Network Working Group
Request for Comments: 4098
Category: Informational

H. Berkowitz
Gett Communications & CCI Training
E. Davies, Ed.
Folly Consulting
S. Hares
Nexthop Technologies
P. Krishnaswamy
SAIC
M. Lepp
Consultant

June 2005

Terminology for Benchmarking BGP Device Convergence in the Control Plane

Status of This Memo

Copyright Notice

Abstract

This document 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 'A Border Gateway Protocol 4 (BGP-4)' [RFC1771], referred to as RFC 1771 in the remainder of the document.) 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. However, this document is primarily targeted for BGP-4 in IPv4 networks. IPv6 will require the use of MP-BGP [RFC2858], as described in RFC 2545 [RFC2545], but this document will not address terminology or issues specific to these extensions of BGP-4. Also terminology and issues specific to the extensions of BGP that support VPNs as described in RFC 2547 [RFC2547] are out of scope for this document.

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

- The principal objective is to derive methodologies that 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 these 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 in characterizing BGP convergence.

For path-vector protocols, such as BGP, the primary initial focus will therefore be on network and system control-plane [RFC3654] activity consisting of the arrival, processing, and propagation of routing information.

We note that for testing purposes, all optional parameters SHOULD be turned off. All variable parameters SHOULD be at their default setting unless the test specifies otherwise.

Subsequent documents will explore the more intricate aspects of convergence measurement, such as the impacts of the presence of Multiprotocol Extensions for BGP-4, policy processing, simultaneous traffic on the control and data paths within the Device Under Test (DUT), and other realistic performance modifiers. Convergence of Interior Gateway Protocols (IGPs) will also be considered in separate documents.

1.1. Overview and Road Map

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 separate categories:

- Components and characteristics of routing information
- Routing data structures and route categories
- Descriptions of the constituent elements of a network or a router that is undergoing convergence

- Characterization of sets of update messages, types of route-change events, as well as some events specific to BGP operation

- Descriptions of factors that impact the performance of convergence processes

1.2. Definition Format

The definition format is equivalent to that defined in 'Requirements for IP Version 4 Routers' [RFC1812], 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. This item may not be applicable (N.A.).

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. Components and Characteristics of Routing Information

2.1. (Network) Prefix

Definition:
"A network prefix is a contiguous set of bits at the more significant end of the address that collectively designates the set of systems within a network; host numbers select among those systems." (This definition is taken directly from section 2.2.5.2, "Classless Inter Domain Routing (CIDR)", of RFC 1812.)

Discussion:
In the CIDR context, the network prefix is the network component of an IP address. In IPv4 systems, the network component of a complete address is known as the 'network part', and the remaining part of the address is known as the 'host part'. In IPv6 systems,

the network component of a complete address is known as the 'subnet prefix', and the remaining part is known as the 'interface identifier'.

Measurement units: N.A.

Issues:

See also:

2.2. Network Prefix Length

Definition:
The network prefix length is the number of bits, out of the total constituting the address field, that define the network prefix portion of the address.

Discussion:
A common alternative to using a bit-wise mask to communicate this component is the use of slash (/) notation. This binds the notion of network prefix length in bits to an IP address. For example, 141.184.128.0/17 indicates that the network component of this IPv4 address is 17 bits wide. Similar notation is used for IPv6 network prefixes; e.g., 2001:db8:719f::/48. 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.

2.3. Route

Definition:
In general, a 'route' is the n-tuple <prefix, nexthop [, other routing or non-routing protocol attributes]>. A route is not end-to-end, but is defined with respect to a specific next hop that should take packets on the next step toward their destination as 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.

2.4. 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, where they are known as Path Attributes, and they are the qualifying data that define the route. 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."

Measurement units: N.A.

Issues:

See also:
Route, Prefix, Adj-RIB-In, Network Level Reachability Information (NLRI)

2.5. Network Level Reachability Information (NLRI)

Definition:
The NLRI consists of one or more network prefixes with the same set of path attributes.

Discussion:
Each prefix in the NLRI is combined with the (common) path attributes 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.

2.6. BGP UPDATE Message

Definition:
An UPDATE message contains an advertisement of a single NLRI field, possibly containing multiple prefixes, and multiple withdrawals of unfeasible routes. See RFC 1771 for details.

