VSI Discovery and Configuration

Definitions, Semantics and State Machines

802.1Qbg Presentation

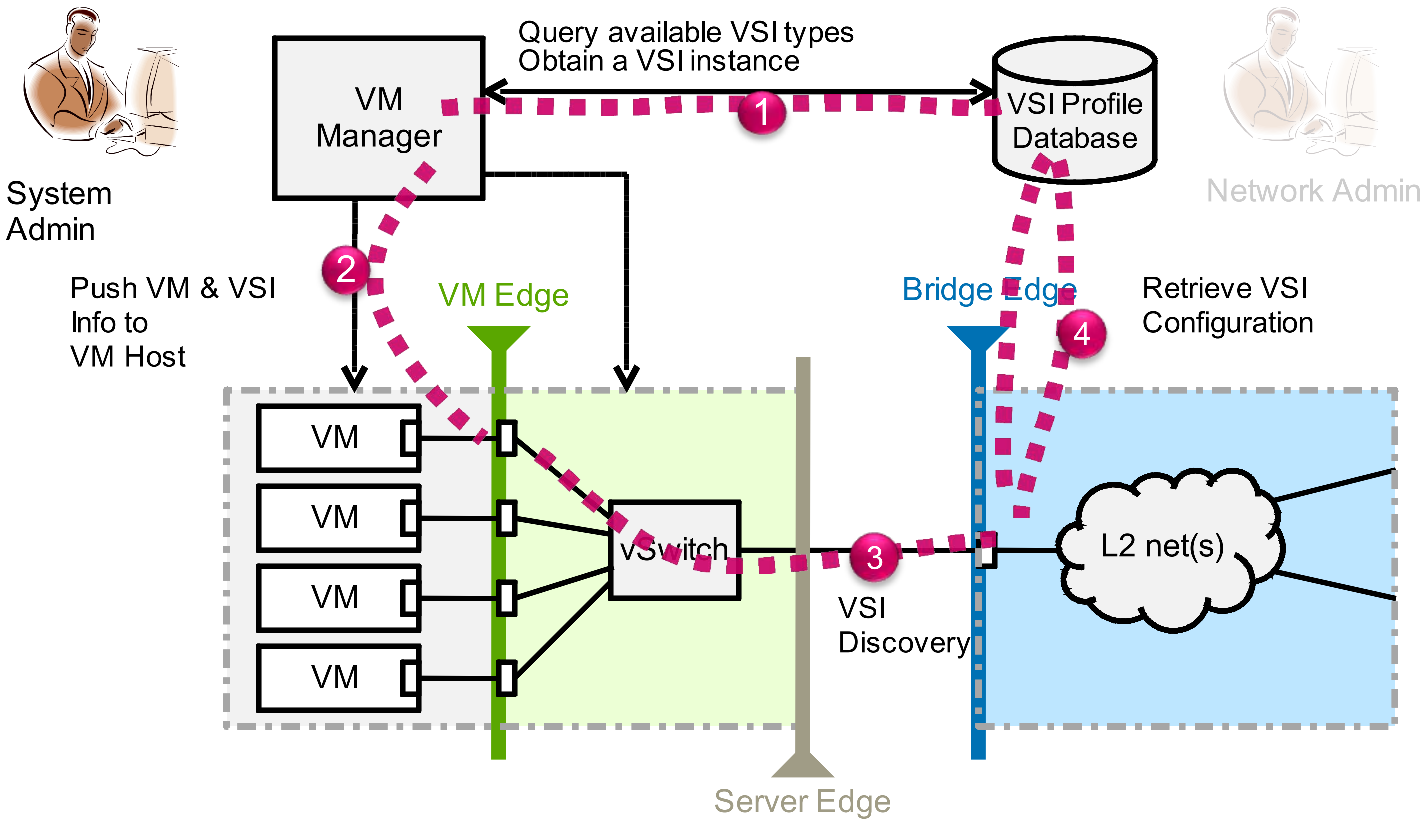
Rakesh Sharma (IBM), Chuck Hudson (HP), Renato Recio (IBM), Manoj Wadekar (Qlogic), Anoop Ghanwani (Brocade), Srikanth Kilaru (Juniper), Vivek Kashyap (IBM), Vijoy Pandey (BNT), Michael Krause (HP)

01/21/2010

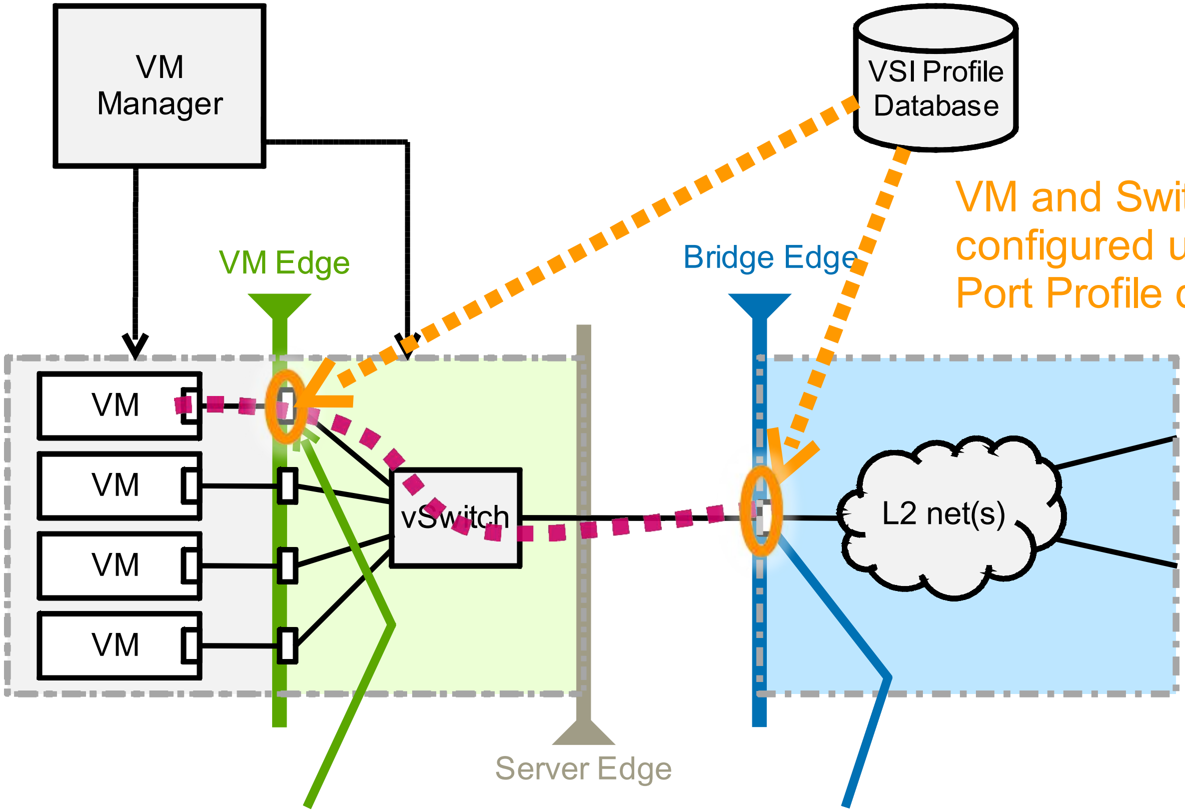
Summary
Proposed EVB-related TLVs

- Multichannel
  - Negotiate whether to use multichannel mode
  - Share enough information to setup the channels
- EVB Discovery
  - Negotiate aggregation mode (VEB, VEPA, etc.)
  - Negotiate whether to use VSI discovery protocol
  - Negotiate whether Hypervisor Authentication is supported
- VSI Discovery/Configuration
  - Announce the arrival status of a (Virtual) Station Interface and communicate information to allow the edge bridge to retrieve the appropriate configuration for the connection.

