Published in: "Trends in High Performance & Large Scale Computing" Lucio Grandinetti and Gerhard Joubert editors, IOS Press publisher

The Evolution of Research and Education Networks and their Essential Role in Modern Science

William JOHNSTON, Evangelos CHANIOTAKIS, Eli DART, Chin GUOK, Joe METZGER, Brian TIERNEY
ESnet, Lawrence Berkeley National Laboratory[†]

Abstract. ESnet – the Energy Sciences Network – has the mission of enabling the aspects of the US Department of Energy's Office of Science programs and facilities that depend on large collaborations and large-scale data sharing to accomplish their science. The Office of Science supports a large fraction of all U.S. physical science research and operates many large science instruments and supercomputers that are used by both DOE and University researchers. The network requirements of this community have been explored in some detail by ESnet and a long-term plan has been developed in order to ensure adequate networking to support the science. In this paper we describe the planning process (which has been in place for several years and was the basis of a new network that is just now being completed and a new set of network services) and examine the effectiveness and adequacy of the planning process in the light of evolving science requirements.

Keywords. Energy Sciences Network (ESnet), networks for large-scale science, network planning.

1. Background

The US Department of Energy's Office of Science ("SC") is the single largest supporter of basic research in the physical sciences in the United States providing more than 40 percent of total funding for this area.

SC manages its research portfolio through program offices, each of which has a specific set of goals:

[†] This work was supported by the Director, Office of Science, Office of Advanced Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

- Advanced Scientific Computing Research
  - Deliver Computing for the Frontiers of Science
- Basic Energy Sciences
  - Advance the Basic Sciences for Energy Independence
- Biological and Environmental Research
  - Harness the Power of Our Living World
- Fusion Energy Sciences
  - Bring the Power of the Stars to Earth
- High Energy Physics
  - Explore the Fundamental Interactions of Energy, Matter, Time, and Space
- Nuclear Physics
  - Explore Nuclear Matter – from Quarks to Stars

The Office of Science supports researchers at more than 300 colleges and universities across the United States. SC balances its support for big science and interdisciplinary teams with investments in basic research projects conducted by leading university and laboratory investigators. The DOE FY2006 university research funding exceeded $US 800 million.

The National Laboratories are the heart of SC's science programs. The DOE national laboratory system is the most comprehensive research system of its kind in the world – and the backbone of American science. The Office of Science is the steward of 10 of these 17 laboratories with world-class capabilities for solving complex interdisciplinary scientific problems. The 10 DOE Office of Science National Laboratories are:

- Ames Laboratory
- Argonne National Laboratory (ANL)
- Brookhaven National Laboratory (BNL)
- Fermi National Accelerator Laboratory (FNAL)
- Thomas Jefferson National Accelerator Facility (JLab)
- Lawrence Berkeley National Laboratory (LBNL)
- Oak Ridge National Laboratory (ORNL)
- Pacific Northwest National Laboratory (PNNL)
- Princeton Plasma Physics Laboratory (PPPL)
- SLAC National Accelerator Laboratory (SLAC)

At the Scientific User Facilities the Office of Science builds and operates some of the world's most important scientific facilities and instruments that researchers depend on to extend the frontiers of science. The Office of Science facilities include particle accelerators, synchrotron light sources, neutron scattering facilities, nanoscale science research centers, supercomputers, high-speed networks, and genome sequencing facilities.

In the 2007 fiscal year, these Office of Science facilities were used by more than 21,000 researchers and students from universities, national laboratories, private industry, and other federal science agencies.

(The material in this section was adapted from Office of Science documents at http://science.doe.gov/about/SCBrochure (September_2008).pdf and http://science.doe.gov/Research_Universities/index.htm)

2. DOE Office of Science and ESnet – ESnet and its Mission

ESnet is an Office of Science facility in the Office of Advanced Scientific Computing Research ("ASCR"). ESnet's primary mission is to enable the science goals of the Office of Science (SC) and that depend on:

- Sharing of massive amounts of data
- Thousands of collaborators world-wide
- Distributed data processing
- Distributed data management
- Distributed simulation, visualization, and computational steering

In order to accomplish its mission ESnet provides high-speed networking and various collaboration services to Office of Science laboratories as well as for many other DOE programs (on a cost recovery basis). The goal is to ensure that they have robust communication with their science collaborators in the U.S. and world-wide.

To this end ESnet builds and operates a high-speed national network specialized to serving tens of thousands of Department of Energy scientists, support staff, and collaborators worldwide. This network provides high-bandwidth, reliable connections from the DOE science community and to all of the world's major research and education (R&E) networks that serve U.S. and international R&D institutions. This enables researchers at national laboratories, universities and other institutions to communicate with each other to accomplish the collaboration needed to address some of the world's most important scientific challenges.

The ESnet architecture and capacity are driven by DOE's involvement in some of the world's largest science experiments. The Large Hadron Collider (LHC) comes online in 2009, resulting in an urgent demand for guaranteed very high-bandwidth connectivity (greater than 10 Gigabits (10,000 Megabits/sec) for huge data transfers. Such service is also identified as required by other Office of Science mission areas as a result of a formal requirements gathering process. This process has shown that similar needs exist for the climate community in the near term, for ITER (international fusion reactor experiment) when it comes online a decade from now, for the huge volumes of data generated in the rapidly evolving SC scientific supercomputing centers and their next generation numerical models (e.g. for climate), and a new generation of instruments such as those associated with the study of dark matter/dark energy cosmology. As DOE's large-scale science moves to a distributed national and international model, ESnet must provide the innovation and expertise to meet these networking needs.

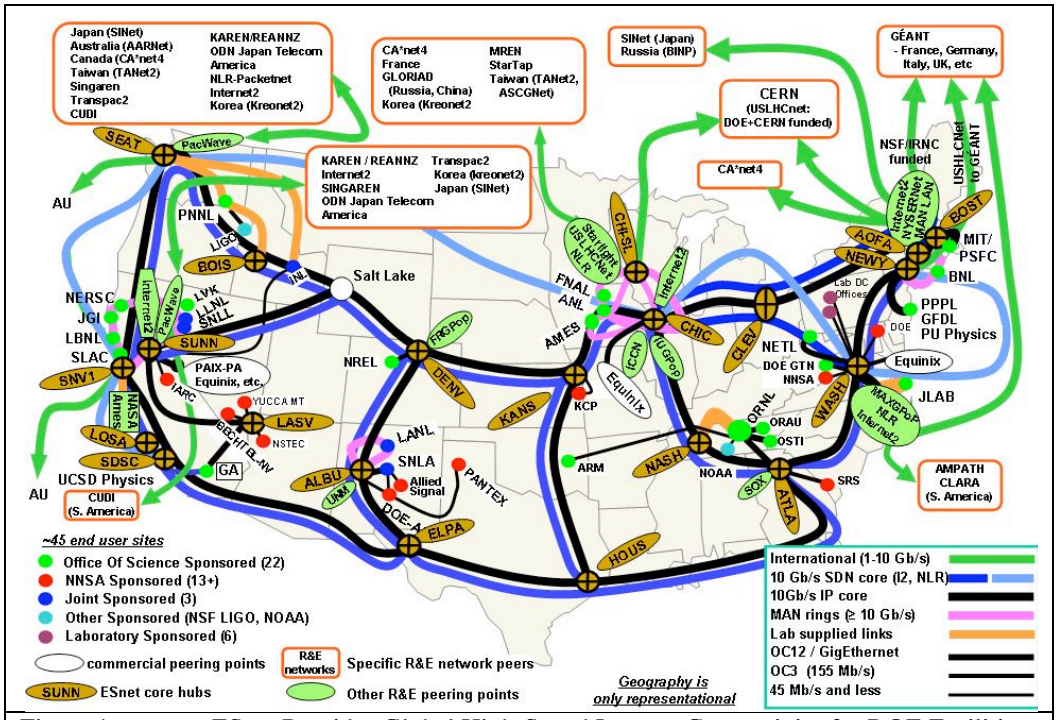


Figure 1. ESnet Provides Global High-Speed Internet Connectivity for DOE Facilities and Collaborators (12/2008). Much of the utility (and complexity) of ESnet is in its high degree of interconnectedness.

To meet specific Office of Science requirements, in 2008 ESnet completed the initial rollout of ESnet4 (Figure 1) which involved deploying a 30 Gb/s core network on a footprint that covers the country in six interconnected rings – this network will grow to 50Gb/s by 2010. The overall ESnet network consists of two core networks that address two different types of traffic. The IP core network carries all of the "commodity" IP traffic from the Labs, including much of the small and medium scale science traffic. The IP core connects to commercial and R&E peering points. The SDN core is primarily designed to carry the large data flows of science and connects primarily to similar high-speed R&E networks (e.g. Internet2's Dynamic Circuit Network and NLR's FrameNet). However, under some circumstances the IP core will carry SDN traffic and the SDN core will carry "commodity" IP traffic. One 10 Gb/s wave/optical circuit is used for the IP core and all of the rest of the waves/optical circuits, including the ones added over the next several years, will be incorporated into the SDN core.

In addition to high-speed connectivity with all of the US research and education (R&E) networks, ESnet4 provides connections to high-speed networks in Europe, Japan and SE Asia, as well as to more than 100 other R&E networks around the world. This ensures that DOE researchers can have high-speed connectivity with all collaborating national and international researchers. In addition to the considerable end-to-end bandwidth and the high reliability of ESnet's operational model, ESnet is actively involved with other domestic and international R&E networks in developing and deploying cutting-edge technologies and services needed to

produce a seamless interoperable infrastructure that will allow the advancement of DOE's large-scale science. ESnet services allow scientists to make effective use of unique DOE research facilities and computing resources, independent of time and geographic location.

ESnet provides highly reliability through redundancy in all critical network components and an engineering staff that is situated in four locations across the country.

ESnet has several types of stakeholders and each has a somewhat different role in setting ESnet priorities:

SC/ASCR provides input to ESnet in the form of high-level strategic guidance through the budgeting process; indirect long term input is through the process of ESnet observing and projecting network utilization of its large-scale users; and direct long term input is through the SC Program Offices Requirements Workshops (described below).

The SC Labs provide extensive engineering input to ESnet with short term input coming through many daily (mostly) email interactions and long term input coming through bi-annual ESnet Coordinating Committee meetings (ESCC consists of all of the Lab network principals).

SC science collaborators are those SC supported or SC facility using scientists who are not at the DOE Labs. This large community provides input to ESnet through numerous community meetings that ESnet participates in. This community is represented to ESnet primarily by the R&E network organizations (such as Internet2, DANTE/GÉANT, the U.S. regional networks, the European NRENS, etc.) that serve this science collaboration community.

3. Strategic Approach to Accomplishing ESnet's Mission

ESnet has a three-pronged approach to planning for meeting its mission:

I. Determine the ESnet requirements by a) working with the SC community to identify the networking implication of the instruments, supercomputers, and the evolving process of how science is done, and b) considering the historical trends in the use of the network.

II. 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.

III. Anticipate future network capabilities that will meet future science requirements with an active program of research and advanced development.

The remainder of section 3 is structured as follows:

3.1 Strategy I: ESnet requirements from examining SC plans for future instruments and science and from historical trends

- 3.1.1 Examine SC plans for future instruments and science
- 3.1.2 Requirements from observing current and historical network traffic patterns

3.2 Strategy II: Build a network environment that will enable the distributed aspects of SC science and then continuously reassess and update the approach

- 3.2.1 Planning process 1: Provide the basic, long-term bandwidth
- 3.2.2. Planning process 2: Undertake continuous reexamination of the long-term
- 3.2.3 Planning process 3: Identify required new network services and implement them

3.2.4. Planning process 4: Develop outreach approaches that identify what communities are not being well served by the network

3.3 Strategy III. Anticipate future network capabilities that will meet future science requirements

3.1. Strategy I: Requirements from examining SC plans for future instruments and science and from historical trends in the network

Science requirements are determined by considering two aspects of the origins of network traffic. First, by exploring the plans and processes of the major stakeholders – the data characteristics of scientific instruments and facilities (what data will be generated by instruments and supercomputers coming on-line over the next 5-10 years?) and by examining the future process of science (how and where will the new data be analyzed and used, what sorts of distributed systems are involved, etc. – that is, how will the process of doing science change over 5-10 years?). And second, by observing current and historical network traffic patterns – what do the trends in network patterns predict for future network needs?

3.1.1. Examine SC plans for future instruments and science

The primary mechanism for determining science community plans is through the Office of Science Network Requirements Workshops. These workshops are organized by the SC Program Offices and are conducted at the rate of two requirements workshops per year, repeating starting in 2010. That is:

- Basic Energy Sciences (2007 – published)
- Biological and Environmental Research (2007 – published)
- Fusion Energy Science (2008 – published)
- Nuclear Physics (2008 – published)
- IPCC (Intergovernmental Panel on Climate Change) special requirements (BER) (August, 2008)
- Advanced Scientific Computing Research (Spring 2009)
- High Energy Physics (Summer 2009)

Additionally several earlier workshops (2002/3 and 2006) looked at specific HEP facilities and ASCR supercomputer centers[‡].

The Workshops have now examined a fair cross section of the Country's major science facilities operated by DOE's Office of Science

Basic Energy Sciences supports fundamental research in materials sciences and engineering, chemistry, geosciences, and molecular biosciences. The BES program also supports world-class scientific user facilities, including four synchrotron radiation light sources, three neutron scattering facilities, and four electron-beam micro characterization centers. Annually, 8,000 researchers from academia, industry, and Federal laboratories perform experiments at these facilities.

