Investigating the behavior of network aware applications with flow- based path selection.

A. Bobyshev, M. Crawford, V. Grigaliunas, M.Grigoriev, R. Rechenmacher , FNAL, Batavia, IL 60510, USA

ABSTRACT

To satisfy the requirements of US-CMS, D0, CDF, SDSS and other experiments, Fermilab has established an optical path to the StarLight exchange point in Chicago. It gives access to multiple advanced research networks, such as UltraScience Net, UltraLight, UKLight, and others. These networks offer very high bandwidth capacity that is generally unavailable via production network paths. The Lambda Station project is developing a path forwarding system for interfacing production mass storage clusters with these experimental networks to enable efficient bulk data movement. The goal is to design a system capable of alternate path forwarding on a per flow basis. One important aspect of this project is investigating the behavior of end-node operating systems and applications in the presence of per-flow rerouting. This article introduces our findings and current status of the research in this area. Our focus is on Linux as the operating system, and SRM (Storage Resource Manager), GridFTP, and dCache as network aware applications.

PROBLEM DESCRIPTION

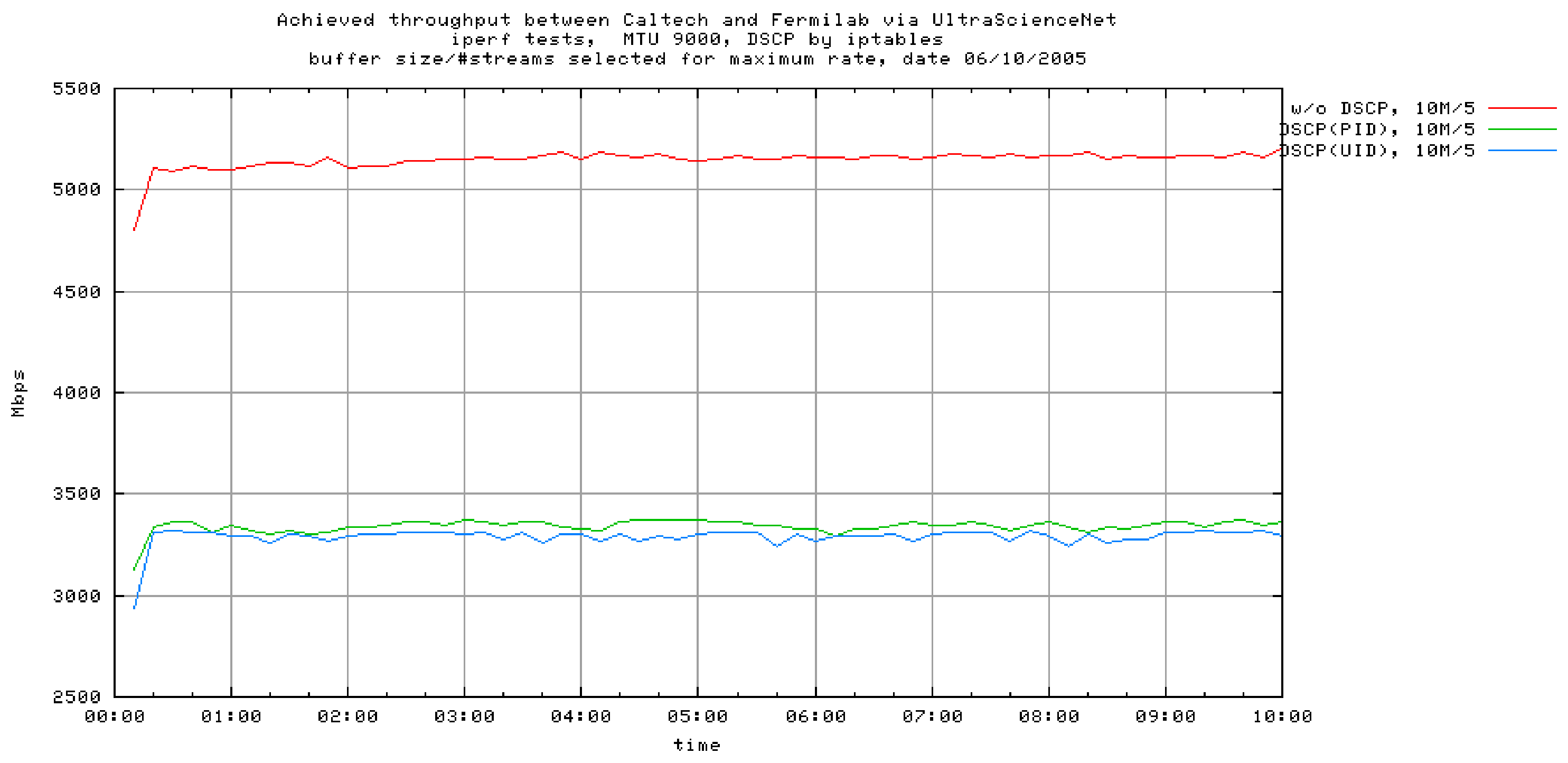
Over the past several years a lot funding, research and deployment efforts have been put into optical-based, advanced research networks, such as National Lambda Rail, UltraLightNet, DataTag, CAnet4, Netherlight, UKLight and very recently DOE UltraScienceNet . These advanced infrastructures have a great potential for data movement required by the Particle Physics Collaborations. However, there are a number reasons why access to these networks is not available for general ISP-like connectivity and hence require admission and forwarding mechanisms . Here are some of them :

