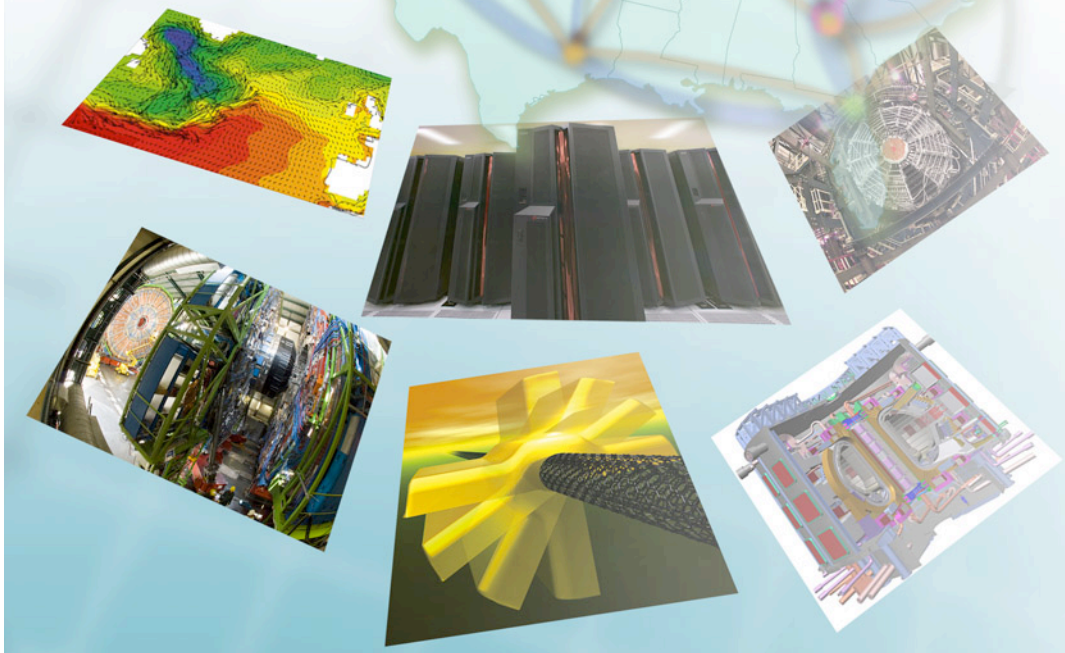
ESnet
Energy Sciences Network

ESnet Network Measurement Current Status & Future Plans

Joe Metzger

- with slides copied from from Jeff Boote, Brian Tierney, Aaron Brown & others on the perfSONAR team

July 24th 2008
ESCC meeting



Networking for the Future of Science

Berkeley Lab

Office of Science
U.S. Department of Energy

Overview

- ESnet Services
  - New content
- LHC Requirements
  - Presented by me at Joint Techs yesterday
- PerfSONAR Knoppix distribution
  - Presented by Jeff Boote yesterday
- PerfSONAR Knoppix distribution experiences
  - Presented by Brian yesterday
- PerfSONAR Futures

Bandwidth Test Platforms

- Deployment
  - Plan: a tester at every hub and every 10GE site
  - Currently ~20 in the field and ~20 at LBL
    - Deploying & Upgrading them with the SDN equipment deployment.
- Current Uses
  - Acceptance Testing
  - Debugging network problems
- Additional uses once we resolve some problems and get more systems deployed
  - Regular scheduled bandwidth testing
    - Internal to ESnet IP & SDN Network
    - External to important peers, and ESnet sites that are interested
  - Dynamic end-user testing via
    - BWCTL & perfSONAR
    - ESnet Performance Center (maybe)
    - OSCARS

Bandwidth Test Platform Configuration

- Interfaces
  - nfe0: GE for management traffic only
  - Vlan911: A vlan on mxge0 for normal tests
  - Vlan3600: A vlan on mxge0 for OSCARS testing
    - Will have a known address in 10.0.0.0/8
    - OSCARS users can create dynamic circuits terminating at test systems
- Proposed Test Management Plan
  - BWCTL will be used to handle locking and short term schedule conflicts
  - There will be a graduated scale of different user classes that can perform tests with different parameters based on remote IP prefix
    - ESnet staff can initiate unlimited tests.
    - ESnet sites can run up to 60 second bi-directional TCP, and 60 second UDP receive only.
    - Research & Education community can run up to 60 second TCP tests capped around 1.5 Gbps.
    - Sessions coming from Commodity Internet prefixes will be discarded
    - We strongly encourage using perfSONAR based tools for running regularly scheduled tests.
  - The testers are a shared resource that consume limited bandwidth.
    - Please use responsibly
    - Contact us if you want to run regularly scheduled tests.

Latency Test Platforms

- Deployment Status
  - 11 in field, 10 at LBL & 9 more on order.
- Most are using CDMA clocks
  - Signal not available at all locations
  - Signal degrades over time as POPs are populated
- IRIG
  - Original plan to get IRIG sources from Level 3 at several locations is not going smoothly.

