IDR
Internet-Draft
Updates: 4271, 4360, 7153 (if approved)
Intended status: Standards Track
Expires: September 4, 2018

Z. Li
China Mobile
J. Dong
Huawei Technologies

March 3, 2018

Carry congestion status in BGP community
draft-li-idr-congestion-status-extended-community-07

Abstract

To aid BGP receiver to steer the AS-outgoing traffic among the exit links, this document introduces a new BGP community, congestion status community, to carry the link bandwidth and utilization information, especially for the exit links of one AS. If accepted, this document will update RFC4271, RFC4360 and RFC7153.

The introducd congestion status community is not used to impact the decision process of BGP specified in section 9.1 of RFC4271, but can be used by route policy to impact the data forwarding behavior.

Status of This Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on September 4, 2018.

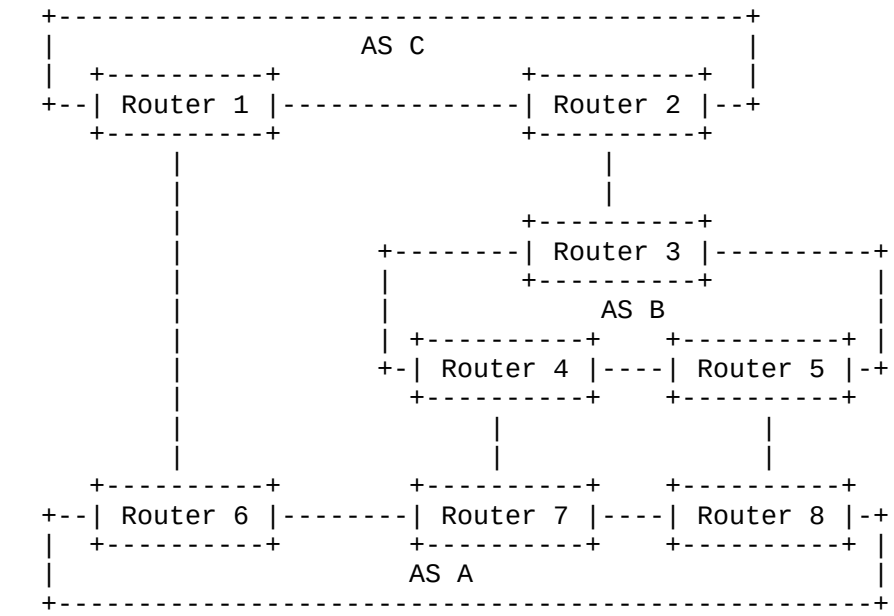
Copyright Notice

Table of Contents

1. Introduction

Knowing the congestion status (bandwidth and utilization) of the AS exit links is useful for traffic steering, especially for steering the AS outgoing traffic among the exit links. Section 7 of [I-D.gredler-idr-bgplu-epe] explicitly specifies this kind of requirement, which is also needed in our field network.

The following figure is used to illustrate the benefits of knowing the congestion status of the AS exit links. AS A has multiple exit links connected to AS B. Both AS A and B has exit link to AS C, and AS B provides transit service for AS A. Due to cost or some other reasons, AS A prefers using AS B to transmit its' traffic to AS C, not the directly connected link between AS A and C. If the exit routers, Router 7 and 8, in AS A tell their iBGP peers the congestion status of the exit links, the peers in turn can steer some outgoing traffic toward the less loaded exit link. If AS A knows the link between AS B and AS C is congested, it can steer some traffic towards AS C from AS B to the directly connected link by applying some route policies.



This document introduces new BGP extensions to deliver the congestion status of the exit link to other BGP speakers. The BGP receiver can then use this community to deploy route policy, thus steer AS outgoing traffic according to the congestion status of the exit links. This mechanisum can be used by both iBGP and eBGP.

In this verion, we provide three solution alternatives according to the discussion in the face to face meetings and mail list. After adoption, one solution will be selected as the final solution based on the working group consensus.

