Application of the Border Gateway Protocol in the Internet

Status of this Memo

This document specifies an Internet standards track protocol for the
Internet community, and requests discussion and suggestions for
improvements. Please refer to the current edition of the "Internet
Official Protocol Standards" (STD 1) for the standardization state
and status of this protocol. Distribution of this memo is unlimited.

Abstract

This document, together with its companion document, "A Border
Gateway Protocol 4 (BGP-4)", define an inter-autonomous system
routing protocol for the Internet. "A Border Gateway Protocol 4
(BGP-4)" defines the BGP protocol specification, and this document
describes the usage of the BGP in the Internet.

Information about the progress of BGP can be monitored and/or
reported on the BGP mailing list (bgp@ans.net).

Acknowledgements

1. Introduction

This memo describes the use of the Border Gateway Protocol (BGP) [1] in the Internet environment. BGP is an inter-Autonomous System routing protocol. The network reachability information exchanged via BGP provides sufficient information to detect routing loops and enforce routing decisions based on performance preference and policy constraints as outlined in RFC 1104 [2]. In particular, BGP exchanges routing information containing full AS paths and enforces routing policies based on configuration information.

As the Internet has evolved and grown over in recent years, it has become painfully evident that it is soon to face several serious scaling problems. These include:

- Exhaustion of the class-B network address space. One fundamental cause of this problem is the lack of a network class of a size which is appropriate for mid-sized organization; class-C, with a maximum of 254 host addresses, is too small while class-B, which allows up to 65534 addresses, is too large to be densely populated.

- Growth of routing tables in Internet routers are beyond the ability of current software (and people) to effectively manage.

- Eventual exhaustion of the 32-bit IP address space.

It has become clear that the first two of these problems are likely to become critical within the next one to three years. Classless inter-domain routing (CIDR) attempts to deal with these problems by

proposing a mechanism to slow the growth of the routing table and the
need for allocating new IP network numbers. It does not attempt to
solve the third problem, which is of a more long-term nature, but
instead endeavors to ease enough of the short to mid-term
difficulties to allow the Internet to continue to function
efficiently while progress is made on a longer-term solution.

BGP-4 is an extension of BGP-3 that provides support for routing
information aggregation and reduction based on the Classless inter-
domain routing architecture (CIDR) [3]. This memo describes the
usage of BGP-4 in the Internet.

All of the discussions in this paper are based on the assumption that
the Internet is a collection of arbitrarily connected Autonomous
Systems. That is, the Internet will be modeled as a general graph
whose nodes are AS's and whose edges are connections between pairs of
AS's.

The classic definition of an Autonomous System is a set of routers
under a single technical administration, using an interior gateway
protocol and common metrics to route packets within the AS and using
an exterior gateway protocol to route packets to other AS's. Since
this classic definition was developed, it has become common for a
single AS to use several interior gateway protocols and sometimes
several sets of metrics within an AS. The use of the term Autonomous
System here stresses the fact that, even when multiple IGPs and
metrics are used, the administration of an AS appears to other AS's
to have a single coherent interior routing plan and presents a
consistent picture of which destinations are reachable through it.

AS's are assumed to be administered by a single administrative
entity, at least for the purposes of representation of routing
information to systems outside of the AS.

2. BGP Topological Model

When we say that a connection exists between two AS's, we mean two
things:

   Physical connection: There is a shared Data Link subnetwork
   between the two AS's, and on this shared subnetwork each AS has at
   least one border gateway belonging to that AS. Thus the border
   gateway of each AS can forward packets to the border gateway of
   the other AS without resorting to Inter-AS or Intra-AS routing.

   BGP connection: There is a BGP session between BGP speakers in
   each of the AS's, and this session communicates those routes that
   can be used for specific destinations via the advertising AS.

