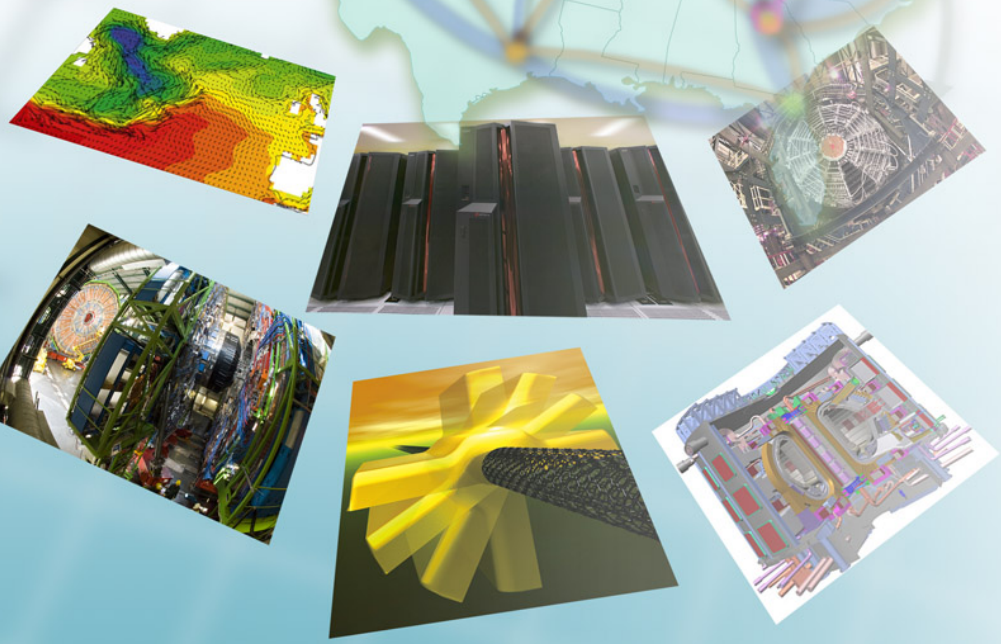
Network Architecture and Services to Support Large-Scale Science: An ESnet Perspective

ESnet
Energy Sciences Network



Joint Techs
January, 2026

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

Networking for the Future of Science

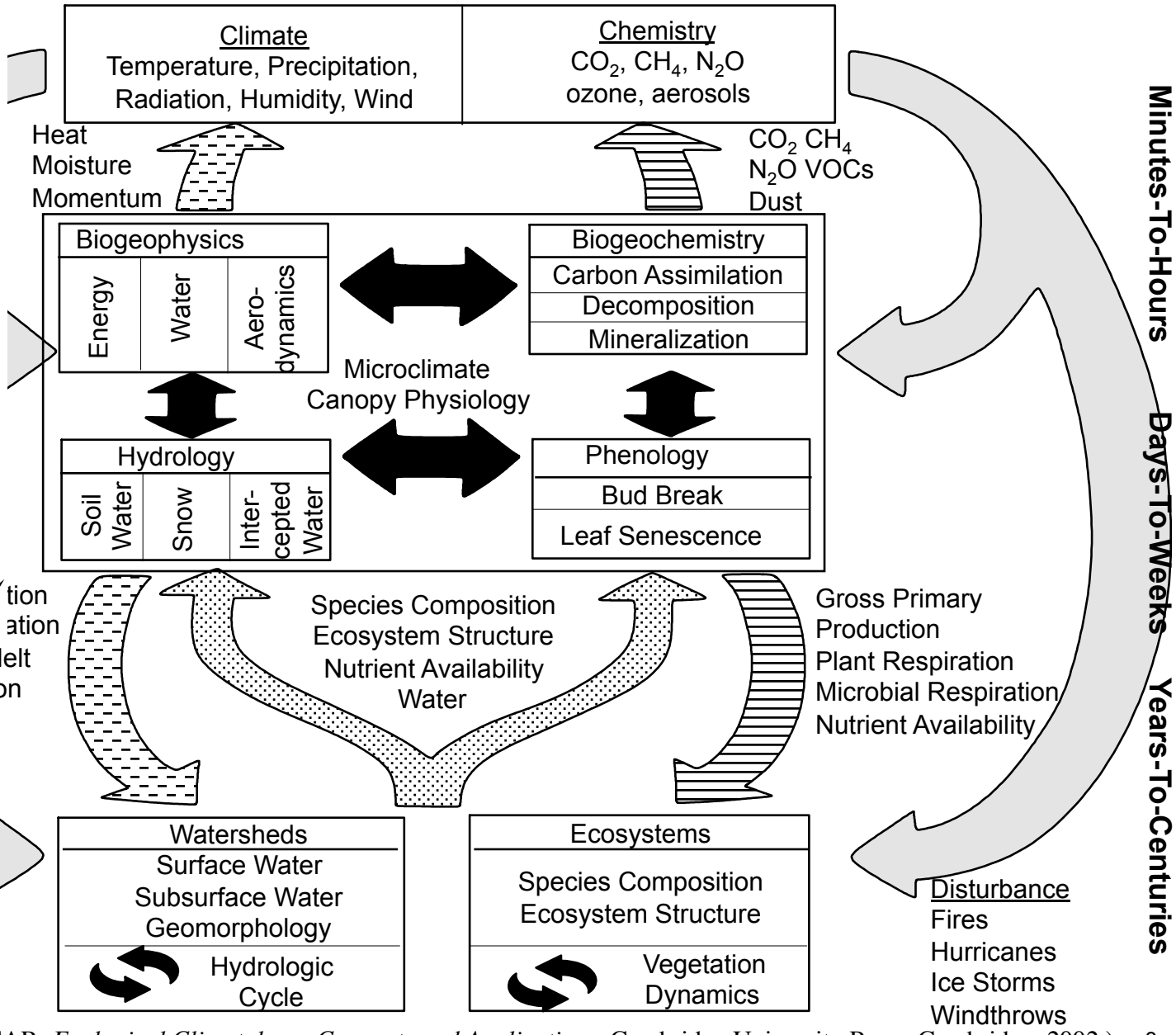
Berkeley Lab
Office of Science
U.S. Department of Energy

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
- ***Distributed data analysis and simulation is the emerging approach*** for these complex problems
- 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

A "*Systems of Systems*" Approach for Distributed 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.)

