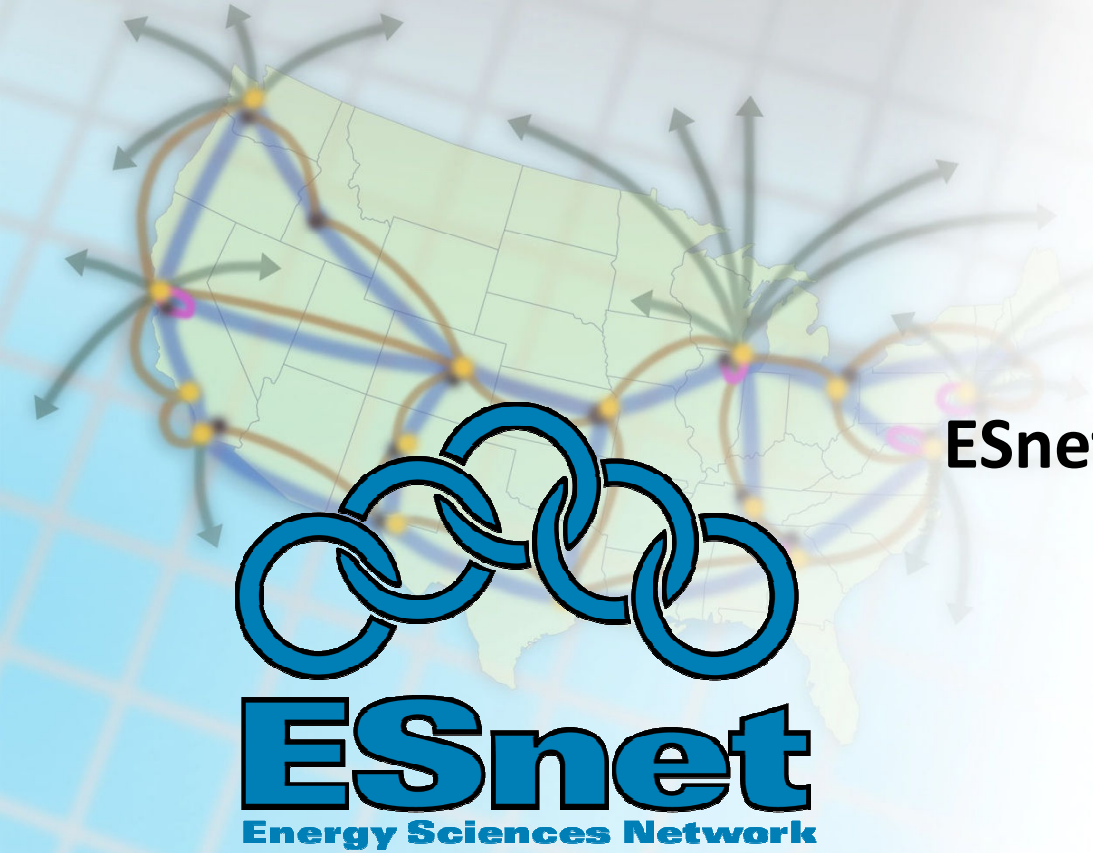


ESnet
Energy Sciences Network

Supporting Advanced Scientific Computing Research • Basic Energy Sciences • Biological and Environmental Research • Fusion Energy Sciences • High Energy Physics • Nuclear Physics

HEPiX Workshop

ESnet: Networking for Science

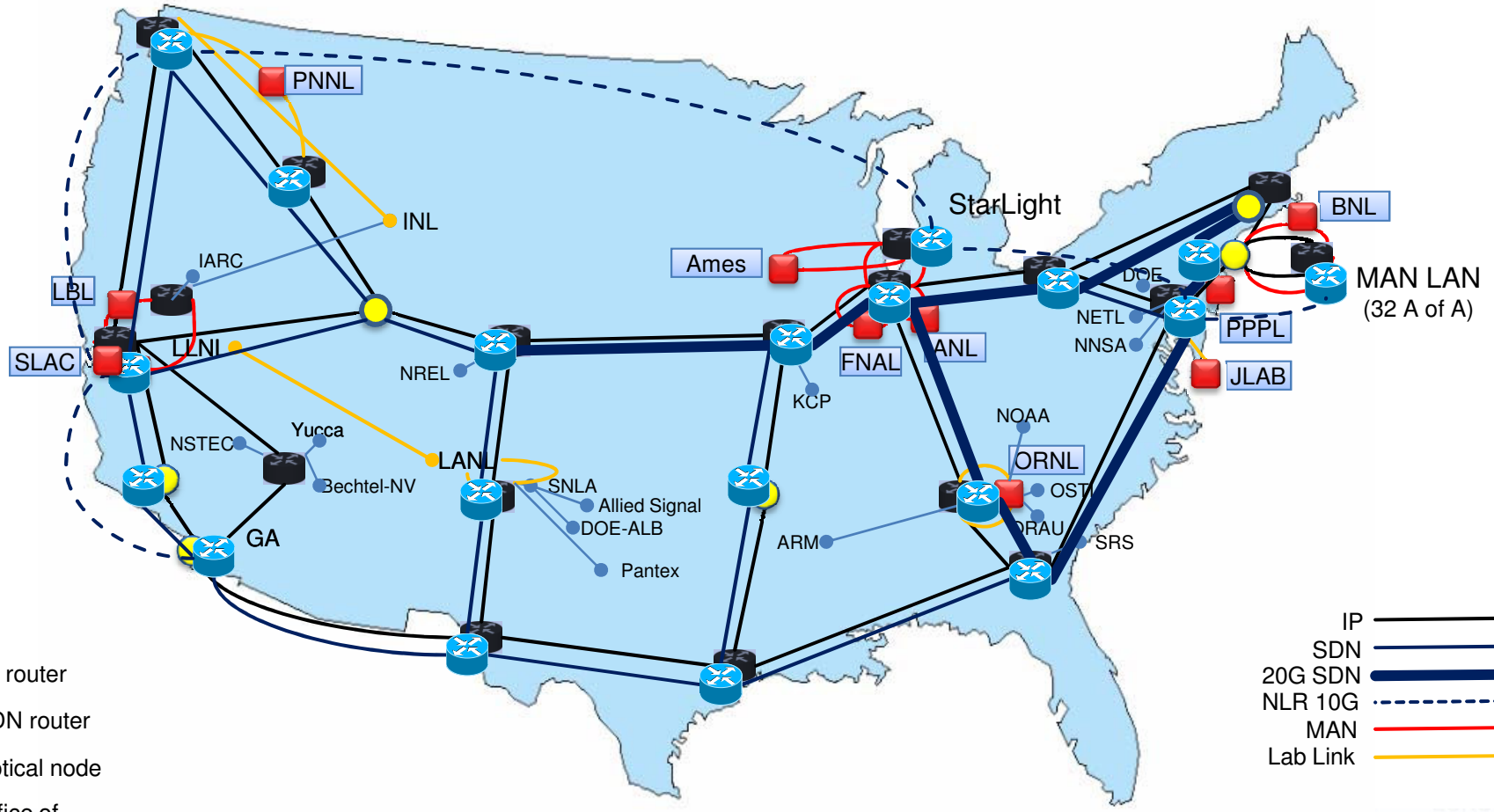
Oct 30, 2009

Joe Burrescia, General Manager
Energy Sciences Network
Lawrence Berkeley National Lab

