Modularization of Edge Virtual Bridging – proposal to move forward

July 2009

Paul Congdon (HP) ptcongdon@ucdavis.edu
Chuck Hudson (HP) chuck.hudson@hp.com

Agenda

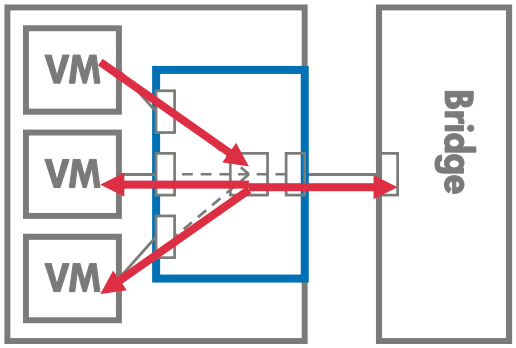
- Refresh on Basic VEPA proposal
- Reflective Relay PAR
- Adding MultiChannel Support
- Supporting Port Expanders
- Adding Remote Replication Services
- Proposed Roadmap to convergence

EVB Yahoo Group
http://tech.groups.yahoo.com/group/evb/

- Unofficial ad hoc group working to develop concepts and proposals related to Edge Virtual Bridging for consideration by the IEEE 802.1 working group.

- Membership
  - 100+ members have joined Yahoo group
  - Affiliated with 20+ companies (including server, switch, NIC, hypervisor & OS companies)
- Weekly Conference Calls
  - Tuesdays 1PM Central
  - Since February 20th 2009
  - 25-30 Attendees Weekly
- Actively working on converging proposals

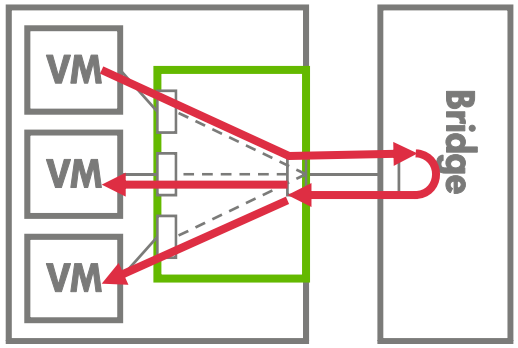
Approaches



Virtual Ethernet Bridge (VEB)

uses MAC+VID to steer frames

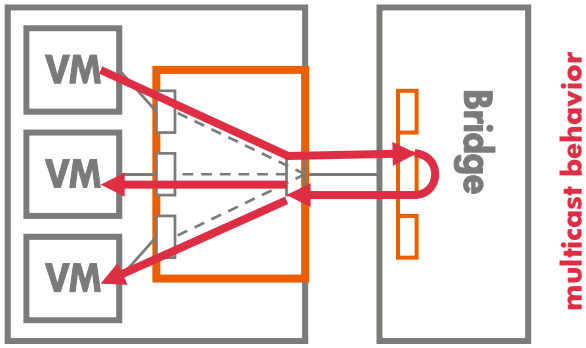
- Emulates 802.1 Bridge
- Works with all existing bridges
- No changes to existing frame format.
- Limited bridge visibility
- Limited feature set
- Best performance.
- Will always be there



Basic VEPA (tagless)

uses MAC+VID to steer frames

- Exploits 802.1 Bridge
- Works with many existing bridges
- No changes to existing frame format.
- Full bridge visibility
- Access to bridge features
- Constrained performance
- Leverages VEB



VEPA +Multichannel, VN-Tag

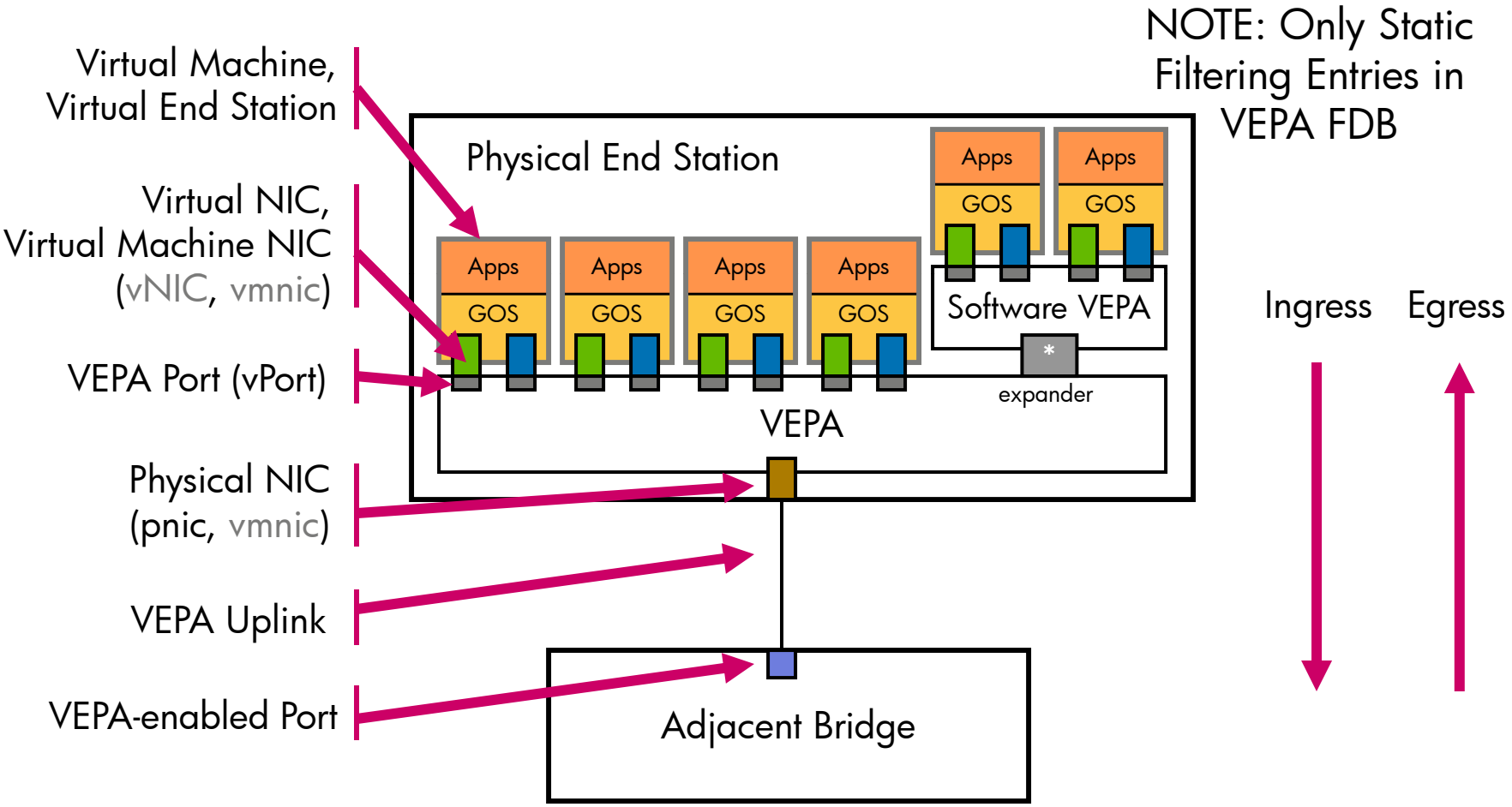
uses tag to help steer frames

- Extends 802.1 Bridge
- Works with few or no existing bridges
- Changes existing frame format.
- Full bridge visibility
- Access to bridge features
- Constrained performance
- Unclear how well it leverages VEB

Limitations of VEBs (today)

