LAMBDASTATION: A FORWARDING AND ADMISSION CONTROL SERVICE TO INTERFACE PRODUCTION NETWORK FACILITIES WITH ADVANCED RESEARCH NETWORK PATHS

PHIL DEMAR, DON PETRAVICK FNAL,
BATAVIA, IL 60510, USA

Abstract:

Over the past several years, there has been a great deal of research effort and funding put into the deployment of optical-based, advanced technology wide-area networks. Fermilab and CalTech have initiated a project to enable our production network facilities to exploit these advanced research network facilities. Our objective is to forward designated data transfers across these advanced wide area networks on a per-flow basis, making use our capacious production-use storage systems connected to the local campus network. To accomplish this, we intend to develop a dynamically provisioned forwarding service that would provide alternate path forwarding onto available wide area advanced research networks. The service would dynamically reconfigure forwarding of specific flows within our local production-use network facilities, as well as provide an interface to enable applications to utilize the service. We call this service LambdaStation. If one envisions wide area optical network paths as high bandwidth data railways, then LambdaStation would functionally be the railroad terminal that regulates which flows at the local site get directed onto the high bandwidth data railways. LambdaStation is a DOE-funded SciDac research project in its very early stage of development.

Problem Area:

Advanced research networks, such as National Lambda Rail, CAnet4, Netherlight, UKLight, and many others, have been deployed in recent years to develop and exploit emerging optical network technologies. These advanced networks are typically not intended to replace production use wide-area network (WAN) facilities. Rather, they are intended to support and nurture research applications whose requirements exceed the capabilities of today's production WANs. Such advanced network infrastructures have the capacity and capability to meet the extremely large data movement requirements of the Particle Physics Collaborations. However, to date, the primary focus of research efforts in the advanced network area has been to provision, dynamically configure & control, and monitor the wide area optical network infrastructure itself. Application use of these facilities has been largely limited to demonstrations using test stands or small prototype high performance computing systems. The issue of integrating existing production-use local computing facilities with these advanced, high bandwidth research networks has remained largely unaddressed. As a result, customized local network access is typically required. A commonly used approach (figure 1) is to multi-home local systems to provide a host-routable backdoor path to the advanced research network.

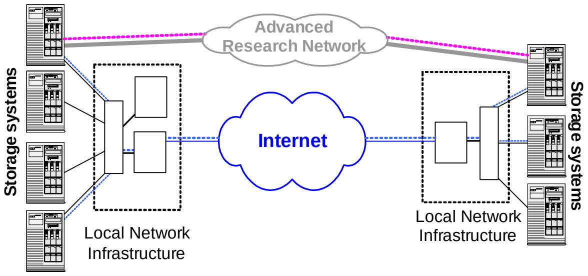

Figure 1: Typical Advanced Network Connectivity

This type of approach has many problems. It adds complexity for the end system(s). It scales poorly in terms of number of end systems, and doesn't scale at all if multiple advanced network paths become available. Other approaches to customized local access create similar scaling and complexity problems. Fundamentally, the issue of connecting production LANs to advanced research networks is a last mile problem. The LambdaStation project is intended to deal with this last mile problem, and enable local network facilities to utilize these advanced network paths via the local network infrastructure.

LambdaStation:

LambdaStation is an alternate path selection service. It is intended to act as an agent that facilitates requests from local applications for high bandwidth WAN paths that aren't available over the normal production WAN network path. Having received such a request, LambdaStation would negotiate the availability of the alternate WAN path, and if available, coordinate its setup & teardown. Since these advanced research networks are typically restricted-use resources, LambdaStation could schedule their use, if necessary. Finally, LambdaStation would coordinate dynamic reconfiguration of the local network infrastructure to selectively forward the application's traffic over the alternate WAN path. Figure 2 depicts the interaction of LambdaStation with the local system(s), local network infrastructure, and advanced WAN path.

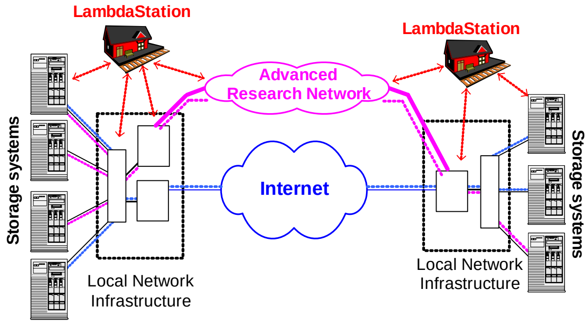
LambdaStation
LambdaStation
Advanced Research Network
Internet
Storage systems
Storage systems
Local Network Infrastructure
Local Network Infrastructure

Figure 2: LambdaStation

Unlike figure 1, traffic bound for the alternate network path would utilize the same interface as normal production traffic. Only within the local network infrastructure would the designated traffic be forwarded differently. The project goal of LambdaStation is to be able to selectively forward traffic on a per-flow basis. While routine traffic between two systems would follow the normal production WAN network path, designated traffic flows belonging to a specific application could simultaneously follow the alternate WAN path, with its presumably higher bandwidth.

