INTERNET-DRAFT Danny McPherson
draft-mcpherson-bgp4-experience-00.txt Keyur Patel
Category Informational
Expires: October 2003 April 2003

Experience with the BGP-4 Protocol

<draft-mcpherson-bgp4-expereince-00.txt>

Status of this Document

Copyright Notice

Abstract

The purpose of this memo is to document how the requirements for advancing a routing protocol from Draft Standard to full Standard have been satisfied by Border Gateway Protocol version 4 (BGP-4). This report satisfies the requirement for "the second report", as described in Section 6.0 of RFC 1264. In order to fulfill the requirement, this report augments RFC 1773 and describes additional knowledge and understanding gained in the time between when the protocol was made a Draft Standard and when it was submitted for Standard.

Table of Contents

1. Introduction 4
2. BGP-4 Overview 4
2.1. A Border Gateway Protocol 4
2.2. BGP version 2 5
2.3. BGP version 3 5
2.4. BGP version 4 6
3. Management Information Base (MIB). 7
4. Implementations. 7
5. Operational Experience 8
6. Metrics. 9
6.1. MULTI_EXIT_DISC (MED) 9
6.1.1. Sending MEDs to BGP Peers. 10
6.1.2. MED of Zero Versus No MED. 10
6.1.3. MEDs and Temporal Route Selection. 10
7. LOCAL_PREF 10
8. Internal BGP In Large Autonomous Systems 11
9. Internet Dynamics. 12
10. BGP Routing Information Bases (RIBs). 12
11. Update Packing. 13
12. Limit Rate Updates. 13
13. Ordering of Path Attributes 14
14. AS_SET Sorting. 14
15. Control over Version Negotiation. 14
16. Reciept of Non-Transitive Attributes from eBGP Peer. 14
17. Security Considerations 15
17.1. TCP MD5 Signature Option 15
17.2. BGP Over IPSEC 15
17.3. Miscellaneous. 16
17.4. Acknowledgements 16
18. References. 17
19. Authors' Addresses. 18
20. Full Copyright Statement. 18

1. Introduction

The purpose of this memo is to document how the requirements for advancing a routing protocol from Draft Standard to full Standard have been satisfied by Border Gateway Protocol version 4 (BGP-4). This report satisfies the requirement for "the second report", as described in Section 6.0 of RFC 1264. In order to fulfill the requirement, this report augments RFC 1773 and describes additional knowledge and understanding gained in the time between when the protocol was made a Draft Standard and when it was submitted for Standard.

2. BGP-4 Overview

BGP is an inter-autonomous system routing protocol designed for TCP/IP internets. The primary function of BGP is to exchange network reachability information with other BGP systems. This information is sufficient to construct a graph of loop-free AS connectivity and some policy decisions at the AS level may be enforced.
The initial version of the BGP protocol was published in RFC 1105. Since then BGP Versions 2, 3, and 4 have been developed and are specified in [RFC 1163], [RFC 1267], and [RFC 1771], respectively. Changes since BGP-4 went to Draft Standard [RFC 1771] are listed in Appendix N of [BGP4].

2.1. A Border Gateway Protocol

BGP [RFC 1105]

Appendix D of [BGP4] Comparison with 1105:

- o Changes to FSM to accommdate BSD 4.3 TCP UI.
- o Notion of Up/Down/Horizontal relations have been removed.
- o Message format changes:
  - Hold Timer removed from BGP Header and added to OPEN Message
  - Version field removed from BGP Header and added to OPEN Message
  - Link Type field removed from OPEN Message
  - OPEN CONFIRM message deprecated and replaced with implicit confirmation provided by KEEPALIVE message.

- UPDATE Message format changed. New fields were added to support multiple path attributes.
- The Marker field was expanded and its role broadened to support authentication

2.2. BGP version 2

BGPv2 [RFC 1163]

Appendix C of [BGP4] Comparison with RFC 1163

- o BGP Identifier introduced to deal with collision detection.
- o Removed restriction that border router of NEXT_HOP path attribute had to be part of same AS.
- o Optimized and simplified exchange of information about reachable routes.

