Network Working Group   Vishwas Manral
Internet Draft   Netplane Systems
                                                                   Russ White
                                                                Cisco Systems
                                                                  Aman Shaikh
Expiration Date: Novemeber 2004   University of California
File Name: draft-bmwg-ospfconv-intraarea-08.txt   May 2004

Benchmarking Basic OSPF Single Router Control Plane Convergence
draft-ietf-bmwg-ospfconv-intraarea-08.txt

1. Status of this Memo

Abstract

This draft provides suggestions for measuring OSPF single router control plane convergence. Its initial emphasis is on the control plane of single OSPF routers. We do not address forwarding plane performance.

NOTE: Within this document, the word convergence relates to single router control plane convergence only.

2. Introduction

There is a growing interest in routing protocol convergence testing, with many people looking at various tests to determine how long it takes for a network to converge after various conditions occur. The major problem with this sort of testing is that the framework of the tests has a major impact on the results; for instance, determining when a network is converged, what parts of the router's operation are considered within the testing, and other such things will have a major impact on what apparent performance routing protocols provide.

This document attempts to provide a framework within which Open Shortest Path First [OSPF] performance testing can be placed, and provide some tests with which some aspects of OSPF performance can be measured. The motivation of the draft is to provide a set of tests that can provide the user comparable data from various vendors with which to evaluate the OSPF protocol performance on the devices.

3. Specification of Requirements

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 [RFC2119].

4. Overview & Scope

While this document describes a specific set of tests aimed at characterizing the single router control plane convergence performance of OSPF processes in routers or other boxes that incorporate OSPF functionality, a key objective is to propose methodologies that will prdouce directly comparable convergence related measurements.

Things which are outside the scope of this document include:

- The interactions of convergence and forwarding; testing is restricted to events occurring within the control plane. Forwarding performance is the primary focus in [INTERCONNECT] and it is expected to be dealt with in work that ensues from [FIB-TERM].

- Inter-area route generation, AS-external route generation, and simultaneous traffic on the control and data paths within the DUT. While the tests outlined in this document measure SPF time, flooding times, and other aspects of all OSPF convergence performance, it does not provide tests for measuring external or

summary route generation, route translation, or other OSPF inter-area and external routing performance. These are expected to be dealt with in a later draft.

Other drafts in the future may cover some of the items noted as not covered in the scope of this draft. For a discussion of the terminology used in this draft (in relation to the tests themselves), refer to [TERM]. For a discussion of the applicability of this draft, refer to [APPLICABILITY].

While this draft assumes OSPFv2, which only carries routing information for IPv4 destinations, the tests described in this document apply to OSPFv3, which carries IPv6 destinations.

5. Test Conditions

In all tests, the following test conditions will be assumed:

- o The link speed should be high enough so that does not become a bottleneck. Link speeds of 10MBps or higher are recommended. The link speed between routers should be specified in the test report.

- o For all point-to-point links, it is assumed that a link failure results in an immediate notification to the operating system, and thus to the OSPF process; this is explained thoroughly in [MILLISEC].

- o No data traffic will be running between the routers during these tests.

- o Optional capabilities which can reduce performance, such as authentication, should be noted in the test results if they are enabled.

- o Optional changes in the default timer values, such as the SPF, hello, router dead, and other intervals, should be noted in the test results.

- o All places where injecting a set of LSAs is referenced, the set can include varying numbers of LSAs of varying types representing a varying number of reachable destinations. See [TERM] for further information about issues with LSA sets and network topologies.

  Tests should be run more than once, since a single test run

cannot be relied on to produce statistically sound results. The number of test runs and any variations between the tests should be recorded in the test results (see [TERM] for more information on what items should be recorded in the test results).

