The Evolution of Research and Education Networks and their Essential Role in Modern Science

TERENA Networking Conference 2009

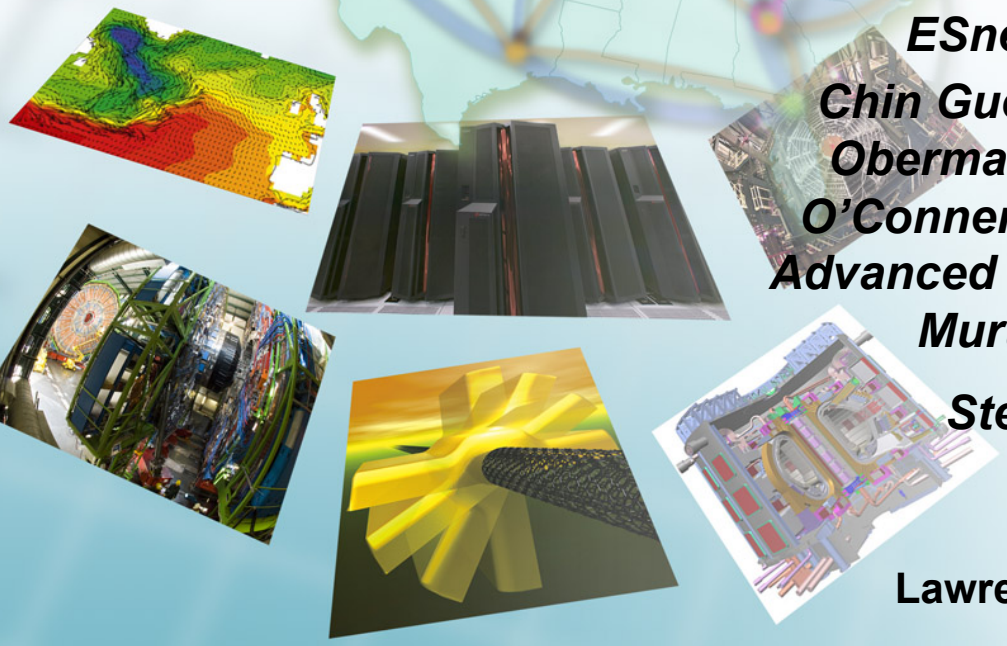
William E. Johnston,
ESnet Adviser and Senior Scientist

Chin Guok, Evangelos Chaniotakis, Kevin Oberman, Eli Dart, Joe Metzger and Mike O'Conner, Core Engineering, Brian Tierney, Advanced Development, Mike Helm and Dhiva Muruganantham, Federated Trust

Steve Cotter, Department Head

Energy Sciences Network
Lawrence Berkeley National Laboratory

wej@es.net, this talk is available at www.es.net



Networking for the Future of Science

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 US research programs in high-energy physics, nuclear physics, and fusion energy sciences. (www.science.doe.gov) – SC funds 25,000 PhDs and PostDocs
- A primary mission of SC's National Labs is to build and operate very large scientific instruments - particle accelerators, synchrotron light sources, very large supercomputers - that generate massive amounts of data and involve very large, distributed collaborations
- **ESnet - the Energy Sciences Network - is an SC program whose primary mission is to enable the large-scale science of the Office of Science 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
- In order to accomplish its mission SC/ASCAR funds ESnet to provide high-speed networking and various collaboration services to Office of Science laboratories

ESnet Approach to Supporting of the Office of Science Mission

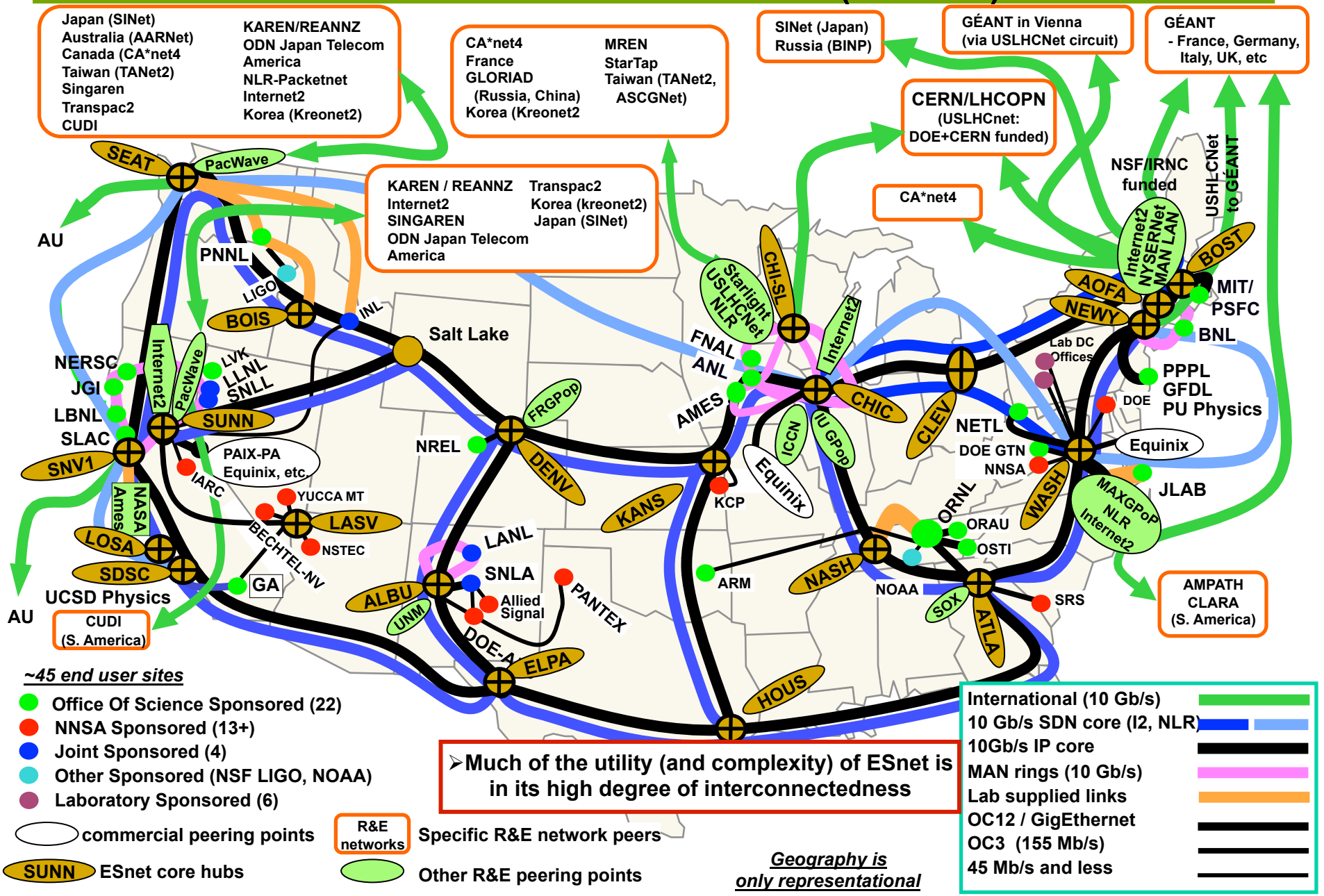
- The ESnet approach to supporting the science mission of the Office of Science involves
  - i) Identifying the networking implications of scientific instruments, supercomputers, and the evolving process of how science is done
  - ii) Developing approaches to building the network environment that will enable the distributed aspects of SC science, and
  - iii) Continually anticipating future network capabilities that will meet future science requirements
- This approach has lead to a high-speed network with highly redundant physical topology, services providing a hybrid packet-circuit network, and certain predictions about future network requirements.

- *What is ESnet?*

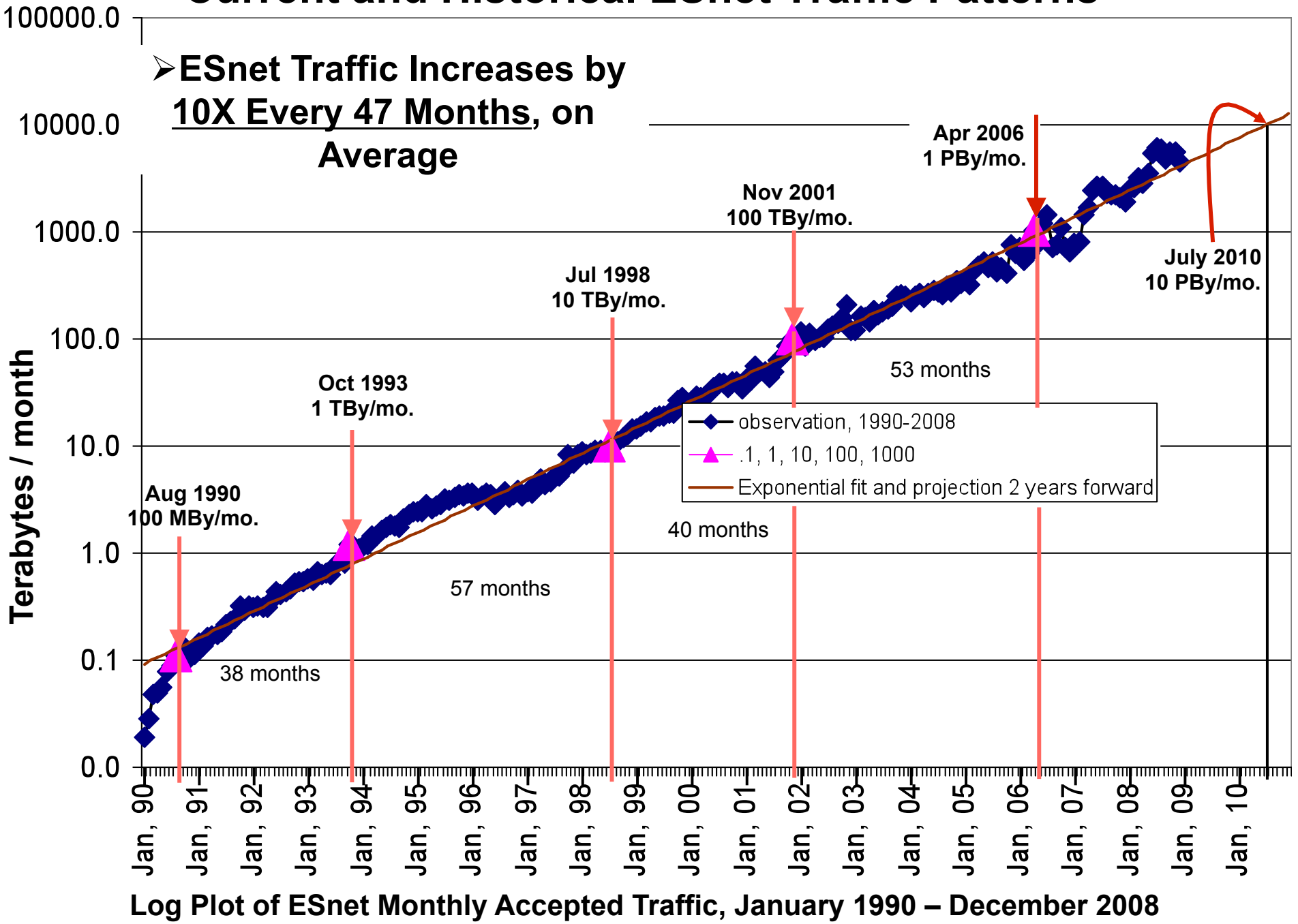
ESnet Defined

- A national optical circuit infrastructure
  - ESnet shares an optical network on a dedicated national fiber infrastructure with Internet2 (US national research and education (R&E) network)
    - ESnet has exclusive use of a group of 10Gb/s optical channels on this infrastructure
  - ESnet's two core networks – IP and SDN – are built on more than 125 10Gb/s WAN circuits
- A large-scale IP network
  - A tier 1 Internet Service Provider (ISP) (direct connections with all major commercial networks providers – "default free" routing)
- A large-scale science data transport network
  - With multiple 10Gb/s connections to all major US and international research and education (R&E) networks in order to enable large-scale science
  - Providing virtual circuit services specialized to carry the massive science data flows of the National Labs
- A WAN engineering support group for the DOE Labs
- An organization of 35 professionals structured for the service
  - The ESnet organization designs, builds, and operates the ESnet network based mostly on "managed wave" services from carriers and others
- An operating entity with an FY08 budget of about $30M
  - 60% of the operating budget is for circuits and related, remainder is staff and equipment related

ESnet Provides Global High-Speed Internet Connectivity for DOE Facilities and Collaborators (12/2008)

