of construction is in the second half of FY 2014; construction will continue throughout the specified 2–5 year time frame.

21.5.1 Instruments and Facilities

During this period, telescope and site construction activities will proceed on the summit, in La Serena, and in Tucson, as well as at contractor facilities.

Camera construction activities will be concentrated at SLAC and BNL.

National computing facilities, including NERSC and various Extreme Science and Engineering Discovery Environment (XSEDE) sites, are providing computing capacity to perform a variety of simulations and to apply the Project's prototype software to their results.

21.5.2 Software Infrastructure

A prototype release of much of the data reduction pipeline software already exists and is the subject of active development, supported by NSF R&D funds and in-kind contributions from collaborating institutions and private donors. Development to date has been focused on risk reduction for the advanced scientific algorithms and database technologies required.

With the start of construction, a larger software development effort will begin. By the end of the specified 5-year period, we expect the Project software to be nearing readiness to support the commissioning activities that will immediately follow.

21.5.3 Process of Science

Scientific activities during the next 2–5 years will be dominated by the analysis of test data from the Project's components, notably the CCD sensors and electronics assemblies as they come together, and the performance and analysis of simulations.

21.6 Remote Access Science Drivers — the Next 2–5 Years

Apart from the distributed software development activity mentioned above, we do not expect a large requirement for remote access for non-Project users during this period.

21.7 Beyond 5 Years — Future Needs and Scientific Direction

At this point, the Project will be entering its commissioning and operations phases, and will reach the form described in Sections 21.1 through 21.4 and in the Summary Table (Section 21.12).

21.8 Network and Data Architecture

For the terabyte-size volumes of LSST data and data products that must be transferred — with 2-second latency — every night and day from the telescope on Cerro Pachón to the Base Center in La Serena, Chile, and then onward to the Archive Center in Champaign, Illinois, the LSST networks must achieve 98% or higher mean data transfer reliability. The maintenance of the low latency is essential to the astronomical usefulness of the transient event detections performed by LSST. Therefore, path diversity, buffering, and redundancy are required to the extent economically feasible, and there must be spare capacity on the order of the normal bandwidth to "catch up" in the event of a failure or slowdown.

21.9 Collaboration Tools

The Project makes heavy use of desktop sharing and audioconferencing tools. For desktop sharing we use GoToMeeting and ReadyTalk (primarily through ESnet-provided accounts), and for audioconferencing we use AT Conference and ReadyTalk, and sometimes GoToMeeting.

Videoconferencing has thus far been found to be less useful, though it is occasionally used for small groups and for particular kinds of meetings for which it seems most helpful. We have found GoToMeeting's service to be easy to use and of reasonable quality.

21.10 Data, Workflow, Middleware Tools, and Services

We are currently using Globus and are evaluating a variety of other tools. We have baselined the use of HTCondor for workflow management, with an LSST steering layer on top, but we are evaluating a variety of other existing tools.

LSST is expected to require state-of-the-art tools in all these areas to support its data rates and user access requirements.

The Project does not currently plan to use cloud services, except in public outreach efforts, but will continue to consider them if cost savings or performance improvements could be achieved.

21.11 Outstanding Issues

None at this time.

21.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• Data come from existing astronomical datasets (e.g., SDSS Stripe-82) and public data from precursor surveys (e.g., DECcam). • Additional data come from simulations of the LSST system at a variety of national computing facilities. • NCSA is the main data archive. Most production processing will occur at XSEDE sites. • Principal software base is the developing open-source LSST image processing code base.	• Analyze simulated datasets to estimate the performance of the Observatory and LSST software by comparison to known simulation inputs. • Analyze precursor datasets to compare with results from existing software, and to enable new astronomical research with the new features of the LSST code.	• Single LSST exposures are 6 GB. • Chunk size of LSST data is a single CCD (32 MB). SDSS file sizes are in the range of 0.3–2.5 MB, with the total Stripe-82 dataset containing ~20 M files. • Datasets range from tens of GB to 13 TB (SDSS Stripe-82). • Largest LSST simulation runs will contain a few hundred files. • Returned data from image analysis, in catalog form: ~2 TB for SDSS Stripe-82.	LAN times for production are derived from the need to process a full precursor dataset (~13 TB) in about a month. This includes intermediate data products about 10 times larger than the input data.	• WAN transfers arise primarily from the movement of data from NCSA to XSEDE sites for processing and back. • In addition, we will need to perform international network bandwidth testing, aimed at the eventual LSST requirement to transfer 6 GB images in ~4 seconds from CTIO (La Serena, Chile) to NCSA.
2–5 years				
• Very similar. There will be increased availability to public data from the DES survey and the Subaru HSC instrument. LSST simulations will increase in fidelity and volume to support software development. • Camera I&T data will need to be transferred between SLAC and BNL.	• LSST software will reach an advanced state of development, with world-leading image analysis capabilities. • We will progress toward occasional near-full-scale tests of data movement and processing.	Testing the Alert Production function of LSST will require the processing of one 6 GB raw image every 18 seconds, including WAN transfers.		

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
5+ years				
• Beginning in 2019, data from the LSST Commissioning Camera will be available. • Initial data from the full camera will be available in 2020. • The French CC-IN2P3 computing center in Lyon has proposed taking on a role in performing roughly half the annual data reprocessing.	Two major activities: • Verifying that the LSST software meets its design requirements, primarily using simulation-based tests. • Facilitating the commissioning of the Observatory using the Commissioning Camera, at first, and then the full LSSTCam. This includes commissioning of the LSST software against real data and its unique artifacts.	• Commissioning Camera images will contain 9 CCDs, with a 32-MB raw image file for each CCD. • The full camera images will contain 201 CCDs (189 in the science array). • Images from either camera will be acquired roughly once every 18 seconds for extended periods during the night and day. • Each image acquisition will involve the transport of the raw data and also a crosstalk-corrected version of the image, of the same size. • A full night of Commissioning Camera operations will produce about 600 GB each of raw and crosstalk-corrected data.	• For live data processing the 6 GB image datasets will be required to move in about 2 seconds within sites and from the mountaintop to the La Serena, Chile project facility. • In annual reprocessings, intra-site bandwidth needs will begin at about 120 GB/sec by the first year of the survey (2022) and reach about 600 GB/sec by the end of the 10-year survey.	• For live data processing, the 6 GB image datasets will be required to move from La Serena, Chile, to NCSA in ~4 seconds. • The final multisite complex is described in the narrative. • WAN transfers of data to CC-IN2P3 will require moving the entire raw dataset (~15 TB/night) to Lyon on a regular basis. Production outputs returned to NCSA will total roughly 3.6 PB in 2022, with a similar amount copied from NCSA to CC-IN2P3 to complete a mirroring. This should increase to about 9 PB by the end of the survey.

