Internet Engineering Task Force
Internet Draft
Intended status: Informational
Expires: December 18, 2014

J. Pelissier, Ed.
Cisco

P. Thaler
Broadcom

P. Bottorff
HP

June 18, 2014

NVO3 VDP Gap Analysis - VM-to-NVE Specific Control-Plane Requirements
draft-pt-nvo3-vdp-vm2nve-gap-analysis-00.txt

Status of this Memo

Copyright Notice

Abstract

[I-D.kreeger-nvo3-hypervisor-nve-cp-01] discusses requirements for Hypervisor-to-NVE Control Plane Protocol Functionality. The IEEE has developed a protocol called VSI Discovery Protocol (VDP) specified in [IEEE8021Qbg]. This protocol is intended to address the same basic problems at layer two as the Hypervisor-to-NVE protocol needs to address at layer three. Simply by adding the ability to carry layer three addresses to VDP using the extensibility features built into the protocol, VDP may be used as the Hypervisor-to-NVE Control Plane protocol.

Table of Contents

1. Introduction

[I-D.kreeger-nvo3-hypervisor-nve-cp-01] discusses requirements for Hypervisor-to-NVE Control Plane Protocol Functionality. The IEEE has developed a protocol called VSI Discovery Protocol (VDP) specified in [IEEE8021Qbg]. This protocol is intended to address the same basic problems at layer two as the Hypervisor-to-NVE protocol needs to address at layer three. Simply by adding the ability to carry layer three addresses to VDP using the extensibility features built into the protocol, VDP may be used as the Hypervisor-to-NVE Control Plane protocol.

This document provides a summary of the data formats and operation of VDP. It then provides an analysis of the requirements of the Hypervisor-to-NVE protocol and summarizes VDP's ability to meet these requirements.

2. Terminology and Conventions

2.1. 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 RFC-2119 [RFC2119].

2.2. Conventions

In sections providing analysis of requirements defined in referenced documents, section numbers from each referenced document are used as they were listed in that document.

In order to avoid confusing those section numbers with the section numbering in this document, the included numbering is parenthesized.

2.3. Terms and Abbreviations

This document uses terms and acronyms defined in [IEEE8021Qbg] and [I-D.kreeger-nvo3-hypervisor-nve-cp-01]:

ECP: Edge Control Protocol [IEEE8021Qbg]

VDP: Virtual Station Interface (VSI) Discovery and Configuration Protocol [IEEE8021Qbg]

VNIC: Virtual Network Interface Card [I-D.kreeger-nvo3-hypervisor-nve-cp-01]

VSI: Virtual Station Interface [IEEE8021Qbg]

This document uses the following additional general terms and abbreviations:

PDU: protocol data unit

TLV: type, length, value

3. VDP Operational Summary

3.1. Introduction

VDP associates a Virtual Station Interface (VSI) with its virtually or physically attached bridge port. While the standard assumes the use of a virtual station, the protocol is actually agnostic as to whether the station is virtually or physically instantiated.

The term VSI as used in [IEEE8021Qbg] is equivalent to the term VNIC used in [I-D.kreeger-nvo3-hypervisor-nve-cp-01].

In addition, VDP automates station configuration during the movement of a VSI from one station to another or from one bridge to another.

3.2. Data Formats

This section provides a descriptive overview of the data formats and the definition of the fields within these formats. For the detailed specification, see [IEEE8021Qbg].

The VDP data formats are defined in terms of type, length, value tuples (TLVs). There are three TLVs defined for VDP:

