Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: April 22, 2012

I. Varlashkin
Easynet Global Services
R. Papneja
Huawei Technologies (USA)
B. Parise
Cisco
T. Van Unen
Ixia

October 20, 2011

Convergence benchmarking on contemporary routers
draft-varlashkin-router-conv-bench-00

Abstract

This document specifies methodology for benchmarking convergence of routers without making assumptions about relation and dependencies between data- and control-planes. Provided methodology is primary intended for testing routers running BGP and some form of link-state IGP with or without MPLS. It may also be applicable for environments using MPLS-TE or GRE, however they're beyond scope of this document and such application is left for further study.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 22, 2012.

Copyright Notice

Table of Contents

1. Introduction

Ability of the network to restore traffic flow when primary path fails has always been important subject for network engineers, researchers and equipment manufacturers. Time to recover from a link or node failure has often been linked to routeing protocols convergence; and benchmarking of a routeing protocol convergence has often been considered sufficient for quantifying recovery performance. As long as routers could obtain new best path only after relevant routeing protocols perform their calculations such methodology was reasonable. However continuous improvements in hardware and software result in more and more routers being able to restore traffic flow even before routeing protocols converge. Methodology described in this document takes such fact into account.

When a failure occurs on the network a router needs to:

1. select new best path so that the packets, which already arrived to the router, can be forwarded

2. let other routers know about new network state so they can find new best path from their perspective

How fast a router can perform these two functions characterise router's performance with regards to convergence. Note that in general case each of these characteristics may or may not be related to the other. For example, some platform may need to perform calculations to find new best path and only then update local FIB and send relevant protocol updates to other routers, another platform can update local FIB without waiting for calculations to complete but still needs to wait for calculations before sending routeing protocol updates, third platform can use different optimisation for both FIB changes and routeing protocol updates without waiting for completion of the calculations. Other variations are also possible. This document makes no assumption about whether local FIB changes and routeing protocol updates dependencies on each other or on routeing protocol calculations.

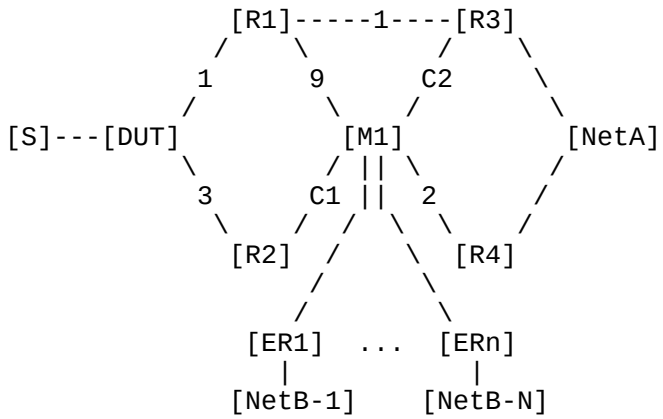
Since it is not known whether local FIB is updated before or after routeing protocol calculations, forwarding-plane method is proposed to benchmark local convergence. And because it is not known whether routeing protocol updates are linked to FIB modification or not the control-plane approach is used to benchmark how fast updates are propagated. However both characteristics are benchmarked using very similar test topologies and procedures. Also, an attempt is made to to minimise dependency on performance on non-DUT elements involved in the tests.

At the time of writing of this document it is not known whether existing network testers and protocol emulators are able to execute described tests out of the box. Nevertheless the authors believe that required functionality can be added with reasonable effort. Alternatively the tests can be performed with help of physical routers to create necessary test topology, which may have impact on time required to perform the test but expected to provide same degree of the test results accuracy. This also means that tests performed using a protocol simulator can be repeated using physical routers and results expected to be comparable.

This document complements draft-papneja-bgp-basic-dp-convergence.

2. Test topology

Unless specified otherwise all tests use same basic test topology outlined below:



S is source of test traffic for data-plane tests, while for control-plane tests S is an emulated or physical router with packet capturing (sniffing) capability.

Unidirectional test traffic goes from Source to NetA.

