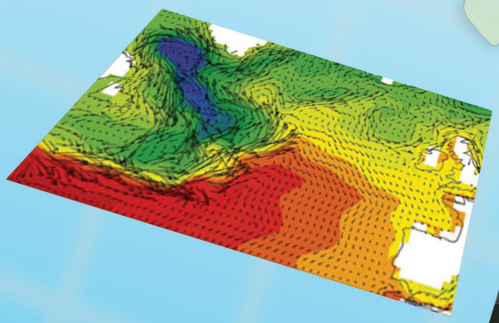
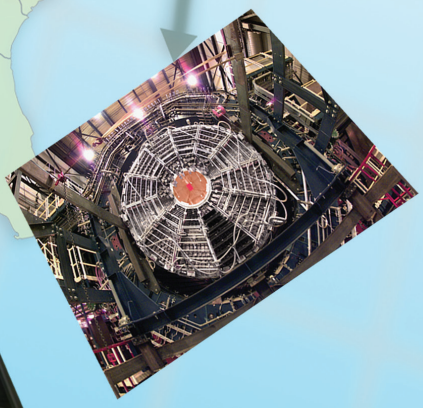
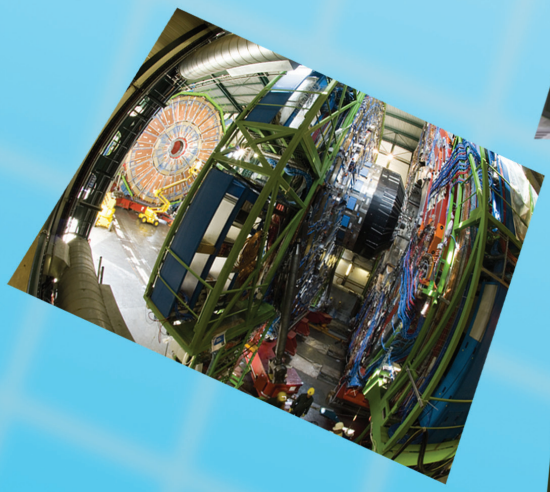
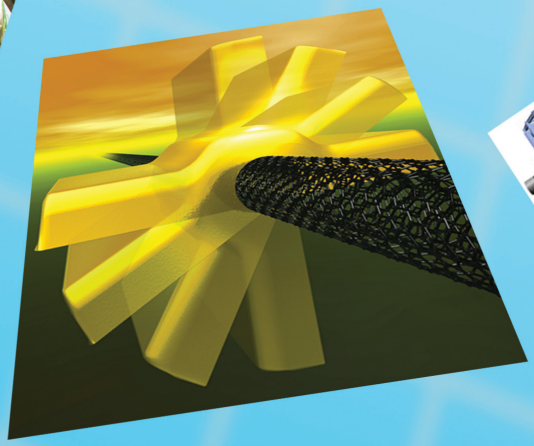
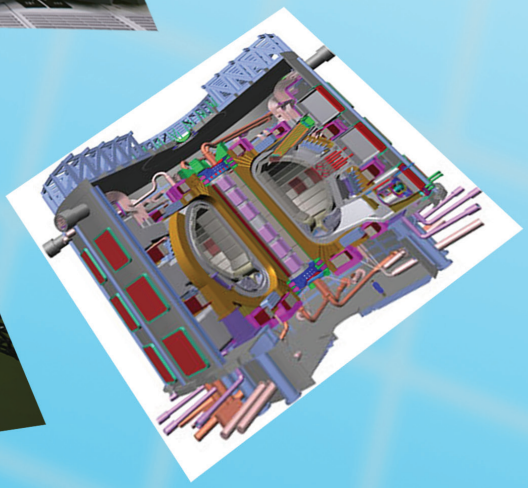
HEP Science Network Requirements

ESnet
Energy Sciences Network

Office of High Energy Physics
Network Requirements Workshop
Conducted August 27 and 28, 2009

Final Report



Berkeley Lab

U.S. DEPARTMENT OF ENERGY
Office of Science

HEP Network Requirements Workshop

Office of High Energy Physics, DOE Office of Science
Energy Sciences Network
Gaithersburg, MD — August 27 and 28, 2009

ESnet is funded by the US 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 Office of High Energy Physics.

This is LBNL report LBNL-3397E

Participants and Contributors

Jon Bakken, FNAL (LHC/CMS)

Artur Barczyk, Caltech (LHC/Networking)

Alan Blatecky, NSF (NSF Cyberinfrastructure)

Amber Boehnlein, DOE/SC/HEP (HEP Program Office)

Rich Carlson, Internet2 (Networking)

Sergei Chekanov, ANL (LHC/ATLAS)

Steve Cotter, ESnet (Networking)

Les Cottrell, SLAC (Networking)

Glen Crawford, DOE/SC/HEP (HEP Program Office)

Matt Crawford, FNAL (Networking/Storage)

Eli Dart, ESnet (Networking)

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

Michael Ernst, BNL (HEP/LHC/ATLAS)

Ian Fisk, FNAL (LHC/CMS)

Rob Gardner, University of Chicago (HEP/LHC/ATLAS)

Bill Johnston, ESnet (Networking)

Steve Kent, FNAL (Astroparticle)

Stephan Lammel, FNAL (FNAL Experiments and Facilities)

Stewart Loken, LBNL (HEP)

Joe Metzger, ESnet (Networking)

Richard Mount, SLAC (HEP)

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

Harvey Newman, Caltech (HEP/LHC/Networking)

Jennifer Schopf, NSF (NSF Cyberinfrastructure)

Yukiko Sekine, DOE/SC/ASCR (NERSC Program Manager)

Alan Stone, DOE/SC/HEP (HEP Program Office)

Brian Tierney, ESnet (Networking)

Craig Tull, LBNL (Daya Bay)

Jason Zurawski, Internet2 (Networking)

Editors

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

Table of Contents

1 Executive Summary

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

In August 2009 ESnet and the Office of High Energy Physics (HEP), of the DOE Office of Science, organized a workshop to characterize the networking requirements of the programs funded by HEP.

The International HEP community has been a leader in data intensive science from the beginning. HEP data sets have historically been the largest of all scientific data sets, and the communty of interest the most distributed. The HEP community was also the first to embrace Grid technologies.

The requirements identified at the workshop are summarized below, and described in more detail in the case studies and the Findings section:

- There will be more LHC Tier-3 sites than orginally thought, and likely more Tier-2 to Tier-2 traffic than was envisioned. It it not yet known what the impact of this will be on ESnet, but we will need to keep an eye on this traffic.
- The LHC Tier-1 sites (BNL and FNAL) predict the need for 40-50 Gbps of data movement capacity in 2-5 years, and 100-200 Gbps in 5-10 years for HEP program related traffic. Other key HEP sites include LHC Tier-2 and Tier-3 sites, many of which are located at universities. To support the LHC, ESnet must continue its collaborations with university and international networks.
- While in all cases the deployed "raw" network bandwidth must exceed the user requirements in order to meet the data transfer and reliability requirements, network engineering for trans-Atlantic connectivity is more complex than network engineering for intra-US connectivity. This is because transoceanic circuits have lower reliability and longer repair times when compared with land-based circuits. Therefore, trans-Atlantic connectivity requires greater deployed bandwidth and diversity to ensure reliability and service continuity of the user-level required data transfer rates.
- Trans-Atlantic traffic load and patterns must be monitored, and projections adjusted if necessary. There is currently a shutdown planned for the LHC in 2012 that may affect projections of trans-Atlantic bandwidth requirements.
- There is a significant need for network tuning and troubleshooting during the establishment of new LHC Tier-2 and Tier-3 facilities. ESnet will work with the HEP community to help new sites effectively use the network.
- SLAC is building the CCD camera for the LSST. This project will require significant bandwidth (up to 30Gbps) to NCSA over the next few years.

- The accelerator modeling program at SLAC could require the movement of 1PB simulation data sets from the Leadership Computing Facilities at Argonne and Oak Ridge to SLAC. The data sets would need to be moved overnight, and moving 1PB in eight hours requires more than 300Gbps of throughput. This requirement is dependent on the deployment of analysis capabilities at SLAC, and is about five years away.
- It is difficult to achieve high data transfer throughput to sites in China. Projects that need to transfer data in or out of China are encouraged to deploy test and measurement infrastructure (e.g. perfSONAR) and allow time for performance tuning.

2 Workshop Background and Structure

The strategic approach of the Office of Advanced Scientific Computing Research (ASCR – ESnet is funded by the ASCR Facilities Division) and ESnet for defining and accomplishing ESnet's mission involves three areas:

1. Work with the SC community to identify the networking implication of the instruments, supercomputers, and the evolving process of how science is done
2. Develop an approach to building a network environment that will enable the distributed aspects of SC science and then continuously reassess and update the approach as new requirements become clear
3. Keep anticipating future network capabilities that will meet future science requirements with an active program of R&D and Advanced Development

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

> A) Exploring the plans and processes of the major stakeholders, including the data characteristics of scientific instruments and facilities, regarding what data will be generated by instruments and supercomputers coming on-line over the next 5-10 years. Also by examining the future process of science: how and where will the new data be analyzed and used, and how the process of doing science will change over the next 5-10 years.
>
> B) Observing current and historical network traffic patterns and trying to determine how trends in network patterns predict future network needs.

The primary mechanism of accomplishing (A) is the Office of Science (SC) Network Requirements Workshops, which are sponsored by ASCR and organized by the SC Program Offices. SC conducts two requirements workshops per year, in a cycle that will repeat starting in 2010:

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

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

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

In August 2009 ESnet and the Office of High Energy Physics (HEP), of the DOE Office of Science, organized a workshop to characterize the networking requirements of the science programs funded by HEP.

Workshop participants were asked to codify their requirements in a case study format that included a network-centric narrative describing the science, the instruments and facilities

currently used or anticipated for future programs, the network services needed, and the way in which the network is used. Participants were asked to consider three time scales in their case studies — the near term (immediately and up to 12 months in the future), the medium term (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 time scale and columns for network bandwidth and services requirements. The case study documents are included in this report.

3 Office of High Energy Physics (HEP)

The mission of the High Energy Physics (HEP) program is to explore and to discover the laws of nature as they apply to the basic constituents of matter, and the forces between them. The core of the mission centers on investigations of elementary particles and their interactions, thereby underpinning and advancing Department of Energy missions and objectives through the development of key cutting-edge technologies and trained manpower that provide unique support to these missions.

4 Large Hadron Collider (LHC) Experiments

4.1 Background

All the Large Hadron Collider (LHC) experiments chose to utilize a distributed computing environment where the majority of the processing and storage resources are located away from CERN. The data distribution architecture for the LHC is organized in "tiers." CERN hosts the Tier-0 center, which is used for prompt reconstruction and the first archival copy of the data; there are 10 Tier-1 centers for ATLAS and 7 Tier-1 centers for CMS that reprocess and serve data as well as provide the second custodial copy; and each experiment has between 40 and 50 Tier-2 computing centers that provide analysis and simulated event generation resources.

In the United States, a critical component of the LHC data distribution from CERN to the Tier-1 centers (Brookhaven National Laboratory for ATLAS and Fermi National Accelerator Laboratory for CMS) is trans-Atlantic network connectivity. US LHCNet provides the trans-Atlantic connectivity between the US and CERN. US LHCNet and ESnet connect in multiple locations to ensure that there are no single points of failure in the Tier-0 to Tier-1 data delivery. ESnet receives the data from US LHCNet and delivers it to the Tier-1 centers at Fermilab and BNL.

In the LHC Community the Tier-1s have unique storage challenge in that they ingest data from CERN for custodial storage and from multiple Tier-2 centers to archive simulated data. The Tier-1s have reprocessing and skimming responsibilities in both CMS and ALTAS, and a significant contribution to the total simulated event production in ATLAS. In order to meet these obligations, the total processing at the Tier-1s is larger than the processing available at CERN for both experiments. In some sense, the Tier-1s have the most challenging network problem of all the LHC Tiers because they are expected to ingest data at predictable rates from multiple sources, synchronize reprocessed data to other Tier-1 centers in scheduled bursts, and send data samples to Tier-2 sites for analysis based on reasonably chaotic user requests.

The Tier-2 centers provide the majority of the analysis resources and capacity for simulated event production. Tier-2 centers are predominantly located at universities and have a high level of diversity in size, connectivity, and experience. The US based Tier-2s for both ATLAS and CMS, on the spectrum of Tier-2s, are large and reliable.

In February 2010, CERN released a revised LHC running schedule. In this schedule, collider operations will commence in spring 2010 at 7 TeV and end in 2011 with a planned integrated luminosity of 1 inverse femtobarn (1 fb^{-1}). In 2012, there will be a maintenance period of approximately one year to finish the repairs required to obtain design energy of 14 TeV. Collider operations will resume in 2013 through 2014/2015 with another extended maintenance period expected around 2015/2016. The changes to the running schedule and to the luminosity profile may impact the LHC networking requirements relative to those presented at the time of the workshop.

