Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 27, 2011

H. Grover
D. Rao
D. Farinacci
Cisco Systems
October 24, 2010

Overlay Transport Virtualization
draft-hasmit-otv-01

Abstract

In today's networking environment most enterprise networks span multiple physical sites. Overlay Transport Virtualization (OTV) provides a scalable solution for L2/L3 connectivity across different sites using the currently deployed service provider and enterprise networks. It is a very cost-effective and simple solution requiring deployment of a one or more OTV functional device at each of the enterprise sites. This solution is agnostic to the technology used in the service provider network and connectivity between the enterprise and the service provider network. This document provides an overview of this technology.

Table of Contents

1. Overview

OTV is a new "MAC in IP" technique for supporting L2 VPNs over an L2/L3 infrastructure. OTV provides an "over-the-top" method of doing virtualization among a large number of sites where the routing and forwarding state is maintained at the network edges, but not within the site or in the core.

OTV can be incrementally deployed and reside in a small number of devices at the edge between sites and the core. We call these devices "Edge Devices" which perform typical layer-2 learning and forwarding functions on their site facing interfaces (internal interfaces) and perform IP-based virtualization functions on their core facing interfaces (for which an overlay network is realized).

Traditional L2VPN technologies rely heavily on tunnels. Rather than creating stateful tunnels, OTV encapsulates layer 2 traffic with an IP header ("MAC in IP"), but does not create any fixed tunnels. Based on the IP header, traffic is forwarded natively in the core over which OTV is being deployed. This is an important feature as the native IP treatment of the encapsulated packet allows optimal multi-point connectivity as well as optimal broadcast and multicast forwarding, plus any other benefits the routed core may provide to native IP traffic. OTV virtualization is independent of the technology deployed in the core; the core network may be a layer-2 metro Ethernet core, a layer-3 IP network core, or an MPLS network core.

Layer-2 traffic which requires traversing the overlay to reach its destination, is prepended with an IP header which ensures the packet is delivered to the edge boxes that provide connectivity to the Layer-2 destination in the original MAC header. As shown in figure 1, if a destination is reachable via Edge Device X2 (with a core facing IP address of IPB), other Edge Devices forwarding traffic to such destination will add an IP header with a destination IP address of IPB and forward the traffic into the core. The core will forward traffic based on IP address IPB, once the traffic makes it to Edge Device X2 it will be stripped of the overlay IP header and it will be forwarded into the site in the same way a regular bridge would forward a packet at layer-2. Broadcast or multicast traffic is encapsulated with a multicast header and follows a similar process.

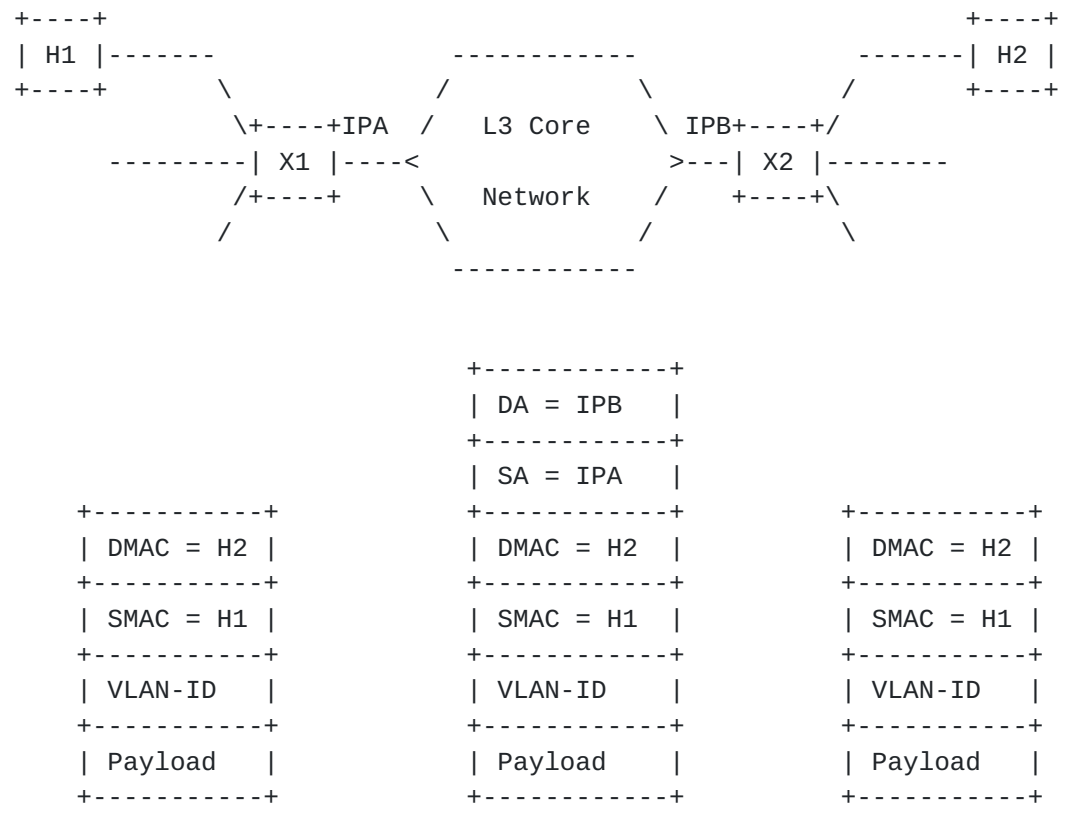


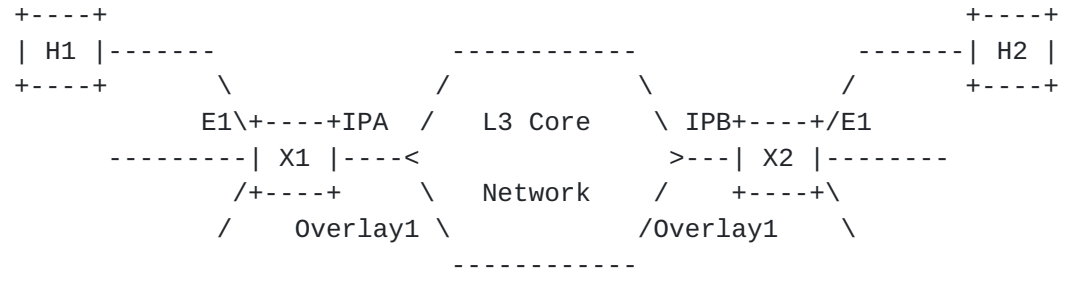
Figure 1. Traffic flow from H1 to H2 with encapsulation in the core.

