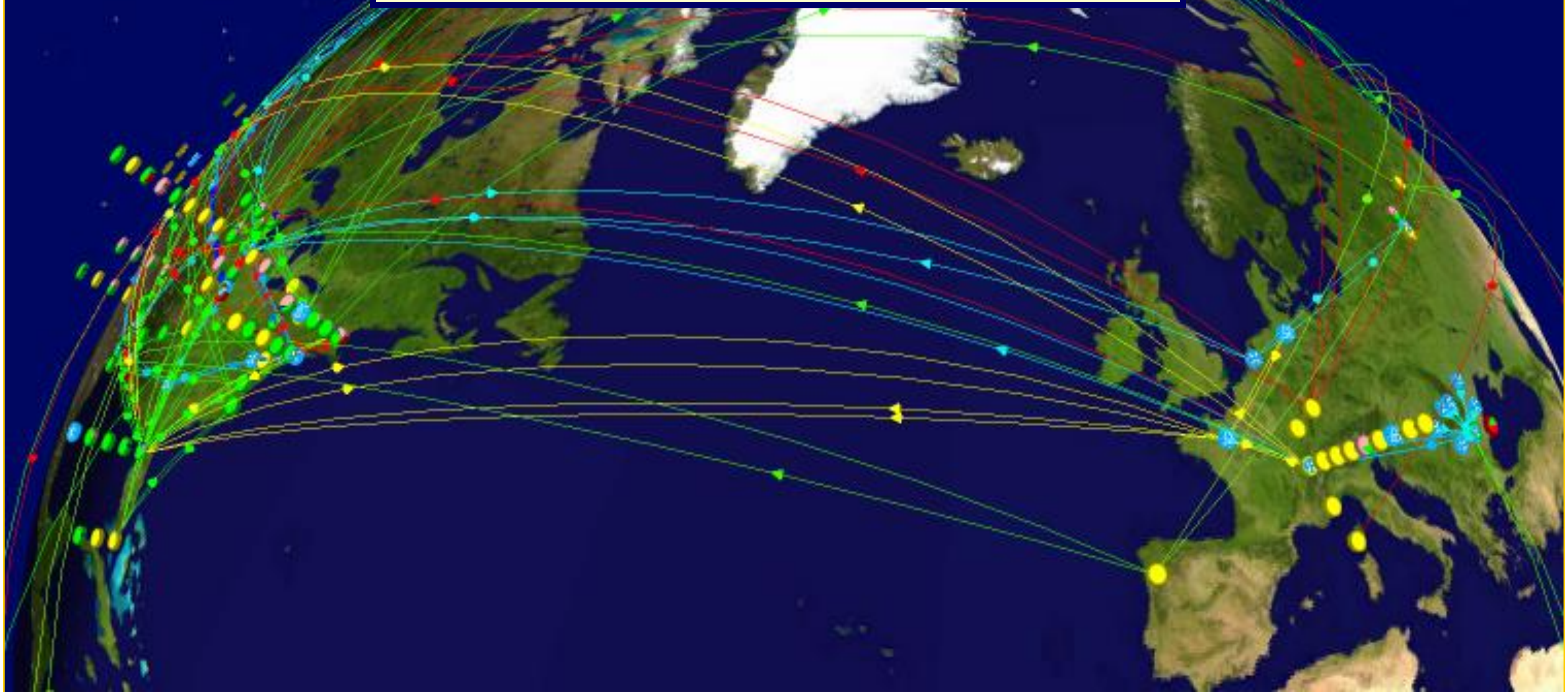
US LHCNet

Harvey Newman
California Institute of Technology

Transatlantic Network Workshop
CERN, June 10, 2010



US LHCNet

- Transatlantic mission-oriented network managed by Caltech and CERN
- Funded by DOE/OHEP with contribution from CERN
- Program: to provide resilient, cost-effective transatlantic networking adequate to support the principal needs of the LHC physics program, with a focus on the US
  - ➔ In partnership with ESnet, Internet2, NLR, SURFnet, GEANT and the NRENs in Europe
- Primary mission: to provide highly reliable, dedicated, high bandwidth connectivity between the US Tier1 centers and CERN [Uptime goal: 99.9+%]
- Further, to support high bandwidth traffic flows between US LHC Tier1 and European Tier2 centers as well as between US Tier2 centers and European Tier1s
- Development, deployment and integration of advancing network and high throughput technologies, to meet the advancing needs

US LHCNet Network

- **US LHCNet is a Multi-vendor, Multi-layer network**
  - ➔ Path-diverse transatlantic links on (currently) five undersea cables, with terrestrial interconnects in the US (NY – CHI) & Europe (GVA – AMS)
  - ➔ Core Equipment: Ciena optical muxes, Force10 switch routers
  - ➔ Offering Layer 1, 2, 3 resilient services to the users
- **A Real-time System designed for Non-stop Operation**
  - ➔ Real-time systems for monitoring and some automated operations
  - ➔ A carefully managed set of virtual circuits with automated fallback provides graceful degradation in case of single or multiple outages
- **US LHCNet NOC:**
  - ➔ 24x7x365 Coverage (on-call 3-line support); Office hours in CET, PDT
  - ➔ Distributed NOC (main locations: CERN/Geneva, Caltech/Pasadena)
  - ➔ A small, talented team (4 engineers) with full range of skills
  - ➔ NOC engineers able to perform tasks from any location world-wide
- **Equipment and US PoP diversity to mitigate effects of equipment or site outages**

Working Methodology

US LHCNet

Production Network

Develop and build next generation networks

High performance
High bandwidth
Reliable network

Pre-Production

- Transatlantic testbed
- Lightpath technologies: DCNSS, OSCARS, DRAC, AutoBAHN
- New transport protocols; Interface & kernel settings
- DICE / Ultralight / λStation / Terapaths; Vendor Partnerships

HEP & DOE Roadmaps

Testbed for Network Services Development

Networks for HEP Research

- LHC and Other Experiments; LHC OPN
- GRID applications: WLCG, DISUN, OSG
- Interconnection of US and EU Grid domains
- EVO …

R&D efforts tailored for the specific needs of the HEP community, with direct feed-back into the high-performance production network.

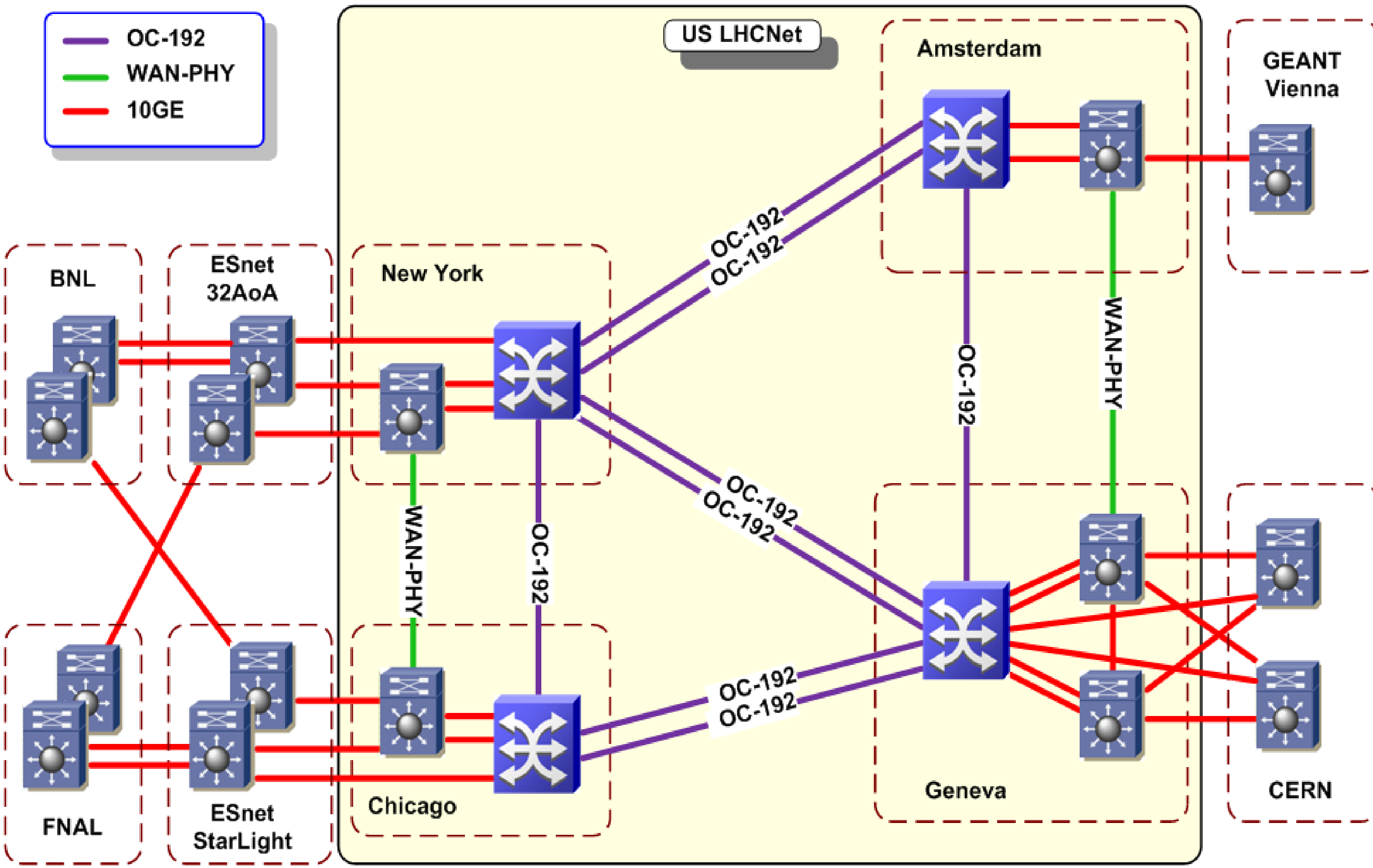
TA Link Capacity vs. "Bandwidth"

US LHCNet

- Various Colloquial definitions/uses of "Bandwidth"
  - ➔ Need to better define capacity to set requirements and roadmaps
- In reality: what is referred to as "10Gbps" provides data rates lower than the link capacity (9.4 Gbps) !
- Application data encapsulated in several layers of network/protocol overhead:
  - ➔ SONET overhead (fixed)
  - ➔ Ethernet overhead (fixed per frame, depends on MTU)
  - ➔ IP (fixed per packet, depends on fragmentation (MTU))
  - ➔ TCP (fixed overhead per frame)
  - ➔ FTP / HTTP / …
- Additional inefficiencies due to e.g. TCP Stack and Tuning (especially for long RTT), Transfer-applications, Schedulers, etc.
- Achievable application data rate is always less than link capacity.
- "10G" SONET link $\Rightarrow$ ~8 Gbps data throughput rate; usually less!