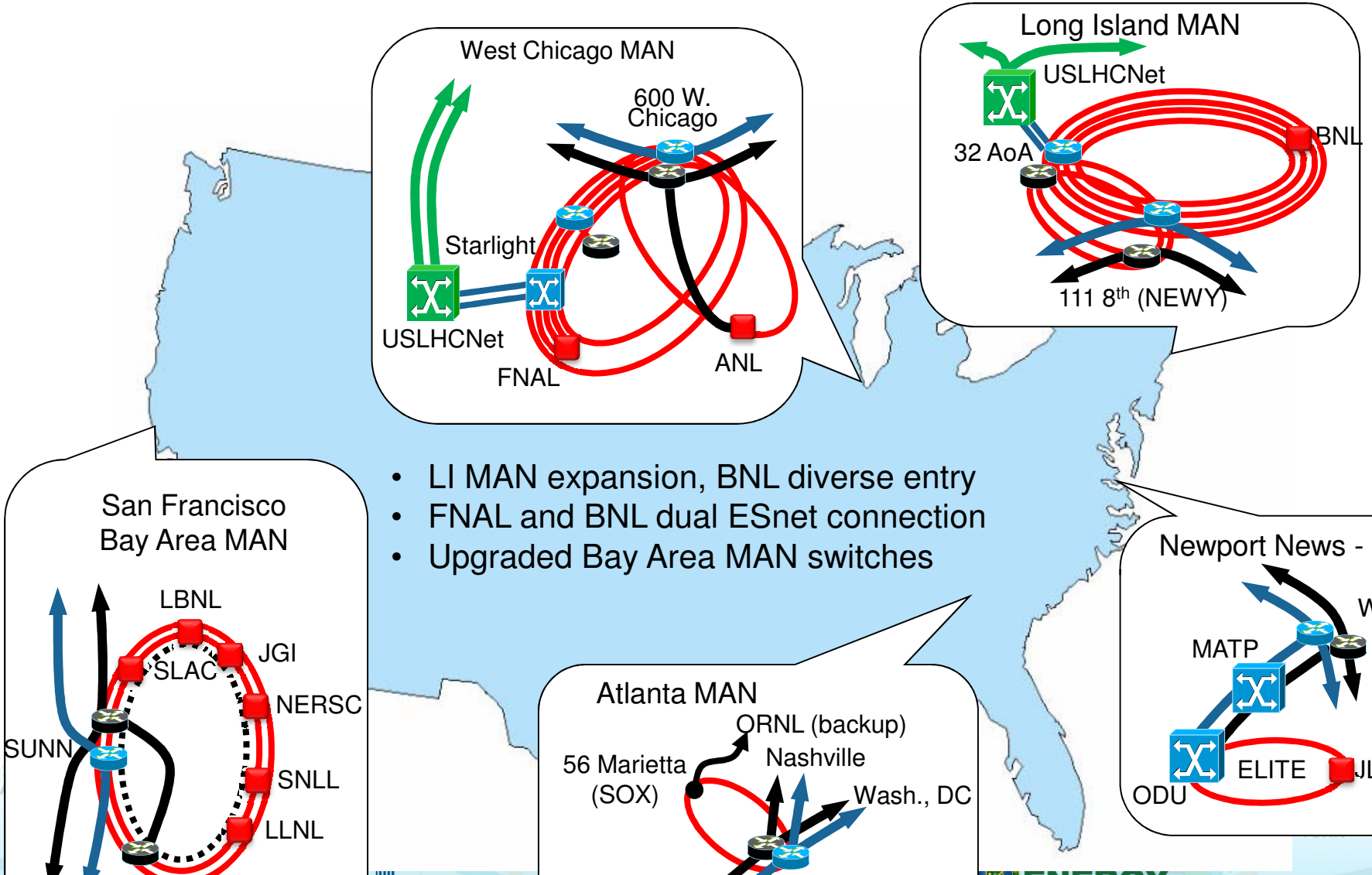
Berkeley Lab

U.S. Department of Energy
Office of Science

ESnet4 Topology Late 2009

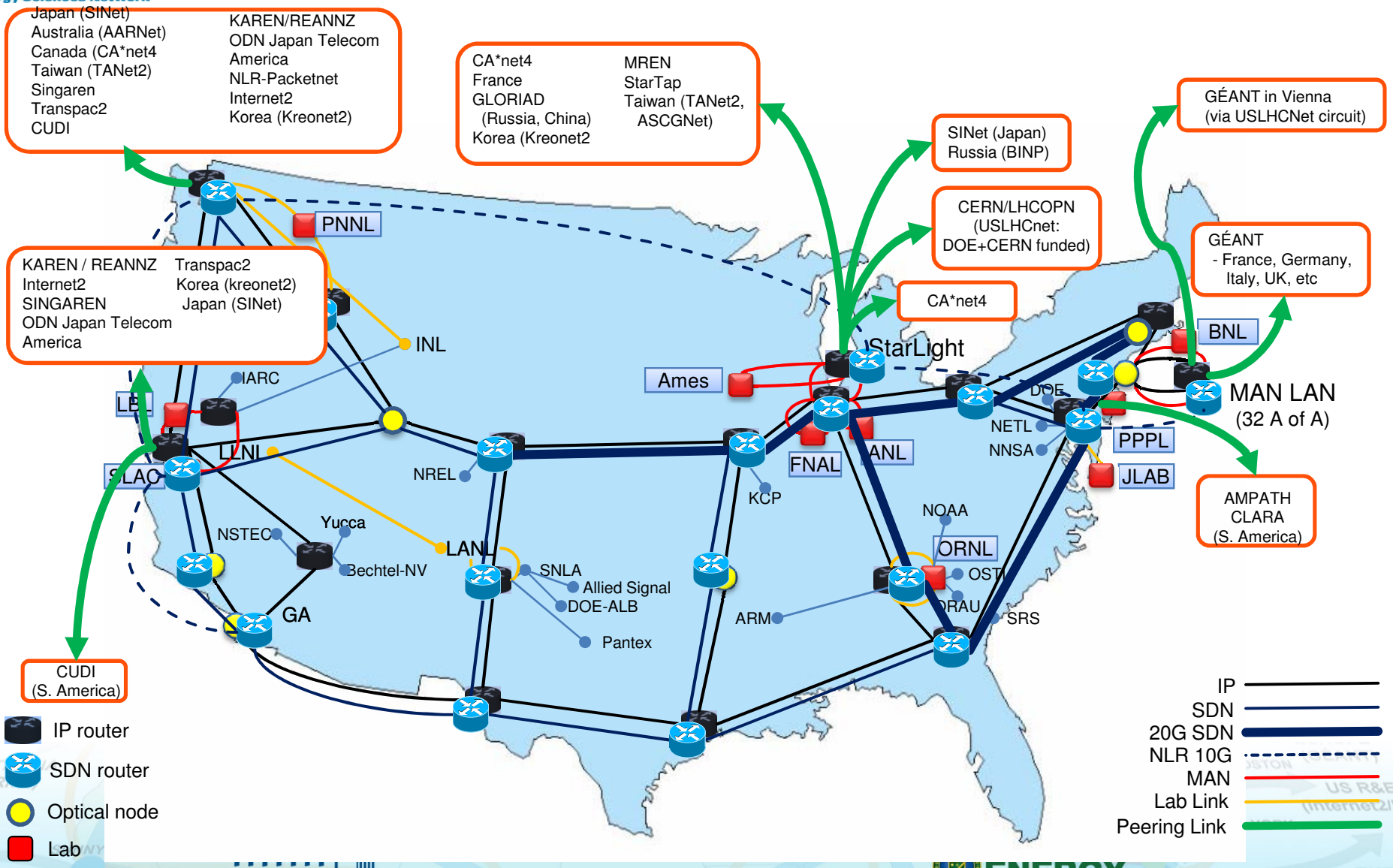


