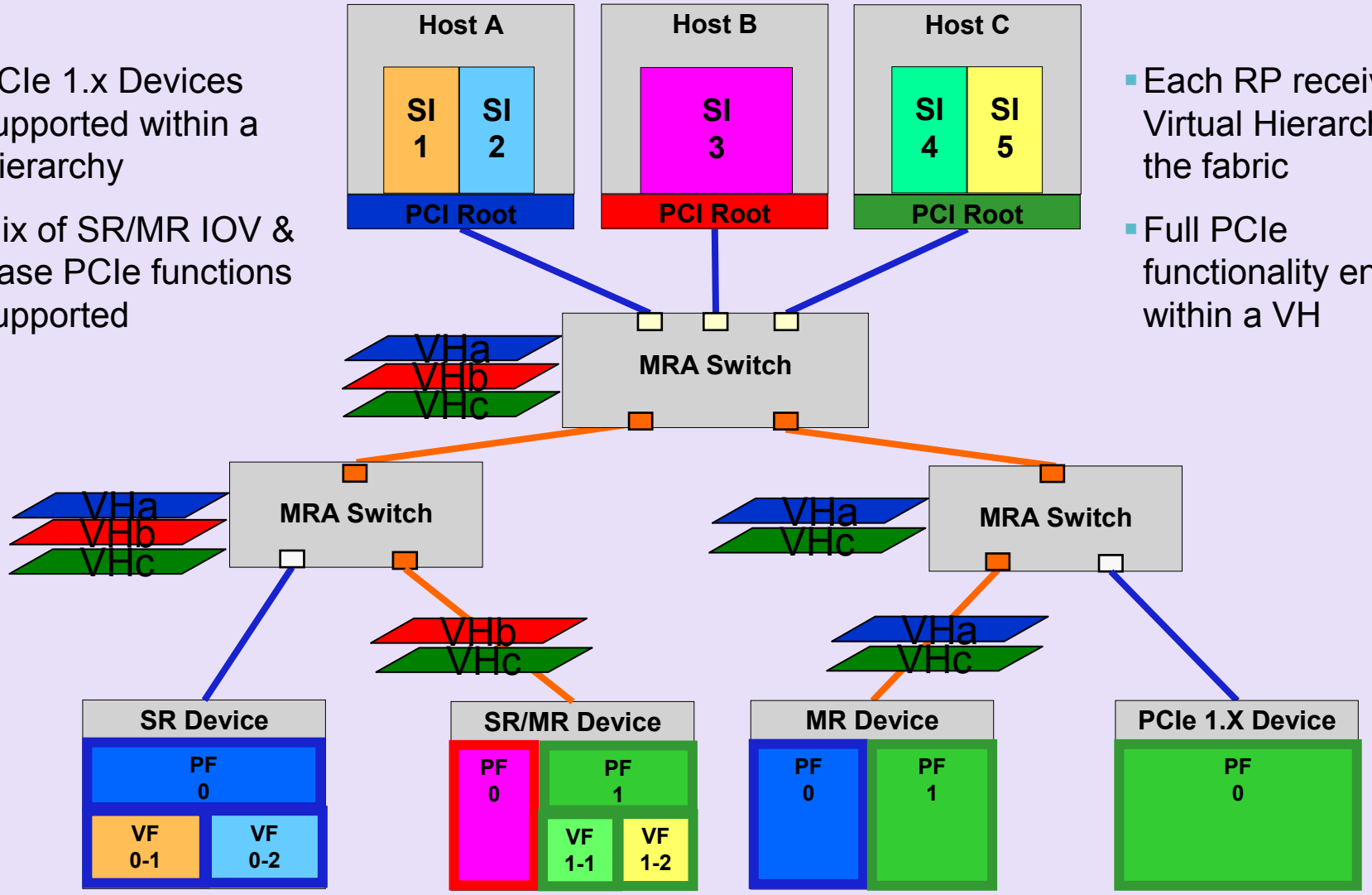

PCI
SIG®

Multi-Root Protocol

Steve Glaser (NextIO)

