IEEE 802 Tutorial:
Edge Virtual Bridging

November 2009

Atlanta, GA

Contributors and Supporters

Siamack Ayandeh	(3Com)	Charles R. (Rick) Maule	(consultant)
Guarav Chawla	(Dell)	Menu Menuchehry	(Marvell)
Paul Congdon	(HP)	Shehzad Merchant	(Extreme)
Dan Daly	(Fulcrum)	Vijoy Pandey	(BNT)
Claudio DeSanti	(Cisco)	Joe Pelissier	(Cisco)
Uri Elzur	(Broadcom)	Peter Phaal	(InMon)
Norm Finn	(Cisco)	Renato Recio	(IBM)
Ilango Ganga	(Intel)	Rakesh Sharma	(IBM)
Anoop Ghanwani	(Brocade)	Jeelani Syed	(Juniper)
Leonid Grossman	(Neterion)	Patricia Thaler	(Broadcom)
Chuck Hudson	(HP)	Neil Turton	(Solarflare)
Brian L'Ecuyer	(PMC-Sierra)	Manoj Wadekar	(QLogic)
Pankaj K Jha	(Brocade)	Martin White	(Marvell)
Jeffry Lynch	(IBM)	Robert Winter	(Dell)
David Koenen	(HP)		

Agenda

➢ Introduction:	Pat Thaler; Broadcom Chair IEEE 802.1 Data Center Bridging Task Group
➢ Background:	Anoop Ghanwani, Brocade
➢ Problem Statement:	Manoj Wadekar, QLogic
➢ Edge Virtual Bridging:	Paul Congdon, HP
➢ Port Extender:	Joe Pelissier, Cisco
➢ Summary, Q&A:	Pat Thaler

EVB PARs

- Two PARs for EVB work
  - Both PARs are amendments to IEEE 802.1Q
  - Both PARs have been submitted for IEEE 802 approval to forward at this meeting
  - This tutorial will describe the work we intend to do in each of these projects

- P802.1Qbg Edge Virtual Bridging
- P802.1Qbh Bridge Port Extension

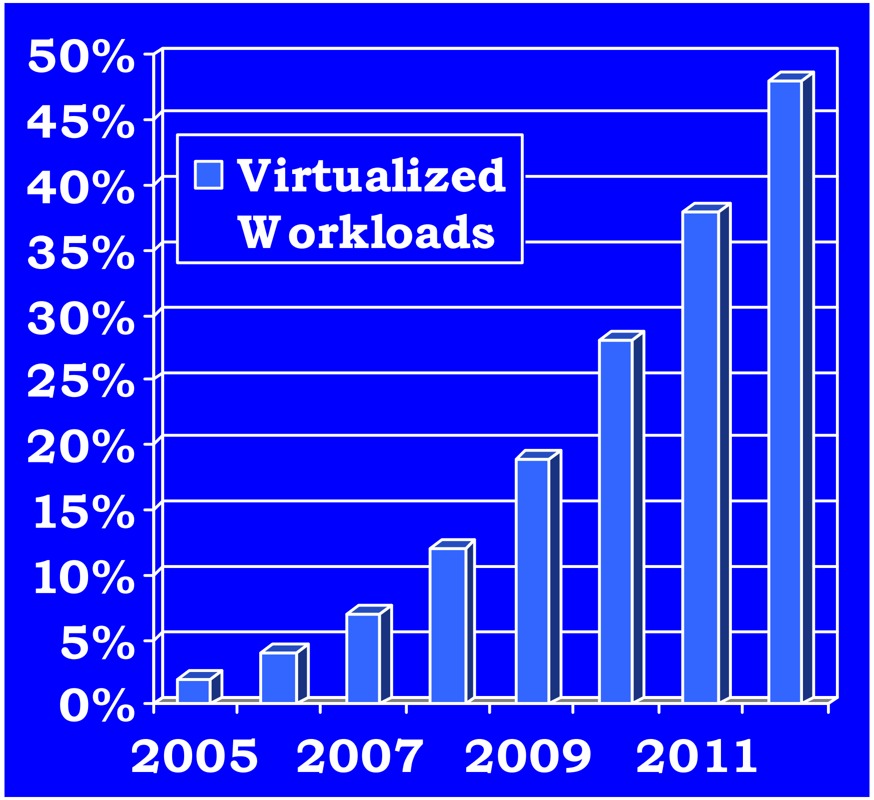
EVB Tutorial

Background:
Server Virtualization

Anoop Ghanwani (Brocade)

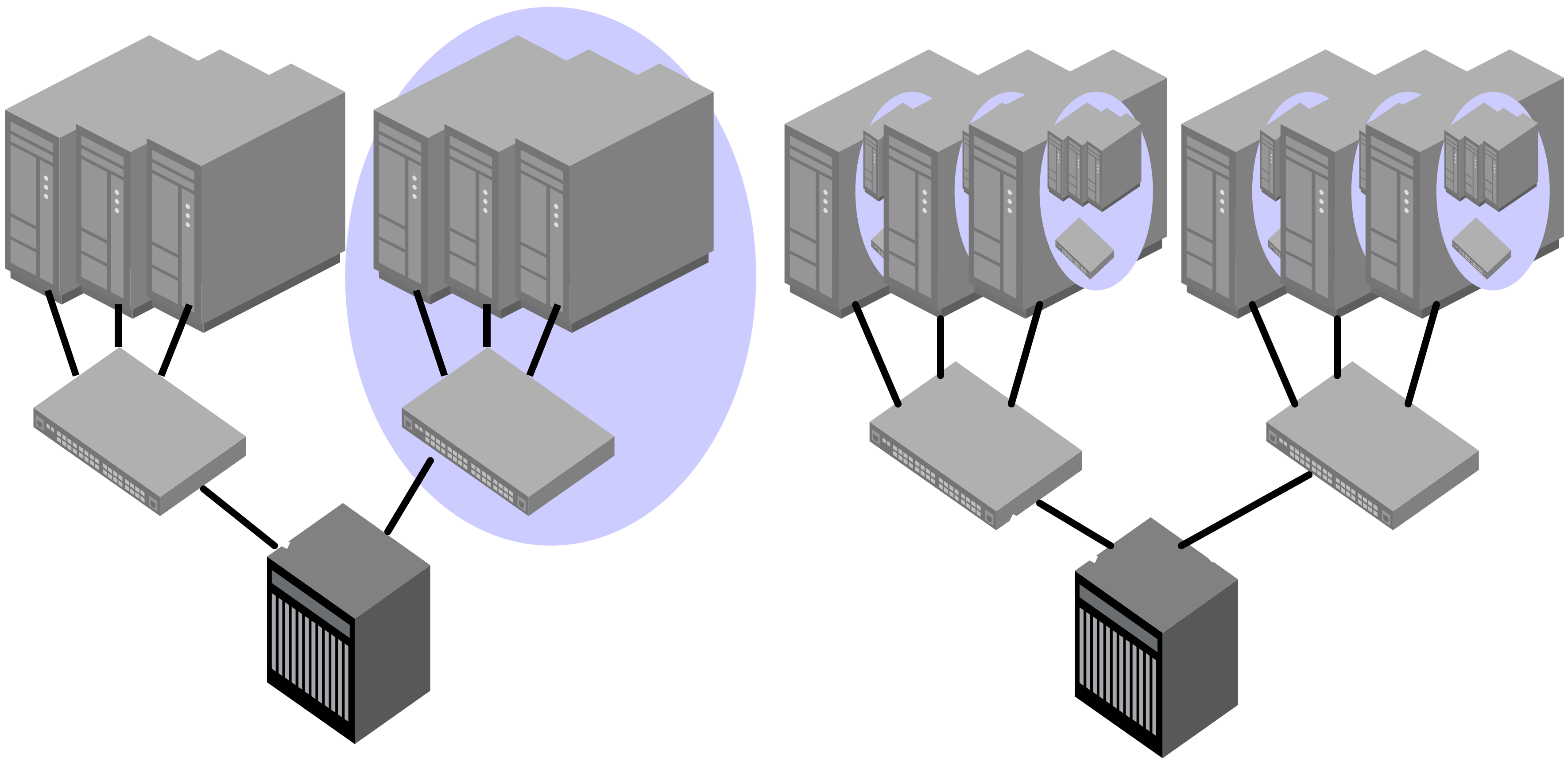
Server Virtualization is Growing Rapidly

- 50% of workloads will be virtualized by 2012
- Affects markets beyond current server virtualization vendors
  - Storage
  - Backup and Recovery
  - Application and service level management
  - Capacity planning
  - Desktop Virtualization
  - …



Source: Gartner – "Virtual Machines and Market Share Through 2012" October 2009

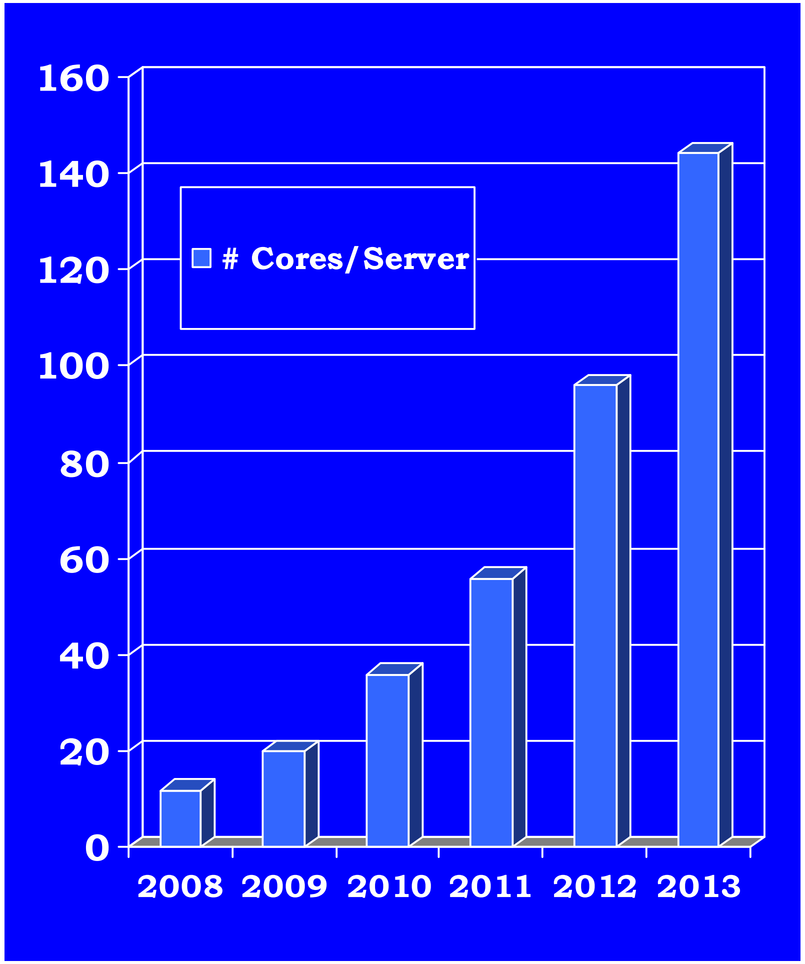
Server Virtualization and the Network



- A physical server
  - Runs multiple virtual servers called *Virtual Machines*
  - Incorporates an internal bridge for inter-VM traffic

Technology Enablers

- Processors
  - Multi-core CPUs
  - Elimination of the CPU - I/O bottleneck
  - Virtualization-enhanced processors
- Software
  - Virtualization software
  - OS/Hypervisor APIs
- Standards
  - PCI SIG SR-IOV enables high-performance IO for virtual servers



Source: TechAlpha – “Ripple Effects of Virtualization” January 2009

Drivers for Data Center Server Virtualization
Cost Savings by Server Consolidation

- Power & cooling
  - Limits # servers in a rack
  - Limits # of blades in a blade center chassis