One Scenario for Configuring Edge Connections (VSIs)



Result: Dynamic & Coordinated Configuration of Edge Connections



VM and Switch edge now configured using the same Port Profile definition!

VM Edge provisions and enforces some connection settings (VLAN, BW, etc.)

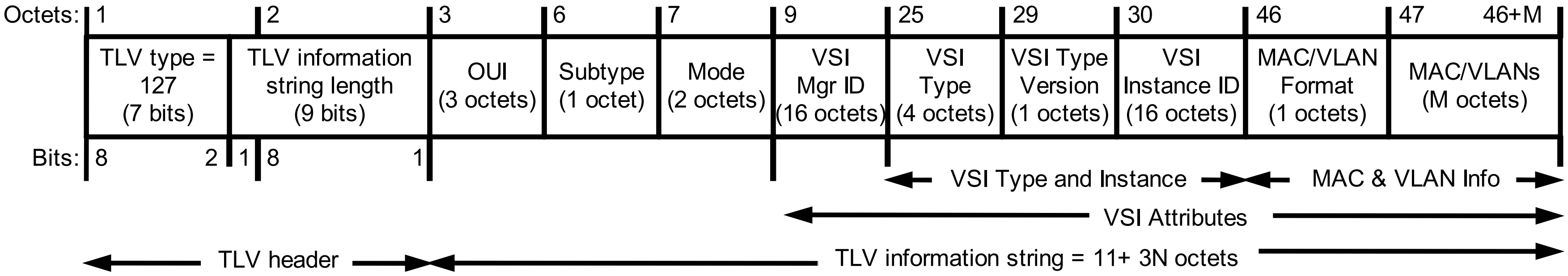
Bridge edge provisions and polices coordinated set of server connection settings.

Port Discovery TLV
Supports Multiple Usage Models

- Database-driven, Instance Aware
  - Driven primarily by Port Profile Domain & Profile Instance
  - Allows database to track instance location, etc.
- Profile-driven
  - Driven primarily by Port Profile ID and Version (Instance ID ignored)
  - Allows for smaller, simplified port profile databases
  - Port Profile definitions may be fully cached on edge switch
- SMB or Lab Config
  - Directly uses VLAN ID contained within vPort discovery LLDP frame to configure VLANs on the edge switches

Proposed VSI Discover/Configuration TLV

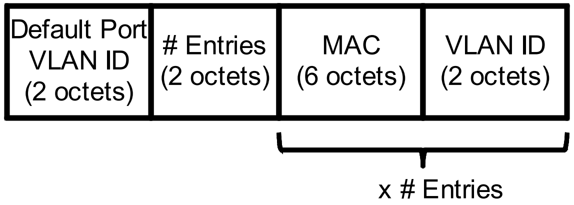
Transport: TBD



Octets:	1	2	3	6	7	9	25	29	30	46	47 46+M
	TLV type = 127 (7 bits)	TLV information string length (9 bits)	OUI (3 octets)	Subtype (1 octet)	Mode (2 octets)	VSI Mgr ID (16 octets)	VSI Type (4 octets)	VSI Type Version (1 octets)	VSI Instance ID (16 octets)	MAC/VLAN Format (1 octets)	MAC/VLANs (M octets)
Bits:	8 2	1 8 1									

- Mode – Indicates VSI TLV Mode
  - First octet identifies a pre-associate, associate, de-associate, or the corresponding confirmation or rejection for each.
  - Second octet is used during a rejection to indicate the reason for the pre-assoc or assoc rejection.
- Port Manager ID – Identifies the Port Manager with the Database that holds the detailed port/VSI type and or instance definitions. May be the IP address of the management server.
- Port Type ID (PTID)– The integer identifier of the port/VSI profile type.
- Port Type ID Version – The integer identifier designating the expected/desired version of the PTID.

- VSI Instance ID – A globally unique ID for the connection instance. The ID shall be done consistent with IETF RFC 4122.
- Format – identifies the format of the MAC and VLAN information that follows in the TLV.
- MAC/VLANs – Listing of the MAC/VLANs associated with the Port Instance (VSI). If Format =1, this would be a simple listing of MAC/VLAN pairs as shown below.

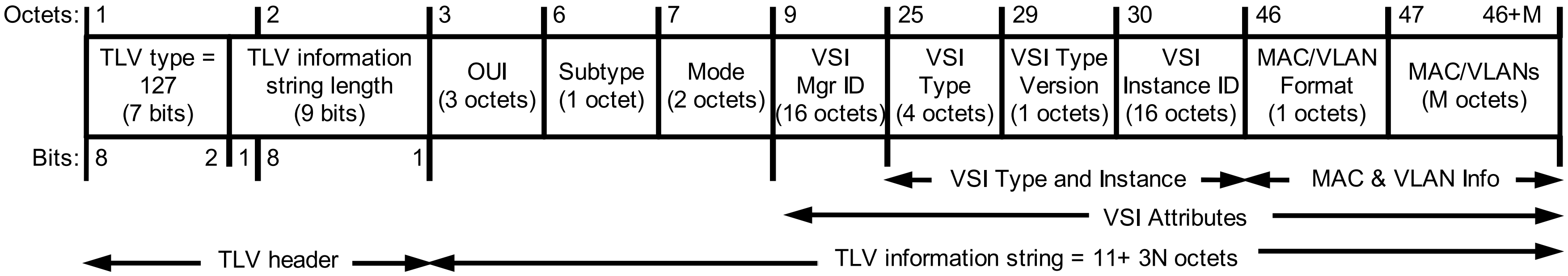


Default Port VLAN ID (2 octets)	# Entries (2 octets)	MAC (6 octets)	VLAN ID (2 octets)

- If Format =2, this would be a simple listing of MAC/VLAN/VSI-state-map-offset pairs (adds 2 octets to end of above, for the VSI state map).

NOTE: The station and bridge environments and their common understanding of the meaning of a port profile ID is outside the scope of this effort

Notes on VSI TLV



Octets:	1	2	3	6	7	9	25	29	30	46	47 46+M
	TLV type = 127 (7 bits)	TLV information string length (9 bits)	OUI (3 octets)	Subtype (1 octet)	Mode (2 octets)	VSI Mgr ID (16 octets)	VSI Type (4 octets)	VSI Type Version (1 octets)	VSI Instance ID (16 octets)	MAC/VLAN Format (1 octets)	MAC/VLANs (M octets)
Bits:	8 2	1 8 1									