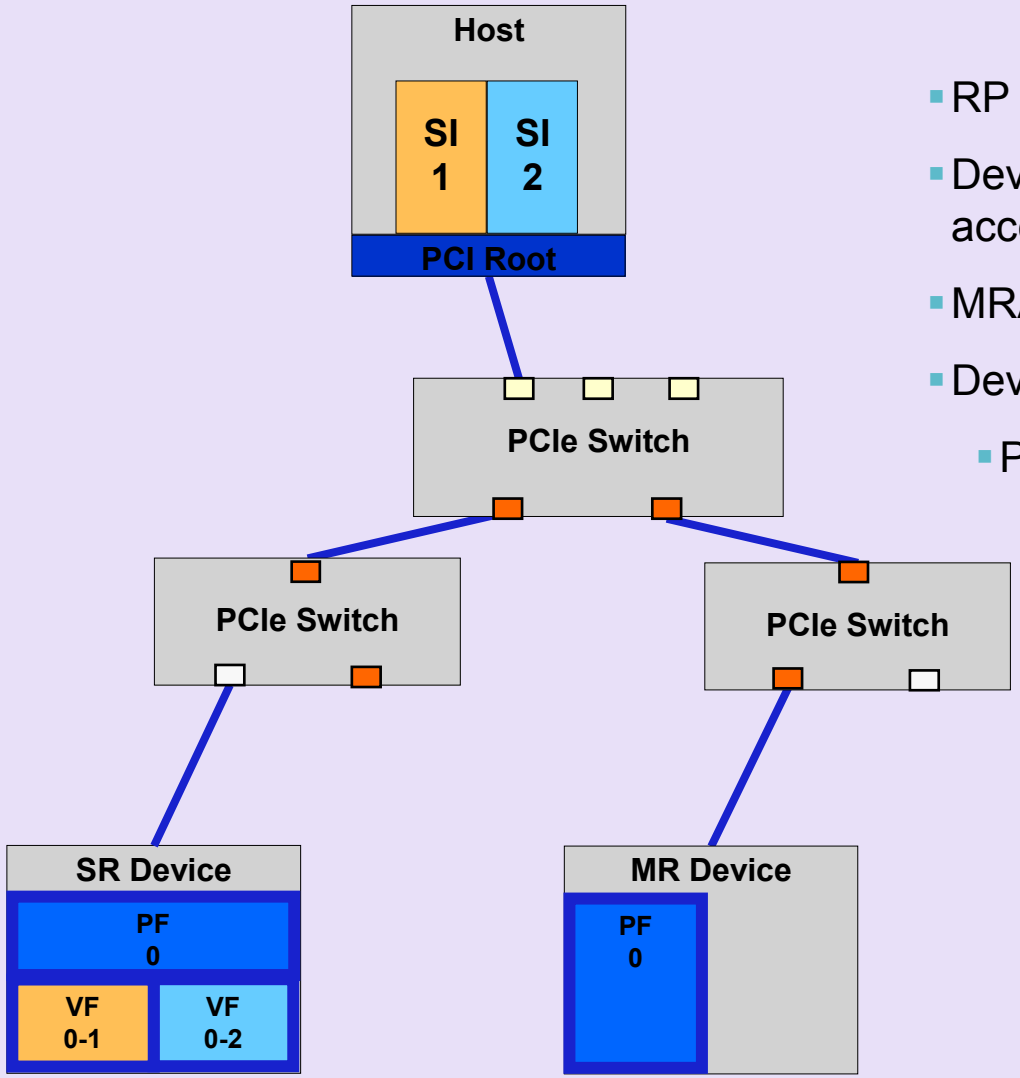
Virtual Hierarchy Overview

- PCIe 1.x Devices supported within a Hierarchy
- Mix of SR/MR IOV & Base PCIe functions supported



- Each RP receives a Virtual Hierarchy in the fabric
- Full PCIe functionality enabled within a VH

RP View within a VH



- RP "sees" only devices within its Hierarchy
- Devices in other hierarchies are not accessible
- MRA switches appear as PCIe 1.X switches
- Devices appear as SR equivalent devices
  - PCIe 1.X devices with IOV capabilties

Virtual Hierarchy Overview

- Independent Virtual PCIe Hierarchy per RP
  - ✓ Virtual PCIe Hierarchy extends from RP to Devices
  - ✓ All valid PCIe operations exist within Virtual Hierarchy
    - – Operations may only have meaning in virtual context – No physical action
  - ✓ Hierarchies are separate, independently controlled resources
    - – Each RP only has access to the resources of its VH
- Protocol Changes Required for implementation
  - ✓ Tag each PCIe TLP on link with VH identifier
    - – Added and removed at DLL
  - ✓ Per VH reset within DLL
- Component impact
  - ✓ RP – No (uses PCIe base)
  - ✓ Switch – Yes (implements virtual PCIe switch per VH)
  - ✓ Device – Yes (implements Type 0 CFG per VH)
  - ✓ RP Software – No (utilizes PCIe SW programming model)

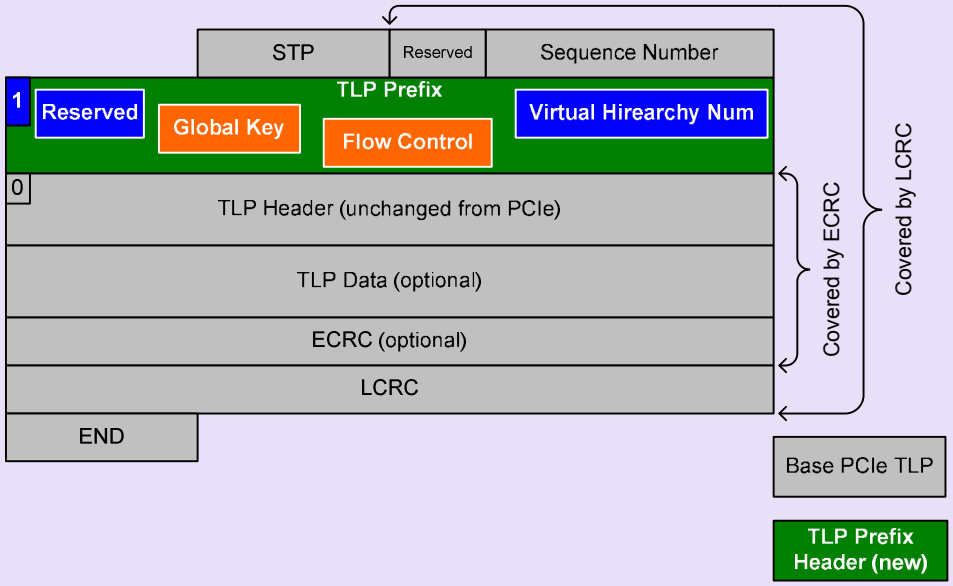
Virtual Hierarchy Protocol

- TLP Tag
  - ✓ Inserted/removed at DLL
    - – No TL impact
  - ✓ Targeted for support of at least 256 VH (exact size TBD)
- RESET DLLP
  - ✓ Per VH RESET DLLP
    - – Guaranteed progress under all fabric conditions
    - – TLP messages can stall due to FC congestion
  - ✓ Propagates using RESET logical rules
    - – Provides a mirror of PCIe RESET within a VH
- Virtual Link Flow Control
  - ✓ Provides an expansion to PCIe Base FC
  - ✓ Solves Congestion Management issues with large VH counts