  Increased server density

- Better resource utilization
  - CPU in servers is underutilized

  Server placement based on available server/network resources

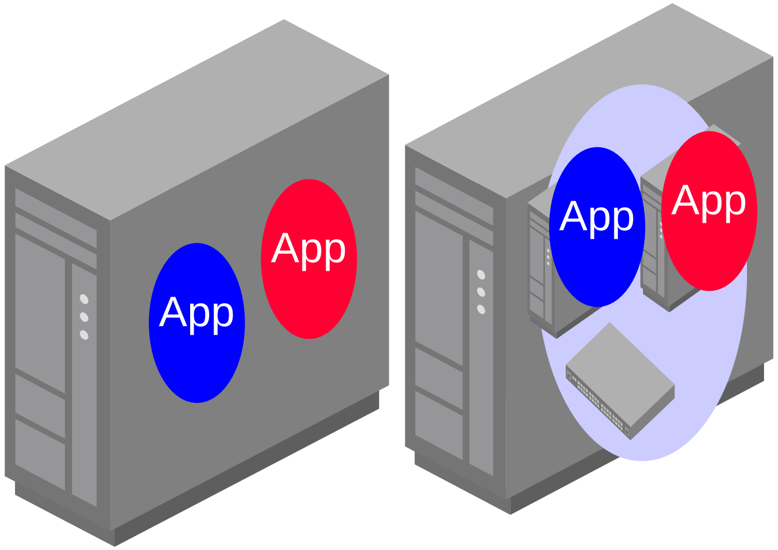
- Server administration
  - Less hardware for a given number of servers

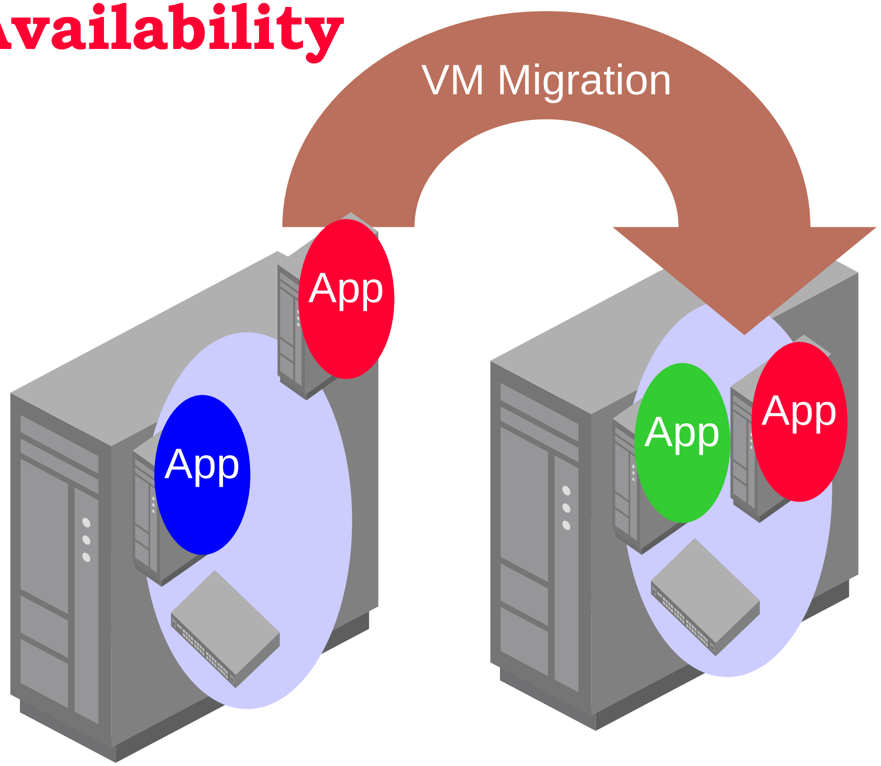
  More servers per server administrator

IEEE 802

Drivers for Data Center Server Virtualization

High Availability





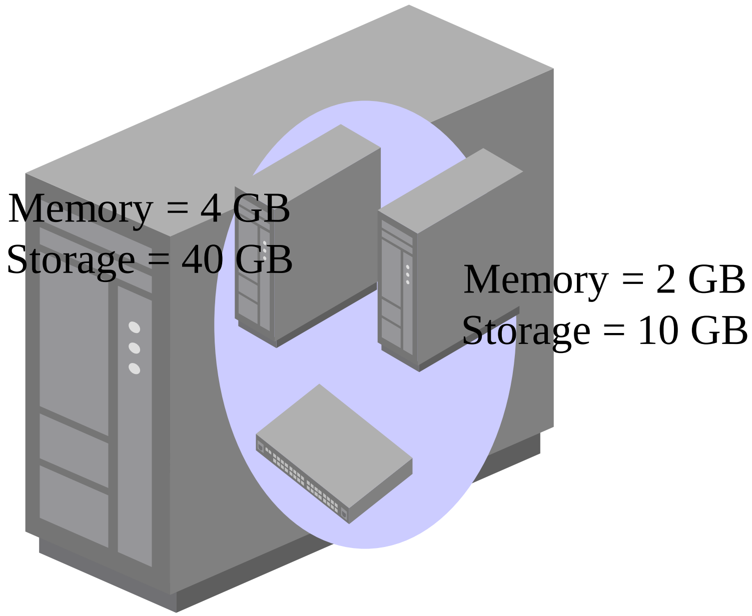
- Better application isolation
- One application per server
- Application crashing the OS becomes a non-issue

- Entire VM can be replicated even across geographical boundaries
- Transparent to users of the server
- Easier disaster recovery

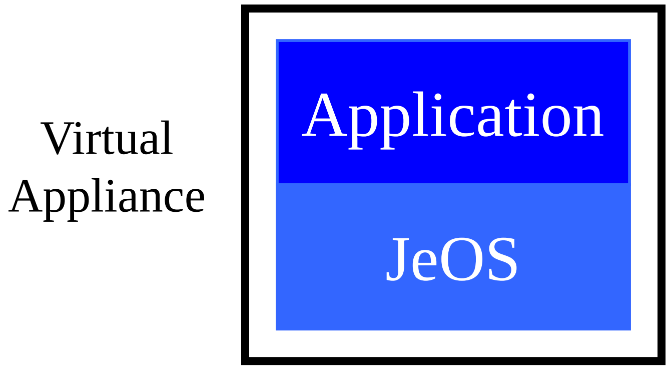
Drivers for Data Center Server Virtualization

New Service and Product Opportunities

- Cloud computing
  - Servers on demand
  - Configurable memory/hard drives
  - Pricing by the hour



- Appliance vs application
  - Application plus "just enough OS"



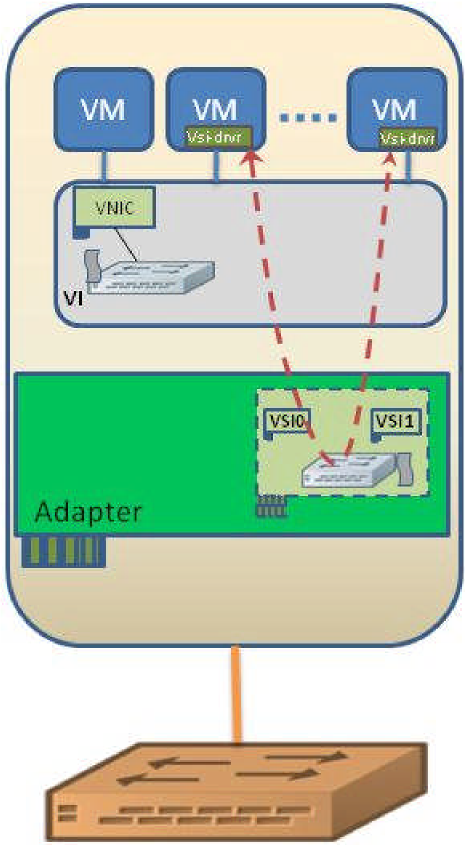
Current Offerings for Server Virtualization

- KVM (linux-kvm.org)
- VMWare
- Xen/Citrix
- Microsoft
- IBM LPARS, VPARS
- HP IVM
- Sun Solaris Containers
- …

Problem Statement

Manoj Wadekar, QLogic

IO Virtualization: Performance Challenges

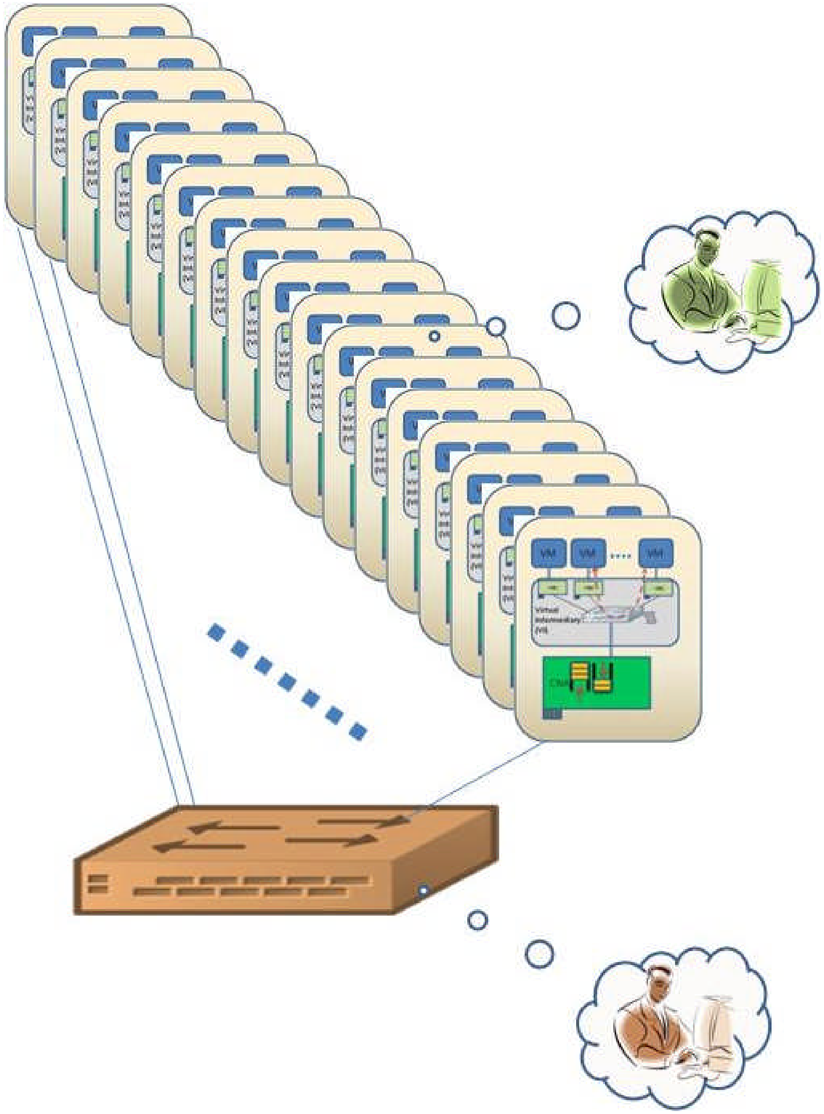


Virtual Ethernet Bridging (VEB)

- Station (desktop and server) virtualization is introducing a proliferation of Virtual Machines (VMs) that share access to a network through an embedded bridge
- IO Performance requirements have driven needs for HW assistance from IO Adapter
  - SR-IOV
  - MR-IOV
  - Embedded bridging in adapters (SW based bridging, HW based bridging in adapters)
    - Also known as Virtual Ethernet Bridging (VEB)

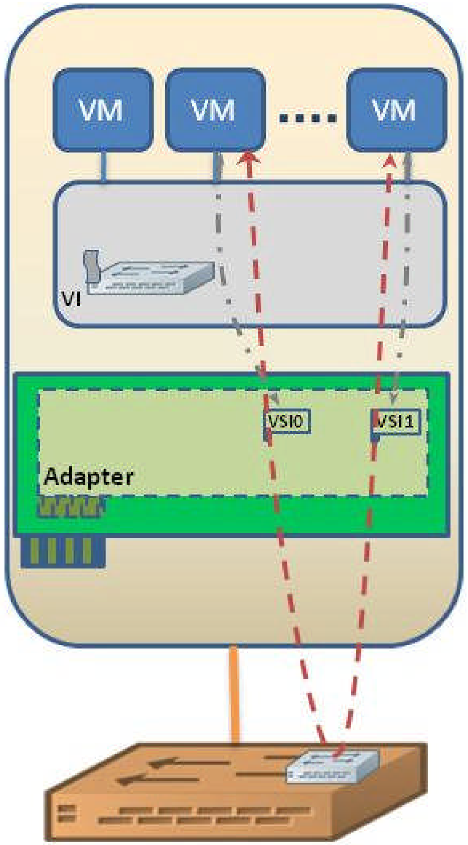
VI: Virtual Intermediary
VSI: Virtual System Interface

