Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 24, 2014

K. Patel
Cisco Systems
R. Raszuk
NTT I3
B. Pithawala
E. Osbourne
A. Sajassi
Cisco Systems
J. Uttaro
ATT
L. Jalil
VeriZon

October 21, 2013

BGP vector routing.
draft-patel-raszuk-bgp-vector-routing-01

Abstract

Network architectures have begun to shift from pure destination based routing to service aware routing. Operator requirements in this space include forcing traffic through particular service nodes (e.g. firewall, NAT) or segments. This document proposes an enhancement to BGP to accommodate these new requirements.

This document proposes a pure control plane solution which allows traffic to be routed via an ordered set of transit points (links, nodes, or services) on the way to traffic's destination, with no change in the forwarding plane. This approach is in contrast to other proposal in this space which provide similar capabilities via modifications to the forwarding plane.

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 RFC 2119 [RFC2119].

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 http://datatracker.ietf.org/drafts/current/.

Table of Contents

1. Introduction

This document addresses two problems. The first is traffic engineering - by providing specific paths over which traffic must flow, an operator can modify the traffic pattern on their network to better address congestion. While typically this has been accomplished by constructing MPLS-TE LSPs and mapping traffic on them, the overhead of the MPLS control plane and the requirement to use the MPLS data plane can pose an operational issue for some service providers such as data center providers. The emerging

solution of segment routing simplifies the control plane but is limited to intra-domain topologies only.

The second can be thought of as services engineering - by providing an ordered list of services nodes through which a particular destination's traffic must traverse, an operator can add services (e.g. NATs, load balancers, firewalls) along the forwarding path towards a specific destination. As services such as NAT, Firewalls and Load Balancers move to the cloud based model, a need to discover, prioritize and chain these services is needed. The draft draft-keyupate-bgp-services-02 describes extensions to BGP that facilitates auto discovery of services within the network. This draft proposes an extension to BGP that facilitates prioritizing and chaining of services within a network. Since service chaining is facilitated using the BGP control plane, it can readily be applied to IP-only tunneling encapsulations for network virtualization such as VXLAN and NVGRE.

In either case, this document refers to the use of the proposed BGP extension as Service Chaining.

To facilitate Service Chaining, this document defines a new BGP attribute known as a BGP Vector Node attribute. The BGP Vector Node attribute consist of an ordered list of IP transit hops that needs to be traversed before the packet is forwarded to its BGP NEXT HOP. The information carried in the ordered list of Vector Node is used towards augmenting the NEXT HOP information for the BGP prefixes as carried in the MP_REACH attribute. This draft specifies rules for BGP-speaking traffic forwarders (i.e. PEs and midpoint nodes) to replace the NEXT HOP information in their RIB/FIB with an intermediate node supplied by the BGP Vector Node attribute.

2. Protocol Extensions

This document describes a BGP attribute known as BGP Vector Node attribute, along with rules for identifying an intermediate next-hop from tthe BGP Vector Node attribute.

2.1. BGP Vector Node Attribute

The BGP Vector Node attribute is a new BGP optional transitive attribute. The attribute type code for the Vector Node attribute is to be assigned by IANA. The value field of the Vector Node attribute is defined as a set of one or more Vector Node TLVs.

A Vector Node TLVs within a Vector Node Attribute are defined as follows:

Type 1 TLV:

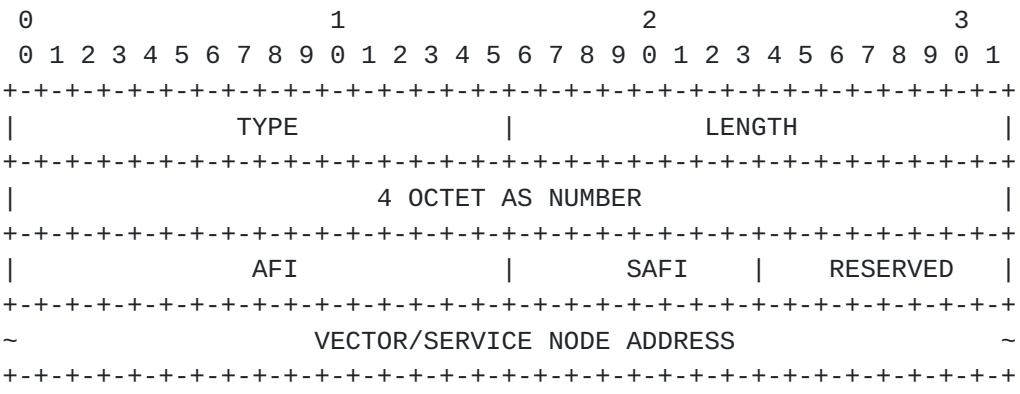


Figure1: Vector Node TLV Type 1

TYPE: Two octets encoding the Vector Node TLV Type. Type 1 contains vector or service node address which packets should traverse. Such address is part of the IGP. Such node is part of BGP mesh.

LENGTH: Two octets encoding the length in octets of the Vector Node TLV, excluding the type and length fields. The Length is encoded as an unsigned binary integer.

4 OCTET AS NUMBER: 4 octet AS number or zero padded 2 octet AS number of the autonomous system Vector Node Address belongs

AFI: Address Family Identifier (16 bits).

SAFI: Subsequent Address Family Identifier (8 bits). Should be set to 1 (unicast)

RESERVED: One octet reserved for special flags

VECTOR/SERVICE NODE ADDRESS: The IPv4 or IPv6 unicast (or anycast) address of transit router.

Type 2 TLV:

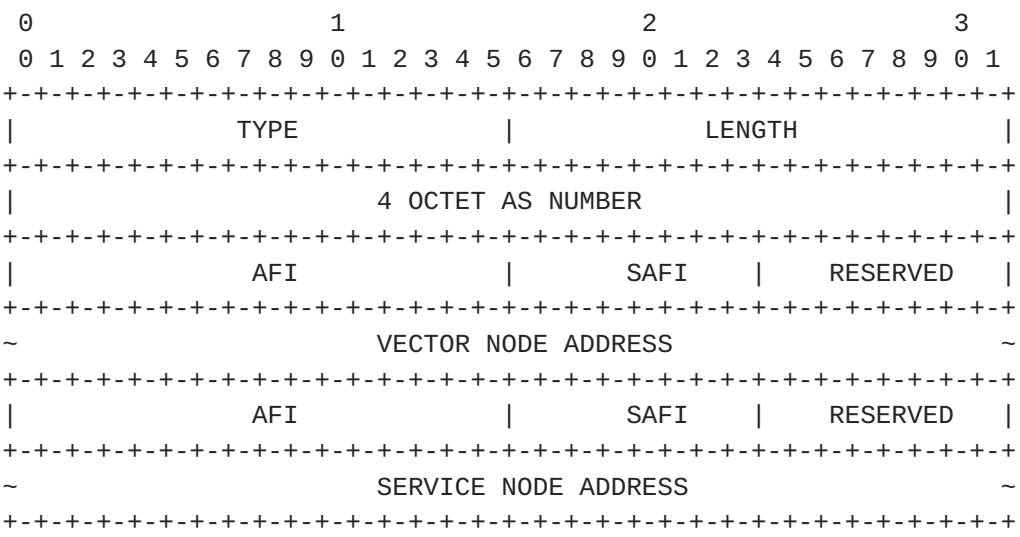


Figure2: Vector Node TLV Type 2

TYPE: Two octets encoding the Vector Node TLV Type. Type 2 contains vector node address and service node address which packets should traverse. Vector node address is part of the IGP and such node is part of BGP mesh. Service node is directly attached to a vector node, is reachable from vector node and does not run any BGP sessions.

