August 2023

GROUP-BASED POLICIES FOR MICRO-SEGMENTATION WITH STANDARD VIRTUAL EXTENSIBLE LOCAL AREA NETWORK (VXLAN) OVERLAY

Radha Krishnaiah Pusapati

Rajagopalan Janakiraman

Muru Panchalingam

Muralidhar Annabatula

Ayan Banerjee

GROUP-BASED POLICIES FOR MICRO-SEGMENTATION WITH STANDARD VIRTUAL EXTENSIBLE LOCAL AREA NETWORK (VXLAN) OVERLAY

AUTHORS:
Radha Krishnaiah Pusapati
Rajagopalan Janakiraman
Muru Panchalingam
Muralidhar Annabatula
Ayan Banerjee

ABSTRACT

Standard Virtual Extensible Local Area Network (VXLAN) headers do not provide the space to carry policy information from ingress Network Virtual Ethernet (NVE) devices to egress NVE devices. The option to provide the bits/bytes to carry this information is provided with different header formats, however, some bases are unable to parse these other frame formats. According to techniques described herein, group-based policies may be implemented over standard VXLAN overlays, thereby eliminating the cost of implementing custom protocol modifications and helping data center customers by remaining with their existing fabrics without costly upgrades.

DETAILED DESCRIPTION

The standard Virtual Extensible Local Area Network (VXLAN) header provides Virtual Network Identifiers (a 24-bit VNI) to designate the VXLAN overlay network for multi-tenancy. In standard VXLAN overlay fabrics, Access Control Lists (ACLs) are used to apply policies (e.g., allow, deny, audit, redirect, etc.) in ingress Network Virtual Ethernet (NVE) devices. The standard VXLAN header does not provide the bits/bytes to carry the policy information from the ingress NVE device to the egress NVE device. The option to provide the bits/bytes to carry this information is provided with a different header format, such as the VXLAN-Group-based Policy Option (GPO) or VXLAN-General Protocol Extension (GPE). However, some bases are unable to parse these new frame formats. It is important to support the policy framework with such hardware.

Existing policy schemes use traffic direction (e.g., network interface vs. core interface) to distinguish the policy enforcement in NVE devices. Due to the virtual machine (VM) mobility in data center fabrics, a given NVE device works as an ingress

NVE device and an egress NVE device in the fabric. This imposes a need for more hardware entries in the NVE device to distinguish traffic directions. The existing Access Control List (ACL)-based methods use endpoint identity information (like Media Access Control (MAC), Internet Protocol (IP) address/prefix, Interface/virtual LAN (VLAN)) in the hardware rules to apply the policy. This again imposes a need for more hardware entries in the NVE device for multi-fabric deployments. Because the egress NVE device is unaware of the ingress NVE device policy information, the policy redirection with service appliances can cause packet loops in multi-tier (e.g., leaf tier, spine tier, super-spine tier) fabrics and multi-domain fabric deployments.

As discussed above, viable options to support scalable policy actions for a variety of scenarios with the standard header is limited or infeasible with the standard header. Techniques discussed herein provide viable options for various customer use cases at scale.

Traditional enterprise data center customers using switch manufacturers generally prefer to use standard-based deployments like VXLAN overlays and BGP EVPN control planes in their fabrics. The customers typically avoid any vendor-specific protocol modifications for fear of lock-in and violating their dual vendor strategy. With the advent of hardware based micro-segmentation, some vendors have started building switches that natively support micro-segmentation through ‘Group Based Policies’ to increase the security in their networks. Standardization of these VXLAN protocol modifications, such VXLAN-GPO and VXLAN-GPE, are ongoing and there have been implementations on other overlay protocols (e.g., Generic Network Virtualization Encapsulation (GENEVE) and Segment Routing IPv6 (SRv6)). Many enterprise data center customers would like to use the micro-segmentation feature but would like the feature to be compatible with their existing equipment.

Micro-segmentation Group Based Policies

In micro-segmentation, instead of using the traditional ACLs to enforce security, the source and destination endpoints that need to communicate are classified into security groups (Endpoint Security Groups (ESGs)) by using network identifiers in the packet (e.g., incoming interface, VLAN, MAC addresses, IP addresses, etc.). The security policies are then applied between the source and destination tags (source group tag (SGT) and

destination group tag (DGT), respectively). The source and destination endpoints are first classified into SGTs and DGTs and then a policy lookup is performed to enforce the matched policy.