Throughout this document we place an additional restriction on the
BGP speakers that form the BGP connection: they must themselves
share the same Data Link subnetwork that their border gateways
share. Thus, a BGP session between adjacent AS's requires no
support from either Inter-AS or Intra-AS routing. Cases that do
not conform to this restriction fall outside the scope of this
document.

Thus, at each connection, each AS has one or more BGP speakers and
one or more border gateways, and these BGP speakers and border
gateways are all located on a shared Data Link subnetwork. Note that
BGP speakers do not need to be a border gateway, and vice versa.
Paths announced by a BGP speaker of one AS on a given connection are
taken to be feasible for each of the border gateways of the other AS
on the same shared subnetwork, i.e. indirect neighbors are allowed.

Much of the traffic carried within an AS either originates or
terminates at that AS (i.e., either the source IP address or the
destination IP address of the IP packet identifies a host internal to
that AS). Traffic that fits this description is called "local
traffic". Traffic that does not fit this description is called
"transit traffic". A major goal of BGP usage is to control the flow
of transit traffic.

Based on how a particular AS deals with transit traffic, the AS may
now be placed into one of the following categories:

    stub AS: an AS that has only a single connection to one other AS.
    Naturally, a stub AS only carries local traffic.

    multihomed AS: an AS that has connections to more than one other
    AS, but refuses to carry transit traffic.

    transit AS: an AS that has connections to more than one other AS,
    and is designed (under certain policy restrictions) to carry both
    transit and local traffic.

Since a full AS path provides an efficient and straightforward way of
suppressing routing loops and eliminates the "count-to-infinity"
problem associated with some distance vector algorithms, BGP imposes
no topological restrictions on the interconnection of AS's.
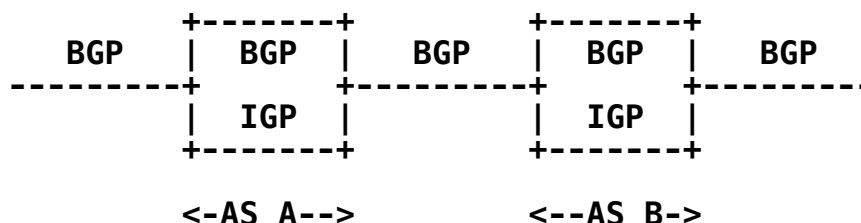
3. BGP in the Internet

3.1 Topology Considerations

   The overall Internet topology may be viewed as an arbitrary
   interconnection of transit, multihomed, and stub AS's. In order to
   minimize the impact on the current Internet infrastructure, stub and
   multihomed AS's need not use BGP. These AS's may run other protocols
   (e.g., EGP) to exchange reachability information with transit AS's.
   Transit AS's using BGP will tag this information as having been
   learned by some method other than BGP. The fact that BGP need not run
   on stub or multihomed AS's has no negative impact on the overall
   quality of inter-AS routing for traffic that either destined to or
   originated from the stub or multihomed AS's in question.

   However, it is recommended that BGP be used for stub and multihomed
   AS's as well. In these situations, BGP will provide an advantage in
   bandwidth and performance over some of the currently used protocols
   (such as EGP). In addition, this would reduce the need for the use
   of default routes and in better choices of Inter-AS routes for
   multihomed AS's.

3.2 Global Nature of BGP

   At a global level, BGP is used to distribute routing information
   among multiple Autonomous Systems. The information flows can be
   represented as follows:

```
                 +-------+          +-------+
        BGP    |  BGP  |   BGP  |  BGP  |   BGP
         ---------+       +---------+       +---------
                 |  IGP  |          |  IGP  |
                 +-------+          +-------+

            <-AS A-->          <--AS B->
```

   This diagram points out that, while BGP alone carries information
   between AS's, both BGP and an IGP may carry information across an AS.
   Ensuring consistency of routing information between BGP and an IGP
   within an AS is a significant issue and is discussed at length later
   in Appendix A.