- these networks are typically community oriented, deployed to serve the needs of certain user's groups and have peering agreements only with a limited number of networks
- in general, the support in such networks is sub-production level and hence organizations can not entirely trust their business to it.
- modern applications require links with different characteristics, e.g voice over IP is delay sensitive, data movement needs high bandwidth but propagation delays are not very critical

The issue of integrating existing production-use computing facilities to the local network infrastructure is not yet addressed.

EFFECT OF DSCP TAGGING BY USING IPTABLES

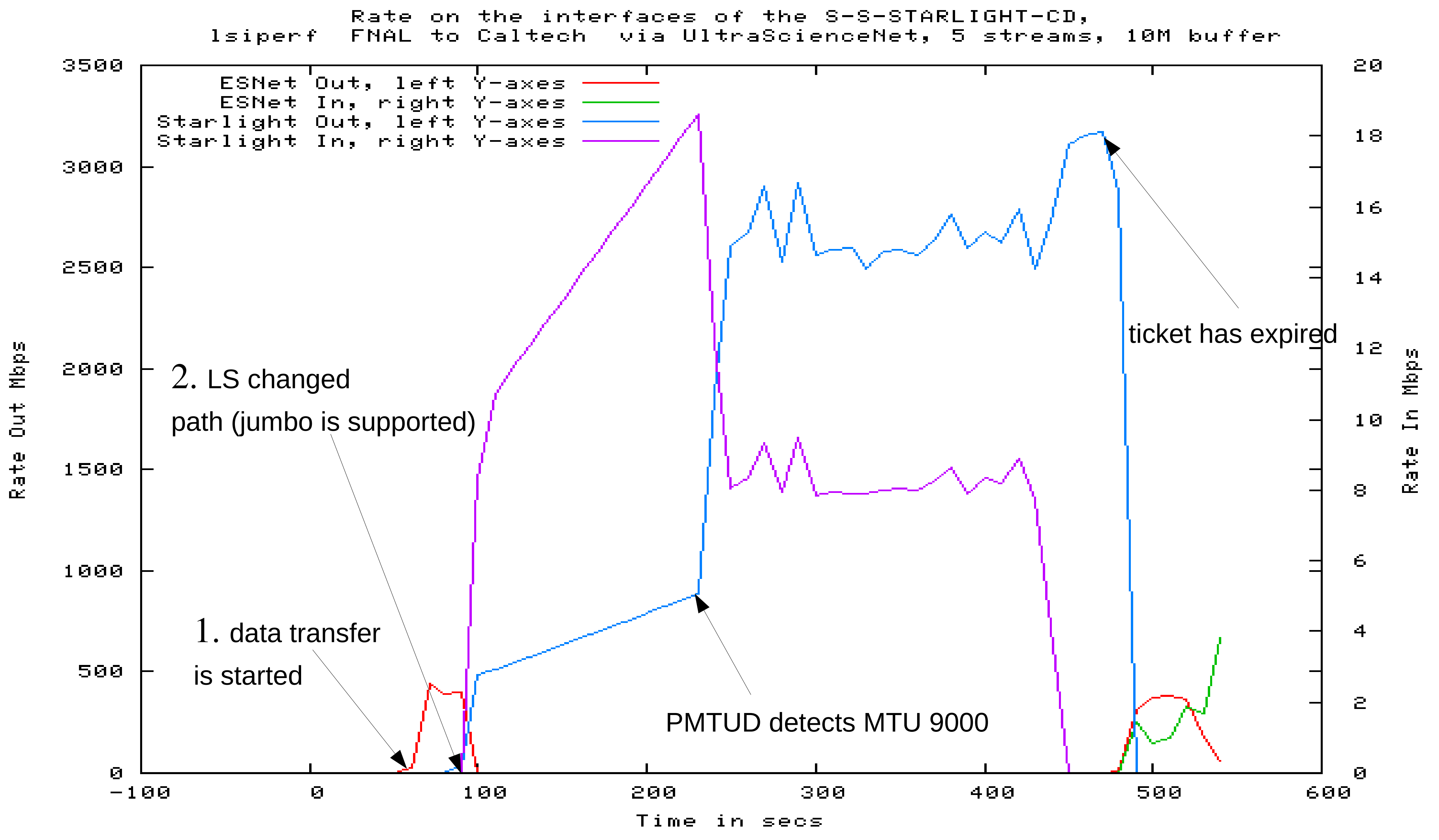
IPTables is a tool that allows user's programs to configure the Linux 2.4 and 2.6 kernel for packet filtering and altering IPv4 packet headers. Setting DSCP tags can be configured for selective traffic based on several criteria such as UID, PID, SID and the name of a command (first 16 characters). This feature is very useful when applications can not be modified to support network awareness or, in our case, a Lambda Station awareness and a proxy service will act on behalf of the application to tag its traffic when required. Obviously, it should be expected that overall end-to-end performance will be affected by this external tagging process. We investigated such an effect of UID- and PID- based traffic selection for data transfer across a 10G UltraScienceNet path. The following hardware was used: Dual XEON 2.6GHz, 4GB RAM, 100MHz PCX-1.1, Intel/PRO 10G NIC and 2x1.9GHz AMD Opteron, 4GB RAM, S2IO 10G NIC. The graphs below introduce the results of throughput measurements between Fermilab and Caltech via the 10Gbps path provided by UltraScienceNet. Measurements were taken by using the IPERF tool for different buffer sizes and number of parallel streams. Traffic for DSCP tagging was selected based on UID or PID. Very accurate tunning of computing systems, both hardware and software, and applications is required to achieve high throughput via a pipe with a long delay. It is not necessarily true that the same parameters will be optimal for transferring with or without DSCP tagging. That is why to compare the effect of DSCP tagging, we considered the best achieved results from multiple tests. As can be seen, performance dropped by 5-20%.



