

Network Communication as a Service-Oriented Capability

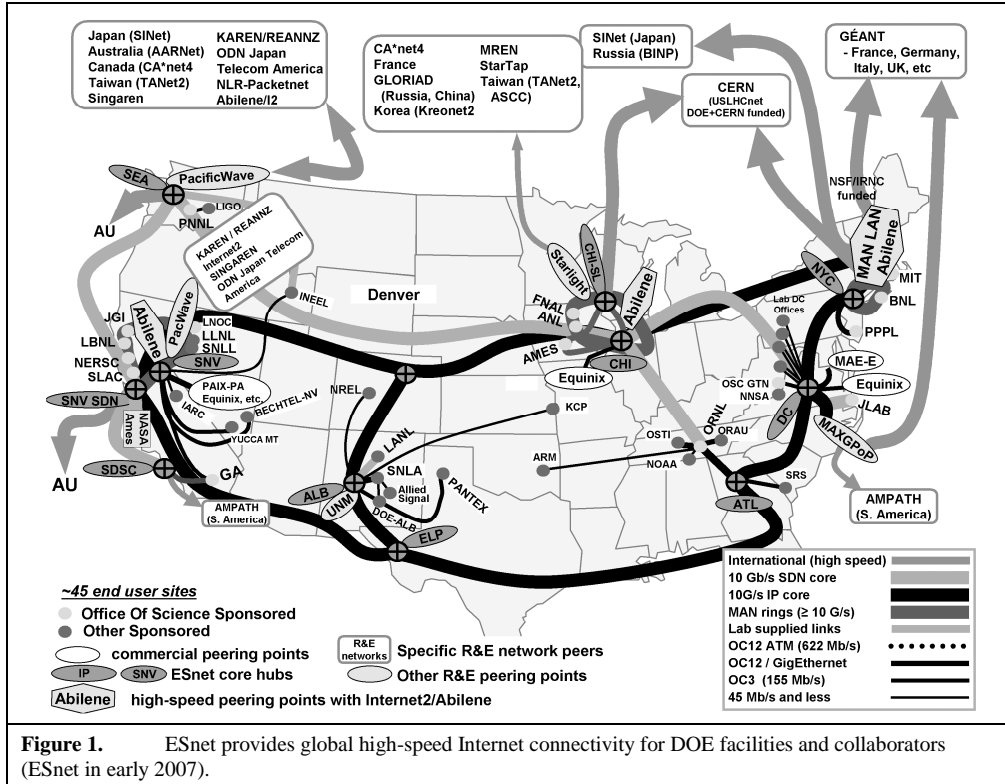
William JOHNSTON, Joe METZGER, Mike O'CONNOR, Michael COLLINS, Joseph
BURRESCIA, Eli DART, Jim GAGLIARDI, Chin GUOK, and Kevin OBERMAN
ESnet, Lawrence Berkeley National Laboratory

Abstract. In widely distributed systems generally, and in science-oriented Grids in particular, software, CPU time, storage, etc., are treated as "services" – they can be allocated and used with service guarantees that allows them to be integrated into systems that perform complex tasks. Network communication is currently not a service – it is provided, in general, as a "best effort" capability with no guarantees and only statistical predictability.

In order for Grids (and most types of systems with widely distributed components) to be successful in performing the sustained, complex tasks of large-scale science – e.g., the multi-disciplinary simulation of next generation climate modeling and management and analysis of the petabytes of data that will come from the next generation of scientific instrument (which is very soon for the LHC at CERN) – networks must provide communication capability that is service-oriented: That is it must be configurable, schedulable, predictable, and reliable. In order to accomplish this, the research and education network community is undertaking a strategy that involves changes in network architecture to support multiple classes of service; development and deployment of service-oriented communication services, and; monitoring and reporting in a form that is directly useful to the application-oriented system so that it may adapt to communications failures

In this paper we describe ESnet's approach to each of these – an approach that is part of an international community effort to have intra-distributed system communication be based on a service-oriented capability.

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



1. The Network Today

1.1. ESnet's Mission

ESnet's mission is to provide an interoperable, effective, reliable, high performance network communications infrastructure, along with selected leading-edge Grid-related and collaboration services in support of the large-scale, collaborative science that is integral to the mission of DOE's Office of Science (SC).

ESnet must provide services that enable the SC science programs that depend on:

- Sharing of massive amounts of data
- Supporting thousands of collaborators world-wide
- Distributed data processing
- Distributed data management
- Distributed simulation, visualization, and computational steering
- Collaboration with the US and International Research and Education community

To this end, ESnet provides network and collaboration services to DOE laboratories. ESnet also serves programs in most other parts of DOE.

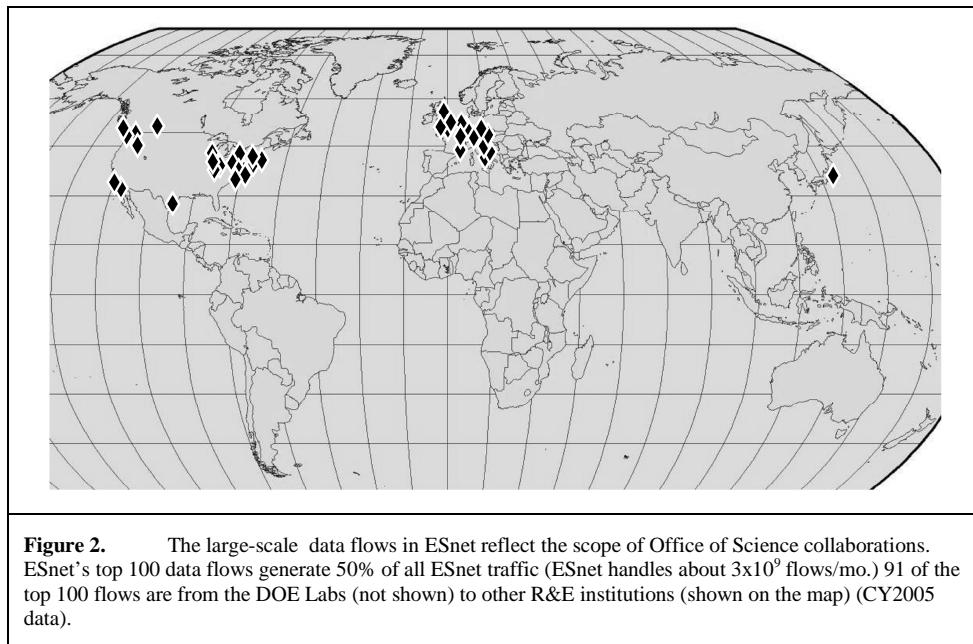
1.2. ESnet Defined

ESnet is:

- A large-scale IP network built on a national circuit infrastructure with high-speed connections to all major US and international research and education (R&E) networks
- An organization of 30 professionals structured for the service
- An operating entity with an FY06 budget of \$26.6M
- A tier 1 ISP providing direct peering with all major networks – commercial, government, and research and education (R&E)
- The primary DOE network providing production Internet service to almost all of the DOE Labs and most other DOE sites. This results in ESnet providing an estimated 50,000 - 100,000 DOE users and more than 18,000 non-DOE researchers from universities, other government agencies, and private industry that use SC facilities with global Internet access.

1.3. ESnet's Place in U. S. and International Science

A large fraction of all of the national data traffic supporting U.S. science is carried by three networks – ESnet and Internet2, and National Lambda Rail. These three entities fairly



well represent the architectural scope of science-oriented networks.

ESnet is a network in the traditional sense of the word. It connects end-user sites to various other networks. Internet2 is primarily a backbone network. It connects U.S. regional networks to each other and International networks. NLR is a collection of light paths or lambda channels that are used to construct specialized R&E networks.

ESnet serves a community of directly connected campuses – the Office of Science Labs. In essence ESnet interconnects the LANs of all of the Labs to the outside world. ESnet also provides the peering and routing needed for the Labs to have access to the global Internet. Internet2 serves a community of regional networks that connect university campuses. These regional networks – NYSERNet (U.S. northeast), SURAnet (U.S. southeast), CENIC (California), etc., – have regional aggregation points called GigaPoPs and Internet2 interconnects the GigaPoPs. Internet2 is mostly a transit network – the universities and/or the regional networks provide the peering and routing for end-user Internet access. This is very similar to the situation in Europe where GÉANT (like Internet2) interconnects the European National Research and Education Networks (NRENs) that in turn connect to the LANs of the European science and education institutions. (The NRENs are like the US regional networks, but organized around the European nation-states).

The top-level networks – ESnet, Internet2, GÉANT, etc. – work closely together to ensure that they have adequate connectivity with each other so that all of the connected institutions have high-speed end-to-end connectivity to support their science and education missions. ESnet and Internet2 have had joint engineering meetings for several years (Joint Techs) and ESnet, Internet2, GÉANT, and CANARIE (Canada) have also formed an international engineering team that meets several times a year.

An ESnet goal is that connectivity from DOE Lab to US and European R&E institutions should be as good as Lab to Lab and University to University connectivity. The key to ensuring this is engineering, operations, and constant monitoring. ESnet has worked with the Internet2 and the international R&E community to establish a suite of monitors that can be used to continuously check a full mesh of paths through all of the major interconnection points.

2. Next Generation Networks

2.1. Evolving Science Environments Drive the Design of the Next Generation ESnet

Large-scale collaborative science – big facilities, massive amount of data, thousands of collaborators – is a key element of DOE's Office of Science. The science community that participates in DOE's large collaborations and facilities is almost equally split between SC labs and universities, and has a significant international component. Very large international (non-US) facilities (e.g., the LHC particle accelerator at CERN in Switzerland and the ITER experimental fusion reactor being built in France) and international collaborators participating in US based experiments are now also a key element of SC science, requiring the movement of massive amounts of data between the SC labs and these international facilities and collaborators. Distributed computing and storage systems for data analysis, simulations, instrument operation, etc., are becoming common, and for data analysis in particular, Grid-

style distributed systems predominate. (See, e.g., the Open Science Grid – an SC led distributed Grid computing project – <http://www.opensciencegrid.org/>)

This Grid-based science environment is very different from that of a few years ago and places substantial new demands on the network. High-speed, highly reliable connectivity between labs and US and international R&E institutions is required to support the inherently collaborative, global nature of large-scale science. Increased capacity is needed to accommodate a large and steadily increasing amount of data that must traverse the network to get from instruments to scientists and to analysis, simulation, and storage facilities. High network reliability is required for interconnecting components of distributed large-scale science computing and data systems and to support various modes of remote instrument operation. New network services are needed to provide bandwidth guarantees for data transfer deadlines, remote data analysis, real-time interaction with instruments, coupled computational simulations, etc.

There are many stakeholders for ESnet. Foremost are the science program offices of the Office of Science: Advanced Scientific Computing Research, Basic Energy Sciences, Biological and Environmental Research, Fusion Energy Sciences, High Energy Physics, and Nuclear Physics – see <http://www.science.doe.gov/>. ESnet also serves labs and facilities of other DOE offices (e.g., Energy Efficiency and Renewable Energy, Environmental Management, National Nuclear Security Administration, and Nuclear Energy, Science and Technology). Other ESnet stakeholders include SC-supported scientists and collaborators at non-DOE R&E institutions (more than 85% of all ESnet traffic comes from, or goes out to non-DOE R&E organizations), and the networking organizations that provide networking for these non-DOE institutions.

Requirements of the ESnet stakeholders are primarily determined by three approaches: 1) Instruments and facilities that will be coming on-line over the next 5–10 years and will connect to ESnet (or deliver data to ESnet sites in the case of LHC and IETR) are characterized by considering the nature of the data that will be generated and how and where it will be stored, analyzed, and used. 2) The process of science in the disciplines of direct interest to SC is examined to determine how the process of that science will change over the next 5–10 years and how these changes will drive demand for new network capacity, connectivity, and services. 3) ESnet traffic patterns are analyzed based on the use of the network in the past 2–5 years to determine the trends, and then projecting this usage forward in time, thus determining how the network must change to accommodate the future traffic patterns implied by these trends.

2.2. A Case Study: *The Data Analysis for the Large Hadron Collider*^a

The major high energy physics (HEP) experiments of the next twenty years will break new ground in our understanding of the fundamental interactions, structures and symmetries that govern the nature of matter and space-time. Among the principal goals are to find the mechanism responsible for mass in the universe, and the “Higgs” particles associated with

^a Material for this sections is drawn from the “Report of the Standing Committee on Inter-Regional Connectivity (SCIC), Networking for High Energy Physics,” February 8, 2007 [7], and from conversations between WEJ and Harvey Newman of Caltech.