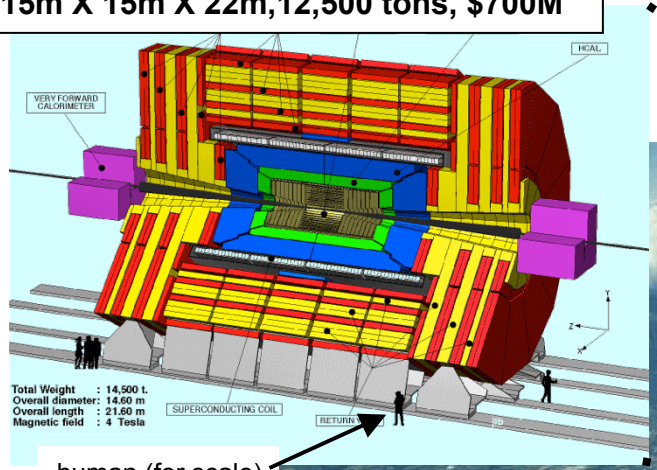
Large-Scale Science: High Energy Physics' Large Hadron Collider (Accelerator) at CERN

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

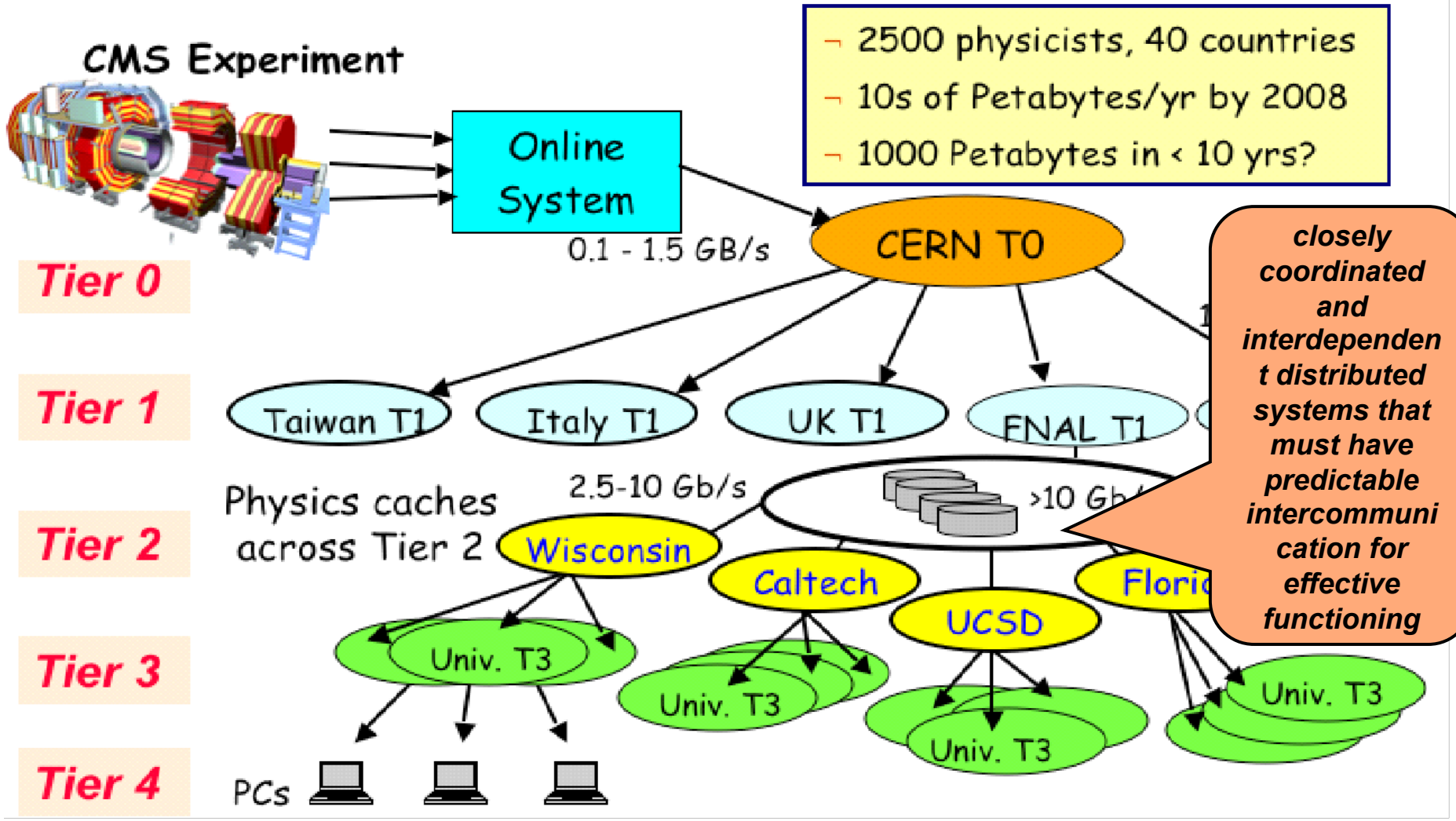


human (for scale)

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

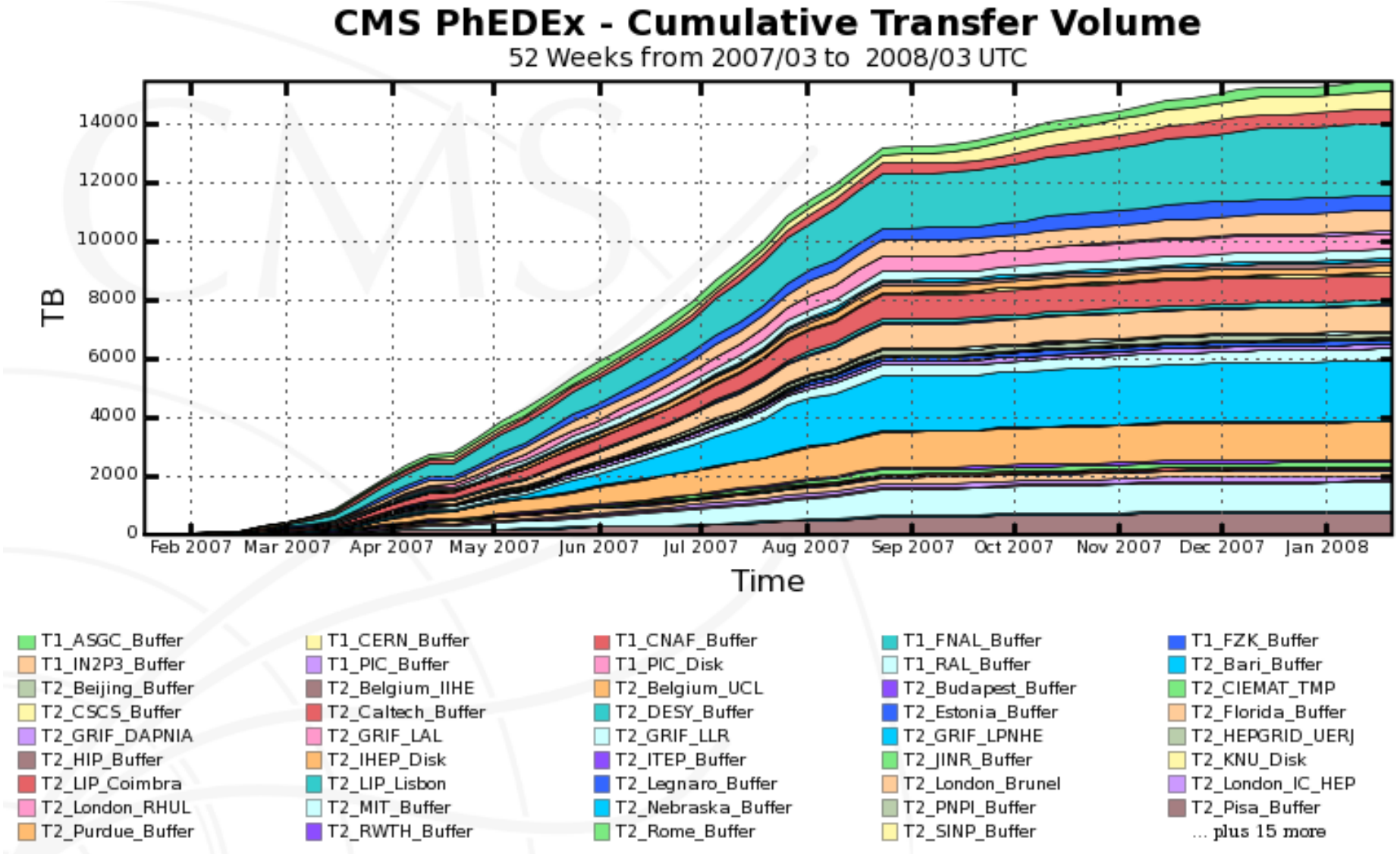
Two counter-rotating, 7 TeV proton beams, 27 km circumference (8.6 km diameter), collide in the middle of the detectors