LENGTH: Two octets encoding the length in octets of the Vector Node TLV, excluding the type and length fields. The Length is encoded as an unsigned binary integer.

4 OCTET AS NUMBER: 4 octet AS number or zero padded 2 octet AS number of the autonomous system Vector Node Address belongs

AFI: Address Family Identifier (16 bits).

SAFI: Subsequent Address Family Identifier (8 bits). Should be set to 1 (unicast)

RESERVED: One octet reserved for special flags

VECTOR/SERVICE NODE ADDRESS: The IPv4 or IPv6 unicast (or anycast) address of respectively a transit node and service appliance. Vector and service node may belong to different AFs.

3. Operation

The BGP Vector Node attribute is used to augment prefix or set of prefixes carried in given BGP UPDATE message with set of nodes information which are intended to be used to influence computation of forwarding paths for those destinations. The Vector Node attribute can be used within a provider's IBGP network and across EBGP networks. The BGP Vector Node attribute is an optional transitive attribute that can be applied to any address family within BGP where there is need for routing the traffic through ordered list of transit nodes.

The BGP Vector Node attribute consists of one more Vector Node TLVs. The ordered list of Vector Node TLVs indicates an ordered list of nodes that need to transit or process the data packets sent towards the destination prefix. The creation of the list of Vector Nodes is outside the scope of this document, but is expected to be created either through a Command Line Interface (CLI) on a router, or using an orchestrator system or by some other automated SDN computing engines.

The Vector Node attribute may be advertised by either an egress BGP speaker or injected by a non-egress node such as a BGP Route Reflector. It must be noted that in the event of non egress injection (e.g. a route reflector) extra assurance must be taken to achieve routing consistency.

Each BGP speaker which supports the BGP Vector Node attribute needs to process the attribute upon receipt and modify the NEXT HOP that node uses when installing the prefix in its local RIB/FIB. The rules to modify the NEXT HOP using the Vector Node attribute are as follows:

1 - Each BGP speaker involved in BGP Vector Routing only examines those TLVs which contains its own AS number. In an event where the BGP Vector node attribute is missing or if no Vector Routing TLVs with an AS number matching to BGP speaker's AS is present (BGP speaker fails the AS check criteria), a BGP speaker MUST use as the NEXT HOP from the received BGP MP_REACH attribute or a BGP NEXT HOP attribute in absense of a MP_REACH attribute.

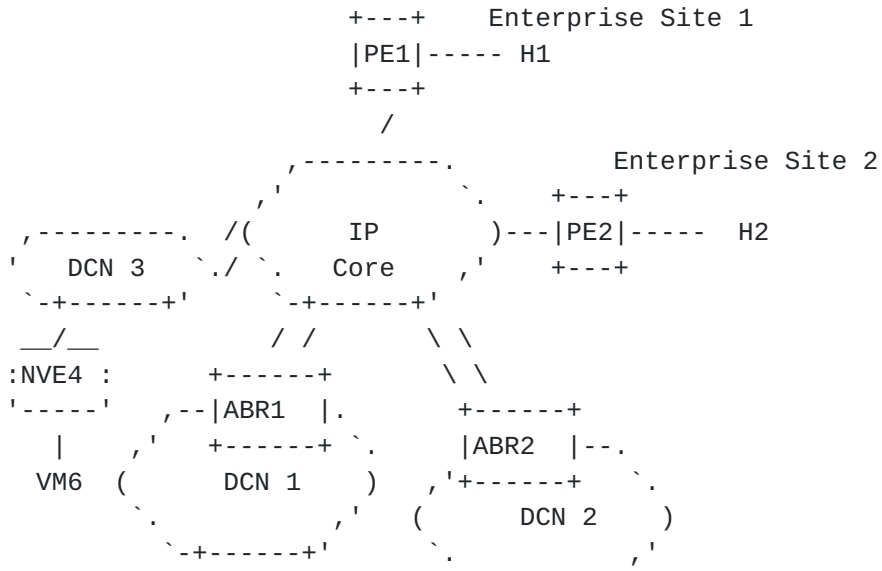
2 - In an event where the BGP speaker passes the AS check criteria for a given Vector TLV, a BGP speaker MUST use as the NEXT HOP of the prefix the first Node Address from the Vector Node Attribute TLVs if it does not find its own IGP node address (typically a loopback address) or if none of the Vector Node addresses belong to any of its connected interface subnets or are covered by any of the locally configured static routes when installing the route in its local RIB/ FIB.