[‡] The workshop reports are available at www.es.net/hypertext/requirements.html

BES facilities include the Advanced Photon Source at ANL, the Advanced Light Source at LBNL, the Combustion Research Facility at Sandia National Lab, the Linac Coherent Light Source at SLAC, the Center for Functional Nanomaterials at BNL, the Spallation Neutron Source (a $1.4 next-generation neutron-scattering facility) at ORNL, the Molecular Foundry at LBNL, the National Center for Electron Microscopy at LBNL, and the National Synchrotron Light Source at BNL. Requirements gathering also included input from the computational chemistry community.

Biological and Environmental Research supports fundamental research in climate change, environmental remediation, genomics, systems biology, radiation biology, and medical sciences. BER funds research at public and private research institutions and at Department of Energy (DOE) laboratories. BER supports leading edge research facilities used by public and private sector scientists across range of disciplines: structural biology, DNA sequencing, functional genomics, climate science, the global carbon cycle, and environmental molecular science.

BER Facilities include the Atmospheric Radiation (solar) Measurement (ARM) Program and the ARM Climate Research Facility (ACRF), Bioinformatics and Life Sciences Programs, Climate Sciences Programs (Large Simulations and Collaborative Tools), the Environmental Molecular Sciences Laboratory at PNNL, the Joint Genome Institute (JGI).The National Center for Atmospheric Research (NCAR) also participated in the workshop and contributed a section to this report due to the fact that a large distributed data repository for climate data will be established at NERSC (SC supercomputer center), ORNL and NCAR.

Fusion Energy Sciences supports advances in plasma science, fusion science, and fusion technology—the knowledge base needed for an economically and environmentally attractive fusion energy source. FES is pursuing this goal through an integrated program of research based in U.S. universities, industry, and national laboratories, augmented by a broad program of international collaboration.

FE facilities include Tokamaks (fusion devices) at General Atomics and MIT, the National Spherical Torus at PPPL, the Fusion Simulation Project, and major collaborating sites including the EAST Tokamak in China and the KSTAR Tokamak in South Korea, and ITER in France.

Nuclear Physics provides most of the Federal support for nuclear physics research, delivering new insights into our knowledge of the properties and interactions of atomic nuclei and nuclear matter.

NP facilities include the Relativistic Heavy Ion Collider (RHIC) at Brookhaven National Laboratory is used by almost 1000 physicists from around the world to study what the universe may have looked like in the first few moments after its creation. JLab's Continuous Electron Beam Accelerator Facility (CEBAF) is the first large-scale application of superconducting electron-accelerating technology and is used by approximately 1,200 scientists from around the world. Other facilities include the Argonne Tandem Linac Accelerator System (ATLAS) at ANL, the Holifield Radioactive Ion Beam Facility (HRIBF) at ORNL and the Spallation Neutron Source (SNS) at ORNL. LBNL's 88-Inch Cyclotron is being supported to test electronic circuit components for radiation "hardness" to cosmic rays by the National Reconnaissance Office (NRO) and U.S. Air Force (USAF). University-based research is

supported by accelerator operations at Texas A&M University (TAMU), the Triangle Universities Nuclear Laboratory (TUNL) at Duke University, at Yale University, and at the University of Washington.

Advanced Scientific Computing Research (ASCR) has as its mission to underpin DOE's world leadership in scientific computation by supporting research in applied mathematics, computer science, and high-performance networks and providing the high-performance computational and networking resources that are required for world leadership in science.

ASCR major resources include NERSC supercomputer center at LBNL, NLCF supercomputer center at ORNL, ACLF supercomputer center at ANL, and ESnet – the Energy Sciences Network

High Energy Physics provides most of the Federal support for research in high energy physics, which seeks to understand the fundamental nature of matter, energy, space, and time.

HEP major facilities include The Tevatron collider at Fermi National Accelerator Laboratory (FNAL) and the Large Hadron Collider (LHC) at CERN, in which the U.S. has a fundamental role.

Requirements from the Workshops

Table 1 summarizes the quantitative and functional requirements that have been determined from the Workshops to date. The information represents a mix of the characteristics of the instruments / systems involved and the process of the associated science – that is, how the instrument / system is used, by whom, the size of the collaborations, the locations of the collaborators, etc. The reflected bandwidths are generally aggregate bandwidths, so they lack the necessary level of detail to actually build out the network infrastructure to get to specific locations. This issue is addressed below.

Table 1. Quantitative and functional requirements that have been determined from the requirements workshops to date.

Science Drivers **Science Areas / Facilities**	**End2End Reliability**	**Near Term End2End Band width**	**5 years End2End Band width**	**Traffic Characteristics**	**Network Services**
ASCR: ALCF supercomputer	-	10Gb/s	30Gb/s	• Bulk data • Remote control • Remote file system sharing	• Guaranteed bandwidth • Deadline scheduling • PKI / Grid
ASCR: NERSC supercomputer	-	10Gb/s	20 to 40 Gb/s	• Bulk data • Remote control • Remote file system sharing	• Guaranteed bandwidth • Deadline scheduling • PKI / Grid
ASCR: NLCF supercomputer	-	Backbone Bandwidth Parity	Backbone Bandwidth Parity	• Bulk data • Remote control • Remote file system sharing	• Guaranteed bandwidth • Deadline scheduling • PKI / Grid
BER: Climate	-	3Gb/s	10 to 20Gb/s	• Bulk data • Rapid movement of GB sized files • Remote Visualization	• Collaboration services • Guaranteed bandwidth • PKI / Grid

Science Drivers Science Areas / Facilities	End2End Reliability	Near Term End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
BER: EMSL/Bio	-	10Gb/s	50-100Gb/s	• Bulk data • Real-time video • Remote control	• Collaborative services • Guaranteed bandwidth
BER: JGI/Genomics	-	1Gb/s	2-5Gb/s	• Bulk data	• Dedicated virtual circuits • Guaranteed bandwidth
BES: Chemistry and Combustion	-	5-10Gb/s	30Gb/s	• Bulk data • Real time data streaming	• Data movement middleware
BES: Light Sources	-	15Gb/s	40-60Gb/s	• Bulk data • Coupled simulation and experiment	• Collaboration services • Data transfer facilities • Grid / PKI • Guaranteed bandwidth
BES: Nanoscience Centers	-	3-5Gb/s	30Gb/s	• Bulk data • Real time data streaming • Remote control	• Collaboration services • Grid / PKI
FES: International Collaborations	-	100Mbps	1Gb/s	• Bulk data	• Enhanced collaboration services • Grid / PKI • Monitoring / test tools
FES: Instruments and Facilities	-	3Gb/s	20Gb/s	• Bulk data • Coupled simulation and experiment • Remote control	• Enhanced collaboration service • Grid / PKI
FES: Simulation	-	10Gb/s	88Gb/s	• Bulk data • Coupled simulation and experiment • Remote control	• Easy movement of large checkpoint files • Guaranteed bandwidth • Reliable data transfer
HEP: LHC (CMS and Atlas)	99.95+% (Less than 4 hours per year)	73Gb/s	225-265Gb/s	• Bulk data • Coupled analysis workflows	• Collaboration services • Grid / PKI • Guaranteed bandwidth • Monitoring / test tools
NP: CMS Heavy Ion	-	10Gb/s (2009)	20Gb/s	• Bulk data	• Collaboration services • Deadline scheduling • Grid / PKI
NP: CEBF (JLAB)	-	10Gb/s	10Gb/s	• Bulk data	• Collaboration services • Grid / PKI
NP: RHIC	Limited outage duration to avoid analysis pipeline stalls	6Gb/s	20Gb/s	• Bulk data	• Collaboration services • Grid / PKI • Guaranteed bandwidth • Monitoring / test tools

The last four entries in the list for High Energy Physics and Nuclear Physics, HEP in particular, provided many of the specific requirements in 2004-5-6 that were the basis of the detailed design of ESnet4.

Requirements for services emerge from detailed discussions about how the analysis and simulation systems actually work: where the data originates, how many systems are involved in the analysis, how much data flows among these systems, how complex is the work flow, what are the time sensitivities, etc.

The result of this process is that fairly consistent requirements are found across the large-scale sciences. Large-scale science uses distributed systems in order to:

- couple existing pockets of code, data, and expertise into "systems of systems"
- break up the task of massive data analysis into elements that are physically located where the data, compute, and storage resources are located

Such systems

- are data intensive and high-performance, typically moving terabytes a day for months at a time
- are high duty-cycle, operating most of the day for months at a time in order to meet the requirements for data movement
- are widely distributed – typically spread over continental or inter-continental distances
- depend on network performance and availability, but these characteristics cannot be taken for granted, even in well run networks, when the multi-domain network path is considered

The nodes of the distributed systems must be able to get guarantees from the network that there is adequate bandwidth necessary to accomplish the task at hand. The systems must also be able to get information from the network to support graceful failure and auto-recovery and adaptation due to unexpected network conditions that are short of outright failure.

General Network Technology and Capabilities Requirements from Instruments and Facilities

Considering the overall requirements from the Workshops one can identify a generic, but important, set of goals for any network and network services implementation:

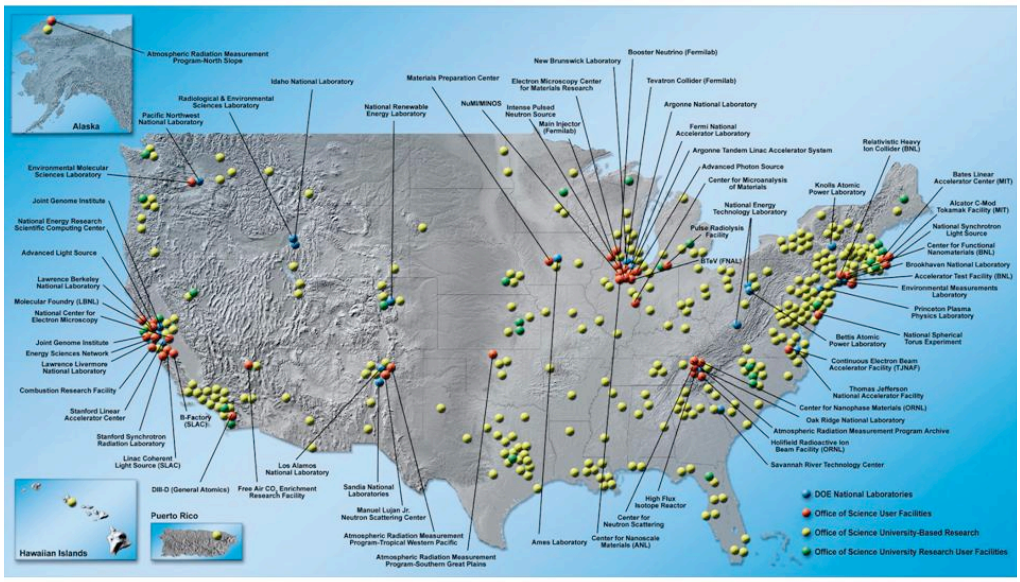
- Bandwidth: Adequate network capacity to ensure timely movement of data produced by the facilities
- High reliability is required for large instruments and "systems of system" (large distributed systems) which now depend on the network to accomplish their science
- Connectivity: The network must have the geographic reach – either directly or through peering arrangements with other networks – sufficient to connect users and collaborators and analysis systems to SC facilities
- Services that provide guaranteed bandwidth, traffic isolation, end-to-end monitoring, etc., are required and these services must be presented to the users in a framework of Web Services / SOA (service oriented architecture) / Grid / "Systems of Systems" that are the programming paradigms of modern science.

As an aside it is worth noting that at present, ESnet traffic is dominated by data flows from large instruments – LHC, RHIC, Tevatron, etc. Supercomputer traffic is currently a small part

of ESnet's total traffic, though it has the potential to increase dramatically over the next 5 years, and could end up being the biggest use of the network. However this will not happen until appropriate system architectures are in place to allow high-speed communication from the supercomputers and their associated storage systems over the wide area network to other supercomputers and storage systems. This is not currently the case; however there are R&D efforts at both DOE and NSF to address this issue.

Connectivity

Much of the design of the ESnet network footprint is intended to accommodate the high degree of connectivity that is needed to meet the requirements of the science collaborations. The footprint must not only accommodate connecting the approximately 50 ESnet sites, but also ensure that high-speed connections can be made to all of the major U.S. and international R&E networks. Although much of this is accomplished by ensuring a presence at the half dozen or so R&E exchange points (MAN LAN in New York, MAX in Washington, DC, Starlight in Chicago, PNWGigaPoP in Seattle, etc.) even at those locations there may be several local points of presence that must be accommodated. The major U.S. and international locations associated with SC programs are illustrated in the next two figures.