IO Virtualization: Management Challenges



- Management Scaling:
  - Embedded bridge in each server needs management
  - So total number of bridges requiring management in DC increases significantly
- Multiple Management Domains:
  - Different management domains for embedded bridges in servers and bridges in adjacent network
- Extended capabilities
  - Disparity between adjacent and embedded Bridge capabilities
  - Flexibility of options for allowing use of capabilities of adjacent bridge for inter-VM traffic

Gap 1: Hairpin Mode

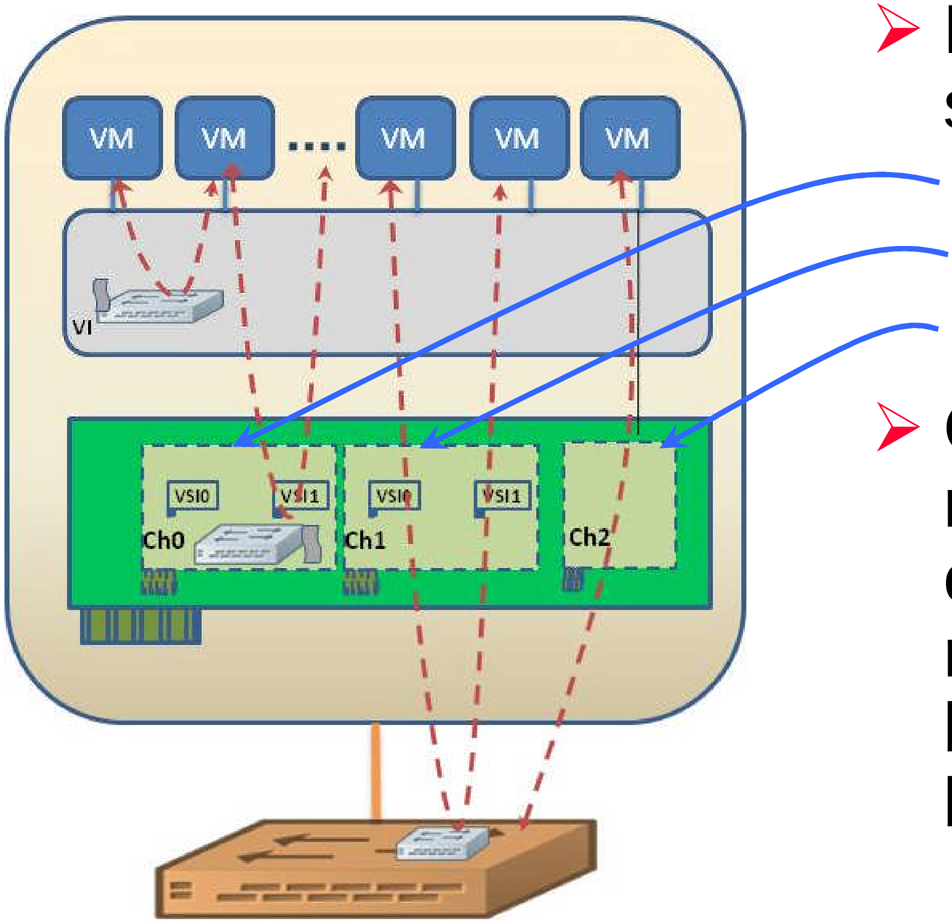


Adjacent Bridge Assist
(e.g. VEPA, PE)

- Management Challenges and need for extended capabilities can be addressed
  - By allowing that inter-VM traffic to be exposed to the relay in the adjacent bridge
- But..
  - Current 802.1 bridges do not allow packet to be sent back to same port within same VLAN
  - Current 802.1 bridges do not have visibility into identity of virtual station interfaces within physical stations

IEEE 802

Gap 2: Multi-channel Capability



- Host may be required to support multiple services
  - Embedded Bridge
  - Adjacent Bridge Assist
  - Dedicated bridge link
- Currently there is no mechanism to discover, configure and control multiple virtual links between station and bridge
  - To enable coexistence of multiple services on station-resident ports

Edge Virtual Bridging
A Definition

Edge Virtual Bridging (EVB) is the environment where physical end stations contain multiple virtual end stations that participate in the bridged LAN.

Note: EVB environments are unique in that virtual NIC configuration information is available to EVB devices that is not normally available to an 802.1Q bridge.

Technical Overview

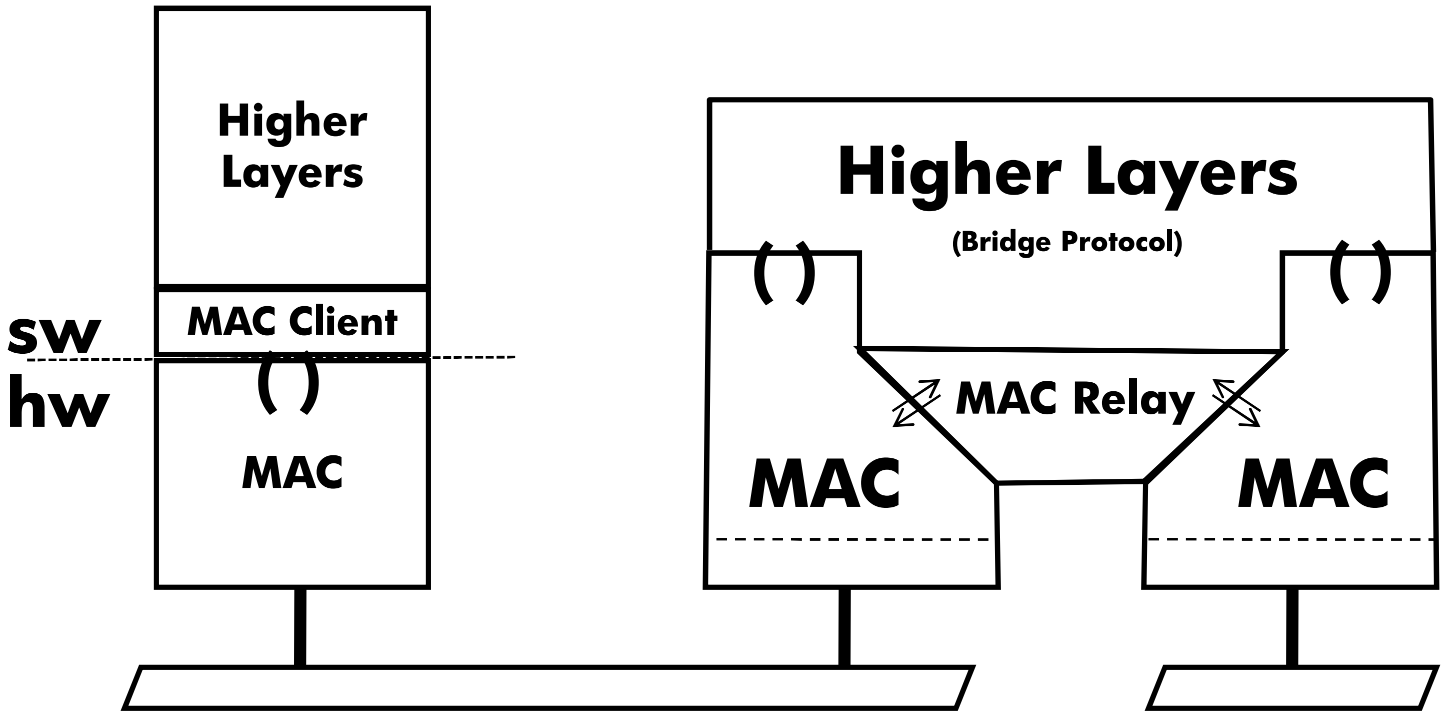
Paul Congdon (HP)
Joe Pelissier (Cisco)

Agenda

- Networking in a Virtualized Environment
- Problems in the Environment
- Solutions
  - VEBs –Virtual Ethernet Bridge
  - VEPAs – Virtual Ethernet Port Aggregator
  - Multichannel Ethernet
  - Remote Replication Services
  - PE – Port Extension
  - Discovery
- PAR Overview

