Benchmarking Working Group H.Berkowitz, Gett Communications
Internet Draft S.Hares, Nexthop
Document: draft-ietf-bmwg-bgpbas-01.txt A.Retana, Cisco
Expires August 2002 P.Krishnaswamy
M.Lepp, Juniper Networks
E.Davies, Nortel Networks
February 2002

# Benchmarking Methodology for Basic BGP Device Convergence

## Status of this Memo



## Abstract

This draft begins the process of establishing standards for measuring the performance of the BGP routing subsystem in a network. Its initial emphasis is on the control plane of single BGP devices. We do not address forwarding plane performance.

## Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [2].

Table of Contents

## 1. Introduction

This document describes a specific set of tests aimed at characterizing the convergence performance of BGP-4 processes in devices that incorporate BGP functionality as described in [10] and subsequent additions. Such devices include conventional routers, route servers, "layer 3 switches" with external path determination engines (e.g. Ethernet Switch/Routers), and controllers of sub-IP path creation and management. A key objective is to propose methodology that will standardize the conducting and reporting of convergence-related measurements.

The convergence performance of BGP-4 processes is important to the effectiveness and efficiency of IP networks. Poor performance can make the network slow to respond to changes and failures, unnecessary processing when updates are delivered over a longer period than is desirable and consequent misdirection or delay of traffic.

Although both convergence and forwarding are essential to basic device operation, this document does not consider the forwarding performance in the Device Under Test (DUT), for two reasons:

- Forwarding performance is being treated separately: Methodology for forwarding performance is the primary focus in [5] and it is expected to be further covered in work that ensues from the definitions of terminology for Forwarding Information Bases in [9].
- The additional time taken to establish new forwarding behavior after the BGP-4 processes have determined new routes and generated adverts to downstream devices should not affect the overall time for convergence of the network

Further, as convergence characterization is a complex process, we deliberately restrict this document to basic measurements aimed at characterizing BGP convergence in an isolated device receiving inputs on two interfaces and generating outputs on a third interface.

Subsequent versions of this document and other document will explore more complex interconnections, interactions of several devices and the more intricate aspects of convergence measurement, such as the presence of policy processing, simultaneous traffic on the control and data paths within the DUT, and other realistic performance modifiers. Convergence of Interior Gateway Protocols will be considered in separate drafts.

### 1.1 Overview and Roadmap

Measurements of protocols can be classified either as internal or external. Internal measurements are derived from time-stamps applied within the Device Under Test (DUT). External measurements infer the timing of a process in the DUT from measurements made on externally observable phenomena such as transmission of packets to or reception of packets from the DUT by connected test devices. In the case of BGP convergence, the DUT is stimulated by sending BGP UPDATE packets to one or more of its interfaces: If measurements of control plane





behavior are required, the progress of the BGP processes can be gauged by observing the UPDATEs generated and transmitted by the DUT in response to the stimuli as they are received by test receivers connected to the interfaces. An alternative type of external measurement, involving both the data and control planes, is to test for data forwarded to the downstream device that relies upon the new route(s) just installewd in the FIB of the Device Under Test.

We focus here on external measurements based only on control plane phenomena, thus facilitating black box comparisons of the routing subsystem in devices with diverse internal architectures and functions.

If alternative internal measurements were adopted, correlating the DUT's time stamps with those from the rest of the test system is a key problem: The requisite Network Time Protocol (NTP) functionality may not be present and it may be difficult to reach the precision needed for these measurements.

For the purposes of this paper, the external technique is more readily applicable. However, external measurements have their own problems because they include the time to advertise the new route downstream and transmission times for the advertisement within the device under test. If data forwarding were to feature in the measurement methodology it too would include some extraneous latency - that of the forwarding lookup process in the DUT at the minimum. This document deals only with external measurements limited to route propagation.

Characterization of the BGP convergence performance of a device SHOULD take into account all distinct stages and aspects of BGP functionality. This requires that the relevant terms and metrics be as specific as possible. A terminology that meets this objective was presented in "Terinology for Benchmarking External Routing Convergence Measurements" [13].

### 1.2 Scope

This document deals with eBGP convergence of a single deviceDevice Under Test (DUT). It restricts the measurement of convergence to events in the control plane, and does not consider the interactions of convergence and forwarding.

Convergence measurements among multiple iBGP-connected devices in an AS, and Internet-wide convergence measurements, are also outside the scope of this document.

