LHC Path Monitoring Tools Deployment Planning

Jeff Boote
Internet2/R&D

May 27, 2008
US ATLAS T2/T3 Workshop at UM

Overview

- Value of network performance monitoring and diagnostic tools
- Recommended tools and services
  - Software distribution mechanisms
- Hardware recommendations
  - Number and types of hosts
- Network issues
  - Location
  - firewalls

INTERNET®

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
  - Line rate flows don't aggregate nicely
  - Debugging problems can be difficult

INTERNET®

Measurement Requirements

- You must have the ability to easily determine the status of the set of paths you rely on for your 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?**
- Do you know if your use of the network is affecting others?
  - Are you getting more, less, or exactly your fair share?

INTERNET2®

Monitoring vs Diagnostics

Monitoring: Constantly/Consistently ‘doing something’ to ensure things are working as you expect

Diagnostics: Performing some individual action to determine if there is a problem, or to determine the cause of a problem

- The same (or very similar) tools are used to perform these actions
- Regular monitoring can trigger alarm, analysis, and then diagnosis
- Diagnosis is aided by historical monitoring

- Both of these activities are required, but they can have slightly different best practices
- Need to make an engineering trade-off

Implementation Considerations

Constraints

1. Different LHC participants are interested in performance over different paths
   - Sites must be able to monitor from their site to other sites of interest (T2 want to actively probe upstream T1s and downstream T3s)
2. Must support monitoring as well as diagnostic interactions
3. Must gracefully degrade given lower levels of participation from sites
4. Diagnostics should aid applications to make performance choices
5. Different analysis should be made available to different users
6. LHC sites have different influence over their ‘local area’ network (T3’s administrative boundary is likely the physics department)

Implications

1. A single monitoring appliance or service is impractical
   - Scaling issues become intractable beyond the T0/T1 paths (mostly due to support issues, but data is more difficult as well)
   - Locally configurable system per-site with prescribed minimal support
2. Solution needs to integrate on-demand with scheduled probes
3. Solution needs ability to determine what diagnostics or tools are available from remote sites
4. Application interfaces must be open, available and usable
5. Analysis (GUI) applications must be available and usable (and new analysis tools must be easy to create)
6. Measurement host placement can be varying topological distances from the LHC compute servers

INTERNET2®

Proposed System Architecture

- perfSONAR
  - Autonomous (federated) measurement deployments
  - Global Discovery
  - Web services SOA solution exposing existing diagnostics and monitoring tools

INTERNET®

What is perfSONAR

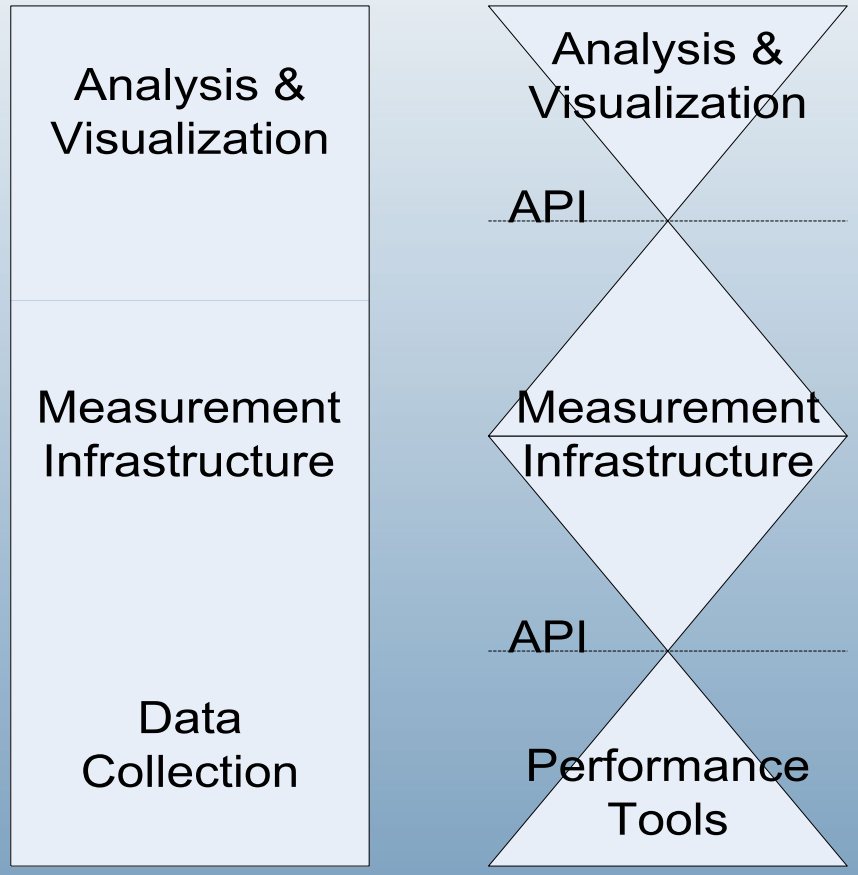
- A collaboration
  - Production network operators focused on designing and building tools that they will deploy and use on their networks to provide monitoring and diagnostic capabilities to themselves and their user communities.
- An architecture & a set of protocols
  - Web Services Architecture
  - Protocols based on the Open Grid Forum Network Measurement Working Group Schemata
- Several interoperable software implementations
  - Java, Perl, Python…
- A Deployed Measurement infrastructure

