Multicast PCI Express®

Jack Regula

PLX Technology

Agenda

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

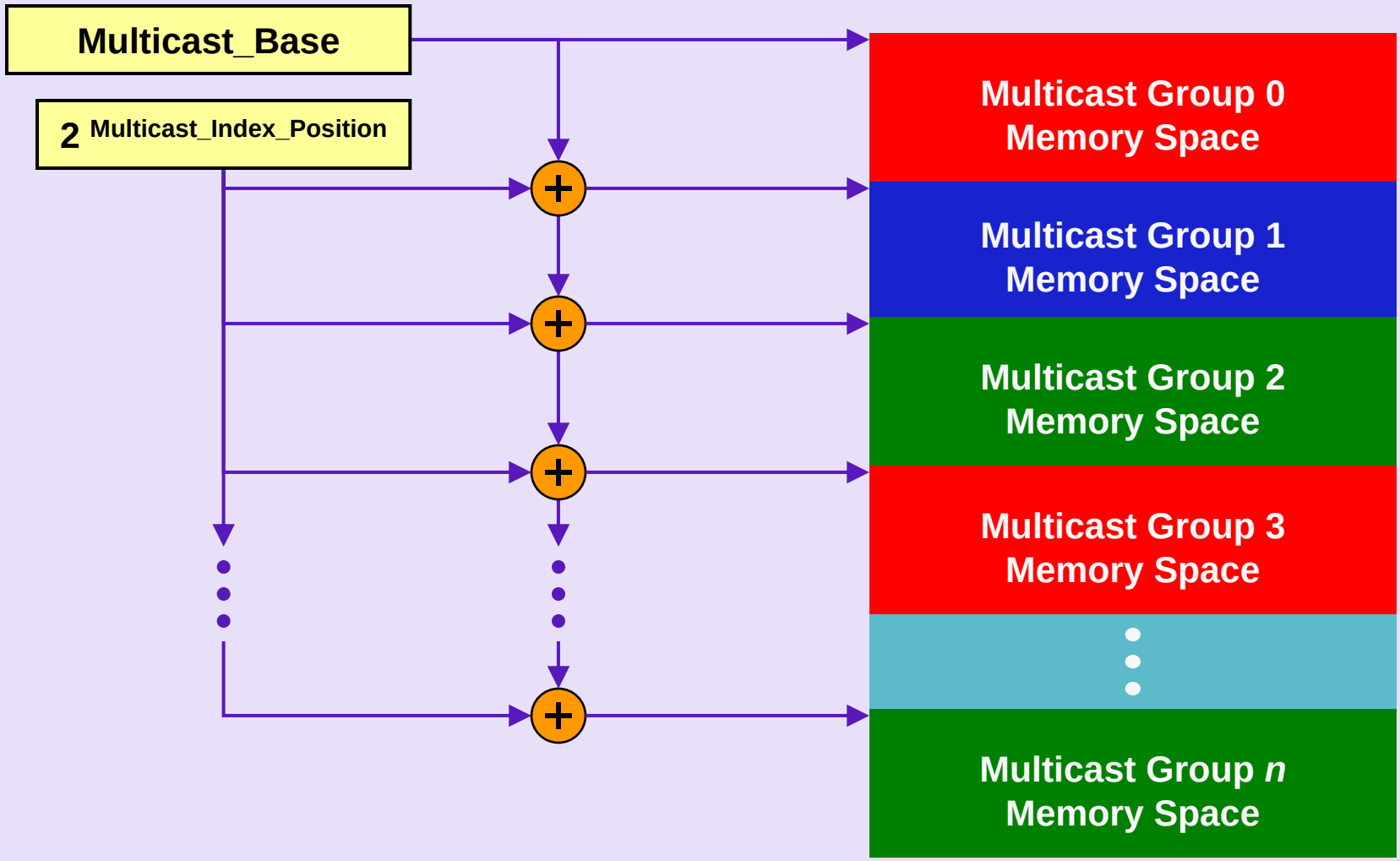
What is Multicast

- ***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 MC Extensions 8/2004
  - ✓ ASI included multicast
  - ✓ Infiniband supports unreliable MC
- Benefits:
  - ✓ Unblock bottleneck of PCIe link at source and between switches
  - ✓ Unblock bottleneck of memory BW at source
  - ✓ Make PCIe more suitable for backplane applications

