Network Virtualization Using Extreme Fabric Connect

Concept and Design

Abstract: This document serves as a design reference for implementing a virtualized fabric network architecture based on Shortest Path Bridging (SPB) named Fabric Connect. It is intended for network architects and engineering staff responsible of the technical design and implementation of the network infrastructure.

This document draws parallels with MPLS and EVPN VXLAN based architectures. MPLS has traditionally been the only way to achieve the same level of network virtualization that SPB offers while EVPN VXLAN offers equivalent functionality in data center deployments.

Published: August 2019

Extreme Networks, Inc.
Phone / +1 408.579.2800
Toll-free / +1 888.257.3000
www.extremenetworks.com

Document # NN48500-649 Rev. 03

Table of Contents

Table of Figures

Table of Tables

Conventions

This section describes the text and command conventions used in this document.

Tip

Highlights a solution benefit.

Note

Highlights important information.

Caution

Highlights important factors that need to be accounted for in solution designs.

Introduction

Over the years, campus and data center networks have had to dramatically evolve to keep up with new trends in the industry. Multi-tenancy requirements, once a prerogative of carrier networks, are becoming more common in large enterprises, particularly in outsourcing environments, while privately owned data centers increasingly need to handle network virtualization. The Internet of Things (IoT) is pervasively becoming a reality, which increases the need for network virtualization at the edge and raises security concerns due to the increased attack surface. Video surveillance is becoming an essential component for property management and physical security in every industry sector, while video telephony and video streaming are becoming ubiquitous. Efficient support of IP multicast is key in delivering both.

Software Defined Networking (SDN) is much-touted as the holy grail for solving network manageability challenges that every organization should aspire to. However, SDN has come to mean different things to different people and translates into a multitude of product offerings and features from analysts and network suppliers.

The phrase “network fabric” has also become an over-hyped buzz word with the expectation that a fabric makes the network behave as a logical entity capable of configuring itself without requiring endless tuning and maintenance by experts.

Historically, virtualization and multi-tenancy have long been requirements in carrier networks. The MLPS networking technology, which relies on a complex layer of protocols, was developed to meet these requirements. Understandably the industry at large has over the years been busy re-packaging and elaborating those complex networking technologies for the enterprise market. Hence it is not unusual to see MPLS positioned for large enterprise backbones and Ethernet VPN (EVPN) solutions claiming to be “IP fabrics” increasingly for enterprise-owned large data centers. Ironically, for many of those vendors, the terms “fabric” and SDN are being used to disguise overlays and automation of the underlying complexity of the resulting heavily loaded protocol stacks via provisioning scripts, which ultimately shortchange the original goals of SDN.

Yet the enterprise network market has needs and requirements that were never a challenge for carrier networks. IoT is driving the need for networks that can be hyper-segmented and elastically stretch and secure IoT devices. While video surveillance and some data applications require IP multicast, which is scalable and can be separated into virtualized domains, neither of which can be easily done with MPLS- and EVPN-based solutions.

Also, carrier networks are ultimately revenue-generating assets that can justify the higher levels of complexity and can be staffed by skilled personnel accordingly. The same is not true in the enterprise space where the network is viewed as an operational cost that enables the enterprise to achieve its business goals.

At Extreme Networks, we have a focus on the needs of enterprise customers - a network solution that is:

- Simple to implement
- Easy to manage and troubleshoot
- Natively secure
- Robust and resilient
- Able to adapt to changes without downtime
- Standards-based
- Able to interact with and migrate from traditional protocols
- Efficient for any L2, L3, or IP multicast based applications
- Dynamic - users and applications can be automatically and elastically assigned to service enabled virtualized networks
- Cost-effective by using a single networking infrastructure

It is for these reasons that Extreme has developed Fabric Connect (FC), which is fundamentally a new networking foundation based on Ethernet. FC utilizes a single control plane that combines essential multicast capabilities and state-of-the-art link state routing with virtualization capabilities to match and exceed those of MPLS- or EVPN-based solutions while significantly reducing complexity.

This document provides a detailed insight into Extreme Fabric Connect and explains how it can meet enterprise network requirements. Because Fabric Connect is fundamentally different from other traditional network architectures, the Reference Architecture section provides comparisons between Fabric Connect and traditional architectures, including benefits and disadvantages. Readers who are familiar with these traditional technologies may find the comparisons helpful when trying to apply their experience to Fabric Connect principles.

The Guiding Principles section introduces at a high-level how the various Fabric Connect components come together to address campus, branch office, data center, and automated edge deployments. The Architecture Components section provides details on the fundamental pieces of Fabric Connect. Several major sections are then dedicated for greater depth into the relevant FC functionalities. The consolidated design overview section then cuts again across the fabric architecture to cement what should now be fully understood by the reader.

Throughout the document, insights are provided into the relevant Extreme networking platforms used to build Fabric Connect as well as warning on platform limitations and correct positioning. However, it is not the intent of this document to provide an all-inclusive design guide. The Extreme product portfolio experiences constant innovation and improvement, and current product specification design guidelines should be referenced when designing Fabric Connect solutions.

Reference Architecture

The goal of network virtualization is to decouple the physical infrastructure from the network services used to interconnect distinct user communities and their applications. Users and devices connected to the network will only see the virtual network to which they belong and are allowed to communicate only with other devices in the same virtual network. User communities can be kept separate from one another and made to access only the applications that they need, increasing the security of the network. Each virtual network holds the addressing (IP routes, MAC addresses), QoS parameters, and security and access policies that pertain to that user community, increasing the scalability of the network.

Apart from these obvious benefits, true network virtualization brings about greater benefits in terms of agility in adapting the network to new applications, new users, and business needs. Much like server virtualization has brought about a transformation in the way applications are managed and deployed in the data center, a virtualized network infrastructure fundamentally changes the way networks are managed, providing the ability to dynamically create, modify, or remove services without affecting other services or requiring maintenance windows.

Extreme's Fabric Connect, which offers virtualization capabilities for L2, L3, and multicast, is based on the Shortest Path Bridging (SPB[1]) protocol, delivering the above-mentioned benefits. It provides a scalable architecture over a dramatically simplified protocol stack (as compared to MPLS or EVPN), which in turn results in efficiency gains in network design, operation, and maintenance.

An SPB network uses a single instance of Intermediate System to Intermediate System (IS-IS) routing protocol whereas MPLS and EVPN require and depend on multiple protocols (OSPF, BGP, LDP, PIM, etc.). As only one protocol is used in the core all service types benefit from the same fast resiliency without any protocol dependencies.

Shortest Path Bridging Fabric

Shortest Path Bridging (SPB) is rapidly becoming one of the leading network technologies to deliver an Ethernet based fabric where all networking services, whether IPv4, IPv6, IP Multicast and/or simply L2 VLANs can be decoupled from the physical infrastructure and virtualized to meet the needs and demands of typical mid to large enterprises.

SPB was originally defined for carrier Ethernet networks to complement and extend carrier MPLS backbones. The attributes that made SPB valuable for carrier network providers provide a solid foundation for virtualized enterprise networks based on a dramatically simplified architecture.

SPB is the combination of three IEEE standards (Table 1) that deliver a new paradigm to the way in which Ethernet-based networks can operate. SPBM uses Mac-in-Mac encapsulation with the edge device source and destination IP/MAC addressing encapsulated by the backbone MAC addresses of the nodes on the edge of the fabric servicing the edge devices. This provides a simple forwarding function because the core only needs to know the shortest path to the target fabric edge node. The MAC address of the edge node (switch) is referred to as the Backbone MAC (BMAC), which is used for reachability to other SPBM nodes using IS-IS as the Interior Gateway Protocol (IGP). A node in the core has a very simple job to do as it only has to look at the backbone MAC address and forward the packet. The Mac-in-Mac encapsulation mechanism has no IP addressing required in the core, providing stealth for the network infrastructure.

The IEEE802.1ah Mac-in-Mac encapsulation used by SPBM brings an addressing hierarchy to Ethernet where the network addressing of end-stations and user devices (whether at L2 with MAC addresses or at L3 with IPv4 or IPv6 addressing) are always seen as being reachable via a fabric node using the node's MAC address (BMAC). From any given source node, the destination BMAC uniquely defines a cut-through forwarding path without any label swapping (MPLS) or hop-by-hop IP routing.

[1] IEEE 802.1aq/IETF RFC 6329

The SPB (IEEE802.1aq) standard leverages this addressing hierarchy, replacing the Spanning Tree protocols, and bringing to Ethernet-based networks the well-known and much appreciated IS-IS Link State Routing protocol. IS-IS computes the shortest paths between all BMACs within the Ethernet fabric. As such, an SPB fabric behaves with similar properties that network administrators expect to see from a traditional IP-based network using either OSPF or IS-IS as the IGP.

Tip

SPB is an Ethernet based fabric and is sometimes referred to as an L2 fabric. It is important to understand that L2-specific flooding and learning mechanisms, which naturally have a negative connotation for L2 networks, are not used in the SPB fabric infrastructure. Like MPLS, SPB uses a non-IP header to route packets through the fabric between fabric nodes. Like a traditional IP routed network, it is using an IGP routing protocol (IS-IS) to calculate shortest paths for which forwarding entries (FDB) are then programmed into the data plane.

Table 1 – SPB IEEE Relevant Standards

Original IEEE Standard Amendment	Year Published	Incorporated into IEEE Standard	Standard Name
802.1ag	2007	802.1Q-2011	Connectivity Fault Management (CFM)
802.1ah	2008	802.1Q-2011	Provider Backbone Bridging (Mac-in-Mac)
802.1aq	2012	802.1Q-2014	Shortest Path Bridging (SPBM & SPBV)*

* SPBM uses Mac-in-Mac Ethernet encapsulation and SPBV uses 802.1ad Q-in-Q encapsulation.

Note

Extreme Networks only implements SPBM in its current products. SPBV was implemented in Extreme EOS S/K series legacy products.

The technology discussed in this design document makes use of SPBM, which leverages 802.1ah Mac-in-Mac encapsulation. Throughout this document, any reference to Shortest Path Bridging or SPB is always referring to SPBM.

Perhaps the most important innovation that IEEE802.1ah introduces to Ethernet is the addition of a new 24-bit Service-ID (I-SID) field in the Mac-in-Mac encapsulation. The I-SID is a fabric-wide global service identifier, in contrast to the VLAN identifier, which in SPBM becomes only locally significant to the node or access port. This brings the ability to virtualize and transport any of the network service types that previously only MPLS-based backbones were capable of supporting, directly over an Ethernet-based network.

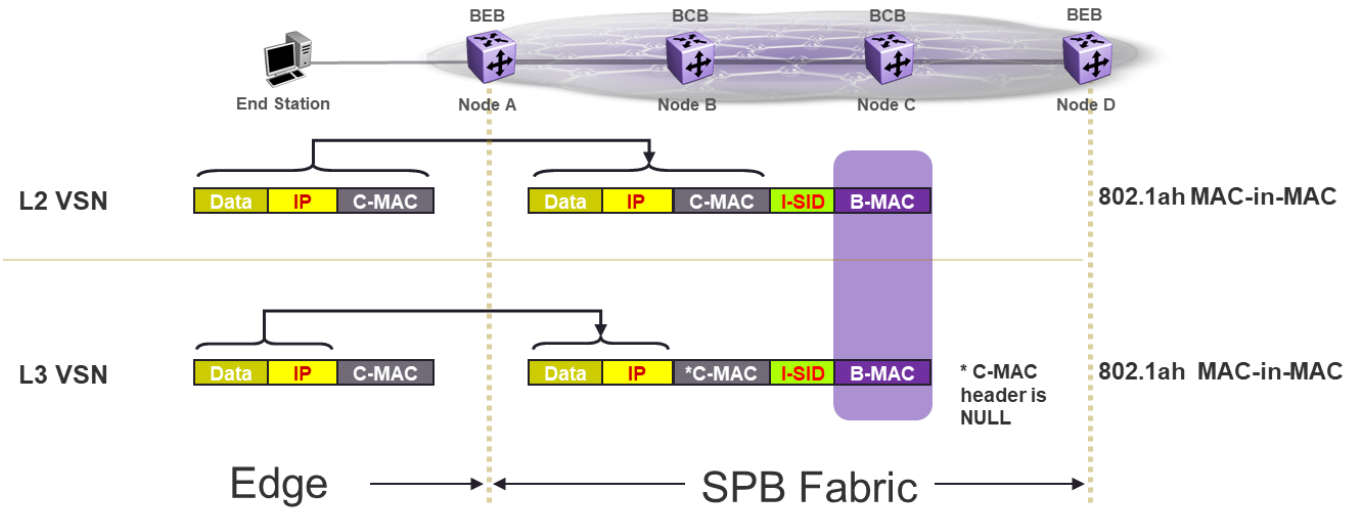


Figure 1 SPBM's Mac-in-Mac Encapsulation

Note

As a comparison, MPLS always pushes two or more labels onto an Ethernet (or other L2 technology) packet. But not all MPLS labels are the same. The outer-most label is a packet forwarding label that is used for label switching the packet across the MPLS backbone. (In an SPB-based Fabric, this role is undertaken by the BVLAN id + BMAC destination address.)

The inner MPLS labels are purely used as VPN IDs; that is, once the packet has reached its destination across the MPLS backbone, the inner label will determine whether the payload is handed off to one or another VRF/VPLS instance. (In an SPB-based Fabric this function is taken over by the I-SID.)

The ability of IS-IS to compute shortest path trees at the Ethernet layer is also capable of producing service-specific shortest-path multicast trees for use with L2 service types as well as for IP Multicast streams. L2 service types need to transport broadcast, (non-snooped) multicast, unknown unicast packets efficiently over the SPB backbone for delivery at every end-point in the same service. IP Multicast streams need to be replicated across the fabric only where IGMP receivers exist. For the sake of comparison, the ability to natively support multicast trees simply does not exist with IP. With IP, dealing with IP Multicast is complex because it was implemented as an afterthought requiring additional protocols such as PIM-SM which can be completely eliminated in an SPB-based Ethernet Fabric.

An SPB Fabric consists of two types of nodes, Backbone Edge Bridge (BEB) and Backbone Core Bridge (BCB). BEBs are generally deployed at the edge of the fabric. BEBs terminate fabric services (L2, L3, multicast) and provide interfaces to networked devices or other non-fabric network services. BCBs are generally deployed in the center of the network. BCBs only perform a transport function along the shortest path towards the destination BMAC and have no knowledge of the transported service types. Only BEB nodes add and remove Mac-in-Mac encapsulation to traffic as it enters/leaves the service it belongs to.

This is fundamentally different from traditional enterprise networks but draws parallels with MPLS-based architectures where service configuration is done solely at the edge of the network without any scaling impact or need to "touch" the core of the network.

Tip

MPLS networks have a similar concept of Provider Edge (PE) nodes and Provider (P) nodes.

The third standard is IEEE802.1ag, which is the new foundation for Operations, Administration & Management (OAM) over Ethernet-based networks for Connectivity and Fault Management (CFM). Defined by carriers for use on carrier-grade networks (including MPLS-based ones), this standard brings to Ethernet and SPB a far more sophisticated troubleshooting toolkit than Enterprise customers are used to with IP. Ethernet-based CFM can test for the most basic connectivity tests (ping) to path tracing (traceroute) over both unicast and service-specific multicast trees (tracetree & tracemroute), as well as network performance monitoring (latency/jitter measurements) via Continuity Check Message (CCM) and Y.1731 extensions.

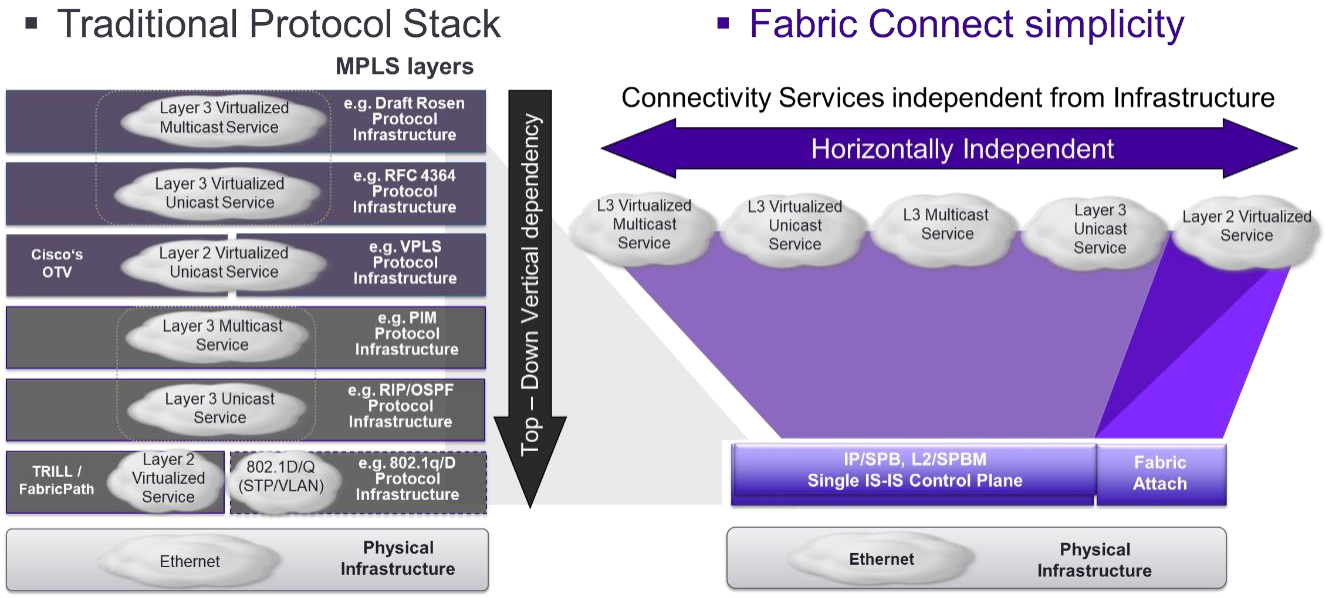


Figure 2 Comparison of SPB's Simplicity with Traditional Protocol Stack

The benefits of delivering MPLS service types over an SPB-based Fabric are many. To start, MPLS is complex and relies on a multitude of control plane protocols each with its own complexities and protocol layer dependencies. Being able to deliver an Ethernet-based Fabric with a single control plane (IS-IS) supporting all of the same service types as MPLS/EVPN but without needing to engineer the backbone with OSPF, Multiprotocol BGP, BGP Route Reflectors, LDP, and PIM-SM makes for an easier life for enterprise network administrators, both in terms of design and maintenance, but inevitably in terms of cost.

Layer 3 Virtualization Overview

L3 Virtualization is designed to support the concept of multi-tenancy. Enterprise networks can separate portions of the network into separately addressed "communities of interest," or logical segments. This improves enterprise security by isolating unrestricted communications to only those members of the specific logical segment. Inter-segment communications may be controlled by using stateful firewalls with interfaces residing in each of the network segments, or through carefully applied inter-segment route leakage policies, or a combination of both.

The deployment of virtualized L3 routing domains over an SPB Ethernet Fabric is achieved in a manner that bears many similarities with the MPLS-based IPVPNs. Virtualized L3 routing domains in Fabric Connect are called Layer 3 Virtual Service Networks (L3 VSNs). Every L3 routing domain is terminated on a Virtual Router and Forwarding (VRF) entity on a BEB, usually located on the distribution layer nodes. These nodes exchange I-SID-IPv4 (or I-SID-IPv6) routes via IS-IS well defined Type Length Values (TLVs). These TLVs exist in the IS-IS Link State Data Base (LSDB) of the SPB Fabric, but will only be inspected by nodes where the same L3 VSN I-SID is terminated.

Core nodes (BCBs) take no notice of these TLVs and simply forward packets based on the shortest path towards the destination BMAC. The BEB distribution nodes participating in the L3 VSN service have directly connected interfaces for access layer subnets that can be extended to edge switches or Fabric Attached switches. If attached to a Fabric Attached switch, the corresponding L2 VLAN can be derived either via

configuration or identity-based networking authentication. These VLANs are associated with a VRF on the BEB nodes where a given VLAN can belong to one and only one VRF. Multiple VRFs can be configured, each belonging to a different Fabric wide service identifier (I-SID), thus providing multi-tenancy in L3 VSN.

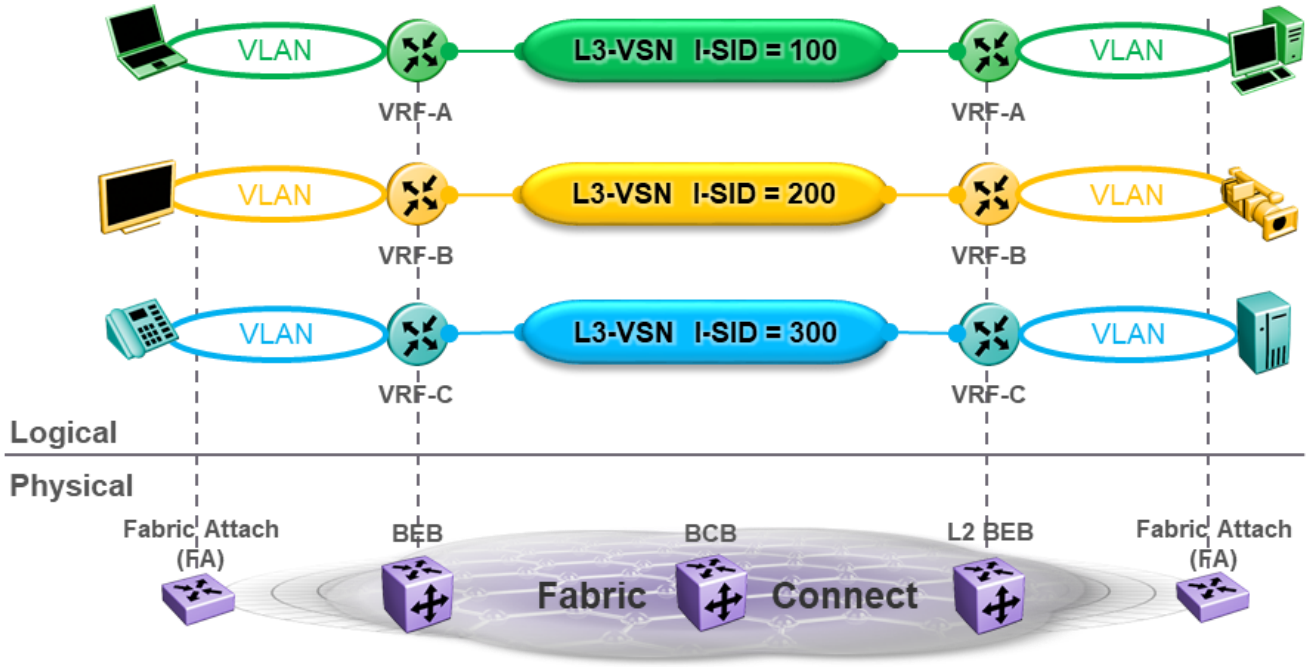


Figure 3 Virtualization with SPB L3 VSNs

Tip

Benefits of SPB L3 VSNs over MPLS-VPNs are:

- Simple service definition via Service Identifier (I-SID) configuration on end-point VRF instead of having to define multiple complex import and export BGP Route Targets and BGP Route Descriptors.
- The same I-SID is also used in the Mac-in-Mac packet encapsulation, whereas with MPLS-VPNs, the inner MPLS labels (used as VPN-id) only have an indirect correlation to BGP Route Target configuration.
- No need for any BGP (and therefore no need for BGP Route Reflectors either).
- No need for MPLS (and therefore no need for an underlying IP IGP or LDP).
- No IP interfaces/subnets inside the Ethernet Fabric. IP interfaces exist only on VLANs where end-stations connect; that is, IP interfaces only exist as gateways for the VSN services they terminate.
- L3 VSNs can be IP Multicast enabled with one command per termination node (no need for complex IETF Draft Rosen or Next-Gen MVPNs support).
- Sub-second convergence because SPB relies on a single link state routing protocol (IS-IS), whereas MPLS is reliant on a BGP-LDP-OSPF protocol stack, which is significantly slower to reconverge.

A comparison of the traditional designs used to deliver Layer 3 virtualization with the SPB architecture can be found in Table 2.

Table 2 – SPB vs Traditional L3 Virtualization Technologies

	Multi-VRF (VRF-Lite)	MPLS (RFC 4364)	Fabric Connect
Applicability	Small networks	Service Provider & Carrier Networks Some large enterprises	Small to Large enterprise networks, campus, core, IoT Small service providers
VRF Scalability	Few VRFs (2-4) because of the need to support an instance of the IGP within each VRF	As many as PE device supports (typically 512-4000 on carrier platforms)	As many as BEB device supports (typically 24-512 on enterprise platforms)
IP route scalability	Limited by maximum number of IP routes supported on Core nodes	No real limit imposed by BGP. In practice limited by maximum number of IP routes supported on PE	Maximum of 20000 IPv4 routes per BEB (limited to maximum size of IS-IS LSP)
Core Foundation	IP hop-by-hop routing	IP hop-by-hop routing + MPLS label switching	Ethernet Switched Shortest Paths
Control Plane	As many instances of OSPF or RIP as there are VRFs	IGP (OSPF or IS-IS) and LDP/RSVP-TE on P and PE nodes Full mesh of MP-iBGP peerings on PE nodes The latter requires BGP Route Reflectors to scale	Single instance of IS-IS
Traffic Engineering	Not possible	Yes, with RSVP-TE	Standardized under 802.1Qca Path Control and Reservation. Not currently supported by Extreme
Control Plane extensions for IP Multicast	PIM in each VRF where IP Multicast required	MVPN draft-Rosen required PIM-SM in core IGP as well as in the VPN VRFs where IP Multicast required with GRE tunnels. Next-Gen MVPN uses MBGP and can support MPLS P2MP LSPs	None required (leverages IS-IS shortest path trees)
Control Plane extensions for IPv6	Requires additional OSPFv3 instances (or IPv6 tunnelling over IPv4)	Requires BGP+ support (RFC 2545) or native BGPv6	None required (Same as for IPv4; simply uses different IS-IS TLVs)
Control Plane extensions for L2 VPNs	Not possible (would require Spanning Tree core)	Possible but requires additional VPLS capability	Native (L2 VSNs)
Operational Complexity	Simple for just two VRFs Complex for more VRFs	High complexity	Simple (to design, to provision, to manage, and maintain)
Virtualization over WAN	Yes Using VRF enabled GRE	Yes Using MPLS over GRE	Yes Using Fabric Extend (SPB over IP or WAN E-LINE)

Layer 2 Virtualization Overview

SPB natively offers very flexible L2 connectivity over the Ethernet Fabric to achieve Transparent LAN Services (TLS). Transparent LAN Services provide the ability to connect Ethernet segments that are geographically separate. These networks appear as a single Ethernet or L2 VLAN domain. This basically allows any edge L2 VLAN to be extended to any other node in the SPB fabric. This can be applied to any VLAN, including the edge VLANs used in the previous L3 virtualization deployment model (IP routed as part of a given L3 domain), as well as in isolated (not routed) L2 VLAN segments. These Layer 2 Virtual Services Networks (L2 VSNs) offer an any-to-any (E-LAN) service type, which means that an L2 VSN can have any number of end-points. End device MAC address are called Customer MAC (CMAC) in Fabric Connect. CMAC learning is performed within the VSN service only on the BEBs where it is terminated.

SPB L2 VSN also offers an E-LINE (point-to-point) service, which is essentially just an E-LAN service with only two end-points, as well as E-TREE (private-VLAN) service that allows extension over the Ethernet fabric of edge private-VLANs where access ports can be configured as promiscuous or isolated.

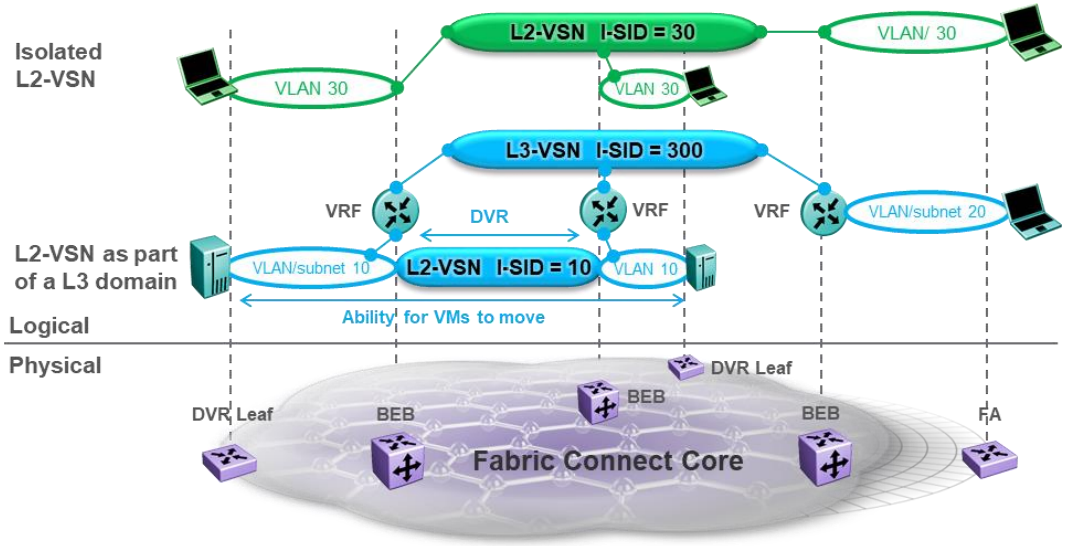


Figure 4 L2 Virtualization with SPB L2 VSNs

In Extreme Fabric Connect, L2 virtualization can also be tightly coupled with L3 virtualization so that a tenant can be allocated a number of L2 VSN segments, which are IP routed within an L3 VSN construct, as illustrated in Figure 4. This is not easily achieved with MPLS equivalent L2 & L3 virtualization technologies, but is addressed by EVPN.

SPB ensures that there cannot be any loops within the Ethernet Fabric backbone via Reverse Path Forwarding Check (RPFC), which checks all incoming packets are properly sourced. Extreme's implementation also fully supports Multi-chassis Link Aggregation Group (MLAG) on the SPB edge nodes (BEBs), which is referred to as Split Multi-Link Trunk (SMLT). This enables L2 VSNs to be redundantly terminated on access ports that are loop-free, without Spanning Tree and where all links are utilized in an active-active fashion.

The I-SID provides the glue to interconnect L2 VSN end-points. As such, VLAN IDs have only local significance at the BEB. VLANs can be re-mapped to different VLAN-ID values at the terminating end-points (BEB). An edge network interface in SPB is referred to as User Network Interface (UNI). UNIs come in three forms depending on the edge network service being connected to the L2 VSN. A VSN can be mapped to a VLAN ID (CVLAN UNI), a raw Ethernet port (Transparent UNI), or to combination of VLAN ID and Ethernet port (Switched UNI). The latter allows the same VLAN ID on different Ethernet ports of the same switch to be assigned to different L2 VSNs.

Tip

Benefits of SPB L2 VSNs over EoMPLS and MPLS-VPLS:

- Simple service definition via Service id (I-SID) configured on end-point VLAN and/or UNI port combination (literally one CLI command), instead of specifying a Virtual Circuit (VC) on EoMPLS or a Virtual Switch Instance (VSI) on VPLS.
- SPB L2 VSNs natively offer E-LAN (any-to-any), E-LINE (point-point) and E-TREE (private-VLAN) service types. In contrast, EoMPLS is only point-point and VPLS is an extension of EoMPLS that dynamically creates a full mesh of EoMPLS circuits to provide an any-to-any service type. This has a number of disadvantages.
- SPB L2 VSNs have no issue handling packet replication across the Fabric, which is needed to deliver broadcast and multicast traffic within the service. This is performed by allocating service-specific shortest-path trees. Whereas VPLS's primary shortcoming is that all broadcast and multicast packets need to be replicated by the ingress PE node multiple times on the same physical interface (each time with a different MPLS label), which becomes exponentially inefficient as the number of end-points in the VPLS VSI increases
- SPB combined with Extreme's SMLT/MLAG/vIST offers an active-active solution with redundant distribution (BEB-SMLT) nodes. VPLS has a major drawback with dual homing an access VLAN into 2 redundant distribution PE nodes as this results in an L2 loop which neither VPLS nor Spanning Tree can prevent. The common approach is to let only one of the PEs (Primary N-PE) do the traffic forwarding and the Standby N-PE only forwards traffic in case of failure; Primary and Standby PEs are usually staggered across the VSI instances
- No need for any BGP (and therefore no need for BGP Route Reflectors either).
- No need for MPLS (and therefore no need for an underlying IP IGP or LDP).
- L2 VSNs can be IP Multicast snoop-enabled with one click per end-point node. This is simply not possible with VPLS, which will always flood IP multicast traffic with the above-mentioned flaw that multicast traffic is inefficiently ingress replicated
- Ability to tightly integrate L2 virtualization with L3 virtualization within the same Fabric architecture. Something that is hard to achieve combining MPLS-VPNs and MPLS-VPLS.

Table 3 – SPB vs Traditional L2 Virtualization Technologies

	Spanning Tree	MPLS VPLS	Fabric Connect
Applicability	Very small networks	Service Provider & Carrier Networks	Small to Large enterprise networks, campus, core, IoT Small service providers
Shortest Path	No Path always along root tree	Yes Path can be traffic engineered	Yes, always
Robustness	Poor Spanning Tree needs to talk to block loops, which is then prone to meltdowns when things go wrong	Strong	Strong
Service Scalability	Limited to 4095 VLANs	4 billion if VPLS used with BGP signalling or auto-	16 million I-SID (24bits) theoretical limit

		discovery (VSI-ID is 32bits); no theoretical limit if VPLS is used with T-LDP; in practice the number of Pseudowires a PE can handle will be the limit	
Service Type Flexibility	E-LAN only and impossible to re-map VLAN-IDs	E-LINE, E-LAN, E-TREE combined with rich selection of UNI type interfaces	E-LINE, E-LAN, E-TREE combined with rich selection of UNI type interfaces
Control Plane	Spanning Tree (MSTP)	IGP (OSPF or IS-IS) and LDP on P and PE nodes. BGP used for VPLS auto-discovery; BGP or LDP used for VPLS Signalling. BGP requires use of Route Reflectors to scale	Single instance of IS-IS
Equal cost paths	No	No	Yes
Operational Complexity	Simple	High complexity	Simple

Data Center Virtualization Overview

Data center requirements have long been the most demanding in terms of network virtualization, and over the years have driven the development of a number of fabric technologies.

Multitenancy in a large data center environment requires a tight integration of L2 Virtualization and L3 Virtualization. L2 is needed to accommodate the virtualized nature of applications. Virtual Machines (VMs) can be dynamically moved within the data center or across data centers without interrupting the application, which means the VM IP address cannot change. L3 is required so that virtualized L2 server segments belonging to one tenant can be bundled within an IP routed domain (VRF). This ensures IP routing between those L2 segments as well as isolation from other L2 segments and L3 domains belonging to other tenants.

Traffic patterns in modern data centers have changed to be predominantly east-west requiring network designs to maximize the bandwidth and minimize the latency between the Top of Rack (ToR) leaf switches. For large-scale data centers, this brings a requirement for the fabric technology to support scalable multi-pathing in spine-leaf architectures where the bandwidth can be equal cost distributed over many ToR uplinks and spines. Whereas for smaller enterprise data centers where the east-west bandwidth requirements can be managed within single 40G/100G connections, adding links directly between ToR switches provides a cost-effective way to minimize the latency for east-west traffic while off-loading the traffic from the data center core.

Shortest path and lowest latency in the data center needs to be ensured for both L2 server traffic (always east-west) and IP routed server traffic (typically both east-west and north-south). IP routed flows within the data center pose a problem because they need to hit their L2 segment's default gateway for the IP routing to occur and the resulting path may no longer be optimal. To support this, the fabric needs to present a distributed anycast gateway on the ToR leaf nodes as well as distributing the knowledge about the location of host IP addresses within the data center. These requirements can only be met by shifting and replicating the IP routing capability as well as host IP knowledge across all the ToR Leaf nodes. From an operational perspective, this is undesirable as it is more manageable to keep the L3 configuration on a few data center Core/Spine layer nodes rather than having to replicate it across all the Leaf nodes.

The Extreme Networks Fabric Connect solution provides an ideal multitenant architecture with tight integration of L2 Virtualization and L3 Virtualization as covered in the preceding sections. It is also an

architecture that ensures shortest path and therefore lowest latency across any data center network topology. When the bandwidth demands from the single ToR leaf node are such that traffic needs to be distributed across multiple 40GbE or 100GbE core links, as is the case in large scale data centers, the topology of choice is spine-leaf. In a spine-leaf topology, SPB is capable to meet the multi-pathing requirements on condition that SPB is deployed with a number of BLVANs (which determines the number of equal cost paths an SPB Fabric can support) to match the number of Spine nodes.

Caution

The SPB standard (IEEE 802.1aq) defines a maximum of 16 BVLANs. Extreme's SPB implementation currently supports 2 BVLANs. This could be increased to 16 BVLANs support in the future.

With the addition of Distributed Virtual Routing (DVR), the SPB Fabric is further enhanced to deliver a distributed anycast gateway directly on the ToR DVR leaf nodes, which can now perform IP host routing for L3 data center flows. This, combined with knowledge of host IPs within the DVR domain, ensures shortest path and lowest latency for any traffic flow within the data center, whether at L2 or L3.

The DVR architecture has also been designed in such a way that the L3 configuration only needs to be provisioned and maintained on a few core/spine nodes acting as DVR controllers. The DVR controllers then automatically propagate the anycast gateway functionality across the DVR leaf nodes. This provides an ideal model which delivers on the data plane requirements yet preserves an architecture that is simple and easy to manage.

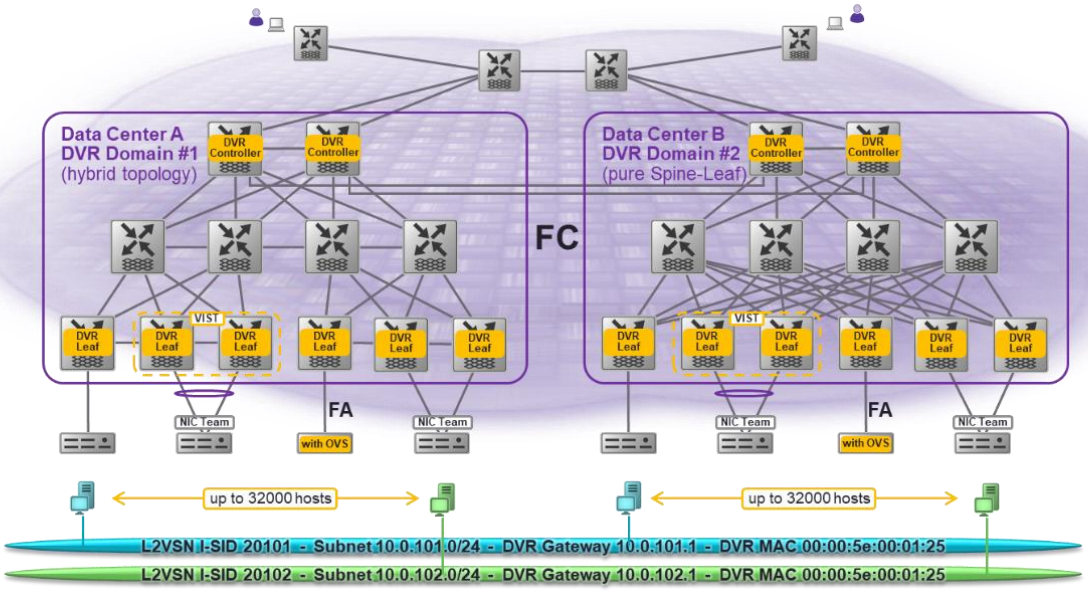


Figure 5 Data Center Virtualization with SPB and DVR

Servers and hypervisors can also be dual homed into redundant ToR leaf nodes in an active-active fashion. Hypervisors offer a wide range of NIC teaming hashing capabilities, some requiring the network side to present equivalent link aggregation capabilities while others do not. The selection of the right hypervisor NIC teaming mode is entirely dictated by the number of VMs supported on the hypervisor as well as the expected traffic load from those VMs.

Tip

The Extreme DVR leaf nodes are able to support NIC teaming requirements with and without the use of Virtual-IST SMLT clustering and with and without LACP.

Note

Extreme's SMLT clustering supports dual homing of server/hypervisors when the NIC teaming configuration requires link aggregation on the network side. Multiple homing (> 2) is only possible when the NIC teaming configuration does not require link aggregation on the network side.

For IP overlay data center solutions such as VMware's NSX, SPB can also provide a scalable and robust underlay infrastructure providing easy to deploy IP multicast for optimized NSX VXLAN deployments, avoiding any ingress replication requirements on the NSX overlay. This combined with the SPB-VXLAN Gateway and OVSDB functionality can be leveraged to connect bare metal servers operating in the SPB fabric underlay into the NSX overlay infrastructure.

Fabric Connect Positioning

Many fabric architectures are offered by different vendors, mostly for use in the data center but in some cases in the campus as well. This section is intended as an overview of that fabric landscape and to allow the reader to better understand how the Extreme Fabric Connect architecture compares to other fabric constructs.

TRILL and Derivatives

TRILL (Transparent Interconnection of Lots of Links) is an Ethernet Fabric standard developed by the IETF at roughly the same time as SPB was standardized by the IEEE. It is similar to SPBM in many respects as it also leverages IS-IS directly over Ethernet. It is notably different from SPBM in some design choices that were made and constitute a trade-off.

TRILL was designed to offer unlimited equal cost multi-pathing where individual TRILL switches can decide to hash egress traffic across any number of egress interfaces that all provide a lowest equal cost path to the destination, much like is possible with OSPF. However, this design choice precludes TRILL-based networks from leveraging the 802.1ag CFM since it becomes impossible for an ingress node to predict what path is used across the network.

Tip

With SPBM, equal cost shortest paths are only possible across an equal number of BVLANs. But within each BVLAN there is a single predictable shortest path from source to destination and from the ingress node 802.1ag CFM can be leveraged to validate that path.

The same design choice also implies that TRILL-based networks cannot make any assumptions on the validity of packets arriving on ingress interfaces (since those packets could have been hashed over any number of equal cost paths). So transient loops can only be dealt with by including a TTL field in the packets so that looping packets extinguish themselves after looping a few times, much like IP does with OSPF.

Tip

Within an SPBM BVLAN there can only be one predictable shortest path to any destination and source alike. SPBM therefore deals with transient loops by applying Reverse Path Forwarding Checks (RPFC) on every ingress packet. This ensures that transient loops are instantly suppressed.

The fact that TRILL requires a TTL field for transient loop suppression is one of the reasons why TRILL defines a completely different packet encapsulation from IEEE 802.1 Mac-in-Mac. The TRILL designers however made a mistake to simply use the VLAN-ID 12-bit field in the original implementation of TRILL

(RFC6325), as this limited the technology to: a) only extending L2 VLANs and b) being limited to 4095 VLANs at it. A later TRILL standard (RFC7172) inroduced Fine-Grained Labelling, which introduces an Inner label high part and low part each encoded over 12 bits, but at the price of defining yet another packet encapsulation incompatible with the previous one.

Tip

SPBM's Mac-in-Mac encapsulation does away with the VLAN used as a service ID field and instead replaces it with a more scalable 16-bit Service-ID (I-SID) which can be extended to offer any service type, much like on an MPLS network, as well as theoretically scaling up to 16 million services.

TRILL's unlimited equal cost multi-pathing capability is the reason why it has been developed and proposed in the data center market, as it fits well in large spine-leaf architectures (where SPBM instead requires as many BVLANs as spines), and where the virtualization requirements are limited to L2 only.

There are two TRILL implementations to note:

- Cisco FabricPath: Uses TRILL's IS-IS control plane but a Cisco proprietary packet encapsulation.
- Extreme Networks VCS (Virtual Cluster Switching): Formerly developed by Brocade, this implementation does use the TRILL packet encapsulation (supporting TRILL Fine-Grained Labelling) but uses Fibre Channel's FSPF (Fabric Shortest Path First) instead of IS-IS as the TRILL control plane.

None of the TRILL implementations are standards-based and inter-operable.

In summary, TRILL based fabrics offer solutions only for the data center and only covering L2 Virtualization without scaling beyond the 4095 VLAN-IDs and completely lacking multitenancy L3 Virtualization.

Ethernet VPN

Ethernet VPN (EVPN) running over a VXLAN overlay, brings together much of the MPLS VPLS and MPLS-VPN functionality in a new architecture focused on the data centre. EVPN tightly integrates L2 Virtualization with L3 Virtualization and at the same time allows active-active multi-homing of hosts (which was not possible with VPLS). The EVPN control plane remains BGP and is almost identical to the MPLS-VPN architecture based on Route Distinguishers and Route Targets but introduces new BGP NLRIs to advertise host IP and MAC addresses.

Tip

EVPN is a huge improvement over VPLS as it builds a consensus and cooperation between different router vendors and service providers to interoperate. VPLS had several different operating modes which made it prone to interoperability issues.

The EVPN standard is defined for operation over traditional MPLS, but in practice all implementations deploy the EVPN over a VXLAN overlay which eliminates the MPLS complexity.

The underlay network only needs to provide IP reachability for the VXLAN tunnel end points (VTEP). The BGP control plane remains complex and requires MP-BGP to run on every ToR leaf node. Most EVPN vendors use BGP not only as the EVPN Control plane but also as IGP for the underlay. BGP is designed to be scalable but it is not naturally fast and therefore must always be combined with BFD to achieve the kind of resilience expected in the data center. The resulting BGP configuration is fairly complex and becomes a spine-leaf of either eBGP peerings (where each leaf is a separate AS number) or iBGP peerings towards BGP Route Reflectors running on the spine nodes.

In large self-service oriented data centers where automation becomes a key element of the design, the EVPN provisioning complexity can be also automated.

Tip

Extreme Networks does offer data center EVPN-based solutions and these can be fully automated via Extreme Workflow Composer (EWC) as well as with integrated Embedded Fabric Automation (EFA) on SLX platforms.

Use of IP ECMP in the underlay underpins the ability for the VXLAN overlay to provide equal cost multipath between the leaf VTEPs. In fact, the EVPN model also ensures load-balancing towards multiple leaf VTEPs to which hosts are multi-homed, but this form of multi-path is only guaranteed to be shortest path if the underlay topology is spine-leaf (and hence all distant leaf VTEPs are an equal number of hops away). This is why all EVPN deployment models are always spine-leaf and EVPN is only positioned for the data center.

VXLAN brings its own constraints to the EVPN model in the way flooded traffic is handled for L2 Broadcast, Unknown, and Multicast (BUM) traffic. In theory, use of IP multicast in the underlay IP network is possible to allow the VXLAN overlay to replicate packets using more efficient underlay replication. But use of IP Multicast in what is a traditional IP routed underlay presents such scaling and operational challenges that all vendors proposing EVPN solutions recommend instead to deploy VXLAN using ingress replication, which is far less efficient.

This results in EVPN not being a suitable architecture to handle IP Multicast in the data center.

Note

IP Multicast is typically not a requirement for the large data centers deploying EVPN.

The same constraints are applicable to the way in which EVPN handles L2 broadcast and unknown packets. One Mbps worth of BUM traffic ingressing a leaf node on one access port could end up consuming 100 Mbps of the uplinks if those L2 BUM packets need to be replicated to 100 other leaf VTEPs; the effect is quickly compounded if many end-points emit BUM traffic. Therefore, EVPN goes to considerable length to reduce BUM traffic by making BGP advertise the host MAC on the one hand (so that unknown traffic can be minimized) and implementing ARP-suppression mechanisms to reduce the number of ARP broadcasts. Yet in scaled environments this can result in the BGP control plane containing more MAC/ARP entries than can be programmed in the hardware forwarding tables, so these mechanisms are often combined with conversational learning techniques to limit them to just active conversations or traffic flows.

The EVPN model also needs extra complexity to handle BUM replication towards multi-homed hosts by implementing a Designated Forwarder and Local Bias rules, which would otherwise result in duplication and reflection of BUM traffic.

Still, even if out of a necessity, EVPN's ability to control and reduce Ethernet broadcasts brings new capabilities that can be relevant in some environments. Using control plane IP/MAC learning can provide a consistent forwarding database in any size network instead of relying on flooding and learning. Control plane learning also offers greater control over MAC learning in its ability to apply policies, that is, "who learns what." This provides benefits in Data Center Interconnect (DCI) over WAN connections where the level of L2 broadcasts can be greatly reduced.

Note

Extreme Fabric Connect does not currently offer any mechanisms to suppress or eliminate Ethernet broadcast beyond the traditional multicast and broadcast rate limiters. Within SPB L2 VSN services, MAC learning occurs in the data plane via the flooding of BUM traffic because this is the most effective and scalable approach for SPB. Unlike EVPN it does not pose any operational issues and the BUM traffic is in not magnified when transmitted across the Fabric.

In summary EVPN is an architecture that is almost exclusively aimed at the data center and should only be deployed in spine-leaf architectures. EVPN's greatest benefits are in the very large data centers where the use of BGP allows not only to scale beyond what an SPB based data center could scale to, but also interconnect private cloud with public cloud.

Tip

Extreme Networks offers compelling EVPN-based data center solutions around the SLX & VDX & MLX product families. Extreme Workflow Composer (EWC) is a vital component of these solutions allowing automation and SDN orchestration of every aspect of the data center.

Table 4 – SPB vs Competing Data Center Fabric Technologies

	TRILL/FabricPath/VCS	EVPN	Fabric Connect
Applicability	Small to Medium data center networks	Large and very large data centers	Small to Large data centers
Virtualization Capabilities	L2 only (L3 needs to be provided from "outside" the fabric)	L2 and L3 tightly integrated	L2 and L3 tightly integrated
Service Scaling	Limited to 4095 VLAN ids for L2 virtualization. Extended up to 2^24 with TRILL Fine-Grained Labelling.	16 Million theoretical limit (24 bit VXLAN VNI)	16 Million theoretical limit (24 bit I-SID)
Fabric Type	Ethernet Fabric	VXLAN Overlay (IP Fabric)	Ethernet Fabric
Control Plane	IS-IS (TRILL / FabricPath) FSPF (VCS)	MP-BGP	IS-IS
Shortest Path	For L2 flows only	For both L2 & L3 flows	Always for L2 flows, with DVR for L3 flows as well
Spine-Leaf Topology	Yes	Yes, superspine / spine-leaf is the only topology which can be used with EVPN	Yes, but SPB needs to be deployed with as many BVLANs as there are Spines
Other Topologies	Yes	No	Yes
Distributed Anycast Gateway	No, L2 only. Possible with Extreme VCS Fabric	Yes	Yes, with DVR
L3 capable Leaf	No	Yes	Yes, incl. DVR leaf
L3 capable Spine	No	No, in the EVPN model Spines are not VTEPs	Yes, incl. DVR controller
L3 Provisioning	n/a, done outside Fabric	Across all Leaf nodes and needs to be consistent	On DVR controllers only (Distribution/Spine layer)
Data Center External Connectivity	Via Core/Spine layer	Via dedicated Border Leaf (not recommended from Spine)	Via DVR controllers (acting as either Distribution/Spine layer or Border Leaf)
Dual Homing of Host (NIC teaming) into two ToRs	Yes	Yes	Yes

	TRILL/FabricPath/VCS	EVPN	Fabric Connect
Multi Homing of Host (NIC teaming) into >2 ToRs	No, if MC-LAG/SMLT is required on ToR Yes, with Extreme VCS Fabric Yes, if Hypervisor hashing mode negates use of MC-LAG/SMLT on ToR	Yes (but not all EVPN vendors offer this capability) (Extreme SLX/VDX platforms do not offer this capability)	No, if MC-LAG/SMLT is required on ToR Yes, if Hypervisor hashing mode negates use of MC-LAG/SMLT on ToR
Flooding of Broadcast, Unknowns and Multicast (BUM)	Via elected Root bridge. Not shortest path, but using efficient replication.	Inefficient ingress replication performed on the VXLAN overlay	Efficient service-specific shortest path multicast trees
Control Plane advertises host MACs	No, data plane MAC learning	Yes, via BGP EVPN Route Type 2	No, data plane MAC learning performed within L2 VSN service
Control Plane advertises host IPs	No, L2 only	Yes, via BGP EVPN Route Type 2	Yes, with DVR
Technology Scalability Limit	4059 VLANs	BGP scaling limits	Extreme SPB Fabric is typically limited to a maximum of 500 FC nodes (though some VSP platforms can scale higher)

Cisco's Campus Fabric

On the campus side, another fabric technology worth mentioning is Cisco's Campus Fabric as part of Cisco's Digital Network Architecture (DNA) framework. This fabric technology is essentially very similar to EVPN as it is also based on a VXLAN overlay but differs from EVPN in its choice of the overlay's signalling control plane, which is not based on BGP but on LISP (Locator Identifier Separation Protocol). According to Cisco, LISP offers an alternative to BGP that can run with smaller tables and less CPU power. But where EVPN only had VTEPs, these now become LISP Tunnel Routers in the Cisco fabric, which also requires a LISP Map Server/Resolver as well as LISP Proxy Tunnel Routers for external connectivity that add complexity to the solution and require these functions to deployed in a redundant fashion to avoid single point of failures.

Like EVPN, Cisco's Campus Fabric can support a tight integration of L2 and L3 virtualization as well as the ability to support distributed anycast gateway. Unlike EVPN, there is no requirement for spine-leaf architectures and all topologies are supported and IP Multicast is supported (both PIM-SM and PIM-SSM), but is again implemented using inefficient ingress replication in the underlay.

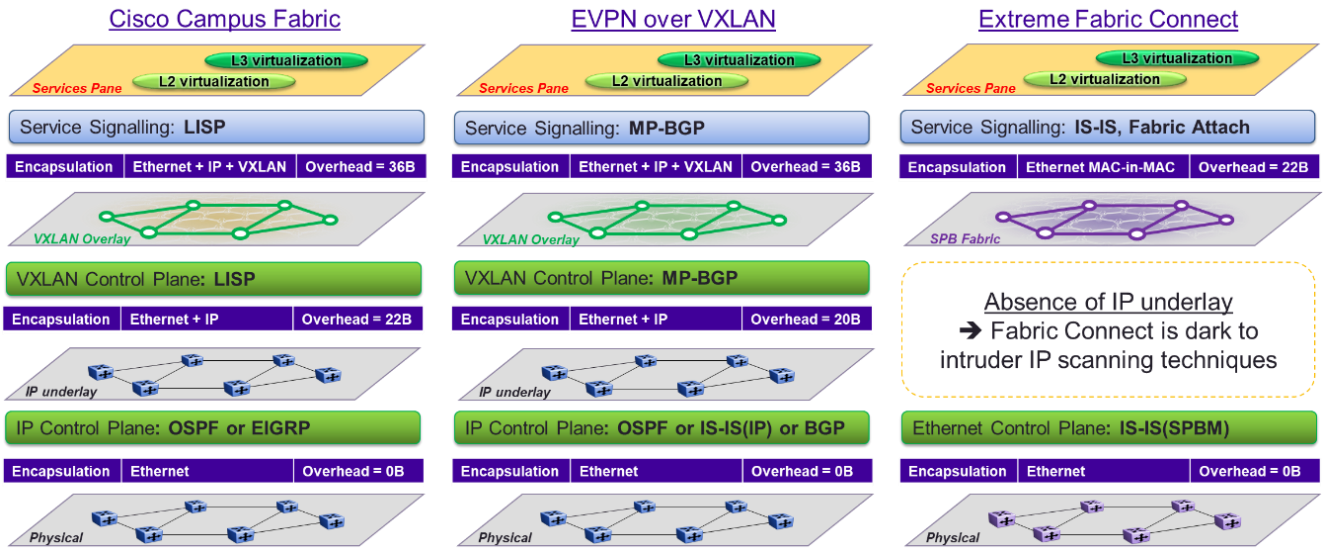


Figure 6 Overview of Fabric Layers and Overlays

Guiding Principles

The goal of network virtualization is to decouple the network users and their applications from the underlying infrastructure so that those users and applications can be segmented from other users and applications while still sharing the same physical network infrastructure resources (bandwidth and connectivity). One network must therefore support many different virtual networks.

These networks become logical and their addressing (IP routes and user-VLANs where end-station MACs are learned) is only applicable within the Virtual Services Network (VSN) to which they belong. Leveraging network virtualization to segregate network users into different domains is always far superior than trying to achieve and maintain the same goal by leaving all users and services in a single Global Routing Table (GRT) with a common addressable IP address space, and then attempting to manage connectivity with extensive use of Access Control Lists (ACLs) or by making all VLANs non-routable within the network and performing all routing across firewall interfaces.

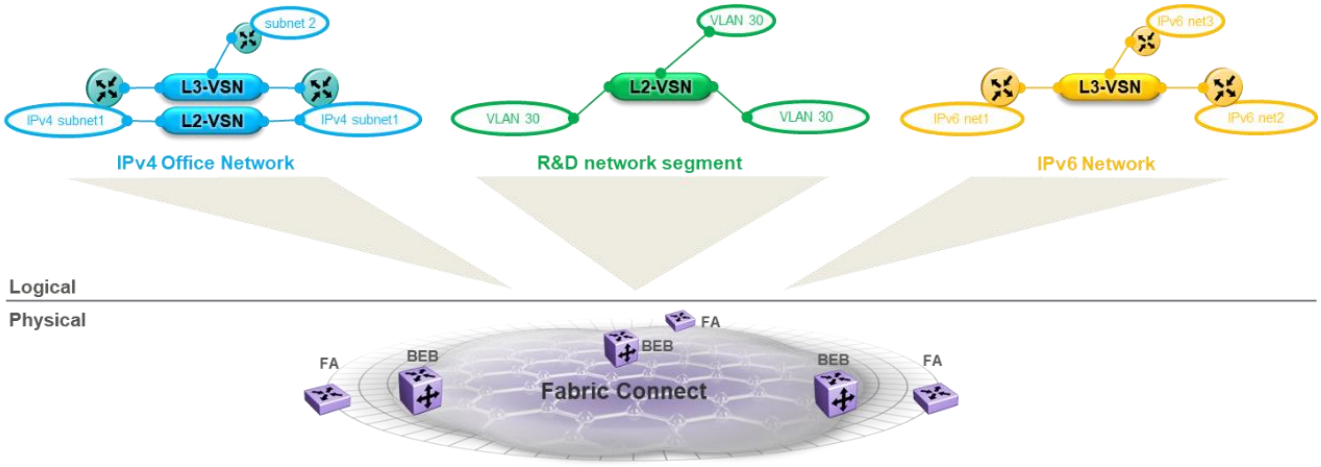


Figure 7 Virtualization of Logical Networks over SPB

The demarcation of physical and virtual networks has never been as rigorous as it is with SPB where the Ethernet Fabric deals with SPB Backbone MAC addresses (BMAC) tied to the physical infrastructure and used by IS-IS to compute the shortest path across the physical topology. End-user MAC addresses are handled within L2 VSN service types and IP routes (IPv4 or IPv6) are advertised within L3 VSN service types. In short, with SPB, IP routing is no longer the foundation of the network backbone, but becomes purely a service above it.

Tip

For the sake of comparison, MPLS needs an underlying IPv4 IGP in order to operate, which in carrier networks is contained inside the core and used only by MPLS (and LDP and iBGP). But in Enterprise networks, this is seldom the case. MPLS deployed in an enterprise environment is thus inconsistent in its use of VRF-0 (Global Router), which plays a dual role of foundation of the MPLS's heavy control plane protocol stack as well as acting as a legitimate L3 domain for some network users and applications.

When virtualization is combined with a simple and consistent end-point provisioning based on a common Service-ID (I-SID) concept, as is the case with Fabric Connect, we have a powerful solution for creating virtual networks on demand.

This makes the Extreme Networks Fabric Connect architecture an ideal SDN framework, which is not based on overlay technologies and is not limited to the data center alone but instead becomes an end-to-end network foundation and enables a highly automated and programmable SDN edge[2] with IEEE 802.1Qcj

2 See Fabric and VSN Security on page 154.

based Fabric Attach signalling. Fabric Attach (FA) is the foundation for the elastic nature of the Fabric Connect architecture as it allows SPB's service I-SID to be seamlessly extended to users, applications and IoT devices without any manual intervention.

Fabric Attach may be deployed in a number of different manners, from an edge-only provisioning model with automated core service connectivity. It may also be SDN enabled through the use of Open vSwitch,[3] which supports Fabric Attach signalling and is deployed for IoT devices using the Extreme Defender for IoT security solution. This solution leverages ExtremeCloud Appliance and uses the Extreme Defender for IoT device as a form of "security proxy device" associated with the IoT device to extend SDN functionality and security to devices that would otherwise be unable to participate in the SDN automation and security functionality provided through the use of the Fabric Connect network.

The Fabric Connect network can also successfully coexist with software-based SDN overlays (e.g., VMware's VXLAN and NSX SDN architecture) while also significantly simplifying and streamlining the underlying infrastructure.

Figure 8 depicts the typical campus + branch + data center deployment model for the Extreme Fabric Connect architecture. The SPB fabric edge devices are the Backbone Edge Bridges (BEB) nodes, which are typically acting as distribution layer switches within the network. This is where the VRFs are defined and thus where the IP interfaces are located, acting as default gateways on the respective user-VLANs which are L2 extended out to the Fabric Attach access layer. Both L3 and L2 VSNs are terminated on the BEB nodes but end-point provisioning of those user-VLANs can either be done directly on the Fabric Attach access switches or can be obtained via Identity based networking where a user is assigned to a VLAN+I-SID directly via 802.1X authentication.

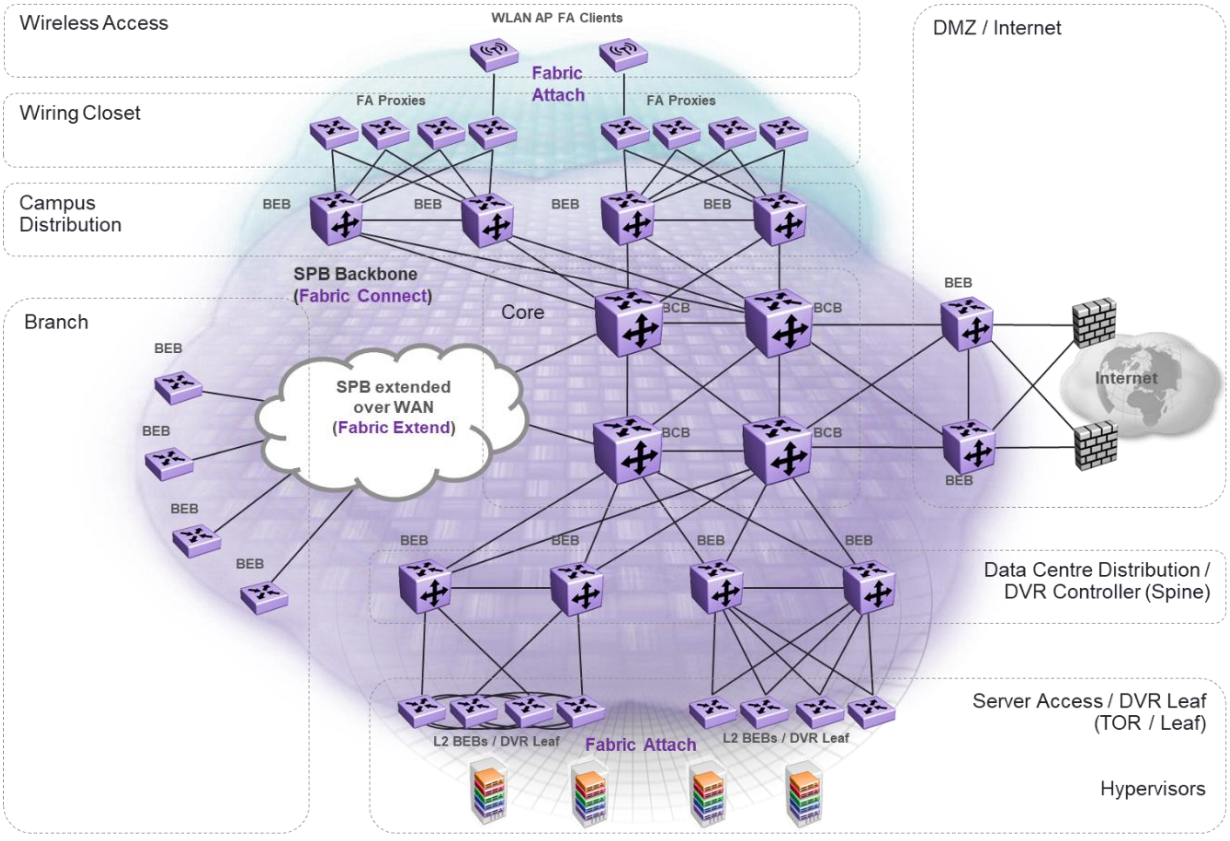


Figure 8 Fabric Connect Reference Deployment Model

3 See Fabric Area Network on page 45.

Tip

If we compare this to a traditional IP/MPLS design, we note that we no longer need any dedicated redundant IP routers to act as BGP Route Reflectors and we also no longer need any dedicated WAN IP routers.

The wiring closet access switches remain L2 switches where the user-VLANs are defined. While these switches could also be made to be part of the SPB backbone, there is little advantage in doing so, since these switches will always be dual homed into a pair of distribution layer nodes and as such will never act as transit nodes for IS-IS and SPB. A much more efficient design for the wiring closet access switches is to use link aggregation (static or LACP-based) on their uplinks using SMLT on the distribution BEB nodes. Extreme's Fabric Attach can then also be used to seamlessly extend the I-SID service provisioning to these wiring closet switches. The outcome is that users can be directly assigned to the correct VLAN/I-SID by simply configuring the access port where they connect (or via RADIUS assigned VLAN/I-SID during 802.1X EAPoL authentication) without any need to manually configure or manage user VLANs on the Fabric Attach uplinks into the BEB distribution.

Fabric Connect can also be extended to the data center all the way to the access L2 ToR switches, which can terminate L2 VSNs directly and make any server VLAN appear on any ToR across any data center. The benefit of extending the Ethernet SPB fabric to the ToR switches is that in small to medium data centers these can now be interconnected (or fully meshed) with high speed Ethernet links to provide very high bandwidth and very low latency precisely where it is needed, across the ToR switches. Data center east-west (server-server) traffic, which is increasingly becoming predominant over north-south (client-server) traffic, can now be delivered along the IS-IS computed shortest path leveraging the ToR fast interconnects. For large scale data centers a spine-leaf architecture will help ensure that all the ToR switches are an equal number of hops away from each other and use of SPB ensures that the east-west traffic can be load balanced across as many equal cost shortest paths as there are spine nodes. In all cases, Fabric Attach completes the picture by extending the benefits of the fabric all the way into the hypervisor vSwitches, allowing applications/VMs to automatically attach to the correct VSN service.

Because all the VSN service termination is performed at the Access (FA) & Distribution (BEB) layers, this leaves us with the high-speed Core devices (Backbone Core Bridges – BCBs) that have a number of desirable attributes. To start, these BCB nodes are completely agnostic of the virtual networks they transport. By definition, no user VLAN or VRF is defined on a BCB and thus the BCB has no knowledge of user MAC addresses transported in L2 VSNs or of IP routes used in L3 VSN VRFs. Note that with Fabric Connect, services may be deployed anywhere in the network, and any node may then become a BEB – even if it does reside at what would traditionally be considered the core or distribution layer. However, when used as SPB BCBs at the core, these nodes simply perform a transport function, delivering SPB's Mac-in-Mac packets to their BMAC destination along the IS-IS computed shortest path.

This makes for a highly scalable design, as there is virtually no scaling impact on the Core BCBs in relation to the number of virtual network VSNs provisioned on the network. It's also a highly robust and reliable design where any creation or removal of a virtual network can be done via end-point provisioning without any need to touch or re-configure critical core nodes.

The SPB Ethernet fabric is also extended to branch offices over the WAN (various options exist to achieve this, including running SPB over IP if the WAN provider is offering an IPVPN service), which allows network users located in the branches to be seamlessly integrated within the same L3 VSN domains, for IPv4 and/or IPv6 and with or without IP Multicast routing support (extending L2 VSNs is equally possible though usually not a requirement).

In summary, Fabric Connect is an Ethernet fabric solution, covering all of campus core, data center, wiring closet and branch.

Fabric Connect and Fabric Attach

The Fabric Connect name is often used to designate the entire Extreme SPB-based fabric solution, including the other variants, Fabric Attach & Fabric Extend, touched on below. However, in this section, Fabric Connect (FC) is the Extreme Networks name given to the core SPB-based fabric technology that is based on IS-IS and Mac-in-Mac (SPBM). A node running in Fabric Connect mode will need to run the IS-IS control plane protocol, form IS-IS adjacencies with its Fabric Connect neighbors, and will then be knowledgeable about the physical topology and how all Fabric Connect nodes are interconnected. This will allow it to calculate shortest paths to every other Fabric Connect node in the network and to react fast and recalculate those shortest paths should any change be detected in the active topology. Given that Mac-in-Mac is the encapsulation used by SPBM, any device running in Fabric Connect mode will need to have Mac-in-Mac capable hardware/chipsets. The IS-IS SPB control plane also allows it to advertise a desire to participate in a service type by announcing the corresponding I-SID for the relevant service type.

As the naming implies, Fabric Attach is all about attaching users and devices to the inner Ethernet SPB Core Fabric, that is, Fabric Connect. You can think of Fabric Attach as a subset of Fabric Connect, where the IS-IS control plane is removed (and with it its ability to calculate shortest paths) and traditional Ethernet based MAC encapsulation and MAC learning are used (without any requirements for Mac-in-Mac). Only the ability to attach users to I-SID based services is preserved and this service attachment signalling is now handled by LLDP. A further restriction of Fabric Attach is that only L2 I-SIDs can be signalled; that is, users and applications must always be Fabric Attached to some L2 VSN segment that needs to exist on at least one Fabric Connect node somewhere.

Tip

L3 I-SIDs for L3 VSN attachment are only relevant and will only exist on Fabric Connect nodes that are acting as IP routers with VRF support. It is perhaps on these nodes that L2 I-SID VSN segments have been defined and bound to a particular VRF and thus L3 VSN. Fabric Attach at the edge can be used to place users and applications in the desired L2 VSN segment belonging to the desired L3 VSN tenant IP routing domain.

Table 5 – Properties of Fabric Connect vs. Fabric Attach

Property	Fabric Connect	Fabric Attach
Able to pre-provision Fabric access port based on edge user/device type	no	yes
Allows node to compute shortest path/trees across Fabric using IS-IS/SPB	yes	no
Allows node to signal end-point Service Attachment to Fabric L2 I-SID	yes	yes
Allows node to signal end-point Service Attachment to Fabric L3 I-SID	yes	no
Node counts towards maximum scaling limit of SPB Nodes per Region	yes	no
Requires node's hardware chipset to be SPB (Mac-in-Mac) capable	yes	no

At the same time, Fabric Attach enhances the Fabric Connect end-point provisioning by being able to detect what edge device type is being connected, which allows for automatic configuration of the fabric access port to accept the end device into the correct VSN. Therefore, edge devices, such as virtual switches, ExtremeWireless Access Points, video surveillance cameras, etc., can be Fabric Attach enabled and act as FA Clients, which then allows them to advertise their identities securely. This information can either be used directly by the Fabric Attach access switch or relayed to the RADIUS server for secured MAC based authentication and authorization onto the desired Fabric I-SID via Fabric Attach RADIUS outbound attributes.

Tip

Fabric Attach includes secure message authentication, which results in a secure deployment where IoT device MACs cannot be spoofed by attackers trying to penetrate the network.

Clearly for certain devices that are not network switches (server hypervisors, ExtremeWireless Access Points, end stations), the benefits of using Fabric Attach are obvious. There would be little point or value in making these devices run IS-IS or use Mac-in-Mac encapsulation, but the benefits of allowing these devices to automatically place or authenticate themselves into the correct VSN are tremendous. The same trade-off is also usually true for network switches acting as wiring closet access switches. In most enterprises, wiring closet switches are often stacks, only operate at Layer 2, and are dual homed with both uplinks connecting into the same pair of distribution nodes. The traditional and most effective way of configuring those uplinks is as a LAG bundle with Multi-chassis LAG (Extreme SMLT clustering) running on the distribution nodes. As such the wiring closet switch only has one logical uplink into the network (i.e., it is stub connected), and again there is little benefit to be gained in making that switch run IS-IS in full Fabric Connect mode. Besides, running a link-state routing protocol, such as IS-IS, in stacking architectures and ensuring fast failovers in case of unit failures, would require sophisticated high availability software capability, which is typically found only on more expensive core/distribution platforms.

