Network Working Group
Internet Draft

Vishwas Manral
Netplane Systems
Russ White
Cisco Systems
Aman Shaikh
University of California

Expiration Date: March 2003
File Name: draft-ietf-bmwg-ospfconv-term-01.txt

October 2002

OSPF Benchmarking Terminology and Concepts
draft-ietf-bmwg-ospfconv-term-01.txt

1. Status of this Memo

This document is an Internet-Draft and is in full conformance with all provisions of Section 10 of RFC2026.

Internet Drafts are working documents of the Internet Engineering Task Force (IETF), its Areas, and its Working Groups. Note that other groups may also distribute working documents as Internet Drafts.

Internet Drafts are draft documents valid for a maximum of six months. Internet Drafts may be updated, replaced, or obsoleted by other documents at any time. It is not appropriate to use Internet Drafts as reference material or to cite them other than as a "working draft" or "work in progress".

The list of current Internet-Drafts can be accessed at http//www.ietf.org/ietf/1id-abstracts.txt

The list of Internet-Draft Shadow Directories can be accessed at http//www.ietf.org/shadow.html.

2. Abstract

This draft explains the terminology and concepts used in [2] and future OSPF benchmarking drafts.

3. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [1].

4. Motivation

This draft is a companion to [2], which describes basic Open Shortest Path First (OSPF [3]) testing methods. This draft explains terminology and concepts used in OSPF Testing Framework Drafts, such as [2].

5. Definitions

o Internal Measurements

- Definition

  Internal measurements are measurements taken on the Device Under Test (DUT) itself.

- Discussion

  These measurement rely on output and event recording, along with the clocking and timestamping available on the DUT itself. Taking measurements on the DUT may impact the actual outcome of the test, since it can increase processor loading, memory utilization, and timing factors. Some devices may not have the required output readily available for taking internal measurements, as well.

  Note: Internal measurements can be influenced by the vendor's implementation of the various timers and processing models. Whenever possible, internal measurements should be compared to external measurements to verify and validate them.

o External Measurements

- Definition

External measurements infer the performance of the DUT through observation of its communications with other devices.

- Discussion

One example of an external measurement is when a downstream device receives complete routing information from the DUT, it can be inferred that the DUT has transmitted all the routing information available. External measurements suffer in that they include not just the protocol action times, but also propagation delays, queuing delays, and other such factors.

For the purposes of this paper, external techniques are more readily applicable.

o Multi-device Measurements

- Definition

Multi-device measurements require the measurement of events occurring on multiple devices within the testbed.

- Discussion

For instance, the timestamp on a device generating an event could be used as the marker for the beginning of a test, while the timestamp on the DUT or some other device might be used to determine when the DUT has finished processing the event.

These sorts of measurements are the most problematic, and are to be avoided where possible, since the timestamps of the devices in the test bed must be synchronized within milliseconds for the test results to be meaningful. Given the state of network time protocol implementation, expecting the timestamps on several devices to be within milliseconds of each other is highly optimistic.

o Point-to-Point links

- Definition

A network that joins a single pair of routers is called a point-to-point link. For OSPF [3], point-to-point links are those on which a designated router are not elected.

- Discussion

A point-to-point link will take lesser time to converge than a broadcast link of the same speed because it does not have the overhead of DR election. Point-to-point links can be either numbered or unnumbered. However in the context of [2], the two can be regarded the same.

o Broadcast Link

- Definition

Networks supporting many (more than two) attached routers, together with the capability to address a single physical message to all of the attached routers (broadcast). In the context of [2] and [3], broadcast links are taken as those on which a designated router is elected.

- Discussion

The adjacency formation time on a broadcast link can be more than that on a point-to-point link of the same speed, because DR election has to take place. All routers on a broadcast network form adjacency with the DR and BDR.

Async flooding also takes place thru the DR. In context of convergence, it may take more time for an LSU to be flooded from one DR-other router to another DR-other router, because the LSA has to be first processed at the DR.

o Shortest Path First Time

- Definition

The time taken by a router to complete the SPF process.

- Discussion

This does not include the time taken by the router to give routes to the forwarding engine.

o Measurement Units

The SPF time is generally measured in milliseconds.

o Hello Interval

- Definition

The length of time, in seconds, between the Hello Packets that the router sends hello packets on the interface. The typical hello interval is 10 seconds on broadcast networks, and 30 seconds for point-to-multipoint and point-to-point networks. On multicast capable media, hellos are sent to a multicast address; on non-multicast capable

media, they are sent unicast to each neighbor.

