Multicast Over PCI Express®

Derek Percival
Application Engineer, PLX Technology

Agenda

- What is Multicast?
- Overview of PCIe® Multicast (MC) ECN
- Multicast Traffic Patterns
  - ✓ From Host to multiple Downstream (DS) ports
  - ✓ Upstream to redundant Hosts
  - ✓ From Peer to Multiple Downstream ports
  - ✓ From Peer to Peer plus Host
- Multicast Register Settings and routing example
- Application Examples
  - ✓ Acceleration
    - Dual-headed graphics
  - ✓ Redundancy
    - Storage mirroring
  - ✓ Protocol Emulation
    - Tunneling Ethernet over PCIe
- Timeframe

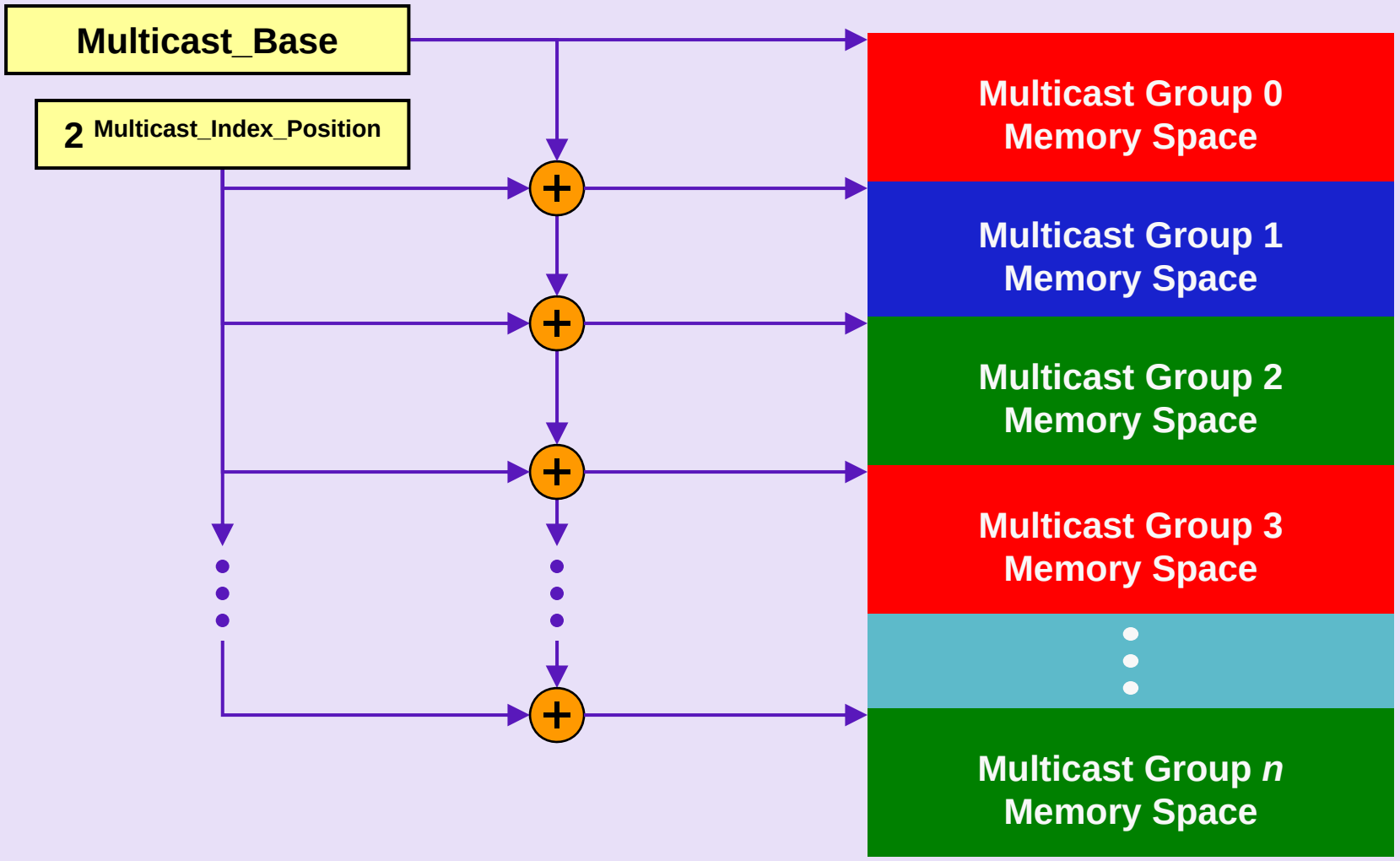
What is Multicast

- ***Wiki***
  - ✓ **Multicast** is the delivery of information to a group of destinations simultaneously using the most efficient strategy to deliver the messages over each link of the network only once, creating copies only when the links to the destinations split.
- History
  - ✓ IP Multicast
  - ✓ RapidIO Multicast Extensions 8/2004
  - ✓ ASI included Multicast
  - ✓ Infiniband supports unreliable Multicast
- Benefits:
  - ✓ Unblock bottleneck of PCIe link at source and between switches
  - ✓ Unblock bottleneck of memory bandwidth (BW) at source
  - ✓ Make PCIe more suitable for backplane applications