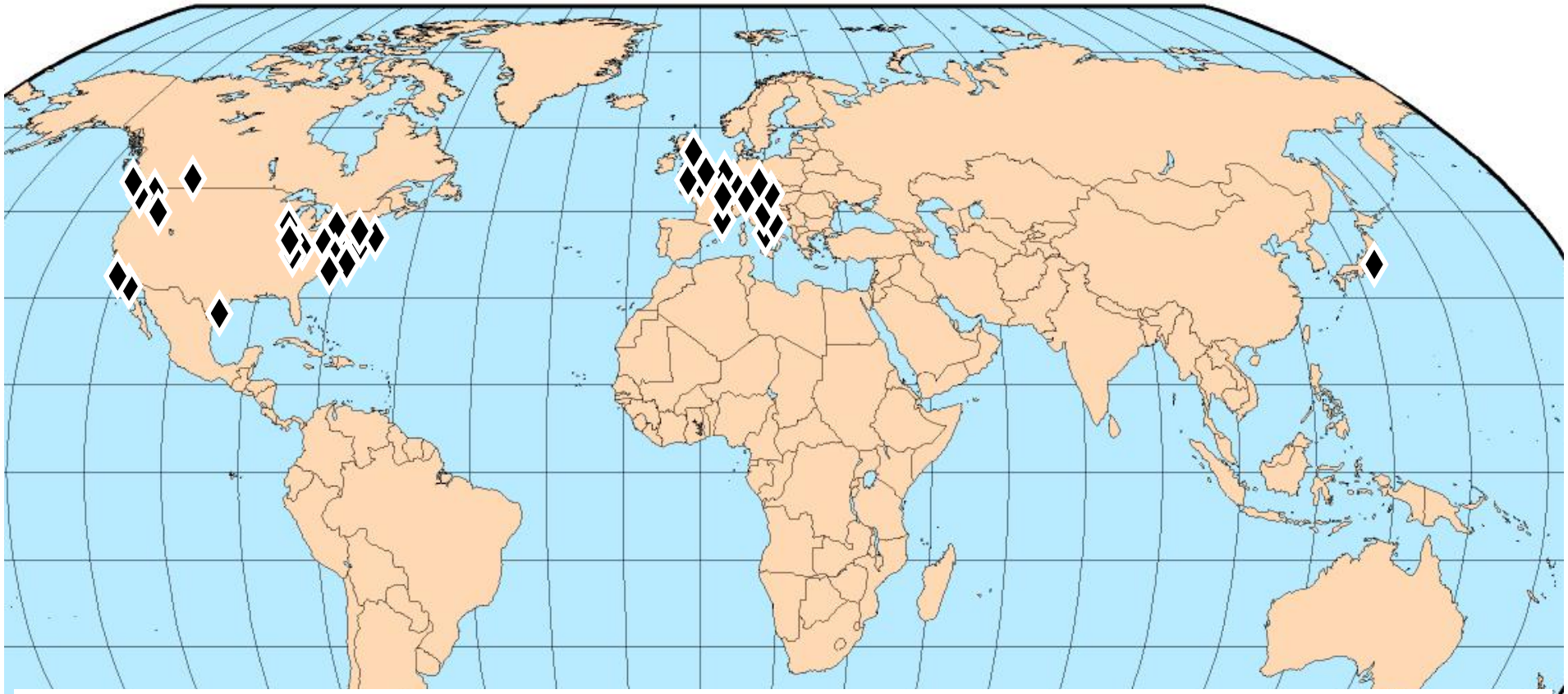


Current and Historical ESnet Traffic Patterns

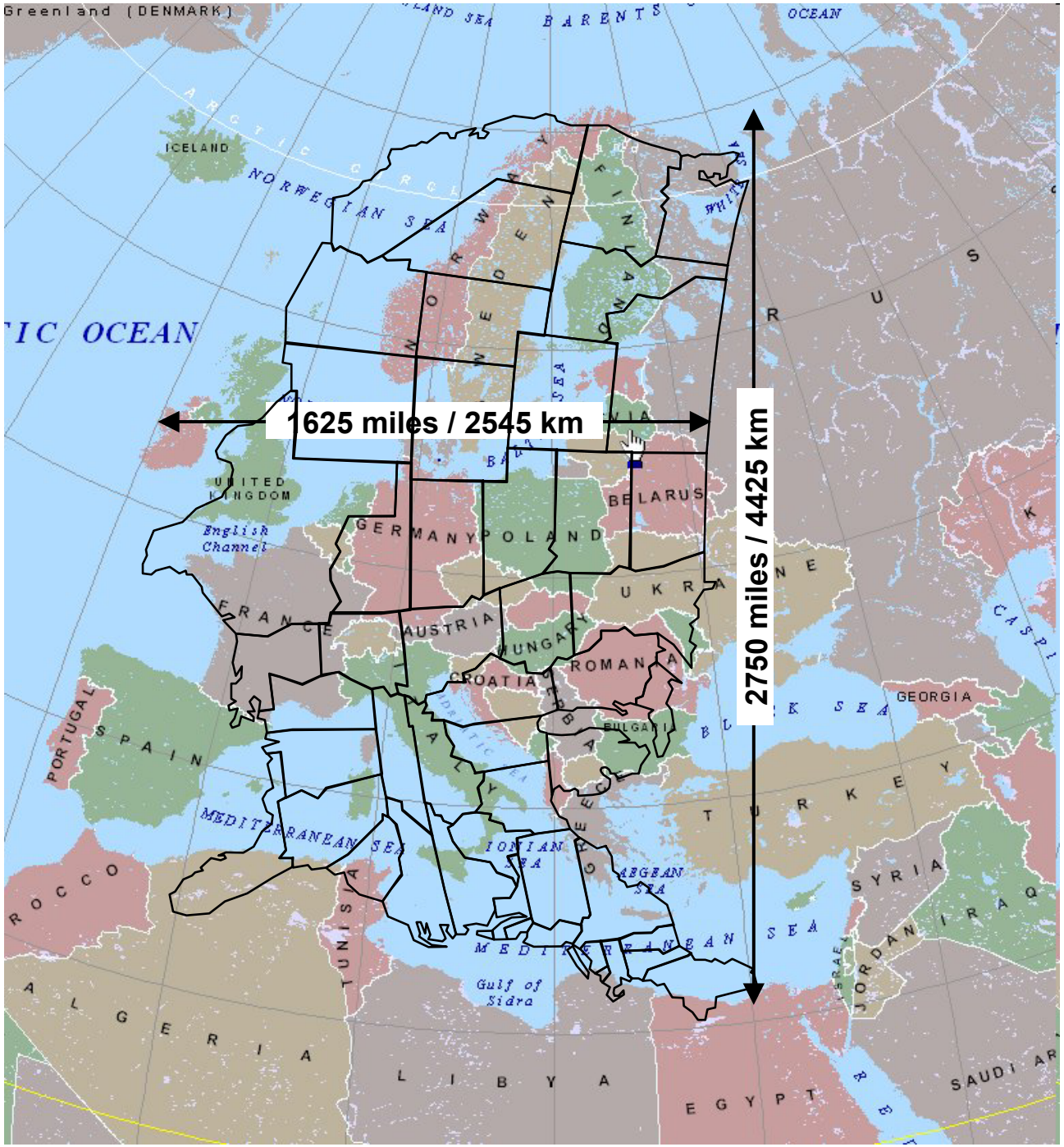


Log Plot of ESnet Monthly Accepted Traffic, January 1990 – December 2008

Most of ESnet's traffic (>85%) goes to and comes from outside of ESnet. This reflects the highly collaborative nature of the large-scale science of DOE's Office of Science.



⧫ = the R&E source or destination of ESnet's top 100 traffic generators / sinks, all of which are research and education institutions (the DOE Lab destination or source of each flow is not shown)



The Operational Challenge

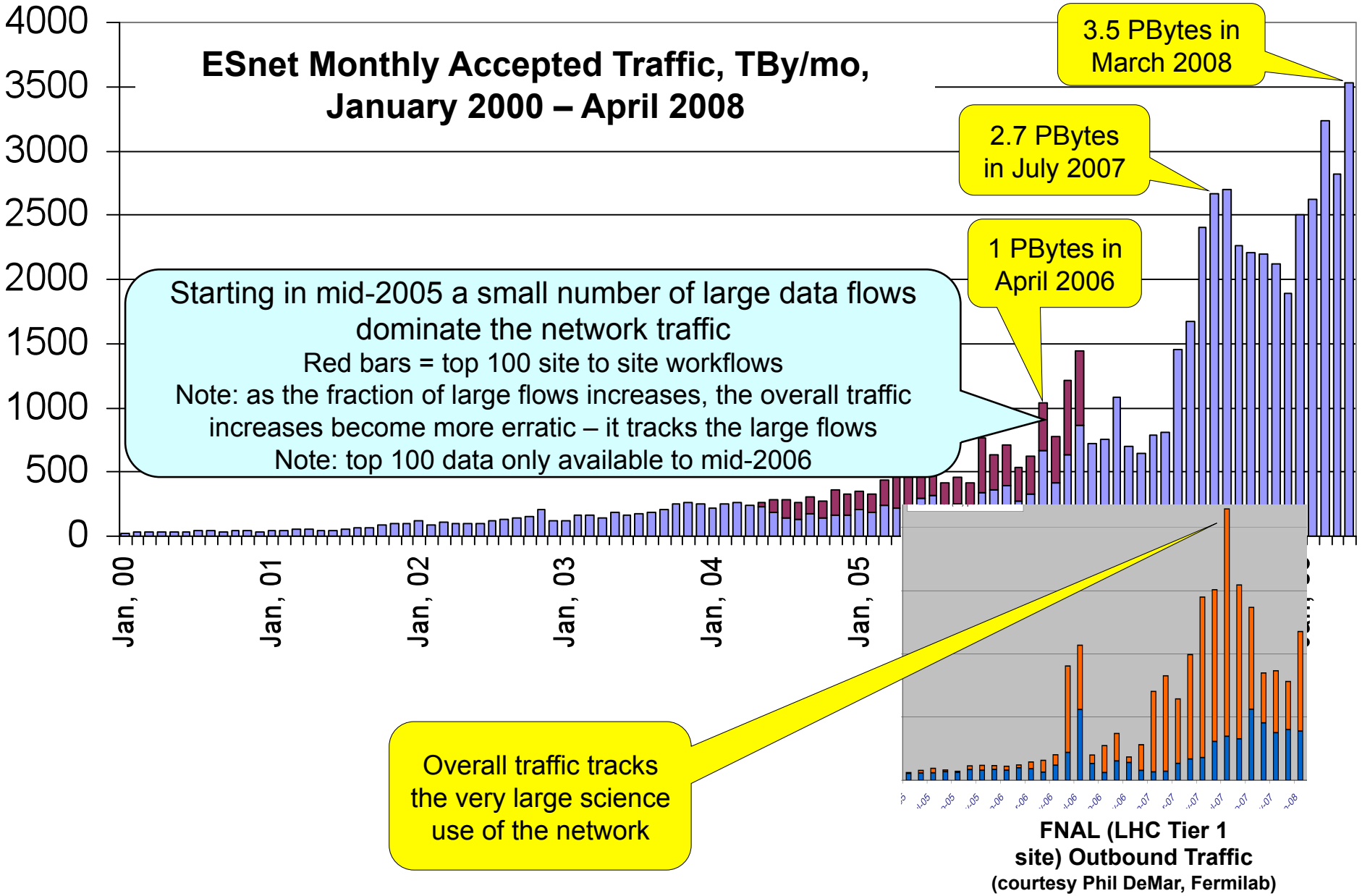
The relatively large geographic ESnet scale of makes it a challenge for a small organization to build, maintain, and operate the network.

➢ *The ESnet Planning Process*

How ESnet Determines its Network Architecture, Services, and Bandwidth

- Requirements are determined by
  1) **Observing current and historical network traffic patterns**
     - What do the trends in network patterns predict for future network needs?
  2) **Exploring the plans and processes of the major stakeholders** (the Office of Science programs, scientists, collaborators, and facilities):
     1a) Data characteristics of scientific instruments and facilities
        – What data will be generated by instruments and supercomputers coming on-line over the next 5-10 years?
     1b) Examining the future process of science
        – How and where will the new data be analyzed and used – that is, how will the process of doing science change over 5-10 years?

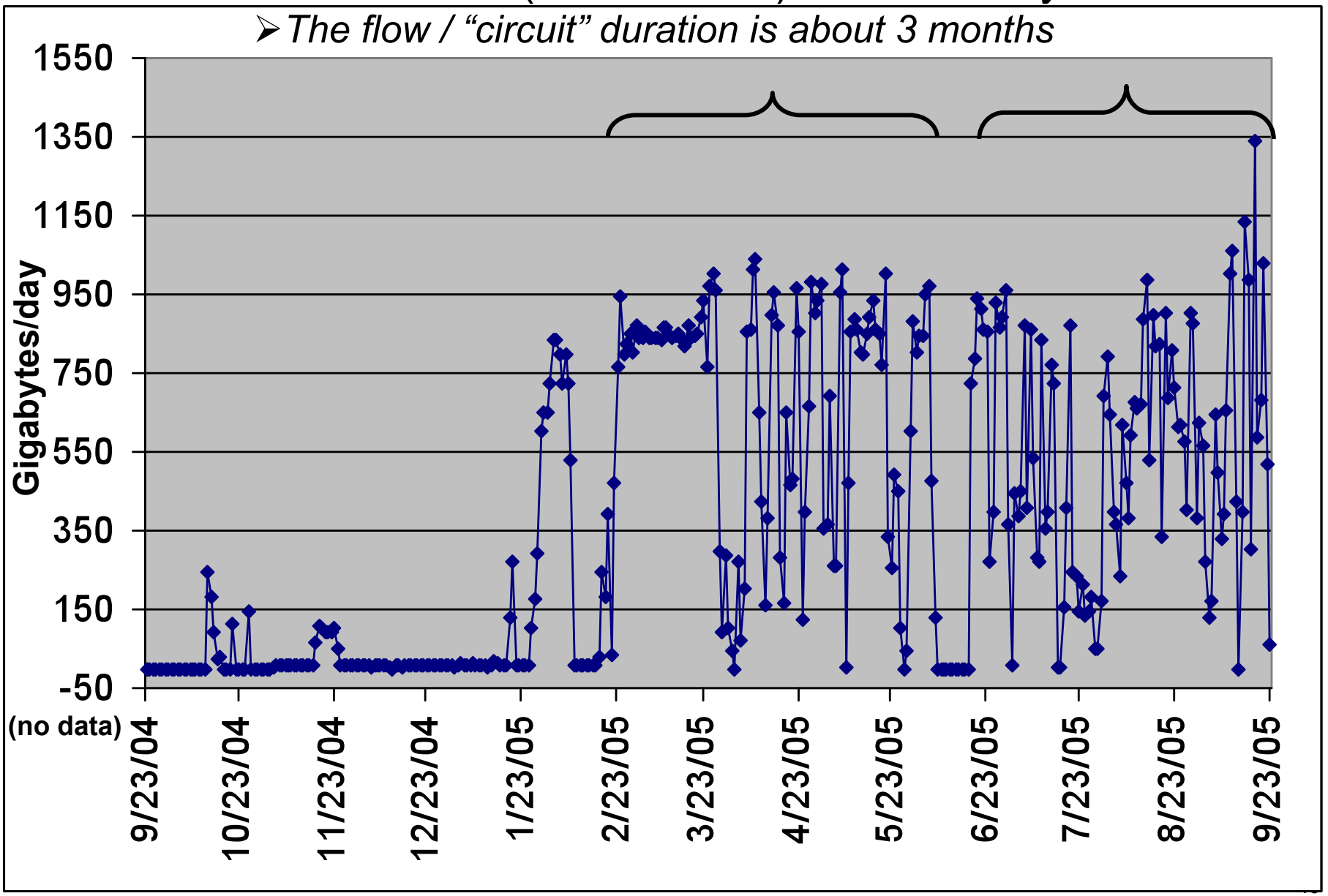
➢ **Observing the Network:** A small number of large data flows now dominate the network traffic – this is one motivator for virtual circuits as a key network service



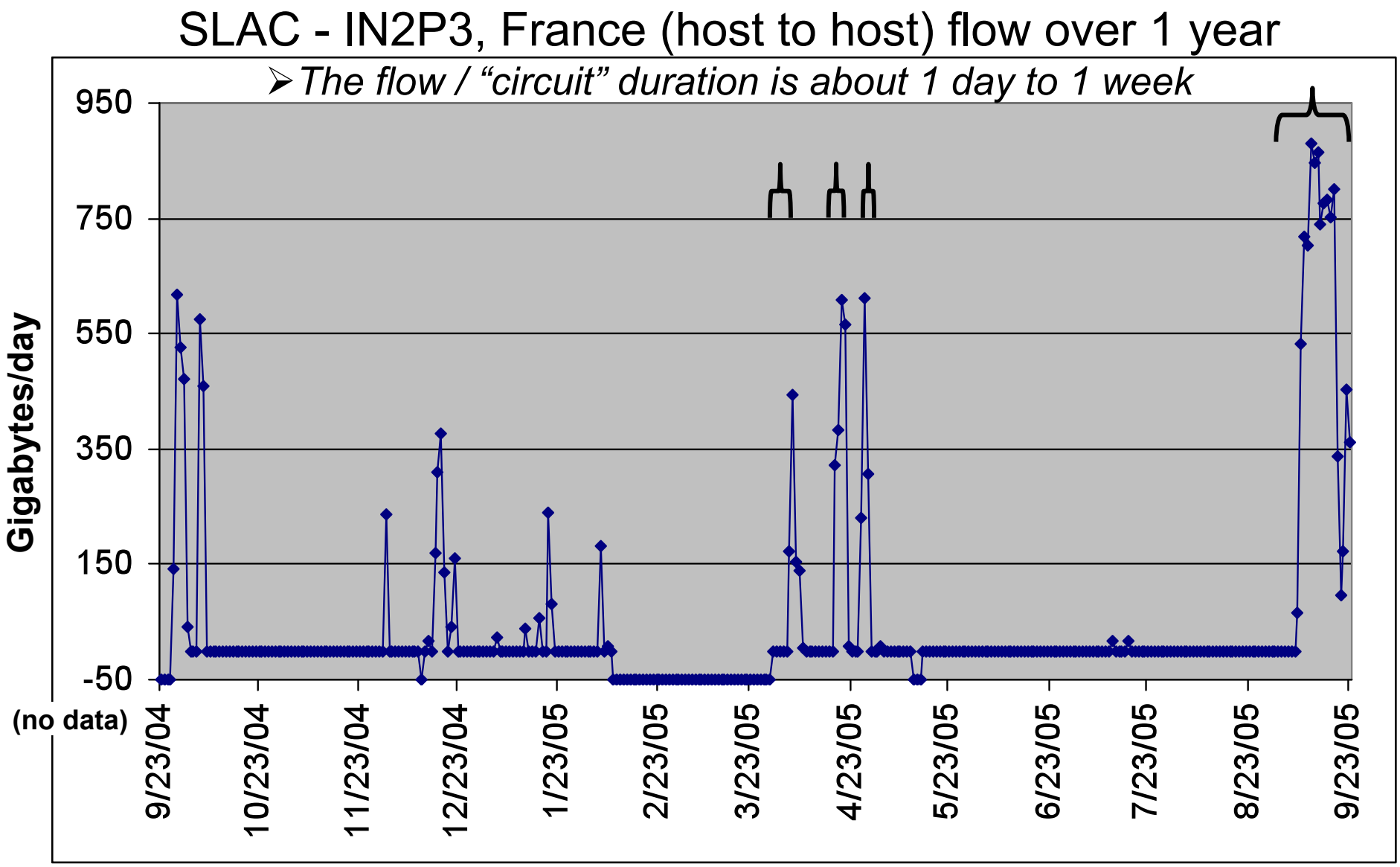
Most of the Large Flows Exhibit Circuit-like Behavior

