Use of Alternate Path WAN Circuits at Fermilab

Phil DeMar / Matt Crawford

CHEP 2007

Office of Science
U.S. DEPARTMENT OF ENERGY

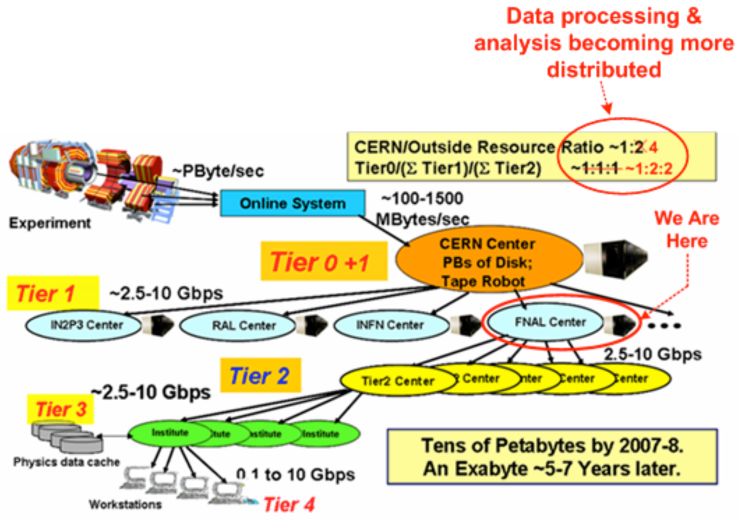
Fermilab

Why end-to-end circuits?

- Convergence of need, capability, & strategic direction
- Sometimes just because our stakeholders ask for them
  - They anticipate better WAN performance with circuits

Need

- Emerging CMS high impact data movement requirements
- Predictable network performance requirements:
  - ❑ Distributed DAQ function
  - ❑ Distributed analysis model
- Data movement thru CMS Tier structure is flexible, not hierarchical
  - ❑ Significant trans-oceanic traffic
- LHC traffic projections call for rapid increase in traffic levels


Data processing & analysis becoming more distributed

We Are Here

Tens of Petabytes by 2007-8.
An Exabyte ~5-7 Years later.

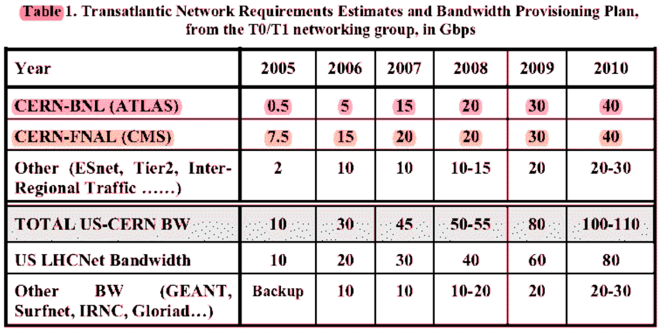


Table 1. Transatlantic Network Requirements Estimates and Bandwidth Provisioning Plan, from the T0/T1 networking group, in Gbps

