The Second International Conference on Parallel, Distributed, Grid and Cloud Computing for Engineering
12-15 April 2011, Ajaccio - Corsica – France
In "Trends in Parallel, Distributed, Grid and Cloud Computing for Engineering,"
Edited by: P. Iványi, B.H.V. Topping, Civil-Comp Press.

Network Services for High Performance Distributed Computing and Data Management

W. E. Johnston, C. Guok, J. Metzger, and B. Tierney
ESnet and Lawrence Berkeley National Laboratory, Berkeley California, U.S.A

Keywords: high performance distributed computing and data management, high throughput networks, network services, science use of networks

Much of modern science is dependent on high performance distributed computing and data handling. This distributed infrastructure, in turn, depends on high performance operation of high speed networks and services – especially when the science infrastructure is widely distributed geographically. This is true for small science groups in fields such as materials, nanotechnology, molecular biology and genomics, etc., that involve remote instruments and/or large amounts of data, as well as for the obvious cases of the large science collaborations in fields such as high energy physics, astronomy and cosmology, large-scale computational science, etc. In all of these cases sophisticated and highly tuned network services are needed 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." That is, the very low packet loss at very high data rates (typically 10 gigabits/second) that is necessary for high network throughput over long (national and intercontinental) distances.

This paper discusses an implementation of both of these services in a large production 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 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.

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. A major source 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, several high energy physics and nuclear physics accelerators.

The workshops' approach is to examine how the science community believes that the process of doing science has 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 is 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. See, e.g., [3].

The next section describes a science scenario that drives the networking requirements.

---

[a] A synchrotron light source is a particle accelerator that is specialized to producing high-intensity, mono-energetic, and 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.

1.1 Collaboration and Data Management in High Energy Physics

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. These experiments are typically in the form of different types of event detectors that use the colliding beams of particles generated by accelerators such as the Large Hadron Collider program (LHC [4]) at the European Center for Nuclear Research (CERN) in Switzerland. The LHC is almost certainly the largest scientific instrument ever constructed. The collaborations centered on the two largest experiments – the CMS [5] and ATLAS [6] detectors – each encompass some 2000 physicists from 150 institutions in more than 30 countries.

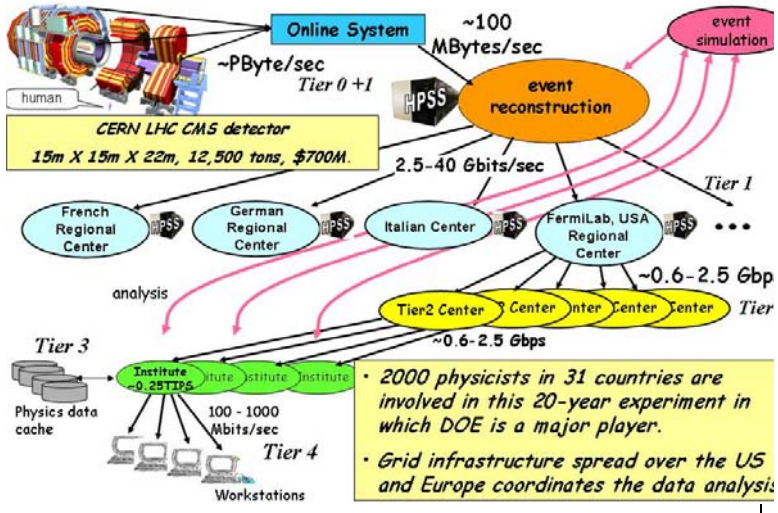

Figure 1. High Energy Physics Data Analysis

This science application epitomizes the need for collaboratories 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.)

The LHC is a synchrotron accelerator some 27km in diameter and currently capable of accelerating protons to an energy of 7 TeV (tera electron volts). The LHC accelerator took 30 years to build and cost about US$6.5B. The detectors and computing equipment cost another US$2B. The annual operating cost of the LHC is about US$1B.

Atlas and CMS each start by generating several tens of petabytes/year[a] and rapidly ramp up to hundreds of petabytes per year. (The Atlas experiment has collected more than 7 petabytes since the LHC started running in March, 2010.)

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

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[a], 39 petabytes of disk, and 50 petabytes of tape storage. The tape archive systems are at the "Tier 1" centers of Figure 1. One copy of the data is kept at CERN for archival purposes, and the Tier 1 center, in combination, hold another complete set of the data generated by the detectors. The Tier 1 centers provide the "working dataset" and supply this data to the Tier 2 centers 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 11 Atlas Tier 1 centers in Europe, the U.S., Canada, and Taiwan, and about 70 Tier 2 centers.

The HEP problems are among the most data-intensive known. Hundreds to thousands of scientist-developers around the world continually develop software to better select candidate physics signals from particle accelerator experiments such as Atlas, 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.

Equally as important for collaborations and distributed data processing on a global scale 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.

