Use of Alternate Path WAN Circuits at Fermilab

Phil DeMar, Andrey Bobyshev, Matt Crawford, Vyto Grigaliunas

Fermilab, PO BOX 500, Batavia, IL 60510, USA

demar@fnal.gov

Abstract. Fermilab hosts the American Tier-1 Center for the LHC/CMS experiment. In preparation for the startup of CMS, and building upon extensive experience supporting Tevatron experiments and other science collaborations, the Laboratory has established high bandwidth, end-to-end (E2E) circuits with a number of US-CMS Tier2 sites, as well as other research facilities in the collaboration. These circuits provide preferred network paths for movement of high volumes of CMS data and represent a departure from the traditional approach of utilizing the general research and education (R&E) network infrastructure for movement of science data. All circuits are statically configured and are based on a variety of underlying network technologies. These circuits are presumed to provide more predictable performance, and they avoid the traffic contention concerns of general-use R&E network links. But the circuits also add significant complexity and effort for the Laboratory's wide area network support. This presentation will discuss Fermilab's experiences with deploying, managing, and utilizing E2E circuits as preferred network paths in parallel with the general IP R&E network infrastructure. Alternate path routing techniques, monitoring issues, troubleshooting, and failover concerns will be covered.

1. Introduction

Fermilab moves extremely large amounts of data offsite in support of its US-CMS Tier-1 Center, as well as for other active experiments based at the Laboratory. These bulk data transfers are typically characterized by high data transfer rates over long duty cycles to a modest number of predictable remote locations. This type of bulk data transfer, common to large-scale, collaborative science projects, is frequently labeled high impact data movement. Using a facility's general internet network path for a project's high impact data transfers can cause problems for both the project and the facility's general user community, if that path isn't sufficiently capacious to accommodate both types of traffic simultaneously. An alternative approach is to separate the high impact data traffic from the facility's general internet traffic by sending it over a different network path, one that can be configured to meet the specific requirements for the project. For the past two years, Fermilab has been establishing and supporting end-to-end circuits to facilitate high impact data movement with a select number of remote sites collaborating on CMS or Run-II experiments. This paper describes Fermilab's experiences with alternate path wide-area circuits, including deployment and support issues, concerns about manageability, and future directions.

1.1. What is an Alternate Path Data Circuit?

There is no concise definition of what constitutes an alternate path data circuit. A number of different terms have been used within the Research & Education (R&E) community, including end-to-end

(E2E) circuits and light paths, to describe virtual point-to-point network connections between two sites. Ethernet is assumed to be the connection interface type. The underlying network infrastructure utilized to establish one of these connections may well be a mix of network technologies, and bandwidth may be shared in some manner. But from the perspective of the end sites, it will appear to be a direct ethernet connection between the two, with no intermediate routing. It is also worth noting that an 'end site' can be a special use network, and doesn't necessarily have to be a facility. However, end-to-end data circuits are assumed to bypass the R&E routed infrastructure, made up of backbone networks such as ESnet, Internet2, GEANT, and other national research networks (NRENs). For purposes of this paper, the term E2E circuit will be used to refer to one of these data circuits.

1.2. Motivation for Alternate Path WAN Circuits

The motivation for deploying and utilizing E2E circuits at Fermilab was a convergence of three things, need, capability, and strategic direction.

1.2.1. *Need:* Fermilab hosts the US Tier-1 Center for CMS. The Tier-1 Center has an obligation to receive and store a significant portion of the raw data generated at the experiment. It is functionally part of a distributed data acquisition system for the experiment. Data movement requirements are steady and predictive. In addition, the Tier-1 Center is obligated to process its share of the raw data, and make the processed data available for the experiment's Tier-2 Centers for analysis. While the projected Tier-1/Tier-2 traffic is less predictive and more bursty in nature than the raw data traffic, the projected aggregate data movement is sufficiently large that reliance on use of the general internet infrastructure was not deemed prudent. In the case of the raw data movement between the Tier-0 (CERN) and all the Tier-1 facilities, a collaborative decision was made to implement a dedicated network infrastructure, with E2E circuits between the Tier-0 and each Tier-1. This dedicated network infrastructure is named the LHC Optical Private Network (LHCOPN).

