Application of the Border Gateway Protocol and IDRP for IP in the Internet

Status of this Memo

Abstract

   This document, together with its companion documents, "A Border
   Gateway Protocol 4 (BGP-4)" and "IDRP for IP", define an inter-
   autonomous system routing protocol for the Internet. "A Border
   Gateway Protocol 4 (BGP-4)" defines the BGP protocol specification.
   "IDRP for IP" defines the use of IDRP for IP in the Internet. The
   IDRP specification [6] defines the IDRP protocol. This document
   describes the usage of the BGP and IDRP for IP in the Internet.

   Information about the progress of BGP can be monitored and/or
   reported on the BGP mailing list (bgp@ans.net). Information about
   the progress of IDRP for IP can be monitored and/or reported on the
   IDRP for IP mailing list (idrp-for-ip@merit.edu).

Acknowledgements

1. Introduction

   This memo describes the use of the Border Gateway Protocol (BGP) and
   "IDRP for IP" in the Internet environment. IDRP and BGP are inter-
   Autonomous System routing protocols. IDRP and BGP have common roots
   in version 2 of BGP. However, IDRP has been standardized within ISO
   as a multi-protocol inter-domain routing protocol. IDRP for IP and
   BGP-4 can be considered, for the most part, interchangeable. IDRP
   has a few additional features, and some minor differences in

encoding. The usage of these additional features will be discussed
in [Section 10](#).

Hereafter in this memo, "BGP" will refer to both BGP-4 and IDRP for
IP. BGP-4 will refer only to version 4 of BGP, and IDRP will refer

to only the IDRP for IP 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, it has become 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 organizations; 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 is 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) [4]. BGP-4 contains many of the
features initially put into the multi-protocol ISO IDRP protocol.
This memo describes the usage of both BGP-4 and IDRP for IP 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 networks 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 network between the two
   AS's, and on this shared network 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 networks 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 network 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 network. 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 network, 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 on a
network 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. BGP
speakers in each AS communicate with each other to exchange network
reachability information based on a set of policies established
within each AS. Routers that communicate directly with each other via
BGP are known as BGP neighbors. BGP neighbors can be located within
the same AS or in different AS's. For the sake of discussion, BGP
communications with neighbors in different AS's will be referred to
as External BGP, and with neighbors in the same AS as Internal BGP.

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 BGP, the BGP neighbors must belong to
different AS's, but share a common network. This common network
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 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 terminology such a set is called the Network Layer
Reachability Information or NLRI) over which it has administrative
control, even when not all of them are reachable at the same time.

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 neither route

     A conformant implementation may also support other options when

dealing with overlapping routes, as specified in Clause 7.16.3.1 of
[6].

By default a BGP speaker should aggregate NLRI representing subnets
to the corresponding network.

Injecting NLRI representing arbitrary subnets into BGP without
aggregation to the corresponding network shall be controlled via
configuration parameters.

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:

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

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

IDRP for IP has two features (DIST_LIST_INCL and DIST_LIST_EXCL)
which allow additional constraints to be placed on the propagation of
routing information, by restricting the group of AS's which may
receive certain information. See Section 10 for further details.

6. **Path Selection with BGP**

One of the major tasks of a BGP speaker is to evaluate different
paths to a destination network from its border gateways, 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 networks reachable through each path.
For purposes of precise discussion, it's useful to consider the set
of feasible paths for a given destination network. In many cases, we
would expect to find only one feasible path. However, when this is
not the case, all feasible paths should be maintained, as 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 partial
order over the set of all feasible paths to a given destination
network. One way to define this partial 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 lowest 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:

Information explicitly present in the full AS path. 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 set of 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. **Recommended 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.

While 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:

BGP implementations should allow an AS to control announcements of
BGP-learned routes to adjacent AS's. Implementations should support
such control with at least the granularity of a single network.
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. BGP implementations
should allow an AS to prefer a particular path to a destination (when
more than one path is available). This function may be implemented

   by allowing system administrators to assign "weights" to AS's, and by
   having the route selection process select a route with the lowest
   "weight" (where "weight" of a route is defined as a sum of "weights"
   of all AS's in the AS_PATH path attribute associated with that
   route). 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

   This section presents guidelines for routing information exchange
   between BGP and BGP-3 or EGP-2, as well as between BGP-4 and IDRP.
   The suggested guidelines are consistent with the guidelines presented
   in [3], [4], and [5].

