Internet Engineering Task Force (IETF)
Request for Comments: 7747
Category: Informational
ISSN: 2070-1721

R. Papneja
Huawei Technologies
B. Parise
Skyport Systems
S. Hares
Huawei Technologies
D. Lee
IXIA
I. Varlashkin
Google

April 2016

Basic BGP Convergence Benchmarking Methodology for Data-Plane Convergence

Abstract

BGP is widely deployed and used by several service providers as the default inter-AS (Autonomous System) routing protocol. It is of utmost importance to ensure that when a BGP peer or a downstream link of a BGP peer fails, the alternate paths are rapidly used and routes via these alternate paths are installed. This document provides the basic BGP benchmarking methodology using existing BGP convergence terminology as defined in RFC 4098.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc7747.

Table of Contents

1. Introduction

This document defines the methodology for benchmarking data-plane Forwarding Information Base (FIB) convergence performance of BGP in routers and switches using topologies of three or four nodes. The methodology proposed in this document applies to both IPv4 and IPv6, and if a particular test is unique to one version, it is marked accordingly. For IPv6 benchmarking, the Device Under Test (DUT) will require the support of Multiprotocol BGP (MP-BGP) [RFC4760] [RFC2545]. Similarly, both Internal BGP (iBGP) and External BGP (eBGP) are covered in the tests as applicable.

The scope of this document is to provide methodology for BGP FIB convergence measurements with BGP functionality limited to IPv4 and IPv6 as defined in [RFC4271] and MP-BGP [RFC4760] [RFC2545]. Other BGP extensions to support Layer 2 and Layer 3 Virtual Private Networks (VPNs) are outside the scope of this document. Interaction with IGPs (IGP interworking) is outside the scope of this document.

1.1. Benchmarking Definitions

The terminology used in this document is defined in [RFC4098]. One additional term is defined in this document as follows.

FIB (data-plane) convergence is defined as the completion of all FIB changes so that all forwarded traffic then takes the newly proposed route. RFC 4098 defines the terms 'BGP device', 'FIB', and 'forwarded traffic'. Data-plane convergence is different than control-plane convergence within a node.

This document defines methodology to test

- data-plane convergence on a single BGP device that supports the BGP functionality with a scope as outlined above; and

- using test topology of three or four nodes that are sufficient to recreate the convergence events used in the various tests of this document.

1.2. Purpose of BGP FIB (Data-Plane) Convergence

In the current Internet architecture, the inter-AS transit is primarily available through BGP. To maintain reliable connectivity within intra-domains or across inter-domains, fast recovery from failures remains most critical. To ensure minimal traffic losses, many service providers are requiring BGP implementations to converge the entire Internet routing table within sub-seconds at FIB level.

Furthermore, to compare these numbers amongst various devices, service providers are also looking at ways to standardize the convergence measurement methods. This document offers test methods for simple topologies. These simple tests will provide a quick high-level check of BGP data-plane convergence across multiple implementations from different vendors.

1.3. Control-Plane Convergence

The convergence of BGP occurs at two levels: Routing Information Base (RIB) and FIB convergence. RFC 4098 defines terms for BGP control-plane convergence. Methodologies that test control-plane convergence are out of scope for this document.

1.4. Benchmarking Testing

In order to ensure that the results obtained in tests are repeatable, careful setup of initial conditions and exact steps are required.

This document proposes these initial conditions, test steps, and result checking. To ensure uniformity of the results, all optional parameters SHOULD be disabled and all settings SHOULD be changed to default; these may include BGP timers as well.

2. Existing Definitions and Requirements

"Benchmarking Terminology for Network Interconnect Devices" [RFC1242] and "Benchmarking Terminology for LAN Switching Devices" [RFC2285] SHOULD be reviewed in conjunction with this document. WLAN-specific terms and definitions are also provided in Clauses 3 and 4 of the IEEE 802.11 standard [IEEE.802.11]. Commonly used terms may also be found in RFC 1983 [RFC1983].

For the sake of clarity and continuity, this document adopts the general template for benchmarking terminology set out in Section 2 of [RFC1242]. Definitions are organized in alphabetical order and grouped into sections for ease of reference. The following terms are assumed to be taken as defined in RFC 1242 [RFC1242]: Throughput, Latency, Constant Load, Frame Loss Rate, and Overhead Behavior. In addition, the following terms are taken as defined in [RFC2285]: Forwarding Rates, Maximum Forwarding Rate, Loads, Device Under Test (DUT), and System Under Test (SUT).

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].

3. Test Topologies

This section describes the test setups for use in BGP benchmarking tests measuring convergence of the FIB (data-plane) after BGP updates have been received.

These test setups have three or four nodes with the following configuration:

1. Basic test setup

2. Three-node setup for iBGP or eBGP convergence

3. Setup for eBGP multihop test Scenario

4. Four-node setup for iBGP or eBGP convergence

Individual tests refer to these topologies.

Figures 1 through 4 use the following conventions:

- AS-X: Autonomous System X

- Loopback Int: Loopback interface on a BGP-enabled device

- HLP, HLP1, HLP2: Helper routers running the same version of BGP as the DUT

- All devices MUST be synchronized using NTP or some other clock synchronization mechanism

3.1. General Reference Topologies

Emulator acts as one or more BGP peers for different test cases.

