Title

ESnet4: next generation network strategy, architecture, and implementation for DOE Science

Journal

Journal of Physics Conference Series, 46(1)

ISSN

1742-6588

Authors

Collins, Michael
Burrescia, Joseph
Dart, Eli
et al.

Publication Date

2006-09-01

DOI

10.1088/1742-6596/46/1/071

Peer reviewed

ESnet4: Next Generation Network Strategy, Architecture, and Implementation for DOE Science

Michael Collins, Joseph Burrescia, Eli Dart, Jim Gagliardi, Chin Guok,
William Johnston (wej@es.net), Joe Metzger, Kevin Oberman, and Mike O'Connor

Introduction

DOE's Office of Science is the largest supporter of basic research in the physical sciences in the U.S. It directly supports the research of 15,000 PhDs, PostDocs and Graduate Students, and operates major scientific facilities at DOE laboratories that serve the entire U.S. research community: other Federal agencies, universities, and industry, as well as the international research and education (R&E) community. ESnet's mission is to provide the network infrastructure that supports the mission of the Office of Science (SC).

Based both on the projections of the science programs obtained from several SC workshops, and on changes in observed network traffic patterns over the past few years, it is clear that ESnet must evolve substantially in order to continue meeting the Office of Science mission needs.

ESnet has successfully served SC science in the past by creating IP networks composed of a single national core ring, with sites connected by single tail circuits. This model is not scalable however, and must change in order to meet the future science bandwidth and reliability requirements. A new model must provide multiple redundant paths from the core to the sites and the national core must be implemented as two core networks – one core tailored to production IP traffic and a second core tailored to the data flows of SC's large-scale science. Further, this second network must be scalable in capacity so that it can easily increase capacity over the next five years to meet increasing science needs.

This document discusses the development of ESnet's strategy to meet these requirements through a new network architecture and implementation approach.

1 ESnet Today

Today ESnet consists of a national ring that is mostly 10 Gb/s with 2.5 Gb/s across the south. ESnet has six major hubs around the country that house the high-speed core network IP routers and peering routers. The traditional 600 Mb/s tail circuits to the SC Labs have, in the past year, mostly been replaced by metro-area optical rings (MANs) providing the Labs with 20-40 Gb/s connections, not only to the ESnet core but also to major research and education (R&E) network peering points. The Labs connect to the ESnet core directly through the core routers as do the major R&E networks – Internet2/Abilene (US), NLR (US), Dante/GEANT (Europe), SInet (Japan), CANARIE (Canada), AARnet (Australia), AMPATH (S. America), and GLORIAD (Russia and China). The commercial networks that provide general, global Internet connectivity connect through the peering routers at the hubs.

The current generation of ESnet is the third major version of the network – the first was based on 45 Mb/s communications lines, the second on a national, 155 Mb/s ATM network, and the third is the current 10/2.5 Gb/s packet over SONET ring. ESnet4 will be a configurable optical infrastructure built in cooperation with the US research and education community.

2 Requirements for the Next Generation ESnet

A series of SC workshops ([1],[2],[3]) established the requirements for most of the SC programs by looking at flagship applications and facilities. The case studies and requirements developed in the workshops were updated for DOE's Lehman Baseline Review of ESnet in 2006 [4]. The primary network requirements to come out of this process were:

- A continuation of the high-quality, high-speed, Internet service for access to university and international collaborators, and the global Internet;
- Network bandwidth must increase substantially, not just in the core but all the way to the sites and to the attached computing and storage systems to accommodate the massive data flows of the new SC facilities, new approaches to computation simulation, and new, large-scale collaborations;
- Rich and diverse connectivity is needed to accommodate the expanding geographic reach of the SC collaborations;
- The 5 and 10 year bandwidth requirements mean that current network bandwidth has, on average, to more than double every year;
- A highly reliable network is critical for science – when large-scale experiments depend on the network for success, the network may not fail;
- There must be network services that can guarantee various forms of quality-of-service (e.g., bandwidth guarantees) and provide for multiple, reservable, guaranteed bandwidth circuits for transporting high-impact traffic supporting science, coupled computing and storage systems, etc.

In order to accommodate this growth and the change in the types of traffic, the architecture of ESnet must evolve.

3 ESnet4: A New Architecture to Meet the Science Requirements

The strategy for the next generation ESnet consists of a new network architecture that organizes four major network elements and a new network service for managing large data flows.