In data center networks, the source and destination endpoints are behind different network switches due to east-west packet communication and VM mobility. The group policy identifiers (IDs) are available in network switches where endpoints are locally connected and learned. For a given data packet, the ingress network switch has a source group policy ID (SGT) and the egress switch has a destination group policy ID (DGT).

To achieve ESG-based end-to-end policy enforcement in data center fabrics, Policy Applied (PA) information is required for service chaining requirements and a Source Group Policy ID (SGT) is needed at the egress network device for egress policy enforcement. The focus of VXLAN enhancements like VXLAN-GPO and GPE differ in the details used to carry SGT and PA information in the data packet. To natively support this functionality using standard VXLAN as overlay, the problem can then be broken down to finding out how these two pieces of information can be supported in the existing frame format.

The PA status provides for an ability for the ingress switch to signal the egress switch in the data packet, regardless of whether the ingress switch already applied the policy. The PA status is needed for the egress switch to not apply (i.e., bypass) any service redirection rules. The SGT provides for an ability to carry the SGT to the egress switch and is applicable to enforce policies in the egress switch for flood and learn topologies. It is possible to support only one of the PA status and the SGT and thereby support a limited set of use cases.

Intelligent Overloading of Policy Applied in VXLAN Overlay Outer Source IP Field

The "Policy Applied" status being "Yes" or "No" needs to be passed via some existing field in the overlay VXLAN header. The Overlay IP Header's Source IP Address field may be chosen for this purpose. When the ingress switch needs to pass "Policy Applied = Yes" status to rest of the fabric, the ingress switch uses a special/reserved IP address known as "PA-SIP" and it is configured as a well-known tunnel source address in

the rest of the fabric. Any egress switch receiving the packet with tunnel source as "PA-SIP" will transmit "Policy Applied = yes" in its pipeline.

Figure 1, below, illustrates an example in which an egress switch drives Policy Applied = yes processing based on receiving a packet with a tunnel source PA-SIP to bypass performing a policy lookup.

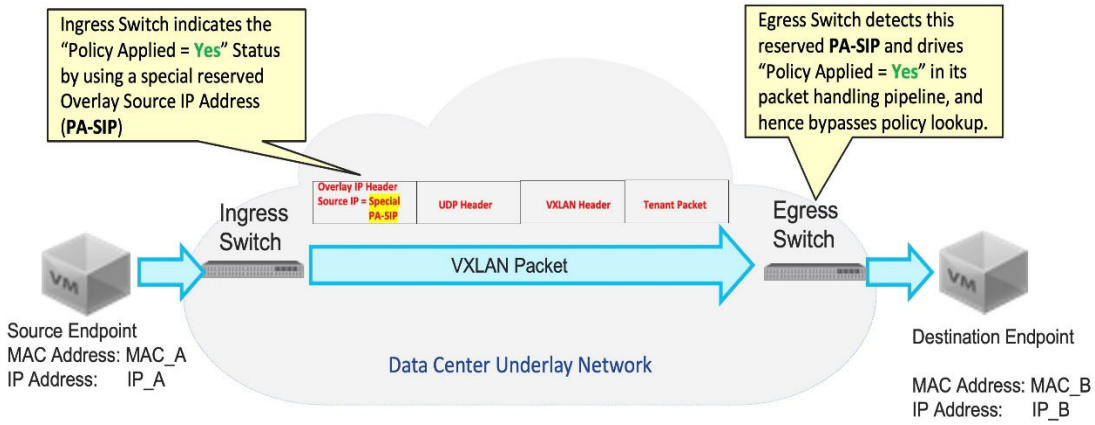

Figure 1: Example in which the Policy Applied = Yes

Figure 2, below, illustrates the handling of "Policy Applied = No" in which the ingress switch uses a local Virtual Tunnel Endpoint (VTEP) as an overlay source IP address and the egress switch defaults to "Policy Applied = No" when the local VTEP is the tunnel source.