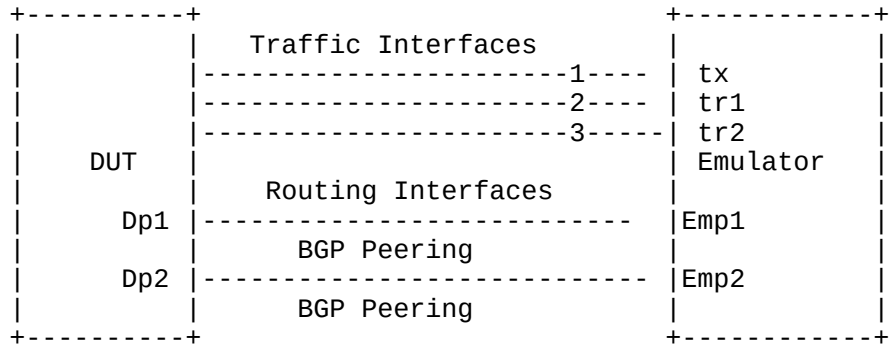

Figure 1: Basic Test Setup

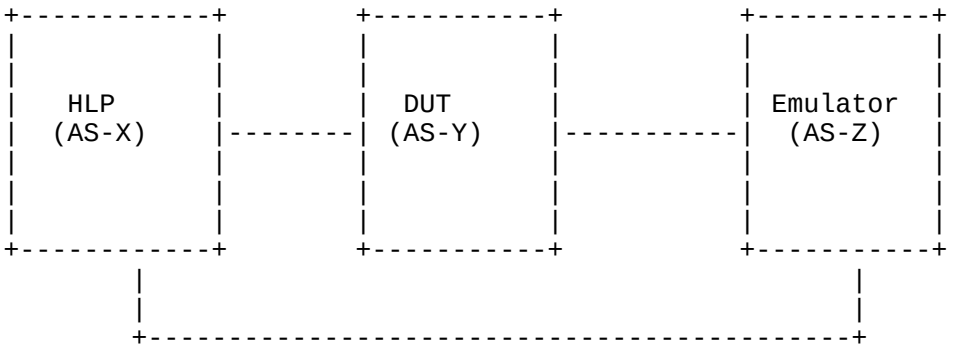

Figure 2: Three-Node Setup for eBGP and iBGP Convergence

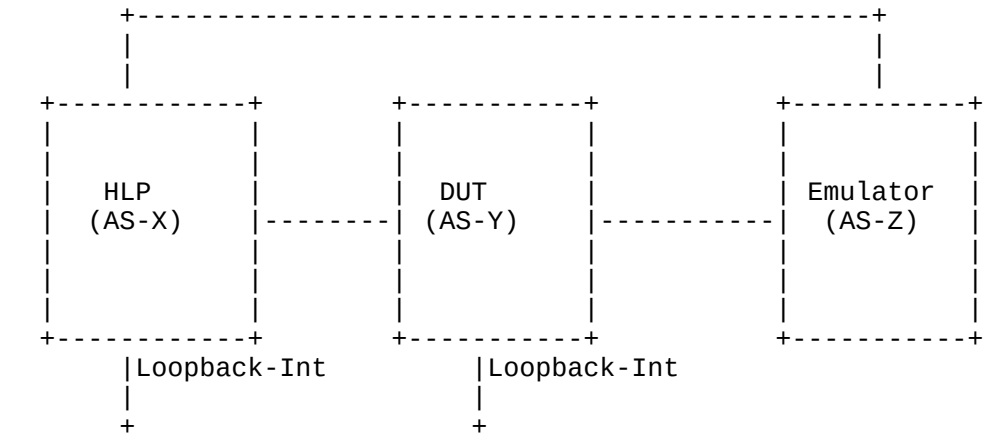

Figure 3: BGP Convergence for eBGP Multihop Scenario

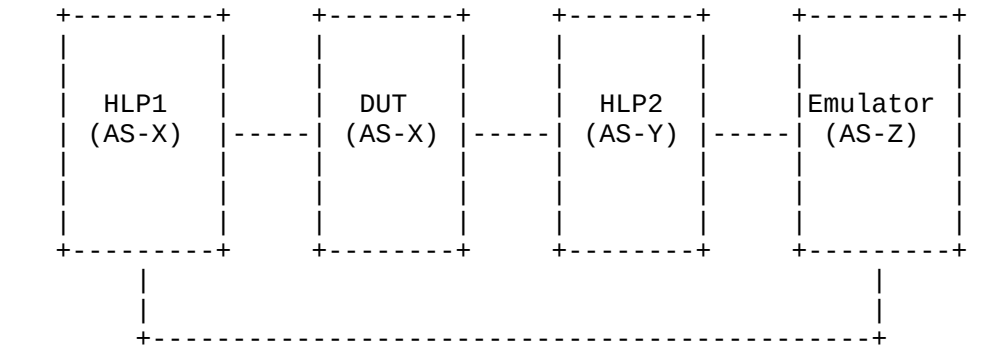

Figure 4: Four-Node Setup for eBGP and iBGP Convergence

4. Test Considerations

The test cases for measuring convergence for iBGP and eBGP are different. Both iBGP and eBGP use different mechanisms to advertise, install, and learn the routes. Typically, an iBGP route on the DUT is installed and exported when the next hop is valid. For eBGP, the route is installed on the DUT with the remote interface address as the next hop, with the exception of the multihop test case (as specified in the test).

4.1. Number of Peers

"Number of Peers" is defined as the number of BGP neighbors or sessions the DUT has at the beginning of the test. The peers are established before the tests begin. The relationship could be either iBGP or eBGP peering depending upon the test case requirement.