The key piece that OTV adds is the state to map a given destination MAC address in the L2 VPN to an IP address of the OTV Edge Device behind which that MAC address is located. OTV forwarding is a function of mapping a destination MAC address in the VPN site to an Edge Device IP address in the overlay network.

To achieve all this, a control plane is required to exchange the reachability information among the different OTV Edge Devices. We will refer to this control plane as the oURP and oMRP (Overlay Unicast Routing Protocol and Overlay Multicast Routing Protocol). OTV does not flood unknown unicast traffic among Edge Devices and therefore precludes data-plane learning on the "overlay interface". Data-plane learning continues to happen on the "internal interfaces" to provide compatibility and transparency within the layer-2 sites connecting to the OTV overlay. The Edge Devices appear to each VPN site to be providing L2 switched network connectivity amongst those sites.

This document describes the use of IS-IS as an IGP capable of carrying both MAC unicast and multicast and IP multicast group addresses, thereby serving as both the oURP and oMRP. However, any

other suitable routing protocol can be used as the OTV control protocol. The information carried in IS-IS LSPs will be MAC unicast addresses and multicast addresses with their associated VLAN IDs and IP next hops. The MAC addresses are those of the hosts connecting to the network and the IP next hops are the addresses of the Edge Devices through which these are reachable in the core. Figure 2 shows what the resulting tables would look like in a simple two site example.



At X1

Destination	Interface/NH
H1	E1
H2	Overlay1:IPB

At X2

Destination	Interface/NH
H1	Overlay1:IPA
H2	E1

Figure 2. OTV Forwarding Tables.

Edge Devices will have an IP address reachable through their core facing interface(s), and these nodes join a configured ASM/Bidir multicast group in the core transport network. The core or the provider network relies on a provider Unicast Routing Protocol (pURP) and a provider Multicast Routing Protocol (pMRP) to connect the Edge Devices to one another. It is not strictly required that the Edge Devices participate in the pURP/pMRP. They typically connect as hosts to the core network. This is compatible and consistent with today's interconnection policies. However, the solution also supports the scenario where the Edge Devices do actively participate at Layer-3 in the pURP/pMRP.

The multicast group that the Edge Devices join is referred to as the "Provider Multicast Group (pMG)". The pMG will be used for Edge Devices to become adjacent with each other to exchange their IS-IS Hellos, LSPs and CSNPs. Thus, by virtue of the pMG, all Edge Devices will see each other as if they were directly connected to the same multi-access multicast-capable segment for the purposes of IS-IS peering. The pMG also defines a VPN; thus, when an Edge Device joins

a pMG the site becomes part of a VPN. Multiple pMGs can be defined to define multiple VPNs.

The pMG can also be used to broadcast data traffic to all Edge Devices when necessary. Broadcast transmission will not incur head-end replication overhead. OTV allows the pMRP to efficiently distribute broadcast traffic by the provider ASM/Bidir group.

When forwarding of VPN multicast is required, new multicast state will be used in order to tailor the distribution trees to the optimal group of receivers, these multicast groups are to be created in the provider control plane (pMRP). For instance, each core device will resort to using SSM multicast in the core by having the Edge Device IGMPv3/ MLDv2 join a {source, group} pair.

Edge Devices must combine data-plane learning on their bridged internal interfaces with control-plane learning on their overlay interfaces. The key to this combination is a series of rules through which data-plane events can trigger control-plane advertisements and/or learning events.

OTV supports L2 multi-homing for sites where one or more of the bridge domains may be connected to multiple Edge Devices. It supports both active-backup and active-active multi-homing capabilities to sites. OTV provides loop elimination for multi-homed "sites" and does not require the extension of STP across sites. This means each site can run it own STP rather than have to create one large STP domain across sites.

1.1. Terminology

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.

Site - A Site is a single or multi-homed connected network which is typically under the control of a single organization. Sites are connected together via Edge Devices that operate in an overlay network. The Edge Devices provide layer-2 connectivity among the sites. A site will not be used by IS-IS as a transit network. A layer-2 site is one that is mostly made up of hosts and switches. Routers may exist but the majority of the topology to the Edge Devices are L2 switched. The MAC addresses advertised on the overlay network are all the hosts and routers connected to the L2 devices at the site. The site typically has several VLANs or bridging domains being actively used. A layer-3 site is one that is mostly made up of routers connecting to hosts via switches. The majority of the topology to the Edge Devices are L3 routed.

The number of MAC addresses advertised on the overlay network are limited to the router devices at the site.

VPN - A VPN is a collection of sites which are controlled by a single administration. The addressing plan, router and switch configuration is consistent as it would be if the sites were physically at the same location. There is one overlay network per VPN which connects all sites. Each VPN uses a dedicated ASM/Bidir provider multicast group allocated by the core network, which provides the separation from other VPNs for the control plane, as well as in the data plane.

Edge Device - A modified L2 switch that performs OTV functions. It will run as an L2 device on the site side, but performs L3 functions on the core facing interfaces. When OTV functionality is described, this functionality only occurs in an Edge Device.

Internal Interface - These are Layer-2 interfaces connected to site based switches or site based routers. The internal interface is layer-2 regardless if it connects to a switch or a router.

Overlay Interface - This is a logical multi-access multicast-capable interface. The overlay interface can replicate broadcast and multicast packets efficiently. The overlay interface provides an IP unicast or multicast encapsulation for L2 frames transmitted from the site. The overlay interface is realized by one or more physical core facing interfaces. The core facing interfaces are assigned IP addresses out of the core provider's address space.

MAC Table - This is a forwarding table of 48-bit MAC addresses. The table can contain unicast or multicast MAC addresses. The table is populated by two sources. One being traditional data-plane learning on internal interfaces and the other by the URP/MRP at the control-plane on the overlay interface. A MAC table is scoped by VLAN therefore allowing the same MAC address to be used in different VLANs, and potentially in different VPNs.