Latency Test Platform Configuration

- All systems will run NTP, OWAMP, & perfSONAR BOUY
  - Several are also providing NTP Stratum 1 time sources for ESnet sites
  - ESnet and Internet2 are each going to provide 2-4 Stratum 1 Time sources for ESnet Site, University and partner OWAMP probe deployment.
- Scheduled Tests using PerfSONAR BOUY
  - A sparse mesh between points on the ESnet backbone
  - A sparse mesh between points on the ESnet backbone and to important peers.
- Some ESnet systems will be available for testing to sites
  - Add-hoc using owping
  - Scheduled tests using perfSONAR BOUY

ESnet Statistics Collection Systems

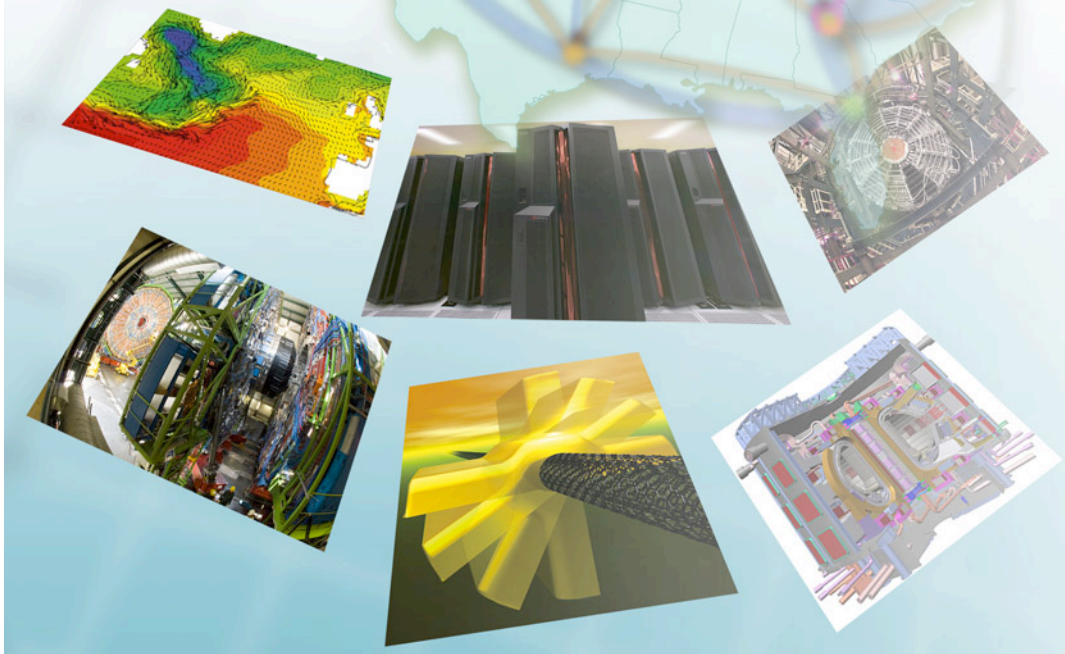
- SNMP collection Systems
  - Raw Collectors
    - Used for monthly statistics reports
  - MRTG
    - Several different collections for:
      - ESnet Info
      - NetInfo
  - ESxSNMP & TSDB
    - ESxSNMP - SNMP collection system
    - TSDB Time Series Data Base
    - Developed in-house by Jon Dugan
    - Main focus is improved meta-data management
    - Open source (code.google.com)
    - Should replace all other ESnet SNMP stats collections systems in the next 6 months or so.
- NetFlow collection system

Network Monitoring

Joe Metzger

Energy Sciences Network
Lawrence Berkeley National Laboratory

July 22 2008
Joint Techs Workshop



Networking for the Future of Science

Current Network Environment

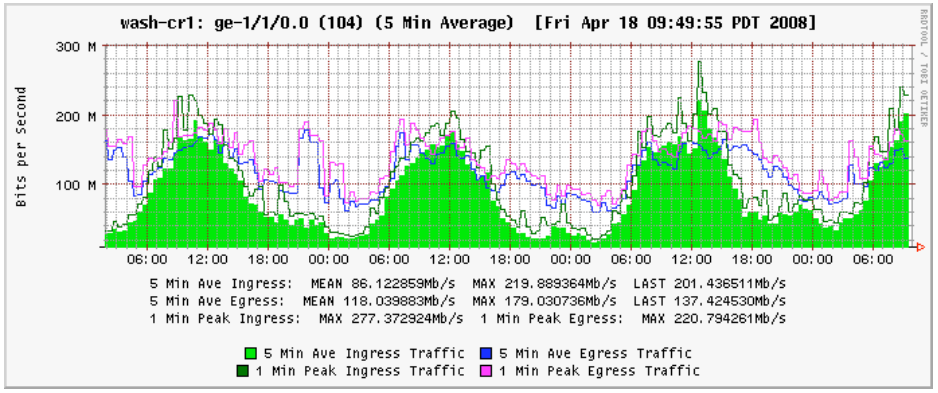
- Most R&E network backbones are composed of 10Gbps links
- The LHC community has the tools, techniques, infrastructure & capability to transfer data at 10Gbps.
- But…
  - Network topology is **constantly** changing!
  - LHC data transfer flows are not typical internet flows
  - Many network operators don't have a lot of experience with large flows
  - Most physics flows cross multiple domains
  - Many cross-domain links haven't been tested at capacity
  - **Large flows don't aggregate nicely**
  - Debugging problems can be difficult

Measurement Requirements

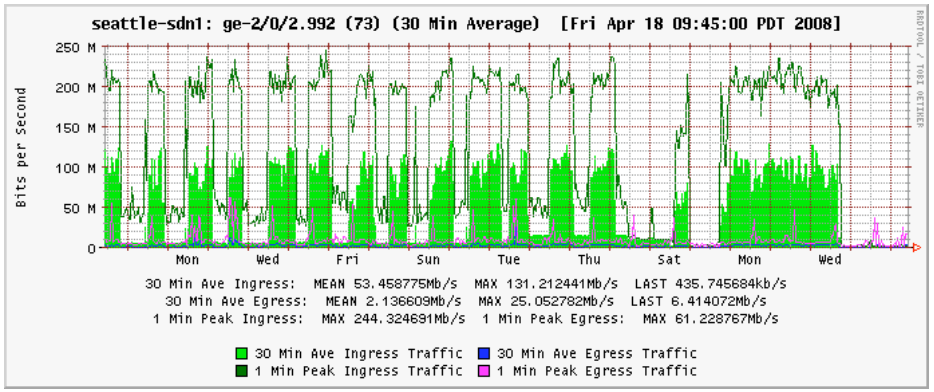
- Users must have the ability to easily determine the status of the set of paths they rely on for their critical missions.
  - Up and working correctly?
    - **How do you prove it?**
  - Down
    - Is there a known problem that is being worked on?
      - Are you seeing a symptom of the problem or something else?
    - Is part of the network down or the applications down?
    - How do you prove the problem is, or is not in your cluster/campus/regional?
    - Who do you call and **what hard data can you provide to help them quickly identify the problem and fix it?**
  - Up but not performing as expected.
    - Is there a known problem?
    - Who do you call and **what hard data can you provide to help them quickly identify the problem and fix it?**
- Users need to understand the amount network resources they are consuming.
  - Are they getting more, less, or approximately their fair share?
  - Do they know if/how their use of the network is affecting others?