These additional topics are unquestionably of interest, and it is the intention of this document to form a stepping stone toward them

### 1.3 Types of Single-Device Convergence

There are two major types of convergence time that tend to be lumped together in product specifications:





- The time needed for a BGP speaker to build a full table after initialization, or for a particular peering session to rebuild its table after a hard reset (see [12],[13] and section 3.5).
- The time needed for a device to respond to a new announcement or withdrawal. This second time has two subtypes: the time to reconverge a update with a single prefix, and the time to reconverge after receiving a small train of updates. See sections 3.6 and 3.7.

External measurements start with the delivery of a stimulus or the first of several route advertisement stimuli from one or more "upstream" devices (identified as TD1 to Tdn) and end when the BGP process(es) in the device have returned to equilibrium as indicated by all advertisements resulting from the stimuli having been sent to a "downstream" peer (TDrx). In the reference configurations below, external measurements are defined with respect to TDrx as the downstream device.

## 2. Reference Configurations

For tests when the number of peers is not a performance parameter of interest, use the configuration in Figure 1:

```
                              +---------+
```

```
TD1==========|         |==========TDrx
|            |         |
D1           |         |
|            |   DUT   |
TD2==========|         |
             +---------+
```

Figure 1.  Basic Test Configuration.

D1 is a prefix reachable by both TD1 and TD2. Neither TD1 nor TD2 is the originating AS for the announcement of D1.  Stimuli will typically be generated by one or both of TD1 and TD2 according to a timed schedule.  The DUT will propagate consequent adverts towards TDrx where their arrival will be recorded and timed.

More complex peering arrangements will involve up to n Test Routers, as shown in Figure 2.  It is recommended that the Figure 1 configuration always be tested as a baseline, and then additional reports made that show the effect on performance of increasing the number of peers.   Again stimuli would be expected from one or more of the TD1 to TDn.





```
             +---------+
TD1==========|         |==========TDrx
|            |         |
D1           |         |
|            |   DUT   |
TD2==========|         |
             |         |
                  ...
TDn==========+---------+
```

Figure 2. Test Configuration with n Peers.

Interface speeds and types MUST be specified as part of the test report.  At least 100 Mbps is recommended, so media delays are not a significant component of convergence times.

In the absence of other route selection criteria, TD1 SHALL have an IP address that makes it most preferred.

## 3. Basic eBGP tests

All devices in this configuration SHOULD have a policy of ADVERTISE ALL/ACCEPT ALL [6]. Tests with prefix filtering, community-based preferences, authentication, etc., as well as performance under route flap conditions are TBD.

Not all eBGP applications are alike. While the tests in this section are applicable to a wide range of configurations, testers MAY select configurations that are most relevant to the intended product use. Such configurations include:

1. Interprovider peering, characterized by an exchange of customer routes, which, in the case of major providers, may be in the tens of thousands of routes but smaller than the full default-free table.

2. Provider/Subscriber edge peering, where transit service implies the subscriber advertises relatively few routes to the provider but may take, variously, a full set of default-free routes, a limited subset of the full set, or just a default route from the provider.

### 3.1 Connection Conditions

The DUT SHOULD be physically connected to the test devices over a medium sufficiently fast that propagation time is not a significant factor. A medium of at least 100 Mbps is recommended.

Multiple peers MAY be connected to a single physical interface using 802.1q VLANs or another appropriate multiplexing scheme, such as a channelized interface. If so, this MUST be documented in the test results because it may change the arrival times of UPDATEs by





serializing packets which might otherwise arrive in parallel where truly separate, asynchronous interfaces are used.

TCP connections SHOULD NOT use slow start. Any nonstandard initial or maximum window sizes SHALL be indicated in the test report.

### 3.2 Test Streams

Update Packet trains presented to the DUT SHOULD in general be random (see definition of random update train in [13]) with respect to selection of prefixes, prefix length, ordering of prefixes, and time of delivery to DUT. Note that this does not preclude prior, offline creation of sets of update trains with the required randomness that can then be used in running the tests multiple times to determine the reproducibility of results, and for use in comparison tests between products. There may also be circumstances such as are described in Section 3.3 where specific ordering of prefixes may be appropriate for some tests.

The degree of update packing SHALL be specified. When long update trains are being sent, the usual case is that the maximum possible number of prefixes are packed into an UPDATE packet subject to the MTU size of the link over which they are being sent.

### 3.3 Route Mixtures

As shown by measurements of routers in actual deployment, such as are documented in 'The CIDR Report' [14] and similar monitoring projects, both the prefix length distribution and the clustering of prefixes take on characteristic values in the mixture of routes seen.