The architectural principles are:
A) Use ring topologies for path redundancy in every part of the network – not just in the core;
B) Provide multiple, independent connections everywhere to guard against hardware failures;
C) Provision one core network – the IP network – specialized for handling the huge ($3*10^9$/mo.) number of small data flows (hundreds to thousands of bytes each) of the general IP network;
D) 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 account for 30% of all ESnet traffic and will completely dominate it in the near future).

The architecture principles provide structure for the four major elements of the network:
1) A high-reliability IP core network based on high-speed, highly capable IP routers to support
- o Internet access for both science and Lab operational traffic, and some backup for the science data carried by SDN
- o Science collaboration services
- o Peering with all of the networks needed for reliable access to the global Internet

2) Metropolitan Area Network (MAN) rings to provide
- o More reliable (ring) and higher bandwidth (multiple 10 Gb/s circuits) site to core connectivity
- o Support for both production IP and large-scale science traffic
- o Multiple connects between the Science Data Network core, the IP core, and the sites

3) Loops off the core rings to provide
- o For dual site connections where MANs are not practical

4) A Science Data Network core network based on layer 2 (Ethernet) and/or layer 1 (optical) switches for
- o Multiple 10 Gb/s circuits with a rich topology for very high total bandwidth to support large-scale science traffic
- o Dynamically provisioned, guaranteed bandwidth circuits to manage large, high-speed science data flows
- o A second connection to the MAN rings for protection against hub failure in the other core network
- o Dynamic sharing of some optical paths with the R&E community
- o An alternate path for production IP traffic

4 Implementing ESnet4

Many implementation choices were considered, however we describe here the approach that was chosen along with some of the rationale.

4.1 The high-reliability IP core

Delivering IP networking to an ESnet site requires three things:
- o a highly reliable IP core network,
- o site connections to this core IP network
- o connections to all of the other networks needed to provide access to the rest of the world

The strategy to build a highly reliable IP core network has four elements.

The point-to-point circuits

The underlying circuit infrastructure will consist of 10Gb/s point-to-point circuits that interconnect the IP core routers. This circuit infrastructure provisioned and maintained by a commercial telecommunications carrier in order to ensure a high level of reliability through regular maintenance, 7x24 nation-wide coverage, etc.

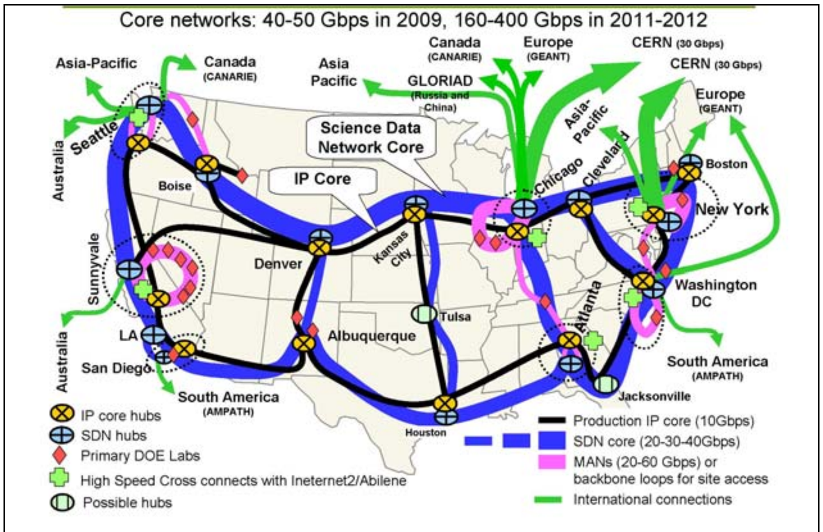

Figure 1. ESnet4 Configuration

The circuits will be provided by optical multiplexing on a pair of fibers that have the required national footprint. The totality of circuits is shared within the R&E community, but ESnet will have some number of them dedicated to its exclusive use. Each circuit is an independent, point-to-point path that can support various layer 2 framing, mostly Ethernet for ESnet.

The core topology

Organizing the circuits in a ring topology produces a cost effective and reliable IP core. A ring architecture is resilient to a single failure, such as a fiber cut or core router interface failure – the IP traffic will simply reroute in the opposite direction around the ring. This ability to reroute past a single failure allows the use of less costly, unprotected circuits. ESnet has used this architecture successfully for the past five years to provide a core network availability of greater than 99.99%.