New Network Traffic Profiles

- Old Typical Traffic Pattern
  - Lots of small flows

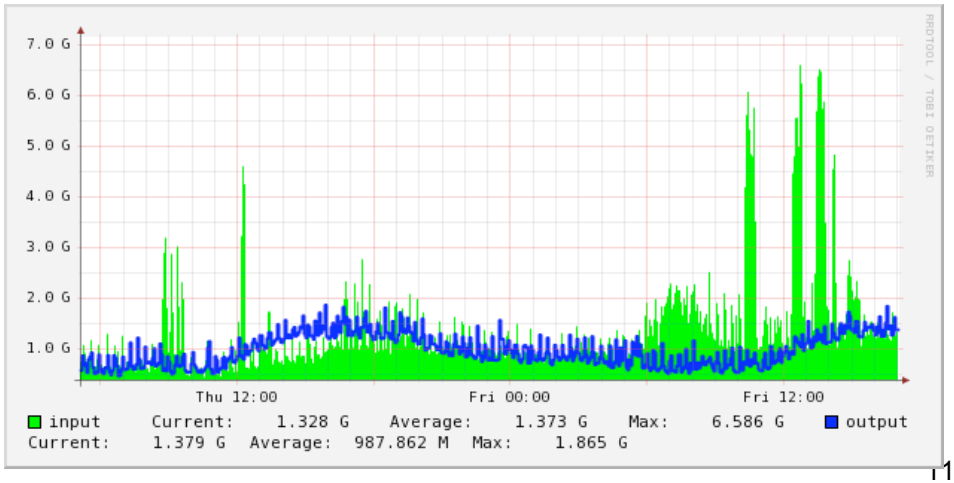


- Steady State Instrument Output or Bulk Data Transfer



- Tuned Bulk Data Transfer

How many 5-7 Gbps flows can you aggregate on a 10 Gbps backbone?



Community Progress

- There has been a lot of work in the Network Measurement space
  - Developing frameworks for exchanging measurement data
  - Developing & improving measurement tools
  - Defining diagnostic methodologies
  - Analysis techniques
- There is a small community that understands how to use these tools and techniques for network performance analysis, verification and debugging
- The LHC community is taking advantage of these capabilities
  - Metcalfe's law - The value of a network is proportional to the square of the number of users.

LHC US Tier 1/2/3 Measurement Document(s)

- Introduction
  - A general discussion of the value of network measurement to the community.
- Best Practices
    - What Measurements to support
      - Delay, Bandwidth, Interface Utilization, Errors & Discards, etc
    - Protocols
      - For measurement collection: ICMP, OWAMP, iperf etc.
      - For measurement Publication & Sharing - perfSONAR
    - Schedules & parameters
      - For regularly scheduled tests
    - Data sharing guidelines
- Implementation Guide
  - What tools to use
  - How to configure them
- Usage Guide
  - How to use the Measurement Infrastructure to:
    - Verify important network infrastructure is running properly
    - Debug problems efficiently

Technical Requirements Latency

1. Continuously measure end-to-end **delay**
   A. What
      - Run continuous tests and store results in an MA
      - Publish results via a standardized web service interface
      - Provide a tool to visualize the data
      - Provide tools to automatically analyze data and generate NOC alarms

   B. Why
      - Measure & document actual availability
      - Provide time references for when problems occurred and when they were fixed
      - Detect & assist in diagnosing common causes of performance degradation
        a. Packet Loss
           - Congestion related
           - Non-Congestion related
        b. Queuing & Jitter caused by congestion
        c. Routing Issues: changes, asymmetry, flapping, etc

Technical Requirements Bandwidth

2. Make regular scheduled **bandwidth** measurements across paths of interest

   A. What
      - Run regularly scheduled tests and store results in an MA
      - Publish results via a standardized web service interface
      - Provide a tool to visualize the data
      - Provide tools to automatically analyze data and generate NOC alarms

   B. Why
      - Detect performance problems
      - Identify when problems appeared
      - Document performance delivered

Technical Requirements Circuit Status

3. Monitor up/down status of cross domain circuits
   A. What
      - Determine the status of a circuit
      - Publish status via a web services interface
      - Provide tools to visualize state
      - Generate NOC alarms when circuits change states
   B. Why
      - Determine when circuits are available
      - Simplify debugging of end to end circuit problems

Technical Requirements Interface Statistics

4. Monitor Link/Circuit Capacity, Utilization & Errors

A. What

- Publish statistics via a web services interface
- Provide tools to visualize the data
- Generate NOC alarms when thresholds are crosses

B. Why

- Allow determining usage patterns
- Simplify throughput problem diagnosis
- Capacity Planning

Technical Requirements Topology

5. Measure & Publish Topology of primary and backup paths
   A. What
      - Publish logical topology via a web services interface
      - Provide tools to visualize the data over time
   B. Why
      - Set user expectations
      - Facilitate network problem diagnosis
      - Allow correlating logical topology to measurements of the physical topology
      - Understand …

Technical Requirements Discovery

6. Find deployed measurement servers

A. What

- Be able to find new measurement servers
- Be able to find available services

B. Why

- Measurement services will be dynamic.
  - Redundant architecture
- Some measurement services are connected to dynamic networks.
- The servers we deploy for any particular project are useful to a much larger community.