Virtual Hierarchy Operation

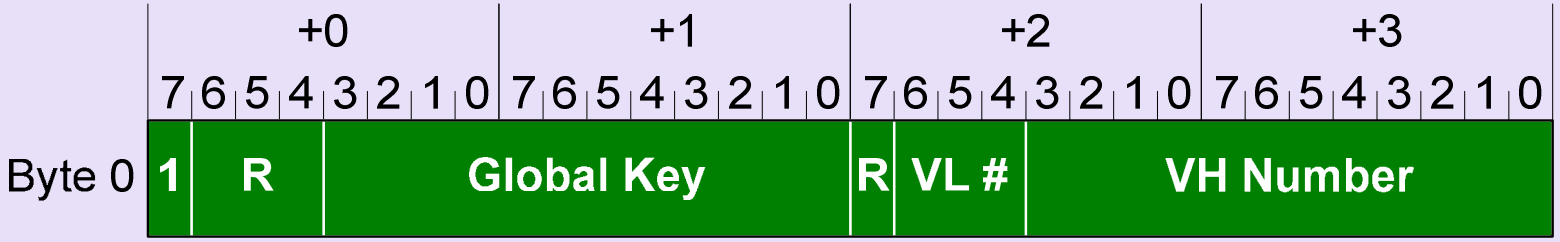
- Initialization & Enumeration
  - ✓ MR-PCIM discovers MR and Base devices
  - ✓ MR-PCIM creates assignment of devices to RP
  - ✓ MR-PCIM programs MR Switch and Device tables with MR assignments
  - ✓ PCIe / SR-PCIM SW enumerates within its Virtual Hierarchy
- Traffic Flow
  - ✓ Base RP & Base/SR Device utilize PCIe Base protocol
  - ✓ MR Device inserts VH tag at DLL for appropriate PF on Base TLP
  - ✓ Switch utilizes VH tag to index correct Type1 CFG headers
  - ✓ PCIe Base routing rules utilized within a Virtual Hierarchy
- RESET
  - ✓ Base RP asserts RESET as TS1 or Fundamental RESET
  - ✓ Switch propagates on MR link as RESET DLLP within a VH
  - ✓ Switch propagates on Base link as TS1
  - ✓ MR Switch or Device receiving RESET DLLP flushes according to PCIe rules
  - ✓ RESET state is tracked per VH by each link partner

TLP Prefix Header



- Header for all TLPs on MR link
  - ✓ Not included on PCIe 1.X links
- Header included on all TLPs
  - ✓ Stable during retransmissions
- Located between Sequence # and TLP Hdr
- ACK / Sequence # concept remains per link
  - ✓ Not affected by Virtual Hierarchy changes
  - ✓ Like VC today
- Header covered by LCRC
  - ✓ Header **not** covered by ECRC

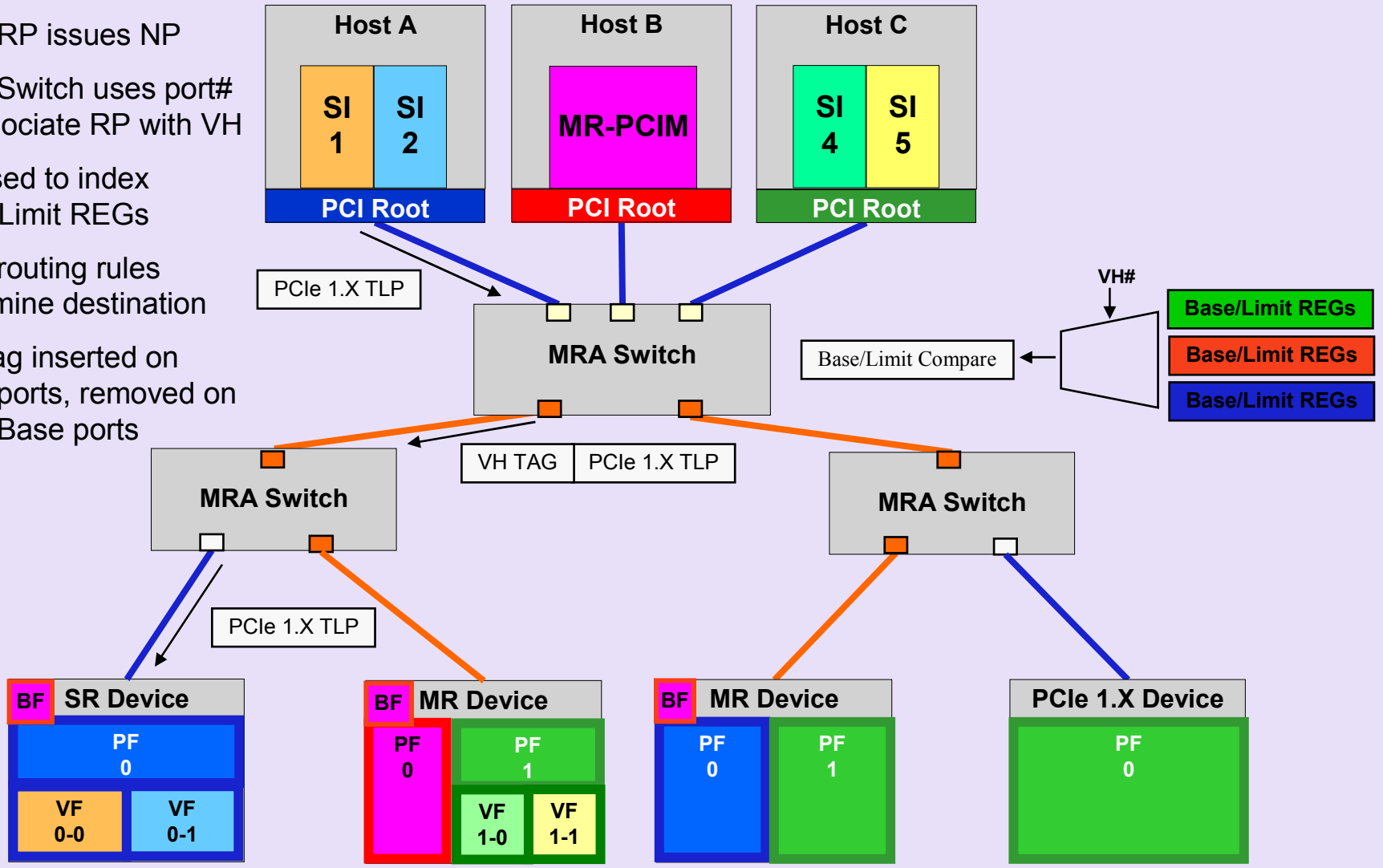
TLP Prefix Header Format (tentative)