To make the relationship to the network clear, we briefly consider the distributed systems that perform the various analysis and data management operations. For specificity the software developed in the Atlas experiment collaboration is discussed. The overall functionality of the software of the other experiments is similar, but the design and implementations differ.

1.1.1 The Atlas distributed computing and data management systems

The Atlas distributed system has two primary components: The Panda job management system and the DDM data management system.

The Panda architecture and workflow is illustrated in Figure 2 and described in [7].

---

[a] 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.

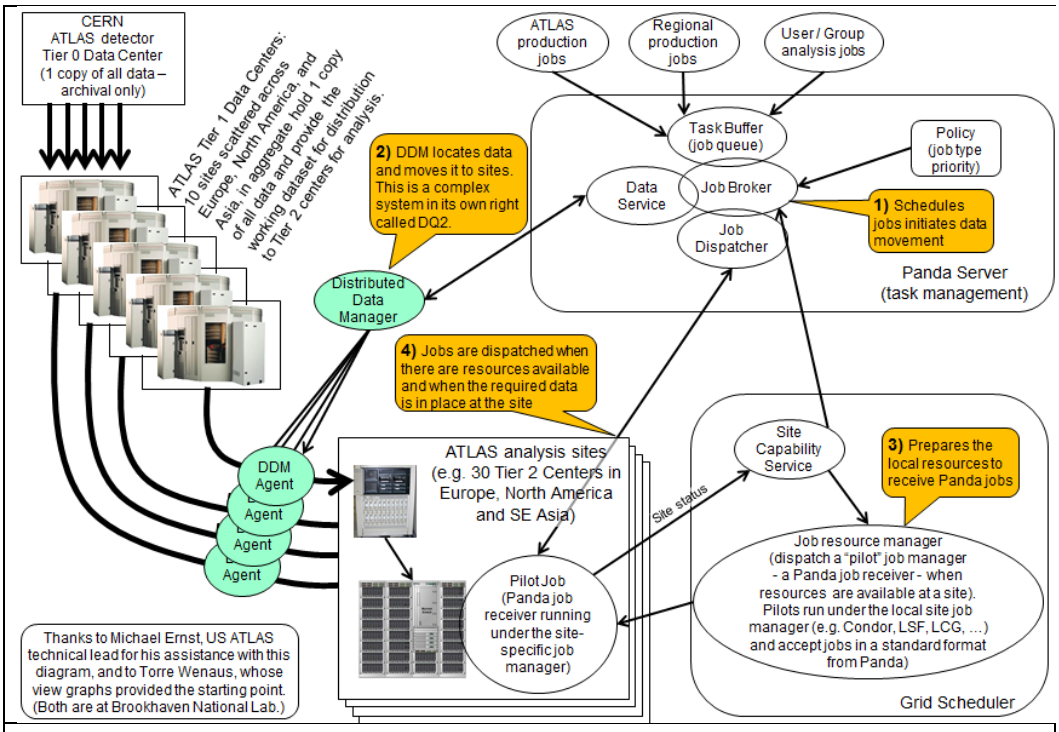

Figure 2. The Panda job management system.

The Panda sever 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. "Pilots" are local Panda job managers that accept jobs from the Panda server. Pilots, in turn, are initiated by various local job management systems such as CondorG, PBS, etc[a]. Once jobs are dispatched to sites, requests are made of DDM – the data management system – to move the required data to the site where the job has been scheduled. The Pilots execute jobs that are queued and whose data is available; do housekeeping related to various sorts of job failure; send output files to their destinations, etc. All of these components operate as a job pipeline, the 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 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 2 are the institutions that provide the computing and disk storage resources – primarily U.S. universities scattered across the country. In the U.S. part 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[a]. The amount of data moved among the 17 institutions involved in the analysis is considerable, averaging about 116 terabytes/day (Figure 3), 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.

The U.S. Atlas collaboration accounts for about 25% of the total Atlas computing and storage resources.

