NV03 working group   L. Dunbar
Internet Draft   D. Eastlake
Category: Informational   Huawei

Expires: April 4 2014   September 20, 2013

NV03 NVA Gap Analysis

draft-dunbar-nvo3-nva-gap-analysis-01

Status of this Memo

Copyright Notice

Abstract

The intent of the draft is to identify the gaps of existing solutions against NVO3's NVE <-> NVA control plane requirement. Through the gap analysis, the document provides a basis for future works that develop solutions for NVE<->NVA control plane.

Table of Contents

1. Introduction

The intent of the draft is to identify the gaps of existing solutions against NVO3's requirement for Network Virtualization Authority (NVA). Through the gap analysis, the document provides a basis for future works to develop solutions for (NVA).

The existing solutions analyzed in draft include the LISP mapping database system and TRILL's directory mechanism.

Section 4.5 of [nvo3-problem-statement] describes the back-end Network Virtualization Authority (NVA) that is responsible for distributing the mapping information for entire overlay system. [nvo3-nve-nva-cp-req] defines the requirement for the control plane between NVA and NVE.

There are many similarities between LISP, TRILL [RFC6325] and NVO3, e.g. LISP using IP header to achieve overlay, TRILL using TRILL header to achieve overlay, and NV03 using L3 headers plus VNID to achieve overlay. This draft analyzes the TRILL directory mechanisms along with some LISP mapping database system that are applicable to NVO3's NVA<->NVE and summarize the gaps.

2. Terminology

The following terms are used interchangeably in this document:

- The terms "Subnet" and "VLAN" because it is common to map one subnet to one VLAN.
- The term ''Directory'' and ''Network Virtualization Authority (NVA)''
- The term ''NVE'' and ''Edge''

Bridge: IEEE Std 802.1Q-2011 compliant device [802.1Q]. In this draft, Bridge is used interchangeably with Layer 2 switch.

DA: Destination Address

DC: Data Center

EoR: End of Row switches in data center. Also known as aggregation switches.

End Station: Guest OS running on a physical server or on a virtual machine. An end station in this document has at

least one IP address and at least one MAC address, which could be in DA or SA field of a data frame.

LISP: Locator/ID Separation Protocol

RBridge: ''Routing Bridge'', an alternative name for a TRILL switch.

NVA: Network Virtualization Authority

NVE: Network Virtualization Edge

SA: Source Address

Station: A node, or a virtual node, with IP and/or MAC addresses, which could be in the DA or SA of a data frame.

ToR: Top of Rack Switch in data center. It is also known as access switches in some data centers.

TRILL: Transparent Interconnection of Lots of Links [RFC6325]

TRILL switch: A device implementing the TRILL protocol [RFC6325]

TS: Tenant System

VM: Virtual Machines

VN: Virtual Network

VNID: Virtual Network Instance Identifier

3. Overall Requirement for NVE<->NVA Control Plane

Section 3.1 of [nvo3-cp-req] describes the basic requirement of inner address to outer address mapping for NVO3. A NVE needs to know the mapping of the Tenant System destination (inner) address to the (outer) address (IP) on the Underlying Network of the egress NVE, in the same way as a TRILL Edge node needing to know how the inner MAC/VLAN is mapped to the egress TRILL edge.

Section 3.1 of [nvo3-cp-req] states that a protocol is needed to provide this inner to outer mapping and VN Context to each NVE that requires it and keep the mapping updated in a timely manner.

Timely updates are important for maintaining connectivity between Tenant Systems.

TRILL's directory mechanism and LISP mapping database system are to achieve the same goal as NVO3's NVE-NVA control plane, i.e. distributing the mapping table that edge nodes use to tunnel traffic across the underlying network. Therefore it is worthwhile to examine the TRILL's directory mechanism and LISP mapping database system, and analyze the gaps.

4. Existing Directory Components

For the ease of description, we match the terminologies used by TRILL/LISP to NVO3. The document will use the NVO3's terminologies as much as possible throughout the document to describe TRILL's directory assistance mechanism.

NVO3	LISP	TRILL
NVE	Edge	Edge, TRILL Edge or RBridge Edge
NVA	MapServer	Directory

4.1. Types of NVA:

NVAs can be centralized or distributed with each NVA holding the mapping information for a subset of VNs. Centralized NVA could have multiple entities for redundancy purpose. A NVA could be instantiated on a server/VM attached to a NVE, very much like a TS attached to a NVE, or could be integrated with a NVE. When a NVA is a standalone server/VM attached to a NVE, it has to be reachable via the attached NVE by other NVEs. A NVA can also be instantiated on a NVE that doesn't have any TSs attached. The NVE-NVA control plane for NVA being attached to NVE will require additional functions on NVEs than NVA being instantiated on NVE.