The core nodes and IP routers

The ESnet core nodes are located in telecom-grade hub facilities that are highly secure, provide redundant sources of uninterruptible power, and provide remote technician services for equipment repair. The core nodes will consist of internally redundant, carrier-class, IP routers. Dual route processors will allow network maintenance with minimal downtime and automatic failover in the event of a processor failure. In addition, there will be secure, out-of-band access to the routers for emergency maintenance.

The approach of using redundant routers in the core nodes was rejected as not being cost effective. ESnet's experience over the past five years using a single, carried-grade core router with a high degree of internal redundancy has been that they are extremely reliable. Additionally, in the new architecture there are changes in the way that sites connect to ESnet (described below) that will eliminate any one hub router as a single point of failure.

4.2 The Science Data Network – Meeting the Bandwidth Requirements of Large-Scale Science

The identified requirements, as well as the current traffic trends, indicate an increasing need for bandwidth dedicated to specific end points (systems). Traffic predictions by the science community and from projections based on the current traffic trends, indicate multiple 10Gb/s connections between experiments and collaboration sites and between computing systems will be required to meet the science needs in the coming years. A large portion of the science data generated and analyzed will be in high bandwidth flows (system to system data transfers) with durations lasting from hours to months. The existence of such high-bandwidth, long duration flows – which we refer to as circuit-like traffic – has been observed in the current network.

Aspects considered when developing a strategy to provide this functionality included:

- The characteristics of the core network
- Meeting the bandwidth requirements
- A virtual circuit reservation and claiming mechanism (described in a later section)

The Science Data Network core

The increases in the total network bandwidth required in the next several years are expected to be mostly for circuit-like traffic involving specific experiments and computing systems. Given the expense of high bandwidth IP routers, and the fact that most of the large-scale science traffic end-points will be relatively static in nature (and therefore the packets in the network all tend to go to the same places and do not have to be routed), alternatives to traditional IP routers can be used to save money.

Layer-2 switches, which are a fifth, or less, the cost of carrier-grade core routers, will be used to manage the circuits involved with voluminous science data traffic. The strategy to build the SDN core network is to deploy a network separate from the IP core and that is comprised of 10Gb/s lambdas terminating in layer 2 switches.

Meeting Bandwidth Requirements

The approach to meeting annually increasing science bandwidth requirements is to build an expandable optical infrastructure on dedicated fiber. The expandability comes from using dense wave (frequency) division multiplexing (DWDM) equipment that has pluggable line cards that "activate" individual 10Gb/s waves (called "lambdas"). Adding these line cards is relatively inexpensive. Such an infrastructure will be built in cooperation with the US R&E community.

4.3 Metropolitan Area Networks – Connecting the Labs

The previous hub and spoke architecture employed by ESnet for site access, that is, point-to-point circuits from the site to the nearest hub router, will not satisfy requirements for reliable, high-speed, scalable connectivity to ESnet sites: the cost of 10Gb/s telecom circuits to sites is prohibitive and they represent a single point of failure. To meet the site connectivity needs ESnet has developed a new strategy for connecting sites to the core networks. The strategy has the following components:

- The MAN architecture
- The MAN switches
- Meeting bandwidth requirements
- Connecting the MANs to the two national networks (IP core and SDN core)
- Loops off of the core rings for remote sites

The MAN Architecture and Implementation

The general strategy for Metropolitan Area Networks (MAN) involves building an optical fiber configured in a ring topology that provides redundant paths to the site. The fiber is provisioned with optical wave division multiplexers capable of instantiating 16-64 10Gb/s channels on the fiber. As in the case of the IP core, the rerouting available at layer-3 permits the use of less expensive unprotected circuits[a].

Diverse physical routing for the east and west[b] lambdas is desired, although a partially collapsed ring may be dictated due to the cost of physically diverse paths. This occurs when some portion of the east and west fiber is carried in the same conduit, typically for the "last mile" to the site – that is, the path from the metro area fiber ring to the site telecom building.

The MAN switches

The MANs are built using less costly layer 2 switches of the same general class as used for the SDN core. The reliability strategy is to use separate port cards for east, west, and site connections. (See lower left of **Error! Reference source not found.** ~~Figure 2.~~) This eliminates an interface card as a single point of failure.

The switches are provisioned with dual supervisor cards, dual power supplies and split power feeds to minimize the downtime associated with maintenance and hardware failures. They are located on the site premise and it is the site's responsibility to provide physical security and an appropriate operating environment.

