Benchmarking Working Group H.Berkowitz, Gett Communications
Internet Draft S.Hares, Nexthop
Document: draft-ietf-bmwg-conterm-03.txt A.Retana, Cisco
Expires January 2003 P.Krishnaswamy, Consultant
M.Lepp, Juniper Networks
E.Davies, Nortel Networks

July 2002

Terminology for Benchmarking BGP Device Convergence in the Control Plane

Status of this Memo

Copyright Notice

Abstract

This draft establishes terminology to standardize the description of benchmarking methodology for measuring eBGP convergence in the control plane of a single BGP device. Future documents will address iBGP convergence, the initiation of forwarding based on converged control plane information and multiple interacting BGP devices. This terminology is applicable to both IPv4 and IPv6. Illustrative examples of each version are included where relevant.

Table of Contents

1. Introduction

This document defines terminology for use in characterizing the convergence performance of BGP processes in routers or other devices that instantiate BGP functionality (see RFC1771 [1]). It is the first part of a two document series, of which the subsequent document will contain the associated tests and methodology. This terminology is applicable to both IPv4 and IPv6. Illustrative examples of each version are included where relevant.

The following observations underlie the approach adopted in this, and the companion document:

- The principal objective is to derive methodologies to standardize conducting and reporting convergence-related measurements for BGP.
- It is necessary to remove ambiguity from many frequently used terms that arise in the context of such measurements.
- As convergence characterization is a complex process, it is desirable to restrict the initial focus in this set of documents to specifying how to take basic control plane measurements as a first step to characterizing BGP convergence.

For path vector protocols, such as BGP, the primary initial focus will therefore be on network and system control-plane activity consisting of the arrival, processing, and propagation of routing information
Subsequent drafts will explore the more intricate aspects of convergence measurement, such as the impacts of 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 also be considered in separate drafts.

1.1 Overview and Roadmap

Characterizations of the BGP convergence performance of a device must take into account all distinct stages and aspects of BGP functionality. This requires that the relevant terms and metrics be as specifically defined as possible. Such definition is the goal of this document.

The necessary definitions are classified into two separate categories:

- Descriptions of the constituent elements of a network or a router that is undergoing convergence
- Descriptions of factors that impact convergence processes

1.2 Definition Format

The definition format is equivalent to that defined in [3], and is

repeated here for convenience:

X.x Term to be defined. (e.g., Latency)

Definition:

One or more sentences forming the body of the definition.

Discussion:
A brief discussion of the term, its application and any restrictions that there might be on measurement procedures.

Measurement units:
The units used to report measurements of this term, if applicable.

Issues:
List of issues or conditions that could affect this term.

See Also:
List of related terms that are relevant to the definition or discussion of this term.

2. Constituent elements of a router or network of routers.

Many terms included in this list of definitions were originally described in previous standards or papers. They are included here because of their pertinence to this discussion. Where relevant, reference is made to these sources. An effort has been made to keep this list complete with regard to the necessary concepts without over definition.

2.1 BGP Instance or Device

Definition:
A BGP instance is a process with a single Loc-RIB that runs on a BGP device.

Discussion:
We have chosen to use "device" as the general case, to deal with the understood [e.g. [9]] and yet-to-be-invented cases where the control processing may be separate from forwarding [12]. A BGP device may be a traditional router, a route server, a BGP-aware traffic steering device, a device using BGP to exchange topology information with a GMPLS environment, etc. A device such as a route server, for example, never forwards traffic, so forwarding-based measurements would be meaningless for it.

Measurement units: N/A

Issues:

See Also:

2.2 BGP Peer

Definition:

A BGP peer is another BGP instance to which the Device Under Test (DUT) has established a TCP connection over which a BGP session is active. In the test scenarios in the methodology discussion that will follow this draft, peers send BGP advertisements to the DUT and receive DUT-originated advertisements.

Discussion:

This is a protocol-specific definition, not to be confused with another frequent usage, which refers to the business/economic definition for the exchange of routes without financial compensation.

It is worth noting that a BGP peer, by this definition is associated with a BGP peering session, and there may be more than one such active session on a router or on a tester. The peering sessions referred to here may exist between various classes of BGP routers (see section 2.3).