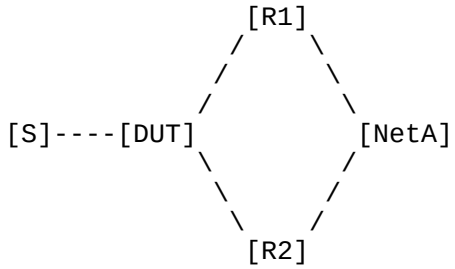
IGP between DUT and R1-R4; BGP between DUT and R3, R4; no BGP between R3 and R4 (important). If tunnelling (e.g. MPLS or GRE) is used then R1 and R2 do not need to run BGP, otherwise they MUST run BGP. Source has static default to DUT; R3 and R4 have static to NetA. NetA is in BGP but not in IGP. M1 is K*M matrix of internal routers. Metrics C1 is used to control whether R2 is LFA for DUT to NetA. Metric C2 is used to control whether R3 or R4 are best exit towards NetA. All other metrics are fixed for all tests and MUST be set to

exact values provided in the above diagram. IGP metrics from M1 to ER1 throughout ERn can be set arbitrarily, their exact values are irrelevant to this test as long as they're valid for given IGP.

Routers ER1 throughout ERn together with prefixes NetB-1 throughout NetB-N are presented to create realistic environment but not used directly in measurements. NetB-1 throughout NetB-N are distinct single-prefix sets.

Traffic restoration depends on ability of R2 and M1 to forward traffic after failure. To eliminate this dependency R2 is set to always forward traffic to R3 and NetA via M1 which in turn always forwards traffic directly via R3 or R2 depending on the test. One possibility to achieve this is to use static routes. Another alternative is to use different IGP between R2 and R3 from the one used by DUT and make routes learned via this IGP preferred on R2. E.g. DUT uses OSPF, then in addition to it R2&R3 also run ISIS and prefer ISIS routes over OSPF ones. A protocol simulator can have internal mechanism to provide required behaviour. There are no other dependencies on non-DUT devices in this tests.

For evaluating eBGP performance following topology is used:



Test topology for eBGP

In "Link failure without LoS" test direct cable between DUT and R1 is replaced with connection over an L2 switch as follow:

[DUT]---[SW1]---[R1]

3. TEST PARAMETERS

3.1. Packing ratios

Routes with different prefixes but same attributes can potentially be packed into single update message. Since both number of update messages and number of prefixes per update can affect convergence time, the tests SHOULD be performed with various prefix packing ratios. This document does not specify values of individual BGP attributes used to control packing ratio.

3.2. Test traffic

Traffic is sent from single source address located at the Source port of the tester to one address in each prefix in NetA set. Packets are sent at rate 1000 per second, which provides 1ms resolution of the convergence time as measured by tests in this document. All packets SHOULD be 64 bytes at IP layer, that is IP header plus IP payload.

3.3. IGP metrics

Basic test topology specifies fixed IGP metrics for some links. These metrics SHOULD be used verbatim. There are also two variable metrics - C1 and C2 - intended for controlling whether R2 is Loop-Free-Alternate (LFA) for DUT towards NetA, and whether R3 remains best exit towards NetA after path failure between DUT and R3. Following values SHOULD be used for C1 and C2 depending on required behaviour:

R2 is LFA?	R3 best?	C1	C2
yes	yes	1	1
yes	no	1	3
no	yes	5	1
no	no	5	3

3.4. Internal routers matrix

Basic test topology has N*K grid of internal routers denoted as M1. When N>1 or K>1 the cost of all links within grid MUST be set to 1 (one). This matrix is intended for controlling topology size, which has affect on particularly SPF run-time.

If traffic is forwarded using a tunneling mechanism, such as MPLS or GRE, the internal routers only need to have reachability information about tunnel end-points. However if traditional hop-by-hop forwarding is used, then internal routers MUST have routes to each and every prefix within NetA set.

This document does not specify how internal routers should obtain necessary reachability information. The only requirement is that after primary DUT-NetA path failure internal routers are able to forward traffic to NetA instantly. Using values of IGP metrics as described earlier addresses this requirement. Also, protocol simulator may have built-in mechanism to achieve desired behaviour.

3.5. Number of next-hops

Basic test topology has set of N edge routers ER1 throughout ERn, each advertising unique prefix. Some BGP implementations may exhibit different performance depending on number of next-hops for which IGP cost has changed after failure. By varying overall number of next-hops such dependency can be detected.

Note that prefixes NetB-1 throughout NetB-n are not used as destinations for test traffic, they’re only present for creating "background environment".

3.6. ’e’ - Failure and Restoration start entropy

Tests described in this document use fixed time T2 and variable offset ’e’ as starting point for simulating failure or restoration event.

Fixing time T2 is necessary as reference point to which variable offset e is added for each iteration of the test. Introduction of such variable offset allows better analysis of the test results. For example, DUT may run FIB changes at certain intervals. If failure introduced close to the end of such interval, shorter outage will be observed, and if introduced close to the beginning of such interval longer outage will be observed. Running test multiple times each time using different offset will help to profile DUT better.