USLHCNet in 2010

Non-stop Operation; Circuit-oriented Services



Core: Optical multiservice Switches [*] that provide resilience

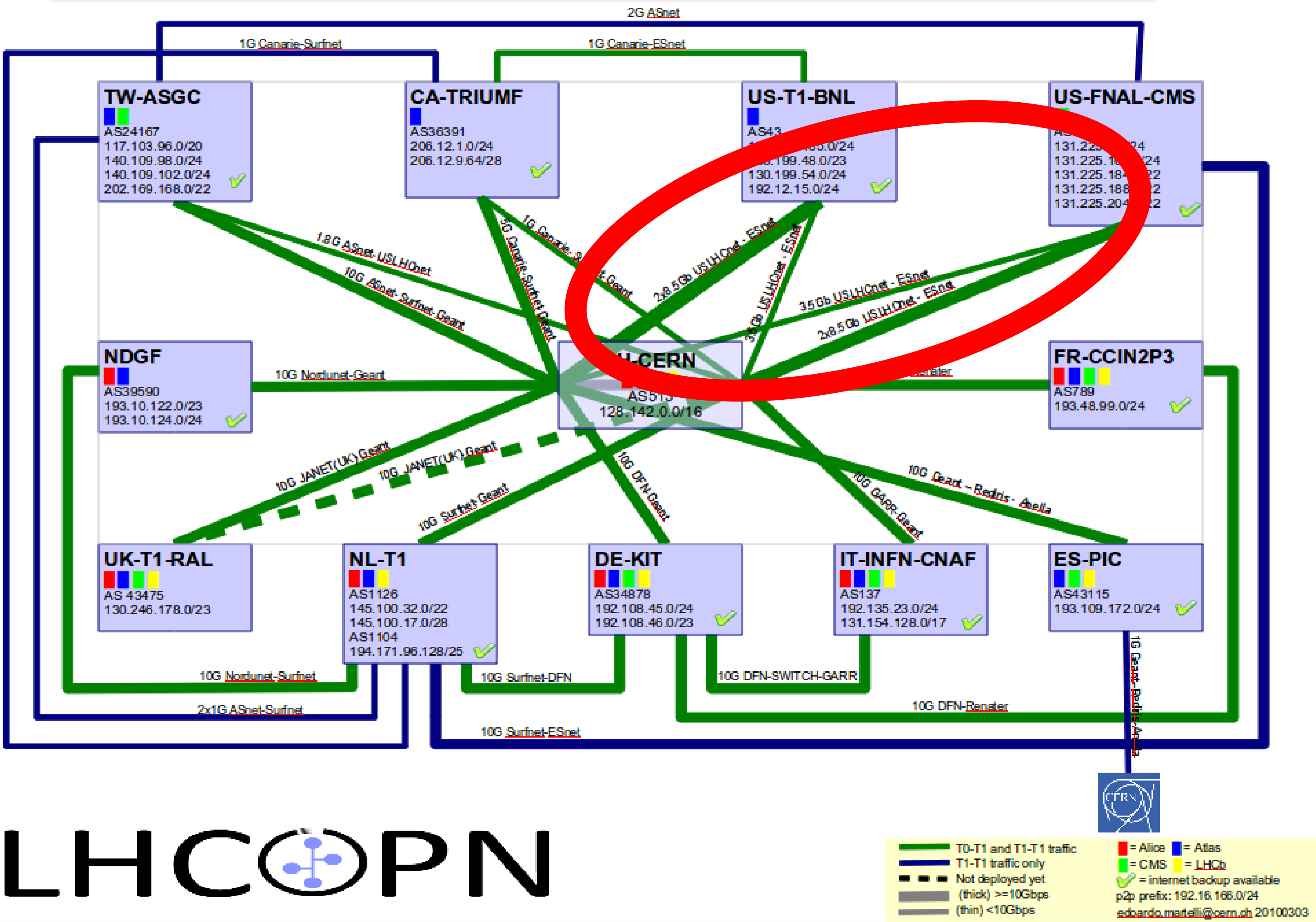
Performance enhancing Standard Extensions: VCAT, LCAS

USLHCNet, ESnet, BNL & FNAL: equipment and link redundancy

[*] Dynamic **circuit-oriented network services with BW guarantees, with robust fallback at layer 1: Hybrid optical network**

US LHCNet: An Integral Part of the LHCOPN

"A Network Within a Network"

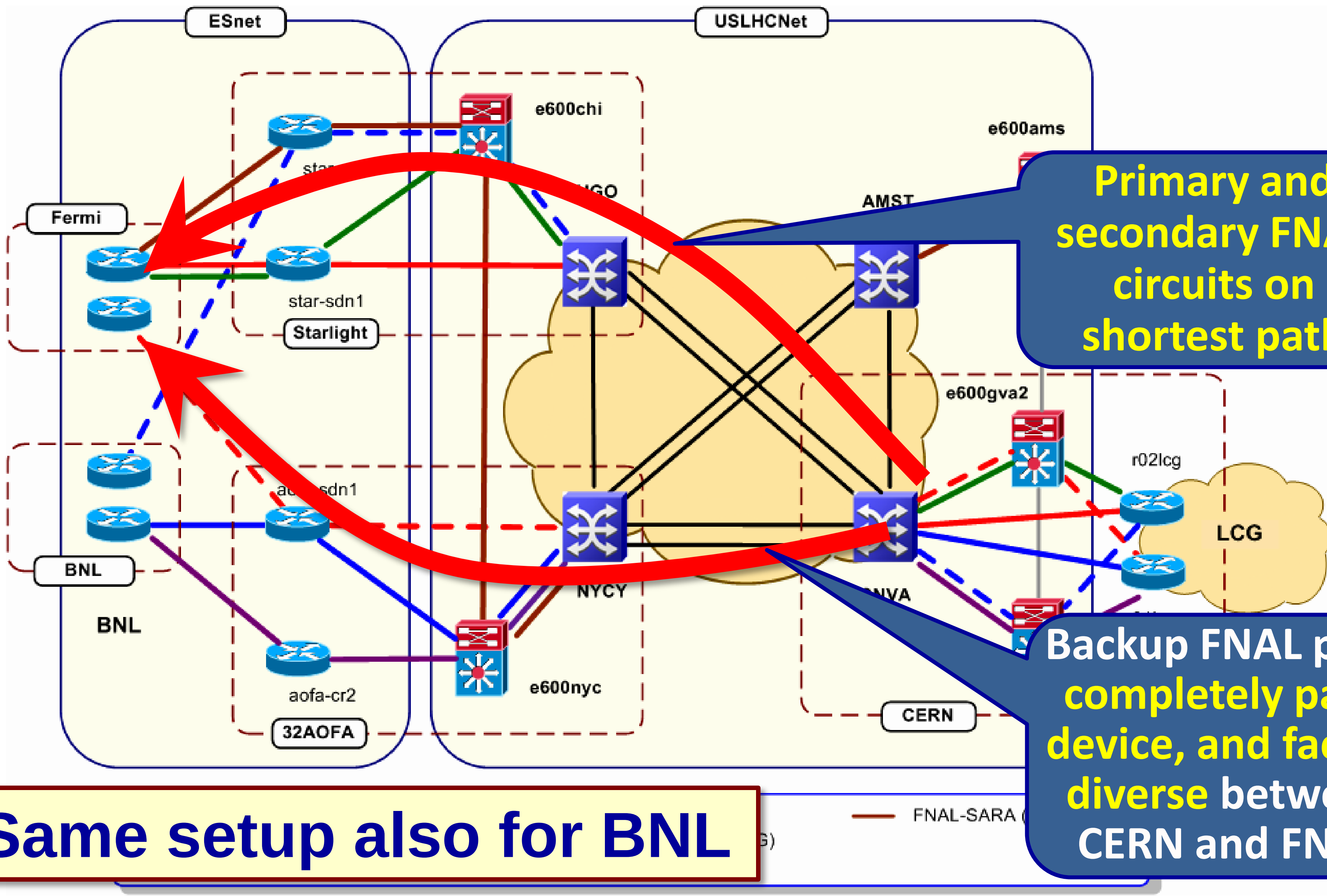


LHCOPN

Bandwidth Allocation

- **Primary Mission (virtual) Circuits**
  US LHCNet provides to each US Tier1:
  1. **One primary circuit (8.567 Gbps, or STS-3c-57v)**
  2. **One secondary circuit (8.567 Gbps, or STS-3c-57v)**
  3. **One explicit backup circuit (3 Gbps guaranteed, expandable up to 4.1 Gbps)**
- **Other Virtual Circuits**
  - ➔ **ESnet-GEANT peering support (4.810 Gbps, STS-3c-32v)**
  - ➔ **Dedicated FNAL-DE-KIT virtual circuit (1.050 Gbps, STS-3c-7v)**
- **US Tier1 Circuits are Protected in US LHCNet**
  - ➔ **Single link outage is transparent to the LHCOPN**

US LHCNet + ESnet: Redundant Tier1 Paths



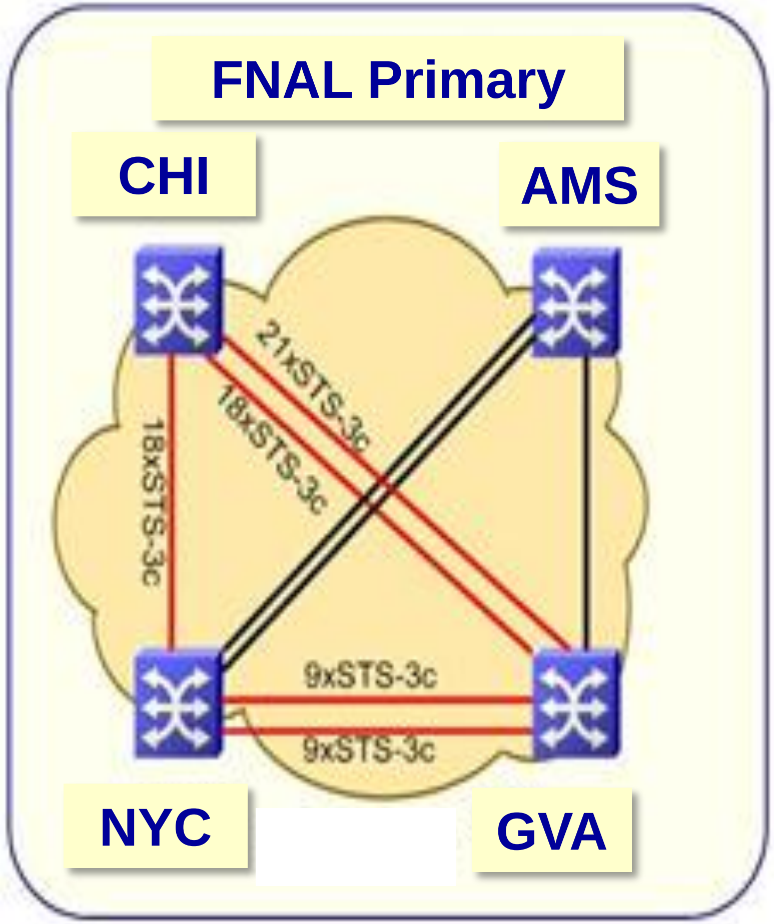
Primary and secondary FNAL circuits on shortest path

Backup FNAL path, completely path, device, and facility diverse between CERN and FNAL

Same setup also for BNL

Increasing Efficiency: Advanced protocol features

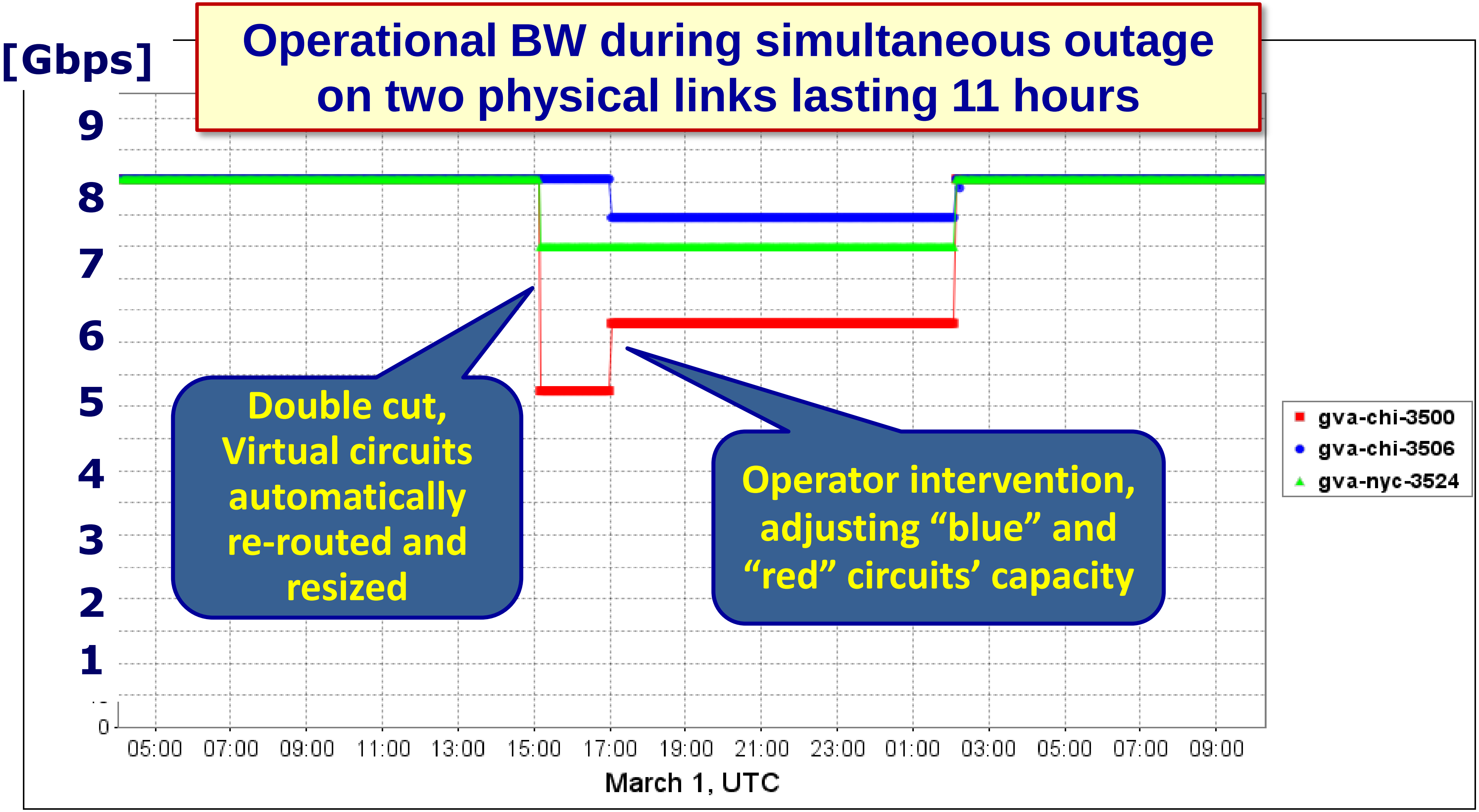
- Mesh protection using Ciena OSRP
- VCAT: Virtual Circuits Across Multiple Links
  - ➔ Only a fraction of a virtual circuit is affected by an outage
  - ➔ End-sites see only lower capacity in case of a link outage



