LAMBDA STATION: A NETWORK PATH FORWARDING SERVICE TO INTERFACE PRODUCTION NETWORK FACILITIES TO ADVANCED RESEARCH NETWORK PATHS

PIs: Don Petravick; Fermilab
Harvey Newman; Caltech

Organizations: Fermilab
Caltech

Project Description:

The Lambda Station service enables our very large, production-use mass storage systems, which are running full-scale SciDAC applications, to fully exploit advanced circuits-based networks. The project work has involved development and deployment of innovative techniques in local networks that enable systems to send traffic, on a per-flow basis, across advanced circuit-based networks such as USNet. These new techniques allow us to selectively forward designated data flows between capacious storage systems, across local networks, to target systems over a dynamically provisioned path. Concurrently, other traffic flows from the same storage systems are forwarded across the same local network infrastructure onto conventional, routed wide area network paths.

Project Accomplishments To Date:

Building a WAN testbed infrastructure:

A reliable 10 Gbps connection between the Caltech campus and the USNet PoP at Sunnyvale has been in operation since March 2005. At Sunnyvale, our equipment is connected to USNet's network via two 10 Gbps connections. On the Chicago side, FNAL is connected to USNet via a 10 Gbps connection. A BGP peering has been established across the USNet infrastructure between our Cisco 6509 at Sunnyvale and FNAL's 6503 at Chicago.

Using policy based routing (PBR) technology, high impact traffic is switched from the default production path onto the high speed path. As shown in ***Figure 1***, the default production path crosses the Caltech campus network, CENIC backbone, ESNet backbone and the FNAL campus network. The bandwidth is limited to 622 Mbps by the ESNet-FNAL link. The high speed 10 Gbps path consists of the Caltech-SNV 10 Gbps wave, the SNV-CHI USNet 10 Gbps connection and the FNAL 10 Gbps local loop to StarLight.

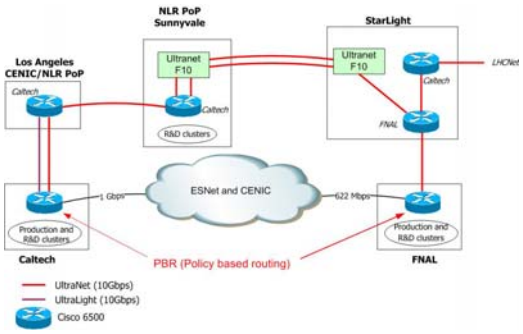

Figure 1 Lambda Station Testbed Infrastructure

The Lambda Station service is coordinated with USNet for network path availability, scheduling, and setup. Software being developed provides applications with the necessary information to utilize the high bandwidth USNet path.

Since our focus is on the experiments and facilities able to exploit the DOE.Science USNet research network, the project is centered on Fermilab facilities, and the joint CDF/CMS Tier 2 center facilities at Caltech and UCSD. We will work to include DOE funded Edge Computing for US CMS at CERN.

MonALISA Modules to Control Optical Switches

As part of the development of end-to-end circuit-oriented network management services, we have developed dedicated MonALISA modules and agents to monitor, administer and control Optical Switches in real time; specifically the purely photonic switches from Calient and Glimmerglass. The modules use TL1 commands to monitor the connectivity matrix of each switch, as well as the optical power on each port. Any change in the state of any link is reported to dedicated software agents. If a switch is connected to the network, or if it ceases to operate, or if a port's light level changes, these state changes are detected immediately and are reflected in the topology presented by the MonALISA Graphical User Interface (GUI). By using the GUI, an authorized administrator can manually construct any desired light path, and monitor the optical power on each new link as it is created.

The distributed set of MonALISA agents can be used to control large systems of optical switches and to create an optical path for end-user applications on demand. The agents use MonALISA's discovery mechanism to find each other, and then communicate among themselves autonomously by using Proxy services. Each Proxy service can handle more than 1,000 messages per second, and several such services are typically used in parallel. This ensures that the communications among the agents is highly reliable, even at very high message-passing rates.

The set of agents also is used to create a global path or tree, as the state and performance of each local area and wide area network link is known as are the state of the cross connections in each switch. The routing algorithm provides global optimization by considering the "cost" of each link or cross-connect. The optimization algorithm is capable of adaptation to handle various priority policies and reservation schemes. The time to determine and construct an optical path (or a multicast tree) end-to-end is typically less than one second, independent of both the number of links in the path and its total length.

End Systems performance evaluation

There is a large gap in the current state of the available technology, when comparing memory-to-memory and disk-to-disk transfers. This is mainly due to limited resources in the end-hosts (CPU power, bus bandwidth, and I/O memory bandwidth on the motherboard). These resources are necessarily shared by both transmission and read/write tasks. Thus we observe that the performance achievable from host memory to host disk transfers, and that from host memory to host memory across the network, when combined, do not result in the hoped for performance transferring files from disk to disk across the network.

In fact, the best disk-to-disk performance we have measured so far is for a single TCP stream between hosts at CERN and Caltech; we were able to transfer large files from disk to disk over this 11,000 km path at 300 MBytes/s[1]. Transferring from disk to memory ran at 700 Mbytes/s. On the CERN side, we used an HP 4-way 1.5 GHz Itanium2 systems running the Linux 2.6.6 kernel, equipped with 3ware controllers. On the Caltech side, we used a 2.4 GHz Opteron system with the same kernel, and Supermicro controllers. We continue to work on ways to improve these numbers and move towards tests in a production setting as part of Lambda Station in the near future. For example, we recently obtained a pair of PCI-X v2 network adaptors for testing early this fall. At the same time, we are evaluating the latest versions of transfer applications such as bbcp and Gridftp.

[1] Using the Microsoft Window 2003 server operating system, we could transfer 1 TByte of data at 536 Mbytes/s between CERN and Caltech.

Project Impact:

The project covers three areas that must be successfully addressed before optical-based advanced network technologies will fulfil their promise of significant benefits to the Particle Physics community, and the broader scientific research community in general:

Dealing with the local network last-mile problem. Lambda Station proposes a network architecture to adapt existing, production-use local network facilities to support access to advanced research networks

Bringing real-world, production-use facilities and applications into the advanced research network environment. The project promotes the necessary synergy between large scale computing facilities and the capabilities of advanced network technologies.

Developing flow-based alternative network path selection capabilities. A preliminary step has been made towards enabling special or alternate forwarding of specific flows for performance or policy reasons.

We intend to follow closely the operational aspects of USNet, particularly the use of CIENA Core Directors supporting fine-grained circuit bandwidth allocation through VCAT/LCAS and the authentication/authorization services, as we may wish to use similar services in the LHCNet production network in the future.

Additional information available at:

http://www.lambdastation.org