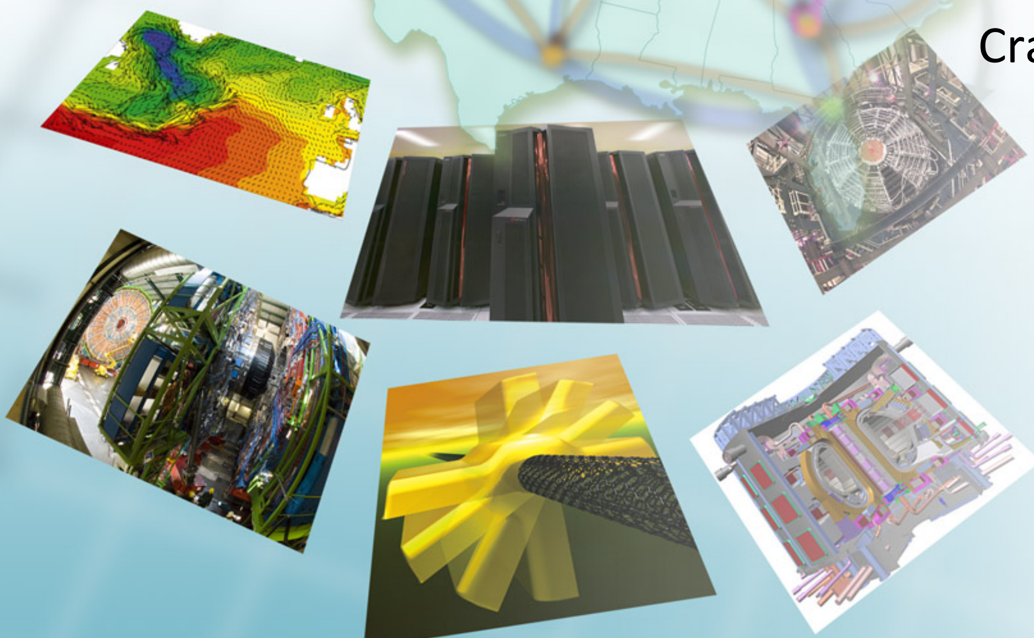
Progress in Integrating Networks with Service Oriented Architectures / Grids

The Evolution of ESnet's Guaranteed Bandwidth Service



Cracow ’09 Grid Workshop
Oct 12, 2009

William E. Johnston, Senior Scientist
Energy Sciences Network
Lawrence Berkeley National Lab

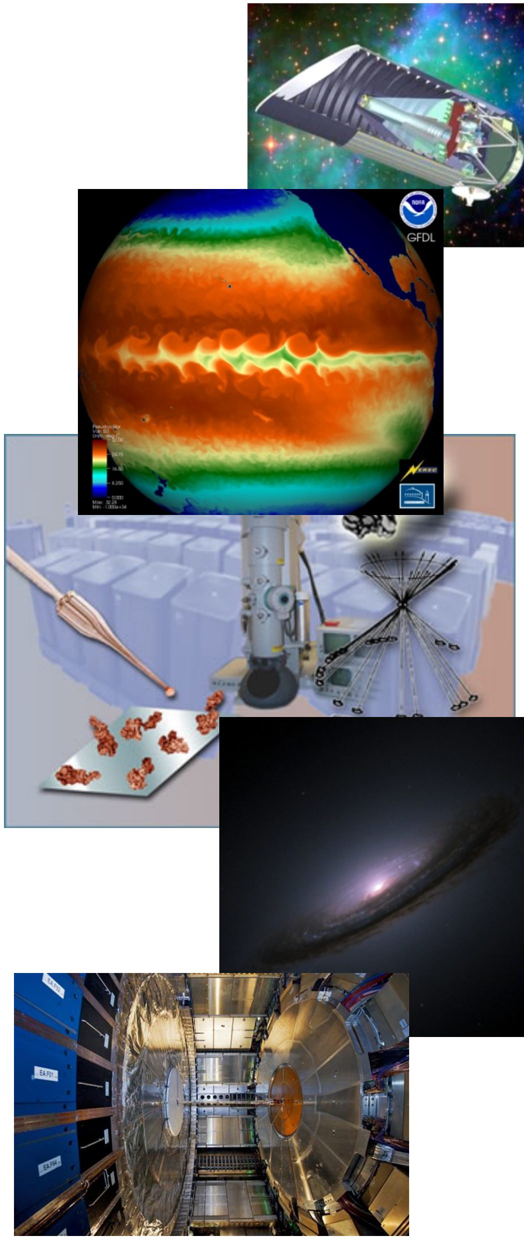
Networking for the Future of Science

DOE Office of Science and ESnet – the ESnet Mission

- The US Department of Energy's 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

DOE Office of Science and ESnet – the ESnet Mission

- **ESnet - the Energy Sciences Network - is an SC program whose <u>primary mission is to enable the large-scale science</u> 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 <u>SC/ASCAR funds ESnet to provide high-speed networking</u> and various collaboration services to Office of Science laboratories
  - ESnet servers most of the rest of DOE as well, on a cost-recovery basis

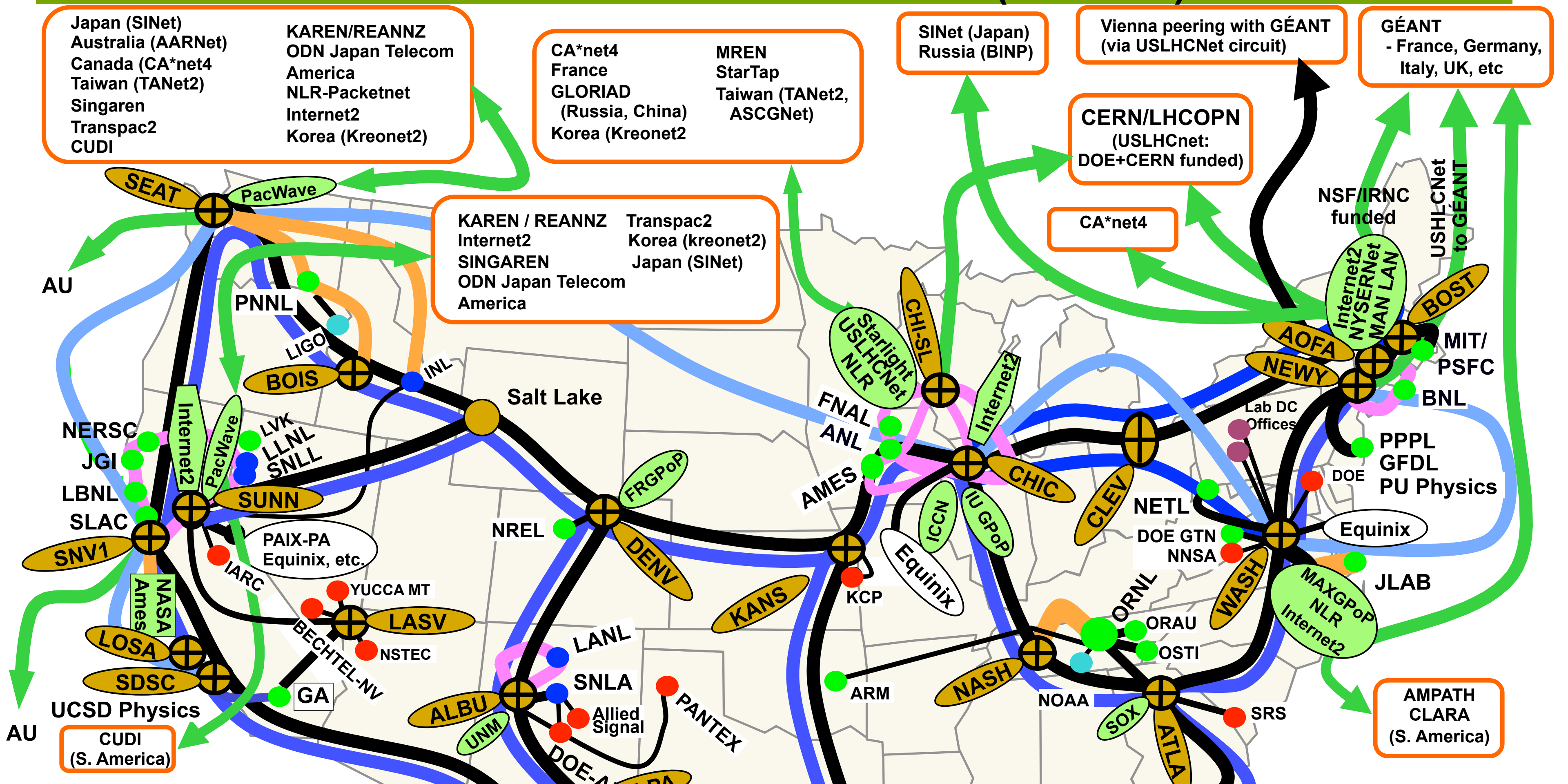


- *What is ESnet?*

ESnet Defined

- A national optical circuit infrastructure
  - ESnet shares an optical network with Internet2 (US national research and education (R&E) network) on a dedicated national fiber infrastructure
    - ESnet has exclusive use of a group of 10Gb/s optical channels on this infrastructure
  - ESnet has two core networks – IP and SDN – that are built on more than 100 x 10Gb/s WAN circuits
- A large-scale IP network
  - A tier 1 Internet Service Provider (ISP) (direct connections with all major commercial networks providers)
- 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, collaborative 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 circuits and related, remainder is staff and equipment related

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



~45 end user sites

- Office Of Science Sponsored (22)
- NNSA Sponsored (13+)
- Joint Sponsored (4)
- Other Sponsored (NSF LIGO, NOAA)
- Laboratory Sponsored (6)

commercial peering points

SUNN ESnet core hubs

R&E networks Specific R&E network peers

Other R&E peering points

➢Much of the utility (and complexity) of ESnet is in its high degree of interconnectedness

Geography is only representational

International (10 Gb/s)
10-20-30 Gb/s SDN core (I2, NLR)
10Gb/s IP core
MAN rings (10 Gb/s)
Lab supplied links
OC12 / GigEthernet
OC3 (155 Mb/s)
45 Mb/s and less

- *The ESnet Planning Process*

How ESnet Determines its Network Architecture, Services, and Bandwidth

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

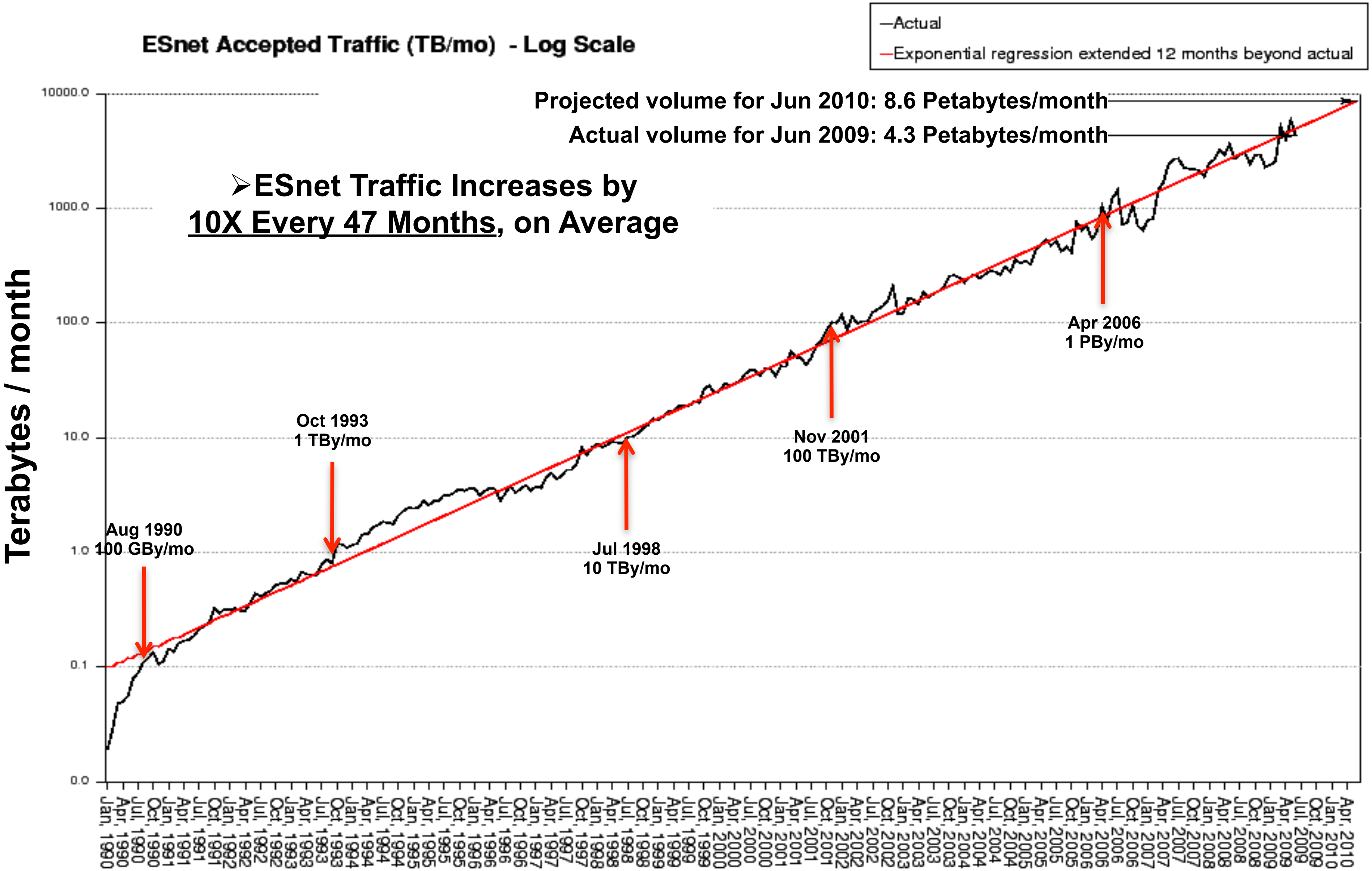
- What data will be generated by instruments and supercomputers coming on-line over the next 5-10 years?