The DUT establishes one or more BGP peer sessions with one or more emulated routers or Helper Nodes. Additional peers can be added based on the testing requirements. The number of peers enabled during the testing should be well documented in the report matrix.

4.2. Number of Routes per Peer

"Number of Routes per Peer" is defined as the number of routes advertised or learned by the DUT per session or through a neighbor relationship with an emulator or Helper Node. The Tester, emulating as a BGP neighbor, MUST advertise at least one route per BGP peer.

Each test run must identify the route stream in terms of route packing, route mixture, and number of routes. This route stream must be well documented in the reporting stream. RFC 4098 defines these terms.

It is RECOMMENDED that the user consider advertising the entire current Internet routing table per peering session using an Internet route mixture with unique or non-unique routes. If multiple peers are used, it is important to precisely document the timing sequence between the peer sending routes (as defined in RFC 4098).

4.3. Policy Processing/Reconfiguration

The DUT MUST run one baseline test where policy is the Minimal policy as defined in RFC 4098. Additional runs may be done with the policy that was set up before the tests began. Exact policy settings MUST be documented as part of the test.

4.4. Configured Parameters (Timers, etc.)

There are configured parameters and timers that may impact the measured BGP convergence times.

The benchmark metrics MAY be measured at any fixed values for these configured parameters.

It is RECOMMENDED these configure parameters have the following settings: a) default values specified by the respective RFC, b) platform-specific default parameters, and c) values as expected in the operational network. All optional BGP settings MUST be kept consistent across iterations of any specific tests

Examples of the configured parameters that may impact measured BGP convergence time include, but are not limited to:

1. Interface failure detection timer
2. BGP keepalive timer
3. BGP holdtime
4. BGP update delay timer
5. ConnectRetry timer
6. TCP segment size
7. Minimum Route Advertisement Interval (MRAI)
8. MinASOriginationInterval (MAOI)
9. Route flap damping parameters
10. TCP Authentication Option (TCP AO or TCP MD5)
11. Maximum TCP window size
12. MTU

The basic-test settings for the parameters should be:

1. Interface failure detection timer (0 ms)
2. BGP keepalive timer (1 min)
3. BGP holdtime (3 min)
4. BGP update delay timer (0 s)
5. ConnectRetry timer (1 s)
6. TCP segment size (4096 bytes)
7. Minimum Route Advertisement Interval (MRAI) (0 s)

8. MinASOriginationInterval (MAOI) (0 s)

9. Route flap damping parameters (off)

10. TCP Authentication Option (off)

4.5. Interface Types

The type of media dictates which test cases may be executed; each interface type has a unique mechanism for detecting link failures, and the speed at which that mechanism operates will influence the measurement results. All interfaces MUST be of the same media and throughput for all iterations of each test case.

4.6. Measurement Accuracy

Since observed packet loss is used to measure the route convergence time, the time between two successive packets offered to each individual route is the highest possible accuracy of any packet-loss-based measurement. When packet jitter is much less than the convergence time, it is a negligible source of error, and hence, it will be treated as within tolerance.

Other options to measure convergence are the Time-Based Loss Method (TBLM) and Timestamp-Based Method (TBM) [RFC6414].

An exterior measurement on the input media (such as Ethernet) is defined by this specification.

4.7. Measurement Statistics

The benchmark measurements may vary for each trial due to the statistical nature of timer expirations, CPU scheduling, etc. It is recommended to repeat the test multiple times. Evaluation of the test data must be done with an understanding of generally accepted testing practices regarding repeatability, variance, and statistical significance of a small number of trials.

For any repeated tests that are averaged to remove variance, all parameters MUST remain the same.

4.8. Authentication

Authentication in BGP is done using the TCP Authentication Option [RFC5925]. (In some legacy situations, the authentication may still be with TCP MD5). The processing of the authentication hash, particularly in devices with a large number of BGP peers and a large amount of update traffic, can have an impact on the control plane of

the device. If authentication is enabled, it MUST be documented correctly in the reporting format.

Also, it is recommended that trials MUST be with the same Secure Inter-Domain Routing (SIDR) features [RFC7115] [BGPsec]. The best convergence tests would be with no SIDR features and then to repeat the convergence tests with the same SIDR features.

4.9. Convergence Events

Convergence events or triggers are defined as abnormal occurrences in the network, which initiate route flapping in the network and hence forces the reconvergence of a steady state network. In a real network, a series of convergence events may cause convergence latency operators desire to test.

These convergence events must be defined in terms of the sequences defined in RFC 4098. This basic document begins all tests with a router initial setup. Additional documents will define BGP data-plane convergence based on peer initialization.

The convergence events may or may not be tied to the actual failure. A soft reset [RFC4098] does not clear the RIB or FIB tables. A hard reset clears BGP peer sessions, RIB tables, and FIB tables.

4.10. High Availability

Due to the different Non-Stop-Routing (sometimes referred to High-Availability) solutions available from different vendors, it is RECOMMENDED that any redundancy available in the routing processors should be disabled during the convergence measurements. For cases where the redundancy cannot be disabled, the results are no longer comparable and the level of impact on the measurements is out of scope of this document.

