NETWORKING FOR HIGH ENERGY AND NUCLEAR PHYSICS AS GLOBAL E-SCIENCE

Harvey B Newman, California Institute of Technology,
Pasadena, CA 91125, USA

On behalf of the ICFA Standing Committee on Inter-Regional Connectivity (SCIC)

Abstract

Wide area networks of sufficient, and rapidly increasing end-to-end capability are vital for every phase of high energy physicists' work. Our bandwidth usage, and the typical capacity of the major national backbones and intercontinental links used by our field have progressed by a factor of more than 1000 over the past decade, and the outlook is for a similar increase over the next decade, as we enter the era of LHC physics served by Grids on a global scale. Responding to these trends, and the emerging need to provide rapid access and distribution of Petabyte-scale datasets, physicists working with network engineers and computer scientists are learning to use networks effectively in the 1-10 Gigabit/range, placing them among the leading developers of global networks. I review the network requirements and usage trends, and present a bandwidth roadmap for HEP and other fields of "data intensive science". I give an overview of the status and outlook for the world's research networks, technology advances, and the problem of the Digital Divide, based on the recent work of ICFA's Standing Committee on Inter-regional Connectivity (SCIC)*.

INTRODUCTION

In an era of global collaborations, and data intensive Grids, advanced networks are required to interconnect the physics groups seamlessly, enabling them to collaborate throughout the lifecycle of their work. For the major experiments, networks that operate with reliable, quantifiable high performance and known characteristics are required to create Data Grids capable of processing and sharing massive physics datasets, rising from the Petabyte (10^{15} byte) to the Exabyte (10^{18} byte) scale within the next decade.

The need for global network-based systems that support our science has made the HENP community a leading early-adopter, and more recently a key co-developer of leading edge wide area networks. Over the past few years, several groups of physicists and engineers in our field, in North America, Europe and Asia, have worked with computer scientists to make significant advances in the development and optimization of network protocols, and methods of data transfer. During 2003-4 these developments, together with the availability of 2.5 and 10 Gigabit/sec wide area links and advances in data servers and their network interfaces (notably 10 Gigabit Ethernet) have made it possible for the first time to utilize networks with relatively high efficiency in the 1 to 10 Gigabit/sec (Gbps) speed range over continental and transoceanic distances.

HENP Network Traffic and Capacity Growth

The rate of growth in HENP network usage is reflected in the total accepted traffic of the U.S. Energy Science Network (ESNet). It grew by 70% per year between 1992 and 1999, and this accelerated to 100% per year (or 1000 times per decade) for the last 5 years [1]. SLAC's network traffic has been growing in steps since 1983 at an average annual rate of increase of 80%, and forward projection of this trend indicates that the traffic could reach 2 Terabits/sec by approximately 2014 [2].

These rates of expansion are paralleled by the growth of Internet traffic in the world at large. The traffic flowing through the Amsterdam Internet Exchange (AMS-IX [3]) for example, has been growing by a factor of 2 in each of the last few years, with growth spurts in the Summer and Fall. One of the most remarkable examples of growth is in China, where the number of Internet users increased from 6 to 78 million in the first half of 2004[#].

The growth of international network capacity on major links used by the HENP community is illustrated by the "LHCNet" link between the U.S. and CERN, which grew from a 9.6 kbps microwave and satellite link in 1985, to a 10 Gbps (OC-192) link today. This represents an increase in network capacity by a factor of 10^6 over the last 20 years, including a 5000-fold increase over the last decade.

R&E Network Growth and Transition to Next Generation Optical Infrastructures

These developments have been paralleled by upgrades in the national, and continental core network infrastructures, as well as the key transoceanic links used for research and education, to typical bandwidths in North America, Western Europe as well as Japan and Korea of 2.5 and now 10 Gbps [4]. The transition to the use of "dense wavelength division multiplexing" (DWDM) to support multiple optical links on a single fiber has made these links increasingly affordable, and this has resulted in a substantially increased number of these links coming into service.

