Network Working Group
Request for Comments: 1774
Category: Informational

P. Traina, Editor
cisco Systems

March 1995

BGP-4 Protocol Analysis

Status of this Memo

This memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

Introduction

The purpose of this report is to document how the requirements for advancing a routing protocol to Draft Standard have been satisfied by the Border Gateway Protocol version 4 (BGP-4). This report summarizes the key features of BGP, and analyzes the protocol with respect to scaling and performance. This is the first of two reports on the BGP protocol.

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. Since then BGP versions 2, 3, and 4 have been developed. Version 2 was documented in RFC 1163. Version 3 is documented in RFC1267. The changes between versions are explained in Appendix 2 of [1].

Possible applications of BGP in the Internet are documented in [2].

Please send comments to iwg@ans.net.

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 employed by the various autonomous systems. Specifically, BGP does not require all autonomous systems to run the same intra-autonomous system routing protocol.

BGP is a real inter-autonomous system routing protocol. 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 some

routing policy decisions at the autonomous system level may be enforced.

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 fashion as new options are added to the Telnet protocol, for instance.

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.

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.

To conserve bandwidth and processing power, BGP uses incremental updates, where after the initial exchange of complete routing information, a pair of BGP routers exchanges only changes (deltas) to that information. Technique of incremental updates requires reliable transport between a pair of BGP routers. To achieve this functionality BGP uses TCP as its transport.

In addition to incremental updates, BGP-4 has added the concept of route aggregation so that information about groups of networks may represented as a single entity.

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.

To allow graceful coexistence with EGP and OSPF, BGP provides support for carrying both EGP and OSPF derived exterior routes BGP also allows to carry statically defined exterior routes or routes derived by other IGP information.

BGP performance characteristics and scalability

In this section we'll try to answer the question of how much link bandwidth, router memory and router CPU cycles does the BGP protocol consume under normal conditions. We'll also address the scalability of BGP, and look at some of its limits.

BGP does not require all the routers within an autonomous system to participate in the BGP protocol. Only the border routers that provide connectivity between the local autonomous system and its adjacent autonomous systems participate in BGP. Constraining the set of participants is just one way of addressing the scaling issue.

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 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).

# NLRI	Mean AS Distance	# AS's	Bandwidth
10,000	15	300	49,000 bytes
20,000	8	400	86,000 bytes *
40,000	15	400	172,000 bytes
100,000	20	3,000	520,000 bytes

* the actual "size" of the Internet at the the time of this document's publication

Note that most of the bandwidth is consumed by the exchange of the Network Layer Reachability Information (NLRI).

BGP-4 was created specifically to reduce the amount of NLRI entries carried and exchanged by border routers. BGP-4, along with CIDR [4] has introduced the concept of the "Supernet" which describes a power-of-two aggregation of more than one class-based network.

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 BGP3. 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.

In January of 1994, router carrying a full routing load for the globally connected Internet had approximately 19,000 network entries (this number is not exact due to local policy variations). The BGP deployment working group estimated that the growth rate at that time was over 1000 new networks per month and increasing. Since the widespread deployment of BGP-4, the growth rate has dropped significantly and a sample done at the end of November 1994 showed approximately 21,000 entries present, as opposed to the expected 30,000.

CPU cycles consumed by BGP 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. If the Internet is unstable, then only the changes to the reachability information (that are caused by the instabilities) are passed between routers (via the 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 6.1 of [1]).

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, but are completely independent on the number of networks that compose the Internet, it follows that BGP should have no scaling problems in the areas of link bandwidth and router CPU utilization, as the Internet grows, provided that the overall stability of the inter-AS connectivity (connectivity between ASs) of the Internet can be controlled. Stability issue could be addressed by introducing some form of dampening (e.g., hold downs). Due to the nature of BGP, such dampening should be viewed as a local to an autonomous system matter (see also Appendix 6.3 of [1]). It is important to point out, that regardless of BGP, one should not underestimate the significance of the stability in the Internet.

Growth of the Internet has made the stability issue one of the most crucial ones. It is important to realize that BGP, by itself, does not introduce any instabilities in the Internet. Current observations in the NSFNET show that the instabilities are largely due to the ill-behaved routing within the autonomous systems that compose the Internet. Therefore, while providing BGP with mechanisms to address the stability issue, we feel that the right way to handle the issue is to address it at the root of the problem, and to come up with intra-autonomous routing schemes that exhibit reasonable stability.

