VSI Discovery and Configuration Protocol (VDP)

Proposed Resolutions to subset of comments against
802.1Qbg Draft-1

Chait Tumuluri (Emulex), Jeffrey Lynch (IBM), Vijoy Pandey (BNT), Rakesh Sharma (BM),
Renato Recio (IBM), Srikanth Kilaru (Juniper)

Version 01

07/12/2010

Introduction

- Presentation “802.1Qbg VSI Discovery and Configuration State Machine” by Paul Bottorff and Paul Congdon proposes a new VDP state machine as proposed resolution to comments against IEEE 802.1Qbg Draft-1.
- We observe that comments do not identify any functional issue or defect.
  - Change need some value associated with them.
  - Comments can be resolved without recreating the state machine.
- This presentation contains proposed resolutions which include explanation, descriptive text addition and editorial updates to existing (IEEE 802.1Qbg Draft-1) VDP state machines.

Bridge Resourcing States

- ***Originator’s Description:***
  - The bridge state machine must hand the command to another machine which may query network management system before sending a response.
  - The Bridge state machine needs to break the command request into two states to allow the command hand-off
    - Need a timeout for the resource request from the Bridge
    - Need to build the response for the station from the resource request response
- ***Resolution Proposal:***
  - IEEE 802.1Qbg Draft-1 state machine does that by incorporating two states, e.g.: Pre-Associate Processing and Pre-Associated. Additional description text to be added for the function ProcRxandSetCfg().

Reverting to Last Successful Associate

- ***Originator's Description:***
  - The current VDP machines reverts to the previous successful associate in the event of an associate failure.
  - Even though the machine includes the transition they are not very specific about retaining the existing state.
- ***Resolution Proposal:***
  - In draft 1, if, while in associate, there is a mismatch, the state machine keeps the VSI in Associate State, see Associate-Processing to Associate: (Assoc_NAK_Rx && VsiState== Assoc). Additional descriptive text to be added.

Local Change Events

- ***Originator's Description:***
  - Current VDP state machines uses the shorthand term "localChange-" to indicate a local event from the Hypervisor or the Bridge control plane.
  - Other 802.1 state machines use variable tests to identify events.
- ***Observation:***
  - IEEE 802.1Qbg Draft-1 use of "localChange-" to indicate a local event is consistent with LLDP and DCB-X.

Command-Response TLV Fields

- ***Originator's Description:***
  - State machine may references TLV fields as follows:
    - operCmd TLV: operCmd.Mode1 and operCmd.Mode2
    - rxCmd TLV: rxCmd.Mode1 and rxCmd.Mode2
    - Resp TLV: Resp.Mode1 and Resp.Mode2
    - rxResp TLV: rxResp.Mode1 and rxResp.Mode2
    - Note: TLV.Mode1 and TLV.Mode2 = NULL when TLV = NULL
    - Note: Current state machine uses shorthand for the content of the TLV mode field, ASSOC_ACK_rx, ASSOC_NACK_rx, PREASSOC_ACK_rx, PREASSOC_NAK_rx, DEASSOC_ACK_rx
- ***Resolution Proposal:***
  - Accept as editorial clarification to existing machine (IEEE 802.1Qbg Draft-1), if consensus exists in TG.

Error handling Variables

- ***Originator's Description:***
  - Current VDP machine illustrates error checks on procedure calls, however these are not conventionally illustrated in our machines unless we specify the specific errors possible and how they affect the machine outcomes.
  - We suggest removing these from the illustration unless we fully specify the operation of the procedures.
  - Also we normally want to remove the if…then…s from state blocks to fully expand the state machine. Current Station state machine uses two timers for an response timeout and for a keep alive.
- ***Resolution Proposal:***
  - Error checks are for the machine operation
  - Accept as update to existing machine (IEEE 802.1Qbg Draft-1) if consensus exists in TG.