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?

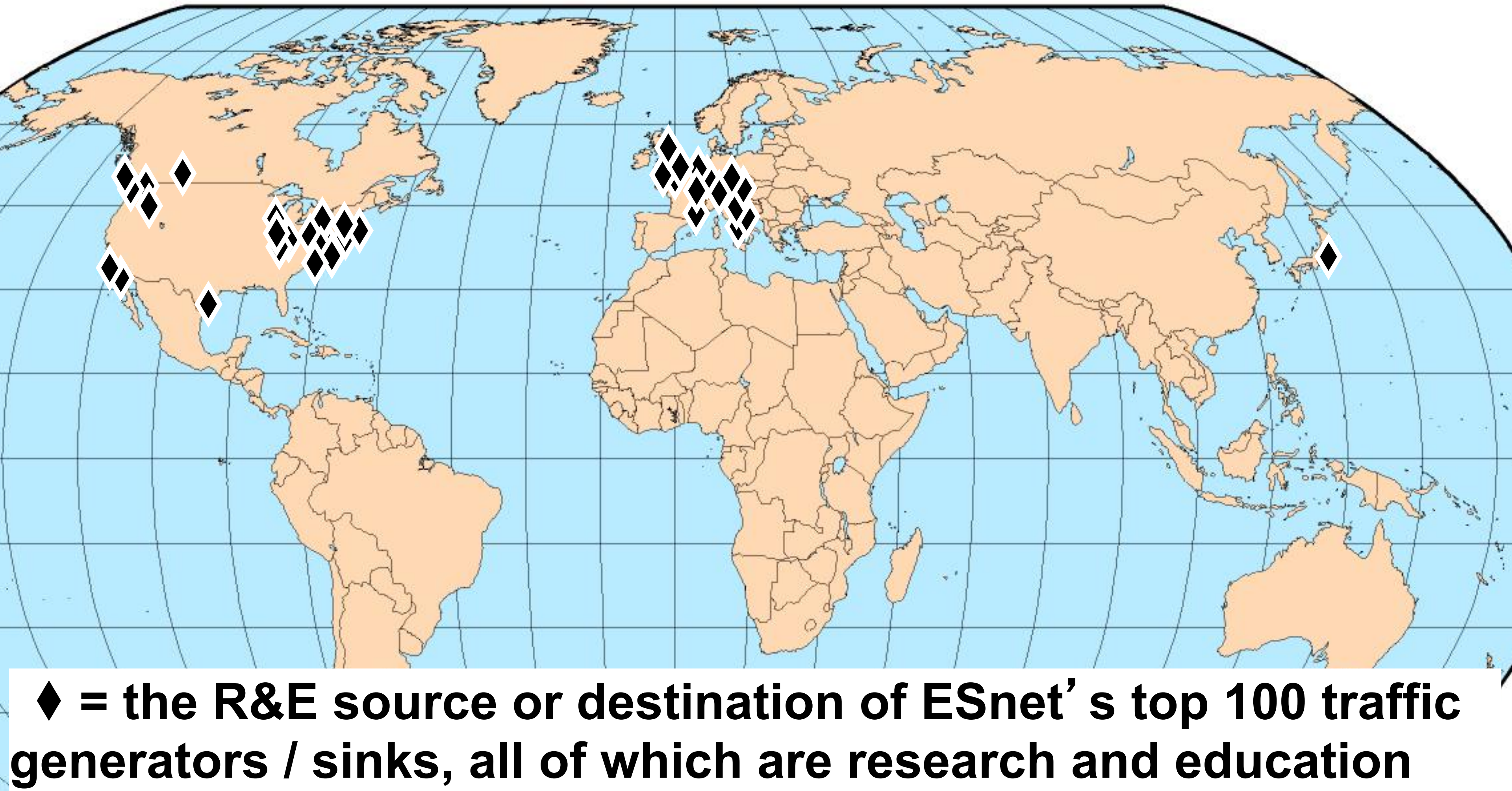
➢Observation: Current and Historical ESnet Traffic Patterns

Current and Historical ESnet Traffic Patterns



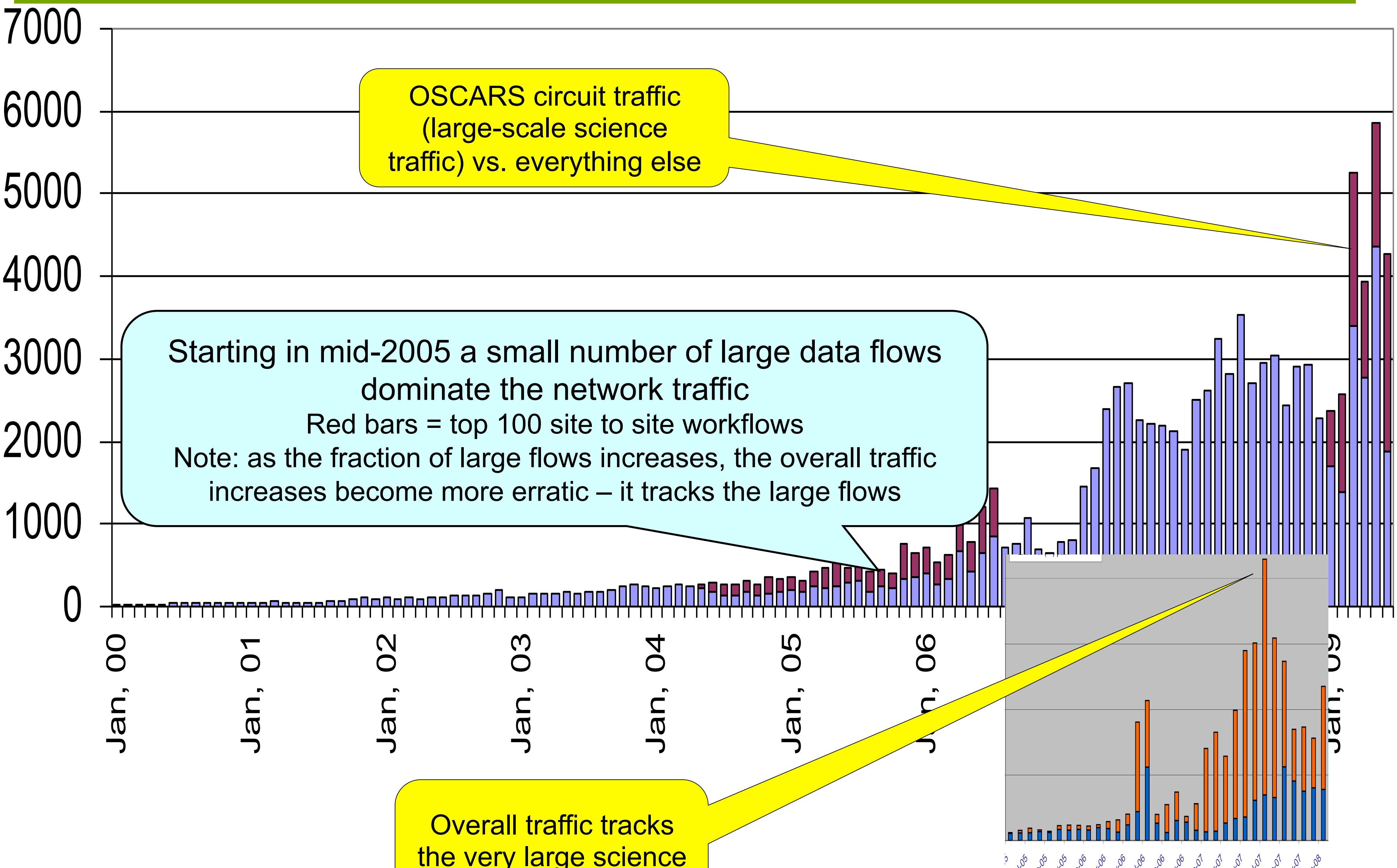
Log Plot of ESnet Monthly Accepted Traffic, January 1990 – June 2009

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)

Observing the Network: A small number of large data flows now dominate the network traffic – this motivates virtual circuits as a key network service

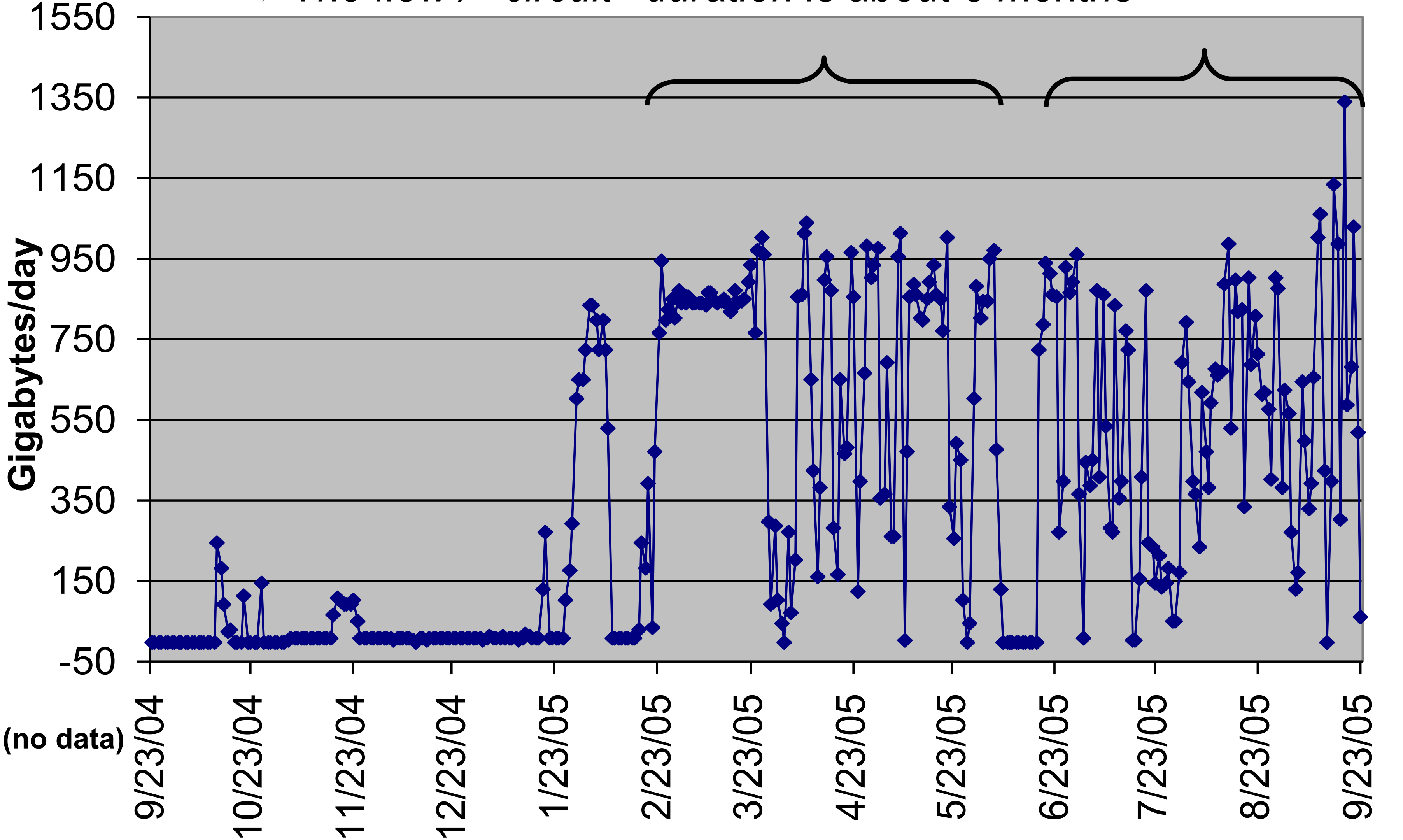


FNAL (LHC Tier 1 site) Outbound Traffic
(courtesy Phil DeMar, Fermilab)

Observing the Network: 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*



Services Requirements from Instruments and Facilities

Fairly consistent requirements are found across the large-scale sciences

- ***Large-scale science uses distributed applications 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 distributed application 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

Services Requirements from Instruments and Facilities (cont.)

