June, 2010

PCI Express® Gen 2 Deep Dive on Power Architecture® Based Products

FTF-NET-F0685

Richard Nie

Sr. Systems and Application Engineer, NMG, NSD

freescale™ semiconductor

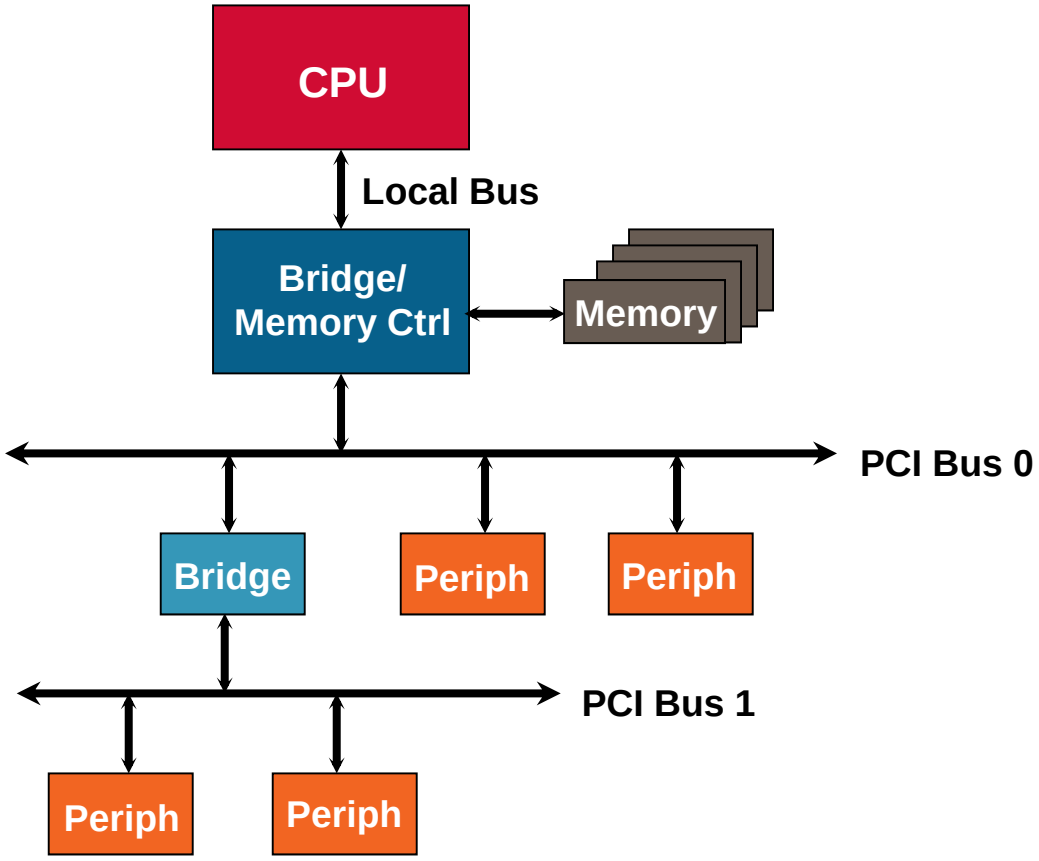
Agenda

- Legacy PCI and PCI Express Technology Overview
- PCI Express Gen2 vs. Gen1 Highlights
- PCI Express Gen2 on Freescale P4080 and P4040