Measurement units: number of BGP peers

Issues:

See Also:

2.3 Default Route, Default Free Table, and Full Table

An individual router's routing table may not necessarily contain a default route. Not having a default route, however, is not synonymous with having a full default-free table(DFT).
It should be noted that the references to number of routes in this section are to routes installed in the loc-RIB, not route instances, and that the total number of route instances may be 4 to 10 times the number of routes.

The actual path setup and forwarding of MPLS speaking routers are outside the scope of this document. A device that computes BGP routes, which a sub-IP device can use to set up paths, has its BGP aspects within scope.

2.3.1 Default Route

Definition:

A Default Route is a route entry that can match any prefix. If a router does not have a route for a particular packet's destination address, it forwards this packet to the next hop in the default route entry, provided its Forwarding Table (Forwarding Information Base (FIB) contains one. The notation for a default route for IPv4 is 0.0.0.0/0 and for IPv6 it is 0:0:0:0:0:0:0:0 or ::/0.

Discussion:

Measurement units: N.A.

Issues:

See Also: default free routing table, route, route instance

2.3.2 Default Free Routing Table

Definition:

A default free routing table has no default routes and is typically seen in routers in the core or top tier of routers in the network.

Discussion:

The term originates from the concept that routers at the core or top tier of the Internet will not be configured with a default route (Notation in IPv4 0.0.0.0/0 and in IPv6 0:0:0:0:0:0:0:0 or ::/0). Thus they will forward every prefix to a specific next hop based on the longest match on the IP addresses.

Default free routing table size is commonly used as an indicator of the magnitude of reachable Internet address space. However, default free routing tables may also include routes internal to the router's AS.

Measurements: The number of routes

See Also: Full Default Free, Default Route

2.3.3 Full Default Free Table

Definition:

A full default free table is a set of BGP routes generally accepted to be the complete set of BGP routes collectively announced by the complete set of autonomous systems making up the public Internet. Due to the dynamic nature of the Internet, the exact size and composition of this table may vary slightly depending where and when it is observed.

Discussion:

Several investigators ([17],[18],[19]) measure this on a daily and/or weekly basis; June 2001 measurements put the table at approximately 105,000 routes, growing exponentially.

It is generally accepted that a full table, in this usage, does not contain the infrastructure routes or individual sub-aggregates of routes that are otherwise aggregated by the provider before announcement to other autonomous systems.

Measurement Units: number of routes

Issues:

See Also: Routes, Route Instances, Default Route

2.3.4 Full Provider Internal Table

Definition:

A full provider internal table is a superset of the full routing table that contains infrastructure and non-aggregated routes.

Discussion:

Experience has shown that this table can contain 1.3 to 1.5 times the number of routes in the externally visible full table. Tables of this size, therefore, are a real-world requirement for key internal provider routers.

Measurement Units: number of routes

Issues:

See Also: Routes, Route Instances, Default Route

2.4 Classes of BGP-Speaking Routers

A given router may perform more than one of the following functions, based on its logical location in the network.

2.4.1 Provider Edge Router

Definition:
A provider edge router is a router at the edge of a provider's network, configured to speak BGP, which peers with a BGP speaking router operated by the end-user. The traffic that transits this router may be destined to, or originate from non-contiguous autonomous systems.

Discussion:
Such a router will always speak eBGP and may speak iBGP.

Measurement units:

Issues:

See Also:

2.4.2 Subscriber Edge Router

Definition:
A subscriber edge router is a BGP-speaking router belonging to an end user organization that may be multi-homed, and which carries traffic only to and from that end user AS.

Discussion:
Such a router will always speak eBGP and may speak iBGP.

Measurement units:

Issues:

See Also:

2.4.3 Inter-provider Border Router

Definition:
An inter-provider border router is a BGP speaking router which maintains BGP sessions with another BGP speaking router in another provider AS. Traffic transiting this router may be directed to or from another AS that has no direct connectivity with this provider's AS.

Discussion:

Such a router will always speak eBGP and may speak iBGP.

Measurement units:

Issues:

See Also:

2.4.4 Intra-provider Core Router

Definition:

An intra-provider core router is a provider router speaking iBGP to the provider's edge routers, other intra-provider core routers, or the provider's inter-provider border routers.

Discussion:

Such a router will always speak iBGP and may speak eBGP.

Measurement units:

Issues:

MPLS speaking routers are outside the scope of this document. It is entirely likely, however, that core Label Switched Routers, especially in the P router role of RFC 2547 [10], may contain little or no BGP information.