It also may 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 (see also Section 5.2). Therefore, even if one assumes extreme instabilities of BGP, its worst case behavior will be the same as the steady state behavior of EGP.

Operational experience with BGP showed that the incremental updates 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).

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

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

In the current NSFNET Backbone (N = 2110, A = 59, and M = 5) if each network is stored as 4 octets, and each autonomous system is stored as 2 octets then the overhead of storing the AS path information (in addition to the full complement of exterior routes) is less than 7 percent of the total memory usage.

It is interesting to point out, 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 NSFNET routers after the introduction of BGP is less than 7 percent.

Since a mean AS distance grows very slowly with the total number of networks (there are about 60 autonomous systems, well over 2,000 networks known in the NSFNET backbone routers, and the mean AS distance of the current Internet is well below 5), 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. Therefore, 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 4 bytes, each AS is encoded as 2 bytes, and each networks is reachable via some fraction of all of the peers (# BGP peers/per net).

# Networks	Mean AS Distance	# AS's	# BGP peers/per net	Memory Req
2,100	5	59	3	27,000
4,000	10	100	6	108,000
10,000	15	300	10	490,000
100,000	20	3,000	20	1,040,000

To put memory requirements of BGP in a proper perspective, let's try to put aside for a moment the issue of what information is used to construct the forwarding tables in a router, and just focus on the forwarding tables themselves. In this case one might ask about the limits on these tables. For instance, given that right now the forwarding tables in the NSFNET Backbone routers carry well over 20,000 entries, one might ask whether it would be possible to have a functional router with a table that will have 200,000 entries. Clearly the answer to this question is completely 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 [4] and BGP-4, 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 [8]. 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.

Applicability of BGP

In this section we'll try to answer the question of what environment is BGP well suited, and for what is it not suitable? Partially this question is answered in the Section 2 of [1], 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."

While BGP is well suitable for the current Internet, it is also almost a necessity for the current Internet as well. Operational experience with EGP showed that it is highly inadequate for the current Internet. Topological restrictions imposed by EGP are unjustifiable from the technical point of view, and unenforceable from the practical point of view. Inability of EGP to efficiently handle information exchange between peers is a cause of severe routing instabilities in the operational Internet. Finally, information provided by BGP is well suitable for enforcing a variety of routing policies.

Rather than trying to predict the future, and overload BGP with a variety of functions that may (or may not) be needed, the designers of BGP took a different approach. The 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. Since BGP was designed with flexibility and expandability in mind, we think it should be able to address new or evolving requirements with relative ease. 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 current Internet that is based on IP (RFC 791) as the Internet Protocol and "hop-by-hop" routing paradigm. It is hard to speculate whether BGP will be suitable for other environments where internetting is done by other than IP protocols, or where the routing paradigm will be different.

Security Considerations

Security issues are not discussed in this memo.

Acknowledgments

The BGP-4 protocol has been developed by the IDR/BGP Working Group of the Internet Engineering Task Force. I would like to express thanks to Yakov Rekhter for providing RFC 1265. This document is only a minor update to the original text. I'd also like to explicitly thank Yakov Rekhter and Tony Li for their review of this document as well as their constructive and valuable comments.

Editor's Address

Paul Traina
cisco Systems, Inc.
170 W. Tasman Dr.
San Jose, CA 95134

EMail: pst@cisco.com

References

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

[2] Rekhter, Y., and P. Gross, Editors, "Application of the Border Gateway Protocol in the Internet", RFC 1772, T.J. Watson Research Center, IBM Corp., MCI, March 1995.

[3] Willis, S., Burruss, J., and J. Chu, "Definitions of Managed Objects for the Fourth Version of the Border Gateway Protocol (BGP-4) using SMIv2", RFC 1657, Wellfleet Communications Inc., IBM Corp., July 1994.

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

[6] Moy J., "Open Shortest Path First Routing Protocol (Version 2)", RFC 1257, Proteon, August 1991.

[7] Varadhan, K., Hares S., and Y. Rekhter, "BGP4/IDRP for IP---OSPF Interaction", Work in Progress.

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