freescale semiconductor

The Legacy PCI Bus Quick Overview

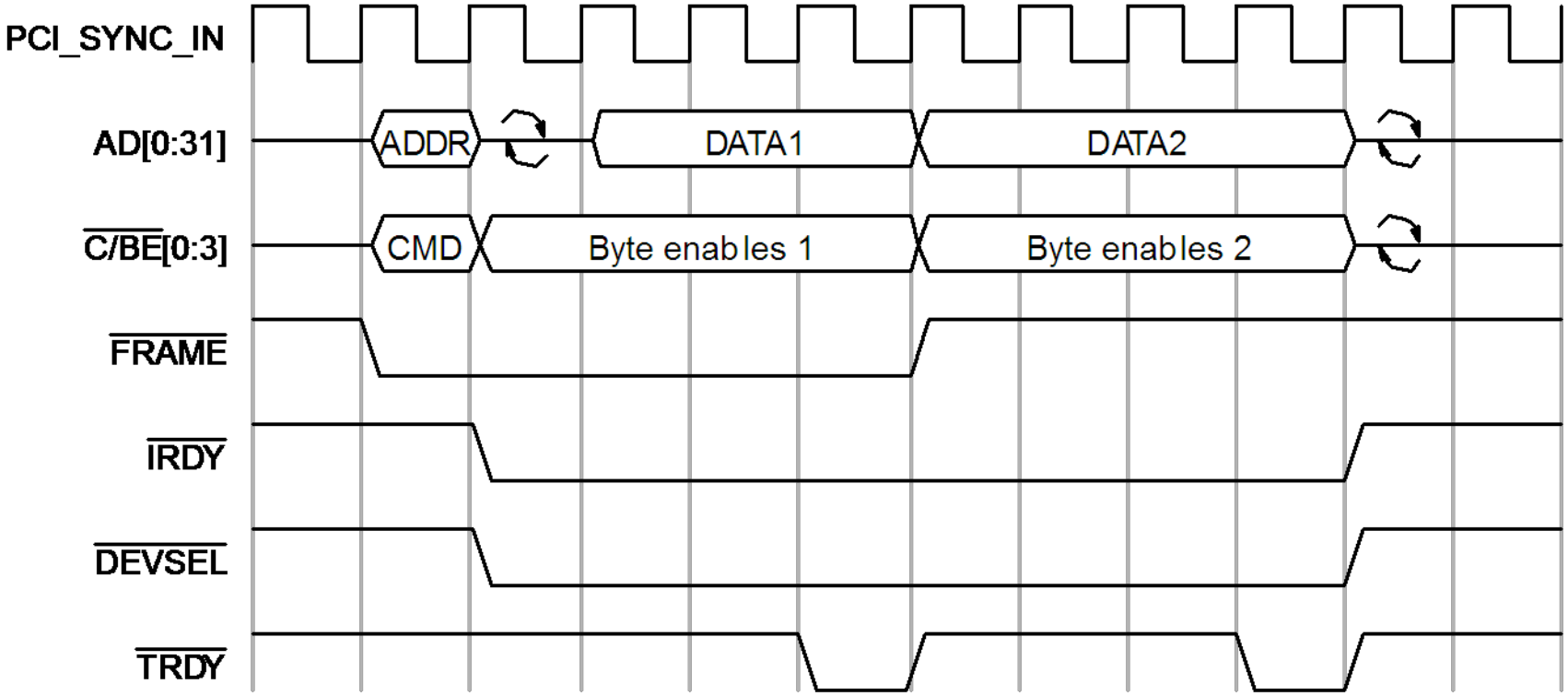
- Parallel shared bus
  - 32-/64-bit multiplexed bus
  - Defined in early 1990s (Rev.2.0 Spec released in 1993)