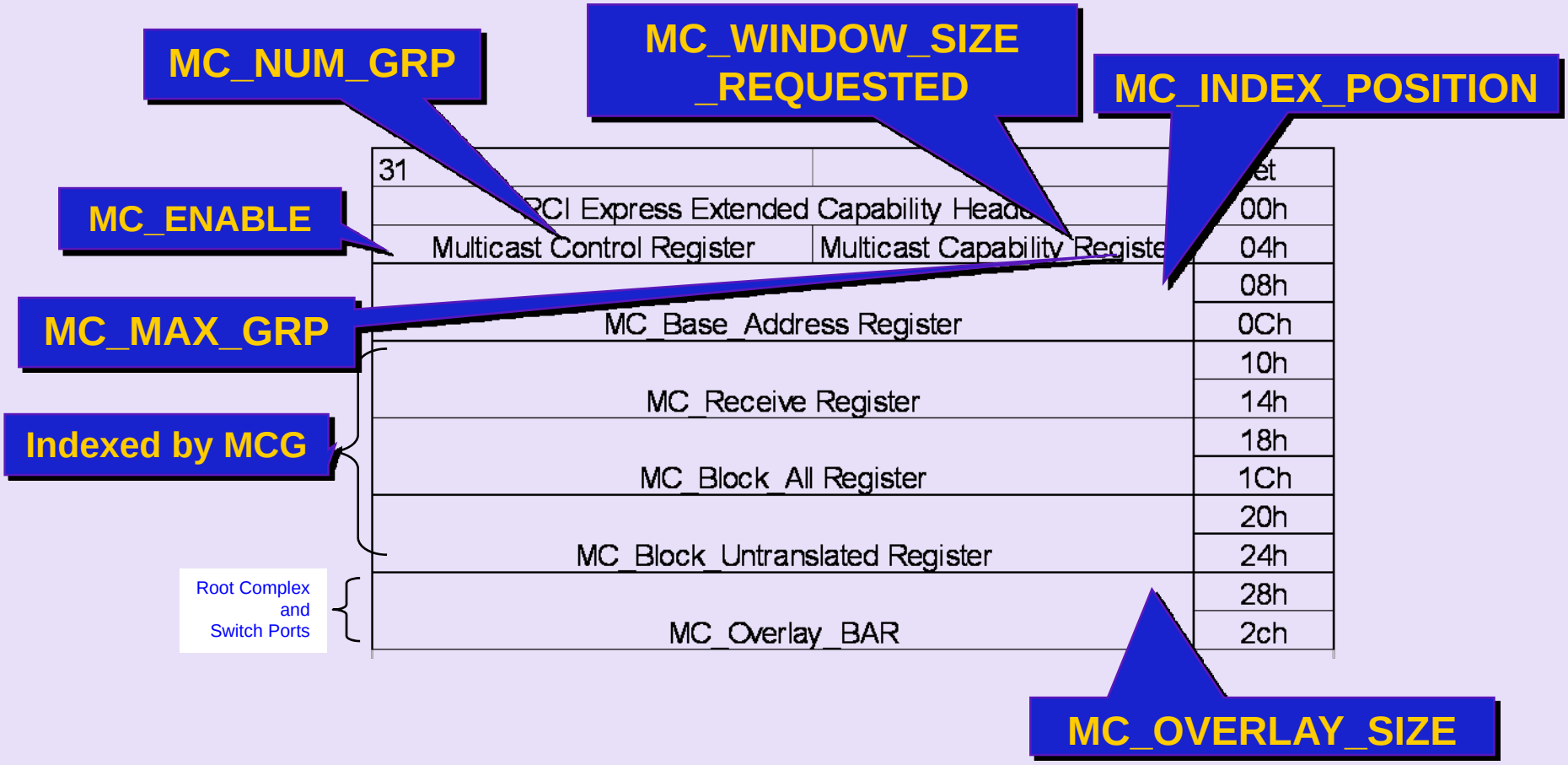
PCIe Multicast

- Address based
  - ✓ Multicast BAR creates an Multicast space
  - ✓ Posted packets that hit in the Multicast BAR are multicast
- "Unreliable" – e.g. no end to end ACKs
  - ✓ Nevertheless, highly reliable because PCIe has low error rate and hop by hop error free transmission
- Supports legacy
  - ✓ Any source can send posted packet in Multicast space
  - ✓ Legacy devices can be multicast targets
- Multicast ECN
  - ✓ Specifies how PCIe components – switches, Root Complexes (RCs), and Endpoints (EPs) - implement Multicast
  - ✓ Improves Multicast flexibility and protection for EPs that participate in Multicast but also allows use of legacy EPs

Multicast Memory Space



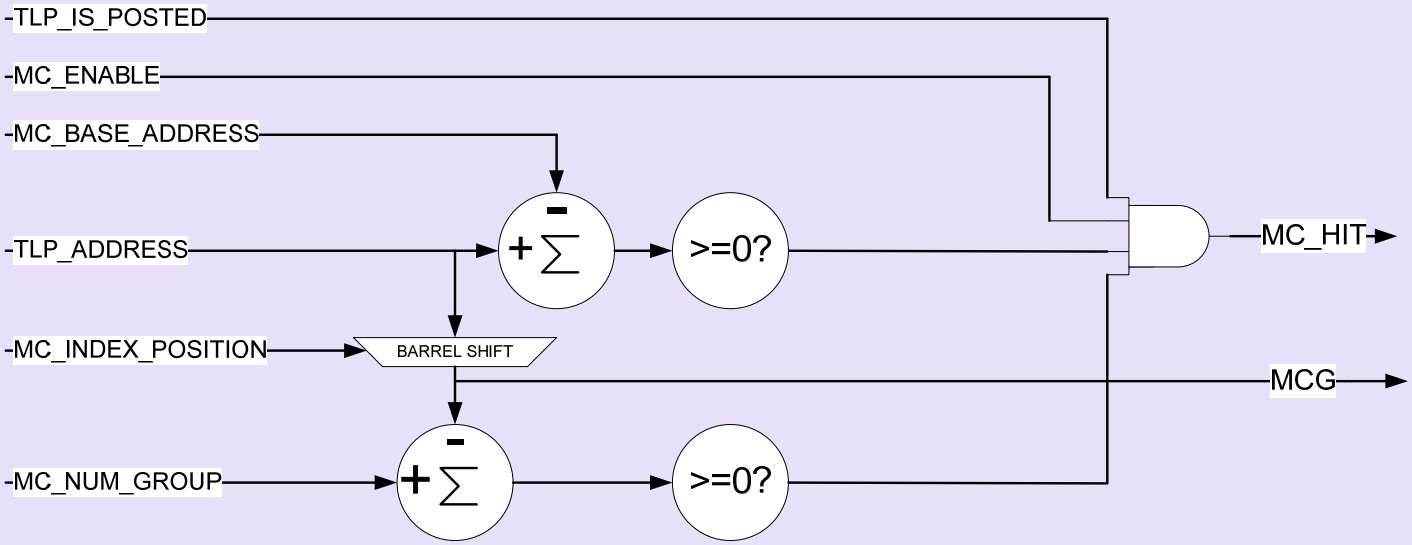
MC Capability Structure



A capability structure in each function/port

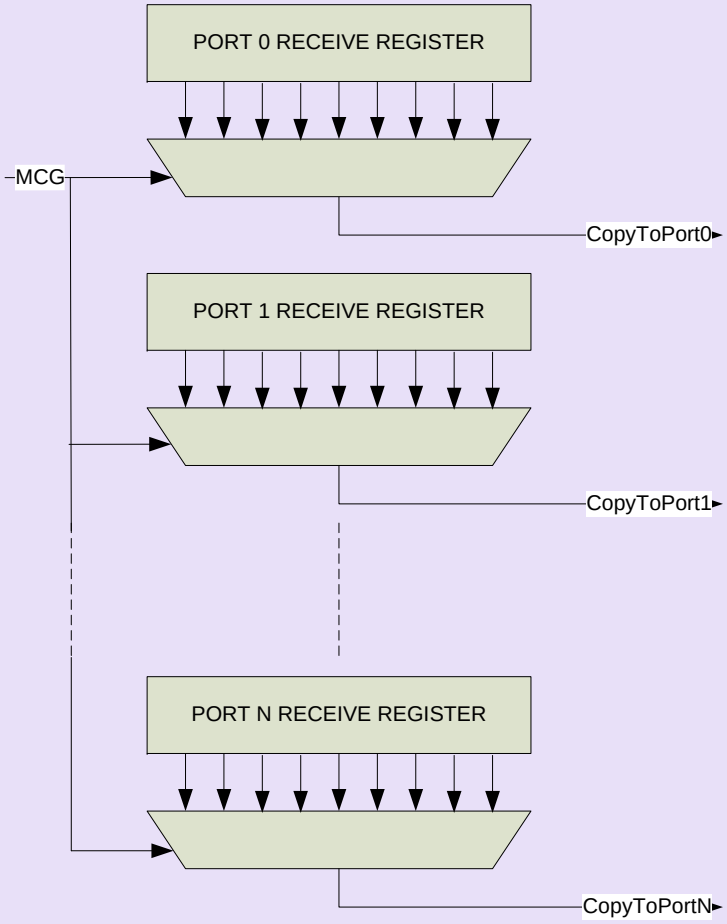
Implementation efficiencies in Multi-function Device since fields must be configured identically

