Contents

Figures

1. Network Definitions

Gigabit Ethernet – Ethernet media consisting of 802.3 standards with various speeds including 1 GbE, 10 GbE, 25 GbE, 40 GbE, 100 GbE, 200 GE, 400 GbE, etc. The standard covers a range of standards and protocols for twister copper wires, optical fibers, auto link speed negotiation, link-based full duplex flow-control, link aggregation, power over Ethernet, per flow control, enhanced transmission selection for bandwidth assignment based on priorities, data center bridging capabilities exchange protocol, etc. The subject is vast, and the author assumes basic and reasonable understanding of the topic.

L2 Switching – When Ethernet, Token-Ring, or Fiber Channel data frames are switched based on a media-based protocol frame carrying destination and source addresses. L2 networks consist of L2 bridges or switches, and they switch L2 frames based on destination address and use spanning tree to prevent a traffic storm. These can be physical LAN or Virtual LAN (VLAN) which follows the 802.1Q standard. The L2 frames can be unicast-, multicast-, broadcast-based traffic.

Spanning Tree – Built to avoid looping and storming of L2 frames. The Spanning-Tree Protocol and its variations run to establish a Spanning Tree.

L3 Routing – When IPv4 or IPv6 L3 packets are routed in a media-independent packet carrying destination and source addresses based on a route table lookup of L3 destination address in a packet. The packet forwarding follows time-to-live or hop-count tracking mechanism and uses class of service for QoS and queueing mechanisms. Security checks of unicast traffic and flood avoidance of multicast traffic happen using reverse path checks.

VLAN – Virtual Local Area Network is a technique of isolating LANs based on the 802.1Q standard with a 16-bit protocol identifier tag and a 16-bit control information tag, which is broken further into a 12-bit VLAN identifier, 3-bit 802.1p-based priority code point, and 1-bit 802.1Q-based drop eligible indicator (indicating congestion).

SVI – Switched Virtual Interface is where a VLAN termination on a physical port is configured with an L3 IP router interface and where the VLAN terminates. Here is where L2 switching terminates and L3 routing starts.

Traditional Network – It is a three-tiered tree-like organization of switches or routers, as shown in Figure 1, interconnected using L2 or L3 networks with three layers for access, aggregation, and core. The access layer fans out to many, while the core layer serves as the back-bone of the network, and the aggregation layer serves the hybrid purpose of fanning out as well as a back-bone, interconnecting the core and the access layers. This is how a traditional enterprise network looks.

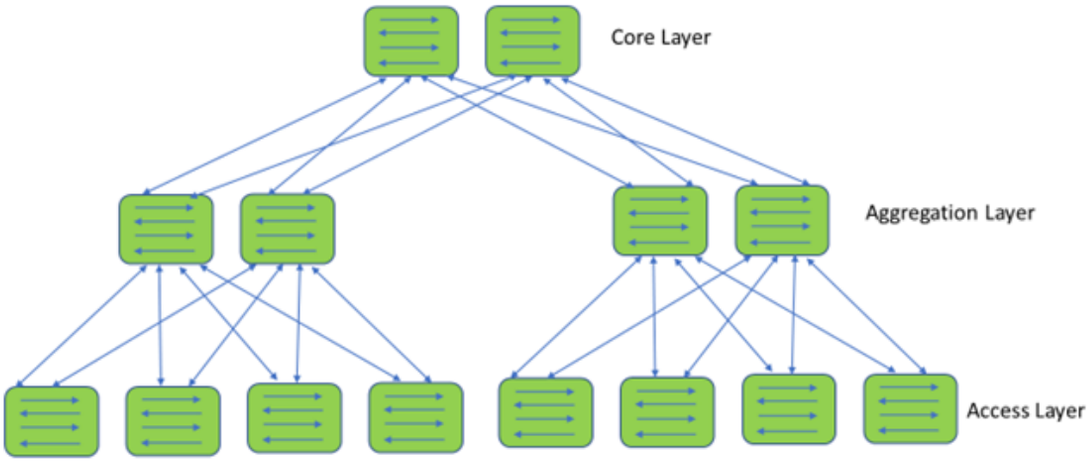

Figure 1: Traditional Enterprise Network

Leaf-and-Spine Network – It is a two-tiered tree-like structure, as shown in the Figure 2, with a leaf layer and a spine layer. One can think of the access and south-bound aggregation layers merging to become the leaf layer and the core and north-bound aggregation layers merging to become the spine layer. This is how a Data Center (DC) Network looks.