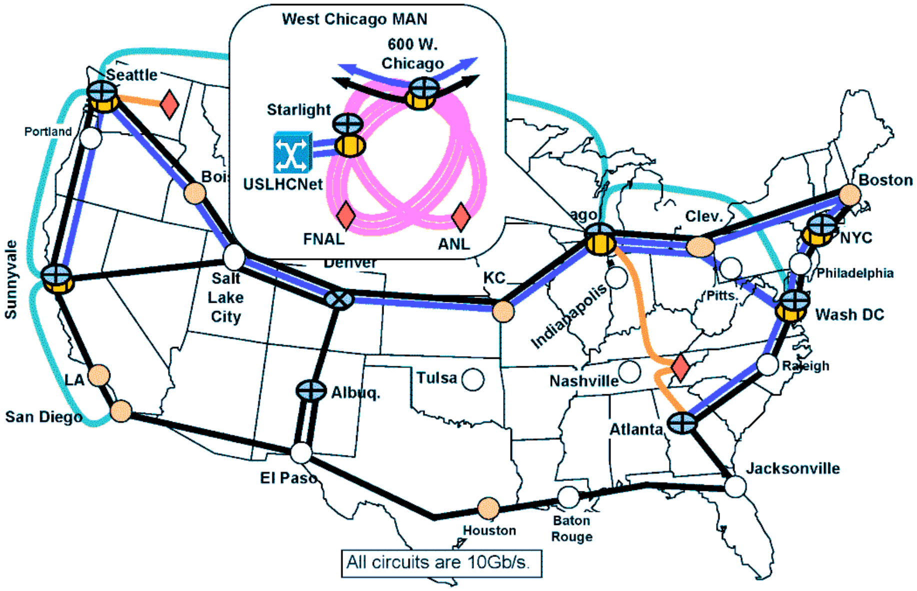
Year	2005	2006	2007	2008	2009	2010
CERN-BNL (ATLAS)	0.5	5	15	20	30	40
CERN-FNAL (CMS)	7.5	15	20	20	30	40
Other (ESnet, Tier2, Inter-Regional Traffic ……)	2	10	10	10-15	20	20-30
TOTAL US-CERN BW	10	30	45	50-55	80	100-110
US LHCNet Bandwidth	10	20	30	40	60	80
Other BW (GEANT, Surfnet, IRNC, Gloriad…)	Backup	10	10	10-20	20	20-30

Capability

- Fermi LightPath:
  - Optical network infrastructure between FNAL & StarLight:
    - Leased dark fiber
    - Dense Wave Division Multiplexing equipment (Ciena Metro)
  - Initial (2004) configuration: 1x10GE & 2x1GE channels
  - Current configuration: 6x10GE & 2x1GE channels
- Direct fiber to StarLight provides a plethora of network connectivity opportunities
  - Wide spectrum of possible peering partners available
  - L2 technology options become available (L1 someday?)
- Optical network infrastructure offers flexible, economical upgrade options