There are instead significant benefits in letting that node run simply Fabric Attach (in FA Proxy mode). Wiring closet access switches can be built with more cost-effective hardware with less powerful CPUs and non-necessarily MAC-in-MAC capable chipsets. Also, in very large enterprises the wiring closet access switches constitute the biggest proportion of network-wide deployed switches and letting them run in Fabric Attach mode ensures that the Fabric Connect core remains well within the maximum number of supported SPB nodes per region scaling limit.

Caution

An Extreme SPB Fabric can currently scale to a maximum of 500 nodes per region (area) assuming a range of SPB VSP platforms are in use (Note, some VSP platforms can scale higher than 500; refer to product Release Notes). This figure is dictated by the current generation of ASICs, but is expected to rise considerably as newer chipsets become available in the years to come. Future support of Multi-Area Fabric will also allow designs to exceed this limit.

On the other hand, in some environments the above-mentioned premises for the wiring access switches may not hold true and then use of Fabric Connect might indeed be a better option than Fabric Attach. If for instance the access switches were physically deployed in a ring or chain topology, or even dual homed to a variety of different core/distribution nodes which cannot easily be paired into SMLT clusters, then use of Fabric Connect would be required. In such topologies, the access switches might be deployed as standalone switches and not stacks, or if deployed as stacks, in case of Base/Maser unit failure, it might be acceptable for them not to re-converge in sub-second times.

Other considerations may involve the use of:

- IP Multicast applications. An FA Proxy switch remains essentially an L2 switch, which, to handle IP multicast, will need to have traditional L2 IGMP snooping enabled. The use of IGMP snooping is highly effective in IPTV type deployments where the end stations connecting to the FA Proxy switch are acting as multicast receivers. Use of Fabric Connect would not provide any gains in this case. However, in video surveillance deployments, where the end stations (cameras) are acting as multicast senders, an FA Proxy IGMP snooping switch is somewhat less efficient as all multicast streams will always be flooded towards the IGMP querier node (i.e., the distribution Fabric Connect BEB node acting as FA Server),whether or not active receivers exist for those streams. In these video surveillance environments, a Fabric Connect solution to the edge delivers the best results.

- Requirement to terminate a large number of I-SIDs on a per-access switch basis. Since Fabric Attach leverages LLDP for service signalling, the Fabric Attach TLV cannot be larger than a given size.

Caution

LLDP's TLV maximum size equates to the ability to request a maximum of 94 I-SIDs with Fabric Attach and not more. Should there be a requirement for an access switch to terminate more than 94 I-SIDs then the Fabric Connect mode could be used.

- Availability of advanced UNI types on the access switch. Fabric Attach caters only for the most common CVLAN UNI type. Use of the more sophisticated Switched-UNI and Transparent UNI modes would require use of Fabric Connect mode.

Fabric Extend

Given the powerful service virtualization capabilities of the Extreme Fabric Connect architecture there was a strong demand from the outset to be able to extend the same capabilities to the branch or between geographically distant networks. This led to the development of Fabric Extend.

Fabric Extend is able to extend the SPB Ethernet Fabric, as an overlay, over the WAN. This then allows the L2 and L3 VSN service types to be extended with ease all the way to the branch office using the same powerful end-point provisioning of Fabric Connect.

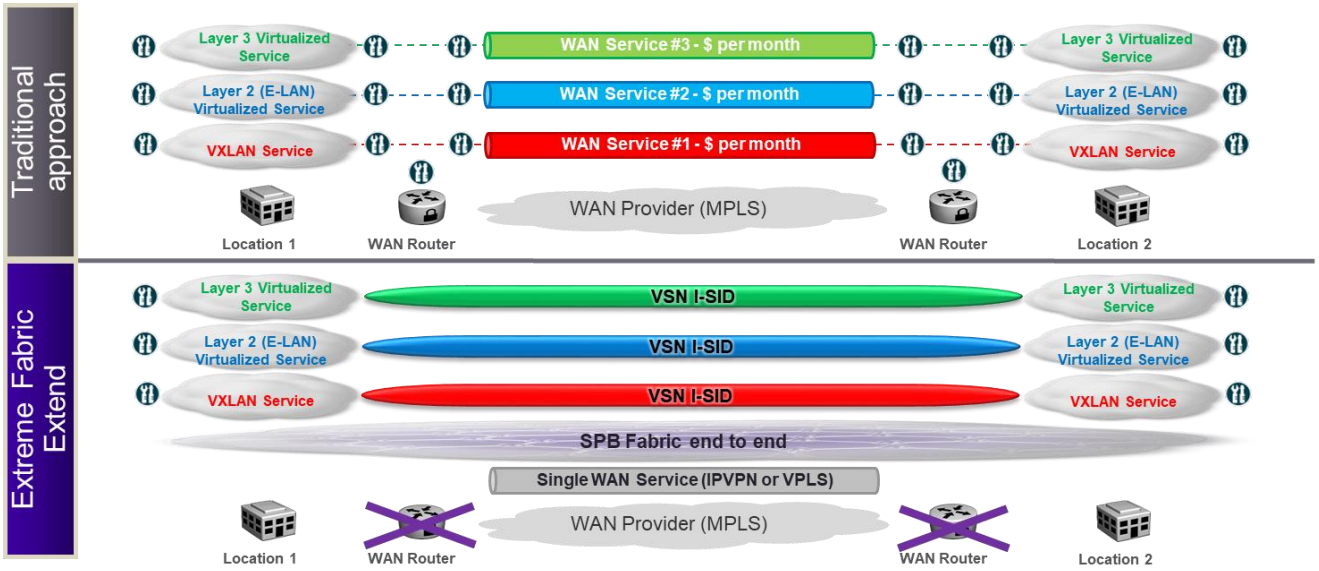


Figure 9 Benefits of Extending Fabric Services with Fabric Extend

As depicted in Figure 9, this brings a number of important benefits. To start, the traditional WAN router is no longer required as the WAN circuits can be directly connected to Extreme's Fabric Extend capable switching platforms.

Tip

With Fabric Extend the traditional WAN Routers are no longer required.

But the real benefit comes when there is a need to extend virtualization (typically in the form of L3 VSNs, but L2 VSNs are also possible) all the way to the branch office. In a traditional WAN architecture, the only way to provide separation for different user groups or tenants would be to purchase separate WAN services for each tenant and map these to the customer's VRF instances. This mapping needs to occur on the WAN router, requires participation from the WAN operator, and thus hinders the ability to deploy those services end-to-end on-demand as every new service would require touching the WAN routers and waiting for the WAN provider to provision the new WAN service.

Because Fabric Extend is in effect extending the SPB Ethernet Fabric foundation, the WAN provider is no longer aware of the customer virtualized VSN services and, in the case of IPVPN WAN service types, is no longer required to participate in IP route advertising with the end customer.

Tip

With Fabric Extend only one WAN service needs to purchased.

Tip

If the WAN service is IP routed (IPVPN MPLS-VPN) with Fabric Extend there is no longer any need for the WAN provider to learn and exchange routes (via OSPF or BGP) with the customer's network.

A further benefit of Fabric Extend is the simplicity with which IP Multicast applications can be extended to the branch office, just as if the branch office users were connected to the central office network. In a traditional WAN architecture, transporting IP Multicast over the WAN is riddled with so many challenges to make the WAN router run PIM, and the WAN operator to offer an IP Multicast capable service that in the practice most enterprises decide not to do so.

Tip

IP Multicast applications can easily be extended to the branch office with Fabric Extend.

Fabric Extend will be covered in greater detail in a separate section, including the various modes in which it can be deployed. But essentially the way in which Fabric Extend creates an overlay over the WAN is by using an overlay of point-to-point tunnels (IP tunnels using either a VXLAN or IPsec encapsulation, or native L2 tunnels) that rely on additional packet encapsulation, which can result in Fabric Extend traffic to exceed the standard Ethernet maximum frame size of 1518 (untagged) or 1522 (tagged) bytes. From a technical standpoint, virtually all WAN operators use MPLS-based networks, which are perfectly capable of supporting Ethernet oversized frames, but it is important to verify that the WAN operator will allow the use of oversized frames in the services it offers.

Caution

Fabric Extend in IP (VXLAN) mode requires the WAN to support frame sizes of 1600 bytes. Fabric Extend in L2 mode requires the WAN to offer point-to-point circuits where the maximum frame size will be 1544 bytes.

Note

If it was desired to support jumbo frame sizes (IP MTU of 9000 bytes) between geo-redundant data centers for instance, this is perfectly possible with Fabric Extend and the Extreme VSP platforms (which can handle maximum frame sizes of up to 9600 bytes), but would require the WAN operator to also support jumbo frame sizes in the WAN service they offer and be able to handle maximum frame sizes of 9100 bytes.

The same is not true for the Internet, which strictly enforces an IP MTU of 1500 bytes. Deploying Fabric Extend over an Internet connection is possible but will require the use of IP Fragmentation on the Fabric Extend nodes deployed at either end of the Internet. The Internet also raises secutity considerations and any Fabric Extend deployment over the Internet will need to be encrypted using an IPsec encapsulation. These requirements will ultimately determine which VSP or XA platform will need to be deployed.

Note

Fabric Extend with IP fragmentation is supported with the Fabric Connect VPN XA1400 platforms as well as with the Extreme Networks VSP4450/4850 platforms (latter require use of the Open Network Adapter - ONA).

Note

Fabric Extend with IPsec encryption is supported with the Fabric Connect VPN XA1400 platforms.

Data Center Architecture

The Extreme Networks Fabric Connect provides an ideal multitenant architecture for the data center where both L2 and L3 VSN service types can be tightly integrated, fully supports IP Multicast applications, and remains topology agnostic to adapt to small- and large-scale data centers alike.

For smaller size data centers, the ToR switches can be intermeshed using high speed 40GbE interconnects which in turn reduces the overall cost of the solution. This is because the data center distribution layer no longer needs to aggregate the same amount of ToR uplinks as the east-west traffic flows are handled via the ToR high speed interconnects.

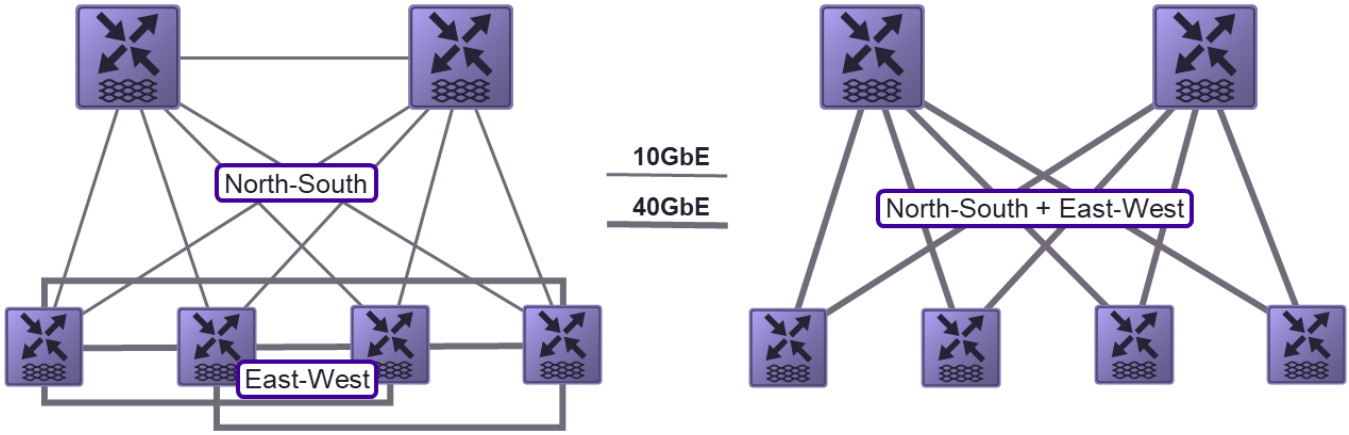


Figure 10 Smaller (Meshed) vs Larger (Spine-Leaf) Topologies

For larger scale data centers where it becomes impractical to fully mesh or interconnect the ToR switches or where the east-west bandwidth requirements exceed the ToR single high speed Ethernet uplink speed, the topology of choice is spine-leaf, which ensures a consistent lowest equal cost path and multi-path capability between any pair of ToR switches. To achieve the multi-path capability in a spine-leaf topology, the Fabric Connect will need to be deployed with as many BVLANs as there are spines.

Caution

Extreme's SPB implementation currently supports two BVLANs. This could be increased to 16 BVLANs in the future.

In larger data centers and those where multi-tenancy is required, east-west flows will often be IP routed L3 flows where the communicating VMs are located in different IP subnets of the same "tenant" VRF domain, yet VMs can be highly mobile and are prone to being moved to any hypervisor within or across different data centers. Having those L3 east-west flows redirected to a centralized default gateway node in the Spine/Distribution does not necessarily ensure the shortest path and lowest latency (both VMs could even be connected to the same ToR switch on different or the same hypervisor).

The only way to ensure that such east-west L3 flows are also switched along the shortest path, like L2 flows, is for the ToR switches to implement a distributed anycast gateway function whereby the ToR switch

becomes not only the default gateway for any traffic flow that is not L2 but also has knowledge about every other host IP within the data center. In the Extreme Fabric Connect architecture these enhanced capabilities for the data center come under the name of Distributed Virtual Routing (DVR).

Caution

Currently DVR is only supported with IPv4 but will be extended to IPv6 in the future. See DVR limitations and Design Alternatives on page 130 for more details.

The table below provides a brief summary of the additional properties which DVR brings to a Fabric Connect data center.

Table 6 – Data Center Fabric Properties with and without DVR

Property	With DVR	Without DVR
ToR switches implement distributed anycast gateway for all server/VMs	yes	no
East-west L2 flows will always take shortest path (lowest latency)	yes	yes
East-west L3 flows will always take shortest path (lowest latency)	yes	no
North-south flows can be made to always take shortest path	yes	no
ToR switch data plane performs L2 VSN BEB switching function	yes	yes
ToR switch data plane performs L3 VSN (and IP Shortcuts) BEB IP routing function	yes	no
ToR switch is managed as L2 access switch (no IP configuration)	yes	yes
ToR switch supports Fabric Attach	yes	yes
ToR switch supports Virtual-IST SMLT clustering	yes	yes

DVR offers a scalable architecture where data centers can be associated with one or more DVR domains. DVR domains are made up of DVR controllers and DVR leaf nodes and can currently scale up to a maximum of 40,000 hosts. Within each DVR domain two or more DVR controllers must exist that can manage up to 250 DVR leaf nodes. The DVR controller function can be located either in the spine nodes or in the distribution layer or indeed in a dedicated data center border nodes since all north-south traffic leaving the data center DVR domain will always transit via the DVR controllers. All 10-25GbE ToR switches become DVR leaf nodes, while smaller 1GbE ToR switches can be connected as FA-Proxy switches either directly to DVR leaf nodes (co-located in the same server rack) or to the DVR controller nodes, as depicted in Figure 11.

Note

Data center design often calls for some 1GbE connectivity within the server racks. However, this connectivity is aimed either for server lights-out management or older non-virtualized bare metal servers. The benefits of bringing these connections into a DVR leaf node are not significant.

Tip

Extreme offers both capabilities, either by deploying the VSP 4k platforms as a DVR leaf (Note: scaling limitations apply for DVR on the VSP4k platform) or using ERS or ExtremeOS platforms in FA Proxy mode, which can be redundantly connected to the 10GbE DVR leaf nodes located in the same server rack.

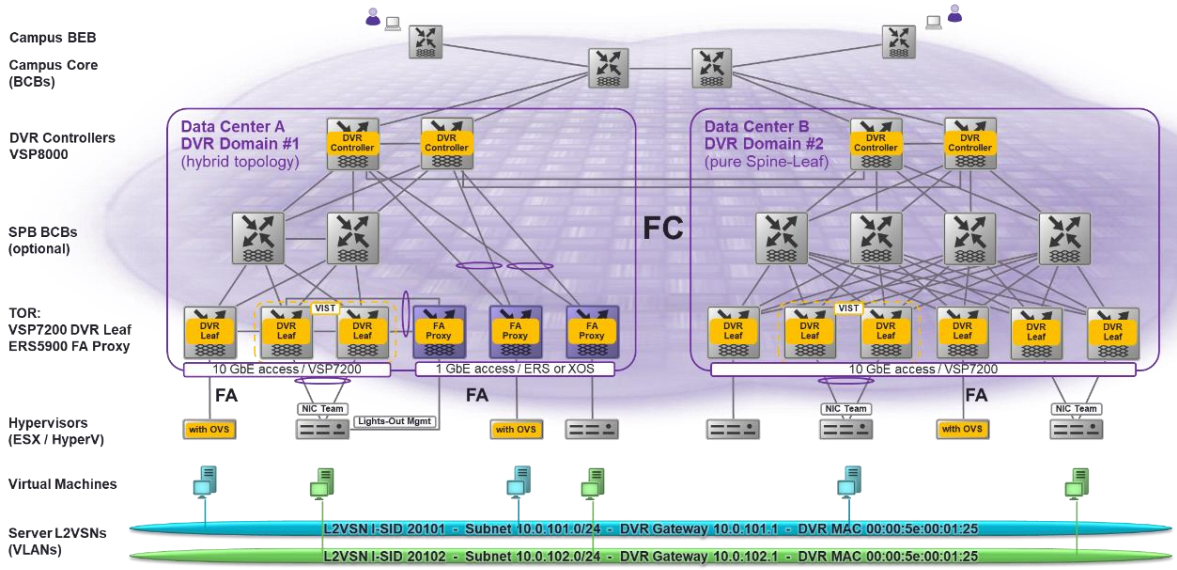


Figure 11 DVR Architecture

Caution

If deploying FA Proxy switches in DVR architectures, these can only be connected either to the DVR controllers or to DVR leaf nodes. In both cases Extreme's vIST SMLT clustering can be leveraged.

From the same diagram above it should be noted that for larger data center designs the DVR architecture can be scaled out into 5-Stage (Three-Tier) topologies by inserting a middle tier of pure BCB devices acting as first spine layer. These BCB nodes do not need to be DVR aware as they will simply transport traffic already MAC-in-MAC encapsulated by the ToRs to their final intended destination along the shortest path, as well as provide equal cost multi-path. Whereas in smaller 3-Stage (Two-Tier) data center designs, the DVR controllers can act as spine nodes.

DVR can be easily activated on top of Fabric Connect and indeed combines IS-IS based signalling with the Fabric Connect I-SID based multicast trees providing a reliable and scalable control plane to disseminate both the anycast gateway across all the DVR leaf nodes and propagate server/VM host IP information within the DVR domain with a single update and a reliable acknowledging mechanism.

Tip

Lightning speed, compared to the EVPN approach with MP-BGP having to push the same update to all meshed peers or via Route Reflectors.

From a management perspective, all L3 IP and VRF configurations are centralized on the DVR controllers, which need to hold a consistent configuration. It is on the DVR controllers that server VLAN segments are created as L2 VSNs, assigned an IP interface belonging to a VRF / "tenant" L3 VSN, and IP Multicast enabled (if need be).

Tip

The superior IP Multicast capabilities of the SPB fabric are in no way compromised by DVR. If anything, they are enhanced as a DVR enabled server segment can be IP Multicast enabled across all DVR leaf nodes by simple configuration on the DVR controllers. By comparison, a data center based on an EVPN IP fabric architecture faces a number of challenges to properly handle IP Multicast applications.

Configuration of the DVR Gateway IP for the segment has similarities with VRRP in that each DVR controller has a unique physical IP interface on the segment and all DVR controllers share the same virtual DVR Gateway IP, which ultimately is configured as default gateway on the servers and VMs residing in the segment. The DVR controllers will then automatically activate the distributed anycast gateway for the L2 segment on all the DVR leaf nodes in the same DVR domain. This will allow the ToR switches to be able to act as default gateways for any connected hosts on the segment.

From a management perspective, the DVR leaf nodes remain purely L2 devices in that the only configuration allowed on them is to assign access ports to server VLAN(s) / L2 VSNs and can ultimately even be offloaded and automated via use of Extreme Management Center ExtremeConnect integration with VMware or Microsoft HyperV or via the use of Fabric Attach on the server hypervisor. This is a very attractive approach for very large data centers, as it would be cumbersome to have to manage the L3 configuration across a large number of ToR switches.

Table 7 – Popular Hypervisor NIC Teaming Hashing Modes

Hypervisor NIC-Teaming mode	VMware ESX Configuration	Microsoft Hyper-V Configuration	SMLT/MLT on ToR	LACP on SMLT
	Route based on originating virtual port	Switch Independent / Hyper-V port	No	n/a
	Route based on IP hash	Switch Dependant / Address Hash	Yes	Yes (ESXi v5.1)
	Route based on source MAC hash	n/a	No	n/a
	Route based on NIC load (Load Based Teaming – LBT)	n/a	No	n/a
	n/a	Switch Independent / Address Hash	No	n/a
	n/a	Switch Dependent / Hyper-V port	Yes	Yes

Server NIC teaming on virtualized server hypervisor has become more complex than it once was. The days when NIC teaming meant that the ToR switches had to do the same link aggregation are long gone. Most hypervisors on the market offer a wide range of hashing schemes to use in conjunction with NIC teaming, and not all of these have a requirement for link aggregation on the ToR switch side. Indeed, some of these schemes are effectively assigning VMs to one of the hypervisor's NICs and expecting all traffic sent and received by the VM to use only that NIC. In these hashing schemes, it is essential that link aggregation (SMLT/MLT) should not be used on the corresponding ToR ports so that returning traffic to the VM will always be delivered on the connection where the VM's MAC was learned. Table 7 gives an overview of some common NIC teaming hashing schemes used with VMware and Microsoft hypervisors.

No scheme is better than the other, but rather each scheme offers advantages and trade-offs based on the number of VMs running on the hypervisor as well as the traffic load that each of those VMs delivers to the network. A vport-based hash scheme is well suited for hypervisors with high number of VMs that are not expected to transmit high traffic loads. An address-based hash is well suited for hypervisors with a low number of VMs that are expected to transmit high levels of bandwidth, which is best to hash across more NICs. VMware's Load Based Teaming provides a popular alternative that is based on vport hashing but where the hypervisor is able to re-allocate VMs to different NICs based on the amount of bandwidth they consume.

The Extreme Networks ToR switch platforms are able to support any of the above-mentioned hashing schemes. It is however recommended to provision the ToR switches in vIST SMLT clustering from the start, so that both link aggregation (SMLT with or without LACP) and non-link aggregation server connectivity can be supported.

Tip

Extreme Networks supports all server hypervisor NIC Teaming schemes.

Finally, the ways in which server or VMs can be attached to their designated server VLAN/segment are illustrated in Figure 12 and fall essentially into three categories:

- Manual configuration of the ToR switch, via CLI or graphical interface. In a DVR architecture, the DVR leaf implements the most flexible Switched-UNI for terminating server L2 VSNs on ToR access ports.
- Via Extreme Management Center ExtremeConnect integration with VMware and Microsoft HyperV. Extreme Management Center is able to discover and track all MAC addresses in the data center via the use of MAC based authentication on the DVR leaf TOR access while at the same time obtaining information on the existence, creation, movement, and assigned Port Group (VLAN id) of all VMs via ExtremeConnect APIs. By combing this knowledge, Extreme Management Center automates the assignment of required server L2 VSNs on the ToR access ports in real time while allowing the network administrator to retain control of the desired level of automation with rule-based criteria.

  Caution

  This functionality is currently only supported on Extreme Fabric Orchestrator.

- Via use of Fabric Attach (802.1Qcj Auto-Attach), if the hypervisor supports Open vSwitch (OVS). The server administrator is now fully in control of requesting the desired L2 I-SID attachments for the VMs needed to deploy.

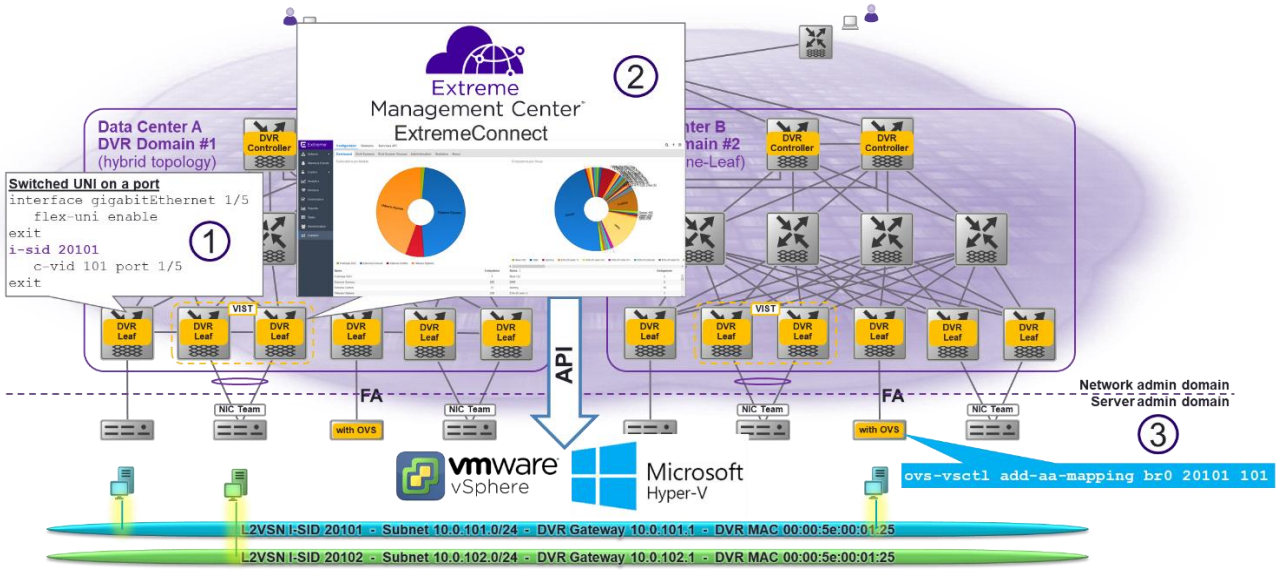


Figure 12 VM Attachment to Server VLAN (L2 VSN)

Architecture Components

This section covers the various components making up an SPB Ethernet fabric as well as the terminology.

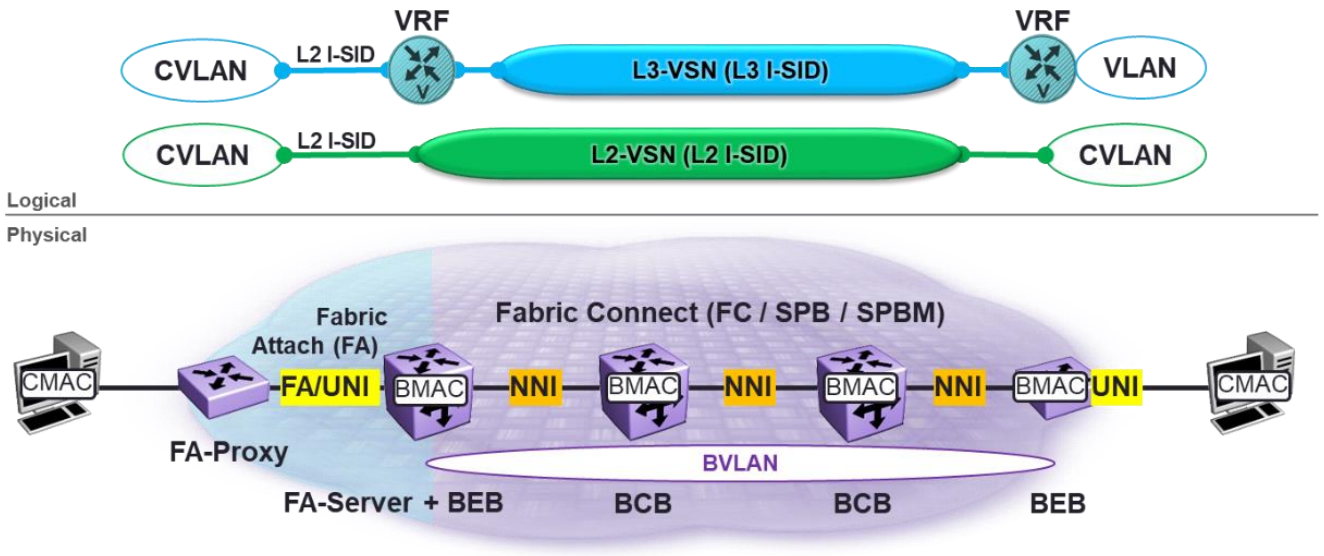


Figure 13 SPB Fabric Architecture Components

User to Network Interface

These are the interfaces at the edge of the SPB fabric where regular Ethernet traffic is sent and received to and from end stations. There is no MAC-in-MAC encapsulation, nor any IS-IS, used on User to Network Interfaces (UNIs). UNIs may be untagged or tagged and will belong to one or more user VLANs, respectively. When traffic enters from an UNI port and needs to be delivered across the SPB Fabric, a Mac-in-Mac encapsulation is added as the packets are switched out of the NNI interface corresponding to the shortest path towards the remote destination.

Network to Network Interface

Network to Network Interfaces (NNIs) are interfaces inside the SPB Ethernet fabric that have been configured for SPB and IS-IS and over which an IS-IS adjacency forms. Once configured, they are automatically turned into q-tagged interfaces and assigned the Backbone VLANs (BVLAN or BVID) with which SPB was globally configured. IS-IS control plane messages are transported in the Primary BVLAN id, while user Mac-in-Mac encapsulated VSN traffic is distributed across all the available BVLANs to achieve load balancing across equal cost fabric paths. An NNI is typically made of an individual Ethernet port, but with Extreme Networks VSP series platforms, can also be made of an MLT bundle comprising two to eight ports that behave as one logical NNI.

Tip

As of VOSS 7.0, it is possible to have more than one IS-IS adjacency between the same pair of SPB nodes (NNI parallel links). In this case, all links will be able to form an IS-IS adjacency but only one of these NNI links will be active and advertised in the IS-IS LSDB topology. The other NNI links will be in a backup state. The SPB interface metric will determine which of the NNI links will be the active one.

This can be useful when separate NNI connections exist between the same pair of SPB nodes using different link speeds (for instance a backup connection via a slower radio link). If instead the connections are running at the same speed, then an NNI MLT interface should be preferred.

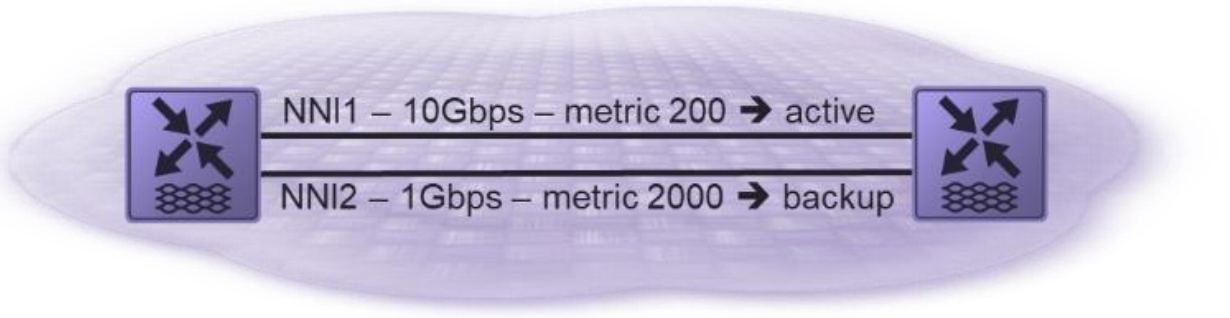


Figure 14 IS-IS NNI Parallel Links

In the Extreme Networks implementation, IS-IS interface authentication is also possible using either simple password authentication or HMAC-MD5 or HMAC-SHA256 authentication. This increases security and ensures that only trusted SPB nodes are allowed to form an IS-IS adjacency in the SPB fabric.

Backbone Core Bridge

A Backbone Core Bridge (BCB) is an SPB node with only NNIs. It is a core node whose role is purely to transport traffic to the destination BMAC along the IS-IS shortest path. As such, a BCB only looks at the outer MAC header of the Mac-in-Mac encapsulated traffic it switches; it does not look at the I-SID information and is completely unaware of the information and addressing within the virtual network (VSN) traffic it transports. There are no VRFs or user-VLANs configured on a BCB node, and therefore the BCB node does not learn any end-station MAC addresses nor does it hold any IP routes for the L3 VSNs it transports. Hence the Ethernet Fabric SPB Core can scale to a virtually unlimited number of services.

Tip

An SPB BCB node is conceptually identical to an MPLS P node.

Typically, a BCB node is provisioned with one single IP interface in the default Global Routing Table (GRT), on a circuit-less loopback, purely for management (e.g., SSH, SNMP) purposes.

Backbone Edge Bridge

A Backbone Edge Bridge (BEB) is an SPB node with both UNI and NNI interfaces. Its role is to terminate virtual networks (VSNs) onto a locally defined VRF (in the case of L3 VSN) or user-VLANs (in the case of L2 VSN). It is the BEB's responsibility to add or remove the Mac-in-Mac encapsulation for traffic entering or leaving the SPB fabric. A BEB must take an active role in the existence of all the VSNs it terminates by holding all the necessary IP routes in the VRFs as well as learning all the user MAC addresses it sees in the user-VLANs defined on it. Naturally a BEB will have a scaling limit with the number of VRFs it supports (and hence the number of L3 VSNs it can terminate), the number of IP routes it can scale to (and hence the number of L3 VSNs and their cumulative number of IP routes), the number of VLANs it supports, and the number of MAC addresses it can learn across those VLANs (and hence L2 VSNs). It is therefore important to select the right platform to meet the required scaling needs. Note however that even if a BEB reaches its VSN scaling limit, this is in no way limiting the wider Ethernet Fabric, as one can always provision additional BEB nodes to scale to more VSNs over the same fabric.

In general, a BEB also performs a BCB function in that it can act as transit node and be part of the shortest path between other SPB BEB nodes (i.e., it can also switch traffic between its NNIs). The BCB function of a BEB node can be manually turned off by enabling the IS-IS overload bit, which will result in IS-IS not computing any shortest paths via this BEB node.

Tip

An SPB BEB node is conceptually identical to an MPLS PE node.

Customer MAC Address

A Customer MAC address (CMAC) is a MAC address that belongs to an end station, server/VM or in general any device that is not running SPB. CMACs only exist within Fabric Connect VSN services and are never seen or used within the SPB Backbone VLANs.

Backbone MAC Address

A Backbone MAC address (BMAC) is a MAC address that belongs to an SPB node in the Ethernet Fabric and only exists inside the SPB fabric (as opposed to end-station user-MAC addresses, which only exist inside user-VLANs or L2 VSNs). In the Mac-in-Mac encapsulation, the outer Ethernet header will contain a Destination BMAC (which determines the egress SPB node, or service-specific multicast tree, across the Fabric for the packet) and a Source BMAC (which indicates which BEB added the Mac-in-Mac encapsulation to the packet when it entered the fabric). Every SPB node allocates one (or more) BMACs and IS-IS announces these BMACs to every other node in the Ethernet fabric so that all nodes can compute the shortest path towards each BMAC in the fabric.

Tip

There are not that many BMACs in an SPB fabric. In the Extreme Networks implementation SPB nodes allocate these BMACs:

- One BMAC for every BEB or BCB node (same as IS-IS System ID).
- One SMLT-Virtual-BMAC shared across both BEBs forming an SMLT cluster.
- One BMAC for the encapsulation of SPB IP Multicast traffic on nodes acting as ingress BEBs for the IP Multicast streams.

Note that with SPB, one of the node's BMAC addresses (the unique globally assigned one) is also the node's IS-IS System ID and in fact is configured as such.

Note

It usually makes sense to configure the IS-IS System ID / node BMAC into an easily recognizable format, such as Locally Administered Address, or LAA. For example, 02yy.yyxx. xxss[4] where the yy.yy bytes can reflect the location of the node in the SPB network while ensuring that xx.xx remains a fully unique number across the SPB network, which can then also be reused to generate a unique SPBM nick-name and ss is used for the SMLT Virtual-BMAC. Alternatively, the burnt in MAC address will be used to derive it.

SMLT-Virtual-BMAC

When two BEB nodes are combined into a Multi-chassis Link Aggregation Group (MLAG), this is done via configuration of Extreme's SMLT clustering, which requires an IST or Virtual-IST (vIST) connection between the two BEBs. This will be further covered in a later section.

The two BEBs forming the SMLT cluster operate as one logical L2 BEB in that they can share SMLT connections and share and synchronize their respective L2 VSN MAC tables. To be seen as one single logical BEB by the rest of the SPB fabric, the SMLT cluster will allocate a unique SMLT-Virtual-BMAC that will be used by both BEBs as source BMAC, instead of the nodal BMAC, to encapsulate any CVLAN traffic received from local UNI ports and Mac-in-Mac encapsulated to cross the SPB fabric. When this traffic is received by distant egress BEBs, reverse MAC learning will ensure that those distant BEBs will associate the source CMACs with the SMLT-Virtual-BMAC of the SMLT cluster where those CMACs are located. This will

4 For guidelines on allocating SPB unique identifiers refer to reference documentation [8].

then ensure that equal cost shortest path load balancing can be leveraged across the SPB Fabric to reach the SMLT cluster BEB nodes.

Tip

The SMLT-Virtual-BMAC can be either auto generated or manually provisioned. If auto provisioned, the system will simply use the SMLT's Primary nodal BMAC with the least significant byte set to 0x80.

Tip

Extreme recommends that the SMLT-Virtual-BMAC configuration should follow the same criteria used for nodal BMAC / IS-IS System ID. If the latter is manually provisioned, then so should the SMLT-Virtual-BMAC. If instead auto configuration is used, then both should be allocated in that manner.

IS-IS Area

SPB operates in unison with IS-IS, which acts as a link state routing protocol for the Ethernet Fabric. IS-IS needs to operate within a given IS-IS area and as such every SPB BEB or BCB node will need to be configured with the same IS-IS area.

Note

The SPB standard only defines operation over a single IS-IS area.

Tip

Extreme Networks will support Multi-Area Fabric Connect in future. When Multi-Area becomes available the IS-IS Area configuration will be leveraged by new functionality on Multi-Area BEB nodes to interconnect services across different areas.

Note

The IS-IS area format is <AFI>.<AreaID>, where AFI is two hex digits and Area ID is four hex digits. Extreme recommends to select SPB IS-IS area IDs 49.xxxx[5] where the AFI = 49 indicates locally administered NSAP addresses.

Tip

Extreme Networks VSP platforms can automatically assign the IS-IS area if deployed in Zero Touch Fabric (ZTF) mode and are connected to an already existing SPB fabric.

IS-IS System ID

Like every IS-IS router, a BEB or BCB node must be allocated a unique IS-IS System ID. In the Extreme Networks SPB implementation, the IS-IS System ID (six octets) also becomes the nodal BMAC of the SPB node. Refer back to Backbone MAC Address on page 41.

5 For guidelines on allocating SPB unique identifiers refer to reference documentation [8].

IS-IS Overload Function

The Overload bit is special in the IS-IS LSP and used to inform the network that the advertising router is not yet ready to forward transit traffic. The IS-IS Overload bit can be automatically set on SPB nodes during their boot up phase, to ensure that they are not considered by other fabric nodes when computing Shortest Path First (SPF) path calculations until after they have completed their initialization.

The same IS-IS Overload function can also be manually activated at any time and is a simple and handy technique to isolate a specific SPB node in the network before any maintenance work is performed on it.

Tip

Extreme Networks VSP and ERS platforms both allow manually setting IS-IS Overload as well as on startup by specifying a delay time.

Note

Note that the IS-IS Overload bit only has effect on SPB "transit" traffic; that is, traffic that would be switched between NNI ports (e.g., on a BCB node). Traffic destined to a SPB (BEB) node will still lead to that node even if the IS-IS Overload bit is set on it.

SPB Bridge ID

The Bridge ID of an SPB node is defined as the concatenation of the node's Bridge Priority (2 bytes) and the node's IS-IS System ID (BMAC). The SPB Bridge ID is used when evaluating equal shortest paths using the Equal Cost Tree (ECT) Algorithms.

Caution

The SPB Bridge Priority field is not currently configurable on Extreme Networks VSP platforms and is hard coded to 00:00.

SPBM Nick-name

More commonly referred to as the SPB node's nick-name, but referred to as the Shortest Path Source Identifier (SPSourceID) in the IEEE standard, the nick-name is a 20-bit (two and a half octets) identifier that must be unique to every BEB/BCB node in the SPB Fabric.

Note

In the Extreme implementation of SPBM, the nick-name can be auto-generated or else must be manually provisioned. If provisioned, Extreme recommends to allocate the nick-name in the format of 0.xx.xx[5] where xx.xx has the same value from the IS-IS System ID 02yy.yyxx.xxss. If the IS-IS System ID was auto-generated, then use a format 1.zz.zz, where zz.zz must be unique across the SPB fabric.

The nick-name comes into play when a fabric service needs to leverage a service-specific I-SID multicast tree. Every I-SID multicast tree is rooted at one particular BEB node and once the shortest path tree is computed by IS-IS, there is a need to program such tree into the data plane of all the SPB nodes forming such tree: the root BEB, the receiving leaf BEBs, and any transit BCB node which happens to be along the shortest path forming such tree. Every I-SID multicast tree translates into a fabric-wide unique multicast BMAC that is simply programmed by IS-IS into the BVLAN MAC table.

The multicast BMAC is formed by combining the root BEB's nick-name (20 bits) with the I-SID (24 bits), plus setting of the MAC address Locally Administered Address (LAA) and Multicast bits.

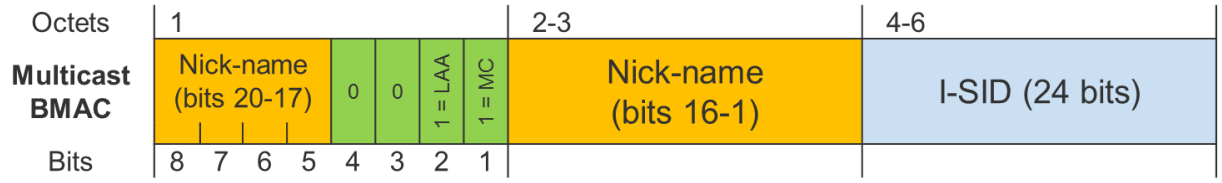


Figure 15 How the SPBM Nick-name is Used to Construct a Multicast BMAC

Another use of the SPBM nick-name is, when configuring IS-IS IP Accept policies, it is possible to single out IP routes advertised by specific BEB nodes by specifying the BEB's nick-name. This provides a similar approach to OSPF Accept policies where OSPF Router IDs are specified.

Dynamic Nick-name Assignment

Dynamic Nick-name assignment (DNN) is the ability for newly provisioned SPB platforms to automatically obtain an SPB Nick-name if deployed in Zero Touch Fabric (ZTF) mode into an already existing SPB fabric. The existing SPB fabric needs to be already deployed with one or more SPB core nodes acting as nick-name servers that can allocate nick-name ids from one of six possible ranges. The newly provisioned node connects to the reserved Fabric Area Network (FAN) as a nick-name client and obtains a valid nick-name id from one of the available nick-name servers.

Note

Dynamic Nick-name Assignment is currently only supported on Extreme Networks VSP platforms.

Note

The six valid Dynamic Nick-name ranges available are:

- Range A: A.00.0A – A.FF.FE (default range)
- Range B: B.00.0A – B.FF.FE
- Range C: C.00.0A – C.FF.FE
- Range D: D.00.0A – D.FF.FE
- Range E: E.00.0A – E.FF.FE
- Range F: F.00.0A – F.FF.FE

Customer VLAN

The Customer VLAN (CVLAN) is a regular VLAN where end-stations and users are located and has all the traditional VLAN properties that the reader should be familiar with. A CVLAN does MAC learning, belongs to a Spanning Tree instance and floods Broadcast, Unknown unicasts, and non-snooped Multicast (BUM) traffic. In the Extreme Fabric Connect architecture, the CVLAN is augmented with the ability to be assigned to an L2 VSN service I-SID, in which case its L2 broadcast domain is extended across the fabric to reach other CVLANs located on distant BEBs. In this case, the CVLAN MAC table is also able to do reverse MAC learning of traffic arriving from the fabric and maintain entries mapping distant CMACs against the distant BMAC from which they can be attained.

Backbone VLAN

The Backbone VLAN (BVLAN or BVID) is a special VLAN that exists only on SPB NNI ports and where Ethernet bridging works in a completely novel way. There is no MAC learning performed in a BVLAN and there is no more flooding of broadcast, multicast, and unknown-unicast packets. SPB forwarding database entries are instead derived from the node and service identifiers contained within SPB's IS-IS Link State

Database. Also, a BVLAN takes no notice of the Spanning Tree state of member ports (e.g., if Spanning Tree was for some reason enabled on NNIs).

The unicast MAC table of a BVLAN will contain every BMAC in the Ethernet fabric and these entries will point to the local NNI port that corresponds to the IS-IS computed shortest path to reach that BMAC across the fabric and from which all traffic from that BMAC is expected to arrive. Indeed, every packet received on NNIs is subjected to Reverse Path Forwarding Check (RPFC), whereby only packets that are received on the same port that corresponds to the shortest path to reach the Source BMAC of the packet are accepted. RPFC is how SPB ensures loop suppression within the fabric.

The BVLAN also populates a multicast MAC table that contains every I-SID service-specific corresponding multicast MAC for I-SIDs that terminate or transit on the local SPB node. Every multicast MAC entry in this table will map to one or more egress ports in such a way as to efficiently replicate packets along the service-specific shortest-path multicast trees.

Both the unicast and multicast MAC tables contained in the BVLAN are programmed by IS-IS and only need changing when either physical infrastructure changes (link or node failure; new NNI link or SPB node provisioned, etc.) or a new I-SID service is provisioned (in the case of the multicast MAC table).

Tip

The combination of SPB BVLAN and IS-IS has the same desirable attributes and is conceptually similar to the IP routing table with OSPF. Just like OSPF programs the IP routing table with the results of its Shortest Path First (SPF) runs, so IS-IS programs the BVLAN MAC table with the results of its SPF runs. The BVLAN can be thought of as just a MAC table repository of IS-IS calculated shortest paths.

There can be more than one BVLAN in an SPBM fabric. One very important property of SPB, is that the IS-IS computed paths within a BVLAN are always congruent. This means that, in the presence of equal cost shortest paths, if a given path is chosen for traffic in one direction, the very same path will be taken in the reverse direction in the same BVLAN.

Tip

This property is important for Ethernet OAM (802.1ag CFM) as it allows Ethernet-based path tracing (traceroute) to follow the same path as VSN traffic for troubleshooting purposes as well as jitter and latency measurement and reporting.

An SPB fabric provisioned with one single BVLAN is thus only able to hold one shortest path if multiple equal cost shortest paths exist, whereas an SPB fabric provisioned with more BVLANs will be able to hold as many equal cost paths as it has BVLANs. Virtual networks (VSNs), or flows within them, are then assigned to one or another BVLAN in a deterministic manner.

Note

The SPB 802.1aq standard defines a maximum number of 16 BVLANs.

In most Enterprise network topologies, the use of two BVLANs offers the best compromise for leveraging equal cost multipath. Currently, Extreme's implementation supports two BVLANs. Load balancing is covered in greater detail in a separate section.

Virtual Services Networks

Virtual Services Networks (VSNs) represent the logical networks which can be created as a service on top of an SPB Ethernet Fabric. End-stations and users only exist within VSNs and therefore end-station MAC addresses, ARP entries, and IP routes are only significant within the VSN to which they belong.

Furthermore, the resources to store those forwarding records are only consumed on the BEB nodes where the VSN services terminate.

A VSN can also be thought of as a VPN; however, because VPN often has a connotation of a point-to-point service type and often implies encryption (often on software based routers), Extreme prefers the term VSN because it naturally reflects an any-to-any service type and it is entirely implemented on Ethernet switching platforms.

A VSN in the SPB architecture can either be an L2 VSN (which gives a VLAN type service across the fabric) or L3 VSN (which gives a VRF type service across the fabric). In both cases a VSN comes into existence the moment two or more BEBs allocate the same I-SID for the same service.

I-SID

The I-SID is defined in the IEEE standard as the SPBM Service ID for the I-Component. The I-SID is a 24-bit numerical service ID which uniquely identifies every service type available in Fabric Connect.

Tip

The 24-bit I-SID can take values between 1 and 16777215; hence the claim that SPB and Fabric Connect can theoretically scale up to 16 million services.

In the Extreme Fabric Connect implementation, I-SIDs can be user-assigned to L2 VSN services and/or L3 VSN services alike and must also be user-allocated when provisioning Virtual-IST (vIST) SMLT clusters. Fabric Connect reserves all I-SIDs above 16 million that are dynamically allocated and used for IP Multicast stream delivery trees as well as for multicast based signalling with some advanced Fabric Connect features such as DVR and PIM Gateway.

Tip

It is worth defining an I-SID allocation scheme to easily identify I-SID values by type, so that for example L3 VSN I-SIDs use an easily recognisable value range from L2 VSN I-SIDs. A section with guidelines and a suggested allocation scheme is available in reference documentation [8].

Inter-VSN Routing

An L2 VSN is an SPB virtualized Layer 2 segment. When an L2 VSN is terminated on an SPB node and an IP interface is assigned to the VLAN that is used to terminate the L2 VSN, we can IP route traffic to and from that L2 VSN segment.

Tip

The IP interface can be bound to any VRF and therefore the L2 VSN segment can be made to belong to any L3 VSN routing domain (or GRT IP Shortcuts).

In short, the node becomes a default gateway for any device located in the L2 VSN. When the node terminates two or more L2 VSNs in this manner, we can also perform IP routing between these L2 VSNs and this capability is referred to as Inter-VSN routing.

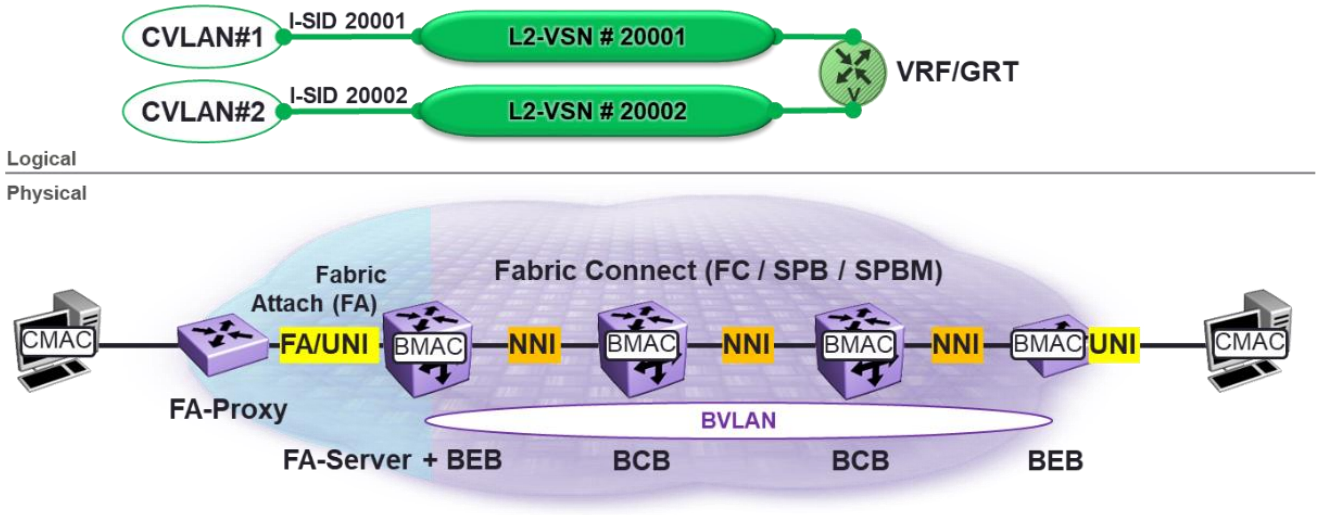


Figure 16 SPB Fabric Inter-VSN Routing

Conceptually this is no different from a switch that IP routes between VLANs where an IP address was configured on each VLAN (in the same VRF). However, with L2 VSN this implies that the SPB node must be capable to handle (de-capsulate and then re-encapsulate) the Mac-in-Mac encapsulation before and after performing the IP routing function.

Tip

The Extreme Networks VSP series of SPB platforms, which support L3 VSNs and IP-Shortcuts, can do Inter-VSN routing for both IPv4 and IPv6.

Tip

Performing IP routing between different VPLS VSI segments is not common, within the MPLS backbone; however, EVPN VXLAN implementations can perform IP routing between different VXLAN VNIs on the VTEP.

Note

Note that just because it is possible to IP route L2 VSN segments just about anywhere in the SPB fabric, this should not lead to the temptation of IP routing all user L2 segments onto the same centrally located SPB node(s). A good design will distribute most IP routing across the network and locate the routing instances where it makes most sense, for example, where those user IP subnets are located. This will ensure a more robust architecture and where the host routes / ARP tables are distributed across all the IP routing instances for the L3 IP routing domain (L3 VSN or IP Shortcuts).

Fabric Area Network

The Fabric Area Network (FAN) is a reserved L2 domain I-SID used by SPB nodes for enhanced fabric discovery and auto-configuration functionality, such as the Dynamic Nick-name (DNN).

Tip

Reserved I-SID 16777001 is used and membership of the Fabric Area Network is announced in TLV 147.

Fabric Attach / Auto-Attach

Fabric Attach (FA) extends the Fabric Connect VSN service I-SID all the way to the end user or IoT device in the campus access as well as to the application in the data center onto devices for which it would make no sense to implement Fabric Connect and IS-IS (like wireless APs, server hypervisors, etc.).

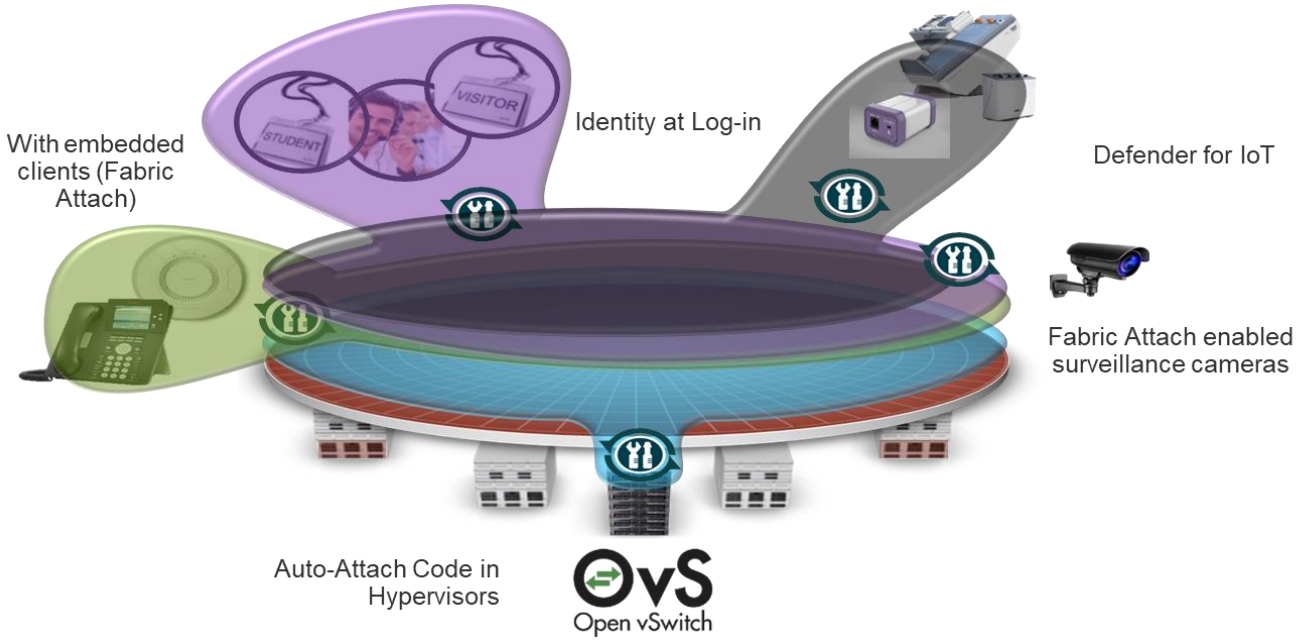


Figure 17 Fabric Attach Foundation of Elastic Campus Architecture

Fabric Attach encompasses a range of mechanisms and operational modes which allow for all these use cases:

- Ability of FA Client devices to self-request their attachment VSN I-SIDs.
- Ability to automatically onboard FA Client devices based on their type or application.
- Ability to automatically extend as well as retract the service attachment to any client type (wired or wireless user, IoT, FA Client, application VM) and thus creating the elastic nature of the Extreme Fabric Connect solution.
- Ability to augment the security of the network against attackers via FA message authentication to increase the security of MAC-based authentication of IoT devices.
- Operation of Ethernet access switch FA Proxy mode (as an alternative to direct Fabric Connect access).

Fabric Attach was initially developed by Extreme Networks and has since been submitted as IEEE standard 802.1Qcj where it is named Auto-Attach and has been implemented in Open vSwitch (OVS)[6].

The Fabric Attach architecture defines a number of different device roles: FA Server, FA Proxy, FA Standalone Proxy, and FA Client. These are further covered in the following sections as well as in

6 IEEE 802.1Qcj Automatic Attachment to Provider Backbone Bridges (PBB). Auto-Attach code has been added to Open vSwitch (OVS) since version 2.4. OVS can be deployed in KVM, Xen and Hyper-V hypervisors to massively simplify VM mobility in the data center

Fabric Attach on page 77.

FA Server

FA Servers are Extreme Networks switches that connect directly to the SPB Fabric Connect core network and thus act as BEB nodes. FA Client devices can be connected directly to the FA Server or via an FA Proxy switch. In both cases, the FA Server can accept FA Client or FA Proxy service binding requests to attach the user/IoT/application to a VLAN:I-SID pair. The FA Server then dynamically attaches the VLAN to the Fabric Connect L2 I-SID.

FA Server can be deployed as a single switch (typically in topologies where Fabric Connect extends to the access switch) as well as with SMLT clustering where the vIST peers logically act as one redundant and logical FA server (typically in topologies where the fabric access is implemented on FA Proxy switches redundantly connected to the FA Server SMLT cluster distribution).

Note

The Extreme VOSS VSP platforms can support FA Server in both single switch and SMLT configurations.

The Extreme ERS platforms can only support FA Server in single switch mode (both in stacking and standalone modes).

FA Client

FA Clients are devices with ability to communicate desired fabric configuration parameters to the FA Server. These parameters include:

- FA Client type: currently can be any of those listed in Table 8.
- Ethernet port tagging mode: whether the FA Client expects to transmit all traffic tagged, untagged or both tagged and untagged.
- VLAN:I-SID service bindings: one or more L2 VSN I-SID bindings which the FA Client will need to communicate with over the network.

FA clients can be either directly connected to the FA Server or they can be connected via an FA Proxy switch. Not all FA Clients types will need to request VLAN:I-SID service bindings.

Table 8 – Available FA Client Types

Id	Type Name	Description
6	Wap-type1	Wireless Access Point with wireless user traffic switched locally on the attachment switch
7	Wap-type2	Wireless Access Point where all wireless user traffic is tunnelled to a central controller
8	Switch	Ethernet switch with basic FA Client functionality
9	Router	IP router
10	Phone	IP phone
11	Camera	IP surveillance camera
12	Video	Video monitoring device
13	Security-dev	Security device
14	Virtual-switch	Virtual Switch (OVS)
15	Server-endpoint	Server endpoint

16	ONA-SDN	Open Network Adapter in SDN/Surge solution
17	ONA-spb-over-ip	Open Network Adapter used on VSP4000 for Fabric Extend

FA Proxy

A FA Proxy is an Ethernet switch which allows Fabric Connect L2 I-SID services to be extended to locally attached clients. The FA Proxy switch is able to proxy for FA Client VLAN:I-SID bindings requests towards the FA Server and to automatically provision the data plane VLAN on its uplink to the FA Server as well as on the access port where the FA Client is connected. The FA Proxy can also initiate itself VLAN:I-SID bindings towards the FA Server when manually configured (as if it was a Fabric Connect BEB) or when clients are EAP/NEAP authenticated and authorized onto the network via Fabric Attach RADIUS VLAN:I-SID attributes.

Use of FA Proxy is of interest in the campus wiring closet where only L2 switching is required and it allows the use of more cost-effective switching platforms that would either not be able to support SPB's Mac-in-Mac encapsulation with Fabric Connect or would be detrimental towards the maximum fabric network scaling size which is dictated by the number of nodes running SPB/IS-IS.

Caution

A FA Proxy switch can only be connected to one logical FA Server which can be implemented as either a single switch or a SMLT cluster.

Extreme Networks does not support FA Proxy chaining whereby a FA Proxy switch is connected behind another FA Proxy switch.

Tip

Both the Extreme ERS and ExtremeXOS platforms support FA Proxy

FA Standalone Proxy

FA Standalone Proxy is a hybrid mode where the FA Proxy switch can operate without the presence of an FA Server. This mode of operation is only useful in situations where the wiring closet access switch is perhaps deployed in a non-fabric architecture or, more in general, in cases where the distribution layer is not capable of providing the FA Server functionality on Extreme legacy SPB platforms.

The FA Standalone Proxy switch can perform all the functions of a FA Proxy switch in terms of Fabric Attach signalling to attached FA Clients and can also accept Fabric Attach RADIUS attributes for NAC.

The FA Standalone Proxy differs from the FA Proxy switch in that it cannot automatically infer its uplinks into the network via the presence of a FA Server and will thus need to be statically configured as to which ports to use as uplinks. The FA Standalone Proxy also supports VLAN:I-SID bindings, except that the I-SID component of these bindings must always be 0. Any VLAN:I-SID bindings requested will simply result in the VLAN component being assigned to the static uplink ports as well as to the relevant client access port.

Note

Only the Extreme Networks ERS platforms currently support FA Standalone Proxy mode

VPN Routing and Forwarding Instance

The VRF original acronym (VPN Routing and Forwarding Instance) originates from MPLS-based IPVPNs, but a more general acronym is Virtual Routing and Forwarding instance. A VRF is a virtualized IP routing table that allows multiple instances of IP routing tables to exist on the same physical router. In the context

of IPVPNs or L3 VSNs, a VRF is thus a repository of all the IP routes that belong to a given virtualized routing domain.

A VRF consists of an IP routing table, a forwarding table, and interfaces assigned to it. Common routing protocols such as OSPF, RIP, BGP, and IS-IS can be used to advertise and learn routes within the VRFs. In a scalable virtualized multi-tenant architecture one single backbone routing protocol is used to advertise and learn routes across all VRFs within their respective virtual domain (IPVPN or L3 VSN).

Tip

In an SPB fabric this role is performed by IS-IS using I-SID-IPv4 (or I-SID-IPv6) routes, which then associate those IP routes with the respective L3 VSN I-SID.

In an MPLS backbone, this role is performed by MP-BGP using VPN-IPv4 (or VPN-IPv6) routes over iBGP peerings which have to render those IP routes unique to BGP via addition of the Route Distinguisher (RD) and specific to the VPN-id via addition of Export Route Targets (RT).

A proper implementation of VRFs must be able to support overlapping IP routes across different VRFs/L3 VSNs, which can be essential when integrating networking infrastructure from different entities following mergers and acquisitions.

Note

A given router will support a maximum number of VRFs and a maximum number of IP routes which can be held across all VRFs as well as in the non-virtualized GRT.

Tip

In the Extreme Networks implementation of Fabric Connect using SPB, a VRF can be software limited to a defined maximum number of IP routes such that it cannot consume all available record entries at the expense of other VRFs operating on the same router.

Global Router Table

The Global Router Table (GRT) constitutes the non-virtualized IP routing instance that all routers operate in by default. It is often also referred to as VRF-0, to differentiate it from VRFs that are numbered starting from VRF-1.

In an SPB fabric there are no underlying dependencies on the GRT, and the IGP used within it, and it is conceptually treated just like any other VRF. IP routing is seen as a service provided by SPB, whether it is performed inside the GRT or a VRF.

This contrasts with MPLS, where the GRT and its IGP take on a foundational role without which neither MPLS's label distribution protocols (LDP or RSVP) nor BGP (with its full mesh of iBGP peerings) would be able to operate. While this is never an issue in carrier networks (where the GRT and its IGP is restricted to the MPLS backbone alone), it becomes less ideal in Enterprise networks where some or most user traffic naturally resides in the Global Routing Table.

There are however, even with SPB, some important properties that distinguish the GRT from VRFs. In the Extreme Networks SPB implementation, the GRT can equally be extended above the Ethernet fabric (as a service); however, this is done via explicit configuration rather than via assigning an I-SID and the resulting domain is referred to as GRT IP Shortcut Routing (rather than as an L3 VSN).

Also in the Extreme Networks product implementation, all in-band management traffic (Telnet, SSH, SNMP, etc.) is only processed if received on an IP interface belonging to the GRT. So in an SPB fabric deployment, the IP Shortcuts are always enabled for management purposes.

Distributed Virtual Routing

Distributed Virtual Routing (DVR) is an extension to Fabric Connect that allows the creation, on DVR controller nodes, of anycast default gateway IPs for fabric L2 VSN segments and their distribution across the fabric access layer on DVR leaf nodes. This functionality is aimed at the data center environment where it allows the definition of L3 multitenancy VRF and IP address gateways on centralized DVR controller nodes (acting as data center core layer, or spine layer or border edge), but the data plane is automatically activated on the DVR leaf nodes, which are the ToR access switches.

DVR controllers and DVR leaf nodes operate within a DVR domain construct where all L2 VSN segment IP gateways are distributed and all DVR nodes are aware and keep track of every host IP within that domain. DVR effectively performs host-based IP routing (IP host route to SPB BMAC mapping) within the domain and thus ensures that under every circumstance IP flows will follow the SPB shortest path, even if those host IPs are mobile (as happens in data centers when VMs are vmotion-ed).

DVR Domain

A DVR domain consists of one or more DVR controller nodes and many DVR leaf nodes. For redundancy purposes, a minimum of two DVR controllers should exist in the DVR domain.

All DVR nodes within the same DVR domain communicate via highly efficient point-multipoint IS-IS control plane I-SID based signalling which allows the DVR controllers to push down VRF & IP layer data plane configuration to all DVR leaf nodes via single updates. The same signalling mechanism is used by any DVR node (controller and leaf nodes alike) to update every other DVR node in the domain about the location or movement of a host IP address. Hence all DVR nodes in the same DVR domain have knowledge of all the host IPs locally connected to it and can perform host based IP routing over SPB.

The DVR domain construct is a construct that allows DVR to scale even to large data centers. Smaller data centers will typically be implemented as a single DVR domain while larger data centers can deploy separate DVR domains within data center pods. L2 VSN segments can span multiple DVR domains and in that case all DVR controllers of those DVR domains can have the same DVR Gateway IP on the same L2 VSN segment. This provides an anycast default gateway for the segment across multiple DVR domains.

Note

Currently DVR can scale up to 250 leaf nodes and up to 40000 Host IPs.

DVR Controller

The DVR controller is where IP and VRF configuration is applied to user L2 VSN segments within a DVR domain. The DVR controller then disseminates such configuration across all DVR leaf nodes within the DVR domain using highly efficient point-multipoint IS-IS control plane I-SID based signalling.

The DVR controller is an L3 BEB with full functionalities and can even have users/servers directly attached to it (as if it was a DVR leaf). Depending on the data center physical design, the DVR controllers are typically positioned either as data center core nodes or spine nodes or border edge nodes. It is worth keeping in mind that all IP routed traffic within the DVR domain and destined to an IP address, which does exists as a host IP in that domain, will be load-balanced towards the nearest DVR controllers. In other words, the DVR controllers will always act as point of exit (and thus entry as well) for all traffic leaving or entering the DVR domain and should therefore be placed in the physical topology in such a way to always ensure shortest path forwarding.

Where multiple DVR domains exist, the DVR controllers will also maintain a DVR Backbone across all DVR domains. Communication over the DVR Backbone uses a separate instance of the highly efficient point-multipoint IS-IS control plane I-SID based signalling. DVR controllers use the DVR Backbone to advertise

reachability of DVR host IP routes across different DVR domains as well as to campus BEBs that may have been activated to join the same DVR Backbone.

DVR Leaf

The DVR leaf is typically deployed as a data center ToR Fabric Connect switch. It is a switch with a dual personality. From a provisioning and management perspective, the DVR leaf is an L2 BEB where the only configuration possible is to associate server access ports to L2 VSN segments. Though even this configuration can be eliminated via the use of Extreme Management Center ExtremeConnect integration with VMware and Microsoft HyperV or the use of Fabric Attach where the server hypervisor supports Open vSwitch (OVS).

Note

A DVR leaf node only supports Switched-UNI on its access ports. Platform VLANs and hence CVLAN UNI cannot be configured on a DVR leaf.

However, from a data plane perspective, the DVR leaf is a VRF-aware IP router capable of acting as a first-hop default gateway for any server or station locally attached to it. The DVR leaf has no IP address to do this, but it does have shared ownership of the DVR Gateway MAC address which was configured on the DVR controller(s) within the same DVR domain. This is how the DVR anycast gateway is implemented.

Tip

There are only three possible IP addresses that can be configured on a DVR leaf:

- IS-IS management IP: used for in-band management of the node.
- Out-of-band IP: used for out-of-band management of the node.
- vIST IP: used if the node is forming an SMLT cluster with another DVR leaf node.

Whenever a new VRF and DVR IP Gateway are provisioned on the DVR controllers, all the DVR leaf nodes within the same DVR domain will be instructed to allocate a VRF and MAC routing interface to match. If SPB Multicast was enabled on the DVR IP Gateway created on the DVR controller, then the DVR leaf will also automatically enable its interface for SPB Multicast.

From a data plane IP routing perspective, the DVR leaf has IP host routes for every server host IP connected to the DVR domain, which allows it to perform shortest path host-based IP routing within any of the L3 SPB service types (IP Shortcuts and L3 VSNs). When servers or VMs ARP for their default gateway (the DVR Gateway IP) the locally attached DVR leaf will respond by providing the DVR Gateway MAC. The DVR leaf nodes are thus immediately aware of the host IPs that are locally attached to them and they advertise these host IP routes to every other DVR node (controller and leaf) within the DVR domain using the highly efficient point-multipoint IS-IS control plane I-SID based signalling. The DVR leafs are also able to handle hypervisor GARP and RARP messages to detect when a host VM IP has moved which allows DVR to constantly keep track of host IPs even when they are mobile.

Note

VMware Vmotion uses RARP packets. Microsoft Hyper-V uses GARP packets. In both cases, it is the hypervisor, not the VMs involved, that generates those packets in order to facilitate the VM move.

DVR Backbone

The DVR Backbone is a dedicated point-multipoint IS-IS control plane I-SID based signalling tree that allows all DVR controllers belonging to different DVR domains to communicate DVR host IP routes. This allows DVR controllers, for their L2 VSN interfaces configured for DVR, to have full visibility of all host IP routes for the L2 VSN segment including host routes which are located in different DVR domains.

It is also possible to let campus L3 BEBs join the same DVR Backbone to get control plane visibility of all available data center host routes. In some architectures with dual data centers where each data center uses separate DVR domains, it is possible to create a redistribution policy on the campus BEB to install selected DVR host routes into the BEB data plane IP routing table in order to solve north-south traffic tromboning problems.

Zero Touch Fabric (ZTF)

Zero Touch Fabric is the ability to deploy a new switch in an existing SPB fabric and for that switch to automatically derive the necessary SPB parameters, such IS-IS area and SPBM nickname, directly from the already operating fabric network.

Tip

Zero Touch Fabric is supported on Extreme VSP platforms.

