Cabled PCI Express® – Implementation Considerations

Lee Mohrmann

R&D Engineer

National Instruments

Session Outline

- Overview of Cabled PCIe®
  - ✓ Why external cabling
  - ✓ Applications of Cabled PCIe
  - ✓ Interoperability Considerations

Disclaimer

- Material contained in this presentation are the recommendations and opinions of National Instruments and not the PCI-SIG®.

Why PCI Express® Cabling?

- PCI-SIG® members were surveyed in April 2003
  - ✓ Responses representing several market segments indicated that cabling was required
    - – Extend PCI Express protocol / functionality across arbitrary distances and packaging
- As a result the Cabling Working Group was formed
  - ✓ Charter is to create a specification that focuses on
    - – Standard cable connectors
    - – Copper cabling attributes and electrical characteristics
    - – Connector retention
    - – Identification/labeling
  - ✓ This is **NOT** a replacement for cabling to USB or 1394 peripherals!

P1394a is an IEEE standard
USB is a standard of the USBIF

Advantages of Cabled PCIe

- Native to System Architecture
- No Bridging Required
- High Speed/High Bandwidth
- Compatible with PCI™ Software Model

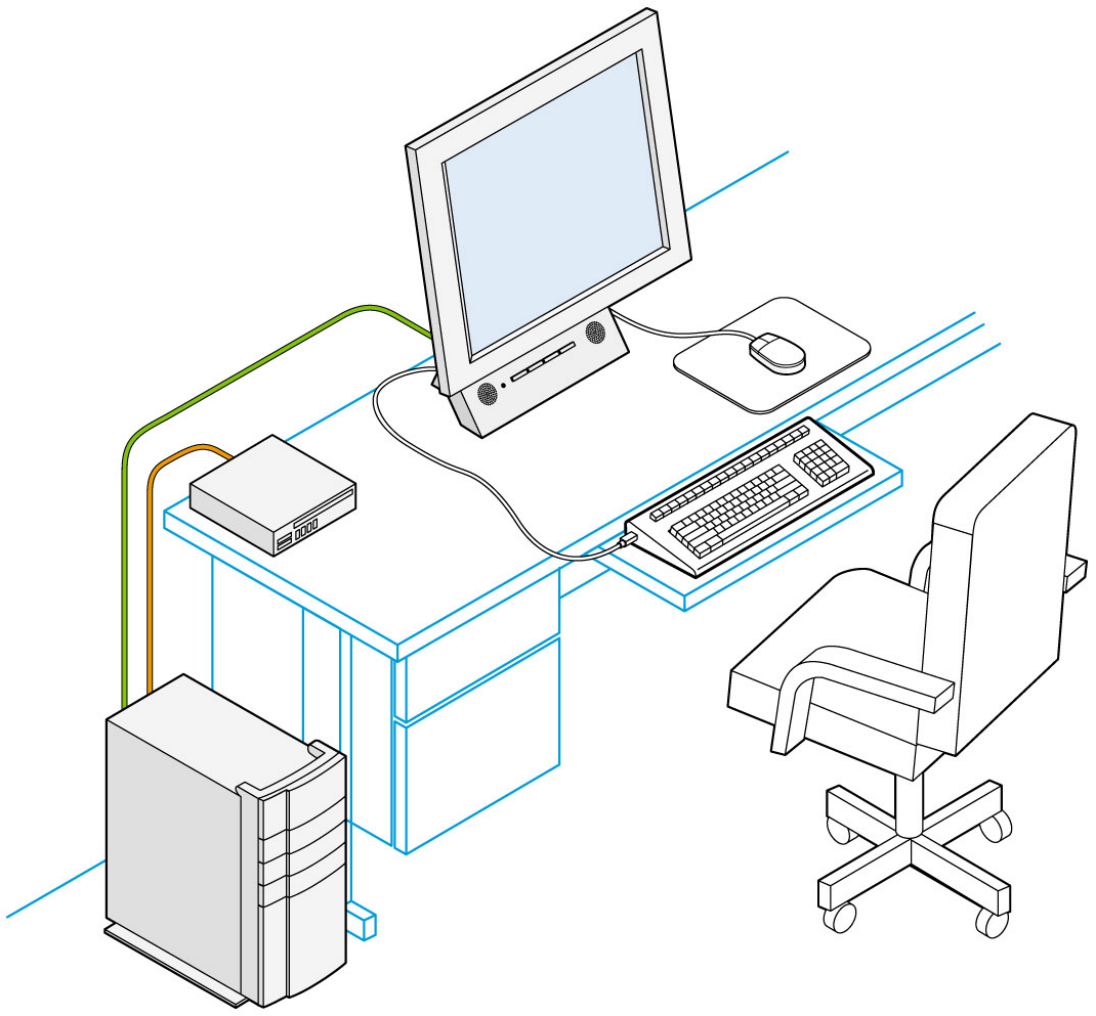
Applications of Cabled PCI Express

Examples of Cabled PCIe Usage Models

- Expansion I/O
  - ✓ Potential Implementations:
    - Mobile / Desktop / Server platform
    - ExpressModule
    - Test & Measurement chassis
    - ExpressCard®
- Split-system (disaggregate) desktop
- Tethered docking for mobile platforms
- External graphics controllers
- Communication equipment
- Embedded applications
  - High speed data transfer within large office equipment