4.2 Key Local Science Drivers

4.2.1 Local Instruments and Facilities - CERN:

The CERN computing facility is located primarily in one building on the CERN Meyrin Site and connected to the LHC detectors over dedicated network links to the experimental halls. The raw data from the detectors is reformatted, calibrated and reconstructed using processing resources in IT, which increases the volume of the data. The samples are buffered on disk and archived to tape based storage using the CASTOR hierarchical mass storage, which was developed by CERN.

Processing resources in the LHC are measured in thousands of HepSpec 2006 (kHS06), which is based on SpecInt 2006. The CERN computing facility located in IT will provide 55 kHS06 for ATLAS and 48 kHS06 for CMS at the start of the run, and growing to 66 kHS06 and 102 kHS06 respectively in 2010. CERN will provide 3.5PB of disk for ATLAS and 1.9PB of disk for CMS at the start of the run, growing to 4.1PB and 4.2PB of disk storage in 2010. CERN also will provide archival storage for the two experiments of 5.1PB for ATLAS and 5.5PB for CMS at the start of the run, growing to 9.0PB and 10.4PB in 2010. These numbers are based on the projections made to the LHC Computing Scrutiny Group in June of 2009.

Under running conditions with collisions ATLAS will collect roughly 200Hz of 1.6MB events, while CMS collects 300Hz of data of 1.5MB events. Reconstructing the events increases the size by approximately one third for both experiments. The CERN computing infrastructure has been demonstrated to be able to archive data to tape at a total rate of about 4GB/s. The components driving the wide area networking have successfully exported 25Gbps of experiment data to the aggregated Tier-1s.

4.2.2 Local Instruments and Facilities – Tier-1:

In the US, Tier-1 computing is provided by Brookhaven National Laboratory (BNL) for ATLAS and Fermi National Accelerator Laboratory (FNAL) for CMS. Both facilities are large in absolute size and in relative size when compared to other Tier-1 computing centers for the LHC. The Tier-1s are connected to CERN and data from the LHC via an Optical Private Network (OPN).

By 2010 both BNL and FNAL will have more than 5PB of disk in production and 40kHS06 of processing. The two experiments have different estimates for the amount of tape needed at Tier-1s but both need multiple petabytes.

Both the BNL and FNAL Tier-1 facility utilize dCache to virtualize the large number of physical devices into a storage system. The LAN traffic between the disk servers and the worker nodes on average is between 1-2MB/sec per processor core, which corresponds to 5-10GB/sec (40Gbps-80Gbps) at FNAL and currently up to 5GB/sec at BNL.

4.2.3 Local Instruments and Facilities – Tier-2:

In the US there are 12 Tier-2 computing facilities: 5 for ATLAS and 7 for CMS. Four of the ATLAS Tier-2s are distributed facilities with hardware and operations support located at two, or three, campuses.

In early 2009 a nominal Tier-2 configuration was 4kHS06 of processing for CMS and 10 kHS06 of processing for ATLAS, which is roughly 500 and 1200 processor cores respectively, with 200TB (CMS) and 400 TB (ATLAS) of usable disk. The available storage space is spread over many physical storage devices and there are several technologies used to make them a coherent storage system. In the US dCache and XRootd are currently in production, with Hadoop currently in commissioning for CMS. Internationally, several Tier-2s also use DPM developed at CERN.

The CMS estimate for local area networking was 1-2MB/sec per processor core, or 500MB/sec -1GB/sec (4Gbps – 8Gbps) aggregated over the cluster. With 2-2.5MB/sec corresponding numbers for ATLAS analysis jobs running at Tier-2 centers are slightly higher. The technologies deployed in the US have been able to achieve these rates using the number of physical devices available.

4.2.4 Local Process of Science - CERN:

Scientific discovery at the LHC is a massive data-mining problem. The design rate of collisions at the experiment is 40MHz with a data collection rate of a few hundred Hz. Only a tiny fraction of the events contain interesting physics and a smaller fraction contain evidence of new discoveries. It is the job of the experiment's data acquisition systems to preferentially select the 1 in 10^5 events that can practically be kept at the T0 and T1 centers.

The job of the CERN computing facility is to archive the collected data and reconstruct the samples with a sufficiently good calibration that the experimenters can verify that they are actually selecting interesting events and that new discoveries are not being rejected by the triggering system. The data reconstruction and archiving activities lead to large but reasonably predictable usage of the CERN local area network of between 600MB/sec and 1000MB/sec per experiment. The data rate is defined by the experiment trigger rate, which then defines the processing and storage resources needed.

4.2.5 Local Process of Science – Tier-1:

The problem of event selections continues with the Tier-1s. The Tier-1s are responsible for updating the data samples by reprocessing with improved calibration, and for creating analysis samples for users. The events are skimmed to attempt to make smaller samples focusing on particular physics processes and thinned to concentrate objects relevant to a particular analysis. Each experiment will collect a few billion events per year and, except in the most fortuitous cases, a new discovery will be based on a few hundred, or less, of very carefully selected events.

4.2.6 Local Process of Science – Tier-2:

The Tier-2 centers were designed to support roughly 40 physicists performing analysis. The physicists make requests for data samples of selected events. These samples will vary from highly selected and summarized data of a few TB to much larger samples of more complex data formats that will result in samples up to tens of TB. The process of science is to analyze these samples calculating discriminating quantities that distinguish the process being investigated from other physics processes. Highly summarized data formats are often created and transferred back to user-controlled space. Data samples can be very dynamic as improved calibration and software is available. The LHC experiments expect to reprocess the complete data samples 3-6 times during the first year of running, which in turn requires an update of the user analysis samples.

4.3 Key Remote Science Drivers

4.3.1 Remote Instruments and Facilities - CERN:

CERN exports data over the wide area to the collection of Tier-1 computing centers. In the US, Tier-1 computing is provided by Brookhaven National Laboratory (BNL) for ATLAS and Fermi National Accelerator Laboratory (FNAL) for CMS. Both US Tier-1 centers are connected via ESnet to an Optical Private Network (OPN) to connect to CERN, the host laboratory for the Large Hadron Collider (LHC). The OPN connections across the Atlantic are managed by US LHCNet, and by October 2009 these connections will consist of 40Gbps of networking shared between the two experiments.

By 2010 ATLAS and CMS will have roughly 40 kHS06 of processing resources and approximately 5PB of disk storage at CERN. The two experiments have different estimates for the amount of remote tape needed but both need multiple petabytes.

The rate of wide area transfer to Tier-1s from CERN is predictable and stable within a range. The rate of data export from CERN is defined largely by the trigger rate of the experiment. A reasonably conservative provisioning factor is assigned to allow for prompt recovery from service loss on either the CERN or Tier-1 end. The rate during sustained data taking during collisions is 2.4Gbps to BNL for ATLAS and 4Gbps to FNAL for CMS. The peak rate is assumed to be a factor of 2 higher to provide for recovery from service failures and changing experiment conditions.

4.3.2 Remote Instruments and Facilities – Tier-1:

The Tier-1 centers receive and process the data samples that are used for most experiment analyses and subsequently have a large number of remote connections. The Tier-1 centers are responsible for archiving data from CERN as described above. Tier-1 centers in CMS receive a roughly equivalent sample averaged over a year from Tier-2 computing facilities to archive simulated event samples, in ATLAS simulated event data amounts about 20% of collision data. The estimate for the data input rate from CERN into FNAL is 2.2Gbps and 1.2Gbps into BNL for custodial data with 1.8Gbps of other samples during the first year, while from Tier-2s the rate is expected to be 0.5Gbps per experiment. While the total volume of data is similar, the Monte Carlo rate is averaged over accelerator downtimes and other periods without beam.

Tier-1s are also responsible for synchronizing reprocessed reconstruction and analysis data objects to other Tier-1s. The Tier-1 with the custodial data is responsible for reprocessing, but the reconstructed and analysis objects are served from multiple locations. The rate of Tier-1 to Tier-1 transfers is driven by how long the experiments are willing to wait to synchronize the reprocessed samples. At FNAL the export rate on average to Tier-1s is expected to be 3.3Gbps and an import rate of 1.2Gbps. At BNL the export rate on average to Tier-1s is expected to be 1.0Gbps and an import rate of 1.5Gbps.

The Tier-1s serve samples to Tier-2 centers for analysis. This rate is driven by user access needs and will have significant bursts. For CMS the goal is to be able to achieve 50MB/s (400Mbps) to the worst connected Tier-2 and 500MB/sec (4Gbps) to the best-connected Tier-2. This allows the replication of a 5TB to 50TB sample in one day. The aggregate peak rate from FNAL to Tier-2 centers for user analysis is expected to be 11Gbps.

For ATLAS the goal is to be able to achieve 100MB/sec on average and a peak rate of 400MB/sec to Tier-2 centers with 10Gbps connectivity. The peak rate from BNL to Tier-2 centers for user analysis is expected to be > 10Gbps.

One of the primary differences between the ATLAS and CMS computing models is the number of connections between Tier-1 and Tier-2 centers. In the ATLAS model connections between Tier-1 and Tier-2 centers are organized into clouds, where most connections to locations in the cloud and connections between clouds are handled at the Tier-1 level. In CMS data is served from the hosting site, which means a Tier-2 center can theoretically connect to any Tier-1. In practice for samples that exist in multiple places the most efficient source is typically the closest Tier-1.

The total import rate into the US Tier-1s is the combination of ingest rates from CERN and the Tier-2s plus synchronization data from other Tier-1s. The data coming from Tier-1s will only come in bursts, but the total for FNAL is 4Gbps for custodial data and potentially as high as 6Gbps when non-custodial samples are included. For BNL the total is expected to reach 11-15Gbps. The export rate also includes synchronization and serving data to Tier-2 users, but will likely reach peaks of 15Gbps for FNAL and >10Gbps for BNL.

4.3.3 Remote Instruments and Facilities – Tier-2:

Tier-2 centers are connected to Tier-1 facilities to receive updated reprocessed data and to serve as archival storage for the locally produced simulated events. The simulated event archiving is roughly steady throughout the year and at a predictable rate based on the number of resources used for simulation. In CMS and ATLAS each Tier-2 site sends data to a Tier-1 at between 50Mbps-100Mbps on average. The rate into Tier-2s is much less predictable and driven by user needs and availability of new data samples. In CMS and most of ATLAS (one Tier-2 site is missing) each Tier-2 site has access to a 10Gbps WAN link. The CMS computing model calls for the worst connected Tier-2 to be able to download data at 50MB/sec (400Mbps) and the best connected should be able to drive bursts to 500MB/sec (4Gbps). The same holds for the ATLAS sites Currently in CMS 6 of 7 Tier-2s have demonstrated a daily average of greater than 250MB/sec (2Gbps) and

one has demonstrated daily average of 4Gbps. In ATLAS daily rates of 100-200MB/sec to each of the Tier-2 centers are frequently observed as part of regular production activities and during exercises.

4.3.4 Remote Process of Science:

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-2s from the US Tier-1s.

The Tier-1 centers continue the process of data mining in the LHC experiments. Once the data is collected it is continually reprocessed. The events are skimmed to attempt to make smaller samples focusing on particular physics processes and thinned to concentrate objects relevant to a particular analysis. Each experiment will collect a few billion events per year and, except in the most fortuitous cases, a new discovery will be based on a few hundred, or less, of very carefully selected events.

The scientific process primarily resides at the remote Tier-2 centers, which are the bulk of the analysis resources for both ATLAS and CMS. Smaller event samples are processed comparing the expected signal from the predicted background. In this case the signal can be a source of new physics, or the standard model physics being investigated.

The Tier-2s make requests for data samples from Tier-1 sites. The disk space available at Tier-2s is large, but has to support analysis groups and user communities. The data will be frequently refreshed and the experiment will refine the selections. The Tier-2 disk space is expected to be treated like a dynamic cache.

4.4 Local Science Drivers – the next 2-5 years

4.4.1 Local Instruments and Facilities – CERN, next 2-5 years:

During the next 2-5 years the LHC will go from startup to operating at design luminosity. The complexity of events, the event processing times, and the 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 facility complexity. Processing and storage nodes will be replaced with faster nodes and larger nodes, though the number of nodes should remain roughly constant.

The LHC plans to operate during 2010 starting at 7TeV center of mass energy and increasing to 10TeV center of mass energy as they gain confidence in the machine performance. In 2011 a reasonably long machine shutdown is anticipated to complete the machine modifications needed to reach the design energy of 14TeV.