INTERNET®

perfSONAR Architecture

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

Decouple 3 phases of a Measurement Infrastructure



perfSONAR Services

- Measurement Point Service
  - Enables the initiation of performance tests
- Measurement Archive Service
  - Stores and publishes performance monitoring results
- Transformation Service
  - Transform the data (aggregation, concatenation, correlation, translation, etc)

These services are specifically concerned with the job of network performance measurement and analysis

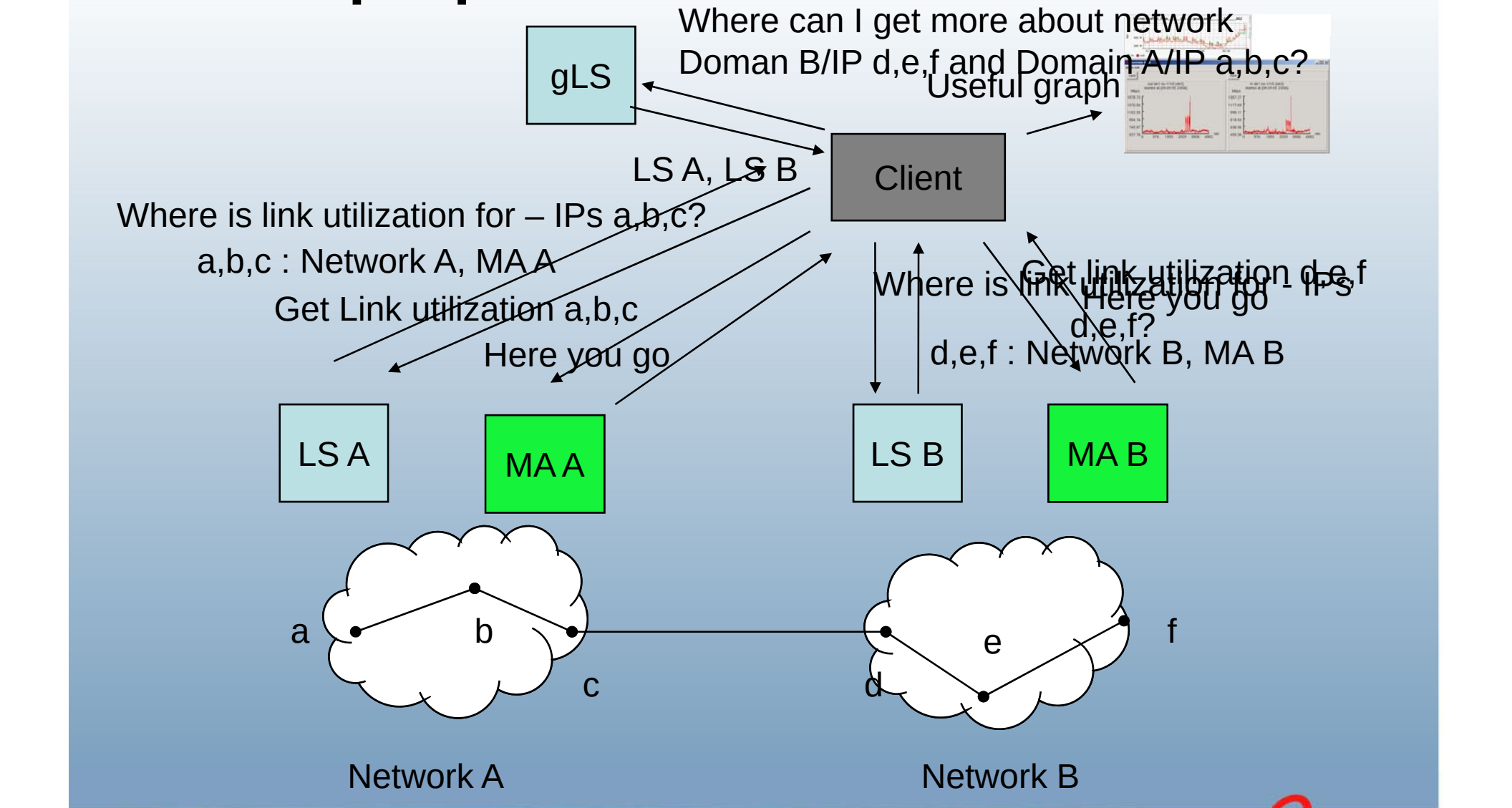
Information Services

- Lookup Service
  - Allows the client to discover the existing services and other LS services.
  - Dynamic: services registration themselves to the LS and mention their capabilities, they can also leave or be removed if a service goes down.
- Topology Service
  - Make the network topology information available to the framework.
  - Find the closest MP, provide topology information for visualisation tools
- Authentication Service*
  - Based on Existing efforts: Internet2 MAT, GN2-JRA5
  - Authentication & Authorization functionality for the framework
  - Users can have several roles, the authorization is done based on the user role.
  - Trust relationship between networks

These services are the infrastructure of the architecture concerned with the job of federating the available network measurement and diagnostic tools

* Proposed deployment does not include Auth Service – too much work on the political side to be practical in the short term