Multicast Hit Processing



- A TLP has a Multicast Hit if:
  - ✓ MC_Enable is Set, and
  - ✓ The TLP is a Posted Request, and
  - ✓ Address targets a Multicast address range
    - Extract the Multicast Group (MCG) number from the target address

Route Table for Switch



Multicast Blocking

- Perform Multicast Blocked TLP checking based on the Multicast Group (MCG) :
  - ✓ If the Multicast TLP came in an Ingress Port containing a Multicast Capability structure, block the Multicast TLP if:
    - The associated MC_Block_All bit is Set, or
    - The associated MC_Block_Untranslated bit is Set and the address is Untranslated
  - ✓ If the Multicast TLP is being sent by an Endpoint Function containing a Multicast Capability structure, block the Multicast TLP if:
    - The associated MC_Block_All bit is Set, or
    - The associated MC_Block_Untranslated bit is Set and the address is Untranslated
  - ✓ Blocking function reports MC_Blocked_TLP via Advanced Error Reporting (AER)
- Allows switches to police Endpoints and Endpoints to police themselves against use of unauthorized MCGs
- Assures compatibility with systems that use Address Translation Services (ATS)

Multicast Replication

- Perform the Multicast based on the Multicast Group (MCG) number:
  - ✓ In Switches or Endpoint components, forward the TLP to each Upstream/Downstream Port or Endpoint Function whose associated MC_Receive bit is Set
  - ✓ In Root Complexes (RCs), forward the TLP to each Root Port or RC Integrated Endpoint whose associated MC_Receive bit is Set
  - ✓ Do not forward the TLP back to the source Port/Endpoint
    - Allows all members of a MCG to use the same MCG to multicast to the group
- Single copy of packet sits in switch buffer until DL_ACK is received from each port whose MC_Receive bit is set
  - ✓ Standard DLL ACK/NACK protocol independently on each egress link
  - ✓ De-allocate buffer location when all ACKs have been received
  - ✓ With appropriate switch buffer architecture, single slow port with a Multicast backlog won't block other ports until/unless ALL buffer memory is utilized

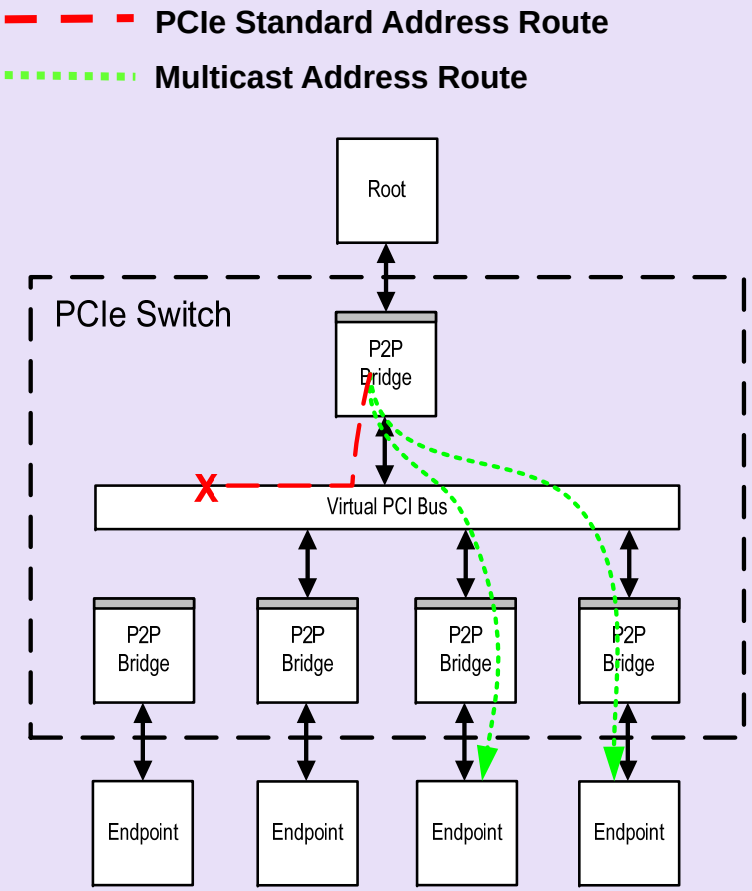
MC Overlay Mechanism

- Allows address in multicast packet to be translated in the egress port of a switch or Root Complex
  - ✓ If enabled, replace TLP_Address with MC_Overlay_BAR for bits >= MC_Overlay_Size
- At port above legacy Endpoint, translate Multicast (MC) TLP address to hit in the device's BAR
- At upstream port of switch, translate MC TLP address into local memory space
- If MC packet has End to end CRC (ECRC) and overlay is enabled:
  - ✓ Regenerate ECRC (robustly)
    - – Invert regenerated ECRC if find error in ECRC check prior to regeneration
  - ✓ Drop ECRC

ECRC Rules for MC Overlay

MC_Overlay Enabled	TLP has ECRC	ECRC Regeneration Supported	Action if ECRC Check Passes	Action if ECRC Check Fails
No	x	x	Forward TLP unmodified	
Yes	No	x	Forward modified TLP	
Yes	Yes	No	Forward modified TLP with ECRC dropped and TD bit clear	
Yes	Yes	Yes	Forward modified TLP with regenerated ECRC	Forward modified TLP with inverted regenerated ECRC