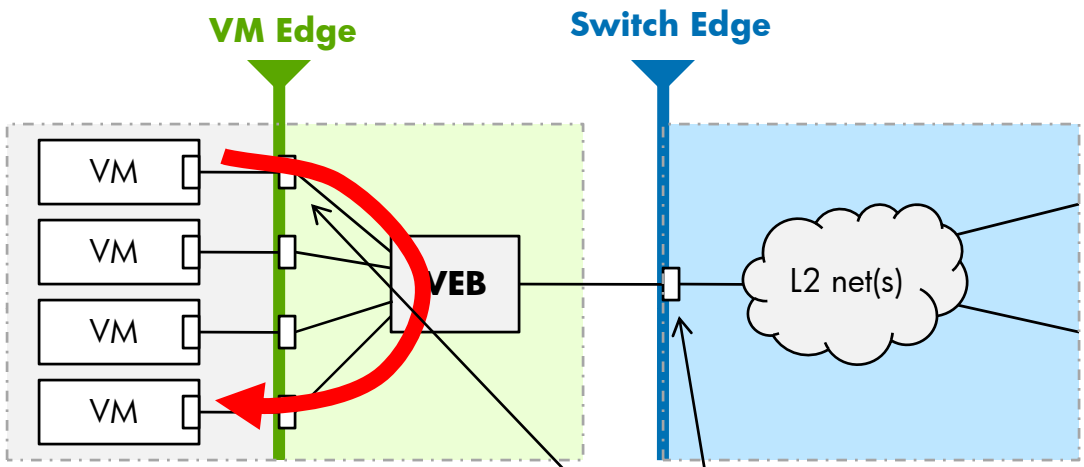
- Limited feature set compared to external switches
  - Limited or no packet processing (TCAMs, ACLs, etc.)
  - Limited support for security features (e.g., DHCP guard, ARP monitoring, source port filtering, dynamic ARP protection/inspection, etc.)
- Limited monitoring capabilities
  - Limited support for statistics and switch MIBs
  - No NetFlow, sFlow, rmon, port mirroring, etc.
- Limited integration with external network management systems
- Limited support for promiscuous ports (typically no learning)
- Limited support for 802.1 protocols (e.g., STP, 802.1X, LLDP)

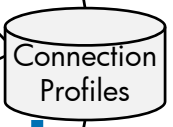
Basic VEPA Anatomy and Terms



Enabling Adjacent Bridge Policy

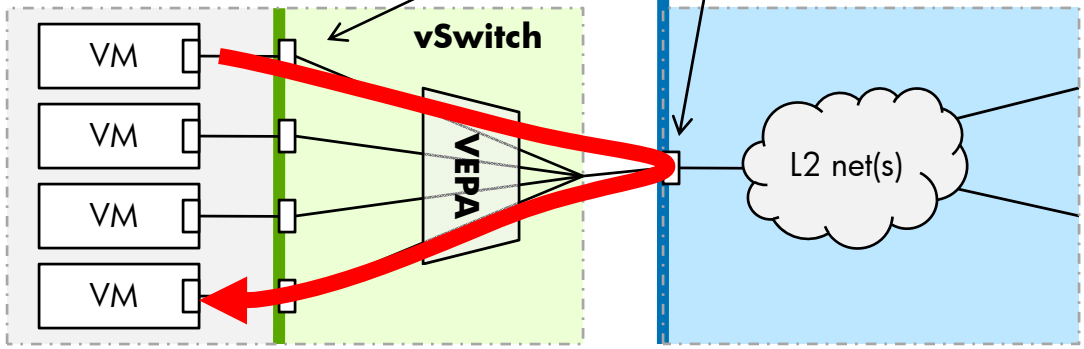
VEB mode allows VM-to-VM traffic with limited policy enforcement





Connection settings communicated to both VM Edge and Switch Edge.

VEPA mode forces all traffic to fully-capable edge for full policy enforcement



Benefits VEPA adds to VEB/vSwitch

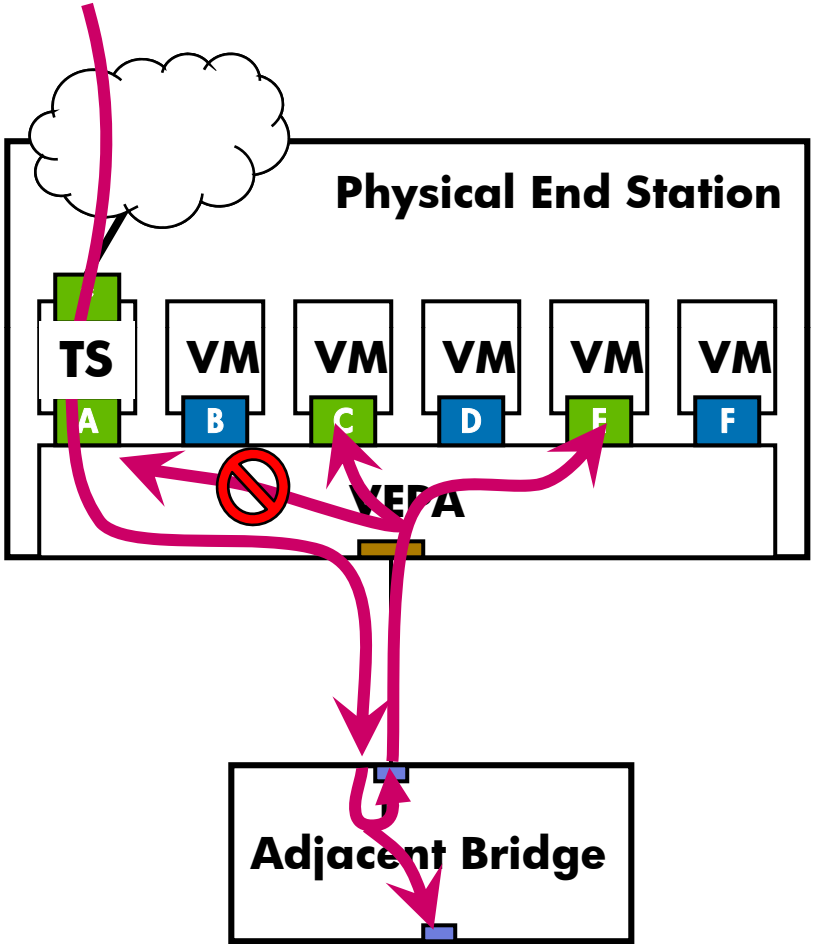
- Port configuration and management is the same for VEB and VEPA ports
- Gains access to external switch features
  - Packet processing (TCAMs, ACLs, etc.)
  - Security features such as: DHCP guard, ARP monitoring, source port filtering, dynamic ARP protection/inspection, etc.
- Enhances monitoring capabilities
  - Statistics
  - NetFlow, sFlow, rmon, port mirroring, etc.