1.2.2. *Capability:* The ability to deploy E2E circuits is largely governed by two factors, the facility's network infrastructure that's available for offsite connectivity, and facility's topological proximity to other R&E network infrastructures of interest. Fermilab was fortunate to be within reasonable proximity (~96km) to the StarLight optical network exchange in Chicago. In 2004, the Laboratory leased an optical fiber pair down to StarLight, and deployed an optical network infrastructure, called Fermi LightPath, that was based on dense wave division equipment (DWDM). The DWDM infrastructure provided multiple 10Gb/s ethernet channels between the Laboratory and StarLight. Some of those channels were available for establishment of E2E circuits. As an international network exchange, StarLight offered a plethora of connection opportunities with other national and international networks to create E2E circuits.

1.2.3. *Strategic Direction:* In 2003, the US Department of Energy held a workshop to decide its strategic network plans for the next 5-10 years. The workshop [1] recognized the high impact data movement requirements for emerging large scale science projects, and adopted a dual network strategy. The classic routed IP infrastructure, provided by ESnet, would be maintained with high bandwidth, highly reliable service. In addition, a parallel network, would be established specifically to support high impact data movement. That network, called the Science Data Network (SDN), would also be maintained by ESnet, and would support E2E circuits for the high impact data. The planning and deployment of the SDN by ESnet provided the Laboratory with the confidence to move ahead with its E2E circuit support.

2. E2E Circuits

At their core, E2E circuits are typically based on extended Ethernet VLans. The underlying network technology supporting the extended ethernet connection can vary. Native ethernet is the obvious,

simple, and most common technology employed. However, Ethernet connections can be supported across MPLS WAN infrastructures, or channelized SONET infrastructures as well. It's also not uncommon for an E2E circuit to be comprised of a mixture of different underlying technologies, if the circuit crosses multiple wide area network infrastructures. VLan trunking allows different E2E circuits to share a common physical network link. Under that scenario, an E2E circuit may appear to be a dedicated pipe, but may in fact be sharing bandwidth with other E2E circuits across certain network segments.

2.1. Typical Circuit Structure

The basic components of an E2E circuit are routers at each end point of the circuit, a concatenation of ethernet links that comprise the layer-2 path between the two routers, and a common VLan configured across all the ethernet segments. The ethernet frame (MTU) size is standardized across all the segments, normally to either the default ethernet size of 1500Bytes, or 9000Bytes, if jumbo frames are utilized. Fermilab E2E circuits currently are all configured for 1500Byte ethernet frames. Figure 1 is an example of an E2E circuit currently supported between Fermilab and Brookhaven National Laboratory. The circuit is a manually-configured concatenation of ethernet segments between switches at five locations, and crosses three different intermediate network domains. While most of the circuit segments use native ethernet for underlying transport, the segment across the Internet2 domain uses channelized SONET service for underlying transport. A common VLan carried across all these segments provides the virtual end-to-end ethernet connection between the routers (red) at both ends. The ethernet segments within the two metropolitan area networks (MANs) support multiple E2E circuit VLans, so the bandwidth across those segments is shared.