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

- The distributed applications 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

- These services must be accessible within the Web Services / Grid Services paradigm of the distributed applications systems

See, e.g., [ICFA SCIC]

- *ESnet Response to the Requirements*

ESnet4 - The Response to the Requirements

I) A new network architecture and implementation strategy

- Provide two networks: IP and circuit-oriented Science Data Network
  - IP network for commodity flows
  - SDN network for large science data flows
  - Logical parity between the networks so that either one can handle both traffic types
- 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) Develop and deploy 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

III) Develop and deploy service-oriented, user accessible network monitoring systems

IV) Provide "consulting" on system / application network performance tuning

Response Strategy II) A Service-Oriented Virtual Circuit Service

Multi-Domain Virtual Circuits as a Service – Service Requirements

- Guaranteed, reservable bandwidth with resiliency
  - User specified bandwidth and time slot
  - Explicit backup paths can be requested
  - Paths may be either layer 3 (IP) or layer 2 (Ethernet) transport
- Requested and managed in a Web Services framework
- Traffic isolation
  - Allows for high-performance, non-standard transport mechanisms that cannot co-exist with commodity TCP-based transport
- End-to-end, cross-domain connections between Labs and collaborating institutions in other networks
- 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
  - If the sites trust the circuit service model of all of the involved networks (which, in practice, is the same as that of ESnet) then the circuits do not have to transit the site firewall
- 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.

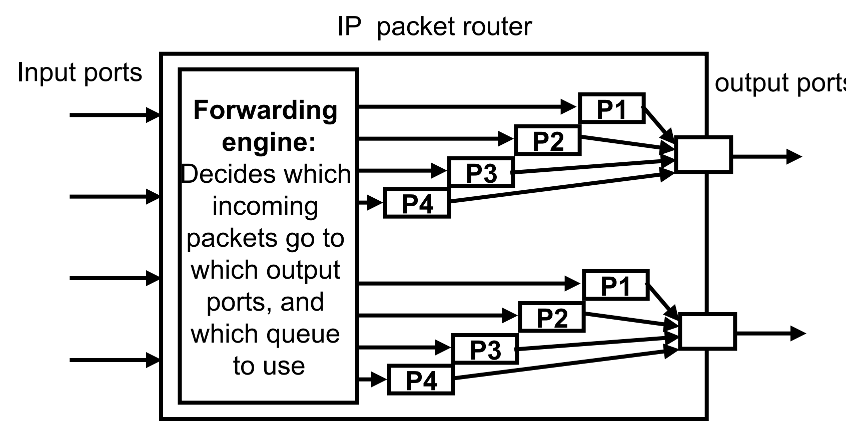
What are the “Tools” Available to Meet the Requirements?

- Ultimately, basic network services depend on the capabilities of the underlying routing and switching equipment.
  - Some functionality can be emulated in software and some cannot. In general, any capability that requires per-packet action will almost certainly have to be accomplished in the routers and switches.

T1) Providing guaranteed bandwidth to some applications and not others is typically accomplished by preferential queuing

- Most IP routers have multiple queues, but only a small number of them – four is typical:

- P1 – highest priority, typically only used for router control traffic
- P2 – elevated priority; typically not used in the type of “best effort” IP networks that make up most of the Internet
- P3 – standard traffic – that is, all ordinary IP traffic which competes equally with all other such traffic
- P4 – low priority traffic – sometimes used to implement a “scavenger” traffic class where packets move only when the network is otherwise idle



What are the “Tools” Available to Meet the Requirements?

T2) RSVP-TE – the Resource ReSerVation Protocol-Traffic Engineering – is used to define the virtual circuit (VC) path from user source to user destination

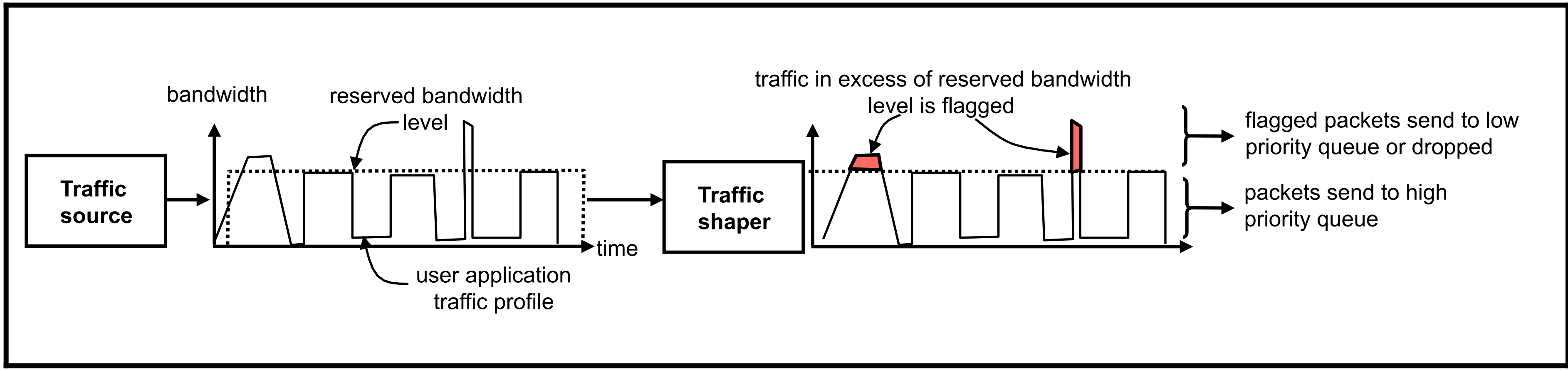
- Sets up a path through the network in the form of a forwarding mechanism based on encapsulation and labels rather than on IP addresses
  - Path setup is done with MPLS (Multi-Protocol Label Switching)
  - MPLS encapsulation can transport both IP packets and Ethernet frames
  - The RSVP control packets are IP packets and so the default IP routing that directs the RSVP packets through the network from source to destination establishes the default path
    - RSVP can be used to set up a specific path through the network that does not use the default routing (e.g. for diverse backup pahts)
- Sets up packet filters that identify and mark the user’s packets involved in a guaranteed bandwidth reservation
- When user packets enter the network and the reservation is active, packets that match the reservation specification (i.e. originate from the reservation source address) are marked for priority queuing

What are the “Tools” Available to Meet the Requirements?

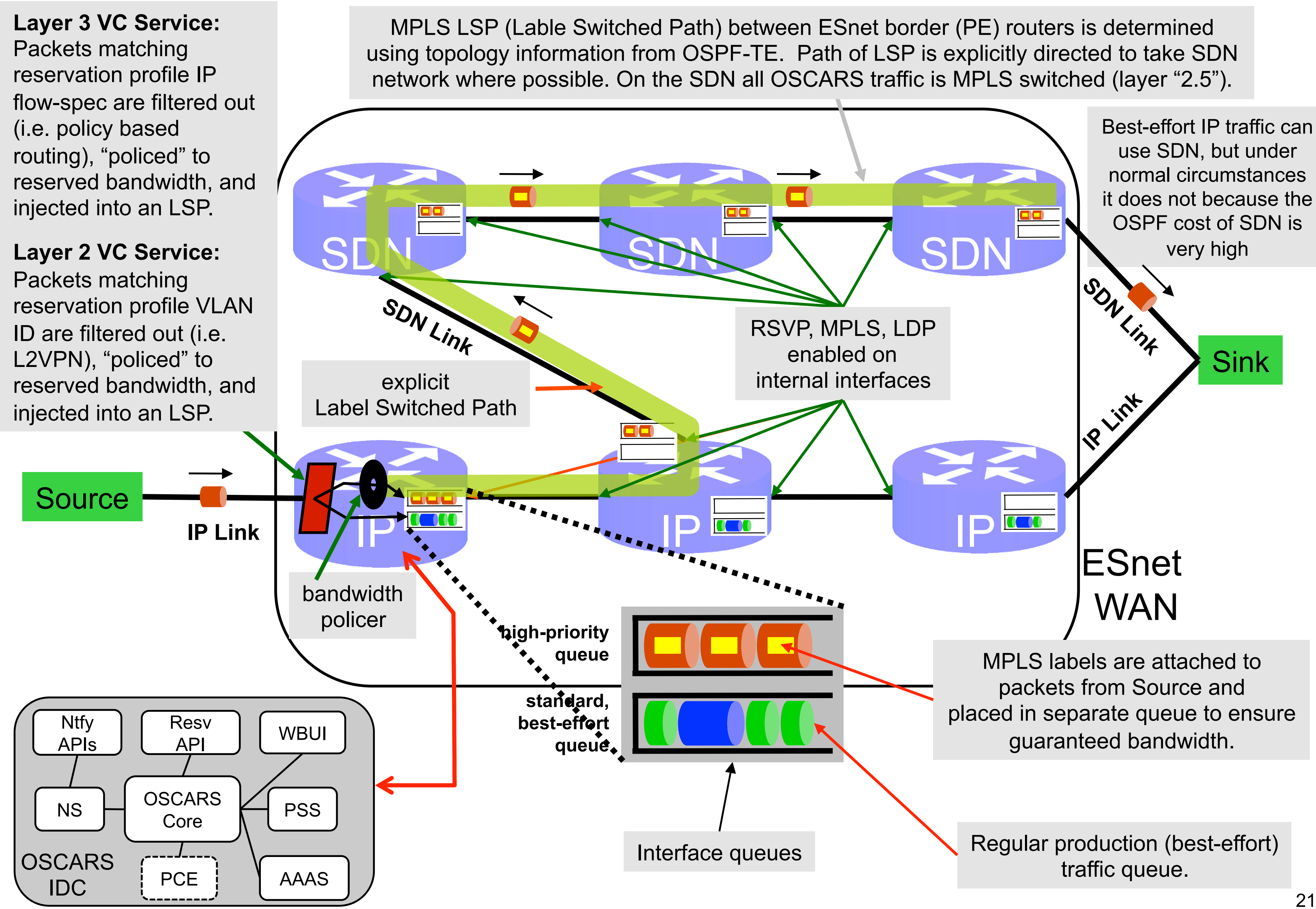
T3) Packet filtering based on address

- the “filter” mechanism in the routers along the path identifies (sorts out) the marked packets arriving from the reservation source and sends them to the high priority queue

T4) Traffic shaping allows network control over the priority bandwidth consumed by incoming traffic



Network Mechanisms Underlying OSCARS



OSCARS Approach

"OSCARS" – ESnet's InterDomain Controller

- Chin Guok (chin@es.net) and Evangelos Chaniotakis (haniotak@es.net)

The general approach of OSCARS is to

- Allow users to request guaranteed bandwidth between specific end points for specific period of time
  - User request is via SOAP or a Web browser interface
  - The assigned end-to-end path through the network is called a virtual circuit (VC)
- Manage available priority bandwidth to prevent over subscription
  - Each network link has an allocation of permitted high priority traffic depending on what else the link is used for
    - For example, a production IP link may historically have some fraction of the link that is always idle. Some fraction of this always idle bandwidth can be allocated to high priority traffic
  - Maintain a temporal network topology database that keeps track of the available and committed priority bandwidth along every link in the network to ensure that priority traffic stays within the link allocation
    - The database is temporal because it must account for all committed bandwidth over the lifetime of all reservations
    - Requests for priority bandwidth will be checked on every link of the end-to-end path over the entire lifetime of the request window
  - The request will only be granted if it can be accommodated within whatever fraction of the allocated bandwidth remains for high priority traffic after prior reservations are taken into account

OSCARS Approach

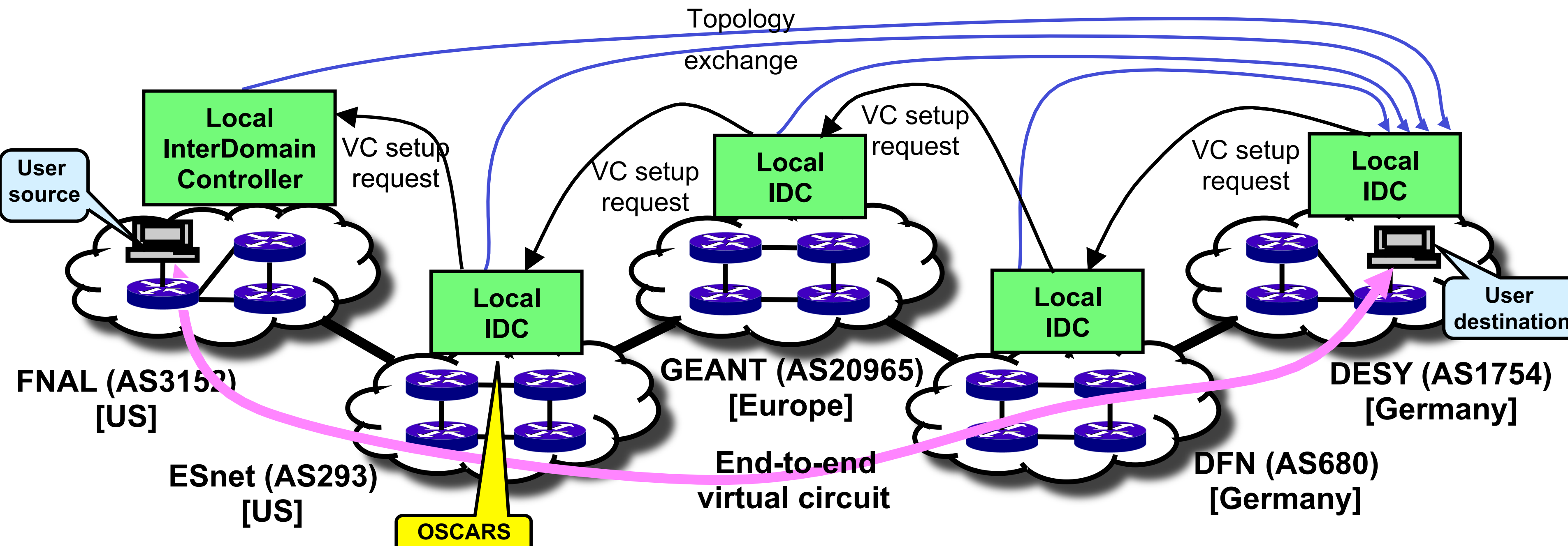
- If the reservation is granted, then at the start time of the reservation:
  - A "tunnel" (MPLS path) is established through the network on each router along the path of the VC using RSVP
    - The normal situation is that RSVP will set up the VC path along the default path as defined by IP routing.
    - User requested path constraints (e.g. that this VC not take the same physical path as its backup VC) are accommodated
  - Incoming packets from the reservation source are identified by using the router address filtering mechanism and "injected" into the MPLS tunnel
    - This provides a high degree of transparency for the user since at the start of the reservation all packets from the reservation source are automatically moved into a high priority path at the time of the reservation start
  - The incoming user packet stream is policed at the requested bandwidth in order to prevent oversubscription of the priority bandwidth

