PCIe® 2.x vs. PCIe 1.x

Joe Cowan
Computer Systems Architect
Hewlett-Packard Company

Today's Topics

- PCIe® 2.0 Electrical & Protocol Changes Over 1.x
- PCIe 2.0 Errata Items Summary
- PCIe 2.1 Protocol Extensions Summary

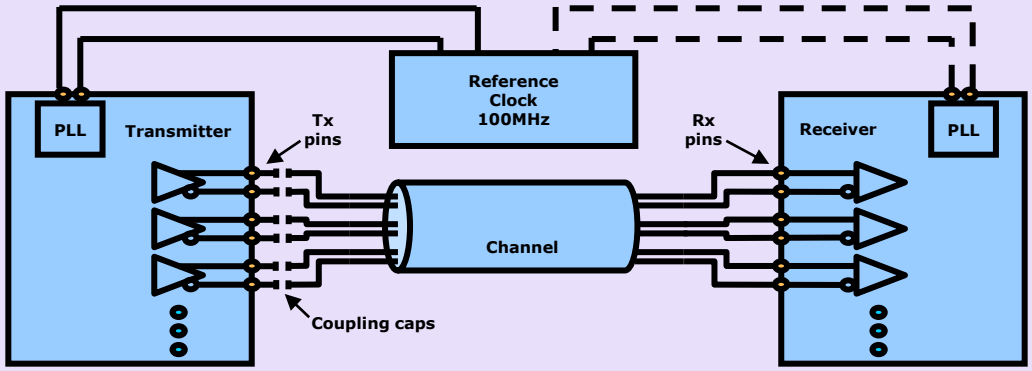
PCIe 2.0 Electrical & Protocol Changes Over 1.x

PCIe 2.0 Changes

- 5 GT/s Signaling Speed
  - ✓ LTSSM Extensions
  - ✓ Link Speed Management & Controls
- Link Bandwidth Notification Mechanism
- Function Level Reset (FLR)
- Access Control Services (ACS)
- Completion Timeout Control
- Retracted: Trusted Configuration Space (TCS)

5 GT/s Signaling Speed

- Tight budgets remove all guard bands!
- All interconnect components specified for enhanced interoperability
  - PCIe 1.0a: Transmitter
  - PCIe 1.1: PCIe 1.0a + Reference Clock
  - PCIe 2.0: PCIe 1.1 + Channel + Receiver
- CEM specification provides electrical interoperability for system board/add-in card



PCIe 2.0 LTSSM Extensions

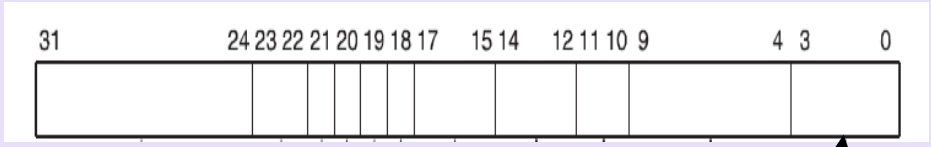
Extension	Explanation	Benefits
Speed Negotiation	Capability to upgrade or downgrade Link speed	RAS (improved Link uptime), dynamic Link speed optimization, power savings (25%+)
Compliance Speed	Programmable as well as in-band mechanism to select compliance pattern speed	Flexibility to perform compliance testing at multiple speeds with low cost
Electrical Idle Entry and Exit	Protocol changes to facilitate circuit design	Enhanced robustness, yield, power savings, ease of design (TTM)
Link Width Up-configure	Capability to increase the Link width up to the initial trained Link width	Power savings
Compliance Entry/Exit	Device Configuration despite Link failures	High availability, enhanced robustness

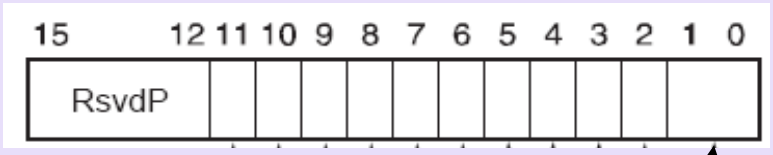
Link Speed Management

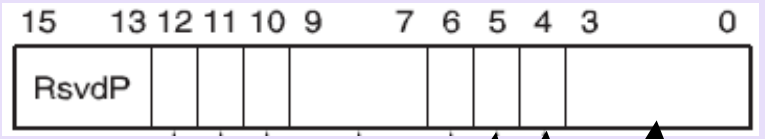
- Default: trains to the greatest common speed
  - ✓ Software can set an upper bound on the speed
  - ✓ Hardware can limit speed for Link reliability
- Hardware is permitted to change the speed autonomously
  - ✓ E.g.: power management
  - ✓ Software can disable
- New mechanism supporting software control for entering/exiting Compliance Mode

