Wide Area Network Project Briefing

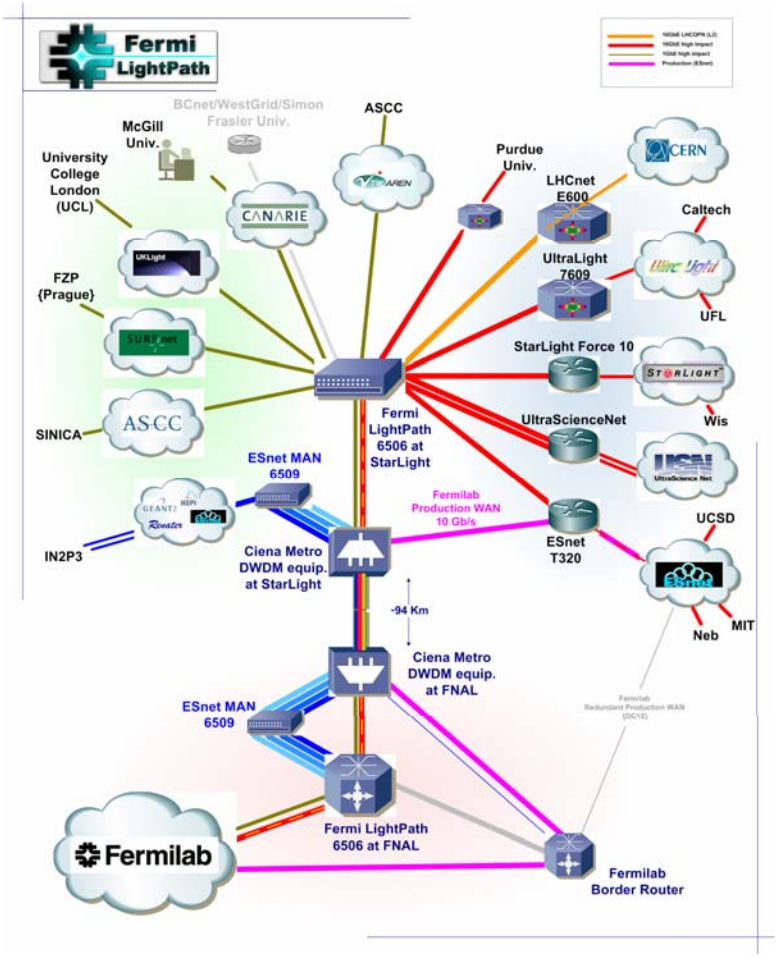
Phil DeMar

January 23, 2007

WAN Infrastructure Support

Current WAN configuration – Off-Site Connectivity

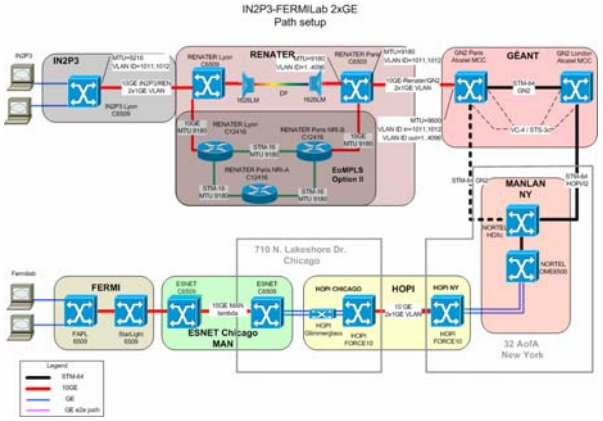
- All traffic now carried on Fermi LightPath (StarLight fiber):
  - Production IP (ESnet) path carried on one 10GE channel
  - 2nd 10GE channel carries:
    - CMS Tier0/Tier1 traffic (CERN)
    - US-CMS Tier1/Tier2 traffic
    - Misc CDF, D0, R&D traffic
  - ESnet OC12 still serving a (tertiary) redundancy role
- Four ESnet MAN 10GE channels now available for use:
  - Planning to migrate LHCOPN traffic to one of these channels soon



Current WAN configuration – End-to-end (E2E) circuits

- 14 E2E circuits currently supported
  - CMS LHCOPN & US-CMS Tier1 / Tier2 circuits heavily utilized
  - Other circuits lightly to moderately utilized
  - WestGrid circuit decommissioned due to acceptable performance across routed (ESnet) path
- Newest circuit(s) to IN2P3 reflect complexity of multi-domain E2E links
  - Monitoring & troubleshooting are especially difficult
- E2E circuit to FZK (Karlsruhe) planned for CMS Tier1/Tier1 path evaluations

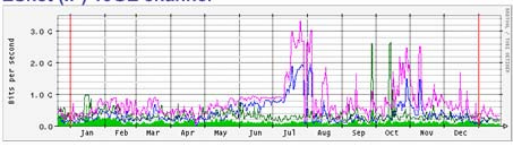
Remote Site	Experiment	Transit Provider(s)	Max B.W.	Status
UCL	CDF	UKLight	1 Gb/s	Moderate use
CERN	CMS	US-LHCnet	10 Gb/s	LHCOPN
Caltech	R & D	UltraScience Network	1 & 10 Gb/s	Intermittent use
Simon Fraser	D0	CAnet4; WestGrid (BC)	1 Gb/s	decommissioned
Caltech	CMS	UltraLight	10 Gb/s	T1/T2 data
Sinica, 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; ?	1 Gb/s	Intermittent use
UCSD	CMS	ESnet (MPLS & 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 {new}	D0 (CMS?)	ESnet,HOPI,GEANT	Two 1Gb/s	#1 – Intermittent use #2 - pending
FZK {new}	CMS	US-LHCnet, DFN	1-10Gb/s	Pending



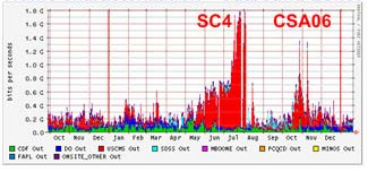
WAN Traffic – Utilization Graphs

- ESnet 10GE channel utilization remains ‘comfortably’ utilized:
  - Peak outbound traffic periods from non-US CMS-Tier2 sites
  - No indication of sharply increasing traffic levels from the Run-II or other experiments
  - Inbound traffic levels remain quite low
- High impact 10GE channel (principally CMS) has periods of heavily utilization:
  - A few instances of traffic congestion detected:
    - Don’t quite correspond to peak utilization periods
    - Need to investigate further
    - High impact traffic to be split across three 10GE channels with deployment of ESnet MAN

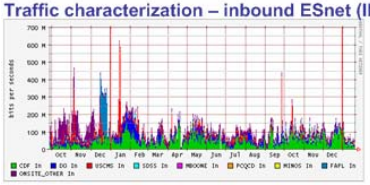
ESnet (IP) 10GE channel



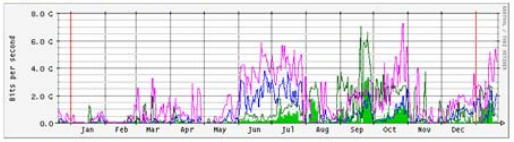
Traffic characterization – outbound ESnet (IP) path



Traffic characterization – inbound ESnet (IP) path



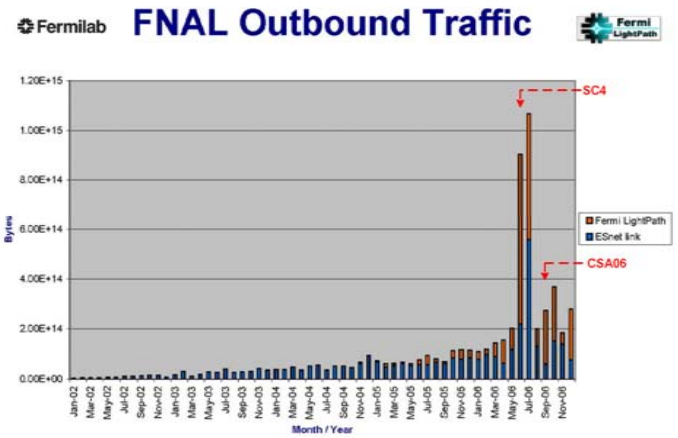
High Impact 10GE channel

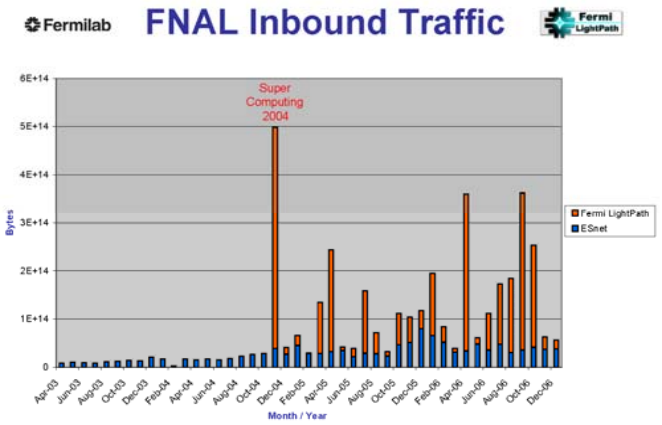


Discards - High Impact 10GE channel

WAN Traffic – Long Term Trends:

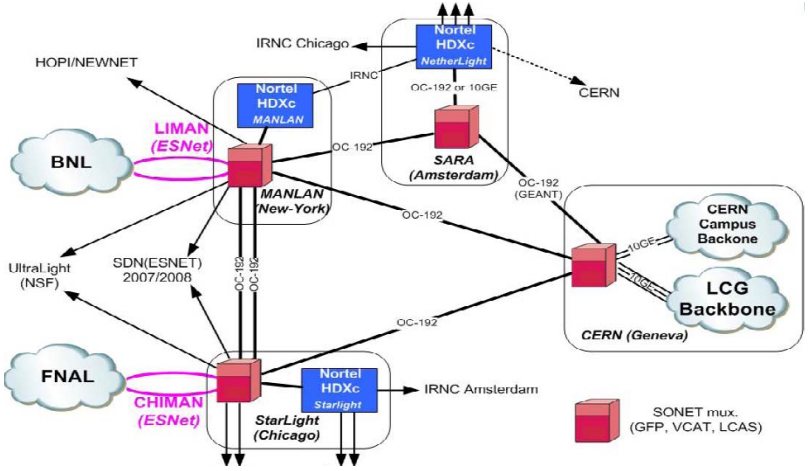
- Outbound traffic:
  - CMS ramp-up clearly causing sharply increasing (exponential?) traffic levels:
    - SC4 peak remains more a demonstration of capability than anything else
    - But Jan numbers should be up
  - High impact data paths account for a majority & increasing % of traffic
- Inbound:
  - Tier0/Tier1 traffic (CERN) responsible for most of the inbound traffic
    - LHCOPN connection on high impact 10GE
    - No apparent significant trend with inbound routed traffic levels