LIGO – CalTech (host to host) flow over 1 year

- *The flow / "circuit" duration is about 3 months*



Most of the Large Flows Exhibit Circuit-like Behavior



Requirements from Observing Traffic Flow Trends

- **ESnet must have an architecture and strategy that allows scaling of the bandwidth available to the science community by a facto of 10x every 3-4 years**

➢ **Most ESnet traffic has a source or sink outside of ESnet**

- Drives requirement for high-bandwidth peering
- Reliability and bandwidth requirements demand that peering be redundant
- 10 Gbps peerings must be able to be added flexibly, quickly, and cost-effectively

➢ **Large-scale science is now the dominant user of the network and this traffic is circuit-like (long duration, same source/destination)**

- Will consume 95% of ESnet bandwidth
- Since large-scale science traffic is the dominant user of the network, and the network must be architected to serve large-scale science as a first consideration
  - Traffic patterns are very different than commodity Internet – the "flows" are circuit-like and vastly greater than all commodity traffic
- Even apart from user services requirements, large-scale science traffic inherently exhibits circuit-like behavior
  - This circuit-like behavior of the large flows of science data requires ESnet to be able to do traffic engineering to optimize the use of the network

➢ Exploring the plans of the major stakeholders

- Primary mechanism is Office of Science (SC) network Requirements Workshops, which are organized by the SC Program Offices; Two workshops per year - workshop schedule, which repeats in 2010
  - Basic Energy Sciences (materials sciences, chemistry, geosciences) (2007 – published)
  - Biological and Environmental Research (2007 – published)
  - Fusion Energy Science (2008 – published)
  - Nuclear Physics (2008 – published)
  - IPCC (Intergovernmental Panel on Climate Change) special requirements (BER) (August, 2008)
  - Advanced Scientific Computing Research (applied mathematics, computer science, and high-performance networks) (Spring 2009)
  - High Energy Physics (Summer 2009)
- Workshop reports: http://www.es.net/hypertext/requirements.html
- The Office of Science National Laboratories (there are additional free-standing facilities) include
  - Ames Laboratory
  - Argonne National Laboratory (ANL)
  - Brookhaven National Laboratory (BNL)
  - Fermi National Accelerator Laboratory (FNAL)
  - Thomas Jefferson National Accelerator Facility (JLab)
  - Lawrence Berkeley National Laboratory (LBNL)
  - Oak Ridge National Laboratory (ORNL)
  - Pacific Northwest National Laboratory (PNNL)
  - Princeton Plasma Physics Laboratory (PPPL)
  - SLAC National Accelerator Laboratory (SLAC)

Science Network Requirements Aggregation Summary

Science Drivers Science Areas / Facilities	End2End Reliability	Near Term End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
ASCR: ALCF	-	10Gbps	30Gbps	• Bulk data • Remote control • Remote file system sharing	• Guaranteed bandwidth • Deadline scheduling • PKI / Grid
ASCR: NERSC	-	10Gbps	20 to 40 Gbps	• Bulk data • Remote control • Remote file system sharing	• Guaranteed bandwidth • Deadline scheduling • PKI / Grid
ASCR: NLCF	-	Backbone Bandwidth Parity	Backbone Bandwidth Parity	•Bulk data •Remote control •Remote file system sharing	• Guaranteed bandwidth • Deadline scheduling • PKI / Grid
BER: Climate	Note that the climate numbers do not reflect the bandwidth that will be needed for the 4 PBy IPCC data sets shown in the Capacity comparison graph below	3Gbps	10 to 20Gbps	• Bulk data • Rapid movement of GB sized files • Remote Visualization	• Collaboration services • Guaranteed bandwidth • PKI / Grid
BER: EMSL/Bio		10Gbps	50-100Gbps	• Bulk data • Real-time video • Remote control	• Collaborative services • Guaranteed bandwidth
BER: JGI/Genomics	-	1Gbps	2-5Gbps	• Bulk data	• Dedicated virtual circuits • Guaranteed bandwidth

Science Network Requirements Aggregation Summary

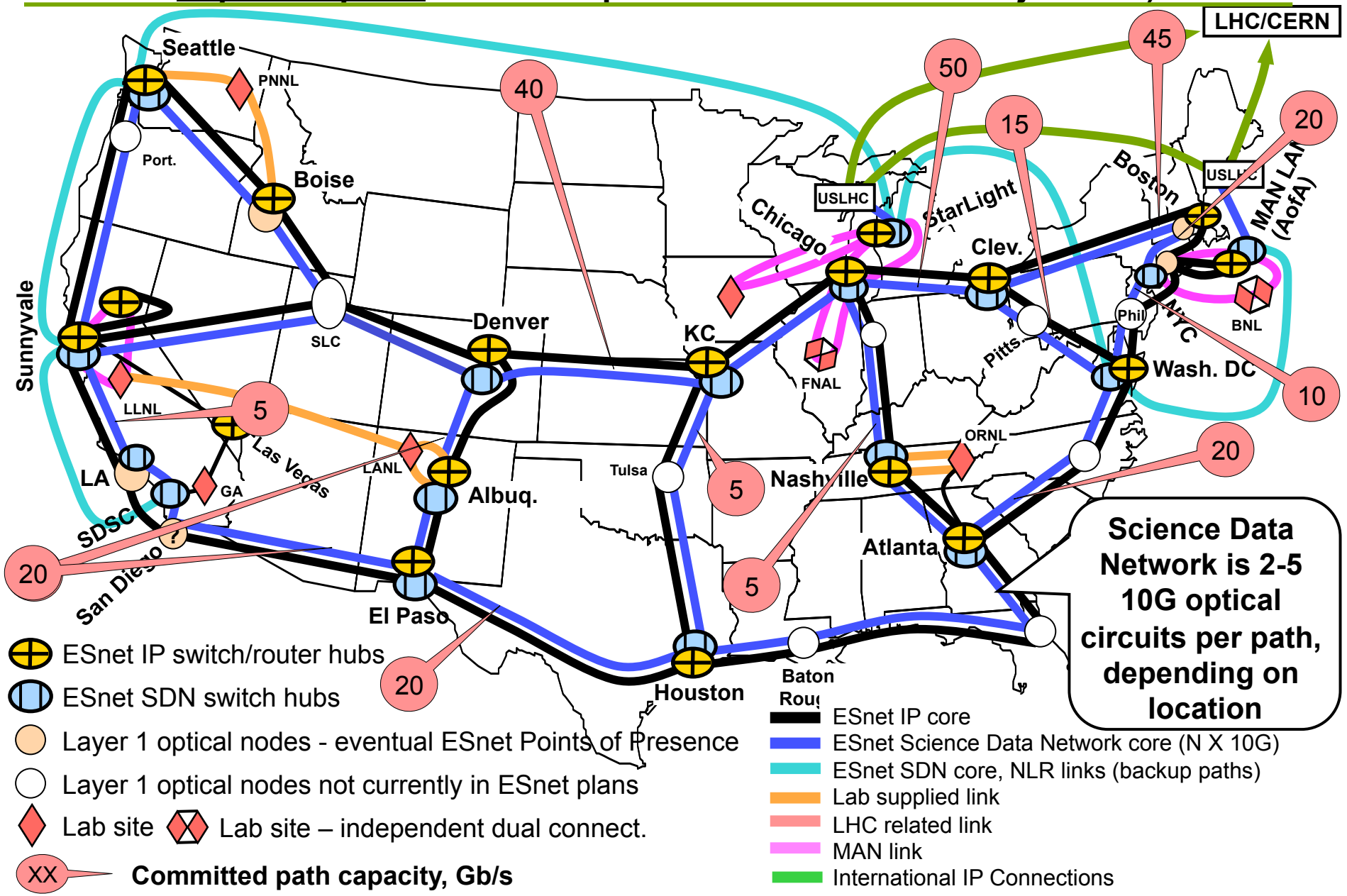
Science Drivers Science Areas / Facilities	End2End Reliability	Near Term End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
BES: Chemistry and Combustion	-	5-10Gbps	30Gbps	• Bulk data • Real time data streaming	• Data movement middleware
BES: Light Sources	-	15Gbps	40-60Gbps	• Bulk data • Coupled simulation and experiment	• Collaboration services • Data transfer facilities • Grid / PKI • Guaranteed bandwidth
BES: Nanoscience Centers	-	3-5Gbps	30Gbps	•Bulk data •Real time data streaming •Remote control	• Collaboration services • Grid / PKI
FES: International Collaborations	-	100Mbps	1Gbps	• Bulk data	• Enhanced collaboration services • Grid / PKI • Monitoring / test tools
FES: Instruments and Facilities	-	3Gbps	20Gbps	• Bulk data • Coupled simulation and experiment • Remote control	• Enhanced collaboration service • Grid / PKI
FES: Simulation	-	10Gbps	88Gbps	• Bulk data • Coupled simulation and experiment • Remote control	• Easy movement of large checkpoint files • Guaranteed bandwidth • Reliable data transfer

Science Network Requirements Aggregation Summary

Science Drivers Science Areas / Facilities	End2End Reliability	Near Term End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
	Immediate Requirements and Drivers for ESnet4				
HEP: LHC (CMS and Atlas)	99.95+% (Less than 4 hours per year)	73Gbps	225-265Gbps	• Bulk data • Coupled analysis workflows	• Collaboration services • Grid / PKI • Guaranteed bandwidth • Monitoring / test tools
NP: CMS Heavy Ion	-	10Gbps (2009)	20Gbps	•Bulk data	• Collaboration services • Deadline scheduling • Grid / PKI
NP: CEBF (JLAB)	-	10Gbps	10Gbps	• Bulk data	• Collaboration services • Grid / PKI
NP: RHIC	Limited outage duration to avoid analysis pipeline stalls	6Gbps	20Gbps	• Bulk data	• Collaboration services • Grid / PKI • Guaranteed bandwidth • Monitoring / test tools

Bandwidth – Path Requirements

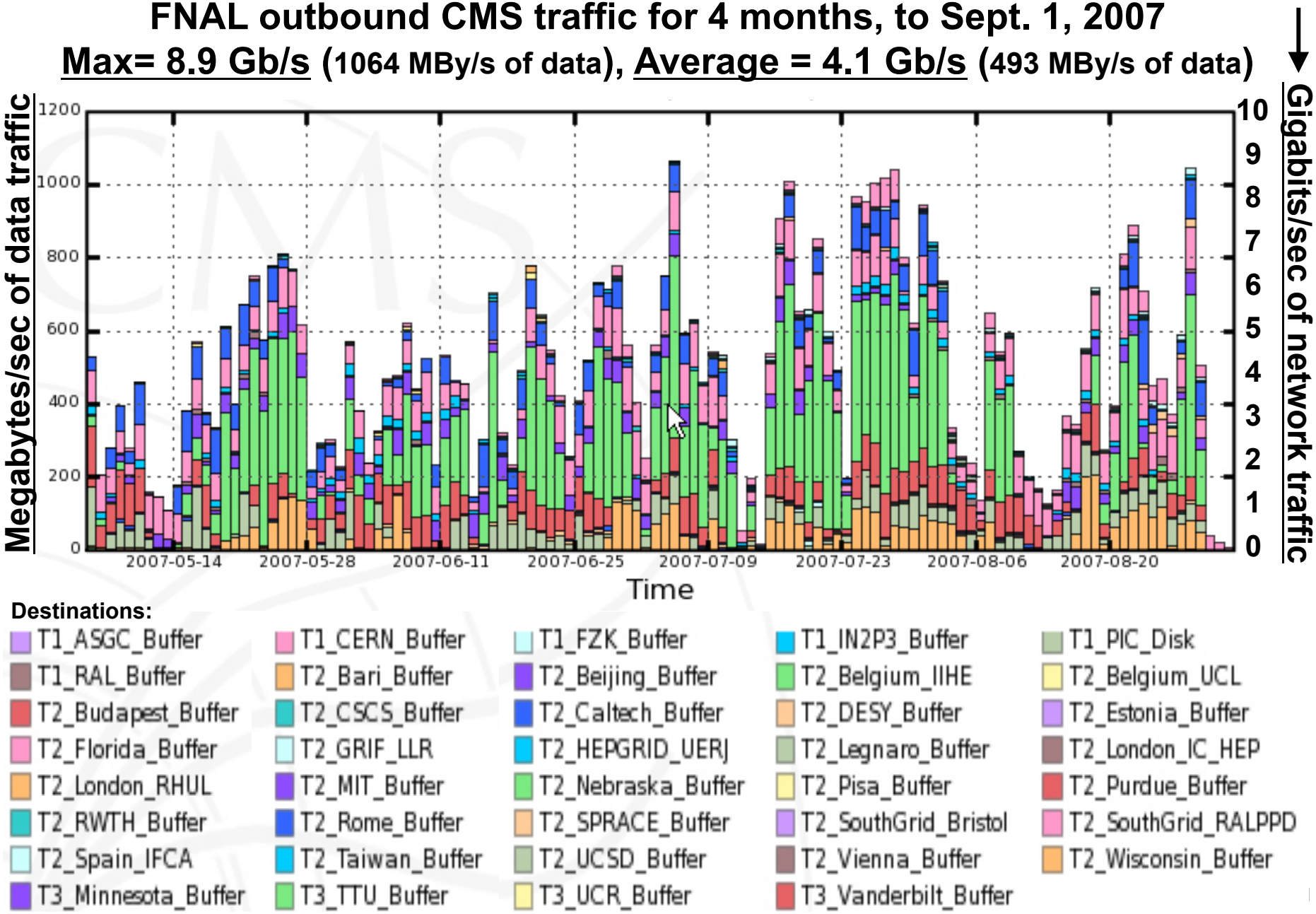
Mapping to the Network for the 2010 Network (Based only on *LHC, RHIC, and Supercomputer* Stated Requirements and Traffic Projections)



Are These Estimates Realistic? Yes.