---

* See the Feb. and Aug. 2004 Reports by the ICFA Standing Committee on Inter-Regional Connectivity (SCIC) at http://cern.ch/icfa-scic .

Source: J.P. Wu, APAN Conference, July 2004

2003-4 saw the emergence of some community-owned or leased wide area fiber infrastructures, managed by non-profit consortia of universities and regional network providers, to be used on behalf of research and education. This trend, pioneered by CA*net4 in Canada [5], is continuing and spreading to other regions.

A prime example is "National Lambda Rail" [6] (Figure 1) covering much of the US, which is now coming into service with many 10 Gbps links, in time for the SC2004 conference. The NLR infrastructure, built to accommodate up to 40 10 Gbps wavelengths in the future, is carrying wavelengths for Internet2's next-generation Hybrid Optical and Packet Infrastructure (HOPI) project [7], and for Caltech on behalf of HENP. This is accompanied by initiatives in 18 states (notably Illinois, California, and Florida). ESNet also is planning to use NLR wavelengths as part of the implementation of its next generation network.

Figure 1 National Lambda Rail (NLR)

GN2, the next generation pan-European backbone succeeding GEANT [8] is now being planned, and national initiatives are already underway in several European countries (notably in the Netherlands [9], Poland [10] and the Czech Republic [11]). In Japan, the SuperSINET 10 Gbps backbone [12] has been in operation since 2001, and a new 20 Gbps optical backbone JGN2 [13] came into service in April 2004. In Korea, the Kreonet 2.5-10 Gbps backbone is transitioning to Kreonet2, and is complemented by the SuperSIREN 10-40 Gbps optical research network [14].

These trends have also led to a forward-looking vision of much higher capacity networks based on many wavelengths in the future. The visions of advanced networks and Grid systems are beginning to converge, where future Grids will include end-to-end monitoring and tracking of networks as well as computing and storage resources [15], forming an integrated information system supporting data analysis, and more broadly research in many fields, on a global scale. The trend is also towards "hybrid" networks where the most demanding applications are matched to dynamically constructed optical paths to help ensure high end-to-end throughput. The development of this new class of hybrid networks is led by such projects as UltraLight [16], Translight [17], Netherlight [18], and UKLight [19].

Scientists and advanced network providers around the world have joined together to form GLIF [20], "a world scale lambda-based lab for application and middleware development, where Grid applications ride on dynamically configured networks based on optical wavelengths ... coexisting with more traditional packet-switched network traffic". The GLIF World Map (Figure 2) shows the proliferation of (typically) 10 Gbps links for research and education across the Atlantic and Pacific oceans. Notable recent additions include the GLORIAD project [21] linking Russia, China, Korea and Japan to the US and Europe, the planned Aarnet links [22] between Australia and the US, and the WHREN links [23] planned for 2005 between Latin America and the US.

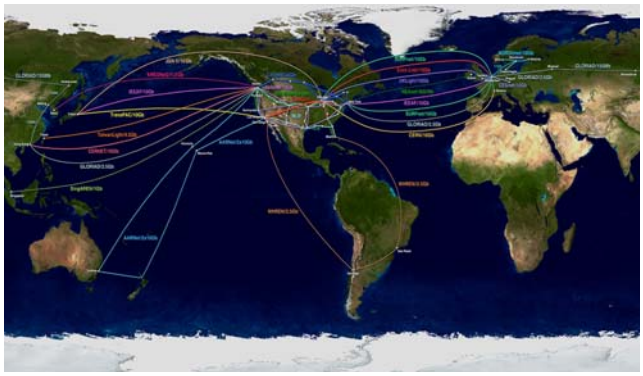

Figure 2 GLIF, the Global Lambda Integrated Facility

In some cases high energy physics laboratories or computer centers have been able to acquire leased "dark fiber" to their site, where they are able to connect to the principal wide area networks they use with one or more wavelengths. The link between Fermilab and Starlight, carrying 10 Gbps as of September 2004, is one example. Most recently DOE's Pacific Northwest Lab (PNNL) acquired 561 route-miles of dark fiber from Fiberco [24] to connect to UltraScience Net [25], Internet2's Abilene backbone [26] and other research networks[§].