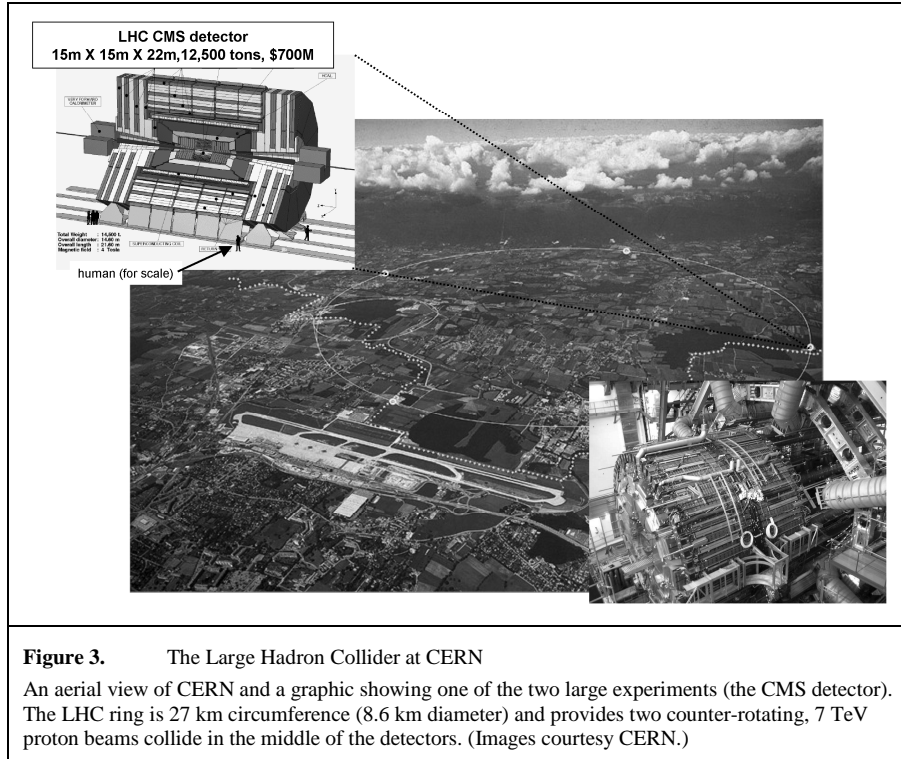


Figure 3. The Large Hadron Collider at CERN
 An aerial view of CERN and a graphic showing one of the two large experiments (the CMS detector). The LHC ring is 27 km circumference (8.6 km diameter) and provides two counter-rotating, 7 TeV proton beams collide in the middle of the detectors. (Images courtesy CERN.)

mass generation, as well as the fundamental mechanism that led to the predominance of matter over antimatter in the observable cosmos.

The largest collaborations today, such as CMS [12] and ATLAS [13], which are building experiments for CERN's Large Hadron Collider program (LHC [14]), each encompass some 2000 physicists from 150 institutions in more than 30 countries. The current generation of operational experiments at Stanford Linear Accelerator Center (SLAC) (BaBar [15]) and Fermilab (D0 [16] and CDF [15]), as well as the experiments at the Relativistic Heavy Ion Collider (RHIC, [18]) program at Brookhaven National Lab, face similar challenges. BaBar, for example, has already accumulated datasets approaching a petabyte.

The HEP problems are among the most data-intensive known. Hundreds to thousands of scientist-developers around the world continually develop software to better select candidate physics signals from particle accelerator experiments such as CMS, better calibrate the detector and better reconstruct the quantities of interest (energies and decay vertices of particles such as electrons, photons and muons, as well as jets of particles from quarks and gluons). These are the basic experimental results that are used to compare theory and experiment. The globally distributed ensemble of computing and data facilities (e.g., see Figure 4), while large by any standard, is less than the physicists require to do their work in an unbridled way. There is thus a need, and a drive, to solve the problem of managing global

resources in an optimal way in order to maximize the potential of the major experiments to produce breakthrough discoveries.

Collaborations on this global scale would not have been attempted if the physicists could not assume the existence of reliable, high capacity, feature-rich networks: to interconnect the physics groups throughout the lifecycle of the experiment, and to make possible the construction of Data Grids capable of providing access, processing and analysis of massive datasets. These datasets will increase in size from petabytes to exabytes (10^{18} bytes) within the next decade. Equally as important is highly capable middleware (the Grid data management and underlying resource access and management services) that is used to facilitate the management of world wide computing and data resources that must all be brought to bear on the data analysis problem of HEP [6].

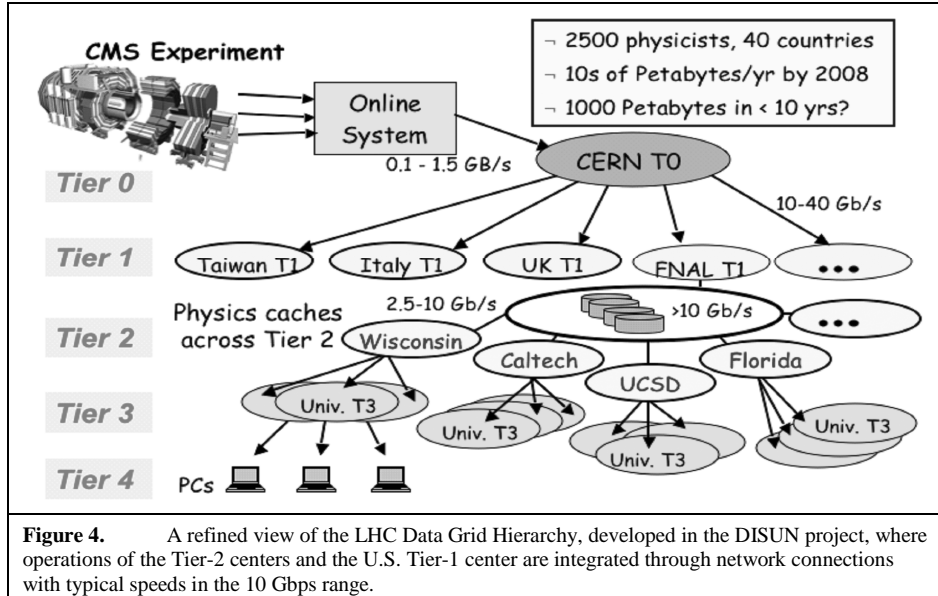
Tiered Model of Regional Computing and Analysis Centers

Building on developments in the early HEP grid projects (PPDG and GriPhyN/iVDGL in the US, and the EU DataGrid), the LHC experiments have adopted the Data Grid Hierarchy of four “Tiers” of globally distributed computing and storage resources. Data at the experiment are stored at the rate of 200-1500 Mbytes/sec throughout the year, resulting in many Petabytes per year of stored and processed binary data that are accessed and processed repeatedly by worldwide collaborators.

Referring to Figure 4, processing and analyzing the data requires the coordinated use of the entire ensemble of Tier-N facilities. The relatively few large Tier-0 and Tier-1 facilities are best suited for the high priority large-scale tasks of systematic data processing, archiving and distribution, and data curation. Moving down the hierarchy to the smaller and more numerous Tier-2 and Tier-3 facilities, individuals and small groups have greater control over how these resources are allocated to small and medium-sized tasks of special interest to them. The Tier-2s, which comprise an estimated 40% of the overall CPU resources, are also foreseen to be the source of most of the simulated data and where most of the later-stage data analysis will take place.

The basic effectiveness of the grid hierarchy concept in a large-scale production setting is being shown clearly in the large-scale grid-based production operations of the LHC experiments, in partnership with the major grid projects OSG and EGEE [19]. The increasing scale and efficiency of these operations supporting the LHC and other major HEP experiments, as well as other science communities, has been accompanied by an increasing efficiency and scale of network usage.

While the top-down picture of the hierarchical computing model and its use in the LHC service challenges has been relatively simple until now, effective use of the compute and storage resources at Tier-2s would benefit greatly from more opportunistic data distribution and local data access. There will therefore be a tendency towards more dynamic data flow among the Tiers, as a growing number of physics groups learn to use the production-oriented and standalone tools effectively. In the longer run, when the community of thousands of physicists will share both local and more remote resources to analyze their data, dynamic and efficient use of the network would enable the community to balance its resource usage, and to



make more effective use of local and regional resources where a group may have higher priority.

Refined View of the LHC Computing Model

At the start of LHC data-taking in 2007-2008, a typical Tier-2 site is expected to comprise of order 500-1000 kSi2000^a of CPU power, and 100-300 TBytes of useable disk space for each experiment served.

Given the scale and nature of storage at the Tier-2s, none of the individual Tier-2 sites will have sufficient resources to host all the relevant data samples for its regional user community. Instead, there will be a need to dynamically move data and user applications among the collection of Tier-2 sites and the corresponding Tier-1 center (for example, the U.S. Tier-2s and U.S. Tier-1s at Fermilab or Brookhaven) in order to optimally exploit the physics potential of the experiment. Accordingly, there will be a corresponding need for the Tier-2 centers to be connected by high bandwidth networks.

Responding to this vision and the corresponding needs, four of the U.S. CMS Tier-2s have initiated the DISUN project, illustrated in Figure 4. The physics data caches depicted at the center of the ring in the figure are distributed across Tier-2 sites, and are made available to scientists as a managed and high-throughput data resource supported by high throughput data transport services which are currently under development. It is also important to note that the diagram is schematic. The European Tier-2s are connected via the GÉANT2 network

^a CINT2000 is a measure of compute-intensive integer performance; kSi2000 is units of a thousand times the CINT2000 metric. An Intel P4 Xeon at 2.8GHz is approximately 1 kSi2000. See www.spec.org

infrastructure, while the US Tier 1s and Tier 2s are interconnected via high-bandwidth peerings between ESnet and Internet2 at the major points of presence in Chicago (StarLight) and New York (MANLAN).

Nature of the Distributed Data Management and Analysis Systems

The LHC data management system has several characteristics that result in requirements for the network and its services.

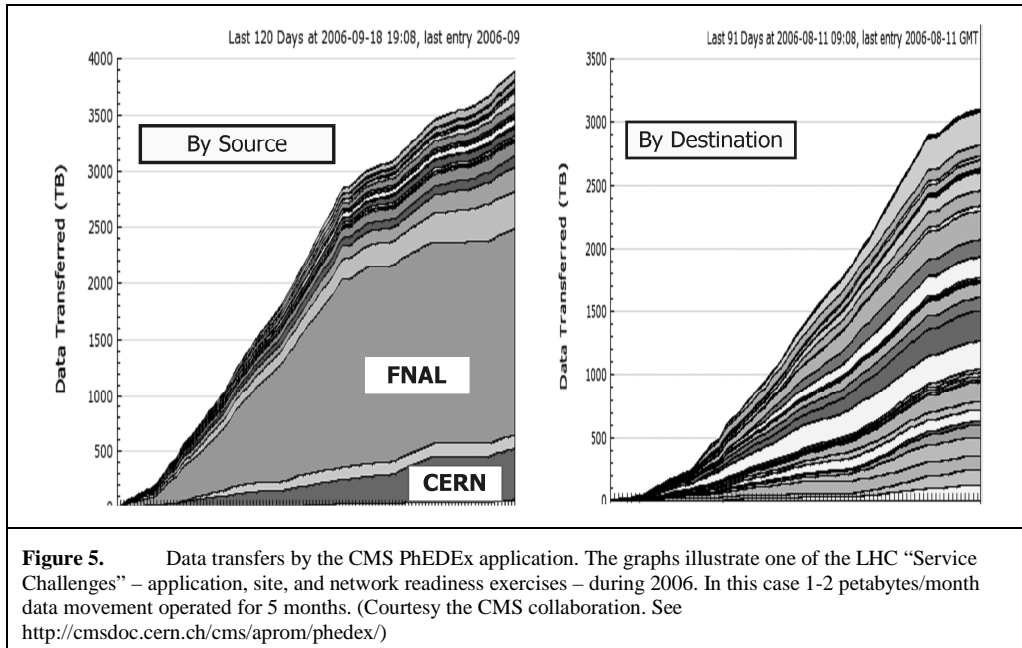
- The systems are widely distributed – typically spread over continental or inter-continental distances. The systems are data intensive and high-performance, typically moving terabytes a day for months at a time. (See Figure 5.)
- The systems are high duty-cycle, operating most of the day for months at a time in order to meet the requirements for data movement.
- Such systems clearly 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. In fact, they cannot be taken for granted even within a single well-run, high-capacity network.
- The applications must be able to get guarantees from the network that there is adequate bandwidth to accomplish the task at hand. The applications must be able to get information from the network that allows graceful failure and auto-recovery and adaptation to unexpected network conditions that are short of outright failure (which is much more common than complete failure).

