Editor: Joe Metzger
Status: DRAFT (v1.0)
Date: 2010-12-1

1 General Service Description for DICE Network Diagnostic Services

The DICE collaboration network diagnostic service will simplify the process of debugging, isolating, and correcting multi-domain network performance problems. The diagnostic service will allow users to measure network characteristics across multi-domain network paths.

The service is designed to support network engineers in situations where a customer is experiencing performance problems on a multi-domain network path, and it is not easy to identify the cause of the problem.

The service is offered collaboratively by DICE partners and a set of adjacent domains (NRENs, Regional or other external partners) that adhere to the requirements of the service. These joint networks form a multi-domain area where the service is provided. There is not a universal service obligation for all adjacent domains. Initial target users are network operations engineers in NRENs, Regional Networks and DICE partners.

This service is based on the perfSONAR architecture and will be implemented using perfSONAR tools and services.

2 Service Functionality Description

The diagnostic service will allow network engineers to:

- Discover network measurement points in multiple domains.
- Make new network measurements across multi-domain paths
- Retrieve previous measurements results across a path

2.1 Discovery of Measurement Points

The initial process for discovering network measurement services will be very simple. Each domain will register their measurement points in their portion of a shared perfSONAR information services infrastructure. Each domain will develop their own tools for querying the shared information service and presenting the resulting information to their users.

The types of measurement services that will be registered include OWAMP and BWCTL services.

2.2 Making Measurements

2.2.1 Bandwidth Measurements

Each domain will establish bandwidth measurement points in their network.

The anticipated uses of active bandwidth measurements are:

- Identify paths that can not sustain high bandwidth TCP sessions
- Demonstrate paths can sustain high bandwidth TCP 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.

The DICE members will maintain 10G measurement points near their interconnection points. Other domains should place measurement points close to their connection to their upstream.

The bandwidth measurement points will be configured to accept measurement requests to and from the bandwidth measurement points of all other participants in the collaboration.

The diagnostic service will utilize IPERF for performing achievable bandwidth measurements between measurement points. The IPERF tests will be invoked with BWCTL. BWCTL will handle short duration scheduling to serialize multiple requests and prevent overlapping tests. In some domains, 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 recommended default TCP stack is Reno. The reason for Reno is that it provides a more sensitive measurement than more aggressive TCP stacks such as BIC.

All of the measurement points will accept at least 10 second long UDP requests at rates up to 100 Mbps.

By default, bandwidth tests should not specify a QOS tag, and the traffic should be treated as best effort within a domain.

2.2.1.1 *On-Demand Bandwidth Measurements*

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

2.2.1.2 *Regularly Scheduled Bandwidth Measurements*

The diagnostic service will support regularly scheduled achievable bandwidth measurements between all of the participating bandwidth measurement points.

Inter-domain measurement schedules should be developed with consideration for the bandwidth and utilization of the cross-connects***. The maximum agreed to test schedule is that each domain will not setup scheduled tests that will consume more that 0.1% of the total aggregate capacity to a neighbouring domain on a daily basis.*** The 0.1% number is a not to be exceeded without prior negotiation parameter. There are no current technical controls to prevent domains from exceeding this.

The following example works out the maximum schedules between ESnet and GEANT in Fall 2010.

- There is 25 Gigabits of capacity between the 2 domains.
- 25 Gigabits * 60 seconds * 60 minutes * 24 hours = 2.16 Petabits.
- 2.16 Petabits * 0.001 (one tenth of one percent) = 2.16 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 ESnet to GEANT test points, and tests from ESnet to NREN's in Europe that cross the ESnet/GEANT interconnections.

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 it's 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, and from western Europe to the eastern North 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.

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

Regularly scheduled IPV6 tests should be negotiated on a case-by-case basis.

2.2.2 One Way Delay Measurements

Each domain 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 domains 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

2.2.2.1 *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 the DICE Diagnostic Service, we are primarily concerned with characterizing packet loss, and identifying significant queuing artefacts on multi-domain paths. Therefore, the target clock accuracy required for the initial roll-out of the DICE diagnostic service **is 1 millisecond**.