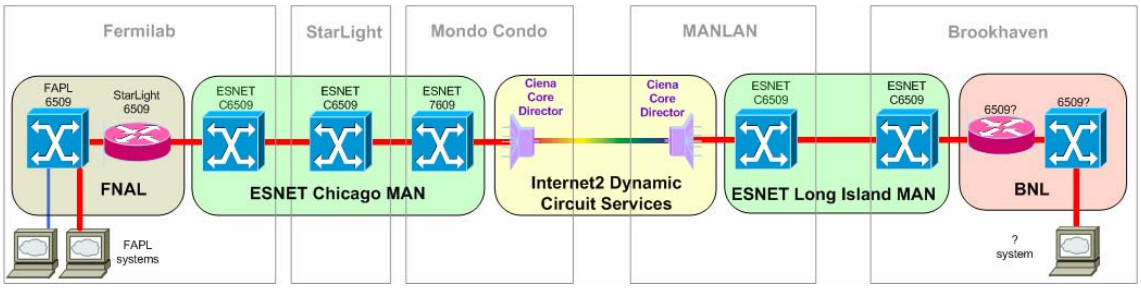

Figure 1: Example of E2E Circuit Components

2.2. Making the Routing Work

E2E circuits provide alternate network paths for movement of data between two sites. Fermilab's model for use of E2E circuits is selective forwarding of designated high impact traffic. An objective is to be able to send high impact traffic between two sites across the E2E circuit while concurrently sending other network traffic between the same two sites across the general routed IP infrastructure. High impact traffic flows are identified by source/destination address pairs, or netblocks in the case of computing clusters. Policy-based routing (PBR) [2] is utilized to forward the data flows with those souce/destination address pairs along the alternate path. Figure 2 (left) depicts the routing from the CMS Tier-1 Center at Fermilab to several US-CMS Tier-2 sites. By default, traffic sourced from the Tier-1 follows the general IP routed infrastructure (blue). The traffic egresses through the facility border router, and is subsequently routed by ESnet to other R&E networks,

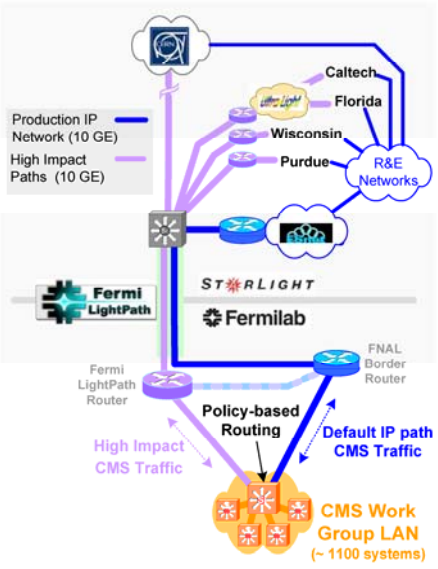

Figure 2: Policy-based routing

eventually reaching the Tier-2s. E2E circuits to the Tier-2 sites terminate in a separate router, also on the facility network perimeter. Traffic bound for any of those Tier-2 facilities is rerouted within the CMS Tier-1 LAN via PBR, using the source netblock for the Tier-1 storage system and destination address netblocks for the respective Tier-2 sites, to a separate, dedicated network link (purple) that connects to the facility E2E circuit router. The router forwards the traffic on to the remote site via the E2E circuit. If the E2E circuit happens to be down, the traffic is simply redirected over to the facility border router, where it is then routed across the general IP internet.

Routing symmetry isn't necessarily a requirement, but is considered highly desirable, particularly for E2E circuits. It is our policy that configuration of an E2E circuit with another site include reciprocal routing across the same circuit back to Fermilab. The implementation of the reciprocal routing is left to discretion of the remote site. However, PBR is implemented inbound on the E2E circuit, where it terminates on the Fermilab router. Arriving network traffic that doesn't match the expected source/destination address combinations gets rerouted over to the facility border router, where is routed into the site, subject to the normal restrictions and access controls.

3. E2E Circuit Deployment at Fermilab

Fermilab has been deploying E2E circuits ever since its optical fiber infrastructure down to StarLight was deployed in 2004. The circuits have served a wide spectrum of Laboratory experiments, involved a mixture of underlying network technologies, and varied in available bandwidth. Table 1 (right) depicts the list of E2E circuits supported. A total of 15 circuits have been supported, of which three have been subsequently torn down. The reasons for the decommissioning those circuits is revealing. In two instances, the level of service across the general internet path rose to a point where comparable network performance was provided to the performance across the circuit path. In absence of any performance boost from using the E2E circuit, the end users were perfectly willing to revert back to general IP connectivity. Note that the bandwidth allocated to those circuits was modest. Very high bandwidth data movement might still be preferred across an E2E circuit, even if comparable network performance is available via the routed IP network. The third E2E circuit was decommissioned due to end of a centrally-funded project. The circuit user apparently felt that while the circuit may have been useful, it was not sufficiently useful to pay for.