- Discussion

  The hello interval should be the same for all routers on the network

  Decreasing the hello interval can allow the router dead interval (below) to be reduced, thus reducing convergence times in those situations where the router dead interval timing out causes an OSPF process to notice an adjacency failure. Very small router dead intervals accompanied by very small hello intervals can produce more problems than they resolve, as described in [4] & [5].

o Router Dead interval

- Definition

  After ceasing to hear a router's Hello Packets, the number of seconds before its neighbors declare the router down. The default dead interval is four times the hello interval; 40 seconds on broadcast networks, and 120 seconds on

non-broadcast networks.

- Discussion

  This is advertised in the router's Hello Packets in the RouterDeadInterval field. The router dead interval should be some multiple of the HelloInterval (say 4 times the hello interval), and must be the same for all routers attached to a common network.

o Incremental SPF

- Definition

  The ability to recalculate a small portion of the SPF tree, rather than the entire SPF tree, on receiving notification of a change in the network topology. At worst, incremental SPF should perform no worse than a full SPF. In better situations, an incremental SPF run will rebuild the SPF tree in much shorter time than a full SPF run.

6. Concepts

6.1. The Meaning of Convergence

A network is termed to be converged when all of the devices within the network have a loop free path to each possible destination. Since we are not testing network convergence, but performance for a particular device within a network, however, this definition needs to be narrowed somewhat to fit within a single device view.

In this case, convergence will mean the point in time when the DUT has performed all actions needed to react to the change in topology represented by the test condition; for instance, an OSPF device must flood any new information it has received, rebuild its shortest path first (SPF) tree, and install any new paths or destinations in the local routing information base (RIB, or routing table).

Note that the word convergence has two distinct meanings; the process of a group of individuals meeting the same place, and the process of a single individual meeting in the same place as an existing group. This work focuses on the second meaning of the word, so we consider the time required for a single device to adapt to a network change to

be SR-Convergence, or Single Router Convergence.

6.2. Measuring Convergence

Obviously, there are several elements to convergence, even under the definition given above for a single device. We will try to provide tests to measure each of these:

- The time it takes for the DUT to pass the information about a network event on to its neighbors.

- The time it takes for the DUT to process information about a network event and calculate a new Shortest Path Tree (SPT).

- The time it takes for the DUT to make changes in its local rib reflecting the new shortest path tree.

6.3. Types of Network Events

A network event is an event which causes a change in the network topology.

- Link or Neighbor Device Up

  The time needed for an OSPF implementation to recoginize a new link coming up on the device, build any necessarily adjacencies, synchronize its database, and perform all other needed actions to converge.

- Initialization

  The time needed for an OSPF implementation to be initialized, recognize any links across which OSPF must run, build any needed adjacencies, synchronize its database, and perform other actions needed to converge.

- Adjacency Down

  The time needed for an OSPF implementation to recognize a link down/adjacency loss based on hello timers alone,

propogate any information as necessary to its remaining adja-

cencies, and perform other actions needed to converge.

- Link Down

  The time needed for an OSPF implementation to recognize a link down based on layer 2 provided information, propogate any information as needed to its remaining adjacencies, and perform other actions needed to converge.

6.4. LSA and Destination mix

In many OSPF benchmark tests, a generator injecting a number of LSAs is called for. There are several areas in which injected LSAs can be varied in testing:

- The number of destinations represented by the injected LSAs

  Each destination represents a single reachable IP network; these will be leaf nodes on the shortest path tree. The primary impact to performance should be the time required to insert destinations in the local routing table and handling the memory required to store the data.

- The types of LSAs injected

  There are several types of LSAs which would be acceptable under different situations; within an area, for instance, type 1, 2, 3, 4, and 5 are likely to be received by a router. Within a not-so-stubby area, however, type 7 LSAs would replace the type 5 LSAs received. These sorts of characterizations are important to note in any test results.

- The Number of LSAs injected

  Within any injected set of information, the number of each type of LSA injected is also important. This will impact the shortest path algorithms ability to handle large numbers of nodes, large shortest path first trees, etc.

- The Order of LSA Injection

The order in which LSAs are injected should not favor any given data structure used for storing the LSA database on the device under test. For instance, AS-External LSA's have AS wide flooding scope; any Type-5 LSA originated is immediately flooded to all neighbors. However the Type-4 LSA which announces the ASBR as a border router is originated in an area at SPF time (by ABR's on the edge of the area in which the ASBR is). If SPF isn't scheduled immediately on the ABRs originating the type 4 LSA, the type-4 LSA is sent after the type-5 LSA's reach a router in the adjacent area. So routes to the external destinations aren't immediately added to the routers in the other areas. When the routers which already have the type 5's receive the type-4 LSA, all the external routes are added to the tree at the same time. This timing could produce different results than a router receiving a type 4 indicating the presence of a border router, followed by the type 5's originated by that border router.