VSI Type and Instance

MAC & VLAN Info

VSI Attributes

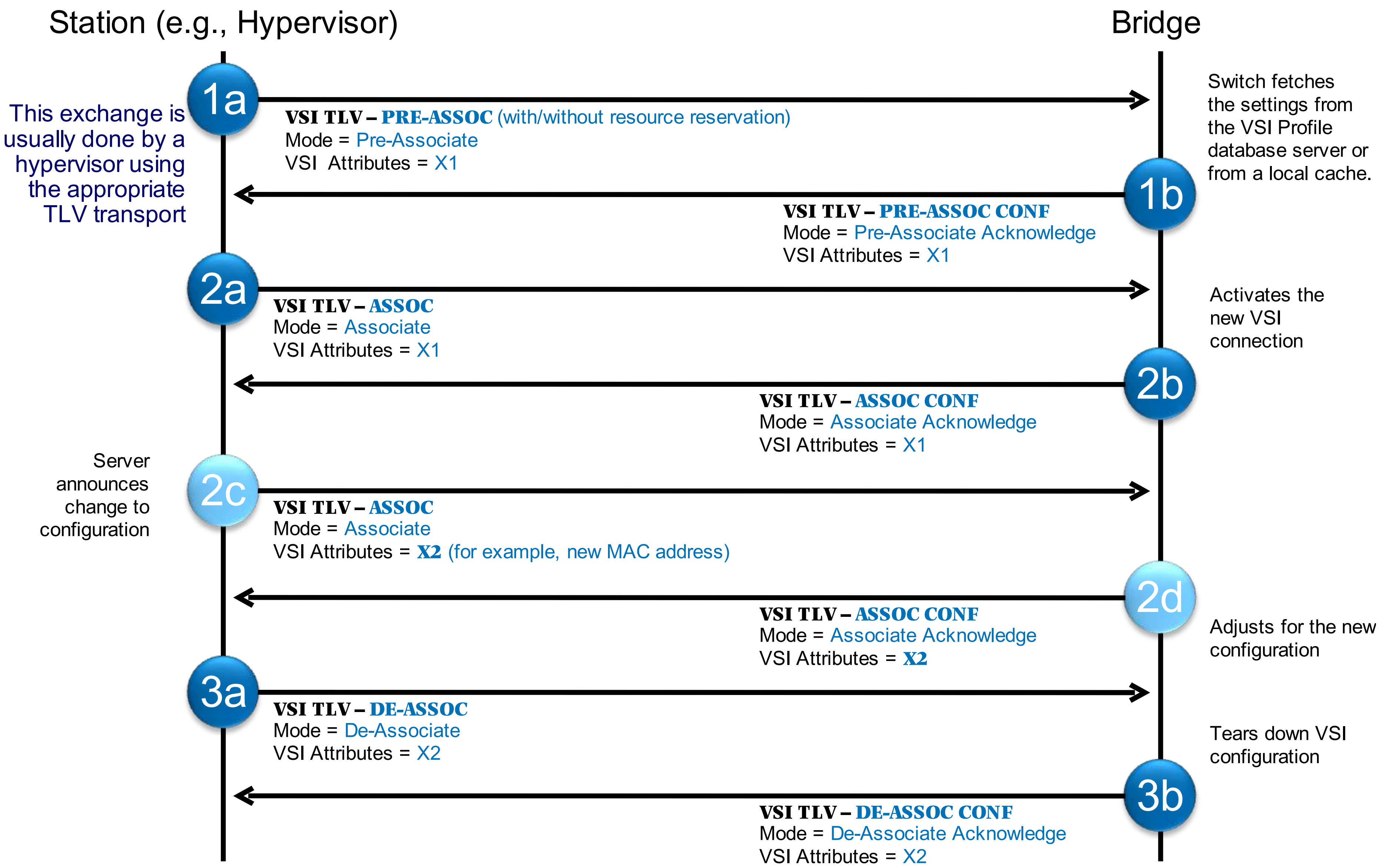
TLV header

TLV information string = 11+ 3N octets

Notes:

- The station and switch environments and their common understanding of the PTID meaning is outside the scope of this TLV.
- VSI TLV is transported in LLDP/T3P PDU. LLDP/T3P PDUs use Physical Station MAC Address (e.g. Hypervisor MAC).
- LLDP/T3P PDUs carry Chassis ID TLV for the Physical Station (Hypervisor).
- The Physical Station port’s VLAN ID uses the VLAN TLV in the same transport (LLDP or T3P) PDU and is not contained in this TLV.
- Format field - VSI TLV allows multiple formats of this information to optimize frame space usage and functionality. Further, it makes possible extensions in future.

VSI Discovery/Config. TLV Example



Station (e.g., Hypervisor)

Bridge

This exchange is usually done by a hypervisor using the appropriate TLV transport

1a

VSI TLV – PRE-ASSOC (with/without resource reservation)
Mode = Pre-Associate
VSI Attributes = X1

Switch fetches the settings from the VSI Profile database server or from a local cache.

1b

VSI TLV – PRE-ASSOC CONF
Mode = Pre-Associate Acknowledge
VSI Attributes = X1

2a

VSI TLV – ASSOC
Mode = Associate
VSI Attributes = X1

Activates the new VSI connection

2b

VSI TLV – ASSOC CONF
Mode = Associate Acknowledge
VSI Attributes = X1

Server announces change to configuration

2c

VSI TLV – ASSOC
Mode = Associate
VSI Attributes = **X2** (for example, new MAC address)

2d

VSI TLV – ASSOC CONF
Mode = Associate Acknowledge
VSI Attributes = **X2**

Adjusts for the new configuration

3a

VSI TLV – DE-ASSOC
Mode = De-Associate
VSI Attributes = X2

Tears down VSI configuration

3b

VSI TLV – DE-ASSOC CONF
Mode = De-Associate Acknowledge
VSI Attributes = X2

VSI Discovery TLV - Mode and Mode Response

- Mode field purpose: Identifies VSI Discovery TLV type:
  - VSI TLV Request field: 1st octet

Pre-Associate:	0x00
Pre-Associate with resource reservation:	0x01
Associate:	0x02
De-Associate:	0x03

  - VSTI TLV Response field: 2nd octet

Success:	0x00
Invalid Format:	0x01
Insufficient PT Resources:	0x02
Unused PTID:	0x03
PTID Violation:	0x04
PTID Version Violation:	0x05
Associate received out of order:	0x06
Out of Sync:	0x07
Duplicate Associate	0x08
Reserved	

- Usage:
  - Used under the control of VDP state machines.

Mode Requests and Responses

- Requests:
  1. Pre-Associate
     - Pre-associate VSI Instance Identifier to a PTID.
     - Validate parameters.
     - Notify bridge to prep for Association.
  2. Pre-Associate with resource reservation.
     - Same as Pre-Associate. Reserves resources in addition.
  3. Associate –
     - Associate VSI Instance Identifier to a PTID.
     - Allow resources are allocated and VSI is active.
  4. De-Associate
     - De-associate a VSI Instance Identifier from the associated PTID.
- Responses:
  - Each of the above Modes has an associated Response.

Pre-associate (0x0000) Semantics

- Pre-Associate VSI Instance Identifier to a PTID
- If required, should obtain Port Type Definition from the Port Manager Database.
- Validate the request and fail it in case of errors.
- Successful Pre-Association does not enable any traffic from VSI.
  - Note that VSI may still be associated at another station.
- Pre-association is required step.
- Makes Associate response faster. Important for VM mobility and failover.

Pre-associate Completion (0x00nn) Semantics

- Second Mode octet contains the results of the Pre-Associate request performed for the VSI Identifier.
  - Success x0000 - Pre-Associate was successful. The switch shall permit a subsequent Associate or De-Associate by the VSI referenced by the VSI Identifier.
  - The following are all unsuccessful Pre-Associate Completions. For each of these, the switch shall not permit a subsequent Associate or De-Associate by the VSI referenced by the VSI Identifier.
    - Invalid Format 0x0001 - The VSI Format is not supported by the switch.
    - Insufficient PT Resources 0x0002 - The switch does not have enough resources to complete the Pre-Association successfully.
    - Unused PTID 0x0003 - The VSI referenced by the VSI Identifier does not exist in the Port Manager database referenced by the Port Manager Identifier.
    - PTID Violation 0x0004 - The VSI referenced by the VSI Identifier is not allowed to be associated with the PTID.
    - PTID Violation 0x0005 - The VSI referenced by the VSI Identifier is not allowed to be associated with the PTID Version.

Pre-Associate with resource reservation (0x0100) Semantics

- Pre-association of a VSI Instance Identifier to a Port Type Identifier
  - Same steps as Pre-Associate
  - Additionally:
    - Bridge should validate required resources and place reservation.
    - Enable pre-Associate timer to conserve resources.
- Does not allow any traffic from VSI.
- Pre-association or Pre-Association with resource reservation is required step.

