TERAPATHS: A QOS-ENABLED COLLABORATIVE DATA SHARING INFRASTRUCTURE FOR PETA-SCALE COMPUTING RESEARCH

S. Bradley, F. Burstein, B. Gibbard, D. Katramatos,
R. Popescu, D. Stampf, D. Yu, BNL, Upton, NY 11793, USA
L. Cottrell, Y. Li, SLAC, Menlo Park, CA 94025, USA
S. McKee, University of Michigan, Ann Arbor, MI 48109, USA

Abstract

TeraPaths, a DOE MICS/SciDAC funded project, deployed and prototyped the use of differentiated networking services to support the global movement of data in the high energy physics distributed computing environment. While this MPLS/LAN QoS work initially targeted networking issues specifically at BNL, the experience acquired and expertise developed is globally applicable to the ATLAS experiment and the high energy physics community in general. TeraPaths dedicates fractions of the available network bandwidth to ATLAS Tier 1 data movement and limits its disruptive impact on BNL's heavy ion physics program and other more general laboratory network needs. We developed a web service-based software system that automates the QoS configuration in LAN paths and negotiates network bandwidth with remote network domains on behalf of end users. Our system architecture can be easily integrated with other network management tools to provide a complete end-to-end QoS solution. We demonstrated the effectiveness of TeraPaths in data transfer activities within and/or originating from the Brookhaven National Laboratory. Our continued work focuses on strategically scheduling network resources to shorten the transfer time for mission critical data relocation, thus reducing the network error rates, which are proportional to transfer times. Such network resources typically span several administrative domains and exhibit unique management difficulties. Overall, our goal remains the provisioning of a robust and effective network infrastructure for high energy and nuclear physics research.

INTRODUCTION

The extreme demands in networking capacity encountered in the world of modern high energy and nuclear physics make evident the need for the capability to distinguish between various data flows and enable the network to treat each flow differently. Not all network flows are of equal priority and/or importance, however, the default network behavior is to treat them as such. Thus, the competition among flows for network bandwidth can cause severe slow downs for all flows, independent of importance, and furthermore cause some applications to fail.

As an example, the Brookhaven National Laboratory (BNL) routinely carries out Relativistic Heavy Ion Collider (RHIC) [1] production data transfers and Large Hadron Collider (LHC) [2] Monte Carlo challenges between the laboratory and several remote collaborators. The aggregate of the peak network requirements of such transfers is well beyond the capacity of the BNL network. To ensure that RHIC production data transfers are not affected, it is necessary to constrain LHC data transfers to opportunistically utilize available bandwidth.

The TeraPaths project enables data transfers with speed and reliability guarantees - crucial to applications with deadlines, expectations, and critical decision-making requirements - through the use of differentiated networking services. TeraPaths offers the capability to selectively and/or collectively configure network equipment to dedicate fractions of the available bandwidth to various data movements and/or replications, thus assuring adequate throughput and limiting the disruptive impact upon each other. This capability is deemed essential for the ATLAS [3] distributed data environment (see Figure 1).

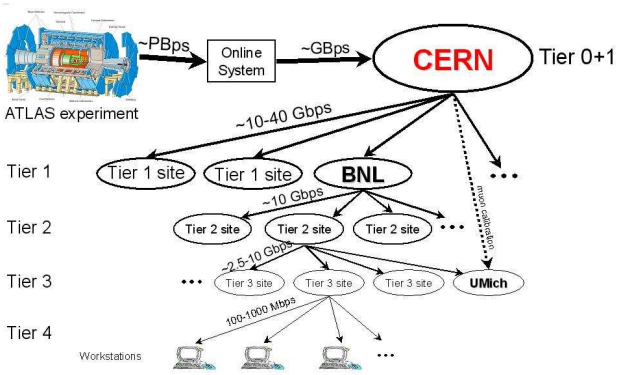

Figure 1: ATLAS data distribution.

The network managing capabilities enabled by this project can be integrated into the infrastructure of grid computing systems and allow the scheduling of network resources along with CPU and storage resources, to enhance the overall performance and efficiency of DOE computing facilities.