The ordering can be changed in various tests to provide insight on the efficiency of storage within the DUT. Any such changes in ordering should be noted in test results.

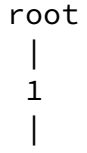
6.5. Tree Shape and the SPF Algorithm

The complexity of Dijkstra's algo depends on the data structure used for storing vertices with their current minimum distances from the source. The simplest structure is a list of vertices currently reachable from the source. Finding the minimum cost vertex then would take O(size of the list). There will be O(n) such operations if we assume that all the vertices are ultimately reachable from the source. Moreover, after the vertex with min cost is found, the algo iterates thru all the edges of the vertex and updates cost of other vertices. With an adjacency list representation, this step when iterated over all the vertices, would take O(E) time. Thus, overall running time is:

O(sum(i:1, n)(size(list at level i) + E).

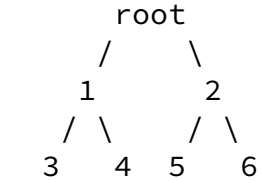
So, everything boils down to the size(list at level i).

If the graph is linear:



and source is a vertex on the end, then size(list at level i) = 1 for all i. Moreover, E = n - 1. Therefore, running time is O(n).

If graph is a balanced binary tree:



size(list at level i) is a little complicated. First it increases by 1 at each level upto a certain number, and then goes down by 1. If we assumed that tree is a complete tree (like the one in the draft) with k levels (1 to k), then size(list) goes on like this: 1, 2, 3,

Then the number of edges E is still n - 1. It then turns out that the run-time is O(n^2) for such a tree.

If graph is a complete graph (fully-connected mesh), then size(list at level i) = n - i. Number of edges E = O(n^2). Therefore, run-time is O(n^2).

shortest path first algorithm to compute the best paths through the network need to be aware that the construction of the tree may impact the performance of the algorithm. Best practice would be to try and make any emulated network look as much like a real network as possible, especially in the area of the tree depth, the meshiness of the network, the

number of stub links verses transit links, and the number of connections and nodes to process at each level within the original tree.

7. Topology Generation

As the size of networks grows, it becomes more and more difficult to actually create a large scale network on which to test the properties of routing protocols and their implementations. In general, network emulators are used to provide emulated topologies which can be advertised to a device with varying conditions. Route generators either tend to be a specialized device, a piece of software which runs on a router, or a process that runs on another operating system, such as Linux or another variant of Unix.

Some of the characteristics of this device should be:

- The ability to connect to the several devices using both point-to-point and broadcast high speed media. Point-to-point links can be emulated with high speed Ethernet as long as there is no hub or other device in between the DUT and the route generator, and the link is configured as a point-to-point link within OSPF.

- The ability to create a set of LSAs which appear to be a logical, realistic topology. For instance, the generator should be able to mix the number of point-to-point and broadcast links within the emulated topology, and should be able to inject varying numbers of externally reachable destinations.

- The ability to withdraw and add routing information into and from the emulated topology to emulate links flapping.

- The ability to randomly order the LSAs representing the emulated topology as they are advertised.

- The ability to log or otherwise measure the time between packets transmitted and received.

- The ability to change the rate at which OSPF LSAs are transmitted.

- The generator and the collector should be fast enough so that they are not bottle necks. The devices should also have a degree of granularity of measurement atleast as small as desired from the test results.

8. Acknowedgements

The authors would like to thank Howard Berkowitz (hcb@clark.net), Kevin Dubray, (kdubray@juniper.net), and Randy Bush (randy@psg.com) for their discussion, ideas, and support.

9. References

[1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC2119, March 1997.

[2] Manral, V., "Benchmarking Methodology for Basic OSPF Convergence", draft-bmwg-ospfconv-intraarea-00, May 2002

[3] Moy, J., "OSPF Version 2", RFC 2328, April 1998.

[4] Ash, J., "Proposed Mechanisms for Congestion Control/Failure Recovery in OSPF & ISIS Networks", October, 2001

[5] draft-ietf-ospf-scalability-00.txt Choudhury, G., et al, "Explicit

Marking and Prioritized Treatment of Specific IGP Packets for Faster IGP Convergence and Improved Network Scalability and Stability", April 2002

10. Authors' Addresses

Vishwas Manral,
Netplane Systems,
189 Prashasan Nagar,
Road number 72,
Jubilee Hills,
Hyderabad.

vmanral@netplane.com

Russ White
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709

riw@cisco.com

Aman Shaikh
University of California
School of Engineering
1156 High Street
Santa Cruz, CA 95064

aman@soe.ucsc.edu