

# **The Advanced Networks and Services Underpinning Modern, Large-Scale Science: DOE's ESnet**

*William E. Johnston*

*ESnet Manager and Senior Scientist,*

*DOE Lawrence Berkeley National Laboratory, MS 50B-2239, Berkeley, California, U.S.A.*

## **Abstract**

---

Modern large-scale science requires networking that is global in extent, rock solid in reliability, adaptable to changing requirements, capable of providing bandwidth bounded only by the latest technology and able to support large volumes of sustained traffic. These were some of the conclusions of a series of workshops conducted by the US Dept. of Energy's Office of Science that examined the networking and middleware requirements of the major science disciplines supported by the Office of Science. The requirements from the workshops have resulted in a new approach and architecture for DOE's Energy Sciences Network (ESnet), which is the network that serves all of the major DOE facilities. This new architecture includes elements supporting multiple, high-speed national backbones with different characteristics, redundancy, quality of service and circuit oriented services, all the while allowing interoperation of these elements with the other major national and international networks supporting science. This paper describes the motivation, architecture, and services of the new approach. The approach is similar to, and designed to be compatible with, other research and education networks such as Internet2/Abilene in the United States and DANTE/GÉANT in Europe, consequently the descriptions given here are at least somewhat representative of the general directions of the networking for the research and education community.

## **1 ESnet's Role in the DOE Office of Science**

---

"The Office of Science of the US Dept. of Energy 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 vital area of national importance. It oversees – and is the principal federal funding agency of – the Nation's research programs in high-energy physics, nuclear physics, and fusion energy sciences. [It also] manages fundamental research programs in basic energy sciences, biological and environmental sciences, and computational science. In addition, the Office of Science is the Federal Government's largest single funder of materials and chemical sciences, and it supports unique and vital parts of U.S. research in climate change, geophysics, genomics, life sciences, and science education." [1]

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

## **2 Office of Science Drivers for Networking**

---

ESnet is driven by the requirements of the science Program Offices in DOE's Office of Science. Several workshops ([2], [3], [4]) examined these requirements as they relate to networking and middleware.

In the first Office of Science workshop (August, 2002) the goal was to examine the networking needs of major OSC science programs. The programs considered were climate simulation, the Spallation Neutron Source facility, the Macromolecular Crystallography facility, high energy physics experiments, magnetic fusion energy sciences, chemical sciences, and bioinformatics. Except for nuclear physics and the two DOE supercomputer facilities (which were considered separately), this is a fairly complete representation of the major OSC programs.

The workshop approach was to examine how the science community believed that the process of doing science had to change over the next 5-10 years in order to make significant advances in the various science disciplines. The resulting future environment and practice of science was then analyzed to characterize how much network bandwidth and what new network and collaboration services would be needed to enable the future environment of science.

Qualitatively, the conclusions were that modern, large-scale science is completely dependent on networks. This is because unique scientific instruments and facilities are accessed and used remotely by researchers from many institutions. Further, these facilities create massive datasets that have to be archived, catalogued, and analyzed by distributed collaborations. The analysis of such datasets is accomplished, e.g., using the approach of Grid managed resources that are world-wide in scope. [5]

The next sections describe two typical science scenarios that drive the networking requirements. They are abstracted from [2].

## **2.1 Distributed Simulation**

To better understand climate change, we need better climate models providing higher resolution and incorporating more of the physical complexity of the real world. Over the next five years, climate models will see a great increase in complexity, for example in work such as the North American Carbon Project (NACP), which endeavors to fully simulate the terrestrial carbon cycle.

These advances are driven by the need to determine future climate at both local and regional scales as well as changes in climate extremes—droughts, floods, severe storm events, and other phenomena. Over the next five years, climate models will also incorporate the vastly increased volume of observational data now available (and even more in the future), both for hind casting and intercomparison purposes. The result is that instead of tens of terabytes of data per model instantiation, hundreds of terabytes to a few petabytes ( $10^{15}$  bytes) of data will be stored at multiple computing sites, to be analyzed by climate scientists worldwide. Middleware systems like the Earth System Grid [9], and its descendents, must be fully utilized in order access and manage such large, distributed, and complex pools of observational and simulation data.

In the period five to ten years out, climate models will again increase in resolution, and many more components will be integrated. These enhanced simulations will be used to drive regional-scale climate and weather models, which require resolutions in the tens to hundreds of meters range, instead of the hundreds of kilometers resolution of today's Community Climate System Model (CCSM) and Parallel Climate Model (PCM).

Better climate modeling requires that the many institutions working on various aspects of the climate be able to easily describe, catalogue, and seamlessly share the knowledge and the vast amounts of data that underlay the knowledge in order to facilitate the required interdisciplinary collaboration. Further, all of the sub-models must interoperate in ways that represent how the elements that make up the climate interact.

As climate models become more multidisciplinary, scientists from oceanography, the atmospheric sciences, and other fields, will collaborate on the development and examination of more realistic climate models. Biologists, hydrologists, economists, and others will assist in the creation of additional components that represent important but as-yet poorly understood influences on climate that must be coupled with the climate models.

There will be a true carbon cycle component, where models of biological processes will be used, for example, to simulate marine biochemistry and fully dynamic vegetation. These scenarios will include human population change, growth, and econometric models to simulate the potential changes in natural resource usage and efficiency. Additionally, models representing solar processes will be integrated to better simulate the incoming solar radiation.

The many specialized scientific groups that work on the different components that go into a comprehensive, multi-disciplinary model, build specialized software and data environments that will almost certainly never all be homogenized and combined on a single computing system. Almost all such multidisciplinary simulation is inherently distributed, with the overall simulation consisting of software and data on many different systems combined into a virtual system by using tools and facilities for building distributed systems.

This paradigm relies on high bandwidth networks for managing massive data sets, quality of service to ensure smooth interoperation of widely distributed computational Grid components, and various (Grid) middleware to interconnect and manage this widely distributed system.

## 2.2 Collaboration and Data Management

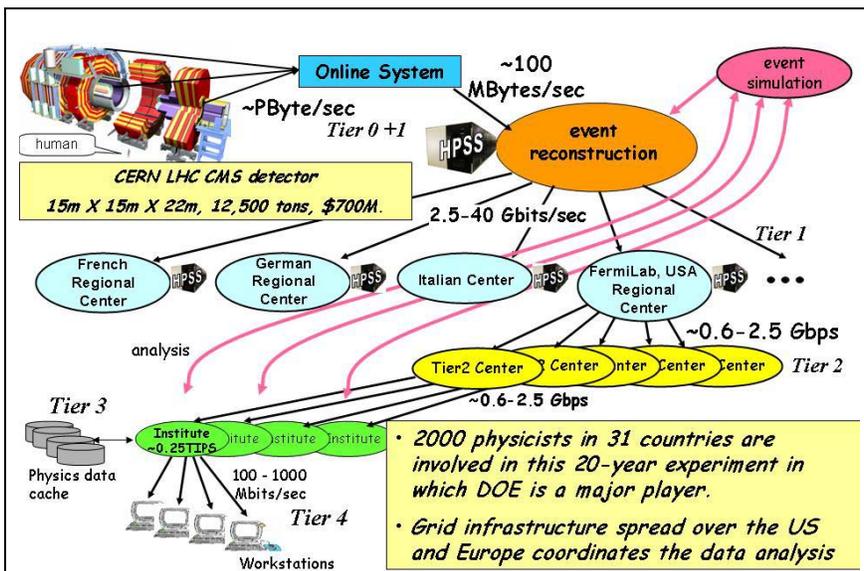
The major high energy physics (HEP) experiments of the next twenty years will break new ground in our understanding of the fundamental interactions, structures and symmetries that govern the nature of matter and space-time. Among the principal goals are to find the mechanism responsible for mass in the universe, and the “Higgs” particles associated with 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 [10] and ATLAS [11] that are building experiments for CERN’s Large Hadron Collider program (LHC [12]), 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 [13]) and Fermilab (D0 [14] and CDF [15]), as well as the experiments at the Relativistic Heavy Ion Collider (RHIC, [16]) program at Brookhaven National Lab, face similar challenges. BaBar, for example, has already accumulated datasets approaching a petabyte.

The HEP (or HENP, for high energy and nuclear physics) 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 1), 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 plan on high capacity networks: to interconnect the physics



**Figure 1. High Energy Physics Data Analysis**

This science application epitomizes the need for laboratories supported by Grid computing infrastructure in order to enable new directions in scientific research and discovery. The CMS situation depicted here is very similar to Atlas and other HEP experiments. (Adapted from original graphic courtesy Harvey B. Newman, Caltech.)

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) 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 [5].

### 2.3 Requirements for Networks and Services Supporting Science

The results of these workshops are surprisingly uniform across the scientific programs represented. Their networking requirements fell into four major areas: bandwidth, reliability, quality of service, and network embedded data cache and computing elements.