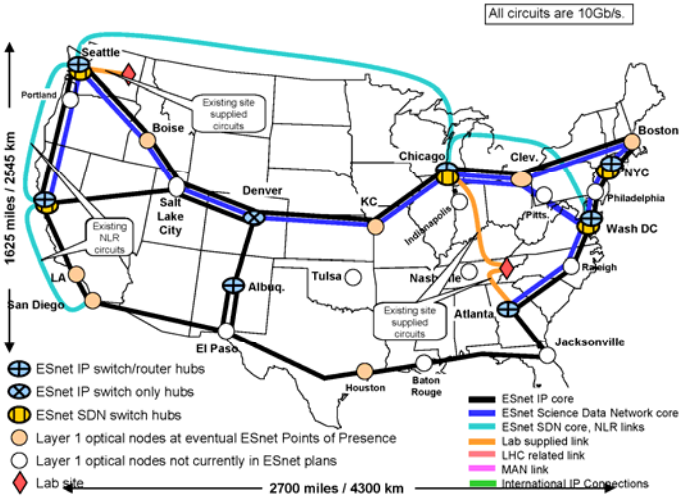
Service Providers – USLHCnet

- Provides trans-Atlantic network service for US LHC collaboration
  - 10GE links interconnecting StarLight/FNAL, ManLan/BNL, & CERN
  - Supports US part of LHCOPN, a virtual network for moving raw data from CERN to Tier1 centers
  - Secondary role of supporting general HEP trans-Atlantic network needs
- Being upgraded to SONET-based infrastructure
  - Capability to channelize 10GE links for different classes of service
  - Raises concerns about mid-path bottlenecks
  - May be used to provide bandwidth for LHC trans-Atlantic Tier1/Tier2 data movement



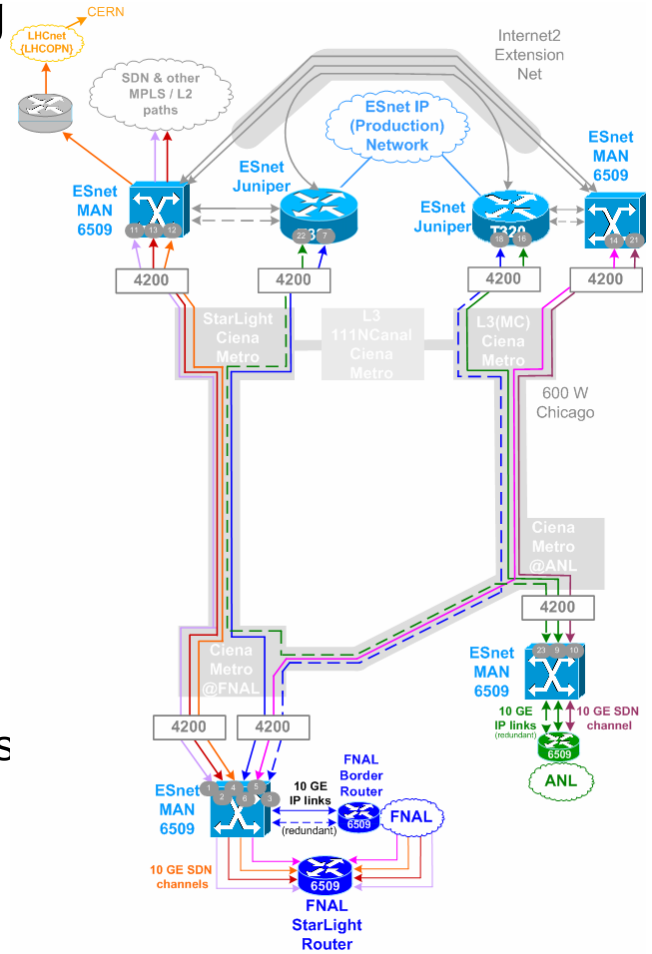
Service Providers - ESnet

- Internet2 has partnered with Level3 for its next generation network
  - Direct access to & control over dedicated optical fiber infrastructure
  - Deploying inexpensive DWDM equipment
- ESnet has contracted with Internet2 for its next backbone (ESnet4)
  - Circuit-oriented network services in addition to raw bandwidth
  - Deployment to start in March
- ESnet4:
  - All backbone links to be 10GE
  - Major backbone sites homed to two PoPs
    - StarLight & Level3 PoP in Chicago…
  - Two classes of backbone circuits:
    - Classic IP service
    - Science Data Network for high impact data movement



ESnet MAN – A Configuration Change:

- Original MAN configuration = three node ring
  - FNAL, ANL, & StarLight
    - with spur to ESnet Chicago (Qwest) PoP
  - But ESnet4 changes ESnet backbone service provider from Qwest to Level3
    - Transition to ESnet4 starts in March
- New ESnet Chicago MAN topology will use Level3 PoP as a 4th node on MAN ring
- FNAL channel config is (largely) unchanged:
  - Two 10GE IP paths, one to StarLight & one to new ESnet Chicago PoP
    - Both paths utilized, depending on routing metrics
  - Four SDN channels for high impact data movement
    - One SDN channel will terminate at Level3 PoP

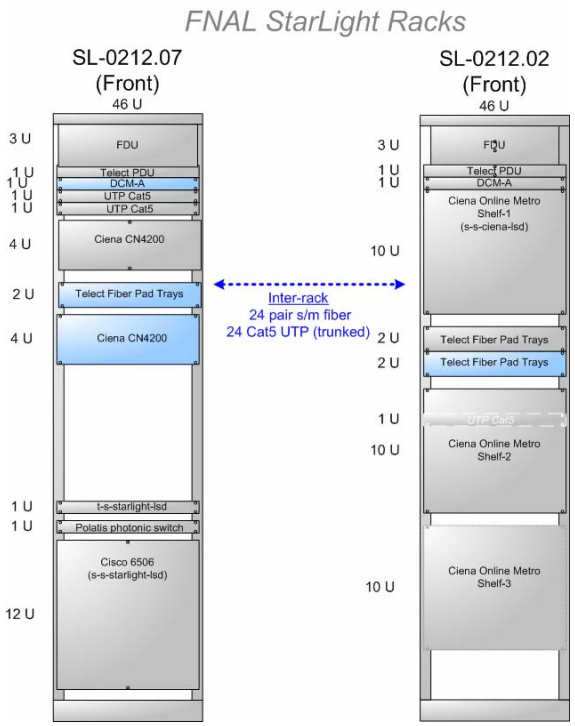


ESnet MAN – Impact of Configuration Change

- Additional Ciena Metro (DWDM) equipment is required:
  - Not funded in initial work proposal
  - Not quickly forthcoming under continuing resolution
- Fiber to the new Level3 PoP is required:
  - Available, but needs to be cut & re-spliced at several locations
- Co-location space is needed within the Level3 PoP
  - New facility just being brought into production; racks not available
  - Contractual issues need to be sorted out
- Implementation plan being modified for initial MAN deployment at FNAL, ANL, & StarLight only
  - Level3 PoP to be incorporated when above issues are sorted out

ESnet MAN Pre-deployment Work

- ANL / FNAL fiber segment is finally in place
  - But apparently won't be characterized or dispersion compensated
- Rack reorganization at StarLight completed:
  - Third Ciena Metro shelf installed
  - Power work & Ciena 4200 chassis to be installed this week
- ESnet MAN 6509s at at both ends are online
- Four 10GE channels are ready for use (using the SC2005 Ciena equip)
  - One channel in use for IN2P3 E2E circuits
  - Will migrate LHCOPN channel over soon
- Out of band management capability deployed:
  - Via both NWU StarLight LAN & console dial-in



ESnet MAN Partial Deployment Schedule

- Tentatively scheduled for week of 2/25
  - FNAL, ANL & StarLight Ciena metro equipment will be upgraded
    - Will be a hardware & software upgrade
      - Includes new management capability
    - Channel reconfiguration with provide some path redundancy
      - But inclusion of Level3 PoP needed to provide intended level of redundancy
  - At least one work day outage of the MAN channels is anticipated
    - We'll be operating on the OC12 during the outage period
    - Looking into contingency connection to ANL ESnet router for redundancy for that day

