
US CMS Tier1 Facility Network

Andrey Bobyshev (FNAL)
Phil DeMar (FNAL)

CHEP 2010
Academia Sinica
Taipei, Taiwan



Outline of the talk :

- USCMS Tier1 Facility Resources
- Data Model/requirements
- Current status
- Circuits/CHIMAN/USLHC Net
- Network tools graphs and snapshots



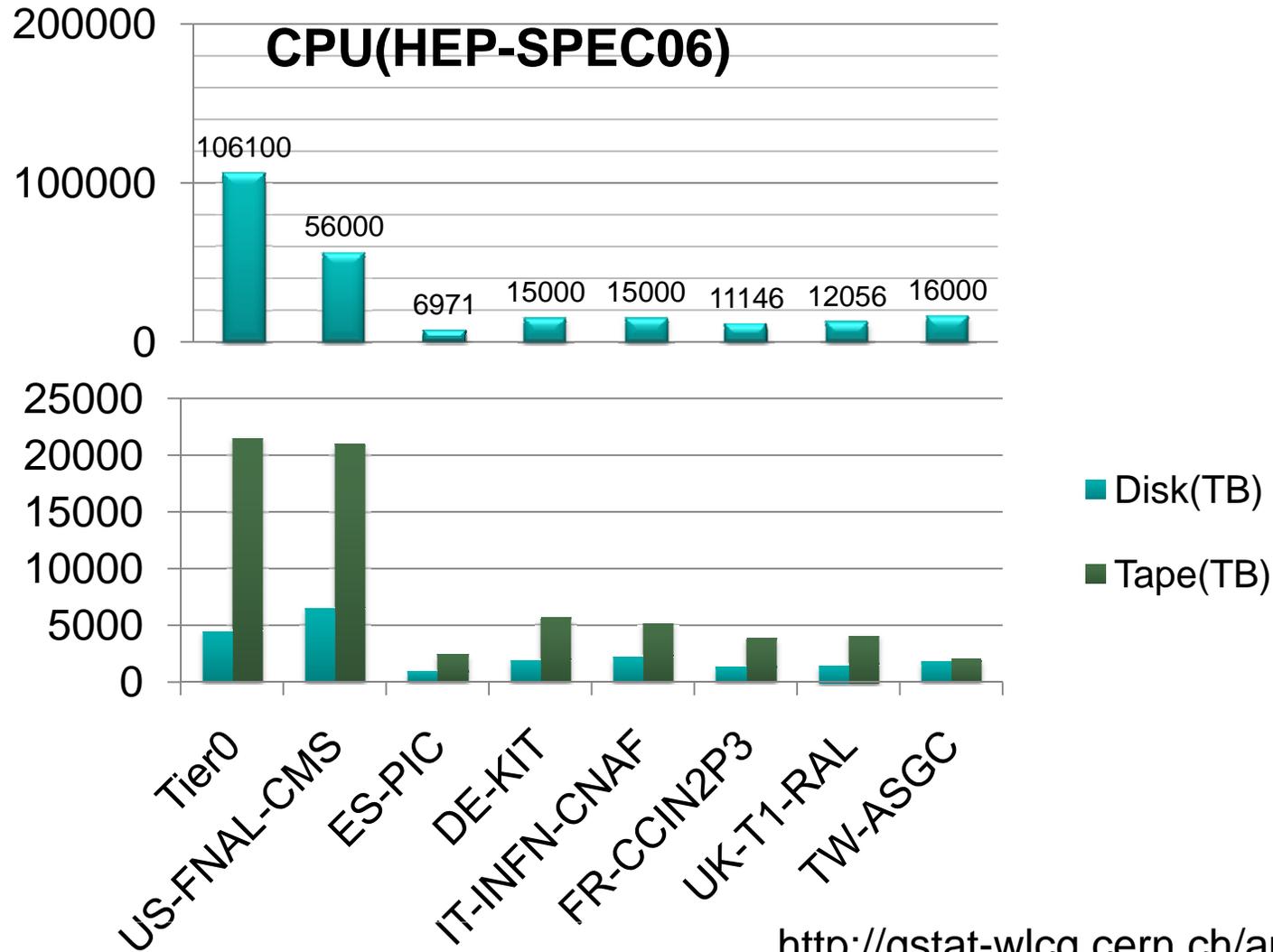
Summary of CMS resources

		Tape (TB)	Disk (TB)	CPU (kHS06)
Switzerland	CH-CERN	21600	4500	106100
France	FR-CCIN2P3	3876	1342	11146
Germany	DE-KIT	5700	1950	15000
Italy	IT-INFN-CNAF	5200	2200	15000
Spain	ES-PIC	2424	902	6971
Taiwan	TW-ASGC	2000	1800	16000
UK	UK-T1-RAL	4192	1560	12058
USA	US-FNAL-CMS	21000	6500	56000