It is understood that many of the participating domains 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.)

2.2.2.2 *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.

2.2.2.3 *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 per second.

Each domain should limit the number of regularly scheduled test streams to any particular measurement point in another domain to less than or equal to 100 packets per second.

2.3 Historical Measurement Results

The 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 participating domains.

Jason Zurawski 10/18/10 7:36 AM

Comment [1]: May want to indicate how much space 1 years worth of data for each test pair may take up. OWAMP data can get chunky as the number of test partners increase…

2.3.1 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

2.3.2 Historical Latency Measurements

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

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

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.

2.3.3 Looking Glass

Domains participating in the diagnostic service **will** provide public 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

Joe Metzger 12/1/10 4:54 PM
Comment [2]: Chris Welty pointed out that a looking glass service should be an important part of any diagnostic service on December 1st 2010. The DICE perfSONAR group/DICE Product Mangers have not made a decision if this should be a mandatory or optional part of the service yet. So this section may be modified in the next week. Input from the DICE-OPS group is welcome.

2.4 User interface & procedures

Each participating domain will be responsible for developing and deploying user interfaces for their network engineers that meets their needs. These user interfaces will access information from other domains using the perfSONAR protocols.

The participating DICE domains should develop and maintain a set of procedures describing how to use the diagnostic service to simplifying diagnosing inter-domain performance problems. The edupert knowledge base, and the http://fasterdata.es.net websites will be used to maintain this information initially.
Conditions of use – To Be Determined

Unknown
Field Code Changed

3 Service Level Specification

3.1 Service level specification for On-Demand Measurements

This will be developed within the first 6 months of service.

3.2 Service level specification for Scheduled Measurements

This will be developed within the first 6 months of service.

3.3 Service level specification for support

Each DICE partner will offer comprehensive user support to their own community of service users. The support channel may be contacted 24x7. If a support query is wrongly directed to the incorrect partner, they will forward the issue to the correct partner and notify the user of the error and the correct support channel.

The support channel will perform the following functions where they are not already carried out by automated interfaces:

3.3.1 Problem Management

This relates to situations where normal service is not delivered to the user. It may include a failure of web interface or API, or of elements of the deployed infrastructure which prevents it from being usable.

The support channel accepts and “owns” all trouble tickets involving their connected users (including problems at the far end of a connection) until they are resolved.

3.3.2 Change Management

This section will contain pointers from the Operations group that describes how one would negotiate tests that do not comply with the limits described above. IE, a high bandwidth UDP tests. Possibly borrow text Ann develops for Dynamic Circuit Service.

Insert paragraph about blacklisting here.

3.3.3 User support

Participating domains will provide training and documentation and handle queries regarding using provided interfaces to the service during business hours. They will endeavour to answer these queries within 24 hours.

4 Operational Level Agreements

The delivery of the diagnostic service is dependent on infrastructure and services delivered by DICE partners and by connecting RONs and NRENs. In order to deliver consistently to meet the service level specification, service delivery should be ensured by means of operating level agreements (OLA) between the main DICE partners. The following areas shall be regulated via OLA.

4.1 Individually Operated Service Elements

This section sets out the responsibilities of partners towards ensuring that infrastructure and support within their domain is available and in good health towards the service.

4.1.1 Network Infrastructure

Participating domains will not intentionally obscure their internal topology by blocking normal traceroute responses.

4.1.2 Measurement Infrastructure

Participating domains have the following obligations to participate in service delivery

- Operate the measurement infrastructure with a target of 99% monthly availability for measurement archives excluding hardware failures & scheduled maintenance.

- Operate the measurement infrastructure with a target of 99% monthly availability for measurement points excluding hardware failures & scheduled maintenance.

- Measurement services, including all interfaces supplied to users, must be registered with the perfSONAR lookup service.

4.1.3 Monitoring the Measurement Infrastructure

Participating domains will deploy and operate the measurement infrastructure in support of the diagnostic service. Planned and unplanned outages will be announced to a global service operations mailing list (perfsonar-ops@perfsonar.net?) Domains will monitor and raise alarms with their service desk on the following:

- Availability of the BWCTL service to perform measurements on the bandwidth measurement points.
- Availability of the OWAMP service to perform measurements on the latency measurement points.
- Availability of the measurement archive service to return results on the measurement archive nodes.
- Presence of the domains measurement service access points in the global information services.