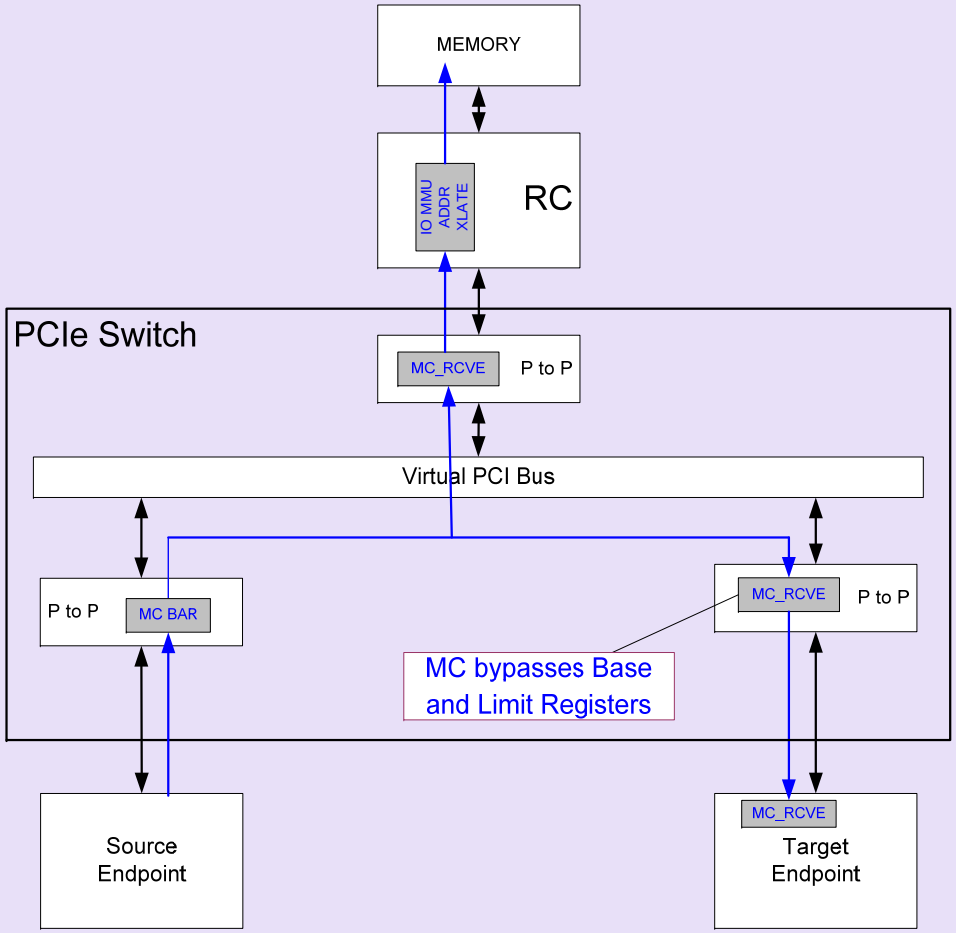
Multicast and Address Routing



- Request that hits a Multicast address range is routed unchanged to Ports that are part of the Multicast Group derived from Request address
- PCIe standard address route not used for multicast
  - ✓ Including default upstream route

Upstream Multicast

- Posted packet is Multicast if its address hits in MC BAR
  - ✓ Goes Upstream (US) if and only if US Port's Receive bit is set
- Each port with MC_Receive Set takes a copy of MC HIT packet
  - ✓ bypassing base and limit register address decoding
- Translate MC TLP Address to hit in local memory
  - ✓ IO MMU -Can scatter
  - ✓ Overlay mechanism -Need contiguous block of RAM

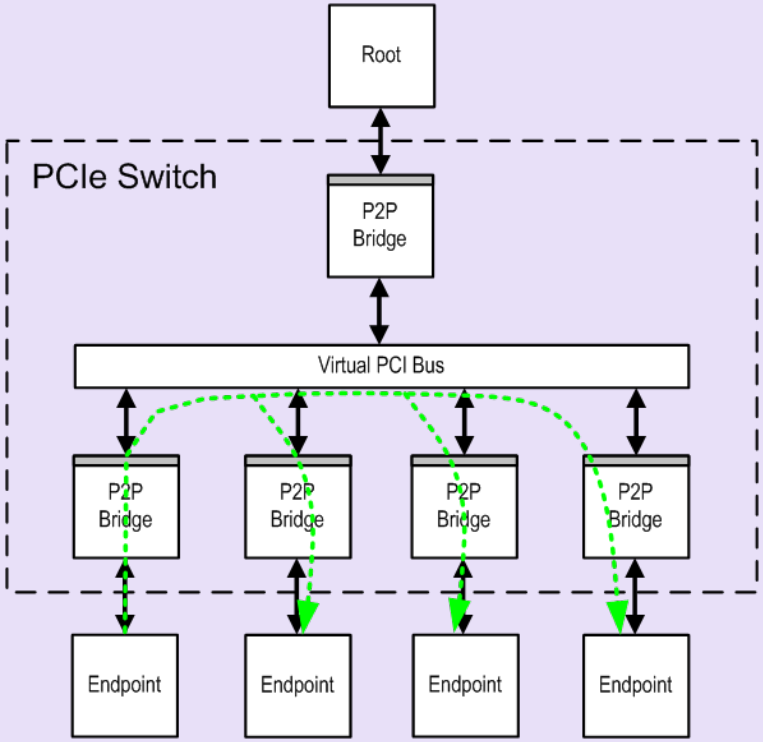


Downstream Multicast

- Potential congestion
  - ✓ Fat upstream port sending to narrow downstream ports
    - Traffic builds up in switch due to rate mismatch
    - Mitigate by source rate control, which may be implicit in the application
  - ✓ Can also be a problem in other flow patterns but aggravated by wider source port
- Multiple applications
  - ✓ Route table update in communications
  - ✓ Multi-headed graphics
  - ✓ Redundancy
    - Send copy to 2nd I/O device
    - Send copy along redundant path in hopes that at least one gets through

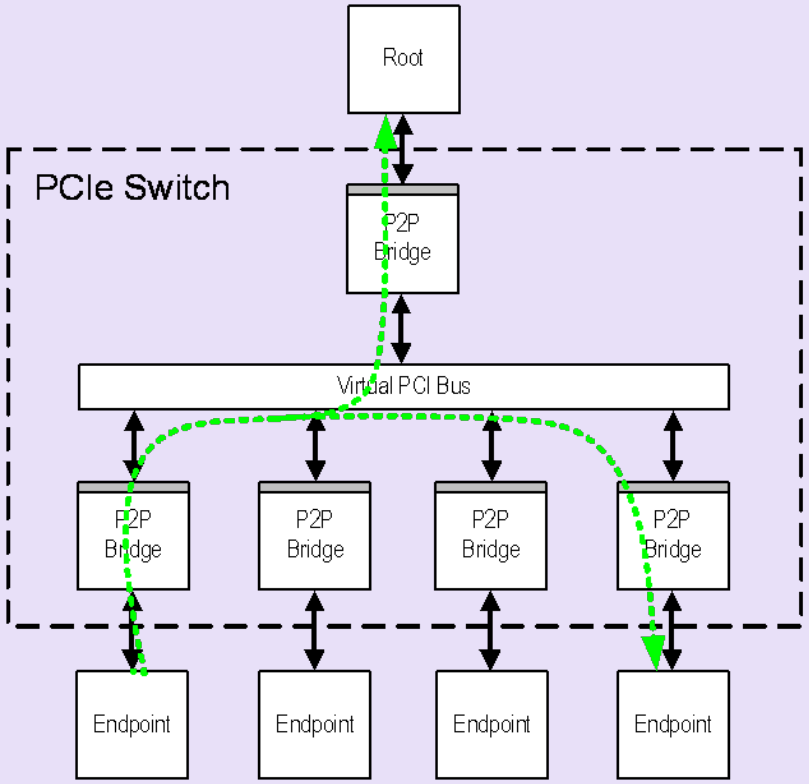
Peer-to-Peer Multicast

- From Downstream (DS) to Multiple Downstream ports
- Communications Example
  - Satellite receiver input
  - Multiple decoder cards operate in parallel
- Instrumentation Example
  - Computing FFT of data from DAC card using multiple DSP cards
  - Each DSP needs to see all the data to compute cross products