3.3 BGP Neighbor Relationships

   The Internet is viewed as a set of arbitrarily connected AS's.
   Routers that communicate directly with each other via BGP are known
   as BGP speakers. BGP speakers can be located within the same AS or in
   different AS's. BGP speakers in each AS communicate with each other

to exchange network reachability information based on a set of
policies established within each AS. For a given BGP speaker, some
other BGP speaker with which the given speaker communicates is
referred to as an external peer if the other speaker is in a
different AS, while if the other speaker is in the same AS it is
referred to as an internal peer.

There can be as many BGP speakers as deemed necessary within an AS.
Usually, if an AS has multiple connections to other AS's, multiple
BGP speakers are needed. All BGP speakers representing the same AS
must give a consistent image of the AS to the outside. This requires
that the BGP speakers have consistent routing information among them.
These gateways can communicate with each other via BGP or by other
means. The policy constraints applied to all BGP speakers within an
AS must be consistent. Techniques such as using a tagged IGP (see
A.2.2) may be employed to detect possible inconsistencies.

In the case of external peers, the peers must belong to different
AS's, but share a common Data Link subnetwork. This common subnetwork
should be used to carry the BGP messages between them. The use of BGP
across an intervening AS invalidates the AS path information. An
Autonomous System number must be used with BGP to specify which
Autonomous System the BGP speaker belongs to.

4. Requirements for Route Aggregation

A conformant BGP-4 implementation is required to have the ability to
specify when an aggregated route may be generated out of partial
routing information. For example, a BGP speaker at the border of an
autonomous system (or group of autonomous systems) must be able to
generate an aggregated route for a whole set of destination IP
addresses (in BGP-4 terminology such a set is called the Network
Layer Reachability Information or NLRI) over which it has
administrative control (including those addresses it has delegated),
even when not all of them are reachable at the same time.

A conformant implementation may provide the capability to specify
when an aggregated NLRI may be generated.

A conformant implementation is required to have the ability to
specify how NLRI may be de-aggregated.

A conformant implementation is required to support the following
options when dealing with overlapping routes:

- Install both the less and the more specific routes

- Install the more specific route only

    - Install the less specific route only

    - Install neither route

Certain routing policies may depend on the NLRI (e.g. "research"
versus "commercial"). Therefore, a BGP speaker that performs route
aggregation should be cognizant, if possible, of potential
implications on routing policies when aggregating NLRI.

5. Policy Making with BGP

BGP provides the capability for enforcing policies based on various
routing preferences and constraints. Policies are not directly
encoded in the protocol. Rather, policies are provided to BGP in the
form of configuration information.

BGP enforces policies by affecting the selection of paths from
multiple alternatives and by controlling the redistribution of
routing information. Policies are determined by the AS
administration.

Routing policies are related to political, security, or economic
considerations. For example, if an AS is unwilling to carry traffic
to another AS, it can enforce a policy prohibiting this. The
following are examples of routing policies that can be enforced with
the use of BGP:

    1. A multihomed AS can refuse to act as a transit AS for other
        AS's. (It does so by only advertising routes to destinations
        internal to the AS.)

    2. A multihomed AS can become a transit AS for a restricted set of
        adjacent AS's, i.e., some, but not all, AS's can use the
        multihomed AS as a transit AS. (It does so by advertising its
        routing information to this set of AS's.)

    3. An AS can favor or disfavor the use of certain AS's for
        carrying transit traffic from itself.

A number of performance-related criteria can be controlled with the
use of BGP:

    1. An AS can minimize the number of transit AS's. (Shorter AS
        paths can be preferred over longer ones.)

    2. The quality of transit AS's. If an AS determines that two or
        more AS paths can be used to reach a given destination, that AS
        can use a variety of means to decide which of the candidate AS

paths it will use. The quality of an AS can be measured by such
things as diameter, link speed, capacity, tendency to become
congested, and quality of operation. Information about these
qualities might be determined by means other than BGP.

