Deploying perfSONAR at DOE Laboratories

Editor:[1] Joe Metzger, Metzger@es.net

DRAFT Version1.7

July 22, 2009

I. ESCC Recommendations

DOE Laboratories should deploy a perfSONAR-based network measurement infrastructure to support and simplify diagnosing wide area network performance problems and provide better monitoring wide area service levels.

A Lab with a gigabit or faster wide area connection should deploy 2 dedicated active measurement hosts (latency & bandwidth) and a perfSONAR router/switch interface measurement archive service. The interface measurement archive can be installed on the active bandwidth test system, or some labs may choose to locate it elsewhere. The active measurement servers should be deployed as close to administrative boundaries as possible to make the resulting data more actionable.

DOE Laboratories with lower speed connections may be able get by with a single server that runs all of the services. The drawback to this configuration is that active bandwidth tests will perturb the latency test results, making analysis and automated problem detection more complex.

Facilities that have significant WAN network performance dependencies, and that are located deep within a laboratory or campus network should consider deploying their own active measurement systems. This can be very helpful when debugging local problems that are round trip time dependent, such as buffer queue depth problems, cross traffic, and switches that are not capable of line rate forwarding. Multiple test hosts also help to "triangulate" a problem – if the test server connected to the site border router achieves excellent wide area performance and the test server co-located with the large-scale science resource can only achieve mediocre results, this usually points to a problem within the site's local network that is difficult to diagnose with just one test server.