BGP version 2 removed from the protocol the concept of "up", "down", and "horizontal" relations between autonomous systems that were present in version 1. BGP version 2 introduced the concept of path attributes. In addition, BGP version 2 clarified parts of the protocol that were "under-specified".

2.3. BGP version 3

BGPv3 [RFC 1267]

Appendix B of [BGP4] Comparison with RFC 1267:

- o Set of destination via single IP prefix. Concept of network classes, or subnetting is foreign to BGP-4. To accommodate these capabilities BGP-4 changes the semantics and encoding associated with the AS_PATH attribute. New text has been added to define semantics associated with IP Prefixes. These abilities allow BGP to support the proposed supernetting scheme [RFC 1518] (BGP4 Draft reference to [9] needs to be fixed).
- o LOCAL_PREF intrduced to facilitate route selection procedures.
- o INTER_AS_METRIC renamed to MULTI_EXIT_DISC
- o ATOMIC_AGGREGATE introduced to ensure that certain aggregates are not deaggregated.
- o Introduced AGGREGATOR.

- o Holdtimer neogtiation per-connection for symmetry. Lower value used. Hold Times of zero now supported.

BGP version 3 lifted some of the restrictions on the use of the NEXT_HOP path attribute, and added the BGP Identifier field to the BGP OPEN message. It also clarifies the procedure for distributing BGP routes between the BGP speakers within an autonomous system.

2.4. BGP version 4

BGP v4 [RFC 1771] [BGP4]

Appendix A of [BGP4] Comparison with RFC 1771:

- o Changes to reflect use of the TCP MD5 Signature Option, Route Reflectors, AS Confederations for BGP and BGP Route Refresh.
- o Clarified use of BGP Identifier in AGGREGATOR Attribute
- o Procedures for imposing upper bound of prefixes a speaker will accept from a peer.
- o Ability to include more than one instance of it's own AS in the AS_PATH attribute for the purpose of inter-AS traffic engineering.
- o Clarified various types of NEXT_HOPS
- o Claried use of ATOMIC_AGGREGATE attribute
- o Discussed relationship be BGP NEXT_HOP attribute and immediate next hop.
- o Clarified tie-breaking procedures
- o Clarified route advertisement frequency text.
- o Deprecated Optional Parameter Type 1 (Authentication Information)
- o UPDATE Message Error subcode 7 (AS Routing Loop) deprecated.
- o Use of Marker field for authentication has been deprecated.

BGP version 4 redefines the (previously defined class-based) network layer reachability portion of the updates to specify prefixes of arbitrary length in order to represent multiple classful networks in a single entry as discussed in [RFC 1519]. BGP version 4 has also modified the AS_PATH attribute so that sets of autonomous systems, as well as individual ASs may be described. BGP version 4 has redescribed the INTER-AS METRIC attribute as the MULTI_EXIT_DISC and added new LOCAL_PREF and AGGREGATOR attributes.

BGP version 4 defines procedures for imposing an upper bound on the number of prefixes that a BGP speaker may accept from its peer. BGP version 4 has modifed the AS_PATH attribute to have an ability to

include more than one instance of its own AS for the purpose of inter-AS traffic engineering.

BGP version 4 deprecates the use of OPTIONAL PARAMETER Type 1 (Authentication Information). BGP version 4 also depecates the use of UPDATE MESSAGE Error subcode 7 (AS Routing Loop).

BGP version 4 provides clarifications on use of BGP Identifier in the AGGREGATOR attribute and use of the ATOMIC_AGGREGATOR attribute. BGP version 4 also provides clarifications on various types of NEXT_HOPs, BGP tie-breaking procedures and frequency of route announcements in BGP.