   The routing information exchange has the following aspects: how a
   route received via EGP2/BGP-3 gets injected into BGP how a route
   received via BGP gets injected into EGP2/BGP-3 how a route received
   via BGP-4 gets injected into IDRP how a route received via IDRP gets
   injected into BGP-4

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

8.1 Exchanging Information With EGP2

   The following guidelines are suggested for exchanging routing
   information between BGP and EGP2.

   To provide for graceful migration, a BGP speaker may participate in
   EGP2 as well as in BGP. Thus, a BGP speaker may receive IP
   reachability information by means of EGP2 as well as by means of BGP.
   It is strongly recommended that the exchange of routing information
   via EGP2 between a BGP speaker participating in BGP and a pure EGP2
   speaker occur only at Autonomous System boundaries.

   The information received by EGP2 can be injected into BGP-4 with the
   ORIGIN path attribute set to 1. It can likewise be injected into
   IDRP with the EXT_INFO path attribute.

   Likewise, the information received via BGP can be injected into EGP2.
   In the latter case, however, one needs to be aware of the potential
   information explosion when a given IP prefix received from BGP

denotes a set of consecutive A/B/C class networks. Injection of BGP
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. Specifically, a
conformant implementation is required to support all of the following
options when injecting BGP 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

Additional constraints on injecting information received via IDRP
into EGP2 are listed in Section 8.4. Additional constraints on
injecting information received via BGP-4 into EGP2 are listed in
Section 8.5.

8.2 Exchanging information with BGP-3

The following guidelines are suggested for exchanging routing
information between BGP and BGP-3.

To provide for graceful migration, a BGP speaker may participate in
BGP-3 as well as in BGP. Thus, a BGP speaker may receive IP
reachability information by means of BGP-3 as well as by means of
BGP. It is strongly recommended that the exchange of routing
information via BGP-3 between a BGP speaker participating in BGP-3
and a pure BGP-3 speaker occur only at Autonomous System boundaries.

When injecting BGP-3 routes into BGP-4, the AS_SEQUENCE information
shall be injected as an AS_SET. For IDRP, the AS_SEQUENCE
information shall be injected as an RD_SET within the RD_PATH
attribute.

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. It may affect policies if
they are based on the content or ordering of the AS_PATH attribute.

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

IDRP's equivalent for the AS_PATH attribute is RD_PATH path
attribute, where RD stands for Routing Domain. A Routing Domain has
an identifier which for IDRP for IP is a fixed prefix catenated with

the AS number. Given this translation, as defined in the "IDRP for
IP" document, the IDRP RD_PATH information containing RD_SEQUENCEs
and RD_SETs is translated into BGP-3 AS_SEQUENCES attribute in the
same way BGP-4's AS_SEQUENCEs and AS-SETs are. The resulting BGP-3
AS_PATH attribute contains all the domains listed in the RD_PATH

attribute. Again, this does not affect loop suppression, but may
affect policies.

While injecting BGP 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 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. Specifically, a conformant implementation is required
to support all of the following options when injecting BGP 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

Additional constraints on injecting information received via IDRP
into BGP-3 are listed in Section 8.4. Additional constraints on
injecting information received via BGP-4 into BGP-3 are listed in
Section 8.5.

8.3 Exchanging Information Between IDRP and BGP-4

To provide for graceful migration, an IDRP speaker may participate in
BGP-4. Thus, an IDRP speaker may receive IP reachability information
by means of BGP-4, as well as by means of IDRP. If IDRP for IP and
BGP-4 routers restrict themselves to the set of functions which is
common to both protocols, translation between the protocols can be
done. Within these restrictions, routers can participate in both
IDRP and BGP-4 conversations on domain boundaries, and within routing
domains.

When passing a BGP-4 route with the ATOMIC_AGGREGATE path attribute
to IDRP, the IDRP for IP ATOMIC_AGGREGATE shall be included in the
IDRP route. The ATOMIC_AGGREGATE attribute is defined in [7].

Note that any IDRP for IP router receiving a route with the
ATOMIC_AGGREGATE option shall not deaggregate that route.