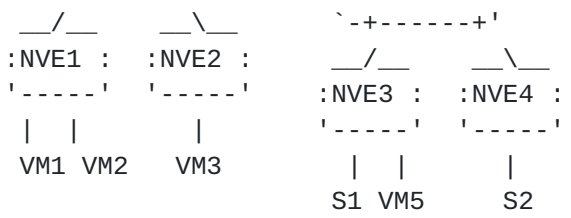
3 - In an event where the BGP speaker passes the AS check criteria for a given Vector TLV and if a BGP speaker finds its IGP node address (typically a loopback address) as one of the Vector node address, or if a BGP speaker finds its connected address as one of the Vector node address, or if the Vector node address is covered by any of the locally configured static route, then it MUST use as a NEXT HOP the next eligible Vector Node address from the Vector Node TLVs when installing the route in the RIB/FIB. In addition depending on the type of Vector Node TLV it may need to flag such a RIB/FIB entry with local punt or redirection for example to force Service Processing of type 2 Vector Node TLV.

4 - In an event where the BGP speaker passes the AS check criteria for a given Vector TLV and if a BGP speaker finds its IGP node address (typically a loopback address) as one of the Vector node address, or if a BGP speaker finds its connected address as one of the Vector node address, or if the Vector node address is covered by any of the locally configured static route, and if the found Vector node address is the last address in the TLV, then the BGP speaker MUST use NEXT HOP as a NEXT HOP address from the received BGP MP_REACH attribute or a BGP NEXT HOP attribute in absense of a MP_REACH attribute.

4. Use case example

As an example, consider the following scenario where VM1 attached to NVE1 needs to communicate with H1 attached to PE1. However, packets from VM1 to H1 need the services of S1 off of NVE3 and S2 off of NVE4 respectively. Therefore, the service chain of VM1 -> S1 -> S2 -> H1 needs to be formed for packets from VM1 to H1.





Lets assume VM1, VM3, S1, S2, and H1 are part of the same VPN and a same Autonomous System. PE1 advertises host route H1 with Vector Node Attribute of [I1, I2]; where I1 and I2 are interface subnet addresses corresponding to service nodes S1 and S2 respectively.

When NVE1 or NVE2 receives this advertisement, it applies rule (2) and subsequently setting the next hop address of H1 to I1 corresponding to service node S1. Therefore, when it receives packets destined to H1, it encapsulates the packets using any existing tunneled mechanisms and forwards them to the I1 address in NVE3.

When NVE3 receives this advertisement, it applies rule (3) by identifying its interface subnet I1 in the Vector Node attribute and subsequently setting the next hop address of H1 to I2 corresponding to service node S2. Therefore, when it receives packets from the network it forwards them to S1 and when it receives packets from its attached service node S1, destined to H1, it encapsulates the packets using any existing tunneled mechanisms and forwards them to the I2 address in NVE4.

When NVE4 receives this advertisement, it applies rule (4) by identifying its interface subnet I2 in the Vector Node attribute and since it is the last address in the Vector Node attribute list, it sets the next hop address of PE1 (received in the BGP advertisement) as the Next Hop for the prefix. Therefore when it receives packets from the network it forwards them to S2 when it receives packets destined to H1 from its attached service node S2, it encapsulates the packets using any existing tunneled mechanisms and forwards them to the PE1.