Implementation Details Latency

1. Continuously measure end-to-end **delay**
   A. Tools
      - OWAMP/perfSONARBOUY
      - Pinger

   B. Configuration
      - Each Domain deploys a Measurement Point at the edge of their domain
        - OWAMP + Pinger
      - Deploy Scheduler & MA
      - Run OWAMP tests to remote sites that have an OWAMP server
      - Run Pinger tests to remote sites that are not able or willing to maintain stable Owamp MPs.

Implementation Details Bandwidth

2. Make regular scheduled **bandwidth** measurements across paths of interest

A. Tools

- BWCTL & BWCTL MP
- perfSONAR-BUOY

B. Configuration

- Deploy 1 GE connected MP at the edge of their domain
- Deploy 1 Scheduler & MA per cluster of MP's
  a. One for the LHCOPN
  b. One per Tier 1 that wants to measure their Tier 2 service

Implementation Details Circuit Status

3. Monitor up/down status of cross domain circuits

   A. Tools

      – Topology + Transform service or E2Emp or SQLMA
      – E2Emon

   B. Configuration

      – Each Domain publishes the status of their portions of cross domain circuits.
      – E2Ecu monitors all LHC circuits?
      – Any NOC can run E2Emon to monitor the subset of circuits that they have responsibility for

Implementation Details Interface Statistics

4. Monitor Link/Circuit Capacity, Utilization & Errors
   A. Tools
      – PS-SNMP MA
   B. Configuration
      – Each domain sets up a Measurement Archive publishing statistics about their network interfaces supporting LHC
        a. Capacity
        b. Utilization
        c. Input Errors
        d. Output Drops

Implementation Details Topology

5. Measure & Publish Topology of primary and backup paths
   A. Tools Still Under Development
      – Internet2 Topology Service

This is not a significant concern for the LHCOPN as long as it continues to be a well defined static topology fully described with the E2Emon tools.

This is an issue when considering Tier 2 traffic which will stress the global R&E Networking Infrastructure!

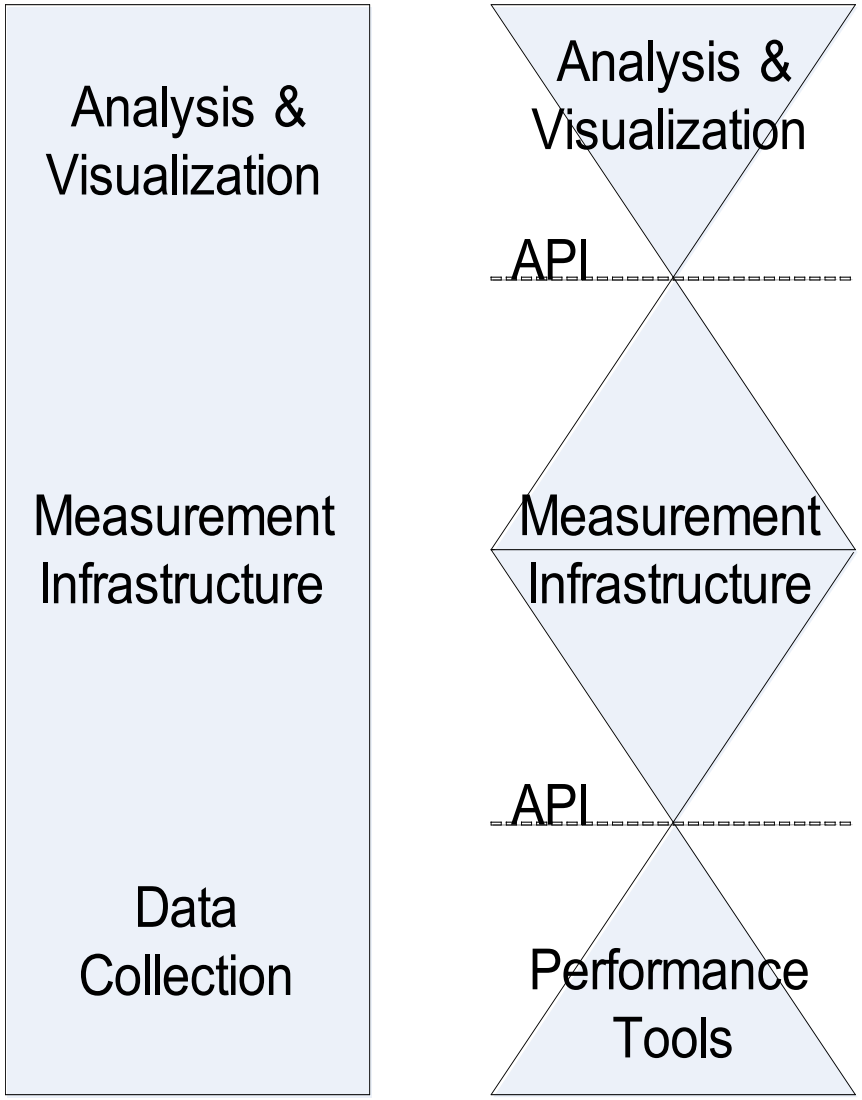
Implementation Details Discovery

6. Find deployed measurement servers
   A. Tools
      - LS: Lookup Service
        - All perfSONAR services register with 1 or more LS's in their domain.
          - Service access points
          - Meta-data about the services they provide
      - GLS
        - All LSs will register with 1 or more GLSs
        - ESnet, Internet2, GEANT & RNP will run the 'root' GLSs.
   B. Configuration
      - Services can automatically discover a close LS and register with it (or be manually configured.)