Advances in End-to-End Network Performance

HENP groups working with computer scientists and network engineers, notably those at Caltech, CERN, SLAC, Manchester and Amsterdam, have made rapid advances in achieving efficient use of transcontinental and transoceanic network links over the last few years. The first single stream data transfer using TCP at a rate exceeding 100 Mbps took place between Caltech and CERN (11,000 km) in September 2001. We are now able to transfer data (memory to memory) at speeds up to 7.5 Gbps in a single standard TCP stream [27] over 15,700 km using 10 Gigabit Ethernet (10 GbE) network interfaces. 7.4 Gbps was achieved using FAST [28], an

[§] Source: Bill St. Arnaud, CA*net4 Newsletter, October 15, 2004.

advanced TCP protocol stack that quickly reaches equilibrium after packet losses and in the presence of congestion, while maintaining fair-sharing among multiple network streams.

HENP's progress is reflected in the Internet2 Land Speed Records [29], which are judged on the basis of the product of the sustained data transfer speed and the length of the network path, using standard TCP protocols over a production network. The Caltech-CERN team has compiled 7 such records since 2002 (Figure 3) and has reached more than 100 Petabit-meters per second with both Linux and Windows (memory to memory). Disk to disk transfers also are progressing rapidly, and 536 Mbytes/sec over 15,700 km disk to disk was recently achieved.

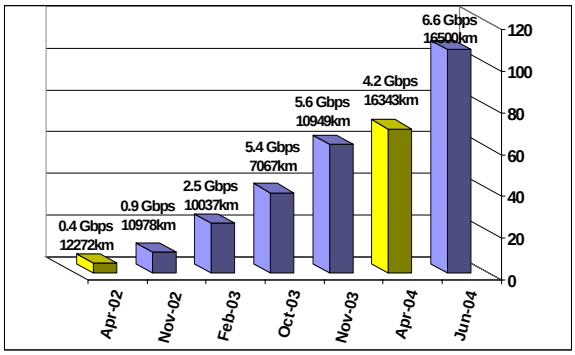

Figure 3 The IPv4 Single Stream Internet2 Land Speed Records. 5 of the 7 records shown are to HENP.

HENP Network Roadmap

The trends and developments summarized have led to an increasingly dynamic "system" view of the network (and the Grid system built on top of it), where physicists at remote locations could extract Terabyte-sized subsets of the data drawn from multi-petabyte data stores on demand, and if needed rapidly deliver this data to their home-sites.

Delivering this data in a short "transaction" lasting minutes, rather than hours, would enable remote computing resources to be used more effectively, while making physics groups remote from the experiment better able to carry out competitive data analyses. Such short transactions also are necessary to avoid the bottlenecks and fragility of the Grid system that would result if hundreds to thousands of such requests were left pending for long periods, or if a large backlog of requests was permitted to build up over time.

It is important to note that transactions on this scale, while still representing very small fractions of the data, correspond to throughputs across networks of 10 Gbps and up. A 1000 second-long transaction shipping 1 TByte of data corresponds to 8 Gbps of net throughput. Larger transactions, in the 10-100 TByte range, could be foreseen within the next decade.

These considerations led to a roadmap for HENP networks in the coming decade, shown in Table 1. Using the US-CERN production and research network links[**] as an example of the possible evolution of major network links in our field, the roadmap[††] shows progressive upgrades every 2-3 years, going from the present 10 Gbps range to the Terabit/sec (Tbps) range within approximately the next 10 years. The column on the right shows the progression towards the use of multiple wavelengths on an optical fiber, and the increasingly dynamic provision of end-to-end network paths through optical circuit switching.

It should be noted that the roadmap in Table 8 is a "middle of the road" projection. The rates of increase fall in between the rate experienced between 1985 and 1995, before deregulation of the telecommunications industry, and the current decade.