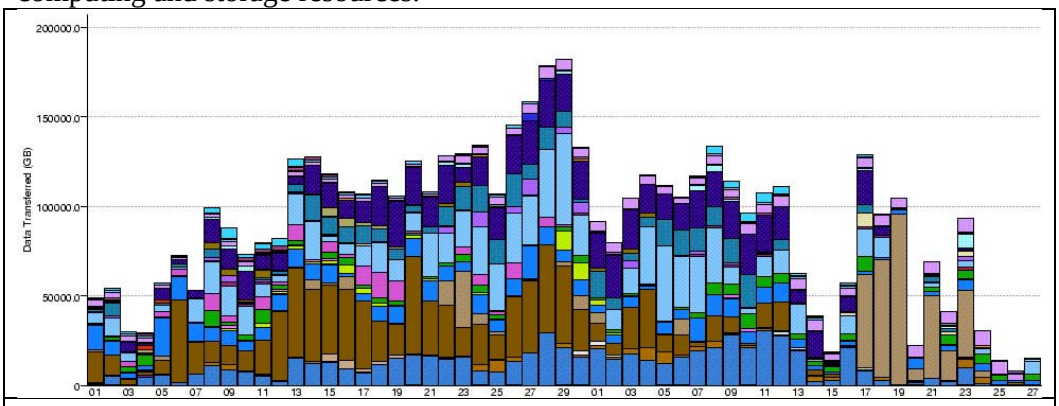

Figure 3. Data transferred (gigabytes) among the 17 U.S. Atlas sites, per day, from Nov. 1 to Dec. 27, 2010. (From the dashboard (system monitor) for the "BLN Cloud" (U.S. Atlas). (The bars are broken down by function and institution.) (http://dashb-atlas-data.cern.ch/dashboard/request.py/site)

1.1.2 Network implications

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

- are data intensive and high-performance, frequently moving terabytes a day for months at a time ;
- are high duty-cycle, operating most of the day for months at a time in order to meet the requirements for data movement;
- have components that are typically spread over continental or inter-continental distances;
- are always distributed across multiple network administrative domains, and;

[a] The LHC accelerator operates about 9 moonths/year. It shuts down during the winter to conserve power and for maintenance. However, even during shut-down computing continues almost unabated. (CERN's power consumption is 120-180 MWatts, which drops to about 35 MW when the accelerator is not running. CERN accounts for about 10% of the total energy consumption of the greater Geneva City area. (http://lhc-machine-outreach.web.cern.ch/lhc-machine-outreach/faq/lhc-energy-consumption.htm)

- are depend on network performance and availability to ensure the efficient functioning of the distributed workflow systems that coordinate the data management and analysis tasks.

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

- get guarantees from the network in order to ensure that there is adequate bandwidth to accomplish the task at the requested time;
- interoperate with compatible services in other domains in order to provide an end-to-end service;
- get 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;
- do 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 the distributed analysis applications systems.

2 ESnet: A Brief Overview

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

ESnet is a single network administrative domain. That is, the equipment and telecommunication circuit infrastructure that is managed, operated, provisioned, secured, etc., by ESnet staff defines the ESnet domain. This is true for essentially all networks[a]. Different domains connect with each other through carefully managed "peerings," which provide both the logical and physical network-network interface. The peering is typically done at exchange points where many networks come together. In the U.S., for example, there are major R&E exchange points (called GigaPoPs) at dedicated facilities in New York, Chicago, Seattle, Sunnyvale (San Francisco), Los Angles, Atlanta, and Maryland (Washington DC). These exchange points are operated by universities, or related organizations. A similar situation exists for the commercial Internet. There are many exchange points worldwide where commercial network operators come together to peer. (Many of these in the U.S. and Europe are operated by a company called Equinix.)

To address the above science needs, ESnet4 – the fourth generation of the technology of ESnet – was designed as a hybrid packet-circuit network consisting of two core networks: (1) an IP (Internet Protocol) core that carries all the commodity

[a] We shall use the term "network" to mean an autonomous / administratively homogeneous network domain. In other words, the domain operated by a network provider like ESnet, GÉANT, AT&T, Level3, etc.

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]. Both the ESnet cores peer with almost all other R&E networks in order to provide global connectivity for science.