Authoritative Edge Device (AED) - This is an Edge Device that forwards Layer-2 frames in and out of a site from and to the overlay interface. Depending on the multi-homing granularity in use, there will be a single AED in the site for a given VLAN or for a given MAC-level flow.

Site-ID - Each Edge Device which resides in an OTV site will advertise over the overlay network the same site-id. The site-id may be determined dynamically or by static configuration.

(VLAN, uMAC) - This is the designation of layer-2 network reachability information as encoded in the URP and as stored in the MAC table. This notation describes a given unicast MAC address within a particular VLAN.

(VLAN, mMAC, mIP) - This is the designation of layer-2 network reachability information as encoded in the MRP and as stored in the MAC/IP table. This notation describes a given multicast MAC/IP address within a particular VLAN. The 'mIP' part of the 3-tuple is provided so both Layer-2 switching and the SSM based tree joins can occur based on the IP group address (since 32-to-1 aliasing can happen for IPv4 group address to MAC mappings and worse for IPv6).

2. Control Plane

This section discusses the control plane hierarchy. At the very base of the hierarchy we find the provider control plane, which enables unicast reachability among the edge boxes and also provides the multicast group that makes edge boxes adjacent from the overlay control plane perspective. The provider control plane also provides the multicast trees in the core that will be used for optimal forwarding of the layer-2 site data traffic.

At the next level, the overlay control plane provides discovery of the Edge Devices that are part of the overlay and conveys client-MAC-address reachability and client-multicast group information between the edge devices.

In general, the control planes are independent of each other. However, in order to optimize multicasting, multicast control-plane events (reports, joins, leaves) that occur in one MRP may initiate events in another MRP so that the optimal tree is always being used to forward traffic. Also, events in the overlay control plane are triggered by forwarding events in the client data plane (however both client and overlay control planes remain independent of each other).

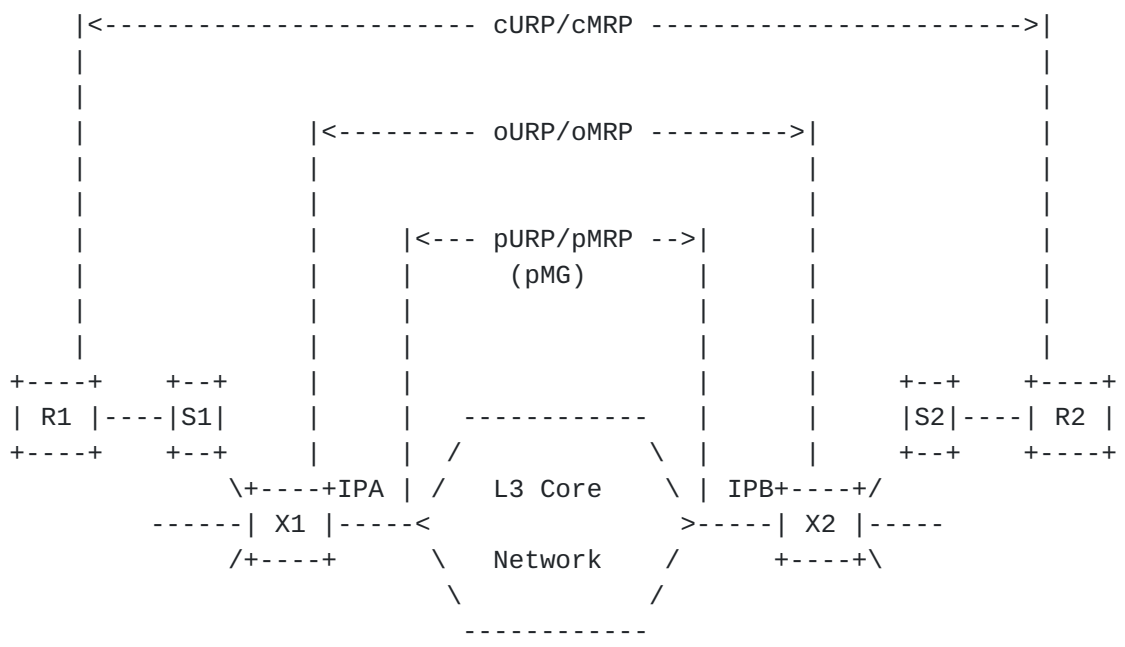


Figure 3. OTV Control Plane Hierarchy

2.1. Provider Control Plane

The provider control plane is the set of routing protocols which run in the core infrastructure to be able to deliver packets sourced from the site networks. There is no required coordination of routing protocols between the site and the core. That is, no more than typically necessary to connect to a core service. In terms of addressing, the Edge Device is allocated an IP address out of the core block of addresses.

For each VPN the Edge Device is to support, a multicast group is required to be allocated from the provider core at a minimum. This multicast group is typically ASM/BiDir. In addition, the multicast state created in the client site network will map to some amount of state in the core network. However, it is not required to provision a unique group for every client data group. The Edge Device takes a client multicast packet and encapsulates it in a core-deliverable multicast packet.

2.2. Overlay Control Plane

The overlay control plane provides auto-discovery of the Edge Devices that are members of an Overlay VPN. It also conveys Layer-2 unicast and multicast reachability information from a site to Edge Devices in other sites and the VLANs or layer-2 bridge domains being extended.

The MAC addresses that are locally connected to an Edge Device are advertised in the overlay URP to other Edge Devices in the VPN. Thus, MAC learning on the overlay is not based on data plane flooding, but is based on explicit advertisements of MAC addresses done by the overlay control plane. Similarly, the multicast groups that a site has receivers or sources for are advertised in the overlay MRP to other Edge Devices in the VPN.