- LCAS: Dynamic VC Adjustment
  - ➔ VC capacity-adjustment leads to a “smaller hit” during extended outages

VCAT/LCAS in use

- Provides hit-less capacity adjustment



Granular Bandwidth Allocation

US LHCNet

Available Capacity is Divided in Virtual Circuits

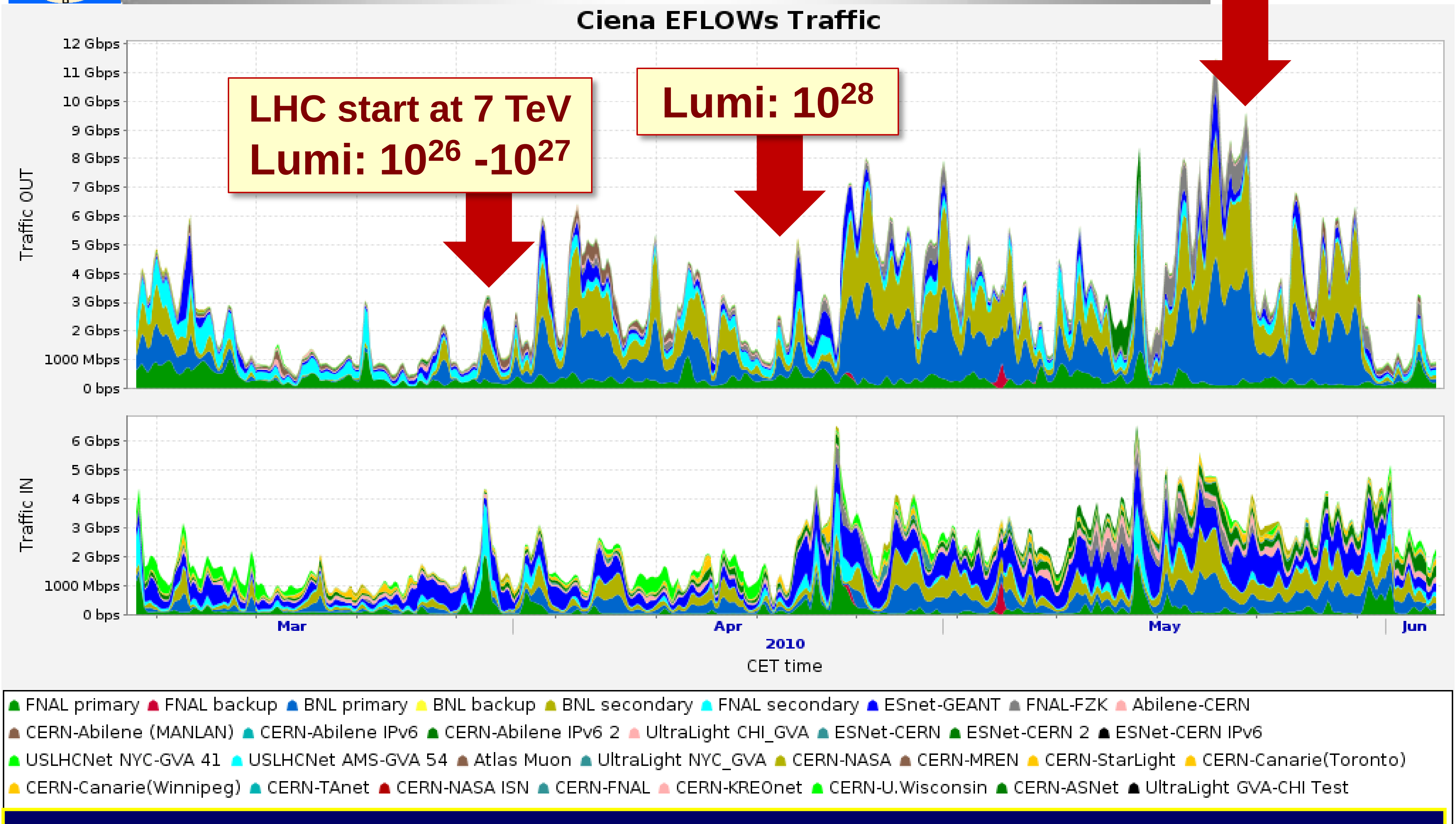
Purpose		Endpoint A	Endpoint B	Allocated Bandwidth [OC-192 links]	Allocated Bandwidth [Gbps]
Tier0-Tier1 Tier1-Tier1 (primary, secondary)	CERN-FNAL	Geneva	Chicago	2×0.9	2×8.567
	CERN-BNL	Geneva	New York	2×0.9	2×8.567
	FNAL-FZK	Chicago	Amsterdam	0.1	1.050
Tier1-Tier2	ESnet-GEANT peering	New York	Amsterdam	0.5	4.810
	Internet2-GEANT peering	New York	Amsterdam	0.3	3.156
Tier1 backup, GPN and other peerings	GPN / FNAL backup	Geneva	New York	0.4	4.208
	GPN / BNL backup FNAL-TIFR	Geneva	Chicago	0.4	4.208
TOTAL ALLOCATION				5.3	51.700

High Service Availability

SIMULTAN-EOUS No. of Failed TA links	Effect on US Tier1 services (primary and secondary)	Effect on Tier2 and other unprotected services	Expected average duration within one year
1 Link	No impact, service protected 16.8 Gbps operational per Tier1	Degraded, operational	22 Days/year
2 Links	Degraded, available bandwidth per Tier1: 9.4 – 16.8 Gbps	Degraded OR not operational	6 Days/year
3 Links	Degraded, at least 8.4 Gbps bandwidth available per Tier1	Degraded OR not operational	< 1 Hour/year (8 minutes/year observed)

Small amount of protection capacity in US LHCNet is enough to protect its highest priority services against single link outages