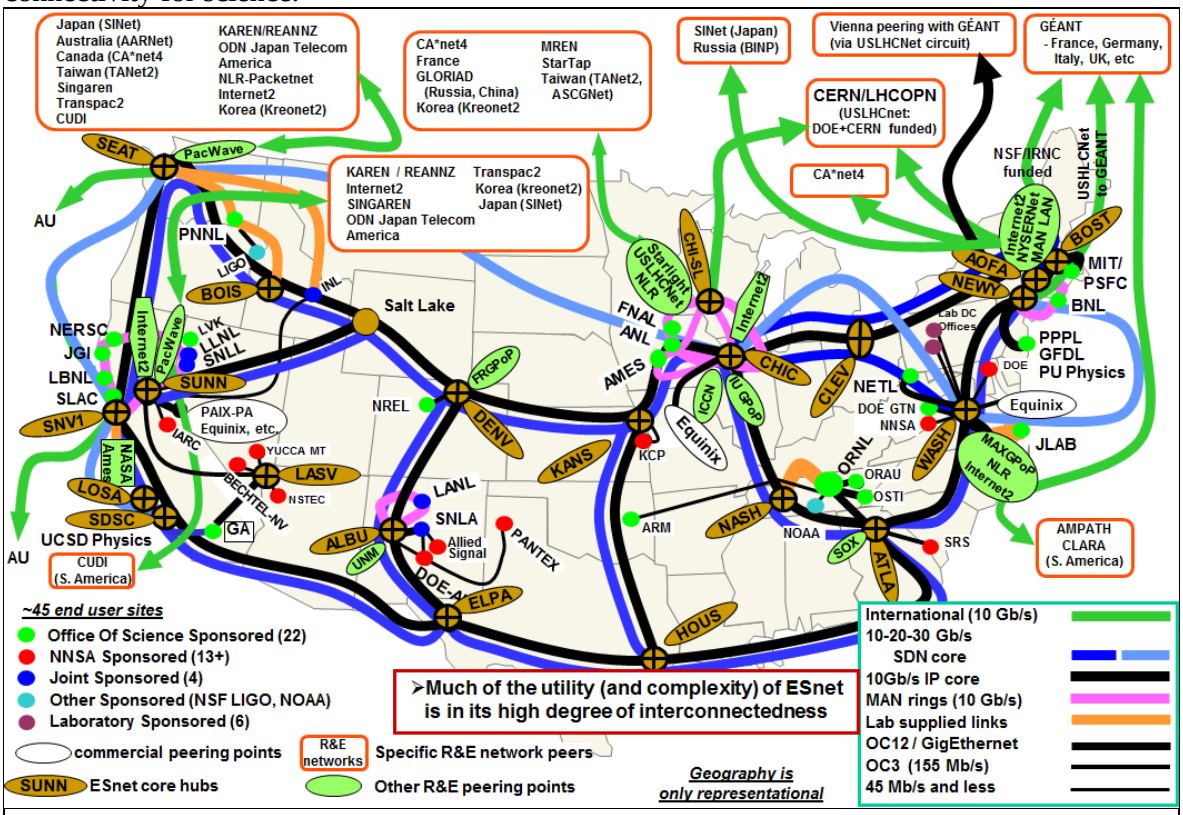

Figure 4. ESnet4

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 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[a] 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. Additionally, all of the DOE national laboratories 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 (Figure 5).

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

[a] We use the term "link" to refer to a physical or otherwise fixed path between network routers and switches. In ESnet these are primarily 10 Gbps optical channels multiplexed on long distance optical fiber between cities.

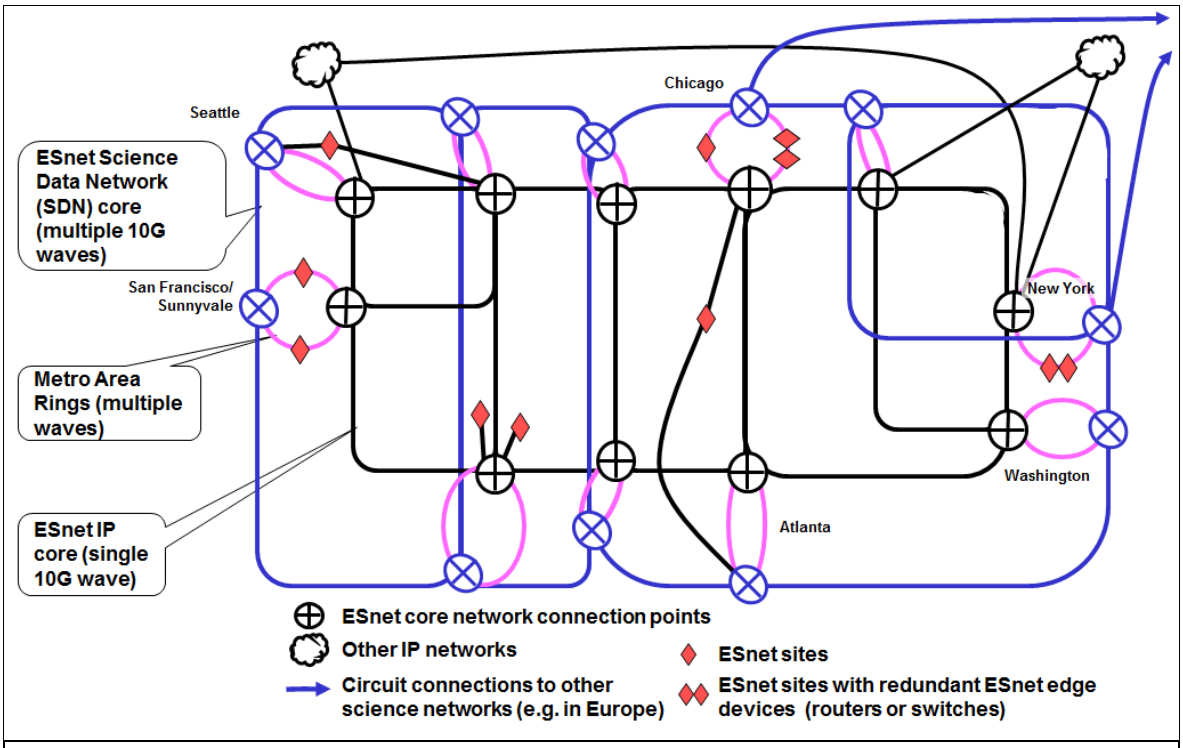

Figure 5. ESnet4 architecture

3 Guaranteed bandwidth network service

Our understanding of the requirements of modern science for new network services has emerged from 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 [2]. 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]:

- 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 inter-node communication.
- Connectivity: The network must have the geographic reach — either directly or through peering arrangements with other networks — sufficient to connect users and collaborators and analysis systems to experiment sites.