ExpressCard is a trademark of the PCMCIA

Example: Split System Application



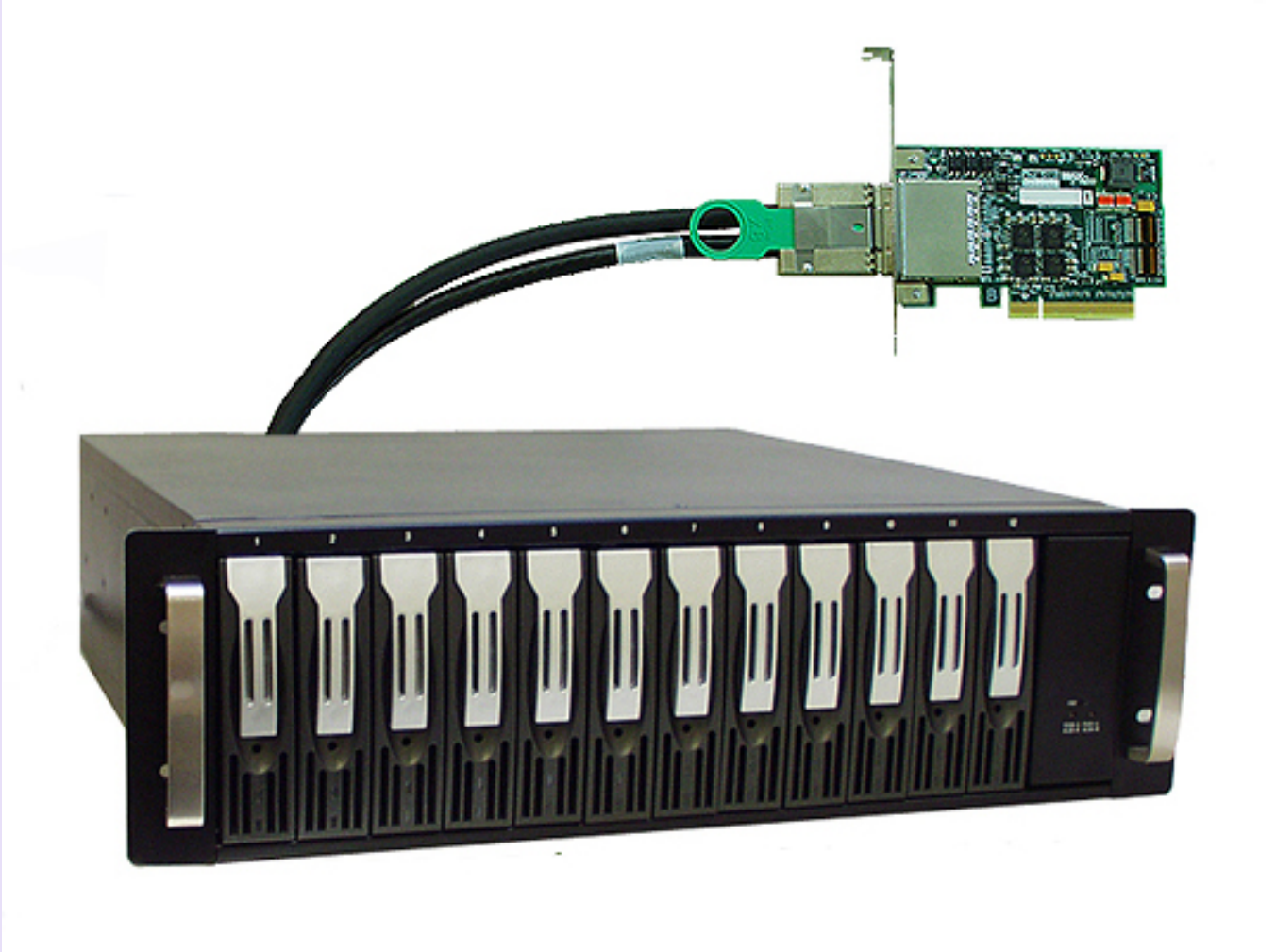
Example: Desktop x1 Expansion

Example: Desktop x4 Expansion

Example: ExpressCard Expansion

Example: External Switch

Example: x8 External RAID Array



PCI Express Cabling Interoperability Considerations

Interoperability Concerns

- BIOS Enumeration of PCI
- Hot Plug Issues
- Electrical Repeaters
- ASPM
- Long Haul Cables
- Clocking