VDP Draft 1.0: State consolidation

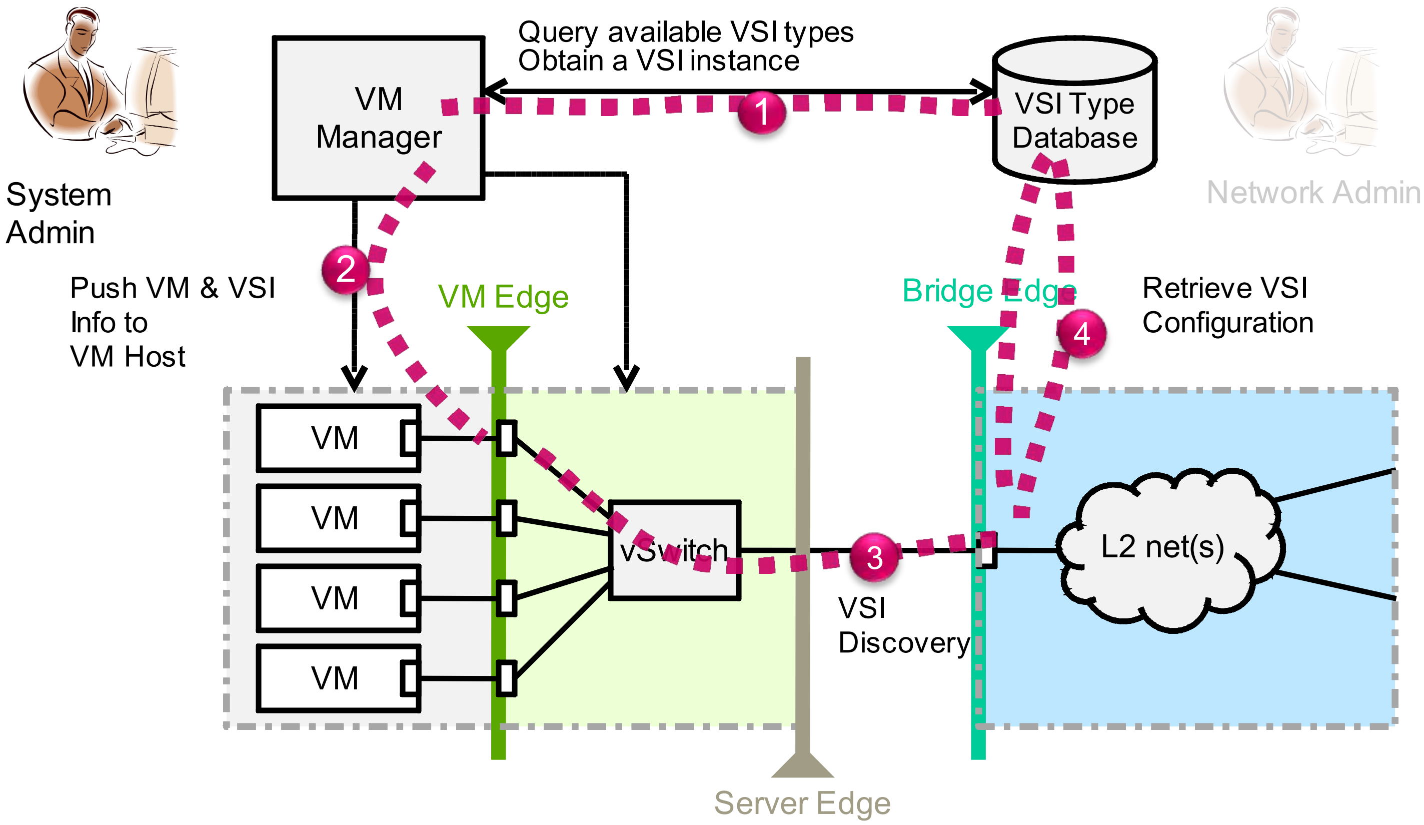
- ***Originator's Description:***
  - Some of the states in VDP state machines can be combined.
- ***Resolution Proposal:***
  - No change is required. In fact, current approach has several advantages
    - Current state machines states match operational state of station.
    - Simplifies state definition.
    - Improves problem determination and robustness.

