High Performance VxLAN Network Design

Tech Con 2025 Abstract 444

HPE Edge;

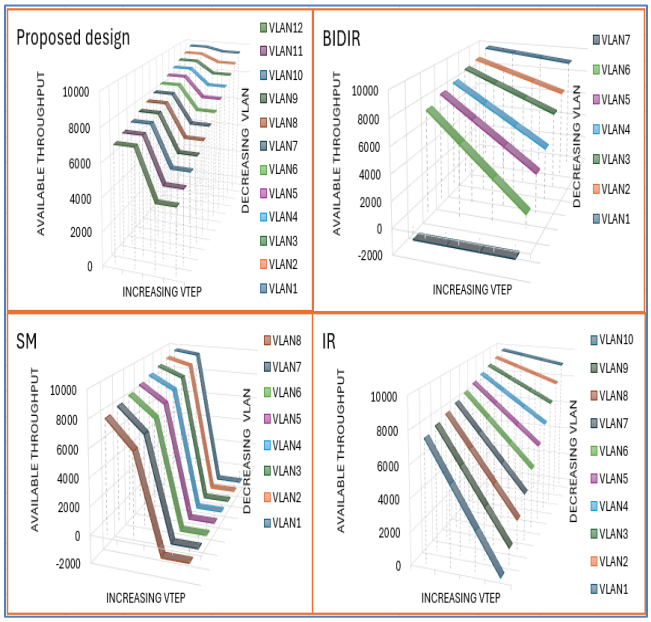
High Performance VxLAN Network Design

Abstract

High Performance Networks aims to provide a) high throughput b) low latency c) high scale for workloads in the network. This paper focuses on achieving high performance for the workloads running in the Overlay Networks. With the increasing adoption of AI/ML/Cloud/IoT/5G/Financial & Media Services etc the workload today is multicast heavy. Legacy networks are optimized (using ECMP load balance) for point-to-point distribution and with the increase in the volume and the replication fanout of Overlay workloads the point-to-point centric network designs perform poorly. The industry standard approach is to move the replication overhead into the underlay network and with increased endpoint scale the design performs poorly. In this paper we propose High performance VxLAN[1] Network Design which is built on the fundamentals of multipoint distribution, virtualization and scale-out architecture to achieve a) High throughput and b) Lower Latency by distributing the traffic into point-to-multipoint (single source multi receiver) and multipoint-to-multipoint (bi-directional) trees on virtualized interfaces called sub-interface [2] c) Higher Scale by offloading the tunnel encapsulation fanout to the large replication ASIC table and by virtualizing the termination such that it can fold multiple sources into a single termination entry there by achieving very high scale fan-in(s) in the Overlay Networks for tunnel termination. The high performance VxLAN Network Design is fundamental for the emerging Multicast deployments in Virtual Private Cloud, Large Enterprise, and High-Performance Computing over Cloud for multidimensional dataset.

Problem statement

Gartner Forecasts Worldwide Public Cloud End-User Spending to reach $679 Billion in 2024 [3]. Multiple Enterprises have started to migrate to “Cloud” and Gartner predicts that by 2028 it would become a business necessity [4]. With GenAI services, multicast in cloud has already become a reality [5]. VxLAN is the standards needed to build Datacenter Cloud (Public and Private). Traffic like Broadcast(B), Unknown Unicast(U), Unknown Multicast(M), IP Directed Broadcast (IPDB [6]) require multipoint-to-multipoint distribution and Known Multicast traffic require point-to-multipoint distribution (typically overlay pim-sm). The industry standard way to achieve multipoint distribution is by Ingress Replication (IR) [7] and Underlay Multicast Replication [8].