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]

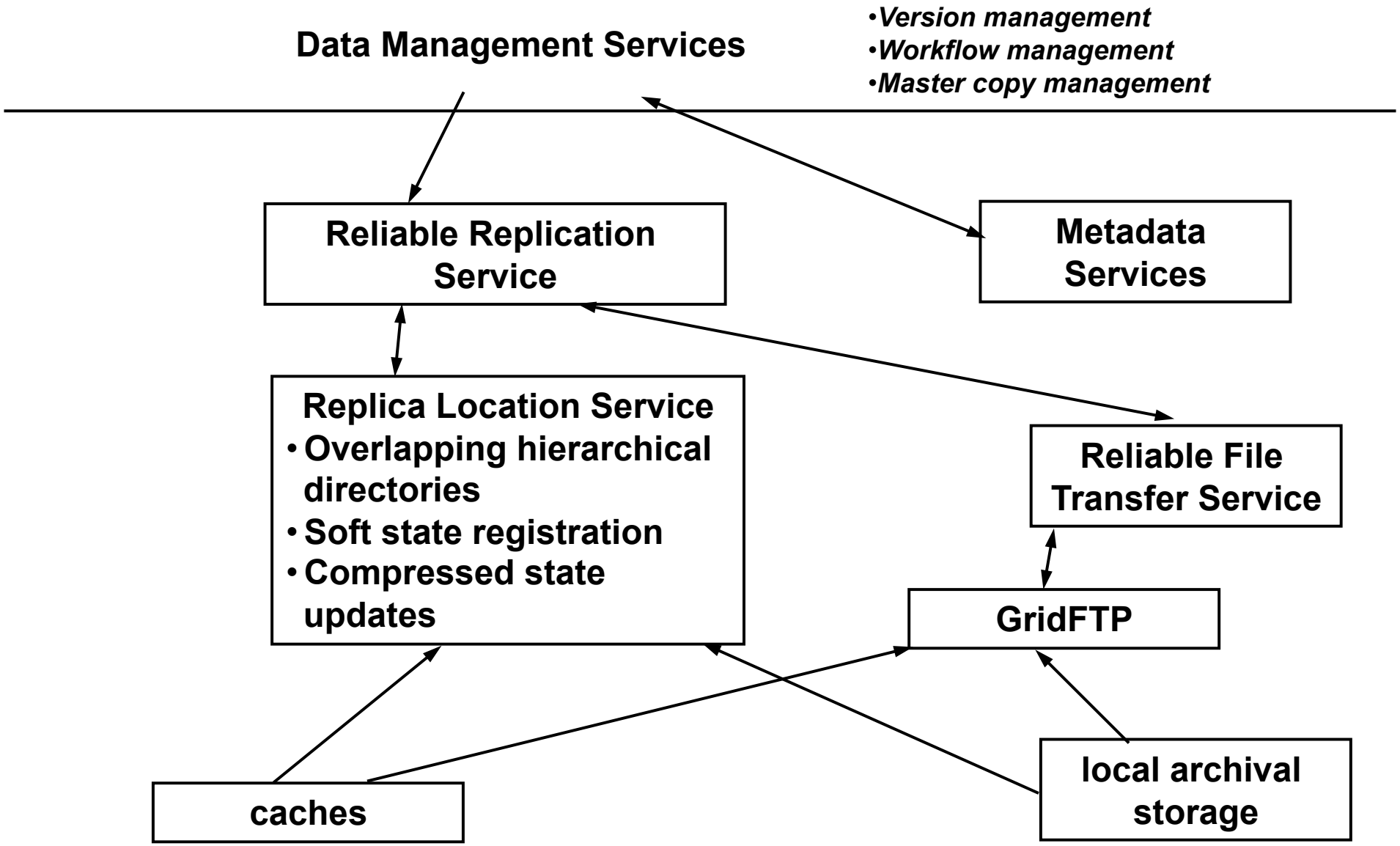


Accumulated data (Terabytes) received by CMS Data Centers ("tier1" sites) and many analysis centers ("tier2" sites) during the past 12 months (15 petabytes of data) [LHC/CMS]