Example perfSonar client interaction

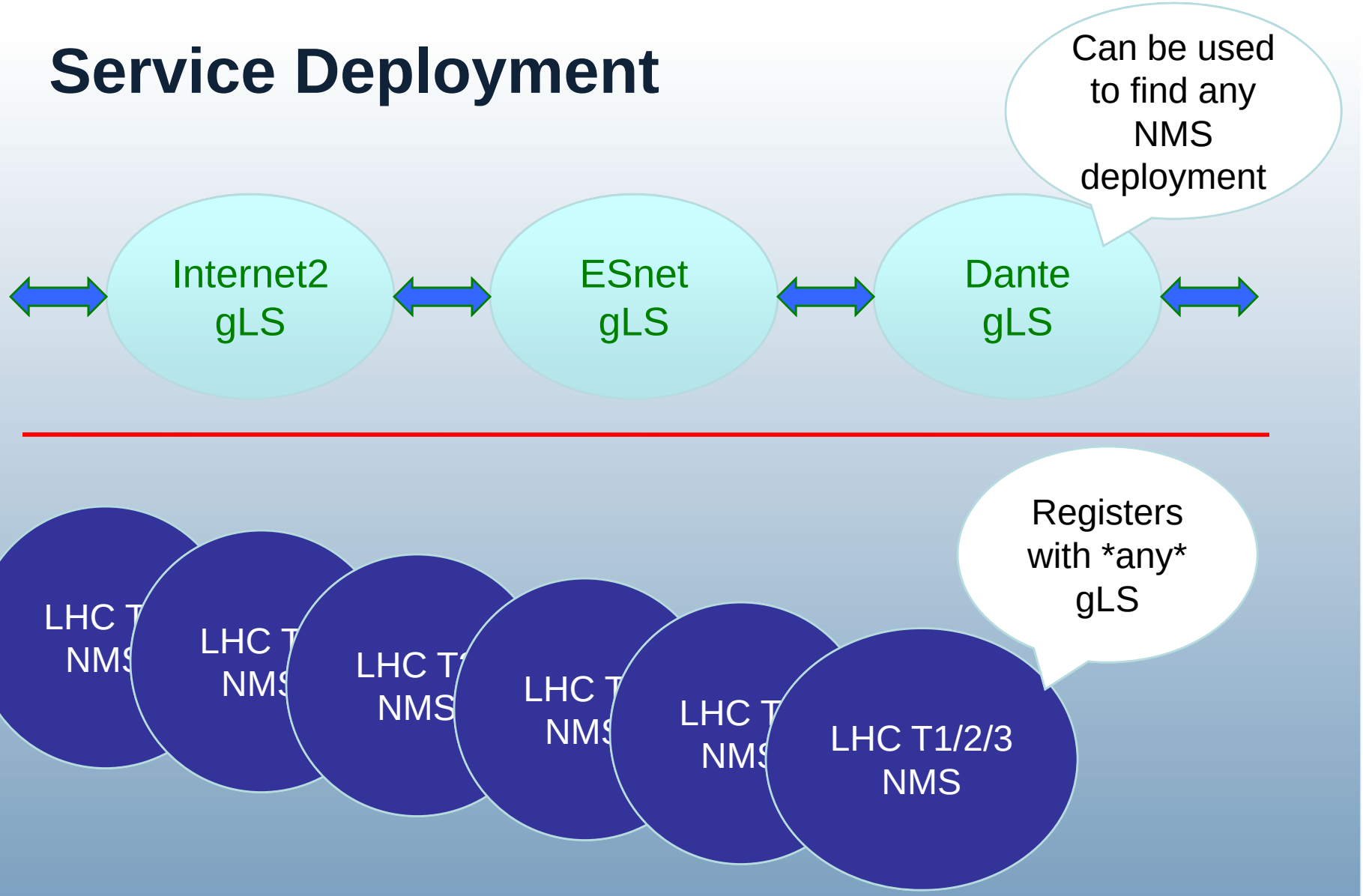


LHC Deployments

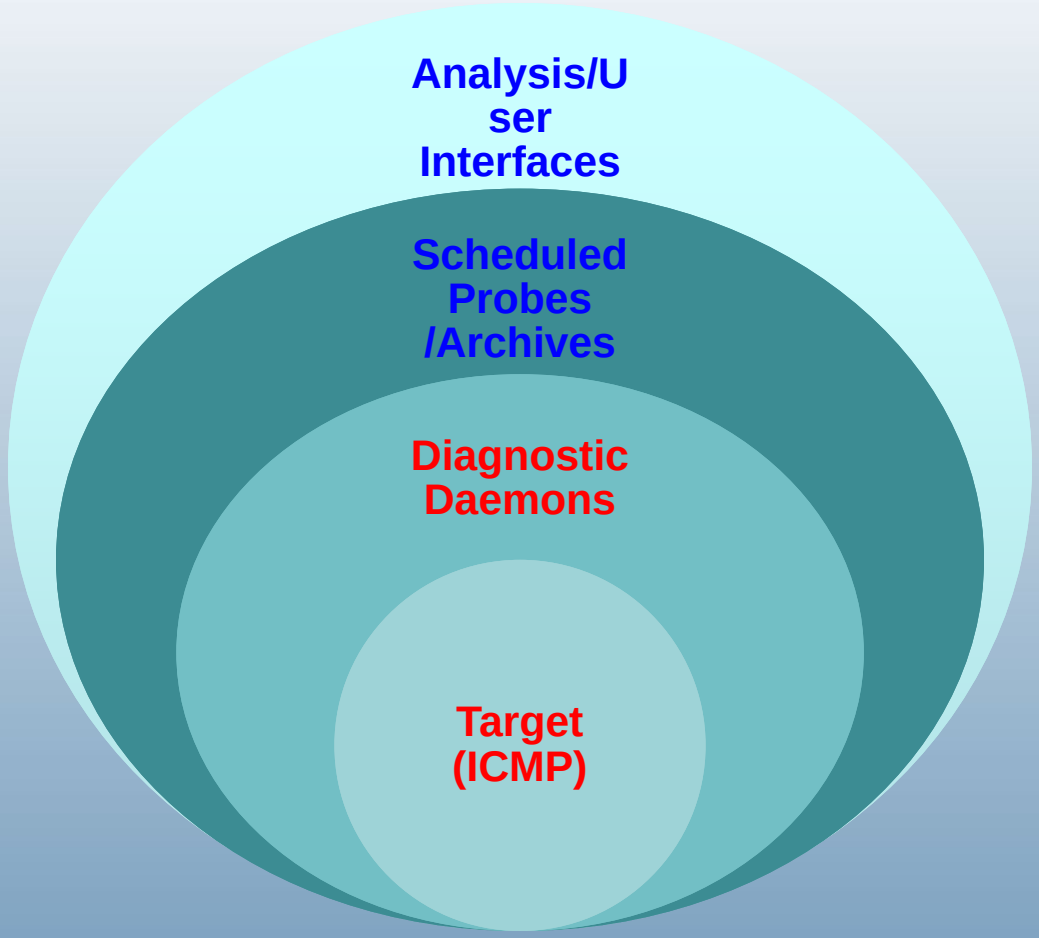
- Specific set of available services at each participant site (Network Measurement System – NMS)
- Small number of ‘support’ services supported by backbone network organizations
- Complimentary diagnostic services at backbone network locations
- Analysis clients
  - Some directly available to end researchers, some specifically designed for NOC personnel

INTERNET®

Service Deployment



LHC Site NMS Deployment



* Required
* Optional

INTERNET2®

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

INTERNET®

Optional Deployment

Functionality

- Data archiving
  - Regularly scheduled probes
    - pingER/owamp/bwctl
  - Circuit Status/Utilization

Resources required

- linux system with reasonable amount of disk space (possibly two)
- Configuration to interact with existing infrastructure

