IDR Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 25, 2010

R. Raszuk, Ed.
K. Patel, Ed.
Cisco Systems
March 24, 2010

Transport Instance BGP
draft-raszuk-ti-bgp-01

Abstract

BGP4 protocol is a well established single standard of an inter-domain routing and non-routing information distribution today. For many applications it is also the protocol of choice to disseminate various application based information intra-domain. It's popularity and it's wide use has been effectively provided by it's reliable transport, session protection as well as loop free build in mechanism.

It has been observed in both intra-domain as well as inter-domain applications that reliable information distribution is an extremely desired tool for many services. Introduction of Multiprotocol Extensions to BGP even further attracted various sorts of new information to be carried over BGP4.

The observation proves that amount and nature of information carried by BGP increases and diverges from the original goal of interconnection for IP Internet Autonomous Systems at a rather fast pace.

This draft proposes BGP to divide information into two broad categories: Internet routing critical and non Internet routing critical that would also include information carried by BGP which is not related directly to routing. For the purpose of this document we will refer to the latter case as second BGP instance.

This draft proposes that the current BGP infrastructure will continue to be used to disseminate Internet routing related information while non routing information or private routing data is recommended to be carried by independent transport instance BGP.

Status of this Memo

This Internet-Draft is submitted to IETF in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that

other groups may also distribute working documents as Internet-Drafts.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at http://www.ietf.org/shadow.html.

This Internet-Draft will expire on September 25, 2010.

Copyright Notice

Table of Contents

1. Contributors

The below is the list of contributors to this document:

Bruno Decraene, France Telecom, 38 rue du General Leclerc, Issy Moulineaux cedex 9, France, Email: bruno.decraene@orange-ftgroup.com

Jakob Heitz, Ericsson, 100 Headquarters Dr., San Jose, CA 95134, US Email: jheitz@redback.com

Thomas D. Nadeau, BT, 81 Newgate Street, London, EC1A 7AJ, United Kingdom, Email: tom.nadeau@bt.com

Jie Dong, Huawei Technologies Co.,Ltd, KuiKe Building, No.9 Xinxi Rd., Beijing, Hai-Dian District, 100085, P.R. China, Email: dongjie_dj@huawei.com

Yoshinobu Matsuzaki, Internet Initiative Japan Inc., Jinbocho Mitsui Bldg., 1-105 Kanda Jinbo-cho, Tokyo, Chiyoda-ku, Japan, Email: maz@iij.ad.jp

2. Introduction

BGP4 [RFC4271] protocol is practically a single standard today for the distribution of an inter-domain routing information. Under many applications it is also used as the protocol of choice when disseminating various application-based information intra-domain.

It's popularity and it's wide use has been effectively provided by it's extensibility, reliable transport, session protection as well as built in loop prevention mechanisms.

It has been observed in both intra-domain as well as inter-domain applications that reliable information distribution is an extremely desired tool for many applications. The introduction of Multiprotocol extensions to BGP [RFC4760] made it appealing for new kinds of information to be carried over BGP4.

While these extensions have proven to be useful, they however have increased the load of information as well as the type of information that BGP was originally envisioned to carry.

This draft proposes BGP to divide information into two broad categories: Internet routing critical and non Internet routing critical. The latter would also include information carried by BGP which is not related directly to routing. For the purpose of this document we will refer to the latter case as second BGP instance.

This draft proposes that the current BGP infrastructure will continue to be used to disseminate Internet routing related information while non Internet routing information or private routing data are recommended to be carried by independent transport instance BGP.

For all currently defined and deployed AFI/SAFIs the mapping on which plane of BGP (routing or transport) such information may be carried is left to the choice of the implementation flexibility and the operator's decision. For all subsequent new AFI/SAFIs it is RECOMMENDED that implementations would have them supported on both instances and that authors of new specifications provide a guidance on which BGP plane they should be carried. It is expected that both instances while running independently from each other will be executed from the same bgp code base.

Authors would like to also observe that the idea of separation routing from non routing related information to be carried over routing protocol is not only limited to BGP. As example one could notice proposed OSPF Transport Instance document [I-D.acee-ospf-transport-instance] where the idea of safely reusing reliable flooding has been recently proposed. We do admit that it has also been some form of inspiration for this proposal.

Another point of view in favor of BGP instance separation is the aspect of service protection. One could see BGP process responsible for global routing due to it's global nature much more exposed to control plane errors and attacks then potentially private only BGP instance contained to one or few ASes, possibly under common administration. In the same way one could also observe that by fully separating global Internet BGP from any local BGP based services the Internet itself can be fully isolated from any issues caused by local service provider's services.

3. Today's operation

In today's networks BGP4 operates per BGP specification [RFC4271]. This model of operation has proven to have number of disadvantages when it comes to concurrent support of multiple applications when amount of transported number of entries is already non trivial, when is not bounded by application architecture and when it is continuously growing.

There are many examples where major router vendors recommend to separate route reflectors into disjoined clusters so Internet routes are not affected by L3VPN routes and vice-versa. To put things into right perspective one needs to observe that local per box scaling numbers have already reached millions of VPN routes. Such scaling

provides real challenge for CPU as well as addressable memory space in 32-bit operating systems when all of such applications use single instance of BGP.