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



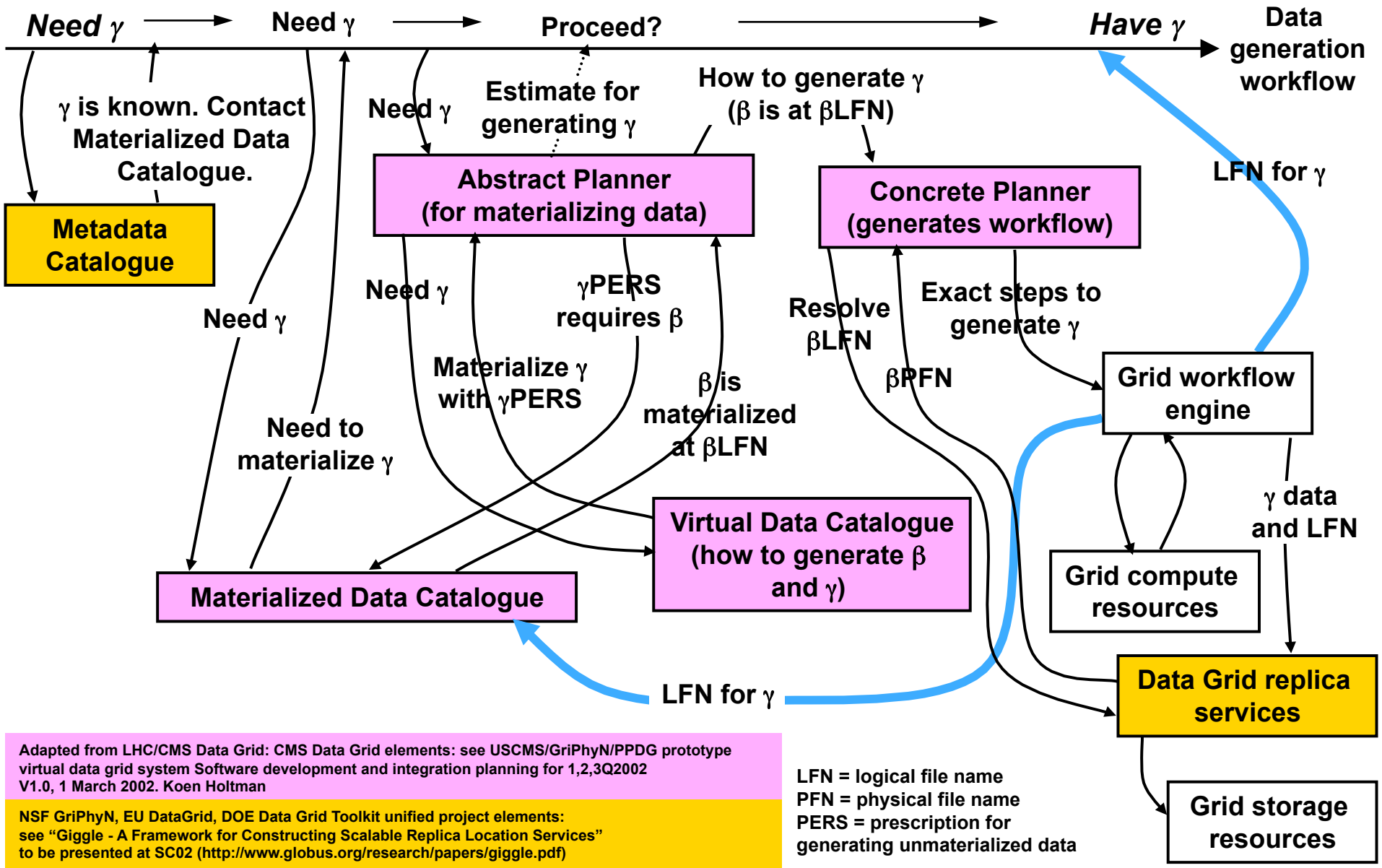
"Service Oriented Architecture" Data Management Service



See: "Giggle: Framework for Constructing Scalable Replica Location Services." Chervenak, et al. http://www.globus.org/research/papers/giggle.pdf

Workflow View of a Distributed Data Management Service

Elements of a Service Oriented Architecture application may *interact in complex ways that make reliable communication service important to the overall functioning of the system*



Adapted from LHC/CMS Data Grid: CMS Data Grid elements: see USCMS/GriPhyN/PPDG prototype virtual data grid system Software development and integration planning for 1,2,3Q2002 V1.0, 1 March 2002. Koen Holtman

NSF GriPhyN, EU DataGrid, DOE Data Grid Toolkit unified project elements: see "Giggle - A Framework for Constructing Scalable Replica Location Services" to be presented at SC02 (http://www.globus.org/research/papers/giggle.pdf)

LFN = logical file name
PFN = physical file name
PERS = prescription for generating unmaterialized data

Service Oriented Architecture / Systems of Systems

- Two types of systems seem to be likely

  1) Where the *components are them selves standalone elements* that are frequently used that way, but that can also be integrated into the types of systems implied by the complex climate modeling example

  2) Where the elements are normally used integrated into a distributed system, but the elements of the system are distributed because of compute, storage, or data resource availability

     - this is the case with the high energy physics data analysis

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**
- **informative**
- **and the network and its services must be scalable and geographically comprehensive**

Networks Must Provide Communication Capability that is Service-Oriented

- Configurable
  - Must be able to provide multiple, specific “paths” (specified by the user as end points) with specific characteristics
- Schedulable
  - Premium service such as guaranteed bandwidth will be a scarce resource that is not always freely available, therefore time slots obtained through a resource allocation process must be schedulable
- Predictable
  - A committed time slot should be provided by a network service that is not brittle - reroute in the face of network failures is important
- Reliable
  - Reroutes should be largely transparent to the user
- Informative
  - When users do system planning they should be able to see average path characteristics, including capacity
  - When things do go wrong, the network should report back to the user in ways that are meaningful to the user so that informed decisions can about alternative approaches
- Scalable
  - The underlying network should be able to manage its resources to provide the appearance of scalability to the user
- Geographically comprehensive
  - The R&E network community must act in a coordinated fashion to provide this environment end-to-end

The ESnet Approach

- Provide *configurability, schedulability, predictability*, and reliability with a flexible virtual circuit service - OSCARS
  - User* specifies end points, bandwidth, and schedule
  - OSCARS can do fast reroute of the underlying MPLS paths
- Provide useful, comprehensive, and meaningful *information on the state of the paths*, or potential paths, to the user
  - perfSONAR, and associated tools, provide real time information in a form that is useful to the user (via appropriate network abstractions) and that is delivered through standard interfaces that can be incorporated in to SOA type applications
  - Techniques need to be developed to monitor virtual circuits based on the approaches of the various R&E nets - e.g. MPLS in ESnet, VLANs, TDM/grooming devices (e.g. Ciena Core Directors), etc., and then integrate this into a perfSONAR framework

