Advancement in Networks for HEP Community

Harvey B. Newman, Artur Barczyk, Azher Mughal

1200E California Blvd
Pasadena, CA 91125
+1 626 395 6656
Harvey.Newman@cern.ch

Abstract. The key role of networks has been brought into focus as a result of the worldwide-distributed computing model adopted by the four LHC experiments, as a necessary response to the unprecedented data volumes and computational needs of the LHC physics program. As we approach LHC startup and the era of LHC physics, the focus has increased as the experiments develop the tools and methods needed to distribute, process, access and cooperatively analyze datasets with aggregate volumes of Petabytes of simulated data even now, rising to many Petabytes of real and simulated data during the first years of LHC operation.

1. Introduction

This paper gives an updated overview of the status and outlook for the world's research networks and major international links used by the high energy physics and other scientific communities, network technology advances on which our community depends and in which we have an increasingly important role, and the problem of the Digital Divide, which is a primary focus of ICFA's Standing Committee on Inter-regional Connectivity (SCIC) [1].

With the emergence of optical networks supporting multiple links of 10 Gigabits/sec (Gbps) on each fiber-pair since the mid-1990’s, HEP’s bandwidth usage, and the typical capacity of the major national backbones and intercontinental links used by our field have progressed together by roughly a factor of 1000 times over the past decade, and the outlook is for a further major increase in capacity over the next decade as both the speed and performance per unit cost of optical networks continue to progress. The exponential growth in affordable network capacity, often outstripping the growth rates in other areas of information technology, is reflected in the capacity growth of the major mission-oriented networks serving our field (such as ESnet [2] and USLHCNet [3] in the US and across the Atlantic), as well as the major networks broadly serving research and education (such as GEANT2 [4] throughout Europe; Internet2, and National Lambda Rail in the US, and APAN in Asia), along with the leading national networks such as SURFNet in the Netherlands and CANARIE in Canada, which have progressed from a 10 Gigabits/sec (Gbps) backbone to multiple 10 Gbps links in their core within the past two years.

2. Scale of Networks

The scale of the required networks has been set by the distribution of data from the “Tier0” at CERN to the 11 “Tier1” centers, as well as the needs for further distribution to 140 Tier2 centers located at
sites throughout the world where more than 40% of the computing and storage resources will be located, and where the majority of the data analysis as well as the production of simulated data are foreseen to take place. This is complemented by hundreds of computing clusters (Tier3s) serving individual physics groups, where there will also be a demand for (1 to a few) Terabyte-scale datasets once the LHC program is underway. During the “data challenges” leading up to the start of the LHC, sustained dataflows among CERN, the Tier1s and Tier2s amounting to up to 5 Petabytes/month and peak transfer rates well above 20 Gbps have been observed, along with individual Tier1-Tier2 transfers at 10 Gbps and in some cases higher. Further increases are expected as the tools mature and the level of familiarity and expertise in transmitting data among the sites increases.

As summarized in this report, more advanced tools [5] capable of very high transfer rates (up to the 100 Gbps range in aggregate) [6] among relatively small sets of storage servers have continued to develop and mature in the last year, including the first tests of the next generation 100 Gbps links [7] following emerging ITU standards that are expected to be completed by the Summer of 2010. This promises to give the LHC program greater agility in distributing and sharing data, while at the same time tending to drive the continued expansion of network requirements over the coming years.

3. LHC OPN Network

In order to respond to the highest priority needs of the experiments, for data distribution from the CERN Tier0 to the Tier1s as well as data exchange among the Tier1s, the LHC Optical Private Network (OPN) [8] has been formed. As a “private network” where only designated flows between specified source and destination addresses are allowed, the OPN will serve to guarantee adequate, secure connectivity to and among the Tier1s, as long as the links continue to evolve in future years to meet the bandwidth needs. Each of the Tier1s is connected to CERN at a minimum bandwidth of 10 Gbps, with some Tier1s (such as BNL and Fermilab which connect to CERN over ESnet and USLHCNet; and the SARA Tier1 in the Netherlands) connecting using multiple 10 Gbps links. The initial configuration of the OPN was largely a “star” network centered on CERN, and this is being supplemented by an increasing number of additional links among the Tier1s in order to ensure that the OPN can provide round-the-clock, nonstop operation as required. Some of the major lateral paths among the Tier1s in the OPN also are progressing from one to several 10 Gbps links.

The Tier2s also have very important roles to play in the overall LHC Computing Model, providing much of the resources as well as being focal points for analysis. As part of this role, each Tier2 requires connectivity to its corresponding Tier1 (in the ATLAS version of the Model) or to the ensemble of Tier1s (in the CMS version of the Model). The ability of Tier2s to get at the datasets needed implies substantial data flow within as well as beyond the limits of the OPN, across GEANT2, Internet2 [9], NLR [10], and other national research and education networks (NRENs). The Tier2 connectivity requirements have been variously estimated in the range from 1 to 10 Gbps, depending on the availability and affordability of bandwidth in each region. The Tier2s in the US, for example, already have one or more 10 Gbps connections in place, in time for LHC startup.