Discussion:
From RFC 1771: "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.

See also:

3. Routing Data Structures and Route Categories

3.1. Routing Information Base (RIB)

The RIB collectively consists of a set of logically (not necessarily physically) 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.

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 among 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 that need to be considered include internal BGP, external BGP, interface, static, and IGP routes.

Issues:

Measurement units:
Number of routes.

See also:
Route, BGP Route, Route Instance, Adj-RIB-In, Adj-RIB-Out, FIB.

3.2. Prefix Filtering

Definition:
Prefix Filtering is a technique for eliminating routes from consideration as candidates for entry into a RIB by matching the network prefix in a BGP Route against a list of network prefixes.

Discussion:
A BGP Route is eliminated if, for any filter prefix from the list, the Route prefix length is equal to or longer than the filter prefix length and the most significant bits of the two prefixes

match over the length of the filter prefix. See 'Cooperative Route Filtering Capability for BGP-4' [BGP-4] for examples of usage.

Measurement units:
Number of filter prefixes; lengths of prefixes.

Issues:

See also:
BGP Route, Network Prefix, Network Prefix Length, Routing Policy, Routing Policy Information Base.

3.3. Routing Policy

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

Discussion:
RFC 1771 further constrains policy to be within the hop-by-hop routing paradigm. Policy is implemented using filters and associated policy actions such as Prefix Filtering. Many ASes formulate and document their policies using the Routing Policy Specification Language (RPSL) [RFC2622] 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:
Routing Policy Information Base, Prefix Filtering.

3.4. Routing Policy Information Base

Definition:
A routing 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 definition of these phases. Incoming policies are applied in Phase 1 of the BGP selection process 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 includes route prefixes, AS paths, communities, etc. The action on match may be to drop the update and not to 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 that accepts all routes and sends them according to a 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:

3.5. Forwarding Information Base (FIB)

Definition:
According to the definition in Appendix B of RIPE-37 [RIPE37]: "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. It is a data plane construct and is used for the forwarding of each packet. 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.

3.6. BGP Instance

Definition:
A BGP instance is a process with a single Loc-RIB.

Discussion:
For example, a BGP instance would run in routers or test equipment. A test generator acting as multiple peers will typically run more than one instance of BGP. A router would typically run a single instance.

Measurement units: N.A.

Issues:

See also:

3.7. BGP Device

Definition:
A BGP device is a system that has one or more BGP instances running on it, each of which is responsible for executing the BGP state machine.

Discussion:
We have chosen to use "device" as the general case, to deal with the understood (e.g., [GLSSRY]) and yet-to-be-invented cases where the control processing may be separate from forwarding [RFC2918]. A BGP device may be a traditional router, a route server, a BGP-aware traffic steering device, or a non-forwarding route reflector. BGP instances such as route reflectors or servers, for example, never forward traffic, so forwarding-based measurements would be meaningless for them.

Measurement units: N.A.

Issues:

See also:

3.8. BGP Session

Definition:
A BGP session is a session between two BGP instances.

Discussion:

Measurement units: N.A.

Issues:

See also:

3.9. Active BGP Session

Definition:
An active BGP session is one that is in the established state. (See RFC 1771.)

Discussion:

Measurement units: N.A.

Issues:

See also:

3.10. BGP Peer

Definition:
A BGP peer is another BGP instance to which the DUT is in the Established state. (See RFC 1771.)

Discussion:
In the test scenarios for the methodology discussion that will follow this document, peers send BGP advertisements to the DUT and receive DUT-originated advertisements. We recommend that the peering relation be established before tests begin. It might also be interesting to measure the time required to reach the established state. 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 4.2).

Measurement units:
Number of BGP peers.

Issues:

See also:

3.11. BGP Neighbor

Definition:
A BGP neighbor is a device that can be configured as a BGP peer.

Discussion:

Measurement units:

Issues:

See also:

3.12. MinRouteAdvertisementInterval (MRAI)

Definition:
(Paraphrased from RFC 1771) 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 routing plane convergence by measuring from 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.

3.13. 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:
It is not known what, if any, relationship exists between the settings of MRAI and MAOI.