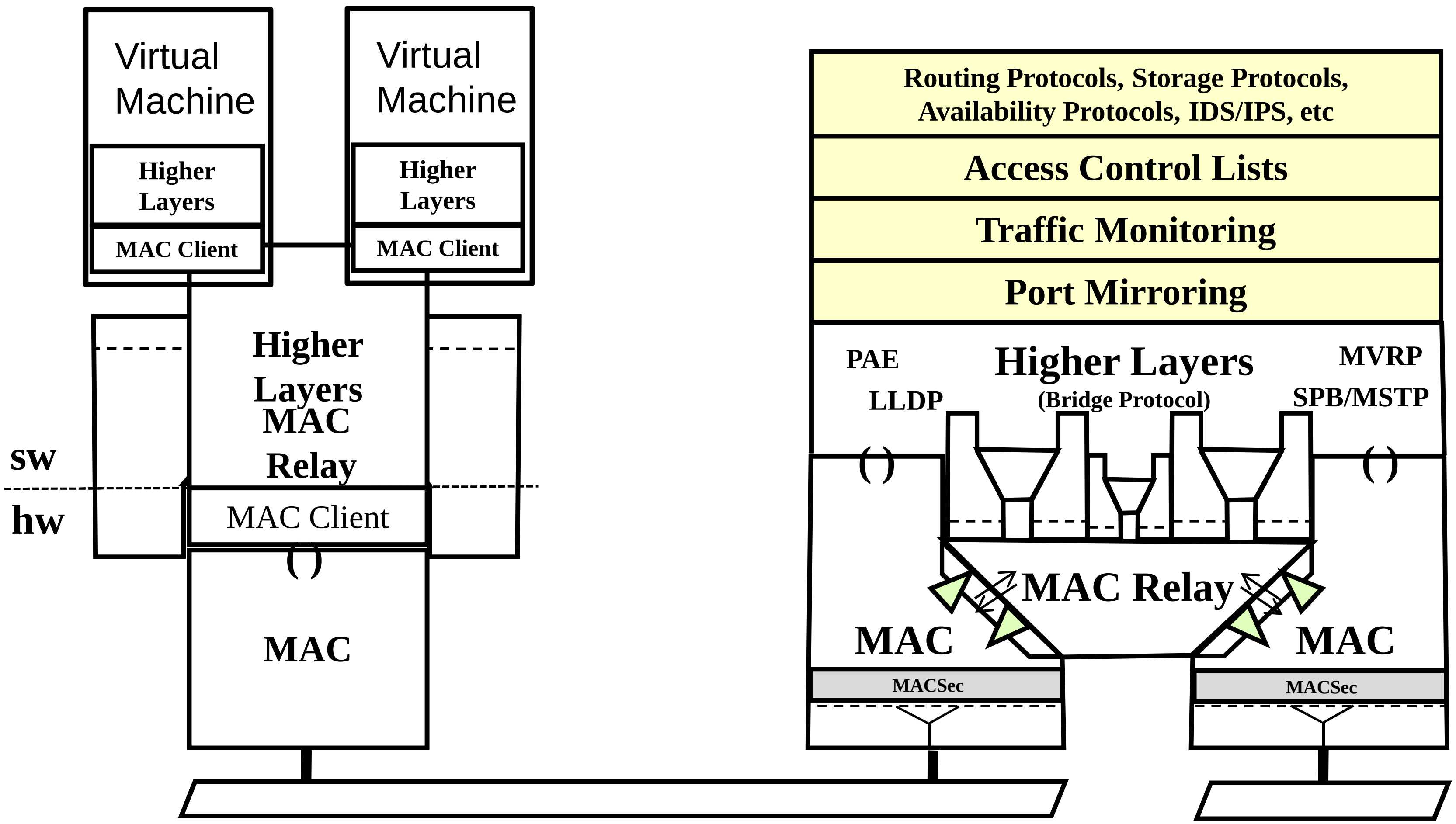
Traditional Networking

The end-station and bridge



Modern Networking

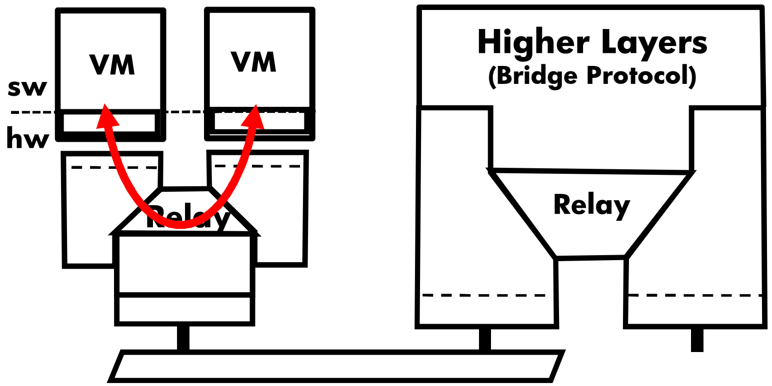
The end-station and bridge



Getting traffic to flow the way you want

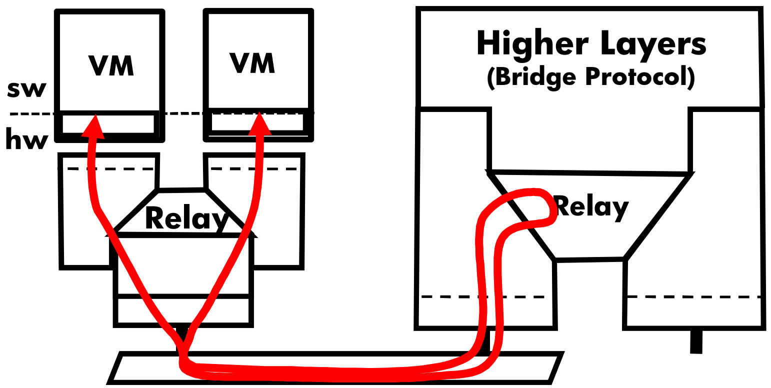
- If you prefer this…

  Fine.. It's called a "bridge" and we have standards for that, but embedded versions frequently result in difficult trade-offs between cost and capability

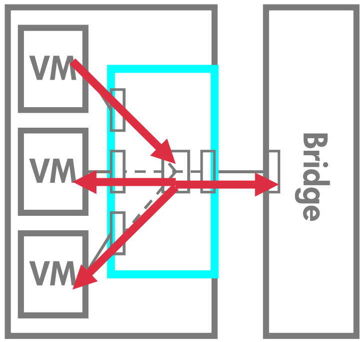


- If you prefer this…

  New forwarding modes need to be defined, and the topology is constrained



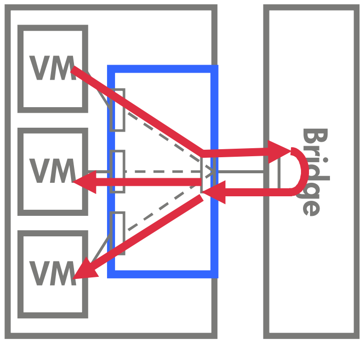
Solution Space



Virtual Ethernet Bridge (VEB)

MAC+VID to steer frames

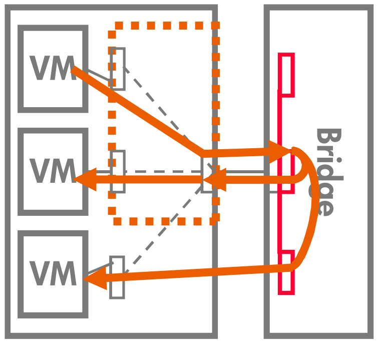
- Emulates 802.1 Bridge
- Existing implementations (vSwitch, SR-IOV bridge)
- Works with all existing bridges
- No changes to existing frame format.
- Limited bridge visibility
- Limited feature set
- Best local performance.
- Legacy, pervasive solution



Virtual Ethernet Port Aggregation (VEPA)

MAC+VID to steer frames

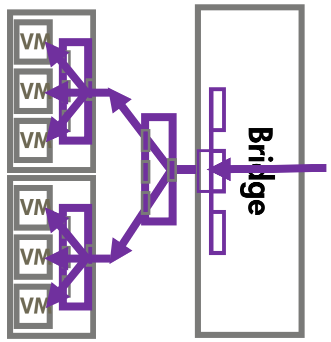
- Exploits 802.1 Bridge
- Works with many existing bridges (hairpin)
- No changes to existing frame format.
- Full bridge visibility
- Access to bridge features
- Constrained performance
- Leverages VEB resources



Multichannel

uses tag for remote ports

- Exploits Provider Bridge
- Similarities to Remote Service Interface
- Uses existing frame formats (S-tags).
- Creates bridge virtual ports
- Defines restricted S-Component
- Access to bridge features
- Adjacent bridge multicast replication (constrained performance)



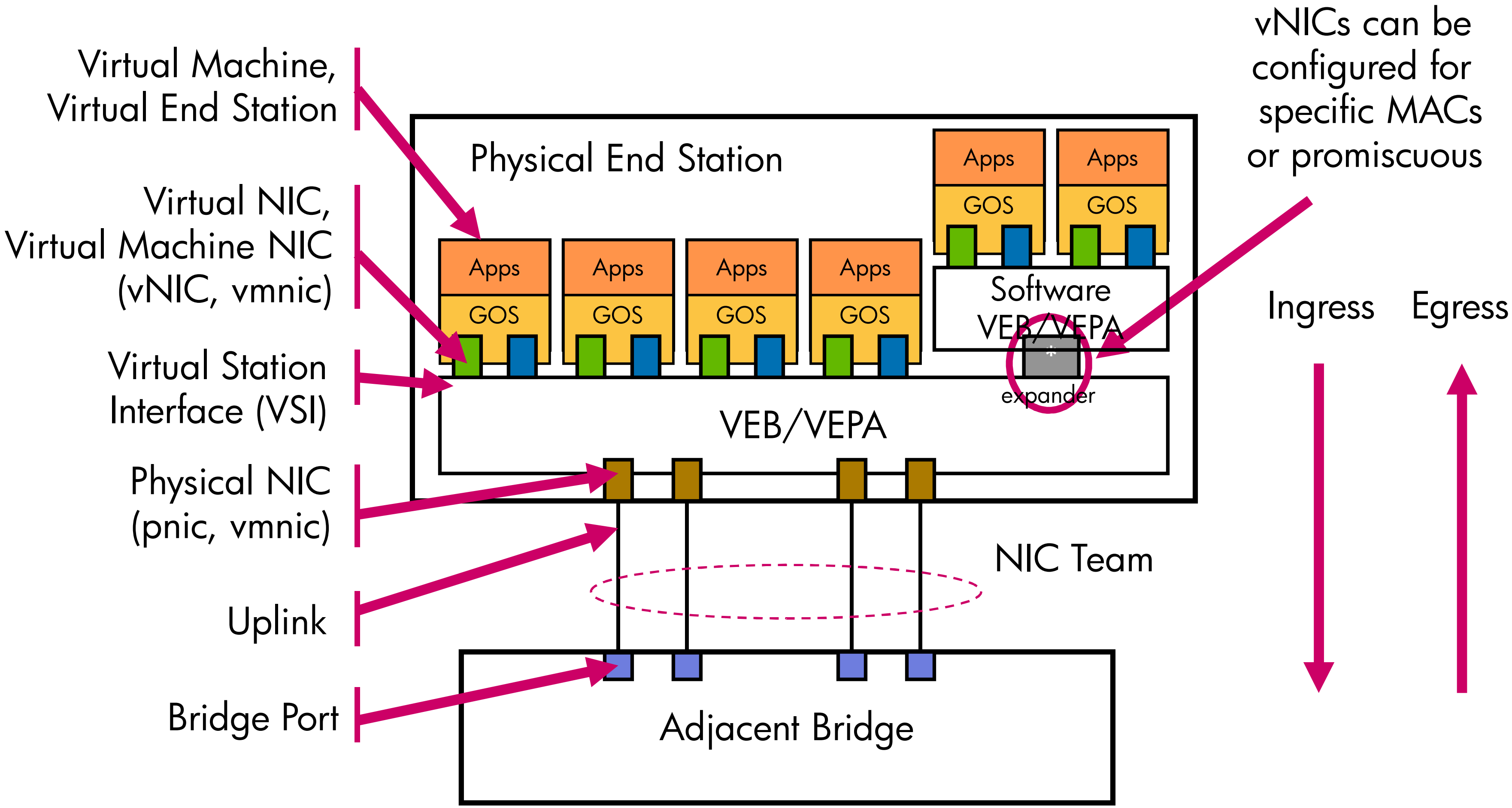
Remote Replication

uses tag to replicate packets

- Extends Multichannel
- Optimizes multicast delivery
- Enables External Cascading
- Defines new tag format
- Defines new name space

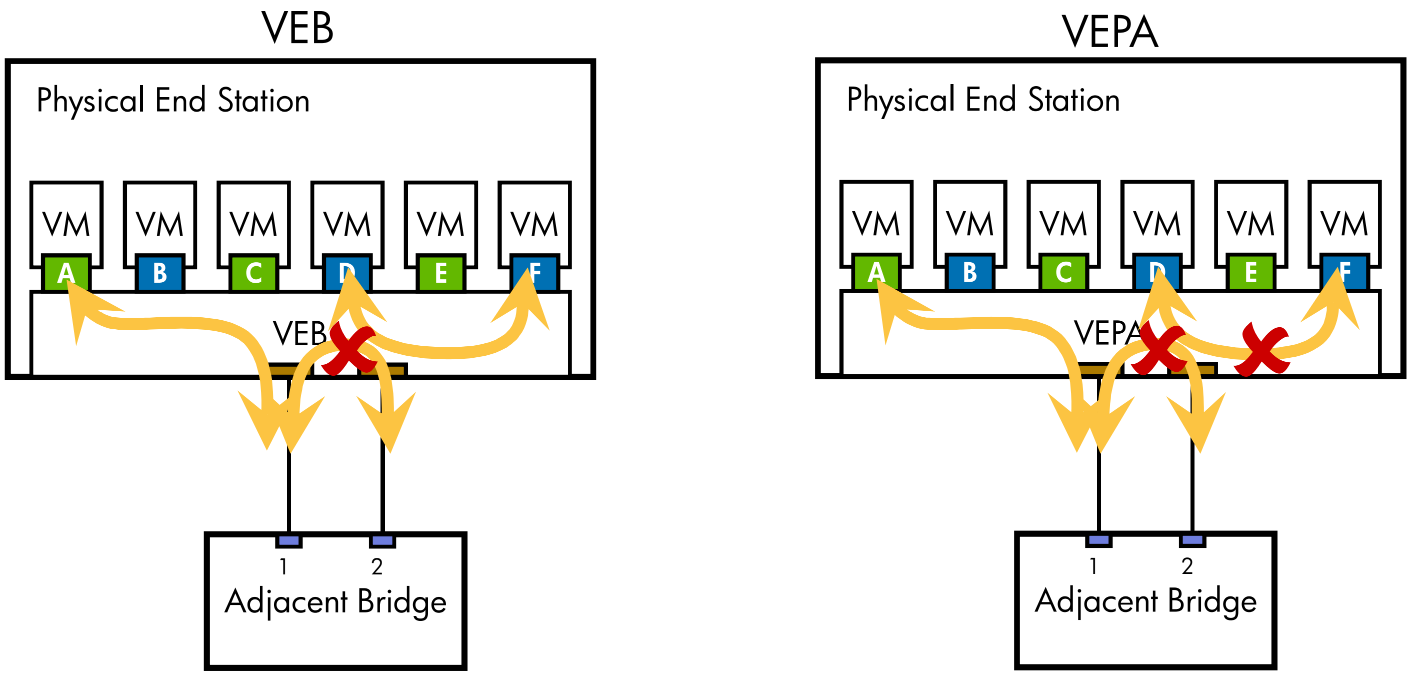
Virtual Ethernet Bridges (VEBs) Virtual Ethernet Port Aggregators (VEPAs)