Possible applications of BGP in the Internet are documented in [RFC 1772].

The BGP protocol was developed by the IDR Working Group of the Internet Engineering Task Force. This Working Group had a mailing list, idr@merit.edu, where discussions of protocol features and operation are held. The IDR Working Group meets regularly during the Internet Engineering Task Force meetings. Reports of these meetings are published in the IETF's Proceedings.

3. Management Information Base (MIB)

The BGP-4 Management Information Base (MIB) has been published [BGP-MIB]. The MIB was updated from previous versions documented in [RFC 1657] and [RFC 1269], respectively.

Apart from a few system variables, the BGP MIB is broken into two tables: the BGP Peer Table and the BGP Received Path Attribute Table. The Peer Table reflects information about BGP peer connections, such as their state and current activity. The Received Path Attribute Table contains all attributes received from all peers before local routing policy has been applied. The actual attributes used in determining a route are a subset of the received attribute table.

4. Implementations

There are numerous independent interoperable implementations of BGP

currently available. Although the previous version of this report provided an overview of the implementations currently used in the operational Internet, at this time it has been suggested that a separate BGP Implementation Report [BGP-IMPL] be generated.

It should be noted that implementation experience with Cisco's BGP-4 implementation was documented as part of [RFC 1656].

For all additional implementation information please reference [BGP-IMPL].

5. Operational Experience

This section discusses operational experience with BGP and BGP-4.

BGP has been used in the production environment since 1989, BGP-4 since 1993. Production use of BGP includes utilization of all significant features of the protocol. The present production environment, where BGP is used as the inter-autonomous system routing protocol, is highly heterogeneous. In terms of the link bandwidth it varies from 64 Kbs to 10 Gbs. In terms of the actual routes that run BGP it ranges from a relatively slow performance PC/RT to a very high performance RISC-based CPUs, and includes both the special purpose routers and the general purpose workstations running various UNIX derivatives and other operating systems.

In terms of the actual topologies it varies from a very sparse to quite dense. Full-mesh IBGP topologies have been largely mitigated by BGP Route Reflection [RFC 2796], Autonomous System Confederations for BGP [RFC 3065], or some combination of the two, in many networks.

At the time of this writing BGP-4 is used as an inter-autonomous system routing protocol between ALL Internet-attached autonomous systems, with nearly 15k active autonomous systems in the global Internet routing table.

BGP is used both for the exchange of routing information between a transit and a stub autonomous system, and for the exchange of routing information between multiple transit autonomous systems. There is no protocol distinction between sites historically considered "backbones" versus "regional" networks.

Within transit networks, BGP is typically used as the exclusive carrier of exterior routing information.

The full set of exterior routes that is carried by BGP is well over 115,000 aggregate entries, representing several times that number of connected networks. The number of active paths in some service provider core routers has exceeded 2 million. Native AS_PATH lengths are as long as 10 for some routes, and "padded" path lengths of 25 or more ASs exist.

6. Metrics

This section discusses different metrics used within the BGP protocol. BGP has a seperate metric parameter for IBGP and EBGP. This allows policy based metrics to overwrite the distance based metrics; allowing each autonomous systems to define their independent policies in Intra-AS as well as Inter-AS. BGP Multi Exit Discriminator (MED) is used as a metric by EBGP peers while BGP Local Preference is used by IBGP peers.

6.1. MULTI_EXIT_DISC (MED)

BGP version 4 re-defined the old INTER-AS metric as a MULTI_EXIT_ DISC (MED). This value may be used in the tie-breaking process when selecting a preferred path to a given address space, and provides BGP speakers with the capability to convey to a peer AS the optimal entry point into the local AS.

Although the MED was meant to only be used when comparing paths received from different external peers in the same AS, many implementations provide the capability to compare MEDs between different ASs as well.

