BGP-4 Protocol Analysis
<draft-ietf-idr-bgp-analysis-00.txt>

Status of this Document

This document is a product of an individual. Comments are solicited and should be addressed to the author(s).

Copyright Notice

Abstract

The purpose of this report 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 [RFC1264]. In order to fulfill the requirement, this report augments RFC 1774 [RFC1774] and summarizes the key features of BGP protocol, and analyzes the protocol with respect to scaling and performance.

Table of Contents

1. Introduction

BGP-4 is an inter-autonomous system routing protocol designed for TCP/IP internets. Version 1 of the BGP protocol was published in RFC 1105 [RFC1105]. Since then BGP versions 2, 3, and 4 have been developed. Version 2 was documented in RFC 1163 [RFC1163]. Version 3 is documented in RFC 1267 [RFC1267]. Version 4 is documented in the [BGP4]. The changes between versions are explained in Appendix A of [BGP4]. Possible applications of BGP in the Internet are documented in [RFC1772].

2. Key Features and algorithms of the BGP-4 protocol

This section summarizes the key features and algorithms of the BGP protocol. BGP is an inter-autonomous system routing protocol; it is designed to be used between multiple autonomous systems. BGP assumes that routing within an autonomous system is done by an intra-autonomous system routing protocol. BGP does not make any assumptions about intra-autonomous system routing protocols deployed within the various autonomous systems. Specifically, BGP does not require all autonomous systems to run the same intra-autonomous system routing protocol (i.e., interior gateway protocol or IGP).

Finally, note that BGP is a real inter-autonomous system routing protocol, and as such it imposes no constraints on the underlying Internet topology. The information exchanged via BGP is sufficient to construct a graph of autonomous systems connectivity from which routing loops may be pruned and many routing policy decisions at the autonomous system level may be enforced.

2.1. Key Features

The key features of the protocol are the notion of path attributes and aggregation of network layer reachability information (NLRI). Path attributes provide BGP with flexibility and expandability. Path attributes are partitioned into well-known and optional. The provision for optional attributes allows experimentation that may involve a group of BGP routers without affecting the rest of the Internet. New optional attributes can be added to the protocol in much the same way that new options are added to, say, the Telnet protocol [RFC854].

One of the most important path attributes is the AS-PATH. As reachability information traverses the Internet, this information is augmented by the list of autonomous systems that have been traversed thus far, forming the AS-PATH. The AS-PATH allows straightforward suppression of the looping of routing information. In addition, the AS-PATH serves as a powerful and versatile mechanism for policy-based routing.

BGP-4 enhances the AS-PATH attribute to include sets of autonomous systems as well as lists. This extended format allows generated aggregate routes to carry path information from the more specific routes used to generate the aggregate. It should be noted however, that as of this writing, AS-SETs are in rarely used in the Internet [ROUTEVIEWS].

2.2. BGP Algorithms

BGP uses an algorithm that cannot be classified as either a pure distance vector, or a pure link state. Carrying a complete AS path in the AS-PATH attribute allows to reconstruct large portions of the overall topology. That makes it similar to the link state algorithms. Exchanging only the currently used routes between the peers makes it similar to the distance vector algorithms.

BGP-4 uses an incremental update strategy in order To conserve bandwidth and processing power. That is, after initial exchange of complete routing information, a pair of BGP routers exchanges only changes (deltas) to that information. Such an incremental update design requires reliable transport between a pair of BGP routers to function correctly. BGP uses TCP as its reliable transport.

In addition to incremental updates, BGP-4 has added the concept of route aggregation so that information about groups of networks may be aggregated and sent as a single Network Layer Reachability (NLRI) Attribute.

Finally, note that BGP is a self-contained protocol. That is, it specifies how routing information is exchanged both between BGP speakers in different autonomous systems, and between BGP speakers within a single autonomous system.

2.3. BGP Finite State Machine (FSM)

The BGP FSM is a set of rules that are applied to a BGP speaker's set of configured peers for the BGP operation. A BGP implementation requires that a BGP speaker must connect and listen to tcp port 179 for accepting any new BGP connections from it's peers. The BGP FSM must be initiated and maintained for each new incoming and outgoing peer connections. However, in steady state operation, there will be only one BGP FSM per connection per peer.

There may exist a temporary period where in a BGP peer may have separate incoming and outgoing connections resulting into two different BGP FSMs for a peer (instead of one). This can be resolved following BGP connection collision rules defined in the [BGP4].