The primary network requirements to come out of the Office of Science workshops were

- o Network bandwidth must increase substantially, not just in the backbone but all the way to the sites and to the attached computing and storage systems
- o The 5 and 10 year bandwidth requirements mean that current network bandwidth has, on average, to more than double every year
- o A highly reliable network is critical for science – when large-scale experiments depend on the network for success, the network may not fail
- o There must be network services that can guarantee various forms of quality-of-service (e.g., bandwidth guarantees).

The bandwidth required by DOE’s large-scale science projects over the next 5 years is characterized in the Roadmap workshop (June, 2003) [4]. Programs that have currently defined requirement for high bandwidth include High Energy Physics, Climate (data and computations), NanoScience at the Spallation Neutron Source , Fusion Energy, Astrophysics, and Genomics (data and computations), and Nuclear Physics , and the OSC supercomputer centers. A summary of the bandwidth requirements are given in Table 1.

Additionally, the applications involved in these science areas must move massive amounts of data in a predictable way. That is, both network reliability and quality of service are critical. Further, achieving high bandwidth end-to-end for distributed applications, on-line instruments,

<b>Science Areas</b>	<b>Today <i>End2End</i> Throughput</b>	<b>5 year timeframe <i>End2End Documented</i> Throughput Requirements</b>	<b>5-10 year timeframe <i>End2End Estimated</i> Throughput Requirements</b>	<b>Remarks</b>
High Energy Physics	0.5 Gb/s	100 Gb/s	1000 Gb/s	high bulk throughput
Climate (Data & Computation)	0.5 Gb/s	160-200 Gb/s	N x 1000 Gb/s	high bulk throughput
SNS NanoScience	Not yet started	1 Gb/s	1000 Gb/s + QoS for control channel	remote control and time critical throughput
Fusion Energy	0.066 Gb/s (500 MB/s burst)	0.198 Gb/s (500MB/20 sec. burst)	N x 1000 Gb/s	time critical throughput
Astrophysics	0.013 Gb/s (1 TBy/week)	N*N multicast	1000 Gb/s	computational steering and collaborations
Genomics Data & Computation	0.091 Gb/s (1 TBy/day)	100s of users	1000 Gb/s + QoS for control channel	high throughput and steering

etc., requires new network services and extensive monitoring and diagnosis in the network in order to provide feedback for both debugging and operation.

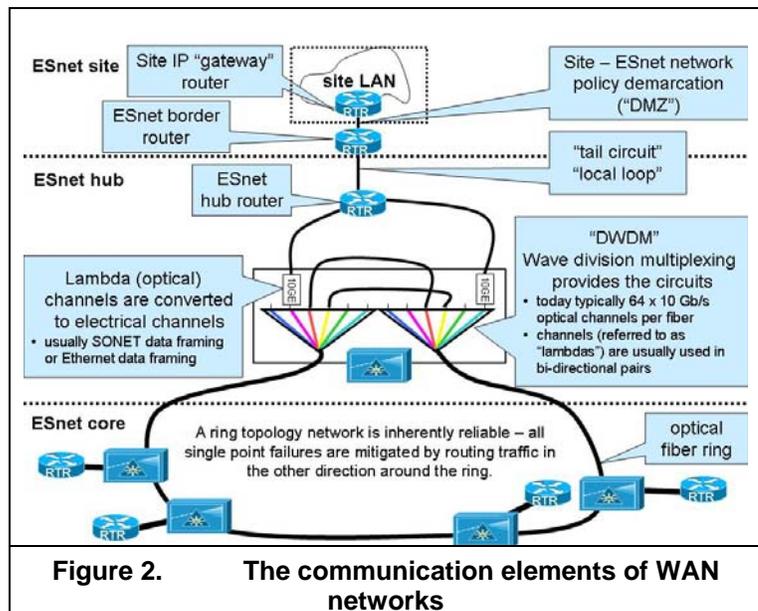
### 3 Interlude: How Do IP Networks Work?

There are four basic aspects to wide-area networking:

The physical architecture, the communication elements and how they are organized, the basic functioning of IP networks, and the logical architecture (how are the interfaces to other networks organized and how is reachability information managed).

Looking at these in this order, consider first the physical architecture. There are two aspects of this – what is the relationship of the communication elements to each other, and what is the overall structure of the communication elements.

Most wide-area networks make use of the communication elements as indicated in Figure 2.



The wide area communication is accomplished by wave (frequency) division multiplexing of optical signal channels that are mixed and carried through optical fiber. Equipment that is typical today multiplexes 64, 10 Gb/s to 40 Gb/s data channels onto a single optical fiber. For data communication networks the channels are almost always used in pairs to provide full duplex communication. The electrical signals are put onto the optical channels using transponders/transmitters that have particular electrical interfaces (e.g. OC192 SONET (United States) / STM-64 SDH (Europe) or 10 Gigabit Ethernet). The optical signal for each channel is about 0.5 nm wide, the mixed signals (channels) are 17.5-32 nm wide and are carried on a base frequency of 1530-1560 nm (near infrared). This process is called Dense Wave Division Multiplexing. Ignored in this discussion are the optical amplifiers and dispersion correctors that are needed for the long distance (hundreds to thousands of km) propagation of the optical signals.

The relationship of groups of communication elements to each other (e.g. systems of interconnected, self-healing rings) is described in the section on the new ESnet architecture, below.

The SONET / SDH or Ethernet circuits implemented by DWDM equipment are point-to-point circuits that connect switches and/or IP routers.

The basic functions of IP networks are to provide addressing, routing, and data transport. IP packets are used by higher level protocols, such as TCP, to provide reliable, stream-oriented data communication. Routers forward IP packets received on one interface to another interface that will get the packet “closer” to its destination. This apparently stateless routing works because the routers exchange reachability information with their neighbors that specifies which interface actually will get the packet closer to its destination. The reachability information is managed by special protocols of which the Border Gateway Protocol (BGP) is most common in the wide area.

In terms of the network protocol architecture stack, application level protocols like HTTP and FTP utilize transport protocols such as TCP that implement semantics such as reliable streams, that, as noted, use IP (as do all Internet protocols) for addressing and routing. While not fundamental, the Domain Name

Service (DNS) that translates between human readable addresses and IP addresses is sufficiently ubiquitous and deeply integrated into applications, that it is now also “essential” for the Internet.

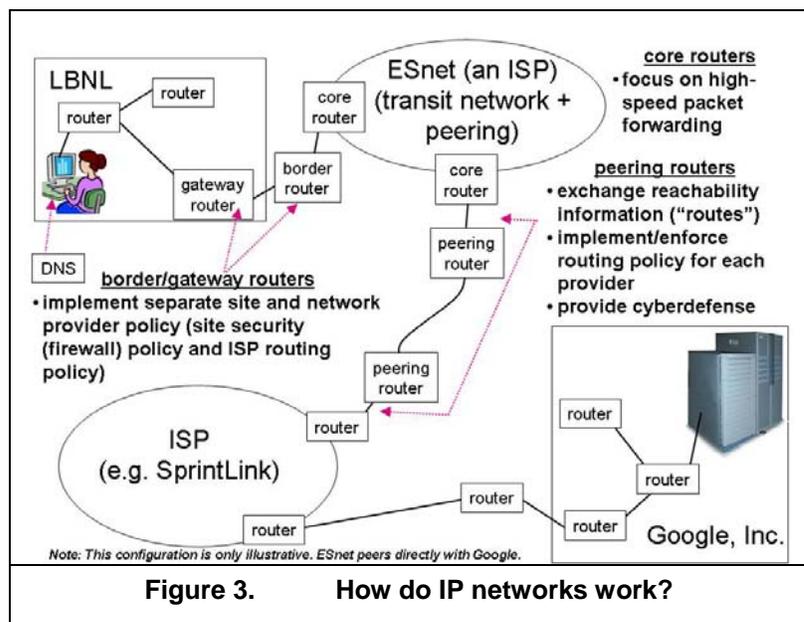
For example, describing Figure 3

- o Accessing a Grid service, a Web server, FTP server, physics data analysis server, etc., from a client application and client computer (e.g. a Web browser on a laptop) involves
  - Providing the target host names (e.g. google.com) where the service is located
  - Determining the IP addresses (e.g. 131.243.2.11 – via DNS)
  - Identifying the service on the target machine (the “port” number of the application server that distinguishes it from other servers that might be running on the same machine)
  - Routing (getting the packets from source, through the Internet, to destination – the service)
- o The Internet transports data packets from source to destination (multiple destinations in the case of multicast)
- o The overall architecture of the Internet is focused on connectivity, routing, and inter-domain routing policy management
  - Domains (“Autonomous Systems”) = large sites and Internet Service Providers (transit networks like ESnet)
  - Policy = Internet Service Provider (ISP) routing strategy and site security

In Figure 3 there are several instances of paired routers. In the case of the site gateway-ISP border router, the site gateway router typically implements security policy (e.g. via firewall rules) and the border router implements the ISP routing policy (e.g., only routes from the site that are for systems at the site will be accepted by the ISP).