OSCARS Approach

- In the case of the user VC being IP based, when the reservation ends the packet filter stops marking the packets and any subsequent traffic from the same source is treated as ordinary IP traffic
- In the case of the user circuit being Ethernet based, the Ethernet circuit is torn down at the end of the reservation
- In both cases the temporal topology link loading database is automatically updated by virtue of the fact that this commitment no longer exists from this point forward
- This reserved bandwidth, virtual circuit is also called a “dynamic circuits” service

➢Environment of Science is Inherently Multi-Domain

- Inter-domain interoperability is crucial to serving science
- An effective international R&E collaboration (ESnet, Internet2, GÉANT, USLHCnet, several European NRENs, etc.) has standardized an inter-domain (inter-IDC) control protocol – "IDCP" – that requests inter-domain circuit setups
- 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 IDC protocol) is initiated at one end of the circuit and passed from domain to domain as the VC segments are authorized and reserved



Example – not all of the domains shown support the VC service

OSCARS Approach

- The ESnet circuit manager (OSCARS) can accept reservation requests from other Domain Controllers (IDC) as well as from users
- The IDCs exchange sufficient topology information to determine the egress and ingress points between domains
- The intra-domain circuits are "terminated" at the domain boundaries and then explicitely cross-connected to the circuit termination point in the domain where the path continues
  - This is so that the local domain can maintain complete control over the portion of the circuit that is within the local domain

➢OSCARS Virtual Circuit Security

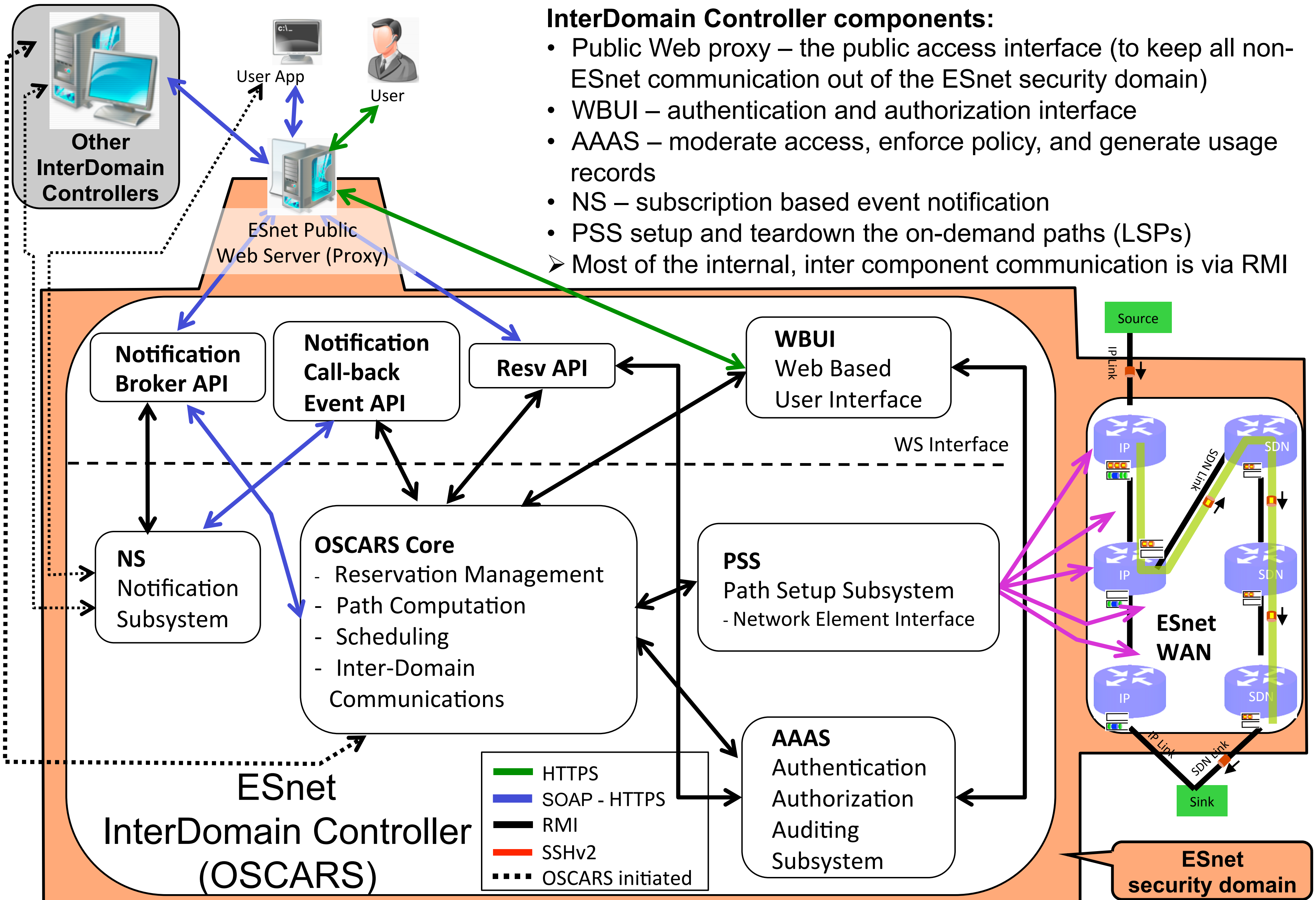
- Virtual circuit security is only guaranteed within the ESnet domain
- User VC transits ESnet as an MPLS path which is explicitly defined hop-by-hop
  - Integrity of the VC is thus a function of the ESnet router control plane integrity, which is closely guarded
- RSVP and MPLS are not enabled on ESnet edge routers
  - ESnet edge routers cannot accept RSVP packets from or send RSVP packets to non-ESnet nodes
  - External MPLS packets are discarded at the ESnet WAN border
- Inter-domain VCs are terminated at domain boundaries and regenerated for the intra-domain VC – that is, inter-domain circuits are piece-wise, with MPLS paths only within each domain

OSCARS Version 2 Service Implementation

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)

➢ Most of the internal, inter component communication is via RMI

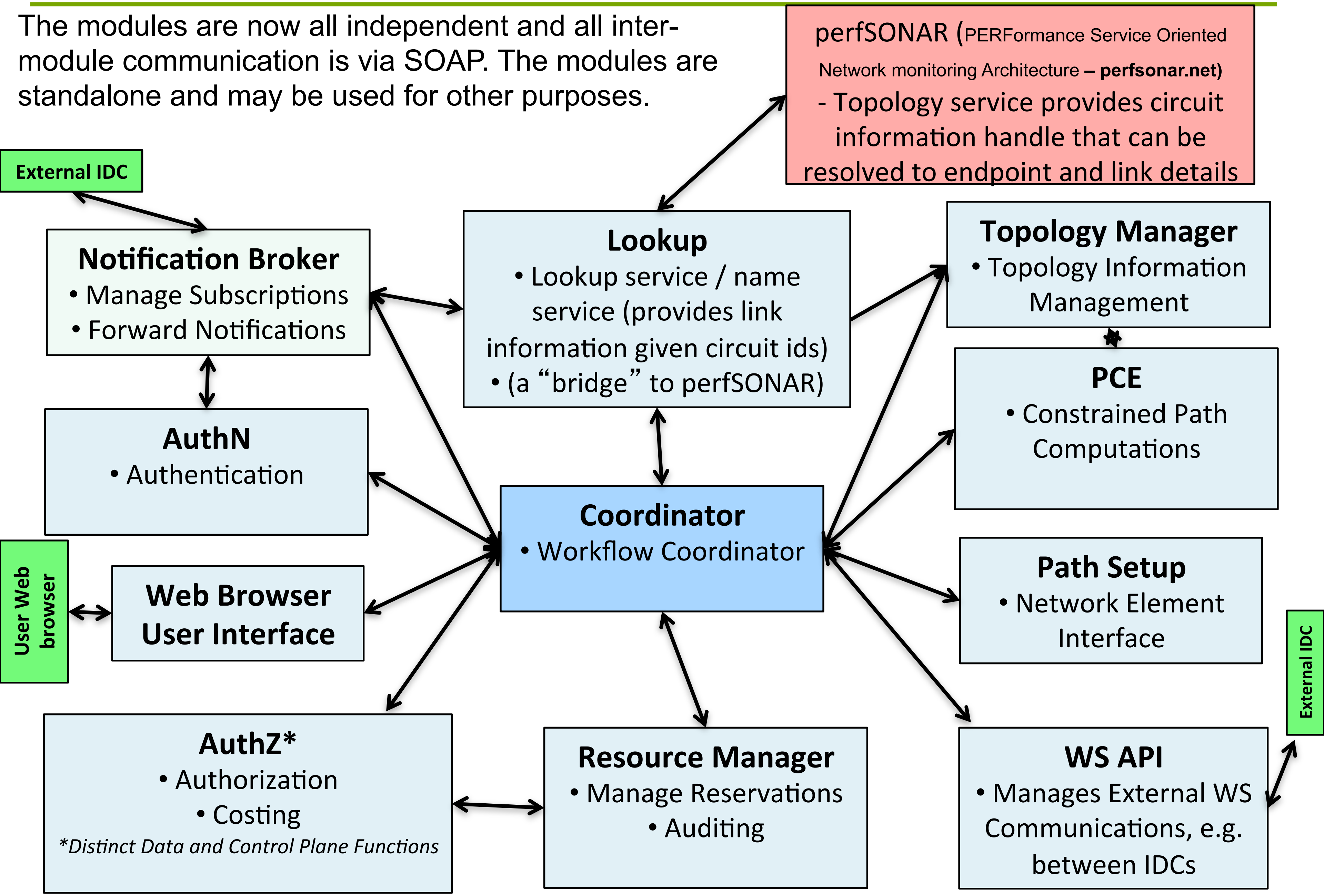


OSCARS 0.6 (Version 3) Design / Implementation Goals

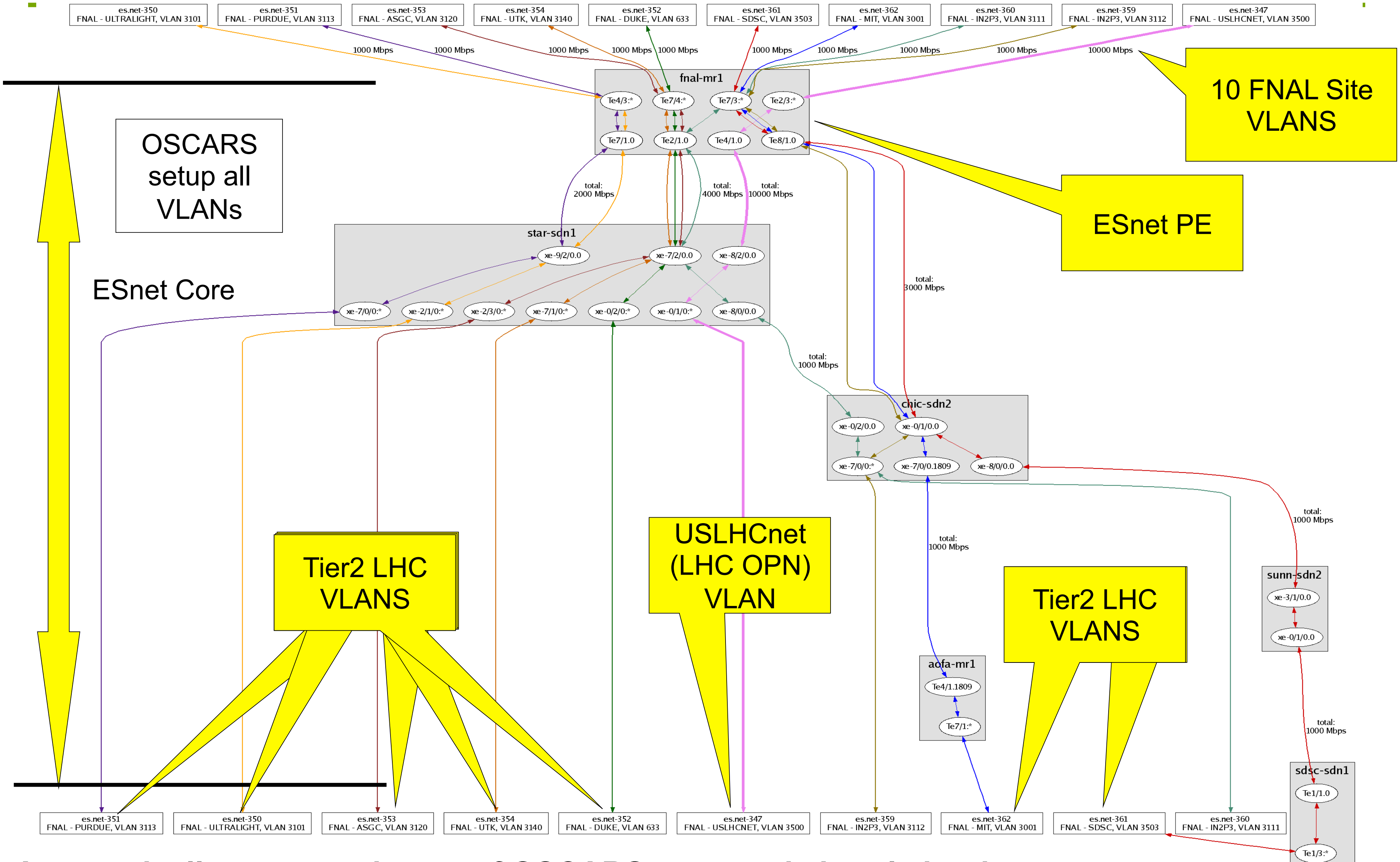
- Support production deployment of service and facilitate research collaborations
  - Distinct functions in stand-alone modules
    - Supports distributed model
    - Facilitates module redundancy
  - Formalize (internal) interface between modules
    - Facilitates module plug-ins from collaborative work (e.g. PCE)
    - Customization of modules based on deployment needs (e.g. AuthN, AuthZ, PSS)
  - Standardize external API messages and control access
    - Facilitates inter-operability with other dynamic VC services (e.g. Nortel DRAC, GÉANT AutoBAHN)
    - Supports backward compatibility of IDC protocol

