TERENA Networking Conference
16 - 19 May, 2011
Prague, Czech Republic

Motivation, Design, Deployment and Evolution of a Guaranteed Bandwidth Network Service

William E. Johnston, Chin Guok, Evangelos Chaniotakis
ESnet and Lawrence Berkeley National Laboratory, Berkeley California, U.S.A

Paper type

Technical paper

Abstract

Much of modern science is dependent on high performance distributed computing and data handling. This distributed infrastructure, in turn, depends on high speed networks and services – especially when the science infrastructure is widely distributed geographically – to enable the science because the science is dependent on high throughput so that the distributed computing and data management systems will be able to analyze data as quickly as instruments produce it.

Two network services have emerged as essential for supporting high performance distributed applications: Guaranteed bandwidth and multi-domain monitoring. Guaranteed bandwidth service – typically supplied as a virtual circuit – is essential for time critical distributed applications, as most science applications are. Detailed monitoring and active diagnosis are critical to isolating degraded network elements that inhibit "high performance use of the network."

This paper discusses the design of the OSCARS virtual circuit service, its evolution as user experience showed some issues with the original design, and its deployment in a large production network. The result of the deployment is that ESnet is a fully integrated, hybrid packet-circuit network infrastructure.

Keywords

high performance distributed computing and data management, guaranteed bandwidth, high throughput networks, network services, science use of networks, hybrid packet-circuit network

1 Motivating applications

"The Office of Science of the U.S. 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 mission of the Energy Sciences Network – ESnet – is to provide a nation-wide, 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.

ESnet is driven by the requirements of the science Program Offices in DOE's Office of Science. The ESnet Science Requirements Workshops [2] examine the networking needs of major OSC science programs. The science areas requiring high-performance networking include, e.g., climate modeling, chemistry and combustion research, magnetic fusion simulation, and several areas in physics and astrophysics. Major sources of data are the OSC national science facilities: three supercomputer centers, a major environmental lab, a major genomics institute, several nanotechnology centers, several synchrotron light sources[a], the nation's several Tokamak fusion reactors, and several high energy physics and nuclear physics accelerators.

[a] A synchrotron light source is a particle accelerator that is specialized to producing high-intensity, mono-energetic, and
(continued next page)

By mid-2005 ESnet (U.S. national networking serving the Dept. of Energy's Office of Science) was experiencing a fundamental and major change in the traffic patterns on the network: A small number of large data flows were rapidly evolving into a dominate feature of the traffic. These flows were between a relatively small number of sites (institutions involved in large-scale science) and were frequently composed of correlated, parallel data flows. At the same time the science community started requesting service guarantees in order to support time-constrained data movement driven by scientific instruments or by workflows involving large numbers of distributed systems.

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 worldwide. 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. See, for example, [3].

The next section describes an example science environment that drives the networking requirements.

1.1 Data Management in High Energy Physics

One of the major experiments associated with the Large Hadron Collider program (LHC [4]) at CERN is the ATLAS [6] detector.

ATLAS will start by generating several tens of petabytes/year[a] and rapidly ramp up to hundreds of petabytes per year. (The ATLAS experiment has already collected more than 7 petabytes since the LHC started running in March, 2010.) Analysis of this data is performed at research and education ("R&E") institutions around the world. These institutions contribute large numbers computing systems, disk farms, and tape systems, currently providing about 28,000 multi-core computers[b], 39 petabytes of disk, and 50 petabytes of tape storage. The Tier 1 Data Centers receive the detector data from CERN, and in combination, hold a complete set of the data generated by the detectors which constitutes the "working dataset." The role of the Tier 1 centers is perform an initial processing step called reconstruction and then to make this data available to the Tier 2 centers (typically at major universities) for science analysis. The CPUs and disks are distributed among the Tier 1 and 2 systems. The numbers above are for 2010 and will increase by 75-100% over the next two years. There are currently 11 ATLAS Tier 1 centers in Europe, the U.S., Canada, and Taiwan, and about 70 Tier 2 centers.

To understand the demands placed on the network by ATLAS, we briefly consider the distributed system that performs the various analysis and data management operations. The overall functionality of the software of CMS [5], the other major LHC experiment is similar, but the design and implementation differs.

The ATLAS distributed system has two primary components: The Panda job management system [7], which is centralized at CERN, and the DDM distributed data management system [8].

The Panda architecture and workflow is illustrated in Figure 1 and described in [7].

Panda accepts several types of physics data processing and analysis jobs for execution. Analysis sites – Tier 2 and Tier 3 centers – report current status (availability of CPUs) to Panda, and DDM knows where the data is located and is responsible for moving the data to the site where the job will be executed.

The job broker examines the data required for each job and instructs DDM to move that data to a site where CPU resources will be available. When the data is in place, the broker dispatches jobs to the site to be executed.

---

nearly coherent beams of light, usually in the X-ray spectrum. Each accelerator will have 10s of beam ports where science groups set up experiments. The experiments involve all sorts of ultra-high resolution imaging, e.g. 3D imaging of biological sub-cellular structures and nano-lithography for advanced semiconductors.

[a] 1 petabyte = 1,000,000 gigabytes

[b] This number is based on http://lcg.web.cern.ch/LCG/Resources/WLCGResources-2010-2012_04OCT2010.pdf, which gives the CPU resources in terms of HEP-SPEC06 units - the new HEP-wide benchmark for measuring CPU performance. Modern systems seem to be about 8 HEP-SPEC06 per core, so a quad core system will deliver about 32 HEP-SPEC06. From this the number of computing systems involved is estimated, assuming an average of 4 cores / system.