See Also:

3. Routing Data Structures

3.1 Routing Information Base (RIB)

The RIB collectively consists of a set of logically (not necessarily literally) distinct databases, each of which is enumerated below. The RIB contains all destination prefixes to which the router may forward, and one or more currently reachable next hop addresses for them.

Routes included in this set potentially have been selected from several sources of information, including hardware status, interior routing protocols, and exterior routing protocols. RFC 1812 contains a basic set of route selection criteria relevant in an all-source context. Many implementations impose additional criteria. A common implementation-specific criterion is the preference given to different routing information sources.

3.1.1 Adj-RIB-In and Adj-RIB-Out

Definition:

Adj-RIB-In and Adj-RIB-Out are "views" of routing information from the perspective of individual peer routers.

The Adj-RIB-In contains information advertised to the DUT by a specific peer. The Adj-RIB-Out contains the information the DUT will advertise to the peer.

See RFC 1771[1].

Discussion:

Issues:

Measurement Units: Number of route instances

See Also: Route, BGP Route, Route Instance, Loc-RIB, FIB

3.1.2 Loc-RIB

Definition:

The Loc-RIB contains the set of best routes selected from the various Adj-RIBs, after applying local policies and the BGP route selection algorithm.

Discussion:

The separation implied between the various RIBs is logical. It does not necessarily follow that these RIBs are distinct and separate entities in any given implementation.

Types of routes can include internal BGP, external BGP,interface, static and IGP routes.

Issues:

Measurement Units: Number of route instances.

See Also: Route, BGP Route, Route Instance, Adj-RIB-in, Adj-RIB-out,FIB

3.2 Policy

Definition:

Policy is "the ability to define conditions for accepting, rejecting, and modifying routes received in advertisements"[9].

Discussion:

RFC 1771 [1] further constrains policy to be within the hop-by-hop routing paradigm. Policy is implemented using filters and associated policy actions. Many AS's formulate and document their policies using the Routing Policy Specification Language (RPSL) [6] and then automatically generate configurations for the BGP processes in their routers from the RPSL specifications.

Measurement Units: Number of policies; length of policies

Issues:

See Also: Policy Information Base.

3.3 Policy Information Base

Definition:

A policy information base is the set of incoming and outgoing policies.

Discussion:

All references to the phase of the BGP selection process below are made with respect to RFC 1771 [1] definition of these phases.

Incoming policies are applied in Phase 1 of the BGP selection process [1] to the Adj-RIB-In routes to set the metric for the Phase 2 decision process. Outgoing Policies are applied in Phase 3 of the BGP process to the Adj-RIB-Out routes preceding route (prefix and path attribute tuple) announcements to a specific peer.

Policies in the Policy Information Base have matching and action conditions. Common information to match include route prefixes, AS paths, communities, etc. The action on match may be to drop the update and not pass it to the Loc-RIB, or to modify the update in some way, such as changing local preference (on input) or MED (on output), adding or deleting communities, prepending the current AS in the AS path, etc.

The amount of policy processing (both in terms of route maps and filter/access lists) will impact the convergence time and properties of the distributed BGP algorithm. The amount of policy processing may vary from a simple policy which accepts all routes and sends all routes to complex policy with a substantial fraction of the prefixes being filtered by filter/access lists.

Measurement Units: Number and length of policies

Issues:

See Also:

3.4 Forwarding Information Base (FIB)

Definition:

As according to the definition in Appendix B of [4]: "The table containing the information necessary to forward IP Datagrams is called the Forwarding Information Base. At minimum, this contains the interface identifier and next hop information for each reachable destination network prefix."

Discussion:

The forwarding information base describes a database indexing network prefixes versus router port identifiers.

The forwarding information base is distinct from the "routing table" (the Routing Information Base or RIB), which holds all routing information received from routing peers. The Forwarding Information Base is generated from the RIB. For the purposes of this document, the FIB is effectively the subset of the RIB used by the forwarding plane to make per-packet forwarding decisions.

Most current implementations have full, non-cached FIBs per router interface. All the route computation and convergence occurs before entries are downloaded into a FIB.

Measurement units: N.A.