Remote Site	Experiment	Transit Provider(s)	Max. B.W.	Status
UCL. UK	CDF	UKLight	1 Gb/s	Moderate use
CERN (LHC)	CMS	US-LHCnet	10 Gb/s	LHCOPN
Simon Fraser	D0	CAnet4; WestGrid (BC)	1 Gb/s	decommissioned
Caltech	CMS	UltraLight	10 Gb/s	T1/T2 data
Apache Pt (NM)	SDSS	ESnet (MPLS)	<< 1Gb/s	decommissioned
Sinica, Taiwan	CDF	ASnet	2.5 Gb/s	Intermittent use
Florida	CMS	UltraLight; FLR	10 Gb/s	T1/T2 data
McGill	CDF / D0	CAnet4	1 Gb/s	Intermittent use
NCHC, Taiwan	SDSS	Twaren	1 Gb/s	Intermittent use
IoP; Prague, Cz	D0	Surfnet; CESnet	1 Gb/s	Intermittent use
UCSD	CMS	ESnet (SDN)	10Gb/s	T1/T2 data
Wisconsin	CMS	WISnet	10 Gb/s	T1/T2 data
Purdue	CMS	Purdue	10 Gb/s	T1/T2 data
IN2P3 , France	D0 (CMS?)	ESnet,HOPI,GEANT	Two x 1Gb/s	Intermittent use
BNL	LHC	Internet2 Dynamic Circuit Service	N x 1Gb/s	Testing

Table 1: E2E Circuit Deployment

3.1. Current E2E Circuit Topology

The Fermi LightPath optical network infrastructure provides the underlying data channels that support the E2E circuits. The data channels provided by the initial Fermi LightPath configuration have since been augmented by four 10Gb/s ESnet MAN channels, including three Science Data Network (SDN) channels. The SDN channels have been deployed specifically to support high impact data movement, and are intended to provide network paths for E2E circuits.

The original Fermi LightPath configuration provides 1Gb/s and 10Gb/s channels between the E2E circuit router at Fermilab and a Fermilab-managed high performance switch down at StarLight. Most of the original E2E circuits were implemented across those two data channels. Fiber jumper cables were run between the switch at StarLight and cooperating R&E network infrastructures, providing the physical connectivity to enable establishment of E2E network paths.

The SDN topology between Fermilab and StarLight is very similar. ESnet, manager of the MAN SDN channels, has deployed two high performance MAN switches, one at Fermilab and the other at StarLight. The three 10Gb/s SDN channels run between those two switches. A fourth 10Gb/s data channel, also running between the two switches, supports the general routed IP service. The Fermilab-based router that supports the E2E circuits has three 10Gb/s connections to it's adjacent ESnet MAN switch, one for each of the three SDN channels. On the StarLight end of the SDN channels, the ESnet switch has physical connections to cooperating R&E network infrastructures that have a presence at StarLight. E2E circuits can then be established, using one of the SDN channels and the physical connection to the cooperating R&E network. Figure 3 displays the Laboratory's E2E circuit topology.

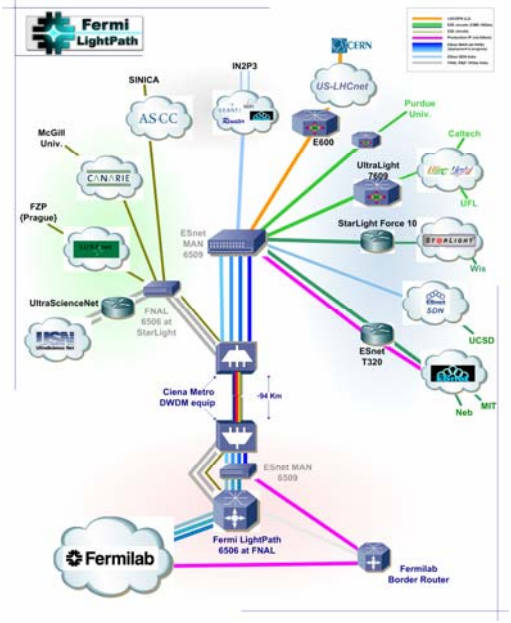

Figure 3: E2E Circuit Topology

The Laboratory's goal has been to utilize the SDN channels for all its E2E circuits. The long term objective is to free up the original Fermi LightPath channels for network R&D projects. As a result, a transition is underway to move existing E2E circuits over to the SDN channels. At the current time, all of the 10Gb/s circuits, essentially the CMS E2E circuits, have been moved over to SDN channels. One of the three SDN channels is dedicated to the E2E circuit between LHC Tier-0 Center (CERN) and the Tier-1 Center at the Laboratory. The other two SDN channels support E2E circuits to the US Tier-2 Centers. E2E circuits that remain on the original Fermi LightPath channels include several legacy 1Gb/s circuits, and 10Gb/s circuit to the UltraScience Network for network research projects