6. Reference Topologies

Several reference topologies will be used throughout the tests described in the remainder of this document. Rather than repeating these topologies, we've gathered them all in one section.

o   Reference Topology 1 (Emulated Topology)

```
                     (                   )
DUT----Generator----(  emulated topology  )
                     (                   )
```

A simple back-to-back configuration. It's assumed that the link between the generator and the DUT is a point-to-point link, while the connections within the generator represent some emulated topology.

o   Reference Topology 2 (Generator and Collector)

```
                                  (                   )
Collector-----DUT-----Generator--(  emulated topology  )
       \              /          (                   )
        \------------/
```

All routers are connected through point-to-point links. The cost of all links is assumed to be the same unless otherwise noted.

o   Reference Topology 3 (Broadcast Network)

```
DUT     R1     R2
 |      |      |
-+------+------+-----.....
```

Any number of routers could be included on the common broadcast network.

o   Reference Topology 4 (Parallel Links)

```
  /--(link 1)-----\           (                   )
DUT               Generator--(  emulated topology  )
  \--(link 2)-----/           (                   )
```

In all cases the tests and topologies are designed to allow performance measurements to be taken all on a single device, whether the DUT or some other device in the network. This eliminates the need for syncronized clocks within the test networks.

7. Basic Process Performance Tests

These tests will measure aspects of the OSPF implementation as a process on the device under test, including:

- Time required to process an LSA
- Flooding time
- Shortest Path First computation

7.1. Time required to process an LSA

- Using reference topology 1 (Emulated Topology), begin with all links up and a full adjacency established between the DUT and the generator.

  Note: The generator does not have direct knowledge of the state of the adjacency on the DUT. The fact the adjacency may be in Full on the generator does not mean that the DUT is ready. It may still (and is likely to) be requesting LSAs from the generator. This process, involving processing of requested LSAs, will affect the results of the test. The generator should either wait until it sees the DUT's router-LSA listing the adjacency with the generator or introduce a configurable delay before starting the test.

- Send an LSA that is already there in the DUT (a duplicate LSA), note the time difference between when the LSA is sent to when the ack is received. This measures the time to propagate the LSA and the ack, as well as processing time of the duplicate LSA. This is dupLSAprocTime.

- Send a new LSA from the generator to the DUT, followed immediately by a duplicate LSA (LSA that already resides in the database of DUT, but not the same as the one just sent).

- The DUT will acknowledge this second LSA immediately; note the

time of this acknowledgement. This is newLSAprocTime.

The amount of time required for an OSPF implementation to process the new LSA can be computed by subtracting dupLSAprocTime from newLSAprocTime.

Note: The duplicate LSA cannot be the same as the one just sent because of the MinLSInterval restriction.[RFC2328] This test is taken from [BLACKBOX].

7.2. Flooding Time

- o Using reference topology 2 (Generator and Collector), enable OSPF on all links and allow the devices to build full adjacencies. Configure the collector so it will block all flooding towards the DUT, although it continues receiving advertisements from the DUT.

- o Inject a new set of LSAs from the generator towards the collector and the DUT.

- o On the collector, note the time the flooding is complete across the link to the generator. Also note the time the flooding is complete across the link from the DUT.

The time between the last LSA is received on the collector from the generator and the time the last LSA is received on the collector from the DUT should be measured during this test. This time is important in link state protocols, since the loop free nature of the network is reliant on the speed at which revised topology information is flooded.

Depending on the number of LSAs flooded, the sizes of the LSAs, the number of LSUs, and the rate of flooding, these numbers could vary by some amount. The settings and variances of these numbers should be reported with the test results.

7.3. Shortest Path First Computation Time

- o Use reference topology 1 (Emulated Toplogy), beginning with the DUT and the generator fully adjacent.

- o The default SPF timer on the DUT should be set to 0, so that any new LSA that arrives, immediately results in the SPF calculation

[BLACKBOX].

- The generator should inject a set of LSAs towards the DUT; the DUT should be allowed to converge and install all best paths in the local routing table, etc..

- Send an LSA that is already there in the DUT (a duplicate LSA), note the time difference between when the LSA is sent to when the ack is received. This measures the time to propagate the LSA and the ack, as well as processing time of the duplicate LSA. This is dupLSAprocTime.

- Change the link cost between the generator and the emulated network it is advertising, and transmit the new LSA to the DUT.