Test report must contain value of T2 (same for all iterations) and values of e for each iterations. This document recommends to use T2=T1+8s and e from 0 to 1s in 0.01s (10ms) increments.

4. TEST PROCEDURES

This section provides generic steps that are used in all tests.

4.1. Initialisation time

The objective of this test is to measure time that must elapse between starting protocols and ability of the test topology to forward traffic. This test is not intended to reflect DUT

performance but used only as a way to find time T1 that is used in all subsequent tests.

To execute test perform following steps:

1. Configure DUT and protocol simulator (or auxiliary nodes)

2. At T0 start traffic and then immediately start routeing protocols

3. When traffic starts arriving Sink Port 1 stop test.

The time of arrival of the first packet is T1.

4.2. Generic data-plane failure test

The purpose of failure test is to measure time required by DUT to resume traffic flow after best path to destination fails. Following steps are common for all failure tests:

1. Start protocols and mark time as T0

2. At time T1 start traffic to each prefix in set NetA

3. At T2+e simulate failure or restoration event (see Section 5)

4. From T2+e until T3 packets do not arrive to NetA

5. After packets are seen again at NetA (T3) wait until time T4

6. Stop traffic

7. Measure total number of lost packets and calculate outage knowing packet-per-second

4.3. Generic test procedure for

1. At T0 bring up all interfaces and protocols, and start capturing BGP packets at RS1

2. At T1+e simulate failure/restoration event (see Section 5)

3. At T2-d1 first UPDATE message is sent by DUT and at T2 it will be observed at RS1

4. At T3-d2 last UPDATE message is sent by DUT and at T3 it will be observed at RS1

d1 and d2 represent serialisation and propagation delay and can be

disregarded unless DUT-RS1 link has large delay. With this in mind, T2-(T1+e) and T3-(T1+e) represent convergence time for the first and last prefix respectively.

5. Failure and restoration scenarios

This section defines set of various failure and restoration scenarios used in step 3 of the generic test procedures described in previous section. Unless otherwise specified all scenarios are applicable to both data- and control-plane test procedures.

5.1. Loss of Signal on the link attached to DUT

This scenario simulates situation where link attached to DUT fails and Loss of Signal (LoS) can be observed by DUT. In other words link fails and results in interface on the DUT going down.

To simulate LoS failure at the time defined by the test procedure shut down R1 side of the link to DUT.

To simulate LoS restoration at the time defined by the test procedure re-activate R1 side of the link to DUT.

5.2. Link failure without LoS

This scenario simulates situation where link between DUT and adjacent node fails but DUT does not observe LoS. In practice such failure can occur when, for example, link between DUT and adjacent node is implemented via carrier equipment that does not shut link down when remote side of the link fails.

DUT can use various methods to detect such failures, including but not limited to protocol HELLO or Keep-alive packets, BFD, OAM. This document does not restrict methods which DUT can use, but requires use of particular method to be recorded in the test report.

Basic network topology is modified for the purpose of this test only as follow: rather than using direct cabling between DUT and R1 the link is implemented via intermediate L2 switch that supports concept of VLAN’s. Initially switch ports connected to DUT and R1 are placed into the same VLAN (same L2 broadcast domain).

To simulate failure at the time defined by the test procedure move switch port connected to R1 to a VLAN different from the one used for switch port connected to DUT.

To simulate restoration at the time defined by the test procedure

move switch port connected to R1 back to the same VLAN as the one used for switch port connected to DUT.

5.3. Non-direct link failure

This scenario simulates situation where a link not directly connected to DUT but located on the primary path to destination fails. Unmodified basic network topology is used.

Depending on technologies used in the setup different failure detection techniques can be employed by DUT. This document assumes that DUT relies exclusively on IGP information to learn about failure and that nodes adjacent to the failed link flood this information within D seconds since the event. If required exact value of D can be obtained through simple additional test, but in this document D is assumed to be 0 (zero).

It is possible, though undesirable, that some traffic and protocol simulators may contunue accepting packets coming through the port that leads to simulated failed link. It is essential to assert such behaviour prior to the tests and if confirmed, exclude packets received after failure from calculations in step 7 of the test.

Failure event is triggered by simulating shutdown of R3 side of the link to R1 at the time defined by the test procedure. R1 MUST send IGP update (depending on which protocol is used) to DUT within D seconds.