ESnet4 Metro Area Rings

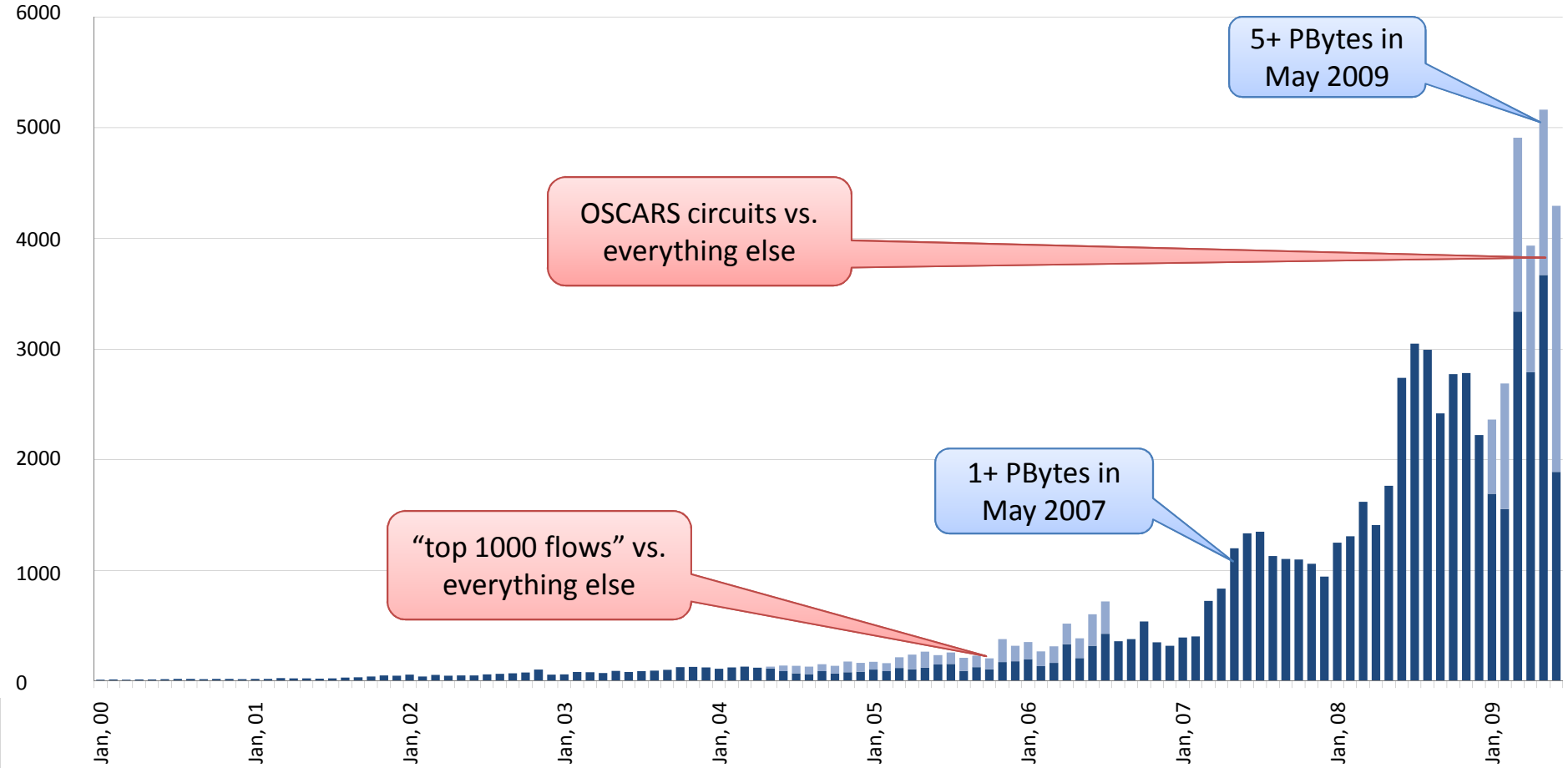


- LI MAN expansion, BNL diverse entry
- FNAL and BNL dual ESnet connection
- Upgraded Bay Area MAN switches