FNAL outbound CMS traffic for 4 months, to Sept. 1, 2007

Max= 8.9 Gb/s (1064 MBy/s of data), Average = 4.1 Gb/s (493 MBy/s of data)



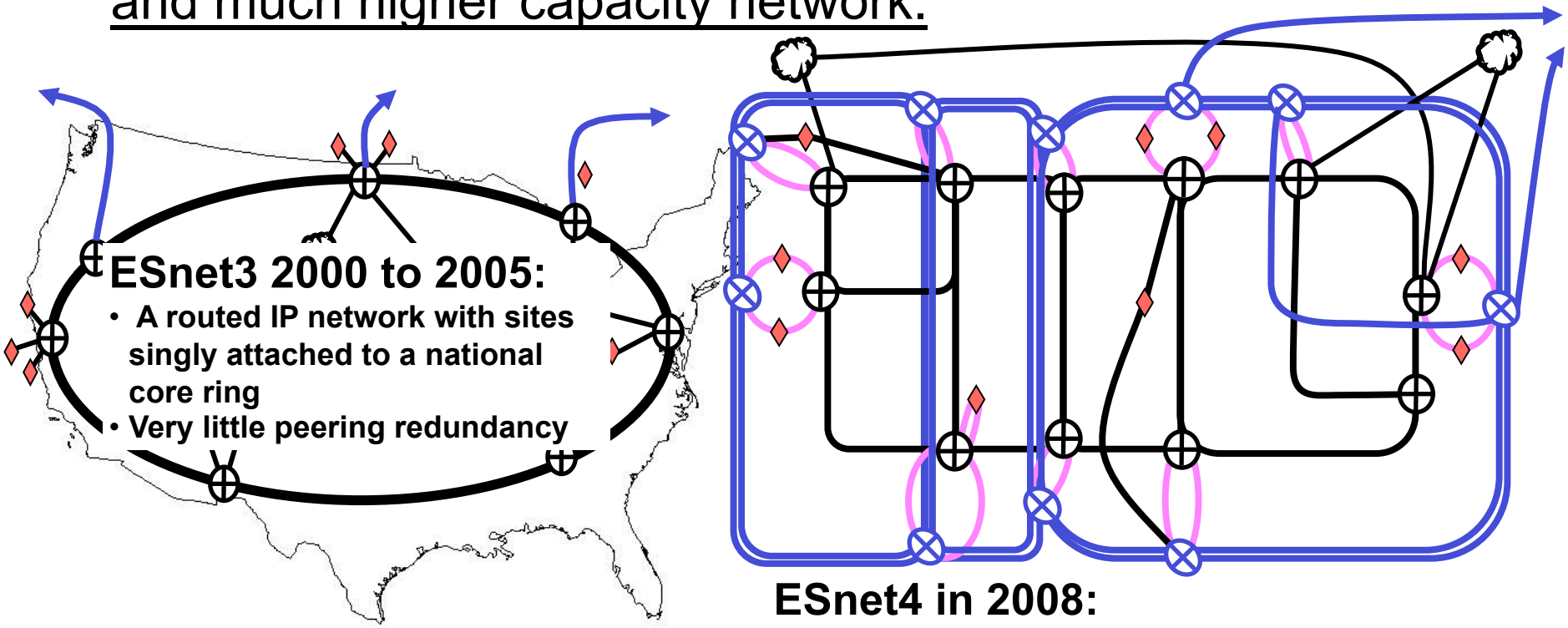
Do We Have the Whole Picture?

- ***However – is this the whole story? (No)***
  - More later ……

- *ESnet Response to the Requirements*

Strategy I) Provide the basic, long-term bandwidth requirements with an adequate and scalable infrastructure

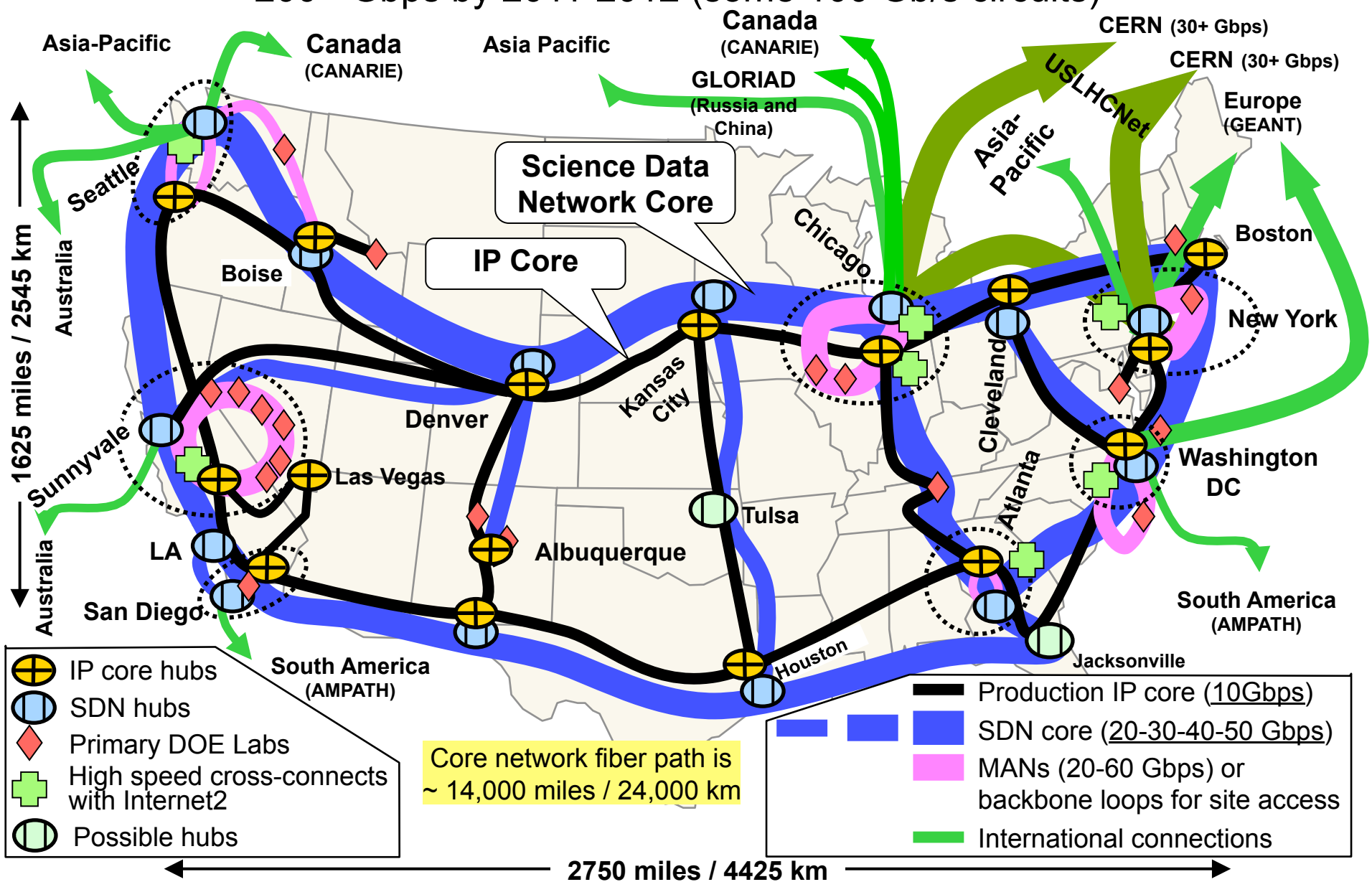
- ESnet4 was built to address specific Office of Science program requirements. The result is a much more complex and much higher capacity network.



- **The new Science Data Network (blue) uses MPLS to provide virtual circuits with guaranteed bandwidth for large data movement**
- **The large science sites are dually connected on metro area rings or dually connected directly to core ring for reliability**
- **Rich topology increases the reliability and flexibility of the network**

2012 Planned ESnet4

Core networks 50-60 Gbps by 2009-2010 (10Gb/s circuits),
200+ Gbps by 2011-2012 (some 100 Gb/s circuits)



Strategy II) A Service-Oriented Virtual Circuit Service

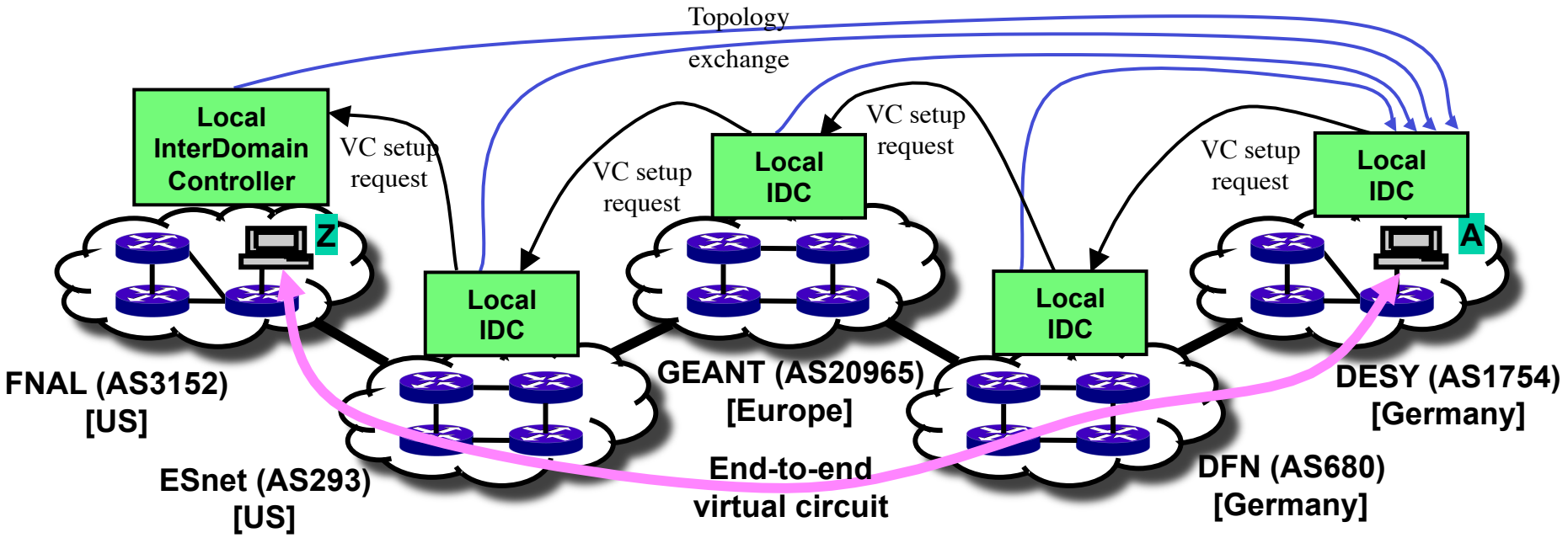
Multi-Domain Virtual Circuits as a Service – "OSCARS" – ESnet's InterDomain Controller

Service Characteristics:

- Guaranteed bandwidth with resiliency
  - User specified bandwidth - requested and managed in a Web Services framework
  - Explicit backup paths can be requested
- Traffic isolation
  - Allows for high-performance, non-standard transport mechanisms that cannot co-exist with commodity TCP-based transport
- Traffic engineering (for ESnet operations)
  - Enables the engineering of explicit paths to meet specific requirements
    - e.g. bypass congested links; using higher bandwidth, lower latency paths; etc.
- Secure connections
  - 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 highly secure and isolated from general traffic
- End-to-end, cross-domain connections between Labs and collaborating institutions

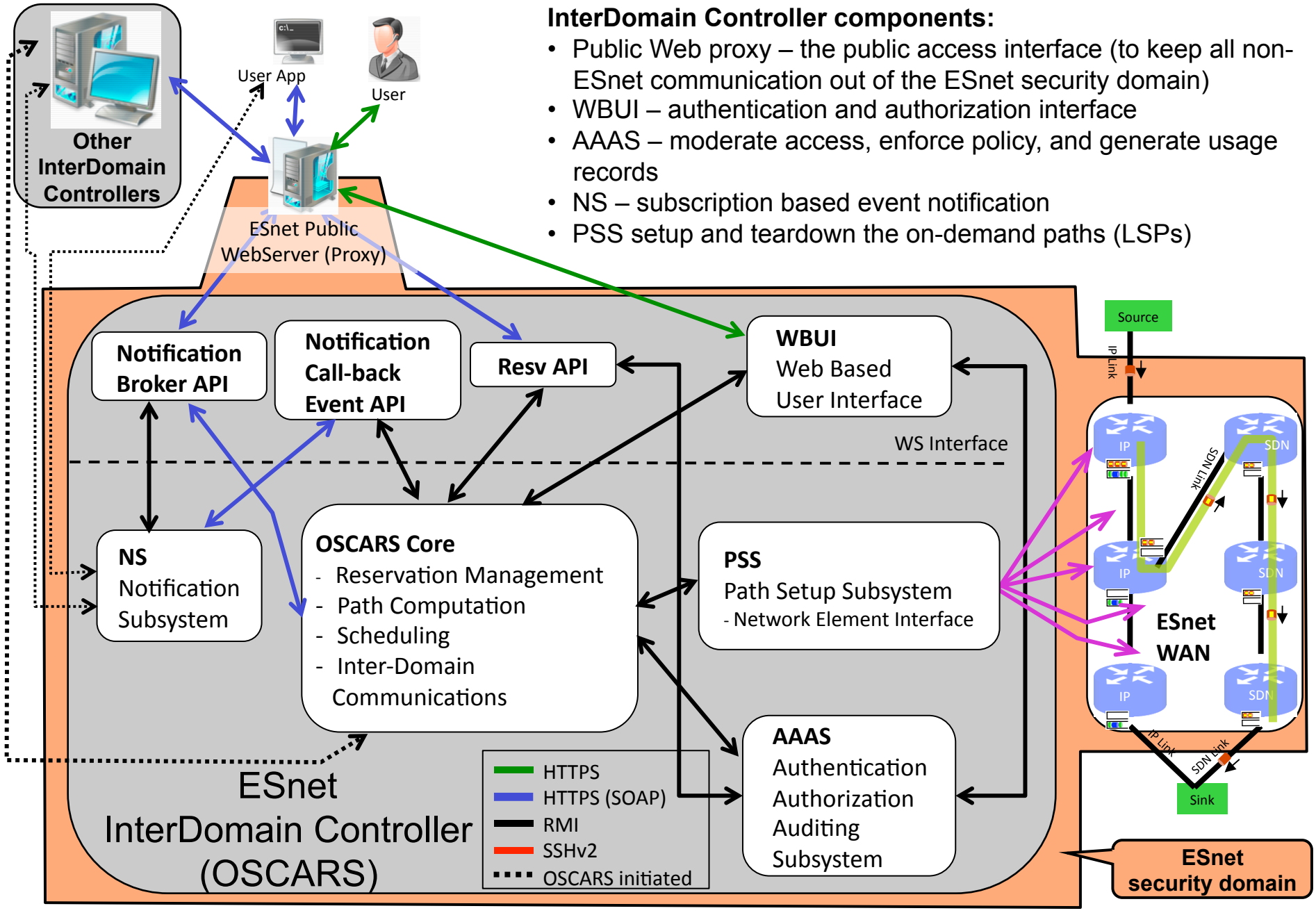
➢Environment of Science is Inherently Multi-Domain

- Inter-domain interoperability is crucial to serving science
- An effective international R&E collaboration has standardized inter-domain (inter-IDC) control protocol – "IDCP" (ESnet, Internet2, GÉANT, USLHCnet, several European NRENs, etc.)
- In order to set up end-to-end circuits across multiple domains:
  1. The domains exchange topology information containing at least potential VC ingress and egress points
  2. VC setup request (via IDCP protocol) is initiated at one end of the circuit and passed from domain to domain as the VC segments are authorized and reserved



