Networking for the Future of DOE Science: High Energy Physics / LHC Networking

February 26, 2007
(revised 3/20/2007)

William E. Johnston

ESnet Department Head and Senior Scientist
Lawrence Berkeley National Laboratory

wej@es.net, www.es.net

➢ DOE Office of Science and ESnet – the ESnet Mission

- "The Office of Science (SC) is the single largest supporter of basic research in the physical sciences in the United States, … providing more than 40 percent of total funding … for the Nation's research programs in high-energy physics, nuclear physics, and fusion energy sciences." (http://www.science.doe.gov)
- In FY2008 SC will support
  - 25,500 PhDs, PostDocs, and Graduate students
  - 21,500 users of SC facilities, half of which come from universities

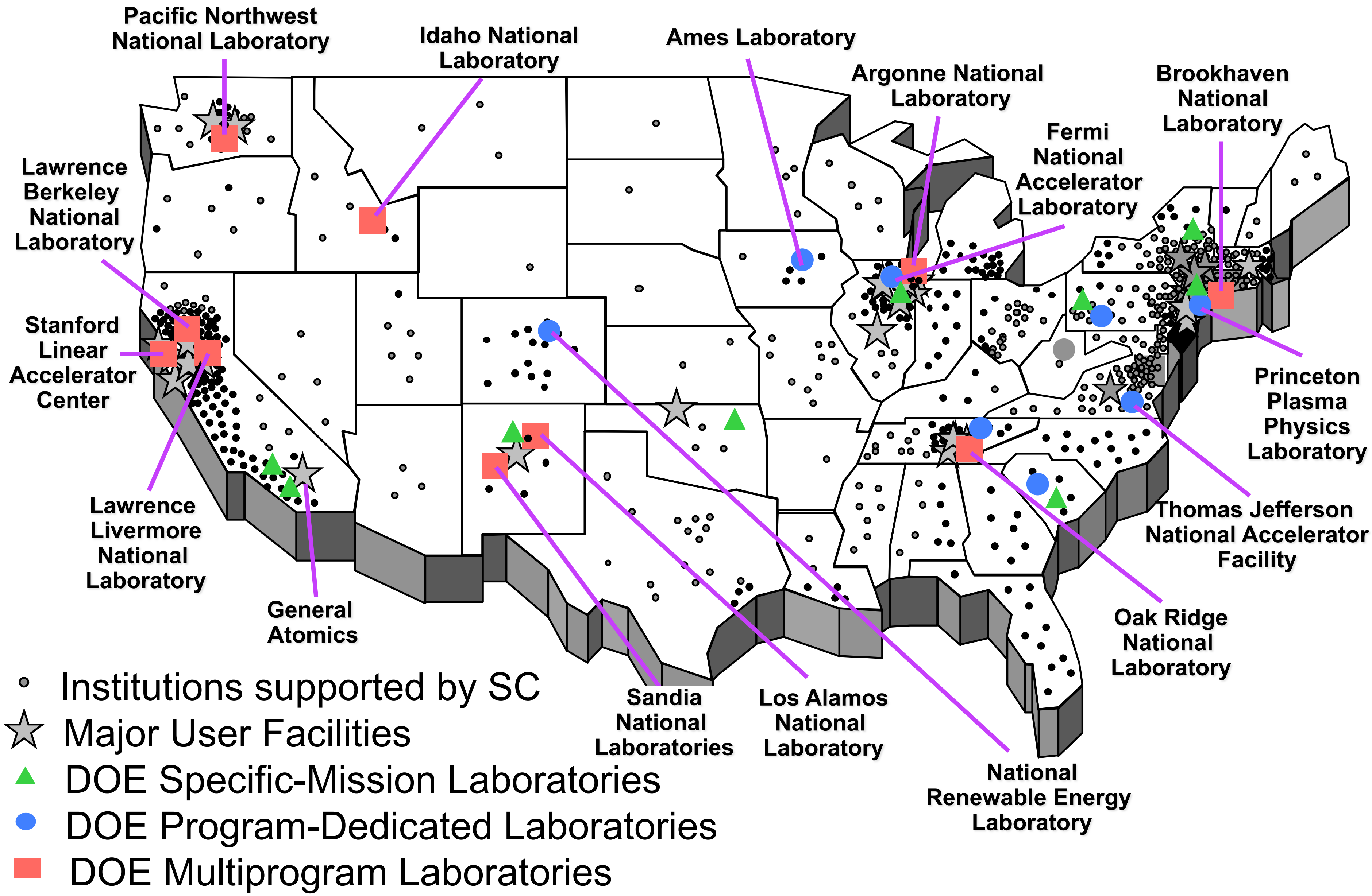
(From the FY2008 Budget Presentation of
Dr. Ray Orbach, Under Secretary for Science,
US Dept. of Energy)

DOE Office of Science and ESnet – the ESnet Mission

- **ESnet's primary mission is to enable the large-scale science that is the mission of the Office of Science (SC) and that depends on:**
  - Sharing of massive amounts of data
  - Supporting thousands of collaborators world-wide
  - Distributed data processing
  - Distributed data management
  - Distributed simulation, visualization, and computational steering
  - Collaboration with the US and International Research and Education community
- ESnet provides network and collaboration services to Office of Science laboratories and many other DOE programs in order to accomplish its mission

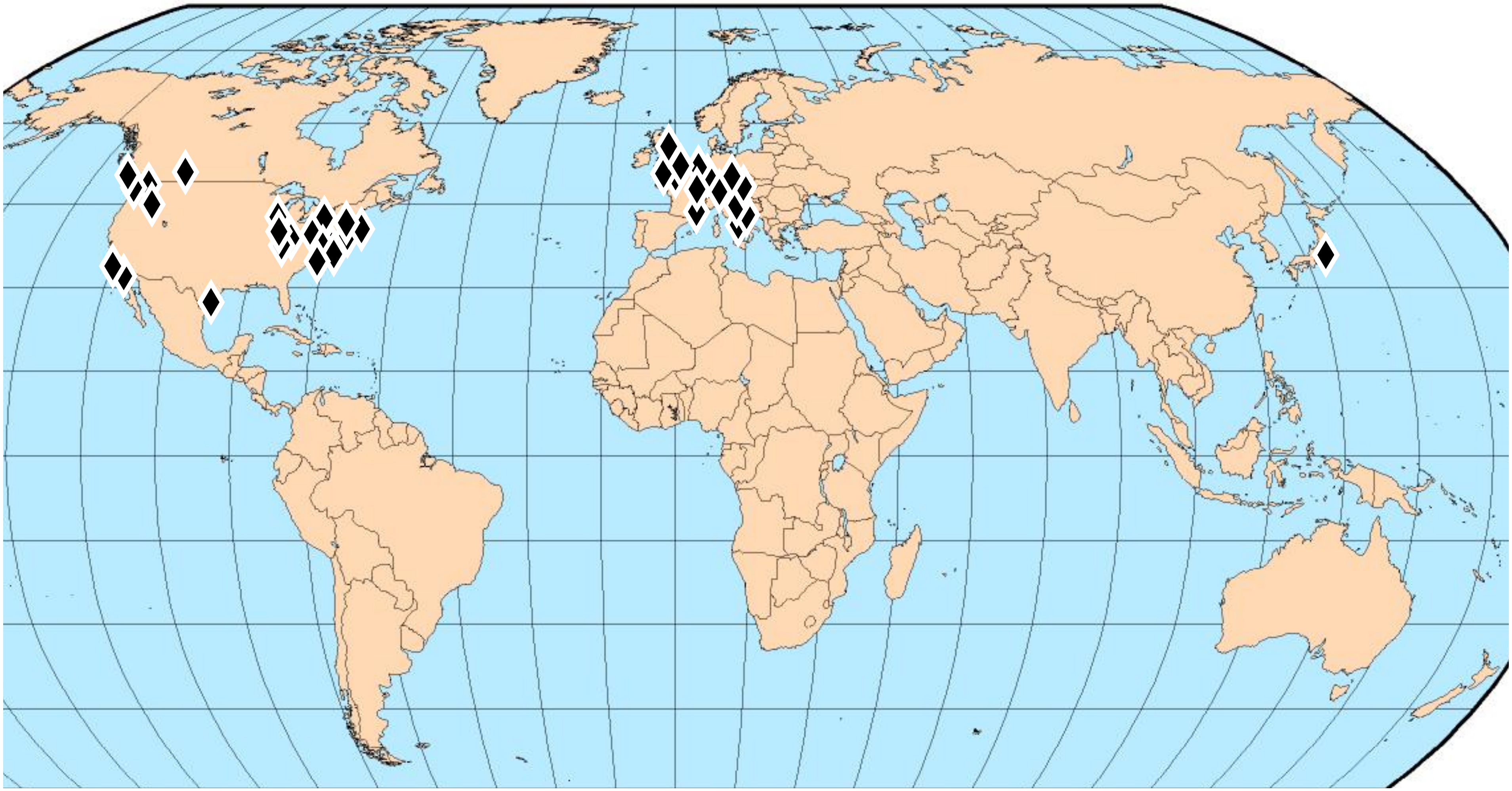
Office of Science US Community

Supporting Physical Sciences Research in the Universities



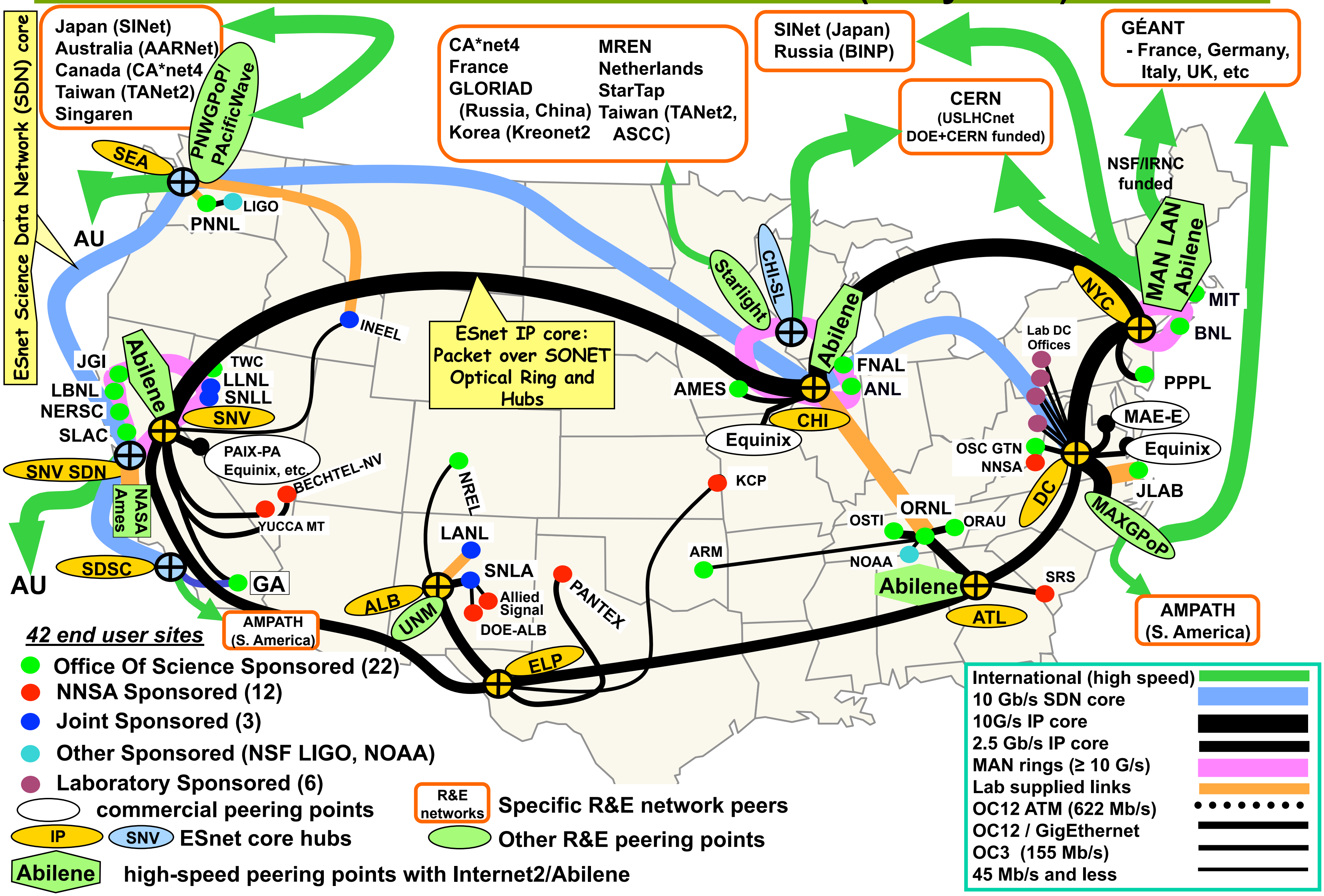
Footprint of Largest SC Data Sharing Collaborators

The Large-Scale Science Instruments of DOE's Office of Science Labs Send Much of their Data to the Research and Education Communities of the US and Europe



- Top 100 data flows generate 50% of all ESnet traffic (ESnet handles about 3×10^9 flows/mo.)
- 91 of the top 100 flows are from the Labs to other institutions (shown) (CY2005 data)

ESnet3 Today Provides Global High-Speed Internet Connectivity for DOE Facilities and Collaborators (Early 2007)



➢A Changing Science Environment is the Key Driver of the Next Generation ESnet

- Large-scale collaborative science – big facilities, massive data, thousands of collaborators – is now a significant aspect of the Office of Science (“SC”) program

- SC science community is almost equally split between Labs and universities
  - SC facilities have users worldwide

- Very large international (non-US) facilities (e.g. LHC and ITER) and international collaborators are now a key element of SC science

- Distributed systems for data analysis, simulations, instrument operation, etc., are essential and are now common (in fact dominate data analysis that now generates 50% of all ESnet traffic)

Planning for Future of Science: The Office of Science's Long Term Networking Requirements

- **Requirements of the Office of Science and their collaborators are primarily determined by**

  1) Data characteristics of instruments and facilities that will be connected to ESnet

  - What data will be generated by instruments coming on-line over the next 5-10 years?
  - How and where will it be analyzed and used?

  2) Examining the future process of science

  - How will the processing of doing science change over 5-10 years?
  - How do these changes drive demand for new network services?

  3) Studying the evolution of ESnet traffic patterns

  - What are the trends based on the use of the network in the past 2-5 years?
  - How must the network change to accommodate the future traffic patterns implied by the trends?

(1) Requirements from Instruments and Facilities

DOE SC Facilities that are, or will be, the top network users

- Advanced Scientific Computing Research
  - National Energy Research Scientific Computing Center (NERSC) (LBNL)*
  - National Leadership Computing Facility (NLCF) (ORNL)*
  - Argonne Leadership Class Facility (ALCF) (ANL)*
- Basic Energy Sciences
  - National Synchrotron Light Source (NSLS) (BNL)
  - Stanford Synchrotron Radiation Laboratory (SSRL) (SLAC)
  - Advanced Light Source (ALS) (LBNL)*
  - Advanced Photon Source (APS) (ANL)
  - Spallation Neutron Source (ORNL)*
  - National Center for Electron Microscopy (NCEM) (LBNL)*
  - Combustion Research Facility (CRF) (SNLL) *
- Biological and Environmental Research
  - William R. Wiley Environmental Molecular Sciences Laboratory (EMSL) (PNNL)*
  - Joint Genome Institute (JGI)
  - Structural Biology Center (SBC) (ANL)
- Fusion Energy Sciences
  - DIII-D Tokamak Facility (GA)*
  - Alcator C-Mod (MIT)*
  - National Spherical Torus Experiment (NSTX) (PPPL)*
  - ITER
- High Energy Physics
  - Tevatron Collider (FNAL)
  - B-Factory (SLAC)
  - Large Hadron Collider (LHC, ATLAS, CMS) (BNL, FNAL)*
- Nuclear Physics
  - Relativistic Heavy Ion Collider (RHIC) (BNL)*
  - Continuous Electron Beam Accelerator Facility (CEBAF) (JLab)*

*14 of 22 are characterized by current case studies

(2) Requirements from Examining the Future Process of Science

- In a major workshop [1], and in subsequent updates [2], requirements were generated by asking the science community how their process of doing science will / must change over the next 5 and next 10 years in order to accomplish their scientific goals
- Computer science and networking experts then assisted the science community in
  - analyzing the future environments
  - deriving middleware and networking requirements needed to enable these environments
- These were complied as case studies that provide specific 5 & 10 year network requirements for bandwidth, footprint, and new services

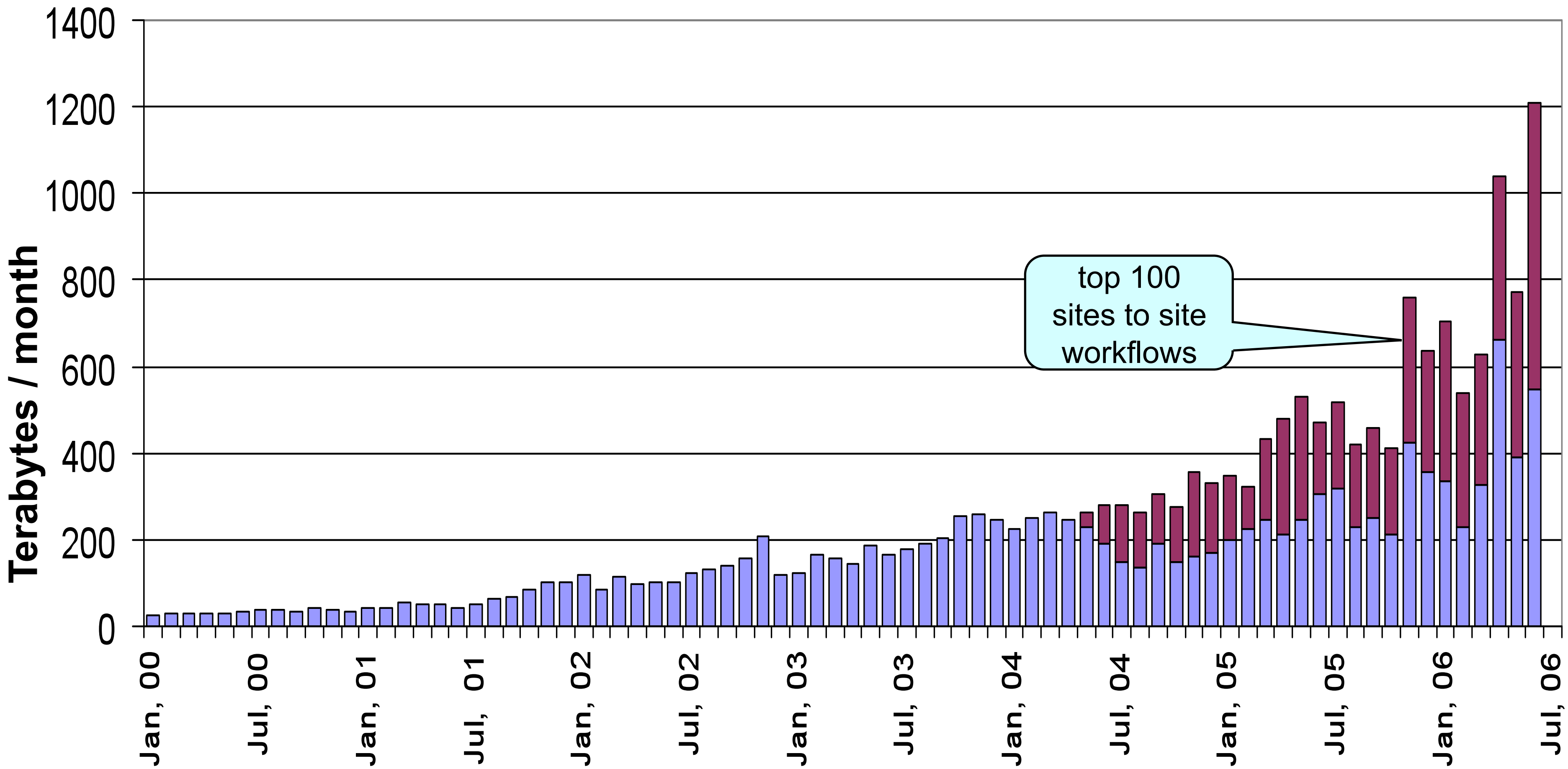
Science Networking Requirements Aggregation Summary

Science Drivers Science Areas / Facilities	End2End Reliability	Connectivity	Today End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
Magnetic Fusion Energy	99.999% (Impossible without full redundancy)	• DOE sites • US Universities • Industry	200+ Mbps	1 Gbps	• Bulk data • Remote control	• Guaranteed bandwidth • Guaranteed QoS • Deadline scheduling
NERSC and ACLF	-	• DOE sites • US Universities • International • Other ASCR supercomputers	10 Gbps	20 to 40 Gbps	• Bulk data • Remote control • Remote file system sharing	• Guaranteed bandwidth • Guaranteed QoS • Deadline Scheduling • PKI / Grid
NLCF	-	• DOE sites • US Universities • Industry • International	Backbone Band width parity	Backbone band width parity	• Bulk data • Remote file system sharing	
Nuclear Physics (RHIC)	-	• DOE sites • US Universities • International	12 Gbps	70 Gbps	• Bulk data	• Guaranteed bandwidth • PKI / Grid
Spallation Neutron Source	High (24x7 operation)	• DOE sites	640 Mbps	2 Gbps	• Bulk data	

Science Network Requirements Aggregation Summary

Science Drivers Science Areas / Facilities	End2End Reliability	Connectivity	Today End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
Advanced Light Source	-	• DOE sites • US Universities • Industry	1 TB/day 300 Mbps	5 TB/day 1.5 Gbps	• Bulk data • Remote control	• Guaranteed bandwidth • PKI / Grid
Bioinformatics	-	• DOE sites • US Universities	625 Mbps 12.5 Gbps in two years	250 Gbps	• Bulk data • Remote control • Point-to-multipoint	• Guaranteed bandwidth • High-speed multicast
Chemistry / Combustion	-	• DOE sites • US Universities • Industry	-	10s of Gigabits per second	• Bulk data	• Guaranteed bandwidth • PKI / Grid
Climate Science	-	• DOE sites • US Universities • International	-	5 PB per year 5 Gbps	• Bulk data • Remote control	• Guaranteed bandwidth • PKI / Grid
		Immediate Requirements and Drivers				
High Energy Physics (LHC)	99.95+% (Less than 4 hrs/year)	• US Tier1 (FNAL, BNL) • US Tier2 (Universities) • International (Europe, Canada)	10 Gbps	60 to 80 Gbps (30-40 Gbps per US Tier1)	• Bulk data • Coupled data analysis processes	• Guaranteed bandwidth • Traffic isolation • PKI / Grid

