Network Working Group
Internet Draft
Expires: December 2006

Gargi Nalawade
Ruchi Kapoor
Dan Tappan
Scott Wainner
Simon Barber
Chris Metz

Cisco Systems

# BGP Tunnel SAFI

draft-nalawade-kapoor-tunnel-safi-05.txt

## Status of this Memo



## Abstract

There is a growing requirement for network operators to support multi-address familiy routing and forwarding services across their backbone networks. In general this is accomplished by constructing a mesh of tunnels between the backbbone provider edge routers and then advertising reachability to prefixes through specific tunnels. This

document defines a new subsequence address family identifier associated with a tunnel end-point information. This enables a single egress provider edge router to use the border gateway protocol as a scalable and efficient means to distribute its tunnel end-point information to many ingress provider edge routers. The result is that the mesh of tunnels is in place and packets can be forwarded through these tunnels based on advertised reachability.

## 1. Introduction

There is a growing requirement for network operators to support multi-address familiy routing and forwarding services across their backbone networks. In the context of network-based IP VPN, this is accomplished today using the mechanisms defined in RFC4364. More recently the softwires effort has emerged as a generalized, network-based routing and forwarding solution supporting connectivity of address familiy islands (e.g. IPv4, IPv6, VPNv4, VPNv6) across a uniform IPv4 or IPv6 backbone network [SW-MESH-FMWK]. In both cases the establishment of tunnels (IP or MPLS) between ingress and egress provider edge (PE) routers must be in place before packets of one address familiy can be tunneled across the backbone network.

Two end-points of a tunnel need to agree upon the end-point information and its binding to a network address at the remote point. Normally, this information can be manually shared and statically configured when the number of tunnels to manage is relatively small. In the case of a network such as an MPLS VPN where there is a need for a tunnel between every ingress and egress PE, the number of tunnel end-points that need to be exchanged and maintained grows dramatically as the network becomes large. The egress PE already defines reachability information for the private routing information as well as the NLRI of the PE itself. This information is distributed via MP-BGP to any number of potential ingress PE. The extent of distribution of egress PE's NLRI and next-hop is unknown by the egress PE; therefore, egress PE cannot feasibly know the tunnel attributes for any potential ingress PE unless the egress PE assigns these attributes. The egress PE needs to advertise it's capability to receive tunneled packets, the types of tunnels supported, the preference for the various tunnel methods, and the attributes associated with the tunnels. The tunnel information then needs to be distributed and maintained using MP-BGP such that every potential ingress PE knows the appropriate tunnel method and attributes of the egress PE. The tunnel capabilities are uniquely defined for a given PE and may or may not correlate with the capabilities of any other potential ingress PE. For this reason, the ingress PE may select the most appropriate tunneling mechanism based on the compability of the

tunnel capabilities between the ingress and egress PE's and their preferences.

## 2. The Tunnel SAFI

This document defines a new BGP SAFI called the Tunnel SAFI. The <AFI, SAFI> [IANA-AFI] [IANA-SAFI] value pair used to identify this SAFI are: AFI=1, SAFI=64, for the IPv4 Tunnel address family; and AFI=2, SAFI=64 for the IPv6 Tunnel address family.

For BGP Speakers supporting [BGP-4], the tunnel end point address will be carried as an NLRI in the MP_REACH attribute for the Tunnel SAFI.

The NLRI will be encoded as a 2-octet Identifier followed by the NLRI format as specified by the respective AFI. The Identifier will identify the tunnel end point being advertised. This Identifier enables multiple tunnels to share the same network address, thus conserving the number of addresses needed to be configured by the operator on each of the Tunnel-endpoints. The network address contained in the Tunnel SAFI NLRI is the network address of the tunnel end point.

The network address contained in the BGP Tunnel SAFI NLRI SHOULD be the same as the network address carried in the 'Network Address of Next Hop' field of the BGP Softwire Nexthop Attribute [BGP-SW-NEXT-HOP]. The BGP Softwire Nexthop Attribute will be carried separately in BGP advertisements, as described in [BGP-SW-NEXT-HOP].