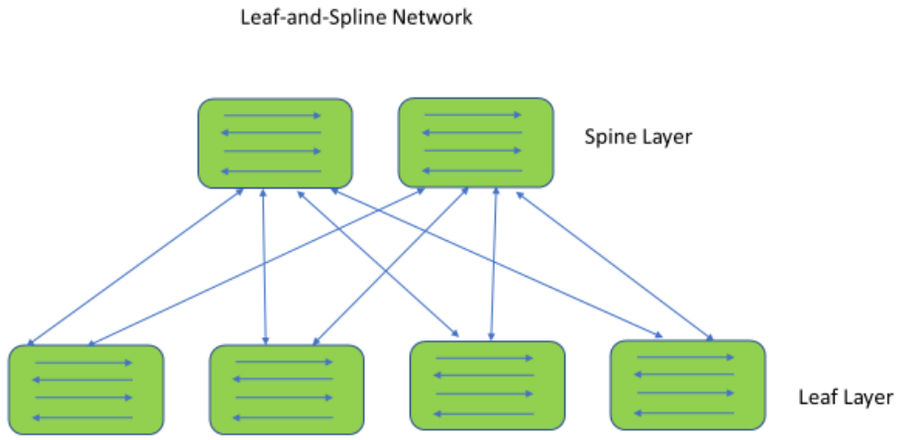

Figure 2: Spine and Leaf Data Center Network

Access Port – The host- or client/server-facing router/switch ports are called access ports. Also referred to as downstream ports.

Network Port – The network provider-facing router/switch ports are called network ports. Also referred to as upstream ports.

Routing Protocol – They are either link-state based like OSPF and ISIS or distance vector based for enterprise networks. To connect different network domains, distance vector based BGP protocol is used. BGP comes in two types, eBGP for the ISP space and iBGP to connect many domains in large enterprises. BGP has been extended to support other protocols besides IP, like Ethernet and MPLS addresses, and it is then called Multiprotocol BGP or MBGP. The link-state protocols have each network router broadcast to all discovered

neighboring-routers directly connected and the metrics describing the links connecting to them, e.g., speed, bandwidth delay, etc. Then each router runs the Dijkstra algorithm to build a network tree like graph with itself as a root and other nodes as network path-based leaves. For any network address it figures out the directly connected next-hop address to forward an L3 packet based on the network destination address. It then modifies the L2 destination address and forwards it on the L2, decreasing its TLL or hop counts. It also applies QoS-related attributes. Distance vector protocols tell only directly connected neighbors about their view of distance-hops or vectors to every network router. It suppresses forwarding on the link from where they have received the routing vector packet update, which is known as split horizon. Each router runs Bellmen-Ford algorithm on accumulated vectors.

Unicast – Point to point L3 packet or L2 frame flow of traffic between sender and receiver endpoints. The sender and receiver addresses are unicast addresses.

Multicast – Point to multi-point L3 packet or L2 frame flow of traffic between a sender and multiple receivers. The sender has a unicast address and sends packets or frames to a multicast address. The receivers join this multicast group address through a subscribe/publish protocol like IGMP for IPv4 and MLD for IPV6 on LAN sub-nets. The routers in WAN run multicast protocols to form distribution trees, which can be sender-based source rooted with the rest of the receiving routers as leaves, like in later part of PIM-sparse, or flood and prune based like in PIM dense mode, or sending to a rendezvous address over tunnel where the rendezvous address is rooted with the rest of recovers as leaves, like the earlier part of PIM sparse mode, or following the reverse split horizon and forwarding to all receiving routers except where the packet/frame is coming from.

Anycast – When many endpoints have the same address and anyone can serve a request, then it is called anycast. There is a range of IPv4 and IPv6 addresses with the same name but not used as such. This concept is used in data centers for scaling of web traffic with layer 4 or layer 7 based server load balancers which act as a proxy for many hundreds of servers. The client sends a request to a virtual IP address which goes to the SLB device, and the SLB device sends the packet to a server that is less loaded and ready to serve the request. The concept is also used in the Static Anycast Gateway (SAG) to support MC-LAG Active-Active.

Broadcast – All “FFs” in an L2 frame or L3 packet destination address is a broadcast address. It means that all networking endpoints including the sender will the frames or packets being sent to that address. By default, L2 frames are flooded in a spanning tree but L3 forwarded by routers.