‘Basic VEPA’ Limitations

- Basic VEPA is challenged by promiscuous ports
  - Must have complete address table and learning is discouraged
  - Difficult to create proper destination mask to account for promiscuous ports
  - Useful to support transparent services
- Want mix of VEPA and VEB ports on single physical link
  - Allow for optimized performance configuration

Problem with Dynamic Addresses

SRC = Z; DST = MulticastC



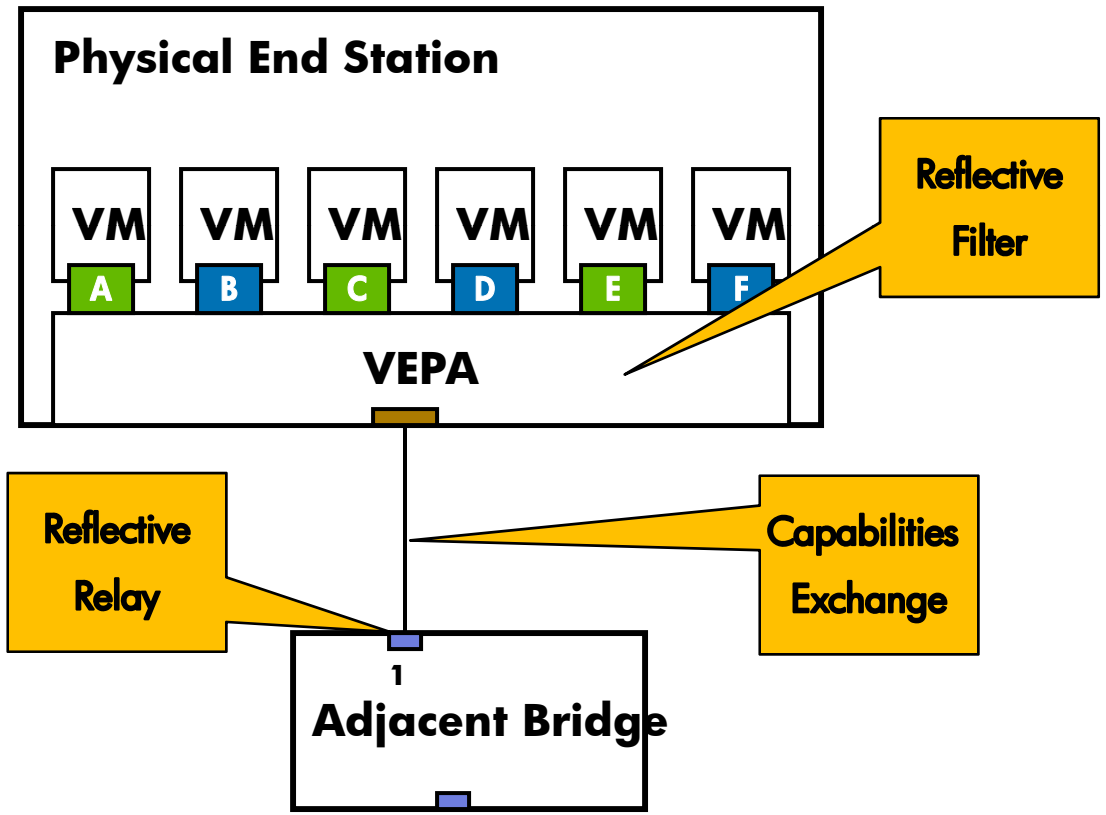
VEPA Address Table

DST MAC	VLAN	Copy To (ABCDEF)
A	1	100000
B	2	010000
C	1	001000
D	2	000100
E	1	000010
F	2	000001
Bcast	1	101010
Bcast	2	010101
MulticastC	1	101010
Unk Mcast	1	100010
Unk Mcast	2	010101
Unk Ucast	1	000000
Unk Ucast	2	000000

Edge Virtual Bridging Acceptable Constraints

- Primary use model is the connection of individual virtual end-stations
- Individual Virtual Machine vNIC configuration is known a head of time.
- VEPAs can be cascaded in software to increase size of address tables and number of supportable VMs
- Primary communication model is north-south, but optimized east-west traffic can be combined using same resources.

Specification Needs for Basic VEPA Operation



Static VEPA Address Table

DST MAC	VLAN	Copy To (ABCDEF)
A	1	100000
B	2	010000
C	1	001000
D	2	000100
E	1	000010
F	2	000001
Bcast	1	101010
Bcast	2	010101
MulticastC	1	101010
Unk Mcast	1	100010
Unk Mcast	2	010101
Unk Ucast	1	000000
Unk Ucast	2	000000

Reflective Relay PAR ('hairpin' mode)

Scope of Proposed Standard

This standard specifies protocols, procedures, and managed objects that:

- Provides for discovery, configuration, and control of a reflective mode of operation for the bridge relay service of a bridge port when connected to an external filtering service.
- Define the requirements of an external filtering service required to allow the safe operation of a reflective relay service.
- Define the requirements of operation for the reflective relay function.

Purpose

The purpose of this standard is to allow multiple stations attached to a common bridge port to obtain the services of bridge relay for that port without requiring the services of a separate port.

VEPA+MultiChannel is a simple extension that allows basic VEPAs, VEBs, and isolated vPorts to share a single physical port.

VEPA+MultiChannel
A Definition

VEPA+MultiChannel (Virtual Ethernet Port Aggregator Plus MultiChannel Ethernet) is a capability that uses S-VIDs to identify multiple virtual switch ports on a single physical switch port. Each virtual switch port may then be associated with a VEB, basic VEPA, or an isolated vPort within the physical station.

This capability is enabled by an S-Component within the physical station to identify the separate station-side elements (VEBs, basic VEPAs, or isolated vPorts) and a corresponding S-Component in the adjacent bridge that maps those elements to virtual switch ports in the adjacent bridge.

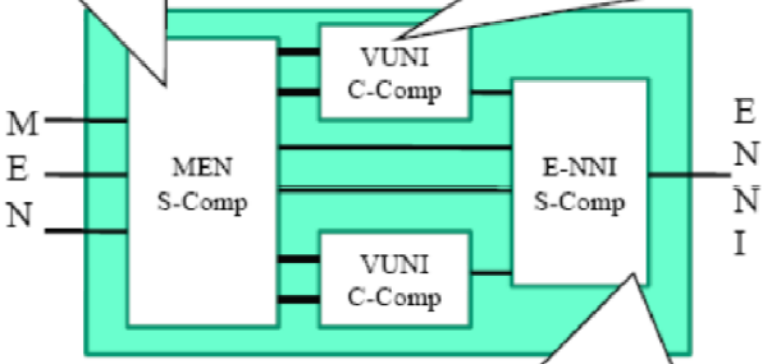
The MultiChannel capability leverages the existing Remote Customer Service Interface work (802.1bc)

Proposed directions for 802.1bc

A more detailed look

This is the S-component that would normally be at an E-NNI, even if were not doing any hairpin switching or VUNI functions.

A C-component to perform the same functions at the "Virtual UNI" for a single customer service interface that the C-component of a Provider Edge Bridge would perform at a normal UNI.