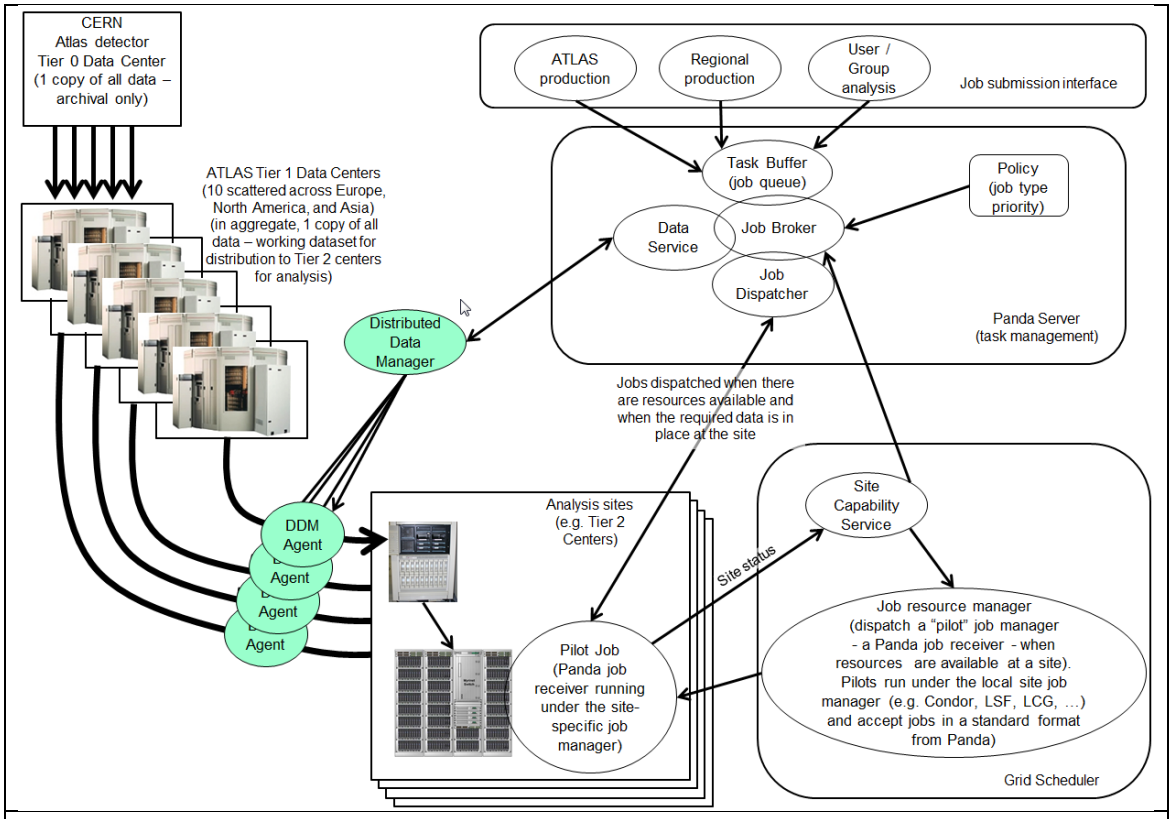

Figure 1. The Panda job management system.

The Panda server is the job coordinator. All job requests are submitted to the server which prioritizes and assigns work to the computing systems pool based on job type (e.g. "production" jobs do the first level of analysis that produces the data needed by all other jobs, and so get a high priority), other priorities (e.g. as assigned by the physics groups that provide the computing resources), and the availability and location of the input data.

At the site, the CPU resources (large clusters) are managed by instantiating a "pilot" via the cluster job manager - such as CondorG, PBS, etc.[a] Pilots are local Panda job managers that accept jobs in a standard format from the Panda server. The pilots are job managers (for Panda jobs) within a job manager (the local cluster job manager). Pilots execute as long as there is work available from the Panda server. When there are no jobs to execute, the pilot exits and makes the cluster available for other purposes. As the Panda server indicates that there is more work to be done, a Grid scheduler re-instantiates a pilot job on the cluster.

As DDM – the data management system – moves the required data to the site where Panda has dispatched jobs that require that data. The Pilots execute jobs that are queued and whose data is available; perform housekeeping tasks related to various sorts of job failures; send output files to their destinations, etc. All of these components operate as a job pipeline, many elements of which operate in parallel.

The data management part of the system ("DQ2" [8]) provides the logical organization of the data into datasets (associated collections of files), does data discovery, keeps track of dataset replicas, data transfer coordination, and does monitoring. Like Panda, DQ2 is designed to make use of several underlying site or collaboration specific data transfer mechanisms. Once the "best" copy of the requested dataset is located, DQ2 dispatches an agent to coordinate data transfer from storage to the computing facility where the requesting job has been scheduled.

[a] CondorG and PBS are both local queue management systems that manage administratively homogenous resources – e.g. all of the computing systems managed by the physics group at an institution hosting a Tier 2 center. CondorG: http://www.cs.wisc.edu/condor/condorg/; PBS: http://en.wikipedia.org/wiki/Portable_Batch_System.

The "sites" in Figure 1 are the institutions that provide the computing and disk storage resources – in the U.S. these are primarily universities scattered across the country. For the U.S. portion of the ATLAS collaboration, Panda manages about 15,800 cores (which is the computing element to which work is assigned). These approximately 4,000 computing systems are essentially fully loaded all of the time that the LHC and the ATLAS detector operate – about 9 mo/yr. The U.S. ATLAS collaboration accounts for about 25% of the total ATLAS computing and storage resources.

The amount of data moved among the 70+ institutions involved in the analysis is considerable, averaging about 116 terabytes/day (Figure 2), or about 1.4 gigabytes/sec average steady state transfer rate, for a cumulative total of almost 3.5 petabytes (350,000 gigabytes). (These quantities based on 30 days starting Nov. 15, 2010.) The data rate numbers cited here are application throughput, not network bandwidth. The 1.4 gigabytes/sec of application throughput requires about 15 gigabits/sec of network bandwidth.

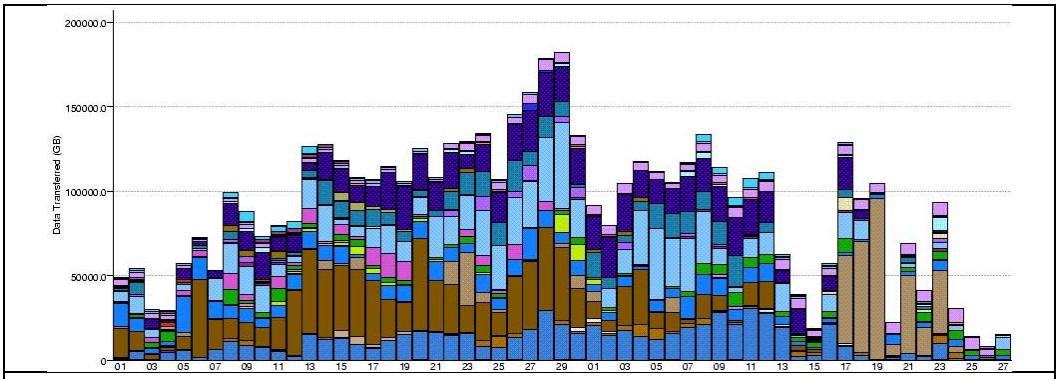

Figure 2. Data transferred (gigabytes) among all ATLAS sites, per day, from Nov. 1 to Dec. 27, 2010. (From the ATLAS dashboard (system monitor). Note that the LHC shutdown for the winter maintenance period on Dec. 13. The bars are broken down Tier 1 data centers.)

In 2010 the accumulative data volume moved by ATLAS Panda and DDM was almost 7 Petabytes. (http://dashb-ATLAS-data.cern.ch/dashboard/request.py/site)

The data transfers shown in Figure 2 provided input for between 35,000 and 50,000 jobs executing at any one time.

The intent of this discussion is to clearly demonstrate that the scale of the LHC experiments data processing (using high-speed networks and widely distributed Grid-based data and CPU resources) is the largest attempted by the science community up to this point. (The other major LHC experiment – CMS – is comparable to ATLAS in the CPU and data resources and data movement.)