4.2. Key components of the information kept in the NVA

The information held by the TRILL directories is inner-outer address mapping information as well as hosts' VLAN IDs. Same is true for NVO3's NVA. For each TS (or VM), TRILL directory has the following attributes:

1. Inner Address: TS (host) Address family (IPv4/IPv6, MAC, virtual network Identifier MPLS/VLAN, etc)
2. Outer Address: The list of locally attached edges (NVEs); normally one TS is attached to one edge, TS could also be attached to 2 edges for redundancy (dual homing). One TS is rarely attached to more than 2 edges, though it could be possible;
3. Timer for NVEs to keep the entry when pushed down to or pulled from NVEs.
4. Optionally the list of interested remote edges (NVEs). This information is for NVA to promptly update relevant edges (NVEs) when there is any change to this TS' attachment to edges (NVEs). However, this information doesn't have to be kept per TS. It can be kept per VN.

NVO3's NVA will need one additional attribute: VN Context (VN ID and/or VN Name).

4.3. Mapping Entries Distribution Mechanism

A directory can offer services in a Push, Pull mode, or the combination of the two.

4.3.1. Push Mode

Under this mode, Directory Server(s) push the inner-outer mapping for all the entries of the VNs that are enabled on an edge node (NVE). If the destination of a data frame arriving at the Ingress Edge (NVE) can't be found in its inner-outer mapping database that are pushed down from the Directory Server(s) (or NVA), the Ingress edge could be configured with one or more of the following policies:

- simply drop the data frame,

- Using legacy method(s) to forward the data frames to other edges, or
- start the ''pull'' process to get information from Pull Directory Server(s) (or NVA)
  When the edge is waiting for reply from Pull process, the edge can either drop or queue the packet.

Again, the VN Context (VNID or VN name) needs to be added for NVO3.

One drawback of the Push Mode is that it usually will push more mapping entries to an edge (NVE) than needed. Under the normal process of edge cache aging and unknown destination address flooding, rarely used entries would have been removed. It would be difficult for Directory Servers (NVA) to predict the communication patterns among TSs within one VN. Therefore, it is likely that the Directory Servers will push down all the entries for all the VNs that are enabled on the NVE.

And with Push there really can't be any source-based policy. It's all or nothing.

4.3.1.1. Requesting Push Service

In the Push Mode, it is necessary to have a way for an edge node (NVE) to request directory server(s) (NVA) to start the pushing process, e.g. when the NVE is initialized or re-started. Or it can be like a routing protocol where it just happens automatically.

Push Directory servers (NVAs) advertise their availability to push mapping information for a particular virtual network to all edges who participate in the VN. There could be multiple directories (NVAs), with each having mapping information for a subset of VNs.

TRILL edge uses modified Virtual Network scoped instances of the IS-IS reliable link state flooding protocol, a.k.a. the ESADI protocol mechanism, to announce all the Virtual Networks in which it is participating to directories (NVAs) who have the mapping information for the VNs. An edge subscribes to push directory information.

The subscription is VN scoped, so that a directory server doesn't need to push down the entire set of mapping entries. Each Push Directory server also has a priority. For robustness, the one or two directories with the highest priority are considered as Active in pushing information for the VN to all edges who have subscribed for that VN.

4.3.1.2. Incremental Push Service

Whenever there is any change in TS' association to an edge (NVE), which can be triggered by TS being added, removed, or de-commissioned, an incremental update can be sent to the edges that are impacted by the change. Therefore, sequence numbers have to be maintained by directory servers (NVA) and edges (NVEs).

If the Push Directory server is configured to believe it has complete mapping information for VN X then, after it has actually transmitted all of its ESADI-LSPs for X it waits its CSNP time (see Section 6.1 of [ESADI]), and then updates its ESADI-Parameters APPsub-TLV to set the Complete Push (CP) bit to one. It then maintains the CP bit as one as long as it is Active.

4.3.2. Pull Mode

Under this mode, an NVE pulls the mapping entry from the directory servers (or NVA) when its cache doesn't have the entry.

The main advantage of Pull Mode is that state is stored only where it needs to be stored and only when it is required. In addition, in the Pull Mode, edge nodes (NVEs) can age out mapping entries if they haven't been used for a certain period of time. Therefore, each edge (NVE) will only keep the entries that are frequently used, so its mapping table size will be smaller than a complete table pushed down from NVA.

