Network Working Group Y. Rekhter, Editor
Request for Comments: 1266 T.J. Watson Research Center, IBM Corp.
October 1991

Experience with the BGP Protocol

1. Status of this Memo.

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

2. Introduction.

The purpose of this memo is to document how the requirements for advancing a routing protocol to Draft Standard have been satisfied by Border Gateway Protocol (BGP). This report documents experience with BGP. This is the second of two reports on the BGP protocol. As required by the Internet Activities Board (IAB) and the Internet Engineering Steering Group (IESG), the first report will present a performance analysis of the BGP protocol.

The remaining sections of this memo document how BGP satisfies General Requirements specified in Section 3.0, as well as Requirements for Draft Standard specified in Section 5.0 of the "Internet Routing Protocol Standardization Criteria" document [1].

This report is based on the work of Dennis Ferguson (University of Toronto), Susan Hares (MERIT/NSFNET), and Jessica Yu (MERIT/NSFNET). Details of their work were presented at the Twentieth IETF meeting (March 11-15, 1991, St. Louis) and are available from the IETF Proceedings.

Please send comments to iwg@rice.edu.

3. Acknowledgements.

The BGP protocol has been developed by the IWG/BGP Working Group of the Internet Engineering Task Force. We would like to express our deepest thanks to Guy Almes (Rice University) who was the previous chairman of the IWG Working Group. We also like to explicitly thank Bob Hinden (BBN) for the review of this document as well as his constructive and valuable comments.

4. Documentation.

BGP is an inter-autonomous system routing protocol designed for the TCP/IP internets. Version 1 of the BGP protocol was published in RFC 1105. Since then BGP Versions 2 and 3 have been developed. Version 2 was documented in RFC 1163. Version 3 is documented in [3]. The changes between versions 1, 2 and 3 are explained in Appendix 3 of [3]. Most of the functionality that was present in the Version 1 is present in the Version 2 and 3. Changes between Version 1 and Version 2 affect mostly the format of the BGP messages. Changes between Version 2 and Version 3 are quite minor.

BGP Version 2 removed from the protocol the concept of "up", "down", and "horizontal" relations between autonomous systems that were present in the Version 1. BGP Version 2 introduced the concept of path attributes. In addition, BGP Version 2 clarified parts of the protocol that were "underspecified". 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. Possible applications of BGP in the Internet are documented in [2].

The BGP protocol was developed by the IWG/BGP Working Group of the Internet Engineering Task Force. This Working Group has a mailing list, iwg@rice.edu, where discussions of protocol features and operation are held. The IWG/BGP Working Group meets regularly during the quarterly Internet Engineering Task Force conferences. Reports of these meetings are published in the IETF’s Proceedings.

5. MIB

A BGP Management Information Base has been published [4]. The MIB was written by Steve Willis (swillis@wellfleet.com) and John Burruss (jburruss@wellfleet.com).

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.

The BGP MIB is quite small. It contains total of 27 objects.

6. Security architecture.

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. As part of the BGP authentication mechanism, the protocol allows to carry encrypted digital signature in every BGP message. All authentication failures result in sending the NOTIFICATION messages and immediate termination of the BGP connection.

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.

7. Implementations.

There are multiple interoperable implementations of BGP currently available. This section gives a brief overview of the three completely independent implementations that are currently used in the operational Internet. They are:

- cisco. This implementation was wholly developed by cisco. It runs on the proprietary operating system used by the cisco routers. Consult Kirk Lougheed (lougheed@cisco.com) for more details.

- "gated". This implementation was developed wholly by Jeff Honig (jch@risci.cit.cornell.edu) and Dennis Ferguson (dennis@CAnet.CA). It runs on a variety of operating systems (4.3 BSD, AIX, etc...). It is the only available public domain code for BGP. Consult Jeff Honig or Dennis Ferguson for more details.

- NSFNET. This implementation was developed wholly by Yakov Rekhter (yakov@watson.ibm.com). It runs on the T1 NSFNET Backbone and T3 NSFNET Backbone. Consult Yakov Rekhter for more details.

To facilitate efficient BGP implementations, and avoid commonly made mistakes, the implementation experience with BGP in "gated" was documented as part of RFC 1164. Implementors are strongly encouraged to follow the implementation suggestions outlined in that document.

Experience with implementing BGP showed that the protocol is relatively simple to implement. On the average BGP implementation takes about 1 man/month effort.

Note that, as required by the IAB/IESG for Draft Standard status, there are multiple interoperable completely independent implementations, namely those from cisco, "gated", and IBM.

8. Operational experience.

This section discusses operational experience with BGP.

BGP has been used in the production environment since 1989. This use involves all three implementations listed above. 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 56 Kbits/sec to 45 Mbits/sec. In terms of the actual routes that run BGP it ranges from a relatively slow performance PC/RT to a very high performance RS/6000, and includes both the special purpose routers (cisco) and the general purpose workstations running UNIX. In terms of the actual topologies it varies from a very sparse (spanning tree or a ring of CA*Net) to a quite dense (T1 or T3 NSFNET Backbones).