SYSTEM DESIGN

Modern networking hardware offers a range of architectures for providing QoS guarantees to data flows. We chose to design TeraPaths around the DiffServ architecture [4] because with this architecture traffic needs to be conditioned (policed/shaped) only at the network boundary. DiffServ is thus highly scalable. Up to 64 traffic categories - classes - are supported, using six bits of the Type of Service (ToS) byte, known as DSCP bits. Treatment of data is determined on a per-packet basis. In contrast, the IntServ architecture (RSVP protocol) determines treatment on a per-flow basis and thus requires the main-

tenance of flow information in all involved network devices.

The TeraPaths software configures and manages LAN QoS paths from end hosts to BNL border routers. Each such path can dedicate a percentage of the available site bandwidth to its assigned data traffic. Distinction between data packets is done by means of their DSCP markings (see Figure 2).

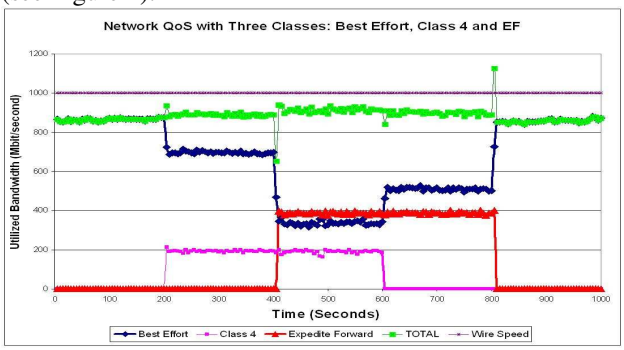

Figure 2: DiffServ traffic class example; Class 4 and Expedite Forward traffic streams remain at predetermined bandwidth levels at the expense of Best Effort traffic.

TeraPaths controls which traffic goes into each of the configured classes at the data flow level (a data flow is defined by a source and a destination IP address and port number pairs). Access to QoS paths is further controlled by advance reservations.

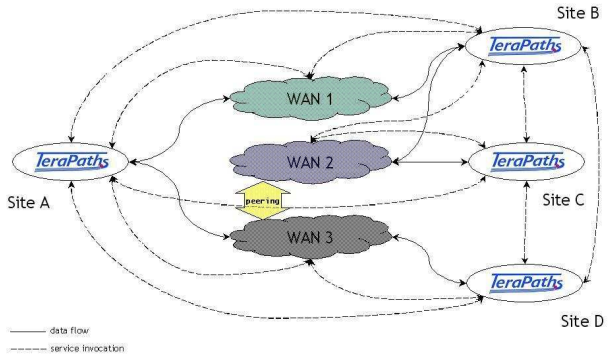

Figure 3: Conceptual view of the network.

Figure 3 shows the conceptual view of the network that TeraPaths utilizes for data transfers. End sites (A, B, C, and D) run the TeraPaths software for configuring and controlling their LANs. Communication between each pair of sites is done through one or more alternative WAN routes. WANs may have peering points, however, inter-WAN route setup is not in the list of TeraPaths responsibilities. End-to-end route setup for a data flow entails first setting up the LAN of the source site, then negotiating and selecting an MPLS [5] tunnel through a WAN, and finally setting up the LAN of the destination site, again through remote invocation of that site's TeraPaths software.

Figure 4 offers a more detailed view of the TeraPaths architecture, which is realized with the help of web services. Without loss of generality, it is assumed that site A initiates a data transfer to site B through a WAN cloud. Each site runs its own instance of the TeraPaths system. The system is comprised of a set of core services that cooperate with the necessary databases to provide user Authentication, Authorization, and Accounting (AAA), advance reservation scheduling, negotiate remote requests, and distribute network configuration commands to management nodes. Management nodes are hosts that supervise network devices (routers, switches) and are responsible for their configuration and monitoring. The configuration is performed through the invocation of a subset of hardware controlling services available at each management node. This subset of services offers a layer of abstraction between configuring requests and hardware configuration and invokes, in turn, suitable hardware drivers to "speak" the actual hardware language (e.g. Cisco IOS commands) and setup the necessary network equipment accordingly.