4. Network Trends

4.1 Transition

A transition began in 2005 to the use of “hybrid” networks, where the general purpose backbones of the major research and education networks are complemented by the use of point-to-point “light paths” to support the most demanding applications, including the large-scale data transfers required by high energy physics. As seen throughout this report and many of the accompanying Appendices, this trend, is spreading throughout the U.S., Europe and Asia. The move to hybrid networks is driven by two key technologies that can provide major increases in the available bandwidth per unit cost: “dense wavelength division multiplexing” (DWDM) where multiple wavelengths of light on a single optical
fiber pair act as logically separate channels that can carry data at speeds of up to 10 Gbps each (and 40 Gbps or 100 Gbps in a growing number of cases), and “dark fiber” where organizations in or working on behalf of the research and education community acquire or lease the fiber infrastructure and operate and manage it themselves. The largest examples of this trend include National Lambda Rail and Internet2 in the U.S., whose deployment was completed in 2006 and in 2007 respectively, and GEANT2 across continental Europe. This is paralleled by similar developments in a growing list of countries through Europe and parts of Asia, as well as more than 30 U.S. states. ESnet4 (partly hosted on the Internet2 infrastructure) provides networking for the U.S. HEP labs and the other major science programs supported by the DOE Office of Science.

4.2 Mission-oriented networks
Mission-oriented networks serving high energy physics have been following a similar trend to the national and continental research and education networks, to help meet some of the highest priority needs of the LHC program in particular. The LHC OPN has been formed of continental and transoceanic links which are predominantly 10 Gbps each. The Energy Sciences Network has deployed ESnet4, where its general purpose 10 Gbps backbone will be complemented by a Science Data Network (SDN) whose backbone capacity is expected to reach approximately 50 Gbps within the next two years, followed by a migration to 100 Gbps links that will bring a major capacity increase. USLHCNet has migrated from four to six 10 Gbps links across the Atlantic, and is expected to provide eight 10 Gbps to support the transatlantic network needs of the LHC experiments and other major programs by approximately 2010, followed by further capacity increases across the Atlantic as it migrates to 40 Gbps and then 100 Gbps links.

These trends are complemented by the deployment of multi-wavelength metropolitan networks, as in the case of ESnet’s connections to the labs, or point-to-point links. Fermilab has a dark fiber connection to the “Starlight” peering point in Chicago where it connects to ESnet and US LHCNet, allowing it to add additional 10 Gbps links as needed, at relatively low cost. Most of the US Tier2 centers have installed 10 Gbps links to the Chicago (StarLight) or New York (MANLAN) peering point, giving them adequate connectivity to the Tier1 centers during the LHC startup phase.

4.3 Bandwidth Roadmap
The 5-10 year outlook is thus for the exponential growth of the networks used by our field, and the largest research and education networks in general, to continue to the Terabit/sec range by the middle of the next decade. A milestone in this continued growth will be the transition to 40 Gbps and 100 Gbps wavelengths, which is expected to be in full swing by 2011. CANARIE, the research and education network of Canada, is planning a transition to a network of up to 72 wavelengths of 40 Gbps, and the ESnet4 plan foresees a network backbone capacity upgrade by a factor of 10, to a backbone capacity of approximately 500-600 Gbps using 100 Gbps wavelengths, in the 2012 time frame. USLHCNet recently developed a five year roadmap that foresees a transition to 40 Gbps and then 100 Gbps links as they become available and cost-effective, with a transatlantic network bandwidth growth rate that is similar to the rate of growth of computing power and storage foreseen by the LHC experiments, resulting in 400 Gbps across the Atlantic by 2014.

It has now long been recognized that bandwidth itself on an increasing scale is not enough to meet the needs of the field. Realizing the scientific wealth of the LHC and our other major scientific programs depends crucially on our field’s ability to use the bandwidth efficiently and reliably, with high data throughput rates, and effectively, where many parallel large-scale data transfers serving the community complete with high probability, often while coexisting with many other streams of network traffic. Responding to these needs, physicists working with network engineers and computer scientists have made substantial progress in the development of protocols and systems that promise to meet these needs, demonstrating data flows in the 10 Gbps range between pairs of servers, and
aggregate data flows over multiple links at rates exceeding 100 Gbps. This has placed our community among the world leaders in the development as well as use of large-scale long range networks.