22 DOE HEP Cosmic Frontier Simulations

22.1 Background

Large-scale simulations are essential for analyzing and interpreting results from cosmological surveys, as well as aiding in survey design and optimization, and in propagating errors through nonlinear processes. In this use case, supercomputers function as data-generating instruments. Typically, the data flow consists of two streams, (1) within the facility where the simulation is run (supercomputer, storage, analysis engine) and (2) from the host facility to a remote analysis/archive site. The data motion may be staged in a scheduled manner, or can be highly bursty, depending on the use case. In future, control of the data flow may be centralized or distributed to a few "trunk" sites.

22.2 Collaborators

All the HEP laboratories (ANL, BNL, Fermilab, LBNL, SLAC) participate in this work, as well as other Labs (Los Alamos National Laboratory [LANL], LLNL) and a number of collaborating universities including Berkeley, Chicago, Harvard, Illinois, Penn State, Stanford, Washington, and Yale. Computational facilities used include the Argonne Leadership Computing Facility (ALCF), NERSC, and the Oak Ridge Leadership Computing Facility (OLCF) along with institutional resources available at the participating units. The number of users varies from institution to institution, on the order of 10 to a few. The data products generated by simulation groups are used downstream by a large number of users who are members of science working groups in cosmological surveys and experiments. In future, the ratio of the user community to the size of the simulation groups will only increase as the role of simulations becomes more important. Thus the data flow pattern of case (1) above will likely become more significant and will need to be actively managed.

22.3 Key Local Science Drivers

22.3.1 Instruments and Facilities

The major compute facilities used are primarily those at the ALCF (Mira, Tukey, Eureka), NERSC (Carver, Edison, Hopper), and OLCF (Titan). Smaller local resources are available at ANL, BNL, Fermilab, SLAC, and Yale. Institutional storage ranges from approximately 100 TB at SLAC, 100 TB at ANL, to less than 10 TB at other sites. Internal data transfer rates range from hundreds of gigabytes per second for supercomputer I/O bandwidths to as low as 10 Gbps internal institutional links.

22.3.2 Software Infrastructure

This response also covers 22.4.2.

Software infrastructure varies widely. The use of code repositories is now widespread. Higher-level workflow tools to manage simulations (e.g., PDACS, SMAASH) and simulation analysis (including remote analysis) are slowly emerging, although most of this

work is still performed by handwritten scripts. Data transfers are performed primarily through GridFTP. Use of Globus as a SaaS for file transfer has proven very effective both within institutions and externally. A number of tools are used to process datasets, including embedded capabilities (e.g., data compression within I/O). The use of data containers (HDF5, PnetCDF) is sometimes limited by their reliance on MPI-IO; HDF5 is also considered too complex by some users. Native I/O tools written for simulation codes still obtain the best performance.

22.3.3 Process of Science

This response also covers 22.4.3.

The simulation results can be viewed as corresponding to three levels. Level 1 is the raw data from the simulations, Level 2 is intermediate-level analysis data, and Level 3 is Level 2 data reduced to the level of catalogs/databases (when possible). Level 1 analysis can be performed both in *in situ* or post-processing modes, while Level 2 and Level 3 analyses are primarily in post-processing mode. All three levels are part of the local and remote "process of science." Level 1 and 2 data analyses involve batch processing, while Level 3 analyses can have a significant level of interactivity, so the data access patterns can be quite different.

In situ analysis uses the supercomputer's own network, whereas Levels 2 and 3 analyses can be conducted locally or remotely and may involve moving data from file systems back to the host supercomputer or to analysis resources. Data at Levels 2 and 3 may also be moved over in batches to remote sites where it can be locally analyzed. Note that the actual mass of data at the three levels is roughly similar, except that the granularity increases significantly from Level 1 to Level 3.

22.4 Key Remote Science Drivers

22.4.1 Instruments and Facilities

As already stated, the major compute facilities used are primarily those at the ALCF (Mira, Tukey, Eureka), NERSC (Carver, Edison, Hopper), and OLCF (Titan). Storage is provided at the facilities at the level of tens of terabytes per subproject of disk and substantially more on tape (but this is not used much because of latency issues). Exceptions include a data-intensive science pilot project at NERSC (300 TB of disc), special dispensations for ALCC and INCITE projects (e.g., at ANL). The major data sources are the supercomputers and other associated compute resources (analysis and visualization clusters, data-intensive computers). The use of cloud resources may also be considered in the near future, both commercial and institutional. Networking is typically via Internet2 and ESnet at 100 Gbps.

The current datastream involves moving about 10–100 TB of bulk data. The much smaller "user" downloads to local storage typically involve less than 1 TB chunks of data. Progress in remote visualization methods should allow (almost) real-time visualization of many large datasets. Examples of this already exist.

22.4.2 Software Infrastructure

See 22.3.2.

22.4.3 Process of Science

See 22.3.3.

22.5 Local Science Drivers — the Next 2–5 Years

22.5.1 Instruments and Facilities

No major changes beyond incremental growth. There will be a major growth in simulated data and the need for analyzing this will almost certainly be passed on to a few facilities. Local growth will be limited (sustainability argument).

22.5.2 Software Infrastructure

This also contains a partial response to 22.6.2.

There will be an evolution in workflow tools to manage local and remote simulation data analysis (along the lines of IPython, PDACS, and other tools). This area is still in flux, and community desires are only now being captured in a design process. Our current major activity is in PDACS (Portal-based Data Analysis services for Cosmological Simulations), a data flow programming model for managing both local and remote data analysis tasks.

22.5.3 Process of Science

No major qualitative changes, except in data size and much more use of simulated data by projects.

22.6 Remote Science Drivers — the Next 2–5 Years

22.6.1 Instruments and Facilities

As to the supercomputers themselves, we expect significant architectural changes, but not disruptive changes with respect to the networks.

Over this time span, we expect and require a few major simulation data archive/analysis centers to emerge; most of the data will eventually be hosted there, They should also have substantial local analysis computing available (since computing must follow the data). In this case, the data stream will consist of two major types: (1) “feeders” to the data centers, moving about a petabyte, and (2) data-center-to-“user” links that would typically move less than 10 TB chunks of the data itself. Remote visualization methods should allow real-time visualization of many large datasets.

22.6.2 Software Infrastructure

See also 22.5.2.

We expect some changes to the simulation software but they are unlikely to be very significant (at least partly due to inertia, because of the size of the current software

base). Major changes should be expected beyond 2018, however. We expect only evolutionary changes in the tool infrastructure.