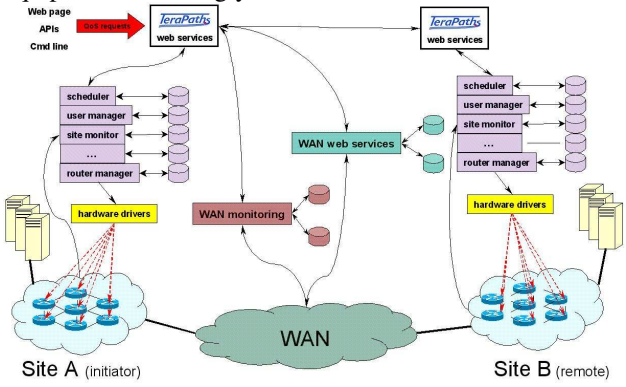

Figure 4: The TeraPaths architecture.

The initiating site is responsible for arranging an end-to-end LAN QoS/MPLS path from the source host at site A through the WAN to the destination host at site B. If the requested bandwidth can be reserved at the specified time locally (at site A), the system proceeds to request an MPLS tunnel through the WAN and a compatible QoS path through the LAN of site B. This is done by remotely invoking the corresponding interface of the WAN provider and the TeraPaths instance of site B and request appropriate reservations. Only when all three reservations can be obtained does the system proceed to actually put the reservations in place so that the end-to-end QoS path configuration can be guaranteed at the requested time. If one or more reservations cannot be obtained, the system responds to the user with a "counter offer", an alternative set of reservations similar to the originally requested. It is up to the user to accept or reject this counter offer.

TeraPaths can be invoked using a web interface, which graphically displays the available bandwidth classes and the existing reservations and facilitates the placement of new reservations. Alternatively, the services can be invoked through Application Programming Interfaces (APIs), further enabling the use of Command Line Interfaces (CLIs) and direct TeraPaths invocation from within applications.

Properly configured, TeraPaths can achieve a partitioning of a site's available bandwidth in statically and/or dynamically allocated slots, to accommodate the needs of a large number of data flows. Each slot, according to its type, is assigned to a class of service from the set of services classes pre-configured within the LAN perimeter. All LAN hardware knows how to treat packets belonging to each such service class, however, the actual policing/shaping of flows is done at the first piece of equipment encountered when leaving the source host. The network configuration module of TeraPaths can automatically reconfigure the entire network or part of the network that is under its control and modify the role and bandwidth assignments of service classes. This is, however, an infrequent administrative task, as the combination of statically and dynamically allocated bandwidth slots can satisfy a wide array of flow requirements.

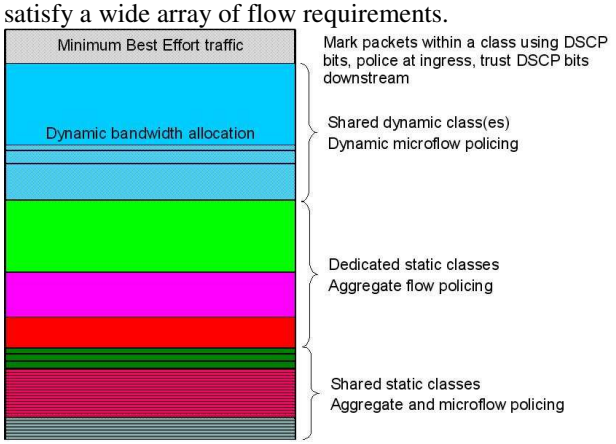

Figure 5: TeraPaths bandwidth partitioning scheme.