Issues:

See Also: Route, RIB

4. Components and characteristics of Routing information

4.1 (Network) Prefix

Definition:

"A network prefix is . . . a contiguous set of bits at the more significant end of the address that defines a set of systems; host numbers select among those systems."
(This definition is taken directly from section 2.2.5.2, "Classless Inter Domain Routing (CIDR)", in [3].)

Discussion:

In the CIDR context, the network prefix is the network component of an IP address.

Measurement Units: N.A.

Issues:

See Also

4.2 Network Prefix Length

Definition:

The network prefix length is the number of bits used to define the network prefix.

Discussion:

A common alternative to using a bit-wise mask to communicate this component is the use of "slash (/) notation." Slash notation binds the notion of network prefix length (see 4.2) in bits to an IP address. E.g., 141.184.128.0/17 indicates the network component of this IPv4 address is 17 bits wide. Similar notation is used for IPv6 network prefixes e.g. :FF02:20::/24

When referring to groups of addresses, the network prefix length is often used as a means of describing groups of addresses as an equivalence class. For example, 'one hundred /16 addresses' refers to 100 addresses whose network prefix length is 16 bits.

Measurement units: bits

Issues:

See Also: network prefix

4.3 Route

Definition:

In general, a 'route' is the n-tuple <prefix, nexthop[, other non-routing protocol attributes]> A route is not end-to-end, but is defined with respect to a specific next hop that will move traffic closer to the destination defined by the prefix. In this usage, a route is the basic unit of information about a target destination distilled from routing protocols.

Discussion:

This term refers to the concept of a route common to all routing protocols. With reference to the definition above, typical non-routing-protocol attributes would be associated with diffserv or traffic engineering.

Measurement Units: N.A.

Issues: None.

See Also: BGP route

4.4 BGP Route

Definition:

A BGP route is an n-tuple <prefix, nexthop, ASpath [, other BGP attributes]>.

Discussion:

BGP Attributes, such as Nexthop or AS path are defined in RFC 1771[1], where they are known as Path Attributes, and are the qualifying data that accompanies the network prefixes in a BGP route UPDATE message. (An UPDATE message may contain multiple prefixes that share a common set of attributes).

From RFC 1771: " For purposes of this protocol a route is defined as a unit of information that pairs a destination with the attributes of a path to that destination... A variable length sequence of path attributes is present in every UPDATE. Each path attribute is a triple <attribute type, attribute length, attribute value> of variable length."

Measurement Units: N.A.

Issues:

See Also: Route, prefix, Adj-RIB-in, NLRI.

4.5 BGP Route Attributes and BGP Timers

The definitions in this section refer to items that are originally defined in RFC 1771 [1] and are repeated here for convenience, and to allow for some discussion beyond the definitions in RFC 1771.

4.5.1 Network Level Reachability Information (NLRI)

Definition:

The NRLI consists of one or more network prefixes that share all other BGP path attributes and are distributed in the update portion (as opposed to the unfeasible routes portion) of a BGP UPDATE message.

Discussion:

Each prefix in the NLRI is combined with the (common) path attributes in the UPDATE message to form a BGP route. The NLRI encapsulates a set of destinations to which packets can be routed (from this point in the network) along a common route described by the path attributes.

Measurement Units: N.A.

Issues:

See Also: Route Packing, Network Prefix, BGP Route, NLRI

4.5.2 MinRouteAdvertisementInterval (MRAI)

Definition:

(Paraphrased from 1771[1]) The MRAI timer determines the minimum time between advertisements of routes to a particular destination (prefix) from a single BGP device. The timer is applied on a pre-prefix basis, although the timer is set on a per BGP device basis.

Discussion:

Given that a BGP instance may manage in excess of 100,000 routes, RFC 1771 allows for a degree of optimization in order to limit the number of timers needed. The MRAI does not apply to routes received from BGP speakers in the same AS or to explicit withdrawals.

RFC 1771 also recommends that random jitter is applied to MRAI in an attempt to avoid synchronization effects between the BGP instances in a network.

In this document we define RIB convergence by measuring the time an NLRI is advertised to the DUT to the time it is advertised from the DUT. Clearly any delay inserted by

the MRAI will have a significant effect on this measurement.

Measurement Units: seconds.

Issues:

See Also: NLRI, BGP route

4.5.3 MinASOriginationInterval (MAOI)

Definition:
The MAOI specifies the minimum interval between advertisements of locally originated routes from this BGP instance.

