Networking for the Future of Large-Scale Science: An ESnet Perspective

Joint Techs
July, 2007

William E. Johnston
ESnet Department Head and Senior Scientist

Energy Sciences Network
Lawrence Berkeley National Laboratory

wej@es.net, www.es.net

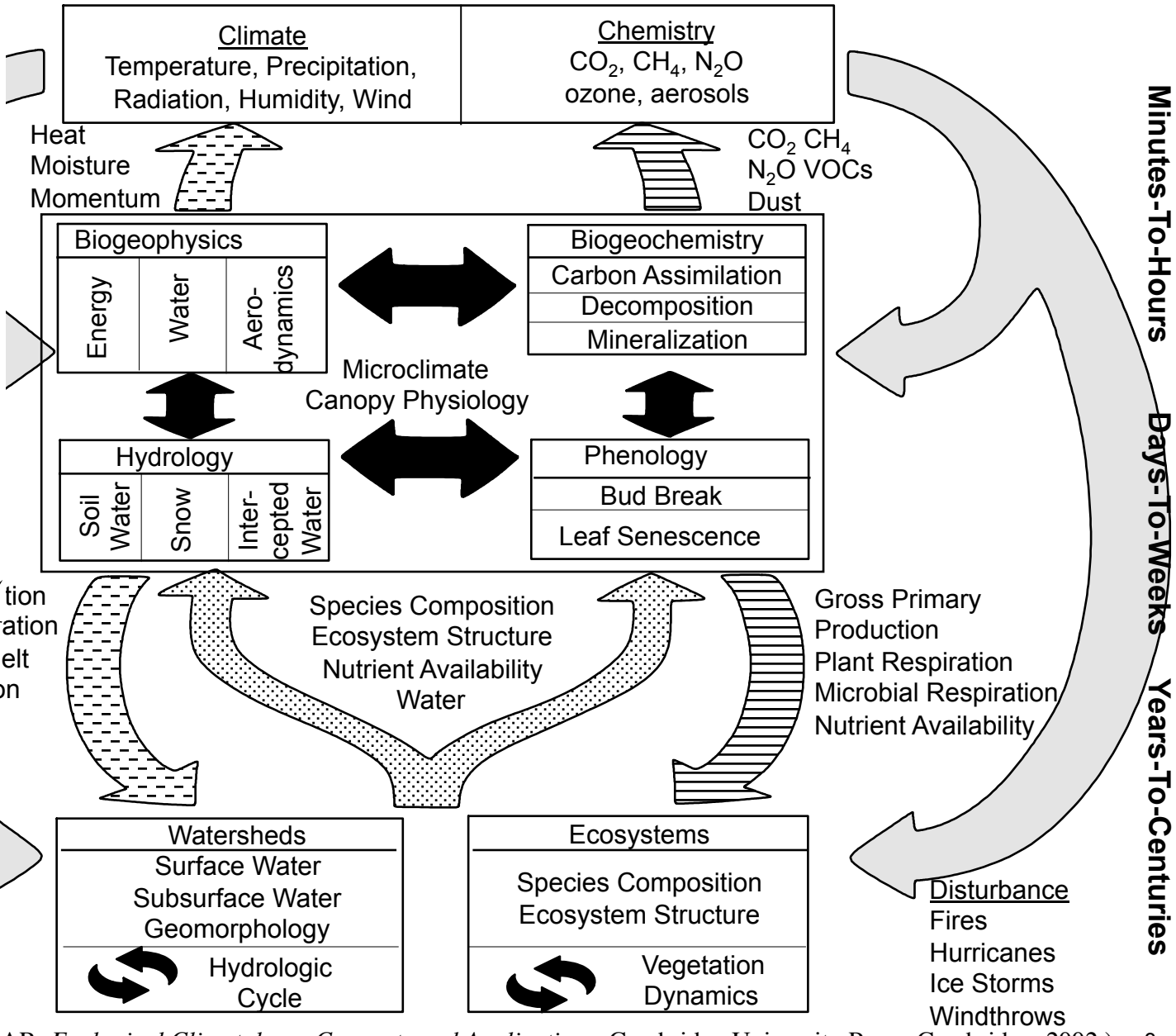
This talk is available at www.es.net/ESnet4

DOE's Office of Science: Enabling Large-Scale Science

- The Office of Science (SC) is the single ***largest supporter of basic research in the physical sciences in the United States***, … providing more than 40 percent of total funding … for the Nation's research programs in high-energy physics, nuclear physics, and fusion energy sciences. (http://www.science.doe.gov) – 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 is an SC program whose primary mission is to enable the large-scale science of the Office of Science (SC) 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

Distributed Science Example: Multidisciplinary Simulation

A "complete" approach to climate modeling involves many interacting models and data that are <u>provided by different groups at different locations</u>

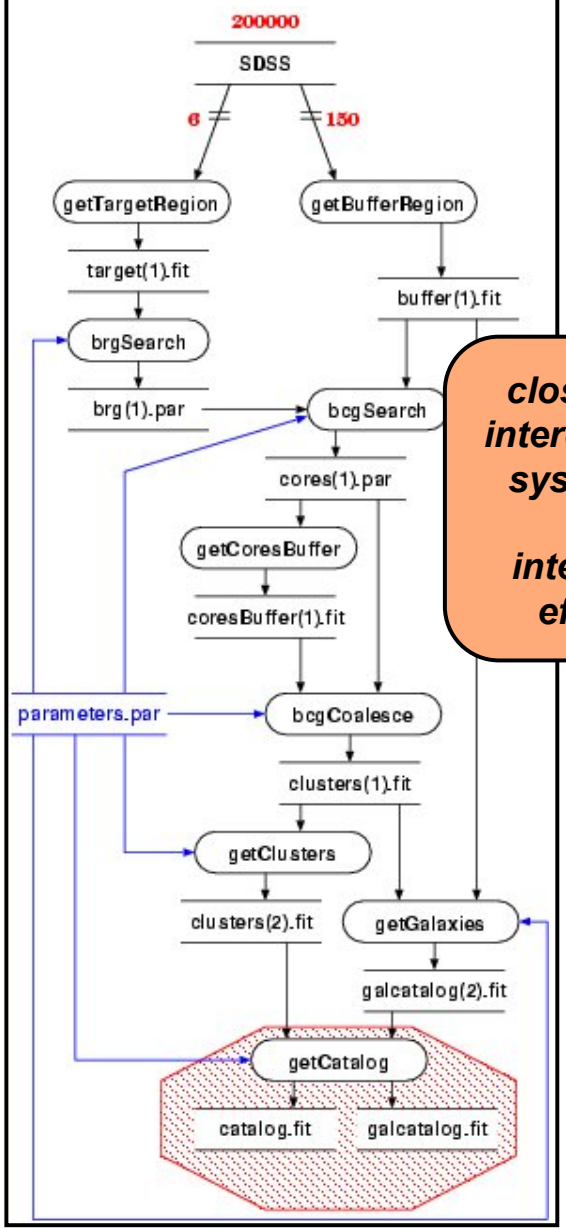


closely coordinated and interdependent distributed systems that must have predictable intercommunication for effective functioning