Also note that any IDRP router not recognizing the ATOMIC_AGGREGATE
option shall set the Partial bit in the Flag field of the attribute
to 1, as required by clause 7.11.1.a in [6].

In exporting reachability information from IDRP for IP to BGP-4, if
the IDRP for IP ATOMIC_AGGREGATE attribute is present and the Partial
bit is set to 0, the BGP-4 ATOMIC_AGGREGATE attribute shall be
included in the BGP-4 route. If the IDRP for IP ATOMIC_AGGREGATE

attribute is present and the Partial bit is set to 1, the BGP-4
ORIGIN attribute shall be set to INCOMPLETE.

The following table specifies mapping between BGP-4 and IDRP for IP
path attributes.

```
                        BGP4            IDRP
                        -------------------------------

                        ORIGIN          EXT_INFO

                        AS_PATH         RD_PATH

                        NEXT_HOP        NEXT_HOP

                        MULTI_EXTI_DISC MULTI_EXIT_DISC

                        LOC_PREF        ROUTE_SEPARATOR
```

Additional constraints on injecting information received via IDRP
into BGP-4 are listed in Section 8.4.

8.4 Additional Constraints on Exchange of IDRP Routes

The following IDRP attributes cannot be passed into EGP2, BGP-3, or
BGP-4:

CAPACITY RD-HOP-COUNT

However, IDRP routes with these attributes can be passed into EGP2,
BGP-3, or BGP-4 after stripping the attributes.

IDRP routes with the following path attributes cannot be passed into
BGP-3, BGP-4, and EGP2:

DIST_LIST_INCL DIST_LIST_EXCL HIERARCHICAL_RECORDING TRANSIT DELAY
RESIDUAL_ERROR EXPENSE LOCALLY DEFINED QoS SECURITY PRIORITY

When passing a route received via EGP2 or a route received via BGP-3
or BGP-4, such that the value of the ORIGIN attribute of the route is
anything but IGP, to IDRP, the corresponding IDRP route shall have
the EXT_INFO path attribute. If an IDRP route carries EXT_INFO path
attribute then the corresponding BGP-3 or BGP-4 route shall have
value of its ORIGIN attribute set to INCOMPLETE.

When passing a BGP-3 or BGP-4 route to IDRP, the IDRP RD-HOP-COUNT
attribute shall be constructed by counting the number of ASs in the
AS-PATH attribute of the route.

If an IDRP route carries ENTRY_SEQ or ENTRY_SET path segments, then
before passing this route to BGP-3 or BGP-4 the BIS shall assume that
the route exited all the confederations denoted in ENTRY_SET or

ENTRY_SEQ and update the RD_PATH of the route accordingly.

8.5 **Additional Constrains on Exchange of BGP-4 Routes**

   The following BGP-4 attribute can not be passed into EGP2, BGP-3, or
   IDRP:

   AGGREGATOR

   However, BGP-4 routes with this attribute can be passed into IDRP.

   A route that carries the BGP-4 ATOMIC_AGGREGATE path attribute shall
   not be exported into EGP2 or BGP-3, unless such export can be
   accomplished without deaggregating the NLRI of the route.

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 eliminate periodic sending of
   KEEPALIVE messages.

   This section describes how to operate without periodic KEEPALIVE
   messages to minimize SVC usage when using an intelligent SVC circuit
   manager. This 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 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 within a
   bounded time of the occurance of such a failure. 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 used with BGP-4) or to the IDRP Finite State

Machine (if used with IDRP for IP) If the connection is re-
established before the dead timer expires it should: cancel the dead
timer. If the connection is re-established after the dead timer has
expired (that is, after a DEAD indication has been sent) it should:
send an internal circuit UP indication to TCP (if used with BGP-4) or
to the IDRP Finite State Machine (if used with IDRP for IP).

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 IDRP Finite State Machine Properties

A small modification must be made to the IDRP Finite State Machine to
process internal notifications from the circuit manager: DEAD:
Generate the DEACTIVATE event UP: Generate the ACTIVATE event

9.5 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 in the case of BGP-4 or close all relevant
peer sessions in the case of IDRP. 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. IDRP for IP Differences

The additional attributes and protocol semantics IDRP contains
besides those in BGP-4 fall into three categories: QoS related,
Distribution List related, and Routing Domain Confederations related.