   3. Preference of internal routes over external routes.

For consistency within an AS, equal cost paths, resulting from
combinations of policies and/or normal route selection procedures,
must be resolved in a consistent fashion.

Fundamental to BGP is the rule that an AS advertises to its
neighboring AS's only those routes that it uses. This rule reflects
the "hop-by-hop" routing paradigm generally used by the current
Internet.

6. Path Selection with BGP

One of the major tasks of a BGP speaker is to evaluate different
paths from itself to a set of destination covered by an address
prefix, select the best one, apply appropriate policy constraints,
and then advertise it to all of its BGP neighbors. The key issue is
how different paths are evaluated and compared. In traditional
distance vector protocols (e.g., RIP) there is only one metric (e.g.,
hop count) associated with a path. As such, comparison of different
paths is reduced to simply comparing two numbers. A complication in
Inter-AS routing arises from the lack of a universally agreed-upon
metric among AS's that can be used to evaluate external paths.
Rather, each AS may have its own set of criteria for path evaluation.

A BGP speaker builds a routing database consisting of the set of all
feasible paths and the list of destinations (expressed as address
prefixes) reachable through each path. For purposes of precise
discussion, it's useful to consider the set of feasible paths for a
set of destinations associated with a given address prefix. In most
cases, we would expect to find only one feasible path. However, when
this is not the case, all feasible paths should be maintained, and
their maintenance speeds adaptation to the loss of the primary path.
Only the primary path at any given time will ever be advertised.

The path selection process can be formalized by defining a complete
order over the set of all feasible paths to a set of destinations
associated with a given address prefix. One way to define this
complete order is to define a function that maps each full AS path to
a non-negative integer that denotes the path's degree of preference.
Path selection is then reduced to applying this function to all
feasible paths and choosing the one with the highest degree of
preference.

In actual BGP implementations, the criteria for assigning degree of preferences to a path are specified as configuration information.

The process of assigning a degree of preference to a path can be based on several sources of information:

1. Information explicitly present in the full AS path.

2. A combination of information that can be derived from the full AS path and information outside the scope of BGP (e.g., policy routing constraints provided as configuration information).

Possible criteria for assigning a degree of preference to a path are:

- AS count. Paths with a smaller AS count are generally better.

- Policy considerations. BGP supports policy-based routing based on the controlled distribution of routing information. A BGP speaker may be aware of some policy constraints (both within and outside of its own AS) and do appropriate path selection. Paths that do not comply with policy requirements are not considered further.

- Presence or absence of a certain AS or AS's in the path. By means of information outside the scope of BGP, an AS may know some performance characteristics (e.g., bandwidth, MTU, intra-AS diameter) of certain AS's and may try to avoid or prefer them.

- Path origin. A path learned entirely from BGP (i.e., whose endpoint is internal to the last AS on the path) is generally better than one for which part of the path was learned via EGP or some other means.

- AS path subsets. An AS path that is a subset of a longer AS path to the same destination should be preferred over the longer path. Any problem in the shorter path (such as an outage) will also be a problem in the longer path.

- Link dynamics. Stable paths should be preferred over unstable ones. Note that this criterion must be used in a very careful way to avoid causing unnecessary route fluctuation. Generally, any criteria that depend on dynamic information might cause routing instability and should be treated very carefully.

7. Required set of supported routing policies

Policies are provided to BGP in the form of configuration
information. This information is not directly encoded in the
protocol. Therefore, BGP can provide support for very complex routing
policies. However, it is not required that all BGP implementations
support such policies.

We are not attempting to standardize the routing policies that must
be supported in every BGP implementation; we strongly encourage all
implementors to support the following set of routing policies:

    1. BGP implementations should allow an AS to control announcements
        of BGP-learned routes to adjacent AS's. Implementations should
        also support such control with at least the granularity of a
        single address prefix. Implementations should also support
        such control with the granularity of an autonomous system,
        where the autonomous system may be either the autonomous system
        that originated the route, or the autonomous system that
        advertised the route to the local system (adjacent autonomous
        system). Care must be taken when a BGP speaker selects a new
        route that can't be announced to a particular external peer,
        while the previously selected route was announced to that peer.
        Specifically, the local system must explicitly indicate to the
        peer that the previous route is now infeasible.