4.4.2 Local Instruments and Facilities – Tier-1, next 2-5 years:

The Tier-1 centers will maintain custodial copies of all the data and will be expected to periodically perform a reprocessing of all of the collected data. The original raw data is

generally stored on archival tape storage and will need to be staged for reprocessing. This is a process model common to HEP, though maintaining high CPU efficiency often requires careful operations.

4.4.3 Local Instruments and Facilities – Tier-2, next 2-5 years:

One area where 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-2s 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 2-3 fold increase in the required rate from local storage to accommodate the growing number of cores.

4.4.4 Local Process of Science, next 2-5 years:

The scientific process for the LHC will run in cycles over the next 2-5 years. At the start of the new energy frontier there is the opportunity for rapid discovery as thresholds for production are crossed. Some of these, like some Super Symmetry channels, turn on extremely fast and may, provided there is a good understanding of the detector and the background, lead to early scientific discoveries. As more data is analyzed the process of discovery turns to signals that occur less frequently and require analyzing larger quantities of data. The LHC will have at least 3 opportunities to cross energy frontiers: 7TeV, 10TeV, and 14TeV. This will require rapid assessment of the data looking for obvious new physics. As the volume of data increases there will be very careful and detailed analysis of large datasets looking for more subtle physics.

The scientific process employed at the Tier-2 centers in the out years will be similar to process used in the first year, but with larger data samples. Some analysis will search for new physics in the data from the current year, but many will seek to analyze the entire integrated sample and will access progressively larger samples.

4.5 Remote Science Drivers – the next 2-5 years

4.5.1 Remote Instruments and Facilities – CERN, next 2-5 years:

The primary difference from a networking perspective will be that the average rates observed in the out years will approach the peak rates observed in the first year. The live time of the experiments is expected to increase as the accelerator operations become more stable. The Tier-1 computing capacities will increase like the capacity increase expected at CERN with technology improvements.

4.5.2 Remote Instruments and Facilities – Tier1, next 2-5 years:

The Tier-1 centers will produce large samples when the whole collected data is reprocessed. These larger products will need to be synchronized to other Tier-1s. The samples selected by physics groups to be served to Tier-2s 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 bandwidth requirement for both Tier-1 to Tier-1 and

Tier-1 to Tier-2 traffic will increase. Compared to the first year of LHC operations an increase of the peak rate of 3 to 5 times is expected in 2-5 years. This is a rough estimate.

4.5.3 Remote Instruments and Facilities – Tier2, next 2-5 years:

As the LHC machine switches to higher luminosity the event complexity and size will increase. The simulated event samples will also increase in complexity and size to match the running conditions. The rate of data from Tier-2 to Tier-1 for archival purposes will at least double. The sizes of samples requested by the analysis community will increase as the integrated luminosity increases, though the total time desired to refresh samples for analysis is similar to year one. The target window of burst transfer rates will slide to at least 100MB/sec for the worst connected sites to 1GB/sec for the best connected sites.

4.5.4 Remote Process of Science, next 2-5 years:

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 Tier-1 centers will be performing similar actions as in the first year except with larger data samples as the integrated data collected grows. The data collected in a year will increase as the accelerator live time improves, but the Tier-1s will also be asked to reprocess previously collected data to provide consistent samples. More data will be recovered from archival storage in these cases.

The Tier-2 centers will be performing similar actions as in the first year except with larger data samples as the integrated data collected grows. The data collected in a year will increase as the accelerator live time improves.

The primary change for the process of science for remote networking will be the size of the transfers. The burst rates will increase to handle the larger samples.

4.6 Beyond 5 years – future needs and scientific direction

Looking beyond 5 years is firmly in the time of the Super LHC upgrades (SLHC). The SLHC is expected to have instantaneous luminosities of 10 times higher than the initial design. The trigger rates expected at the experiments will not be 10 times higher, but could increase by factors of 2-3 and the event size and complexity will increase dramatically. Significant improvements will be needed in the design of the software to handle the reprocessing. Improvements in the design of storage systems to handle the data volume and processing access are required.

4.7 Comparison of the CMS and ATLAS computing models

The two LHC experiments discussed here, CMS and ATLAS, have different computing models with different implications for the networks that support them. ATLAS has a structured model wherein Tier-2 centers download data from their local Tier-1, and send their results back to that Tier-1. In the US, the ATLAS Tier-1 at BNL will receive a full copy of the ATLAS data set from CERN. This means that ATLAS Tier-2s in the US are not considered likely to add significantly to the load on trans-Atlantic network circuits.

In contrast, CMS has a much more dynamic computing model, wherein a CMS Tier-2 is considered likely to fetch data from any CMS Tier-1. Also, in contrast to ATLAS, the CMS Tier-1 at FNAL is not expected to have a complete copy of the CMS data set from CERN. This means that CMS Tier-2 traffic will add (perhaps significantly) to the load on trans-Atlantic network circuits.

4.8 Ad-Hoc and Opportunistic Analysis

In addition (and in complement) to the structured analysis of LHC data by the Tier-1 and Tier-2 centers, ad-hoc analysis, driven by scientific inquiry and determined by aspects of the LHC data, will occur. It is difficult to quantify this exactly, because there is currently no LHC data to analyze. However, an attempt was made at the workshop to get a rough idea of the scope of this analysis and its potential impact on the network.

It was estimated by workshop participants that the opportunistic analysis of LHC data would comprise several different science-driven efforts, and that each of these might be equal in volume to 10% of the data set of a typical LHC Tier-2 center. In aggregate, the opportunistic analysis might consume an amount of data equal to the Tier-2 centers' structured data volumes.

It was also estimated that this data volume might increase by a factor of six over the next five years.

4.9 Impact of extended transatlantic circuit outages

The LHC relies heavily on a set of trans-Atlantic circuits to provide data connectivity between US and European sites – the circuits provide the means by which the data from CERN are brought to the US, and the means for exchanging results between US and European sites. The CMS and ATLAS experiments have disk buffers to allow for recovering from network outages, however these buffers have a limited capacity which governs how long a network outage can be tolerated.

An analysis of the ATLAS disk buffer capacity shows that the Tier-0 disk buffer is designed to hold at least two days of data during a network outage which can then be transferred in addition to current data when the network has recovered. If the outage were to last longer then the data would be transferred from the Tier-0 to Tier-1 centers in France and Germany. For traffic between European Tier-1 centers and the US Tier-1 at BNL, the Tier-1 disk buffers are designed for two to three days, though this is less of an issue because the data are kept on disk by the Tier-1 centers for other purposes anyway. Similar logic holds for the data output of the ATLAS Tier-1 at BNL – the BNL Tier-1 maintains the data, and so the notion of disk buffer overflow is not critical.

However, in the event of a week or multi-week trans-Atlantic outage, there could be significant negative impact to the experiment if all circuits were cut. This argues strongly for highly diverse trans-Atlantic circuits. This diversity exists currently, and it will be important to maintain that diversity for the foreseeable future. Since a dramatic reduction in bandwidth can cause problems similar to a complete outage, the diversity of connectivity should provide not just for basic connectivity, but for adequate service continuity during partial outages (a reduced number of available paths).

As far as recovery after the outage is concerned, in relevant exercises Tier-1 sites have shown that they can ingest traffic at rates up to, and some of them (typically the smaller Tier-1s) well above, 5 times the nominal rate - at least to disk, to tape is a matter of concurrent read operations competing for tape drives which may result in longer recovery times. The sending rate was observed to be at least 3 times the nominal rate. It was demonstrated that sites can presently recover from a 2 day backlog in 1 day or less. This might take significantly longer if the center supported both CMS and ATLAS and the links to both Fermilab and BNL were affected.

4.10 Network bandwidth impacts due to congestion, maintenance or development

In considering network requirements it is important to distinguish between the desired throughput between two points in the network and the capacity and character of the network to be installed. In addition to bandwidth to support the end-to-end throughput levels between various pairs of points in the network as required by users and site operators, installed link capacity must include non-payload bandwidth for the headroom needed to accommodate protocol overhead and congestion handling.

Installed bandwidth on diverse links must be sufficient to ensure a specified level of availability of the network, including during maintenance and unplanned downtimes. Further, some additional bandwidth is needed for operational/management/development functions. The network topology, beyond point to point links, has to include physically diverse paths and cross links in order to provide high availability end to end, including when one or more links go down, or are taken down for maintenance or short-duration tests, in which case automated fallback using an alternative end-to-end path in order to sustain non-stop operation of the network.

These sorts of engineering parameters that are derived from user bandwidth and reliability requirements are incorporated into the design and deployment of all modern R&E networks.

4.11 Summary Table – LHC Near Term (0-2 years)

Feature	Key Science Drivers		Anticipated Network Requirements	
LHC Aspect	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Tier-0 CERN and Trans-Atlantic	• Startup of the LHC • CERN to Tier-1 data transfers	• Data Mining for early discovery at the energy frontier	• Local area networking at CERN requires 800MB/sec to 1000MB/sec from local data storage per experiment for prompt reconstruction and calibration	• Wide area transfers of 2.4+4Gbps from CERN to the US on average with peak rates of 2*(2.4+4)Gbps
Tier-1	• Tier-1 data ingest from CERN • Tier-1 data reprocessing • Tier-2 data serving	• Archival storage for data from CERN and Tier-2 centers • Reprocess data with improved calibration for finer selection • Data serving for Tier-2s for detailed analysis	• 3Gbps to tape (CMS) • 1.2Gbs to tape (ATLAS) • 40Gbps to 80Gbps from disk for FNAL • 30Gbps to 50Gbps from disk for BNL • 11Gbps from disk per experiment toward the Tier-2 centers	• 2.2Gbps to 4Gbps from CERN and 0.5Gbps from local Tier-2s per experiment • 1.2Gbps to FNAL and 3.2Gbps to remote Tier-1s • 1.5Gbps to BNL and 1Gbps to remote Tier-1s • 11Gbps per experiment toward Tier-2s
Tier-2	• Tier-2 data export • Tier-2 data import	• Export of simulated event production to Tier-1 centers • Refresh samples for analysis by users	• 50Mbps to 100Mbps to disk per Tier-2 • 4Gbps to 12Gbps from disk to worker nodes	• 50Mbps to 100Mbps to archival Tier-1 sites • 400Mbps to worst connected Tier-1, 4Gbps to best connected Tier-1 for data transfers

4.12 Summary Table – LHC Medium Term (2-5 years)

Feature	Key Science Drivers		Anticipated Network Requirements	
LHC Aspect	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Tier-0 CERN and Trans-Atlantic	• LHC at design luminosity • CERN to Tier-1 data transfers	• Data Mining for low probability discovery physics at the design energy of the LHC	• Local area networking at CERN requires 800MB/sec to 1000MB/sec from local data storage per experiment for prompt reconstruction and calibration	• Wide area transfer rates of 2*(2.4+4)Gbps on average from CERN to the US with peak rates roughly 2 times higher
Tier-1	• Tier-1 data ingest from CERN • Tier-1 data reprocessing • Tier-2 data serving	• Archival storage for data from CERN and Tier-2 centers • Reprocess data with improved calibration for finer selection • Data serving for Tier-2s for detailed analysis	• 7Gbps to tape (CMS) • 3Gbps to tape (ATLAS) • 80Gbps to 120Gbps from disk per experiment • 30Gbps from disk per experiment	• 6Gbps from CERN and 1Gbps from local Tier-2 centers • 5Gbps to remote Tier-1 centers and 2Gbps to FNAL • 2Gbps to remote Tier1 centers and 3Gbps to BNL • 30Gbps to Tier-2 centers per experiment
Tier-2	• Tier-2 data export • Tier-2 data import	• Export of simulated event production to Tier-1 centers • Refresh samples for analysis by users	• 100Mbps to 200Mbps to disk • 8Gbps to 16Gbps from disk to worker nodes	• 100Mbps to 200Mbps to archival Tier-1 sites • 800Mbps to worst connected Tier-1, 8Gbps to best connected Tier-1 for data transfers

4.13 Summary Table – LHC Long Term (5+ years)

Feature	Key Science Drivers		Anticipated Network Requirements	
LHC Aspect	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Tier-0 CERN and Trans-Atlantic	• Super LHC Operations	• Processing of extremely complex high luminosity events.	• Local area networking of 3GB/sec to 5GB/sec per experiment on average - data transfer from local data storage for prompt reconstruction and calibration.	• Wide area transfer rate from CERN of 60Gbps to the US on average.
Tier-1	• Super LHC Operations	• Archival Storage from CERN and Tier-2 Centers	• Local area networking of 10Gbps to 15Gbps to tape	• Wide area transfer rate from CERN of 30Gbps to the US on average per experiment
Tier-2	• Tier-2 data export • Tier-2 data import	• Export of simulated event production to Tier-1 centers • Refresh samples for analysis by users	• 500Mbps to disk • Large data transfers to parallel applications	• 500Mbps to archival Tier-1 sites • 100TB samples updated routinely by local analysis users

5 LHC Data Movement

5.1 Background

The ability to move large data sets is crucial to many areas of 21st century science, but in no field is it more fundamental than in High Energy Physics. The basic blocks of LHC data to users are measured in terabytes, and a month's product, in petabytes. Datasets from future machines should be expected to be as much larger than these, as LHC datasets are than those from the preceding machines.