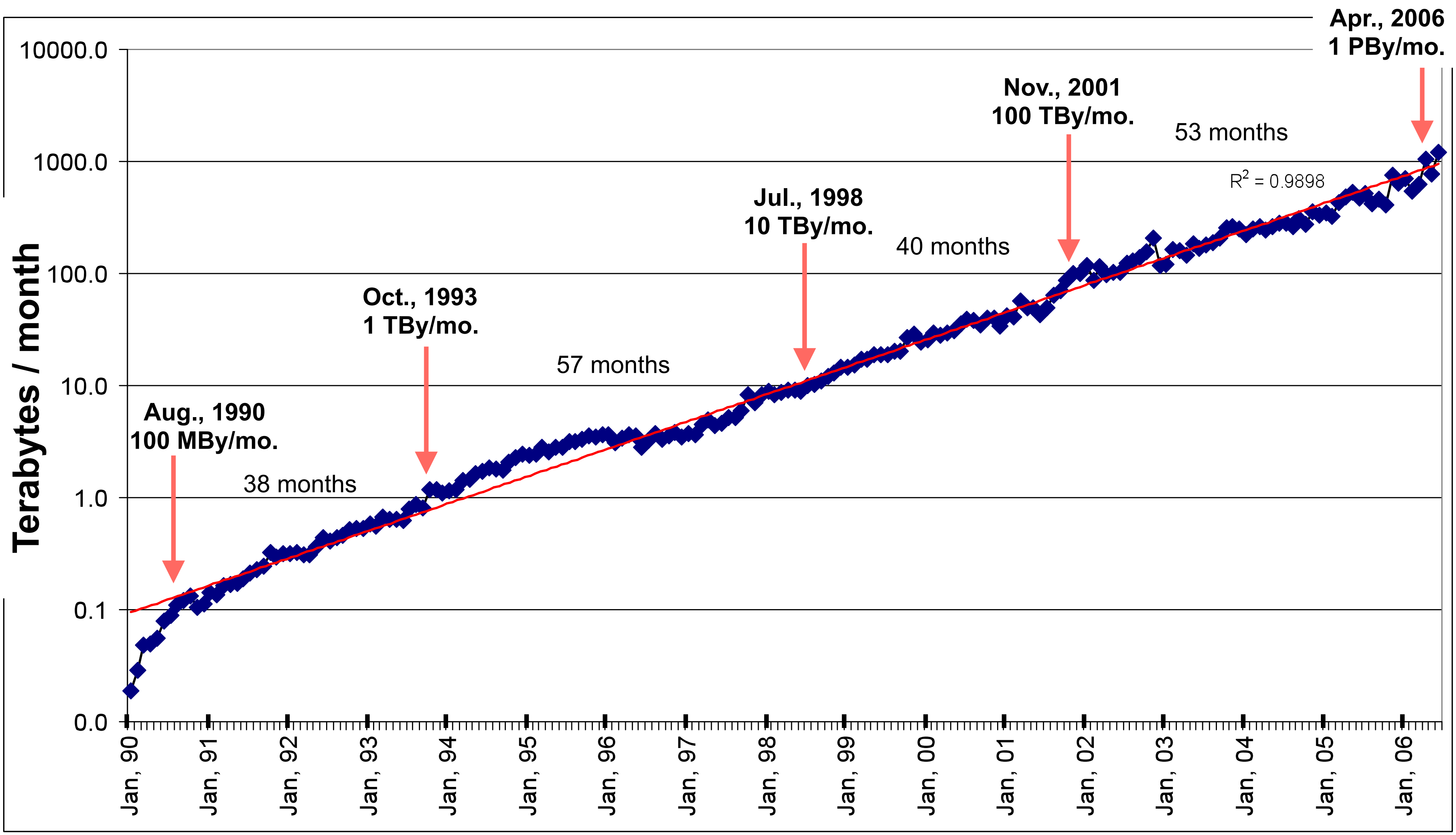
(3) The Science Trends are Seen in Observed Evolution of Historical ESnet Traffic Patterns



ESnet Monthly Accepted Traffic, January, 2000 – June, 2006

- ESnet is currently transporting more than1 petabyte (1000 terabytes) per month
- More than 50% of the traffic is now generated by the top 100 sites — large-scale science dominates all ESnet traffic

Reflecting the Growth of the Office of Science Large-Scale Science, ESnet Traffic has Increased by 10X Every 47 Months, on Average, Since 1990

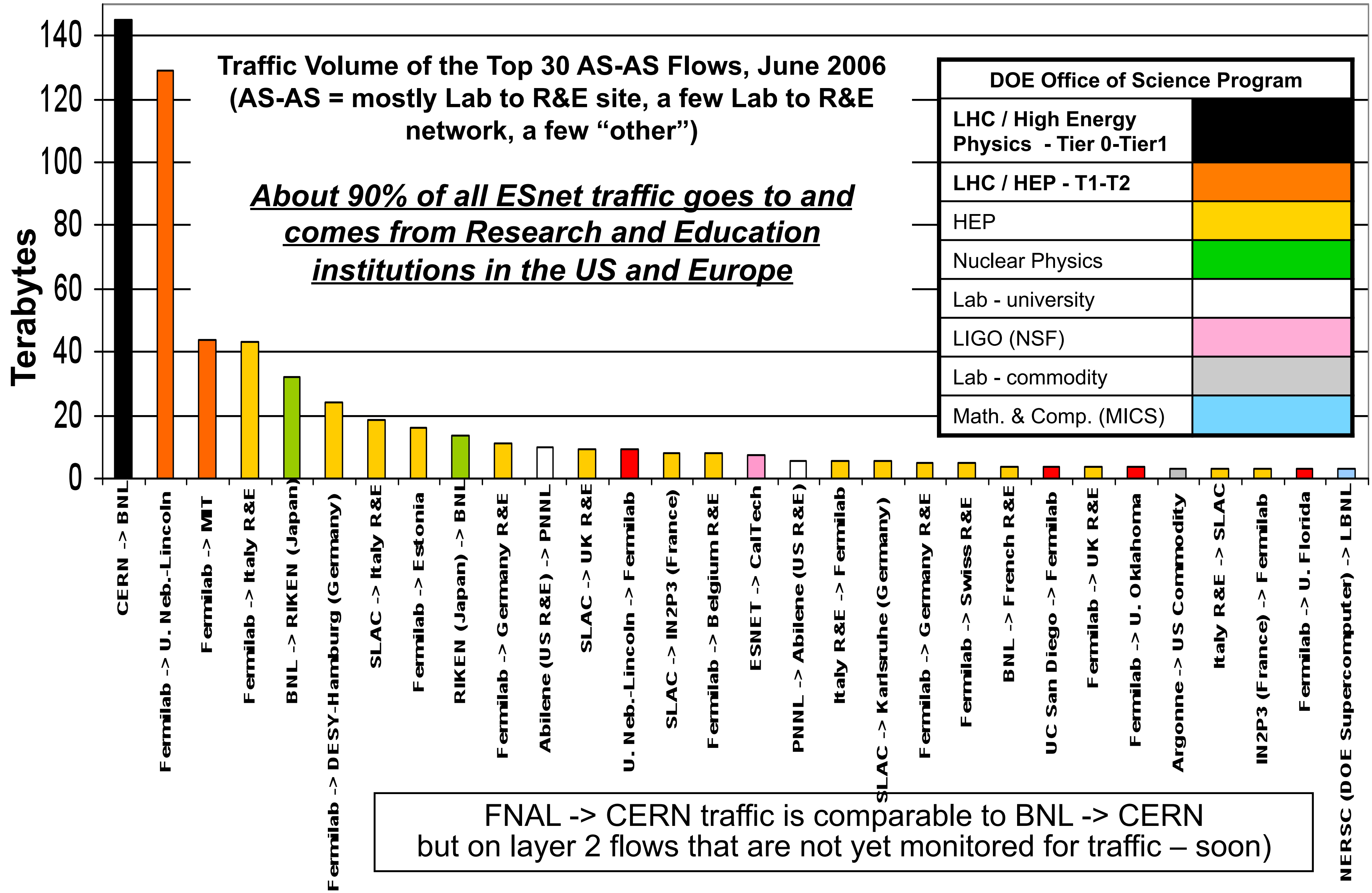


Log Plot of ESnet Monthly Accepted Traffic, January, 1990 – June, 2006

Requirements from Network Utilization Observation

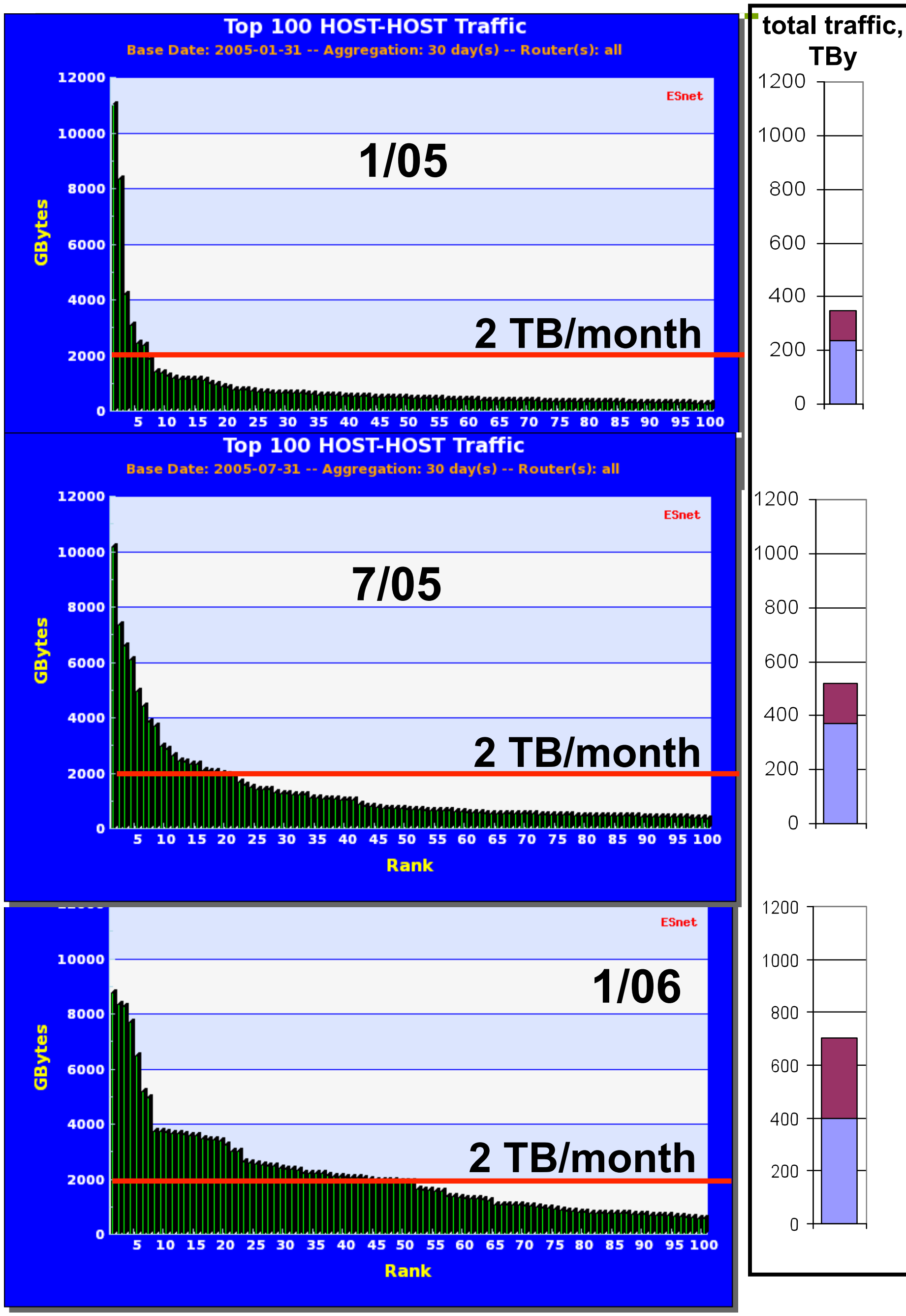
- In 4 years, we can expect a 10x increase in traffic over current levels *without the addition of production LHC traffic*
  - Nominal average load on busiest backbone links is ~1.5 Gbps today
  - In 4 years that figure will be ~15 Gbps based on current trends
- Measurements of this type are science-agnostic
  - It doesn't matter who the users are, the traffic load is increasing exponentially
  - Predictions based on this sort of forward projection tend to be conservative estimates of future requirements because they cannot predict new uses
- Bandwidth trends drive requirement for a new network architecture
  - New architecture/approach must be scalable in a cost-effective way

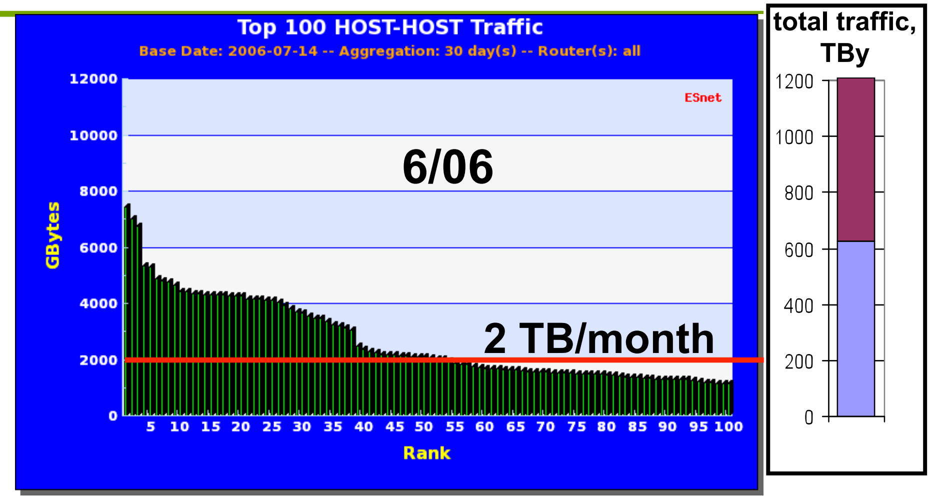
Large-Scale Flow Trends, June 2006
(Subtitle: “Onslaught of the LHC”)



FNAL -> CERN traffic is comparable to BNL -> CERN
but on layer 2 flows that are not yet monitored for traffic – soon)

➢Traffic Patterns are Changing Dramatically

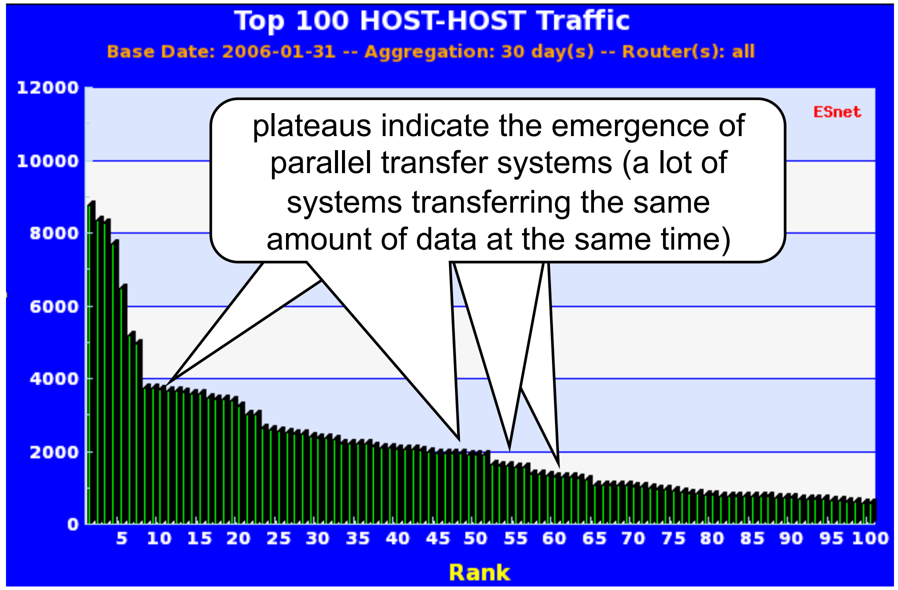




- While the total traffic is increasing exponentially
  - Peak flow – that is system-to-system – bandwidth is decreasing
  - The number of large flows is increasing

The Onslaught of Grids

Question: Why is peak flow bandwidth decreasing while total traffic is increasing?

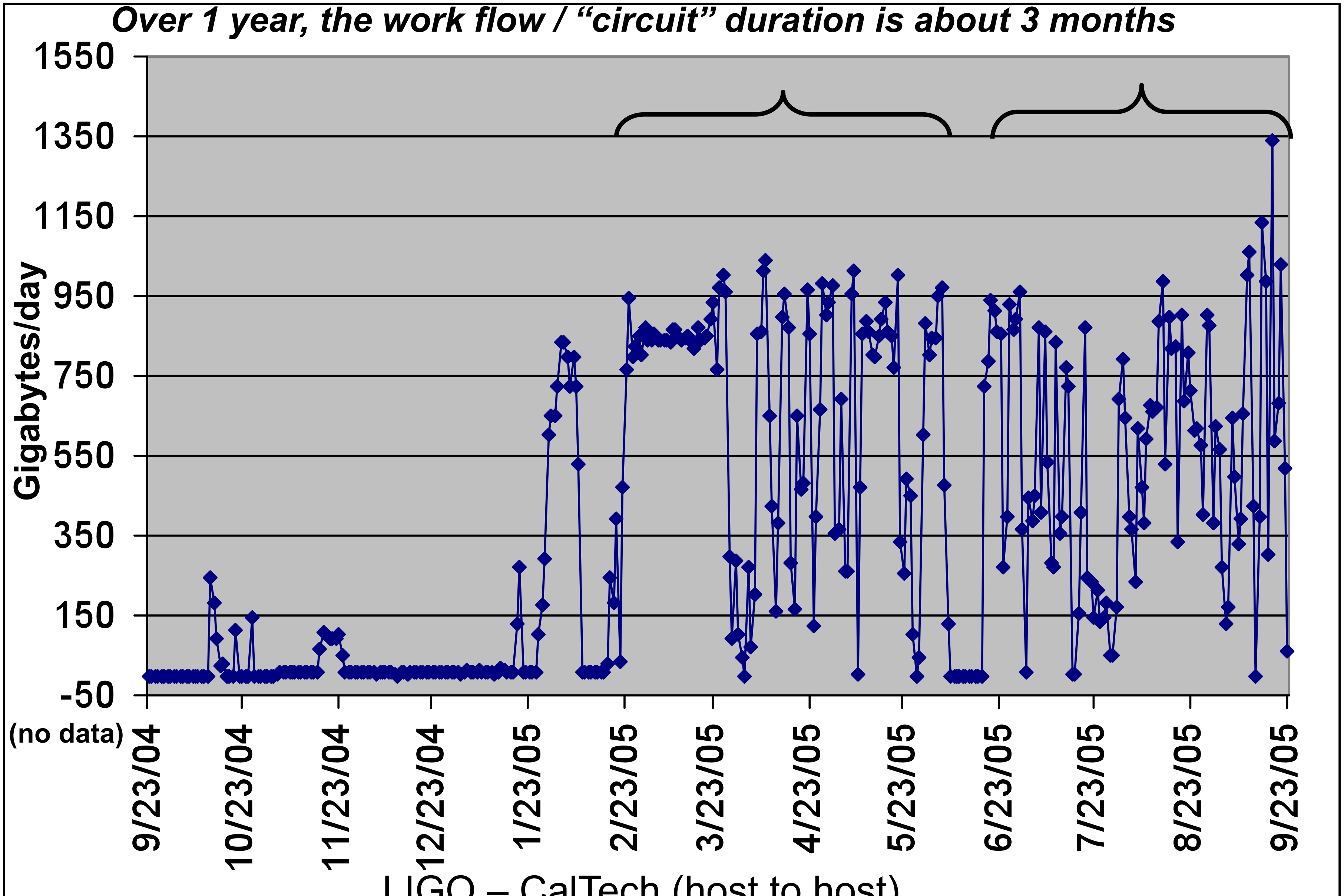


Answer: Most large data transfers are now done by parallel / Grid data movers

- In June, 2006 ***<u>72%</u>*** of the hosts generating the top 1000 flows were involved in parallel data movers (Grid applications)
- ***This is the most significant traffic pattern change in the history of ESnet***
- This has implications for the network architecture that favor path multiplicity and route diversity

➢Network Observation – Circuit-like Behavior

Look at Top 20 Traffic Generator's Historical Flow Patterns



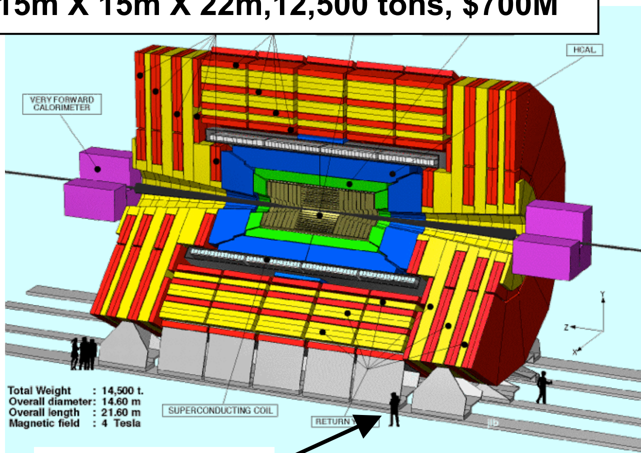
Prototype Large-Scale Science: High Energy Physics' Large Hardon Collider (Accelerator) at CERN

LHC Goal - Detect the Higgs Boson

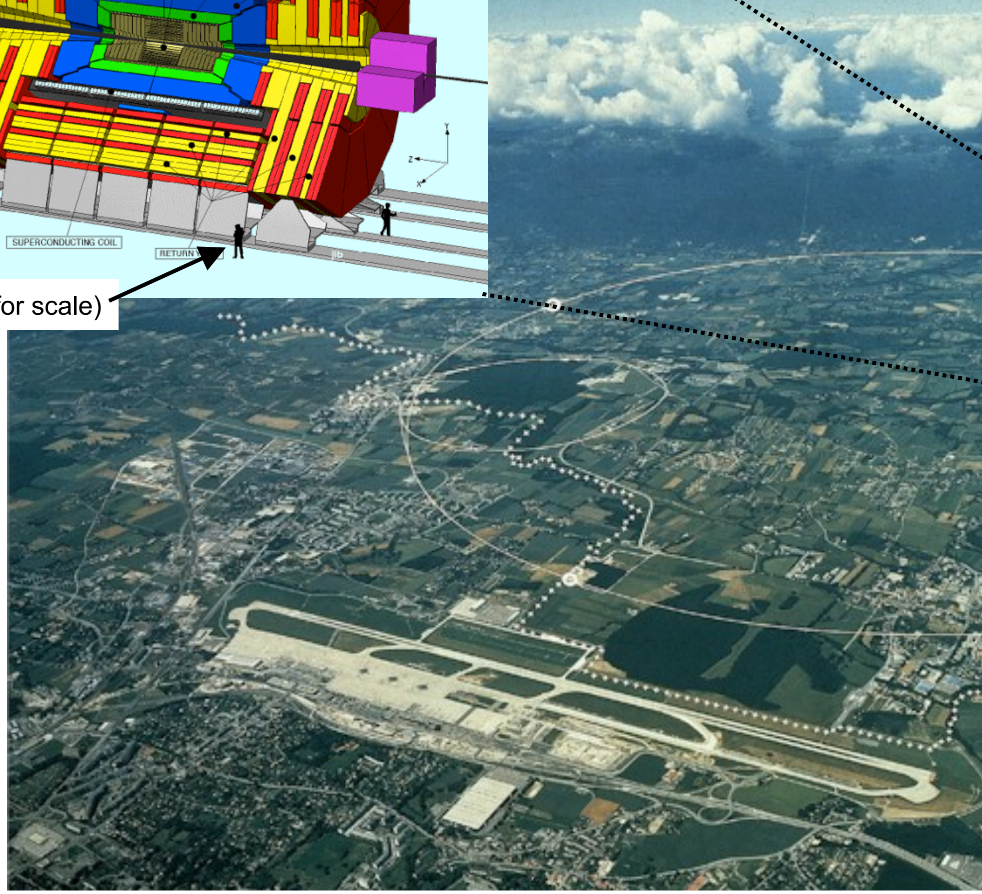
The Higgs boson is a hypothetical massive scalar elementary particle predicted to exist by the Standard Model of particle physics. It is the only Standard Model particle not yet observed, but *plays a key role in explaining the origins of the mass* of other elementary particles, in particular the difference between the massless photon and the very heavy W and Z bosons. Elementary particle masses, and the differences between electromagnetism (caused by the photon) and the weak force (caused by the W and Z bosons), are critical to many aspects of the structure of microscopic (and hence macroscopic) matter; thus, if it exists, the Higgs boson has an enormous effect on the world around us.