- VSI Manager ID
- VDP Association
- VDP Organizationally Defined

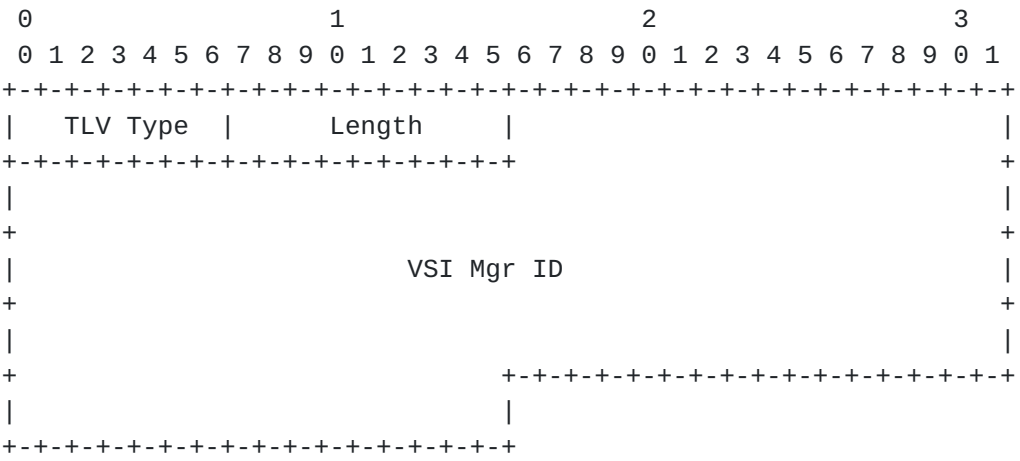
These TLVs are carried in a PDU using ECP. It should be noted that ECP is independent of VDP and that VDP may be transported utilizing any protocol capable of reliably transporting a PDU.

Each PDU contains exactly one VSI Manager ID TLV that is the first TLV in the PDU. The VSI Manager ID TLV is followed by one or more VDP Association TLVs and zero or more VDP Organizationally Defined TLVs. The TLVs following the VSI Manager ID TLV occur in any order.

3.2.1. VSI Manager ID TLV

The VSI Manager ID TLV provides a way for a hypervisor to indicate the address of a manager that contains network configuration information for the VSIs in the PDU.

The following illustrates the format of the VSI Manager ID TLV:



Field descriptions:

TLV Type: Set to 5.

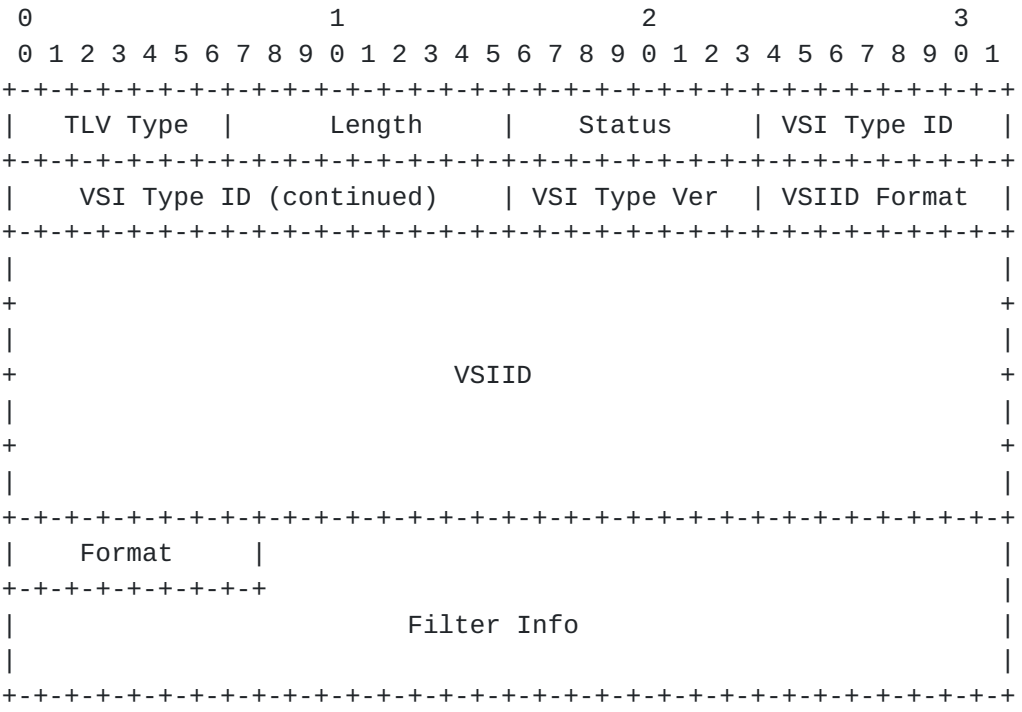
Length: Contains 16, the length of the information field in octets.