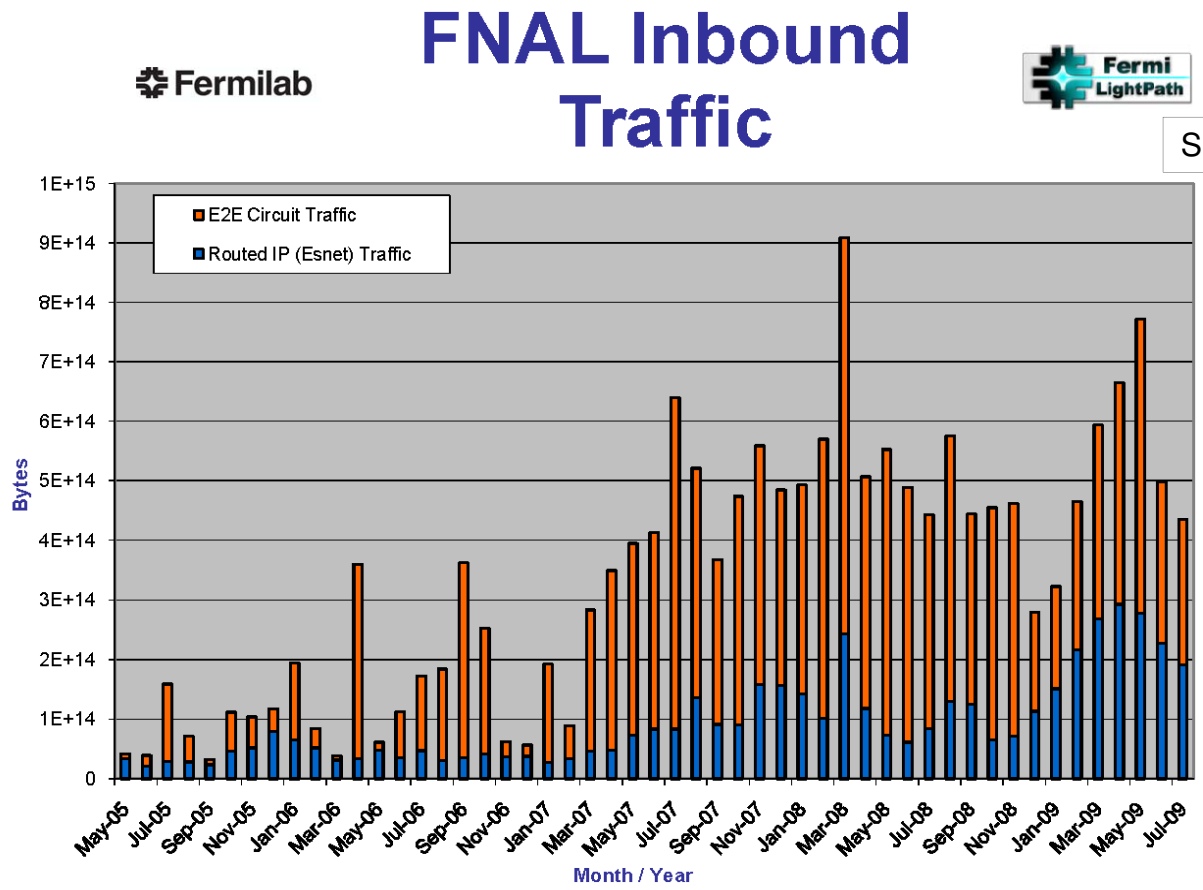
ESnet R&E Peerings Late 2009



ESnet Total Accepted Traffic, TBy/mo

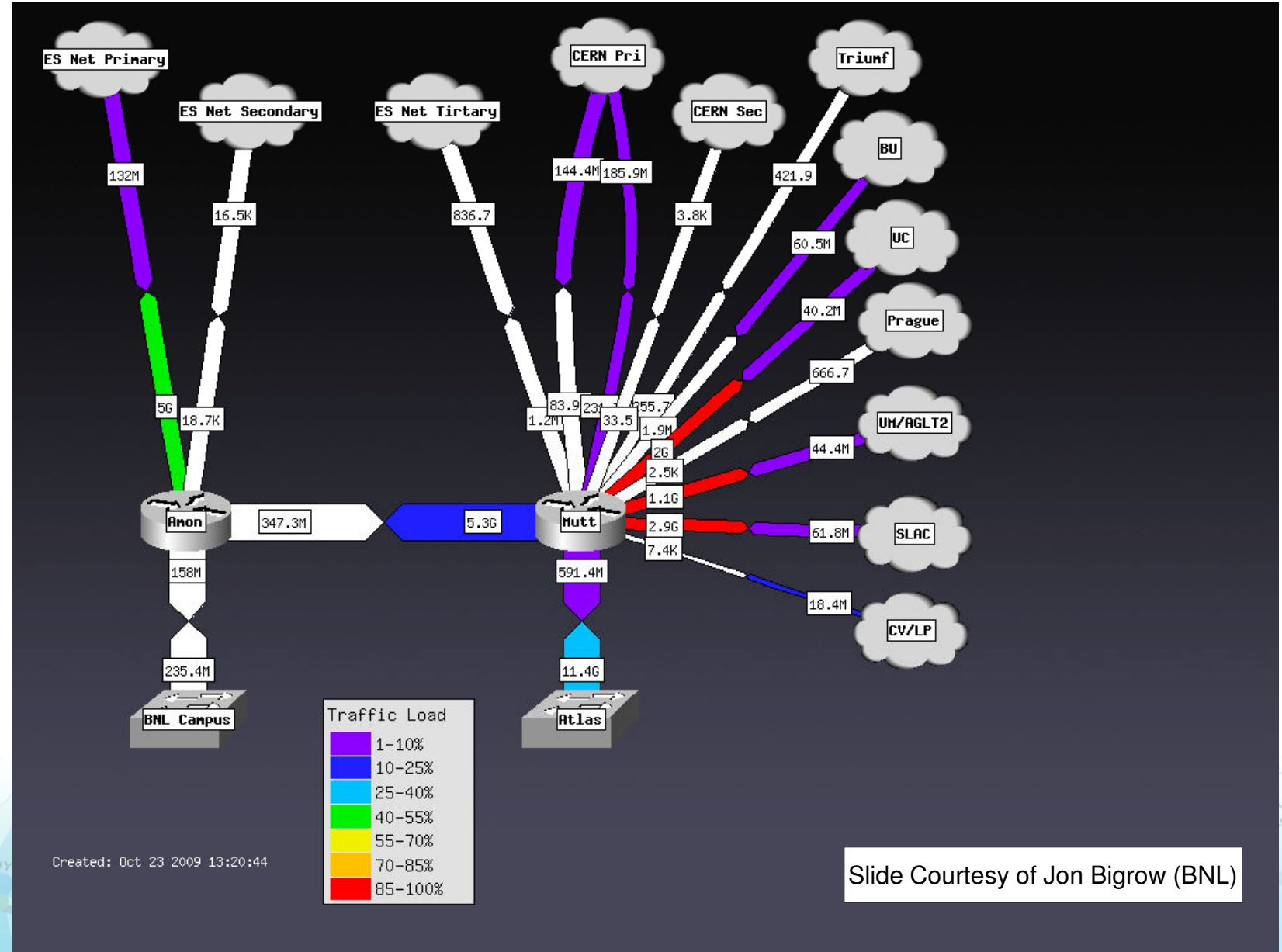


Growing Use of Circuits for Science Within ESnet



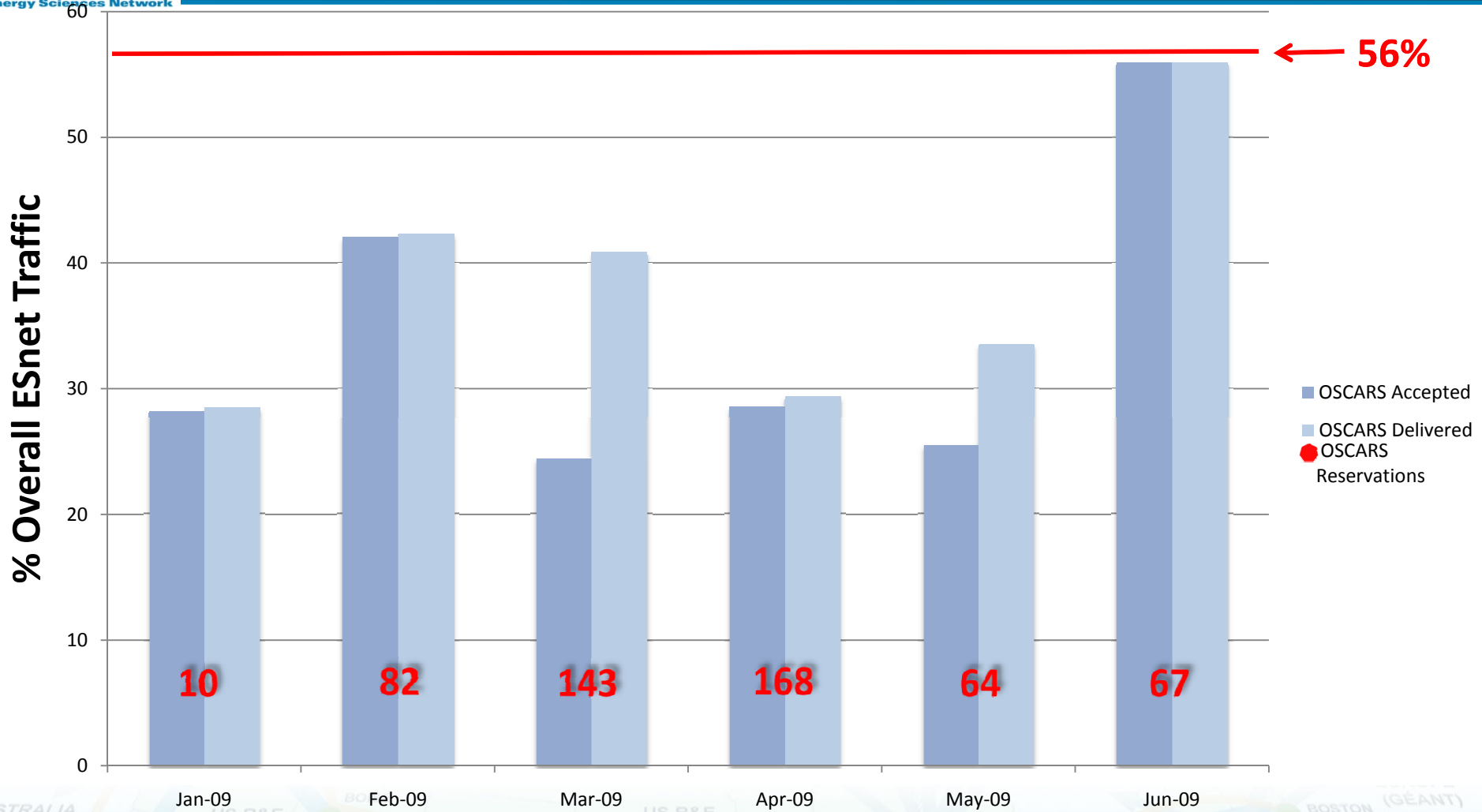
FNAL traffic over five years – IP traffic is a mix of science and business, the growing "LightPath" traffic is all science traffic