Year	*Production*	*Experimental*	*Remarks*
2001	**0.155**	**0.622-2.5**	**SONET/SDH**
2002	**0.622**	**2.5**	**DWDM; GbE Integ.**
2003	**2.5**	**10**	**DWDM; 1+10GbE**
2005	**10**	**2-4 X 10**	**λ Switch; λ Provisioning**
2007	**2-4 X 10**	**~10 X 10; 40 Gbps**	**1st Gen. λ Grids**
2009	**~10 X 10 or 1-2 X 40**	**~5 X 40 or ~20-50 X 10**	**40 Gbps λ Switching**
2011	**~5 X 40 or ~20 X 10**	**~25 X 40 or ~100 X 10**	**2nd Gen λ Grids Tbit Networks**
2013	**~Terabit**	**~MultiTbps**	**~Fill One Fiber**

Table 1. A Roadmap for major links used by HENP. Projections follow the trend of affordable bandwidth increases over the last 20 years: by a factor of ~500-1000 per decade.

As shown in Table 2, taken from DOE's High Performance Network Workshop [30], the Roadmap shown above is compatible with the outlook for many other fields of "data intensive science". While the applications for each field vary, each of them foresees network needs in the Terabit/sec range within the next 5-10 years.

[**] Jointly funded by the US DOE and NSF, CERN and the European Union.

[††] Source: H. Newman. Also see "Computing and Data Analysis for Future HEP Experiments" presented by M. Kasemann at the ICHEP02 Conference, Amsterdam (7/02).
See http://www.ichep02.nl/

Science	Today End2End Thruput	5 years Thruput	5-10 Years Thruput	Remarks
HEP	0.5 Gbps	100 Gbps	1000 Gbps	Bulk thruput
Climate	0.5 Gbps	160-200 Gbps	N x 1000 Gbps	Bulk thruput
SNS Nano-Science	Not yet started	1 Gbps	1000 Gb/s + QoS for Control Channel	Remote control + time critical throughput
Fusion Energy	500 MB/sec burst	500MB/ 20 sec. burst	N x 1000 Gbps	Time critical throughput
Astrophys.	1 TB/wk	N*N multicast	1000 Gbps	Comp. steering + collaboration
Genomics	1 TB/day	100s of Users	1000 Gbps + QoS for Control Channel	High Throughput and steering

Table 2 Roadmap for fields of data intensive science, from the DOE High Performance Network Workshop

ICFA Standing Committee on Inter-Regional Connectivity (SCIC)

ICFA's involvement in network issues began in 1996, when it issued the following Statement [31]:

ICFA urges that all countries and institutions wishing to participate even more effectively and fully in international high energy physics collaborations should:

- *review their operating methods to ensure that they are fully adapted to remote participation*
- *strive to provide the necessary communication facilities and adequate international bandwidth*

ICFA commissioned the SCIC in 1998 with the charge to:

- Track network progress and plans, and connections to the major HENP institutes and universities in countries around the world;
- Monitor network traffic, and end-to-end performance in different world regions
- Keep track of network technology developments and trends, and use these to "enlighten" network planning for our field
- Focus on major problems related to networking in the HENP community; determine ways to mitigate or overcome these problems; bring these problems to the attention of ICFA, particularly in areas where ICFA can help.

Three SCIC working groups were formed in 2002:

- ***Monitoring***, Chaired by Les Cottrell of SLAC
- ***Advanced Technologies,*** Chaired by Richard Hughes-Jones (Manchester) and Olivier Martin (CERN)
- ***The Digital Divide,*** Chaired by Alberto Santoro (UERJ, Rio de Janeiro)

As documented at http://cern.ch/icfa-scic, the SCIC has reported annually to ICFA and has been actively engaged in working to close the Digital Divide separating the HEP groups in the most technologically and economically favored regions from the rest of the world. The Monitoring and Digital Divide working groups have been active in bilateral and multilateral projects to assist groups in Southeast Europe, Asia, Latin America and the Middle East, and to encourage scientific development in Africa. The SCIC also has been very active in the World Summit on the Information Society [32], where CERN and Caltech hosted a Forum and an Online Stand hosting real-time demonstrations of advanced networking, remote collaboration, and education and outreach.

