Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 19, 2016

Z. Li
Z. Zhuang
Huawei Technologies
S. Lu
Tencent

October 17, 2015

BGP Extensions for Service-Oriented MPLS Path Programming (MPP) draft-li-idr-mpls-path-programming-02

Abstract

Service-oriented MPLS programming (SoMPP) is to provide customized service process based on flexible label combinations. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path to the transport path. This document defines BGP extensions to support Service-oriented MPLS path programming.

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/.

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 April 19, 2016.

Copyright Notice

Table of Contents

1. Introduction

The label stack capability of MPLS would have been utilized well to implement flexible path programming to satisfy all kinds of service requirements. But in the distributed environment, the flexible programming capability is difficult to implement and always confined to reachability. As the introducing of central control in the network, the flexible MPLS programming capability becomes possible owing to two factors: 1. It becomes easier to allocate label for more purposes than reachability; 2. It is easy to calculate the MPLS path in a global network view. Moreover, the MPLS path programming capability can be utilized to satisfy more requirements of service bearing in the service layer which is defined as Service-oriented MPLS path programming. BGP will play an important role for MPLS path programming to download programmed MPLS path and map the service path

to the transport path. This document defines BGP extensions to support Service-oriented MPLS path programming.

2. Terminology

BGP: Border Gateway Protocol

EVPN: Ethernet VPN

L2VPN: Layer 2 VPN

L3VPN: Layer 3 VPN

MPP: MPLS Path Programming

MVPN: Multicast VPN

RR: Route Reflector

SR-Path: Segment Routing Path

NLRI: Network Layer Reachability Information

3. Architecture and Usecases of SoMPP

3.1. Architecture

The architecture of BGP-based MPLS path programming is shown in the Figure 1. Central control plays an important role in MPLS path programming. It can extend the MPLS path programming capability easily. The central controller can calculate path in a global network view and implement the MPLS path programming to satisfy different requirements of services. The result of MPLS path programming can be advertised from the central controller to the client nodes through BGP extensions to the ingress PEs. When client nodes receives the result of MPLS path programming, it will install the MPLS forwarding entry for the specified BGP prefix to implement the service process.

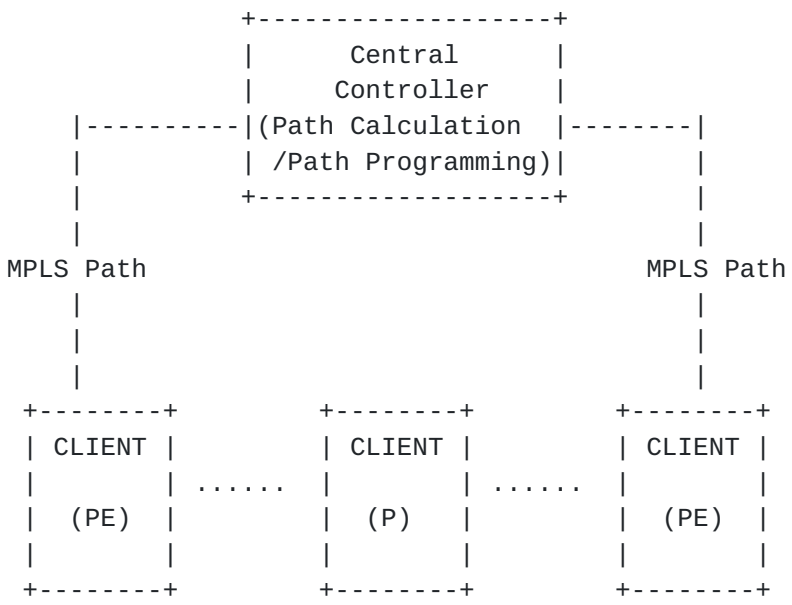


Figure 1 BGP-based MPLS Path Programming

3.2. Usecases

3.2.1. Deterministic ECMP

Entropy Label[RFC6790] is introduced to improve the ECMP capability by encapsulate the entropy label in the MPLS label stack. The existing implementation is always to calculate the entropy label based on the header of packets by specific hash algorithm in the ingress node. That is, the entropy label is determined locally by the ingress node. The method can improve the hash of packets in the network for load-sharing. But since the ingress node lacks the knowledge of the global traffic pattern of the network and calculates the entropy label by itself it may be not able to improve the ECMP capability accurately and in some cases it may deteriorate the imbalance of load-sharing.