3.2. Current Use Patterns

For the past two years, offsite traffic loads at the Laboratory have been driven by data movement out of and into the CMS Tier-1 Center. As previously noted, many of the E2E circuits, in particular the circuits across 10Gb/s infrastructure, were put in place to support CMS data movement. Figure 4 shows the outbound monthly data rates over the past two years. From the graph, it is apparent that the bulk (84%) of the network traffic going offsite was carried on the E2E circuits. This traffic is almost entirely out of the CMS Tier-1 Center. As a point of reference, the 2.1 petabyes of data moved offsite in July, 2007, is equivalent to a continuous data stream

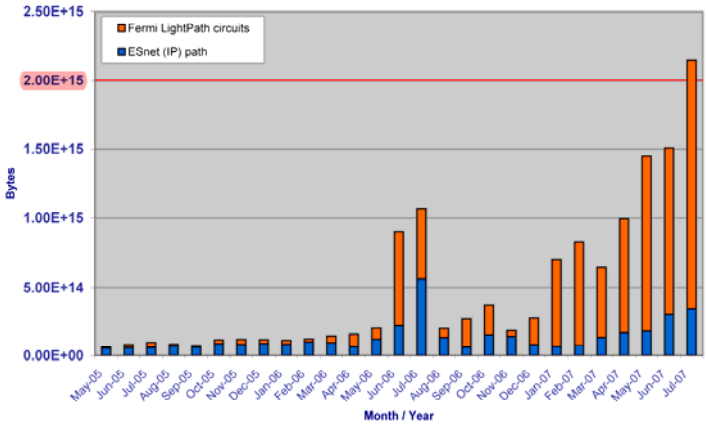

Figure 4: Offsite Traffic Levels (Outbound)

of 6.7Gb/s on a 24/7 basis for the entire month.

The Laboratory's inbound data rates are normally significantly lower than outbound data rates. However, the same general pattern of E2E circuits carrying the bulk of the data movement still holds. In July, 2007, the Laboratory's inbound traffic level was 0.64 petabytes for the month, of which 87% was carried on E2E circuits.

4. E2E Circuit Support Issues

While E2E circuits have demonstrated their usefulness in supporting high impact data movement, there are costs associated with implementing and supporting them. There will be a tangible cost for the underlying network infrastructure. The infrastructure cost may initially be part of a centrally-funded project to advance the state of network technology, but eventually cost recovery processes get invoked. Beyond the hard costs for underlying network infrastructure and associated data services, there are significant effort costs associated with E2E circuits. Given the current state of the technology, establishment of a circuit is manual and can require extensive debugging effort. The circuits add complexity in terms of management and operational support, particularly when the circuit crosses a number of administrative domains. Proper documentation can require significant effort, both initially and on an ongoing basis. Perhaps an even greater concern is that the documentation gets done inadequately or not at all, leaving support personnel in a very difficult position when a circuit isn't working properly. Above all, basic network support activities of monitoring, troubleshooting, and understanding failure modes are much more complex than with a conventional routed IP service.

4.1. Monitoring

Monitoring of E2E circuits requires significantly more effort than simply relying on the general IP routed infrastructure. By definition, the scope of monitoring is extended to covering the end-to-end path. The end-to-end path is typically made up of a concatenation of layer-2 segments, often traversing multiple different administrative domains. Existing tools designed for monitoring a network path across the routed IP infrastructure work at layer-3, and will only provide monitoring information on the circuit end points. As a result, commonly used network monitoring tools are largely ineffective for monitoring an E2E circuit.