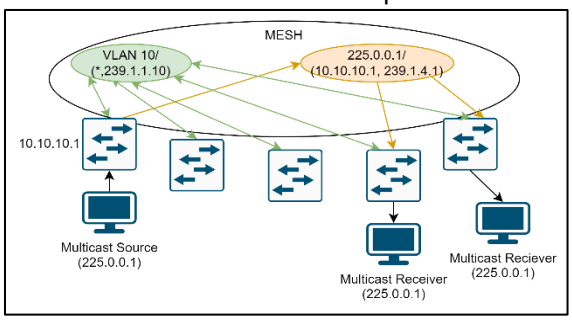
In IR mode the distribution is achieved via point-to-point tunnels by source vtep generating copies per remote endpoint and with increase in the remote endpoint scale the network performs poorly. In Underlay Multicast mode distribution is achieved by multicast tree. Most or all the implementations today build point-to-multipoint tree via pim-sm [9] (or) multipoint-to-multipoint tree via pim-bidir [9] but not both. With pim-bidir, the tree is rooted on fixed node called “rendezvous point (rp)” and thus cannot scale and require an ASIC roll-out to increase the link speed. With pim-sm, the tree is rooted initially on a fixed rp called root-path-tree (RPT) and then shifts to dynamic nodes nearest to the receiver called shortest-path-tree (SPT). pim-sm performs well for high bandwidth workloads, however poorly when the endpoint increases as the source endpoint has fixed tunnel encapsulation which restricts the fanout and thereby cannot meet the high-performance requirements. The left figure shows the 3d simulated graph of “available throughput”, “vteps” and “vlans”. The simulation is done on a grid with five-vteps, two-spines, user workload is 250Mbps, uplink bandwidth is 1G. Thus, the total bandwidth is 10G (5x2x1G). The simulation assumes fixed two-tunnel-encap and two-tunnel-decap resources for multipoint. In IR, the bandwidth geometrically drops with increase in vtep and vlan scale. In BIDIR, the link gets oversubscribed faster with increase in vtep and vlan scale. In pim-sm, the system tunnel resources are exhausted with increased vtep scale and requires asic roll-out to increase the tunnel capacity not ideal in cloud. We are creatively implementing novel multicast underlay capabilities on *existing platforms with current ASICs*. This paper proposes *software design*, *asic dataplane design* and *scale-out principles* which can be deployed as a new feature on *existing products without rolling out new asic*. Multicast services in the underlay

network can be spined based on the overlay demands without impacting the on-going user traffic. With improved multicast underlay capabilities, the overall network scalability and performance gets enhanced thereby making it a high-performance network design**.**

Our solution

The paper proposes critical software, asic design elements along with scale-out principles to improve network performance. The following are the summarized elements: **Software Design(s)** - (1) *Traffic Distribution*: The overlay traffic is classified and transmitted on different set of distribution trees (2) *Prefix Group Design*: The underlay multicast address pool follows ip-prefix match design (3) *Virtualized PIM Interfaces*: Grid uplinks are converted to Sub-Interfaces to support dual pim protocol. **ASIC design(s)** - (1) *Improved Tunnel Encapsulation fan-out*: Offloading the replication load within the asic from Tunnel RAM to Replication RAM (2) *Improved Tunnel Termination fan-in*: Folds multiple remote vteps into single virtualized tunnel termination entry.

Traffic Distribution: The multipoint-to-multipoint traffic is typically a flood-and-learn workload which is short-lived and low bandwidth. The design builds a bidirectional tree via pim-bidir between the various endpoints from the configured Virtual Network ID (represents a VxLAN Broadcast Domain). For instance, consider a VNI ID 10; the design builds and generates an auto mapped multicast group 239.1.1.10 and redirects flood, ipdb traffic on VLAN10 to bidirectional tunnel associated with 239.1.1.10. The remote nodes can further filter the traffic based on the configuration and the client(s) learnt. Each traffic class - flood, ipdb have redirection asic controls.



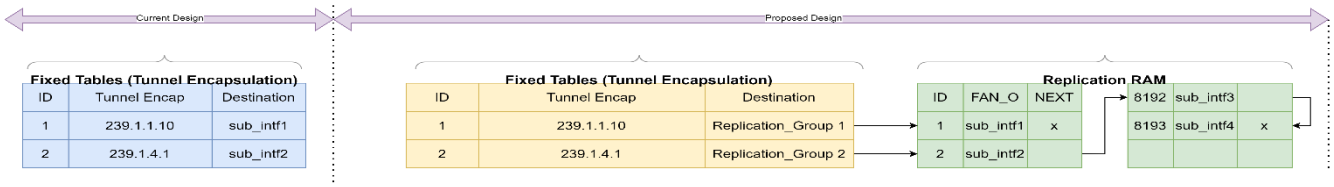
The point-to-multipoint traffic is typically long lived and has high bandwidth like CCTV/IPTV/Live Feeds etc. The design builds a unidirectional distribution tree from the overlay traffic's address. For instance, consider overlay flow is 225.0.0.1; the design builds and generates an auto mapped pim-sm distribution tree 239.1.4.1. The asic redirects the overlay traffic associated with 225.0.0.1 to underlay tree associated with 239.1.4.1. Above Figure depicts the bidir (*,239.1.1.10) and sm (10.10.10.1, 239.1.4.1) trees. As the scale increases/decreases*, the grid can spin the multicast services accordingly to cater to the scale without impacting the existing clients. Thus, the design follows a scale-out/scale-in architecture*. Multiple VNIs can map to a common address, likewise multiple overlay flows can map common address, however the pim-sm and pim-bidir addresses are non-overlapping and enforced by config. Refer Prefix Group Design for more details.