Link Speed Controls

- Link Capability register
  - ✓ *Maximum Link Speed* field renamed to *Supported Link Speeds*
- Link Status register
  - ✓ *Link Speed* field renamed to *Current Link Speed*
- (New) Link Control 2 register
  - ✓ *Hardware Autonomous Speed Disable* bit
  - ✓ *Enter Compliance* bit
  - ✓ *Target Link Speed* field





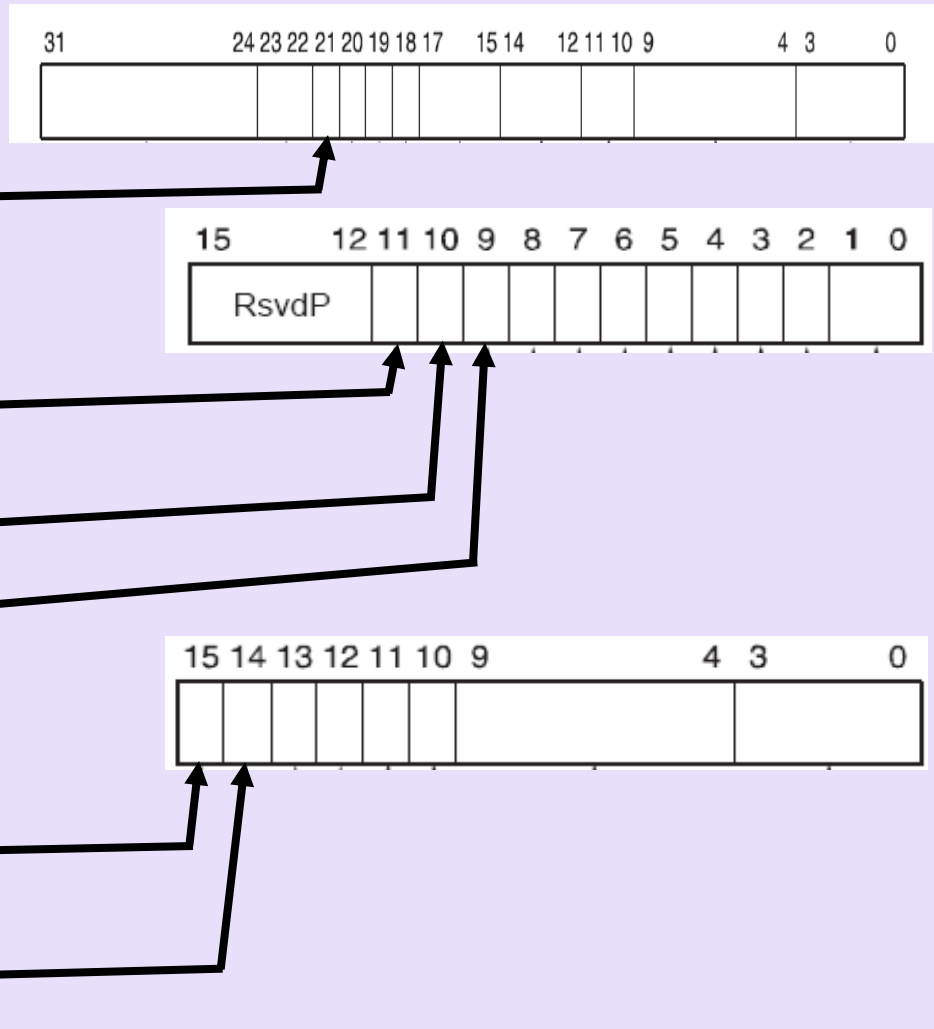


Bandwidth Notification

- Mechanism for PCIe-aware software to be notified when Link bandwidth changes
  - ✓ E.g.: Link retrains to a lower bandwidth due to reliability problem
  - ✓ E.g.: hardware-autonomous Link retraining
- Logically coupled with Link Speed Management
- Required for all Root Ports and Switch Downstream Ports that support wider than x1 and/or multiple Link speeds