ESnet MAN – Post Deployment Issues

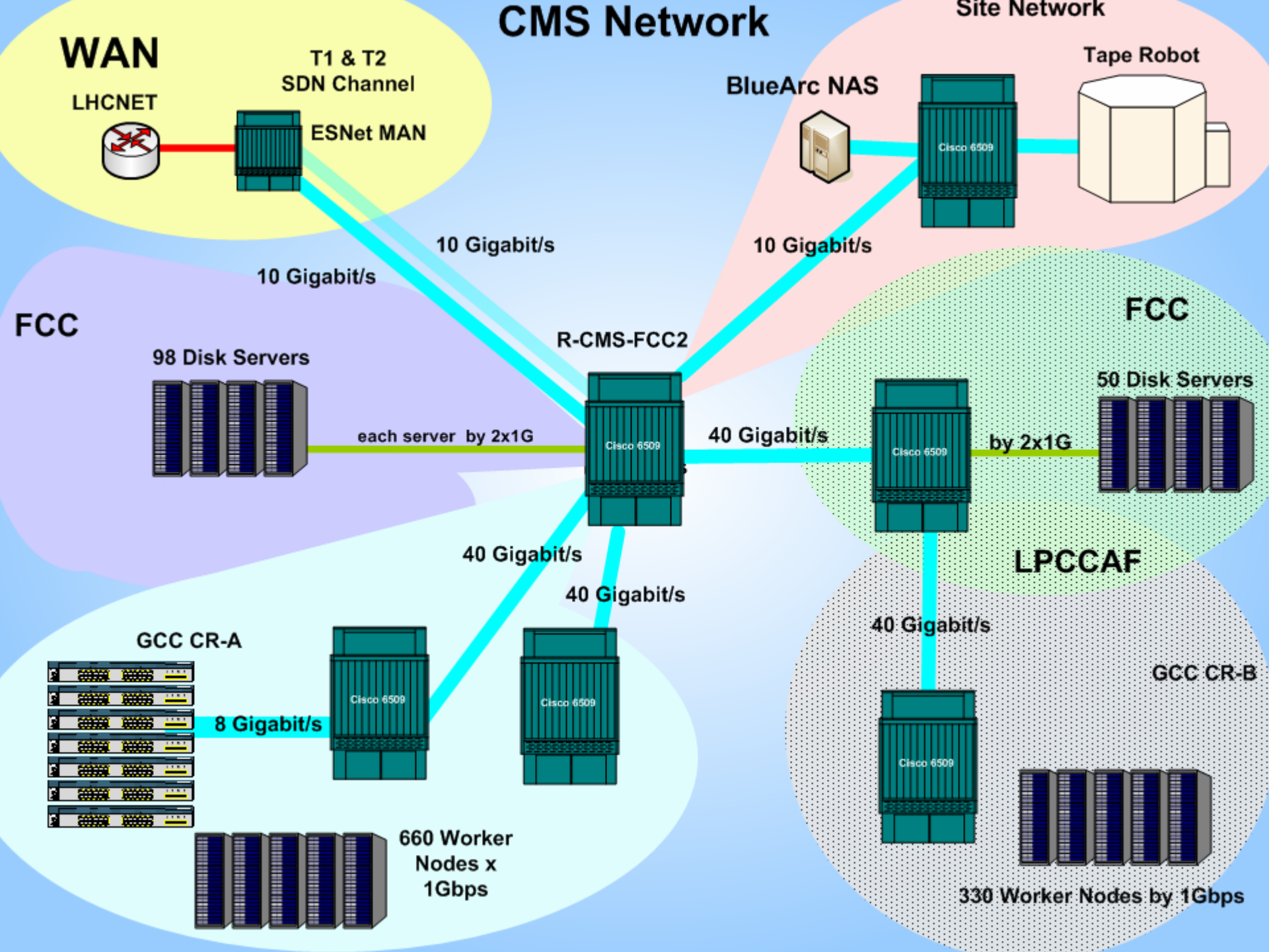
- Migration of existing E2E circuits will need to coordinated & facilitated on a one-by-one basis
- Ongoing costs & maintenance agreement with ESnet needs to be put in place
- Inter-governmental agreement with Batavia on use of their fiber still needs to be pursued
- Tentative channel allocation model needs to be finalized with ESnet:
  - Delayed deployment of Level3 PoP limits level of redundancy
  - IP #2 & SDN #4 temporarily homed to StarLight

Channel	Use
IP #1	StarLight PoP IP
IP #2	Level3 PoP IP
SDN #1	LHCOPN
SDN #2	US-CMS Tier2s
SDN #3	Run-II, others
SDN #4	Redundant SDN

FNAL ANL Network Group

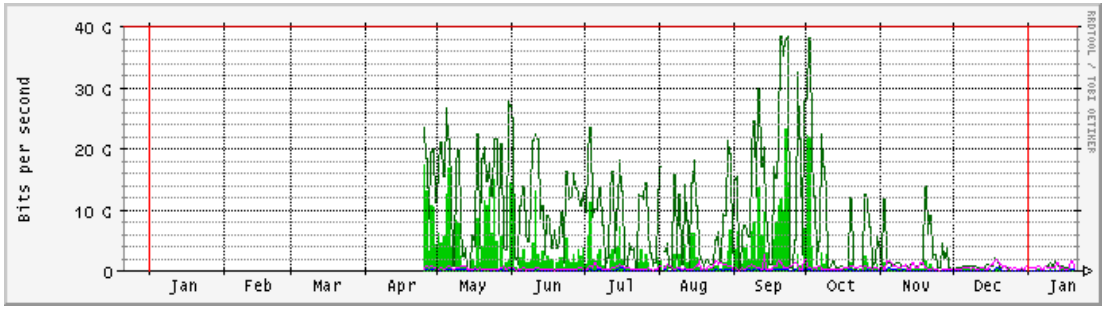
- ANL / FNAL MOU provides joint operations framework
  - FNAL ANL network group (FANG) to be organized to facilitate that…
- FANG primary focus = developing a joint operations support model
  - Deployment of common monitoring capabilities
  - Coordination of on-call procedures & local maintenance operations
  - Development of a common triage process for troubleshooting
  - Definition of a common problem reporting procedures
  - Establishment of common sparing plan
  - Specification & generation of common documentation
- FANG secondary focus = strategic and tactical planning for the MAN
  - Channel allocation and channel resource planning & coordination
  - Topology change planning & implementation
  - Equipment refresh strategies & plans
  - Deployment strategies for new technology
  - Review of provisioning agreements

CMS Tier-1 Support

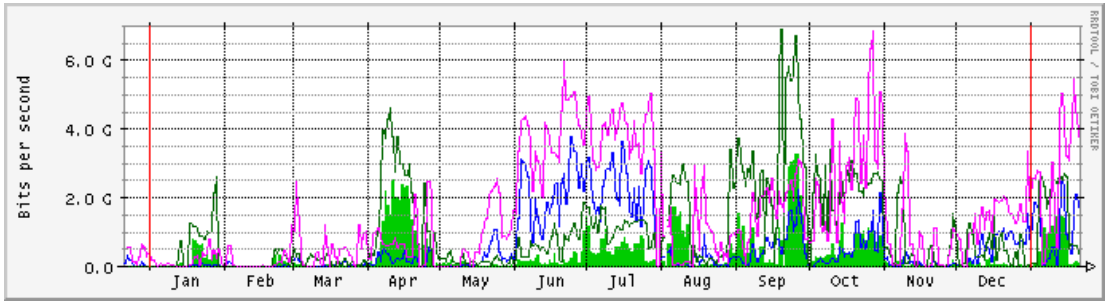

CMS Network
WAN
LHCNET
T1 & T2
SDN Channel
ESNet MAN
Site Network
BlueArc NAS
Tape Robot
Cisco 6509
10 Gigabit/s
10 Gigabit/s
10 Gigabit/s
FCC
98 Disk Servers
each server by 2x1G
R-CMS-FCC2
Cisco 6509
40 Gigabit/s
FCC
50 Disk Servers
by 2x1G
LPCCAF
40 Gigabit/s
40 Gigabit/s
40 Gigabit/s
GCC CR-A
8 Gigabit/s
660 Worker
Nodes x
1Gbps
GCC CR-B
330 Worker Nodes by 1Gbps

Traffic utilization within CMS Tier1 Facility

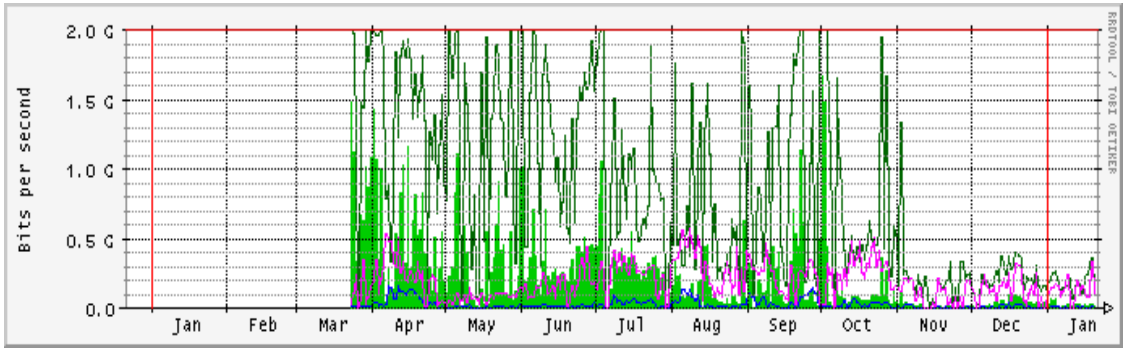
A ~40Gbps onsite rate:



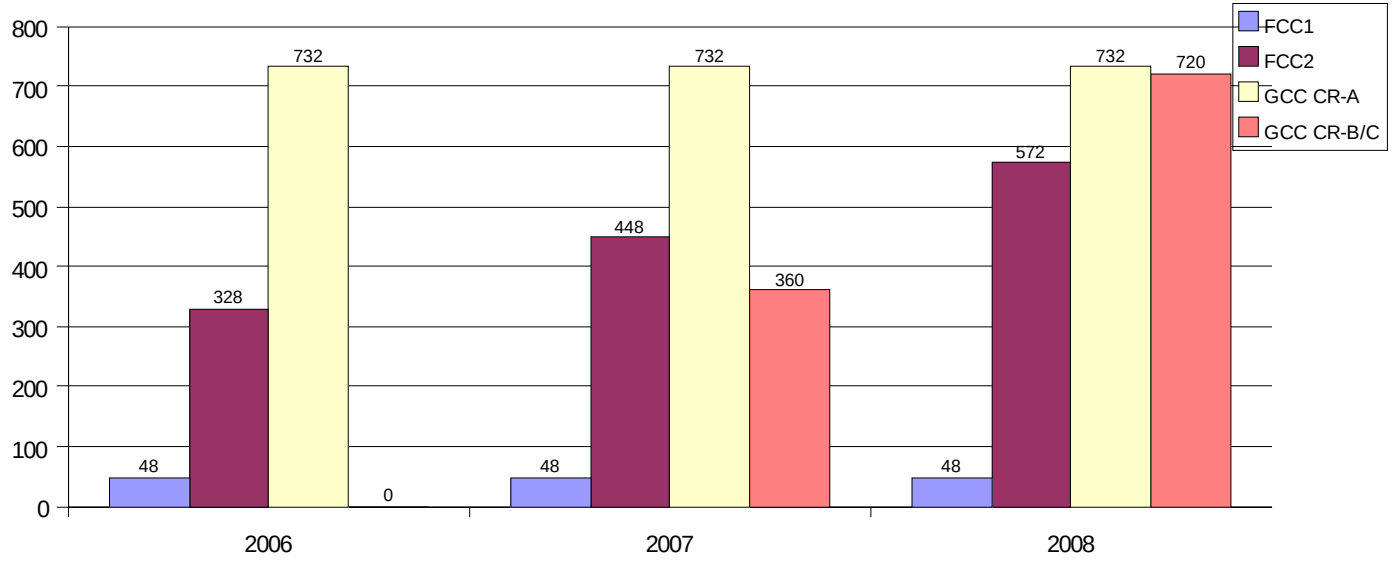
A ~7Gbps offsite rate:



A ~2Gbps disk servers rate:

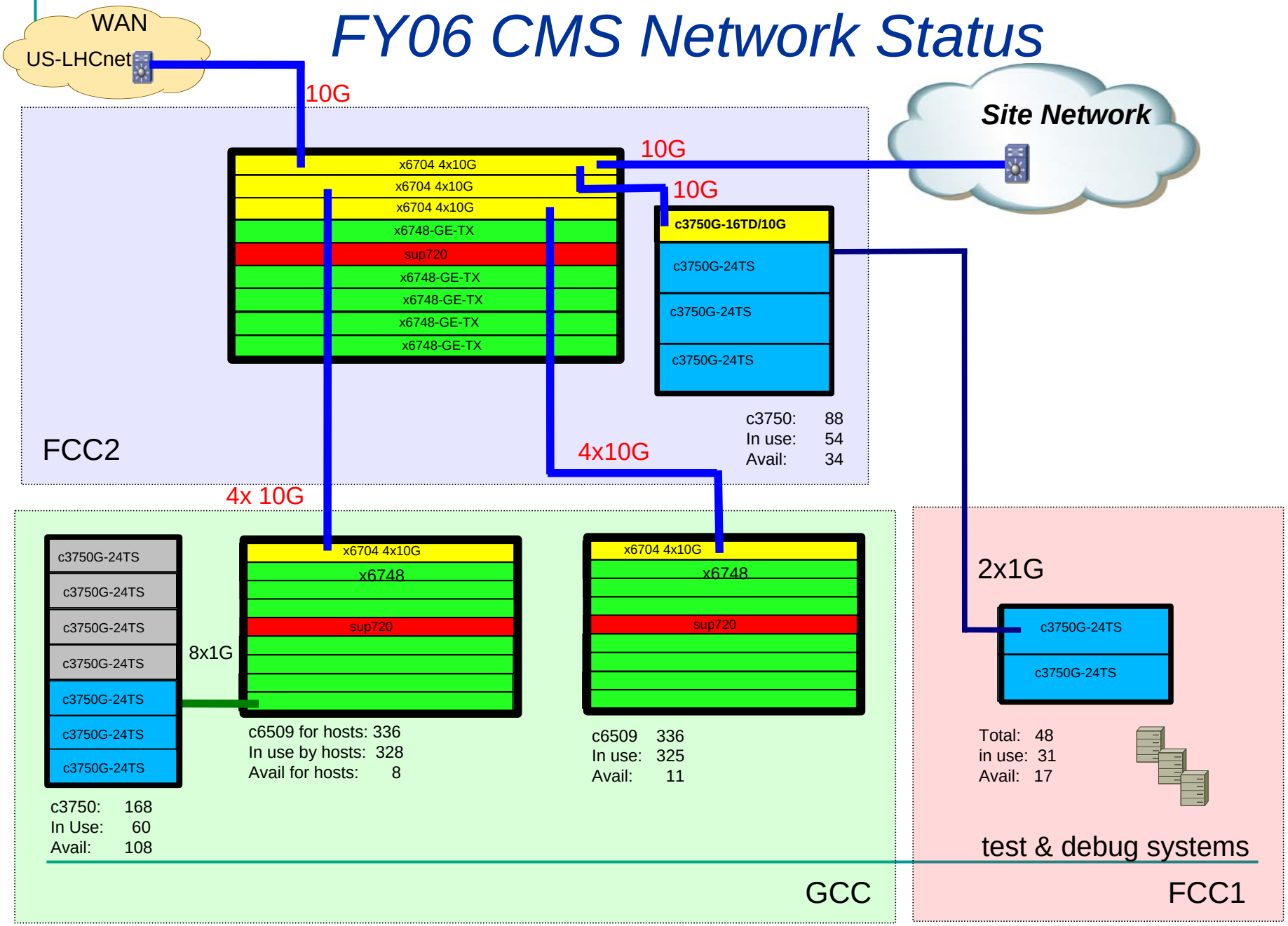


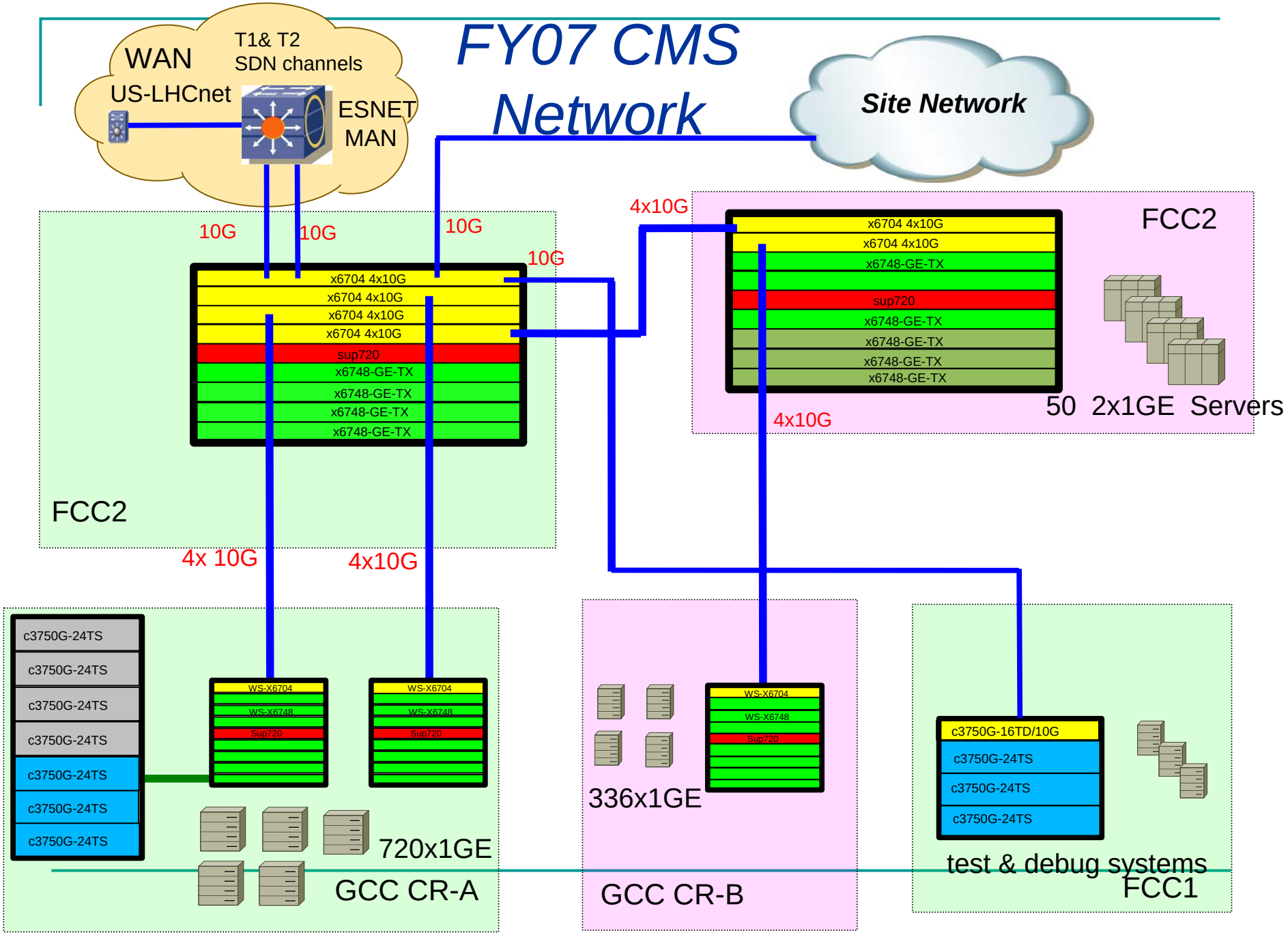
Status and requirements for a 1Gbps ports



- Increasing demands for 2x1Gbps connections in FCC2 CR (bonded servers connections)
- Higher density switching fabric is needed (anticipated by 1HCY08)
- Demands for higher concentration of CMS nodes
- Proliferation of locations restricts possibilities for use of high density switching fabric