At the time of this writing BGP is used as an inter-autonomous system routing protocol between the following autonomous systems: CA*Net, T1 NSFNET Backbone, T3 NSFNET Backbone, T3 NSFNET Test Network, CICNET, MERIT, and PSC. Within CA*Net there are 10 border routers participating in BGP. Within T1 NSFNET Backbone there are 20 border routers participating in BGP. Within T3 NSFNET Backbone there are 15 border routers participating in BGP. Within T3 NSFNET Test Network there are 7 border routers participating in BGP. Within CICNET there are 2 border routers participating in BGP. Within MERIT there is 1 border router participating in BGP. Within PSC there is 1 router participating in BGP. All together there are 56 border routers spanning 7 autonomous systems that are running BGP. Out of these, 49 border routers that span 6 autonomous systems are part of the operational Internet.

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. It covers both the Backbones (CA*Net, T1 NSFNET Backbone, T3 NSFNET Backbone), and the Regional Networks (PSC, MERIT).

Within CA*Net, T3 NSFNET Backbone, and T3 NSFNET Test Network BGP is used as the exclusive carrier of the exterior routing information both between the autonomous systems that correspond to the above networks, and with the autonomous system of each network. At the time of this writing within the T1 NSFNET Backbone BGP is used together with the NSFNET Backbone Interior Routing Protocol to carry the

exterior routing information. T1 NSFNET Backbone is in the process of moving toward carrying the exterior routing information exclusively by BGP. The full set of exterior routes that is carried by BGP is well over 2,000 networks.

Operational experience described above involved multi-vendor deployment (cisco, "gated", and NSFNET).

Specific details of the operational experience with BGP in the NSFNET were presented at the Twentieth IETF meeting (March 11-15, 1991, St. Louis) by Susan Hares (MERIT/NSFNET). Specific details of the operational experience with BGP in the CA*Net were presented at the Twentieth IETF meeting (March 11-15, 1991, St. Louis) by Dennis Ferguson (University of Toronto). Both of these presentations are available in the IETF Proceedings.

Operational experience with BGP exercised all basic features of the protocol, including the authentication and routing loop suppression.

Bandwidth consumed by BGP has been measured at the interconnection points between CA*Net and T1 NSFNET Backbone. The results of these measurements were presented by Dennis Ferguson during the last IETF, and are available from the IETF Proceedings. These results showed clear superiority of BGP as compared with EGP in the area of bandwidth consumed by the protocol. Observations on the CA*Net by Dennis Ferguson, and on the T1 NSFNET Backbone by Susan Hares confirmed clear superiority of BGP as compared with EGP in the area of CPU requirements.

9. Using TCP as a transport for BGP.

9.1. Introduction.

On multiple occasions some members of IETF expressed concern about using TCP as a transport protocol for BGP. In this section we examine the use of TCP for BGP in terms of:

- real versus perceived problems
- offer potential solutions to real problems
- perspective on the convergence problem
- conclusions

BGP is based on the incremental updates. This is done intentionally to conserve the CPU and bandwidth requirements. Extensive operational experience with BGP in the Internet showed that indeed the use of the incremental updates allows significant saving both in terms of the CPU utilization and bandwidth consumption. However, to operate correctly the incremental updates must be exchanged over a reliable

transport. BGP uses TCP as such transport. It had been suggested that another transport protocol would be more suitable for BGP.

9.2. Examination of Problems - Real and "perceived".

Extensive operational experience with BGP in the Internet showed that the only real problem that was attributed to BGP in general, and the use of TCP as the transport for BGP in particular, was its slow convergence in presence of congestion. This problem was experienced in CA*Net. As we mentioned before, CA*Net is composed of 10 routers that form a ring. The routers are connected by 56 Kbits/sec links. All links are heavily utilized and are often congested. Experience with BGP in CA*Net showed that unless special measures are taken, the protocol may exhibit slow convergence when BGP information is passed over the slow speed (56 Kbits/sec) congested links. This is because a large percentage of packets carrying BGP information are being dropped due to congestion. Therefore, there are three inter-related problems: congestion, packet drops, and the resulting slow convergence of routing under congestion and packet drops.

Observe, that any transport protocol used by BGP would have difficulty preventing packets from being dropped under congestion, since it has no direct control over the routers that drop the packets, and the congestion has nothing to do with the BGP traffic. Therefore, since BGP is not the cause of congestion, and cannot directly influence dropping at the routers, replacing TCP (as the BGP transport) with another transport protocol would have no effect on packets being dropped due to congestion. We think that once a network is congested, packets will be dropped (regardless of whether these packets carry BGP or any other information), unless special measures outside of BGP in general, and the transport protocol used by BGP in particular, are taken.

If packets carrying routing information are lost, any distributed routing protocol will exhibit slow convergence. If quick convergence is viewed as important for a routing within a network, special measures to minimize the loss of packets that carry routing information must be taken. The next section suggests some possible methods.