4.1.4 AA Infrastructure

The perfSONAR project does not currently support sufficiently advanced AA capabilities in all of the required tools and implementations necessary to meet the diagnostic service requirements. Therefore, a phased approach to deploying AA, and services that require AA will be used.

Phase 1 of the service will be deployed without AA where possible, and with IP address based ACLS where necessary.

It is understood that this plan has scaling problems. But it will meet the critical immediate need, and we expect it to suffice until the requisite perfSONAR tools are able to use an AA infrastructure that includes the current target user community (NOC engineers at R&E networks.)

Future phases of the service will involve deploying AA on some services in some domains, while other domains while still maintaining interoperability, and diagnostic capabilities to domains and/or services without AA. A roadmap will be developed that deals with per domain, per service migration as robust interoperable AA solutions become available.

4.1.4.1 *AA for Measurement Archives*

Measurement Archives in a participating domain must be accessible to visualization and analysis tools from all the other participating domains. This will be achieved by implementing one of the following in order of preference.

1. Participants are encouraged to make their measurement archives publically accessible.
2. Participants may restrict access to their measurement archives to the subset of Internet prefixes serviced by Research and Education Networks including NRENs and Gigapops. Participants may construct this list from their own routing infrastructure, or DICE members may publish a list for the community. Domains using this policy should update their access lists at least weekly.

4.1.4.2 *AA for Measurement Points*

Measurement Points in a participating domain must be accessible to the measurement points in all the other participating domains. They should also be accessible to user tools in other domains. This will be achieved by implementing one of the following in order of preference.

1. Participants are encouraged to make their measurement points publically accessible.
2. Participants may restrict access to their measurement points to the subset of Internet prefixes in the global research and education networks routing table including NRENs Gigapops & Regionals. Domains using this policy should update their access lists at least weekly.
3. Participants may futher restrict access to their measurement points if necessary to the set of measurement points registered in the global information service. Domains using this policy should update their access lists at least weekly.

4.1.5 Security Support

Each DICE partner will operate a team that can handle incident response in relation to security incidents affecting the service.

This team will liase with other security teams in the affected area and will provide feedback to developers and operators to implement mitigating and preventative measures.

4.1.6 Service Desk

Service Desk contact information and an escalation path must be published to DICE partners.

Participating service desks are responsible for receiving and acting on alarms raised relevant to the service to deliver the target SLS. This will occur via the perfsonar-ops@perfsonar.net mailing list.

All participating Service Desks are responsible for sending notices to neighbouring Service Desks participating in service delivery about scheduled events and network or other dependent infrastructure outages. The Neighbouring Service Desks can propagate this information to its neighbours depending on impact of the events and outages on the service.

All participating Service Desks are responsible for answering queries and investigating and resolving problems about the service reported by service desks operated by other DICE partners also delivering the service. This may include problems taking a measurement or retrieving results.

Ann Harding 10/18/10 7:27 AM
Comment [3]: What metrics do we want to put on this?

4.2 Jointly Operated Service Elements

The Internet2, ESnet & GEANT will each support a gLS which is the shared root of the lookup or information service. The gLS will be treated as production infrastructure, and managed according to the standards each domain has for maintaining production infrastructure.

Internet2 will host the perfsonar-ops@perfsonar.net mailing list, as well as other maintenance and announce mailing lists which may be required to support this service. Each DICE domain will have at least 1 designated person who will maintain the mailing list membership for their respective areas.

Unknown
Field Code Changed

The DICE partners will contribute to the maintenance and upkeep of the www.perfsonar.net web site

Unknown
Field Code Changed

5 Bibliography

1. DICE
2. OWAMP http://www.internet2.edu/performance/owamp/
3. BWCTL http://www.internet2.edu/performance/bwctl
4. HADES http://www.hades.cc
5. perfSONAR http://www.perfsonar.net