

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

perfSONAR ESCC Indianapolis IN

July 21, 2009

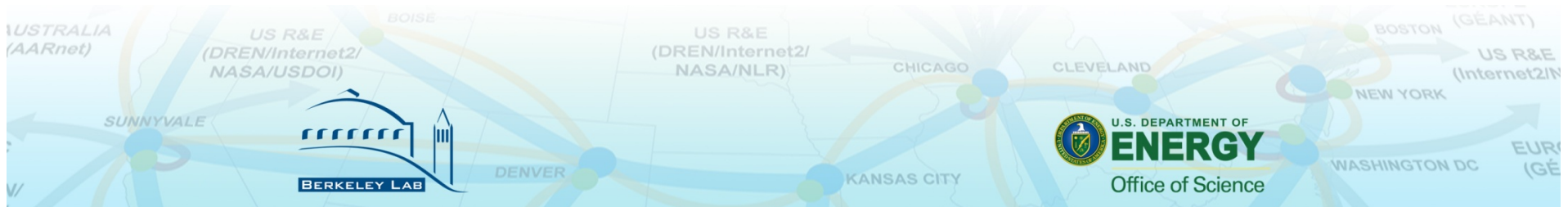
Joe Metzger, Brian Tierney
ESnet/LBNL





Measurement Recommendations

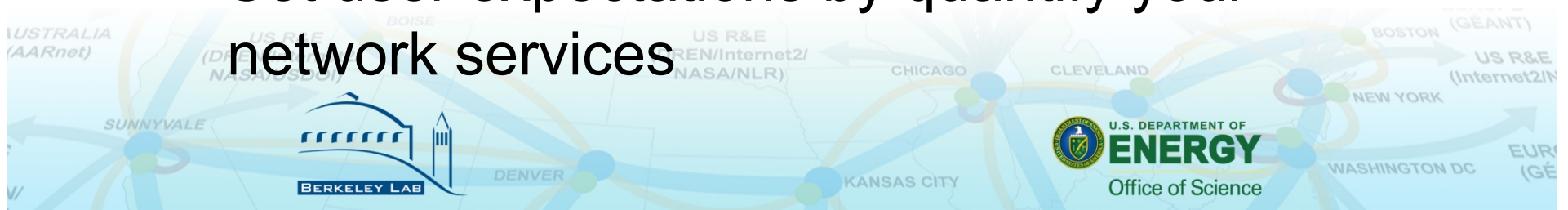
- White Paper has been posted at:
 - http://fasterdata.es.net/ps_howto.html
 - This is a draft
 - I am expecting additional community input and will continually refine this document.





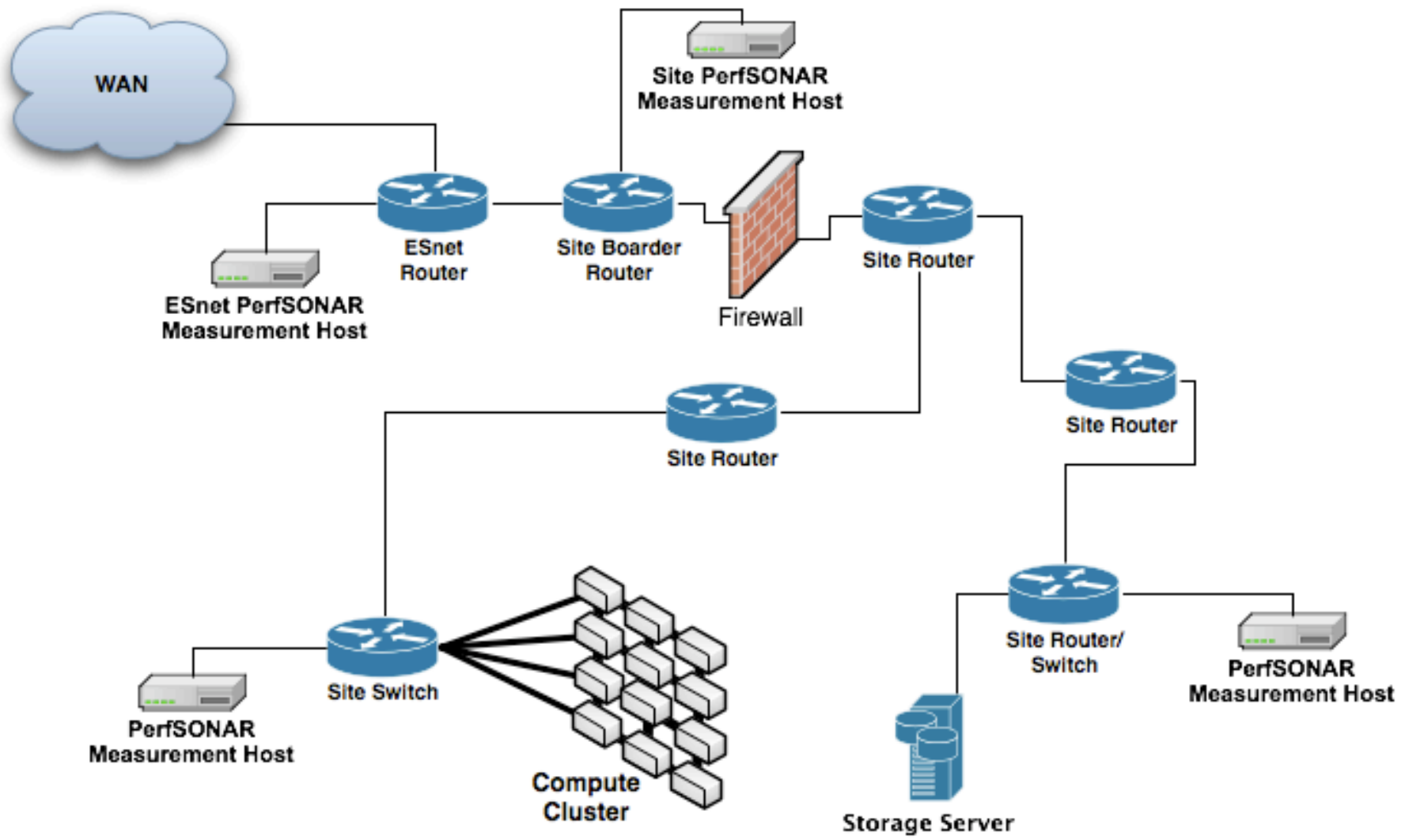
Measurement Recommendations

- Deploy perfSONAR tools
- At Site border:
 - 1 Bandwidth system, 1 latency system & several other services (Utilization, NDT, etc)
- Near significant network resources
- Use it to:
 - Find & fix current local problems
 - Identify when they re-occur
 - Set user expectations by quantify your network services





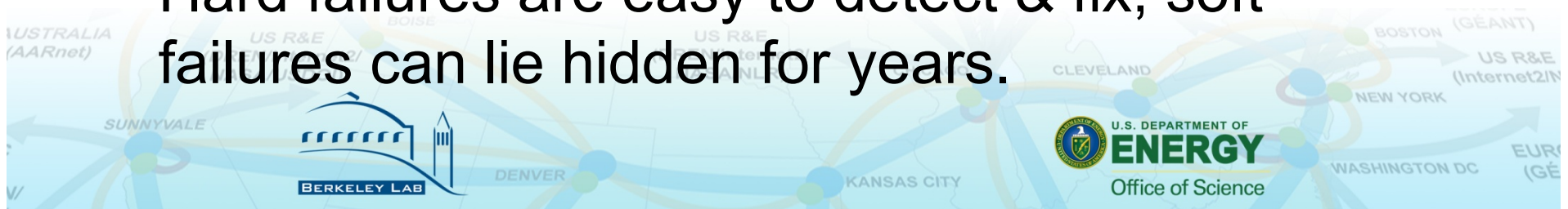
Typical Campus Deployment

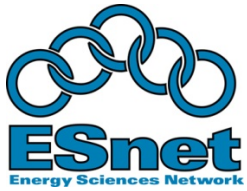




Soft Network Failures

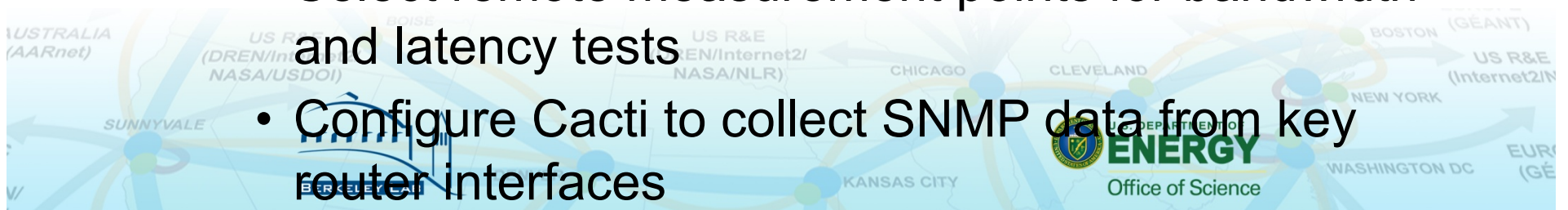
- Soft failures are where basic connectivity functions, but high performance is not possible.
- TCP was intentionally designed to hide all transmission errors from the user:
 - “As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users.” (From IEN 129, RFC 716)
- **Some soft failures only affect high bandwidth long RTT flows.**
- Hard failures are easy to detect & fix, soft failures can lie hidden for years.