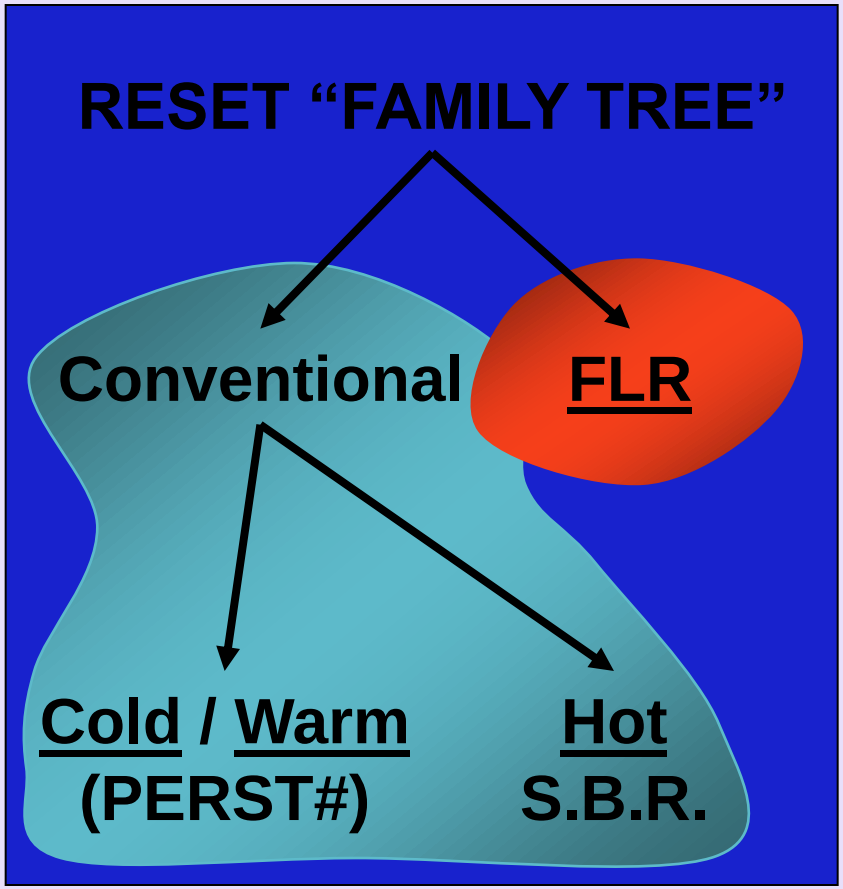
BW Notification Mechanism

- Link Capability register
  - ✓ *Link Bandwidth Notification Capability* bit
- Link Control register
  - ✓ *Link Autonomous Bandwidth Interrupt Enable* bit
  - ✓ *Link Bandwidth Management Interrupt Enable* bit
  - ✓ *Hardware Autonomous Width Disable* bit
- Link Status register
  - ✓ *Link Autonomous Bandwidth Status* bit
  - ✓ *Link Bandwidth Management Status* bit



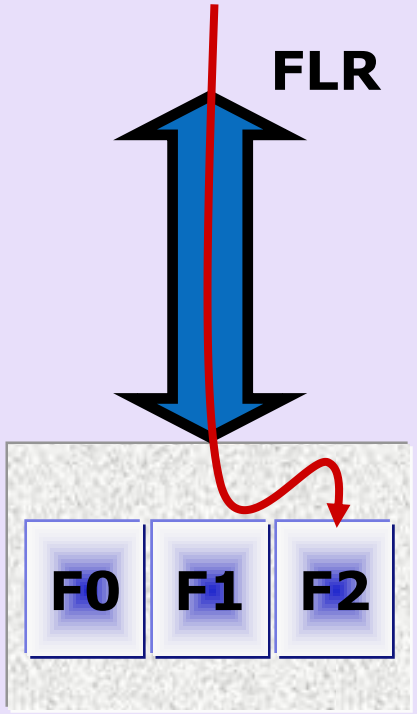
Function Level Reset

- New type of reset
  - ✓ Existing resets may (but are not required to) reset Function internals
  - ✓ FLR definition requires Function internal reset
- SW initiated Function-specific reset



FLR Details

- Endpoints only
  - ✓ All types: legacy, native, integrated
- Register interface is simple
- Implementation & effects are potentially complex
  - ✓ Resets internal Function-specific state
- Not all architected registers are reset
  - ✓ Hardware initialized (HwInit), BIOS set, etc.