Prefix Group Design: The design proposes configurable multicast group ranges to facilitate prefix-based match criteria. This is an important design element to burn very few ACLs, Policy TCAM and RP TCAM rules for Security, QoS, Optimization, respectively. The design automatically picks the underlay group ip based on a pre-determined algorithm fixed in all the endpoints. In the above figure, the pim-bidir pool is 239.1.1.0/24 (256 tunnels inclusive of 239.1.1.0 and 239.1.1.255) and pim-sm pool is 239.1.4.0/22 (1024 tunnels). MSB (most significant bits) masking is the algorithm used here. Dynamic algorithms can be built based on the traffic pattern and communicated via EVPN message(s) to all the endpoints. The common algorithm design is flexible and is out of scope in this paper.

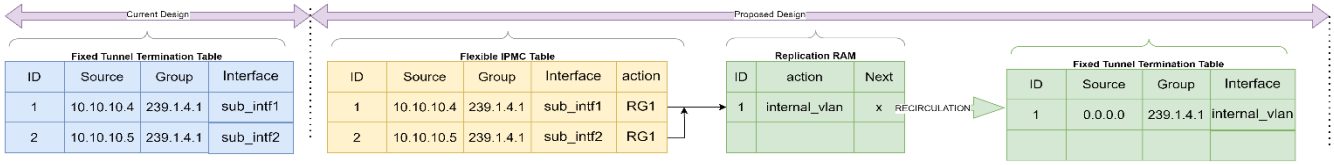
Virtualized PIM Interfaces: Grid network is typically hard wired in mesh form in cloud data center racks, and it is impossible to dynamically re-wire/re-configure the uplink interface(s) for pim-bidir or/and pim-sm. A simple industry standard, yet powerful virtualization technique called "Router on Stick" aka "Sub Interfaces" [10] is proposed in this design. This is an important element as it virtualizes the physical link into multiple virtual interfaces which can be used for pim-sm and pim-bidir isolation. Note: PIM-SM quickly shifts to "SPT" path thereby not over-subscribing the pim-bidir RPT links while sharing the same physical link temporarily.

Improved Tunnel Encapsulation Fan Out: Currently Aruba ASICs contains – 1) Large Replication table (64K entries) that is used to replicate packets to multiple interfaces and 2) Fixed Tunnel Ram (4K entries) for tunnel encapsulation and fanout. When a new endpoint is spined, the multicast services on the connected spine is also spined and the source vtep is required to generate additional copy of the multipoint traffic towards the spined spine (also the shortest-path spine). This is achieved by burning additional tunnel entry with every fan-out increase. The tunnel encapsulation fanout is improved by gluing the replication entires to the tunnel entries thereby significantly improving the fanout capabilities. *Thus, with increase and decrease in the endpoint scale, the replication entries are added and removed without impacting the existing clients and follows a scale-out/scale-in design*. With pim-bidir pool of 239.1.1.0/24 (256 tunnels) and pim-sm pool of 239.1.4.0/22 (1024 tunnels), the fanout that can be achieved is just 4 = floor (4096/1024). With the improved design the fanout that can be achieved is 64 = floor (65536/

(1024). Thus, this is a very good scale for the proposed VxLAN solution (16x improvement). Further the fanout can be easily improved by the fact that most if not all the pim-sm distribution tree would typically have the same list of the fan-outs/ spines to replicate to. By coalescing replication entries into a replication group and gluing them with multiple tunnel entries we can achieve very high fanout (>256x for large packets). Figure below demonstrates the fanout offloading from tunnel ram to replication ram. Note: pim-bidir is a single fan-out protocol design (always rp tree). There can be a large fanout implementation for pim-bidir which can provide fast failover. However, it is out of scope in this paper.