DOE National Laboratories, Office of Science User Facilities, Office of Science University-Based Research, and Office of science University Research User Facilities

(http://science.doe.gov/News_Information/News_Room/2006/University_support/University_Support.pd

Figure 2. Requirements for connectivity: The U.S. collaborators of Doe's Office of Science drive the design of the US network-network peering

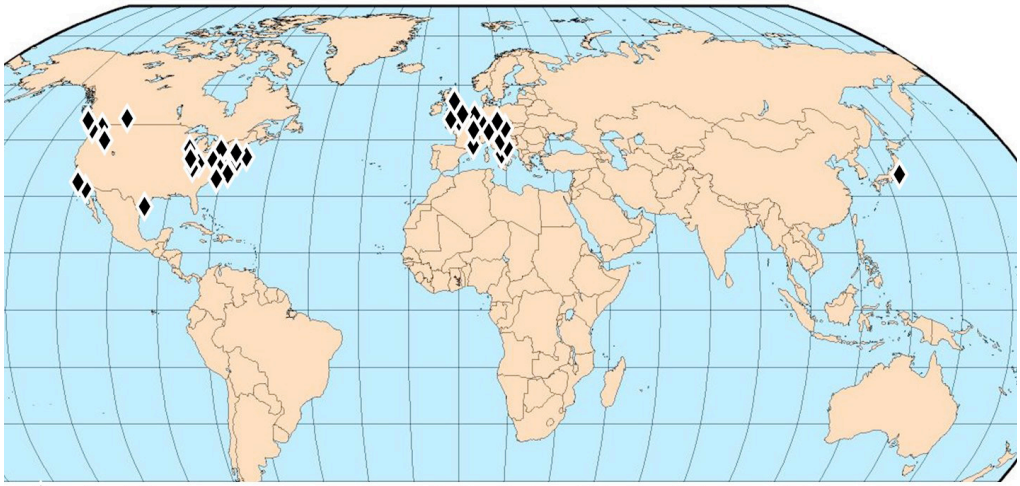


Figure 3. Requirements for connectivity: The international collaborators of DOE's Office of Science, drives ESnet design for international connectivity

Other Requirements

Assistance and services are needed for smaller user communities that have significant difficulties using the network for bulk data transfer. Part of the problem here is that WAN network environments (such as the combined US and European R&E networks) are large, complex systems like supercomputers, and it is simply unrealistic expect to get high performance when using this "system" in a "trivial" way – this is especially true for transferring significant amounts of data over distances of thousands to tens of thousands of km.

Some user groups need more help than others.

- Collaborations with a small number of scientists typically do not have network tuning expertise
  - They rely on their local system and network administrators
  - They often don't have much data to move
  - Their data transfer and storage systems typically have default network tuning parameters and typically suffer poor wide area data transfer performance as a result
  - Therefore, they avoid using the network for data transfer if possible
- Mid-sized collaborations have a lot more data, but similar expertise limitations
  - More scientists per collaboration, much larger data sets (10s to 100s of terabytes)
  - Most mid-sized collaborations still rely on local system and networking staff, or supercomputer center system and networking staff for WAN assistance (where their expertise is typically limited)
- Large collaborations (HEP, NP) are big enough to have their own internal software shops
  - Dedicated people for networking, performance tuning, etc
  - Typically need much less assistance in using wide area networks effectively

- Often held up (erroneously) as an example to smaller collaborations

A semi-quantitative assessment of this situation based on experience is illustrated in Figure 4

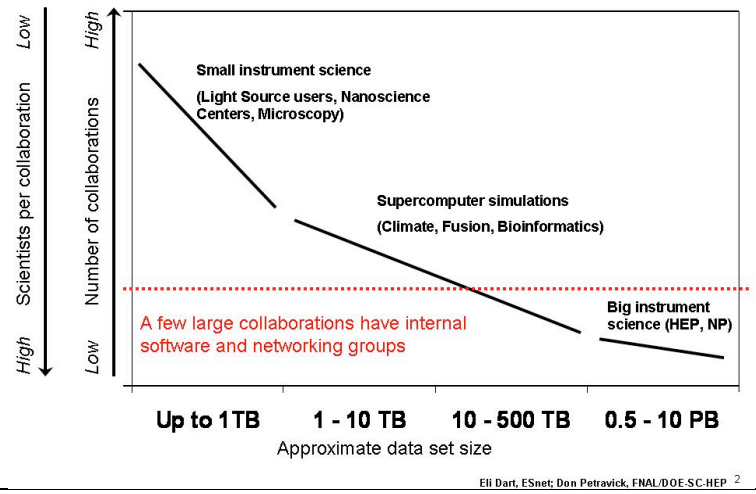


Figure 4. Rough user grouping by collaboration data set size. (Eli Dart, ESnet, and Don Petravick, FNAL/DOE-SC-HEP)

To address some of these issues ESnet has established a Web site for information and best practice on system tuning for high wide area network throughput: fasterdata.es.net.

Observations from the Requirements Workshops

The concept that that grew out of early work with the HEP community – namely that modern science cannot be accomplished without capable, high quality, very high-speed networks interconnecting the collaborators, their instruments, and their data repositories – has been validated, extended, and nuanced based on the findings of the first four (of six) formal requirements workshops.

These workshops use case studies to identify how instruments (including supercomputers) and the process of doing science are changing and how those changes both depend on and drive the network.

One of the fundamental strategic realizations from HEP, and confirmed by the workshops, is that science productivity is now directly related to the amount of data that can be shared / accessed / incorporated / processed / catalogued in the new process of science that is heavily dependent on data analysis and computational models used by individuals, groups, and large collaborations of scientists. This is directly evident from the workshops in the areas of combustion and climate simulation, protein and cellular modeling (Proteomics), and the collaborative data analysis associated with large scientific instruments such as STAR and PHENIX at RHIC and Atlas and CMS at the LHC.

Several observations come from these sorts of specific science use of networks case studies:

- Networks are critical infrastructure for science, but no scientist wants to have to be aware of the network.

- Networks are essential for the functioning of large experiments, but the cost of networks (and computing) is tiny compared to the cost of large instruments (networks typically cost millions of dollars where scientific instruments often cost many billions of dollars – the LHC is a roughly $US 10 billion experiment)
- Networks are just one part of the infrastructure of data management that is key for scientific productivity – middleware that provides semantically relevant operations for data management is critical to provide a more abstract and capable data movement functionality
- For science data movement, all that matters is end-to-end (application to application) performance. The network could be "infinitely" fast and if some other problem (e.g. access to end host or storage) limits throughput to 100 Kb/s, the fast network is almost useless.
- Middleware is critical, but no matter how good it is it will not be effective if it does not fit into the "culture" of any given science community – utility and usability are critical.
- Different science communities have different requirements and different approaches to data movement / sharing. There will be (are) many different types of data movement middleware. This results in a problem that is too diffuse / too big for a "user" assistance team in a networking organization to have much impact on the overall scientific community. (Network organizations like ESnet and Internet2 are very small compared to many scientific collaborations – especially those involving large instruments.) Therefore it may be that all that the networking organizations can do is to try and provide general advice: "best practice" prescriptions for what approaches work, cautions about what not to do, goals for what should be achievable, etc. In other words collect and publicize best practices at several levels. One example of this is ESnet's WAN performance tuning site fasterdata.es.net.
- Just the process of doing outreach, as in ESnet's requirements workshops, can identify potential or extant issues which leads to increased user awareness, which generates conversations about possible mitigation strategies. This frequently results in the science community beginning a process of integrating WAN data movement throughput solutions into the science planning.

3.1.2. Science requirements determination: Requirements from observing current and historical network traffic patterns

ESnet has detailed historical network traffic information that is accurate going back to about 1989. There is a lot of history of the evolution of science use of networking summarized in this data.

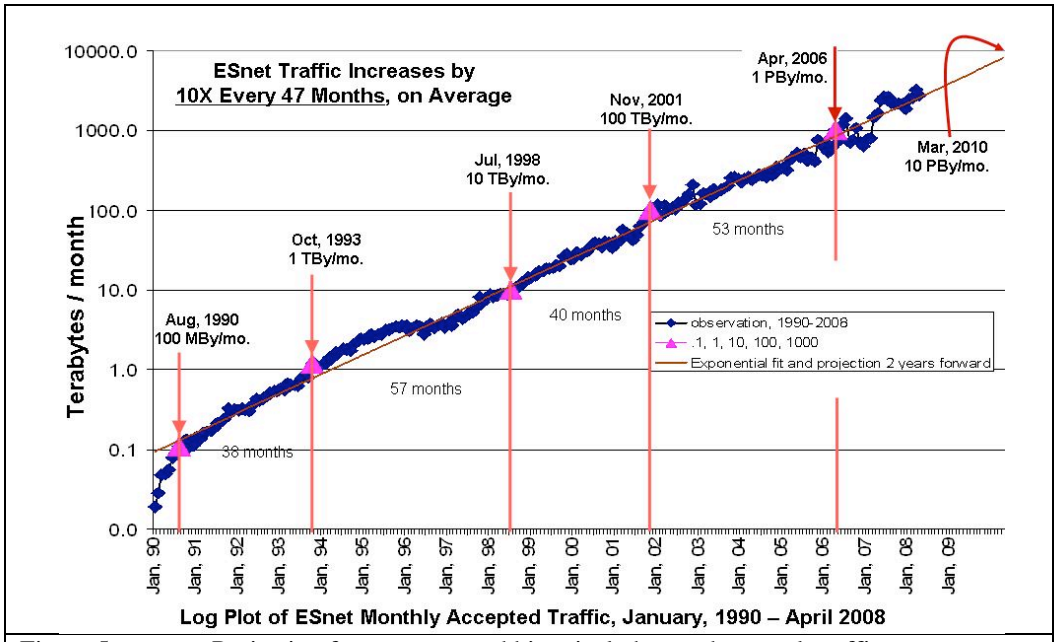


Figure 5. Projection from current and historical observed network traffic patterns

By looking at some of the implications of the historical traffic (Figure 5), together with some link-specific traffic data, we see that in 4 years we can expect a 10x increase in traffic over current levels just based on historical trends:

- nominal average load on the busiest backbone links in June 2006 was ~1.5 Gb/s
- in 2010 average load will be ~15 Gb/s based on current trends and 150 Gb/s in 2014

(These projections are based on early 2006 data were made just as the large-scale science begin to dominate ESnet traffic and so are certainly understated.)

Measurements of this type are science-agnostic – it doesn't matter who the users are or what they are doing: the traffic load is increasing exponentially. Predictions based on this sort of forward projection could produce low estimates of future requirements because they cannot entirely predict new uses of the network (though some of this is built into the "memory" of past uses of the network, which includes new instruments coming on line).

Large-Scale Science Now Dominates ESnet

ESnet carries about 2 x 10^9 flows (connections) per month which account for all of the ESnet traffic. Large-scale science – LHC, RHIC, climate data, etc. – now account for about 90% of all ESnet traffic, but in only a few thousand flows/month. What this implies is that a few large data sources/sinks now dominate ESnet traffic, which means that in the future, overall network usage will follow the patterns of the very large users. (Figure 6) Managing this situation in order to provide reliable and predictable service to large users while not disrupting a lot of small users, requires the ability to isolate these flows to a part of the network designed for them ("traffic engineering").

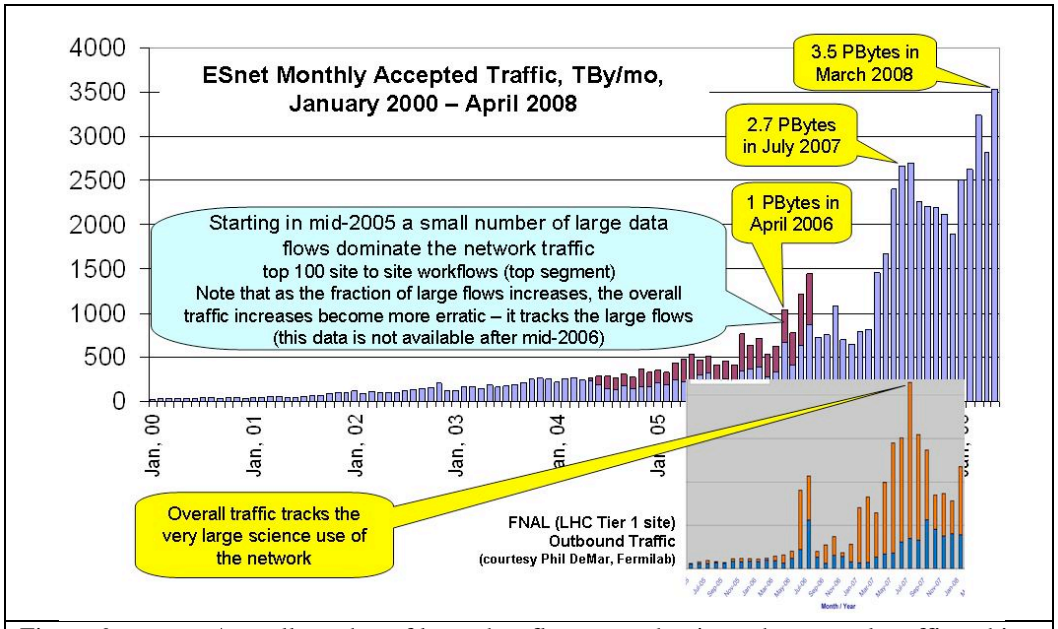


Figure 6. A small number of large data flows now dominate the network traffic – this motivates virtual circuits as a key network service

3.2. Strategy II: 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

ESnet has evolved a planning process that involves:

1. Providing the basic, long-term bandwidth requirements with an adequate and scalable infrastructure;

2. Undertaking continuous reexamination of the long-term requirements because they frequently change;

3. Identifying required new network services and implement them;

4. Developing outreach approaches that identify what communities are not being well served by the network because they are not equipped to make effective use of the network.