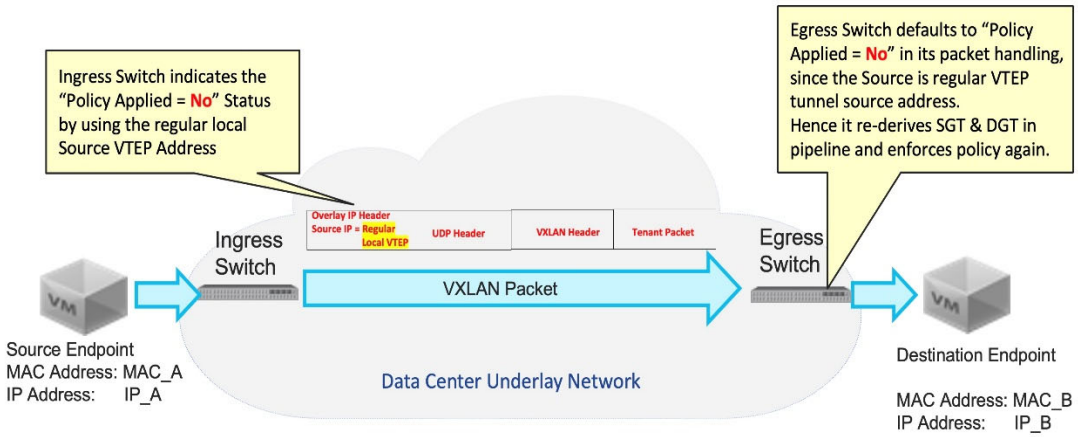

Figure 2: Example in which the Policy Applied = No

The endpoints to the group policy ID information are present in network switches software through user configuration. When locally connected endpoints are learned by network switches, the respective group policy ID is derived and installed in the hardware along with the endpoints. In Border Gateway Protocol (BGP) Ethernet Virtual Private Network (EVPN) control protocol fabrics, the endpoints are distributed in the fabric and imported to respective network switch-based tenant deployments. By advertising the group policy ID in the BGP EVPN protocol, the ingress network switch learns endpoints and respective group policy IDs.

When the ingress network device is aware of the destination endpoint and its policy ID, the ingress network device applies the policy (Ingress Policy). The ingress network switch keeps the policy applied information (PA-SIP) in the VXLAN outer source IP field in the data packet. Based on the policy information present in PA-SIP, the egress network switch takes actions accordingly.

The below examples illustrate the usage of PA-SIP. Below are sample ESG selector configurations to illustrate the policy.

ESG-Selector-100
security group webservers id 100
match ip 10.1.1.0/24

The endpoints matching to the 10.1.1.0/24 subnet are grouped to group policy ID 100.

ESG-Selector-200
security group database id 200
match ip 10.1.2.0/24

The endpoints matching to the 10.1.2.0/24 subnet are grouped to group policy ID 200.

Figure 3, below, illustrates an example of an ESG Permit Policy with a user configured permit policy for customer traffic from endpoint IP1 to endpoint IP2.

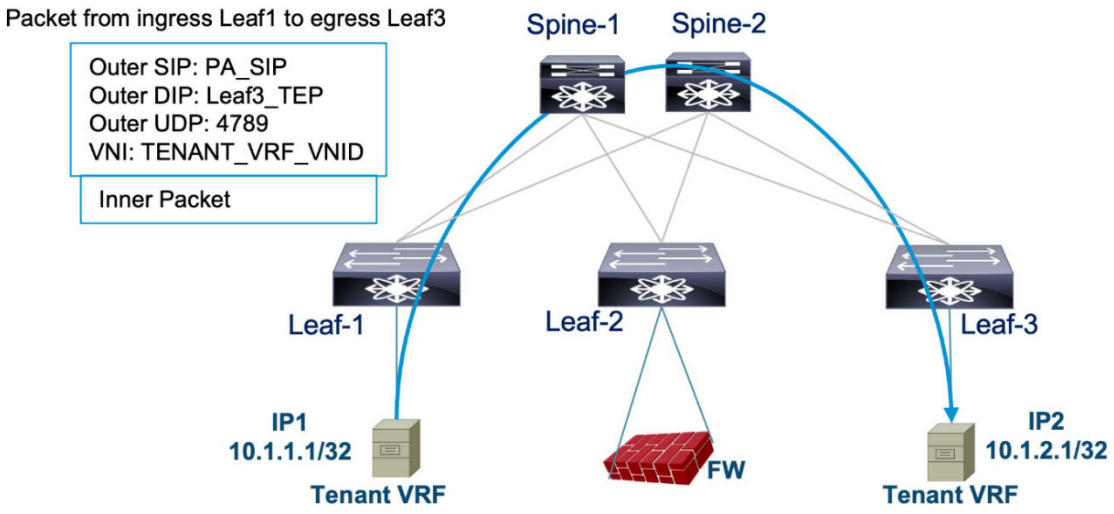

Figure 3: Example ESG Permit Policy

In the example illustrated in Figure 3, a user configured security contract policy allows data traffic between group policy ID 100 (10.1.1.0/24 endpoints) and group policy ID 200 (10.1.2.0/24 endpoints). In Figure 3, the ingress network switch (Leaf1) learned the identify of IP2 along with its group policy ID from the BGP EVPN control protocol. The ingress Leaf1 applies the user configured policy and keeps the policy applied information in the outer SIP of the VXLAN encapsulated data packet because the egress policy is not required. The egress network switch (Leaf3) honors the ingress policy and forwards the packet to endpoint IP2 based on a result of a traditional route lookup.