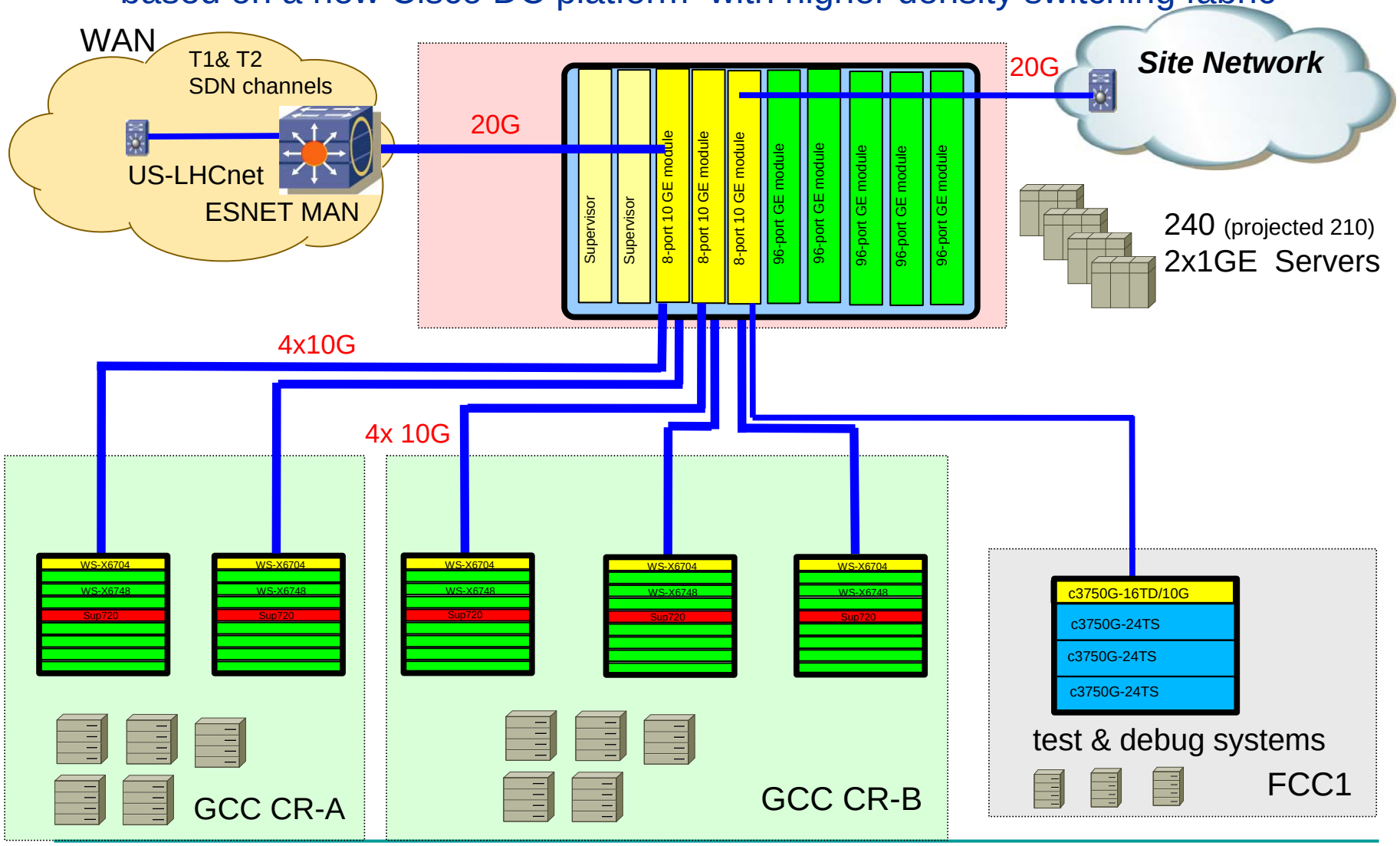
FY06 CMS Network Status




FY07 CMS Network
WAN
T1& T2
SDN channels
US-LHCnet
ESNET
MAN
Site Network
10G
10G
10G
10G
4x10G
x6704 4x10G
x6704 4x10G
x6704 4x10G
x6704 4x10G
sup720
x6748-GE-TX
x6748-GE-TX
x6748-GE-TX
x6748-GE-TX
FCC2
x6704 4x10G
x6704 4x10G
x6748-GE-TX
sup720
x6748-GE-TX
x6748-GE-TX
x6748-GE-TX
x6748-GE-TX
FCC2
50 2x1GE Servers
4x10G
4x 10G
4x10G
c3750G-24TS
c3750G-24TS
c3750G-24TS
c3750G-24TS
c3750G-24TS
c3750G-24TS
c3750G-24TS
WS-X6704
WS-X6748
Sup720
WS-X6704
WS-X6748
Sup720
720x1GE
GCC CR-A
WS-X6704
WS-X6748
Sup720
336x1GE
GCC CR-B
c3750G-16TD/10G
c3750G-24TS
c3750G-24TS
c3750G-24TS
test & debug systems
FCC1

FY08/09 CMS Network Upgrade

based on a new Cisco DC platform with higher density switching fabric



1650 worker nodes (projected 1176)

Conclusion on CMS Networking :

- Preparing requisitions for this year buys, assuming:
  - 50 new servers will go in FCC2
  - 260 worker nodes will be in GCC/CR-B
  - Cost = ***~$252K*** for networking to grow installation in FCC2
- Prepared plans for upgrades in FY08-09 based on different scenarios and higher capacity switch technology availability
- Growing demands for Nx1 Gbps (bonded) connections for servers and worker nodes
- Growing demands for higher density switching fabric (anticipate availability of a new Cisco DC platform in 1HCY08
- Growing needs for a higher concentration of CMS nodes

Monitoring

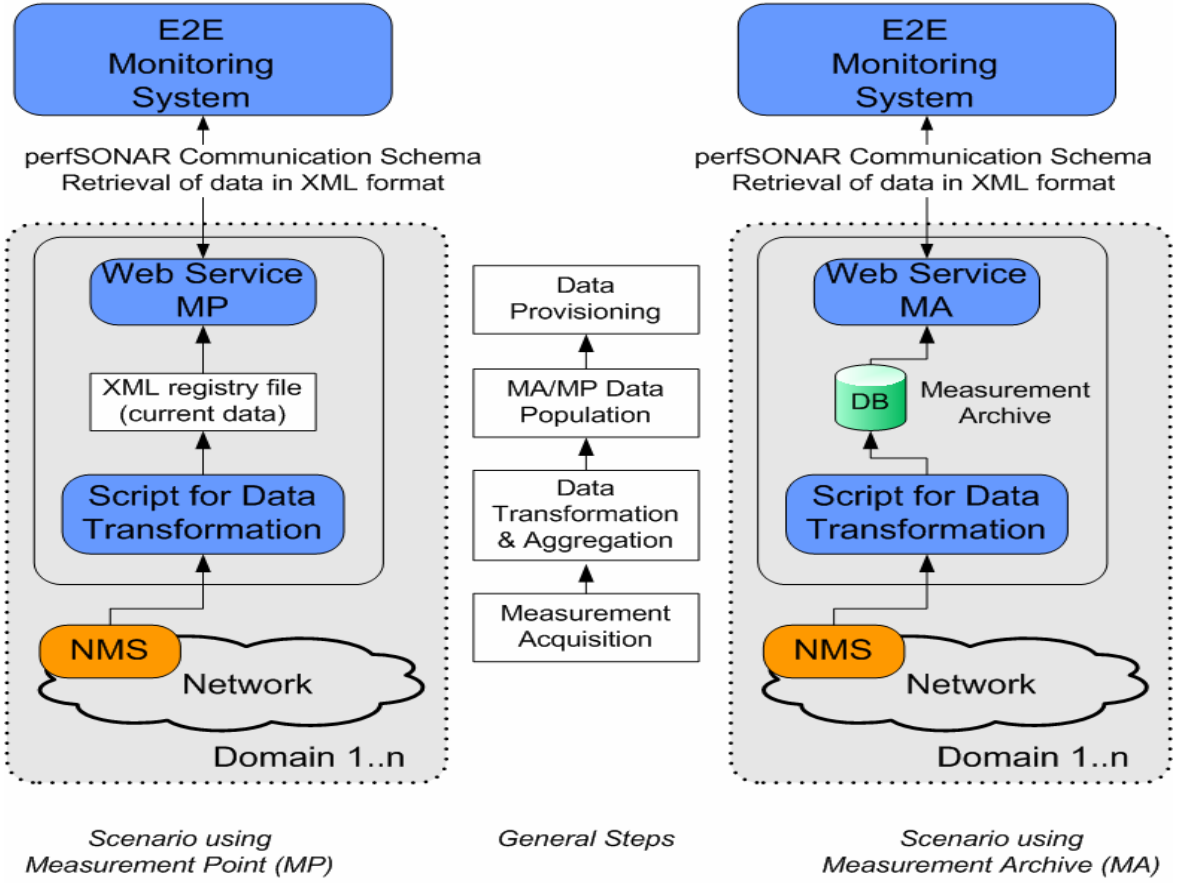
PerfSONAR

- Network monitoring middleware interoperable across platforms
  - based on SOAP API (python, java implementations and perl as separate project)
  - supported and developed by ESnet, Internet2 and all European NRENs
  - decentralized modular collection of discoverable services
- Transparent data exchange between authoritative domains by utilizing SOA
- Messaging protocol is based on SOAP XML messages and following the Open Grid Forum (OGF) Network Measurement Working Group (NM-WG).

Major components of the perfSONAR

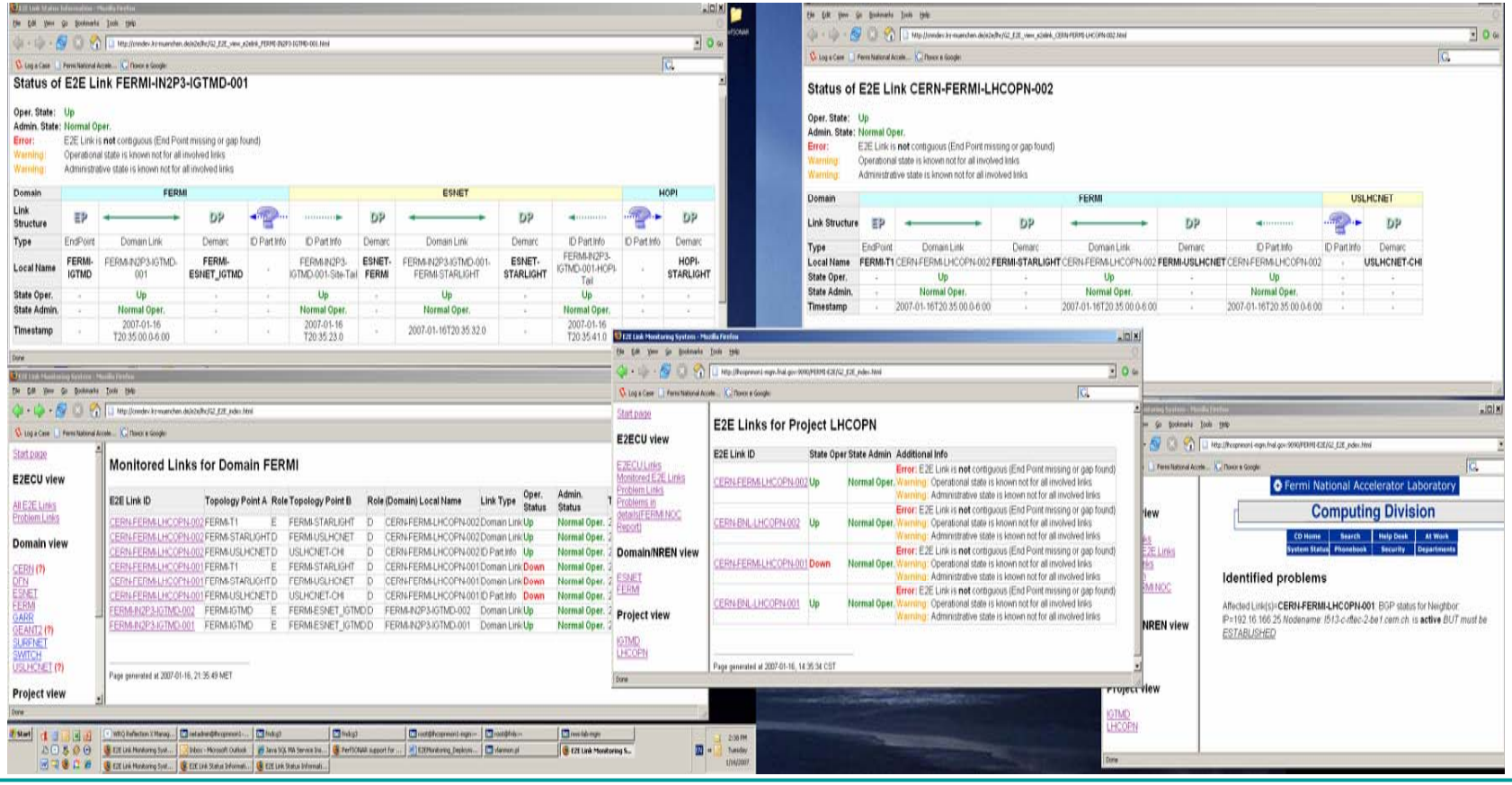
- MP – Measurement Point (standalone web service which publishes measurements)
- MA – Measurement Archive (RRD, SQL data formats) with historical data, accepts XML queries
- GUI java client (webstart based), traceroute visualization (by ESnet)
- LS – Lookup Service (service discovery)
- to be implemented:
  - Topology Service
  - AAA