    2. BGP implementations should allow an AS to prefer a particular
        path to a destination (when more than one path is available).
        At the minimum an implementation shall support this
        functionality by allowing to administratively assign a degree
        of preference to a route based solely on the IP address of the
        neighbor the route is received from. The allowed range of the
        assigned degree of preference shall be between 0 and $2^{(31)}$ -
        1.

    3. BGP implementations should allow an AS to ignore routes with
        certain AS's in the AS_PATH path attribute. Such function can
        be implemented by using the technique outlined in [2], and by
        assigning "infinity" as "weights" for such AS's. The route
        selection process must ignore routes that have "weight" equal
        to "infinity".

8. Interaction with other exterior routing protocols

    The guidelines suggested in this section are consistent with the
    guidelines presented in [3].

An AS should advertise a minimal aggregate for its internal
destinations with respect to the amount of address space that it is
actually using. This can be used by administrators of non-BGP 4 AS's
to determine how many routes to explode from a single aggregate.

A route that carries the ATOMIC_AGGREGATE path attribute shall not be
exported into either BGP-3 or EGP2, unless such an exportation can be
accomplished without exploding the NLRI of the route.

8.1 Exchanging information with EGP2

This document suggests the following guidelines for exchanging
routing information between BGP-4 and EGP2.

To provide for graceful migration, a BGP speaker may participate in
EGP2, as well as in BGP-4. Thus, a BGP speaker may receive IP
reachability information by means of EGP2 as well as by means of
BGP-4. The information received by EGP2 can be injected into BGP-4
with the ORIGIN path attribute set to 1. Likewise, the information
received via BGP-4 can be injected into EGP2 as well. In the latter
case, however, one needs to be aware of the potential information
explosion when a given IP prefix received from BGP-4 denotes a set of
consecutive A/B/C class networks. Injection of BGP-4 received NLRI
that denotes IP subnets requires the BGP speaker to inject the
corresponding network into EGP2. The local system shall provide
mechanisms to control the exchange of reachability information
between EGP2 and BGP-4. Specifically, a conformant implementation is
required to support all of the following options when injecting BGP-4
received reachability information into EGP2:

   - inject default only (0.0.0.0); no export of any other NLRI

   - allow controlled deaggregation, but only of specific routes;
     allow export of non-aggregated NLRI

   - allow export of only non-aggregated NLRI

The exchange of routing information via EGP2 between a BGP speaker
participating in BGP-4 and a pure EGP2 speaker may occur only at the
domain (autonomous system) boundaries.

8.2 Exchanging information with BGP-3

This document suggests the following guidelines for exchanging
routing information between BGP-4 and BGP-3.

To provide for graceful migration, a BGP speaker may participate in
BGP-3, as well as in BGP-4. Thus, a BGP speaker may receive IP

reachability information by means of BGP-3, as well as by means of
BGP-4.

A BGP speaker may inject the information received by BGP-4 into BGP-3
as follows.

If an AS_PATH attribute of a BGP-4 route carries AS_SET path
segments, then the AS_PATH attribute of the BGP-3 route shall be
constructed by treating the AS_SET segments as AS_SEQUENCE segments,
with the resulting AS_PATH being a single AS_SEQUENCE. While this
procedure loses set/sequence information, it doesn't affect
protection for routing loops suppression, but may affect policies, if
the policies are based on the content or ordering of the AS_PATH
attribute.