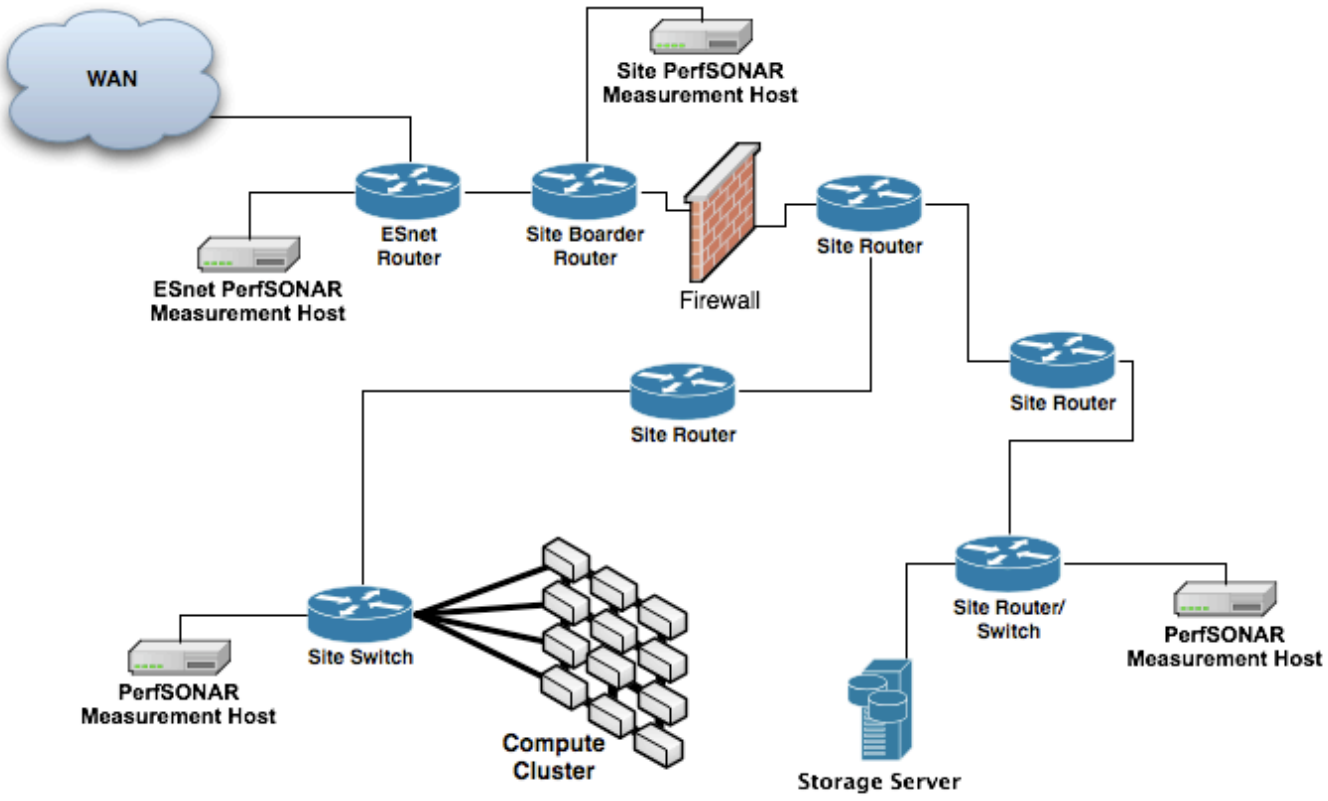

Figure 1: Typical Site Deployment

[1] This document is being developed by the community with contributions from many people.

II. Background

2.1 PerfSONAR Introduction

PerfSONAR[2] is an infrastructure for network performance monitoring, making it easier to solve end-to-end performance problems on paths crossing several networks. It contains a set of services delivering performance measurements in a federated environment.

PerfSONAR is also an international consortium of organizations that are developing and standardizing network measurement protocols and software that implements the framework.

2.2 Motivation

Many people in our community are not achieving the high levels of network performance that our Research & Education networks are capable of delivering. Some users are not getting good performance because of soft network failures that punish high performance flows while allowing low bandwidth connectivity. Some users don't realize that commonly used tools such as scp and rsync over ssh have known problems over high bandwidth wide area networks that can be easily worked around. Many users don't even realize that the performance they are getting is sub-par, and don't know that they should (or even that they can) improve their performance by configuring their end systems and reporting problems to network support organizations that can implement fixes.

Deploying a perfSONAR network measurement infrastructure can help address these issues by identifying soft failures in the network so they can be fixed, and by setting expectations appropriately by demonstrating the capabilities that exist.

When something is measured, it can be improved. By running continuous or regularly scheduled measurements and analyzing/reporting the results, you know when the network is performing properly, and when it is broken. Often knowing the time that performance degraded provides critical information that makes the diagnosis of soft failures easier.

Soft Failures

In the arena of high-performance networking, it's easy to track down "hard failures" such as when someone breaks or cuts through a fiber link. But identifying "soft failures," like dirty fibers or router processor overload, has been hard. Such soft failures still allow network packets to get through, but can cause a network to run 10 times slower than it should.

The most commonly used networking protocol, TCP, was intentionally designed to gracefully recover from failures and to work (albeit at significantly reduced speeds) in environments with very high error rates. The result of this design decision is that TCP provides low throughput rates across paths with a multitude of problems and hides all of the errors.[3] These soft failures can be hard to detect and track down, in part because one error can mask all of the others. Another common characteristic of soft failures is their affects are proportional to round trip time.

Common examples of soft errors are: devices that are not forwarding packets at line rate, devices at bottleneck links with insufficient queue depths, packet losses from failing hardware, etc.

Network Design Issues

There are several LAN design considerations that need to be taken into consideration to support high performance wide area traffic flows. These include segregating high bandwidth wide area flows from local traffic, tuning switch buffers to support wide area flows, and considering high bandwidth requirements when deploying network security infrastructure such as firewalls. LANs designed without considering these issues frequently deliver good local area performance and poor wide area performance. *The result is that users assume that the problems are in the wide area and outside of their control, instead of identifying and fixing the local problems.*

[2] http://www.perfsonar.net

[3] Modern WAN circuits have bit error rates better than 10^{-12}, TCP doesn't fail completely until error rates exceed 10^{-2}.

Please note that the current active measurement tools require a stable system clock with NTP time synchronization. This means the tools do not run well in a virtual server environment, or in an environment with significant temperature fluctuations.