5. Deployment considerations

The BGP Vector Routing can be deployed for both Intra and Inter-domain networks without any restriction on version of IP address used as a Vector Node.

When using BGP Vector Routing and BGP multipath feature it is mandatory to assure consistent imposition of BGP Vector Node Attribute for a given prefix or group of prefixes from any imposition point in the network. When BGP speaker detects inconsistency across content of BGP Vector Routing Attribute across paths of the same prefix it is mandated to ignore such attribute and log a system warning.

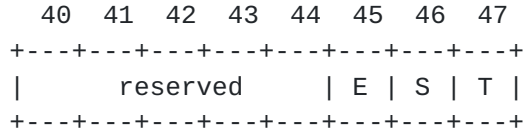
When using BGP Vector Routing marking from any points within the domain it is mandatory to assure consistency of application of BGP Vector Routing Attribuite in all injection points.

Use of mixed TLV types (type 1 and type 2 is allowed).

Reachability to BGP Vector Routing Nodes is resolved in exactly same manner as a reachability to traditional BGP Next Hops are resolved with the help of IGP routing. As such, BGP Vector Routing can use IGP Segment Routing rules to reach next BGP Vector Node.

This specification for its deployment simplicity assumes that BGP Vector Routing must be used with some form of IGP encapsulation between ingress, egress and all transit or service nodes. In particular IP encapsulation, MPLS encapsulation or Segment Routing can be used to transit packets within any IGP domain where BGP may not be present or BGP routers are not upgraded with new functionality.

In the presence of requirement for more selective then to entire IP destination packet handling (example separate port 80 http traffic from delay sensitve packets) the BGP Vector Node attribute can be attached to BGP update containing Dissemination of Flow Specification Rules RFC 5575 [RFC5575] where traffic action is defined as new E bit (Encapsulate).



E-bit - defines new action which results in encapsulation of matching packets to the next vector node as specified in the BGP Vector Node Attribute.

The rest of the encoding as well as validation rules remain unchanged as defined in RFC 5575 [RFC5575].

6. IANA Considerations

This document defines a new BGP attribute known as a BGP Vector Node attribute. The code point for a new BGP Vector Node attribute has to be assigned by IANA from the BGP Path Attributes registry.

7. Security considerations

No new security issues are introduced to the BGP protocol by this specification.

8. Acknowledgements

Authors would like to acknowledge Brian Field, Bruno Decraene and Ahmed Bashandy for their valuable input, review and comments.

9. References

9.1. Normative References

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

[RFC4223] Savola, P., "Reclassification of RFC 1863 to Historic", RFC 4223, October 2005.

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

9.2. Informative References

[I-D.keyupate-bgp-services]
Patel, K., Medved, J., and B. Pithawala, "Service Advertisement using BGP", draft-keyupate-bgp-services-02 (work in progress), April 2013.

[I-D.previdi-filsfils-isis-segment-routing]
Previdi, S., Filsfils, C., Bashandy, A., Horneffer, M., Decraene, B., Litkowski, S., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., and J. Tantsura, "Segment Routing with IS-IS Routing Protocol", draft-previdi-filsfils-isis-segment-routing-02 (work in progress), March 2013.

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

Authors' Addresses

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

Email: keyupate@cisco.com

Robert Raszuk
NTT I3
101 S. Ellsworth Ave
San Mateo, CA 94401
US

Email: robert@raszuk.net

Burjiz Pithawala
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: bpithaw@cisco.com

Eric Osborne
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: eosborne@cisco.com

Ali Sajassi
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
US

Email: sajassi@cisco.com

James Uttaro
ATT
200 S. Laurel Ave
Middletown, NJ 07748
USA

Email: uttaro@att.com

Luay Jalil
VeriZon
1201 E Arapaho Rd
Richardson, Texas 75081
USA

Email: luay.jalil@verizon.com