22.6.3 Process of Science

No major changes, except in size as noted in 22.5.3.

22.7 Beyond 5 Years — Future Needs and Scientific Direction

It is expected that beyond 5 years, the simulated datasets will be very large, with Level 1 data volumes hitting 20 PB/year, with possible reduction by a factor of 4 due to compression and analytics post-processing. Significant high-throughput computing will need to be done on these datasets, but because they cannot (most often) be broken up into small chunks, the Grid computing model will not be effective. To address problems of this kind, data-intensive computing facilities are expected to undergo a sea change. This will drive disruptive changes in the hardware and software environments.

22.8 Network and Data Architecture

High-performance data transfer will be of significant importance, especially with regard to scheduled transfer of large datasets from computational to storage/analysis centers. The shared data volume that would need to be transferred is estimated to be about 2 PB/year. It is likely that individual institutions will develop their own responses as well, so it is important to have a unified strategy to the extent possible. In particular, the ideas behind the Science DMZ appear quite attractive.

22.9 Collaboration tools

Videoconferencing has been found to be very useful and Adobe Connect, Polycom, and Skype have all been good performers. (We have no experience with ECS.) Any improvements in this area will have a significant impact. Teleconferencing through free services and providers such as ReadyTalk is reasonably effective.

22.10 Data, Workflow, Middleware Tools, and Services

Much of this has already been covered. As far as cloud computing and storage is concerned, it is primarily a question of latency, bandwidth, amount of associated computing, and cost. Primarily due to cost issues, we do not see this option as viable currently. But this situation could change with time. There are several attractive features of the cloud model, including virtualization (not something that we normally exploit).

22.11 Outstanding Issues (if any)

The security protocols used at a few centers make the use of Globus impractical. This should be fixed.

22.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• Carver, Edison, Eureka, Hopper, Mira, Titan, Tukey. • Simulation codes: ART, Gadget, HACC, Nyx; many analysis tools. Data management: SMAASH, PDACS, scripts. Data movement: Globus Online, GridFTP.	In-place analysis of data on supercomputers, next-level analysis on clusters, final analysis on local resources (if needed).	• Datasets range in size from few TB to few PB per simulation. • Typical number of files in Level 1 dataset is equal to the number of I/O nodes used, so varies from tens of files at ~100 GB/file to thousands of files at similar sizes.	• Tens of minutes to transfer large files to the file systems at ~100 GB/sec. • Local transfers of various sizes at ~1–100 Gbps.	• Typically, few TB/hour, intermittent data transfers on weekly/monthly timescales of tens of TB. • Most major data exchanges have been between ANL and LBNL (~100 TB total).
2–5 years				
• New supercomputers on the 2017–2018 time frame. • No major change in software environment.	Aim to set up a few data storage /analysis sites to function as data hubs.	• Size of datasets will go up by a factor of 10. • File sizes will not increase much, so file numbers will likely increase.	Time to transfer large files will remain unchanged even as files get larger	• Order-of-magnitude increase in throughput expected. • Collaborating sites will increase to ANL, BNL, Fermilab, LBNL, ORNL, SLAC, university transfers will be subdominant.
5+ years				
Exascale systems in the early 2020s.	Difficult to predict in any detail, expect the data hub model to remain viable, most analysis will be remote.	• Further order-of-magnitude increase. • Too hard to say what will happen with files, depends on the I/O hardware among other things.	Hard to predict	• Not clear what throughput will be needed, if hardware changes radically. • Collaborating sites will be ANL, BNL, Fermilab, LBNL, ORNL, SLAC.

23 Computational Cosmology

23.1 Background

The computational cosmology community at SLAC participates in multi-institution collaborations that generate simulation data products measured in tens of terabytes and increasing in size as computational resources grow. These simulation activities are needed to support extracting the science from observational surveys (DES, LSST) as well as laboratory and accelerator experiments. These data require routine but infrequent transfers between institutions and SLAC. That is, work in this field requires local and remote management of many such datasets.

23.2 Collaborators

Participating DOE collaborators are at SLAC, Fermilab, BNL, and LBNL. Many other collaborators and institutions worldwide are involved. A partial list is below:

- Tom Abel — Kavli Institute for Particle Astrophysics and Cosmology
- Marcelo Alvarez — CITA, University of Toronto, Ontario, Canada
- Raul E. Angulo — CEFCA, Spain
- Michael Busha — Institute for Theoretical Physics, University of Zürich, Switzerland
- August E. Evrard — University of Michigan
- Oliver Hahn — Department of Physics, ETH Zurich, CH-8093 Zürich, Switzerland
- Ralf Kaehler — Kavli Institute for Particle Astrophysics and Cosmology
- Ji-hoon Kim — University of California, Santa Cruz
- Mark R. Krumholz — University of California, Santa Cruz
- Tony Li — Kavli Institute for Particle Astrophysics and Cosmology
- Yuexing Li — Pennsylvania State University
- Jonathan Mckinney – University of Maryland
- Michael L. Norman — University of California, San Diego
- Britton D. Smith — Michigan State University
- Matthew J. Turk — Columbia University
- Risa Wechsler — Kavli Institute for Particle Astrophysics and Cosmology
- John Wise — Georgia Institute of Technology
- Hao-Yi Wu — University of Michigan
- Hidenobu Yajima — Pennsylvania State University
- Qirong Zhu — Pennsylvania State University

23.3 Key Local Science Drivers

23.3.1 Instruments and Facilities

Simulations are being carried out at SLAC, NERSC, and at NSF XSEDE machines. Storage and data curation takes place at SLAC and NERSC.

23.3.2 Software Infrastructure

File systems: standard Linux (ext4), Lustre, XFS, ZFS

Transfers: GridFTP (mainly via Globus), bbcp, fdt

23.3.3 Process of Science

The main phases of producing science are the initial bulk simulations, storage management, post-processing, and visualization. The datasets are at the 10 TB scale (and growing) and these phases take place at multiple institutions. Hence the role of the network is paramount in enabling science. Providing convenient, routine access to data during these phases is key to future progress.

23.4 Key Remote Science Drivers

23.4.1 Instruments and Facilities

Researchers routinely transfer simulation datasets from one institution to another as part of the workflow from raw simulation to final results. Datasets are currently multi-terabyte with expected growth to multiple tens of terabytes. Storage at SLAC for these datasets is currently of order 500 TB.

23.4.2 Software Infrastructure

The collaboration primarily uses Globus currently, but has historically used bbcp, fdt and others for *ad-hoc* transfers.

23.4.3 Process of Science

Software development and midscale (less than 1,000 cores) simulations are done locally. Larger scale simulations are done at remote sites, such as NERSC or NSF XSEDE sites. Transfer is done to/from storage at SLAC to remote sites.