VSI Mgr ID: 16 octet field identifying the IPv6 [RFC4291] address of the manager from which to obtain the VSI type. A value of 0 indicates that the device does not know this address.

3.2.2. VDP Association TLV

The VDP Association TLV identifies a VSI and the Filter Info for packets from that VSI. Filter Info is information from which a filter for packets from that VSI can be constructed.

The following illustrates the format of the VDP association TLV:



Field definitions:

TLV Type: Set to one of the following values based on the type of TLV:

Value	TLV Type
1	Pre-Associate
2	Pre-Associate with resource reservation
3	Associate
4	De-associate

Length: Contains the length of the TLV information string which is 23 plus the number of octets in the Filter Info field.

Status: The status field contains four flags encoded one each in bits 16-19 and an error type encoded bits 20-23:

Bit 16: Reserved for future standardization.

Bit 17: Req/Ack - set to zero to indicate that the TLV contains a request.

Bit 18: S-bit - indicates that the VSI user is suspended (S-bit = 1) or no information (S-bit = 0).

Bit 19: M-bit - indicates that the VSI user is migrating (M-bit = 1) or no information (M-bit = 0).

Bits 20-23:

Value	Error Type
0	Success
1	Invalid Format
2	Insufficient Resources
3	Unable to contact VSI manager
4	Other failure
5	Invalid VID, GroupID, or MAC address
all others	Reserved for future standardization

VSI Type ID: An integer used to identify the type of the VSI. The type of VSI is used by the VSI manager to obtain the configuration for a VSI and its scope is limited to an individual VSI manager.

VSI Type Ver: The version of a VSI type. This allows a VSI database to maintain multiple versions of a VSI type.

VSIID format: Indicates the format of the VSIID field. The allowed values are:

Value	Description
1	An IPv4 address encoded as specified in [RFC4291]
2	An IPv6 address encoded as specified in [RFC4291]
3	An IEEE 802 MAC address (6 octets) with 10 leading octets of all zeros
4	The format is locally defined
5	A UUID as specified in [RFC4122]
All others	Reserved for future standardization

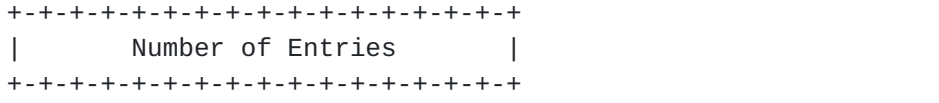
VSIID: An identifier of the VSI instance in the format specified by VSIID format.

Format: Indicates the format of the Filter Info field. The allowed values are:

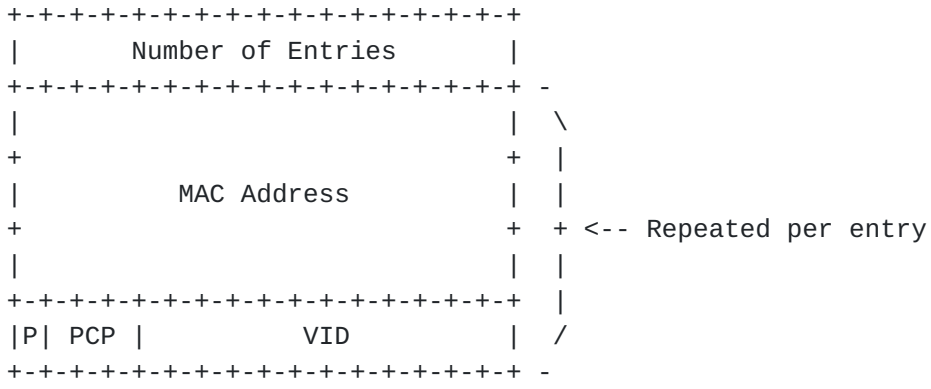
Value	Description
1	VID
2	MAC/VID
3	GroupID/VID
4	GroupID/MAC/VID
All others	Reserved for future standardization

Filter Info: The contents of this field vary depending on the value of the Format field.