Discussion:
Random jitter is applied to MAOI in an attempt to avoid synchronization effects between BGP instances in a network.

Measurement Units: seconds

Issues:

See Also: MRAI, BGP route

4.6 Route Instance

Definition:
A route instance is a single occurrence of a route sent by a BGP Peer for a particular prefix. When a router has multiple peers from which it accepts routes, routes to the same prefix may be received from several peers. This is then an example of multiple route instances.

Discussion:
Each route instance is associated with a specific peer. The BGP selection algorithm may reject a specific route instance due to local policy.

Measurement Units: Number of route instances

Issues:
The number of route instances in the Adj-RIB-in bases will vary based on the function to be performed by a router. An inter-provider router, located in the default free zone will likely receive more route instances than a provider edge router, located closer to the end-users of the network.

See Also:

4.7 Active Route

Definition:
Route for which there is a FIB entry corresponding to a RIB entry.

Discussion:

Measurement Units: Number of routes.

Issues:

See also: RIB.

4.8 Unique Route

Definition:
A unique route is a prefix for which there is just one route instance across all Adj-Ribs-In.

Discussion:

Measurement Units: N.A.

Issues:

See Also: route, route instance

4.9 Non-Unique Route

Definition:
A Non-unique route is a prefix for which there is at least one other route in a set including more than one Adj-RIB-in.

Discussion:

Measurement Units: N.A.

Issues:

See Also: route, route instance, unique active route.

4.10BGP UPDATE message

Definition:
An UPDATE message is an advertisement of a single NLRI, possibly containing multiple prefixes, and multiple withdrawals of unfeasible routes. See RFC 1771 ([1]) for details.

Discussion:

Measurement Units: N.A.

Issues:

4.11Characterization of sets of update messages

This section contains a sequence of definitions that build up to the definition of an Update Train, a concept originally introduced by Jain and Routhier [11]. This is a formalization of the sort of test stimulus that is expected as input to a DUT running BGP. This data could be a well-characterized, ordered and timed set of hand-crafted BGP UPDATE packets. It could just as well be a set of BGP UPDATE packets that have been captured from a live router.

Characterization of route mixtures and Update Trains is an open area of research. The particular question of interest for this work is the identification of suitable Update Trains, modeled or taken from live traces that reflect realistic sequences of UPDATEs and their contents.

4.11.1 Route Packing

Definition:

Route packing is the number of route prefixes accommodated in a single Routing Protocol UPDATE Message either as updates (additions or modifications) or withdrawals.

Discussion:

In general, a routing protocol update may contain more than one prefix. In BGP, a single UPDATE may contain two sets of multiple network prefixes: one set of additions and updates with identical attributes (the NLRI) and one set of unfeasible routes to be withdrawn.

Measurement Units:

Number of prefixes.

Issues:

See Also: route, BGP route, route instance, update train, NLRI.

4.11.2 Route Mixture

Definition:

A collection of routes such as an NLRI, a set of UPDATE messages or a RIB.

Discussion:

A route mixture is the input data for the benchmark. The particular route mixture used as input must be selected to suit the question being asked of the benchmark.

Data containing simple route mixtures, such as 100,000 /32 routes might test the performance limits of the BGP device.

Using live data, or input that simulates live data, should improve understanding of how the BGP device will operate in a live network. The data for this kind of test must be route mixtures that model the patterns of arriving control traffic in the live Internet.

To accomplish that kind of modeling it is necessary to identify the key parameters that characterize a live Internet route mixture. The parameters and how they interact is an open research problem. However, we identify the following as affecting the route mixture:

- Path length distribution
- Attribute distribution
- Prefix distribution
- Packet packing
- Probability density function of inter-arrival times of UPDATES

Each of the items above is more complex than a singlenumber. For example, one could consider the distribution of prefixes by AS or distribution of prefixes by length.

Measurement Units: Probability density functions

Issues:

See Also: NLRI, RIB.

4.11.3 Update Train

Definition:

An update train is a set of Routing Protocol UPDATE messages sent by a router to a BGP peer.

Discussion:

The arrival pattern of UPDATEs can be influenced by many things, including TCP parameters, hold-down timers, BGP header processing, a peer coming up or multiple peers sending at the same time. Network conditions such as a local or remote peer flapping a link can also affect the arrival pattern.