Following are different states of BGP FSM for its peers:

IDLE: State when BGP peer refuses any incoming connections.

CONNECT: State in which BGP peer is waiting for its TCP connection to be completed.

ACTIVE: State in which BGP peer is trying to acquire a peer by listening and accepting TCP connection.

OPENSENT: BGP peer is waiting for OPEN message from its peer.

OPENCONFIRM: BGP peer is waiting for KEEPALIVE or NOTIFICATION message from its peer.

ESTABLISHED: BGP peer connection is established and exchanges UPDATE, NOTIFICATION, and KEEPALIVE messages with its peer.

3. BGP Performance characteristics and Scalability

In this section, we provide "order of magnitude" answers to the questions of how much link bandwidth, router memory and router CPU cycles the BGP protocol will consume under normal conditions. In particular, we will address the scalability of BGP and its limitations.

It is important to note that BGP does not require all the routers within an autonomous system to participate in the BGP protocol. In particular, only the border routers that provide connectivity between the local autonomous system and their adjacent autonomous systems need participate in BGP. Constraining this set of participants is just one way of addressing the scaling issue.

3.1. Link bandwidth and CPU utilization

Immediately after the initial BGP connection setup, the peers exchange complete set of routing information. If we denote the total number of routes in the Internet by N, the mean AS distance of the Internet by M (distance at the level of an autonomous system, expressed in terms of the number of autonomous systems), the total number of autonomous systems in the Internet by A, and assume that the networks are uniformly distributed among the autonomous systems, then the worst case amount of bandwidth consumed during the initial exchange between a pair of BGP speakers is

MR = O(N + M * A)

The following table illustrates the typical amount of bandwidth consumed during the initial exchange between a pair of BGP speakers based on the above assumptions (ignoring bandwidth consumed by the BGP Header). For purposes of the estimates here, we will calculate MR = 4 * (N + (M * A)).

# NLRI	Mean AS Distance	# AS's	Bandwidth (MR)
40,000	15	400	184,000 bytes
100,000	10	10,000	800,000 bytes
120,000	10	15,000	1,080,000 bytes
140,000	15	20,000	1,760,000 bytes

[note that most of this bandwidth is consumed by the NLRI exchange]

BGP-4 was created specifically to reduce the size of the set of NLRI entires carried and exchanged by border routers. The aggregation scheme, defined in RFC 1519 [RFC1519], describes the provider-based aggregation scheme in use in today's Internet.

Due to the advantages of advertising a few large aggregate blocks instead of many smaller class-based individual networks, it is difficult to estimate the actual reduction in bandwidth and

processing that BGP-4 has provided over BGP-3. If we simply enumerate all aggregate blocks into their individual class-based networks, we would not take into account "dead" space that has been reserved for future expansion. The best metric for determining the success of BGP-4's aggregation is to sample the number NLRI entries in the globally connected Internet today and compare it to projected growth rates before BGP-4 was deployed.

At the time of this writing, the full set of exterior routes carried by BGP is approximately 120,000 network entries [ROUTEVIEWS].

3.1.1. CPU utilization

An important (and fundamental) feature of BGP is that BGP's CPU utilization depends only on the stability of the Internet. If the Internet is stable, then the only link bandwidth and router CPU cycles consumed by BGP are due to the exchange of the BGP KEEPALIVE messages. The KEEPALIVE messages are exchanged only between peers. The suggested frequency of the exchange is 30 seconds. The KEEPALIVE messages are quite short (19 octets), and require virtually no processing. Therefore, the bandwidth consumed by the KEEPALIVE messages is about 5 bits/sec. Operational experience confirms that the overhead (in terms of bandwidth and CPU) associated with the KEEPALIVE messages should be viewed as negligible.

During periods of Internet instability, changes to the reachability information are passed between routers in UPDATE messages. If we denote the number of routing changes per second by C, then in the worst case the amount of bandwidth consumed by the BGP can be expressed as O(C * M). The greatest overhead per UPDATE message occurs when each UPDATE message contains only a single network. It should be pointed out that in practice routing changes exhibit strong locality with respect to the AS path. That is routes that change are likely to have common AS path. In this case multiple networks can be grouped into a single UPDATE message, thus significantly reducing the amount of bandwidth required (see also Appendix F.1 of [BGP4]).