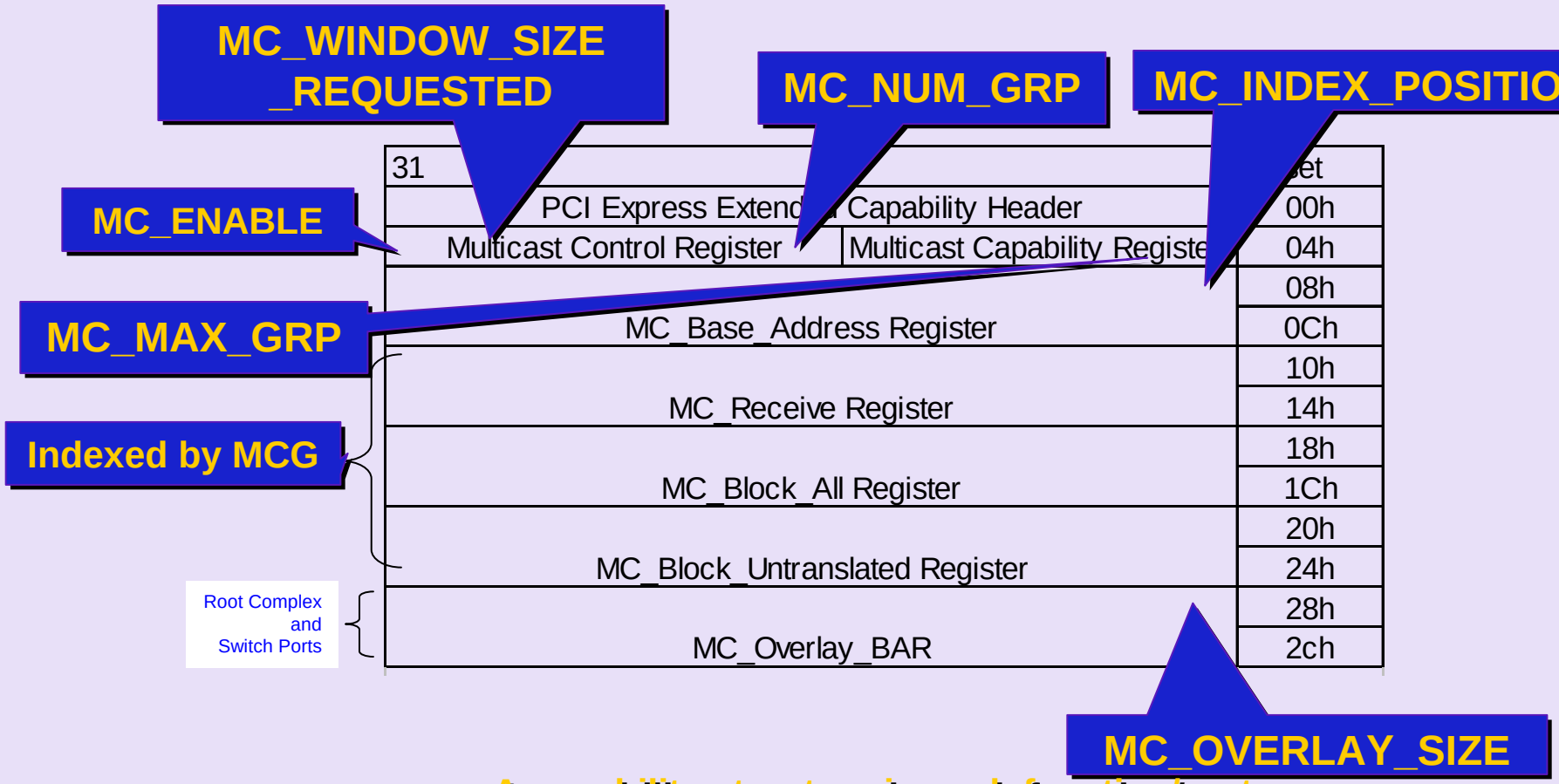
PCIe Multicast

- Address based
  - ✓ MC BAR creates an MC space
  - ✓ Posted packets that hit in the MC 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 MC space
  - ✓ Legacy devices can be multicast targets
- Multicast ECN
  - ✓ Specifies how PCIe components – switches, RCs, and EPs - implement MC
  - ✓ Improves MC flexibility and protection for EPs that participate in MC but also allows use of legacy EPs