Flow Based Traffic Statistics

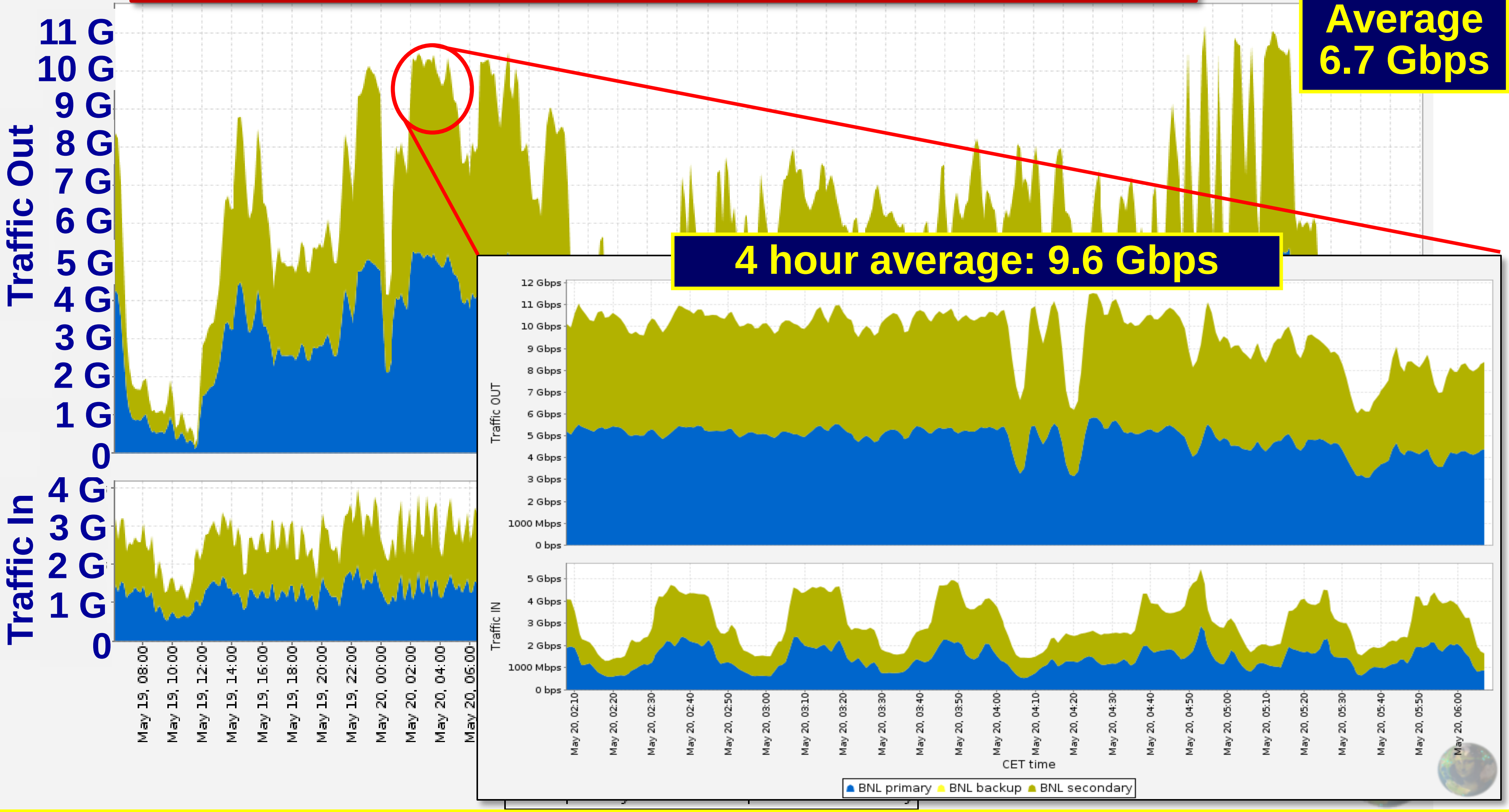


Clearly visible correlation between luminosity and data rates

BNL Traffic: May 19-22

Extended periods of high network utilization

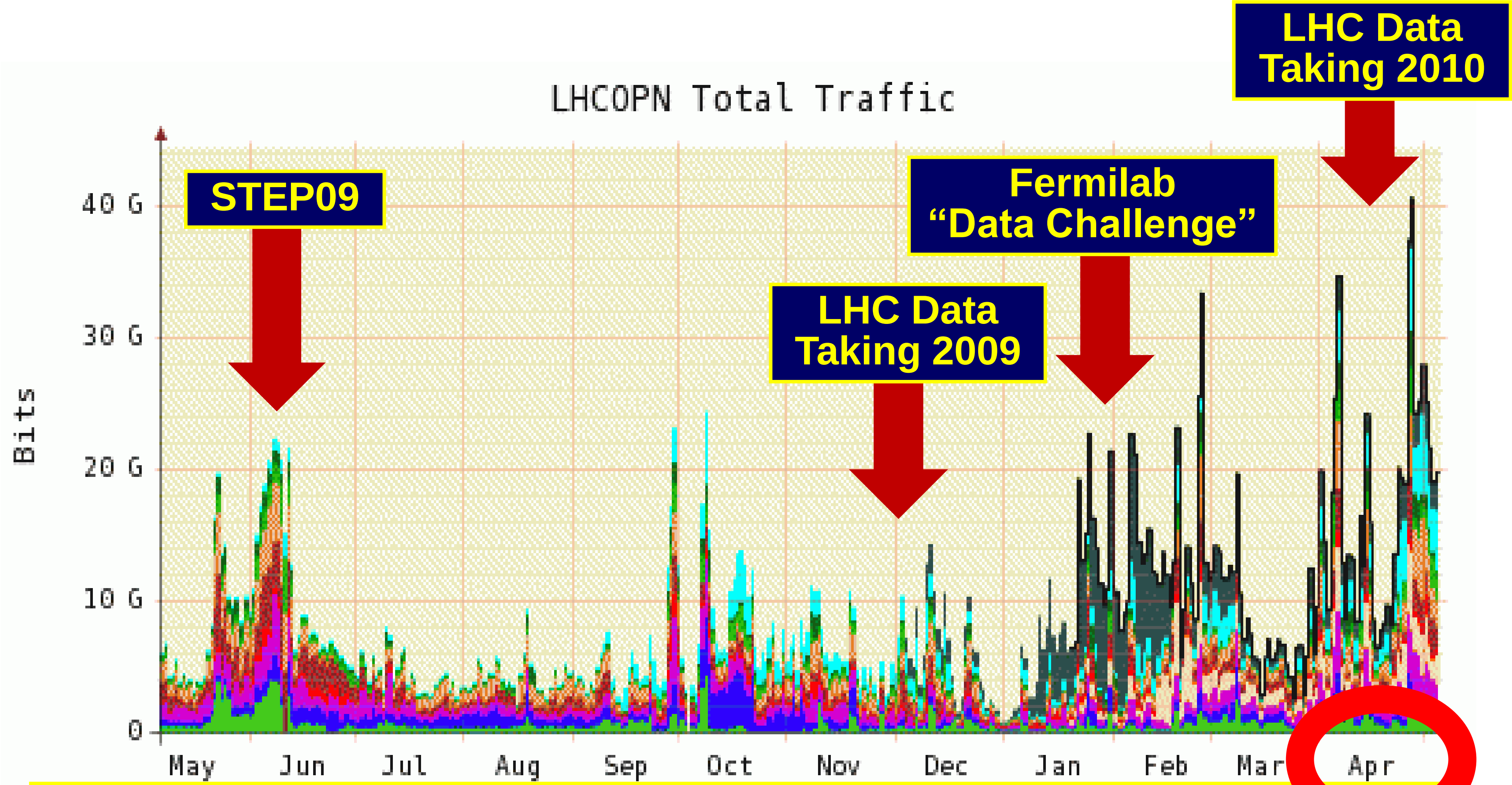
4 Day Average 6.7 Gbps



Ceiling observed at ~11 Gbps: due to end-system limitations ?

LHCOPN 2009/10 Statistics

US LHCNet



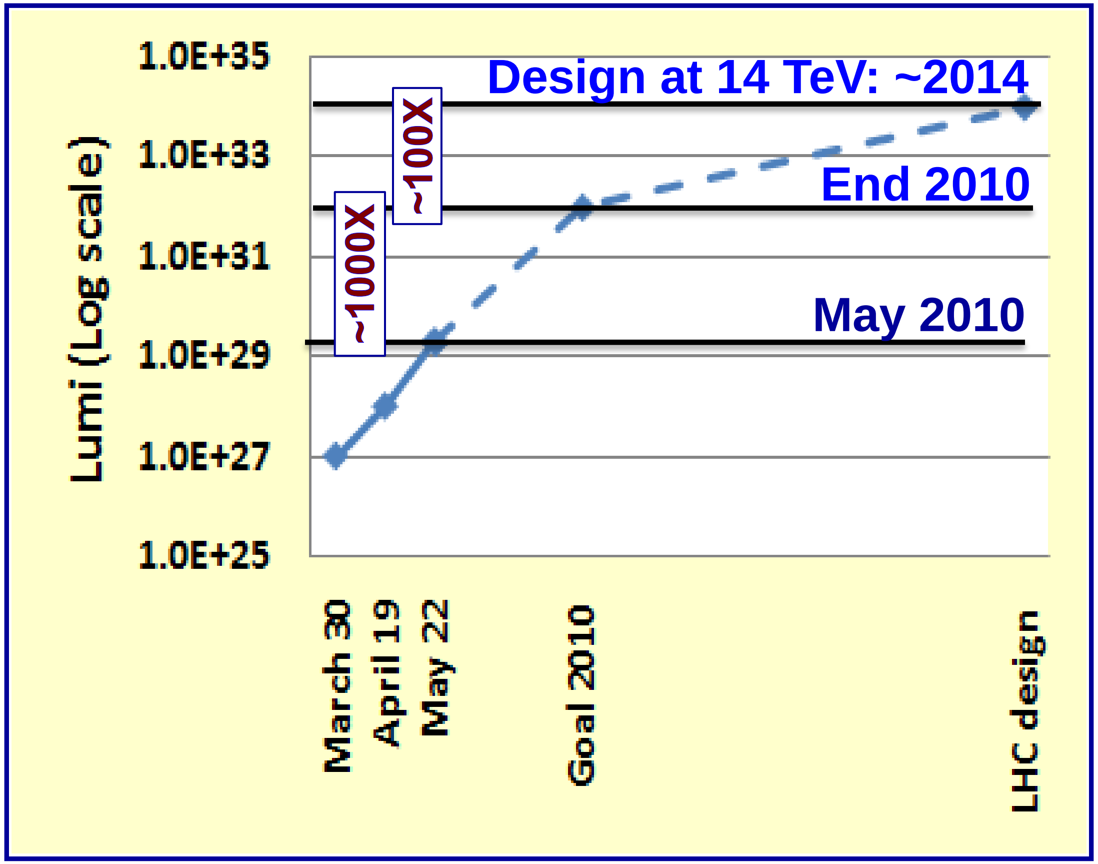
LHC Luminosity still ramping up, but observed data rates well above previous estimates !

Future LHC Data Rates?

- March 30, 2010: LHC started at 7 TeV collision energy
  - ➔ Low Luminosity, 10^{26} to 10^{27}, i.e. Low Data Rates
- April 19, 2010: 10-fold increase in luminosity (10^{28})
- May 22: another 10-fold increase (10^{29})
- Goal for 2010: 10^{32} !
- [LHC design lumi: 10^{34}]

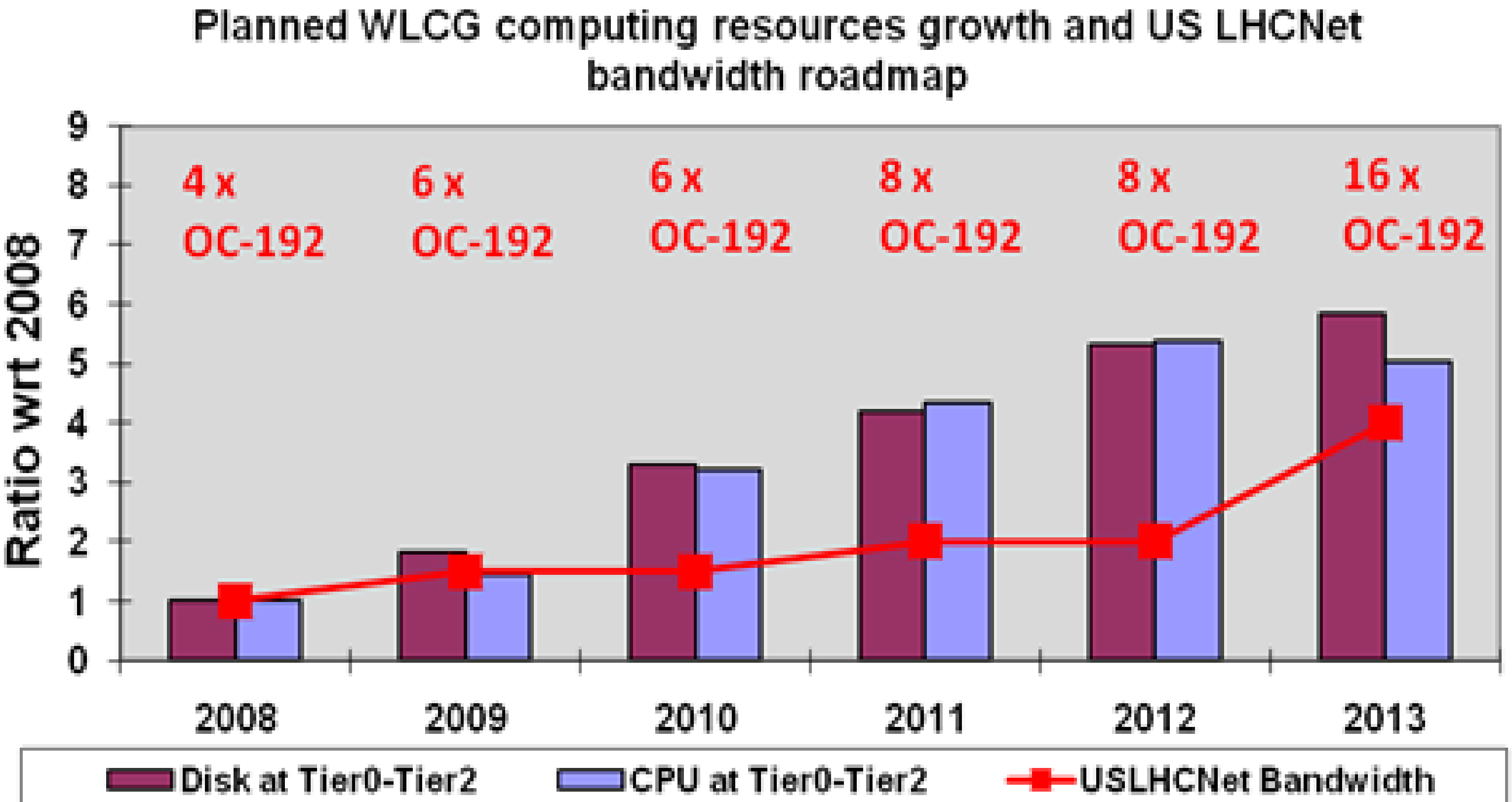
There is still a 3 order of magnitude improvement expected **THIS YEAR**

But we need to understand how it will translate into network utilization!



USLHCNet 2008-13 Bandwidth Roadmap
Versus WLCG CPU and Disk Storage Roadmap

US LHCNet



40G in 2008
to
400G in 2015

10X in 7 Yrs.

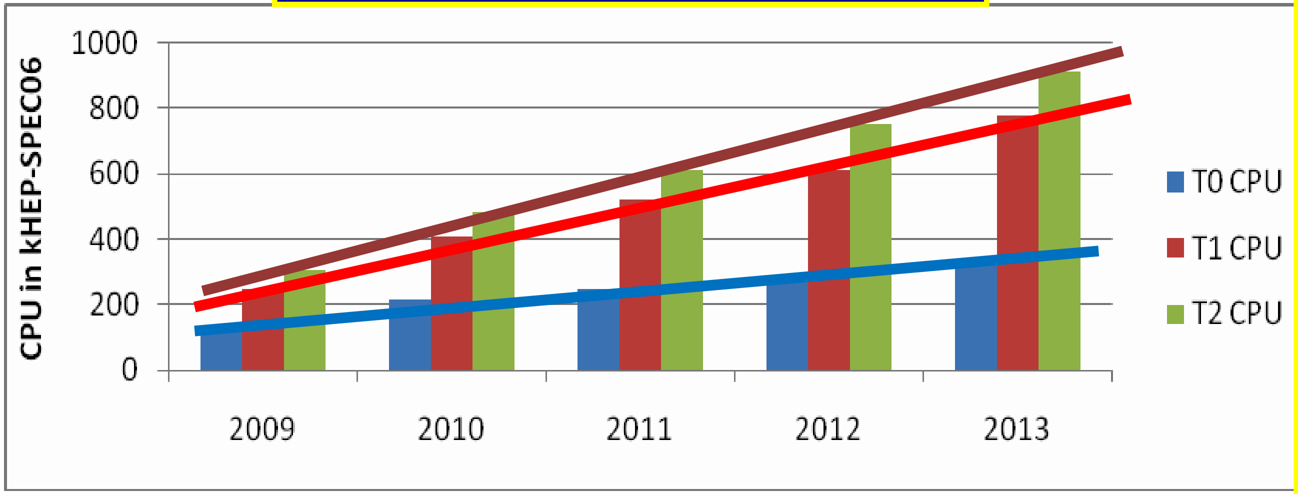
Slower **than historical trends [~50X]**

	2008	2009	2010	2011	2012	2013	2014	2015
Disk at Tier0-Tier2 (PB)	39	70	127	162	206			
CPU at Tier0-Tier2 (kHEP-SPEC06)	400	577	1281	1733	2158			
USLHCNet Bandwidth (OC-192s)	4	6	6	8	12	16	28	40