FLOW SWITCHING BETWEEN TWO PATHS WITH DIFFERENT A SMALLER MTU

The support of jumbo frames by network infrastructure is typically essential to get a reasonable high throughput between hosts connected to the network by 10G NICs. While deployment of support for jumbo frames in R&D networks is manageable and even commonplace, jumbo frame support in production networks, both WAN and LAN, is still problematic. That is why flows rerouted on application's demand may be switched between alternative paths with different Maximum Transmissions Unit (MTU) sizes in routers along the traffic's path. This is not a new problem for IP networks. Path Maximum Transmission Unit Discovery (PMTUD) was developed to adapt the Maximum Segment Size (MSS) for a smaller MTU between two endpoints. However, in our research we found that most current implementations of the PMTUD algorithm are not very robust. When policy based routing is utilized instead of conventional destination IP address routing tables, replies for probes can be lost (in addition to usual packet loss in the Internet) or re-routed by intermediate routers. The recent IETF draft[1] of the PMTUD algorithm has addressed most of our observations, however, it is not yet implemented.

Graphs in the figure below show a typical behavior for switching selective flows between two paths with different characteristics. A test application (lsiperf) starts transferring data via production net with support for a 1500 byte MTU. For the two same points, Caltech and Fermilab, there is an alternative 10Gb/s path via UltraScienceNet that supports a 9000 byte MTU. In the background, the application opens a ticket for the alternative path via UltraScienceNet and watches the progress of its provisioning for selective flows. When the new path is established, as indicated by the ticket's status, the application starts tagging traffic by DSCP and selected flows will be forwarded to the high bandwidth network infrastructure that supports jumbo frames (point 2 in figure below). The end systems can now use jumbo frames. However, it takes up to 5 minutes for current implementations of PMTUD to determine it. When the ticket is expired and networks are reconfigured to forward the same flows back via the production path, the transmitter changes its MSS according to an MTU of 1500. This situation is detected by PMTUD practically immediately because no data segments bigger than 1460 bytes can traverse the end-to-end path.