III. Deployment Guidance

3.1 Hardware Considerations

It is important to have reliable, known good hardware for the measurement systems so that there is confidence that issues in the measurement results are due to network issues, and not server constraints. Current low-end servers are sufficient for measurement archive services and latency testers. The hardware requirements for the active bandwidth testers depend on the capabilities of the network they are measuring. Current low-end servers are sufficient for gigabit or slower tests.

Many types of network problems can be discovered by doing tests using gigabit Ethernet test hosts. Facilities that are required to support network flows that regularly exceed 1Gbps should deploy a 10 gigabit active bandwidth test infrastructure.

Examples of the hardware used successfully by ESnet, and others in the community can be found at http://fasterdata.es.net/ps_howto.html.

The active network measurement systems should be run directly on physical systems. Running them in a virtual environment will generate virtually meaningless data due to the virtual nature of the network interfaces and the system clock.

3.2 perfSONAR Software

The perfSONAR software is available in several different form-factors. By far the easiest get up and running is the PS-Performance Toolkit put together by Internet2. The toolkit is a bootable CDROM that easily turns a blank machine into a fully functioning perfSONAR appliance running a recent Linux operating system. This distribution includes tools to configure the local system, setup scheduled latency and bandwidth tests, query routers for interface data and export that data via an perfSONAR measurement archive, and register all of the perfSONAR services with the global lookup service. This approach works well in environments where there is a need for a simple to use, hands off, contained solution. The PS-Performance Toolkit also contains several additional network diagnostic tools such as NDT, which can help when diagnosing network problems.

The other approach is to install packages of individual perfSONAR services on a system with an existing operating system. This is a bit more involved, and the WEB GUI configuration options are not available. This approach may be required in environments where there are policy or management constraints for systems providing public network services (e.g.: requirements for a single, common OS version or automated security patches).

3.3 perfSONAR System Types

Active Bandwidth Tester

The Active Bandwidth tests systems are primarily used for debugging throughput problems, documenting capabilities, and identifying network design or implementation problems. The Active Bandwidth test system should run **bwctl** and perfSONAR-BUOY (**pSB)**.

Bwctl is a wrapper program that coordinates invocation of iperf, nuttcp, or other bandwidth test programs, and serializes test requests to prevent overlapping tests.[4] It also contains mechanisms to limit acceptable test parameters by classes, which are configured based on IP address, or shared keys. DOE labs should allow TCP tests up to 60 seconds in length. Un-authenticated outbound UDP tests should be disabled to limit the potential for these systems to be used in DOS attacks.

pSB is a perfSONAR tool that manages registration of active test capabilities in the perfSONAR lookup service, scheduling of active tests as well as archiving & publication of test results.

[4] http://e2epi.internet2.edu/bwctl

Lookup registration is taken care of automatically when running a PS-Performance Toolkit node. Laboratories not using the PS-Performance Toolkit should register their pSB nodes with the ESnet hLS's (home lookup service).

One concern with scheduling active bandwidth tests is that the tests themselves might consume network resources that are needed by other applications. The following schedule recommendations apply to labs that are not suffering serious capacity challenges. Labs with capacity limitations may want to modify them.

Suggested Test Schedules	
Nearest ESnet Test System	6 to12 Times per day
Farthest ESnet Test System	4 to 6 times per day
Important Collaborators	4 to 6 times per day

All scheduled tests within a continent should be **20 seconds in duration**. This might need to be bumped up to 60 seconds for tests over 1 Gbps *to other continents* if necessary to allow TCP windows to open up. The sum of all bandwidth tests in a day should be well under 1% of WAN access capacity. Tests approaching or exceeding 1% should use 'less than best effort' QOS.

Laboratories that maintain a Nagios infrastructure should use Nagios to monitor perfSONAR processes and utilize the Nagios plugins to generate appropriate alarms if the active bandwidth results fall outside acceptable limits.

Active Latency Tester

The active latency tests are designed to identify and quantify queuing delays, packet loss patterns and characterize congestion. They do this by running continuous tests at very low packet rates. The latency test software applications are **OWAMP** and **pSB**.

OWAMP is a software system that measures delay using a one way ping.[5] This software uses accurate time-stamps in the packets to determine how long it takes the packet to transit the network. Analyzing the distributions of the measurements can provide significant information about the network path between 2 points.

pSB is a perfSONAR tool that manages registration of active test capabilities in the perfSONAR lookup service, scheduling of active tests as well as archiving & publication of test results.