2.2.1. Edge Device Discovery and Adjacency setup

The overlay URP establishes adjacencies only between Edge Devices that are in the same VPN. Edge Devices become part of a VPN when they join a multicast group defined in the core (provider MRP); devices using the same group are members of the same VPN. Thus, the adjacency setup provides a very simple mechanism to automatically discover members of the VPN. The hellos and updates between overlay-URP peers travel over the multicast group defined in the pMRP. Thus, Edge Devices peer with each other as if they were directly connected at layer-2. This peering is possible as all the traffic for the oURP is encapsulated with the pMRP group address and sent into the core. Thus, all Edge Devices in a given VPN receive the oURP multicast traffic as if they were all on the same segment. Similarly, the overlay MRP packets are encapsulated with the pMRP group address corresponding to the VPN. The overlay MRP is used to inform all the Edge Devices that the subscribers to a particular group are reachable over the overlay network.

An Edge Device can support multiple overlay VPNs. Each overlay has its own dedicated provider-multicast group address and a distinct set of adjacencies. There may be multiple overlay adjacencies between the same set of Edge Devices, or the membership may be disjoint for each overlay.

2.2.2. Extended VLANs

Each overlay basically extends a set of VLANs or layer-2 bridge domains among the member sites. On a given Edge Device, a set of VLANs is uniquely extended on a specific overlay. Other VLANs may be extended on other overlays. This entails both advertising and accepting information in the control plane such as VLANs and their associated MAC and group information, as well as forwarding unicast, multicast and broadcast traffic for these VLANs.

To allow scalability of connecting large L2 sites together via the overlay, by default, an Edge Device will not advertise any information for any VLANs. To avoid inadvertent merging of VLANs among sites, Edge Devices will be required to configure the VLANs for which Edge Devices will advertise reachability information for.

2.2.3. Multiple Instances

An Edge Device may support bridging of multiple distinct layer-2 domains with overlapping VLANs which are to be treated as distinct. These VLANs may be extended on the overlay by treating them as separate instances both in advertising control plane information and while forwarding in the data plane. A single overlay VPN can support more than one instance among the Edge Devices in that overlay.

2.2.4. Advertising Unicast MAC Routes

When a MAC address is learned by arrival of a data packet on an internal interface, the Edge Device advertises the MAC address on the overlay URP. In addition to conveying the MAC address reachability to other edge devices, it also provides a mapping to one of the IP addresses of the advertising Edge Device; i.e., the IP next-hop and encapsulation for that MAC address. Typically, even if a site is multi-homed, a unicast MAC address is advertised by a single Edge device, that is the Authoritative Edge Device. Hence, remote Edge Devices will see a single path to reach a given MAC address. However, when active-active multihoming is being used, there will be equal-cost paths to reach a MAC address in a site and the sender Edge Device will load-balance flows among the paths.

2.2.5. Advertising Multicast Routes

An Edge Device learns about the multicast groups that hosts in the site are interested in by snooping IGMP/MLD reports on the internal interfaces. When a multicast MAC or group address is learned, the Edge Device notifies other Edge Devices about it by placing a (VLAN,mMAC,mIP) entry in a multicast control PDU. Thus, the overlay MRP informs all the Edge Devices that the subscribers to a particular group are reachable over the overlay network. This information is used by Edge Devices to populate their multicast oif-list at the source site. As long as there is one site that has a receiver for a multicast group, the Edge Devices at the source site will forward traffic for that group onto the overlay. Edge Devices at the receiving sites will also join the corresponding multicast group in the provider plane (pMRP). Thus, multicast trees are built natively in the core, not on the overlay, and provide optimal delivery of multicast data.

2.2.5.1. Delivery Groups

Delivery groups are multicast groups used in the core network to transport site multicast traffic. Multicast data for various customer data groups are aggregated into a typically smaller set of core multicast trees, without requiring extensive coordination

between OTV edge boxes. Delivery group selection is centralized at each source OTV Edge Device which controls the mapping of a (S,G) to a (DS, DG). It exports this mapping to other Edge Devices so that they can join the (DS, DG) in the core. Link-local site multicast groups may also map to a specific delivery group instead of the provider multicast group used for control packets. Delivery group mapping allows for fair amount of flexibility for the customer sites and the provider to decide control of state versus bandwidth tradeoff in the core.

When a receiver site Edge Device learns a (S, G) to (DS, DG) mapping, it joins the (DS, DG) tree in the core. As an optimization, this join may be done only if there are local receivers for the group. It also installs a layer-3 multicast route for (DS,DG) to decapsulate incoming packets with the appropriate core uplink interface as the RPF interface.

2.2.5.2. Active Source Discovery

An OTV Edge Device will advertise a delivery group mapping for a (*,G) or (S,G) route only when there is an active source sending data in its site. For this, the Edge Device will learn the active sources by snooping multicast data received on the internal interfaces. If a remote receiver interested in this group, a (VLAN, S,G) entry is installed with the overlay as an OIF and the (DS,DG) as outer encapsulation. When IGMP/MLD is being used on the core uplink, the (DS,DG) encapsulated packet may be emitted directly on the uplink interface. The first-hop router on the other end of the core uplink will then forward this packet along the core multicast tree.

2.2.6. Adjacency Server

In case the provider core does not support ASM/Bidir multicast, there is an alternate mechanism to discover the remote Edge Devices which are part of a VPN. In this scenario, an Edge Device is configured as an Adjacency Server. All other Edge Devices inform the Adjacency Server regarding their reachability and capability information via the overlay control plane. Adjacency Server is responsible for informing all the other existing Edge Devices regarding addition or loss of an Edge Device. Based on the reachability information, the Edge Devices can further communicate with one another directly using unicast or multicast data path.

2.3. Connecting an Edge Device to the Overlay

In order to successfully connect to the overlay, the Edge Device has several functions on its different interfaces. These are summarized in this section.

2.3.1. Edge Devices as MAC Routers

The Edge Device need not participate in the provider URP (pURP) as a router, but can simply behave as a host. This keeps its requirements and functionality simple. In this mode, the Edge Device has an IP address which is significant in the core/provider addressing space. The Edge Device joins the multicast groups in the core by issuing IGMPv3/MLDv2 reports, just like a host would. Thus the Edge Device does not have an IGP relationship with the core. This allows for simpler insertion into any type of core network.