With the central controlled MPLS path programming, the central controller can collect the global traffic pattern information of the network and based on the information deterministically calculate the entropy label for specific flows to help improve the load-sharing of the network. Then the central controller can download the label stack information with the deterministic entropy label to the ingress PEs for the specific BGP prefix. The ingress node can install the MPLS forwarding entry shown in the following figure to help optimize the ECMP of the flow specified by the BGP prefix, then optimize the ECMP of the whole network.

3.2.2. Centralized Mapping of Service to Tunnels

In the network there can be multiple tunnels to one specific destination which satisfy different constraints. In the traditional way, the tunnel is set up by the distributed forwarding nodes. As the PCE-initiated LSP setup [I-D.ietf-pce-pce-initiated-lsp]is introduced, the tunnel with different constraints can be set up in the central controlled way. In order to satisfy different service requirements, it is necessary to provide the capability to flexibly map the service to different tunnels which constraints can satisfy the required service requirement. Since the central controller has enough information of the whole network view, it can be an effective way to map the service (such as L3VPN and L2VPN) to the tunnel by the central controller and advertise the mapping information to the ingress PE of the service to guide the mapping in the forwarding node.

There can be two types of behaviors to map service to the tunnel:

1. Specify the tunnel type: with the method BGP will carry the tunnel type information for the BGP prefix. When the ingress PE receives the information, it will use the tunnel type and the nexthop address (or other specified target IP address) to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are more than one tunnels, the ingress PE will load share the traffic across all the tunnels.

2. Specify the specific tunnel: For MPLS TE/SR-TE tunnel, there can be multiple MPLS TE tunnels from one ingress PE to a specific destination with different constraints. BGP can carry the tunnel identifier information for the BGP prefix from the controller to the ingress node. When the ingress PE receives the information, it will use the tunnel identifier information to search the corresponding tunnels to bear the flow specified by the BGP prefix. If there are multiple tunnel identifiers, the ingress PE will load share the traffic across all the tunnels.

4. Download of MPLS Path

According to the service requirements, the central controller can combine MPLS labels flexibly. Then it can download the service label combination for specific prefix.BGP extensions are necessary to advertise label stacks for the prefix in NLRI field.

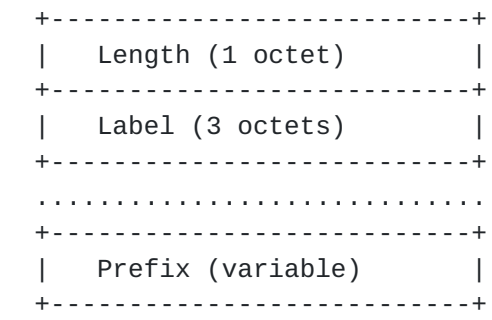


Figure 1: NLRI Definition in RFC3107

[RFC3107] defines above NLRI to advertise label binding for specific prefix. The label field can carry one or more labels. Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack". But for other AFI/SAFIs using label binding such as VPNv4, VPNv6, EVPN, MVPN, etc., it dose not support the capability to carry more labels for the specific prefix. Moreover for the AFI/SAFIs which do not support label binding capability originally, but may possibly adopt MPLS path programming now, there is no label field in the NLRI. In order to support flexible MPLS path programming, this document defines and uses a new BGP attribute called the "Extended Label attribute". This is an optional transitive BGP attribute. The format of this attribute is defined as follows:

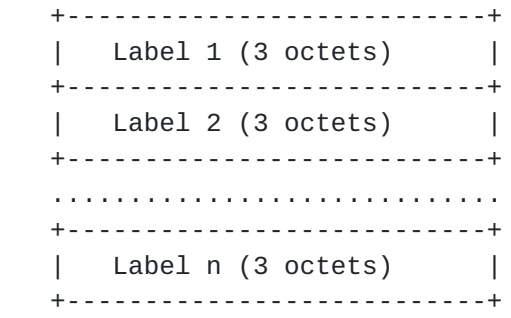


Figure 2: Extended Label Attribute

The Label field carries one or more labels (that corresponds to the stack of labels [[RFC3032]]). Each label is encoded as 3 octets, where the high-order 20 bits contain the label value, and the low order bit contains "Bottom of Stack" (as defined in [[RFC3032]]).

The central controller for MPLS path programming could build a route with Extended Label attribute and send it to the ingress routers.

Upon receiving such a route from the central Controller, the ingress router SHOULD select such a route as the best path. If a packet

comes into the ingress router and uses such a path, the ingress router will encapsulate the stack of labels which is derived from the Extended Label Attribute of the route into the packet and forward the packet along the path.

The "Extended Label attribute" can be used for various BGP address families. Before using this attribute, firstly, it is necessary to negotiate the capability between two nodes to support MPLS path programming for a specific BGP address family. If negotiation fails, a node MUST NOT send this attribute and MUST discard this attribute when it receives.

5. Download of Mapping of Service Path to Transport Path

5.1. Specify Tunnel Type

[I-D.ietf-idr-tunnel-encaps] proposes the Tunnel Encapsulation Attribute which can be used without BGP Encapsulation SAFI to specify a set of tunnels. It defines a series of Encapsulation Sub-TLVs for particular tunnel types. It also defines the Remote Endpoint Attributes Sub-TLV to specify the remote tunnel endpoint address for each tunnel which can be different the BGP nexthop. The Tunnel Encapsulation Attributes can be reused for the MPLS path programming to specify the tunnel types, the encapsulation and the remote tunnel endpoint address which can determine a set of tunnels which the service can map to. Now the limited MPLS tunnel types are defined for the Tunnel Encapsulation Attributes. In order to support MPLS path programming, the following MPLS tunnel types are to be defined:

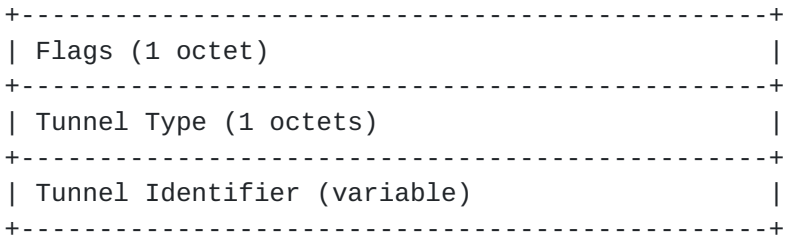
Value	Tunnel Type
TBD	LDP LSP
TBD	RSVP-TE LSP
TBD	MPLS-based Segment Routing Best-effort Path
TBD	MPLS-based Segment Routing Traffic Engineering Path

5.2. Specify Specific Tunnel

Besides specifying the tunnel types to determine the set of tunnels which the service traffic can map to, the specific tunnels can be specified directly by the tunnel identifiers when map the service traffic to the path. BGP extensions is necessary that through the community attribute of BGP the identifier of the transport path can be carried when advertise the specific prefix.

In order to support the application, this document defines a new BGP attribute called the "Extended Unicast Tunnel attribute". This is an

optional transitive BGP attribute. The format of this attribute is defined as follows:



The Flags is reserved and must be set as zero. The Tunnel Type identifies the type of the tunneling technology used for the unicast service path. The tunnel type determines the syntax and semantics of the Tunnel Identifier field. This document defines following Tunnel Types:

+ 0 - No tunnel information present
+ 1 - RSVP-TE LSP
+ 2 - MPLS-based Segment Routing Traffic Engineering Path

Tunnel Specific Attributes contains the attributes of the tunnel. The field is optional. The value depends on the tunnel type. It will be defined in the future versions.