OSCARS 0.6 (ver. 3) Architecture and Implementation

The modules are now all independent and all inter-module communication is via SOAP. The modules are standalone and may be used for other purposes.



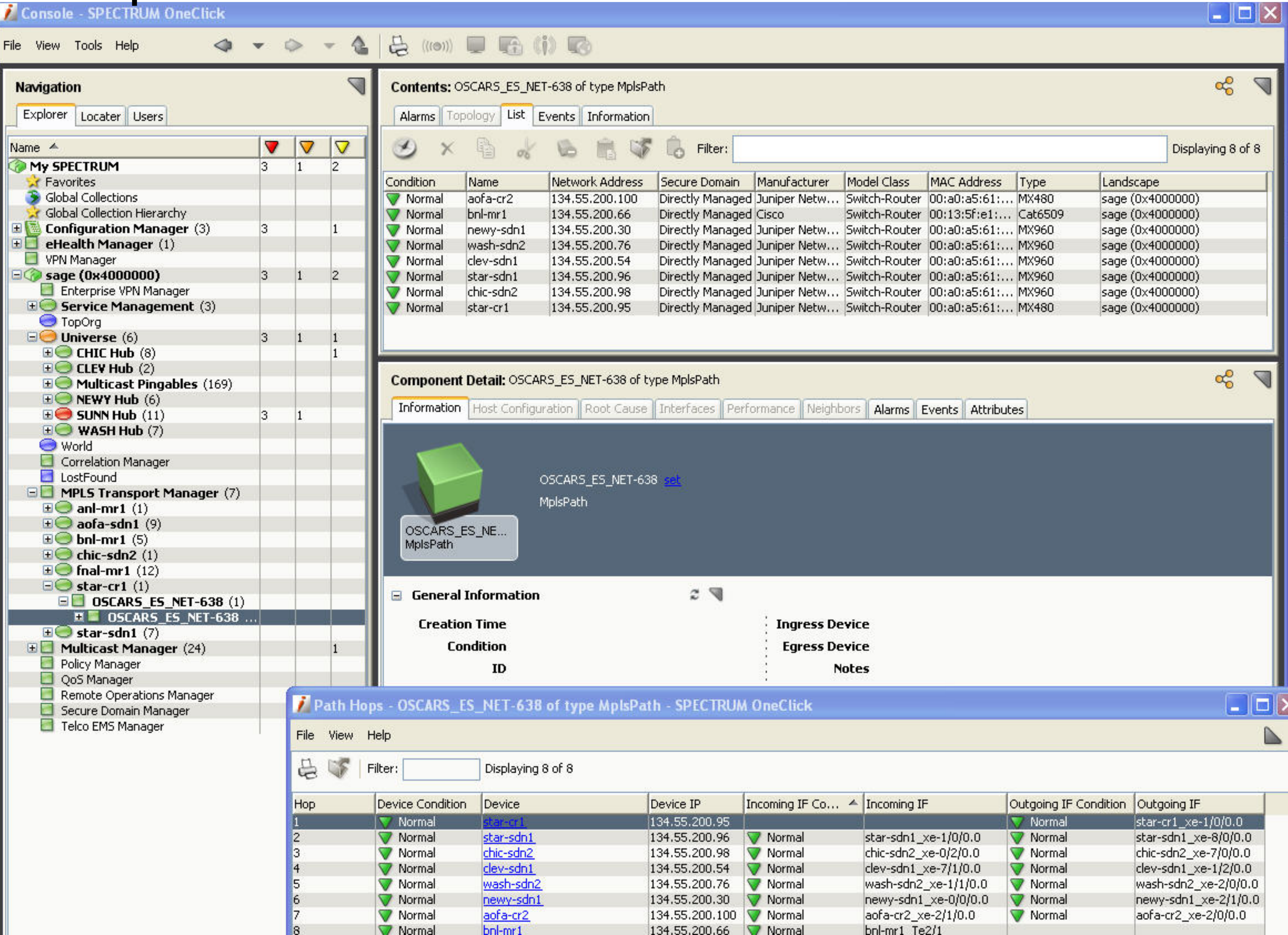
➢OSCARS is a production service in ESnet



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.

Spectrum Network Monitor Can Now Monitor OSCARS Circuits



OSCARS Collaborative Research Efforts

- LBNL LDRD "On-demand overlays for scientific applications"
  - To create proof-of-concept on-demand overlays for scientific applications that make efficient and effective use of the available network resources
- GLIF GNI-API "Fenius" to translate between the GLIF common API to:
  - DICE IDCP: OSCARS IDC (ESnet, I2)
  - GNS-WSI3: G-lambda (KDDI, AIST, NICT, NTT)
  - Phosphorus: Harmony (PSNC, ADVA, CESNET, NXW, FHG, I2CAT, FZJ, HEL IBBT, CTI, AIT, SARA, SURFnet, UNIBONN, UVA, UESSEX, ULEEDS, Nortel, MCNC, CRC)
- DOE Projects:
  - "Virtualized Network Control" to develop multi-dimensional PCE (multi-layer, multi-level, multi-technology, multi-layer, multi-domain, multi-provider, multi-vendor, multi-policy)
  - "Integrating Storage Management with Dynamic Network Provisioning for Automated Data Transfers" to develop algorithms for co-scheduling compute and network resources
  - "Hybrid Multi-Layer Network Control" to develop end-to-end provisioning architectures and solutions for multi-layer networks

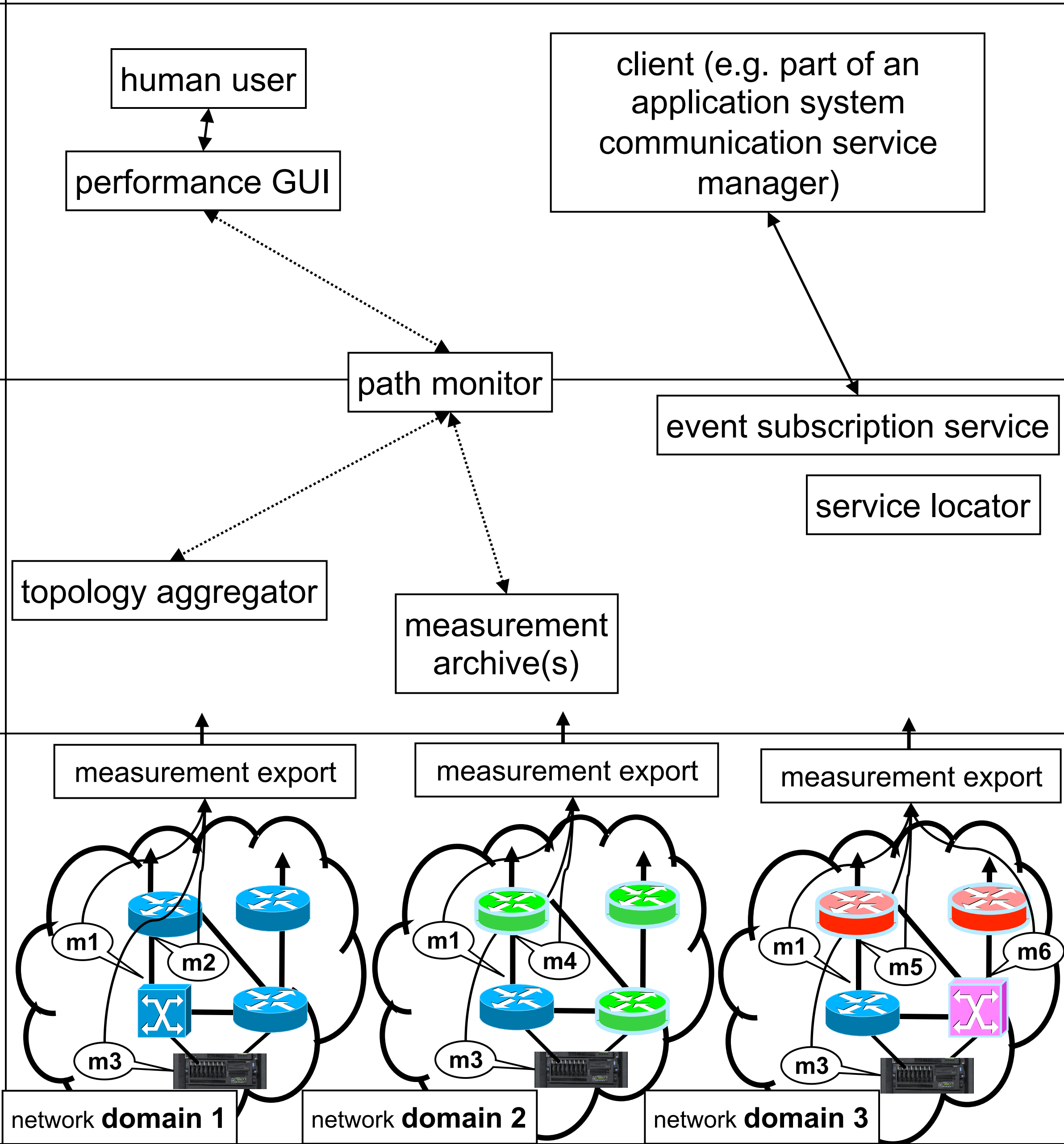
Response Strategy III: Monitoring as a Service-Oriented Communications Service

- 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 follows work of the Open Grid Forum (OGF) Network Measurement Working Group (NM-WG) and is based on SOAP XML messages
- 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

layer	examples
interface	• real-time end-to-end performance graph (e.g. bandwidth or packet loss vs. time) • historical performance data for planning purposes • event subscription service (e.g. end-to-end path segment outage)
service	
measurement point	• The measurement points (m1….m6) are the real-time feeds from the network or local monitoring devices • The Measurement Export service converts each local measurement to a standard format for that type of measurement

architectural relationship



perfSONAR Application: Traceroute Visualizer

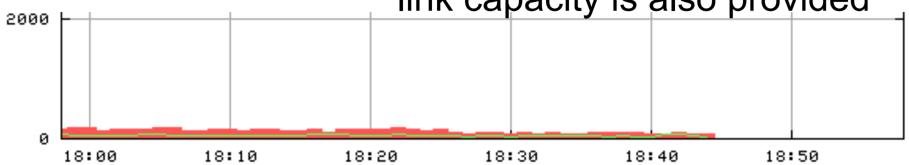
- Multi-domain path performance monitoring is an example of a tool based on perfSONAR protocols and infrastructure
  - provide users/applications with the end-to-end, multi-domain traffic and bandwidth availability
  - provide real-time performance such as path utilization and/or packet drop
  - One 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

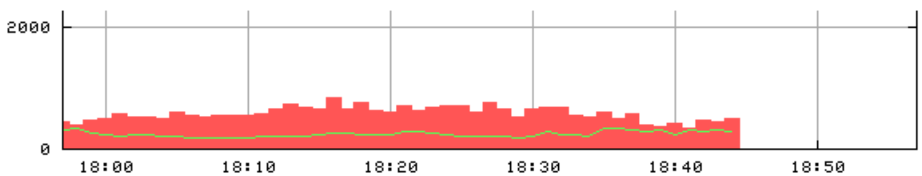
- Forward direction bandwidth utilization on application path from LBNL to INFN-Frascati (Italy) (2008 SNAPSHOT)
  - traffic shown as bars on those network device interfaces that have an associated MP services (the first 4 graphs are normalized to 2000 Mb/s, the last to 500 Mb/s)