<http://gstat-wlwg.cern.ch/apps/pledges/>



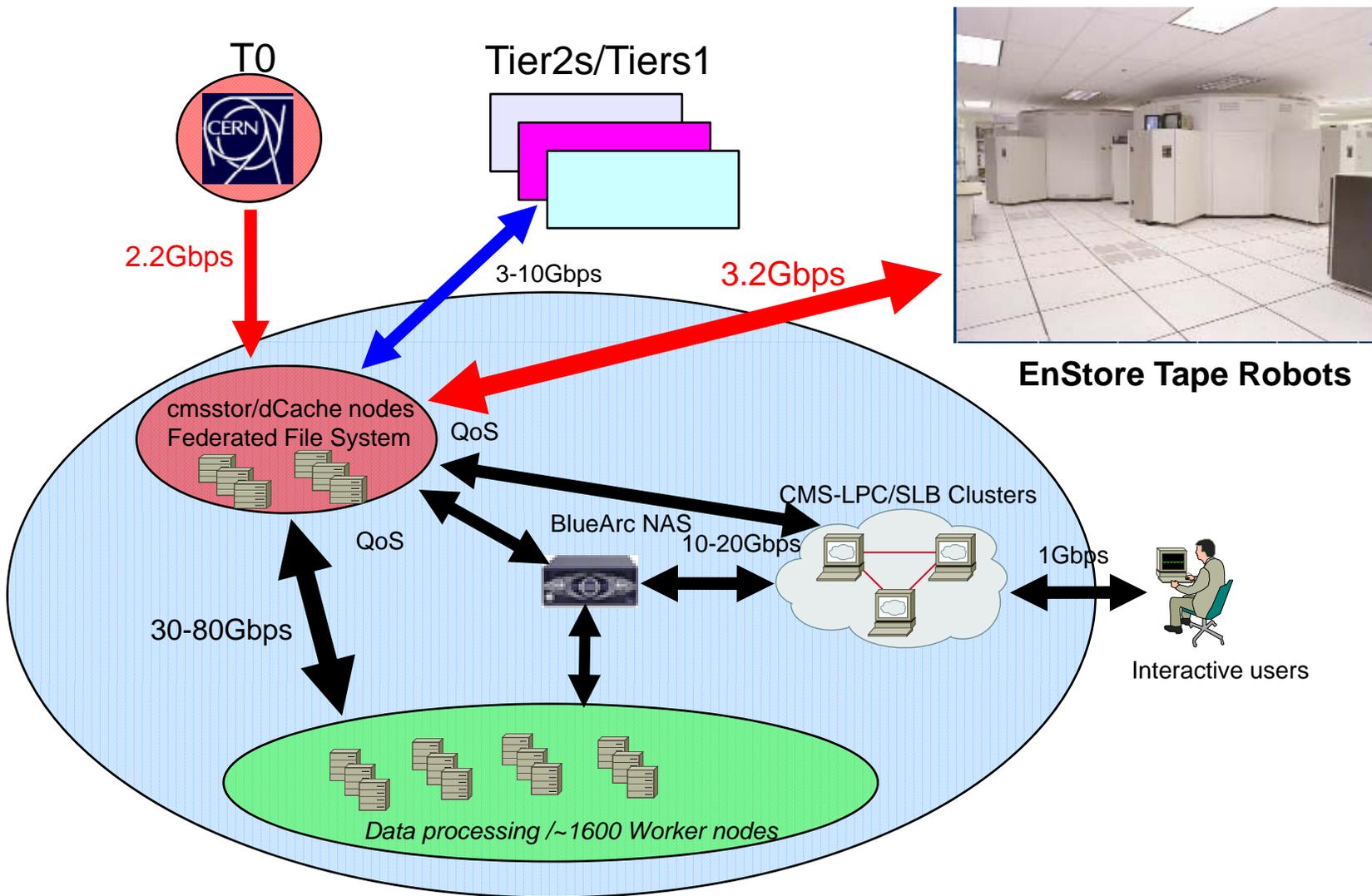
CMS resources (2011 pledges)



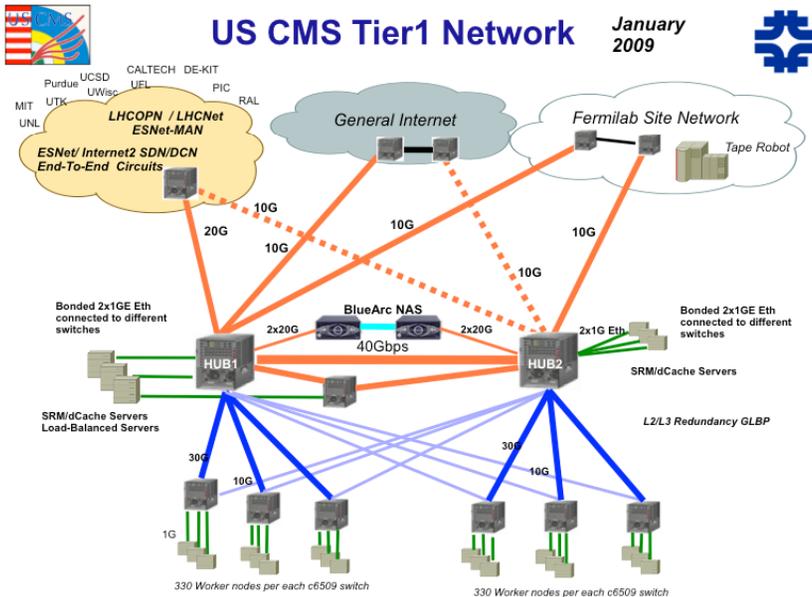
<http://gstat-wlwg.cern.ch/apps/pledges/>