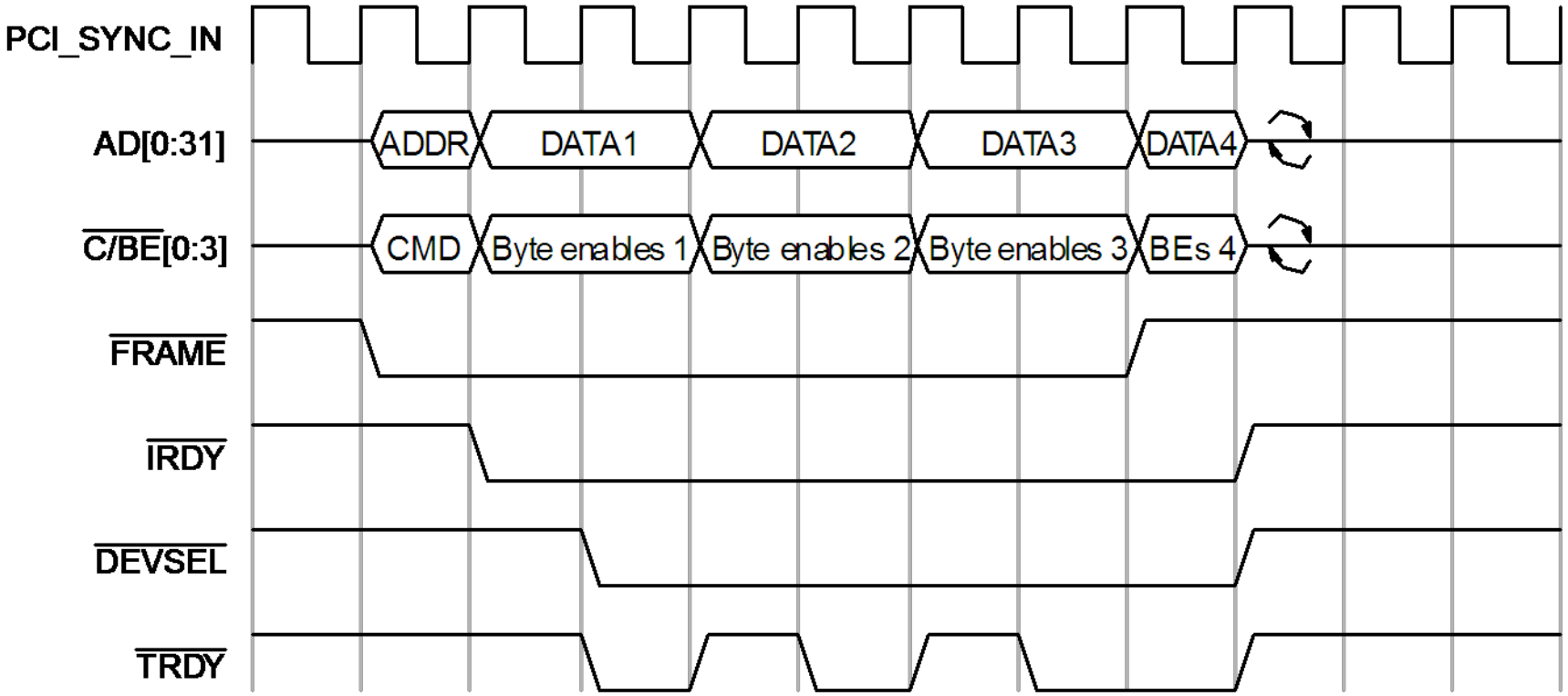
PCI Bus Commands

C/BE#[3:0]	PCI Bus Command
0000	Interrupt Acknowledge
0001	Special Cycle
0010	I/O Read
0011	I/O Write
0110	Memory Read
0111	Memory Write
1010	Configuration Read
1011	Configuration Write
1100	Memory Read Multiple
1101	Dual Address Cycle
1110	Memory Read Line