Foundations for the Service Enabled Fabric

SPB Service primitives

SPB uses IS-IS as its link state routing protocol. IS-IS was originally developed as the routing protocol for the ISO Connectionless Network Protocol CNLP and was later extended to operate with IP around the same time OSPF was first defined. While most enterprises are more used to using OSPF, IS-IS has always been widely used by service providers and in many respects, is considered to be more scalable and extensible than OSPF. When it comes to Ethernet fabrics, IS-IS is the perfect choice as it operates directly onto a Layer 2 Ethernet encapsulation (as opposed to OSPF, which operates on top of IP).

The SPB 802.1aq standard does not alter the way IS-IS operates, but rather defines new IS-IS Type-Length-Values (TLVs) to carry SPB BVLAN, BMAC and I-SID service related information.

In an SPB Ethernet fabric, every node (BEB or BCB) has IS-IS configured globally and on each and every NNI. The SPB BVLANs are also part of the global configuration.

This is enough for SPB to form IS-IS adjacencies and to become an Ethernet fabric where every node has computed the IS-IS shortest path towards every other node in the fabric. At this point the Ethernet OAM can be leveraged to perform L2pings or L2traceroutes in any of the SPB BVLANs to verify and troubleshoot connectivity across the fabric. Note that this can be done even before configuring a single IP address or user-VLAN or virtual network (VSN).

Currently SPB only supports a single IS-IS area and point-to-point IS-IS interfaces. This means that every NNI interconnect must be a direct Ethernet connection or a direct LAG/MLT connection or must be seen as an L2 point-to-point circuit, if transported over some WAN cloud. Use of the Mac-in-Mac encapsulation implies the use of oversized Ethernet frames.

Note

The Mac-in-Mac encapsulation adds 22 bytes to a regular Ethernet frame, which means that a native Ethernet packet with an 802.1Q tag is 1522 bytes and, once the Mac-in-Mac encapsulation is applied, becomes 1544 bytes.

Tip

All Extreme Networks SPB-capable platforms natively support Ethernet oversize frames and can all also operate with Jumbo MTU support in conjunction with SPB.

The SPB fabric offers two service primitives that, combined, produce all the rich VSN service types. The first primitive is equivalent to what OSPF does for IP. Every node in the SPB fabric shares the same IS-IS LSDB. From this, every node performs an SPF run to compute the shortest path to every other node (BMAC) in the fabric. The result is programmed in the BVLAN unicast MAC table (just like OSPF programs IP routes in the IP Routing table).

Tip

In the event of multiple equal cost shortest paths and if the Fabric was deployed with more than one BVLAN, each BVLAN will be programmed with one of the available equal cost shortest paths.

SPB adds to this capability a congruence property whereby, in the presence of equal cost shortest paths between a pair of nodes, both nodes will select the same path in both directions within the same BVLAN (which is not true with OSPF in the presence of ECMP paths). This is an important property that comes in to

play when leveraging 802.1ag (CFM), as it ensures that when injecting Ethernet OAM packets along the path, on a given BVLAN, the responses will come back on the same path.

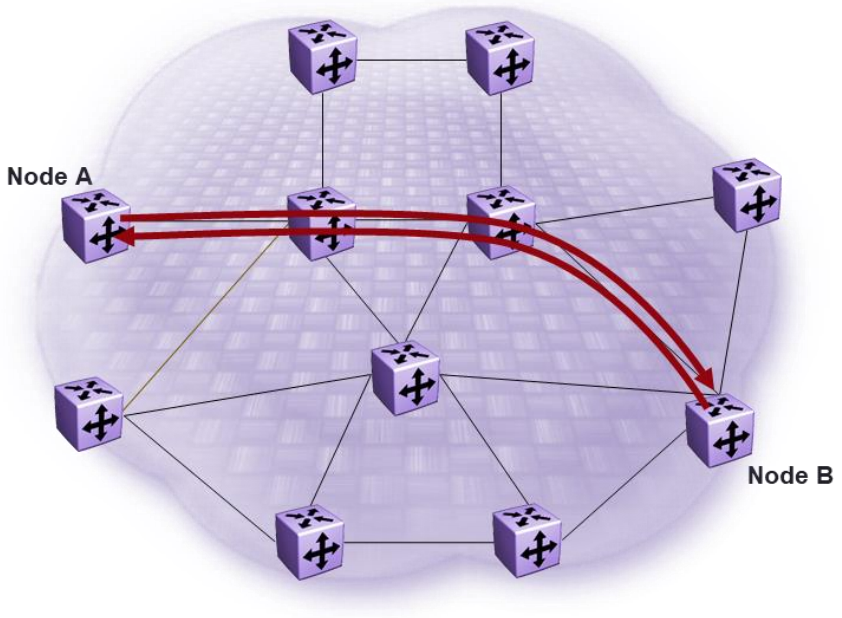


Figure 18 Unicast Shortest Path Between Two Nodes; Forward & Reverse Congruent

The second primitive is what makes SPB so much better for multicast than any other networking protocol to date. For every I-SID, announced by each BEB in the fabric, for which the ingress BEB has set the TX bit in its TLV and one or more egress BEBs have also set the RX bit in their respective TLVs, then all the nodes in the fabric (including transit BCB nodes) will compute the shortest path multicast tree for that I-SID, and if (and only if) they are part of that multicast tree, then they will install a corresponding multicast BMAC in their respective BVLAN multicast table.

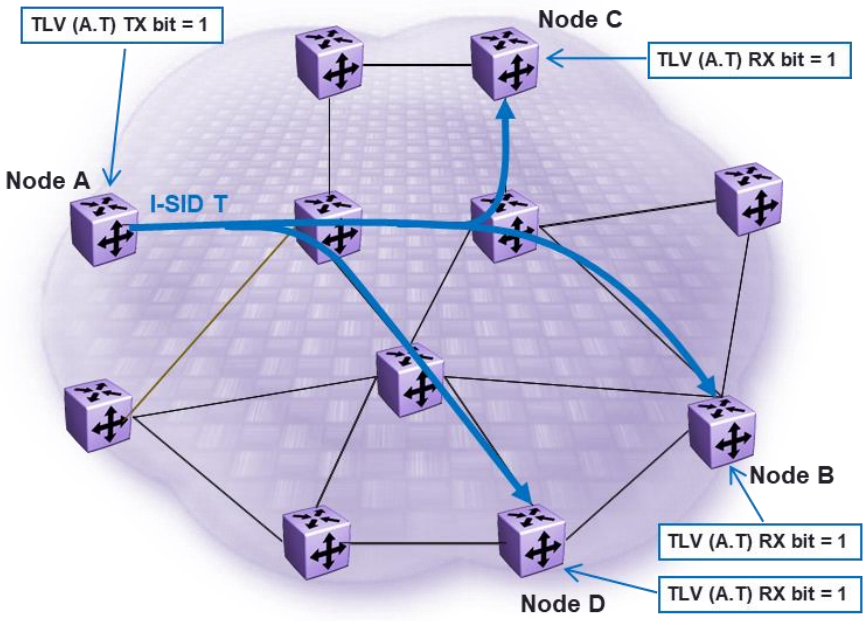


Figure 19 Service-Specific (I-SID) Multicast Shortest Path Tree Rooted at Node A

Clearly an SPF run on an SPB node is not just working out a single shortest path tree for itself, but also has to consider itself as potentially a transit (BCB) node on other node's I-SID multicast trees.

Tip

With today's CPUs, the time it takes to perform these SPF runs is less significant than the time it takes to program the switch hardware.

A further important property of SPB is that the path used by a multicast tree from a root node to a leaf node will always match (be congruent with) the unicast path between those same nodes in the same BVLAN. This is important to ensure that applications running over an L2 VSN service type (where some traffic might get initially flooded as the user MAC addresses are learned within the VSN service on the end-point BEBs) will never suffer from out-of-sequence packet delivery.

In both cases, the state of these shortest paths and trees is stored as MAC addresses in the BVLAN MAC tables and does not need to change unless the SPB physical topology changes or an I-SID service, which uses multicast-trees, is modified.

Tip

In terms of scalability, Extreme Networks VSP series switching platforms support hundreds of thousands of MAC entries in the hardware FIB (Forwarding Information Base).

Caution

Check platform release notes for exact scaling limits around SPB multicast BMACs. These are reported as maximum number of supported multicast streams when operating as a BCB.

An IP routed L3 VSN (or IP Shortcuts) service type always and only leverages the first SPB primitive; that is if only IP unicast routing is required. If, however, the L3 VSN has IP Multicast enabled, then the second primitive will be used to efficiently deliver IP Multicast streams only to where IGMP receivers exist, within that L3 VSN.

Tip

Each IP S,G (Source, Group) Multicast stream is allocated a dedicated I-SID (from a reserved I-SID range beginning at decimal 16,000,001) for optimal delivery across the SPB fabric.

In a traditional network design based on PIM-SM (or Draft Rosen over MPLS-VPNs), dedicated IP Multicast forwarding (S.G) records are used, but these tables are much more limited in size across any and all vendors' platforms, which in turn severely limits the overall IP multicast scaling of the network as a whole.

A L2 VSN service type will always combine both primitives. The first primitive is to deliver unicast traffic for user-MACs which have already been MAC learned within the VSN. The second primitive is to deliver broadcast, non-snooped multicast and unknown-unicasts on an I-SID tree, which will forward and replicate these packets only towards the other BEBs which are also members of the same L2 VSN. In this case, there will be as many I-SID trees as there are BEBs terminating that L2 VSN service.

Tip

It is the combination of I-SID and Nick-name of the BEB acting as root that makes an I-SID multicast tree unique in the SPB fabric. BEBs participating in an L2 VSN therefore all use the same I-SID, which was configured as part of the L2 VSN configuration.

If the L2 VSN has been enabled for IP Multicast (L2 IGMP snooping within the VSN), then L3 IP multicast traffic will also be handled using dedicated multicast I-SID trees. In this case, a stream specific I-SID will be

used, which will only forward the multicast stream to BEBs that have IGMP registered receivers (as opposed to using the L2 VSN I-SID tree, which would multicast the traffic to all BEBs in the L2 VSN).

The second primitive is also leveraged for Fabric Connect control plane enhancements, such as DVR and PIM Gateway (which will be covered later in this document), where reserved I-SID ranges are used to exchange IS-IS multicast based signalling.

Finally, SPB offers the ability to signal service termination at the edge of the Fabric via I-SID end-point provisioning, which in Fabric Connect can consist of any type of L2 or L3 service, type including IP multicast. In the case of L2 I-SIDs this service signalling can be further extended via Fabric Attach to non-SPB capable devices to reach the end user or IoT device as well as the data center application residing in virtualized hypervisor servers.

SPB Equal Cost Trees (ECT)

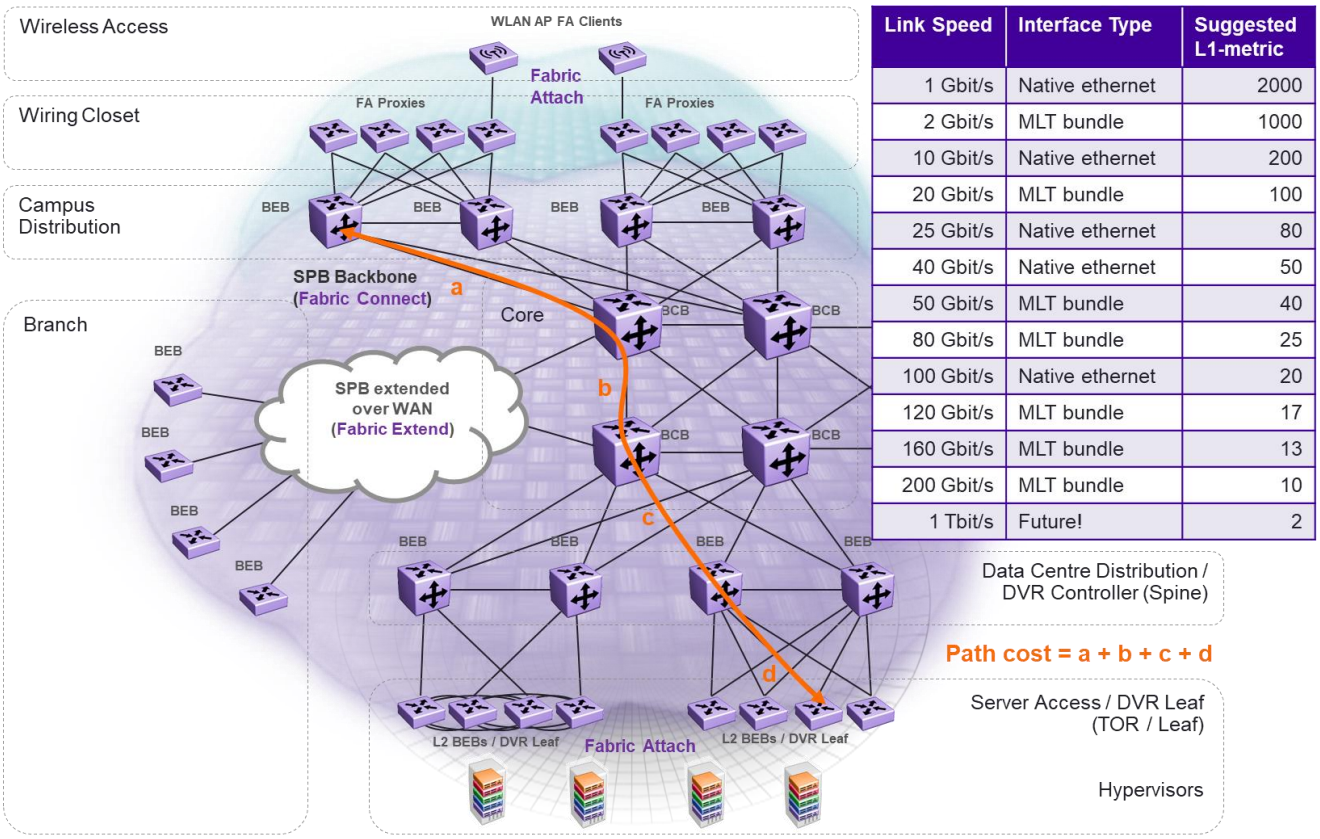
The previous section covered the SPB primitives and it explains how shortest paths for both unicast and multicast trees are then programmed into the BVLANs. This section will focus on how the shortest path is calculated and how equal cost trees are considered for use across the available BVLANs.

IS-IS calculates path cost by simply adding the link metric cost of every NNI link making up the path, which is very much how OSPF works. The congruence properties of SPB however demand that a given path must have the same path cost in both directions. Since an NNI link is terminated by two SPB nodes there is always the possibility that the link metric for the same link might be configured with different values on the two SPB nodes. SPB resolves this by ensuring that when IS-IS considers a link with inconsistent metrics, it will always use the worst metric for that link.

Tip

In the Extreme Networks Fabric Connect implementation, whenever an NNI link is configured with inconsistent metrics, the network administrator is alerted about this in the switch log file.

When a SPB Fabric is made up of NNI interconnects of different speed, it may become necessary to have the metric on the link reflect the speed of the NNI link. So a 100GbE NNI link should be configured with a link metric 10 times smaller than a 10GbE NNI link. Generally in most topology designs redundant NNI links will tend to have the same link speed, but if this was not the case, SPB can only prefer the path with the faster links if this is reflected in the link metrics.



Link Speed	Interface Type	Suggested L1-metric
1 Gbit/s	Native ethernet	2000
2 Gbit/s	MLT bundle	1000
10 Gbit/s	Native ethernet	200
20 Gbit/s	MLT bundle	100
25 Gbit/s	Native ethernet	80
40 Gbit/s	Native ethernet	50
50 Gbit/s	MLT bundle	40
80 Gbit/s	MLT bundle	25
100 Gbit/s	Native ethernet	20
120 Gbit/s	MLT bundle	17
160 Gbit/s	MLT bundle	13
200 Gbit/s	MLT bundle	10
1 Tbit/s	Future!	2

Figure 20 SPB path calculation and suggested link metrics

Caution

Extreme Networks VSP platforms currently do not automatically set the SPB metric on NNI ports based on link speed. Instead a default metric of 10 is used on all IS-IS interfaces when initially created.

When more than one path is found with shortest and equal cost, SPB can provision each available BVLAN into one of these paths. It is important to understand how SPB allocates paths to the available BVLANs so that this can be taken into account when designing a Fabric Connect network.

The SPB BVLANs act as different data planes for the same Fabric Connect network. In the absence of any equal cost shortest paths, all BVLANs will be programmed identically, whereas in the presence of equal cost shortest paths, each BVLAN can be programmed with one of those available shortest paths. VSN services are then allocated to one or another BVLAN in a deterministic manner (how VSN services are allocated to BVLANs is covered in section “Load Sharing Over Fabric Connect VSNs” on page 152).

Hence a SPB network can distribute traffic across as many equal cost shortest paths as it has BVLANs. On every SPB node, IS-IS computes the shortest path tree from that node towards every other node in the Fabric. Where equal cost trees are found IS-IS uses precise ECT algorithms to select a path in each available BVLAN.

Caution

The SPB standard (IEEE 802.1aq) defines a maximum of 16 BVLANs. Extreme’s SPB implementation currently supports 2 BVLANs. This could be increased to 16 BVLANs support in the future.

Hop count, is an important consideration when evaluating equal cost shortest paths. SPB considers that a path with a greater hop count will have a higher latency than a path with a lower hop count even if both

paths have the same path cost. Hence any paths with a higher hop count will not be considered if a path exists with a lower hop count and same metric.

In the event that IS-IS still sees a number of equal cost shortest paths (with the same metric and hop count) there has to be a deterministic way for all nodes in the SPB fabric to unequivocally agree which path to program in BVLAN#1 and which to program in BVLAN#2. This is an absolutely vital function for an SPB network because SPB relies on Reverse Path Forwarding Check (RPFC) to suppress transient loops and hence all nodes must agree on a same unique path between any two given nodes within the same BVLAN.

The SPB standard defines a set of 16 Equal Cost Tree (ECT) algorithms to be used to symmetrically calculate a unique shortest path into as many BVLANs. These ECT algorithms work by inspecting the concatenation of all Bridge IDs which define each path and masking every Byte of these IDs with a Mask value which is unique to each algorithm. The IDs are then re-ordered in numerical order to form PATHIDs. The numerically lowest PATHID is then selected and assigned and programmed into the respective BVLAN.

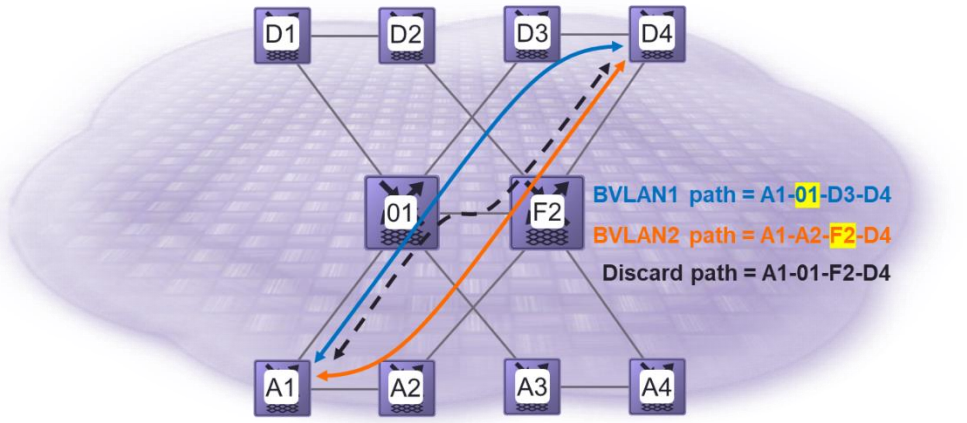


Figure 21 Example of SPB's ECT algorithms to select shortest paths

In the example depicted in Figure 21, all NNI links have the same metric and thus there are three possible equal cost shortest paths between each pair of diametrically opposite nodes. Extreme Networks Fabric Connect currently supports two BVLANs which is more than sufficient in the topology at hand. Therefore only two of those paths will be installed in the available BVLANs. All three available paths will thus be evaluated using the ECT Algorithm defined for BVLAN#1 then again using the ECT Algorithm defined for BVLAN#2.

Each path is defined by the concatenation of the SPB Bridge IDs (Bridge Priority + IS-IS System ID) and only those IDs which differ need to be compared. The ECT Algorithm for BVLAN#1 uses a null mask, so the path chosen in this case will always have the lowest numerical PATHID (01-D3 in the diagram). Whereas the ECT Algorithm for BVLAN#2 uses a mask of all 1s which flips every bit of the Bridge IDs involved before re-arranging them into a sorted PATHID from which to choose the lowest. In practice the ECT Algorithm for BVLAN#2 will be the highest numerical PATHID (A2-F2 in the diagram).

The other 14 ECT Algorithms defined for BVLANs 3-16 have non-null masks which would result in a different shuffle of the Bridge IDs for the paths being compared.

Tip

The SPB standard defined 16 ECT Algorithms defined under OUI 00-80-C2. However the standard allows SPB implementations to define new and alternative ECT Algorithms should this be required.

In general, the SPB ECT Algorithms provide path diversity wherever possible in a given network topology, but can on occasions not select the paths that one might expect or desire. If there is a need to engineer an SPB Fabric so that ECT Algorithms will always select a set of given paths, for example each traversing a

core node in the topology, then the network administrator can influence the ECT Algorithms only by intervening on the NNI link metrics or on the SPB node Bridge IDs.

In the example depicted in Figure 21, the less desirable path was discarded by ensuring that one core node had the lowest numerical Bridge ID (01) in the network, while the other core node the highest (F2). The same could also easily be achieved by setting a slightly higher NNI link metric on the interconnect between the two core nodes.

Caution

Changing the SPB Bridge ID on Extreme Networks Fabric Connect VSP platforms can only be done via manipulation of the IS-IS System ID. The SPB Bridge Priority field is not currently configurable and is hard coded to 00:00 in IS-IS TLVs.

Staying on the same physical topology example, a typical deployment scenario will have the distribution SPB BEBs paired in SMLT (MLAG) Clusters. If these BEBs are not fully meshed with the core nodes, as shown in Figure 22, it makes sense to deploy the VSP SMLT BEBs in a regular Primary/Secondary split-BEB allocation.

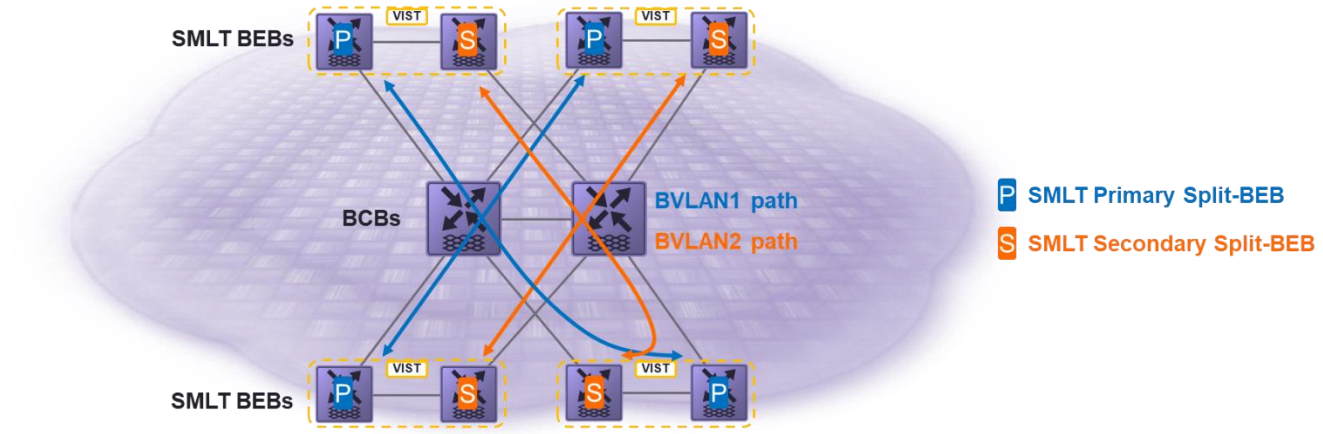


Figure 22 SPB shortest path optimization via SMLT Primary/Secondary BEB positioning

This is because, in the absence of any failure condition, the Primary SMLT BEB will always advertise the SMLT-Virtual-BMAC on BLVAN#1 (this determines where traffic to SMLT-Virtual-BMAC will be sent to) and will always transmit on the same BVLAN#1 (and vice-versa for the Secondary SMLT BEB for BVLAN#2).

So traffic between two SMLT Cluster BEB pairs, will normally (under no failure conditions) always flow from Primary to Primary and Secondary to Secondary. In the above diagram the traffic flow is somewhat less optimal for the lower right hand SMLT Cluster BEBs as these are wired into the topology in the opposite order.

Tip

Primary split-BEB is the VSP node forming the SMLT Cluster with the lowest IS-IS System ID; the other VSP node in the SMLT Cluster being the Secondary split-BEB. Primary/Secondary allocation can be inspected by looking at the BEB's ISIS SPBM properties.

IP Routing and L3 Services over Fabric Connect

Core IGP / GRT IP Shortcuts

With Fabric Connect, IP routing (whether IPv4 or IPv6) is just a service and is in no way acting as a foundation to the architecture. The Global Router Table (GRT) represents the default IP routing instance of a router and will usually be used for user traffic, sometimes even in multi-tenancy deployments leveraging VRFs and L3 VSNs. When the GRT is extended over SPB with IS-IS as the routing protocol, the resulting service type is referred to as IP Shortcut Routing rather than L3 VSN. This is because there can only be one instance of IP Shortcuts (there is only one GRT) and because IP Shortcuts do not need an I-SID to be defined but are instead deployed by IP enabling SPB on the nodes.

The term IP Shortcut also refers to an important property that applies to IP routing over an SPB fabric. Unlike conventional IP routing (e.g., with OSPF as the IGP) where IP routing is hop-by-hop and where the next-hop of every IP route in the IP routing table is the IP address of the immediate next-hop router and the IP TTL is decremented at every hop, with SPB the ingress BEB performs the first IP routing hop and the egress BEB performs the second and last IP routing hop (IP TTL decrements by 2). In between, an SPB L2 shortcut is taken along the IS-IS computed shortest path across the Ethernet fabric. This is similar to MPLS-VPN routing where the end-point VRFs perform two IP routing hops with an MPLS LSP in between.

IP Shortcuts are conceptually identical to VRF L3 VSNs, but they do differ in some minor respects. Since IP Shortcut routing doesn't use an I-SID, the packet encapsulation used does not need to use Mac-in-Mac and the IP encapsulation is mapped directly onto an SPB BMAC Ethernet header. Effectively, IP Shortcuts encapsulation is like a Mac-in-Mac encapsulation where the inner Ethernet MAC header (and I-SID field) has been removed, thus leaving just IP in BMAC. Figure 23 is an update of Figure 1, but now includes IP Shortcuts. IP Shortcut IP routes are not associated with an I-SID when advertised by IS-IS, which thus uses the traditional IS-IS TLV 135 (Extended IP Reachability) and IS-IS TLV 236 (IPv6 Reachability).

One other aspect where IP Shortcuts differ from L3 VSNs is that in the Extreme Networks VSP series platforms, all in-band management traffic (e.g., SSH, SNMP, HTTPS) is only processed on IP interfaces that belong to the GRT. Therefore, in an SPB fabric deployment, the GRT IP shortcuts are always deployed, if not for user traffic, at least for management traffic.

In a "greenfield" multi-tenant design it therefore makes a lot of sense to dedicate the GRT and IP Shortcuts exclusively for IP based management (SSH, SNMP, TACACS, RADIUS, etc.) and to only designate Layer 3 VSNs to carry user traffic and business applications in order to isolate user traffic from network control traffic as a means to improve network security. This makes for a more robust architecture where the risk of unauthorized access onto the network devices is greatly reduced and the use of management access policies can be greatly simplified.

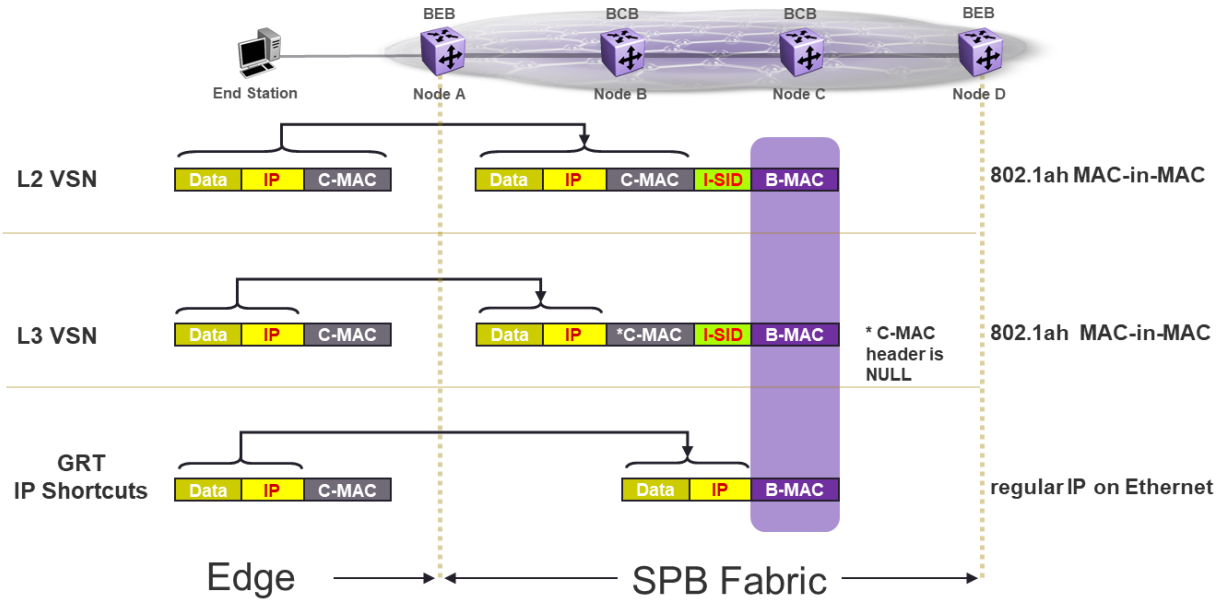


Figure 23 Different Encapsulation Used by GRT IP Shortcuts

Virtualized L3 VPNs / L3 VSNs

L3 VSNs represent virtualized IP routing domains as a service over an SPB fabric. Conceptually they offer the same service as MPLS-based IPVPNs but over a considerably simpler architecture. An L3 VSN is terminated in VRF instances located on the BEB nodes typically acting as distribution layer. In a campus environment, the VRF IP subnets will typically be local; that is, directly connected IP interfaces corresponding to end-station and server subnets. If this is the case, these VRFs will just need to redistribute into IS-IS direct routes only. Though it is often necessary for VRFs to hold Static, OSPF, BGP, RIP IP routes obtained from or pointing to a traditional IP router outside of the Ethernet fabric, in which case these IP routes will also need redistributing into IS-IS.

The service id (I-SID) configuration on the VRFs determines which VRFs belong to the same L3 VSN. When a VRF with an I-SID assigned redistributes an IP route into IS-IS, those routes are announced using newly defined TLVs for use with SPB (I-SID-IPv4 or I-SID-IPv6). These TLVs announce reachability of a given IP subnet for a given I-SID (L3 VSN). Other VRFs configured with the same I-SID (thus belonging to the same L3 VSN) will automatically inspect the IS-IS LSDB and install any IP routes associated with the same I-SID. These routes will thus be installed in the VRF IP routing table with as next-hop the BMAC of the BEB node which originated the TLV.

IP routing over an L3 VSN is therefore equivalent to IP Shortcut routing in that IP routing is performed only on the ingress BEB and on the egress BEB, with an SPB L2 shortcut to the egress BEB BMAC in between (just as MPLS-VPNs use an LSP between the ingress and egress PEs). Figure 24 shows the relevant forwarding tables that make L3 VSNs possible with SPB.

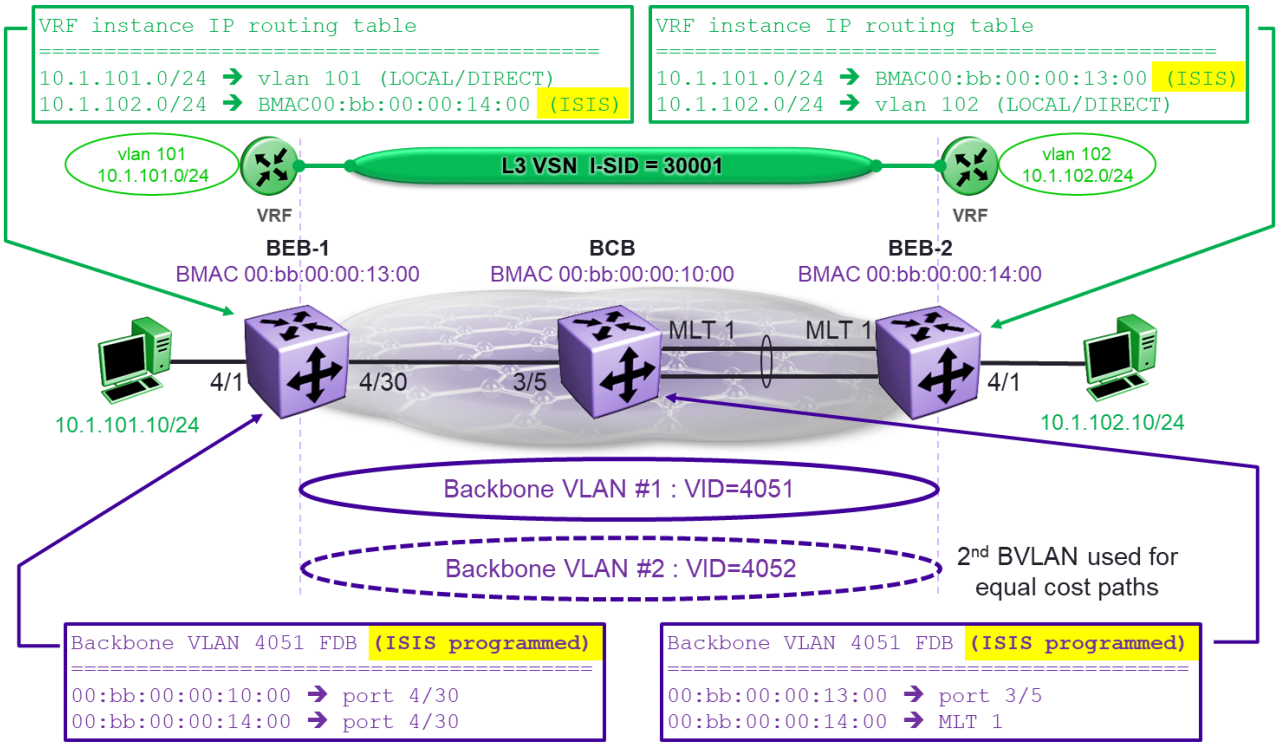


Figure 24 Relevant SPB L3 VSN Forwarding Tables

Tip

One of the nice properties of L3 VSNs (and IP Shortcuts also) is that inspection of the VRF (GRT) IP routing table will show IS-IS routes with next-hop an easily recognizable IS-IS ASCII system-name of the egress BEB (instead of just an IP address as is the case with a traditional OSPF IGP, or MPLS-VPN).

The example in Figure 23 is showing next-hop BMACs to explain how L3 VSNs actually work. In practice, the IS-IS ASCII system-name associated with the node's BMAC would be shown instead.

VPN Security Zones / IP IS-IS Accept Policies

Each L3 VSN represents a virtualized IP routing domain that segregates and isolates the IP devices within it from other IP devices belonging to other L3 VSNs. The ability of segregating the network users into fully isolated virtual networks raises the security and availability of the network. Communication between users belonging to different functions can thus be prevented and different applications can be made to run in dedicated virtual networks.

However, because these virtual networks ultimately belong to the same physical network, it is usually desirable for them to share certain services such as Internet access, DHCP services, DNS services, or Shared servers delivering some other application across different L3 VSNs. These services can reside either outside of the Ethernet fabric or within a VRF/L3 VSN of their own.

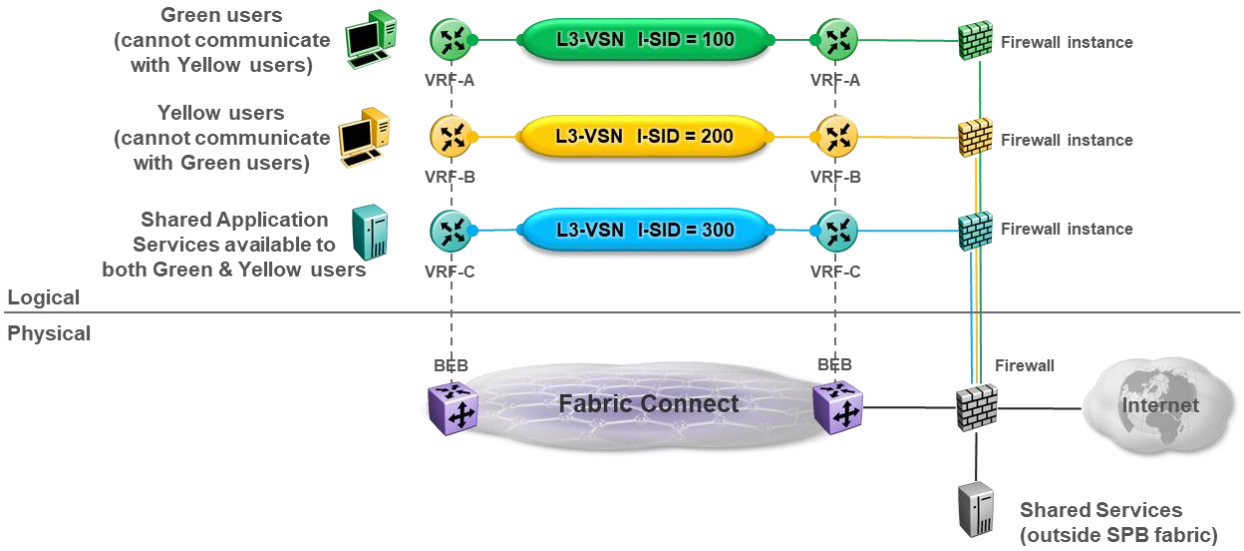


Figure 25 Security Zones with Common Services

There are two design approaches that are equally possible and can complement each other. If the requirement is for all inter L3 VSN communications to be secured by a firewall, then each L3 VSN will need to be head-ended by a dedicated firewall instance, which will allow the creation of security policies specific to all traffic leaving or entering the L3 VSN (including NAT, if necessary). This approach can be used to provide shared access to the Internet as well as towards shared services located outside of the virtualized Ethernet fabric. Of course, this approach can also be used to allow different L3 VSNs to communicate together via the same Firewalls.

The other approach is applicable where there is a desire to let certain IP networks in one L3 VSN communicate with some other IP networks in different L3 VSNs, and there is no requirement to firewall that traffic or further congest the existing firewalls with this additional traffic. This approach consists in "redistributing," in a controlled manner, selected IP routes from one L3 VSN to another. For bi-directional traffic communication, it is always necessary to redistribute the necessary routes in both directions.

In an SPB fabric design this design can be easily achieved via IS-IS IP accept policies. A VRF belonging to an L3 VSN will, by default, only ever accept and import IP routes which belong to the same I-SID service id. However, this behavior can be overridden via IS-IS accept policies whereby a VRF can be made to accept IP routes from any other available I-SID in the SPB fabric (including IP routes on the same local BEB but in a different VRF). Installing the IP route in the VRF control plane is the key to allowing traffic to follow that IP route in the data plane. In the Extreme Networks Fabric Connect implementation this capability is possible between VRF/L3 VSNs as well as GRT/IP Shortcuts and VRF/L3 VSNs. The ability to "redistribute" SPB with IS-IS Accept policies is similar to the "leaking" of IP routes in between MPLS-VPNs using BGP Route Target manipulation.

Tip

Benefits of SPB IS-IS Accept policies over MPLS-VPN Route Target manipulation

- In SPB an I-SID-IPv4 (or I-SID-IPv6) route can only belong to one and only one I-SID. This is not true with MPLS-VPNs where the same BGP VPN-IPv4 (or VPN-IPv6) route can be tagged with any number of export Route Targets (RT) as Extended Communities.
- The MPLS-VPN approach is more sophisticated as one can define multiple import and export RTs. However, this boils down to different ways to configure the same behavior that makes the MPLS approach more complex to design and manage.
- An SPB L3 VSN has a well-defined default behavior of not communicating with any other L3 VSNs (if no IS-IS accept policies are defined). With MPLS-VPNs a VPN which does not "leak" routes with other VPNs is really just a special case of a VPN which needs to be configured with the same Route Target (RT) as both export and import.

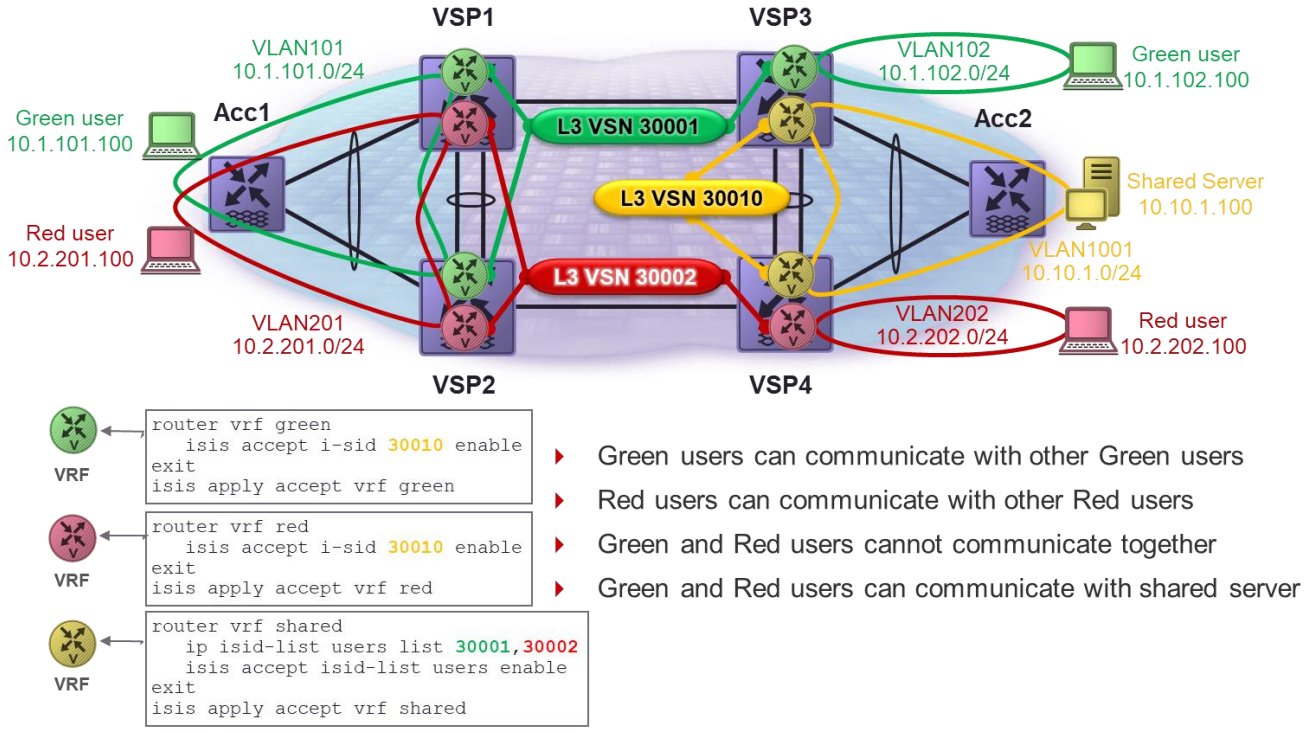


Figure 26 Simplicity of Using IS-IS Accept Policies with SPB

Figure 26 shows the simplicity how IP routes can be accepted from one L3 VSN (or IP Shortcuts) to another. This will work equally with IS-IS learned IP routes of a different L3VSN from a distant BEB as well as with IP routes of a different L3VSN which are present on the same local BEB in a different VRF (in Figure 26 this is happening on the VSP3 & VSP4 nodes).

It should be noted that in the Extreme L3 VSN BEB implementation, traffic that is received from the fabric (with Mac-in-Mac encapsulation) can be IP routed back into the same fabric, if an applicable IP route for that traffic requires that. This means that in the example depicted, if the shared L3 VSN is announcing a default route and this default route is accepted by both the green and red user L3 VSNs, then communication becomes possible between the green and red users since that traffic will now follow the default route to reach the shared VSN VRF. This also means that it will get IP routed across as the shared VRF is accepting IP routes from both the green and red L3 VSNs.

Caution

Extreme Networks Fabric L3 VSN BEBs are capable of IP routing traffic received from the fabric back into the fabric in a single IP routed hop.

ISIS IP Route Types and Protocol Preference

The Extreme Fabric Connect L3 service types (L3 VSN and IP Shortcuts) are able to announce two different types of IP routes in IS-IS using the both the I-SID-IPv4 and I-SID-IPv6 TLVs as well as with the regular TLV 135 (Extended IP Reachability) for IPv4 and TLV 236 for IPv6 routes. These are referred to as IS-IS Internal IP routes and IS-IS External IP routes and in many ways, offer a similar distinction as what OSPF offers with LSA-5 external Type1 and Type2 routes.

Note

Technically, unlike OSPF, all IP routes are "external" to IS-IS, since IS-IS does not operate on top of IP. Yet we shall refer to "internal" IS-IS routes as those routes which derive their metric from the calculated SPB Fabric path cost as opposed to the route metric carried in the IS-IS TLV.

Every IP route advertised in SPB IS-IS TLVs has a metric associated with the route and carried in the TLV itself. This metric is typically obtained from the original route metric on the IP routing table of the BEB that redistributed the route into IS-IS. We will refer to this metric as the "external" metric of the route.

Tip

The external metric of an IS-IS IP route can be manipulated via use of redistribution route policies.

When a different BEB, in the same VSN routing domain (L3 I-SID) considers that IS-IS IP route for inclusion in its IP routing table, a separate "internal" metric for that IP route is computed. The internal metric is no other than the SPB's shortest path cost to reach the BEB that originated the TLV containing the IP route. That is, the internal metric is in fact the SPB path cost metric.

In the original Extreme Fabric Connect implementation, all IS-IS IP routes were considered "internal" type route and were considered for inclusion in the IP routing tables based on their internal metric alone, with the external metric only being used as a tie breaker.

Since then, the Fabric Connect implementation has been enhanced to support a new "external" type route that can now be considered for inclusion in the IP routing table based on its external metric alone

Note

Encoding of the IS-IS external route type is included within the I-SID-IPv4 and I-SID-IPv6 TLVs used with L3 VSNs and is included in a sub-TLV of TLVs 135 (IPv4) and 236 (IPv6) used with GRT IP Shortcuts.

The following table provides a breakdown of the IS-IS IP route selection criteria when the same IP route is seen advertised by more than one distant BEB in the IS-IS LSDB.

Table 9 – IS-IS Internal and External IP Route Tie Breaking

Tie breaking criteria when the same IP route is advertised by more than one distant BEB
• Same IP route seen as both internal and external type. ➢ IS-IS Internal route is preferred over IS-IS External route. • Both IP routes are internal type. ➢ The IP route with the lowest internal metric is preferred. • Both IP routes are internal type AND have the same internal metric. ➢ The IP route with the lowest external metric is preferred. ➢ Both IP routes are internal type AND have the same internal AND external metric. ➢ The IP routes can be considered for IP ECMP. • Both IP routes are external type. ➢ The IP route with the lowest external metric is preferred. • Both IP routes are external type AND have the same external metric. ➢ The IP routes can be considered for IP ECMP.

The availability to distinguish between internal and external IS-IS IP routes allows for greater flexibility in achieving IP routing design goals. As an example, a common design requirement when deploying redundant firewalls is to designate one firewall as primary and the other as secondary, whereby traffic should always be forwarded to the primary firewall while it is present and operational.

In the example depicted in Figure 27, the BEBs, to which the firewalls are connected, both announce via IS-IS a default route. If these IS-IS IP routes were of internal type, it would result in other BEBs always preferring the default route towards the nearest firewall. In contrast, using IS-IS external routes and suitable manipulation of the external route metric can ensure that all BEBs will always install the default route towards the primary firewall.

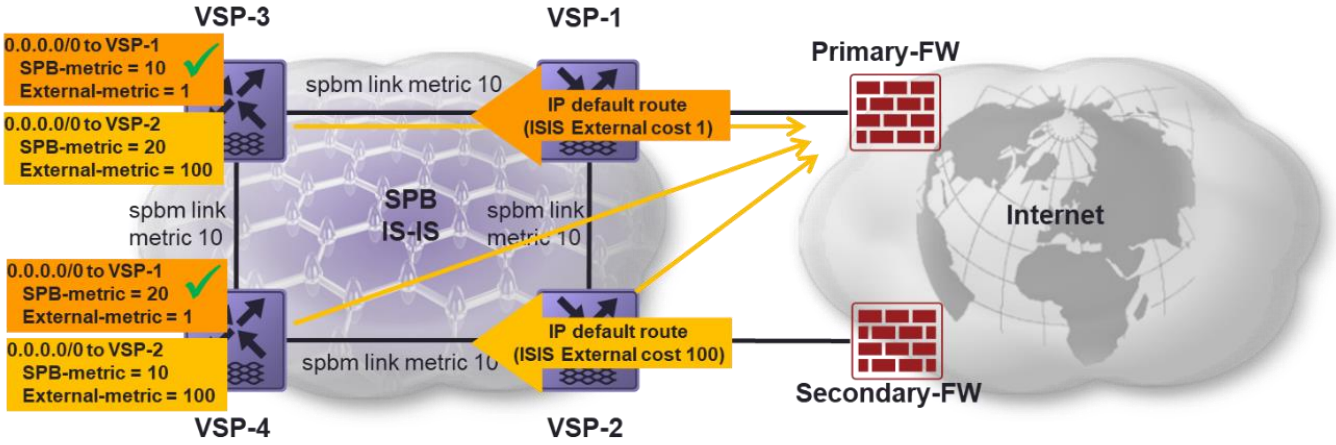


Figure 27 IS-IS External Routes Prefer Lower Route Metric over SPB's Shortest Path

Another useful application of IS-IS external routes in conjunction with IS-IS Accept policies is when a fabric L3 service (L3 VSN or IP Shortcuts) needs to be redundantly connected into an IP routed cloud using traditional IP routing protocols such as OSPF, RIP, or BGP, as illustrated in Figure 28. This is a delicate exercise where care needs to be taken that IP routes redistributed from one cloud into the other by the first border router do not get redistributed back into the originating cloud by the other border router.

Note

The same delicate exercise equally applies when redistributing between any two IP routing protocols. The very same challenges had to be dealt with in the days when RIP networks were connected or migrated to OSPF.

The first point to be aware of and to take into consideration is that every IP router (and a fabric L3 BEB is no different) maintains a ranking of protocol preferences for every IP routing protocol susceptible to install IP routes in the IP routing table. If the very same IP route is available from two different IP routing protocols, then only the one from the protocol with the highest protocol preference will be installed in the IP routing table.

Tip

Extreme Networks VOSS VSP platforms allow configuration of the protocol preferences on a per VRF/GRT basis.

Note

In Cisco terminology, the protocol preference is referred to as "Administrative Distance."

The Extreme protocol preference implementation uses a numerical value between 0 – 255. The lower the value, the higher the priority of the protocol. The same is true with Cisco's "Administrative Distance."

Tip

In some implementations, like in the Extreme Networks VOSS VSP platforms, the same IP route from the routing protocol with the second highest protocol preference is not simply discarded but is also installed in the IP routing table as an "Alternative" IP route (except of course if IP Alternative routes has been disabled on the IP routing instance).

Table 10 – Extreme VOSS VSP Default Protocol Preferences

IPv4	IPv6	Default Preference Value
Local / Direct	Local / Direct	0
Static	Static	5
SPBM_L1 (IS-IS)	SPBM_L1 (IS-IS)	7
OSPF_INTRA	OSPFv3_INTRA	20
OSPF_INTER	OSPFv3_INTER	25
eBGP	eBGP	45
RIP	RIPng	100
OSPF_E1 (External type1)	OSPFv3_E1 (External type1)	120
OSPF_E2 (External type2)	OSPFv3_E2 (External type2)	125
iBGP	iBGP	175

When redistributing between two separate IP routing protocols, it is always important to determine which protocol will have the highest preference, as this will determine in what way the necessary routing policies will have to be put in place. Generally, the routing protocol of the core network should have the highest

preference. To put it another way, if there really was an IP route conflict and the same IP route was to be seen in both clouds, would you want the smaller IP cloud to interfere and potentially replace that same IP route of your core network?

In the example at hand we will assume that the core network is the Extreme Fabric Connect SPB cloud and that it should therefore use a higher protocol preference than OSPF/RIP or BGP. This is step (1).

Tip

By default IS-IS already has a higher protocol preference than other IP routing protocols; the default protocol preferences are shown in Table 10. These values can be changed on a per VRF/GRT basis.

When the actual protocol redistribution is configured on the border routers, OSPF/RIP/BGP routes should be redistributed into IS-IS as external routes and, in the reverse direction, only IS-IS internal routes will be redistributed into OSPF/RIP/BGP. This provides a neat and elegant way to prevent the same IP route being reflected back towards the same routing protocol from which it originated.

As soon as the IP route redistribution is initiated, if there are two or more border routers between the two clouds we can observe the protocol preferences at play. What will typically happen is that one border router will start redistributing the OSPF/RIP/BGP routes into IS-IS before the other one. The slower (or last configured) border router will thus be presented with the same IP routes from both OSPF/RIP/BGP on one side and from IS-IS on the other. Since we have determined that IS-IS is to have the higher preference, the border router would thus replace the OSPF/RIP/BGP route it had in its routing table with the IS-IS route obtained from the other border router (and would keep the OSPF/RIP/BGP route as alternate).

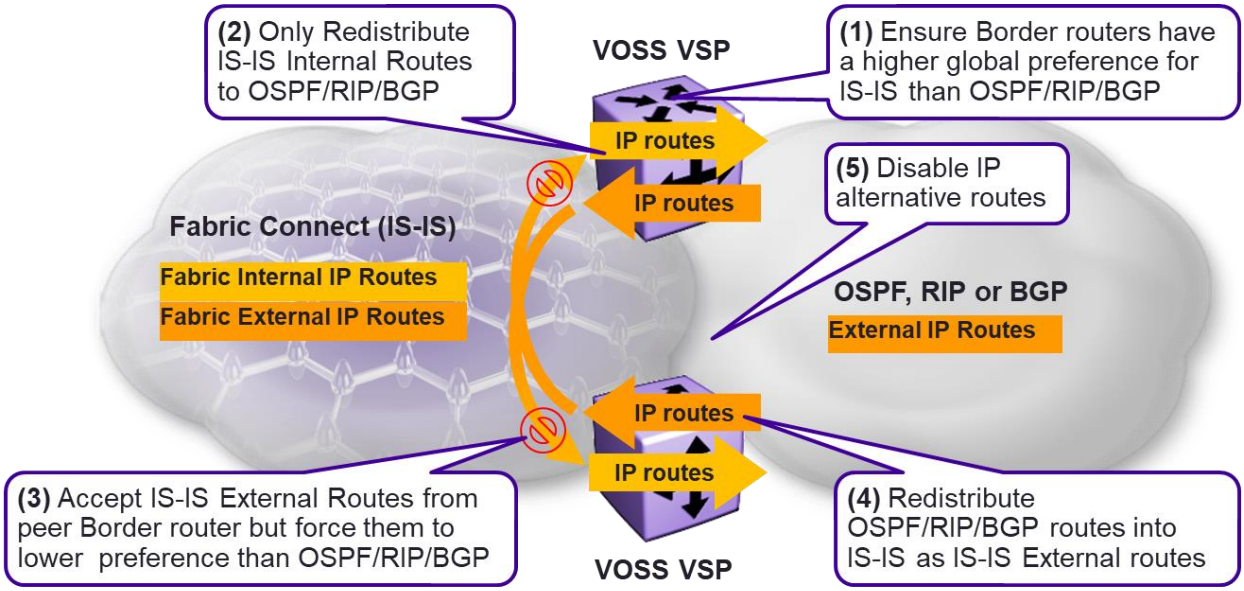


Figure 28 Redundantly IP Routing Fabric VSN with External OSPF/RIP/BGP Network

There are no routing loops and this becomes a steady state for the IP routing tables, but clearly this is sub-optimal, since we know that the IP routes in question originate from the OSPF/RIP/BGP cloud. The correct approach to remediate is to apply on the border routers an IS-IS Accept policy, which will act only on IS-IS external routes (if these are always IP routes from the OSPF/RIP/BGP cloud in question) or specifically only IS-IS external routes received from the peer border routers (otherwise) with an override action to use a different protocol preference value for these IP routes. The modified protocol preference value will need to be such that these IS-IS external routes will not be able to displace the original OSPF/RIP/BGP routes already in the IP routing table.

The protocol preference plays also on the reverse side, namely an IS-IS Internal route, once redistributed into OSPF/RIP/BGP by one border router will be seen by other border routers from both protocols: IS-IS on one side and OSPF/RIP/BGP on the other. However, in this case the globally set (and default in this case) protocol preferences will always ensure that the IS-IS Internal routes cannot be replaced by an OSPF/RIP/BGP one.

It is worth noting however that these lower preference and discarded IP routes are in fact still installed in the IP routing table, but they are installed as alternative routes. The intention of alternative IP routes is that, should the active IP route become unavailable, the alternative IP route immediately becomes the new active IP route. This is done without waiting for the routing protocols to converge, and in this redistribution use case becomes undesirable.

Imagine a valid IS-IS Internal IP route suddenly becomes unavailable, and both border routers suddenly decide to replace it with their OSPF/RIP/BGP alternate IP route. This route will not survive very long because the same border routers will stop redistributing it into OSPF/RIP/BGP, but it creates a transient state during which, ironically, the same border routers will try to redistribute it back into IS-IS as an external route. The correct approach is to simply disable the IP alternative route functionality on the border routers.

L2 Services Over SPB IS-IS Core

E-LAN / L2 VSNs with CVLAN/Switched UNI

An L2 VSN is the interconnection at Layer 2 (bridging) of distant Ethernet segments into one single Ethernet or L2 VLAN domain. A L2 VSN is natively an any-to-any (E-LAN) service type which means that it can have any number of end-points and that MAC learning is performed within the VSN service. This MAC learning is confined to the edge BEB nodes where the service terminates and is not applicable in any way to the core BCB nodes or to IS-IS (which is only concerned about BMACs). Furthermore, BEBs are able to do reverse MAC learning in the sense that any user MAC seen from a distant BEB will be learned against that BEB's BMAC during the Mac-in-Mac de-capsulation. This will then allow the end-point to send traffic toward that end-user MAC using a unicast Mac-in-Mac encapsulation. This is illustrated in Figure 29.

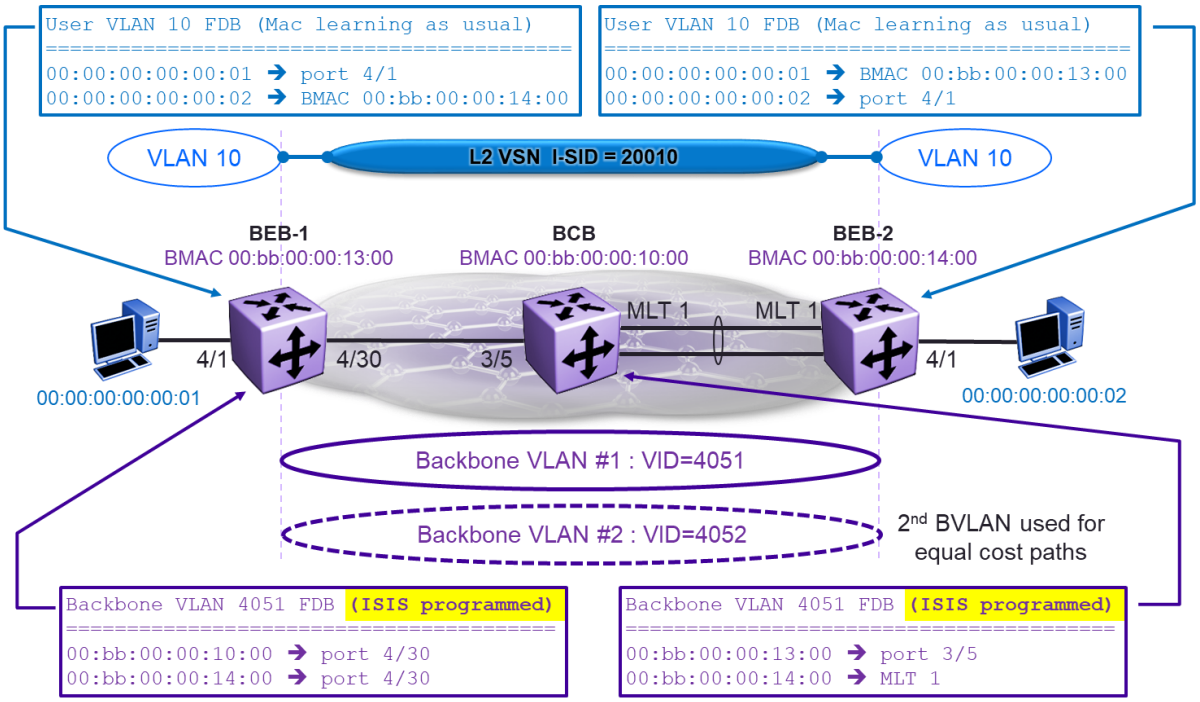


Figure 29 Relevant SPB L2 VSN Forwarding Tables

Tip

The VLAN-ID is only end-point (BEB) significant and can be re-mapped to a different VLAN-ID across the L2 VSN service.

IS-IS is not in any way involved in advertising user MAC addresses. The only role that IS-IS plays for L2 VSNs is to exchange I-SID membership information within the Ethernet fabric, which in turn allows the fabric to automatically provision I-SID multicast trees for the delivery of L2-multicast, broadcast, and unknown-unicast to MACs which have not yet been learned by the VSN. Since service-specific (I-SID) multicast trees are tied to a single root node, SPB will automatically provision as many multicast trees as the L2 VSN has end-points. Each multicast tree consumes one entry in the BVLAN Forwarding Information Base (FIB) of BEB or BCB nodes which happen to be along the shortest path of that multicast tree.

Tip

Use of the SPB multicast trees means that L2 VSNs are able to correctly handle flooded traffic within the VSN and perform replication along the multicast tree in an efficient manner.

This is not true with MPLS VPLS, which uses a full mesh of EoMPLS circuits to deliver an any-to-any service and hence needs to ingress replicate flooded traffic across all EoMPLS circuits. This becomes exponentially inefficient as the number of end points in the VPLS VSI increases. The same ingress replication inefficiencies are true for EVPN.

Configuration of the UNI end-points is done by assigning an I-SID service id to the Ethernet segments or VLAN-IDs or combination of both that are to be bridged together by the L2 VSN. There are a number of possible UNI types depending on how the Ethernet segment or VLAN is presented to the end-point BEB node. In the case of E-LAN L2 VSNs, the most commonly used UNI types are CVLAN and Switched.

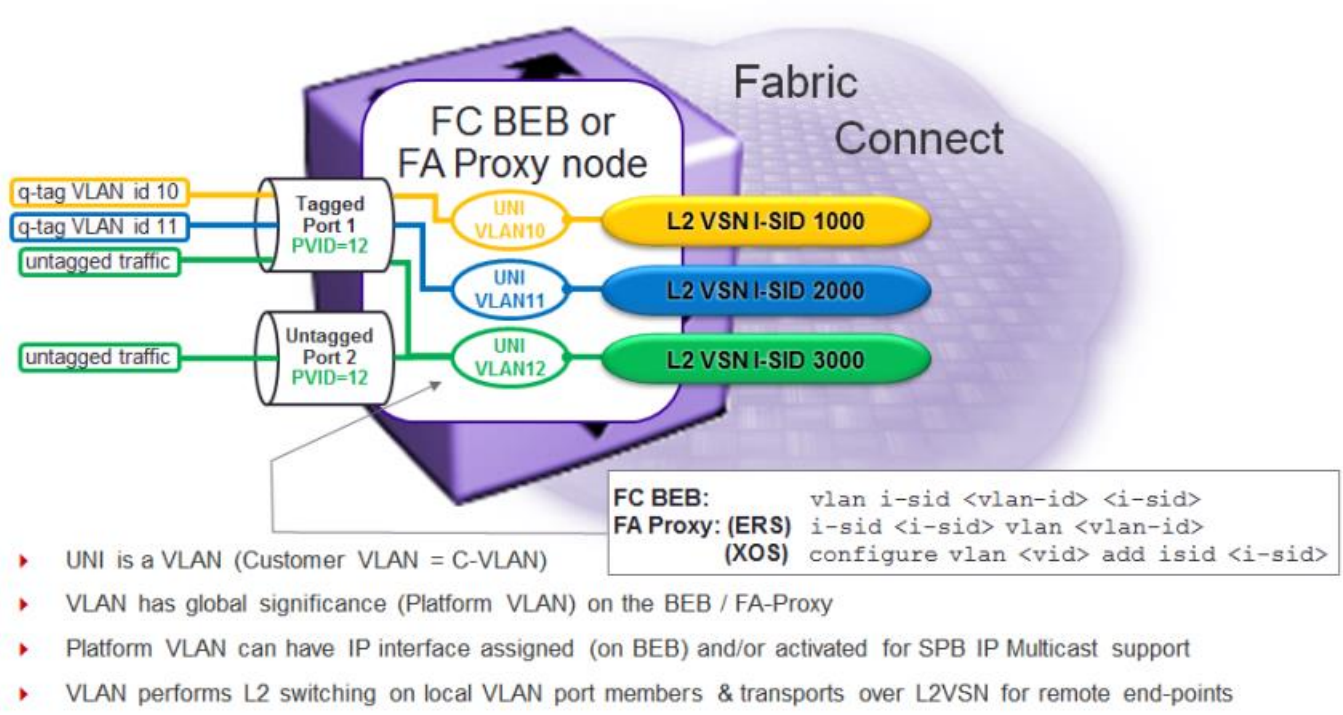


Figure 30 CVLAN UNI

In a campus environment, the CVLAN UNI, depicted in Figure 30, is the most commonly used UNI type. This allows the L2 VSN to terminate into a regular user-VLAN that can be locally bridged across more than one of the BEB local Ethernet (UNI) ports or terminated via Fabric Attach on FA Proxy access switches. The BEB may also have an IP interface defined on the VLAN in order to act as default gateway for traffic originating from that L2 VSN segment or to activate SPB IP Multicast support for the VSN. The CVLAN UNI type is also equivalent to the L2 I-SID service attachment offered by a FA Proxy node.

The Switched UNI, depicted in Figure 31, is more commonly used when the Ethernet segment to be assigned to the L2 VSN is discretely identified to a VLAN-ID on a specific Ethernet port and/or there is a need to be able to leverage the whole VLAN-ID range (1-4095) independently on each UNI port. As well this feature has the ability to perform VLAN-ID mapping across the same BEB UNI ports. These types of requirements are more common in a service provider role where the SPB fabric is used to provide L2 TLS connectivity to other end-customer networks.

In the case of Switched UNI, the L2 VSN MAC table held on the end-point BEB node is no longer a VLAN MAC table but rather a MAC table directly associated to the I-SID itself. It is quite possible to assign CVLAN

UNIs and Switched UNIs to the same I-SID, in which case the user-VLAN that constitutes the CVLAN UNI will be bridged together with the Switched UNIs. This is also possible on the same local BEB.

Note

Switched-UNI is in fact what is being used with Fabric Attach by the FA Server BEB and is also the only available UNI type on a DVR leaf node.

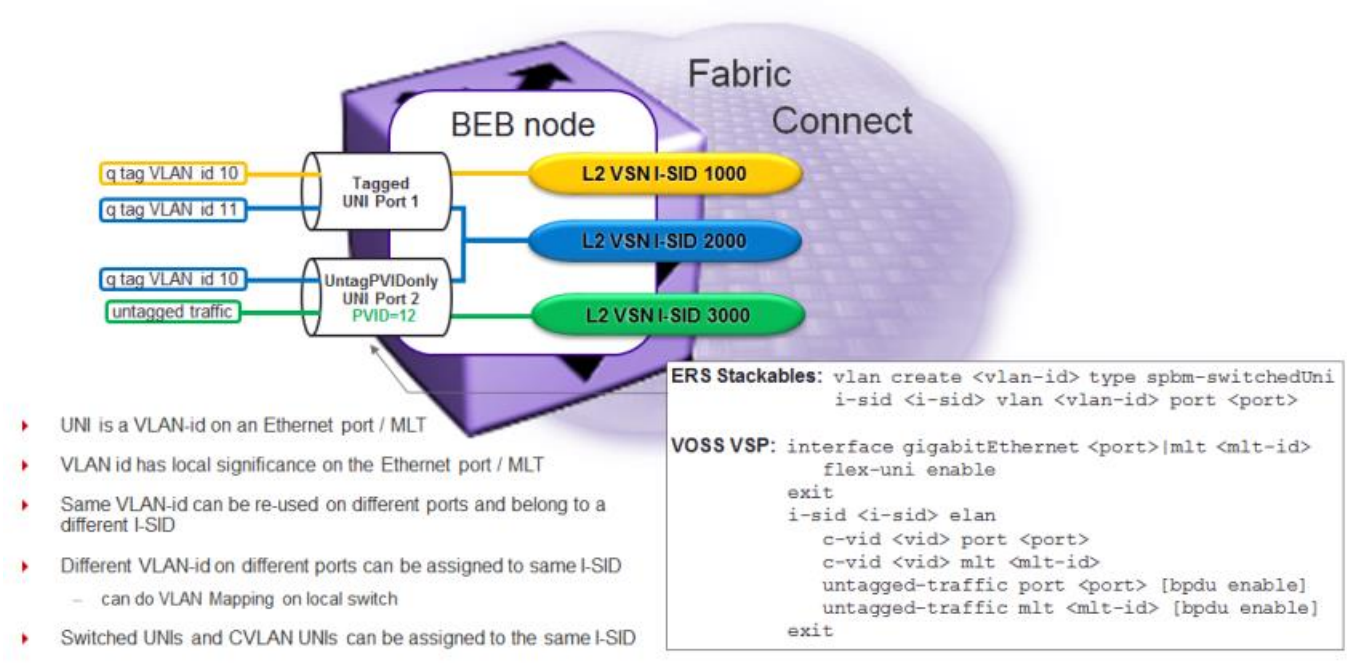


Figure 31 Switched UNI

Tip

Extreme VSP platforms also allow Switched-UNI bindings to be defined specifically for untagged traffic. In this case, there is a BPDU configurable switch that will determine whether or not Spanning Tree BPDUs will be transported into the L2 VSN.

E-LINE / L2 VSNs with Transparent UNI

Delivering point-to-point (E-LINE) L2 tunnels over SPB is just a special case of an L2 VSN that was defined with only two end-points. This is most commonly done in conjunction with a Transparent UNI type that is able to associate the L2 VSN with all traffic received on a raw Ethernet port. Whatever traffic is received on that port, with or without a q-tag, and even if the traffic belongs to a signalling protocol using link-local PDUs (e.g., Spanning Tree BPDUs, LACP, LLDP), this traffic will be tunnelled across the SPB fabric.

Tip

Only Ethernet Flow Control Pause frames cannot be transported across the L2 VSN with Transparent UNIs.

Tip

Transparent UNIs are also capable of performing VSN MAC learning, so they can also be used for E-LAN type L2 VSNs where three or more Transparent UNIs are assigned to the same I-SID.

Caution

Transparent UNIs should not be assigned to the same I-SID as CVLAN or Switched UNI.