Some of the main conclusions of SCIC's 2004 report to ICFA are summarized below:

A key issue for our field is to close the Digital Divide in HENP, so that scientists from all regions of the world have access to high performance networks and associated technologies that will allow them to collaborate as full partners: in experiment, theory and accelerator development. While the throughput obtainable on networks, monitored by SLACs Internet2 End-to-end Performance Monitoring (IEPM) project [33] is improving year by year in all world regions, Central Asia, Russia, Southeast Europe, Latin America, the Middle East and China remain 4-5 years behind the US, Canada, Japan and most of Europe. India and Africa are 7-8 years behind, and are in danger of falling farther behind.

The rate of progress in the major networks has been faster than foreseen (even 1 to 2 years ago). The current generation of network backbones, representing an upgrade in bandwidth by factors ranging from 4 to more than several hundred in some countries, arrived in the last two years in the US, Europe and Japan. This rate of improvement is faster, and in some cases many times the rate of Moore's Law[‡‡]. This rapid rate of progress, confined mostly to the US, Europe, Japan and Korea, as well as the major transoceanic routes, threatens to open the Digital Divide further, unless we take action.

Reliable high End-to-end Performance of networked applications such as large file transfers and Data Grids is required. Achieving this requires:

- ***End-to-end monitoring extending to all regions serving our community.*** A coherent approach to

[‡‡] Usually quoted as a factor of 2 improvement in performance at the same cost every 18 months.

monitoring that allows physicists throughout our community to extract clear, unambiguous and inclusive information is a prerequisite for this.

- ***Upgrading campus infrastructures. While National and International backbones have reached 2.5 to 10 Gbps speeds in many countries, campus network infrastructures are still not designed to support Gbps data transfers in most HEP centers.*** A reason for the underutilization of National and International backbones is the lack of bandwidth to groups of end users inside the campus.
- ***Removing local, last mile, and national and international bottlenecks end-to-end, whether the bottlenecks are technical or political in origin.*** Many HEP laboratories and universities situated in countries with excellent network backbones are not well-connected, due to limited access bandwidth to the backbone, or the bandwidth provided by their metropolitan or regional network, or through the lack of peering arrangements between the networks with sufficient bandwidth. This problem is very widespread in our community, with examples stretching from China to South America to the northeast region of the U.S.. Root causes vary from lack of local infrastructure to unfavorable pricing policies.
- ***Removing Firewall bottlenecks.*** Firewall systems are so far behind the needs that they won't match the data flow of Grid applications. The maximum throughput measured across available products is typically limited to a few hundred Mbps. It is urgent to address this issue by designing new architectures that eliminate/alleviate the need for conventional firewalls. For example, Point-to-point provisioned high-speed circuits as proposed by emerging Light Path technologies could remove the bottleneck. With endpoint authentication, the point-to-point paths are private and intrusion resistant circuits, so they should be able to bypass site firewalls if the endpoints (sites) trust each other.
- ***Developing and deploying high performance (TCP) toolkits in a form that is suitable for widespread use by users.*** Training the community to use these tools well and wisely also is required.

ICFA Digital Divide Workshops

In 2003 the SCIC proposed to ICFA, given the importance of the Digital Divide problem to our field, that this and related issues be brought to our community for discussion. This led to ICFA's approval of the First Digital Divide and HEP Grid Workshop [34] that took place at UERJ in Rio de Janeiro in February 2004 with the theme: *Global Collaborations, Grids and Their Relationship to the Digital Divide.* A central focus of the workshop was Digital Divide issues in Latin America, and their relation to problems in other world regions.