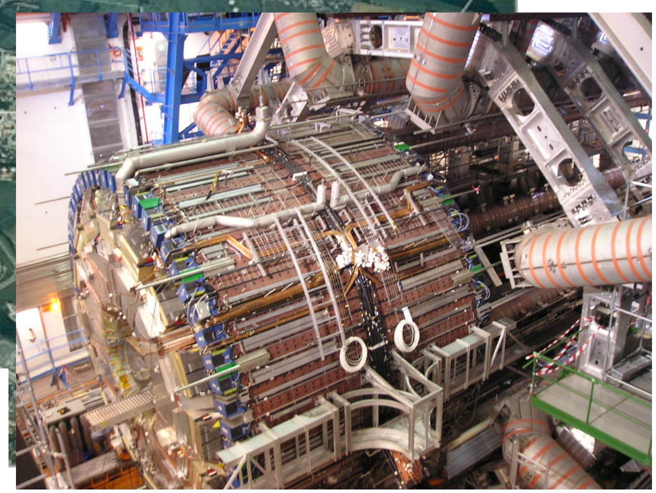
The Largest Facility: Large Hadron Collider at CERN

LHC CMS detector
15m X 15m X 22m,12,500 tons, $700M



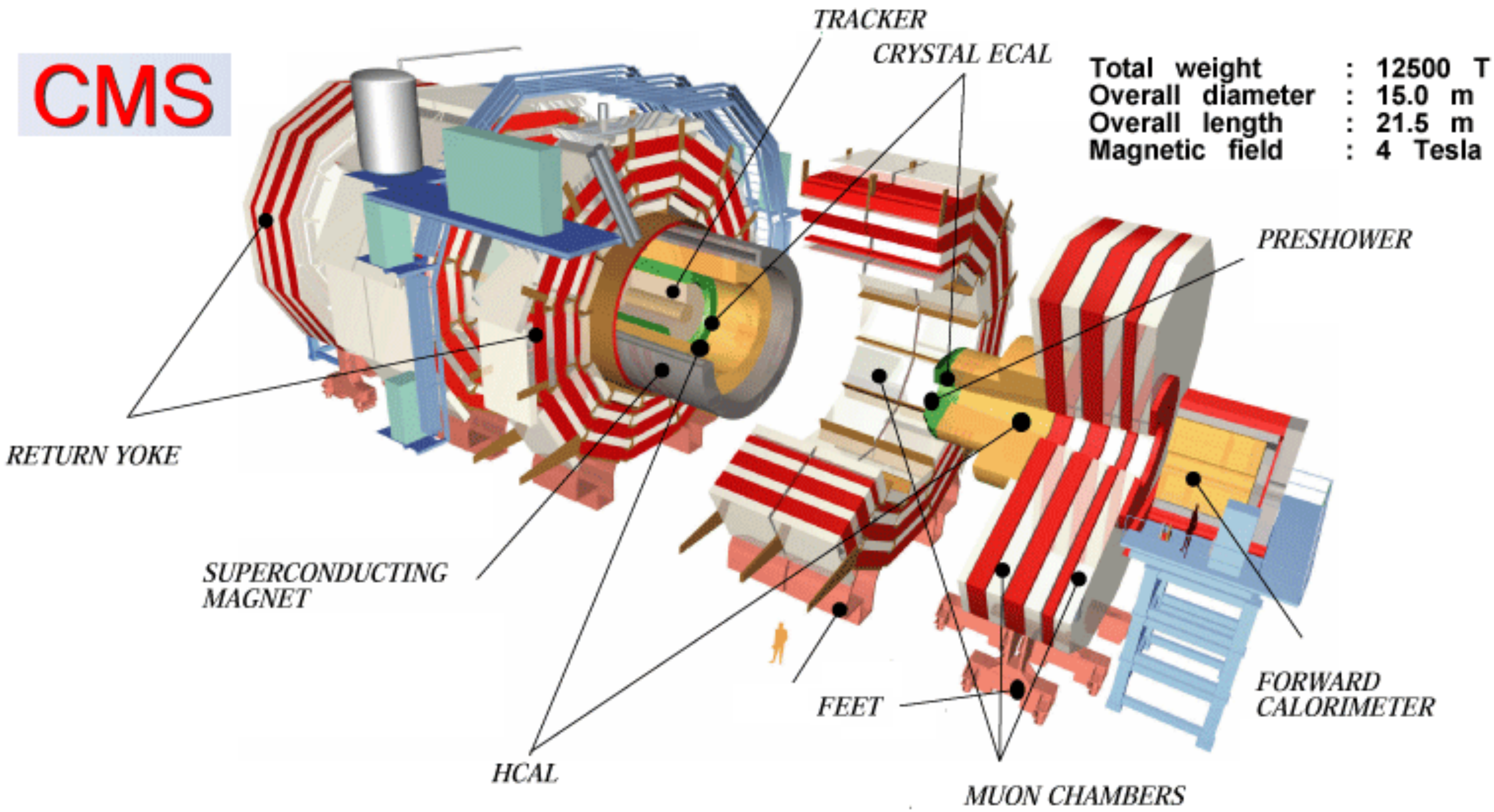
human (for scale)





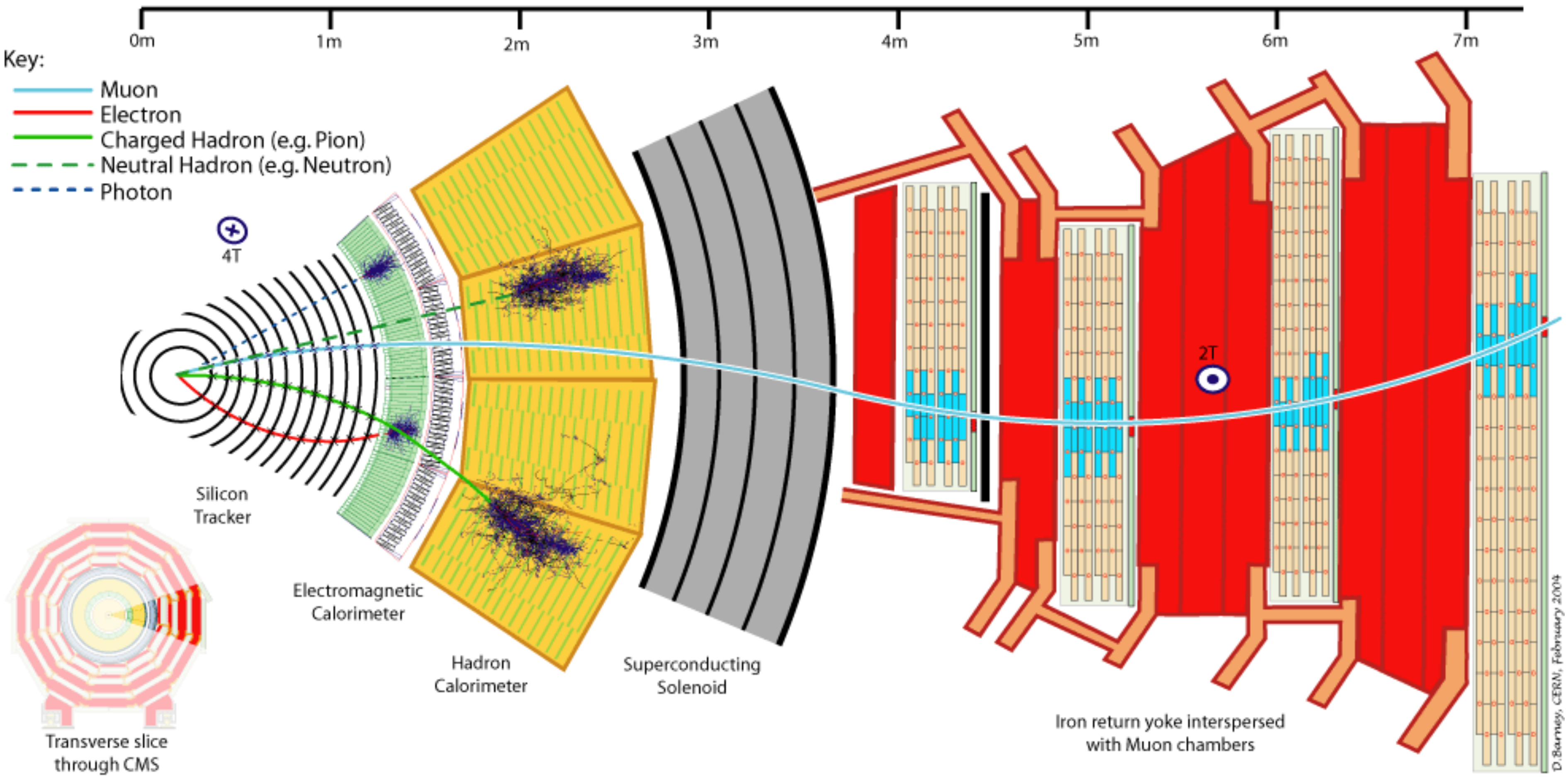
Two counter-rotating, 7 TeV proton beams, 27 km circumference (8.6 km diameter), collide in the middle of the detectors

One of the two Primary Experiments at the LHC



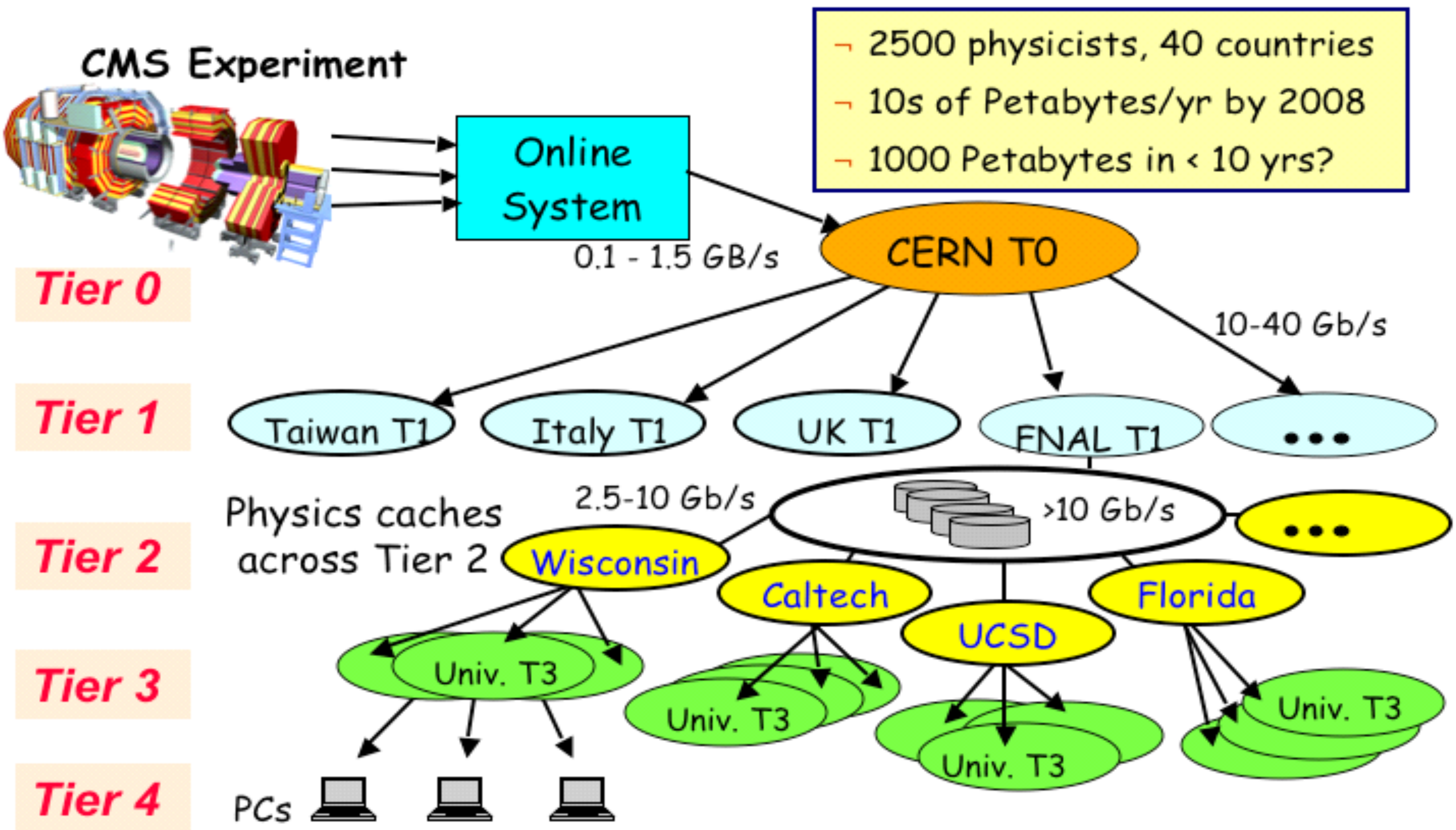
The set up of the Compact Muon Solenoid (CMS). In the middle, under the so called barrel there is a man for the scale. (HCAL=hadron calorimeter, ECAL=electromagnetic calorimeter)

One of the two Primary Experiments at the LHC



A slice of the CMS detector.

Data Management Model: A refined view of the LHC Data Grid Hierarchy where operations of the Tier2 centers and the U.S. Tier1 center are integrated through network connections with typical speeds in the 10 Gbps range. [ICFA SCIC]

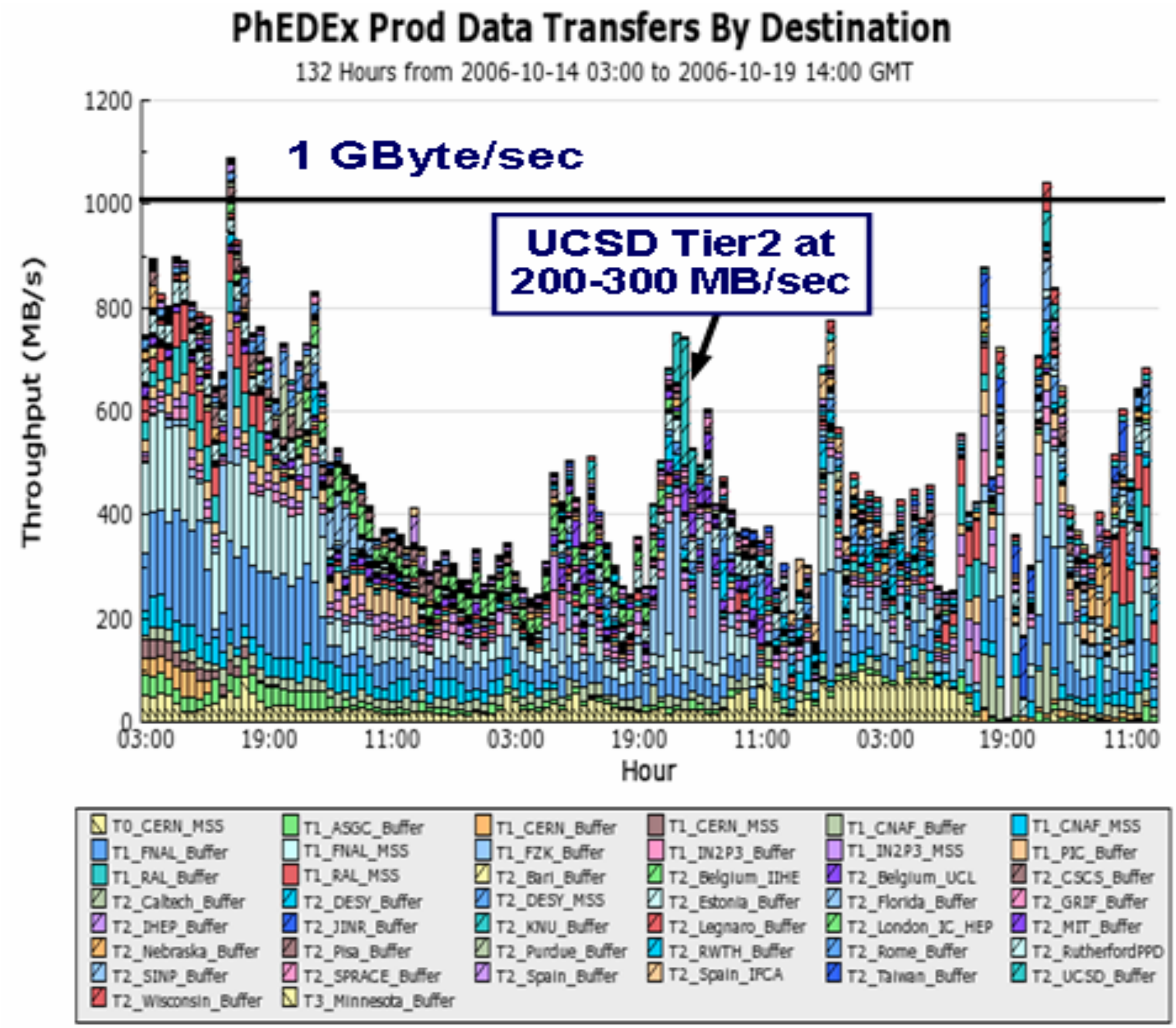


Roadmap for major links used by HEP. Projections follow the trend of affordable bandwidth increases over the last 20 years: by a factor of ~400 to 1000 times per decade. The entries marked in yellow reflect past or present implementations. [ICFA SCIC]

Year	Production	Experimental	Remarks
2001	0.155	0.622-2.5	SONET/SDH
2002	0.622	2.5	SONET/SDH DWDM; GigE Integ.
2003	2.5	10	DWDM; 1 + 10 GigE Integration
2005-6	10-20	2-4 X 10	λ Switch; λ Provisioning
➢ 2007-8	30-40	~100 or 2 X 40 Gbps	1st Gen. λ Grids
2009-10	60-80	~5 X 40 or ~2 X 100	40 or 100 Gbps λ Switching
2011-12	~5 X 40 or ~2 X 100	~25 X 40 or ~10 X 100	2nd Gen λ Grids Terabit Networks
2013-14	~Terabit	~MultiTbps	~Fill One Fiber

Readiness

132 Hours of CMS data transfers among sites in the US and Europe using PhEDEx, by destination, during October 2006 [ICFA SCIC]

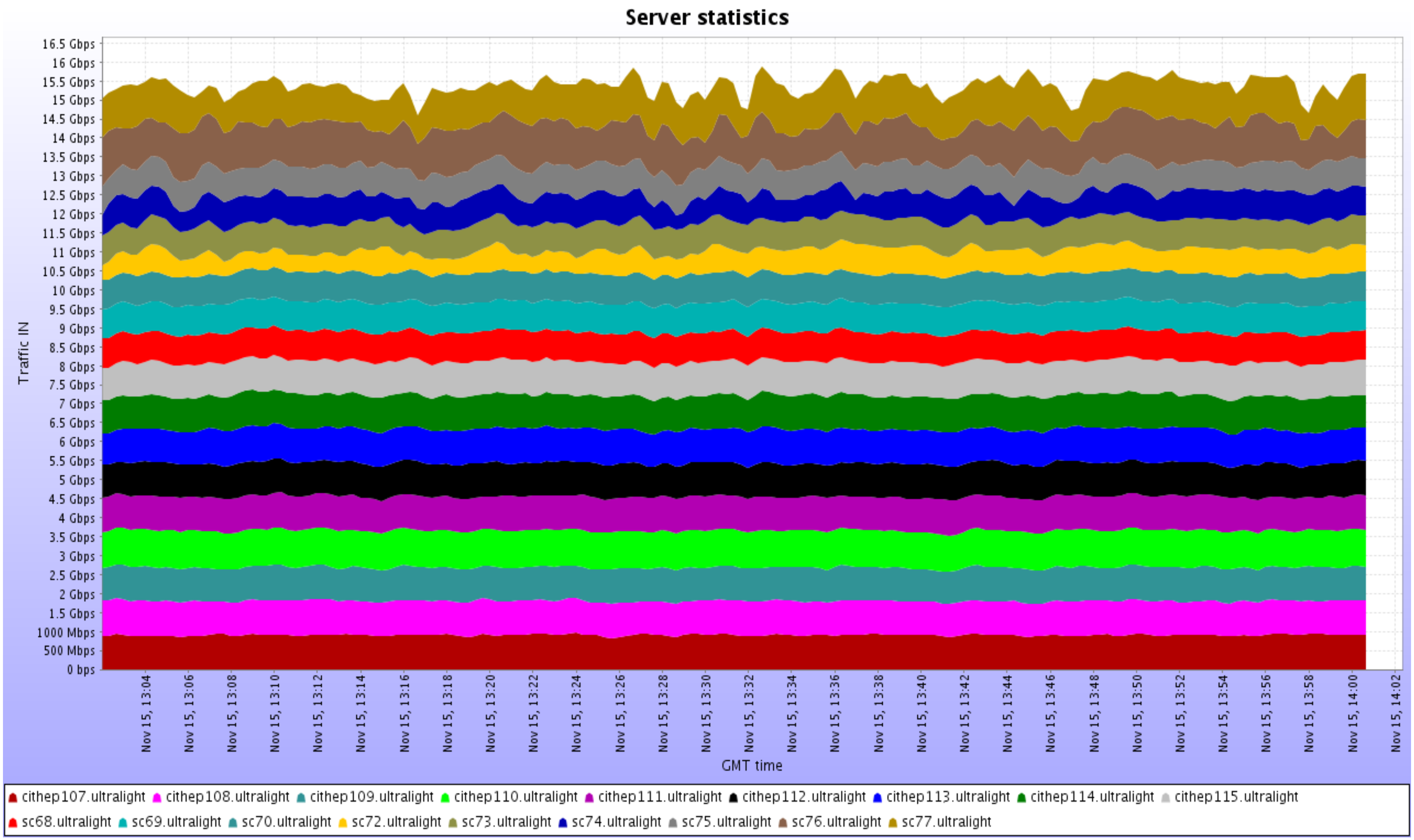


[The] production tools themselves have been shown to scale to [high] data rates over short distances. A PhEDEx performance validation test[1] in December 2006 showed scalability up to 1.1 Petabytes per hour. Fermilab's dCache also was shown to be able to transfer data at speeds approaching 40 Gbps (equivalent to more than 10 Petabytes/month) over the local area network at the lab[2].