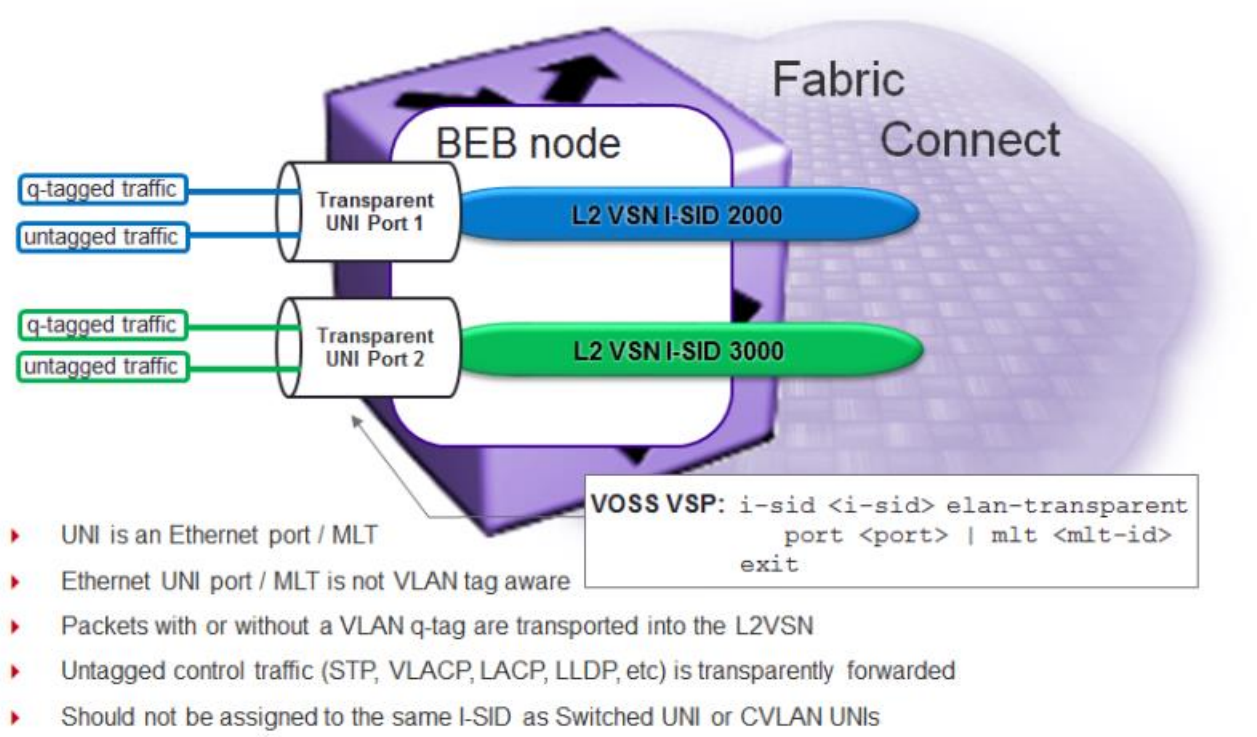


Figure 32 Transparent UNI

E-TREE / L2 VSNs with Private-VLAN UNI

An E-TREE service type is an enhanced version of an E-LAN L2 VSN that uses a CVLAN UNI where the VLAN used is a Private-VLAN. A Private-VLAN is a special VLAN construct where port members can be assigned one of three possible roles:

- Promiscuous role: Devices connected to these ports are able to communicate with every other device in the same L2 segment. Promiscuous ports are usually untagged ports that are only members of the Private-VLAN.
- Isolated role: Devices connected to these ports are only able to communicate with devices hanging off Promiscuous ports. Isolated ports are usually untagged ports that are only members of the Private-VLAN.
- Trunk role: These ports must be used to q-tag extend a Private-VLAN between two switches. A Trunk port is always q-tagged and can carry one or more Private-VLANs as well as other regular VLAN ids.

When a Private-VLAN is created, it is necessary to provide two VLAN-IDs: a Primary VID and a Secondary VID. In general, a Promiscuous port will always transmit traffic on the primary VID but will receive traffic from both the primary and secondary VIDs. An Isolated port will transmit traffic into the secondary VID and will only receive traffic from the primary VID. Both VIDs share the same MAC table and are both q-tagged on Trunk ports. This is the industry standard IEEE 802.1Q implementation of Private-VLANs used by many networking vendors as well as server hypervisor vendors such as VMware with ESX.

Extreme Networks Fabric Connect conforms to this implementation and enhances it to seamlessly operate over the SPB Fabric by being able to assign an L2 VSN I-SID to a Private-VLAN with end-point provisioning.

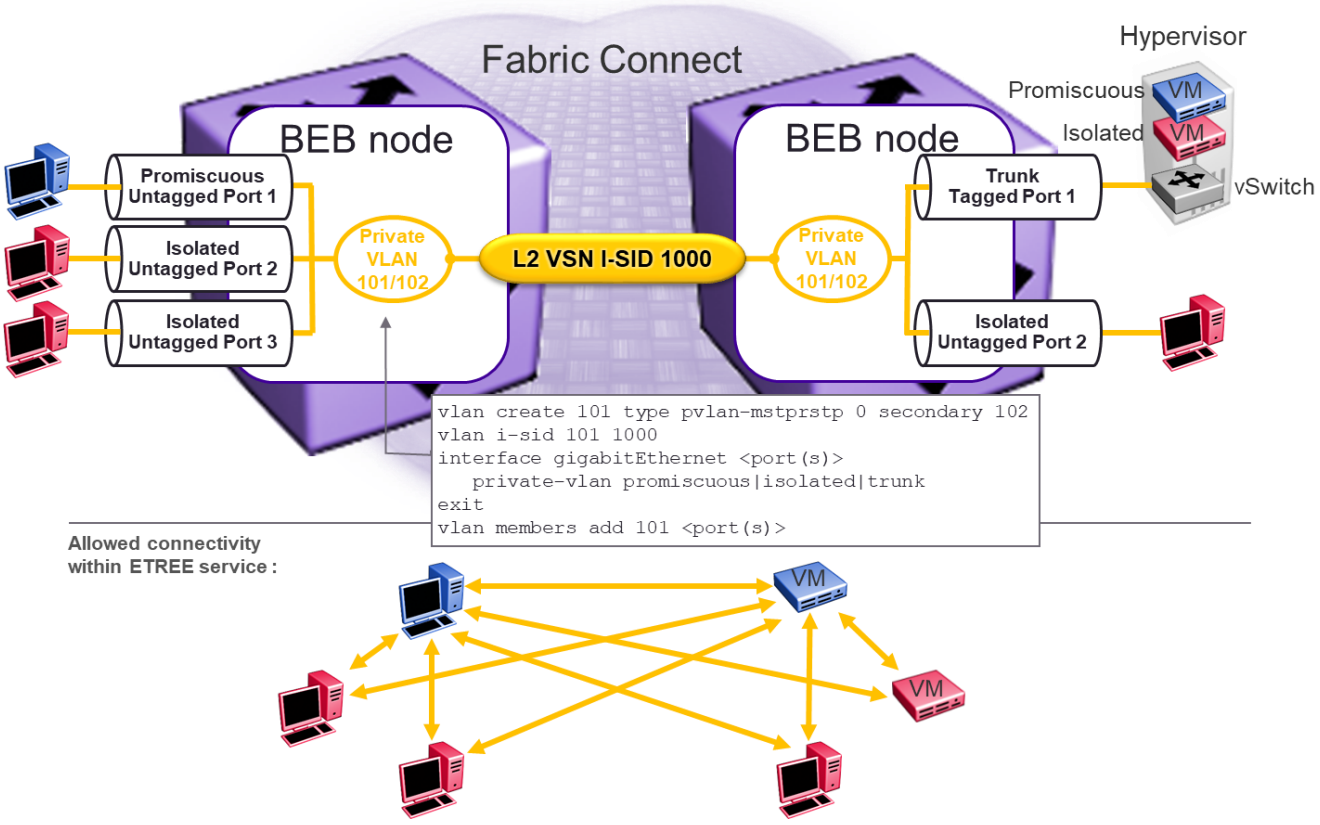


Figure 33 E-TREE Private-VLAN L2 VSN

Tip

An E-TREE L2 VSN can mix end-point BEBs using Private-VLAN UNI as well as regular CVLAN UNIs. This means all devices connected to the CVLAN UNI will have Promiscuous-like connectivity within the L2 segment. If an IP interface needs to be configured to IP route the traffic on and off the E-TREE segment, that IP interface should be configured on a CVLAN UNI that belongs to the same service.

Note

Because Private-VLANs use the two VLAN-IDs, they are configured with to enforce no connectivity between Isolated ports, so every BEB of an E-TREE L2 VSN must use the same VLAN-IDs and it is not possible to re-map VLAN-IDs for this service type. Likewise, if a CVLAN UNI BEB is part of the same E-TREE L2 VSN, it must also use the VLAN-ID corresponding to the Private-VLAN Primary VID.

Fabric Attach

Fabric Attach (FA) complements and extends the Fabric Connect architecture to bring the Fabric services directly to the end-users in the wired and wireless access as well as to the business applications located in the data center.

Note

Fabric Attach is the Extreme naming of IEEE 802.1Qcj Auto Attach standard.

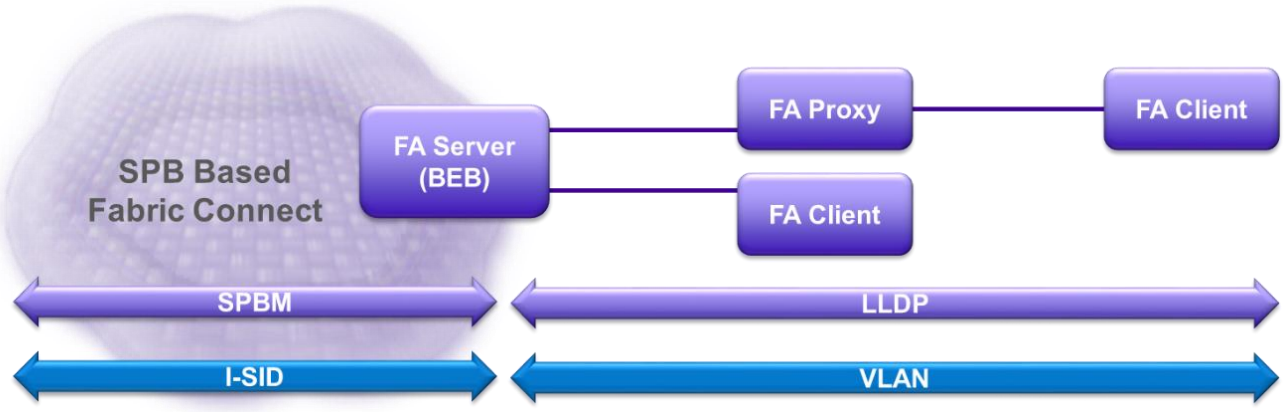


Figure 34 Fabric Attach Ecosystem

LLDP is the mechanism by which end-stations and networking devices communicate and discover from each other about their neighbors and capabilities and Fabric Attach extends LLDP by adding new Fabric Attach TLVs.

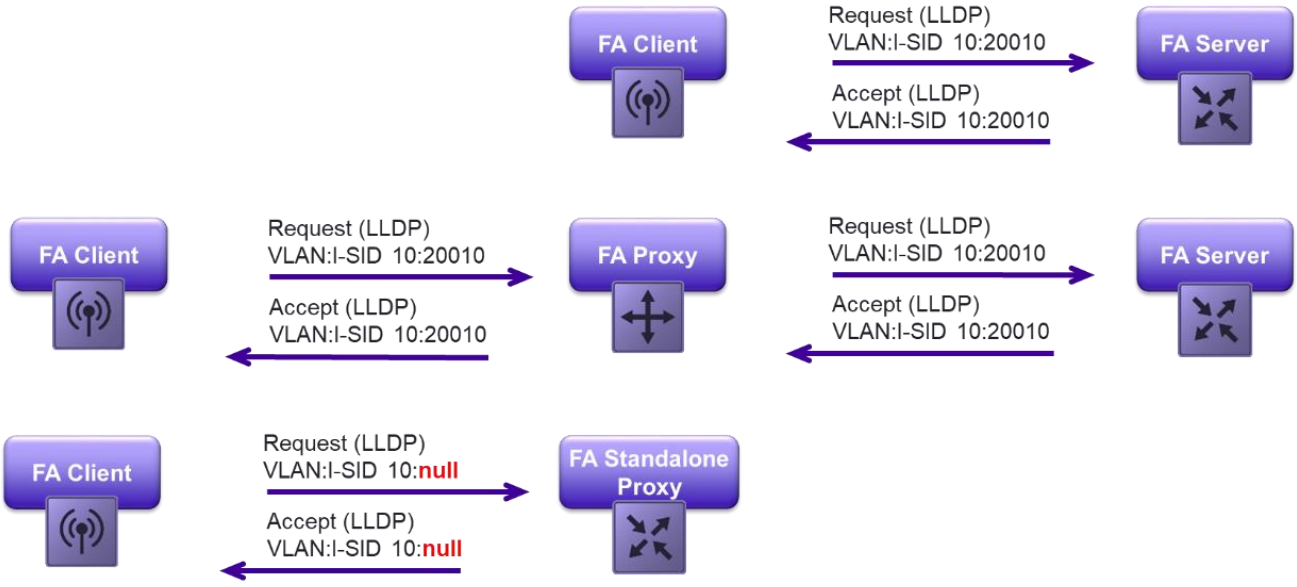


Figure 35 Fabric Attach Model

The Fabric Attach model defines four possible roles, as depicted in Figure 35.

- FA Server: This node is by definition an SPB BEB node that is part of Fabric Connect and is the node in charge of requesting and signalling any Fabric Attach L2 VSN I-SID requests via SPB's IS-IS protocol.

  Tip

Extreme Networks VOSS VSP platforms provide FA Server support, including in redundant MLAG SMLT clustering configurations.

ERS 5900/4900/4800 platforms can support FA Server but do not support MLAG / SMLT clustering.

- FA Client: This is an end-device supporting Fabric Attach LLDP extensions and is therefore able to automatically negotiate directly with the fabric access the terms of its virtual service connection. Examples of FA clients are: Extreme Wireless APs, Open vSwitch (OVS)[7], Extreme's Defender for IoT, video surveillance cameras from Pelco and Axis, and Industrial Ethernet switches from third-party vendors such as MicroSens, Hirshman, and Nexans.
- FA Proxy: This is a non-SPB Ethernet access switch acting as a Fabric Attach Proxy (Client relay) function between the FA Server and FA Clients. In the Extreme Fabric Connect reference model, this is the recommended configuration mode of the wiring closet switch, which is typically dual homed via MLT link aggregation into an SMLT cluster distribution layer acting as one logical FA Server. An FA Proxy switch also has full FA Client functionality in that it can also initiate FA signalling with the FA Server.

  Note

  Only Extreme Networks ERS and ExtremeXOS platforms support FA Proxy mode
- FA Standalone Proxy: FA Standalone Proxy is a mode where the FA Proxy switch operates without the presence of an FA Server. This mode is only useful in situations where the wiring closet access switch is deployed in a non-fabric architecture or in cases where the distribution layer is not capable of providing the FA Server functionality. In some cases, use of FA Standalone Proxy mode can be useful for migration purposes. Either way, the FA signalling can only request VLAN-IDs and the I-SID value must be left at 0 (null).

  Note

  Only Extreme Networks ERS platforms support FA Standalone Proxy mode.

Fabric Attach in the data center will usually consist of Open vSwitch (OVS) acting as an FA Client and the DVR leaf ToR switch acting as the FA Server (with or without SMLT clustering).

FA Proxy switches will typically be in the wiring closet and in some cases in the data center on 1 gigabit ToR access switches which would then be connected into either DVR leaf node(s) or DVR controllers acting as FA Servers.

Caution

Note that FA Proxy chaining is not supported. A FA Proxy switch must be directly connected to an FA Server.

FA Element Signalling

Fabric Attach is as much about attaching end-stations and applications to I-SID based services as it is about discovering the capabilities of the neighboring Fabric Attach enabled elements.

Fabric Attach extends LLDP with two vendor-specific TLVs, the first of which is the Element TLV. Every Fabric Attach enabled device, whether it is acting as an FA Server or FA Proxy or FA Client, will advertise information about itself via the Element TLV. The information provided includes the FA Element type, the desired or operational 802.1Q tagging mode of its Ethernet interface, management VLAN information, and a Fabric Attach System ID to uniquely identify the FA Element. Use of the discovered FA Element type

7 Open vSwitch (OVS) version 2.4 and above.

becomes particularly important in implementing onboarding policies for newly connected devices and achieving the goal of an automated edge fabric.

Therefore, if an ExtremeWireless AP FA Client is detected, its connecting port can automatically be configured for 802.1Q tagging (or perhaps in untag-PVID-only mode which will ensure that the AP's management traffic remains untagged) and be added to the local management VLAN (or other Fabric-wide I-SID) such that the AP can obtain an IP address via DHCP and connect into the WLAN Controller / WLAN Management system to obtain its final configuration. Once the WLAN AP has obtained its final configuration, this will include VLANs required for SSID termination, which the AP can then FA signal to be added as required. Likewise, an FA Client video surveillance camera can be automatically placed into its final IP multicast enabled L2 or L3 VSN, on an untagged port, where it can obtain its IP address via DHCP.

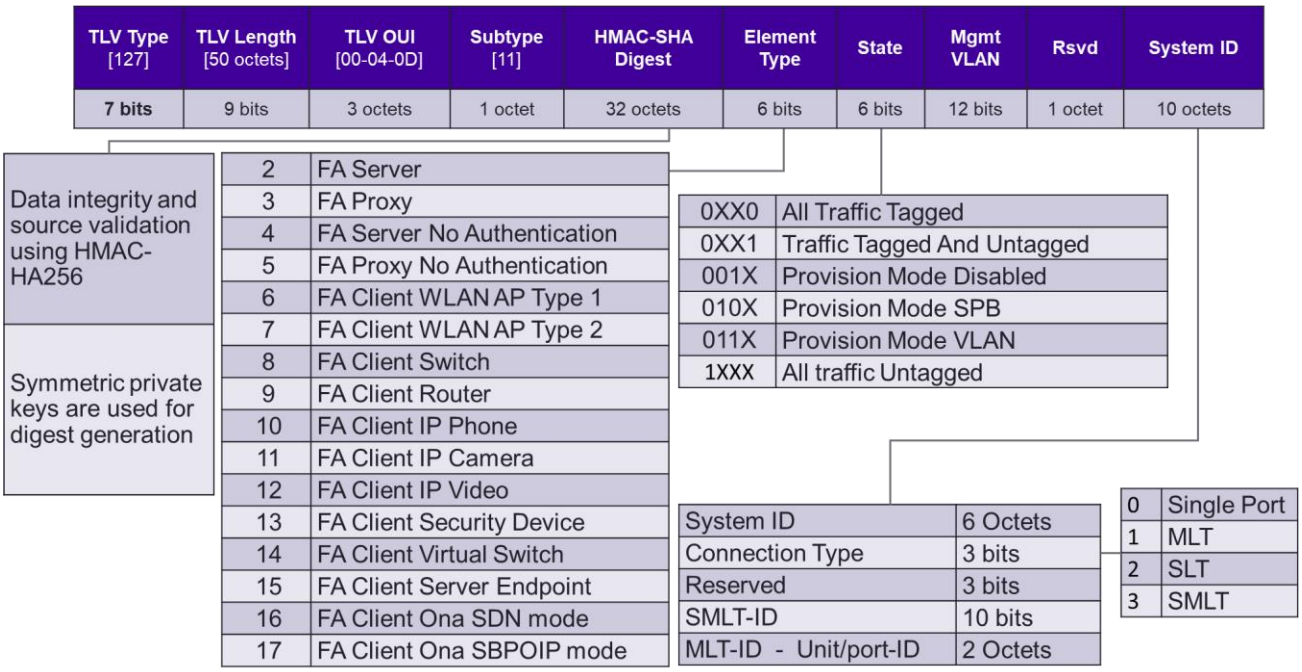


Figure 36 Fabric Attach LLDP Element Signalling TLV

The FA system ID includes the element's MAC address as well as port/interface type used by the FA element. The inclusion of the MAC address increases the security of the solution where 802.1X NAC is used, as it becomes almost impossible for an attacker to remove an FA Client device and attempt to spoof the device's own MAC inside LLDP (if FA message authentication is used). Inclusion of the port/interface used provides an elegant way for detecting a redundant active-active connection into an FA Server using either link aggregation (MLT) or where the FA Server is constituted by MLAG SMLT cluster.

Tip

An SMLT cluster FA Server will use its SMLT-Virtual-BMAC as system ID MAC.

A breakdown of the TLV is shown in Figure 36. The authenticity of the information provided is guaranteed by message authentication provided by a HMAC-SHA digest (this will be covered below).

FA Service Assignment (I-SID) Signalling

The other LLDP TLV provided by Fabric Attach is the Assignment TLV where an FA Client or FA Proxy device can request a VLAN-ID & I-SID binding to the FA Server. The assignment status will reflect the state of the service binding which will start off in pending state and will either become active after being accepted or else rejected by the FA Server.

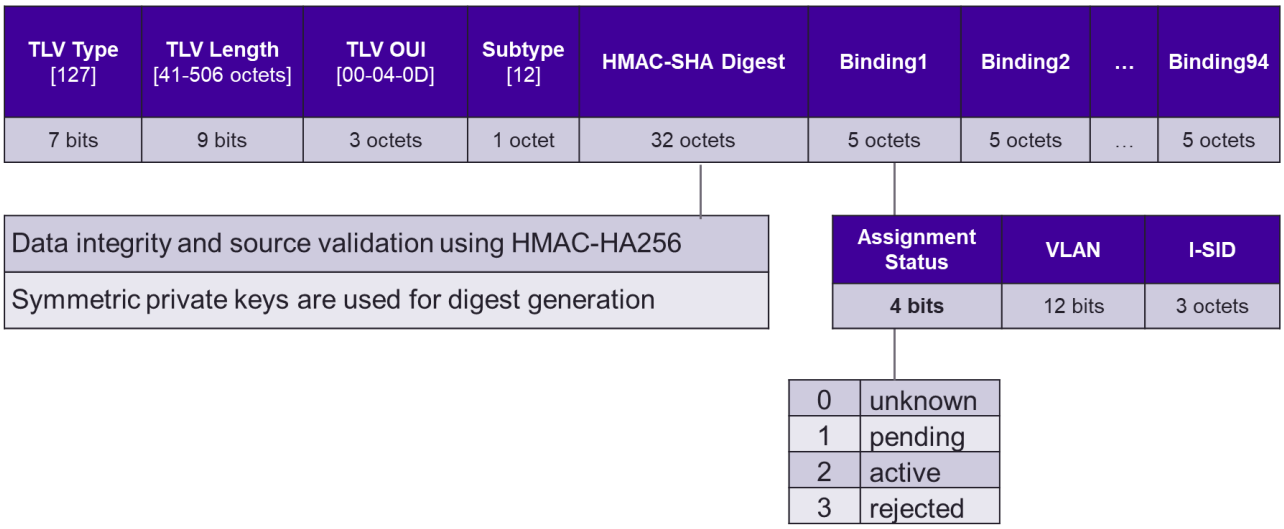


Figure 37 Fabric Attach LLDP Service Signalling TLV

Note

A FA Standalone Proxy node will only process FA signalling TLVs where the I-SID value is null

There are a number of possible ways in which a Fabric Attach I-SID binding can be assigned:

- FA Client expressly asks for one, or more, VLAN:I-SID bindings. The FA Proxy or FA Server will then handle the assignment request. This is typically the case with ExtremeWireless FA Client Access Points where the ExtremeCloud Appliance WLAN controller programs the FA mode configuration in the AP triggering FA assignment requests for the required VLAN:I-SID bindings.

Figure 38 FA Client Requests VLAN:I-SID Binding

- FA Proxy or FA Server, configured with Zero Touch FA Auto Client Attach, detects that an FA Client of type X has been connected to an access port. If a Zero-Touch-Client policy exists for the FA Client type, this will determine the VLAN:I-SID binding to which the device will be attached. If the FA Client was connected to an FA Proxy, this will request the service binding from the FA Server; if instead the FA Client is directly connected to the FA Server, the FA Server can automatically provision the VSN on the corresponding port. This is a typical approach for onboarding video surveillance IP cameras.

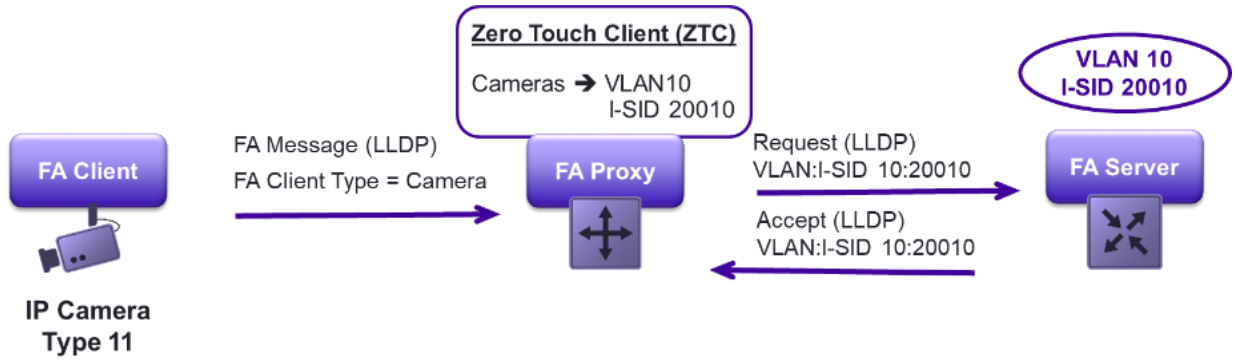


Figure 39 FA Zero-Touch-Client Assigns VLAN:I-SID Binding to Discovered FA client

Note

Only Extreme Networks ERS and VSP platforms support FA Zero Touch Client.

- FA Proxy or FA Server have 802.1X NAC enabled on the access port where the FA Client is detected. RADIUS MAC-based authentication is performed, which is augmented with new Fabric Attach inbound RADIUS attributes that provide information about the FA Client type, FA Client ID, and FA operational mode of the authenticator. If the policy decision is to authorize the FA Client, then one, or more, VLAN:I-SID bindings can be sent back as RADIUS attributes to authorize the FA Client access port with. The FA Proxy, or FA Server, will then assign those VLAN IDs to the access port and handle the FA signalling back to the FA Server.

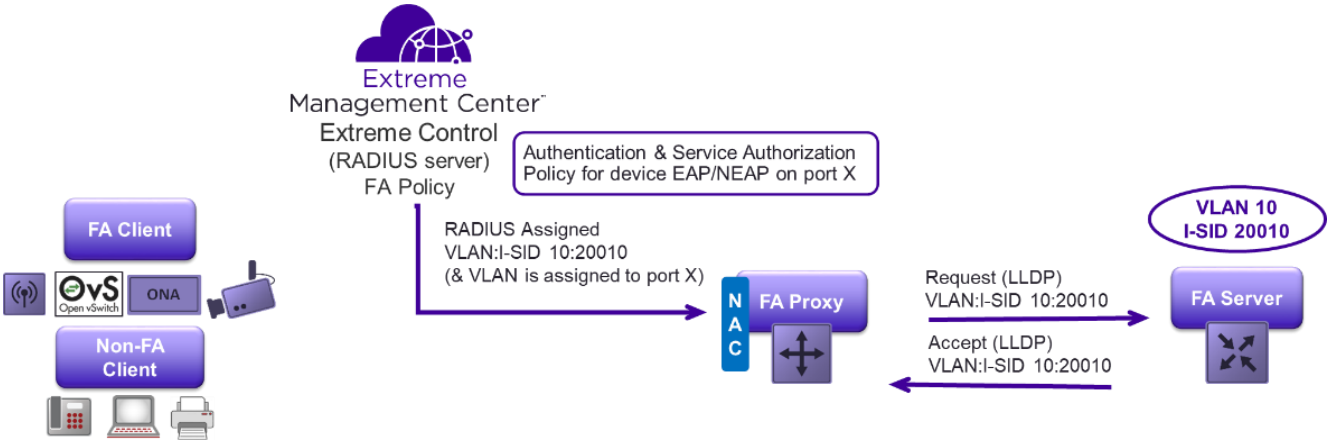


Figure 40 VLAN:I-SID Binding is RADIUS Assigned via NAC

- Manual configuration on the FA Proxy device. A VLAN object can be created on the FA Proxy switch and assigned to the relevant I-SID. The FA Proxy will then signal this VLAN:I-SID binding back to the FA Server. This provides the equivalent CVLAN UNI functionality that would have been available if the access FA Proxy node had been deployed directly with SPB Fabric Connect.

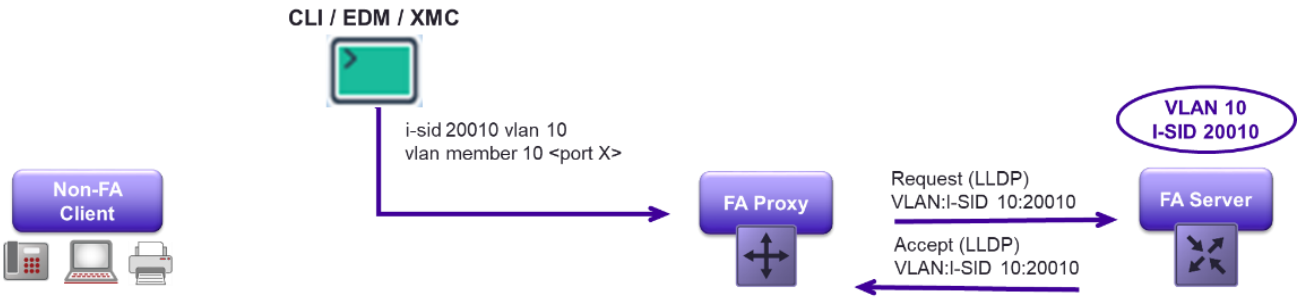


Figure 41 VLAN:I-SID Binding via Manual Configuration

In all cases, the VLAN-ID requested via FA signalling remains locally significant to the FA Server access port, where all FA bindings are implemented using Switched-UNI, and only the I-SID is globally significant.

If the FA Server is providing IP routing and or IP Multicast support for the L2 I-SID, then a platform CVLAN must be provisioned and bound to the same L2 I-SID on the FA Server.

Tip

Technically the CVLAN-ID does not need to be the same as the FA signalled VLAN-ID, since FA is running on a Switched-UNI port. However, for the sake of clarity it makes sense to use the same VLAN-ID for both.

A FA Proxy or Client device can originate as many service assignment bindings as it has VLAN:I-SIDs to request from the FA Server. The upper limit of how many VLAN:I-SID bindings can be requested via Fabric Attach is determined by the number of service bindings that can be packed into the Assignment TLV, which, like all LLDP TLVs, is limited to a maximum value field size of 511 bytes. That limit works out to 94 VLAN:I-SID bindings. Only one Assignment TLV can exist in an LLDP PDU.

Note

An FA Server supports a maximum of 94 VLAN:I-SID bindings on each Fabric Attach port/interface.

Finally, just like the FA Element TLV, the authenticity of the VLAN:I-SID bindings requested via the Service Assignment TLV is also guaranteed by message authentication provided by a HMAC-SHA digest (this will be covered below).

FA Message Authentication

As already mentioned in the preceding sections, both the FA Element and Assignment TLVs can be message authenticated. This is achieved via an HMAC-SHA256 algorithm that is used to calculate the message authentication code (i.e., digest) involving a cryptographic hash function (SHA-256) in combination with a secret Pre-Shared Key (PSK). The resulting digest is conveyed in the appropriate field within the TLV.

Every Extreme Networks Fabric Attach device comes pre-configured with a secret pre-shared key that is not publicly available. This pre-configured secret FA key can then be complemented or replaced with a customer-defined secret FA key.

When FA message authentication is enabled, which it is by default, the (pre) configured FA key is used to generate a HMAC digest that is included in FA TLVs. On the node receiving the FA TLV, the HMAC digest is recomputed for the TLV data and compared against the digest included in the TLV. If the digests are the same it means that both the device which generated the TLV information as well as the device reading the received TLV share the same key. Hence, the data is considered valid. If not, the data is considered invalid and is “silently” ignored.

Note

FA message authentication does not encrypt and hide any of the information conveyed by the FA LLDP TLVs, and thus all information conveyed in the LLDP PDU will still be visible in a packet capture. What FA message authentication ensures is that the FA signalled information is only processed, and acted upon, if the authentication was successful.

The importance of FA message authentication is in securing the Fabric boundaries and preventing unauthorized devices pretending to be valid Fabric Attach elements in order to gain unauthorized access to VSN L2 I-SID services. This becomes particularly important in IoT deployments where the risk is that an attacker will attempt to remove an IoT device in order to gain access to the network. Not only will FA

message authentication prevent such a rogue device from getting access to Fabric VSN services, but even those services which were accessible to the displaced IoT device will be retracted as the rogue device will not be accepted as a valid FA Client device.

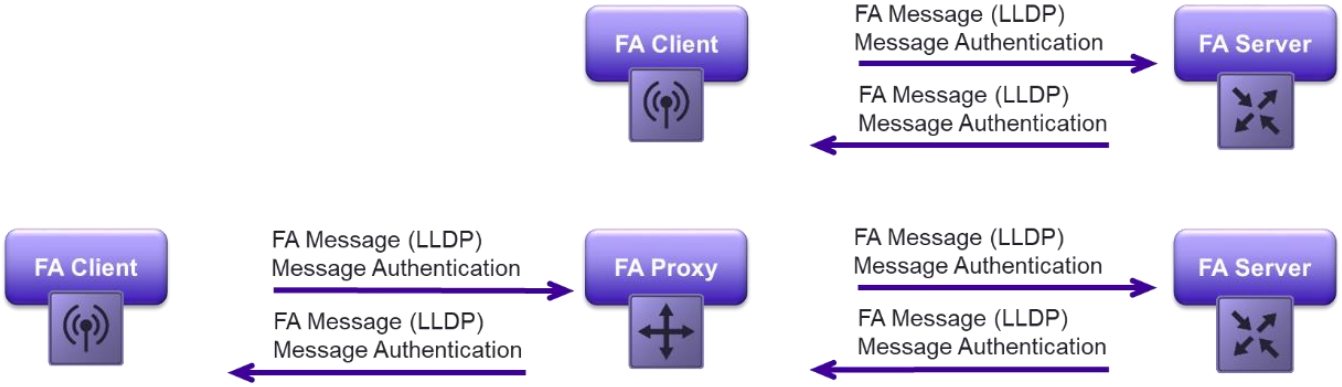


Figure 42 Fabric Attach Message Authentication

In environments where Network Access Control (NAC) is deployed at the network access layer, FA message authentication further hardens and enhances any MAC-based RADIUS authentication performed against recognized FA Client devices. Let's take the example of a video surveillance deployment where security is a key aspect and all video surveillance cameras must be successfully authenticated and authorized onto the surveillance network. An EAP-TLS deployment is generally considered as being the most secure approach, yet this approach is often avoided due to the complexity of managing PKI and rolling out certificates to every deployed video camera.

For this reason, MAC-based RADIUS authentication is often preferred to EAP-TLS because it requires less effort to put in place and manage. Yet it is fairly easy for an attacker to get around MAC-based RADIUS authentication by simply spoofing the MAC address of whatever IoT device is being displaced (and most devices clearly label their MAC address on a handy sticker). In a video surveillance deployment where the video cameras are FA client enabled, the FA Proxy or FA Server switch acting as RADIUS authenticator can complement normal MAC based authentication with Fabric Attach inbound RADIUS attributes, such as FA Client ID or FA Client Type.

Because these FA attributes are protected by FA message authentication, the NAC security is greatly increased without the logistical burden of a full EAP-TLS deployment. A simple rule on the RADIUS server to verify that the FA Client ID matches the device's reported MAC address is sufficient to ensure the authenticity of the FA Client device. Now an attacker attempting to displace a video surveillance camera to obtain access to the network would still be able to spoof the camera's own MAC address but would find it much harder to spoof the same LLDP FA TLVs that the camera was originating. Further, without knowledge of the FA pre-configured secret key, it would fail to be recognized as a valid FA Client by the Fabric access switch and consequently refused network access by the RADIUS server.

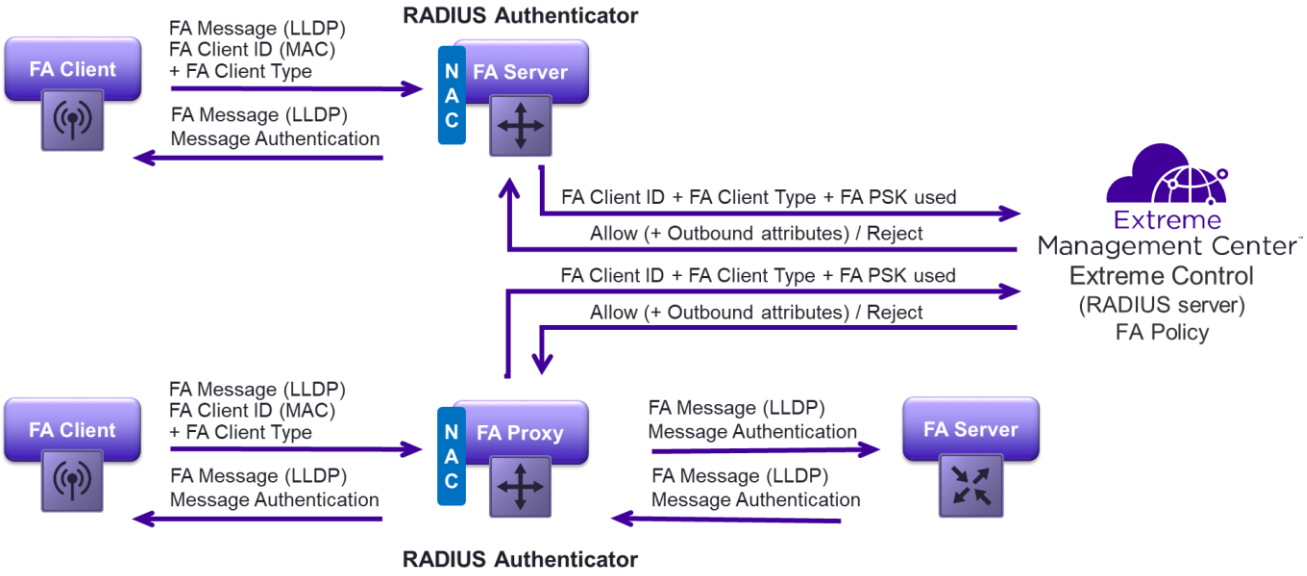


Figure 43 FA Message Authentication Hardening RADIUS MAC-Based Authentication

For many deployments, the Extreme Networks FA message authentication pre-configured key can be used as long as it remains secret and cannot be found in the public domain. This has the advantage that onboarding newly deployed FA Client devices can be done without any prior pre-staging of those devices to pre-set on them the customer-defined FA secret key.

Where it is desired to replace the Extreme Networks pre-configured secret key with a user-defined secret key, this can be easily provisioned across all FA Server and FA Proxy devices already provisioned on the network. To avoid having to pre-stage FA Client devices with the same user-defined secret key prior to deployment in the network, the FA access switches can be configured to support dual FA message authentication keys. In this mode, a user-defined secret key is provisioned for already deployed devices and the Extreme pre-configured secret key can still be used for FA Client onboarding purposes.

That means a factory shipped Extreme Networks WLAN AP can be zero touch provisioned by simply plugging the AP directly into the production network where it will be onboarded using the Extreme Networks pre-configured secret key, but which will immediately be overwritten with the correct user-defined secret key as soon as the AP obtains its final configuration. If NAC is in use, the RADIUS server can accommodate the onboarding policy for a given FA Client type as the FA Proxy or FA Server switch, acting as RADIUS authenticator, will also include an FA RADIUS attribute specifying which FA message authentication key was used by the FA Client attempting to gain access.

Note

Currently only Extreme Networks ERS platforms have dual FA message authentication key support.

FA Zero Touch Provisioning

The main focus of Fabric Attach is to attach end users and applications or IoT devices to Fabric I-SID services and in the examples covered above we have seen that this is achieved by either the FA Client or the FA Proxy requesting service attachment to some L2 I-SID and VLAN pair towards the FA Server.

However, the same Fabric Attach LLDP mechanisms can also be leveraged to ease provisioning of FA Proxy access switches and FA Client management connectivity in the reverse direction.

FA Server switches will typically be in a core/distribution role and from a provisioning perspective must have already been deployed before any FA Proxy or FA Clients are connected to the network. Part of the FA Server provisioning configuration will determine what FA management VLAN should be put in place and advertised to attached FA Proxy and FA Client devices. In some cases, the FA Server will be acting as default gateway for that management VLAN, while in other cases the FA Server will simply act as an L2VSN BEB for accessing a fabric-wide management segment.

Note

In both cases, the FA Server will need to associate to a Fabric Attach enabled port a management I-SID. This I-SID will be used to either map to a local platform VLAN where an IP Gateway interface is created or it can simply be the I-SID of a fabric-wide L2VSN management segment.

The Fabric Attach Element TLV optionally allows the FA Server to announce the management VLAN to downstream FA Proxy switches and FA Clients. From Figure 44, note that only the management VLAN ID is advertised by FA messages. This is because the corresponding I-SID for the management VLAN only has significance on the FA Server and is of no use to FA Proxy or FA Client devices.

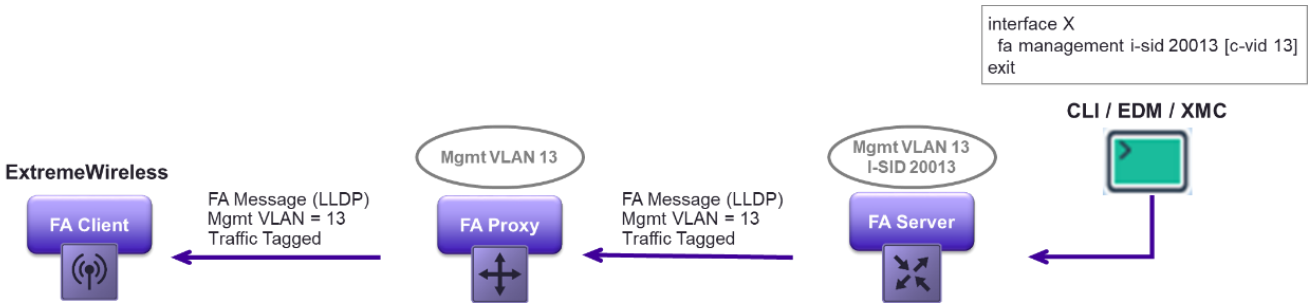


Figure 44 FA Signalling of Management VLAN from FA Server

The advertised FA management VLAN can take one of the possible values listed in Table 11.

In addition to the FA management VLAN, every Fabric Attach device is able to indicate whether the traffic that it will transmit on its FA port is either all tagged, all untagged, or a combination of both tagged and untagged using the Element TLV State field.

Table 11 – Fabric Attach Management VLAN ID Values

FA Mgmt VLAN ID	Description of management VLAN
0	No management VLAN exist; the FA Server port was not configured with one.
1..4094	A tagged management VLAN exists and the FA Server will only accept q-tagged traffic for it on the corresponding VLAN ID.
4095	An untagged management VLAN exist and the FA Server will accept untagged traffic for it.

Consequently, a freshly deployed wiring closet switch can automatically:

- Discover the FA Server.
- Become an FA Proxy device.

  Tip

  Extreme ERS and ExtremeXOS platforms operate in FA Proxy mode by default.

- Configure its uplink ports to the FA Server for VLAN tagging.

If a management VLAN is advertised by the FA Server, it can automatically:

- Create the management VLAN
- Assign the management VLAN on the uplink to the FA Server.
- Obtain an IP address via DHCP.

The objective is to ease deployment of the FA Proxy wiring closet switches which can be deployed with minimal initial configuration (if not in factory defaults) and obtain their final configuration once discovered and onboarded via Extreme Management Center.

Caution

Current FA Zero Touch Provisioning capability on FA Proxy devices (ERS and ExtremeXOS) does not automatically create an MLT/LAG when the discovered FA-Server is an SMLT cluster. In this case, only one of the uplinks will be used until the final config is pushed down from the management station which will then have to include the MLT/LAG config.

Caution

VLACP configuration, if in use on the FA Server ports, will also not be automated by FA. In this case, it is necessary to pre-configure VLACP on the FA Proxy switch before it can even communicate with the rest of the network.

Fabric Attach will also advertise the management VLAN to any attached FA clients, but whether or not this information is used by the FA Client will depend on the FA Client type and the desired deployment model. The scope of the management VLAN is to be able to access and manage all devices which are VLAN aware and part of the network infrastructure. If the FA Client is an ExtremeWireless AP, it could make sense for those APs to obtain their DHCP IP management addresses via the same management VLAN (which remains separate from any other VLAN the AP uses for wireless users). Whereas if the FA Client is a video surveillance camera, which typically uses a single untagged network connection, there is no use for a management VLAN.

Tip

On Extreme ERS platforms it is possible to disable propagation of the FA management VLAN towards the access ports. Hence the FA management VLAN can still be used by the ERS (as it is advertised by the FA Server), but is not propagated to any FA Client devices.

Even in the case of an ExtremeWireless FA Client it may sometimes be preferred to let the APs obtain their IP addresses via DHCP over a separate VLAN/IP subnet from the one being used by the FA Proxy wiring closet switches (the latter typically have statically assigned IPs while APs always use DHCP).

In the example depicted in Figure 44, as soon as the ExtremeWireless AP boots up it will discover the FA advertised management VLAN and will therefore perform DHCP on that VLAN (using q-tagged frames). If instead no FA management VLAN was discovered, then the AP will perform DHCP untagged and the switch it is connected to will need to be able to handle all AP management traffic as untagged.

The FA Client device can also use the Element TLV State field to indicate to the upstream FA Proxy or FA Server switch whether it intends to transmit all traffic tagged or untagged or a combination of both tagged and untagged. Provided that FA zero-touch is globally enabled, the FA Proxy/Server switch will automatically configure the appropriate tagging mode on ports where the FA Client is connected.

Caution

Currently only ERS access switches are capable of automatically configuring the appropriate tagging mode based on what the FA Client has advertised. This is true whether the ERS is running in FA Proxy or FA Server mode.

On the Fabric Attach access node (which could be a FA Proxy or FA Server) where FA Client devices are going to be connected, there are also a number of FA zero-touch-options available for use that are designed to ease the onboarding of the FA Client device. Each of these options can be configured globally for all possible FA Client types or individually for a specific FA client type. That is, it is possible to activate one option for FA Client Wireless APs and a different option for FA Client IP cameras.

The following FA zero-touch-options are available:

- auto-port-mode-fa-client: When this option is activated for certain FA Client types, whenever an FA client of that type is discovered on an access port, the access port is automatically pre-configured for EAP/NEAP in mode Multiple-Hosts-Single-Authentication (MHSA). The FA Client will thus need to authenticate against a RADIUS server using either EAPoL or RADIUS MAC-based authentication (NEAP). In the latter case, the authentication is rendered more secure by inclusion of FA Client id, FA Client type and FA message authentication pre-shared-key (PSK) used by the FA Client (Extreme default secret key or custom key) inbound RADIUS attributes as described in FA Message Authentication on page 82.

  This option can be a more secure way of authorizing FA Clients onto the network. It can even be used to authenticate wireless APs as the EAP mode is MHSA, which will allow wireless user traffic (from different source MACs) once the AP MAC has been authorized on the port.

  For a given FA Client type, this option is incompatible with options auto-pvid-mode-fa-client, auto-mgmt-vlan-fa-client, and auto-client-attach.

- auto-pvid-mode-fa-client: When this option is activated for certain FA Client types, whenever an FA client of that type is discovered on an access port, the access port will be automatically assigned to the FA management VLAN. The port PVID is also set to the FA management VLAN ID. This is required in case the FA Client requested, via the FA Element TLV, both tagging and untagged traffic which would result in the FA access port being automatically configured as untagPvidOnly.

  This option can be useful for onboarding FA Client devices that are VLAN aware (and are thus likely to signal for additional VLAN:I-SIDs), but for which management traffic needs to remain untagged on the same access link.

  For a given FA Client type, this option is incompatible with options auto-port-mode-fa-client, auto-mgmt-vlan-fa-client, and auto-client-attach.

- auto-mgmt-vlan-fa-client: This option is almost identical to the auto-pvid-mode-fa-client option above, in that the access port will be automatically assigned to the FA management VLAN, but with the exception that the PVID on the port is not changed.

This option can be useful for onboarding FA Client devices that are also VLAN aware (and are thus likely to signal for additional VLAN:I-SIDs) but where all communication, including over the FA management VLAN, needs to remain tagged.

For a given FA Client type, this option is incompatible with options auto-port-mode-fa-client, auto-pvid-mode-fa-client, and auto-client-attach.

- auto-client-attach: When this option is activated for certain FA Client types, whenever an FA client of that type is discovered on an access port, the access port will automatically be assigned as an untagged member of a VLAN:I-SID from the FA zero-touch-client configuration profile, if a profile exists for that FA client type. FA Zero-Touch-Client (ZTC) are policies that are pre-configured on the FA access switch and simply map an FA Client type to a single VLAN:I-SID.

  This option can be useful for onboarding non VLAN aware FA Client devices (which will typically not request any VLAN:I-SID bindings themselves) onto their destination VSN. The FA access switch will automatically assign the VLAN on access port and FA signal the VLAN:I-SID binding to the FA Server.

  For a given FA Client type, this option is incompatible with options auto-port-mode-fa-client, auto-pvid-mode-fa-client, and auto-mgmt-vlan-fa-client.

- auto-trusted-mode-fa-client: When this option is activated for certain FA Client types, whenever an FA client of that type is discovered on an access port, the access port will be automatically made QoS trusted. This is useful if the IP DSCP QoS markings from the FA Client can be trusted. A typical use case is with FA Client IP surveillance cameras which, once authenticated and authorized onto the network via the auto-port-mode-fa-client option, can be trusted to transmit video surveillance multicast with a certain IP DSCP and it is desired for that traffic to have a better QoS than best effort.

  For a given FA Client type, this option is compatible with all other zero-touch-options.

Caution

All the above FA zero-touch-options are currently only available on ERS access switches. This is true whether the ERS is running in FA Proxy or FA Server mode.

The exception is FA Zero-Touch-Client (ZTC), which is also available on VOSS FA Server nodes (which can be configured directly without the use of FA zero-touch-options).

IP Multicast Enabled VSNs

Initial Considerations

Multicast allows information to be efficiently forwarded from a source device to many receivers who have expressed an interest in joining the multicast group. There are many examples of applications that benefit from or require multicast support, such as Video Surveillance, IPTV, Video Conferencing, Financial market data distribution on trading floors, and Ghost distribution of backup disk images to multiple computers simultaneously.

Every multicast packet that the source generates, the network must be able to replicate to every receiver in the group. This should be done in an efficient manner such that packet replication occurs at every branch in the multicast delivery tree and delivery to each and every receiver is along the shortest path towards that receiver. This is exactly what Fabric Connect does.

Note

Without Multicast support, the only other alternatives are to use Unicast or Broadcast.

Using Unicast on the source to send the same information to many receivers is inefficient and does not scale because the source device has to generate the same packet as many times as there are receivers, which can quickly exhaust the processing capabilities of the source device as well as congesting its connection into the network.

Using Broadcast is only an option within a Layer2 segment (VLAN) but is also inefficient as the information is flooded to all devices in the segment whether or not they have registered to receive it.

IP Multicast Over SPB

Shortest Path Bridging is the only networking technology to date that has been engineered to properly handle multicast from the ground up. This contrasts with IP and MPLS, where IP Multicast was retro-fitted as an afterthought and is the reason why the IP multicast control planes defined for operation over IP (PIM), and MPLS IPVPNs (draft Rosen) as well as EVPN are so complex and inefficient.

The Extreme Networks Fabric Connect SPB Fabric is able to abstract IP Multicast streams into dedicated stream-specific multicast trees that are completely independent from the IP routing table used by IP unicast traffic, while still constraining the multicast traffic to the virtual network (VSN) to which it belongs. That is, the dynamic service id I-SID allocated to an IP Multicast stream from a source in a given VSN can only be received by receivers located in the same VSN.

Note

In fact, with SPB it would be very easy to unconstrain IP Multicast traffic from the VSN to which source and receivers belong. In this way, it would be possible for receivers to add themselves to a multicast stream tree originating from a source located in a different VSN. However, Extreme Networks has not yet productized this capability.

The default IP multicast mode of operation for any VSN type is disabled, which will result in no IP Multicast being forwarded or receivers registered within the VSN. Once an SPB VSN has been activated for IP Multicast the BEBs will start registering multicast stream senders and receivers independently.

Upon receiving an IP multicast packet on a UNI port, the BEB will record the Class D Group Address of the stream, the Source IP of the Sender, and the VSN (I-SID) membership of the source. It will then dynamically allocate to the stream an I-SID from a reserved range.

Tip

In the Extreme Networks Fabric Connect SPB implementation, each BEB reserves I-SIDs in the range 16,000,001 – 16,600,000 for IP Multicast streams. Each allocated I-SID in that range defines one unique IP Multicast stream (1 Group address + 1 Source IP address) rooted at that BEB node. I-SIDs in that range are not allowed for use when creating L2 VSNs or L3 VSNs.

Note

The same I-SID, say 16,000,001, on different BEBs designates a completely different IP Multicast tree, which could well have been dynamically assigned to a different multicast group address and even to a different VSN.

Tip

The same Group Addresses (as well as Source IP addresses) can be reused in different VSNs.

The BEB will then IS-IS announce to the Ethernet Fabric the existence of the IP Multicast stream using an IP Multicast TLV containing the above-recorded information.

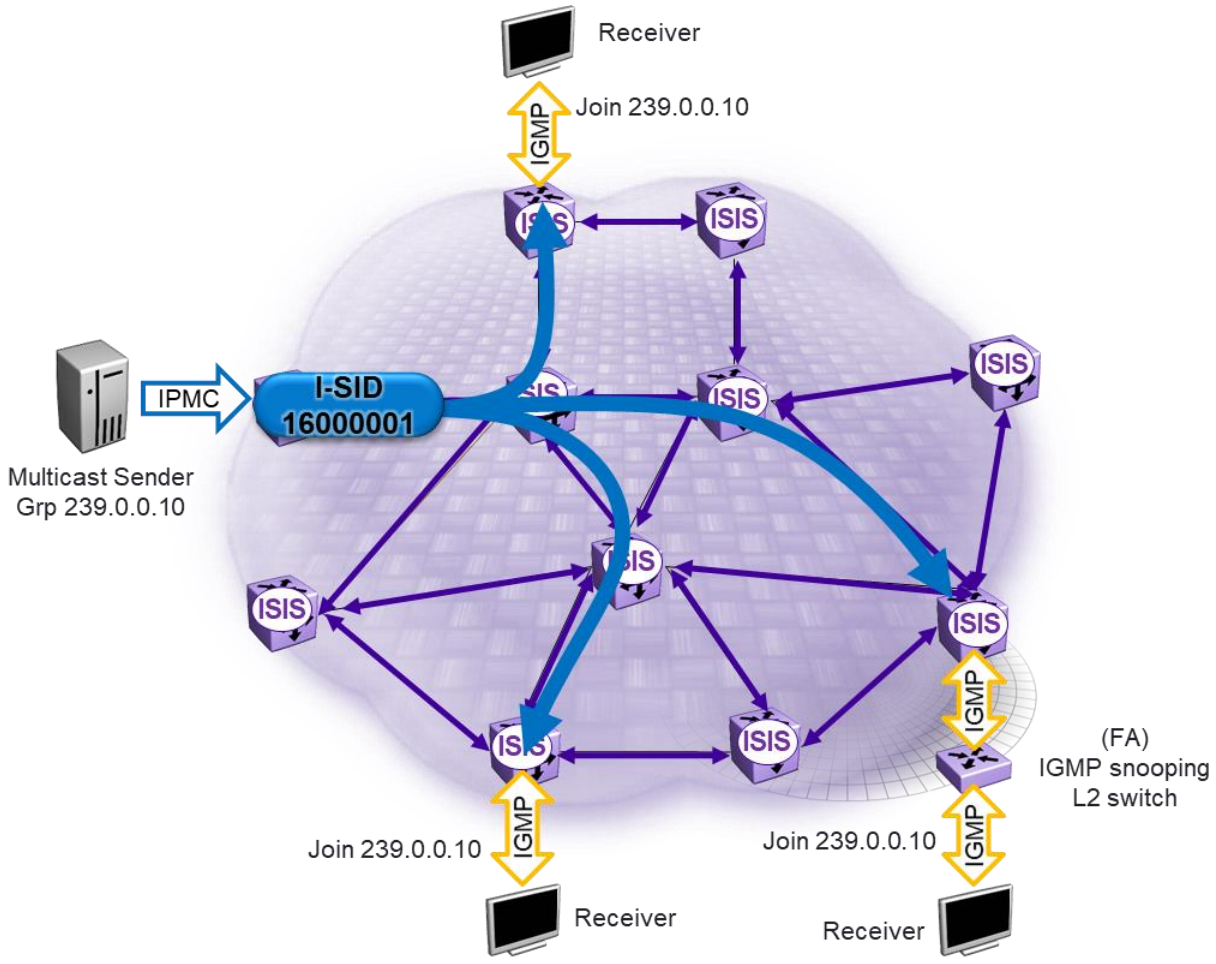


Figure 45 IP Multicast with SPB

Tip

In a design where the source is SMLT connected to two ingress BEBs, both BEBs forming the SMLT cluster will IS-IS announce availability of the IP multicast stream.

In the absence of receivers, the SPB Fabric does not set up any multicast tree and the ingress BEB silently discards the multicast packets it receives from the source.

On the receiver side, the protocol used for registering receivers is still IGMP, which is automatically enabled on the BEB VSN segments where SPB IP multicast was activated. On these segments the IP Multicast enabled BEB will become the IGMP querier.

Tip

In the Extreme Networks implementation, IGMP versions 1 to 3 are supported. By default, IGMP version 2 is used. By using IGMPv3, it is possible to support Source-Specific IP Multicast over SPB.

Tip

External IGMP queriers and hence external IP multicast routers, should not exist on Fabric Connect access segments enabled for IP multicast. Otherwise a single IGMP querier will be elected and Fabric Connect multicast may not work as expected.

Note

The main benefit of using IGMP version 3 is that receivers can register for source-specific groups (i.e., instead of just requesting to join Multicast stream 239.0.0.10 as IGMP v2 would do, IGMP v3 could ask to join only the Multicast stream 239.0.0.10 sent from source with IP 172.16.10.100), which is more secure in the sense that it becomes harder to try and spoof a multicast stream using a different source IP address. IGMP version 3 was defined to allow these source-specific joins so that it could be used in conjunction with Source-Specific PIM (PIM-SSM), which was an attempt to simplify PIM-SM by removing the Rendezvous Point (RP) functionality.

While this did somewhat simplify PIM, on the other hand, source-specific multicast burdens the application on the receivers with a required prior knowledge of the source IP of the desired multicast streams.

When an IGMP join is received, an entry is added to the IGMP membership table. The IS-IS LSDB is then checked by the BEB to see whether or not a matching IP multicast stream exists somewhere in the Ethernet Fabric. If a match is found, constrained to the same VSN (i.e., the IGMP receiver is on the same VSN as the source of the stream), the BEB will update its I-SID TLV. This will indicate that it wishes to be added to the multicast I-SID tree for the corresponding IP multicast stream rooted at the BEB where the source of the stream is located. If no multicast tree was yet set up, it is now set up. If a multicast tree was already in place, a new branch is added to it.

Tip

The time to set up the SPB multicast tree is in the order of 100 milliseconds, which in the case of video distribution or video surveillance allows the applications to rapidly flick across the available channels without the typical latency observed on traditional PIM IP Multicast deployments. This should be combined with the use of IGMP immediate leave feature.

Tip

In a design where the receiver is SMLT connected to two egress BEBs, the IGMP receiver table is automatically synchronized between the BEBs over the IST (or vIST). If a multicast tree is set up, both egress BEBs will set up a tree, each over one of the available BVLANs. This ensures that SPB IP Multicast traffic can utilize SPB's equal cost shortest paths and multicast stream recovery is sub-second in the event of SMLT BEB or link failure at either end. (Source SMLT BEBs as well as Receiver SMLT BEBs.)

Multicast Services

The SPB IP Multicast functionality described above can be activated on any of the available VSN service types already covered:

- IP Multicast over GRT IP Shortcuts: Both Sources and Receivers can exist in the default Global Routing domain. SPB Multicast is enabled at the IP interface level and need only be activated on the IP interfaces where sources or receivers exist. Also, if L2 VSN segments are part of the IP Shortcuts routing domain and IP interfaces (clearly in the same subnet) exist on their BEB end-points, IP Multicast will automatically be snooped within the L2 VSN as well.
- IP Multicast over L3 VSN: Both Sources and Receivers can exist in VRFs connected to the same L3 VSN. It is not required for the L3 VSN to advertise any IP routes, if the L3 VSN is to be used exclusively for IP Multicast. SPB Multicast is enabled at the IP interface level and need only be activated on the IP interfaces where sources or receivers exist. Also, if L2 VSN segments are part of the L3 VSN routing domain and IP interfaces (clearly in the same subnet) exist on the L2 VSN BEB end-points, IP Multicast will automatically be snooped within the L2 VSN as well.
- IP Multicast over L2 VSN: Sources and Receivers can only exist within the same L2 segment (and will thus have IP interfaces in the same subnet). In this case the SPB fabric is only providing an L2 snoop type functionality. The VLAN used to terminate the L2 VSN service is simply enabled for IGMP snooping. There need not be any IP address configured on any of the BEBs, though it is advisable to configure an IGMP query source IP in case non-SPB L2 IGMP-snooping switches are aggregated into the BEB (otherwise the BEBs will use a 0.0.0.0 source IP in the IGMP Query messages they transmit).

Where L2 VSNs are IP routed into an L3 domain (L3 VSN or IP Shortcuts), as depicted in Figure 46, there are clearly two possible ways to activate SPB IP Multicast on them. On the one hand, the L2 VSN BEBs can simply be IGMP snoop enabled. This will allow IP multicast to be efficiently snooped by SPB within the L2 VSN segment but will result in any IP multicast traffic within that L2 VSN segment being restricted to the L2 VSN alone. In other words, if the L2 VSN is IP routed as part of an L3 domain, no IP multicast traffic will ever be able to enter or leave the L2 VSN segment. This can be useful in some scenarios but typically is not the most useful.

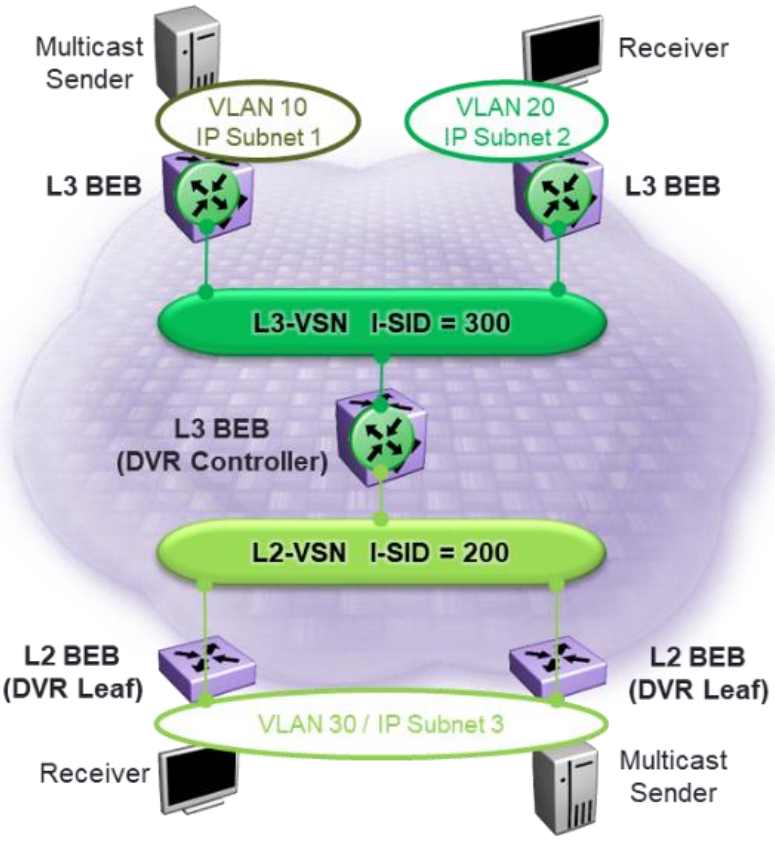


Figure 46 IP Multicast Over L3 VSN that Comprises L2 VSNs

A more common requirement is for the L2 VSN segment to be allowed to perform IP multicast routing to any other BEB in the L3 domain (L3 VSN or IP Shortcuts) as well as efficient snooping within the L2 VSN segment. This is possible but requires, on the L2 BEB, configuration of an IP interface on which SPB Multicast is enabled. This IP interface is thus exclusively used for VRF awareness and the ability perform SPB Multicast rather than being used as a gateway for IP unicast traffic.

Caution

If the L2 BEB must participate in L3 VSN domain SPB multicast, then the platform used needs to be L3 VSN capable. That is, it needs to be a VOSS VSP platform.

An ERS platform can only be used as an L2 BEB and be activated for SPB multicast in an L3 domain if IP Shortcuts are being used.

Tip

If the L2 BEBs in the example topology depicted in Figure 46 were DVR leaf nodes, then it would be sufficient to SPB multicast enable the DVR IP interface on the DVR controller and automatically IP multicast routing will be performed across the L3 domain as well as snooped within the L2 VSN domain.

One common misconception when trying to understand SPB multicast is to assume that IP multicast traffic must follow the same path as IP unicast traffic, as is the case with traditional IP multicast routing such as PIM. In the example illustrated in Figure 46, if IP multicast was activated across the entire L3 VSN domain (including the L2 VSN), then the middle L3 BEB will IP forward all unicast traffic between the top and bottom IP subnets. Yet from an IP multicast perspective, that same L3 BEB will not perform any BEB function for IP multicast traffic being forwarded between the same IP subnets (and indeed does not even need to be SPB multicast enabled). That is because SPB multicast is entirely performed by the ingress and egress BEBs alone by leveraging SPB I-SID delivery trees. If the core node happens to be on the shortest

path for the multicast traffic its involvement in forwarding it would purely be from a BCB transport perspective.

In each of the deployment scenarios described, IGMP Access policies can be applied on the BEB VLAN interfaces to specify which streams will be accepted from a source device as well as which streams a receiver is allowed to join.

Tip

Benefits of SPB IP Multicast over PIM as well as MPLS-VPNs with Rosen Draft:

- All legacy IP Multicast protocols to date (including PIM-SM and PIM-SSM) need to store IP Multicast Source IP + Group IP (S.G) state on all routers in the network. The network as a whole cannot scale to more IP Multicast streams than the Core switches can handle S.G record entries. With Extreme Networks Fabric Connect, S.G record entries are only consumed on the ingress BEB node where each stream is mapped to an SPB multicast tree. The SPB Core only needs to maintain state by consuming one multicast BMAC entry in the BVLAN Forwarding Information Base (FIB). Multicast scaling over an Fabric Connect is thus much more scalable by several orders of magnitude.
- The Rosen Draft architecture allows IP Multicast to work within MPLS IPVPNs. This model however relies on an IP Multicast enabled IGP in the core (typically using PIM-SSM) as well as in the VRFs terminating the IPVPNs and IP (or GRE) encapsulates client side VRF PIM and multicast traffic over the IGP Multicast Distribution Trees (MDT). Ironically, this model cannot even use MPLS labels and is a trade-off between limiting the scalability impact of IP Multicast on the IGP P routers (by multiplexing VPN Multicast streams over IGP MDTs) and providing efficient Multicast, which only delivers the streams where receivers exist. In summary, an architecture using MPLS IPVPNs with Draft Rosen can either be made to scale or to forward IP Multicast in an efficient manner, not both like Fabric Connect.
- SPB IP Multicast can set up multicast forwarding trees in a few hundred milliseconds and offers resiliency of the same order. This is typical of any and all of the VSN service types that can be deployed over SPB, as they all depend on the same underlying IS-IS control plane instance. Suffice to say that with PIM based networks, multicast streams can take over a minute to be restored after a core link or router failure.

SPB Multicast PIM Gateway

Compared to the traditional protocol stack approach, Extreme's Fabric Connect architecture offers a superior networking technology, which delivers powerful virtualization and unmatched IP Multicast support over a single efficient SPB IS-IS control plane. These benefits are available across all VSN services deployed over the Fabric. Yet at the same time the Extreme Fabric has been designed to be able to support all traditional standardized protocols to make sure that Fabric Connect VSNs can always be connected into legacy architectures for co-existence or migration purposes.

Ensuring this goal for IP Multicast has required the development of the PIM Gateway functionality, which allows a Fabric Connect network to forward IP Multicast applications between a traditional IP PIM based network.

PIM-SM has become the de-facto standard protocol for forwarding IP Multicast over traditional IP routed networks. PIM-SSM is another variant of PIM which, although less commonly used, is able to offer a simplified PIM implementation (using no Rendezvous Point – RP), but at the price of making the IP multicast applications, source aware via use of IGMPv3 source-specific joins. The Extreme PIM Gateway functionality is able to operate with either PIM-SM or PIM-SSM, although a different approach is required for each.

Tip

Fabric Connect PIM Gateway is able to interoperate with both PIM-SM and PIM-SSM.

We'll start by looking at the most common approach, dealing with PIM-SM, and then revisit PIM-SSM in a later section below.

PIM-SM is a very complex protocol. At the heart of PIM-SM is the Rendezvous Point (RP), which is an IP router designated to consolidate knowledge of available Multicast Sources and is the focus of all PIM Joins when a receiver expresses a wish to join a stream. When a new Multicast stream is received by a PIM Router, it immediately PIM Registers (encapsulates) that stream towards the RP. Likewise, if a new Multicast Receiver expresses a wish to receive a given Multicast stream, the PIM router where the IGMP entry is added will trigger a chain of PIM *.G Joins, hop-by-hop by every PIM Router along the reverse forwarding path towards the RP.

Forwarding of the multicast stream requires two conditions: the stream sender exists and a multicast receiver has joined the stream. When both these conditions are met, the multicast stream is initially forwarded via the RP. To this effect, if the RP is no longer receiving the stream, it also triggers a chain of PIM S.G Joins, hop-by-hop by every PIM Router along the reverse forwarding path, this time towards PIM Router where the sender is located. Forwarding IP multicast via the RP is not efficient as the path is unlikely to be optimal (except if the RP is close to the sender) and software forwarding is typically used on the RP. It is however necessary to get the stream going to start with, because the last hop PIM router (where the receiver is located) has no idea what the source IP of the requested stream is (the receiver is not aware of the source IP and only asked for the Class D Group address). However, once the stream gets going, the last hop PIM router will be able to infer the source IP of the sender, and at this point (also depending on vendor implementations of PIM), the last hop PIM router will trigger yet again a chain of PIM S.G Joins, hop-by-hop by every PIM Router along the reverse forwarding path, this time directly towards PIM Router where the sender is located. Not surprisingly, PIM-SM suffers from a very high latency when first establishing multicast streams.

The key to interact with PIM-SM is to have direct access to the RP function where all the knowledge of multicast sources in the PIM cloud is concentrated. The IETF has defined the Multicast Source Discovery Protocol (MSDP) precisely for this purpose. This is yet another protocol which needs to be piled on top of the protocol stack but is the only way that separate PIM clouds can be made to interoperate.

The Extreme Fabric Connect PIM Gateway therefore has no other choice but to match the same complexity and implement both PIM signalling and MSDP.

Note

Fabric Connect PIM Gateway requires MSDP in order to interface with a PIM-SM network. Ensure that PIM-SM RP router operating in PIM cloud is able to support MSDP.