Doing all the processing of a major experiments' data in a single location is a politico-economic impossibility, verging, with just a bit of exaggeration, on a physical impossibility due to the energy requirements of that much computing power. Data sets must be moved to, from, and among numerous storage and analysis centers at an aggregate rate several times greater than the rate at which the raw data is produced.

The raw capacity of networks themselves is growing extremely rapidly. The only response to "How much bandwidth can my facility get?" is "How much can you afford?" But the ability of applications and cost-effective end systems to move the data is not growing as fast.

The tolerable clock time for completing the move of a data set varies for different workflows, but human factors limit it to no more than the order of one day, and for some purposes significantly less. Complex and sometimes brittle layered data management systems move data on schedule or on demand, but are prone to unpredictable performance or outright failures.

Distant storage and computing facilities are linked over the open Internet, and those facilities themselves are only as secure as they can afford to be. Authentication and authorization systems that mitigate these exposures are complex beyond the full understanding of the general user, and perhaps beyond the general users' tolerance level.

5.2 Key Local Science Drivers

5.2.1 Instruments and Facilities:

State of the art scientific storage systems hold data on spinning media, with redundant copies, a tape archive back-end, or both. The data may be striped across disks or even across servers for faster access. Servers face local clients over the local area network (LAN) through Ethernet or, sometimes, InfiniBand. In HEP computing, where the I/O rates of a single computational job are usually modest, Ethernet predominates as the network fabric. Striping across disk spindles is common, and striping files across servers is sometimes found.

Storage clusters with hundreds to a thousand servers have one-gigabit (1 Gbps) interfaces, or a few bonded gigabit interfaces per server. These are well-matched to the speed of a current RAID array. More expensive servers have ten-gigabit (10 Gbps) network interfaces and faster aggregate I/O capacity, but clusters built on such technology tend to comprise tens rather than hundreds of servers.

Computational clusters are located in the same or different buildings, depending on the scale of the installation and the age and capacities of the buildings and the electrical and cooling systems housing the machines. At the largest sites, the power needs of compute elements have exceeded the capacity of 20th century buildings, so the compute elements themselves sprawl over multiple buildings. Optical fiber connections among buildings are relatively cheap if the operating institution has control of all the land to be traversed. But if the site is in an urban setting, as many universities are, acquiring fiber right-of-way between buildings may be costly.

5.2.2 Process of Science:

Local data flows run between storage systems and computational clusters. The analysis jobs typically have very low Amdahl numbers (bits of I/O per instruction executed) and so the transfer capacities of local storage clusters has been able to keep pace, although not without considerable investment in capacity and throughput. The hourly average data dates between CMS storage and compute elements at Fermilab has been as high as 12 GB/sec (100 Gbps). The network capacity to support this today involves aggregating large numbers of 10 Gbps network switch interfaces and dedicating tens of pairs of fibers between buildings.

5.3 Key Remote Science Drivers

5.3.1 Instruments and Facilities:

Remote transfer of data is generally supported by the same storage systems that give local access to the data. At some sites, a local or global shortage of IP addresses causes storage nodes to be given private-use (RFC 1918) addresses and WAN transfers to or from them are funneled through a small number of application gateway hosts. Application gateways to storage systems without routable IP addresses introduce their own set of failures. The alternative, Network Address Translation (NAT) is generally incompatible with end-to-end security.

5.3.2 Process of Science:

HEP sites without custodial responsibility for data—in LHC terms, the Tier-2 and Tier-3 sites—tend to rely on another site to replace lost data, rather than keeping redundant copies or operating tertiary storage systems. (Use of redundant copies may be reserved only for data generated locally, until those data are delivered to a custodial site.) In the US, a single 10 Gbps wide area network (WAN) link is the norm, and this is sometimes shared with the rest of the university. Dedicated or on-demand 10 Gbps for HEP use have proven themselves very useful where they are available. For example, a 50 TB data set was replaced at a USCMS Tier-2 site in 32 hours from the Tier-1 site over an on-demand network circuit crossing ESnet and Internet2. As pleasing as this was to the Tier-2, it represented less than 50% efficiency in use of that circuit.

Reasons for low transfer efficiency include TCP's standard behavior of throttling up the transmission rate until congestion occurs, then cutting the rate in half, as well as contention for resources in the host. Foremost among the latter is disk contention. The data rates of even the fastest disk subsystems plummet drastically when two or more

large files are read from different areas of the disk. And only during technology demonstrations do concurrent file transfers between sites A and B neatly pair disk A_1 with disk B_1, A_2 with B_2, and so on.

Recent research has also pinpointed and, in some cases, found solutions for, throughput bottlenecks in end systems themselves. Process scheduling, interrupt dispatching, and bad CPU core and cache assignments all can prevent a host from making maximal use of a high-speed network.

Concurrent parallel TCP streams for each file are used to smooth out the TCP speed "sawtooth" behavior over the wide area. This is a technique whose day may be passing soon, as various single-stream problems are overcome. Certainly when many file transfers are proceeding in parallel, some to or from the same hosts, it is redundant to create more parallelism at the TCP level. Between two storage systems with file striping, parallel streams might be a natural fit. However, with sites independently building and managing their facilities, it will so rarely be possible to match the striping at both endpoints that such a capability may never be implemented.

5.4 Local Science Drivers – the next 2-5 years

5.4.1 Instruments and Facilities:

Worldwide storage and processing of HEP data will remain the norm for the foreseeable future. The speeds of CPU cores have hit a wall and will not grow much further until physically new fabrication methods or architectures are found. Massively multicore systems will be the norm in the next two to five years. Unless HEP analysis changes in ways that involve more instructions per bit of data (still lower Amdahl numbers), the pressure will be on the system bus, the I/O subsystem, and the network interface to keep all the cores busy.

Speeds of facility core network and WAN devices continue to grow at a satisfactory pace. The challenges of the next half-decade will be to exploit them fully to keep CPUs hard at work.

RAID arrays are facing problems of concurrent multiple failures, as disks get larger and the time to rebuild a set after a single failure increases. Some other configuration of disk storage will be needed, or even "volatile" storage systems will need redundant copies of files in order to achieve satisfactory uptime.

5.4.2 Process of Science:

Some parts of the HEP analysis workflow are amenable to distributed processing in the MAP/REDUCE model. If that mode of processing is useful when applied to a subset of the complete dataset, such as is found at a single site, then some hybrid compute+storage nodes will be part of the facilities' offerings.

5.5 Remote Science Drivers – the next 2-5 years

5.5.1 Instruments and Facilities:

To alleviate disk contention during simultaneous file transfers, file-sized (multi-gigabyte) solid-state caches may be used between disk and network. This may be part of main memory or separate solid-state drives attached to the system. (Since the persistence of flash memory is not needed for this purpose, and it adds cost and has a short useful life when updated often, flash drives are not a good choice.)

By the time application gateways to storage clusters on private networks become unbearable, the addressing problem must be solved. (Fermilab has already had requests from some sites to makes its storage available on the nascent IPv6 Research and Education backbone networks.)

5.5.2 Process of Science:

Whether or not the disk contention issue is solved, scientists will seek to get maximal throughput from network links. On-demand virtual circuits can be an effective part of the solution so long as any TCP sessions sending rate never exceeds the allocated bandwidth. Hosts can perform traffic shaping to avoid TCP's "give 'til it hurts" behavior, *if* the hosts know how much bandwidth has been allocated to each flow.

If, on the other hand, congestion feedback remains the active control for data flow, then all active network devices in data paths must have packet buffer space of the order of the bandwidth-delay product of the network paths they are part of. That much buffer memory in a large number of switches and routers will be expensive. Barring some architectural breakthrough, extraordinary means would be needed to reduce the number of devices on long paths to make them affordable.

5.6 Beyond 5 years – future needs and scientific direction

At five years from the present, or soon after, the local area network trunks in large facilities will be in the terabit or multi-terabit per second range through aggregation. Congestion feedback may become completely unacceptable as a means to regulate high-rate data flows.

5.7 Summary Table – LHC Data Movement

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• Hundreds of 1-2Gbps and/or tens of 10Gbps servers. • Application gateways or NAT at some sites.	• Low I/O to instruction ratio. • Volatile storage at non-custodial sites. • 50% network utilization is considered very good.	• Large sites near the limits of bonded 10Gbps links.	• University-scale facilities need dedicated or on-demand 10Gbps.
2-5 years	• Move beyond RAID. • Multicore challenges bus and network. • Dedicated host per transfer or solid-state file buffers.	• New processing models for some workflows?	• Small sites approach 100Gbps trunks. • Large sites approach 1Tbps trunks.	• IPv6 needed to avoid NAT or application gateways. • Match applications' offered network load to reserved circuit bandwidth.
5+ years	•	•	• Terabit local networks.	• Congestion prevention by making the end systems aware of network path characteristics.

6 Trans-Atlantic Networking

The close collaboration between HEP sites in the U.S. and in Europe means that special consideration must be given to trans-Atlantic network connectivity and bandwidth capacity. The LHC will be a heavy user of trans-Atlantic bandwidth over the next few years, but several other HEP projects rely on trans-Atlantic networking as well, including BaBar and the Tevatron, and also including upcoming collaborations such as Super-B. The LHC use of the network has some fairly stringent reliability requirements in order to ensure full US participation in the experiments.

In addition to the issues addressed in section 4.10 "Network bandwidth impacts due to congestion, maintenance or development" on the relationship between required and deployed bandwidth, trans-Atlantic circuits are typically less reliable than land-based circuits, and the time to repair outages can be significantly longer (days to weeks). Therefore, it is critical to maintain sufficient diversity and capacity in trans-Atlantic circuits (e.g. by procuring connectivity from multiple vendors and/or ensuring that different circuits traverse different undersea cables and come onshore at different locations) to meet the required data transfer rates and reliability.

The need for diversity and deployed bandwidth in excess of the user-level requirements is driven by user reliability and service continuity requirements, and is not something that typically enters into the calculations of the network user community. However, this is something that must be considered by the engineering teams that build the network to provide service to the user community. Given the impact of extended outages (disk buffers for the experiments are measured in single-digit days while outages on trans-Atlantic circuits can last several weeks), provisioning additional bandwidth on diverse paths is necessary in order to protect against the potential impact of extended outages.

Table 1, below, shows an estimate of trans-Atlantic bandwidth needs for the near term, while table 2 (next page) provides more detail in a case study summary table. The bandwidth figures below are derived from the needs for bandwidth, redundancy, and protocol overhead in light of the increased likelihood of outages on trans-Atlantic circuits. While these numbers may in aggregate appear higher than those described in other case studies, they represent the addition of engineering considerations to ensure the availability of trans-Atlantic network service adequate for the needs of the science – even in the face of multiple link failures. In addition, protocol overhead and the effects of congestion are accounted for.

Table 1: Deployed Transatlantic Network Bandwidth Estimates and Provisioning Plan to meet User-Level Requirements (in Gbps) (From the T0/T1 networking group. Source: Harvey Newman, Caltech.)

Year	2006	2007	2008-9	2009-10	2011
CERN-BNL (ATLAS)	5	15	20	30	40
CERN-FNAL (CMS)	15	20	20	30	40
Other (ALICE, LHCb, LARP, Tier1-Tier2, Development, Inter-Regional Traffic, etc.)	10	10	10-15	20	20-30
TOTAL US-CERN BW	30	45	50-55	80	100-110
US LHCNet Bandwidth	20	30	40	60	80
Other bandwidth (GEANT2, IRNC, SURFnet, WHREN, CANARIE, etc.)	10	10	10-20	20	20-30

Table 2: Short, mid and long term bandwidth projections. The entries in the table cover the anticipated needs of LHC Tier0 and Tier1 operations, and a share of the Tier2 needs. (Source: Harvey Newman, Caltech)

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	Science Instruments and Facilities	Process of Science	Local Area Network Bandwidth and Services	Wide Area Network Bandwidth and Services
Near-term (0-2 years)	• LHC	• Detector data processing • Simulation data processing • Real-time control (remote control center) • Remote collaboration	• Each Tier1: 30 Gbps to/from storage (disk and tape) • Traffic isolation from General Purpose Network	• 80 Gbps transatlantic bandwidth • Traffic isolation from General Purpose Network • Guaranteed bandwidth, dynamic allocation
2-5 years	• LHC	• Detector data processing • Simulation data processing • Real-time control (remote control center) • Remote collaboration	• Each Tier1: 100 Gbps to/from storage (disk and tape) • Traffic isolation from General Purpose Network	• 280 Gbps transatlantic bandwidth • Traffic isolation from General Purpose Network • Guaranteed bandwidth, dynamic allocation
5+ years	• LHC	• Detector data processing • Simulation data processing • Real-time control (remote control center) • Remote collaboration	• Each Tier1: 120+ Gbps to/from storage (disk and tape) • Traffic isolation from General Purpose Network	• 400+ Gbps transatlantic bandwidth • Traffic isolation from General Purpose Network • Guaranteed bandwidth, dynamic allocation

7 ATLAS Tier-3 analysis center at ANL

7.1 Background

The ATLAS analysis center (ANL) was built to support ATLAS physics analyses, in particular for ATLAS physicists at US mid-west Institutes. The center is one of the three Analysis Support Centers (ACS) in the US. The center offers ATLAS users: (1) A model Tier-3 (T3g) for ATLAS analysis; (2) Meeting and office space for visitors; (3) Computer accounts; (4) analysis and software expertise and consultation; (5) T3g setup expertise and consultation.