See also:
MRAI, BGP Route.

3.14. 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.

3.15. 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.

3.16. 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.

3.17. Route Instance

Definition:
A route instance is one of several possible occurrences of a route for a particular prefix.

Discussion:
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. Each route instance is associated with a specific peer. The BGP algorithm that arbitrates between the available candidate route instances 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 border router, located in the default-free zone (see Section 4.1.4), will likely receive more route instances than a provider edge router, located closer to the end-users of the network.

See also:

4. 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.

4.1. 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). Also, a router that has a full set of routes as in a DFT, but that also has a 'discard' rule for a default route would not be considered default free.

Note that in this section the references to number of routes are to routes installed in the loc-RIB, which are therefore unique routes, not route instances. Also note that the total number of route instances may be 4 to 10 times the number of routes.

4.1.1. Default Route

Definition:
A default route can match any destination address. If a router does not have a more specific route for a particular packet's destination address, it forwards this packet to the next hop in the default route entry, provided that its Forwarding Table (Forwarding Information Base, or 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.

4.1.2. 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 packet to a specific next hop based on the longest match between the destination IP address and the routes in the forwarding table.

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.

Measurement units:
The number of routes.

See also:
Full Default-Free Table, Default Route.

4.1.3. Full Default-Free Table

Definition:
A full default-free table is the union of all sets of BGP routes taken from all the default-free BGP routing tables 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 on where and when it is observed.

Discussion:
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:
The full default-free routing table is not the same as the union of all reachable unicast addresses. The table simply does not

contain the default prefix (0/0) and does contain the union of all sets of BGP routes from default-free BGP routing tables.

See also:
Routes, Route Instances, Default Route.

4.1.4. Default-Free Zone

Definition:
The default-free zone is the part of the Internet backbone that does not have a default route.

Discussion:

Measurement units:

Issues:

See also:
Default Route.

4.1.5. 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 might 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.

4.2. 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.

4.2.1. Provider Edge Router

Definition:
A provider edge router is a router at the edge of a provider's network that speaks eBGP to a BGP speaker in another AS.

Discussion:
The traffic that transits this router may be destined to or may originate from non-adjacent autonomous systems. In particular, the MED values used in the Provider Edge Router would not be visible in the non-adjacent autonomous systems. Such a router will always speak eBGP and may speak iBGP.

Measurement units:

Issues:

See also:

4.2.2. Subscriber Edge Router

Definition:
A subscriber edge router is router at the edge of the subscriber's network that speaks eBGP to its provider's AS(s).

Discussion:
The router belongs to an end user organization that may be multi-homed, and that carries traffic only to and from that end user AS. Such a router will always speak eBGP and may speak iBGP.

Measurement units:

Issues:
This definition of an enterprise border router (which is what most Subscriber Edge Routers are) is practical rather than rigorous. It is meant to draw attention to the reality that many enterprises may need a BGP speaker that advertises their own routes and accepts either default alone or partial routes. In such cases, they may be interested in benchmarks that use a partial routing table, to see whether a smaller control plane processor will meet their needs.

See also:

4.2.3. Inter-provider Border Router

Definition:
An inter-provider border router is a BGP speaking router that maintains BGP sessions with other BGP speaking routers in other providers' ASes.

Discussion:
Traffic transiting this router may be originated in or destined for another AS that has no direct connectivity with this provider's AS. Such a router will always speak eBGP and may speak iBGP.

Measurement units:

Issues:

See also:

4.2.4. Core Router

Definition:
An core router is a provider router internal to the provider's net, speaking iBGP to that 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:
By this definition, the DUTs that are eBGP routers aren't core routers.

See also:

5. Characterization of Sets of Update Messages

This section contains a sequence of definitions that build up to the definition of an update train. The packet train concept was originally introduced by Jain and Routhier [PKTTRAIN]. It is here adapted to refer to a train of packets of interest in BGP performance testing.

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 on or taken from live traces that reflect realistic sequences of UPDATEs and their contents.

5.1. 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 as 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.

5.2. Route Mixture

Definition:
A route mixture is the demographics of a set of routes.