There are two sets of statistics which exhibit related but different characteristics:
- The distribution in typical default free routing table
- The distribution in the dynamic UPDATEs arriving at a device

The characteristics are reasonably consistent although there are significant bursts of activity from time to time that distort the normal situation.

In creating update trains as test stimuli, these characteristics SHOULD be used to drive:
- The distribution of prefix lengths
- The clustering of prefixes in the total prefix space

The characteristics used should be appropriate for the sort of test in which the update train is to be used.  Initial table load trains should reflect the structure of a default free routing table whereas trains for incremental updates should typically reflect the characteristics of the dynamic UPDATEs.

Future versions of this document may suggest specific profiles for these characteristics, but these remain TBD at present.





### 3.4 Order of Received Updates

For the fairest testing of update trains the order of the prefixes

SHALL include one randomized test.  It Should also include one test sorted by prefix size, and one radix tree implementation.

Assume we have a Adj-RIB-out that consists of

```
          1.0.0.0/8
          2.0.0.0/8
          3.0.0.0/8
          1.1.0.0/16
          2.1.0.0/16
          3.1.0.0/16
          3.2.0.0/16
          1.1.1.0/24
          1.1.2.0/24
          2.1.2.0/24
```

If it were sent in this order, top to bottom, it would be sorted by prefix size and prefix value within size.  A radix tree implementation might like to receive this very much.

But if it were sent out in the following order

```
          1.0.0.0/8
          1.1.0.0/16
          1.1.1.0/24
          1.1.2.0/24
          2.0.0.0/8
          2.1.0.0/16
          2.1.2.0/24
          3.0.0.0/8
          3.1.0.0/16
          3.2.0.0/16
```

It would favor an implementation that orders its routing table as a strict tree, implemented as a linked list.

A 'fair' test train would be

```
          1.0.0.0/8
          2.1.0.0/16
          1.1.0.0/16
          3.0.0.0/8
          1.1.1.0/24
          2.0.0.0/8
          1.1.2.0/24
          3.1.0.0/16
          2.1.2.0/24
          3.2.0.0/16
```

which is random, and equally fair to any particular implementation.

On the other hand, when dealing with a network of devices from a single vendor, in the updates forwarded from a device as a result of a set of stimuli, particularly during a complete table load, the prefixes may be ordered, both within the route packing in a single UPDATE and across the update train which results from the stimuli, in such a way as to be advantageous to downstream devices of the same type: Hence, it MAY be desirable to measure both randomized and 'friendly' orders so as to get a more realistic view of the real world behavior of the devices.  Note that this can only apply to update trains where the individual update packets are delivered close together in time.  If the spacing is too great(greater than the MIN_ADVERT_TIME) the packets will become separate stimuli that are processed individually.

Measurement units:  A metric of randomness,TBD

### 3.5 Initial Convergence

While this is relatively simple to measure, and often is the basis of product specifications, it is operationally far less significant than reconvergence after changes.  A "carrier-grade" device should not initialize often, and the proposed soft reset option reduces the need to rebuild views. The initialization time, therefore, can be amortized over a long period of time and may disappear into the noise when compared to reconvergence (See [12] for details of soft restart standards proposals.  Proprietary implementations already exist).

#### 3.5.1 Single Peer Initial Convergence Time

This basic reference test uses a representatively sized and populated target RIB and the device SHOULD be configured for as basic behavior as possible, thus minimizing variable influences (eg authentication off, filters off, no policy, slow start off).

The test begins with OPEN requests sent from TD1 and TD2 to the DUT. Each Test Router sends a standard routing table with a number, NR, of routes, designated first route (FR) to last route (LR). The value of NR should be reported with every test. There are perfectly valid reasons to test with a small NR, such as testing a device intended as a small multihomed enterprise gateway to the Internet.  In contrast, a large NR would be appropriate for a device intended as a major interprovider gateway.

Conceptually, the test ends when the DUT begins to advertise the last route, LR, in the routing table to TDrx. Since individual implementations may vary in the order in which they construct their outgoing updates i.e., different ordering, packing, etc.), it is possible that the received LR will not necessarily be the last update advertised by the DUT. Note that the routes FR and LR are not in any respect special. They are identified so that progress of the stimulus generator can be monitored and the corresponding events that might be logged in the DUT can be identified.