To elaborate:

1. Providing the basic, long-term bandwidth requirements with an adequate and scalable infrastructure:

The HEP requirements as articulated in 2004/5 were the basis of the architecture, capacity (and therefore budget), and services of the next generation network (ESnet4). However, because the aggregate HEP requirements were large and the specific connectivity requirements were geographically diverse, ESnet was able to design a network that was generically "large," relatively high aggregate capacity, had a comprehensive network footprint, and was flexible

and scalable in several dimensions. The implementation of the first phase of ESnet4 (based on an Internet2-ESnet partnership for the underlying optical network) was done from mid-2006 to late 2008 (now complete – see Figure 1). This approach of "a rising tide lifts all boats" has proven to be effective in addressing most of the network requirements of the other SC programs that have been identified after the 2004/5 planning process.

2. Undertaking continuous reexamination of the long-term requirements because they frequently change:

In spite of a methodical approach to determining the long-term requirements (mostly via the requirements workshops), unanticipated requirements show up in what appears to be an inherently hap-hazard way, with surprises coming from all quarters. There are two probable reasons for this:

- The SC science Programs are large and while the workshops try and characterize the major science in each program some areas are missed
- The processes of science are constantly refined and updated which frequently produces new requirements

Is there sufficient flexibility and scalability in the original "new network" (ESnet4) design and implementation to accommodate the "surprises"? Probably, but not certainly.

3. Identify required new network services and implement them:

Intuiting the actual, useful implemented form of a new network service from the user articulated requirements (e.g. guaranteed bandwidth) is very hard. It runs the danger of using detailed requirements from too small a portion of the community (probably necessary initially in order to implement anything) and subsequently discovering that one community's service semantic does not solve for another community what appeared to be the same issue.

Further, no network that is useful to the science community operates in isolation. Almost any new service that is going to be useful must be implemented in all of the networks that are involved in (typically) global science collaborations. This means that most of the world's R&E networks must agree on the service description and then implement it in the particular architecture and implementation (hardware + circuits) of their network. Just getting the required level of coordination – even after there is in-principle agreement that the service is necessary – is a labor intensive and lengthy process. Even so, in two important areas – the guaranteed bandwidth / circuit services and end-to-end monitoring – this cooperative process is working well.

4. Develop outreach approaches that identify what communities are not being well served by the network because they are not equipped to make effective use of the network:

Even when the network tries to deal with particular communities (science collaborations) the number of different solutions (typically middleware that not only solves the network interface problem but works with all of the applications that the community already has developed) is of a scale that rapidly exceeds the manpower that can reasonably be made available in the network organizations. This is also a lengthy process that has to be approached with considerable "social skills" in order to be effective in raising the community's awareness to the point that they see that science resources must be allocated to the problem of effective use of the network

3.2.2. Planning Process 1. Provide the basic, long-term bandwidth requirements with an adequate and scalable infrastructure

ESnet4 was built to address specific Office of Science program requirements. The result is a much more complex and much higher capacity network than in the past (Figure 7).

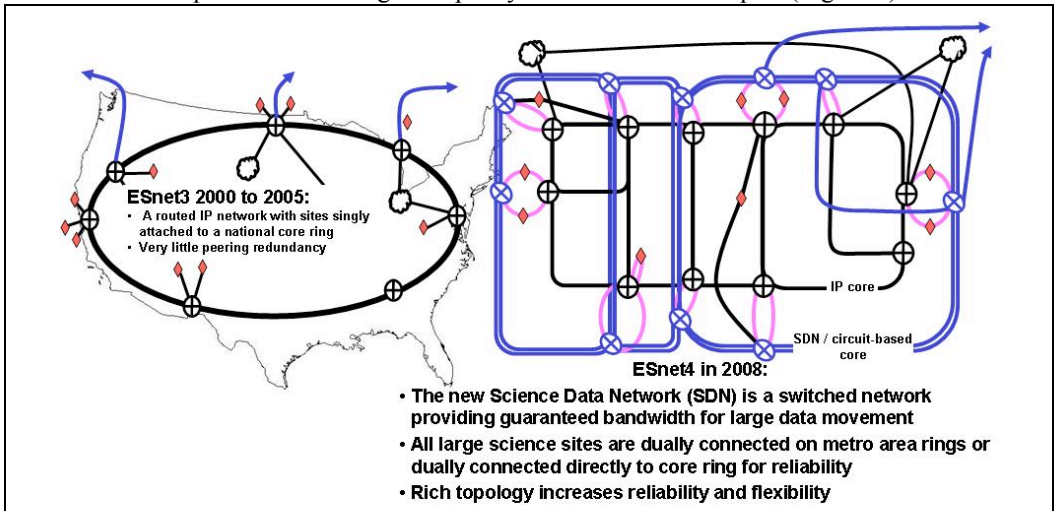


Figure 7. Over its 20+ year history there have four major architectural changes in ESnet: from 56 Kb/s satellite links to 45 Mb/s X.25 to 155 Mb/s ATM to 2.5 and 10 Gb/s SONET to the current multiple 10 Gb/s optical circuits (ESnet4)

Building the Network as Opposed to Planning the Budget

Aggregate capacity requirements like those shown in Table 1 indicate how to budget for a network but do not tell you how to build a network. To actually build a network you have to look at where the traffic originates and ends up and how much traffic is expected on specific paths. In 2006, when the construction of ESnet4 started, we had more or less specific bandwidth and path (collaborator location) information for LHC (CMS, CMS Heavy Ion, and Atlas), the SC Supercomputers, CEBF/JLab, and RHIC/BNL. This specific information has lead to the current and planned configuration of the network for the next several years. (Figure 8 and Figure 9 illustrate the resulting path configuration and the anticipated path loadings (Gb/s) in 2010.)

As a gauge of how realistic the 40-50 Gb/s traffic predictions are we can look at the full-scale analysis system trials conducted by the LHC, CMS detector collaboration. During one four month period these trials were generating 9 Gb/s of traffic – about 4 Gb/s average sustained over the entire four months (Figure 10). This clearly indicates that 1) the bandwidth predictions are not unreasonably high, and 2) that a new generation of scientific instruments and distributed data analysis is generating levels of network traffic never before seen in the R&E community.

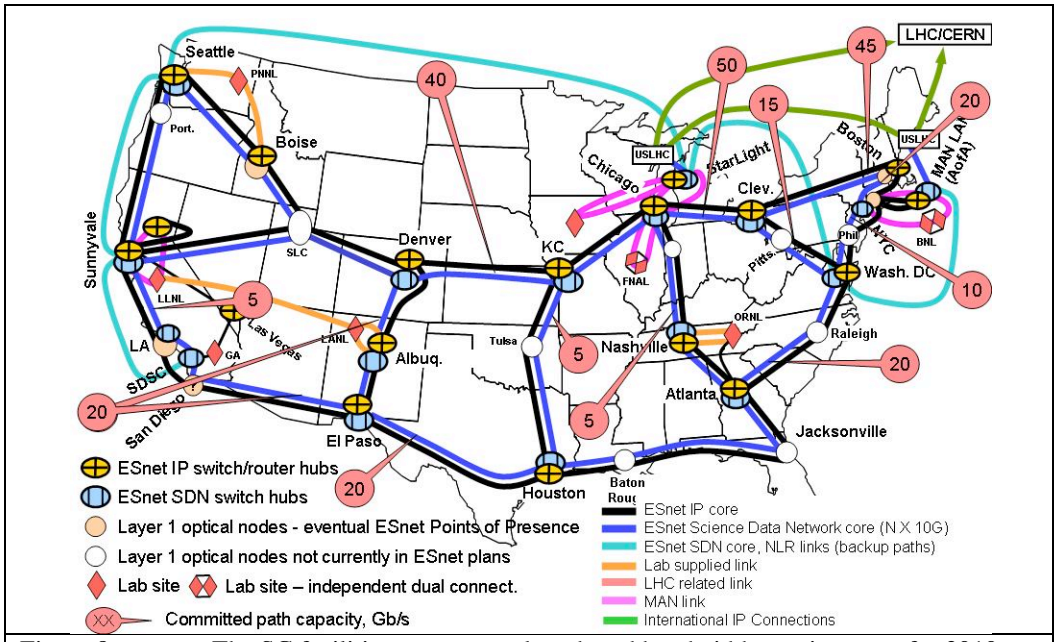


Figure 8. The SC facilities core network path and bandwidth requirements for 2010 as identified in 2006. The network in 2008 has two 10Gb/s optical circuits on the footprint illustrated. One 10G circuit will be added each year until there are a total of six – the current planned capacity of ESnet4.

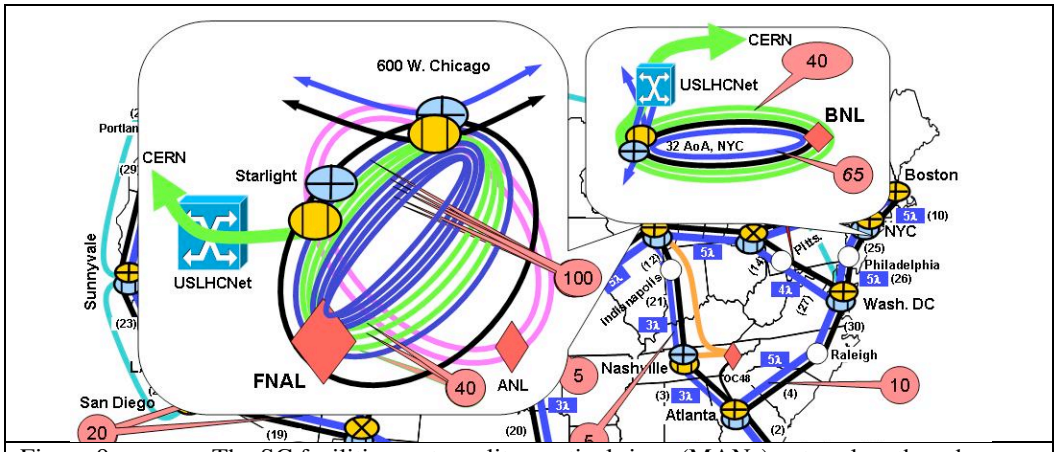


Figure 9. The SC facilities metropolitan optical rings (MANs) network path and bandwidth requirements for 2010 as identified in 2006 for the LHC and RHIC.

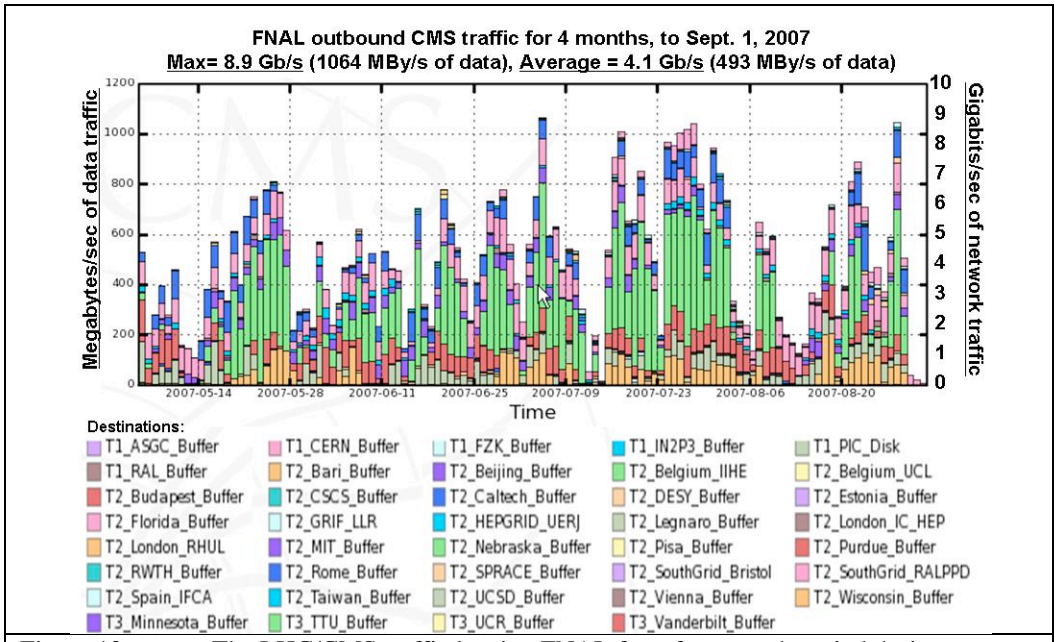


Figure 10. The LHC/CMS traffic leaving FNAL for a four month period during a "service trial" (a coordinated exercise where all of the analysis systems involved in CMS data analysis operate at full data rates on synthetic data in order to debug and determine the readiness of the systems).