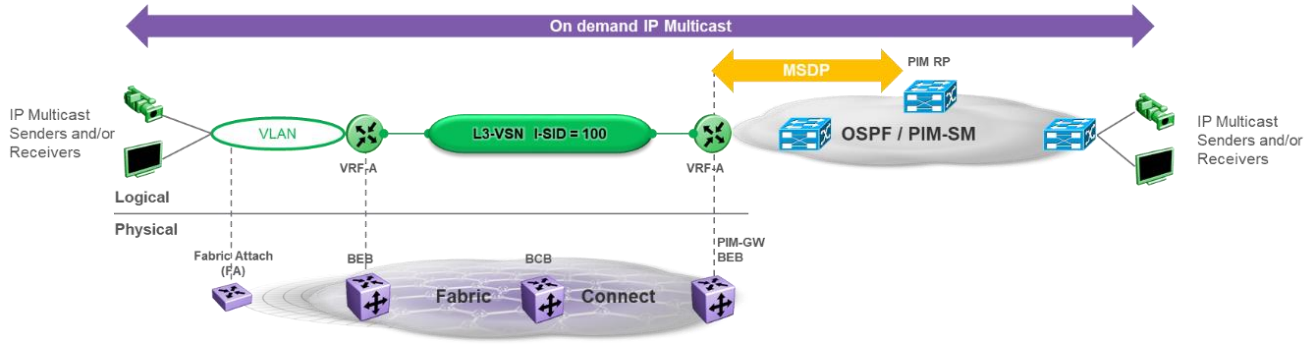


Figure 47 PIM Gateway to Legacy IP Multicast Routed Networks

Figure 47 illustrates at a high level how PIM Gateway interacts with a PIM-SM network. The first striking difference is that on the Fabric Connect side, the PIM Gateway functionality is in reality running off one of possibly many L3 VSNs or IP Shortcut service instances, whereas PIM-SM is almost never virtualized in VRFs (except in Draft Rosen Multicast VPNs).

Note

It is also possible to have two (or more) L3 VSNs (or IP Shortcuts) service instances interface via PIM Gateway to the same PIM-SM cloud, which would allow all those VSNs to be able to forward IP multicast traffic to and from PIM. However, it is never efficient to have the same multicast streams being forwarded over separate VSNs, potentially over the same physical infrastructure. If users across different VSNs are to receive the same multicast streams, a more efficient approach is to designate a single L3 VSN for the Multicast and "leak" the multicast stream on the final access hop towards the receivers in different VSNs using Multicast VLAN Registration (MVR) as explained in Inter VSN IP Multicast on page 99.

MSDP runs as a TCP peering between the PIM Gateway Controller node and the PIM RP router and serves two purposes. The first purpose is to extract all the information about PIM side sources from the PIM RP and announce those same sources into the Fabric Connect IS-IS LSDB by using the appropriate TLV and allocating to each source a dynamically allocated I-SID. Once knowledge of the PIM sources is available in Fabric Connect, a receiver on a Fabric BEB can request for the I-SID multicast tree to be built all the way to the designated PIM Gateway BEB node, which will then trigger a PIM Join on that same node into the PIM cloud in order to start receiving the multicast stream from the PIM side.

Tip

The PIM Gateway BEB immediately has knowledge of the stream's source IP (this was extracted from the RP via MSDP), and therefore is able to initiate PIM S.G Joins directly towards the PIM Router where the source is located.

However, making a Fabric Connect source available to PIM receivers requires a different approach. PIM receivers can only join streams for which the PIM RP has knowledge of the source. Hence the second purpose of MSDP is to prime the RP with information about fabric-side sources. Once this has happened, a PIM Router where a receiver expresses the desire to join a fabric source will trigger the usual PIM-SM mechanics whereby that router will initiate a chain of PIM *.G joins all the way to the RP. The RP, having knowledge of the Fabric source (received from MSDP), will then trigger its own chain of PIM S.G joins all the way towards the PIM Gateway BEB which, in turn, will then add itself to the multicast I-SID tree within Fabric Connect in order to start receiving the stream. As usual, once the stream starts flowing and the last hop PIM router can infer the source IP of the stream, this router will most likely initiate another chain of PIM S.G joins this time directly towards the PIM Gateway BEB, so as to stop forwarding the stream via the RP, which is inefficient.

It should be noted that, unlike with Fabric Connect, PIM requires an underlying unicast IP routing table to be in place, without which it cannot operate. It is essential that PIM be able to look up in the IP routing table routes corresponding to the active RP as well as every IP subnet corresponding to PIM sources. For fabric sources it is therefore essential that the relevant IP subnets must have been redistributed, by the PIM Gateway BEB, into whatever IP routing protocol is used in the PIM cloud (typically OSPF). The same is true in the reverse direction and correct redistribution into IS-IS of PIM-side source IP subnets will also be required for the PIM Gateway Controller function to allocate a PIM Gateway node for the PIM source. In short, any deployment of PIM Gateway will require a correct redistribution of IP routes between the Fabric and PIM clouds as detailed in ISIS IP Route Types and Protocol Preference on page 67.

In many cases it will be desirable to keep control of which multicast sources are made available between the fabric and PIM clouds, rather than making all sources available in both directions, as would be the

default behavior. To this effect, MSDP offers the ability to associate route policies to limit which multicast sources are advertised.

Tip

The Extreme Networks PIM Gateway MSDP implementation allows for both MSDP redistribution and SA-filters. The former allows policies to control, at a global level, what sources are announced by the PIM Controller over MSDP. The latter SA-filters offer similar capability but SA-filters are defined on an MSDP peer basis and can be applied either in the 'in' or 'out' direction, so can be used to filter out sources sent from the PIM RP.

Deployment Model with PIM-SM

The Fabric Connect PIM Gateway functionality offers a flexible architecture that defines two separate functionalities: PIM Controller and PIM Gateway. The PIM Controller runs MSDP peerings into the PIM cloud RPs and assigns both fabric sources and PIM sources to one of the available PIM Gateways. The PIM Gateway is controlled by the PIM Controller and for the multicast sources assigned to it has to perform the following functions:

- Advertise into SPB's IS-IS the appropriate multicast TLV to make the Fabric VSN aware of the availability of PIM side multicast sources.
- Initiate PIM-SM S.G joins into the PIM cloud in order to start receiving multicast streams from PIM sources.
- Process PIM-SM S.G joins received from any PIM Router (including the RP) requesting multicast streams from Fabric Sources.
- Adding itself to the multicast I-SID tree of a Fabric source, in order to forward this into PIM.

These functional blocks can be co-located into a single Fabric BEB node, or can be distributed into different fabric nodes.

Tip

The PIM Gateway role is supported in any of the Extreme Networks VOSS VSP platforms.

The PIM Controller role is only supported on Extreme Networks VSP 8k and 7200 platforms.

Using a single BEB node for both roles or separating the roles into separate fabric nodes is a design choice. In most cases it is probably easier to co-locate the functions within the same node.

Note

A good reason to separate the functions is when the PIM Gateway BEB is an Extreme Networks VSP 4k platform, which does not support the PIM Controller function.

If the PIM Controller and PIM Gateway functions are running on separate nodes and also if more than one PIM Controller or PIM Gateway exist, all communication between them will be using reserved I-SIDs for efficient IS-IS multicast based control plane signalling (much like used with DVR).

The diagram in Figure 48 shows a typical PIM Gateway design featuring redundant interconnects between SPB and PIM as well as redundant PIM Gateways and PIM RPs. The PIM Gateway roles are also shown as running on separate nodes so as to provide a clear overview of how the solution works.

It is true that PIM can operate with only one RP for a given stream. However, it is possible to optimize PIM-SM deployments in such a way that different multicast streams are handled by different RPs (this optimization sometimes will seek to locate the RP function as close as possible to the sources). Furthermore, most PIM-SM deployments are designed such that another PIM Router can take over the RP

role should the active RP fail. All this essentially means that when implementing PIM Gateway, it might be necessary to provision MSDP peerings with more than one PIM router. In the figure, we are assuming that two RPs exist.

Since this is a redundant design, we also have two PIM Controllers and two PIM Gateways. Each PIM Gateway owns one interconnect into the PIM cloud and route redistribution is performed over these links in both directions (IS-IS ←→ OSPF in this case). The PIM Gateway will thus typically be running OSPF over the interconnect. Since both PIM Gateways will be redistributing IS-IS routes into OSPF and OSPF routes into IS-IS, it is important that this redistribution be done properly and according to the guidelines detailed in ISIS IP Route Types and Protocol Preference on page 67. Failure to do so could result in route reflection whereby a PIM source IP route gets redistributed into IS-IS by one PIM Gateway and then redistributed back to OSPF by the other PIM Gateway. This would not prevent PIM Gateway from operating but would compromise the ability to load balance available streams across both interconnects.

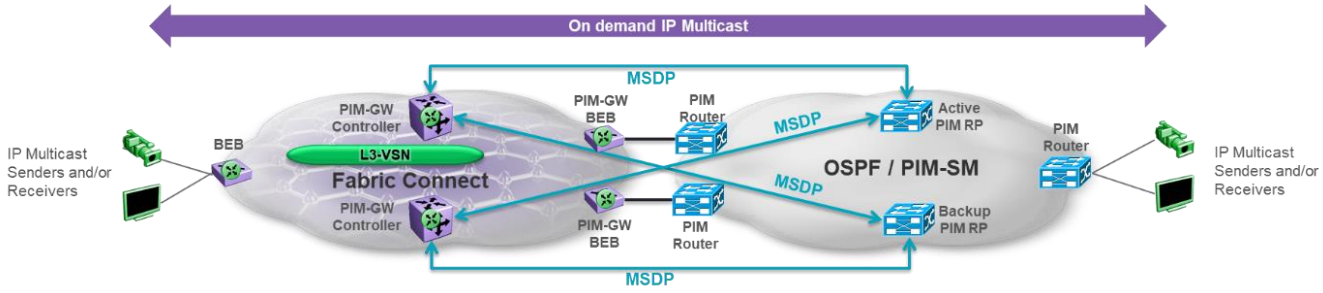


Figure 48 Redundant PIM Gateway Deployment Model

Each PIM Controller should have MSDP peerings into both PIM RPs. This offers the most resilient design, as it ensures that the PIM Gateway functionality can survive failure of one PIM Controller + one RP at both ends. It is worth noting that the MSDP protocol was defined with the intent of interconnecting separate PIM-SM clouds. Since each PIM cloud is in effect an OSPF Autonomous System, the use of BGP would often go hand-in-hand with MSDP. MSDP is therefore able to leverage the BGP AS-path information in order to avoid loops.

However, the typical use case of Fabric Connect PIM Gateway is unlikely to make use of BGP. In the example at hand it is much easier to let the PIM Gateways redistribute directly between OSPF and IS-IS instead of adding BGP to the recipe. The challenge is to ensure that a PIM Controller, when it learns about a PIM source from one PIM RP, does not re-advertise that same source to the other PIM RP. To avoid this, both MSDP peerings defined on each PIM Controller will need to be made part of the same MSDP peer-group.

Tip

Extreme Networks PIM Controller can run MSDP with BGP. However, a simpler approach is not to use BGP and use MSDP peer-groups instead.

Caution

If MSDP with BGP is used, a BGP peering will be required on the PIM Controller in order to be able to inspect multicast sources against BGP path attributes. In practice, this will require the PIM Controller function to be co-located on the PIM Gateway so that a single eBGP peering can be terminated there.

From a control plane perspective, the PIM Controllers will learn about PIM sources from MSDP. The PIM Controllers will then examine which of the available PIM Gateways are redistributing into IS-IS the relevant source IP subnets for the VSN at hand (the PIM Controller can simply inspect its own copy of the IS-IS LSDB). If more than one PIM Gateway is available as candidate, a hash will be applied to allocate the PIM

source to one of the PIM Gateways. The hash ensures a distribution of PIM sources across all the available PIM Gateways. The selected PIM Gateway for a given PIM source will then be responsible for advertising and making available the PIM multicast stream into the IS-IS LSDB using the appropriate multicast TLV. So, for PIM sources it is the PIM Controller that determines the allocation of PIM Gateway.

In the reverse direction, for fabric sources, it is IP routing in the PIM cloud that determines the PIM Gateway to use. Once a PIM S.G join sequence is initiated on the PIM side for a fabric source, these PIM Joins will be forwarded along the reverse path towards the IP subnet of the fabric source, using the IP routing tables of the PIM routers to determine the path. The path will inevitably lead to one or the other PIM Gateway and this will determine which PIM Gateway will then be used to obtain the multicast stream from the Fabric source.

PIM Gateway with PIM-SSM

PIM-SSM is a simplified version of PIM-SM for Source Specific Multicast (SSM). The premise of PIM-SSM is that the multicast receiver application needs to know not just the Class D Group address it wishes to join but also the source IP of the sender, which means IGMPv3 must be used with PIM-SSM.

This means that the function of the PIM RP is no longer needed, as the PIM router where the receiver is located can simply issue PIM S.G joins directly towards the requested source IP.

Use of Fabric Connect PIM Gateway with a PIM-SSM network is also possible and, in some ways, is simpler since there is no need to use MSDP because there is no RP in the PIM-SSM network.

Provided that the Fabric source IP subnets have been correctly redistributed into the PIM-SSM IGP (typically OSPF) by the PIM Gateways, in one direction things will just work out of the box. A PIM-SSM router issuing PIM S.G joins towards a fabric source will end up hitting one of the PIM Gateway BEBs, which can then request the Fabric multicast stream and forward it into the PIM-SSM network.

However, in the reverse direction, the Fabric IS-IS LSDB needs to be aware of the PIM-SSM sources before any BEB can add itself as a receiver. The information about PIM sources can no longer be obtained from MSDP, and will therefore need to be statically provisioned instead. To this effect, on the PIM Controller, it is possible to manually enter multiple source bindings each of which will have to include the class D group address as well as the PIM-SSM source subnet. Once this information has been entered the PIM Controller will select an available PIM Gateway, where the given source IP subnet can be reached from, and will allocate the PIM-SSM source to it.

Inter VSN IP Multicast

Extreme's Fabric Connect delivers unparalleled virtualization of IP Multicast. No other networking technology is able to virtualize IP Multicast into multiple separate logical (VSN) domains with the same simplicity, efficiency, and scalability. And no other networking technology is able to define VSNs defined exclusively for IP Multicast transport and where no IP unicast forwarding and addressing is even required.

It is therefore extremely easy to create separate "tenant" VSNs where each VSN can be activated for separate IP Multicast applications. Different VSNs are completely separate and only use the same Ethernet Fabric infrastructure. Across different L3 VSNs (and L2 VSNs), not only can the same IP unicast addressing be used within each VSN, but also the multicast Groups addresses can be the same.

In some environments, it is sometimes desired to have a flexible approach where different VSNs can be kept separate for unicast connectivity but are allowed to share a selection of IP Multicast streams from a different VSN. Or, alternatively, the VSNs are interconnected via a firewall to allow some unicast traffic, but it would not make any sense to force the IP multicast traffic across that same firewall.

Possible examples would include a video surveillance deployment, fully contained and virtualized within its own L3 VSN, within which video recorders and security screens operate with IP Multicast, but with some security personnel located in a different VSN (in order to reach other different applications) requiring

access to certain video surveillance multicast streams. As both VSNs ultimately operate on the same Ethernet Fabric infrastructure, it would not make sense to duplicate the surveillance cameras IP Multicast streams across both VSNs. This would be inefficient, as it would potentially result in the same IP Multicast packets being transmitted twice over the Ethernet Fabric core. With IP multicast, the most efficient approach is always to create a replication branch on the furthest node along the shortest path. This is what Fabric Connect does in the core and this is what can also be achieved on the last Fabric Attach hop using Multicast VLAN Registration (MVR).

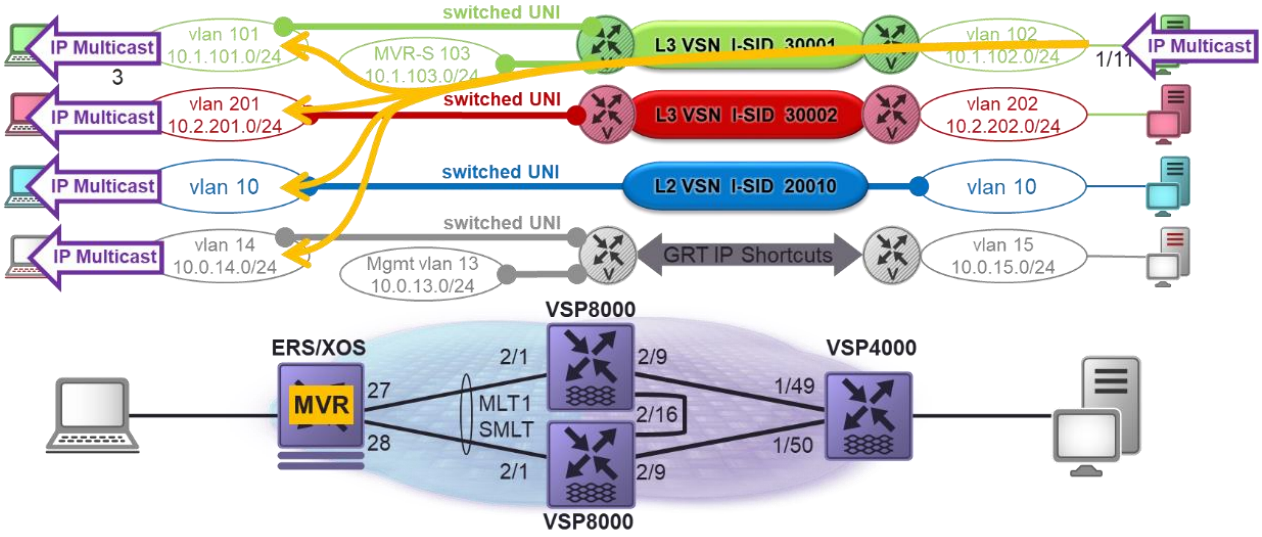


Figure 49 Inter-VSN IP Multicast with MVR on FA Proxy

The MVR function operates on a standard L2 Ethernet switch, which in the Extreme Fabric architecture will be operating as an FA Proxy, and designates one MVR source VLAN (MVR-S) and a number of MVR Receiver VLANs. The MVR source VLAN only needs to be present on the FA Proxy uplinks while the MVR Receiver VLANs are regular user CVLANs; both can be provisioned using Fabric Attach I-SID signalling. Any IGMP request from users within the MVR Receiver VLANs can be made to trigger an IGMP Report (Join) on the MVR Source VLAN. Once the requested IP multicast stream is received on the upstream MVR source VLAN, it will also be IGMP snooped on the receiver ports of the relevant MVR receiver VLANs where the end-users had IGMP signalled an interest in receiving the multicast stream.

Tip

Extreme Networks supports MVR on the ExtremeXOS and ERS 4900 and 5900 platforms.

The MVR global configuration on the access switch will define an IP Multicast Class D range of authorized IP Multicast streams, which will thus be allowed to traverse the VSN domain to which they belong into any of the local CVLANs designated as MVR Receiver VLANs.

Tip

An MVR Receiver VLAN can receive IP multicast via the MVR Source VLAN (if multicast Group address requested falls into MVR range), as well as natively on the same VLAN (if multicast Group address requested does not fall into MVR range). This means that the red or blue or grey users in Figure 49 can receive IP multicast from the green VSN, but can also receive other IP multicast streams from within their own VSNs.

In the example shown in Figure 49, notice that the intent is that the VSN, which naturally owns the IP multicast source, is also able to have users and multicast receivers locally on the same VSN. MVR comes with one limitation, namely that no IGMP receivers are allowed on the MVR source VLAN, so the correct design approach to achieve the design goal is to designate a separate L2 segment (VLAN) for the MVR

source VLAN and another L2 segment (VLAN) for the VSN users attached to the FA Proxy access switch performing MVR. Note that this implies that the VSN that owns the IP multicast source is an L3 VSN (or IP Shortcut instance) which can thus have multiple L2 segments on the VRF/GRT that terminates the VSN.

Caution

The Extreme Networks MVR implementation on ERS platforms does not allow IGMP receivers on the MVR source VLAN.

Another related use case, depicted in Figure 50, would be a deployment where it is desired that all virtualized VSNs be allowed to receive IPTV channels as a shared fabric service. In this case a simpler approach is to place the IPTV source into a dedicated IP multicast enabled L2 VSN. There will be no receivers or end-stations within the IPTV L2 VSN (so a single L2 segments is all that is needed). Instead, the L2 VSN itself becomes the MVR source VLAN across all access FA Proxy switches.

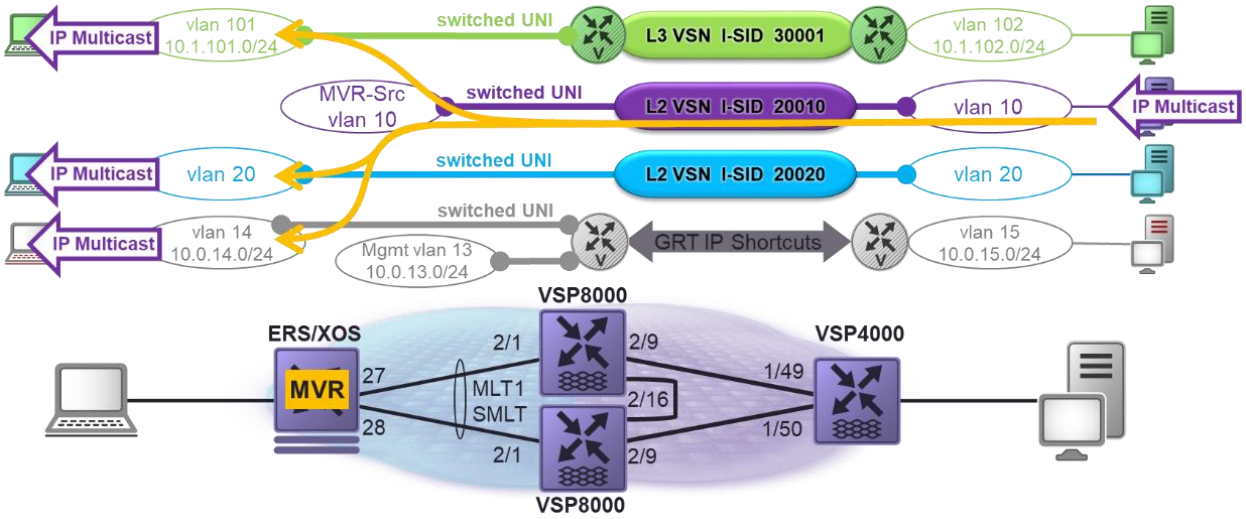


Figure 50 Inter-VSN IP Multicast with Fabric-wide MVR L2 VSN

Now any user in any VSN or VSN type across the fabric can be made to receive the shared IPTV multicast application.

Extending the Fabric Across the WAN

This section will explore the possible design choices available for extending the Extreme Networks Fabric Connect architecture across the WAN / Internet. Applications include extending the fabric from the campus to the branch offices as well as interconnecting larger locations, for example geo-redundant data centers for Data Center Interconnect (DCI).

There are essentially two possible design approaches with the Extreme Fabric Connect architecture, as depicted in Figure 51. The most compelling option is Fabric Extend, where the underlying SPB Fabric is extended all the way. This presents the advantage that all VSN Fabric services, regardless of type (L2, L3, E-TREE, IP Multicast), can now be seamlessly extended to any part of the fabric with end-point provisioning. The Fabric VSN services are transported over the fabric and the WAN operator has no visibility of those services and no need to participate in IP route advertising if those services are L3-based. On the other hand, Fabric Extend will result in the Fabric Connect being extended to all nodes across different geographical locations and the scaling limits of SPB nodes per region will have to be observed.

Caution

An Extreme SPB Fabric can currently scale to a maximum of 500 nodes per region (area), assuming a range of SPB VSP platforms are in use. (Some VSP platforms can scale higher than 500; refer to product Release Notes). This figure is dictated by the current generation of ASICs, but is expected to rise considerably as newer chipsets become available in the years to come. Future support of Multi-Area Fabric will also allow designs to exceed this limit.

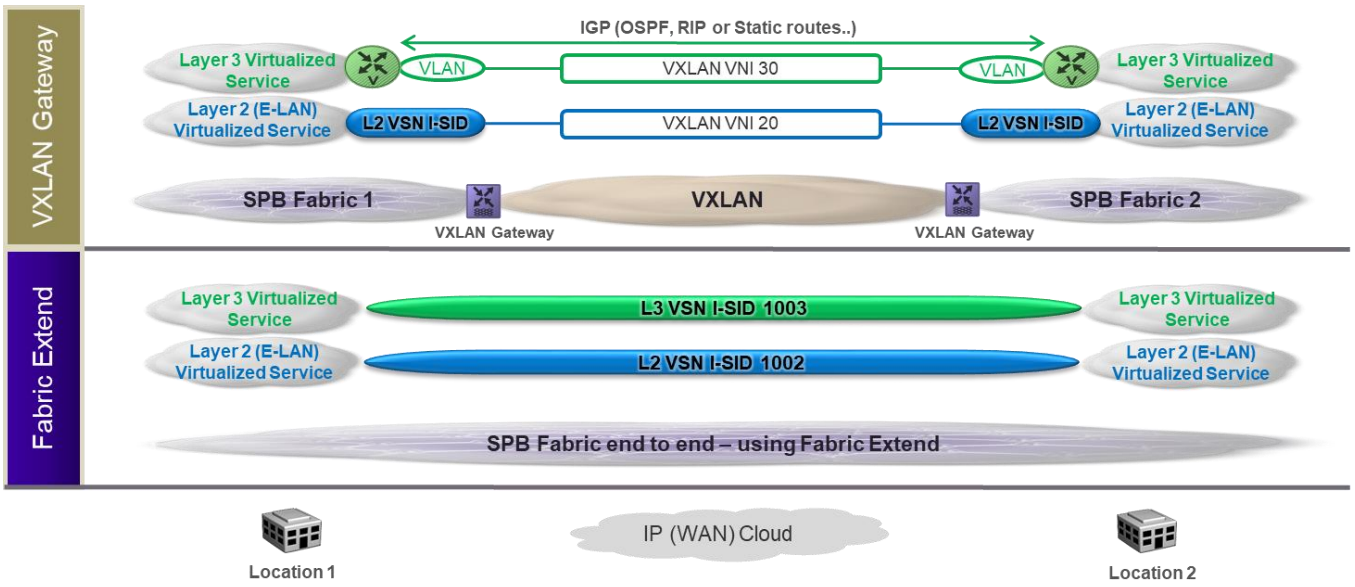


Figure 51 WAN Extending the Fabric or the VSN Services

The other design approach consists of extending only selected VSNs instead of the SPB Fabric foundation. This is achieved via the VXLAN Gateway functionality, which allows a Fabric Connect BEB node to map a SPB L2 VSN (or VLAN) segment into a VXLAN VNI.

Tip

Extreme Networks VSP 4900, 7200, 7400, 8200 & 8400 platforms support the VXLAN Gateway functionality, which is implemented in hardware switching.

This has the advantage that the SPB Fabrics across the distant locations remain separate and the scaling limits of SPB nodes per region will only need to be observed within each location.

Both Fabric Extend and the use of VXLAN Gateway have the advantage of being able to traverse a WAN L3 cloud without having to request multiple circuits from the WAN provider. However, use of the VXLAN Gateway approach does present some limitations with respect to Fabric Extend. Extending L2 VSNs with VXLAN Gateway is fairly straightforward, but extending L3 service types (L3 VSNs or IP Shortcuts) will require running a traditional IP routing protocol (BGP, OSPF, RIP, or Static routes) over the VXLAN VNI interconnecting segment. The end-point provisioning simplicity of Fabric Connect is somewhat lost as any new service deployed over VXLAN will also require touching the configuration of the VXLAN Gateways. And finally, IP Multicast will not be viable over VSN services extended with VXLAN Gateway.

Note

IP Multicast over VXLAN extended L2 VSN will flood, using inefficient ingress replication, neither of which should be considered over WAN circuits.

IP Multicast over VXLAN extended L3 VSN, while technically feasible, would require running PIM and PIM-Gateway functionality, which becomes highly complex if compared to using Fabric Extend.

The following sections will cover all the possible Fabric Extend operational modes as well as VXLAN Gateway in greater depth.

Fabric Extend

Extending the SPB Ethernet fabric over WAN circuits presents a number of challenges which in part depend on the type of WAN services used. For a start, the SPB Fabric can only operate if the Fabric Connect nodes are interconnected with Ethernet point-point links, over which point-point IS-IS adjacencies form.

Tip

The IS-IS protocol would technically support broadcast interfaces but the SPB standard only defines the use of IS-IS point-point interfaces. An SPB implementation using IS-IS broadcast interfaces would result in less efficient multicast trees. Extreme's SPB implementation only support IS-IS point-point interfaces as per the SPB specification.

Preserving the point-to-point nature of interconnections between Fabric Connect nodes is straightforward when the WAN services in question are E-LINE point-to-point Ethernet links. But when the WAN service types are any-to-any like with L2 E-LAN (VPLS typically used by WAN provider) or IPVPN (RFC4364 MPLS-VPNs), there needs to be a way to preserve the point-to-point nature of SPB interconnects. In the case of an L3 type WAN service (IPVPN), there is also the challenge that the SPB Mac-in-Mac encapsulation cannot be sent natively over the WAN service but needs to be IP-encapsulated.

Fabric Extend addresses both these challenges by using an IP encapsulation which allows the creation of IP point-to-point tunnels as well as being capable of traversing a WAN IP cloud.

Tip

The Extreme implementation of Fabric Extend, when operating with IP tunnels, uses either a VXLAN or an IPsec encapsulation. Other IP type encapsulations would have been equally possible, but VXLAN was chosen because of its support in the hardware silicon used on Extreme Networks VSP platforms.

Extreme Networks VSP 4900, 7200, 7400, 8200 & 8400 platforms have VXLAN capable hardware and can thus natively support Fabric Extend in IP VXLAN mode with hardware switching.

Extreme Networks Fabric Connect VPN XA1400 (XA1440 & XA1480) platforms natively support Fabric Extend in IP VXLAN mode but can also support Fabric Extend with an IPSec encapsulation; both are implemented with software switching.

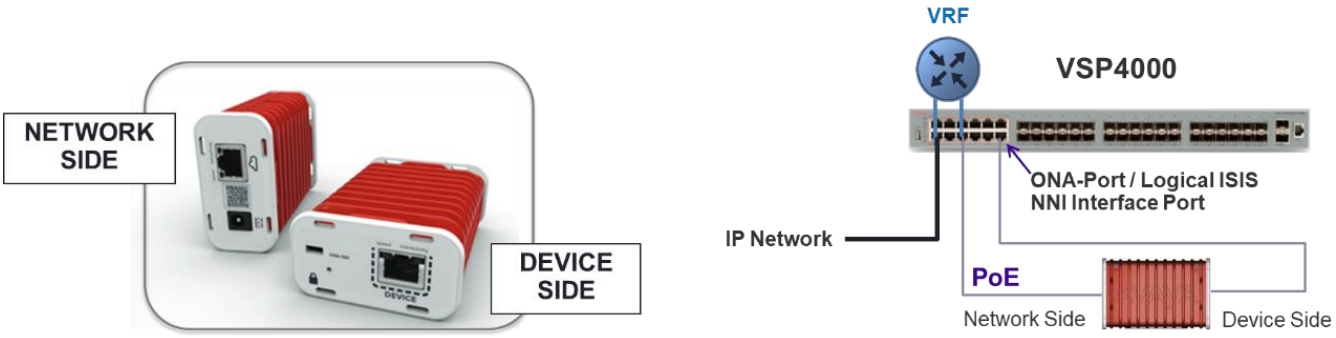


Figure 52 Fabric Extend Pairing of ONA with VSP4000

Note

Extreme Networks VSP 4450/4850 platforms do not have VXLAN capable hardware and must therefore be paired with an Open Network Adapter (ONA) to support Fabric Extend in IP VXLAN mode. See Figure 52.

Clearly the use of a VXLAN or IPsec encapsulation has implications with regards to the maximum frame sizes that can be expected to transit over the WAN service. SPB's Mac-in-Mac encapsulation already adds 22 bytes to a normal Ethernet frame size, and a VXLAN encapsulation will add a further 50 bytes, as can be seen in Figure 53. Hence the largest possible Ethernet packet with frame size of 1518 (untagged) / 1522 (tagged) bytes (for an IP MTU of 1500 bytes) will reach a maximum size of 1594 bytes. With an IPsec encapsulation the overhead is even larger, though IPsec will typically be used over the Internet and will thus require fragmentation and reassembly anyway.

Caution

Fabric Extend in IP (VXLAN) mode requires the WAN to support frame sizes of 1600 bytes.

When planning a Fabric Extend deployment, it is therefore important to verify with the WAN provider that the WAN service will be able to support such MTU requirements.

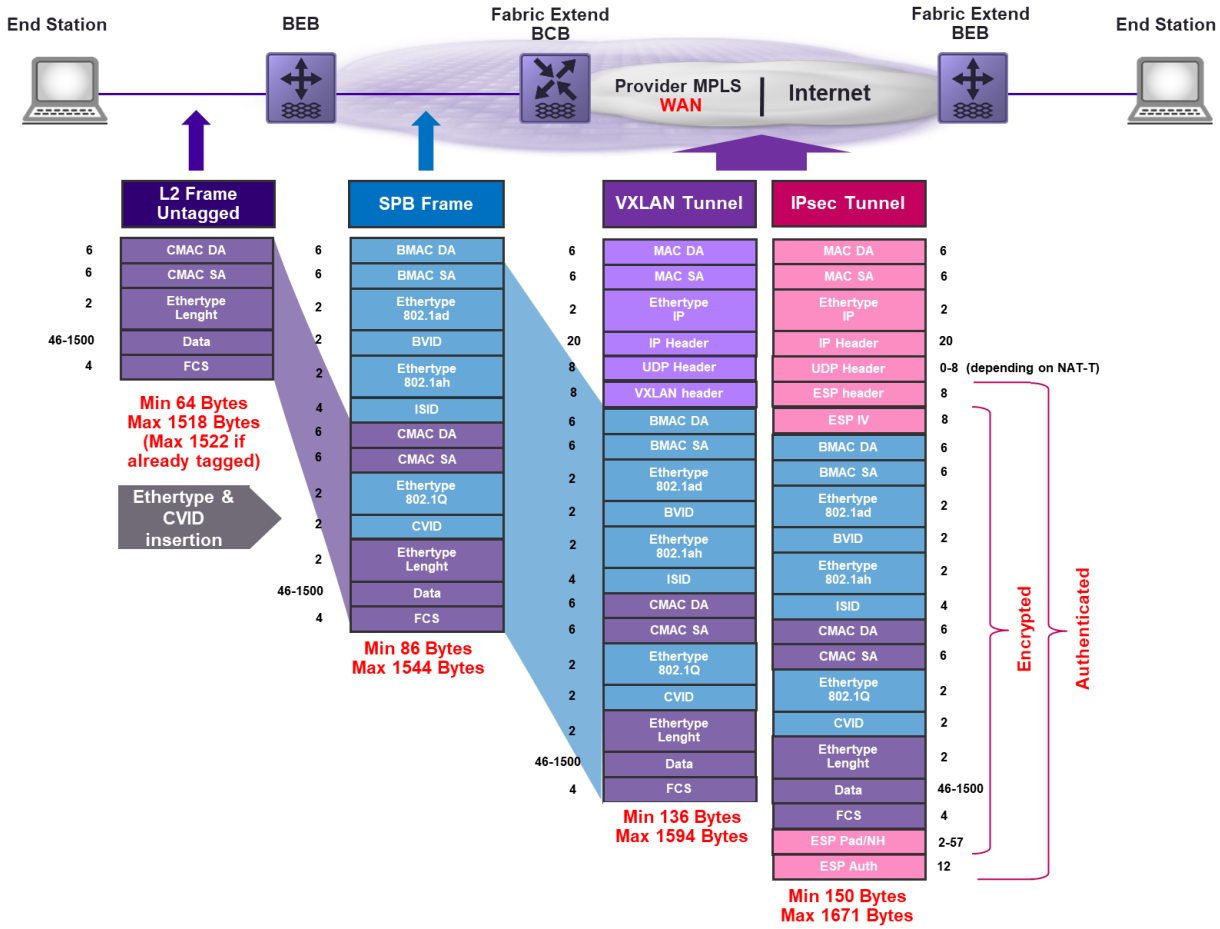


Figure 53 Fabric Extend IP Mode (VXLAN / IPsec) MTU Considerations

Tip

WAN providers will typically use MPLS to deliver their WAN services. MPLS equipment is perfectly capable to handle oversize frames in excess of 1600 bytes. However, the WAN provider may choose whether or not to allow larger MTUs in their service offerings.

For Data Center Interconnect (DCI), it is also possible to deploy Fabric Extend with jumbo frame sizes. The same encapsulation overheads of 22 (Mac-in-Mac) + 50 VXLAN) bytes apply, but the original Ethernet frame generated by servers/VMs will now have an IP MTU of 9000 bytes, which will result in a maximum Fabric Extend frame size just short of 9100 bytes. Again, the WAN provider must be able to offer a jumbo capable service to meet these MTU requirements.

Tip

When using jumbo frame sizes, the IP MTU is set (on the end-stations) to 9000 bytes, resulting in a maximum Ethernet frame size that is well below the absolute maximum that jumbo capable equipment can support (typically between the 9200 – 9800 mark). Extreme VSP platforms can handle maximum frame sizes of up to 9600 bytes.

A special case to consider is using Fabric Extend over the Internet (as opposed to a WAN). The Internet will not handle any traffic with an IP MTU greater than 1500 bytes so, at first sight, this would not be an option for Fabric Extend (or the VXLAN Gateway approach). The only way to run Fabric Extend over the Internet (or a WAN where IP MTUs greater than 1500 are not allowed) is for the Fabric Extend terminating equipment to support IP fragmentation and re-assembly. Extreme Networks offer this capability on the Fabric Connect VPN XA 1400 platforms and the VSP4450/4850 platforms (latter must be associated with an Open Network Adapter - ONA).

Note

Fabric Connect VPN is a licensed software application hosted on XA1400 hardware and based on VSP Operating System Software (VOSS). Two licensing tiers are available offering either WAN Bandwidth of 100Mbps (applicable to both XA1440 and XA1480) or 500Mbps (applicable to XA1480 only).

Note

The ONA is a software-driven device attached to VSP4450/4850 used to apply and remove the VXLAN encapsulation when used in Fabric Extend mode and can therefore do IP fragmentation.

Use of Fabric Extend with VXLAN fragmentation and reassembly will not, however, scale as much as a non-fragmented implementation.

Caution

The fragmentation and reassembly is handled in software and throughput rates will be greatly reduced if the majority of traffic being VXLAN encapsulated requires fragmentation and reassembly.

Note

Fabric Extend with IP Fragmentation is only possible if an Extreme XA1440/1480 or VSP4450/4850 with Open Network Adapter (ONA) is deployed at both ends of the connection. Other Extreme VSP platforms will not support IP fragmentation.

Note

Fabric Extend with IPsec encapsulation is only supported with the Fabric Connect VPN XA1400 platforms.

The Fabric Extend (VXLAN or IPsec) IP tunnels constitute effectively an IP overlay above the WAN cloud. Where they run to and from becomes arbitrary, yet they ultimately will determine what fabric topology is seen by IS-IS. It is important to assess what the expected traffic flows will be over the WAN and then to deploy the Fabric Extend IP tunnels accordingly. Typical deployments would be hub-star, dual-hub-star, fully meshed, or any combination thereof. Traffic flows for which no direct IP tunnel exists, will end up crossing the WAN more than once if they use multiple IP tunnels.

Caution

Extreme Management Fabric Manager supports the Fabric Extend Manager tool to automate and push down the IP tunnel configuration based on the required Fabric Extend IP tunnel overlay topology.

The final challenge that Fabric Extend needs to address is that typically WAN deployments have a few head offices and many (possibly hundreds of) branch offices. In Fabric Connect mode, there is a tight coupling of physical Ethernet port (or MLT bundle) and IS-IS/SPB interface which are mapped 1:1. For any-to-any type WAN services (L2 E-LAN or IPVPN), it is clear that all the resulting Fabric Extend IP tunnels will need to be terminated on a single Ethernet port connecting to the WAN service. This is also true in the case of point-to-point WAN service types since the WAN provider, at the head office end, will bundle one end of all the point-to-point circuits into a single Ethernet connection and use 802.1Q tags to differentiate between them.

Fabric Extend addresses this requirement by allowing the creation of logical IS-IS interfaces where many Fabric Extend IS-IS adjacencies can be terminated on the single Ethernet port (one-to-many mapping).

Tip

Extreme Networks VSP platforms can support up to 255 IS-IS adjacencies over Fabric Extend IS-IS logical interfaces.

The following sections will cover each of the possible ways in which Fabric Extend can be deployed. It should be noted that while certain technical trade-offs exist between the various modes, each mode is designed to adapt over a particular WAN service type. There are many factors both technical and economical that will determine what the most appropriate WAN service type is for a given customer. Ultimately it is the chosen WAN service type which will dictate which Fabric Extend mode needs to be used, rather than the inverse.

Tip

Extreme Networks Fabric Extend offers a solution to adapt to any WAN service type.

Fabric Extend over IPVPN Service

The diagram in Figure 54 shows a typical Fabric Extend deployment over an L3 WAN service, which is typically delivered by the WAN provider using RFC4364 MPLS BGP IPVPNs. In the example shown, a hub-and-spoke topology was used for the Fabric Extend IP tunnels, but could equally have been fully meshed, or a combination. All IP tunnels terminate on Fabric Extend logical IS-IS interfaces that are capable of terminating one or more IP tunnels (and thus IS-IS interfaces) on the same physical Ethernet port.

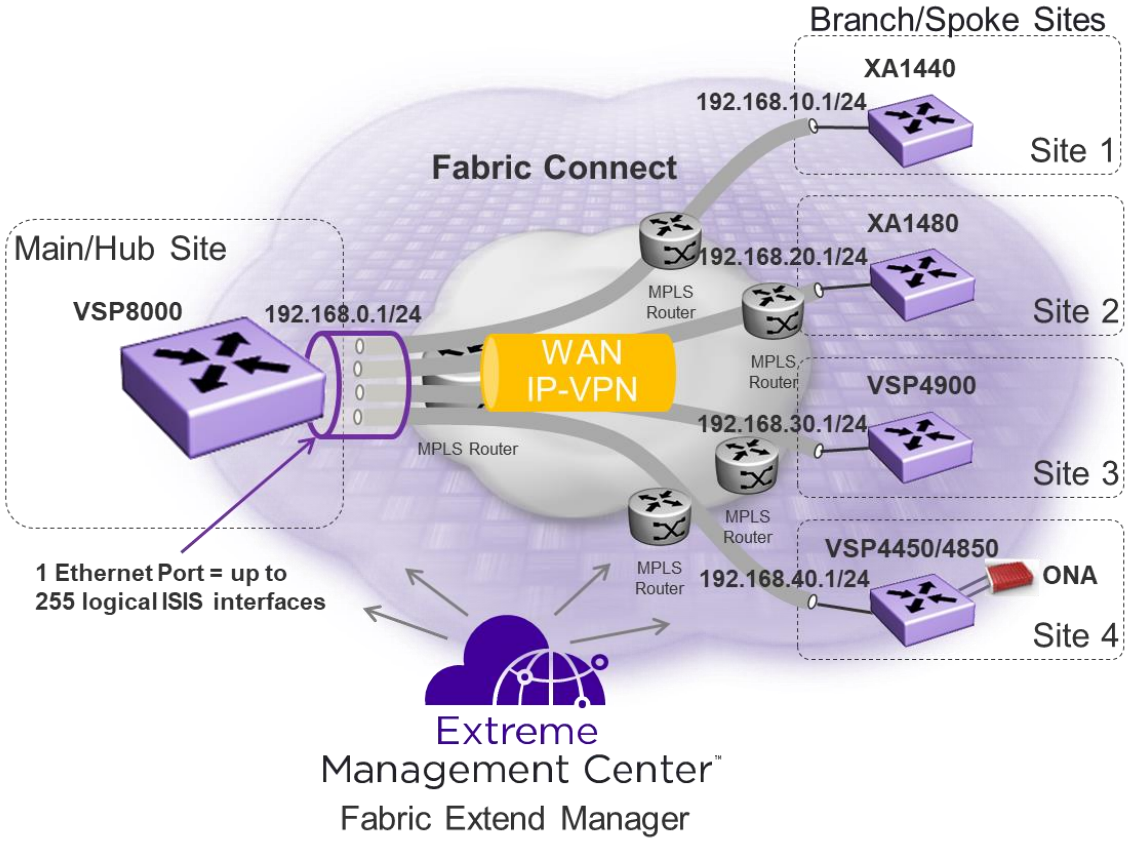


Figure 54 Fabric Extend over WAN L3 Any-to-Any IPVPN Service

At each site, the WAN provider will allocate an IP subnet and IP address to be used by the customer's CE (Customer Edge) equipment. In a traditional non-fabric deployment, the customer would also have to agree with the WAN provider on what IP routing protocol to use (typically BGP or OSPF, if not IP static

routes) such that the customer IP subnets can be re-advertised by the WAN IPVPN between the distant sites.

With Fabric Extend the IP address provided by the WAN provider will become the Fabric Extend source IP and will be used to originate and terminate all IP (VXLAN) tunnels. Thus the only IP traffic which will be seen by the WAN provider will be VXLAN encapsulated traffic between the IP addresses it supplied. The need to run an IP routing protocol with the provider goes away.

Because the WAN service becomes an underlay to Fabric Connect, it is necessary that any IP interfaces used by Fabric Extend to build the IP tunnels should be isolated from any other IP interface used to carry VSN services above the Fabric. The correct design approach is to allocate a VRF for Fabric Extend and use this VRF to allocate the Fabric Extend IP interfaces used to send and receive traffic from the WAN.

This is illustrated in Figure 55. The Fabric Extend VRF should not be used to terminate any L3 VSN (via I-SID assignment) and thus any IP routes used by Fabric Extend will remain isolated from IP service domains running above the Fabric (IP Shortcuts and L3 VSNs).

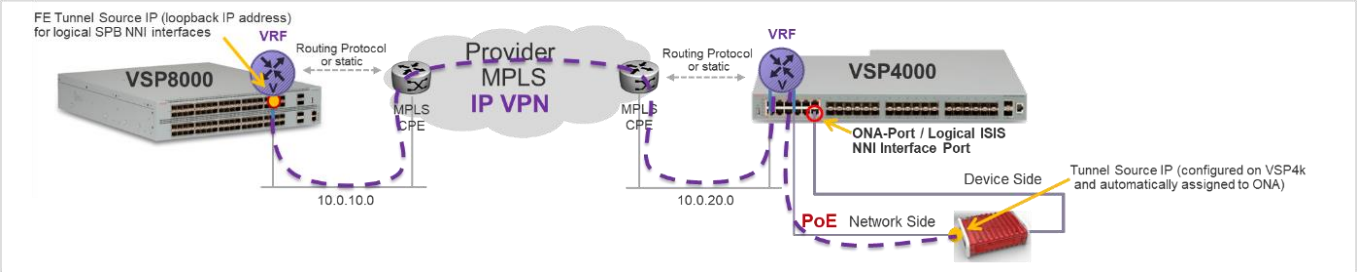


Figure 55 Fabric Extend Deployment Model over WAN L3 IPVPN Service

Note

Use of the GRT (VRF-0) or an L3 VSN enabled VRF for terminating Fabric Extend tunnels is technically feasible and supported. Such a deployment could be envisaged where there is a need to preserve IP management to distant Fabric Extend equipment natively over the WAN service or where Fabric Extend is deployed over a campus LAN architecture and there is a need to preserve connectivity between the Fabric Extend overlay and the campus IP routed underlay.

However, such designs should be avoided as they are fraught with additional complexity to ensure that the IP routes necessary to form the Fabric Extend overlay tunnels must never end up being replaced with IS-IS IP routes from the overlay, which would result in the overlay falling apart.

Fabric Extend over the Public Internet with IPSec

The diagram in Figure 56 shows how Fabric Extend can be performed over the public Internet. Only the Fabric Connect VPN software (e.g. running on XA1400 platforms) can be used in this mode as Fabric Extend over the Internet requires IP fragmentation and, for security reasons, IPsec must be used to encrypt the traffic.

Tip

The Extreme Networks Fabric Connect VPN software (e.g. running on XA1400 platforms) supports the following IPsec parameters:

- IPSec Mode: Tunnel mode
- Key Exchange Protocol: IKEv2
- Authentication methods: Pre shared key
- IPSec Security Protocol: ESP
- ESP Encryption: aes128gcm16 sha256

With the Fabric Connect VPN software the XA1400 designated WAN port is hardened and can be directly connected to the Internet without having to go through any firewall. Nevertheless, in many deployments the Fabric Extend terminating XA1400 nodes will typically be connected behind the enterprise firewall and in this case two possible deployment models exist. Either the Fabric Connect VPN software is configured with a public Internet IP address (and no NAT is performed by the Firewall) or the Fabric Connect VPN software is configured with private IP addressing and the Firewall perfoms NAT. In this latter case the Fabric Connect VPN will need to be enabled for NAT Traversal (NAT-T) which will add an additional UDP header to the IPsec ESP encapsulation as shown in Figure 53.

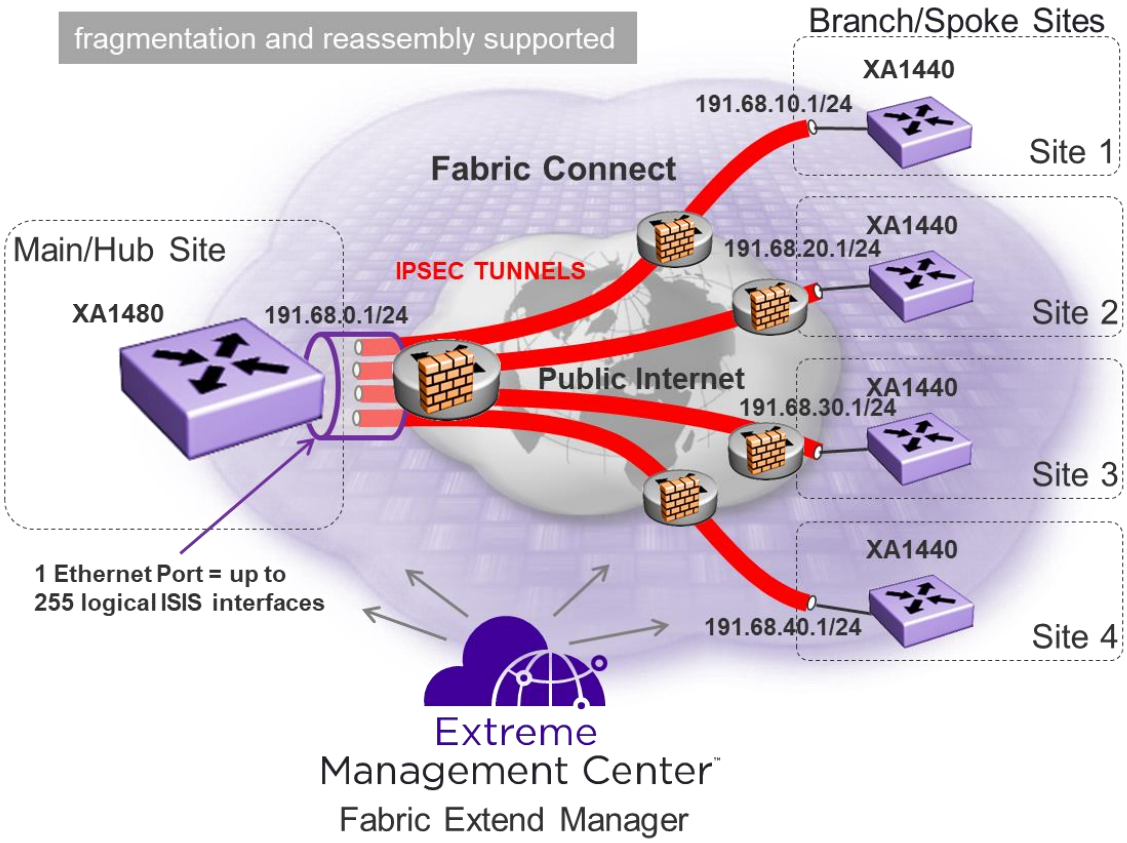


Figure 56 Fabric Extend over the Public Internet with IPSec

Caution

Extreme Networks Fabric Connect VPN software does not support NAT Traversal (NAT-T) in the first release.

IP interfaces used by Fabric Extend to build the IP tunnels must be isolated from any other IP interface used to carry VSN services above the Fabric. The correct design approach is to allocate a VRF for Fabric Extend

and use this VRF to allocate the Fabric Extend IP interfaces used to send and receive traffic from the Internet. This VRF will never be L3 VSN enabled.

Note

Use of the GRT (VRF-0) or an L3 VSN enabled VRF for terminating Fabric Extend tunnels should be avoided when deploying Fabric Extend over the Internet. The Fabric Connect VPN interface used for Fabric Extend will most likely need to be located in the enterprise DMZ and only the Fabric Extend IP will be allowed out of the Firewall. Fabric Internet access should be performed using alternative GRT or L3VSN services which will have a separate connectivity via the enterprise Firewall.

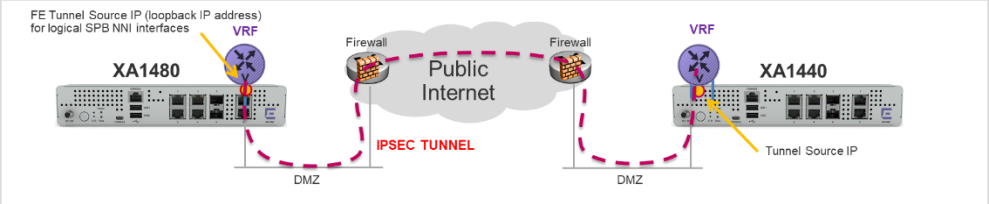


Figure 57 Fabric Extend Deployment Model over Public Internet with IPSec

Fabric Extend over E-LAN/VPLS Service

The diagram in Figure 58 shows a typical Fabric Extend deployment over an L2 any-to-any WAN service which is typically delivered by the WAN provider using MPLS VPLS. In the example shown, a hub-and-spoke topology was used for the Fabric Extend IP tunnels, but could equally have been fully meshed or a combination. All IP tunnels terminate on Fabric Extend logical IS-IS interfaces that are capable of terminating one or more IP tunnels (and thus IS-IS interfaces) on the same physical Ethernet port.

The WAN provider is offering an L2 service and it is up to the customer to allocate a single IP subnet to use across the geographical L2 segment so that the Fabric Extend IP tunnels can be created as required.

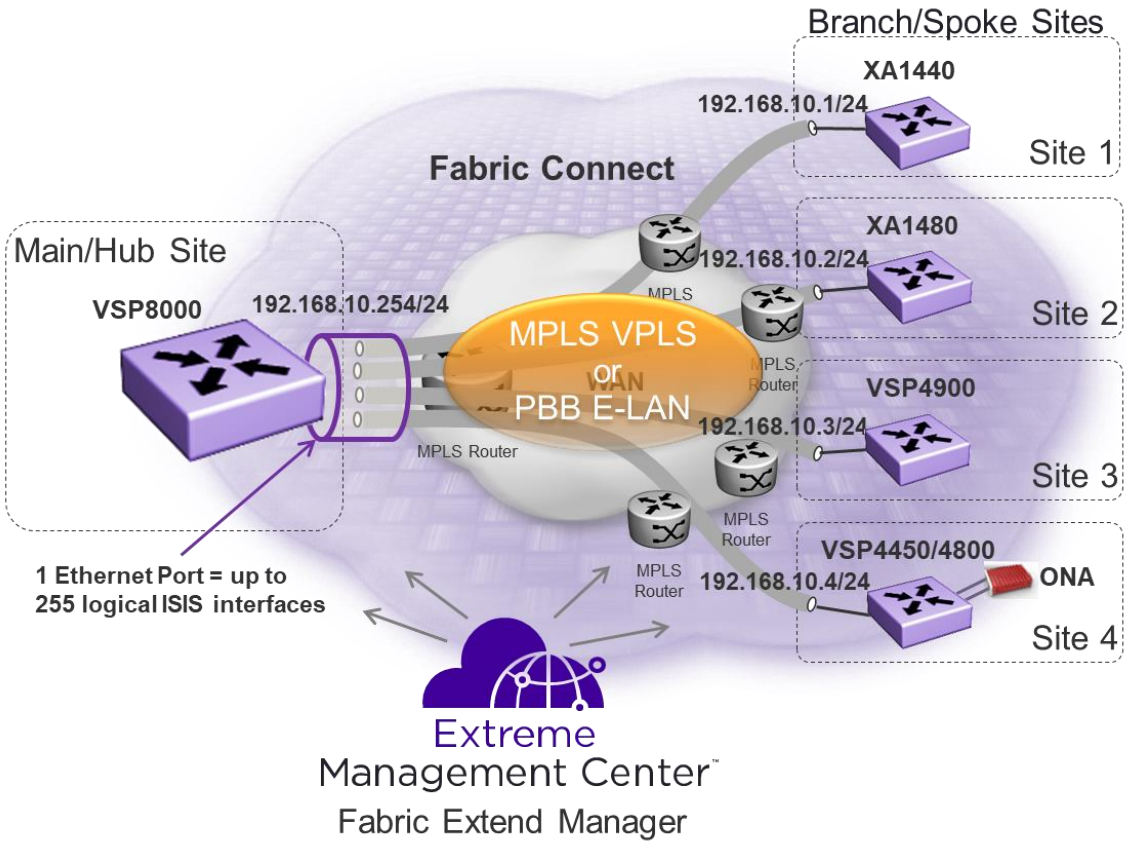


Figure 58 Fabric Extend over WAN L2 Any-to-Any E-LAN Service

The WAN service remains an underlay to Fabric Connect, so it is again necessary that the IP interfaces used by Fabric Extend to build the IP tunnels should be isolated from any other IP interface used to carry VSN services above the fabric. The correct design approach is again to allocate a VRF for Fabric Extend and use this VRF to allocate the Fabric Extend IP interfaces used to send and receive traffic from the WAN. This is illustrated in Figure 59.

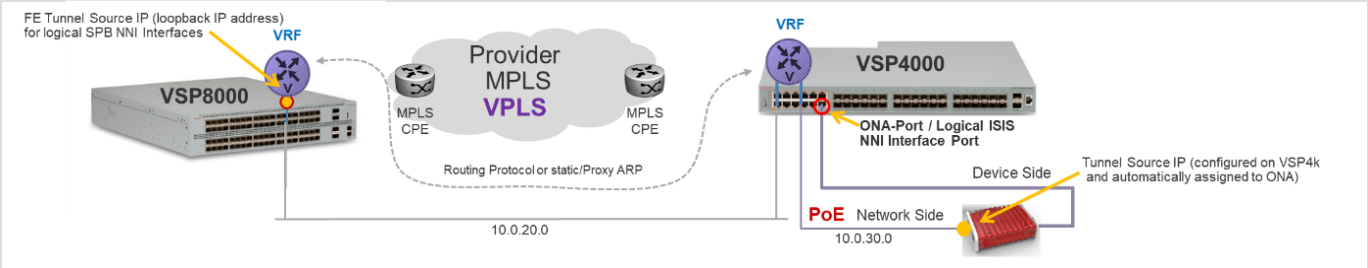


Figure 59 Fabric Extend Deployment Model over WAN L2 E-LAN Service

One notable difference of operating Fabric Extend over an L2 E-LAN service type (vs. an L3 WAN service type) is that all the IP tunnel end points are IP addresses in the very same IP subnet and will thus consume ARP entries on the VXLAN encapsulating end-point. (Whereas with an L3 WAN service all end-points are in distant subnets and only one ARP entry is used on the VXLAN encapsulating end point, corresponding to its default gateway.)

Tip

Extreme Networks VSP 4900, 7200, 7400, 8200, 8400 & XA1400 platforms can handle sufficient ARP entries to support the maximum scaling of Fabric Extend IP tunnels.

Caution

The same is not true with the Extreme VSP 4450/4850 platform and Open Network Adapter (ONA), where the ONA cannot handle more than one ARP entry for all IP tunnels. To work around this limitation, it is necessary to front end the ONA Tunnel IP with a VRF which locally exists on the VSP. The ONA Tunnel IP will thus be in a different IP subnet from that used over the wider WAN L2 segment and will thus present the ONA with a single default gateway (and hence ARP entry) for all Fabric Extend IP tunnels. However, use of IP Static Routes will be required to make the ONA IP Tunnel address reachable. This is illustrated in Figure 59.

Fabric Extend over E-LINE Service

The diagram in Figure 60 shows a Fabric Extend deployment over WAN L2 point-to-point services. The WAN provider would typically provide these via use of MPLS Pseudowires which would be aggregated in some form at the central location in order to be handed off over a single Ethernet connection.

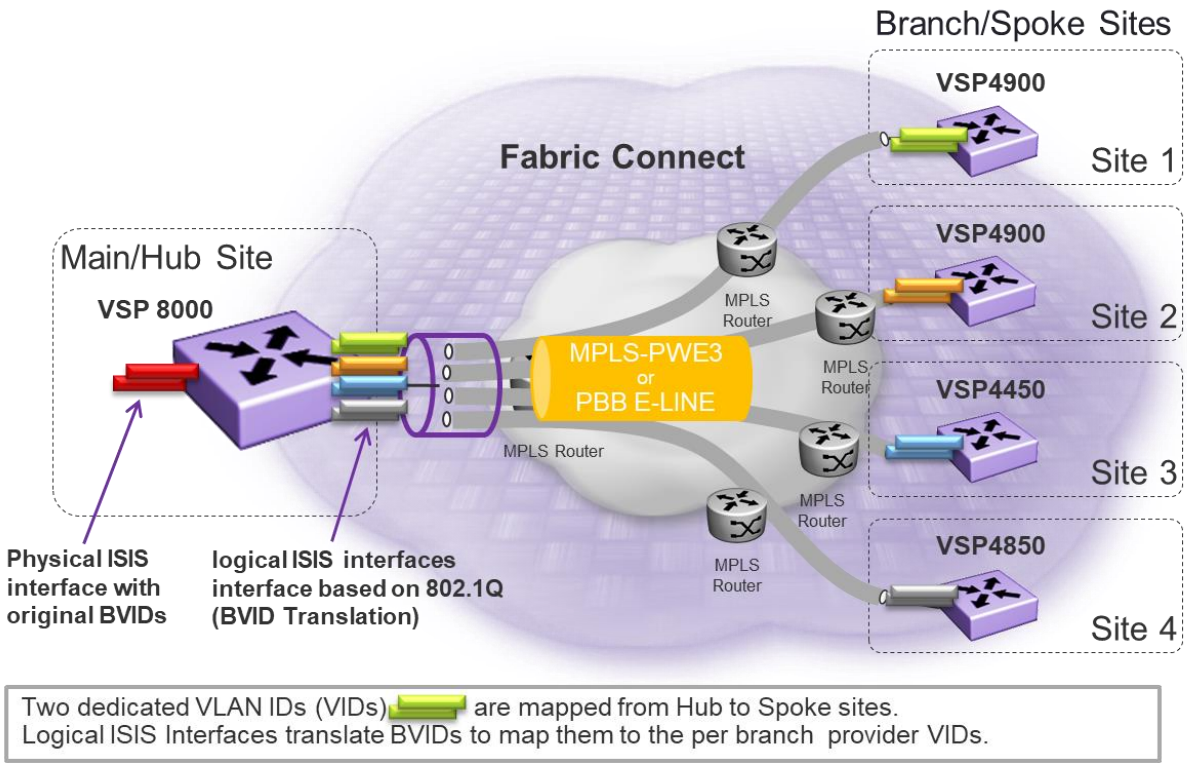


Figure 60 Fabric Extend over WAN L2 Point-to-Point E-LINE Services

This Fabric Connect mode is different from the other modes covered above in that there is no need to use an IP (VXLAN) encapsulation and tunnels because the nature of the WAN service is already L2 and point-to-point, which means that IS-IS and SPB's Mac-in-Mac encapsulation can be run natively directly over these WAN circuits. We shall refer to this as the Fabric Extend L2 mode. This also means that there is no IP tunnel overlay in this Fabric Extend mode and therefore there is no possibility to craft IP tunnels to match the traffic patterns. It is the WAN point-to-point circuits that need to be provisioned instead to match those desired traffic flows.

Caution

This Fabric Extend mode is not supported on Fabric Connect VPN XA1400 platforms.

Tip

Because this Fabric Extend mode does not use a VXLAN encapsulation, it is possible to deploy it with Extreme Networks VSP 4450/4850 platforms without the use of an associated Open Network Adapter (ONA).

Fabric Extend end-points terminating more than one WAN point-to-point circuit will need to be provisioned as Fabric Extend logical IS-IS interfaces that can terminate one or more L2 circuits (and thus IS-IS interfaces) on the same physical Ethernet port.

There are two possible Ethernet encapsulations that WAN providers will typically use to aggregate all circuits onto a single Ethernet connection at the head end site: IEEE 802.1ad Q-in-Q or native IEEE 802.1Q-tags. It is important to note that the Extreme Networks Fabric Extend logical IS-IS interfaces only support 802.1Q encapsulation and cannot support Q-in-Q.

Caution

Fabric Extend in L2 mode does not support IEEE 802.1ad Q-in-Q; only IEEE 802.1Q-tags are supported.

This has some implications in that we know that an SPB Ethernet Fabric will emit traffic with 802.1Q-tags for any and all of the BVLANs it was provisioned with. The first problem is that the SPB BVLAN IDs are globally configured and have to be the same across the entire SPB Fabric, which does not square with the need to change that same Q-tag to differentiate between the various WAN circuits arriving on the same port. This first problem is resolved by the Fabric Extend logical IS-IS interface capability to implement VLAN translation when used in Fabric Extend L2 mode. In other words, for a given point-to-point WAN circuit, the BVLAN Q-tags will be translated to a different pair of Q-tag values that will be unique to that circuit and different from other Q-tag value pairs used for other WAN circuits terminating on the same logical IS-IS Ethernet port.

The second problem is that use of an 802.1Q encapsulation means that the WAN provider is not able to hide the various SPB BVLAN Q-tags behind a single outer Q-in-Q tag service. Instead a number of circuits equal to the number of BVLANs will need to be provided in parallel between each pair of Fabric Extend nodes. This is depicted in Figure 60 by the double colored bricks terminating each circuit. Configuration of a Fabric Extend end-point on a logical IS-IS interface will thus require two VLAN-tags, one for the Primary BVLAN and one for the Secondary BVLAN.

Note

Extreme's Fabric Connect currently support a maximum of two BVLANs, but support of up to 16 BVLANs will become available in future software versions. However, when 16 BVLAN support is added, it is expected that this will be able to co-exist with other parts of the same SPB Fabric where only two BVLANs are in use. Use of more than two BVLANs will typically only provide benefits in the data center for spine-leaf architectures and will not be used in the wider campus and branch fabric.

A further point of consideration has to do with the outer Ethernet Ethertype encoding used. The correct SPB Mac-in-Mac encapsulation Ethertype value is hexadecimal 88:a8, which is the same Ethertype used with IEEE 802.1ad Q-in-Q but is different from the 81:00 hexadecimal value that is typically expected on IEEE 802.1Q interfaces. The WAN circuits may thus not accept SPB traffic arriving with an 88:a8 Ethertype on an IEEE 802.1Q interface.

Tip

All Extreme Networks SPB capable platforms are able to process and receive Mac-in-Mac traffic using either 88:a8 or 81:00 Ethertype.

All Extreme Networks SPB-capable platforms can be globally configured to use either 88:a8 or 81:00 Ethertype when (acting as a BEB) they generate Mac-in-Mac encapsulated traffic. By default, 81:00 Ethertype is used. The Ethertype can be changed dynamically without any operational impact.

Clearly Fabric Extend L2 mode is not as flexible as the Fabric Extend L3 modes where an IP encapsulation (VXLAN or IPsec) is used. It does however present an advantage over the other Fabric Extend modes in that the Extreme Networks VSP4450/4850 platforms could be deployed without the additional cost of an Open Network Adapter (ONA).

VSN Extend with VXLAN Gateway

The VXLAN Gateway functionality brings the ability to interconnect virtualized L2 segments in a VXLAN overlay with virtualized L2 segments in the Extreme Fabric Connect architecture.

The VXLAN Gateway BEB thus also becomes a VXLAN termination point (VTEP) where a Circuitless IP address is assigned to be the VTEP source IP and a number of static VXLAN tunnels can be provisioned towards remote VTEPs across the VXLAN cloud. The VXLAN virtual segments (VXLAN Network Identifiers - VNI) are then simply mapped to a list of remote VTEP tunnels on one side and to a Fabric L2 I-SID on the other.

In the Extreme implementation, the VXLAN Gateway functionality is able to operate on SMLT clustered BEBs (which share a vIST) and is essential in order to provide redundant active-active and Spanning Tree free gateway functionality for interconnecting what are in effect virtualized L2 broadcast segments.

In an SMLT cluster configuration, both VXLAN Gateways sharing the vIST will need to be configured with the same Circuitless IP address as VTEP source IP. Both Gateways will announce the same VTEP IP into the VXLAN IP underlay thus ensuring that both Gateways are used to load balance traffic arriving from the VXLAN cloud and entering the Extreme Fabric. In the reverse direction load balancing is ensured via the usual SMLT clustering mechanism which presents a Virtual-BMAC for the SMLT cluster.

A typical deployment scenario is depicted in Figure 61, which uses redundant VXLAN Gateways with a vIST in between them and maps a VXLAN VNI on one side with a Fabric L2 I-SID on the other. The VXLAN Gateway will also work with third party VXLAN VTEPs but currently requires all remote VTEPs to be statically provisioned.

Note

VMware NSX uses a VXLAN overlay but requires OVSDB to discover remote VTEPs and automate the creation of the VXLAN tunnels. Extreme Networks VSP platforms do support OVSDB but this is not yet covered in this document.

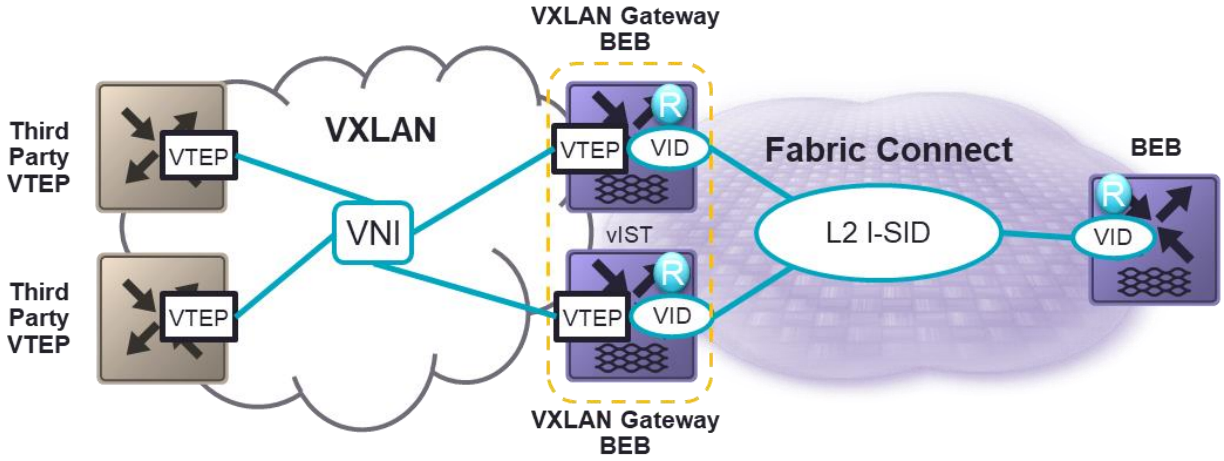


Figure 61 VXLAN Gateway Capabilities

It should be noted that in the Extreme Networks implementation of VXLAN Gateway the VXLAN VNIs by virtue of being interconnected with Fabric L2 I-SID can thus be fully integrated into the Fabric L2 & L3 VSNs. Much like an L2 VSN, the VXLAN VNI can be assigned to an IP address, which in turn can belong to a VRF, which in turn can belong to an L3 VSN.

Like with most other vendors offering VXLAN overlay capabilities, the Extreme VXLAN Gateway will handle broadcast, unknown, and multicast (BUM) packets using ingress replication, which has the advantage of not requiring the IP underlay transport network to be IP Multicast capable, but results in a less efficient way of dealing with BUM traffic. Every BUM packet received from the Fabric side L2 I-SID will thus need to be ingress replicated by the VXLAN Gateway VTEP as many times as that VTEP has remote VTEPs defined for the given VNI.