BUM – Broadcast, Unknown Unicast, and Multicast traffic is called BUM traffic. When an L2 switch does not have an L2 destination address in its forwarding database (L2 FDB), meaning it has not learned the forwarding path towards the L2 destination. In such case the traffic is sent as BUM traffic as if it was a broadcast traffic to all members of a given VLAN or subnet without VLANs over a spanning tree.

Tunneling – Every L3 IP packet is encapsulated with an L2 Ethernet frame. L2 switching happens between directly connected switches. The direct connection is a physical connection over some media. L3 connected IP routers need not be directly connected. These routers can be connected over a layer 2 network of switches. The packets between routers will be routed and frames over an L2 network will be switched. These are just the terminologies used. One can say that the L3 IP packet is in a way tunneled inside an L2 Ethernet frame. But the word “tunnel” is dropped and given standard for IP networking. Similarly, any L2 frame or L3 packet can be

tunneled over any L3 packet, like IPv4 in IPv6, IPv6 in IPv4, MPLS over IPv4 or IPv6, Ethernet over MPLS, VLAN in VXLAN over UDP in IPv4 or IPv6.

Underlay Network – The network that is tunneling the L3 packets or L2 frames and connecting and bridging as L2 or L3 network transport for many remote network domains. The outer L3 packet represents this underlay network. This can be considered as the backbone fabric of the network.

Overlay Network – The network that exists on the edges and is the client to the underlay network. The L2 frames or L3 packets are part of the overlay network and are tunneled using the underlay network.

DC – A data center provides units of computing, networking, and storage resources, usually using virtual hypervisor, virtual networking, and virtual storage technologies.

EVPN – Ethernet Virtual Private Network leverages BGP to share different kinds of L2 and L3 addresses between its peers. BGP routers are interconnected as a mesh or through reflectors and exchange routes learned locally with the remote BGP peers. The L2 frames are tunneled in L3 IPv4/v6 Packet using VXLAN format over UDP. It is a BGP based control plane for Layer 2 bridging and Layer 3 routing VPNs for multi-tenant environments. It unifies L2 and L3 networks with an equivalent of a traditional L3-only MPLS based VPN control plane.

LAG – Link Aggregation Group is basically a port-channel as a collection of many physical ports on each switch or router, interconnected with parallel Gigabit Ethernet cables, acting as one for redundancy, load-balancing, and bandwidth scaling. They run the Link Aggregation Control Protocol (LACP) to give the appearance of one virtual logical port on each end. Here, the assumption is that each end router or switch is a single chassis-based system running one NOS or multi chassis-based system running only one master NOS along with many slave NOSes in many line cards to scale the number of ports offered for large scale routers and switches.

MC-LAG – It is the evolution of LAG for multi-chassis with two independent Master NOSes based switches/routers where Inter-Chassis Control Protocol (ICCP) is run to give the impression to the remote downstream host end, with its NIC, that it is acting as a one giant Virtual Logical Port but with High Availability because of two local upstream chassis acting as one. It means the upstream MC-LAG pairs must keep sending keep-alive messages to each other, and share L2 forwarding tables, forward BUM traffic to each other as well as forward ARP packets to the MC-LAG peer for learning. And the peer-end has to suppress forwarding them to the remote downstream LAG-enabled host. This is done by post isolation group feature where the downstream connected port is traffic isolated. Only when one upstream MC-LAG end discovers the other upstream peer end's LAG link to its remote downstream host is down, then it starts forwarding BUM, ARP traffic and its forwarded packets to the remote host. Each MC-LAG peer upstream node is responsible for forwarding packets to its remote downstream host when packets/frames are received on non MC-LAG ports and vice-versa to non-MC-LAG ports from its remote downstream host.

FRR – FR Routing is a free and open-source Internet routing protocol suite for Linux environments, which implements BGP, OSPF, RIP, IS-IS, PIM, LDP, BFD, Babel, PBR, Open Fabric, VRRP, EIGRP, and NHRP.