The drawback of Pull Mode is that it might take some time for NVEs to pull the needed mapping from NVA. Before NVE gets the response from NVA, the NVE has to buffer the subsequent data frames with destination address to the same target. The buffer could overflow before the NVE gets the response from NVA. However, this scenario should not happen very often in data center environment because most likely the TSs are end systems

which have to wait for (TCP) acknowledgement before sending subsequent data frames. Another option is forward, not flood, subsequent frames to a default location, i.e. forward to a re-encapsulating NVE.

The practice of an edge waiting and dropping packets upon receiving an unknown DA is not new. Most deployed routers today drop packets while waiting for target addresses to be resolved. It is too expensive to queue subsequent packets while resolving target address. The routers send ARP/ND requests to the target upon receiving a packet with DA not in its ARP/ND cache and wait for an ARP or ND responses. This practice minimizes flooding when targets don't exist in the subnet. When the target doesn't exist in the subnet, routers generally re-send an ARP/ND request a few more times before dropping the packets. The holding time by routers to wait for an ARP/ND response when the target doesn't exist in the subnet can be longer than the time taken by the Pull Mode to get mapping from NVA.

4.3.2.1. Pull Requests

Here are some events that can trigger the pulling process:

- An edge node (NVE) receives an ingress data frame with a destination whose attached edge (NVE) is unknown, or
- The edge node (NVE) receives an ingress ARP/ND request for a target whose link address (MAC) or attached edge (NVE) is unknown.

Each Pull request can have queries for multiple inner-outer mapping entries.

4.3.2.2. Pull Response

There are several possibilities of the Pull Response:

1. Valid inner-outer address mapping, coupled with the valid timer indicating how long the entry can be cached by the edge (NVE).
   The timer for cache should be short in an environment where VMs move frequently. The cache timer can also be configured.

2. The target being queried is not available. The response should include the policy if requester should forward data frame in legacy way, or drop the data frame.

3. The requestor is administratively prohibited from getting an informative response.

If no response is received to a Pull Directory request within a configurable timeout, the request should be re-transmitted with the same Sequence Number up to a configurable number of times that defaults to three.

4.3.2.3. Cache Consistency

It is important that the cached information be kept consistent with the actual placement of VMs. Therefore, it is highly desirable to have a mechanism to prevent NVEs from using the staled mapping entries.

When data at a Pull Directory changes, such as entry being deleted or new entry added, and there may be unexpired stale information at a querying edge (NVE), the Pull Directory MUST send an unsolicited message to the edge (NVE).

To achieve this goal, a Pull Directory server MUST maintain one of the following, in order of increasing specificity.

1. An overall record per VN of when the last returned query data will expire at a requestor and when the last query record specific negative response will expire.

2. For each unit of data (IA APPsub-TLV Address Set) held by the server and each address about which a negative response was sent, when the last expected response with that unit or negative response will expire at a requester.

Note: It is much more important to cache negative reply, because there are many invalid address queries. Study has shown that for each valid ND query, there are 100's of invalid address queries.

3. For each unit of data held by the server and each address about which a negative response was sent, a list of Edges that were sent that unit as the response or sent a negative

response to the address, with the expected time to expiration at each of them.

4.3.2.4. Pull Request Errors

If errors occur at the query level, they MUST be reported in a response message separate from the results of any successful queries. If multiple queries in a request have different errors, they MUST be reported in separate response messages. If multiple queries in a request have the same error, this error response MAY be reported in one response message.

4.3.2.5. Redundant Pull Directories (NVAs)

There could be multiple directories (NVA) holding mapping information for a particular VN for reliability or scalability purposes. Pulling Directories (NVAs) advertise themselves by having the Pull Directory flag on in their Interested VNs sub-TLV [rfc6326bis].

A pull request can be sent to any of them that is reachable but it is RECOMMENDED that pull requests be sent to a server (NVA) that is least cost from the requesting edge (NVE).

4.3.3. Hybrid Mode

For some edge nodes that have great number of VNs enabled and combined number of hosts under all those VNs are large, managing the inner-outer address mapping for hosts under all those VNs can be a challenge. This is especially true for Data Center gateway nodes, which need to communicate with a majority of VNs if not all.

For those Edge nodes, a hybrid mode should be considered. That is the Push Mode being used for some VNs, and the Pull Mode being used for other VNs. It is the network operator's decision by configuration as to which VNs' mapping entries are pushed down from directories (NVA) and which VNs' mapping entries are pulled.

In addition, directory can inform the Edge to use legacy way to forward if it doesn't have the mapping information, or the

Edge is administratively prohibited from forwarding data frame to the requested target.

5. Redundancy

For redundancy purpose, there should be more than one directory (NVAs) that hold mapping information for each VN. At any given time, only one or a small number of push directories is considered as active for a particular VN. All NVAs should announce its capability and priority to all the edges.