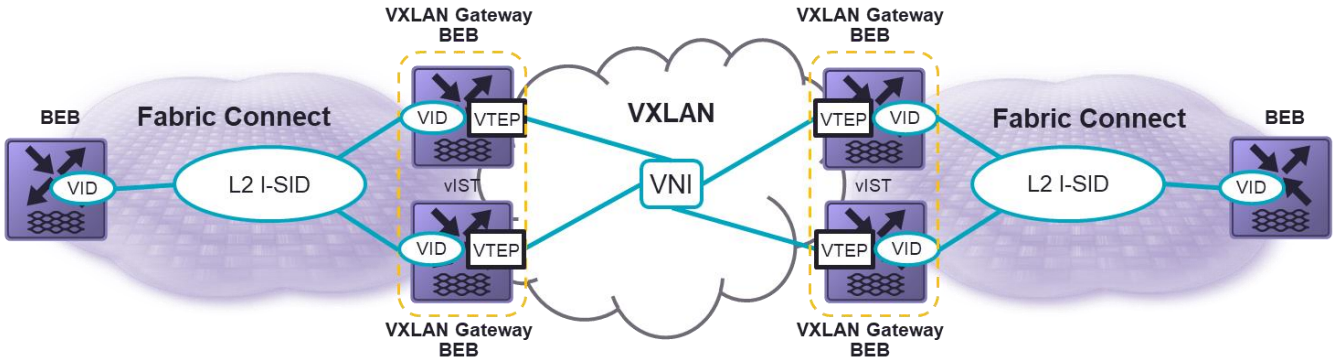


Figure 62 Extending an L2 VSN Across Fabrics with VXLAN Gateway

Figure 62 shows a possible use of VXLAN Gateway to extend a Fabric L2 VSN across two separate Extreme SPB Fabrics where the VTEPs are Extreme VSP platforms at both ends. The L2 VSN L2 I-SIDs are mapped to the VXLAN VNI on the VXLAN Gateways so the I-SID value can be different at both ends. The VXLAN cloud is by definition an IP overlay and can be transported over any L3 capable transport network including a WAN cloud but on condition that this is capable of handling oversize frames.

Note

The VXLAN encapsulation adds 50 bytes to the maximum size of an Ethernet packet, 1518 (untagged) or 1522 (tagged). The Extreme Networks VXLAN Gateway functionality only encapsulates in VXLAN the native Ethernet frame and does not include the SPB Fabric Mac-in-Mac encapsulation.

Caution

The VXLAN Gateway functionality is only supported on Extreme Networks VSP 4900, 7200, 7400, 8200, and 8400 platforms, and these platforms do not support IP fragmentation of VXLAN frames. The VXLAN underlay network must therefore be able to support oversized or jumbo frames sizes. It is not possible to run VXLAN Gateway over the Internet as that would require IP fragmentation.

It should be noted running IP multicast on the resulting end-to-end L2 segment will present challenges. The Fabric L2 VSNs can be IP Multicast enabled, but no IGMP Sender or Receiver information will be exchanged on the VXLAN VNI segment. IGMP static entries would thus need to be provisioned in order to always flood the required IP Multicast streams across the VXLAN cloud. The other alternative is not to IP Multicast enable the Fabric L2 VSNs, which will result in all IP Multicast traffic to be treated like L2 multicast and be flooded everywhere. Both approaches will be handled using inefficient ingress replication on the VXLAN overlay.

Tip

Prefer a Fabric Extend approach if there is a requirement for IP Multicast.

Figure 63 shows the use of the same VXLAN Gateways and VXLAN interconnecting cloud to interconnect an L3 VSN service between the same Extreme SPB Fabrics. This approach essentially simply defines an L2 segment over the VXLAN overlay alone and places IP interfaces on this segment at both ends on the VXLAN Gateways. Those IP interfaces can belong to a VRF, and thus to an L3 VSN, and can be used to exchange IP routes for the interconnecting VRFs using a conventional IP or IPv6 routing protocol (RIP, OSPF, BGP). On the VXLAN Gateway, the VXLAN VNI is still mapped to a Fabric L2 I-SID, which in this case is used to simply map to a local CVLAN on the same node where the routing IP interface resides.

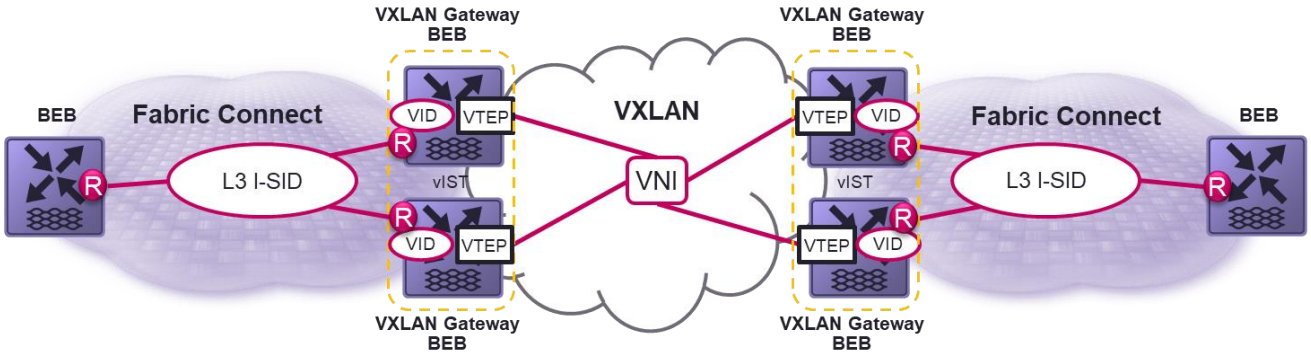


Figure 63 Extending an L3 VSN across Fabrics with VXLAN Gateway

Running IP Multicast across the extended L3 domain also presents challenges. It would require IP Multicast enabling the Fabric L3 VSN at each end and running PIM + PIM Gateway on top of the VXLAN Gateways.

Tip

Prefer a Fabric Extend approach if there is a requirement for IP Multicast.

Distributed Virtual Routing

This section will go into greater depth about Distributed Virtual Routing (DVR) in the context of the Extreme Fabric Connect architecture. DVR is an enhancement of Fabric Connect that allows both the use of a distributed anycast gateway and the ability to compute the shortest path to the individual host IP. This becomes hugely important in environments where the host IPs are mobile. The data center is the most challenging environment in this respect since the advent of server virtualization. VMs can be moved with ease, without even being stopped, from one physical hypervisor to the next and indeed across geo-redundant data centers. When those VMs migrate, they always take their IP address with them; this is necessary lest the applications running on those VMs would lose all their connections (open sockets) in the process.

Traffic Tromboning Challenges

The best way to understand the benefits of DVR is to first of all understand what would be the limitations of a Data Canter Fabric Connect architecture if it was not DVR enabled.

We already know that the SPB Fabric by definition always calculates the shortest path. But it will always calculate the shortest path towards some already defined Backbone MAC (BMAC) which constitutes the destination BEB for the traffic at hand. But since all traffic carried over the SPB Fabric is part of a service type (VSN), and the addressing used in these virtual networks is not the BMAC but some other L2 or L3 addressing scheme, there has to be a mapping between the two.

Hence if we look at L2 VSNs, these services learn end-user MAC addresses (CMAC) and associate these with the BEB's BMAC from which they have been learned. The assumption is that a given CMAC is behind a given BMAC and thus taking the shortest path to that BMAC also ensures the shortest path to that CMAC. This works fine for L2 flows that remain within the same L2 segment (source and destination are in same IP subnet) and is equally applicable to L2 flows within the data center.

Similarly with L3 VSNs, they exchange IP routes via IS-IS and install these IP routes in the relevant VRF IP routing table. In this case, it is an IP network that is being associated with the BEB's BMAC that has a local IP interface on that IP network. Again, the assumption is that any host with an IP address belonging to that network is residing behind that BEB and that taking the shortest path to that BMAC also ensures the shortest path to that IP address. This assumption generally holds true in the campus, but does not hold true in the data center where VM hosts are highly mobile in the east-west direction along server VLAN L2 VSNs.

Figure 64 illustrates the worst-case scenario of what could happen in an architecture where two geo-redundant data centers have been made into a single SPB Fabric, leveraging Fabric Extend, and where DVR is not being used. The server VLAN is L2 extended across both data centers using an L2 VSN. Both core routers in both data centers have an IP interface on the server L2 segment and act as default gateways for that same segment. They thus all announce the server subnet northwards toward the wider campus and the branch offices in the example. They will also have to use some form of gateway redundancy protocol, like standard VRRP, southwards toward the data center hosts (VMs). This can result in two forms of traffic tromboning.

In the first case, traffic from a branch office destined to a data center VM will make a forwarding decision based on the lowest cost to reach one of the four IP routers announcing the VM's IP network, but since there is no knowledge about where the VM is actually located, there is a 50% chance that the traffic will arrive to the wrong data center and then need to take a second high latency hop over potentially the same WAN (if the L2 VSN is Fabric Extended between the two data centers) to reach its destination.

In the second case, any data center L3 flow that needs to be IP routed either to reach another host in the data centers (L3 east-west flow) or to reach the wider campus (south-north) will need to hit the default gateway for the server segment. With standard VRRP, only one IP router will be acting as that default

gateway, and these flows might be forced to go across to the other data center unnecessarily resulting in a sub-optimal forwarding path, which is no longer the desired shortest path.

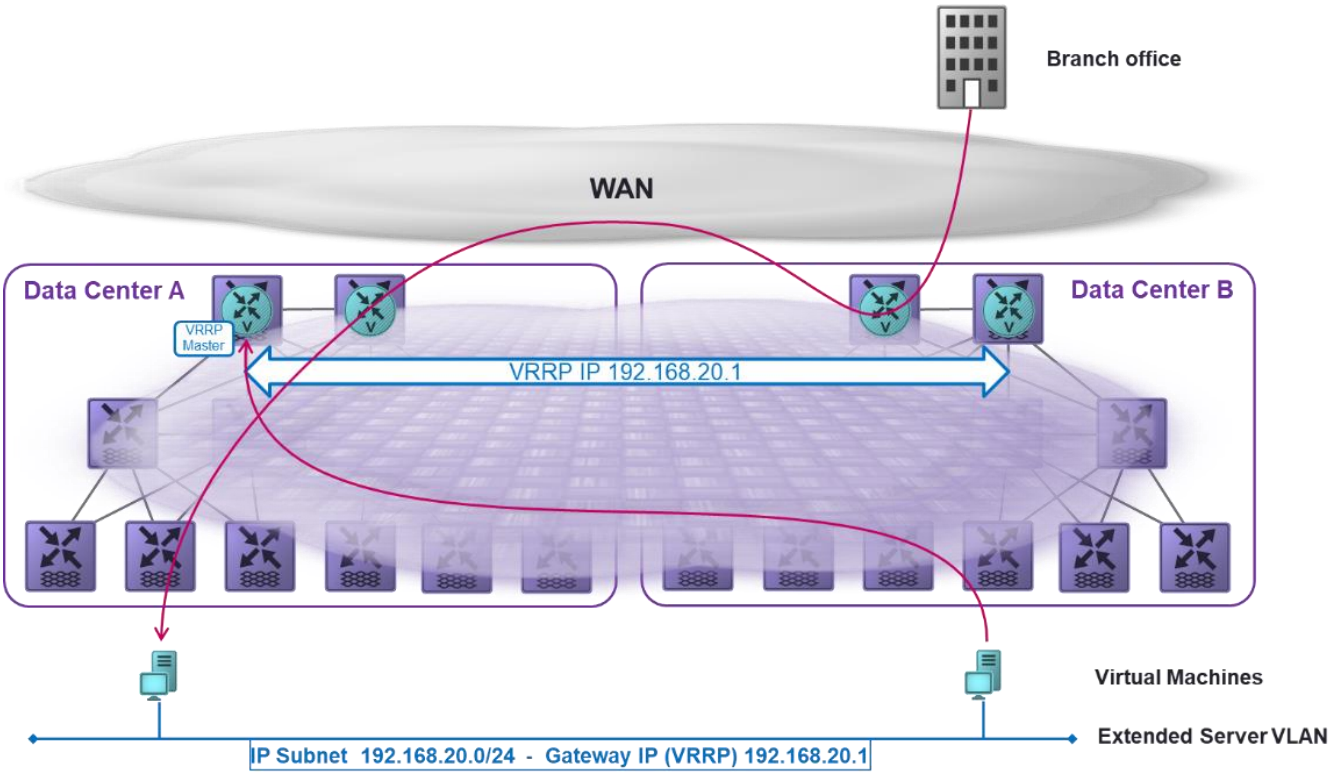


Figure 64 Traffic Tromboning Challenges in a Non-DVR Enabled Data Center

DVR solves these traffic tromboning issues by pushing the distributed anycast gateway functionality right on the ToR switches and by distributing the knowledge of host IP routes not only in the in the ToR switches but also where needed in the wider network.

DVR Deployment Model

DVR offers a highly scalable deployment model based around DVR domains. Within each DVR domain one or more DVR controllers must exist and these DVR controllers are then able to drive a number of DVR leaf nodes within that DVR domain. The size of a DVR domain is dictated by the maximum number of DVR leaf nodes supported within that domain as well as the maximum number of required server subnets/segments and server/VM hosts that are expected to be located across those DVR enabled server segments. Within a DVR domain, the DVR controllers and the DVR leaf nodes maintain knowledge of every host IP.

Note

A single DVR domain can currently scale up to eight DVR controllers, 250 DVR leaf nodes, up to 500 DVR enabled segments (L2 VSNs).

A DVR deployment will be limited to the maximum number of host records the VSP platform acting as DVR Controller can support, currently 32000 IPv4 hosts on VSP7200 and VSP8x00 and 40000 on VSP7400.

Premier license is required on DVR Controllers.

Based on the above scaling requirements, there are many ways to deploy DVR domains. A single DVR domain can span multiple smaller data centers provided that the overall host count, segment count, and DVR leaf count does not exceed the scaling requirements. More likely, each data center will become a DVR domain in its own right, as is depicted in Figure 65. Whereas for very large data centers a DVR domain can

become a “pod” building block within that data center and thus the single data center can be built with multiple DVR domains.

The ToR switches become DVR leaf nodes while the DVR controllers should strategically be placed in a position where all traffic is expected to transit in and out of the DVR domain. This is because any L3 flow which cannot be handled via host based routing within the DVR domain will automatically make its way to one of the nearest DVR controllers. However, the architecture allows complete freedom on how the DVR controllers and DVR leaf nodes are interconnected amongst each other. In smaller data center designs the DVR leaf nodes can be meshed between themselves with high speed interconnects and the DVR controllers can act as a distribution layer, only handling north-south traffic. It is also possible to have some of the ToRs switches themselves acting as DVR controllers.

For larger spine-leaf designs the DVR controllers can act as the second-tier spines in three-stage architectures or as third tier spines in a five-stage topology. In the latter case, the second-tier spines will be constituted by non-DVR aware SPB BCB nodes, which simply act as a transport layer. This is as shown in Figure 65. In large scale spine-leaf designs it is also possible to locate the DVR controllers on border leaf nodes that connect into the spine layers much like any other leaf nodes. This is commonly done with EVPN based designs, which helps ensure equal cost multipath towards those border leaf nodes from all the other leaf nodes.

If the DVR controllers are acting as spine nodes, it is not necessary to make all the spine nodes DVR controllers. In general, the DVR controller needs to be redundant so there should always be a minimum of two DVR controllers per DVR domain. However, the reason for scaling up more than two DVR controllers will be dictated by the number of exit points from the DVR domain as well as the number of equal cost multi-path required to distribute north-south traffic leaving the DVR domain. Also in the presence of traffic destined for unknown IP hosts within DVR enabled segments the DVR controllers are responsible for performing the necessary ARP-ing, which will be distributed across the available DVR controllers. However, as will be explained later on, the chances of having unknown host IPs within DVR are insignificant provided that all server/VMs have been configured with a Default Gateway IP.

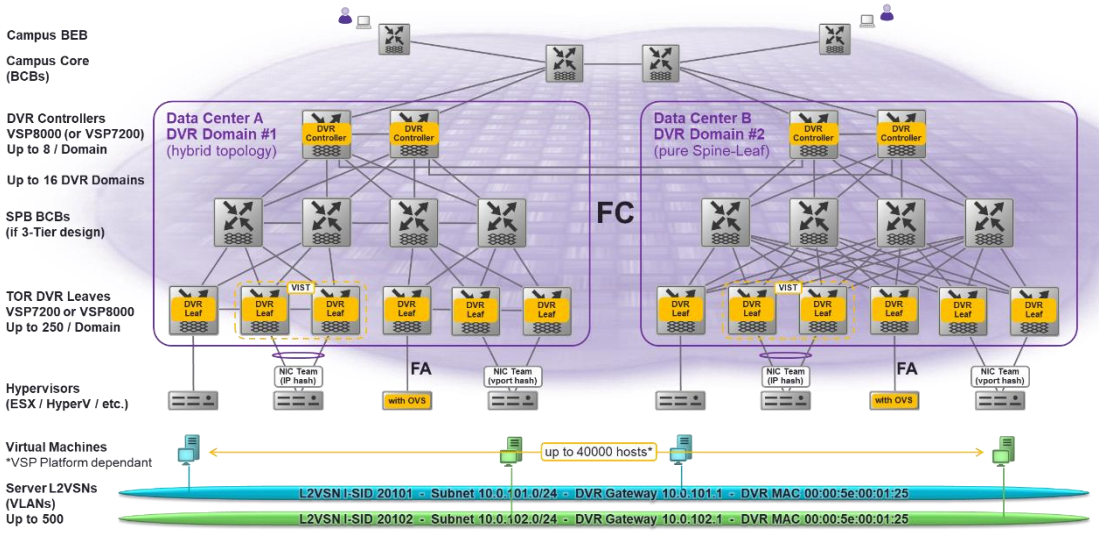


Figure 65 DVR Model and Scaling

The DVR leaf ToR switches will almost always be switches offering 10GbE or 25GbE to the access. In some Enterprise class data centers it is still common to also deploy a number of GbE access ToR switches, yet these are typically destined for older bare metal servers or for Lights-Out management connectivity of hypervisors. In both cases the resulting host IPs are not mobile and there is no significant benefit in making these gigabit ToRs, DVR leaf nodes. Instead the Extreme DVR architecture can accommodate gigabit ToRs to be connected as FA Proxy switches, either directly into a pair of DVR controllers, or directly into a pair of DVR leaf nodes, as was shown in Figure 11.

Caution

The DVR leaf will most likely be an Extreme Networks 10GbE VSP 7200 or 25GbE VSP 7400 or VSP 8k platform, which provides the maximum scaling for DVR. For GbE access connectivity, the VSP 4k platform can be used as a DVR leaf but with a reduced scaling of not more than 6000 IPv4 hosts within the DVR domain.

A better approach for GbE connectivity would be to deploy ERS or ExtremeXOS platforms in an FA Proxy role, which can be either connected into DVR leaf or DVR controller nodes.

Note

The FA Proxy ToR must connect to a DVR enabled node, either a DVR controller or a DVR leaf that will be acting as FA Server. Also for redundancy it is desirable to connect the FA Proxy using MLT uplinks into an SMLT cluster, which means the DVR nodes would need to be provisioned with a vIST. This will typically be the case for DVR leaf nodes but not necessarily for the DVR controllers.

Tip

In the Extreme Networks DVR implementation, a pair of DVR controllers can be SMLT clustered via vIST. While this has no impact on DVR scaling figures, it only makes sense if there is a requirement to dual home GbE ToR FA Proxy switches into the DVR controllers.

From a configuration point of view, a DVR domain comes into existence once a node is configured to act as DVR controller for a DVR domain ID. Once a DVR controller has been defined, a number of DVR leaf nodes can be provisioned into the same DVR domain id.

Note

The DVR domain ID is a number from 1 to 255.

Tip

In the Extreme Networks implementation, a node can be made a DVR controller without any system restart or interruption.

Caution

Configuring a ToR switch as DVR leaf requires a system restart and a purge of any VLAN and L3 configuration since this will thereafter be driven by the DVR controllers.

Within a DVR domain, the DVR controllers and DVR leaf nodes communicate over a highly efficient DVR domain reserved I-SID that is used to carry point-to-multipoint IS-IS control plane signalling. The DVR domain ID is used to allocate a unique reserved domain I-SID for the DVR domain. A separate reserved DVR Backbone I-SID is also reserved for communication between DVR controllers belonging to different DVR domains, which is why a limit also exists in the total number of DVR domains supported.

Note

The current DVR architecture can scale to a maximum of 16 DVR domains.

Note

The DVR Backbone I-SID uses reserved value 16678216.

The DVR domain I-SID uses reserved range 16678216 + DVR domain ID; hence the DVR domain ID 1 I-SID will take value 16678217, domain ID 2, 16678218 and so on.

Also, if DVR leaf nodes are SMLT enabled, I-SID range 16677215 + Cluster ID is used for the vIST connection.

Tip

There is no need to configure the DVR domain & Backbone I-SIDs. These are automatically inferred once the node has been configured as a DVR controller or leaf for a given DVR domain.

Once the DVR controllers and DVR leaf nodes are in place, it becomes possible to create DVR-enabled segments. From a configuration perspective, these are created much in the same way as an L2 VSN is created and to which a local IP interface is assigned. The IP interfaces can belong to one of many VRFs which in turn may belong to fabric-wide L3 VSN tenants. The configuration needs to be consistent across the DVR controllers in the sense that all DVR controllers within the same DVR domain should have the same DVR segments.

On these segments, each DVR controller will have its own local and unique IP interface as well as virtual DVR Gateway IP which needs to be the same IP address across all DVR controllers. This is conceptually similar to how VRRP is configured. It is this DVR Gateway IP address that will need to be used as Default Gateway IP by all server/VM hosts residing on the DVR segment.

Note

If a DVR segment is only configured on one of the DVR controllers, it will be immediately active across all the DVR leaf nodes and will be operational. However, for resiliency and load balancing purposes, that DVR segment needs to be provisioned across all DVR controllers in the DVR domain.

DVR segments will typically also get created across multiple DVR domains, as shown in Figure 66, based on the data center requirements for how far a server VLAN/segment needs to stretch. As such the DVR segment will need creating across all DVR controllers for all the DVR domains where it is required. The same DVR Gateway IP will need to be configured on all DVR controllers.

As soon as a DVR segment has been configured on a DVR controller, the DVR controller immediately advertises the existence of the segment via the Domain DVR I-SID with a single point-to-multipoint update to reach all DVR leaf nodes within the DVR domain. These use IS-IS based LSP updates which are reliably acknowledged and contain information about the DVR segment including:

- L2 I-SID
- L3 I-SID (0 if IP Shortcuts)
- DVR Gateway IP
- IP Mask
- IP routes
- IP Multicast state and IGMP version to use

The information exchanged via the DVR domain I-SID will be stored by all DVR nodes within the DVR domain in a separate LSDB from the SPB's regular IS-IS LSDB so that the latter does not get overloaded. The DVR leaf nodes will use this information to activate the distributed anycast gateway in their data plane, which will then allow them to handle ARP/RARP/GARP directly with the server/VM hosts and to perform

first-hop IP routing for any L3 flow received from those hosts. In case IP Multicast for the segment was enabled on the DVR controller, the DVR leaf nodes will also automatically activate IGMP on the same segment using the requested IGMP version. When the DVR leaf activates a new DVR interface for the DVR segment, it will allocate a local VLAN-ID and VRF-ID out of its local pool.

Tip

Each DVR leaf has a local pool of 4000 VLAN-IDs and 255 VRFs which are no longer user configurable but instead get allocated sequentially to the DVR interfaces created on the DVR controllers.

Caution

Currently a DVR leaf will support a maximum of 500 DVR interfaces and hence will at most consume 500 VLAN-IDs from the available pool.

Note

Extreme Networks VSP 7200 and 8k platforms by default support a maximum of 24 VRFs. In order to scale to 255 VRFs it is necessary to activate the vrf-scaling boot flag, which in turn will reduce the VLAN-ID pool to 3500. The vrf-scaling boot flag, if set, should be enabled across all DVR leaf and controller nodes.

By default, the DVR controllers will also push to the DVR leaf nodes a default route to themselves, which will result in every DVR leaf also installing a default route towards the DVR controllers. IP ECMP will always be enabled on the DVR leaf nodes and if more than one DVR controller is present in the DVR domain, and with SPB equal shortest path, then these default routes will leverage IP ECMP to distribute traffic needing to leave the DVR domain across the available DVR controllers.

In DVR designs where multiple exit points are desired from a given DVR domain, the DVR controllers placed at those exit points can be configured to not advertise a default route and/or instead advertise a more specific IP route leveraging DVR redistribution.

It should be noted that not all server segments operating in a DVR domain need be DVR enabled segments. It is perfectly acceptable to have some segments created with VRRP on the DVR controllers and for these segments to be simply terminated on DVR leaf nodes using Switched-UNI end-points. In this case, these segments will not be handled by DVR signaling and will operate in the traditional way, meaning that the DVR leaf will only be able to perform L2 switching for these segments and any IP routing will need to be performed on the DVR controllers. This flexibility is important to ensure that data center DVR architectures can be made to work even for applications that are currently not supported by DVR.

Note

Currently DVR segments can only be created for IPv4. The DVR architecture is already defined also for IPv6 operation but IPv6 support will become available in a future software release. Should IPv6 be required in a server segment, the segment will need to be provisioned using traditional VRRP v3 for both IPv4 and IPv6.

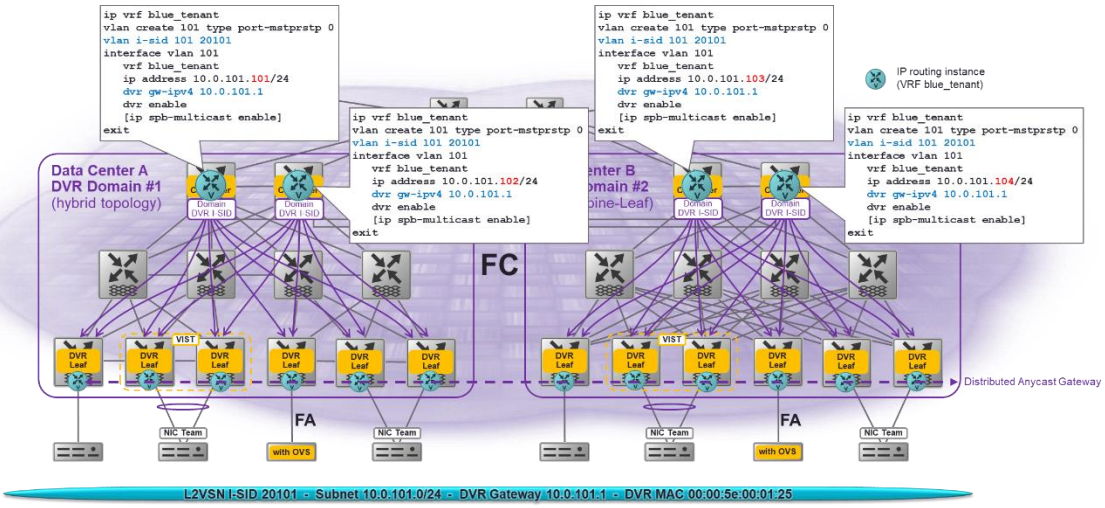


Figure 66 – DVR Gateway IP Provisioning on DVR Controllers Only

The configuration pushed down to the DVR leaf will be limited to the creation of a DVR interface and associated IP routes for the VRF context (L3 VSN). This configuration will not include configuration of the DVR leaf access ports which will need to be provisioned separately to associate those ports with one or more DVR segments via Switched-UNI (VLAN-ID + L2 I-SID pair) bindings. These can be provisioned on the DVR leaf node itself, or automated via Extreme Management Center ExtremeConnect (with VMware or Microsoft HyperV integration) or via Fabric Attach where Open vSwitch (OVS) is used in the hypervisors. These schemes were already shown in Figure 12.

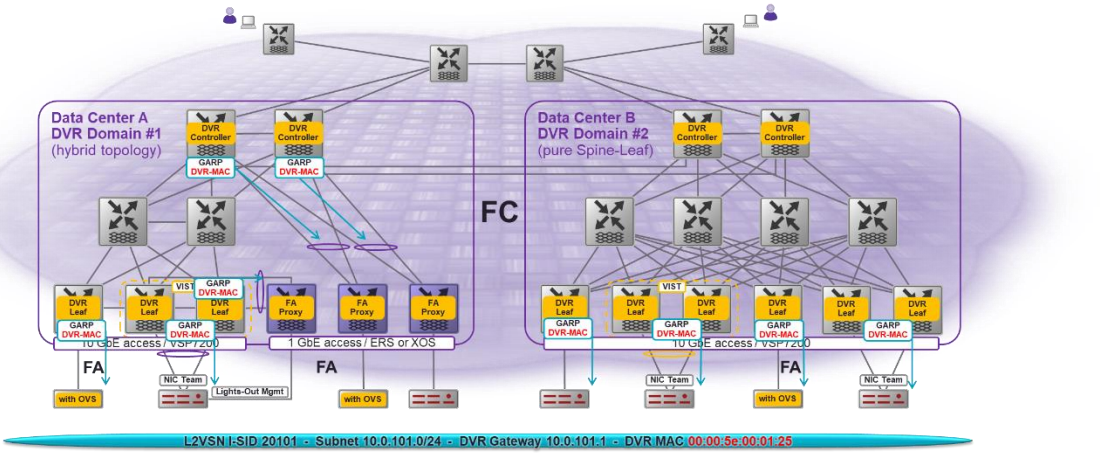


Figure 67 – How DVR Ensures MAC Learning of DVR Gateway MAC

A further point of consideration is that the DVR Gateway virtual IP defines a corresponding virtual MAC address, much like VRRP does. This DVR MAC address becomes jointly owned by all DVR-enabled nodes (controller and leaf nodes) so as to ensure the distributed anycast gateway functionality. Yet there needs to be a mechanism for this MAC address to be properly MAC learned by any L2 switch that might be in use in front of the DVR controller or DVR leaf nodes. This includes the software-based vSwitch located inside the server hypervisors as well as any FA Proxy switch connected to DVR-enabled nodes. Failure to properly learn the DVR MAC would result in these L2 switches flooding all traffic destined to the DVR distributed anycast gateway.

DVR does not use or need a chatty hello protocol to achieve this (as VRRP does) and instead ensures that the DVR MAC is always properly learned by generating Gratuitous ARP (GARP) messages for the DVR Gateway IP every 100 seconds across all DVR segments, but only on UNI ports (not on NNI ports). This is depicted in Figure 67.

DVR Host Tracking and Traffic Forwarding

It is worth looking at how DVR learns and tracks data center hosts as well as how it handles traffic forwarding for L3 flows within the data center from which it will become apparent how powerful DVR is compared to competing data center architectures.

Pretty much every operating system, once configured with a Default Gateway IP, will proceed out of its own initiative to perform ARP resolution for that gateway IP in order to obtain the gateway's MAC address which can then be used to transmit any traffic destined outside of the host's own IP subnet.

If the DVR leaf has an active DVR interface for the server segment on which the ARP request is received, it is the DVR leaf which will handle the ARP exchange directly, for the DVR Gateway IP, by providing the host with the DVR Gateway MAC in use on the segment.

Tip

In the DVR implementation, the DVR Gateway MAC is always the same across all DVR segments: 00:00:5e:00:01:25 for IPv4 (will be 00:00:5e:00:02:25 for IPv6). This allows a better use of silicon hardware resources on the DVR leaf nodes to scale to many more routed MACs than would be otherwise possible.

Responding to host ARP requests for the DVR Gateway IP, allows the immediately attached DVR leaf node to discover the existence of the host. The very same DVR domain I-SID which was used by the DVR controllers to provision the DVR interface on the DVR leaf nodes is used by the DVR leaf nodes to advertise the existence of the server/VM host IP, in one efficient single update to all other DVR leaf nodes and DVR controllers. This allows all nodes within the DVR domain to build and maintain IP host tables covering every host within that domain. The DVR controllers will in their turn use the DVR Backbone I-SID to advertise the existence of host IPs across other DVR domains. The result is that a DVR leaf will be aware of all host IPs within its own DVR domain, while the DVR controllers will be aware of all host IPs across all DVR domains.

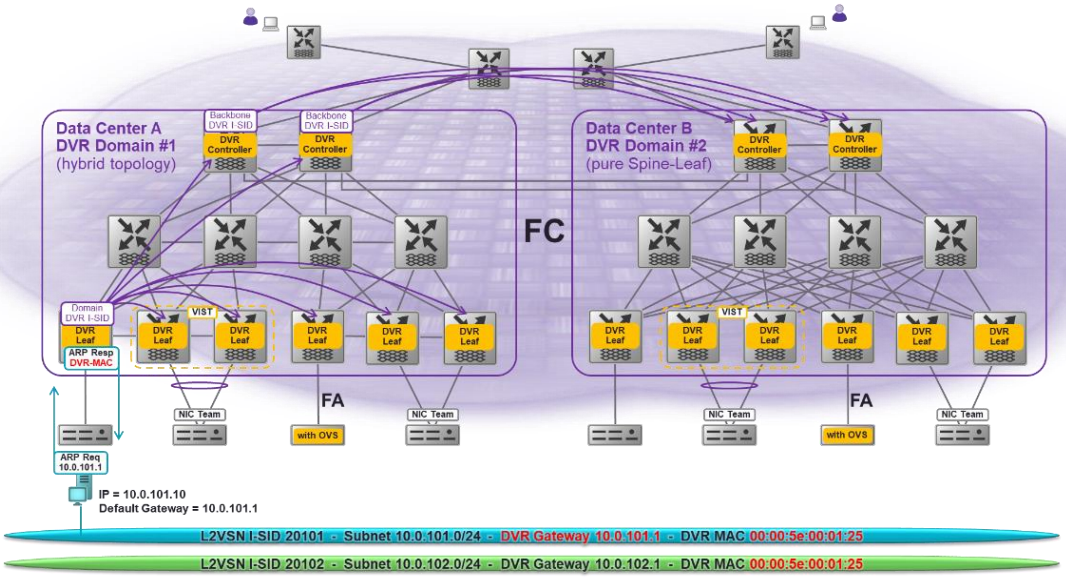


Figure 68 DVR's Distributed Anycast Gateway in Action

This raises an important point around the maximum host scaling of DVR. A DVR leaf will only need to program host IP routes for IP hosts located within the DVR domain. Whereas a DVR Controller will not only program host IP routes for IP hosts located within the DVR domain but also every host which is seen in other DVR domains if these hosts are located on L2 VSN segments for which the DVR Controller has a local DVR gateway IP interface. Hence for DVR domains which share the same DVR enabled L2 VSNs the maximum DVR host scaling will be network wide. Whereas if DVR enabled L2 VSNs are not extended across DVR doamins then the maximum DVR host scaling limit will apply to the DVR domain only.

When it comes to traffic forwarding, any L3 flow emitted by the server/VMs will always have destination MAC the DVR Gateway MAC and will thus benefit from efficient data plane IP routing from the immediately attached DVR leaf. In the example shown in Figure 69, all VMs are located in one of two DVR segments which both belong to the same L3 VSN tenant. When VM1 sends traffic for VM2, which is another host in the same DVR domain but on a different segment and thus different IP subnet, the DVR leaf to which VM1 is connected will be able to perform shortest path host based IP routing directly towards the DVR leaf where VM2 is connected. In the case where both VM1 and VM2 were connected to the same DVR leaf, the ToR switch would be able to perform local routing.

When VM3 sends traffic to VM6, which is again in a different segment but this time is a host located in a different DVR domain, the DVR leaf to which VM3 is connected will not have an IP host route for VM6. The IP routing lookup performed will thus result in the DVR leaf forwarding the traffic via an IP hash towards one of the nearest available default routes it has towards its DVR controllers. The DVR controller receiving the traffic will now be able to perform host based IP routing directly to the DVR leaf in the destination DVR domain. Note that the DVR controller defines the exit point for the DVR domain which is thus also the entry point for the receiving DVR domain from a physical topology perspective. In the diagram, the path between the DVR controller from domain 1 and the destination DVR leaf in domain 2 is the shortest path, which happens to transit via the DVR controller of domain 2 but is not required to do so.

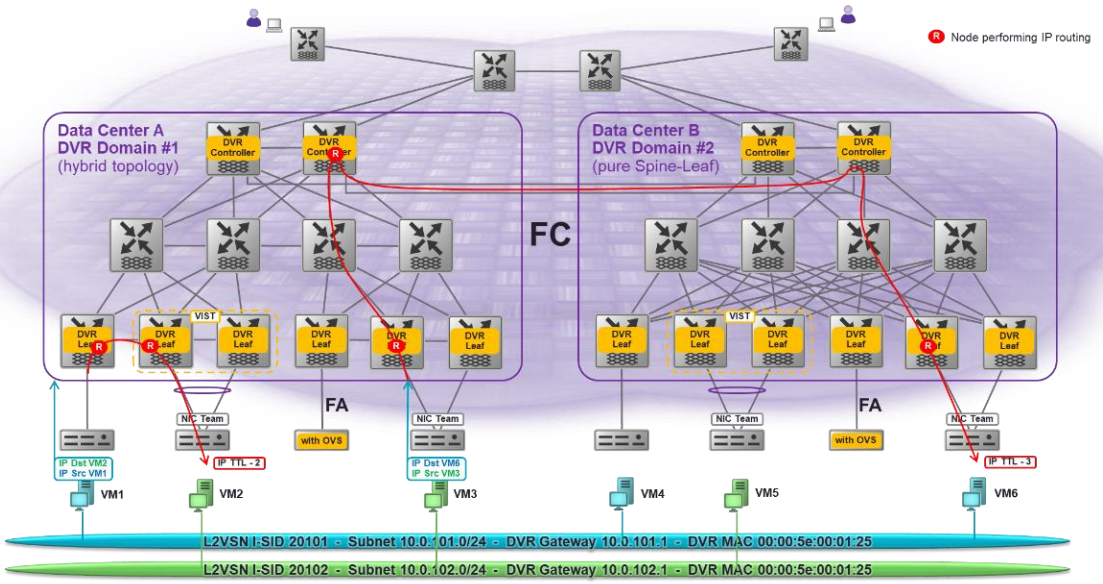


Figure 69 DVR East-West Traffic Forwarding for L3 Flows

A few words about what would happen in the eventuality that a host emitted L3 traffic destined for a yet unknown host IP located somewhere in the data center. In this case the DVR leaf would again forward the traffic to one of its available DVR controllers, which would deem the traffic as destined to a local DVR segment for an end-station for which no DVR host IP is yet known. The DVR controller will thus have to perform ARP resolution for that host. In the case where the host exists and does reply, its locally attached DVR leaf will intercept the ARP response and the usual DVR signalling process will repeat itself so that the DVR leaf will synchronize the host IP within the DVR domain and the DVR controllers will synchronize it across the DVR Backbone I-SID. Such ARP resolution activity can be load balanced across all the DVR controllers within the DVR domain. Yet this scenario is unlikely because most host operating systems will always pro-actively perform ARP resolution for their configured default gateway IP and not simply wait to do so until they have L3 traffic flows to transmit.

So, DVR uses host-based IP routing to guarantee shortest path and lowest latency of data center L3 flows. Yet one of the challenges of data center with virtualized servers is that VMs are mobile and that DVR needs to be able to keep track of VM movements within and across DVR domains. Different hypervisor vendors have come up with slightly different schemes but all provide mechanisms to aid the network infrastructure to keep track of VM movements so as to update host MAC and IP records. There are two approaches used:

one makes use of GARP, which allows an IP host binding to be changed or updated and the other makes use of Reverse ARP (RARP), which allows a host MAC to be seen to move. In both cases, it is the destination hypervisor itself that generates these GARP/RARP messages after completion of a VM move and on behalf of that VM. It is not the VM itself that generates these packets.

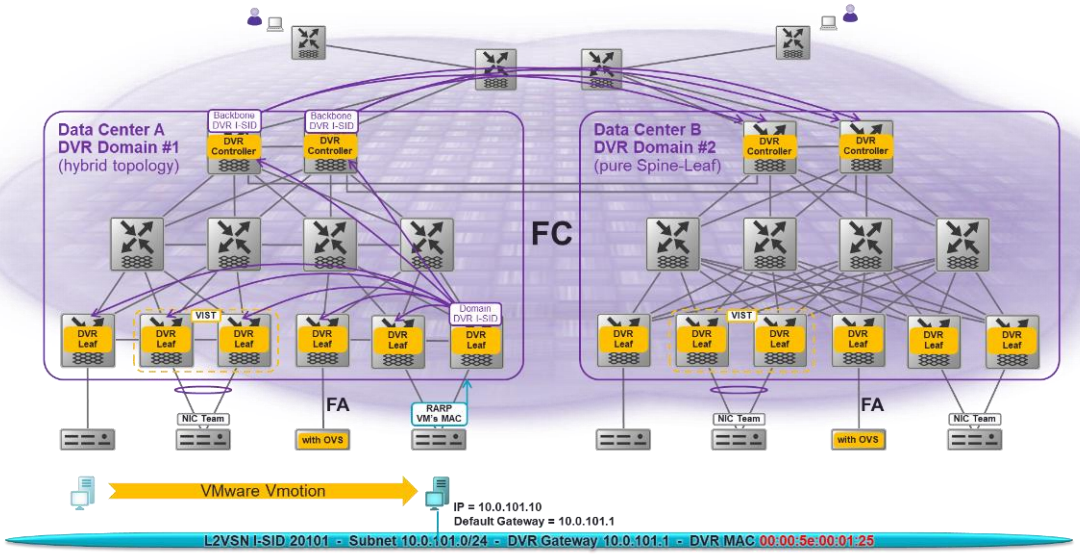


Figure 70 DVR Using RARP with VMware Vmotion

From the most popular hypervisor vendors we will note VMware, which uses a RARP scheme, and Microsoft Hyper-V, which uses a GARP scheme. A DVR leaf is able to process both GARP and RARP messages and will thus work with both. Figure 70 and Figure 71 show how DVR handles a VM move event with VMware Vmotion (uses RARP) and Microsoft Hyper-V Live Migration (uses GARP), respectively.

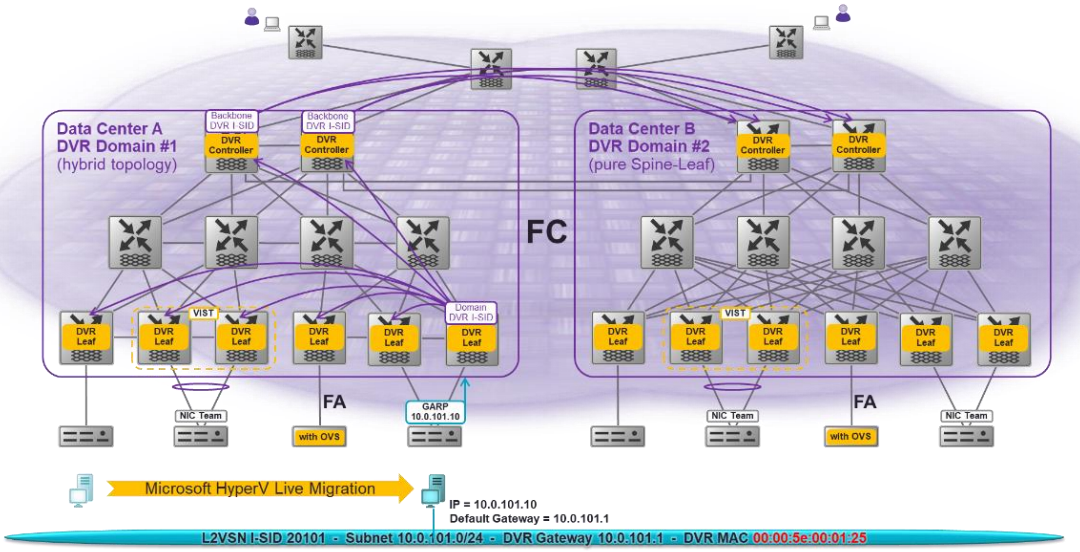


Figure 71 DVR Using GARP with Microsoft Hyper-V Live Migration

In both cases the DVR leaf intercepting the GARP/RARP packet has enough information to identify the host IP which has moved and to signal the move via the same DVR domain I-SID signaling as before. In reality, it is all the DVR leafs present on the DVR segment that will receive the host GARP/RARP (these are L2 broadcast packets which are naturally flooded across the L2 VSN segment) and each node is capable of caching a short-lived ARP entry to reflect the update to the host IP route. The original DVR leaf owner for the host will at this point withdraw its ownership of the host via DVR domain I-SID signaling. Only one DVR leaf will receive the GARP/RARP packet on a UNI port and this DVR leaf will become the new owner of the host IP, and will therefore signal this host IP route via DVR domain I-SID signaling. The short-lived ARP

entries will age out fairly quickly but by that time all DVR hosts in the DVR domain will have an updated host IP route for the VM.

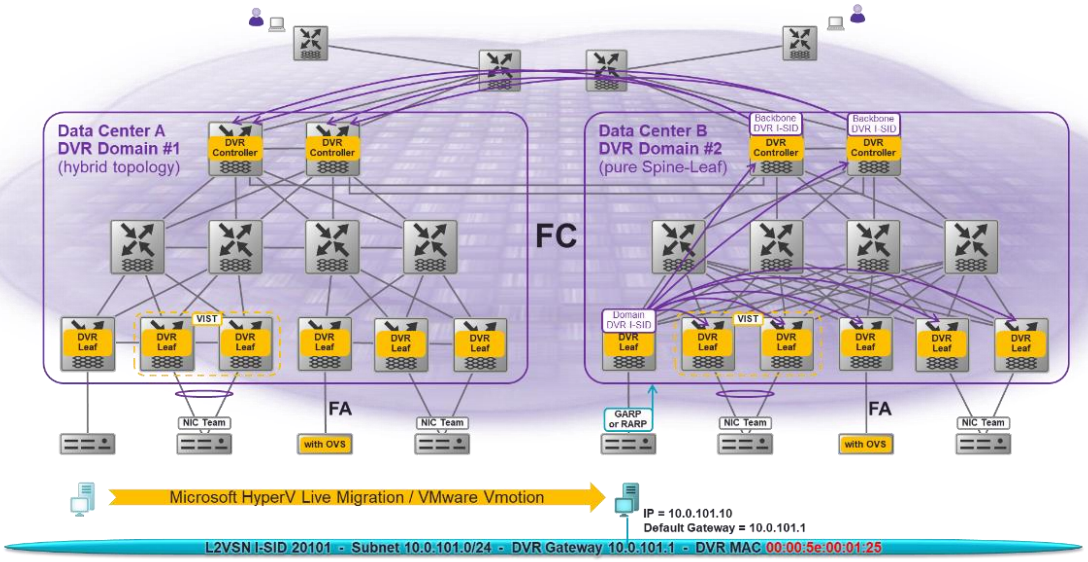


Figure 72 DVR with VM Migration Across DVR domains

In the case of a VM migration across DVR domains, the mechanism is much the same. Again, the GARP/RARP will be received by all DVR nodes, across both DVR domains. The DVR leaf, which used to be the owner of the host route in the source DVR domain, will withdraw its host route while the DVR leaf, which becomes the new owner of the host route in the target DVR domain, will signal the host route within the target DVR domain only. All DVR leaf nodes across both DVR domains will have installed short-lived ARP entries for the host in the process. Except that now, once these short-lived ARP entries are aged out, knowledge of the VM host will disappear in the source DVR domain and subsequently any traffic destined to that host from this DVR domain will have to follow the default routes toward the DVR controllers. The result is a robust and efficient host-based routing architecture that is able to follow and keep track of all data center host IPs.

Eliminating North-South Tromboning

North-south traffic tromboning is a phenomenon that can occur in the presence of geo-redundant data centers where server subnets span both data centers and thus both data centers are advertising the corresponding IP subnet of that server segment into the wider campus/branch network. The routers operating in the wider campus/branch network typically have to make a routing decision based on the IP subnet alone and will typically always choose the path to the nearest data center, which is not necessarily the shortest path to the server host/VM that is mobile and could be located in either data center. Depending on the network design, this can result in sub-optimal traffic forwarding and a higher latency inflicted on the applications running from the data center.

Extreme Networks DVR architecture offers the capability of minimizing these traffic tromboning effects by allowing the network administrator to identify those critical data center host IPs that act as termination points for all client-server interactions for a given business application and then allowing those selected host routes to be taken into consideration by campus/branch routers in order to always ensure the shortest path to reach those data center resources.

Tip

Note that in modern data centers, not all host IPs communicate north-south with users. A majority of those hosts are either backend databases or batteries of servers operating behind load balancers that are used exclusively for east-west communication within the

data center(s). So only a small proportion of data center host IPs will need to be optimized against north-south traffic tromboning.

There are two possible approaches with DVR. In the first case, the Extreme's Fabric Connect is end-to-end extended from the data centers all the way to the wider campus, as depicted in Figure 73.

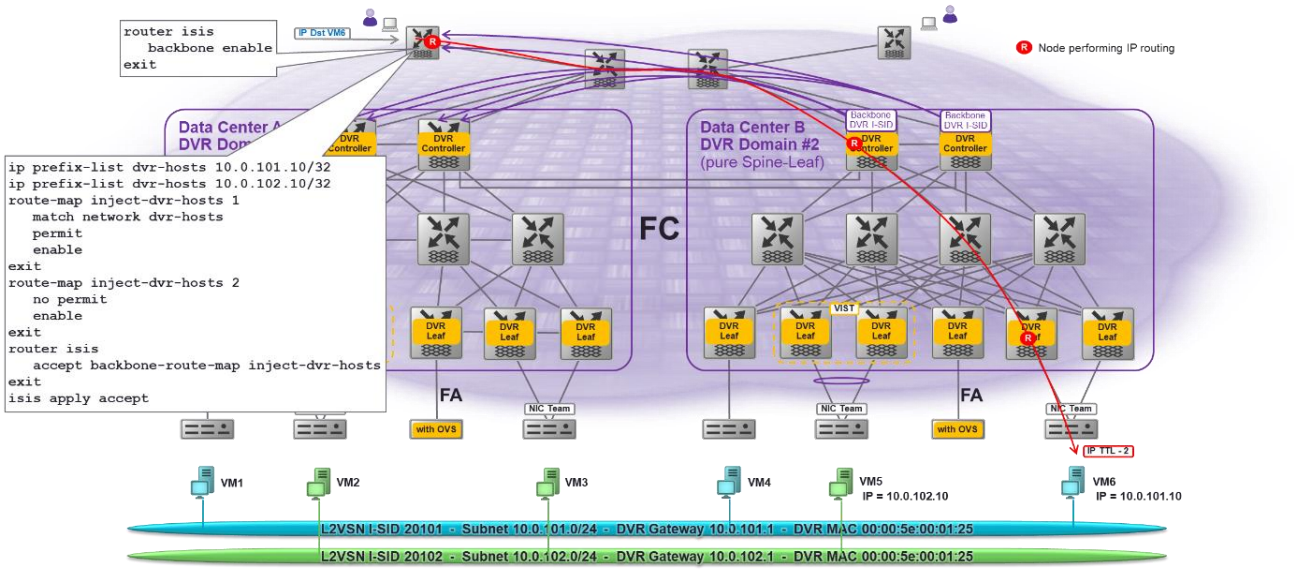


Figure 73 Eliminating North-South Tromboning with DVR, Over Campus Fabric

The users will be located in client IP subnets where a campus BEB will be acting as default gateway for those users. On these campus IP Routers, Extreme offers the ability to let these BEBs join the DVR Backbone I-SID tree in a listening only role. This is achieved by activating the IS-IS Backbone global switch, which does not enable DVR on the node, but simply allows that BEB to obtain and maintain a software only table of all data center hosts, across all DVR domains.

Tip

The ability to let an L3 BEB join the DVR Backbone is possible on all of the Extreme Networks VSP platforms running VOSS software.

The obtained software table of data center hosts is not programmed in hardware and is not inserted in the IP routing table and will not have any effect in altering traffic forwarding paths at first. For that to happen the network administrator will need to take an additional step by defining an accept policy to inject a selection of those hosts into the corresponding VRF IP routing table. Once the backbone accept policy is applied, the BEB's VRF IP routing table will be augmented to include host routes for the injected data center hosts.

Caution

The backbone accept policy should be defined in such a way that only a selection of relevant hosts is injected for which traffic tromboning effect needs to be optimized. Do not inject all data center hosts as there could be thousands of hosts and this could exhaust the BEB's IP routing table scaling resources.

The resulting host routes on the campus BEBs will thus always point to the DVR controllers of the DVR domain where the host is located. These routes will also automatically update themselves should the host VM be migrated to a different DVR domain.

Caution

Because these hosts routes will point to the nearest DVR controller of the DVR domain where the host is located, this form of traffic tromboning optimization will only work if the two data centers have been deployed into different DVR domains. If the two data centers are part of one single DVR domain, then this optimization is not possible.

In the second case, we shall consider a network deployment where the Extreme Fabric Connect does not extend beyond the data centers and instead the campus/branch locations are interconnected via a traditional WAN or customer owned MPLS network. In this case, the campus/branch IP routers will not be fabric-enabled but will be running traditional IP routing protocols such as BGP, OSPF, or RIP.

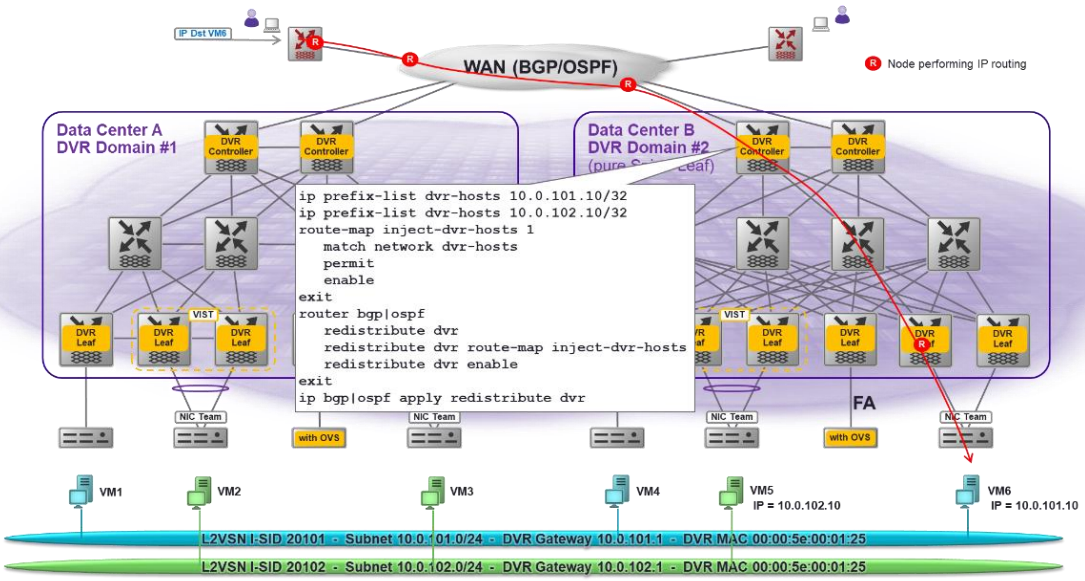


Figure 74 Eliminating North-South Tromboning with DVR, Over Legacy WAN

The assumption here is that the DVR controllers acting as gateways for all traffic into and out of the data center DVR domains will also be running those same traditional IP routing protocols facing the WAN or customer owned MPLS backbone. In this case, the design approach is to again identify which data center host IPs can benefit from traffic tromboning optimization, but this time action will be taken on the DVR controllers so they can redistribute only those host IP routes into either BGP or OSPF.

Caution

DVR host route redistribution currently is supported towards BGP or OSPF. RIP is currently not supported.

This is achieved via DVR redistribution which will result in the DVR controller injecting selected host IP routes into either BGP or OSPF but only if the hosts in question are located in the same DVR domain.

DVR limitations and Design Alternatives

To date there are still some limitations in the use of DVR, so this section will detail what those limitations are and how to design around them. Some of these limitations have already been mentioned in the preceding sections but will all be covered again here in order.

The limitations and the corresponding design alternative to use are the following:

- Only IPv4 support and no IPv6 support: If IPv6 is required on some server segments, these cannot be configured as DVR segments. Instead they will need to be configured using traditional VRRPv3 for both IPv4 and IPv6. The DVR leaf will then only perform L2 switching for the segment and the DVR controller will perform IPv6 routing with VRRP.
- DVR segments can only terminate on DVR enabled BEBs: An L2 VSN on which a DVR Gateway IP interface has been defined must only be terminated on a DVR controller or a DVR leaf BEB. Allowing a non-DVR BEB to terminate the same L2 VSN will result in connectivity problems for all devices behind that non-DVR BEB. Do not create a DVR Gateway IP on an L2 VSN segment until all terminating BEBs are DVR enabled.
- FA Proxies terminating DVR L2 VSN I-SIDs must always be connected to a DVR enabled BEB: This is a direct extension of preceding limitation. Because only a DVR enabled node can terminate a DVR enabled segment, any FA Proxy needing to terminate those same segments will need to connect to a DVR enabled FA Server BEB. Both the DVR controller and the DVR leaf can act as FA Server BEB.
- No Microsoft NLB support: NLB is not supported on a DVR segment. If NLB is required, a VRRP interface should be configured on the DVR controller instead.
- No L2 VSN IP Multicast support: DVR fully supports IP Multicast on DVR segments; however these DVR segments always constitute IP routed segments operating under IP Shortcuts or L3 VSN; meaning that the data center server segment does not extend outside the data center and is IP routed towards the wider campus. Extending a non-IP routed L2 segment from a campus BEB all the way to a DVR BEB is possible, but it is not possible to IP Multicast enable this segment on the DVR BEB (as this would require IGMP snoop enabling a CVLAN on the DVR leaf for which no CVLAN configuration is allowed). If L2 VSN IP Multicast is required to extend into the data center, terminate the L2 VSN on a DVR controller instead of a DVR leaf.
- No E-TREE support on DVR leaf: No CVLAN configuration is allowed on a DVR leaf, and by definition an E-TREE L2 VSN needs to be terminated on a Private VLAN CVLAN. If an E-TREE L2 VSN is required, this will need to be terminated on a DVR controller.
- No IP routing protocols allowed on DVR leaf: It is not possible for a DVR leaf to run any IP routing protocol. IP routing protocols can be run on a DVR controller.

Quality of Service

Initial Considerations

The objective of network Quality of Service (QoS) is to allow different type of traffic to contend inequitably for shared network resources. The goal is to converge applications such as voice, video and data over the same network infrastructure. Voice is low constant bandwidth but is a real-time application and thus does not tolerate delay (latency). Data is bursty and can tolerate high levels of latency, while video is usually somewhere in between, depending if it is used for real-time video conferencing or IPTV video streaming.

QoS is all about managing buffer resources as well as buffer congestion in a different way for the different traffic classes. The main consequences of buffer congestion being delay, jitter, and packet drops.

The IETF reference model for QoS is the Differentiated Services (DiffServ) Model (RFC 2475), which defines a way to mark IP packets with a Differentiated Services Code Point (DSCP) and Per Hop Behaviors (PHB). Each DSCP has a corresponding PHB such that traffic entering the network with a certain DSCP class must be given the same PHB by every node it traverses across the network. Traffic is assigned to one of the discrete DSCP classes either by the application on the source or by network policies defined at the network access, and the corresponding PHB is derived from the DSCP marking by subsequent hops in the network.

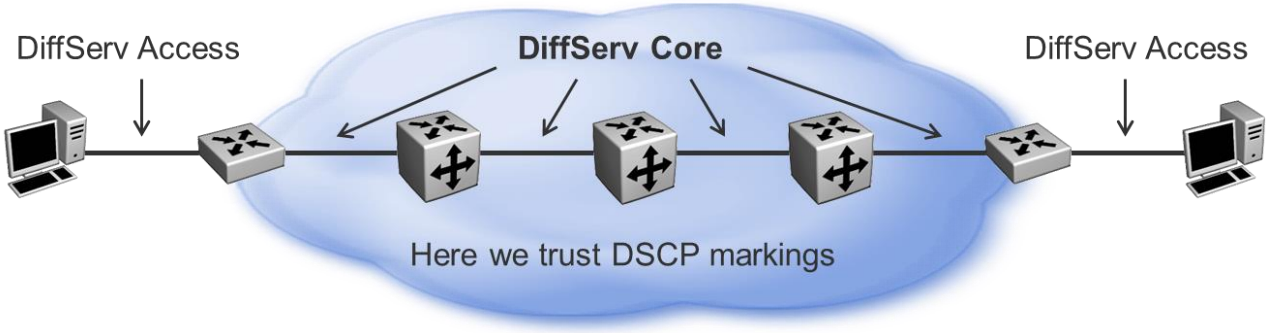


Figure 75 QoS DiffServ Model

In this model, a router IP interface can be configured either as DiffServ Core (Trusted) or DiffServ Access (Untrusted). A DiffServ Core port will always derive the QoS and hence the PHB from the DSCP marking recorded in the packet IP header. Whereas a DiffServ Access port will derive the QoS exclusively from access policies defined on the port and will then update the DSCP for the packets accordingly (or reset DSCP to zero if no policy could be applied to a given packet).

The DSCP field is part of the Layer3 IP header (in IPv4 as well as IPv6) so while every IP router or L3 switch will have no problem deriving the PHB from the DSCP field, the same cannot be expected from all Layer2 Ethernet switches that will not even be looking at the IP header to forward packets. For this reason, and provided that q-tagging is in use, the PHB can also be derived from the Ethernet 802.1p bits. In general, the Ethernet 802.1p bits should always be set to a consistent value with respect to the DSCP marking and every IP router or L3 switch will automatically update (using egress mapping tables) the 802.1p bits to a DSCP consistent value when IP routing a packet onto Ethernet q-tagged segments.

Tip

All of Extreme Networks ERS, VSP and ExtremeXOS series of Ethernet access switches are able to derive QoS from the DSCP marking.

With today's platforms, which can act both as L2 and L3 switches, when an ingress packet is received carrying both DSCP and 802.1Q-Tag 802.1p-bit QoS markings, the default behavior is to QoS classify using the markings that correspond to the way in which the packet will be forwarded. Hence a packet that is L2 switched (bridged) within the VLAN (or L2 VSN) will use the p-bits to derive the QoS PHB, while a packet that is L3 switched (IP routed) will use the DSCP markings to derive the QoS PHB.

Tip

On Extreme Networks ERS, VSP and ExtremeXOS platforms, a port can be independently configured as L2-Trusted or L2-Untrusted, and L3-Trusted or L3-Untrusted. An L2-Untrusted + L3-Trusted configuration will ensure that DSCP is used to derive QoS even on L2 switched traffic.

Table 12 illustrates the DiffServ DSCP & PHB defined in RFC 4594 as well as their corresponding mapping to 802.1p priority levels and how these service classes are normally handled in the Extreme Networks switching platforms.

Table 12 – QoS Markings and Queuing Profiles

Class of Service (CoS) and ERS and VSP Naming	Classification			Description
	PHB	DSCP	802.1p	
CoS 7 Network/Critical	CS7, CS6	48, 56	7	Network Control – Strict Queue, 5-10% shaped
CoS 6 Premium	EF, CS5	46, 40	6	Real Time Voice – Strict Queue, 50% shaped
CoS 5 Platinum	AF4x[1], CS4	34,36,38,32	5	Real Time Video – WRR Queue
CoS 4 Gold	AF3x[1], CS3	26,28,30,24	4	Non-Real Time Streaming – WRR Queue
CoS 3 Silver	AF2x[1], CS2	18,20,22,16	3	Non-Real Time – WRR Queue
CoS 2 Bronze	AF1x[1], CS1	10,12,14,8	2	Non-Real Time – WRR Queue
CoS 1 Standard (Default)	DF, CS0	0 - 4	0	Best Effort – WRR Queue
CoS 0 Custom	n/a	n/a	1	Scavenger – Low Priority Queue

[1] where x can take value 1,2,3 depending on the Drop Precedence

It should be noted that not all QoS classes need be used. The DiffServ model is a template that defines as many classes of traffic with unique QoS requirements as possible and the Extreme Networks VSP, ERS, and ExtremeXOS series QoS product implementation allows up to an eight-class model. In this model, the Network/Critical class is always present and reserved exclusively for network control protocols (for instance, IS-IS with Fabric Connect), and in the absence of any QoS policies all application and user traffic will use the default Best Effort queue.

The network design should take into consideration the applications, traffic types, and virtual networks to be transported as well as how critical each one of those is for the business. Based on that analysis, the QoS policy will define the class to use for each virtual network or application within.

Tip

All of Extreme Networks VSP series of Core and Distribution platforms are pre-configured for eight QoS classes.

Extreme Networks ERS series access platforms can be configured to use QoS queue-sets with one to eight QoS classes (the default queue-set uses two queues)

ExtremeXOS platforms can support up to eight Egress QoS Profiles (QP1-8) of which only QP1 & QP8 are configured by default.

QoS Implementation Over SPB

In an SPB backbone, all packet forwarding is performed at Layer 2 using the BMAC outer Ethernet header of the Mac-in-Mac encapsulation. The Mac-in-Mac header always carries in the B-TAG a Backbone Priority Code Point (PCP; i.e., p-bits in the Backbone VLAN Q-tag) and a Drop Eligible Indicator (DEI) bit. These are the QoS markings that will be used to derive the PHB by every transport (BCB) node encountered along the SPB shortest path as well as the terminating egress BEB node.

The 802.1ah Mac-in-Mac encapsulation Service instance I-TAG can also carry a Priority Code Point (I-PCP) 3-bit field and Drop Eligible Indicator (I-DEI) 1-bit (the I-TAG field also carries the I-SID information), which can provide an extra mechanism to tunnel QoS marking across the SPB fabric, when those markings have been removed or are missing in the encapsulated payload.

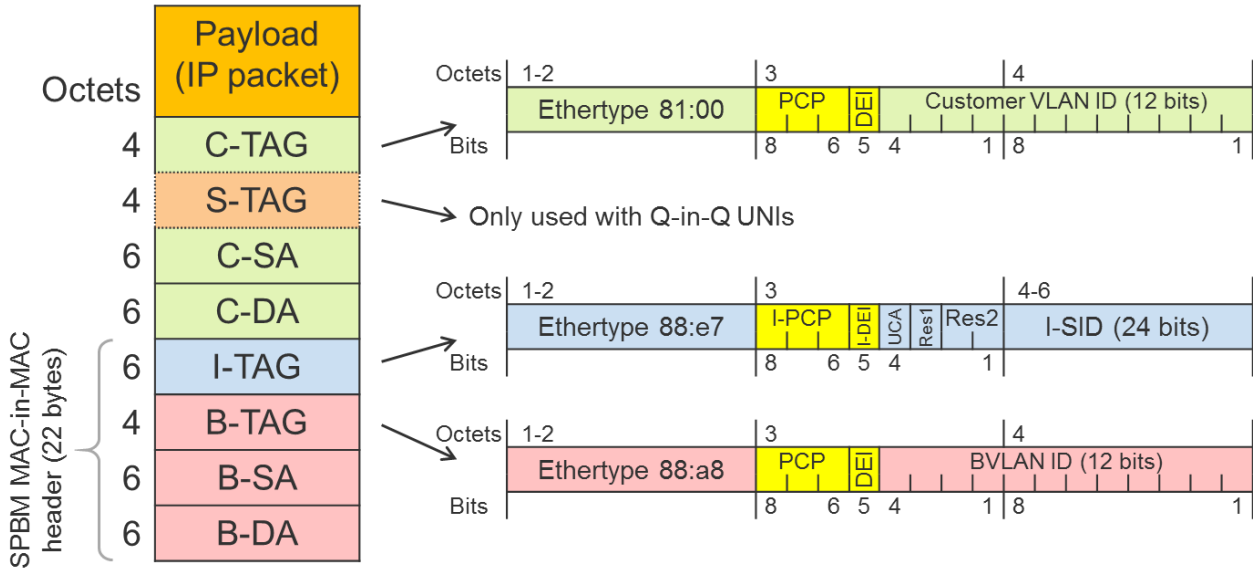


Figure 76 QoS Fields in an SPBM Mac-in-Mac Frame

Note

The use of I-PCP and I-DEI bits is defined in IEEE802.1ah (section 25.4)[8] and is needed when interconnecting Q-in-Q networks over an SPBM backbone where the S-TAG may be stripped before applying the Mac-in-Mac encapsulation and thus the I-PCP bits are needed to relay the S-TAG PCP and DEI bits.

Tip

An MPLS label carries the QoS marking in the Experimental bits field, which is also a 3-bit field like the Ethernet PCP fields. There is no equivalent of the DEI bit in MPLS.

8 See Reference Documentation [4].

Caution

Extreme Networks ERS and VSP series platforms do not currently make use of the DEI bit.

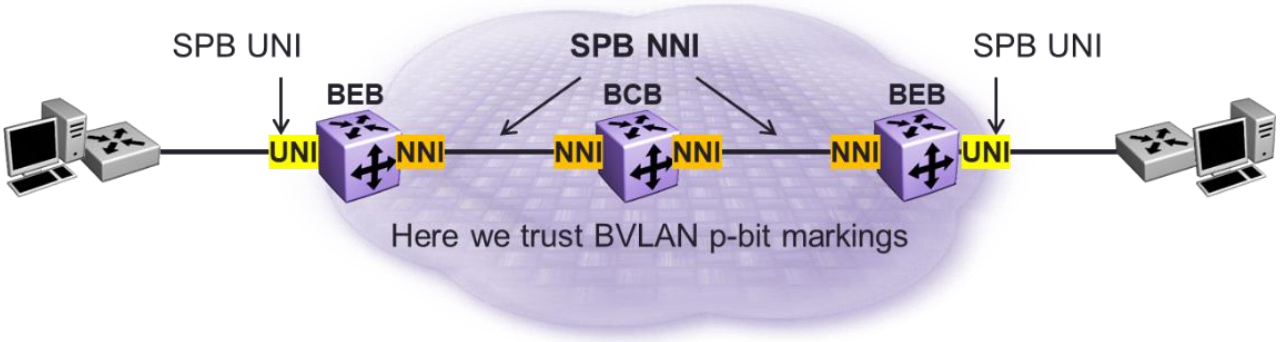


Figure 77 QoS SPB Model

Because SPB uses a transport encapsulation with its own QoS markings, there are two possible approaches to combining the DiffServ model with an SPB backbone; we shall refer to these as the Provider and Uniform models.

In the Provider model, it is possible to "tunnel DiffServ" across the SPB backbone over a VSN service without altering those DSCP markings (or I-TAG I-PCP) or deriving any PHB from them. Instead the SPB backbone will apply a QoS policy on the UNI ports for that VSN service so as to derive a PHB using the SPB QoS markings which will reflect the Service Level Agreement (SLA) for that VSN service provided to that end-customer. This model is illustrated in Figure 78 and is only applicable when there is an end-customer network that does not share the same DiffServ domain as the SPB backbone.

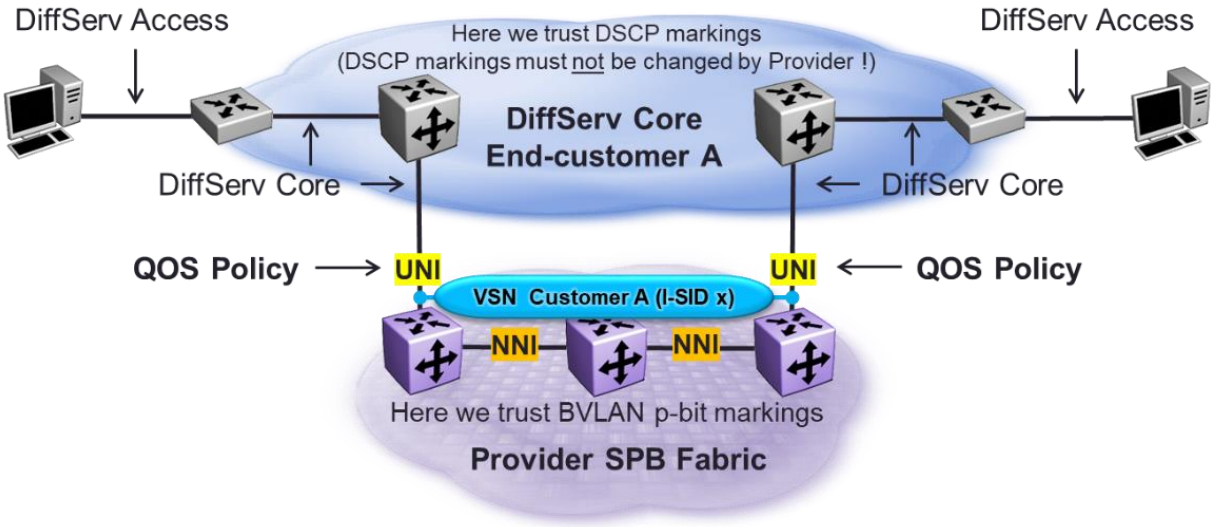


Figure 78 SPB QoS Provider Model

Note

The Provider model is equivalent to the Pipe Model described for MPLS in RFC 3270.