- Immediately inject another LSA which is a duplicate of some other LSA the generator has previously injected (preferrably a stub network someplace within the emulated network).

  Note: The generator should make sure that outbound LSA packing is not performed for the duplicate LSAs and they are always sent in a separate Link-state Update packet. Otherwise, if the LSA carrying the topo change and the duplicate LSA are in the same packet, the SPF will be started the duplicate LSA is acked.

- Measure the time between transmitting the second (duplicate) LSA and the acknowledgement for that LSA; this is the totalSPFtime. The total time required to run SPF can be computed by subtracting dupLSAprocTime from totalSPFtime.

The accuracy of this test is crucially dependant on the amount of time between the transmission of the first and second LSAs. If there is too much time between them, the test is meaningless because the SPF run will complete before the second (duplicate) LSA is received. If there is too little time between the LSAs being generated, then they will both be handled before the SPF run is scheduled and started, and thus the measurement would only be for the handling of the duplicate LSA.

This test is also specified in [BLACKBOX].

Note: This test may not be accurate on systems which implement OSPF as a multithreaded process, where the flooding takes place in a separate process (or on a different processor) than shortest path first computations.

It is also possible to measure the SPF time using white box tests

(using output supplied by the OSPF software impelemtor). For instance:

- o Using reference topology 1 (Emulated Topology), establish a full adjacency between the generator and the DUT.

- o Inject a set of LSAs from the generator towards the DUT. Allow the DUT to stabilize and install all best paths in the routing table, etc.

- o Change the link cost between the DUT and the generator (or the link between the generator and the emulated network it is advertising), such that a full SPF is required to run, although only one piece of information is changed.

- o Measure the amount of time required for the DUT to compute new shortest path tree as a result of the topology changes injected by the generator. These measurements should be taken using available show and debug information on the DUT.

Several caveats MUST be mentioned when using a white box method of measuring SPF time; for instance, such white box tests are only applicable when testing various versions or variations within a single implementation of the OSPF protocol. Futher, the same set of commands MUST be used in each iteration of such a test, to ensure consistent results.

There is some interesting relationship between the SPF times reported by white box (internal) testing, and black box (external) testing; these two types of tests may be used as a "sanity check" on the other type of tests, by comparing the results of the two tests.

See [APPLICABILITY] for further discussion.

8. Basic Intra-Area OSPF tests

These tests measure the performance of an OSPF implementation for basic intra-area tasks, including:

- o Forming Adjacencies on Point-to-Point Link (Initialization)

- o Forming Adjacencies on Point-to-Point Links

- o Link Up with Information Already in the Database

- o Initial convergence Time on a Designated Router Electing (Broadcast) Network

- o Link Down with Layer 2 Detection

- o Link Down with Layer 3 Detection

- o Designated Router Election Time on A Broadcast Network

8.1. Forming Adjacencies on Point-to-Point Link (Initialization)

This test measures the time required to form an OSPF adjacency from the time a layer two (data link) connection is formed between two devices running OSPF.

- o Use reference topology 1 (Emulated Topology), beginning with the link between the generator and DUT disabled on the DUT. OSPF should be configured and operating on both devices.

- o Inject a set of LSAs from the generator towards the DUT.

- o Bring the link up at the DUT, noting the time that the link carrier is established on the generator.

- o Note the time the acknowledgement for the last LSA transmitted from the DUT is received on the generator.

The time between the carrier establishment and the acknowledgement for the last LSA transmitted by the generator should be taken as the total amount of time required for the OSPF process on the DUT to react to a link up event with the set of LSAs injected, including the time required for the operating system to notify the OSPF process about the link up, etc.. The acknowledgement for the last LSA transmitted is used instead of the last acknowledgement received in order to prevent timing skews due to retransmitted acknowledgements or LSAs.

8.2. Forming Adjacencies on Point-to-Point Links

This test measures the time required to form an adjacency from the time the first communication occurs between two devices running OSPF.

- o Using reference topology 1 (Emulated Topology), configure the DUT and the generator so traffic can be passed along the link

between them.

- o Configure the generator so OSPF is running on the point-to-point link towards the DUT, and inject a set of LSAs.