5. Recent Developments in High Speed Networks

In 2009, continued major advances were made in the ability of the latest low cost compute nodes and small data servers (of the type commonly used at Tier1 and Tier2 centers) to transport data between storage systems at high speeds. This was achieved by using new data transport applications, along with optimized end-system, network interface and protocol settings, so that a single rack of such systems could match the capacity of four to eight 10 Gbps links. The integration of these advanced tools with the production systems at the Tier1 and Tier2 centers began in the Fall of 2007, and is now well-advanced. The trend to higher speeds was facilitated in 2008 by the appearance of higher density compute nodes (more than 600 compute cores per rack) and servers with embedded 10 Gigabit Ethernet network interfaces. In 2010, further throughput increases are expected by the appearance of mass-market storage devices with I/O interface speeds at and above 100 MBytes/sec.

A complementary set of developments concerns the use of a new generation of optical multiplexers with emerging standard software that can create multiple logical sub-channels with robust bandwidth guarantees on a link. This helps ensure the timely delivery of the highest priority datasets, while isolating the high performance flows from the general network traffic. The transition to these new architectures recently took a major step forward in our field, as Internet2 USLHCNet, and SINET3 in Japan installed these multiplexers as their core equipment. A broad-based consortium of the major networks and optical network development projects serving our field also took a major step forward towards making “circuit-oriented” networks production ready. Within the past two years the GLIF and DICE consortia has agreed on the key concepts, developed a design, and begun to deploy initial implementations of secure circuits crossing multiple administrative domains. A great deal of work remains to fully realize these developments in support of the LHC experiments, and is continuing at an accelerating pace.

Advances in our field in the past year have been paralleled by remarkable ongoing developments in several fields of information technology, as well as cultural trends fueling rising network demand and large scale plans for continued network capacity expansion throughout the world. The rise in the pervasiveness and speed of “broadband” to homes and businesses, using both fixed line and mobile connections is accelerating, and a transition to “fiber to the home” supporting 100 Mbps and soon 1 Gbps speeds has begun in many areas. This is being driven by the distribution of high resolution video/audio, streaming content and high resolution images, as well as social networking and the overall expansion of the size of the “Digital World” (now encompassing more than 1022 bits of digital information stored and in flight). This is spurred on by increasingly affordable mass-produced computers, and Terabyte-scale internal and external disk storage, along with the rise of solid state disks and 10-gigabit/sec network interfaces in servers, all geared to move data at a gigabit/sec or more.

As a result of this accelerating demand, network core infrastructures have been under pressure. Additional fibers on both land-based and transoceanic cables have been lit to meet the demand, and new undersea cable projects across the Pacific, to Latin America and the Middle East have proliferated, with several coming into service in 2009. There is also a drive to put more 40 Gbps links into production now, and strong anticipation of the coming 100 Gbps links as soon as the standards are completed and adopted. Although no new transatlantic cables are foreseen yet, the electronics upgrade of amplifiers on several of these cables has begun, in order to accommodate 40 Gbps links already this year.
6. Conclusion

All of these new trends, as well as the technical developments, point to the central role of networks with larger and larger capacity, and their use to transport data with increasing agility. As in the past, we expect HEP to remain at the forefront of these trends when it comes to the transport of the largest datasets, putting technology advancements in storage, networks and network interfaces, and advances in data transport applications to the test as they become available, and then to full use. This will help maintain the exponential growth of high energy physics’ network bandwidth requirements and use, as has been experienced by our field over the last 25 years.

What is perhaps new is that the exponential growth of network demand and use is taking hold throughout the world, in those economies that can afford it. Many countries and cities are giving greater priority to access to broadband, recognizing it as an important part of societal development, economic advancement, and competitiveness in the global marketplace, so that the number of countries and municipalities moving towards universal broadband connectivity is increasing.

As we advance in all these areas, often with great rapidity, there is a growing danger that we are leaving behind our collaborators in regions with less-developed networks, less-developed general infrastructure, or regulatory frameworks and business models that put the required networks financially as well as technologically out of reach. This threatens to further open the Digital Divide that already exists among the regions of the world.

Since 2002, recognizing that this Divide is a crisis that effects society and the world economy, as well as the scientific community, the SCIC has worked assiduously through its monitoring efforts, workshops, and technical assistance and training on the ground, to reduce or eliminate the Divide, both within our scientific community, and more broadly in the world research community of which HEP is a part. Through years of effort we have made progress in several countries, where we have helped foster significant progress in the IT infrastructure as well as the methods and tools available for scientific research. But problems persist in many regions of the world, and so a great deal of work remains to be done.

References

[1] http://icfa-scic.web.cern.ch/icfa-scic
[2] http://es.net
[3] http://www.uslhcnt.org
[4] http://www.geant2.net
[5] http://monalisa.cern.ch/FDT
[6] http://supercomputing.caltech.edu
[7] http://www.ciena.com
[8] https://twiki.cern.ch/twiki/bin/view/LHCOPN/WebHome
[9] http://www.internet2.org
[10] http://www.nlr.net