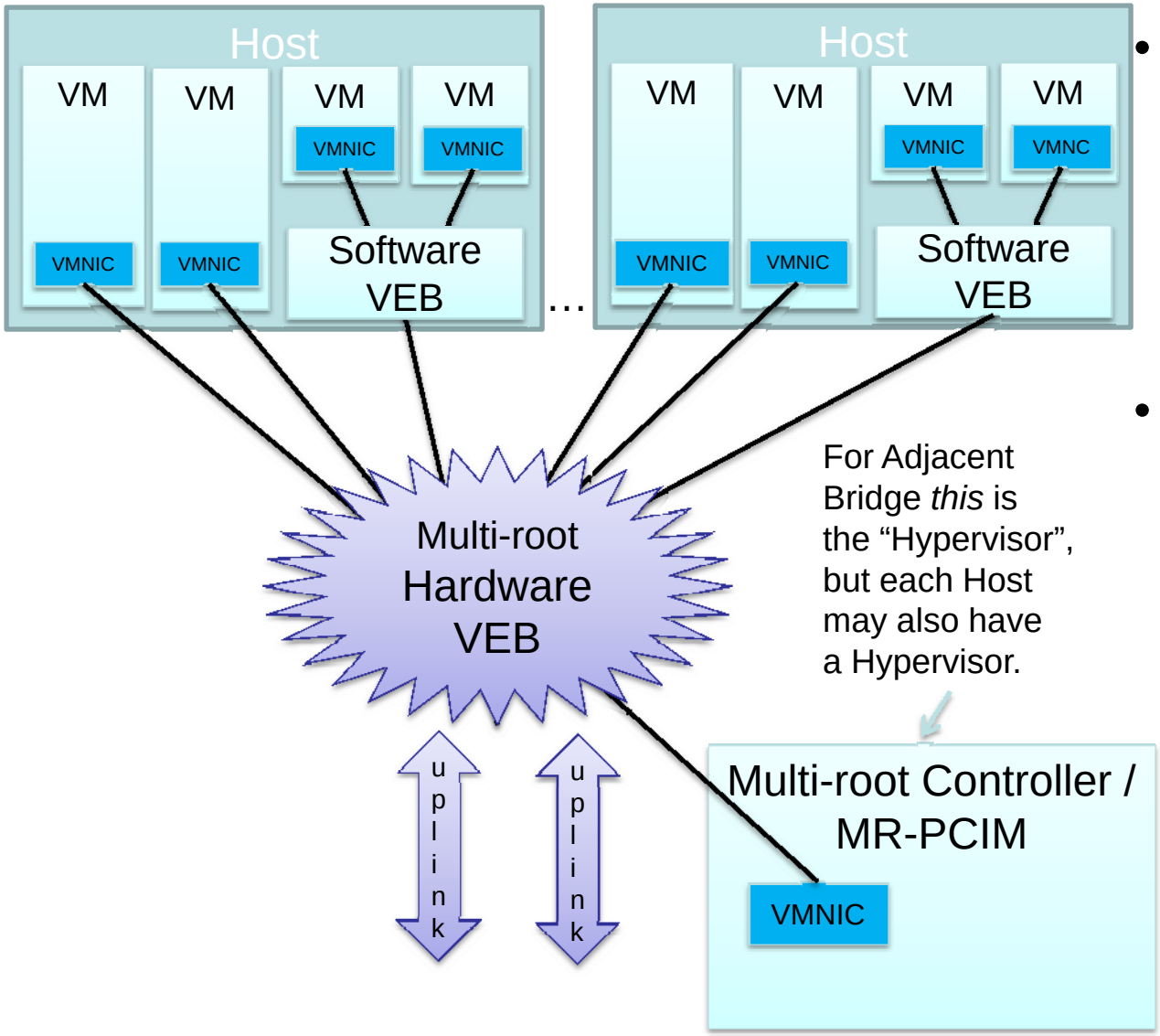
Multi-Root VEBs

Caitlin Bestler

Caitlin.bestler@aprius.com

21 Jan 2010

Multi-Root VEB

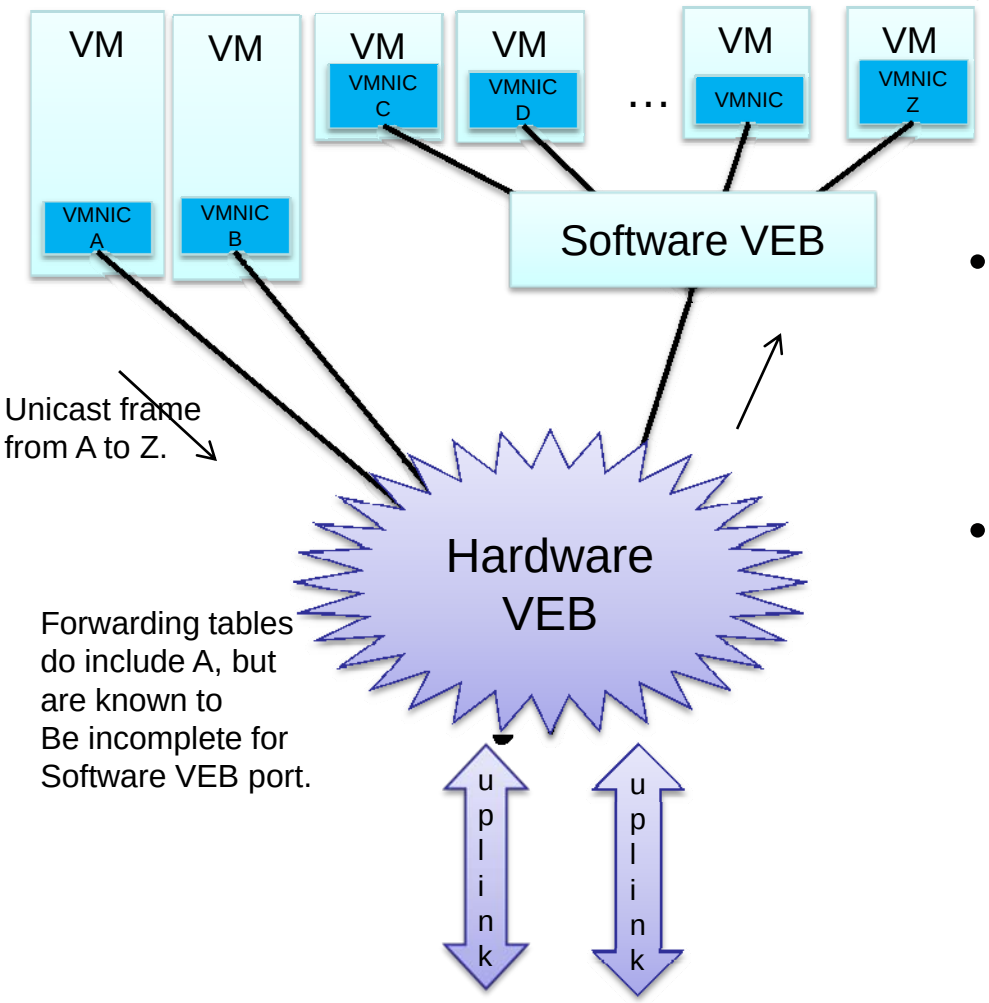


- As with a typical Hardware VEB, one or more uplinks are Bridged with local ports that are PCIe Functions.
- However there is only one multi-root master function, and one local master per host.

Multi-Root Issues

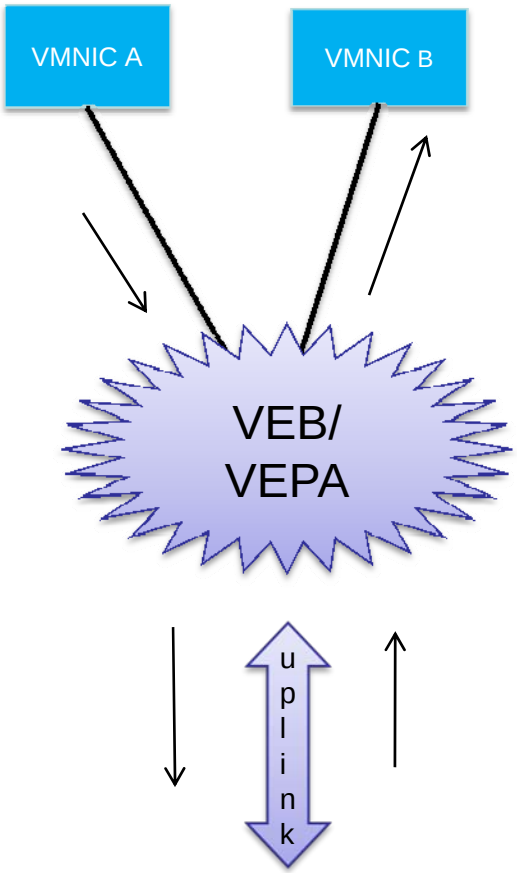
- Terminology: "Hypervisor" is not the entity that admits/controls the VSIs in a Multi-Root VEB.
  - A term that emphasizes the role (assigning VSIs) rather than the usual occupant would be better.
  - "Virtualization Intermediary" works, but is vague.
- VEB Stacking
  - A Multi-Root Hardware VEB has Single-Root VEBs in the role of End Stations.
  - The definition of a "VEB" should be compatible with this.
- Hairpin Reflection
  - Without Hairpin reflection, the multi-root Hardware VEB must know *all* VMNICs supported by *all* Software VEBs.
  - Avoid forcing an "all or nothing" choice on Hairpin Reflection.