One (deliberate) consequence of ESnet's new architecture is that site availability (as seen from other sites) is increasing. (Figure 11)

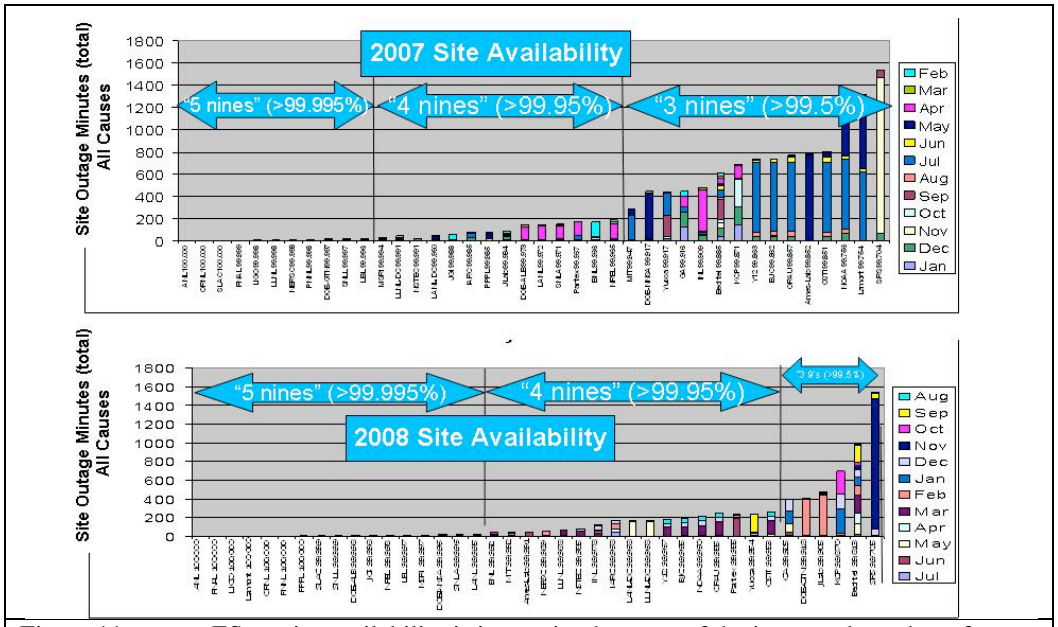


Figure 11. ESnet site availability is increasing because of the increased number of redundant connections to the sites and within the network. (The SC Labs are mostly on the left side – high reliability. Note that there is very little that ESnet can do to improve the availability for the high-outage sites (right side). These are mostly due to carrier outages on long, low bandwidth (copper) circuits to small sites / offices.)

3.2.3. Effectiveness of the approach: Re-evaluating the Strategy and Identifying Issues

The current strategy (that lead to the ESnet4, 2010-12 plans) was developed primarily as a result of the information gathered in the 2002 and 2003 network workshops, and their updates in 2005-6 that had input from the LHC, climate, RHIC, SNS , Fusion Energy sites, supercomputing, and a few others) [1]. So far, the more formal requirements workshops have largely reaffirmed the ESnet4 strategy developed based on these early workshops. However – is this the whole story?

How do the Science Program Identified Requirements Compare to this Capacity Planning?

We can get a feeling for whether the planning process is adequate, at least in aggregate, by looking at the aggregate requirements as articulated and comparing them to 1) the planned network, 2) an estimate of the growth rates of the planned capacity and the projected growth of science data.

Table 2 shows the aggregate bandwidth requirements ("5 year end-to-end") from the general descriptions of the instruments and the amount of data that the science community knows will have to be put onto the network (from Table 1); the bandwidth accounted for in path planning (because there was enough information available about who would be using the data to define network paths for that data), and; the difference (requirements unaccounted for in the actual network path planning).

Table 2. Synopsis of Known Aggregate Requirements, 6/2008

Science Areas / Facilities	**5 year end-to-end bandwidth requirements (Gb/s)**	**accounted for in current ESnet path planning**	**Unacc'ted for**
ASCR: ALCF	30	30	
ASCR: NERSC	40	40	
ASCR: NCLF	50	50	
BER: Climate	20		20
BER: EMSL/Bio	100		100
BER: JGI/Genomics	5		5
BES: Chemistry and Combustion	30		30
BES: Light Sources	60		60
BES: Nanoscience Centers	30		30
Fusion: International Collaborations	1		1
Fusion: Instruments and Facilities	20		20
Fusion: Simulation	88		88
HEP: LHC	265	265	
NP: CMS Heavy Ion	20		20
NP: JLAB	10		10
NP: RHIC	20	20	
total	**789**	**405**	**384**

This difference was anticipated in the network design and the planned (and budgeted for) network capacity is considerably larger than the 405 Gb/s "accounted for" plans, and appears adequate to meet the projected needs (Table 3).

Table 3. ESnet Planned Aggregate Capacity (Gb/s) Based on 5 yr. Budget vs. 5 yr. Science Network Requirements Aggregation Summary

year	2006	2007	2008	2009	2010	2011	2012	2013
ESnet planned aggregate capacity	57.50	192	192	842	1442	1442	1442	2042
Requirements (aggregate Gb/s)								789

Note that the "planned aggregate capacity" measure is the sum of the link capacity on all of the major inter-city links in the core network. There is enough peering point and topological diversity (that is, there is traffic flowing into and out of the network in almost every one of the cities that define the inter-city links) that this has proven to be at least a somewhat useful measure.

Therefore, the planned aggregate capacity growth of ESnet matches the know requirements as understood in 2006/8. The "extra" capacity indicated in the table (e.g. 2042 Gb/s planned vs. 789 Gb/s required in 2013) is needed to account for the fact that there is less than complete

flexibility in mapping specific path requirements to the aggregate capacity planned network and specific paths specific paths will not be known until several years into building the network. Whether this approach works is to be determined, but indications are that it probably will.

3.2.4. Planning Process 2. Undertaking continuous reexamination of the long-term requirements because they frequently change

Recently ESnet has undertaken an effort to identify trends that would, at least somewhat independently, validate or change the projections developed by the methodologies described above. This was done by a combination of soliciting user estimates and looking at the 15 year historical traffic trends in the network.

One hypothesized way to develop an independent measure of growth is to assume that the science processes of large instruments (and probably supercomputers in the next few years) have changed forever and large scale distributed collaborations are now the norm. Because of this, just looking at the growth of the volume of data that must be managed by any discipline will provide a measure of the growth of network traffic.

As an indication of the validity of this assumption (the changing process of science is continuing to drive more use of the network) consider what is happening to high energy physics data analysis.

When the network planning was done for the two major LHC experiments (CMS and ATLAS) in 2005-6 the assumption (by the HEP community) was a fairly strict hierarchical model with the LHC at the apex (tier 0) sending data to nation-based tier 1 data centers (e.g. FNAL and BNL in the U.S.). The data analysis would mostly be done at tier 2 centers (universities with significant computing capability) and that the tier 2 centers would mostly generate 3-5 Gb/s network flows from one tier 1 center. Tier 3 centers (smaller universities) would get data mostly form the tier 2 centers at about 1 Gb/s.

However, what has happened is that several Tier2 centers are capable of 10Gb/s now and many Tier2 sites are building their local infrastructure to handle 10Gb/s. Further, many Tier3 sites are also building 10Gb/s-capable analysis infrastructures. Both Tier2 and Tier3 sites will be accessing data from multiple Tier1 data centers. This was not in LHC plans even a year ago.

It has been said[§] that scientific productivity will follow high-bandwidth access to large data volumes and this provides a powerful incentive for science groups to upgrade network capacity and to use it.

We are already seeing the "onslaught" of this traffic from the Tier 2 centers (not anticipated in 2006) and it now seems likely that this will cause a second onslaught in 2009 as the Tier3 sites all upgrade their network and computing capacity to handle 10Gb/s of LHC traffic. In other words it is possible that the USA installed base of LHC analysis hardware will consume significantly more network bandwidth than was originally estimated.

There are already indications of this when looking at the historical traffic record in a little more detail than just the 15 year aggregate. If we project that traffic from 1990 to 2004 (which we

[§] Harvey Newman (HEP, Caltech) predicted this eventuality several years ago

refer to as the "old" era vs. "new" era because of the clear change in traffic patters in 2004-5 (Figure 6)) as one projection and the traffic from 2004-2008 as a second projection, what we see is that, indeed, there is a significant difference, especially as the projections are on a log plot. (Figure 12)

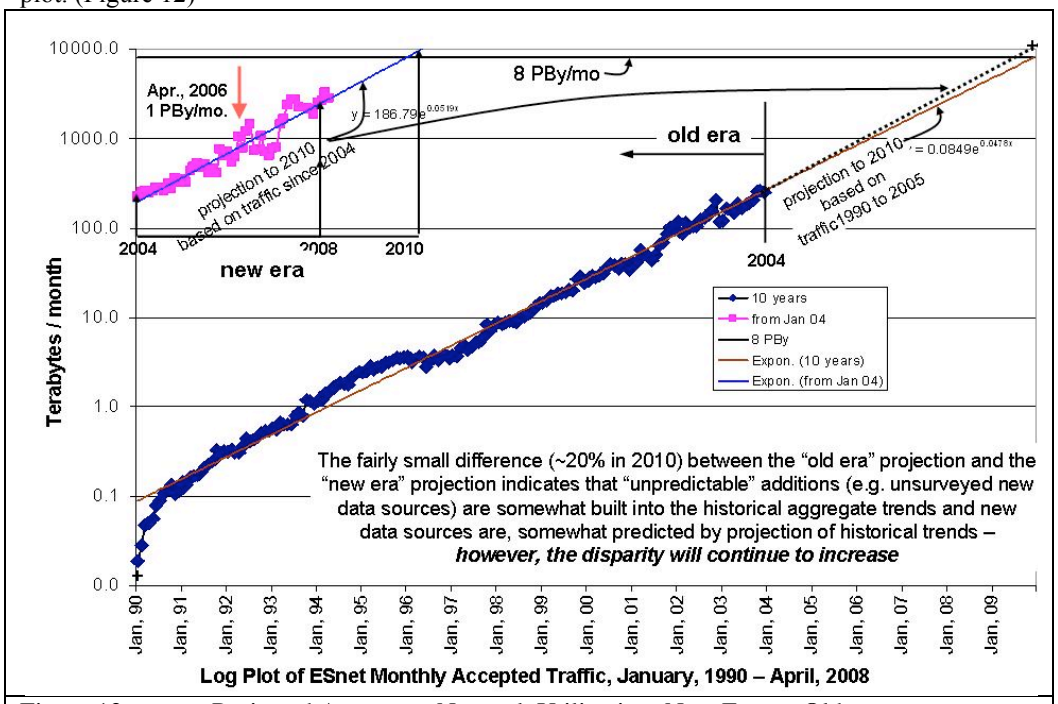


Figure 12. Projected Aggregate Network Utilization: New Era vs. Old

As another indicator, if we consider the growth of the volume of scientific data that will almost certainly be moved over the network in the future (as it is today in the high energy physics community) we had a basis for comparing this growth with the planned capacity of the network. The result (Figure 13) is that three measures (projected traffic, projected HEP data, and projected climate simulation data) are growing faster than the planned capacity of the network.

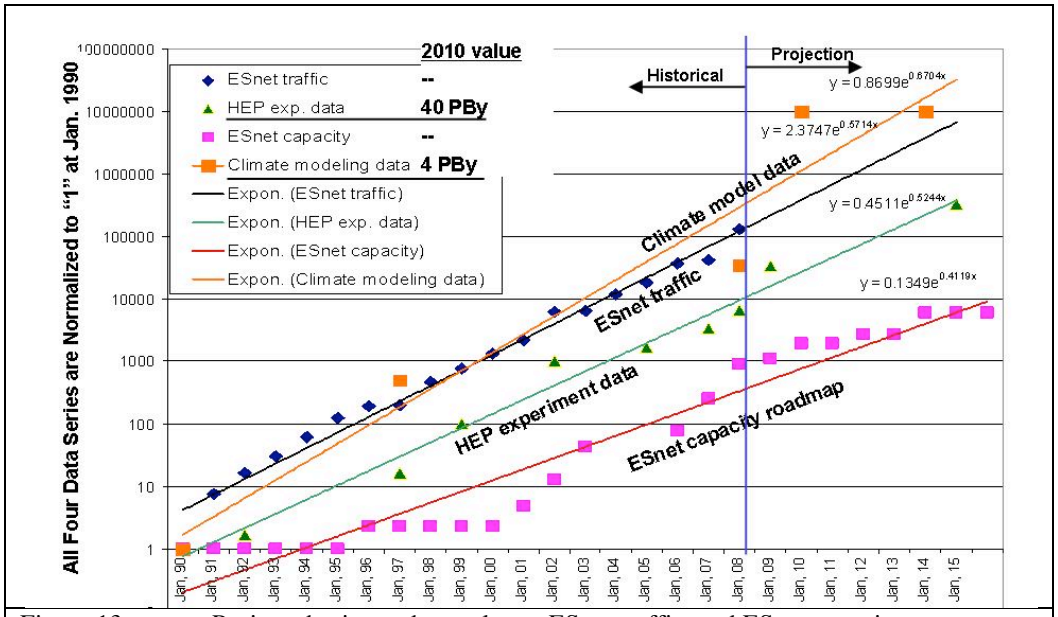


Figure 13. Projected science data volume, ESnet traffic, and ESnet capacity.

(Ignore the units of the quantities being graphed they are normalized to 1 in 1990, just look at the long-term trends: All of the "ground truth" measures are growing significantly faster than ESnet projected capacity.) *(HEP data courtesy of Harvey Newman, Caltech, and Richard Mount, SLAC. Climate data courtesy Dean Williams, LLNL, and the Earth Systems Grid Development Team.)*

Issues for the Future Network

The current estimates from the LHC experiments and the supercomputer centers have the currently planned ESnet 2011 wave configuration operating at capacity and there are several other major sources that will be generating significant data in that time frame (e.g. Climate). The significantly higher exponential growth of traffic (total accepted bytes) vs. total capacity (aggregate core bandwidth) means traffic will eventually overwhelm the capacity – "when" cannot be directly deduced from aggregate observations, but if you add the fact that the nominal average load on busiest backbone paths in June 2006 was ~1.5 Gb/s - In 2010 average load will be ~15 Gb/s based on current trends and 150 Gb/s in 2014 then one could guess that capacity problems will start to occur by 2015-16 without increasing the planned capacity of the network. Further, the "casual" increases in overall network capacity based on straightforward commercial channel capacity that have sufficed in the past are less likely to easily meet future needs due to the (potential) un-affordability of the hardware. For example the few existing commercial examples of >10G/s router/switch interfaces are ~10x more expensive than the 10G interfaces which is not a practical solution unless prices come down.