The MED was purposely designed to be a "weak" metric that would only be used late in the best-path decision process. The BGP working group was concerned that any metric specified by a remote operator would only affect routing in a local AS if no other preference was specified. A paramount goal of the design of the MED was ensure that peers could not "shed" or "absorb" traffic for networks that they advertise.

6.1.1. Sending MEDs to BGP Peers

[BGP4] allows MEDs received from any EBGP peers by a BGP speaker to be passed to its IBGP peers. Although advertising MEDs to IBGP peers is not a required behavior, it is a common default. MEDs received from EBGP peers by a BGP speaker MUST NOT be sent to other EBGP peers.

6.1.2. MED of Zero Versus No MED

An implementation MUST provide a mechanism that allows for MED to be removed. Previously, implementations did not consider a missing MED value to be the same as a MED of zero. No MED value should now be equal to a value of zero.

6.1.3. MEDs and Temporal Route Selection

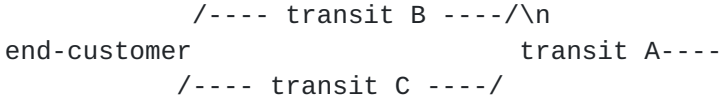
Some implementations have hooks to apply temporal behavior in MED- based best path selection. That is, all other things being equal up to MED consideration, preference would be applied to the "oldest" path, without preferring the lower MED value. The reasoning for this is that "older" paths are presumably more stable, and thus more preferable. However, temporal behavior in route slection results in non-deterministic behavior, and as such, is often undesirable.

7. LOCAL_PREF

The LOCAL_PREF attribute was added so a network operator could easily configure a policy that overrode the standard best path determination mechanism without independently configuring local preference policy on each router.

One shortcoming in the BGP-4 specification was a suggestion for a default value of LOCAL-PREF to be assumed if none was provided. Defaults of 0 or the maximum value each have range limitations, so a common default would aid in the interoperation of multi-vendor routers in the same AS (since LOCAL_PREF is a local administration knob, there is no interoperability drawback across AS boundaries).

LOCAL_PREF MUST be sent to IBGP Peers.
LOCAL_PREF Attribute MUST NOT be sent to EBGP Peers. (Handling errors) Spec says Notification with Error Code UPDATE Message Error.
The common default value for LOCAL_PREF is 100. No default value is defined.
Another area where more exploration is required is a method whereby an originating AS may influence the best path selection process. For example, a dual-connected site may select one AS as a primary transit service provider and have one as a backup.



In a topology where the two transit service providers connect to a third provider, the real decision is performed by the third provider and there is no mechanism for indicating a preference should the third provider wish to respect that preference.
A general purpose suggestion that has been brought up is the possibility of carrying an optional vector corresponding to the AS-PATH where each transit AS may indicate a preference value for a given route. Cooperating ASs may then chose traffic based upon comparison of "interesting" portions of this vector according to routing policy.
While protecting a given ASs routing policy is of paramount concern, avoiding extensive hand configuration of routing policies needs to be examined more carefully in future BGP-like protocols.

8. Internal BGP In Large Autonomous Systems

While not strictly a protocol issue, one other concern has been raised by network operators who need to maintain autonomous systems with a large number of peers. Each speaker peering with an external router is responsible for propagating reachability and path information to all other transit and border routers within that AS. This is typically done by establishing internal BGP connections to all transit and border routers in the local AS.

In a large AS, this leads to a full mesh of TCP connections (n * (n-1)) and some method of configuring and maintaining those connections. BGP does not specify how this information is to be propagated, so alternatives, such as injecting BGP attribute information into the local IGP have been suggested. Also, there is effort underway to develop internal BGP "route reflectors" or a reliable multicast transport of IBGP information which would reduce configuration, memory and CPU requirements of conveying information to all other internal BGP peers.

BGP "Route Reflector" extensions has been defined in RFC 1966 to alleviate the the need for "full mesh" IBGP.