## 3. BGP Encapsulation Attributes

The BGP Tunnel SAFI will carry the tunnel end-point information inside a BGP encapsulation attribute. The encapsulation attribute used can be either the BGP Tunnel Encapsulation Attribute [BGP-TUN] or the BGP Softwire Mesh encapsulation attribute [BGP-SW-ENCAP]. The egress PE may support one or more tunnel methods. The egress PE MUST advertise all tunnel types for which it will support tunnel termination. The egress PE MAY advertise one or more tunnel types.

If a BGP Speaker supports the BGP Tunnel SAFI then it MUST understand the Tunnel Encapsulation attribute [BGP-TUN]. A BGP update for the Tunnel SAFI MUST contain either the BGP Tunnel Encapsulation Attribute [BGP-TUN] or the BGP Softwire Mesh encapsulation attribute [BGP-SW-ENCAP]. A BGP update for the Tunnel SAFI MUST NOT contain both the BGP Tunnel Encapsulation Attribute [BGP-TUN] and the BGP Softwire Mesh encapsulation attribute [BGP-SW-ENCAP] in the same update message. If such an update message is received by a BGP speaker, the message should be ignored.

The details of the contents of the BGP Tunnel Encapsulation Attribute

[BGP-TUN] are described in the section below.

### 3.1 Contents of BGP Tunnel Encapsulation Attrubite.

As defined in [BGP-TUN], the first bit of the TYPE field in the BGP Tunnel Encapsulation Attribute is the 'transitive bit'. If the bit value is 1, implies that this tunnel is transitive. If the bit value is 0, it implies this specific tunnel is not transitive.

The Value Field of the BGP Tunnel Encapsulation Attribute, MUST contain at least one of the following valid Type codes for this SAFI.
It MAY contain one or more TLVs with these Type codes.

Type 1: L2TPv3 Tunnel information

Type 2: mGRE Tunnel information

Type 3: IPSec Tunnel information

Type 4: MPLS Tunnel information

Type 5: L2TPv3 in IPSEC Tunnel information

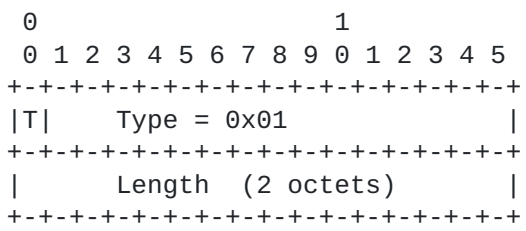
Type 6: mGRE in IPSEC Tunnel information

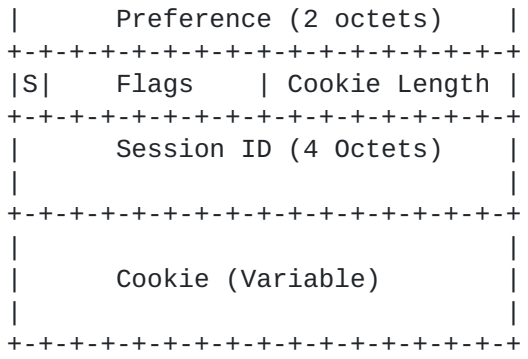
#### 3.1.1 L2TPv3 Tunnel information TLV

The L2TPv3 Tunnel Information TLV has a type value of 1. The value part of the L2TPv3 Tunnel Information Type contains the following :

- Preference (2 Octets)
- Flags (1 Octet)
- Cookie Length (1 Octet)
- Session ID (4 Octets)
- Cookie (Variable)

The L2TPv3 Tunnel Information TLV looks as follows :






where

Length - A 2 Octet field that specifies the length of the L2TPv3 attribute in octets.