5. Test Cases

All tests defined under this section assume the following:

a. BGP peers are in Established state.

b. BGP state should be cleared from Established state to Idle prior to each test. This is recommended to ensure that all tests start with BGP peers being forced back to Idle state and databases flushed.

c. Furthermore, the traffic generation and routing should be verified in the topology to ensure there is no packet loss observed on any advertised routes.

d. The arrival timestamp of advertised routes can be measured by installing an inline monitoring device between the emulator and the DUT or by using the span port of the DUT connected with an external analyzer. The time base of such an inline monitor or external analyzer needs to be synchronized with the protocol and traffic emulator. Some modern emulators may have the capability to capture and timestamp every NLRI packet leaving and arriving at the emulator ports. The timestamps of these NLRI packets will be almost identical to the arrival time at the DUT if the cable distance between the emulator and DUT is relatively short.

5.1. Basic Convergence Tests

These test cases measure characteristics of a BGP implementation in non-failure scenarios like:

1. RIB-IN Convergence

2. RIB-OUT Convergence

3. eBGP Convergence

4. iBGP Convergence

5.1.1. RIB-IN Convergence

Objective:

This test measures the convergence time taken to receive and install a route in RIB using BGP.

Reference Test Setup:

This test uses the setup as shown in Figure 1

Procedure:

A. All variables affecting convergence should be set to a basic test state (as defined in Section 4.4).

B. Establish BGP adjacency between the DUT and one peer of the emulator, Emp1.

C. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

D. Start the traffic from the emulator tx towards the DUT targeted at a route specified in the route mixture (e.g., routeA). Initially, no traffic SHOULD be observed on the egress interface as routeA is not installed in the forwarding database of the DUT.

E. Advertise routeA from the peer (Emp1) to the DUT and record the time.

   This is Tup(Emp1,Rt-A), also named XMT-Rt-time(Rt-A).

F. Record the time when routeA from Emp1 is received at the DUT.

   This is Tup(DUT,Rt-A), also named RCV-Rt-time(Rt-A).

G. Record the time when the traffic targeted towards routeA is received by the emulator on the appropriate traffic egress interface.

   This is TR(TDr,Rt-A), also named DUT-XMT-Data-Time(Rt-A).

H. The difference between the Tup(DUT,RT-A) and traffic received time (TR (TDr, Rt-A) is the FIB convergence time for routeA in the route mixture. A full convergence for the route update is the measurement between the first route (Rt-A) and the last route (Rt-last).

   Route update convergence is

   TR(TDr, Rt-last)- Tup(DUT, Rt-A), or

   (DUT-XMT-Data-Time - RCV-Rt-Time)(Rt-A).

Note: It is recommended that a single test with the same route mixture be repeated several times. A report should provide the standard deviation and the average of all tests.

Running tests with a varying number of routes and route mixtures is important to get a full characterization of a single peer.

5.1.2. RIB-OUT Convergence

Objective:

This test measures the convergence time taken by an implementation to receive, install, and advertise a route using BGP.

Reference Test Setup:

This test uses the setup as shown in Figure 2.

Procedure:

A. The Helper Node (HLP) MUST run same version of BGP as the DUT.

B. All devices MUST be synchronized using NTP or some local reference clock.

C. All configuration variables for the Helper Node, DUT, and emulator SHOULD be set to the same values. These values MAY be basic test or a unique set completely described in the test setup.

D. Establish BGP adjacency between the DUT and the emulator.

E. Establish BGP adjacency between the DUT and the Helper Node.

F. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

G. Start the traffic from the emulator towards the Helper Node targeted at a specific route (e.g., routeA). Initially, no traffic SHOULD be observed on the egress interface as routeA is not installed in the forwarding database of the DUT.

H. Advertise routeA from the emulator to the DUT and note the time.

   This is Tup(EMx, Rt-A), also named EM-XMT-Data-Time(Rt-A).

I. Record when routeA is received by the DUT.

   This is Tup(DUTr, Rt-A), also named DUT-RCV-Rt-Time(Rt-A).

J. Record the time when routeA is forwarded by the DUT towards the Helper Node.

   This is Tup(DUTx, Rt-A), also named DUT-XMT-Rt-Time(Rt-A).

K. Record the time when the traffic targeted towards routeA is received on the Route Egress Interface. This is TR(EMr, Rt-A), also named DUT-XMT-Data Time(Rt-A).

FIB convergence = (DUT-XMT-Data-Time -DUT-RCV-Rt-Time)(Rt-A)

RIB convergence = (DUT-XMT-Rt-Time - DUT-RCV-Rt-Time)(Rt-A)

Convergence for a route stream is characterized by

a) individual route convergence for FIB and RIB, and

b) all route convergence of

FIB-convergence = DUT-XMT-Data-Time(last) - DUT-RCV-Rt-Time(first), and

RIB-convergence = DUT-XMT-Rt-Time(last) - DUT-RCV-Rt-Time(first).

5.1.3. eBGP Convergence

Objective:

This test measures the convergence time taken by an implementation to receive, install, and advertise a route in an eBGP Scenario.

Reference Test Setup:

This test uses the setup as shown in Figure 2, and the scenarios described in RIB-IN and RIB-OUT are applicable to this test case.

5.1.4. iBGP Convergence

Objective:

This test measures the convergence time taken by an implementation to receive, install, and advertise a route in an iBGP Scenario.

Reference Test Setup:

This test uses the setup as shown in Figure 2, and the scenarios described in RIB-IN and RIB-OUT are applicable to this test case.

5.1.5. eBGP Multihop Convergence

Objective:

This test measures the convergence time taken by an implementation to receive, install, and advertise a route in an eBGP Multihop Scenario.

Reference Test Setup:

This test uses the setup as shown in Figure 3. The DUT is used along with a Helper Node.