- Global Key provides mechanism for protection across VH
  - ✓ Unique number for each VH
  - ✓ Inserted at MR Ingress, Checked at MR Egress
- VL # provides FC information in TLP
  - ✓ Ensures correct TLP association with FC credit bucket
- VH Number indicates the VH of the TLP
  - ✓ Changes link by link
  - ✓ Direct index into tables at Receiver

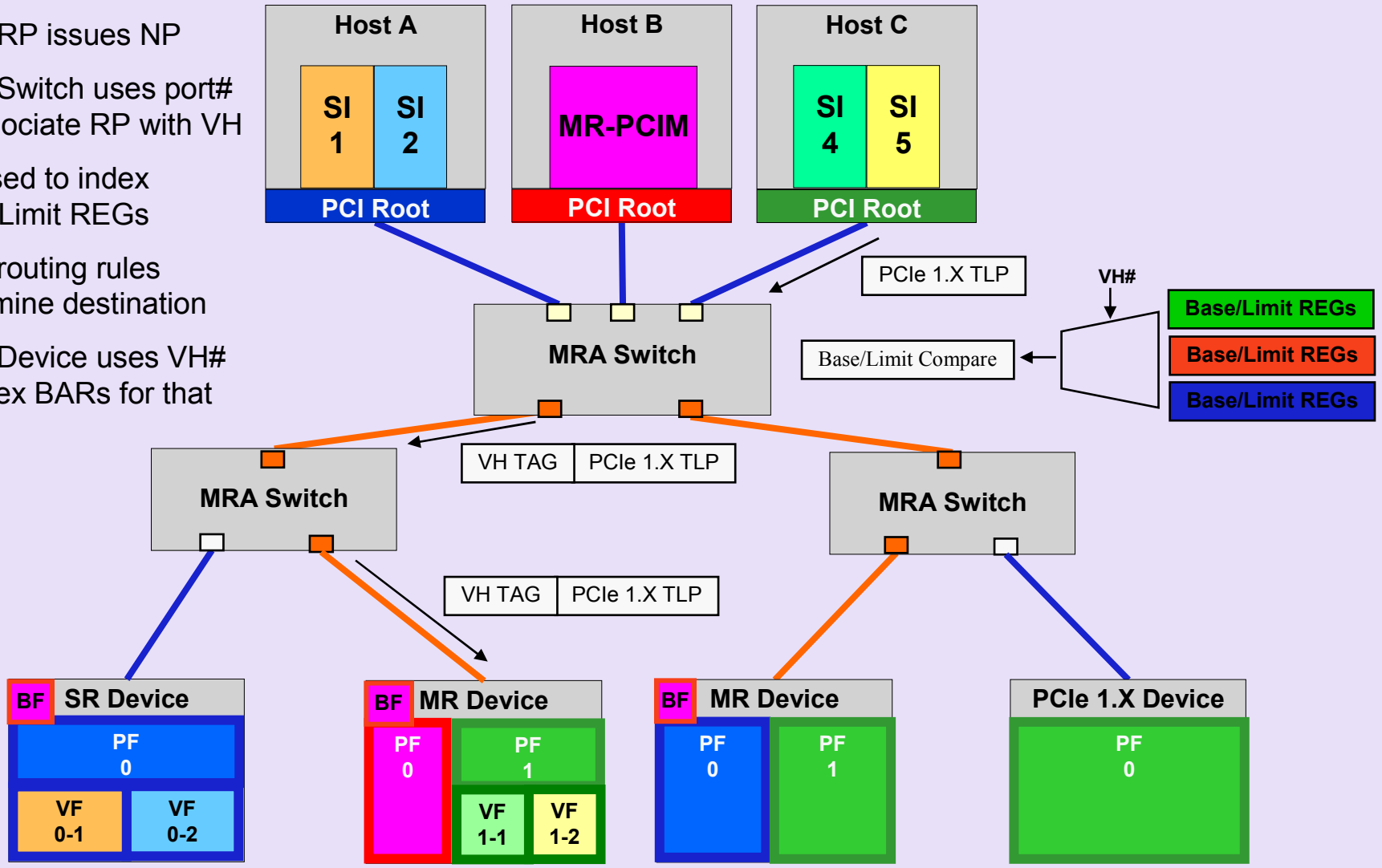
Non-Posted Request from PCIe Base to PCIe Base