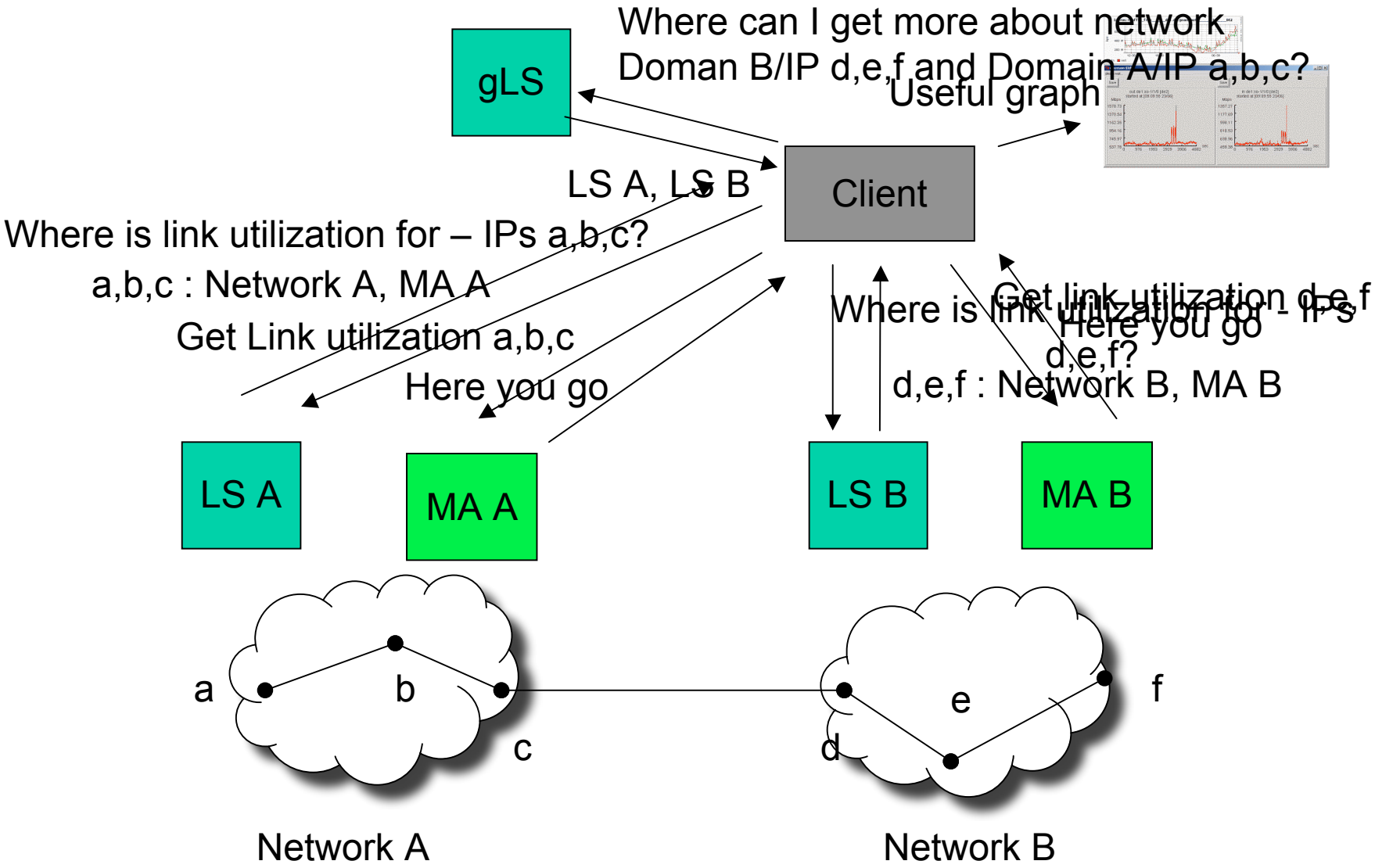
Next Steps

- Identify community representatives to participate in writing & reviewing document
  - Backbone Network Measurement specialists
  - Physics Users
  - Tier 1 center network operators
  - Campus Networking Person
- Present idea & current status at Internet2 Spring Member Meeting HENP-SIG
- Present draft recommendations to the US LHC community at the T1/T2/T3 meeting at BNL May 2008
  - https://wiki.internet2.edu/confluence/display/PSPS/Tier-2-BCP
- Evaluate the 'US Recommendations' applicability to the global environment at LHCOPN meeting in June
- Present recommendations & pilot implementations at Joint Techs in July
- US LHC community using infrastructure before end of summer