- Services: Guaranteed bandwidth, traffic isolation, end-to-end monitoring, etc., are required as network services and these must be presented 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.

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

Based on the analysis of a number of science application areas [9], the virtual circuit service has as its goals to:

- Allow users to request a virtual circuit providing guaranteed bandwidth between specific end points for a specific period of time
  - User request is via a Web Services for applications or a Web browser interface for human users
- Provide traffic isolation so that non-standard transport protocols may be used for maximizing throughput over very long (international) distances
- Provide the network operators with a flexible mechanism for traffic engineering (traffic management) in the core network
  - E.g. controlling how the large science data flows use the available network capacity
- Support the inherently multi-domain environment of large-scale science
  - OSCARS must interoperate with similar services in other network domains in order to set up cross-domain, end-to-end virtual circuits

The OSCARS service is a "virtual circuit" service. It is called a "virtual circuit" because the user sees a dedicated path that has circuit-like properties: It is not shared, it provides traffic isolation, it has fixed bandwidth, etc. This type service is sometimes also called a pseudowire service.

OSCARS uses MPLS[a] to transport the circuit payload and is essentially a management and control system ("plane") for MPLS. However, OSCARS supports capabilities that are beyond existing MPLS control plane mechanisms. In particular, OSCARS manages routing constraints that arise from the temporal nature of the circuit reservation capability (with a specified bandwidth, start and end time) that are outside of the scope of the standard MPLS configuration tools.

[a] MPLS – multi-protocol label switching – is a highly scalable, protocol agnostic, data-carrying mechanism that is supported in many modern network routers. MPLS can transport IP packets and Ethernet frames, so from the user's point of view the virtual circuit is either an IP circuit or an Ethernet VLAN (virtual local area network).

3.2 OSCARS implementation

3.2.1 Architecture

The OSCARS architecture is illustrated in Figure 6. Most of the functions illustrated there are described in the following text.

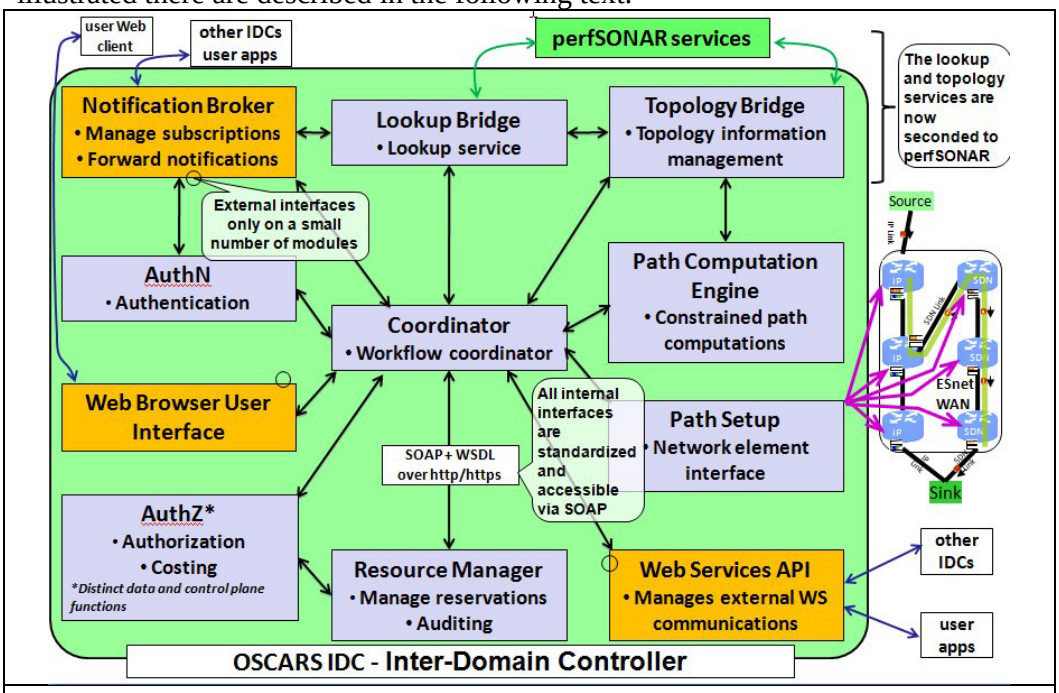

Figure 6. OSCARS architecture

3.2.2 Design and constraints

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.

Design constraints:

There are a number of constraints that are imposed by the user requirements and from operational considerations in the network.