Procedure:

A. The Helper Node MUST run the same version of BGP as the DUT.

B. All devices MUST be synchronized using NTP or some local reference clock.

C. All variables affecting convergence, like authentication, policies, and timers, SHOULD be set to basic settings.

D. All three devices, the DUT, emulator, and Helper Node, are configured with different ASs.

E. Loopback interfaces are configured on the DUT and Helper Node, and connectivity is established between them using any config options available on the DUT.

F. Establish BGP adjacency between the DUT and the emulator.

G. Establish BGP adjacency between the DUT and the Helper Node.

H. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test

I. Start the traffic from the emulator towards the DUT targeted at a specific route (e.g., routeA).

J. Initially, no traffic SHOULD be observed on the egress interface as routeA is not installed in the forwarding database of the DUT.

K. Advertise routeA from the emulator to the DUT and note the time (Tup(EMx,RouteA), also named Route-Tx-time(Rt-A).

L. Record the time when the route is received by the DUT. This is Tup(EMr,DUT), also named Route-Rcv-time(Rt-A).

M. Record the time when the traffic targeted towards routeA is received from the egress interface of the DUT on the emulator. This is Tup(EMd,DUT) named Data-Rcv-time(Rt-A)

N. Record the time when routeA is forwarded by the DUT towards the Helper Node. This is Tup(EMf,DUT), also named Route-Fwd-time(Rt-A).

FIB Convergence = (Data-Rcv-time - Route-Rcv-time)(Rt-A)

RIB Convergence = (Route-Fwd-time - Route-Rcv-time)(Rt-A)

Note: It is recommended that the test be repeated with a varying number of routes and route mixtures. With each set route mixture, the test should be repeated multiple times. The results should record the average, mean, standard deviation.

5.2. BGP Failure/Convergence Events

5.2.1. Physical Link Failure on DUT End

Objective:

This test measures the route convergence time due to a local link failure event at the DUT’s Local Interface.

Reference Test Setup:

This test uses the setup as shown in Figure 1. The shutdown event is defined as an administrative shutdown event on the DUT.

Procedure:

A. All variables affecting convergence, like authentication, policies, and timers, should be set to basic-test policy.

B. Establish two BGP adjacencies from the DUT to the emulator, one over the peer interface and the other using a second peer interface.

C. Advertise the same route, routeA, over both adjacencies with preferences so that the Best Egress Interface for the preferred next hop is (Emp1) interface.

D. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

E. Start the traffic from the emulator towards the DUT targeted at a specific route (e.g., routeA). Initially, traffic would be observed on the best egress route, Emp1, instead of Emp2.

F. Trigger the shutdown event of Best Egress Interface on the DUT (Dp1). This time is called Shutdown time.

G. Measure the convergence time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface (Dp2).

   Time = Data-detect(Emp2) - Shutdown time

H. Stop the offered load and wait for the queues to drain. Restart the data flow.

I. Bring up the link on the DUT’s Best Egress Interface.

J. Measure the convergence time taken for the traffic to be rerouted from Dp2 to Best Egress Interface, Dp1.

   Time = Data-detect(Emp1) - Bring Up time

K. It is recommended that the test be repeated with a varying number of routes and route mixtures or with a number of routes and route mixtures closer to what is deployed in operational networks.

5.2.2. Physical Link Failure on Remote/Emulator End

Objective:

This test measures the route convergence time due to a local link failure event at the Tester’s Local Interface.

Reference Test Setup:

This test uses the setup as shown in Figure 1. The shutdown event is defined as a shutdown of the local interface of the Tester via a logical shutdown event. The procedure used in Section 5.2.1 is used for the termination.

5.2.3. ECMP Link Failure on DUT End

Objective:

This test measures the route convergence time due to a local link failure event at the ECMP member. The FIB configuration and BGP are set to allow two ECMP routes to be installed. However, policy directs the routes to be sent only over one of the paths.

Reference Test Setup:

This test uses the setup as shown in Figure 1, and the procedure used in Section 5.2.1.

5.3. BGP Adjacency Failure (Non-Physical Link Failure) on Emulator

Objective:

This test measures the route convergence time due to BGP Adjacency Failure on the emulator.

Reference Test Setup:

This test uses the setup as shown in Figure 1.

Procedure:

A. All variables affecting convergence, like authentication, policies, and timers, should be set to basic-policy.

B. Establish two BGP adjacencies from the DUT to the emulator: one over the Best Egress Interface and the other using the Next-Best Egress Interface.

C. Advertise the same route, routeA, over both adjacencies with preferences so that the Best Egress Interface for the preferred next hop is (Emp1) interface.

D. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

E. Start the traffic from the emulator towards the DUT targeted at a specific route (e.g., routeA). Initially, traffic would be observed on the Best Egress Interface.

F. Remove BGP adjacency via a software adjacency down on the emulator on the Best Egress Interface. This time is called BGPadj-down-time, also termed BGPpeer-down.

G. Measure the convergence time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface. This time is Tr-rr2, also called TR2-traffic-on.

   Convergence = TR2-traffic-on - BGPpeer-down

H. Stop the offered load and wait for the queues to drain and restart the data flow.

I. Bring up BGP adjacency on the emulator over the Best Egress Interface. This time is BGP-adj-up, also called BGPpeer-up.

J. Measure the convergence time taken for the traffic to be rerouted to the Best Egress Interface. This time is Tr-rr1, also called TR1-traffic-on.

