LHCONE Diagnostic Service

This is a strawman proposal to engender community discussion.

Eric Boyd 1/30/12 6:00 PM
Comment [1]: Note, we might want to extend the service definition to include recommended hardware specs.

Goal
The goal of the LHCONE Diagnostic Service is to enable:
1) Monitoring the health of LHCONE.
2) Diagnosing and fixing problems detected by LHONE.

Objectives
The specific objectives of the LHCONE Diagnostic Service include:
1) On-demand Layer 3 end-to-end performance tests (IPv4 and, where supported, IPv6) between any two LHCONE sites, and two intermediate networks, or any LHCONE site and any intermediate network;
2) A sparse mesh of regularly scheduled Layer 3 tests (IPv4 and, where supported, IPv6) between any two LHCONE sites;
3) A full mesh of regularly scheduled Layer 3 tests (IPv4 and, where supported, IPv6) between any two directly connected networks (including both LHCONE site networks and intervening networks)
4) Collection of passive Layer 3 measurement data applicable to LHCONE links in use by the LHCONE VRF service.
5) Collection of passive Layer 2 measurement data applicable to LHCONE links in use by the LHCONE VRF service.
6) Collection of passive measurement data applicable to dynamically created VLANS created through the LHCONE Point-to-Point VLAN service.

Implementation Strategy
The implementation strategy for the LHCONE Diagnostic Service is to:
1) Get something good up and running quickly.
2) Not let "perfect" be the enemy of the good.
3) Iterate and improve.
4) Start with the DICE Diagnostic Service definition in use and deployed by ESnet, GEANT, and Internet2. (Implicitly this means address Objectives #1, #2, and #3 and don't address Objectives #4, #5, and #6.)
5) Extend the DICE Diagnostic service to cover all participating networks and end sites participating in LHCONE (both the LHCONE VRF service and the LHCONE Point-to-Point Service).
6) Revise the combined LHCONE / DICE Diagnostic Service definition as operational experience accumulates.
7) Extend the LHCONE Diagnostic Service definition to address Objectives #4, #5, and #6.

Bandwidth Measurements

All LHCONE domains will establish bandwidth measurement points in their network. The anticipated uses of active bandwidth measurements are:

- Identify paths that cannot sustain high bandwidth TCP or UDP sessions.
- Demonstrate paths can sustain high bandwidth TCP or UDP sessions.
- Generate test data streams that can be analyzed to characterize network performance problems.

The ideal configuration is to have measurement points capable of sourcing and sinking link capacity (currently 10G) TCP and UDP measurement streams from other service participants near each network border. This will simplifying diagnosing a class of problems that occurs between 1G and 10G that are of growing importance to our community.

LHCONE will maintain 10G measurement points at the interconnection points. It's recommended that downstream domains should place measurement points close to their upstream connection. Where possible, domains place the equipment with as minimal amount of network infrastructure between the measurement point and the circuit interconnect as possible.

The bandwidth measurement points will be configured to accept measurement requests to and from the bandwidth measurement points of the other participants service. The diagnostic service will utilize IPERF for performing achievable bandwidth measurements between measurement points. The IPERF tests will be invoked solely with BWCTL by relevant organizations. BWCTL will handle short duration scheduling to serialize multiple requests and prevent overlapping tests. BWCTL may in turn be invoked by perfSONAR tools.

All of the measurement points will accept at least 60 second long inbound and outbound TCP requests. Requests should allow window sizes of at least 32 MB. The bandwidth measurement points will be configured to use a modern TCP stack which is more aggressive than RENO, and which is representative of the TCP stacks used by and recommended to a domains user community. The recommended default TCP stack is CUBIC or HTCP.

Eric Boyd 1/30/12 6:02 PM

Comment [2]: Need at least 10% of the link capacity. More would be better for debugging. Use BWCTL limits to protect.

All of the measurement points will accept at least 10 second long UDP requests at rates up to 100 Mbps. The relevant tool will notify the user if they request a test with a duration that is not supported by the appropriate networks. The envisioned use case involves active measurement archives that mirror common video streams.