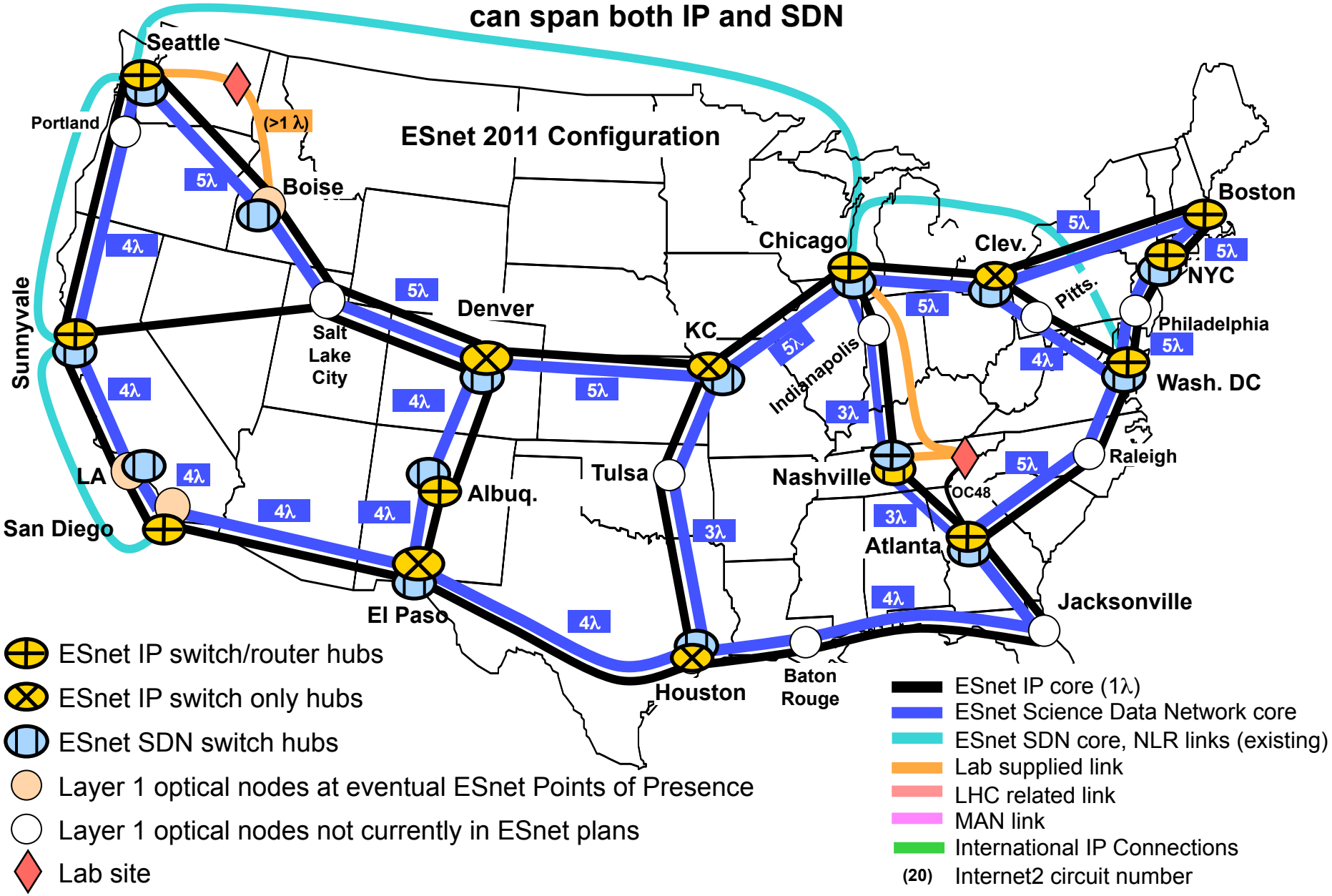
* User = human or system component (process)

The ESnet Approach

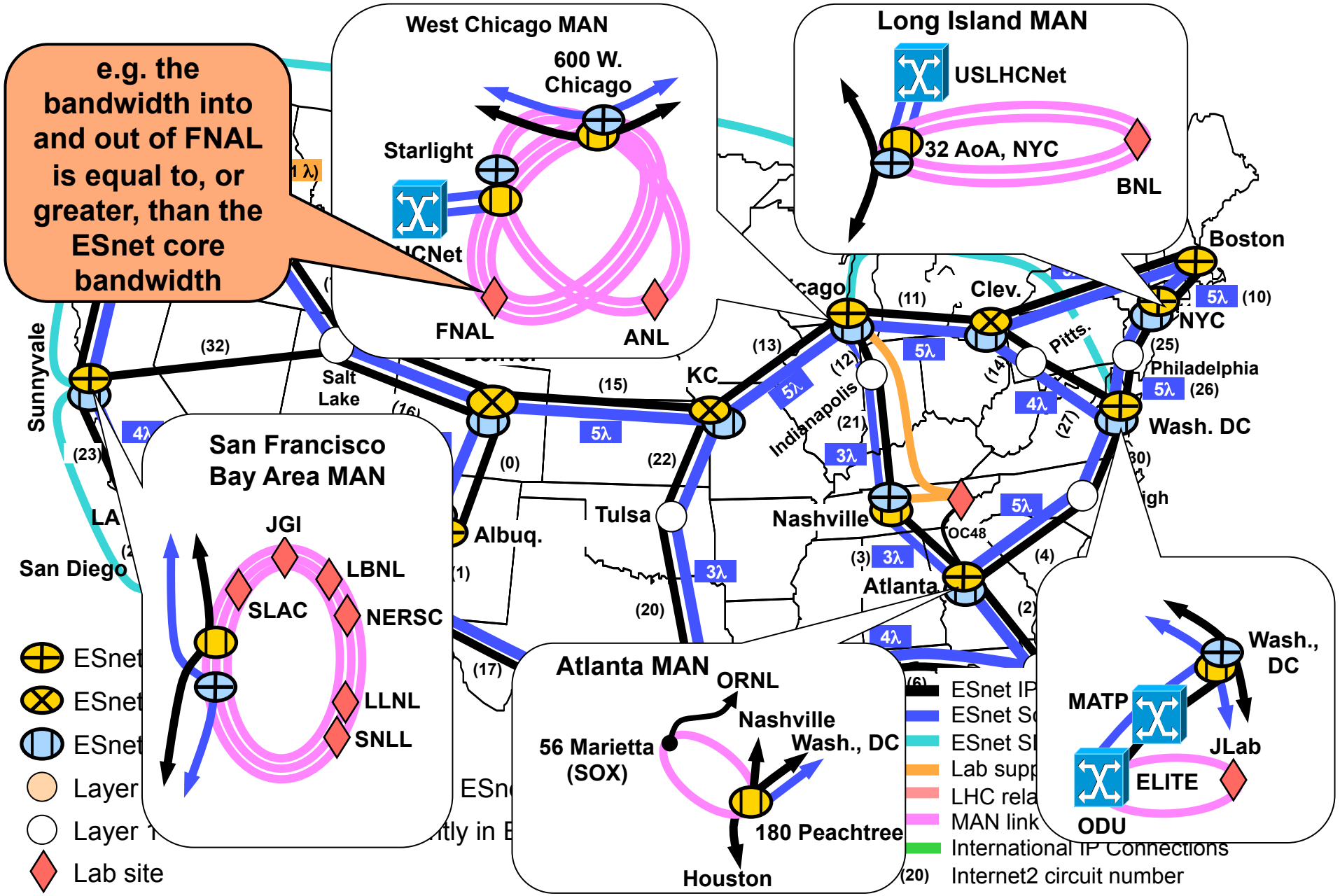
- *Scalability* will be provided by new network services that, e.g., provide dynamic wave allocation at the optical layer of the network
  - Currently an R&D project
- *Geographic ubiquity* of the services can only be accomplished through active collaborations in the global R&E network community so that all sites of interest to the science community can provide compatible services for forming end-to-end virtual circuits
  - Active and productive collaborations exist among numerous R&E networks: ESnet, Internet2, CANARIE, DANTE/GÉANT, some European NRENs, some US regionals, etc.