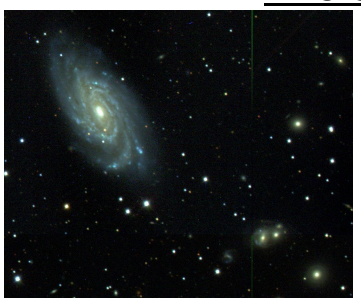
(Courtesy Gordon Bonan, NCAR: *Ecological Climatology: Concepts and Applications*. Cambridge University Press, Cambridge, 2002.)

Distributed Science Example: Sloan Galaxy Cluster Analysis

The science "application"



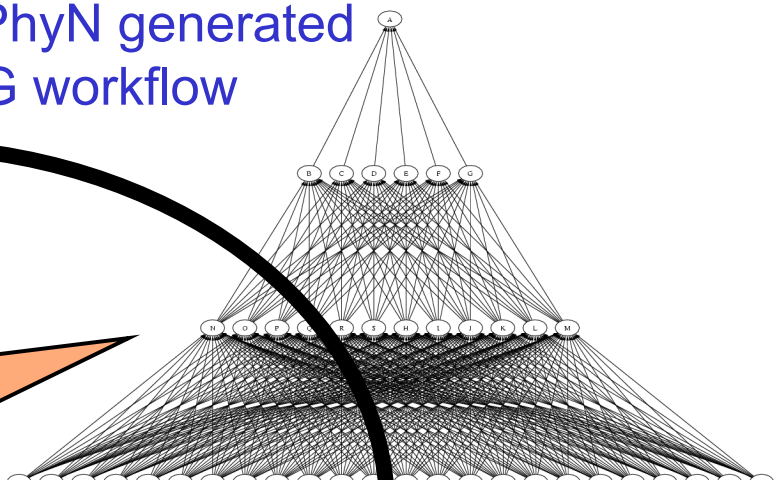
The science process and results



Sloan Data

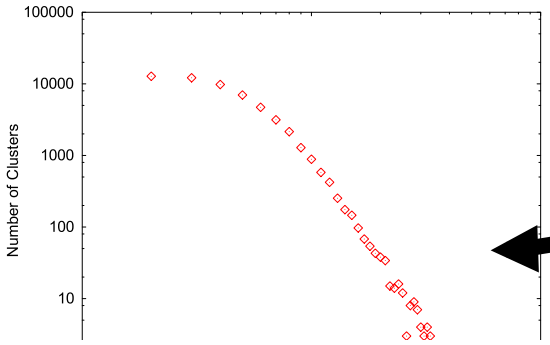
closely coordinated and interdependent distributed systems that must have predictable intercommunication for effective functioning

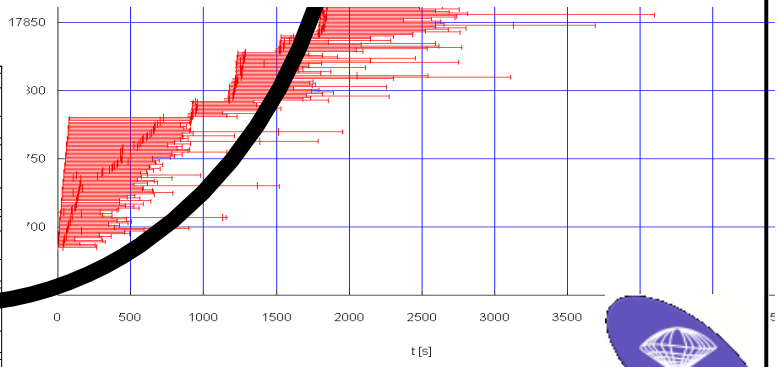
GriPhyN generated DAG workflow



A DAG representation of the workflow for 48 and 60 searches over 600 datasets (each node represents a process on a machine) executed in 2402 seconds on 62 hosts.

Galaxy cluster size distribution





***From "Applying Chimera Virtual Data Concepts to Cluster Finding in the Sloan Sky Survey," J. Annis, Y. Zhao, J. Voeckler, M. Wilde, S. Kent and I. Foster. In SC2002. 2002. Baltimore, MD.http://www.sc2002.org/paperpdfs/pap.pap299.pdf**

SDSS

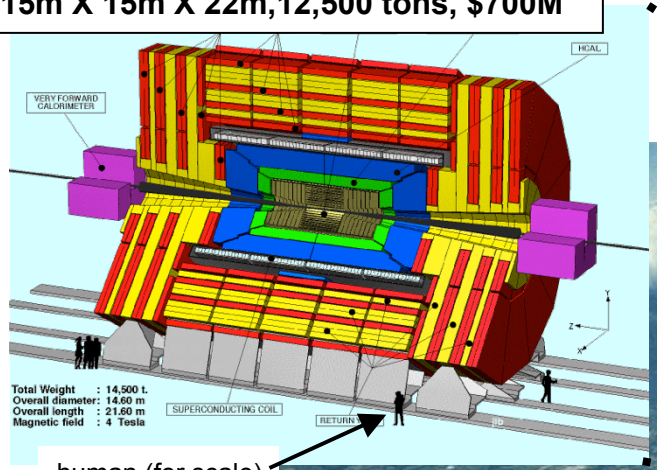
Large-Scale Science: High Energy Physics' Large Hadron Collider (Accelerator) at CERN