By default, bandwidth tests should not specify a QOS tag, and the traffic should be treated as best effort within a domain. QoS tagging should be allowed if specifically requested, though there is no expectation that the end-to-end path will support or honor the QoS tagging.

The bandwidth measurement points will register the information about their configurations that are necessary to interpret test results. This includes test interface capacity, TCP algorithms in use, version information about the OS, BWCTL, IPERF.

On-Demand Bandwidth Measurements

The diagnostic service will support on-demand achievable bandwidth measurements (BWCTL) between all of the participating bandwidth measurement points.

Regularly Scheduled Bandwidth Measurements

The diagnostic service will support regularly scheduled achievable bandwidth measurements (BWCTL) between all of the participating bandwidth measurement points. The goal is to develop historical information for diagnosing future problems. It is expected that a domain will run a limited number of tests (~4 per day) to a moderate number of remote domains of particular interest to their community.

Eric Boyd 1/30/12 6:03 PM

Comment [3]: Probably not the best way to say this. If a domain is testing to 12 or so other domains and running tests every 6 hours (e.g. the estimate of 4 a day) they still run 6x12=72 tests. And all of the other domains are testing to them as well, adding in more.

Measurement schedules between LHCONE domains should be developed with consideration for the bandwidth and utilization of the cross-connects. The participants agree to limit the amount of regularly scheduled tests to a 'reasonable' amount. Because this is a complex distributed problem, it is difficult to define reasonable. The following definition is an attempt to define reasonable: The normal agreed test schedule is that each domain will not set up scheduled tests that will consume more that 0.1% of the total aggregate capacity to a neighbouring domain on a daily basis. These suggested limits apply to adjacent domains. Any downsteam networks will exercise care to avoid unreasonable traffic levels. The 0.1% figure is not to be exceeded without prior negotiation. There are no current technical controls to prevent domains from exceeding this hence we are working on a trust model at present. Each domain will take steps to measure usage of test data using whatever means at their disposal.

It is expected that the perfSONAR tools provide suitable logging facilities to enable the network operators to analyse the impact of the use of them on their network.

The following example works out the maximum schedules between two hypothetical sites.

- There is 20 Gigabits of capacity between the 2 domains.
- 20 Gigabits * 60 seconds * 60 minutes * 24 hours = 1.72 Petabits.
- 1.72 Petabits * 0.001 (one tenth of one percent) = 1.72 Terabits
- Assuming tests average 1 Gigabit per second, the maximum schedule between should be a sum of 2160 seconds across all tests. This would be a maximum of 36 60 second tests, or 108 tests that are 20 seconds long per day. This includes tests from Net A to Net B test points, and tests from Net A to other connectors behind Net B.

The previous formula provides the maximum agreed to test configuration. Typical test configurations will consume significantly less capacity. The following are the recommended regularly scheduled tests rates for TCP tests.

- Frequency: Bandwidth tests should be scheduled no more than once in a 6 hour interval. The time should be randomized within the interval by 10 %.
- Duration: Tests should be long enough for TCP to achieve its maximum throughput for at least 50% of the duration of the test. The following initial guidelines may need to be adjusted depending on test host capacity, test parameters (window size), and the TCP stacks in use at both ends.
- 30 second long tests should be sufficient for measurements within a continent or from the Southern US to South America.
- Coverage: There should be sufficient tests configured so that each domain measures all of their direct connections to participating adjacent domains 4 times a day in each directions.

Eric Boyd 1/30/12 6:04 PM
Comment [4]: Change to 60 seconds?

If regularly scheduled inter-domain UDP tests are desired, the duration and bandwidth should be negotiated on a case-by-case basis by the NOC engineers involved.

Eric Boyd 1/30/12 6:05 PM
Comment [5]: Recommend testing it 1 time per day to all sites for 10 seconds?

Regularly scheduled IPv6 tests should be negotiated on a case-by-case basis.

A repository should be maintained about the regularly scheduled tests including types of tests between which source and destination IP addresses and detailed timestamps. This should be available for all domains to view.