Timer Events Using Local Variables

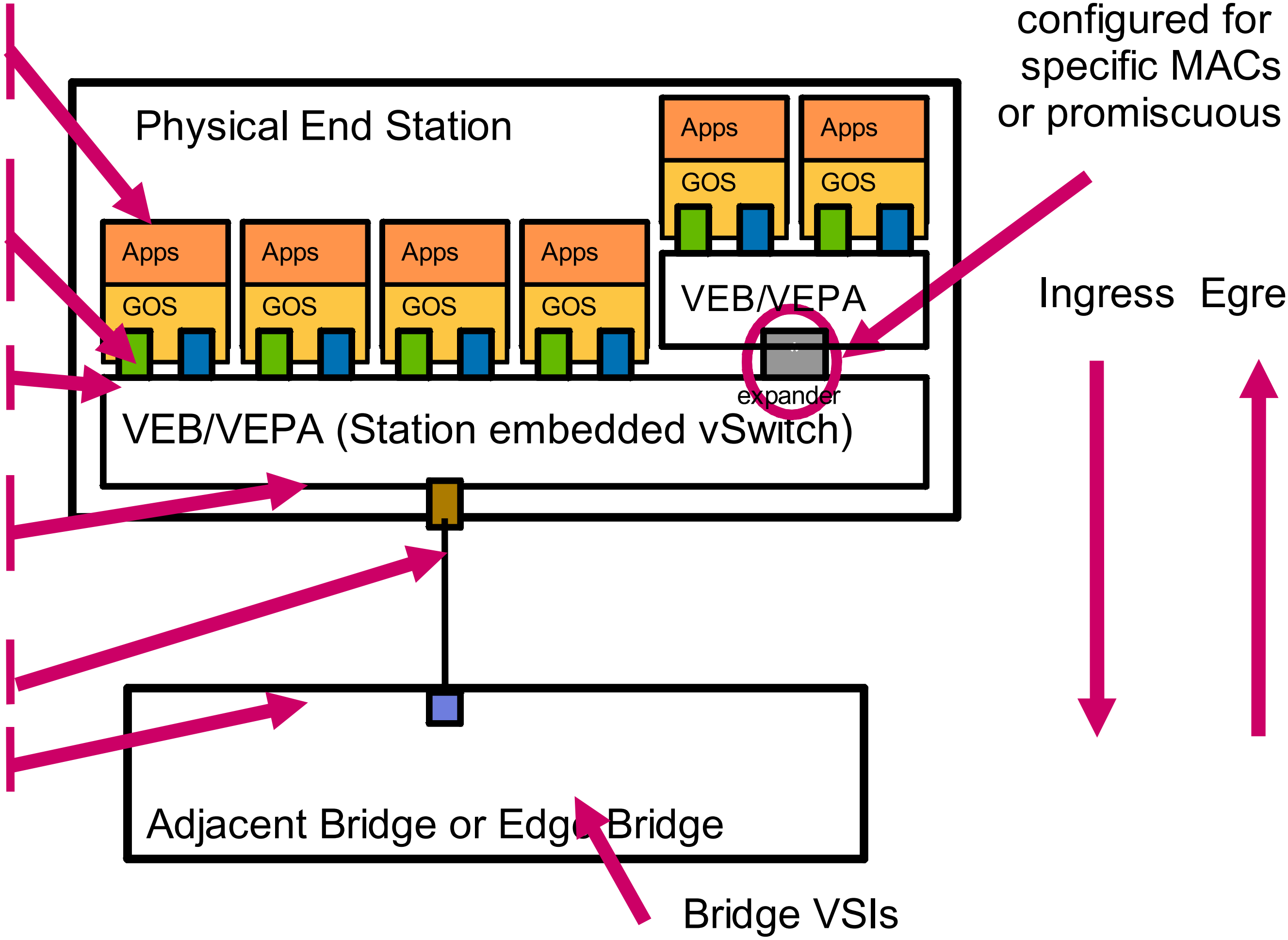
- ***Originator’s Description:***
  - Current Station state machine uses two timers for an response timeout and for a keep alive.
  - Current Bridge state machine uses one timer for a keep alive time out.
    - Change Station to count down timers
    - Change Bridge to count down timers
- ***Resolution Proposal:***
  - Can be an update to existing machine (IEEE 802.1Qbg Draft-1) if recommended by TG.

Backup

One Scenario for Configuring Edge Connections (VSIs)



Basic VEB/VEPA Anatomy and Terms



Virtual Machine, Virtual End Station

Virtual NIC, Virtual Machine NIC (vNIC, vmnic)

Virtual Station Interface (VSI)

Hypervisor controlled Physical NIC (pnic, vmnic)

Uplink (Physical Link or Channel)

Bridge Port or vPort

vNICs can be configured for specific MACs or promiscuous

Ingress Egress

Bridge VSIs

VSI Discovery and Configuration Requirements

1. Support VSI preAssociate (with and without resource reservations), Associate and deAssociate.
2. ASSOCIATE, PreAssociate and DeAssociate are Idempotent i.e. can be repeated.
3. Capability to Associate skipping PreAssociate.
4. VDP will work both for VEPA and VEB environments.
5. VSI TLVs are defined on slide 6-9.
6. Local or remote event can arrive in any state.
7. VSI state machines ETTP as TLV transport and utilize following capabilities of ETTP
    1. Transport will be transmitting TLVs in-order and are received in-order.
    2. Flow control
    3. Transport error from ETTP and LLDP are indicated to VDP
    4. ETTP provides best effort delivery of TLV. Therefore VDP waits for ETTP waits for ACK_Timeout. If no response is received, VSI exits.
        1. ACK_Timeout = 2*T3P retransmission period * MAX-RETRIES)+locally administered wait described informative text.
8. Timeout mechanism to ensure:
    a. Bridge resources are not reserved too long for inactive VSIs (lease semantics)
    b. Allow removing resources from inactive VSIs with the goal of
        - Conserve bridges resources (Number VSIs being hadled by bridge can be large).
        - Prevent inactive or VMs in error state to continue to hold resources.
    - Timeout out values to be negotiated on per channel between station and bridge. One timeout used for all ULPs on the channel negotiated using EVB TLV.

VSI Discovery and Configuration Requirements

Manageability and Robustness

a. Ensure VSI state and configuration between the Station and the Bridge remains consistent.

b. Hard errors at the Bridge or the Hypervisor that can impact individual VSI or Hypervisor/Bridge as a whole (resetAll)

   - Option 1: All VSI configuration goes away.

c. Bridge and Station Errors are detected through one or more of the following mechanisms.

   - VSI KEEP-ALIVE (periodic transmission of VSI TLV from station and response from Bridge)
   - ACK Timer
   - Transport (ETTP and LLDP) status indications.

d. Support switch/hypervisor administrator actions force VSI deAssociate.