The service must provide user access at both layers 2[a] (Ethernet VLAN (Ethernet virtual local area network)) and 3 (IP). The motivation for this constraint was two-fold. First there are sites who will want to use the service that do not have a layer 2 connection to ESnet. Second, for individual end-users it is likely to be

[a] By convention, the functions of an Internet network are partitioned as layers 0/1 (“physical”) (“permenant” network connections (originally wires, today mostly optical channels multiplexed onto an optical fiber), layer 2 (“link”) (the switching infrastructure on top of channels, layer 3 (“Internet”) (the IP packet protocols), layer 4 (“transport”) (TCP), layer 5 (“application”) (e.g. HTTP).

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.

ESnet uses only layer 2, 2.5, and 3 devices (Ethernet switches, MPLS switchs, and IP routers) and therefore the implementation must not require 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 new hardware into the network.

For inter-domain (across multiple networks) circuit setup no RSVP[a]-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.)

In terms of cross-domain circuits, whether to actually set up a requested cross-domain circuit must be at the discretion of the local controller (e.g. OSCARS) in accordance with local policy and available resources.

Available network tools:

The implementation strategy for OSCARS makes use of available network management tools and adds capabilities as necessary to meet the goals.

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

- 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 accommodate “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).
    - Once OSCARS calculates a path then RSVP is used to signal and provision the path on a strict hop-by-hop basis

[a] Resource Reservation Protocol

To these existing tools are added:

- Service guarantee mechanisms using
  - Elevated priority queuing for the virtual circuit traffic to ensure unimpeded throughput
  - Link bandwidth usage management to prevent oversubscription by circuits
- 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)

3.2.3 Routing and utilization management

The general approach to routing a user requested virtual circuit and management of network resources is based on a 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 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 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 does 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.

3.2.4 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 LSP (Label Switched Path) 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]

[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

and the routers that interconnect the links of the path. This process defines, in MPLS parlance, a Label Switched Path (Figure 7). 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 because the network only sees the MPLS packets – it does not see what is being transported. 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).

The bandwidth guarantees are provided (1) by assigning the virtual circuit traffic an elevated queuing priority (MPLS and IP packets go through the same queues in the router so elevating MPLS queuing priority puts it at a higher priority than any other traffic on this link), (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 circuits do not interfere with each other.

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.

As noted, the virtual circuits are rate-limited at the ingress, however in a generalization of the original ideas about the semantics of circuits, the current implementation allows sources using reserved bandwidth to burst above the allocated bandwidth if idle capacity is available. This is accomplished without interfering with other circuits by marking the over-allocation bandwidth as low priority traffic.

---

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.

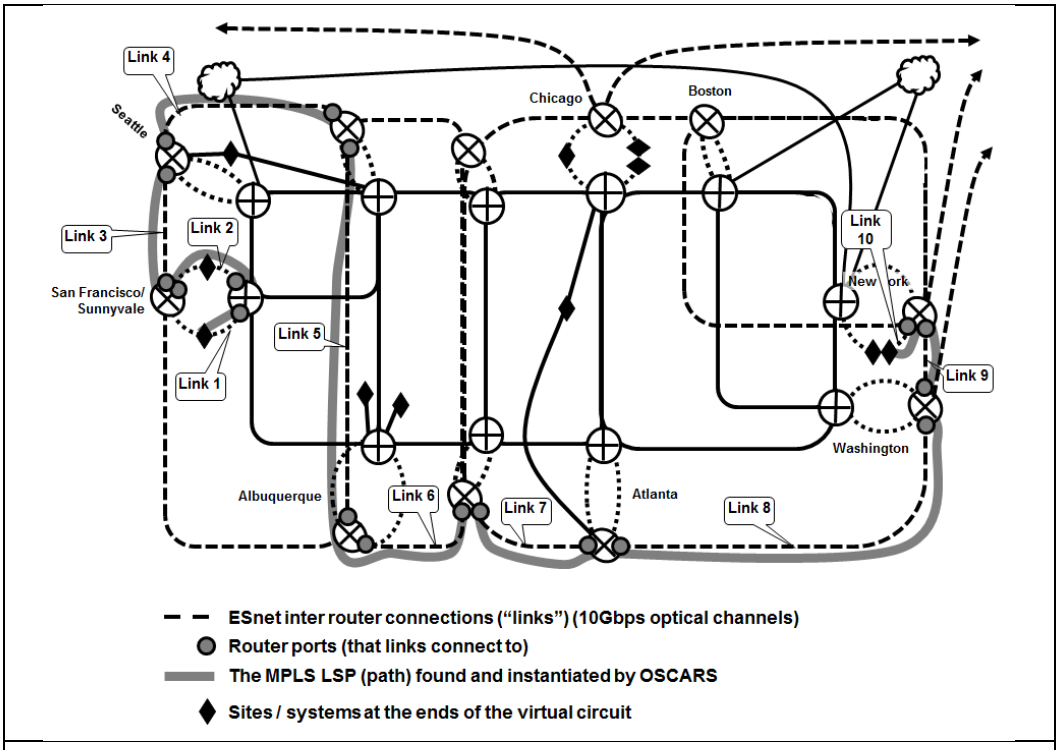

Figure 7. The relationship between links and ports (physical entities) and an OSCARS paths (virtual). Note that the particular routing (chosen links) implies that there is not enough bandwidth on the geographically shorter combination of links in the SF Bay metro ring and on the Sunnyvale – Albuquerque inter-city link.