Instantaneous BNL Circuit Traffic



Slide Courtesy of Jon Bigrow (BNL)

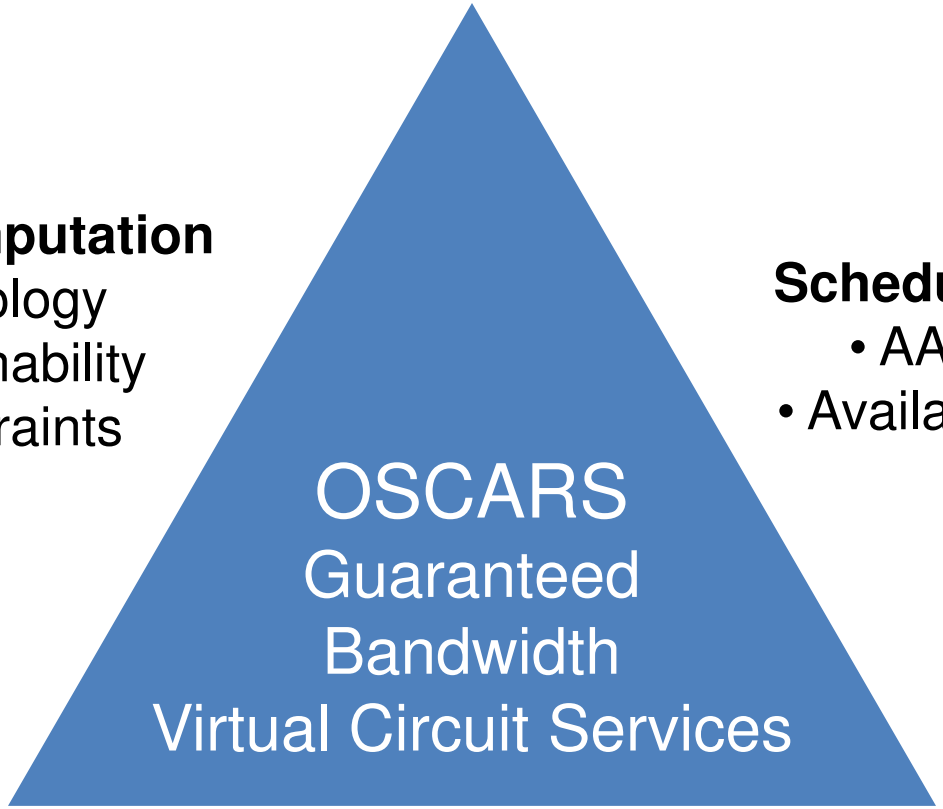
SDN - IP Traffic Breakdown



On-demand Secure Circuits and Advance Reservation System (OSCARS) Overview

Path Computation
- Topology
- Reachability
- Contraints

Scheduling
- AAA
- Availability



Provisioning
- Signaling
- Security
- Resiliency/Redundancy

OSCARS Design Goals

- **Configurable**
  - The circuits must be dynamic and driven by user requirements (e.g. termination end-points, required bandwidth, etc)
- **Schedulable**
  - Premium service such as guaranteed bandwidth will be a scarce resource that is not always freely available and therefore should be obtained through a resource allocation process that is schedulable
- **Predictable**
  - The service should provide circuits with predictable properties (e.g. bandwidth, duration, etc) that the user can leverage.
- **Usable**
  - The service must be easy to use by the target community
- **Reliable**
  - Resiliency strategies (e.g. reroutes) should be largely transparent to the user
- **Informative**
  - The service should provide useful information about reserved resources and circuit status to enable the user to make intelligent decisions
- **Scalable**
  - The underlying network should be able to manage its resources to provide the appearance of scalability to the user
  - The service should be transport technology agnostic (e.g. 100GE, DWDM, etc)
- **Geographically comprehensive**
  - The R&E network community must act in a coordinated fashion to provide this environment end-to-end
- **Secure**
  - The user must have confidence that both ends of the circuit is connected to the intended termination points, and that the circuit cannot be “hijacked” by a third party while in use
- **Provide traffic isolation**
  - Users want to be able to use non-standard/aggressive protocols when transferring large amounts of data over long distances in order to achieve high performance and maximum utilization of the available bandwidth