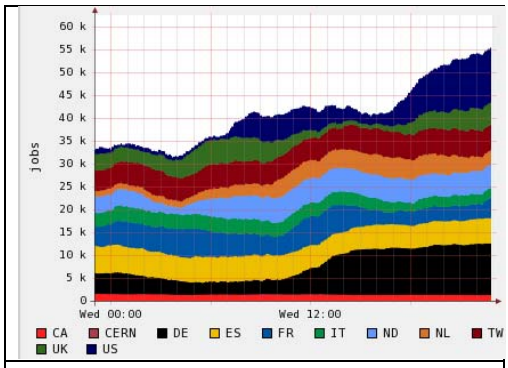

Figure 3. Number of running Panda jobs running world-wide, one day by hour. (http://panda.cern.ch:25880/server/pandamon/query?dash=prod)

It is also intended to demonstrate that network service guarantees are an essential element in these systems that depend on moving very large amounts of data over long distances as an integral part of the workflow that does the physics analysis of the data. One of the important issues for reliable network transfer is that the data rates from the instruments and the time to process that data are used to provision the overall collection of compute resources, and there is not much excess capacity in this collection. Therefore, if the network fails or slows down, even though the data can be buffered for several days at the head end, and even if there is sufficient excess capacity in the network to move the buffered data together with the current data when the network comes back up, with little excess capacity in the compute resources it may take an unacceptably long time to catch up.

1.2 Network implications

Distributed application systems such as the ATLAS and CMS data analysis systems:

- are data intensive and high-performance, normally moving tens to hundreds of terabytes a day;

- are high duty-cycle, operating 24 hours/day for most of the year in order to meet the requirements for data movement;
- have compute and data storage components that are typically distributed over continental or inter-continental distances;
- always involve operation across multiple network administrative domains, and;
- depend on network performance and availability to ensure the efficient functioning of the distributed workflow systems that manage the data movement and analysis tasks.

This requires that the distributed application system components have access to network services that can:

- provide guarantees from the network in order to ensure that there is adequate bandwidth to accomplish the task within the application time constraints;
- interoperate with compatible services in other domains in order to provide an end-to-end service;
- provide real-time performance information from the network that allows for graceful failure and auto-recovery, and adaptation to unexpected network conditions that are short of outright failure, and;
- perform end-to-end monitoring in a multi-domain environment that can view and assess the state of all of the intra-domain segments that make up the multi-domain end-to-end path since problems in any domain will impact end-to-end performance.

These network services must be available in an appropriate programming paradigm; that is, within the Web Services / Grid Services paradigm that is the framework of most distributed analysis, science applications systems.

2 ESnet: A Brief Overview

To provide some context for the environments in which the services being discussed are implemented, we briefly describe the architecture and implementation of ESnet. This is additionally useful because ESnet's architecture and implementation is typical for one of the two styles of implementation of most of the large research and education networks in the U.S. and Europe.

ESnet is a self-contained network administrative domain. That is, the equipment and telecommunication circuit infrastructure are all managed, operated, provisioned, secured, etc., by ESnet staff. ESnet peers with essentially all the world's research and education networks at the major U.S. R&E exchange points (called GigaPoPs) in New York, Chicago, Seattle, Sunnyvale (San Francisco), Los Angles, Atlanta, and Maryland (Washington DC). ESnet also has a Vienna, Austria peering location. ESnet is a Tier 1 ISP and peers directly with most commercial networks, primarily at three Equinix exchanges around the U.S.

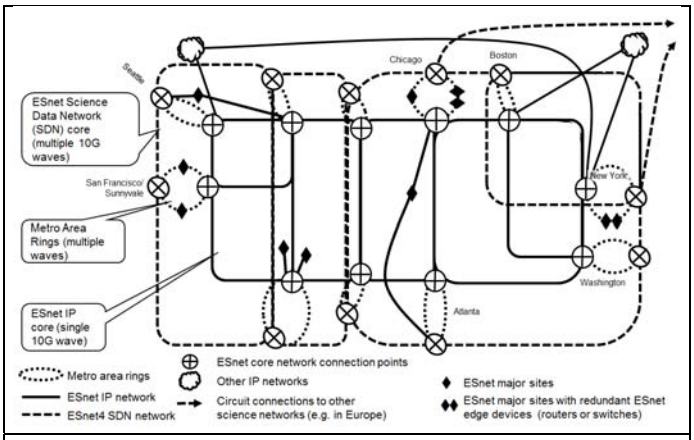

Figure 4. ESnet4 architecture showing the separation of the general IP network traffic and the mostly OSCARS circuit-based Science Data Network traffic. The IP network uses one 10Gb/s optical path and the SDN and metro area rings are multiple 10Gb/s paths.

To address the above science needs, ESnet4 – the fourth technology generation of ESnet – was designed as a hybrid packet-circuit network consisting of two core networks: (1) an IP core that carries all the general/commodity IP traffic; and (2) a circuit-oriented core (called the Science Data Network or SDN) that is primarily designed to carry large scientific data flows [9].

The ESnet network is a national infrastructure with a richly interconnected topology built from multiple 10 Gbps optical circuits. These optical circuits interconnect a collection of PoPs (points of presence) in most major U.S. cities and at national and international R&E exchange (peering) points. The optical infrastructure covers most of the U.S. in six interconnected rings. One 10 Gbps footprint on the core network is dedicated to general IP traffic and all other 10 Gbps links are devoted to the SDN. At the current time SDN provides 20-40 Gbps on the national network, and at the current rate of increase will provide 40–60 Gbps by 2011and 100 Gbps by 2012. (ESnet is currently in transition to ESnet5 which will use 100 Gb/s waves.) Additionally, all of the DOE national

laboratories – ESnet's primary user sites – are dually connected to the core, mostly by a collection of metro area optical rings in the San Francisco Bay area, Chicago area, and New York–Long Island area. Labs not in these metro areas are connected by loops off the core network (see Figure 4 and Figure 5).

Both the OSCARS circuit service described in this paper and the perfSONAR monitoring systems are integrated into the ESnet production network.