In the case of the peering router – peering router pairing between ISPs, the ISPs will accept some routes (claims of reachability for remote networks) and reject others. This selectivity is because the ISP may (frequently does) have multiple ways of getting to a given network and host, and it will use the one that is best suited for its connectivity. (“Best may mean fastest, least expensive, etc.)

“Peering” is the agreements and physical interconnects between two networks that exchange routing information. In order to reach the entire Internet it requires about 160,000 IPv4 routes. There are several dozens of major peering points where networks exchange routing information and packets. Each route provides reachability information for blocks of address space in the network that says “how do I get packets closer to their destination.” The routes are obtained by “peering” (exchanging routing information) with other networks at one or more physical locations. Each peering point provides all of the routes for systems that it believes that it can reach. However, there are frequently multiple ways to reach a given host and the receiver selects the routes based on what it believes is the “best” path to the various networks.

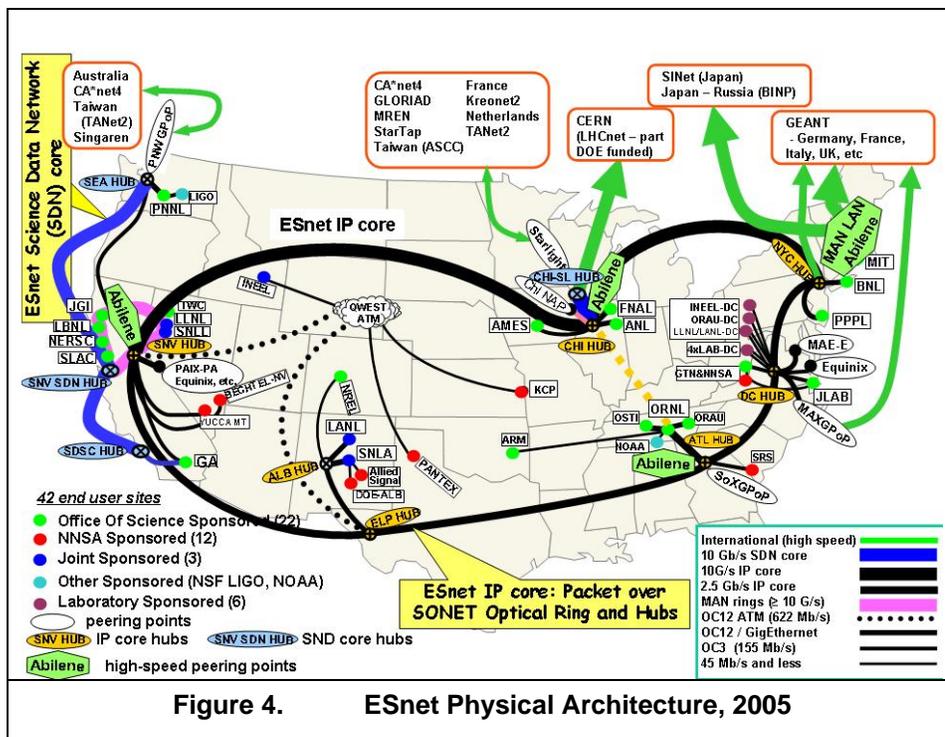


**Figure 3. How do IP networks work?**

## 4 What Does ESnet Provide?

All three of the research and education communities mentioned here – those associated with ESnet, Abilene and the U.S. regional nets, and GÉANT and the NRENs<sup>a</sup> – serve institutions that have requirements for, and access to, a common set of Internet services. However, as mentioned, exactly which of the several networks associated with a given institution provides the services varies. ESnet, while it has a substantial user base (more than 30,000 users in the 42 served sites), is small compared to the U.S. higher education community served by Abilene or the overall European R&E community served by GÉANT. However, ESnet provides a convenient microcosm in which to describe the services provided by various network organizations to the scientific community. (While not excluding the education community, ESnet does not serve that community and the services ESnet provides to its customer base may not map one to one with services offered to higher education institutes.)

One of the characteristics of science oriented networks is that they must provide a relatively small number of sites with very large amount of bandwidth. (As opposed to commodity ISPs like AOL or EarthLink which are tailored to provide a huge number of users a relatively small amount of bandwidth.) In particular, ESnet must provide high bandwidth access to DOE sites and to DOE's primary science collaborators in the science community. This is accomplished by a combination of high-speed dedicated circuits that connect the end sites and by high-speed peerings with the major R&E network partners.



ESnet builds and operates a comprehensive IP network infrastructure (IPv4, IP multicast, and IPv6, peering, routing, and address space management) based on commercial and R&E community circuits. The current physical architecture is shown in Figure 4 which illustrates the extent and diversity of the circuits.

ESnet provides full and carefully optimized access to the global Internet. This is essential, as mentioned above, for the best possible access to the sites where collaborators are located. In order to accomplish this ESnet has peering agreements with many commercial and non-commercial networking. Those agreements result in routes (reachability information) being exchanged between all of the networks needed to provide comprehensive (global) Internet site access.

As noted above, in order to provide DOE scientists access to all Internet sites, ESnet manages the full complement of Global Internet routes. This requires about 160,000 IPv4 routes from 180 peers. (The

<sup>a</sup> In Europe the National Research and Education Networks (NRENs) are the national infrastructure that provide networking to all of the European R&E institutions. The NRENs play a role in Europe very like the regionals (NYSERNet, CENIC, etc.) do in the U.S. European NRENs are not like the "NREN" research network infrastructure of the U.S. 15-20 years ago.

peering policy mentioned above selects these 160,000 routes from about 400,000 that are offered at all of the peering points.) These peers are connected at 40 general peering points that include commercial, research and education, and international networks. With a few of ESnet's most important partner networks (notably Abilene and GEANT), direct peering (core router to core router) is done to provide high performance.

In summary, ESnet provides:

- o An architecture tailored to accommodate DOE's large-scale science
  - Move huge amounts of data between a small number of sites that are scattered all over the world
- o Comprehensive physical and logical connectivity
  - High bandwidth access to DOE sites and DOE's primary science collaborators: Research and Education institutions in the United States, Europe, Asia Pacific, and elsewhere
  - Full access to the global Internet for DOE Labs
- o A full suite of network services
  - IPv4 and IPv6 routing and address space management
  - IPv4 multicast (and soon IPv6 multicast)
  - Prototype guaranteed bandwidth and virtual circuit services
  - Scavenger service so that certain types of bulk traffic will use all available bandwidth, but will give priority to all other traffic when it shows up
- o A highly collaborative and interactive relationship with the DOE Labs and scientists for planning, configuration, and operation of the network
  - ESnet and its services evolve continuously in direct response to OSC science needs
- o Comprehensive user support, including "owning" all trouble tickets involving ESnet users (including problems at the far end of an ESnet connection) until they are resolved – 24 hrs/day x 365 days/year coverage
  - ESnet's mission is to enable the network based aspects of OSC science, and that includes troubleshooting or otherwise managing network problems wherever they occur
- o Cybersecurity in the WAN environment (ESnet sites are responsible for site cybersecurity)
- o Collaboration services and Grid middleware supporting collaborative science
  - Federated trust services with science oriented policy
  - Audio and video conferencing

## **5 Drivers for the Evolution of ESnet**

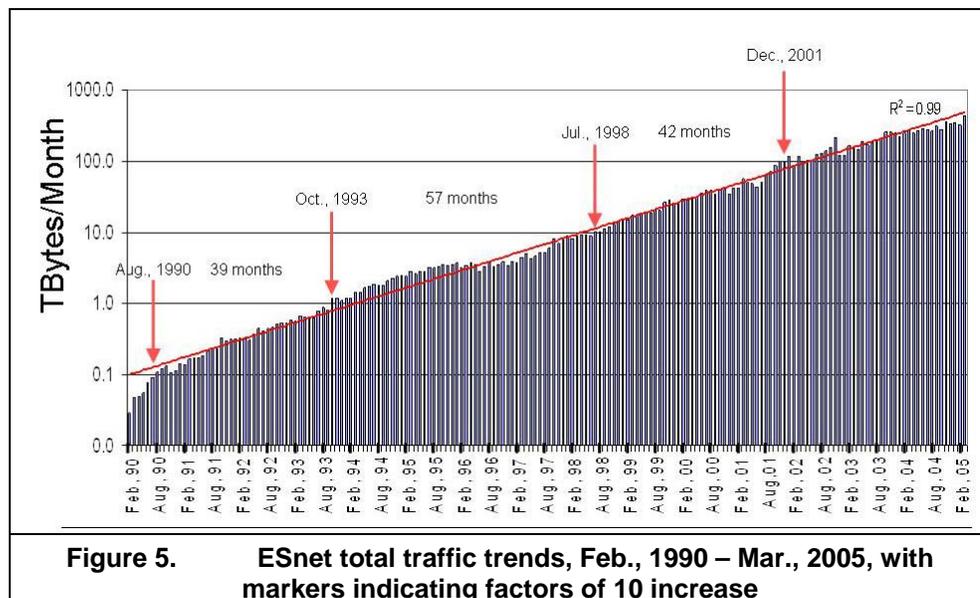
---

The predictive requirements that have been laid out by the scientific community, and that are discussed above, are backed up by observed traffic growth and patterns in ESnet.