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

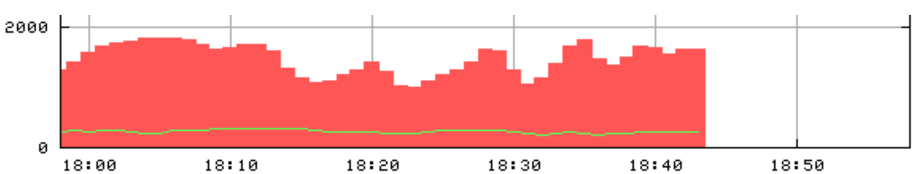
link capacity is also provided



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

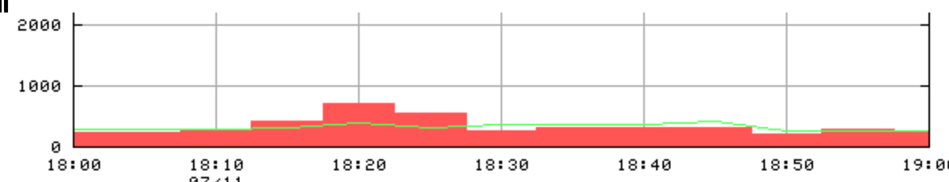


7 chislsdn1-oc192-snv2sdn1.es.net (GRAPH OMITTED)
8 chiccr1-chislsdn1.es.net

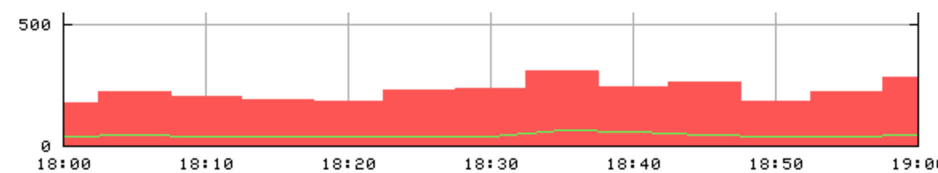


9 aofacr1-chicsdn1.es.net (GRAPH OMITTED)

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***

ESnet

Energy Sciences Network

U.S. DEPARTMENT OF ENERGY | Office of Science

- Some details

- *The ESnet Planning Process*

How ESnet Determines its Network Architecture, Services, and Bandwidth

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):

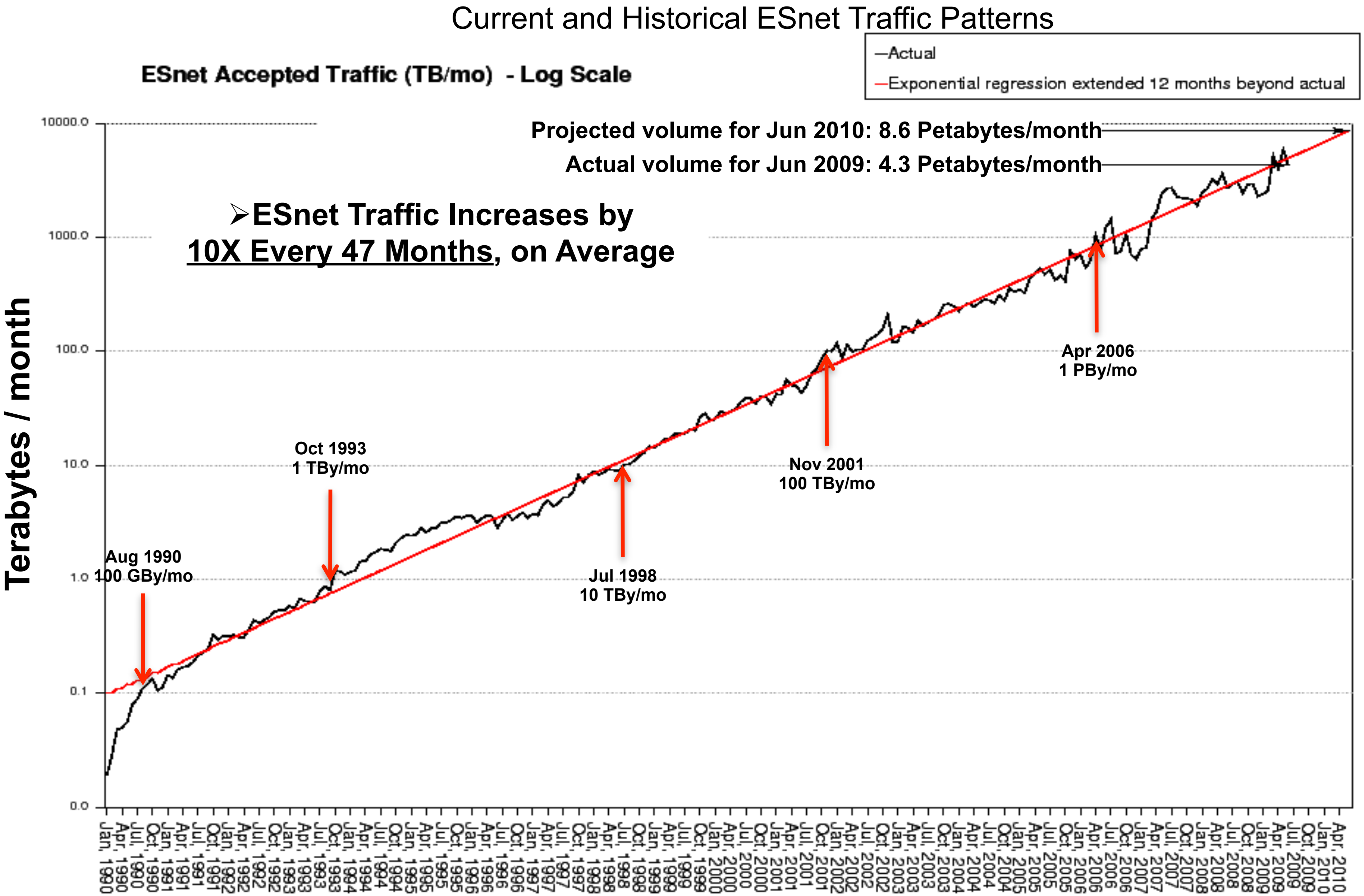
1a) Data characteristics of scientific instruments and facilities

- 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

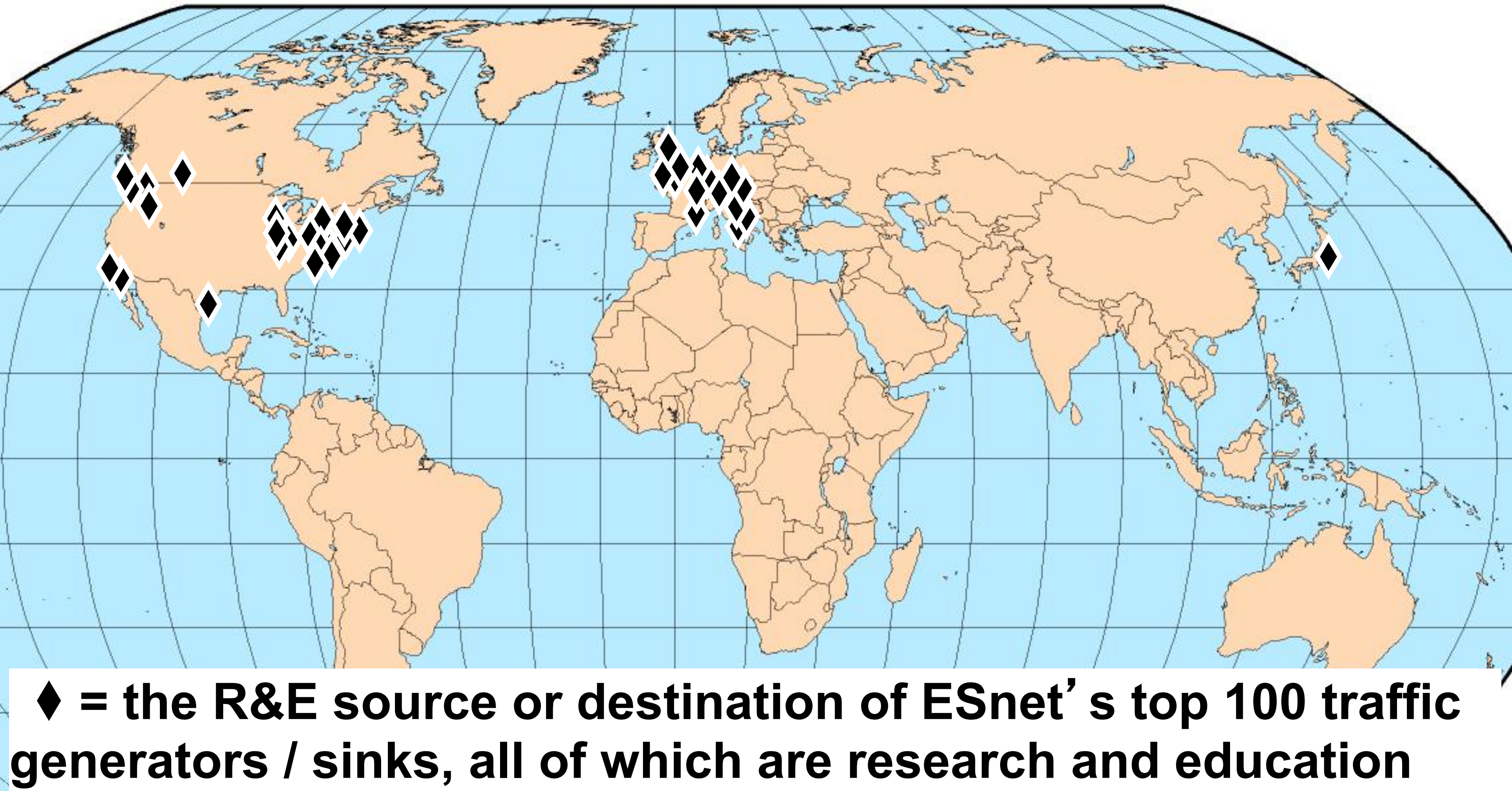
- How and where will the new data be analyzed and used – that is, how will the process of doing science change over 5-10 years?

➢Observation: Current and Historical ESnet Traffic Patterns



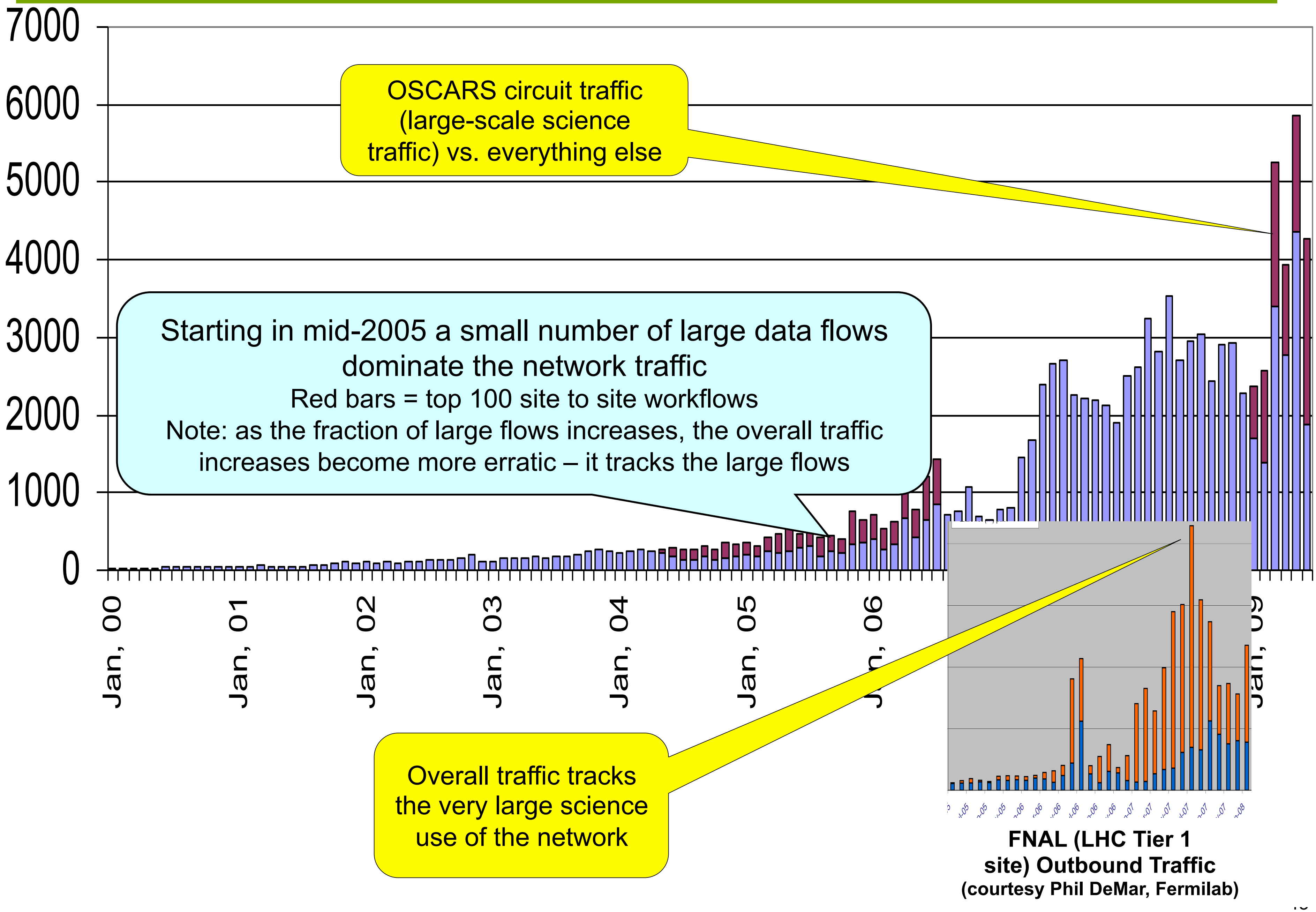
Log Plot of ESnet Monthly Accepted Traffic, January 1990 – June 2009