Multicast Memory Space



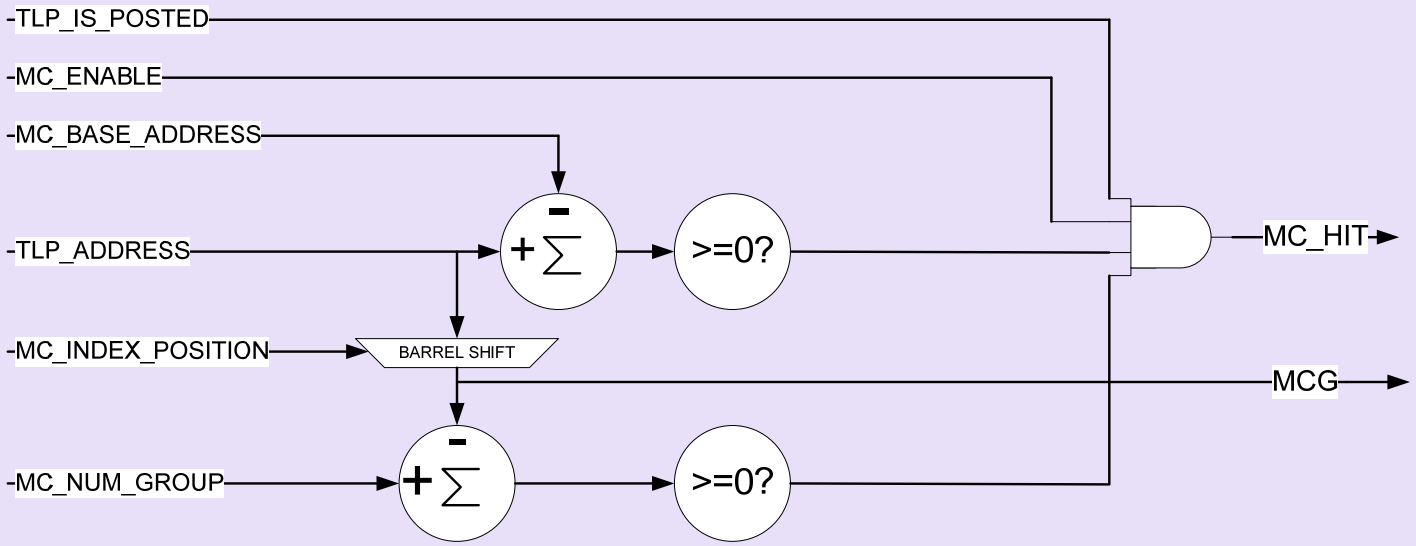
MC Capability Structure



A capability structure in each function/port

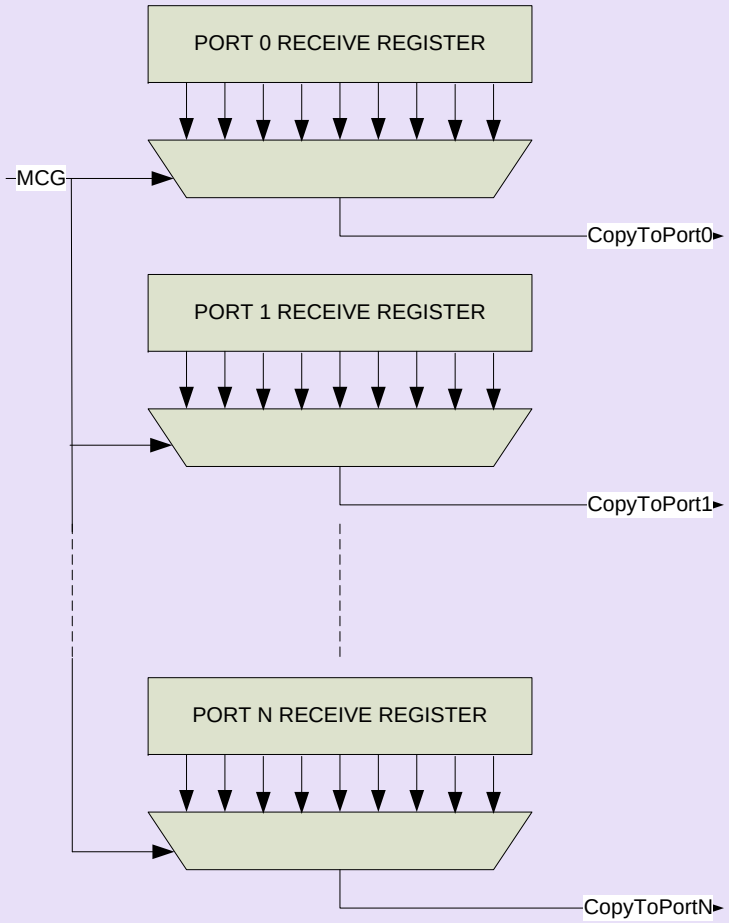
Implementation efficiencies in MFD since fields must be configured identically