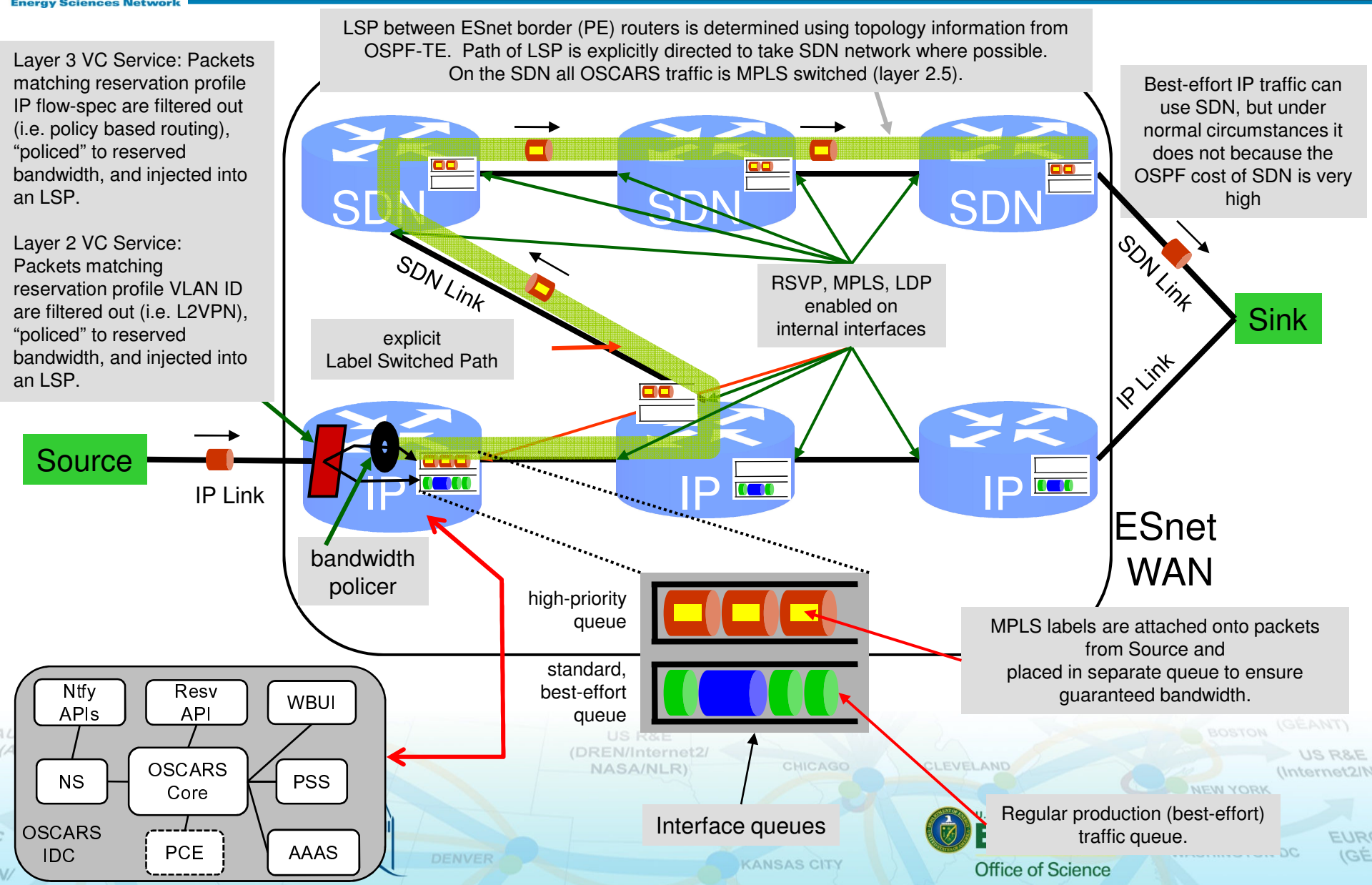
Network Mechanisms Underlying OSCARS

LSP between ESnet border (PE) routers is determined using topology information from OSPF-TE. Path of LSP is explicitly directed to take SDN network where possible. On the SDN all OSCARS traffic is MPLS switched (layer 2.5).

Layer 3 VC Service: Packets matching reservation profile IP flow-spec are filtered out (i.e. policy based routing), "policed" to reserved bandwidth, and injected into an LSP.

Layer 2 VC Service: Packets matching reservation profile VLAN ID are filtered out (i.e. L2VPN), "policed" to reserved bandwidth, and injected into an LSP.

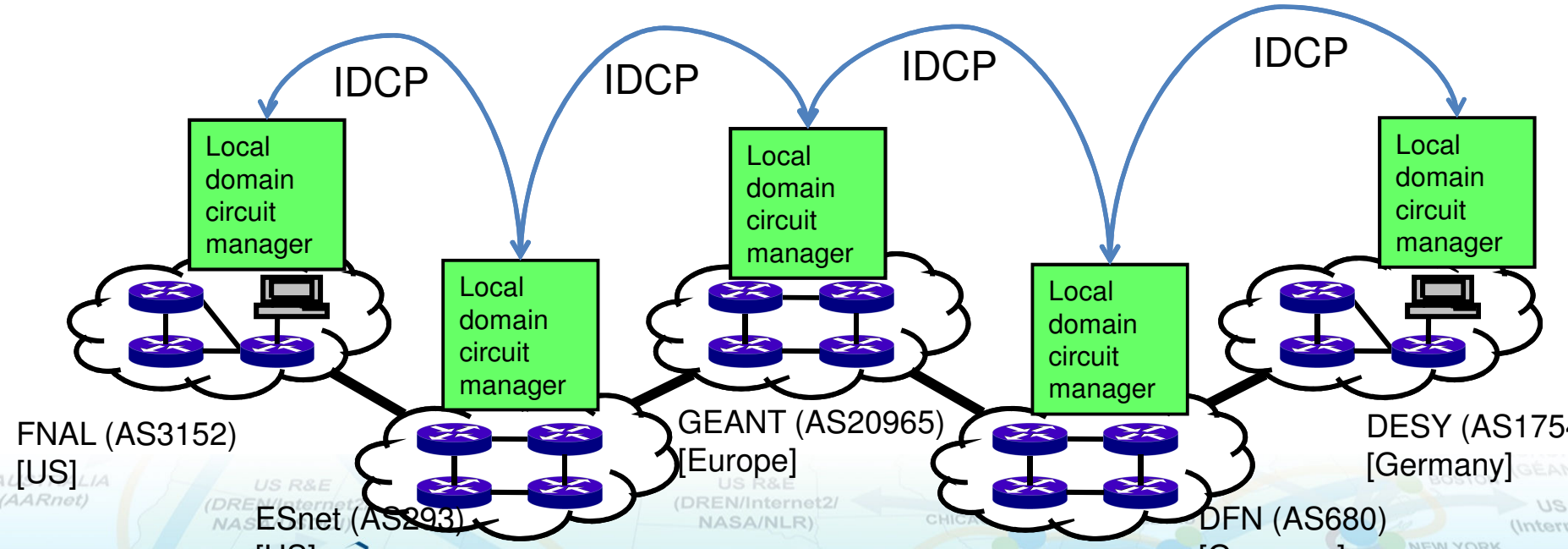
Best-effort IP traffic can use SDN, but under normal circumstances it does not because the OSPF cost of SDN is very high



Inter-Domain Virtual Circuits

In order to set up end-to-end circuits across multiple domains without violating security or allocation management policy of any of the domains, the process of setting up end-to-end circuits is handled by the Inter-Domain Control Protocol (IDCP) running between domains that

1) Negotiates for bandwidth that is available for this purpose
2) Requests that the domains each establish the "meet-me" state so that the circuit in one domain can connect to a circuit in another domain (thus maintaining domain control over the circuits)



Berkeley Lab

U.S. Department of Energy
Office of Science

OSCARS is in Production Now

- OSCARS is currently being used to support production traffic
- Operational Virtual Circuit (VC) support
  - As of 10/2009, there are 26 long-term production VCs instantiated
    - 21 VCs supporting HEP
      - LHC T0-T1 (Primary and Backup)
      - LHC T1-T2
    - 3 VCs supporting Climate
      - GFDL
      - ESG
    - 2 VCs supporting Computational Astrophysics
      - OptiPortal
- Short-term dynamic VCs
    - Between 1/2008 and 10/2009, there were roughly 4600 successful VC reservations
      - 3000 reservations initiated by BNL using TeraPaths
      - 900 reservations initiated by FNAL using LambdaStation
      - 700 reservations initiated using Phoebus

- **The adoption of OSCARS as an integral part of the ESnet4 network was a core contributor to ESnet winning the Excellence.gov "Excellence in Leveraging Technology" award given by the Industry Advisory Council's (IAC) Collaboration and Transformation Shared Interest Group (Apr 2009)**

OSCARS Collaborative Efforts

- As part of the OSCARS effort, ESnet worked closely with the DICE (DANTE, Internet2, CalTech, ESnet) Control Plane working group to develop the InterDomain Control Protocol (IDCP) which specifies inter-domain messaging for end-to-end VCs
- The following organizations have *implemented/deployed systems which are compatible with the DICP IDCP*:
  - Internet2 ION (OSCARS/DCN)
  - ESnet SDN (OSCARS/DCN)
  - GÉANT AutoBHAN System
  - Nortel DRAC
  - Surfnet (via use of Nortel DRAC)
  - LHCNet (OSCARS/DCN)
  - Nysernet (New York RON) (OSCARS/DCN)
  - LEARN (Texas RON) (OSCARS/DCN)
  - LONI (OSCARS/DCN)
  - Northrop Grumman (OSCARS/DCN)
  - University of Amsterdam (OSCARS/DCN)
  - MAX (OSCARS/DCN)
- The following “higher level service applications” have adapted their existing systems to communicate using the DICE IDCP:
  - LambdaStation (FNAL)
  - TeraPaths (BNL)
  - Phoebus (University of Delaware)

PerfSONAR

- An open web-services based framework for collecting, managing and sharing network measurements
- The framework is being deployed across the science community
- Encouraging people to deploy *'known good'* measurement points near domain boundaries
- Using the framework to find & correct soft network failures.