Preference - A 2 Octet field containing a Preference associated with the TLV. The Preference value indicates a preferred ordering of tunneling encapsulations according to the sender (i.e. egress PE). The recipient of the information SHOULD take the sender's preference into account in selecting which encapsulation it will use. A higher value indicates a higher preference.

Flags - A 1 Octet field containing flag-bits. The leftmost bit indicates whether Sequence numbering is to be used or not. The remaining bits are reserved for future use.

Cookie Length - is a 1 Octet field that contains the length of the Variable length Cookie.

Session ID - A 4 Octet field containing a non-zero identifier for a session. The Session ID is used to delineate services on the egress PE. The support for a service such as MPLS VPN MUST have at least one Session ID assigned. Multiple Session ID's may be assigned for the same service instance. The primary motivation for assigning multiple Session ID's for the same service instance is provide a graceful transition when changing cookie values. The egress PE can receive both Session ID's with their unqiue Cookie value thus allowing a graceful roll-over from an old Session ID and Cookie to a new Session ID and Cookie. Alternatively, multiple service instances may be distributed across multiple processes in order to scale. Each service instance may be assigned a unique Session ID and Cookie and advertised by BGP such that packets received from the ingress PE are directed to the appropriate service instance on the egress PE.

Cookie - Cookie is a variable length (maximum 64 bits), value used by
L2TPv3 to check the association of a received data message with the session identified by the Session ID. The Cookie value is tightly coupled with the Session ID. Upon the generation of a Session ID by the egress PE, the associated Cookie MAY be generated such that packets received by the egress PE from an ingress PE can be quickly validated for proper service context.

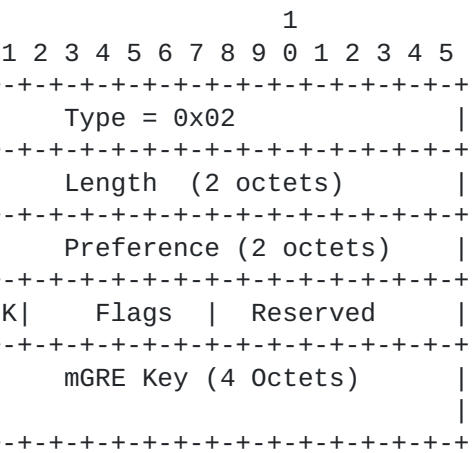
The default value of the Length Field for the L2TPv3 Tunnel information TLV is between 8 and 16 bytes, depending on the length of the Cookie field specified in Cookie length. If the length of the TLV is greater than that value, the subsequent portion of the Value field contains one or more sub-TLVs as defined in [BGP-TUN].

### 3.1.2 mGRE Tunnel Information TLV

The mGRE Tunnel Information Type has a Type 2. The value part of the mGRE Tunnel Information Type contains the following :

- Preference (2 Octets)
- Flags (1 Octet)
- mGRE Key (0 or 4 Octets)

The mGRE Tunnel Information TLV looks as follows :




Length - A 2 Octet field that specifies the length of the mGRE information in octets.

Preference - A 2 Octet field containing a Preference associated with

the TLV. The Preference value indicates a preferred ordering of tunneling encapsulations according to the sender (i.e. egress PE). The recipient of the information (i.e. ingress PE) SHOULD take the sender's preference into account in selecting which encapsulation it will use. A higher value indicates a higher preference.

Flags - A 1 Octet field containing flag-bits. The leftmost bit indicates whether Sequence numbering is to be used or not. The 2nd bit Indicates whether an mGRE Key is present or not. The Remaining bits are reserved for future use.

Reserved - A 1 Octet field reserved for future use

mGRE Key - A 4 Octet field containing an optional mGRE Key. The key value may be generated by the egress PE and advertised by the egress PE to any potential ingress PE. In this case, the key value has unidirectional relevance from all viable ingress PE's to the egress PE. Alternatively, the key value may be statically configured such that all ingress and egress PE's use the same key value.