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



MC Blocking

- Perform MC Blocked TLP checking based on the 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 AER
- Allows switches to police EPs and EPs to police themselves against use of unauthorized MCGs
- Assures compatibility with systems that use ATS

Multicast Replication

- Perform the Multicast based on the 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 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 MC backlog won't block other ports until/unless ALL buffer memory is utilized

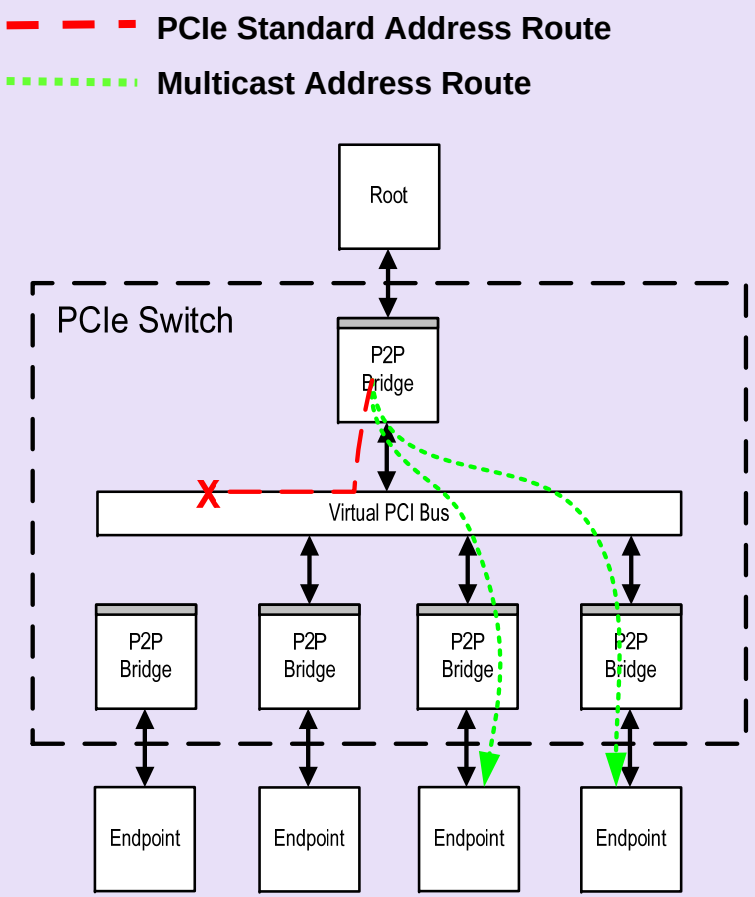
MC Overlay Mechanism

- Allows address in multicast packet to be translated in the egress port of a switch or RC
  - ✓ If enabled, replace TLP_Address with MC_Overlay_BAR for bits >= MC_Overlay_Size
- At port above legacy EP, translate 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 ECRC and overlay is enabled:
  - ✓ Regenerate ECRC (robustly)
    - – Invert regenerated ECRC if find error in ECRC check prior to regen
  - ✓ 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