Components of LambdaStation:

LambdaStation is software. From a design perspective, the LambdaStation software would consist of five modules (figure 3):

- Local Demand Module
- Local Steering Service
- WAN module
- Peer LambdaStation Module
- Authentication/Authorization module

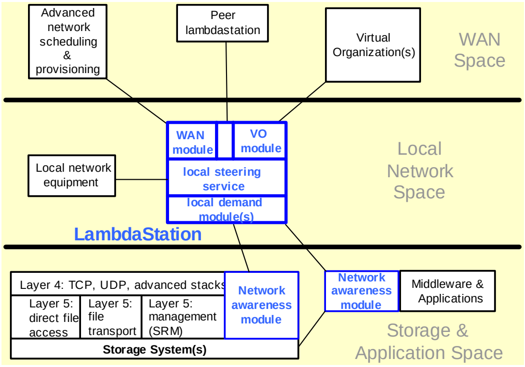

Figure 3: Components of LambdaStation

Local demand module: Interface to host systems or resource manager. Accepts alternate WAN path requests, including transfer characteristics (total bytes, peak rate, expected duration, scheduling parameters, etc.). Coordinates flow identification keys (source/destination addresses and ports, DSCP code points, IP version #, etc.). Provides state information on alternate path to requesting end system (request status, setup status, expected bandwidth, path termination, etc.)

Local Steering Service: Customizable interface to local network infrastructure. Dynamically modifies local forwarding for designated flows, in a graceful, non-disruptive manner. Makes appropriate ACL modifications at site egress point

WAN module: Customizable interface to alternate path networks which are available to the site. Checks alternate path availability; negotiates alternate WAN path setup / teardown, and scheduling, if necessary.

Peer LambdaStation module: Coordinates with peer LambdaStation at remote site to establish path symmetry. Exchanges flow identification selectors. Acknowledges WAN setup/teardown, and LAN forwarding reconfiguration. Verifies alternate WAN path continuity. Provides notification of WAN path or traffic termination.

Authentication/Authorization module: Provides authentication & authorization (AA) for alternate path requests. Initially expected to be internal LambdaStation AA. The longer term goal is to utilize Virtual Organization AA infrastructure.

Technical Approach:

The development model for LambdaStation has four elements that form the technical foundation of the project:

- Flow identification
- Local path forwarding
- WAN interface model
- Graceful switchover & failback

Flow identification: Will be based on either source/destination IP address & host-assigned DSCP code point, or source/destination IP address & source/destination port number. Initial development will be focused on the DSCP code point approach, due to the difficulty in dynamically reconfiguring local forwarding to continually changing source port numbers. The feasibility of utilizing IPv6 and its flow-id field will be investigated

Local path forwarding: Will be based on existing policy routing mechanisms available in commonly deployed switch & router products. The initial test bed will be based on Cisco switches, and dynamic modification of policy routing ACLs will be used to alter forwarding of designated traffic flows. The local steering module in LambdaStation will have a site configurable component that can be adapted to forwarding control capabilities of the local network infrastructure.

WAN interface model: Assumes either a layer-2 site-to-site connection, based on VLAN assignment, or a label-switched path based on an MPLS-type of tagging. In either scenario, a site ingress/egress router with connections to the alternate path networks is assumed. Layer-3 peering is established between ingress/egress routers at collaborating sites. The WAN module in LambdaStation will also have a site configurable component that can be adapted to each site's specific alternate path WANs.

Graceful switchover & failback: Establishment & teardown of alternate WAN paths must be done carefully to avoid disruption to existing data flows. A graceful setup is required in which the WAN component of the path gets created first and tested for continuity, then the local path forwarding is modified to use the new WAN path. The teardown process proceeds in the reverse order. Transparent failover back to the production network WAN path will be necessary to deal with any premature loss of connectivity across the alternate path WAN component. Bandwidth limitations or QBSS may be desirable to limit the impact of advanced network high volume data traffic being rerouted over the production WAN path(s).

Case Study:

Consider a demonstration of updating files for CMS on an interactive analysis cluster at the CalTech Tier-2 center from a master copy at the Fermilab Tier-1 center. This might typically consist of 3.5 TB of data, spread over a storage system of 50 file servers on each side. The data is normally packed in one gigabyte files, so 3500 files would be transferred. Properly aggregated, the file transfer should be completed in slightly less than an hour across a 10Gb/s WAN path.

The SRM (Storage Resource manager) residing in the storage system at Fermilab is presented with the list of the 3500 files to transfer. Using knowledge of the available bandwidth between the sites, the type of network, the bandwidth of the disks, and the loading of the storage system from other uses, file transfers begin. Initially this will use the existing production WAN path.