A new monitoring infrastructure, called PerfSonar [3], is being developed to facilitate network performance monitoring on paths that cross multiple network administrative domains. Functionally, PerfSonar compartmentalizes the network monitoring to individual network domains, so that each domain is responsible for its own data collection and segment monitoring. Aggregation of the monitoring data collected within each network domain then provides the end-to-end perspective. For monitoring E2E circuits, PerfSonar requires deployment, configuration, and data collection by a monitoring system, referred to as a Measurement Point (MP), within each network domain along the circuit's path. Currently, PerfSonar has been implemented across the Laboratory's LHCOPN circuit (Figure 5). There are four network domains, each monitoring its own subset of the end-to-end path.

Status of E2E Link CERN-FERMI-LHCOPN-001

Oper. State: Up
Admin. State: Normal Oper.

Domain	CERN		USLHCNET					ESNET					FERMI			
Link Structure	EP	◄·····	·····►	DP	◄——►	DP	◄·····	·····►	DP	◄——►	DP	◄·····	·····►	DP	◄——►	EP
Type	EndPoint	ID Part.Info	ID Part.Info	Demarc	Domain Link	Demarc	ID Part.Info	ID Part.Info	Demarc	Domain Link	Demarc	ID Part.Info	ID Part.Info	Demarc	Domain Link	EndPoint
Local Name	CERN-T0	5513-C-BE1	CERN-FERMI-LHCOPN-001-GVA-CERN	USLHCNET-GEN	CERN-FERMI-LHCOPN-001-GVA-CHI	USLHCNET-CHI	CERN-FERMI-LHCOPN-001-CHI-ESNET	CERN-FERMI-LHCOPN-001-STARLIGHT-Tail	ESNET-STARLIGHT	CERN-FERMI-LHCOPN-001-FERMI-STARLIGHT	ESNET-FERMI	CERN-FERMI-LHCOPN-001-Site-Tail	md8	FERMI-ESNET	md2	FERMI-T1
State Oper.	-	Up	Up	-	Up	-	Up	Up	-	Up	-	Up	Up	-	Up	-
State Admin.	-	Normal Oper.	Normal Oper.	-	Normal Oper.	-	Normal Oper.	Normal Oper.	-	Normal Oper.	-	Normal Oper.	Normal Oper.	-	Normal Oper.	-
Timestamp	-	2007-01-26 T13:15:22 +01:00	2007-02-08 T17:31:23 +01:00	-	2007-02-10T01:17:03 +01:00	-	2007-01-26 T13:15:19 +01:00	2007-02-10 T00:15:43.0	-	2007-02-10T00:15:43.0	-	2007-02-10 T00:15:43.0	2007-02-09 T17:00:01.0-6:00	-	2007-02-09T17:00:01.0-6:00	-

Page generated at 2007-02-10, 01:17:54 MET

Figure 5: PerfSonar Monitoring of LHCOPN E2E Circuit

There are several limitations with the current level of PerfSonar monitoring on E2E circuits. First, the monitoring tool is in its early stages of development, and its functionality is limited. For example, the LHCOPN monitoring is simply checking on the interface status of the demark links between the network administrative boundaries. This does not provide a true indication that the end-to-end path is up and completely functional. Instead, it indicates when individual problem segments are detected, and assumes that an absence of any detected problem means that the circuit is operating normally. Further development of PerfSonar is expected to improve end-to-end monitoring capabilities.

The limited deployment of PerfSonar MPs also limits the amount of E2E circuit monitoring currently being done. The LHCOPN circuit is the only E2E circuit with full PerfSonar monitoring implemented. As PerfSonar MPs get deployed at other sites and transit networks, end-to-end PerfSonar monitoring of the other E2E circuits is expected to be implemented.

4.2. Troubleshooting

As with monitoring, troubleshooting is more complicated in an E2E circuit environment, due to greater complexity, multiple network administrative domains, and limited tool sets for layer-2 troubleshooting. PerfSonar is being enhanced to support a number of active monitoring tools, which should facilitate faster and more efficient troubleshooting, when available. At the current time, however, troubleshooting remains very much a manual, effort-intensive process.