Access Control Services

- For Downstream Ports and multi-Function devices
- New Extended Capability & Status/Mask/Severity bits in AER
- Source validation
  - ✓ Downstream Ports range check Requester ID BusNum in upstream Request TLPs
- Peer-to-peer controls
  - ✓ Determine whether to forward directly, block, or redirect peer-to-peer requests to the RC for access validation
- Address Translation Services (ATS) controls
  - ✓ Blocking of Requests with translated addresses
  - ✓ Direct P2P of Requests with translated addresses

Completion Timeout

- Required: Architected Disable Bit
  - ✓ "Turns off" timeout
  - ✓ Not to be used in normal operation
- Optional: Completion Timeout Programmability
  - ✓ Devices indicate supported ranges from the four bins defined
  - ✓ Two selectable ranges for each bin

PCIe 2.0 Errata Items Summary

PCIe 2.0 Errata Released July 2008

- Spec errors, clarifications, typos, terminology and formatting grouped into three bins
  - ✓ Bin A: “non-straight-forward changes” – 57
  - ✓ Bin B: “straight-forward changes” – 35
  - ✓ Bin C: “trivial editorial changes” – 38
- 130 total

PCIe 2.0 Errata Highlights

- Ordering Rules Summary Table and Descriptions overhauled to create a more suitable base for AtomicOps & ID-Based Ordering ECRs
  - ✓ Errata do not change ordering semantics, but subsequent ECRs do
- Several misc clarifications on 5.0 GT/s vs 2.5 GT/s operation
- Several misc clarifications on de-emphasis operation
- Several misc clarifications on cross-Link operation
- Various TLP Header fields: Reserved vs “required to be 0”
  - ✓ Generally prefer to keep fields Reserved to avoid squandering bits
  - ✓ A few cases left “required to be 0” in case some Receivers implemented checks, even though such checks were never intended
- Link Capability bits having same value in all Functions of MFD associated with an Upstream Port
  - ✓ Link Cap bits in Upstream Port Functions of MFD all refer to same Link
  - ✓ Link Cap bits in Downstream Port Functions of MFD refer to different Links
- ACS Violation bits in AER not required to be implemented in AER unless ACS is implemented
- Earlier VC Enable bit clarifications for VC Cap struct carried over to same bit in MFVC Cap struct

Ordering Rules Summary Table – Key Problems

- Not all components can distinguish Read Completions from I/O & Configuration Write Completions
- Some row/column labels enumerate individual transactions, causing maintenance problems as new transactions are defined, e.g., AtomicOps
- Associated descriptions are excessively wordy and inconsistent in style

Row Pass Column?		Posted Request	Non-Posted Request		Completion	
		Memory Write or Message Posted Request (Col 2)	Read Request (Col 3)	I/O or Configuration Write Request NPR with Data (Col 4)	Read Completion (Col 5)	I/O or Configuration Write Completion (Col 6)
Posted Request	Memory Write or Message Posted Request (Row A)	a) No b) Y/N	Yes	Yes	a) Y/N b) Yes	a) Y/N b) Yes
Non-Posted Request	Read Request (Row B)	No	Y/N	Y/N	Y/N	Y/N
	I/O or Configuration Write Request NPR with Data (Row C)	No	Y/N	Y/N	Y/N	Y/N
Completion	Read Completion (Row D)	a) No b) Y/N	Yes	Yes	a) Y/N b) No	Y/N
	I/O or Configuration Write Completion (Row E)	Y/N	Yes	Yes	Y/N	Y/N

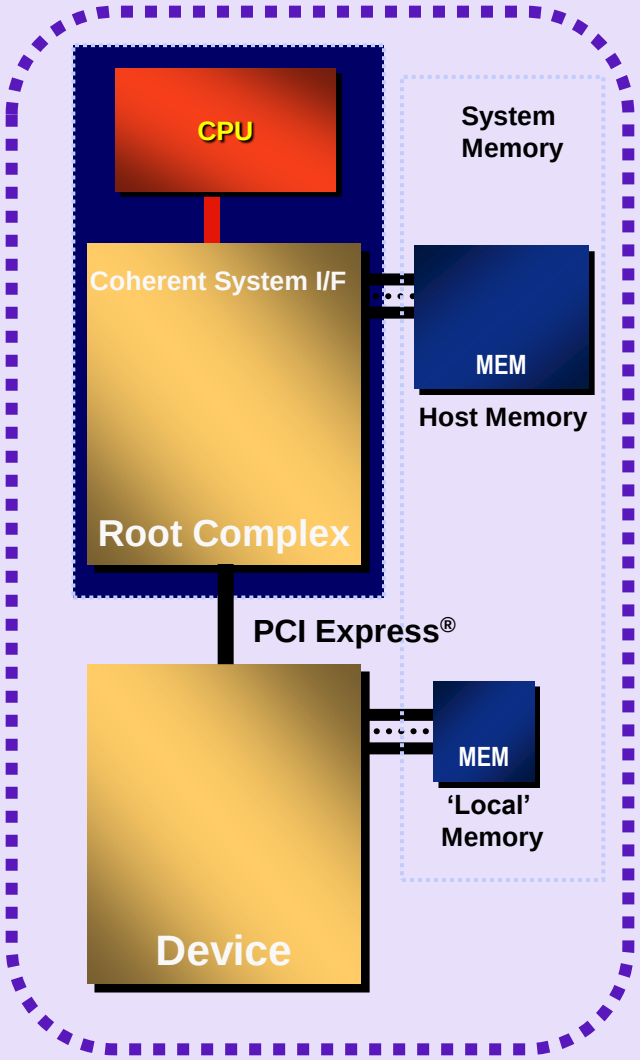
Ordering Rules Summary Table – End Result