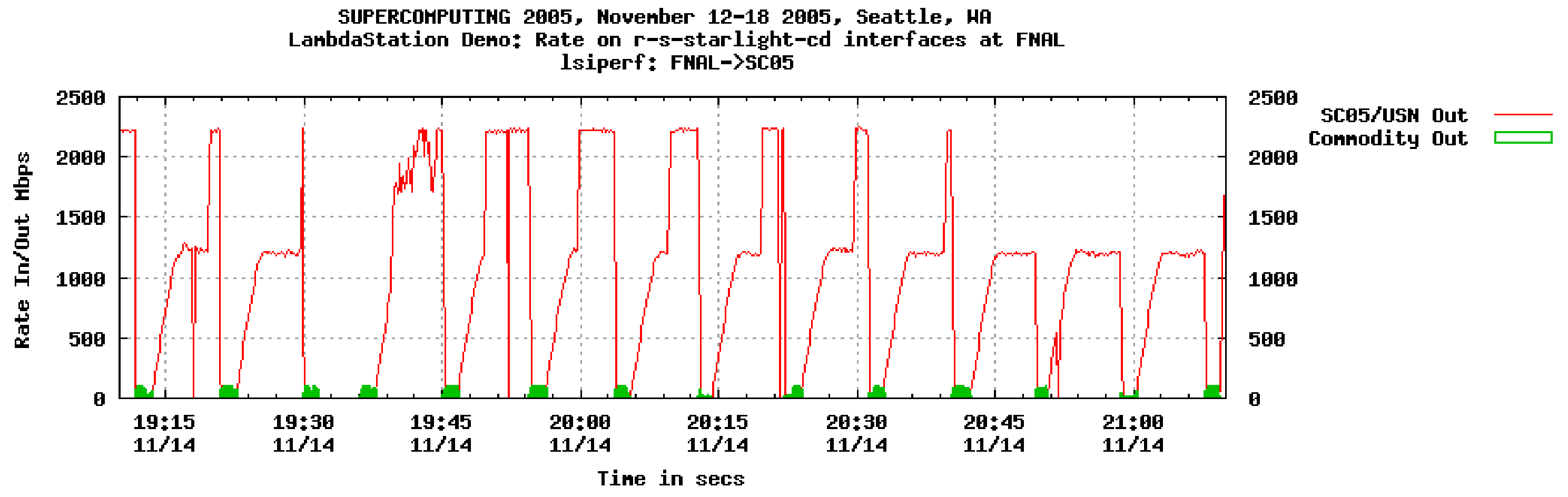
2. LS changed path (jumbo is supported)