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 might be suitable to test the performance limits of the BGP device. Using live data or input that simulates live data will 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 this 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 length distribution

* Packet packing

* Probability density function of inter-arrival times of UPDATES

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

Measurement units:
Probability density functions.

Issues:

See also:
NLRI, RIB.

5.3. 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, upstream 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 in live networks.

See also:
Route Mixture, MRAI, MAOI.

5.4. 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.

5.5. Route Flap

Definition:
A route flap is a change of state (withdrawal, announcement, attribute change) for a route.

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 229 [RIPE229], which contains current guidelines on flap-damping parameters.

Measurement units:
Flapping events per unit time.

Issues:
Specific Flap events can be found in Section 6.1. A bench-marker SHOULD use a mixture of different route change events in testing.

See also:
Route Change Events, Flap Damping, Packet Train

6. Route Changes and Convergence

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

6.1. Route Change Events

A taxonomy characterizing routing information changes seen in operational networks is proposed in RIPE-37 [RIPE37] and Labovitz et al [INSTBLTY]. 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 [INSTBLTY] and given below.

1. AADiff: Implicit withdrawal of a route and replacement by a route different in some path attribute.

2. AADup: Implicit withdrawal of a route and replacement by route that is identical in all path attributes.

3. WADiff: Explicit withdrawal of a route and replacement by a different route.

4. 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 terms to describe the sequence of events from the update train perspective, as listed above, and event indications in the time domain in order to measure activity from the perspective of the DUT. With this in mind, we incorporate and extend the definitions of [INSTBLTY] to the following:

1. Tup (TDx): Route advertised to the DUT by Test Device x

2. Tdown(TDx): Route being withdrawn by Device x

3. Tupinit(TDx): The initial announcement of a route to a unique prefix

4. 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:

1. Tup(1,TDx) means Device x sends 1 route

2. Tup(S,TDx) means Device x sends a train, S, of routes

3. 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 RFC 1771, 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:

1. Tbest -- The current best path.

2. Tbetter -- Advertise a path that is better than Tbest.

3. Tworse -- Advertise a path that is worse than Tbest.

6.2. 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, the convergence resulting from a change that alters the best route instance for a single prefix at a router would be deemed to have occurred when this route is advertised to its downstream peers. By way of contrast, OSPF convergence concludes when SPF calculations have been performed and the required link states are advertised onward. 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
that we describe herein and in 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 (as far as is possible) to exclude any

additional time needed to download and install the forwarding information base in the data plane. This definition is 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. Care also needs to be taken to ensure that the convergence time is not influenced by policy processing on downstream peers. 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:

7. BGP Operation Events

The BGP process(es) in a device might restart because operator intervention or a power failure caused a complete shutdown. In this case, a hard reset is needed. A peering session could be lost, for example, because of action on the part of the peer or a dropped TCP session. A device can reestablish its peers and re-advertise all relevant routes in a hard reset. However, if a peer is lost, but the BGP process has not failed, BGP has mechanisms for a "soft reset."

7.1. Hard Reset

Definition:
An event that 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:

7.2. Soft Reset

Definition:
A soft reset is performed on a per-neighbor basis; it does not clear the BGP session while re-establishing the peering relation and does not stop the flow of traffic.

Discussion:
There are two methods of performing a soft reset: (1) graceful restart [GRMBGP], wherein the BGP device that has lost a peer continues to forward traffic for a period of time before tearing down the peer's routes and (2) soft refresh [RFC2918], wherein a BGP device can request a peer's Adj-RIB-Out.

Measurement units: N.A.

Issues:

See also:

8. Factors That Impact the Performance of the Convergence Process

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

8.1. General Factors Affecting Device Convergence

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

8.1.1. 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, depending on the quantity of updates generated by each additional peer. Increasing the number of peers also increases the processing workload for TCP and BGP keepalives.

8.1.2. 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.

8.1.3. 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.

The following are extreme examples:

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

8.1.4. 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 an understanding of the interactions with both the IGPs and the protocols associated with the physical media, such as Ethernet, SONET, and DWDM.

8.1.5. 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 [RFC2439] and discussed in RIPE-229 [RIPE229]. 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.