23.5 Local Science Drivers — the Next 2–5 Years

23.5.1 Instruments and Facilities

The modernization of cluster and storage architecture at SLAC will support larger datasets.

23.5.2 Software Infrastructure

23.5.3 Process of Science

23.6 Remote Science Drivers — the Next 2–5 Years

23.6.1 Instruments and Facilities

23.6.2 Software Infrastructure

23.6.3 Process of Science

23.7 Beyond 5 Years — Future Needs and Scientific Direction

23.8 Network and Data Architecture

We would like to see tools and capabilities that make data access and transfers transparent and simple for the science user. It may be that dataset curation is best at one location while some post-processing visualization capability is best at another location. Hence the "move the processing to the data" paradigm won't suffice and network capability is the key to advancing the science.

23.9 Collaboration tools

We use Skype and Google+ services as well as ReadyTalk and other ESnet services.

23.10 Data, Workflow, Middleware Tools, and Services

23.11 Outstanding Issues

None at this time.

23.12 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• Simulations currently being carried out at SLAC, NERSC, and NSF XSEDE machines. Storage and data curation takes place at SLAC and NERSC. • File systems: standard Linux (ext4), Lustre, XFS, ZFS. • Transfers: GridFTP (Globus), bbcp, fdt.	Software development and midscale (<1000 c) simulations done locally. Larger-scale simulations done at remote (NERSC, NSF XSEDE) sites. Transfer to/from storage at SLAC to remote sites.	• Size: 1 TB. • Range: 100 GB to 5 TB.	Transfer rate: 300 MB/sec.	• Transfer rate required: 100 MB/sec, 3 hours, 4 times a month. • Collaborating sites: NERSC, NSF XSEDE sites.
2–5 years				
No substantial changes expected except for hardware modernization.	No substantial changes expected.	• Size: 10 TB. • Range: 2–20 TB.	Transfer rate: 2 GB/sec.	• 2 TB/hr, 5 hours, 8 times a month. • Collaborating sites: NERSC, NSF XSEDE sites.
5+ years				
No substantial changes expected except for hardware modernization.	No substantial changes expected.	• Size: 40 TB. • Range: 4–100 TB.	Transfer rate: 5 GB/sec.	• 10 TB/hr, 4 hours, 10 times a month. • Collaborating sites: NERSC, NSF XSEDE sites.

24 Community Accelerator Modeling Using ACE3P

24.1 Background

Advanced accelerator modeling using high-performance computing has provided the capability for high-fidelity and high-accuracy simulations of accelerator structures and systems for the design, optimization, and analysis of accelerators. Running on DOE state-of-the-art supercomputers, parallel electromagnetic computation has enabled the design of accelerator cavities to machining tolerances and the analysis of large-scale accelerator systems to ensure accelerator operational reliability. The applications include existing and planned accelerators in HEP such as the LHC, Project X, the Muon Collider, and the dielectric laser accelerators; and in NP, the CEBAF Upgrade, RHIC, and FRIB.

24.2 Collaborators

The parallel electromagnetic simulation suite ACE3P developed at SLAC has had, for the past 15 years, a wide user base in the accelerator community both within DOE and beyond. ACE3P runs on NERSC computers and has been used by about 50 research scientists, engineers, and graduate students in six DOE national laboratories, four universities, and two private companies in the United States and at CERN for accelerator projects and applications.

24.3 Key Local Science Drivers

24.3.1 Instruments and Facilities

ACE3P is used for modeling of particle accelerators, in operation or under development, including colliders, high-intensity particle sources, and light sources. The ACE3P user community uses remote computers at NERSC for simulation, and local desktops for visualization and data analysis. Large datasets are stored in the NERSC HPSS archive and fetched back to servers at local institutions when analysis is performed. LANs at most institutions have similar transfer bandwidth, which is determined by current cable capabilities.

24.3.2 Software Infrastructure

The daily activities in individual local institutions involve preprocessing and post-processing. In preprocessing, the model is built for input to ACE3P using the third-party mesh-generation software Cubit, developed at Sandia National Laboratories. Post-processing includes the analysis and visualization of datasets generated by ACE3P simulations using the third-party visualization software ParaView distributed by Kitware Inc., and data analysis tools developed at SLAC. There is no need to transfer data between remote collaborators, as individual users perform their activities separately.

24.3.3 Process of Science

Scientists at individual institutions use desktops to visualize and analyze data that are stored on local hard drives or file servers. This process is important to determine the

properties of accelerator cavities, to find out potential problems, and to mitigate these effects to achieve optimized designs. The efficient visualization of fields and particles in complex geometries, in particular for large datasets, can be facilitated by a fast LAN.

24.4 Key Remote Science Drivers

24.4.1 Instruments and Facilities

The application modules of ACE3P run on NERSC computers and the remote access from individual institutions includes the transfer of preprocessing data to the remote compute site and the transfer of data for post-processing to local sites. The rate of data transfer from NERSC is adequate, while that to a local storage is determined by the hardware write capabilities. The write speed on local hardware varies from 100 MB/sec on desktop hard drives to 25 MB/sec on NFS file servers.

24.4.2 Software Infrastructure

The daily activities of using ACE3P for accelerator modeling in the wide area environment involve data transfer from NERSC computers to local storage at individual institutions. Because each institution uses the NERSC compute facilities for its own simulation problems independently, there is no data transfer management requirement between the institutions.

24.4.3 Process of Science

Scientists use the compute resources at NERSC to perform simulation for the design and optimization of accelerator cavities. The data generated from these simulations are transferred back to institutions across the country for analysis; hence, maintaining and enhancing adequate data bandwidth from the remote facility to these institutions are essential to the scientific process.

The data are moved from NERSC to individual institutions through Secure Copy Program (SCP). The data transfer is in the form of individual files, which can be visualized and analyzed independently in most cases. The analysis tools do not directly interact with data movement and the files in a dataset can be analyzed even though the transfer of the whole dataset is not complete. The performance is limited by the capabilities of LAN rather than those of the WAN. Currently the best write rate achieved on local hardware is 100 MB/sec. Therefore, for large datasets at the order of terabytes, the transfer time can be hours. To circumvent this problem, an alternative is being explored to process data and visualize results remotely on NERSC using the visualization tool ParaView running in parallel mode and then to send the images to local desktops for display.

24.5 Local Science Drivers — the Next 2–5 Years

24.5.1 Instruments and Facilities

In addition to those mentioned in 24.3.1, ACE3P may be ported to local clusters at individual institutions. This will shift the dependence of compute resources at NERSC to local institutions, especially for simulation of small- and medium-size problems.

