IDR P. Lapukhov
Internet-Draft E. Aries
Intended status: Standards Track Facebook
Expires: June 12, 2015 P. Marques
Contrail Systems
E. Nkposong
Microsoft Corporation
December 9, 2014

Use of BGP for opaque signaling
draft-lapukhov-bgp-opaque-signaling-00

Abstract

Border Gateway Protocol with multi-protocol extensions (MP-BGP) [RFC4760] enables the use of the protocol for dissemination of virtually any information. This document proposes a new Address Family/Subsequent Address Family to be used for distribution of opaque data. This functionality is intended to be used by applications other than BGP for exchange of their own data on top of BGP mesh. The structure of such data is not to be interpreted by the regular BGP speakers, rather the goal is to use BGP purely as a convenient and scalable communication system.

Table of Contents

1. Introduction

Implementation of Multiprotocol Extensions for BGP-4 [RFC4760] allows to pass aribtrary data in BGP protocol messages. This capability has been leveraged by many for dissemination of non-routing related information over BGP (for example, see "Dissemination of Flow Specification Rules" defined in [RFC5575]). However, there has been no channel defined explicitly to disseminate data with arbitrary opaque payload. The intended use case is for applications other than BGP to leverage the protocol machinery for distribution of their own state in the network domain. Publishers and consumers will use BGP UPDATE messages to exhcnage opaque data. It is up to the BGP implementation to provide a custom API for message producers or consumers if needed.

2. BGP Opaque Data AFI

This document introduces a new AFI known as a "BGP Opaque Data AFI" with the actual value to be assigned by IANA. The purpose of this AFI is to exchange opaque information within a BGP network.

3. BGP Key-Value SAFI

This document introduces a new SAIF known as "BGP Key-Value SAFI". The purpose of this SAFI is exchange of opaque information structured as a Key-Value binding (advertisement).

4. Capability Advertisement

A BGP speaker that wishes to exchange Opaque Data MUST use the Multiprotocol Extensions Capability Code, as defined in [RFC2858], to advertise the corresponding AFI/SAFI pair.

5. Disseminating Key-Value bindings

This document proposes to implement a distributed, eventually consistent Key-Value store on top of existing BGP protocol mechanics. The proposal is for "Key" portion to be encoded as the NLRI part of MP_REACH_NLRI attribute and "Value" encoded using a new optional transitive attribute.

> Publishers, acting as BGP speakers, advertise keys along with associated values into the routing domain. The BGP network synchronizes that state by propagating the encoded data following regular BGP protocol operations.
>
> Consumers, acting as BGP speakers, receive the information via BGP protocol UPDATE messages. Only publishers and consumers of the opaque data are supposed to interpret its contents - the rest of the BGP network acts merely as a dissemination system.

Multiple publishers can advertise the same key (NLRI) associated with different values. It is also possible for the advertised associations to have the same Key-Value pairs but differ in there other BGP attributes. In that case, BGP would follow the best-path selection logic to prevent duplicate information in the network. A consumer will receive the value created by the publisher "closest" in terms of BGP best-path selection logic, based on the policies that exist in the routing domain. This document does not propose any method of achieving global consensus for all published values for a given key. See section Section 5.3 that discusses the means of propagating multiple values for the same key.

5.1. Publishing a Key-Value binding

The encoding scheme proposed below follows the semantics of a Key-Value bindings. The "Key" is stored in the NLRI section of the MP_REACH_NLRI attribute, as shown on Figure 1.

```
+----------------------------------------------------------+
| AFI (2 octets)                                           |
+----------------------------------------------------------+
| SAFI (1 octet)                                           |
+----------------------------------------------------------+
| Length of Next Hop Address (1 octet), must be zero       |
+----------------------------------------------------------+
| Reserved (1 octet), must be zero                         |
+----------------------------------------------------------+
| Opaque Key Length (1 octet)                              |
+----------------------------------------------------------+
| Opaque Key Data (variable)                               |
+----------------------------------------------------------+
```

Figure 1: MP_REACH_NLRI Layout

- The AFI/SAFI values are to be allocated by IANA.
- Length of Next Hop Address: must be zero, since no information is encoded in the next-hop address field.
- Opaque Key Length: identifies the size of the Key field. If field is set to zero, the implementation MUST ignore the advertisement.
- Opaque Key Data: the byte string representing the opaque key contents. This portion SHOULD NOT be interpreted by BGP implementation.

The "Value" portion of a published binding is to be encoded in a new optional transitive attribute as shown on Figure 2:

```
 0                   1                   2                   3
 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
|     Type      |0 0 0 0| Opaque Value Length   |               |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+               |
~                                                               ~
|                   Opaque Value Data (variable)                |
+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+.........................
```

Figure 2: OPAQUE_VALUE attribute layout

- o Type: Identifies the new OPAQUE_VALUE attribute, with the value to be allocated by IANA.