Pre-associate with Resource Reservation Completion Semantics (0x01nn)

- Second Mode octet contains the results of the Pre-Associate with Resource Reservation request performed for the VSI Identifier.
  - Success 0x0100 - Pre-Associate was successful. Prior to issuing this response, the switch shall reserve resources for use in a subsequent Associate or De-Associate by the VSI referenced by the VSI Identifier.
  - The following are all unsuccessful Pre-Associate Completions. For each of these, the switch shall not permit a subsequent Associate or De-Associate by the VSI referenced by the VSI Identifier.
    - Invalid Format 0x0101 - The VSI Format is not supported by the switch.
    - Insufficient PT Resources 0x0102 - The switch does not have enough resources to complete the Pre-Association successfully.
    - Unused PTID 0x0103 - The VSI referenced by the VSI Identifier does not exist in the Port Manager database referenced by the Port Manager Identifier.
    - PTID Violation 0x0104 - The VSI referenced by the VSI Identifier is not allowed to be associated with the PTID.
    - PTID Violation 0x0105 - The VSI referenced by the VSI Identifier is not allowed to be associated with the PTID Version.

Associate (0x02) Semantics

- Sets up the switch port for the VSI/Port Type and Instance
  - Allocates required bridge resources for the VSI/Port.
  - Binds specific MAC/VLAN pairs with the VSI/Port
  - Activates the switch configuration for the VSI/Port
- Associate requires Pre-Associated or Pre-Associated with Resource Reservation VSIs.
  - If the switch receives an Associate before the switch has issued a Pre-Associate or Pre-Associate with Resource Reservation Successful Completion, then the Associate shall complete in error and the VSI exchange shall be terminated.
  - Same VSI may not be successfully Associated more than once.

Associate Completion (0x02nn) Semantics

- Second Mode octet contains the results of the Associate request performed for the VSI Identifier.
  - Success x0200 - Associate was successful. Prior to issuing this response, the switch shall:
    - For a Format 1 TLV, associates the Port Type referenced by the Port Type Identifier and Port Type Version with the MAC Address, VLAN and VSI Identifier.
  - The following are all unsuccessful Associate Completions. For each of these, the switch shall not permit a subsequent De-Associate by the VSI referenced by the VSI Identifier.
    - Invalid Format 0x0201 - The VSI Format is not supported by the switch.
    - Insufficient PT Resources 0x0202 - The switch does not have enough resources to complete the Association successfully. If the Associate was preceeded by a successful Pre-Associate with Resource Reservation, then the switch shall not issue this request.
    - PTID Violation 0x0204 - The VSI referenced by the VSI Identifier is not allowed to be associated with the PTID.
    - PTID Violation 0x0205 - The VSI referenced by the VSI Identifier is not allowed to be associated with the PTID Version.
    - Associate Received Out of Order 0x0206 - The switch received the Associate prior to the successful completion of a Pre-Associate or Pre-Associate with Resource Reservation for the VSI referenced by the VSI Identifier.
    - Out of Sync 0x0207 - The PTID or one of the VSI List fields used in the Associate is not the same as the corresponding field used in the Pre-Associate.

De-Associate (0x03) Semantics

- De-Associate VSI Instance Identifier from a PTID.
  - Pre-Associated and Associated VSIs can be deAssociated.
  - De-Associate releases resources and de-activates VSI configuration
  - VSI may get De-Associated by bridge due to bridge error situation or management action.

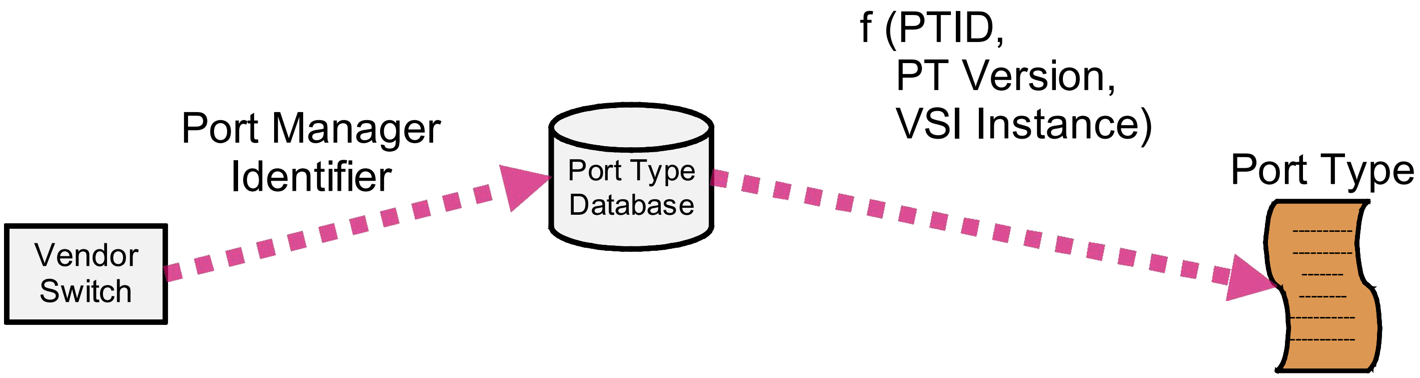
De-Associate Completion (0x03nn) Semantics

- Second Mode octet contains the results of the De-Associate request performed for the VSI Identifier.
  - Success x0300 - De-Associate was successful. Prior to issuing this response, the switch shall:
    - For a Format 1 TLV, de-associates the Port Type referenced by the Port Type Identifier from the the MAC Address, VLAN and VSI Identifier.
  - The following are all unsuccessful De-Associate Completions
    - Invalid Format 0x0301 - The VSI Format is not supported by the switch.
    - PTID Violation 0x0304 - The VSI referenced by the VSI Identifier is not allowed to be de-associated with the PTID.
    - PTID Version Violation 0x0305 - The VSI referenced by the VSI Identifier is not allowed to be de-associated with the PTID Version.

Note: The result of the above semantics is that De-Associate can be issued at any time.

Port Manager Identifier Semantics

- Definition: Identifies the Port Manager with the Database that holds the detailed port/VSI type and or instance definitions.
  - The contents of the Port Manager Database are outside the scope of this specification.
  - The Port Manager Database may use a combination of the following fields to index into the Port Manager Database:
    1. Port Type Identifier
    2. Port Type Version
    3. VSI Instance



Port Type Identifier Semantics

- Definition: Integer value field used to identify a pre-configured set of controls/attributes that are to be associated with a set of VSIs.
  - PTID contents/meaning and the database used to contain the Port Profiles are outside the scope of this effort.
  - One PTID may be describe the port profile configuration of multiple VSIs.
  - The Port Type content referenced by the same PTID may differ between switches and VEBs. For example:
    - Same PTID is used by switches from two different vendors.
    - Same PTID is used by a VEB and vendor switches.

Port Type Identifier Version Semantics

- Definition: The integer identifier designating the expected/desired PTID.
  - The PTID Version enables a Port Manager Database to contain multiple Port Type versions.
  - Allows smooth migration to newer port types.

VSI Instance ID

- Purpose: A globally unique ID for the VSI instance. The ID shall be done consistent with IETF RFC 4122 VSI ID is unique.