Measurement units:

Probability density function for the inter-arrival times of UPDATE packets in the train.

Issues:

Characterizing the profiles of real world UPDATE trains is a matter for future research. In order to generate

realistic UPDATE trains as test stimuli a formal mathematical scheme or a proven heuristic is needed to drive the selection of prefixes. Whatever mechanism is

selected it must generate Update trains that have similar characteristics to those measured from live routers.

See Also: Route Mixture, MRAI, MAOI

4.11.4 Randomness in Update Trains

As we have seen from the previous sections, an update train used as a test stimulus has a considerable number of parameters that can be varied, to a greater or lesser extent, randomly and independently.

A random Update Train will contain:

- A route mixture randomized across
  - NLRIs
  - updates and withdrawals
  - prefixes
  - inter-arrival times of the UPDATEs

  and possibly across other variables.

This is intended to simulate the unpredictable asynchronous nature of the network, whereby UPDATE packets may have arbitrary contents and be delivered at random times.

It is important that the data set be randomized sufficiently to avoid favoring one vendor's implementation over another's.
Specifically, the distribution of prefixes could be structured to favor the internal organization of the routes in a particular vendor's databases. This is to be avoided.

4.12Route Flap

Definition:

RIPE 210 [7] define a route flap as "the announcement and withdrawal of prefixes." For our purposes we define a route flap as the rapid withdrawal/announcement or announcement/withdrawal of a prefix in the Adj-RIB-in. A route flap is not a problem until a route is flapped several times in close succession. This causes negative repercussions throughout the internet.

Discussion:
Route flapping can be considered a special and pathological case of update trains. A practical interpretation of what may be considered excessively rapid is the RIPE recommendation of "four flaps in a row". See Section 6.1.5 on flap damping for further discussion.

Measurement units: Flapping events per unit time.

Issues:
Specific Flap events can be found in Section 5.1Route Change Events. A bench-marker should use a mixture of different route change events in testing.

See Also: Route change events, flap damping, packet train

5. Route Changes and Convergence

The following two definitions are central to the benchmarking of external routing convergence, and so are singled out for more extensive discussion.

5.1 Route Change Events

A taxonomy characterizing routing information changes seen in operational networks is proposed in [4] as well as Labovitz et al[5]. These papers describe BGP protocol-centric events, and event sequences in the course of an analysis of network behavior. The terminology in the two papers categorizes similar but slightly different behaviors with some overlap. We would like to apply these taxonomies to categorize the tests under definition where possible, because these tests must tie in to phenomena that arise in actual networks. We avail ourselves of, or may extend, this terminology as necessary for this purpose.

A route can be changed implicitly by replacing it with another route or explicitly by withdrawal followed by the introduction of a new route. In either case the change may be an actual change, no change, or a duplicate. The notation and definition of individual categorizable route change events is adopted from [5] and given below.

a) AADiff: Implicit withdrawal of a route and replacement by a route different in some path attribute.
b) AADup: Implicit withdrawal of a route and replacement by route that is identical in all path attributes.
c) WADiff: Explicit withdrawal of a route and replacement by a different route.
d) WADup: Explicit withdrawal of a route and replacement by a route that is identical in all path attributes.

To apply this taxonomy in the benchmarking context, we need both terms to describe the sequence of events from the update train perspective, as listed above, and event indications in the time

domain so as to be able to measure activity from the perspective of the DUT. With this in mind, we incorporate and extend the definitions of [5] to the following:

a) Tup (TDx): Route advertised to the DUT by Test Device x
b) Tdown(TDx): Route being withdrawn by Device x
c) Tupinit(TDx): The initial announcement of a route to a unique prefix
d) TWF(TDx): Route fail over after an explicit withdrawal.

But we need to take this a step further. Each of these events can involve a single route, a "short" packet train, or a "full" routing table. We further extend the notation to indicate how many routes are conveyed by the events above:

a) Tup(1,TDx) means Device x sends 1 route
b) Tup(S,TDx) means Device x sends a train, S, of routes
c) Tup(DFT,TDx) means Device x sends an approximation of a full default-free table.

The basic criterion for selecting a "better" route is the final tiebreaker defined in RFC1771, the router ID. As a consequence, this memorandum uses the following descriptor events, which are routes selected by the BGP selection process rather than simple updates:

a) Tbest -- The current best path.
b) Tbetter -- Advertise a path that is better than Tbest.
c) Tworse -- Advertise a path that is worse than Tbest.

