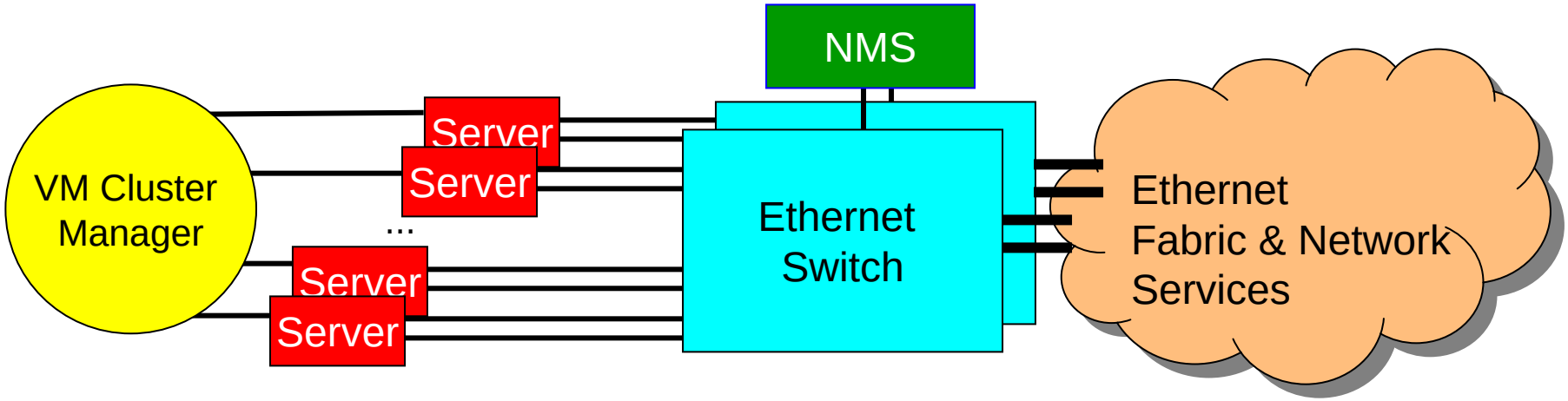
VM Migration Trigger Discussion

P. Congdon M. Krause

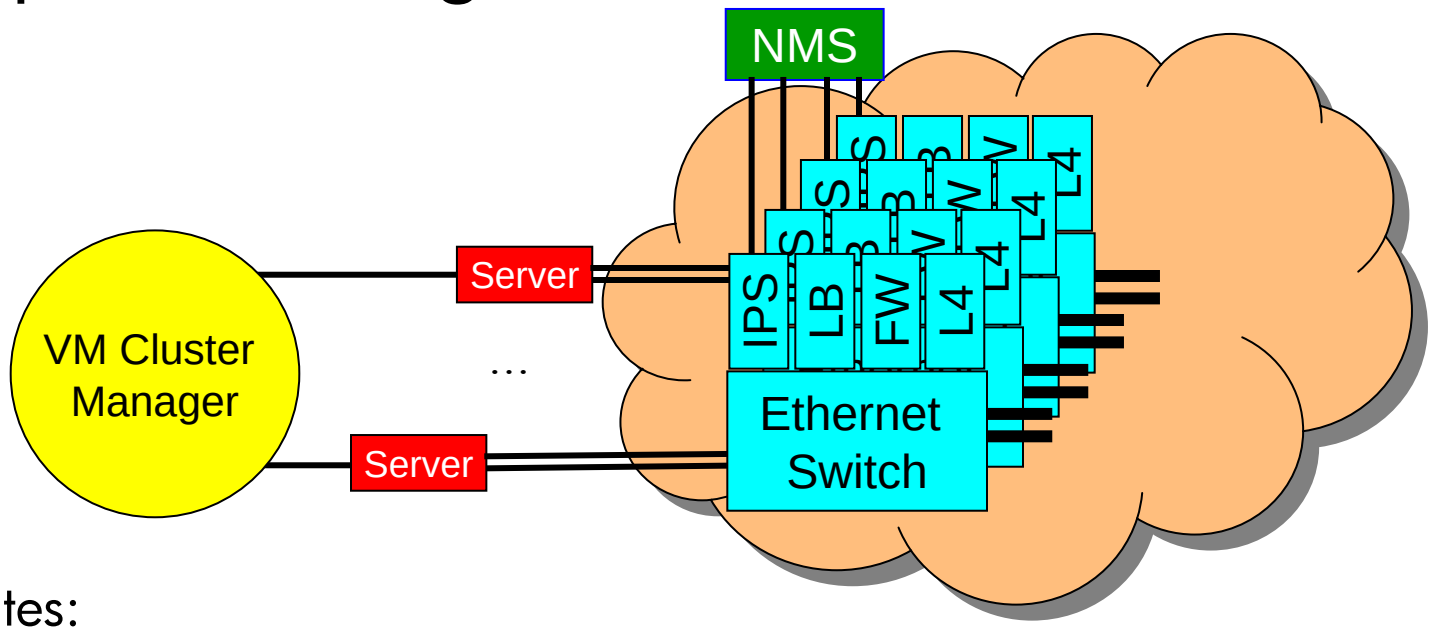
6/12/2011

Example of a Modest VM Cluster



- Attributes:
  - Scale – 8-32 nodes per cluster
  - Two Ethernet switch configuration
  - Shared NMS
  - Software network appliances, e.g. a VM-based firewall
  - Network hardware functionality, e.g. load balancer
- VM migration often has no impact to network due to modest scale and shared components – VM Cluster manager network interactions may be very limited (e.g. comprehend the topology to avoid SPOF).
- Server contain single or distributed vSwitch
  - Distributed vSwitch is a vSwitch than spans all or part of VM cluster