The test receiver (e.g TDrx) SHOULD record the time at which LR is advertised, but also continue monitoring to see if additional routes are advertised. Initial convergence time is the interval between receipt of FR to the later of two events: the reception of re-advertised LR, or the last update received after the stimulus of LR. LR may be, and often will, be the last update, but that is not guaranteed.

### 3.5.2 Multiple Peers

TBD

## 3.6 Incremental Re-convergence with a Single Peer and a Single Update

For all of these measurements, an update train with a single update is used and the device SHOULD be configured for as basic behavior as possible, thus minimizing variable influences (eg authentication off, filters off, no policy, slow start off).

### 3.6.1 Explicit add of single new route

This test measures the time required to add a single route (D1) newly advertised by a peer. At the start of the test the route does not exist in the DUT's RIB, and hence the new route will not displace a route in the RIB.

The DUT has been initialized, with no path to D1. Measurement time begins when TD1 announces D1 to the DUT.

Measurement time stops when the DUT advertises D1 to TDrx.

### 3.6.2 Sequential withdraw and reannounce of a Single Prefix

The DUT has been initialized and has a path to D1 via TD1, but not via TD2. Simultaneously, TD1 sends TDown(S,TD1) and TD2 announces the new route with Tbest(TD2).

Measurement begins when Tbest is received at the DUT. Measurement time stops when the DUT advertises the new route to D1 to TDrx.

### 3.6.3 Time to Change to Alternate Path after Explicit Withdrawal of a Single Route

The DUT has been initialized and has paths to D1 via both TD1 and TD2. TD1's path is preferred, but TD1 withdraws it with TDown(S,TD1)). Re-convergence occurs when a route from the path(s) advertised by TD2 becomes active.

Measurement time stops when the DUT advertises the new route to D1 via TD2 to TDrx.

## 3.7 Incremental Reconvergence with a Single Peer and Small Packet Trains





For all of these measurements, an update train with a small number of updates is used and the device SHOULD be configured for as basic behavior as possible, thus minimizing variable influences (eg authentication off, filters off, no policy, slow start off). The train SHOULD deliver the updates over a short period of time so that the device may deal with them as a batch rather than reconverging separately for each UPDATE packet received.

### 3.7.1 Explicit add of several new routes

This test measures the time required to add a number of routes (Dm) newly advertised by a peer. At the start of the test these routes do not exist in the DUT's RIB, and hence the new routes will not displace any routes in the RIB.

The DUT has been initialized, with no path to any of Dm. Measurement time begins when TD1 announces D1 to the DUT.

Measurement time stops when the DUT advertises the last of the routes Dm to TDrx.

### 3.7.2 Sequential withdraw and reannounce for a small group of prefixes

The DUT has been initialized and has paths to each of Dm via TD1, but not via TD2. Simultaneously, TD1 sends TDown(S,TD1) and TD2 announces the new route with Tbest(S,TD2).

Measurement begins when Tbest(S.TD1) is received at the DUT. Measurement time stops when the DUT advertises the last of the new routes to Dm to TDrx.

### 3.7.3 Time to Change to Alternate Path after Explicit Withdrawal of several routes

The DUT has been initialized and has paths to each of Dm via both TD1 and TD2. TD1's path is preferred in each case, but TD1 withdraws it with TDown(S,TD1). Re-convergence occurs when routes selected from the path(s) advertised by TD2 become active.

Measurement time stops when the DUT advertises the last of the routes to Dm via TD2 to TDrx.

### 3.8 Incremental Re-convergence with Multiple Peers

The number of routes per BGP peer is an obvious stressor to the convergence process. The number, and relative proportion, of multiple route instances and distinct routes being added or withdrawn by each peer will affect the convergence process, as will the mix of overlapping route instances advertised by teo or more peers.





## 4. Route Flaps

The following tests evaluate convergence when route flap exists.

Let TDF be a device that will generate only flapping routes.

```
             +---------+
TD1==========|         |==========TDrx
|            |         |
D1           |         |
|            |   DUT   |
TD2==========|         |
             |         |
                  ...
TDF==========+---------+
```

Figure 3. Test Diagram with a Router, TDF, flapping.

### 4.1 Flap Isolation Test

TDF will advertise a continuously flapping route i.e. repeated advertisements and withdrawals of a single route sent at intervals equal to the MIN_ADVERT_TIME. Repeat the eBGP convergence tests. The objective is to determine whether one route flapping affects the operation of the device.