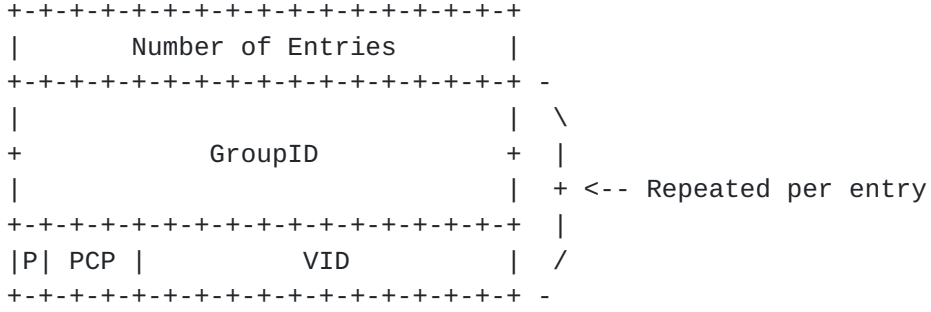
If the Format field indicates the VID format, the format of the Filter Info field is as follows:



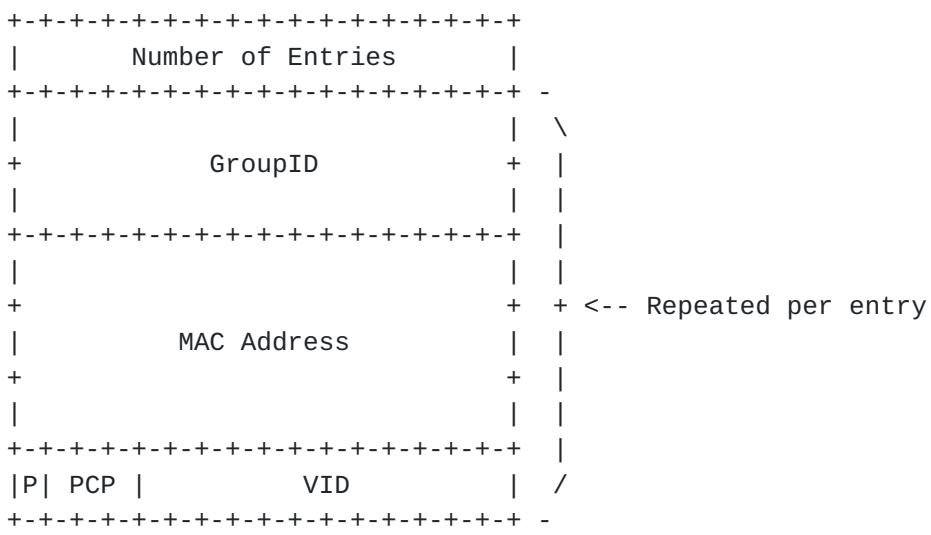
If the Format field indicates the MAC/VID format, then the format of the Filter Info field is:



If the Format field indicates the GroupID/VID format, then the format of the Filter Info field is:



If the Format field indicates the GroupID/MAC/VID format, then the format of the Filter Info field is:



The following field definitions apply to all formats of the Filter Info field in which the defined field appears:

Number of Entries: Contains the number of filter entries in the Filter Info field.

GroupID: Enables the specification of a VLAN when the total number of VLANs exceeds 4095. For Filter Info formats with a GroupID, the hypervisor can send the Null VID. The Bridge then supplies a local VID that it maps to the GroupID. See [IEEE8021Qbg] for details.

MAC Address: An IEEE 802 MAC address.