In other words, the network has to behave like a service that provides guarantees and information to support recovery when the guarantees are not met. The application then must be capable of using such information to implement dynamic reconfiguration strategies and so on.

As more experience is gained with the current generation of applications and prototype network services, several things are becoming clear. One is that the network has inadequate tools to monitor the new services like virtual circuits (“VC”) and report back to the application in sufficient detail for the application to respond in an intelligent way. Another is that because VC services are relatively coarse-grained with respect to applications (VCs are typically set up between sites at this point), the application will have to share the bandwidth of a VC.

2.3. Network Requirements from Data and Collaboration Characteristics of DOE Office of Science Instruments, Facilities, and Science Practice

There are some 20 major instruments and facilities currently operated or being built by SC, plus the LHC (CERN, Switzerland) and ITER (France). To date, ESnet has characterized 14 of these for their future requirements. Facilities such as DOE’s big accelerators (RHIC at Brookhaven, SNS at Oakridge) and supercomputer centers (NERSC at Lawrence Berkeley, NLCF at Oak Ridge, and ALCF at Argonne), as well as the LHC at CERN, are typical of the hardware infrastructure of the science supported by the Office of Science. These facilities generate four types of network requirements: bandwidth, connectivity and geographic footprint, reliability, and network services.



In order to determine the requirements of SC science based on how the process of conducting scientific research will change, a set of case studies were developed in which the science communities were asked to describe how they expected to have to be doing their science in five and ten years in order to make significant progress. Computer scientists then worked with the scientists to translate the new processes into network requirements – in particular those related to collaboration, data sharing and remote analysis, remote instrument control, and large-scale simulations coupled with each other and/or with external sources of data (e.g., operating instruments)[2]. Bandwidth needs are determined by the quantity of data produced and the need to move the data for remote analysis. Connectivity and geographic footprint are determined by the location of the instruments and facilities, and the locations of the associated collaborative community, including remote and/or distributed computing and storage used in the analysis systems. These locations also establish requirements for connectivity to the network infrastructure that supports the collaborators (e.g., ESnet connectivity to Internet2 and the US regional R&E networks, and GÉANT and the European national R&E networks – the NRENs).

The reliability requirements are driven by how closely coupled the facility is with remote resources. For example, off-line data analysis – where an experiment runs and generates data and the data is analyzed after the fact – may be tolerant of some level of network outages. On the other hand, when remote operation or analysis must occur within the operating cycle time of an experiment (“on-line” analysis, e.g., in magnetic fusion experiments), or when other critical components depend on the connection (e.g., a distributed file system between supercomputer centers), then very little network downtime is acceptable. The reliability issue is critical and drives much of the design of the network. Many scientific facilities in which

DOE has invested hundreds of millions to billions of dollars, together with their large associated science communities, are heavily dependent on networking. Not surprisingly, when the experiments of these facilities depend on the network, then these facilities and scientists demand that the network provide very high availability (99.99+%), in addition to very high bandwidth.

The fourth requirement is in the area of types of service. In the past, networks typically provided a single network service – best-effort delivery of data packets^a – on which are built all of today’s higher-level applications (FTP, email, Web, socket libraries for application-to-application communication, etc.), and best-effort IP multicast (where a single outgoing packet is, sometimes unreliably, delivered to multiple receivers). In considering future uses of the network by the science community, several other network services have been identified as requirements, including bandwidth guarantees^b, traffic isolation^c, and reliable multicast.

Bandwidth guarantees are typically needed for on-line analysis, which always involves time constraints. Another type of application requiring bandwidth guarantees is distributed workflow systems such as those used by high energy physics data analysis. The inability of one element (computer) in the workflow system to adequately communicate data to another will ripple through the entire workflow environment, slowing down other participating systems as they wait for required intermediate results, thus reducing the overall effectiveness of the entire system.

^a Packet management by IP networks is not deterministic, but rather statistical. That is, the IP packets that make up, e.g., a TCP stream are injected into the network from many computers that are all connected to a single router – e.g. a typical large SC Lab will have many internal “subnets” all of which connect through different interfaces to a single site gateway router that provides connectivity to the outside world. The packets are queued in the router in whatever order they reach the routing processor (also called the forwarding processor). The packets in the queue waiting to be forwarded to their next-hop destination are intermixed indiscriminately by virtue of being queued immediately from several different input connections. As long as the queue does not overflow this is not an issue (in fact it is the norm) since every packet is routed through the network independently of every other packet. If the packets come into a router through several interfaces and they are all processed out through a single interface – as is typical, e.g., for a site gateway router that has several connections on the site side and a single connection on the Wide Area Network side – then it is possible for the forwarding processor to fall behind. This can happen either because the forwarding processor is not fast enough to keep up with the routing (which is rare in modern routers) or because the aggregate input traffic bandwidth exceeds the bandwidth of the single output interface (a circumstance that, in principle, is easily realized). When this happens the input queue for the forwarding engine will fill and “overflow” – this is called network congestion. The overflow process is a random discard of the incoming packets, and the overall effect is that there is no guarantee that a packet sent to a router is forwarded on to its next hop toward its destination – packet forwarding is a “best-effort” process. (Users typically see congestion as a slowdown in the network – they do not see the packet loss directly because most applications use TCP as a reliable transport protocol. TCP uses IP packets to move data through the network and it detects packet loss and automatically resends the lost IP packets in order to ensure reliable data delivery.)

^b Bandwidth guarantees are provided in IP networks by doing two things: First, the packets in a bandwidth-guaranteed connection are marked as high priority and are forwarded ahead of any waiting best-effort packet. Second, the bandwidth-guaranteed connections are managed so that, in aggregate, they never exceed the available bandwidth anywhere in the path to their destination. This entails limiting the input bandwidth of a bandwidth-guaranteed connection to an agreed upon value, and then by limiting the number of such connections so as not to exceed the available bandwidth along the path.

^c Traffic isolation is provided in a way similar to bandwidth guarantees in that the packets are queued and forwarded in such a way that they do not interact with other classes of traffic such as best-effort.

Traffic isolation is required because today's primary transport mechanism – TCP – is not ideal for transporting large amounts of data across large (e.g., intercontinental) distances. There are protocols better suited to this task, but these protocols are not compatible with the fair-sharing of TCP transport in a best-effort network, and are thus typically penalized by the network in ways that reduce their effectiveness. A service that can isolate the bulk data transport protocols from best-effort traffic is needed to address this problem.

Reliable multicast is a service that, while not entirely new, must be enhanced to increase its effectiveness. Multicast provides for delivering a single data stream to multiple destinations without having to replicate the entire stream at the source, as is the case, e.g., when using a separate TCP-based connection from the source to each receiver. This is important when the data to be delivered to multiple sites is too voluminous to be replicated at the source and sent to each receiving site individually. Today, IP multicast provides this capability in a fragile and limited way (IP multicast does not provide reliable delivery as TCP-based transport does). New services may be required to support reliable and robust multicast.

In the case studies that have been done to date [5], one or more major SC facilities have identified a requirement for each of these network capabilities.

The case studies of [2], [4], and [5] were picked both to get a good cross-section of SC science and to provide realistic predictions based on highly probable changes in the scientific process in the future. The case studies were conducted over several years and included the following Office of Science programs and associated facilities: Magnetic Fusion Energy, NERSC, ACLF, NLCF, Nuclear Physics (RHIC), Spallation Neutron Source, Advanced Light Source, Bioinformatics, Chemistry / Combustion, Climate Science, and High Energy Physics (LHC).

Summary of the conclusions of the case studies

There is a high level of correlation between network requirements for large and small-scale science – the primary difference being bandwidth – and so meeting the requirements of the large-scale stakeholders will generally provide for the requirements of the smaller ones, provided the required services set is the same.

Some of the non-bandwidth findings from the case studies included:

- The geographic extent and size of the user base of scientific collaboration is continuously expanding. As noted, DOE US and international collaborators rely on ESnet to reach DOE facilities, and DOE scientists rely on ESnet to reach non-DOE facilities nationally and internationally (e.g., LHC, ITER). Therefore, close collaboration with other networks is essential in order to provide high-quality end-to-end service, diagnostic transparency, etc.
- Robustness and stability (network reliability) are essential. Large-scale investment in science facilities and experiments makes network failure unacceptable when the experiments depend on the network.
- Science requires several advanced network services for different purposes. There are requirements for predictable latency and quality of service guarantees to support remote real-time instrument control, computational steering, and interactive

visualization. Bandwidth guarantees and traffic isolation are needed for large data transfers (potentially using TCP-unfriendly protocols), and network support for deadline scheduling of data transfers.

The aggregation of requirements from the 14 case studies (see [5]) results in:

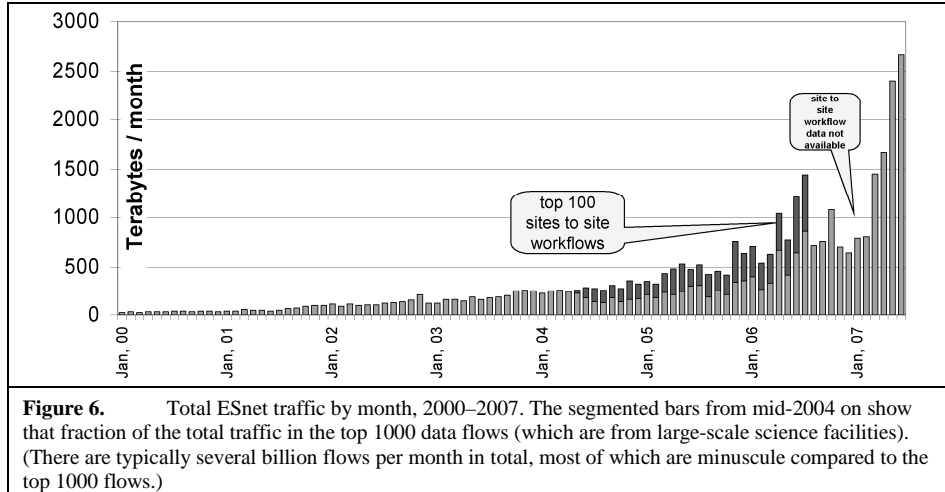
- Reliability
 - The Fusion requirements of 1 minute of down time during an experiment that runs 8–16 hours a day, 5–7 days a week, implies a network availability of 99.999%. LHC data transfers can only tolerate a small number of hours of outage in streams that operate continuously for 9 months per year, otherwise the analysis of the data coming from the LHC will fall too far behind to ever catch up. This implies a network availability of 99.95%.
 - These needs result in a requirement for redundancy (which is the only practical way to achieve this level of reliability) both for site connectivity and within ESnet.
- Connectivity
 - The geographic reach of the network must be equivalent to that of the scientific collaboration. Multiple peerings with the other major R&E networks are needed to add reliability and bandwidth for inter-domain connectivity. This is critical both within the US and internationally.
- Bandwidth
 - A bandwidth of 10 Gb/s site-to-site connectivity is needed now, and 100 Gb/s will be needed by 2010. Multiple 10 Gb/s peerings (interconnections) with the major R&E networks will be needed for data transfers. The network must have the ability to easily deploy additional 10 Gb/s circuits and peerings as needed by new science projects.

Bandwidth and service guarantees are needed end-to-end, so all R&E networks must interoperate as one seamless fabric. Flexible rate bandwidth guarantees are needed – that is, a project must be able to ask for the amount of bandwidth that it needs and not be forced to use more or less.

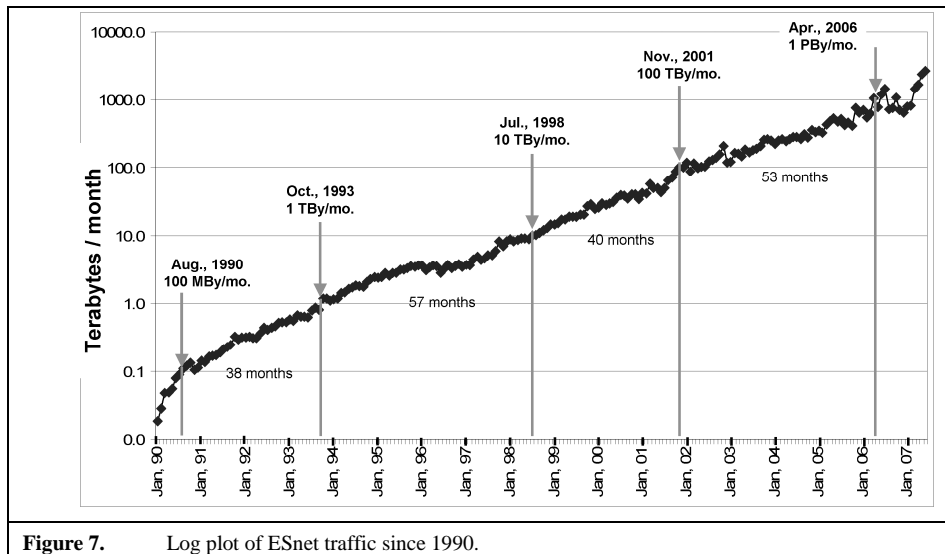
The case studies include both quantitative and qualitative requirements.