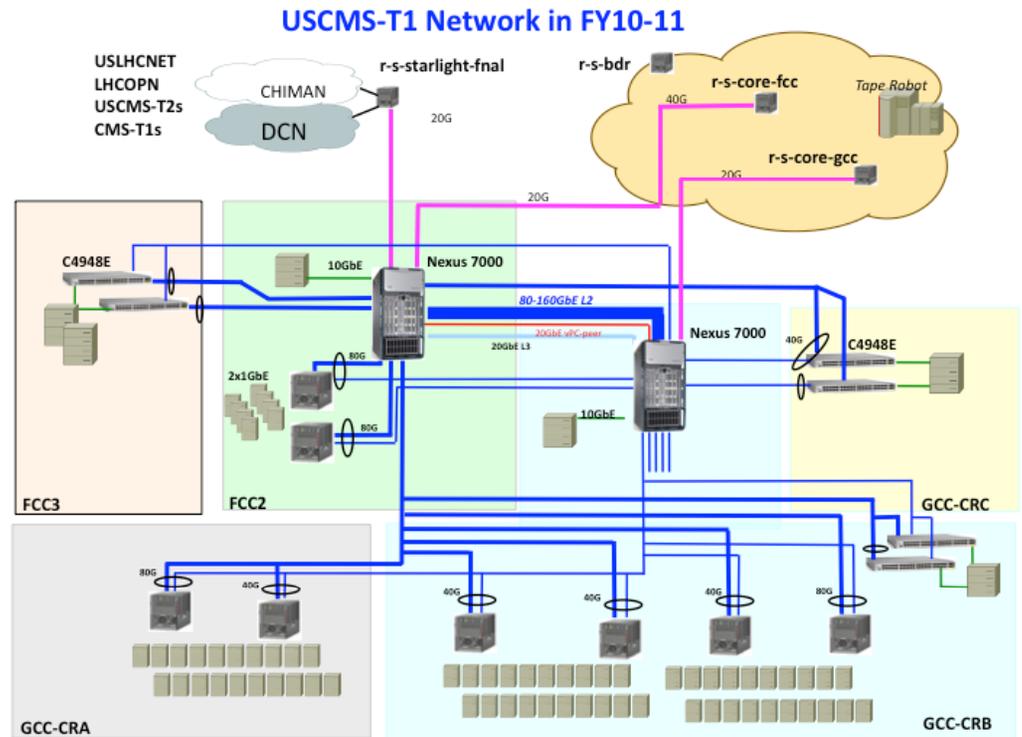
Model of USCMS-T1 Network traffic



Upgrade of Core Fabric Recently Completed



Cisco 6509-based core

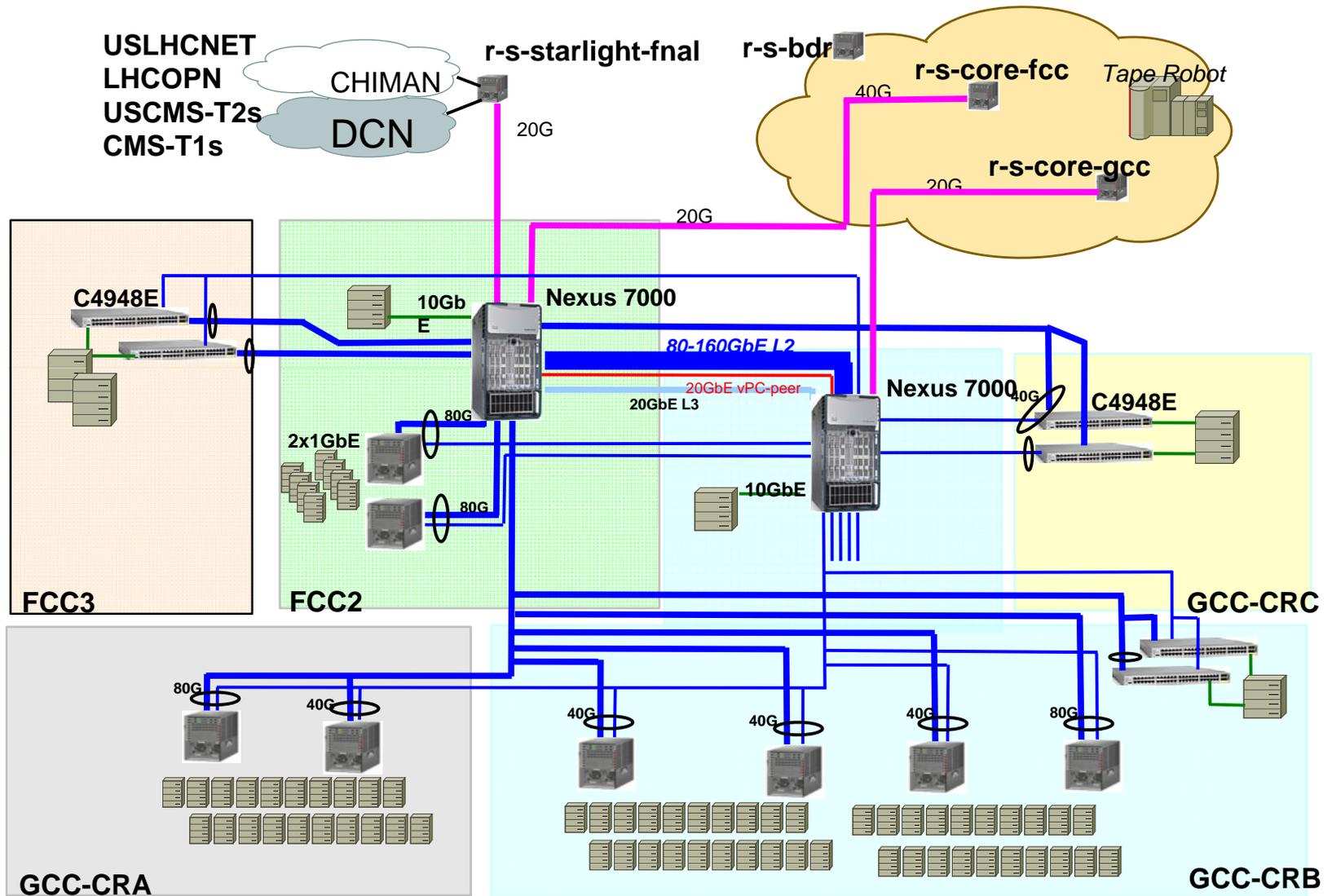


Cisco Nexus 7000-based core

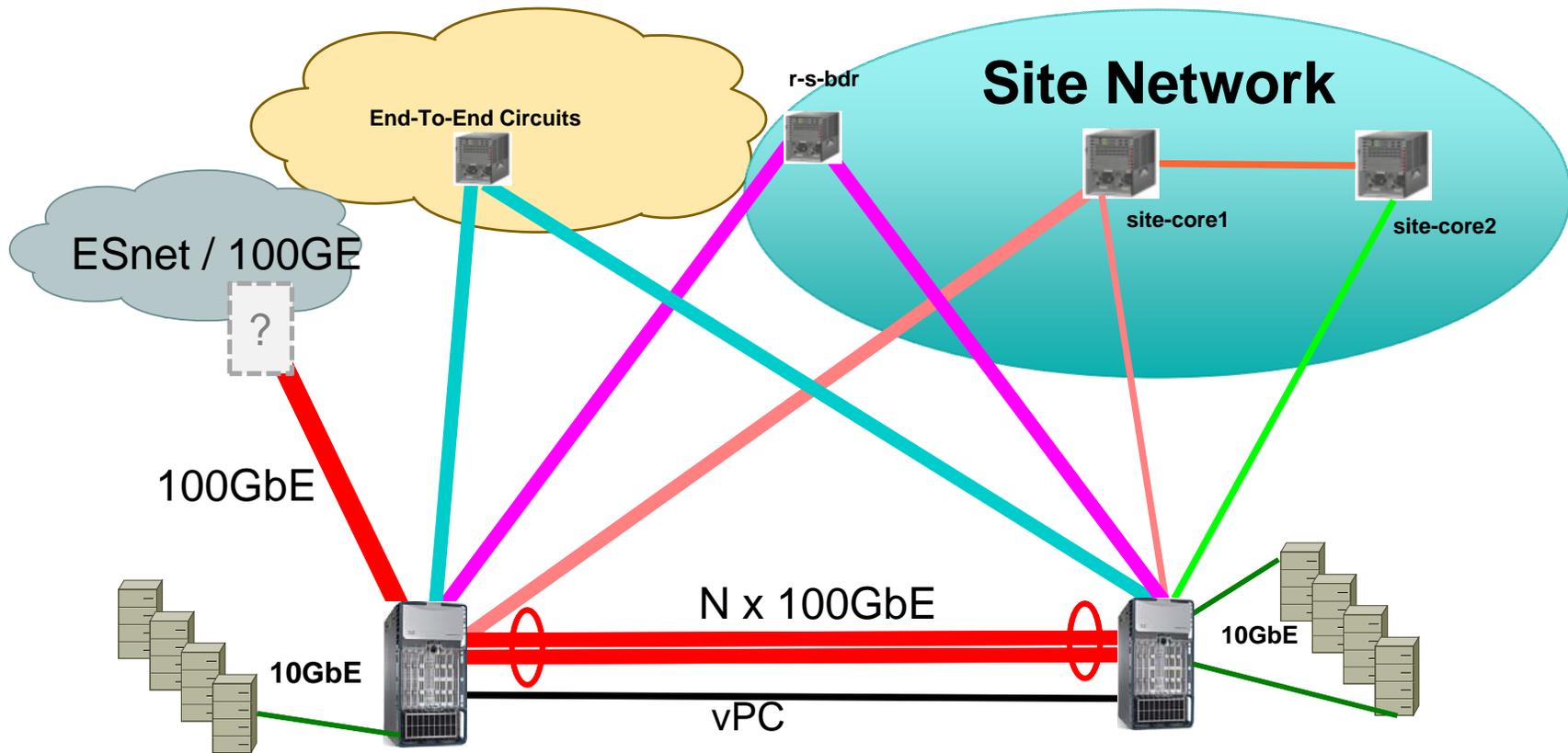