While injecting BGP-4 derived NLRI into BGP-3, one needs to be aware
of the potential information explosion when a given IP prefix denotes
a set of consecutive A/B/C class networks. Injection of BGP-4 derived
NLRI that denotes IP subnets requires the BGP speaker to inject the
corresponding network into BGP-3. The local system shall provide
mechanisms to control the exchange of routing information between
BGP-3 and BGP-4. Specifically, a conformant implementation is
required to support all of the following options when injecting BGP-4
received routing information into BGP-3:

   - inject default only (0.0.0.0), no export of any other NLRI

   - allow controlled deaggregation, but only of specific routes;
     allow export of non-aggregated NLRI

   - allow export of only non-aggregated NLRI

The exchange of routing information via BGP-3 between a BGP speaker
participating in BGP-4 and a pure BGP-3 speaker may occur only at
the autonomous system boundaries. Within a single autonomous system
BGP conversations between all the BGP speakers of that autonomous
system have to be either BGP-3 or BGP-4, but not a mixture.

9. Operations over Switched Virtual Circuits

When using BGP over Switched Virtual Circuit (SVC) subnetworks it may
be desirable to minimize traffic generated by BGP. Specifically, it
may be desirable to eliminate traffic associated with periodic
KEEPALIVE messages. BGP includes a mechanism for operation over
switched virtual circuit (SVC) services which avoids keeping SVCs
permanently open and allows it to eliminates periodic sending of
KEEPALIVE messages.

This section describes how to operate without periodic KEEPALIVE
messages to minimise SVC usage when using an intelligent SVC circuit
manager. The proposed scheme may also be used on "permanent"
circuits, which support a feature like link quality monitoring or
echo request to determine the status of link connectivity.

The mechanism described in this section is suitable only between the
BGP speakers that are directly connected over a common virtual
circuit.

9.1 Establishing a BGP Connection

The feature is selected by specifying zero Hold Time in the OPEN
message.

9.2 Circuit Manager Properties

The circuit manager must have sufficient functionality to be able to
compensate for the lack of periodic KEEPALIVE messages:

- It must be able to determine link layer unreachability in a
  predictable finite period of a failure occurring.

- On determining unreachability it should:

     - start a configurable dead timer (comparable to a
       typical Hold timer value).

     - attempt to re-establish the Link Layer connection.

- If the dead timer expires it should:

     - send an internal circuit DEAD indication to TCP.

- If the connection is re-established it should:

     - cancel the dead timer.

     - send an internal circuit UP indication to TCP.

9.3 TCP Properties

A small modification must be made to TCP to process internal
notifications from the circuit manager:

- DEAD: Flush transmit queue and abort TCP connection.

- UP: Transmit any queued data or allow an outgoing TCP call to
  proceed.

9.4 Combined Properties

Some implementations may not be able to guarantee that the BGP
process and the circuit manager will operate as a single entity; i.e.
they can have a separate existence when the other has been stopped or
has crashed.

If this is the case, a periodic two-way poll between the BGP process
and the circuit manager should be implemented. If the BGP process
discovers the circuit manager has gone away it should close all
relevant TCP connections. If the circuit manager discovers the BGP
process has gone away it should close all its connections associated
with the BGP process and reject any further incoming connections.

10. Conclusion

The BGP protocol provides a high degree of control and flexibility
for doing interdomain routing while enforcing policy and performance
constraints and avoiding routing loops. The guidelines presented here
will provide a starting point for using BGP to provide more
sophisticated and manageable routing in the Internet as it grows.

Appendix A. The Interaction of BGP and an IGP

   This section outlines methods by which BGP can exchange routing
   information with an IGP. The methods outlined here are not proposed
   as part of the standard BGP usage at this time. These methods are
   outlined for information purposes only. Implementors may want to
   consider these methods when importing IGP information.

   This is general information that applies to any generic IGP.

   Interaction between BGP and any specific IGP is outside the scope of
   this section. Methods for specific IGP's should be proposed in
   separate documents. Methods for specific IGP's could be proposed for
   standard usage in the future.