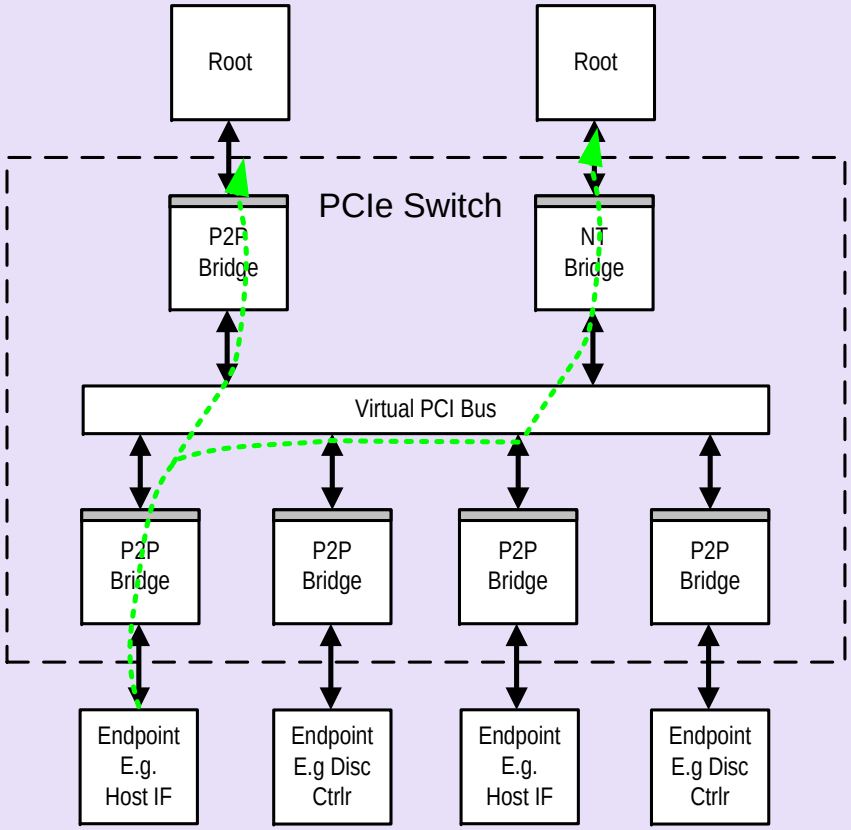
Peer plus Host Multicast

- From DS port to DS port plus host
- Note: Can only route to Root Complex via Receive register bit – default upstream route does not apply
- Instrumentation example
  - ✓ Data acquisition input
  - ✓ To host for processing
  - ✓ To peer for logging



MC to Redundant Hosts

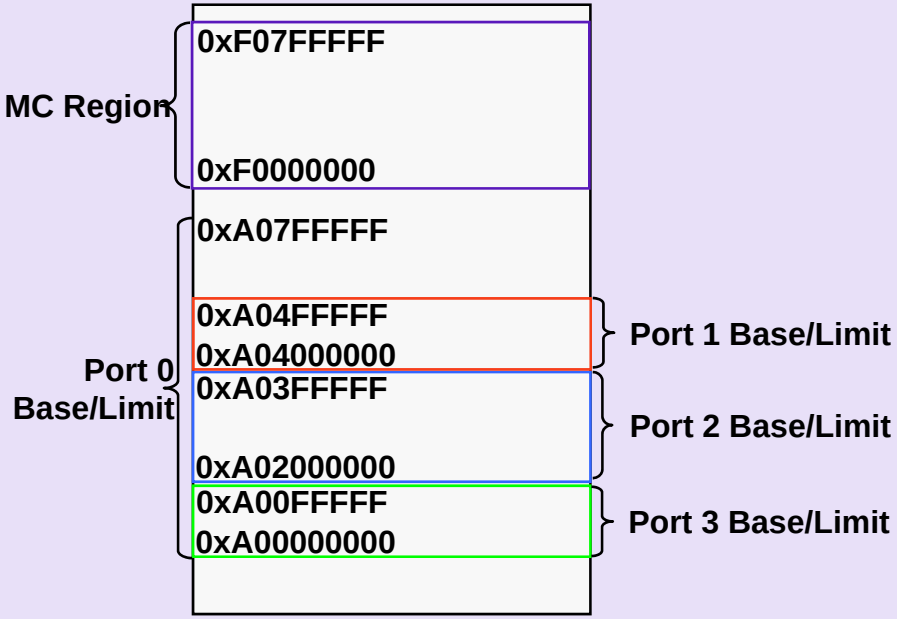
- Upstream to redundant hosts for host failover
  - ✓ Storage mirroring
- DMA writes from each Endpoint are Multicast to both hosts
  - ✓ Else endpoint's PCIe link would need to be 2x BW of external connection
- Redundant host just buffers the data
- Active host communicates its progress to Redundant host
  - ✓ E.g. via completion queue
- If active root fails, redundant root can take over seamlessly
  - ✓ Data written to disk is safe even if just in write buffer in Root Complex

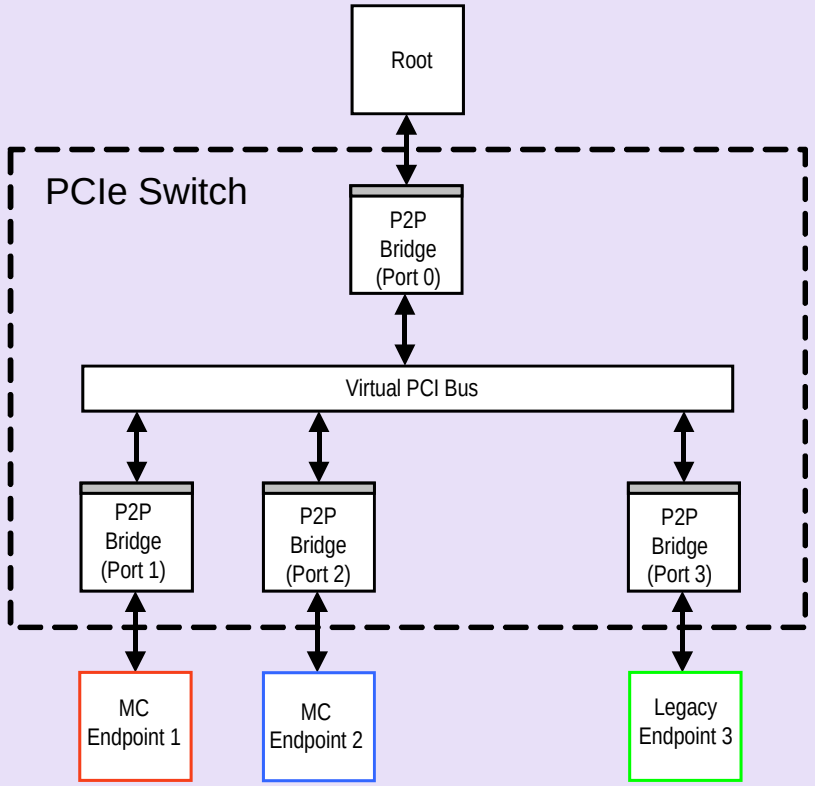


Multicast Register settings and routing Example

Multicast Worked Example

- The example system uses a mix of EPs with MC capability and one EP without MC capability (labeled “Legacy EP”)
- Assume that there are 8 MCGs (not all associated with this switch, i.e. another switch may be attached to the RC)
- Assume group window size of 1MB
- Memory Map