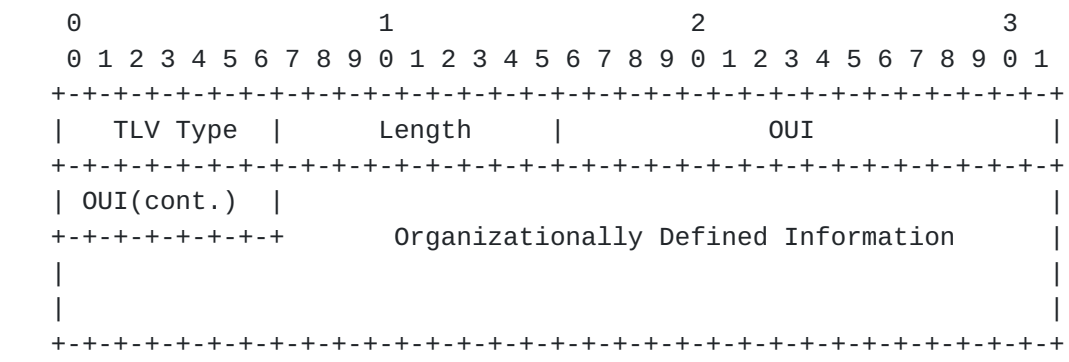
P: Set to one to indicate that the PCP field is significant, 0 otherwise.

PCP: Priority code point. If P is set to zero, this field is ignored and the Filter Info entry applies to all Priority code points.

VID: VLAN Identifier.

3.2.3. Organizationally Defined TLV

The following illustrates the format of the VDP association TLV:



Field descriptions:

TLV Type: Set to 0x7f.

Length: Contains the length of the TLV information string in octets which is 3 (the length of the OUI) plus the length of the Organizationally Defined Information.

OUI: An organizationally unique identifier assigned by the IEEE registration authority that identifies the organization that defined the content of the Organizationally Defined Information field.

Organizationally Defined Information: Information that is defined by the organization identified by the OUI field.

3.3. VDP Operations

VDP provides for four fundamental operations:

1. Pre-Associate
2. Pre-Associate with Resource Reservation
3. Associate
4. De-Associate

These operations are described in detail with their associated state machines in [IEEE8021Qbg]. The following sub-paragraphs provide a general description of each of these operations. Each operation may be initiated by a hypervisor or other entity within a station and responded to by the bridge. In addition, the bridge may initiate the De-Associate operation.

3.3.1. Pre-Associate

The Pre-Associate operation informs the bridge that the station may initiate an Associate operation with the same parameters in the

future. This allows the bridge to validate the operation and inform the station whether or not an identical Associate operation would have succeeded. However, this provides no guarantee that the same or similar Associate operation will succeed in the future.

Additionally, this operation allows the bridge to prepare for a future Associate operation, e.g. caching configuration information from a management server, thereby potentially decreasing the time required to process the future Associate operation.

It is not necessary to perform a Pre-Associate operation prior to an Associate operation.

3.3.2. Pre-Associate with Resource Reservation

The Pre-Associate with Resource Reservation operation is identical to the Pre-Associate operation with the additional step of the bridge reserving its necessary resources in order to increase the probability that a future identical Associate operation will succeed. Additionally, the reservation of resources may further decrease the time required to process a future Associate operation.

It is not necessary to perform a Pre-Associate with Resource Reservation prior to performing an Associate operation.

3.3.3. Associate

The Associate operation creates and activates an associate between the VSI and the bridge port to which it is connected. The bridge allocates the necessary resources to create this association and applies any necessary configuration associated with the VSI Type ID.

[IEEE8021Qbg] does not specify the mechanism by which the bridge determines the resources and configuration required by a VSI Type ID. VSI Type ID simply acts as a handle to identify the configuration information to be retrieved from a repository that is outside the scope of [IEEE8021Qbg].

A station may issue an Associate without having previously issued a Pre-Associate or Pre-Associate with Resource Reservation.

During normal operations a VSI is associated with one bridge port. During network transitions (e.g., virtual station migration) a VSI might be associated with more than one port.

The bridge uses only the information in the Associate operation to establish the association. Any resource reservation that may have

been created based on a previous Pre-Associate or Pre-Associate with Resource Reservation that is not required for the Associate operation is released.