Further, we cannot just keep adding capacity by provisioning more 10 Gb/s optical circuits for several reasons. First is that the cost of the sort of managed wave (circuit) service that ESnet uses (and feels is necessary for the required level of reliability based on observing the reliability of wide area networks are operated by the R&E community) is not practical for two

reasons. A 10 Gb/s circuit across ESnet's entire footprint (as in the planned network) is more than $US 1 million/year. Second, there are not enough waves available to the R&E community. The several national fiber footprints that exist today are unlikely to be able to increase because of projected capacity demands by the telecom industry.

Where Will the Capacity Increases Come From?

This leaves us with two likely approaches to increase capacity.

ESnet4 planning assumes technology advances will provide100 Gb/s optical waves (they are 10 Gb/s now) which gives a *potential* 5000 Gb/s aggregate (in the sense of summing link capacities as described above) core network by 2012. The ESnet4 SDN switching/routing platform is designed to be upgradable to 100 Gb/s network interfaces, so not all of the core network equipment would have to be replaced at the same time. With capacity planning based on the ESnet 2010 wave count, together with some considerable reservations about the affordability of 100 Gb/s network interfaces, we can probably assume some fraction of the 5000 Gb/s of potential, aggregate core network capacity by 2012 depending on the cost of the equipment – perhaps 20% – about 2000 Gb/s of aggregate path capacity (which is the assumption in the number in Table 3 and Figure 13). Increases beyond this will depend on 100G waves, and the associated network equipment, becoming affordable[**].

The second approach involves using a more dynamic use of the underlying optical infrastructure than is possible today.

The Internet2-ESnet partnership optical network that both organization use today is build on dedicated fiber and optical equipment that is configured with 10 X 10G waves / fiber path. The optical equipment allows for more waves to be added in groups of 10 up to 80 waves.

The current wave transport topology is essentially static or only manually configured – our current network infrastructure of routers and switches assumes this. However, assuming that some fraction of the 80 waves can be provisioned (i.e. at an affordable cost), then with completely flexible traffic management extending down to the optical transport level we should be able to extend the life of the current infrastructure by moving significant parts of the capacity to the specific routes where it is needed. This entails integrating the management of the optical transport with the "network" and providing for dynamism / route flexibility at the optical level in order to make optimum use of the available capacity. To understand the issues and the approach consider the next several figures.

Figure 14 illustrates the configuration of the optical node in the Internet2-ESnet network today. A fixed set of waves is assigned to each organization which then uses those waves as static circuits between interfaces on routers and switches.

[**] Unlike a year ago, there are now some encouraging indicators regarding the possibility of affordable 100G circuits. For example, at the Supercomputing 2008 conference Internet2, ESnet, Infinera, Ixia, Juniper Networks and Level 3 Communications will aggressively develop and test emerging 100 Gigabit Ethernet (GbE) technologies over the next year with the aim of deploying them on the nationwide Internet2 and ESnet networks. Further, ESnet is now (7/2009) funded to build a 100 Gb/s testbed across the country.

Figure 15illustrates the fact that, even though the Infinera optical nodes have internal wave switching capability, currently the path of a wave through the network is static – the same as though it were a physical circuit.

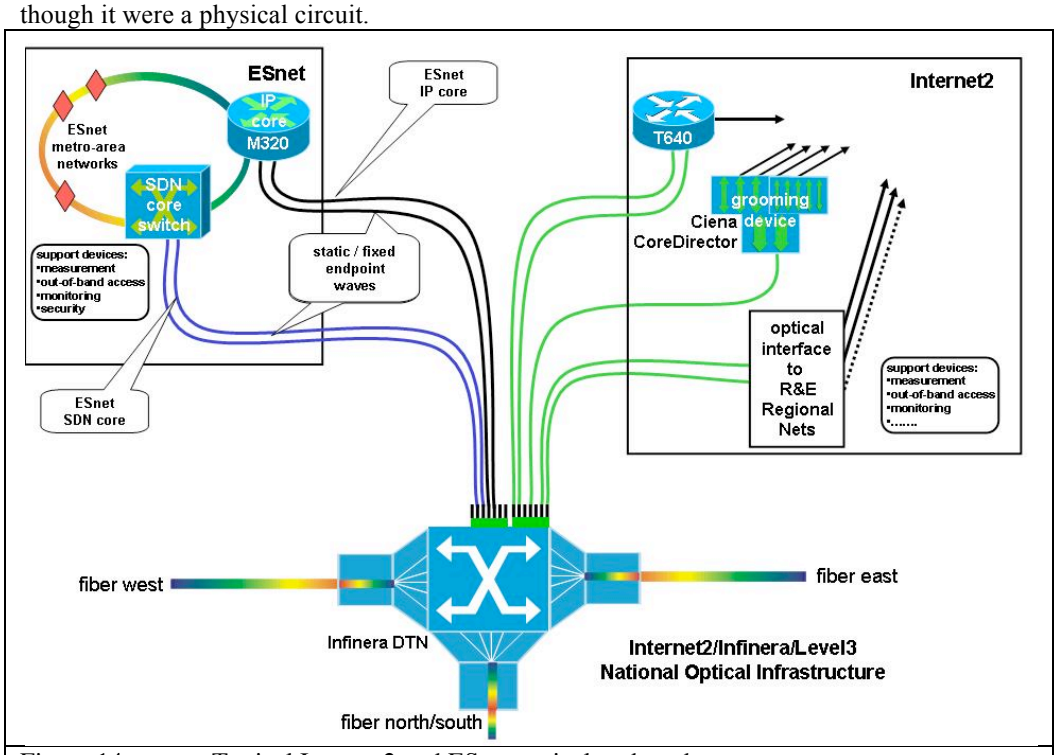


Figure 14. Typical Internet2 and ESnet optical node today

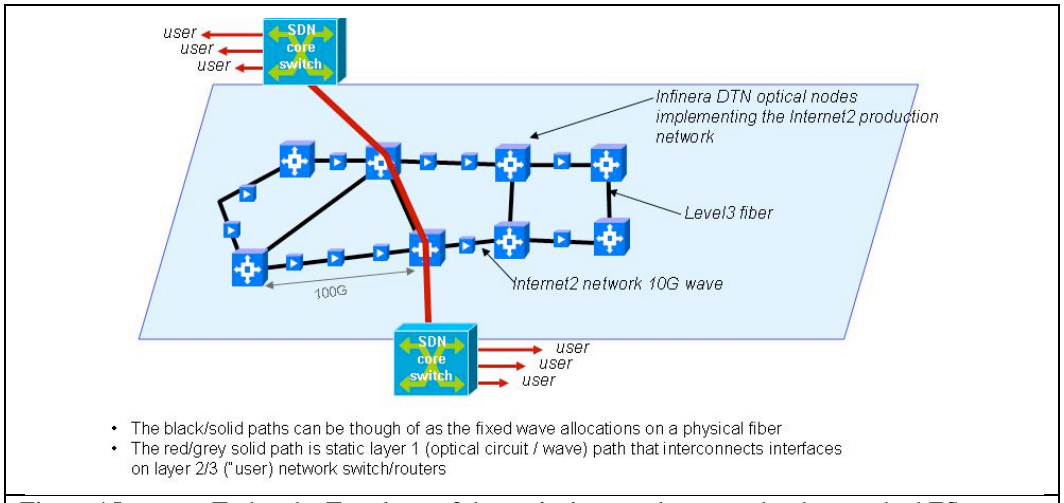
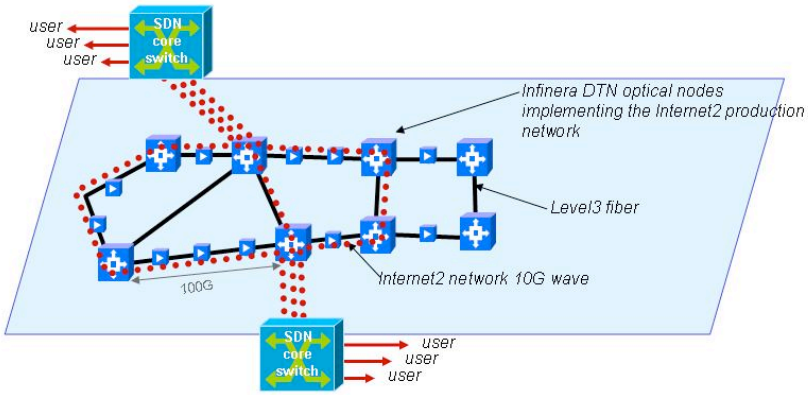


Figure 15. Today the Topology of the optical network as seen by the attached ESnet switches and routers is determined by a static configuration

Dynamically Routed Optical Circuits for Traffic Engineering

The Infinera optical devices ("DTN") convert all network traffic (Ethernet or SONET) to G.709 framing internally and the DTNs include a G.709 crossbar switch that can map any input (user network facing) interface to any underlying wave and therefore to any other network interface. By adding a layer 1 (optical) control plane that is managed by Internet2 and ESnet, the G.709 switch can be controlled in concert with topology information communicated to the switches and routers, and the underlying topology of the optical network can be changed as needed for traffic management, including the use of topological path redundancy for increasing point to point capacity (Figure 16).

This layer 1 control plane is in the planning phase and a testbed to develop the control plane management that is isolated from the production network is in the late stages of planning. The control plane manager will be based on an extended version of the OSCARS [2] dynamic circuit manager.



- The black/solid paths can be though of as the fixed wave allocations on a physical, fiber
- The red/grey/dotted paths are dynamically switched layer 1 (optical circuit / wave) paths that can provide:
  i) transient paths between the layer 2/3 ("user) network switch/router interfaces for congestion avoidance (traffic engineering)
  ii) multiple paths between layer 2/3 network switch/router interfaces for a higher capacity optical connection across the network, or
  iii) direct connections between sites if the intermediate networks can carry optical circuits
  all of this is transparent to the

Figure 16. Dynamically routed optical circuits allow for traffic engineering (load balancing / rerouting)

As more waves are provisioned in the Internet2-ESnet optical network these new waves can be shared and dynamically allocated to maximize the utilization of the available waves. This has the potential to double or triple the effective capacity of the network because of the rich and redundant physical topology.

Conclusions

The warning of the trends illustrated in Figure 13 ("Projected science data volume, ESnet traffic, and ESnet capacity.") is that even though you cannot make direct predictions as to the traffic resulting from the science data volume increases; it is growing exponentially faster than the projected network capacity which is guaranteed to cause problems in the future. (My guess, based on looking at traffic growth and link loadings, is that problems will occur by 2015, or so, if nothing is done – WEJ).

On the other hand, there are two technology advances, both of which are being pursued – 100G optical circuits and dynamic wave / optical circuit management – that should provide new network capabilities that will carry us to the next full new technology generation. (And we have no particular insights into what that will look like, but it will be based on research going on in physics laboratories today, probably related to new uses of quantum mechanical effects.)

3.2.5. Planning Process 3. Identify required new network services and implement them

The capabilities that are users' top priority are reservable, guaranteed bandwidth and user-level, end-to-end monitoring. The monitoring capability is being done as an international community effort to refine and deploy perfSONAR and is described in [3] and [4]. The reservable, guaranteed bandwidth service is described here.

OSCARS: ESnet's Guaranteed Bandwidth Virtual Circuit Services

To support large-scale science, networks must provide communication capability that is service-oriented. In the case of the virtual circuit service, the capabilities must be:

- Configurable – Be able to provide multiple, specific "paths" (specified by the user as end points) with specific characteristics.
- Schedulable – Premium service such as guaranteed bandwidth will be a scarce resource that is not always freely available, therefore time slots obtained through a resource allocation process must be schedulable.
- Predictable – A committed time slot should be provided by a network service that is not brittle, i.e. reroute in the face of network failures is important.
- Reliable – The service should be transparently self-correcting, e.g. virtual circuit reroutes should be largely transparent to the user.
- Informative – When users go to configure distributed systems they (or their cyber agents) should be able to see average path characteristics, including capacity, etc. When things do go wrong, the network should report back to the user in ways that are meaningful to the user so that informed decisions can about alternative approaches.
- Scalable – The underlying network should be able to manage its resources to provide the appearance of scalability to the user
- Geographically comprehensive – The R&E network community must act in a coordinated fashion to provide this environment end-to-end.
- Secure – The use of guaranteed virtual circuits consumes a valuable resource (guaranteed bandwidth) and connects directly into the user environment. Therefore the user must have confidence that the other end of the circuit is connected where it is supposed to be connected and that the circuit cannot be "hijacked" by a third party while in use.
- Provide traffic isolation – Users want to be able to use non-standard/aggressive protocols when transferring large amounts of data over long distances in order to achieve high performance and maximum utilization of the available bandwidth.

The ESnet Approach for Required Virtual Circuit Capabilities

ESnet's OSCARS system provides configurability, schedulability, predictability, and reliability with a flexible virtual circuit (VC) service: user specifies end points, bandwidth, and schedule and OSCARS can dynamically traffic engineer the underlying paths.