548_2010-10-01



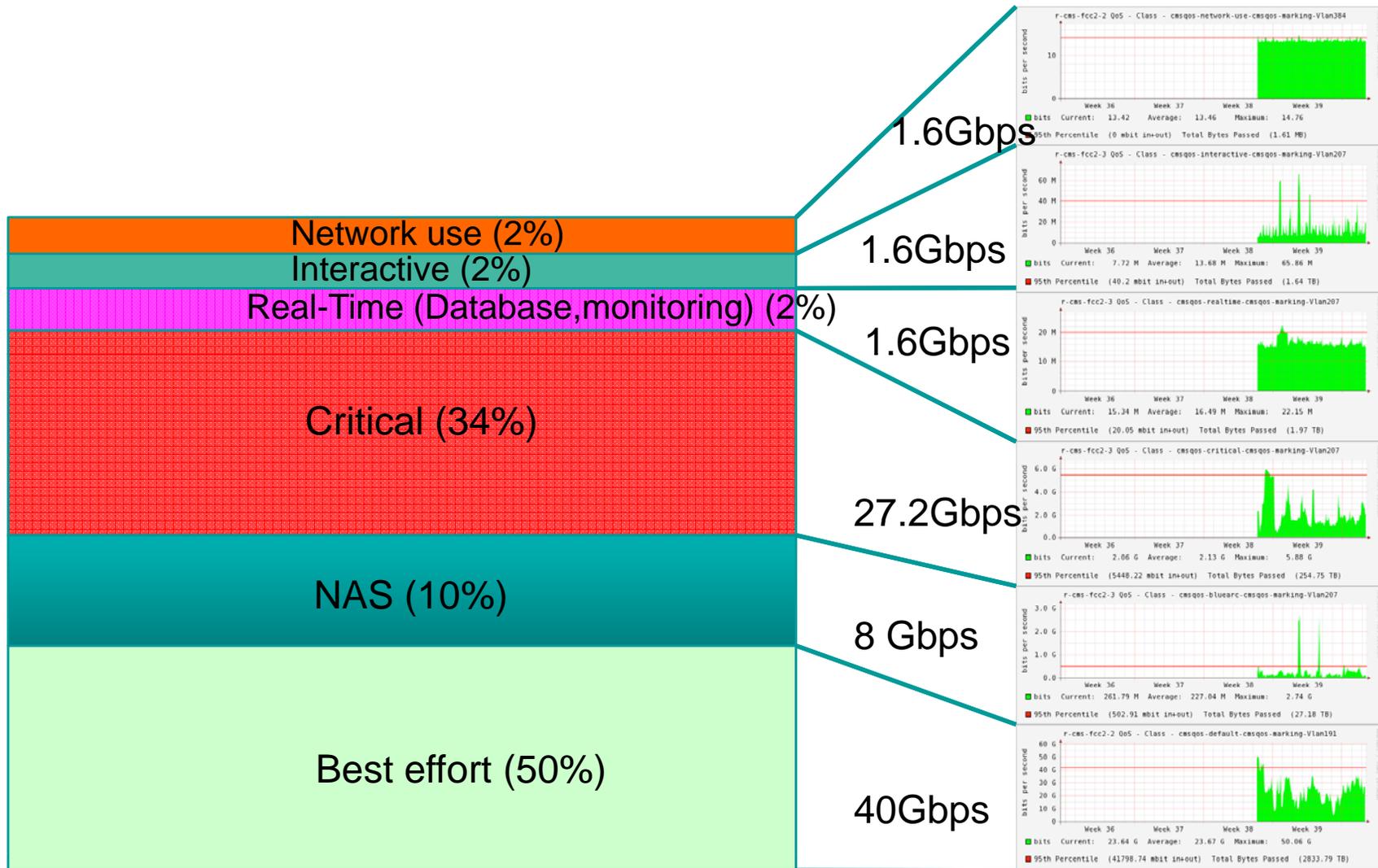
USCMS-T1 Network in 2010-11



USCMS Tier1 Network in 2013-14



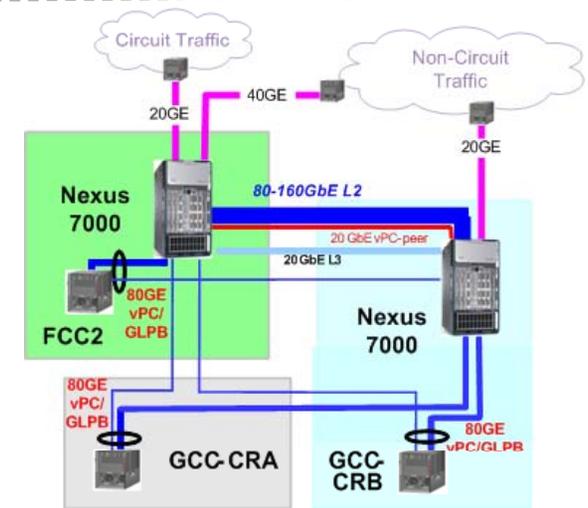
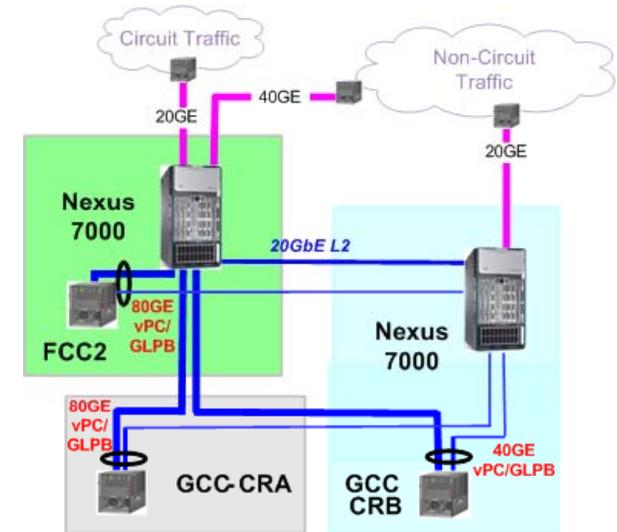
QoS Classes of Traffic in USCMS-T1



Redundancy Within the Tier-1

- Today:
 - FCC2 Nexus is core switch fabric
 - GCC-CRB Nexus = redundant core
 - Switches connected at 4 (or 8) x 10GE
 - Virtual Port channel (vPC)
 - Gateway Load-balancing protocol (gLBP) for failover to redundant core