In a network deployed SDN (Software Defined Network) controller, congestion status extended community can be used by the controller to steer the AS outgoing traffic among all the exit links from the perspective of the whole network.

For the network with Route Reflectors (RRs) [RFC4456], RRs by default only advertise the best route for a specific prefix to their clients. Thus RR clients has no opportunity to compare the congestion status among all the exit links. In this situation, to allow RR clients learning all the routes for a specific prefix from all the exit links, RRs are RECOMMENDED to enable add-path functionality [RFC7911].

To emphasize, the introduced new BGP extensions have no impact on the decision process of BGP specified in section 9.1 of [RFC4271], but can be used by route policy to impact the data forwarding behavior.

2. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

3. Previous Work

In [constrained-multiple-path], authors from France Telecom also specified the requirement to know the congestion status of a link.

To aid a router to perform unequal cost load balancing, experts from Cisco introduced Link Bandwidth Extended Community in [link-bandwidth-community] to carry the cost to reach the external BGP neighbor. The cost can be either configured per neighbor or derived from the bandwidth of the link that connects the router to a directly connected external neighbor. This document was accepted by the IDR working group, but expired in 2013.

Link Bandwidth Extended Community only carries the link bandwidth of the exit link. The method provided in our document can carry the link bandwidth together with the link utilization information. What the BGP receiver needs to impact its traffic steering policy is the up-to-date unused link bandwith, which can be derived from the link bandwith and link utilization. Since Link Bandwidth Extended Community is expired, the BGP speaker who receives update message with both Link Bandwidth Extended Community and Congestion Status Community SHOULD ignore the Link Bandwidth Extended Community and use the Congestion Status Community.

4. Solution Alternative 1: Extended Community

As described in [RFC4360], the extended community attribute is an 8-octet value with the first one or two octets to indicate the type of this attribute. Since congestion status community needs to be delivered from on AS to other ASes, and used by the BGP speakers both in other ASes and within the same AS as the sender, it MUST be a transitive extended community, i.e. the T bit in the first octet MUST be zero.

We only define the congestion status community for four-octet AS number [RFC6793], since all the BGP speakers can handle four-octet AS number now and the two-octet AS numbers can be mapped to four-octet

AS numbers by setting the two high-order octets of the four-octet field to zero, as per [RFC6793].

Congestion status community is a sub-type allocated from Transitive Four-Octet AS-Specific Extended Community Sub-Types defined in section 5.2.4 of [RFC7153]. Its format is as Figure 1.

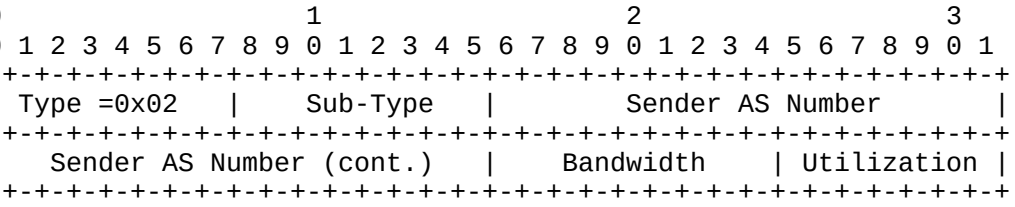


Figure 1: Congestion status extended community

Type: 1 octet. This field MUST be 0x02 to indicate this is a Transitive Four-Octet AS-Specific Extended Community.

Sub-Type: 1 octet. It is used to indicate this is a Congestion Status Extended Community. Its value is to be assigned by IANA.

Sender AS Number: 4 octets. Its value is the AS number of the BGP speaker who generates this congestion status extended community. If the generator has 2-octct AS number, it MUST encode its AS number in the last (low order) two bytes and set the first (high order) two bytes to zero, as per [RFC6793].