Format

- Definition: Used to identify the VSI’s MAC information.
  - Format 1: 0x00
  - Format 2: 0x01

VSI List for Formats

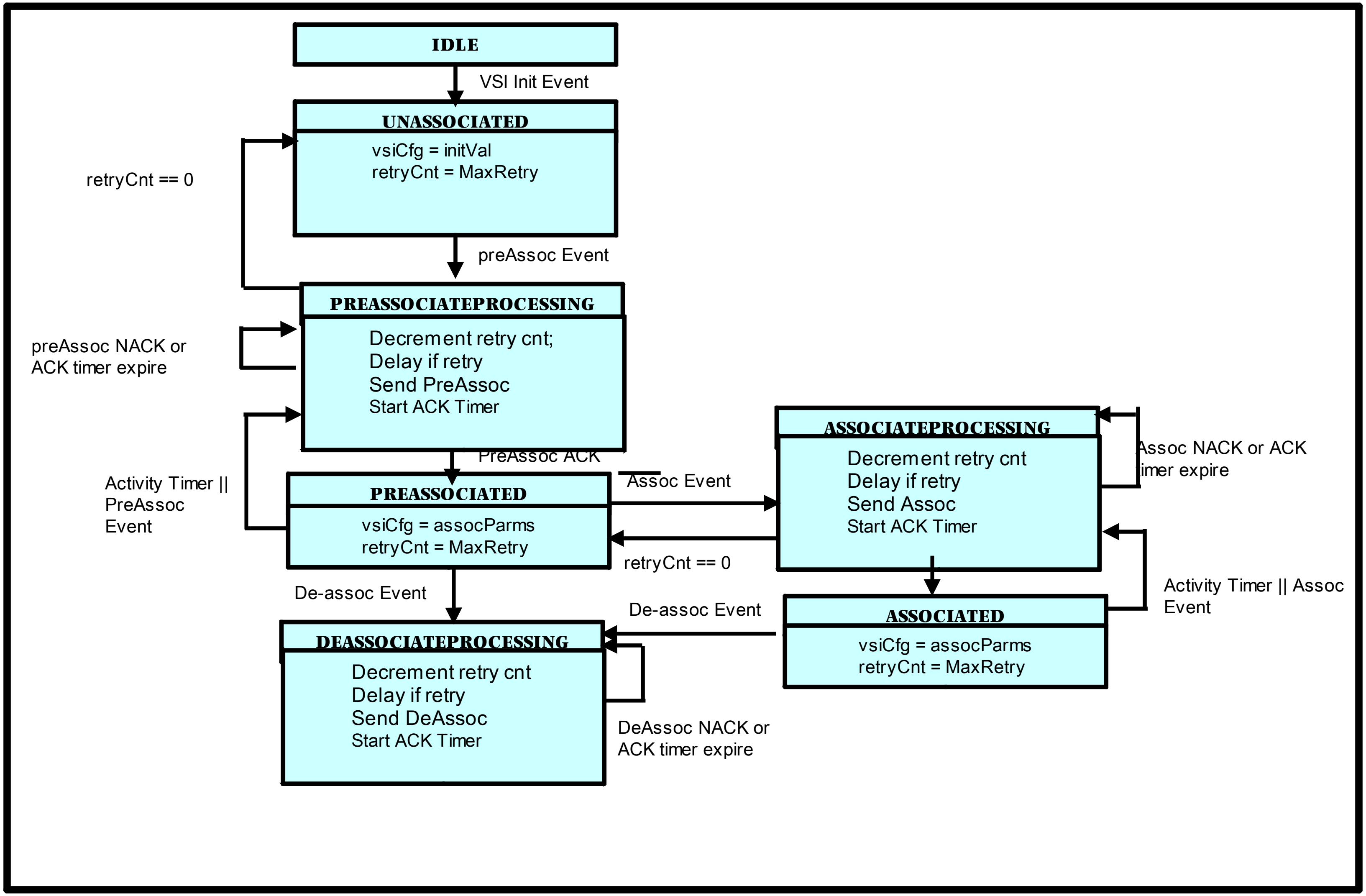
- Format 1
  - Definition: Contains the Port VLAN ID and set of MAC Addresses and VLANs to be associated with the VSI.
    - Note the bridge uses MAC+VID to identify traffic from VSI and to steer the frames.
  - Field:
    - PVID: 2 octets
    - #MAC-VLAN pairs: 2 octets

      Per Pair Content:
      - MAC address: 48 bits
      - VID: 12 bits
- Format 2
  - Definition: Contains the Port VLAN ID and set of MAC Addresses and VLANs to be associated with the VSI.
    - Note the bridge uses MAC+VID to identify traffic from VSI and to steer the frames.
  - Field:
    - Offset into VSI state map: 2 octets
    - PVID: 2 octets
    - #MAC-VLAN pairs: 2 octets

      Per Pair Content:
      - MAC address: 48 bits
      - VID: 12 bits

VSI State Machine – Hypervisor

(One Instance per VSI)



IDLE

VSI Init Event

UNASSOCIATED

vsiCfg = initVal
retryCnt = MaxRetry

retryCnt == 0

preAssoc Event

PREASSOCIATEPROCESSING

Decrement retry cnt;
Delay if retry
Send PreAssoc
Start ACK Timer

preAssoc NACK or
ACK timer expire

PreAssoc ACK

Activity Timer ||
PreAssoc
Event

PREASSOCIATED

vsiCfg = assocParms
retryCnt = MaxRetry

Assoc Event

retryCnt == 0

ASSOCIATEPROCESSING

Decrement retry cnt
Delay if retry
Send Assoc
Start ACK Timer

Assoc NACK or ACK
timer expire

Activity Timer || Assoc
Event

ASSOCIATED

vsiCfg = assocParms
retryCnt = MaxRetry

De-assoc Event

De-assoc Event

DEASSOCIATEPROCESSING

Decrement retry cnt
Delay if retry
Send DeAssoc
Start ACK Timer

DeAssoc NACK or
ACK timer expire

Station VSI State Machine Definitions

State	Description
IDLE	VSI not operational, No resources
UNASSOCIATED	VSI created but no configuration and resources
PREASSOCIATED	VSI is configured and optionally has resources assigned but is not active.
ASSOCIATED	VSI is configured (PTID), has resources and active.
XXXPROCESSING	VSI TLV Exchange with reliability.

Events	Source	Description
TLV Rx	Bridge	TLV received from transport
TLV Tx	Station	TLV Tx to station over the transport
Activity Timer Expire	Local	Timer cause VSI TLV TX to keepalive VSI.
VSI Init Event	Local	VM Manager/Hypervisor instantiates VSI
PreAssoc Event	Local	VM Manager/Hypervisor initiates VSI PreAssociate
Assoc Event	Local	VM Manager/Hypervisor initiates VSI Associate

Notes:

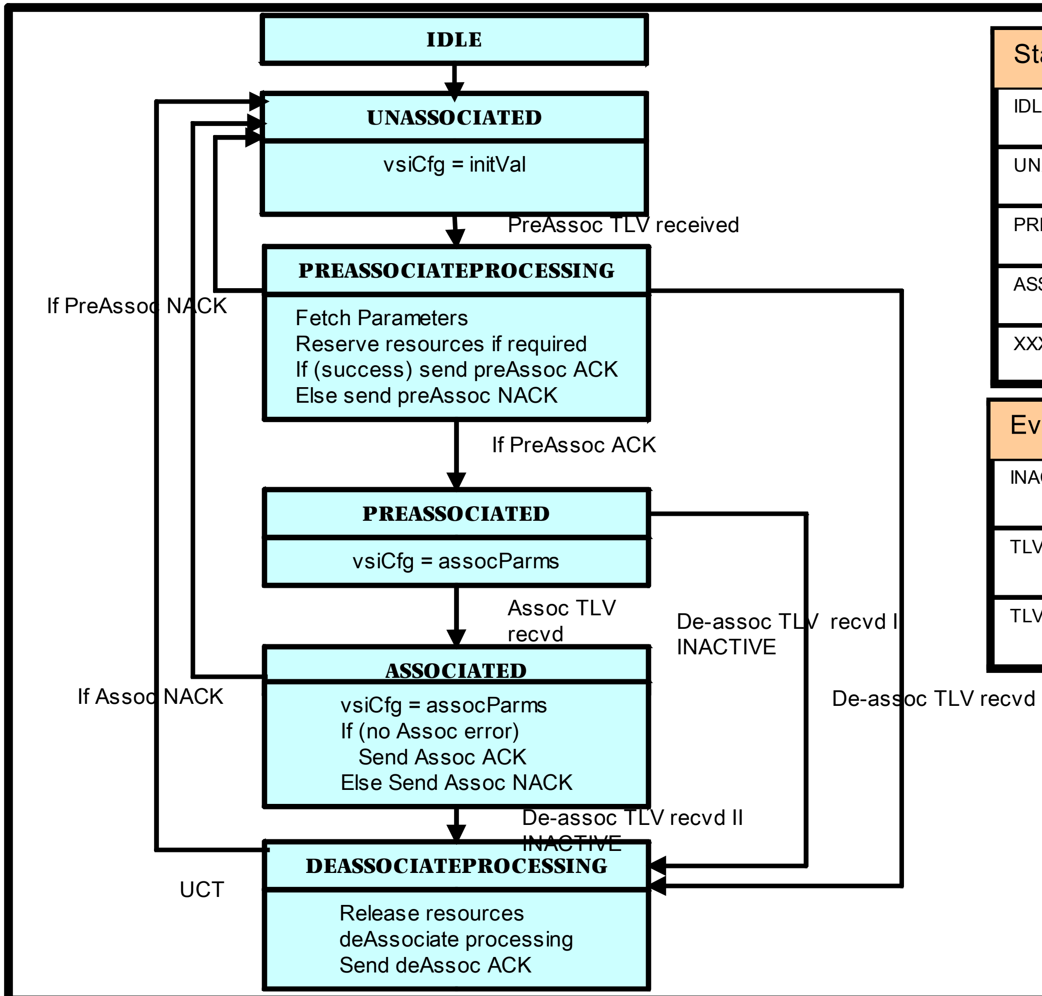
1. Station VSI State Machine supports acknowledged TLV exchanges
2. Performance optimization
3. Resource conservation on bridge and station side
4. Focus on simplicity specially for bridge side implementation.
5. Station and Bridge periodic synchronization is built in state machine

Periodic Synchronization of State and Configuration.

- Option1: Periodic TLV Exchanges synchronizing state and configuration
  - Advantages:
    - Integrated with State Machine
    - Handles new and deleted VSIs and configuration changes efficiently.
    - Accuracy in conveying current state and configuration.
    - Scales with number of VSIs
  - Disadvantages
    - VSI TLV is exchanged and carries complete state and configuration information.
- Option 2: Send bitmap containing Acknowledged state (i.e. previously Sync'd) of all VSI instances and digest of all VSIs to indicate current configuration and state information.
  - Advantages:
    - Reduces data to be transferred on periodic transmissions.
  - Disadvantages:
    - New and deleted VSIs can fragment bit map making it inefficient and complex.
    - Digest based configuration change is not exact and could cause issues.
    - Number of VSIs get limited be bit-map size which needs to be determined in advance.

Note: VSI State Machines show option 1

VSI State Machine – Adjacent Bridge
(One Instance per VSI)



IDLE

UNASSOCIATED

vsiCfg = initVal

PreAssoc TLV received

PREASSOCIATEPROCESSING

Fetch Parameters
Reserve resources if required
If (success) send preAssoc ACK
Else send preAssoc NACK

If PreAssoc NACK

If PreAssoc ACK

PREASSOCIATED

vsiCfg = assocParms

Assoc TLV recvd

De-assoc TLV recvd ll INACTIVE

De-assoc TLV recvd

ASSOCIATED

vsiCfg = assocParms
If (no Assoc error)
  Send Assoc ACK
Else Send Assoc NACK

If Assoc NACK

De-assoc TLV recvd ll INACTIVE

DEASSOCIATEPROCESSING

Release resources
deAssociate processing
Send deAssoc ACK

UCT

State	Description
IDLE	VSI not operational, No resources
UNASSOCIATED	VSI created but no configuration and resources
PREASSOCIATED	VSI is configured and optionally has resources assigned but is not active.
ASSOCIATED	VSI is configured (PTID), has resources and active.
XXXPROCESSING	VSI TLV Exchange with reliability.

Event	Source	Description
INACTIVE	Local	Inactive VSI
TLV Rx	Station	TLV received from transport
TLV Tx	Bridge	TLV Tx to station over the transport

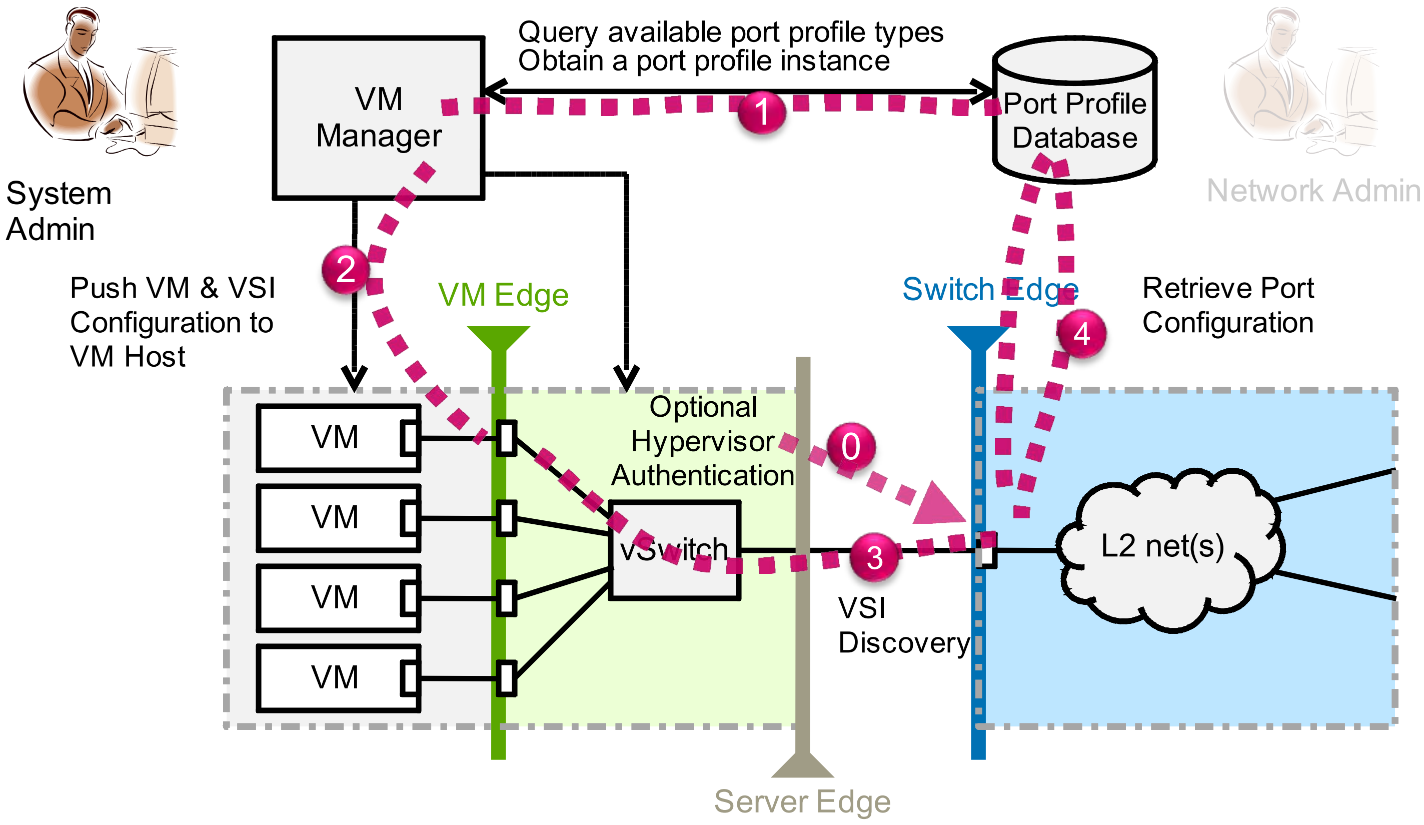
VSI Discovery and Configuration Protocol (VDP) Module (Implementation Note)