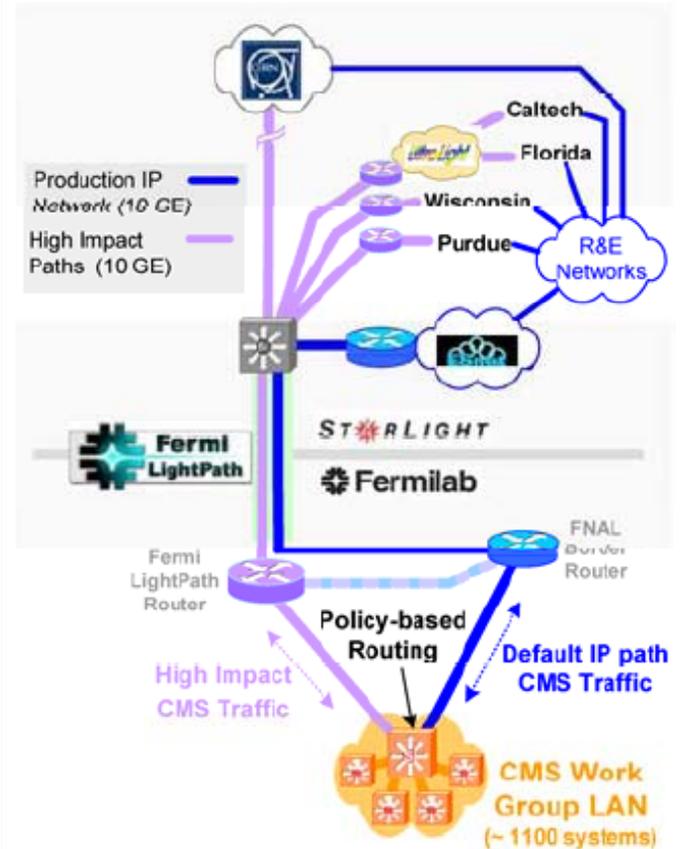
- Near Future:
 - Interconnected Nexus's @ 80-160Gb/s
 - Function as distributed fabric
 - Switches still configured w/ vPC & gLBP
 - Most connections to nearest core switch



Off-Site Traffic: End-to-End Circuits

USCMS-T1 has a long history of using ESNNet/SDN and Internet2 DCN circuits

Circuit	Country	Affiliation	BW
LHCOPN	Switzerland	T0	8.5G
LHCOPN Secondary	Switzerland	T0	8.5G
LHCOPN Backup	Switzerland	T0	3.5G
DE-KIT	Germany	T1	1G
IN2P3	France	T1	2x1G
ASNet/ASGC	Taiwan	T1	2.5G
CALTECH	USA	T2	10G
Purdue	USA	T2	10G
UWISC	USA	T2	10G
UFL	USA	T2	10G
UNL	USA	T2	10G
MIT	USA	T2	10G
UCSD	USA	T2	10G
TIFR	India	T2	1G
UTK	USA	T3	1G
McGill	Canada	CDF/D0	1G
Cesnet, Prague	Czech	D0	1G



SLA monitor, IOS Track objects to automatically fail over traffic if circuit is down



PerfSonar Monitoring

- Monitoring status of circuits
- Alert on a change of link status
- Utilization
- PingEr RTT measurements
- PerfSonar-BUOY – Active measurements, BWCTL & OWAMP
 - Two NPI Took kit boxes
- Two LHCOPN/MDM monitoring boxes



Circuits SLA monitoring (Nagios)