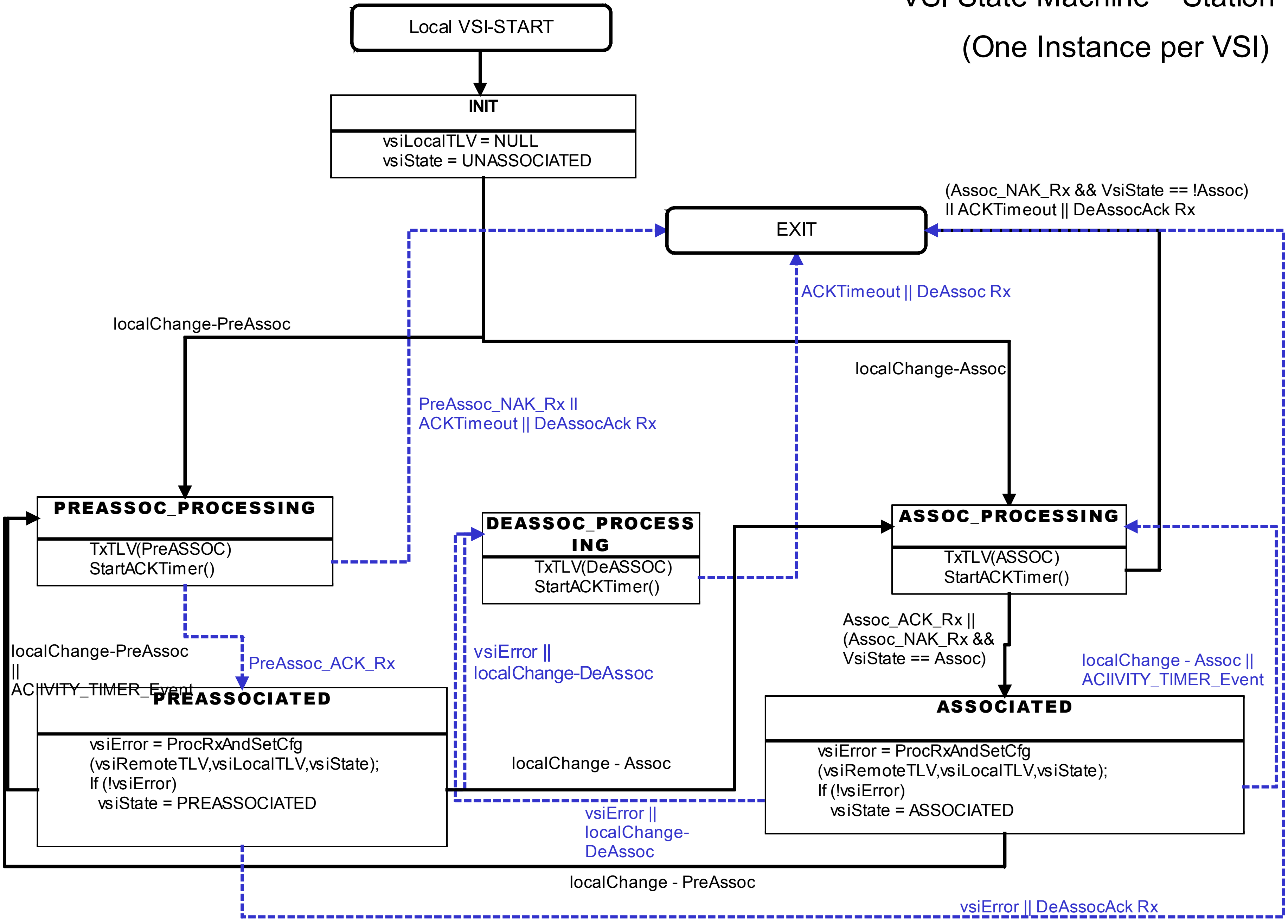
e. Statistics and logging support (need specific proposal)