24.5.2 Software Infrastructure

Similar to 24.3.2.

24.5.3 Process of Science

Similar to 24.3.3.

24.6 Remote Science Drivers — the Next 2–5 Years

24.6.1 Instruments and Facilities

Similar to 24.4.1.

24.6.2 Software Infrastructure

Similar to 24.4.2.

24.6.3 Process of Science

As mentioned in 24.4.3, the possibility of processing and visualizing data on a remote compute facility is being explored, without the need to transfer data from the remote site to local computer hardware. If this proves to be an efficient and robust process, it will improve the performance of analyzing simulation results, especially for the increasing size of datasets generated from large-scale accelerator modeling at the system level.

24.7 Beyond 5 Years — Future Needs and Scientific Direction

Accelerator modeling using ACE3P will continue to support the simulation needs of existing and planned accelerators. The simulation capabilities will be improved to perform integrated simulations, including multiphysics simulations to address design and engineering issues at the system scale. The simulation and the analysis of data will be carried out at remote supercomputer centers, and the software and local network infrastructure will be enhanced to facilitate this process.

24.8 Outstanding Issues

The process of remote processing and visualization of data will alleviate the demand for high-bandwidth requirements of data transfer, both in the LAN and WAN. To achieve this, a robust and streamlined interactive environment for this process at off-site computer centers should be developed to provide users a workflow superior to the current process of “first transfer and then analyze” at their local institutions.

24.9 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near Term (0–2 years)				
• Existing and future particle accelerators. • Parallel electromagnetic simulation suite ACE3P.	• ACE3P simulation on NERSC computers. • Preprocessing and post-processing on hardware at local institutions.	• Maximum size 1 TB/set • Range of dataset sizes from 50 GB to 1 TB (depending on simulation). • Dataset composition: 500,000 files, 2 MB each.	Time to transfer a dataset at 1 TB in 1 hour.	Time to transfer a dataset from remote site at 1 TB in 1 hour, a couple of times per day.
2–5 years				
• Existing and future particle accelerators. • Parallel electromagnetic simulation suite ACE3P.	• ACE3P simulation on NERSC and local computers. • Preprocessing and post-processing on hardware at local institutions. • Post-processing of large datasets on NERSC computers.	• Maximum size 5 TB/set. • Range of dataset sizes from 1 TB to 5 TB (depending on simulation). • Dataset composition: 10,000 files, 300 MB each.	• Time to transfer a dataset at 1 TB in 1 hour.	Time to transfer a dataset from remote site at 1 TB in 1 hour, a couple times per day.
5+ years				
• Existing and future particle accelerators. • Parallel electromagnetic simulation suite ACE3P.	• ACE3P simulation on NERSC and local computers. • Preprocessing and post-processing on hardware at local institutions. • Post-processing of large datasets on NERSC computers.	• Maximum size 25 TB/set. • Range of dataset sizes from 1 TB to 25 TB (depending on simulation). • Dataset composition: 50,000 files, 500 MB each.	Time to transfer a dataset at 2 TB in 1 hour.	Time to transfer a dataset from remote site at 2 TB in 1 hour, a couple times per day.

25 Lattice Gauge Theory

25.1 Background

The foremost goals of high energy physicists are to perform precise tests of the Standard Model (SM) of subatomic physics and to search for physical phenomena that require theories that go Beyond the Standard Model (BSM) for their understanding. Furthermore, despite the many successes of the SM, high energy physicists believe that a more general theory is needed to explain physics at the shortest length scales, or highest energies. Two complementary approaches are being used in these studies. One is to work at the Intensity Frontier (IF), where particle beams with the highest available intensities are used to make precise determinations of a wide range of physical quantities. The other is to work at the Energy Frontier (EF), where accelerators with the highest available energy (the LHC at present) are used to directly search for particles or other physical phenomena not included in the SM. Work done by lattice field theorists in the United States will support experiments at both of these frontiers. Precise lattice QCD calculations are required to determine some of the fundamental parameters of the SM, and in many cases to understand whether experiments at the IF are in agreement with the SM. Lattice investigations of theories that have been proposed for BSM physics are needed to determine which, if any, of them are compatible with the data coming from EF experiments at the LHC.

Quark-flavor experiments at the IF have historically played a key role, because they can probe energy scales far greater than those reached directly by the accelerators. In the coming decade, quark-flavor experiments will continue both at electron positron ($e^- e^+$) machines (BES III in China and Belle II in Japan) and the LHC, which has a dedicated *b*- and *c*-quark experiment, LHCb, as well as some b-physics in the central detectors, ATLAS and CMS. Furthermore, a new set of kaon experiments is being mounted: NA62 at CERN, KOTO at J-PARC in Japan, and the proposed ORKA experiment at Fermilab. To interpret these experiments, we need lattice QCD calculations of hadronic properties that are as precise as the experiments. Thus, lattice QCD calculations have provided and will continue to provide essential theoretical input to the experimental high energy physics program.

Other experiments at the IF will also depend on lattice QCD. The Muon g-2 Experiment at Fermilab expects a fourfold reduction in the experimental uncertainty of the muon magnetic moment. The leading theoretical uncertainties, stemming from hadronic contributions to the muon magnetic moment, will dominate the total uncertainty unless they are improved, and lattice QCD offers the only feasible path for doing so. Matrix elements of protons and neutrons are needed to interpret constraints on CP violation from limits on electric dipole moments, to aid the search for baryon-number violation in proton decay and neutron-antineutron oscillations, and even to guide searches for dark matter and axions at the Cosmic Frontier. Increased precision on the determination of the b-quark mass and the strong coupling constant α_s are essential for accurate prediction of Higgs production and decay at a future International Linear Collider (ILC)

and, thus, to test whether the 126 GeV Higgs is the SM Higgs or whether it has non-standard couplings. Lattice QCD calculations currently provide the most precise values of the b-quark mass and α_s, so it is essential to continue to improve our calculations.

25.2 Key Local and Remote Science Drivers

25.2.1 Instruments and Facilities

Lattice field theorists in the U.S. Quantum Chromodynamics (USQCD) use a wide variety of computers (mostly) in the United States. The DOE computing centers at ANL, LBNL (NERSC), and ORNL are all heavily utilized. In addition, USQCD has a PRAC (Petascale Resource Computing Allocations) grant to use Blue Waters at NCSA and a number of groups have additional allocations on XSEDE resources. Most significantly, the USQCD Computing Project supports hardware at BNL, Fermilab, and JLab dedicated to the study of lattice field theory. The USQCD Scientific Program Committee manages the allocation of these resources.