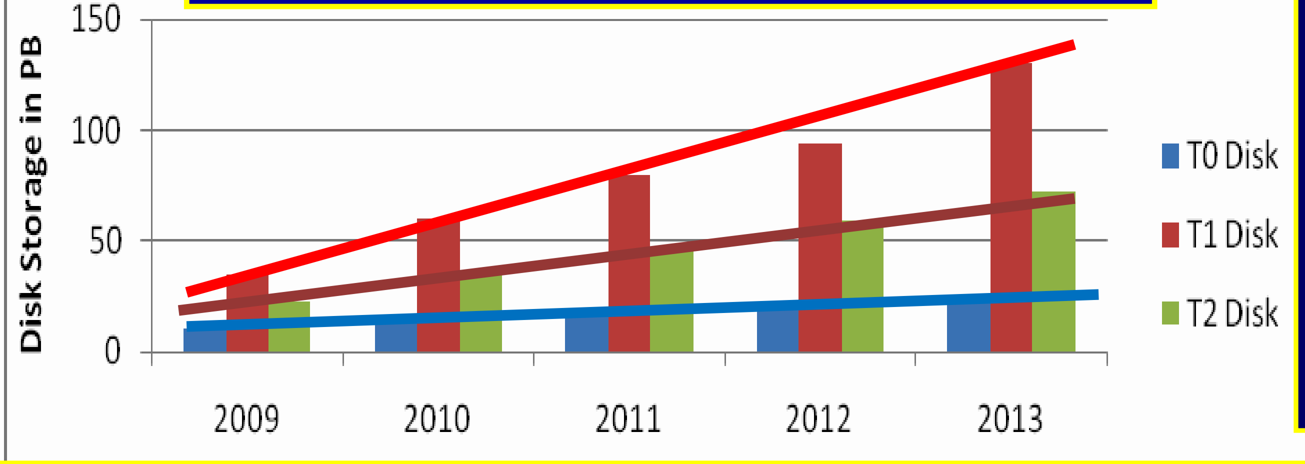
US LHCNet BW Roadmap shifted to respond to delay in LHC Startup
Outlook: now lags behind disk storage and CPU roadmap

WLCG CPU and Disk Storage Roadmap: Tier0 vs Tier1s vs Tier2s

WLCG CPU Resources Growth



WLCG Disk Resources Growth



More Significant CPU and Disk Increase At Tier1s and Tier2s

➡ Increased Reliance On, and Need for Additional National, Regional, and Transatlantic Network Resources

http://lcg.web.cern.ch/LCG/Resources/WLCGResources-2009-2010_24FEB10.pdf

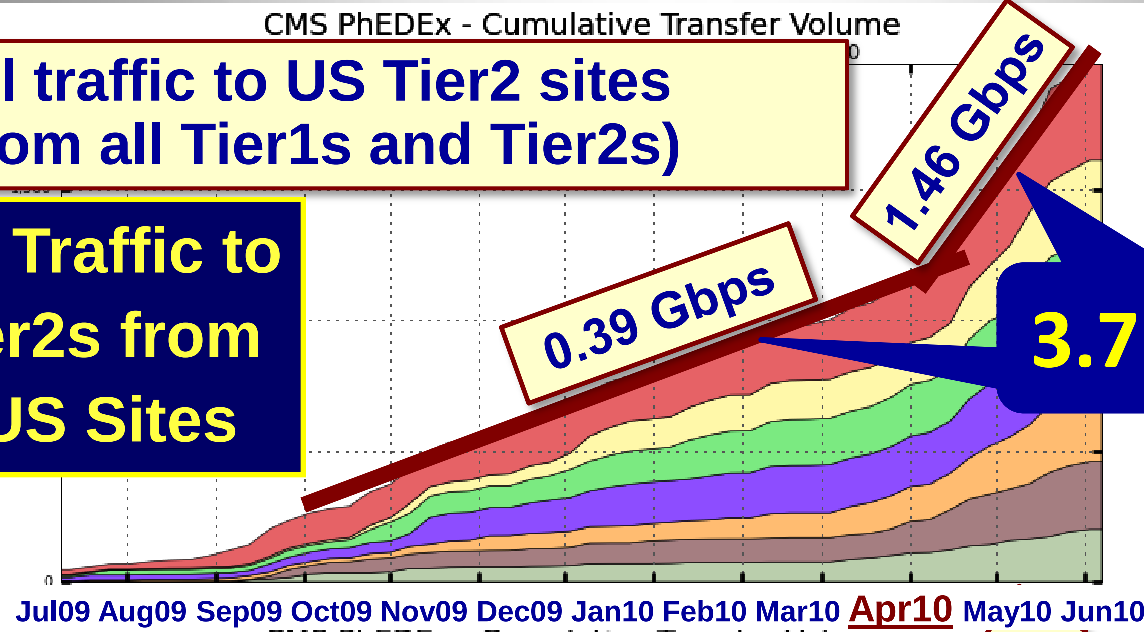
Do Tier2s Matter for Transatlantic Networks?

US CMS Tier2s
US National vs Transatlantic Traffic

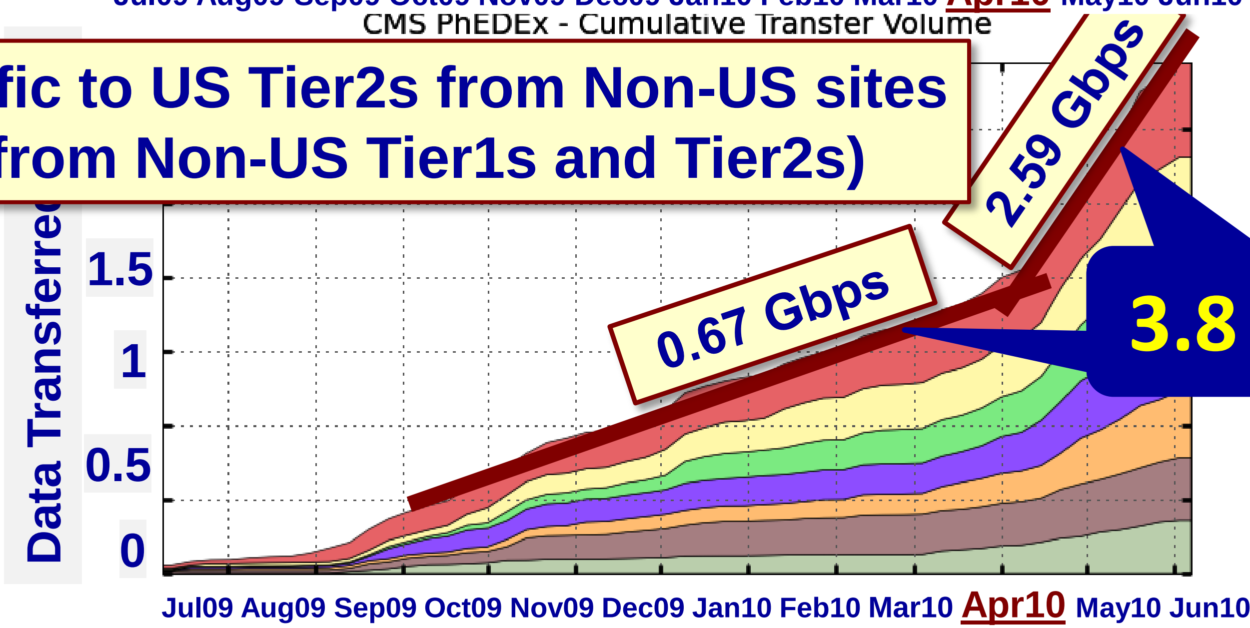
US LHCNet

All traffic to US Tier2 sites (from all Tier1s and Tier2s)

44% of Traffic to US Tier2s from Non-US Sites



Traffic to US Tier2s from Non-US sites (from Non-US Tier1s and Tier2s)



<u>Outlook:</u> More TA Traffic From Tier2s

Hope for Insights from This Workshop

Scheduling Network Resources in US LHCNet (a.k.a. Dynamic Circuits)

Dynamic Resource Allocation for HEP

- US LHCNet has deployed the Internet2 DCN Software Suite since April 2008 as part of our R&D efforts
- Dynamic circuit allocation optimizes bulk data transfers
  - ➔ Guaranteed bandwidth between source-destination sites for a requested time period
  - ➔ Traffic isolation
  - ➔ Predictable transfers
  - ➔ User/application control [eventually also system-level controls]
- Several US Tier2s as well as Tier1s are reachable via Internet2 ION and ESnet OSCARS Services
- Works with LambdaStation (CMS) and Terapaths (Atlas)
- US LHCNet is collaborating with European partners to bring dynamic resource allocation also to LHC sites in Europe
  - ➔ Which are the European LHC (Tier1/Tier2) sites willing to connect ?

Managing Storage & Network Resources
US LHCNet Progressive Approach

Hadoop: Distributed File System

FDT: High-performance data transfer tool

High performance data transfers between LHC sites

MonALISA: Distributed Monitoring Framework

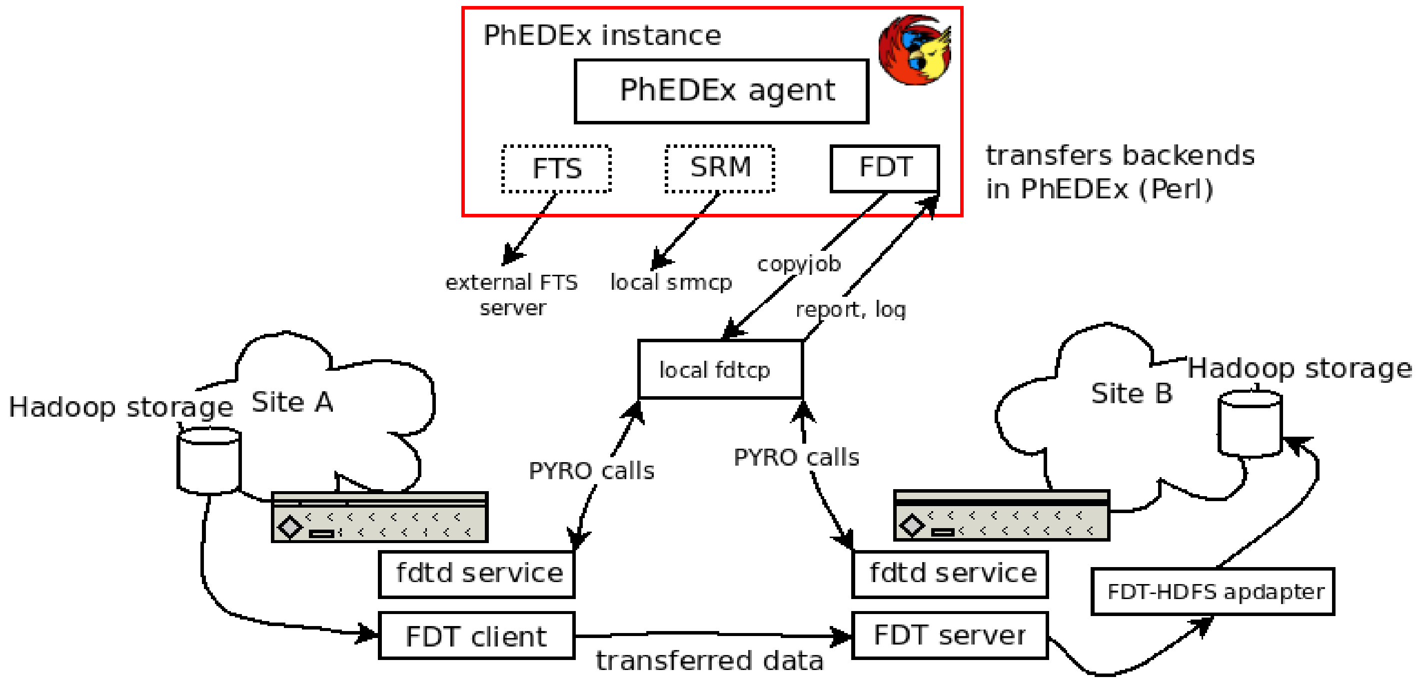
DCNSS: Dynamic Circuit capability

ION: DCN-based service in Internet2

Managed high performance data transfers between LHC sites over user-controlled, dedicated infrastructure