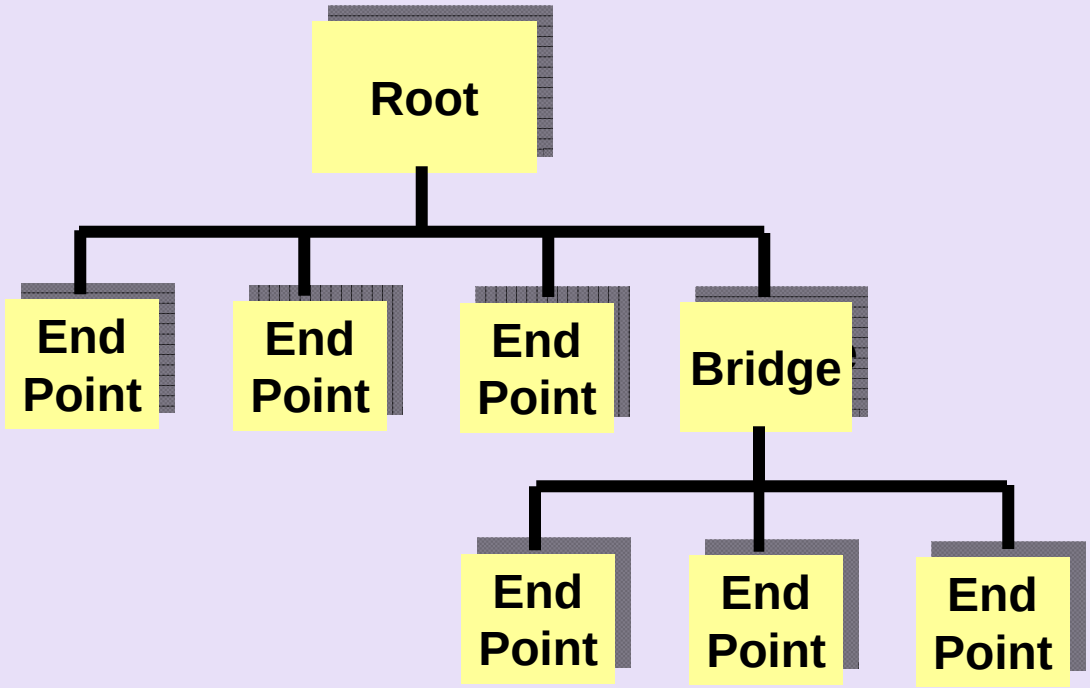
PCI Bus Enumeration

- System BIOS must enumerate PCI bus chain during system startup
- Not all system firmware implementations perform a complete PCI bus scan in the interest of saving timing during the boot process
  - ✓ Some busses are not enumerated and loss of functionality may occur.

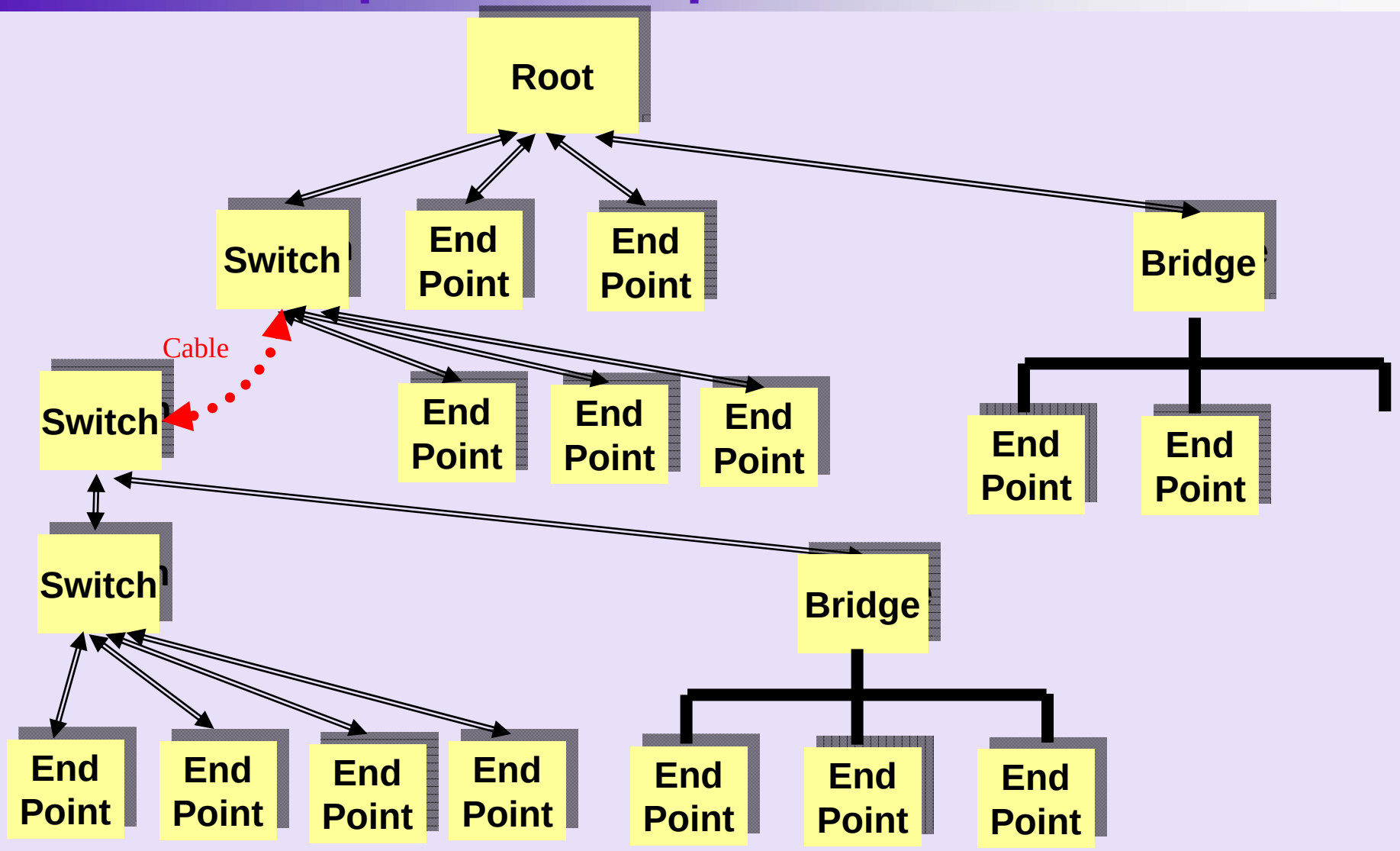
PCI Enumeration Background

- PCs usually implement significantly fewer PCI buses/devices than allowed by the spec
- Enumeration often limited to a few expansion scenarios
  - ✓ Notebook docking stations
  - ✓ Cardbus/ExpressCard modules with simple PCI topologies
- BIOS configures PCI hierarchy dynamically or statically (based on assumptions)
- Many OS versions assume BIOS configuration of all levels of PCI-PCI bridging
- PCI Express contributes to PCI bus proliferation and complexity