Meeting bandwidth requirements

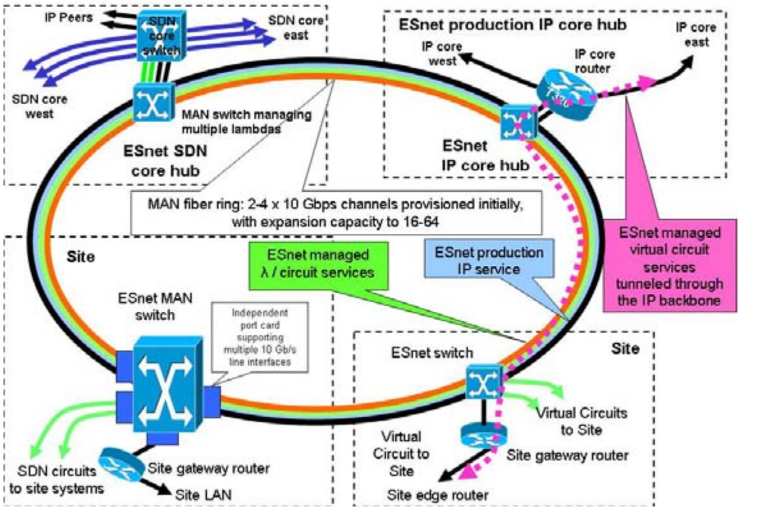
IP Peers
SDN core switch
SDN core east
SDN core west
MAN switch managing multiple lambdas
ESnet SDN core hub
ESnet production IP core hub
IP core west
IP core router
IP core east
ESnet IP core hub
MAN fiber ring: 2-4 x 10 Gbps channels provisioned initially, with expansion capacity to 16-64
ESnet managed virtual circuit services tunneled through the IP backbone
Site
ESnet managed λ / circuit services
ESnet production IP service
ESnet MAN switch
Independent port card supporting multiple 10 Gb/s line interfaces
ESnet switch
Site
Virtual Circuits to Site
Virtual Circuit to Site
Site gateway router
SDN circuits to site systems
Site gateway router
Site LAN
Site edge router

Figure 2. Generic MAN Architecture

The bandwidth requirements for a MAN may be comparable to (or even greater than) the bandwidth requirements of the SDN core (e.g. where a site uses the MAN to reach both ESnet and an R&D network). Bandwidth on a MAN is incrementally added by provisioning additional wavelengths on the ring, which requires a relatively modest investment in transceivers and layer 2 switch interfaces. Adding bandwidth to sites attached to a MAN is typically much less costly than finding additional carrier supplied circuits from the hub to the site. If the requirement for bandwidth exceeds what can be met by adding lambdas, the path forward is to rely on next generation interface technology. It is anticipated that there will be a 100Gb/s interface standard developed in the next five or so year time frame, and prior to that possibility a vendor specific 40Gb/s interface, or an early release of a 100Gb/s interface. A switch upgrade may be necessary to utilize the new interfaces.

Connecting the MANs to the two national core networks (IP and SDN)

The MAN is the local extension of both national cores, IP and SDN. The usual initial implementation is two lambdas (two of the independent 10 Gb/s channels obtained by optical multiplexing) on the fiber ring. One lambda ring is configured as a 10 Gb/s IP ring and the other configured as two 10 Gb/s SDN paths for circuit based SDN traffic.

Ideally, two of the nodes on a MAN ring are separately connected to the IP and SDN core network (as illustrated in Figure 1) This provides increased reliability by having two separate and independent paths from sites to both cores by using separate equipment, and by allowing for failover between the IP and SDN services. This strategy ensures that no single failure, even of a core router/switch, will disrupt all of the connectivity of any ESnet MAN connected site.

Loops off of the core rings for remote sites

Where MANs are not feasible, providing two independent connections to one or both core networks from an ESnet site achieves similar reliability enhancements. See, e.g., the PNNL site in eastern Washington State as shown in Figure 1.

4.4 Managing Large-Scale Data Flows – Virtual Circuit Services

The need for virtual circuit services and quality of service (QoS – e.g. guaranteed bandwidth) was identified as one of the most important new network service by the science requirements workshops ([1], [3]). Additional rationale for deploying circuit services are 1) the high cost to provision dedicated long distance 10Gb/s lambdas, and, 2) the science requirements for virtual circuits typically do not require the full 10Gb/s of bandwidth of a lambda path. To be cost effective the national core capacity available on the SDN must be sharable between many circuits.