- o Configure the DUT so OSPF is initialized, but not running on the point-to-point link between the DUT and the generator.

- o Enable OSPF on the interface between the DUT and the generator on the DUT.

- o Note the time of the first hello received from the DUT on the generator.

- o Note the time of the acknowledgement from the DUT for the last LSA transmitted on the generator.

The time between the first hello received and the acknowledgement for the last LSA transmitted by the generator should be taken as the total amount of time required for the OSPF process on the DUT to build a FULL neighbor adjacency with the set of LSAs injected. The acknowledgement for the last LSA transmitted is used instead of the last acknowledgement received in order to prevent timing skews due to retransmitted acknowledgements or LSAs.

8.3. Forming adjacencies with Information Already in the Database

- o Using reference topology 2 (Generator and Collector), configure all three devices to run OSPF.

- o Configure the DUT so the link between the DUT and the generator is disabled .

- o Inject a set of LSAs into the network from the generator; the DUT should receive these LSAs through normal flooding from the collector.

- o Enable the link between the DUT and the generator.

- o Note the time of the first hello received from the DUT on the generator.

- o Note the time of the last DBD received on the generator.

- o Note the time of the acknowledgement from the DUT for the last LSA transmitted on the generator.

The time between the hello received from the DUT by the generator and the acknowledgement for the last LSA transmitted by the generator should be taken as the total amount of time required for the OSPF process on the DUT to build a FULL neighbor adjacency with the set of LSAs injected. In this test, the DUT is already aware of the entire network topology, so the time required should only include the processing of DBDs exchanged when in EXCHANGE state, the time to build a new router LSA containing the new connection information, and the time required to flood and acknowledge this new router LSA.

The acknowledgement for the last LSA transmitted is used instead of the last acknowledgement received in order to prevent timing skews due to retransmitted acknowledgements or LSAs.

8.4. Designated Router Election Time on A Broadcast Network

- o Using reference topology 3 (Broadcast Network), configure R1 to be the designated router on the link, and the DUT to be the backup designated router.

- o Enable OSPF on the common broadcast link on all the routers in the test bed.

- o Disble the broadcast link on R1.

- o Note the time of the last hello received from R1 on R2.

- o Note the time of the first network LSA generated by the DUT as received on R2.

The time between the last hello received on R2 and the first network LSA generated by the DUT should be taken as the amount of time required for the DUT to complete a designated router election computation. Note this test includes the dead interval timer at the DUT, so this time may be factored out, or the hello and dead intervals reduced to make these timers impact the overall test times less. All changed timers, the number of routers connected to the link, and other variable factors should be noted in the test results.

Note: If R1 sends a "goodbye hello," typically a hello with its neighbor list empty, in the process of shutting down its interface, using the time this hello is received instead of the time of the last hello received would provide a more accurate measurement.

8.5. Initial Convergence Time on a Broadcast Network, Test 1

- o Using reference topology 3 (Broadcast Network), begin with the DUT connected to the network with OSPF enabled. OSPF should be enabled on R1, but the broadcast link should be disabled.
- o Enable the broadcast link between R1 and the DUT. Note the time of the first hello received by R1.
- o Note the time the first network LSA is flooded by the DUT at R1.
- o The differential between the first hello and the first network LSA is the time required by the DUT to converge on this new topology.

This test assumes that the DUT will be the designated router on the broadcast link. A similar test could be designed to test the convergence time when the DUT is not the designated router as well.

This test may be performed with varying numbers of devices attached to the broadcast network, and varying sets of LSAs being advertised to the DUT from the routers attached to the broadcast network. Variations in the LSA sets and other factors should be noted in the test results.

The time required to elect a designated router, as measured in Designated Router Election Time on A Broadcast Network, above, may be subtracted from the results of this test to provide just the convergence time across a broadcast network.

Note all the other tests in the document include route calculation time in the conergence time, as described in [TERM], this test may not include route calculation time in the resulting measured convergence time, because initial route calculation may occur after the first network LSA is flooded.