1111	Memory Write and Invalidate

PCI Read Transaction



PCI Write Transaction



PCI Bus Transaction Termination

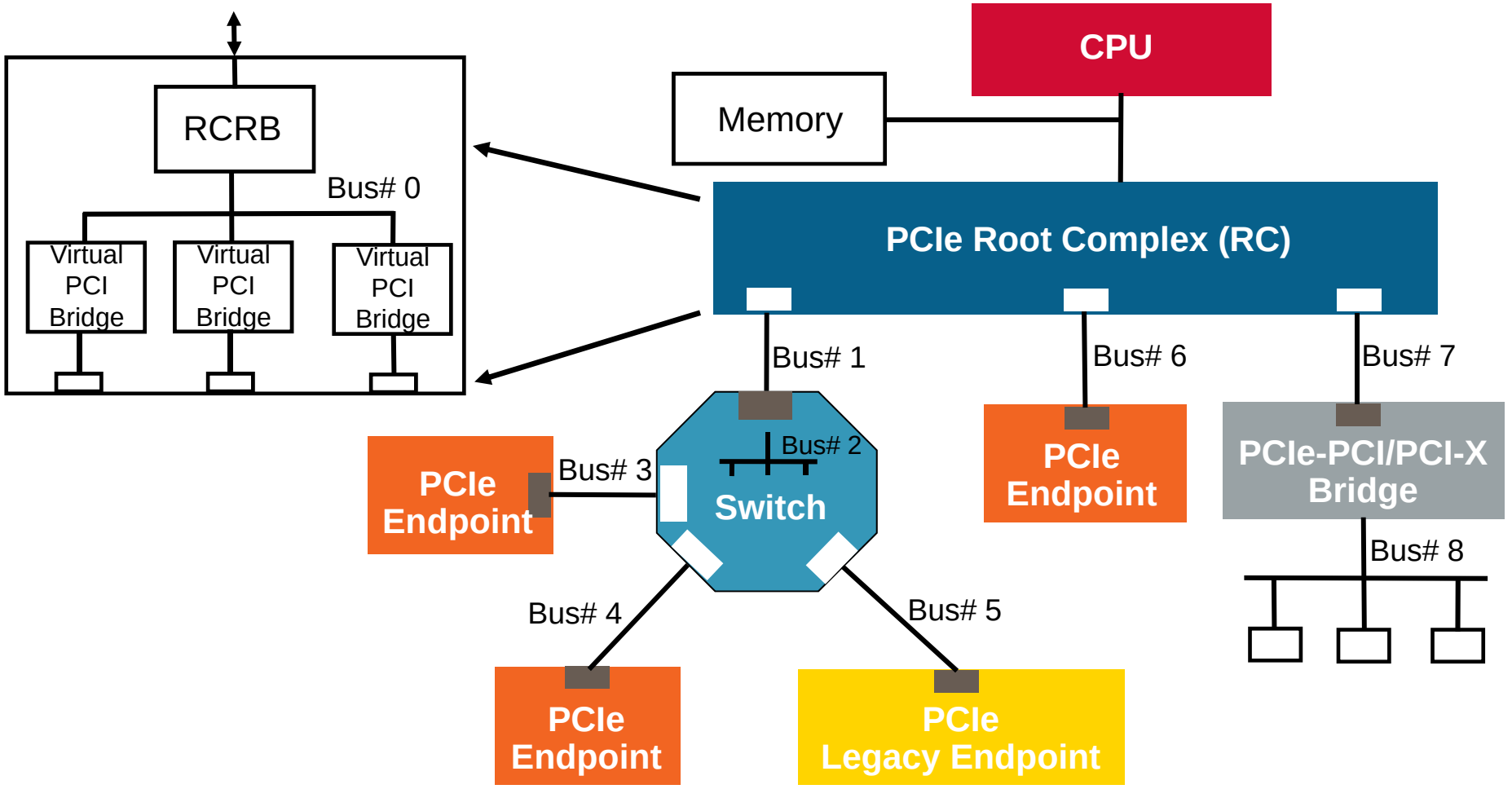
- Master-initiated
  - Completion—normal termination
  - Timeout—controlled by latency timer
  - Master-abort—no target device claimed the transaction