Overview

   By definition, all transit AS's must be able to carry traffic which
   originates from and/or is destined to locations outside of that AS.
   This requires a certain degree of interaction and coordination
   between BGP and the Interior Gateway Protocol (IGP) used by that
   particular AS. In general, traffic originating outside of a given AS
   is going to pass through both interior gateways (gateways that
   support the IGP only) and border gateways (gateways that support both
   the IGP and BGP). All interior gateways receive information about
   external routes from one or more of the border gateways of the AS via
   the IGP.

   Depending on the mechanism used to propagate BGP information within a
   given AS, special care must be taken to ensure consistency between
   BGP and the IGP, since changes in state are likely to propagate at
   different rates across the AS. There may be a time window between the
   moment when some border gateway (A) receives new BGP routing
   information which was originated from another border gateway (B)
   within the same AS, and the moment the IGP within this AS is capable
   of routing transit traffic to that border gateway (B). During that
   time window, either incorrect routing or "black holes" can occur.

   In order to minimize such routing problems, border gateway (A) should
   not advertise to any of its external peers a route to some set of
   exterior destinations associated with a given address prefix X via
   border gateway (B) until all the interior gateways within the AS are
   ready to route traffic destined to these destinations via the correct
   exit border gateway (B). In other words, interior routing should
   converge on the proper exit gateway before/advertising routes via
   that exit gateway to external peers.

A.2 Methods for Achieving Stable Interactions

   The following discussion outlines several techniques capable of
   achieving stable interactions between BGP and the IGP within an
   Autonomous System.

A.2.1 Propagation of BGP Information via the IGP

   While BGP can provide its own mechanism for carrying BGP information
   within an AS, one can also use an IGP to transport this information,
   as long as the IGP supports complete flooding of routing information
   (providing the mechanism to distribute the BGP information) and one
   pass convergence (making the mechanism effectively atomic). If an IGP
   is used to carry BGP information, then the period of
   desynchronization described earlier does not occur at all, since BGP
   information propagates within the AS synchronously with the IGP, and
   the IGP converges more or less simultaneously with the arrival of the
   new routing information. Note that the IGP only carries BGP
   information and should not interpret or process this information.

A.2.2 Tagged Interior Gateway Protocol

   Certain IGPs can tag routes exterior to an AS with the identity of
   their exit points while propagating them within the AS. Each border
   gateway should use identical tags for announcing exterior routing
   information (received via BGP) both into the IGP and when propagating
   this information to other internal peers (peers within the same AS).
   Tags generated by a border gateway must uniquely identify that
   particular border gateway--different border gateways must use
   different tags.

   All Border Gateways within a single AS must observe the following two
   rules:

     1.   Information received from an internal peer by a border gateway
          A declaring a set of destination associated with a given
          address prefix to be unreachable must immediately be propagated
          to all of the external peers of A.

     2.   Information received from an internal peer by a border gateway
          A about a set of reachable destinations associated with a given
          address prefix X cannot be propagated to any of the external
          peers of A unless/until A has an IGP route to the set of
          destinations covered by X and both the IGP and the BGP routing
          information have identical tags.

   These rules guarantee that no routing information is announced
   externally unless the IGP is capable of correctly supporting it. It

also avoids some causes of "black holes".

One possible method for tagging BGP and IGP routes within an AS is to use the IP address of the exit border gateway announcing the exterior route into the AS. In this case the "gateway" field in the BGP UPDATE message is used as the tag.

An alternate method for tagging BGP and IGP routes is to have BGP and the IGP agree on a router ID. In this case, the router ID is available to all BGP (version 3 or higher) speakers. Since this ID is already unique it can be used directly as the tag in the IGP.

A.2.3 Encapsulation

Encapsulation provides the simplest (in terms of the interaction between the IGP and BGP) mechanism for carrying transit traffic across the AS. In this approach, transit traffic is encapsulated within an IP datagram addressed to the exit gateway. The only requirement imposed on the IGP by this approach is that it should be capable of supporting routing between border gateways within the same AS.