If the DUT implements the BGP-4 route flap damping capability described in [4], then the capability SHOULD be disabled for this test. Testing of the route flap damping capability is FFS.

### 4.2 Authentication

Repeat all tests above with MD5 authentication if the DUT implements the capabilities described in [11].

Repeat all tests with Ipsec authentication turned on.

## 5. Acknowledgements

Thanks to Francis Ovenden for review and Abha Ahuja for encouragement. Much appreciation to Jeff Haas, Matt Richardson, and Shane Wright at Nexthop for comments and input.





## 6. References

[1] Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996

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

[3] Ahuja, A., Jahanian, F., Bose, A. and Labovitz, C., "An Experimental Study of Delayed Internet Routing Convergence", RIPE 37 - Routing WG.

[4] Villamizar, C., Chandra, R. and Govindan, R., "BGP Route Flap Damping", RFC 2439, November 1998.

[5] Bradner, S. and McQuaid, J., "Benchmarking Methodology for Network Interconnect Devices",

RFC 2544, March 1999

[6] Alaettinoglu, C., Villamizar, C., Gerich, E., Kessens, D., Meyer, D., Bates, T., Karrenberg, D. and Terpstra, M., "Routing Policy Specification Language (RPSL)", RFC 2622, June 1999.

[7] Ferguson, P. and Senie, D., "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", RFC 2827, May 2000

[8] Chen, E., "Route Refresh Capability for BGP-4", RFC 2928, DATE NEEDED

[9] Trotter, G., "Terminology for Forwarding Information Base(FIB)-based Router Performance", RFC 3222, December 2001

[10] Rekhter, Y. and Li, T., "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, . March 1995.

[11] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

[12] Ramachandra, S., Rekhter, Y., Fernando, R., Scudder, J.G. and Chen, E., "Graceful Restart Mechanism for BGP", draft-ietf-idr-restart-02.txt, January 2002, work in progress.





[13] Berkowitz, H., Hares, S., Retana, A., Krishnaswamy, P. and Lepp, M., "Terminology for Benchmarking External Routing Convergence Measurements", draft-ietf-bmwg-conterm-01.txt, Febtruary 2002, Work in progress

[14] Bates, T., "The CIDR Report", http://www.employees.org/~tbates/cidr-report.html Internet statistics relevant to inter-domain routing updated daily

## 7. Acknowledgments

Thanks to Francis Ovenden for review and Abha Ahuja for encouragement. Much appreciation to Jeff Haas, Matt Richardson, and Shane Wright at Nexthop for comments and input. Debby Stopp and Nick Ambrose contributed the concept of route packing. Thanks to Martin Biddiscombe for trying out the tests.

## 8. Author's Addresses

Howard Berkowitz
Gett Communications
5012 S. 25th St
Arlington VA 22206
Phone: +1 703 998-5819
Fax: +1 703 998-5058
EMail: hcb@gettcomm.com

Elwyn Davies
Nortel Networks
London Road
Harlow, Essex CM17 9NA
UK
Phone: +44-1279-405498
Email: elwynd@nortelnetworks.com

Susan Hares
Nexthop Technologies
517 W. William
Ann Arbor, Mi 48103
Phone:
Email: skh@nexthop.com

Padma Krishnaswamy
Email: kri1@earthlink.net

Marianne Lepp
Juniper Networks
51 Sawyer Road
Waltham, MA 02453
Phone: 617 645 9019
Email: mlepp@juniper.net





Alvaro Retana
Cisco Systems, Inc.
7025 Kit Creek Rd.
Research Triangle Park, NC 27709
Email: aretana@cisco.com

## Appendix A. Representative Scenarios

The following describes sample BGP applications positioned at various points in the network.

### A.1 Default-free interprovider peering

The DUT exchanges 0.3 to 0.5 D with a small number of peers. Typically, devices in this application are limited by bandwidth rather than route processing

### A.2 Interprovider peering with transit

The DUT exchanges 1.3 D routes with a small number of peers.

### A.3 Provider edge device

The DUT has a large number (>10) of eBGP peers.

To 10% of the peers, the DUT advertises 1.3 D.
To 20% of the peers, the DUT advertises 0.3 D.
To 70% of the peers, the DUT advertises default.

50% of the peers advertise an aggregate and a more-specific route to the DUT.
20% of the peers advertise 10 or more routes to the DUT.

30% of the peers advertise a single route to the DUT.

### A.4 Multihomed subscriber edge device

The DUT connects to 2 peers.  It advertises an aggregate and a more-specific to each.