

ESnet4: Networking for the Future of DOE Science

November 7, 2006

William E. Johnston

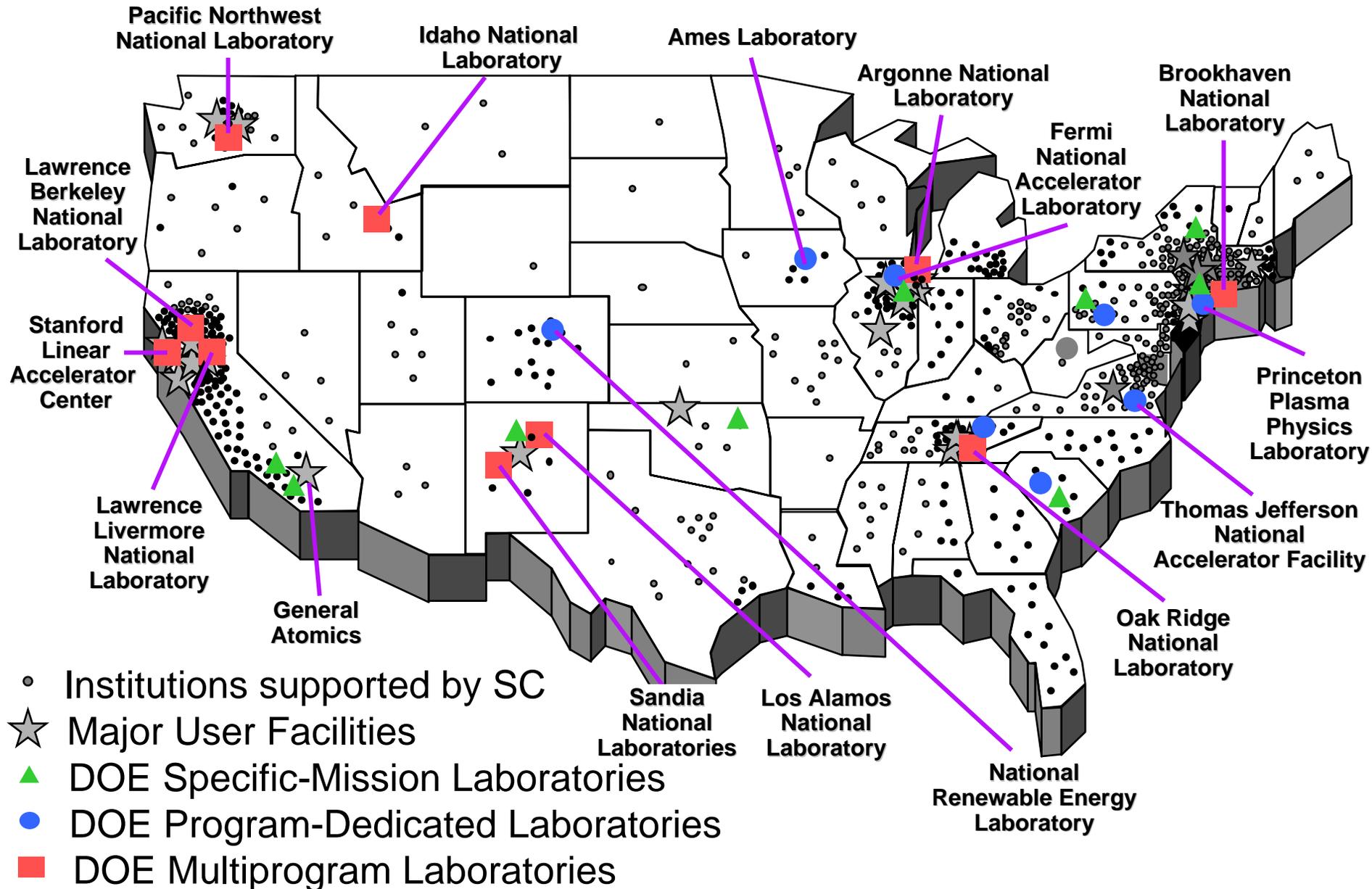
ESnet Department Head and Senior Scientist
Lawrence Berkeley National Laboratory

➤ DOE Office of Science and ESnet – the ESnet Mission

- “The Office of Science 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>)
- **ESnet’s primary mission is to enable the large-scale science that is the mission of the Office of Science (SC):**
 - Sharing of massive amounts of data
 - Supporting thousands of collaborators world-wide
 - Distributed data processing
 - Distributed data management
 - Distributed simulation, visualization, and computational steering
- ESnet also provides network and collaboration services to DOE laboratories and other DOE programs in cases where this is cost effective.

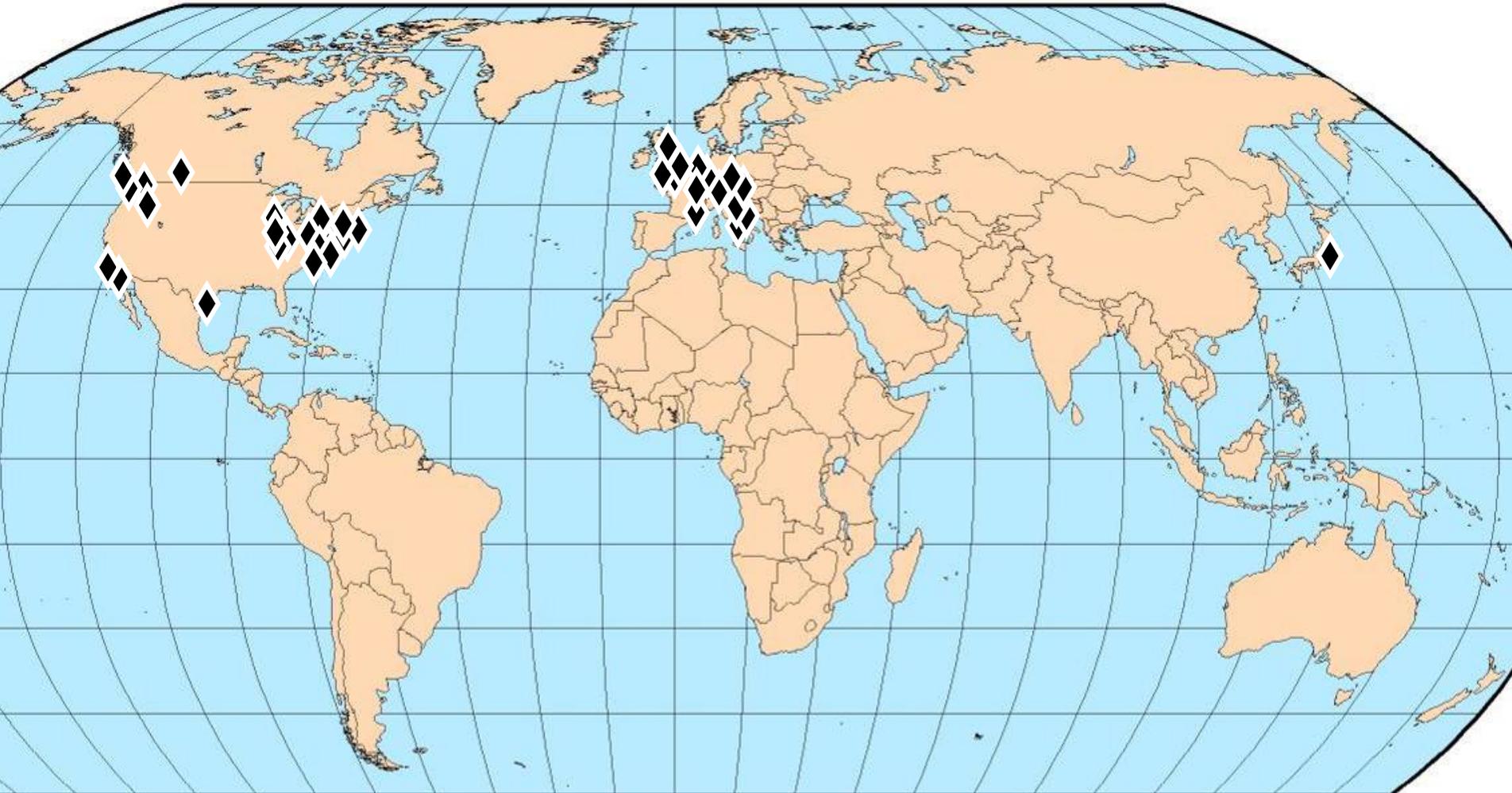
Office of Science US Community

Drives ESnet Design for Domestic Connectivity



Footprint of Largest SC Data Sharing Collaborators

Drives the International Footprint that ESnet Must Support



- 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)

What Does ESnet Provide? - 1

- An architecture tailored to accommodate DOE's large-scale science
 - Move huge amounts of data between a small number of sites that are scattered all over the world
- Comprehensive connectivity
 - High bandwidth access to DOE sites and DOE's primary science collaborators: Research and Education institutions in the US, Europe, Asia Pacific, and elsewhere
- Full access to the global Internet for DOE Labs
 - ESnet is a tier 1 ISP managing a full complement of Internet routes for global access
- Highly reliable transit networking
 - Fundamental goal is to deliver every packet that is received to the "target" site

What Does ESnet Provide? - 2

- A full suite of network services
 - IPv4 and IPv6 routing and address space management
 - IPv4 multicast (and soon IPv6 multicast)
 - Primary DNS services
 - Circuit services (layer 2 e.g. Ethernet VLANs), MPLS overlay networks (e.g. SecureNet when it was ATM based)
 - Scavenger service so that certain types of bulk traffic can use all available bandwidth, but will give priority to any other traffic when it shows up
 - Prototype guaranteed bandwidth and virtual circuit services

What Does ESnet Provide? - 3

- New network services
 - Guaranteed bandwidth services
 - Via a combination of QoS, MPLS overlay, and layer 2 VLANs
- Collaboration services and Grid middleware supporting collaborative science
 - Federated trust services / PKI Certification Authorities with science oriented policy
 - Audio-video-data teleconferencing
- Highly reliable and secure operation
 - Extensive disaster recovery infrastructure
 - Comprehensive internal security
 - Cyberdefense for the WAN

What Does ESnet Provide? - 4

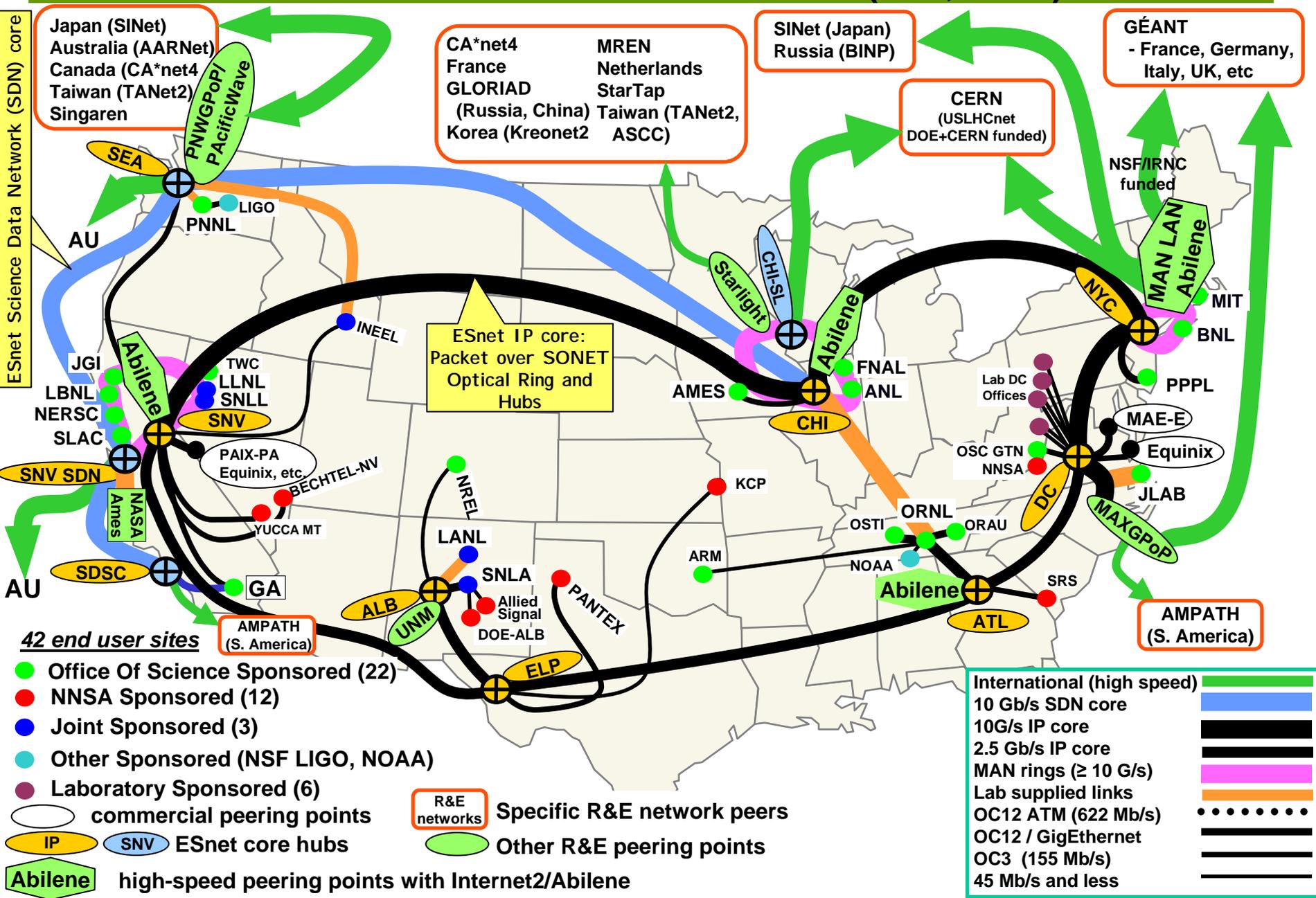
- Comprehensive user support, including “owning” all trouble tickets involving ESnet users (including problems at the far end of an ESnet connection) until they are resolved – 24x7x365 coverage
 - ESnet’s mission is to enable the network based aspects of OSC science, and that includes troubleshooting network problems wherever they occur
- A highly collaborative and interactive relationship with the DOE Labs and scientists for planning, configuration, and operation of the network
 - ESnet and its services evolve continuously in direct response to OSC science needs
 - Engineering services for special requirements