Lost Digest Sync Semantics for State Machine, for each VSI Instance,

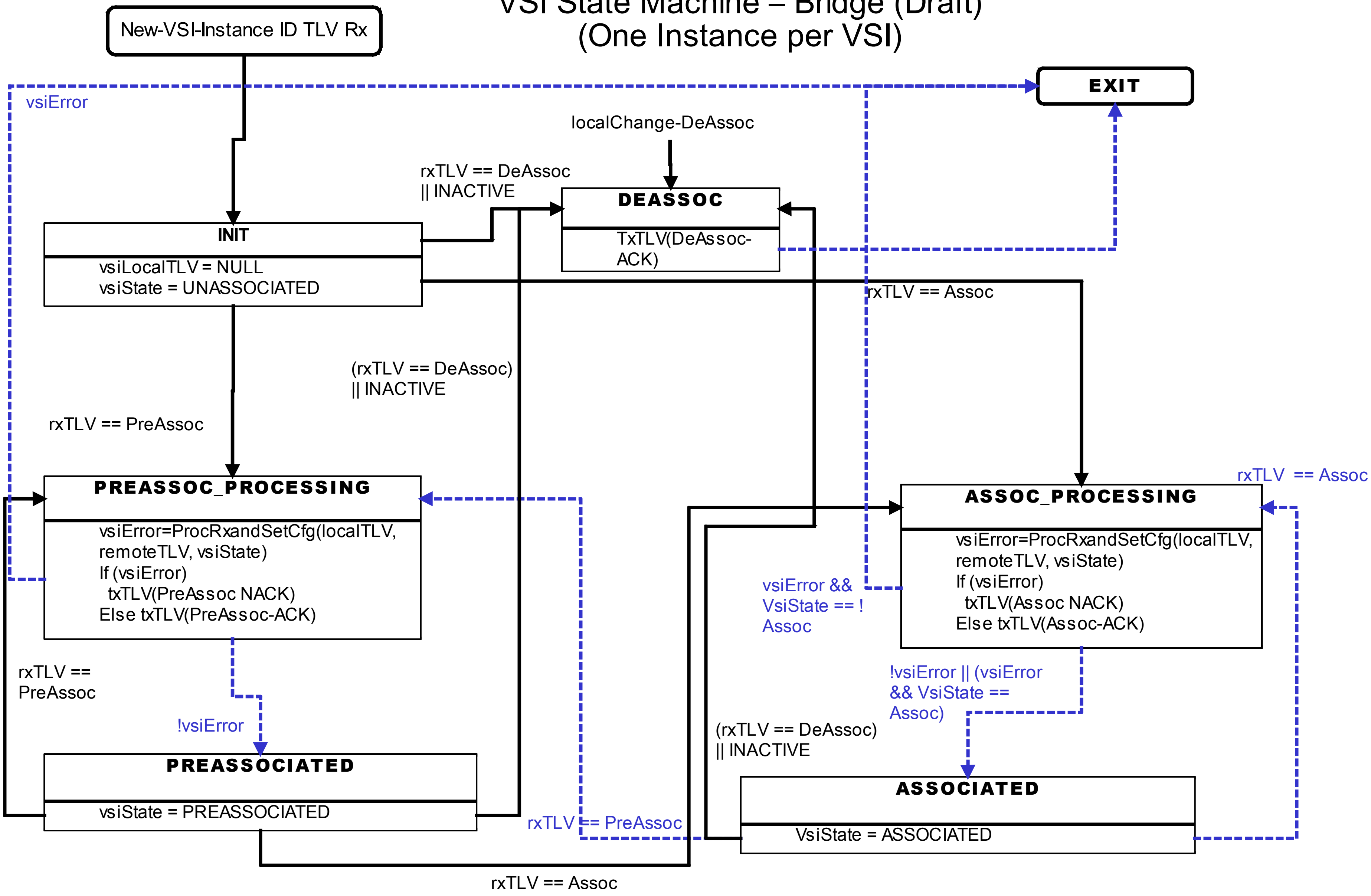
a. TLV Digest based sync is future enhancement and is not covered here.

VSI State Machine – Station

(One Instance per VSI)



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



New-VSI-Instance ID TLV Rx

EXIT

vsiError

localChange-DeAssoc

rxTLV == DeAssoc
|| INACTIVE

DEASSOC

TxTLV(DeAssoc-
ACK)

INIT

vsiLocalTLV = NULL
vsiState = UNASSOCIATED

rxTLV == Assoc

(rxTLV == DeAssoc)
|| INACTIVE

rxTLV == PreAssoc

rxTLV == Assoc

PREASSOC_PROCESSING

vsiError=ProcRxandSetCfg(localTLV,
remoteTLV, vsiState)
If (vsiError)
txTLV(PreAssoc NACK)
Else txTLV(PreAssoc-ACK)

ASSOC_PROCESSING

vsiError=ProcRxandSetCfg(localTLV,
remoteTLV, vsiState)
If (vsiError)
txTLV(Assoc NACK)
Else txTLV(Assoc-ACK)

vsiError &&
VsiState == !
Assoc

rxTLV ==
PreAssoc

!vsiError

!vsiError || (vsiError
&& VsiState ==
Assoc)

(rxTLV == DeAssoc)
|| INACTIVE

PREASSOCIATED

vsiState = PREASSOCIATED

rxTLV == PreAssoc

ASSOCIATED

VsiState = ASSOCIATED

rxTLV == Assoc

Notes on VSI State Machine

1. **vsiState:** Local variable for current state.
2. **localTLV:** Current local (active) TLV (configuration)
3. **AdminTLV:** TLV from local administration. In addition appropriate localChange variable is set. It allows mode change
4. **RemoteTLV:** TLV received from remote.
5. **TxTLV(*vsiTLV*):** Transmits AdminTLV using TLV transport (ETTP) service interfaces. Includes support for aggregation of VSI TLVs.
6. **ProcRxAndSetCfg(***vsiRemoteTLV,vsiLocalTLV,vsiState***):** Processes receive TLV and Sets local TLV variable based on received Remote TLV and vsiState. In case of error, returns error. This function handles PreAssociate with and without resource reservation case as well as accessing VSI Type definition fetch, if required
7. **StartACKtimer**(): Resets ACKTimeout local variable to FALSE and Starts ACK timer. Response (ACK or NACK) is expected before timer expires.
8. **ACKTimeout**: This local variable is set to true, if ACK timer expires
9. **vsiErrorPerm**(*vsiRemoteTLV*): processes the vsiRemoteTLV and returns TRUE is response code is unrecoverable (permanent) error.