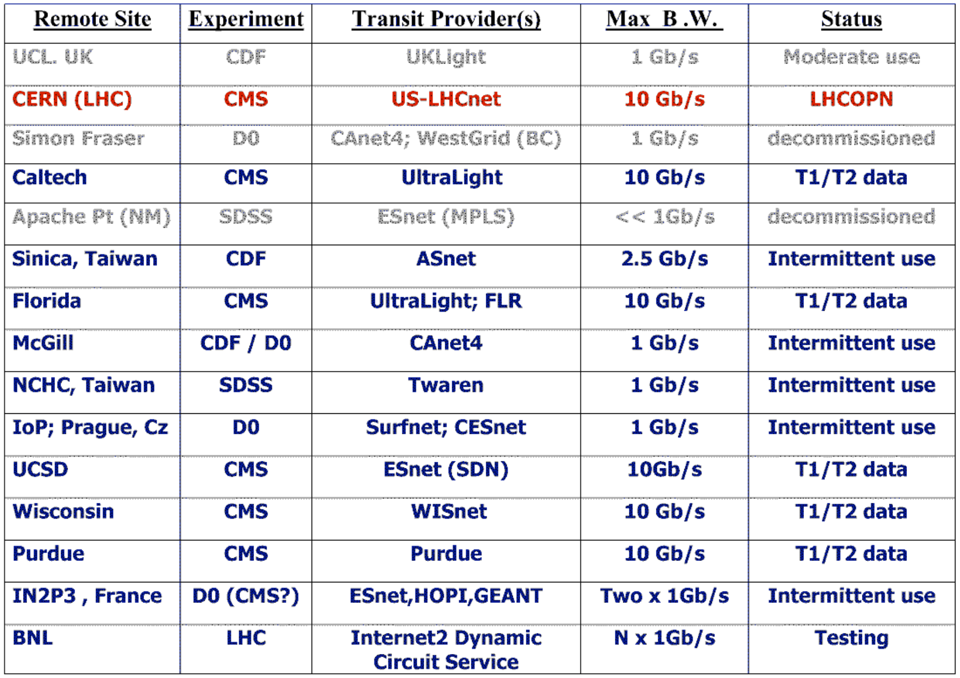
Strategic Direction

- DOE High Performance Network Planning Workshop established a strategic model to follow:
  - High bandwidth backbones for reliable production IP service
    - ESnet
  - Separate high-bandwidth network paths for large scale science data flows
    - Science Data Network
  - Metropolitan Area Networks (MAN) for local access
    - Fermi LightPath a cornerstone for Chicago area MAN



FNAL Alternate Path Circuits

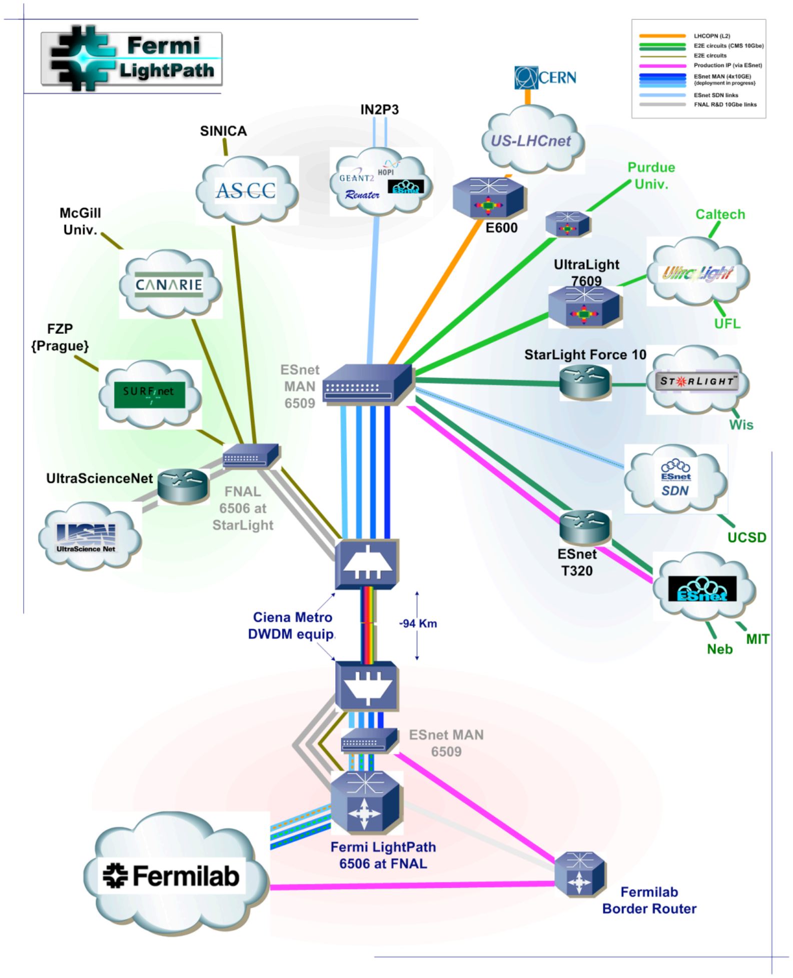
- Supported since 2004
- Serve a wide spectrum of experiments
  - CMS Tier-2s are heavy users
- Implemented on multiple technologies
  - But based on end-to-end layer-2 paths
- Usefulness has varied