The Path Computation module deals with the topology database and the routing that is described below.

The Path Setup module is effectively a device driver. Its implementation in ESnet manages an MPLS infrastructure as described. Other networks use OSCARS and they replace the Path Setup module with an implementation that is specific to their network devices and management.

The functions of the remaining modules are primarily adjucts to the basic OSCARS function and are noted briefly in Figure 6.

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 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). While OSCARS is 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].

Reliable circuits:

In order to provide high reliability, the user can request a second circuit that is diversely routed from the first circuit. These circuits pairs are typically used as virtual private networks (VPNs) that interconnect private IP routing "clouds." With IP routing managing both circuits, if on fails then the IP packets are just 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 Network monitoring

As noted, many modern science applications involve instruments, people, and institutions from all over the world. Connecting the required locations entails the cooperation of many different networks – campus, regional/national and international networks – that operate as independent administrative domains, but must cooperate closely in order to support the high throughput communication of distributed scientific computing and data management. Network failures are easily detected and repaired. Degraded network elements are much more difficult to identify. Degradation of network elements generally manifests itself as increased, or even irregular, bit errors. Bit errors manifest themselves as packet loss as the network protocols detect and deal with bit errors. Many transport protocols like TCP assure data integrity and completeness in ways that effectively amplify the effects of bit errors in terms of degrading end-to-end throughput, especially when operating over long distances. Therefore locating and eliminating sources of bit errors is crucial for high throughput.

Therefore, the second critical service for supporting high-performance wide-area distributed systems is end-to-end monitoring that can test and evaluate the decomposed network elements from application interface to application interface across all of the intervening network domains.

4.1 perfSONAR

PerfSONAR[a] is intended as a significant first step in cross-domain monitoring and testing by both network operators and users. perSONAR has been widely deployed[b] in the international R&E networking community and the networks that support the LHC data management and analysis. perfSONAR has succeeded because 1) it is widely deployed in the R&E community and provides an invaluable

[a] For more information on perfSONAR see www.perfsonar.net.

[b] There are currently (Jan. 2011) 353 instances of perfSONAR testers and measurement archives deployed by 88 institutions in N. America and SE Asia, based on the list complied at http://ndb1.internet2.edu/perfAdmin/directory.cgi Additionally, there are a substantial number in Europe.

cross-domain monitoring and testing functionality, and 2) it is widely deployed because the design explicitly allows for federated deployment: That is, the domain operators deploy perfSONAR in ways that are consistent with the policies in their domain (e.g. what is tested, what data is distributed and to whom, etc.).

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

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

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

4.1.1 Architecture

A service oriented architecture (SOA) has been adopted by the community that consists of three layers and a collection of defined service functions. (See Figure 8.)

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

The currently extant services fall into seven categories:

- Measurement Point (MP) service: Creates and/or publishes monitoring information related to active and passive measurements
- Measurement Archive (MA) service: Stores and publishes monitoring information
- Lookup service (LS): Registers all participating services and their capabilities
- Topology service (TS): provides network topology information
- Authentication service (AS): Manages domain-level access to services
- Transformation service (TrS): performs manipulation (aggregation, statistics) on available data sets

- Resource Protector (RP) service: arbitrates the use of limited measurement resources based on the policy of the resource owner