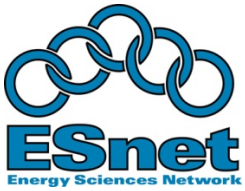




Deploying a perfSONAR measurement host in under 30 minutes

- Using the PS Performance Toolkit is very simple
 - Boot from CD
 - Use command line tool to configure
 - what disk partition to use for persistent data
 - Network address and DNS
 - User and root passwords
 - Use Web GUI to configure
 - Select which services to run
 - Select remote measurement points for bandwidth and latency tests
 - Configure Cacti to collect SNMP data from key router interfaces

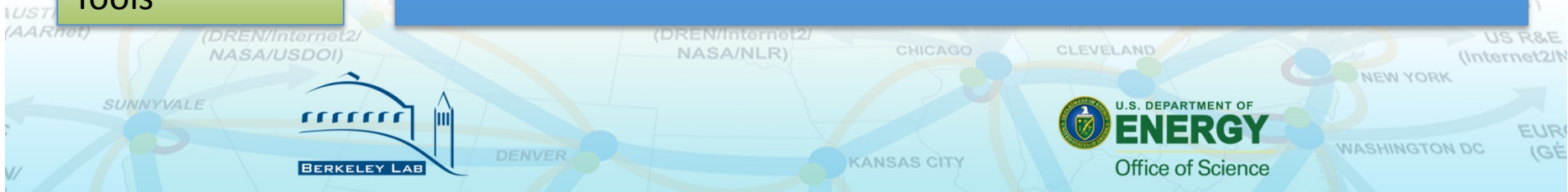
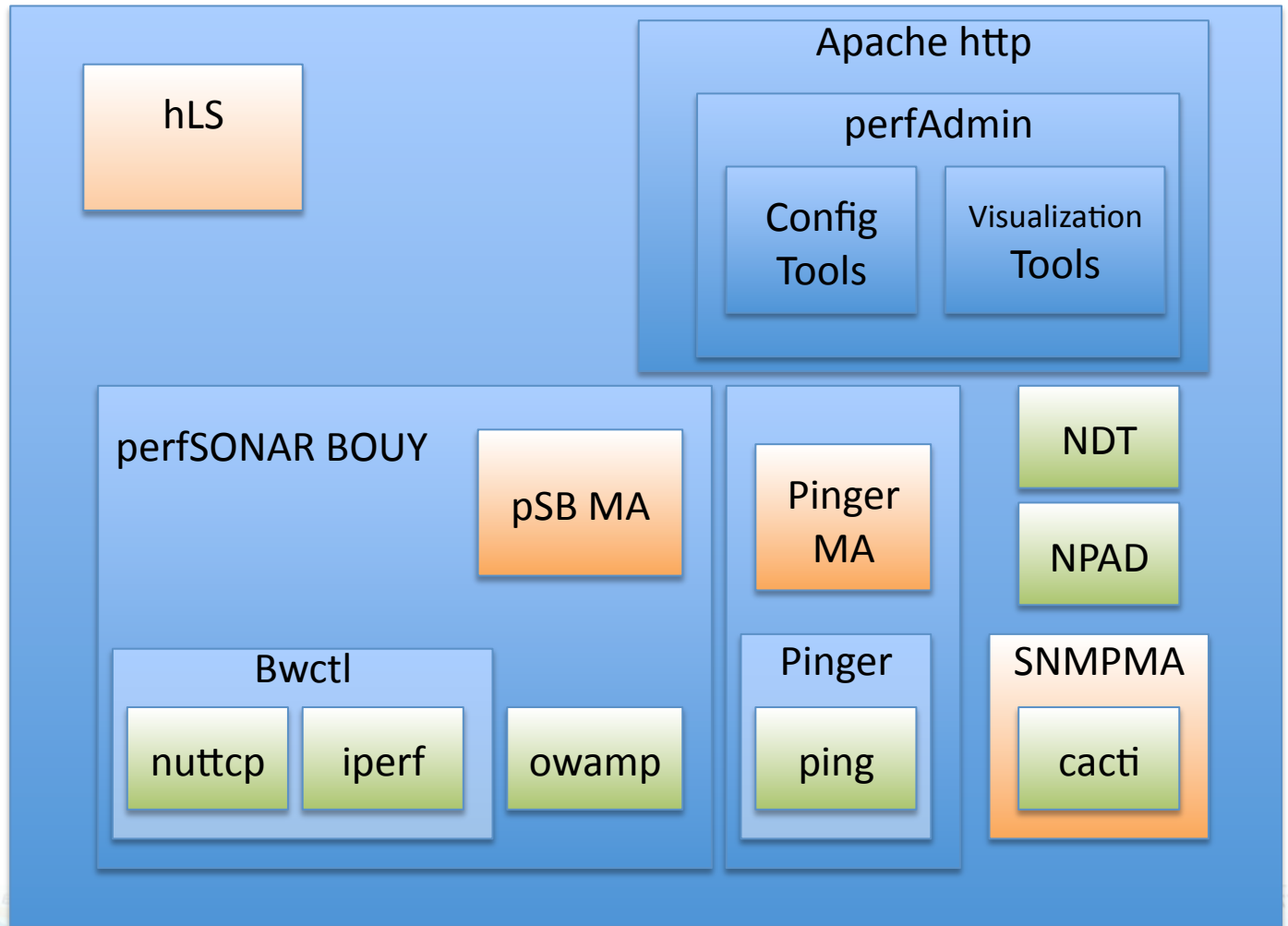




PS-Performance Toolkit Components

ing
ure?

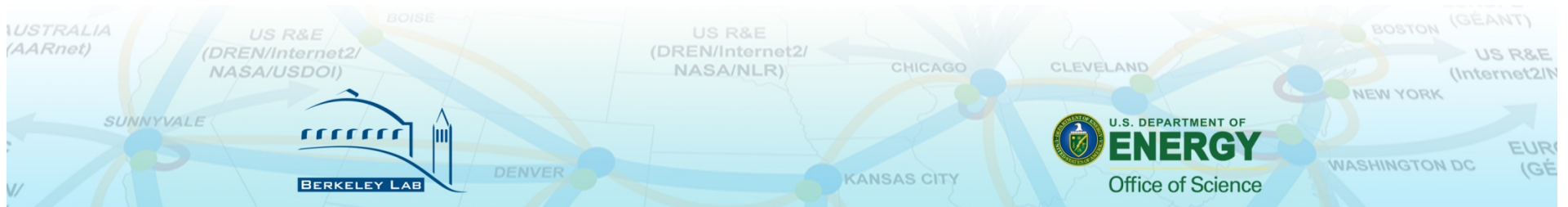
- Support Applications
- Provides a perfSONAR Web Service
- Low Level Measurement Tools





Backup Slides

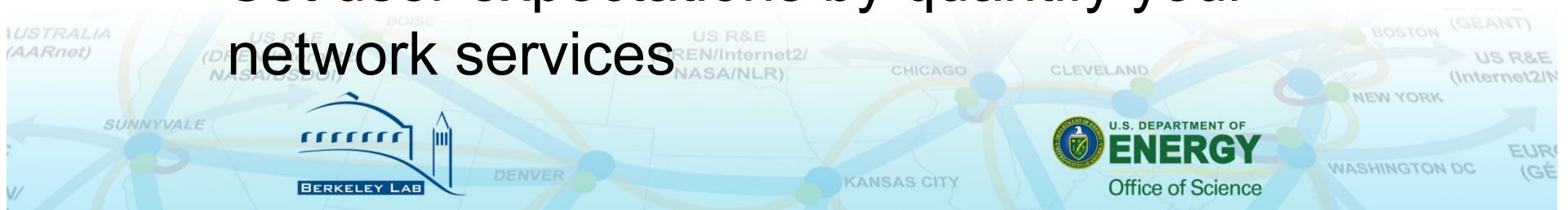
- The following slides were from the measurement BOF at Joint Techs.





Measurement Recommendations

- Deploy perfSONAR tools
- At Site border:
 - 1 Bandwidth system, 1 latency system & several other services (Utilization, NDT, etc)
- Near significant network resources
- Use it to:
 - Find & fix current local problems
 - Identify when they re-occur
 - Set user expectations by quantify your network services





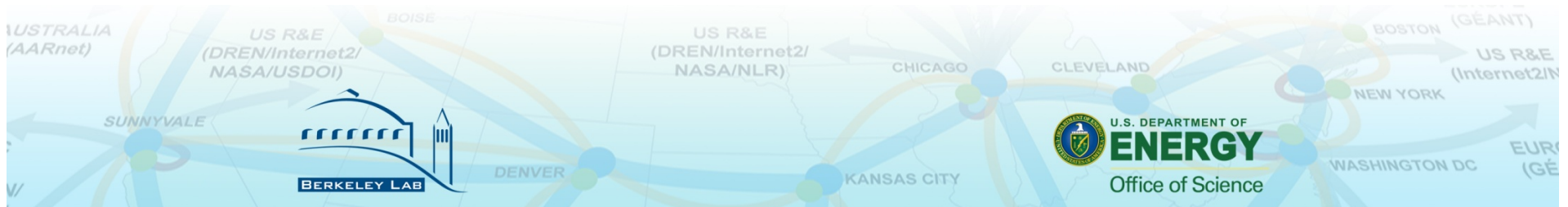
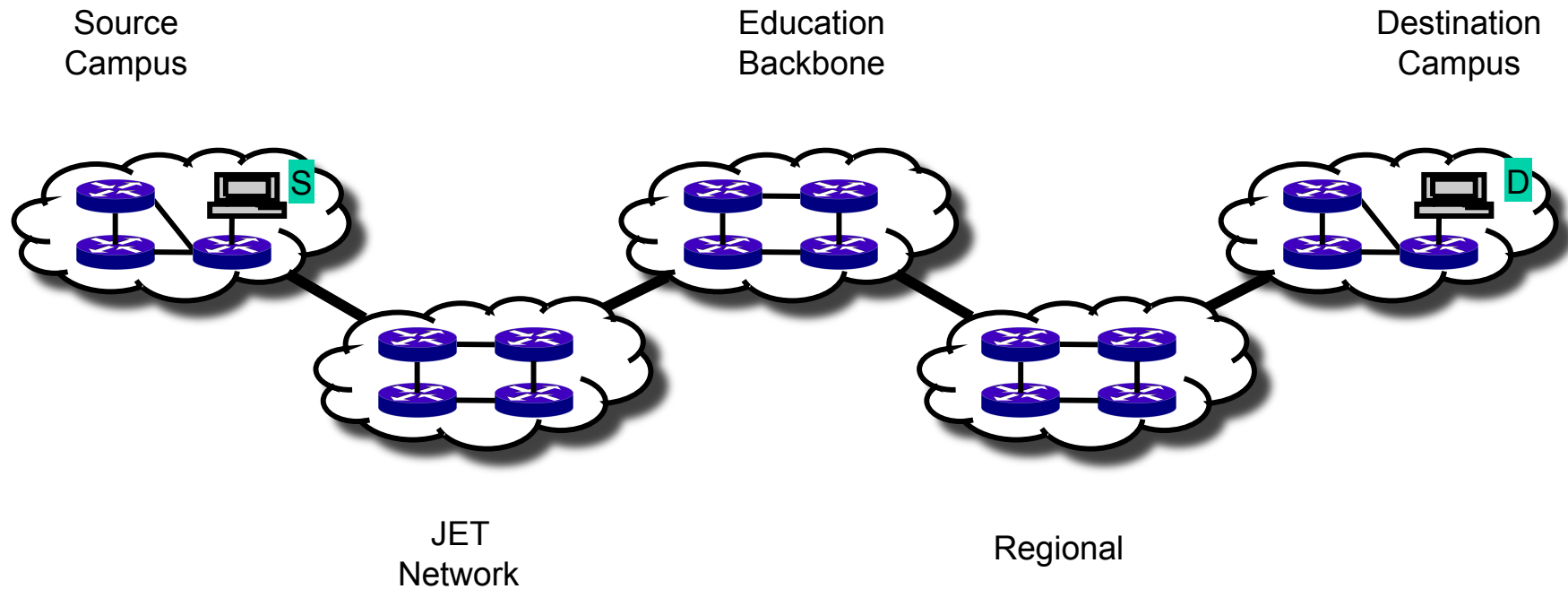
High Performance Networking