1. PCIe RP issues NP
2. MRA Switch uses port# to associate RP with VH
3. VH used to index Base/Limit REGs
4. PCIe routing rules determine destination
5. VH Tag inserted on MRA ports, removed on PCIe Base ports



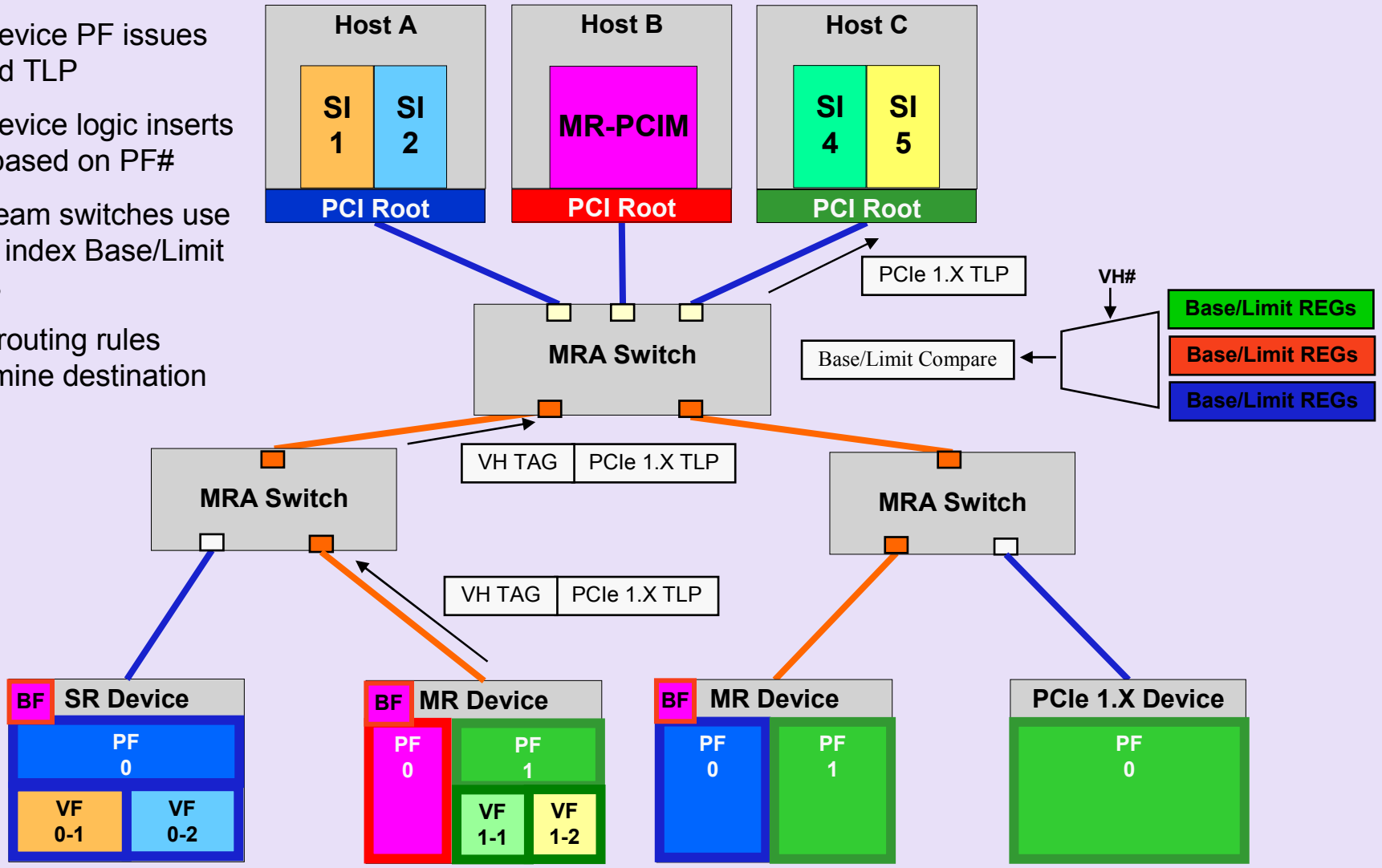
Non-Posted Request from PCIe Base to MRA

1. PCIe RP issues NP
2. MRA Switch uses port# to associate RP with VH
3. VH used to index Base/Limit REGs
4. PCIe routing rules determine destination
5. MRA Device uses VH# to index BARs for that VH

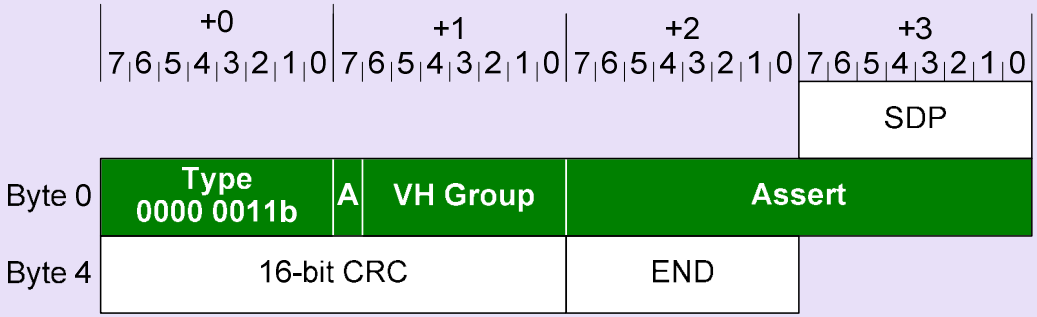


Posted from MRA to PCIe Base

1. MR Device PF issues Posted TLP
2. MR Device logic inserts VH# based on PF#
3. Upstream switches use VH to index Base/Limit REGs
4. PCIe routing rules determine destination