9. Internet Dynamics

As discussed in [BGP4-ANALYSIS], the driving force in CPU and bandwidth utilization is the dynamic nature of routing in the Internet. As the net has grown, the number of route changes per second has increased.

We automatically get some level of damping when more specific NLRI is aggregated into larger blocks, however this isn't sufficient. In Appendix F of [BGP4] are descriptions of damping techniques that should be applied to advertisements. In future specifications of BGP-like protocols, damping methods should be considered for mandatory inclusion in compliant implementations.

Route changes are announced using BGP UPDATE messages. The greatest overhead in advertising UPDATE messages happens whenever route changes to be announced are inefficiently packed.Announcing routing changes sharing common attributes in a single BGP UPDATE message [13.1] also helps save considerable bandwidth.

Persistent BGP errors may cause BGP peers to flap persistently if peer dampening is not implemented. This would result in significant CPU utilization. Implementors may find it useful to implement peer dampening to avoid such persistent peer flapping [BGP4].

10. BGP Routing Information Bases (RIBs)

[BGP4] states "Any local policy which results in routes being added to an Adj-RIB-Out without also being added to the local BGP speaker's

forwarding table, is outside the scope of this document".
However, several well-known implementations do not confirm that Loc-RIB entries were used to populate the forwarding table before installing them in the Adj-RIB-Out. The most common occurrence of this is when routes for a given prefix are presented by more than one protocol and the preferences for the BGP learned route is lower than that of another protocol. As such, the route learned via the other protocol is used to populate the forwarding table.
It may be desirable for an implementation to provide a knob that permits advertisement of "inactive" BGP routes.
It may be also desirable for an implementation to provide a knob that allows a BGP speaker to advertise BGP routes that were not selected by descision process.

11. Update Packing

The BGP4 protocol permits advertisement of multiple prefixes with a common set of path attributes to be advertised in a single update message. When possible, Update packing is recommended as it reduces overhead in receivers in terms of:

- o System overhead due to receipt of fewer Update messages.
- o Overhead for scanning the routing table for BGP Updates to its peers and other routing protocols (and the redistribution of this information as well).

The BGP protocol suggests that withdrawal information should be packed in the begining of Update message along with information about more or less specific reachable routes in a single UPDATE message. This would help alleviate excessive route flapping in BGP.

12. Limit Rate Updates

The BGP protocol defines different mechanisms to rate limit the Updates. The BGP protocol defines MinRouteAdvertisementInterval parameter that determines the minimum time that must be elsape between the advertisement of routes to a particular destination from a single BGP speaker. This value is set on a per BGP peer basis.

13. Ordering of Path Attributes

The BGP protocol suggests that BGP speakers sending multiple prefixes per an UPDATE message should sort and order path attributes according to Type Codes. This would help their peers to quickly identify sets of attributes from different update messages which are semantically different.

Implementers may find it useful to order path attributes according to Type Code so that sets of attributes with identical semantics can be more quickly identified.

14. AS_SET Sorting

AS_SETs are commonly used in BGP route aggregation. They reduce the size of AS_PATH information by listing AS numbers only once regardless of any number of times it might appear in process of aggregation. AS_SETs are usually sorted in increasing order to facilitate efficient lookups of AS numbers within them. This optimization is entirely optional.

15. Control over Version Negotiation

Because pre-BGP-4 route aggregation can't be supported by earlier version of BGP, an implementation that supports versions in addition to BGP-4 should provide the version support on a per-peer basis.

16. Reciept of Non-Transitive Attributes from eBGP Peer

E.g., LOCAL_PREF, RR or confed, etc..

NEEDS MORE WORK

17. Security Considerations

BGP provides flexible and extendible mechanism for authentication and security. The mechanism allows to support schemes with various degree of complexity. All BGP sessions are authenticated based on the BGP Identifier of a peer. In addition, all BGP sessions are authenticated based on the autonomous system number advertised by a peer.

Since BGP runs over TCP and IP, BGP's authentication scheme may be augmented by any authentication or security mechanism provided by either TCP or IP.