The ANL ASC is operated by the ANL ATLAS group of the HEP division (ANL).

ASTRO group performs a simulation of a supernova using a dedicated cluster, but their requirements are not as high as for the ATLAS group.

7.2 Key Local Science Drivers

7.2.1 Instruments and Facilities:

Overall file storage is 20 TB. There are three clusters with 40 (ASTRO), 50 (ATLAS), 24 (HEP) CPU cores, respectively. 6 TB is allocated to distributed file storage (2TB per Linux box). Data uploads are done using the “dq2-get” tool. In case of the distributed file storage, data are copied using multiple threads of “dq2-get” (based on the arcond package). The dq2-get comes with the OSG-client tool installed for the ATLAS cluster.

All our Linux computers are based on Scientific Linux 5.3 (default kernel 2.6.18) and 4.7 (default kernel 2.6.9). The computers are connected by 1 Gbps Netgear switches. The uplink is 2 Gbps (fibers). All Linux servers have 1 Gbps network cards.

7.2.2 Process of Science:

The main workflow for the ASC is to run over data files. Users submit jobs to the grid (Tier1/2) where they skim data (ATLAS AOD format) or create ROOT N-tuples. The data files are copied from Tier1/2s to the local cluster. File downloads are performed by random users at random time (depends on many factors).

At present, we are working with Monte Carlo files (90% of all downloaded files). Files are typically downloaded from BNL (Tier1) or Univ. of Chicago (Tier2).

7.3 Key Remote Science Drivers

7.3.1 Instruments and Facilities:

Data from the ATLAS experiment will be delivered to Tier-1 and Tier-2 sites. They will then be skimmed using the pAthena tool and copied to ANL using "dq2-get".

The size of each downloaded file is 50-200 MB.

Data will be downloaded from BNL (Tier-1) or other ATLAS Tier-2 sites.

7.3.2 Process of Science:

Data will be processed at BNL or at a Tier-2 site (the closest is at University of Chicago). Then data will be copied to ANL for data analysis using a local cluster. Data are copied from the grid sites after selecting of events of interest.

7.4 Local Science Drivers – the next 2-5 years

7.4.1 Instruments and Facilities:

At any given moment, ASC will store 20-50 TB of data. Most of the data will be redistributed between many computers ("pre-staged") to be ready for data analysis.

70% of the data will be fully reconstructed Monte Carlo files which will be reloaded from Tier-1 and Tier-2 sites approximately every 6 months.

The center will use pAthena to submit jobs to the grid and dq2-get to download the results. dq2-get will be used with multiple threads (for each PC farm box in parallel).

It is expected that the center will install a file server with about 100 TB of storage space, and use SRM for data subscription from Tier-1 and Tier-2 sites. It is likely that other Tier-3 sites will deploy similar capabilities.

All the ASC Linux computers are based on Scientific Linux 5.3 (default kernel 2.6.18) and 4.7 (default kernel 2.6.9). The computers will be connected by Netgear switches with 10 Gbps uplink. All Linux servers will have 1 Gbps network cards.

7.4.2 Process of Science:

The science process consists of analysis of ATLAS data from the LHC – copying data from Tier-1/Tier-2 and processing the data using a local cluster. Data are copied from the grid sites after selection of events of interest. Data will be downloaded chaotically at random times by random users (depends on the ATLAS reprocessing schedule and many other factors). The size of each downloaded file can be up to 1 GB.

7.5 Remote Science Drivers – the next 2-5 years

7.5.1 Instruments and Facilities:

ATLAS Tier-1 and Tier-2 sites.

7.5.2 Process of Science:

Data from the ATLAS experiment will be delivered to Tier-1 and Tier-2 sites.

7.6 Beyond 5 years – future needs and scientific direction

It is likely that the ASC will be converted into an "analysis facility", and so will be able to export data (not only import data from Tier-1 and Tier-2 sites). The ASC should be able to handle ~100-200 TB downloads and uploads. Expected download times should be within several days. Thus, 10 Gbps connectivity to remote Tier-1 and Tier-2 sites will be necessary.

7.7 Outstanding Issues:

Downloading 100 TB of data annually is currently possible only from the University of Chicago Tier-2 (Midwestern Tier-2). The data rate is ~400 Mbps, and it takes 20 days for downloads (assuming 4.5 TB/per day). The data transfer is practical for other Tier-2s and the Tier-1 at BNL. At present, ASC gets ~100-150 Mbps of throughput to BNL, SLAC and other remote sites (for a single thread), although ASC is connected to the ANL campus network at 1 Gbps, and all switches and host network interfaces are 1 Gbps.

One network performance issue has been identified (with the help from ESnet and Eli Dart): TCP has to be tuned on Linux boxes used for downloads. Using custom kernels for Scientific Linux 4.7 (default 2.6.9) is not practical (it is centrally supported by CERN). But we are moving towards Scientific Linux 5.3 (kernel 2.6.18).

It seems the main problem now is the10 Gbps to 1 Gbps speed transition in the network switches (due to output queue overrun on the switches). It is important to be able to make full use of the current 1 Gbps network connection for data transfers from BNL and Tier-2 sites. At the moment, ASC can achieve 1 Gbps data transfer throughput only to hosts inside the ANL laboratory.

The next question is how to start using a 10 Gbps network connection (ASC is presently working on this). What hardware should be used to achieve 10 Gbps download speeds assuming modest budget (tens of thousands of dollars)? This is a common problem for the vast majority of Tier-3 sites.

It would be good to have a tool that could download files in parallel on multiple Linux boxes. One such tool has been developed at ANL (arcond package) but needs to be tested.

7.8 Summary Table:

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• Download data to ANL from Tier-1 and Tier-2 sites • Use dq2-get for multi-threaded downloads • 20TB of local storage	• Processing of downloaded data sets • Keep up to 20TB of data locally	• 1Gbps	• 2Gbps
2-5 years	• 50TB of local storage • 70% of local data will be fully reconstructed Monte Carlo files which will be refreshed every 6 months	• Same as above, but with 50TB of local capacity	• 1Gbps	• 10Gbps
5+ years	• Same, plus possible data export (analysis facility)	• Downloads of 100-200TB data sets for processing at ANL • Export of data to other sites	• 10Gbps	• 10Gbps

8 Tevatron Experiments

8.1 Background

The Tevatron is a proton-antiproton collider at Fermilab west of Chicago. The collider has been in operation since the ‘80s and has two general-purpose detectors, CDF and D0, that record the interactions, producing data at about 2 TB/sec each. Online filtering reduces the data volume to between 0.5 to 1.2 PB/year. After reconstruction about 500 TB/year are used for scientific analysis by each experiment. Another 500 TB/year of data is generated with Monte Carlo programs at Grid sites in America, Europe and Asia.

Each detector is operated by a collaboration of 60 to 90 institutions with a total of about 600 scientists from around the world. The experiments are in their final stage of data collection.

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

8.2.1 Instruments and Facilities:

Data are produced by each detector and sent to the Fermilab computer center on site for tape storage and processing. Each experiment has dedicated multi-gigabit-per-second network links to the compute center for this.

Grid farms at Fermilab are employed for Monte Carlo generation. Inter-switch connectivity is at 10 Gbps, sized for data analysis. (Monte Carlo generation requires less network bandwidth than data analysis.)

Grid farms of each experiment process the detector data and store physics data. The detector data are fetched from tape to a disk-based caching system and from there transferred to the worker node. Tape subsystem, disk caches and Grid farms are connected via 10 Gbps links. Physics data are stored on tape analogous to detector data. OSG middleware is used for workflow/job management. Enstore, SAMcache and dCache are used for data storage/access.

The same Grid farms (and for D0 also the desktop cluster) are used in the early stages of the analysis process. In this case, physics data is either transferred to the worker node or read over the network from the disk-based cache. Output is sent to disk servers for temporary storage.

Desktop and workgroup servers are used in the late stages of the analysis. Selections of 100s to 10s of GB are transferred and analyzed locally. The network speed in the offices is 100 Mbps or 1 Gbps.

The primary data archive for each experiment is the Fermilab Enstore managed tape-library system.

8.2.2 Process of Science:

Data analysis consists of repeated selections and data versus Monte Carlo comparison. Selections start from physics and time subsets of data. Collisions/events with unwanted characteristics are filtered out and unnecessary information in the events dropped.

Results are compared, verified and reviewed in small groups for scientific correctness.

Each experiment has about 5 to 10 video equipped conference rooms.

8.3 Key Remote Science Drivers

8.3.1 Instruments and Facilities:

Collision data originate at Fermilab and are stored at Fermilab.

Monte Carlo data are generated at Fermilab and Grid sites in America (MIT, Florida, Wisconsin, UCSD, Michigan State, Oklahoma, Gaia/Canada, and others), Europe (Bologna/Italy, Karlsruhe/Germany, IN2P3/France, and others) and Asia (Academia Sinica/Taiwan) and sent to Fermilab for storage (no input, just output data). OSG and LCG tools are used for workflow/job management. Kerberized rcp, scp, rsync and GridFTP are used for output file transfer.

Grid sites in North America and Europe re-process D0 collision data (new D0 data are processed at Fermilab) and send the result back to Fermilab for storage (input and output data from/to Fermilab using GridFTP, SRM storage at some sites).

Desktop and departmental analysis clusters in universities and laboratories in North America, Europe, Asia and South America are used to analyze the data. This is done either by working over the network on Fermilab computers or fetching small subsets of data from Fermilab to the remote sites.

8.3.2 Process of Science:

Fermilab is the data repository for both Tevatron experiments. All data intensive computing is performed at Fermilab. Software development, detector monitoring, work over-the-network on Fermilab computers and local analysis are performed around the globe.

About a quarter of the detector monitoring shifts at CDF are done from remote. About half of the CDF offline system monitoring is done from outside Fermilab.

8.4 Local Science Drivers – the next 2-5 years

8.4.1 Instruments and Facilities:

The Tevatron collider is currently scheduled to switch off in 2011. Analysis is expected to tail out about 3 years later. (Shutdown of the Tevatron experiments depends on LHC startup and performance.)

8.4.2 Process of Science:

No change.

8.5 Remote Science Drivers – the next 2-5 years

8.5.1 Instruments and Facilities:

New Grid sites are expected to be integrated (LCG sites for D0, KISTI, Korea for CDF). CPU available to CDF and D0 at Grid sites is expected to diminish as LHC turns on.

8.5.2 Process of Science:

No change.

8.6 Beyond 5 years – future needs and scientific direction

Legacy data analysis could continue well into the LHC era.

8.7 Summary Table

Feature	**Key Science Drivers**		**Anticipated Network Requirements**	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• CDF and D0 detector data • Monte Carlo data generated at Grid sites • Fermilab based Grid farms • Grid sites in America, Europe and Asia.	• repeated selections with data versus Monte Carlo comparison • collaborative work	• 10 Gbps switch interconnect • 100 Mbps, 1 Gbps server, worker node and desktop connectivity • Kerberized login and file transfer services	• 500 Mbps to and from Fermilab from CDF, D0 institutions and Grid sites in America and Europe (about 25% on Starlight)
2-5 years	• Tevatron and detectors switching off • use of less CPU at more Grid sites	• no change	• no change	• about the same
5+ years	• ceased need for Monte Carlo data	• legacy analyses	• reduced needs	• reduced needs

9 Neutrino Program at Fermilab

9.1 Background

The neutrino program of Fermilab is based on 8 GeV protons from the Booster and 120 GeV protons from the Main Injector. First generation experiments are in a mature data taking state (MiniBooNE, ArgoNeuT, MINOS) or have completed the operations phase (SciBooNE). MiniBooNE has the largest data volume, accumulating about 100 TB/year. Next generation experiments are under construction (Minerva), or approved (NOνA), and plans for the long-term future (DUSEL) are being worked on. Both Minerva and NOνA expect data volumes of about 30 TB/year. The goals of the experiments are neutrino oscillation and neutrino-nucleus cross-section measurements.