8.1.6. Churn

In theory, a BGP device could receive a set of updates that completely define 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 it re-advertises this information, these are normal constituents of churn. Routine updates can range from single announcements 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 [RFC3345]. 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 on whether the same processor(s) are used for forwarding and control.

8.2. 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.

8.2.1. 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 would be presented if only control traffic were present.

8.2.2. 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 MUST be reported if they use non-default values.

8.2.3. TCP Parameters Underlying BGP Transport

Because all BGP traffic and interactions occur over TCP, all relevant parameters characterizing the TCP sessions MUST be provided; e.g., slow start, max window size, maximum segment size, or timers.

8.2.4. Authentication

Authentication in BGP is currently done using the TCP MD5 Signature Option [RFC2385]. 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.

9. Security Considerations

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

10. 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. Debby Stopp and Nick Ambrose contributed the concept of route packing.

Alvaro Retana was a key member of the team that developed this document, and made significant technical contributions regarding route mixes. The team thanks him and regards him as a co-author in spirit.

11. References

11.1. Normative References

[RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, November 1998.

[RFC1812] Baker, F., "Requirements for IP Version 4 Routers", RFC 1812, June 1995.

[RIPE37] Ahuja, A., Jahanian, F., Bose, A., and C. Labovitz, "An Experimental Study of Delayed Internet Routing Convergence", RIPE-37 Presentation to Routing WG, November 2000, <http://www.ripe.net/ripe/meetings/archive/ripe-37/presentations/RIPE-37-convergence/>

.

[INSTBLTY] Labovitz, C., Malan, G., and F. Jahanian, "Origins of Internet Routing Instability", Infocom 99, August 1999.

[RFC2622] Alaettinoglu, C., Bates, T., Gerich, E., Karrenberg, D., Meyer, D., Terpstra, M., and C. Villamizar, "Routing Policy Specification Language (RPSL)", RFC 2280, January 1998.

[RIPE229] Panigl, C., Schmitz, J., Smith, P., and C. Vistoli, "RIPE Routing-WG Recommendation for coordinated route-flap damping parameters, version 2", RIPE 229, October 2001.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

[GLSSRY] Juniper Networks, "Junos(tm) Internet Software Configuration Guide Routing and Routing Protocols, Release 4.2", Junos 4.2 and other releases, September 2000, <http://www.juniper.net/techpubs/software/junos/junos42/swcmdref42/html/glossary.html>

.

[RFC2547] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, March 1999.

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

11.2. Informative References

[RFC2918] Chen, E., "Route Refresh Capability for BGP-4", RFC 2918, September 2000.

[GRMBGP] Sangli, S., Rekhter, Y., Fernando, R., Scudder, J., and E. Chen, "Graceful Restart Mechanism for BGP", Work in Progress, June 2004.

[BGP-4] Chen, E. and Y. Rekhter, "Cooperative Route Filtering Capability for BGP-4", Work in Progress, March 2004.

[RFC3654] Khosravi, H. and T. Anderson, "Requirements for Separation of IP Control and Forwarding", RFC 3654, November 2003.

[RFC3345] McPherson, D., Gill, V., Walton, D., and A. Retana, "Border Gateway Protocol (BGP) Persistent Route Oscillation Condition", RFC 3345, August 2002.

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 Multiprotocol Extensions for IPv6 Inter-Domain Routing", RFC 2545, March 1999.

Authors' Addresses

Howard Berkowitz
Gett Communications & CCI Training
5012 S. 25th St
Arlington, VA 22206
USA

Phone: +1 703 998-5819
Fax: +1 703 998-5058
EMail: hcb@gettcomm.com

Elwyn B. Davies
Folly Consulting
The Folly
Soham
Cambs, CB7 5AW
UK

Phone: +44 7889 488 335
EMail: elwynd@dial.pipex.com

Susan Hares
Nexthop Technologies
825 Victors Way
Ann Arbor, MI 48108
USA

Phone: +1 734 222-1610
EMail: skh@nexthop.com

Padma Krishnaswamy
SAIC
331 Newman Springs Road
Red Bank, New Jersey 07701
USA

EMail: padma.krishnaswamy@saic.com

Marianne Lepp
Consultant

EMail: mlepp@lepp.com

Full Copyright Statement

Intellectual Property

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.