1) Network Architecture Tailored to Circuit-Oriented Services

ESnet4 is a hybrid network: IP + L2/3 Science Data Network (SDN) - OSCARS circuits can span both IP and SDN



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

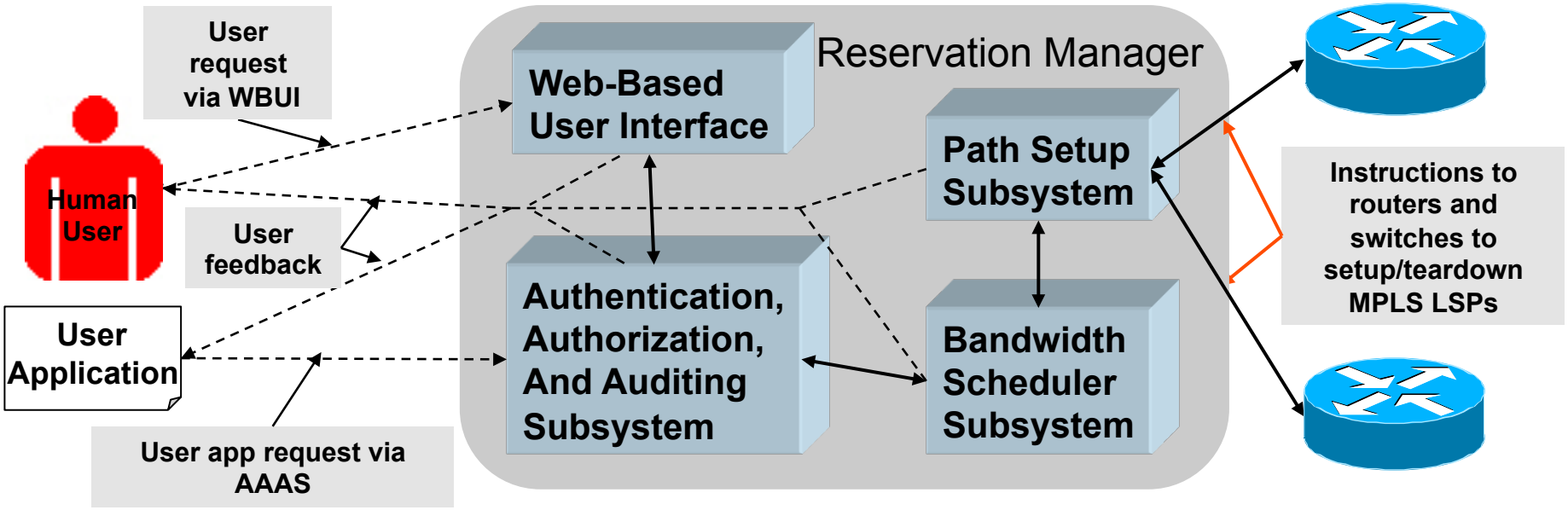


2) Multi-Domain Virtual Circuits

ESnet OSCARS [OSCARS] 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
  - GÉANT: 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) perfSONAR Monitoring Applications 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]

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

Paths are not always up, of course - especially international paths that may not have an easy alternative path

Status of E2E Link FERMI-IN2P3-IGTMD-002

Oper. State: Down
Admin. State: Normal Oper.
Error: E2E Link is **not** contiguous (End Point missing or gap found)
Warning: Operational state is known not for all involved links
Warning: Administrative state is known not for all involved links

IN2P3					GEANT2			
	DP		EP			DP		DP
ID Part.Info	Demarc	Gap	EndPoint	ID Part.Info	ID Part.Info	Demarc	Domain Link	Demarc
-	**GEANT2-NY**	-	**IN2P3-IGTMD2**	IN2P3-RENATER_LYO_FERMI2	RENATER-LYO-FERMI2	**RENATER-LYO**	RENATER-PAR-LYO-02	**RENATER-PAR**
-	-	-	-	Down	Up	-	Up	-
-	-	-	-	Normal Oper.	Normal Oper.	-	Normal Oper.	-
-	-	-	-	2008-01-20 T01:06:01.0-6:00	2008-1-20 T2:9:01.0+0000	-	2008-1-20 T2:9:01.0+0000	-

[http://lhcopnmon1.fnal.gov:9090/FERMI-E2E/G2_E2E_view_e2elink_FERMI-IN2P3-IGTMD-002.html]

Path Performance Monitoring

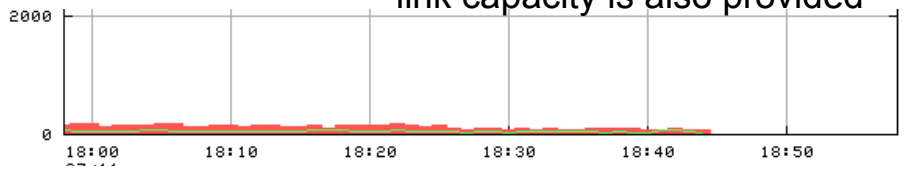
- Path performance monitoring needs to provide users/applications with the end-to-end, multi-domain traffic and bandwidth availability
  - should also provide real-time performance such as path utilization and/or packet drop
- Multiple path performance monitoring tools are in development
  - One example – 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

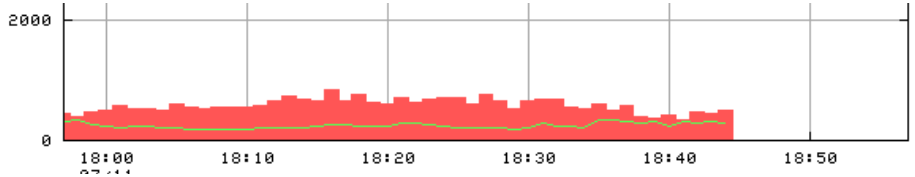
- Forward direction bandwidth utilization on application path from LBNL to INFN-Frascati (Italy)
  - traffic shown as bars on those network device interfaces that have an associated MP services (the first 4 graphs are normalized to 2000 Mb/s, the last to 500 Mb/s)