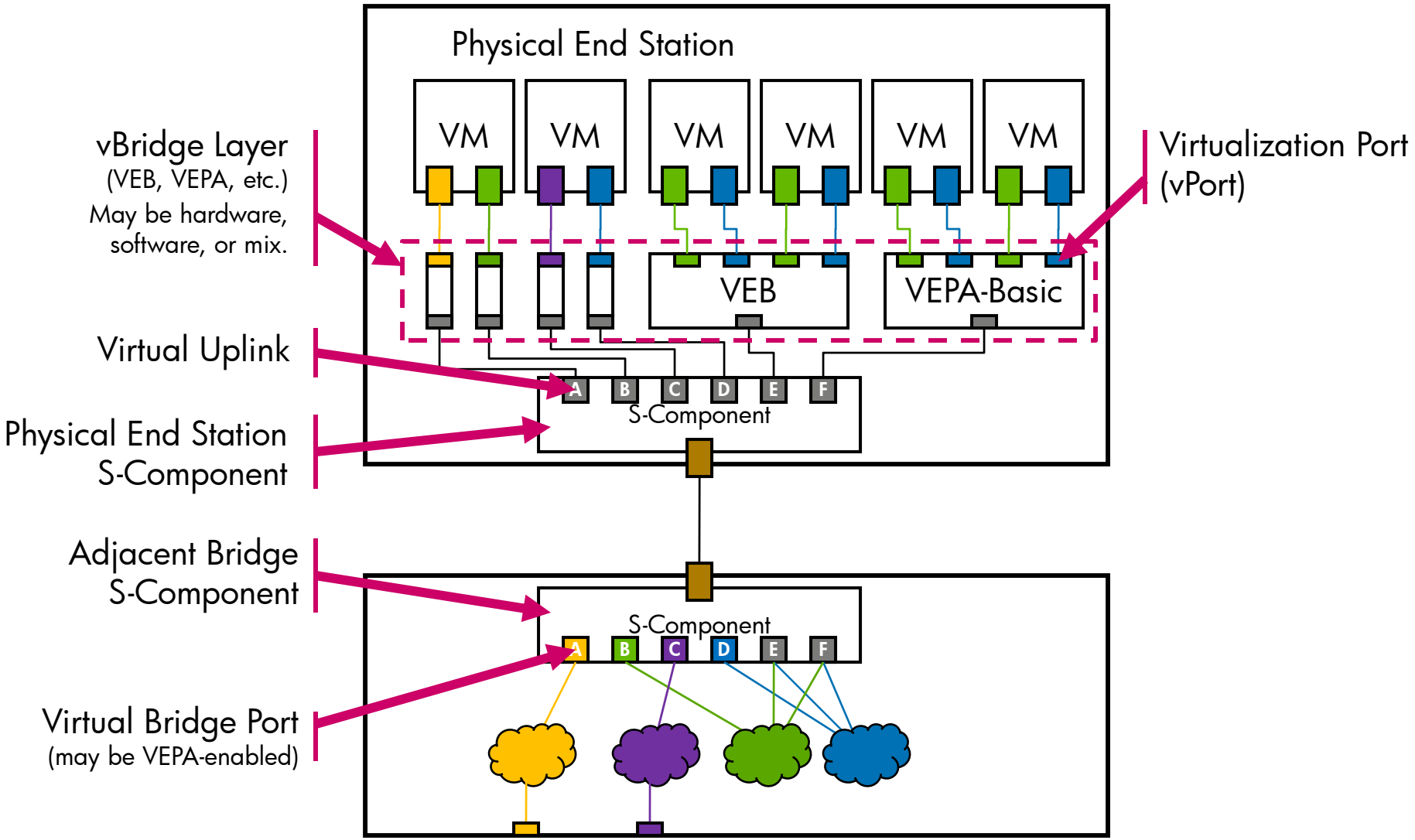


S-component (or new demultiplexing entity) dedicated to demultiplexing ingress frames from E-NNI based on the S-VID, and tagging egress frames to the E-NNI with the appropriate S-VID.