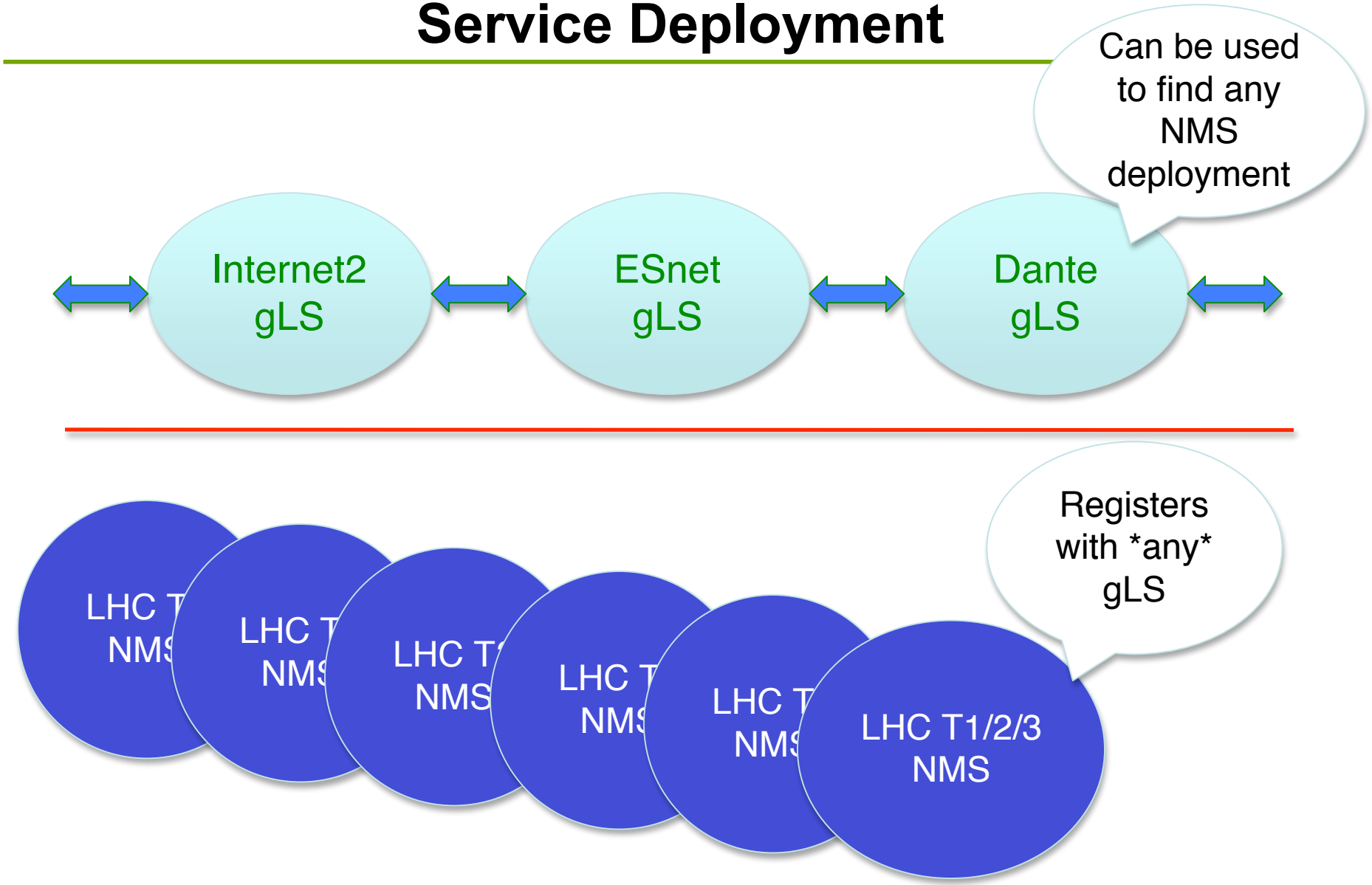
Decouple 3 phases of a Measurement Infrastructure



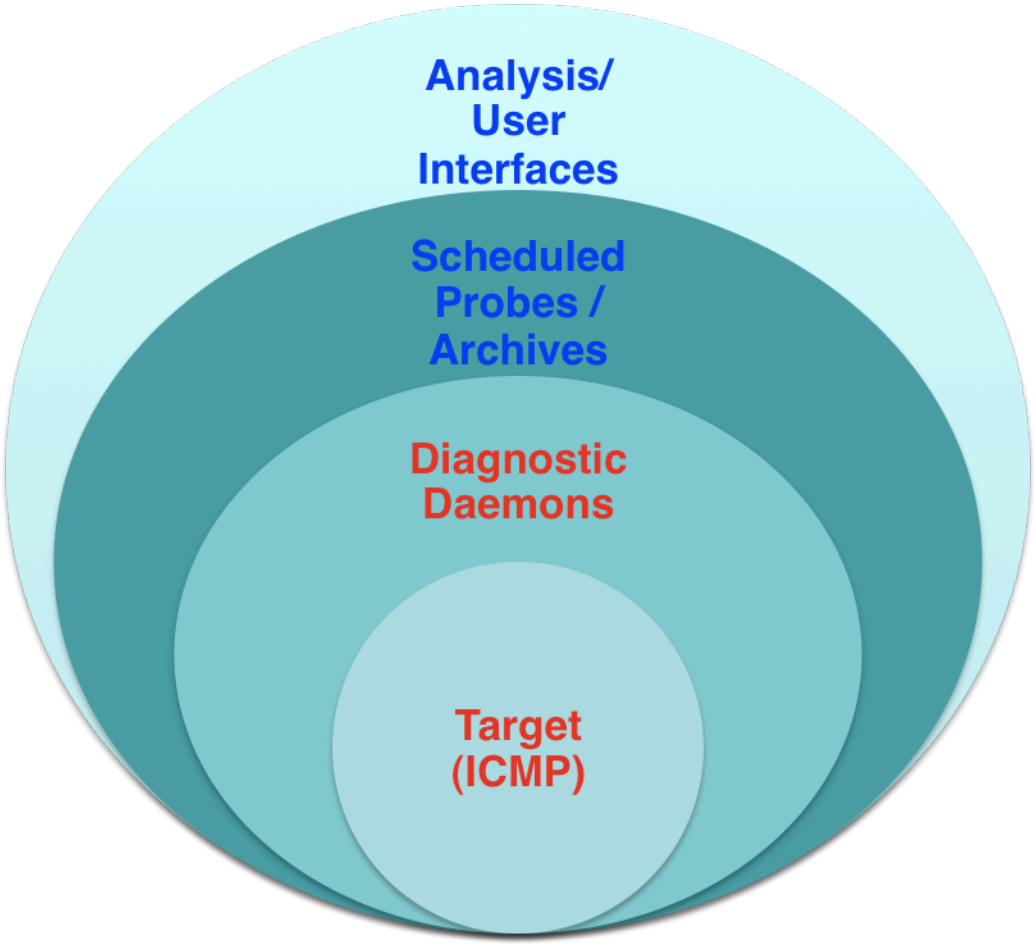
Example perfSonar client interaction



Service Deployment



LHC Site NMS Deployment



* Required
* Optional

Required Deployment

Functionality

- Host with ICMP access
  - Need to be able to ‘ping’ and ‘traceroute’ to somewhere on the site
- Diagnostic Daemons
  - NDT
  - OWAMPD
  - BWCTLD
- Registration of availability