perfSONAR for E2E monitoring



E2E Monitoring for LHCOPN and beyond at Fermilab

Virtual circuits status monitoring



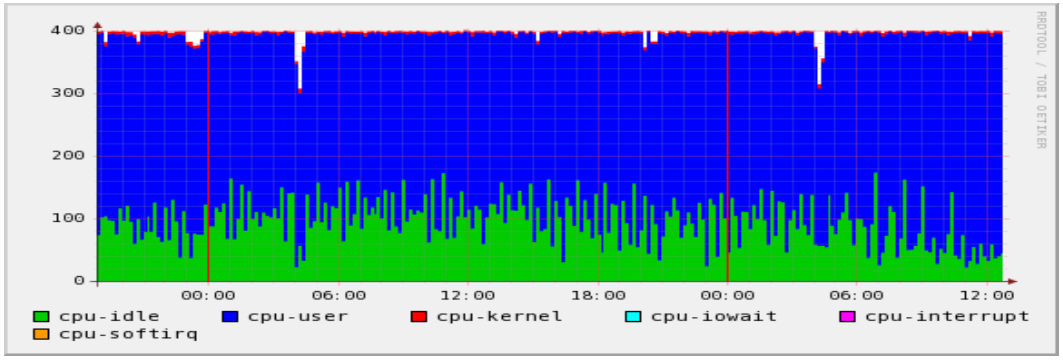
WAN monitoring

- PerfSONAR integration for LHCOPN E2E monitoring and beyond (IGTMD project with IN2P3)
  - E2E Measurement Point (MP)
    - BGP session monitoring
    - VLAN state monitoring
    - physical interfaces monitoring
  - SQL Measurement Archive (MA) installed and configured for E2E L2 status monitoring
- further development for WAN E2E monitoring and troubleshooting
  - contributing to perfSONAR project (perfSONAR PS – perl implementation)
  - integration with other monitoring frameworks (MonALISA)
  - analysis tools
  - configuration management and NOC interface
- 10 years of PingER data archive

Netflow Collection and Analysis

- Increasing demands for netflow data analysis, both for computer security purpose and network traffic characterization
- Migrated to the distributed netflow collection system using *BlueArc NAS* service supported by CSS department.
  - Current storage allocation is 1TB (~ one year's worth)
- Stopped gathering data in the format of Cisco Netflow collector at end of 2006 (no use for that now)
- 7 devices: border, startlight, 2 core routers, CMS workgroup, CDF and D0. Monthly volume is ~ 10GB/device.
- 3 processing systems (2.4GHz,2GB RAM, ~200GB disk)

A typical CPU utilization

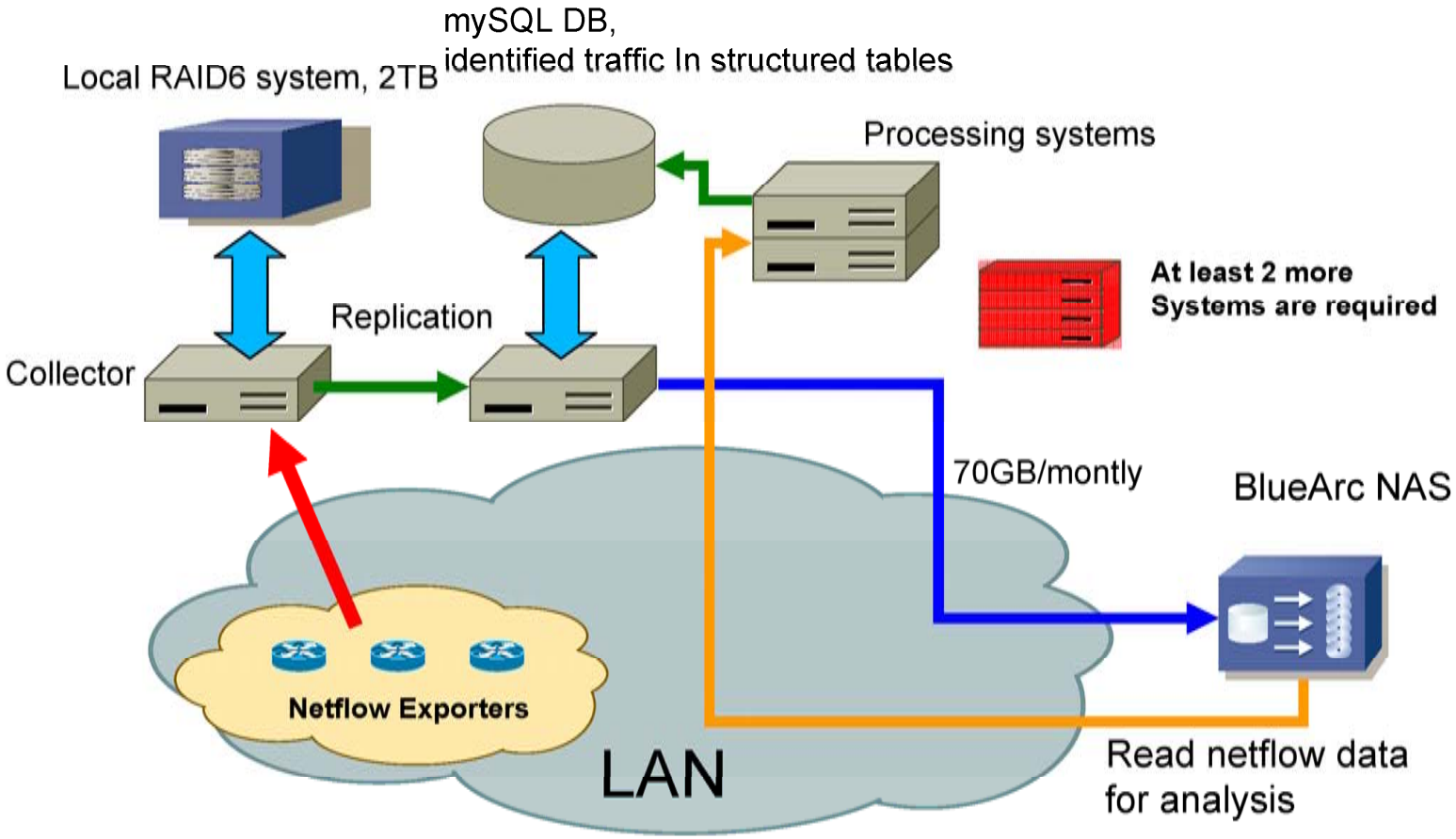


Netflow Analysis

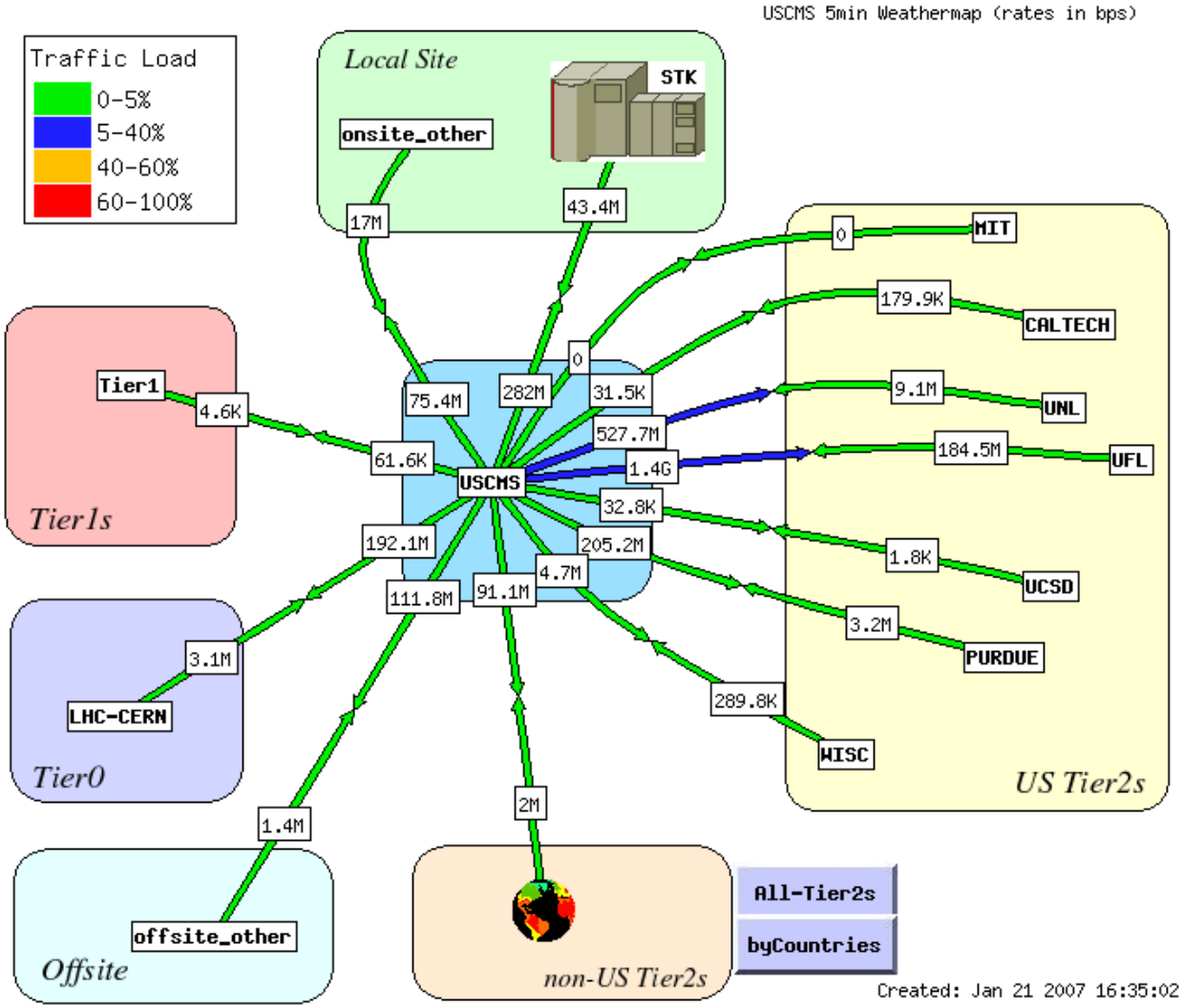
- **AutoBlocker** (*4 years 91* days guarding the Fermilab network, (without breaks), using self controlled mechanism):
  - Added some capabilities for traffic recognition, Bittorent, a few typical profiles for CDF and CMS clusters
  - Improved notifications are sent to NIMI
  - New improved exception system helps avoid blocking valid traffic.
    - Rate of false positive blocks is very low.