ARRA Advanced Networking Initiative

ESnet received ~$62M in ARRA funds from DOE for an Advanced Networking Initiative

- Build an end-to-end prototype network to address our growing data needs while accelerating the development of 100 Gbps networking technologies
- Build a network testbed facility for researchers and industry

DOE is also funding $5M in network research that will use the testbed facility with the goal of near-term technology transfer to the production ESnet network

Separately, DOE has funded Magellan, an associated DOE computing project, at $33M that will utilize the 100 Gbps network infrastructure

- A research and development effort to establish a nationwide scientific mid-range distributed computing and data analysis testbed
- It will have two sites (NERSC / LBNL and ALCF / ANL) with multiple 10's of teraflops and multiple petabytes of storage, as well as appropriate cloud software tuned for moderate concurrency.

ANI Overview

Prototype network goals are to accelerate the deployment of 100 Gbps technologies and strengthen US competitiveness

- Key step toward DOE's vision of a 1-terabit networking linking DOE supercomputing centers and experimental facilities
- Build a persistent infrastructure that will transition to the production network ~2012

Testbed goal is to build an experimental network research environment at sufficient scale to usefully test experimental approaches to next generation networks

- Funded by ARRA for 3 years, then roll into the ESnet program
  - Project start date: Sept 2009
- Breakable, reservable, configurable, resettable
- Enable R&D at 100 Gbps

100 Gbps Prototype Overview

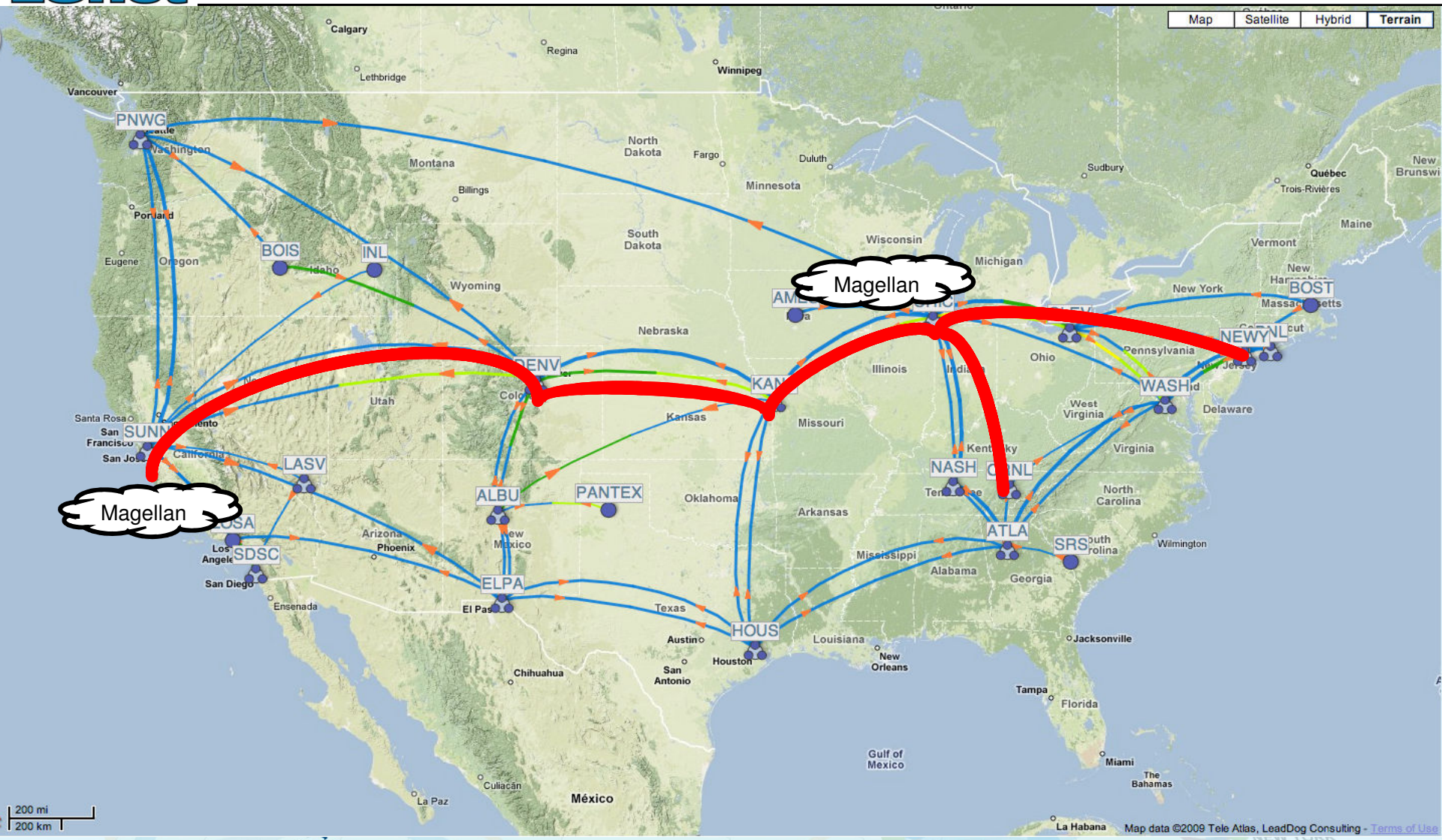
Started technology research and evaluation process

- Meetings / briefings with vendors to ensure we understand the technology and direction companies are going
- Begun design / planning process with ESnet metropolitan network operators

Transport and routing/switch technologies are at different stages of development requiring separate RFPs

- Transport: reasonable number of technology options by early 2010
  - Looking for 100 Gbps wave service in the wide area – can be shared
  - Don't need to own / control optical gear
  - Plan to run OSCARS layer 2 / 3 services across network
  - Dark fiber is part of DOE's long-term research agenda
- Routing / Switching: limited options, not ready till end of 2010 at earliest
  - ESnet will purchase this equipment
  - Will conduct testing / evaluation as part of selection process

Advanced Networking Initiative Topology



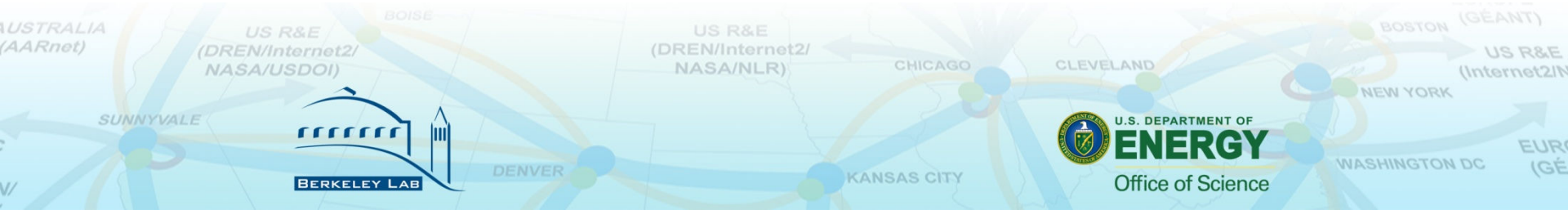
Testbed Overview

Capabilities:

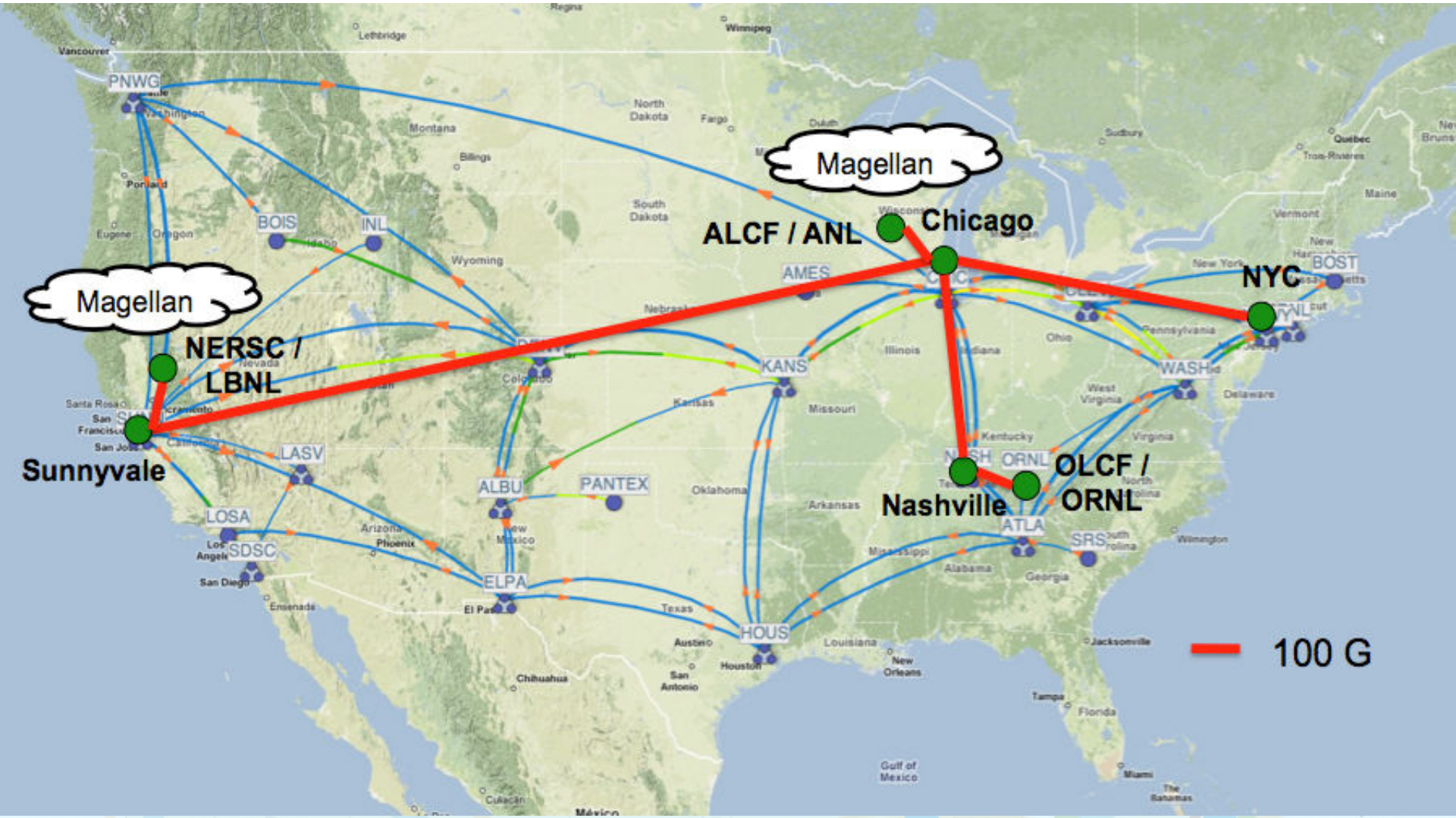
- Ability to support end-to end experiments at 100 Gbps
- Dynamic network provisioning
- Plan to acquire and use dark fiber on a portion of testbed footprint
  - Ability to do hybrid (layer 0-3) networking
- Use Virtual Machine technology to support protocol and middleware research
- Detailed Monitoring
  - Researchers will have access to all possible monitoring data from the network devices

Each node will consist of:

- DWDM device (Layer 0-1)
- Programmable Ethernet Switch (layer 2)
- Standard and programmable Router (layer 3)
- Test and measurement hosts
  - VM based test environments



Nationwide 100 Gbps Testbed Network



Supported DOE ARRA-funded Research Projects

Climate 100 (LLNL, LBL, ANL)

- Project Goal: Scaling the ESG to 100 Gbps
- Testbed role: provide interconnect between “Magellan project” resources at ANL and NERSC

Advanced Network and Distributed Storage Laboratory (OSG: U Wisc, FNAL, etc)

- Project Goal: enhance VDT data management tools to effectively utilize 100 Gbps networks
- Testbed role: provide interconnect between “Magellan project” resources at ANL and NERSC

100 Gbps NIC (Acadia and Univ New Mexico)

- Project Goal: produce a host NIC capable of 100 Gbps
- Testbed role: Provide test environment for this device

100 Gbps FTP (BNL/Stony Brook University)

- Project Goal: design 100 Gbps transport protocols and tools
- Testbed role: Provide a 100 Gbps test environment

Supported DOE Network Research Projects

"Resource optimization in hybrid core networks with 100 Gbps links" (Univ Virginia)

- Project Goal: Develop methods for automatic classification of flows that should be moved from the IP network to a dynamic virtual circuit
- Testbed role: Provide a test environment for validating these methods

"Integrating Storage Resource Management with Dynamic Network Provisioning for Automated Data Transfer" (BNL, LBL)

- Project Goal: Integration of dynamic virtual circuits into BestMAN – Berkeley Storage Manager
- Testbed role: Provide a 100 Gbps test environment for verifying this work

"Provisioning Terascale Science Apps using 100 Gbps Systems" (UC Davis)

- Project Goal: Advanced path computation algorithms
- Testbed role: Provide a control plane test environment for these algorithms using OSCARS

Supported DOE Network Research Projects

Virtualized Network Control (ISI, ESnet, Univ New Mexico)

- Project Goal: multi-layer, multi-technology dynamic network virtualization
- Testbed role: provide a control plane test environment for experiments using hybrid networking

"Sampling Approaches for Multi-domain Internet Performance Measurement Infrastructures to Better Serve Network Control and Management" (Ohio State Univ)

- Project Goal: Develop new network measurement sampling techniques and policies
- Testbed role: Provide a network environment for initial deployment

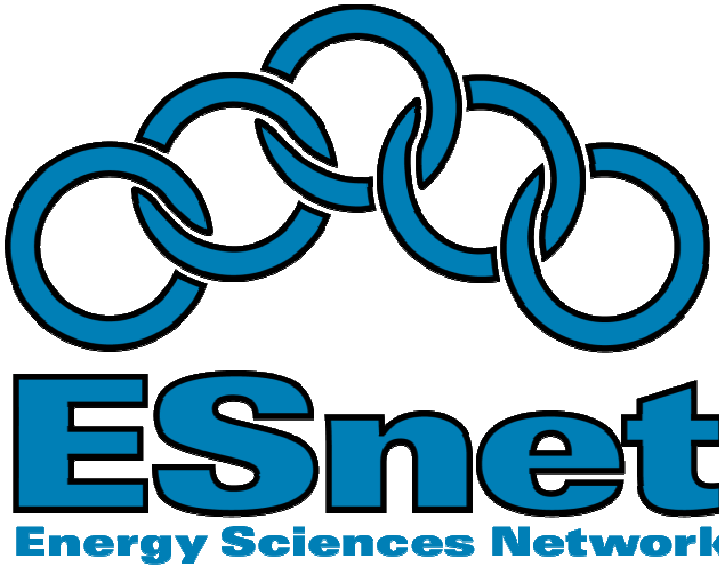
OSCARS Collaborative Research Efforts

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, Internet2)
- 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)


ESnet
Energy Sciences Network


U.S. DEPARTMENT OF
ENERGY
Office of
Science