5.2 Device Convergence in the Control Plane

Definition

A routing device is said to have converged at the point in time when the DUT has performed all actions in the control plane needed to react to changes in topology in the context of the test condition.

Discussion:

For example, when considering BGP convergence, a change that alters the best route instance for a single prefix at a router would be deemed to have converged when this route is advertised to its downstream peers. Similarly, OSPF convergence concludes when SPF calculations have been performed and the required link states advertised onwards.

The convergence process, in general, can be subdivided into three distinct phases:

- convergence across the entire Internet,
- convergence within an Autonomous System,
- convergence with respect to a single device.

Convergence with respect to a single device can be
- convergence with regard to data forwarding process(es)
- convergence with regard to the routing process(es), the focus of this document.

It is the latter, convergence with regard to the routing process, that we describe in this and the methodology documents.

Because we are trying to benchmark the routing protocol performance which is only a part of the device overall, this definition is intended (so far as is possible) to exclude any additional time such as is needed to download and install the forwarding information base in the data plane. This definition should be usable for different families of protocols.

It is of key importance to benchmark the performance of each phase of convergence separately before proceeding to a composite characterization of routing convergence, where implementation-specific dependencies are allowed to interact.

The time resolution needed to measure the device convergence depends to some extent on the types of the interfaces on the router. For modern routers with gigabit or faster interfaces, an individual UPDATE may be processed and re-advertised in very much less than a millisecond so that time measurements must be made to a resolution of hundreds to tens of microseconds or better.

Measurement Units:
Time period.

Issues:

See Also:

6. BGP Operation Events

The BGP speaker process(es) in a device restarts completely, for example, because of operator intervention or a power failure, or fails partially because a TCP session has terminated for a particular link. Until recently the BGP process would have to re-advertise all relevant routes on reestablished links potentially triggering updates across the network. Recent work is focused on limiting the volume of updates due to operational events and the amount of processing resulting from these events: This work includes soft refresh[12], a graceful restart mechanism [13] and cooperative route filtering (e.g.[14]).

6.1 Hard reset

Definition:
An event which triggers a complete re-initialization of the routing tables on one or more BGP sessions, resulting in exchange of a full routing table on one or more links to the router.

Discussion:

Measurement Units: N/A

Issues:

See Also:

6.2 Soft reset

Definition:
An event which results in a complete or partial restart of the BGP session(s) on a BGP device, but which avoids the exchange of a full table by maintaining state across the restart.

Discussion:

Measurement Units: N/A

Issues:

See Also:

7. Factors that impact the performance of the convergence process

While this is not a complete list, all of the items discussed below have a significant affect on BGP convergence. Not all of them can be addressed in the baseline measurements described in this document.

7.1 General factors affecting device convergence

These factors are conditions of testing external to the router Device Under Test (DUT).

7.1.1 Number of peers

As the number of peers increases, the BGP route selection algorithm is increasingly exercised. In addition, the phasing and frequency of updates from the various peers will have an increasingly marked effect on the convergence process on a router as the number of peers grows. Increasing the number of peers also increases the processing

workload for TCP and BGP keepalives.

7.1.2 Number of routes per peer

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, and IGP routes.

7.1.3 Policy processing/reconfiguration

The number of routes and attributes being filtered, and set, as a fraction of the target route table size is another parameter that will affect BGP convergence.

Extreme examples are
- Minimal Policy: receive all, send all,
- Extensive policy: up to 100% of the total routes have applicable policy.

7.1.4 Interactions with other protocols.

There are interactions in the form of precedence, synchronization, duplication and the addition of timers, and route selection criteria. Ultimately, understanding BGP4 convergence must include understanding of the interactions with both the IGPs and the protocols associated with the physical media, such as Ethernet, SONET, DWDM.

7.1.5 Flap Damping

A router can use flap damping to respond to route flapping. Use of flap damping is not mandatory, so the decision to enable the feature, and to change parameters associated with it, can be considered a matter of routing policy.