Row Pass Column?		Posted Request (Col 2)	Non-Posted Request		Completion (Col 5)
			Read Request (Col 3)	NPR with Data (Col 4)	
Posted Request (Row A)		a) No b) Y/N	Yes	Yes	a) Y/N b) Yes
Non-Posted Request	**Read Request** (Row B)	No	Y/N	Y/N	Y/N
	NPR with Data (Row C)	No	Y/N	Y/N	Y/N
Completion (Row D)		a) No b) Y/N	Yes	Yes	a) Y/N b) No

- No longer separate rows/columns for Read Completions vs Config/IO Write Completions
- Row/column labels now consistently based on classes of transactions instead of enumerated individual transactions
- Associated descriptions now more consistently and concisely worded

PCIe 2.1 Protocol Extensions Summary

PCIe Protocol Extensions



- Performance Improvements
  - ✓ **TLP Processing Hints** – hints to optimize system resources and performance
  - ✓ **TLP Prefix** – mechanism to extend TLP headers for TLP Processing Hints, MR-IOV, and future extensions
  - ✓ **ID-Based Ordering** – Transaction-level attribute/hint to optimize ordering within RC and memory subsystem
  - ✓ **Extended Tag Enable Default** – permits default for Extended Tag Enable bit to be Function-specific
- Software Model Improvements
  - ✓ **Atomic Operations** – new atomic transactions to reduce synchronization overhead
  - ✓ ***Page Request Interface*** *– mech in ATS 1.1 for a device to request faulted pages to be made available (not covered)*
- Communication Model Enhancements
  - ✓ **Multicast** – mechanism to transfer common data or commands sent from one source to multiple recipients
- Power Management
  - ✓ **Dynamic Power Allocation** – support for dynamic power operational modes through standard configuration mech
  - ✓ **Latency Tolerance Reporting** – Endpoints report service latency requirements for improved platform power mgmt
  - ✓ **Optimized Buffer Flush/Fill** – Mechs for devices to align DMA activity for improved platform power mgmt
- Configuration Enhancements
  - ✓ **Alternative Routing-ID Interpretation** (ARI) – mechanism for a device to support up to 256 Functions
  - ✓ **Resizable BAR** – Mechanism to support BAR size negotiation
  - ✓ **Internal Error Reporting** – Extend AER to report component internal errors and record multiple error logs

Protocol Extensions Summary

Extension	Description	Status
Alternative RID Interpretation (ARI)	Enables a device to support up to 256 Functions	PCIe 2.1 spec
Atomic Operations (AtomicOps)	Atomic Read-Modify-Write mechanism	PCIe 2.1 spec
Internal Error Reporting	Extend AER to report component internal errors and record multiple error logs	PCIe 2.1 spec
Resizable BAR	Mechanism to support BAR size negotiation	PCIe 2.1 spec
Multicast	Address-Based Multicast of Posted Request TLPs	PCIe 2.1 spec
ID-Based Ordering (IDO)	New type of relaxed ordering semantics to improve performance	PCIe 2.1 spec
Dynamic Power Allocation (DPA)	Dynamic power mgmt for substates of D0 (active state)	PCIe 2.1 spec
Latency Tolerance Reporting (LTR)	Endpoints report service latency requirements, enabling improved platform power mgmt	PCIe 2.1 spec
Extended Tag Enable Default	Permits default for Extended Tag Enable bit to be Function-specific instead of 0b	PCIe 2.1 spec
TLP Processing Hints (TPH)	Hints for optimized TLP processing within host memory/cache hierarchy	PCIe 2.1 spec
Optimized Buffer Flush/Fill (OBFF)	Mechanisms for devices to align DMA activity for improved platform power mgmt	ECN Finalized 5/2009
TLP Prefix	Mechanism to extend TLP headers	PCIe 2.1 spec

Thank you for attending the
PCI-SIG Developers Conference 2010.

For more information please go to
www.pcisig.com