INTERNET®

Hardware Requirements

- 2 Dedicated hosts (~$400/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)

INTERNET®

Software Deployment Issues

- Difficult software prerequisites
  - Web100 kernel (for NDT)
  - Oracle XMLDB (for archives and info services)
    - This is the free open-source XMLDB formerly known as SleepyCat
- Deployment options
  - NPT (Network Performance Toolkit) knoppix disk
    - Current version has most ‘required’ functionality
      - Only lacks ‘registration’ for lookup-service
    - Intend to have LHC related perfSONAR services on disk by July Jt Techs conference
  - RPM installs
    - pS-PS development team can support 32-bit RHEL5 RPMS directly
    - Looking for community involvement to support additional OS/hardware architectures
  - Source installs

Network Issues

- Should be deployed as ‘close’ to the administrative boundary as possible
  - Administrative boundary can be the physics department. The point is to differentiate network issues from computer server issues. This allows tests to be run from the computer servers to the ‘local’ nms hosts to do this
  - Aids in path dissection
    - Backbone networks already deploying, Some regional research networks as well (Pushing deployments from the middle out, and the edges in to support this)
- NMS hosts will likely need specialized firewall rules
  - Throughput/latency/diagnostic tests can be run on specific range of ports (even specific to local policies)
  - Web services must be able to be contacted
  - In general, any outgoing traffic must be allowed – incoming traffic can be more specific to local policies

More details… the diagnostic tools

pingER (RTT latency)

Description

- Regularly run ping and collect results

Provides

- Availability
- Time reference for problems
- Some insight into reasons for performance degradation

OWAMP (One-Way latency)

- Description
  - Daemon to request and run one-way latency tests
- Provides
  - Diagnostic
    - Additional insight into reasons for performance degradation (direction helps, more sensitive to jitter)
    - Some routing issue insight (hops/directional latency jumps)
  - Regular probes with archive
    - Availability
    - Time reference for problems

INTERNET®

More details… the diagnostic tools

BWCTL

- Description
  - Daemon to request/run iperf tests (now supports multiple streams)
- Provides
  - Diagnostic
    - Detect problems by using the network as the user would
  - Regular probes w/archive
    - Document what is possible
    - Document ‘when’ performance issues start

NDT

- Description
  - Web browser invoked advanced diagnostic testing to indicate why a particular performance was achieved. (detailed diagnostic information is available to pass on to network engineers)
- Provides
  - User accessible diagnostic tool
  - From the ‘client’ perspective – give useful results to someone that can do something about it
  - Provides the user with a more accurate expectation of performance by informing them of the bottleneck

INTERNET®

More details… the archives

Link/Circuit status

- Using whatever backend is appropriate (SNMP/TL1 etc…) archive the up/down state of ‘important’ circuits
  - Topology service hooks to correlate paths to circuits under development
- Generate NOC alarms for multi-domain circuits

SNMP MA

- Archive utilization/errors
- Capacity planning
- Simplify thoughput problem diagnoses
- Insight into usage patterns

Optional services – only useful in specific contexts

INTERNET®

More details… the archives

Topology Service

- Publish local topology of interest to remote users
- Used to determine paths of interest
- Visualize topology for knowledgeable user
- Still under development, but useful as is

perfSONAR Client Developments

- Most tuned to specific services currently
- Different user focus (micro vs macro view)
- These represent what is possible – I would expect that LHC participants would want something more tuned to what they care about

- Client applications
  - perfSONAR-UI (acad.bg)
  - Fusion (Internet2)
- Web Based
  - GMAPS (SLAC)
  - Domain Utilization Browser (ESnet)
  - pS-PS Weathermap (Internet2)
  - pingER Analysis (FNAL)
  - perfAdmin (Internet2)
  - CNM (DFN)
  - E2EMon (DFN)

Please see my presentation at the LHC T1/T2/T3 Networking Workshop for more details on these:
https://indico.bnl.gov/conferenceDisplay.py?confId=80

INTERNET2®

And Finally…

Summary

- A scalable infrastructure for providing network performance information to interested LHC participants (humans as well as applications)
- Open Source licenses and development model
- Multiple deployment options
- Interfaces for any application to consume the data
- Internet2 (and ESnet and partners are committed to supporting these tools)

More information

http://www.internet2.edu/performance/pS-PS

http://e2epi.internet2.edu/bwctl/

http://e2epi.internet2.edu/ndt/

http://e2epi.internet2.edu/owamp/

http://www-iepm.slac.stanford.edu/pinger/

Internet2 Community Performance WG

https://mail.internet2.edu/wws/info/performance-announce

INTERNET®

www.internet2.edu