Figure 5 shows an example qualitative partitioning scheme of a site's available bandwidth. Dedicated static classes are policed on an aggregate bandwidth basis. That is, while a single flow can utilize all of the pre-determined bandwidth, any additional flow will cause the bandwidth to be equally shared among all flows assigned to the same class. Shared static classes are policed on an aggregate and a per-flow bandwidth basis. In this case, the bandwidth allocated for a class is further divided to a number of sub-slots defined by the specified per-flow bandwidth fraction. While the number of flows is less or equal to the number of sub-slots, each flow receives exactly one bandwidth fraction. If the number of flows exceeds the number of available sub-slots the total allocated bandwidth will still be observed by reducing the bandwidth fraction accordingly. Finally, shared dynamic classes are classes selected for their widely recognizable DSCP markings, e.g. the Expedite Forward (EF) class, and thus bound to be honored even by older generation equipment. These classes are assigned a portion of a site's bandwidth which can be further distributed to a number of data flows dynamically, i.e. the per-flow bandwidth is not predetermined but allocated to flows according to requests as site policy permits. Summarizing, site bandwidth partitioning under TeraPaths provides the mechanism to satisfy a variety of bandwidth allocation policies. Thus, frequent, high-priority flows can have their own, dedicated, class; groups of hosts can share a dedicated class without affecting other traffic; flows that require small bandwidth amounts can be funneled into the same shared static class, thus reducing the number of in-use classes (recall that a total of only 64 classes with corresponding DSCP markings is possible); finally, shared dynamic classes can be utilized to cover the needs of flows that cannot be otherwise satisfied. It should be pointed out that bandwidth partitioning occurs only when privileged flows are present. In absence of such flows, the network resumes best effort behavior. Nevertheless, there should always be a minimal fraction of bandwidth allocated for class 0 (best effort) so that common traffic can always proceed through the network.

Network monitoring is necessary as a reliable means of determining if and how well QoS paths are working. Furthermore, although monitoring information is not directly necessary for carrying out QoS path setup requests, it is anticipated that, in the future, monitoring information will be utilized in an MPLS-capable version of TeraPaths that will also provide route selection options within a site's complex LAN, based on specified criteria (e.g. bottleneck avoidance).

DEVELOPMENT AND EXPERIMENTS

The TeraPaths software is being developed using the proven underlying network communications technology of web services. The web services technology is secure, reliable, freely available, and permits the designer to specify the services offered by each administrative domain without specifying how they will be implemented. The BNL implementation of TeraPaths uses Java-based web services and a MySQL database to program Cisco routers. However, end-users can only see the fact that BNL provides a service that permits them to negotiate for a fraction of bandwidth, at a particular time and for a specific duration. For TeraPaths to be easily adopted by other end-users and deployed at their sites, we only use freely available software (e.g. Java and GlassFish or Jboss application servers) and standard distribution techniques (WAR files). Although TeraPaths is designed to request WAN MPLS tunnels by invoking the web services of WAN providers implementing mutually agreed-upon interfaces, we can use the façade design pattern to wrap any services a provider makes available and match our software's requirements.

For testing purposes, we put together a fully featured test bed using the same Cisco hardware as in the BNL production network [6]. This test bed allows for all kinds of experiments without the risk of adversely affecting the production network. The TeraPaths software is developed and tested on the test bed's private network, then gradually extended to larger and more public network segments at BNL to verify its robustness and ability to co-exist with existing traffic. Figure 6 displays a simple test bed experiment. Two iperf streams initially share the bandwidth of the gigabit link between the two test bed routers (ap-

proximately 60MB/s each stream). While a TeraPaths reservation for the Class 2 iperf stream is active, the bandwidth that stream occupies falls to the reserved 30MB/s, conceding the rest 90MB/s to the competing iperf stream.

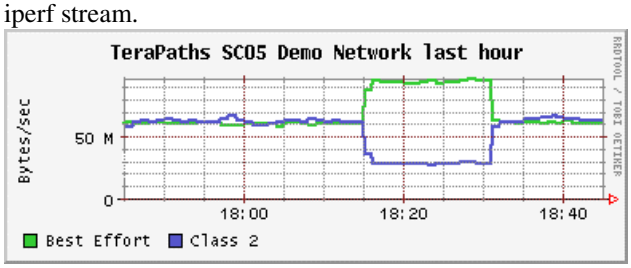

Figure 6: Protected vs. best effort traffic on test bed.

Figure 7 depicts a more complex experiment demonstrated at SuperComputing 2005. Here, two bbcp disk-to-disk data transfers, one at 200Mb/s and one at 400Mb/s, are protected from background competing iperf traffic through an ES-net MPLS tunnel. Only the iperf, best effort, traffic gets affected by the bbcp transfers which do not interfere with each other and maintain constant, predetermined bandwidth throughout each transfer cycle.