Example

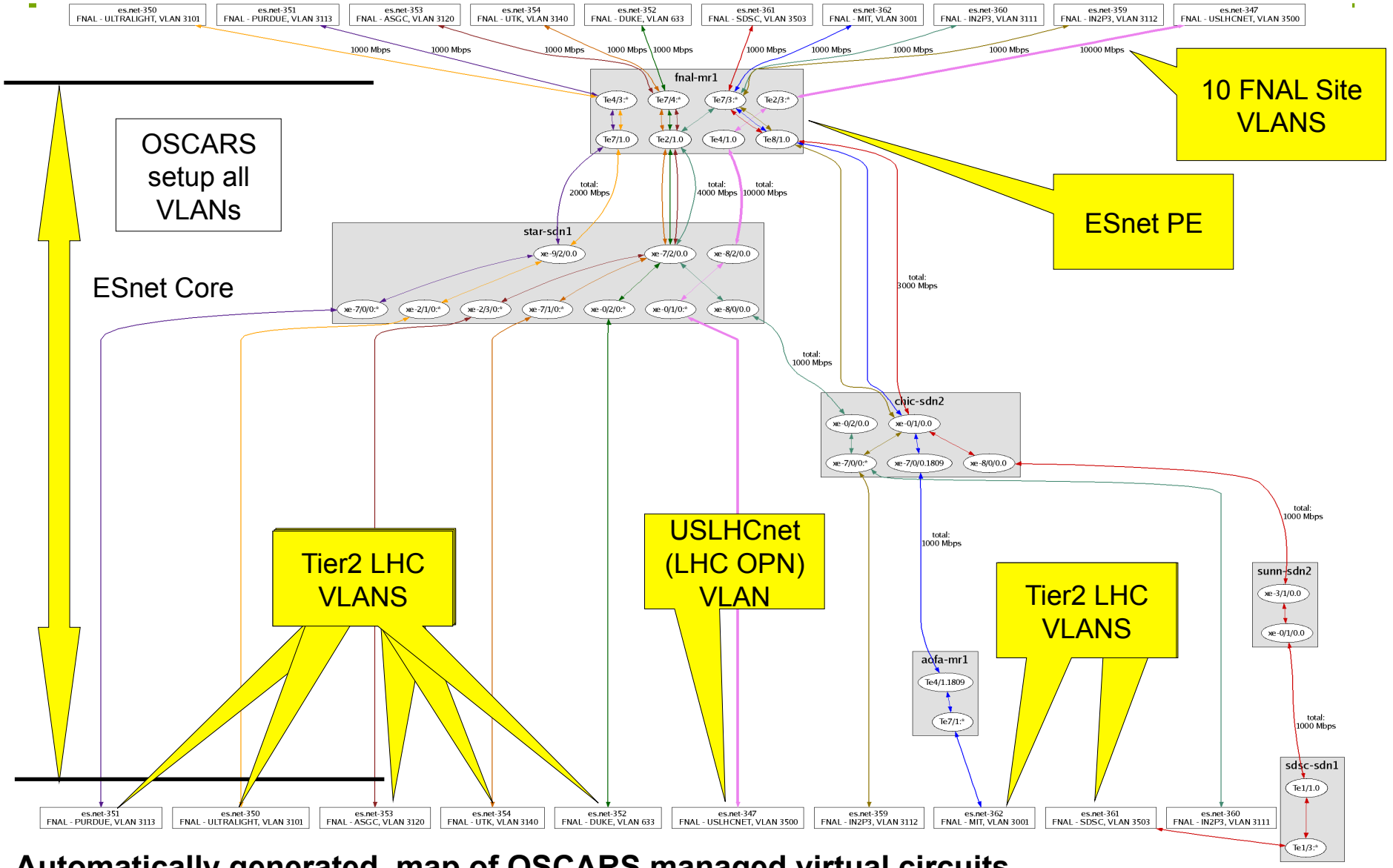
OSCARS Services Overview



InterDomain Controller components:

- Public Web proxy – the public access interface (to keep all non-ESnet communication out of the ESnet security domain)
- WBUI – authentication and authorization interface
- AAAS – moderate access, enforce policy, and generate usage records
- NS – subscription based event notification
- PSS setup and teardown the on-demand paths (LSPs)

➢OSCARS is a production service in ESnet



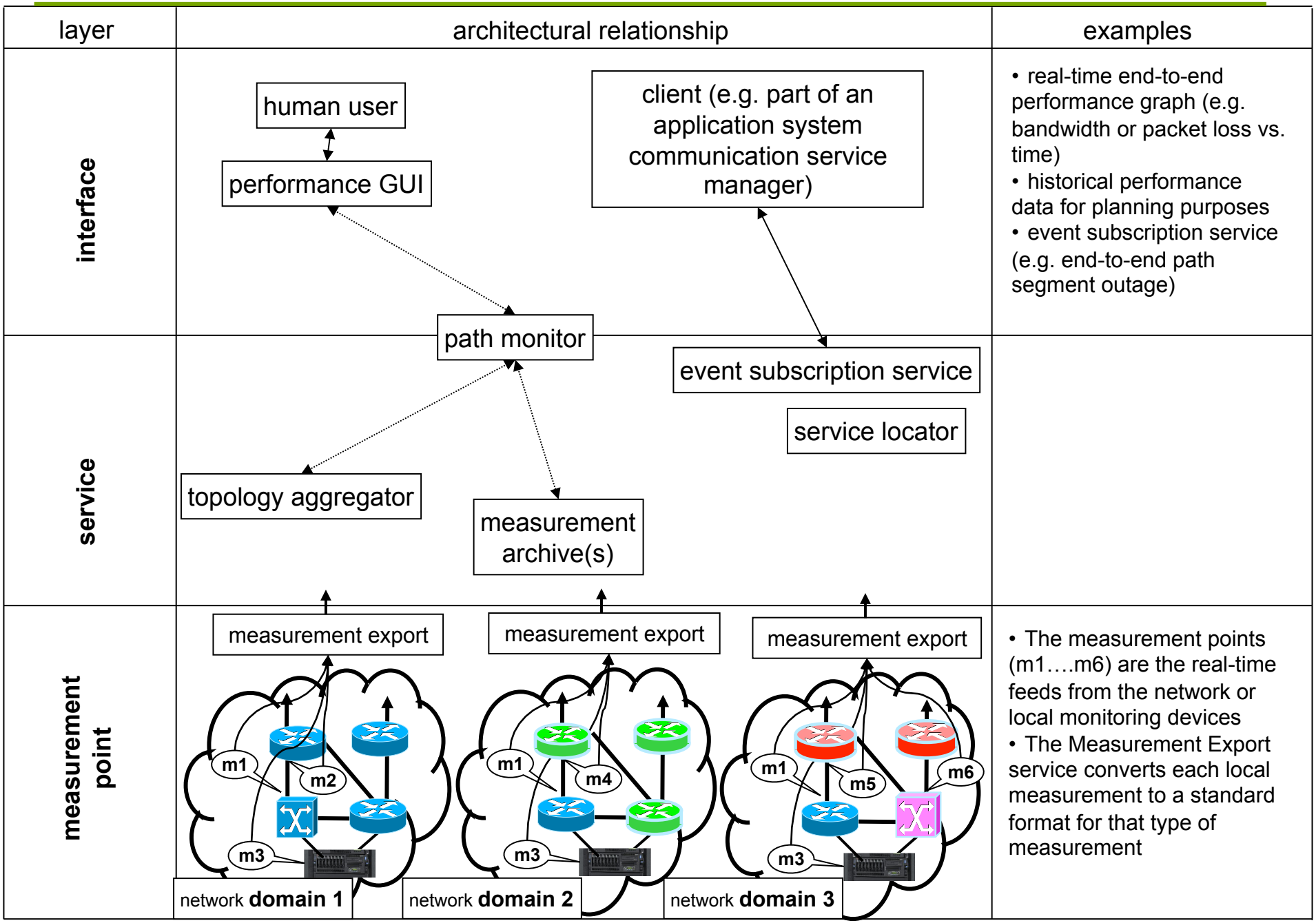
Automatically generated map of OSCARS managed virtual circuits

E.g.: FNAL – one of the US LHC Tier 1 data centers. This circuit map (minus the yellow callouts that explain the diagram) is automatically generated by an OSCARS tool and assists the connected sites with keeping track of what circuits exist and where they terminate.

Strategy III: Monitoring as a Service-Oriented Communications Service

- perfSONAR is a community effort to define network management data exchange protocols, and standardized measurement data gathering and archiving
  - Widely used in international and LHC networks
- The protocol is based on SOAP XML messages and follows work of the Open Grid Forum (OGF) Network Measurement Working Group (NM-WG)
- Has a layered architecture and a modular implementation
  - Basic components are
    - the “measurement points” that collect information from network devices (actually most anything) and export the data in a standard format
    - a measurement archive that collects and indexes data from the measurement points
  - Other modules include an event subscription service, a topology aggregator, service locator (where are all of the archives?), a path monitor that combines information from the topology and archive services, etc.
  - Applications like the *traceroute visualizer* and *E2EMON* (the GÉANT end-to-end monitoring system) are built on these services

perfSONAR Architecture



Traceroute Visualizer

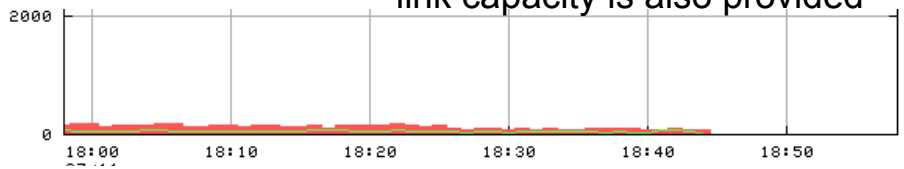
- Multi-domain path performance monitoring is an example of a tool based on perfSONAR protocols and infrastructure
  - provide users/applications with the end-to-end, multi-domain traffic and bandwidth availability
  - provide real-time performance such as path utilization and/or packet drop
  - One example – Traceroute Visualizer [TrViz] – has been deployed in about 10 R&E networks in the US and Europe that have deployed at least some of the required perfSONAR measurement archives to support the tool

Traceroute Visualizer

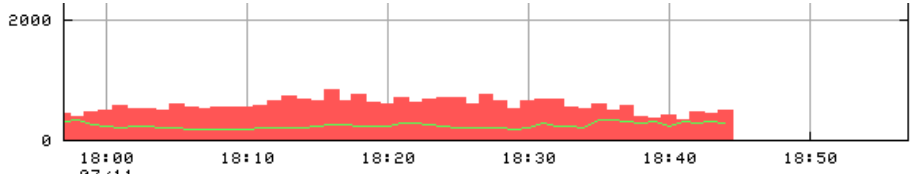
- Forward direction bandwidth utilization on application path from LBNL to INFN-Frascati (Italy) (2008 SNAPSHOT)
  - traffic shown as bars on those network device interfaces that have an associated MP services (the first 4 graphs are normalized to 2000 Mb/s, the last to 500 Mb/s)

1 ir1000gw (131.243.2.1)
2 er1kgw
3 lbl2-ge-lbnl.es.net

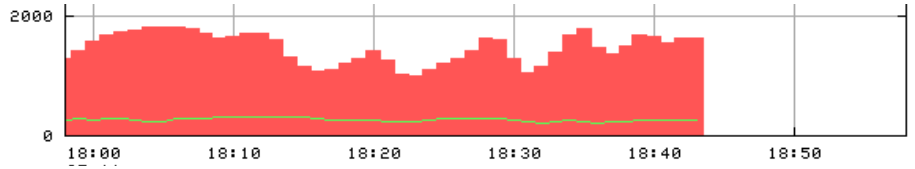
link capacity is also provided



4 slacmr1-sdn-lblmr1.es.net (GRAPH OMITTED)
5 snv2mr1-slacmr1.es.net (GRAPH OMITTED)
6 snv2sdn1-snv2mr1.es.net

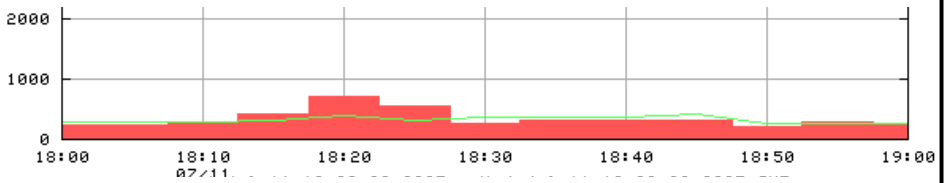


7 chislsdn1-oc192-snv2sdn1.es.net (GRAPH OMITTED)
8 chiccr1-chislsdn1.es.net

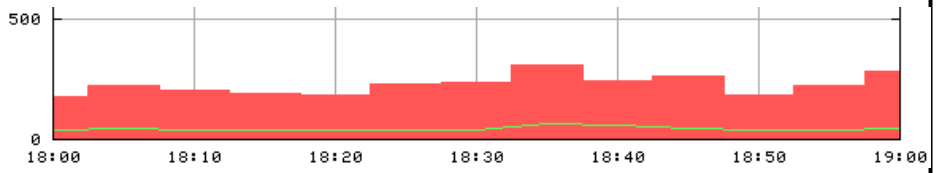


9 aofacr1-chicsdn1.es.net (GRAPH OMITTED)

10 esnet.rt1.nyc.us.geant2.net (NO DATA)
11 so-7-0-0.rt1.ams.nl.geant2.net (NO DATA)
12 so-6-2-0.rt1.fra.de.geant2.net (NO DATA)
13 so-6-2-0.rt1.gen.ch.geant2.net (NO DATA)
14 so-2-0-0.rt1.mil.it.geant2.net (NO DATA)
15 garr-gw.rt1.mil.it.geant2.net (NO DATA)
16 rt1-mi1-rt-mi2.mi2.garr.net



17 rt-mi2-rt-rm2.rm2.garr.net (GRAPH OMITTED)
18 rt-rm2-rc-fra.fra.garr.net (GRAPH OMITTED)
19 rc-fra-ru-lnf.fra.garr.net (GRAPH OMITTED)



20
21 www6.lnf.infn.it (193.206.84.223) 189.908 ms 189.596 ms 189.684 ms

(GARR was s front-runner in deploying perfSONAR)

ESnet PerfSONAR Deployment Activities

- ESnet is deploying OWAMP and BWCTL servers next to all backbone routers, and at all 10Gb connected sites
  - 31 locations deployed
  - Full list of active services at:
    - http://www.perfsonar.net/activeServices/
- Instructions on using these services for network troubleshooting:
  - http://fasterdata.es.net

➢ ***These services have already been extremely useful to help debug a number of problems***

  - ***perfSONAR is designed to federate information from multiple domains***
  - ***provides the only tool that we have to monitor circuits end-to-end across the networks from the US to Europe***

- PerfSONAR measurement points are deployed at dozens of R&E institutions in the US and more in Europe
  - See https://dc211.internet2.edu/cgi-bin/perfAdmin/serviceList.cgi

➢ ***The value of perfSONAR increases as it is deployed at more sites***

- *What Does the Network Situation Look Like Now?*

ESnet Status as of 12/2008

- ESnet is set to provide bandwidth and connectivity adequate for all know uses of the network, including the LHC, for the next several years
  - There is adequate capacity in the metro area networks that connect the LHC Tier1 Data Centers to get LHC data to the core network
  - There is adequate capacity in all national core paths
  - There is full redundancy of connections to the Tier 1 centers
  - There is adequate capacity and redundancy in the connections to the US R&E networks serving the university community in order to get data to the Tier 2 and 3 sites at the maximum rates that have been observer (which is substantially higher than the HEP planning documents indicate)
  - There is adequate capacity and redundancy in the connections to the international R&E networks serving traffic to and from the European Tier 1 and Tier 2 centers and visa versa (this is apart from the LHCOPN Tier 0 to Tier 1 capacity provided by USLHCNet)
  - There is a functioning and capable virtual circuit service providing guaranteed bandwidth (primarily from the US Tier 1 to Tier 2 centers, but also from US Tier 1 to European Tier 2 centers)

What Does the Situation Look Like Now?
Re-evaluating the Strategy and Identifying Issues

- The current strategy (that lead to the ESnet4, 2012 plans) was developed primarily as a result of the information gathered in the 2003 and 2004 network workshops, and their updates in 2005-6 (including LHC, climate simulation, RHIC (heavy ion accelerator), SNS (neutron source), magnetic fusion, the supercomputers, and a few others) [workshops]
- So far the more formal requirements workshops have largely reaffirmed the ESnet4 strategy developed earlier
- ***However – is this the whole story*? (No)***

(* Details may be found in "The Evolution of Research and Education Networks and their Essential Role in Modern Science." November, 2008. To be published in *Trends in High Performance & Large Scale Computing,* Lucio Gandinetti and Gerhard Joubert editors. Available at http://www.es.net/pub/esnet-doc/index.html)