Bandwidth: 1 octet. Its value is the bandwidth of the exit link in unit of 10 gbps (gigabits per second). The link with bandwidth less than 10 gbps is not suitable to use this feature. To reflect the practice that sometimes the traffic is rate limited to a capacity smaller than the physical link, the value of the bandwidth can be the configured capacity of the link. The available configured capacity can be calculated from this field together with Utilization field. Zero means the bandwidth is unknown or is not advertised to other peers.

Utilization: 1 octet. Its value is the utilization of the exit link in unit of percent. A value bigger than 100 means the incoming traffic is higher than the link capacity. We can use the "Utilization" field together with the "Bandwidth" field to calculate the traffic load that we can further steer to this exit link.

5. Solution Alternative 2: Large Community

As described in [RFC8092], the BGP large community attribute is an optional transitive path attribute of variable length, consisting of 12-octet values. The BGP large community attribute is mainly used to extend the size of BGP Community [RFC1997] and Extened Community [RFC4360], thus to accommodate at least two four-octet ASNs [RFC6793]. As shown in the following figure, the format of the 12-octet BGP Large Community value is not suitable to be used to define new type for congestion status community.

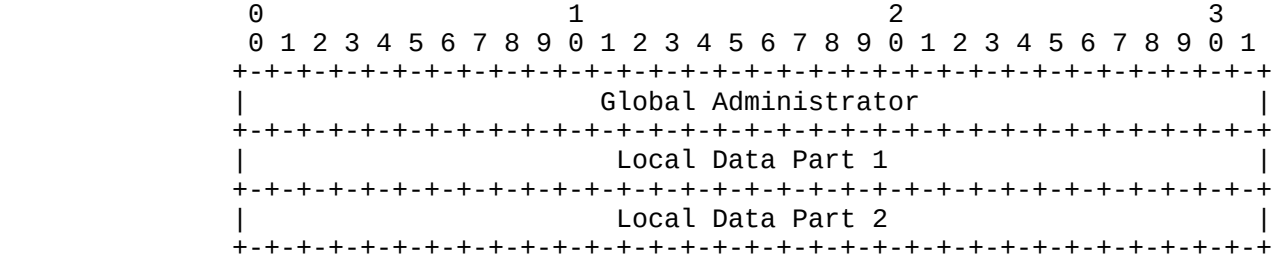


Figure 2

Global Administrator: A four-octet namespace identifier.

Local Data Part 1: A four-octet operator-defined value.

Local Data Part 2: A four-octet operator-defined value.

6. Solution Alternative 3: Community Container

As described in [I-D.ietf-idr-wide-bgp-communities], the BGP Community Container has flexible encoding format, which we can use to define the congestion status community.

A new type of the BGP Community Container is defined for the congestion status community, which has the same common header as the BGP Community Container with the following encoding format.

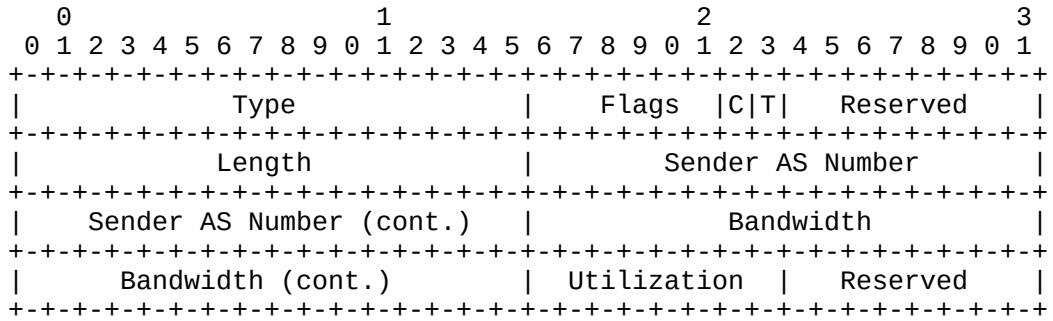


Figure 3

Type: 2 octets. Its value is to be assigned by IANA from the registry "BGP Community Container Types" to indicate this is the Congestion Status Community.