FDT-PhEDEx Integration

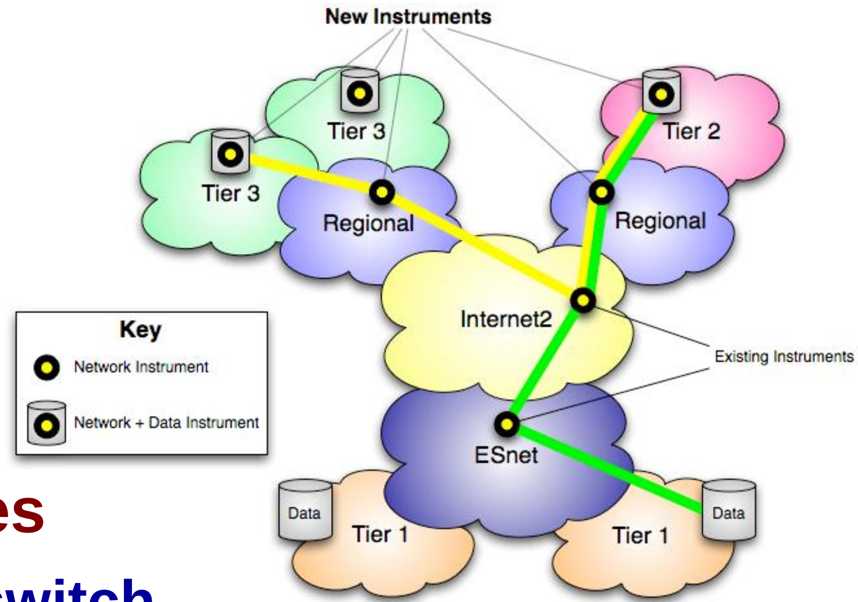
- **FDT uses the IDC API to request dedicated bandwidth between end-systems for the duration of a bulk transfer**
  - **Demonstrated at GLIF 2009 conference in Daejeon**
- **Work ongoing on integrating FDT as transfer tool in PhEDEx**
- **Will allow PhEDEx to transparently reserve network resources in the future, leading to truly managed transfers**

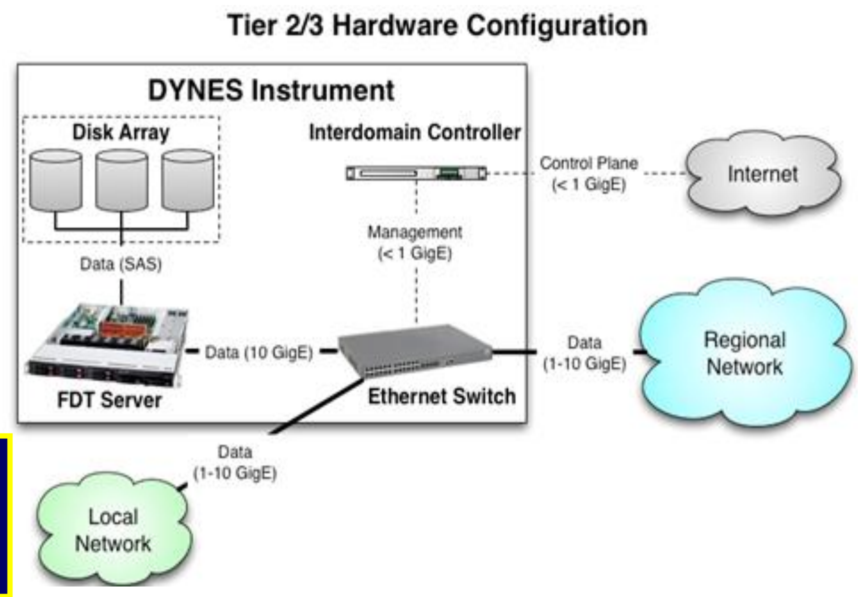


DYNES Project (US NSF)
Internet2, Caltech, Michigan, Vanderbilt

US LHCNet

- **Goal: Connect US LHC sites to the Dynamic Circuit Infrastructure**
  - ➔ Deploy enabling hardware at
    - ❐ ~39 campuses (LHC Tier2 and Tier3 sites) in the US
    - ❐ Involving ~16 regional networks
- **Deploy ION-enabled software at sites**
  - ➔ IDC server, FDT server + Ethernet switch
- **Support dynamic circuit operation and integration with higher-level tools through provided API**
  - ➔ Terapaths, LambdaStation, PhEDEx, …





Scheduling for optimal use of limited dynamic resources: follow-on project

Architectural Considerations

Transatlantic Issues

High-Availability solutions require multiple links with carefully planned path redundancy

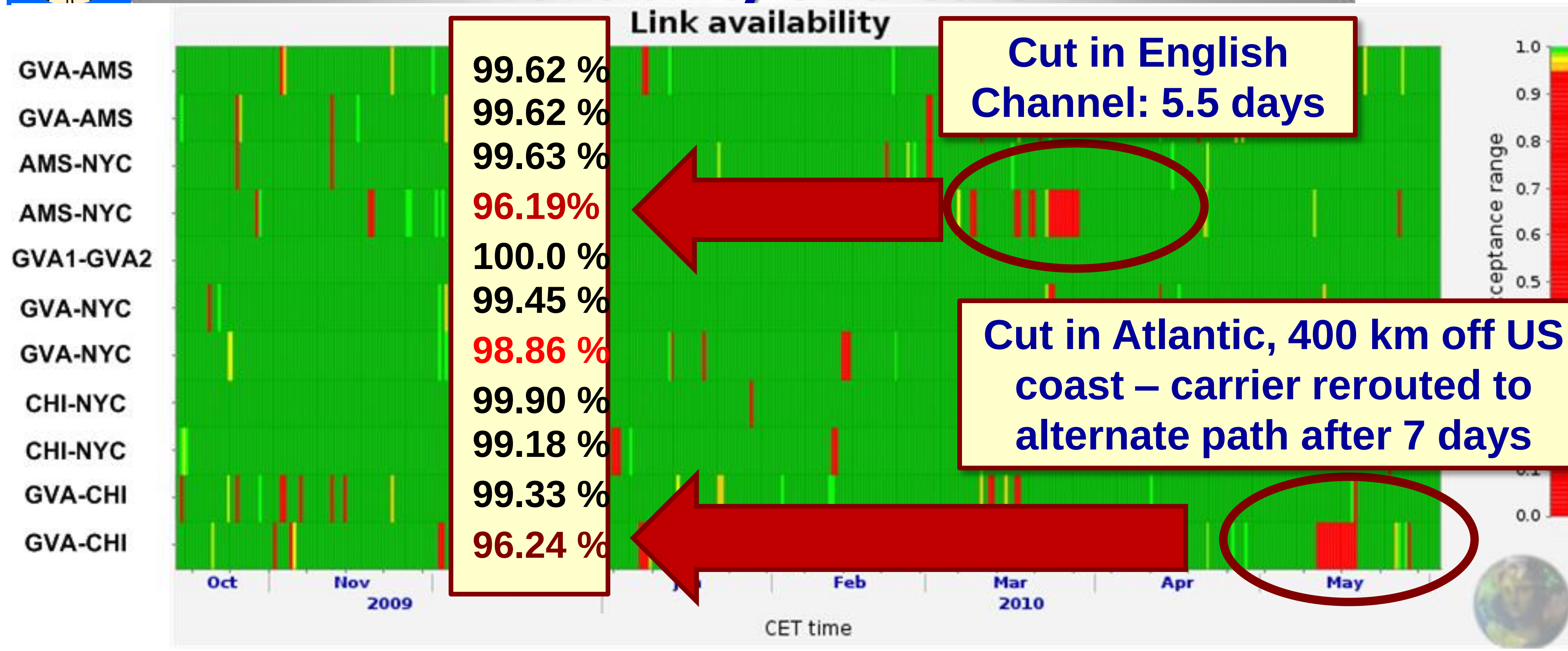
- TA links more complex than purely terrestrial ones
  - ➔ Longer distance - more fibre, more equipment
  - ➔ Typically constructed from segments from multiple owners
- Submarine segments: hostile environment, hard access

Comparative Remarks on Outages

- Terrestrial spans: more frequent, shorter TTR
  - ➔ "Easier access" for repairs; but also for diggers, copper thieves,…
  - ➔ Complex equipment – from amplifiers to add/drop multiplexers
- Submarine segments: less frequent, much longer TTR
  - ➔ "Difficult access" ➡ longer repair time; Potential hazards: ship anchors, trawlers, geological events, sharks
  - ➔ Repair speed depends on Time to Arrival of repair fleet, problem location, weather conditions

US LHCNet Transatlantic Link

Availability and Cuts



- NB: a single submarine cut can reduce availability significantly !
- Two options to provide real robustness:
  - Buy protected circuits (expensive)
  - *Construct* protected services from unprotected elements
- Cost and efficiency mandates the second approach!

Technology Choices

- Current US LHCNet design matches well LHC requirements
- Best Practices Guidelines:
  - ➔ E.g. "Switch where you can, route where you must"
- We started evaluation of the next generation architectural design
  - ➔ Evaluating new and emerging technologies and standards: functionality and features, performance and cost
  - ➔ Fit new developments with future requirements
  - ➔ In Collaboration with partners (ESnet, Internet2, NLR, SURFnet, …)

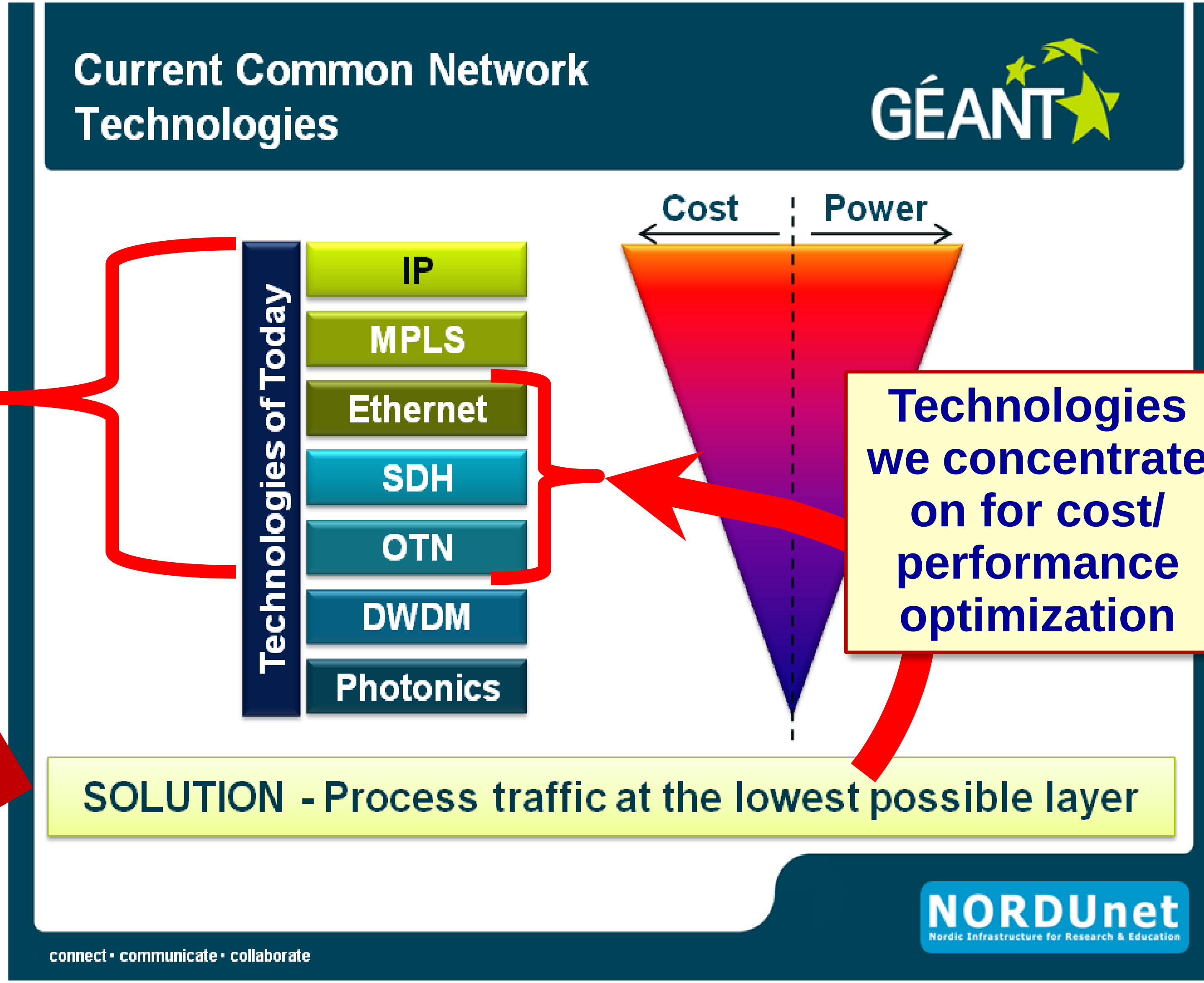
Technology Choices

US LHCNet

Tony Breach, GN3 JRA1: Presented at TNC2010 Conference Last week

Of main interest for deployment in US LHCNet

!