LHC Goal - Detect the Higgs Boson

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

The Largest Facility: Large Hadron Collider at CERN

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

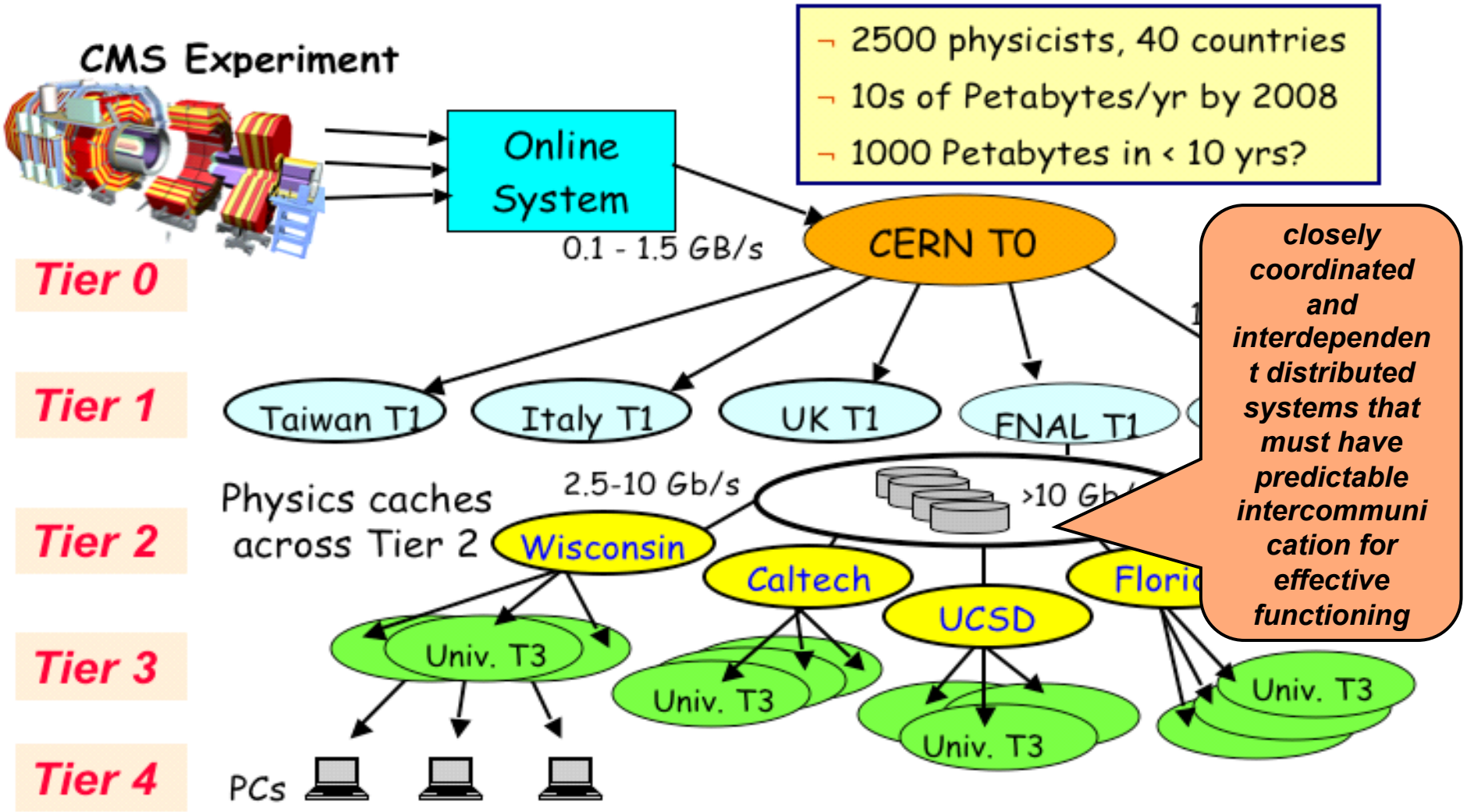


human (for scale)

CMS is one of several major detectors (experiments). The other large detector is ATLAS.

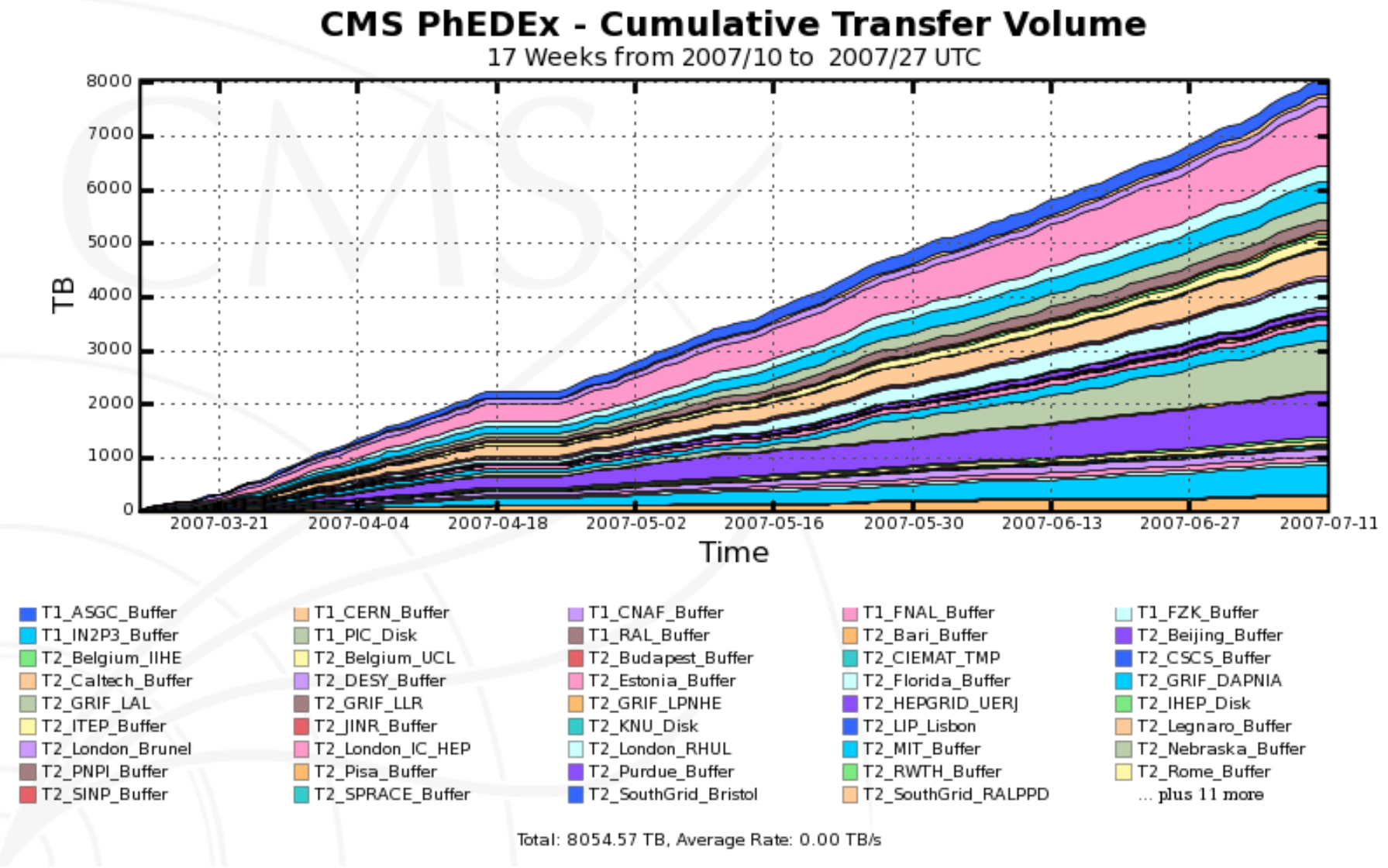
Two counter-rotating, 7 TeV proton beams, 27 km circumference (8.6 km diameter), collide in the middle of the detectors

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



Accumulated data (Terabytes) received by CMS Data Centers ("tier1" sites) and many analysis centers ("tier2" sites) during the past four months (8 petabytes of data) [LHC/CMS]

This sets the scale of the LHC distributed data analysis problem.