In the Uniform model, the VSN services and the SPB backbone share the same DiffServ domain and thus the DSCP markings are simply converted to Backbone p-bit markings for transport over the SPB fabric. The ingress BEB node will therefore derive the PHB from the IP DSCP markings (or Ethernet VLAN Tag p-bits markings, for an L2 VSN service) for traffic received on UNI ports and when it applies the Mac-in-Mac encapsulation on the egress NNI port will automatically set the Backbone p-bit to the corresponding QoS

class marking, so that subsequent transport (BCB) hops along the SPB path will be able to derive the correct PHB by only looking at those Backbone p-bits. This will include the egress BEB node, which will remove the Mac-in-Mac encapsulation, and at which point the DSCP markings will continue to be relevant to any subsequent IP routing hops (either external or internal to the SPB Fabric) encountered.

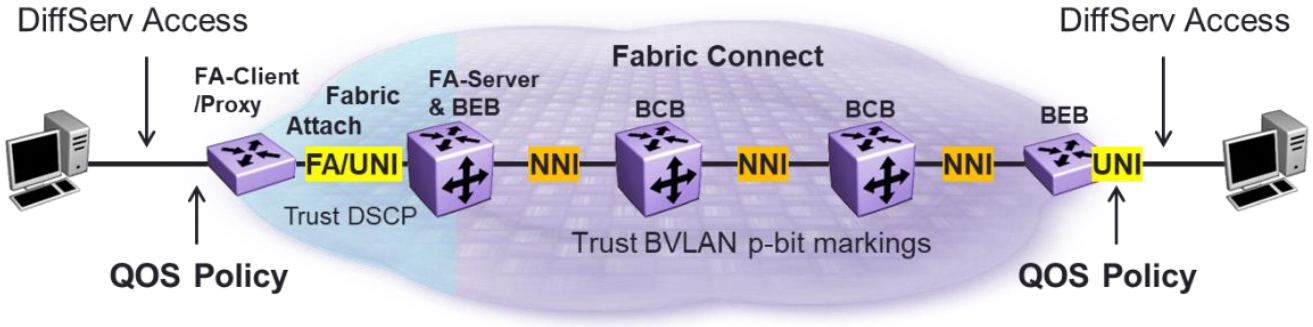


Figure 79 SPB QoS Uniform Model

Note

The Uniform model takes its name from the same model described for MPLS in RFC 3270.

Both models can be combined on the same SPB backbone for different VSN services. However, in an Enterprise environment, the Uniform model is the most applicable and is therefore the model we will look at in more detail.

In the Extreme Networks Fabric Connect solution, the SPB fabric can also be extended to non-Mac-in-Mac capable devices using Fabric Attach, as illustrated in Figure 79. With Fabric Attach, both the FA-Client/Proxy and FA-Server will automatically Trust DSCP (and native VLAN tag p-bits) QoS markings on their FA/UNI interconnect and the FA-Server will automatically map those markings into the corresponding Backbone p-bit marking for the PHB class. There is, in fact, nothing to configure to make this happen and the only QoS configuration required will be on the access ports where users connect (typically located on an FA Proxy switch access ports) or where servers connect in the data center (typically on an SPB DVR BEB node) or where firewalls connect (either in the DMZ or in the data center). The QoS policy will determine whether the traffic received from these end devices can be trusted (in which case the access ports are configured as Trusted ports) or not (in which case the access ports are configured as Untrusted ports with access policies to selectively assign the appropriate QoS service class based on ACL filter hits).

In the case of an ExtremeXOS access layer, advanced policies can be used to automatically assign the correct QoS markings to traffic profiles defined in Extreme Management Center and either statically assigned to ports or assigned via LLDP with CEP (Convergence End-Point Detection), or via ExtremeCloud with dynamic assignment.

Tip

Benefits of SPB's QoS model over MPLS QoS based on the Experimental QoS bits:

With MPLS QoS there is a lot of complexity, and additional configuration required on the egress PE to accommodate the use of Penultimate Hop Popping (PHP), whereby the penultimate hop in the MPLS backbone removes the MPLS labels from the packets before these reach the last LSP hop (i.e., the PE node.) This is done to avoid packet recycling (double lookup) on the egress PE node, which would otherwise have a forwarding performance impact on the PE. The problem is that the MPLS QoS Experimental bits are contained in the label and are thus thrown away by PHP. Much of RFC 3270 discusses how to resolve this problem as this applies to both the Pipe and Uniform models. With SPB, there is no such inconvenience as PHP.

QoS Considerations with Fabric Extend

When using Fabric Extend with IP (VXLAN) encapsulation, we are yet again adding an additional packet header to transparently cross a WAN provider cloud. In terms of QoS, every time a new encapsulation is added, there is a need to reflect the QoS of the packet payload in that new header because that new header will be used by downstream routers to forward the packet and we need to ensure a consistent PHB.

Note

The Fabric Extend L2 mode does not add any additional encapsulation. Instead it performs VLAN tag translation on the existing Mac-in-Mac encapsulation. Therefore, the Fabric Extend L2 mode is not applicable to this discussion.

In the Extreme Fabric Extend IP (VXLAN), implementation the p-bits in the Backbone VLAN Q-tag are automatically re-mapped to IP DSCP values with a consistent PHB.

Note

Fabric Extend QoS mappings are as follows:

p-bit 0 ➔ dscp 0, p-bit 1 ➔ dscp 0, p-bit 2 ➔ dscp 10, p-bit 3 ➔ dscp 18,

p-bit 4 ➔ dscp 26, p-bit 5 ➔ dscp 34, p-bit 6 ➔ dscp 46, p-bit 7 ➔ dscp 46

Caution

Currently Fabric Extend IP (VXLAN) QoS mappings are static and cannot be changed.

It is worth noting that the WAN provider may or may not even look at the IP DSCP marking for the traffic it receives. This will depend on the Service Level Agreement (SLA) in place and the capabilities of the WAN provider equipment. The WAN provider will typically use an MPLS backbone to deliver WAN services to their customers of which they will have many. All their customers will be transported over their same backbone but not all customers will have the same SLA. The WAN provider will most likely allocate each of their customers into one of their own QoS levels, but this classification will be independent of any DSCP marking present in the packets sent to them.

Where the customer DSCP marking may come to play is if the WAN provider is able to use more complex hierarchical QoS queueing in their equipment, which would allow them to maintain two levels of QoS granularity across their equipment. In practice, this complexity is often avoided and it is best practice for the customer to egress shape their connection to the WAN provider to match whatever bandwidth has been purchased for the service. Any traffic in excess of purchased WAN bandwidth can thus be queued in the customer's Fabric Extend egress logical IS-IS Ethernet port where any drops will be QoS aware based on the customer's own application marking.

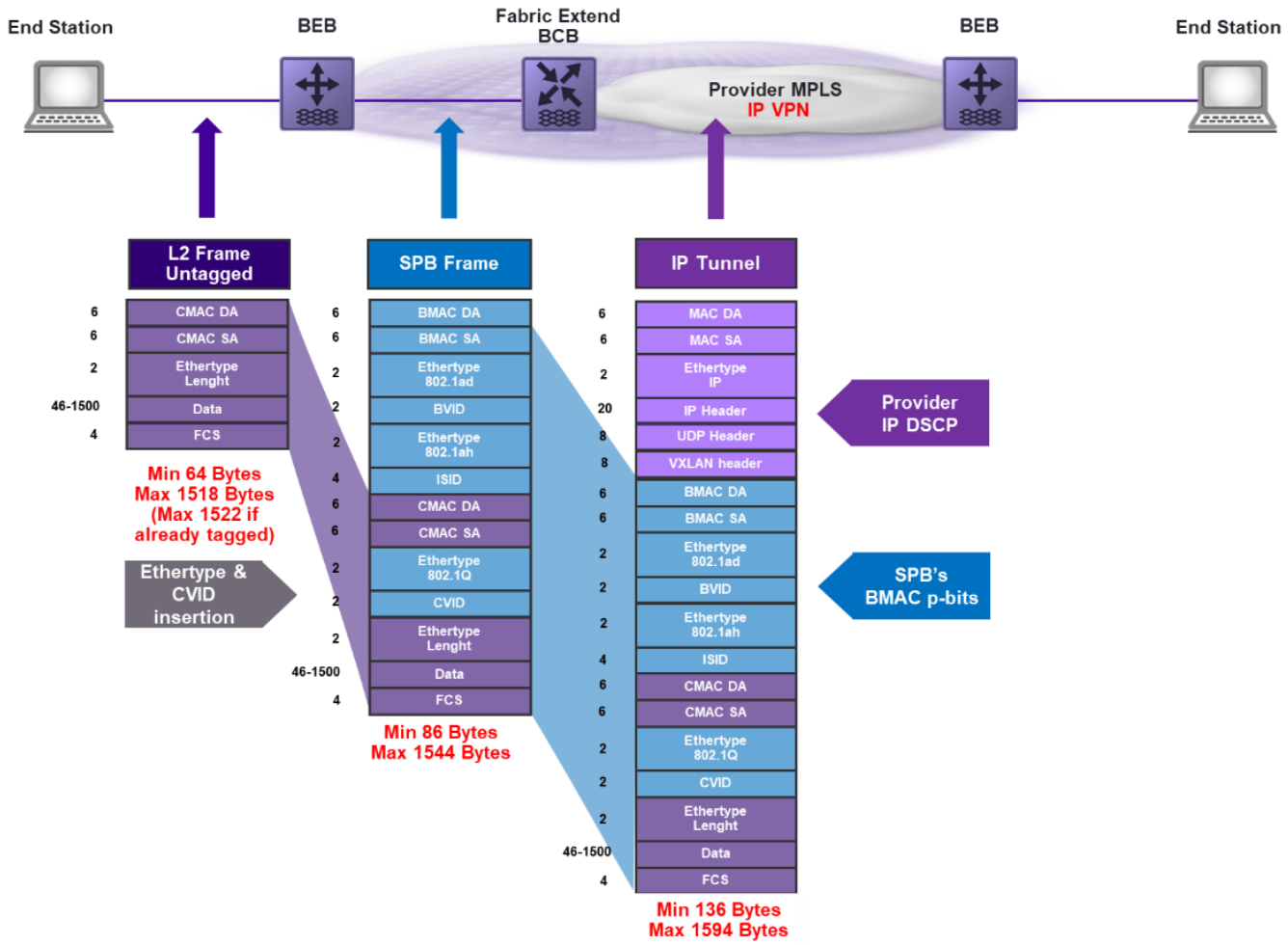


Figure 80 QoS Marking Over Fabric Extend

The Fabric Extend DSCP QoS markings become more relevant when Fabric Extend is not used to extend Fabric Connect over a WAN provider, but instead is used to extend Fabric Connect over an enterprise's self-owned campus IP core. This approach can be used as a migration strategy to fabric-enable the access of the network before SPB-enabling the core of the network. In this scenario, the DSCP markings become important in ensuring that the IP underlay network is able to correctly derive QoS for the Fabric Extend traffic.

In either case, the IP DSCP marking of the Fabric Extend traffic should be preserved by the WAN provider or campus IP underlay. This marking will be used when the traffic is received at the other side of the WAN connection by the receiving Fabric Extend ingress logical IS-IS Ethernet port to derive QoS classification for the VXLAN de-encapsulation process but the DSCP marking are not used to modify the SPB BMAC p-bits within. Once the VXLAN encapsulation has been removed, QoS will again be inferred from the SPB BMAC p-bits.

Consolidated Design Overview

This section will illustrate how the various VSN service types should come together into the reference architecture outlined in the Guiding Principles section on page 27.

Campus Distribution

From a logical point of view, it makes sense to define the L3 VSN networks such that the routing instances are located in VRFs on the nearest distribution nodes to where the end-stations are located, and from here user VLANs / L2 VSN segments are extended to reach out to those users, located either in the wiring closet or in data center ToR switches.

Tip

With SPB, we have the flexibility to perform IP routing for a VSN on any IP/VRF capable node located anywhere in the Ethernet fabric and likewise to stretch a user/server VLAN anywhere across the fabric. While this flexibility is an obvious boon for handling exceptions and adapting to unexpected requirements (as previous technologies would not easily allow it), at the same time it is important to maintain an overall clean and elegant VSN architecture and not get into bad habits that could ultimately render the overall network design hard to comprehend.

Figure 81 illustrates how L3 VSNs are terminated into VRFs on a distribution BEB. The same BEB then associates IP interfaces to VLANs onto a specific VRF and then extends those VLANs out of UNI ports to reach the end-stations located in Layer 2 wiring closet switches or other device directly connected to the BEB node (e.g., a firewall).

Tip

With Fabric Attach, the wiring closet access switch can automatically attach a user to the desired VLAN/I-SID via RADIUS based authentication without any need to manually provision and manage what VLAN ids need to be tagged on the BEB UNI ports acting as uplinks for the wiring closet switches.

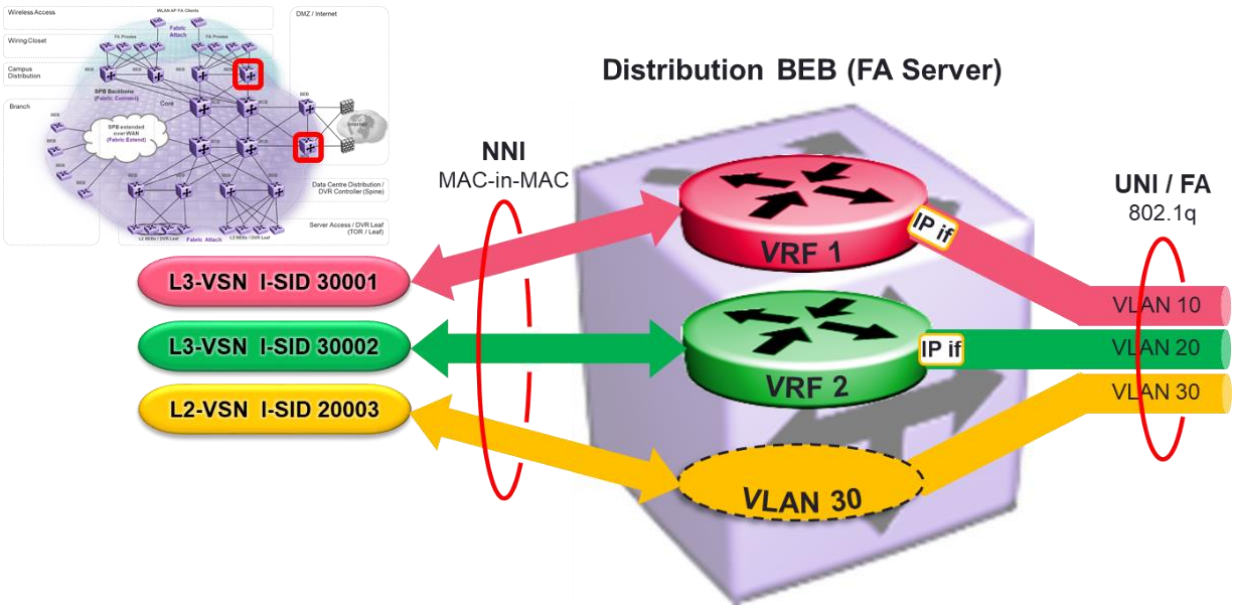


Figure 81 Zoom on Distribution BEB with UNI/FA Interfaces

The same picture illustrates how at the same time some other fabric-wide L2 VSN segments (e.g., for Guest WLAN users where the default gateway is implemented by a captive portal in the data center and thus no IP interface is needed in the SPB fabric) can also be extended to the same wiring closet switches in the same exact manner. Note that in this case there need not be a platform VLAN defined on the distribution BEB node as Fabric Attach uses Switched UNI functionality, which can directly map the L2 connection into the desired fabric L2 VSN without any additional provisioning required on the BEB FA Server.

Note

For a Fabric Attach VLAN:I-SID binding, the only reason to have a platform VLAN defined on the FA Server BEB is if either an IP interface is required on the BEB to act as default gateway for the segment and/or if there is a need to activate SPB Multicast on that segment.

Where the distribution layer nodes connect to Extreme Networks Fabric Attach access switches, the BEBs can be aggregated into an SMLT cluster (Multi-chassis Link Aggregation Group – MLAG) in such a way that the Layer 2 access switches can be connected with simple MLT (or LACP or EtherChannel) link aggregation. This allows user VLANs to be redundantly q-tagged to any number of access switches (on as many SMLT links) with all uplinks actively used for traffic forwarding and no Spanning Tree required. This is illustrated in Figure 82.

Clearly any VSN service terminated on one of the SMLT BEBs will also need terminating on the IST (or Virtual-IST) peer BEB as well. Hence, for an L3 VSN, both SMLT BEBs will need to have a VRF configured and the same user VLANs associated to that VRF. For gateway redundancy within those VRF VLANs, either standards VRRP (with Extreme Networks VRRP-Backup-Master extensions) or Extreme's RSMLT-Edge functionality or even DVR can be leveraged. If the SMLT attached device is Fabric Attach capable there is no need to manage what VLANs are assigned to which SMLT links as the FA Client-Proxy device will automatically provision these based on their end-point provisioning or via RADIUS authentication of the connected devices.

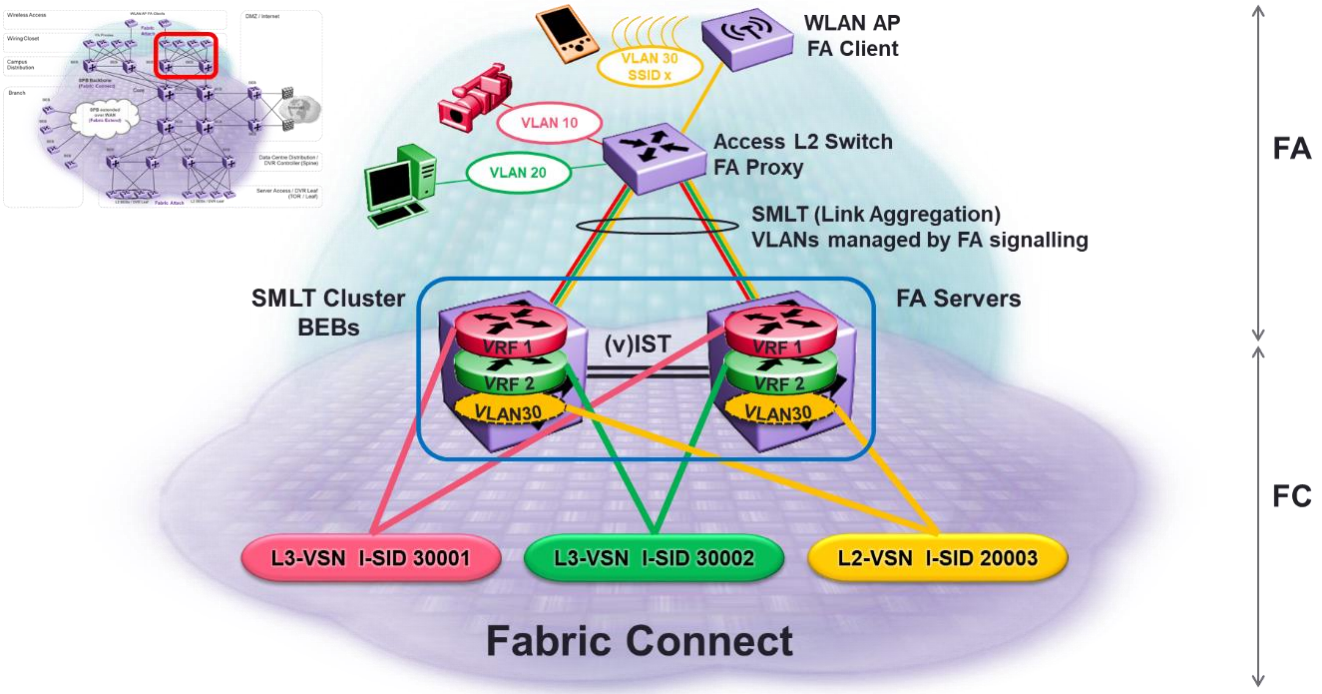


Figure 82 Distribution BEB with SMLT Clustering

Data Center Distribution

The data center distribution model is slightly different as here there is a real benefit in extending Fabric Connect all the way to the ToR switches. This allows to high speed interconnect the ToRs in smaller data center designs as well as the use of DVR, which in all cases ensures shortest path and lowest latency for all data center east-west and north-south traffic flows.

In a DVR design the data center distribution will be performing the role of DVR controllers and as such all L3 VSNs for the data center will be provisioned on these nodes as well as the corresponding VRF instances and server L2 VSN segments.

Tip

If the data center topology is spine-leaf, then either the spines become DVR controllers or a pair of border leaf nodes are made the DVR controllers. Either way, the DVR controllers represent the point of entry and exit for data center traffic.

As illustrated in Figure 83, these Distribution nodes can thus use inter-VSN routing capability whereby traffic can be IP routed on or off a server L2 VSN segment as well as providing IP routing between these server L2 VSN segments.

Tip

In a DVR architecture, the DVR controller is not likely to do much inter-VSN routing as the DVR leaf nodes would have already IP routed the flow directly at the ToR access layer.

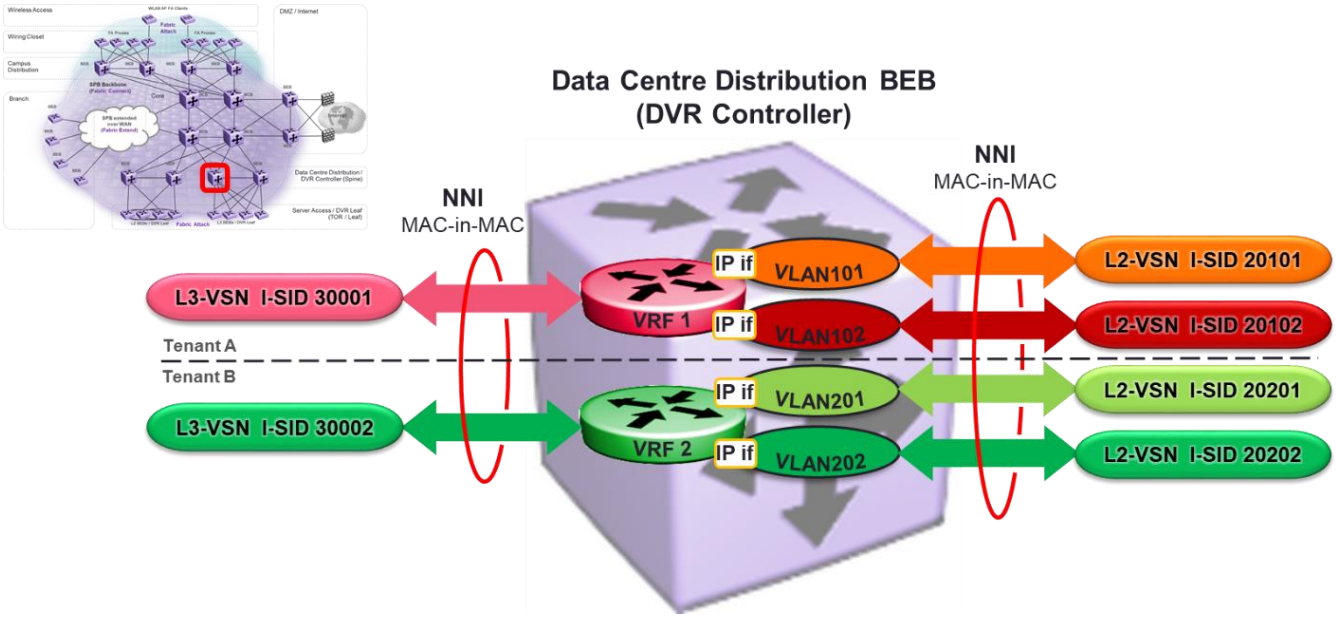


Figure 83 Zoom on Data Center Distribution BEB / DVR Controller

Data Center Access

To complete the picture, we will also zoom into the data center access ToR switches, which will be acting as DVR leaf nodes. These switches are part of the SPB fabric and will typically be deployed as SMLT cluster pairs so that servers can be dual-homed with active-active LACP LAG SMLT links where necessary. As illustrated in Figure 84, the DVR leaf ToR switches will have much the same awareness of VRFs and L3VSNs as well as all the available server L2 VSNs within the DVR domain since this is automatically pushed to them from the DVR controllers. The ToR switches will also present a distributed anycast gateway for the attached servers and will perform the first IP routing hop for any non-L2 data center flow.

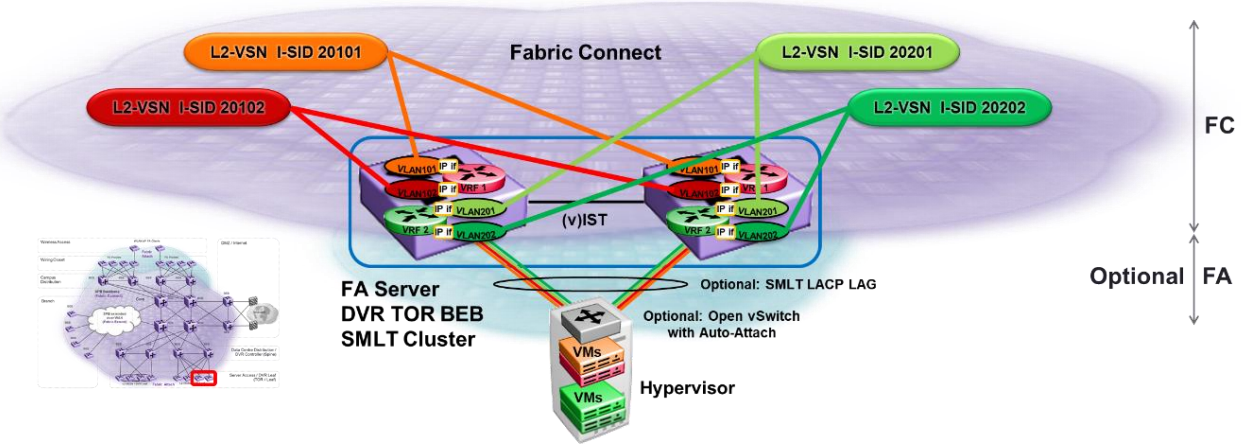


Figure 84 Data Center Top of Rack (ToR) BEB with SMLT clustering

From a configuration perspective, the DVR leaf ToR switch remains purely an L2 switch and can only be provisioned with which L2 VSN I-SID to attach to a given server access port (even this configuration can be automated if the server hypervisor is capable of Fabric Attach via use of OVS).

High Availability

High Availability (HA) is the function of eliminating single points of failure and detection of failures as they occur.

System Level Resiliency

System level resiliency includes having nodes with redundant hardware components and enhanced software features to detect system and hardware failures by providing recovery mechanisms.

Hardware Component Redundancy

Some of the hardware components for which redundancy is required follows:

- Power Supplies (PSUs)
- Switch Fabric Modules
- Cooling Fans
- Control and Management Modules (CPUs)

A system comprising a chassis-based solution will maximize the levels of hardware component redundancy across all components, including the Control and Management Modules.

Lower form factor platforms (non-chassis based) will also provide Power Supply and Fan Cooling modules though will typically provide no redundancy in terms of Control Modules and Switching Fabric.

In a properly designed SPB Fabric, engineered with an adequate number of BCB nodes and NNI core links, deploying two or more lower form factor platforms instead of one large chassis platform can become a viable and more cost-effective alternative. This is because these platforms become a less critical component within the overall SPB Fabric.

Tip

Extreme Networks offers a compact form factor chassis as well as lower format modular platforms to fit every Ethernet fabric design.

High-Availability Mode

High-Availability (HA) mode is available on Extreme Networks VSP 8600 chassis based platform and ensures that in dual CPU configuration the Standby CPU is kept in sync with the Master CPU, in such a way as to immediately fail over should the latter fail. HA mode failover ensures that the control protocols used in the network do not need to restart or re-converge which in turns ensures no impact to traffic forwarding.

Caution

Protocols not supported in HA mode will operate in Warm-Standby mode, meaning that they will restart if a CPU switchover occurs.

Network Level Resiliency

Fabric Connect Fast-Rerouting

Fabric Connect is capable of providing fast traffic recovery upon a link or node failure (typically around 200 ms) for all traffic and VSN service types, including unicast and multicast. This property is virtually impossible to achieve in a traditional architecture where different layers of the protocol stack (e.g, PIM-SM, or BGP, or LDP, etc.) are dependent on other protocols in the stack beneath them.

Clearly for Fabric Connect fast-rerouting to be possible, the SPB core needs to be architected with sufficient redundant paths and switching capacity such that loss of a component (link or node) does not impact connectivity and/or switching capacity after a failure. If this premise holds, then the ability of Fabric Connect to heal around failed links or nodes becomes an operational asset whereby any given SPB node can be taken out of operation from the fabric for maintenance work or software updates without impacting any applications running over the network. The virtualized network now offers the same operational capability that virtual servers do, where Virtual Machines can be hot-migrated away from a physical server in order to perform maintenance on the hardware.

Tip

When taking an Fabric Connect node offline for maintenance work, this can be done gracefully by activating IS-IS overload, whereby the node will signal to the rest of the SPB Fabric that it is no longer to be considered as a transit node for path calculations by other nodes. If the node is a BEB using VLACP, this can be subsequently deactivated to ensure SMLT/Fabric Attach links are also gracefully deactivated before the node is taken fully offline.

Virtual LACP

VLACP is an extension of LACP (Link Aggregation Protocol) used exclusively for link integrity and neighbor control plane keep alive detection.

This is accomplished by each switch transmitting VLACP PDUs on switch-switch links at a set timer interval in order for a link to maintain a 'link-up' state. It allows the switch to detect unidirectional or bidirectional link failures irrespective of intermediary devices as well as an indication of the control plane liveliness of the adjacent node, which gigabit Ethernet auto-negotiation Far End Fault Detection (FEFI) and 10 gigabit Ethernet Remote Fault Indication (RFI) cannot detect.

Note

LACP offers the same link integrity capability combined with the ability to dynamically aggregate interfaces into Link Aggregation Groups (LAG), i.e., Multi-Link Trunks (MLTs)

The differences between LACP and VLACP are:

- VLACP does not have any link aggregation capabilities.
- VLACP is optimized to run on significantly faster timers, as low as 200 ms.
- VLACP can use customized multicast MAC address and ethertypes, which allows it to be used beyond link local (i.e., across a Layer 2 Ethernet cloud).

VLACP is not required if LACP is already enabled on a given set of ports.

With Extreme Networks Fabric Connect, best practice design guidelines call for VLACP to be activated on all switch-switch Ethernet physical (or logical) links. Because VLACP runs on fast timers, loss of link integrity can be detected sub-second, without having to rely on IS-IS Hello timers which run at a much slower pace (every nine seconds with three retries).

VLACP thus ensures that SPB's fast-reroute capabilities are ensured under these extreme failure scenarios:

- Logical link failure. The link to the adjacent switch is transported over some L2 transport cloud (in this case physical link status is no longer correlated with logical link status).
- Software lockup or not responding on adjacent switch. This is a rare but possible failure scenario in moderns Ethernet switching platforms where the hardware is able to continue switching traffic without software/CPU intervention, for example after a software failure/seizure.

- Unidirectional link. A link that is only able to Transmit but not Receive (or vice-versa) and would thus result in traffic getting black-holed in one direction. In the case of a single fibre strand failure, Ethernet's auto-negotiation Far End Fault Detection (FEFI) and 10 gigabit Ethernet Remote Fault Indication (RFI) will be able to detect this and take the Ethernet interface down even on the side where light is still received. However, the above-mentioned Ethernet mechanisms will fail if the unidirectional link status is the result of some other chipset lockup condition inside the node, but behind the MAC chipset. VLACP will detect these.

In regard to the VLACP PDU, the Extreme Networks VSP and ERS series design guidelines specify the use of MAC address of 01:80:c2:00:00:0f for all directly connected point-to-point Ethernet links.

Tip

MAC addresses 01:80:c2:00:00:0X are IEEE reserved as link-local multicast addresses. Packets sent to these multicast MAC addresses will remain link-local and no Ethernet bridge / L2 switch is allowed to forward these packets onto other segments.

In a scenario where node1-node2-node3 are chain connected, in the event of loss of configuration of node2 (which could result in node2 simply forwarding VLACP PDUs between node1 and node3), it is desirable that VLACP on node1 and node3 interfaces isolate node2. This is possible provided that VLACP was configured using 01:80:c2:00:00:0f.

Caution

ExtremeXOS does not support VLACP. However, the similar benefits of VLACP can be obtained when the ExtremeXOS platform is deployed as a FA Proxy with LACP enabled Load Sharing (MLT) towards the FA Server.

Whereas if the link in question is logical (transported over some L2 transport cloud), the Extreme Networks VSP and ERS series design guidelines specify the use of MAC address 01:80:c2:00:11:xx, which is a normal Ethernet Multicast address which can be forwarded by intermediate Ethernet bridges.

Tip

Conceptually, VLACP does at Layer 2 what Bidirectional Failure Detection (BFD) has to do at Layer 3 in an MPLS-based architecture. VLACP allows IS-IS to react faster to "non-clean" link failure types described above, just like BFD allows BGP or OSPF or IP Static Routes to react faster than they would if they had to wait for their own timers to expire.

VLACP is very similar to Cisco's Unidirectional Link Detection Protocol (UDLD).

Link Aggregation / Multi-Link Trunking (MLT)

The ability to bundle two or more Ethernet links into one logical link is useful for two reasons. The first reason is to provide link resiliency so that loss of one physical link does not mean loss of the logical link. This is typically the main reason for using link aggregation on access links or towards servers. The other reason is to increase the available bandwidth of the overall logical link by hashing traffic across all available links forming the link aggregation. Thus, an aggregation of two 40GbE interfaces will result in an 80GbE logical link. This use is more common for switch-switch links to increase the available bandwidth in the core and is an easy way to address congestion hotspots in the network topology.

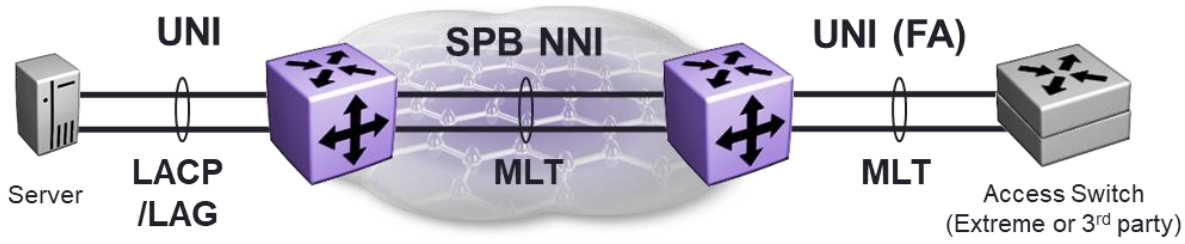


Figure 85 Multi-Link Trunking (MLT) Used in Core and Access

In the Extreme Networks VSP and ERS series terminology, this is referred to as Multi-Link Trunking (MLT) when the link aggregation is statically configured and as Link Aggregation Groups (LAG) when the aggregation is done dynamically by the LACP protocol. In the case of a static MLT, we also refer to Distributed-MLT (DMLT) when the MLT is configured either on a stacked switch comprising multiple units or a chassis, and where the DMLT Ethernet links are distributed across the different slots of the chassis. This provides additional protection against a unit failure in a stack or a line card failure in the case of a chassis.

In the ExtremeXOS platforms, the ability to perform link aggregation is referred to as load sharing.

Tip

LACP is generally recommended on access links, i.e. when doing link aggregation towards servers and/or non-Extreme Networks networking devices (e.g., firewalls), as this combines both the link aggregation and link integrity capabilities of the LACP protocol.

Tip

When doing link aggregation between Extreme Networks VSP and ERS series devices on core links, it is recommended to use static MLTs with VLACP configured on each of the underlying physical Ethernet ports. This provides a simpler configuration (than LACP), greater control of which core links are to be aggregated together, as well as benefitting of much faster VLACP timers for detecting any link faults, as described in the preceding section.

Tip

When doing link aggregation between Extreme Networks VSP FA Servers and ExtremeXOS FA Proxy access switches, it is recommended to use LACP SMLT / load sharing. This provides the similar benefits to a static MLT config with VLACP.

Tip

In the Extreme Networks Fabric Connect SPB implementation, an MLT can be used as a logical NNI, where IS-IS is configured on the MLT bundle and only sees the logical aggregate in shortest path computations.

Tip

Most Extreme Networks VSP, ERS, and ExtremeXOS series switching platforms support the ability to aggregate up to eith Ethernet ports into either an MLT or LAG, with the exception of the lower end, lower cost access switches which will only support four.

Every device doing link aggregation is responsible for implementing a hashing algorithm to distribute egress traffic across all available links forming the MLT/LAG.

Caution

All links forming the link aggregation (MLT/LAG) must have the same interface speed (with LACP this is a mandatory requirement; with static MLT, some platforms may tolerate different speeds, but it is not recommended). Otherwise this would undermine the hashing algorithm's ability to distribute load on the aggregate links (slower links will congest, while faster links will be under-utilized).

There are many hashing algorithms available and different devices will support different hashing algorithms. While some hashing algorithms are better at evenly distributing traffic across all available links than others, all hashing algorithms must ensure that all packets belonging to a given flow (e.g., a TCP connection between host IP1 TCP port1 and host IP2 TCP port2) in a given direction will always egress on the same link forming the link aggregation. This is important to eliminate the risk of the network delivering out of sequence packets for a given flow.

An IP-based hash is always superior to a MAC-based hash in its ability to evenly distribute traffic. This is because an IP flow is unique to the two end-stations, whereas a MAC flow is unique between two IP routers/firewalls or between the end-station and its default gateway. Likewise, an IP L3+L4 hash is superior to an IP hash because IP+TCP/UDP port flow is unique to an application running on the two end-stations. Equally important is the ability to properly hash broadcast and multicast traffic flows, where the latter can make use of a significant amount of bandwidth, particularly in video based applications.

Tip

All Extreme Networks VSP and ERS series switching platforms are capable of supporting an L3 IP hash as well as a Broadcast / IP Multicast hash.

Tip

Extreme Networks VSP series switches are capable of performing an IP L3+L4 hash by hashing source IP address + TCP/UDP port and destination IP address + TCP/UDP port. This provides for the finest possible spread of traffic across available interfaces within the MLT/LAG.

Tip

Extreme Networks ERS series stackable switches use an enhanced hashing algorithm that will prefer the use a local DMLT link(s) on the same unit where the traffic arrived on as well as preferring the shortest path around the stacking cables. This is done to optimize the usage of the stacking backplane.

Tip

ExtremeXOS platforms support configurable address-based load sharing offering any of L2 or L3 or L3+L4 hash.

Multi-chassis Link Aggregation

Multi-chassis Link Aggregation (MLAG) is the capability of distributing MLT or LAG links across two independent chassis or switches, thus adding a further level of resiliency to access and distribution layers as well as negating the use of Spanning Tree in a topology that would otherwise require it. In the Extreme Networks VSP series terminology, this capability is better known as Split Multi-Link Trunking (SMLT), which was invented by an earlier company incorporated into Extreme Networks and later copied by most vendors in the industry.

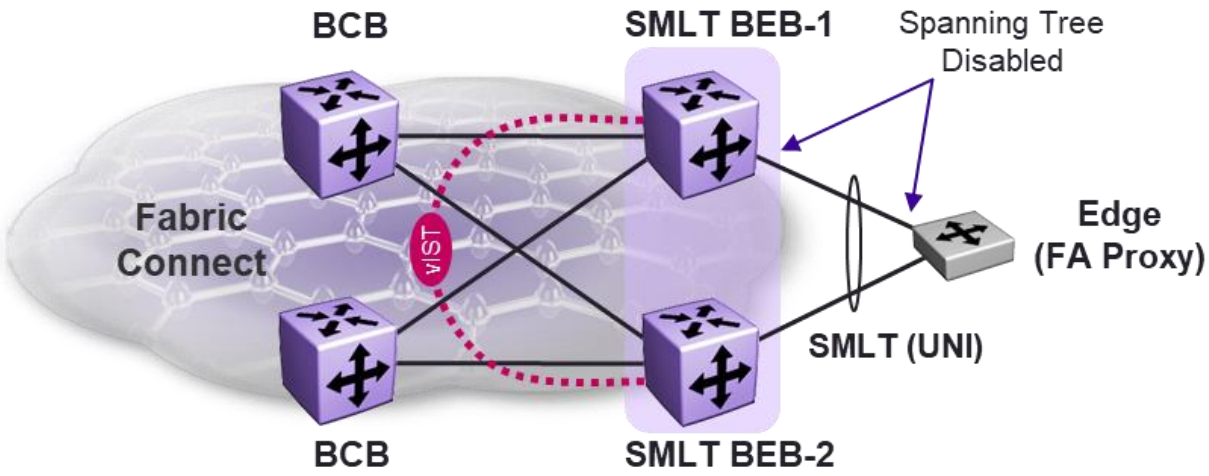


Figure 86 Split Multi-Link Trunking (SMLT) Used in SPB Fabric Access

SMLT provides the ability of having two active/active switches in a cluster. The two SMLT cluster switches appear as a single switch to the attached device such as a server, L2 switch, or L3 switch or router, which in all cases is configured with either a static MLT (or third-party static link aggregation scheme such as Cisco Ether-Channel) or LACP. The key benefit of SMLT being that L2 VLANs can be extended, loop-free, over the SMLT links, with all links actively forwarding traffic thus eliminating the need for Spanning Tree.

Tip

All Extreme Networks Fabric Connect VSP platforms support SMLT in conjunction with SPB BEB functionality.

Note

Without SMLT (MC-LAG) capability:

- Wiring closet switches would need to use Spanning Tree on their uplinks, which would condemn all uplinks but one to a discarding state.
- Dual-homed servers would have to resort to an active-standby NIC teaming arrangement.

In the Extreme Networks SMLT implementation, both switches forming the SMLT cluster operate in an active/active fashion. This is true in the data plane where both switches are able transmit and receive and forward traffic on the SMLT links they own as well as in the control plane where both switches remain independent from a protocol and management perspective.

An Inter-Switch Trunk (IST) TCP based connection is used between the cluster switches to exchange and synchronize L2 related information such as MAC addresses in the VLAN FDBs, LACP System-ID, ARP entries, and IGMP Group membership tables. Essentially, if an end-station MAC address is learned by one switch on its SMLT link, the IST will ensure that the other switch in the cluster will program the same MAC address against its corresponding SMLT link. Likewise, ARP entries and IGMP group members pointing to SMLT interfaces are synchronized across both switches. And for SMLT connections running LACP, the cluster needs to advertise the same LACP System-ID so that the attached devices will form a LAG.

In the original SMLT implementation, the IST required physical ports between the two cluster switches plus the two switches in the cluster had to be the same type of switch. The need for the IST to always be up mandated that the IST be implemented using a DMLT direct connection between the IST peers. Yet if all edge devices are dual-homed using SMLT links there is usually very little traffic using the IST DMLT. In practice, traffic usage of the IST DMLT is dependent on the number of single attached devices (or devices for which SMLT links have failed).

Tip

On earlier platforms using the original IST implementation (before vIST), the IST DMLT would be provisioned to also act as an SPB NNI MLT interface.

Taking advantage of the SPB Fabric, the above IST restrictions have been removed with the introduction of Virtual IST (vIST) on the latest Extreme Networks VSP series platforms. With a vIST, the IST is no longer tied to any physical MLT instance, but is instead associated with an L2VSN I-SID. Hence, provided that both IST peers are still connected to the same SPB Fabric, IS-IS will always be able to compute a shortest path connection for the IST connection to use. Each SMLT cluster switch now will mostly just need redundant NNI connections into the SPB Fabric.

Tip

With vIST, there is no requirement for the SMLT cluster switches to share a direct NNI connection (as shown in Figure 86). In most cases, however, it will still make sense to have a direct SPB NNI connection between the two switches, but this connection will be no different from any other SPB NNI interface and there is no longer any need for it to be of DMLT type.

Tip

Also, with vIST the SMLT cluster pair do not have to be the same switch model (though it usually makes sense for them to have the same number of interfaces).

A further enhancement of SMLT when operating with SPB is that the SMLT cluster operates with an SMLT-Virtual-BMAC which ensures that any traffic ingressing the SPB Fabric via the SMLT cluster nodes will be Mac-in-Mac encapsulated with a source BMAC, which is not the node's individual BMAC, but instead the SMLT cluster's SMLT-Virtual-BMAC. This ensures that on the distant egress BEBs, where the same traffic egresses the SPB Fabric, reverse CMAC learning will learn the source CMACs as reachable via the SMLT cluster SMLT-Virtual-BMAC and return traffic can then be load-balanced back toward both SMLT cluster nodes leveraging whatever BVLAN allocation is used across all the distant BEBs.

Note

The SMLT-Virtual-BMAC can be manually provisioned or auto-generated.

Thus, SMLT clustering not only provides active-active load balancing of traffic ingressing the Fabric Connect, but also provides multi-path load balancing within the SPB Fabric for traffic egressing the Fabric Connect on the SMLT cluster.

Active/Active IP Gateway Redundancy with SMLT

As stated in the previous section, the SMLT IST performs synchronization of L2 related tables to make both switches in the SMLT cluster appear as one switch. However, from an L3 IP perspective, both switches remain independent IP routers, each with their own IP addresses.

Note

An MLAG implementation where both nodes share the same IP interfaces is to be avoided as it requires an Active/Standby control plane (where the software on one switch controls both switches in the MLAG cluster) with all the disadvantages that entail in case of node failure. For example, loss of interconnecting communication channel between MLAG peers becomes catastrophic and results in duplicate IP conditions.

For user VLANs that need to be IP routed on the Distribution SMLT cluster, it is therefore necessary to implement some form IP Gateway Redundancy on the IP interfaces of that VLAN on each switch forming the cluster. It is also desirable that both nodes in the SMLT cluster can perform IP routing in tandem for traffic that they receive from their own SMLT links, so that the active/active traffic forwarding enjoyed for L2 flows translates into active/active forwarding for IP routed flows also.

Extreme Networks VSP series offers three deployment models to achieve this:

- Virtual Router Redundancy Protocol (VRRP) with Backup-Master extensions.
- Routed-SMLT (RSMLT) with Edge extensions.
- Distributed Virtual Routing (DVR).

Note

The VRRP and RSMLT options are supported with both IPv4 and IPv6. DVR is currently only supported with IPv4.

VRRP with Backup-Master Extensions

VRRP is the IETF standard for IP Gateway Redundancy (RFC 3768 VRRPv2 for IPv4; RFC 5798 VRRPv3 for IPv4 and IPv6) whereby two or more IP routers can share a VRRP IP Gateway interface which at any given time is active on the VRRP Router elected as Master. End user devices should be configured with the VRRP IP address as default gateway.

The role of the VRRP elected Master is to:

(i) Generate VRRP Hellos.
(ii) Reply to ARP requests for the VRRP IP by providing the corresponding VRRP MAC address.
(iii) IP route traffic received with Destination MAC address the VRRP MAC.

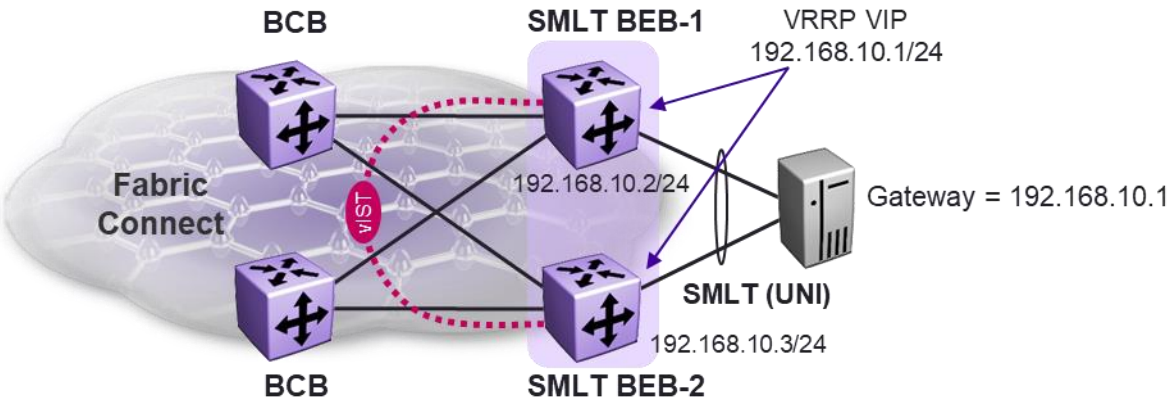


Figure 87 SMLT with VRRP Backup-Master

Extreme Networks fully complies with the VRRP standard (both for IPv4 and IPv6 and supporting both VRRPv2 and VRRPv3), but can enhance its operation on SMLT clustering by activating the Backup-Master functionality. The Backup-Master functionality should be enabled on the VRRP instance of both switches forming the SMLT cluster. One of the switches will naturally be elected as VRRP Master for the VRRP instance, as per the standard. Backup-Master has no effect on the VRRP Master switch, which will continue to perform all of (i), (ii) and (iii) listed above. Backup-Master will instead allow the VRRP Backup switch to also take “ownership” of the VRRP MAC and thus to perform (iii) in tandem.

Tip

An IP router will only IP route a packet if the packet was sent to its MAC address (or any other MAC address it has been programmed to "own"). By letting both VRRP routers "own" the VRRP MAC address, both routers are able to perform IP routing for it in an active/active fashion.

Tip

Extreme Networks supports VRRP with Backup-Master for both IPv4 (VRRPv2 or VRRPv3) and IPv6 (VRRPv3) in the VSP series of Fabric routing switches.

To use VRRP Backup-Master, the VRRP IP address must be defined as a third virtual IP address different from the IP address configured as VLAN IP.

VRRP with Backup-Master is completely independent and does not require any IST signalling to operate. It is therefore possible to have the same VRRP instance extended across two or more SMLT clusters. In which case, there will still be one single VRRP Master Router for the instance, located in one SMLT cluster, while all the other routers will all perform Backup-Master functionality. This provides an active/active solution for delivering the same default gateway IP address across multiple SMLT clusters.

Tip

Always use VRRP with Backup-Master for VLAN segments that are prone to be L2 VSN extended to more than one SMLT cluster.

Caution

Extending VRRP to multiple SMLT clusters will ensure active-active forwarding only for L2 segments where end-stations are connected via UNI or Fabric Attached interface types. If the end-stations are connected to distant L2 BEB nodes, which would typically be ToR switches in the case of a data center, then all traffic would be sent to the VRRP Master node only and DVR should be used instead of VRRP.

RSMLT with Edge Extensions

Routed-SMLT (RSMLT) was originally developed to marry an OSPF routed design over an SMLT Core design in the days when SPB Fabric core designs were not yet possible. RSMLT in that original form is not useful in an Fabric Connect design and will thus not be covered here.

RSMLT with Edge extensions does however offer a valid alternative to using VRRP with Backup-Master because it offers some advantages:

- Simpler to configure: No need to configure a third "Virtual" IP address. RSMLT-Edge makes use of the two VLAN IP addresses configured on each SMLT cluster switch.
- More efficient on CPU resources in scaled environments: RSMLT leverages IST signalling and hence there is no need for a VRRP Hello protocol (which can impact CPU load in scaled environments where hundreds of IP interfaces are configured with VRRP on the same device).
- Can scale beyond the maximum number of supported VRRP instances. Every Extreme Networks VSP series switch is quoted as supporting up to a maximum number of VRRP instances (typically 255).

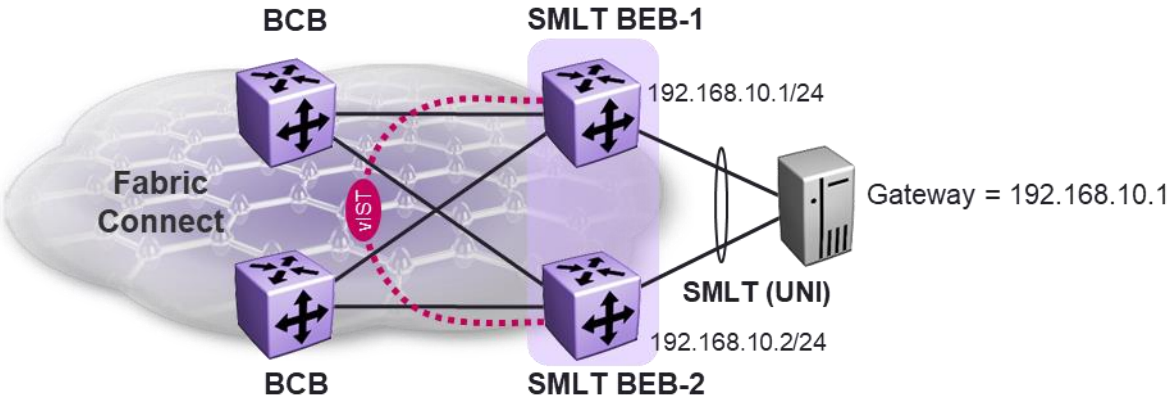


Figure 88 SMLT with RSMLT-Edge

With RSMLT, each switch in the SMLT cluster possesses its own IP address on a given VLAN. End user devices should be configured with one or the other (does not matter which one) IP address as default gateway.

The VLAN IP interface of the first switch is tied to a VLAN allocated MAC address derived from the chassis MAC of that switch. Likewise, on the second switch in the cluster.

RSMLT uses IST signalling to make the first switch take “ownership” of the corresponding VLAN MAC address of the second switch, and vice-versa. Therefore, unlike VRRP Backup-Master, RSMLT can only operate on the two routers forming the SMLT cluster.

Tip

An IP router will only IP route a packet if the packet was sent to its MAC address (or any other MAC address it has been programmed to “own”). By letting both routers “own” eachother’s VLAN MAC address, both routers are able to perform IP routing for eachother’s IP in an active/active fashion.

With RSMLT-Edge each switch is also capable of responding to ARP (and ICMP in case of peer failure) requests for its IST peer’s IP address and the peered VLAN MAC addresses are automatically saved to the config file, thus ensuring that the scheme will keep working even if both nodes are powered off and only one of them is then restarted.

Tip

Extreme Networks supports RSMLT-Edge for both IPv4 and IPv6 on the VSP series. For IPv6, RSMLT-Edge is currently only supported with GRT IPv6 interfaces not VRF IPv6 interfaces.

Tip

Prefer RSMLT-Edge over VRRP Backup-Master when the user VLAN will always remain localized to a single SMLT cluster (i.e., it will never be L2 VSN extended to other SMLT clusters).

DVR On Campus SMLT Distribution

DVR is mostly geared around providing anycast default gateway functionality and host-based IP routing within the Fabric Connect data center. It can however be used as an alternative to either VRRP or RSMLT-Edge gateway redundancy.

Configuration-wise it is similar to VRRP in that both a local IP and a virtual DVR Gateway IP need to be configured (the latter to be used as default gateway by end-stations). Like VRRP, DVR does not have any SMLT IST dependencies and thus can also be used on user segments which span multiple SMLT clusters. Like RSMLT, DVR does not use or need a Hello protocol and can therefore scale to the maximum number of IP interfaces that the platform supports.

However, DVR works with a concept of DVR domains and within and across these domains DVR announces IP host routes using the DVR domain and Backbone reserved signalling I-SIDs. All of this functionality could be useful in campus environments where the users are mobile (e.g., high density of wireless users), but is of limited use otherwise.

The other important thing to remember is that DVR-enabled L2 VSN/VLANs can only be extended to other DVR-enabled (controller or leaf) nodes or to FA Proxy switches which are themselves connected into DVR-enabled (controller or leaf) nodes.

Caution

A DVR-enabled L2 VSN cannot be extended to a non-DVR BEB. Doing so would result in no IP routed connectivity for hosts located behind the non-DVR BEB.

Load Sharing Over Fabric Connect VSNs

The ability of SPB to load balance traffic along equal cost shortest paths is an important property that goes hand in hand with its fast rerouting capability to ensure that any link or node failure in the core will not just heal very rapidly but will only impact a fraction of transit traffic traversing the core.

The SPB Fabric is able to store as many equal cost shortest paths between a source and destination node as it has BVLANs available. In the Extreme Networks Fabric Connect implementation, currently only two BVLANs are supported and this section is going to explore how effective an implementation using two BVLANs really is and how it translates for the different VSN types as well as for IP Multicast running within those VSNs.

We shall start by examining how IP routed flows can leverage SPB's equal cost shortest paths with IP ECMP. The example in Figure 89 represents an SPB Fabric where we have four paths between BEB-1 (or BEB-2) and BEB-3, which are all of equal cost. Fabric Connect is using two BVLANs, so the fabric itself will only cater for two shortest paths between a pair of two nodes. This example is equally applicable to L3 VSNs and IP Shortcuts.

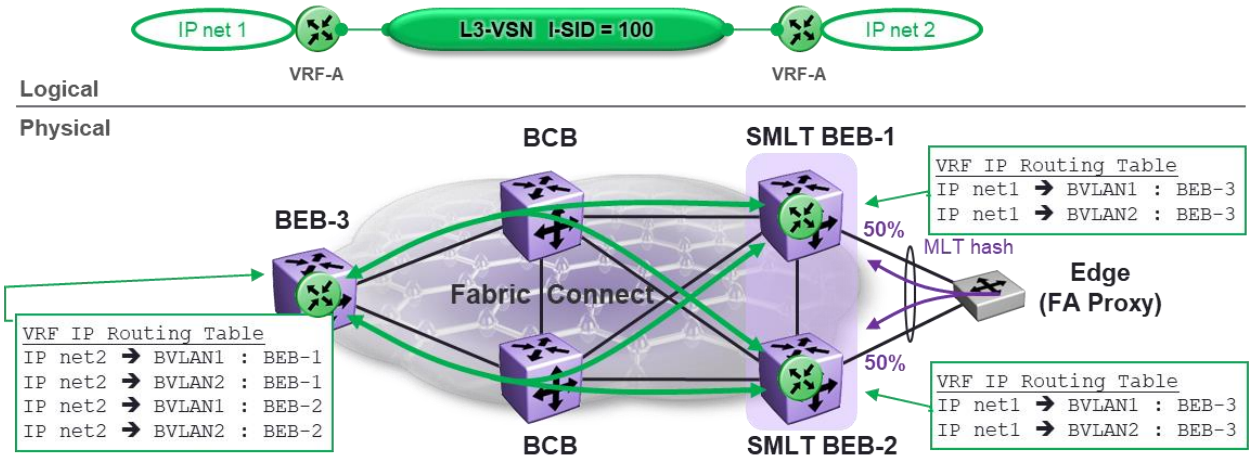


Figure 89 L3 ECMP Translation into SPB Equal Cost Shortest Paths

Both BEB-1 and BEB-2, acting as an SMLT cluster (with IP gateway redundancy, VRRP Backup-Master, or RSMLT-Edge), will announce into IS-IS their point of presence for the same IP net 2. BEB-3 has IP ECMP

enabled, and will thus learn from IS-IS the availability of IP net 2 via both BEB-1 and BEB-2. As the SPB Fabric is operating with two BVLANs the two IP paths multiply by the number of BVLANs and become 4 distinct IP ECMP paths which will create as many entries in the VRF (or GRT) IP routing table.

Tip

Use of SMLT on the edge BEBs allows equal cost paths provided by the SPB BVLANs to be multiplied by a factor of two.

In the reverse direction, BEB-3 is the only node to IS-IS advertise availability of IP net 1. Therefore, both BEB-1 and BEB-2 will only be able to install two distinct IP ECMP routes towards BEB-3; one for each BVLAN. The example illustrates how the combination of SPB equal cost shortest paths with SMLT clustering at the edge and the activation of IP ECMP results in an optimal distribution of traffic flows across all available core links.

Tip

In the Extreme Networks Fabric Connect implementation of CFM, L2 ping and traceroute performed against a target IP subnet, will automatically cater for all used IP ECMP paths.

Let us now consider how L2 VSN flows can be made to leverage SPB's equal cost shortest paths. This is illustrated in Figure 90 where again the four paths between BEB-1 (or BEB-2) and BEB-3 are all of equal cost. The normalized way to achieve L2 VSN load sharing with SPB is to distribute L2 VSN I-SIDs across the available BVLANs. The assignment could be statically provisioned but is always dynamically assigned in Extreme Networks Fabric Connect. A given I-SID could also be assigned to different BVLANs on different BEB nodes (SPB would still be able to handle this and build the necessary forwarding paths in the necessary BVLANs). In the example at hand, BEB-3 is using BVLAN1 for I-SID 2001 and BVLAN2 for I-SID 2002. The resulting load balancing spread is therefore only possible if that BEB has many L2 VSN I-SIDs and these are equally distributed across the two available BVLANs. It follows that all traffic for L2 VSN I-SID 2001 entering BEB-3 will be switched across the SPB Fabric using BVLAN1, whereas all traffic for I-SID 2002 will be switched using BVLAN2.

Note

On Extreme Networks Fabric Connect SPB platforms, a standalone BEB (not part of an SMLT cluster) will dynamically assign odd-numbered L2 VSN I-SIDs to BVLAN1 and even-numbered I-SIDs to BVLAN2. There is currently no override to this behavior.

Caution

In this scenario, it is important to make sure that the I-SID numbering scheme gives a good distribution of even and odd I-SIDs. For example, avoid numbering schemes where all I-SIDs have values 10, 20, 30, etc., since these are all even numbers.

In practice, deployment of an Extreme Networks Fabric Connect SPB architecture will leverage SMLT clustering on the BEB nodes to obtain a better load balancing behavior. In the case of BEB-1 and BEB-2, which are part of an SMLT cluster, it is no longer the L2 VSN I-SID that gets assigned to the BVLANs, but rather the BEB nodes themselves. Hence BEB-1 will always transmit flows into the fabric using BVLAN1 for all L2VSNs it terminates while BEB-2 will always make use of BVLAN2 for the same L2VSNs.

Note

The fabric SMLT IST BEB peers will automatically take on one of two possible roles:

- Primary split-BEB: Node with the lowest IS-IS System ID; will always transmit traffic into BVLAN#1.

- Secondary split-BEB: Node with the highest IS-IS System ID; will always transmit traffic into BVLAN#2.

Tip

On the receive side, both nodes forming the SMLT cluster are able to receive from either BVLAN. In fact, all traffic received from an SMLT interface is Mac-in-Mac encapsulated using a source SMLT-Virtual-BMAC which is jointly "owned" by both nodes forming the SMLT cluster. The SMLT-Virtual-MAC will thus be reverse-learned by remote BEBs as they de-capsulate Mac-in-Mac traffic and will thereafter be used by them to forward unicast traffic across the fabric directly to the SMLT cluster. Under normal operation BEB-1 will "own" the SMLT-Virtual-MAC on BVLAN1 while BEB-2 will own it on BVLAN2. In case of SMLT node failure, the remaining node will own the SMLT-Virtual-MAC on both BVLANs.

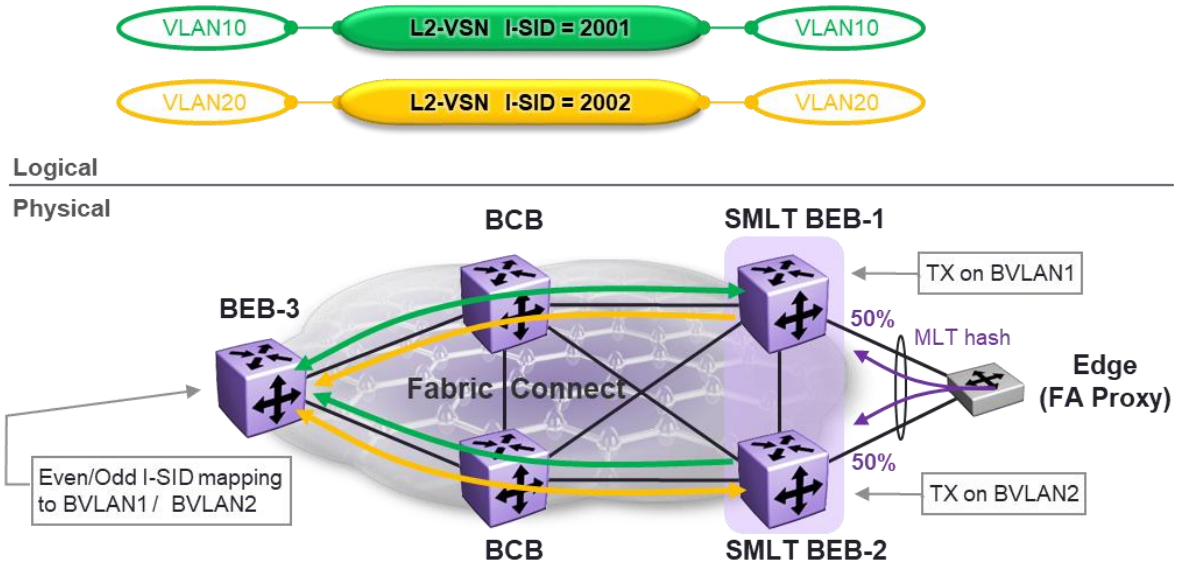


Figure 90 L2 VSN Load Balancing into SPB Equal Cost Shortest Paths

The MLT hash function of the SMLT attached devices (which could be servers in the data center or wiring closet access switches with Fabric Attach) will automatically hash flows to both SMLT BEB nodes providing a good 50% spread, which is thus directly translated into a spread of the same traffic into both BVLAN1 and BVLAN2.

Tip

Use of SMLT on the edge BEBs allows the SMLT edge MLT flow hash to directly translate into the two available BVLAN equal cost shortest paths.

Note

In the SMLT case (traffic from right to left), both BEB-1 and BEB-2 will receive traffic for the same L2 VSN VLAN and hence some flows of L2 VSN I-SID 2001 will be forwarded by BEB-1 in BVLAN1, while other flows in the same L2 VSN will be forwarded by BEB-2 in BVLAN2.

Tip

From a CFM standpoint, performing L2 ping or traceroute or tracetree on the ingress BEB needs to be performed on the corresponding BVLAN. This remains deterministic as required by SPB, whereby BEB-1 uses BVLAN1 while BEB-2 uses BVLAN2.

We'll conclude this section by looking at how IP Multicast load balancing is handled with Fabric Connect. As already mentioned in IP Multicast Over SPB on page 89, each and every multicast stream (defined by a unique combination of Group IP address, Source IP address and ingress BEB) is allocated an I-SID that defines the shortest path multicast tree to deliver that stream to any BEB that has signalled its interest in being part of it (based on whether or not it has IGMP receivers). These I-SID trees can be created independently in one or the other BVLAN. Let us consider the example in Figure 91, where again the four paths between BEB-1 (or BEB-2) and BEB-3 are all of equal cost.

Source-Group streams 1 & 2 originate from BEB-3. Stream 1 was the first stream detected by BEB-3, hence it was dynamically assigned I-SID 16000001 on BVLAN1. Stream 2 was detected next, and was assigned to I-SID 16000002 on BVLAN2. And so on. Essentially, the Multicast reserved I-SIDs 16000001-16600000 are still assigned to BVLAN1 and BVLAN2 on an even/odd basis. However, they are assigned sequentially, as new IP Multicast streams are detected by the ingress BEB. This therefore produces a good spread of multicast streams across both BVLANs and we can see that spread translating into an effective load balancing for Streams 1 & 2.

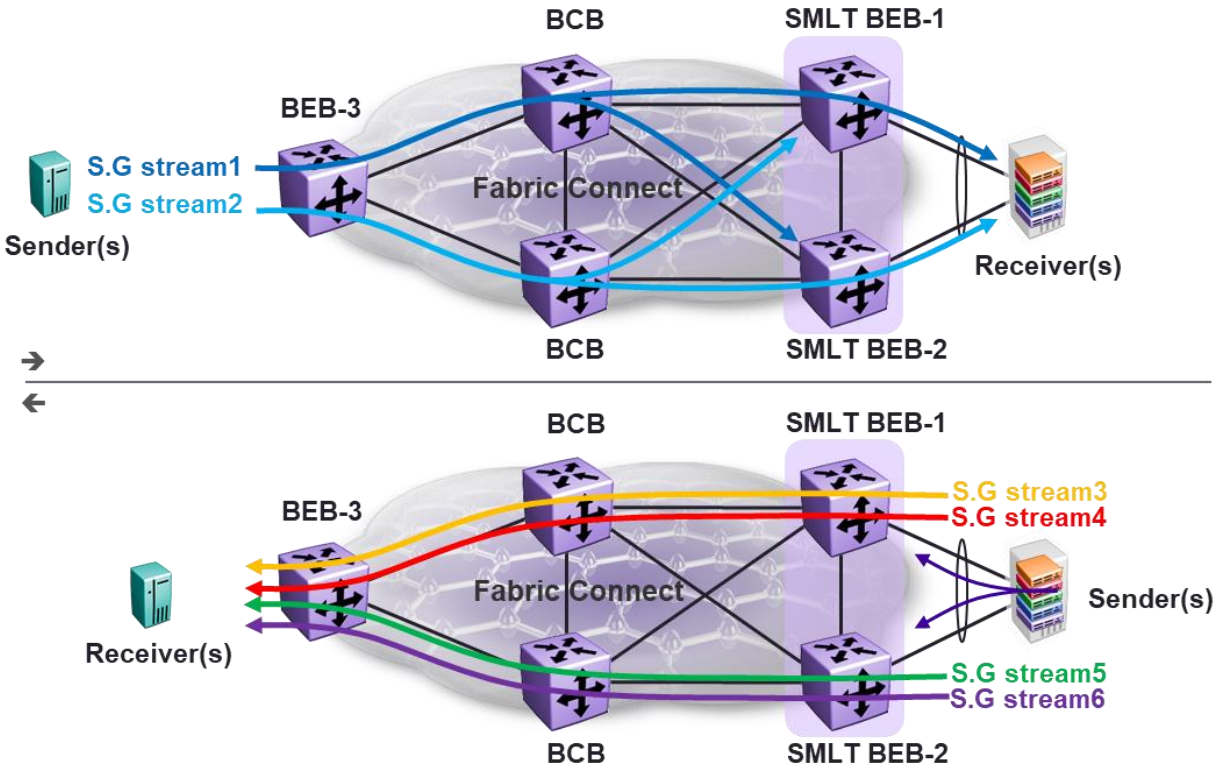


Figure 91 IP Multicast Load Balancing into SPB Equal Cost Shortest Path Trees

Tip

BEB-1 will always forward stream 1 onto the SMLT links towards the receiver, because BEB-1 is primarily operating on BVLAN1 within the SMLT cluster. Likewise, BEB-2 for stream 2 on BVLAN2. However, BEB-2 will also add itself to the multicast tree of stream 1 on BVLAN1 (and likewise BEB-1 for stream 2 on BVLAN2), because it might have to forward stream 1 on its SMLT link, should the SMLT link of BEB-1, or BEB-1 itself, suddenly fail. The solution consistently delivers sub-second failover.

In the reverse direction, Source-Group streams 3 through 6 originate from the SMLT cluster side. Provided that the SMLT edge device is capable of performing an MLT hash for Multicast traffic, we can expect on average that BEB-1 will receive 50% of those streams and BEB-2 the other 50%. As streams 3 & 4 are hashed to BEB-1, they will get assigned to I-SIDs 16000001 and 16000002, respectively, both on BVLAN1.