Technology Roadmap

- **Main requirement: Fast and Robust Protection Switching**
  - ➔ Candidate technologies: SONET, OTN, Carrier Ethernet (G.8031)
- **Upgrade to 40G/100G**
  - ➔ Ethernet: 40 and 100Gb Ethernet will already be available in 2010
  - ➔ Transport 40G (OTN-3) is already available and deployed in some networks – transatlantic by 2011/2012
  - ➔ Transport 100G: transatlantic routes probably by 2013/2014
- **Main Challenge: Cross the Atlantic**
- **End-to-end Dynamic Circuit Support**
  - ➔ Continued use of Dynamic Circuit Control Plane
  - ➔ Build-out of dynamic network resources on both sides of the Atlantic: work with partners (DICE, SURFnet, …)
  - ➔ Connecting to Tier2 sites: collaborate with regional network and local site admins (Expanded direction in DYNES)

US LHCNet Bandwidth Roadmap

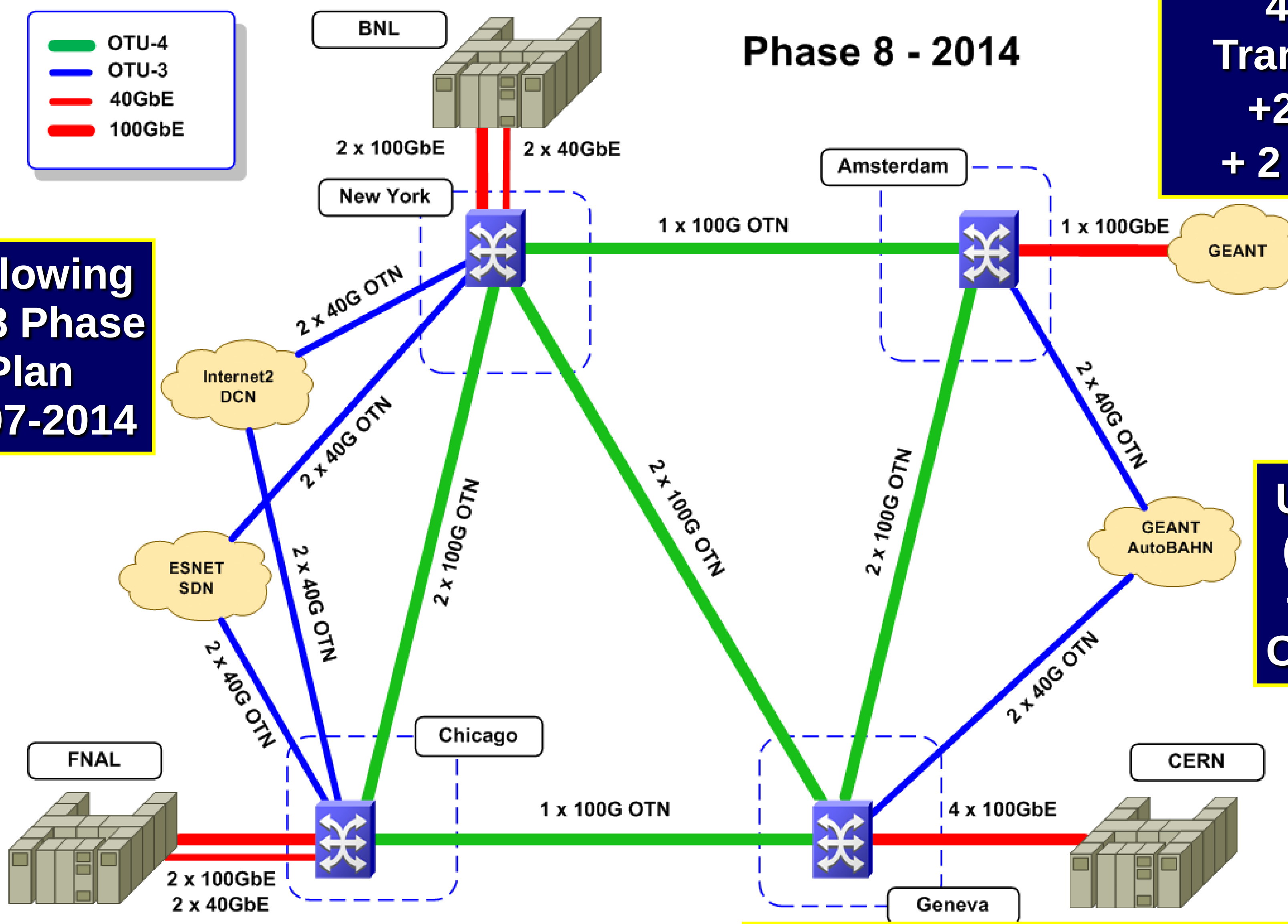
- Current US LHCNet bandwidth matches CMS and ATLAS requirements for 2010/2011 LHC run period
- US LHCNet is prepared for an upgrade in 2011
  - ➔ If LHC reaches the 2010-11 luminosity goals
  - ➔ Depending on the Experiments' Data Model evolution

	2009	2010	2011	2012	2013
No. of OC-192 links (or equivalent)	4 – 6	6	8	8	16
Line rate [Gbps]	37.6 – 56.4	56.4	75.2	75.2	150.4
Expected Application Payload Bandwidth [Gbps]	32-48	48	64	64	128

Implementation Scenario: USLHCNet Phase 8 (2015 or 2014 ?): Transition to Full Use of 100G

4 X 100G Trans-Atlantic +2 NY-CHI + 2 AMS-GVA

Following an 8 Phase Plan 2007-2014



Using OTU-4 (100G) Links + Next-Gen. Optical Muxes

Total 25 100G and 16 40G Mux. ports

US LHCNet: More than a Carrier

US LHCNet

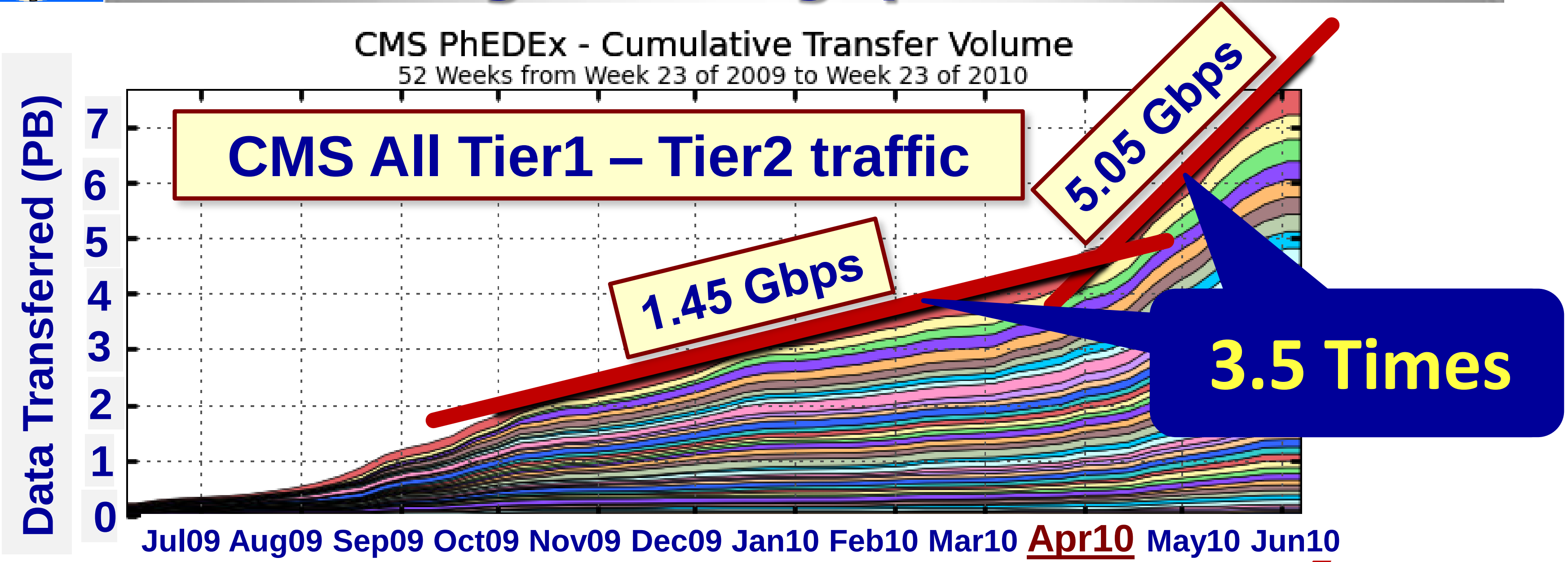
- US LHCNet is not "just a carrier", added value:
  - Proximity to the HEP community
    - Understand requirements from direct involvement
    - Expertise in high throughput data transport, global scale real-time monitoring infrastructures, LHC Computing Model issues, etc.
  - Fast response to (changing) requirements
  - Active R&D for HEP networking
    - Ultralight and PLaNetS (past NSF-funded projects)
    - End-to-end resource monitoring (MonALISA)
    - Dynamic resource allocation (e.g. DYNES)
- Mission orientation results in cost optimized high-performance, high-availability services to HEP, and in particular to the LHC Community

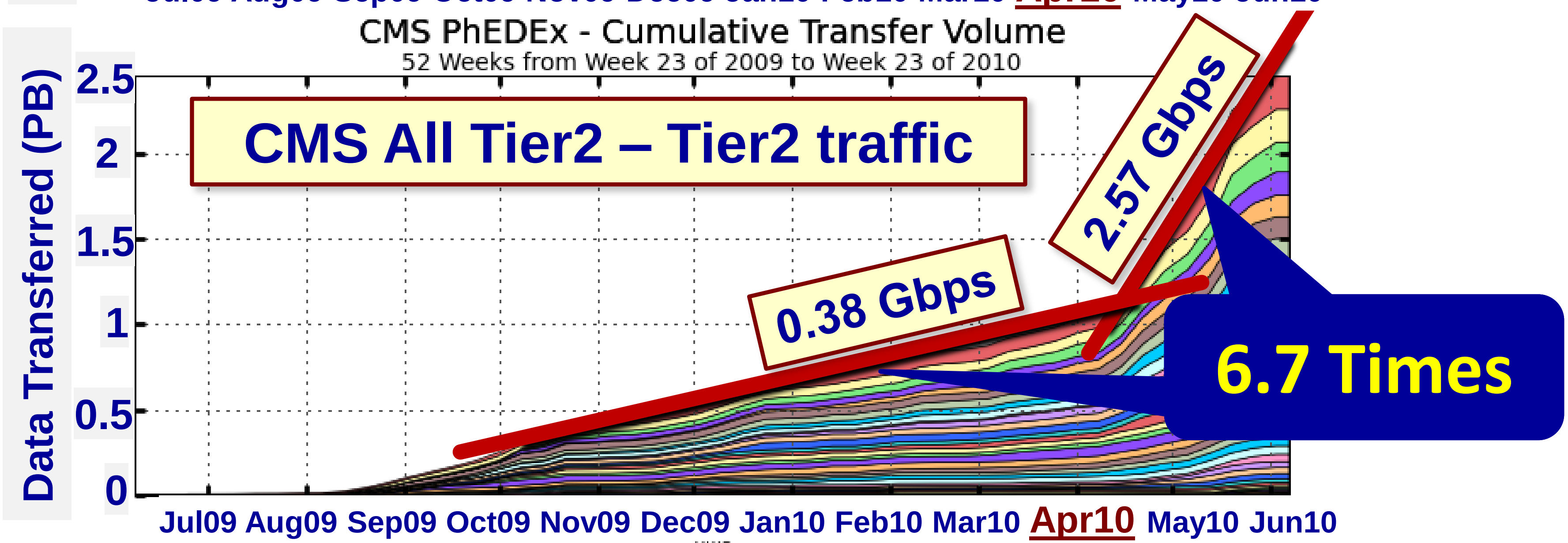
Conclusions

- US LHCNet provides high-availability, high-performance transatlantic networking to the LHC program
  - ➔ Mission Orientation Coupled to a Multi-layer View
    - ➔ Best value for investment
- LHC has started 7 TeV operation, expected to last to late 2011, with data rates expected to increase significantly still in 2010
- US LHCNet's Technology roadmap is tailored to provide the services & availability required by the LHC experiments
- US LHCNet's Bandwidth roadmap is designed to match the LHC experiments' requirements [just adequate for this run]
  - ➔ Need to stay agile; respond to the evolving needs
- Trends indicate: Tier2 networking is growing in importance
  - ➔ We need to consider a coordinated effort, led by the HEP community, working with the network community