The detectors are operated by collaborations of up to 200 scientists from 35 institutions from North and South America and Europe (France, Germany, Greece, Russia, UK).

9.2 Key Local Science Drivers

9.2.1 Instruments and Facilities:

Data are produced by neutrino detectors on the Fermilab site. When the detector is deployed underground (underground deployment is the production configuration), the data rates are small – about 3 GB/day. Data are sent to the Fermilab compute center for storage and processing.

Dedicated clusters and Grid farms at Fermilab are used to process the detector data and produce physics data. For MINOS production output is 5 times larger than detector data but a subsequent N-tuple stage reduces the data volume used for analysis by a factor 2.5.

Physics and/or N-tuple data are copied to fileservers and high-performance NAS disks. Processed data is archived in the Fermilab Enstore managed tape-library system.

Most experiments have a dedicated cluster at Fermilab for analysis and plan to perform any high data volume analysis on-site. MINOS uses also the Fermilab Grid farms for analysis.

The NAS disks have a 10 Gbps connection while fileservers, nodes in the analysis clusters and Grid farms have 1 Gbps connections.

MiniBooNE uses the Fermilab Grid facilities also for Monte Carlo generation.

MINOS uses the Fermilab Grid farms for Monte Carlo mixing, reconstruction and N-tupling. The resulting data, about 3 TB/year, is archived on tape and stored on NAS disks for analysis.

The primary data archive for all experiments is the Fermilab Enstore managed tape-library system.

9.2.2 Process of Science:

Data analysis consists of repeated selections and data-versus-Monte Carlo comparison.

Collaborative work is mainly telephone based. Most experiments use video conferencing infrequently. Minerva has the largest use with less than 10 meetings per week.

9.3 Key Remote Science Drivers

9.3.1 Instruments and Facilities:

Far-side detectors are located in the Soudan mine and at Ash River south of Voyageurs National Park, both in Minnesota. MINOS sends DAQ, detector control and beam signals from Fermilab to Soudan. Also NOvA plans to trigger both near and far detectors on beam coincidence. The MINOS far detector produces data at about 1 GB/day. The NOvA far detector is just below the surface and yields 55 GB/day of data.

For backup, the University of Minnesota fetches a copy of all MINOS detector data from Fermilab, about 1 TB/year.

Depending on analysis needs individual institutions may fetch the whole MINOS N-tuple data from Fermilab and re-process them locally.

Monte Carlo generation and simulation for MINOS is done on departmental clusters at collaborating institutions (mainly Caltech, Minnesota, Tufts, College of William and Mary, Rutherford Laboratory, UK), about 4 TB of data per year.

Data are transferred via Kerberized ftp, rcp and http.

9.3.2 Process of Science:

Software development and data analysis is done either over-the-network on the analysis cluster at Fermilab or locally on a small amount of manually transferred data.

9.4 Local Science Drivers – the next 2-5 years

9.4.1 Instruments and Facilities:

MINOS and MiniBooNE will most likely end data taking around 2011.

Minerva is scheduled to start data taking in 2010. The Minerva detector will produce about 10 TB/year.

NOvA will get prototype detector data in 2010 and is scheduled to start data taking in 2013. The NOvA near detector will produce about 1 TB/year.

9.4.2 Process of Science:

No change.

9.5 Remote Science Drivers – the next 2-5 years

9.5.1 Instruments and Facilities:

Minerva will generate about 4 TB/year of Monte Carlo data on clusters at collaborating institutions (North and South America and Europe) and will sent those to Fermilab for storage and analysis (Kerberized ftp, rcp, scp).

NOvA will get first data from a few-module far detector at Ash River, MN in 2012. The full far detector will produce about 19 TB/year.

NOvA will produce about 10 TB/year of Monte Carlo data at Grid facilities of Fermilab and other collaborating institutions (Southern Methodist University, TX).

9.5.2 Process of Science:

No change.

9.6 Beyond 5 years – future needs and scientific direction

The long term plans of the neutrino program focus on the Deep Underground Science and Engineering Laboratory, DUSEL. A large water Cherenkov or liquid Argon detector could start operation in DUSEL at Homestake, SD between 2019 and 2021.

9.7 Summary Table

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• Near detectors at Fermilab, far detector at Soudan, MN • Fermilab analysis cluster and Grid farms • Departmental clusters at collaborating institutions	• Repeated selections with data versus Monte Carlo comparison	• 10 Gbps NAS disk • 1 Gbps servers and worker nodes • Kerberized login and file transfer services	• 1 Mbps link to far detector in Soudan • 100 Mbps off-site link from Fermilab
2-5 years	• Far detector at Ash River, MN	• No change	• No change	• 10 Mbps link to far detector in Ash River • 5 Mbps link to/from Fermilab to collaborating institutions
5+ years	• Possible large scale detector at DUSEL, SD	• No change	•	•

10 SLAC Experimental HEP Programs

10.1 Background

The BaBar experiment at SLAC acquired a petabyte-scale dataset that is now under intense analysis at SLAC and European computer centers.

SLAC Joined the ATLAS experiment in 2006 and has been running an ATLAS Tier 2 center since 2007. SLAC is seeking DOE support for a substantial expansion of ATLAS analysis capacity at SLAC in 2011 and beyond.

SLAC is also helping plan a possible Super B-Factory in Italy that could start taking data as early as 2015.

10.2 Key Science Drivers

10.2.1 Instruments and Facilities:

BaBar has an active data sample of close to a petabyte that is under analysis at SLAC and European Tier-A centers. Intense analysis of this sample will continue through 2010, to be followed by two years of "steady analysis" on a lower scale. Activity at SLAC will stay relatively constant through 2011 – the ramp down of European Tier-A centers will offset the 2011 decline in total BaBar activity. A lower level of "steady analysis" will continue until the end of 2012, followed by a period of much lower activity where the focus will be on data curation.

SLAC runs an ATLAS Tier 2 center. The organized production data flows between BNL and SLAC are well understood. The expected "chaotic" physics analysis workflow is very poorly understood. SLAC expects that the physics analysis workload will be very high and will justify expansion of ATLAS computing at SLAC to create a Western Analysis Facility

If a SuperB collider is funded in Italy, SLAC will likely be a major participant and the center of US SuperB activity. The SuperB accelerator should deliver 100 times the luminosity of the PEP-II B-Factory at SLAC.

10.3 Summary Table: SLAC Experimental HEP Programs

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• BaBar	• Analysis of a ~1PB dataset + some new simulation production in the US and Europe	• No problem	• 1TB/day (100Mbps), mainly to and from Europe
	• ATLAS	• Operate SLAC Tier2 center (move data to and from BNL, move data to Universities)		• 13TB/day in and 8TB/day out (total 2.1Gbps sustained), mainly to and from Europe
	• SuperB	• Construction, etc.		• Minimal
2-5 years	• BaBar	• Ramping down	• No problem	• Less than 1TB/day
	• ATLAS	• Tier2 ramp-up with increasing data • Possible Western Analysis Facility	• Computer center will need Terabits per second	• ~30TB/day in and 30TB/day out (6Gbps sustained) • 150 to 300TB/day possible (15-30Gbps)
	• SuperB	• Construction, etc.	• No problem	• Minimal
5+ years	• ATLAS	• ATLAS Tier2 alone • Western Analysis Facility	• No problem • Computer center will need Terabits per second	• 50 to 100TB/day • 500+TB/day (50Gbps)
	• SuperB	• Major SuperB analysis center at SLAC	• No problem	• 50-300TB/day, rising with time (30Gbps)

11 Daya Bay

11.1 Background

Recent discoveries in neutrino physics have shown that the Standard Model of particle physics is incomplete. The observation of neutrino oscillations has unequivocally demonstrated that the masses of neutrinos are nonzero. The small magnitude of neutrino masses (<2 eV) and the surprisingly large size of the two mixing angles thus far measured have provided important clues and constraints to extensions of the Standard Model. The third mixing angle, theta13, is small and has not yet been determined.

The Daya Bay collaboration will perform a precision measurement of this mixing angle by searching for the disappearance of electron antineutrinos from the nuclear reactor complex in Daya Bay, China, one of the most prolific sources of antineutrinos in the world.

11.2 Key Science Drivers

The PDSF Cluster at NERSC is the US Tier 1 center for Daya Bay simulation and data processing. The HPSS mass storage system at NERSC is our main US 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 fall 2011, we will reach full raw data rate of approximately 260 GB per day resulting in 150 TB of storage usage annually. We will be taking data at 25% rate starting in summer 2010 and at relatively modest rates until then. Near-real-time delivery of data to the US from the detector site in China is necessary to ensure US scientists access to data equal to Chinese scientists.

The data sets consist of 1GB data files. There is additional metadata to be 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 must also support collaboration services such as videoconferencing. Outage recovery is expected to take place within the duration of the outage (e.g. 2x capacity – phrase this better). 1GB data sets are expected to be transferred at a nominal rate of about 50Mbps, resulting in a transfer time of approximately 3 minutes per data set. Analysis and simulation will occur at both IHEP and NERSC, and the resultant data sets will be exchanged between IHEP and NERSC.

Daya Bay uses DCS videoconferencing between US and Chinese institutions. We have recently installed videoconferencing hardware on site at the Daya Bay nuclear power plant.

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

11.3 Science Drivers – the next 2-5 years

No change.

11.4 Beyond 5 years – future needs and scientific direction

Network usage will ramp down as the experiment concludes.

11.5 Outstanding Issues

Trans-Pacific end-to-end networking from IHEP in Beijing to LBNL in Berkeley across GLORIAD have shown much lower bandwidth than theoretically possible (see plot). Our diagnostics of the problem are still awaiting coordination between perfSONAR nodes in China and Hong Kong with nodes in the US.

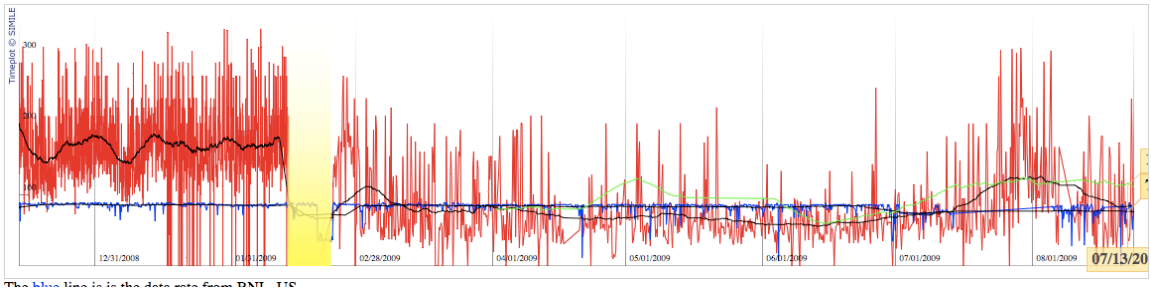


Figure 1: Disk to disk transfers from IHEP to NERSC. Individual transfers are in red, and a moving average is in black. The green line is the moving average for a different reference host in China, and the blue line is a reference test between BNL and NERSC.

Changes in network configuration implemented by CSTNet or IHEP engineers are often surprises that take time to recognize and diagnose. Establishing personal relations with Chinese and Hong Kong network engineers has started to help, but a more formal avenue would be beneficial.

Instabilities in network connections between IHEP and CSTNet are more frequent and longer than those typically seen in the US.

Also, US collaborators routinely meet problems with Chinese content filters when in China at the experimental site or collaborating institutions.

11.6 Summary Table

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
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) • Data sets are 1GB files • Derived data sets are 10% of raw data set size • Simulated data sets are 40% of raw data set size • Database synchronization and other traffic in addition to data traffic	• Analysis of raw, derived, and simulated data to determine the theta-13 mixing angle • Transfer of raw data from detectors to IHEP in Beijing, and from IHEP to NERSC • Transfer of simulated and derived data sets between IHEP and NERSC	• N/A	• OC1 between Daya Bay and IHEP • OC3 between Daya Bay and IHEP by summer 2011 • 100Mbps throughput expected • Collaboration services, including videoconferencing
2-5 years	• As above	• No change	• N/A	• No change
5+ years	• As above	• No change	• N/A	• No change

12 Astrophysics/Astroparticle

12.1 Background

The sky surveys map and measure light spectra at night with large telescopes. The Sloan Digital Sky Survey, SDSS, started mapping the southern sky in 1999. During the 9 years of SDSS-1 and SDSS-2 data taking, 70 TB of data were collected. Now SDSS-3 (BOSS) is collecting data until 2015. The SDSS collaboration of about 150 scientists is making the data publically accessible about a year after recording/processing.