One Way Delay Measurements

LHCONE sites will establish latency measurement points in their domain. The latency measurement points should be placed close to the egress points of the network.

The latency measurement points should be configured to accept measurement requests from all other participating latency measurement points.

The latency measurements will support the OWAMP Control and OWAMP Test protocols as defined in RFC 4656.

The anticipated use of the one-way delay measurements in order of importance are:

- Characterizing loss on a path.
- Characterizing queuing delay on a path
- Identifying asymmetric routing on a path
- Characterizing duplication, reordering and hop-count on a path
- Identifying re-routing events on a path

One Way Delay Measurement Point Clock Issues

One-way latency measurements require accurate stable clocks to produce accurate results. However accuracy required for different uses varies significantly. For example, re-routing within a metropolitan area might introduce 5-10 microsecond changes. Re-routing on trans- oceanic paths may introduce 20-30 millisecond changes. Queuing delays can be anywhere between 10s of microseconds up to 10s of milliseconds. On the other hand, loss rate measurements are most meaningfully characterized on seconds to hour time scales.

For the purpose of this service, we are primarily concerned with characterizing packet loss, and identifying significant queuing artifacts on multi-domain paths. Therefore, the target clock accuracy required is 1 millisecond. (It is understood that achieving better than 1 millisecond accuracy may be impossible or financially unfeasible.)

It is understood that participants may have deployed one-way delay measurement infrastructure with significantly tighter standards. We are not advocating relaxing those standards, but instead want to encourage deployment of one way latency measurement points at domain boundaries where designing for sub 100 microsecond levels of accuracy may not be financially feasible.

There are recommendations about how to configure NTP to obtain sub 1 millisecond accuracy on the OWAMP web site. (Many Unix systems are configured by default to get their time from a pool of servers behind the DNS domain pool.ntp.org. This configuration will not achieve the required precision.)

Eric Boyd 1/30/12 6:05 PM

Comment [6]: Note the use of 4 clock peers, and also note there is no hard requirement for a physical clock (e.g. GPS, CDMA) to be attached to the measurement machine.

On-Demand One Way Delay Measurements

The diagnostic service will support on-demand one-way delay measurements between all of the participating one-way delay measurement points. On demand tests to a single destination should not exceed 10 packets per second without prior negotiation.

Regularly Scheduled One Way Delay Measurements

The diagnostic service will support scheduled one-way delay measurements between all of the participating one-way delay measurement points. Regularly scheduled tests should be configured at 10 packets every second. LHCONE sites should limit the number of regularly scheduled test streams to any particular measurement point in each others domain to less than or equal to 100 packets per second.

Historical Measurement Results

The LHCONE Diagnostic Service will provide access to historical network measurement results via perfSONAR. Historical measurement results should be maintained for at least 12 months.

Each domain should ensure that the results of their regularly scheduled bandwidth and latency tests are published via perfSONAR measurement archives. These perfSONAR measurement archives will be available to the target user community: the NOC Engineers from all participants.

Historical Bandwidth Measurements

The perfSONAR measurement archive containing bandwidth measurement data will support querying for the following information:

- Time the test started
- Duration of the test
- Average bandwidth achieved over the full duration of the test

Historical Latency Measurements

The perfSONAR measurement archives containing latency data will support querying the following information regarding each test interval:

- Number of packets sent
- Packets lost
- Minimum latency
- Median latency
- Maximum latency
    - Note: maximum measured latency in an interval is known to be a very poor indicator of network performance because it is dominated by host artifacts in some domains.

The perfSONAR measurement archives containing latency data may support querying the following information:

- The 25th & 75th percentiles in the interval.
- The perfSONAR measurement archives containing latency data will support querying for statistics on 60 second intervals.

Looking Glass

LHCONE domains should provide web access to a looking glass.

The looking glass should provide the following capabilities:

- Ability to see router interface details & counters including discards, queue drops, etc.
- Ability to see BGP routes and their attributes
- Ability to ping arbitrary destinations
- Ability to traceroute to arbitrary destinations