However, the Edge Device does participate in the overlay URP and its IP address is used as a router ID and a next-hop address for unicast traffic by the overlay URP. However, the Edge Device does not build an IP routing table with the information received from the oURP, but rather builds a hybrid table where MAC address destinations are reachable via IP next-hop addresses. This may be termed as a MAC router because it can route packets based on MAC addresses.

Thus, Edge Devices are IP hosts in the provider plane, MAC routers in the overlay plane and bridges in the client bridging plane. It should be noted that Edge Devices can also support full IP routing functionality and participate in the pURP/pMRP as routers.

2.3.2. Internal Interface Behavior

The internal interfaces on an Edge Device are bridged interfaces and are indifferent to whether the site itself is L2 or L3. These interfaces behave as regular switch interfaces and learn the source MAC addresses of traffic they receive. Spanning tree BPDUs are received, processed and sourced on internal interfaces as they would on a regular 802.1d, 802.1s and 802.1w switch. IGMP/MLD and data snooping is enabled on internal interfaces to discover local receivers and sources in the site. Additionally, traffic received on internal interfaces may trigger oURP/oMRP advertisements and/or pMRP group joins as described earlier.

Traffic received on an internal interface will be forwarded according to the MAC and multicast tables either onto other internal interfaces (regular bridging) or onto the overlay (OTV forwarding). This is explained in detail in the Forwarding section.

2.3.3. Overlay Interface Behavior

An overlay interface is a logical interface which is associated with an IP address in the provider/core address space. Traffic out of these interfaces is encapsulated with an IP header, and traffic received on these interfaces must be de-capsulated to produce a L2

frame. The encapsulated packets exit the Edge Device on one or more underlying physical or logical L3 interfaces.

STP BPDUs are not sourced from overlay interfaces, therefore there should not be STP BPDUs in the core, nor do the overlay interfaces participate in the spanning tree protocol.

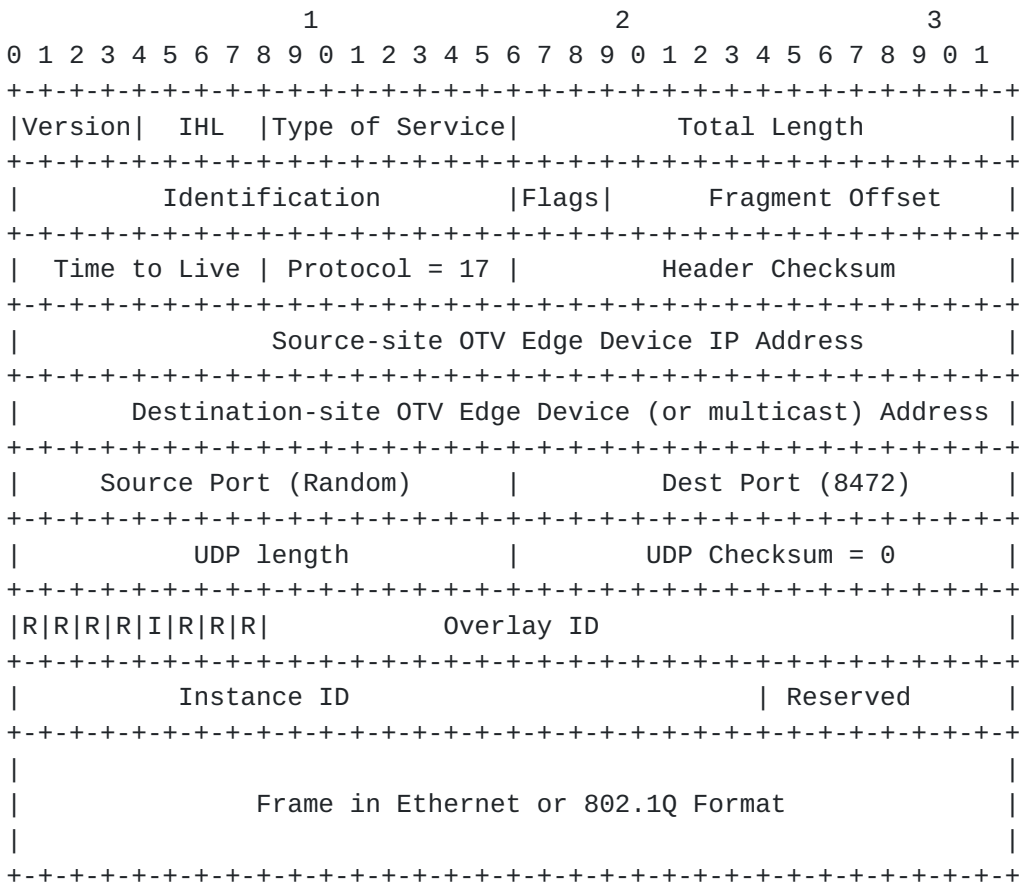
The IP addresses assigned to the overlay interfaces are used as next-hop addresses by the overlay-URP, therefore the MAC table for the overlay interface will include a remote IP address as the next-hop information for remote MAC addresses.

3. Data Plane

3.1. Encapsulation

The overlay encapsulation format is a Layer-2 ethernet frame encapsulated in UDP inside of IPv4 or IPv6.

The format of OTV UDP IPv4 encapsulation is as follows:



IPv4 Header:

Version: Set to value 4 in decimal.

IHL: Set to value 5 in decimal meaning there are no IP options present in an OTV encapsulated packet.

Type of Service: The 802.1P bits from the Ethernet Frame are copied to this field.

Total Length: The total length of the IP datagram in bytes. This includes the IP header, the UDP header, the OTV header, and the L2 frame without the preamble and CRC fields.

Identification: Set randomly by the OTV Edge Device.

Flags: The DF bit should be set to 1.

Time to Live: Set by the OTV Edge Device and is configurable.

Protocol: Since the packet is UDP encapsulated, this field is set to 17 decimal.

Header Checksum: Must be computed by the OTV Edge Device over the IP header fields.

Source Address: The IP address of the OTV Edge Device doing the encapsulation of the L2 frame.

Destination Address: The IP unicast or multicast address set by the OTV Edge Device which is encapsulating the L2 frame. The Edge Device decides when the address is set to a unicast or multicast addresss.