Flags: 1 octet. C and T bits MUST be set to indicate the Congestion Status Community is transitive across confederation and AS boundaries. The other bits in Flags field MUST be set to zero when originated and SHOULD be ignored upon receipt.

Reserved: Reserved fields are reserved for future definition, which MUST be set to zero when originated and SHOULD be ignored upon receipt.

Length: 2 octets. This field represents the total length of a given container’s contents in octets.

Sender AS Number: 4 octets. Its value is the AS number of the BGP speaker who generates this congestion status community. If the generator has 2-octct AS number, it MUST encode its AS number in the last (low order) two bytes and set the first (high order) two bytes to zero, as per [RFC6793].

Bandwidth: 4 octets. Its value is the bandwidth of the exit link in IEEE floating point format (see [IEEE.754.1985]), expressed in bytes per second. Zero means the bandwidth is unknown or is not advertised to other peers. Appendix A lists some typical bandwidth values, most of which are extracted from Section 3.1.2 of [RFC3471].

To reflect the practice that sometimes the traffic is rate limited to a capacity smaller than the physical link, the value of the bandwidth can be the configured capacity of the link. The available configured capacity can be calculated from this field together with Utilization field.

Utilization: 1 octet. Its value is the utilization of the exit link in unit of percent. A value bigger than 100 means the incoming traffic is higher than the link capacity. We can use the "Utilization" field together with the "Bandwidth" field to calculate the traffic load that we can further steer to this exit link.

7. Deployment Considerations

- To avoid route oscillation

  The exit router SHOULD set a threshold. When the utilization change reaches the threshold, the exit router SHOULD generate a BGP update message with congestion status community.

  Implementations SHOULD further reduce the BGP update messages trigered by link utilization change using the method similar to BGP Route Flap Damping [RFC2439]. When link utilization change by small amounts that fall under thresholds that would cause the announcement of BGP update message, implementations SHOULD suppress the announcement and set the penalty value accordingly.

  To reduce the update churn introduced, when one BGP router needs to re-advertise a BGP path due to attribute changes, it SHOULD update its Congestion Status Community at the same time. Supposing there are N ASes on the way from the far end egress BGP speaker to the final ingress BGP speaker, this allows reducing the update churn as the final ingress BGP speaker will receive a single UPDATE refreshing the N communities, rather than N UPDATEs, each refreshing one community.

- To avoid traffic oscillation

  Traffic oscillation means more traffic than expected is attracted to the low utilized link, and some traffic has to be steered back to other links.

  Route policy is RECOMMENDED to be set at the exit router. Congestion status community is only conveyed for some specific routes or only for some specific BGP peers.

  Congestion status community can also be used in a SDN network. The SDN controller uses the exit link utilization information to steer the Internet access traffic among all the exit links from the perspective of the whole network.

- Other Conserns

To avoid forwarding loops incremental deployment issues, complications in error handling, the reception of such community over IBGP session SHOULD NOT influence routing decision unless tunneling is used to reach the BGP Next-Hop.

8. Security Considerations

This document defines a new BGP community to carry the congestion status of the exit link. It is up to the BGP receiver to trust the congestion status communities or not. Following deployment models can be considered.

The BGP receiver may choose to only trust the congestion status communities generated by some specific ASes or containing bandwidth greater than a specific value.

You can filter the congestion status communities at the border of your trust/administrative domain. Hence all the ones you receive are trusted.

You can record the communities received over time, monitor the congestion e.g. via probing, detect inconsistency and choose to not trust anymore the ASes which advertise fake news.

9. IANA Considerations

For solution alternative 1, one sub-type is solicited to be assigned from Transitive Four-Octet AS-Specific Extended Community Sub-Types registry to indicate the Congestion Status Community defined in this document.

For solution alternative 3, one community value is solicited to be assigned from the registry "Registered Type 1 BGP Wide Community Community Types" to indicate the Congestion Status Community defined in this document.