BAR0 Range = 1MB
0xA04000000 to 0xA04FFFFF

MCG’s 1,2,3

BAR0 Range = 2MB
0xA02000000 to 0xA03FFFFF

MCG 2,4

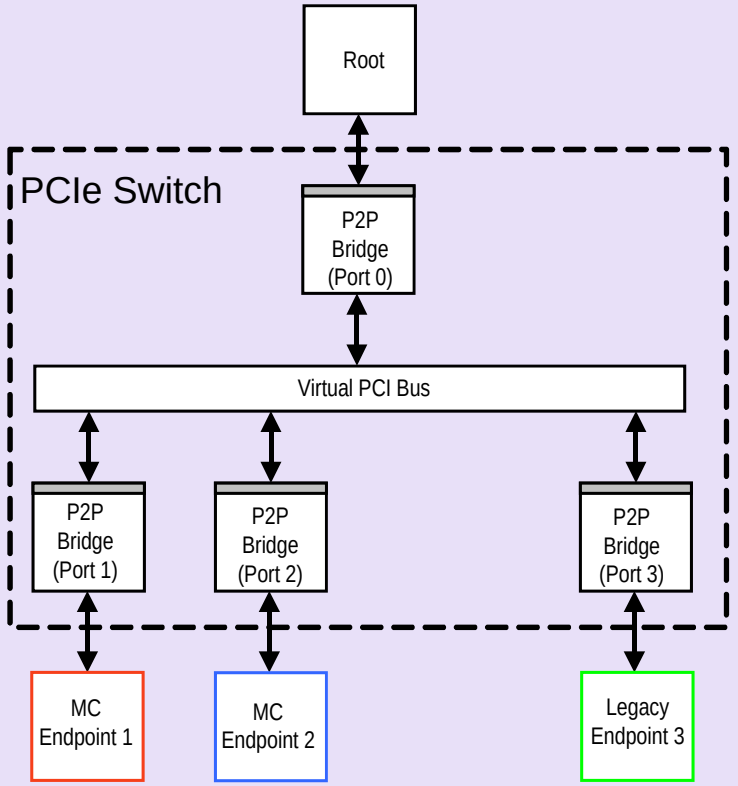
BAR0 Range = 1MB
0xA00000000 to 0xA00FFFFF

No MC capability structure

Multicast Worked Example Continued – Switch MC Regs.

- Switch Common Port Values:

Register	Value	Comment
PCIe Extended Cap. ID	= 0012h	
Capability Version	= 1	
MC_Max_Group	= 3Fh	64 groups
MC_Window_Size_Rqstd	= 0	Rsrvd in switches
MC_ECRC_Regen._Support	= 1	
MC_Num_Group	= 7	8 groups
MC_Enable	= 1	
MC_Index Position	= 10h	1MB
MC_Base_Address[31:12]	= 0xF0000	
MC_Base_Address[63:32]	= 0x00000000	
MC_Block_All[7:0]	= 00000000b	
MC_Block_Untranslated[7:0]	= 00000000b	
MC_Overlay_Size	= 0	Except port 3
MC_Overlay_BAR	= 0	Except port 3



BAR0 Range = 1MB 0xA04000000 to 0xA04FFFFF

MCG's 1,2,3

BAR0 Range = 2MB 0xA02000000 to 0xA03FFFFF

MCG 2,4

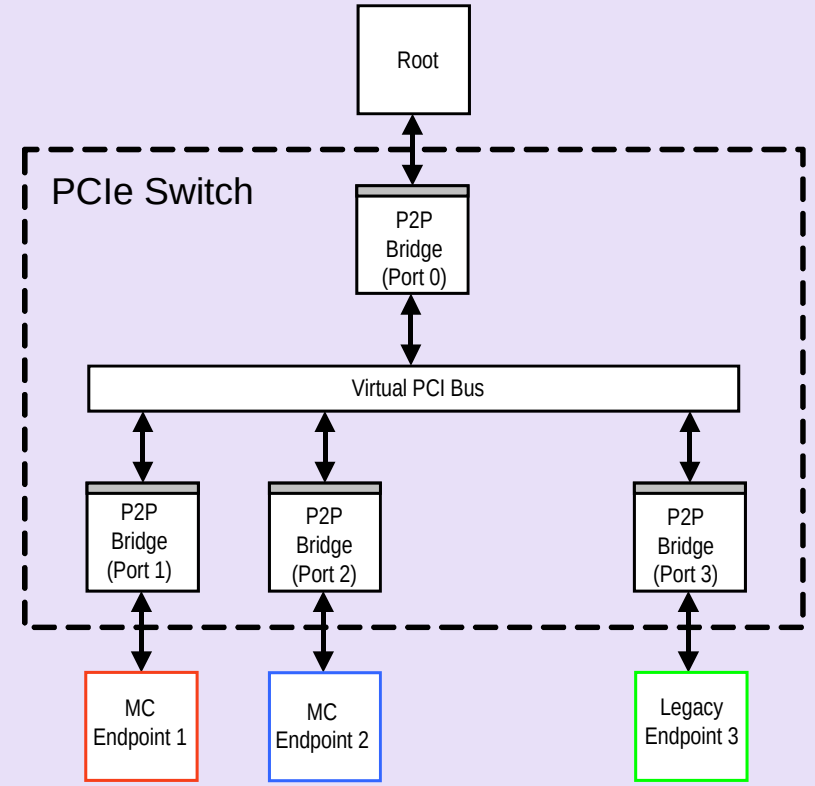
BAR0 Range = 1MB 0xA00000000 to 0xA00FFFFF

Not MC capable

Multicast Worked Example Continued – Switch MC Regs.