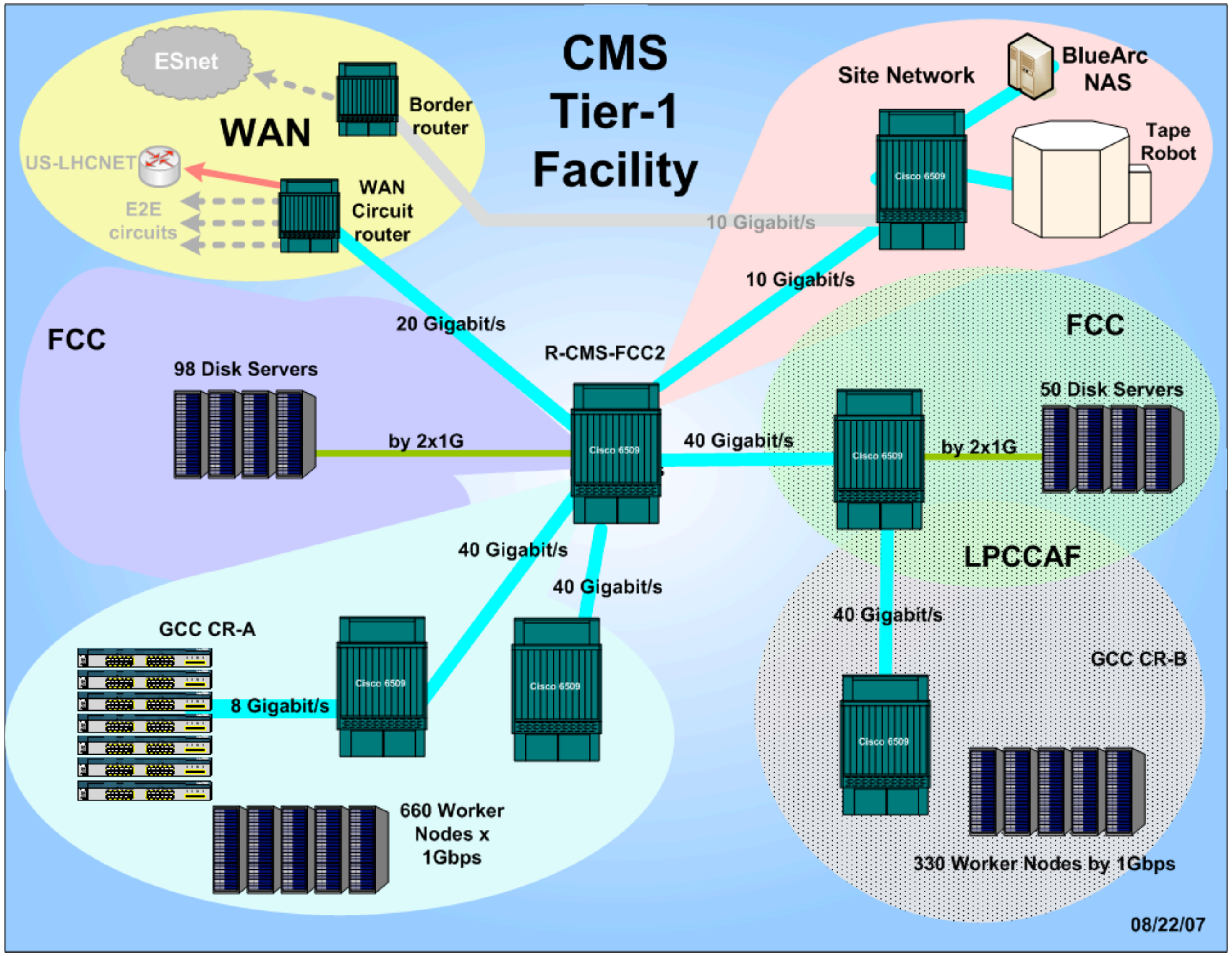
Remote Site	Experiment	Transit Provider(s)	Max B.W.	Status
UCL. UK	CDF	UKLight	1 Gb/s	Moderate use
CERN (LHC)	CMS	US-LHCnet	10 Gb/s	LHCOPN
Simon Fraser	D0	CAnet4; WestGrid (BC)	1 Gb/s	decommissioned
Caltech	CMS	UltraLight	10 Gb/s	T1/T2 data
Apache Pt (NM)	SDSS	ESnet (MPLS)	<< 1Gb/s	decommissioned
Sinica, 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