8.6. Initial Convergence Time on a Broadcast Network, Test 2

- o Using reference topology 3 (Broadcast Network), begin with the DUT connected to the network with OSPF enabled. OSPF should be enabled on R1, but the broadcast link should be disabled.
- o Enable the broadcast link between R1 and the DUT. Note the time of the first hello transmitted by the DUT with a designated router listed.

- o Note the time the first network LSA is flooded by the DUT at R1.

- o The differential between the first hello with a designated router lists and the first network LSA is the time required by the DUT to converge on this new topology.

8.7. Link Down with Layer 2 Detection

- o Using reference topology 4 (Parallel Links), begin with OSPF in the full state between the generator and the DUT. Both links should be point-to-point links with the ability to notify the operating system immediately upon link failure.

- o Disable link 1; this should be done in such a way that the keepalive timers at the data link layer will have no impact on the DUT recognizing the link failure (the operating system in the DUT should recognize this link failure immediately). Disconnecting the cable on the generator end would be one possibility, or shutting the link down.

- o Note the time of the link failure on the generator.

- o At the generator, note the time of the receipt of the new router LSA from the DUT notifying the generator of the link 2 failure.

  The difference in the time between the initial link failure and the receipt of the LSA on the generator across link 2 should be taken as the time required for an OSPF implementation to recognize and process a link failure, including the time required to generate and flood an LSA describing the link down event to an adjacent neighbor.

8.8. Link Down with Layer 3 Detection

- o Using reference topology 4 (Parallel Links), begin with OSPF in the full state between the generator and the DUT.

- o Disable OSPF processing on link 1 from the generator. This should be done in such a way so it does not affect link status; the DUT MUST note the failure of the adjacency through the dead interval.

- o At the generator, note the time of the receipt of the new router LSA from the DUT notifying the generator of the link 2 failure.

The difference in the time between the initial link failure and the receipt of the LSA on the generator across link 2 should be taken as the time required for an OSPF implementation to recognize and process an adjacency failure.

9. Security Considerations

This doecument does not modify the underlying security considerations in [OSPF].

10. Acknowledgements

Thanks to Howard Berkowitz, (hcb@clark.net), for his encouragement and support. Thanks also to Alex Zinin (zinin@psg.net), Gurpreet Singh (Gurpreet.Singh@SpirentCom.COM), and Yasuhiro Ohara (yasu@sfc.wide.ad.jp) for their comments as well.

11. Normative References

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

[TERM]Manral, V., "OSPF Convergence Testing Terminiology and Concepts", draft-ietf-bmwg-ospfconv-term-08, May 2004

[APPLICABILITY]
Manral, V., "Benchmarking Applicability for Basic OSPF Convergence", draft-ietf-bmwg-ospfconv-applicability-05, May

[RFC2119]
Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

12. Informative References

[INTERCONNECT]
Bradner, S., McQuaid, J., "Benchmarking Methodology for Network Interconnect Devices", RFC2544, March 1999.

[MILLISEC]
Alaettinoglu C., et al., "Towards Milli-Second IGP Convergence" draft-alaettinoglu-isis-convergence

[FIB-TERM]
Trotter, G., "Terminology for Forwarding Information Base (FIB) based Router Performance", RFC3222, October 2001.

[BLACKBOX]
Shaikh, Aman, Greenberg, Albert, "Experience in Black-Box OSPF measurement"

13. Authors' Addresses

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

vmanral@netplane.com

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

riw@cisco.com

Aman Shaikh
AT&T Labs (Research)
180, Park Av
Florham Park, NJ 07932

ashaikh@research.att.com

Network Working Group   Vishwas Manral
Internet Draft   Netplane Systems
                                                              Russ White
                                                           Cisco Systems
                                                             Aman Shaikh
Expiration Date: November 2004   University of California
File Name: draft-ietf-bmwg-ospfconv-applicability-05.txt   May 2004

Benchmarking Applicability for Basic OSPF Convergence

draft-ietf-bmwg-ospfconv-applicability-05.txt

Status of this Memo

Copyright Notice

Abstract

This document discusses the applicability of various tests for measuring single router control plane convergence, specifically in regards to the Open Shortest First (OSPF) protocol. There are two general sections in this document, the first discussing specific advantages and limitations of specific OSPF convergence tests, and the second discussing more general pitfalls to be considered when testing routing protocols convergence testing.