9.3. Solutions to the problem.

Two possible measures could be taken to reduce the drop of BGP packets which slows convergence of routing:

1) alleviate the congestion

2) reduce the percentage of BGP packets that are dropped due

to congestion by marking BGP packets and setting policies to routers to try not to drop BGP packets

Alleviating the network congestion is a subject outside the control of BGP, and will not be discussed in this paper.

Operational experience with BGP in CA*Net shows that reducing the percentage of BGP packets dropped due to congestion by marking them, and setting policies to routers to try not to drop BGP packets completely solves the problem of slow convergence in presence of congestion.

The BGP packets can be marked (explicitly or implicitly) by the following three methods:

a) by means of IP precedence (Internetwork Control)

b) by using a well-known TCP port number

c) by identifying packets by just source or destination IP address.

Appendix 4 of the BGP protocol specification, RFC 1163, recommends the use of IP precedence (Internetwork Control) because the precedence provides a well-defined mechanism to mark BGP packets. The method of a well-known TCP port number to identify packets is similar to the one that was used by Dave Mills in the NSFNET Phase I. Dave Mills identified Telnet traffic by a well known TCP port number, and gave it priority over the rest of the traffic. CA*Net identified BGP traffic based on it’s source and destination IP address. Packets receive a priority if either the source or the destination IP address belongs to CA*Net.

If packets that carry the routing information are being dropped (because of congestion), one also may ask about how does a particular routing protocol react to such an event. In the case of BGP the packets are retransmitted using the TCP retransmission mechanism. It seems plausible that being more aggressive in terms of the retransmission should have positive effect on the convergence. This can be done completely within TCP by adjusting the TCP retransmission timers. However, we would like to point out that the change in the retransmission strategy should not be viewed as a cure for the problem, since the root of the problem lies in the way how packets that carry the BGP information are handled within a congested network, and not in how frequently the lost packets are retransmitted.

It should also be pointed out that the local system can control the

amount of data to be retransmitted (in case of a congestion or losses) by adjusting the TCP Window size. That allows to control the amount of potentially obsolete data that has to be retransmitted.

9.4. Perspective on the Convergence Problem.

To put the convergence problem in a proper perspective, we'd like to point out that much of the Internet now uses EGP at AS borders, ensuring that routing changes cannot be guaranteed to propagate between ASes in less than a few minutes. It would take huge amount of congestion to slow BGP to this pace. Additionally, the problems of EGP in the face of packet loss are well known and far exceed any imaginable problem BGP/TCP might ever suffer. Therefore, the worst case behavior of BGP is about the same as the steady case behavior of EGP.

Within an AS the speed of convergence of the AS's IGP in the face of congestion is of far greater concern than the propagation speed of BGP, and indeed avoiding loss of packets carrying IGP, and a more aggressive transport is similarly of much greater importance for an IGP than for BGP.

The issue of BGP convergence is of exaggerated importance to CA*Net since CA*Net carries no information about external routes in its IGP. CA*Net uses BGP to transfer external routes for use in computing internal routes through the CA*Net network. The reason CA*Net does this has nothing to do with BGP. Under more ordinary circumstances an IGP carries external routing information for use in computing internal routes. CA*Net shows that BGP can work under extreme stress. However, it's results should not be taken as the norm since most networks will use BGP in a different (and less stressful) configuration, where information about external routes will be carried by an IGP.

9.5. Conclusion.

The extensive operational experience with BGP showed that the only problem attributed to BGP was the slow convergence problem in presence of congestion. We demonstrated that this problem has nothing to do with BGP in general, or with TCP as the BGP transport in particular, but is directly related to the way how packets that carry routing information are handled within a congested network. The document suggests possible ways of solving the problem. We would like to point out that the issue of convergence in presence of congested network is important to all distributed routing protocol, and not just to BGP. Therefore, we recommend that every routing protocol (whether it is intra-autonomous system or inter-autonomous system) should clearly specify how its behavior is affected by the

congestion in the networks, and what are the possible mechanisms to avoid the negative effect of congestion (if any).

10. Bibliography.

[1] Hinden, B., "Internet Routing Protocol Standardization Criteria", RFC 1264, BBN, October 1991.

[2] Rekhter, Y., and P. Gross, "Application of the Border Gateway Protocol in the Internet", RFC 1268, T.J. Watson Research Center, IBM Corp., ANS, October 1991.

[3] Lougheed, K., and Y. Rekhter, "A Border Gateway Protocol 3 (BGP-3)", RFC 1267, cisco Systems, T.J. Watson Research Center, IBM Corp., October 1991.

[4] Willis, S., and J. Burruss, "Definitions of Managed Objects for the Border Gateway Protocol (Version 3)", RFC 1269, Wellfleet Communications Inc., October 1991.

Security Considerations

Security issues are discussed in section 6.

Author's Address

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

IETF BGP WG mailing list: iwg@rice.edu
To be added: iwg-request@rice.edu