3.3.4. De-Associate

The De-Associate operation removes an association between a VSI and a bridge port. The bridge may de-allocate any resources that were reserved as part of the association.

In addition, a De-Associate operation may be issued to inform a bridge that resources may be de-allocated that were reserved as a result of a previous Pre-Associate or Pre-Associate with Resource Reservation.

A bridge may initiate a De-Associate operation. This could be necessary, for example, in the case of a change in the bridge's configuration or operational status.

3.4. VDP Extensibility

3.4.1. Transport of VDP

ECP is defined in IEEE 802.1Qbg to transport VDP TLVs. It is a simple protocol operating over layer 2. It allows for one PDU to be outstanding at a time. Acknowledgement of an ECP PDU indicates that the PDU contents were received. Processing and responses to TLVs in the PDU can take place after the acknowledgement.

Currently, IEEE 802.1Qbg specifies that the Nearest Customer Bridge group MAC address is used as the destination in ECP PDUs carrying VDP.

NVO3 is likely to want to use a different destination address as the NVE is not necessarily the nearest customer bridge. There have been other protocols that initially required a certain destination address and the requirement was modified when new uses required new addresses. For instance ECP could be used with individual destination addresses instead of a group address.

Alternatively, a different reliable transport could be identified for carrying VDP TLVs for NVO3.

3.4.2 Enhancing the VDP Association TLV

The VDP Association TLV Filter Info is currently specified using layer 2 addressing (MAC address, VLAN, etc.). It is likely that the

IETF would need to extend this to include IPv4 and IPv6 addressing mechanisms and tenant IDs. There are at least two straight forward ways to do this.

The method preferred by the authors would be to request the IEEE to add additional Filter Info formats to cover the needed extensions. There are currently 252 Format identifiers that are reserved for future standardization. With this method there are two alternatives. An IEEE 802.1 project could be initiated to add the additional Filter Info formats to IEEE 802.1Q. Alternatively, IETF could ask IEEE 802.1 to assign some of the Filter Info format identifiers to IETF for definition in an RFC.

Alternatively, the IETF could autonomously define the desired extensions using the Organizationally Defined TLV. The contents of the Organizationally Defined Information Field could be defined by the IETF to be identical to that of the VDP association TLV with the addition of IETF defined Filter Info formats.

3.4.2. Enhancing Migration Support

The VDP TLV contains two status bits to help in migrating state when a VSI is migrating. The M-bit indicates that the VSI is migrating as opposed to a new VSI or one not known to be migrating. The S-bit indicates when the VSI is known to have been suspended for migration. NVO3 could provide guidance on using these bits.

4. Gap Analysis

[I-D.kreeger-nvo3-hypervisor-nve-cp-01] discusses requirements for Hypervisor-to-NVE Control Plane Protocol Functionality. This section summarizes the requirements and describes VDP's ability (or lack thereof) to meet the requirements.

The requirements from [I-D.kreeger-nvo3-hypervisor-nve-cp-01] are summarized in the table below. The column labeled "VDP Support & Additional Discussion" indicates whether VDP supports the requirement. The notation "SBF" in this column indicates that the VDP framework supports the operation; however, and additional Filter Info format or other minor extension is required for a complete implementation. A section number in this column indicates a section in this document that provides additional discussion of the particular requirement and how VDP achieves it.

Paragraph in I-D. kreeger-nvo3-hypervisor-nve-cp-01]	Requirement	VDP Support & Additional Discussion
(4.)	"...identifies the Tenant System (TS) VNIC addresses and VN Name (or ID)..."	SBF 4.1.
(4.)	"...identify a locally significant tag (e.g., an 802.1Q VLAN tag) that can be used to identify the data frames that flow between the TS VNIC and the VN."	Yes 4.2.
(4.1.)	"The NVE must be notified when an End Device requires connection to a particular VN and when it no longer requires connection."	Yes
(4.1.)	"...the external NVE must provide a local tag value for each connected VN to the End Device to use for exchange of packets between the End Device and the NVE (e.g. a locally significant 802.1Q tag value)."	Yes 4.2.
(4.1.)	"The Identification of the VN in this protocol could either be through a VN Name or a VN ID."	Yes 4.3.
(4.2.)	"...the "hypervisor-to-NVE" protocol requires a means to allow End Devices to communicate new tenant addresses associations for a given VNIC within a VN."	Yes
(4.3.)	"When a VNIC within an End Device terminates function..., all addresses associated with that VNIC must be disassociated with the End Device on the connected NVE."	Yes
(4.3.)	"If the VNIC only has a single address associated with it, then this can be	Yes