Is ESnet Planned Capacity Adequate? E.g. for LHC and climate? (Maybe So, Maybe Not) – Must undertake continuous reexamination of the long-term requirements because they frequently change

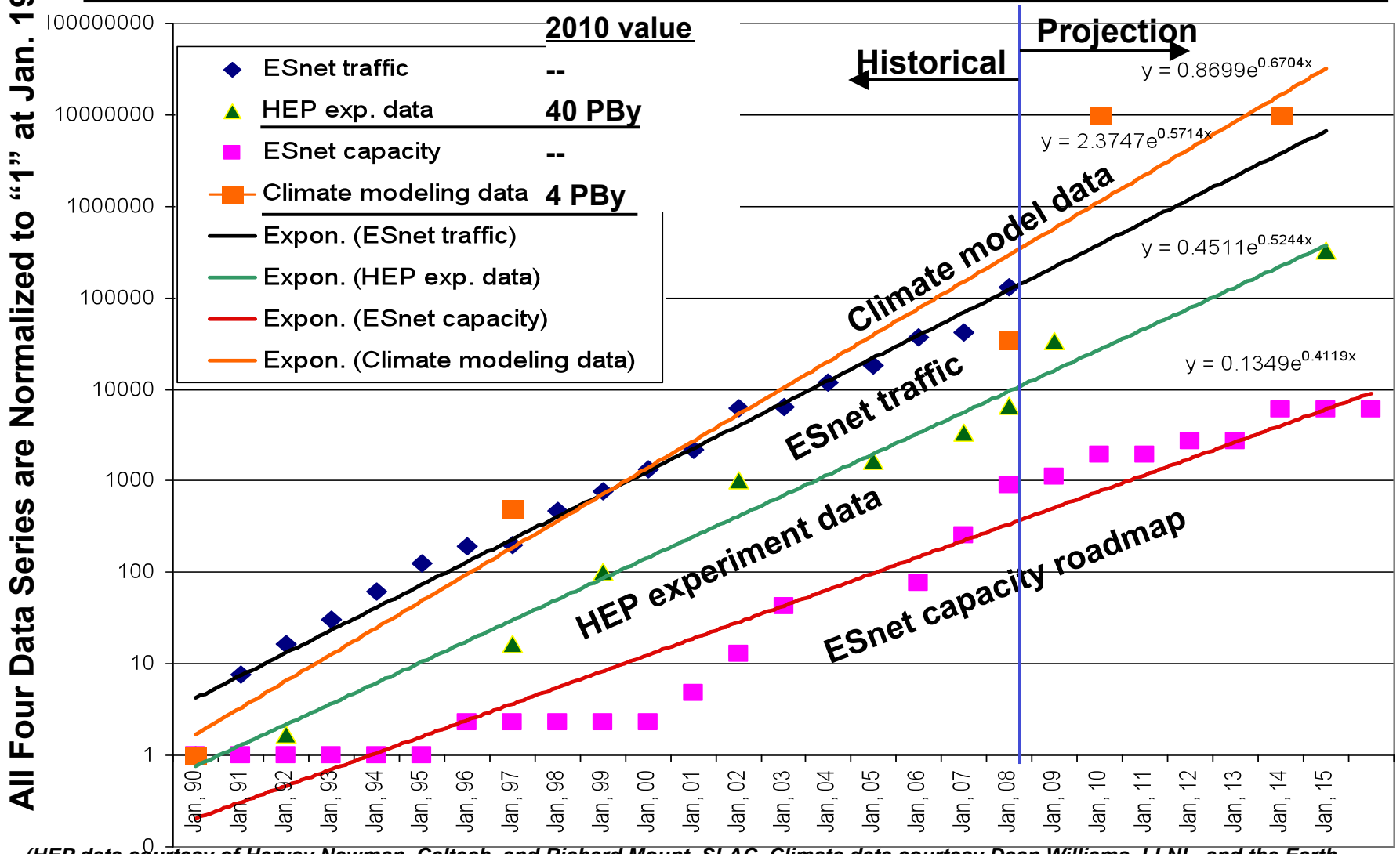
- Several Tier2 centers (mostly at Universities) are capable of 10Gbps now
  - Many Tier2 sites are building their local infrastructure to handle 10Gbps
  - We won't know for sure what the "real" load will look like until the testing stops and the production analysis begins
  - ➢ Scientific productivity will follow high-bandwidth access to large data volumes ⇒ incentive for others to upgrade
- Many Tier3 sites are also building 10Gbps-capable analysis infrastructures – this was not in LHC plans a year ago
  - Most Tier3 sites do not yet have 10Gbps of network capacity
  - It is likely that this will cause a "second onslaught" in 2009 as the Tier3 sites all upgrade their network capacity to handle 10Gbps of LHC traffic
- ➢ ***It is possible that the USA installed base of LHC analysis hardware will consume significantly more network bandwidth than was originally estimated***
  - N.B. Harvey Newman (HEP, Caltech) predicted this eventuality several years ago
- The needs of the climate modeling community are just emerging (and were not predicted in the requirements studies) and based on data set size are likely to equal those of the LHC
- ITER is not accounted for at all

Predicting the Future

- How might we “predict” the future without relying on the practitioner estimates given in the requirements workshops?
- Consider what we know – not just about historical traffic patterns, but also look at data set size growth
  - The size of data sets produced by the science community has been a good indicator of the network traffic that was generated
    - The larger the experiment / science community the more people that are involved at diverse locations and the more that data must move between them

Network Traffic, Science Data, and Network Capacity

Ignore the units of the quantities being graphed they are normalized to 1 in 1990, just look at the long-term trends: ***All of the "ground truth" measures are growing significantly faster than ESnet projected capacity based on stated requirements***



(HEP data courtesy of Harvey Newman, Caltech, and Richard Mount, SLAC. Climate data courtesy Dean Williams, LLNL, and the Earth Systems Grid Development Team.)

Issues for the Future Network

- The significantly higher exponential growth of science dataset size vs. total capacity (aggregate core bandwidth) means traffic will eventually overwhelm the capacity – "when" cannot be directly deduced from aggregate observations, but if you add this fact
  - Nominal average load on busiest backbone paths in June 2006 was ~1.5 Gb/s - In 2010 average load will be ~15 Gbps based on current trends and 150 Gb/s in 2014

My (wej) guess is that capacity problems will develop by 2015-16 without new technology approaches

Where Do We Go From Here?

- It seems clear that ESnet in the future will have to have both
  - capacity well beyond the 2004-6 projections, and
  - the ability to more flexibly map traffic to waves (traffic engineering in order to make optimum use of the available capacity)
- To obtain more capacity ESnet will have to go to 100Gb/s waves as there is not enough wave capacity to satisfy newly projected needs by just adding more 10Gb/s waves on the current fiber and it does not appear feasible to obtain a second national fiber footprint

- *What is the Path Forward?*

1) Optimize the use of the existing infrastructure

Dynamic Wave / Optical Circuit Management:

- The current path/wave/optical circuit topology is rich in redundancy
- The ***current wave transport topology is essentially static*** or only manually configured - our current network infrastructure of routers and switches assumes this
- With completely flexible traffic management extending down to the optical transport level we ***should be able to extend the life of the current infrastructure*** by moving significant parts of the capacity to the specific routes where it is needed

➢ We must integrate the optical transport with the “network” and provide for dynamism / route flexibility at the optical level in order to make optimum use of the available capacity

2) 100Gb/s Waves

- ESnet is actively involved in the development and deployment of 100Gb/s per channel optical transport equipment and 100Gb/s routing equipment
  - ESnet has received special funding (almost $US 60M!) to build a national 100G/wave testbed
  - The testbed will connect (at least) the three Office of Science supercomputers involved in climate modeling (Argonne – near Chicago, IL; Oak Ridge – east of Nashville, Tennessee; NERSC – Berkeley, CA)
    - Two other major players in US climate modeling are Lwarence Livermore Lab – east of Berkeley, CA, and NCAR (National Center for Atmospheric Research, Boulder, Colorado) may be added later if the initial testbed is successful in driving 100G component cost down
  - See Steve Cotter's talk ("ESnet's Approach to Enabling Virtual Science") in session 2A – "Support infrastructure – 'All change – introducing GN3 and ESNET4'"

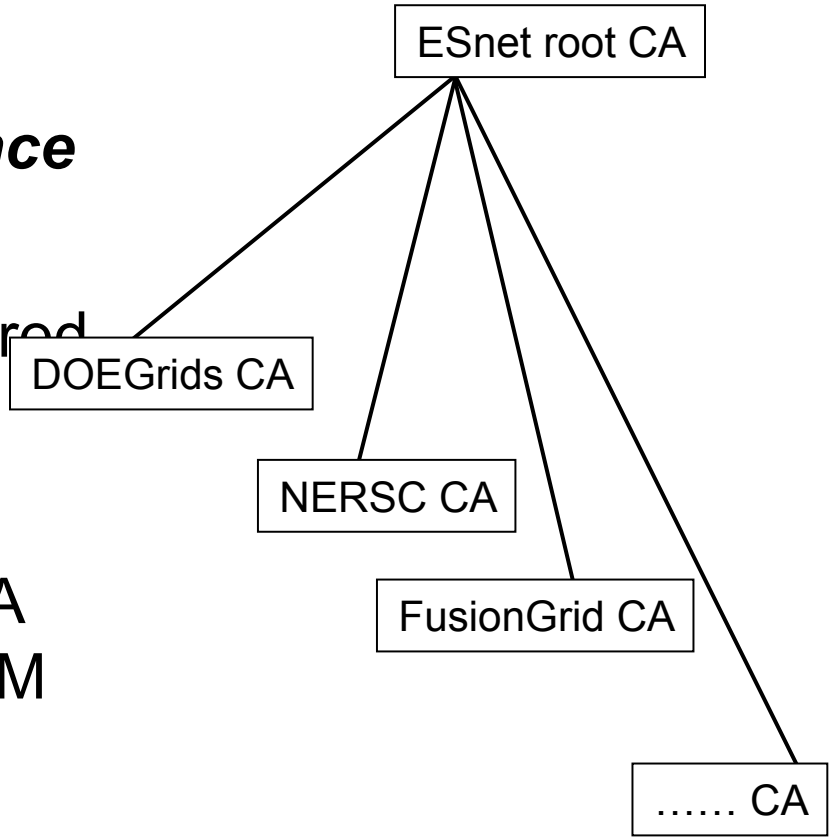
- *Science Support / Collaboration Services*

Federated Trust Services – Support for Large-Scale Collaboration

- Remote, multi-institutional, identity authentication is critical for distributed, collaborative science in order to permit sharing widely distributed computing and data resources, and other Grid services
- Public Key Infrastructure (PKI) is used to formalize the existing web of trust within science collaborations and to extend that trust into cyber space
  - The function, form, and policy of the ESnet trust services are driven entirely by the requirements of the science community and by direct input from the science community
- International scope trust agreements that encompass many organizations are crucial for large-scale collaborations
  - ESnet has lead in negotiating and managing the cross-site, cross-organization, and international trust relationships to provide policies that are tailored for collaborative science
    - ➢ This service, together with the associated ESnet PKI service, is the basis of the routine sharing of HEP Grid-based computing resources between US and Europe

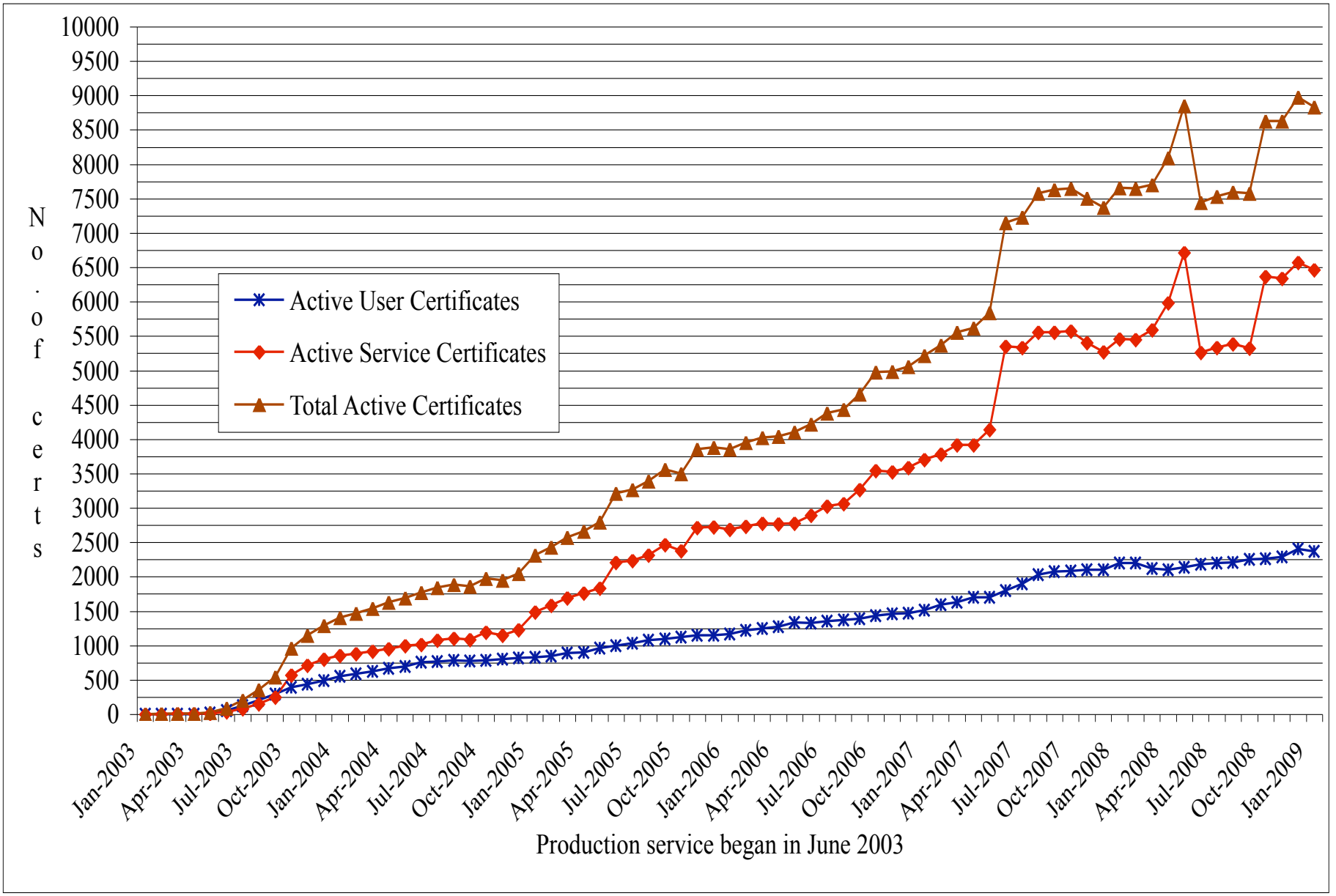
ESnet Public Key Infrastructure

- ***CAs are provided with different policies as required by the science community***
  - DOEGrids CA has a policy tailored to accommodate international science collaboration
  - NERSC CA policy integrates CA and certificate issuance with NIM (NERSC user accounts management services)
  - FusionGrid CA supports the FusionGrid roaming authentication and authorization services, providing complete key lifecycle management