Common PCI Implementation



Complex PCI Implementation With PCIe



PCI Enumeration Problem

- Problem
  - ✓ Systems that limit PCI enumeration can cause interoperability failures with modules that include complex PCI hierarchies
    - In the worst case, system will hang during POST when configuring complex modules
    - NI has observed this issue with several released BIOSes
      - Occurs in both Desktop and Notebook PC's
  - ✓ This type of failure is fatal for the general expansion use case

PCI Enumeration Recommendation

- Recommendation
  - ✓ BIOS should traverse and configure all possible PCI buses
    - – Need to ensure that we do not cause a poor user experience
      - Loss of Functionality
      - System Lockup
    - – Possible option bits can be presented to user in configuration for bus scans
      - Full Bus scan
      - Limit Bus scan
- Notes on ExpressCard
  - ✓ System BIOS should configure all busses subordinate to a PCI Express ExpressCard interface

Hot Plug Capability

- Hot Plug signaling supported through PCIe native detection and CPRSNT#
  - ✓ PCIe ports can detect a status change following cable insertion/removal once both subsystems have a valid power condition and exit PERST#
  - ✓ Hot Plug notification can also be achieved through sideband signaling using Cable Presence Detect
    - Note that CPRSNT# going active indicates a downstream subsystem is ready to receive a reference clock
    - Such downstream subsystem will remain in a reset condition for a defined minimum time period

Hot-plug Resource Allocation Background

- Background
  - ✓ Hot-inserted PCI devices must draw memory, IO, and bus number resources from existing, boot-time pool
  - ✓ Some OS implementations use BIOS-assigned PCI-PCI bridge resources to PCI devices hot-inserted subordinate to the bridge
    - – Memory/IO resources drawn from pool associated with bridge's base/limit registers
    - – PCI bus number resources drawn from pool associated with bridge's secondary/subordinate bus number range
  - ✓ If system supports hot-plug, BIOS must assign adequate resources to parent PCI-PCI bridge to ensure child devices receive adequate resources when hot-inserted

Hot-plug Resource Allocation Problem

- Problem
  - ✓ External connections increase potential for hot-inserted PCI systems to exceed bridge's allocated resources
    - PCI Express switches
    - Cabled PCI Express expansion solutions
  - ✓ If OS cannot assign adequate resources, reboot required
    - Yellow "!" in Device Manager
    - Bad customer experience

Hot-plug Resource Allocation Recommendation

- BIOS offers configuration option for tweaking memory, IO, and bus number pre-allocations
  - ✓ Allows advanced users to enable hot-plug on systems with complex PCI hierarchies
  - ✓ Some BIOSes already implementing options for adjusting pre-allocated resource amounts
- BIOS assigns reasonable resources to ExpressCard Root Port bridges

Other Hot-Plug Considerations

- Inconsistent FW/OS level support can lead to a variety of user experience issues from reported errors to system failure

- Understand the hardware, FW, and OS support for hot-plug events for your implementation
  - ✓ Include indicators and user documentation that clearly specifies how your system will behave to set reasonable expectations

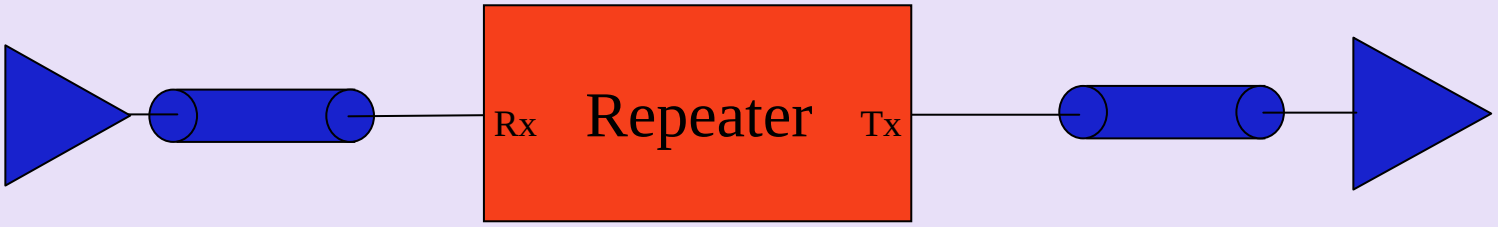
Electrical Repeaters Background

- Repeaters are common to improve signal integrity in SERDES links
- PCI Express is somewhat unique in the SERDES space
  - ✓ Rx Presence Detect Mechanism
  - ✓ Electrical IDLE support
  - ✓ Beacon Support

Electrical Repeaters
Rx Presence Detect Problem

- In-band Presence Detect
  - ✓ Transmitters detect load using Receiver's 50ohm load and link capacitance
  - ✓ Repeater will sit between Transmitter and Receiver, isolating the receiver resistive load
    - In essence becomes the "receiver" for this mechanism
  - ✓ Transmitter detects the Repeater's load and enters Polling.Compliance if Electrical Idle Exit is not received