  - Multiple vSwitch (each type) may exist in a cluster

Example of a Larger Scale VM Cluster



- Attributes:
  - Scale – 32-256 nodes per cluster
  - Multi-Ethernet switch configuration or larger dual-switch configuration
  - Shared NMS
  - Software network appliances, e.g. a VM-based firewall
  - Network hardware functionality, e.g. load balancer
- VM migration may occur across disparate components and in turn impact network services which may not be visible to the VM Cluster manager
- Server contain single or distributed vSwitch
  - Distributed vSwitch is a vSwitch than spans all or part of VM cluster
  - Multiple vSwitch (each type) may exist in a cluster

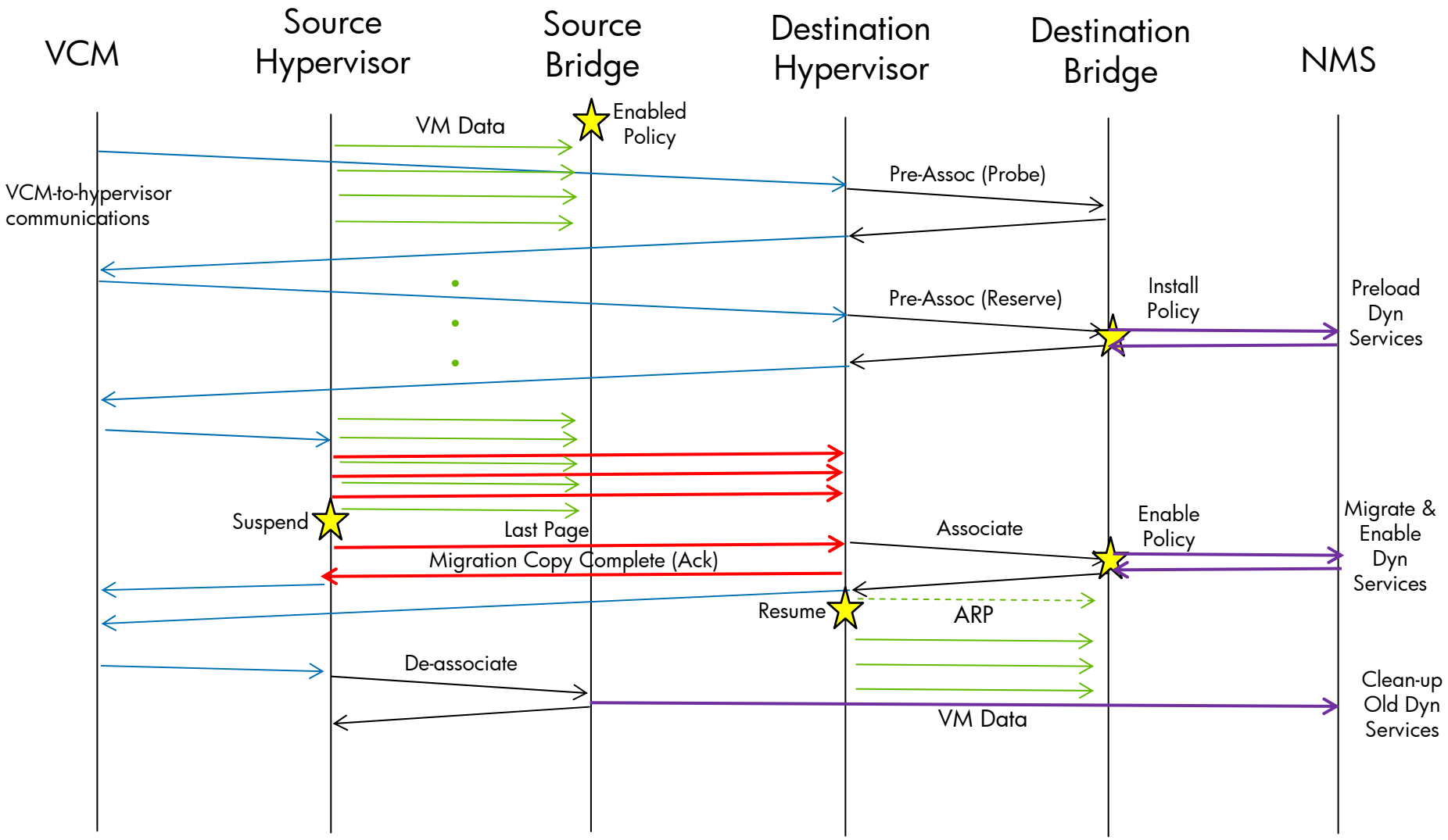
VM Cluster Manager (VCM) Responsibilities