Another common complain is that by default all address families are carried today over single TCP session and any major protocol error or local system failure may results in full BGP instance reset affecting all applications carried between such pair of BGP speakers.

4. Related work

To address the session separation without forcing users to manually bound each session or group of session to a different BGP peering address Multisession BGP [I-D.scudder-bgp-multisession] solution has

been proposed. It is our opinion that Multisession BGP is an excellent tool to automatically bound selected group of applications to different TCP BGP sessions. But this is only limited to session separation.

All BGP OPEN messages would still end up going to the same BGP TCP port number 179. Furthermore, all the incoming sessions are handled by the same BGP process. Even in distributed BGP systems today single speaker is still tasked to handle all address families exchanged with a set of peers it is serving.

Multisession is an excellent way to easily separate different address families and bound them to different TCP sessions within each BGP instance. Such separation would be done at the micro level (session level) while separation of BGP instances could be seen as macro level division (BGP process/thread, memory space, internal queuing and buffering etc ...).

5. Transport Instance Proposal

In order to minimize impact between different classes of applications carried today or to be carried by BGP in the future to those of critical nature for Internet connectivity, this draft proposes to run two separate instances of BGP one for each of them.

The separation of concurrent, but not necessarily congruent BGP instances will be complete. It will include both the router side and network side.

5.1. Router's resource separation

There are many ways in modern router's operating systems to separate threads or processes running under single operating system from each other.

We will leave the details to the implementation, but it is assumed that any implementation which complies with this document will allow

to differentiate the amount of control plane CPU processing time allowed for specific BGP instance in it's scheduler's prioritization. It is recommended that prioritization of one instance over another in terms of CPU processing will be left to the local operator's decision. The proposed separation may also very much allow to run each BGP instance on separate core of multi core CPU or different RP where applicable.

It is also observed that such instance isolation will allow to use memory separation as well as different LC/RP communication channels/ queues resulting in even greater instance isolation and minimizing any potential impact between one another.

5.2. Protocol changes

The proposed here Transport Instance BGP does not require any changes to BGP4 protocol mechanism, state machine, error handling or operation. The exact same procedures and semantics apply in the same way for routing instance as well as transport instance BGP. The operational advantage in the instance separation is the ability to apply different Hold Time interval in each instance fitting to the operator's needs.

The only protocol change proposed in this document is the new TCP port number Transport Instance BGP will be waiting on for BGP OPEN Messages. Such new port number is to be allocated by IANA.

5.3. AFI/SAFI numbering

With the introduction of MP-BGP extension to BGP [RFC4760] protocol has been enhanced with the ability to carry different sets of information each separated by it's own AFI/SAFI value as listed in IANA's Subsequent Address Family Identifiers (SAFI) registry.

For Transport Instance BGP authors decided not to create a new IANA registry which would specify new SAFI pool. Instead we recommend that single AFI/SAFI pool to be used by both BGP instances.

The main motivation for this choice is to prevent any confusion on which SAFIs are allowed to be transported over which BGP instance as

well as to allow for customer configuration choice based on the

actual network needs and amount of information carried in each address family.

Another valid reason for single SAFI pool and no SAFI bonding to any particular BGP instance is the easy migration requirement from one instance to the other in smooth and not service impacting fashion. In order to perform such migration between instances operator will be free to run during a migration window given address family on both instances and when the target instance already populates the application database with the data terminate the originally deployed distribution of such information. Such process is bi-directional ie. rollback can be also supported gracefully.

5.4. BGP Identifier & BGP peering address

When running both independent instances on the same platform question arises on the recommended choice for BGP Identifier [I-D.ietf-idr-bgp-identifier] as well as BGP peering address to be used.

It needs to be observed that since via different BGP OPEN TCP port number and then different session ports if only implementation allows there is no requirement this specification would enforce to make any of those different between both instances.

Never the less this draft would like to encourage that such freedom of choice is given to the network administrator and that any dual instance BGP implementation should accommodate it.

Another advantage of sharing the same peering address of BGP sessions between instances is that in the event of operator's choice to use fast failure detection tools like BFD [I-D.ietf-bfd-base] the same event can be passed to both instances without any additional need to run two parallel and independent BFD sessions.

5.5. IP Precedence

On the network side all today's BGP messages are send with IP precedence value of Internetwork Control of 110000, which is used for high-priority routing traffic.

Transport Instance BGP SHOULD use as default the same IP precedence, but implementations MAY allow configuring a different one to reflect the real purpose of the new BGP instance.

6. Summary of benefits

Below is a combined list of main benefits provided by Transport Instance BGP:

- Mutual isolation and independence from protocol or process failures caused by any instance.

- Independence in: CPU usage, memory space and internal router buffering.

- Different port for BGP OPEN messages allowing the same BGP router_id or peering address sharing between instances.

- Different and fully isolated TCP sessions between instances. Each instance may still benefit from multisessions BGP proposal within each instance.

- Possibility of different IP precedence BGP message marking for more fair and accurate PHB treatment.