Providing useful, comprehensive, and meaningful information on the state of the paths, or potential paths, to the user is based on perfSONAR, and associated tools. PerfSONAR provides real time information in a form that is useful to the user (via appropriate abstractions) and that is delivered through standard interfaces that can be incorporated in to SOA/Web Services style applications. Techniques still need to be developed to monitor virtual circuits based on the approaches of the various R&E nets – e.g. MPLS in ESnet, VLANs, TDM/grooming devices (e.g. Ciena Core Directors), etc., and then integrate this into a perfSONAR framework.

Reliability approaches for Virtual Circuits are still under investigation and are topics for R&D – there are many ramifications to "blindly" rerouting a VC and the subtleties (e.g. how to minimally interfere with other circuits) need to be worked out.

Scalability will be provided by new network services that, e.g., provide dynamic wave allocation at the optical layer of the network in order to increase the effective bandwidth between end points by taking advantage of the path redundancy possible in a topologically rich core network (see section 3.2.4 "Where Will the Capacity Increases Come From?," above), and reduce routing overhead.

Geographic ubiquity of the services can only be accomplished through active collaborations in the global R&E network community so that all sites of interest to the science community can provide compatible services for forming end-to-end virtual circuits. To accomplish this active and productive collaborations exist among numerous R&E networks: ESnet, Internet2, Caltech, DANTE/GÉANT, some European NRENs, some U.S. regional networks (RONs), etc.

OSCARS circuits are "secure" to the edges of the network (the site boundary/DMZ) because they are managed by the control plane of the network which is highly secure and isolated from the general traffic.

The end-to-end provisioning of OSCARS virtual circuits is provided by explicit Label Switched Paths (LSP) using MPLS (Multi-Protocol Label Switching) and RSVP (Resource Reservation Protocol). The explicit LSPs that are automatically traffic engineered to prefer paths on the SDN supports both layer 3 (IP) and layer 2 (Ethernet VLANs) end-user VC services. From a user's perspective, the layer 3 VC service provides an isolated IP "tunnel", and the layer 2 VC service provides an Ethernet VLAN path, both of which have bandwidth guarantees. Having the ability to traffic engineer LSPs over both the SDN and IP has the advantage that virtual circuits can be set up to sites that only have IP connections to the network, which is true for many smaller institutions.

For guaranteed bandwidth there are several ways that the requirements may be met. Most R&E networks use a hardware device (sometimes called a "grooming[††] device") that provides the required capabilities between hardware interfaces on an Ethernet or SONET based path. ESnet took a software approach based on several standard protocols in order to provide end-to-end virtual circuits. OSCARS uses MPLS and RSVP in combination with QoS to provide guaranteed bandwidth within the SDN and IP core network. MPLS and RSVP is used in provisioning explicit LSPs which are traffic engineered hop-by-hop between the ESnet ingress and egress edge points. QoS is used to ensure service guarantees (or lack thereof) for different classes of traffic, such as guaranteed VC traffic, and best effort IP traffic. Policing on a per VC basis is done at the ESnet ingress edge points to prevent over-subscription of reservations, by ensuring that each user can only get the requested and reserved bandwidth. Reservations are also tracked link by link and over time by OSCARS to prevent over-booking of available link bandwidth (admission control). (However, see below on how user experience has led to somewhat relaxing this restriction.)

[††] Traffic grooming is the process of grouping many small network flows into larger units, which can be processed as single entities. Dedicated hardware circuits such as Ethernet circuits are used to connect the grooming devices together.

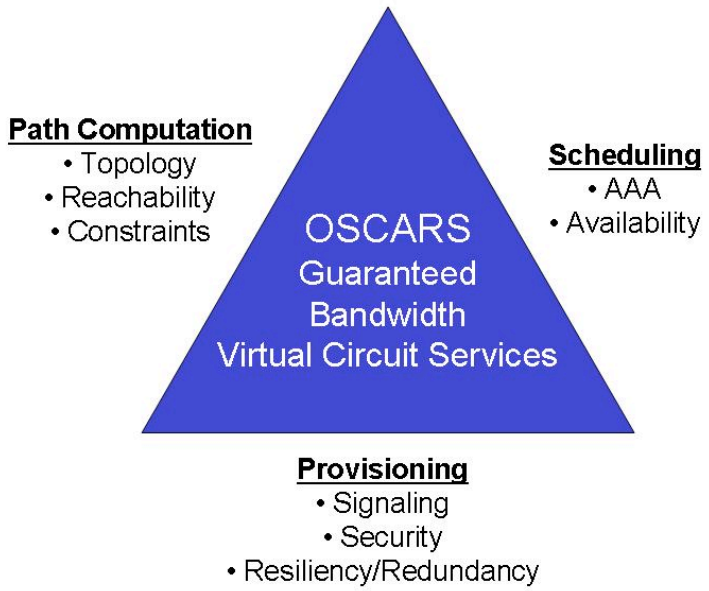


Figure 17. OSCARS is an extensible framework that incorporates various functions.

Configurability, schedulability, predictability (bandwidth guarantees), and reliability needs identified in the science requirements are part of OSCARS by design since this is the problem that OSCARS was intended to address.

OSCARS Status

ESnet developed the original OSCARS code base working with the Internet2. In the ESnet centric deployment of the service in the network, a prototype layer 3 (IP) guaranteed bandwidth virtual circuit service was deployed in ESnet in early 2005. A prototype layer 2 (Ethernet VLAN) virtual circuit service deployed in ESnet in third quarter 2007. Support for "soft reservations" (reservations that are guaranteed the user requested bandwidth under congestion scenarios but can burst higher if excess bandwidth) was provided in the second quarter, 2008. Automatic diagramming of the VCs associated with a given site was also introduced in second quarter, 2008, as was the capability for sites to administer their own circuits.

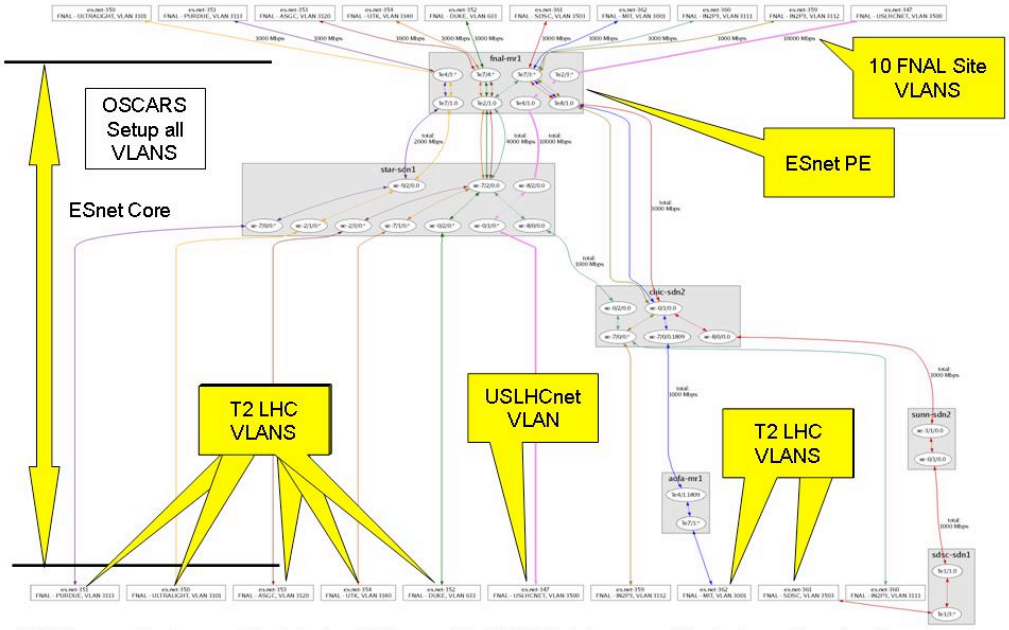


Figure 18: OSCARS generated and managed virtual circuits at FNAL – one of the US LHC Tier 1 data centers. This circuit map (minus the yellow callouts that explain the diagram) is automatically generated by an OSCARS tool and assists the connected sites with keeping track of what circuits exist and where they terminate.

OSCARS is a production service in ESnet

OSCARS is now deployed as a prototype production service in ESnet and has the status of all other ESnet services. It is, however, a very new service and still evolving.

As a production service, OSCARS is undergoing a continual reassessment and is adapting to user experiences with the service.

User experience in the first year of OSCARS operation has revealed several new capabilities that are required. In spite of the fact that one of OSCARS primary goals was to provide hard bandwidth guarantees, it has turned out that this needed to be relaxed. It is useful to permit over-subscribing a path in order to accommodate backup circuits and to allow for site managed load balancing.

Another area that has become clear is that there is a need to direct routed IP traffic onto SDN in a way transparent to the user. There are many issues here and this is an area of research and experimentation at the moment.

Inter-domain collaborative efforts are extensive and well developed

Efforts to ensure that the virtual circuit service will operate across network domains (e.g. between ESnet, Internet2, and GÉANT) are extensive and well developed.

Terapaths developed at BNL and LambdaStation developed by FNAL are both systems designed to provide guaranteed bandwidth within the site / Lab domains. OSCARS has been designed to interoperate with these existing systems and inter-domain interoperability for layer 3 virtual circuits was demonstrated third quarter 2006 (Terapaths) and for layer 2 virtual circuits demonstrated at SC07 (fourth quarter 2007) for both Terapaths and LambdaStation.

Seamlessly setting-up optical circuits across independently operated networks is the goal that will make the virtual circuit service useful to the science community and this requires the coordination of multiple administrative domains. This is achieved through compatible control plane software enabling provisioning across domain boundaries with the appropriate authentication and authorization. Compatible control plane software is under development through several on-going projects, including the NSF-funded DRAGON project, the ESnet OSCARS program, and the GÉANT2 AutoBAHN project. Interoperability of circuits set up across these domains was demonstrated at SC07 (fourth quarter 2007).

The DICE working group (DANTE/GEANT, Internet2, Caltech, ESnet) coordinates the work and issue resolution (both software and operational) for cross-domain virtual circuits. Recent DICE accomplishments include a draft of topology exchange schema being formalized in collaboration with OGF's Network Measurement Working Group (NMWG), and interoperability was demonstrated 3Q07. (This capability is related to discovering available paths through networks outside of your own domain. Likewise, an initial implementation of cross-domain reservation and signaling messages (the DICE InterDomain Controller Protocol (IDCP)) demonstrated at SC07. Other multi-organization activity includes interoperability testing with the Nortel (Canadian telecom equipment manufacturer) DRAC (Dynamic Resource Allocation Controller) which implemented the IDCP on their platforms.

3.2.6. Planning Process 4. Develop outreach approaches that identify what communities are not being well served by the network because they are not equipped to make effective use of the network

Assistance and services are needed for smaller user communities that have significant difficulties using the network for bulk data transfer. This issue cuts across several SC Science Offices, and these problems must be solved if a broad category of scientists are to effectively analyze the data sets produced by petascale machines.

Consider some case studies.

Light Source case study: Light sources[‡‡] (ALS, APS, NSLS, etc) serve many thousands of users. A typical user team is one scientist plus a few graduate students that use 2-3 days of beam time per year. They take data, then go home and analyze the data. Data set sizes are up to 1TB, typically 0.5TB. There is widespread frustration with network-based data transfer among light source users, and a few hundred kb/s of throughput on a network path that is 10 Gb/s end-

[‡‡] "Light sources" in this context refer to a specialized sort of particle accelerator that is designed to produce beams of light that are highly monochromatic and/or highly phase coherent and usually in the frequency range of x-rays. These instruments are key to research in life science imaging, semiconductor device development, materials research, etc. See, e.g., http://www-als.lbl.gov/

to-end is not an unusual experience. There are many reasons for this: Appropriate WAN transfer tools are not installed, user end systems (typically Windows PCs) are not tuned for WAN transfers, and there is a lack of available expertise for fixing these problems. There can also be network problems at the "other end" – typically in a small part of a university network.

The result is that today users copy data to portable hard drives or burn stacks of DVDs, but data set sizes will probably exceed hard disk sizes in the near future.

Combustion simulation case study: Combustion simulations generate large data sets. A user awarded an INCITE allocation (large CPU time allocation) on a NERSC supercomputer generates 10TB data sets. The user is then awarded an INCITE allocation at ORNL and needs to move data set from NERSC to ORNL. There are persistent data transfer problems here involving a lack of a common toolset, unreliable transfers, and low performance. In one case the data was moved, but it took almost two weeks of babysitting the transfer by the user.

Fusion Energy case study: Large-scale fusion simulations (e.g. GTC - a 3D particle-in-cell code developed for studying turbulent transport in magnetic confinement fusion) are run at both NERSC and ORNL, Users wish to move data sets between supercomputer centers for various sorts of analysis. Data transfer performance is low, and workflow software is unavailable or unreliable.

Common issue: Cybersecurity policies are often an obstacle to scientific collaboration. Firewalls, OTP, and PKI requirements can all make it difficult to share data, and often scientists give up and send DVDs instead of "fighting" with site security personnel over appropriate approaches to high-speed WAN data transfers.

Addressing these problems is essential

Persistent performance problems exist throughout the DOE Office of Science (and everywhere else in the science community except a few large collaborations like those associated with the LHC). Existing tools and technologies (e.g. GridFTP, TCP tuning) are not deployed on end systems or are inconsistently deployed across major resources. The resulting performance problems impede scientific productivity – unreliable data transfers soak up scientists' time, because, e.g., they must babysit transfers.