4.3. Failover Issues

E2E circuits will fail. At Fermilab, the expected result of E2E circuit failure is that the circuit traffic would fail over to the general routed IP network. The path of failure will be the 10Gb/s connection between the E2E circuit router to the facility border router (see Figure 2). Since the routed IP path is also via a 10Gb/s channel, no throttling of throughput on the circuit-based traffic has been put in place. However, the capability to do so exists, and would be relatively simple to implement.

The other concern about failover situations is making sure the consequences of an E2E circuit failure are the intended ones. Analysis of potential failure modes and their consequences is done on each E2E circuit, and failure modes are tested as opportunities to do so present themselves. Note that there are typically multiple failure modes for any particular E2E circuit. An ethernet segment failure may provide a different result than a failure by the router supporting the E2E circuit. Using scheduled network maintenance outages to monitor and verify failover modes is our preferred practice.

4.4. Configuration Changes Affecting PBR

One of the consequences of using PBR to implement E2E circuit routing is that it's relatively easy for routing configurations to get changed. Asymmetric paths may result if the routing configuration is not the same on both ends of the circuit. An erroneous or misapplied PBR access list may result in unexpected rerouting of circuit traffic back to the general IP routed path. Worse, any changes in routing may not be obvious. In order to preserve path symmetry, we have begun analyzing network flow data patterns for clear inconsistencies. If the number of flows in one direction of an E2E circuit is significantly skewed from the number in the opposite direction, we flag the circuit for possible routing asymmetry, and investigate. Currently, this detection and subsequent analysis is manual, but we expect to have the flow analysis automated in the near future.

E2E circuit routing may also be indirectly impacted by configuration changes on end systems. Figure 6 (following page) shows what can occur when end system address blocks are changed. In this instance, CMS traffic between the Tier-1 Center and a Tier-2 site (University of Nebraska, Lincoln) was being routed via an E2E circuit. The Tier-2 site readdressed its systems into a different address block, rendering the PBR source/destination address block configuration for the circuit incorrect.

Since the high impact data traffic no longer matched up with the PBR configuration, traffic reverted to traversing the general routed IP path. The Laboratory's general use offsite network link, normally sustaining 1-2Gb/s, was suddenly supporting upwards of 8Gb/s and encountering a modest level of link congestion. The problem was discovered and corrected within a few hours by modifying the PBR configuration to add the new address block in use at the Tier-2 site. The instance highlights the type of complications that may result when changes are made that affect system addressing or network routing. As with asymmetric paths, analysis network flow data may provide automated means to detect, and even correct unexpected routing changes that involved E2E circuits.

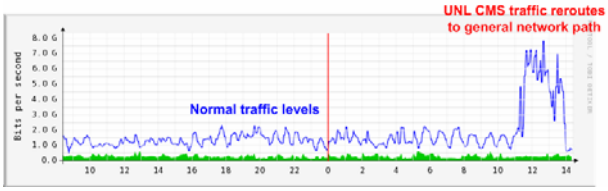

Figure 6: Rerouted E2E Circuit Traffic

5. Future Directions

The E2E circuits currently supported at the Laboratory are all statically configured. They also remain in place continuously. This is manageable, given the current availability of alternate network paths and the level of demand for those paths. However, using static E2E circuits will not scale well, as demand grows. Leaving a circuit up all the time may be inefficient and costly, particularly if the circuit isn't in use continuously.

There is considerable research and experimentation within the R&E community on dynamically-provisioned data circuits. Our intention is to track the development of dynamic data circuit technology, and support dynamic end-to-end circuits as the technology matures and the demand for that type of service emerges. To that end, we have developed the capability to dynamically implement routing changes within our local network infrastructure through the Lambda Station project [4]. A Lambda Station server will accept alternate path service requests from users and applications, coordinate with wide area networks for setup of dynamic circuit services, and make the necessary PBR reconfigurations to enable use of the wide area dynamic circuits by locally-connected facilities. A major long term goal is to work closely with wide area network service providers so that as dynamic circuit services mature, we will be able to make use of them through local services such as Lambda Station.

References

[1] http://www.es.net/hypertext/welcome/pr/Roadmap/index.html
[2] http://www.cisco.com/warp/public/732/Tech/plicy_wp.htm
[3] http://www.perfsonar.net/
[4] http://www.lambdastation.org/