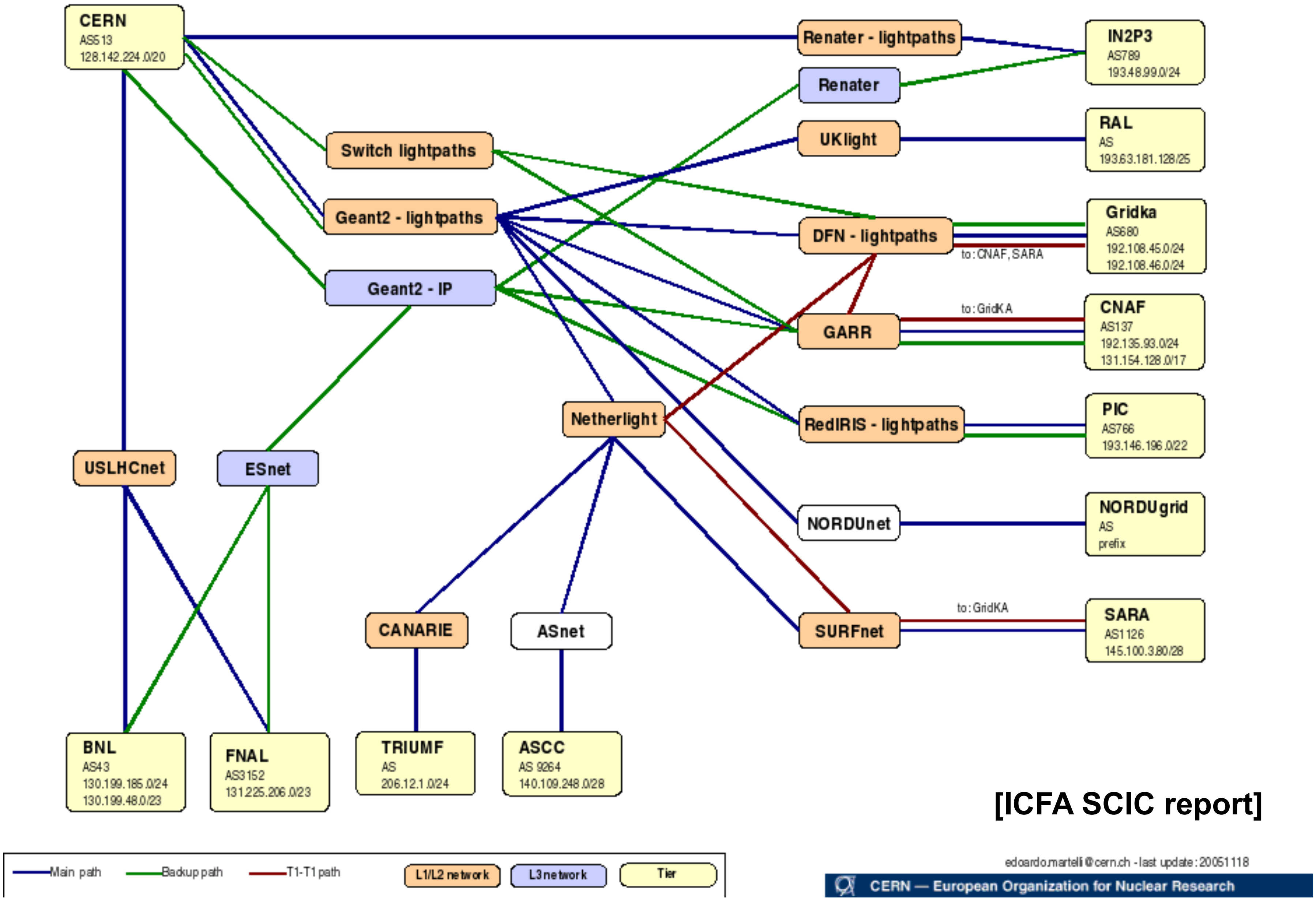
[1] See https://twiki.cern.ch/twiki/bin/view/CMS/PhedexValidation20061213
[2] Source: M. Crawford, FNAL.

FDT disk-to-disk data flows between SC06 and Caltech using 10 nodes sending and 8 nodes receiving data on a single 10 Gbps link. The stable flows in the figure continued overnight.

[ICFA SCIC]

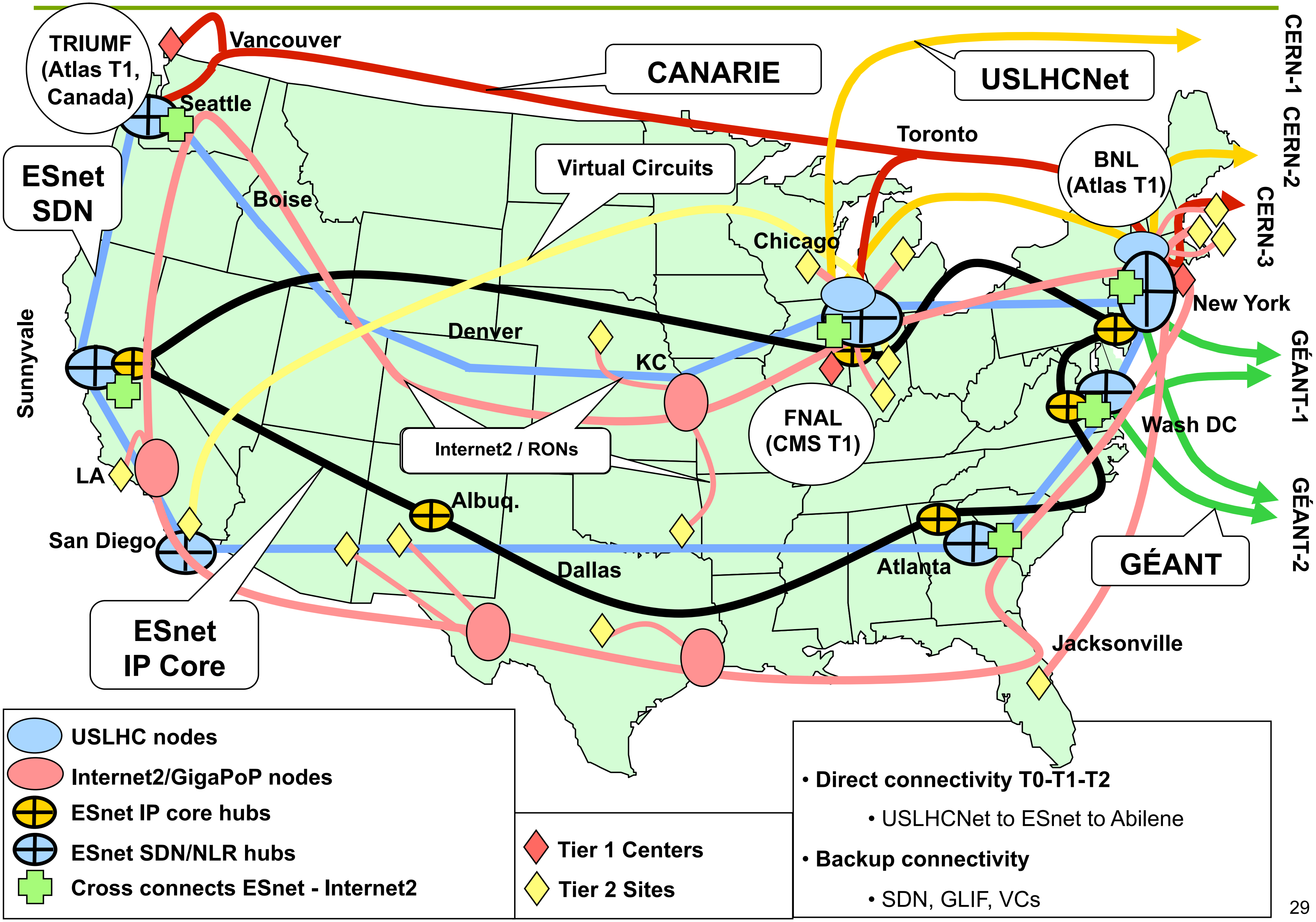


Production:
LHC Optical Private Network (OPN) connecting CERN to TIER-1 centres



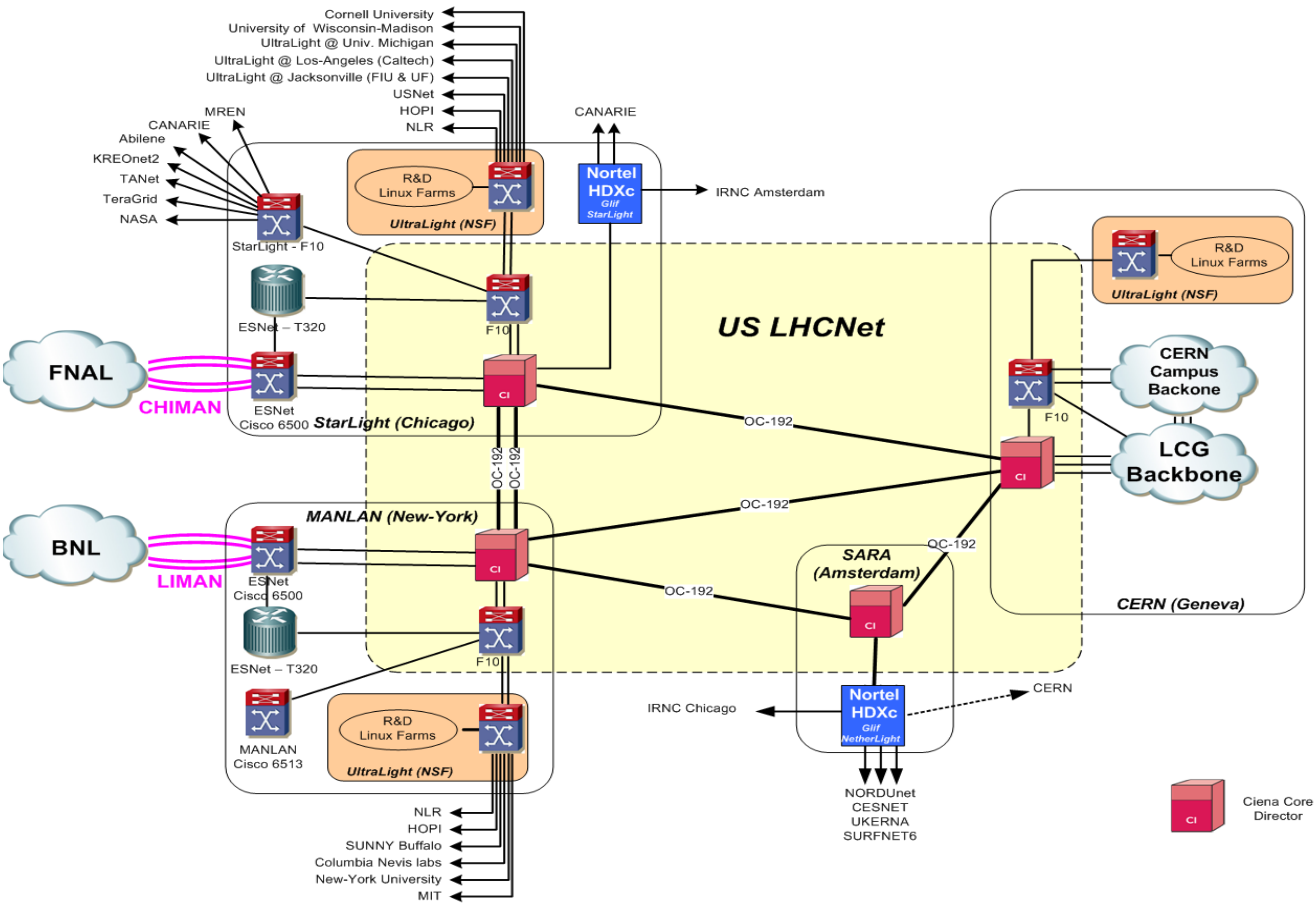
[ICFA SCIC report]

LHC Tier 0, 1, and 2 Connectivity Requirements Summary

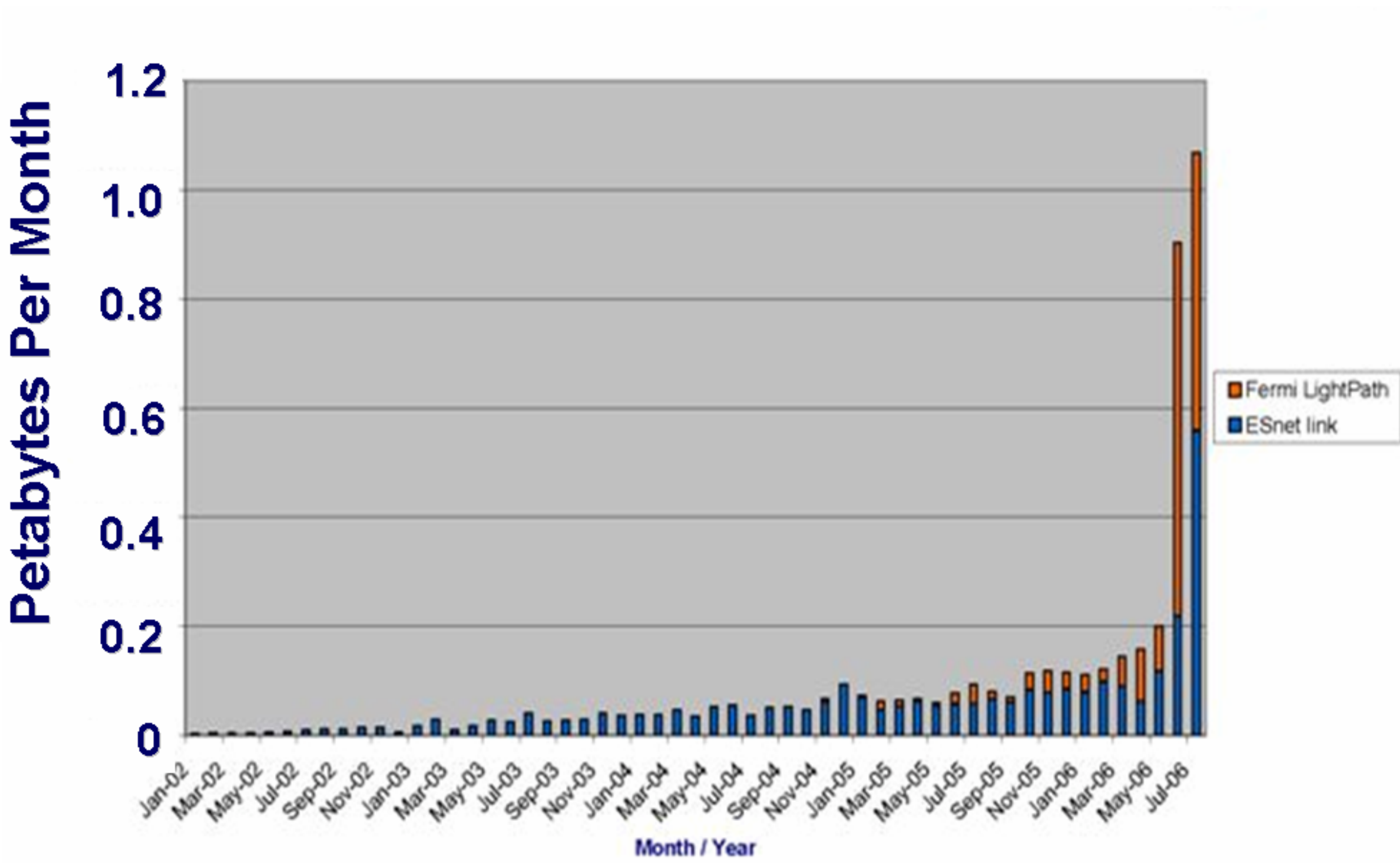


- **Direct connectivity T0-T1-T2**
    - USLHCNet to ESnet to Abilene
- **Backup connectivity**
    - SDN, GLIF, VCs

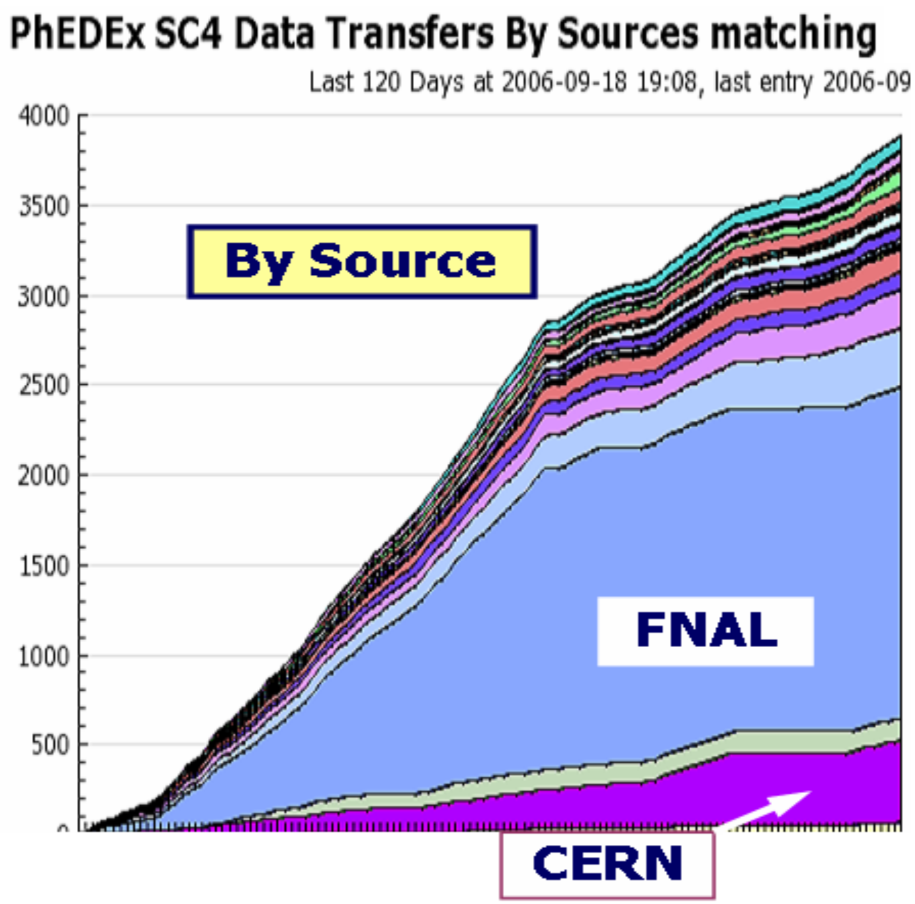
US-CERN backbone ("US LHCNet") [ICFA SCIC]

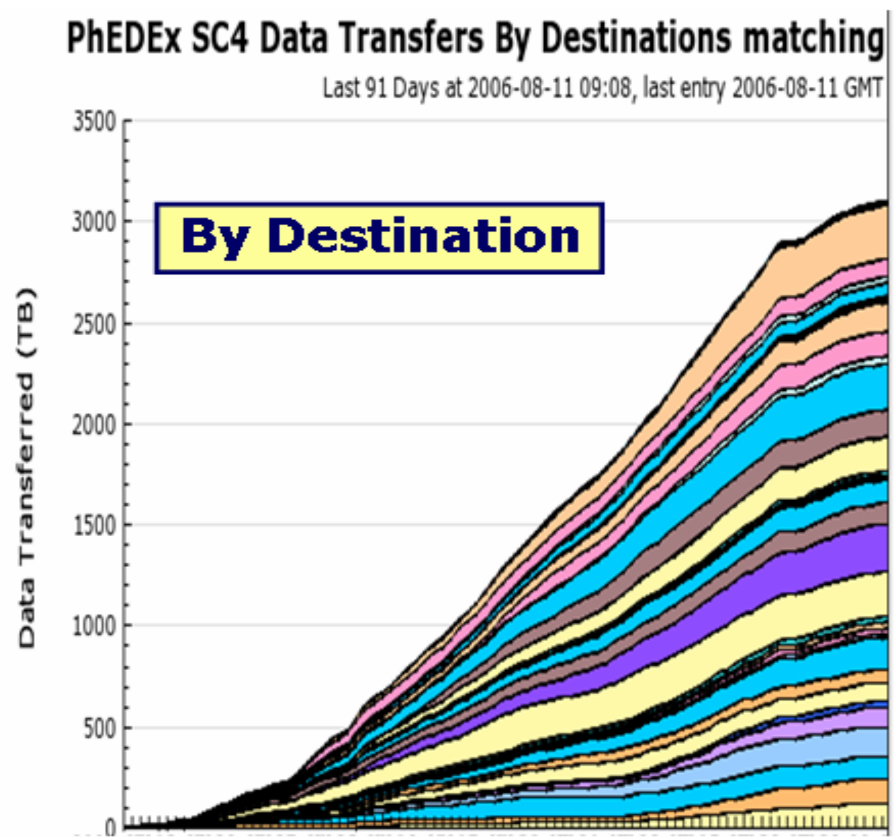


Fermilab outbound traffic (Petabytes/month) through July 2006, showing the onset of LHC Service Challenge 4 in May 2006 [ICFA SCIC]



Accumulated data (Terabytes) sent and received by CMS Tier1s and Tier2s during LHC Service Challenge 4, starting in May 2006 [ICFA SCIC]





Changing Science Environment ⇒ New Demands on Network

Science Networking Requirements Summary

- **Increased capacity**
  - Needed to accommodate a large and steadily increasing amount of data that must traverse the network
- **High network reliability**
  - Essential when interconnecting components of distributed large-scale science
- **High-speed, highly reliable connectivity between Labs and US and international R&E institutions**
  - To support the inherently collaborative, global nature of large-scale science
- **New network services to provide bandwidth guarantees**
  - Provide for data transfer deadlines for
    - remote data analysis, real-time interaction with instruments, coupled computational simulations, etc.