May 2009 802.1 meeting, Pittsburgh 20

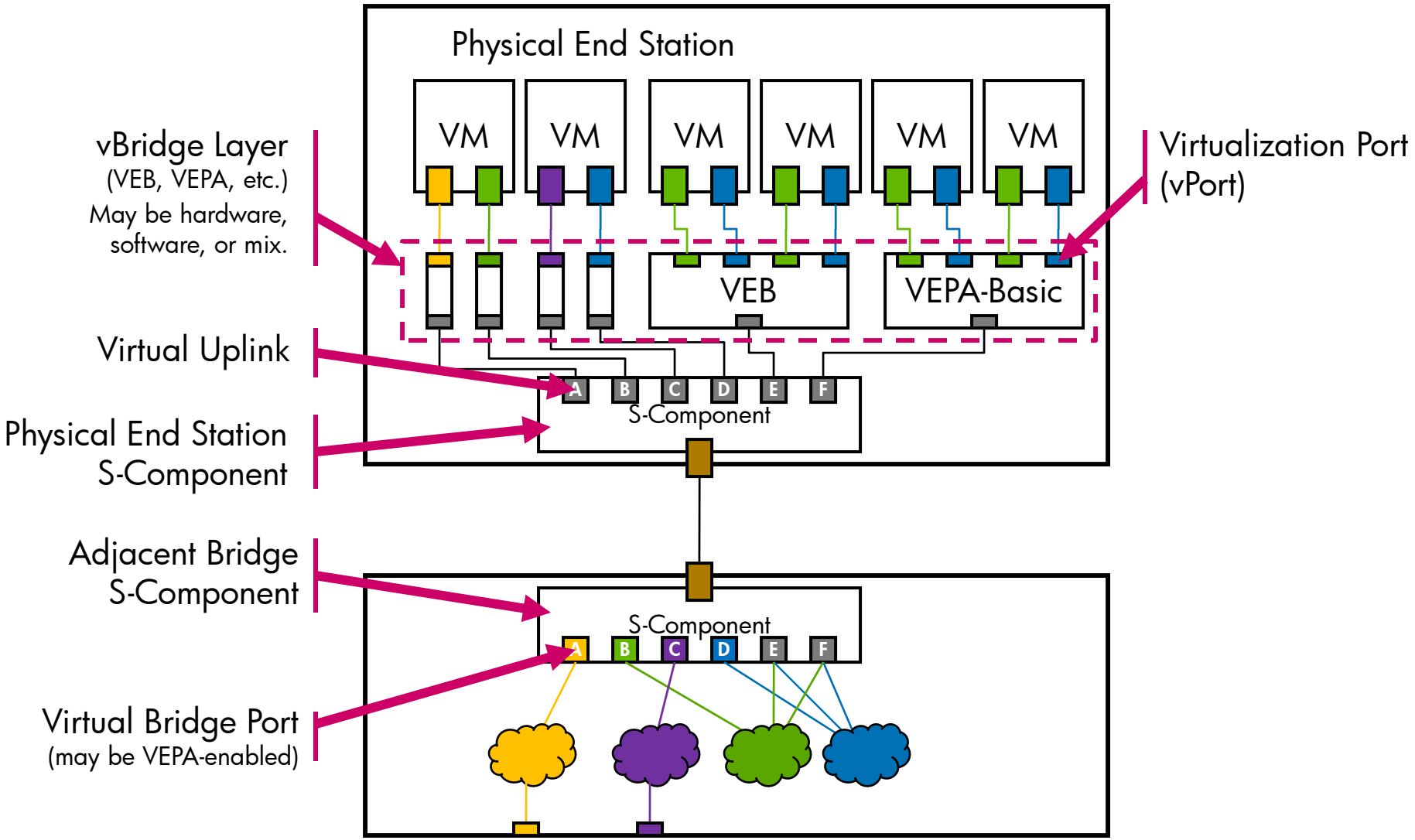
VEPA+MultiChannel

New Anatomy and Terms



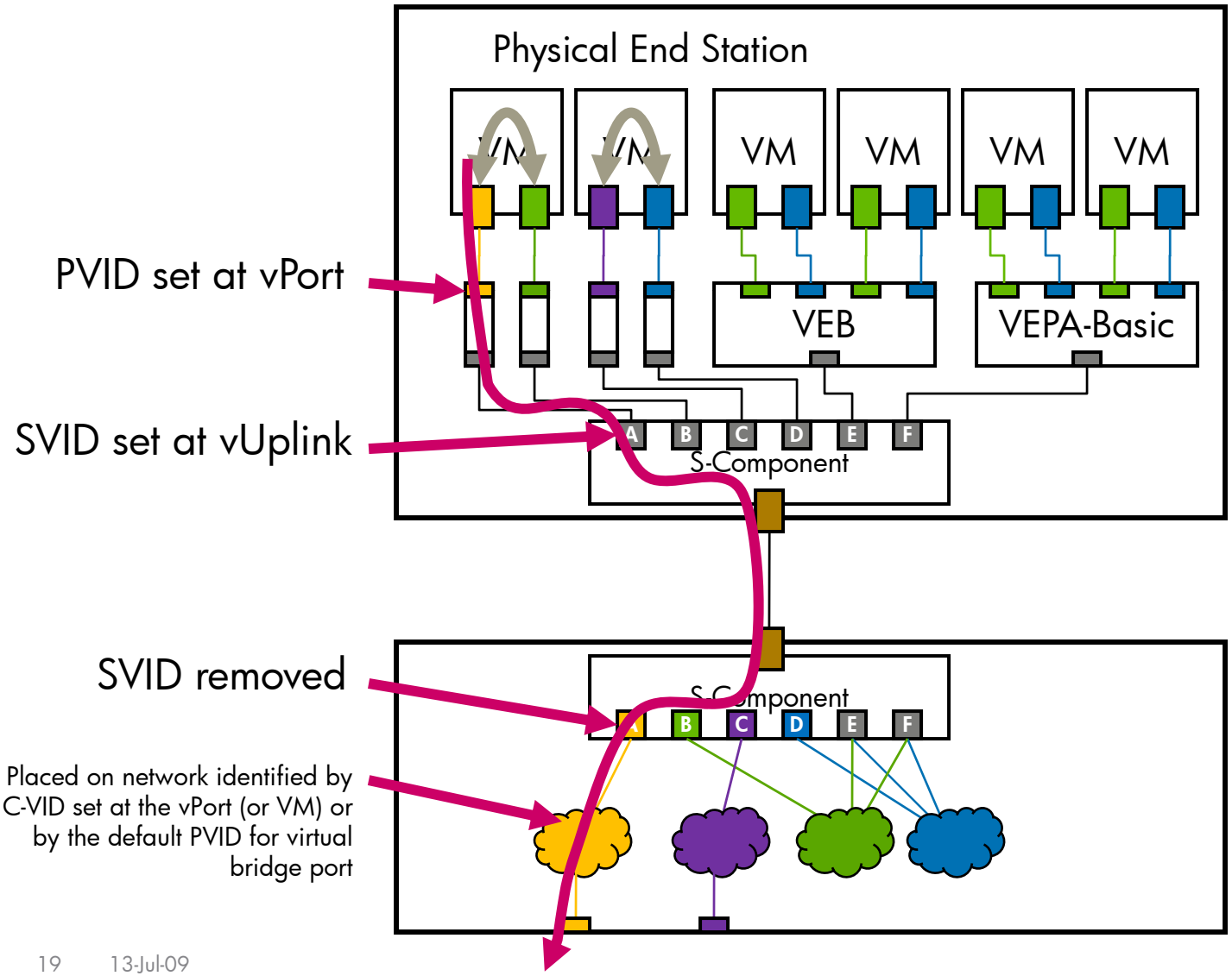
VEPA+MultiChannel

New Anatomy and Terms



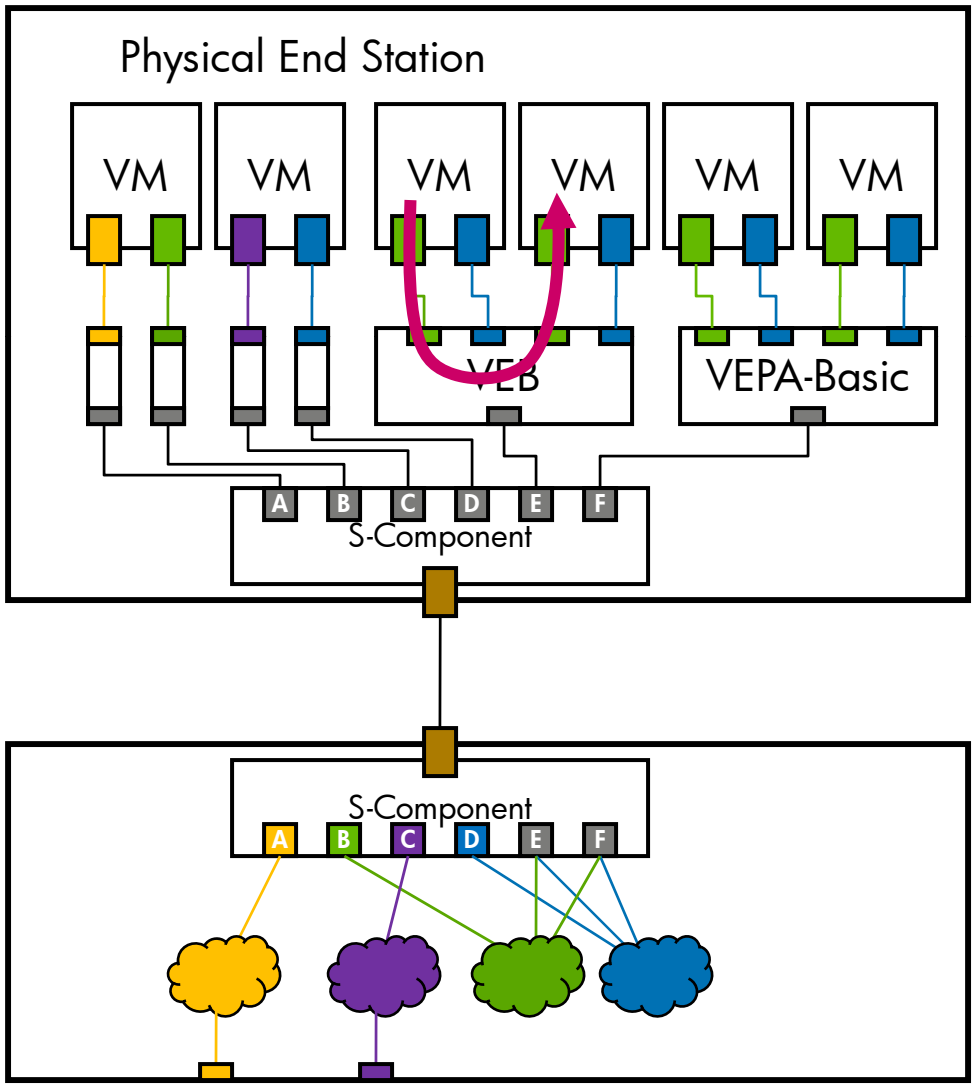
VEPA+MultiChannel Approach

Isolation vPort



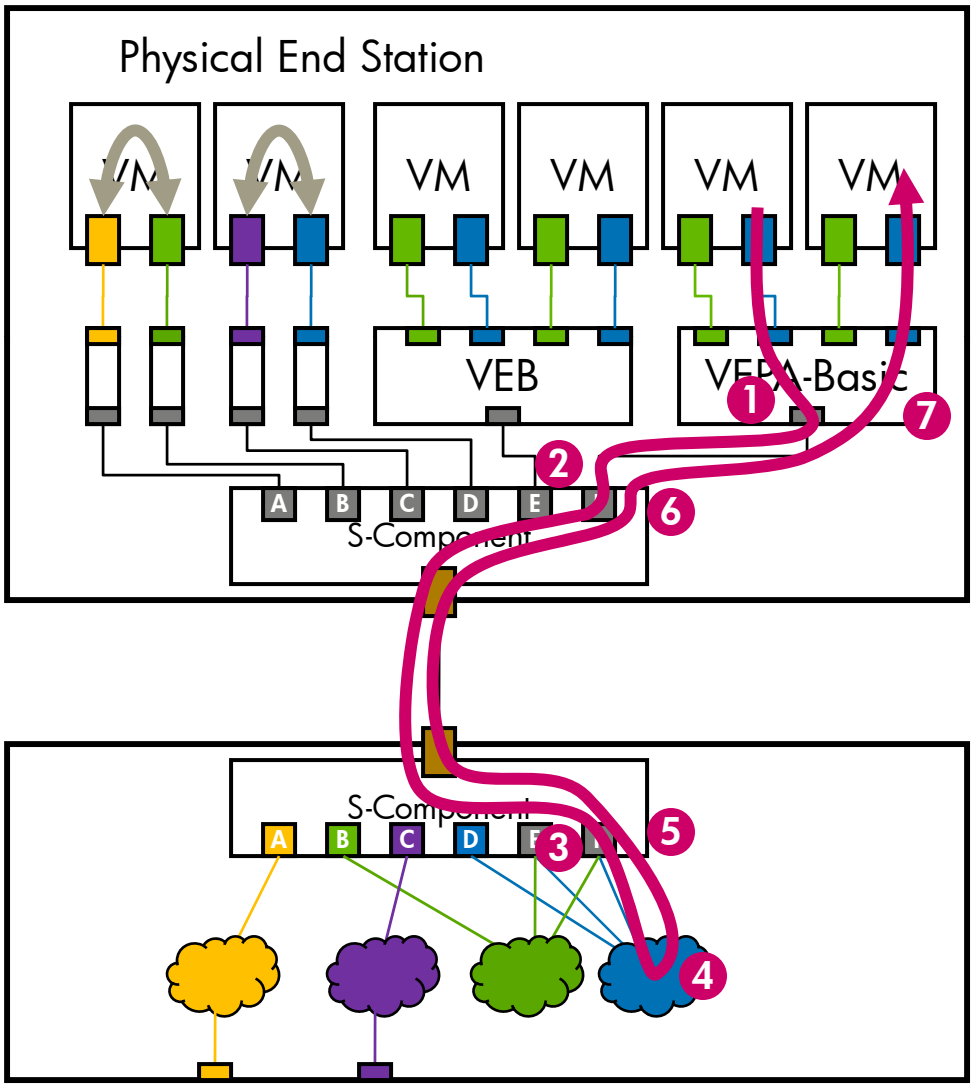
VEPA+MultiChannel Approach

Example: Basic VEB Unicast to Local VM



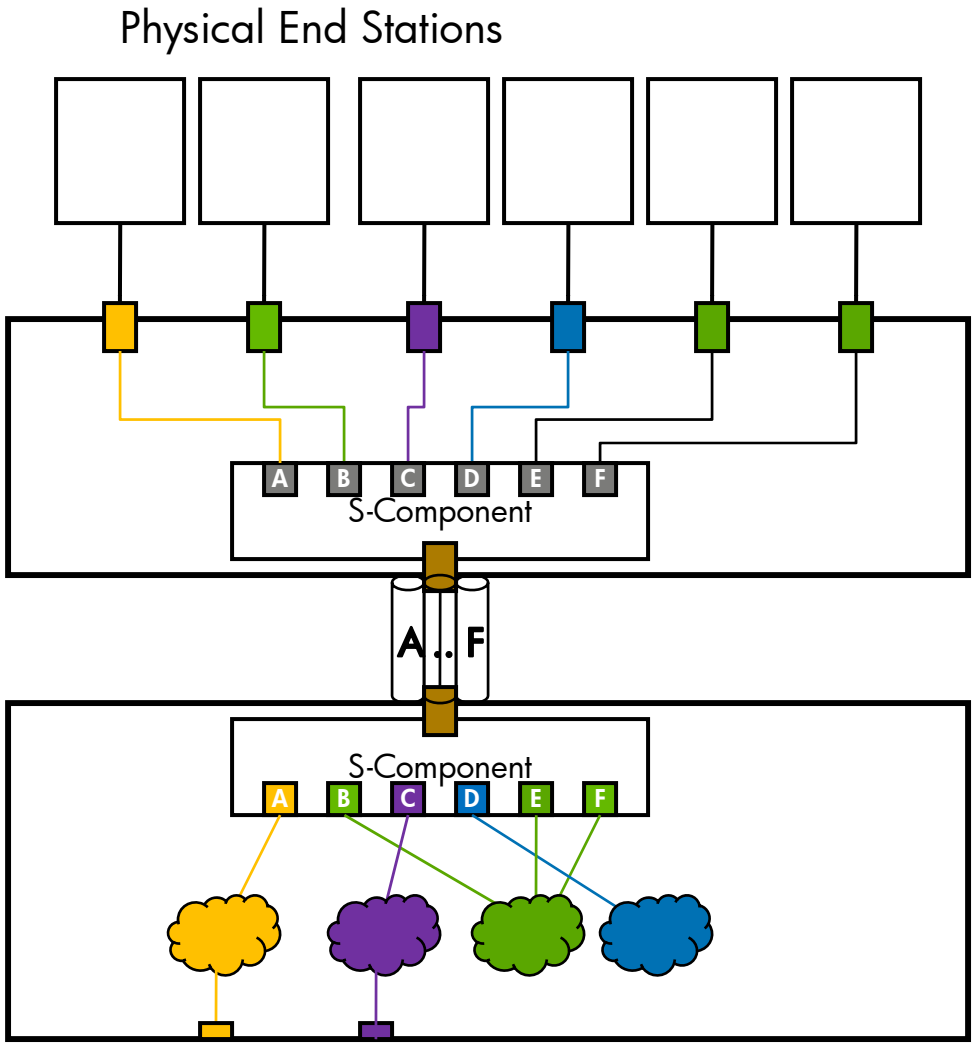
VEPA+MultiChannel Approach

Example: Basic VEPA Unicast to Local VM