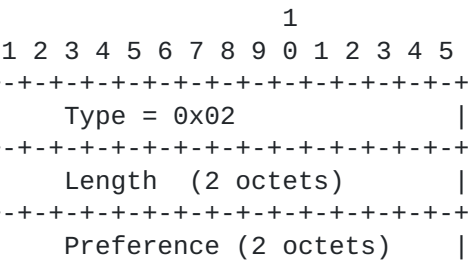
If the Length field of the TLV contains a value greater than 3 Octets plus the value specified in the Key Length, the subsequent portion of the Value field contains one or more sub-TLVs as defined by [BGP-TUN].

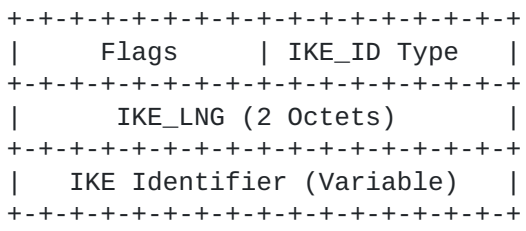
### 3.1.3 IPSec Tunnel Information TLV

The IPSec Tunnel Information Type has a Type 3. The value part of the IPSec Tunnel Information Type contains the following :

- Preference (2 Octets)
- Flags (1 Octet)
- IKE ID Type (1 Octets)
- IKE ID Length (2 Octets)
- IKE Identifier (Variable)

The IPSec Tunnel Information TLV looks as follows :






Length - A 2 Octet field that specifies the length of the IPSec information in octets.

Preference - A 2 Octet field containing a Preference associated with the TLV. The Preference value indicates a preferred ordering of tunneling encapsulations according to the sender. The recipient of the information SHOULD take the sender's preference into account in selecting which encapsulation it will use. A higher value indicates a higher preference.

Flags - A 1 Octet field containing flag-bits.

IKE_ID Type - This 1 Octet field identifies the type of IKE Identifier used by the egress PE

IKE_LNG - This 2 Octet field indicates the length of the IKE Identifier.

IKE Identifier - A variable length field containing an IKE Identifier of the egress PE.

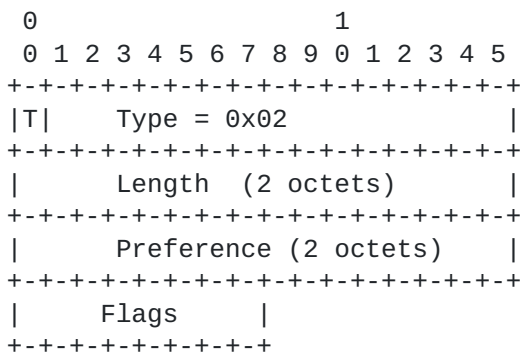
If the Length field of the TLV contains a value greater than 11 Octets plus the value specified in the Key Length, the subsequent portion of the Value field contains one or more sub-TLVs as defined by [BGP-TUN].

### 3.1.4 MPLS TLV

The MPLS TLV has a Type 4. The value part of the MPLS TLV contains the following :

- Preference (2 Octets)
- Flags (1 Octet)

The MPLS Tunnel Information TLV looks as follows :




Length A 2 Octet field that specifies the length of the MPLS TLV in octets.

Preference A 2 Octet field containing a Preference associated with the TLV. The Preference value indicates a preferred ordering of tunneling encapsulations according to the sender. The recipient of the information SHOULD take the sender's preference into account in selecting which encapsulation it will use. A higher value indicates a higher preference.

Flags - A 1 Octet field containing flag-bits.

### 3.1.5 L2TPv3 in IPSEC TLV

When the value in the Type field is 5, the Value portion of the SAFI-Specific Attribute TLV will carry an IPSec TLV followed by an L2TPv3 TLV.

### 3.1.6 mGRE in IPSEC TLV

When the value in the Type field is 6, the Value portion of the SAFI-Specific Attribute TLV will carry an IPSec TLV followed by an mGRE TLV.