- Most of the R&E community has access to 10 Gbps networks.
- Naive users with the right tools should be able to easily get:
 - **200 Mbps/per stream between properly maintained systems**
 - **2 Gbps aggregate rates between significant computing resources**
- Most users are not experiencing this level of performance
 - “There is widespread frustration involved in the transfer of large data sets between facilities, or from the data’s facility of origin back to the researcher’s home institution. “ From the BES network requirements workshop:
<http://www.es.net/pub/esnet-doc/BES-Net-Req-Workshop-2007-Final-Report.pdf>
- We can increase performance by measuring the network and reporting problems!

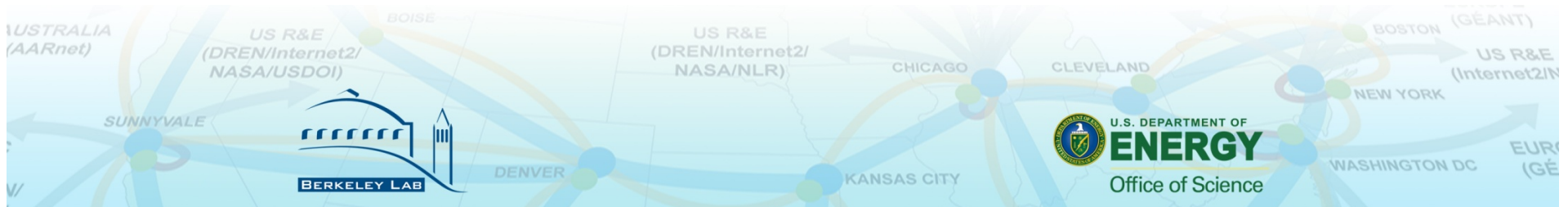
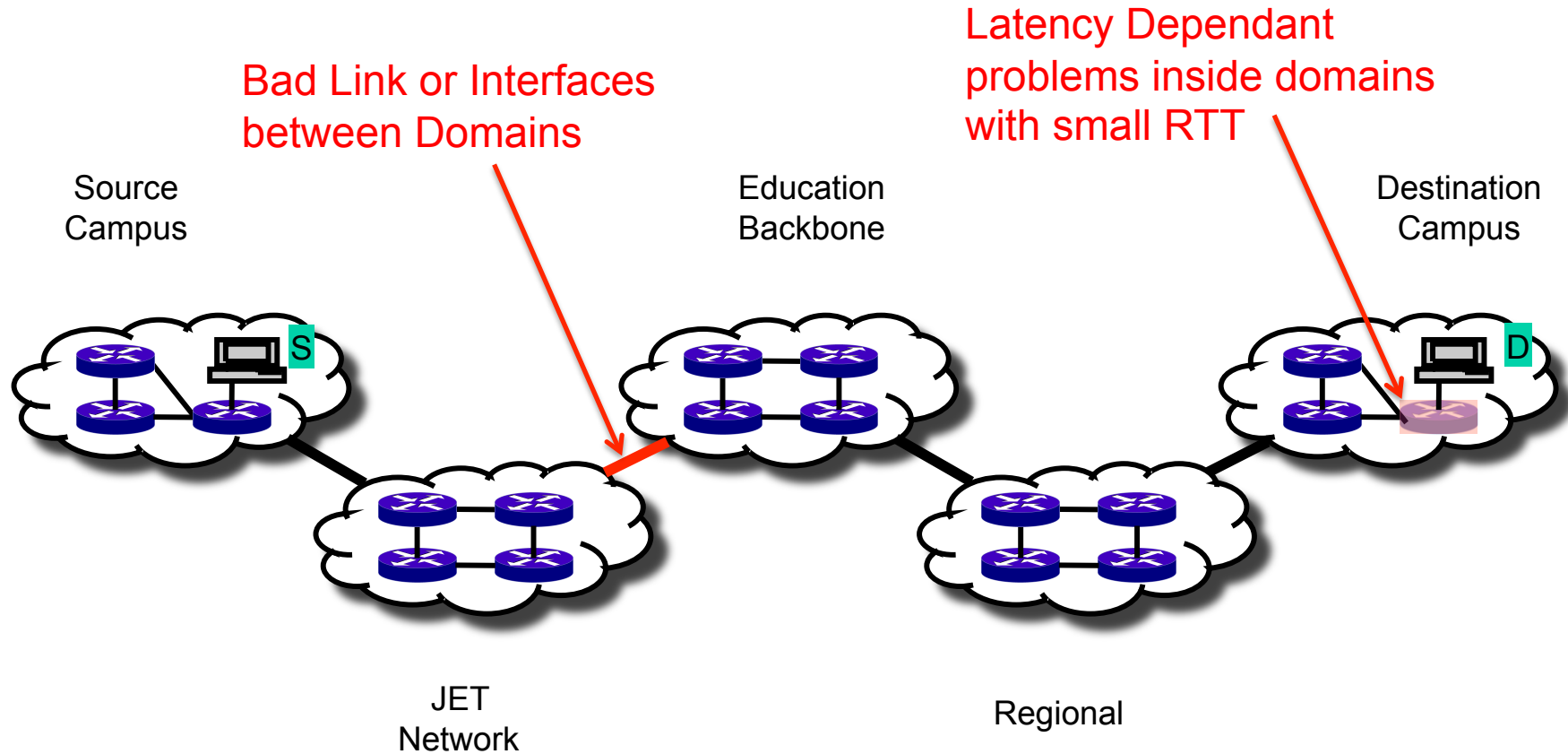




Network Troubleshooting is a Multi-Domain Problem



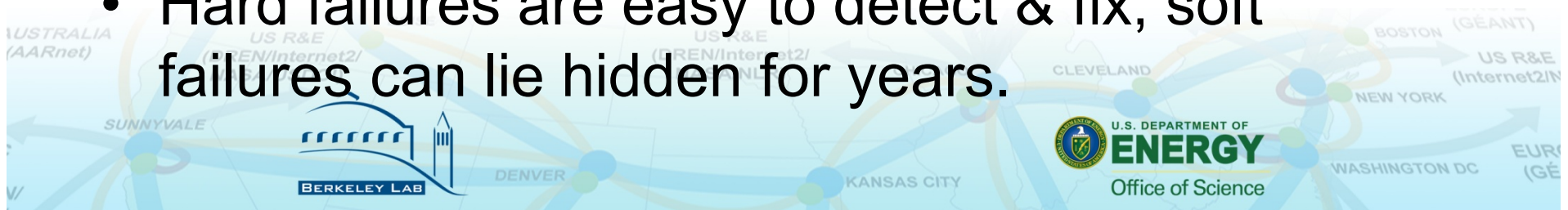
Where are common problems?

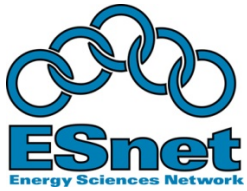




Soft Network Failures

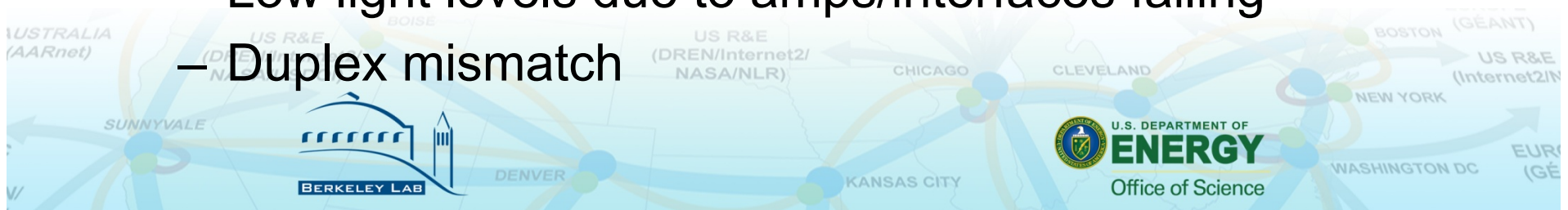
- Soft failures are where basic connectivity functions, but high performance is not possible.
- TCP was intentionally designed to hide all transmission errors from the user:
 - “As long as the TCPs continue to function properly and the internet system does not become completely partitioned, no transmission errors will affect the users.” (From IEN 129, RFC 716)
- **Some soft failures only affect high bandwidth long RTT flows.**
- Hard failures are easy to detect & fix, soft failures can lie hidden for years.