The QoS related attributes can be thought of as an extension of the

TOS functions already defined in IP. These functions have been left
by IDRP for IP for future study and experimentation. Anyone
interested in IDRP's QoS features should contact the BGP/IDRP for IP
working groups.

The Distribution List attributes are DIST_LIST_INCL and

   DIST_LIST_EXCL. Their function can be described as follows: BGP
   speakers receive information from BGP speakers in other AS's, which
   we call the "upstream" AS's. They then select routes from this
   information and propagate them to other AS's, which we call the
   "downstream" AS's. The distribution list attributes provide
   mechanisms for the router originating some reachability information
   (or any router downstream of it) to constrain the downstream
   propagation of that information. If distribution list attributes are
   included, downstream AS's are required to restrict distribution of
   the routing information -- in the case of DIST_LIST_INCL, the routing
   information may only be distributed to the specified set of AS's, and
   in the case of DIST_LIST_EXCL, the routing information my be
   distributed to any AS's but the specified set. If no distribution
   list attributes are included, the information may be distributed
   without constraint.

   Routing Domain Confederations is a mechanism to group together
   routing domains with compatible policies, in effect providing
   "aggregation" of Routing Domains. By using RDCs, AS paths can be
   compacted considerably.

   With a Confederation, several AS's can be grouped together. From the
   point of view of AS's external to the Confederation, the AS path
   information (RD_PATH) can be replaced by a single identifier, the RDC
   identifier. For example, if 10 associated AS's containing 30 IP
   networks decide to form a Confederation (they might be the members of
   an academic consortium, for example), they could advertise all 30 of
   their networks with a single entry in the RD_PATH, abstracting the
   internal topology of their confederation.

[11]. Conclusion

   The BGP protocols, BGP-4 and IDRP, provide 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
   will 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, unless
   encapsulation is used (see Section A.2.3).

   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 a route to some exterior network X via border gateway
   (B) to all of its BGP neighbors in other AS's until all the interior
   gateways within the AS are ready to route traffic destined to X 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 other AS's.

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 into Internal
   BGP when propagating this information to other border gateways 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:

   Information received via Internal BGP by a border gateway A declaring
   a network to be unreachable must immediately be propagated to all of
   the External BGP neighbors of A. Information received via Internal
   BGP by a border gateway A about a reachable network X cannot be
   propagated to any of the External BGP neighbors of A unless/until A
   has an IGP route to 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 network X is
specified in the BGP identifier field of the BGP OPEN message
received from gateway A via Internal BGP by all other border gateways
within the same AS. In order to route traffic to network 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:

Information received via Internal BGP by a border gateway A declaring
a network to be unreachable must immediately be propagated to all of
the External BGP neighbors of A. Information received via Internal
BGP by a border gateway A about a reachable network X cannot be
propagated to any of the External BGP neighbors 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] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4),
Internet Draft, cisco Systems, T.J. Watson Research Center, IBM
Corp., September 1993.

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

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

[4] Rekhter, Y., Li, T., "An Architecture for IP address Allocation
with CIDR", RFC1518, September 1993

[5] Rekhter, Y., Topolcic, C. "Exchanging Routing Information Across
Provider/Subscriber Boundaries in the CIDR Environment", RFC1520,
September 1993

[6] ISO/IEC IS 10747 - Information Processing Systems -
Telecommunications and Information Exchange between Systems -
Protocol for Exchange of Inter-domain Routeing Information among
Intermediate Systems to Support Forwarding of ISO 8473 PDUs, 1993.

[7] Hares, S., Scudder, J., "IDRP for IP", Internet Draft, Merit
Network Inc., September 1993.

[8] ISO/IEC JTC 1 "Protocol for Exchanging of Inter-Domain Routeing
Information among Intermediate Systems to Support Forwarding of ISO
8473 PDUs", IS10747 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 218
     Yorktown Heights, NY 10598

     Phone: (914) 945-3896
     EMail: yakov@watson.ibm.com

     Susan Hares
     Merit, Inc
     1071 Beal Avenue
     Ann Arbor, MI 4810x

     Phone: (313)936-2095
     Email: skh@merit.edu

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

     IETF IDRP for IP WG mailing list: idrp-for-ip@merit.edu
     To be added: idrp-request@merit.edu