	a single address disassociate message to the NVE."		
(4.3.)	"...if the VNIC had hundreds of addresses associated with it, then the protocol with the NVE would be better optimized to simply disassociate the VNIC with the NVE, and the NVE can automatically disassociate all addresses that were associated with the VNIC."	SBF 4.4.	
(4.4.)	"...the NVE can make optimizations if it knows which addresses are associated with which VNICs within an End Device and also is notified of state changes of that VNIC, specifically the difference between VNIC shutdown/startup and VNIC migration arrival/departure."	Yes	

4.1. VDP Addressing support

VDP as currently defined is fundamentally layer two. It supports addresses composed of an IEEE 802 style MAC address optionally combined with a VLAN identifier. These addresses are carried in the Filter Info field of the VDP Association TLV, see 3.2.2. The framework of VDP allows for the communication of various formats of the Filter Info field and additional formats may be added that support layer three addresses such as IPv4 and IPv6 addresses, see 3.3. Alternatively, using the organizationally defined TLV mechanism, an IETF defined TLV may be used.

4.2. VDP Support of VLAN Identification

VDP supports two mechanisms for expressing a locally significant tag. One is to express a 802.1Q VLAN ID explicitly. The other is to use a GroupID which has local significance and can be mapped to an actual VLAN by the network controlling entities (see [IEEE8021Qbg] for details

4.3. VDP Support of VN Identification

The GroupID is a 32-bit value that may be used for VN identification. Additional Filter Info formats may be defined to support a GUID or other forms of a name. Alternatively, using the organizationally defined TLV mechanism, an IETF defined TLV may be used.

4.4. Removal of all Addresses Associated with a VNIC

VDP currently does not support the mass removal of all addresses associated with a VNIC. Instead, these must be removed individually. However, such a capability may be defined by creating a Format code that indicates no Filter Info entry is present in the VDP Association TLV. On a de-associate operation, this would indicate the need to remove all addresses.

5. Summary and Conclusions

VDP meets most of the requirements to support the VM-to-NVE control plane protocol. With the addition of a few Filter Info formats, all of the requirements may be met within the framework of VDP.

6. Security Considerations

TBD

7. References

7.1. Normative References

[IEEE8021Qbg] IEEE Std 802.1Qbg(TM)-2012, IEEE Standard for Local and metropolitan area networks-Media Access Control (MAC) Bridges and Virtual Bridged Local Area Networks-Amendment 21: Edge Virtual Bridging.

[I-D.kreeger-nvo3-hypervisor-nve-cp-01] Kreeger, L., Narten, T., and D. Black, "Network Virtualization Hypervisor-to-NVE Overlay Control Protocol Requirements", draft-kreeger-nvo3-hypervisor-nve-cp-01, August, 2013.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4122] Leach, P., Mealling, M., and R. Salz, "A Universally Unique IDentifier (UUID) URN Namespace", RFC 4122, July 2005.

[RFC4291] Hinden, R. and S. Deering, "IP Version 6 Addressing Architecture", RFC 4291, February 2006.

8. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Joseph Pelissier
Cisco
170 Tasman Drive
San Jose, CA 95134
USA

Email: jopeliss@cisco.com

Patricia Thaler
Broadcom Corporation
5300 California Ave
Irvine, California 92617
USA

Email: pthaler@broadcom.com

Paul Bottorff
8000 Foothills Blvd., M/S 1421
Roseville, CA 95747

Email: paul.bottorff@hp.com