The continuous active latency test systems consume negligible network resources. So, the decision process about which tests to set up should be based on the network traffic characteristics and should cover the paths to the top 6-12 most common network destinations. In general the list of remote test systems should include:

- Nearest ESnet test system.
- 2-3 Distant ESnet test systems.
- 1-6 Important collaborators or network supporting important collaborators.

To monitor the latency to sites not running owamp, one can install and configure the perfSONAR **PingER** service.

Router Interface Measurement Archives

The router interface measurement archive service allows you to publish utilization, errors and discard data about switch and router interfaces. This allow network engineers and end users to easily locate bottleneck links, and identify, or rule out probable sources of problems when all domains along a path publish this information. (e.g.: see the ESnet Traceroute Visualizer: (URL HERE). The tool used for this is the perfSONAR SNMP MA.

The perfSONAR SNMP MA is primarily a publishing tool. It relies on other tools such as cacti[6], or mrtg[7] to gather the actual statistics in most situations. This separation of functionality allows easy integration with existing network management practices.

DOE laboratories should publish as much router & switch interface data as possible. The minimal set of information should include the labs external interfaces, and any obvious bottleneck links near the external border. Ideally all of the router and switch interfaces between the external border and common data transfer nodes should be published.

[5] http://e2epi.internet2.edu/owamp
[6] http://www.cacti.net/
[7] http://oss.oetiker.ch/mrtg/

3.4 Deployment Roadmap

Deploy Infrastructure

1. Download the latest PS-Performance Toolkit distribution, or perfSONAR-PS distribution from http://www.perfsonar.net.
2. Follow the instructions to install.
    a. Using the PS-Performance Toolkit
        i. DOE Laboratories should configure their systems with the *keyword* "DOE-SC-LAB".
    b. Other distributions:
        i. Register individual perfSONAR services with the ESnet hLS systems at:
        http://ps1.es.net:8095/perfSONAR_PS/services/hLS
        http://ps4.es.net:8095/perfSONAR_PS/services/hLS
        ii. DOE Laboratories should configure their systems with the *keyword* "DOE-SC-LAB".
    c. NTP Configuration
        i. ESnet sites should configure their latency measurement points to NTP peer with all of the following:
            1. Local stratum 1 or 2 time sources
            2. 2 closest ESnet owamp servers
            3. Chronos.es.net
            4. Saturn.es.net
        ii. Bandwidth measurement points should NTP peer with
            1. Local Stratum 1 or 2 time sources
            2. Chronos.es.net
            3. Saturn.es.net

Identify & Correct initial problems

1. Verify all services are up and running, and are registered in the perfSONAR Global Lookup Service. You can do this by opening up the administration interface in a NPT node, or by going to the *Active PerfSONAR Services* link on http://fasterdata.es.net. Note: It may take several hours for the information to propagate to all the global lookup services.
2. Verify that NTP time synchronization is working on all measurement points. Look at the ntp state using ntpq –p localhost
    a. Offset from the system peer (*) should be within 0.1
    b. Jitter or dispersion from the system peer should be less than 0.1
    c. There should be at least 5 peers.
3. Frequently initial bandwidth test results will be significantly lower than expected. This typically indicates that there is a problem with the local infrastructure or TCP tuning on the server. See http://fasterdata.es.net for information about diagnosing and correcting the problems.

Setup Continuous Testing & Monitoring to quickly identify future problems

Setup the regularly scheduled tests, and Nagios alarms to alert processes go down and when results are outside of expected parameters.

Debugging Problems

http://fasterdata.es.net has a wealth of information and pointers for debugging network performance problems.

IV. Security Considerations

The security environments at DOE Laboratories can make deploying any network service including measurement infrastructures challenging. Engineers at several laboratories are working on security plans for deploying perfSONAR systems at their facilities. We will be collecting as much useful information from those processes as possible and adding it to future versions of this white paper, and posting it on the www.perfsonar.net website.