ESnet4 - The Response to the Requirements

I) A new network architecture and implementation strategy

- Rich and diverse network topology for flexible management and high reliability
- Dual connectivity at every level for all large-scale science sources and sinks
- A partnership with the US research and education community to build a shared, large-scale, R&E managed optical infrastructure
  - a scalable approach to adding bandwidth to the network
  - dynamic allocation and management of optical circuits

II) Development and deployment of a virtual circuit service

- Develop the service cooperatively with the networks that are intermediate between DOE Labs and major collaborators to ensure and-to-end interoperability

➢ Next Generation ESnet: I) Architecture and Configuration

- **Main architectural elements and the rationale for each element**

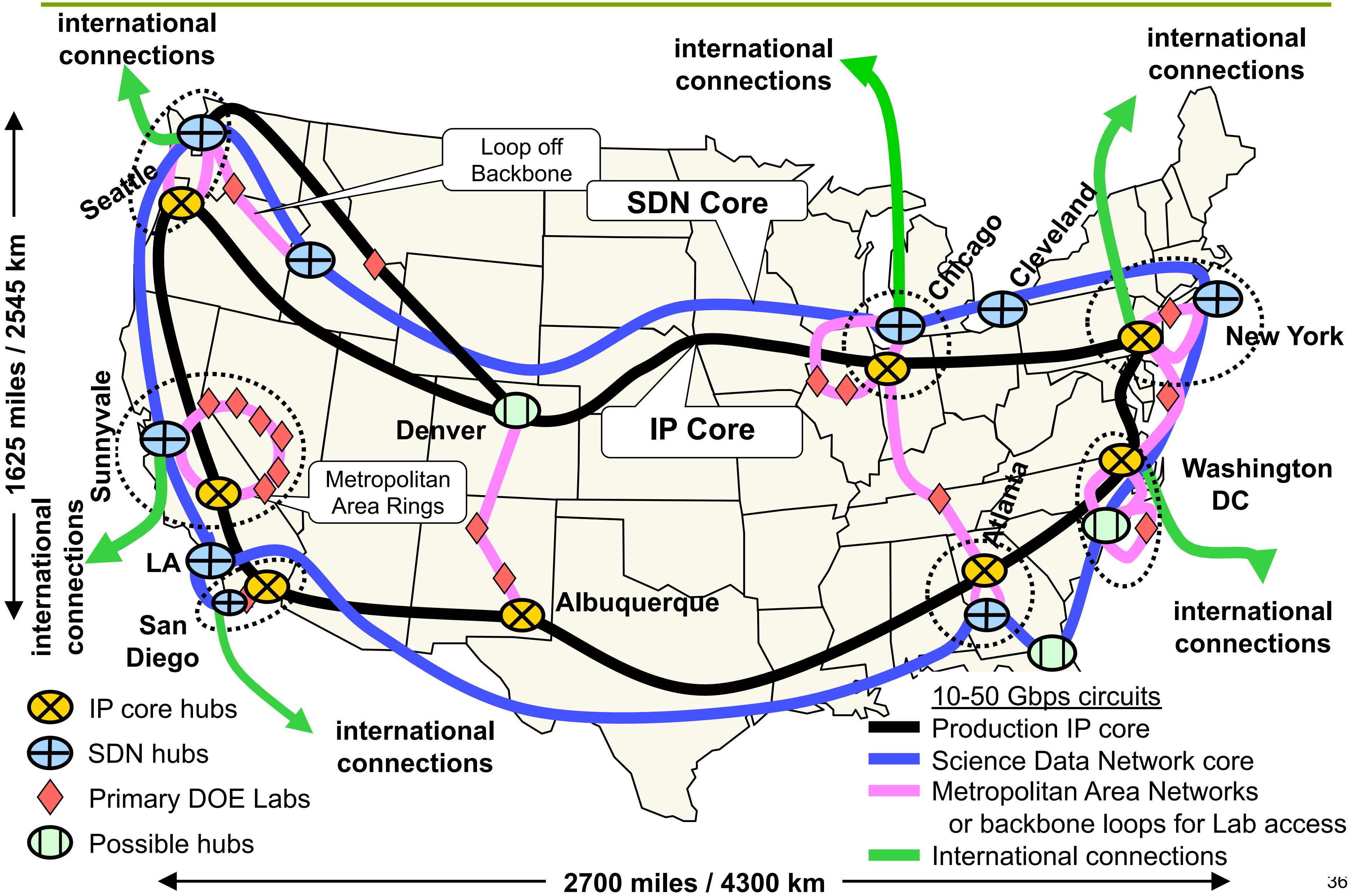
  1) A ***High-reliability IP core*** (e.g. the current ESnet core) to address
     - General science requirements
     - Lab operational requirements
     - Backup for the SDN core
     - Vehicle for science services
     - Full service IP routers

  2) ***Metropolitan Area Network*** (MAN) rings to provide
     - Dual site connectivity for reliability
     - Much higher site-to-core bandwidth
     - Support for both production IP and circuit-based traffic
     - Multiply connecting the SDN and IP cores

  2a) ***Loops off of the backbone*** rings to provide
     - For dual site connections where MANs are not practical

  3) A ***Science Data Network*** (SDN) core for
     - Provisioned, guaranteed bandwidth circuits to support large, high-speed science data flows
     - Very high total bandwidth
     - Multiply connecting MAN rings for protection against hub failure
     - Alternate path for production IP traffic
     - Less expensive router/switches
     - Initial configuration targeted at LHC, which is also the first step to the general configuration that will address all SC requirements
     - Can meet other unknown bandwidth requirements by adding lambdas

ESnet Target Architecture: IP Core+Science Data Network Core+Metro Area Rings



ESnet4

- Internet2 has partnered with Level 3 Communications Co. and Infinera Corp. for a dedicated optical fiber infrastructure with a national footprint and a rich topology - the “Internet2 Network”
  - The fiber will be provisioned with Infinera Dense Wave Division Multiplexing equipment that uses an advanced, integrated optical-electrical design
  - Level 3 will maintain the fiber and the DWDM equipment
  - The DWDM equipment will initially be provisioned to provide10 optical circuits (lambdas - λs) across the entire fiber footprint (80 λs is max.)
- ESnet has partnered with Internet2 to:
  - Help support and develop the optical infrastructure
  - Develop new circuit-oriented network services
  - Explore mechanisms that could be used for the ESnet Network Operations Center (NOC) and the Internet2/Indiana University NOC to back each other up for disaster recovery purposes

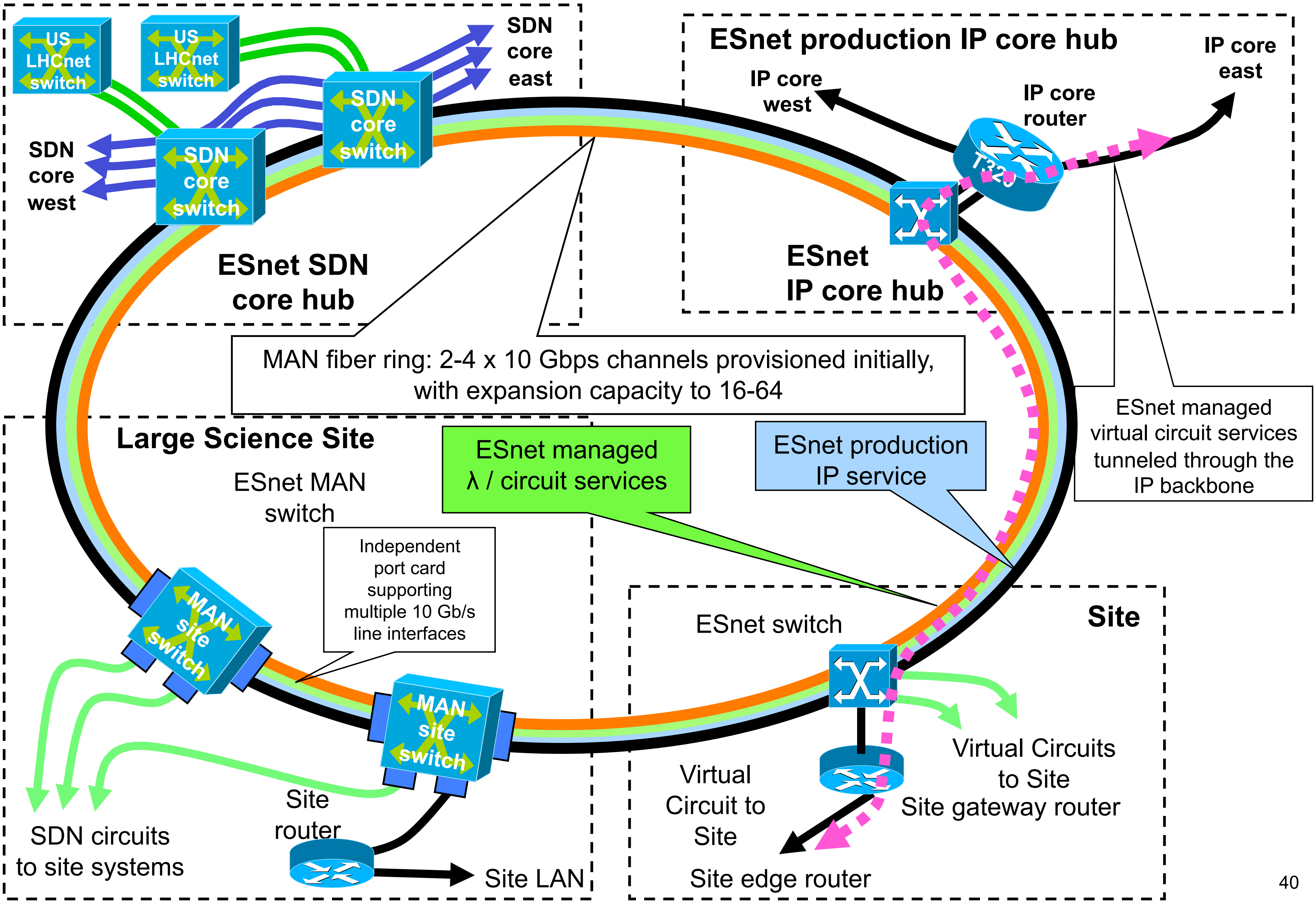
ESnet4

- ESnet will build its next generation IP network and its new circuit-oriented Science Data Network primarily on the Internet2 circuits (λs) that are dedicated to ESnet, together with a few National Lambda Rail and other circuits
  - ESnet will provision and operate its own routing and switching hardware that is installed in various commercial telecom hubs around the country, as it has done for the past 20 years
  - ESnet’s peering relationships with the commercial Internet, various US research and education networks, and numerous international networks will continue and evolve as they have for the past 20 years

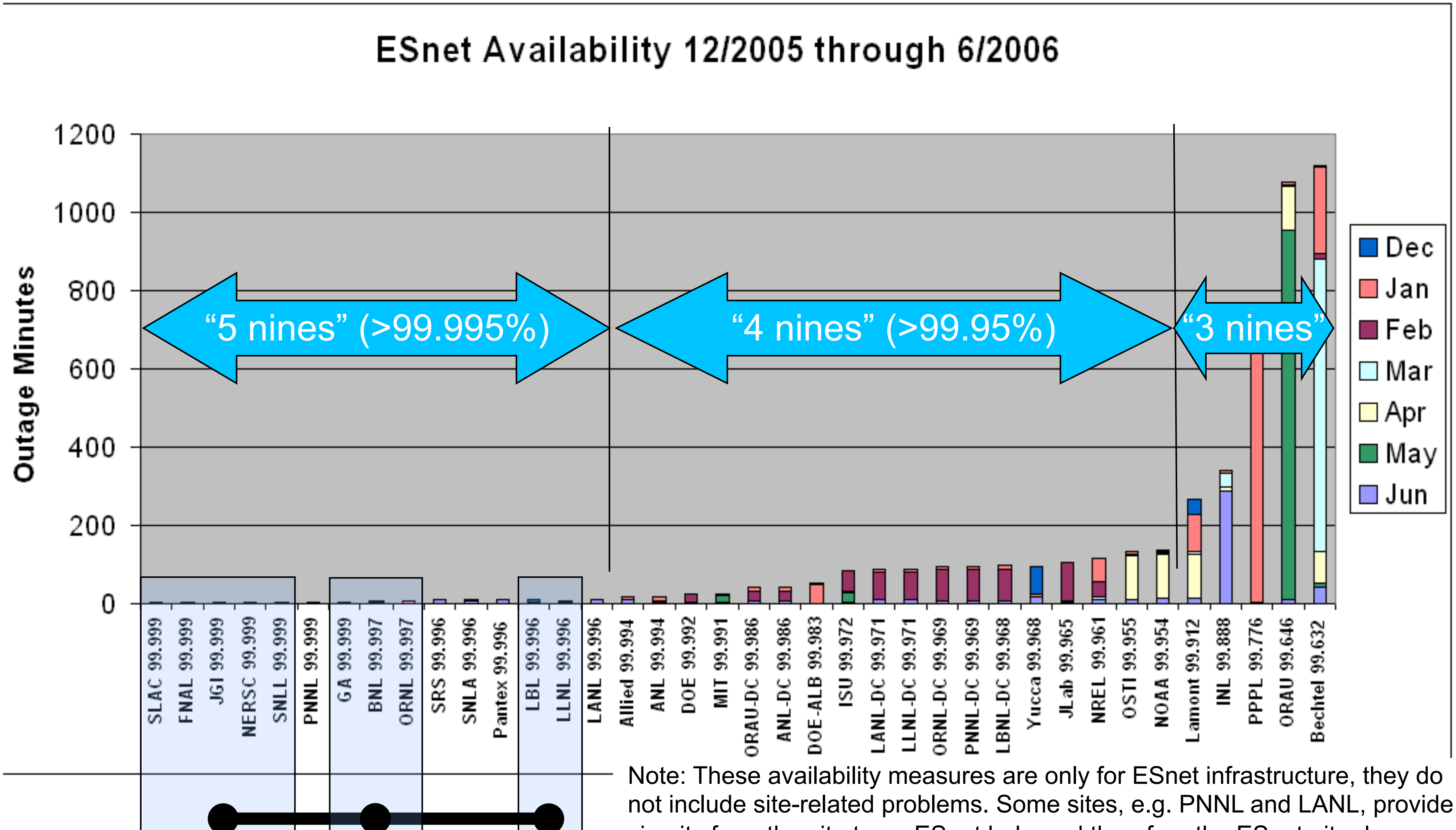
ESnet4

- ESnet4 will also involve an expansion of the multi-10Gb/s Metropolitan Area Rings in the San Francisco Bay Area, Chicago, Long Island, Newport News (VA/Washington, DC area), and Atlanta
  - provide multiple, independent connections for ESnet sites to the ESnet core network
  - expandable
- Several 10Gb/s links provided by the Labs that will be used to establish multiple, independent connections to the ESnet core
  - currently PNNL and ORNL

ESnet Metropolitan Area Network Ring Architecture for High Reliability Sites



ESnet is a Highly Reliable Infrastructure



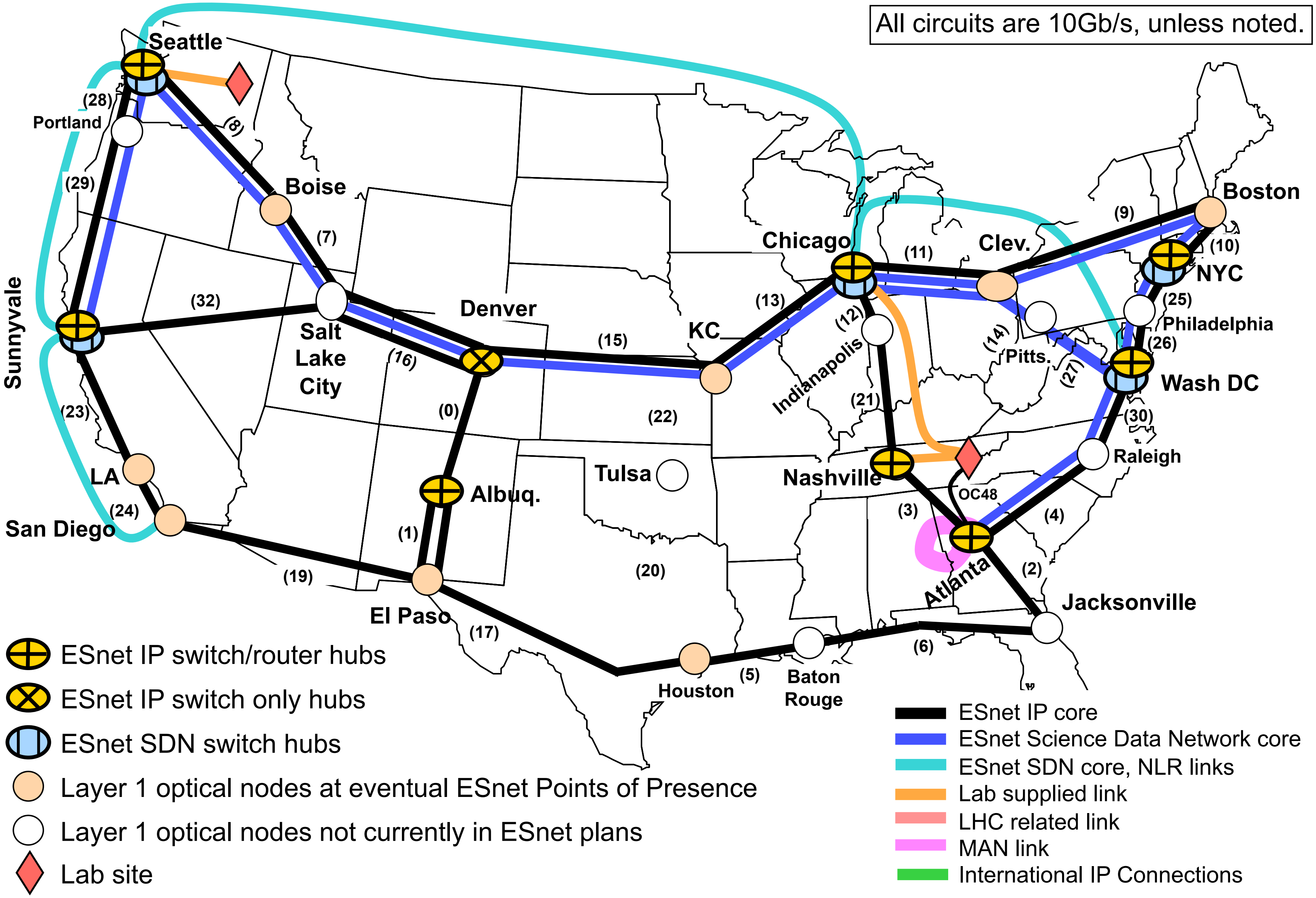
Dually connected sites

Note: These availability measures are only for ESnet infrastructure, they do not include site-related problems. Some sites, e.g. PNNL and LANL, provide circuits from the site to an ESnet hub, and therefore the ESnet-site demarc is at the ESnet hub (there is no ESnet equipment at the site. In this case, circuit outages between the ESnet equipment and the site are considered site issues and are not included in the ESnet availability metric.