UDP Header:

Source Port: Is chosen by the OTV Edge Device which is encapsulating the L2 frame based on a hash of the L2 frame. This allows packets to be load-split evenly over LAGs on routers in the core, responsible for delivering these IP encapsulated packets.

Destination Port: This is an IANA assigned well-known user port number. Packets encapsulated by an OTV Edge Device put value 8472 in the destination port field.

UDP Length: Is the length in bytes of the UDP header, the OTV header, and the L2 frame without the preamble and CRC fields.

UDP Checksum: This is set to 0 by the OTV Edge Device when doing

encapsulation and ignored by the OTV Edge Device which is decapsulating at the destination site.

OTV Header:

Flags:

'I' - Instance-ID bit. When set to 1, it indicates the Instance ID should be used in the forwarding lookup.

'R' - Reserved bits.

Overlay ID: Is used only for control plane packets such as the URP/ MRP (IS-IS) to identify packets for a specific overlay.

Instance ID: Set by the OTV Edge Device doing the encapsulation to specify a logical table that should be used for lookup by the OTV Edge Device at the destination site.

L2 Ethernet Frame:

The L2 Frame minus the preamble and CRC received on an internal link by an OTV Edge Device.

The addition of OTV encapsulation headers increases the size of an L2 packet received on an internal interface such that the core uplinks on the Edge Device as well as the routers in the core need to support an appropriately larger MTU. OTV encapsulated packets must not get fragmented as they traverse the core, and hence the IP header is marked to not fragment by the Edge Device. The Edge Device drops packets that exceed the core uplink MTU.

The following tables enumerates how MAC level packets are encapsulated in the OTV header.

MAC-level Frame	OTV IP Encapsulation
Unicast Frame	IP unicast packet
Broadcast Frame	ASM/Bidir IP multicast packet
Link-local Multicast Frame	ASM/Bidir IP multicast packet
Data Multicast Frame	SSM IP multicast packet

3.2. Forwarding Process

Most of the interesting forwarding cases happen when a packet comes from the Overlay Link to be forwarded to an Internal Link, or vice versa. But for completeness, forwarding between internal links is also described

3.2.1. Forwarding between Internal Links

When an Edge Device has internal links, it operates like a traditional L2 switch. That is, it will send unicast packets on a port where the MAC was learned; it will send multicast packets on the ports it has IGMP/MLD-snooped; and it will send broadcast packets out all ports for a given VLAN or layer-2 bridge domain

3.2.2. Forwarding from an Internal Link to the Overlay

An Edge Device will decide to forward a Layer-2 unicast, multicast, or broadcast packet over the overlay interface when the overlay control plane has put the logical port of the overlay interface in the forwarding table, such as for the corresponding unicast or multicast address. When a packet is sent over the overlay interface, it is first prepended with an OTV header that includes the IP address of the overlay next-hop. The packet as received from the internal interface is not touched other than to remove the preamble and FCS from the frame. The IP address, outer MAC address and other encapsulation information are all installed in the forwarding hardware by the control plane so the OTV header can be prepended and the packet forwarded at high rate.

The Edge Device has to be eligible to forward this packet as per the control plane, such as being the Authoritative Edge Device. Multi-homing of sites imposes additional rules on the forwarding of traffic as described later in this document.

3.2.3. Forwarding from the Overlay to an Internal Link

When a packet is received on the overlay interface, it will need to be IP decapsulated to reveal the inner MAC header for forwarding. The inner MAC header SA and DA addresses and VLAN-ID will used for forwarding actions. For any type of packet received on the overlay interface, it will be accepted only if the Edge Device is the Authoritative Edge Device as determined by an inspection of the received packet header.

When a unicast packet is received on the overlay interface, the outer OTV IP header is removed, and the VLAN-ID and the MAC DA from the inner header is used to do the MAC table lookup. Here onwards, this is a regular bridging operation, whether the MAC address entry is present or not.

When a multicast packet is received on the overlay interface, the outer OTV IP header is removed. The VLAN-ID and inner MAC header SA and DA or inner IP header SA and DA are used to do a Layer-2 multicast table lookup and forward the packet on the right internal

interfaces. A multicast packet received from the overlay will not be sent back out on the overlay.

When a broadcast packet is received on the overlay interface, the outer OTV IP header is removed and the packet is then flooded on all internal interfaces.

3.2.4. Unicast Packet Flows

Hosts typically generate ARP requests and learn the MAC addresses of other hosts from ARP requests and replies. Switches learn the source MACs from packet headers and store this state to optimally forward traffic destined to these MACs. The OTV Edge Devices will also learn the MACs locally on their site facing interfaces, and will install remote MACs received over the overlay control plane into the local MAC table with the appropriate remote Edge Devices as next-hops.

Once these actions take place, every switch will forward the L2 packet based on the MAC table entry. The OTV Edge Device at the source site will also do a MAC table lookup which will yield a next-hop entry pointing to a remote Edge Device. Once the OTV header with the IP address is prepended, the packet is then forwarded to the destination Edge Device at Layer-3 as a regular IP packet.

The Edge Device as well as the core routers may load-balance these encapsulated packets among equal-cost multiple Layer-3 paths, with packets belonging to a single Layer-2 flow being hashed to a specific equal-cost path.

3.2.5. Unknown Unicast Packet Handling

When the switched network at an OTV site has no state for a MAC address, it will flood the unicast packet on the spanning tree throughout the site. The Edge Devices are on the spanning tree (like any other switch at the site) so they will receive these unknown unicast packets.

It is imperative that the Edge Devices hold previously learned MAC addresses for an extended period of time so that remote Edge Devices can get reachability to these local MACs. So the cache timers will be longer than the traditional MAC aging timers on switches. In fact, the Edge Device MAC aging timers generally need to be greater than the ARP request interval from any host. Either an unknown flood or a broadcast packet could cause an update of the MAC entries in the Edge Device. And when MACs go inactive, an Authoritative Edge Device must withdraw the MAC address from the overlay control plane. Traffic to these unknown destinations will not be forwarded onto the overlay. Thus, OTV does not flood unknown unicasts. In an OTV

network unknown destinations become known the moment the host emits at least one packet. The assumption is that no host on the network is completely silent.