Topology of circuit connections

- Circuits utilize MAN infrastructure:
  - One 10GE channel reserved for routed IP service (purple)
  - One supports LHCOPN circuit (orange) to CERN
  - Two support end-to-end circuits to CMS Tier-2 (shades of green)
- Circuits based on end-to-end vLANs
  - Direct BGP peering with remote site
- Multiple provider domains is the norm
  - Deployed technology varies by domains involved
  - Complexity is higher than IP service

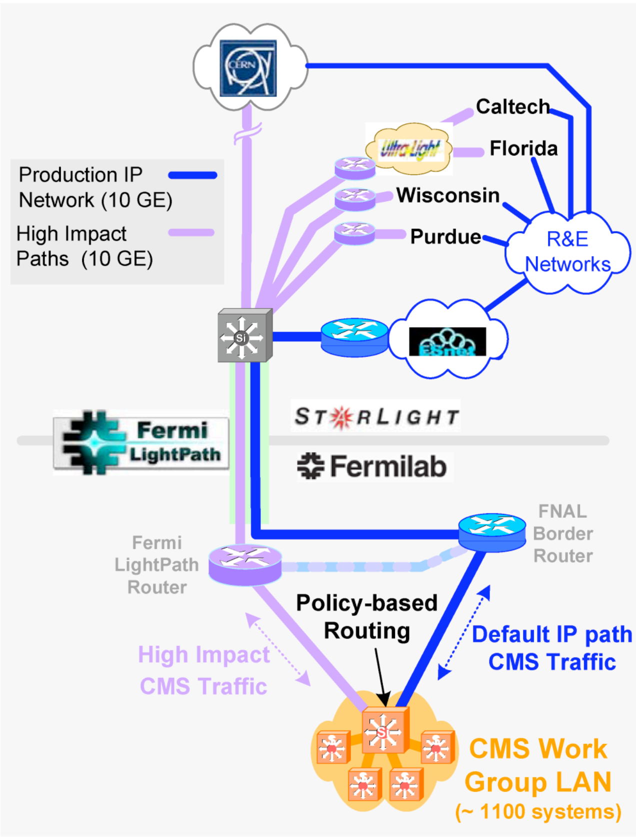


Internal US-CMS Tier-1 LAN



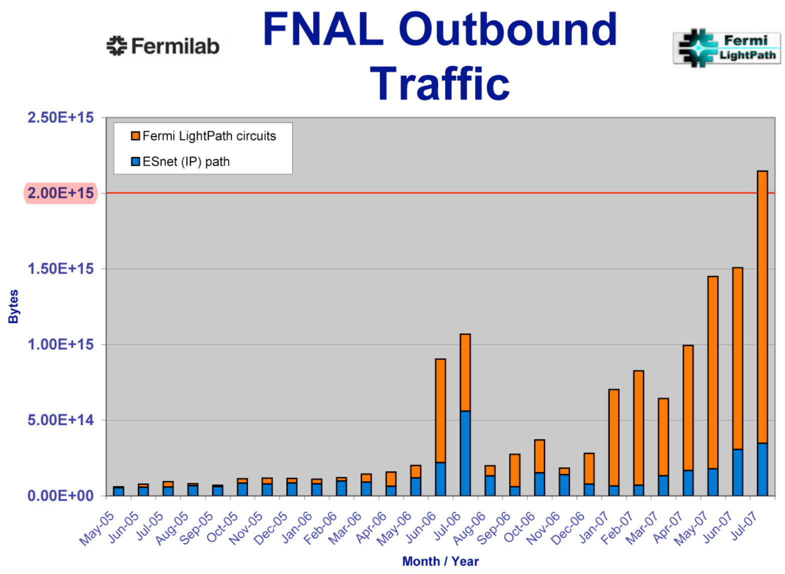
Making the E2E circuit routing work

- Define high impact traffic flows:
  - Minimal-size source/dest. netblock pairs
    - US-CMS Tier-1 / CERN T-0 address pairs follow LHCOPN circuit path (purple)
    - Other FNAL-CERN traffic on routed path (blue)
- Establish E2E circuits on alternate path border router
  - BGP peer across VLAN-based circuits, advertising only source netblock
- Policy route internally on source/dest pairs
- Inbound routing depends on policies of remote end
  - Prefer comparable PBR for symmetry
  - But implement inbound PBR locally



Usefulness of E2E Circuits

- Monthly FNAL outbound traffic
- Recent spikes exclusively due to CMS ramp-up testing
  - ❑ Supports CMS traffic projections
  - ❑ Traffic levels indicate performance capabilities, not trend



- Relative ratio of circuit-based traffic to routed traffic is also more an indication of performance capability
  - ❑ US Tier-2s (circuit-based) routinely sustain 2-3 Gb/s and higher
  - ❑ In CSA06 European T2s (routed) were sustaining 100Mb/s-900Mb/s