Default system configuration is inadequate and most system administrators don't know how to properly configure a computer for WAN data transfer. Scientists and system administrators typically don't know that WAN data transfer can be high performance, so they don't ask for help.

Tools and technologies for high performance WAN data transfer exist today. TCP tuning documentation exists, and tools such as GridFTP are available and are used by sophisticated users. DOE has made significant contribution to these tools over the years and one of the things that ESnet has done is to set up a web site devoted to information / best practice on bulk data transfer, host tuning, etc. (fasterdata.es.net)

Sophisticated users and programs are able to get high performance – user groups with the size and resources to "do it themselves" get good performance (e.g. HEP, NP). Smaller groups do not have the internal staff and expertise to manage their own data transfer infrastructures, and so get low performance. The WAN is the same in the high and low performance cases but the end system configurations are different.

Apart from tuning and WAN tool deployment, other potential approaches include various latency hiding forwarding devices in the network which is a current topic of R&D and experimentation in ESnet and in Internet2 (see subsection "Transparent Acceleration of Data Transfers" in section 3.3, below).

As a result of the experience ESnet has gained in the requirements workshops a path forward has been identified (which would probably not be a lot different for any similar set of users, e.g. in university environments):

- Task one entity (within the SC community) with development, support and advocacy for WAN data transfer software
- Port GridFTP to new architectures – we need these tools to work on petascale machines and next-generation data transfer hosts
- Increase the usability of the tools – scientific productivity must be the goal of these tools, so they must be made user-friendly so scientists can be scientists instead of working on data transfers
- Work toward a consistent deployment – all major DOE facilities must deploy a common, interoperable, reliable data transfer toolkit (NERSC, ORNL, light sources, nanocenters, etc)
- Workflow engines are needed to automate the transfer of large data sets or large numbers of files
- Test and monitoring infrastructure (e.g. iperf+PerfSONAR) must be available at every site

However, several of these have the caveat that middleware must have a long term support infrastructure or scientists will be leery of becoming dependent on something that might become unsupported in the future.

These problems MUST be solved if scientists are to effectively analyze the data sets produced by petascale machines.

3.3. Strategy III: Anticipating future network capabilities that will meet future science requirements with an active program of R&D and Advanced Development

The next generation networks and services that are needed to support the new science environment are being defined by examining three types of requirements: the data generating and use characteristics of new instruments and facilities; changes in the process of doing science in this new environment, and examination of the future implications of historical trends in network traffic.

Additionally, advances in the state of telecommunications and network technology over the past several years enable revolutionary approaches to providing national-scale network services. Concurrently, middleware and applications architectures are increasingly supporting wide area distributed systems, introducing fundamental changes in the nature and profile of network traffic as observed by network providers. Hybrid network architectures - involving both traditional Internet connectivity and scheduled, targeted (circuit-like) capacity - offer the potential to provide a radically different set of interaction modes between the network and applications and middleware.

ESnet4 - the next generation SC network - responds to both the need for dramatic increases in bandwidth and for new circuit based network services. Initially, ESnet4 is a hybrid network that mixes a conventional IP packet network and a circuit oriented network. While work has

been going on for several years to define the circuit services, control mechanisms, and AAA mechanisms, many questions remain as to how various aspects of this service should be configured and managed in a production network environment. Additionally, there is increasing demand for services supporting end-to-end performance, such as monitoring on a per-application basis, as well as other new services that are needed to move the network from a communal, best-effort service to a schedulable, dedicated service that can be incorporated into the managed resource environment of large-scale science experiments.

To move beyond proof-of-concept demonstrations and initial deployment for sophisticated users toward persistent, reliable services it will be necessary to harvest the best concepts shown to be feasible and systematically move them into production services. To support this new set of services, and these new application and middleware systems, there must also be deep understanding of the behavior of the systems and the interactions between applications, middleware, and network services.

In April, 2007 the DOE Office of Science organized a workshop to bring together a small group of experts to work with the ESnet team to examine the current roadmap, roadmaps of similar enterprises, user requirements, and new technology options. The objective of the workshop was to create a new, multi-year technology roadmap for ESnet that identifies milestones and partners. The document from which this section is abstracted[§§] is an updated and condensed version of the report from that workshop. The workshop report is available at www.es.net/hypertext/ESnetRD-Workshop-Report.pdf and a prioritized list of tasks is available in the appendix of that report.

These topics include:

- Guaranteed network bandwidth
- End-to-end monitoring
- Transparent acceleration of data transfers
- Federated Trust

R&D topics to secure the future

To address the needs of guaranteed network bandwidth and end-to-end monitoring, a number of research and development topics were identified.

Virtual Circuits

There are a number of issues that must be solved to enable the production use of virtual circuits. The control plane must allow for the provisioning of services in a multi-service, multi-layer, multi-domain, multi-vendor environment. Control plan issues include topology schema design and exchange for inter-domain setup of virtual circuits, finding and scheduling optical network paths, intelligent service selection, secure control plane exchange, and mechanisms for gracefully handling circuit outages. The management plane must support an authentication and authorization infrastructure that includes support for service level agreements and a circuit charge model. The data plane should support coordinated data plane handoffs at domain boundaries, and provide efficient end-to-end virtual circuits using the lowest common network transport layer.

[§§] This section is abstracted from "ESnet Research Priorities, August 2008" which is available at http://www.es.net/hypertext/Adv-Dev-Projects.html

Dynamic Wave Management

ESnet4 operates on an optical infrastructure shared with Internet2. The majority of the optical circuits are used individually by each network, however some are expected to be used jointly by the two networks – that is, ESnet will have optical circuits dedicated to its use, Internet2 will have optical circuits dedicated to its use, and there will be some circuits that are shared in order to support the new dynamic provisioning services.

Dynamic joint management of a collection of waves that are designated for this purpose will be an important capability for, e.g., traffic engineering (the automatic provisioning of additional capacity on heavily (perhaps transiently heavily) used paths) and for managing scheduled 10G circuits where entire waves will form the links in a circuit path.

In order to accomplish the R&D needed to build and test the control plane for dynamic wave management it is essential to have a testbed environment where the R&D can be performed on the DWDM equipment in a way that is guaranteed to be non-interfering with the production waves on the optical network.

Automatic Network Management

R&E networks in general, and ESnet in particular, have become much more complex in recent years, and this trend will continue into the future. Better tools for network management will be essential in order for these networks to scale without greatly increasing the operations staffing levels.

Network management systems should be able to anticipate problems based on a continual analysis of monitoring data in order to detect potential problems before they happen. The goal of such a system would be to anticipate problems based on knowledge-based analysis and projections of network state. For example, if packet rate counters for a router interface were steadily – but irregularly enough to escape manual inspection – growing over time, an automated analysis system would predict that the router would become saturated and warn the engineering staff. Another example is the case where a relatively low bandwidth interface, say with a commercial peering partner, were to rapidly saturate and thus disrupt traffic to that peer. This circumstance should be automatically detected in order to allow for defensive action such as automatically rerouting traffic to an alternate path in order to avoid problems.

Analysis across many geographically diverse interfaces could also be used to detect coordinated stealth attacks against the network, against a group of sites, or against a particular network application; allowing cybersecurity actions to be initiated.

Such an automated analysis and management capability would allow network engineers to specify 'what if' scenarios, and then specify what to do if such a situation starts to develop. Even if this did not result in the "permanent" or production fix of the problem, it would give engineers breathing room to better analyze the situation in detail and design a long-term solution. This type of system could also be used to detect paths with particularly large number of high-speed flows where a dedicated wavelength could be used to better manage the network traffic. In the long term this could be integrated with the virtual circuit management system to automatically move such traffic off of the IP network and into the circuit infrastructure.

In order to create such an automated system several issues must be addressed. Improved networking monitoring data and monitoring data archives such that described in the next

section will be required; signal/anomaly detection techniques will have to be identified or developed; ontologies will have to be developed to describe the semantics of network functionality; rule-based systems will need to be interfaced with the signal/anomaly detection system in order to generate knowledge-base actions; the consequential actions will have to be passed to systems that interact with the network control plane to make changes in the network configuration; and so forth.

End-to-End Monitoring Services

Currently there is no standardized way to determine what performance levels the network is capable of delivering that works across multiple domains, what portions of the network are up and working correctly and what parts are broken or down for maintenance, etc. This leads to problems distinguishing between network issues, problems in the applications, problems with the protocols, and congestion caused by other network users.

Cross-domain network measurement and monitoring services must be developed in order to facilitate effective use of advanced networks by complex, widely distributed applications. These services should allow users, applications and other network middleware to determine realistic performance expectations, document the services that are delivered, verify that the capabilities committed are actually provided, and easily discriminate between network problems and limitations, and application problems. These services need to be provided in a seamless fashion to support paths that cross many different domains in support of globally distributed applications.

We need a standardized way to determine what performance levels the network is capable of delivering across multiple domains. This is needed to distinguish between network issues, problems in the applications, problems with the protocols, and congestion caused by other network users. As advanced services such as virtual circuits are deployed this problem will only get harder.

Network measurement and monitoring services, including a network topology discovery service, must be developed in order to facilitate effective use of advanced networks by complex distributed applications. Further research and development is needed in the areas of the deployment of a network measurement framework, improved end-to-end monitoring tools, better integration with Grid Middleware, and tools for monitoring virtual circuits.

Transparent Acceleration of Data Transfers

On a LAN one can obtain reasonable transfer speeds using the default host settings and standard file transfer tools such as *scp* and *sftp*. However, host tuning and specialized tools are still required to obtain good performance on a WAN. We need to eliminate the need for scientists and their students to become networking experts in order to obtain good WAN file transfer rates.

One way to approach this that is a transparent way to accelerate data transfer speeds. A number of hardware appliance-based approaches to aspects of this problem exist, such as the Cisco "Wide Area Application Services" (WAAS) (http://www.cisco.com/web/go/waas), Phoebus (http://www.internet2.edu/performance/phoebus/), and REDDNet (http://www.reddnet.org), yet none of these fully solve the problem of moving large data sets from a typical MS Windows-based desktop or laptop to/from a data-store at a university or National Laboratory.

Research is needed to determine how and when to redirect traffic to the acceleration appliance, and what the best, easy to use approach to problem of authentication and authorization on the appliance is.

PKI and Federated Trust Issues

The cross-site and international nature of DOE Office of Science collaborations demands a well managed, scalable, flexible, and federated approach to authentication, authorization, and the creation of virtual organizations to manage the collaboration resources. Current approaches are ad hoc and impose a high overhead on the scientists – and security professionals. What is needed is a system for cross-site authentication and attribute-based authorization to allow sites and communities involved in cross-site collaboration to meet the policy needs of federal government computing sites, without imposing unmanageable procedures on users of the resources. The goal of such a system would be to build on federation of identity, by allowing a user's identity to be initially sourced from a trusted organization, such as a university IT department or DOE laboratory, which performs authentication and vetting of the user. ESnet is a logical trusted third party to manage such a system.

While this topic is very important to the functioning of large-scale collaborations, this paper has not discussed these issues in order to constrain the scope of the paper. Please see the R&D topics report for more information.

4. Summary

ESnet and the Office of Science have put in considerable effort in working with the SC community to identify the networking implication of the instruments, supercomputers, and the evolving process of how science is done

The resulting understanding of the current and evolving needs of the SC community has resulted in a new network environment that we believe will enable the distributed aspects of SC science. The new network and its services are not a static implementation, but continuously evolving as the reassessment of the effectiveness of the original plans indicate needed changes and because the ways in which the science community uses the network changes with an evolving process of science.

In order to avoid limiting the scope of solutions to engineering practices that are developed and understood, ESnet and SC use an active program of R&D and Advanced Development effort to keep anticipating future network capabilities that will meet future science requirements.

5. Acknowledgements

In addition to the authors, the ESnet senior network engineering staff that is responsible for the evolution of ESnet also contributed many ideas: Joseph H. Burrescia, Michael S. Collins, Jon Dugan, James V. Gagliardi, Yvonne Y. Hines, Kevin Oberman, and Michael P. O'Connor. Steve Cotter is now the head of ESnet.

ESnet is funded by the US Dept. of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program. Dan Hitchcock was the ESnet Program Manager when ESnet4 was

being planned. Vince Dattoria is the current Program Manager. Thomas Ndousse-Fetter is the Program Manager for the network research program that funded the OSCARS project.

ESnet is operated by Lawrence Berkeley National Laboratory, which is operated by the University of California for the US Dept. of Energy. This work was supported by the Director, Office of Science, Office of Advanced Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

6. References

[1] See http://www.es.net/hypertext/requirements.html

[2] For more information contact Chin Guok (chin@es.net). Also see http://www.es.net/oscars

[3] perfSONAR is an infrastructure for network performance monitoring, making it easier to solve end-to-end performance problems on paths crossing several networks. It contains a set of services delivering performance measurements in a federated environment. These services act as an intermediate layer, between the performance measurement tools and the diagnostic or visualization applications. This layer is aimed at making and exchanging performance measurements between networks, using well-defined protocols. See www.perfsonar.net

[4] "Network Communication as a Service-Oriented Capability." March 2008, William E Johnston, Joe Metzger, Mike O'Connor, Michael Collins, Joseph Burrescia, Eli Dart, Jim Gagliardi, Chin Guok, and Kevin Oberman. Published in: "High Performance Computing and Grids in Action," Volume 16 Advances in Parallel Computing, Editor: L. Grandinetti, March 2008, IOS Press, ISBN: 978-1-58603-839-7. Also available at http://www.es.net/pub/esnet-doc/index.html