1. VEPA ingress frame from VM forwarded out VEPA uplink to S-Component
2. Station S-Component adds SVID (F)
3. Bridge S-Component removes SVID (F)
4. Bridge Virtual Port is configured for VEPA mode, so it forwards based on bridge forwarding table (unblocked on virtual switch port F).
5. Bridge S-Component adds SVID (F)
6. Station S-Component removes SVID (F)
7. VEPA forwards frame based on its VEPA address table.

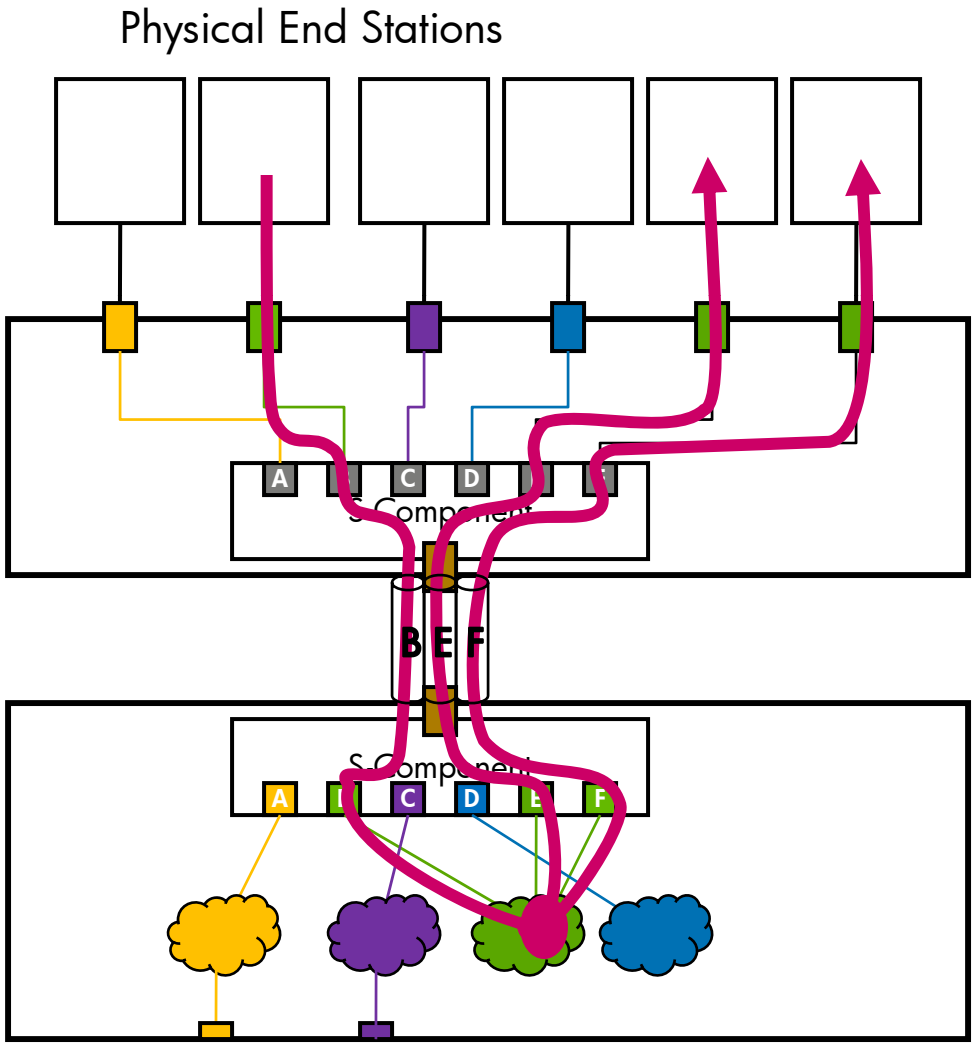
A Port Expander
with adjacent bridge multicast replication