1. Introduction

There is a growing interest in testing single router control plane convergence for routing protocols, with many people looking at testing methodologies which can provide information on how long it takes for a network to converge after various network events occur. It is important to consider the framework within which any given convergence test is executed when attempting to apply the results of the testing, since the framework can have a major impact on the results. For instance, determining when a network is converged, what parts of the router's operation are considered within the testing, and other such things will have a major impact on what apparent performance routing protocols provide.

This document describes in detail the various benefits and pitfalls of tests described in [BENCHMARK]. It also explains how such measurements can be useful for providers and the research community.

NOTE: The word convergence within this document refers to single router control plane convergence [TERM].

2. Advantages of Such Measurement

o To be able to compare the iterations of a protocol implementation. It is often useful to be able to compare the performance of two iterations of a given implementation of a protocol to determine where improvements have been made and where further improvements can be made.

o To understand, given a set of parameters (network conditions), how a particular implementation on a particular device is going to perform. For instance, if you were trying to decide the processing power (size of device) required in a certain location within a network, you can emulate the conditions which are going to exist at that point in the network and use the test described to measure the perfomance of several different routers. The results of these tests can provide one possible data point for an intelligent decision.

  If the device being tested is to be deployed in a running network, using routes taken from the network where the equipment is to be deployed rather than some generated topology in these tests will give results which are closer to the real preformance of the device. Care should be taken to emulate or take routes from the actual location in the network where the device will be (or would be) deployed. For instance, one set of routes may be

taken from an ABR, one set from an area 0 only router, various sets from stub area, another set from various normal areas, etc.

- To measure the performance of an OSPF implementation in a wide variety of scenarios.

- To be used as parameters in OSPF simulations by researchers. It may some times be required for certain kinds of research to measure the individual delays of each parameter within an OSPF implementation. These delays can be measured using the methods defined in [BENCHMARK].

- To help optimize certain configurable parameters. It may some times be helpful for operators to know the delay required for individual tasks so as to optimize the resource usage in the network i.e. if it is found that the processing time is x seconds on an router, it would be helpful to determine the rate at which to flood LSA's to that router so as to not overload the network.

3. Assumptions Made and Limitations of such measurements

- The interactions of convergence and forwarding; testing is restricted to events occurring within the control plane. Forwarding performance is the primary focus in [INTERCONNECT] and it is expected to be dealt with in work that ensues from [FIB-TERM].

- Duplicate LSAs are Acknowledged Immediately. A few tests rely on the property that duplicate LSA Acknowledgements are not delayed but are done immediately. However if some implementation does not acknowledge duplicate LSAs immediately on receipt, the testing methods presented in [BENCHMARK] could give inaccurate measurements.

- It is assumed that SPF is non-preemptive. If SPF is implemented so that it can (and will be) preempted, the SPF measurements taken in [BENCHMARK] would include the times that the SPF process is not running ([BENCHMARK] measures the total time taken for SPF to run, not the amount of time that SPF actually spends on the device's processor), thus giving inaccurate measurements.

- Some implementations may be multithreaded or use a multiprocess/multirouter model of OSPF. If because of this any of the assumptions taken in measurement are violated in such a model, it could lead to inaccurate measurements.

o The measurements resulting from the tests in [BENCHMARK] may not provide the information required to deploy a device in a large scale network. The tests described focus on individual components of an OSPF implementation's performance, and it may be difficult to combine the measurements in a way which accurately depicts a device's performance in a large scale network. Further research is required in this area.

o The measurements described in [BENCHMARK] should be used with great care when comparing two different implementations of OSPF from two different vendors. For instance, there are many other factors than convergence speed that need to be taken into consideration when comparing different vendor's products, and it's difficult to align the resources available on one device to the resources available on another device.