- VCM comprehends all resources within the cluster – servers, storage, network, etc. (this may include VM-based appliance resources)
- VCM comprehends and provisions hypervisors and migration targets
  - VCM acquires per VM OVF data schema which includes network information from the port profile database and other configuration services
    - VCM comprehends each VM's purpose – application, infrastructure, appliance, etc.
  - VCM determines optimal placement and pushes OVF (native or may be packaged as part of hypervisor-specific data object) to targeted hypervisor
    - Hypervisor invokes VDP to acquire network information
    - Hypervisor configures VM via OVF and VDP information
  - Migration target hypervisor may perform probes (Pre-Associate) to determine if migration is possible and report to VCM to make any migration decision
    - VCM may request hypervisor to Pre-Associate with Resources to insure a migration may occur – variable time delay between reservation and migration
      - Resource reservations may be used for planned and unplanned events
- VCM initiates a migration based on external policy, e.g. fail-over event
  - Migration may fail or be halted by the VCM - cannot make any assumption of success until the VCM & hypervisors signal the migration is successful

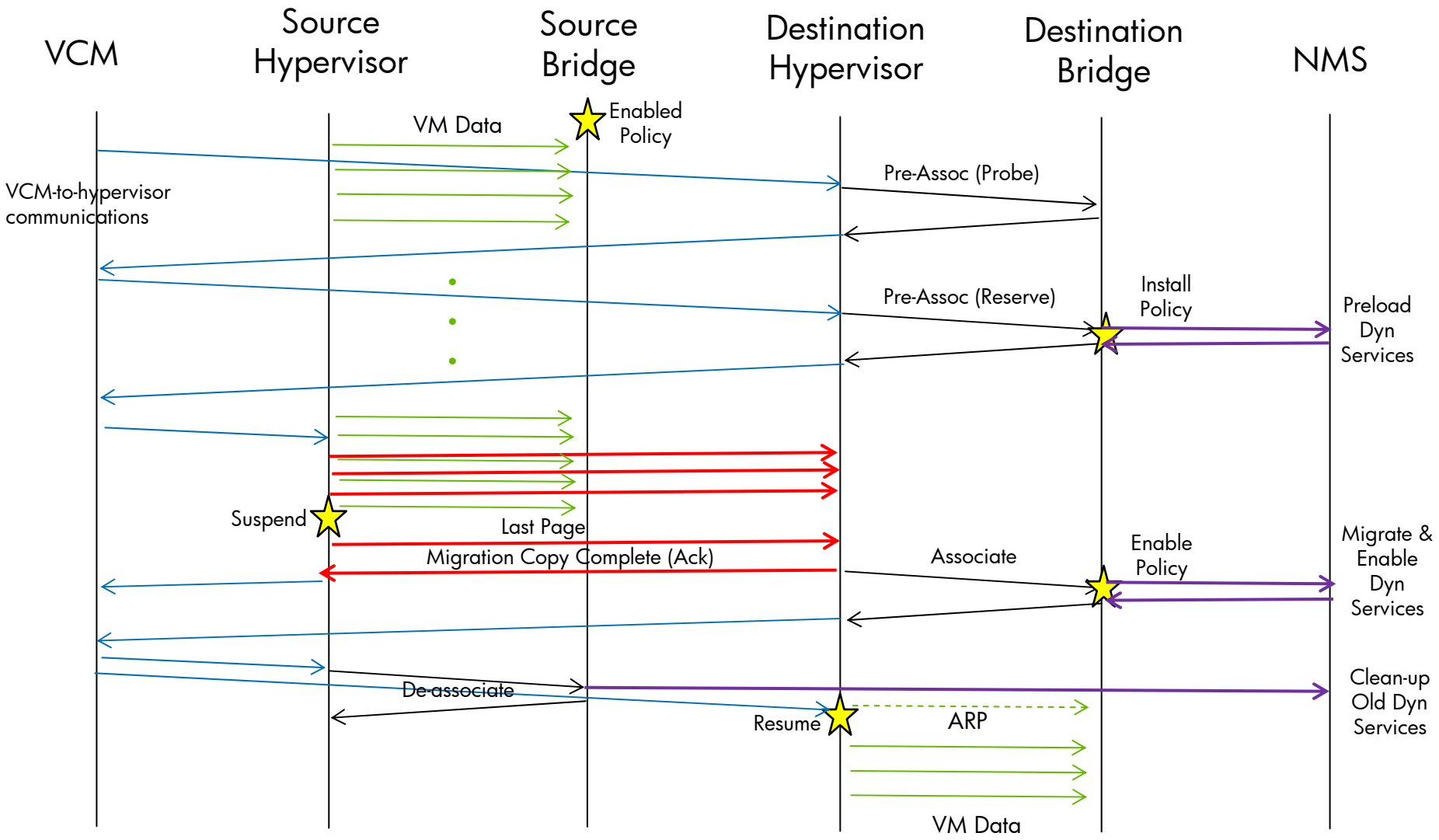
VM Migration

- Basics of VM Migration
  - VCM pushes per VM OVF / resource configuration to destination
    - Destination hypervisor confirms migration possible
  - VCM instructs destination hypervisor to pre-stage migration
    - Destination hypervisor performs Pre-ASSOC with resource reservations to insure network resources are available
    - Bridge allocates and configures resources but does not enable
      - Bridge may optionally acquire and configure additional network state information before or post the Pre-ASSOC completion
      - Non-determinate time between the Pre-ASSOC Request and its completion