- Target-initiated
  - Disconnect—target is temporarily unable to continue the transaction
    - With or without data transfer
  - Retry—target is temporarily unable to begin the transaction
  - Target-abort—target will never be able to service the transaction
    - Fatal error
    - Target does not want the transaction retried

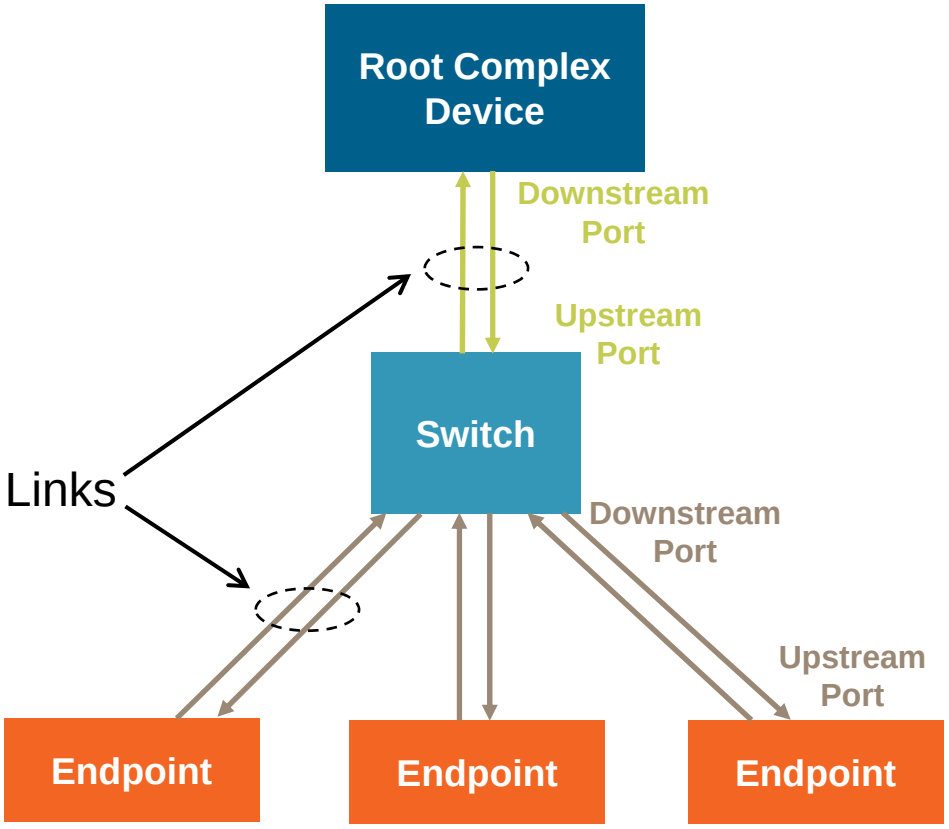
Limitation of PCI Architecture

- Slow device arbitration consumed bandwidth due to wait state insertion
- Transfer size unknown
- Delayed transactions are inefficient
  - Retries use up bus time
  - No transfer count
  - Bus master doesn't identify itself
- Interrupt handling is inefficient
  - Since INT# is wired OR together, it takes quite a lot of time to figure out which device caused interrupt
- Slow clock speed
  - The reflection nature makes it slow, reaching the ceiling limit of its bandwidth
  - To increase performance, increase speed to 66 MHz, which lowers load to 1 to 2 slots
  - If you want to connect more devices, you would see more and more hub-link interface (most likely 64-bit width) on the system ... causing board routing: a headache.
- Detected error results in system shutdown