17.1. TCP MD5 Signature Option

RFC 2385 defines a way in which the TCP MD5 signature option can be used to valid information transmitted between two peers. This method prevents any third party from injecting information (e.g., a TCP RST) into the datastream, or modifying the routing information carried between two BGP peers. RFC ???? provides suggestions for choosing passwords to be used with MD5.

TCP MD5 is not ubiquitously deployed at the moment, especially in inter- domain scenarios, largely because of key distribution issues. Most key distribution mechanisms are considered to be too "heavy" at this point.

17.2. BGP Over IPSEC

BGP can run over IPSEC, either in a tunnel, or in transport mode, where the TCP portion of the IP packet is encrypted. This not only prevents random insertion of information into the data stream between two BGP peers, it also prevents an attacker from learning the data which is being exchanged between the peers.

IPSEC does, however, offer several options for exchanging session keys, which may be useful on inter-domain configurations. These options are being explored in many deployments, although no definitive solution has been reach on the issue of key exchange for BGP in IPSEC.

It should be noted that since BGP runs over TCP and IP, BGP is vulnerable to the same denial of service or authentication attacks that are present in any other TCP based protocol.

17.3. Miscellaneous

Another issue any routing protocol faces is providing evidence of the validity and authority of the routing information carried within the routing system. This is currently the focus of several efforts at the moment, including efforts to define the threats which can be used against this routing information in BGP [draft-murphy, attack tree], and efforts at providing a means to provide validation and authority for routing information carried within BGP [SBGP, soBGP and possibly ATT stuff].

Should information as it relates to IRR derived prefix-based filters be included here?

BTSH & RP Queuing...?

17.4. Acknowledgements

We would like to thank Paul Traina and Yakov Rekhter for authoring previous versions of this document. We would also like to acknowledge Russ White for valuable feedback on this document.

18. References

[RFC 1105] Lougheed, K., and Rekhter, Y, "Border Gateway Protocol BGP", RFC 1105, June 1989.

[RFC 1163] Lougheed, K., and Rekhter, Y, "Border Gateway Protocol BGP", RFC 1105, June 1990.

[RFC 1264] Hinden, R., "Internet Routing Protocol Standardization Criteria", RFC 1264, October 1991.

[RFC 1267] Lougheed, K., and Rekhter, Y, "Border Gateway Protocol 3 (BGP-3)", RFC 1105, October 1991.

[RFC 1519] Fuller, V., Li. T., Yu J., and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", RFC 1519, September 1993.

[RFC 1656] Traina, P., "BGP-4 Protocol Document Roadmap and Implementation Experience", RFC 1656, July 1994.

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

[RFC 1772] Rekhter, Y., and P. Gross, Editors, "Application of the Border Gateway Protocol in the Internet", RFC 1772, March 1995.

[RFC 1773] Traina, P., "Experience with the BGP-4 protocol", RFC 1773, March 1995.

[RFC 2796] Bates, T., Chandra, R., and Chen, E, "Route Reflection - An Alternative to Full Mesh IBGP", RFC 2796, April 2000.

[RFC 3065] Traina, P., McPherson, D., and Scudder, J, "Autonomous System Confederations for BGP", RFC 3065, Febuary 2001.

[RFC 3345] McPherson, D., Gill, V., Walton, D., and Retana, A, "BGP Persistent Route Oscillation Condition", RFC 3345, August 2002.

[BGP4-ANALYSIS] Work in Progress.

[BGP4-IMPL] Work in Progress.

[BGP4] Rekhter, Y., T. Li., and Hares. S, Editors, "A Border Gateway Protocol 4 (BGP-4)", BGP Draft, Work in Progress.

19. Authors' Addresses

Danny McPherson
Arbor Networks
Email: danny@arbor.net
Keyur Patel
Cisco Systems
Email: keyupate@cisco.com

20. Full Copyright Statement