2.4. Requirements from Observing Traffic Patterns

From the analysis of historical traffic patterns, several clear trends emerge that result in requirements for the evolution of the network so it can handle the projected traffic load.



The first, and most obvious, pattern is the exponential growth of the total traffic handled by ESnet (Figure 6 and Figure 7). This traffic trend represents a 10x increase every 47 months on average since 1990 (Figure 7). ESnet traffic just passed the 1 petabyte per month level with about 1.5 Gb/s average, steady-state load on the New York-Chicago-San Francisco path. If this trend continues (and all indications are that it will accelerate), the network must be provisioned to handle an average of 15 Gb/s in four years. This implies a minimum backbone bandwidth of 20 Gb/s, because the network peak capacity must be at least 40% higher than the average load in order for today’s protocols to function properly with bursty traffic (which is the norm). In addition, the current traffic trend suggests that 200 Gb/s of core network bandwidth will be



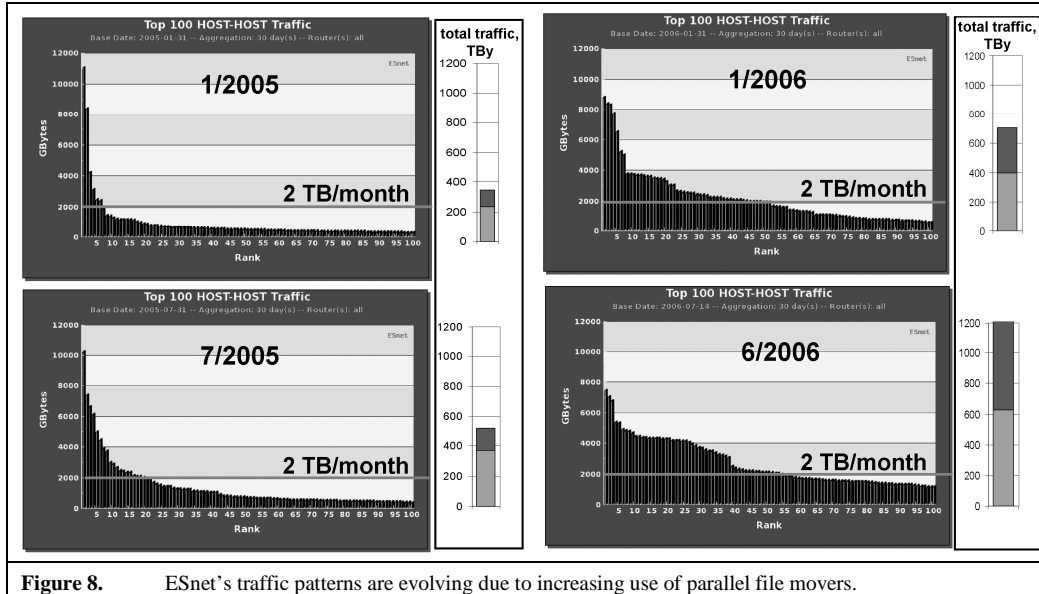


Figure 8. ESnet’s traffic patterns are evolving due to increasing use of parallel file movers.

required in eight years. This can only be achieved within a reasonable budget by using a network architecture and implementation approach that allows for cost-effective scaling of hub-to-hub circuit bandwidth.

The second major change in traffic is the result of a dramatic increase in the use of parallel file mover applications (e.g., GridFTP). This has resulted in the most profound change in traffic patterns in the history of ESnet. Over the past two years, this has resulted in a change from the historical trend where the peak system-to-system (“workflow”) bandwidth of the largest network users increased along with the increases in total network traffic, to a situation where the peak bandwidth of the largest user systems is coming down, and the number of flows that they generate is going up, while the total traffic continues to increase exponentially. This reduction in peak workflow bandwidth, together with an overall increase in bandwidth, is the result of the decomposition of single large flows into many smaller parallel flows. In other words, the same types of changes that happened in computational algorithms as parallel computing systems became prevalent are now happening in data movement – that is, parallel I/O channels operating across the network. This is illustrated in Figure 8, where the top 100 host-to-host data transfers, in one month averages, for a sampling of months over the past 18 months, are represented in the bar charts labeled “Host to Host Traffic.” (The “stair-step” appearance arises from groups of associated parallel file movers that move approximately the same amount of data while operating.) Next to these graphs is the total network traffic for that month, segmented as in Figure 6.

The third clear traffic trend is that over the past two years the impact of the top few hundred workflows – there are of order 6×10^9 flows per month in total – has grown from negligible before mid-2004 to more than 50% of all traffic in ESnet by mid-2006! This is

illustrated in Figure 6, where the top part of the traffic bars shows the portion of the total generated by the top 100 hosts.

The fourth significant pattern comes from looking at the source and destination locations of the top data transfer systems – an examination that shows two things. First is that the vast majority of the transfers can easily be identified as science traffic since the transfers are between two scientific institutions with systems that are named in ways that reflect the name of the science group. Second, for the past several years the majority of the large data transfers have been between institutions in the US and Europe and Japan, reflecting the strongly international character of large science collaborations organized around large scientific instruments (Figure 9).

Finally, Figure 9—only somewhat jokingly referred to as the “onslaught of the LHC” – also illustrates the limitation of using traffic trends alone to predict the future network needs of science. No traffic observations could have predicted the upsurge in LHC data movement, both from CERN to the SC Labs and from the SC Labs to US universities. Obviously traffic trend analysis cannot predict the start of new science projects.

2.5. Network Requirements Summary

The combination of the case studies and the traffic pattern trends adds quantitative aspects to the general requirements that were identified early in this paper.

The aggregate network capacity must reach 100–200 Gb/s in the five- to seven-year time frame. Network reliability must increase from the historical 99.9% to 99.99% to something more like 99.99% to 99.999% availability to the end site. The peerings – external network interconnections between national R&E and international R&E networks and ESnet – must increase both in bandwidth and reliability in a similar fashion. In addition, several specific new network services related to bandwidth guarantees must be introduced into the production network.

A general requirement is that there must be flexibility in provisioning the network capacity. The location of the greatest need for bandwidth within the network will change over time, and the budgetary resources available for the network may also change. It must be possible add and move hub-to-hub capacity as needed and to deploy new capacity on a schedule determined by science needs and funding availability.

3. Enabling Future Science: ESnet’s Evolution over the Next 10 Years

Based both on the projections of the science programs and the changes in observed network traffic and patterns over the past few years, it is clear that the network must evolve substantially in order to meet the needs of DOE’s Office of Science.

The current trend in traffic patterns – the large-scale science projects giving rise to the top 100 data flows that represent about 1/2 of all network traffic – will continue to evolve. As the LHC experiments ramp up in 2006-07, the data to the Tier-1 centers (FNAL and BNL) will increase 200-2000 times. A comparable amount of data will flow out of the Tier-1 centers to

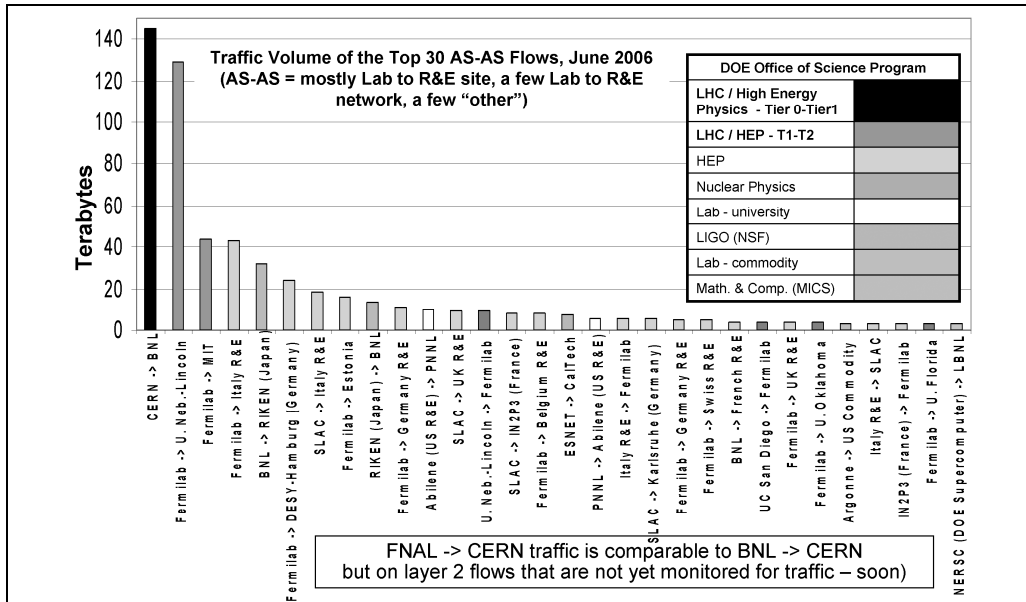


Figure 9. Traffic patterns due to new uses of the network by the LHC. LHC to BNL is the No. 1 traffic generator; FNAL to and from US universities accounts for Nos. 2, 3, 13, 23, 24, and 28.

the Tier-2 centers (U.S. universities) for data analysis. The DOE National Leadership Class Facility supercomputer at ORNL anticipates a new model of computing in which simulation tasks are distributed between the central facility and a collection of remote “end stations” that will generate substantial network traffic. As climate models achieve the sophistication and accuracy anticipated in the next few years, the amount of climate data that will move into and out of the NERSC center will increase dramatically (they are already in the top 100 workflows) Similarly, the experiment facilities at the new Spallation Neutron Source and Magnetic Fusion Energy facilities will start using the network in ways that require fairly high bandwidth with guaranteed quality of service.

This evolution in traffic patterns and volume will result in the top 100 - 1000 flows accounting for a very large fraction of all the traffic in the network, even as total ESnet traffic volume grows: The large-scale science data flows will overwhelm everything else on the network.

By 2009/2010 the current, few gigabits/sec of average traffic on the backbone will increase to 40 Gb/s (LHC traffic) and then increase to probably double that amount as the other science disciplines move into a collaborative production simulation and data analysis mode on a scale similar to the LHC. This will get the backbone traffic to 100 Gb/s in 2010-2012 as predicted by the science requirements analysis three years ago.

The old ESnet hub and spoke architecture (through 2004) would not have let ESnet meet these new requirements. The current core ring cannot be scaled to handle the anticipated large

science data flows at affordable cost. Point-to-point, commercial telecom tail circuits to sites are neither reliable nor scalable to the required bandwidth.

3.1. ESnet4: A New Architecture to Meet the Science Requirements

In order to accommodate this growth, and the change in the types of traffic, the architecture of the network must change. The general requirements for the new architecture are that it provide:

- High-speed, scalable, and reliable production IP networking, connectivity for University and international collaboration, highly reliable site connectivity to support Lab operations as well as science, and Global Internet connectivity
- Support for the high bandwidth data flows of large-scale science including scalable, reliable, and very high-speed network connectivity to DOE Labs
- Dynamically provisioned, virtual circuits with guaranteed quality of service (e.g. for dedicated bandwidth and for traffic isolation)

In order to meet these requirements, the capacity and connectivity of the network must increase to include fully redundant connectivity for every site, high-speed access to the core for every site (at least 20 Gb/s, generally, and 40-100 Gb/s for some sites) and a 100 Gb/s national core/backbone bandwidth by 2009/2010 in two independent backbones.

The strategy for the next-generation ESnet is based on a set of architectural principles that lead to four major network elements and a new network service for managing large data flows.

The architectural principles are:

- Use ring topologies for path redundancy in every part of the network – not just in the WAN core.
- Provide multiple, independent connections everywhere to guard against hardware and fiber failures.
- Provision one core network – the IP network – specialized for handling the huge number (3×10^9 /mo.) of small data flows (hundreds to thousands of bytes each) of the general IP traffic.
- Provision a second core network – the Science Data Network (SDN) – specialized for the relatively small number (hundreds to thousands) of massive data flows (gigabytes to terabytes each) of large-scale science (which by volume already accounts for 50% of all ESnet traffic and will completely dominate it in the near future).