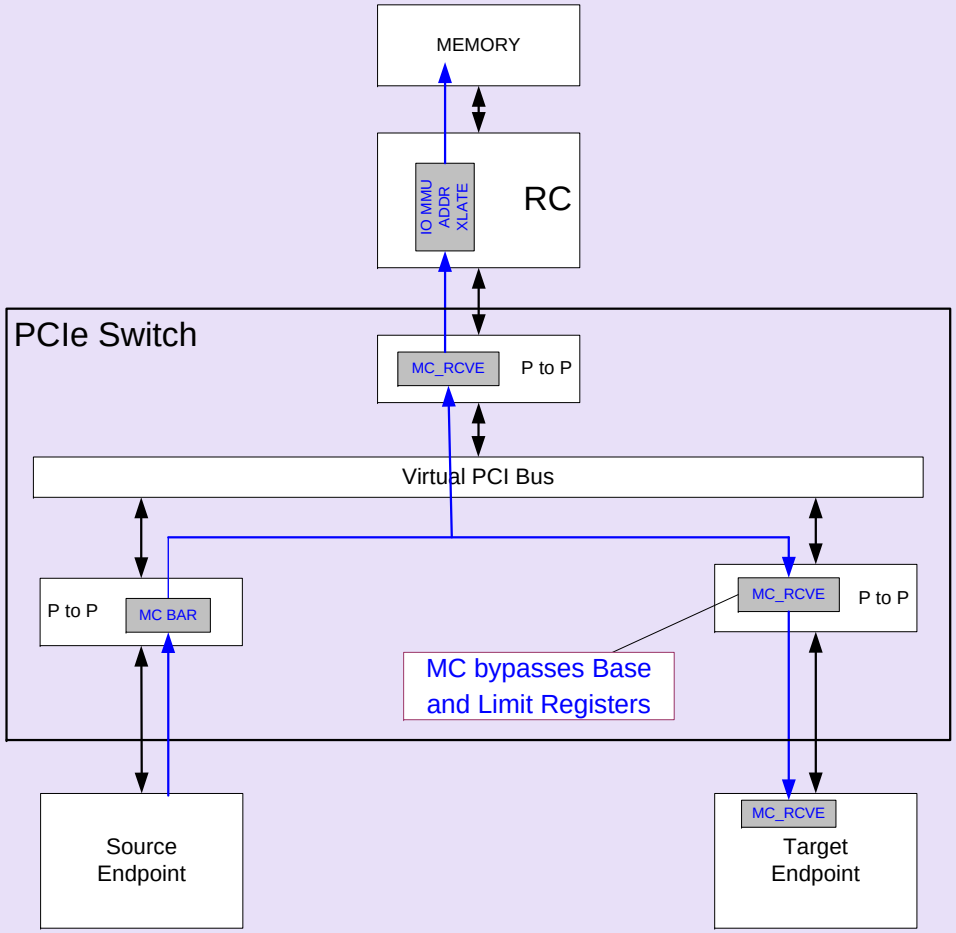
Multicast and Address Routing



- 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 MC if its address hits in MC BAR
  - ✓ Goes upstream iff 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