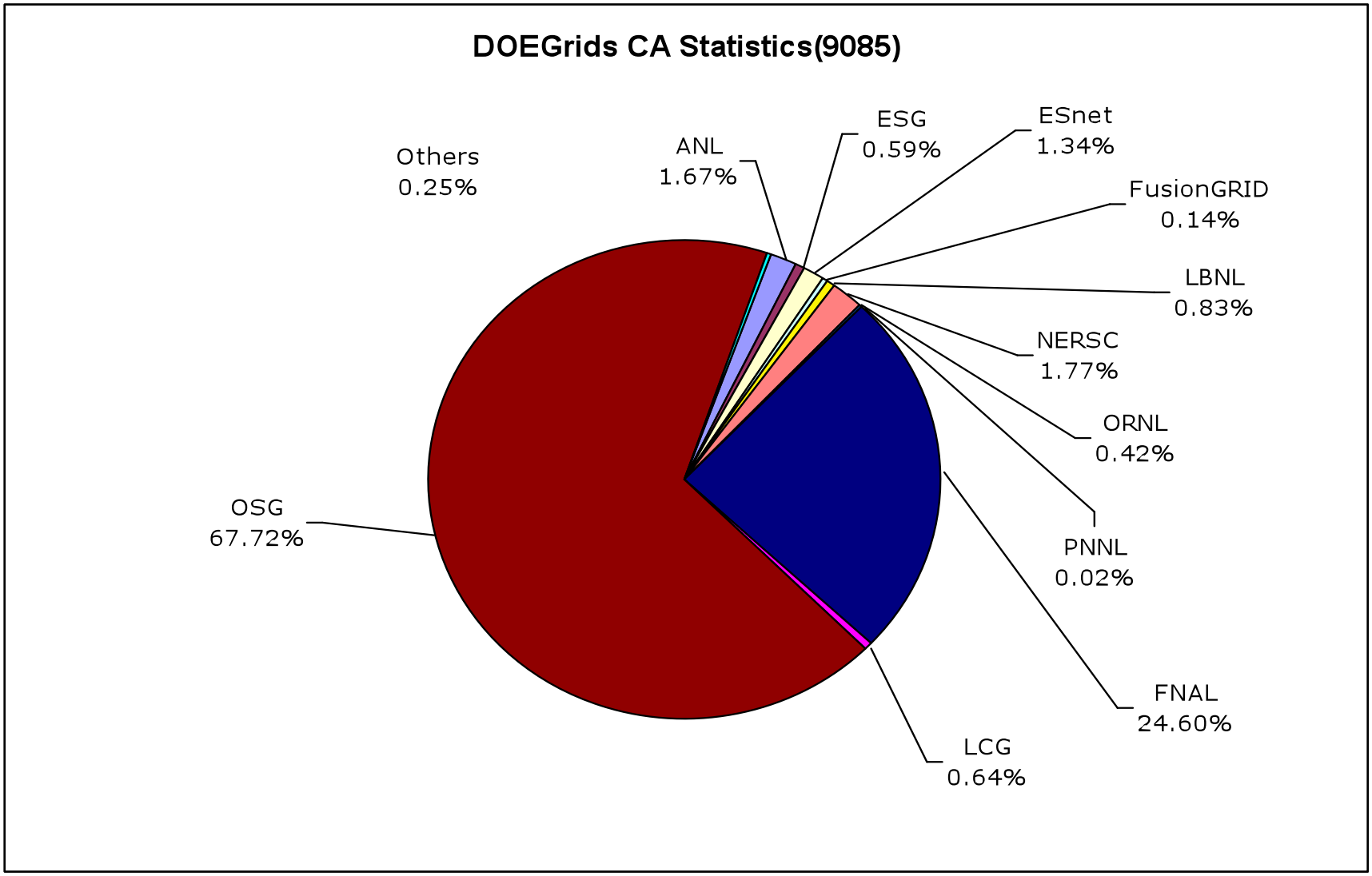
See www.doegrids.org

DOEGrids CA (Active Certificates) Usage Statistics



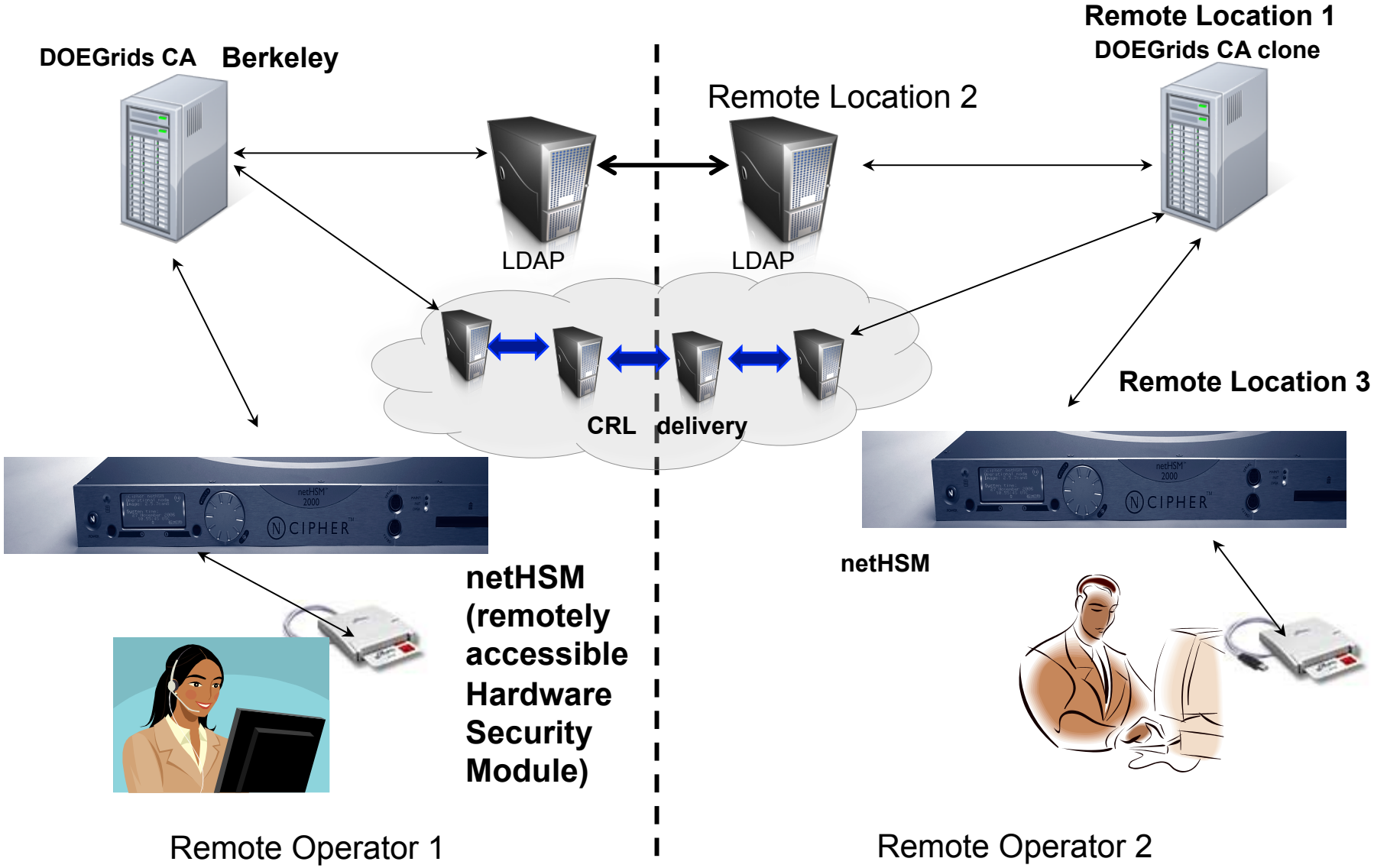
* Report as of Feb 29, 2009

DOEGrids CA Usage - Virtual Organization Breakdown



OSG Includes (BNL, CDF, CIGI,CMS, CompBioGrid, DES, DOSAR, DZero, Engage, Fermilab,GADU, geant4, GLOW, GPN, GRASE GUGrid, i2u2, ILC, JLAB, LIGO, mariachi, MIS, nanoHUB, NWICG, NYSGrid, OSG, OSGEDU, SBGrid, SLAC, STAR & USATLAS)

In development: DOEGrids CA with High Availability



OpenID

- What about new services?
  - Caveat emptor – Mike Helm has thought a lot about this, but does not have concrete plans yet – these slides were "invented" by WEJ
- OpenID does not provide any assurance of (human) identity…… neither does PKI

OpenID

- What DOEGrids CA provides is a community-driven model of “consistent level of assuredness of human identity associated with a cyber auth process” – to wit:
  - PMA (Policy Management Authority) sets the policy for the minimum “strength” of personal / human identity verification prior to issuing a certificate
  - Providing a level of identity assurance consistent with the requirements of a given science community (VO) is accomplished by certificate requests being vetted a VO-nominated Registration Agent (RA) who validates identity before issuing a cert.
  - Relying Parties (those services that require PKI certs in order to provide service) use Public Key Infrastructure to validate the cert-based identity that wa s vetted by the RA

OpenID

- ESnet might be an OpenID Service Provider based around, e.g., DOEGrids CA for communities that require some consistent level of assuredness of human identity associated with a cyber auth process
- DOEGrids would issue OpenID URL credentials based on DOGrids certs
  - A third-party could probably do the same thing by only issuing OpenID URL credentials based on DOGrids certs – these OpenID credentials would inherit their assuardness uniformity from DOEGrids CA
    - The “I (wej) assume” is due to the fact that WEJ does not know what machinery is involved in generating an OpenID credential – presumably the credential would have to cryptographically protected and validated by some mechanisms akin to PKI. The Service Provide would have to provide and operate these mechanisms

ESnet Conferencing Service (ECS)

- A highly successful ESnet Science Service that provides audio, video, and data teleconferencing service to support human collaboration of DOE science
  - Seamless voice, video, and data teleconferencing is important for geographically dispersed scientific collaborators
  - Provides the central scheduling essential for global collaborations
  - ESnet serves more than a thousand DOE researchers and collaborators worldwide
    - H.323 (IP) videoconferences (4000 port hours per month and rising)
    - audio conferencing (2500 port hours per month) (constant)
    - data conferencing (150 port hours per month)
    - Web-based, automated registration and scheduling for all of these services
  - Very cost effective (saves the Labs a lot of money)

➢ Conclusions

1) The US national and pan-European networks are in reasonably good shape for meeting requirements for the next TWO years.

2) To extend the current infrastructure to meet requirements through 2013-2015 requires research, development, and deployment in the areas of (at least)

   i. dynamic management of waves and integration of this with the layer 2 and 3 control planes;

   ii. 100G/wave transport technology;

   iii. transparent and dynamic re-routing of flows on the IP networks to the virtual circuit networks (SDN, DCN, etc.), and;

   iv. highly capable, "universally" deployed, end-to-end monitoring.

3) It is important to be looking at the technology for the next generation of network which must be designed and deployed by 2015-2017

References

[OSCARS] – "On-demand Secure Circuits and Advance Reservation System"
For more information contact Chin Guok (chin@es.net). Also see

http://www.es.net/oscars

[Workshops]
see http://www.es.net/hypertext/requirements.html

[LHC/CMS]
http://cmsdoc.cern.ch/cms/aprom/phedex/prod/Activity::RatePlots?view=global

[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/

[E2EMON] Geant2 E2E Monitoring System –developed and operated by JRA4/WI3, with implementation done at DFN
http://cnmdev.lrz-muenchen.de/e2e/html/G2_E2E_index.html
http://cnmdev.lrz-muenchen.de/e2e/lhc/G2_E2E_index.html

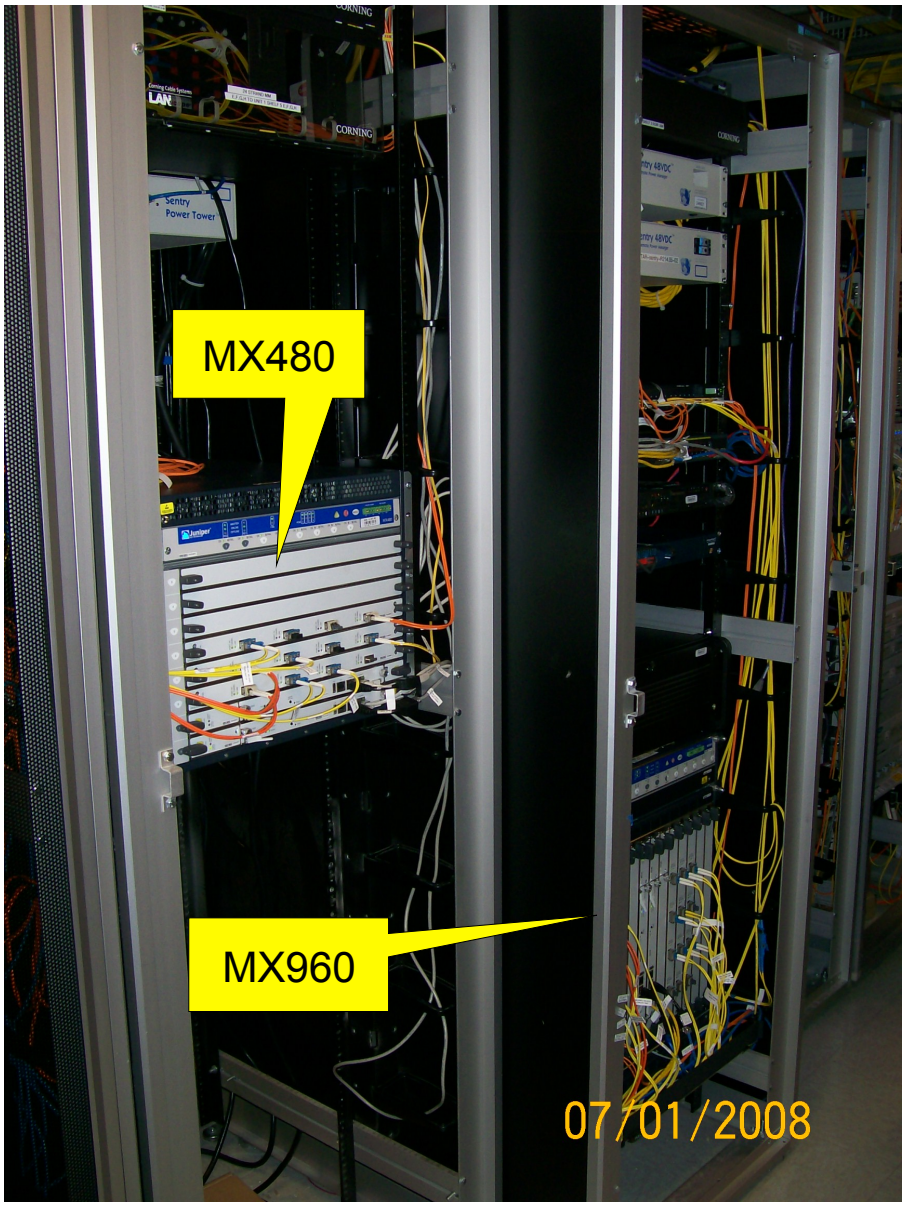
[TrViz] ESnet PerfSONAR Traceroute Visualizer
https://performance.es.net/cgi-bin/level0/perfsonar-trace.cgi