The LHC Data Management System has Several Characteristics that Result in Requirements for the Network and its Services

- The ***systems are data intensive and high-performance***, typically moving terabytes a day for months at a time
- The ***system are high duty-cycle***, operating most of the day for months at a time in order to meet the requirements for data movement
- The ***systems are widely distributed*** – typically spread over continental or inter-continental distances
- Such ***systems 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 applications ***must be able to get guarantees from the network*** that there is adequate bandwidth to accomplish the task at hand
- The applications ***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

This slide drawn from [ICFA SCIC]

Enabling Large-Scale Science

- These requirements are generally true for systems with widely distributed components to be reliable and consistent in performing the sustained, complex tasks of large-scale science

➢ **Networks must provide communication capability that is service-oriented: configurable, schedulable, predictable, reliable, and informative – and the network and its services must be scalable**

The LHC is the First of Many Large-Scale Science Scenarios

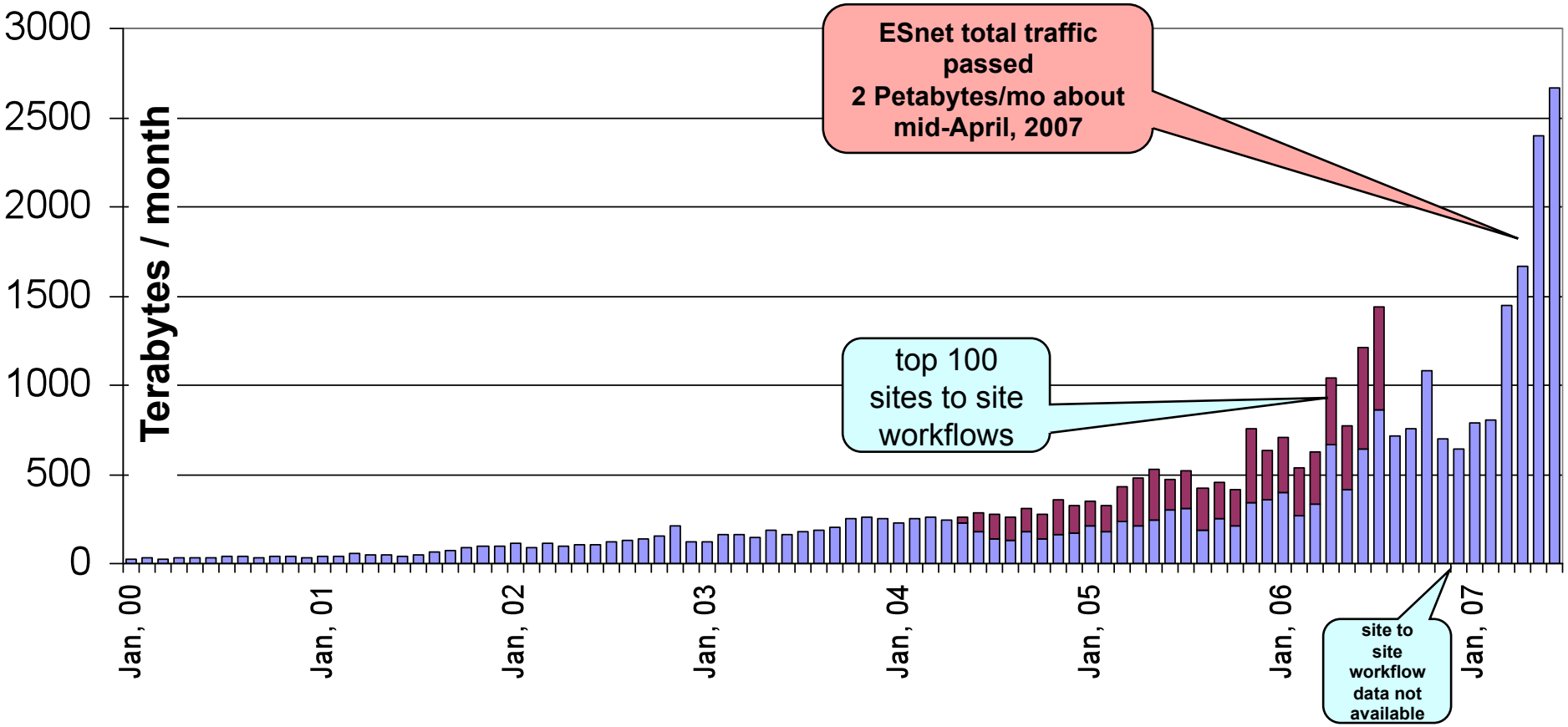
Science Drivers Science Areas / Facilities	End2End Reliability	Connectivity	Today End2End Band width	5 years End2End Band width	Traffic Characteristics	Network Services
Magnetic Fusion Energy	99.999% (Impossible without full redundancy)	• DOE sites • US Universities • Industry	200+ Mbps	1 Gbps	• Bulk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

(See refs. [1], [2], [3], and [4].)