Restoration event is triggered by simulating recovery of R3 side of the link to R1 at the time defined by the test procedure. R1 MUST send IGP update (depending on which protocol is used) to DUT within D seconds.

5.4. Best route withdrawal

This scenario sumulates situation where best AS exit path to a destination is no longer valid and ASBR sends BGP UPDATE to its iBGP peers. Unmodified basic network topology is used.

Disconnecting R3 from NetA implies that R3 will send BGP WITHDRAW for this prefixes in its update to DUT. It is possible, though undesirable, that some protocol simulator and traffic generators will still count packets received at sink port 1 even after prefixes were withdrawn. To correctly execute this test it’s mandatory that traffic received at sink port 1 after withdrawing prefixes is ignored and not counted as delivered. If traffic generator is not able to assure such functionality (should be asserted prior to the test), then packets received at the sink port 1 MUST be excluded from

calculation in step 7 of the test.

Failure event is triggered by simulating failure of the link between R3 and NetA and immediate withdrawal of all corresponding prefixes by R3.

Restoration event is triggered by simulating recovery of the link between R3 and NetA and immediate BGP UPDATE for all corresponding prefixes by R3.

5.5. iBGP next-hop failure

This scenario simulates situation where ASBR used as best exit to a destination unexpectedly fails both at control and forwarding plane. Both R1 and a router within M1 connected to R3 MUST send appropriate IGP update message to the rest of the network within D seconds. To detect failure DUT MAY rely on IGP information provided by rest of the network or it MAY employ additional techniques. This document does not restrict what detection mechanism should DUT use but requires that particular mechanism is recorded in the test report.

Failure event is triggered by simulating removal of R3 from the test topology at the time defined by the test procedure, followed by IGP update as described in previous paragraph.

Recovery event is triggered by re-introducing R3 into the test topology, followed by IGP update as described in first paragraph of this section and immediate re-activation of BGP session between R3 and DUT. Note that recovery time calculated by this method depends on DUT performance in respect to bringing up new BGP session. This is intentional. Control plane convergence benchmarking can be performed separately by a method that is outside of the scope of this document and two results can be correlated netto data-plane convergence value should that be necessary.

6. Test report

TODO: Report format is to be discussed.

Test report MUST contain following data for each test:

1. T1 and ’e’

2. Number of prefixes NetA and NetB

3. Size of M1 (recored as N*K)

4. Traffic rate, in packets per second, and packet size at IP layer in octets

5. Number of lost packets during falure, and number of lost packets during restoration

7. Link bundling and Equal Cost Multi-Path

Scenarios where DUT can balance traffic to NetA across multiple best paths is explicitly excluded from scope of this document. There are two reasons.

First, two different DUT may choose different path (out of all equal) to forward given packet, which makes it unreasonably difficult to define generic traffic that would produce comparable results when testing different platforms.

Second, mechanisms used to handle failures in ECMP (but not necessarily in link-bundling) environment are similar to those handling single-path failures. Therefore it’s expected that convergence in ECMP scenario will be of the same order as in single-path scenario.

8. Graceful Restart and Non-Stop Forwarding

While Graceful Restart and Non-Stop Forwarding mechanisms are related to DUT ability to forward traffic under certain failure conditions, the test covering DUT own ability to restore or preserve traffic flow already covered in RFC6201.

9. Security considerations

The tests described in this document intended to be performed in isolated lab environment, which inheretently has no security implication on the live network of the organisation or Internet as whole.

Authors foresee that some people or organisations might be interested to benchmark performance of the live networks. The tests described in this document are disruptive by their nature and will have impact at least on the network where they’re executed, and depending on the role of that network effect can extend to other parts of the Internet. Such tests MUST NOT be attempted in live environment without careful consideration.

The fact of publishing this document does not increase potential negative consequences if tests are executed in live environment because information provided here is mere recording of widely known and used techniques.

10. IANA Considerations

None.

11. Acknowledgments

Authors would like to thank Gregory Cauchie, Rob Shakir, David Freedman, Anton Elita, Saku Ytti, Andrew Yourtchenko, for their valuable contribution and peer-review of this work.

12. Normative References

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

Authors' Addresses

Ilya Varlashkin
Easynet Global Services

Email: ilya.varlashkin@easynet.com

Rajiv Papneja
Huawei Technologies (USA)

Email: rajiv.papneja@huawei.com

Bhavani Parise
Cisco

Email: bhavani@cisco.com

Tara Van Unen
Ixia

Email: TVanUnen@ixiacom.com