Figure 4, below, illustrates an example of an ESG security policy with service chain insertion.

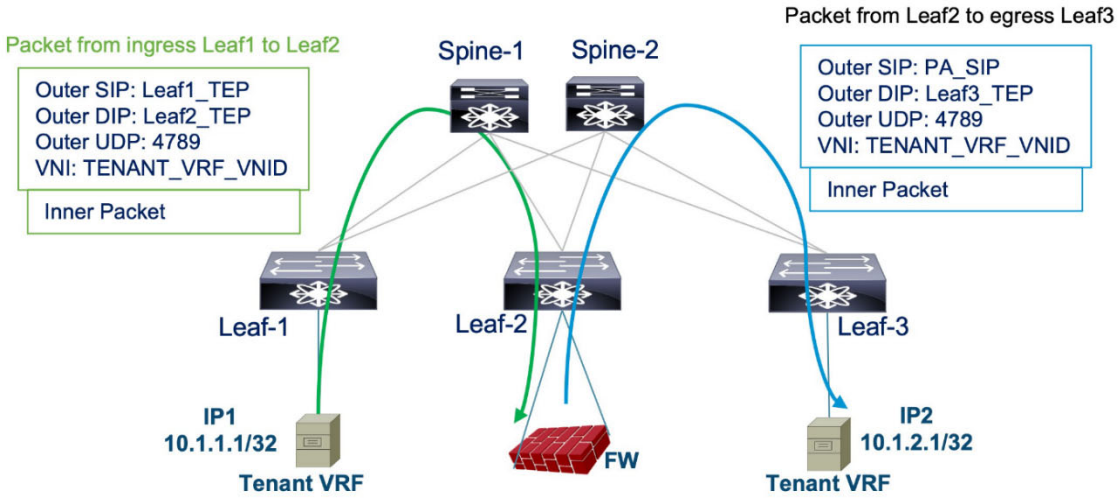

Figure 4: Example ESG Service Policy

In the example illustrated in Figure 4, a user configured policy redirects the customer data traffic between group policy ID 100 and group policy ID 200 to the firewall appliance present behind the Leaf2. In a first step (shown in green in Figure 4), the ingress network switch (Leaf1) applies the service redirect policy and then redirects the packet to the service appliance (Leaf2). For service redirected packets, the ingress leaf does not keep the PA-SIP in the data packet because the redirection policy is required in the service leaf to forward the packet to the firewall.

In a second step (shown in blue in Figure 4), when a packet is coming from the firewall into the fabric, the Leaf2 applies the ESG policy and keeps the PA information in the outer SIP of the VXLAN encapsulated data packet. The egress network switch (Leaf3) honors the policy applied information received in the data packet and forwards the packet to IP2 based on a result of a traditional route lookup. Without having the PA information in the blue color flow, the Leaf3 tries to apply the user configured redirect policy so that the packet is looped in the fabric. According to techniques described herein, with the PA-SIP in the data packet, the egress policy can be skipped selectively.

Intelligent Overloading of SGT in VXLAN Overlay UDP Source Port Field

Figure 5, below, illustrates the overlay packet format of the traffic from the source endpoint hosted behind the ingress switch to the destination endpoint hosted behind the egress switch.