USQCD makes many of its gauge configurations available through the International Lattice Data Grid (ILDG) and the Gauge Connection at NERSC. Ensembles are also sometimes shared with international colleagues before they are available through ILDG or the Gauge Connection. This usually requires volunteer effort by a USQCD member to manage the transfer.

25.2.2 Process of Science

Quantum chomodynamics (QCD) is formulated in the four-dimensional space-time continuum; however, to carry out numerical calculations one must reformulate it on a lattice or grid. It should be emphasized that the lattice formulation of QCD is not merely a numerical approximation to the continuum formulation. Like most four-dimensional quantum field theories, QCD must be regularized to obtain its physical predictions, and the lattice regularization of QCD is every bit as valid as continuum ones. The lattice spacing a establishes a momentum cutoff π/a that removes ultraviolet divergences. Standard renormalization methods apply, and in the perturbative regime they allow a straightforward conversion of lattice results to any of the standard continuum regularization schemes.

Lattice QCD calculations proceed in two steps. In the first, one uses importance sampling techniques to generate gauge configurations, which are representative samples from the Feynman path integral that defines QCD. These configurations are saved, and in the second step they are used to calculate a wide variety of physical quantities. To obtain physical results, one carries out calculations for a range of small lattice spacings, and then performs extrapolations to the zero lattice spacing (continuum) limit. Furthermore, the computational cost of calculations rises as the masses of the quarks, the fundamental constituents of strongly interacting matter, decrease. Until recently, it has been too expensive to carry out calculations with the masses of the two lightest quarks — the up and the down quarks — set to their physical values.

Instead, calculations have been performed for a range of up and down quark masses, and extrapolated to their physical values guided by chiral perturbation theory, an effective field theory that determines how physical quantities depend on the masses of the lightest quarks. The extrapolations in lattice spacing (continuum extrapolation) and quark mass (chiral extrapolation) are the major sources of systematic errors in QCD calculations, and both must be under control to obtain trustworthy results. In current simulations, several groups are, for the first time, working at or near the physical masses of the up and down quarks. The gauge configurations produced in these simulations will greatly reduce, and eventually eliminate, the systematic errors associated with the chiral extrapolation.

A number of different formulations of QCD on the lattice are in use, all of which are expected to give the same results in the continuum limit. In recent years, major progress has been made in the field through the development of improved formulations (improved actions), which reduce finite lattice spacing artifacts. In the United States, the actions being used to study QCD include domain-wall quarks, the highly improved staggered quark (HISQ), and Wilson/Clover quarks.

Table 32 contains details of the MIMD Lattice Computation (MILC) HISQ ensembles, the most extensive set currently in use by USQCD. Most of the configurations have been generated except for those in the last line. Ensemble generation does not usually require much bandwidth as the configurations can normally be archived at the center at which they are generated. However, a second copy is made, and the ensemble is likely to be copied to another facility where one or more measurement codes will be run. It would be convenient to be able to move an ensemble to a new center in two weeks. The largest ensemble is 1.55×10^{14} bytes. This would require a transfer rate of 128 MB/sec or about 1 Gbps. It would not appear that lattice field theory network requirements will be taking much more bandwidth, unless there is a radical change in workflow such as extensive archiving or site-to-site transfers of quark propagators. One should keep in mind that moving an ensemble to a new facility to carry out analysis on the ensemble is not carried out on any fixed schedule, so network demands for lattice field theory tend to come in bursts. However, they also tend to be modest compared with high energy experiments.

Table 32. File sizes and archival storage requirement for MILC HISQ ensembles. The first two columns contain the spatial and temporal grid sizes, respectively. The third column contains the number of grid sites and the fourth column is the number of bytes in a single configuration. Each ensemble will contain 1,000 configurations. The last column contains the number of bytes of archival storage required for the ensemble. The entire dataset requires 235 TB.

N_s	N_t	Sites	Single Configuration (bytes)	Configs	Storage (bytes)
16	48	196608	5.66E+07	1000	5.66E+10
24	48	663552	1.91E+08	1000	1.91E+11
32	48	1572864	4.53E+08	1000	4.53E+11
24	64	884736	2.55E+08	1000	2.55E+11
24	64	884736	2.55E+08	1000	2.55E+11
32	64	2097152	6.04E+08	1000	6.04E+11
40	64	4096000	1.18E+09	1000	1.18E+12
48	64	7077888	2.04E+09	1000	2.04E+12
32	96	3145728	9.06E+08	1000	9.06E+11
48	96	10616832	3.06E+09	1000	3.06E+12
64	96	25165824	7.25E+09	1000	7.25E+12
48	144	15925248	4.59E+09	1000	4.59E+12
64	144	37748736	1.09E+10	1000	1.09E+13
96	192	169869312	4.89E+10	1000	4.89E+13
128	256	536870912	1.55E+11	1000	1.55E+14
Total bytes					2.35E+14

25.3 Local and Remote Science Drivers — the Next 2–5 Years

We certainly hope that U.S. computing capacity and that of the USQCD computing project will continue to grow over the next 5 years. This will allow us to pursue additional physics projects and several groups within USQCD will continue to generate larger ensembles. However, the MILC ensembles are currently the largest, and it is not clear that they will grow significantly beyond what is shown in the table. One significant enhancement would be the addition of dynamical electromagnetic effects. This would slightly increase the size of the ensembles, but it might not be necessary to do these studies with the finest lattice spacing in the table (last line). USQCD is projecting a tenfold increase in its archival storage over the next 5 years.

Significant growth is possible in the study of theories relevant for the breaking of electroweak symmetry. A number of such theories are currently being studied, but it does not seem like there is (yet) a leading candidate for the theory most likely to be applicable to nature.

25.4 Beyond 5 Years — Future Needs and Scientific Direction

Of course, one predicts the future at one's own peril. Algorithmic breakthroughs are impossible to predict, but can change future computing requirements. If computing requirements are reduced, storage requirements and transfer rates are likely to increase. Many quantities are under very good control with the lattice spacing we currently use. However, it is possible that some ensembles with smaller lattice spacing will be required for *b*-quark physics. It does not seem likely that we would decrease the lattice spacing by more that a factor of 2 in the next 5 years. The current spatial sizes are likely larger than needed just for *b*-quark studies, so a factor of 16 in storage beyond current largest ensemble should be an upper limit in the next 5 years. Beyond that period, it might be desirable to decrease the lattice spacing by another factor of 2 with another factor of 16 in storage and bandwidth requirements.

Progress in our area depends on advances in computing capability and capacity. It seems quite likely that new computers with greater capability will be installed in the next few years, and beyond that there is the hope of exascale computers. Such computers would certainly accelerate progress in lattice field theory, allowing generation of ensembles closer to the continuum limit, or greater statistical accuracy on ensembles comparable to what we now have. It would also make it possible to explore new algorithms more quickly.