Resources required

- Accessible host (firewall modifications likely)
- Modest linux systems (two)
- Must run a daemon that registers tool availability to gLS

Minimal Hardware Requirements

- 2 Dedicated hosts (~$500/host)
  - Differentiate network tests from LHC application tests
  - Gives ‘local’ servers to test LHC servers against
  - 1 for latency related tests (and pS infrastructure tasks – will move if this causes problems)
    - Minimal CPU/memory/disk requirements
      - Recommend minimum 2.0 Ghz/1GB
    - Best if power-management disabled, and in temperature controlled environment
    - Nearly any NIC is ok
      - Recommend a 10/100/1000 Mbps NIC (on-board is fine)
  - 1 for throughput related tests
    - CPU/NIC needs to be ‘right sized’ for throughput intensities
      - Recommend minimum 2.0 Ghz/1GB
      - Recommend 10/100/1000 Mbps NIC (on-board is fine)

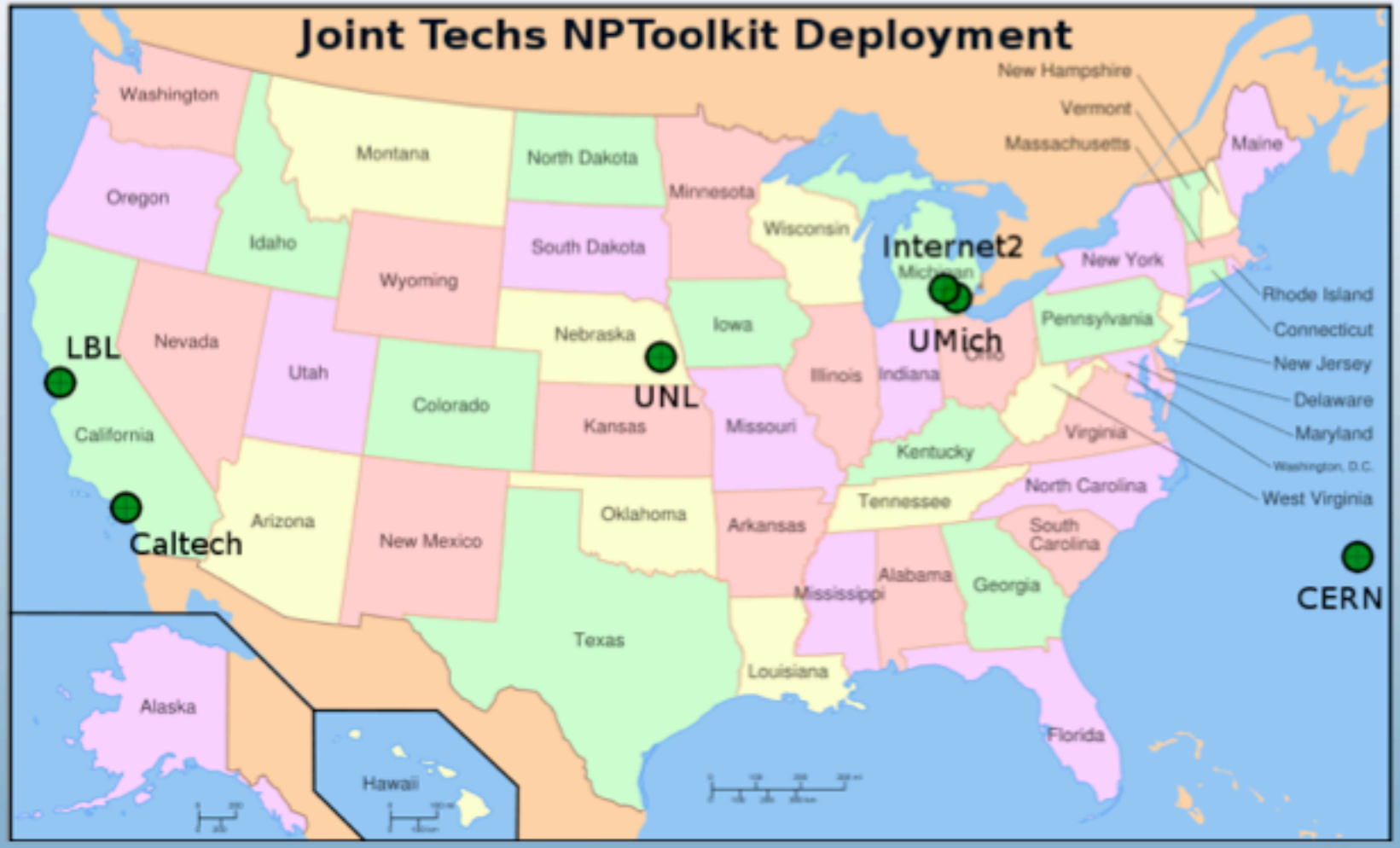
Jt Techs Deployment

- Deployment Locations:
  - UNL, LBL, USLHC (cal-tech), CERN, UM, Internet2(office)