### **5.1 Observed Traffic Growth is Exponential**

The total traffic handled by ESnet has been growing exponentially for the past 15 years and currently ESnet handles more than 400 Terabytes/month of data. This growth is somewhat uneven month to month, but over 10 years the ESnet traffic has increased, on average, by a factor of 10 every 46 months.

This growth is qualitatively consistent with the predictions of the science community, and there should be nothing surprising about this growth. In a sense it is just tracking Moore's law. As computers get bigger, they run larger simulations that generate more data – an exponential process; many sensor based instruments, e.g. telescope CCDs, are also experiencing Moore's law growth of the sensors, and the data volume goes up correspondingly, etc.



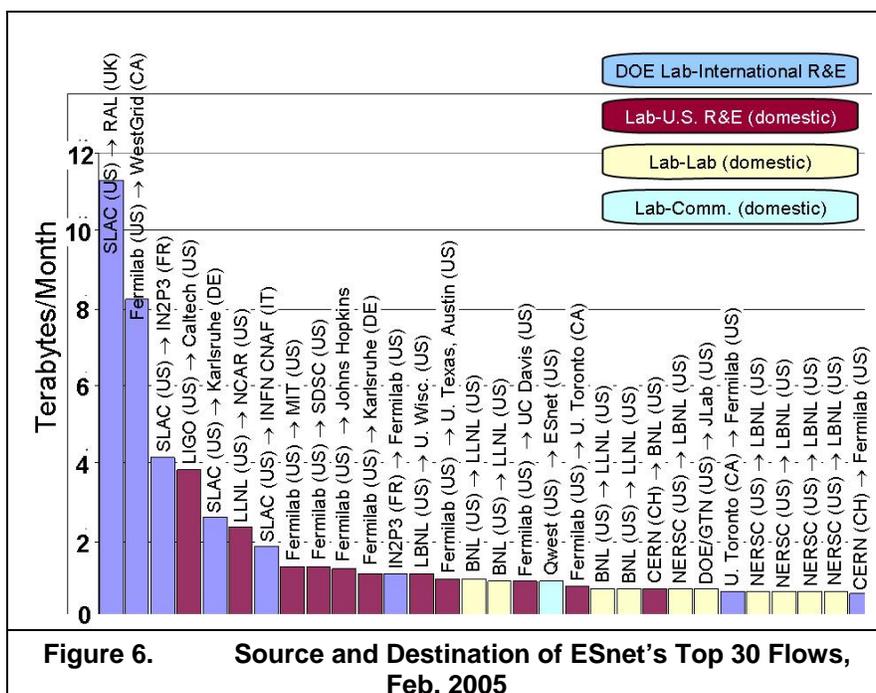
**Figure 5. ESnet total traffic trends, Feb., 1990 – Mar., 2005, with markers indicating factors of 10 increase**

## 5.2 A Small Number of Science Users Account for a Significant Fraction of all ESnet Traffic

Another aspect of large-scale science previously mentioned is that it is inherently collaborative. One reason for this is that the instruments of large scale science are so large and expensive that only one or two may be built and all of the scientist in that field must share the instrument. As noted above, for example, the “instruments” (detectors) at the LHC accelerator at CERN, in Switzerland, cost almost \$U.S. 1 billion each (there are two major detectors on the LHC) and the scientific teams consists of several thousand physicists who are located all over the world.

The effects of this sort of large-scale collaboration are reflected in the data flow patterns in ESnet.

Over the past few years there has been a trend for very large flows (host-to-host data transfers) to be an increasingly large fraction of all ESnet traffic. These very large flows of individual science projects now represent the dominate factor driving the planning for the evolution of the network. This trend is clearly illustrated in Figure 6 which shows host-to-host flows. In fact, 17 of the top 20 flows are from only three ESnet sites, all of which are high energy physics Labs.



**Figure 6. Source and Destination of ESnet's Top 30 Flows, Feb. 2005**

The current predominance of international traffic is due to high-energy physics – the BaBar (SLAC) and D0 (Fermilab) experiments both have major analysis sites in Europe. However, all of the LHC U.S. Tier-2

data analysis centers are at U.S. universities, and as the Tier-2 centers come on-line, the ESnet traffic from U.S. Tier-1 centers (Fermilab and Brookhaven) to U.S. Tier-2 centers at various universities will increase enormously.

High energy physics is several years ahead of the other science disciplines in data generation. Several other disciplines and facilities (e.g. the supercomputer centers, driven by climate modeling) will contribute comparable amounts of additional traffic in the next few years.

## **6 ESnet, Abilene, NLR, GÉANT, and LHC Networking**

---

A large fraction of all of the national data traffic supporting U.S. science is carried by two networks – ESnet and Internet-2/Abilene, and on layer 1 services from 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. Abilene is 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. Esnet also provides the peering and routing needed for the Labs to have access to the global Internet. Abilene 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 Abilene interconnects the GigaPoPs. Abilene is mostly a transit network – the universities and/or the regional networks provide the peering and routing for end-user Internet access. This is also very similar to the situation in Europe where GÉANT (like Abilene) interconnects the European National Research and Education Networks that in turn connect to the LANs of the science and education institutions. (The NRENs are like the US regionals, but organized around the European nation-states).

The top level networks – ESnet, Abilene, 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 Abilene have had joint engineering meetings for several years (Joint Techs) and ESnet, Abilene, GÉANT, and CERN have also formed an international engineering team (“IntTechs”) that meets several times a year.

The goal is that connectivity from DOE Lab to US and European Universities should be as good as Lab to Lab and University to University connectivity. The key to ensuring this is constant monitoring. ESnet has worked with the Abilene and the US University community to establish a suite of monitors that can be use to provide a full mesh of paths that continuously checks all of the major interconnection points. Similarly, the ESnet to CERN connection is monitored, and key European R&E institutions will soon be included.

The networking to support the analysis of the massive data analysis required for the several experiments associated with the LHC (see section 2.2) is a special case. The LHC community is essentially building a three continent, lambda based network for the data distribution. By mid-summer, 2005 there will be two 10 Gb/s paths from CERN to the United States to support the Tier-1 data centers at Fermi National Accelerator Laboratory (Fermilab) and Brookhaven National Lab, both of which are ESnet sites. Many of the immediate requirements for the near-term evolution of ESnet arise out of the requirements of the LHC networking and both United States and international design and operations groups have been established.

## **7 How the Network is Operated**

---

The normal operation of a network like ESnet involves monitoring both the state of the logical network (connectivity to the rest of the Internet) and the state of the physical network (the operational status of the network links, switches, routers, etc.).

Managing the logical network entails ensuring that there are paths from the systems at the DOE Labs to every other system connected to the Internet. This is accomplished by having a comprehensive set of routes to all of the active IP address space through the peering process described above. Managing these routes in order to provide high quality access to the global Internet is a never ending task because ISPs come and go and change their relationship to other ISPs, etc. Automated tools to control which routes are accepted from specific peers help keep this process maintainable.

The physical network is managed largely through extensive, continuous monitoring. The eleven hubs and 42 end sites are monitored minute by minute at more than 4400 physical and logical interfaces. This includes every aspect of the operating state of the equipment and the traffic flowing over every interface. All of this information is simultaneously analyzed by a network monitoring system and entered into a database that is accessible to the ESnet engineers at various locations around the country.

## **7.1 Scalable Operation is Essential**

R&E networks like ESnet are typically operated with a small staff. The key to this is that everything related to the operation of the network and related services must be scalable. The question of how to manage a huge infrastructure with a small number of people dominates all other issues when looking at whether to support new services (e.g. Grid middleware): Can the service be structured so that its operational aspects do not scale as a function of the user population? If not, then the service cannot be offered.

In the case of the network itself, automated, real-time monitoring of traffic levels and operating state of some 4400 network entities is the primary network operational and diagnosis tool. Much of the analysis of this information (generated in real-time with sample intervals as short as minutes or generated asynchronously as alarms) is automatically analyzed and catalogued as to normal or abnormal, urgent or not. Urgent abnormal events filter up through a hierarchy of operational and engineering staff. The entire ESnet network is operated 24x7x365 by about 16 people.

## **7.2 What Does the Network Actually Look Like?**

The ESnet core consists of 11 hubs and sub-hubs. A typical ESnet hub (“AoA” - 32 Avenue of the Americas, NYC, in this case) is illustrated in Figure 7. The core routers have the primary job of high-speed forwarding of packets. They have the high-speed interfaces for the 2.5 and 10 Gb/s cross-country circuits, and for circuits to Abilene and GÉANT. At AoA the core router has a 10 Gb/s circuit to Chicago and 2.5 Gb/s to Washington, DC and a 2.5 Gb/s circuit to Brookhaven Lab, an ESnet site.