These architecture principles lead to four major elements for building the new network:

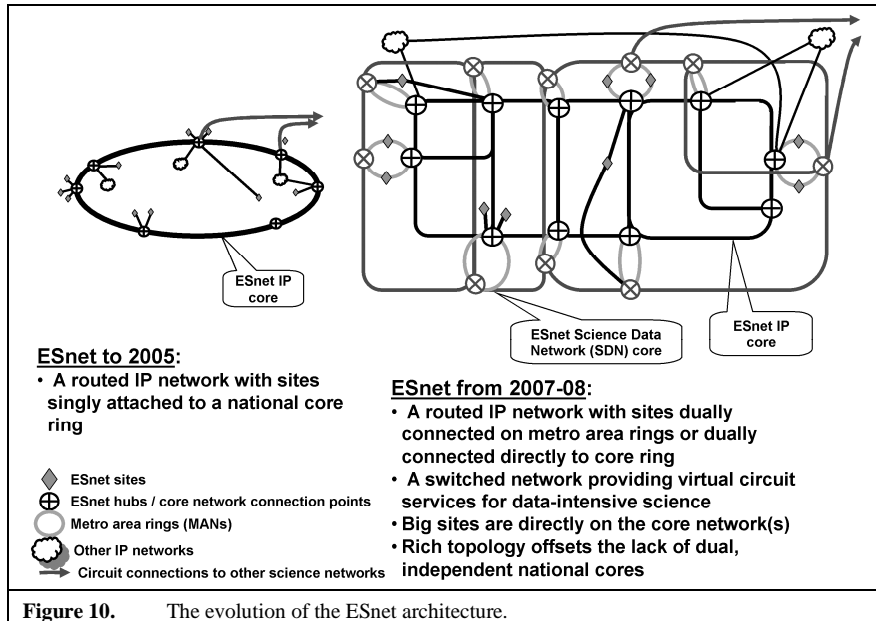
- A high-reliability IP core network based on high-speed, highly capable IP routers to support:
 - Internet access for both science and lab operational traffic, and some backup for the science data carried by SDN
 - science collaboration services
 - peering with all of the networks needed for reliable access to the global Internet.

- A Science Data Network core network based on Ethernet switches that support Multi-Protocol Label Switching (MPLS) and/or layer 1^a (optical) switches for:
 - multiple 10 Gb/s circuits with a rich topology for very high total bandwidth to support large-scale science traffic and for the redundancy needed to high reliability
 - dynamically provisioned, guaranteed bandwidth circuits to manage large, high-speed science data flows
 - dynamic sharing of some optical paths with the R&E community for managing peak traffic situations and for providing specialized services such as all-optical, end-to-end paths for uses that do not yet have encapsulation interfaces (e.g. Infiniband)
 - an alternate path for production IP traffic.
- Metropolitan Area Network (MAN) rings connecting labs to the core(s) to provide:
 - more reliable (ring) and higher bandwidth (multiple 10 Gb/s circuits) site-to-core connectivity
 - support for both production IP and large-scale science traffic
 - multiple connections between the Science Data Network core, the IP core, and the sites.
- Loops off the core rings to provide for dual connections to remote sites where MANs are not practical

These elements are structured to provide a network with fully redundant paths for all of the SC Labs. The IP and SDN cores are independent of each other and both are ring-structured for resiliency. These two national cores are interconnected at several locations with ring-structured metropolitan area networks that also incorporate the DOE Labs into the ring. This will eliminate all single points of failure except where multiple fibers may be in the same conduit (as is frequently the case between metropolitan area points of presence and the physical sites). In the places where metropolitan rings are not practical (e.g. the geographically isolated Labs) resiliency is obtained with dual connections to one of the core rings. (See Figure 10.)

The theoretical advantages of this architecture are clear but it must also be practical to realize in an implementation. That is, how does ESnet get to the 100 Gb/s multiple backbones and the 20-40 Gb/s redundant site connectivity that is needed by the SC community in the 3-5 yr time frame?

^a The “layer” term refers to the Open Systems Interconnect (OSI) standard model. Very briefly, layer 1 refers to the sending and receiving bits at the optical or electrical interface. Layer 2 refers to how a computer gets access to a network – e.g. via an Ethernet interface. Layer 3 refers to routing and switching (e.g. IP routers) and layer 4 refers to data transport (e.g. TCP). The OSI model does not map perfectly onto the IP model, but the terms are used anyway. Likewise referring to an Ethernet switch as a “layer 2” device and an IP router as a “layer 3” is not strictly accurate since almost all modern Ethernet switches can do some IP routing and almost all IP routers can do some Ethernet switching. Again, however, the terms are used anyway.



3.2. Building ESnet4

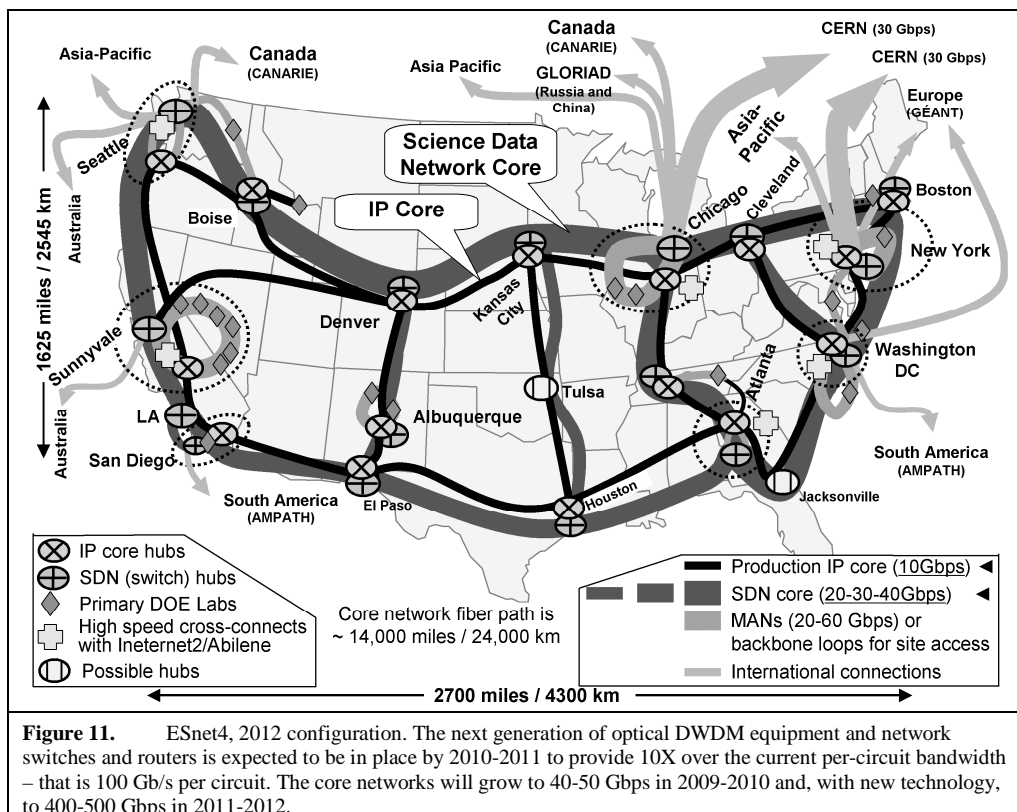
Internet2 – the network that serves the US R&E community–has partnered with Level 3 Communications Co. and Infinera Corp. to build a dedicated optical fiber infrastructure with a national footprint and a rich topology - the “Internet2 Network.”

The fiber will be provisioned with Infinera Dense Wave Division Multiplexing equipment that uses an advanced, integrated optical-electrical design. Level 3 will maintain the fiber and the DWDM equipment as part of its commercial network–a very important consideration for reliability. The DWDM equipment will initially be provisioned to provide 10 optical circuits (lambdas or waves) across the entire fiber footprint (40 waves is the current configuration capacity, 80 is maximum.)

ESnet has partnered with Internet2 to:

- Share the optical infrastructure
- Develop new circuit-oriented network services
- Explore mechanisms that could be used for the ESnet Network Operations Center (NOC) and the Internet2/Indiana University NOC to back each other up for disaster recovery purposes

ESnet will build its next generation IP network and its new circuit-oriented Science Data Network primarily on Internet2 optical circuits that are dedicated to ESnet, together with a few from National Lambda Rail and others. ESnet will provision and operate its own routing and switching hardware that is installed in various commercial telecom hubs around the country, as it has done for the past 20 years. ESnet’s peering relationships with the commercial Internet,



various US research and education networks, and numerous international networks will continue and evolve as they have for the past 20 years.

ESnet4 will also involve an expansion of the multi-10Gb/s Metropolitan Area Rings in the San Francisco Bay Area, Chicago, Long Island, Newport News (VA/Washington, DC area), and Atlanta to provide multiple, independent connections for ESnet sites to the ESnet core network. (Building the Metropolitan Area Networks that get the Labs to the ESnet cores is a mixed bag and somewhat opportunistic – a combination of R&E networks, dark fiber networks, and commercial managed lambda circuits are used.) In fact, in the new architecture all of the big SC Labs are effectively connected directly to both the IP and SDN core networks.

3.3. New Network Services

New network services are also critical for ESnet to meet the needs of large-scale science. (See [2], [3], [4], and [5].)

Dynamically provisioned virtual circuits that provide traffic isolation are needed to enable the use of non-standard transport mechanisms that cannot co-exist with TCP-based transport and provide guaranteed bandwidth.

Guaranteed bandwidth was identified as very important in three specific situations.

The first situation is that it is the only way that we currently have to address deadline scheduling – e.g. where fixed amounts of data have to reach sites on a fixed schedule in order that the processing does not fall so far behind that it could never catch up. This is very important for certain experiment’s data analysis

The second situation is where remote computing elements are involved in control of real-time experiments. Two examples of this were cited in the applications requirements workshop [2] – one from magnetic fusion experiments and the other from the Spallation Neutron Source. The magnetic fusion situation is that theories are tested with experiments in Tokamak fusion reactors. The experiments involve changing the many parameters by which the reactor can operate and then triggering plasma generation. The “shot” (experiment) lasts a few 10s of milliseconds and generates hundreds of megabytes of data. The device takes about 20 minutes to cycle for the next shot. In that 20 minutes the data must be distributed to the remote collaborators, analyzed, and the results of the analysis fed back to the reactor in order to set up the next experiment (shot). In order to have enough time to analyze the data and use the parameters to set up the next experiment, 200-500 Mb/s of bandwidth must be guaranteed for 2-5 minutes to transmit the data and leave enough time to do that analysis. The situation with the SNS is similar.

The third situation is when Grid based analysis systems consist of hundreds of clusters at dozens of universities that must operate under the control of a workflow manager that choreographs complex workflows. This requires network quality of service to ensure a steady flow of data and intermediate results among the systems. Without this, systems with many inter-dependencies could stop and start, causing interruptions that would propagate throughout the entire collection of systems. This would create an unstable and inefficient production environment that would reduce the overall throughput necessary to keep up with the steady generation of data by the experiment. (This is of particular concern with the huge amount of data coming out of the LHC experiments.)

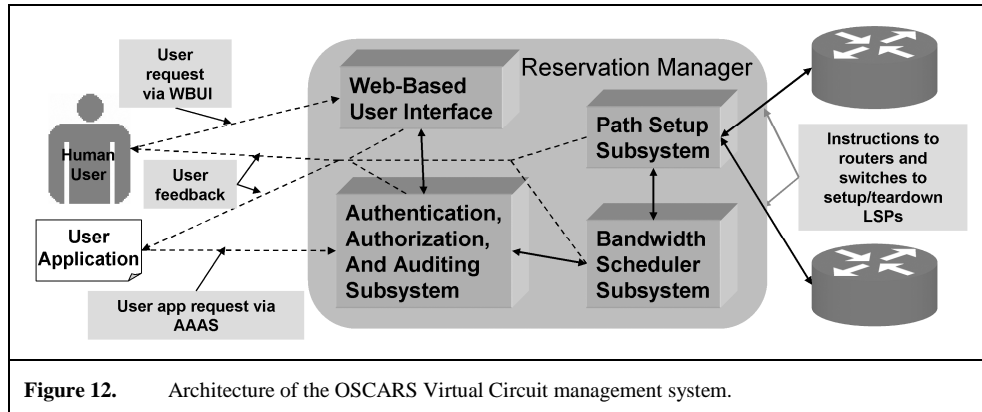
In addition to virtual circuits, another new network service that is essential is an end-to-end monitoring service. As cross-domain virtual circuit services start to be deployed, monitoring is seen as a critical service that is needed both for network operators and users.

4. Development and Deployment of Service-Oriented Communication Services

DOE SC has funded the OSCARS (On-demand Secure Circuits and Advance Reservation System) project to develop and deploy the various technologies that provide dynamically provisioned circuits and quality-of-service (QoS) that can be integrated into a production network environment. Such “circuits” are called “virtual circuits” (VCs) because that are defined in software and thus are mutable (as opposed to hardware established circuits).