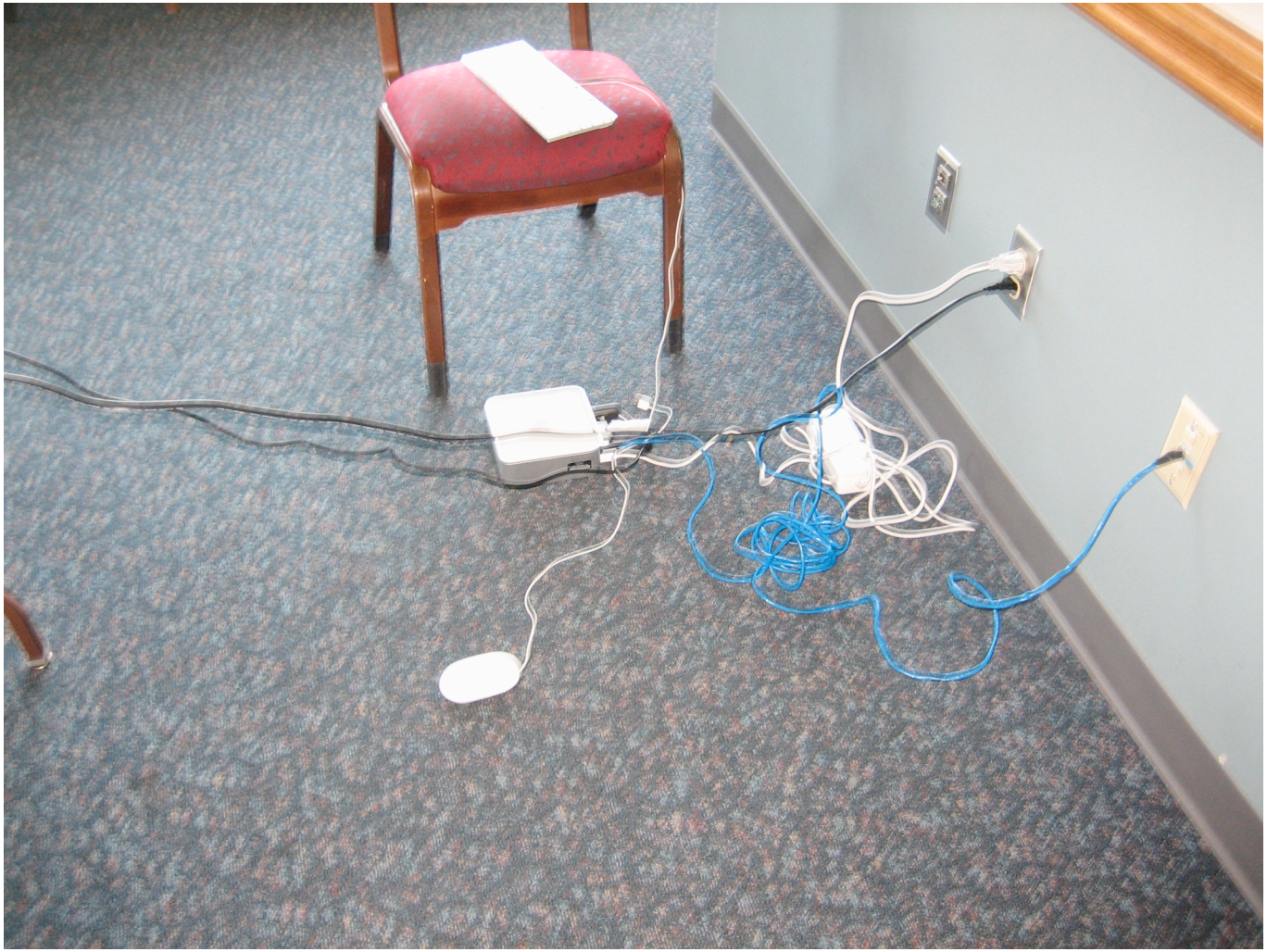
- Diagnostic Services
  - NDT
  - NPAD
  - BWCTL
  - OWAMP

- Services
  - SNMP MA
  - PingER MA
  - perfSONAR-BUOY (BWCTL/OWAMP scheduled)

NPToolkit JT Demo



UNL Demo System



Installation

- Takes < 30 minutes
- Steps
  - Download iso file (5 min)
  - Burn CD (5 min)
  - insert CD, hit reset, and wait for it to boot up (10 min)
  - answer configuration questions (5 min)
- Most services are now up and running!

Additional Configuration

- To run SNMP collector for SNMP MA (measurement archive), need to configure cacti
  - http://code.google.com/p/perfsonar-ps/wiki/CactiConfig
- Might want to add additional pingER hosts
  - web interface for this
- Might want to configure regular throughput or delay tests
  - web interface for this

Current Issues

- Bug where configuration settings not saved across a reboot
  - fix for this in the next release
- Currently come up with both http and https services running
  - I would like an option to force https only

Future needs

- Brian's list of needed missing components
  - Test client to monitor all perfSONAR services that LHC cares about
  - Better support for redundant services
  - Better support configuration update services
  - Better support for grouping services into "communities"

More Information

- FAQ available here:
  - https://wiki.internet2.edu/confluence/display/PSPS/LHC+NPToolkit+FAQ

PerfSONAR Status Summary

- Current Knoppix Solution
  - ✓ Schedule, Manage & Publish Latency Measurements
  - ✓ Schedule, Manage & Publish Bandwidth Measurements
  - ❑ Measure & publish Circuit Status
  - ❑ Measure (or determine) and Publish Topology Data
  - ✓ Support Discovery of Measurement Infrastructure Components
- Deployment Next Steps
  - Deploy Beta Knopix disk at several LHC Centers this summer.
  - Work on configuration & management issues
  - Next beta will be suitiable for larger community deployment
- PerfSONAR group next steps
  - Working on updating perfSONAR Architecture & Vision white papers
  - Topology & Discovery integration with OSCARS/DCN
  - AAA

How can you get involved?

- Think about your network measurement needs.
  - Who are your important customers, providers, users, etc
- Think about your measurement publication issues. Can you publish:
  - Latency Measurements
  - Bandwidth Measurements
  - Border Interface Utilization, Error & Capacity (mrtg)
  - Topology Data
  - Circuit Status Information
- Is there any AAA system that covers your entire network measurement user base?
- Experiment with Knoppix/Live CD

How can you get involved?

- Think about your labs cross domain network measurement needs.
- Think about data publication issues.
  - Foobar
    -
  - Could you make the following data publically accessible