3.2.6. Multicast Packet Flows

A multicast receiver host sends out IGMP/MLD reports for the multicast groups it wants to join. The sites may use either IGMPv2 or IGMPv3. A multicast capable switch will forward these reports to router ports and querier ports. The OTV Edge Device behaves as either a querier or a router in the network and hence receives these reports.

A host in a site may be a source for an (S,G) group and sends data. This data is flooded or forwarded along IGMP/MLD snooped links by the site switches. When an Edge Device receives this packet, it does a Layer-2 multicast table lookup which may yield several OIFs. If the overlay interface is part of the OIF-list, then the Edge Device encapsulates the packet in an OTV IP header which includes the delivery group (DS, DG) IP addresses. It then emits the resulting IP multicast packet into the core which is forwarded along a core multicast tree to the receiver site edge devices.

The receiver site Edge Device also joins one or more (DS, DG) core multicast trees as directed by various source site Edge Devices. This allows it to receive data from other sites. The core multicast trees may either be SSM or ASM though this document focusses on the SSM case.

3.2.7. Broadcast Packet Flows

A broadcast packet originated at an OTV site needs to be delivered to all sites of the same VPN. This is typically done with the ASM/Bidir group encapsulation which is the same group used for the oURP/oMRP (pMG). A different data group can also be used to forward broadcast traffic.

A broadcast packet, sourced in a site, gets to all Edge Devices because each Edge Device is on the site spanning tree. However, duplicates must not be allowed to appear on the overlay network when there are multiple Edge Devices, so the Authoritative Edge Device for the VLAN is the only Edge Device that forwards the packet on the overlay network. All edge devices at a remote site will receive the broadcast packet over the core multicast group. To prevent duplicates going into the site, only the Authoritative Edge Device in that site will forward the packet into the site. And once sent into the site, the packet gets to all switches on the site spanning tree. Because only the AED can forward broadcast packets in or out of the

site, broadcast loops are avoided.

Other types of packets such as link-local multicast packets and non-IP Layer-2 packets may also be sent along the pMG or on a dedicated data group.

3.3. STP BPDU Handling

Since the Edge Device acts as an L2 switch it does participate in the Spanning Tree Protocol if the site has been configured to use it. However, there is no STP activity on the overlay interface. The following are the rules an OTV Edge Device will follow:

o When STP is configured at a site, an Edge Device will send and receive BPDUs on internal interfaces. An OTV Edge Device will not originate or forward BPDUs on the overlay network.

o An OTV Edge Device can become a root of one or more spanning trees.

o An OTV Edge Device will take the typical action when receiving Topology Change Notification (TCNs) messages.

o When on OTV Edge Device detects another Edge Device in it's site has come up or gone down, it may send a TCN so it can gather new state for when its authoritative status changes for a VLAN.

To allow the L2 switch network to scale to larger number of nodes and MAC addresses, it is considered a feature of OTV to maintain and keep the spanning trees small and per site.

4. MAC Address Mobility

In a traditional layer-2 switched network, mobility of a host is easily achievable because each switch in the network tracks the source MAC address in each packet and the interface the last packet was received on. So if that MAC is later seen on another interface, the new interface can be updated at the same time the packet is forwarded. These fast MAC moves need to be achieved when a MAC moves from one OTV site to another. The Authoritative Edge Device for a VLAN determines a MAC move in combination with traditional learning on the internal interfaces and explicit MAC advertisements on the overlay.

If an Authoritative Edge Device has a MAC address stored in the MAC forwarding table which points to the overlay interface, it means that an Edge Device in another site has explicitly advertised the MAC as being local to it's site. Therefore, any packets coming from the MAC

will be coming from the overlay. Once that MAC is heard on an internal interface, it has moved into the site. Since it has moved into a new site, the Authoritative Edge Device in the new site is responsible for advertising it.

When a MAC appears in a new site, the Authoritative Edge Device will advertise the new MAC address with a metric value of 0. When the Edge Device in the site the MAC has moved from hears the advertisement, it will withdraw the MAC address that it had previously advertised. Once the MAC address is withdrawn, the Edge Device where the MAC has moved to will change the metric value to 1. All remote sites sending to this MAC address will start using the new Edge Device as soon as they hear it's MAC advertisement with metric 0.

5. Multi-homing

A site typically will be multi-homed with multiple Edge Devices connecting to the overlay. This provides the site with increased network redundancy and resilience to failures.

When sites are multi-homed, there is a potential for loops to be created between the OTV overlay and the layer-2 domains at different sites. One option to address such loops is to transport STP BPDUs on the overlay and rely on STP to break any loops that may form when multi-homed sites connect to the overlay. However, this is not desirable as it leads to very large or complex STP domains. OTV multi-homing avoids loops through a combination of techniques in the control plane and data plane.

OTV does not transport STP BPDUs over the core. As a result, each site will have its own STP domain, which is separate and independent from the STP domains in other sites, even though all sites will be part of a common broadcast or Layer-2 domain. It also does not flood unknown unicast traffic on the overlay.

5.1. Authoritative Edge Device Selection

An Authoritative Edge Device is an Edge Device that forwards Layer-2 frames in and out of a site from and to the overlay network. When a site is multi-homed to the overlay, a proper Authoritative Edge Device selection ensures that traffic crossing the site-overlay boundary does not get duplicated, create loops or cause any churn in the MAC tables of switches within the local and remote sites.

The Authoritative Edge Device (AED) may be statically assigned or determined via an election among the devices in the same site. A

unique AED may be selected for each VLAN or it may be on a finer MAC-level granularity. In either case, for a given MAC-level flow, the data path will be symmetric.

An Authoritative Edge Device has the primary responsibility to advertise locally learned source MAC addresses and IGMP/MLD-snooped multicast addresses in the oURP and oMRP.

When done per-VLAN, an AED will be authoritative for all unicast and multicast addresses within a single VLAN. The authoritative responsibility can be shared with other Edge Devices for other VLANs so traffic can be load balanced among all Edge Devices across different VLANs.

For the particular scenario of all-active multi-homing and load balancing, AEDs may be elected on a finer granularity. Thus there may be several AEDs in any given VLAN in this case and different flows can use different Edge Devices.