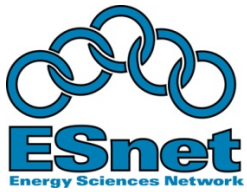




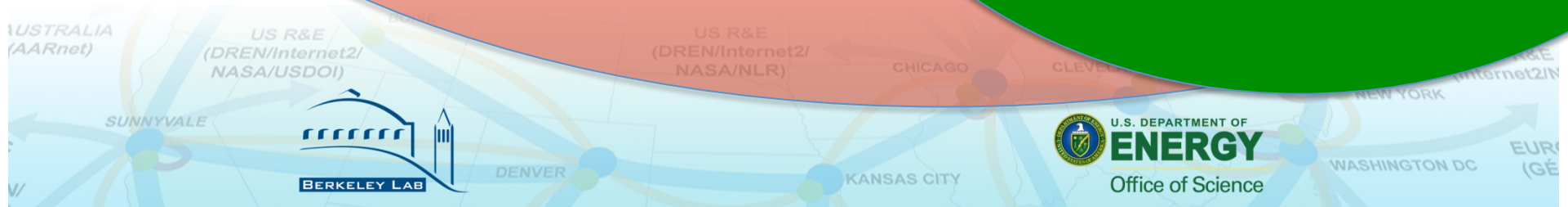
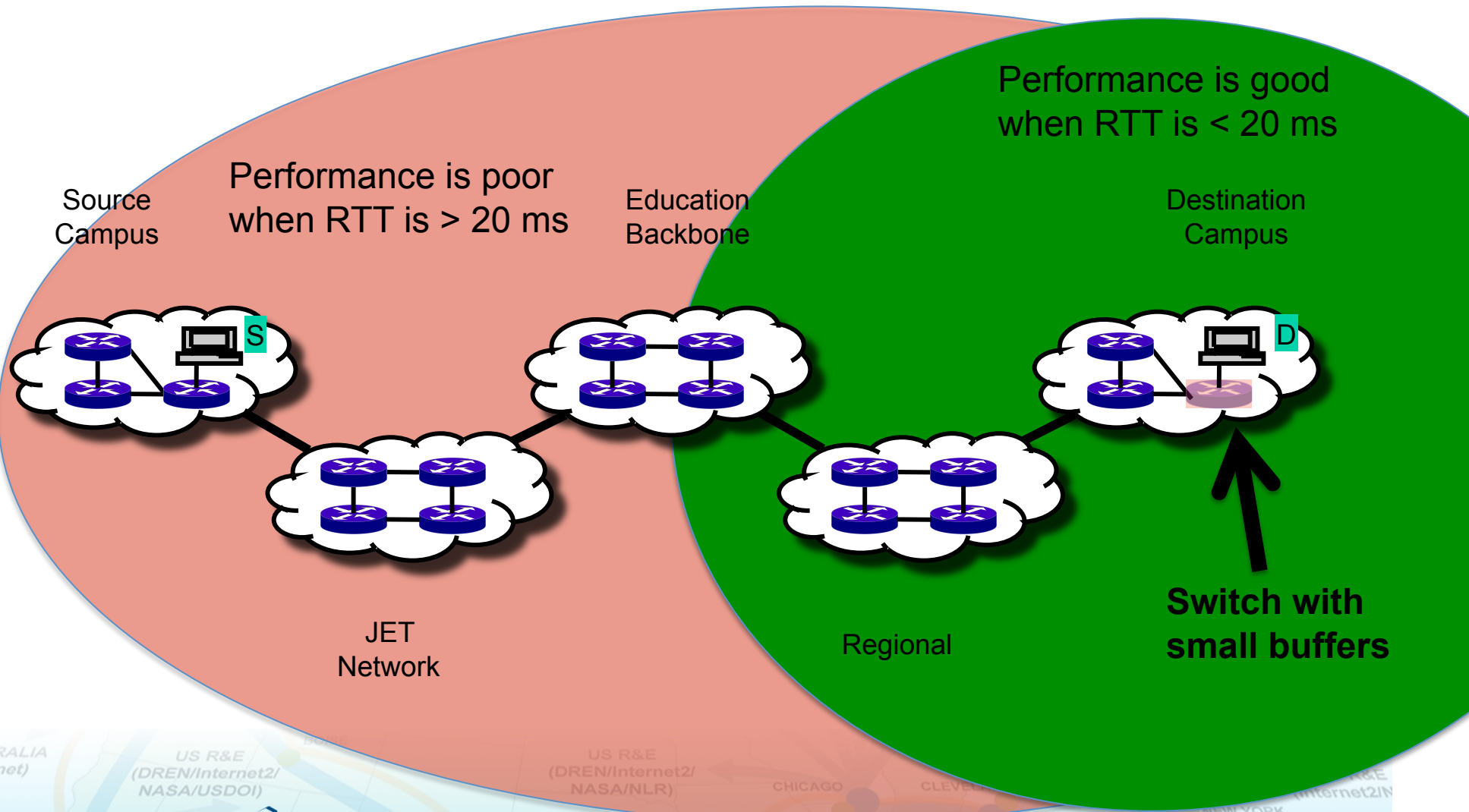
Common Soft Failures

- Small Queue Tail Drop
 - Switches not able to handle the long packet trains prevalent in long RTT sessions and cross traffic at the same time
- Un-intentional Rate Limiting
 - Process Switching on Cisco 6500 devices due to faults, acl's, or mis-configuration
 - Security Devices...
- Random Packet Loss
 - Bad fibers or connectors
 - Low light levels due to amps/interfaces failing
 - Duplex mismatch





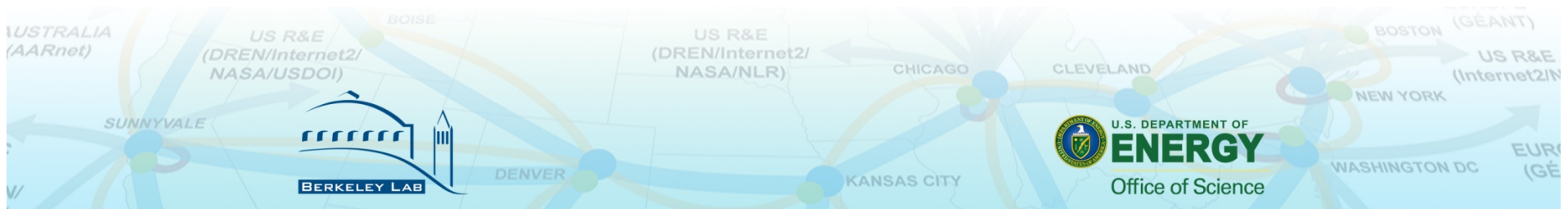
Local testing will not find some problem.





Addressing the Problem: perfSONAR

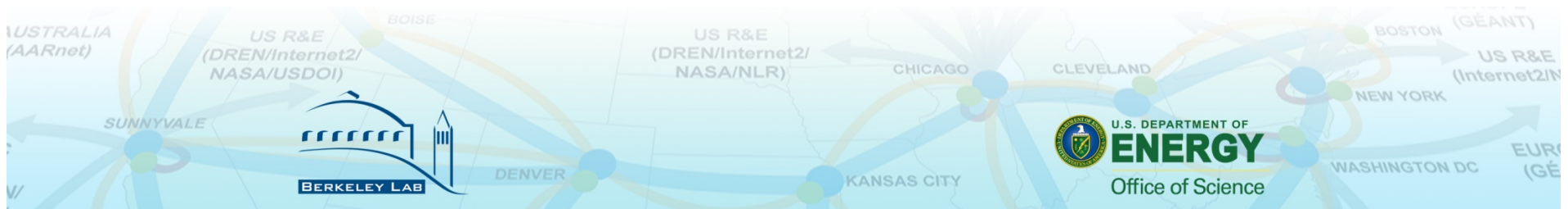
- Developing an open web-services based framework for collecting, managing and sharing network measurements
- Deploying the framework across the science community
- Encouraging people to deploy '*known good*' measurement points near domain boundaries





What is perfSONAR?