## 4. Capability Advertisement

A BGP speaker MAY participate in the distribution of the IPv4 Tunnel address family or IPv6 Tunnel address family information. A BGP speaker that wishes to exchange the IPv4 Tunnel address family or the IPv6 Tunnel address family, MUST use the MP_EXT Capability Code as defined in [BGP-MP], to advertise the corresponding (AFI, SAFI) pair.

## 5. Operation

A BGP Speaker that receives the Capability for the IPv4 Tunnel address family or the IPv6 Tunnel address family, MAY advertise the IPv4 Tunnel address family or IPv6 Tunnel address family prefixes to that peer.

The BGP Tunnel Encapsulation attribute is defined only to be used in UPDATE messages for the IPv4 tunnel address family or the IPv6 Tunnel address family. If the BGP Tunnel Encapsulation Attribute is received in an UPDATE message for any other AFI/SAFI, it MUST be ignored.

If a BGP Speaker receives an unrecognized Transitive Tunnel Encapsulation TLV as part of the BGP Tunnel Encapsulation Attribute, it MUST accept it and propagate it to other peers.

## 6. Deployment Considerations

In order for the Tunnels to come up between two end-points, the BGP Speakers advertising the Tunnel end-points using the IPv4/IPv6 Tunnel SAFI, MUST exchange at least one common encapsulation option.

## 7. Applicability

### 7.1. IPSec Tunnels Applicability

IPSec protection of IP routed packets requires the establishment of an IPSec proxy that specifies the source and destination range of addresses that require protection. The synchronization of the IPSec proxy and the viability of the path to the destination IP address range has been a persistent problem in the deploy of IPSec solutions.
The IPSec proxy must be associated with an IKE end-point identifier.
IPSec is inherently a tunneling protocol; however, it has no means of synchronizing the viability of the destination path in the IPSec proxy. One approach to synchronizing the IPSec proxy, the IKE end-point and the path viability is to leverage BGP Tunnel SAFI. The BGP protocol provides a means of distributing the destination address range of the IPSec proxy via the NLRI. The IKE end-point identifier may be consistent with the BGP next-hop and may be specified by the TLVs in the BGP Tunnel Encapsulation Attribute [BGP-TUN] in the BGP tunnel SAFI. An IPSec end-point that receives a BGP announcement may qualify the update and use the NLRI prefix as the destination range in the IPSec proxy. The IPSec end-point may learn the remote peer's IKE identity that is defined by the next-hop attribute of the Tunnel SAFI. The route viability is Inherently conveyed via the BGP protocol. The combination of the traditional IP NLRI and the Tunnel NLRI allows IPSec to automatically establish the connection attributes required to protect IP traffic between the two end-

points.

### 7.2. IP Tunnels Applicability

Multiprotocol Label Switching (MPLS) VPN introduced a peer-to-peer model that enables large scale IP VPN implementations. Traditional MPLS VPNs rely on an MPLS transport network to implement this peer-to-peer model. the MPLS transport with an IP transport. VPN traffic is carried by an IP tunnel instead of an MPLS Label Switched Path (LSP). The VPN customer receives the same service experience regardless of the transport choice used by the service provider.

MPLS VPN uses the same mechanisms for VPN route distribution regardless of the backbone transport choice (IP or MPLS). Customer edge (CE) devices exchange routing information with the provider edge (PE) devices using BGP or an Interior Gateway Protocol (IGP) protocol. This routing information is exchanged between PEs using Multi-Protocol BGP (MP-BGP). VPN routing information is carried by MP-BGP as VPNv4 addresses. As part of this VPN route exchange, PEs learn the nexthop (egress PE) and a VPN label to be associated with each VPN route.

Before proper VPNv4 BGP next hop resolution can take place, each PE needs to know which other PEs (i.e. Tunnel endpoints) are reachable via the IP tunnel.