RESET DLLP (tentative)



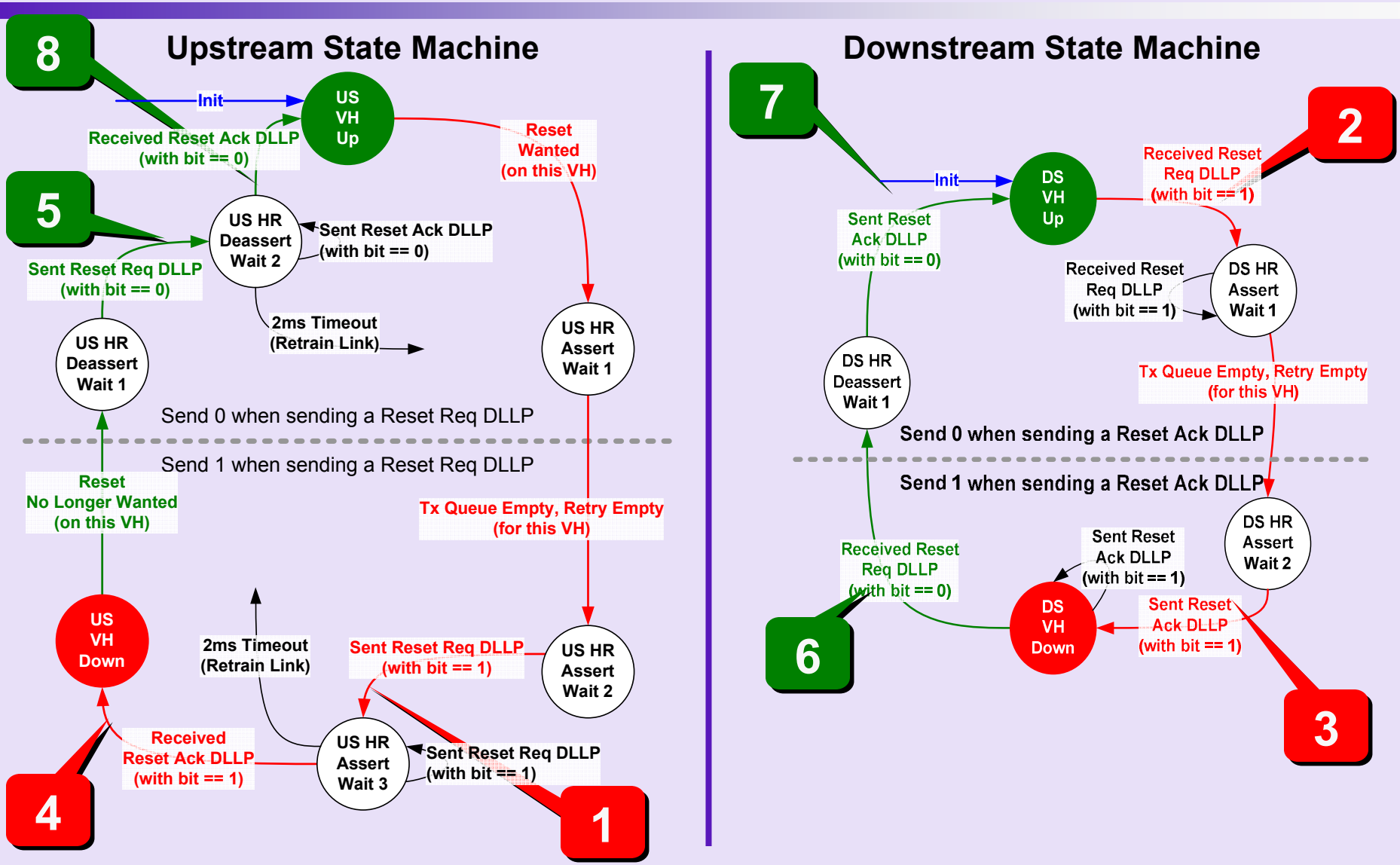
A: 0 = Request
1 = Acknowledge

VH Group: upper bits of VH number representing a group of 16 VHs

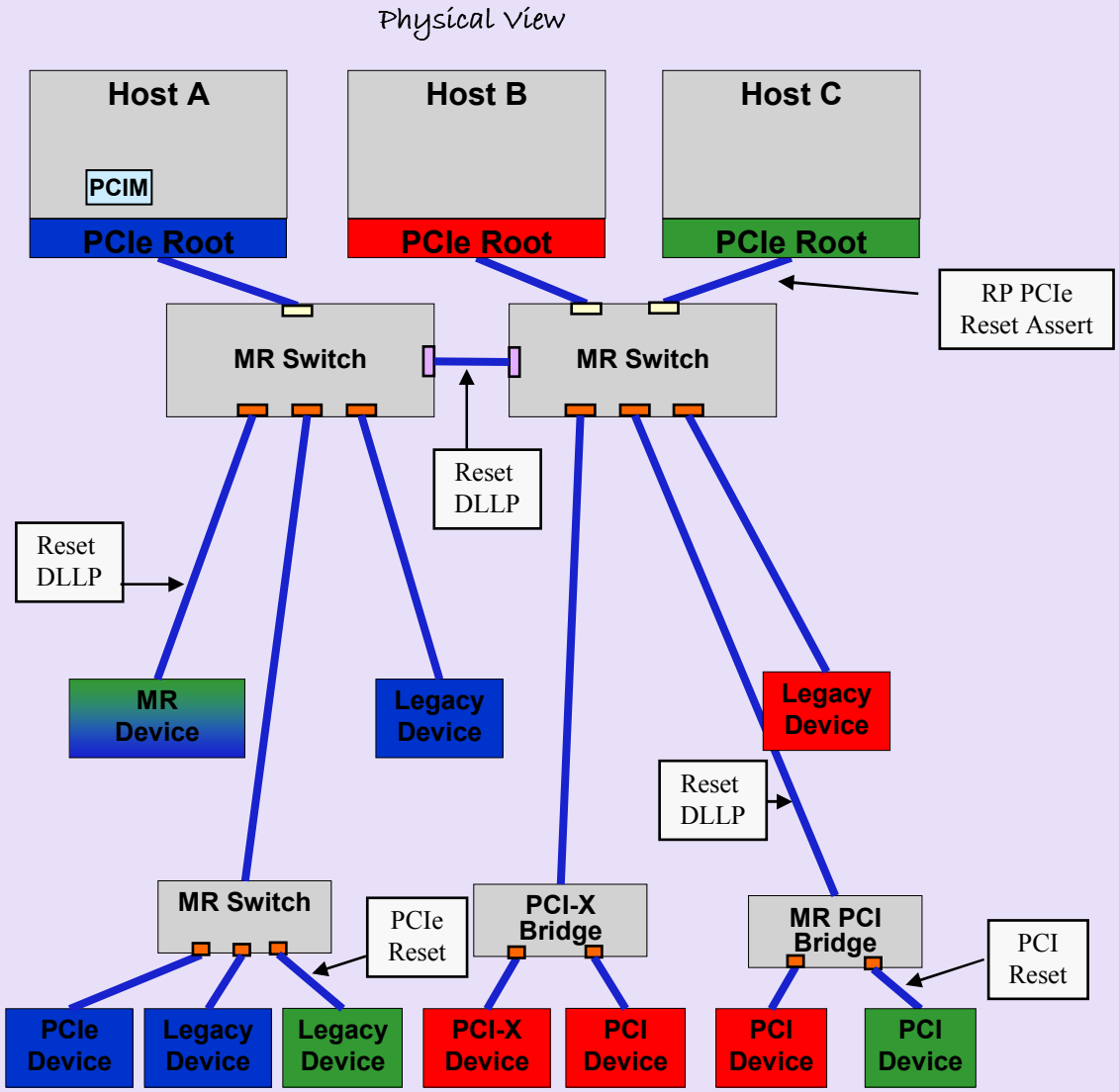
Assert: Bit field representing the VHs within group of 16.
1 = Assert Hot Reset
0 = Deassert Hot Reset

- Provides RESET assert/clear for up to 16 VH in single DLLP