The next generation Dark Energy Survey, DES, with its 500 Megapixel camera, will start data taking in 2011. It will record about 300 GB per night. The 5-year data taking and processing will yield about 4 PB of data including a 350 TB database. Scientists from 25 institutions in North, South America and Europe collaborate on DES.

The goal of CDMS (Cryogenic Dark Matter Search) is to detect dark matter via cryogenic Germanium and Silicon detectors. Data taking started in 2001. An upgrade of the experiment, SuperCDMS, to a 25 kg detector is being worked on. The collaboration consists of about 50 scientists from 14 institutions (US plus one from Germany).

12.2 Key Remote Science Drivers

12.2.1 Instruments and Facilities:

Data are produced by the telescope at Apache Point Observatory in New Mexico (SDSS) and Cerro Tololo Inter-American Observatory in Chile. In the first 5 years of operation SDSS sent the data to Fermilab via tape. With the upgrade of the microwave link of the observatory data is now transferred via network within a day to Fermilab. (The data enter ESnet in Albuquerque, NM.) DES expects to transfer its data over the network to NCSA in Urbana, IL. The microwave link from the mountains is the critical and most bandwidth-limited path.

Dedicated computing resources at Fermilab do the processing of the SDSS data. DES data will be processed on TeraGrid at NCSA. The outputs of the processing are flat files in a directory structure (as db) and a smaller SQL database. Fermilab is the primary repository of the SDSS-1 and SDSS-2 data. NERSC at LBNL will receive the SDSS-3 data, do the processing, and host the primary repository. The SDSS-1 and SDSS-2 data will be included in the SDSS-3 repository. Secondary archives of SDSS data in India and the UK exist. Secondary archives for DES data are planned in the UK and Spain.

Fermilab is serving SDSS-1 and SDSS-2 data at a rate of about 350 GB/day. It will continue serving data until ~2013. The data are accessed via http, rsync (flat files) and via SQL queries from around the world. (During the last quarter the SDSS repository at Fermilab was accessed from over 78,000 distinct IP addresses.)

The University of Portsmouth, UK provides an educational web service called “Galaxy Zoo” that uses the SDSS repository at Fermilab.

The CDMS detector produces data at a rate of 10 to 20 TB/year. It is located at the Soudan mine in Minnesota. The data are sent in quasi real-time to Fermilab.

Fermilab stores and processes the detector data. The size of the resulting physics data is about 3 to 7 TB/year. The physics data are sent from Fermilab to Stanford University and the University of Minnesota and from there distributed to the other institutions of the CDMS collaboration.

12.2.2 Process of Science:

For the sky surveys data are recorded during the night at the observatories and sent for storage and processing to the primary repository during the following 24 hours. Processed data is accessed on demand from desktop PCs of scientists around the world.

Software development is done over the network or locally via remote code repositories.

Collaborative work is telephone based with occasional video conferencing.

For the dark matter search data is recorded DC (Direct Current, i.e. continuously), processed, distributed via a two-tier system and analyzed at collaborating institutions.

12.3 Remote Science Drivers – the next 2-5 years

12.3.1 Instruments and Facilities:

SDSS-3 will record spectroscopy data (SDSS-1 and SDSS-2 was mostly imaging data), which results in half to a third times smaller data.

The libraries of the University of Chicago and Johns Hopkins University will handle long-term preservation of the data.

DES data taking is scheduled to start in autumn 2011. The primary repository will be at NCSA in Urbana, IL with secondary archives at Fermilab, Spain and the UK.

CDMS plans to distribute the physics data from Fermilab directly to all 14 collaborating institutions. SuperCDMS will increase the data volume by a factor of ten.

12.3.2 Process of Science:

No change.

12.4 Beyond 5 years – future needs and scientific direction

The CDMS collaboration is planning two more upgrades beyond the 25 kg detector: In about five years a 100 to 150 kg detector to be located in SNOLAB, Sudbury, Canada will produce about 300 TB per year. A 1-ton detector at DUSEL in South Dakota would go into operation in about 10 years with a data rate of 1 to 2 petabytes per year.

12.5 Summary Table

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• Apache Point Observatory, NM • CDMS at the Soudan mine, MN • Fermilab for SDSS-1 and SDSS-2 and CDMS • NERSC at LBNL, CA for SDSS-3 • Stanford and Minnesota universities are data hubs of CDMS	• Recording of imaging and spectroscopy data with telescopes • Processing to flat files and databases • On demand analysis by scientist around the world • Central processing and analysis at institutions	• N/A	• 20 Mbps from Apache Point Observatory to NERSC at LBNL • 100 Mbps from Fermilab to desktops around the world • 10 Mbps from Fermilab to Stanford and Minnesota universities • 100 Mbps from each Stanford and Minnesota universities • 1 Gbps to fill replica in a week
2-5 years	• Cerro Tololo Inter-American Observatory • NCSA in Urbana, IL	• No change	•	• 100 Mbps from Cerro Tololo Inter-American Observatory to NCSA in Urbana, IL • 200 Mbps from Soudan to Fermilab • 200 Mbps from NCSA to desktops around the world • 1 Gbps from NCSA to secondary archives at each Fermilab, in Spain and the UK • 1 Gbps from Fermilab to CDMS institutions
5+ years	• SuperCDMS 100kg detector at SNOLAB • SuperCDMS 1 ton detector at DUSEL	•	•	• 200 Mbps from SNOLAB to Fermilab • 10 Gbps from Fermilab to SuperCDMS sites • 1 Gbps from DUSEL to Fermilab

13 Cosmology (Low Redshift Supernova Studies)

13.1 Background

This case study describes several Low Redshift Supernova studies: the Supernova Factory (concluding), the Low-Redshift Supernova Program (in installation), and the Palomar Transient Factory (commissioning).

The goal is to collect a large sample of data from type Ia supernovae to better understand their characteristics and to use them to study the evolution of the universe. Low redshift corresponds to the most recent, or present time and is compared to high redshift or early time in the history of the universe.

13.2 Key Science Drivers

13.2.1 Instruments and Facilities: Supernova Factory

To search for supernovae we look at the sky each night and then compare the images with older images and determine which objects, if any, have brightened considerably. To do the processing, the images are sent from the telescope (Palomar in California, or soon, La Silla in Chile) to the National Energy Research Scientific Computing Center (NERSC). It is critical to get immediate feedback to schedule follow-up on other telescopes, principally the University of Hawaii 2.2 meter telescope on Mauna Kea. The follow-up data are sent to Lyon France for processing and results then come to LBNL.

13.2.2 Instruments and Facilities: Baryon Oscillation Spectroscopic Survey (BOSS)

This new program uses large galaxy surveys as a new and powerful approach to understanding cosmology and dark energy. Until recently, Type Ia supernovae were the only proven technique for measuring the expansion history of the Universe, and hence the geometrical effects of dark energy. In 2005, the complementary technique of baryon acoustic oscillations (BAO) was demonstrated. The BAO scale was measured in large galaxy surveys, and this scale (unlike supernovae) can be tied directly to the scale of the CMB in the early Universe. The first BAO detections were made in the Sloan Digital Sky Survey (SDSS). This survey has imaged 100 million galaxies in 5 filters and taken spectroscopy for 1 million galaxies, from which we generate 3-dimensional maps of the Universe. This is the largest galaxy survey to date. The SDSS at Apache Point telescope remains the premier instrument for measuring BAO. A collaboration led by LBNL is upgrading the SDSS spectrographs and pursuing a dedicated BAO experiment from 2009 through 2014.

13.2.3 Process of Science:

The BOSS collaboration develops a target list in advance so that there is no searching required and the processing is not time-critical. The spectroscopic data are transferred to the Reimann Cluster at LBNL for processing.

The experiment is now getting test data from the telescope over network connections. There are currently no bottlenecks. Commissioning will continue through December 2009 at which time higher rate transfers will begin. No problems are anticipated but they should conduct some bandwidth tests and, if necessary, identify bottlenecks.

13.3 Science Drivers – the next 2-5 years

13.4 Beyond 5 years – future needs and scientific direction

Two new programs are on the horizon:

- BigBOSS: Similar to BOSS at the NOAO telescope.
- DomeA: Supernova Studies (low and high redshift) using telescope on Antarctica (but not South Pole). Might do computing for search on-site but would need at least sample data transmitted to LBNL. This is a collaboration with China.

13.5 Outstanding Issues

The current network works well for Palomar with wireless to SDSC and ESnet beyond. Chile will be harder. There are other DOE programs also observing in Chile: DES, LSST.

The BOSS collaboration currently uses telephone conferencing with slides on the web (WebEx). Some members find this unsatisfactory as it does not allow visual cues of people who wish to speak etc. For this they would like HD Video supported in the hubs. They tried EVO but found it unsatisfactory because of set-up time and frequent disruptions.

13.6 Summary Table

Table not provided

14 Large Synoptic Survey Telescope

14.1 Background

The Large Synoptic Survey Telescope will take one large-aperture image of the Chilean night sky every 15 seconds for ten years. The rich science program will include studies using gravitational lensing and a search for dark energy. The images will be processed in real time in Chile and at NCSA to generate alerts for other observatories. The full image data will be stored at NCSA. NCSA will process the full image data creating derived catalogs at various levels of detail. Derived catalogs, plus limited image data, will be distributed to Data Access Centers. Specialized Analysis Centers may also exist.

14.2 Key Science Drivers

14.2.1 Instruments and Facilities:

SLAC expects to construct the 3 gigapixel LSST camera in 2011-2013. SLAC will also propose to host a Data Access Center and possible a center specializing in data-intensive dark energy analysis.

14.2.2 Process of Science:

- Real-time transmission and processing of images to generate alerts.
- Off line analysis by members of the LSST Collaboration using both the catalogs and raw or processed (e.g. co-added) image data.
    - Object catalog (up to 1 PB/year) and analyzing CPUs must be co-located;
    - Location of image storage (10 PB/year) may depend more on relative costs of network and storage.
- Public availability of data.

14.3 Summary Table: LSST at SLAC

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)		• Software development and small-scale tests	• Minimal	• Minimal
2-5 years	• Camera construction	• Test real-time transmissions to NCSA of 6 Gigabyte images in 2 seconds	• 30 gigabits/s	• 30 Gbps
5+ years	• 2015 on, operation of Data Access Center and/or Dark Energy Analysis Center	• Non-real time transfer of data from NCSA to persistent local storage, or • High speed transfer of data from NCSA to transient local storage	• Computer Center: ~terabits/s • Site: ~10 Gbps	• 10 Gbps (with enough local storage to obviate the need for high-speed transfers from NCSA.), or perhaps • 100s of Gbps with reduced local storage

15 Particle Astrophysics and Cosmology at SLAC

15.1 Background

Particle Astrophysics and Cosmology at SLAC involves both observational and theoretical studies. LSST is the example of the most demanding observational study and has its own SLAC case study.

From the founding of SLAC's Kavli Institute for Particle Astrophysics and Cosmology (KIPAC), the institute has attracted faculty interested in computationally intensive theoretical studies of topics such as star formation in the early universe and colliding galaxies.

15.2 Key Science Drivers

15.2.1 Instruments and Facilities:

Facilities used include remote "Leadership Class" machines for simulation and SLAC-site clusters for both simulation and analysis. Shared memory SMPs have proved valuable in "sandbox" style exploration of theoretical approaches with minimal constraint from the computer architecture.

Scalable, sharable file systems (Lustre is in current use) are very important in support of collaborative work.

15.2.2 Process of Science:

- Day-to-day use of SLAC site facilities
- More occasional use of leadership class facilities, transferring data back to SLAC for analysis.

15.3 Summary Table: Particle Astrophysics and Cosmology at SLAC

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	•	• Simulate on local and leadership class machines (few TB/run); • Analyze at SLAC	• Not a problem	• Transfer a few TB/night • Needs 1 Gbps
2-5 years	•	• Simulate on local and leadership class machines (few tens of TB/run); • Analyze at SLAC	• 10 Gbps	• Transfer ~25 TB/night • 10 Gbps
5+ years	•	• Simulate on local and leadership class machines (up to 100s of TB/run); • Analyze at SLAC	• Computer Center: ~terabits per second • Site: ~40 Gbps	• ~150 Gbps (if data transferred to SLAC for analysis)

16 Accelerator Modeling at SLAC

16.1 Background

Accelerator modeling is an essential part of the initial R&D for, the design of, and the operation of accelerators.