There are also several direct connections for high-speed peerings on the core router: MAN LAN (Manhattan Landing) provides 10 Gb/s connections to several R&E networks – currently Abilene, NYSERNet (New York R&E network), SInet (Japan), CANARIE (Canada), HEAnet (Ireland), and Qatar. Most ESnet hubs also have a peering router that connects to the core router and to the peers that happen to have presence in that hub (“AoA Hub Peer” in Figure 7). The separation of the peering function from the core routing function simplifies management and allow for a more effective cyber security stance.

The router labeled “AoA Hub” is used for low-speed, long-haul interfaces such as OC3 (155 Mb/s) and T3 (45 Mb/s) that have to interface with Telco equipment at the sites.

Supporting and auxiliary equipment consists of a secure terminal server (the triangle and lighting bolt) that provides access of last resort by telephone modem, a power controller that allows for remote power cycling of all of the other equipment, and one or more performance testing systems. At AoA ESnet has two types of performance testing systems: The Performance Center systems provide for interactive

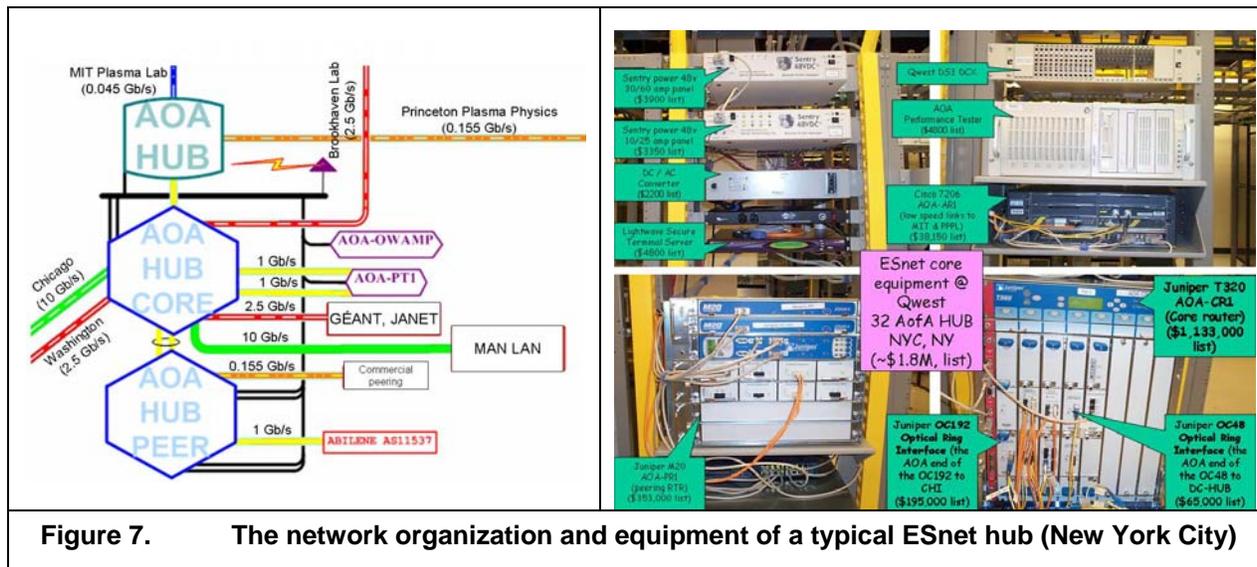


Figure 7. The network organization and equipment of a typical ESnet hub (New York City)

diagnostics and are available to ESnet engineers, to site network engineers, and to end users. The OWAMP (One-Way Active Measurement Protocol) server is used for the DOE Lab to University site testing described in section 6. There is also a local management network.

## 8 Operating Science Mission Critical Infrastructure

ESnet is a visible and critical piece of DOE science infrastructure. There are several tens of thousands users at DOE facilities and many others worldwide who are collaborators. Further, like most modern enterprises, the business operation of DOE Labs depends on network access for many types of interactions with the outside world. This requires high reliability and high operational security in the systems that are integral to the operation and management of the network. Making the network and the science support robust involves both network design and operational issues.

As mentioned elsewhere in this paper, a new network architecture is being implemented that is intended to provide as close to fully redundant network connectivity for the DOE Labs as is practical. Beyond this there are several broad categories of operational functions that must be addressed with the same eye towards fail safe operations.

- o Secure and redundant mail and Web systems are central to the operation and security of ESnet
  - trouble tickets are by email
  - engineering communication by email
  - engineering database interfaces are via Web
- o Secure network access to Hub routers
- o Backup secure telephone modem access to Hub equipment
- o 24x7x365 help desk and 24x7x365 on-call network engineer
- o trouble@es.net (end-to-end problem resolution)

Disaster recovery and stability is essential and part of the basic network architecture. The engineering and operating infrastructure for ESnet is being replicated in a secure telecommunications facility in New York (the primary Network Operations Center – NOC – is in California). The replicated systems currently include

- o Spectrum (network monitoring and alarm system)

- o Engineering web server and databases providing network log management, IPv4 and IPv6 address administration, and other functions
- o Domain Name Server
- o Network device configuration management and load server.
- o Web server www.es.net

The second phase will replicate the Remedy trouble ticket system and e-mail capabilities (scheduled for completion is 3QCY05). When complete this will allow ESnet to be operated independently by any of the engineers at the four NOC locations (two in California, one in Iowa, and one in New York) even in the event of a complete failure of the primary NOC in California.

In order to maintain science mission critical infrastructure in the face of cyber attack ESnet must also provide responsive, tailored, and defensive cybersecurity that results in a coordinated response to cyber attacks that both protects the Labs and keeps them on-line. This is accomplished with a Phased Response Cyberdefense Architecture that is intended to protect the network and the ESnet sites from a class of attacks like large-scale denial-of-service attacks. The phased response ranges from blocking certain site traffic to a complete isolation of the network that allows the sites to continue communicating among themselves in the face of the most virulent attacks. This is accomplished by separating ESnet core routing functionality from external Internet connections by means of a “peering” router that can have a policy different from the core routers, providing a rate limited path to the external Internet that will insure site-to-site communication during an external denial of service attack, and proving “lifeline” connectivity for downloading of patches, exchange of e-mail and viewing web pages (i.e.; e-mail, dns, http, https, ssh, etc.) with the external Internet prior to full isolation of the network.

## **9 Enabling the Future: ESnet’s Evolution over the Next 10-20 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 mission needs.

The current trend in traffic patterns – the large-scale science projects giving rise to the top 100 data flows that represent about 1/3 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 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 flows)

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 most of 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.

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 as predicted by the science requirements analysis three years ago.

The old hub and spoke architecture (to 2004) would not 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.

## 9.1 ESnet's Evolution – The 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:

- 1) 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
- 2) Support for the high bandwidth data flows of large-scale science including scalable, reliable, and very high-speed network connectivity to DOE Labs
- 3) 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 Gbps national core/backbone bandwidth by 2008 in two independent backbones.

## 9.2 ESnet Strategy – A New Architecture

The new architecture that will meet the requirements consists of three elements.

- 1) Metropolitan Area Network (MAN) rings with multiple channels to provide dual backbone connectivity to a site for increased reliability, much higher site-to-core bandwidth, and support for both production IP and circuit-based traffic.
- 2) A Science Data Network (SDN) core for provisioned, guaranteed bandwidth circuits to support large, high-speed science data flows. This new core network will provide very high total bandwidth, and allow multiple connections to the MAN rings for protection against hub failure in the IP core, thus making available an alternate path for production IP traffic.
- 3) A high-reliability IP core (e.g. the current ESnet core) to address general science requirements, Lab operational (business) requirements, provide some level of backup for the SDN core, and act as a vehicle for science collaboration services.

These elements are structured to provide a network with fully redundant paths for all of the OSC 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 8.)