25.5 Outstanding Issues

None at this time.

25.6 Summary Table

See table above.

26 Perturbative QCD and Phenomenology

26.1 Background

The Large Hadron Collider (LHC) is creating high demand for precise predictions for Standard Model reactions as backgrounds to new physics searches. At the same time, parameter spaces in various scenarios need to be scanned to identify possible signatures of Beyond Standard Model (BSM) physics. We typically simulate large event samples using MC event generators to satisfy both these needs. The event samples are stored in compressed text files or ROOT ntuple files, which are exchanged over the LAN or WAN. They are used by theorists and experimentalists to calculate observables without the need to redo the actual theoretical calculation, which is typically much more time-consuming than the analysis of the results.

26.2 Collaborators

One of the collaborations involved in this effort is the BlackHat project. BlackHat provides results of perturbative QCD calculations at the next-to-leading order (NLO) in form of parton level events, which are stored in ROOT ntuples.

Another collaboration is the Sherpa project, which constructs and maintains a particle-level MC event generator for collider physics, in particular LHC and ILC physics. Sherpa is part of MCnet, a European-Community-funded Marie Curie Research Training Network, which spans all major collaborations involved in the construction of MC event generators for current and future collider experiments. A related VO called “pheno” controls the Grid activities of this and related projects (see http://www.phenogrid.dur.ac.uk/).

26.3 Key Local and Remote Science Drivers

26.3.1 Instruments and Facilities

As the above-described projects are of purely theoretical nature, they do not involve any instruments. BlackHat and Sherpa can efficiently be run on tightly coupled computer systems, such as the Cray and IBM BlueGene architectures available at NERSC, OLCF, and ALCF. Other resources are local computing clusters at various DOE- and NSF-funded U.S. institutions and around the world. Sherpa can make use of Grid resources, and it has been successfully run on the OSG and the U.K. Grid for Particle Physics.

26.3.2 Software Infrastructure

The software packages used to manage daily activities are standard ftp and GridFTP tools. We have worked with BeStMan and iRods to transfer files from Grid Storage Elements distributed throughout the United States to a server at SLAC. The amount of data transferred is typically small in this case. We have transferred about 1 TB in the course of a few days.

26.3.3 Process of Science

We generate compressed text files or ROOT ntuple files, which contain particle information such as momenta and decay vertices as well as cross sections and statistical information. The total size of such files is about 1 TB per calculation for a typical process of interest, but it strongly varies with the complexity of the process. Files are transferred between the places where they are stored permanently (CERN, Durham [U.K.], UCLA, SLAC) and the places where they are analyzed (MSU, Baylor, etc.).

26.4 Local and Remote Science Drivers — the Next 2–5 Years and Beyond

26.4.1 Instruments and Facilities

We expect the computing infrastructure relevant to our activities to be qualitatively the same over the next years, with both local and remote facilities available for HEP theory to carry out more demanding calculations. More precise theoretical predictions are needed to enable the science at collider experiments, and therefore we expect a substantial increase in network traffic related to the production of the event files described above.

26.4.2 Software Infrastructure

We expect the relevant software structure to be qualitatively identical to what is available to date.

26.4.3 Process of Science

The process of science will not change in the course of the next years.

26.5 Collaboration tools

We are using collaborative tools, such as audio and video conferencing services (ReadyTalk, Skype, Vidyo).

26.6 Summary Table

Key Science Drivers			Anticipated Network Needs	
Science Instruments, Software, and Facilities	**Process of Science**	**Dataset Size**	**LAN Transfer Time Needed**	**WAN Transfer Time Needed**
Near, Mid and Long Term				
• HPC facilities (NERSC, ALCF, OLCF), local computing clusters, Grid (OSG, GridPP). • Standard ftp and GridFTP tools.	• Generation of event files containing theory predictions for collider experiments. • Transfer of these event files from/to collaborators and users for analysis.	• ~1 TB/set • 250 GB–3 TB per set, depending on complexity of calculation. • 1–2 GB per file.	1 TB in 4 hours, continued streaming for analysis.	• 1 TB in ~8 hours, occasionally. • Data exchange between collaboration sites (SLAC, MSU, Durham [U.K.], CERN, CEA Saclay, Freiburg, Munich, Dresden).

27 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

AAA	Any data, Any time, Anywhere
AAF	ALICE Analysis Facility
AD	antineutrino detector
ALCC	ASCR Leadership Computing Challenge
ALCF	Argonne Leadership Computing Facility
ALICE	A Large Ion Collider Experiment
AliEn	ALICE Environment
AMD	Advanced Micro Devices
AMGA	ARDA Metadata Grid Application
AMQP	Advanced Message Queuing Protocol
AMS	Alpha Magnetic Spectrometer
ANL	Argonne National Laboratory
ANL	Argonne National Laboratory
ANSE	Advanced Network Services for Experiments
AOD	analysis object data
ASCR	Office of Advanced Scientific Computing Research
ASGC	Academia Sinica Grid Computing
AST	NSF Astronomy
ATLAS	A Toroidal LHC Apparatus
AURA	Association of Universities for Research in Astronomy
BES	Beam Energy Scan
BGP	Border Gateway Protocol
BNL	Brookhaven National Laboratory
BOSS	Baryon Oscillation Spectroscopic Survey
BSM	Beyond the Standard Model
BUR	Beam User Request
CA	cooperative agreement
CCD	charge-coupled device
CCJ	Computing Center in Japan
CDMS	Cryogenic Dark Matter Search
CE	Compute Element
CEBAF	Continuous Electron Beam Accelerator Facility

CERN	European Organization for Nuclear Research
CLAS	CEBAF Large Acceptance Spectrometer
CMS	Compact Muon Solenoid
CP	charge parity
CPU	Central Processing Unit
CRAB	CMS Remote Analysis Builder
CVMFS	CERN Virtual Machine Filesystem
DAQ	data acquisition
DBI	Database Interface
DCS	data collection system; detector control system
DDM	distributed data management
DECam	Dark Energy Camera
DES	Dark Energy Survey
DESC	Dark Energy Science Collaboration
DESI	Dark Energy Spectroscopic Instrument
DOE	Department of Energy
DPD	Derived Physics Data
DST	data summary tape
ECS	ESnet Collaboration Services
EF	Energy Frontier
EIC	Electron-Ion Collider
ELIC	Electron Ion Collider
E-LITE	Eastern Lightwave Internetworking Technology Exchange
ESD	event summary data
ESnet	Energy Sciences Network
EVO	Enabling Virtual Organizations
EXO	Enriched Xenon Observatory
FAX	Federated ATLAS XrootD system
FDT	Fast Data Transfer
FEL	free-electron laser
Fermilab	Fermi National Accelerator Laboratory
FGST	Fermi Gamma Ray Space Telescope
FRIB	Facility for Rare Isotope Beams
FTD	File Transfer Daemon
FTP	File Transfer Protocol
GEANT	Gigabit European Advanced Network Technology
GFS	Global File System
GLORIAD	Global Ring Network for Advanced Applications Development
GO	Globus Online
GPN	Great Plains Network
GSFC	Goddard Space Flight Center
GSI	Grid Security Infrastructure
HEP	Office of High Energy Physics
HFT	Heavy Flavor Tracker