SRM sends a request to the local LambdaStation for a high bandwidth, alternate path to CalTech. LambdaStation and SRM coordinate on the flow identification parameters (source/destination IP address and DSCP code point) needed to forward that particular data movement over an alternate path. The coordination has other attributes, such as the virtual organization on whose behalf the file transfer is taking place.

The local LambdaStation negotiates use of a 10 Gb/s path across the DOE UltraScience to CalTech for the next hour. Concurrently, the local LambdaStation coordinates with a peer LambdaStation at CalTech to establish path symmetry. As soon as the alternate WAN path is established between the two sites, and path verification is successfully completed, the two LambdaStations configure their respective local switches and routers to forward the designated flows onto the dynamic link. Each LambdaStation also configures its ingress/egress router to admit the appropriate inbound flows, and reject inappropriate ones. The Fermilab LambdaStation acknowledges creation of the alternate path to SRM, which adjusts its file transfer parallelism to make optimal use of the higher bandwidth path. SRM might even chose to change to a TCP transport algorithm optimized for high bandwidth, low loss paths.

At the end of the hour, very many, but not quite all, files have transferred. The dynamic link is scheduled to be torn down. The local LambdaStations reconfigure local forwarding back to the production network paths. SRM becomes aware of the reduced bandwidth, readjust the file transfer parallelism, and if necessary, reverts to standard TCP transport.

TestBed:

LambdaStation is a US Department of Energy-funded SciDac research project. The principal project sites are Fermilab, the US-CMS Tier-1 center, and CalTech, a US-CMS Tier-2 center. LambdaStation software development & testing will be conducted at the two sites. The project goal is to move CMS experiment data between the two sites, using production CMS facilities at both locations, and having LambdaStation selectively route that data across an alternate high bandwidth path. The Department of Energy's UltraScience Net, another SciDac research project, will serve as the primary high bandwidth research wide area network for alternate path forwarding. UltraScience Net will have up to two 10 Gb/s lambdas available for use between StarLight (Chicago) and Sunnyvale (San Jose, Cal.) on a scheduled basis. Fermilab and CalTech have arranged access to StarLight and Sunnyvale, respectively. Figure 4 depicts the LambdaStation test environment.

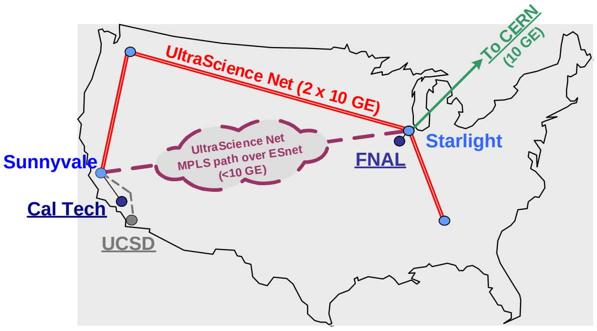
UltraScience Net (2 x 10 GE)
To CERN (10 GE)
Starlight
FNAL
Sunnyvale
UltraScience Net MPLS path over ESnet (<10 GE)
Cal Tech
UCSD

Figure 4: LambdaStation Project Testbed

A second alternate wide area network may be provided by the Energy Sciences Network (ESnet). A sub-10 Gb/s MPLS tunnel within the ESnet production network backbone would enable testing of LambdaStation with label-switched wide area network technologies, such as MPLS

Two other sites are expected to be involved in the testing. High volume CMS data movement between CERN, the CMS Tier-0 center, and Fermilab will also be tested, using the 10 Gb/s CERN-to-StarLight trans-Atlantic link. The University of California, San Diego, another US-CMS Tier-2 site is also expected to participate.

Project Schedule:

The project outline specifies a three-year development plan:

Year 1: Move data between Fermilab & CalTech production use facilities, using alternate WAN path forwarding. LambdaStation is expected to have a dynamic LAN reconfiguration capability, and SRM is expected to have allocated bandwidth awareness.

Year 2: Provide significantly higher WAN performance between Fermilab & Cal via the alternate WAN path. LambdaStation LAN reconfiguration would be fully automated, peer LambdaStation coordination implemented, UltraScience Net WAN path setup/teardown & scheduling automated, and storage systems adapted to schedule transfers to path availability.

Year 3: Harden LambdaStation to production use quality. Full integration with UltraScience Net. Harden application integration to production use quality. Develop Virtual Organization authorization & authentication sensitivity.

Summary:

LambdaStation is a research project intended to enable production-use systems and storage facilities to make use of advanced research networks or other alternate network paths. It is based on the concept of dynamic reconfiguration of production local area networks for select forwarding of specific data flows. If the research project proves to be successful, it could enable data-intensive science research programs, such as high-energy physics, to make effective use of emerging optical network technologies and the very high bandwidth capacity those technologies are beginning to offer.