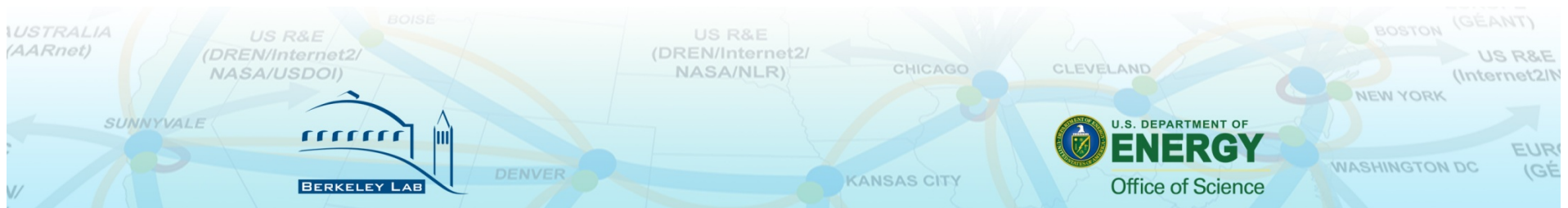
- A collaboration
 - For developing, deploying & utilizing network measurement tools
- An architecture and protocols
- A collection of software

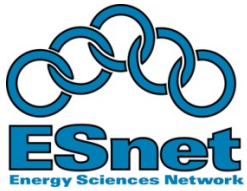




perfSONAR Terminology

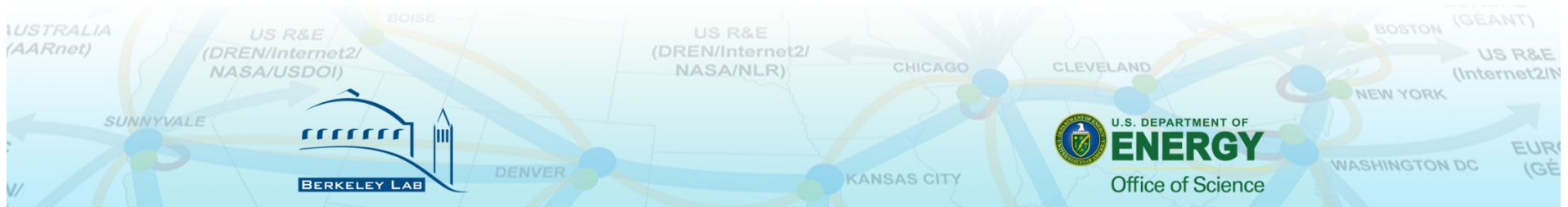
- perfSONAR Collaboration: Collection of groups working on perfSONAR tools
- perfSONAR Schemas: OGF standards for perfSONAR communications
- perfSONAR Bundle: collection of tools into a release
- perfSONAR MDM: A measurement service coordinated by DANTE
- perfSONAR PS: Perl-based tools
- perfSONAR Performance Toolkit: Bootable CD packaging of several tool
- perfSONAR Bandwidth Services: Active bandwidth probe control (bwctl)
- perfSONAR Latency Services: Active latency probe control (owamp/PingER)
- perfSONAR Measurement Archives: Store and publish results / data
- perfSONAR Analysis Tools: data visualization tools
- perfSONAR Troubleshooting Services: NDT and NPAD
- **perfSONAR = all of the above**

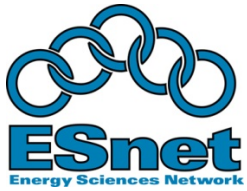




perfSONAR Developers

- ESnet
- GEANT
- Internet2
- RNP
- University of Delaware
- FERMI
- Georgia Tech
- SLAC
- ARIES
- BELNET
- CARNet
- CESNET
- DANTE
- DFN
- FCCN
- Consortium GARR
- GRNET
- IST
- POZNAN Supercomputing Center
- Red IRIS
- Renater
- SURFnet
- SWITCH
- UNINETT



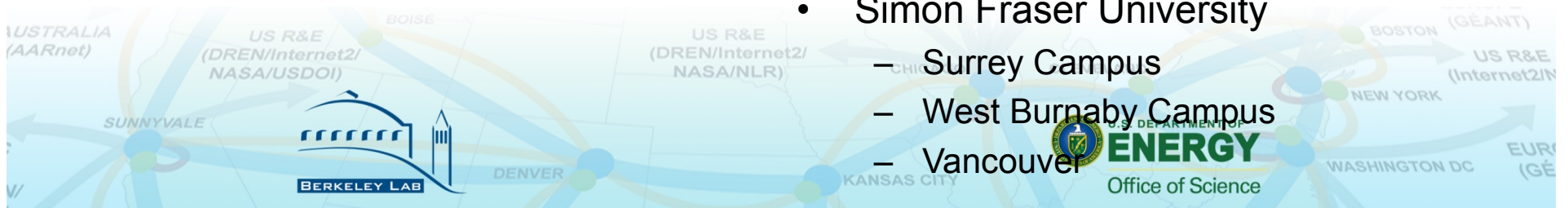


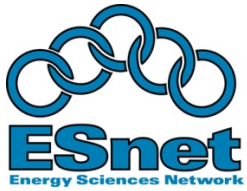
perfSONAR Deployments

- Internet2
- University of Michigan, Ann Arbor
- Indiana University
- Boston University
- University of Texas Arlington
- Oklahoma University, Norman
- Michigan Information Technology Center
- William & Mary
- University of Wisconsin Madison
- Southern Methodist University, Dallas
- University of Texas Austin
- Vanderbilt University
- ESnet
- Argonne National Lab
- Brookhaven National Lab
- Fermilab
- National Energy Research Scientific Computing Center
- Pacific Northwest National Lab

- APAN
- GLORIAD
- JGN2PLUS
- KISTI Korea
- Monash University, Melbourne
- NCHC, HsinChu, Taiwan
- Simon Fraser University

- Surrey Campus
- West Burnaby Campus
- Vancouver





perfSONAR Deployments (2)

- GEANT
- GARR
- HUNGARNET
- PIONEER
- SWITCH
- CCIN2P3
- CERN
- CNAF
- DE-KIT
- NIKHEF/SARA
- PIC
- RAL
- TRIUMF
- ASCC

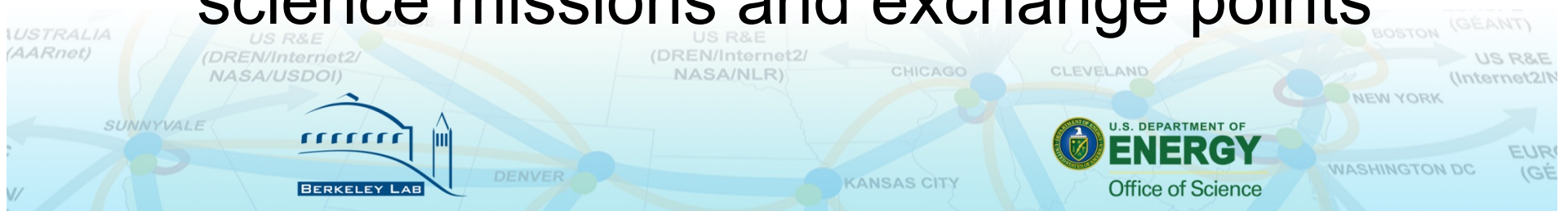
Note: These are just the deployments I know about. There are probably more...

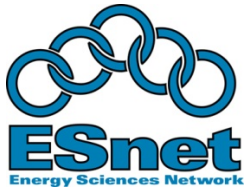




perfSONAR JET deployment

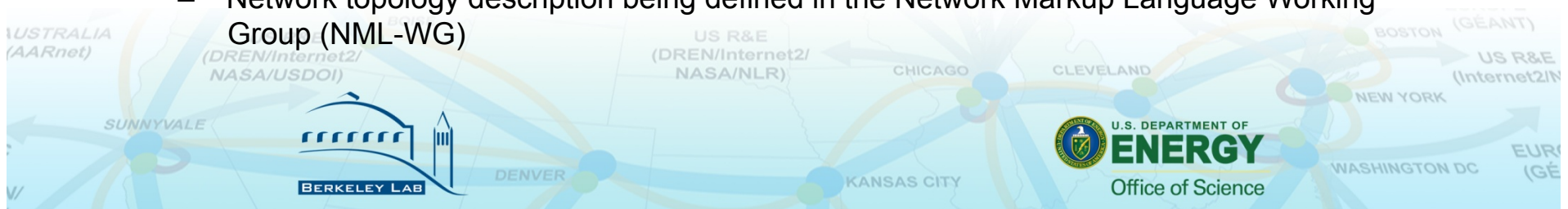
- The Joint Engineering Team is developing a perfSONAR deployment plan
 1. Reviewing the network measurement data each network is willing to share, or would like to access
 2. Reviewing the perfSONAR tools & monitoring functions to evaluate which networks will deploy which ones.
- First deployments in the nets with open science missions and exchange points

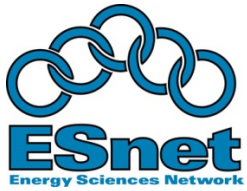




perfSONAR Architecture

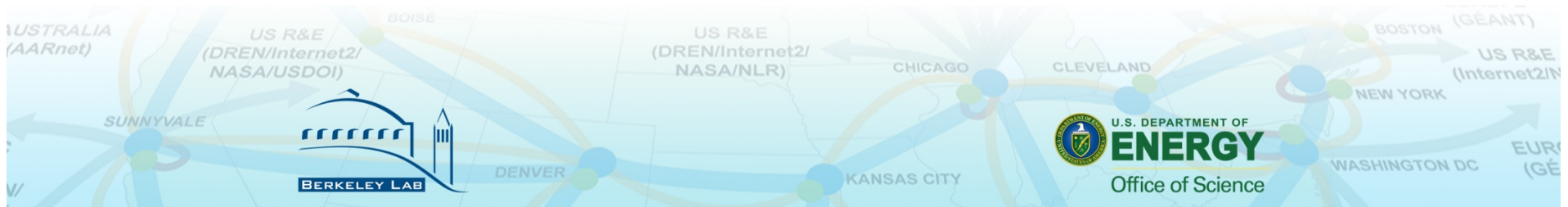
- Interoperable network measurement middleware (SOA):
 - Modular
 - Web services-based
 - Decentralized
 - Locally controlled
- Integrates:
 - Network measurement tools and data archives
 - Data manipulation
 - Information Services
 - Discovery
 - Topology
 - Authentication and authorization
- Based on:
 - Open Grid Forum Network (OGF) Network Measurement Working Group (NM-WG) schema
 - Currently attempting to formalize specification of perfSONAR protocols in a new OGF WG (NMC-WG)
 - Network topology description being defined in the Network Markup Language Working Group (NML-WG)





perfSONAR Protocols

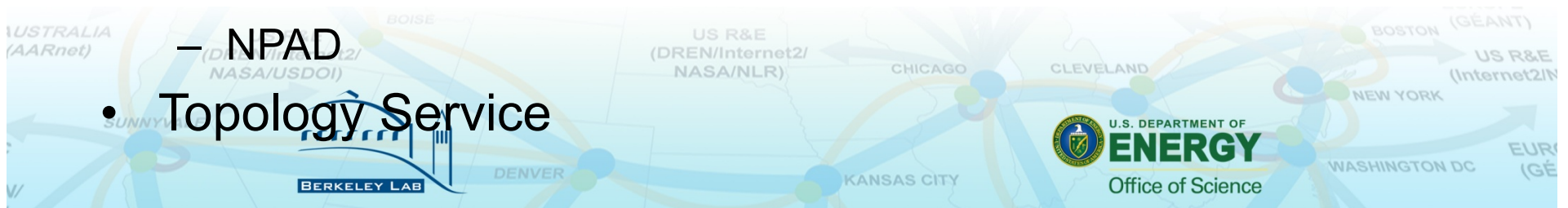
- Web Services based protocols for:
 - Finding measurement services
 - Exchanging measurement data
 - Scheduling measurements
- We are standardizing these protocols in the OGF