   Convergence = TR1-traffic-on - BGPpeer-up

5.4. BGP Hard Reset Test Cases

5.4.1. BGP Non-Recovering Hard Reset Event on DUT

Objective:

This test measures the route convergence time due to a hard reset on the DUT.

Reference Test Setup:

This test uses the setup as shown in Figure 1.

Procedure:

A. The requirement for this test case is that the hard reset event should be non-recovering and should affect only the adjacency between the DUT and the emulator on the Best Egress Interface.

B. All variables affecting the test SHOULD be set to basic-test values.

C. Establish two BGP adjacencies from the DUT to the emulator: one over the Best Egress Interface and the other using the Next-Best Egress Interface.

D. Advertise the same route, routeA, over both adjacencies with preferences so that the Best Egress Interface for the preferred next hop is (Emp1) interface.

E. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

F. Start the traffic from the emulator towards the DUT targeted at a specific route (e.g., routeA). Initially, traffic would be observed on the Best Egress Interface.

G. Trigger the hard reset event of the Best Egress Interface on the DUT. This time is called time reset.

H. This event is detected and traffic is forwarded to the Next-Best Egress Interface. This time is called time-traffic flow.

I. Measure the convergence time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface.

   Time of convergence = time-traffic flow - time-reset

J. Stop the offered load and wait for the queues to drain and restart.

K. It is recommended that the test be repeated with a varying number of routes and route mixtures or with a number of routes and route mixtures closer to what is deployed in operational networks.

L. When varying number of routes are used, convergence time is measured using the Loss-Derived method [RFC6412].

M. Convergence time in this scenario is influenced by failure detection time on the Tester, BGP keepalive time and routing, and forwarding table update time.

5.5. BGP Soft Reset

Objective:

This test measures the route convergence time taken by an implementation to service a BGP Route Refresh message and advertise a route.

Reference Test Setup:

This test uses the setup as shown in Figure 2.

Procedure:

A. The BGP implementation on the DUT and Helper Node needs to support BGP Route Refresh Capability [RFC2918].

B. All devices MUST be synchronized using NTP or some local reference clock.

C. All variables affecting convergence, like authentication, policies, and timers, should be set to basic-test defaults.

D. The DUT and the Helper Node are configured in the same AS, whereas the emulator is configured under a different AS.

E. Establish BGP adjacency between the DUT and the emulator.

F. Establish BGP adjacency between the DUT and the Helper Node.

G. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

H. Configure a policy under the BGP on the Helper Node to deny routes received from the DUT.

I. Advertise routeA from the emulator to the DUT.

J. The DUT will try to advertise the route to the Helper Node; it will be denied.

K. Wait for three keepalives.

L. Start the traffic from the emulator towards the Helper Node targeted at a specific route, say routeA. Initially, no traffic would be observed on the egress interface, as routeA is not present.

M. Remove the policy on the Helper Node and issue a route refresh request towards the DUT. Note the timestamp of this event. This is the RefreshTime.

N. Record the time when the traffic targeted towards routeA is received on the egress interface. This is RecTime.

O. The following equation represents the Route Refresh Convergence Time per route.

Route Refresh Convergence Time = (RecTime - RefreshTime)

5.6. BGP Route Withdrawal Convergence Time

Objective:

This test measures the route convergence time taken by an implementation to service a BGP withdraw message and advertise the withdraw.

Reference Test Setup:

This test uses the setup as shown in Figure 2.

Procedure:

A. This test consists of two steps to determine the Total Withdraw Processing Time.

B. Step 1:

(1) All devices MUST be synchronized using NTP or some local reference clock.

(2) All variables should be set to basic-test parameters.

(3) The DUT and Helper Node are configured in the same AS, whereas the emulator is configured under a different AS.

(4) Establish BGP adjacency between the DUT and the emulator.

(5) To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

(6) Start the traffic from the emulator towards the DUT targeted at a specific route (e.g., routeA). Initially, no traffic would be observed on the egress interface as routeA is not present on the DUT.

(7) Advertise routeA from the emulator to the DUT.

(8) The traffic targeted towards routeA is received on the egress interface.

(9) Now the Tester sends a request to withdraw routeA to the DUT. TRx(Awith) is also called WdrawTime1(Rt-A).

(10) Record the time when no traffic is observed as determined by the emulator. This is the RouteRemoveTime1(Rt-A).

(11) The difference between the RouteRemoveTime1 and WdrawTime1 is the WdrawConvTime1.

WdrawConvTime1(Rt-A) = RouteRemoveTime1(Rt-A) - WdrawTime1(Rt-A)

C. Step 2:

(1) Continuing from Step 1, re-advertise routeA back to the DUT from the Tester.

(2) The DUT will try to advertise routeA to the Helper Node (this assumes there exists a session between the DUT and Helper Node).

(3) Start the traffic from the emulator towards the Helper Node targeted at a specific route (e.g., routeA). Traffic would be observed on the egress interface after routeA is received by the Helper Node.

WATime=time traffic first flows

(4) Now the Tester sends a request to withdraw routeA to DUT. This is the WdrawTime2(Rt-A).

WAWtime-TRx(Rt-A) = WdrawTime2(Rt-A)

(5) DUT processes the withdraw and sends it to the Helper Node.

(6) Record the time when no traffic is observed as determined by the emulator. This is:

TR-WAW(DUT,RouteA) = RouteRemoveTime2(Rt-A)

(7) Total Withdraw Processing Time is:

TotalWdrawTime(Rt-A) = ((RouteRemoveTime2(Rt-A) - WdrawTime2(Rt-A)) - WdrawConvTime1(Rt-A))

5.7. BGP Path Attribute Change Convergence Time

Objective:

This test measures the convergence time taken by an implementation to service a BGP Path Attribute Change.