Protocol adjacencies are set up among the Edge Devices in the same site. The AED is selected from this list of Edge Devices in the same site. The AED selection algorithm tries to ensures an even spread of VLANs across the Edge Devices. A simple mechanism may be via a hash of the VLAN-ID. Alternatively, a static AED assignment may be to use a VLAN range division among all Edge Devices in the site. The local VLAN/AED specific information may be advertised to other Edge Devices.

Each Edge Device keeps track of the other Edge Devices in the same site. If an Edge Device has a failure such that it is incapable of forwarding traffic for its authorized VLANs, other Edge Devices in the same site will detect or be notified of this event and run the AED selection procedure to reassign authority for the failed device's VLANs.

5.2. Site Identifier

All Edge Devices that belong to a single Layer-2 site will advertise a Site-ID on the overlay control plane. This information is used by remote Edge Devices to identify the members of the same site. The Site-ID influences the AED election and path selection from remote Edge Devices to the local site. The Site-ID may be statically assigned or dynamically computed by the devices in the same site.

6. IS-IS as an Overlay Control Protocol

This section describes the use of the IS-IS protocol to serve as the

Overlay URP and MRP. The details of the IS-IS PDUs and TLVs defined for OTV are described in [IS-IS-OTV].

It is highly desired to leverage the native and existing IS-IS protocol functionality where feasible. There are some protocol extensions specific to OTV which are described in this document.

The overlay network serves as a logical multi-access Ethernet LAN connecting the various Edge Devices. Hence, IS-IS hellos and LSPs can be exchanged directly over the overlay network similar to IS-IS operation on a LAN. These IS-IS packets are encapsulated in the OTV IP multicast header and reach other Edge Devices on the core multicast tree. In addition, OTV IS-IS packets use a distinct Layer-2 multicast destination address. Therefore, OTV IS-IS packets do not conflict with IS-IS packets used for other technologies even if they may be sent over the same links in the core or arrive at an Edge Device on the same core uplink interfaces.

IS-IS packets belonging to different overlay VPNs are mutually isolated and distinguished by the OTV control packet header and the use of distinct multicast groups in the core. Standard IS-IS authentication mechanisms may additionally be used to provide further isolation and authentication of VPN membership.

OTV IS-IS employs IS-IS LAN procedures on the overlay network. It forms IS-IS adjacencies with all other Edge Devices in the overlay and elects a Designated Router (DIS). The IS-IS system ID uniquely identifies an Edge Device in the IS-IS control plane.

IS-IS IIHs are sent and received on the overlay by all Edge Devices. The IP addresses assigned to the overlay on an Edge Device is advertised in the IIHs and provides the IP reachability information to the edge device through the core.

CSNPs are sent on the overlay by the DIS and used to achieve reliable delivery of the link state database. This link state database holds LSPs that describe the Edge Device connectivity to the pseudo-node (or the multi-access overlay network). The LSPs also hold the unicast MAC information that is advertised by a site Edge Device. CSNPs are also used to reliably deliver the Group Membership link state database that holds LSPs describing the multicast MAC group addresses. OTV IS-IS only maintains the Level-1 link state database.

Unicast MAC address information is carried in LSPs in the MAC-Reachability (MAC-RI) TLV defined in [IS-IS-Layer2]. All MAC addresses are typically advertised with a metric of 1. When using the MAC move procedures, the metric will be set to 0. Definition of the fields used by OTV is specified in [IS-IS-OTV].

Multicast related information is carried in LSPs in several different TLVs specified in [IS-IS-OTV]. The multicast groups that a site has receivers for are carried in the sub-TLVs of the Group Address TLV. Multicast sources discovered in a site are advertised in a Group Membership Active Source TLV. This TLV includes the list of groups for which the source is sending data along with the core Delivery Groups to which the advertising Edge Device will map the site data groups.

When an Adjacency Server is being used, all Edge Devices inform the Adjacency Server regarding their reachability and capability information by including in their hellos the Adjacency Server TLV. The Adjacency Server includes a list of all the Edge Devices it has heard from, and their capabilities, in its hello PDUs.

The Site-ID information is contained in the Site Identifier TLV and sent in IS-IS IIHs.

7. Acknowledgements

The authors would like to thank many for their careful review. They include Venu Nair, Victor Moreno, Ashok Chippa, Sameer Merchant, Tony Speakman, Raghava Sivaramu, Nataraj Batchu, Sreenivas Duvvuri, Gaurav Badoni, Veena Raghavan, Marc Woolward and Tim Stevenson.

Many have received individual presentations of OTV and provided critical feedback early in the design process. These reviewers include Vince Fuller, Peter Lothberg, Dorian Kim, Peter Schoenmaker, Mark Berly, Scott Kirby, Dana Blair, Tom Edsall, Dinesh Dutt, Parantap Lahiri, and Jeff Jensen.

8. Security Considerations

The specifications in this document do not add any new security issues to Layer-2 bridging technologies. Existing security mechanisms may be used both in the control plane and in data forwarding to achieve any security requirements.

This document specifies the use of IS-IS as a control protocol for OTV. It adds no additional security risks to IS-IS, nor does it provide any additional security for IS-IS.

9. IANA Considerations

There are new IS-IS PDUs and TLVs being proposed for OTV, and are

defined in [IS-IS-OTV].

10. Normative References

[IS-IS] ISO/IEC 10589, "Intermediate System to Intermediate System Intra-Domain Routing Exchange Protocol for use in Conjunction with the Protocol for Providing the Connectionless-mode Network Service (ISO 8473)", 2005.

[IS-IS-Layer-2] Banerjee, A., "Extensions to IS-IS for Layer-2 Systems", 2010.

[IS-IS-OTV] Rao, D., "IS-IS Extensions to support OTV", 2010.

Authors' Addresses

Hasmit Grover
Cisco Systems
170 W Tasman Drive
San Jose, CA 95138
US

Email: hasmit@cisco.com

Dhananjaya Rao
Cisco Systems
170 W Tasman Drive
San Jose, CA 95138
US

Email: dhrao@cisco.com

Dino Farinacci
Cisco Systems
170 W Tasman Drive
San Jose, CA 95138
US

Email: dino@cisco.com