Presence Detect with Repeater



Transmitter is Quiet in Detect State

Transmitter detects load and begins transmitting

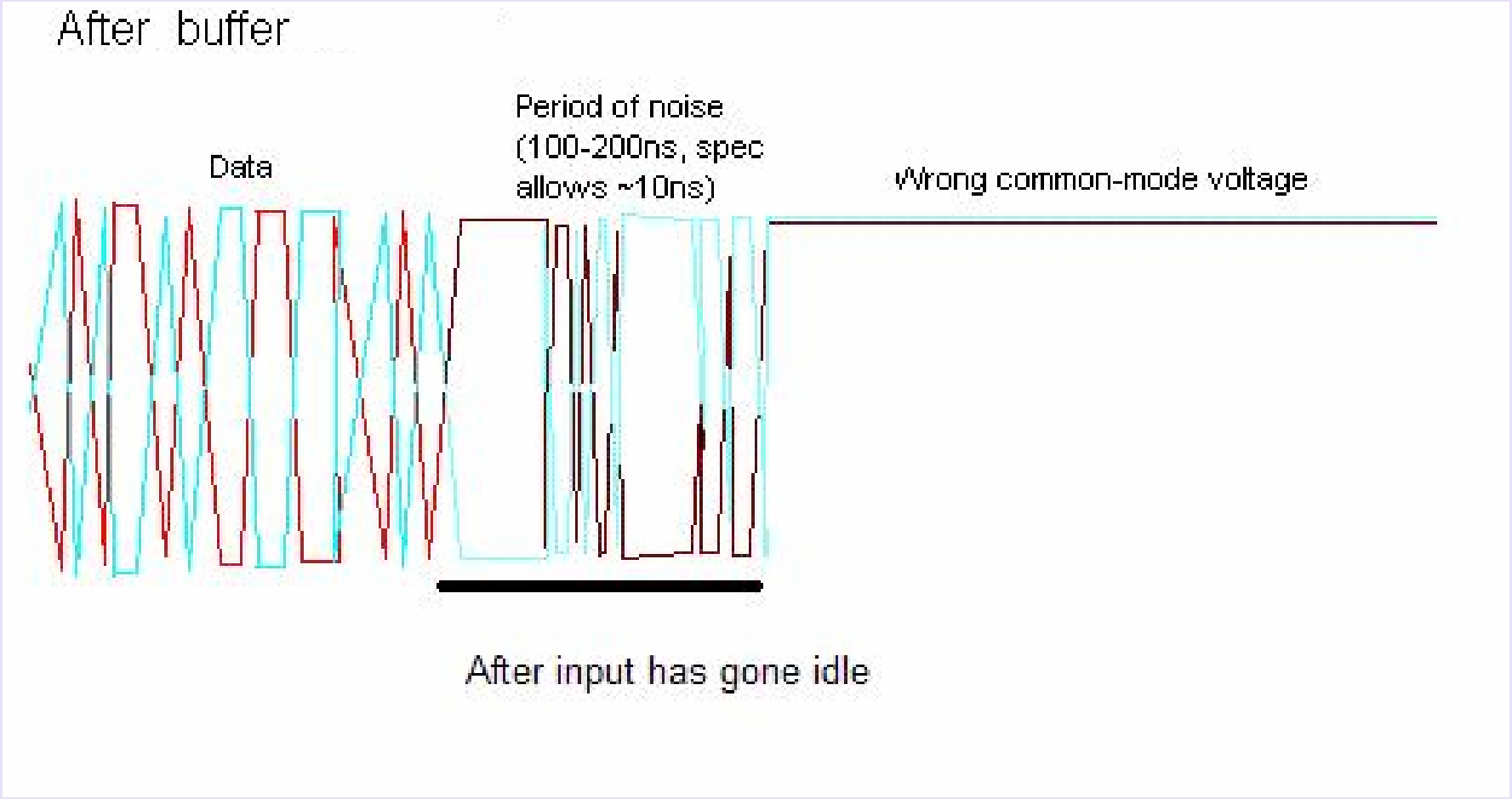
Preferred Behavior is not to begin transmitting until an actual receiver is present

Electrical Repeaters
Rx Presence Detect Recommendation

- Repeater presents Hi-Z load to link transmitter until repeater detects load on its own output
- When link receiver is detected, transition Hi-Z input to 50ohms, effectively passing the receiver detect

Electrical Repeaters Electrical Idle Problem

- Electrical Idle
  - ✓ Defined as no differential voltage across differential pair
  - ✓ Limited Common Mode Voltage change from Active States
  - ✓ Input Sensitivity
    - – Some devices detect “noise” less than 60mV Minimum detection threshold and amplify to a logic level, effectively bringing Receiver out of low power mode to search for IDLE exit sequence
  - ✓ Entry Timing
    - – EIOS is 8ns, followed by up to 8ns of transmitter active time
    - – Repeater needs it’s own time for IDLE detect and then transition to idle
    - – Receivers may start looking for valid data 20ns after receiving EIOS
      - • 16ns window for device to enter IDLE and be ready for IDLE Exit
      - • Not all receivers depend on 20ns delay, threshold detect is allowed…
  - ✓ Exit Timing
    - – Devices EXIT L0s after receiving FTS & SKPs appropriately. The number of FTS may not account for latency of repeater
  - ✓ Timings are not taken into account with Exit Latency Fields. Repeaters add a latency not accounted for!