Main perfSONAR Services

- Lookup Service
 - gLS – Global service used to find services
 - hLS – Home service for registering local perfSONAR metadata
- Measurement Archives
 - SNMP MA – Interface Data
 - pSB MA -- Scheduled bandwidth and latency data
- Measurement Points
 - BWCTL
 - OWAMP
 - PINGER
- Troubleshooting Tools
 - NDT
 - NPAD
- Topology Service

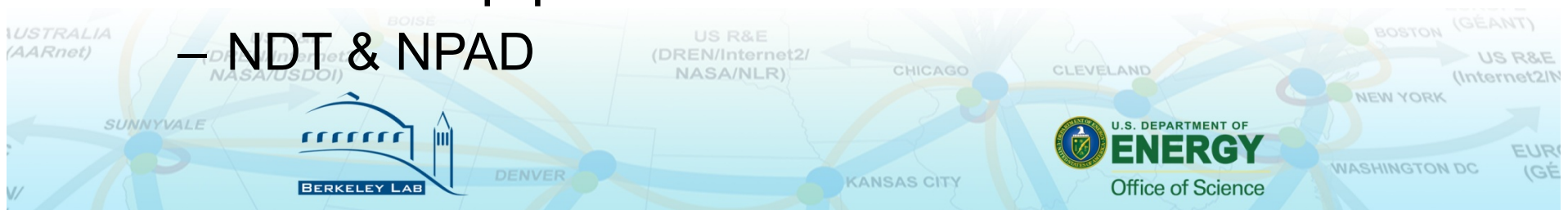




Selecting Network Measurements

- Router Interface Data
 - Utilization, Errors, Discards
 - Border & internal bottleneck links
 - Before & after the security infrastructure
- Active Bandwidth Measurements
 - Identify Important paths to measure
 - Do you need to test 10G paths?
- Latency Measurements
 - Identify important paths to measure
- LAN/Desktop performance

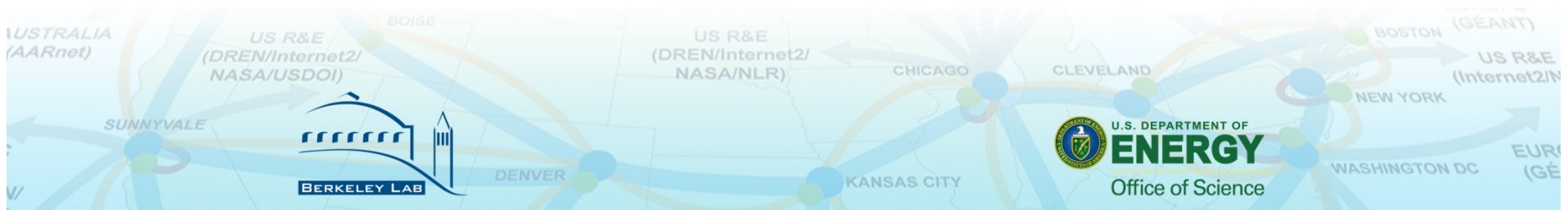
– NDT & NPAD





perfSONAR Software Terminology

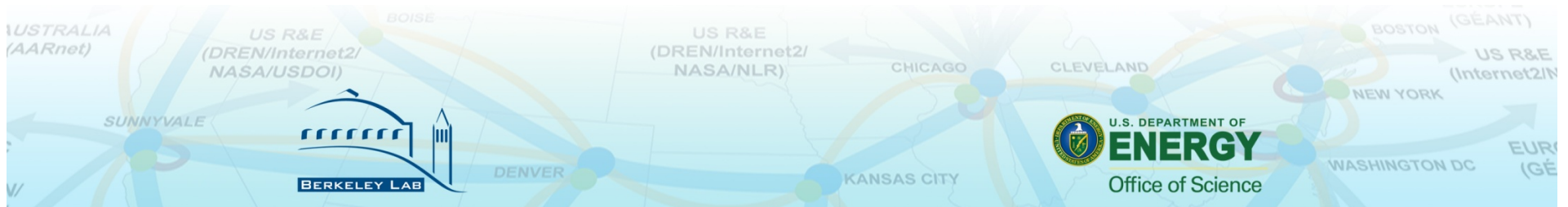
- There are multiple perfSONAR services
 - le, lookup service, measurement archives, measurement points, authentication, etc
- There are multiple code trains
 - perfSONAR-PS
 - perfSONAR MDM
- perfSONAR service bundles
 - Integrated tested releases that may contain services picked from **both** code trains.





perfSONAR PS

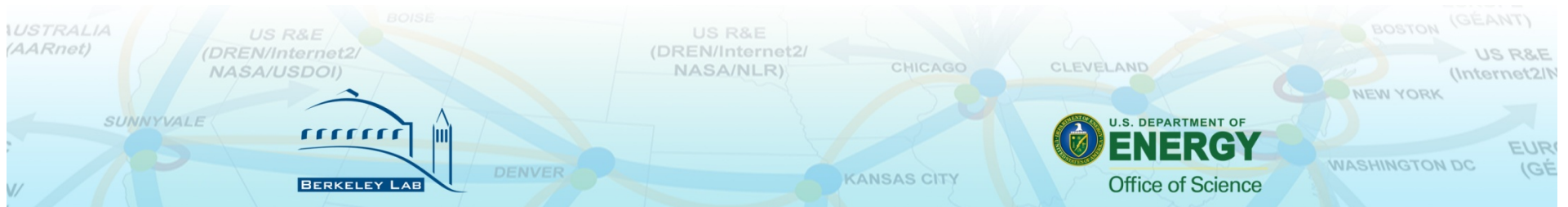
- Primarily written in Perl
- Emphasis on
 - ease of deployment
 - community driven development & support
- Mostly US Developers





perfSONAR MDM

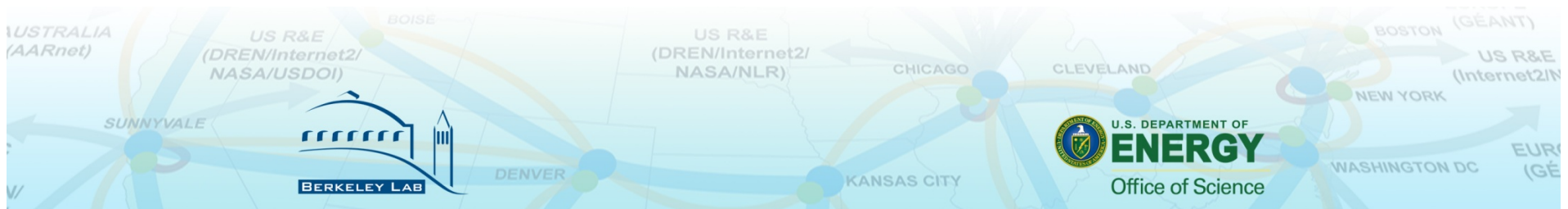
- Heavy reliance on Java
 - Some perl as well
- Emphasis on
 - measurement as a service offering
 - security & access restrictions
- Mostly European developers





perfSONAR Bundles

- PS-Performance Toolkit
 - Based on perfSONAR-PS code train
 - CDROM that automates creating a measurement appliance
- perfSONAR-PS v3.1
 - Packages of individual perfSONAR services
- perfSONAR-MDM v3.1
 - The basis of the LHCOPN perfSONAR MDM service





Selecting a Bundle or Distribution

- Do you need to support NDT & NPAD, or are you looking for a simple measurement appliance?
 - Consider PS-Performance Toolkit
- Does your organization have restrictions on OS's and patching for servers supporting external network services?
 - Consider perfSONAR PS RPM packages
- Is publishing data to restricted groups critical, and are you a member of the eduGAIN federation?
 - Consider MDM release