Hairpin Reflection also an Issue for Single-Root VEBs



- Anytime the Hardware VEB's forwarding tables cannot hold all of the Software VEBs VMNICs Hairpin Reflection is useful.
- Multi-root environment just makes this more likely because more Software VEBs typically means more VMNICs.
- Without Hairpin reflection Where does Hardware VEB send frame from A to Z (not in its tables)?
  - Uplink Only?
    - Probably correct, but not always.
  - Uplink and Software VEB?
    - Software VEB may not have same capacity as the Uplink.
    - PAUSE from Software VEB can delay outbound traffic.

Unnecessary Reflection Costs More than 2X.



Direct: Frame from A to B via VEB

1. Transmit from A to VEB
2. Wait in VEB Output Queue for B
3. Transmit from VEB to B.

- Two Hops
- One Queue Wait

Reflection: Frame from A to B via VEPA

1. Transmit from A to VEPA
2. Wait in VEPA Output Queue to Adjacent Bridge.
3. Transmit from VEPA to Adjacent Bridge
4. Wait in Adjacent Bridge Output Queue for VEPA.
5. Transmit from Adjacent Bridge to VEPA.
6. Wait in VEPA Output Queue for B.
7. Transmit from VEPA to B.

- Four Hops
- Three Queue Waits
- Even worse if any of the extra queues delays trigger any form of Pause, CNM or drop.