ESnet History

ESnet0/MFENet mid-1970s-1986	ESnet0/MFENet	56 Kbps microwave and satellite links
ESnet1 1986-1995	ESnet formed to serve the Office of Science	56 Kbps, X.25 to 45 Mbps T3
ESnet2 1995-2000	Partnered with Sprint to build the first national footprint ATM network	IP over 155 Mbps ATM net
ESnet3 2000-2007	Partnered with Qwest to build a national Packet over SONET network and optical channel Metropolitan Area Networks	IP over 10Gbps SONET
ESnet4 2007-2012	Partner with Internet2 and US Research & Education community to build a dedicated national optical network	IP and virtual circuits on a configurable optical infrastructure with at least 5-6 optical channels of 10-100 Gbps each

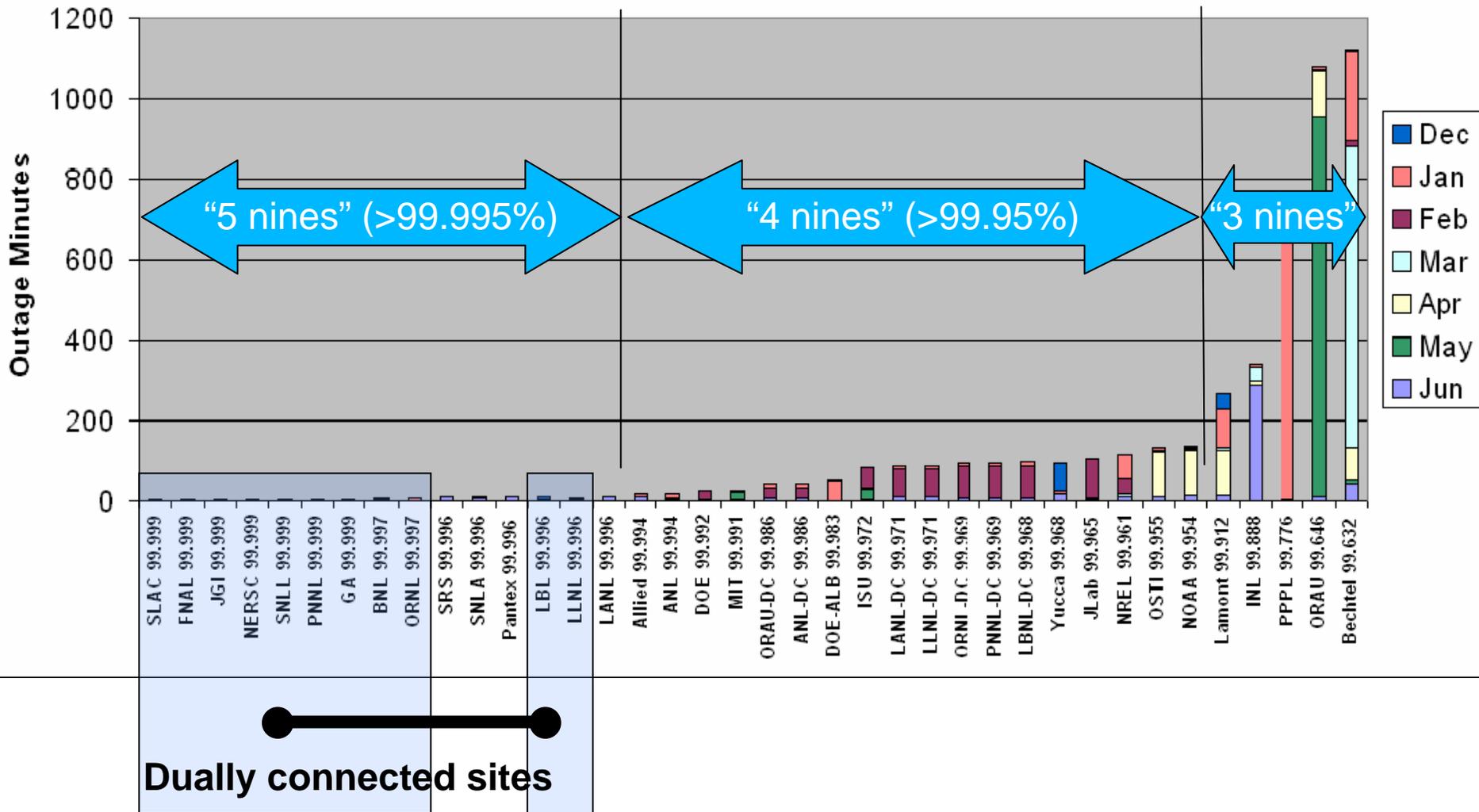
transition in progress

ESnet3 Today Provides Global High-Speed Internet Connectivity for DOE Facilities and Collaborators (Fall, 2006)



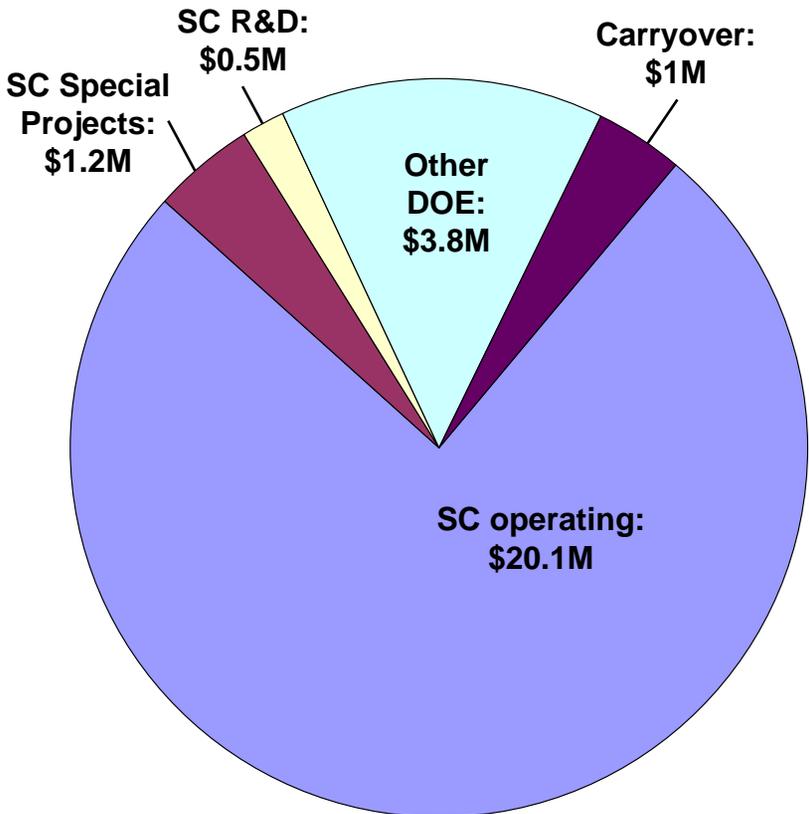
ESnet is a Highly Reliable Infrastructure

ESnet Availability 12/2005 through 6/2006

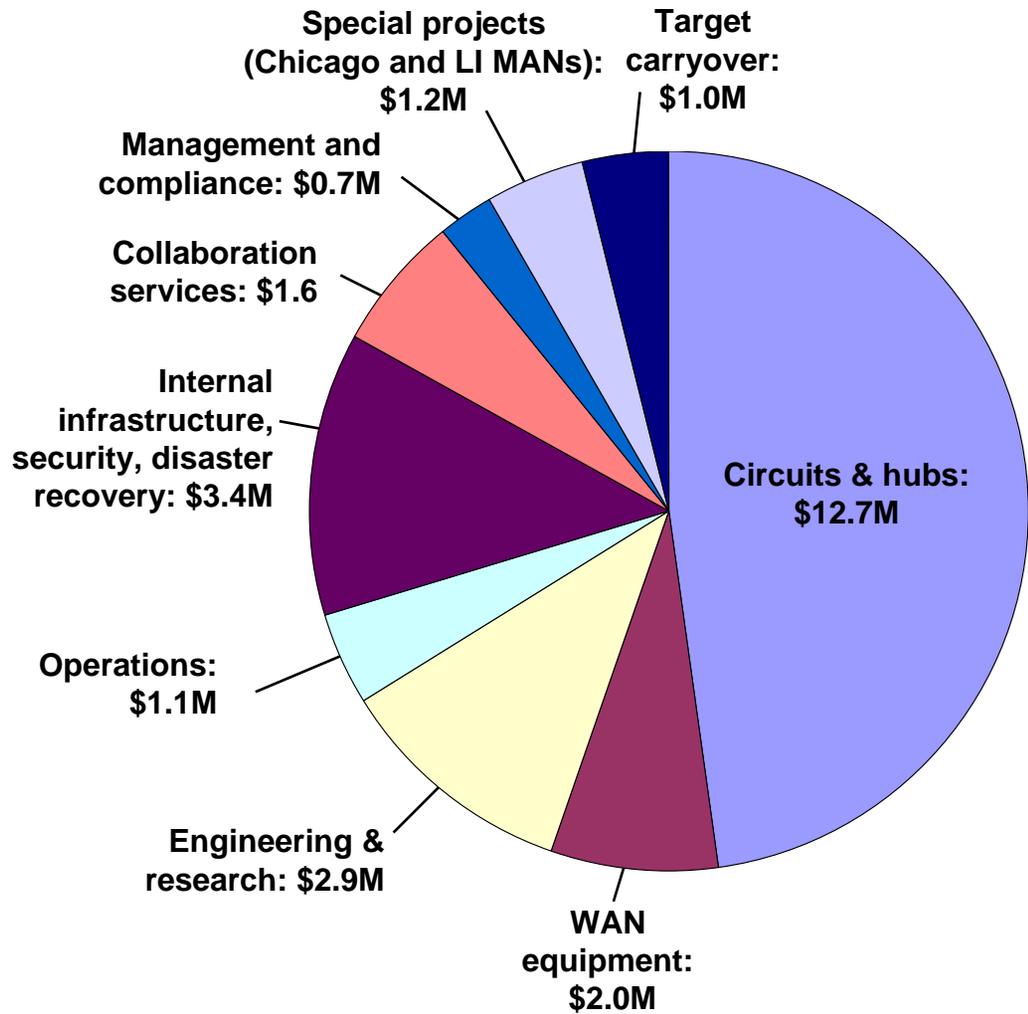


ESnet FY06 Budget is Approximately \$26.6M

Approximate Budget Categories



**Total funds:
\$26.6M**



**Total expenses:
\$26.6M**

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

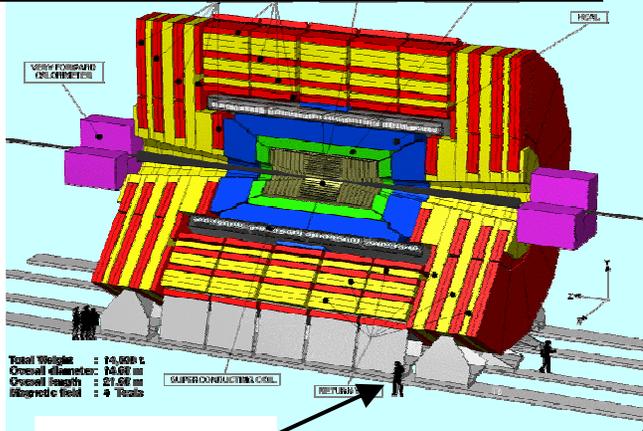
- Large-scale collaborative science – big facilities, massive data, thousands of collaborators – is now a significant aspect of the Office of Science (“SC”) program
- SC science community is almost equally split between Labs and universities
 - SC facilities have users worldwide
- Very large international (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 the Future Network - ESnet4