The LHC is the First of Many Large-Scale Science Scenarios

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

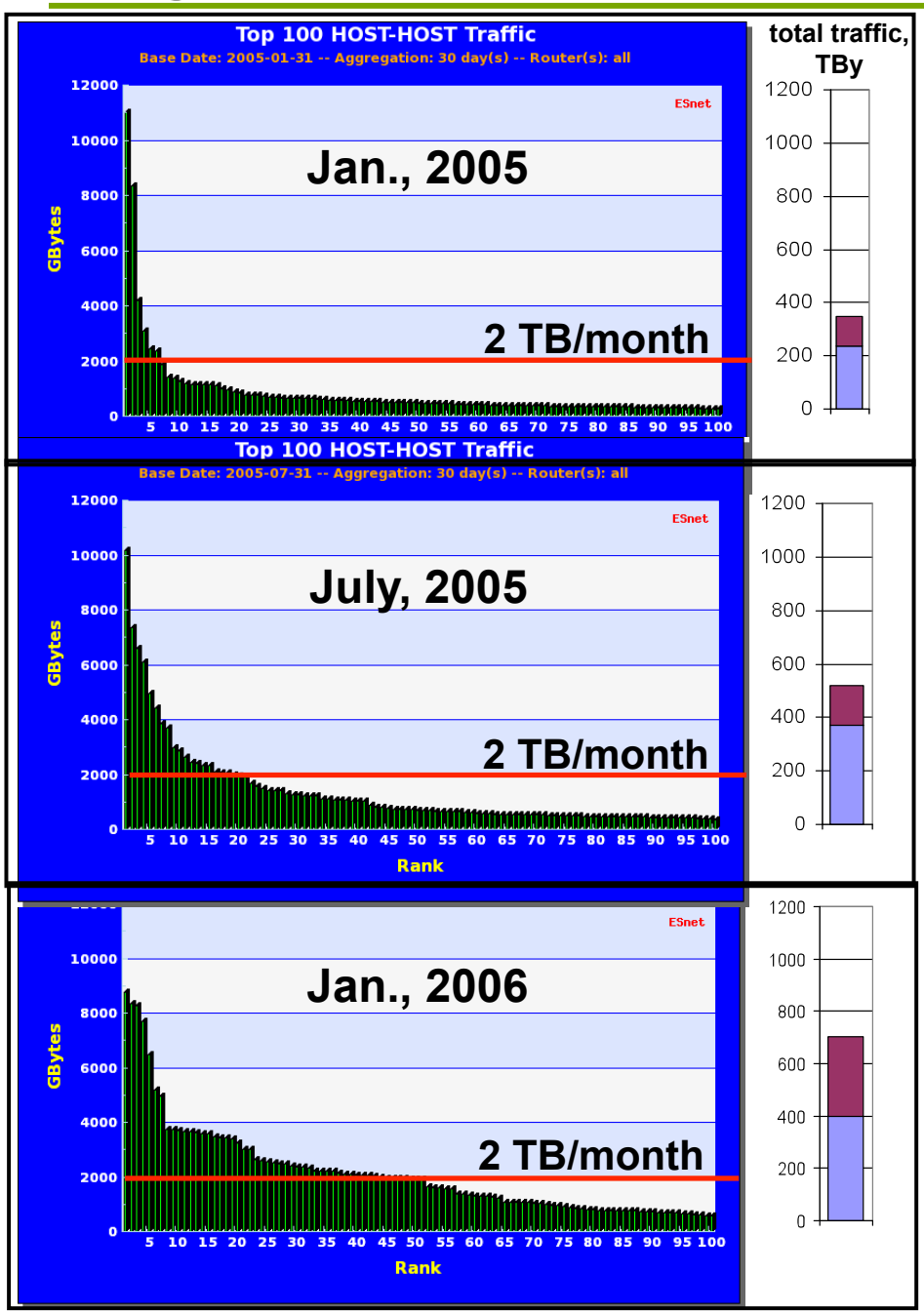
Large-Scale Science is Beginning to Dominate all Traffic

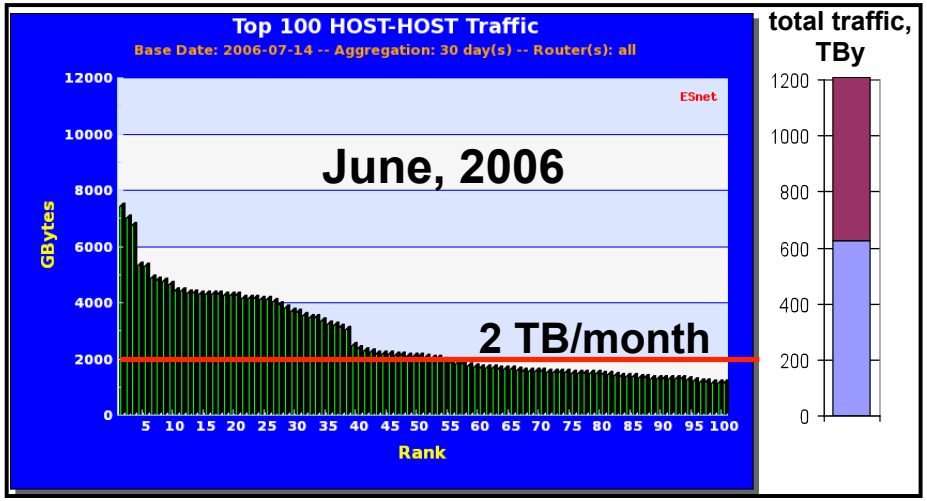


ESnet Monthly Accepted Traffic, January, 2000 – June, 2007

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

Large-Scale Science is Generating New Traffic Patterns

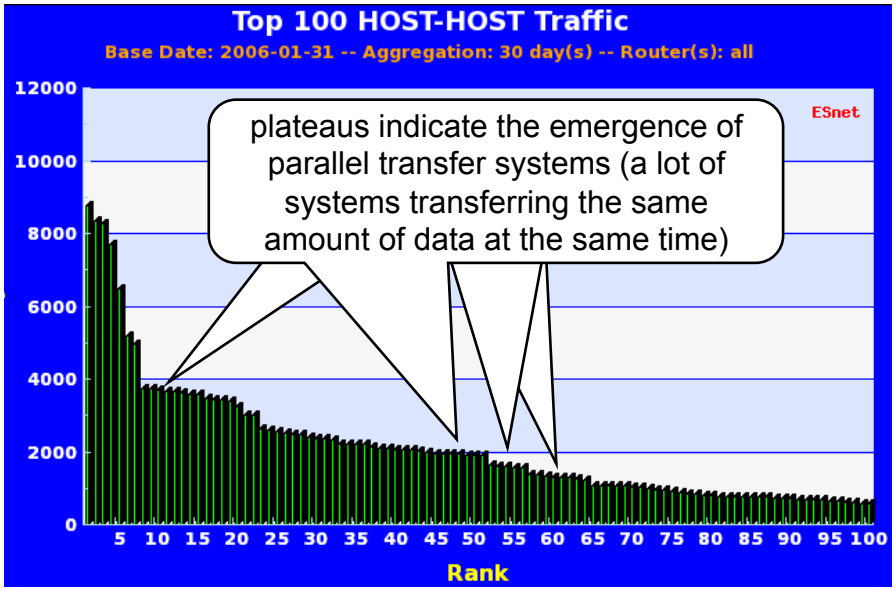




- While the total traffic is increasing exponentially
  - Peak flow – that is system-to-system – bandwidth is decreasing
  - The number of large flows is increasing