The elements of this system (illustrated in Figure 12) are the

- Web-Based User Interface (WBUI) that will prompt the user for a username/password and forward it to the AAAS (Authentication, Authorization, and Auditing Subsystem)



- Authentication, Authorization, and Auditing Subsystem that will authenticate users, handle access authorization, enforce policy, and generate usage records
- Bandwidth Scheduler Subsystem (BSS) that will track reservations and map the state of the network (present and future)
- Path Setup Subsystem (PSS) that will setup and teardown the on-demand paths (VCs)

The end-to-end provisioning of VCs will initially be provided by a combination of Ethernet switch management of optical channel circuits in the MANs and Ethernet VLANs managed as MPLS paths (Multi-Protocol Label Switching and Label Switched Paths - LSPs) in the SDN core and as MPLS VCs in the IP network.

There are two realms in which OSCARS must operate: 1) intra-domain – that is, to establish a schedulable, guaranteed bandwidth circuit service within the boundary of the ESnet network; 2) inter-domain – e.g. to provide end-to-end QoS between DOE Labs and US and European universities.

Setting up inter-domain guaranteed bandwidth circuits is not a trivial task. It typically involves the virtual circuit extending across five to seven autonomous networks: the lab/campus network at each end, the lab/campus service provider (e.g. ESnet, a US RON (Regional Optical Network), or a European NREN) and the US national or pan-European transit network (e.g. ESnet, Internet2, GÉANT) or SINet (Japan). Differences in network infrastructure (e.g. hardware, link capacity, etc.) must be addressed at the inter-domain boundary in order to provide consistent service characteristics (e.g. bandwidth, delay, and jitter) across domains, as must the issues of different policies, such as Acceptable Use Policies (AUPs), Service Level Agreements (SLAs) and security requirements. Nevertheless, inter-domain circuits are essential, especially between ESnet, Internet2, and GÉANT. (Note that OSCARS does not address the important issue of inter-domain brokering policies. Enforcement of such policies, however, are critical to the deployment of OSCARS as a production service. Collaborative work is being done with the GÉANT, Joint Research Activity 5 project to ensure a compatible authentication/authorization framework.)

In the absence of agreed upon standards for the inter-domain interface (called an “ENNI” – external network-network interface) the community is ensuring interoperability by

collaboratively developing the software. This collaboration currently involves joint code development with the Internet2 BRUW project, and is working with HOPI (Internet2), TeraPaths (Brookhaven Lab), and DRAGON (an NSF-funded project) to ensure interoperability between each of these projects. OSCARS is also working with HOPI (Internet2), JRA5 (GÉANT's Joint Research Activity 5 project) to define an appropriate and interoperable AAI framework. OSCARS is working with DICE Control Plane group to determine schemas and methods of distributing topology and reachability information, multi-domain scheduling, and inter-domain signaling (DICE=Internet2, ESnet, GÉANT, CANARIE/UCLP; see <http://www.garr.it/dice/presentation.htm>); and working with Tom Lehman (DRAGON), Nagi Rao (USN), Nasir Ghani (Tennessee Tech) on multi-level, multi-domain hybrid network performance measurements. A number of OSCARS circuits are currently being tested between various institutions.

For more information on the OSCARS implementation see <http://www.es.net/oscars>.

5. The Critical Role of Monitoring and Reporting

In order to build large-scale, widely distributed systems that operate reliably to perform complex data analysis (cf. LHC Case Study, above) or computational simulation tasks, the distributed applications and middleware must be able to learn, in real-time, about unexpected changes in the state of the communication between all of its components. Without this capability human users or system operators are left trying to intuit what has gone wrong. A problem that appears to come from one component may actually be an unreported communications problems from a very different part of the system. A reliable network monitoring service that describes the current state of application communications allows applications to adapt their behavior to changing circumstances, or at least to fail gracefully and accurately announce why it is failing.

An essential change in network services over the next five years will be to provide reliable, comprehensive, timely, and interpretable information about the state of all networks components in the end-to-end path in a manner that can be meaningfully interpreted and used by user-level applications. This ability must be accompanied by a corresponding capability in the applications and middleware to accept the communication services monitoring results and do something intelligent with those results. This may include adapting the functioning of the system to the changed / diminished communication service capability, graceful shutdown of the system, notifying the user what is happening (in terms that are useful to the users involved), and so on. The monitor results must be presented in a way that is meaningful to the user's view of the network.

Together with the new capabilities provided by virtual circuits, monitoring services that can report problems directly to the networked applications and users^a will move network communications toward a managed service model more like the computing environment provides.

^a We will use the term "user" to interchangeably mean an application agent or service or a human user.

5.1. Background

All networks do extensive real-time monitoring which is used for a variety of uses. Short-term monitoring (on the order of minutes) is used for identification and debugging of problems in every element of the network – circuits, interfaces, switching and routing equipment, routing state (logical connectivity), and so on. This monitoring is primarily used to detect failure or failure onset through degraded performance or some aspect of the many network element health indicators. ESnet, for example, monitors almost 5,000 network element characteristics in real time in its national network. A commercial network monitoring system (Spectrum) is used to manage this information, generate operator alerts, and so on.

Intermediate term (hours to days) interface traffic monitoring is done for capacity management: Hotspots can develop in the network due to changes in the user demand or capability, changes in network capacity (augments, or outages), or routing changes. It may be possible to address these hotspots by configuration (routing) changes, as networks are growing more densely meshed internally and more richly connected to each other. This sort of information could also trigger physical reconfiguration of parts of the network – typically by increasing interface bandwidth when possible.

Long term (months to years) traffic trend monitoring supports planning future network configurations, etc.: Traffic trends that show up over months or years (e.g. Figure 6, Figure 7, Figure 8, and Figure 9) are essential in planning future architecture changes and major upgrades that will occur years in the future. These are one of the several metrics that drive the design of the next generation of the network.

Typically, detailed (minute-level granularity) network interface usage is available on-line for about a month and is then archived for future reference. Summary information (monitor data summarized at hourly, daily, or weekly granularity) is available on-line for several years. (ESnet, for example, monitors almost a thousand logical network interfaces on 64 routers and switches, and collects and archives about 325 GBy/month of monitor data.)

5.2. Network Monitoring System Design Goals

Detailed real-time network link state and performance data is routinely collected and archived in almost all production networks. However, what is of interest to the network operators is the behavior of specific router or switch interfaces and the link connecting them. Therefore the data is collected and data archives are organized and indexed in this fashion. Further, the form of this data is typically peculiar to each network, making the information almost useless to the user trying to see end-to-end behavior. In order to be useful to the user for end-to-end monitoring, the information must satisfy an additional set of requirements.

There must be tools to map the user view (as represented, e.g., by a traceroute of the application-application path) to the network view and then collect and map the corresponding network monitoring data back to the user view. That is, the tools must convert the user view to the physical path representation – the sets of interfaces and links that comprise the path at the physical level; extract the related data from the archive; map it back to the user view; and return the results to the user in a format that is standard across all networks.

Further, the entire end-to-end path must be included in the monitoring. In a typical R&E environment such paths involve five to six network domains: the site LAN, the regional or national network, a second national or a pan-national network, back into a regional or national network, and into the site LAN at the other end of the path. Each of these domains must provide the data for the segments of the user path that are part of that domain. This sort of cross-domain monitoring is critical both for high-performance applications that depend on widely distributed components and for network operators who are increasingly required to manage end-to-end paths.

A user should be able to be notified of service outages by subscribing to alerts for a given application path. Further, the report should provide information about the source of the outage – is it due to congestion (to which the user may be contributing), or link errors which is a network problem, or some other problem. Currently network operators log planned outages in a calendar system and this system must also report future outages to the user.

Again, the problem with this from the user point of view is that the descriptions are given in terms of the physical topology of the network. To be useful to the user, physical topology must be mapped into user path descriptions and point failures must be reported in terms of their impact on the user path.

5.3. New Monitoring Services

perfSONAR^a

PerfSONAR is intended as a significant first step in cross-domain monitoring by both network operators and users.

Quoting from the perfSONAR Web site (www.perfsonar.net):

PerfSONAR has three contexts:

1) perfSONAR is first a consortium of organizations who seek to build network performance middleware that is interoperable across multiple networks and useful for intra- and inter-network analysis. One of the main goals is to make it easier to solve end-to-end performance problems on paths crossing several networks.

2) perfSONAR is a protocol. It assumes a set of roles (the various service types), defines the protocol standard (syntax and semantics) by which they communicate, and allows anyone to write a service playing one of those roles. The protocol is based on SOAP XML messages and following the Open Grid Forum (OGF) Network Measurement Working Group (NM-WG).

^a This section draws on Hanemann, A., Boote, J. W., Boyd, E. L., Durand, J., Kudarimoti, L., Lapacz, R., Swany, D. M., Zurawski, J., Trocha, S., "PerfSONAR: A Service Oriented Architecture for Multi-Domain Network Monitoring", In "Proceedings of the Third International Conference on Service Oriented Computing", Springer Verlag, LNCS 3826, pp. 241–254, ACM Sigsoft and Sigweb, Amsterdam, The Netherlands, December, 2005 and Hanemann, A., Liakopoulos, A., Molina, M., Swany, D. M., "A Study on Network Performance Metrics and their Composition" TERENA Networking Conference 2006. - download; also appeared in special edition of Campus-Wide Information Systems (Volume 23 – 4 – 2006 – ISSN 1065-0741), Emerald Publishing Group Ltd. For more information see these and other papers at www.perfsonar.net.

3) *perfSONAR* is, finally, an example set of code (implementation of services) that attempts to implement an interoperable performance middleware framework. Those sets of code are developed by different partners. Some pieces of code are "more important" than others because their goal is to ensure interoperability between domains (e.g. the Lookup Service and the Authentication Service). Different subsets of code are important to each partner, with a great deal of overlap. The services developed acts as an intermediate layer, between the performance measurement tools and the diagnostic or visualization applications.

Functionality: In order to satisfy the needs of the various communities of users of network data – the network operators and engineers, the network support staff at the institutions of the end users, and the end users both in the process of debugging the performance of a distributed application or as part of a service that reports network problems to an application resource manager – there are several aspects of network monitoring that must be addressed.

There are three general categories of performance measurement data – active measurements, passive measurements, and network state variables (SNMP variables) – that can be thought of as data producers. From the network data user's point of view this data must be available in various ways and must have various services associated with it both to homogenize the information from different networks and to present the data in useful ways. Data should be provided as a data flow or via polling.

The analysis tools, threshold alarms, and visualization tools are data consumers that, in turn, need data that is already transformed in various ways. Therefore, between data producers and data consumers there may be a pipeline of aggregators, correlators, filters, and buffer services that can be regarded as data transformers and data archives.

Further, the services – the data producers, consumers, transformers, and archives – are all resources that need to be discovered and almost certainly used within an authentication and authorization framework that maintains the policy prescribed by the network operators that own the measurement data.

Architecture: A service oriented architecture (SOA) has been adopted by the community that consists of three layers and a collection of defined service functions. (See Figure 13.)

- The *Measurement Point layer* is the lowest layer of the architecture. It collects network measurements, transforms the results into a standard format, and publishes the information to a Measurement Archive, or other service.
- The *Service layer* includes data management, manipulation, and transformation services and a collection of "housekeeping" services that provide standard authentication and authorization, service discovery, etc. The service layer is not a simple in-and-out layer, but contains pipeline or compound services like the Measurement Archive are both a service and a consumer of services.
- The *Interface layer* provides the clients that produce human or application useful representations.