The second meeting in this series, the International ICFA Workshop on HEP Networking, Grids and Digital Divide Issues for Global e-Science will take place at Kyungpook National University in Daegu, Korea in May 2005. The workshop goals are to:

- Review the current status, progress and barriers to effective use of major national, continental and transoceanic networks used by HEP
- Review progress, strengthen opportunities for collaboration, and explore the means to deal with key issues in Grid computing and Grid-enabled data analysis, for high energy physics and other fields of data intensive science, now and in the future
- Exchange information and ideas, and formulate plans to develop solutions to specific problems related to the Digital Divide in various regions, with a focus on Asia Pacific, as well as Latin America, Russia and Africa
- Continue to advance a broad program of work on reducing or eliminating the Digital Divide, and ensuring global collaboration, as related to all of the above aspects.

ACKNOWLEDGEMENTS

I am deeply indebted to the members of the SCIC, and particularly the Working Group Chairs as well as Sylvain Ravot (Caltech) and Danny Davids (CERN) for the content of this report. The help and support of ICFA and the leaders of the world's research and education network organization is gratefully acknowledged, with particular thanks to Jonathan Dorfan (ICFA), Doug Van Houweling (Internet2), Tom West (NLR) and Tom de Fanti (UIC; Starlight). The perspectives developed in this report rely on discussions with many individuals over the years, notably Richard Mount (SLAC), David O. Williams and Olivier Martin (CERN), Vicky White (FNAL), Paul Avery (Florida), Bill St. Arnaud (CANARIE), Iosif Legrand (Caltech), Shawn McKee (Michigan), Guy Almes (NSF) and Bill Johnston (LBNL). This work would not have been possible without the strong support of the U.S. Department of Energy and the National Science Foundation.

REFERENCES

[1] Source: W. Johnston (LBNL), ESnet Manager. .
[2] Source: R. Cottrell (SLAC), Assistant Director of Computing Services.
[3]See www.ams-ix.net/about/stats/index.html Peak daily traffic through AMS-IX is currently 36 Gbps.
[4] This is documented in Appendices to the February 2004 ICFA SCIC report (http://cern.ch/icfa-scic) covering many of the major national and international networks and network R&D projects.
[5] See the CA*net4 map at www.canarie.ca/canet4/
[6] See www.nlr.net
[7] See http://networks.internet2.edu/hopi/
[8] See www.dante.net/geant

[9] See http://www.surfnet.nl/info/en/home.jsp
[10] See http://www.pionier.gov.pl
[11] See http://www.ces.net/
[12] See http://www.sinet.ad.jp/english/super_sinet.html
[13] See http://www.jgn.nict.go.jp/
[14] See http://www.kreonet2.net/ and presentations at the 3rd HEP Data Grid Workshop, Kyungpook National University, at http://chep.knu.ac.kr/HEPDGWS04/
[15] The MonALISA network and Grid monitoring and management system, now running at 160 sites, is a prime example. See http://monalisa.caltech.edu and the talk by I. Legrand *et al.* at this conference.
[16] See http://ultralight.caltech.edu
[17] See www.startap.net/translight/
[18] See www.surfnet.nl/info/innovatie/netherlight.jsp
[19] See http://www.uklight.ac.uk/
[20] See www.glif.is The fourth GLIF workshop took place in Nottingham, UK in September 2004.
[21] See www.gloriad.org
[22] See www.aarnet.edu.au/news/sxtransport.html
[23] See www.ampath.fiu.edu/
[24] See www.fiberco.org/
[25] See www.csm.ornl.gov/ultranet/
[26] See http://abilene.internet2.edu/
[27] See the talk by S. Ravot *et al.* at this conference. 7.5 Gbps throughput has been obtained with 10 GbE interfaces from S2io (www.s2io.com)
[28] FAST has been developed by S. Low who heads Caltech's Netlab. See http://netlab.caltech.edu/FAST
[29] See http://lsr.internet2.edu/
[30] See www.doecollaboratory.org/meetings/hpnpw/
[31] See www.fnal.gov/directorate/icfa/icfa_communicaes.html
[32] See http://www.itu.int/wsis/
[33] See www-iepm.slac.stanford.edu/; and R. Cottrell, www.slac.stanford.edu/grp/scs/net/talk03/icfa-aug04.ppt
[34] See www.lishep.uerj.br/