Large-Scale Science is Generating New Traffic Patterns

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

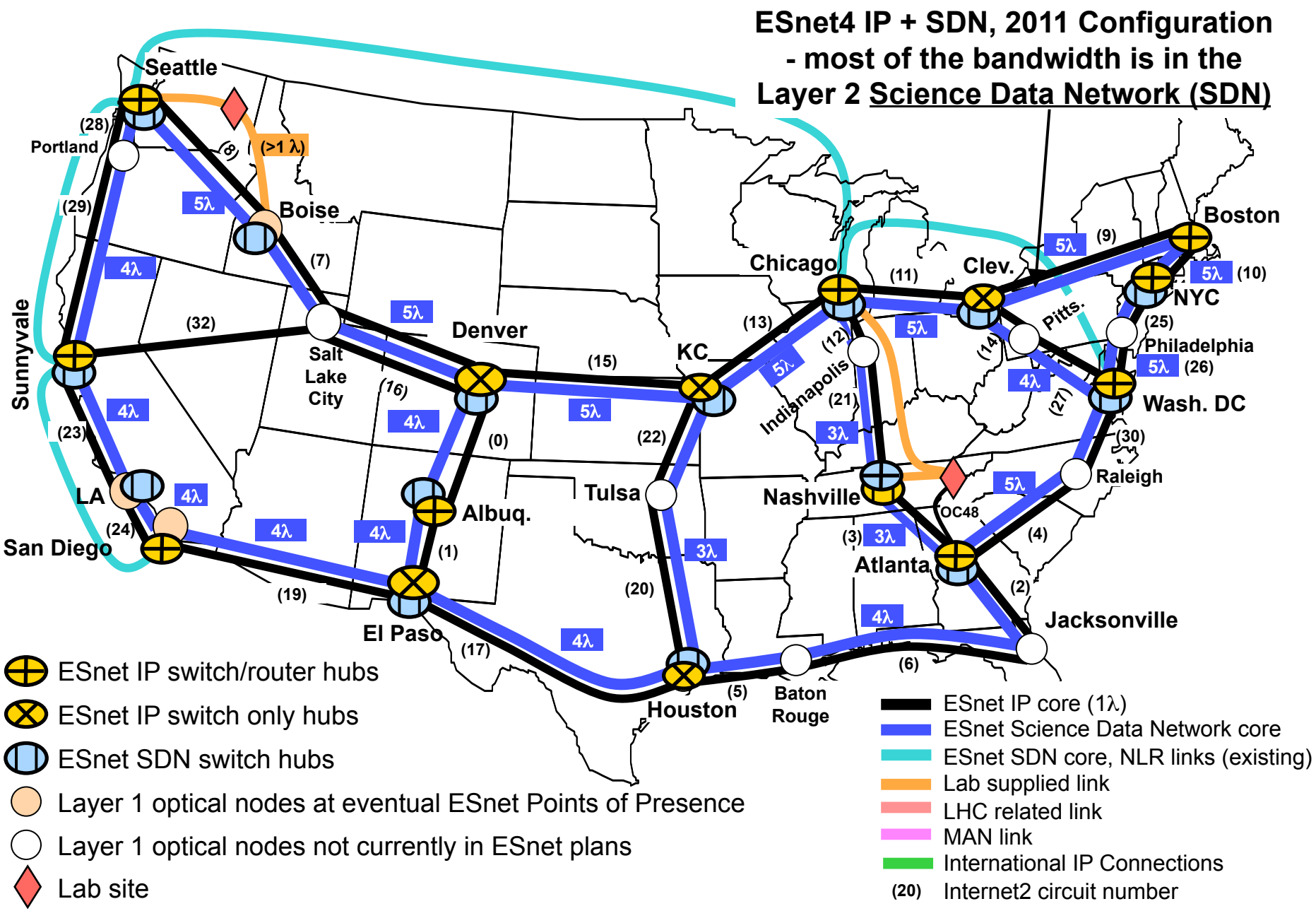
What Networks Need to Do

- The above examples currently only work in carefully controlled environments with the assistance of computing and networking experts
- For this essential approach to be successful in the long-term it must be routinely accessible to discipline scientists - without the continuous attention of computing and networking experts
- In order to
  - facilitate operation of multi-domain distributed systems
  - accommodate the projected growth in the use of the network
  - facilitate the changes in the types of traffic

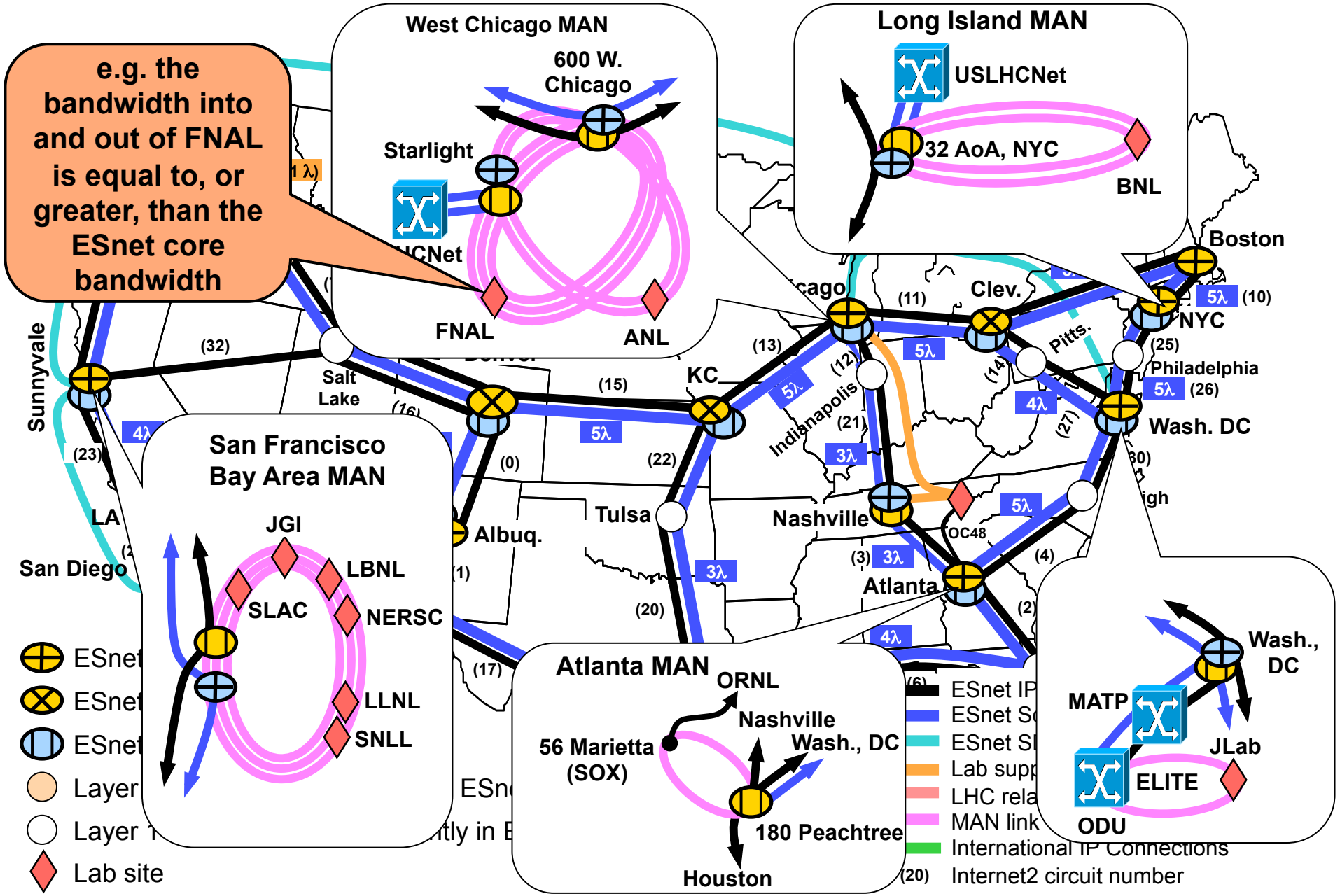
  the architecture and services of the network must change
- **The general requirements for the new architecture are that it provide:**
  1) **Support the high bandwidth data flows of large-scale science including scalable, reliable, and very high-speed network connectivity to end sites**
  2) **Dynamically provision virtual circuits with guaranteed quality of service (e.g. for dedicated bandwidth and for traffic isolation)**
  3) **provide users and applications with meaningful monitoring end-to-end (across multiple domains)**