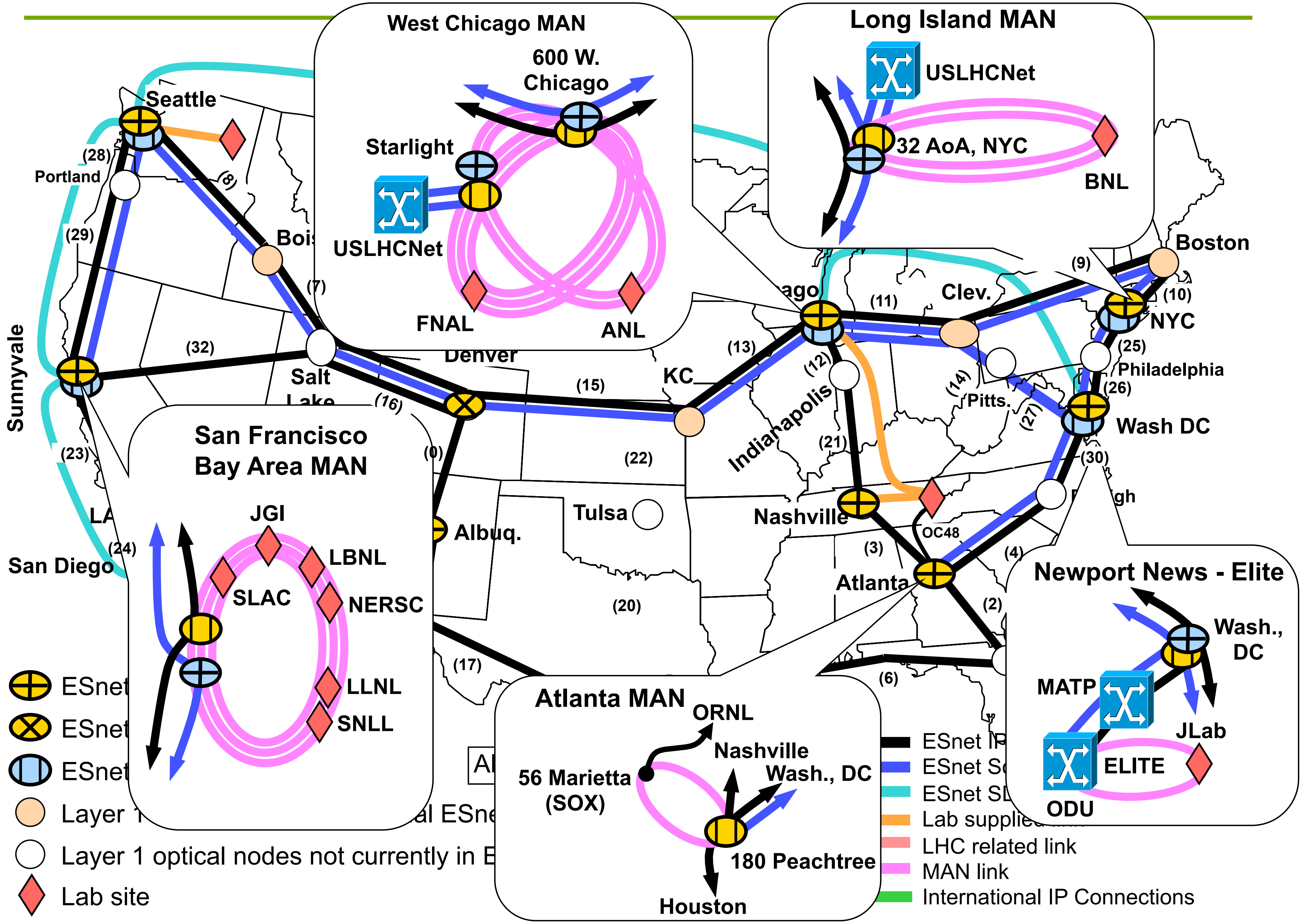
ESnet4 Roll Out

ESnet4 IP + SDN Configuration, mid-September, 2007

All circuits are 10Gb/s, unless noted.

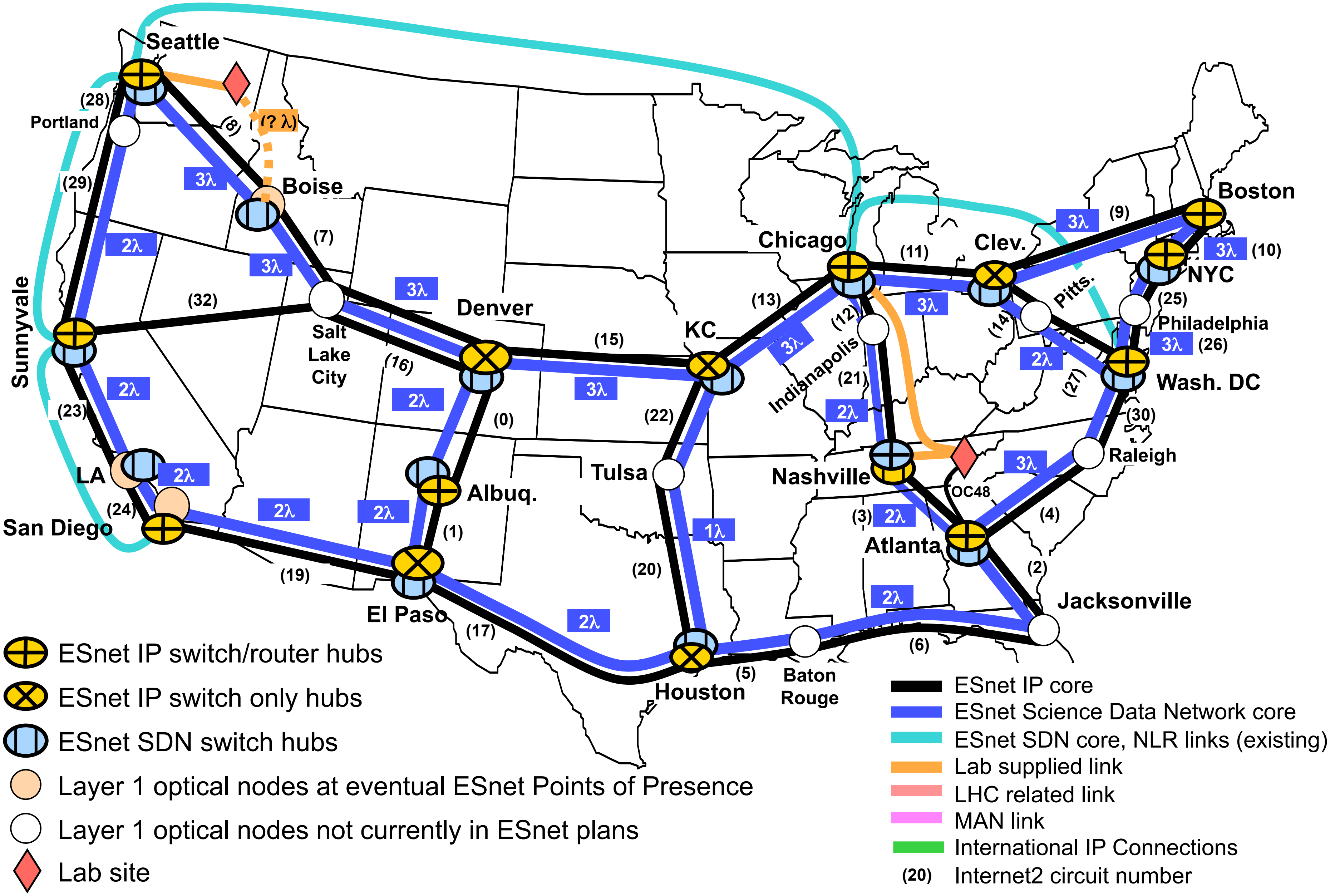


ESnet4 Metro Area Rings, 2007 Configurations

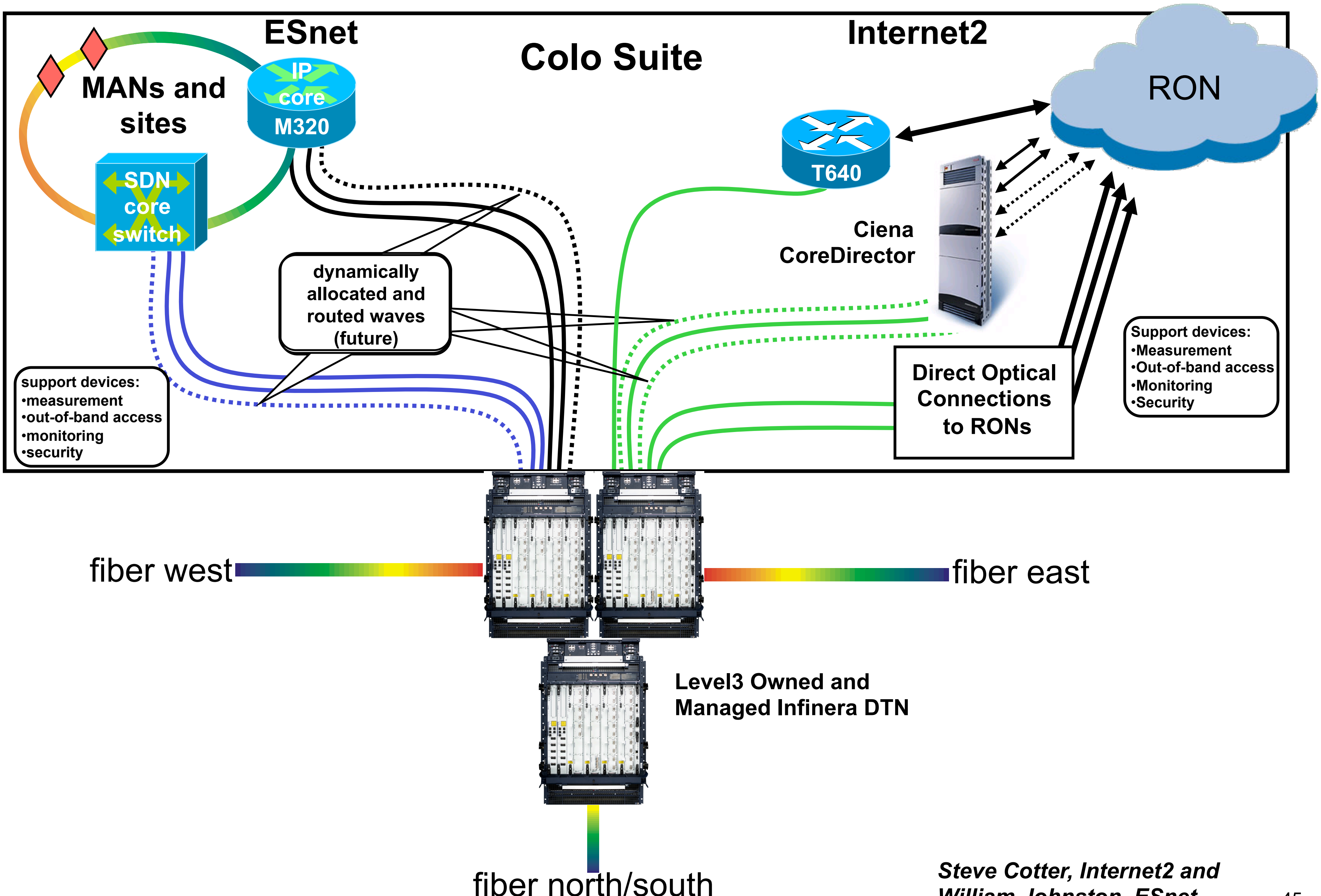


ESnet4 2009 Configuration

(Some of the circuits may be allocated dynamically from shared a pool.)

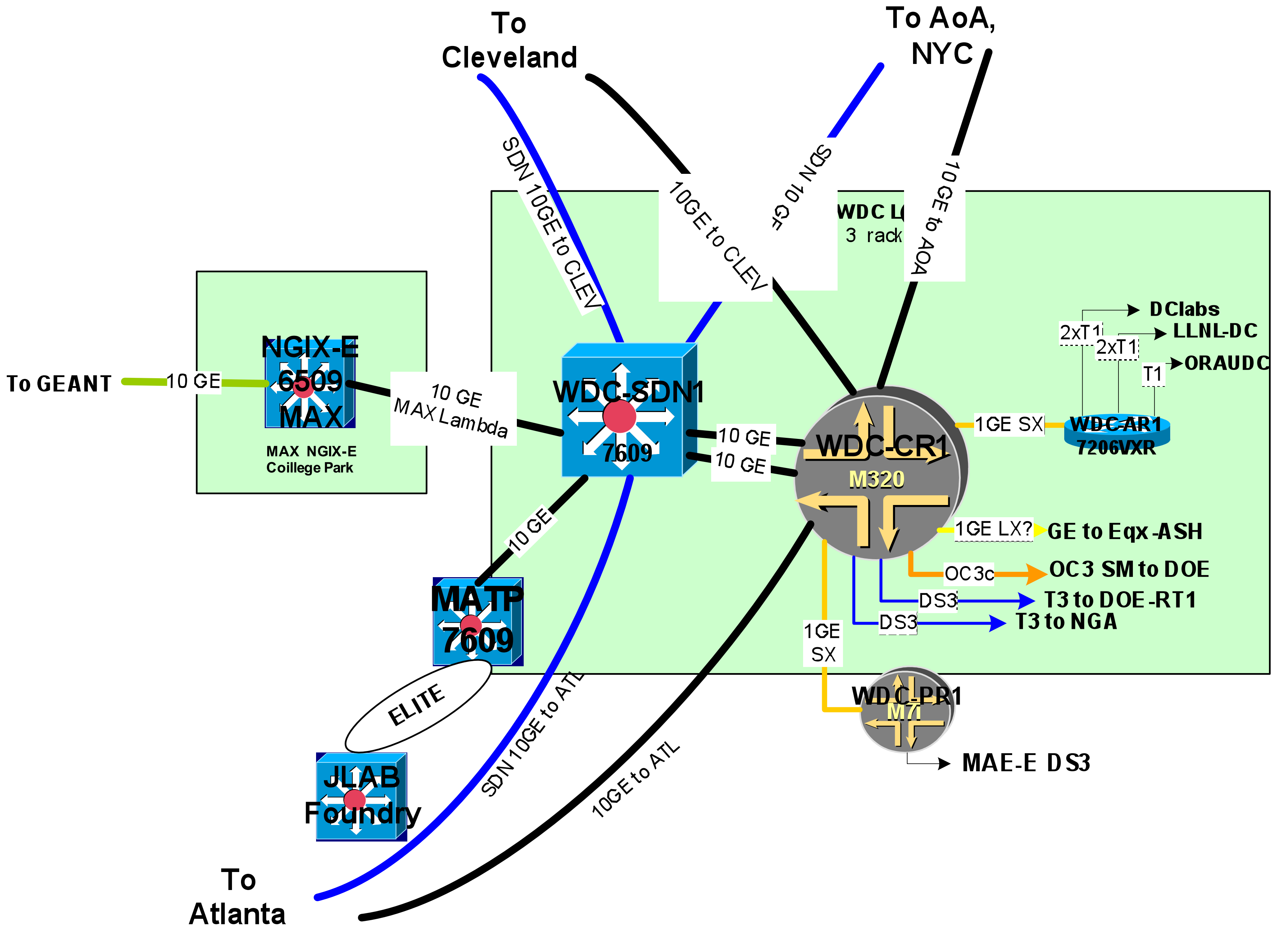


Internet2 and ESnet Optical Node

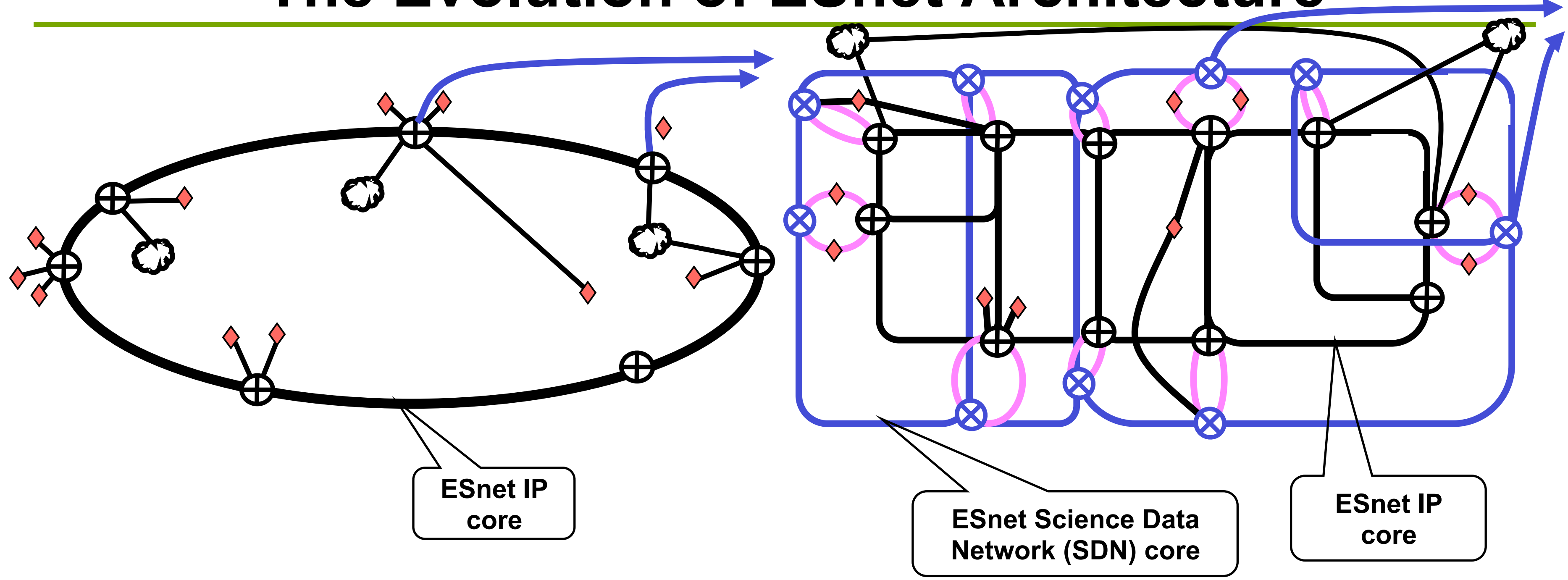


Steve Cotter, Internet2 and William Johnston, ESnet

Typical ESnet4 Hub



The Evolution of ESnet Architecture

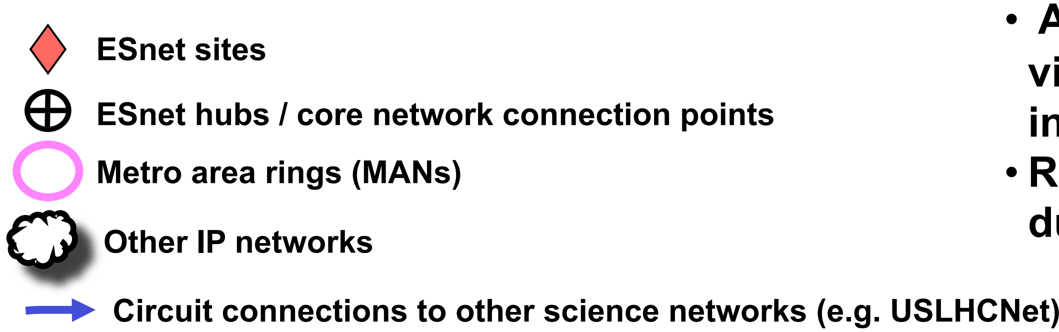


ESnet to 2005:

- A routed IP network with sites singly attached to a national core ring

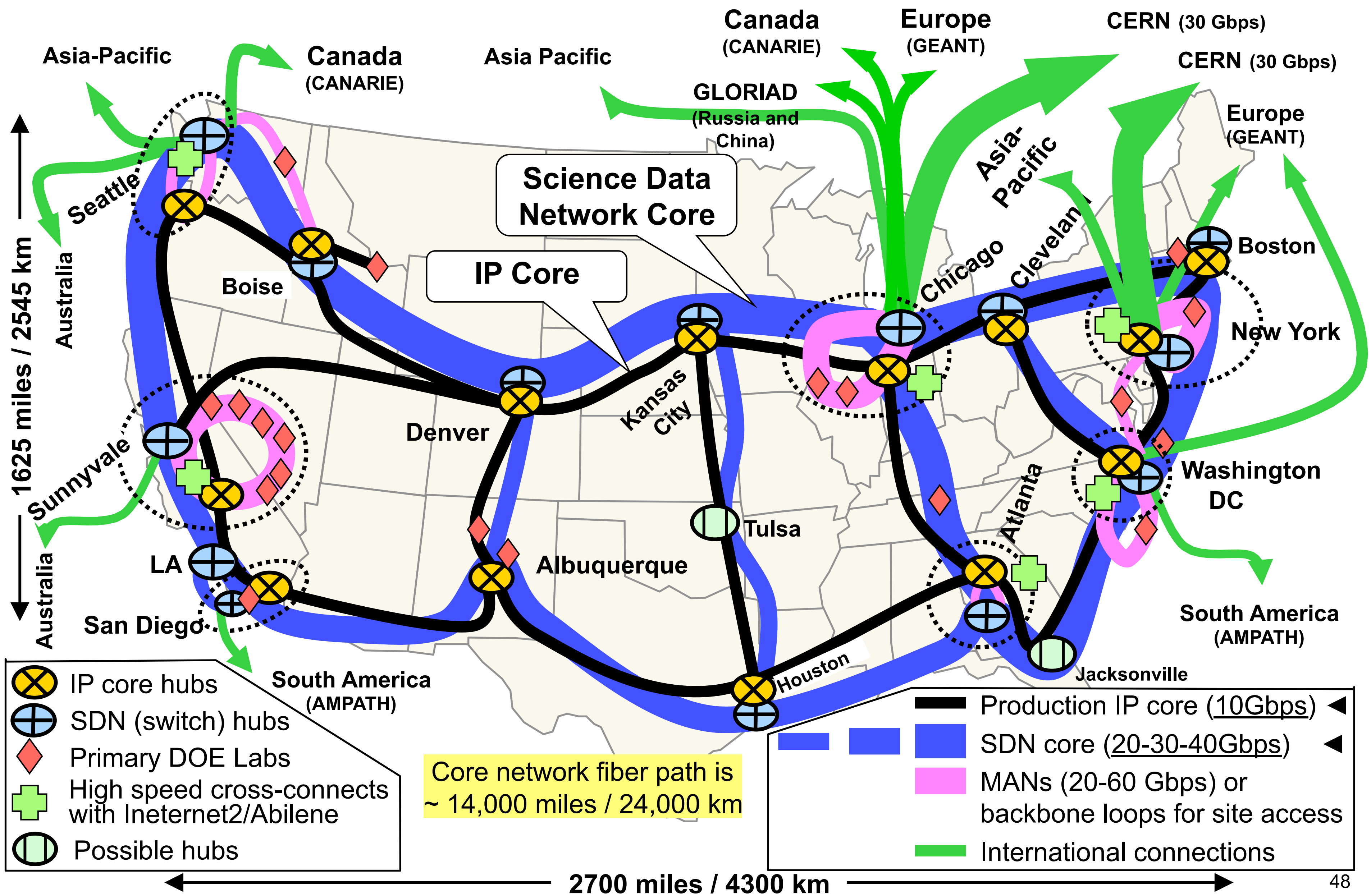
ESnet from 2006-07:

- A routed IP network with sites dually connected on metro area rings or dually connected directly to core ring
- A switched network providing virtual circuit services for data-intensive science
- Rich topology offsets the lack of dual, independent national cores



ESnet4 Planed Configuration

Core networks: 40-50 Gbps in 2009-2010, 160-400 Gbps in 2011-2012



New Network Service: Virtual Circuits

- Traffic isolation and traffic engineering
  - Provides for high-performance, non-standard transport mechanisms that cannot co-exist with commodity TCP-based transport
  - Enables the engineering of explicit paths to meet specific requirements
    - e.g. bypass congested links, using lower bandwidth, lower latency paths
- Guaranteed bandwidth (Quality of Service (QoS))
  - User specified bandwidth
  - Addresses deadline scheduling
    - Where fixed amounts of data have to reach sites on a fixed schedule, so that the processing does not fall far enough behind that it could never catch up – very important for experiment data analysis
- Reduces cost of handling high bandwidth data flows
  - Highly capable routers are not necessary when every packet goes to the same place
  - Use lower cost (factor of 3-5x) switches to relatively route the packets
- Secure
  - The circuits are “secure” to the edges of the network (the site boundary) because they are managed by the control plane of the network which is isolated from the general traffic
- Provides end-to-end connections between Labs and collaborator institutions

Virtual Circuit Service Functional Requirements

- Support user/application VC reservation requests
  - Source and destination of the VC
  - Bandwidth, latency, start time, and duration of the VC
  - Traffic characteristics (e.g. flow specs) to identify traffic designated for the VC
- Manage allocations of scarce, shared resources
  - Authentication to prevent unauthorized access to this service
  - Authorization to enforce policy on reservation/provisioning
  - Gathering of usage data for accounting
- Provide virtual circuit setup and teardown mechanisms and security
  - Widely adopted and standard protocols (such as MPLS and GMPLS) are well understood within a single domain
  - Cross domain interoperability is the subject of ongoing, collaborative development
  - secure and-to-end connection setup is provided by the network control plane
  - accommodate heterogeneous circuit abstraction (e.g.. MPLS, GMPLS, VLANs, VCAT/LCAS)
- Enable the claiming of reservations
  - Traffic destined for the VC must be differentiated from “regular” traffic
- Enforce usage limits
  - Per VC admission control polices usage, which in turn facilitates guaranteed bandwidth
  - Consistent per-hop QoS throughout the network for transport predictability

Oscars Approach

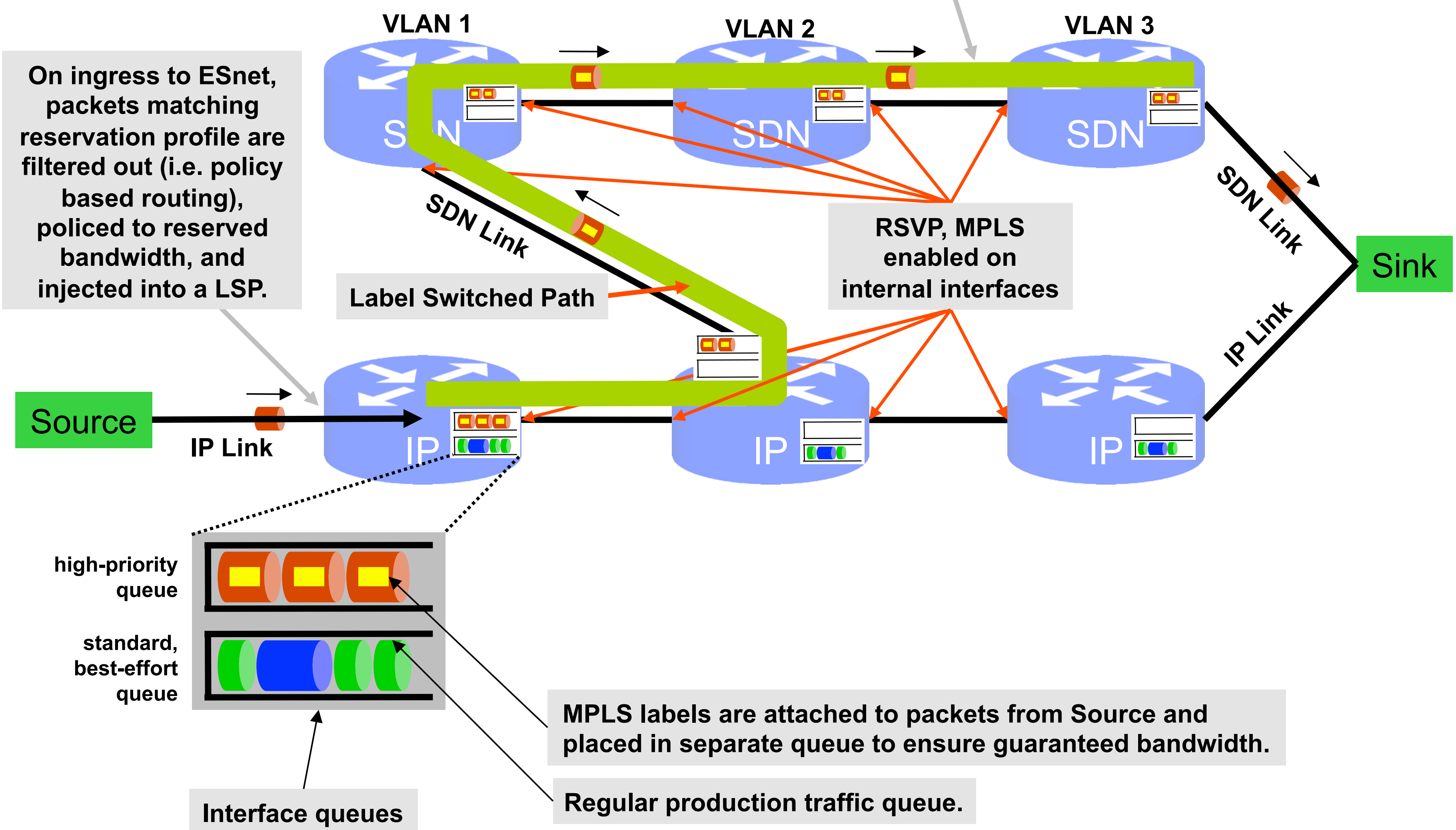
- Based on Source and Sink IP addresses, route of Label Switched Path (LSP) between ESnet border routers is determined using network topology and link usage policy
  - The OSPF configuration of the network is dumped daily into a topology database
  - Path of LSP can be explicitly directed to take SDN network
- On the SDN Ethernet switches all traffic is MPLS switched (layer 2.5)
  - MPLS is used to stitch together a collection of “local” VLANs
- On ingress to ESnet, packets matching reservation profile are “identified” (i.e. using policy based routing), policed to reserved bandwidth, and injected into a LSP
  - link policy will determine the bandwidth available for high priority queuing
  - a bandwidth scheduler keeps track of the assigned vs. available priority traffic
    - the reservation system effectively does admission control to ensure that the available priority bandwidth is never over-subscribed
    - the policer ensures that individual flows do no exceed their alloted/reserved bandwidth

OSCARS Reservations

1. A user submits a request to the RM specifying start and end times, bandwidth requirements, the source and destination hosts
2. Using the source and destination host information submitted by the user, the ingress and egress border routers, and circuit path (MPLS LSP) is determined
3. This information is stored by the BSS in a database, and a script periodically checks to see if the PSS needs to be contacted, either to create or tear down the circuit
4. At the requested start time, the PSS configures the ESnet provider edge (PE) router (at the start end of the path) to create an LSP with the specified bandwidth
5. Each router along the route receives the path setup request via the Reservation Resource Protocol (RSVP) and commits bandwidth (if available) creating an end-to-end LSP. The RM is notified by RSVP if the end-to-end path cannot be established.
6. Packets from the source (e.g. experiment) are routed through the site's LAN production path to ESnet's PE router. On entering the PE router, these packets are identified and filtered using flow specification parameters (e.g. source/destination IP address/port numbers) and policed at the specified bandwidth. The packets are then injected into the LSP and switched (using MPLS) through the network to its destination (e.g. computing cluster).
7. A notification of the success or failure of LSP setup is passed back to the RM so that the user can be notified and the event logged for auditing purposes
8. At the requested end time, the PSS tears down the LSP

The Mechanisms Underlying OSCARS

Based on Source and Sink IP addresses, route of LSP between ESnet border routers is determined using topology information from OSPF-TE. Path of LSP can be explicitly directed to take SDN network. On the SDN Ethernet switches all traffic is MPLS switched (layer 2.5), which stitches together VLANs



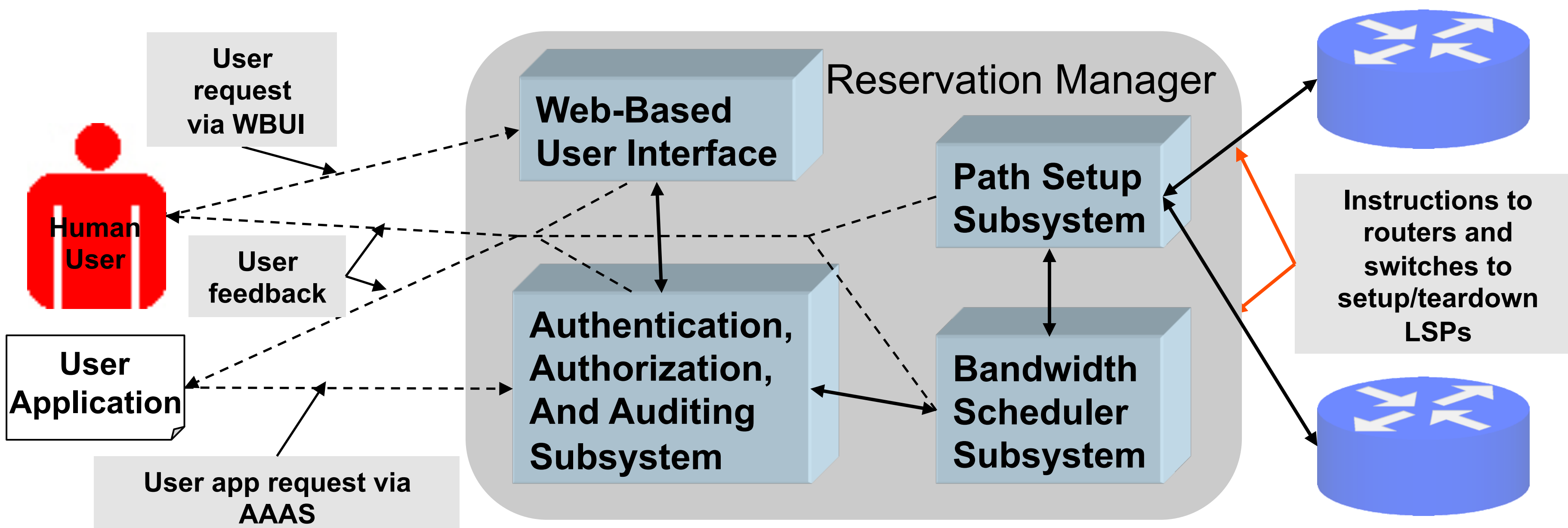
MPLS labels are attached to packets from Source and placed in separate queue to ensure guaranteed bandwidth.

Regular production traffic queue.

ESnet Virtual Circuit Service: OSCARS
(On-demand Secured Circuits and Advanced Reservation System)

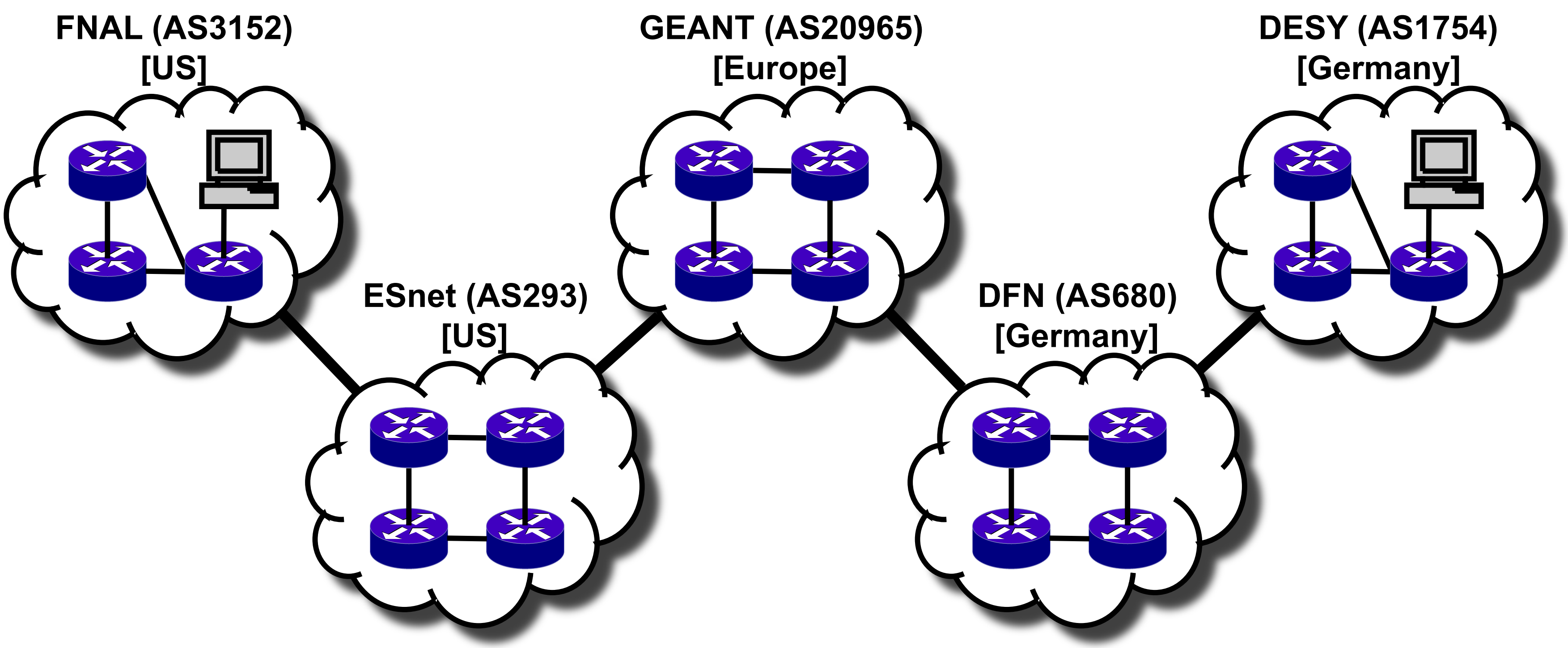
Software Architecture (see Ref. 9)

- Web-Based User Interface (WBUI) will prompt the user for a username/ password and forward it to the AAAS.
- Authentication, Authorization, and Auditing Subsystem (AAAS) will handle access, enforce policy, and generate usage records.
- Bandwidth Scheduler Subsystem (BSS) will track reservations and map the state of the network (present and future).
- Path Setup Subsystem (PSS) will setup and teardown the on-demand paths (LSPs).



Environment of Science is Inherently Multi-Domain

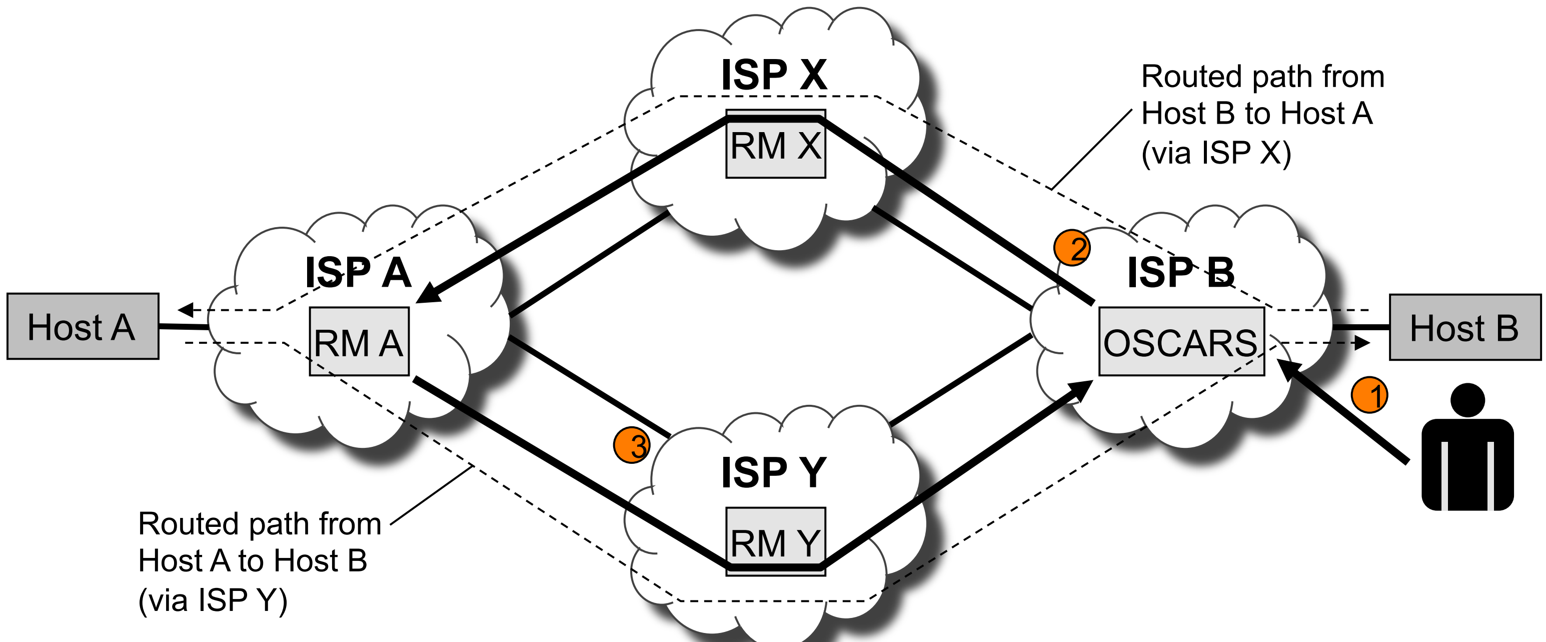
- End points will be at independent institutions – campuses or research institutes - that are served by ESnet, Abilene, GÉANT, and their regional networks
  - Complex inter-domain issues – typical circuit will involve five or more domains - of necessity this involves collaboration with other networks
  - For example, a connection between FNAL and DESY involves five domains, traverses four countries, and crosses seven time zones



Inter-domain Reservations: A Tough Problem

- **Motivation**:
  - For a virtual circuit service to be successful, it must
    - Be end-to-end, potentially crossing several administrative domains
    - Have consistent network service guarantees throughout the circuit
- **Observation:**
  - Setting up an intra-domain circuit is easy compared with coordinating an inter-domain circuit
- **Issues:**
  - Cross domain authentication *and* authorization
    - A mechanism to authenticate and authorize a bandwidth on-demand (BoD) circuit request must be agreed upon in order to automate the process
  - Multi-domain Acceptable Use Policies (AUPs)
    - Domains may have very specific AUPs dictating what the BoD circuits can be used for and where they can transit/terminate
  - Domain specific service offerings
    - Domains must have way to guarantee a certain level of service for BoD circuits
  - Security concerns
    - Are there mechanisms for a domain to protect itself (e.g. RSVP filtering)

Inter-domain Path Setup



1. On receiving the request from the user, OSCARS computes the virtual circuit path and determines the downstream AS (ISP X).
2. The request is then encapsulated in a message forwarded across the network (ISP X) towards Host A, crossing all intervening reservations systems (RM X), until it reaches the last reservation system (RM A) that has administrative control over the network (ISP A) that Host A is attached to.
3. The remote reservation system (RM A) then computes the path of the virtual circuit, and initiates the bandwidth reservation requests from Host A towards Host B (via ISP Y). This can be especially complex when the path back (from Host B to A) is asymmetric and traverses AS's (e.g. ISP Y) that were not traversed on the forward path, causing the local OSCARS to see the path originating from a different AS than it originally sent the request to.