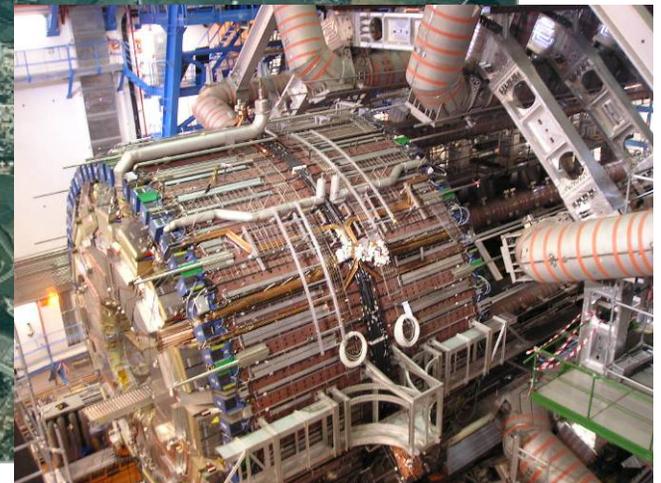
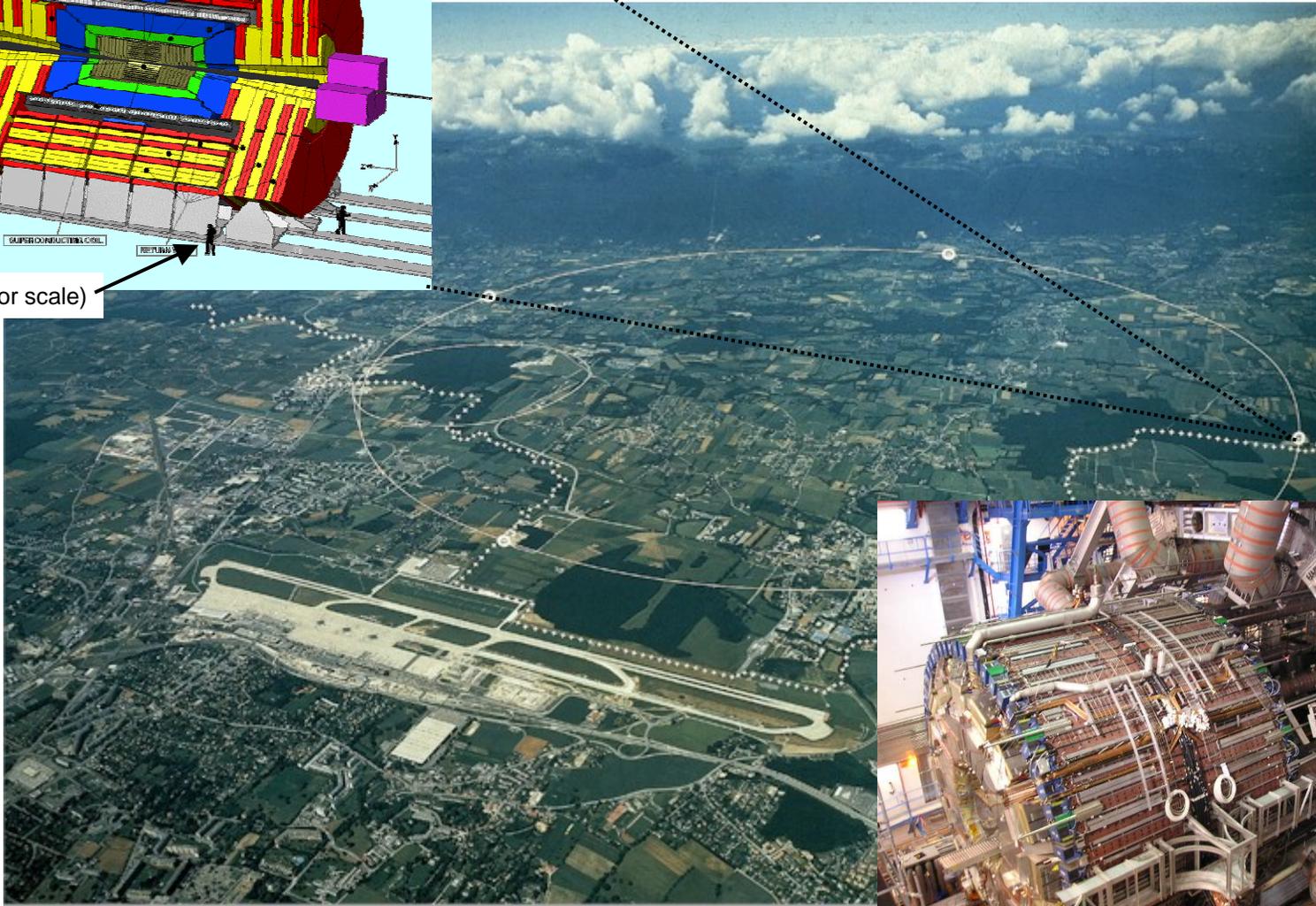
- **Requirements of the ESnet stakeholders are primarily determined by**
 - 1) Data characteristics of instruments and facilities that will be connected to ESnet**
 - What data will be generated by instruments coming on-line over the next 5-10 years?
 - How and where will it be analyzed and used?
 - 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?

The Largest Facility: Large Hadron Collider at CERN

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



human (for scale)



(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 compiled as case studies that provide specific 5 & 10 year network requirements for bandwidth, footprint, and new services

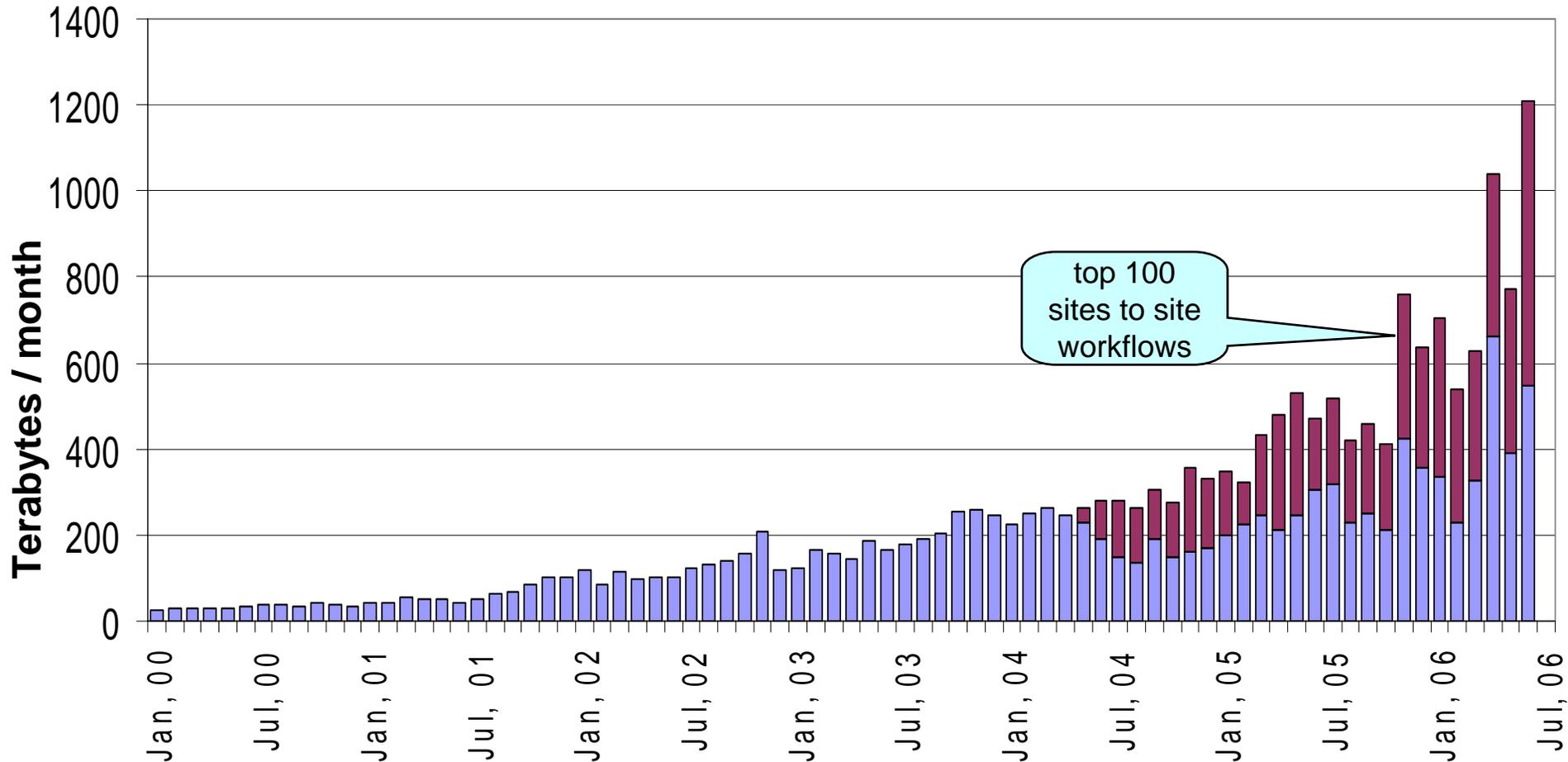
Science Networking Requirements Aggregation Summary

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)	<ul style="list-style-type: none"> • DOE sites • US Universities • Industry 	200+ Mbps	1 Gbps	<ul style="list-style-type: none"> • Bulk data • Remote control 	<ul style="list-style-type: none"> • Guaranteed bandwidth • Guaranteed QoS • Deadline scheduling
NERSC and ACLF	-	<ul style="list-style-type: none"> • DOE sites • US Universities • International • Other ASCR supercomputers 	10 Gbps	20 to 40 Gbps	<ul style="list-style-type: none"> • Bulk data • Remote control • Remote file system sharing 	<ul style="list-style-type: none"> • Guaranteed bandwidth • Guaranteed QoS • Deadline Scheduling • PKI / Grid
NLCF	-	<ul style="list-style-type: none"> • DOE sites • US Universities • Industry • International 	Backbone Band width parity	Backbone band width parity	<ul style="list-style-type: none"> • Bulk data • Remote file system sharing 	
Nuclear Physics (RHIC)	-	<ul style="list-style-type: none"> • DOE sites • US Universities • International 	12 Gbps	70 Gbps	<ul style="list-style-type: none"> • Bulk data 	<ul style="list-style-type: none"> • Guaranteed bandwidth • PKI / Grid
Spallation Neutron Source	High (24x7 operation)	<ul style="list-style-type: none"> • DOE sites 	640 Mbps	2 Gbps	<ul style="list-style-type: none"> • 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	-	<ul style="list-style-type: none"> • DOE sites • US Universities • Industry 	1 TB/day 300 Mbps	5 TB/day 1.5 Gbps	<ul style="list-style-type: none"> • Bulk data • Remote control 	<ul style="list-style-type: none"> • Guaranteed bandwidth • PKI / Grid
Bioinformatics	-	<ul style="list-style-type: none"> • DOE sites • US Universities 	625 Mbps 12.5 Gbps in two years	250 Gbps	<ul style="list-style-type: none"> • Bulk data • Remote control • Point-to-multipoint 	<ul style="list-style-type: none"> • Guaranteed bandwidth • High-speed multicast
Chemistry / Combustion	-	<ul style="list-style-type: none"> • DOE sites • US Universities • Industry 	-	10s of Gigabits per second	<ul style="list-style-type: none"> • Bulk data 	<ul style="list-style-type: none"> • Guaranteed bandwidth • PKI / Grid
Climate Science	-	<ul style="list-style-type: none"> • DOE sites • US Universities • International 	-	5 PB per year 5 Gbps	<ul style="list-style-type: none"> • Bulk data • Remote control 	<ul style="list-style-type: none"> • Guaranteed bandwidth • PKI / Grid
Immediate Requirements and Drivers						
High Energy Physics (LHC)	99.95+% (Less than 4 hrs/year)	<ul style="list-style-type: none"> • US Tier1 (FNAL, BNL) • US Tier2 (Universities) • International (Europe, Canada) 	10 Gbps	60 to 80 Gbps (30-40 Gbps per US Tier1)	<ul style="list-style-type: none"> • Bulk data • Coupled data analysis processes 	<ul style="list-style-type: none"> • Guaranteed bandwidth • Traffic isolation • PKI / Grid