Since in the steady state the link bandwidth and router CPU cycles consumed by the BGP protocol are dependent only on the stability of the Internet, it follows that BGP should have no scaling problems in the areas of link bandwidth and router CPU utilization. This assumes that as the Internet grows, the overall stability of the inter-AS connectivity of the Internet can be controlled. In particular, while the size of the IPv4 Internet routing table is bounded by O(2^32 * M), (where M is a slow-moving function describing the AS

interconnectivity of the network), no such bound can be formulated for the dynamic properties (i.e., stability) of BGP. Finally, since the dynamic properties of the network cannot be quantitatively bounded, stability must be addressed via heuristics such as BGP Route Flap Dampening [RFC2439]. Due to the nature of BGP, such dampening should be viewed as a local to an autonomous system matter (see also Appendix F.2 of [BGP4]).

Growth of the Internet has made the stability issue one of the most crucial issues for Internet operations. BGP by itself does not introduce any instabilities into the Internet. Rather, instabilities are largely due to the the dynamic nature of the edges of the network, coupled with less than optimal aggregation. As a result, stability should be addressed through improved aggregation and isolating the core of the network from the dynamic nature of the edge networks.

It may also be instructive to compare bandwidth and CPU requirements of BGP with EGP. While with BGP the complete information is exchanged only at the connection establishment time, with EGP the complete information is exchanged periodically (usually every 3 minutes). Note that both for BGP and for EGP the amount of information exchanged is roughly on the order of the networks reachable via a peer that sends the information. Therefore, even if one assumes extreme instabilities of BGP, its worst case behavior will be the same as the steady state behavior of it's predecessor, EGP.

Operational experience with BGP showed that the incremental update approach employed by BGP presents an enormous improvement both in the area of bandwidth and in the CPU utilization, as compared with complete periodic updates used by EGP (see also presentation by Dennis Ferguson at the Twentieth IETF, March 11-15, 1991, St.Louis).

3.1.2. Memory requirements

To quantify the worst case memory requirements for BGP, denote the total number of networks in the Internet by N, the mean AS distance of the Internet by M (distance at the level of an autonomous system, expressed in terms of the number of autonomous systems), the total number of autonomous systems in the Internet by A, and the total number of BGP speakers that a system is peering with by K (note that K will usually be dominated by the total number of the BGP speakers within a single autonomous system). Then the worst case memory requirements (MR) can be expressed as

MR = O((N + M * A) * K)

It is interesting to note that prior to the introduction of BGP in the NSFNET Backbone, memory requirements on the NSFNET Backbone routers running EGP were on the order of O(N *K). Therefore, the extra overhead in memory incurred by the modern routers running BGP is less than 7 percent.

Since a mean AS distance M is a slow moving function of the interconnectivity ("meshiness") of the Internet, for all practical purposes the worst case router memory requirements are on the order of the total number of networks in the Internet times the number of peers the local system is peering with. We expect that the total number of networks in the Internet will grow much faster than the average number of peers per router. As a result, scaling with respect to the memory requirements is going to be heavily dominated by the factor that is linearly proportional to the total number of networks in the Internet.

The following table illustrates typical memory requirements of a router running BGP. It is assumed that each network is encoded as four bytes, each AS is encoded as two bytes, and each networks is reachable via some fraction of all of the peers (# BGP peers/per net). For purposes of estimates here, we will calculate MR = ((N*4) + (M*A)*2) * K.

# Networks	Mean AS Distance	# AS's	# BGP peers/per net	Memory Req (MR)
100,000	20	3,000	20	1,040,000
100,000	20	15,000	20	1,040,000
120,000	10	15,000	100	75,000,000
140,000	15	20,000	100	116,000,000

In analyzing BGP's memory requirements, we focus on the size of the forwarding table (ignoring implementation details). In particular, we derive upper bounds for the size of the forwarding table. For example, at the time of this writing, the forwarding tables of a typical backbone router carries on the order of 120,000 entries. Given this number, one might ask whether it would be possible to have a functional router with a table that will have 1,000,000 entries. Clearly the answer to this question is independent of BGP. On the other hand the answer to the original questions (that was asked with respect to BGP) is directly related to the latter question. Very interesting comments were given by Paul Tsuchiya in his review of BGP in March of 1990 (as part of the BGP review committee appointed by

Bob Hinden). In the review he said that, "BGP does not scale well. This is not really the fault of BGP. It is the fault of the flat IP address space. Given the flat IP address space, any routing protocol must carry network numbers in its updates." With the introduction of CIDR [RFC1519] and BGP-4 [BGP4], we have attempted to reduce this limitation. Unfortunately, we cannot erase history nor can BGP-4 solve the problems inherent with inefficient assignment of future address blocks.

