BESS WG
Internet-Draft
Intended status: Standards Track
Expires: 2 January 2023

S. Dikshit
T R. Gadekal
Aruba, HPE

1 July 2022

Secure IP Binding Synchronization via BGP EVPN
draft-saumthimma-evpn-ip-binding-sync-00

Abstract

The distribution of clients of L2 domain across extended, networks leveraging overlay fabric, needs to deal with synchronizing the Client Binding Database. The 'Client IP Binding' indicates the IP, MAC and VLAN details of the clients that are learnt by security protocols. Since learning 'Client IP Binding database' is last mile solution, this information stays local to the end point switch, to which clients are connected. When networks are extended across geographies, that is, both layer2 and layer3, the 'Client IP Binding Database' in end point of switches of remote fabrics should be in sync. This literature intends to align the synchronization of 'Client IP Binding Database" through an extension to BGP control plane constructs and as BGP is a typical control plane protocol configured to communicate across network boundries.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 2 January 2023.

Copyright Notice

Table of Contents

1. Important Terms

BGP: Border Gateway Protocol

VTEP: Virtual Tunnel End Point or Vxlan Tunnel End Point

RD: Route Distinguisher

RT: Route Target

NLRI: Network Layer Reachability Information

EVPN: Etherenet Virtual Private Network

2. Introduction

The distribution of clients of L2 domain across extended, networks leveraging overlay fabric, needs to deal with synchronizing the Client Binding Database. The 'Client IP Binding' indicates the IP, MAC and VLAN details of the clients that are learnt by security protocols. Since learning 'Client IP Binding database' is last mile solution, this information stays local to the end point switch, to which clients are connected. When networks are extended across geographies, that is, both layer2 and layer3, the 'Client IP Binding Database' in end point of switches of remote fabrics should be in sync. This literature intends to align the synchronization of 'Client IP Binding Database" through an extension to BGP control plane constructs and as BGP is a typical control plane protocol configured to communicate across network boundries.

3. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

When used in lowercase, these words convey their typical use in common language, and they are not to be interpreted as described in [RFC2119].

4. Problem Description

Access Switches, build trusted database of L3 clients by snooping into DHCP/ND packets in the client VLAN. This database can be leveraged by security and analytical features. Security features help in protecting the network from unauthorized/untrusted users while analytical features/application gauge the nature of access with the telemetry information for the designated hosts. This information help keeping the network health in a secure and informed way

* Prevent ARP cache depletion attacks

* Allow traffic only from known clients on access ports

* Denial of service and Man in the middle attacks

In Campus networks of Colleges, Branch office, Headquarters and DataCenter DC networks, it is a very common deployment, for networks, to extend Layer-2 across sites and geographies over WAN, Wide Area Network, via well defined overlay fabric solution like EVPN and constructs like Vxlan, GPE, GUE, NVGRE, with Vxlan being the most deployed. The solution for Campus and DC can be combined for enterprise solutions as well.

The campus networks are extended between headquarter and branch offices where as DCs are extended for load sharing, resiliency, hierarchical availability of data across geographies and region. In such deployments, the client database ends up being local to the first-hop access switch or gateway, the client is connected to. Client connected to the first hop access switch or gateway. The other set of access switches within or remotely placed in the extended fabric, treat the client as untruste or unauthorized on the client VLAN. This would result in switches to deny services to legitimate clients.

The following diagram shows the topology of disparate fabrics.