4. Observations on the Tests Described in [BENCHMARK]

Some observations taken while implementing the tests described in [BENCHMARK] are noted in this section.

4.1. Measuring the SPF Processing Time Externally

The most difficult test to perform is the external measurement of the time required to perform an SPF calculation, since the amount of time between the first LSA which indicates a topology change and the duplicate LSA is critical. If the duplicate LSA is sent too quickly, it may be received before the device under test actually begins running SPF on the network change information. If the delay between the two LSAs is too long, the device under test may finish SPF processing before receiving the duplicate LSA. It is important to closely investigate any delays between the receipt of an LSA and the beginning of an SPF calculation in the device under test; multiple tests with various delays might be required to determine what delay needs to be used to accurately measure the SPF calculation time.

Some implementations may force two intervals, the SPF hold time and the SPF delay, between successive SPF calculations. If an SPF hold time exists, it should be subtracted from the total SPF execution time. If an SPF delay exists, it should be noted in the test results.

4.2. Noise in the Measurement Device

The device on which measurements are taken (not the device under test) also adds noise to the test results, primarily in the form of delay in packet processing and measurement output. The largest source of noise is generally the delay between the receipt of packets by the measuring device and the information about the packet reaching the device's output, where the event can be measured. The following steps may be taken to reduce this sampling noise:

o Increasing the number of samples taken will generally improve the tester's ability to determine what is noise, and remove it from the results.

o Try to take time-stamp for a packet as early as possible. Depending on the operating system being used on the box, one can instrument the kernel to take the time-stamp when the interrupt is processed. This does not eliminate the noise completely, but at least reduces it.

o Keep the measurement box as lightly loaded as possible.

o Having an estimate of noise can also be useful.

The DUT also adds noise to the measurement. Points (a) and (c) apply to the DUT as well.

4.3. Gaining an Understanding of the Implementation Improves Measurements

While the tester will (generally) not have access to internal information about the OSPF implementation being tested using [BENCHMARK], the more thorough the tester's knowledge of the implementation is, the more accurate the results of the tests will be. For instance, in some implementations, the installation of routes in local routing tables may occur while the SPF is being calculated, dramatically impacting the time required to calculate the SPF.

4.4. Gaining an Understanding of the Tests Improves Measurements

One method which can be used to become familiar with the tests described in [BENCHMARK] is to perform the tests on an OSPF implementation for which all the internal details are available, such as [GATED]. While there is no assurance that any two implementations will be similar, this will provide a better understanding of the

tests themselves.

5. 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 ABRs 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. Tree Shape and the SPF Algorithm

The complexity of Dijkstra's algorithm depends on the data structure used for storing vertices with their current minimum distances from the source, with the simplest structure being a list of vertices currently reachable from the source. In a simple list of vertices, 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 algorithm 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, with E being the number of edges in the graph. 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:

```
root
 |
 1
 |
 2
 |
 3
 |
 4
 |
 5
 |
 6
```

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:

```
    root
   /    \
  1      2
 / \    / \
3   4  5   6
```

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).

So, the performance of the shortest path first algorithm used to compute the best paths through the network is dependant o the construction of the tree The 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 versus 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 [BROADCAST-P2P].

- 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. Security Considerations

This doecument does not modify the underlying security considerations in [OSPF].

9. Acknowledgements

Thanks to Howard Berkowitz, (hcb@clark.net) and the rest of the BGP benchmarking team for their support and to Kevin Dubray(kdubray@juniper.net) who realized the need of this draft.

10. Normative References

[BENCHMARK]
Manral, V., "Benchmarking Basic OSPF Single Router Control Plane Convergence", draft-bmwg-ospfconv-intraarea-08, May 2004

[TERM]Manral, V., "OSPF Convergence Testing Terminiology and Concepts", draft-bmwg-ospfconv-term-08, May 2004

[RFC2119]
Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997

11. Informative References

[INTERCONNECT]
Bradner, S., McQuaid, J., "Benchmarking Methodology for Network Interconnect Devices", RFC2544, March 1999.

[FIB-TERM]
Trotter, G., "Terminology for Forwarding Information Base (FIB) based Router Performance", RFC3222, October 2001.