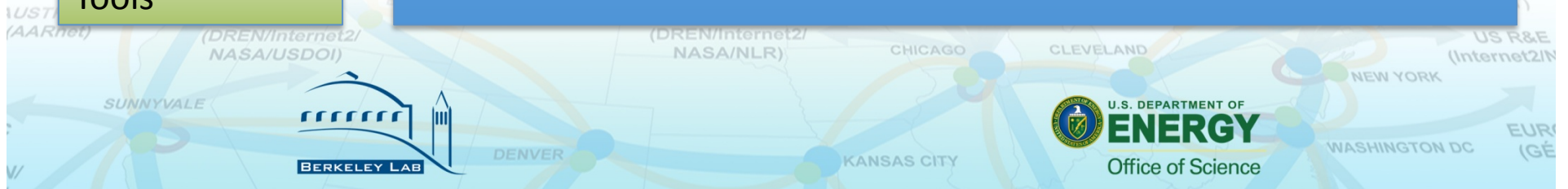
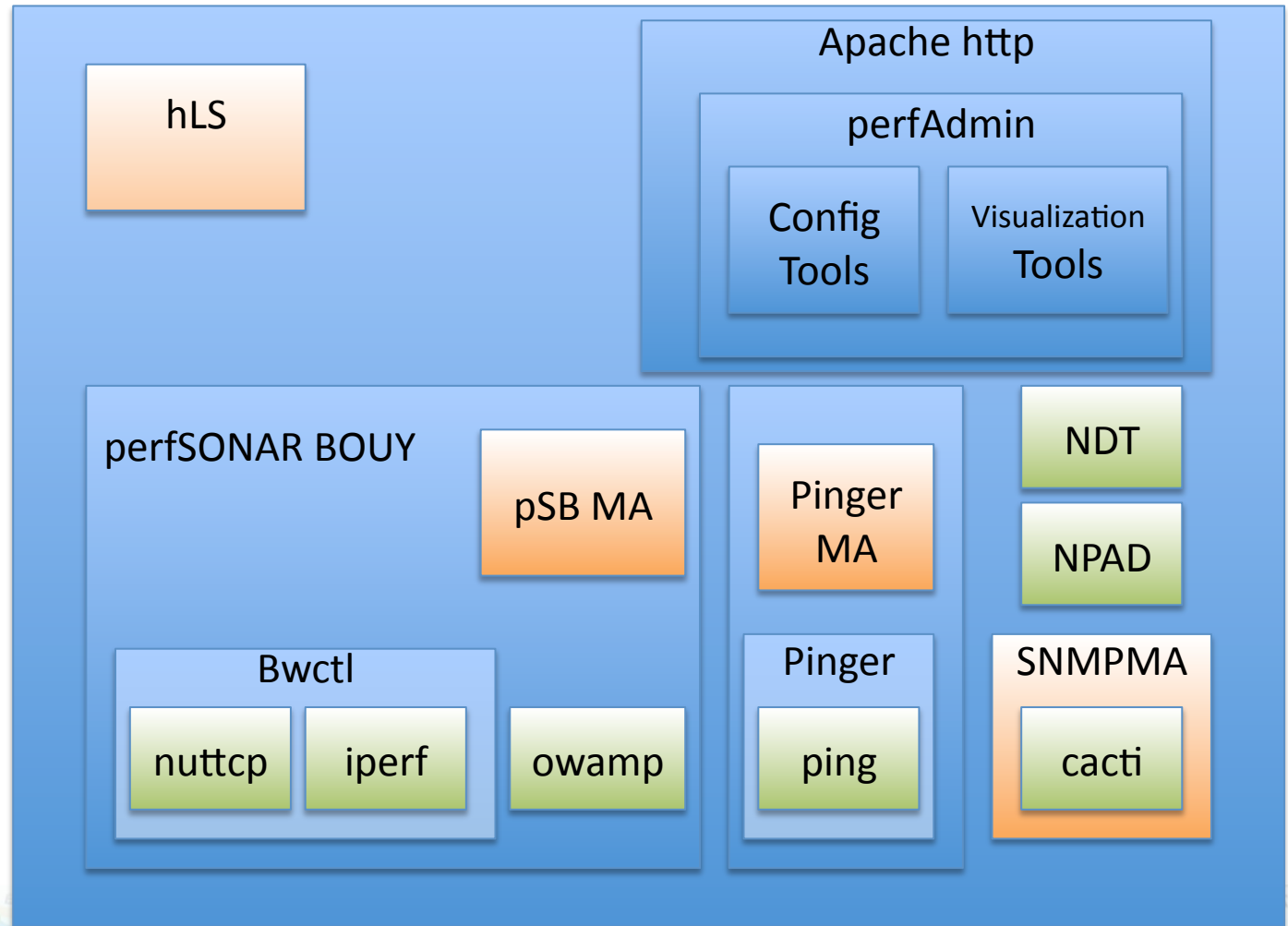




PS-Performance Toolkit Components

ing
ure?

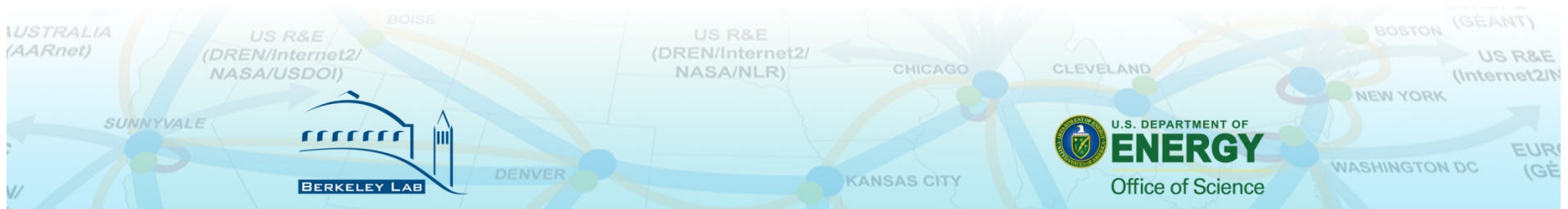
- Support Applications
- Provides a perfSONAR Web Service
- Low Level Measurement Tools

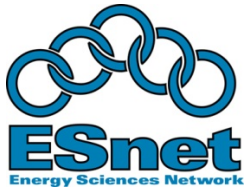




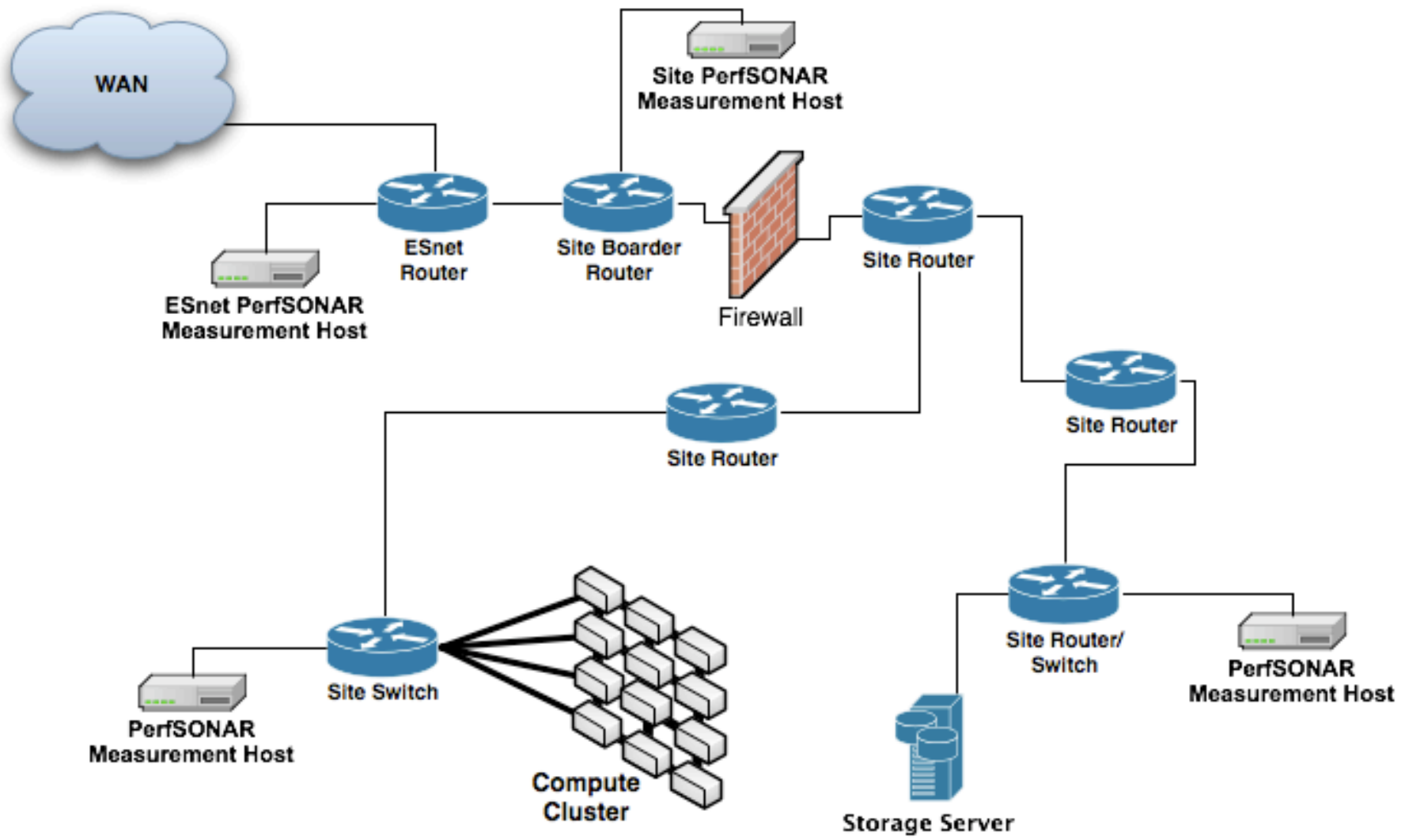
perfSONAR Hardware

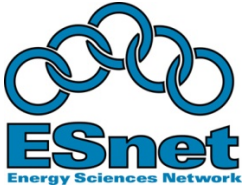
- Requires dedicated hardware (not virtual servers)
- Copy somebody else, or try before you buy...
 - ESnet deployed hardware details at
 - <https://performance.es.net/PMP.html>
 - Sample host configuration for PS Performance Toolkit
 - http://fasterdata.es.net/ps_howto.html
 - Find somebody with the class of machine that your looking for and ask them how it works!