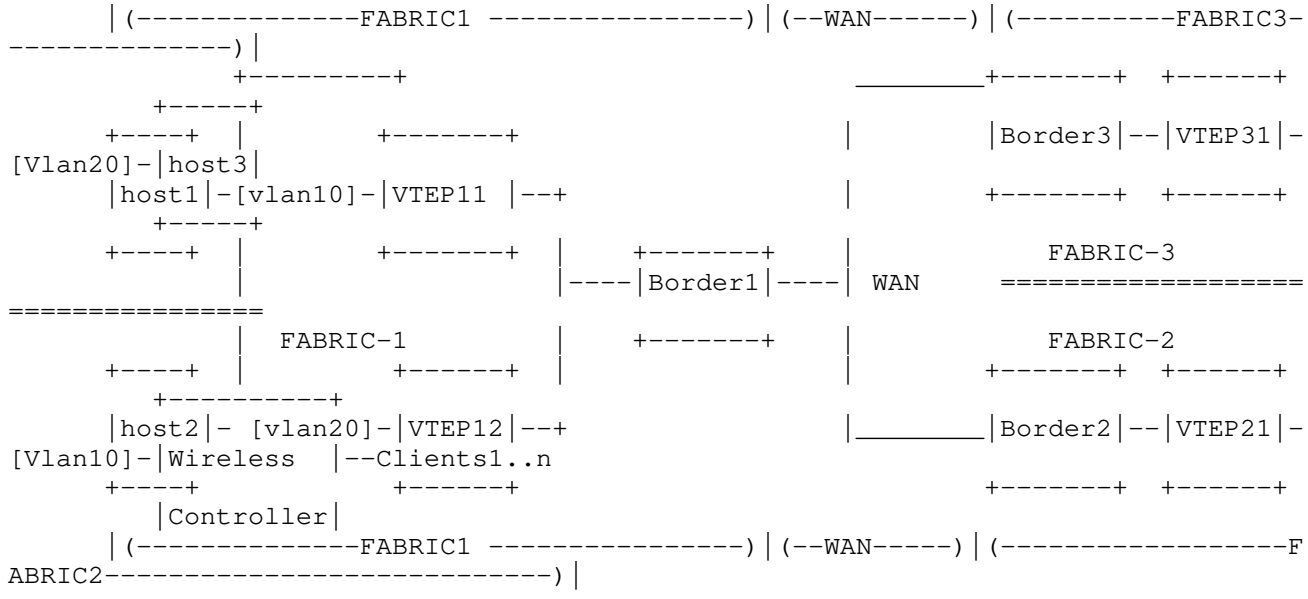


Figure 1: Figure 2: Multifabric Overlay Topology

In the above mentioned topology diagram,

* the Client IP Binding Database is local to client connected to the switches within the fabric. Without explicit solution in place, these details are not known to other switches within and outside the fabric, unless and until explicitly communicated. For example, the Access Vtep11 and Access Vtep12, though in the same fabric, can snoop in to DHCP packets originating or destined from the locally attached hosts, that is, Host1 and Host2 respectively, but not into the other Access hosts. These packets are typically unicasted to and from the DHCP server located in a centralized location. The packets are not leaked to remote fabrics as well.

The DHCP packets generated from Host1 or Host2 are typically not leaked to remote fabrics, Fabric2 or Fabric3 in the above example. Hence this information about host1 and host2 cannot be snooped in by remote Access devices Vtep12 and Vtep13 in the remote fabrics.

* The typical example shown in the above diagram, elaborates on the problem

* There are no known standard techniques to achieve this synchronization between switches, within or across extended network, via an overlay network.

* Security policies that needs knowledge about the connected clients in a VLAN across fabrics, will not work as expected in this model.

For example, to protect from neighbor cache depletion attacks, a switch can be configured to perform resolution only for the known clients in the network. This is not possible if Client IP Binding Database is not synced. NOTE, the above diagram, leverages BGP EVPN provisioned overlays over Vxlan fabric as the network extension instrument. Though the example can be generalized for any BGP provisioned overlay network like VPLS, VPWS, L3VPN etc.

One possible option is to build the client database by sniffing into data plane packets. This class of solution is ingrained with problems.