To reiterate, BGP limits with respect to the memory requirements are directly related to the underlying Internet Protocol (IP), and specifically the addressing scheme employed by IP. BGP would provide much better scaling in environments with more flexible addressing schemes. It should be pointed out that with only very minor additions BGP was extended to support hierarchies of autonomous system [KUZINGER]. Such hierarchies, combined with an addressing scheme that would allow more flexible address aggregation capabilities, can be utilized by BGP-like protocols, thus providing practically unlimited scaling capabilities.

4. Applicability

In this section we answer the question of which environments is BGP well suited, and for which environments it is not suitable. Partially this question is answered in the Section 2 of [RFC1771], where the document states the following:

> "To characterize the set of policy decisions that can be enforced using BGP, one must focus on the rule that an AS advertises to its neighbor ASs only those routes that it itself uses. This rule reflects the "hop-by-hop" routing paradigm generally used throughout the current Internet. Note that some policies cannot be supported by the "hop-by-hop" routing paradigm and thus require techniques such as source routing to enforce. For example, BGP does not enable one AS to send traffic to a neighbor AS intending that the traffic take a different route from that taken by traffic originating in the neighbor AS. On the other hand, BGP can support any policy conforming to the "hop-by-hop" routing paradigm. Since the current Internet uses only the "hop-by-hop" routing paradigm and since BGP can support any policy that conforms to that paradigm, BGP is highly applicable as an inter-AS routing protocol for the current Internet."

Importantly, the BGP protocol contains only the functionality that is essential, while at the same time provides flexible mechanisms within the protocol itself that allow to expand its functionality. For example, BGP capabilities provide an easy and flexible way to introduce new features within the protocol. Finally, since BGP was designed with flexibility and expandability in mind, new or evolving requirements can be addressed via existing mechanisms. The existence proof of this statement may be found in the way how new features (like repairing a partitioned autonomous system with BGP) are already introduced in the protocol.

To summarize, BGP is well suitable as an inter-autonomous system routing protocol for the IPv4 Internet that is based on IP [RFC791] as the Internet Protocol and "hop-by-hop" routing paradigm. Finally, there is no reason to believe that BGP should not be equally applicable to IPv6 [RFC2460].

5. Acknowledgments

We would like to thank Paul Traina for authoring previous versions of this document.

6. References

[BGP4] Rekhter, Y., T. Li., and Hares. S, Editors, "A Border Gateway Protocol 4 (BGP-4)", draft-ietf-idr-bgp4-19.txt. Work in progress.

[KUZINGER] ISO/IEC 10747, Kunzinger, C., Editor, "Inter-Domain Routing Protocol", October 1993.

[RFC791] "INTERNET PROTOCOL", DARPA INTERNET PROGRAM PROTOCOL SPECIFICATION, RFC 791, September, 1981.

[RFC854] Postel, J. and Reynolds, J., "TELNET PROTOCOL SPECIFICATION", RFC 854, May, 1983.

[RFC1105] Lougheed, K., and Rekhter, Y, "Border Gateway Protocol BGP", RFC 1105, June 1989.

[RFC1163] Lougheed, K., and Rekhter, Y, "Border Gateway Protocol BGP", RFC 1105, June 1990.

[RFC1264] Hinden, R., "Internet Routing Protocol Standardization Criteria", RFC 1264, October 1991.

[RFC1267] Lougheed, K., and Rekhter, Y, "Border Gateway Protocol 3 (BGP-3)", RFC 1105, October 1991.

[RFC1519] Fuller, V., Li. T., Yu J., and K. Varadhan, "Classless Inter-Domain Routing (CIDR): an Address Assignment and Aggregation Strategy", RFC 1519, September 1993.

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

[RFC1772] Rekhter, Y., and P. Gross, Editors, "Application of the Border Gateway Protocol in the Internet", RFC 1772, March 1995.

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

[RFC2460] Deering, S, and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December, 1998.

[ROUTEVIEWS] Meyer, David, "The Route Views Project", http://www.routeviews.org

7. Author's Address

David Meyer
Email: dmm@maoz.com

Keyur Patel
Cisco Systems
Email: keyupate@cisco.com

8. Full Copyright Statement