The next several slides present the ESnet response to these requirements

1) A Hybrid Network is Tailored to Circuit-Oriented Services



High Bandwidth all the Way to the End Sites – major ESnet sites are now effectively directly on the ESnet "core" network

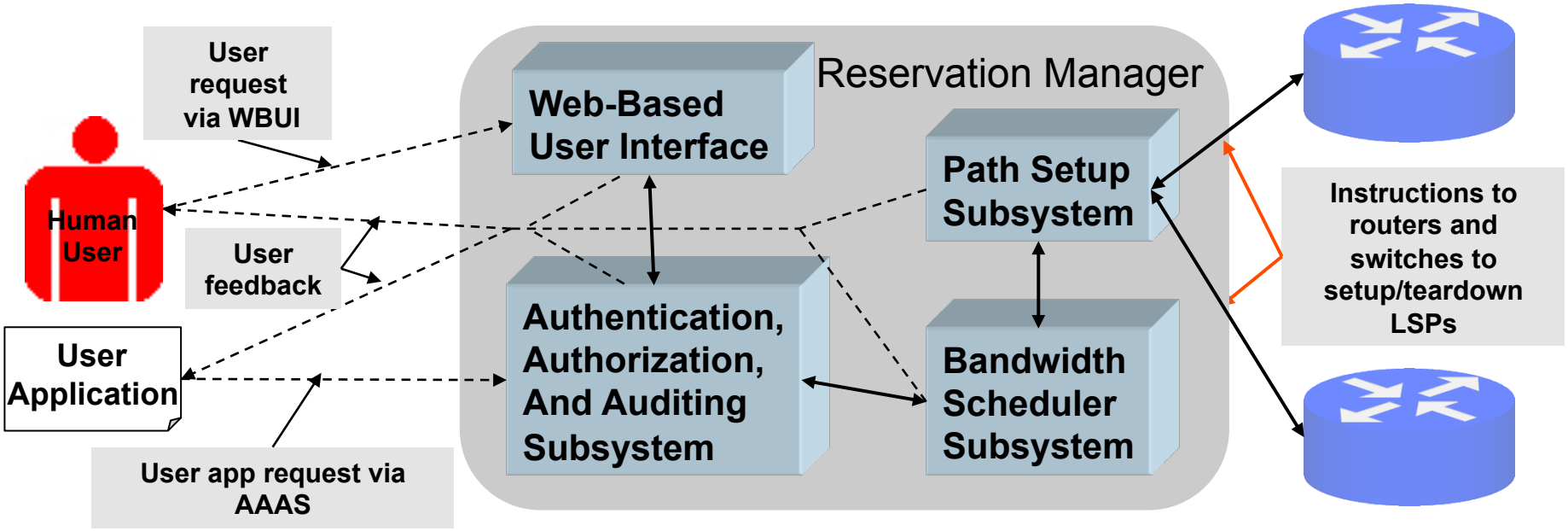


2) Multi-Domain Virtual Circuits

ESnet OSCARS [6] project has as its goals:

- 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 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 isolated from the general traffic
- End-to-end (cross-domain) connections between Labs and collaborating institutions

OSCARS



- To ensure compatibility, the design and implementation is done in collaboration with the other major science R&E networks and end sites
  - Internet2: Bandwidth Reservation for User Work (BRUW)
    - Development of common code base
  - GEANT: Bandwidth on Demand (GN2-JRA3), Performance and Allocated Capacity for End-users (SA3-PACE) and Advance Multi-domain Provisioning System (AMPS) extends to NRENs
  - BNL: TeraPaths - A QoS Enabled Collaborative Data Sharing Infrastructure for Peta-scale Computing Research
  - GA: Network Quality of Service for Magnetic Fusion Research
  - SLAC: Internet End-to-end Performance Monitoring (IEPM)
  - USN: Experimental Ultra-Scale Network Testbed for Large-Scale Science
  - DRAGON/HOPI: Optical testbed

3) Monitoring Applications of the Types that Move Us Toward Service-Oriented Communications Services

- E2Emon provides end-to-end path status in a service-oriented, easily interpreted way
  - a perfSONAR application used to monitor the LHC paths end-to-end across many domains
  - uses perfSONAR protocols to retrieve current circuit status every minute or so from MAs and MPs in all the different domains supporting the circuits
  - is itself a service that produces Web based, real-time displays of the overall state of the network, and it generates alarms when one of the MP or MA's reports link problems.

E2Emon: Status of E2E link CERN-LHCOPN-FNAL-001

Oper. State: Up
Admin. State: Normal Oper.