The timers are defined by RFC 2439 [2] and discussed in RIPE-229 [7]. If this feature is in effect, it requires that the device keep additional state to carry out the damping, which can have a direct impact on the control plane due to increased processing. In addition, flap damping may delay the arrival of real changes in a route, and affect convergence times

7.1.6 Churn

In theory, a BGP device could receive a set of updates that completely defined the Internet, and could remain in a steady state, only sending appropriate keepalives. In practice, the Internet will always be changing.

Churn refers to control plane processor activity caused by announcements received and sent by the router. It does not include

keepalives and TCP processing.

Churn is caused by both normal and pathological events. For example, if an interface of the local router goes down and the associated

prefix is withdrawn, that withdrawal is a normal activity, although it contributes to churn. If the local device receives a withdrawal of a route it already advertises, or an announcement of a route it did not previously know, and re-advertises this information, again these are normal constituents of churn. Routine updates can range from single announcement or withdrawals, to announcements of an entire default-free table. The latter is completely reasonable as an initialization condition.

Flapping routes are a pathological contributor to churn, as is MED oscillation [16]. The goal of flap damping is to reduce the contribution of flapping to churn.

The effect of churn on overall convergence depends on the processing power available to the control plane, and whether the same processor(s) are used for forwarding and for control.

7.2 Implementation-specific and other factors affecting BGP convergence

These factors are conditions of testing internal to the Device Under Test (DUT), although they may affect its interactions with test devices.

7.2.1 Forwarded traffic

The presence of actual traffic in the device may stress the control path in some fashion if both the offered load due to data and the control traffic (FIB updates and downloads as a consequence of flaps)are excessive. The addition of data traffic presents a more accurate reflection of realistic operating scenarios than if only control traffic is present.

7.2.2 Timers

Settings of delay and hold-down timers at the link level as well as for BGP4, can introduce or ameliorate delays. As part of a test report, all relevant timers should be reported if they use non-default value.

7.2.3 TCP parameters underlying BGP transport

Since all BGP traffic and interactions occur over TCP, all relevant parameters characterizing the TCP sessions should be provided: eg Slow start, max window size, maximum segment size, or timers.

7.2.4 Authentication

Authentication in BGP is currently done using the TCP MD5 Signature Option [8]. The processing of the MD5 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.

8. Security Considerations

The document explicitly considers authentication as a performance-affecting feature, but does not consider the overall security of the routing system.

9. References

Normative

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

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

[3] Baker, F.,"Requirements for IP Version 4 Routers", RFC 1812. June 1995.

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

[5] Labovitz, C., Malan, G.R. and Jahanian, F., "Origins of Internet Routing Instability," Infocom 99.

[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] Barber, T., Doran, S., Karrenberg, D., Panigl, C., Schmitz, J., "RIPE Routing-WG Recommendation for coordinated route-flap damping parameters", RIPE 210.

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

[9] Juniper Networks,"Junos(tm) Internet Software Configuration Guide Routing and Routing Protocols, Release 4.2" http://www.juniper.net/techpubs/software/junos42/swconfig-routing42/html/glossary.html#1013039. September 2000 (and other releases).

[10] Rosen, E. and Rekhter, Y., "BGP/MPLS VPNs", RFC 2547, March 1999.

[11] Jain, R. and Routhier, S.A., "Packet trains -- measurement and a new model for computer network traffic," IEEE Journal on Selected Areas in Communication, 4(6)September 1986.

Illustrative

[12] Chen, E., "Route Refresh for BGP-4", RFC 2918, September 2000.

[13] 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.

[14] Chen, E. and Rekhter, Y, "Cooperative Route Filtering Capability for BGP-4", draft-ietf-idr-route-filter-05.txt, January 2002, work in progress.

[15] T. Anderson et al. "Requirements for Separation of IP Control and Forwarding", draft-ietf-forces-requirements-02.txt, February 2002, work in progress.

[16] McPherson, Gill, Walton, Retana, "BGP Persistent Route Oscillation Condition", <draft-ietf-idr-route-oscillation-01.txt>, February 2002, work In progress.

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

[18] Smith, P. (designer), APNIC Routing Table Statistics, http://www.apnic.net/stats/bgp/, Statistics derived from a daily analysis of a core router in Japan.

[19] Huston, G., Telstra BGP table statistics, http://www.telstra.net/ops/bgp/index.html, Statistics derived daily from the BGP tables of Telstra and other AS's routers.

For Internet Draft consistency purposes only

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

10.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.

11. 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

Full Copyright Statement