- Port 0 Specific values
  MC_Receive[7:0] = 00011110b
  MCG 1,2,3,4
- Port 1 Specific values
  MC_Receive[7:0] = 00001110b
  MCG 1,2,3
- Port 2 Specific values
  MC_Receive[7:0] = 00010100b
  MCG 2,4
- Port 3 Specific values
  MC_Receive[7:0] = 00010000b
  MCG 4

MC_Overlay_Size = 10h 1MB
MC_Overlay_BAR[31:20] = 0xA00 [19:6]=0
MC_Overlay_BAR[63:32] = 0x00000000



BAR0 Range = 1MB
0xA04000000 to
0xA04FFFFF

MCG's 1,2,3

BAR0 Range = 2MB
0xA02000000 to
0xA03FFFFF

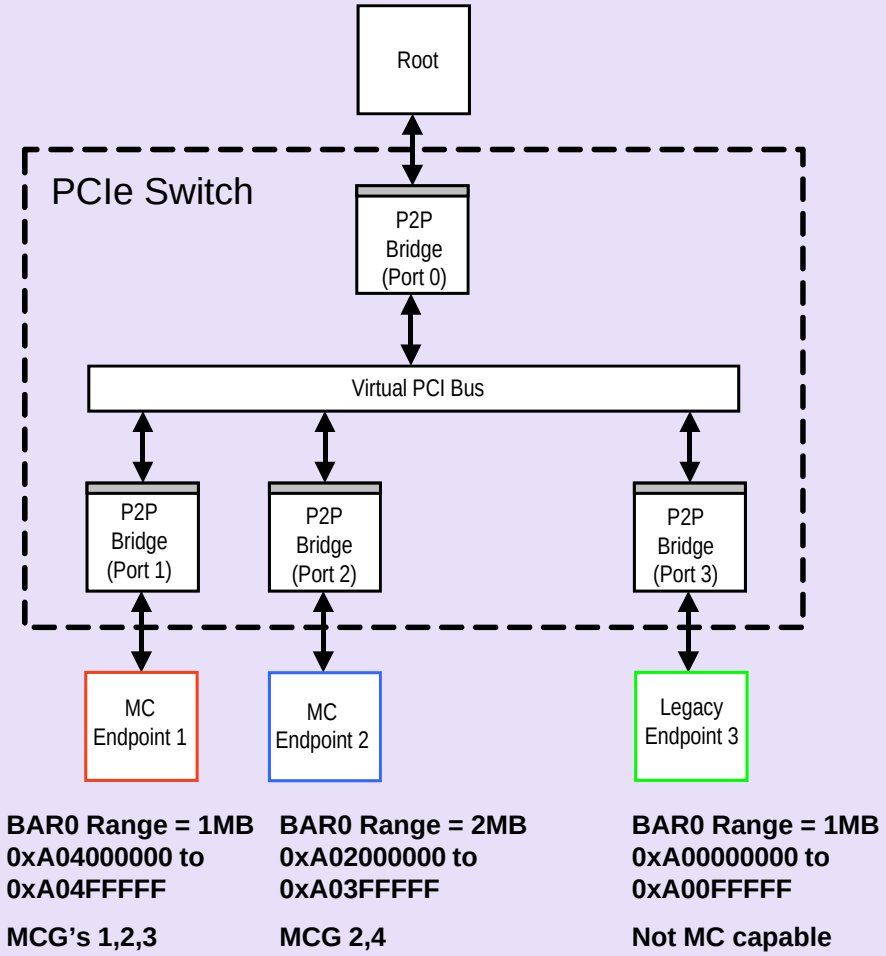
MCG 2,4

BAR0 Range = 1MB
0xA00000000 to
0xA00FFFFF

Not MC capable

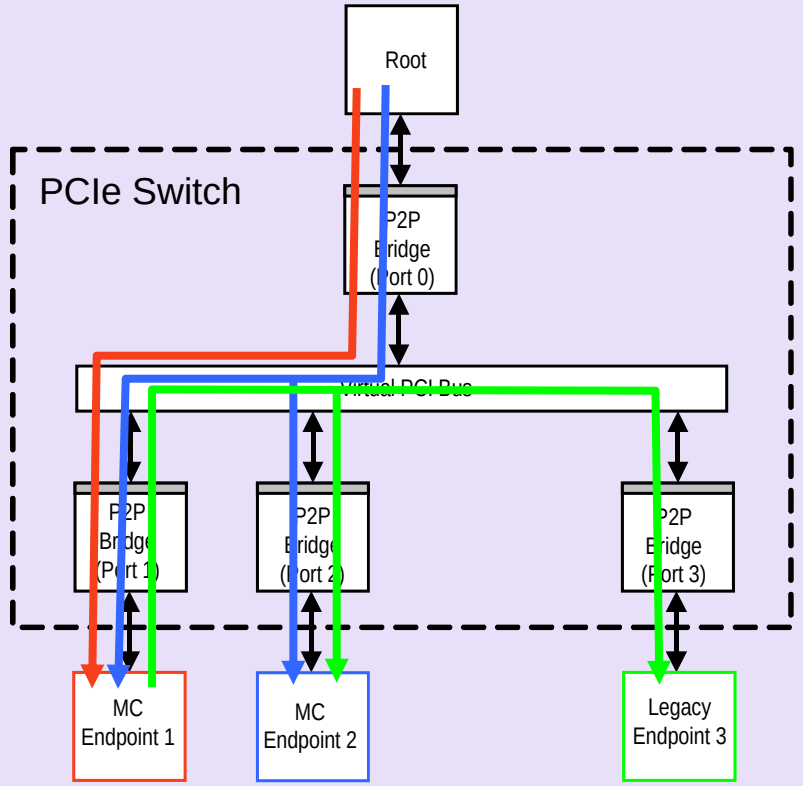
Multicast Worked Example Continued – EP MC Regs.

- Endpoint Multicast Capability Structure values
  - ✓ Most are the same as the switch
  - ✓ MC_Window_Size_Rqstd = 10h
    - 1MB for all EPs
- EP1 Specific values
  MC_Receive[7:0] = 00001110b MCG 1,2,3
- EP2 Specific values
  MC_Receive[7:0] = 00010100b MCG 2,4
- EP3 is not MC capable



Multicast Worked Example Continued – Transactions

- Example transactions
  - ✓ Root Writes to address 0xF0100100
    - 0xF0100100 is within MCG 1 so the write is forwarded to EP1 (red)
  - ✓ Root Writes to address 0xF0200000
    - 0xF0200000 is within MCG 2 so the write is forwarded to EP1 and EP2 (blue)
  - ✓ EP1 Writes to address 0xF0400010
    - 0xF0400010 is within MCG 4 so the write is forwarded to EP2 and using the Overlay function to EP3. The write to EP3 is translated to address 0xA0000010 to hit EP3's BAR0 (green)



BAR0 Range = 1MB 0xA04000000 to 0xA04FFFFF

MCG's 1,2,3

BAR0 Range = 2MB 0xA02000000 to 0xA03FFFFF

MCG 2,4

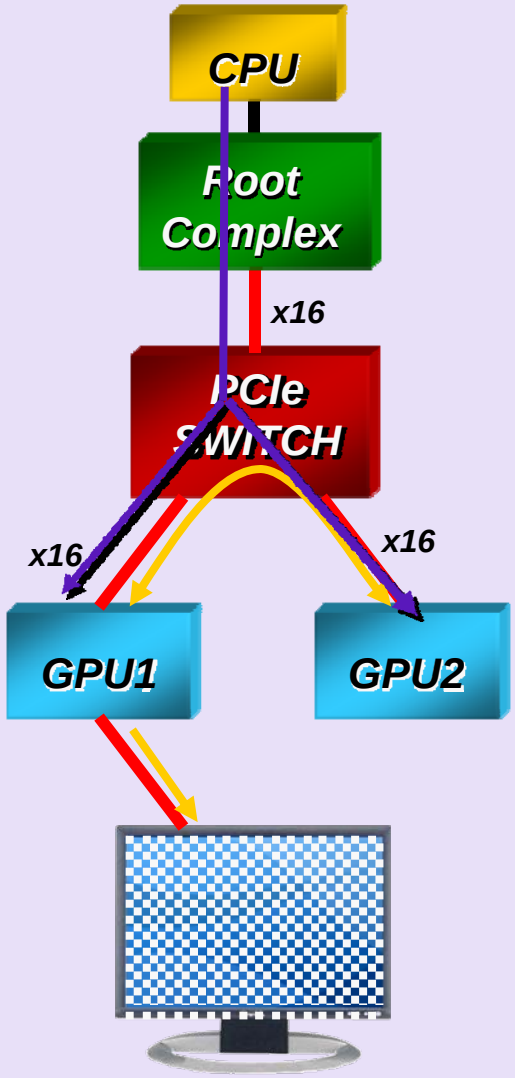
BAR0 Range = 1MB 0xA00000000 to 0xA00FFFFF