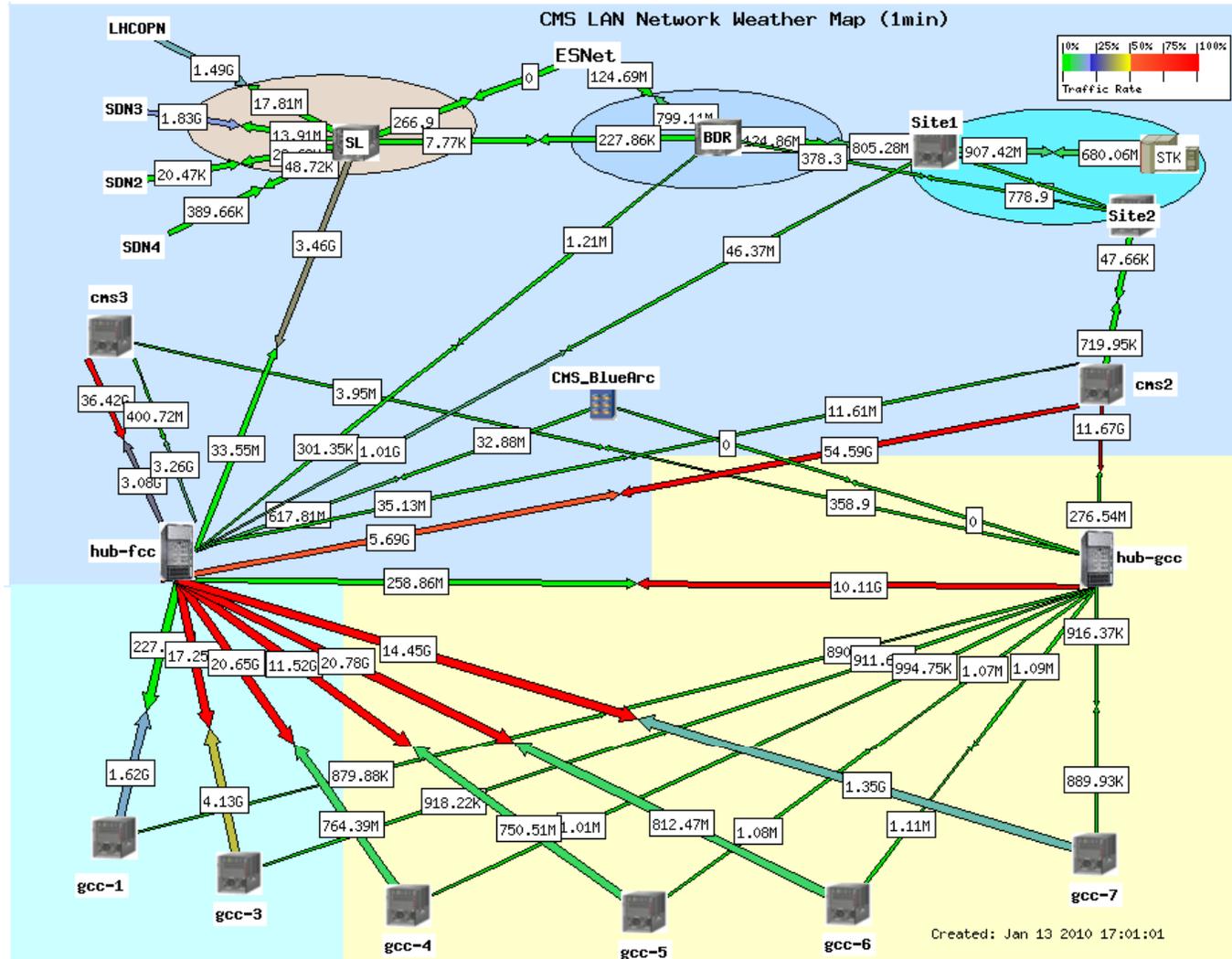
Host ↑↓	Service ↑↓	Status ↑↓	Last Check ↑↓	Duration ↑↓	Attempt ↑↓	Status Information
r-s-starlight-final	ASNet Circuit Status	OK	10-08-2010 07:39:49	2d 16h 15m 45s	1/5	IP SLA 27 # ASNet (ICMP probe to 198.151.133.181) status: OK (RTT=1 ms)
	ASNet-Direct Circuit Status	OK	10-08-2010 07:36:27	1d 14h 24m 7s	1/5	IP SLA 26 # ASNet-Direct (ICMP probe to 198.151.133.177) status: OK (RTT=1 ms)
	DE-KIT Circuit Status	OK	10-08-2010 07:36:04	2d 16h 14m 30s	1/5	IP SLA 21 # DE-KIT (ICMP probe to 198.151.133.141) status: OK (RTT=108 ms)
	ESNET-CHIPOP Circuit Status	OK	10-08-2010 07:36:42	2d 16h 13m 52s	1/5	IP SLA 32 # ESNET-CHIPOP (ICMP probe to 198.151.133.253) status: OK (RTT=1 ms)
	IN2P3-1 Circuit Status	OK	10-08-2010 07:37:57	2d 16h 12m 37s	1/5	IP SLA 23 # IN2P3-1 (ICMP probe to 198.151.133.153) status: OK (RTT=100 ms)
	IN2P3-2 Circuit Status	OK	10-08-2010 07:38:34	2d 16h 12m 0s	1/5	IP SLA 24 # IN2P3-2 (ICMP probe to 198.151.133.157) status: OK (RTT=116 ms)
	LHCOPN-BACKUP Circuit Status	OK	10-08-2010 07:39:37	2d 17h 25m 57s	1/5	IP SLA 17 # LHCOPN-BACKUP (ICMP probe to 192.16.166.29) status: OK (RTT=132 ms)
	LHCOPN-PRIMARY Circuit Status	OK	10-08-2010 07:40:15	2d 17h 36m 39s	1/5	IP SLA 16 # LHCOPN-PRIMARY (ICMP probe to 192.16.166.25) status: OK (RTT=120 ms)
	LHCOPN-SECONDARY Circuit Status	CRITICAL	10-08-2010 07:39:53	0d 8h 19m 41s	5/5	IP SLA 25 # LHCOPN-SECONDARY (ICMP probe to 192.16.166.5) status: Time Out (RTT=0 ms)
	MIT Circuit Status	OK	10-08-2010 07:39:12	2d 16h 11m 22s	1/5	IP SLA 19 # MIT (ICMP probe to 198.151.133.137) status: OK (RTT=23 ms)
	PURDUE Circuit Status	OK	10-08-2010 07:39:50	2d 16h 15m 44s	1/5	IP SLA 20 # PURDUE (ICMP probe to 198.151.133.161) status: OK (RTT=1 ms)
	Prague-D0 Circuit Status	OK	10-08-2010 07:39:30	0d 13h 1m 4s	1/5	IP SLA 29 # Prague-D0 (ICMP probe to 198.151.133.197) status: OK (RTT=120 ms)
	TIFR Circuit Status	OK	10-08-2010 07:38:05	0d 17h 17m 29s	1/5	IP SLA 31 # TIFR (ICMP probe to 198.151.133.249) status: OK (RTT=280 ms)
	UCSD Circuit Status	OK	10-08-2010 07:36:43	2d 16h 13m 51s	1/5	IP SLA 18 # UCSD (ICMP probe to 198.151.133.21) status: OK (RTT=60 ms)
	UFL Circuit Status	OK	10-08-2010 07:37:20	2d 16h 13m 14s	1/5	IP SLA 28 # UFL (ICMP probe to 198.151.133.185) status: OK (RTT=44 ms)
	UNL Circuit Status	OK	10-08-2010 07:39:30	2d 17h 51m 4s	1/5	IP SLA 35 # UNL (ICMP probe to 198.151.133.25) status: OK (RTT=16 ms)
	UTK Circuit Status	OK	10-08-2010 07:37:58	2d 16h 12m 36s	1/5	IP SLA 22 # UTK (ICMP probe to 198.151.133.145) status: OK (RTT=16 ms)
	UWISC Circuit Status	OK	10-08-2010 07:36:08	2d 8h 9m 26s	1/5	IP SLA 15 # UWISC (ICMP probe to 198.151.133.205) status: OK (RTT=4 ms)
	Ultralight Circuit Status	OK	10-08-2010 07:38:35	2d 16h 11m 59s	1/5	IP SLA 30 # Ultralight (ICMP probe to 198.151.133.217) status: OK (RTT=1 ms)