  - VCM informs source hypervisor to migrate state
    - Source hypervisor iteratively flushes VM pages to destination
    - When down to the last pages, the source hypervisor suspends the VM and flushes the last pages to the destination hypervisor
      - Last pages are the final suspend and the VM is inactive at the source
  - Destination hypervisor configures VM leaving it suspended until the last pages are received and determines the migration is successful
  - Destination hypervisor confirms success to VCM and activates VM as instructed

Full Ladder Diagram – Minimize Latency



Full Ladder Diagram – Safest Exchange



Migration Observations

- VCM controls migration process from start to finish
  - Coordinates source and destination hypervisor activities and provides policies and resource definitions required for VM migration
  - VM will see at most 4-5 seconds of disruption during final suspend and move. As long as the total time does not exceed any VM-tracked timers or network / storage timers, then the migration is experience is treated as a momentary performance loss.
  - VMs do not constantly move – migration may never occur for a given VM or may occur in time deltas measured in minutes, hours or even days.
    - At any given time there may be zero to just a handful of migrations occurring within a large cluster.
- Source and destination hypervisors interact with network via VDP
  - VCM interacts with port profile database
  - VCM may interact with network services through (non-) industry standard interfaces / protocols
- Source VM never interacts with the network post final suspend
- Destination VM never interacts with the network until migration is successful, i.e. a successful Association in the VM migration flow which enables the VSI and the bridge-side resources
- Many network services are implementation-specific with no industry standards, e.g. ACL, policies, etc.