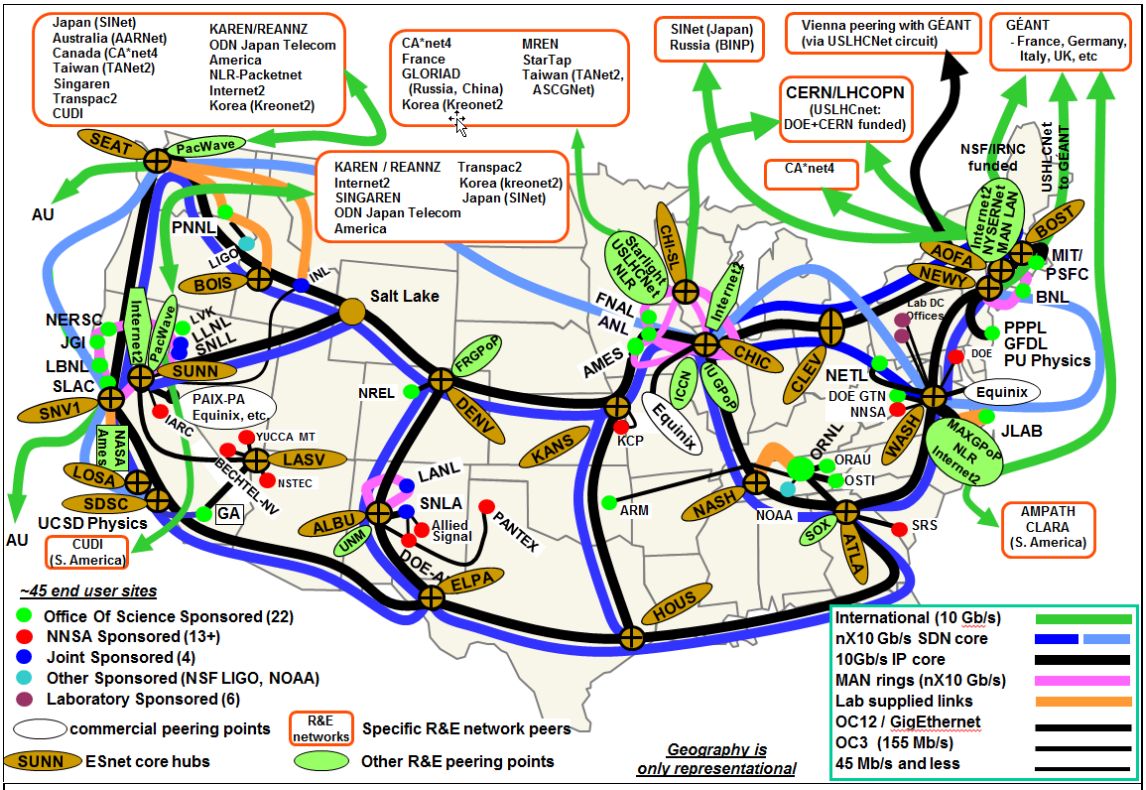

Figure 5. ESnet4 in 2010.

3 Guaranteed bandwidth network service

Our understanding of the requirements of modern science for new network services has emerged from two processes. One is the observation and analysis of historical trends in traffic patterns in ESnet, and other is detailed discussions with science projects about how their analysis and simulation systems actually work (where the data originates, how many systems are involved in the analysis, how the systems and collaborators are distributed, how much data flows among these systems, how complex is the work flow, what are the time sensitivities, and so forth – see [2]).

By 2005 ESnet was seeing the most profound change in traffic patterns since the network was created in the mid-1980s. The traffic was no longer showing a smooth exponential growth reflective of a large number of relatively small flows, but was now being dominated by a small number of flows of large-scale science. The dual impact of this was that first the overall traffic volume over months was no longer at all smoothly increasing, but rather had become very irregular as 1000 or fewer flows (out of the billions of small flows) were now producing almost 50% of the total traffic. (See Figure 6.) Second, these big flows had a very small number of distinct end points and they were very long-lived. In other words, perfect candidates for a circuit-like service rather than having every packet routed by an expensive core router.

As seen above, large experiment applications are typically 1) data intensive, high-performance, and high duty-cycle in order to meet requirements for data movement and analysis, and; 2) widely distributed among multiple institutions that are typically spread over continental or intercontinental distances. Considering the overall requirements, a set of generic, but important, goals can be identified for networks in order to support large-scale science [9]:

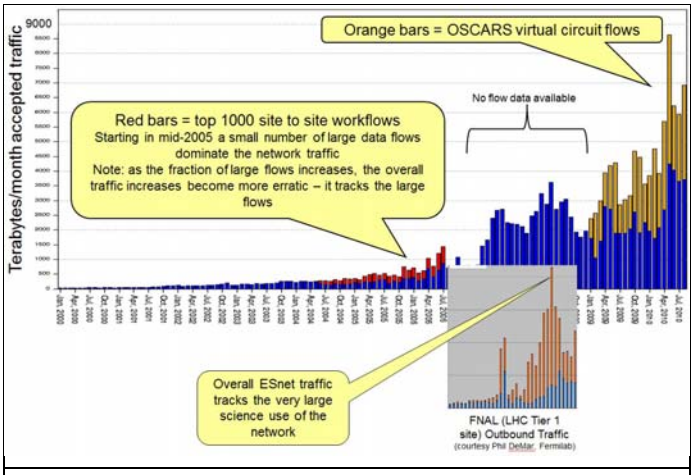

Figure 6. Historical traffic trends: The impact of large science data flows on overall network traffic.

- Bandwidth: Adequate network capacity to ensure timely and high-performance movement of data produced by the facilities.
- Reliability: High reliability is required for large instruments and "systems of systems" (large distributed systems) that now depend on the WAN (wide area network) network for data distribution and communication between computer systems. Further, reliability includes individual paths in the network operating essentially error-free in order to support sustained, high-bandwidth (10 Gb/s), long-distance (thousands of kilometers) data transfer.
- Connectivity: The network must have the geographic reach — either directly or through peering arrangements with other networks — sufficient to connect collaborators and analysis systems to experiment sites. Further, any services that are to be useful to science must be end-to-end services – that is, compatible services must exist in all of the network domains between the end points.
- Services: Guaranteed bandwidth, traffic isolation, end-to-end monitoring, etc., are required as network services and these must be made available to the users in the context of web services, SOA (service oriented architecture), the Grid, and "systems of systems," that are the programming paradigms of modern science.

In addition, the nodes of the distributed application systems must be able to get guarantees from the network that there is adequate network capacity over the entire lifetime of the task at hand. The systems must also be able to get performance and state information from the network to support end-to-end problem resolution, graceful failure, auto-recovery, and adaptation due to unexpected network conditions that are short of outright failure.

4 The OSCARS (On Demand Secure Circuits and Reservation System) virtual circuit service

4.1 Goals and constraints

When the requirements studies were analyzed, use-cases constructed and examined, and constraints for implementation identified, the OSCARS service goals resulted:

- Configurable: The circuits are dynamic and driven by user requirements (e.g. termination end-points, required bandwidth, sometimes routing, etc.).
- Schedulable: Premium service such as guaranteed bandwidth will be a scarce resource that is not always freely available and therefore is obtained through a resource allocation process that is schedulable.
- Predictable: The service provides circuits with predictable properties (e.g. bandwidth, duration, reliability) that the user can leverage.
- Reliable: Resiliency strategies (e.g. re-routes) that can be made largely transparent to the user should be possible.
- Informative: The service must provide useful information about reserved resources and circuit status to enable the user to make intelligent decisions.
- Geographically comprehensive: OSCARS must interoperate with different implementations of virtual circuit services in other network domains to be able to connect collaborators, data, and instruments worldwide.

- Secure: Strong authentication of the requesting user is needed to ensure that both ends of the circuit are connected to the intended termination points; the circuit must be managed by the highly secure environment of the production network control plane in order to ensure that the circuit cannot be "hijacked" by a third party while in use.