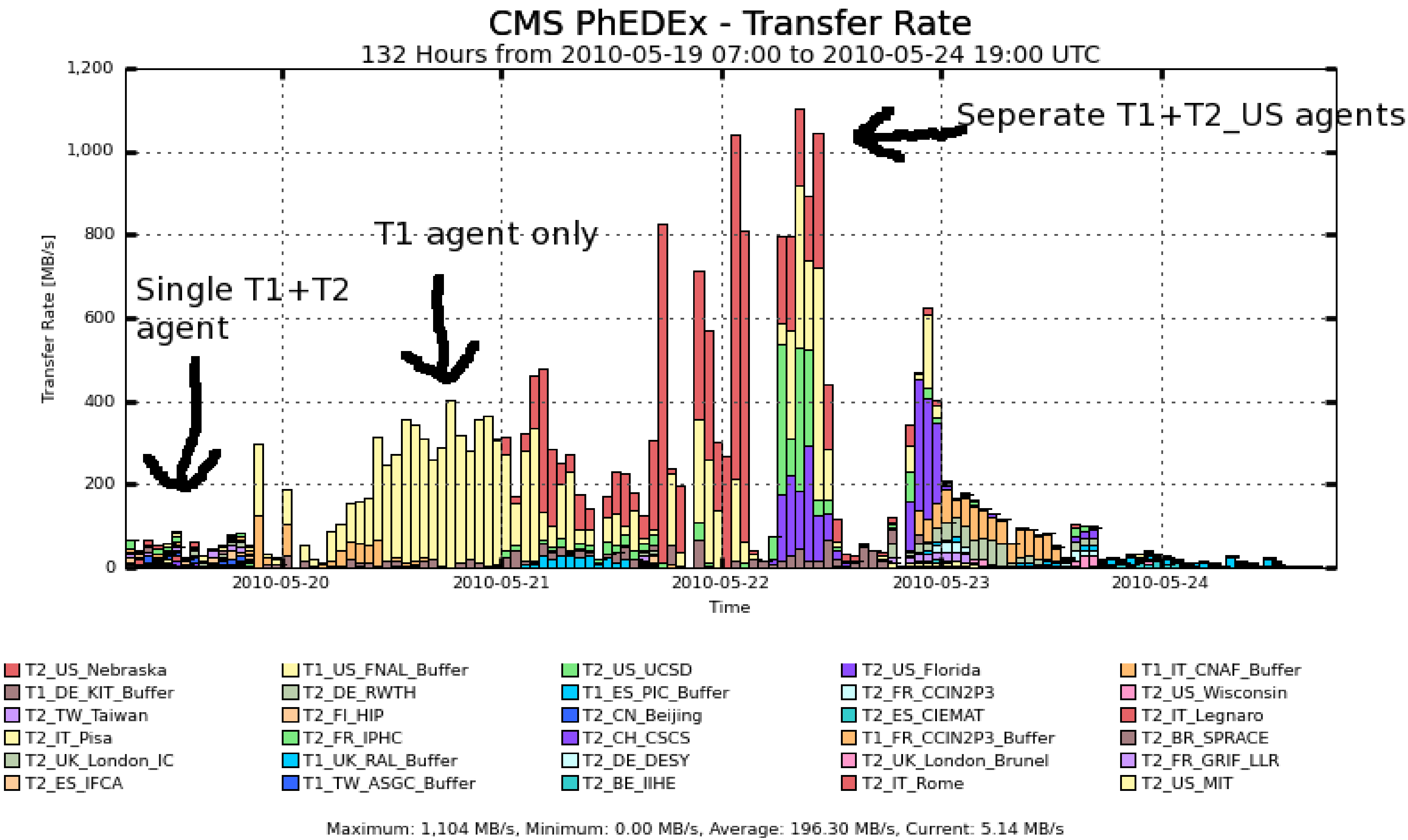
Backup Slides

Tier2 – Recent Trends: Average Throughput Numbers



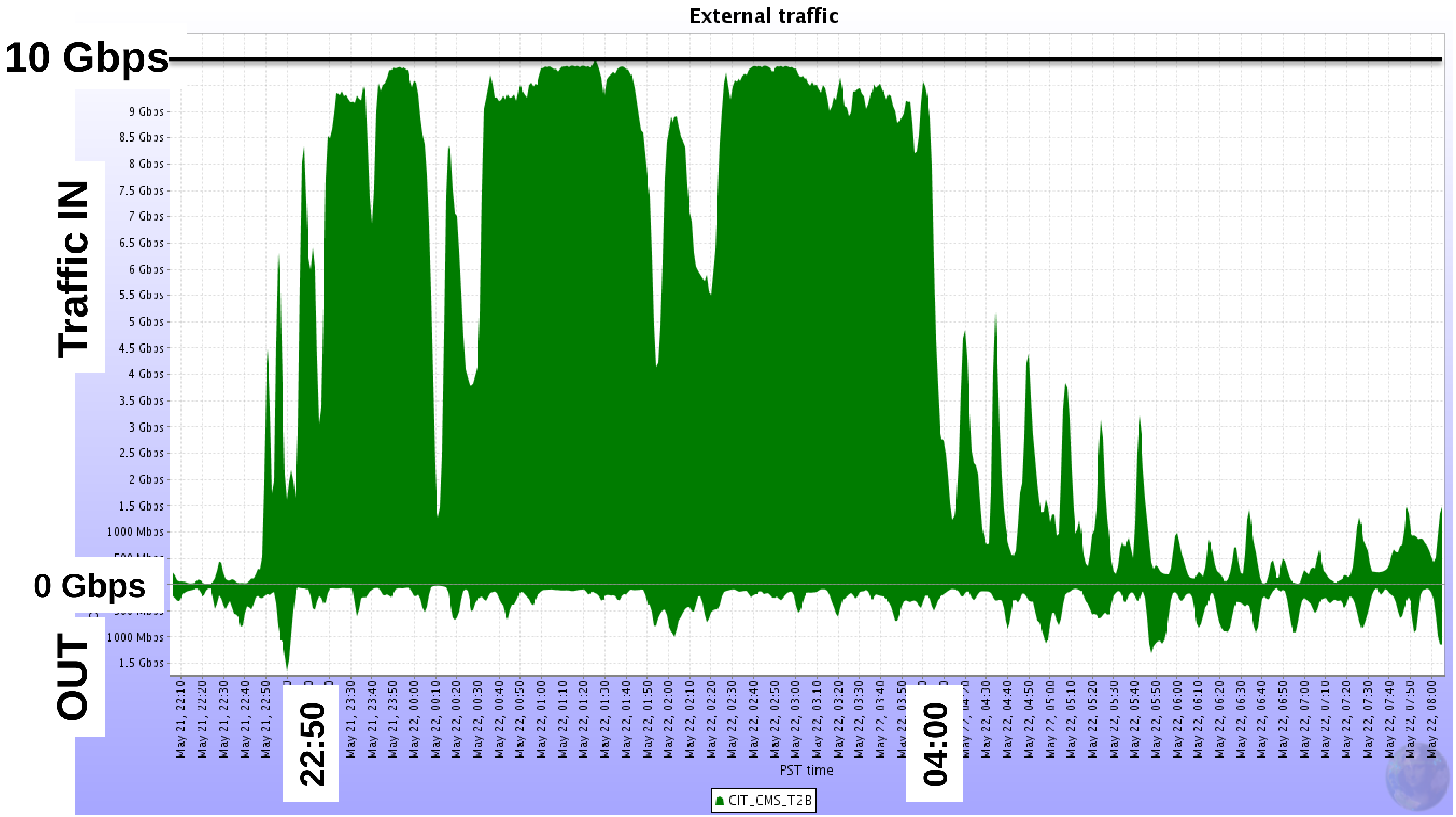


Application Level (PhEDEx) Tuning Example: Caltech Tier2



Tuned Tier2 Data Rates

- Well-tuned Tier 2 cluster (after Phedex tuning) can saturate a 10 Gbps link over an extended period of time:

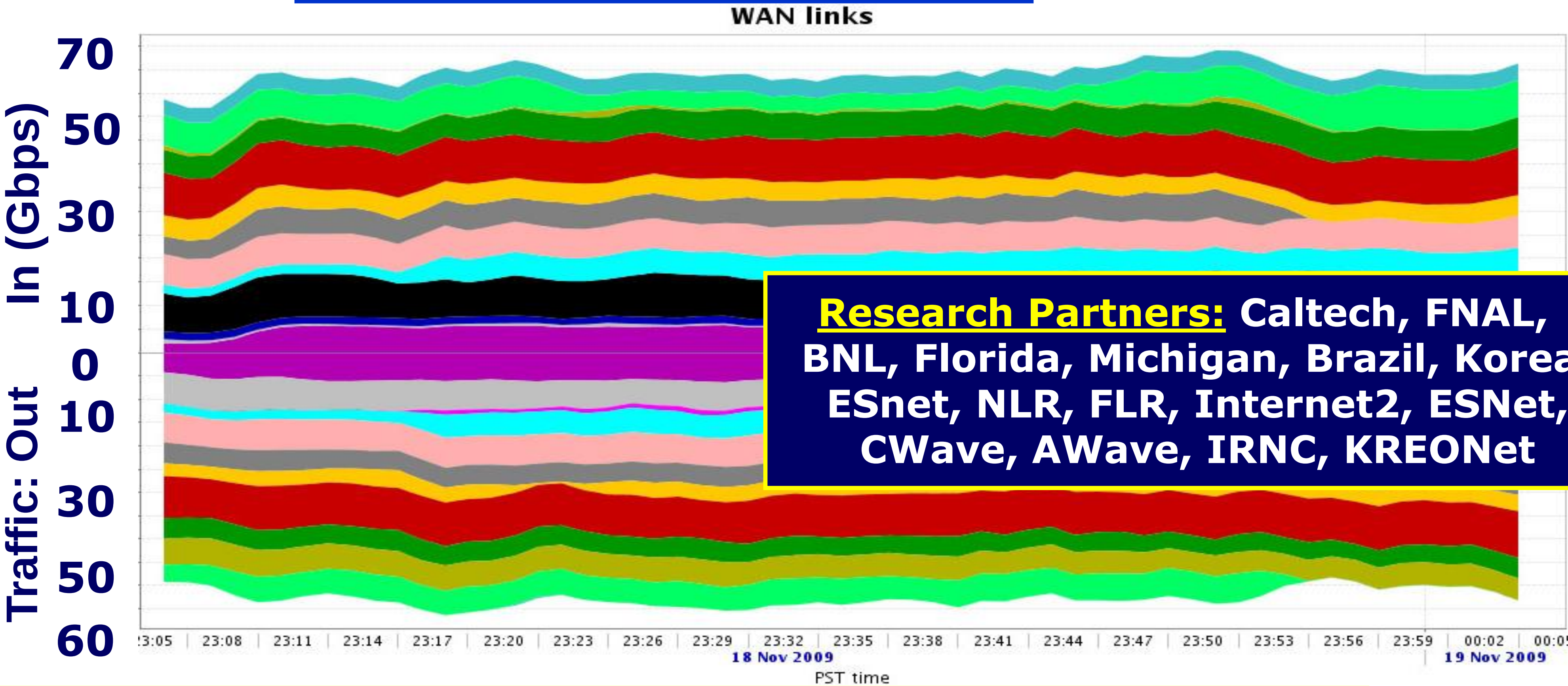


SC09

Tier2 Capabilities

Example: Tier 2 Sized Setup at SuperComputing 2009 Bandwidth Challenge:

~616 CPU Cores and 38 10GE NICs in 1 Rack of Servers
53 10GE Switch Ports
~100 TB Disk



Research Partners: Caltech, FNAL, BNL, Florida, Michigan, Brazil, Korea; ESnet, NLR, FLR, Internet2, ESNet, CWave, AWave, IRNC, KREONet

Max. 119 Gbps; 110 Gbps Sustained; 65 Gbps Outbound Using FDT and FDT/Hadoop Storage to Storage

NB: Sun Limitations