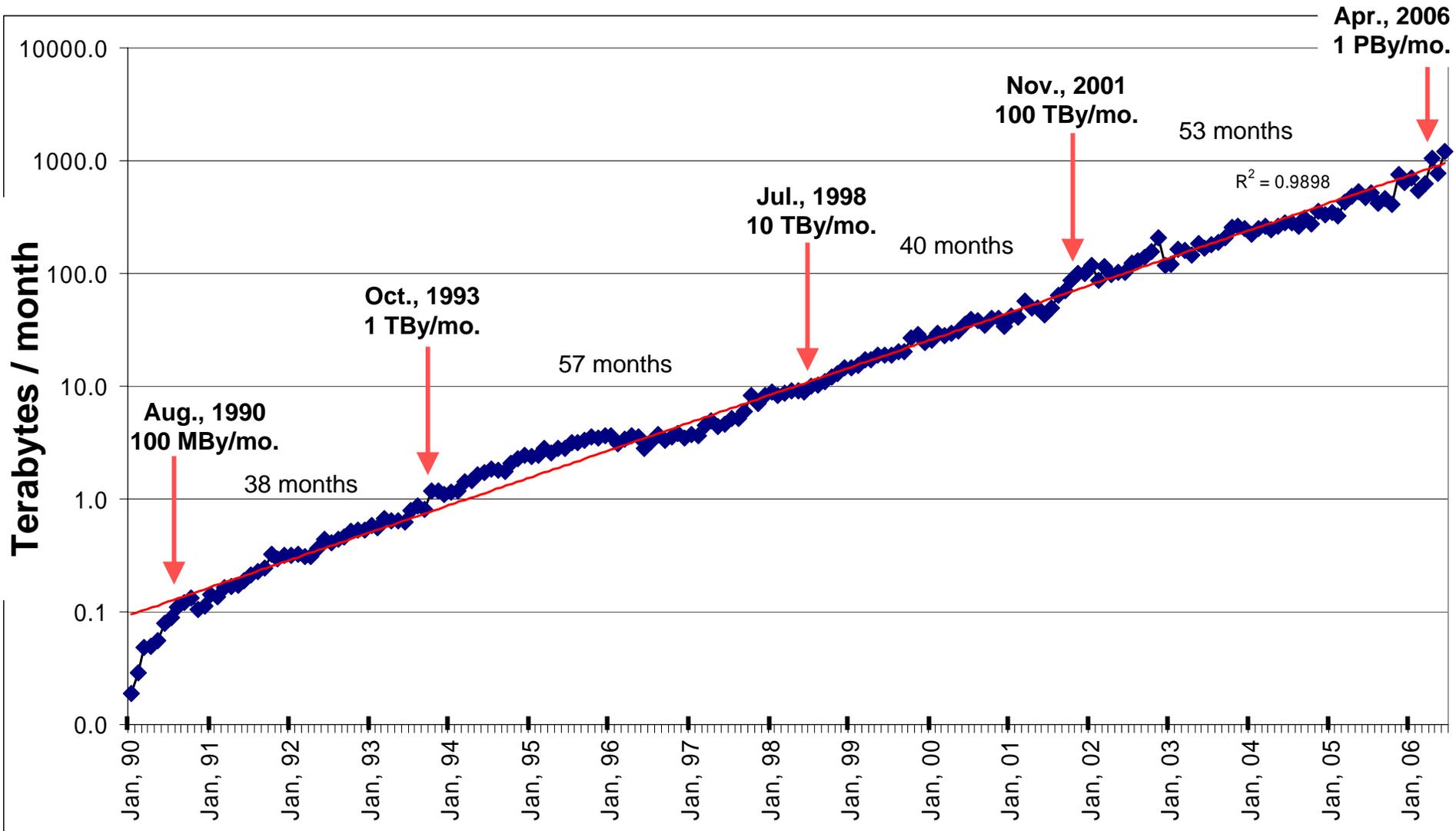
3) These Trends are Seen in Observed Evolution of Historical ESnet Traffic Patterns



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

- ESnet is currently transporting more than 1 petabyte (1000 terabytes) per month
- More than 50% of the traffic is now generated by the top 100 sites

ESnet Traffic has Increased by 10X Every 47 Months, on Average, Since 1990



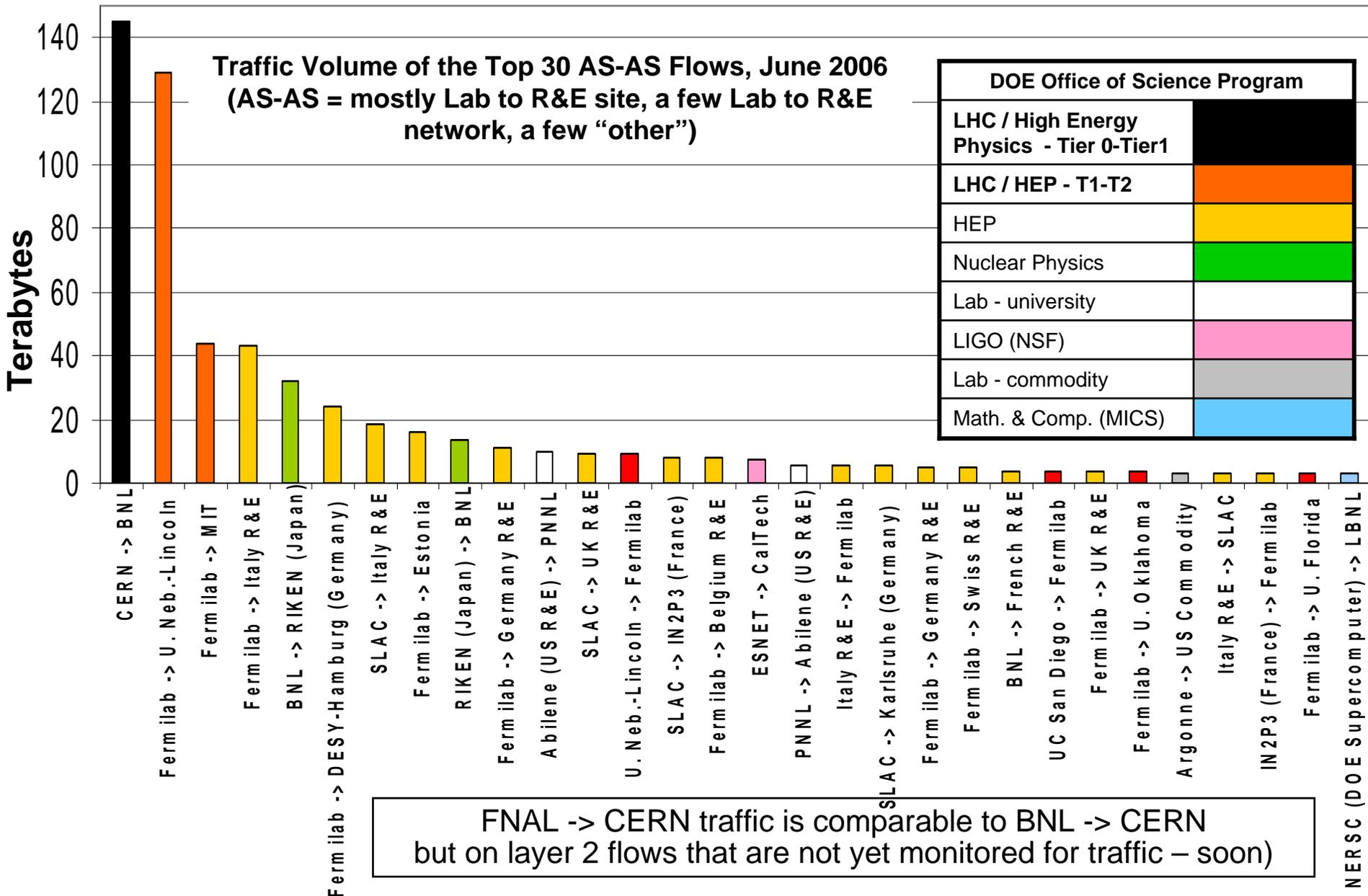
Log Plot of ESnet Monthly Accepted Traffic, January, 1990 – June, 2006

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

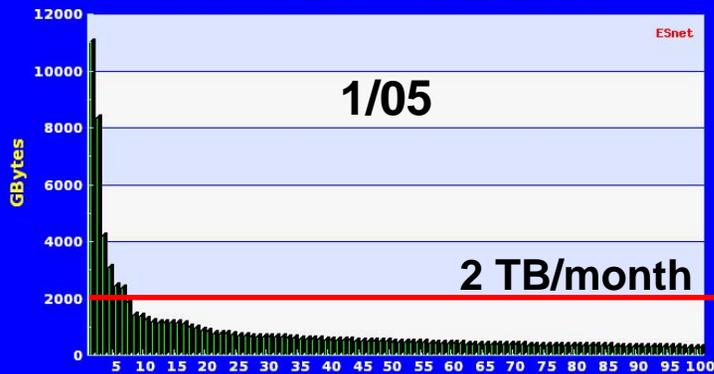
Subtitle: "Onslaught of the LHC")



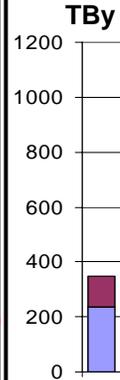
➤ Traffic Patterns are Changing Dramatically

Top 100 HOST-HOST Traffic

Base Date: 2005-01-31 -- Aggregation: 30 day(s) -- Router(s): all

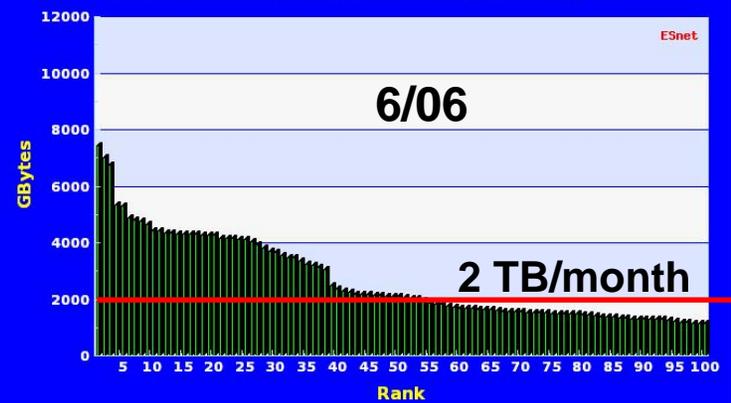


total traffic, TBy

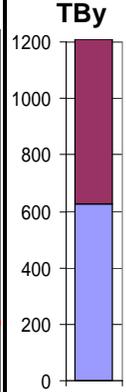


Top 100 HOST-HOST Traffic

Base Date: 2006-07-14 -- Aggregation: 30 day(s) -- Router(s): all

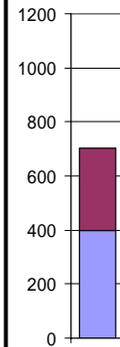
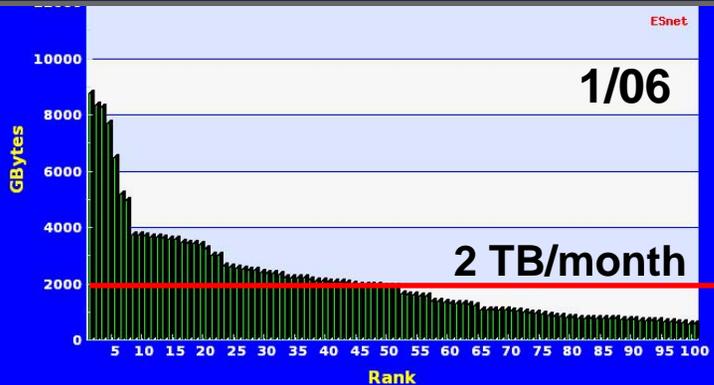
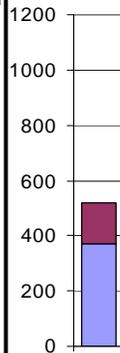
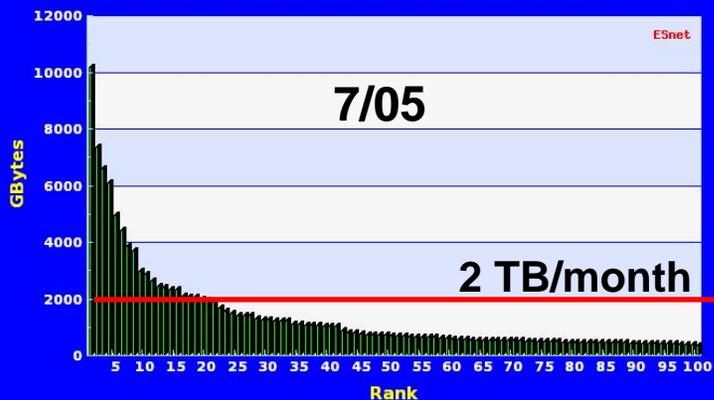