Port Expander

Controlling Bridge

Adjacent Bridge Replicates As Needed

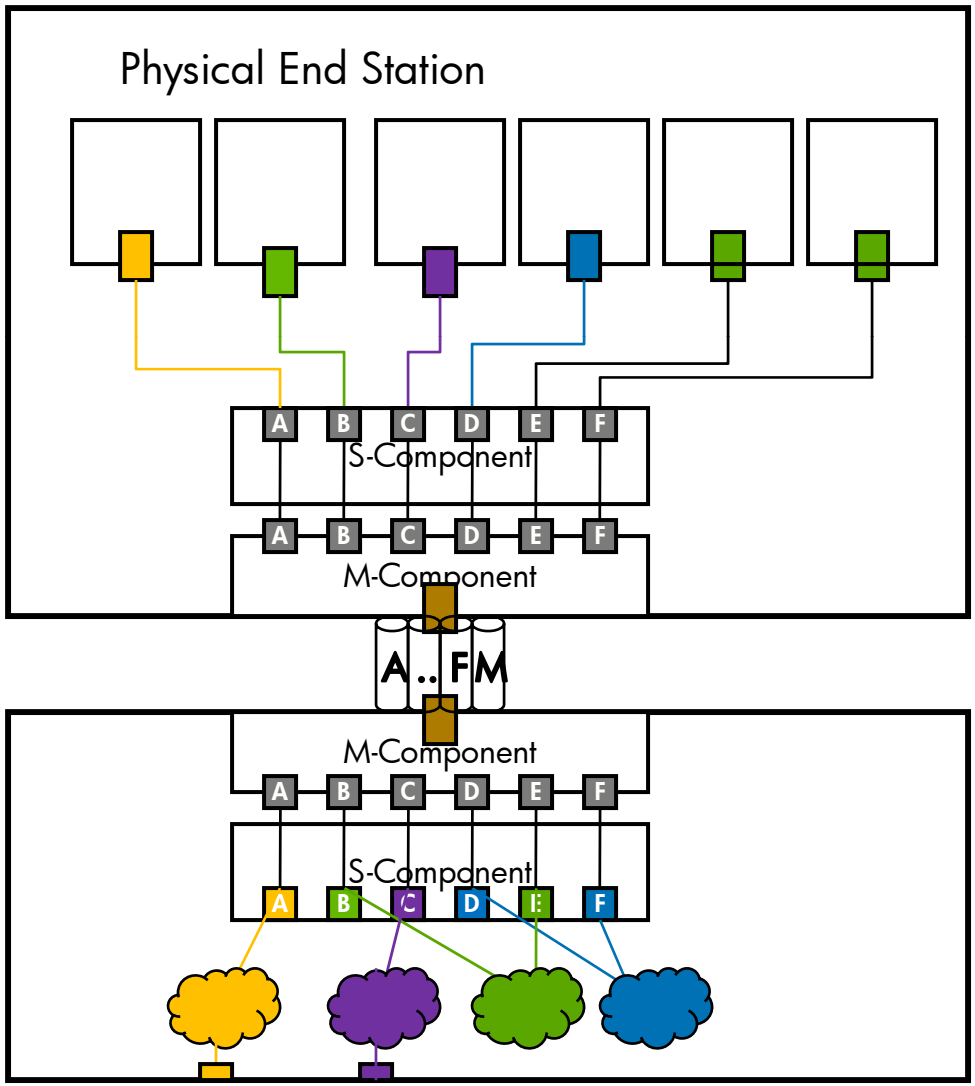


Discussion of Remote Replication Services

- VEPA+Multichannel standardization can be covered by 'hairpin' PAR and leveraging work for Remote Customer Service Interface (RSCI – expected as 802.1bc)
- Port Expander without downstream replication could be covered by RSCI PAR
- Non-address based downstream replication is the only remaining 'unique' item within VN-Tag proposal
  - NOTE: This, however, may be harder to solve than it sounds
- Both explicit Ingress+Egress indicators are only needed to support downstream replication (broadcast/multicast/flood)
- A new bridge 'component' could be envisaged to solve the problem.
  - NOTE: There may be other choices, this is just an idea
- A new 'tag' is required, but can be layered on the existing tag structure.

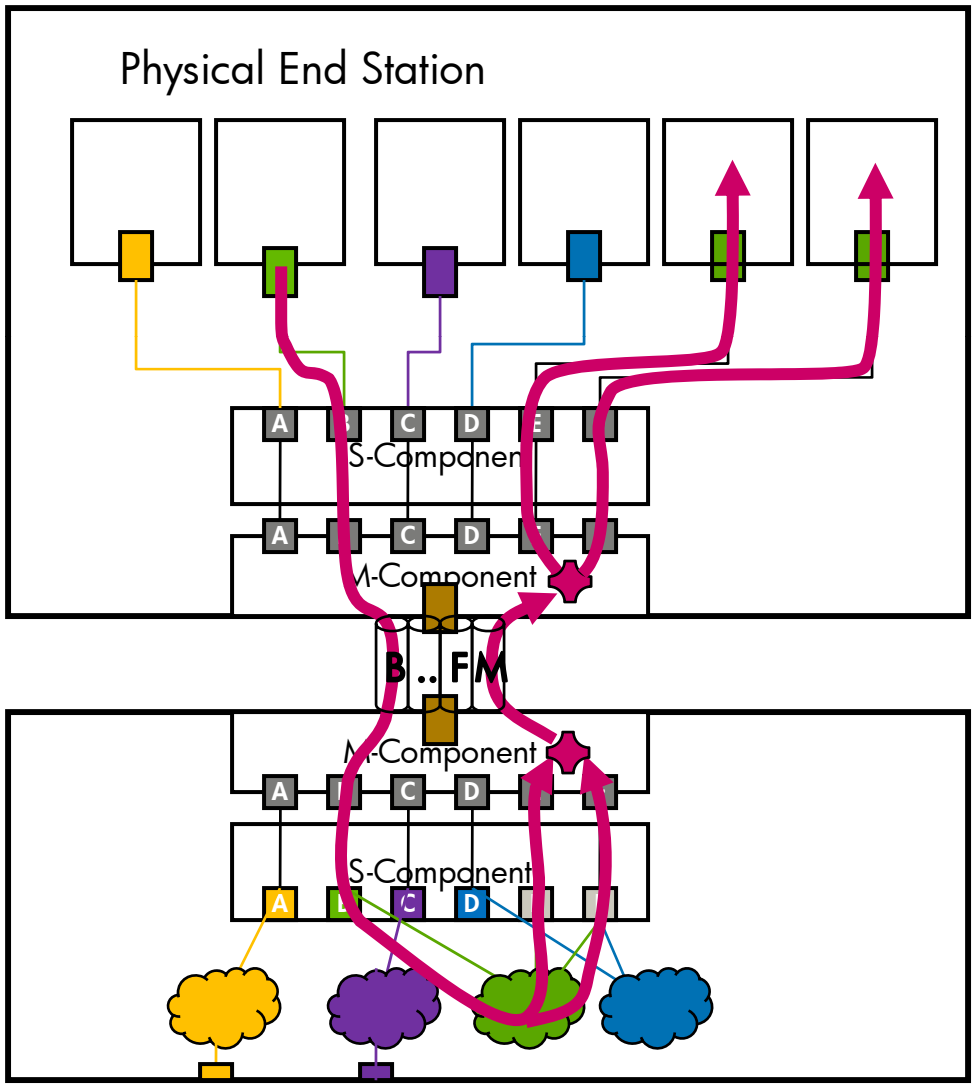
Port Expander
with Downstream Replication