PCI Express Architecture Overview

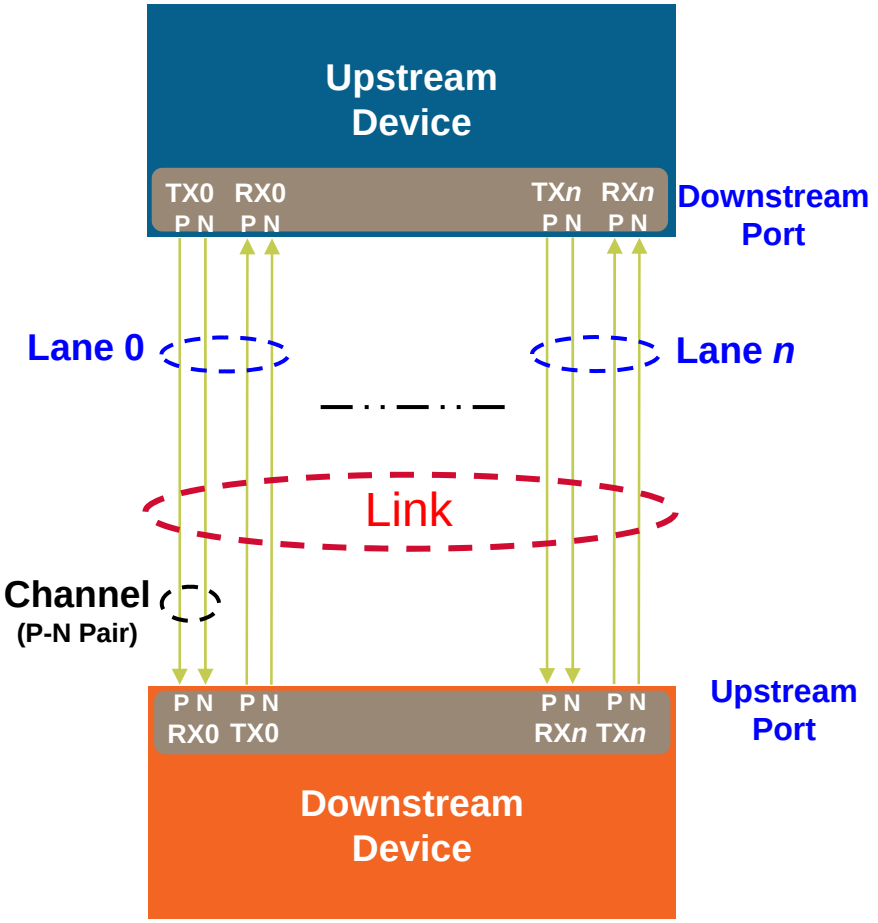


PCI Express Topology-Related Terminology



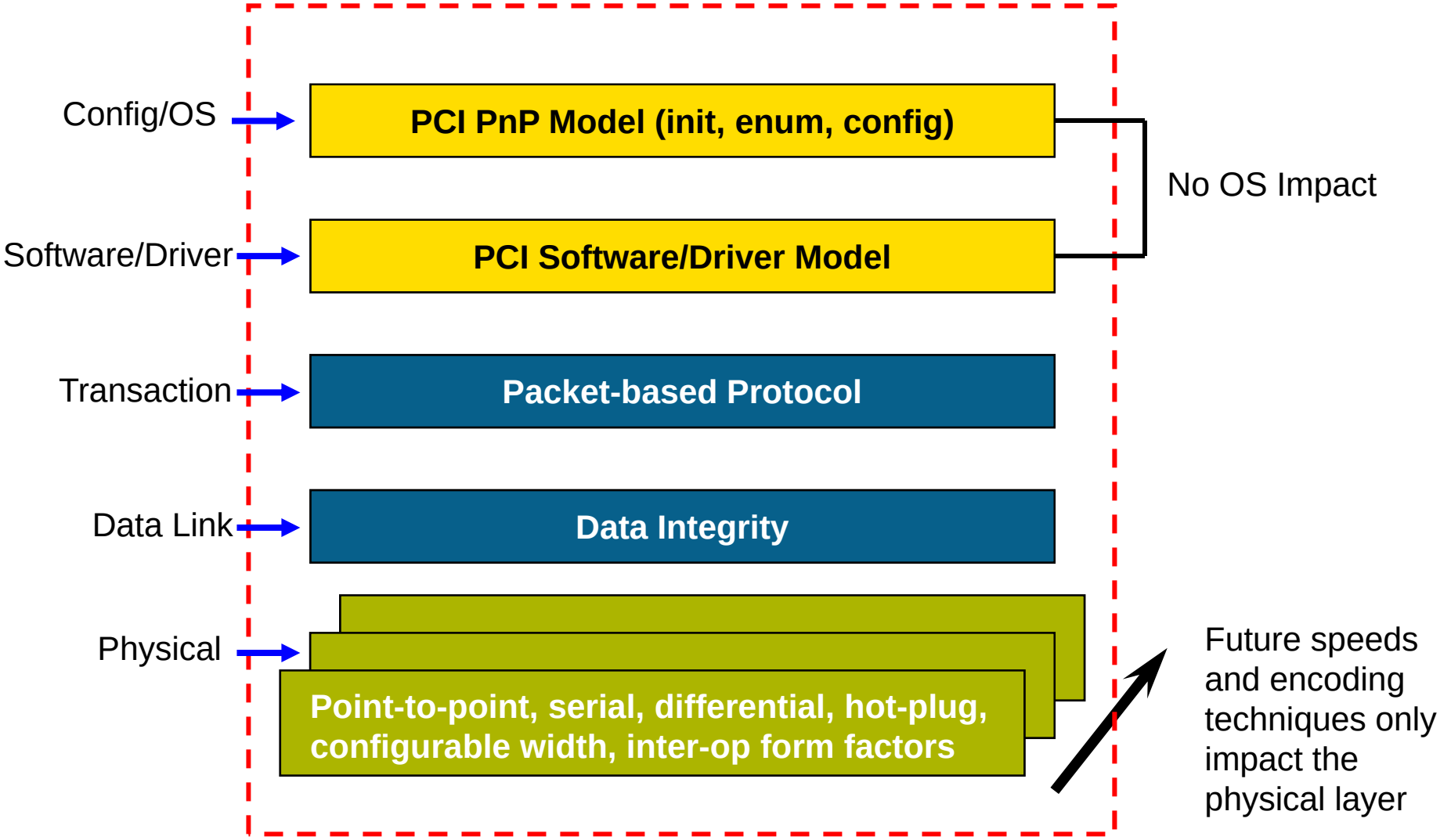
PCI Express Physical Layer Terminology

- Link
  - Collection of two ports and their interconnecting ***lanes***
- Lane
  - A set of differential signal pairs: one pair for Tx and another for Rx.
- Port
  - Physically, ***a group of transmitters and receivers*** located on the ***same chip*** that ***define a link***
  - Logically, an interface between a component and a PCI Express Link
- x1, x2, x4, x8, x16, ….. xN (Link)
  - ***A by-N link is composed of N lanes***



freescale semiconductor

PCI Express – Layered Architecture (Software and Hardware)