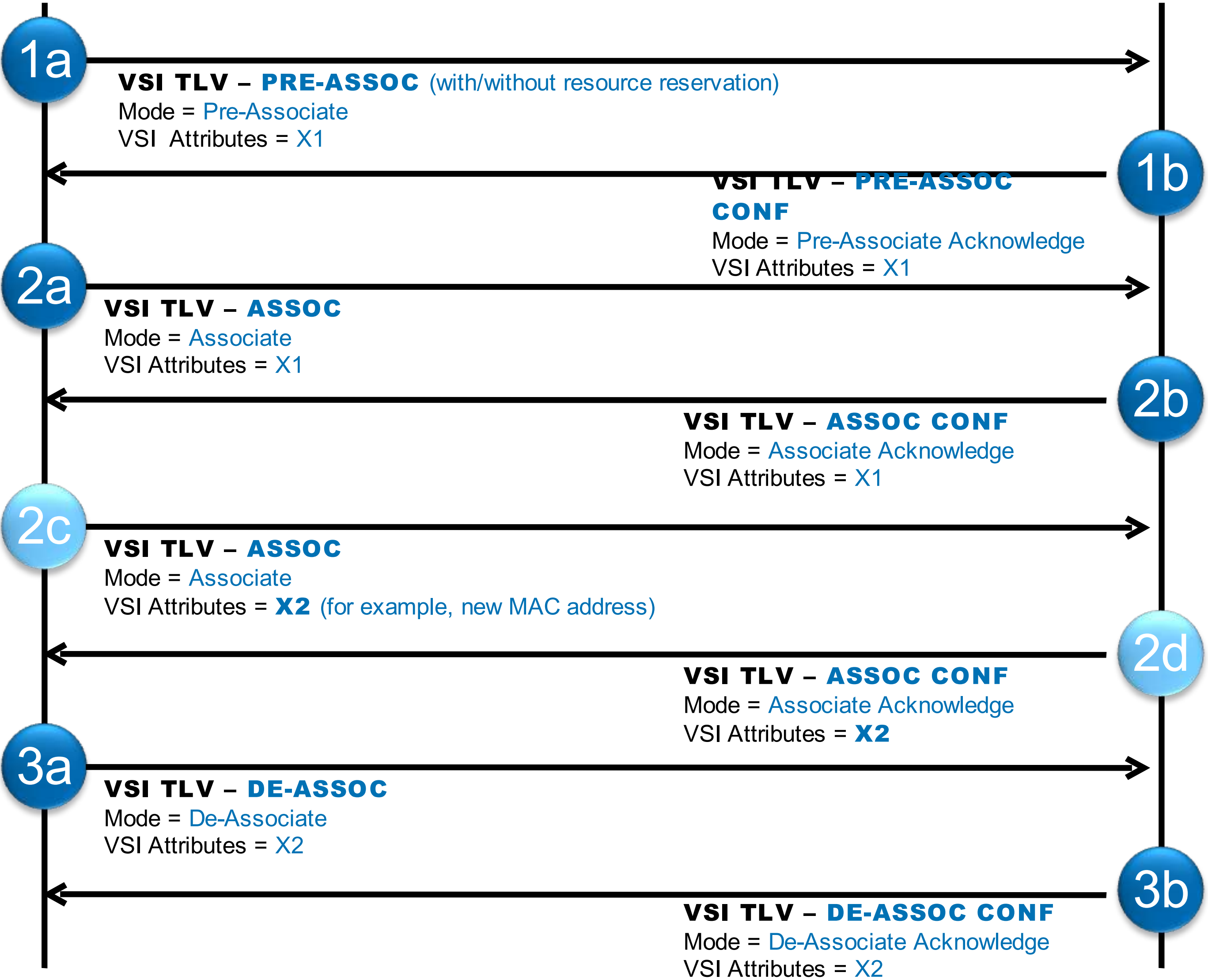
Keep alives during preassociate

- ***Originator's Description:***
  - No KEEP ALIVES during preassociate
- ***Resolution Proposal:***
  - This is not true. The version 0 state machine does have a Keep Alive during pre-associate, it's the ACTIVITY_TIMER_EVENT. When the Activity Timer pops, the state machine transitions from Pre-Associated to Pre-Associate processing, which results in resending the last Pre-Associate.