total traffic, TBy



Top 100 HOST-HOST Traffic

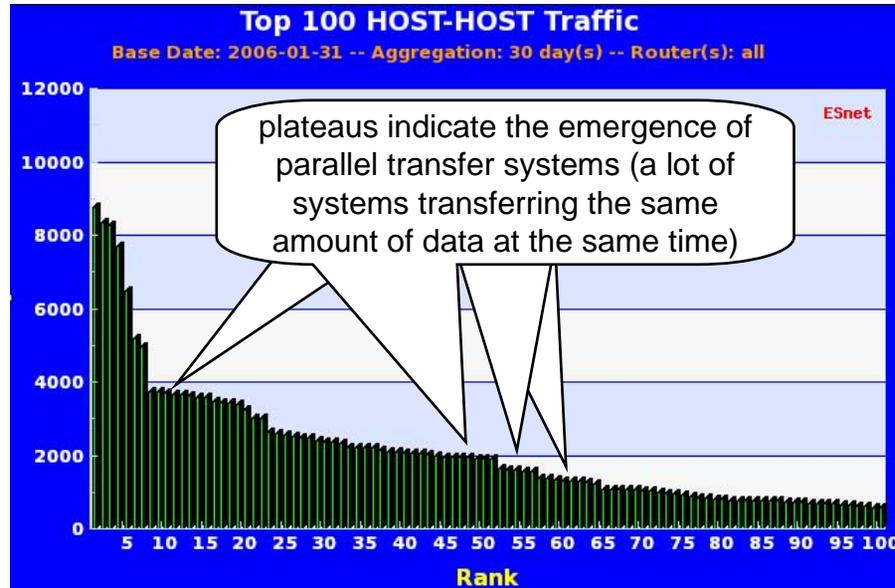
Base Date: 2005-07-31 -- Aggregation: 30 day(s) -- Router(s): all



- 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?

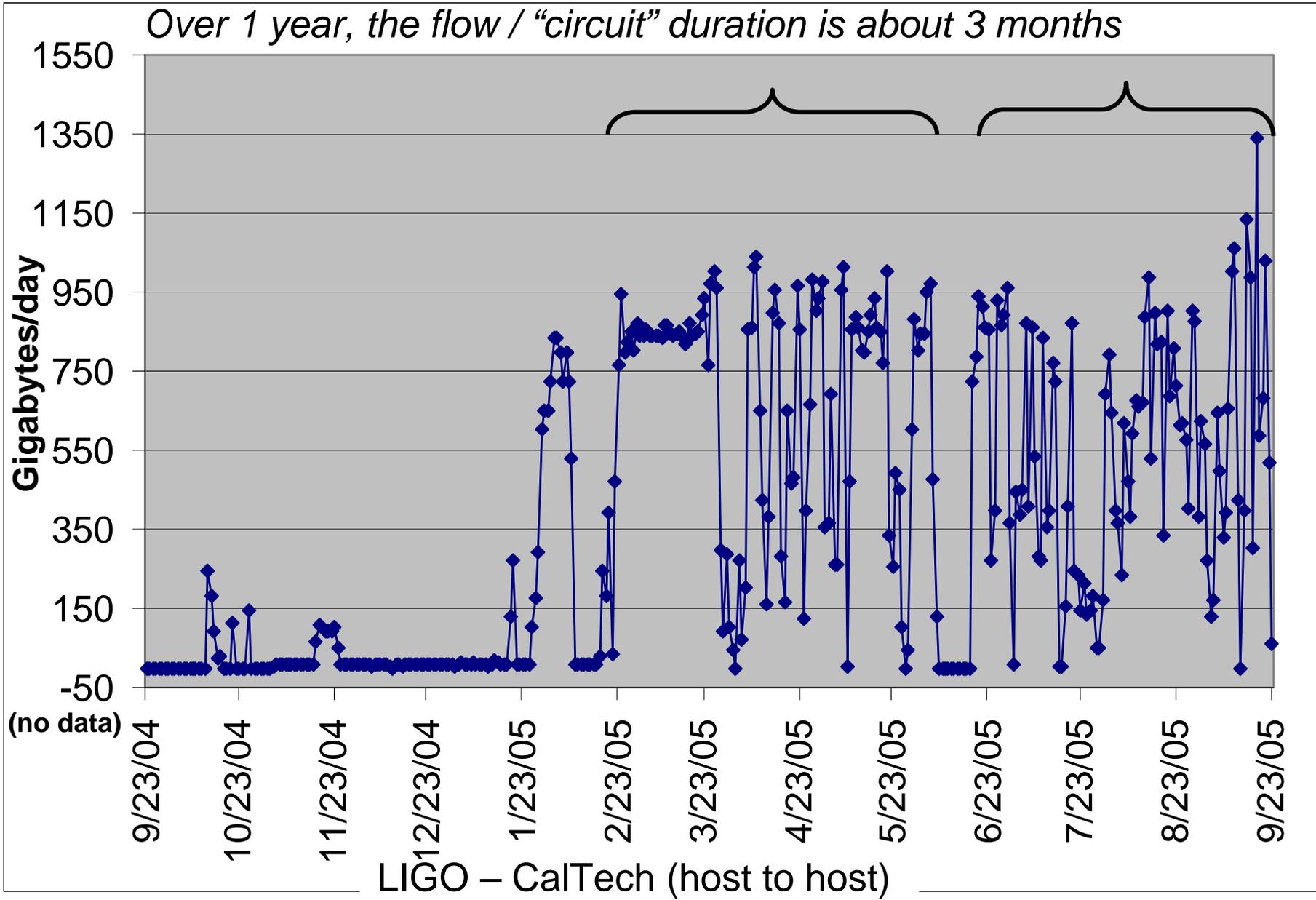


Answer: Most large data transfers are now done by parallel / Grid data movers

- In June, 2006 **72%** 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

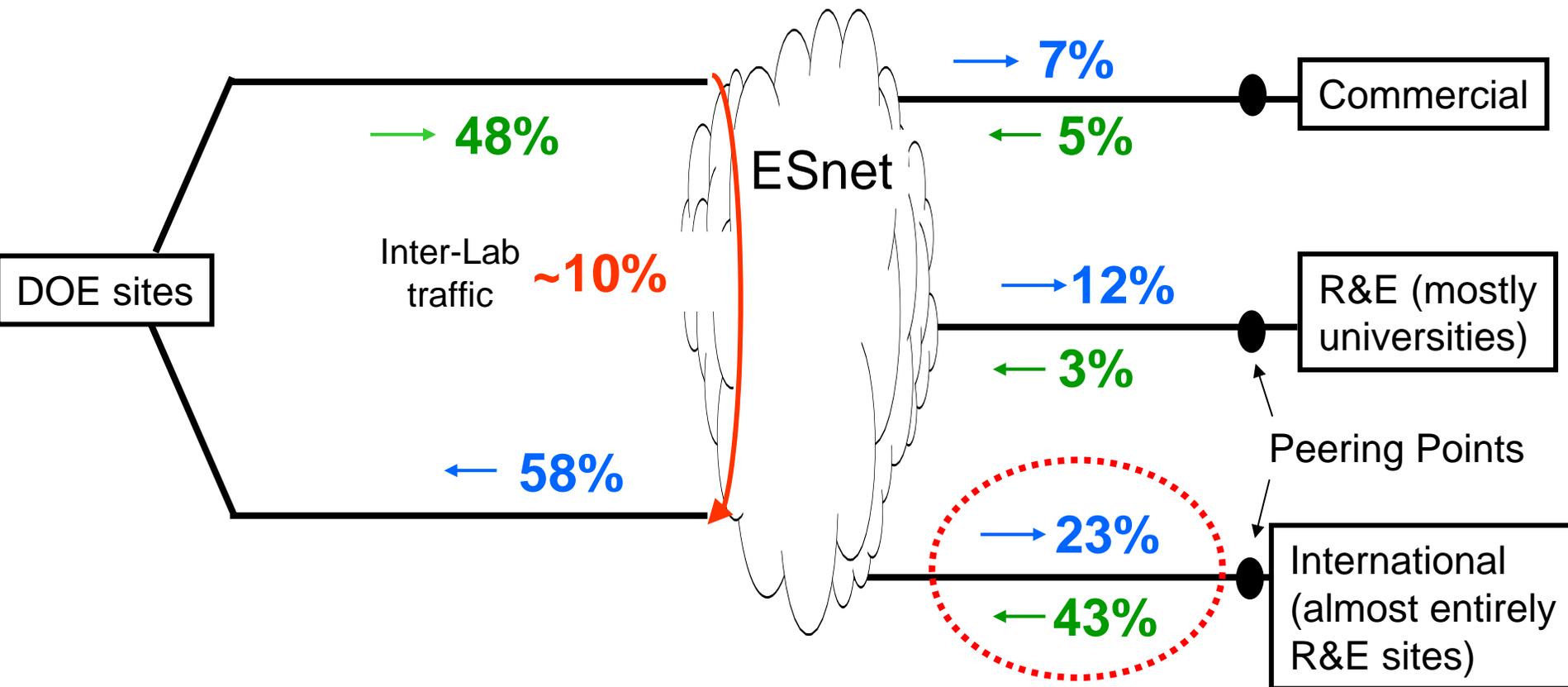
Network Observation – Circuit-like Behavior

Look at Top 20 Traffic Generator's Historical Flow Patterns
Over 1 year, the flow / "circuit" duration is about 3 months



What is the High-Level View of ESnet Traffic Patterns?

ESnet Inter-Sector Traffic Summary, Mar. 2006



Traffic notes

- more than 90% of all traffic Office of Science
- less than 10% is inter-Lab

Traffic coming into ESnet = Green
Traffic leaving ESnet = Blue
Traffic between ESnet sites ↪
% = of total ingress or egress traffic

➤ Requirements from Traffic Flow Observations

- Most of ESnet science traffic has a source or sink outside of ESnet
 - Drives requirement for high-bandwidth peering
 - Reliability and bandwidth requirements demand that peering be redundant
 - Multiple 10 Gbps peerings today, must be able to add more bandwidth flexibly and cost-effectively
 - Bandwidth and service guarantees must traverse R&E peerings
 - Collaboration with other R&E networks on a common framework is critical
 - Seamless fabric
- Large-scale science is now the dominant user of the network
 - Satisfying the demands of large-scale science traffic into the future will require a purpose-built, scalable architecture
 - Traffic patterns are different than commodity Internet

Changing Science Environment ⇒ New Demands on Network

- **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 end-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) 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