In addition to the goals, a number of constraints were identified as well, relating to the available technology, the need to integrate with the operating environment of the existing ESnet, and the fiscal reality that large amounts of new funding could not be required.

The service must provide user access at both layers 2 (Ethernet VLAN) and 3 (IP). There are several reasons for this. First, there are sites that will want to use the service that do not have a layer 2 connection to ESnet, hence the needs for the service to be provided on the IP network. Second, for individual end-users it is likely to be considerably easier to utilize the service at layer 3 in order to provide virtual circuits to individual user systems. In order to keep large data flows off of the general IP network, these circuits will be moved from the IP network to the circuit-based SDN network at the first available opportunity (that is, the first time that a direct path is available from the IP network to the SDN network – likely the first ESnet PoP that the site connection reaches. Most of the use of OSCARS circuits was expected to be Ethernet VLANS between site/user routers that manage access to the circuit with site managed BGP, and this has turned out to be the case.

ESnet uses only layer 2, 2.5, and 3 devices (Ethernet switches, MPLS switches, and IP routers) and therefore the implementation must not require layer 1 TDM (time-division multiplexing) equipment (such as SONET/SDH with VCAT / LCAS) in order to provide bandwidth management. This constraint is due to the large capital and operating costs associated with introducing this type of new hardware into the network.

For inter-domain (across multiple networks) circuit setup no RSVP-style signaling across domain boundaries will be allowed. This is because circuit setup protocols like RSVP do not have adequate (or any) security tools to manage (limit) what RSVP requests from an external network domain can do inside your domain. (RSVP is used internally where there is a uniform policy regime across all devices.) Cross-domain circuit setup has to be accomplished by explicit agreement between autonomous circuit controllers in each domain – whether to actually set up a requested cross-domain circuit is at the discretion of the local controller (e.g. OSCARS) in accordance with local policy and available resources. Inter-domain circuits are terminated at the domain boundary and a separate, data-plane service used to "stitch" the circuits together into an end-to-end path.

4.2 Implementation approach

Given the goals and constraints it was decided that OSCARS would provide an MPLS mediated "virtual circuit" service. The user sees a dedicated path that has circuit-like properties: It is not shared, it provides traffic isolation, it has fixed bandwidth, and is available both in an Ethernet VLAN environment and in the IP network. This type service is sometimes also called a pseudowire service.

A top-down-bottom-up approach to designing and implementing OSCARS starts with the user requirements (given above) and the capabilities of the network devices. Once that it has been determined that there are a set of network capabilities that can provide the user requirements, then those capabilities – together with any other constraints that have to be taken into account – are abstracted to control plane functions.

Given the devices used in the ESnet network, there are a collection of tools, capabilities, and an operational stance that are available to support the service capabilities.

- OSPF-TE (Open Shortest Path First-Traffic Engineering) refers to the traffic engineering tools supported by the OSPF routing that is used in the core of ESnet. OSPF-TE discovers the complete physical topology of the network and then encodes and delivers the topology to a third party (OSCARS, in this case).
- RSVP-TE (Resource Reservation Protocol - Traffic Engineering) is an extension of the resource reservation protocol (RSVP) for traffic engineering. It supports the reservation and provisioning of resources across the network.
- MPLS transport accommodates both layer 3 (IP) traffic and layer 2 (Ethernet) circuits that can encapsulate IP packets and both "typical" Ethernet transport and generalized transport / carrier Ethernet functions such as multiple ("stacked") VLAN tags ("QinQ")
- MPLS-TE is not used to determine the path of the virtual circuit through the network. The Constrained Shortest Path First (CSPF) path routing calculations that typically would be done by MPLS-TE mechanisms are instead done by OSCARS due to additional parameters/constraints that must be accounted for (e.g. future availability of link resources, policy enforcement).
- Multiple levels of priority queuing for the virtual circuit traffic to ensure unimpeded throughput.

A number of services must be added to the available tools to meet the goals for OSCARS.

When a new circuit is requested, OSCARS must not only consider the current state of the network in determining the routing, but also consider all future commitments – e.g. circuits that have been reserved but are not scheduled to be instantiated until some point in the future. In other words, the CSPF (Constrained Shortest Path First) calculations are actually done by OSCARS in order to account for additional parameters/constraints that are beyond the scope of MPLS-TE (e.g. future availability of link resources).

Once OSCARS calculates a path, RSVP is used to signal and provision the circuit using strict hop-by-hop route information to explicitly traffic engineer the path within the network from ingress to egress.

Service guarantee mechanisms require several things. Link bandwidth usage by the reserving party must be managed to prevent link oversubscription by circuits. There must be a way to specify policy, at least in the form of the maximum bandwidth available for circuits on a link, but also allow for other uses of the link. Enforcement of this policy is effectively an admission control mechanism: Once the available bandwidth is committed subsequent requests are denied.

The OSCARS path setup, management, and operational mechanisms are illustrated in Figure 7.

Security services to be added include strong authentication for reservation management and circuit endpoint verification and authorization in order to enforce resource usage policy. Circuit path security/integrity is provided by the high level of operational security of the ESnet network control plane that manages the network routers and switches that provide the underlying OSCARS functions (RSVP and MPLS)