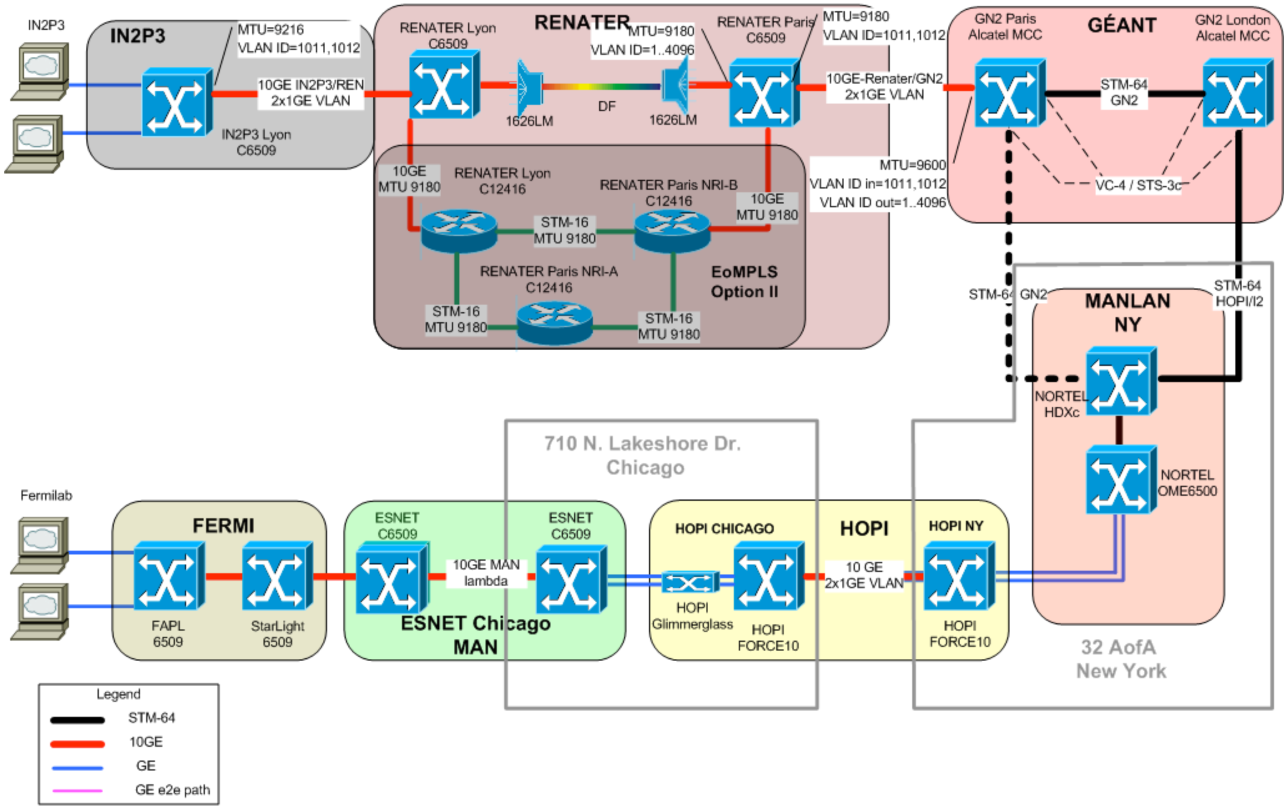
Issues with E2E circuits

- Circuit coordination & establishment can be complex
  - Varies with # of administrative domains and mix of underlying technology
- Monitoring becomes more difficult
- Troubleshooting problems are more difficult, too
  - Likely to be needed more frequently as well
- Failure modes need to be understood and tested
- Proper documentation can be a lot of work
  - Or doesn't get adequately done…

An example of circuit complexity

- IN2P3/FNAL test circuit:
  - ❑ Four service providers
  - ❑ Technology mix
  - ❑ ~2 months to get configured
  - ❑ Monitoring still not complete
  - ❑ Circuit documentation is sparse

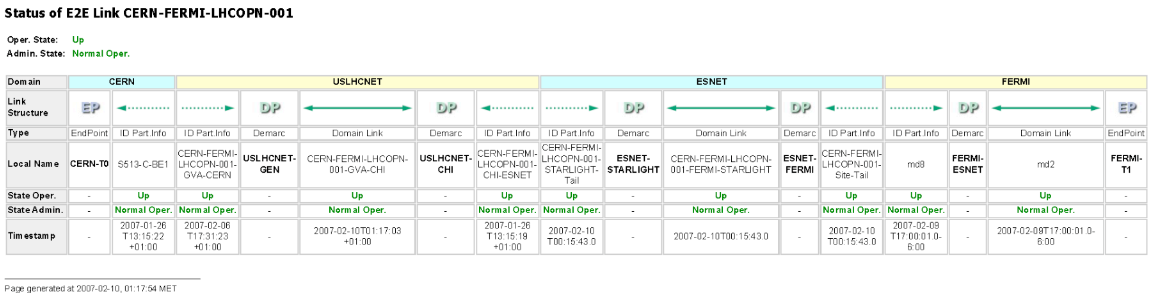