3) **Metropolitan Area Network** (MAN) rings to provide

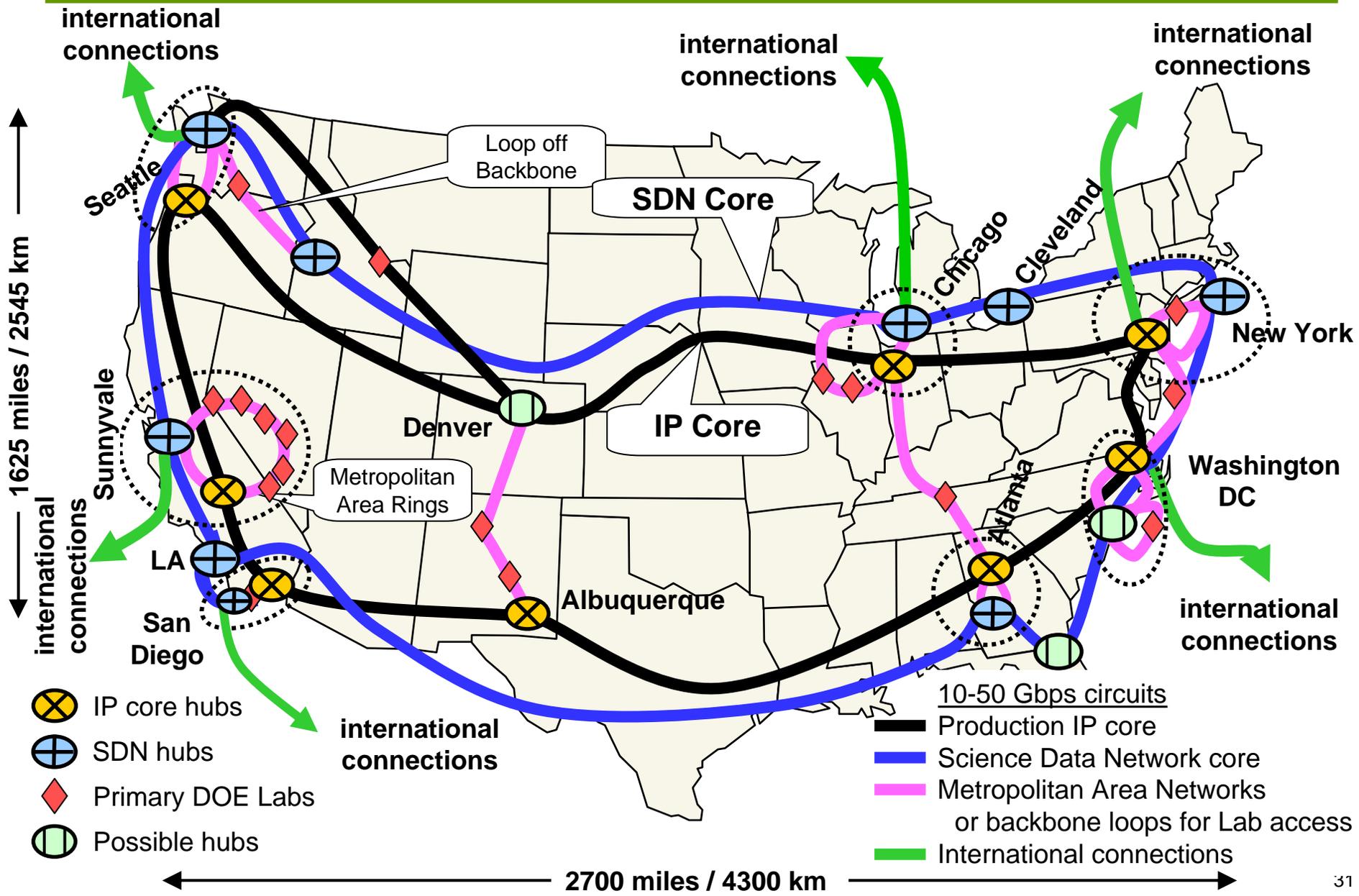
- Dual site connectivity for reliability
- Much higher site-to-core bandwidth
- Support for both production IP and circuit-based traffic
- Multiply connecting the SDN and IP cores

3a) **Loops off of the backbone** rings to provide

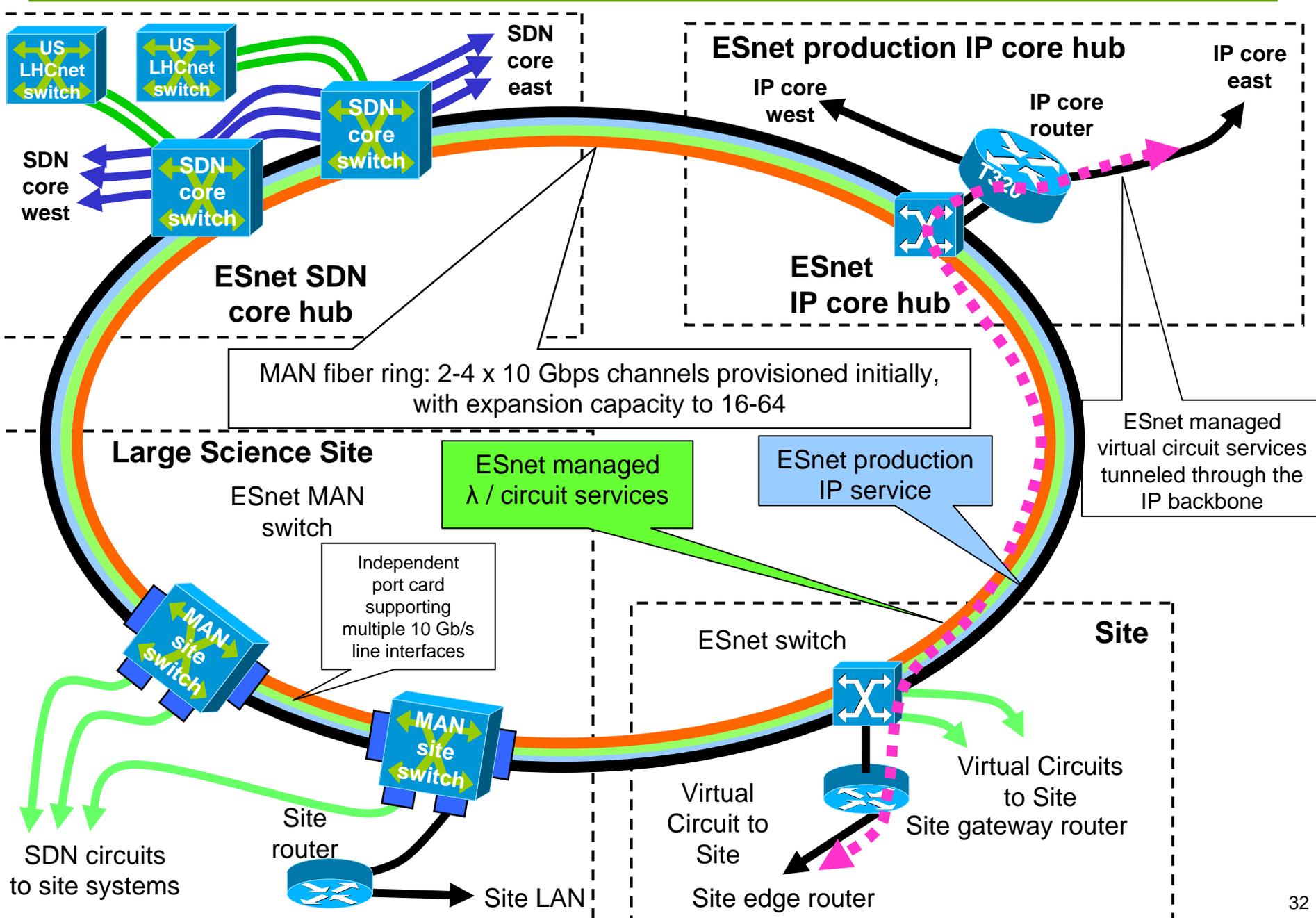
- For dual site connections where MANs are not practical

ESnet Target Architecture:

IP Core+Science Data Network Core+Metro Area Rings

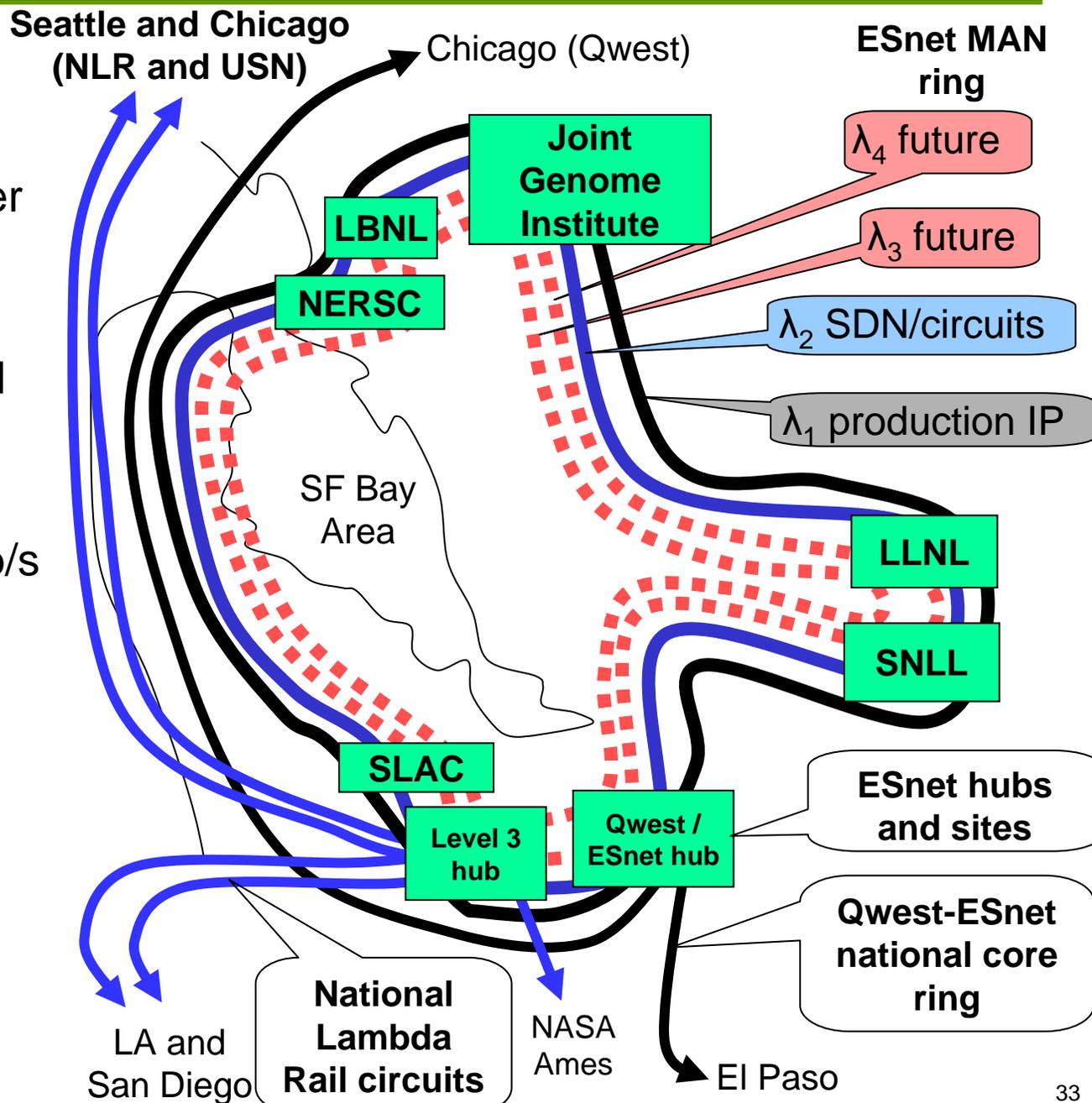


ESnet Metropolitan Area Network Ring Architecture for High Reliability Sites



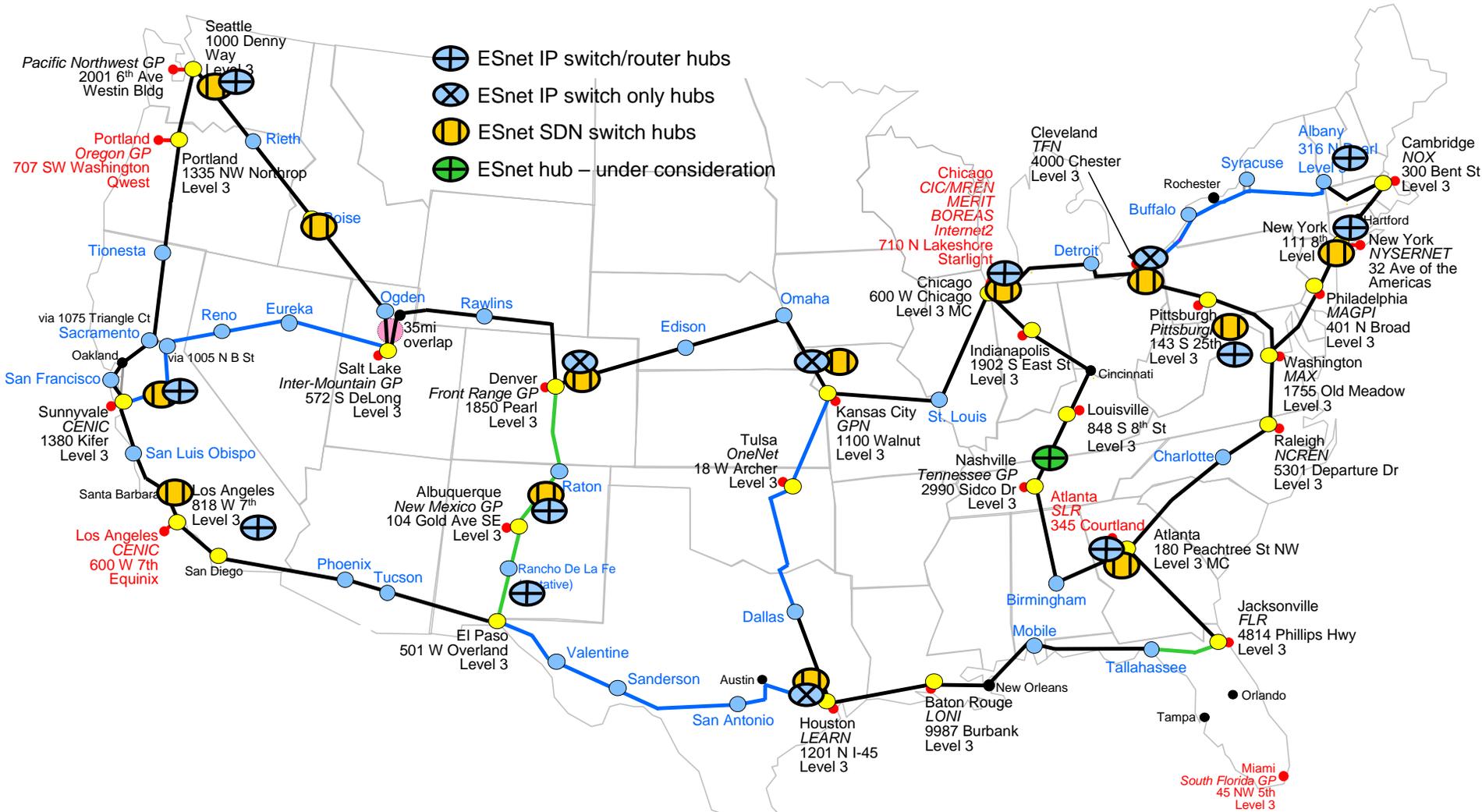
ESnet SF Bay Area MAN

- 2 λ s (2 X 10 Gb/s channels) in a ring configuration, and delivered as 10 GigEther circuits
- Dual site connection (independent “east” and “west” connections) to each site
- Will be used as a 10 Gb/s production IP ring and 2 X 10 Gb/s paths (for circuit services) to each site



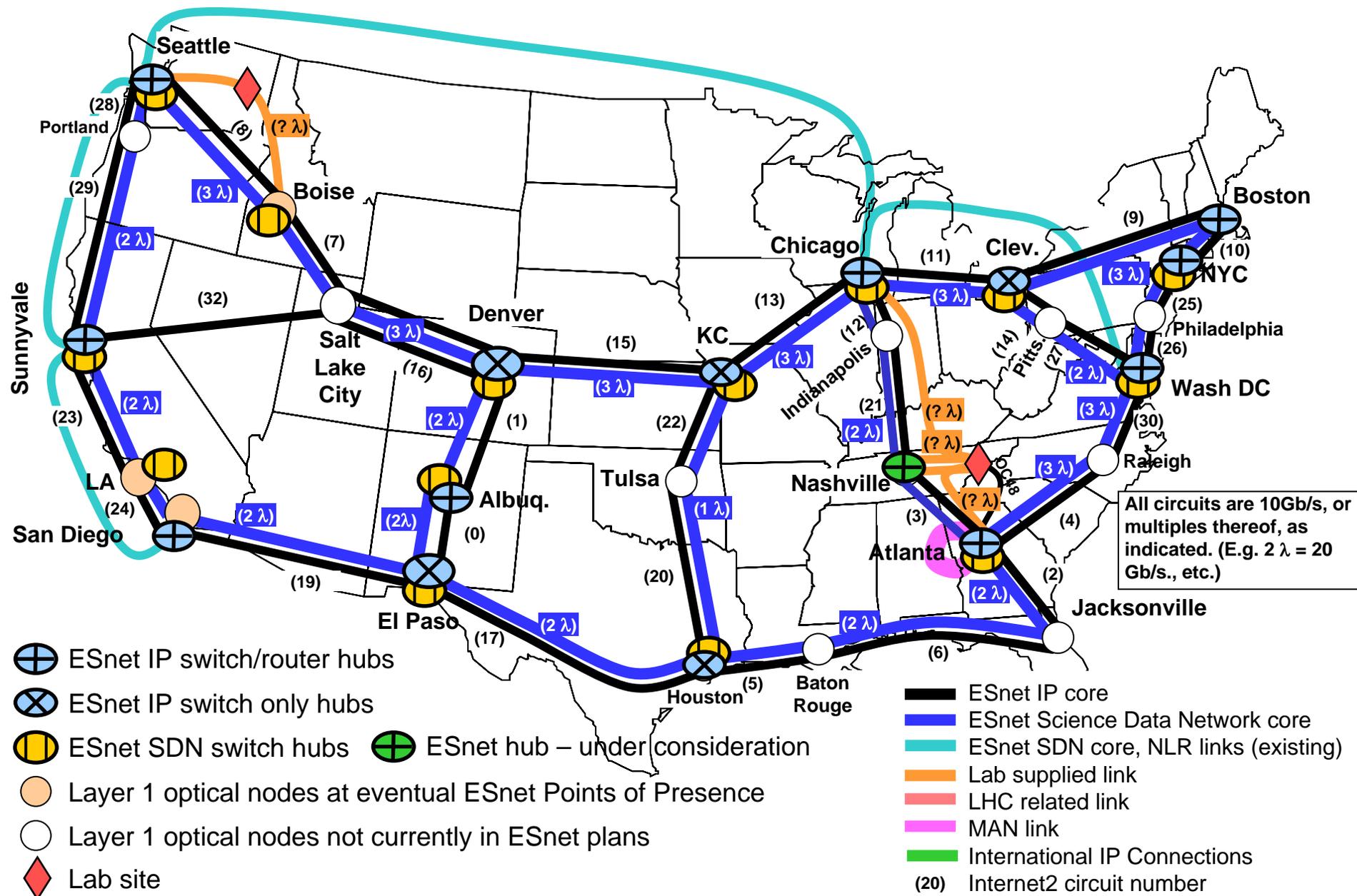
Optical Infrastructure

Internet2, with ESnet as its largest partner, has contracted with Level3 Communications to provide a carrier-maintained optical infrastructure that is based on a dedicated fiber infrastructure that is provisioned with the advanced Infinera, Dense Wave Division Multiplexing equipment.

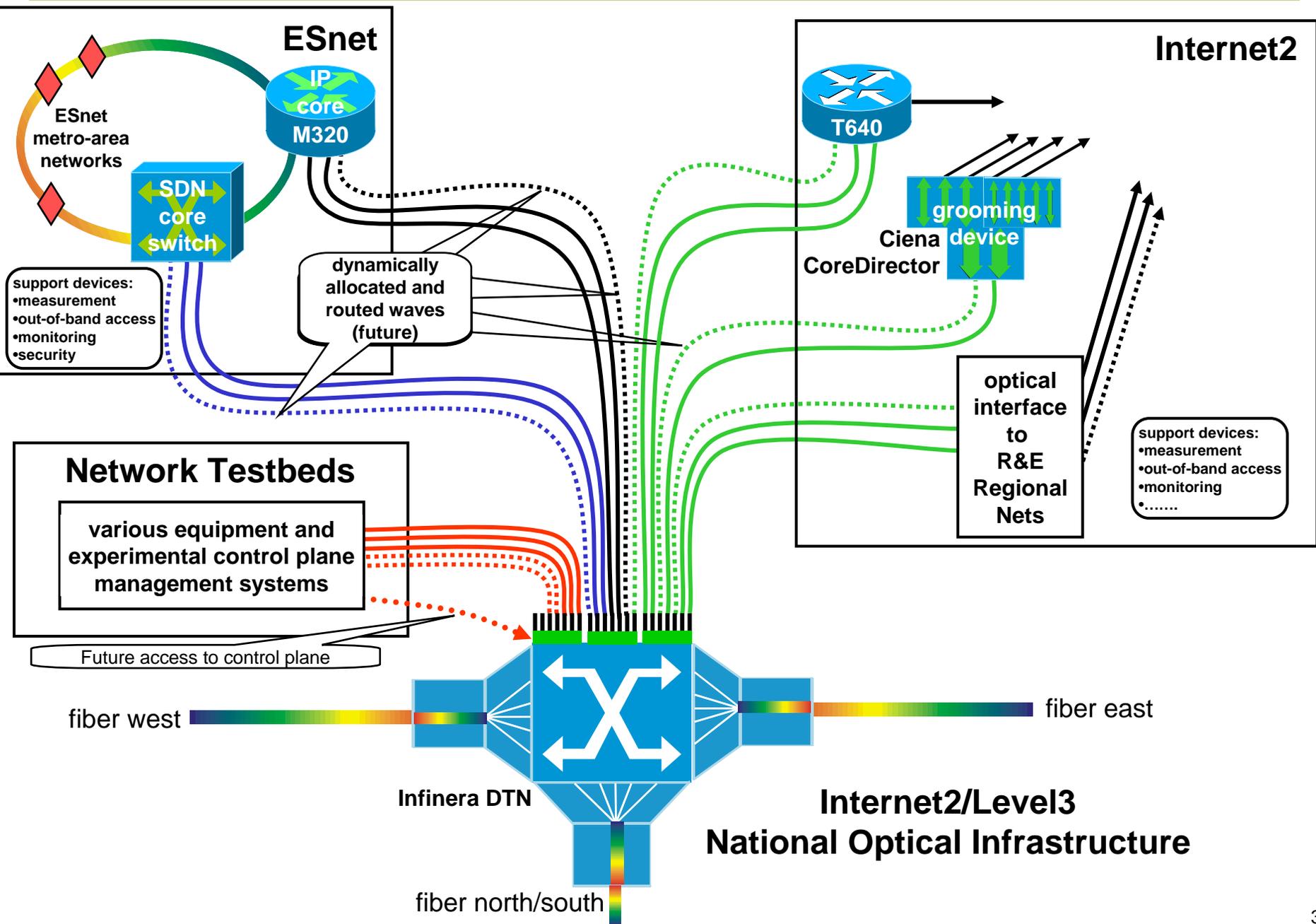


ESnet4 2009 Configuration

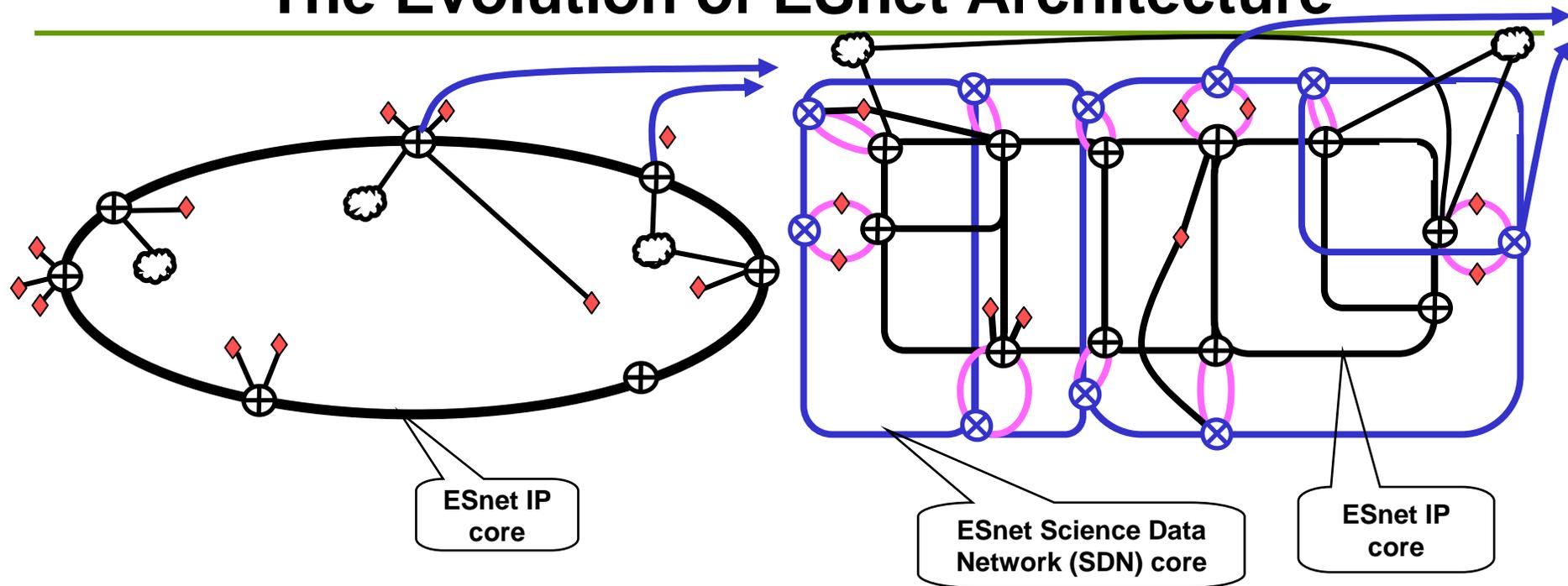
(Some of the circuits may be allocated dynamically from shared a pool.)



