Inter-Domain Routing Working Group
INTERNET DRAFT
<draft-ietf-idr-bgp4-op-experience-02.txt>

Paul Traina
cisco Systems
May 13, 1996

Operational Experience with the BGP-4 protocol

Status of this Memo

Introduction

The purpose of this memo is to document how the requirements for advancing a routing protocol to Full Standard have been satisfied by Border Gateway Protocol version 4 (BGP-4). This report documents experience with BGP. It is the second of two reports on the BGP protocol.

The remaining sections of this memo document how BGP satisfies General Requirements specified in Section 3.0, as well as the Requirements for Full Standard as specified in Section 6.0 of the "Internet Routing Protocol Standardization Criteria" document [1].

Please send comments to bgp@ans.net.

Documentation

BGP is an inter-autonomous system routing protocol designed for TCP/IP networks. 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 4 has re-described 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, bgp@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 (Bay), John Burruss (Bay), 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 the support of 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 the carriage of an encrypted digital signature in every BGP message. All authentication failures result in the sending of a NOTIFICATION message 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.

One method for improving the security of TCP connections for use with BGP has been documented in [7].

Operational experience

This section discusses operational experience with BGP-4, which has involved the use of several independent implementations of BGP.

BGP has been used in the Internet since 1989, BGP-4 since 1993. This use has involved at least three independant implementations. Production use of BGP has included 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.

This environment includes link bandwidths which vary from from 28 Kbits/sec to 150 Mbits/sec.

Routers which run BGP range from relatively low-performanced IBM PC/RTs to those equiped with high performance RISC based CPUs, and includes both the special purpose routers and the general purpose workstations running UNIX.

Topologies in the production environment vary from the very sparse (e.g. the spanning tree of the ICM network) to quite dense (e.g. Sprintlink, Alternet, and MCI backbones).

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, and Sprint. The smallest know backbone consists of one BGP speaker, whereas the largest contains nearly 120 BGP speakers. All together, there are several thousand 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 those considered "local" networks.

Within most transit networks, BGP is used as the exclusive carrier of exterior routing information. At the time of this writing, few sites propogate all exterior routing information into their interior routing protocols.

The full set of exterior routes that is carried by BGP in the production Internet is well over 30,000 distinct classless prefixes representing several times that number of connected networks.

Operational experience with BGP-4 has 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 over EGP when protocol bandwidth consumption is compared. 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 "prefix 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 destinations and the unused portions of advertised CIDR allocations.

The migration path chosen by 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 using the shadow network, BGP-4 implementations were deployed on production transit networks at those sites. BGP-4

capable routers negotiated BGP-4 connections and inter-operated with other sites by speaking BGP-3. Several test aggregate routes were injected into this network in addition to classful destinations 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 destinations 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 supplied motivation for transit providers who had not migrated to either do so, accept a default route, or lose connectivity to several popular destinations.

Currently, BGP-4 is the default choice for carrying exterior routing information in the production Internet.

Metrics

BGP version 4 re-defined the 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 intended to be used only 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 IDR working wanted to insure 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 neighboring autonomous systems could not "shed" or

"absorb" traffic for destinations that they advertise.

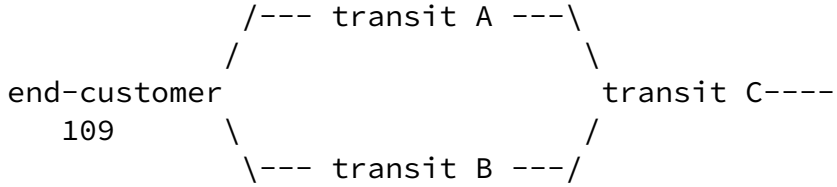
The LOCAL-PREFERENCE attribute was added so a local operator could easily configure a policy that overrode the standard best path determination mechanism without requiring the manual configuration on every router in the AS.

One shortcoming in the BGP-4 specification was a suggestion for a default value of LOCAL-PREFERENCE to be assumed if none was provided. Defaults of 0 or the maximum value each have range limitations, so a common default would have aided in the interoperation of different BGP implementations in the same AS (since LOCAL-PREFERENCE 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 or remote 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 multiple transit service providers connect to additional autonomous systems, there is no formal mechanism for indicating a path selection preference should a remote autonomous system wish to respect that preference.

In BGP implementations where the total length of the sequence portions of the AS path attribute may be used as part of the path selection criteria, one practice in use today is to prepend additional copies of the originator's autonomous system number to the AS path.



Using the example above, if the "end customer" advertises routes originating in its autonomous system as having an AS path of "109" to

transit A, and a path of "109 109" to transit B, transit provider C may be influenced by the difference in AS sequence lengths and prefer the path via transit A.

There has been some discussion of the creation of an optional transitive attribute which would represent a sequence of (AS, preference) entries to indicate a preference value for a given path. Cooperating ASs would chose traffic based upon comparison of "interesting" portions of this sequence according to local routing policy.

Additional suggestions have been made suggesting a less flexible "destination provider selection" attribute to indicate desired preferences.

While protecting a given autonomous system's routing policy is of paramount concern, avoiding extensive hand configuration of routing policies needs to be examined more carefully in future protocol varients.

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.

This practice leads to an O(n^2) mesh of TCP connections and requires some method of configuring and maintaining those connections. BGP does not regulate how this information is to be propagated, so alternatives, such as injecting BGP attribute information into the local IGP have been suggested. Also, internal BGP "route reflectors", and "autonomous system confederation" mechanisms have been implemented and demonstrate a significant improvement in configuration, memory and CPU requirements necessary to convey information to all other BGP peers in an autonomous system.

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 receive some level of damping when more specific reachability information is aggregated into larger blocks, however this isn't sufficient.

At least one current implementation of BGP provides route update dampening that includes routing hysteresis. This allows fast convergence for routes that flap relatively infrequently while suppressing instabilities caused by frequently flapping paths. Operational experience in the Internet shows that large-scale deployment of this dampening technique has proven to be highly beneficial for the stability of the routing system.

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 from which this document is based. I'd like to thank Yakov Rekhter, John Hawkinson, and Vince Fuller for the review of this document as well as constructive and valuable comments. This report is based on the initial work of Peter Lothberg (STUPI), Andrew Partan (UUNET), and several others.

Author's Address:

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

References

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

[2] draft-ietf-idr-bgp4-02.txt
Rekhter, Y., and Li, T., "A Border Gateway Protocol 4 (BGP-4)", January 1996.

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

[4] RFC1657
S. Willis, J. Burruss, J. Chu, "Definitions of Managed Objects for the Fourth Version of the Border Gateway Protocol (BGP-4) using SMIv2", July 1994.

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

[6] RFC1773
Traina P., "Experience with the BGP-4 protocol." March 1995.

[7] RFC1774
Traina P., "BGP Version 4 Protocol Analysis", March 1995.