Electrical Repeaters
Electrical Idle Recommendation

- Add squelch circuits to prevent amplification of noise
- Extremely quick response time
  - ✓ Into and Out of IDLE conditions

- At the system level, allow Latency Fields in PCIe devices to be programmed for longer delays
- Increase N_FTS for cabled links
- Worst case, disable L0s and L1 states via ASPM Control Field

Electrical Repeaters Beacon Support

- Beacon Support
  - ✓ Low frequency Wake mechanism
  - ✓ Pulse Widths range from 2ns to 16ms
  - ✓ Low output impedance
  - ✓ DC balance restoration within 32ms
- Many devices for Gigabit SERDES do not handle this low frequency signal
- Recommend Supporting Beacon Functionality.

Power Provisioning

- 3.3V Optional Power Provisioning to connector receptacle
  - ✓Keying specified for support of this option
  - ✓No cable wiring provided, power is only provided to local connector housing
  - ✓Intended use for active equalization, repeater devices and E-O Transceivers
- Implementation specifics are beyond the scope of the specification
  - ✓Voltage range is as 3.0 to 3.6V
  - ✓Maximum sustained current of 350mA for a lane, but this varies with link width
- A cabled system should provide for the power consumption, some form of current protection, and cooling at both ends

ASPM

- Advanced State Power Management
  - ✓ Power management beyond the Sleep States
  - ✓ Provide power savings at the PHY level
    - – Discreet power modes L0s, L1, and L2 providing various levels of power savings
- ASPM can add latency to a link
  - ✓ The link must power up and re-sync before communication is restored
  - ✓ FTS varies from 16ns to 4us (defined by N_FTS) for L0s Exit
  - ✓ If link does not train in the specified timeframe, link moves to Recovery
    - – If Recovery fails, then link transitions back to Detect
- Default latencies of devices may not be optimal for a cabled application

ASPM Impact to Flow Control

There have been situations in which Transmitters stall (stop sending new data)

- ✓ One direction on link is in Electrical Idle while other is active (asymmetric data transfers). Idle portion must resync before return communication occurs
- ✓ Credits may be exhausted (Receiver can not accept more data so Transmitter stops TLP advancement)
  - Too much time elapses before Link is completely active
  - Transmitter stalls awaiting ACK of previous packets
  - Retry is initiated after timeout
    - limit of 3 Retries before link enters Recovery state
- ✓ Additional communication may continue when after transaction ACK is received and credits are available, but process may repeat reducing performance

Replay Timer and Buffer

- Replay Buffer
  - ✓ Transmitters stores TLPs for Ordering and Transmission
  - ✓ Holds data until ACKnowldedgmenet of receipt
  - ✓ Transmitter stalls (stops processing new TLPs) once buffer is full
- Replay Timer
  - ✓ Timeout for receipt of ACK/NAK
  - ✓ Small timer values initiate Retries and possible Recovery

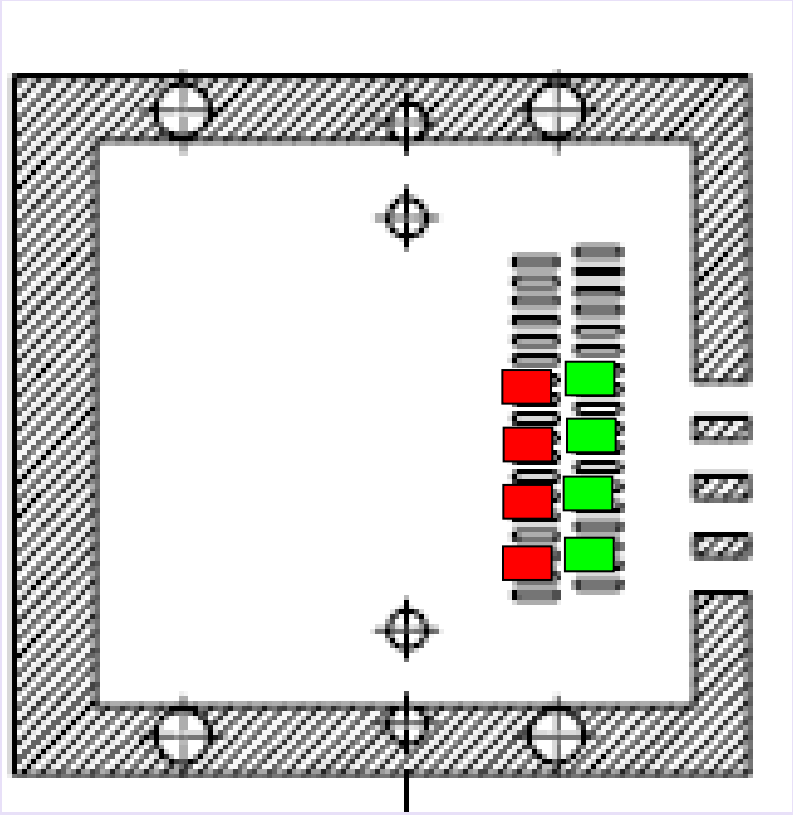
ASPM Issues Recommendations

- Increase Replay Timer
  - ✓ Allow more time for latency and transmission paths
- Add corresponding Reply Buffers
  - ✓ Prevent performance impacts due to Transmitters stalling
- Understand the impact of repeaters in your system