Service Configuration Relative to Migration

Network Service	Self-learning	Explicit Configuration	PRE-ASSOC w/ Resources	Post Migration
DNS	VM initiated	External config	No impact	No impact
DHCP	VM initiated	External config	No impact	No impact
Firewall	No	External config	Configure during	Enable
Load balancer	Yes	External policy	No impact	No impact
ACL	No	External config	Configure during	Enable
Statistics	Yes	No	Extract during – optional NMS managed so no extraction needed	Optional Post Fix-up
IGMP Filter	Self-learning	No	No impact	No impact
IGMP join	No	NMS Configured	Configure during	Fixed up & Enable
TCP layer 4 proxy	Self-learning	No but max # may be NMS configured	Configure during but cannot account for TCP connections created prior / during migration time	Yes – create any new TCP, update TCP window, etc. Must always update final context

Service Migration relative to Trigger

Network Service	Source Trigger	Destination Trigger	Comment
DNS	No need	No need	IP / MAC are part of OVF / VM operation and migrate unchanged
DHCP	No need	No need	IP / MAC migrate unchanged
Firewall	No significant benefit	Enable Service	Should be configured but not enabled as part of PRE-ASSOC with Resources
Load balancer	No significant benefit	Enable Service	Should be configured but not enabled as part of PRE-ASSOC with Resources
ACL	No significant benefit	Enable Service	Should be configured but not enabled as part of PRE-ASSOC with Resources
Statistics	No significant benefit	Populate & Enable	Optional final extraction and population at with Dest. Trigger or dealt with via NMS
IGMP Filter	No significant benefit	No significant benefit	Self-learning at destination
IGMP join	Possible benefit – unclear	Populate & Enable	NMS configured. Initial may be done at PRE-ASSOC with Resources
TCP layer 4 proxy	Value relative to # TCP context	Populate & Enable	Must allocate / update TCP state for any connections created prior / during migration independent of when trigger occurs.

Single vs. Dual-Trigger

Attribute	Single	Dual	Comment
Complexity	Simpler	More complex	Single point of control
Configuration	Simpler	More complex	Single-trigger model relies Pre-ASSOC with Resources to perform time-intensive operations. Dual-trigger may also but primarily relies on first trigger to perform time-intensive operations
Migration "Fix Up" Time for Dynamic Services	Equal	Equal / Service-dependent Faster	Both approaches require post-migration fix-up of dynamic content. Most services would experience the same "fix up" time. For Layer 4 proxy, first trigger starts TCP connection creation but final state must still be upgraded upon source final suspension. Dual-trigger may provide some benefit but the actual time may not be significant.

VDP Semantics – from 802.1Qbg

- Pre-Associate
  - The Pre-Associate TLV type is used to pre-associate a VSI Instance Identifier with a bridge port. The bridge validates the request (see below) and returns a failure reason in case of errors. Successful pre-association does not imply that the VSI Type will be applied to any traffic flowing through the VSI, as the VSI instance may still be associated with another port on the network.
- Pre-Associate with Resource
  - Pre-Associate with Resource Reservation has the same steps as Pre-Associate, but also reserves resources in the Bridge to prepare for a subsequent Associate request.
  - The Bridge validates the required resources and shall reserve resources for a subsequent associate step. Pre-Associate does not allow any traffic from VSI; this is enabled at a later time when the VSI is Associated.
- Associate
  - The Associate TLV Type creates and activates an association between a VSI Instance and a bridge port. The Bridge allocates any required bridge resources for the referenced VSI. The Bridge activates the configuration for the VSI Type ID. This association is then applied to the traffic flow to/from the VSI Instance.

Specification Modifications

- May want to remove the following from Associate:
  - *NOTE—This allows a newly created virtual station to be prepared while the old virtual station is still running, minimizing the transition time from the old virtual station to the new.*
- May want to add a note that states:
  - *NOTE – A Pre-associate with Resource Reservations should be performed at the destination prior to a VM migration being initiated. An Associate should be performed once the VM migration has succeeded to enable the VSI at the destination bridge.*

Back-up

NMS Roles & Responsibilities

- Varies by solution
  - Statistics gathering
  - Service provisioning and configuration
    - E.g. ACL, firewall, load balancer policies, etc.
  - Coordination across network components
    - Acquisition of port profile database elements
    - Migration of network implementation-specific configuration and state information
  - NMS comprehends the VSI ID and can track all attributes and state changes at the source and migrate these to the destination hypervisor as needed.
    - There is no standard proposed to accomplish this