The Services: The currently extant services fall into seven categories:

- Measurement Point (MP) service: Creates and/or publishes monitoring information related to active and passive measurements
- Measurement Archive (MA) service: Stores and publishes monitoring information

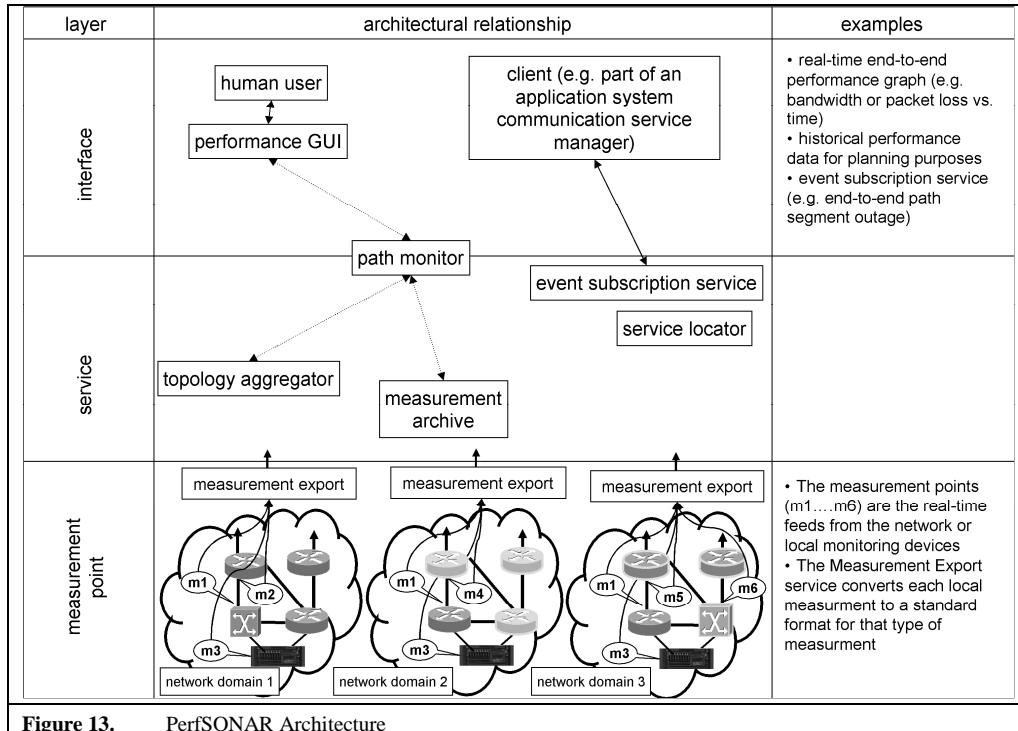
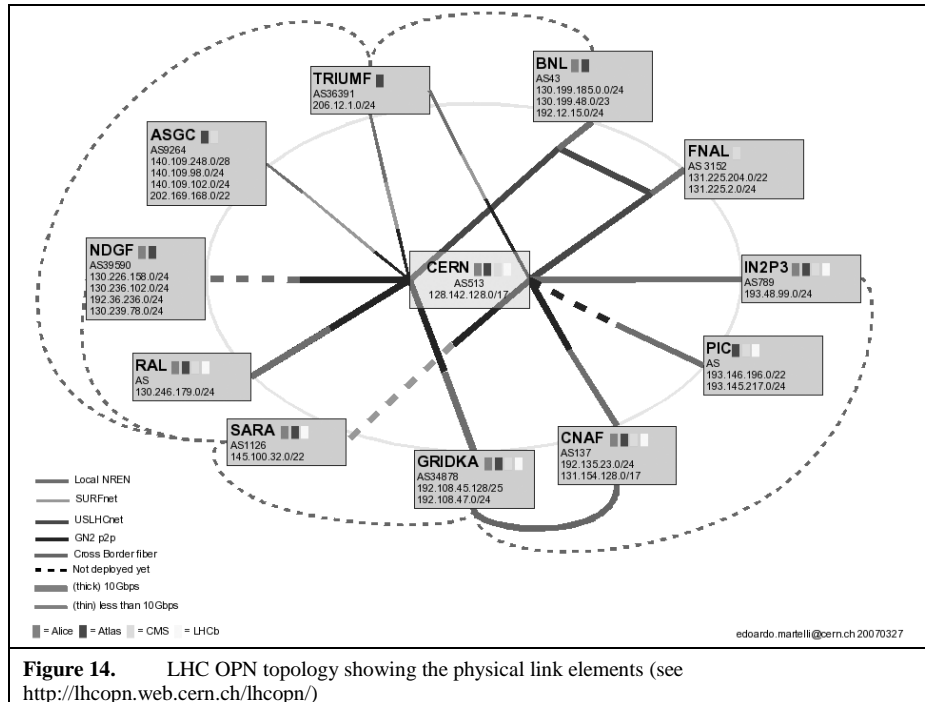


Figure 13. PerfSONAR Architecture

- Lookup service (LS): Registers all participating services and their capabilities
- Topology service (TS): provides network topology information
- Authentication service (AS): Manages domain-level access to services
- Transformation service (TrS): performs manipulation (aggregation, statistics) on available data sets
- Resource Protector (RP) service: arbitrates the use of limited measurement resources based on the policy of the resource owner

Use of the System: The Measurement Point (MP) services at the lowest layer create or collect network measurement data. Network operators frequently maintain exclusive management access to their network devices for operational and security reasons. Network operators can use the perfSONAR framework by deploying MP services that query their network devices for state information and push this information into Measurement Archive (MA) services. This provides an important data abstraction functionality by isolating the method used to obtain the data from the standardized perfSONAR data publication representation. This allows the middle layer of perfSONAR services to process and analyze data from different sources within one domain, or from sources across multiple domains, using a single standardized interface.

This architecture provides a clean separation between the policies regarding how the locally controlled MP accesses the network infrastructure, and the policies governing how internal and external perfSONAR services access the resulting data in the MA services. It has



other benefits such as allowing multiple consumers to share the same data thereby reducing the measurement load on the underlying system.

The middle layer of perfSONAR contains a set of cooperating services, including the Measurement Archive (MA), Lookup Service (LS), Topology Service (TS), Transformation service (TrS), and the Authentication service (AS). These services can be used individually, or together to provide uniform access to network measurements across multiple domains.

All services register their presence and capabilities with their local domain's LS. The LS's cooperate to function as a global registry across all domains. This allows the services to find each other within one domain, and it allows applications to find services across multiple domains. The LS allows MP's to locate MA's that can store their results. It allows user applications to locate the MA that contains data of interest.

The TS service supports automated analysis of the network by identifying the underlying structure in the networks and providing information about how multiple network domains are interconnected. This capability will be essential in future networking environments where circuit services will dynamically alter the underlying network infrastructure used by applications in real time.

The Measurement Archive (MA) can be configured to accept and store setup requests as well as publication requests. The publication request includes a subscription handle, and the results are sent directly to the client (or indirectly via a TrS). As a client, the MA registers its own presence with an LS, subscribes to an MP, other MA, or TS, and publishes measurement

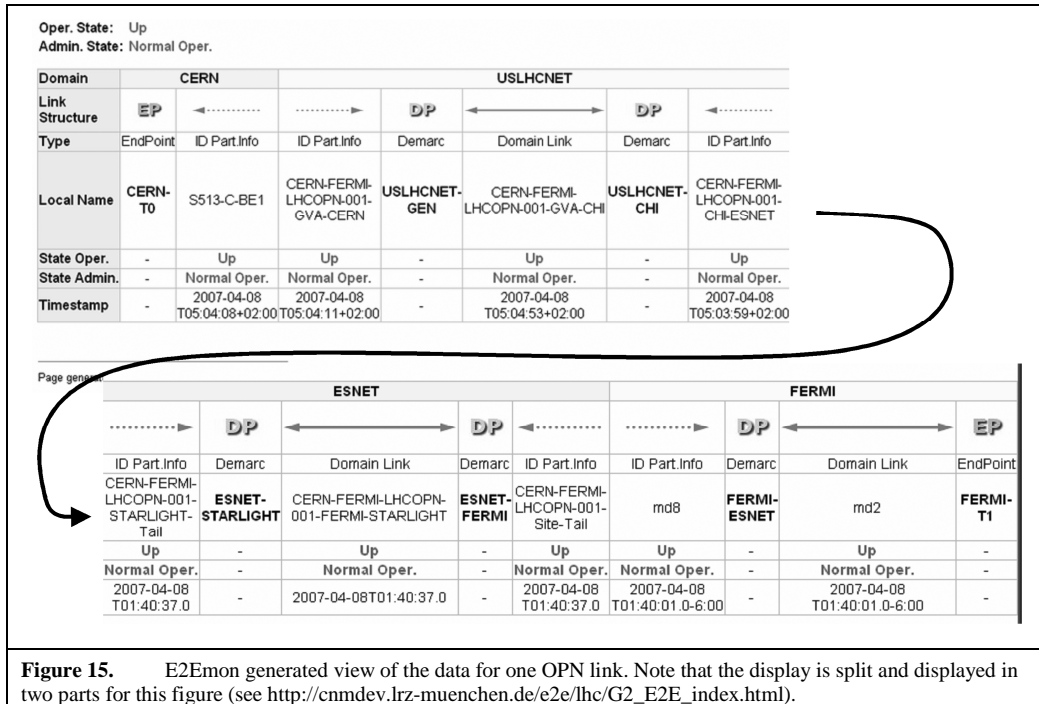


Figure 15. E2Emon generated view of the data for one OPN link. Note that the display is split and displayed in two parts for this figure (see http://cnmdev.lrz-muenchen.de/e2e/lhc/G2_E2E_index.html).

data to subscribers. The MA may send resource availability and authorization requests to the RP.

Multi-Domain Monitoring: The first production deployment of the perfSONAR framework is multi-domain monitoring for the LHC Optical Private Network (LHCOPN or OPN) network (Figure 14). LHCOPN is the network that transfers data from the LHC Tire-0 facility at CERN to the Tier-1 Data Centers in various countries.

In this case perfSONAR provides a set of conventions for representing network data in a common format, together with the SOA approach that allows the various component services of perfSONAR be used to assemble monitoring applications for different purposes.

perfSONAR MP services are deployed inside each network domain to monitor the links related to each domain's OPN. Some domains are providing real-time status information directly from their MP. Other domains have the MP store the data in a MA, which publishes both current and historical information.

The MP in each domain consists of two components. The domain specific component in the various networks typically interfaces with the operational network monitoring system to obtain the link status data for the portion of the end-to-end path within that particular network. Virtually every network does internal monitoring in a different way that has evolved historically along with the network. The perfSONAR component of each MP takes the resulting data, generates a standard XML file, and publishes it via the MP service interface, or

Start page E2ECU view All E2E Links Problem Links Domain view CERN DFN ESNET FERMI GARR GEANT2 IN2P3 RENATER SURFNET SWITCH USLHCNET Project view IGTMD LHCOPN Availability Statistics Current Month Last Month	All E2E Links			
	E2E Link ID	State Oper	State Admin	Additional Info
	CERN-FERMI-LHCOPN-001	Up	Normal Oper.	
	FERMI-IN2P3-IGTMD-002	Down	Normal Oper.	Error: E2E Link is not contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	GRIDKA-SARA-LHCOPN-001	Up	Normal Oper.	Error: E2E Link is not contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	FERMI-IN2P3-IGTMD-001	Down	Trouble Shooting	Error: E2E Link is not contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	GRIDKA-IN2P3-LHCOPN-001	Down	Normal Oper.	Error: E2E Link is not contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	CERN-BNL-LHCOPN-001	Up	Normal Oper.	
	CERN-SARA-LHCOPN-001	Up	Normal Oper.	
	CERN-GRIDKA-LHCOPN-001	Up	Normal Oper.	
	CERN-FERMI-LHCOPN-002	Up	Normal Oper.	Error: E2E Link is not contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	CERN-SARA-LHCOPN-004	Up	Normal Oper.	Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	CERN-CNAF-LHCOPN-001	Up	Normal Oper.	
	CERN-SARA-LHCOPN-002	Up	Normal Oper.	Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	CERN-RAL-LHCOPN-001	Up	Normal Oper.	Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	CERN-BNL-LHCOPN-002	Up	Normal Oper.	Error: E2E Link is not contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
	CERN-SARA-LHCOPN-003	Up	Normal Oper.	Warning: Operational state is known not for all involved links

Figure 16. E2Emon generated view of the data for all OPN links showing the operational state and administrative state of each link. (The first entry - CERN-FNAL – is the one line summary of the information presented in the view of Figure 15.) (see http://cnmdev.lrz-muenchen.de/e2e/lhc/G2_E2E_index.html)

pushes it to an MA for archiving and publishing. This information is used by an application called E2Emon^a.

E2Emon uses perfSONAR protocols to retrieve current circuit status every minute or so from MAs and MPs in all domains supporting the circuits.

E2Emon is itself a service that produces Web based, real-time displays of the overall state of the network, and it generates alarms when one of the MP or MA's reports link problems. The web interface for a single link is shown in Figure 15, and the OPN-wide view is shown in Figure 16. These tools are being used by the E2ECU (End to End Coordination Unit), which is a function of the GÉANT Network Operations Center that provides the overall management of the OPN circuits.

Another important multi-domain application of perfSONAR is for path performance monitoring. This presents not the just the operational state of the path as in the previous example, but also provide real-time performance such as path utilization and/or packet drop.