Routing through the footprint

- Original EMI shield for connector was optimized for microstrip escapes of RX lines for DC blocking capacitors
- Spec was later changed to move blocking caps to transmitters, but shield has not been modified
- Escape routing is slightly more difficult

x4 Connector Footprint Diagram



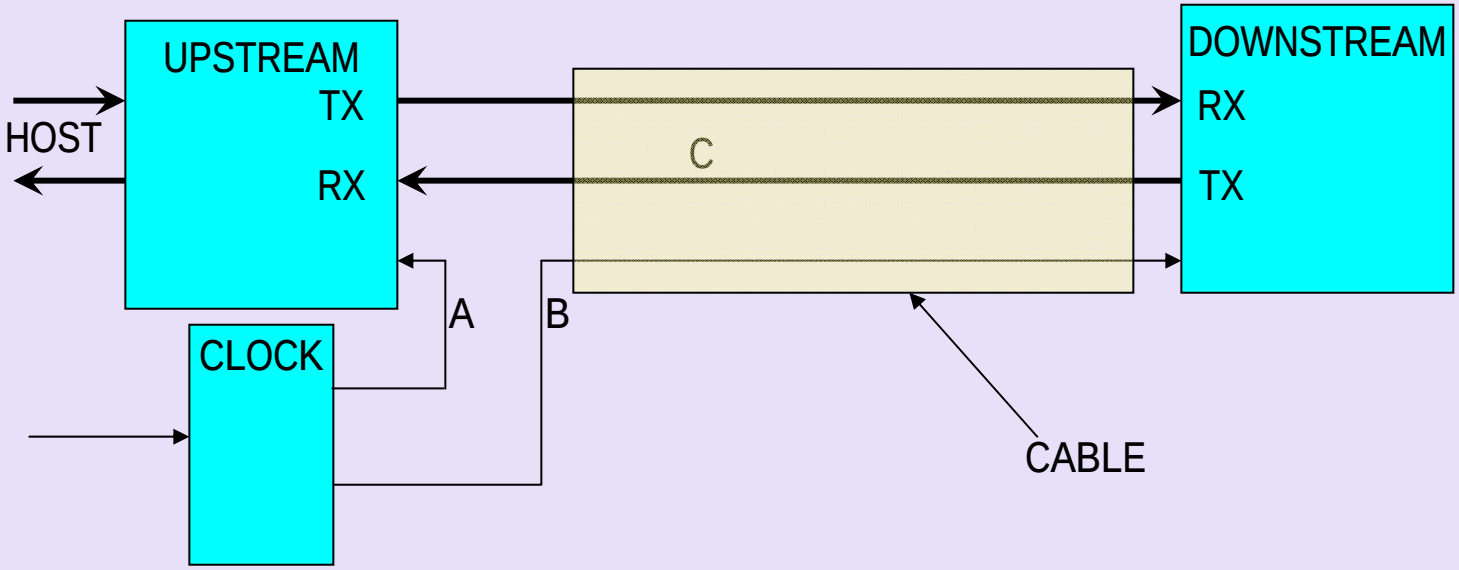
- Specification has TX pairs on the board edge side of connector.
- This can make microstrip escapes for capacitor connections non-trivial

Dealing with SSC and jitter

- PCI Express systems typically utilize a common reference clock in order to support SSC and meet the +/- 300ppm requirement.
- There are limits to the jitter phase delay of the SSC profiles for reliable operation.
  - ✓ PCIe has an approximate limit of 70ns round trip DUE TO EYE CLOSURE. This budget corresponds to the 7m cable length limitation.
- Allowing separate reference clocks for the system link and the cable link allows for longer cable distances and interoperability.
  - ✓ This requires a device to take in the local SSC system clock for the upstream link and a separate clock for the cabled link.

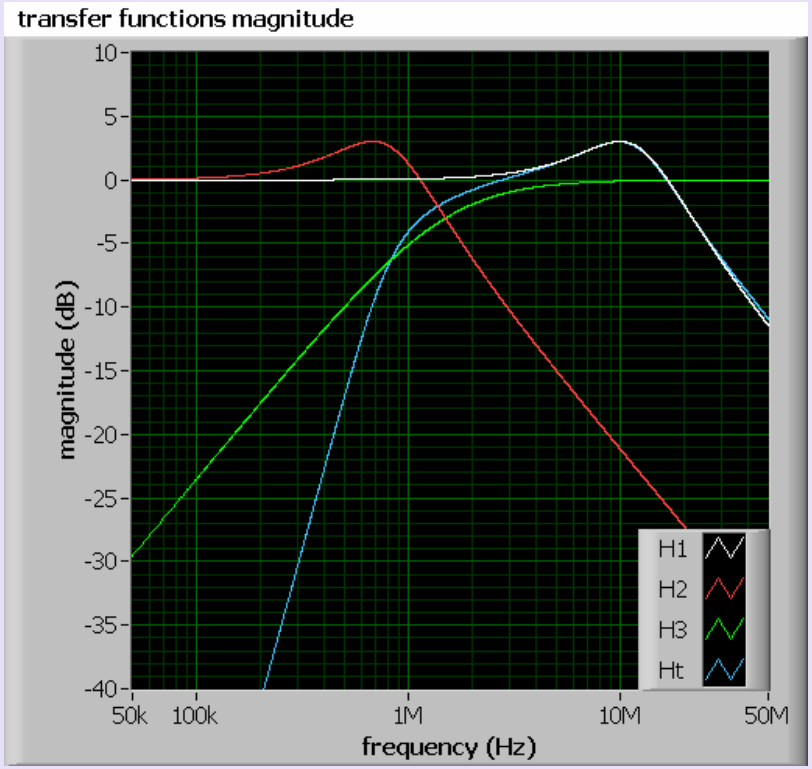
Signal Delay Calculation Evaluating Phase Jitter Impact