Reference Test Setup:

This test uses the setup as shown in Figure 1.

Procedure:

A. This test only applies to Well-Known Mandatory Attributes like origin, AS path, and next hop.

B. In each iteration of the test, only one of these mandatory attributes need to be varied whereas the others remain the same.

C. All devices MUST be synchronized using NTP or some local reference clock.

D. All variables should be set to basic-test parameters.

E. Advertise the same route, routeA, over both adjacencies with preferences so that the Best Egress Interface for the preferred next hop is (Emp1) interface.

F. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

G. Start the traffic from the emulator towards the DUT targeted at the specific route (e.g., routeA). Initially, traffic would be observed on the Best Egress Interface.

H. Now advertise the same route, routeA, on the Next-Best Egress Interface but by varying one of the well-known mandatory attributes to have a preferred value over that interface. We call this Tbetter. The other values need to be the same as what was advertised on the Best-Egress adjacency.

TRx(Path-Change(Rt-A)) = Path Change Event Time(Rt-A)

I. Measure the convergence time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface.

   DUT(Path-Change, Rt-A) = Path-switch time(Rt-A)

   Convergence = Path-switch time(Rt-A) - Path Change Event Time(Rt-A)

J. Stop the offered load and wait for the queues to drain and restart.

K. Repeat the test for various attributes.

5.8. BGP Graceful Restart Convergence Time

Objective:

This test measures the route convergence time taken by an implementation during a Graceful Restart Event as detailed in the terminology document [RFC4098].

Reference Test Setup:

This test uses the setup as shown in Figure 4.

Procedure:

A. It measures the time taken by an implementation to service a BGP Graceful Restart Event and advertise a route.

B. The Helper Nodes are the same model as the DUT and run the same BGP implementation as the DUT.

C. The BGP implementation on the DUT and Helper Node needs to support the BGP Graceful Restart Mechanism [RFC4724].

D. All devices MUST be synchronized using NTP or some local reference clock.

E. All variables are set to basic-test values.

F. The DUT and Helper Node 1 (HLP1) are configured in the same AS, whereas the emulator and Helper Node 2 (HLP2) are configured under different ASs.

G. Establish BGP adjacency between the DUT and Helper Nodes.

H. Establish BGP adjacency between the Helper Node 2 and the emulator.

I. To ensure adjacency establishment, wait for three keepalives to be received from the DUT or a configurable delay before proceeding with the rest of the test.

J. Configure a policy under the BGP on Helper Node 1 to deny routes received from the DUT.

K. Advertise routeA from the emulator to Helper Node 2.

L. Helper Node 2 advertises the route to the DUT and the DUT will try to advertise the route to Helper Node 1, which will be denied.

M. Wait for three keepalives.

N. Start the traffic from the emulator towards the Helper Node 1 targeted at the specific route (e.g., routeA). Initially, no traffic would be observed on the egress interface as routeA is not present.

O. Perform a Graceful Restart Trigger Event on the DUT and note the time. This is the GREventTime.

P. Remove the policy on Helper Node 1.

Q. Record the time when the traffic targeted towards routeA is received on the egress interface.

   This is TRr(DUT, routeA), also called RecTime(Rt-A).

R. The following equation represents the Graceful Restart Convergence Time.

   Graceful Restart Convergence Time(Rt-A) = ((RecTime(Rt-A) - GREventTime) - RIB-IN)

S. It is assumed in this test case that after a switchover is triggered on the DUT, it will not have any cycles to process the BGP Refresh messages. The reason for this assumption is that there is a narrow window of time where after switchover, when we remove the policy from Helper Node 1, implementations might generate Route Refresh automatically and this request might be serviced before the DUT actually switches over and re-establishes BGP adjacencies with the peers.

6. Reporting Format

For each test case, it is recommended that the reporting tables below are completed, and all time values SHOULD be reported with resolution as specified in [RFC4098].

Parameter	Units or Description
Test case	Test case number
Test topology	1, 2, 3, or 4
Parallel links	Number of parallel links
Interface type	Gigabit Ethernet (GigE), Packet over SONET (POS), ATM, other
Convergence Event	Hard reset, soft reset, link failure, or other defined
eBGP sessions	Number of eBGP sessions
iBGP sessions	Number of iBGP sessions
eBGP neighbor	Number of eBGP neighbors
iBGP neighbor	Number of iBGP neighbors
Routes per peer	Number of routes
Total unique routes	Number of routes
Total non-unique routes	Number of routes
IGP configured	IS-IS, OSPF, static, or other
Route mixture	Description of route mixture
Route packing	Number of routes included in an update
Policy configured	Yes, No
SIDR origin authentication [RFC7115]	Yes, No
bgp-sec [BGPsec]	Yes, No

Packet size offered to the DUT	Bytes
Offered load	Packets per second
Packet sampling interval on Tester	Seconds
Forwarding delay threshold	Seconds

Timer values configured on DUT

Interface failure indication delay	Seconds
Hold time	Seconds
MinRouteAdvertisementInterval (MRAI)	Seconds
MinASOriginationInterval (MAOI)	Seconds
Keepalive time	Seconds
ConnectRetry	Seconds

TCP parameters for DUT and Tester

Maximum Segment Size (MSS)	Bytes
Slow start threshold	Bytes
Maximum window size	Bytes

Test Details:

a. If the Offered Load matches a subset of routes, describe how this subset is selected.

b. Describe how the convergence event is applied; does it cause instantaneous traffic loss or not?

c. If there is any policy configured, describe the configured policy.