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)

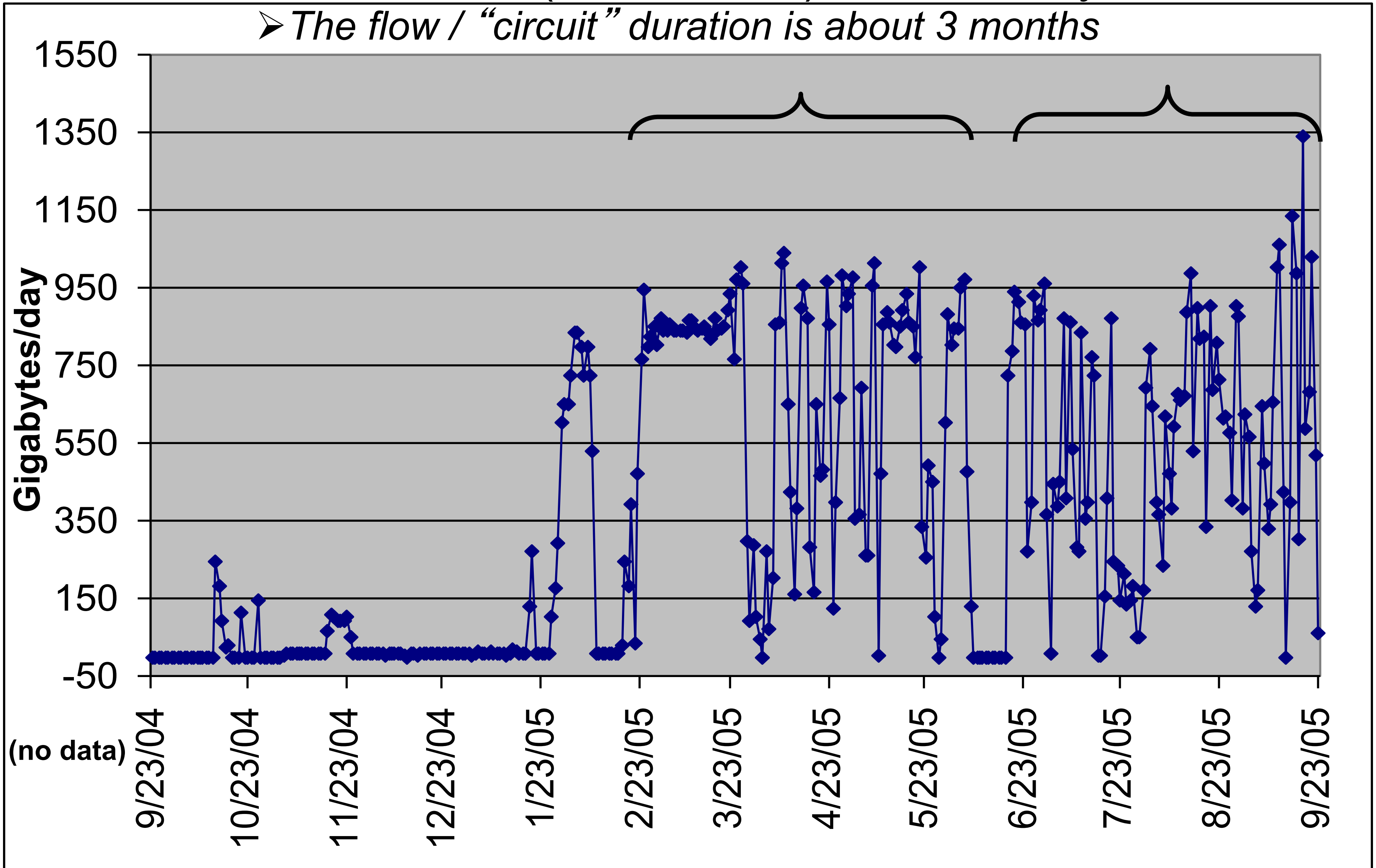
Observing the Network: A small number of large data flows now dominate the network traffic – this motivates virtual circuits as a key network service



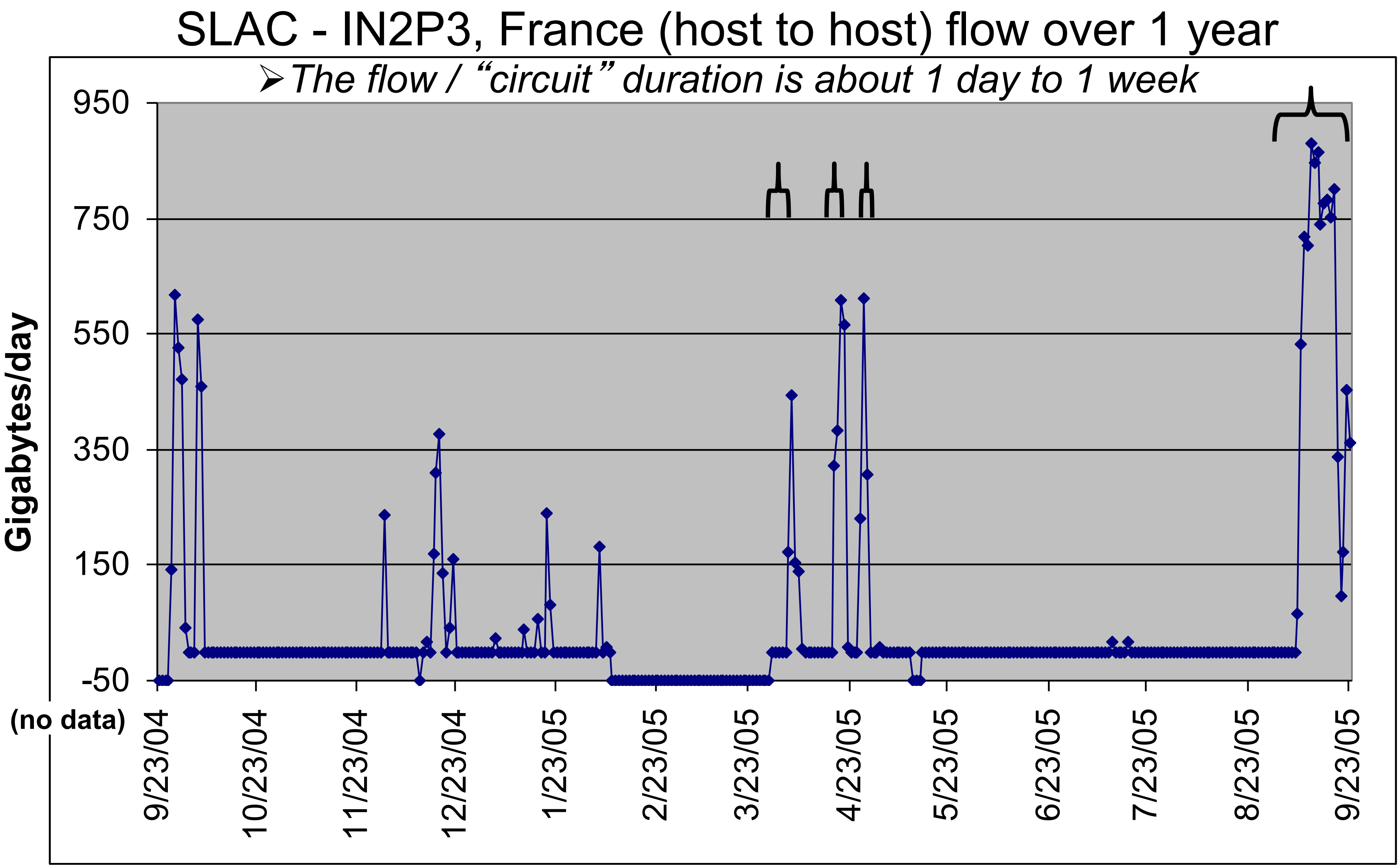
Observing the Network: Most of the Large Flows Exhibit Circuit-like Behavior

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 10X every 3-4 years**
- **Peerings must be built to accommodate the fact that 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 use 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 use of the network 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
    - The 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		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

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

Science Network Requirements Aggregation Summary

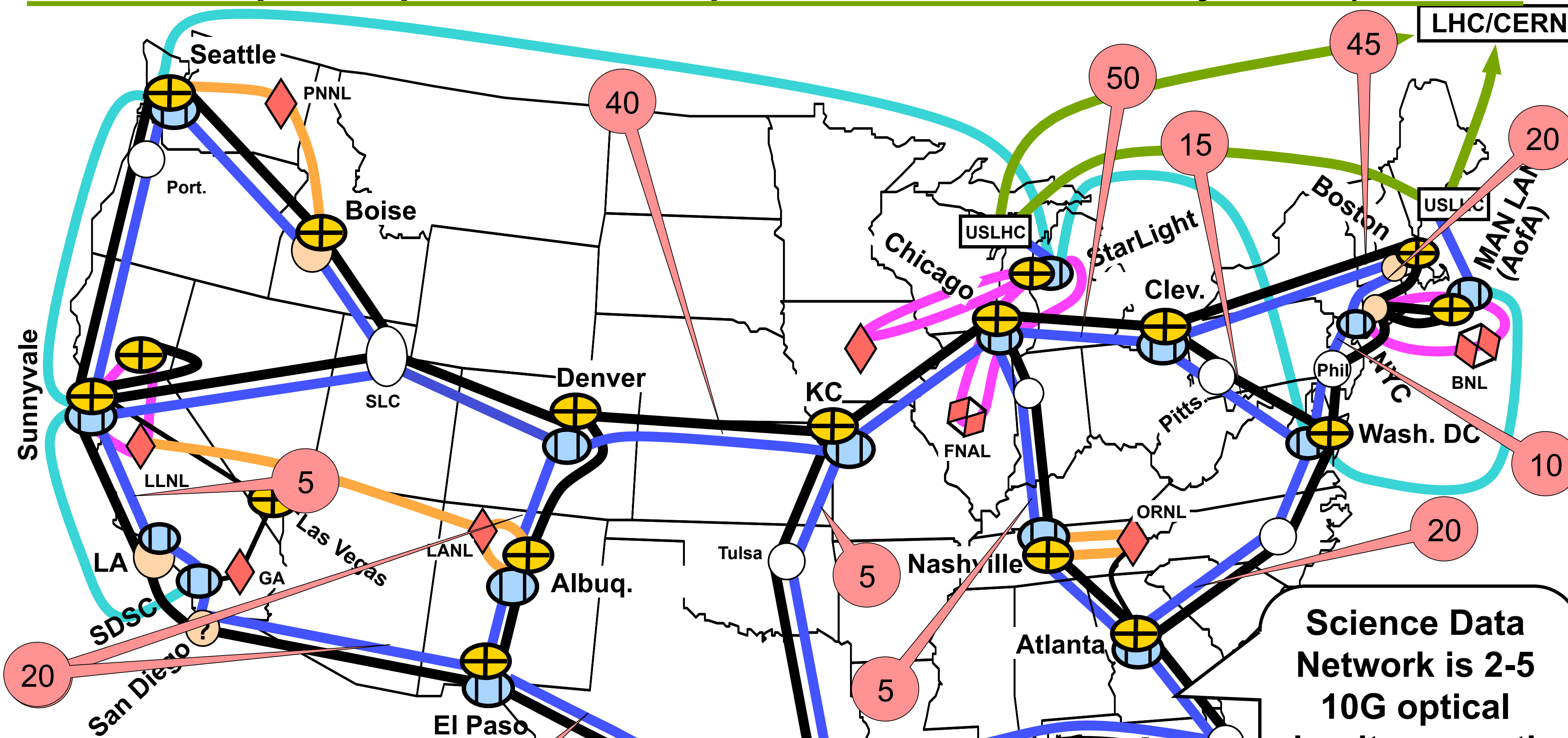
Science Drivers Science Areas / Facilities	End2End Reliability	Near Term End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
BES: Chemistry and Combustion	-	5-10Gbps	30Gbps	• Bulk data • Real time data streaming	• Data movement middleware
BES: Light Sources	-	15Gbps	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)



ESnet IP switch/router hubs

ESnet SDN switch hubs

Layer 1 optical nodes - eventual ESnet Points of Presence

Layer 1 optical nodes not currently in ESnet plans

Lab site

Lab site – independent dual connect.