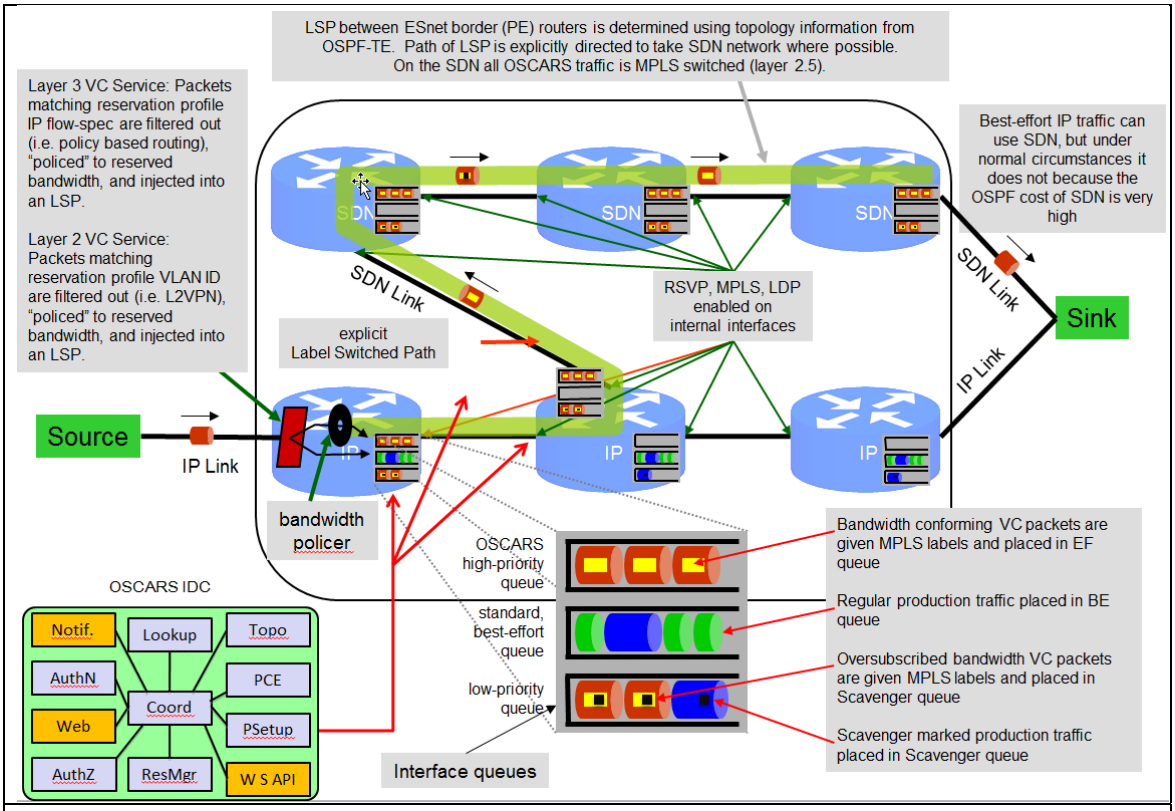

Figure 7. Network mechanisms underlying ESnet's OSCARS

4.3 The OSCARS design

OSCARS uses MPLS to transport the circuit payload and its implementation is essentially an enhanced management and control system ("plane") for MPLS.

The key issues in implementing the OSCARS service are routing and utilization management, path setup, circuit reliability, and inter-domain circuits.

4.3.1 Routing and utilization management

The general approach to routing a user-requested virtual circuit and management of network resources is based on a temporal link topology database. This database contains information on all links available for circuits. The maximum capacity of those links available for circuits (which is typically set by policy rather than technology) and the amount of the link that is committed to existing circuit reservations.

The underlying link topology information is obtained using the OSPF-TE extension of the OSPF routing protocol that is used in the core network.

Requests for new circuits are processed by running a CSPF routing algorithm that identifies the shortest possible path between the virtual circuit endpoints taking into account the link-by-link constraints. These constraints include the bandwidth available on the link for the entire duration of the reservation request. That is, the available bandwidth between reservation start and end taking into account the bandwidth already committed to other reservations, both now and into the future as far out as the end of the longest extant reservation. This implies that the link topology database must have a link-by-link temporal component in order to fully represent all existing reservations. (These reservations are essentially a collection of associated links, bandwidth reserved on those links, and a start and end time.)

This approach provides one element of utilization management: It provides information for admission control, with the "caller" (requestor of the reservation) getting a "busy signal" if there is not path though the network that satisfies the reservation request.

4.3.2 Path setup

Once a path for the virtual circuit is determined (assuming the request is consistent with available link capacities), it is set up link-by-link through the network using RSVP-TE to construct the MPLS Label Switched Path (LSP) that defines the virtual circuit provided to the user. That is, RSVP "walks" the path link by link, and at each router that interconnects the links, it sets up MPLS label switching entries that switch MPLS packets from the input port to the output port[a] on the routers that interconnect the links of the path. This process defines, in MPLS parlance, a Label Switched Path (Figure 8). This LSP is the transport or tunnel mechanism that contains the data flow of the user virtual circuit. The traffic within the LSP is isolated from all other traffic in the network by using distinct logical queues for bandwidth guaranteed circuit MPLS packets, and routed IP packets. This allows, e.g., for the use of an aggressive IP transport protocol for transatlantic data transport, that, if it showed up on the routed IP network, would be dropped by the routers as "dangerous" in the sense that it would not compete fairly with all other IP traffic for queuing resources.

At the data transport layer, which is the transport service offered to the user, the circuit can be established at layer 2 as a tagged VLAN or at layer 3 as special routing applied to the IP address of the source (the science system) that directs the OSCARS reservation packets to the MPLS tunnel.

The bandwidth guarantees are provided (1) by forwarding the virtual circuit traffic into a dedicated logical output queue and elevating its queuing priority, (2) by doing admission control so that no link that carries OSCARS circuit traffic is ever oversubscribed, and (3) by managing the traffic flowing into each virtual circuit (e.g. by rate limiting the virtual circuit input data to the reservation requested bandwidth). Together these ensure that the circuit traffic has priority over any other traffic on the link and that OSCARS circuits do not interfere with each other.

[a] MPLS packets are switched – a relatively low overhead operation – rather than routed as IP packets are. That is, each IP packet passing through a router must have its destination address parsed and then a decision made by the router about where to send the packet to get it closer to its final destination. MPLS switching is a capability that is typically implemented in high-end IP routers rather than in dedicated MPLS switches. This, together with the way that MPLS paths are routed, is why MPLS is sometimes referred to as a layer "2.5" protocol: While it is switched (usually done in a layer 2 device) the switching is typically done in a router (a layer 3 device). MPLS and the new "OpenFlow" switches (www.openflowswitch.org) have many features in common.

The bandwidth that OSCARS can use on any given link is set by a link policy so that the link can be shared with other uses. This allows, e.g., for the IP network to backup the OSCARS circuits-based SDN network (OSCARS is permitted to use some portion of the IP network) and similarly for using the SDN network to backup the IP network, where the SDN link policy might set the maximum circuit reservation-based traffic at, say, 85% of link capacity allowing 15% for IP traffic to backup the IP network.

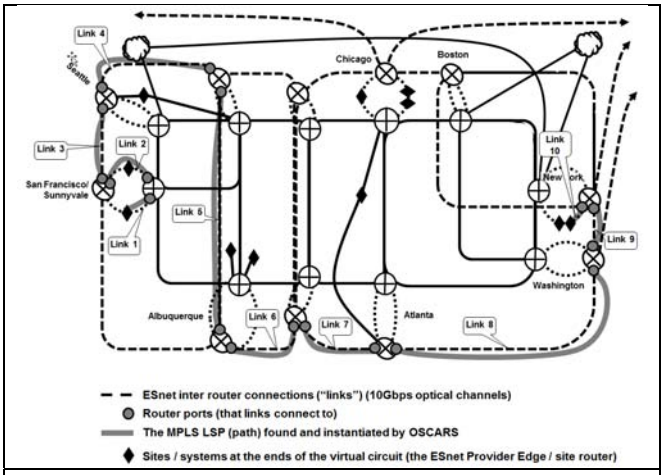

Figure 8. The relationship between links and ports (physical entities) and an OSCARS paths (virtual). Note that the particular routing (chosen links) implies that there is not enough bandwidth on the geographically shorter combination of links in the SF Bay metro ring and on the Sunnyvale – Albuquerque inter-city link.

4.3.3 Circuit reliability

In order to provide high reliability, the user can request a second circuit that is, to the extent possible, diversely routed from the first circuit. These circuit pairs are typically used as virtual private networks (VPNs) that interconnect private IP routing "clouds." By using IP routing to manage connectivity over both the circuits, if one circuit fails the IP packets are simply routed over the remaining path. This is a common way for critical functions such as the LHC tier 1 data centers to use OSCARS virtual circuits.