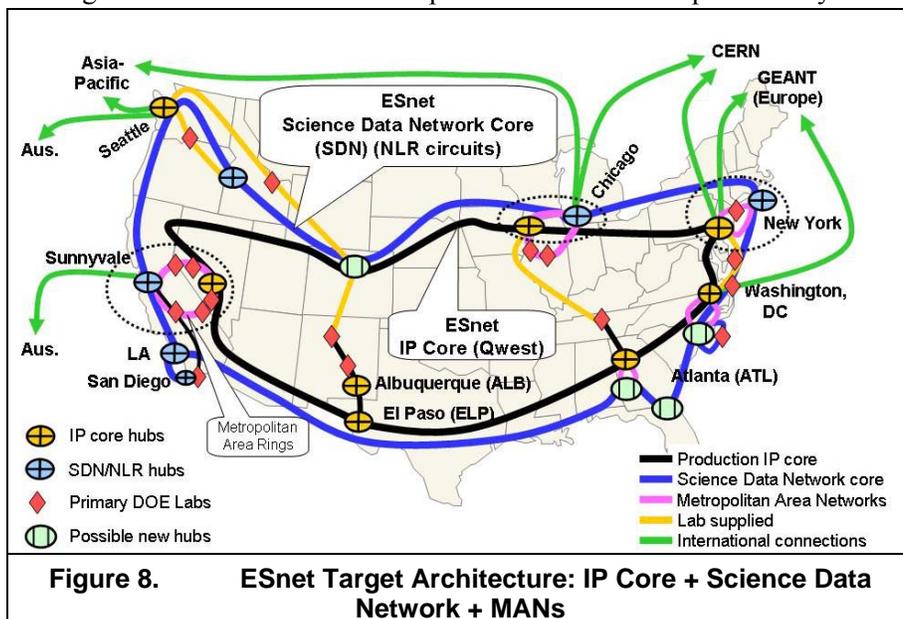
The theoretical advantages of this architecture are clear but it must also be practical to realize in an implementation. That is, how does ESnet get to the 100 Gbps multiple backbones and the 20-40 Gbps redundant site connectivity that is needed by the OSC community in the 3-5 yr time frame?

Only a hybrid approach is affordable.

The core IP network that carries the general science and Lab enterprise traffic should be provided by a commercial telecom carrier in the wide area in order to get the >99.9% reliability that certain types of science and the Lab CIOs demand.

Part, or even most, of the wide area bandwidth for the high impact science networking (SDN) will be provided by National Lambda Rail – an R&E community network that is much less expensive than commercial telecoms.

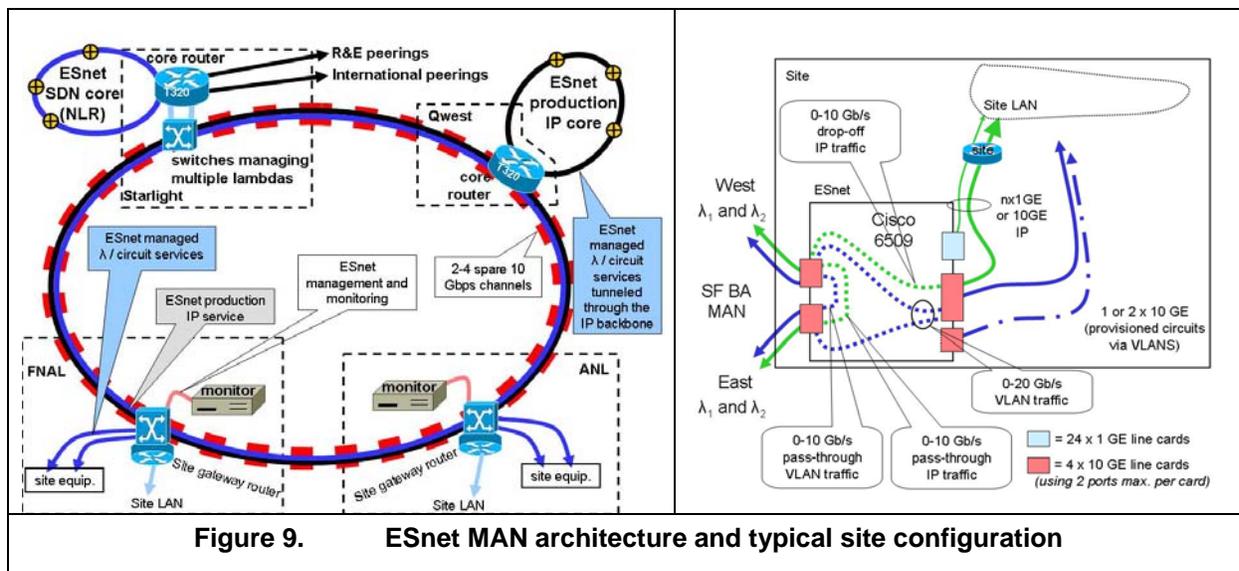
The Metropolitan Area Networks that get the Labs to the ESnet cores are a mixed bag and somewhat opportunistic – a combination of R&E networks, dark fiber networks, and commercial managed lambda circuits will be used.



### 9.3 ESnet Strategy: MANs

The MAN architecture is designed to provide at least one redundant path from sites to both ESnet cores, scalable bandwidth options from sites to the ESnet cores, and support for high-bandwidth point-to-point provisioned circuits.

The MANs are also intended to provide independent connections to each of the two national core networks. This will ensure that no single failure, even of a core router, will disrupt all of the connectivity of any ESnet site. The architecture is illustrated in Figure 9.



In a partnership with Qwest Communications, the ESnet MAN architecture has first been realized in the San Francisco Bay Area. Qwest has provided a fiber network that provides a logical ring consisting of at least two independent fibers (the so-called “east” and “west” directions of the ring) into and out of each of

five Bay Area sites, to the IP core hub and the SDN core hub. Qwest provides a “managed lambda” service and ESnet initially makes use of two lambda rings delivered as 10 Gb/s Ethernet channels (Figure 9). Each site has a layer2/3 switch/router that manages the Ethernet channels of the SDN and provides the production IP service (Figure 9).

## 9.4 Science Data Network

The Science Data Network (SDN) will be ESnet’s second national core ring. SDN is intended to provide most of the bandwidth for high-impact science traffic and serve as a backup for the production IP network.

Most of the bandwidth needed in the next few years is between sites and peering points that are near paths along the West and East coasts and across the northern part of the United States, and so multiple lambda paths are planned for these paths.

National Lambda Rail (NLR) is “a major initiative of U.S. research universities and private sector technology companies to provide a national scale infrastructure for research and experimentation in networking technologies and applications.” (www.nlr.net) ESnet will use NLR lambdas to provide 30-50 Gbps along the path described above and a single 10 Gb/s ring around the country by 2008. Closing the SDN ring in the south to provide resilience at 10 Gbps is a compromise driven by economic considerations that, however, provides a measure of redundancy for the second core.

## 9.5 High Reliability IP core

The ESnet production IP core is currently provided by Qwest Communications and is a highly reliable infrastructure that consistently provides > 99.9% reliability. This level of reliability is the norm for commercial telecommunications infrastructure and results from the redundancy, diversity, and dynamic circuit management at all levels. Such a high reliability infrastructure will probably always be the basis of the ESnet production IP service.

## 10 New Network Services

---

New network services are also critical for ESnet to meet the needs of large-scale science.

One of the most important new network services identified by the Roadmap workshop [5] is dynamically provisioned virtual circuits that provide traffic isolation that will 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 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 experiment 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 workflow. This requires quality of service to ensure a steady flow of data and intermediate results among the systems. Without this, systems with many dependencies and with others dependent on them would stop and start with the interruptions propagating throughout the whole collection of systems creating unstable and inefficient production of analysis results 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.)

## 10.1 OSCARS: Guaranteed Bandwidth Service

DOE has funded the OSCARS (On-demand Secure Circuits and Advance Reservation System) project to determine how the various technologies that provide dynamically provisioned circuits and that provide various qualities of service (QoS) can be integrated into a production net environment.

The end-to-end provisioning will initially be provided by a combination of Ethernet switch management of  $\lambda$  (optical channel) paths in the MANs and Ethernet VLANs and/or MPLS paths (Multi-Protocol Label Switching and Label Switched Paths - LSPs) in the ESnet cores.

The current scope of OSCARS is intra-domain – that is, to establish a guaranteed bandwidth circuit service within the boundary of the ESnet network. The Reservation Manager (RM) has been developed in the first phase of the project pertains only to the scheduling of resources within the ESnet domain.

Setting up inter-domain guaranteed bandwidth circuits is not a trivial task. Differences in network infrastructure (e.g. hardware, link capacity, etc.) may not provide consistent service characteristics (e.g. bandwidth, delay, and jitter) across domains. Differences in policies, such as Acceptable Use Policies (AUPs), Service Level Agreements (SLAs) and security requirements, may not allow the provisioning of inter-domain circuits. None-the-less, inter-domain circuits are of considerable interest, especially between ESnet, Internet2/Abilene, and GÉANT.

The MPLS mechanism provides good control over the path configuration. The Resource Reservation Setup Protocol (RSVP) is used to establish the state in the routers, and RSVP packets may either be routed using the default IP routing, or can set up paths along specified links. The label switching used by MPLS avoids the routing mechanism and QoS may be accomplished by assigning the paths to non-standard queues in the router – for example, expedited forwarding will provide bandwidth guarantees by giving the MPLS packet priority over the best-effort queuing of normal IP packets. Although there are typically only a small number of available queues, multiple virtual circuits with QoS are possible by limiting the total allocated QoS bandwidth to the available QoS bandwidth, and by enforcing QoS circuit bandwidth limits at the ingress points. That is, if a given path can accommodate 5 Gb/s of priority traffic (leaving, say, 5 Gb/s available for best-effort traffic) then 5 – 1 Gb/s, guaranteed bandwidth user virtual circuits can be accommodated by enforcing a limit of 1 Gb/s at each circuit ingress point. The ingress policing is something that is always done because that is the mechanism of enforcing policy based sharing of bandwidth between best-effort and expedited traffic. The enforcing of bandwidth limits on multiple expedited virtual circuits is just an extension of that approach that is managed by the reservation manager that allocates and establishes the circuits in the first place.