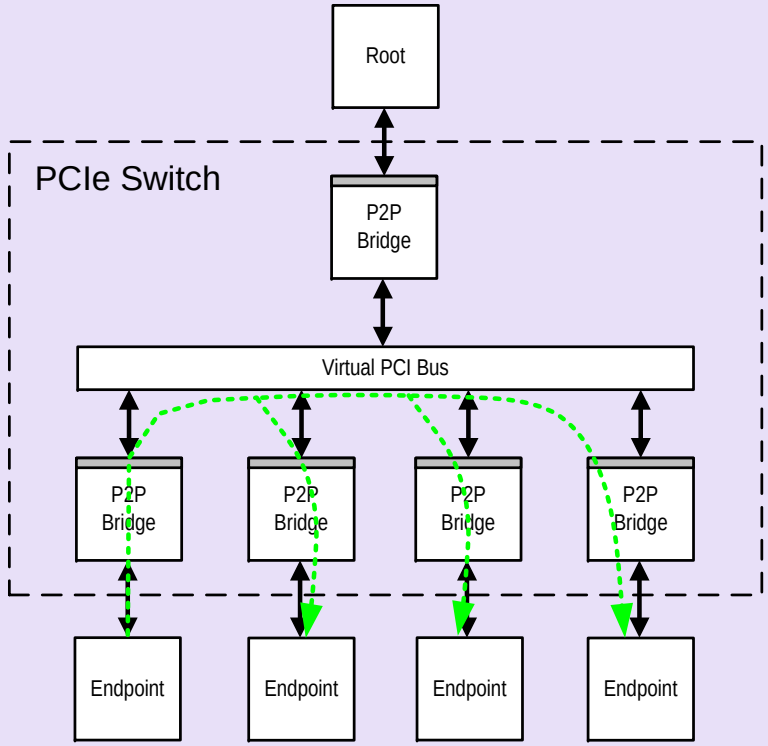


Downstream Multicast

- 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

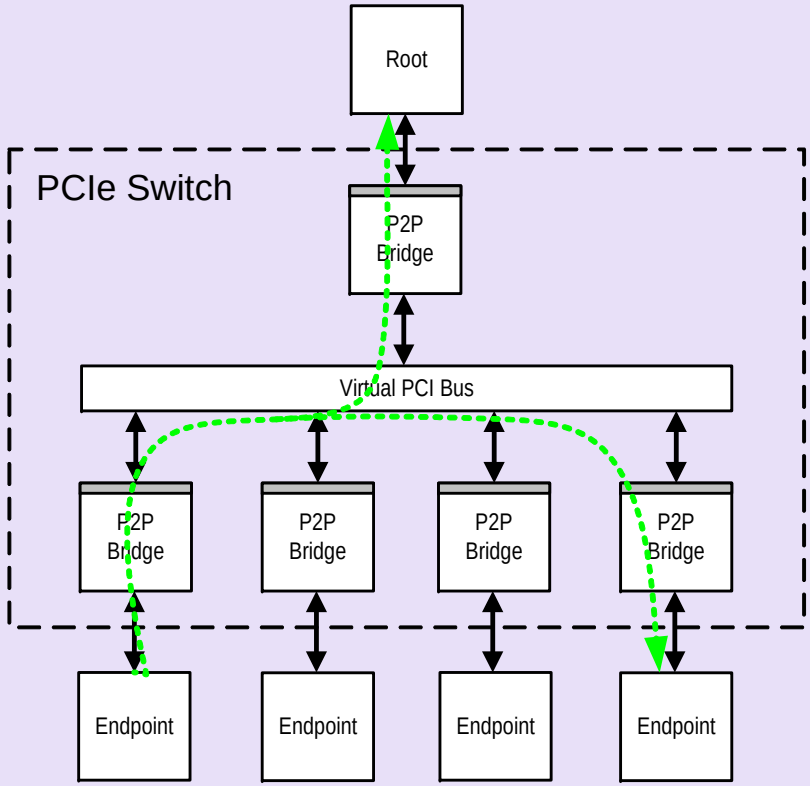
Peer to Peer Multicast

- ✓ From DS to Multiple DS 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



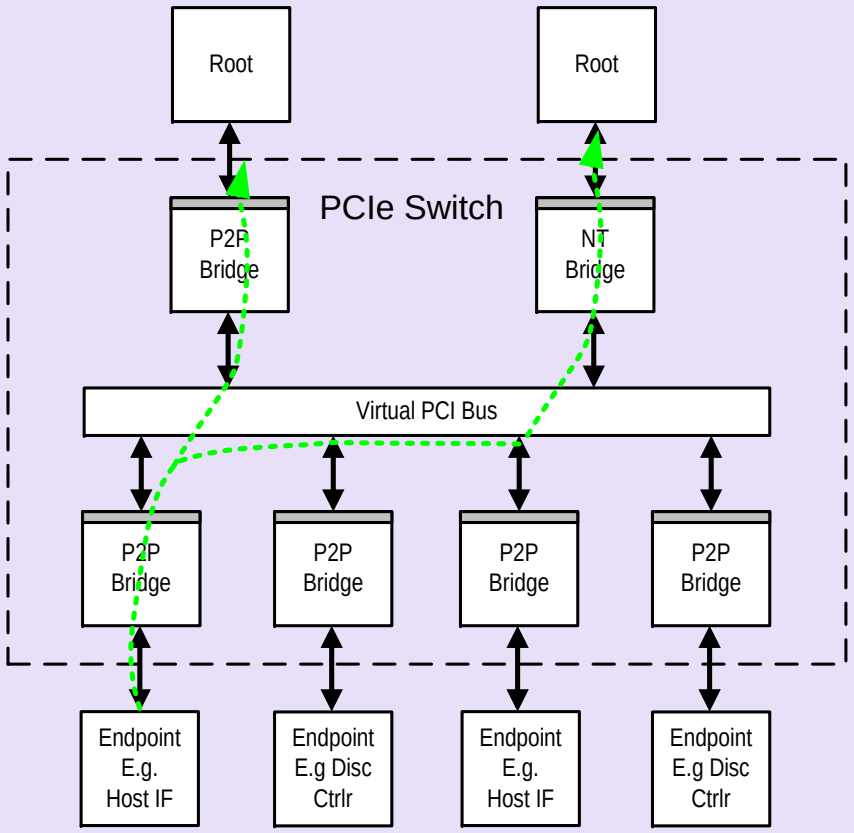
Peer plus Host Multicast

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



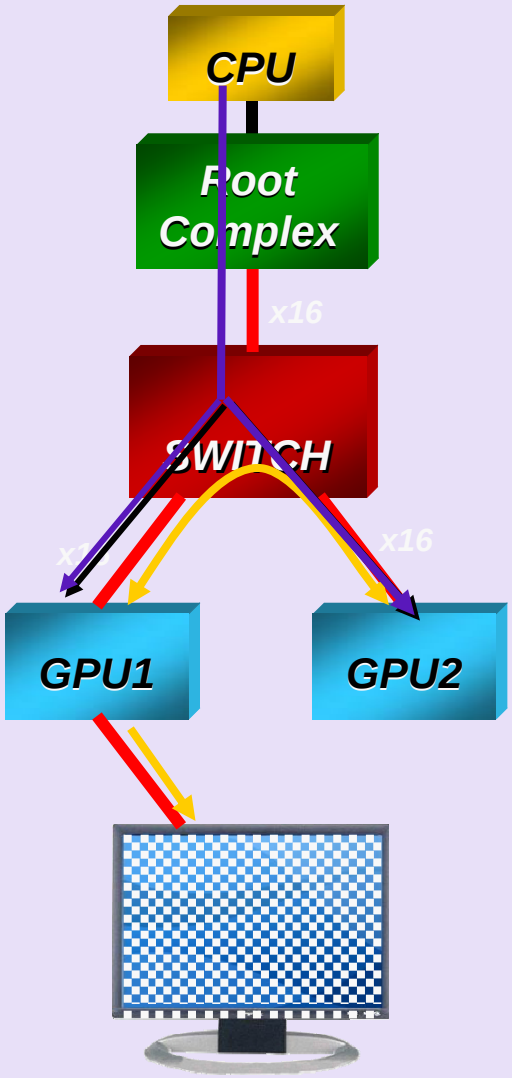
MC to Redundant Hosts

- Upstream to redundant hosts for host failover
  - ✓ Storage mirroring
- DMA writes from each EP are MCed 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 RC



Graphics & FP acceleration

- Dual-headed graphics
  - ✓ Each GPU paints ½ the screen
  - ✓ MC 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



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 MC groups, each of which specifies a tri-cast and is used for sets of 3 adjacent scanlines repeatedly down the frame. Use two more MC groups for first and last scan lines
  - ✓ Alternately
    - If BW available, send entire image to all

Tunneling Ethernet thru PCIe

- Ethernet Tunneling
  - ✓ PCIe payloads are Ethernet packet or fragments
  - ✓ Same upper layer software used as with standard Ethernet physical layer
- MC uses when tunneling Ethernet
  - ✓ ARP broadcast
  - ✓ VLAN support
    - Each VLAN requires a MC 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

How many MCGs Supported?

- How many MC groups should a component support?
  - ✓ MC 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 MC flows
  - ✓ Support for only16 groups might allow several MC 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

- MC ECN should be approved by the time you read this
- Expect MC support in switches in 2009

Thank you for attending the
PCI-SIG Developers Conference 2008

For more information please go to
www.pcisig.com