NOTE: new tag required (new ethertype, mif_list), original S-VID tag left on



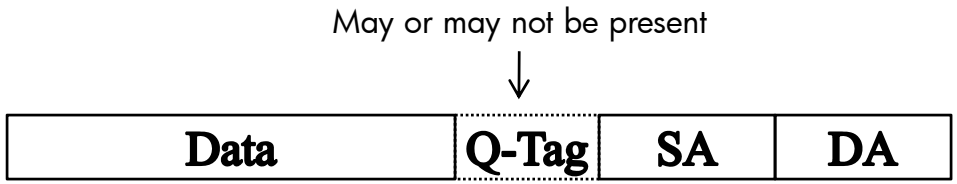
M-Component Collects and Replicates

NOTE: new tag required (new ethertype, mif_list), original S-VID tag left on



Frame Formats

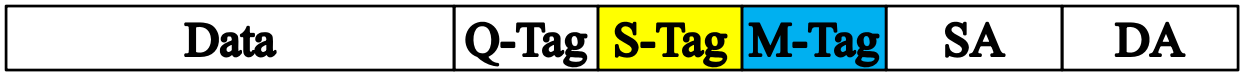
Original
(vPort <-> Port Expander)



Original+
MultiChannel
(CB <-> Port Expander)

S-Tag:
- Src vPort (PE -> CB) **OR**
- Dst vPort (CB -> PE)
- Flags

Original+
MultiChannel+
MultiCast
(CB -> Port Expander)



S-Tag:
- Src vPort
- Flags

M-Tag:
- MIF_list_ID
- Flags

Roadmap to Convergence

- Reflective Relay PAR

VEPA Basic

  - Enables hairpin forwarding on a per-port basis when VEPA is directly attached
  - Independent of MultiChannel and Remote Replication Services
- Leverage Remote Customer Services Interface PAR (802.1Qbc)

Multichannel

  - Defines a MultiChannel service to remote ports
  - Submitted for approval

Replication Tag

- Remote Replication Services
  - Defines a tag to represent a group of remote ports for which a frame is to be replicated
  - Requires a protocol to communicate tag definitions
  - Independent of Reflective Relay
  - May be dependent on MultiChannel support

Summary

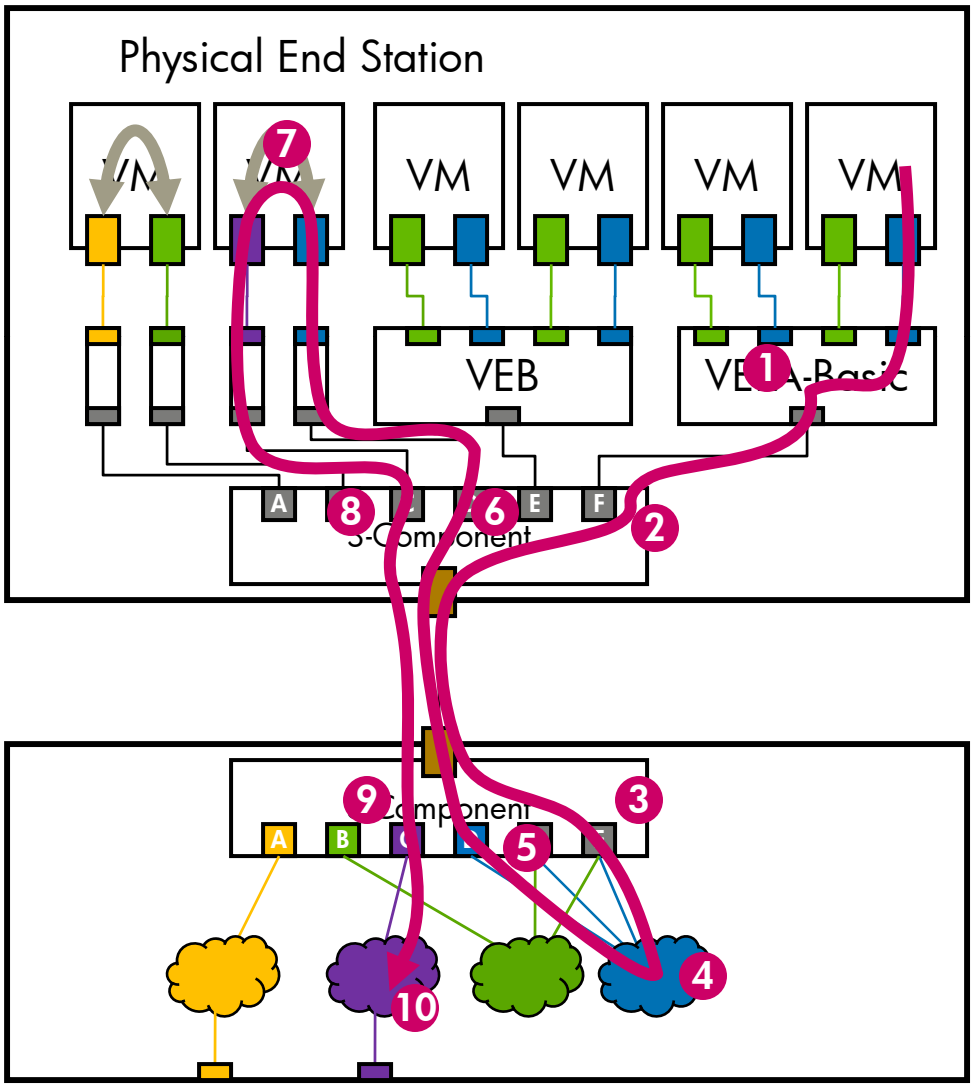
- Modularizing the problem is a good idea
  - Only implement what is needed, where needed
  - Allows use and reuse as needed
  - Allows for phased adoption
- Propose Reflective Relay PAR on Thursday for 'hairpin' forwarding
- Consider methods of supporting Remote Replication Services that are compatible with existing mechanisms

Additional Material

Not Covered in the interest of time

VEPA+MultiChannel Approach

Example: Using Transparent Service Separating Blue & Purple VLANs



1. VEPA ingress frame from VM forwarded out VEPA uplink to S-Component
2. Station S-Component adds SVID (F)
3. Bridge S-Component removes SVID (F)
4. Forwards based on bridge forwarding table to virtual switch port E.
5. Bridge S-Component adds SVID (D)
6. Station S-Component removes SVID (D)
7. Transparent service bridges across to purple VLAN.
8. Station S-Component adds SVID (C)
9. Bridge S-Component removes SVID (C)
10. Bridge forwards frame on purple VLAN.