6. Inconsistency Processing

If an edge (NVE) notices that a Push Directory server (NVA) is no longer reachable [RFCclear], it MUST ignore any Push Directory data from that server because it is no longer being updated and may be stale.

There may be transient conflicts between mapping information from different Push Directory servers (NVAs) or conflicts between locally learned information and information received from a Push Directory server. TRILL associates a confidence level with address table information so, in case of such conflicts, information with a higher confidence value is preferred over information with a lower confidence. In case of equal confidence, Push Directory information is preferred to locally learned information and if information from Push Directory servers conflicts, the information from the higher priority Push Directory server is preferred.

7. Gap Summary

7.1. Features necessary to NVO3 but not present in TRILL

NVO3's NVA will need one additional attribute: VN context (VN ID and/or VN Name).

For data center networks that don't have IS-IS protocol enabled, other mechanism have to be considered.

7.2. Additional detailed requirement applicable to NVO3's NVA

Here are some of the TRILL's directory detailed requirements that should be considered by NVO3 NVA as well:

- Push Mode:
  - o For redundancy purposes, for each VN there should be multiple NVA entities holding the mapping information for the TSs in the VN. At any given time, only one or a small number of the NVAs are considered as Active for a particular VN. All NVAs should announce their capability and priority to all the edges.
  - o If the destination of a data frame arriving at the Ingress Edge (NVE) can't be found in its inner-outer mapping table that are pushed down from the Directory Server(s) (NVA), the Ingress edge could be configured to:

    simply drop the data frame,
    flood it to all other edges that are in the same VN,
    or
    start the ''pull'' process to get information from Pull Directory Server(s) (or NVA)
  - o If an NVE lost its connection to its NVA, it MUST ignore any Push Directory data from that server because it is no longer being updated and may be stale.
  - o When transient conflict occurs: higher priority data take precedence.

- Pull Mode:
  - o The Pull Directory response could indicate that the address being queried is not available in NVA or that the requestor is administratively prohibited from getting an informative response.
  - o The timer for ingress NVE caching should be short in an environment where VMs move frequently. The cache timer could be configured or could be sent along with the Pulled reply from the NVA.
  - o Each Pull request can have multiple queries for different TSs.
  - o It is highly desirable to have a mechanism to prevent NVEs from using the stale mapping entries pulled from NVA.

o While waiting for query response from NVA, the NVE has to buffer the subsequent data frames with destination address to the same target. The buffer could overflow before the NVE gets the response from NVA.
o If no response is received to a Pull Directory request within a configurable timeout, the request should be re-transmitted with the same Sequence Number up to a configurable number of times.

- Hybrid Mode:
o NVE can be configured to get some VN's mapping entries via push mode and other VN's mapping entries via pull mode.

8. Security Considerations

Accurate mapping of inner address into outer addresses is important to the correct delivery of information. The security of specific directory assisted mechanisms will be discussed in the document or documents specifying those mechanisms.

For general TRILL security considerations, see [RFC6325].

9. IANA Considerations

This document requires no IANA actions. RFC Editor: please delete this section before publication.

10. Acknowledgements

Special thanks to Dino Farinacci for valuable suggestions and comments to this draft.

11. References

11.1. Normative References

As an Informational document, this draft has no Normative References.

[nvo3-nve-nva-cp-req] draft-ietf-nvo3-nve-nva-cp-req-00, "Network Virtualization NVE to NVA Control Protocol Requirements", Kreeger, et al. July 31, 3013.

11.2. Informative References

[802.1Q] IEEE Std 802.1Q-2011, "IEEE Standard for Local and metropolitan area networks - Virtual Bridged Local Area Networks", May 2011.

[802.1Qbg] IEEE Std 802.1Qbg-2012, ''Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks -- Edge Virtual Bridging'', July 2012.

[RFC826] Plummer, D., "An Ethernet Address Resolution Protocol", RFC 826, November 1982.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC6325] Perlman, et, al ''RBridge: Base Protocol Specification'', https://datatracker.ietf.org/doc/rfc6325/, July, 2011

[RFC6439] Perlman, et, al ''RBridges: Appointed Forwarders'', https://datatracker.ietf.org/doc/rfc6439/, Nov 2011

Authors' Addresses

Linda Dunbar
Huawei Technologies
5430 Legacy Drive, Suite #175
Plano, TX 75024, USA
Phone: (469) 277 5840
Email: linda.dunbar@huawei.com

Donald Eastlake
Huawei Technologies
155 Beaver Street
Milford, MA 01757 USA
Phone: 1-508-333-2270
Email: d3e3e3@gmail.com