- **Traffic characterization**
  - Flow-tools package is used for analysis
  - New topN, distributed tagging of traffic, results are in mySQL DB
  - Distributed processing for border, StarLight and CMS WG,
    - 3 servers are used
    - need at least 2 more to satisfy minimal demands
  - Evaluating commercial software packages

Netflow Collection and Analysis



CMS Weather Map



WAN Group R & D Activities

Lambda Station Project

On-demand flow based routing for data intensive GRID applications over multi topology networks

- In 3rd year of proposed three year project
- Goals are:
  - a working Java implementation of software
  - testing with production SciDAC applications
- Working prototype used in SC05 demo and other tests between FNAL and Caltech
- **SRM/dCache** has a limited Lambda Station awareness, places ticket, join and extend, recognizes absence of Lambda Station, tries to use it only for specific destinations
- Java-based implementation in progress
- Presented at SC05, CHEP04, CHEP06, GridNets2006 Workshop
- Published in the IEEE proceedings of the International conference Broadnets 2006. Oct 1-2, San Jose, California, USA

Publications

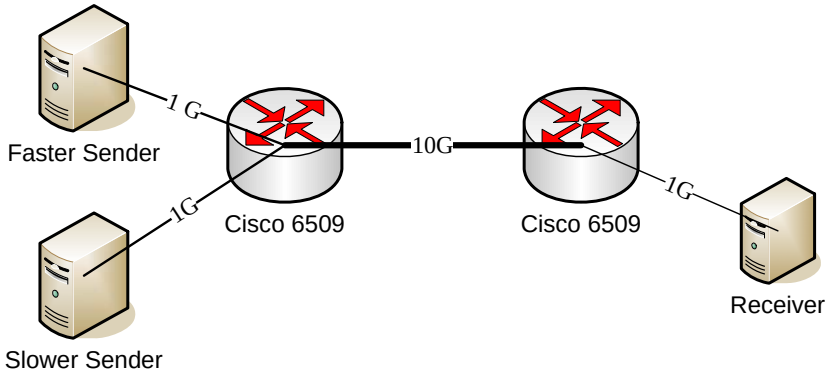
- Wenji Wu, et. al. “Two-layer Hierarchical Wavelength Routing for Islands of Transparency Optical Networks,” International Journal of Computer Communications, vol. 29, pp. 2952-2963, Elsevier, 2006.
- Wenji Wu, Matt Crawford, Mark Bowden, “The Performance Analysis of Linux Networking – Packet Receiving,” to appear in International Journal of Computer Communications, Elsevier, 2007.
- Wenji Wu, Matt Crawford, “Potential Performance Bottleneck in Linux TCP,” to appear in International Journal of Communication Systems, Wiley, 2007.
- Wenji Wu, Matt Crawford, “Interactivity vs. Fairness in Networked Linux Systems”, submitted to International Journal of Computer Networks, Elsevier, 2006.
  - In revision, passed the first-round review

Interactivity vs. Fairness in Networked Linux Systems (cont.)

- Linux classifies processes as “interactive” or “non-interactive” based on their run/sleep characteristics.
- A process that wakes up for a short period of time to handle an I/O event (e.g., pressing a key on the keyboard) and then goes back to sleep is an “interactive” process.
- A process that runs continuously during its timeslice is a “non-interactive” process.
- Depending on the packet arrival rate, a network receiving process can end up being classified as either “interactive” or “non-interactive”.
- The longer the gap between packets, the more likely that the receive process will be classified as “interactive”...this can happen even if the process is running for 90% of it’s timeslice.

Interactivity vs. Fairness in Networked Linux Systems

- Run *iperf* to send data in one direction between two computer systems
- We have added instrumentation within Linux kernel
- Compiling Linux kernel as background system load by running ***make –nj***
- Receive buffer size is set as 40M bytes

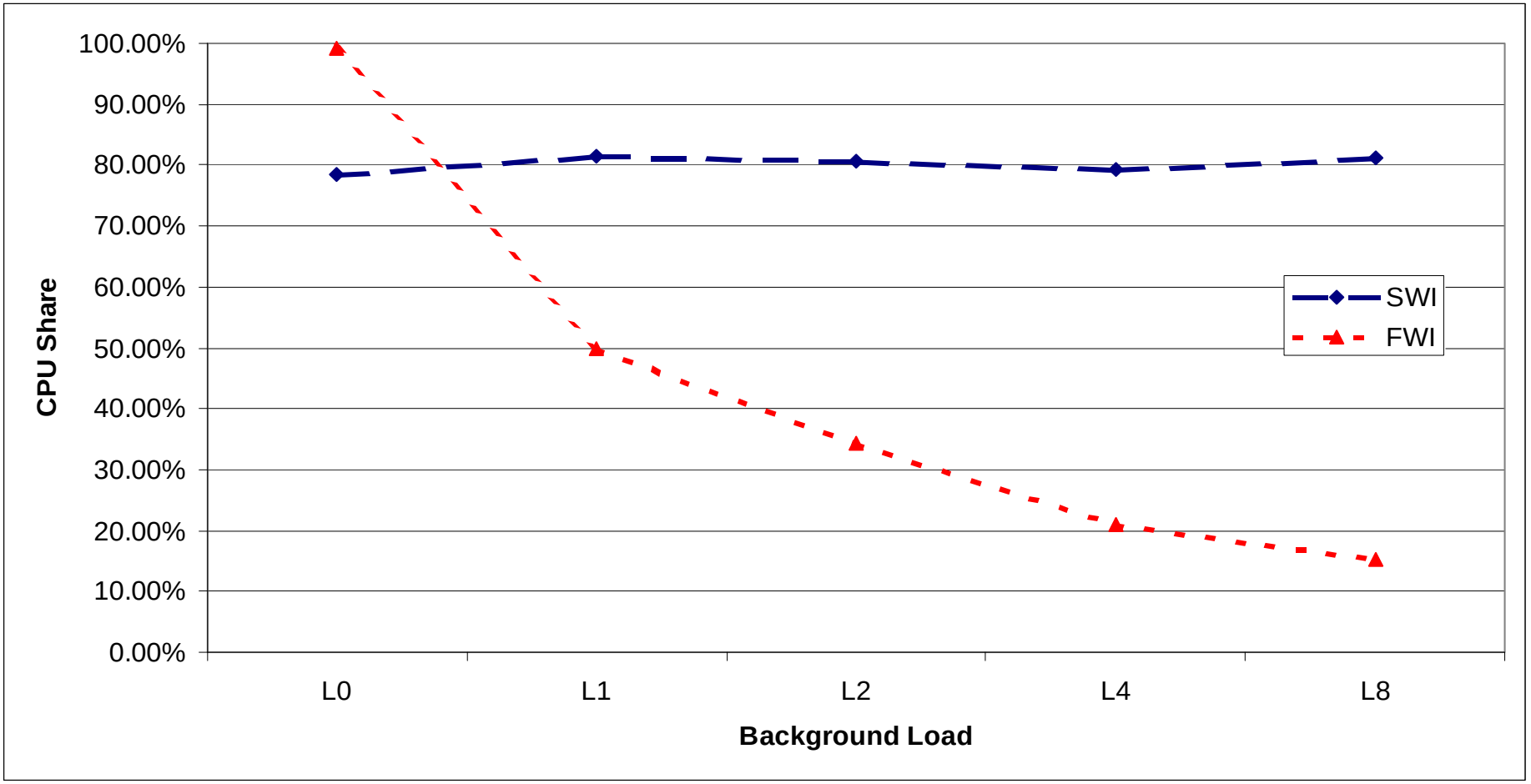


Fermi Test Network

	Fast Sender	Slow Sender	Receiver
CPU	Two Intel Xeon CPUs (3.0 GHz)	One Intel Pentium IV CPU (2.8 GHz)	One Intel Pentium III CPU (1 GHz)
System Memory	3829 MB	512MB	512MB
NIC	Syskonnect, 32bit-PCI bus slot at 33MHz, 1Gbps/sec, twisted pair	Intel PRO/1000, 32bit-PCI bus slot at 33 MHz 1Gbps/sec, twisted pair	3COM, 3C996B-T, 32bit-PCI bus slot at 33MHz, 1Gbps/sec, twisted pair

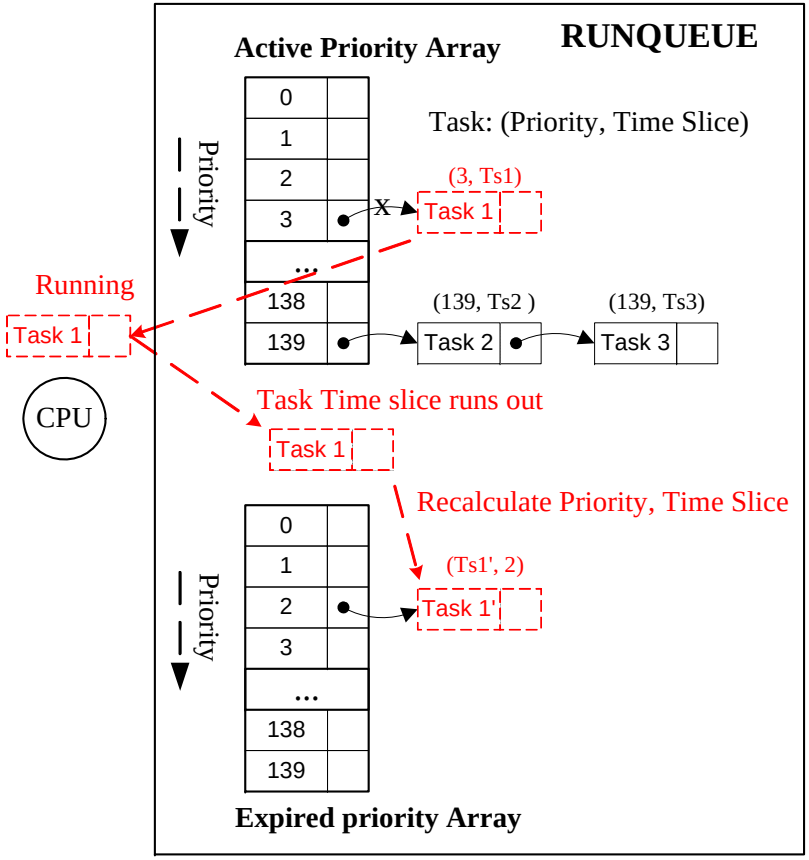
Sender & Receiver Features

Interactivity vs. Fairness in Networked Linux Systems (cont.)