With endpoint authentication, these MPLS paths are private and intrusion resistant circuits, so they should be able to bypass site firewalls if the endpoints trust each other.

MPLS is supported on all ESnet routers (and in most routers used in the cores of the R&E networks) and it allows for “stitching together” of paths across domains.

Since OSCARS facilitates the reservation of valuable shared resources, allocation management must be taken into account. Guaranteed bandwidth circuits that go across the country or beyond will be a scarce resource, and therefore an allocation mechanism including how allocations are assigned and managed must be considered.

When OSCARS is put into production, an allocation management framework (similar to that for allocating supercomputing time) must also be put in place.

The allocation management framework is beyond the scope of the current project, however the AAA subsystem is being designed to facilitate contacting an allocation management systems as part of the authorization process.

### ***The Reservation Manager (RM) Components***

A Web-Based User Interface (WBUI) will prompt the user for a username/password and forward it to the AAA system. The WBUI will operate in an HTTPS mode and the servers will be within the trusted domain of the RM.

WBUI then places the reservation request and user's authenticated identifier into a signed SOAP message and forward it to the AAAS.

The Authentication, Authorization, and Auditing Subsystem (AAAS) will handle access, enforce policy, and generate usage records. It will accept reservation requests in the form of signed SOAP messages, and then extracts the authenticated ID and passes it to a policy server that will interface with an allocation management system to authorize the request.

Authorized requests are sent to the Bandwidth Scheduler Subsystem for scheduling.

The AAAS also generates usage records for auditing, accounting, and allocation management purposes.

The Bandwidth Scheduler Subsystem (BSS) will track reservations and map the state of the network (present and future). BSS will preserve state including the current network topology (links in the network) and bandwidth reservations. The BBS will determine the path that will be taken and reserve the bandwidth by updating the reservation database.

BBS will also reschedule link reservations in the event of planned and unplanned network outages.

Just prior to the reservation time, the BBS will trigger the Path Setup Subsystem to setup the LSP. BBS will also trigger the teardown of the LSP when the reservation has expired.

The Path Setup Subsystem (PSS) will setup and teardown the on-demand paths (LSPs). Prior to path setup, PSS will validate the route the LSP will use, and query routers along the path to check for "illegal" active LSPs that could conflict with bandwidth reservation. To setup and teardown the LSP, PSS will change the router configuration on the start-end of the path.

All of the modules have well defined interfaces and are designed to be called from applications other than the Web-Based User Interface.

### ***Operation of the System – Making A Reservation***

The user will make reservations through the Web-Based User Interface. (Subsequently, user applications may talk directly to the AAAS to make a reservation.)

Information which uniquely identifies the stream (e.g. source/destination IP addresses, source/destination port numbers, protocol), along with the duration and bandwidth requirement are entered.

A notice is returned as to whether the reservation is accepted or denied.

### ***Operation of the System – Claiming the Reservation***

At the time when the reservation becomes active, the user simply sends traffic from the source to the destination (the source address and port are specified as part of the reservation process).

The packets are filtered on the ESnet ingress router and those conforming to a reservation are injected into an LSP that was setup by the PSS when the reservation became active.

## **Implementation**

OSCARS is being implemented in three phases each with duration of about one year.

Phase 1 (completed by Jun 2005) consists of

- o Test and deploy MPLS, RSVP, and QoS in the ESnet production environment
- o Implement web-based user interface.
- o Implement basic access-control security for AAAS.
- o Develop simple scheduling algorithms for BSS.
- o Test and implement access methods for PSS.
- o Test at least one user-level application using the QoS service.

Phase 2

- o Create tools to monitor LSP setup/teardown and bandwidth usage.
- o Test and deploy DOEGrids certificate authentication for AAAS.
- o Evaluate the AAAS with one or more user communities

Phase 3

- o Test and deploy authorization and auditing mechanisms for AAAS.
- o Develop rescheduling algorithms for BSS to address network changes during a reservation.
- o Evaluate the BSS with a user community.
- o Test and develop policy server and client for AAAS and BSS.
- o Test and deploy Generalized MPLS (GMPLS) to include optical cross connect equipment in LSP if applicable.

## **Current Status**

MPLS paths have been established through the core between two pairs of ESnet sites Fermi Lab and Brookhaven, and General Atomics and the NERSC supercomputer center.

These tests have shown – unsurprisingly – that policing has a detrimental effect on unshaped TCP flows. The burstiness of the TCP flow causes the policer to discard multiple back-to-back packets, triggering TCP's congestions control mechanism. Flow shaping to ensure that packets are delivered to the virtual circuit within the allowed bandwidth is something that needs to be done prior to injecting the packets into the network. Unfortunately this cannot be done by the user at the application level because a TCP window's worth of IP packets are injected all at once at the network interface cards (NIC) line rate (and in a TCP instance tuned for cross-country distances this is a lot of packets). Shaping usually must be done in the OS of the system that talks to the network or on the NIC. Approaches to ameliorate this situation are being investigated, but are mostly outside the scope of the OSCARS project.

One possible alternative would be to utilize less loss sensitive (and typically less compatible with commodity TCP) protocols when using OSCARS LSPs (e.g. UDP rate-based transport).

## **Collaborations**

The Bandwidth Reservation for User Work (BRUW) project is a part of Internet2's Hybrid Optical and Packet Infrastructure (HOPI) project that will allow authorized users to reserve bandwidth across the Abilene backbone network with minimal human intervention in the process to support advanced applications and research.

The OSCARS and BRUW projects are jointly developing code, which is possible because OSCARS and BRUW have very similar architectures. Among other things, this is intended to ensure compatibility between Internet2 and ESnet communities. There is also close cooperation with the DANTE/GÉANT virtual circuit project ("lightpaths – Joint Research Activity 3 project).

A demonstration of dynamic setup of an inter-domain LSP circuits between ESnet and Internet2 is targeted for the SC05 conference. One motivation for this is as a prototype for U.S. LHC Tier-1 (DOE Labs) – Tier-2 (U.S. universities) data transfers.

## 10.2 ESnet Grid and Middleware Services Supporting Science

The two key workshops whose results are guiding the evolution of ESnet ([2] and [5]) both identified various middleware services that had to be in place, in addition to the network and its services, in order to provide an effective distributed science environment.

In addition to the high-bandwidth network connectivity for DOE Labs, ESnet provides several of the middleware services that are critical for collaboration. These services are called “science services” – services that support the practice of science. Examples of these services include:

- o Trust management for collaborative science
- o Cross site trust policies negotiation
- o Long-term PKI key and proxy credential management
- o Human collaboration communication
- o End-to-end monitoring for Grid / distributed application debugging and tuning
- o Persistent hierarchy roots for metadata and knowledge management systems

There are a number of such services for which an organization like ESnet has characteristics that make it the natural provider. For example, ESnet is trusted, persistent, and has a large (almost comprehensive within DOE) user base. ESnet also has the facilities to provide reliable access and high availability of services through assured network access to replicated services at geographically diverse locations.

However, given the small staff of an organization like ESnet, a constraint on the scope of such services is that they must be scalable in the sense that as the service user base grows, ESnet interaction with the users does not grow.

There are three types of such services that ESnet offers to the DOE and/or its collaborators.

- o Federated trust
  - policy is established by the international science collaboration community to meet its needs
- o Public Key Infrastructure certificates for remote, multi-institutional, identity authentication
- o Human collaboration services
  - video, audio, and data conferencing

## 10.3 Authentication and Trust Federation Services

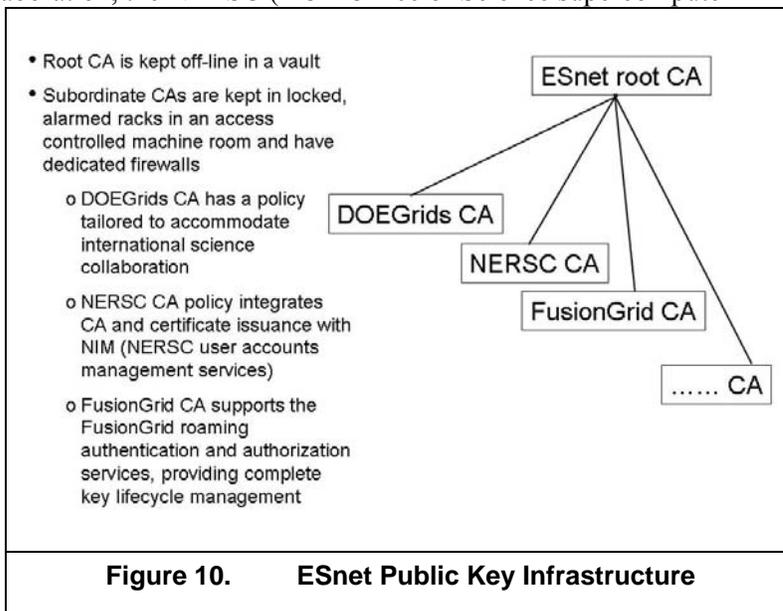
Cross-site identity authentication and identity federation is critical for distributed, collaborative science in order to enable routine sharing computing and data resources, and other Grid services. ESnet provides a comprehensive service to support secure authentication.