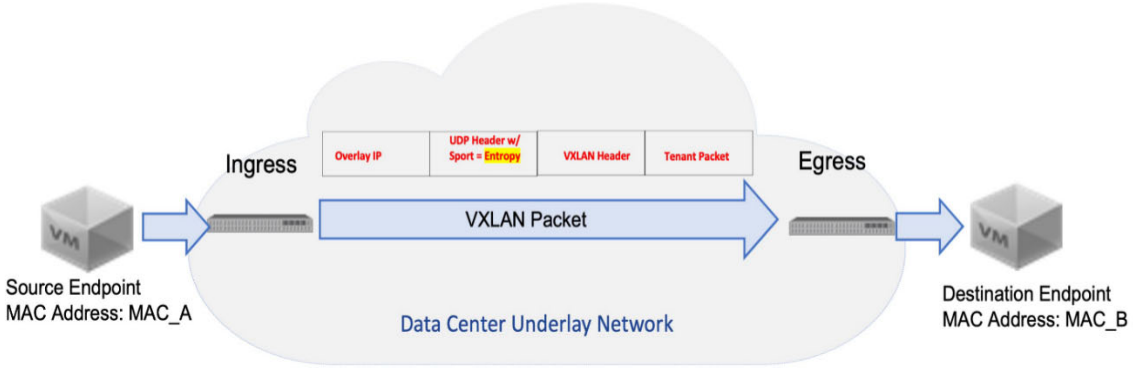

Figure 5: Example Overlay Packet Format of Traffic

As illustrated in Figure 5, the overlay UDP header's source port is 16 bits in length and the entropy hash of the original tenant packet is stored in this field. More details are provided in RFC 7348. The choice is left to the vendor/switch manufacturers to decide which fields from the tenant packet to use to calculate the hash and what kind of hash function (e.g., polynomial, non-linear, etc.) to use for this hash calculation.

In an example shipping data center switch, the following fields are used in the calculation of the entropy: {Source Mac Address, Destination Mac Address, Source IP Address, Destination IP Address, IP Protocol, L4 Source Port (if present), L4 Destination Port(if present)}.

In the above example, the final entropy, referred to as Equation 1, will be:

Entropy = Hash({MAC_A, MAC_B, IP_A, IP_B, IP Protocol, L4 Sport, L4 Dport}) % 65535 (Equation 1)

This 16-bit value is stamped in the packet's overlay UDP header's L4 Source Port field, as illustrated in Equation 2, below.

Outer UDP Sport = Entropy (Equation 2)

The SGT that needs to be carried is also 16 bits in length. Additionally, assume the switch's Application Specific Integrated Circuits (ASIC) pipeline is made to perform the operation in Equation 2 instead as shown below in Figure 3:

Outer UDP Sport = Entropy^SGT (Equation 3)

The ingress switch also XOR's the 16-bit SGT with the packet entropy calculated by Equation 1 and stores this 16-bit final result as the overlay's outer UDP source port. In other words, the ingress switch also encoded the SGT into the 16-bit UDP source port field.

This operation did not reduce the randomness provided by the entropy field, since Equation 3 yielded another 16-bit random number.

Next a determination is made as to whether the egress switch that needs to decapsulate the outer header can re-derive the SGT for its usage. Given that the egress switch also has the similar capabilities and exact hash functions of the ingress switch, the egress switch can also re-derive the original entropy again using same Equation 1.

Additionally, the egress switch knows the identity of the outer UDP source port that came in the packet. Using these two pieces of information, the egress switch can re-derive the SGT as shown below in Equation 4:

SGT = Entropy^Outer UDP Sport (Equation 4)

Once the SGT is re-derived, and together with local DGT, the egress switch can perform the policy lookup and enforce the security policies relevant for this traffic.

According to techniques described herein, by using programming capabilities (the XOR math operation) and since both the ingress and egress switches can recalculate the same entropy (using Equation 1) from the tenant packet, the overloading of the SGT in the outer UDP source port field became transparent and seamless. Using techniques described herein, group based policies may be implemented over standard VXLAN overlays, thereby eliminating the cost of implementing custom protocol modifications and helping data center customers by remaining with their existing fabrics without costly upgrades.

In summary, according to techniques described herein, in the subset of cases where only the PA-SIP option is available, ingress policing is performed, and customers are able to use existing deployments of switches. With the use of the SGT embedded in the source port, the deployment of silicon ASICs is supported, which allows for an increased scale due to egress enforcement and use in flood-and-learn networks.

Techniques described herein provide for an egress policy by carrying a source policy ID in the VXLAN overlay UDP source port field and an ingress policy by carrying PA information in VXLAN overlay outer source IP field. This allows for the use of existing hardware without the need for costly upgrades to newer hardware to have end-to-end policy enforcement. Even in mixed mode environments with existing hardware (e.g., standard VXLAN) and newer hardware (e.g., VXLAN-GBP), the techniques described herein are extendable to provide an end-to-end policy. Techniques described herein can

be adopted to other overlay technologies. Techniques described herein provide for overloading existing fields and carrying SGT and PA information to achieve micro-segmentation while using existing hardware (e.g., standard VXLAN) and not mandating costly hardware upgrades for customers. Some customers may have nodes that can inter-operate with the standard method on some links and can work in the pre-standard PA-IP method on other links that will be connected to the deployed fabrics. Therefore, gateway devices can be used to interoperate between devices that are standard capable and those that are not.