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

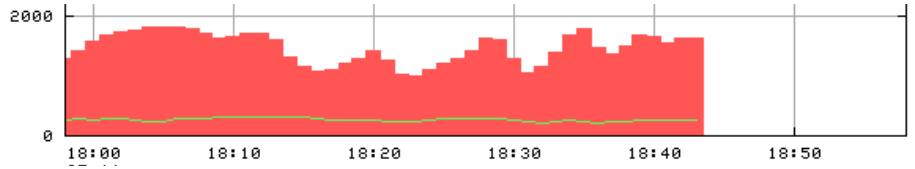
link capacity is also provided



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

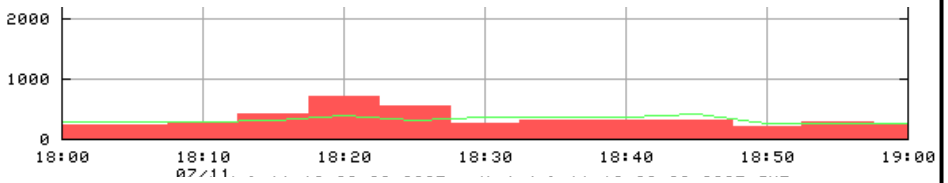


7 chislsdn1-oc192-snv2sdn1.es.net (GRAPH OMITTED)
8 chiccr1-chislsdn1.es.net

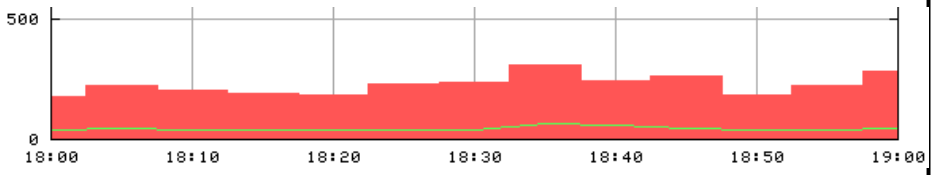


9 aofacr1-chicsdn1.es.net (GRAPH OMITTED)

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

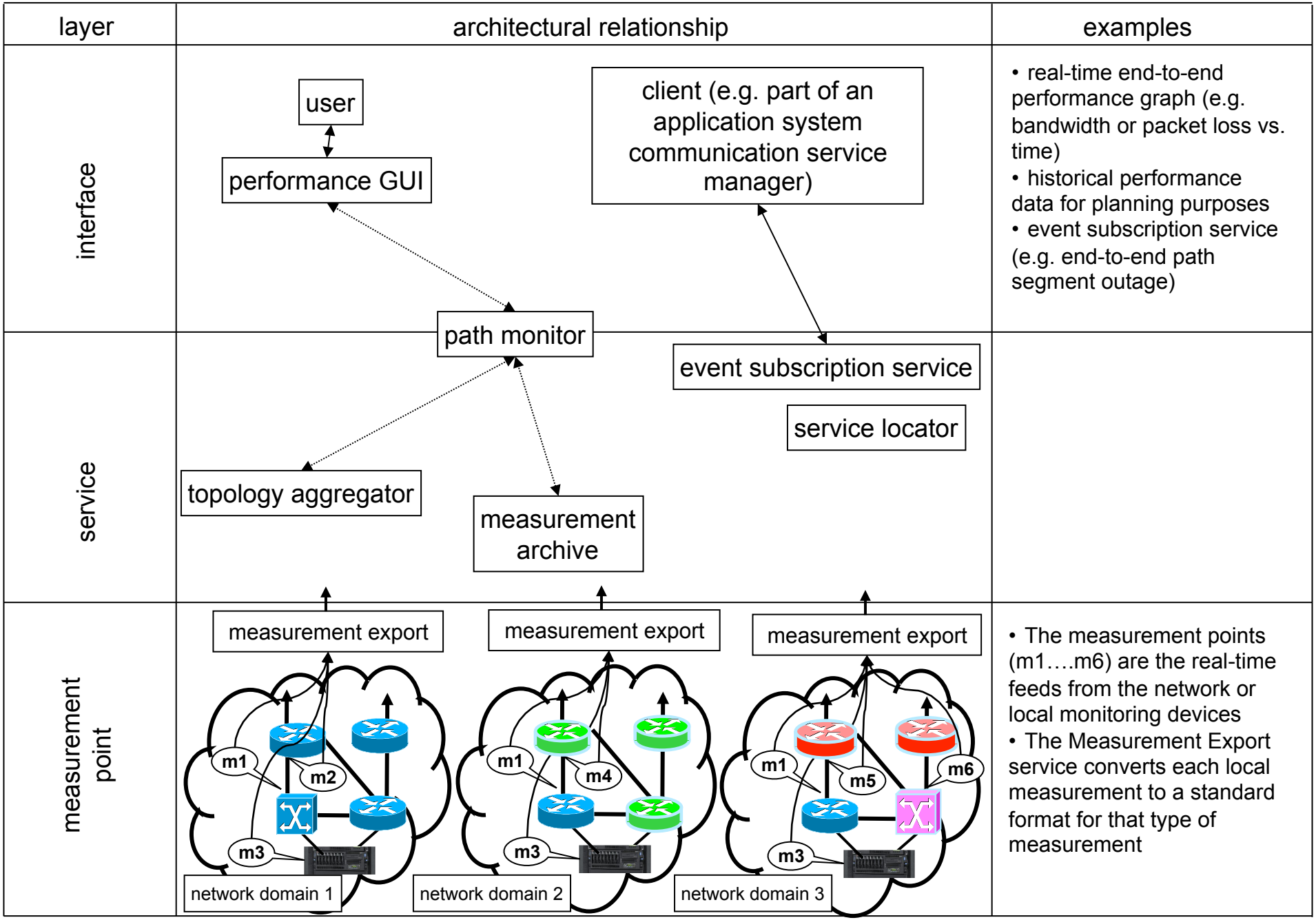


17 rt-mi2-rt-rm2.rm2.garr.net (GRAPH OMITTED)
18 rt-rm2-rc-fra.fra.garr.net (GRAPH OMITTED)
19 rc-fra-ru-lnf.fra.garr.net (GRAPH OMITTED)