10. Acknowledgments

We appreciate the constructive suggestions received from Bruno Decraene. Many thanks to Rudiger Volk, Susan Hares, John Scudder, Randy Bush for their review and comments to improve this document.

11. References

11.1. Normative References

[I-D.ietf-idr-wide-bgp-communities]
Raszuk, R., Haas, J., Lange, A., Decraene, B., Amante, S., and P. Jakma, "BGP Community Container Attribute", draft-ietf-idr-wide-bgp-communities-04 (work in progress), March 2017.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, DOI 10.17487/RFC4360, February 2006, <https://www.rfc-editor.org/info/rfc4360>.

[RFC7153] Rosen, E. and Y. Rekhter, "IANA Registries for BGP Extended Communities", RFC 7153, DOI 10.17487/RFC7153, March 2014, <https://www.rfc-editor.org/info/rfc7153>.

[RFC8092] Heitz, J., Ed., Snijders, J., Ed., Patel, K., Bagdonas, I., and N. Hilliard, "BGP Large Communities Attribute", RFC 8092, DOI 10.17487/RFC8092, February 2017, <https://www.rfc-editor.org/info/rfc8092>.

11.2. Informative References

[constrained-multiple-path]
Boucadair, M. and C. Jacquenet, "Constrained Multiple BGP Paths", October 2010, <https://www.ietf.org/archive/id/draft-boucadair-idr-constrained-multiple-path-00.txt>.

[I-D.gredler-idr-bgplu-epe]
Gredler, H., Vairavakkalai, K., R, C., Rajagopalan, B., Aries, E., and L. Fang, "Egress Peer Engineering using BGP-LU", draft-gredler-idr-bgplu-epe-11 (work in progress), October 2017.

[link-bandwidth-community]
Mohapatra, P. and R. Fernando, "BGP Link Bandwidth Extended Community", January 2013, <https://www.ietf.org/archive/id/draft-ietf-idr-link-bandwidth-06.txt>.

[RFC1997] Chandra, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, DOI 10.17487/RFC1997, August 1996, <https://www.rfc-editor.org/info/rfc1997>.

[RFC2439] Villamizar, C., Chandra, R., and R. Govindan, "BGP Route Flap Damping", RFC 2439, DOI 10.17487/RFC2439, November 1998, <https://www.rfc-editor.org/info/rfc2439>.

[RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, DOI 10.17487/RFC3471, January 2003, <https://www.rfc-editor.org/info/rfc3471>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC6793] Vohra, Q. and E. Chen, "BGP Support for Four-Octet Autonomous System (AS) Number Space", RFC 6793, DOI 10.17487/RFC6793, December 2012, <https://www.rfc-editor.org/info/rfc6793>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

Appendix A. Bandwidth Values

Some typical bandwidth values encoded in 32-bit IEEE floating point format are enumerated below.

Link Type	Bit-rate (Mbps)	Bandwidth Value (Bytes/Sec) (32-bit IEEE Floating point)
E1	2.048	0x487A0000
Ethernet	10.00	0x49989680
Fast Ethernet	100.00	0x4B3EBC20
OC-3/STM-1	155.52	0x4B9450C0
OC-12/STM-4	622.08	0x4C9450C0
GigE	1000.00	0x4CEE6B28
OC-48/STM-16	2488.32	0x4D9450C0
OC-192/STM-64	9953.28	0x4E9450C0
10GigE	10000.00	0x4E9502F9
OC-768/STM-256	39813.12	0x4F9450C0
100GigE	100000.00	0x503A43B7

Authors' Addresses

Zhenqiang Li
China Mobile
No.32 Xuanwumenxi Ave., Xicheng District
Beijing 100032
P.R. China

Email: li_zhenqiang@hotmail.com

Jie Dong
Huawei Technologies
Huawei Campus, No.156 Beiqing Rd.
Beijing 100095
P.R. China

Email: jie.dong@huawei.com