4.1. Broadcast Over WAN

This solution requires learnings over broadcast packets like ND and ARP. The bridge-domain extension over fabrics, will require the broadcasts to travel over the WAN pipe, and needs to be refreshed periodically as and when there is a request for host discovery. With reference to the above diagram, Gratuitous ARP(GARP), generated by Host2, will be required to be flooded to remote fabrics (fabric 2 and 3), in a typical data plane learning. But with EVPN control plane and arp-suppression techniques, there is minimal leak of arp broadcasts in the EVPN fabric, and thus the WAN. Hosts attached to access switches may GARP frequently (too many and too often), thus aggravating the broadcast leaks over WAN. As mentioned in the topology description, DCHP Snooping will not help here, as DHCP packets will not make it to remote fabrics. Even in DCHP Relay based deployments, the Relay configuration is local to a fabric and cannot be leaked to remote fabrics.

4.2. Same Subnet across fabrics over different Vlans

There is a possibility that same subnet allocation for ip address happens across disparate Vlans within or across the fabrics. Vlan 10 in fabric1 and Vlan20 in fabric 2 are mapped to same subnet gateway, lets say, 10.0.0.0/24. The IP binding database learnt via DHCP or ND snooping (hosted on access Vtep11 and Vtep13) needs to be synchronized as the broadcast domain is extended over a different VNI, lets say, 100 and 200 (mapped to vlan 10 and 20 respectively). Thus the problem at hand is that the allocations (IP bindings) are to be synchronized for same or different subnet, but across the Vlans in disparate fabrics. This is not possible with following deployment

* Native Vlan extension and vlan translation based fabric extensions

* Static Vxlan based network extensions.

Note that, In BGP Control plane, Route Targets help go around this anomaly with dataplane learnings.

4.3. Unwarranted and Insecure Information leak

With dataplane solution (via GARP or ND), the flooding over WAN to all remote fabrics will lead to unwarranted learning in fabrics over which there is no operator control.

* Thus potentially leaking client subnet specific information to the untargeted fabrics, creating an information leak and security risk.

* The source fabrics do not have control over the flooding in WAN (as its a flood in Vlan bridge domain).

* Hence there is a necessity to selectively leak or restrict the synchronization between specific fabrics.

In cases where DHCP Relay is configured,

* DHCP exchange happens between the client connected to Access Switch and the server through unicast channel. This is a predominant usecase but inherently prohibits other Access Switches snooping into the client DHCP packets.

* In the reference example, the Access Switches are also overlay tunnel end points, VTEP.

For example, in above diagram, the host1 GARP is flooded over WAN to both fabric2 and fabric3, even though the intention, is to sync the information to fabric2 only.

5. Security of Fast Roaming Clients

If a wireless client, lets say, host1, moves in a lift or elevator across floors and hence across gateways, the security policy synchronization is a must between the Vteps sitting behind the wireless controllers. Rather than waiting for client to speak up behind every wireless controller, the first controller and corresponding vtep can publish the security clearance or security risk of the host to all remote vteps.

6. Solution(s)

This document proposes a solution for synchronizing the IP Bindings between the Access Switches by leveraging the BGP EVPN control plane constructs called Route Type 2 NLRIs (EVPN NLRI, MAC Advertisement Route). The proposal defines a new extended community, which indicates the IP Binding synchronization, to be carried, in BGP Update message carrying the Route Type 2 NLRI (Network Layer Reachability Information). As EVPN is the at the core of fabric deployments to support multihoming and multitenancy, this control plane extension alleviates the Synchronization of IPs which is specific to tenants and applicable to all or selective hosts attached to the Access Vteps. The host can be attached to more than one Vtep (indicating multihoming) over a segment (called Ethernet Segment). BGP EVPN is a goto solution for Vxlan fabric extension across the WAN, as its also easy to realize the native transport (underlay) encapsulation as IP based.