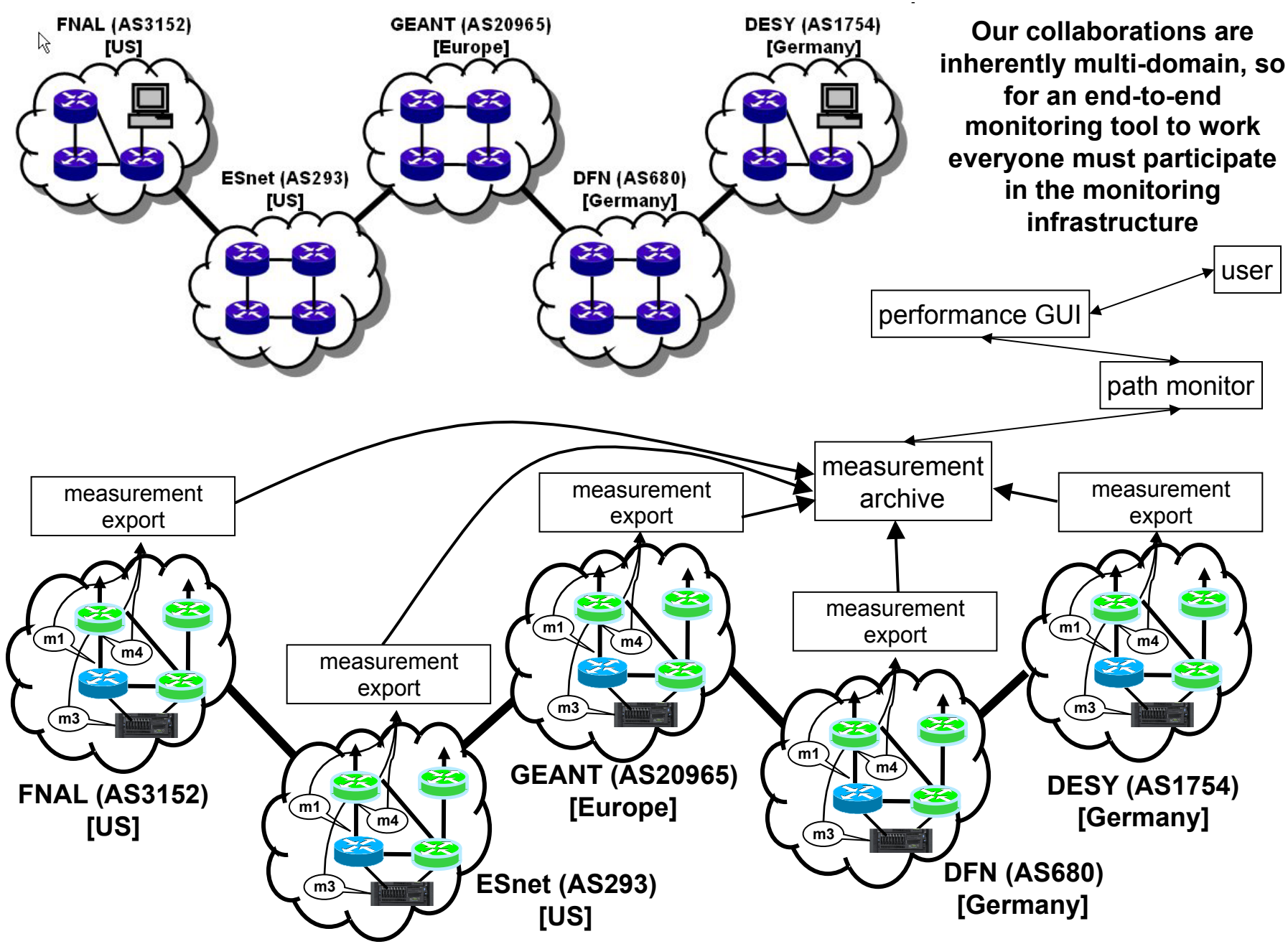


20
21 www6.lnf.infn.it (193.206.84.223) 189.908 ms 189.596 ms 189.684 ms

perfSONAR architecture



perfSONAR Only Works E2E When All Networks Participate



Conclusions

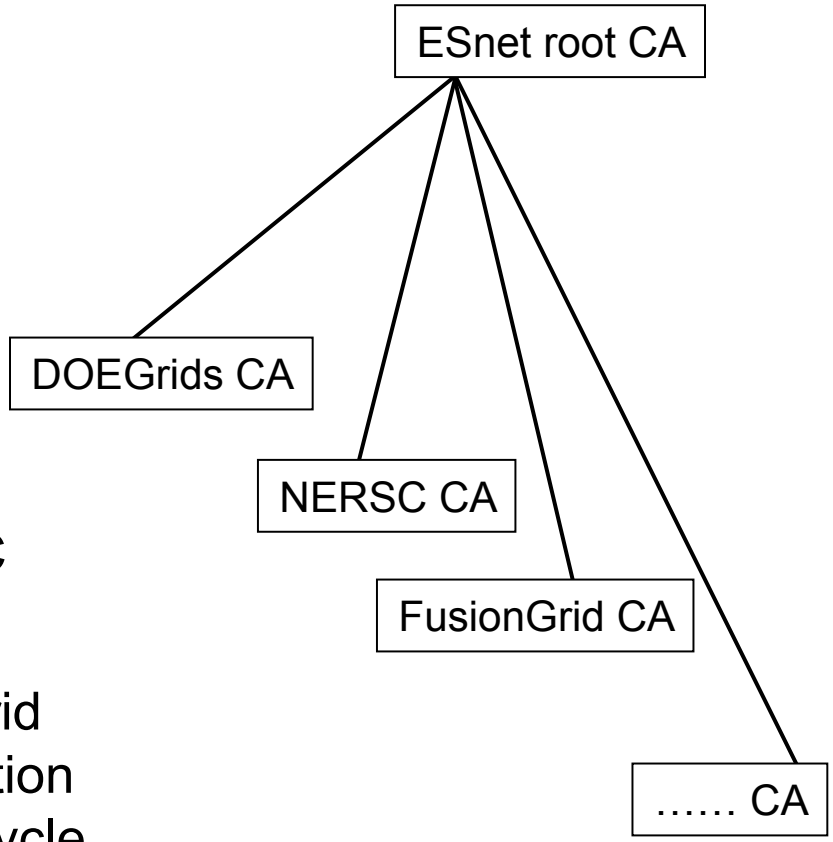
- To meet the existing overall bandwidth requirements of large-scale science *networks must deploy adequate infrastructure*
  - mostly on-track to meet this requirement
- To meet the emerging requirements of how large-scale science software system are built the *network community must provide new services* that allow the network to be a "service element" that can be integrated into a Service Oriented Architecture / System of Systems framework
  - progress is being made in this direction

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
  - Stats:
    - User certificates issued 5237
    - Host & service certificates issued 11704
    - Total no. of currently active certificates 6982



See www.doegrids.org

References

[OSCARS]

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?graph=quantity_cumulative&entity=src&src_filter=&dest_filter=&no_mss=true&period=l52w&upto=

[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://wiki.perfsonar.net/jra1- wiki/index.php/PerfSONAR_support_for_E2E_Link_Monitoring

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