VSI	VSI TLV Fields	VSI state	VSI Timer Ticks	VSI statistics
VSI0	VSI0 TLV fields	VSI0 state	timerTicks	TBD
VSI1	VSI1 TLV fields	VSI1 state	timerTicks	TBD
VSIn	VSIn TLV fields	VSIn state	timerTicks	TBD

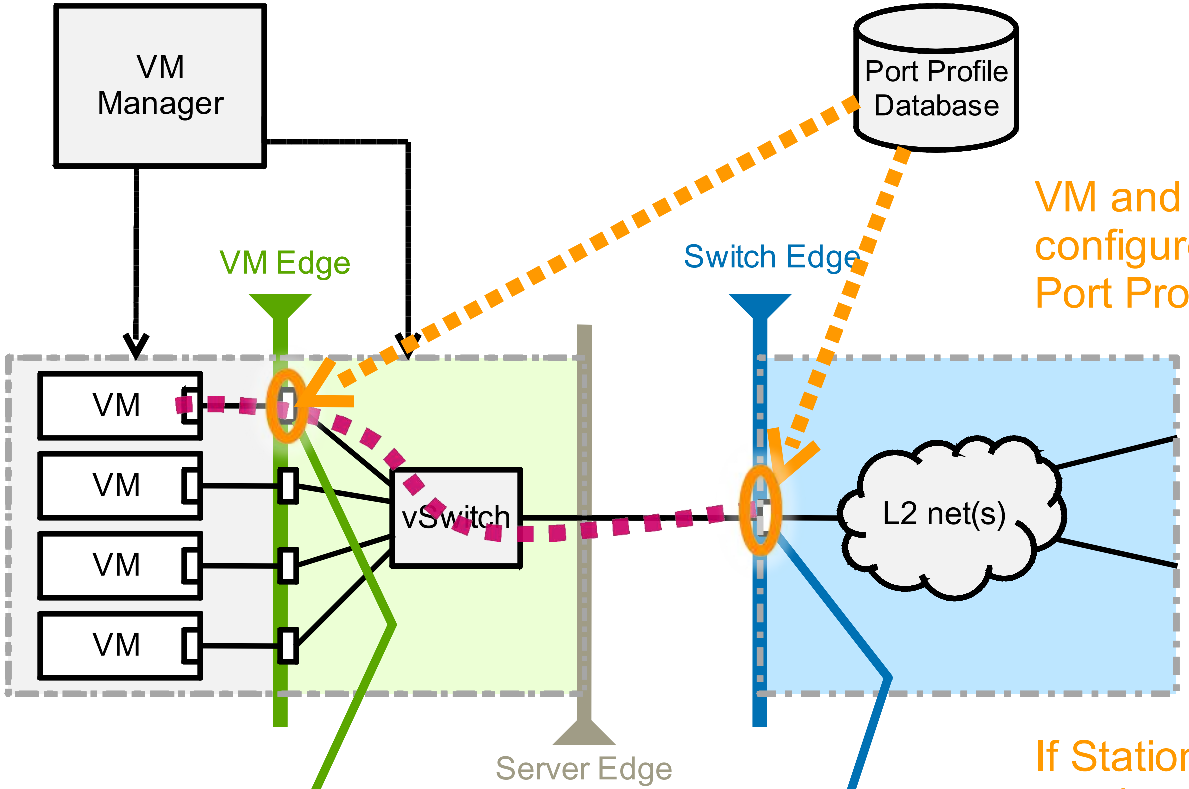
- VDP module implementation can be single module that supports all VSIs that
  - Contains table of VSI variables
  - timerTicks updated by single timer with long duration (for example, every 10 or more seconds
  - VDP module get requests (create VSI, pre-Assoc, Assoc and de-Assoc) from VDP user (Hypervisor/Bridge OS) and sends events to the user.
  - Transmits TLVs by queueing to T3P module and receives TLVs from T3P for processing.

Usage Examples

Steps for Configuring Edge Connections (vPorts)


System Admin
VM Manager
Query available port profile types
Obtain a port profile instance
1
Port Profile Database
Network Admin
2
Push VM & VSI Configuration to VM Host
VM Edge
Switch Edge
Retrieve Port Configuration
4
VM
VM
VM
VM
Optional Hypervisor Authentication
0
vSwitch
3
VSI Discovery
L2 net(s)
Server Edge

Result: Coordinated Configuration of Edge Connections (vPorts)



VM and Switch edge now configured using the same Port Profile definition!

VM Edge provisions basic connection settings (VLAN, BW, etc.)

Switch edge provisions and polices full set of server connection settings.

If Station authentication was used, network had verified the Station is authorized to use the Port Type exchanged in the VSI TLV!

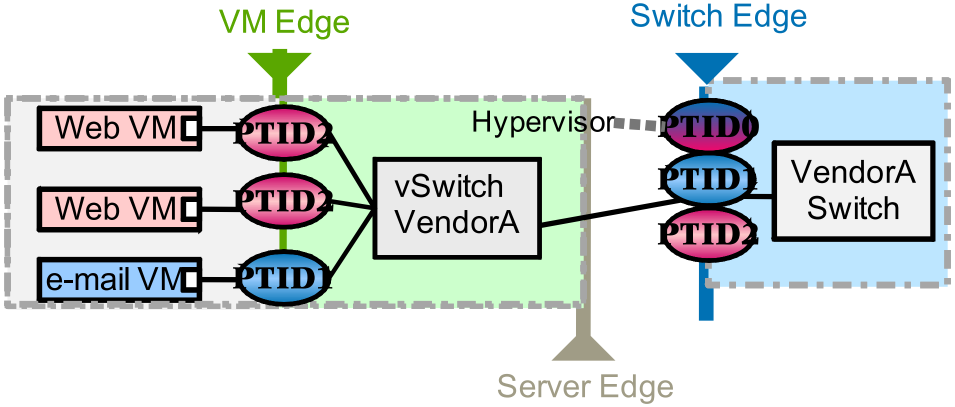
Port Type Identifier Version Examples

- The PTID Version allows multiple Port Type versions, examples include:
  - Enables database to be organized such that all versions of a Port Type have a common set of content, as well as version dependent content.
  - Enables differing generations of Port Types, where each generation is defined by a PTID version.

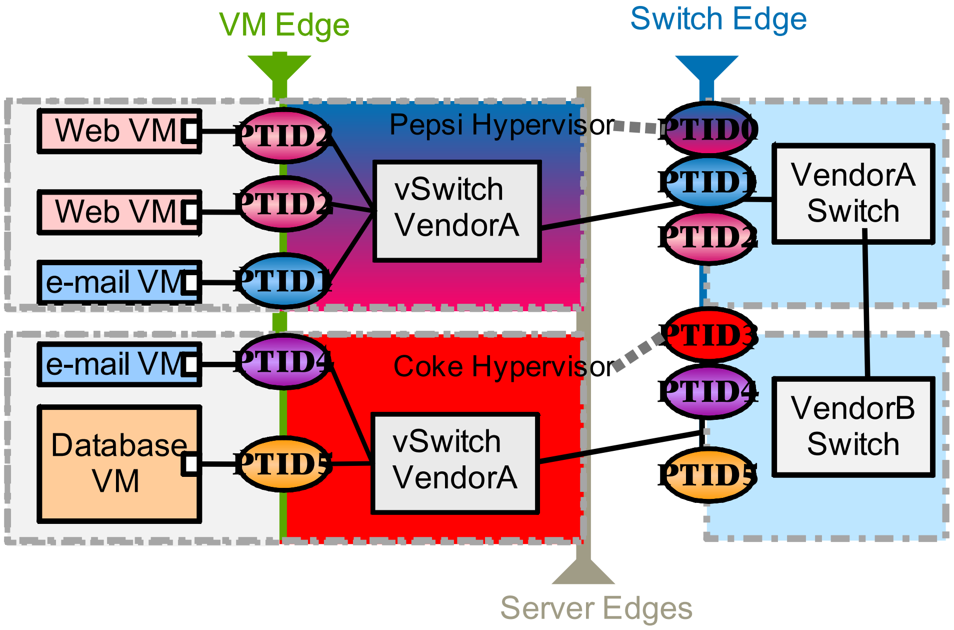
Port Type Identifier Usage Examples

- Usage example, stations in same layer-2, where the Stations are:

1. Hosting different applications.

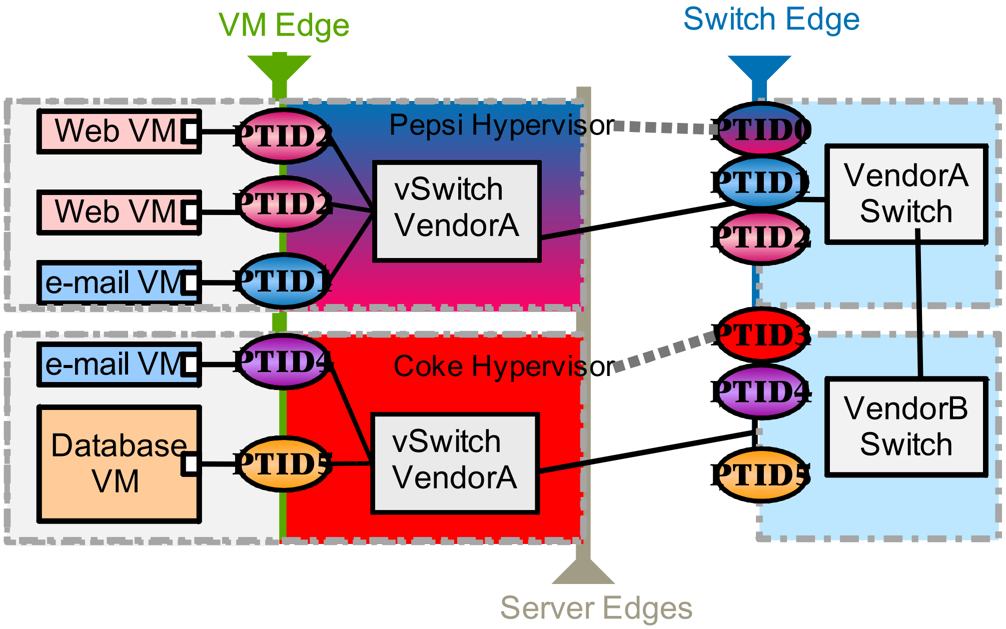


2. Operating in a multi-tenacy environment.



Port Manager Identifier Usage Example

- Stations in same layer-2, where the Stations are used to host cloud Applications from multiple companies and each company is given Hypervisor management control.
  For example:
  - Two Hypervisor Domains (Coke and Pepsi) within the same layer-2 fabric, where each Hypervisor Domain is associated with a range of Port Types.
  - In this scenario, the Pepsi Hypervisor must not be able to associate VMs to the Coke Hypervisor PTIDs.

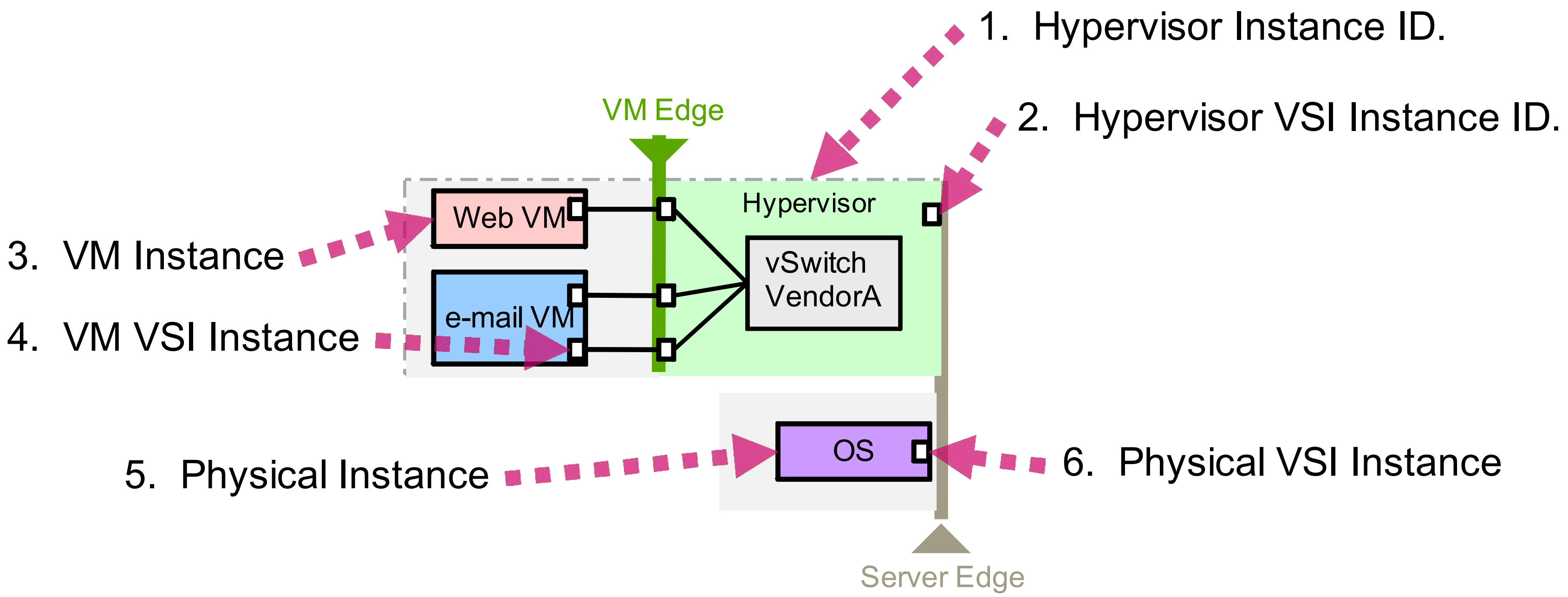


Permissible association for this example

PTID	Hypervisor
0	Pepsi Hyp.
1	Pepsi Hyp.
2	Pepsi Hyp.
3	Coke Hyp.
4	Coke Hyp.
5	Coke Hyp.

VSI Instance ID Usage Examples

- Example uses of the VSI Instance ID include:
  1. A specific Hypervisor instance ID.
  2. A specific Hypervisor VSI instance (e.g. Virtual NIC port) ID.
  3. A specific VM instance ID.
  4. A specific VM VSI instance (e.g. Virtual NIC port) ID.
  5. A specific physical instance ID.
  6. A specific physical VSI instance ID.



Format Field Usage Examples

- Hypervisor can associate VSI to new VLAN dynamically without requiring profile updates. Specially useful for Lab and SMB environments.
- Support VM's connectivity to multiple VLANs e.g. Web tier, Accounting Dept VLAN and Public VLAN.

Port Manager Informative Text

- Authentication of the Hypervisor’s use of PTID is not required.
- If authentication of a Hypervisor’s PTID usage is supported:
  - The PPD database must contain a mapping of:
    - Hypervisor to PTIDs
    - Hypervisor and VM to PTIDs
  - The PPD database mapping must enable:
    - The same PTID mapped to different Hypervisors.
    - The same PTID mapped to different Hypervisor and VM.
  - The Hypervisor authentication mechanism is outside the scope of this effort.
  - The same Station may host two different Hypervisor types, in which case, both Hypervisor must be authenticated.