Complete the table below for the initial convergence event and the reversion convergence event.

Parameter	Unit
Convergence Event	Initial or reversion
Traffic Forwarding Metrics	
Total number of packets offered to the DUT	Number of packets
Total number of packets forwarded by the DUT	Number of packets
Connectivity packet loss	Number of packets
Convergence packet loss	Number of packets
Out-of-order packets	Number of packets
Duplicate packets	Number of packets
Convergence Benchmarks	
Rate-Derived Method [RFC6412]:	
First route convergence time	Seconds
Full convergence time	Seconds
Loss-Derived Method [RFC6412]:	
Loss-Derived convergence time	Seconds
Route-Specific (R-S) Loss-Derived Method:	
Minimum R-S convergence time	Seconds
Maximum R-S convergence time	Seconds
Median R-S convergence time	Seconds
Average R-S convergence time	Seconds
Loss of Connectivity (LoC) Benchmarks	
Loss-Derived Method:	
Loss-Derived loss of connectivity period	Seconds

Route-Specific Loss-Derived Method:	
Minimum LoC period [n]	Array of seconds
Minimum Route LoC period	Seconds
Maximum Route LoC period	Seconds
Median Route LoC period	Seconds
Average Route LoC period	Seconds

7. Security Considerations

Benchmarking activities as described in this memo are limited to technology characterization using controlled stimuli in a laboratory environment, with dedicated address space and the constraints specified in the sections above.

The benchmarking network topology is an independent test setup and MUST NOT be connected to devices that may forward the test traffic into a production network or misroute traffic to the test management network.

Further, benchmarking is performed on a "black-box" basis, relying solely on measurements observable and external to the DUT/SUT.

Special capabilities SHOULD NOT exist in the DUT/SUT specifically for benchmarking purposes. Any implications for network security arising from the DUT/SUT SHOULD be identical in the lab and in production networks.

8. References

8.1. Normative References

[IEEE.802.11] IEEE, "IEEE Standard for Information technology -- Telecommunications and information exchange between systems Local and metropolitan area networks -- Specific requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications", IEEE 802.11-2012, DOI 10.1109/ieeestd.2012.6178212, April 2012, <http://ieeexplore.ieee.org/servlet/opac?punumber=6178209>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, DOI 10.17487/RFC2918, September 2000, <http://www.rfc-editor.org/info/rfc2918>.

[RFC4098] Berkowitz, H., Davies, E., Ed., Hares, S., Krishnaswamy, P., and M. Lepp, "Terminology for Benchmarking BGP Device Convergence in the Control Plane", RFC 4098, DOI 10.17487/RFC4098, June 2005, <http://www.rfc-editor.org/info/rfc4098>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC6412] Poretsky, S., Imhoff, B., and K. Michielsen, "Terminology for Benchmarking Link-State IGP Data-Plane Route Convergence", RFC 6412, DOI 10.17487/RFC6412, November 2011, <http://www.rfc-editor.org/info/rfc6412>.

8.2. Informative References

[BGPsec] Lepinski, M. and K. Sriram, "BGPsec Protocol Specification", Work in Progress, draft-ietf-sidr-bgpsec-protocol-15, March 2016.

[RFC1242] Bradner, S., "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, DOI 10.17487/RFC1242, July 1991, <http://www.rfc-editor.org/info/rfc1242>.

[RFC1983] Malkin, G., Ed., "Internet Users' Glossary", FYI 18, RFC 1983, DOI 10.17487/RFC1983, August 1996, <http://www.rfc-editor.org/info/rfc1983>.

[RFC2285] Mandeville, R., "Benchmarking Terminology for LAN Switching Devices", RFC 2285, DOI 10.17487/RFC2285, February 1998, <http://www.rfc-editor.org/info/rfc2285>.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, DOI 10.17487/RFC2545, March 1999, <http://www.rfc-editor.org/info/rfc2545>.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <http://www.rfc-editor.org/info/rfc4724>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <http://www.rfc-editor.org/info/rfc4760>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <http://www.rfc-editor.org/info/rfc5925>.

[RFC6414] Poretsky, S., Papneja, R., Karthik, J., and S. Vapiwala, "Benchmarking Terminology for Protection Performance", RFC 6414, DOI 10.17487/RFC6414, November 2011, <http://www.rfc-editor.org/info/rfc6414>.

[RFC7115] Bush, R., "Origin Validation Operation Based on the Resource Public Key Infrastructure (RPKI)", BCP 185, RFC 7115, DOI 10.17487/RFC7115, January 2014, <http://www.rfc-editor.org/info/rfc7115>.

Acknowledgements

We would like to thank Anil Tandon, Arvind Pandey, Mohan Nanduri, Jay Karthik, and Eric Brendel for their input and discussions on various sections in the document. We also like to acknowledge Will Liu, Hubert Gee, Semion Lisyansky, and Faisal Shah for their review and feedback on the document.

Authors' Addresses

Rajiv Papneja
Huawei Technologies

Email: rajiv.papneja@huawei.com

Bhavani Parise
Skyport Systems

Email: bparise@skyportsystems.com

Susan Hares
Huawei Technologies

Email: shares@ndzh.com

Dean Lee
IXIA

Email: dlee@ixiacom.com

Ilya Varlashkin
Google

Email: ilya@nobulus.com