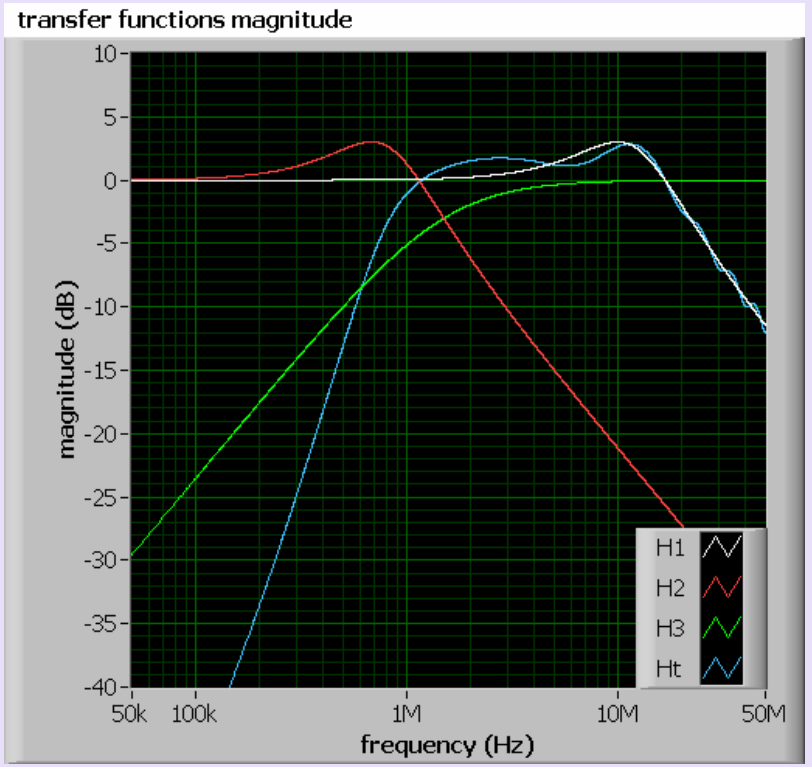
- Phase delay between upstream RX and upstream RefClk
  - ✓= (B + C – A) * Tpd = 60nsec for 7 meter cable
- Additional 10nsec budget for component delay differences
- Negative impact on “Eye Closure” transfer function



Phase Jitter Impact from Cable

- Images show impact from 10ns versus 100ns roundtrip delay
- Increase in eye closure from RefClk phase jitter components up to about 6MHz





Long Cable Considerations

- The 1.1 PCIe Base specification provides Flow Control Update Latency Guidelines (Table 2-28)
  - ✓ Recommendations are provided based on maximum payload and link size
- Silicon provides limited Retry Buffer storage and Credit capability
  - ✓ More of a concern with small packets with increased roundtrip delay due to ACK/NAK latency (Table 3-5)

Long Cable Cons. (cont.)

- The 1.1 PCIe Base specification provides Replay Timer Guidelines (Table 3-4)
  - ✓ Timer is reset any time an ACK/NAK DLLP is received
  - ✓ Direct impact from roundtrip delay
  - ✓ Potential repeat of TLP that has been received
- Can affect bandwidth due to unnecessary replay
- Can confuse state-machines due to unexpected replay
- Understand how cable delay may impact performance by understanding Flow Control updates, ACK/NAK Latency, and Retry Buffer sizing

Further User Experience Considerations

- Connection of two upstream subsystems
  - ✓Power domain isolation no longer present
  - ✓CPRSNT# Always inactive
    - –CREFCLK always disabled
  - ✓CPWRON connected together
    - –High impedance path, not considered an issue
  - ✓CWAKE# Always inactive
  - ✓CPERST# Output Buffers Shorted together
    - –Requires tolerant output buffers
- Connection of two downstream subsystems
  - ✓All isolation circuits inactive due to lack of current source
  - ✓CREFCLK not provided by downstream subsystem

Recommendations for Implementation Over a Cable

- Verify PCI Enumeration
- Understand your system Hot-Plug behavior
- Choose repeaters carefully, if deemed necessary for implementation
- Understand Latencies and their impact to performance with desired cable lengths
  - ✓ IDLEs
  - ✓ ACK/NAK
  - ✓ Timeouts
- Choose devices with larger Replay Timers and Replay Buffers
- Separate clock domains for long cables
- Choose high performance clock sources

Industry Help Needed

- Clock domain separation devices are needed
- Increasing Replay Buffer sizing and Replay Timeouts to account for increased latencies
- Allow Latencies to be programmed at device initialization
- Develop PCIe aware repeaters
- System BIOSes should fully enumerate PCI bus chain and allocate sufficient resources Hot Plug

Thank you for attending the
PCI-SIG Developers Conference 2007.

For more information please go to
www.pcisig.com

Cabled PCI Express – Implementation Considerations

Lee Mohrmann

R&D Engineer

National Instruments