[BROADCAST-P2P]
Shen, Naiming, et al., "Point-to-point operation over LAN in link-state routing protocols", draft-ietf-isis-igp-p2p-over-lan-03.txt, August, 2003.

[GATED]
http://www.gated.org

12. Authors' Addresses

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

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

Network Working Group   Vishwas Manral
Internet Draft   Netplane Systems
                                                              Russ White
                                                           Cisco Systems
                                                             Aman Shaikh
Expiration Date: November 2004   University of California
File Name: draft-bmwg-ospfconv-term-08.txt   May 2004

OSPF Benchmarking Terminology and Concepts

draft-bmwg-ospfconv-term-06.txt

Status of this Memo

Copyright Notice

Abstract

This draft explains the terminology and concepts used in OSPF benchmarking. While some of these terms may be defined elsewhere, and we will refer the reader to those definitions in some cases, we also include discussions concerning these terms as they relate specifically to the tasks involved in benchmarking the OSPF protocol.

1. Specification of Requirements

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 RFC 2119 [RFC2119].

2. Motivation

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

3. Common Definitions

Definitions in this section are well known industry and benchmarking terms which may be defined elsewhere.

- o White Box (Internal) Measurements

  - Definition

    White Box measurements are measurements reported and collected 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: White box 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.

    Because of the potential for variations in collection and presentation methods across different DUTs, white box

measurements MUST NOT be used as a basis of comparison in benchmarks. This has been a guiding principal of Benchmarking Methodology Working Group.

o Black Box (External) Measurements

- Definition

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

- Discussion

  One example of a black box 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 of internal operations may 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 [BENCHMARK], external techniques are more readily applicable.

o Multi-device Measurements

- Measurements assessing communications (usually in combination with internal operations) between two or more DUTs. Multi-device measurements may be internal or external.

4. Terms Defined Elsewhere

Terms in this section are defined elsewhere, and included only to include a discussion of those terms in reference to [BENCHMARK].

o Point-to-Point links

- Definition

  See [OSPF], Section 1.2.

- Discussion

  A point-to-point link can 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 [BENCHMARK] and [OSPF], the two can be regarded the same.

o Broadcast Link

- Definition

  See [OSPF], Section 1.2.

- 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 LSA 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 Execution Time

- Definition

The time taken by a router to complete the SPF process, as described in [OSPF].

- Discussion

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

  Some implementations may force two intervals, the SPF hold time and the SPF delay, between successive SPF calculations. If an SPF hold time exists, it should be subtracted from the total SPF execution time. If an SPF delay exists, it should be noted in the test results.

- Measurement Units

  The SPF time is generally measured in milliseconds.

o Hello Interval

- Definition

  See [OSPF], Section 7.1.

- Discussion

  The hello interval should be the same for all routers on a 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. Further discussion on small hello intervals is given in [OSPF-SCALING].

o Router Dead interval

- Definition

  See [OSPF], Section 7.1.

- Discussion

  This is advertised in the router's Hello Packets in the Router-DeadInterval 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.

5. Concepts

5.1. The Meaning of Single Router Control Plane 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 Single Router Convergence.

This concept does not include the time required for the control plane of the device to transfer the information required to forward packets to the data plane, nor the amount of time between the data plane receiving that information and being able to actually forward traffic.

5.2. Measuring Convergence

Obviously, there are several elements to convergence, even under the definition given above for a single device, including (but not limited to):

o   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.

5.3. 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 adjacencies, 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. Security Considerations

This doecument does not modify the underlying security considerations in [OSPF].

7. Acknowedgements

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

8. Normative References

[BENCHMARK]
Manral, V., "Benchmarking Basic OSPF Single Router Control Plane Convergence", draft-bmwg-ospfconv-intraarea-08, May 2004.

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

9. Informative References

[OSPF-SCALING]
Choudhury, Gagan L., Editor, "Prioritized Treatment of Specific OSPF Packets and Congestion Avoidance", draft-ietf-ospf-scalability-06.txt, August 2003.

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