The address of the exit gateway A for some exterior destination X is specified in the BGP identifier field of the BGP OPEN message received from gateway A (via BGP) by all other border gateways within the same AS. In order to route traffic to destination X, each border gateway within the AS encapsulates it in datagrams addressed to gateway A. Gateway A then performs decapsulation and forwards the original packet to the proper gateway in another AS.

Since encapsulation does not rely on the IGP to carry exterior routing information, no synchronization between BGP and the IGP is required.

Some means of identifying datagrams containing encapsulated IP, such as an IP protocol type code, must be defined if this method is to be used.

Note that, if a packet to be encapsulated has length that is very close to the MTU, that packet would be fragmented at the gateway that performs encapsulation.

A.2.4 Pervasive BGP

If all routers in an AS are BGP speakers, then there is no need to have any interaction between BGP and an IGP. In such cases, all routers in the AS already have full information of all BGP routes. The IGP is then only used for routing within the AS, and no BGP

routes are imported into the IGP.

For routers to operate in this fashion, they must be able to perform
a recursive lookup in their routing table. The first lookup will use
a BGP route to establish the exit router, while the second lookup
will determine the IGP path to the exit router.

Since the IGP carries no external information in this scenario, all
routers in the AS will have converged as soon as all BGP speakers
have new information about this route. Since there is no need to
delay for the IGP to converge, an implementation may advertise these
routes without further delay due to the IGP.

A.2.5 Other Cases

There may be AS's with IGPs which can neither carry BGP information
nor tag exterior routes (e.g., RIP). In addition, encapsulation may
be either infeasible or undesirable. In such situations, the
following two rules must be observed:

   1. Information received from an internal peer by a border gateway
       A declaring a destination to be unreachable must immediately be
       propagated to all of the external peers of A.

   2. Information received from an internal peer by a border gateway
       A about a reachable destination X cannot be propagated to any
       of the external peers of A unless A has an IGP route to X and
       sufficient time has passed for the IGP routes to have
       converged.

The above rules present necessary (but not sufficient) conditions for
propagating BGP routing information to other AS's. In contrast to
tagged IGPs, these rules cannot ensure that interior routes to the
proper exit gateways are in place before propagating the routes to
other AS's.

If the convergence time of an IGP is less than some small value X,
then the time window during which the IGP and BGP are unsynchronized
is less than X as well, and the whole issue can be ignored at the
cost of transient periods (of less than length X) of routing
instability. A reasonable value for X is a matter for further study,
but X should probably be less than one second.

If the convergence time of an IGP cannot be ignored, a different
approach is needed. Mechanisms and techniques which might be
appropriate in this situation are subjects for further study.

References

   [1] Rekhter Y., and T. Li, "A Border Gateway Protocol 4 (BGP-4), RFC
       1771, T.J. Watson Research Center, IBM Corp., cisco Systems,
       March 1995.

   [2] Braun, H-W., "Models of Policy Based Routing", RFC 1104,
       Merit/NSFNET, June 1989.

   [3] Fuller, V., Li, T., Yu, J., and K. Varadhan, "Supernetting: an
       Address Assignment and Aggregation Strategy", RFC1519, BARRNet,
       cisco, MERIT, OARnet, September 1993.

Security Considerations

   Security issues are not discussed in this memo.

Authors' Addresses

   Yakov Rekhter
   T.J. Watson Research Center IBM Corporation
   P.O. Box 704, Office H3-D40
   Yorktown Heights, NY 10598

   Phone: +1 914 784 7361
   EMail: yakov@watson.ibm.com

   Phill Gross
   MCI Data Services Division
   2100 Reston Parkway, Room 6001,
   Reston, VA 22091

   Phone: +1 703 715 7432
   EMail: 0006423401@mcimail.com

   IETF IDR WG mailing list: bgp@ans.net
   To be added: bgp-request@ans.net