4.3.4 Inter-domain virtual circuits

An important aspect of the virtual circuit service is that it is only useful if it provides end-to-end guaranteed throughput across multiple network domains, because essentially all science data flows originate in one domain (e.g., a national lab on ESnet or at CERN) and terminate in another domain (e.g., a science group on a U.S. or European campus).

In order to provide this capability, a group of R&E networks (the "DICE" collaboration) has defined an Inter-Domain Control Protocol (IDCP) [10]. The IDCP has standardized the information and messages needed to set up end-to-end circuits across multiple domains. That is, for the exchange of topology information containing at least potential virtual circuit (VC) ingress and egress points, how to propagate the circuit setup request, and how data plane connections are facilitated across domain boundaries. OSCARS, thus, is an Inter-Domain Controller (IDC).

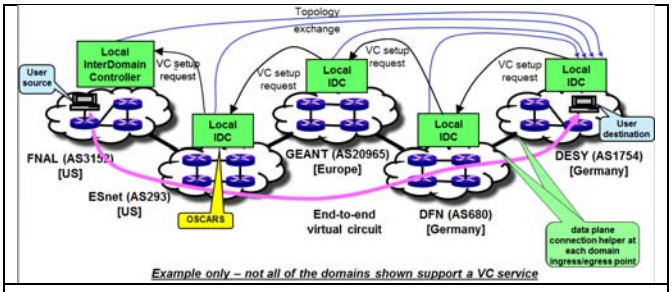

Figure 9. Inter-domain virtual circuits

It should be noted that while the IDCP does coordinate the specifics of the data plane technology that "stitches" two network domains together (e.g. Ethernet VLANs), it does not dictate how a VC is provisioned within a domain. So, for example, for layer 2 data plane connection at the ESnet boundary, the MPLS payload would be de-encapsulated and presented as Ethernet frames.

While OSCARS and its derivatives are fairly widely used in other network domains, there are several IDCs based on different approaches, and all of these have been demonstrated to interoperate within the U.S. and internationally [10]. Standardization of this approach is being undertaken within the Open Grid Forum (OGF) in the Network Services Interface (NSI) working group [12].

4.4 Software architecture

The OSCARS software architecture is illustrated in Figure 10.

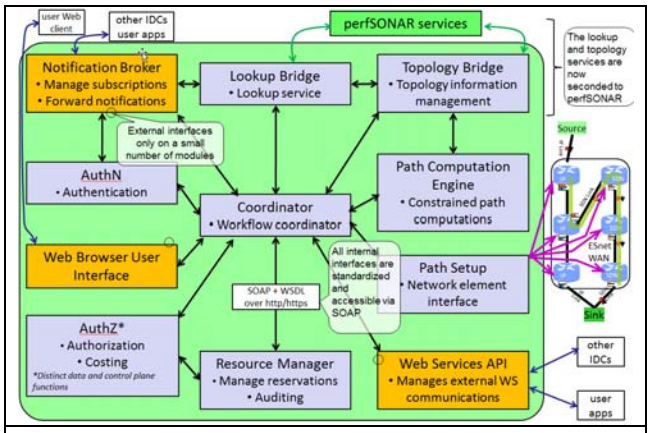

Figure 10. OSCARS v 0.6 architecture

The Notification Broker provides status information such as when a circuit setup is completed. At the present this service is primarily used for inter-domain circuit setup (described above).

The lookup service is used to provide a standard naming mechanism for the circuits. While this was originally provided by OSCARS it has now been seconded to perfSONAR (a monitoring service widely used in the same community as OSCARS – see [13]) and is accessed using the Lookup Bridge.

The topology service provides a standard way to describe the links that are available for OSCARS reservations. This has also been seconded to perfSONAR and is accessed via the Topology Bridge.

The Path Computation Engine is where the constrained routing is done. This takes into account link utilization policy and current and future reservation commitments.

The Path Setup is essentially a device driver for the specific hardware in the network. It takes the route calculated by the path computation engine and uses it to instantiate the circuit in the network. In the case of ESnet's MPLS devices, it uses RSVP to set up the MPLS Label Switched Path. It also installs the rate limiters at the ingress point. Other networks use OSCARS but have replaced the Path Setup module with an implementation that is specific to their network devices and management style.

The Coordinator manages the OSCARS internal workflow.

OSCARS v0.6 is the third rewrite of the code base. It restructures the code to increase the modularity and expose internal interfaces so that the community can start standardizing IDC components. For example there are already several different path setup modules that correspond to different hardware configurations in different networks.

5 Evolution of the service: User-level traffic engineering/management

Initially, user reservations were hard-limited to the requested bandwidth value. It soon became apparent that this was too restrictive. A change was made in the service such that the virtual circuits are rate-limited at the ingress but are permitted to burst above the allocated bandwidth if idle capacity is available. This is accomplished without interfering with other circuits, or other uses of the link, by marking the over-allocation bandwidth as low- priority traffic[a]. This seemingly small change has had surprisingly broad consequences.

Now the users can both provide high reliability through redundancy and they can develop and implement their own capacity usage models. In order to provide high reliability, the user can request a second circuit that is diversely routed from the first circuit. Further, a third circuit that is normally used for other purposes (such as normal IP traffic) can be used to partly backup either or both of the first two. OSCARS backup circuits are frequently configured on paths that have other uses, including backup paths for other OSCARS circuits and/or carrying commodity IP traffic, and so reduce available bandwidth during failover, which is why users want to develop their own capacity models. The extent to which OSCARS circuits can intrude in the normal traffic is configurable by the user. Together with user BGP (Border Gateway Protocol) sessions managing the input to the circuits, the new semantics provide the users with powerful user-level traffic engineering capability, allowing them to establish their own failover, re-purposing, and fair-sharing strategies among several different applications or users.

[a] All ESnet devices support at least four levels of queuing: 1 is low priority (and used for what is called sometimes scavenger traffic), 2 is normal or best effort and is where all usual IP traffic is queued, 3 is elevated priority and is where all OSCARS MPLS packets that are within the bandwidth specification are queued, and 4 is top priority and is used only for network control traffic.

Making and implementing the sharing decisions that allows the user to define the capacity model, is a traffic engineering capability that OSCARS makes available to the experienced site like the LHC Tier 1 Data Centers

The example of Figure 11 is not speculative – it is routinely used by the two ESnet sites (Brookhaven and Fermilab) that are the LHC Tier 1 data centers in the US.