IN2P3-FERMILab 2xGE
Path setup



Monitoring E2E circuits

- Complicated by multi-domain boundaries and layer-2 technology
- PerfSONAR emerging as cross-domain data collection monitoring tool
  - A work-in-progress at this point
  - Minimal level of monitoring capabilities currently available
    - interface status…
  - Active monitoring capabilities being worked on

- PerfSonar currently deployed for LHCOPN E2E circuit monitoring



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**	S513-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	2007-02-06 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	-

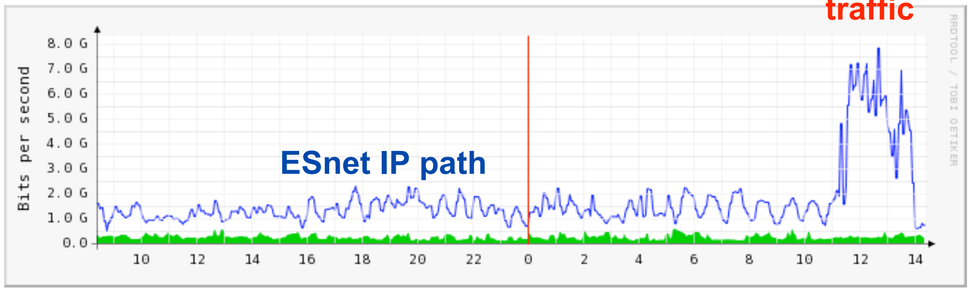
Page generated at 2007-02-10, 01:17:54 MET