Basic VEB/VEPA Anatomy and Terms



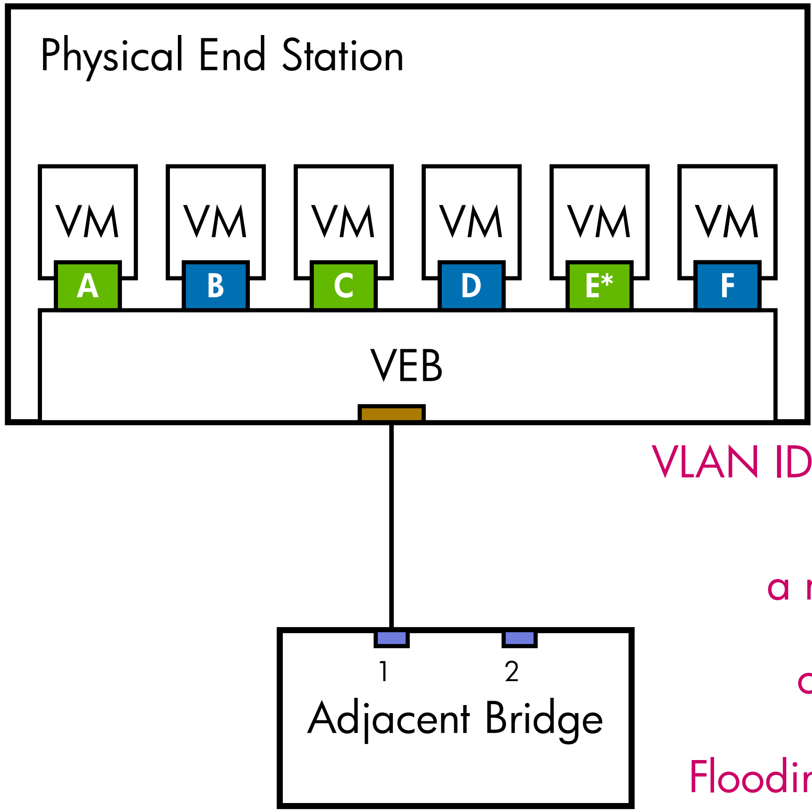
IEEE 802

Loop-free Forwarding Behavior



- Forward based on MAC address (and port group or VLAN)
- Do NOT forward from uplink to uplink
  - Single active logical uplink
  - Multiple uplinks may be ‘teamed’ (802.3ad and other algorithms)
- Do not participate in (or affect) spanning tree

VEB/VEPA Address Table



* Promiscuous VSI

VEB Address Table

via registration

Based on VLAN ID (Port Groups)

C registers a multicast listen

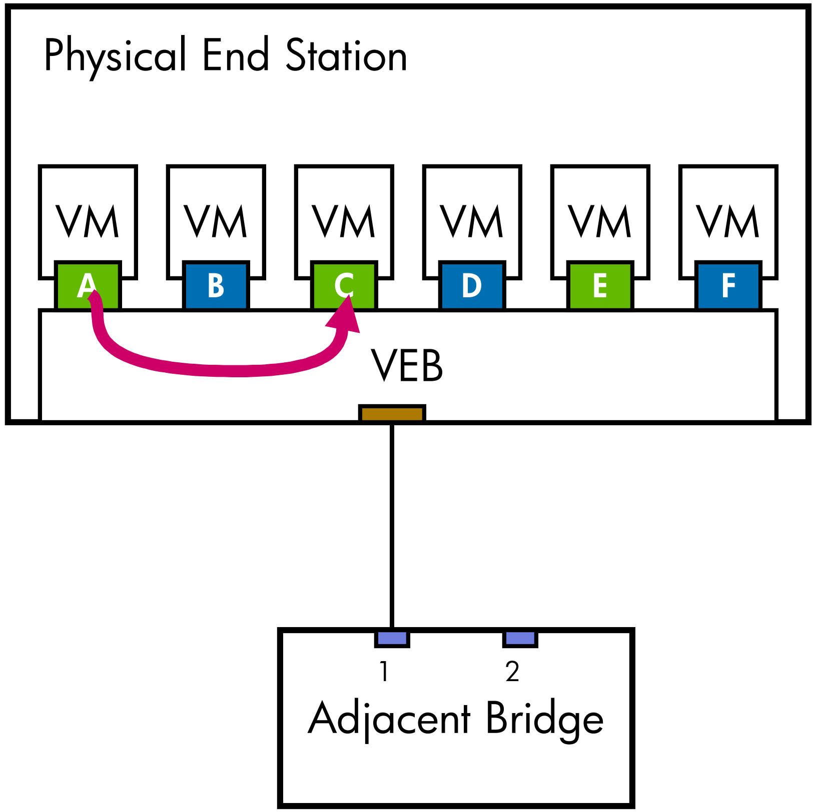
C avoids other multicasts

Flooding of unknown unicast limited to promiscuous ports and uplink

DST MAC	VLAN	Copy To (ABCDEF Up)
A	1	100000 0
B	2	010000 0
C	1	001000 0
D	2	000100 0
E	1	000010 0
F	2	000001 0
Bcast	1	101010 1
Bcast	2	010101 1
MulticastC	1	101010 1
Unk Mcast	1	100010 1
Unk Mcast	2	010101 1
Unk Ucast	1	000010 1
Unk Ucast	2	000000 1

VEB Unicast Example

SRC = A; DST = C

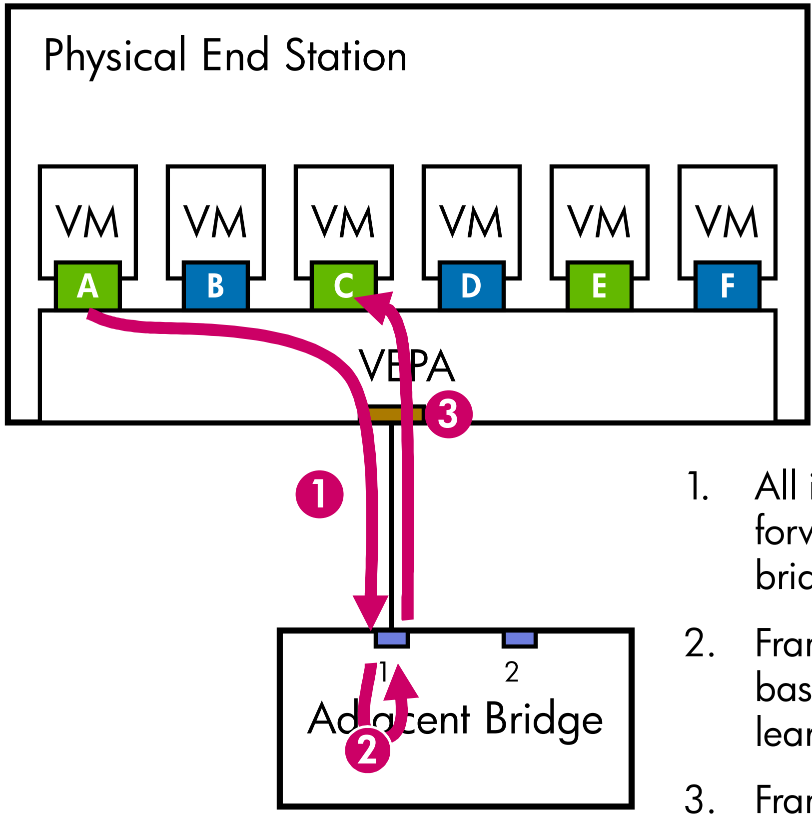


VEB Address Table

DST MAC	VLAN	Copy To (ABCDEF Up)
A	1	100000 0
B	2	010000 0
C	1	001000 0
D	2	000100 0
E	1	000010 0
F	2	000001 0
Bcast	1	101010 1
Bcast	2	010101 1
MulticastC	1	101010 1
Unk Mcast	1	100000 1
Unk Mcast	2	010101 1
Unk Ucast	1	000000 1
Unk Ucast	2	000000 1

VEPA Unicast Example

SRC = A; DST = C



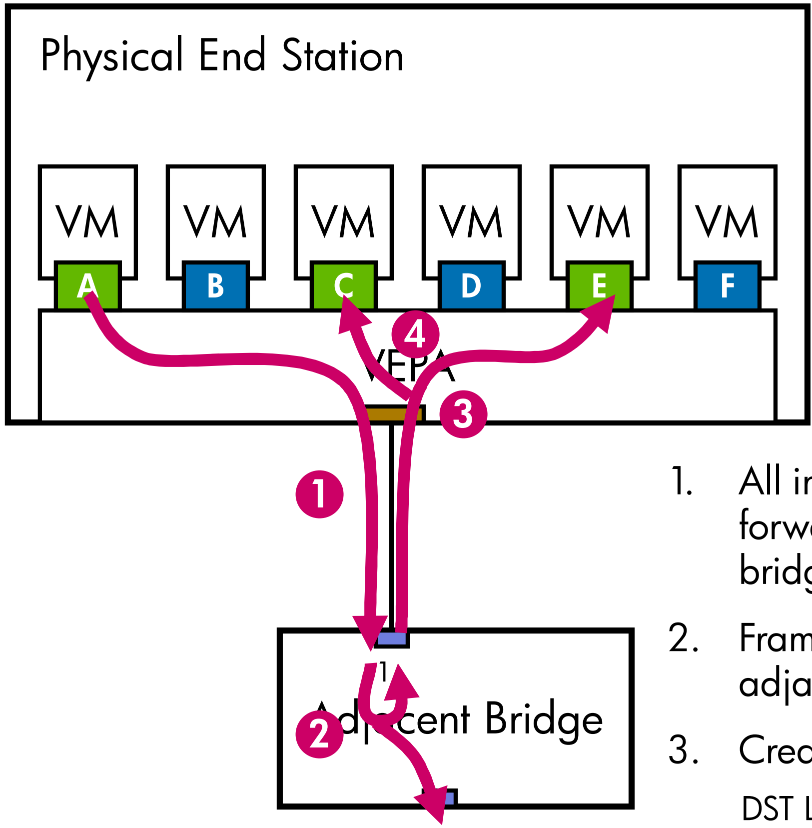
1. All ingress frames forwarded to adjacent bridge
2. Frame forwarded based on adj. bridge learning.
3. Frame forwarded based on delivery mask generated from VEPA address table

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

VEPA Multicast Example

SRC = A; DST = MulticastC



1. All ingress frames forwarded to adjacent bridge
2. Frame forwarded by adjacent bridge.
3. Create delivery mask

   DST Lookup = 101010
   SRC Lookup = 100000
   Delivery Mask = 001010

4. Deliver Frame Copies

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

Benefits of VEB/VEPA Solution

- VEPA is a simple extension to VEB
  - Similar port configuration
  - Similar address table
  - Minor changes to frame forwarding behavior
- VEPA addresses many of the limitations with VEBs
  - Exposes traffic to external bridge
  - Eliminates unnecessary flooding to promiscuous VMs
- Easy migration between VEB and VEPA modes
  - Simultaneous operation of VEB and VEPA
- Straight forward to implement
  - “Hairpin mode” may be implemented in many existing bridges with a firmware upgrade
  - Logical extension to existing vSwitches/VEBs