VSI Discovery/Config. Exchange Example

Station (Hypervisor)/VSI

Bridge



This exchange is usually done by Station using the ECP TLV transport

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

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

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

VSI TLV – ASSOC
Mode = Associate
VSI Attributes = X1

Activates the new VSI connection

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

Server announces change to configuration

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

Adjusts for the new configuration

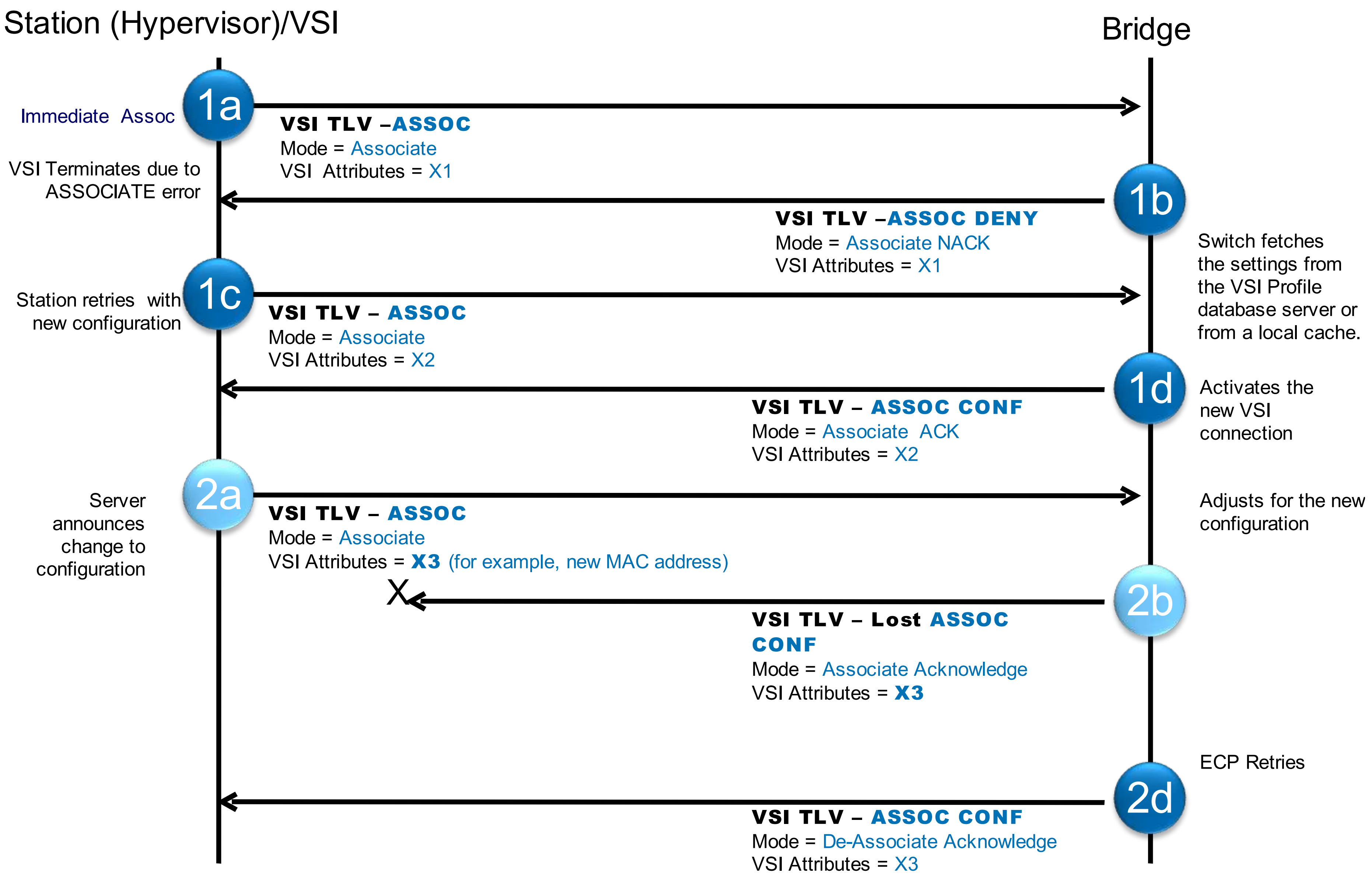
VSI TLV – ASSOC CONF
Mode = Associate Acknowledge
VSI Attributes = **X2**

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

Tears down VSI configuration

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

VSI Exchange Example – Direct Associate with Error Situations



Station (Hypervisor)/VSI

Bridge

1a

Immediate Assoc

VSI TLV –ASSOC
Mode = Associate
VSI Attributes = X1

1b

VSI Terminates due to ASSOCIATE error

VSI TLV –ASSOC DENY
Mode = Associate NACK
VSI Attributes = X1

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

1c

Station retries with new configuration

VSI TLV – ASSOC
Mode = Associate
VSI Attributes = X2

1d

Activates the new VSI connection

VSI TLV – ASSOC CONF
Mode = Associate ACK
VSI Attributes = X2

2a

Server announces change to configuration

VSI TLV – ASSOC
Mode = Associate
VSI Attributes = **X3** (for example, new MAC address)