At the same time, the same I-SIDs will be signalled over the IST connection and BEB-2 will allocate I-SIDs 16000001 and 16000002 for the same streams but on BVLAN2. Both BEB-1 and BEB-2 will create shortest path multicast trees in the fabric, for both I-SIDs on both BVLANs. For a given I-SID, only one of the data plane trees will actually carry the multicast traffic, but in case of SMLT link failure on BEB-1 or failure of BEB-1 itself, streams 3 & 4 will immediately switch to BEB-2, which is thus able to instantly forward the streams without any delay for having to create the multicast tree.

The reverse is true for streams 5 & 6, which are hashed to BEB-2 and get allocated I-SIDs 16000003 and 16000004 respectively, on BVLAN2. Again, these same I-SIDs are signalled by BEB-2 to BEB-1 over the IST connection and BEB-1 will allocate them on BVLAN1 as well.

The end result is a good load balancing spread of IP Multicast streams across the SPB fabric which when combined with SMLT clustering on the BEBs provides unbeatable sub-second resiliency for IP Multicast applications which no other networking technology to date is able to deliver.

Human Level Resiliency

It is a fact that human error is often a cause and/or can contribute toward network failures. Whenever a configuration change is required, whether to add a new service or modify an existing one, there is chance that human error might entail a non-desirable outcome, ranging from either the new service not working properly, or that some other existing service is impacted, up to extreme cases where the entire network and all of its users are impacted by a complete outage.

Complexity and repetitive work greatly increase the chances of human errors occurring. Designing and maintaining a network infrastructure capable of performing network virtualization inherently adds a certain level of complexity which is absent in traditional architectures where no virtualization is required. That said, by its very nature, Fabric Connect represents a huge leap in terms of reducing network complexity as compared to more traditional architectures, including MPLS.

In this context, traditional technologies that attempted to offer network virtualization by provisioning a service on a hop-by-hop basis, on every node including access and distribution and core, are to be avoided. These outdated design approaches would typically extend an L2 VLAN service by configuring that VLAN on every link and every core switch across the network, often mandating the use of Spanning Tree in the Core. Likewise, to extend an L3 routing domain a VRF would be configured on every node, including core nodes, and a separate and independent IP routing protocol is spawned for each VRF in the network (this design approach is often referred to as VRF-Lite). Clearly these acrchitectures maximize the configuration effort as well as the network nodes that need to be touched, and thus increase the chances for human error to take its toll.

In a Fabric Connect architecture, all VSN service types are configured via end-point provisioning. This property is crucial. Whenever a new VSN service needs to be created / modified / removed, only the edge node where the service exists needs to be touched. The core nodes, arguably the most critical components in any network design, do not need to be touched. More importantly, in an end-point provisioning capable architecture such as Fabric Connect (and also MPLS), the core nodes are simply agnostic of the virtual networks they transport. This property goes a long way to limiting the scope of human error as it reduces the impact of any errors to network edge nodes where any adverse impact will be more localized.

In the Extreme Networks Fabric Connect architecture, where SPB is extended from the data center ToR switch to the remote branch office as well as to the wiring closet access switches via Fabric Attach, end-point provisioning is absolute for placing users and servers on the correct L2 segment (VLAN/L2 VSN) and will only require touching the distribution BEB nodes when a new IP routed segment is created or a new

routing domain end-point is provisioned (VRF/L3 VSN). The same is true when activating IP Multicast within a service.

Tip

With MPLS architectures, the end-point provisioning goes only as far as the PE distribution node and in no cases to the access switches. This is true for both IPVPN and VPLS service types. In the case of Draft Rosen Multicast VPNs, the MPLS core will need to be PIM enabled.

Indeed, in an Extreme Networks Fabric Connect architecture, if we assume that all the L3 VSN routing domains are already in place, adding a new application or new user to the network can be completely liberated by any human intervention on the network switches as users can be assigned to the correct L2 segment via identity based routing (RADIUS Authentication with Fabric Attach) and new VMs can be automatically attached to the correct server VLAN by the network management automatic provisioning software or Fabric Attach in the OVS software in the hypervisor.

Loop Detection and Protection Mechanisms

So far, we have covered how human error can have an impact on network provisioning when software configuring the network devices and how Fabric Connect minimizes these risks. There is however another important area which is prone to human errors and which is physical patching of cables, fibres and the patch panels which both make use of. We will also include here human errors originating from personnel configuring the server hypervisors in the data center, since these components have software-based vSwitches within them which, if misconfigured, can ultimately lead to the same network failure conditions.

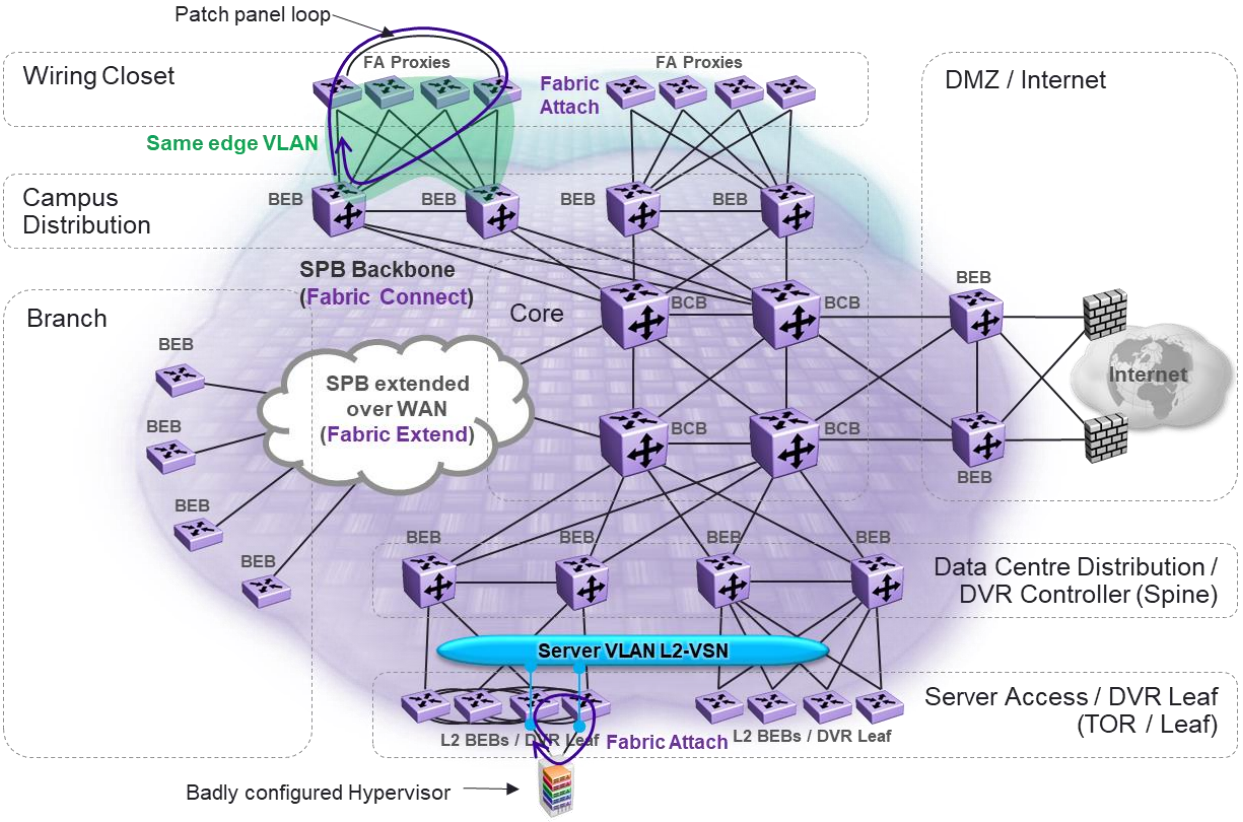


Figure 92 Loop Forming on Access VLAN

These failure conditions fall into two categories; both are very serious:

- Loops forming on access L2 VLANs. A full bridging loop will result in broadcast/multicast traffic being looped around at the rate of the slowest link in the loop. With gigabit in the wiring closet

access and 10Gbps in the data center access ToR, any loop will fully consume those rates on links involved and will impact the CPU utilization on the nodes which own the looped links as the source MAC of the looping packet will be in constant conflict with the real location of that MAC. Half loops (which cause packet reflection) will also result in MAC learning disruption as well as connectivity issues for those MACs, though will not result in a full bridging loop.

Tip

SPB is a loop-free technology. There simply cannot be any loops within the SPB core. However, loops at the access of Fabric Connect are still possible.

- Different VLANs belonging to different VSN services being collapsed together. This can easily happen via incorrect patching on patch panels. The end result is that all users on a VLAN/Subnet will be collapsed (bridged together) with other users on a different VLAN/Subnet that could even belong to a completely different routing domain L3 VSN. This does not constitute a full bridging loop. Yes, both VLANs will learn all the MACs in both the VLANs, but in practice it is not a service affecting fault. All users in both VLANs will keep on operating correctly via their respective default gateway, completely unaware that they are now operating on what is effectively a multi-netted L2 segment. The real implication of such a fault is from a security perspective as users in one VSN can potentially gain unauthorized access to another VSN.

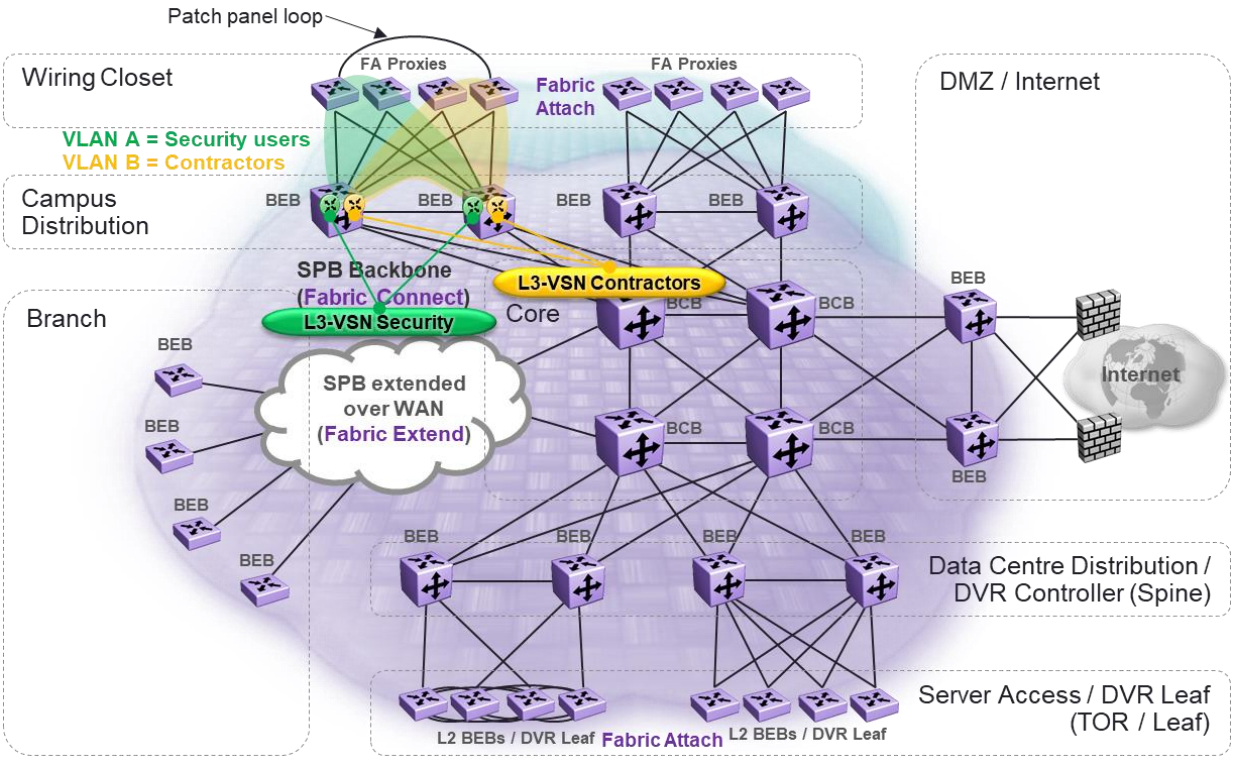


Figure 93 Access VLANs Collapsed Together

It is worth noting that a first step toward making sure that a full bridging loop is constrained in its intensity is by making sure that broadcast rate-limiting is always set on all access ports. There is absolutely no reason for broadcast traffic to be allowed to run line rate on a gigabit access port. The same can be usually said of multicast except if bandwidth-intensive IP multicast applications (such as high definition IPTV) are in use.

Yet ultimately what is needed here is a design approach that can reliably detect and prevent the above access layer loop conditions from occurring in the first place. The Extreme Networks Fabric Connect architecture offers a solution with three layers of defense based on:

- Spanning Tree BPDU Filtering (also known as BPDU Guard) on access ports.
- Simple Loop Prevention Protocol (SLPP) Guard on access ports.

- Simple Loop Prevention Protocol (SLPP) on distribution (IP gateway) nodes.

Extreme Networks' best design guidelines mandate that user and server access ports must always be configured with Spanning Tree enabled as MSTP Edge ports (or STP FastStart). In an Extreme Networks Fabric Connect design, these are the only ports where Spanning Tree should be left running; this is done to ensure they emit BPDUs every second, which can then be used to preempt a patch panel loop condition. The very same access ports should also always be provisioned with both BPDU Filtering and SLPP-Guard enabled.

An access port with BPDU Filtering enabled will immediately go offline if it receives a Spanning Tree BPDU. Likewise, an access port with SLPP-Guard enabled will immediately go offline if it receives an SLPP packet.

SLPP is a simple protocol that consists of generating an L2 broadcast packet (an SLPP PDU, which carries information of the VLAN it was originated on) on a given VLAN at regular intervals. Like for any broadcast packet, the expectation is that these packets will be flooded across all port members of the VLAN but are never seen to come back (on same or different VLAN). SLPP is configured on distribution layer nodes, which is usually where all the VLAN IP interfaces are also configured (though SLPP has no IP dependencies). SLPP PDUs will be flooded on VLANs and L2 VSNs alike.

Let us cover the layers of defense as they would trigger following a patch panel loop condition.

First Line of Defense – BPDU Filtering on Access Ports

1. Patch panel loop is introduced.
2. Link up event on access ports involved.
3. Spanning Tree BPDUs will be the first packets to be sent out of the access ports.
4. If the patch panel loop is just a cable, the access ports will receive eachother's BPDUs and BPDU Filtering will immediately shut down one or both the access ports. A trap is generated and the fault condition is averted.

Second Line of Defense – SLPP Guard on Access Ports

1. If the access loop is via a hypervisor server, or via an IP phone (neither of which is likely to relay or generate BPDUs), BPDU Filtering will not detect anything and both access ports will settle into a Forwarding state where packet forwarding onto the external loop commences.
2. The first packets to make it out onto the external loop segment will be broadcast or multicast packets among these SLPP packets (which are typically generated every 500 milliseconds).
3. The first access port to receive an SLPP PDU over the external loop segment will immediately shut down thanks to SLPP-Guard. A trap is generated, no loop has time to form, and the fault condition is averted.

Third Line of Defense – SLPP on Wiring Closet Distribution

1. This scenario is only applicable if the access switches were not correctly provisioned with SLPP-Guard or there is a problem with the access switch's MLT uplinks into the distribution nodes causing packet reflection. In this case, SLPP-Guard on the access switches cannot be relied on to detect or prevent looped packets.
2. In this case, a loop condition will take hold (either a full loop or a half loop or collapsed edge VLANs)
3. If we are experiencing a full or half loop within the same VLAN:
    a. Distribution nodes that are running SMLT links (with or without Fabric Attach) toward the wiring closet access switches will start seeing SLPP PDUs returning to them on those uplinks, and will count these packets against configured slpp-rx-thresholds.
    b. A first distribution node will disable its uplink to the offending wiring closet switch and generate a trap. If the fault condition was due to MLT misconfiguration on the access switch, the fault condition is corrected and users connected to the access switch retain connectivity. However, that access switch has now lost uplink redundancy.

   c. If instead the fault condition persists, the second distribution node will also disable its uplink, generate a trap, and the fault condition is corrected by isolating the offending wiring closet switch from the rest of the network. However, all users connected to that access switch will lose connectivity.

4. If instead we are experiencing two different VLANs collapsed together:
   a. Both distribution nodes will start sending Trap alerts indicating that SLPP PDU from VLAN X is being received in VLAN Y (and vice-versa). In this case no action is taken on the uplinks since this error condition does not result in a bridging loop, nor does it impair user connectivity (which would instead be impacted if the uplinks were shutdown). The traps are sufficient to alert the network administrator that there is a security breach between two different VLANs.

Fabric and VSN Security

This section covers the security aspects of managing and maintaining a virtualized network architecture such as Extreme's Fabric Connect SPB. We start by covering the benefits that separation and segmentation offer in a virtualized architecture, and how such architecture must conceal itself from outside users to repel potential attacks. SPB and its different VSN types are then analysed in their stealth-ness and robustness to outside attacks.

Address Space, Routing, and Traffic Separation

Traffic separation is an essential component to network security. The ability to segment disparate users and applications into private virtual networks and to prevent communication where this is not warranted helps harden the network against potential attacks. The virtualized network architecture which Fabric Connect delivers represents the best possible way to achieve this separation via the use of L3 VSNs, L2 VSNs as well as the flexibility of doing Inter-VSN routing of L2 VSN segments that can be virtualized in VRFs which in turn may or may not belong to L3 VSNs.

With Fabric Connect, traffic separation is ensured by the use of the I-SID service–ID within the SPB backbone and by a combination of separation of IP routing tables using VRFs (on the L3 VSN BEBs) and MAC address forwarding tables per I-SID (on L2 VSN BEBs).

As a consequence of this address space segmentation, derives the additional flexibility of a virtualized networking architecture whereby the actual address space used (whether MAC addresses or IP address spaces) is only significant within the VSN to which it belongs. It follows that different VSNs can use the same IP address space without any conflict.

With Fabric Connect the VSN is provisioned only on the edge BEB nodes where the VSN is determined by the I-SID value added to the VLAN for L2 VSN or VRF for an L3 VSN. In summary, the VSN service provides the following:

- Any VSN can use the same address space as any other VSN.
- VSN provisioning is done on the edge BEB's switches where required; there is no IP knowledge in the core.
- Transit BCB core nodes have no IP configuration and only forward traffic based on System-ID (BMAC) destination. On these nodes, there is no knowledge of the VSNs they transport.
- Traffic from one VSN can never flow to another VSN.
- Traffic from one L3 VSN cannot be forwarded to another L3 VSN unless it is specifically configured to do so.
- Traffic from one L2 VSN cannot be forwarded to another L2 VSN unless both L2 VSNs have an IP gateway interface defined and belonging to the same routing domain (VRF).
- Each L3 VSN can support multiple protocols with redistribution into IS-IS.
- This includes Direct, OSPF, RIP, Static, and BGP. The L3 VSN BEB terminates the service on a VRF instance. On this VRF any of the above-mentioned IP routing protocols can be instantiated.

Every VSN is crystallized by an I-SID service-ID value; only when two or more BEB switches are provisioned with the same I-SID value (for the same VSN service type, L2 or L3) will they be able to forward traffic between them.

Concealment of the Core Infrastructure

Concealment of the core infrastructure is a very important property for virtualized network architectures, as it makes it much harder for potential outside attackers to gain any useful information which could be used in an attack to compromise the availability and security of the network. Concealment means that the core functions of the network are invisible to outside networks and the Internet to which the virtual VSN networks might be connected.

Note

Traditional IP core based networks using MPLS/BGP are very sensitive to this topic and go to great pains in explaining how these concerns may be mitigated via the use of extensive packet filtering and by limiting the IP routing information which is made available even within VPN networks. Comparisons are often drawn to the concealment properties of ATM and Frame Relay networks to which MPLS architectures aspire.

In this respect, SPB represents yet again a paradigm shift from IP-based core infrastructures. Every IP interface seen in a network is like a door that an attacker will try to pry open or to scan the topology of the network itself. By its very nature SPB runs directly over Ethernet using IS-IS as the control plane protocol and thus does not have any IP dependencies. IP becomes purely a virtualized service running on top of SPB and hence any IP interfaces only exist at the service presentation level of an L3 VSN at the edge of the network.

The Fabric Connect core is thus simply invisible to any IP scanning techniques. Anyone running an IP scan against the environment would get a simple list of IP subnets all showing a single hop to one another. The topological details of the core are totally dark to the scanning attempts because there is simply no IP running in it; it's not required. Each IP network point of presence views all other IP networks not as the next hop to it but as the actual service point of presence on the other side of the Fabric Connect cloud.

So let's take a closer look at Extreme Networks Fabric Connect Stealth Networking.

Extreme's Fabric Connect Stealth Networking

The primary aspect of networking is to establish and maintain an end-to-end path. In the legacy model where IP is used in the core, this creates a 'catch-22' in that the IP protocol is not only the service that is delivered, but it is the utility which at a foundational level establishes the sense of a network path. This means that all other levels of abstraction to provide for service virtualization and hence privacy is built upon it.

A good analogy is brushing one's trail during clandestine operations such as in reconnaissance. The method involves erasing one's trail and backtracking to a place where the trail is effectively obscured such as a creek bed or a rocky surface. If one thinks about it, one has to also ingress the area from that point as well. This creates an ingress point that really cannot change. As one brushes the trail, care must be taken to ensure that:

- There are no visible strokes in the dirt.
- Random debris is placed about that matches neighboring areas in both density and pattern.
- There are no errant footprints or breaking of foliage as one backs out of the area.

This is obviously a very difficult thing to do that only the true master can attain in a reliable fashion. In most instances, the brushed trail looks just as obvious, if not more so than the actual footprints. A seasoned tracker will look for telltale signs and then go from there to nearest vectors and search for other clues. The task is difficult because we are bound to a single plane, in this case the ground. As such, no brushed path will be perfect. It's just a question of whether the tracker will be good enough pick it up or not.

We cannot divorce ourselves from the fact that we have to use the path that we are attempting to conceal. The analogy here is very strong to methods for private networking today. As we are dependent upon IP to

establish the initial service path (sections thereof), all additional path notions such as BGP and MPLS are dependent upon, it meaning that these networks are potentially vulnerable to IP scanning techniques. Strong access control lists can mask the environment from the general routed core, but this carries with it its own set of conundrums in that path behavior is dependent upon reachability, as such there is only so much that can be masked. Certain nodes will need to 'see' the IP reachability information, so all of this leads to a scenario very similar to the trail brushing analogy.

But consider a bird. A bird can arrive at that given location. It will most certainly take a path to get there as well as one to leave. It will also leave footprints where it lands; this would be its 'point of presence' on the ground. Beyond this, however, there is no trace of the path that the bird took even though it did indeed take one. No amount of tracking on the ground will effectively yield the path information.

This is because the paths for the bird are occurring on a different plane. Here the analogy to Fabric Connect is also very strong. In Fabric Connect, path behavior is created at the Ethernet Switched Path level (hereafter referred to as ESPs). All ESP knowledge is handled within resident link-state databases in each Fabric Connect switch node. As a result, IP simply becomes a service around the edge of the Fabric Connect Cloud.

Much like the bird footprints, an IP subnet becomes a 'service point of presence'. Explicit path information however is totally obscured from the perspective of IP because the path is not a routed IP hop-by-hop path; it is held as an ESP at the Ethernet Shortest Path Bridging level.

Tip

The definition of a Stealth Network Topology is as follows:

"A network that is self-contained, with no ingress in or out of it except by strictly controlled secure access points. The network must also be dark and not visible to IP or other topological scanning techniques. As such the potential surface for any such activities is either highly mitigated and protected or totally eliminated due to true isolation."

In the Extreme Networks Fabric Connect architecture the different VSN types have slightly different stealth attributes. These are listed in Table 13. By virtue of running over SPB, they all share the stealth property that the core's topology cannot be inferred.

Table 13 - Stealth Properties for SPB VSN Types

Stealth Properties	L3 VSN I-SID → VRF	L2 VSN I-SID → VLAN	IP Shortcuts GRT (VRF-0)
Core's topology cannot be inferred	✓	✓	✓**
IP interfaces need not exist	✗	✓*	✗
IP interfaces only exist at the edge as gateways for IP end users in subnet	✓	n/a	✓
All remote IP subnets are 1 hop away	✓	n/a	✓
Network IP interfaces are not bound to any socket (e.g., SSH, SNMP, HTTPS)	✓	n/a	✗

* NOTE: Any IP interfaces provisioned will expose the L2 VSN to the routing instance to which that IP interface belongs (i.e., GRT IP Shortcuts or VRF which may or may not be part of an L3 VSN).

** NOTE: While the core topology cannot be determined, the individual switch elements can still be enumerated through the IS-IS Source IP addresses.

The most stealth VSN is an L2 VSN where no IP interface has been defined, as this is a totally closed L2 environment where nothing can enter or exit unless otherwise provisioned. It is invisible to the IP protocol as IP is not being used. IP can still run 'inside' the L2 VSN but with the IP subnet being established by the devices that are attaching into it. This type of service is useful for protocols that are used for control and management of security critical infrastructure such as power grids, subways, and trains as well as automated production and manufacturing floors. These environments are extremely sensitive and providing a totally closed L2 environment for these types of applications is very important.

Note

An L2 VSN where IP address(es) are defined on the VLAN of the terminating BEB(s) will expose the L2 VSN to the routing domain (GRT IP Shortcuts, or VRF which may or may not be part of an L3 VSN) to which that IP interface belongs. In this case, we would consider the stealthness of that routing domain service.

With L3 VSN and IP Shortcuts, each IP subnet is advertised from its point of presence into IS-IS. Within these service types, IP routing happens exclusively at ingress and egress of the SPB fabric. Each subnet views itself as one hop away even though in reality there are a wide variety of potential ESPs involved in the actual data transits. The core of the network is consequently 'dark' to the IP protocol.

As such, IP scanning yields little information, and no sense of topology. Even the IP routing table, which clearly exists in the GRT / VRF where the service is terminated will show all the IP routes known within the service, but those routes which are reachable across the Fabric Connect crucially do not have a next-hop IP address. The next-hop is in fact the IS-IS System-ID (BMAC) of the SPB node where that destination subnet has its point of presence. This is illustrated for the GRT below .

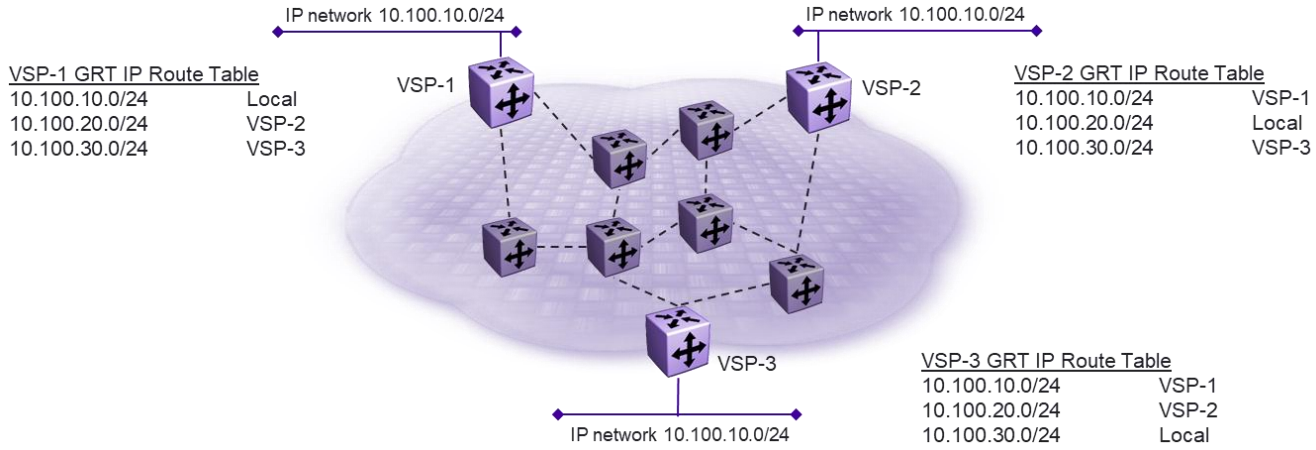


Figure 94 Stealth Networking with IP Shortcuts (L3 VSN)

An L3 VSN will present additional stealth properties due to the fact that, in the Extreme Networks VSP series platforms, VRF IP interfaces are not bound to any higher layer sockets (no management access) and hence cannot be used to gain any additional information from such interfaces and any interaction with them will be mostly limited to ARP and ICMP.

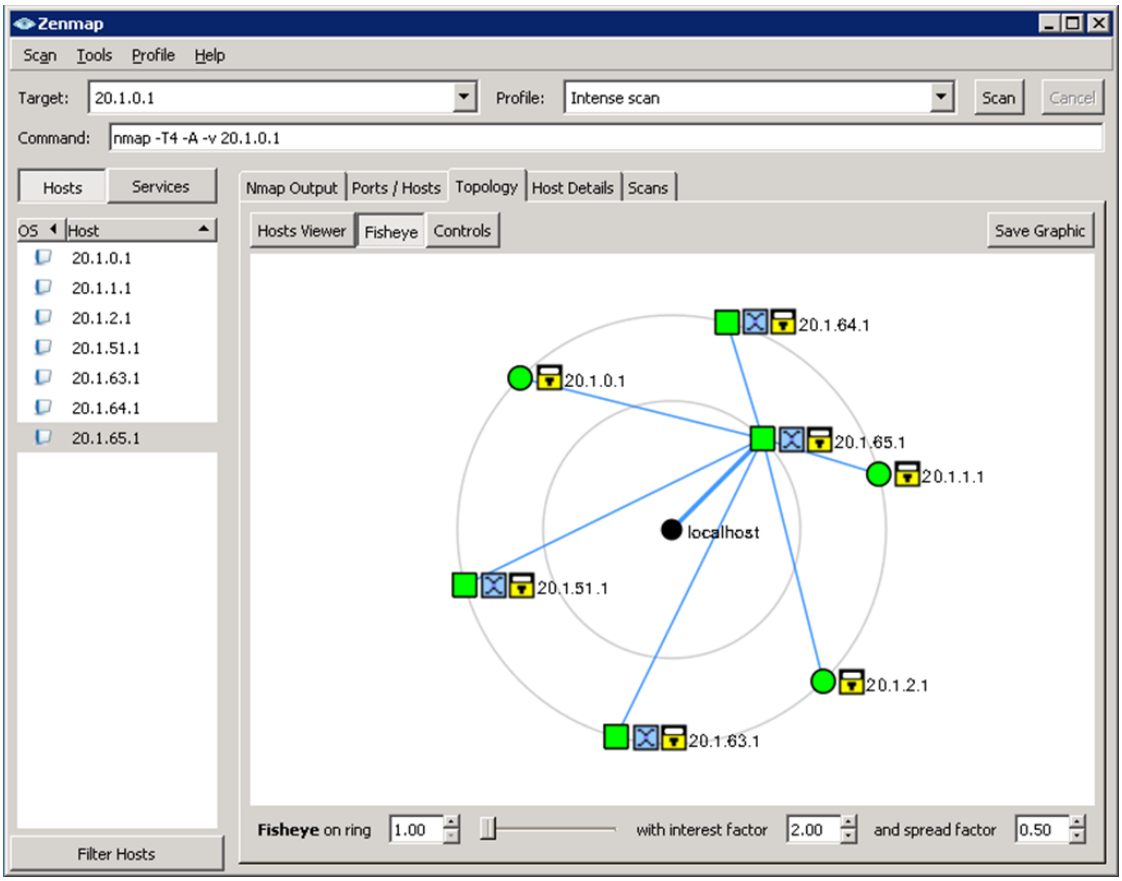


Figure 95 L3 VSN Topology as Seen by IP Scanning Tools

Resistance to Attacks

By resistance to attacks we intend the ability of the core to withstand any denial of service (DoS) attacks from the outside. It is essential that the core itself should not be accessible for such attacks and that if the attack is launched from within a VSN, no other VSNs be attained by it. An attack originating on an L2 VSN service that is not provisioned with any IP interfaces can only target other end-stations within the same VSN service.

An attack originating on an L3 VSN service will be able to target the gateway IP interfaces on the VRF where the VSN service is terminated. However, there are no open ports (sockets) on these interfaces with the exception of DHCP (for the relay agent) and possibly of some routing protocol (e.g., OSPF, RIP, BGP), but only if an instance of these protocols has been created on the VRF and enabled on the IP interface in question. By default, the VRF IP interface will only respond to ARP and ICMP, as such any attack will be limited to ARP poisoning and spoofing (covered in next section).

An attack originating on GRT IP Shortcuts routing domain does instead present some risks, as the network IP interfaces within this domain are activated for device management and bound to many sockets used for management protocols (e.g. SSH, SNMP, HTTPs). Extreme's recommendation is therefore that the GRT IP Shortcuts routing domain be reserved exclusively for management of the Fabric Connect infrastructure by only extending it to a management subnet where the Network Management stations reside on one side and to the SPB nodes on the other. By doing so, this routing domain need not be extended to any user VLAN on the network access ports.

Tip

As a matter of comparison, consider that there are two basic ways an MPLS core can be attacked: by attacking the provider-edge router, or by attacking the signalling mechanisms of MPLS. Both types of attacks require specific router configuration via ACLs to be repelled.

In the Fabric Connect model the latter is simply not applicable nor possible. The former (attacking the L3 VSN BEB) is equally applicable, though by default the VRF IPs have the highest levels of protection enabled, without requiring additional configuration.

Impossibility of Spoofing Attacks

Packet spoofing and replay attacks are a form of impersonation attacks whereby an attacker uses a false identity (or spoofs the identity of another legitimate device) to obtain unauthorized access to a VSN and its associated services. This is equally possible at Layer 2 (in L2 VSNs) and Layer 3 (in L3 VSNs) where the attacker will try to generate packets spoofing either an L2 MAC address or an L3 IP address. If the receiver accepts the spoofed packets, this could allow the attack to either fool the network, or the receiver, into forwarding to it traffic flows that the attacker can then scan for authentication sequences, which ultimately could lead to unauthorized access.

In an L3 VSN service, spoofing VSN IP routes would require the attacker to be able to inject invalid IP routes into the BEB's VRF. For this to happen, the attacker would need to try to poison the VRF routing table using a dynamic routing protocol.

Tip

In the Extreme Networks Fabric Connect implementation, by default, no IP routing protocol instances exist on VRFs; these are not needed for SPB L3 VSNs to operate. Do not activate any IP routing protocols on VRF local VLAN IP interfaces where users may reside (subnet mask smaller than 30 bits).

If a dynamic IP routing protocol (such as OSPF, RIP, BGP) does need to be enabled on a VRF terminating an L3 VSN (because an authorized traditional IP router needs to be connected to the VSN), connect these routers using point-to-point 30-bit IP subnets and always use protocol authentication options (e.g., HMAC-MD5).

Spoofing other IP users' IP addresses within the same access user subnet can be easily done using ARP spoofing techniques. The only way to repel these types of attacks is to ensure that standard edge security features are deployed on access switches.

Tip

Use these Layer 3 security access features:

- IPv4
- DHCP Snooping
- Dynamic ARP Inspection
- IP Source Guard
- IPv6 - First Hop Security (FHS) - RIPE 554
- DHCPv6 guard
- Router Advertisement filtering (RA guard)
- Dynamic Neighbour solicitation or advertisement inspection
- Neighbour reachability detection inspection
- Duplicate Address Detection (DAD) inspection

Spoofing MAC addresses is even easier than spoofing IP addresses, as the dynamic nature of Ethernet transparent learning bridges (which remains true in VLANs and inside L2 VSNs services alike) is that source MAC addresses are inspected and forwarding tables are immediately updated if a given MAC address is seen arriving from a different port. An attacker could try and spoof the MAC address of another user in the same segment or of the router IP acting as default gateway for the segment. The best way to repel such attacks is to enforce user authentication on the access ports via 802.1X EAPoL.

Tip

Extreme Networks offers 802.1X EAPoL with Multiple Host Multiple Authentication (MHMA) and Multi-VLAN support. This implementation leverages MAC-based VLAN support in the underlying hardware such that once a device has successfully EAP authenticated onto the network, only packets with the same source MAC address which was authenticated will be allowed.

Stealth Networking Design Guidelines

While Fabric Connect offers inherent stealth characteristics it is important to understand best practice design guidelines to minimize the exposed IP footprint given to a would-be attacker for any critical information or service. The first thing to address is the separation of the network management control plane from the client user communities. Even though Fabric Connect operates at Layer 2, it requires Layer 3 IP interfaces for communications and management of the individual fabric nodes. These interfaces are connectionless IP interfaces that are associated with VRF0 on the Global Routing Table (GRT). Therefore, it is imperative that this domain of interest be reserved in totality only for IT administrative access for purposes of network control and security.

All other user or device presence needs to be relegated into L2 or L3 VSNs as the case may require. Under no circumstances should any leaks be allowed into the GRT. While we do provide certain instances of code-based protection, it is important to realize that such leaking can occur inadvertently by design. This section is intended to review the best practice guidelines to avoid such a scenario.

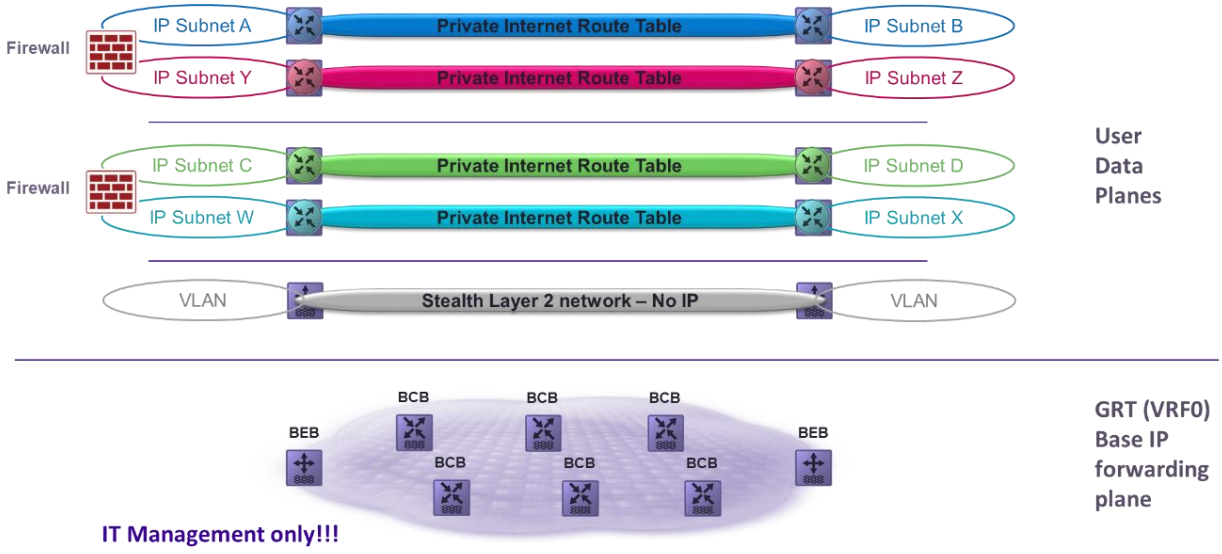


Figure 96 Isolation of the Global Routing Table (VRF0)

By default, all innate IP services are based on the GRT. This is the only viable channel of management communication without the use of dedicated VSNs and physical loopback of management interfaces, which quickly results in an obtuse implementation. By removing the client user and device communities from the GRT we provide for a very clean and dedicated IP environment for the management of the fabric and security infrastructure. No security demarcation interfaces should be allowed between the GRT and the user community domains. (If any are allowed, proper security exception procedures should be followed and maintained.)

Access to the GRT should be based on strong multifactor authentication as required by the environmental security policies of the organization. Ideally, the administrator should possess two separate devices for access. One device would be dedicated for access to the GRT with associated separate user credentials for administrative concerns and another for normal user access with separate user credentials for normal access. While virtualization can be used to achieve same device separation, it is outside of the scope of any separation that the fabric provides and therefore is a potential point of exposure to the GRT. Best security practices for the given virtualization environment should be followed. Given that we have made accommodations for the isolation of the GRT we will now address the various user level services that the fabric has available.

Layer 2 Virtual Service Networks

As covered earlier, L2 VSNs are nothing more than VLANs at the service edge associated with I-SIDs in the service core. Note that at the very basic primitive the L2 VSN is a true Layer 2 phenomenon. It can exist completely on its own without the use of any IP whatsoever. Moreovver, these services can be extended farther and more extensively than any traditional tagged VLAN approach. By default, they have no association to any VRF. As such they result in totally isolated L2 domains of interest that are totally ‘dark’ because there is simply no IP against the context of the topology.

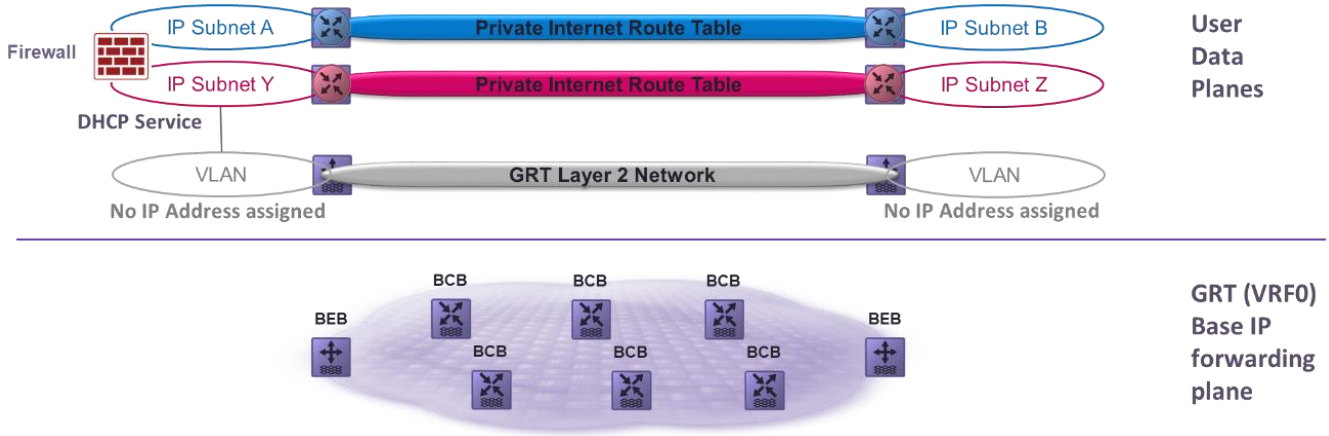


Figure 97 DHCP Services for L2 Virtual Service Networks

As the figure above illustrates, you can even run DHCP services within the L2VSN and provide for a default gateway for a complete IP user environment but the only point of exposure is the default gateway. There is no other way in or out for the service. There is an obvious ‘data corralling’ effect that occurs in this scenario and it allows for a security demarcation for the service and any allowed external traffic. This is a great approach for controlled access environments such as captive portals for guests or contractors or in highly regulated environments such as PCI.

Note however that if any VLAN termination assigned to the given I-SID were provisioned with an IP address, it would appear in the GRT and have access to the network control plane as shown in the illustration below.

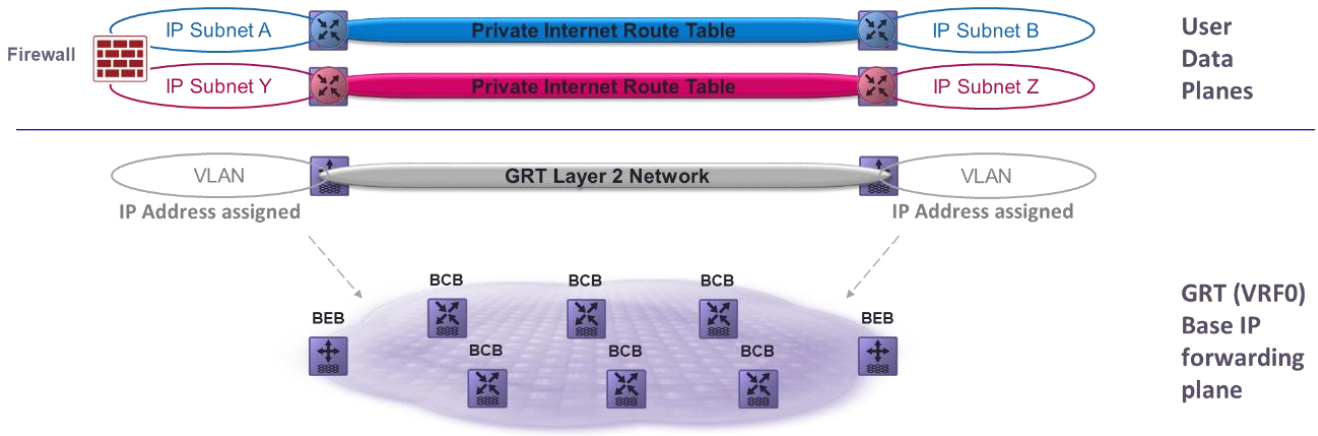


Figure 98 GRT (VRF0) L2 VSN

The L2 VSN is now an extension of the GRT and is such part of that given security domain. If this is done unintentionally, a significant security breach could occur. There may be valid reasons for doing this, so it is important to consider the exceptions and more importantly document, monitor, and audit them as you would any security exception policy. But as a general practice this should be avoided. The best practice is to never assign IP addresses to the VLANs at the L2 VSN service edge. Any IP 'personality' should be provided by DHCP services within the L2 VSN.

Different L2 Service Categories

E-tree

While the typical L2 CVLAN will behave as a typical LAN, there are other modes of behavior that are useful in an overall security model. The first is a service known as E-tree, which is a hub-and-spoke L2 service domain that enforces the prevention of direct spoke-to-spoke (peer-to-peer) traffic. All traffic with the E-tree will follow from the spokes up to the hub. There, communication can be allowed or disallowed as per policy and security infrastructure. As a result, E-tree topologies are useful in linear data footprints such as PCI card holder data environments where peer-to-peer communication between individual points of sale is strictly prohibited.

Transparent UNI

In some instances, there is the need to transport a service and that is the end of it. It could be a service-independent trunk for an agency or a department that you are hosting. It could also be a public access network that you are 'piping' through your network and do not wish any connection but only wish to provide transit. In these instances, the transparent UNI is the service of choice as it will accept any properly formatted Ethernet frame and transport it regardless of category (tagged/untagged). Note that the transport tunnel is in separated service plane that allows for provider like capabilities.

Switched UNI

VLANs are a local service phenomenon to the Fabric Connect edge. As such, they can be associated with different I-SIDs as required there by separating their communities of interest at the fabric edge. In essence, two end stations of the same VLAN ID at the network edge can be associated with two different I-SIDs in the Fabric Connect service core. This is useful for a number of reasons: first, it allows for a true multi-tenancy model where clients can have the same VLAN ID but belong to separate domains of interest; second, it allows for some very sophisticated security practices such as the ability to redirect a suspect system into a 'forensic' I-SID that places them into a sandbox. They still have complete network access and there is no way that they can determine that their network placement has changed and that they are under the microscope.

Layer 3 Virtual Service Networks

There are instances where you require a fully private IP routing environment. This is provided for by L3 Virtual Service Networks. In this service instance, the I-SID is used as the peering mechanism for the individual VRFs that exists within the constricted community of interest. In traditional networking, the VRF is typically peered using an IGP like OSPF or with a protocol like BGP for a true peering model. While these methods work, they are complex and difficult to maintain and troubleshoot. The replacement of these complex protocol constructs with the simple element of the I-SID realizes a ten-fold reduction in overall configuration requirements.

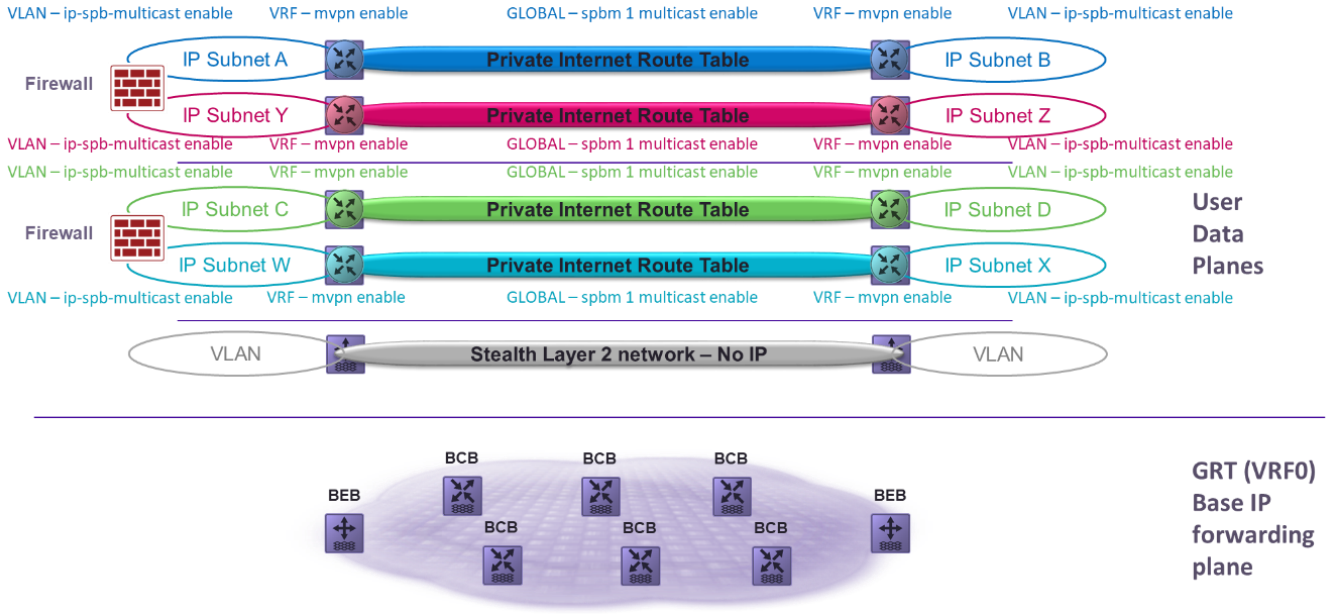


Figure 99 L3 VSN Topologies with Multicast Enabled

Anatomically, this service is a totally separate IP forwarding plane from the GRT. Any VRF1 to VRF*n* is a totally separate IP routing environment from VRF0 (GRT). As such, they are totally separate IP forwarding instances, not only from the GRT but from each other as well. They are constricted within the service paths of the I-SIDs they are associated with. This is important for a number of reasons such as massive operational expense improvements or drastic improvement in troubleshooting methods that would not exist otherwise. But the major point of interest here is that the I-SID is an atomic method of audit record for access and control to a given environment.

There may be instances where you wish to extend a given VLAN and subnet to other locations off a given VRF. While in most instances it is recommended to provision a new VRF at the remote location, there may be the need for an end-to-end L2 environment. This can be accomplished by using L2 VSNs to extend the local VRF VLAN to another location in the fabric. While this is a supportable configuration, IP multicast is not supported in this mode. When IP multicast is required, a new VRF on the I-SID should be provisioned with new IP subnet/VLANs.

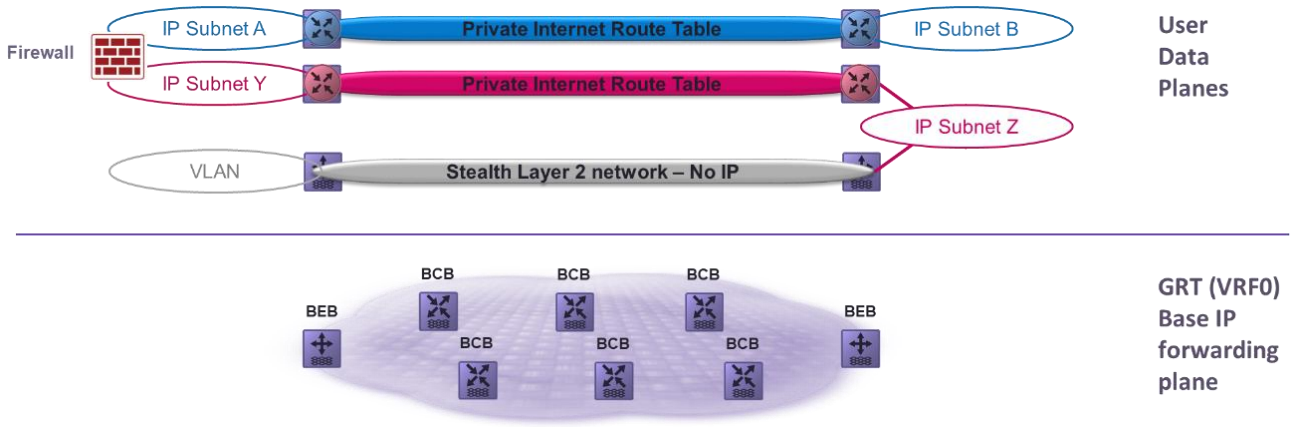


Figure 100 L3 VSN Extension

However, when IP multicast is not required and it is desired to have an end-to-end L2 service, the L2 VSN extension method is often selected. There is an important principle to remember regarding L2 VSNs. Any IP address assigned to the VLAN gets placed onto the GRT. This obviously creates a security issue with segmentation. In the figure above, note that BEB-A has a local VLAN termination to a VRF. It is in turn associated with an L2 VSN that extends over to BEB-B. The proper practice is to ensure that no IP addresses are assigned to the remote VLAN service termination point. By following this practice, any L2 VSN extensions off an L3 VSN will remain off the GRT.

Tip

The following is a stealth design quick checklist:

- Reserve the GRT for IT management and security practices only!
- Never assign IP addresses to L2 VSN service termination VLANs unless you intend on having the VSN be part of the GRT.
- When extending L3 VSN VRF subnets using L2 VSNs, never assign IP addresses to remote L2 VSN service termination VLANs unless you intend on having the L3 VSN be part of the GRT. This should be viewed as temporary measure and a switch capable of proper VRF termination should be deployed.
- Perform regular checks on the GRT to ensure that any inadvertent insertions are caught as soon as possible. Ideally the GRT should be only ISIS Source IDs and administrative systems for network management and security.

Fabric as Best Foundation for SDN

Software-Defined Networking (SDN) offers enterprises significant potential to not only increase efficiency and reduce operational cost, but fundamentally alter the way applications and infrastructure interoperate, consequently improving business agility, innovation, and competitive positioning.

While technically SDN can be described as a separation of control and data plane (and many industry players will have us believe that SDN is all about data center automation), the truth is far from it. The key value proposition of SDN lies in the ability to quickly and effectively integrate the programmable (network) edge with applications that support users and business processes.

The "edge" in this case is not to be understood as the typical network edge (device), but as the place where the network meets the application, i.e., a hypervisor supporting multiple application VMs or Docker containers, an edge voice application device (also known as a phone) or devices purpose-built as SDN edge devices.

This edge differentiates the way applications can use and/or interact with the infrastructure. At this programmable edge, the true power of SDN is unleashed.

Building upon a poor enterprise network foundation is ill advised. As enterprises deploy SDN-based applications, these applications need to enable agile development and reliable and fast deployment, while simultaneously supporting existing enterprise applications.

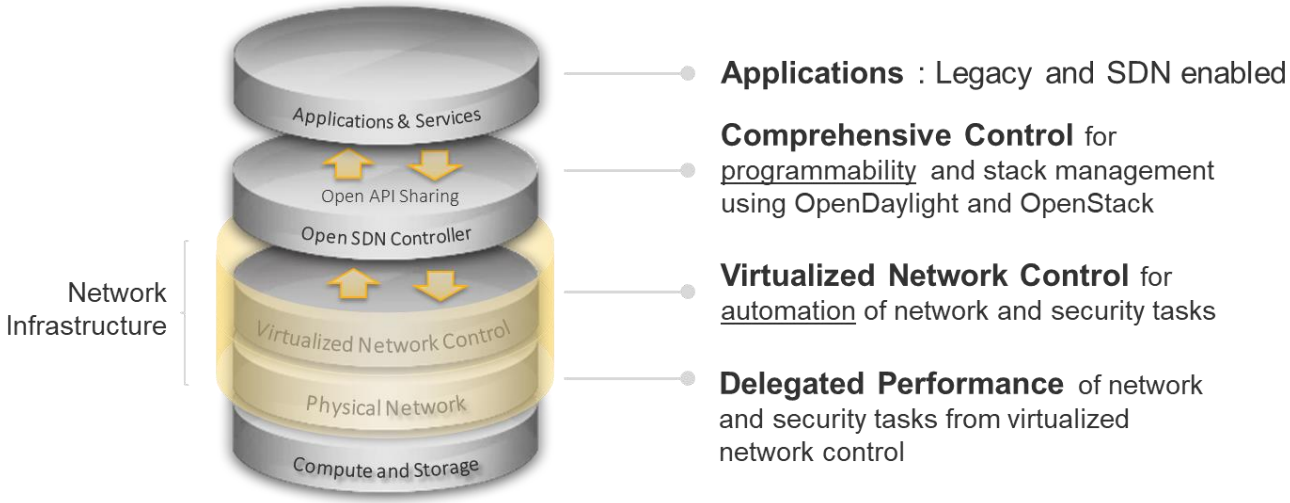


Figure 101 Orchestration of Applications and Services at Cloud Scale

Extreme's Fabric Connect Architecture is specifically designed to meet all these requirements through:

- Providing a fully automated, secure and reliable virtualized core network. The Fabric Connect core network is a true, modern, core network architecture that is drastically simplified, compared to complex, legacy network technologies. Fabric Connect relies on a single, provider-grade, standard core network protocol: Shortest Path Bridging (SPB). Fabric Connect offers customers an automated, zero-touch core network, full network virtualization, supporting up to 16 million service identifiers, sub-second switchovers in case of failures, full topological flexibility, and PCI-compliant stealth networking capabilities. A true enterprise-grade core network designed for maximum reliability and the lowest-possible operational cost.
- Enabling an automated and flexible edge. Building upon the capabilities of the automated, zero-touch core network, Fabric Extend enables extensibility over any third-party network. Fabric Attach enables edge devices, be they switches, wireless access points, hypervisors, cameras, or Internet-of-Things devices, to automatically attach to the fabric and connect to the desired virtual service network, thus removing any need for edge pre-provisioning, reducing operational cost, and increasing ease of use.

- Unleashing the power of the open ecosystem. Building upon the automated core and Fabric Attach-enabled edge, open SDN edge devices, such as any Open vSwitch-based device, can automatically and zero-touch attach to the desired virtual service network. Utilizing the open interfaces and open-source tools and controllers, such as Open Daylight, customers can quickly deploy SDN applications in an agile fashion and benefit from innovations from the community at large.

As with many things in life, a solid foundation is the beginning of great things to come. The Fabric Connect-based SDN architecture provides the industry's leading automated enterprise core, and with it, the foundation to SDN success.

Glossary

ACL	Access Control List
AF	Assured Forwarding (DSCP PHB)
AFI	Authority and Format Identifier (NSAP addresses)
AP	Access Point (WLAN)
ARP	Address Resolution Protocol (IPv4)
AS	Autonomous System
ASCII	American Standard Code for Information Interchange
ATM	Asynchronous Transfer Mode
BCB	Backbone Core Bridge
BEB	Backbone Edge Bridge
BFD	Bidirectional Forwarding Detection
BGP	Border Gateway Protocol (RFC 4271)
BMAC	Backbone MAC (IEEE 802.1ah Mac-in-Mac encapsulation)
BPDU	Bridge Protocol Data Unit (Spanning Tree)
B-TAG	Backbone VLAN tag (IEEE 802.1ah Mac-in-Mac encapsulation)
BUM	Broadcast, Unknown unicast and Multicast
BVID	Backbone VLAN identifier (IEEE 802.1ah Mac-in-Mac encapsulation)
BVLAN	Backbone VLAN (IEEE 802.1ah Mac-in-Mac encapsulation)
CCM	Continuity Check Message (IEEE 802.1ag)
CDP	Cisco Discovery Protocol
CE	Customer Edge
CFM	Connectivity Fault Management (IEEE 802.1ag)
CLI	Command Line Interface
CMAC	Customer MAC
CNLP	Connectionless Network Layer Protocol (ISO 8473)
CoS	Class of Service (QoS)
CPU	Central Processing Unit
CS	Class Selector (DSCP PHB)
CVLAN	Customer VLAN (UNI type)
DAD	Duplicate Address Detection (IPv6)
DCI	Data Center Interconnect
DEI	Discard Eligible Indication (PCP)
DF	Default Forwarding (DSCP PHB)
DHCP	Dynamic Host Configuration Protocol
DMLT	Distributed Multi-Link Trunk (MLT)

DMZ	Demilitarized Zone
DNA	Digital Network Architecture (Cisco)
DNN	Dynamic Nick-Name assignment
DNS	Domain Name System
DoS	Denial of Service (attack)
DSCP	Differentiated Services Code Point
DVR	Distributed Virtual Routing
EAP	Extensible Authentication Protocol
EAPoL	Extensible Authentication Protocol (EAP) over LAN
EAP-TLS	EAP Transport Layer Security
ECT	Equal Cost Tree (SPB)
eBGP	External BGP
ECMP	Equal Cost Multi-Path
EF	Expedite Forwarding (DSCP PHB)
EFA	Extreme Fabric Automation
EFO	Extreme Fabric Orchestrator (legacy Avaya Fabric Orchestrator)
E-LAN	Emulated LAN (MEF service set)
E-LINE	Emulated LINE (MEF service set)
EoMPLS	Ethernet over MPLS
ERS	Ethernet Routing Switch (Extreme stackable platform)
ESP	Ethernet Switched Path
ESX	VMware hypervisor
E-TREE	Emulated TREE (MEF service set)
EVPN	Ethernet VPN (RFC 7432)
EWC	Extreme Workflow Composer
ExtremeXOS	Extreme Networks switching platform
FA	Fabric Attach
FAN	Fabric Area Network
FC	Fabric Connect
FDB	Forwarding Database
FE	Fabric Extend
FEFI	Far End Fault Indication
FHS	First Hop Security (IPv6)
FIB	Forwarding Information Base
FSPF	Fabric Shortest Path First (Fibre Channel)
GARP	Gratuitous ARP
GbE	Gigabit Ethernet

GRE	Generic Routing Encapsulation
GRT	Global Routing Table (VRF-0)
HA	High Availability
HMAC	Hash-based Message Authentication Code
HTTP	Hypertext Transfer Protocol
HTTPS	Hypertext Transfer Protocol Secure
iBGP	Internal BGP
ICMP	Internet Control Message Protocol
I-DEI	Instance Drop Eligible Indictor (IEEE 802.1ah Mac-in-Mac encapsulation)
IEEE	Institute of Electrical and Electronics Engineers
IETF	Internet Engineering Task Force
IGMP	Internet Group Management Protocol (RFC 2236)
IGP	Interior Gateway Protocol
IoT	Internet of Things
IP	Internet Protocol
I-PCP	Instance Priority Code Point (IEEE 802.1ah Mac-in-Mac encapsulation)
IPTV	Internet Protocol Television
IPVPN	IP Virtual Private Network
I-SID	Backbone Service Instance Identifier; IEEE 802.1ah
IS-IS	Intermediate System to Intermediate System
ISO	International Organization for Standardization
IST	Inter Switch Trunk (Extreme Networks SMLT clustering)
IT	Information Technology
I-TAG	Instance Tag (IEEE 802.1ah Mac-in-Mac encapsulation)
L2 VSN	Layer 2 Virtual Services Network
L3 VSN	Layer 3 Virtual Services Network
LAA	Locally Administered (MAC) Address
LACP	Link Aggregation Control Protocol (IEEE 802.1AX)
LAG	Link Aggregation Group
LAN	Local Area Network
LBT	Load Based Teaming
LDP	Label Distribution Protocol (MPLS)
LISP	Locator Identifier Separation Protocol
LLDP	Link Layer Discovery Protocol (IEEE 802.1AB)
LSDB	Link State Data Base
LSP	Link State PDU (IS-IS)
LSP	Label Switched Path (MPLS)

MAC	Media Access Control
MLAG	Multi-chassis Link Aggregation Group
MD5	Message-Digest algorithm
MDT	Multicast Distribution Tree
MEF	Metro Ethernet Forum
MHMA	Multiple Host Multiple Authentication (IEEE 802.1X)
MHSA	Multiple Host Single Authentication (IEEE 802.1X)
MIM	MAC-in-MAC
MLT	Multi-Link Trunk
MLX	Extreme Networks routing platform
MP-BGP	Multi-Protocol BGP
MPLS	Multi-Protocol Label Switching
MSDP	Multicast Source Discovery Protocol (RFC 3618)
MSTP	Multiple Spanning Tree Protocol
MTU	Maximum Transmission Unit
MVPN	Multicast VPN (MPLS)
MVR	Multicast VLAN Registration
MVR-S	MVR Source VLAN
NAC	Network Access Control
NAT	Network Address Translation
NAT-T	NAT Traversal (for IPsec)
NEAP	Non-EAP (MAC based authentication)
NIC	Network Interface Card
NLB	(Microsoft) Network Load (Server) Balancing
NNI	Network – Network Interface
N-PE	Network Provider Edge (MPLS)
NSAP	Network Service Access Point address (ISO/IEC 8348)
NSX	VMware SDN network virtualization and security platform
OAM	Operation, Administration & Maintenance
ONA	Open Network Adapter
OSPF	Open Shortest Path First (RFC 2328)
OVS	Open vSwitch
OVSDB	Open vSwitch Database Management Protocol
PBB	Provider Backbone Bridging
PCI	Payment Card Industry
PCP	Priority Code Point (802.1Q p-bits)
PDU	Protocol Data Unit

PE	Provider Edge (MPLS)
PHB	Per Hop Behavior (DSCP)
PHP	Penultimate Hop Popping (MPLS)
PIM	Protocol Independent Multicast
PIM-SM	Protocol Independent Multicast Sparse Mode
PIM-SSM	Protocol Independent Multicast source specific mode
PKI	Public Key Infrastructure
PoE	Power over Ethernet (IEEE 802.3af)
PSK	Pre-Shared Key
PSU	Power Supply Unit
PVID	Port VLAN ID
QoS	Quality of Service
RA	Router Advertisement (IPv6)
RADIUS	Remote Authentication Dial-In User Service
RARP	Reverse ARP
RD	Route Distinguisher (MPLS)
RFC	Request For Comment
RFI	Remote Fault Indication
RIP	Routing Information Protocol
RIPE	Réseaux IP Européens
RP	Rendezvouz Point (PIM-SM)
RPFC	Reverse Path Forwarding Check
RSMLT	Routed SMLT; extension to Extreme's SMLT Switch clustering for router redundancy
RSTP	Rapid Spanning Tree Protocol
RSVP	Resource Reservation Protocol (MPLS)
RT	Route Target (MPLS)
RX	Receive
SDN	Software Defined Networking
SHA	Secure Hash Algorithm
SLA	Service Level Agreement
SLPP	Simple Loop Prevention Protocol
SLX	Extreme Networks Data Center platform
SMLT	Split MLT (VSP series Switch clustering a.k.a. Multi-Chassis Link Aggregation Group)
SNMP	Simple Network Management Protocol
SONMP	SynOptics Network Management Protocol (Extreme Networks topology discovery)
SPB	Shortest Path Bridging
SPBM	Shortest Path Bridging MAC (using IEEE 802.1ah Mac-in-Mac encapsulation)

SPBV	Shortest Path Bridging VID (uses IEEE 802.1ad Q-in-Q encapsulation)
SPF	Shortest Path First (IS-IS and OSPF Dijkstra's algorithm)
SPSourceID	Shortest Path Source Identifier (IEEE 802.1aq)
SSH	Secure Shell
SSID	Service Set Identifier (WLAN IEEE 802.11)
SSM	Source Specific Multicast
S-TAG	Service VLAN Tag
STP	Spanning Tree Protocol
TACACS	Terminal Access Controller Access Control System
TCG	Technical Configuration Guide
TCP	Transmission Control Protocol (IP)
TLS	Transparent LAN Services
TLV	Type Length Value (IS-IS)
ToR	Top of Rack
TRILL	Transparent Interconnection of Lots of Links (RFC 7176)
TTL	Time To Live
TX	Transmit
UDLD	Unidirectional Link Detection
UDP	User Datagram Protocol (IP)
UNI	User – Network Interface
VC	Virtual Circuit (EoMPLS)
VCS	Virtual Cluster Switching
VDX	Extreme Networks Data Center platform
VID	VLAN Identifier
vIST	Virtual IST (VSP SMLT Clustering)
VLACP	Virtual LACP; Extreme's variant of Unidirectional Link Detection (UDLD)
VLAN	Virtual LAN; IEEE 802.1Q
VM	Virtual Machine
VNI	VXLAN Network Identifier
VOSS	VSP Operating System Software
VPLS	Virtual Private LAN Service (MPLS)
VPN	Virtual Private Network
VRF	Virtual Routing and Forwarding Instance or VPN Routing and Forwarding Instance
VRRP	Virtual Router Redundancy Protocol (RFC 3768, RFC 5798)
VSI	Virtual Switch Instance (VPLS)
VSN	Virtual Services Networks (Extreme Networks Ethernet Fabric)
VSP	Virtual Services Platform (Extreme Networks SPB capable platform)

VTEP	VXLAN Tunnel Endpoint
VXLAN	Virtual Extensible LAN (RFC 7348)
WAN	Wide Area Network
WLAN	Wireless LAN
WRR	Weighted Round Robin
ZTC	Zero Touch Client (Fabric Attach)
ZTF	Zero Touch Fabric

Reference Documentation

Ref	Document Title	Publication / Date / Document	Description
[1]	802.1aq Shortest Path Bridging, Design and Evolution by David Allan & Nigel Bragg	Book 2012 ISBN 978-1-118-14866-2	Book explains both the “what” and “why” of SPB as a technology as well as how the final SPB came about
[2]	Switch Clustering Best Practices	TCG October 2014 NN48500-584	Best practices around deployment of SMLT clustering
[3]	Auto-attach using LLDP with IEEE 802.1aq SPBM networks	IETF Draft January 2015 draft-unbehagen-lldp-spb-00	Fabric Attach extensions to SPB explained
[4]	Media Access Control (MAC) Bridges and Virtual Bridge Local Area Networks Section 25: Support of the MAC Service by Provider Backbone Bridged Networks	IEEE Standard 31 August 2011 IEEE Std 802.1Q™-2011	IEEE Standard for Provider Backbone Bridging (MAC in MAC encapsulation)
[5]	Stealth Networking Technology & Design practices for Fabric Connect by Ed Koehler	TCG July 2014 NN48500-648	Technical Configuration guide which explores the Stealth properties of a Fabric Connect SPB network
[6]	Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks Amendment 20: Shortest Path Bridging	IEEE Standard 29 June 2012 IEEE Std 802.1aq™-2012	IEEE Standard for Shortest Path Bridging
[7]	IS-IS Extensions Supporting IEEE 802.1aq Shortest Path Bridging	IETF Standard April 2012 RFC 6329	Easier reading than IEEE standard gives a good overview of SPB and IS-IS extensions it makes use of
[8]	SPB Deployment Considerations	draft-lapuh-spb-deployment-03	Best practices when implementing IEEE 802.1aq Shortest Path Bridging (SPB) networks

Revisions

Rev #	Date	Authored/Revised	Remarks
01	19 February 2015	Content: Ludovico Stevens, John Vant Erve, Ed Koehler, Andrew Rufener Review: Goeran Friedl, Didier Ducarre	Published as Avaya external document
02	29 September 2018	Content: Ludovico Stevens, Ed Koehler, Didier Ducarre Review: Steve Emert, Gregory Deffenbaugh, Mikael Holmberg, Scott Fincher, Alexander Nonikov, John Hopkins, Roger Lapuh, Goeran Friedl, Stephane Grosjean	Extreme Networks rebranding. Addition of Data Center focused sections, DVR, Fabric Extend, VXLAN Gateway, PIM Gateway, advanced IP routing, re-write of many sections.
03	3 June 2019	Content: Ludovico Stevens, Didier Ducarre Review: Didier Ducarre, Luigi Simonetti, Johnny Hermansen, Goeran Friedl, Fukuo Miyamoto	Correction to DVR Scaling guidelines. Replaced EFO VPS references and replaced with VOSS/XMC VM Endpoint-tracking. New section covering SPB Equal Cost Tree (ECT); also added definition of SPB Bridge id relevant to ECT. Updates to Fabric Extend section to cover new XA1400