1. data transfer is started

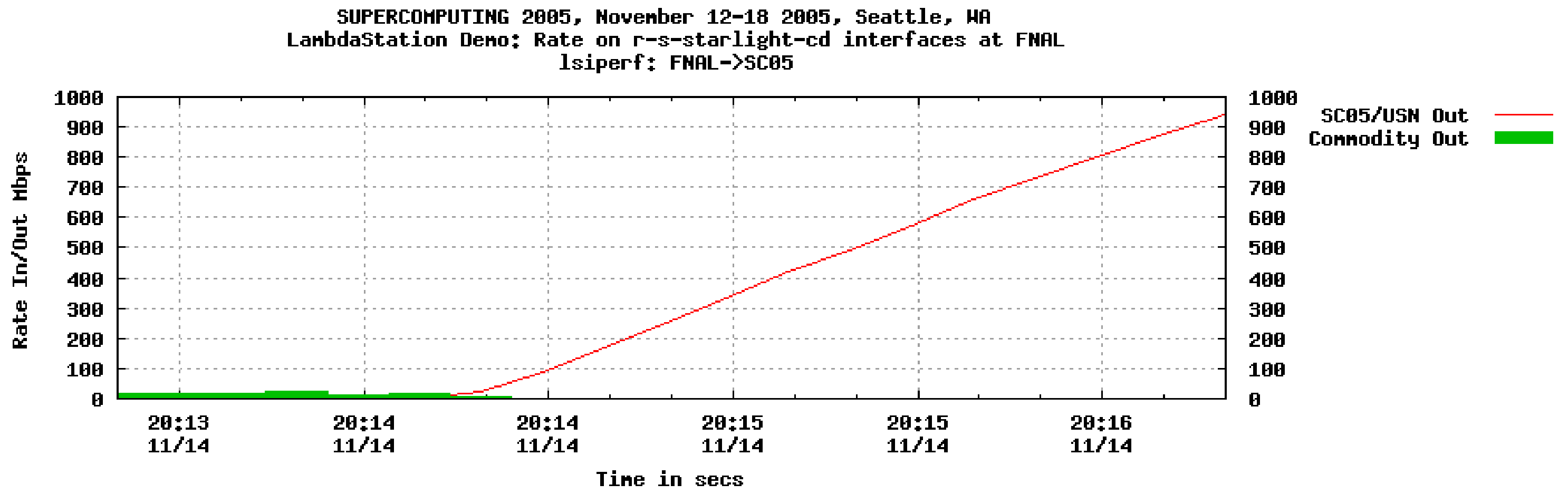
PMTUD detects MTU 9000

ticket has expired

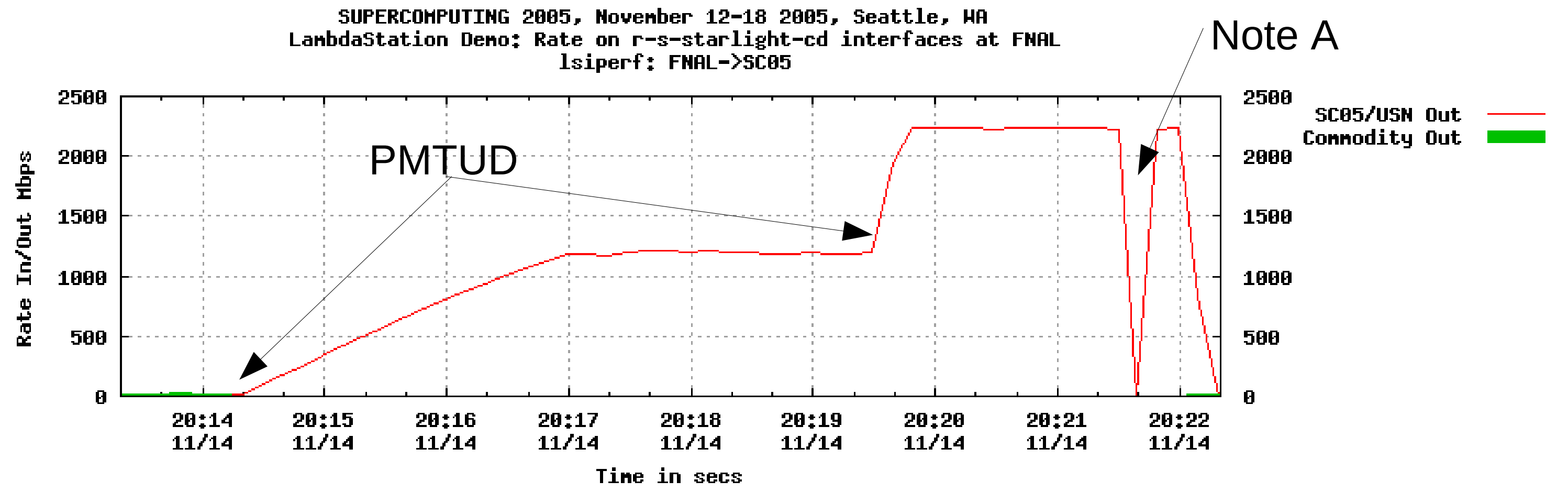
The graphs below show the results of performance measurements for flow based switching during Super Computing 2005, Seattle, WA, November 12 – 18, 2005. Two sites, Fermilab/Chicago and SC05/Seattle, were prepared to run this demo. At each site there were two test machines with 10G interfaces and one server running as a site LambdaStation (LS). Two alternative paths were available, the commodity Internet and the 10G path from the SC05 booth to the StarLight POP in Chicago and then to Fermilab.



Tickets were placed periodically by application to switch its flows via 10G direct path from SC05/Seattle to Fermilab/Chicago.



Initial interval after flows are switched



Note A

PMTUD

Note A: This dip has been identified as statistics gathering anomaly, believe to be an ASIC problem associated with router hardware. Measurement of the same throughput by other means (e.g. applications statistics) does not show that drop of rate..

REFERENCES

1.Path MTU Discovery, draft-ietf-pmtud-method-05, Network Working Group, Internet-Draft, IETFhttp://www.ietf.org/internet-drafts/draft-ietf-pmtud-method-05.txt, M.Mathis, J.Heffner,PSC, October 23, 2005

2. A Lambda Station Project Web Site, http://www.lambdastation.org/

3.Phil DeMar, Donald L.Petravic, LambdaStation: A forwarding and admission control service to interface production network facilities with advanced research network paths, CHEP2004 International Conference on Super Computing, Interlaken, Switzerland, 27th September - 1st October 2004