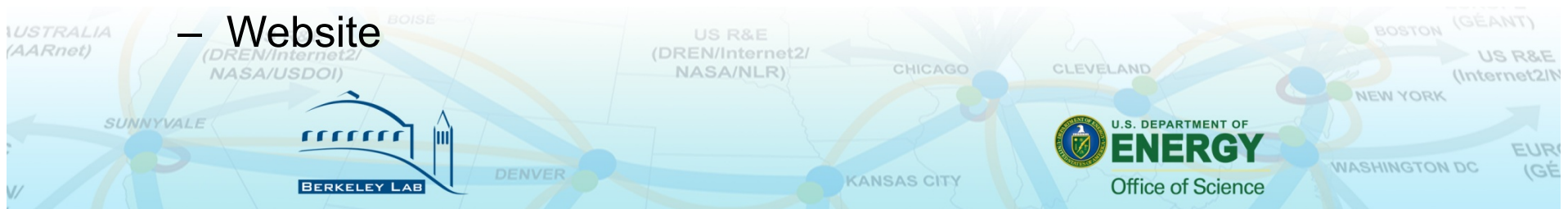
Typical Campus Deployment





Developing a Measurement Plan

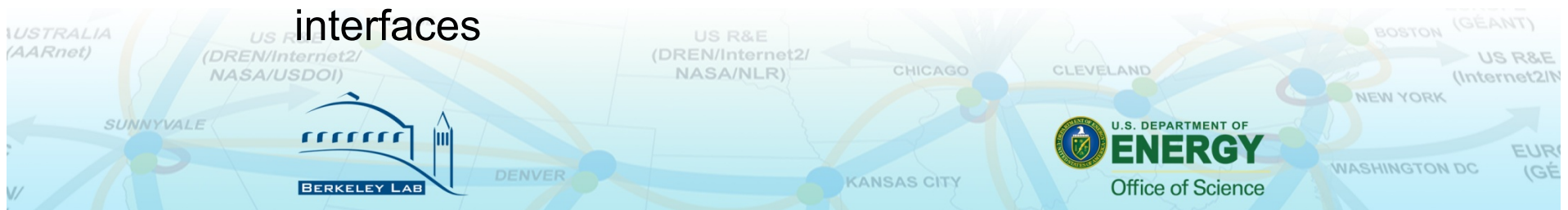
- What are you going to measure?
 - Achievable bandwidth
 - 2-3 regional destinations
 - 4-8 important collaborators
 - 4-12 times per day to each destination
 - 20 second tests within the NA, longer to EU or Asia
 - Latency
 - OWAMP: ~10 collaborators over diverse paths
 - Pinger: Important collaborators who don't support owamp
 - Interface Utilization & Errors
- What are you going to do with the results?
 - NAGIOS Alerts
 - Reports to user community
 - Website





Deploying a perfSONAR measurement host in under 30 minutes

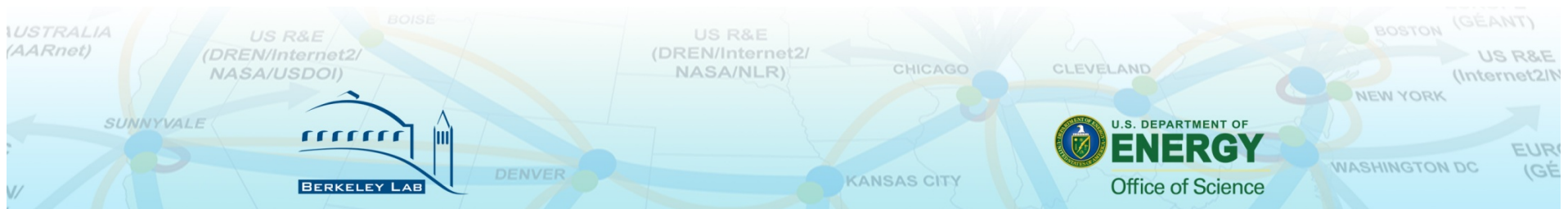
- Using the PS Performance Toolkit is very simple
 - Boot from CD
 - Use command line tool to configure
 - what disk partition to use for persistent data
 - Network address and DNS
 - User and root passwords
 - Use Web GUI to configure
 - Select which services to run
 - Select remote measurement points for bandwidth and latency tests
 - Configure Cacti to collect SNMP data from key router interfaces





Measurement “Communities”

- The PS Performance Toolkit lets you specify which measurement community you measurement host is meant to service
 - Sample communities: LHC, DOE-SC-LAB, Internet2, ESnet, Climate, etc.
- This makes it easier to locate other measurement hosts of interest





Example: US Atlas

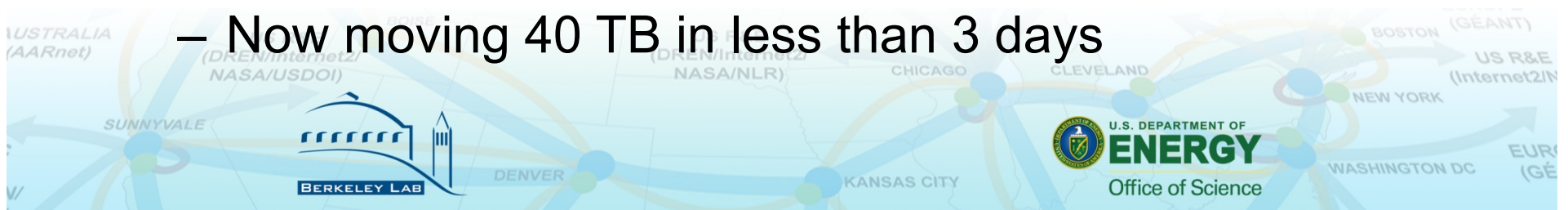
- Tier 1 to Tier 2 Center Data transfer problem
 - Couldn't exceed 1 Gbps across a 10GE end to end path that included 5 administrative domains
 - Used perfSONAR tools to localize problem
 - Identified problem device
 - An unrelated domain had leaked a full routing table to the router for a short time causing FIB corruption. The routing problem was fixed, but the router started process switching some flows after that.
 - Fixed
 - Rebooting device fixed the symptoms of the problem
 - Better BGP filters configured to prevent reoccurrence (of 1 cause of this particular class of soft faults)

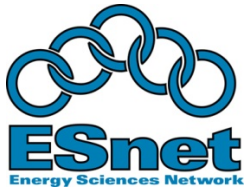




Example: NERSC & OLCF

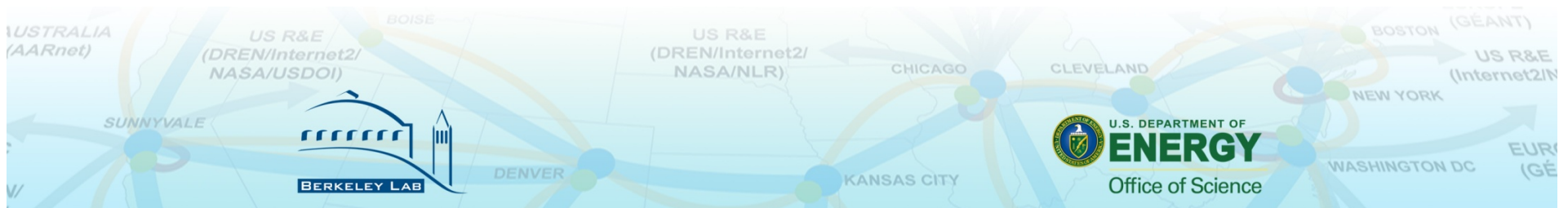
- Users were having problems moving data between supercomputer centers
 - One user was: “waiting more than an entire workday for a 33 GB input file”
- perfSONAR Measurement tools were installed
 - Regularly scheduled measurements were started
- Numerous choke points were identified & corrected
- Dedicate wide area transfer nodes were setup
 - Tuned for Wide Area Transfers
 - Now moving 40 TB in less than 3 days





How to Participate

- Deploy perfSONAR
- Use perfSONAR to find & correct the hidden performance problems in your networks.

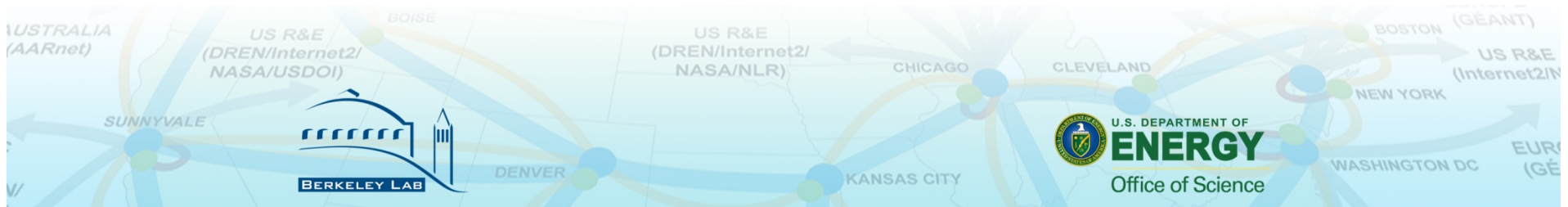


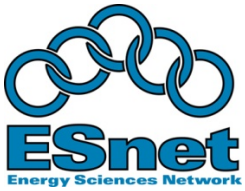


Firewalls

If your server is behind a firewall, you need to open the following ports

- Open to Global perfSONAR Servers
 - Lookup Service -- open port tcp/8095
- Open to perfSONAR Users
 - SNMP MA -- open port tcp/8065
 - PingER -- open port tcp/8075
 - perfSONAR-BUOY -- open port tcp/8085
 - bwctl -- open port tcp/4823, edit `/usr/local/etc/bwctld.conf`, set `peer_port` to a value, open the tcp port for that value, and edit `/usr/local/etc/bwctld.conf`, set `iperf_port`, `thrulay_port` and `nuttcp_port` to a specific range, and open the tcp/udp ports for those ranges.
 - owamp -- open port tcp/861, edit `/usr/local/etc/owampd.conf`, set `testports` to range, open the udp ports for that range
 - NDT -- open port tcp/3001, open port tcp/3002, open port tcp/3003, open port tcp/7123
 - NPAD -- open port tcp/8100, open port tcp/8200
- Open for local management
 - Apache HTTP Server -- open port tcp/80, open port tcp/443
 - SSH -- open port tcp/22





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)