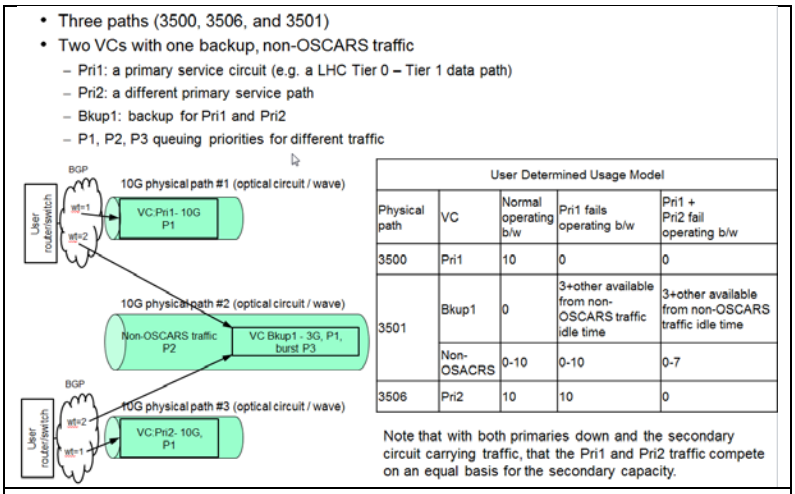

Figure 11. Example of user-defined traffic engineering

6 Deployment of the services

OSCARS is fully deployed in ESnet and can manage resources on all of the roughly 130 routers and switches in ESnet. It is a production service and is integrated into the Spectrum network monitor that ESnet uses. Additionally, it has the necessary tools to provide detailed circuit and site configurations to the users.

OSCARS currently manages 31 long-term circuits that serve four science disciplines: high energy physics (the LHC), climate research, computational astrophysics, and genomics. All of LHC Tier 1 data paths within the U.S. utilize OSCARS circuits. OSCARS also manages thousands of short-lived reservations over the course of a year, most of which are set up by traffic management systems at the sites using the OSCARS Web Services interface.

The virtual circuit service for collaborative science is only useful if it provides end-to-end guaranteed throughput across multiple network domains, because essentially all science data flows originate in one domain (e.g., a national lab on ESnet) and terminate in another domain (e.g., a science group on a U.S. or European campus).

OSCARS and a few other systems that provide virtual circuits are being fairly widely deployed in the campus, regional/national and international networks that support distributed science and the R&E communities, including ESnet. Through the compatibility provided by the DICE ICDP work, the services provided are a key element in the infrastructure that enables large-scale collaborative science projects.

7 Summary

OSCARS has provided a significant new network service that allows the network to be scheduled and provides guarantees in a service-oriented software environment. OSCARS in ESnet uses MPLS as the transport mechanism, but it has been ported to several other network environments that use different transport and bandwidth guarantee mechanisms, for example SONET/SDH with VCAT / LCAS (in USLHCnet) and in GMPLS controlled networks, e.g. using the DRAGON controller [14].

OSCARS is jointly developed in an informal international consortium and several dozen R&E networks around the world use OSCARS or variations of it.

8 Acknowledgements

This work was supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Notes and References

[1] http://www.energy.gov/, Science and Technology tab.

[2] Science Requirements for ESnet Networking, http://www.es.net/hypertext/requirements.html

[3] LHC Computing Grid Project http://lcg.web.cern.ch/LCG/

[4] LHC - The Large Hadron Collider Project. http://lhc.web.cern.ch/lhc/general/gen_info.htm

[5] CMS - The Compact Muon Solenoid Technical Proposal. http://cmsdoc.cern.ch/

[6] The ATLAS Technical Proposal. http://ATLASinfo.cern.ch/ATLAS/TP/NEW/HTML/tp9new/tp9.html

[7] T. Maeno, "PanDA: Distributed production and distributed analysis system for ATLAS." Computing in High Energy and Nuclear Physics (CHEP), 2007. Available at http://iopscience.iop.org/1742-6596/119/6/062036

[8] M. Branco, D. Cameron, B. Gaidioz, V. Garonne, B. Koblitz, M. Lassnig, R. Rocha, P. Salgado, T. Wenaus, on behalf of the ATLAS Collaboration, "Managing ATLAS data on a petabyte-scale with DQ2." Computing in High Energy and Nuclear Physics (CHEP), 2007. Available at http://iopscience.iop.org/1742-6596/119/6/062017.

[9] W. Johnston, E. Chaniotakis, E. Dart, C. Guok, J. Metzger, B. Tierney, The Evolution of Research and Education Networks and their Essential Role in Modern Science, a chapter in "Trends in High Performance & Large Scale Computing" Lucio Grandinetti and Gerhard Joubert editors. Available at http://www.es.net/pub/esnet-doc/index.html

[10] IDCP (2010). See http://www.controlplane.net/

[11] W. Johnston, E. Chaniotakis, C. Guok, "ESnet and the OSCARS VIRTUAL CIRCUIT Service: Motivation, Design, Deployment and Evolution of a Guaranteed Bandwidth Network Service Supporting Large-Scale Science." Available at http://www.es.net/pub/esnet-doc/index.html#oscars100510

[12] Network Services Interface (NSI) working group, Open Grid Forum, http://ogf.org/gf/group_info/view.php?group=nsi-wg

[13] perfSONAR (2010). See http://www.perfsonar.net/

[14] DRAGON: Dynamic Resource Allocation via GMPLS Optical Networks. See http://dragon.maxgigapop.net/twiki/bin/view/DRAGON/WebHome

Biographies

EVANGELOS CHANIOTAKIS is network engineer in the ESnet Engineering Group, and a software developer focusing on dynamic bandwidth provisioning services and systems, network operation and automation, guaranteed bandwidth services, and software design & development. He has worked on the routing team at SCinet, and has been a software engineer at University of Crete. He was educated at Panepistimio Kritis (University of Crete) and has a BSc in Mathematics.

CHIN GUOK joined ESnet in 1997 as a network engineer, focusing primarily on network statistics. He was a core engineer in the testing and production deployment of MPLS and QoS (Scavenger Service) within ESnet. He is the technical lead of the ESnet On-Demand Secure Circuits and Advanced Reservation System (OSCARS) project, which enables end users to provision guaranteed bandwidth virtual circuits within ESnet. He also serves as a co-chair of the Open Grid Forum On-Demand Infrastructure Service Provisioning Working Group.

WILLIAM E. JOHNSTON (wej@es.net) is a senior scientist and advisor to ESnet, the network that serves the US Dept. of Energy Office of Science. He led ESnet between 2003 and 2008, during which time ESnet undertook a complete reanalysis of the requirements of the DOE's science programs that ESnet supports. As a result of this a new network architecture and an implementation approach were defined that would accommodate the massive data flows of science as typified by the movement of petabytes/year from the Large Hadron Collider (LHC). This new network was built in 2007 and 2008. Previously he ran the Lawrence Berkeley National Laboratory's Distributed Systems Department and worked on many projects related to the application of computing in science environments. He also co-founded the Grid Forum (now OGF) with Ian Foster and Charlie Catlett. He has worked in the field of computing for more than 40 years and has taught computer science at the undergraduate and graduate levels. He has a Master's degree in mathematics and physics from San Francisco State University. For more information see www.dsd. lbl.gov/ ~ wej.