MPLS – It is a packet-forwarding technique that uses 20-bit labels to make data forwarding decisions. With MPLS, the layer 3 header analysis occurs just once, when the packet enters the MPLS domain and transitions

from L2, ATM, Optical or IP network to the MPLS network. Control protocols like LDP, RSVP, and MPLS map an Ethernet, ATM, Optical, IPv4 or Ipv6 destination to an MPLS label. An MPLS network consists of Physical Edge (PE) routers, which are edge border routers, and Provider (P) routers, which are core routers. The label inspection drives subsequent packet forwarding in the MPLS network when transiting through Provider (P) routers. MPLS is used in Virtual Private Networking (VPN), Traffic Engineering (TE), Quality of Service (QoS) and Any Transport over MPLS (AToM), such as Ethernet Over MPLS (EoMPLS) for point-to-point PE transport over WAN and Virtual Private Network (VPLS) for multi-point PE routers. It decreases the forwarding overhead of core routers by using a simplified MPLS label that is a priori mapped to an L2 IPv4 or IPv6 destination without repeating the complex IPv4 or IPv6 header processing at every router, for example.

Please note that MPLS is not supported in Edgecore SONiC but is explained here for awareness and completeness.

EVPN instance (EVI) – An EVPN Instance (EVI) is an EVPN routing and forwarding instance spanning all the PE routers participating in that VPN. It is represented by the Virtual Network Identifier (VNI).

Ethernet Segment (ES) – An Ethernet Segment is a set of Ethernet links that connects a multi-homed device.

Ethernet Segment Identifier (ESI) – Ethernet segments are assigned a unique non-zero identifier. ESI consist of 1 byte of ESI Type and 9 bytes of ESI value. E.g., 00 00 00 00 00 00 10 00 00 05.

Designated Forwarder (DF) – Only a single router is allowed to decapsulate and forward the traffic for a given Ethernet segment in active mode. For All-Active or Active-Active mode, DF is responsible for forwarding BUM packets to CE devices.

DF Election – DF selection process using MBGP Type-4 route. It is used to prevent forwarding of the loops. For Single-Active or Active-Standby Mode, in a VLAN-based DF Election, multiple PE will be load-balanced among each VLANs to become primary DF.

Split Horizon – The way to preventing a BUM packet echoing back to the same Ethernet segment.

Redundancy Mode – PE devices participated in multi-homing should configured to work in either Active-Active/All-Active or Active-Standby/Single-Active mode. This will affect the Load Balancing behavior for Unicast packets sending from remote PEs.

Aliasing – PE devices use M-BGP Route Type 1, known as Ethernet Auto Discovery (EAD) or Ethernet segment (EAD/EVI/ES) route to signal that it has reachability to an EVPN instance on a given Ethernet segment.

Mac Mass Withdraw – It is used for fast convergence during access failure scenarios using the M-BGP Route Type 1 Ethernet Auto Discovery route - for Ethernet-Segment (EAD/EVI/ES) route.

BGP/EVPN Route-Types – The routing information can be L2 MAC address or an L3 destination masked based address. On EVPN, a route type is used to classify the various kinds of routing information. The minimum required route types are RT-2, RT-3, and RT-5. For multi-homing, Edgecore also supports RT-1 and RT-4. The five route types are described in the following table:

Route Type	Information Carried	Description
RT-1	Ethernet Auto-Discovery (AD) Route	It is used for fast convergence and aliasing. EVPN fast convergence allows PE devices to change the next-hop adjacencies for all MAC addresses associated with a particular Ethernet segment. EVPN aliasing allows traffic to be balanced across multiple egress points. Few routes are sent per ES. This route carries the list of EVIs that belong to the ES. Used in fast route convergence, redundancy mode, aliasing, and split horizon.
RT-2	MAC/IP Advertisement Route	Advertise MAC, address reachability, advertise IP/MAC binding – It is used to exchange end hosts' IP and MAC addresses advertised within NLRI.
RT-3	Inclusive Multicast Route	Multicast tunnel end point discovery – It is used to forward Broadcast, Unknown Unicast and Multicast (BUM) traffic across EVPN networks.
RT-4	Ethernet Segment Route	Redundancy group discovery, DF election – It is needed in multi-homing scenarios and used for designated forwarder election to send Broadcast, Unknown Multicast and Multicast (BUM) traffic to a CE/Leaf on a particular Ethernet segment.
RT-5	IP Prefix Route	Advertise IP prefixes – It is used to carry IPv4 and IPv6 advertisements in EVPN-only networks. For EVPN Type 2 routes carrying MAC and IP addresses, tight coupling of specific IP addresses and IP prefixes might not be desirable.

Edgecore Networks Corporation
No. 1, Creation Rd. III
Hsinchu Science Park
Hsinchu 300
Taiwan, R.O.C.
www.edge-core.com