- Open platform for carrying non Internet routing information or easy migration path with minimized risk to current BGP infrastructure in new emerging Internet architecture's hierarchical model.

The technique here is quite general. If, in the future, it is found that there is a clearly definable need for yet more separate transports, additional RFCs can be written defining the applicability and the TCP/SCTP port number to be used.

7. Applications

As examples one may notice that carrying router names for easy operational enhancement, carrying free form ADVISORY [I-D.scholl-idr-advisory] Messages or adding flexibility to auto discover IBGP peers [I-D.raszuk-idr-ibgp-auto-mesh] fit nicely into Transport Instance BGP.

Another group of potential candidates for Transport Instance BGP could be any type of auto discovery mechanism for other applications. For example: L2VPN/VPLS or MVPN Auto Discovery [I-D.ietf-l3vpn-2547bis-mcast] are possible candidates.

Along the same lines a service provider may also choose to use

Transport Instance BGP to distribute information about L3VPN route targets as described in RFC4684 [RFC4684].

Another class of applications perfectly fitting the separate BGP instance model for it's global information distribution authors foresee a mapping plane of identifiers to locators in the new evolving internet architecture. As example LISP-ALT [I-D.fuller-lisp-alt] or APT [I-D.jen-apt] are already calling to use BGP as a mapping plane protocol to simplify initial deployment. While it is foreseen that in the future those may migrate to better distribution schemes for example LISP-DHT to get enough of initial traction and momentum a Transport Instance BGP seems like a very good match to the mapping plane requirements.

One may observe that Service Providers may choose to deploy a new instance of BGP to carry their critical services (example L3VPNs) over it for full isolation from Internet BGP. In such application they will be able to prioritize such instance according to their internal policy and offered services prioritization.

Last, but not least to recommend to be enabled on transport instance BGP is RFC5512 [RFC5512] BGP Encapsulation SAFI and BGP Tunnel Encapsulation Attribute.

8. Security considerations

Transport Instance BGP proposed in this document does not introduce any new security concerns as compared to base BGP4 specification [RFC4271]. Also all security work applicable to base routing instance BGP does also apply as is to transport instance BGP.

9. IANA Considerations

The new TCP port number for Transport Instance BGP are to be allocated by IANA from WELL KNOWN PORT NUMBERS registry.

bgp-ti xxx/tcp BGP Transport Instance

While routing instance BGP has also been allocated UDP port 179 authors see no particular reason for UDP port allocation for BGP.

The new SCTP port number for Transport Instance BGP are to be allocated by IANA from WELL KNOWN PORT NUMBERS registry.

bgp-ti xxx/sctp BGP Transport Instance

Specification to use BGP over SCTP can be found here [I-D.zhiyfang-fecai-bgp-over-sctp]

10. Acknowledgments

The authors would like to thank Randy Bush, Tom Scholl and Joel Halpern for their valuable comments.

11. References

11.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

11.2. Informative References

[I-D.acee-ospf-transport-instance] Lindem, A., Roy, A., and S. Mirtorabi, "OSPF Transport Instance Extensions", draft-acee-ospf-transport-instance-03 (work in progress), February 2009.

[I-D.fuller-lisp-alt]
Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "LISP Alternative Topology (LISP+ALT)", draft-fuller-lisp-alt-05 (work in progress), February 2009.

[I-D.ietf-bfd-base]
Katz, D. and D. Ward, "Bidirectional Forwarding Detection", draft-ietf-bfd-base-11 (work in progress), January 2010.

[I-D.ietf-idr-bgp-identifier]
Chen, E. and J. Yuan, "AS-wide Unique BGP Identifier for BGP-4", draft-ietf-idr-bgp-identifier-11 (work in progress), February 2010.

[I-D.ietf-l3vpn-2547bis-mcast]

Aggarwal, R., Bandi, S., Cai, Y., Morin, T., Rekhter, Y., Rosen, E., Wijnands, I., and S. Yasukawa, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10 (work in progress), January 2010.

[I-D.jen-apt]
Jen, D., Meisel, M., Massey, D., Wang, L., Zhang, B., and L. Zhang, "APT: A Practical Transit Mapping Service", draft-jen-apt-01 (work in progress), November 2007.

[I-D.raszuk-idr-ibgp-auto-mesh]
Raszuk, R., "IBGP Auto Mesh", draft-raszuk-idr-ibgp-auto-mesh-00 (work in progress), June 2003.

[I-D.scholl-idr-advisory]
Scholl, T. and J. Scudder, "BGP Advisory Message", draft-scholl-idr-advisory-00 (work in progress), March 2009.

[I-D.scudder-bgp-multisession]
Scudder, J. and C. Appanna, "Multisession BGP", draft-scudder-bgp-multisession-00 (work in progress), November 2003.

[I-D.zhiyfang-fecai-bgp-over-sctp]
Fang, K. and F. Cai, "BGP-4 message transport over SCTP", draft-zhiyfang-fecai-bgp-over-sctp-00 (work in progress), May 2009.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, November 2006.

[RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.

Authors' Addresses

Robert Raszuk (editor)
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: raszuk@cisco.com

Keyur Patel (editor)
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: keyupate@cisco.com