- Each circuit has an SLA monitor running icmp-echo application.
- Status of SLA monitor is tracked by SNMP

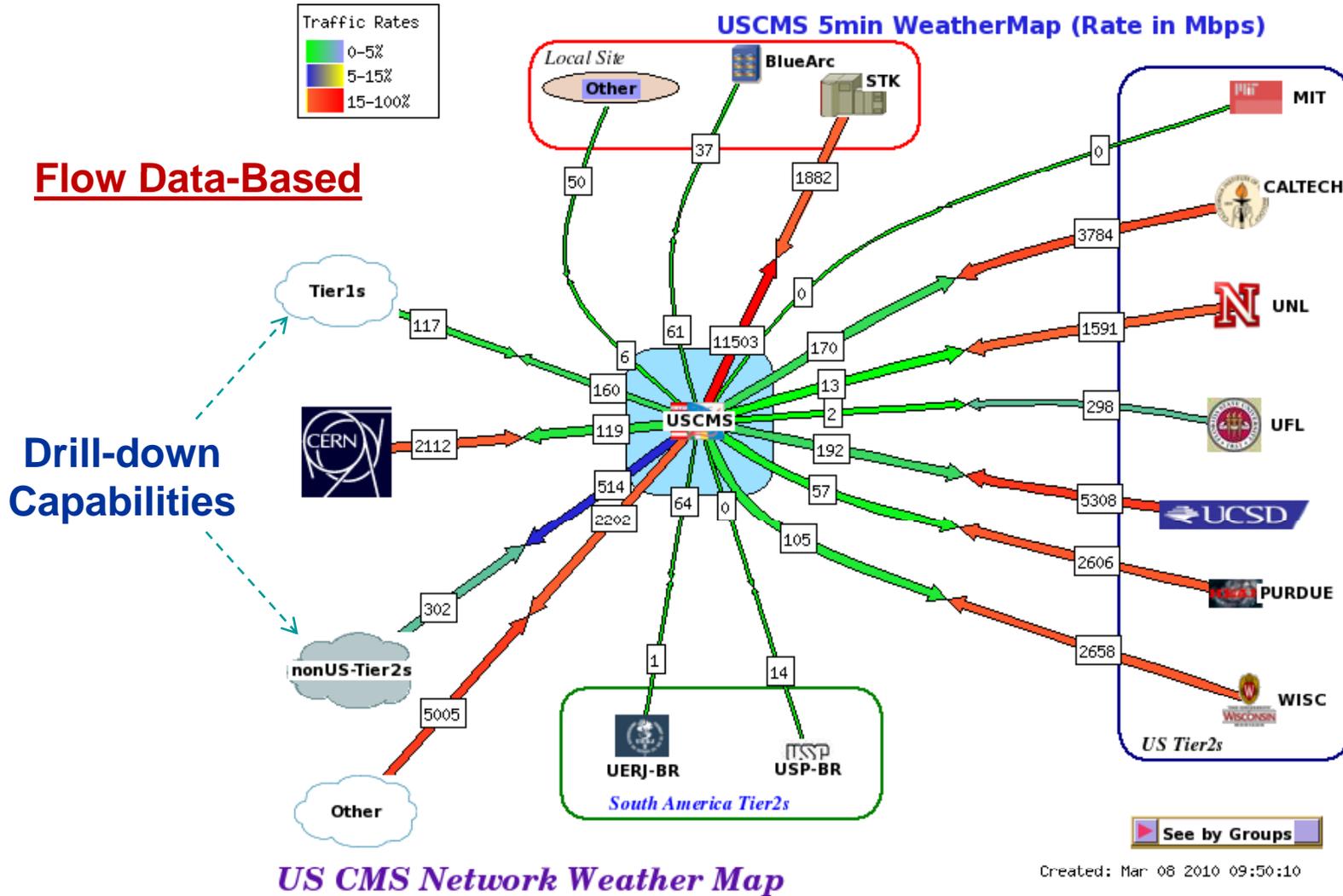


Weather Map Tools (I)

SNMP-Based



Weather Map Tools (II)



Work in progress:

- **New monitoring approaches**
 - Any-to-Any/cloud-to-cloud performance
 - Any production host can become an element of monitoring infrastructure
 - Combination of passive and active measurements
- **10GBase-T/IEEE 802.3an for end system**
 - Intel E10G41AT2 NIC PCI-Express
 - AristaNetworks DCS-7120T-4S (as ToR solution)
 - Directly to Nexus7K (via fiber at the moment)
 - Cat6E regular physical infrastructure deployed ~100m



Summary

- Two buildings, four computing rooms (fifth one is coming)
- Two Nexus 7000 switches for 10G aggregation, interconnected 80-160Gbps
- 2 x10Gbps to the Site Network (read/write data to tapes)
- 10Gbps to the Border Router (non-US Tier2s, other LHC-related traffic)
- 20Gbps toward ESNET CHIMAN and USLHCNET, SDN/DCN/E2E circuits



Summary (continued)

- ~200 dCache nodes with 2x1GE
- ~1600 worker nodes with 1GE
- ~150 various servers
- 2X 20G for *BlueArc NAS storage*
- C6509 access switches connected by 40-80GE
- Redundancy/loadsharing at L2 (vPC) and L3 (GLBP)
- IOS based Server Load Balancing for interactive clusters
- 19 SDN/DCN End-To-End Circuits
- Virtual port channelling (vPC)
- QoS, 5 major classes of traffic



Additional slides



USCMS Tier1: Data Analysis Traffic