SLAC's Beam Physics department uses computation to model electromagnetic structures, and also to model the transport of particles within accelerators. The applications range from end-to-end simulations of the LCLS (Linac Coherent Light Source – BES), to designing the RFQ cavity for the planed FRIB (Facility for Rare Isotope Beams – NP), to modeling the accelerating structures of a future Linear Collider. SLAC is a member of the COMPASS SciDAC project.

16.2 Key Science Drivers

16.2.1 Instruments and Facilities:

Facilities used include remote "Leadership Class" machines for simulation, SLAC-site clusters for simulation, and SLAC-site clusters dedicated to the analysis of simulation runs.

16.2.2 Process of Science:

- Day-to-day simulation of operation facilities needs dedicated, and thus normally on-site, facilities.
- Accelerator R&D and design has fewer near-real-time constraints:
    - SLAC-site clusters are used for medium-scale simulations
    - SLAC-site clusters optimized for interactive analysis are used to analyze medium scale simulation output
    - Remote leadership-class machines are used for large simulation runs, generating terabytes today and expecting petabytes for simulations that reach realistic ILC beam sizes
    - SLAC-site clusters are currently used to analyze leadership-class simulation output

16.3 Outstanding Issues

- Where should the analysis of petabyte-scale simulation output be performed in about 5 years from now? If there is no break-through on remote visualization and analysis, the current practice will continue and petabyte-scale simulation data will have to be transferred to SLAC for local analysis. That will add a significant requirement for the future network.

16.4 Summary Table: Accelerator Modeling at SLAC

Feature	Key Science Drivers		Anticipated Network Requirements	
Time Frame	**Science Instruments and Facilities**	**Process of Science**	**Local Area Network Bandwidth and Services**	**Wide Area Network Bandwidth and Services**
Near-term (0-2 years)	• Existing, planned and projected accelerators	• Simulate on local and leadership class machines (few TB/run); • Analyze at SLAC	• Not a problem	• Transfer a few TB/night • Needs 1 Gbps
2-5 years	• Existing, planned and projected accelerators	• Simulate on local and leadership class machines (few tens of TB/run); • Analyze at SLAC	• 30 Gbps	• Transfer ~50 TB/night • 20 Gbps
5+ years	• Existing, planned and projected accelerators (e.g. realistic ILC)	• Simulate on local and leadership class machines (up to 1 PB/run); • Analyze at SLAC	• Computer Center: ~1 Tbps • Site: ~10 Gbps	• ~350 Gbps (if data transferred to SLAC for analysis)

17 Findings

There were a number of findings from this workshop, many of which are related to the LHC. We present the LHC findings in a group, followed by the others.

LHC Findings

There were several findings related to the Large Hadron Collider (LHC), as this is the largest distributed computing project currently underway in High Energy Physics.

- Some scientists at the workshop felt that data transfers between LHC Tier-2 sites was very likely to occur, and might be significant. These transfers are not accounted for in the standard tiered data distribution model. In the United States, most Tier-2 sites are on university campuses and so ESnet is unlikely to see this traffic – however, this could be a significant contributing factor to high traffic load in campus or regional networks.
- It is likely that ad hoc, science-driven, opportunistic data analysis will generate a significant amount of data movement traffic. These data transfers are difficult to quantify in advance, but an attempt was made to estimate their impact. One likely source is secondary analysis of primary derived data sets of global interest. The consensus at the workshop was that the combined network traffic load generated would be roughly equal to a Tier-2 center. Managing the traffic flows that result from ad-hoc data analysis will require significant flexibility from the networks that transport the data. This will require vigilance on the part of ESnet and LHC sites to identify large traffic flows outside the LHC tiered data distribution model and ensure that they are handled properly end to end (e.g. by moving them to SDN circuits). Some of this traffic will add load to the trans-Atlantic circuits. The bandwidth needs for ad-hoc analysis could increase by a factor of 6 over the next five years.
- LHC Tier-2 and Tier-3 site networks will have a steep learning curve. ARRA funds will add a significant number of additional Tier-3 sites. The funds available are likely to be used for computational and storage resources, not for site networking upgrades. This means that tuning Tier-3 site networks will be very important, and that existing site networks that may or may not be able to accommodate the increased traffic load.
- Trans-Atlantic traffic between Tier-1 and Tier-2 centers may grow quickly, causing congestion on the trans-Atlantic circuits. Trans-Atlantic traffic load needs to be watched, and the traffic between Tier-1 and Tier-2 centers needs to be tracked and understood. Given the lead time for trans-Atlantic circuits (six months for procurement and provisioning is typical), a clear understanding of the traffic dynamics is important.
- Trans-Atlantic circuits, due to their nature (they are carried on sub-sea cables), can have significantly longer repair times than is typical with land-based circuits. Outages on the order of weeks for undersea cables are not unheard of. Given the impact that protracted outages can have, it is imperative that trans-Atlantic connectivity be geographically diverse, and allow for adequate service even in the face of multiple failures. This is currently the case, and it is critical that this diversity be maintained for the foreseeable future.

- There is likely to be a need for performance tuning assistance at the Tier-3 sites. Test and measurement infrastructures such as perfSONAR will be a big help. ESnet has deployed an extensive perfSONAR test and measurement infrastructure that can be used for troubleshooting network performance problems. Tier-3 sites should be encouraged to deploy perfSONAR to help them find and fix network performance issues.
- Virtual circuits are already a critical part of the networking infrastructure that supports the LHC. They are an integral part of the LHCOPN, and the bulk of Tier-1 to Tier-2 connectivity in the United States uses ESnet virtual circuits on the ESnet Science Data Network (SDN). Virtual circuits provide traffic engineering capabilities that are used to ensure diversity and to protect LHC traffic from other traffic (and vice versa). Virtual circuits also provide bandwidth guarantees that, among other things, provide for guaranteed minimum network service levels for access to Tier-1 data resources by Tier-2 centers.
- The approximate data set size for analysis at Tier-2 centers is 30TB. Moving this 30TB data set in a reasonable amount of time (i.e. eight hours) requires a 10Gbps networking capability. This means that all Tier-2 centers may need to upgrade their network infrastructure to accommodate 10Gbps data transfers from Tier-1 centers in the near future, and that some of the current data rate projections for Tier-1 to Tier-2 transfers may be low.

Other Findings

LSST: SLAC is building the camera for the Large Synoptic Survey Telescope (LSST). During the construction and testing of the camera and its data acquisition system, there is a need to support high-speed data flows between SLAC and the National Center for Supercomputing Applications (NCSA) where the LSST data repository will reside. Planning discussions between ESnet, NCSA, and SLAC are ongoing.

China – Getting reasonable network throughput to China continues to be very challenging. Chinese networks are behind US networks in the deployment of test and measurement infrastructures such as perfSONAR. ESnet is engaged in ongoing outreach to our Chinese colleagues to assist them to improve network performance between US and Chinese research institutions. It took over a year of work to achieve acceptable data transfer performance for Daya Bay, and this was largely accomplished by setting up automated tests and plotting the results over time. This allowed changes in performance to be discovered, analyzed, and improvements to be documented.

Video conferencing: The HEP community makes heavy use of videoconferencing facilities for collaboration. Some of the experiments depend on EVO from Caltech, and others depend on the ESnet ECS service. Many people in the community are satisfied with the level of videoconference support. However, there are complaints that debugging problems with the ECS gatekeepers continues to be very difficult if not impossible, and EVO users complain of long setup times and other difficulties. It is expected that videoconferencing service use will increase as the scope of collaborations increases over time. In some cases, it is difficult or impossible to use videoconferencing due to networking problems (e.g. to sites in China).

Future projects: There are several future HEP projects that have significant potential to impact ESnet. These include:

- Accelerator modeling at SLAC – possibility for 1PB/day data transfer from Leadership Computing Facilities to SLAC. The data sets would need to be moved overnight – given that moving 1PB in eight hours requires more than 300Gbps of throughput, this is significant. The transfers would be occasional, are contingent on the deployment of sufficient analysis equipment at SLAC, and are about five years away.
- SuperB detector in Italy would likely add traffic load to ESnet and the trans-Atlantic circuits starting in 2015.
- Particle Astrophysics simulations at the Leadership Computing Facilities (LCFs) could require 10Gbps of bandwidth in 2-5 years for periodic transfer of large data sets from the LCFs to SLAC.

18 Requirements Summary and Conclusions

Many of the network requirements for HEP-funded programs are well understood. However, there are some significant aspects of network usage by HEP that will be dependent on the data movement patterns resultant from science-driven analysis or are contingent on the funding of future projects.

The near-term requirements for HEP are dominated by the LHC. However, while many aspects of the LHC network requirements are well-understood, there are aspects that will only become clear after scientists have begun to analyze the data from the LHC and the network usage patterns for ad-hoc, science-driven analysis become clear.

Some of the significant bandwidth requirements are contingent on future funding, e.g. for the International Linear Collider (ILC) or Super-B in Italy. As the funding picture for these programs (and therefore their timing) becomes clearer, the networking needs for these projects will need to be reconsidered.

Action Items

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

- ESnet will work with the LHC community to help write a performance tuning guide for Tier-3 sites so as to help reduce the time spent on network troubleshooting.
- ESnet will help with network planning for the LSST camera construction at SLAC
- ESnet will continue its involvement with the LHCOPN in support of the LHC
- ESnet will work with sites to ensure that upcoming large data flows can be supported (applications include LSST camera construction, accelerator simulation data sets, etc)
- ESnet will continue to develop and update the fasterdata.es.net site as a resource for the community
- ESnet will continue to assist sites with perfSONAR deployments and will continue to assist sites with network and system performance tuning

In addition, ESnet will continue development and deployment of the ESnet On-demand Secure Circuits and Advance Reservation System (OSCARS) to support virtual circuit services on the Science Data Network.

19 Glossary

Booster (at Fermilab): The Booster is used to accelerate protons from the Linac to an energy of 8GeV

CSTNet: China Science and Technology Network. See http://www.cstnet.net.cn/english/ and http://www.cstnet.net.cn/english/aboutcstnet/milestones.htm

COMPASS: Community Petascale Project for Accelerator Science and Simulation. See https://compass.fnal.gov/

DAQ: Data AcQuisition

dCache: A system for storing and retrieving huge amounts of data, distributed among a large number of heterogeneous server nodes, under a single virtual filesystem tree with a variety of standard access methods. See http://www.dcache.org/

DPM: Disk Pool Manager – a lightweight solution for managing disk storage. See https://twiki.cern.ch/twiki/bin/view/LCG/DpmGeneralDescription

dq2-get: One of the DQ clients. See https://twiki.cern.ch/twiki/bin/view/Atlas/PandaDataService and https://twiki.cern.ch/twiki/bin/view/Atlas/DQ2ClientsHowTo for more information.

Enstore: A mass storage system used at Fermilab. See http://computing.fnal.gov/docs/products/enstore/

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

GLORIAD: Global Ring Network for Advanced Applications Development. GLORIAD provides international network connectivity in support of scientific research. See http://www.gloriad.org/

Hadoop: A software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of nodes and petabytes of data. Hadoop was inspired by Google's MapReduce and Google File System (GFS) papers. See http://hadoop.apache.org/

ILC: International Linear Collider – see http://www.linearcollider.org/

kHS06: A unit of computational resource measurement – thousands of HepSpec 2006, based on SpecInt 2006

Linac (at Fermilab): The Fermilab Linac is a negative hydrogen ion, 400 MeV accelerator. See http://linac.fnal.gov/

Main Injector (at Fermilab): The Main Injector accelerator accepts protons from the Booster and accelerates them to an energy of about 120 GeV

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

NAS: Network Attached Storage

pAthena: A distributed analysis interface to the ATLAS offline software framework Athena, and a component of the PanDA system. See https://twiki.cern.ch/twiki/bin/view/Atlas/PanDA for PanDA, and https://twiki.cern.ch/twiki/bin/view/Atlas/PandaAthena for pAthena and Athena.

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

Reprocessing: Event reconstruction processing that is re-run with improved algorithms, better calibration data, etc.

SciDAC: DOE's Scientific Discovery through Advanced Computing program. See http://www.scidac.gov/

Skimming: The extraction of a subset of a larger data set, e.g. a subset of events with similar event-level attributes that make the events interesting as a group unto themselves

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

Xrootd: The Scalla/XRootD software framework is a fully generic suite for fast, low latency and scalable data access, which can serve natively any kind of data, organized as a hierarchical filesystem-like namespace, based on the concept of directory. See http://xrootd.slac.stanford.edu/

20 Acknowledgements

This work would not have been possible without the contributions and participation of those who provided information and attended the workshop. ESnet would also like to thank the HEP program office for their help in organizing the workshop and providing insight into the facilities supported by the HEP program. In addition, the LBNL conference support and logistics staff was very helpful.

ESnet is funded by the US Department of Energy, Office of Science, 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 Office of High Energy Physics.

This is LBNL report LBNL-3397E