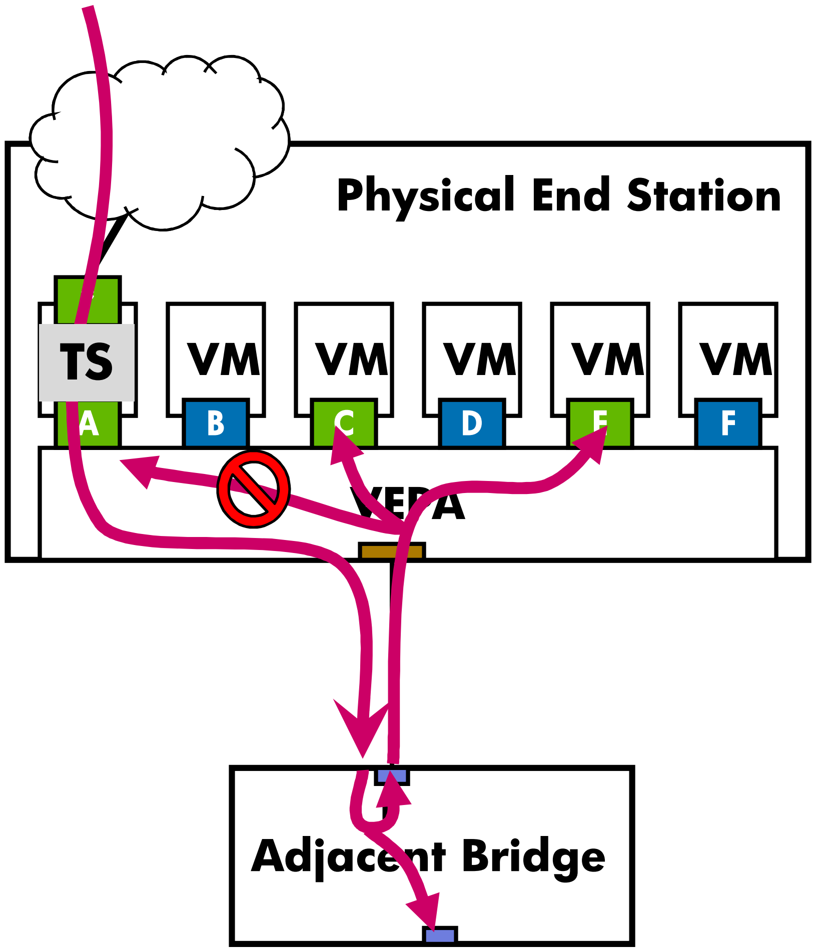
IEEE 802

'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
- Can't mix VEPA, VEB, and directly accessible ports on single physical link
  - Allow for optimized performance configuration
- Doesn't support hierarchy to unrestricted physical ports.

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

Tagging Scheme Extensions

- Filtering conditions is addressed by ‘isolating’ the Virtual Station Interfaces (VSI’s)

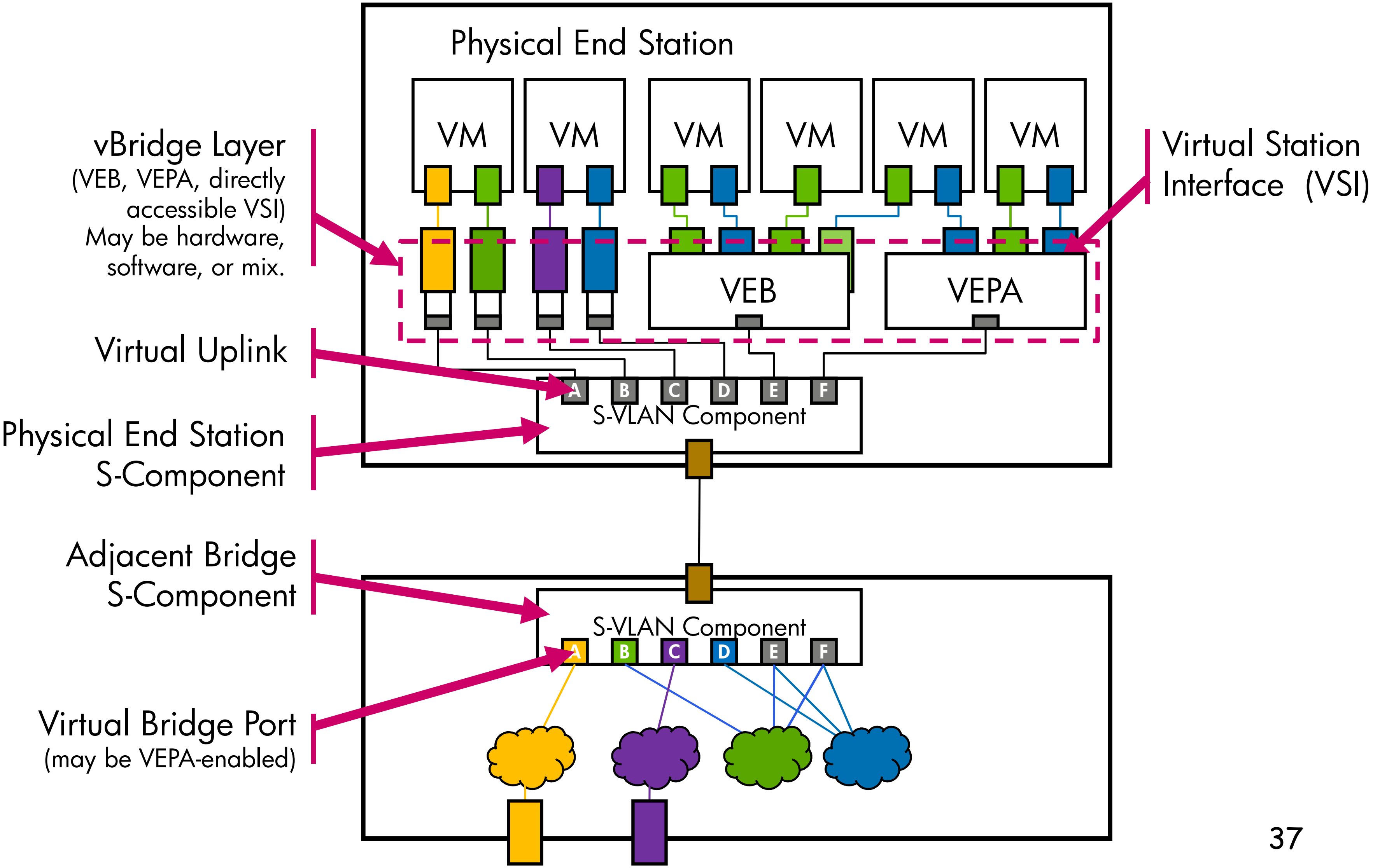
- Tagging schemes provide a virtual port indication for the adjacent bridge

- Normal bridge learning and flooding are extended to isolated VSIs

MultiChannel

MultiChannel

New Anatomy and Terms



IEEE 802

MultiChannel Approach

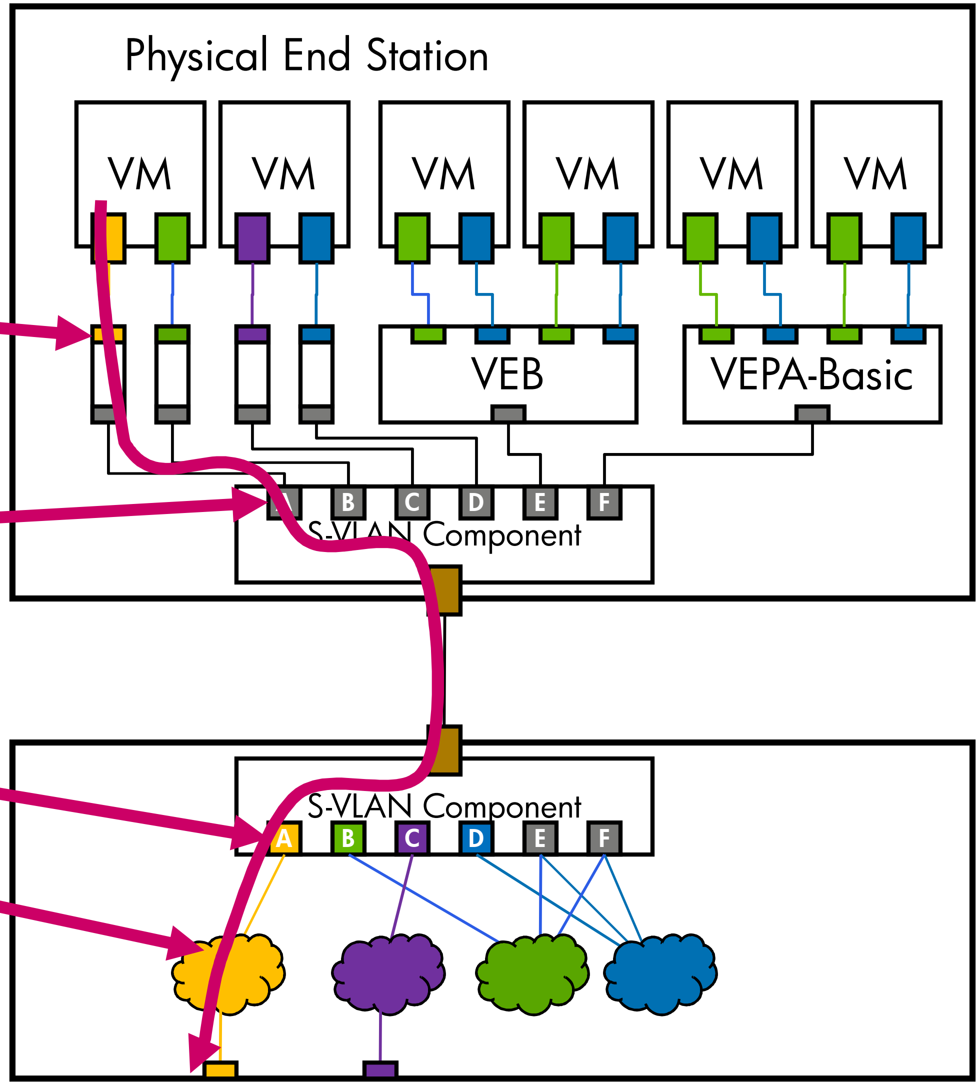
Directly Accessible VSI

PVID set at VSI

SVID set at vUplink

SVID removed

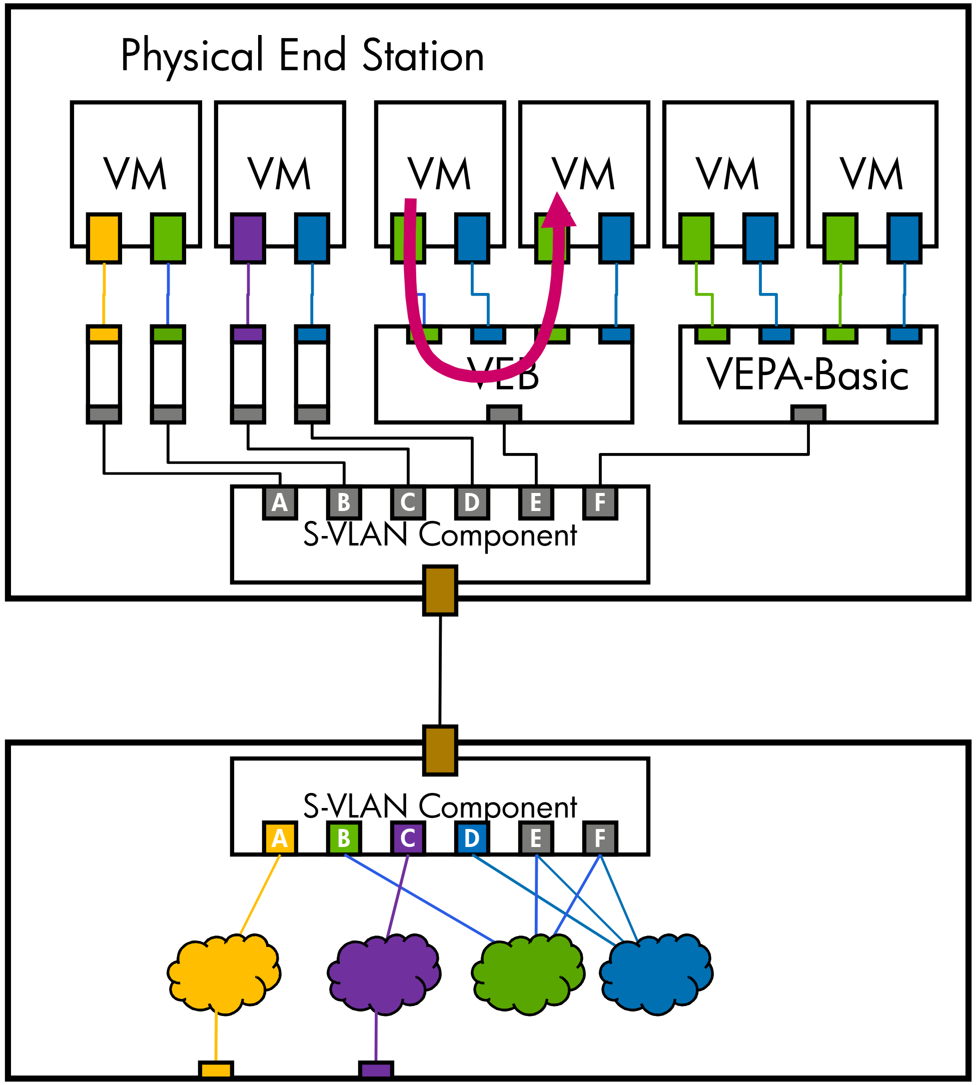
Placed on network identified by C-VID set at the VSI (or VM) or by the default PVID for virtual bridge port



