Network Working Group
Request for Comments: 1773
Obsoletes: 1656
Category: Informational

P. Traina
cisco Systems

March 1995

Experience with the BGP-4 protocol

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 memo is to document how the requirements for advancing a routing protocol to Draft Standard have been satisfied by Border Gateway Protocol version 4 (BGP-4). This report documents experience with BGP. This is the second of two reports on the BGP protocol. As required by the Internet Architecure 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 initial work of Peter Lothberg (Ebone), Andrew Partan (Alternet), and several others. Details of their work were presented at the Twenty-fifth IETF meeting and are available from the IETF proceedings.

Please send comments to iwg@ans.net.

Acknowledgments

The BGP protocol has been developed by the IDR (formerly BGP) Working Group of the Internet Engineering Task Force. I would like to express deepest thanks to Yakov Rekhter and Sue Hares, co-chairs of the IDR working group. I'd also like to explicitly thank Yakov Rekhter and Tony Li for the review of this document as well as constructive and valuable comments.

Documentation

BGP 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 RFC 1267. The changes between versions 1, 2 and 3 are explained in Appendix 2 of [2]. All of the functionality that was present in the previous versions is present in version 4.

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".

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.

BGP version 4 redefines the (previously 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 [5]. BGP version 4 has also modified the AS-PATH attribute so that sets of autonomous systems, as well as individual ASs may be described. In addition, BGP version for has redescribed the INTER-AS METRIC attribute as the MULTI-EXIT DISCRIMINATOR and added new LOCAL-PREFERENCE and AGGREGATOR attributes.

Possible applications of BGP in the Internet are documented in [3].

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

MIB

A BGP-4 Management Information Base has been published [4]. The MIB was written by Steve Willis (Wellfleet), John Burruss (Wellfleet), and John Chu (IBM).

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.

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. 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.

However, 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.

Implementations

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

- cisco Systems
- gated consortium
- 3COM
- Bay Networks (Wellfleet)
- Proteon

To facilitate efficient BGP implementations, and avoid commonly made mistakes, the implementation experience with BGP-4 in with cisco's implementation was documented as part of RFC 1656 [4].

Implementors are strongly encouraged to follow the implementation suggestions outlined in that document and in the appendix of [2].

Experience with implementing BGP-4 showed that the protocol is relatively simple to implement. On the average BGP-4 implementation takes about 2 man/months effort, not including any restructuring that may be needed to support CIDR.

Note that, as required by the IAB/IESG for Draft Standard status, there are multiple interoperable completely independent implementations.

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. This use involves at least two of the 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 28 Kbits/sec to 150 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 RISC based CPUs, and includes both the special purpose routers and the general purpose workstations running UNIX.

In terms of the actual topologies it varies from a very sparse (spanning tree of ICM) to a quite dense (NSFNET backbone).

At the time of this writing BGP-4 is used as an inter-autonomous system routing protocol between ALL significant autonomous systems, including, but by all means not limited to: Alternet, ANS, Ebone, ICM, IIJ, MCI, NSFNET, and Sprint. The smallest know backbone consists of one router, whereas the largest contains nearly 90 BGP speakers. All together, there are several hundred known BGP speaking routers.

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 distinction between sites historically considered backbones vs "regional" networks.

Within most transit networks, BGP is used as the exclusive carrier of the exterior routing information. At the time of this writing within a few sites use BGP in conjunction with an interior routing protocol to carry exterior routing information.

The full set of exterior routes that is carried by BGP is well over 20,000 aggregate entries representing several times that number of connected networks.

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

Specific details of the operational experience with BGP in Alternet, ICM and Ebone were presented at the Twenty-fifth IETF meeting (Toronto, Canada) by Peter Lothberg (Ebone), Andrew Partan (Alternet) and Paul Traina (cisco).

Operational experience with BGP exercised all basic features of the protocol, including authentication, routing loop suppression and the new features of BGP-4, enhanced metrics and route aggregation.

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 Twenty-first 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 the BGP protocol family as compared with EGP in the area of CPU requirements.

Migration to BGP version 4

On multiple occasions some members of IETF expressed concern about the migration path from classful protocols to classless protocols such as BGP-4.

BGP-4 was rushed into production use on the Internet because of the exponential growth of routing tables and the increase of memory and CPU utilization required by BGP. As such, migration issues that normally would have stalled deployment were cast aside in favor of pragmatic and intelligent deployment of BGP-4 by network operators.

There was much discussion about creating "route exploders" which would enumerate individual class-based networks of CIDR allocations to BGP-3 speaking routers, however a cursory examination showed that this would vastly hasten the requirement for more CPU and memory resources for these older implementations. There would be no way internal to BGP to differentiate between known used networks and the unused portions of the CIDR allocation.

The migration path chosen by the majority of the operators was known as "CIDR, default, or die!"

To test BGP-4 operation, a virtual "shadow" Internet was created by linking Alternet, Ebone, ICM, and cisco over GRE based tunnels. Experimentation was done with actual live routing information by establishing BGP version 3 connections with the production networks at those sites. This allowed extensive regression testing before deploying BGP-4 on production equipment.

After testing on the shadow network, BGP-4 implementations were deployed on the production equipment at those sites. BGP-4 capable routers negotiated BGP-4 connections and interoperated with other sites by speaking BGP-3. Several test aggregate routes were injected into this network in addition to class-based networks for compatibility with BGP-3 speakers.

At this point, the shadow-Internet was re-chartered as an "operational experience" network. tunnel connections were established with most major transit service operators so that operators could gain some understanding of how the introduction of aggregate networks would affect routing.

After being satisfied with the initial deployment of BGP-4, a number of sites chose to withdraw their class-based advertisements and rely only on their CIDR aggregate advertisements. This provided motivation for transit providers who had not migrated to either do so, accept a default route, or lose connectivity to several popular destinations.

Metrics

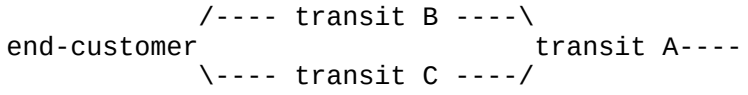
BGP version 4 re-defined the old INTER-AS metric as a MULTI-EXIT-DISCRIMINATOR. This value may be used in the tie breaking process when selecting a preferred path to a given address space. The MED is meant to only be used when comparing paths received from different external peers in the same AS to indicate the preference of the originating AS.

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 insure that peers could not "shed" or "absorb" traffic for networks that they advertise.

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

One shortcoming in the BGP4 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).

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.

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 an n^2 mesh of TCP connections 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.

Internet Dynamics

As discussed in [7], 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 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 6 of [2] are descriptions of dampening 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.

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 1266. 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.

Author's Address

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

EMail: pst@cisco.com

References

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

[2] 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.

[3] 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.

[4] 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.

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

[7] Traina P., "BGP Version 4 Protocol Analysis", RFC 1774, cisco Systems, March 1995.