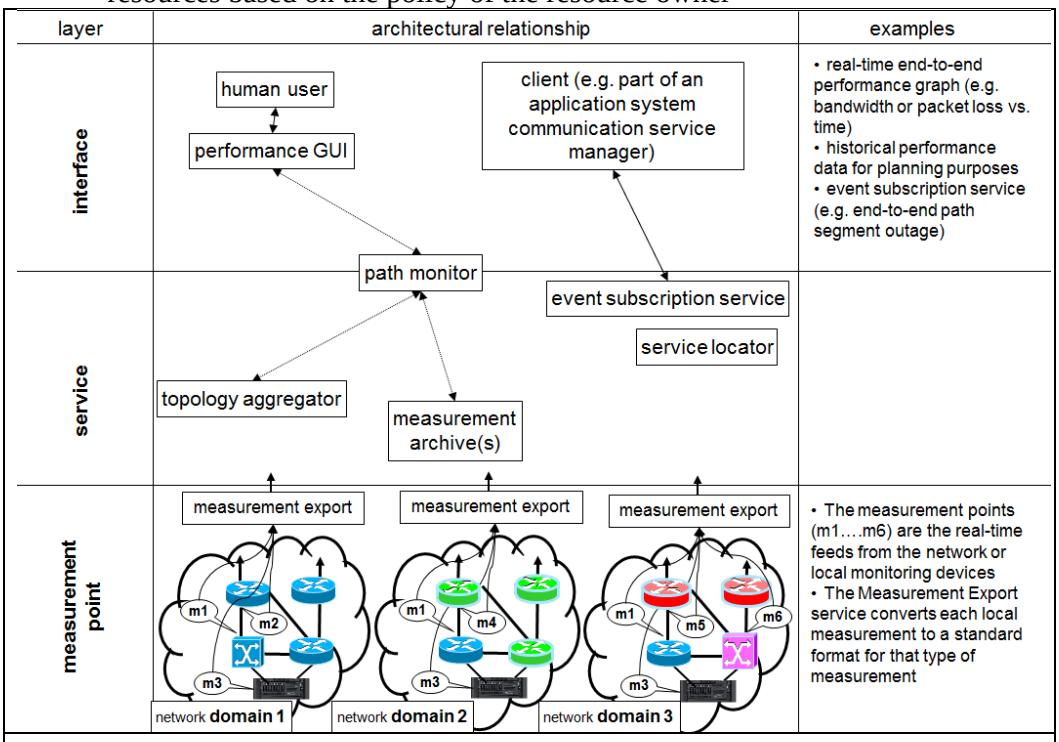

Figure 8. PerfSONAR Architecture

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

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

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

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

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

4.1.2 Multi-Domain Monitoring

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

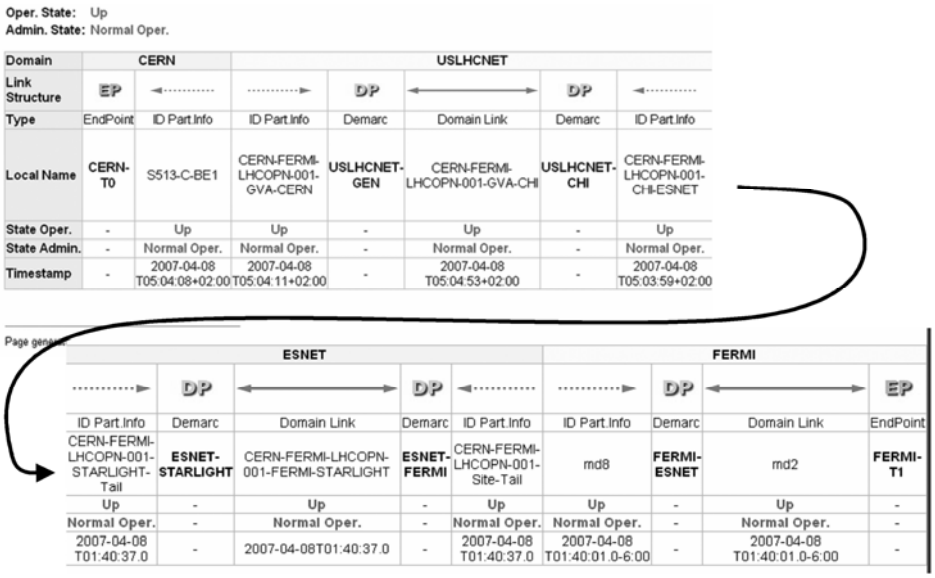

Figure 9. E2Emon generated view of the data for one OPN link. Note that the display is split and displayed in two parts for this figure (see, e.g., http://cnmdev.lrz-muenchen.de/e2e/lhc/mon/G2_E2E_index_PROD.html and select CERN-FERMI-LHCOPN-001).

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

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

The MP in each domain consists of two components. The domain specific component in the various networks typically interfaces with the operational network monitoring system to obtain the link status data for the portion of the end-to-end path within that particular network. Virtually every network does internal monitoring in a different way that has evolved historically along with the network. The perfSONAR component of each MP takes the resulting data generates a standard XML file, and publishes it via the MP service interface, or pushes it to an MA for archiving and publishing. This information is used by an application called E2Emon[a].

E2Emon uses perfSONAR protocols to retrieve current circuit status every minute or so from MAs and MPs in all domains supporting the circuits. See Figure 9.

5 Deployment of the services and summary

Both of the software systems described here – one providing guaranteed bandwidth and one doing network monitoring – are being fairly widely deployed in the campus, regional/national and international networks that support distributed science and the R&E communities, including ESnet. The services provided are a key element in the infrastructure that enables large-scale collaborative science projects.

OSCARS is a production service in ESnet and 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.

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). This capability is provided.

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

The perfSONAR network monitoring system is the first effective multi-domain network test and monitor capability. It is widely deployed in the R&E community and is routinely used to debug network paths (IP and circuit) that are used by the international science community. perfSONAR is also developed in an informal international consortium.

Together, these services have provided critical infrastructure for the functioning of large science experiments such as those at CERN's LHC.

6 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. This document is report number LBNLXXXX.

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

7 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/