Multiple path performance monitoring tools are in development. One example – Traceroute Visualizer^b – has been deployed at about 10 R&E networks in the US and Europe

^a An application developed by the German R&E network DFN for monitoring circuits using perfSONAR protocols

^b <https://performance.es.net/cgi-bin/level0/perfsonar-trace.cgi>

path	domain
traceroute to www6.lnf.infn.it (193.206.84.223), 64 hops max, 40 byte packets	
1 ir1000gw (131.243.2.1) 0.340 ms 0.306 ms 0.271 ms	LBNL LAN
2 er1kgw (131.243.128.5) 2.325 ms 2.551 ms 1.885 ms	
3 lbl2-ge-lbnl.es.net (198.129.224.2) 198.665 ms 1.261 ms 1.447 ms	
4 slacmr1-lblmr1.es.net (134.55.219.10) 1.598 ms 1.409 ms 1.451 ms	ESnet (SF Bay MAN)
5 snv2mr1-slacmr1.es.net (134.55.217.2) 1.886 ms 1.795 ms 1.739 ms	
6 snv2sdn1-snv2mr1.es.net (134.55.207.37) 1.740 ms 1.896 ms 1.742 ms	ESnet WAN core
7 denvcr1-snv2sdn1.es.net (134.55.220.49) 46.050 ms 29.245 ms 28.937 ms	
8 chiccr1-denvcr1.es.net (134.55.209.46) 52.483 ms 52.495 ms 52.484 ms	
9 chiccdn1-a-chiccr1.es.net (134.55.218.101) 52.482 ms 52.496 ms 52.486 ms	GEANT WAN core
10 washsdn1-chiccdn1.es.net (134.55.218.98) 69.152 ms 69.164 ms 69.154 ms	
11 washcr1-sdn2-washsdn1.es.net (134.55.220.53) 69.155 ms 69.016 ms 69.066 ms	GARR (Italian R&E net) core
12 esnet-wash.rtl.fra.de.geant2.net (62.40.125.77) 161.564 ms 161.518 ms 161.568 ms	
13 so-6-2-0.rtl.geant2.net (62.40.112.21) 169.615 ms 169.584 ms 169.669 ms	
14 so-2-0-0.rtl.mil.it.geant2.net (62.40.112.34) 177.056 ms 177.070 ms 176.924 ms	
15 garr-gw.rtl.mil.it.geant2.net (62.40.124.130) 177.070 ms 176.960 ms 176.927 ms	
16 rt1-mi1-rt-mi2.mi2.garr.net (193.206.134.190) 177.363 ms 177.199 ms 177.251 ms	INFN Frascati
17 rt-mi2-rt-rm2.rm2.garr.net (193.206.134.230) 186.649 ms 189.231 ms 186.571 ms	
18 rt-rm2-rc-fra.garr.net (193.206.134.214) 187.135 ms 187.042 ms 187.159 ms	
19 rc-fra-ru-lnf.fra.garr.net (193.206.136.206) 187.161 ms 187.133 ms 187.166 ms	
20 ***	
21 www6.lnf.infn.it (193.206.84.223) 187.324 ms 187.123 ms 187.162 ms	

Figure 17. Application view of an end-to-end path

that have at least some of the required MA services to support the tool. The user input to the tool is a traceroute between elements of a distributed application that defines the path through the IP network. The tool analyzes the path and topology information is retrieved from perfSONAR services; it then queries the MA services in the intervening networks. The MA services returns the requested utilization information, which is passed to a graphing tool. By way of example, the path between Lawrence Berkeley National Laboratory and the Poznan, Poland supercomputer center involves crossing five domain boundaries and is shown in Figure 17 and Figure 18.

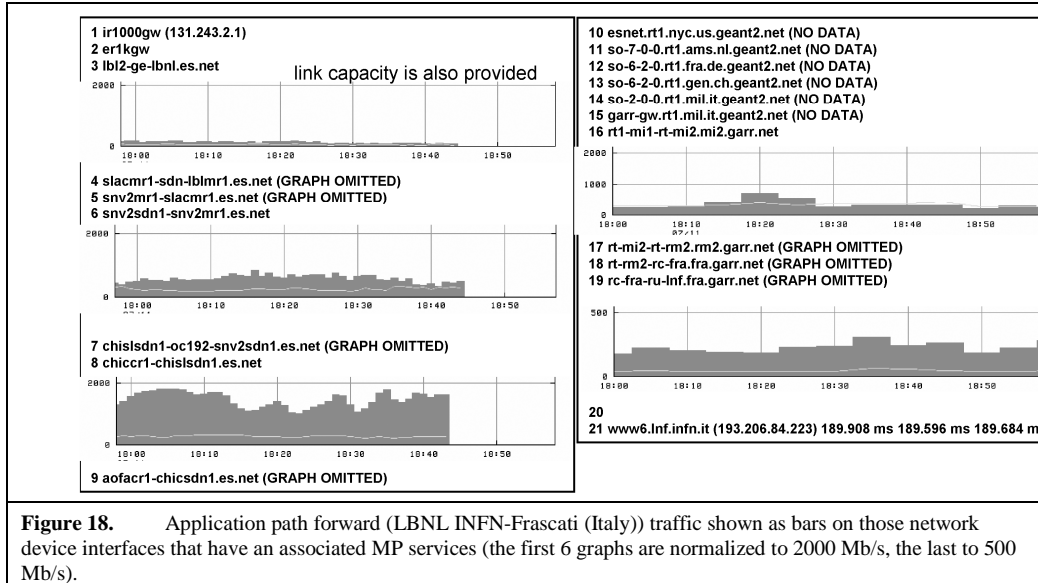
Status: perfSONAR is being developed through collaboration between some 25 network organizations in US and Europe. The basic framework is complete and the protocols are being documented. New services are being developed and deployed. For more information see [21] and [22].

PerfSONAR is still in its development phases and not yet routinely deployed, though it is gaining ground. Perhaps even more important than the current state of perfSONAR is the growing recognition within the networking community that the anonymous, best-effort Internet of 10 years ago is no longer adequate to serve the needs of large-scale, data intensive applications such as large scientific instruments and experiments.

Network Outage Footprint Calculator

It is important to solve the problem of determining the impact caused by the failure of a particular network element, and to provide this information to the application.

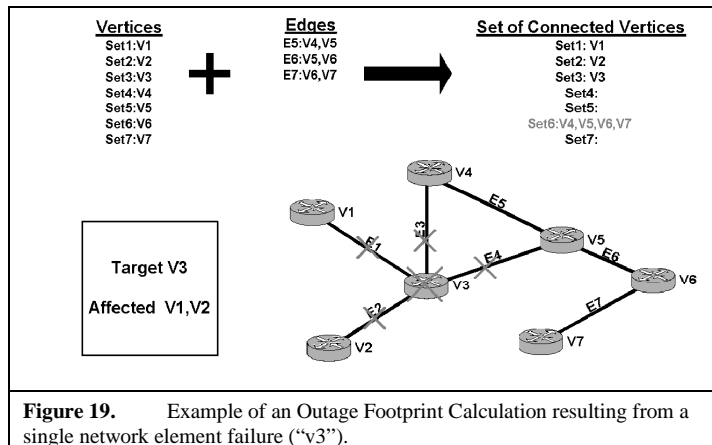
ESnet has been experimenting with an automated approach to solving this problem. The approach involves two issues: 1) accurately determining the dynamic topology of the network and 2) using the topology to determine current state of the overall network.



Topology Mapping: In order to accurately monitor the network one must accurately model the network. This is accomplished by monitoring each network interface and deriving an accurate IP layer connectivity model of the network on an hourly basis. The daily IP layer connectivity changes that occur through the course of regular operations are captured each night and archived so that retrospective questions about connectivity can be answered.

Outage Footprint Calculator: The Outage Footprint Calculator computes the devices (routers, interfaces) that will be isolated from the network given a list of routers and interfaces out of service.

The current network topology is used to create a list of “vertices” and “edges”. A connected graph of the network is derived by inspecting each “edge” in the network topology model and joining the sets of routers at each end of each “edge”. During normal conditions when the network is 100% available, the processing of all edges results in a single set of devices representing the fully connected ESnet network as represented by all



vertices showing up in a single set of connected vertices.

To compute the effect of removing a set of routers or links, each “edge” connected to the given router(s) or interface argument(s) is removed from the “edge” list prior to running the connection algorithm. The resulting affected devices end up in vertex sets that are separate from each other and therefore unreachable. (Each set is a disconnected part of the network.) This is illustrated in Figure 19. This sort of representation can be combined with a path description in much the same way that is done for the perfSONAR “path monitor” service to provide application-view information about the impact of planned outages in the network.

6. Conclusions

The usage of, and demands on, ESnet (and similar R&E networks) are expanding significantly as large-scale science becomes increasingly dependent on high-performance networking. The motivation for the next generation of ESnet is derived from observations of the current traffic trends and case studies of major science applications. The case studies of the science uses of the network lead to an understanding of the new uses of the network that will be required. These new uses require that the network provide new capabilities and migrate toward network communication as a service-oriented capability. This paper has described ESnet’s response to these new directions.

7. Acknowledgements

The ESnet senior network engineering staff that are responsible for the evolution of ESnet consists of Joseph H. Burrencia, Michael S. Collins, Eli Dart, Jon Dugan, James V. Gagliardi, Chin P. Guok, Yvonne Y. Hines, Joe Metzger, Kevin Oberman and Michael P. O’Connor. This group of people contributed to this paper, with Guok contributing the OSCARS material, Metzger contributing the perfSONAR material, and O’Conner contributing the Outage Footprint Calculator material. Harvey Newman, Caltech, has contributed greatly to ideas about service-oriented network services as needed by the Grid analysis systems of the High Energy Physics community. Thanks are due Brian Tierney of the LBL Distributed Systems Group for applying his expertise in high-performance network based systems and TCP stack tuning during the paper editing process.

ESnet is funded by the US Dept. of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program. Dan Hitchcock is the ESnet Program Manager and Thomas Ndousse-Fetter is the Program Manager for the network research program that funds 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 under contract DE-AC03-76SF00098.

8. Notes and References

- [1] <http://www.energy.gov/>, Science and Technology tab.
- [2] High Performance Network Planning Workshop, August 2002
<http://www.doecollaboratory.org/meetings/hpnpw>
- [3] DOE Workshop on Ultra High-Speed Transport Protocols and Network Provisioning for Large-Scale Science Applications, April 2003 <http://www.csm.ornl.gov/ghpn/wk2003>
- [4] DOE Science Networking Roadmap Meeting, June 2003
<http://www.es.net/hypertext/welcome/pr/Roadmap/index.html>
- [5] Science Requirements for ESnet Networking,
<http://www.es.net/hypertext/requirements.html>
- [6] LHC Computing Grid Project <http://lcg.web.cern.ch/LCG/>
- [7] International Committee for Future Accelerators (ICFA), Standing Committee on Inter-Regional Connectivity (SCIC), Professor Harvey Newman newman@hep.caltech.edu, Caltech, Chairperson. ICFA SCIC. Report of the Standing Committee on Inter-Regional Connectivity (SCIC) "Networking for High Energy Physics," February 8, 2007.
- [8] http://www.sc.doe.gov/ascr/20040510_hecrtf.pdf (public report)
- [9] ASCR Strategic Planning Workshop, July 2003 <http://www.fp-mcs.anl.gov/ascr-july03spw>
- [10] Planning Workshops-Office of Science Data-Management Strategy, March & May 2004
<http://www-user.slac.stanford.edu/rmount/dm-workshop-04/Final-report.pdf>
- [11] ESG - Earth System Grid. <http://www.earthsystemgrid.org/> ESG - Earth System Grid. <http://www.earthsystemgrid.org/>
- [12] CMS - The Compact Muon Solenoid Technical Proposal. <http://cmsdoc.cern.ch/>
- [13] The ATLAS Technical Proposal.
<http://atlasinfo.cern.ch/ATLAS/TP/NEW/HTML/tp9new/tp9.html>
- [14] LHC - The Large Hadron Collider Project.
http://lhc.web.cern.ch/lhc/general/gen_info.htm
- [15] The BaBar Experiment at SLAC. <http://www-public.slac.stanford.edu/babar/>
- [16] The D0 Experiment at Fermilab. <http://www-d0.fnal.gov/>
- [17] The CDF Experiment at Fermilab. <http://www-cdf.fnal.gov/>
- [18] The Relativistic Heavy Ion Collider at BNL. <http://www.bnl.gov/RHIC/>
- [19] The Enabling Grids for E-sciencE (EGEE) project is funded by the European Commission and aims to build on recent advances in grid technology and develop a service grid infrastructure which is available to scientists 24 hours-a-day.
(<http://public.eu-egee.org/>)
- [20] <http://csrc.nist.gov/piv-project/>
- [21] "Measuring Circuit Based Networks," Joe Metzger, ESnet. <http://www.es.net/pub/esnet-doc/index.html#JM021307>
- [22] "ESnet Network Measurements," Joe Metzger, ESnet. <http://www.es.net/pub/esnet-doc/index.html#JM021307>