HI	heavy ion
HISQ	highly improved staggered quark
HLT	high-level trigger
HPS	Heavy Photon Search
HPSS	high-performance storage system
HPWREN	High Performance Wireless Research and Education Network
I/O	input/output
IB	InfiniBand
IF	Intensity Frontier
IHEP	Institute of High Energy Physics
ILC	International Linear Collider
ILDG	International Lattice Data Grid
INCITE	Innovative and Novel Computational Impact on Theory and Experiment
ILC	International Linear Collider
IaaS	Infrastructure As A Service
IP	Internet Protocol
IPAC	Infrared Processing and Analysis Center
JLAB	Thomas Jefferson National Accelerator Facility
JSA	Jefferson Science Associates
KISTI	Korean Institute of Science and Technology Information
KUP	keep up production
LAN	local area network
LANL	Los Alamos National Laboratory
LarTPC	liquid argon time projection chamber
LBNE	Long Baseline Neutrino Experiment
LBNL	Lawrence Berkeley National Laboratory
LC	Livermore Computing
LCF	Leadership Computing Facility
LCIO	Linear Collider I/O
LDAP	Lightweight Directory Access Protocol
LHC	Large Hadron Collider
LHCONE	LHC Open Network Environment
LITE	Lightwave Internetworking Technology Enterprise
LLNL	Lawrence Livermore National Laboratory
LNGS	National Laboratory of Gran Sasso
LOI	letter of intent
LQCD	Lattice QCD
LS	Long Shutdown
LSST	Large Synoptic Survey Telescope
MAN	metropolitan area network
MARIA	Mid-Atlantic Research Infrastructure Alliance
MATP	Mid Atlantic Terascale Partnership
MC	Monte Carlo
MIE	Major Item of Equipment

MINOS	Main Injector Neutrino Oscillation Search
M&O	maintenance and operation
MOU	Memorandum of Underestanding
MREFC	Major Research Equipment and Facilities Construction
MSS	mass storage system
MTD	Muon Telescope Detector
MWT2	Midwest Tier-2 Center
NASA	National Aeronautics and Space Administration
NCSA	National Center for Supercomputing Applications
NE	Network Element
NERSC	National Energy Research Scientific Computing Center
NEWT	NERSC Web Toolkit
NFS	network file system
NGC	North Galactic Cap
NLO	next-to-leading order
NOAO	National Optical Astronomy Observatory
NP	Office of Nuclear Physics
NSF	National Science Foundation
NSRL	NASA Space Radiation Laboratory
NuMI	Neutrinos at the Main Injector
O2	online/offline
ODM	Offline Data Monitor
ODU	Old Dominion University
OIM	OSG Information Management
OLCF	Oak Ridge Leadership Computing Facility
ONE	optical network environment
OPN	optical private network
ORNL	Oak Ridge National Laboratory
OSC	Ohio Supercomputer Center
OSCARS	On-Demand Secure Circuits and Advance Reservation System
OSG	Open Science Grid
PanDA	Production and Distribution Analysis
PBS	Portable Batch System
PDACS	Portal-based Data Analysis services for Cosmological Simulations
PDSF	Parallel Distributed Systems Facility
PD2P	PanDA Dynamic Data Placement
perfSONAR	PERformance Service Oriented Network monitoring Architecture
PI	principal investigator
PhEDEx	Physics Event Data Export
PHENIX	Pioneering High Energy Nuclear Interaction eXperiment
PID	Particle Identification Detector
PKI	public key infrastructure
PNNL	Pacific Northwest National Laboratory
pp	proton-proton

PQM	Physics Quality Monitoring
PRAC	Petascale Resource Computing Allocations
PROOF	Parallel ROOT Facility
PTF	Palomar Transient Factory
PWG	Physics Working Group
QCD	quantum chromodynamics
QGP	quark-gluon plasma
R&D	Research and development
RACF	RHIC/ATLAS Computing Facility
RAID	redundant array of independent disks
RCF	RHIC Computing Facility
RDO	raw data object
RHIC	Relativistic Heavy Ion Collider
SAC	STAR Analysis Center
SC	Office of Science
SCIPP	Santa Cruz Institute for Particle Physics
SCP	Secure Copy Program
SDN	software defined networking
SDN	Science Data Network
SDSC	San Diego Supercomputer Center
SE	Storage Element
SIP	Session Initiation Protocol
SLIC	Simulator for the Linear Collider
SM	Standard Model
SNS	Spallation Neutron Source
SoX	Southern Crossroads
SQL	Structured Query Language
SRF	superconducting radiofrequency
SRM	Storage resource manager
STAR	Solenoidal Tracker At RHIC
STEM	science, technology, engineering, and mathematics
SUMS	STARS Unified Meta Scheduler
SURA	Southeastern Universities Research Association
SURF	Sanford Underground Research Facility
SUSY	SuperSymmetry
UC	University of California
UIUC	University of Illinois at Urbana-Champaign
UNH	University of New Hampshire
USQCD	U.S. Quantum Chromodynamics
UTA	University of Texas at Arlington
VCR	Virtual Control Room
VM	virtual machine
VO	virtual organization
VOIP	voice over Internet Protocol

VOMRS	Virtual Organization Management and Registration Service
VOMS	Virtual Organization Membership Service
VORTEX	Virginia Optical Research Technology Exchange
VTX	Silicon Vertex Tracker
WAN	wide area network
WIPP	Waste Isolation Pilot Plant
WISE	Wide-Field Infrared Survey Explorer
WLCG	Worldwide LHC Computing Grid
WMS	workload management system
WSU	Wayne State University
XSEDE	Extreme Science and Engineering Discovery Environment
ZTF	Zwicky Transient Factory

28 Acknowledgements

This work would not have been possible without the contributions and participation of those who provided information and attended the review. ESnet would also like to thank the HEP and NP program offices for their help in organizing the review and providing insight into the facilities supported by the HEP and NP programs.

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 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 Offices of High Energy Physics and Nuclear Physics.

This is LBNL report LBNL-6642E