Operational experiences with circuits

- E2E circuit failure modes are different than for IP service
  - They are more complex
  - Impact of the failure may be severely felt elsewhere
  - Operational failures can be “creative” and difficult to troubleshoot

- Asymmetric paths will occur and will be difficult to detect
  - We’re working on flow data analysis to detect this

- Unexpected consequences of changes
  - UNL moves several T2 systems to a new subnet



UNL CMS traffic

ESnet IP path

Performance Analysis Methodology

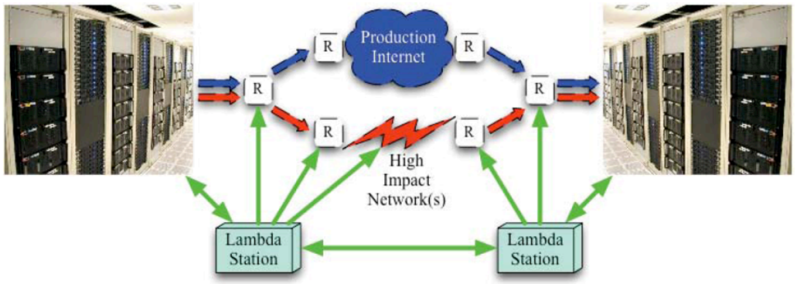
- Problem diagnosis more difficult at layer-2
- Developing structured approach to troubleshooting
- Model for the process is medical diagnosis
  - Collect the physical characteristics
  - Run diagnostic tests
  - Record everything; develop a history of the analysis
- Strategic approach:
  - Sub-divide problem space:
    - Application-related problems
    - End system diagnosis and tuning
    - Network path analysis
  - Then divide and conquer

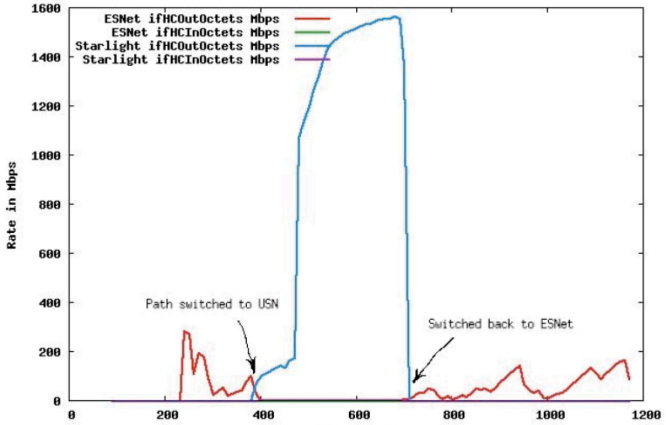
Steps in Performance Analysis

- Definition of the problem space
- Collection of system information & network path characteristics
- Host configuration analysis
- Network path performance analysis
  - Current base tools: NDT & OWAMP
- Evaluate packet flow patterns

Dynamic Circuits on the Horizon

- Dynamic path-selection services under development
  - Lambda Station (FNAL), Terapaths (BNL)
- Lambda Station (LS) project:
  - Based on PBR mechanisms
  - LS called by apps or wrapper scripts
  - Schedules reservable network paths
  - Configures selective rerouting into LAN
  - Only configures local site infrastructure
  - Coordinates with LS on remote end
  - Deployed within Tier-1 SRM service





Winding It Up…

- End-to-end circuits have proven to be useful at FNAL
  - Especially for LHC/CMS high impact data movement
  - In some cases, useful for other experiments & projects as well
- Additional management & support cost involved
  - Complexity is an obvious concern
  - Scalability too…
- We will see a natural selection process play out
  - What works & is worth the effort will remain and grow
  - What doesn’t prove to be worth the effort will disappear
- When will dynamic end-to-end circuits be widely available?
  - The crystal ball is a little cloudy…