- Handshake for complete reliability of RESET propagation
- Guarantees buffer flush of VH within intermediate switches
- Utilizes currently RSVD DLLP
- RESET state machine utilized to track VH state by each link partner

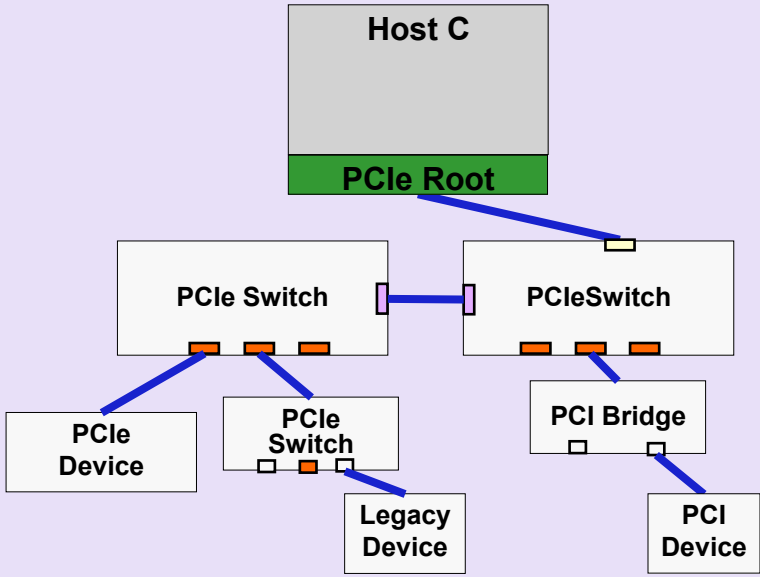
Reset State Machines



RESET Propagation



- RP Reset completely resets Virtual Hierarchy
- Equivalent to PCIe 1.X RESET functionality