IEEE 802

MultiChannel Approach

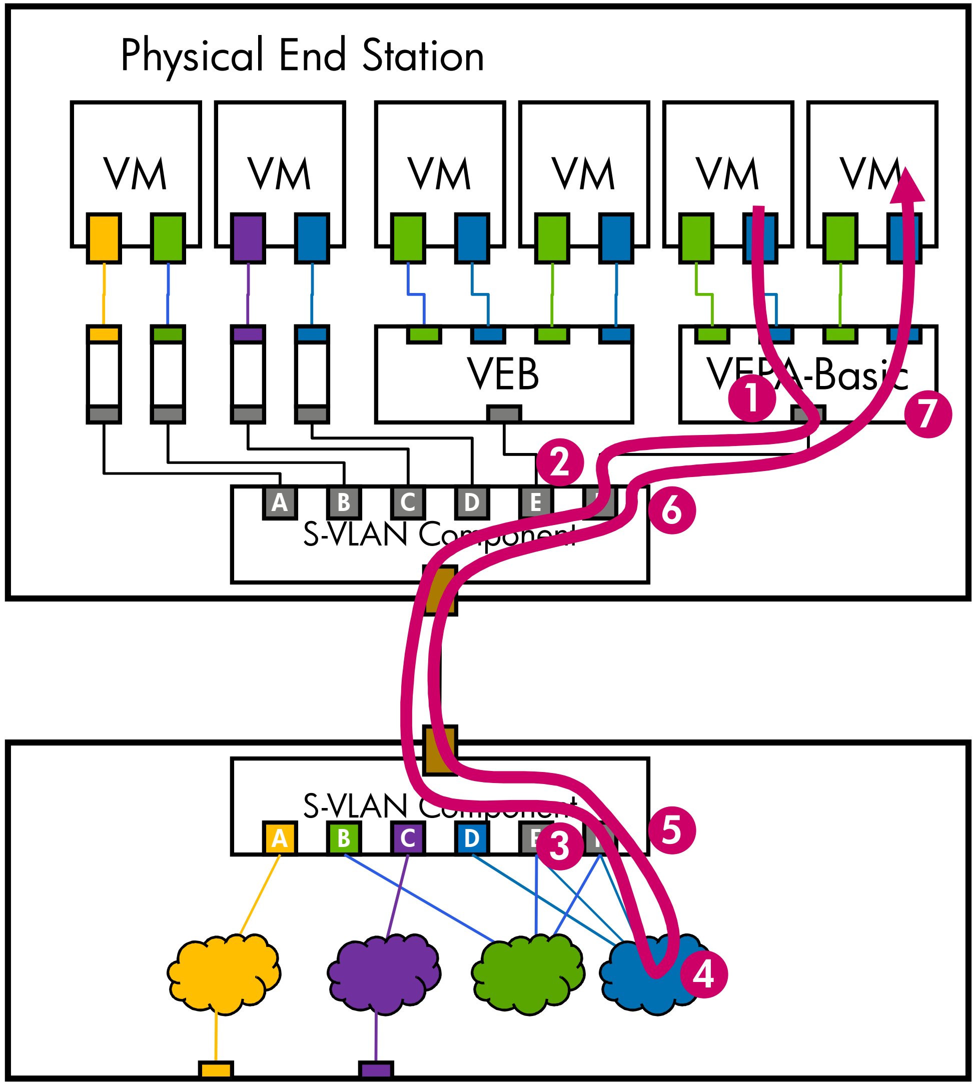
Example: Basic VEB Unicast to Local VM



IEEE 802

MultiChannel Approach

Example: Basic VEPA Unicast to Local VM

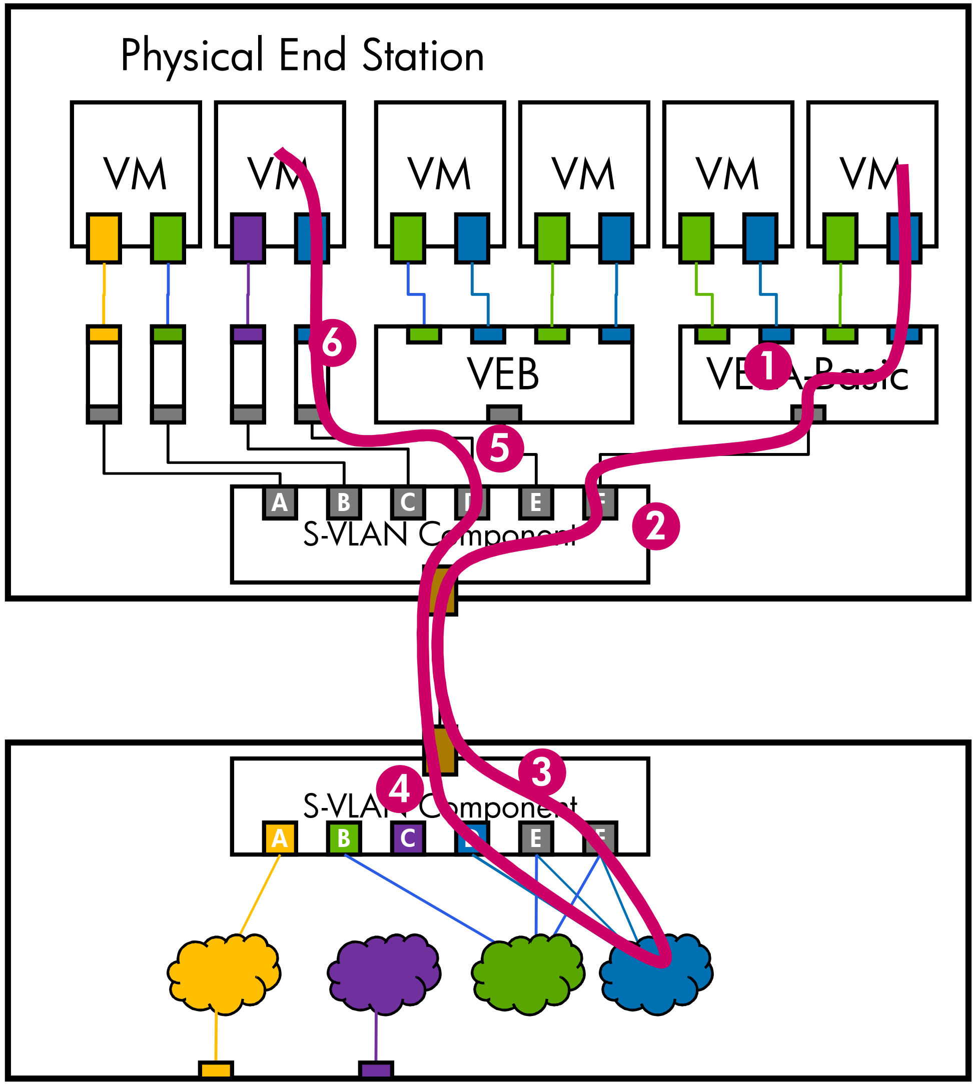


1. VEPA ingress frame from VM forwarded out VEPA uplink to S-Component
2. Station S-Component adds SVID (F)
3. Bridge S-Component removes SVID (F)
4. Bridge Virtual Port is configured for VEPA mode, so it forwards based on bridge forwarding table (unblocked on virtual bridge port F).
5. Bridge S-Component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

IEEE 802

MultiChannel Approach

Example: VM through VEPA to Directly Accessible VSI



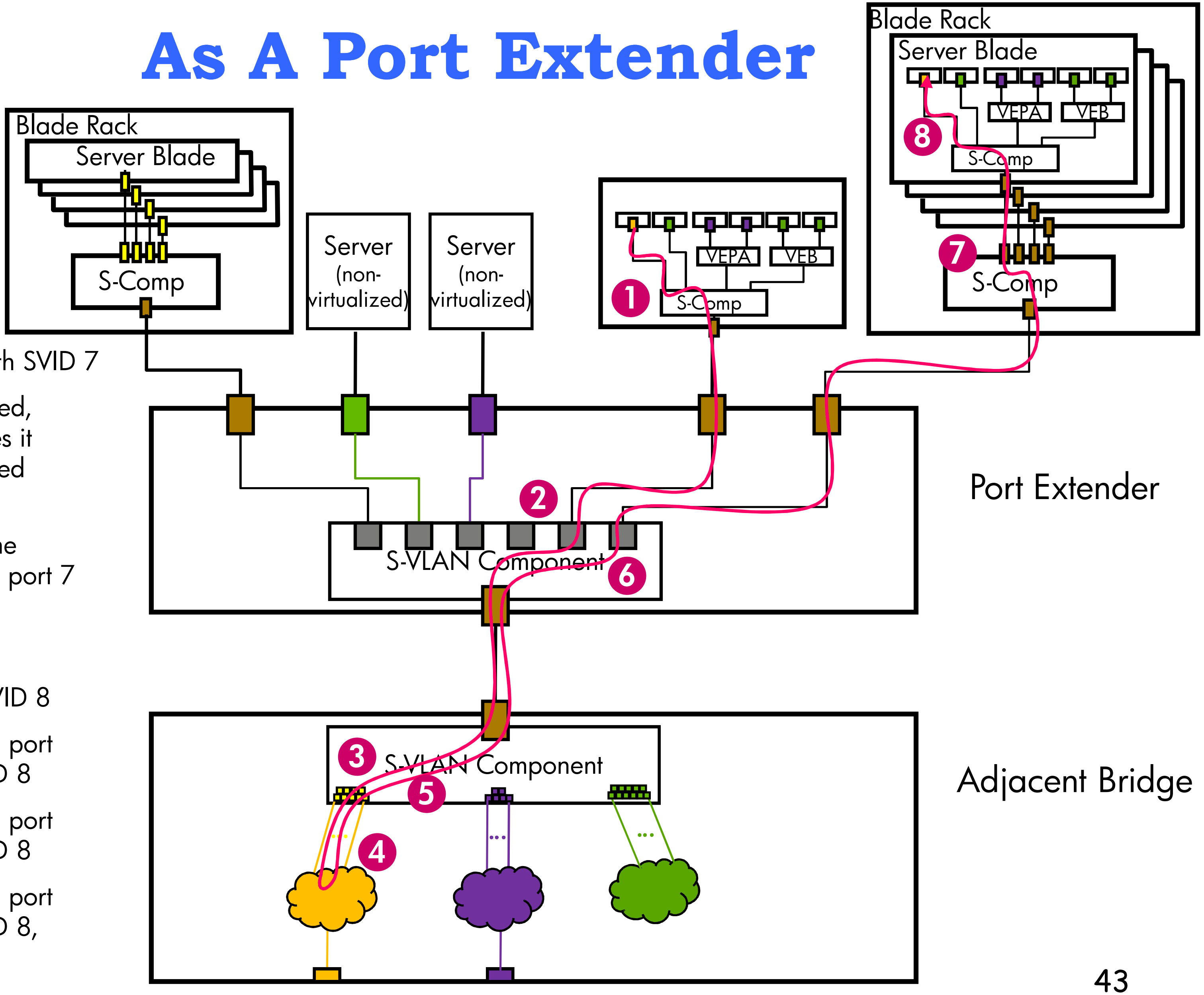
1. VEPA ingress frame from VM forwarded out VEPA uplink to S-Component
2. Station S-Component adds SVID (F)
3. Bridge S-Component removes SVID and forwards to port F
4. Frame is forward back to port D, S-Component adds SVID D
5. Station S-Component removes SVID D
6. S-Component forwards frame on Port D on Blue VLAN.