Domain	CERN		USLHCNET				
Link Structure	EP	◄·····	·····►	DP	◄——►	DP	◄·····
Type	EndPoint	ID Part.Info	ID Part.Info	Demarc	Domain Link	Demarc	ID Part.Info
Local Name	**CERN-T0**	S513-C-BE1	CERN-FERMI-LHCOPN-001-GVA-CERN	**USLHCNET-GEN**	CERN-FERMI-LHCOPN-001-GVA-CHI	**USLHCNET-CHI**	CERN-FERMI-LHCOPN-001-CHI-ESNET
State Oper.	-	Up	Up	-	Up	-	Up
State Admin.	-	Normal Oper.	Normal Oper.	-	Normal Oper.	-	Normal Oper.
Timestamp	-	2007-04-08 T05:04:08+02:00	2007-04-08 T05:04:11+02:00	-	2007-04-08 T05:04:53+02:00	-	2007-04-08 T05:03:59+02:00

Page generat

ESNET					FERMI			
·····►	DP	◄——►	DP	◄·····	·····►	DP	◄——►	EP
ID Part.Info	Demarc	Domain Link	Demarc	ID Part.Info	ID Part.Info	Demarc	Domain Link	EndPoint
CERN-FERMI-LHCOPN-001-STARLIGHT-Tail	**ESNET-STARLIGHT**	CERN-FERMI-LHCOPN-001-FERMI-STARLIGHT	**ESNET-FERMI**	CERN-FERMI-LHCOPN-001-Site-Tail	md8	**FERMI-ESNET**	md2	**FERMI-T1**
Up	-	Up	-	Up	Up	-	Up	-
Normal Oper.	-	Normal Oper.	-	Normal Oper.	Normal Oper.	-	Normal Oper.	-
2007-04-08 T01:40:37.0	-	2007-04-08T01:40:37.0	-	2007-04-08 T01:40:37.0	2007-04-08 T01:40:01.0-6:00	-	2007-04-08 T01:40:01.0-6:00	-

E2Emon generated view of the data for one OPN link [E2EMON]

Path Performance Monitoring

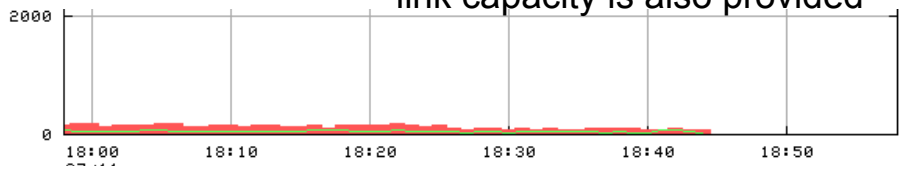
- Path performance monitoring needs to provide users/applications with the end-to-end, multi-domain traffic and bandwidth availability
  - should also provide real-time performance such as path utilization and/or packet drop
- Multiple path performance monitoring tools are in development
  - One example – Traceroute Visualizer [TrViz] – has been deployed at about 10 R&E networks in the US and Europe that have at least some of the required perfSONAR MA services to support the tool

Traceroute Visualizer

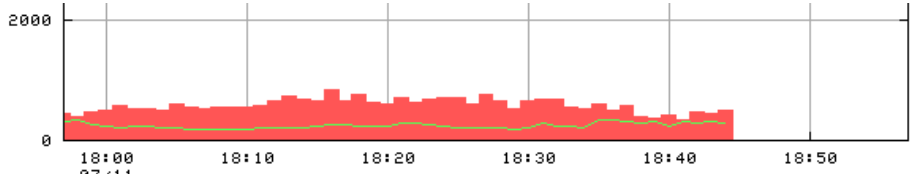
- Forward direction bandwidth utilization on application path from LBNL to INFN-Frascati (Italy)
  - traffic shown as bars on those network device interfaces that have an associated MP services (the first 4 graphs are normalized to 2000 Mb/s, the last to 500 Mb/s)

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

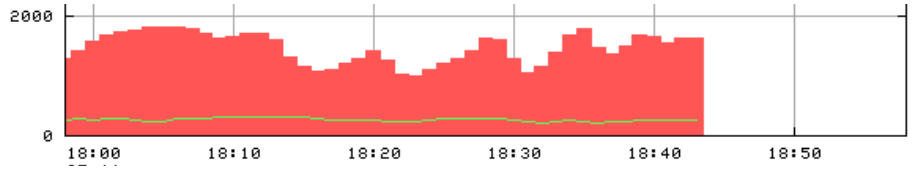
link capacity is also provided



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

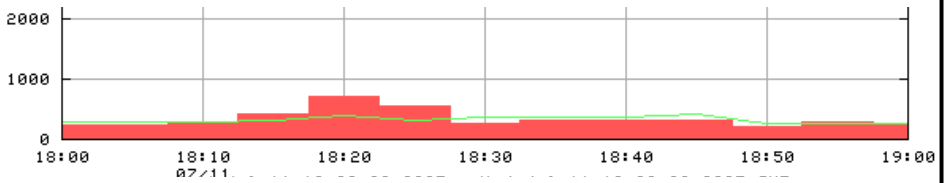


7 chislsdn1-oc192-snv2sdn1.es.net (GRAPH OMITTED)
8 chiccr1-chislsdn1.es.net

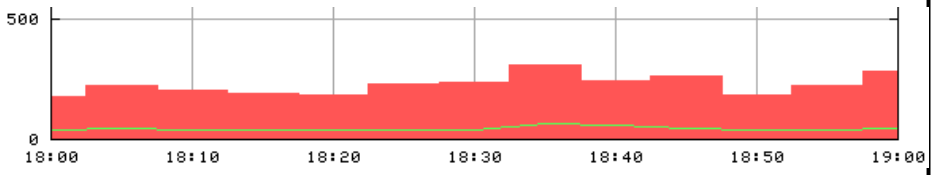


9 aofacr1-chicsdn1.es.net (GRAPH OMITTED)

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

Conclusions (from the ESnet Point of View)

- The usage of, and demands on, ESnet (and similar R&E networks) are expanding significantly as large-scale science becomes increasingly dependent on high-performance networking
- The motivation for the next generation of ESnet is derived from observations of the current traffic trends and case studies of major science applications
- The case studies of the science uses of the network lead to an understanding of the new uses of the network that will be required
- These new uses require that the network provide new capabilities and migrate toward network communication as a service-oriented capability.

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. Science Case for Large Scale Simulation, June 2003
   - http://www.pnl.gov/scales/
5. Planning Workshops-Office of Science Data-Management Strategy, March & May 2004
   - http://www-conf.slac.stanford.edu/dmw2004
6. For more information contact Chin Guok (chin@es.net). Also see
   - http://www.es.net/oscars

[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