Improved Tunnel Termination Fan In: Currently Aruba ASICs contains – 1) Large Dynamically Resizable Multicast Lookup table (48K dynamic HASH entries) and 2) Fixed Tunnel Termination Lookup table (4K entries). As the cloud igmpv3 [10] user's subscription increases, the client vtep joins the source specific underlay distribution tree which typically is on a different spine interface(s). The client vtep (connected to the cloud user), now receives multicast vxlan packet on a different interface and to terminate the packet additional tunnel termination entry is required to be burnt. By protocol design pim-bidir is single fan-in as it always receives the packet from the fixed rp-tree, however pim-sm will have multiple fan-in requirements. The tunnel termination fan-in for pim-sm is improved by three mechanisms - (1) Moving the (interface, source, multicast-group) aka (ISG) lookup from fixed tunnel lookup to dynamic multicast lookup (2) Redirecting the (ISG) lookup hit traffic to boot time created interface called "internal_vlan" (3) Building a (internal_vlan, 0.0.0.0, G) Tunnel Termination, where 0.0.0.0 represent wild card source lookup effectively folding all the sources into single termination. Thus, with the increase and decrease of the multicast receivers, the dynamic multicast lookup table can be updated keeping the tunnel termination fixed, effectively creating a scale-out/scale-in design without affecting on-going receivers. For pim-sm pool of 239.1.4.0/22 (1024 tunnels), the fan-in that can be achieved is just 3 = floor ((4096-256)/1024). In the improved design, the fan-in that can be achieved is 48 = floor(48K/1024). Thus, this is a particularly good scale for the proposed VxLAN solution (16x improvement). The figure below depicts the improved fan-in capability.



Evidence the solution work

Software constructs like pim-bidir, sub-interfaces, prefix-based policy are released features in Aruba CX. We have a halon design in-place with pim-bidir as underlay in 10.14. We are currently designing the pim-sm solution. Dynamically increasing and decreasing the multicast table size is prototype ready.

Competitive approaches

Competitors like Cisco achieves the high scale by reducing the underlay multicast tree, thereby decreasing the network performance because all the overlay traffic gets distributed into single underlay tree causing huge packets in-discard on the remote vteps [11]

Current status

The asic dataplane architecture review is completed and we know how to configure the ASIC to achieve improved fan-out and fan-in. It is in the plan to do some proof-of-concept experiments to prove out the path more fully.

Next steps

This is committed for a future Halon release, and it is in the plan to implement some, or all the improvements proposed in this paper into the Aruba CX releases.

References

[1] – VxLAN - https://en.wikipedia.org/wiki/Virtual_Extensible_LAN

[2] – Sub Interfaces - https://networklessons.com/cisco/ccna-routing-switching-icnd1-100-105/how-to-configure-router-on-a-stick

[3] - https://www.gartner.com/en/newsroom/press-releases/11-13-2023-gartner-forecasts-worldwide-public-cloud-end-user-spending-to-reach-679-billion-in-20240

[4] https://www.gartner.com/en/newsroom/press-releases/2023-11-29-gartner-says-cloud-will-become-a-business-necessity-by-2028

[5] - https://docs.aws.amazon.com/vpc/latest/tgw/tgw-multicast-overview.html

[6] – IPDB https://www.arubanetworks.com/techdocs/AOS-CX/10.14/HTML/ip_route_6300-6400-8100-83xx-9300-10000/Content/Chp_IPDirBroad/ip-dir-bro-cnf-exa53.htm

[7] - https://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst9400/software/release/16-11/configuration_guide/lyr2/b_1611_lyr2_9400_cg/configuring_vxlan_evpn_ingress_replication.html

[8] - https://www.cisco.com/c/en/us/td/docs/switches/lan/catalyst9300/software/release/17-8/configuration_guide/vxlan/b_178_bgp_evpn_vxlan_9300_cg/configuring_optimized_l2_overlay_mutlicast.html#l2-multicast-replication-types

[9] – PIM-SM/PIM-BIDIR - https://en.wikipedia.org/wiki/Protocol_Independent_Multicast

[10] – IGMPv3 - https://en.wikipedia.org/wiki/Internet_Group_Management_Protocol

[11] - https://deliabtech.com/blogs/underlay-multicast-routing-for-vxlan-bum-traffic/