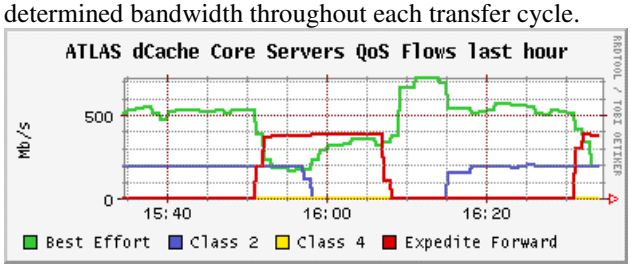

Figure 7: BNL to U. of Michigan: two bbcp disk-to-disk transfers (at 200Mb/s and 400Mb/s) against competing iperf traffic through ES-net MPLS tunnel.

COLLABORATION

We are closely collaborating with ES-net's OSCARS [7] project to allocate end-to-end network paths with specified bandwidth. So far, we have studied the behaviour of LAN QoS/MPLS paths and verified that traffic with specific source and destination IP addresses is properly labelled and tunnelled, and policed to pre-arranged bandwidth limits. In addition, we verified that the MPLS tunnel honors the packet QoS markings. We are currently working on fully automating the setting up of MPLS tunnels, through which QoS flows originating from or destined to BNL pass, by using web services provided by OSCARS and invoked by TeraPaths.

We are also collaborating with SLAC to analyze network monitoring needs and requirements. The SLAC IEPM-BW monitoring suite was installed at BNL and is currently monitoring network connectivity and performance between BNL and various partner sites in both Europe and the United States. Active tests are performed from a high-performance gigabit-enabled host using tools such as ping, iperf, thrulay, and pathchirp. In the future, anomalous event detection will also be performed to aid the automatic discovery of network problems.

SLAC will also monitor the QoS path between a typical Tier-1 to Tier-2 or Tier-3 peering site (such as that of BNL to University of Michigan and BNL to SLAC) across the WAN infrastructure by both passive and active means. To reduce the effect of active measurements upon other traffic on QoS paths, passive measurements will be conducted through a combination of SNMP and Netflow data in order to graphically represent pertinent QoS related network performance information for network management and QoS validation.

In order to gain real experience on the effects of transferring data over multiple disparate QoS paths, SLAC has also worked closely with the OSCARS project. Multiple concurrent flows were initiated over an artificially congested ES-net path between BNL and SLAC and the effectiveness of the implemented EF class in terms of IPDV was proven across real production environments [8].

CONCLUSIONS AND FUTURE WORK

TeraPaths demonstrates that the combination of LAN QoS techniques, based on the DiffServ architecture, combined with WAN MPLS tunnels is a feasible and reliable approach to providing end-to-end, dedicated bandwidth paths to data flows in the demanding environment of high energy and nuclear physics. TeraPaths offers a flexible way to partition a site's available bandwidth into predetermined bandwidth slots to protect various data flows from competing against each other.

A series of experiments at BNL, using both a test bed and the production network, indicate that LAN QoS does not impact the overall network utilization. We did not observe any performance deterioration while the QoS policy was active. We, nevertheless, need to verify that the same observation is also valid for the high-speed network (10 Gbps and up), that currently is in the process of being installed at BNL, to ensure that using QoS/MPLS path configurations does not interfere with other data transfers at high rates.

Our future plans, except for completing the work-in-progress of fully automating the setup of end-to-end QoS paths across different administrative domains, also include supporting Virtual Organizations (VOs), adding monitoring information and/or policy-based route selection within complex LANs using MPLS, and pursuing further enhancements to the TeraPaths system to widen its deployment to higher and lower tier sites.

REFERENCES

[1] http://www.bnl.gov/RHIC/
[2] http://lhc.web.cern.ch/lhc/
[3] http://www.usatlas.bnl.gov/
[4] An architecture for differentiated services, RFC2475
[5] Multiprotocol label switching architecture, RFC3031
[6] http://www.usatlas.bnl.gov/twiki/bin/view/Projects/TeraPaths
[7] http://www.es.net/oscars/
[8] http://www-iepm.slac.stanford.edu/dwmi/oscars