Port Extension and Remote Replication Services

MultiChannel Can Act As A Port Extender

Assume ports are numbered front to back, left to right:

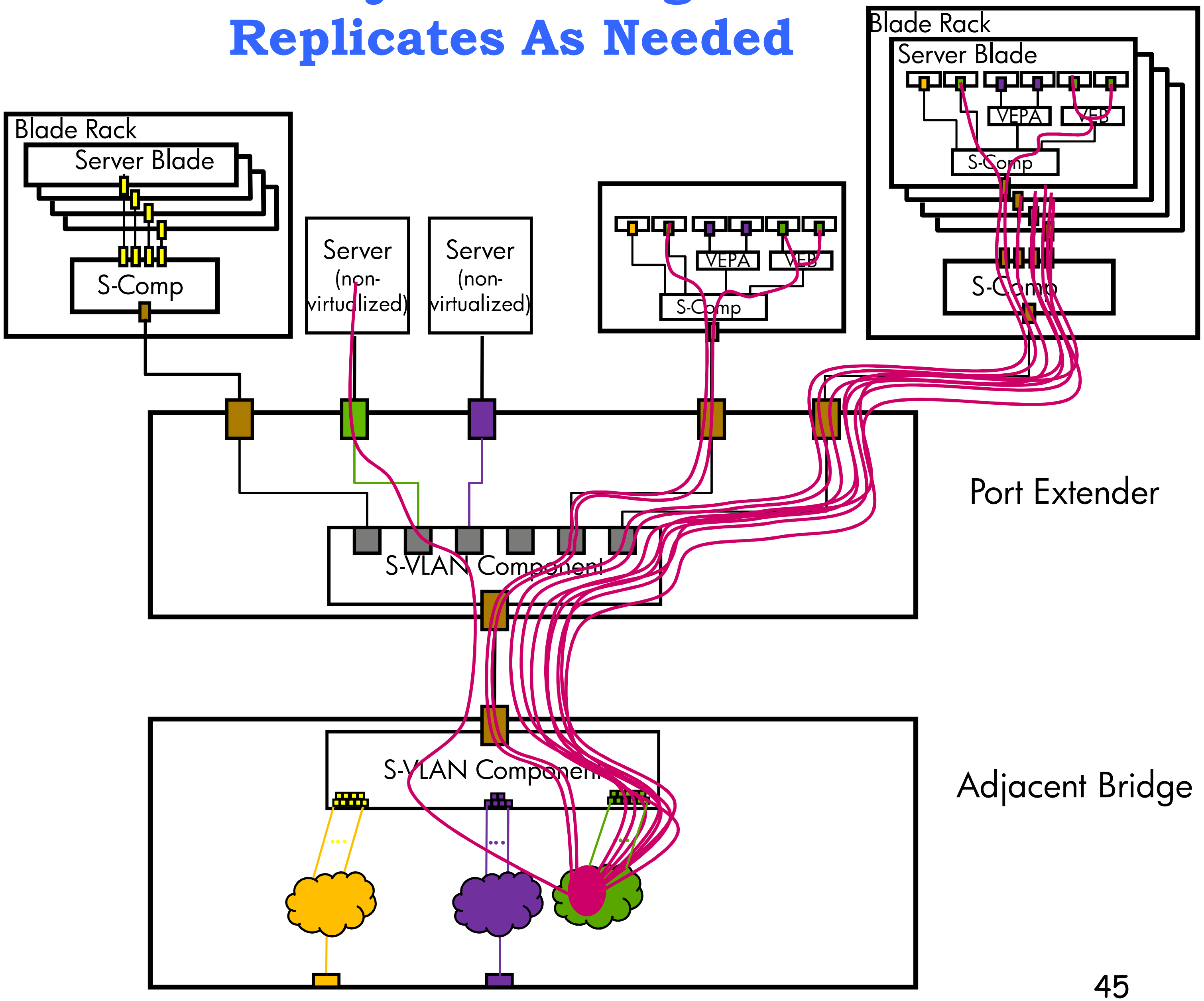
- Frame is tagged with SVID 7
- Since frame is tagged, S-Component passes it through (no cascaded tags)
- STag removed, frame forwarded to virtual port 7
- Frame forwarded to virtual port 8
- Stag added with SVID 8
- Frame forwarded to port that belongs to SVID 8
- Frame forwarded to port that belongs to SVID 8
- Frame forwarded to port that belongs to SVID 8, STag removed



MultiChannel Limitations

- Limited reach
  - Extensions needed to allow effective use of multichannel with cascaded port extenders.
  - Cascading is important to allow for flexibility in the design of network topologies.
- Inefficient bandwidth usage for multicast and flooded frames
  - Replication required for each channel carrying the same VLAN
  - Issue for multicast, broadcast, and flooded unicast frames

Adjacent Bridge Replicates As Needed

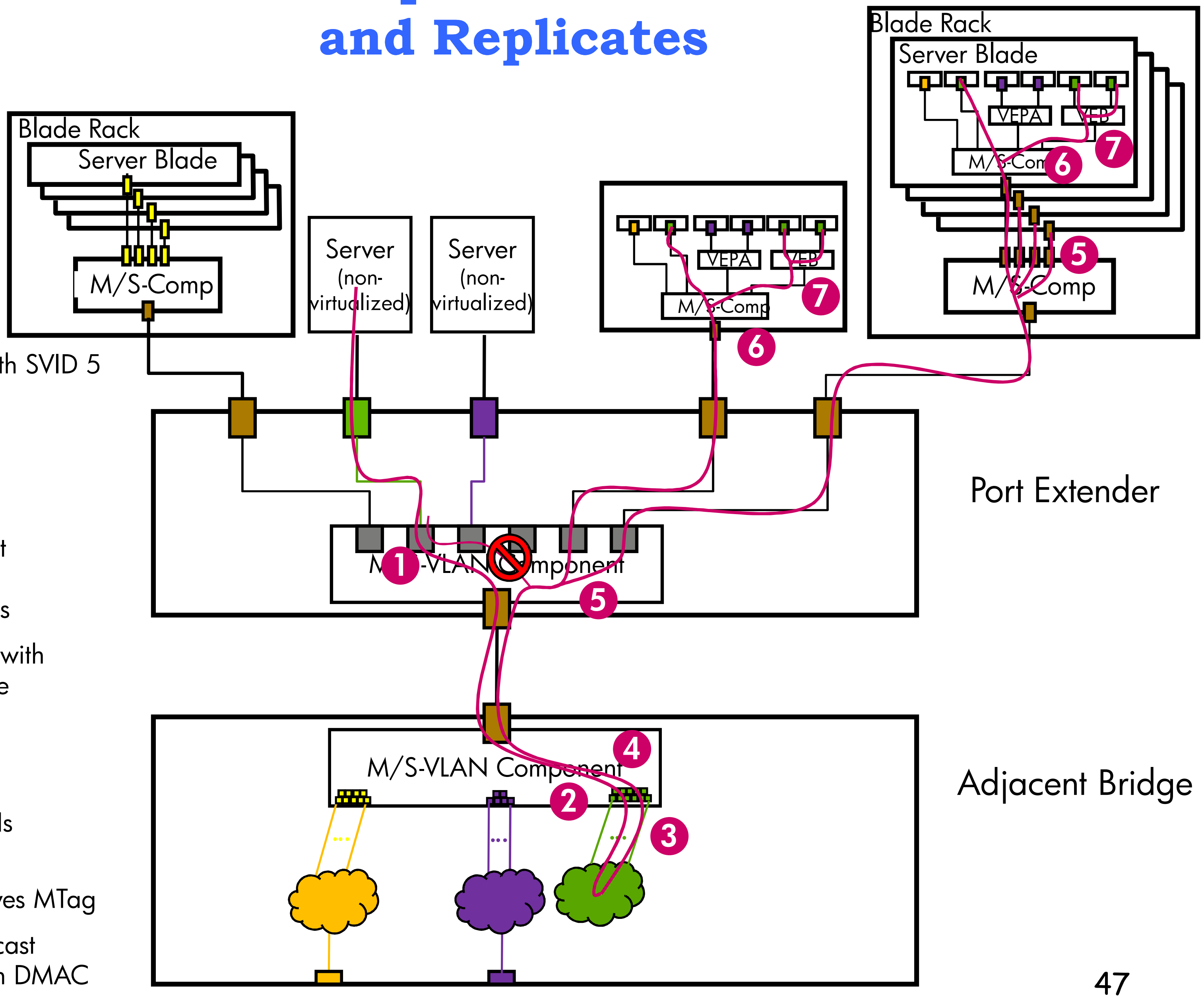


Adjacent Bridge Replication Challenges

- Replication adds excessive latency and consumes excessive bandwidth in environments using lots of multicast (e.g. financial markets)
- Reduces the ability of the adjacent bridge to apply sophisticated filtering rules (e.g. egress ACLs)
- Use of a multicast tag provides:
  - Ability for adjacent bridge to provide complete control of multicast frame delivery (e.g. egress ACL filtering)
  - Support for filtering of multicast frames destined to promiscuous ports
  - Simplified forwarding and filtering logic within the forwarding components

M-Component Collects and Replicates

- Frame is tagged with SVID 5
- Frame is relayed to virtual port 5, STag is removed
- Frame is relayed to multicast virtual port based on flood or group MAC address
- Frame is MTagged with group id and source SVID
- Frame is replicated based on group id, filtered from SVLANs which match SVID
- Last M-Comp removes MTag
- VEBs perform multicast as normal based on DMAC



Discovery

Possible Edge Discovery Exchanges

- Multichannel Configuration (per physical interface)
  - Whether multichannel & remote replication supported
  - Number of channels
  - Channel setup (Channel #, S-Tag)
- EVB Discovery (per channel)
  - Capabilities discovery (VEB, VEPA, PE, etc.)
  - Number of virtual station interfaces (VSI's)
  - Configuration of reflective relay (hairpin)
- Virtual Station Interface Discovery
  - Notify presence of Virtual Station Interfaces
  - Support arrival/departure of specific VSI's
  - Enable physical bridge port configuration based on VSI

Summary and Q & A

Pat Thaler, Broadcom

Summary

- Virtualization in Data Centers is increasing
  - To provide flexible, scalable, efficient, fault tolerant support for applications
- Some extensions to Bridge and End Station behaviors are needed to support virtualization
- Two PARs are proposed to provide this:
  - P802.1Qbg Edge Virtual Bridging
  - P802.1Qbh Bridge Port Extension

IEEE 802

802.1 Standards Roadmap

- Proposed – 802.1bg – Edge Virtual Bridging
  - Enables hairpin forwarding on a per-port basis when VEPA is directly attached
  - Defines a MultiChannel service to remote ports
  - Provides for discovery and coordinated configuration of station embedded components
    - Applies to both 802.1bg and 802.1bh
- Proposed – 802.1bh – Port Extension
  - Defines a tag to represent a group of remote ports for which a frame is to be replicated
  - Builds upon Remote Customer Service Interface and Edge Virtual Bridging

Basic VEPA

MultiChannel

Port Extension & Remote Replication

Next steps

- The proposed PARs are posted for review at:
  - http://ieee802.org/PARs.shtml
- Comments are due by 5 PM Tuesday
- Joint meeting of the Interworking and DCB task groups of IEEE 802.1 to discuss the PARs
  - Wednesday, 9 AM in Regency V
  - Any changes to PARs will be posted by 5 PM Wednesday

Questions?