XX Committed path capacity, Gb/s

ESnet IP core

ESnet Science Data Network core (N X 10G)

ESnet SDN core, NLR links (backup paths)

Lab supplied link

LHC related link

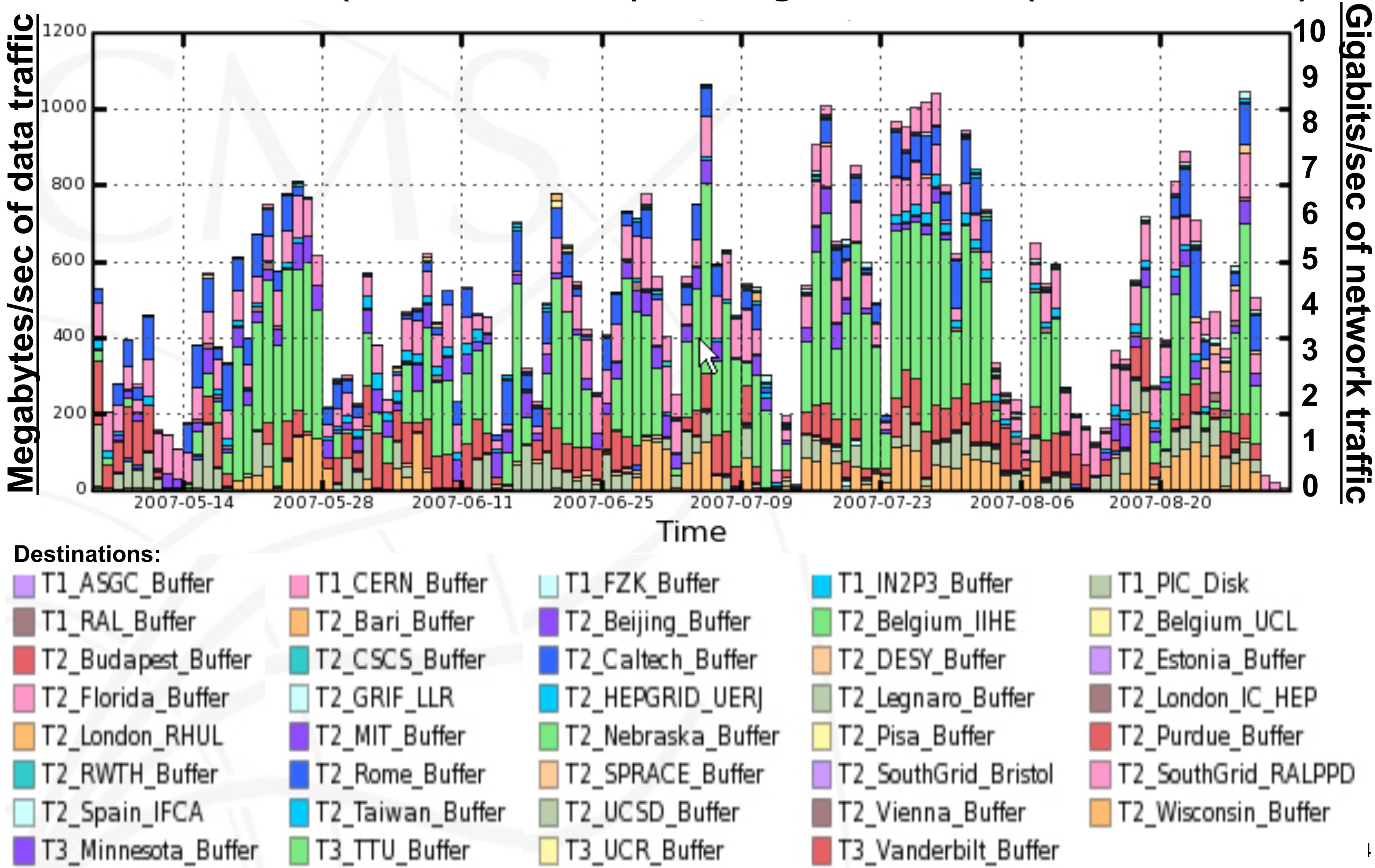
MAN link

International IP Connections

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)



Services Requirements from Instruments and Facilities

Fairly consistent requirements are found across the large-scale sciences

- ***Large-scale science uses distributed applications 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 distributed application 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

Services Requirements from Instruments and Facilities (cont.)

- The distributed application system elements must be able to get guarantees from the network that there is adequate bandwidth to accomplish the task at hand
- The distributed applications 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
- These services must be accessible within the Web Services / Grid Services paradigm of the distributed applications systems

See, e.g., [ICFA SCIC]

Summary Requirements from Instruments and Facilities

- ***Bandwidth – 200+ Gb/s core 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 isolation, service delivery architecture compatible with Web Services / Grid / "Systems of Systems" application development paradigms
    - Implicit requirement is that the service not have to pass through site firewalls which cannot handle the required bandwidth (frequently 10Gb/s)
  - ***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*

ESnet4 - The Response to the Requirements

I) A new network architecture and implementation strategy

- Provide two networks: IP and circuit-oriented Science Data Network
  - IP network for commodity flows
  - SDN network for large science data flows
  - Logical parity between the networks so that either one can handle both traffic types
- 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) Develop and deploy 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

III) Develop and deploy service-oriented, user accessible network monitoring systems

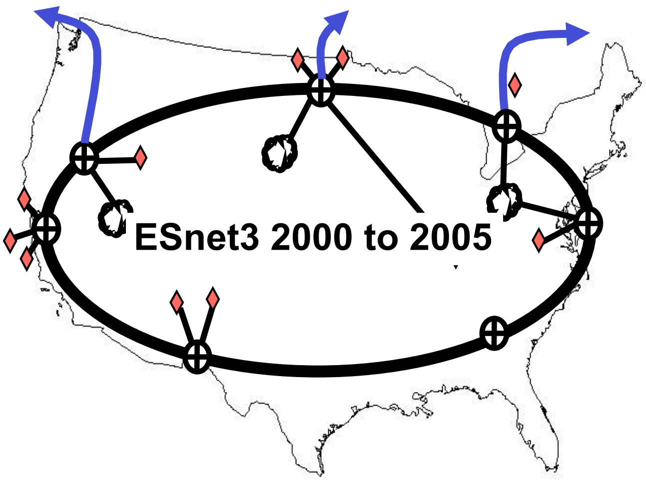
IV) Provide "consulting" on system / application network performance tuning

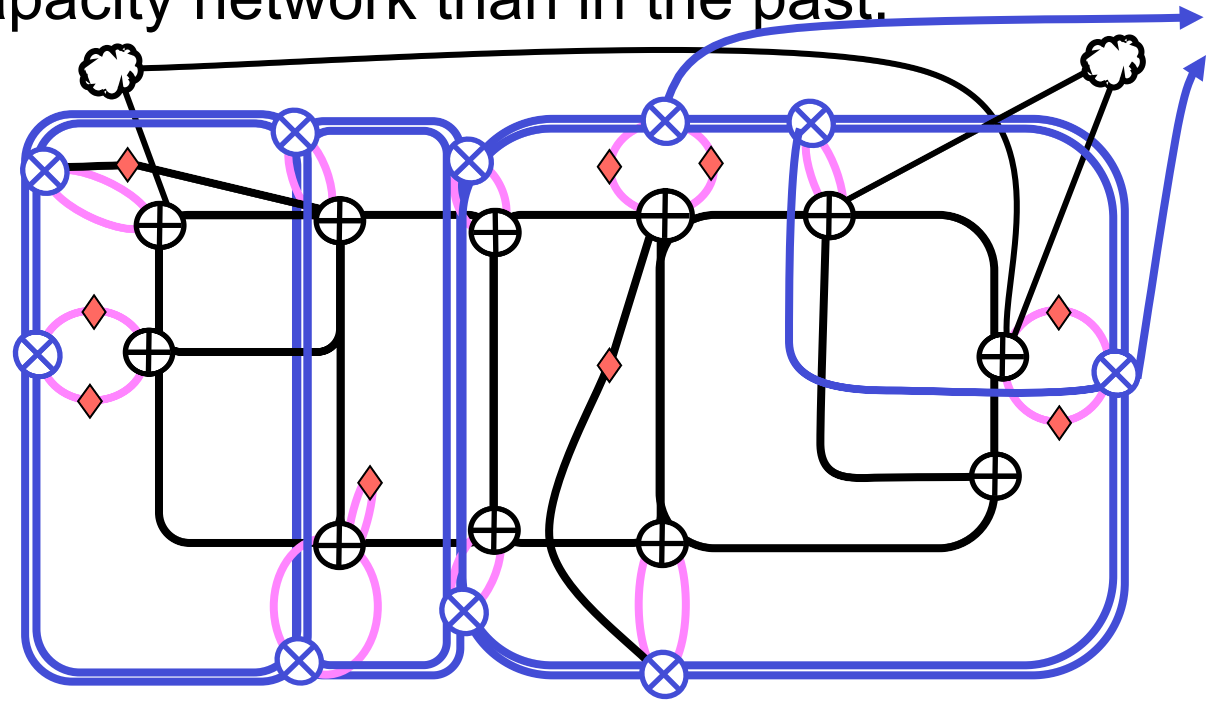
➢ Response Strategy I) ESnet4

- ESnet has built its next generation network as two separate networks:
  - An IP network for general traffic and
  - The new circuit-oriented Science Data Network for large-scale science traffic
- Both the IP and SDN networks are built on an underlying optical infrastructure that is shared between Internet2 (US R&E network) and ESnet

New ESnet Architecture

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

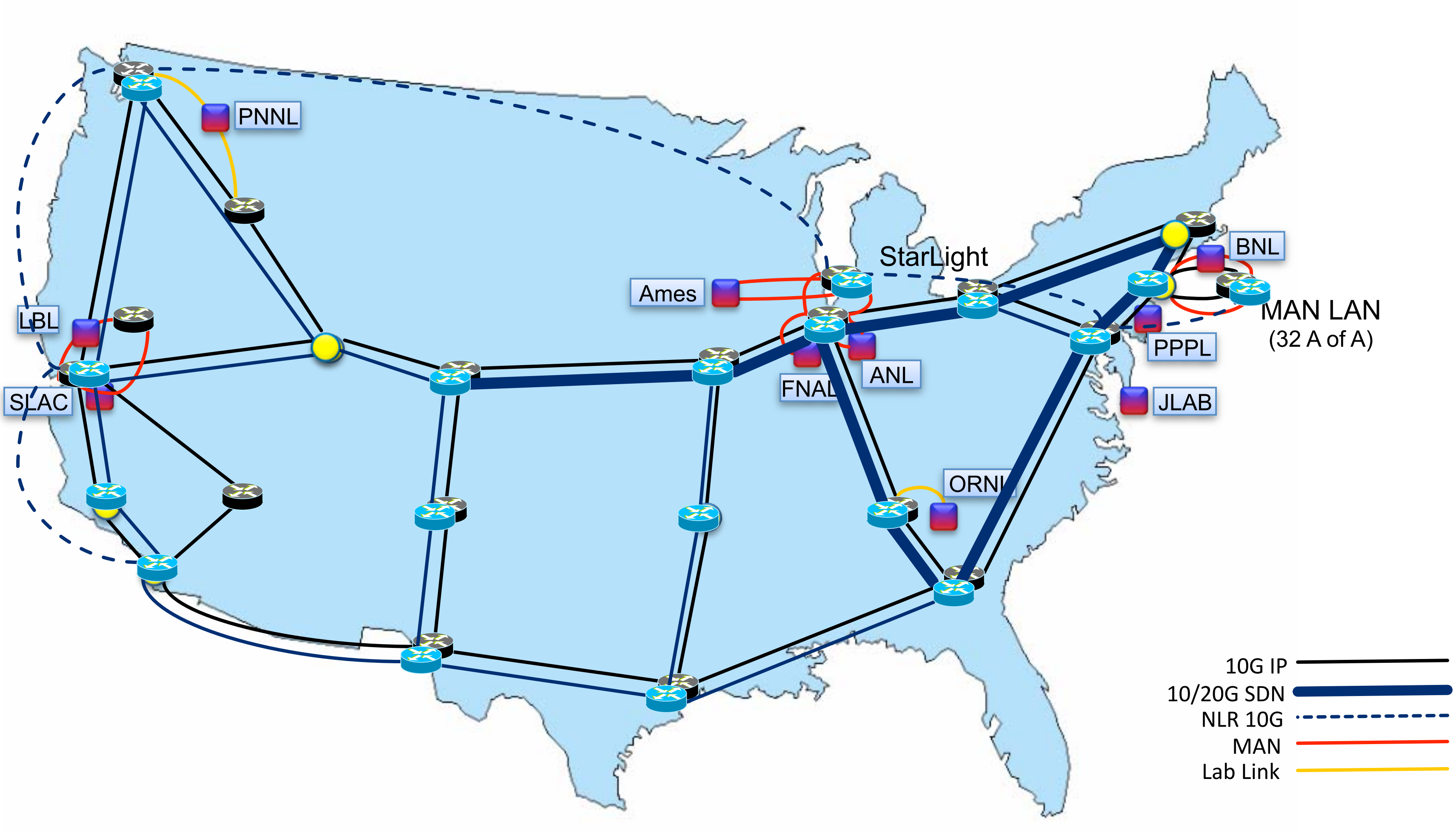




ESnet4 in 2008:

- **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**

ESnet4 Optical Footprint



Typical Internet2 and ESnet Optical Node