Internet2 and ESnet Optical Node



The Evolution of ESnet Architecture



ESnet to 2005:

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

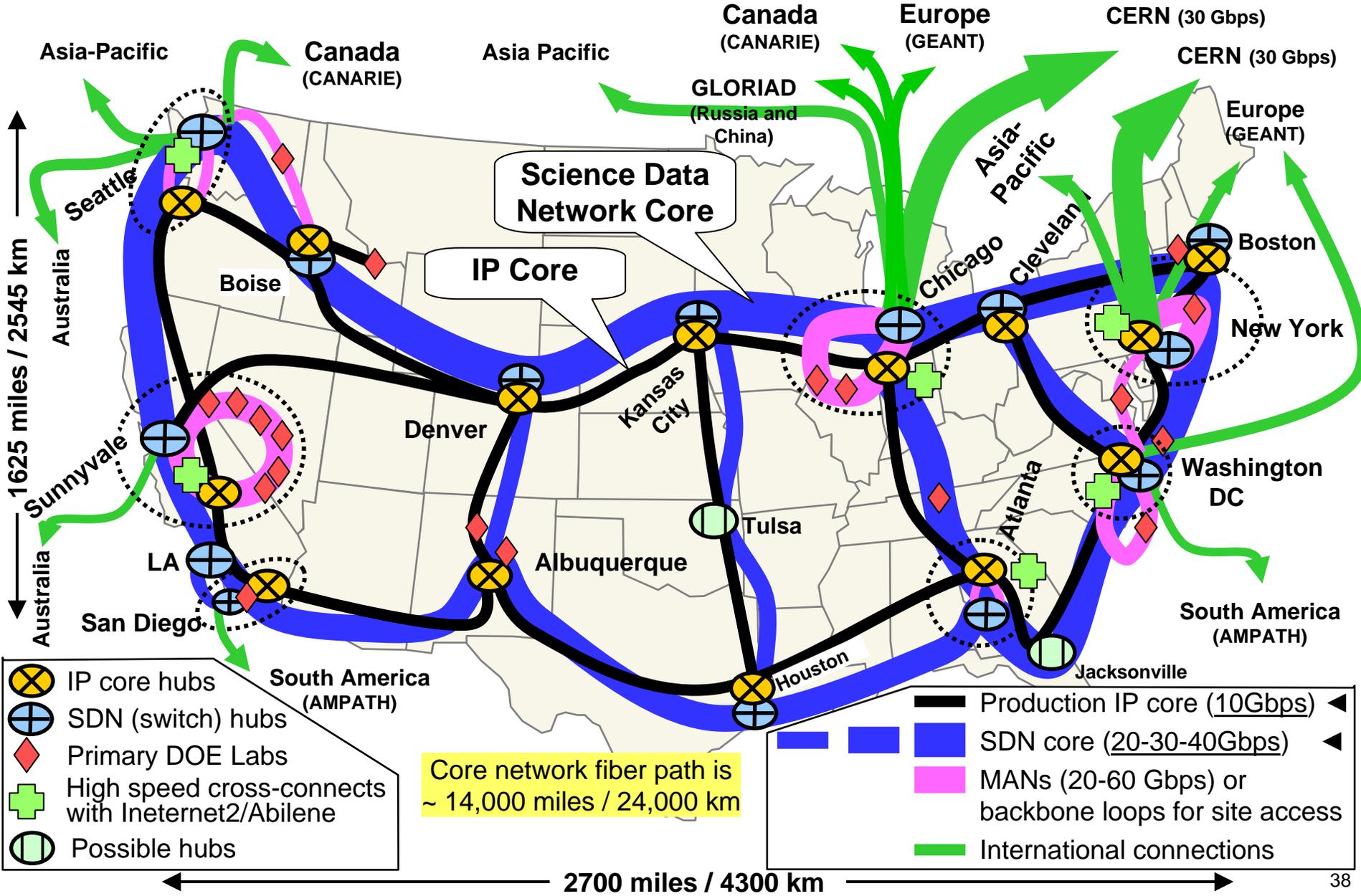
ESnet from 2006-07:

- 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

- ◆ ESnet sites
- ⊕ ESnet hubs / core network connection points
- Metro area rings (MANs)
- ☁ Other IP networks
- ➡ Circuit connections to other science networks (e.g. USLHCNet)

ESnet4 Planed Configuration

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



➤ Next Generation ESnet: II) Virtual Circuits

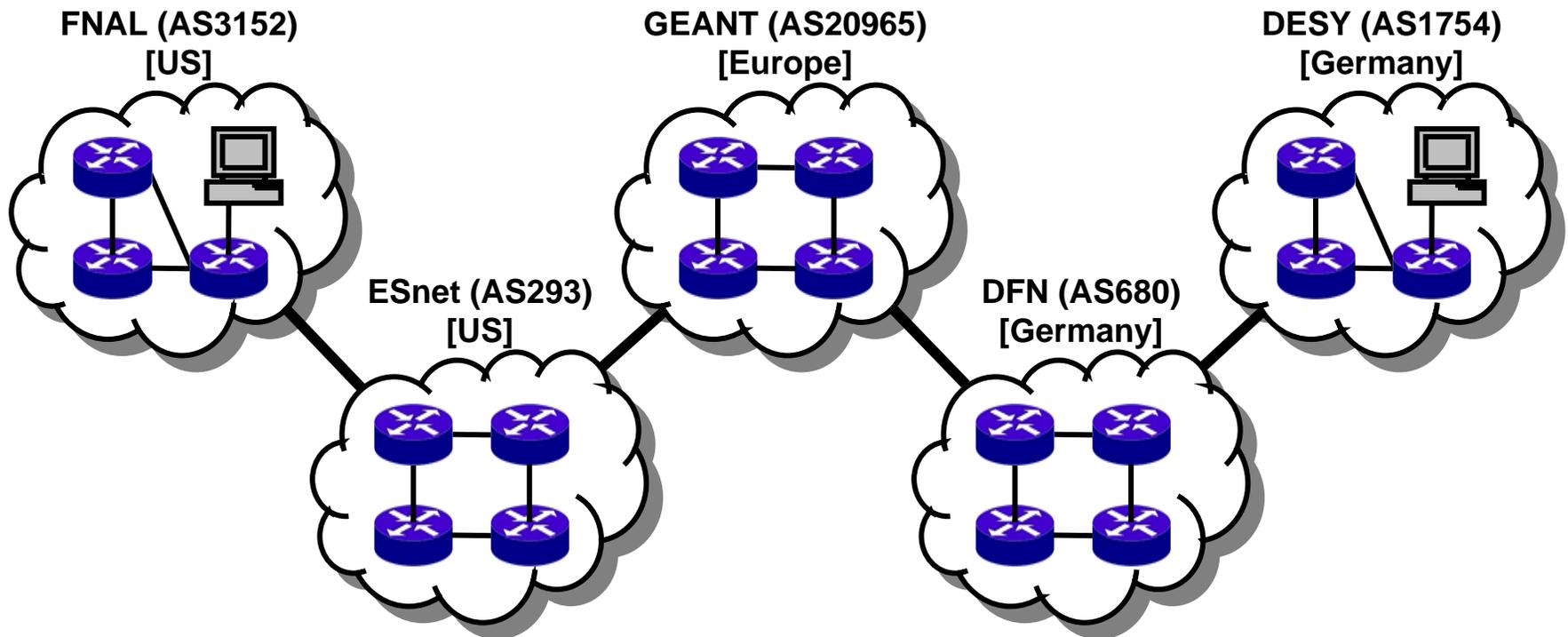
- Traffic isolation and traffic engineering
 - Provides for high-performance, non-standard transport mechanisms that cannot co-exist with commodity TCP-based transport
 - Enables the engineering of explicit paths to meet specific requirements
 - e.g. bypass congested links, using lower bandwidth, lower 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 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

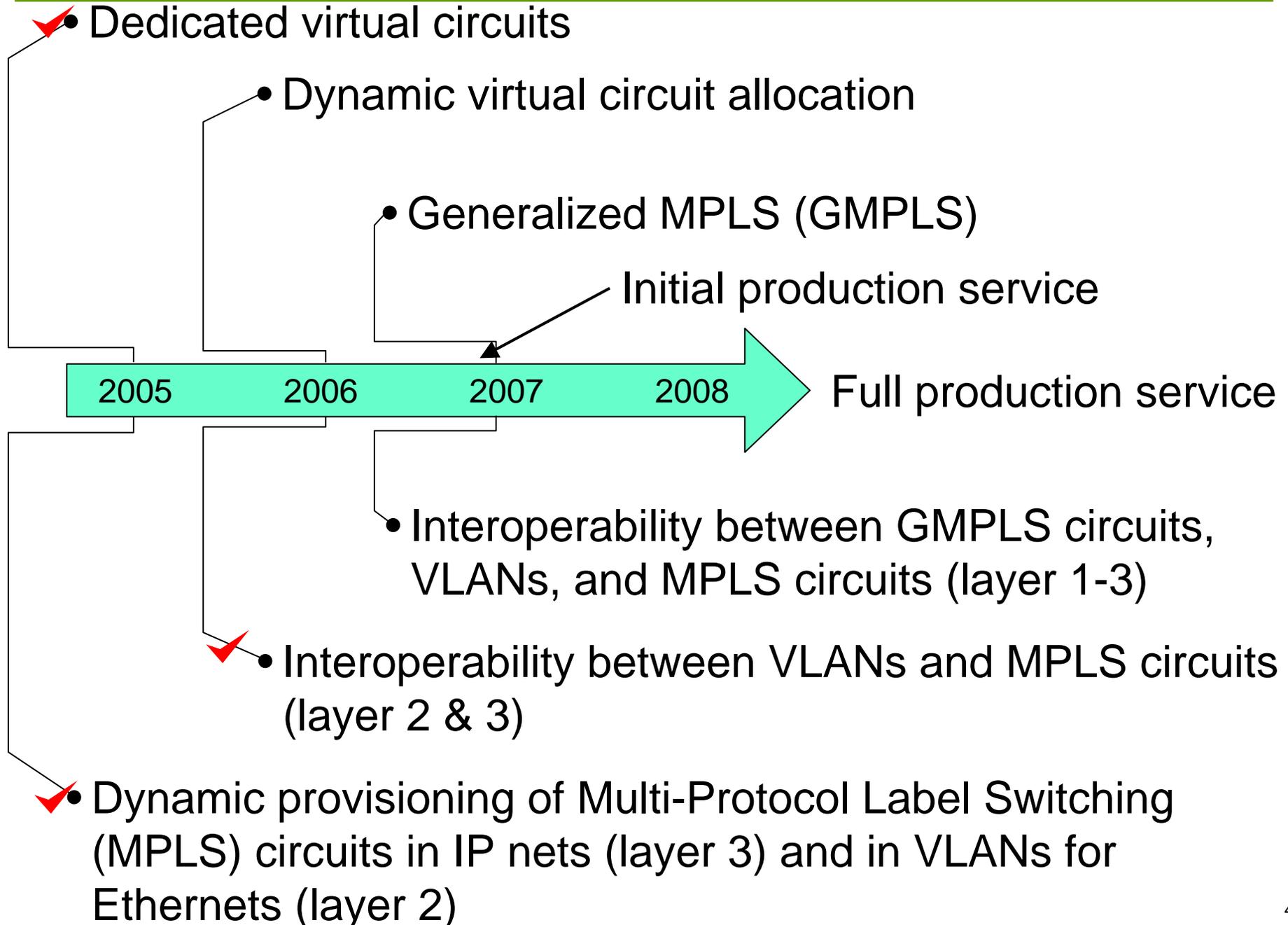
- Support user/application VC reservation requests
 - Source and destination of the VC
 - Bandwidth, start time, and duration of the VC
 - Traffic characteristics (e.g. flow specs) to identify traffic designated for the VC
- Manage allocations of scarce, shared resources
 - Authentication to prevent unauthorized access to this service
 - Authorization to enforce policy on reservation/provisioning
 - Gathering of usage data for accounting
- Provide 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 end-to-end connection setup is provided by the network control plane
- 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

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



ESnet Virtual Circuit Service Roadmap



➤ Summary

- **ESnet is currently satisfying its mission by enabling SC science that is dependant on networking and distributed, large-scale collaboration**
- **ESnet has put considerable effort into gathering requirements from the DOE science community, and has a forward-looking plan and expertise to meet the five-year SC requirements**

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