- o Opaque Value Length: Two octets encoding the total length of the attribute in octets, including the Type and Length fields. The length is encoded as an unsigned binary integer. The four most significant bits of this field MUST be set to zero, due to the limit imposed by maximum BGP message size. Note that the minimum length is 3, indicating that no Opaque Value Data field is present. Such binding, in presence of non-zero length key is still valid, as it informs the consumers that the key "exists".

- o Opaque Value Data: A field containing zero or more octets. This portion SHOULD NOT be interpreted by BGP implementations.

Even when the OPAQUE_VALUE optional transitive attribute is not present in BGP advertisement, the BGP implementation MUST still retain Opaque Key (NLRI) in its LocRIB and propagate it further as usual. This case is to be interpreted as an announcement of the key existence.

5.2. Removing a Key-Value binding

The removal procedure follows the regular MP-BGP route withdrawal, using the MP_UNREACH_NLRI attribute. This section defines the attribute structure for the new AFI/SAFI.

The message shown on Figure 3 instructs the receiving BGP speaker to delete the N bindings corresponding to Key 1, Key 2 ... Key N if the keys have been previously learned from the withdrawing speaker. If any of the Keys is not found in the LocRIB or has not been previously received from the withdrawing BGP peer, such key removal request MUST be ignored.

```
+----------------------------------------------------------+
| AFI (2 octets)                                           |
+----------------------------------------------------------+
| SAFI (1 octet)                                           |
+----------------------------------------------------------+
| Opaque Key 1 Length (1 octet)                            |
+----------------------------------------------------------+
| Opaque Key 1 Data (variable)                             |
+----------------------------------------------------------+
~                                                          ~
| Opaque Key N Length (1 octet)                            |
+----------------------------------------------------------+
| Opaque Key N Data (variable)                             |
+----------------------------------------------------------+
```

Figure 3: MP_UNREACH_NLRI attribute layout

5.3. Propagating multiple values for the same key

It is possible to propagate multiple values associated with the same key using the Add-Path extension defined in [I-D.ietf-idr-add-paths]. The values are differentiated by the Path Identifier field present in the key. If the capability is negotiated between two BGP speakers, the key encoding in NLRI field of MP_REACH_NLRI and MP_UNREACH_NLRI attributes is extended to look as shown on Figure 4. The "Opaque Key Length" and "Opaque Key Data" retain the same meaning as defined previously.

```
+----------------------------------------------------------+
| Path Identifier (4 octets)                               |
+----------------------------------------------------------+
| Opaque Key Length (1 octet)                              |
+----------------------------------------------------------+
| Opaque Key Data (variable)                               |
+----------------------------------------------------------+
```

Figure 4: AddPath Encoding

The above defined encoding is to be used both with MP_REACH_NLRI and MP_UNREACH_NLRI attributes. It is up to the implementation to decide how many additional values to propagate with the key.

Note that the Add-Path extension could also be used to propagate the same Key-Value pairs with different Path Identifiers, assuming that other BGP attributes associated with the same NLRI are different.

6. Message filtering

Limiting the scope of opaque information flooding is an important operational concern. BGP already has the mechanisms needed to control this process, and these mechanisms are briefly reviewed below.

6.1. Automated filtering

One can leverage mechanics presented in [RFC4684] and use the route-target extended community attribute to identify "channels" where key-value bindings are published. The consumers would signal their interest in particular "channel" by advertising the corresponding router-target membership. The publications then need to contain the router-target extended community attribute to constrain information propagation.

6.2. Filtering via policy

Ad-doc message filtering could be implemented using BGP standard (see [RFC4271]) or extended community attributes (see [RFC4360]). The semantic of these attributes is to determined by the policy and publishers/consumers. Filtering could be done locally on receiving speaker, or on remote speaker, by using outbound route filtering feature defined in [RFC5291].

7. IANA Considerations

For the purpose of this work, IANA would be asked to allocate values for the new AFI and SAFI.

8. References

8.1. Normative References

[RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.

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

8.2. Informative References

[RFC4360] Sangli, S., Tappan, D., and Y. Rekhter, "BGP Extended Communities Attribute", RFC 4360, February 2006.

[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.

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

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, August 2008.

[RFC5575] Marques, P., Sheth, N., Raszuk, R., Greene, B., Mauch, J., and D. McPherson, "Dissemination of Flow Specification Rules", RFC 5575, August 2009.

[I-D.ietf-idr-add-paths]
Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", draft-ietf-idr-add-paths-10 (work in progress), October 2014.

Authors' Addresses

Petr Lapukhov
Facebook
1 Hacker Way
Menlo Park, CA 94025
US

Email: petr@fb.com

Ebben Aries
Facebook
1 Hacker Way
Menlo Park, CA 94025
US

Email: exa@fb.com

Pedro Marques
Contrail Systems
440 N Wolfe Rd
Sunnyvale, CA 94085
US

Email: roque@contrailsystems.com

Edet Nkposong
Microsoft Corporation
1 Microsoft Way
Redmond, WA 98052
US

Email: edetn@microsoft.com