Adjusts for the new configuration

2b

X

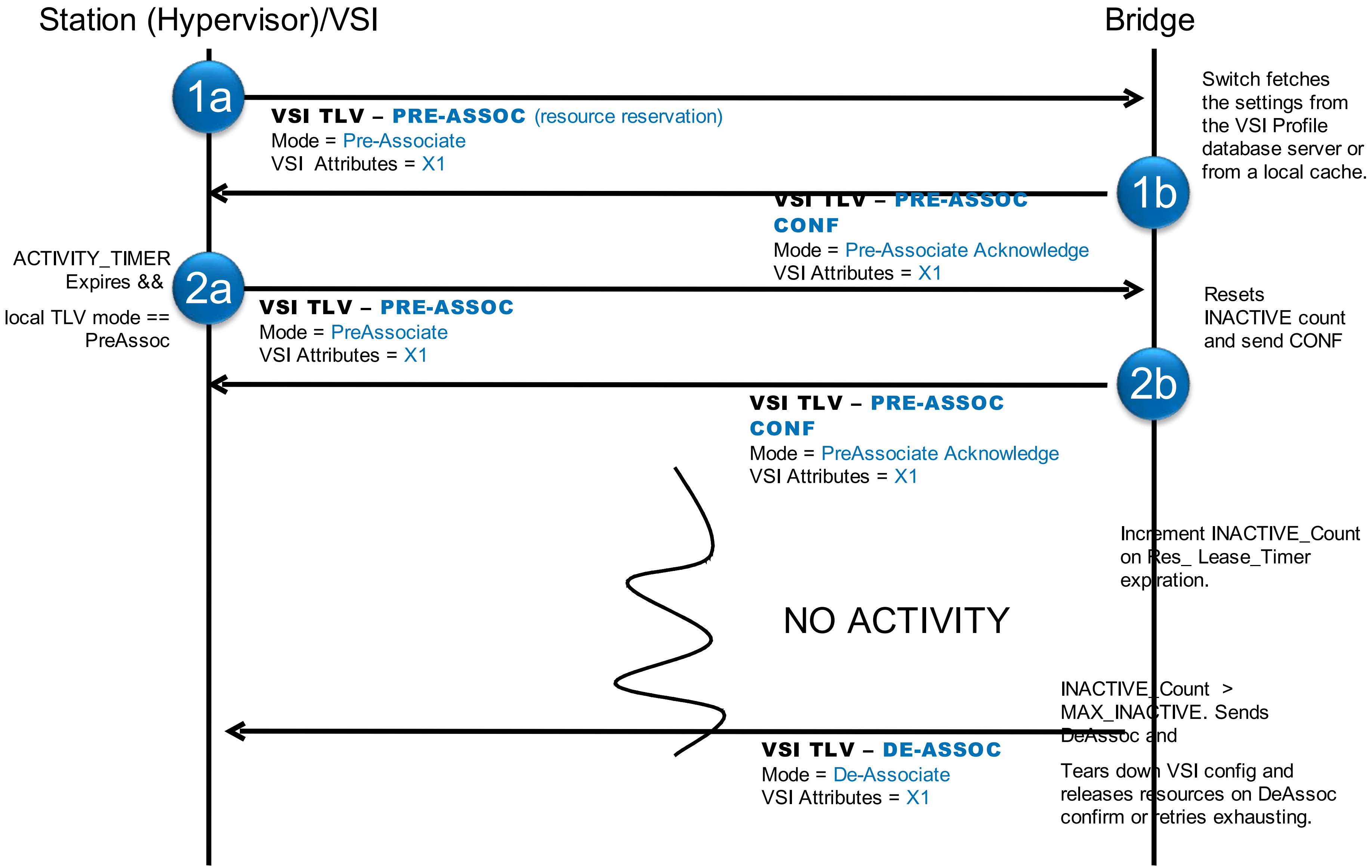
VSI TLV – Lost ASSOC CONF
Mode = Associate Acknowledge
VSI Attributes = **X3**

ECP Retries

2d

VSI TLV – ASSOC CONF
Mode = De-Associate Acknowledge
VSI Attributes = X3

VSI Exchange Example – PreAssoc Resource Lease



Station (Hypervisor)/VSI

Bridge

1a

VSI TLV – PRE-ASSOC (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

ACTIVITY_TIMER Expires &&

local TLV mode == PreAssoc

2a

VSI TLV – PRE-ASSOC
Mode = PreAssociate
VSI Attributes = X1

Resets INACTIVE count and send CONF

2b

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

Increment INACTIVE_Count on Res_ Lease_Timer expiration.

NO ACTIVITY

INACTIVE_Count > MAX_INACTIVE. Sends DeAssoc and

Tears down VSI config and releases resources on DeAssoc confirm or retries exhausting.

VSI TLV – DE-ASSOC
Mode = De-Associate
VSI Attributes = X1