Not MC capable

Multicast Applications and Usage Examples

Example: Graphics & FP Acceleration

- Dual-headed graphics
  - ✓ Each GPU paints ½ the screen
  - ✓ Multicast commands downstream
    - E.g. vector list
  - ✓ Use peer to peer to transfer bit map from GPU2 to GPU1
- General FP acceleration using GPUs where some GPUs need to see the same data
  - ✓ Push data or commands downstream to multiple GPUs/FPUs



Example: Decoding/Decompression

- Encoded/compressed video data comes in peer board and is multicast to multiple peer decoder boards
- System contains N decoder boards operating in parallel that need to see all the data
  - ✓ Each decoder responsible for 1/N scan lines but needs to see adjacent scan lines to perform the computation
  - ✓ Use N Multicast groups, each of which specifies a tri-cast and is used for sets of 3 adjacent scan lines repeatedly down the frame. Use two more Multicast groups for first and last scan lines
  - ✓ Alternately
    - If BW available, send entire image to all

Example: Tunneling Ethernet thru PCIe

- Ethernet Tunneling
  - ✓ PCIe payloads are Ethernet packet or fragments
  - ✓ Same upper layer software used as with standard Ethernet physical layer
- Multicast uses when tunneling Ethernet
  - ✓ Address Resolution Protocol (ARP) broadcast
  - ✓ VLAN support
    - Each VLAN requires a Multicast Group
  - ✓ IP multicast mapped onto PCIe multicast
- Allows PCIe to replace Ethernet in communications system control plane and for host to host communications in blades

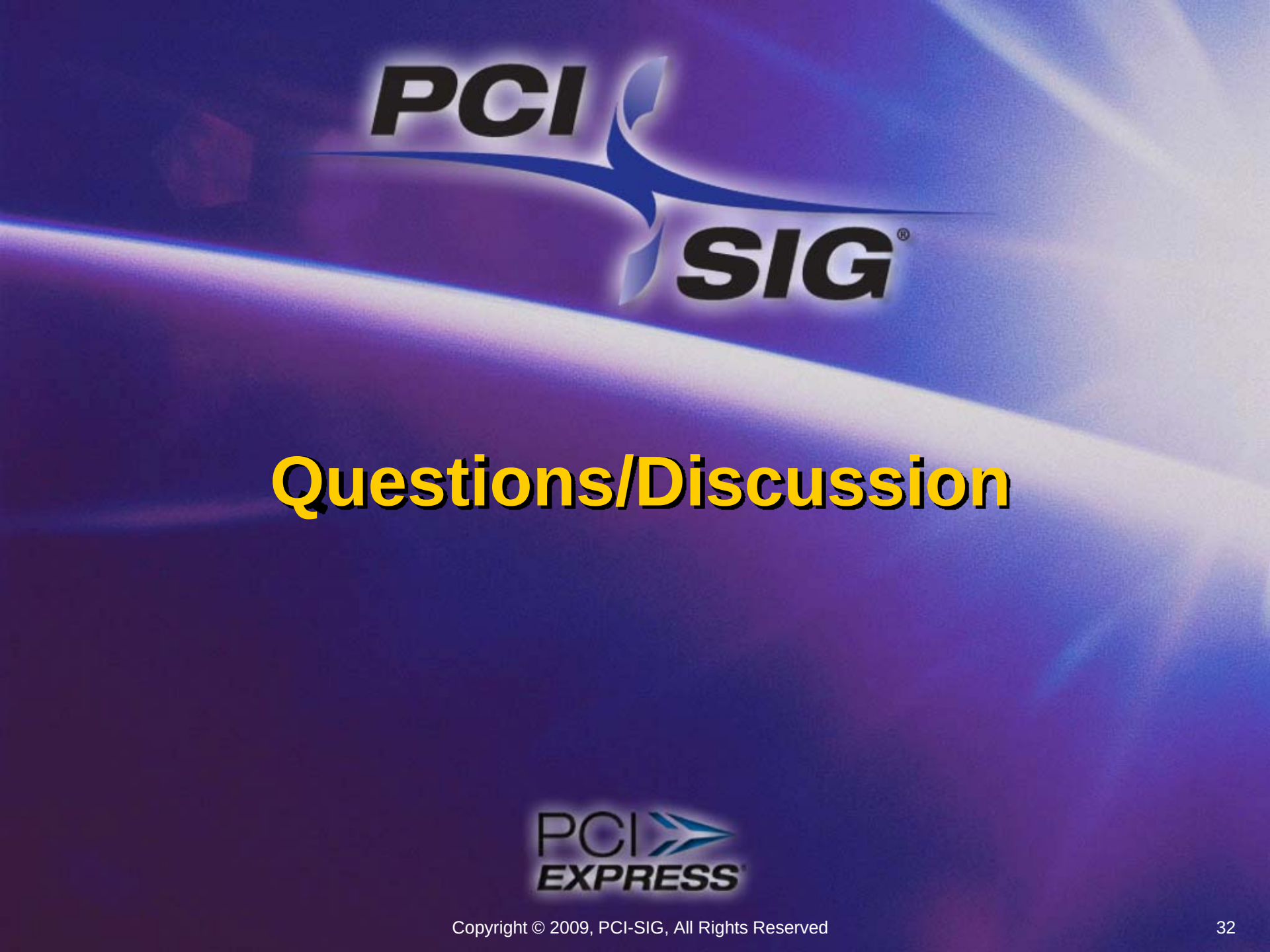
Multicast Implementation Suggestions

How many MCGs Supported?

- How many Multicast Groups (MCGs) should a component support?
  - ✓ Multicast ECN has architectural limit of 64 groups
    - Supporting 64 is trivial cost in context of multi-million gate device
    - 64 groups supports all combinations of a 6-port switch
- Most applications characterized by a small number of Multicast flows
  - ✓ Support for only16 groups might allow several Multicast applications to run through a common switch
  - ✓ Each endpoint would elect to receive 0-several groups from among the total of 16 active in the fabric

Multicast Timeline

- Multicast ECN was approved in May 2008
- Expect Multicast support in switches in 2009



Questions/Discussion

Thank you for attending the
PCI-SIG Developers Conference 2009

For more information please go to
www.pcisig.com