When the Tunnel Type is set to "No tunnel information present", the Tunnel attribute carries no tunnel information (no Tunnel Identifier). when the type is used, the tunnel used for the service path is determined by the ingress router.

When the Tunnel Type is set to RSVP - Traffic Engineering (RSVP-TE) Label Switched Path (LSP), the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address> as specified in [RFC3209] If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The other fields in the RSVP-TE LSP Identifier are the same as specified in [RFC3209].

When the Tunnel Type is set to MPLS-based Segment Routing Traffic Engineering Path, the Tunnel Identifier is <C-Type, Tunnel Sender Address, Tunnel ID, Tunnel End-point Address>. If C-Type = 7, Tunnel Sender Address and Tunnel End-point Address are IPv4 address in 4 octets. If C-Type = 8, Tunnel Sender Address and Tunnel End-point Address are IPv6 address in 16 octets. The tunnel identifier is similar as that of RSVP-TE LSP.

BGP can carry multiple Extended Unicast Tunnel Attributes for specific prefix. If there are multiple tunnel identifiers, the ingress PE will load share the traffic across all the specified tunnels for the service traffic determined by the specific BGP prefix.

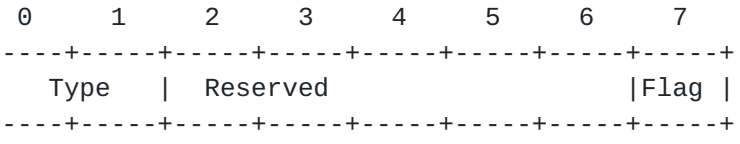
6. Route Flag Extended Community

In order to make the MPLS path programming to take effect, the route advertised by the central controller after the MPLS Path Programming should be selected by the ingress PE over other routes for the same BGP prefix. There are two options of BGP extensions for the purpose:

Option 1: A new BGP Extended Community called as the "Route Flag Extended Community" can be introduced. The Type value is to be assigned by IANA.

The Route Flag Extended Community is used to carry the flag appointed by the BGP central controller.

The format of this extended community is defined as follows:



Flag = 0, Treat as normal route
Flag = 1, Treat as best route

When a router receives a BGP route with a Route Flag Extended Community and the Flag set to "1", it SHOULD use the route as the best route when select the route from multiple routes for a specific prefix.

Option 2: [I-D.ietf-idr-custom-decision] defines a new Extended Community, called the Cost Community, which can be used in tie breaking during the best path selection process. The Cost Community can be reused by the MPLS path programming to set the "Point of Insertion" as 128 to make the route advertised by the central controller to be chosen.

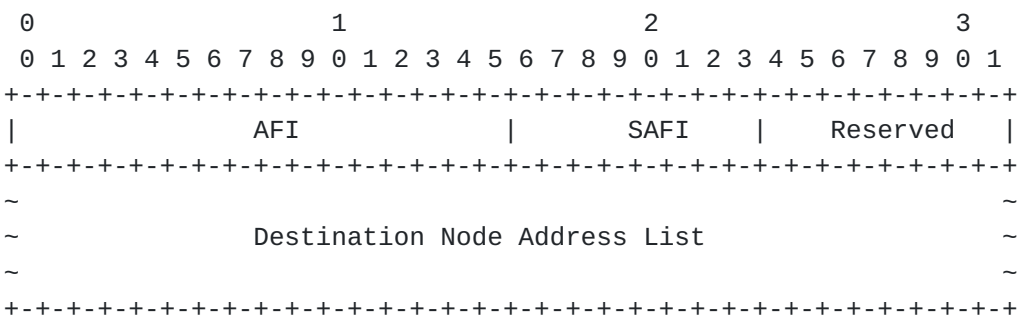
7. Destination Node Attribute

This document defines and uses a new BGP attribute called as the "Destination Node attribute" which Type value is to be assigned by

IANA. The Destination Node attribute is an optional non-transitive attribute that can be applied to any address family.