For example, in the above diagram Vtep11 (also hosting the IP Binding database), publishes BGP update messages, carrying EVPN Route Type 2 for all the local MAC,IP learnings to the remote BGP peers (within or across the fabrics). These MAC,IP learnings are typically learnt via local dynamic learnings that is, host1 sends a GARP towards Vtep11 and is received on Vtep11 over the client (tenant) facing gateway interface (interface vlan 10). This interface can be a Vlan interface mapped to the subnet, for example, interface vlan 10 on Vtep11. All hosts attached over vlan 10 to Vtep11 are allocated the same subnet IP address and monitored by the IP Binding (or security validation) application hosted on Vtep11. Note that the similar validation is performed on all the access switches in a secure network. The dynamic learnings are validated against the IP Binding database (learnt via typically) as indicated in the EVPN Route Type 2. Note that the process of learning the IP Bindings (via snooping of DHCP or ARP) is outside the purview of this document.

6.1. Client IP Binding Sync Extended Communities

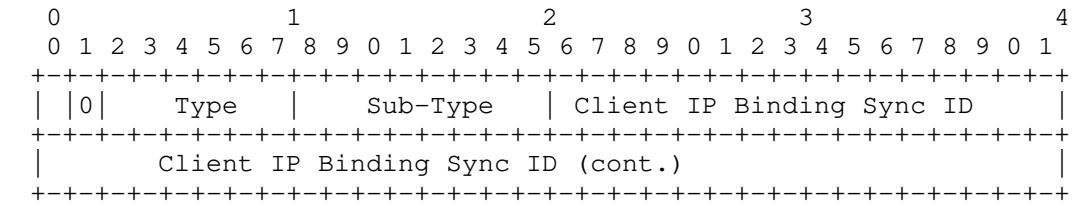
A new extended community Path attribute called Client-IP Binding Sync Extended Communities is defined. This attribute is an optional attribute and also transitive in nature and can be relayed in the control plane path. This attribute, if carried along with BGP update message with MAC, IP bindings (with EVPN NLRI, MAC Advertisement Route), indicates the following to the receiving BGP peer

* MAC, IP data can ALSO be leveraged for Client IP Binding synchronization.

* The data is to be handed over to the Security entity or application for validating the IP allocation

* The handover procedure is implementation specific and outside the purview of this invention.

The following diagram shows the Client IP Binding Sync Extended Communities



Field	Description
Type	1 Octet size, Type field can be newly defined as a proprietary one. Its a transitive Attribute, hence 1 bit in this octet is 0.
Sub-Type	1 Octet size. The value of sub-type will be allocated as per IANA For experimental purposes, it can be used as any unallocated value within the free range. This field when set to appropriate value indicates to the receiver about processing the NLRI for IP-Binding synchronization.
Client IP Binding Sync ID	6 Octets size, This field carries the Identifier configured to categorize the IP Binding instance. The field is applicable to all data carried in MAC IP NLRIs in the same update message. The instance is configurable and value is implementation dependent.

Figure 2: Figure 2 Client-IP Binding Sync Extended Communities

the following section gives a detailed flow of the send and receive side processing the new PDU.

6.1.1. Processing of Client IP Binding

The following set of bullets describe the 'Send Side Processing' flow

(1) the Access Vtep hosting the Client IP Binding Database can be triggered to carry the new extended communities, via a knob .

(2) The configuration can be an explicit 'CLIENT IP BINDING SYNC ID' and carried inside the Path attribute which can be configured on Access Vteps. The guidance for configuration of this ID is that, it indicates a Group of Access Switch spread across fabrics, that intend to share the IP Binding data as they might be part of same client or group of clients. All, catering to IP allocation of same tenant or group of tenants over same or disparate vlans and subnets. The configured value is ALSO leveraged to match the received value in the Route Type 2 Update message.

(3) The configuration scope of the 'CLIENT IP BINDING SYNC ID' can be global for all BGP updates or specific to vlans for which MAC,IP is being published in the UPDATE message. This is implementation specific and based on the 'IP BINDING SYNC IDs' scope within the 'Group of Access Switch'. If Vlan based, then 'CLIENT IP BINDING SYNC IDs' can be inherited or derived from 'Route Target' extended communities, configured on the 'Group of Access Switches' .