The Tunnel SAFI update messages provide a means of distributing the Tunnel endpoint address as the NLRI in the Tunnel SAFI UPDATE. The Tunnel endpoint address should be consistent with the BGP next-hop in the VPNv4 update messages. This information is used to determine which IP tunnel needs to be used for which VPNv4 prefixes.

In addition, each PE needs to know the tunnel attributes (used to define this tunnel) that other PEs expect, so VPN packets can be encapsulated appropriately. Manual configuration of this information is not scalable, as the number of PEs increases. A PE that receives the Tunnel SAFI update may use the tunnel NLRI prefix and the tunnel attributes specified by the other end, and try and establish a tunnel to that endpoint. PEs take advantage of the existing MP-BGP infrastructure to distribute tunnel endpoint information. The Tunnel SAFI UPDATE message is used to signal tunnel attribute and endpoint information amongst PEs. And thus tunnel endpoint discovery is accomplished using MP-BGP updates.

## 8. Security Considerations

This extension to BGP does not change the underlying security issues.

## 9. Acknowledgements

The authors would like to thank Jim Guichard, Francois LeFaucher and David Ward for their contribution. We would like to thank Arjun Sreekantiah, Shyam Suri, Chandrashekhar Appanna, John Scudder and Mark Townsley for their comments and suggestions.

## 10. References

[IANA-AFI] http://www.iana.org/assignments/address-family-numbers

[IANA-SAFI] http://www.iana.org/assignments/safi-namespace

[BGP-4] Rekhter, Y. and T. Li (editors), "A Border Gateway Protocol 4 (BGP-4)", Internet Draft draft-ietf-idr-bgp4-26.txt, April 2005.

[BGP-CAP] Chandra, R., Scudder, J., "Capabilities Advertisement with BGP-4", draft-ietf-idr-rfc2842bis-02.txt, April 2002.

[BGP-TUN] Kapoor R., Nalawade G., "BGPv4 Tunnel Encapsulation Attribute", draft-nalawade-kapoor-idr-bgp-ssa-03.txt, work in progress.

[MULTI-BGP] Bates et al, "Multiprotocol Extensions for BGP-4", draft-ietf-idr-rfc2858bis-02.txt, work in progress.

[SW-MESH-FMWK] Metz, C. et al, "A Framework for Softwire Mesh Signaling, Routing and Encapsulation across IPv4 and IPv6 Backbone Networks", draft-wu-softwire-mesh-framework-00, June 2006.

[BGP-SW-NEXT-HOP] Nalawade G. et al, "BGP Softwire Nexthop Attribute", draft-nalawade-sw-nhop-00.txt, June 2006.

[BGP-SW-ENCAP] Nalawade G., Barber S., Ward D., Kapoor R., Metz C., "BGPv4 Softwire Mesh Encapsulation Attribute", draft-softwire-mesh-encap-attribute-00.txt, June 2006.

## 11. Authors' Addresses

Gargi Nalawade
Cisco Systems, Inc
170 West Tasman Drive
San Jose, CA 95134
mailto:gargi@cisco.com

Ruchi Kapoor
Cisco Systems, Inc
170 West Tasman Drive
San Jose, CA 95134

mailto:ruchi@cisco.com

Dan Tappan
Cisco Systems, Inc
170 West Tasman Drive
San Jose, CA 95134
mailto:tappan@cisco.com

Scott Wainner
Cisco Systems, Inc
13600 Dulles Technology Drive
Herndon, VA 20171
mailto:swainner@cisco.com

Simon Barber
Cisco Systems, Inc
mailto:sbarber@cisco.com

Chris Metz
Cisco Systems, Inc
170 West Tasman Drive
San Jose, CA 95134
mailto:chmetz@cisco.com

## 12. Intellectual Property Statement

Director.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.

## 13. Full Copyright Statement



## 14. Expiration Date

This memo is filed as <draft-nalawade-kapoor-tunnel-safi-05.txt>, and expires December, 2006.