The Destination Node attribute is used to carry a list of node addresses, which are intended to be used to determine the nodes where the route with such attribute SHOULD be considered. If a node receives a BGP route with a Destination Node attribute, it MUST check the node address list. If one address of the list belongs to this node, the route MUST be used in this node. Otherwise the route MUST be ignored silently.

The format of this attribute is defined as follows:



AFI: Address Family Identifier (16 bits).

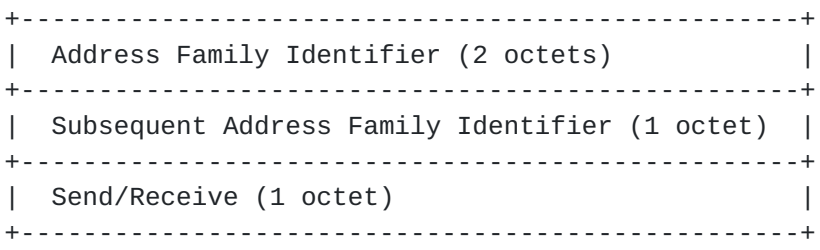
SAFI: Subsequent Address Family Identifier (8 bits).

Reserved: One octet reserved for special flags

Destination Node Address List: The list of IPv4 (AFI=1) or IPv6 (AFI=2) address.

8. Capability Negotiation

It is necessary to negotiate the capability to support MPLS path programming. The MPLS-Path-Programming Capability is a new BGP capability [RFC5492]. The Capability Code for this capability is to be specified by the IANA. The Capability Length field of this capability is variable. The Capability Value field consists of one or more of the following tuples:



The meaning and use of the fields are as follows:

Address Family Identifier (AFI): This field is the same as the one used in [RFC4760].

Subsequent Address Family Identifier (SAFI): This field is the same as the one used in [RFC4760].

Send/Receive: This field indicates whether the sender is (a) willing to receive programming MPLS paths from its peer (value 1), (b) would like to send programming MPLS paths to its peer (value 2), or (c) both (value 3) for the <AFI, SAFI>.

9. IANA Considerations

TBD.

10. Security Considerations

TBD.

11. References

11.1. Normative References

[I-D.ietf-idr-custom-decision]
Retana, A. and R. White, "BGP Custom Decision Process", draft-ietf-idr-custom-decision-06 (work in progress), April 2015.

[I-D.ietf-idr-tunnel-encaps]
Rosen, E., Patel, K., and G. Velde, "Using the BGP Tunnel Encapsulation Attribute without the BGP Encapsulation SAFI", draft-ietf-idr-tunnel-encaps-00 (work in progress), August 2015.

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

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, DOI 10.17487/RFC3032, January 2001, <http://www.rfc-editor.org/info/rfc3032>.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <http://www.rfc-editor.org/info/rfc3209>.

[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, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <http://www.rfc-editor.org/info/rfc4760>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <http://www.rfc-editor.org/info/rfc5036>.

[RFC5492] Scudder, J. and R. Chandra, "Capabilities Advertisement with BGP-4", RFC 5492, DOI 10.17487/RFC5492, February 2009, <http://www.rfc-editor.org/info/rfc5492>.

11.2. Informative References

[I-D.ietf-pce-pce-initiated-lsp] Crabbe, E., Minei, I., Sivabalan, S., and R. Varga, "PCEP Extensions for PCE-initiated LSP Setup in a Stateful PCE Model", draft-ietf-pce-pce-initiated-lsp-04 (work in progress), April 2015.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, DOI 10.17487/RFC3107, May 2001, <http://www.rfc-editor.org/info/rfc3107>.

[RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, DOI 10.17487/RFC6790, November 2012, <http://www.rfc-editor.org/info/rfc6790>.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Sujian Lu
Tencent
Tengyun Building,Tower A ,No. 397 Tianlin Road,Xuhui District
Shanghai 200233
China

Email: jasonlu@tencent.com