OSCARS: Guaranteed Bandwidth VC Service For SC Science

- To ensure compatibility, the design and implementation is done in collaboration with the other major science R&E networks and end sites
  - Internet2: Bandwidth Reservation for User Work (BRUW)
    - Development of common code base
  - GEANT: Bandwidth on Demand (GN2-JRA3), Performance and Allocated Capacity for End-users (SA3-PACE) and Advance Multi-domain Provisioning System (AMPS) extends to NRENs
  - BNL: TeraPaths - A QoS Enabled Collaborative Data Sharing Infrastructure for Peta-scale Computing Research
  - GA: Network Quality of Service for Magnetic Fusion Research
  - SLAC: Internet End-to-end Performance Monitoring (IEPM)
  - USN: Experimental Ultra-Scale Network Testbed for Large-Scale Science
- In its current phase this effort is being funded as a research project by the Office of Science, Mathematical, Information, and Computational Sciences (MICS) Network R&D Program
- A prototype service has been deployed as a proof of concept
  - To date more then 30 accounts have been created for beta users, collaborators, and developers
  - More then 500 user reservation requests have been processed

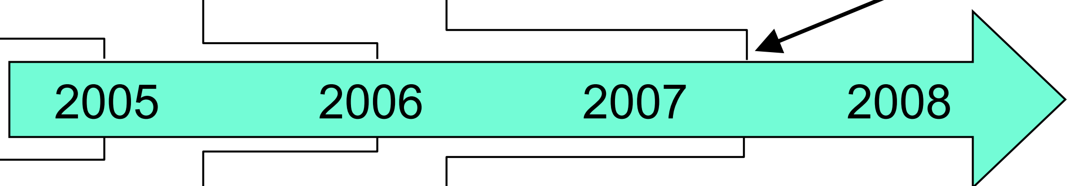
OSCARS Update

- Completed porting OSCARS from Perl to Java to better support web-services
  - This is now the common code base for OSCARS and I2's BRUW
- Paper on OSCARS was accepted by the IEEE GridNets
- Collaborative efforts
  - Working with I2 and DRAGON to support interoperability between OSCARS/BRUW and DRAGON
    - currently in the process of installing an instance of DRAGON in ESnet
  - Working with I2, DRAGON, and TeraPaths (Brookhaven Lab) to determine an appropriate interoperable AAI (authentication and authorization infrastructure) framework (this is in conjunction with GEANT2's JRA5)
  - Working with DICE Control Plane group to determine schema and methods of distributing topology and reachability information
    - DICE=Internet2, ESnet, GEANT, CANARIE/UCLP; see http://www.garr.it/dice/presentation.htm for presentations from the last meeting
  - Working with Tom Lehman (DRAGON), Nagi Rao (USN), Nasir Ghani (Tennessee Tech) on multi-level, multi-domain hybrid network performance measurements
    - this is part of the Hybrid Multi-Layer Network Control for Emerging Cyberinfrastures project funded by Thomas Ndousse)

ESnet Virtual Circuit Service Roadmap

- Dedicated virtual circuits
- Dynamic virtual circuit allocation
- Generalized MPLS (GMPLS)

Initial production service



Full production service

- Interoperability between GMPLS circuits, VLANs, and MPLS circuits (layer 1-3)
- Interoperability between VLANs and MPLS circuits (layer 2 & 3)
- Dynamic provisioning of Multi-Protocol Label Switching (MPLS) circuits in IP nets (layer 3) and in VLANs for Ethernets (layer 2)

ESnet Network Measurements
ESCC Feb 15 2007

Joe Metzger
metzger@es.net

Measurement Motivations

- Users dependence on the network is increasing
  - Distributed Applications
  - Moving Larger Data Sets
  - The network is becoming a critical part of large science experiments
- The network is growing more complex
  - 6 core devices in 05’, 25+ in 08’
  - 6 core links in 05’, 40+ in 08’, 80+ by 2010?
- Users continue to report performance problems
  - ‘wizards gap’ issues
- The community needs to better understand the network
  - We need to be able to demonstrate that the network is good.
  - We need to be able to detect and fix subtle network problems.

perfSONAR

- perfSONAR is a global collaboration to design, implement and deploy a network measurement framework.
  - Web Services based Framework
    - Measurement Archives (MA)
    - Measurement Points (MP)
    - Lookup Service (LS)
    - Topology Service (TS)
    - Authentication Service (AS)
  - Some of the currently Deployed Services
    - Utilization MA
    - Circuit Status MA & MP
    - Latency MA & MP
    - Bandwidth MA & MP
    - Looking Glass MP
    - Topology MA
  - This is an Active Collaboration
    - The basic framework is complete
    - Protocols are being documented
    - New Services are being developed and deployed.

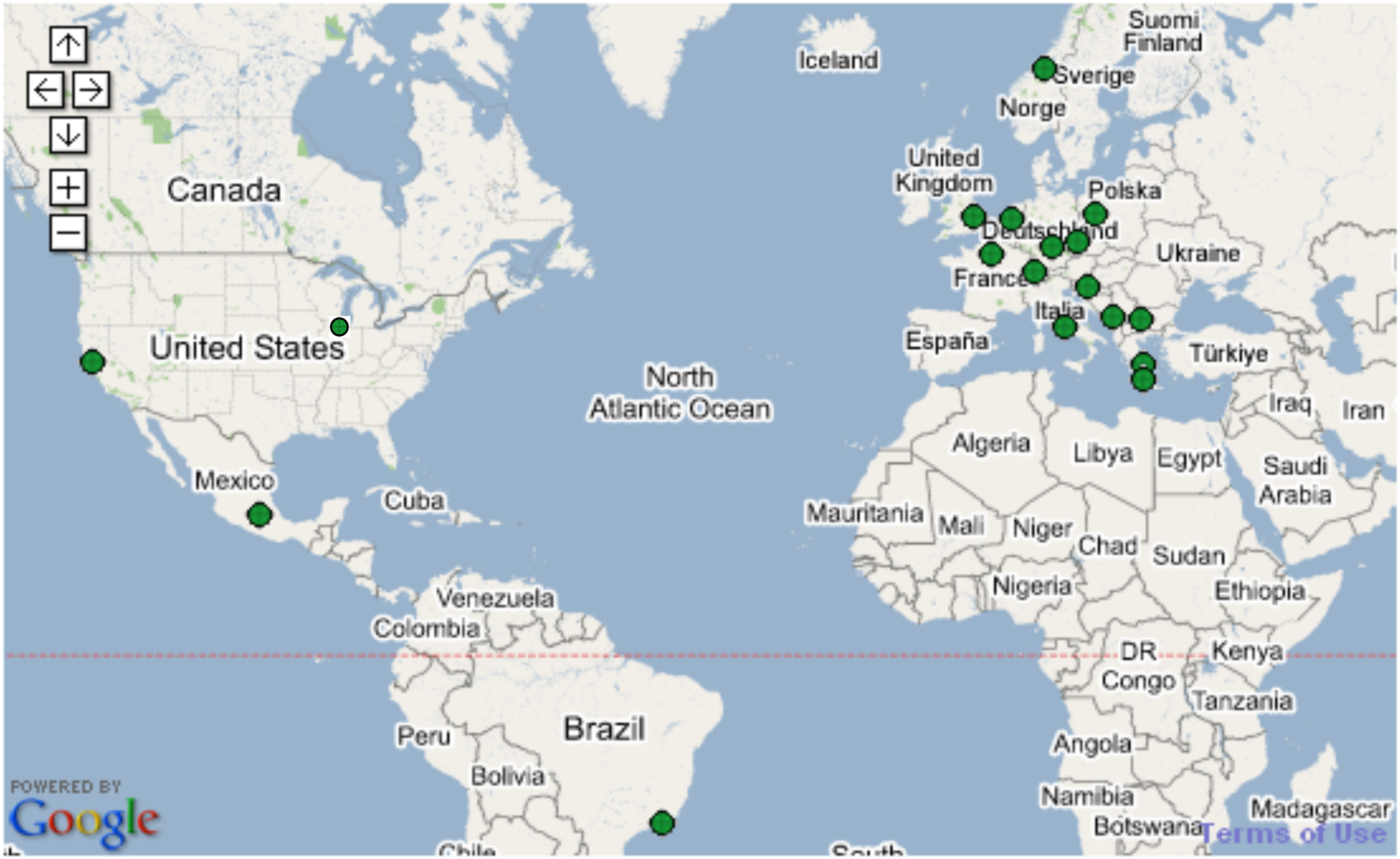
perfSONAR Collaborators

- ARNES
- Belnet
- CARnet
- CESnet
- Dante
- University of Delaware
- DFN
- ESnet
- FCCN
- FNAL
- GARR
- GEANT2
- Georga Tech
- GRNET
- Internet2
- IST
- POZNAN Supercomputing Center
- Red IRIS
- Renater
- RNP
- SLAC
- SURFnet
- SWITCH
- Uninett

* Plus others who are contributing, but haven't added their names to the list on the WIKI.

perfSONAR Deployments

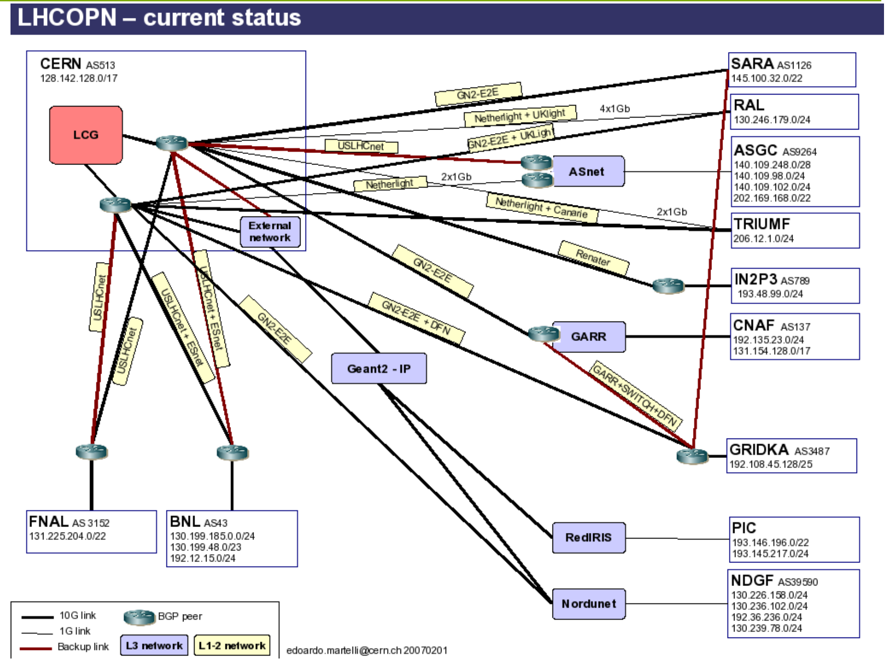
16+ different networks have deployed at least 1 perfSONAR service (Jan 2007)



ESnet perfSONAR Progress

- ESnet Deployed Services
  - Link Utilization Measurement Archive
  - Virtual Circuit Status
- In Development
  - Active Latency and Bandwidth Tests
  - Topology Service
  - Additional Visualization capabilities
- perfSONAR visualization tools showing ESnet data
  - Link Utilization
    - perfSONARUI
      - http://perfsonar.acad.bg/
    - VisualPerfSONAR
      - https://noc-mon.srce.hr/visual_perf
    - Traceroute Visualizer
      - https://performance.es.net/cgi-bin/level0/perfsonar-trace.cgi
  - Virtual Circuit Status
    - E2EMon (for LHCOPN Circuits)
      - http://cnmdev.lrz-muenchen.de/e2e/lhc/G2_E2E_index.html

LHCOPN Monitoring



- LHCOPN
  - An Optical Private Network connecting LHC Teir1 centers around the world to CERN.
  - The circuits to two of the largest Tier1 centers, FERMI & BNL cross ESnet
- E2Emon
  - An application developed by DFN for monitoring circuits using perfSONAR protocols
- E2ECU
  - End to End Coordination Unit that uses E2Emon to monitor LHCOPN Circuits
  - Run by the GEANT2 NOC

E2EMON and perfSONAR

- E2Emon
  - An application suite developed by DFN for monitoring circuits using perfSONAR protocols
- perfSONAR is a global collaboration to design, implement and deploy a network measurement framework.
  - Web Services based Framework
    - Measurement Archives (MA)
    - Measurement Points (MP)
    - Lookup Service (LS)
    - Topology Service (TS)
    - Authentication Service (AS)
  - Some of the currently Deployed Services
    - Utilization MA
    - Circuit Status MA & MP
    - Latency MA & MP
    - Bandwidth MA & MP
    - Looking Glass MP
    - Topology MA
  - This is an Active Collaboration
    - The basic framework is complete
    - Protocols are being documented
    - New Services are being developed and deployed.

E2Emon Components

- Central Monitoring Software
  - Uses perfSONAR protocols to retrieve current circuit status every minute or so from MAs and MPs in all the different domains supporting the circuits.
  - Provides a web site showing current end-to-end circuit status
  - Generates SNMP traps that can be sent to other management systems when circuits go down
- MA & MP Software
  - Manages the perfSONAR communications with the central monitoring software
  - Requires an XML file describing current circuit status as input.
- Domain Specific Component
  - Generates the XML input file for the MA or MP
  - Multiple development efforts in progress, but no universal solutions
    - CERN developed one that interfaces to their abstraction of the Spectrum NMS DB
    - DANTE developed one that interfaces with the Acatel NMS
    - ESnet developed one that uses SNMP to directly poll router interfaces
    - FERMI developed one that uses SNMP to directly poll router interfaces
    - Others under development

E2Emon Central Monitoring Software

http://cnmdev.lrz-muenchen.de/e2e/lhc/G2_E2E_index.html

E2E Link Monitoring System - Mozilla Firefox

File Edit View Go Bookmarks Tools Help

Start page

E2ECU view

All E2E Links
Problem Links

Domain view

CERN
DFN
ESNET
FERMI
GARR
GEANT2
IN2P3
SURFNET
SWITCH
USLHCNET

Project view

IGTMD
LHCOPN

Availability Statistics

Current Month
Last Month

E2E Links for Project LHCOPN

E2E Link ID	State Oper	State Admin	Additional Info
CERN-GRIDKA-LHCOPN-001	Up	Normal Oper.	
CERN-FERMI-LHCOPN-002	Up	Normal Oper.	Error: E2E Link is **not** contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
CERN-CNAF-LHCOPN-001	Down	Maintenance	
CERN-RAL-LHCOPN-001	Up	Normal Oper.	Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
CERN-FERMI-LHCOPN-001	Up	Normal Oper.	
CERN-BNL-LHCOPN-002	Up	Normal Oper.	Error: E2E Link is **not** contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
GRIDKA-SARA-LHCOPN-001	Up	Normal Oper.	Error: E2E Link is **not** contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
CERN-NDGF-LHCOPN-001	Up	Normal Oper.	Error: E2E Link is **not** contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
CERN-IN2P3-LHCOPN-001	Up	Normal Oper.	Error: E2E Link is **not** contiguous (End Point missing or gap found) Warning: Operational state is known not for all involved links Warning: Administrative state is known not for all involved links
CERN-BNL-LHCOPN-001	Up	Normal Oper.	
CNAF-GRIDKA-LHCOPN-001	Up	Normal Oper.	
CERN-SARA-LHCOPN-001	Up	Normal Oper.	

Page generated at 2007-02-09, 17:25:06 MET

Start page

E2ECU view

All E2E Links
Problem Links

Domain view

CERN
DFN
ESNET
FERMI
GARR
GEANT2
IN2P3
SURFNET
SWITCH
USLHCNET

Project view

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	-	2007-02-09 T17:20:33+01:00	-	2007-01-26 T13:15:19+01:00	2007-02-09 T16:20:36.0	-	2007-02-09T16:20:36.0	-	2007-02-09 T16:20:36.0	2007-02-09 T16:20:00.0-6:00	-	2007-02-09 T16:20:00.0-6:00	-

ESnet4 Hub Measurement Hardware

- Latency
  - 1U Server with one of:
    - EndRun Praecis CT CDMA Clock
    - Meinberg TCR167PCI IRIG Clock
    - Symmetricom bc637PCI-U IRIG Clock
- Bandwidth
  - 4U dual Opteron server with one of:
    - Myricom 10GE NIC
      - 9.9 Gbps UDP streams
      - ~6 Gbps TCP streams
      - Consumes 100% of 1 CPU
    - Chelsio S320 10GE NIC
      - Should do 10G TCP & UDP with low CPU Utilization
      - Has interesting shaping possibilities
      - Still under testing…

Network Measurements ESnet is Collecting

- SNMP Interface Utilization
  - Collected every minute
    - For MRTG & Monthly Reporting
- Circuit Availability
  - Currently based on SNMP Interface up/down status
  - Limited to LHCOPN and Service Trial circuits for now
- NetFlow Data
  - Sampled on our boundaries
- Latency
  - OWAMP

ESnet Performance Center

- Web Interface to run Network Measurements
- Available to ESnet sites
- Supported Tests
  - Ping
  - Traceroute
  - IPERF
  - Pathload, Pathrate, Pipechar
    - (Only on GE systems)
- Test Hardware
  - GE testers in Qwest hubs
    - TCP iperf tests max at ~600 Mbps.
  - 10GE testers are being deployed in ESnet4 hubs
    - Deployed in locations where we have Cisco 6509 10GE Interfaces
    - Available via Performance Center when not being used for other tests
    - TCP iperf tests max at 6 Gbps.

ESnet Measurement Summary

- Standards / Collaborations
  - PerfSONAR
- LHCOPN
  - Circuit Status Monitoring
- Monitoring Hardware in ESnet 4 Hubs
  - Bandwidth
  - Latency
- Measurements
  - SNMP Interface Counters
  - Circuit Availability
  - Flow Data
  - One Way Delay
  - Achievable Bandwidth
- Visualizations
  - PerfSONARUI
  - VisualPerfSONAR
  - NetInfo

References

1. High Performance Network Planning Workshop, August 2002
   – http://www.doecollaboratory.org/meetings/hpnpw
2. Science Case Studies Update, 2006 (contact eli@es.net)
3. DOE Science Networking Roadmap Meeting, June 2003
   – http://www.es.net/hypertext/welcome/pr/Roadmap/index.html
4. DOE Workshop on Ultra High-Speed Transport Protocols and Network Provisioning for Large-Scale Science Applications, April 2003
   – http://www.csm.ornl.gov/ghpn/wk2003
5. Science Case for Large Scale Simulation, June 2003
   – http://www.pnl.gov/scales/
6. Workshop on the Road Map for the Revitalization of High End Computing, June 2003
   – http://www.cra.org/Activities/workshops/nitrd
   – http://www.sc.doe.gov/ascr/20040510_hecrtf.pdf (public report)
7. ASCR Strategic Planning Workshop, July 2003
   – http://www.fp-mcs.anl.gov/ascr-july03spw
8. Planning Workshops-Office of Science Data-Management Strategy, March & May 2004
   – http://www-conf.slac.stanford.edu/dmw2004

17. For more information contact Chin Guok (chin@es.net). Also see
   - http://www.es.net/oscars

ICFA SCIC "Networking for High Energy Physics." International Committee for Future Accelerators (ICFA), Standing Committee on Inter-Regional Connectivity (SCIC), Professor Harvey Newman, Caltech, Chairperson.
   - http://monalisa.caltech.edu:8080/Slides/ICFASCIC2007/

Additional Information

Parallel Data Movers now Predominate

Look at the hosts involved in 2006-01-31—— the plateaus in the host-host top 100 flows are all parallel transfers (thx. to Eli Dart for this observation)

A132023.N1.Vanderbilt.Edu	lstore1.fnal.gov	5.847
A132021.N1.Vanderbilt.Edu	lstore1.fnal.gov	5.884
A132018.N1.Vanderbilt.Edu	lstore1.fnal.gov	6.048
A132022.N1.Vanderbilt.Edu	lstore1.fnal.gov	6.39
A132021.N1.Vanderbilt.Edu	lstore2.fnal.gov	6.771
A132023.N1.Vanderbilt.Edu	lstore2.fnal.gov	6.825
A132022.N1.Vanderbilt.Edu	lstore2.fnal.gov	6.86
A132018.N1.Vanderbilt.Edu	lstore2.fnal.gov	7.286
A132017.N1.Vanderbilt.Edu	lstore1.fnal.gov	7.62
A132017.N1.Vanderbilt.Edu	lstore2.fnal.gov	9.299
A132023.N1.Vanderbilt.Edu	lstore4.fnal.gov	10.522
A132021.N1.Vanderbilt.Edu	lstore4.fnal.gov	10.54
A132018.N1.Vanderbilt.Edu	lstore4.fnal.gov	10.597
A132018.N1.Vanderbilt.Edu	lstore3.fnal.gov	10.746
A132022.N1.Vanderbilt.Edu	lstore4.fnal.gov	11.097
A132022.N1.Vanderbilt.Edu	lstore3.fnal.gov	11.097
A132021.N1.Vanderbilt.Edu	lstore3.fnal.gov	11.213
A132023.N1.Vanderbilt.Edu	lstore3.fnal.gov	11.331
A132017.N1.Vanderbilt.Edu	lstore4.fnal.gov	11.425
A132017.N1.Vanderbilt.Edu	lstore3.fnal.gov	11.489
babar.fzk.de	bbr-xfer03.slac.stanford.edu	2.772
babar.fzk.de	bbr-xfer02.slac.stanford.edu	2.901
babar2.fzk.de	bbr-xfer06.slac.stanford.edu	3.018
babar.fzk.de	bbr-xfer04.slac.stanford.edu	3.222
bbr-export01.pd.infn.it	bbr-xfer03.slac.stanford.edu	11.289
bbr-export02.pd.infn.it	bbr-xfer03.slac.stanford.edu	19.973

bbr-xfer07.slac.stanford.edu	babar2.fzk.de	2.113
bbr-xfer05.slac.stanford.edu	babar.fzk.de	2.254
bbr-xfer04.slac.stanford.edu	babar.fzk.de	2.294
bbr-xfer07.slac.stanford.edu	babar.fzk.de	2.337
bbr-xfer04.slac.stanford.edu	babar2.fzk.de	2.339
bbr-xfer05.slac.stanford.edu	babar2.fzk.de	2.357
bbr-xfer08.slac.stanford.edu	babar2.fzk.de	2.471
bbr-xfer08.slac.stanford.edu	babar.fzk.de	2.627
bbr-xfer04.slac.stanford.edu	babar3.fzk.de	3.234
bbr-xfer05.slac.stanford.edu	babar3.fzk.de	3.271
bbr-xfer08.slac.stanford.edu	babar3.fzk.de	3.276
bbr-xfer07.slac.stanford.edu	babar3.fzk.de	3.298
bbr-xfer05.slac.stanford.edu	bbr-datamove10.cr.cnaf.infn.it	2.366
bbr-xfer07.slac.stanford.edu	bbr-datamove10.cr.cnaf.infn.it	2.519
bbr-xfer04.slac.stanford.edu	bbr-datamove10.cr.cnaf.infn.it	2.548
bbr-xfer08.slac.stanford.edu	bbr-datamove10.cr.cnaf.infn.it	2.656
bbr-xfer08.slac.stanford.edu	bbr-datamove09.cr.cnaf.infn.it	3.927
bbr-xfer05.slac.stanford.edu	bbr-datamove09.cr.cnaf.infn.it	3.94
bbr-xfer04.slac.stanford.edu	bbr-datamove09.cr.cnaf.infn.it	4.011
bbr-xfer07.slac.stanford.edu	bbr-datamove09.cr.cnaf.infn.it	4.177
bbr-xfer04.slac.stanford.edu	csfmove01.rl.ac.uk	5.952
bbr-xfer04.slac.stanford.edu	move03.gridpp.rl.ac.uk	5.959
bbr-xfer05.slac.stanford.edu	csfmove01.rl.ac.uk	5.976
bbr-xfer05.slac.stanford.edu	move03.gridpp.rl.ac.uk	6.12
bbr-xfer07.slac.stanford.edu	csfmove01.rl.ac.uk	6.242
bbr-xfer08.slac.stanford.edu	move03.gridpp.rl.ac.uk	6.357
bbr-xfer08.slac.stanford.edu	csfmove01.rl.ac.uk	6.48
bbr-xfer07.slac.stanford.edu	move03.gridpp.rl.ac.uk	6.604