[a] "protected circuits" use a layer 1 technology, SONET, and multiple lambdas or multiple fibers to provide automatic layer 1circuit-by-circuit rerouting

[b] The "east" and "west" directions on the fiber refers to the two segments of the ring that come into each node rather than compass directions.

The identified uses and/or benefits of virtual circuits include:

- Traffic isolation and traffic engineering - Permits the use of transport mechanisms that cannot co-exist with commodity TCP-based transport;
- Guaranteed bandwidth (Quality of Service (QoS)) - Addresses deadline scheduling -so that processing does not fall far enough behind that it could never catch up. Guarantees for time critical traffic – real time control of an instrument based on output of the instrument;
- Reduces cost of handling high bandwidth data flows - Separating the traffic to allow lower cost (factor of 5, or more) switches to be used in the SDN core;
- Secure, end-to-end connections between Lab and collaborator systems - VC setups undergo an authentication and authorization process, and the network routing policy excludes other traffic sources from using the circuit.

The functional requirements for virtual circuits (VCs) include:

- Support for user/application VC reservation requests - Information required in the request includes end addresses, the required bandwidth, start time, and duration of the VC;
- Manage allocations of scarce, shared resources - Guaranteed bandwidth VCs, like supercomputer cycles, will be a scarce resource. Strong authentication is needed to prevent unauthorized access to this service and authorization to enforce policy for VC use. Usage data is required for accounting and feedback to a resource allocation mechanism;
- Provide circuit setup and teardown mechanisms on both the IP and SDN cores - Virtual Circuits may be provisioned in both networks in order to reach end points not directly attached to the SDN;
- Enable the claiming of reservations- Users must be able to easily use the VCs;
- Enforce limits - Admission control and ingress bandwidth policing to prevent oversubscription of scarce resources;
- Cross domain compatibility - The environment of science is inherently multi-domain and the service must accommodate end points that are at institutions served by ESnet, Abilene, GÉANT, and their regional networks.

These requirements are being met in ESnet with the On-demand Secure Circuits and Advance Reservation System (OSCARS). OSCARS examines paths for available bandwidth and then reserves bandwidth in the network between the VC end points. The bandwidth is claimed by the user by virtue of identifying characteristics of the traffic being specified in the reservation – that is, the source and destination hosts.

OSCARS guarantees the bandwidth of circuits by maintaining a network-wide database of all reservable and reserved bandwidth and ensuring that neither the SDN bandwidth nor the limits on priority circuits in the IP network are exceeded. Policy based routing in the IP network is used to separate the circuit based and IP production traffic at the ingress interface. Circuit based traffic to or from specified hosts is routed onto an MPLS LSP in the SDN network. Sites typically will use a separate interface for high bandwidth VC traffic.

Cross domain compatibility presents some complex issues. There is a need for a standard network-network management interface (NNI) but no general standard exists at the present. The issue is addressed by developing the virtual circuit services as a collaboration among a number of the R&E serving networks and organizations. The collaboration involves OSCARS [5], Internet2 [6], GEANT (the European equivalent of Internet2/Abilene) and the European regional networks (NRENs) [7], Brookhaven National Laboratory[8], General Atomics[9], SLAC [10], DRAGON (NSF research testbed) [12]~~[1]~~,and Ultra Science Network (DOE research network) [11].

A prototype service has been deployed in ESnet. To date more then 20 beta accounts have been created. More than 100 reservation requests have been processed. The deployment roadmap has the initial production service to beginning in 2007.

Acknowledgements

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. 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-AC02-05CH11231.

Notes and References

[1] High Performance Network Planning Workshop, August 2002 http://www.doecollaboratory.org/meetings/hpnpw

[2] 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

[3] DOE Science Networking Roadmap Meeting, June 2003 http://www.es.net/hypertext/welcome/pr/Roadmap/index.html

[4] “Science-Driven Network Requirements for ESnet” available from Eli Dart (dart@es.net)

[5] ESnet OSCARS webpage: http://www.es.net/oscars

[6] Internet2 BRUW Project: http://discvenue.internet2.edu/wordpress

[7] GEANT PACE Project: http://pace.geant2.net

[8] BNL TeraPaths Project: http://www.atlasgrid.bnl.gov/terapaths

[9] General Atomics QoS Project: http://www.fusiongrid.org/network

[10] SLAC IEPM Project: http://www-iepm.slac.stanford.edu

[11] UltraScienceNet Testbed: http://www.usn.ornl.gov

[12] http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/WebHome