Managing cross-site trust agreements among many organizations is crucial for authorization in collaborative environments. ESnet assists in negotiating and managing the cross-site, cross-organization, and international trust relationships to provide policies that are tailored to collaborative science.

### 10.3.1 ESnet Public Key Infrastructure

ESnet provides Public Key Infrastructure and X.509 identity certificates that are the basis of secure, cross-site authentication of people and Grid systems. The ESnet root Certification Authority (CA) service supports several CAs with different uses and policies that issue X.509 identity certificates after validating the user request against the policy of the CA. For example, the DOEGrids CA has a policy tailored to accommodate international science collaboration, the NERSC (DOE Office of Science supercomputer center) CA policy integrates CA and certificate issuance with NERSC user accounts management services, and the FusionGrid CA supports the policy of the DOE magnetic fusion program's FusionGrid roaming authentication and authorization services, providing complete key lifecycle management.

This PKI is focused entirely on enabling science community resource sharing and its policies are driven entirely by the science communities that it serves. That is, the trust requirements of the science communities are formally negotiated and encoded in the Certification Policy and Certification Practice Statement of the CAs.



The DOEGrids CA ([www.doeagrids.org](http://www.doeagrids.org)) was the basis of the first routine sharing of HEP computing resources between United States and Europe.

### 10.3.2 Federation and Trust management

ESnet has been working with the international Grid community develop policies and processes that facilitate the establishment of multi-institutional and cross-site trust relationships. This effort led to the development of two key documents used by the community and published by the Global Grid Forum (GGF): CA Policy Management Authority guidelines, and a reference Certificate Policy and Certification Practices Statement (CP/CPS). Policy Management Authorities (PMAs) encode, manage, and enforce the policy representing the trust agreements that are worked out by negotiation. The PMA guidelines outline how to establish a PMA. The CP/CPS guidelines were written to outline issues of trust that must be addressed when setting up a CA.

These documents are used by the regional CA providers to organize their management and to specify their policies. The European, EU Grid PMA, and the Asia Pacific, AP PMA, both use these documents for their communities.

#### **The Americas Grid PMA**

ESnet represents the DOE and NSF Grid user community by participation as a full member on the EUGrid PMA. This is a requirement because of the need to collaborate between the two user communities. However, with the successful deployment of the EU and AP PMAs, there has been pressure on the Americas to form their own regional PMA. The number of users in the Americas has continued to grow, but the only PMA that could handle their trust needs was the European PMA. The increasing number of users from the Americas that need to form similar trust relationships with the Europeans was beginning to stress the capacity of European Grids community. To better serve the Americas Grid

community, and to help off load certification by the EU Grid PMA, ESnet has helped establish The Americas Grid PMA (TAGPMA).

### ***International Grid Trust Federation***

The formation of the EU, AP PMA, and Americas Grid PMAs has created a need to coordinate these regional efforts to insure a common, global trust based federation. The IGTF was fostered by ESnet to help coordinate the global efforts of trust management. In March 2003, ESnet met in Tokyo with a number of international PMAs, this led to the establishment of the IGTF ([www.GridPMA.org](http://www.GridPMA.org)). The IGTF has grown to include the three major regional PMAs: [www.EUGridPMA.org](http://www.EUGridPMA.org), [www.APGridPMA](http://www.APGridPMA.org) and the new [www.TAGPMA.org](http://www.TAGPMA.org) (Americas). It will be the publishing point for various policies and official points of contact (POC).

### ***OCSP Service for Grids***

The Online Certificate Status Protocol (OCSP) (RFC 2560) is a simple query protocol that relieves PKI clients of the burden of managing and maintaining lists of revoked certificates. ESnet engineers are co-authoring an OCSP requirement document for GGF and have built a pilot OCSP service (<http://amethyst.es.net/>) for use by Grids and other users of ESnet PKI services.

### ***Wide Area Net RADIUS and EAP***

ESnet has developed considerable expertise in the use of RADIUS (RFC 2865) across ESnet, and in tandem with RADIUS is investigating the Extensible Authentication Protocol (EAP) (RFC 2864) to support secure authentication services for distributed computing applications. Authentication methods supported include one-time password services, PKI (as used in Grids and possibly in the US government's Personal Identity Verification (PIV) project [17]), and conventional password services. Applications supported include Grid credential stores like MyProxy, sshd, web servers, and UNIX login.

### ***PGP Key Server***

The ESnet PGP keyserver (<http://www.es.net/pgp>) provides an access and distribution service for PGP keys in the ESnet community, and with other PGP users. The PGP key server is a node in a mesh of worldwide PGP key servers, supporting over two million PGP keys.

## **10.4 Voice, Video, and Data Tele-Collaboration Service**

Another important and highly successful ESnet Science Service are audio, video, and data teleconferencing services that are very important for supporting human collaboration in geographically dispersed scientific collaborators.

ESnet provides the central scheduling and ad-hoc conferencing services that are essential for global collaborations, and that serve more than a thousand DOE researchers and collaborators worldwide. The services consist of ad-hoc H.323 (IP) videoconferences (5000 port hours per month are used), scheduled audio conferencing (2000 port hours per month used), and data conferencing (50 to 100 port hours per month).

Web-based, automated registration and scheduling for all of these services provide the central coordination that makes the service valuable to a large, world-wide community (<http://www.ecs.es.net>).

## **11 Conclusions**

---

ESnet is an infrastructure that is critical to DOE's science mission, both directly and in supporting collaborators. It is focused on the Office of Science Labs, but serves many other parts of DOE.

ESnet is implementing a new network architecture in order to meet the science networking requirements of DOE's Office of Science. This architecture is intended to provide high reliability and very high bandwidth.

Grid middleware services for large numbers of users are hard – but they can be provided if careful attention is paid to scaling. ESnet provides PKI authentication services and world-wide video and audio conferencing to DOE scientists and their collaborators.

## 12 Acknowledgements

---

The ESnet senior network engineering staff that are responsible for the evolution of ESnet consists of Joseph H. Burrescia, Michael S. Collins, Eli Dart, James V. Gagliardi, Chin P. Guok, Yvonne Y. Hines, Joe Metzger, Kevin Oberman and Michael P. O'Connor. The staff responsible for Federated Trust includes Tony J. Genovese, Michael W. Helm, and Dhivakaran Muruganatham (Dhiva). The staff responsible for the Tele-Collaboration services includes Stan M. Kluz, Mike Pihlman, and Clint Wadsworth. This group of people contributed to this paper.

ESnet is funded by the US Dept. of Energy, Office of Science, Advanced Scientific Computing Research (ASCR) program, Mathematical, Information, and Computational Sciences (MICS) program. Mary Anne Scott 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.

## 13 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] LHC Computing Grid Project <http://lcg.web.cern.ch/LCG/>
- [6] [http://www.sc.doe.gov/ascr/20040510\\_hecrtf.pdf](http://www.sc.doe.gov/ascr/20040510_hecrtf.pdf) (public report)
- [7] ASCR Strategic Planning Workshop, July 2003 <http://www.fp-mcs.anl.gov/ascr-july03spw>
- [8] Planning Workshops-Office of Science Data-Management Strategy, March & May 2004 <http://www-user.slac.stanford.edu/rmount/dm-workshop-04/Final-report.pdf>
- [9] ESG - Earth System Grid. <http://www.earthsystemgrid.org/> ESG - Earth System Grid.  
<http://www.earthsystemgrid.org/>
- [10] CMS - The Compact Muon Solenoid Technical Proposal. <http://cmsdoc.cern.ch/>
- [11] The ATLAS Technical Proposal. <http://atlasinfo.cern.ch/ATLAS/TP/NEW/HTML/tp9new/tp9.html>
- [12] LHC - The Large Hadron Collider Project. [http://lhc.web.cern.ch/lhc/general/gen\\_info.htm](http://lhc.web.cern.ch/lhc/general/gen_info.htm)
- [13] The BaBar Experiment at SLAC. <http://www-public.slac.stanford.edu/babar/>
- [14] The D0 Experiment at Fermilab. <http://www-d0.fnal.gov/>
- [15] The CDF Experiment at Fermilab. <http://www-cdf.fnal.gov/>
- [16] The Relativistic Heavy Ion Collider at BNL. <http://www.bnl.gov/RHIC/>
- [17] <http://csrc.nist.gov/piv-project/>