Additional Information

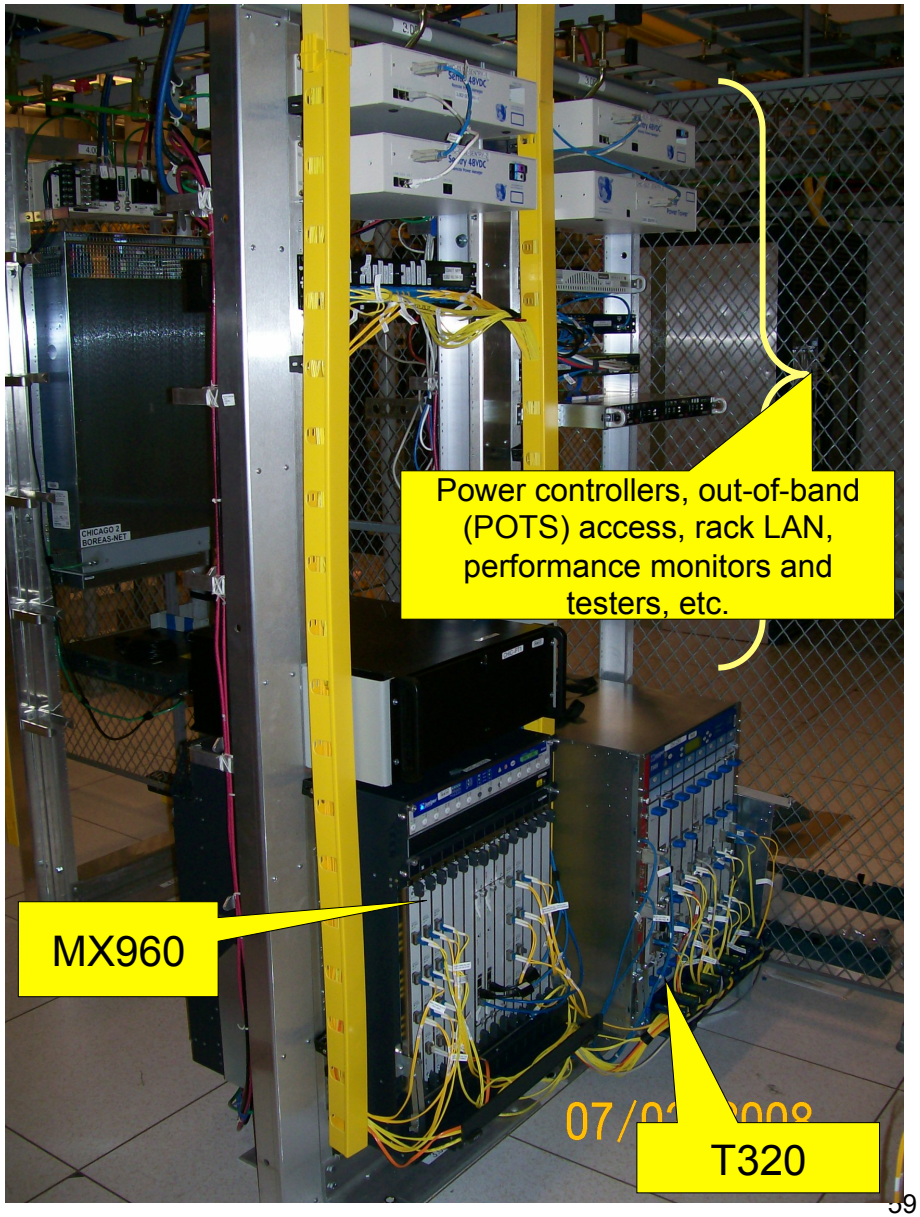
- *What is ESnet?*

ESnet4 Hubs are in Carrier or R&E Collocation Facilities

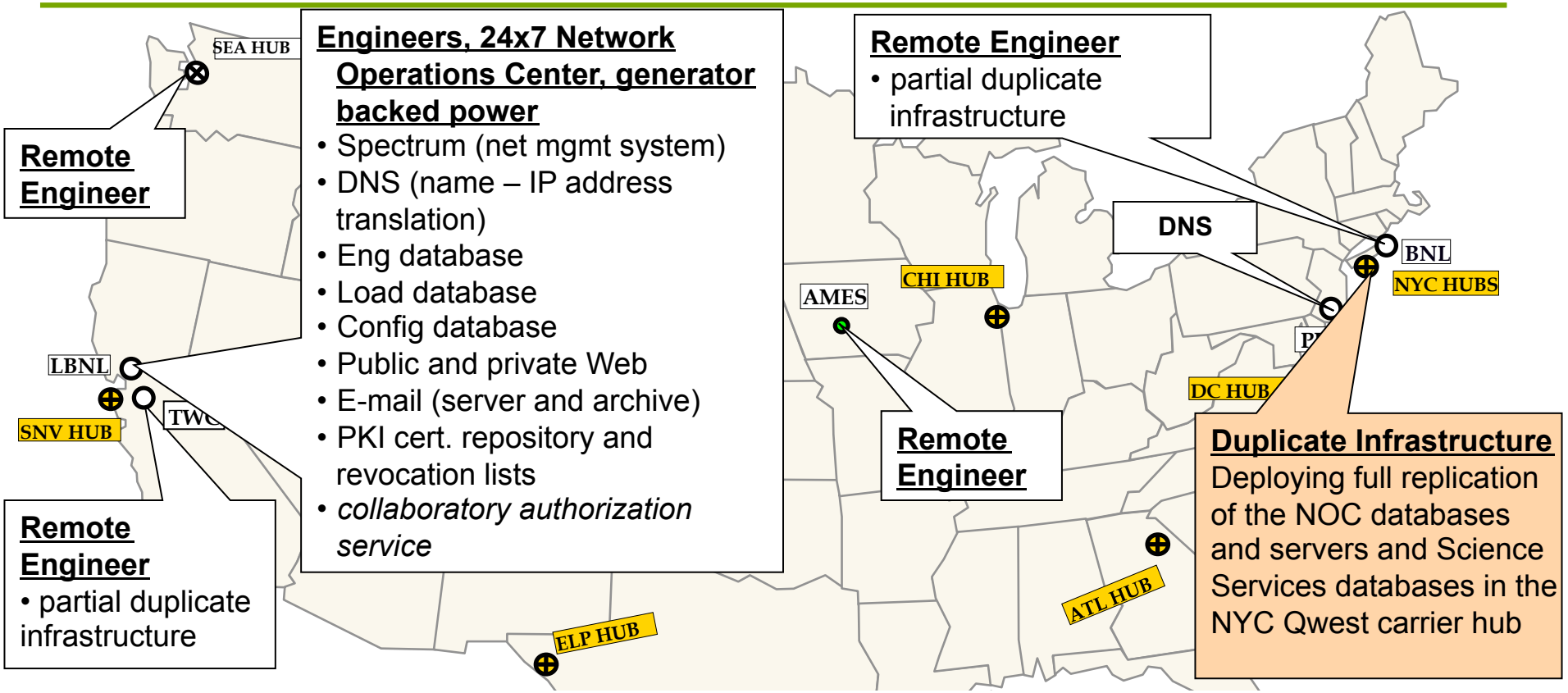
Starlight (Northwestern Univ., Chicago)



600 West Chicago (Level3 MondoCondo)



ESnet Provides Disaster Recovery and Stability



- The network must be kept available even if, e.g., the West Coast is disabled by a massive earthquake, etc.

Reliable operation of the network involves
- remote Network Operation Centers (4)
- replicated support infrastructure
- generator backed UPS power at all critical network and infrastructure locations

- high physical security for all equipment
- non-interruptible core - **ESnet core operated without interruption** through
  - N. Calif. Power blackout of 2000
  - the 9/11/2001 attacks, and
  - the Sept., 2003 NE States power blackout

➢ *The ESnet Planning Process*

Services Requirements from Instruments and Facilities

- Fairly consistent requirements are found across the large-scale sciences
- ***Large-scale science uses distributed systems*** in order to:
  - Couple existing pockets of code, data, and expertise into "systems of systems"
  - Break up the task of massive data analysis into elements that are physically located where the data, compute, and storage resources are located
- Such systems
  - are data intensive and high-performance, typically moving terabytes a day for months at a time
  - are high duty-cycle, operating most of the day for months at a time in order to meet the requirements for data movement
  - are widely distributed – typically spread over continental or inter-continental distances
  - depend on network performance and availability, but these characteristics cannot be taken for granted, even in well run networks, when the multi-domain network path is considered

➢ The system elements must be able to get guarantees from the network that there is adequate bandwidth to accomplish the task at hand

➢ The systems must be able to get information from the network that allows graceful failure and auto-recovery and adaptation to unexpected network conditions that are short of outright failure

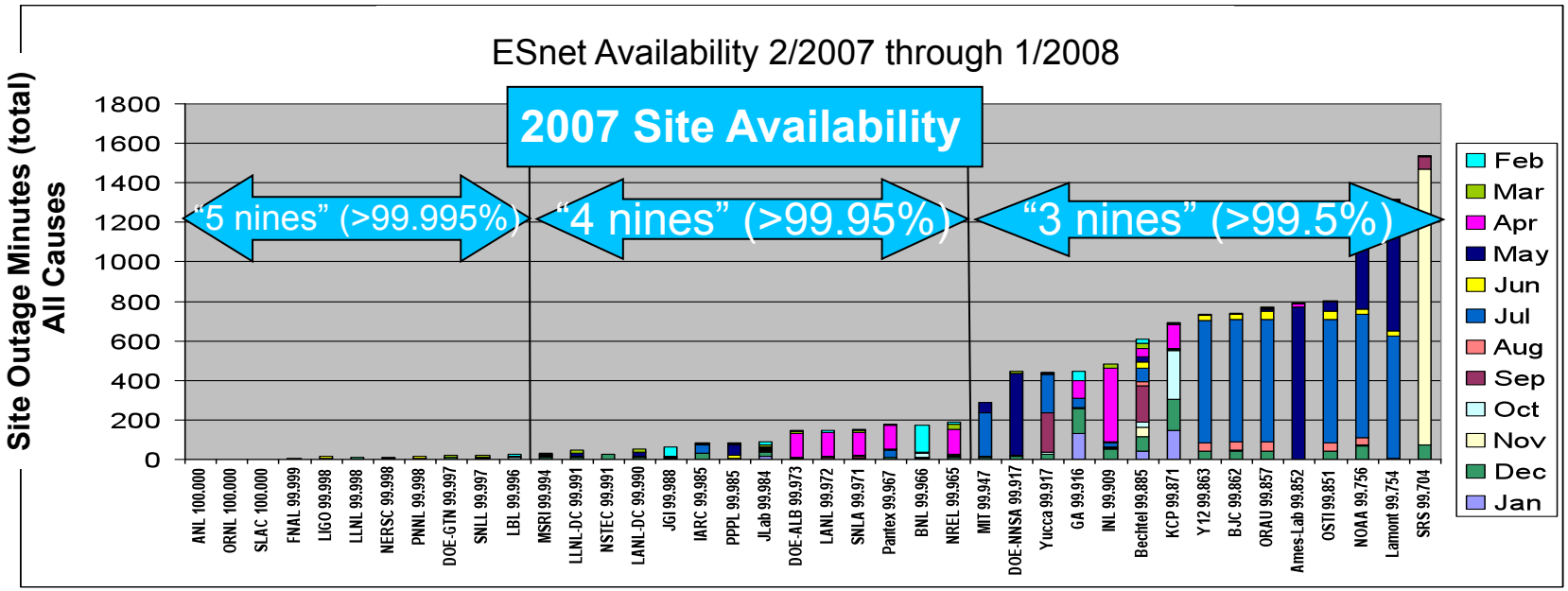
See, e.g., [ICFA SCIC]

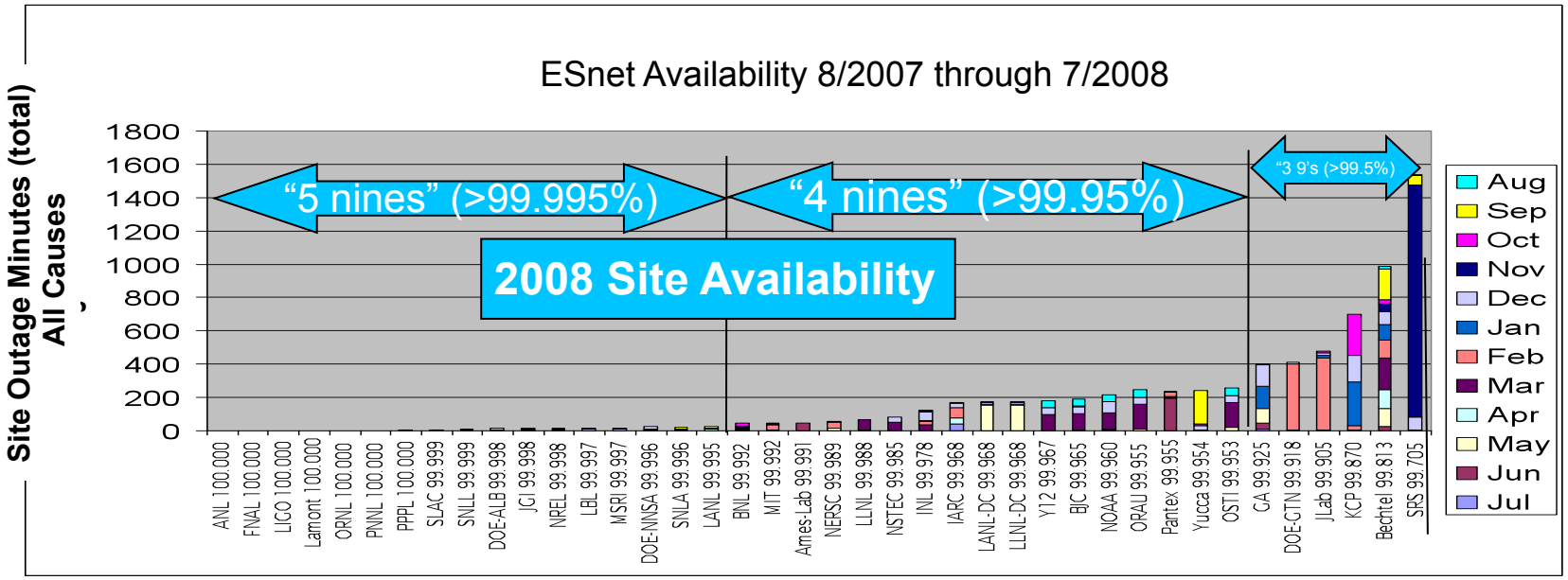
General Requirements from Instruments and Facilities

- ***Bandwidth – 200+ Gb/s core network by 2012***
  - Adequate network capacity to ensure timely movement of data produced by the facilities
- ***Reliability – 99.999% availability for large data centers***
  - High reliability is required for large instruments which now depend on the network to accomplish their science
- ***Connectivity – multiple 10Gb/s connections to US and international R&E networks (to reach the universities)***
  - Geographic reach sufficient to connect users and analysis systems to SC facilities
- Services
  - ***Commodity IP is no longer adequate – guarantees are needed***
    - Guaranteed bandwidth, traffic isolationA service delivery architecture compatible with Web Services / Grid / "Systems of Systems" application development paradigms
  - ***Visibility into the network end-to-end***
  - ***Science-driven authentication infrastructure (PKI)***
- ***Outreach to assist users in effective use of the network***

➢ *ESnet Response to the Requirements*

Reliability: One Consequence of ESnet's New Architecture is that Site Availability is Increasing





- *What Does the Network Situation Look Like Now?*

Where Are We Now?

How do the science program identified requirements compare to the network capacity planning?

Synopsis of "Science Network Requirements Aggregation Summary," 6/2008			
	5 year requirements	Accounted for in current ESnet path planning	Unacc'ted for
Requirements (aggregate Gb/s)	789	405	384

- The current network is built to accommodate the known, path-specific needs of the programs
- However this is not the whole picture: The core path capacity planning (see load-annotated map above) so far only accounts for 405 Gb/s out of 789 Gb/s identified aggregate requirements provided by the science programs
- The planned aggregate capacity growth of ESnet matches the know requirements, at least for the next several years ("aggregate capacity" is a measure based on total capacity of 13 "reference" links)

ESnet Planned Aggregate Capacity (Gb/s) Based on 5 yr. Budget								
	2006	2007	2008	2009	2010	2011	2012	2013
ESnet "aggregate"	57.50	192	192	842	1442	1442	1442	2042

  - The "extra" capacity indicated above is needed to account for the fact that there is much less than complete flexibility in mapping specific path requirements to the aggregate capacity planned network and we won't know specific paths until the science data models are finalized and implemented
- Whether this approach works is TBD, but indications are that it probably will