What, Why, and How?

Interactivity vs. Fairness in Networked Linux Systems (cont.)

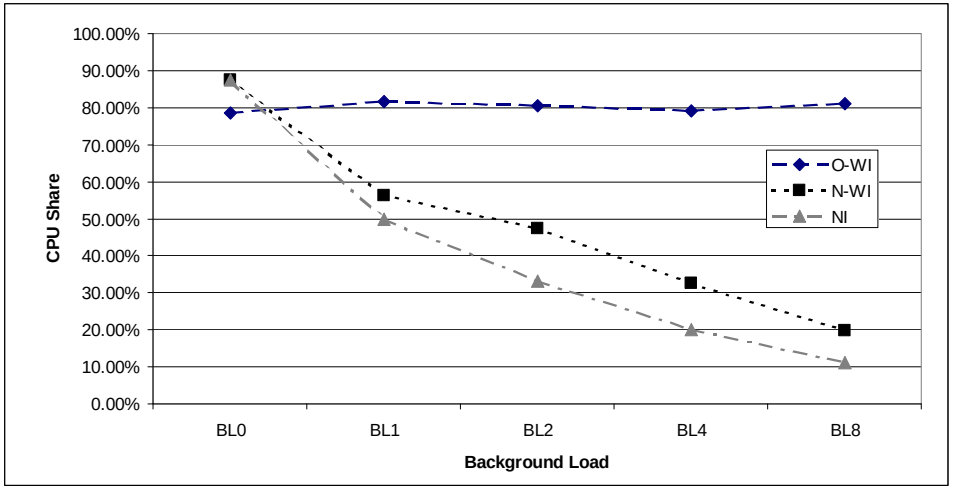


Linux Scheduling Mechanism

Interactivity vs. Fairness in Networked Linux Systems (cont.)

- For a truly interactive application, the wait-for-packet sleeps usually last for tens or hundreds of milliseconds or more; However, the inter-packet sleeps for bulk data transmission application usually last for a few milliseconds or less.
  - A possible solution: When the inter-packet sleep duration does not exceed some minimal value, the sleep is not credited to the process' interactivity estimator.

We have implemented this strategy within Linux kernel!!



Admission Control Needed for Linux-based Networked Systems (cont)

- When a Linux-based network system is overloaded, each TCP-stream's performance is degraded. In the extreme, the system might crash
  - In Fermilab, we have seen over-loaded dCache systems (Linux-based) crash, how about performance?

- Procedures must be taken to prevent the situation that systems are over-loaded.
  - Over-engineering?
  - Admission control?

Admission Control Needed for Linux-based Networked Systems

State	Conditions	Memory Pressure
Hard_memory_pressure	> sysctl_tcp_mem[2]	Yes
Soft_memory_pressure	> sysctl_tcp_mem[1]	Yes
Non_memory_pressure	< sysctl_tcp_mem[0]	No

Linux TCP Memory Pressure

		Per TCP Socket Memory Allocation Status		
		< sysctl_tcp_mem[0]	> sysctl_tcp_mem[0]	> sysctl_tcp_mem[2]
Global TCP Memory Allocation Status	Hard_memory_pressure	F	F	F
	Soft_memory_pressure	S	F	F
	Non_memory_pressure	S	S	F

Relationship between per TCP socket memory allocation status and global TCP memory allocation status

Potential Future Research Projects

- Network performance on 64-bit platforms
  - CPU is moving from 32-bit to 64-bit
  - Network applications on 64-bit platforms
    - Source code issues
    - Library package issues
  - Protocol stack on 64-bit platform
- Virtualization
  - Virtual machine's network performance
  - Interactions between virtual machine and physical networks

Kernel Tuning for Storage Systems

- Problem: Achieving CMS service challenge target rates requires too many concurrent file transfers. Applying the standard formula for high bandwidth × delay led to system crashes.
  - “Low memory” limitations in 32bit Linux make it unstable on systems with many high latency network connections.
  - Linux default setting for total network buffer memory is miscalculated, leading to out-of-memory condition.
  - Switching to 64bit Linux solves some problems, creates others (64bit kernel and JVM are not as well tested).
- Results presented at LHC Computing Tier-2 Workshop
  - Interest expressed in working with Fermilab on network tuning (Taipei, Melbourne, Mumbai)

Pure Optical (“Photonic”) Switching

- Worldwide interest and activities coordinated through GLIF (or gλif) - Global Lambda Integrated Facility.
- Our interest: reservable clear-channel paths to other sites
  - avoid congestion & effects of limited buffering in routers & switches
  - remote switching of redundant optical links, optical loopback
- Status
  - Polatis 16 X 16 optical switch purchased
  - basic tests completed using FAPL cluster
  - now installed at Starlight
    - Connected to control network; not yet in optical path
    - Expect to connect to HOPI for inter-domain channelized wavelength testing

Advanced Host Capabilities

- Apply Linux network traffic control filters to optimize data rates in wide area connections
  - Rate limiting to avoid TCP congestion oscillation
  - Shape outgoing traffic to predefined distributions
  - Emulate wide-area network connections by adding packet delay
  - Emulate network problems by automatically inserting bit errors, missing or duplicate packets, out-of-order packets

- Network simulation
  - Capture network activity (tcpdump, etc)
  - Analyze (tcptrace, xplot, etc)
  - Convert to realistic input for network simulation
  - Identify common problems

Miscellaneous

ESnet Site Requirements Survey

- Intended to provide understanding of current & future networking needs of ESnet sites to forcast service requirements on ESnet
  - Includes large scale data movement needs, but has wider scope:
    - Enhanced network services (QoS, dedicated network paths, etc.)
    - Collaborative tools (video & teleconferencing)
    - Security services (PKI infrastructure)
  - Requesting short & long term forcasts on requirements:
    - What's needed within the next 12 months
    - What is projected over the next 5-10 years
    - What might be needed in the distant (10-15 year) future
- 3 sections, requesting respective responses from the site, instruments & facilities at the site, and heavy users & PIs at the site:
  - Responses are expected to overlap

ESnet Site Requirements Survey - II

- We plan to compile the data for the survey response by:
  - Sending the survey to the CD liaison for each major experiment / project:
    - CMS Tier-1 facility; Run-II experiments, LCQCD, ILC, SDSS2, Neutrino experiments, LHC@FNAL, OSG (any others?)
  - Letting the experiment / project liaison solicit information from PI types, if appropriate
  - Aggregating the results
  - Adding in projected general facility requirements
- Deadline for responses is April
  - Rough cut desired by upcoming ESCC meeting (mid-February)

IPv6

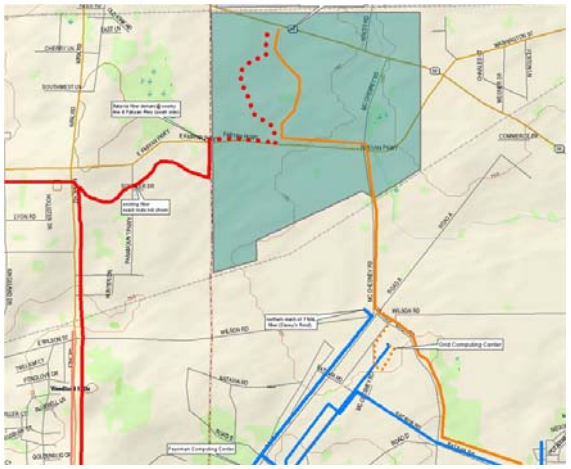
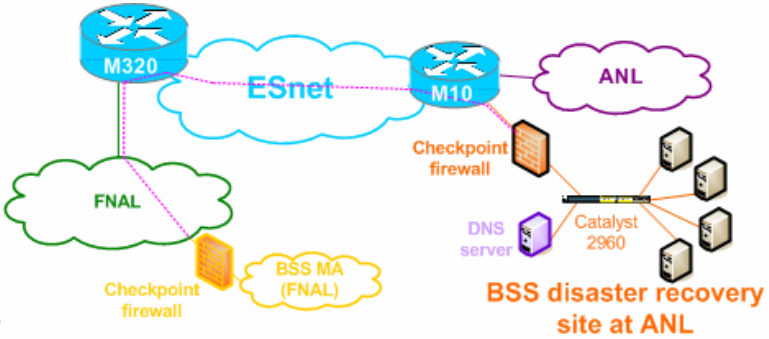
- OMB M-05-22 requires agency backbones to support IPv6 by 6/30/08
  - But extending these requirements to cover site networks seems inevitable
  - Extensive IPv6 deployment in Asia could generate request for support here
- Our current IPv6 activities:
  - Very small testbed in place, using backup border router (no off-site access)
  - Coordination with ESnet on address assignment & peering requirements
    - FNAL Address block = 2001:400:2410::/48
  - (Co)leading ESnet Site Coordinators committee effort to organize working group that would develop roadmap for site IPv6 transition planning
- Tentative outline for next steps in IPv6 support:
  - Expand testbed to new CD-based subnet with IPv6 support, for volunteers
  - Establish native IPv6 peering with ESnet (connectivity to the IPv6 world…)
  - Focus on developing IPv6 knowledge-base, tools, & monitoring capabilities

Wide Area Work Group (WAWG)

- A body for discussing & coordinating wide area network issues
  - Meets alternate Fridays at 9:00am in FCC1 (video conferencing avail.)
  - Open agenda; all wide area network issues or problems are welcomed
  - In existence for almost 3 years:
  - In the last year, outside participation had become spotty-to-nonexistent
- Conducted a self-assessment and decided:
  - FNAL-centric operational issues should be kept outside WAWG activities
  - Must revert to coordination body for inter-organizational WAN activities
    - Need to be proactive on issues & pulling in participation
- Currently focusing on:
  - LHC data movement issues & concerns
  - Monitoring tools & issues, in particular facilitating PerfSonar deployment
  - End-to-end circuit establishment, support, use, & testing

Odds & Ends

- DuPage National Technology Park (DNTP)
  - Ducting to DNTP is in place, including GCC spur
  - Abovenet fiber is in place
  - FNAL GCC/DNTP fiber is being installed now
  - But no racks/facilities yet in DNTP Comm Center



- BSS Disaster Recovery Site at ANL:
  - CIDR address block received from ESnet
    - Subnet allocation strategy adopted
  - Support model modified to include DNS server at ANL
  - Waiting to move forward with deployment