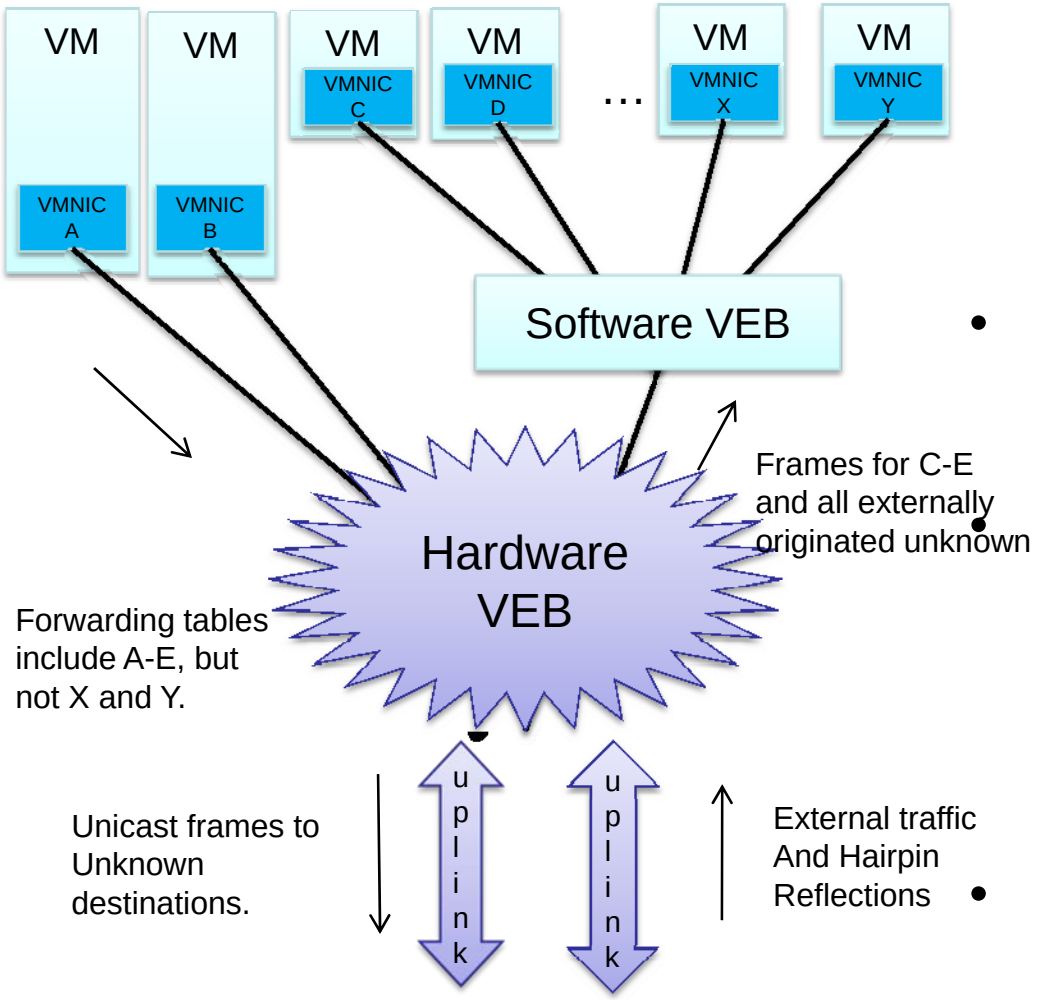
More Optimizations Possible

- Specific VEBs may have other potential optimizations.
  - Local links are not the same as general links.
- Example: the multi-root “switch” may match work requests, and only pull the frame once the destination is ready to receive it.
- Example: a software bridge can frequently “forward” frames by copy-on-write page mapping.

Established Connection Direct Forwarding

- Another example of partial use of Hairpin Reflection – Frames establishing TCP connections (SYN/SYN-ACK) are forwarded to the Adjacent Bridge.
  - This enables the Adjacent Bridge to let a Firewall (internal or external to it) approve and track the connection.
- Only after the connection is established are the frames directly forwarded.
- This is still falls under one simple rule:
  - “the VEB forwards some subset of the frames internally”.
- The benefits of the external firewall can be gained *without* requiring the entire connection flow through it.
  - An “All or Nothing” rule would block this functionality.

Avoiding All or Nothing Trap



- Because of limitations in Hardware VEB's forwarding table size there are VMNICs known to the Software VEB but not to the Hardware VEB.
- Traffic from A to B,C,D and E can still be directly forwarded.
  - **Major** performance improvement.
- Traffic from between VMNICs on the Software VEB can always be direct.
  - And any to Hardware VEB supported VMNIC (A,B) can be direct through Hardware VEB.
- Some frames from each VMNIC can be directly forwarded, even if all cannot be.

Impacts of VEB Direct Forwarding

- There are environments where Direct Forwarding will have major benefits.
- But just forwarding better is not enough, VEBs also have to **NOT** forward.
  - Accounting and Control are part of forwarding.
- Direct Forwarding must not result in:
  - Losing visibility of direct forwarded traffic.
  - Losing control of direct forwarded traffic.
  - Losing opportunity to learn end stations.
  - Duplicate deliveries when frame is direct delivered AND then hairpin reflected.

Maintaining Visibility

- Adjacent Bridge must know that it is connected to a VEB, i.e. something that can do direct forwarding.
  - And hence it knows to query it to obtain statistics on locally forwarded frames.
- Alternately, we could view this as a “VEPA” that has a “short circuit” capability.

Maintaining Control

- VEB has access to Port Profile.
- Does not do direct forwarding that contradicts the Port Profile.
- VEB must implement ACLs in port profile, but may broaden rules to make them simpler and/or reduce the number:
  - Broadened rules must forward via the Adjacent Bridge, which will enforce the correct ACLs.

Maintaining Learning

- Asymmetric Direct Forwarding can result in Adjacent Bridge only seeing traffic from X to Y, but not Y to X.
  - In which case it might forget where Y is.
- VEB must ensure that direct forwarding does not effectively hide any end station from the Adjacent Bridge.
  - One method: only do Direct Forwarding if the reverse flow would also be directly forwarded.

Avoiding Duplicate Delivery

- VEB must prevent duplicate delivery when a frame that was directly delivered is also hairpin reflected.
- This requires:
  - filtering on the Source MAC Address.
  - Refraining from doing Direct Delivery when the Source MAC Address is unknown.

VEB with Hairpin
or VEPA with Direct Forwarding?

- Two ways of describing the same behavior.
- VEB with hairpin matches evolution better.
- But either characterization would work.

No Impact on Adjacent Bridge

- When a VEB requests enabling Hairpin Reflection, simply honor the request.
  - VEB is responsible for properly handling all reflected frames.
    - Never re-deliver a frame to its source: unicast or multicast.
  - VEB must control / track all frames it directly delivers.
    - But there is no simple characterization of what frames will be direct delivered.
    - In example, direct delivery to X,Y was supported for C thru Y, but not for frames originating from A and B.
    - Simply assume that the VEB will directly deliver when convenient.
      - Query its statistics to find out how often it is actually doing so.
  - VEB must ensure that asymmetric direct forwarding does not hide an end station from the Adjacent Bridge.

Proposal

- A “VEB”, as opposed to a “VEPA” is *allowed* to do direct forwarding.
  - It MAY request Hairpin Reflection.
- Alternately, have a method for a VEPA to indicate that it does direct forwarding.
- Either must address the visibility/control issues cited earlier.