(4) The Withdraw of routes, Route Type 2, MAY also carry the extended community to indicate to the BGP peers configured on Group of Access Switches to convey the INVALIDITY of the MAC,IP bindings, published earlier.

(5) The Access Vteps (or BGP peer) on the send side MUST NOT insert the new extended communities in the withdraw message, if the corresponding IP Bindings are still valid.

the following bullets describe the 'Receive Side Processing'

(1) The same 'CLIENT IP SYNC ID' should be configured on the receiving Access Switch (BGP peer) to process or absorb the MAC, IP information within the IP Binding Context. For example, It can be a security application, which validates the IP Bindings.

(2) If the receiving BGP peer DOES NOT SUPPORTS the 'Client IP Binding Sync' Extended Communities in the received Route Type 2,it ignores the processing of this extended community while processing other attributes.Thus, no implication on any 'Client IP Binding' application, if hosted on this BGP Peer. It MUST respond to unsupported attribute as defined in [RFC4379].

(3) If the receiving BGP peer SUPPORTs the Client IP Binding Sync Extended Communities in the received Route Type 2, It decodes the Extended Communities and extracts the "Client IP SYNC ID" value. It matches the value with the locally configured one.

* If match is through, then it imports the MAC, IP NLRIs carried in the same update message to the Security application, which validates the MAC,IP against the security policies andi absorbs or drops after comparing it against the local IP Binding database.

* If nomatch then it MUST NOT import the NLRIs into local IP Binding Database.

(4) In case there are other BGP peers to whom the Update message is to be relayed, then the Client IP Binding Sync Extended Communities are also relayed without any updates in the values as its a transitive attribute. For example, Border1 receives an update message with the Client IP Binding Sync Extended Communities from Vtep11. It processes it as mentioned in above bullets and relays it to its eBGP (external BGP peers), Border2 and Border3 in remote fabrics fabric2 and fabric3 respectively as its an optional transitive attribute. Note that the above example quotes eBGP (different Autonomous System) across fabrics, but same logic will apply when Route Reflector reflects routes between two iBGP peers.

The augmented control plane feature, helps all Access swtiches to be synchronized with respect to allowed or validated client credentials, thus preventing rogue traffic or DoS attacks 'originating in' or 'destined to' the local network.

7. Backward Compatibility

Backward Compatibility for non-support nodes is as per the following standards already defined in [RFC7606], that, BGP speaker should discard the unsupported TLV types

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997, <http://www.rfc-editor.org/rfc/rfc2119.txt>.

[RFC4379] Kompella, K., "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006, <https://www.rfc-editor.org/rfc/rfc4379.html>.

[RFC7153] Rosen, E., "IANA Registries for BGP Extended Communities", RFC 7153, March 2014, <https://www.rfc-editor.org/rfc/rfc7153.html>.

[RFC7348] Mahalingam, M., "Virtual eXtensible Local Area Network (VXLAN): A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", RFC 7348, August 2014, <http://www.rfc-editor.org/rfc/rfc7348.txt>.

[RFC7432] Sajassi, A., "BGP MPLS-Based Ethernet VPN", RFC 7432, February 2015, <http://www.rfc-editor.org/rfc/rfc7432.txt>.

[RFC7606] Chen, E., "Revised Error Handling for BGP UPDATE Messages", RFC 7606, August 2015, <https://www.rfc-editor.org/rfc/rfc7606.html>.

[RFC9014] Rabadan, J., "Interconnect Solution for Ethernet VPN (EVPN) Overlay Networks", RFC 9014, May 2021, <http://www.rfc-editor.org/rfc/rfc9014.txt>.

Saumya Dikshit
Aruba Networks, HPE
Mahadevpura
Bangalore 560 048
Karnataka
India
Email: saumya.dikshit@hpe.com

Gadekal, Thimma Reddy
Aruba Networks, HPE
Mahadevpura
Bangalore 560 048
Karnataka
India
Email: tgadekal@hpe.com