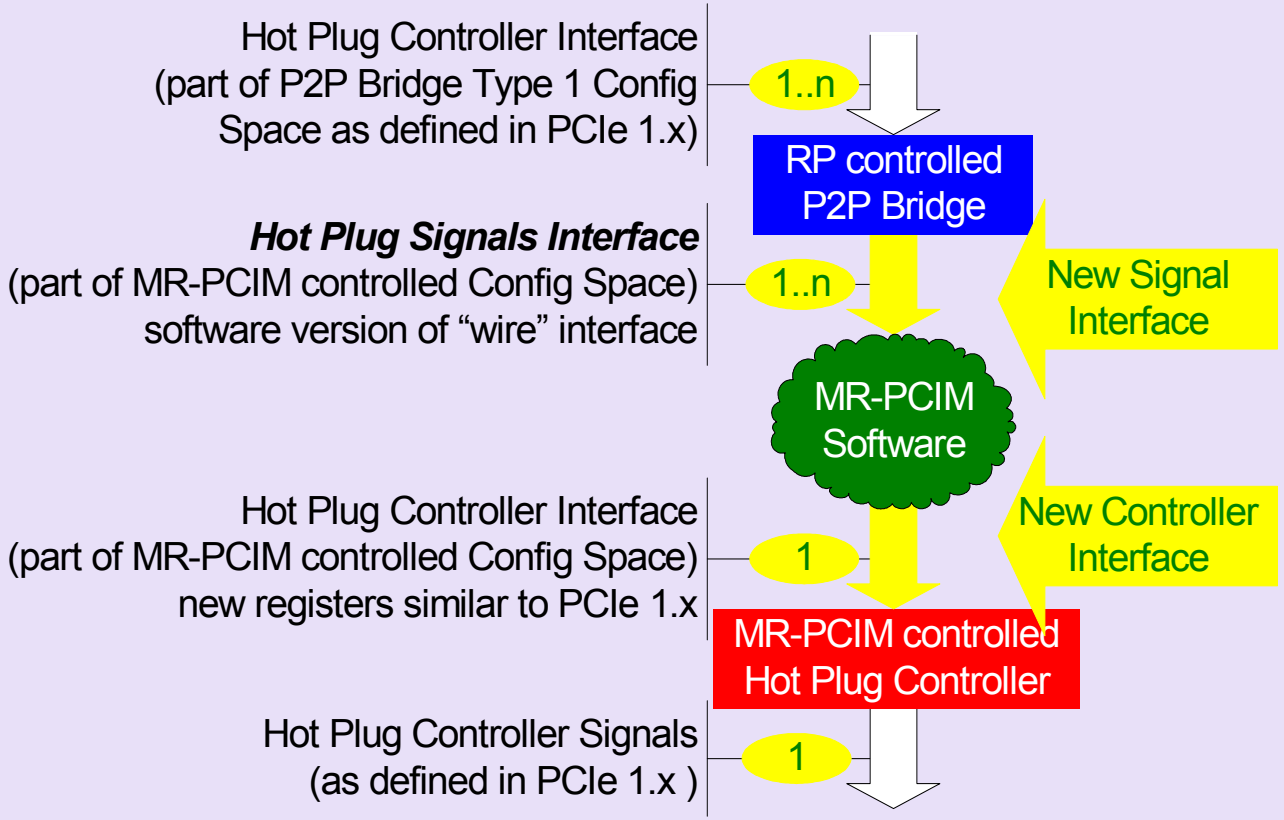


Hot Plug

Hot Plug in MR

- MRA Switches have two Hot Plug Controllers (HPC)
  - ✓ Physical controller (1 per link)
    - – Controls events on the link
    - – Owned by MR-PCIM
  - ✓ Virtual controller (1 per VH per link)
    - – Controls events within a VH
    - – Owned by SR-PCIM, coordinated with MR-PCIM
      - • New registers added for MR-PCIM access point
- Hot Plug in MR consists of two events
  - ✓ Physical events coordinated by MR-PCIM and physical device
  - ✓ Virtual events coordinated by MR-PCIM, SR-PCIM, and virtual HPC
- HPC SW interface same a PCIe Base
  - ✓ Utilizes PCIe Hot Plug specification within switches

MR Hot Plug Interfaces



Power Management

Power Management Philosophy

- MR IOV Power Management will leverage PCIe model
  - ✓ASPM and PCI-PM expanded for MR IOV
  - ✓Each VH controls its own D-state of Device
- Same Link states and transitions as PCIe
  - ✓Handshake for L1 and L2/3 L-state transitions unchanged
- Same wake-up model as PCIe
  - ✓Beacon/WAKE# supported if a fabric Component needs main power restored
  - ✓Functions issue PM_PME TLP to request PM state change
- MR-PCIM controls main power to a Device
  - ✓Components can still be AUX-powered for highest power savings

PCI-PM and ASPM

- Each VH has its own “virtual” D-state in the Device
  - ✓ Allow a VH to maintain same software model of controlling D-states
    - – Function still clears state if written to D3
  - ✓ For link state transitions, utilize same scoreboard technique in PCIe spec for multi-function
    - – All functions in Device must be written to D3 before link transitions to L1
  - ✓ MR-PCIM also controls a virtual D-state of endpoint
- Multi-function Power Management rules in PCIe expanded to encompass VHs
  - ✓ All functions *in all VHs* must be in D3 state before L1 link state transition can be initiated
- Existing L-states preserved
  - ✓ PHY layer unchanged by MR IOV Power Management

Component Power

- Power controlled by MR-PCIM
  - ✓ During fabric initialization, MR-PCIM adds power to appropriate components
  - ✓ MR-PCIM determines if all functions are powered off so that main power can be removed
    - – If a VH is powered off or in reset state, its D-state is not considered in the power management scoreboard
  - ✓ MR-PCIM uses Power Controller Control bit in physical hot plug controller to turn on and off power to Device

L1 wakeup example

- All Functions in all VHs are in D3
- A Function wants to send PM_PME
  - ✓ Downstream Link returns to L0 state
  - ✓ No need to send Beacon / WAKE# since main power still active
- MR switch trains upstream link back to L0 and forwards PM_PME to appropriate RP
- RP services PM_PME and writes appropriate components back to D0
- Function is now active
  - ✓ Very similar to PCIe L1 wakeup

L2 wakeup example

- Main power is off for Device
  - ✓ AUX power present to allow for Beacon / WAKE# initiation
- Device issues Beacon / WAKE# to downstream MR switch port
  - ✓ MR switch propagates Beacon / WAKE# to management port
  - ✓ MR switch will begin link training process on all upstream ports associated with that Device
- MR-PCIM adds main power back to Device
- One or more Functions issue PM_PME TLP(s)
- MR switch forwards PM_PME to appropriate RP
- RP services PM_PME and writes appropriate components back to D0
- Function is now active
  - ✓ Similar to PCIe, except MR-PCIM restores main power instead of any RP's

Questions


PCI
SIG®