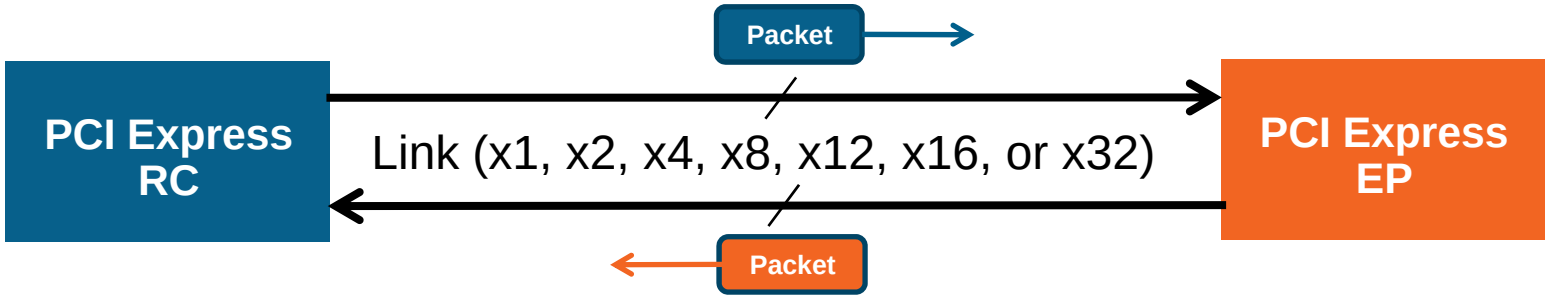


freescale semiconductor

PCI Express Data Transmission Model

- Point-to-point connection
- Serial bus significantly reduces number of pins →requires only 4 pins for a x1 link
- Scalable: x1, x2, x4, x8, x16, x32
- Dual simplex connection
- 2.5 Gbits/s per direction (***Gen 1***)
- Packet based transaction protocol

Link Width	**x1**	**x2**	**x4**	**x8**	**x12**	**x16**	**x32**
Aggregate ***Raw*** Bandwidth (Gbits/sec) ***- Pair(s) of lane (Rx & Tx)***	5	10	20	40	60	80	160



freescale semiconductor

PCI Express Architecture Highlights

- Extension of PCI architecture for PCs and servers
  - Maintains BIOS-level compatibility with PCI
- Load-store architecture
  - 128-4096 bytes maximum payload size
- Layered protocol divided into three layers
  - Transaction (TL)
    - Fulfill the data transaction goal of many types: (Memory, configure, I/O) read/write
  - Data link (DLL)
    - Ensures data integrity with the ack/nack, retry, flow control DLLPs
  - Physical (PL)
    - Establish and maintain the link

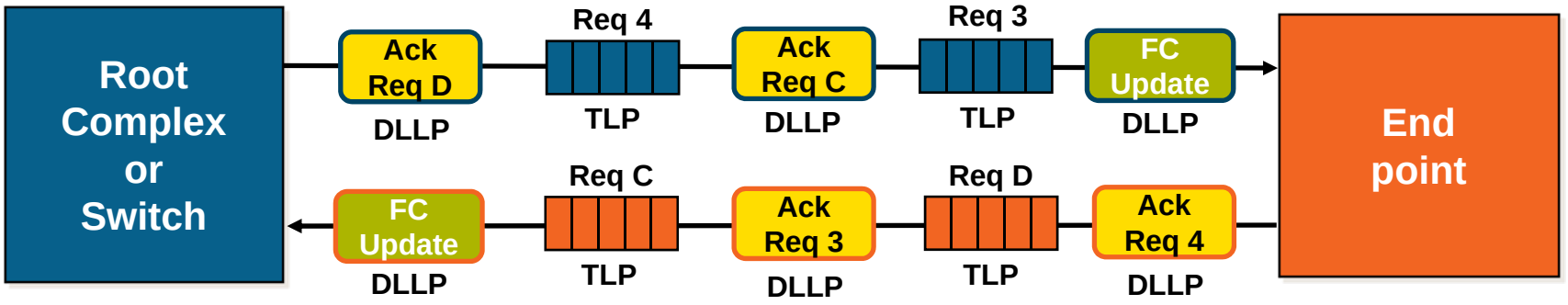
PCI Express Architecture Highlights (Cont.)

- Serial differential interface
  - 2.5 Gbits/sec per lane (Gen 1)
  - Scalable width: x1, x2, x4, x8, x12, x16, x32
  - Embedded clocking; 8B/10B encoding
  - Lane reversal and polarity inversion

- New features over PCI
  - Message transactions
  - Configuration address space extended from 256B to 4KB
  - Improved error handling and data transfer robustness (LCRC, ECRC)
  - Power management and hot plug/swap support
  - QoS support

Transaction Layer Highlights

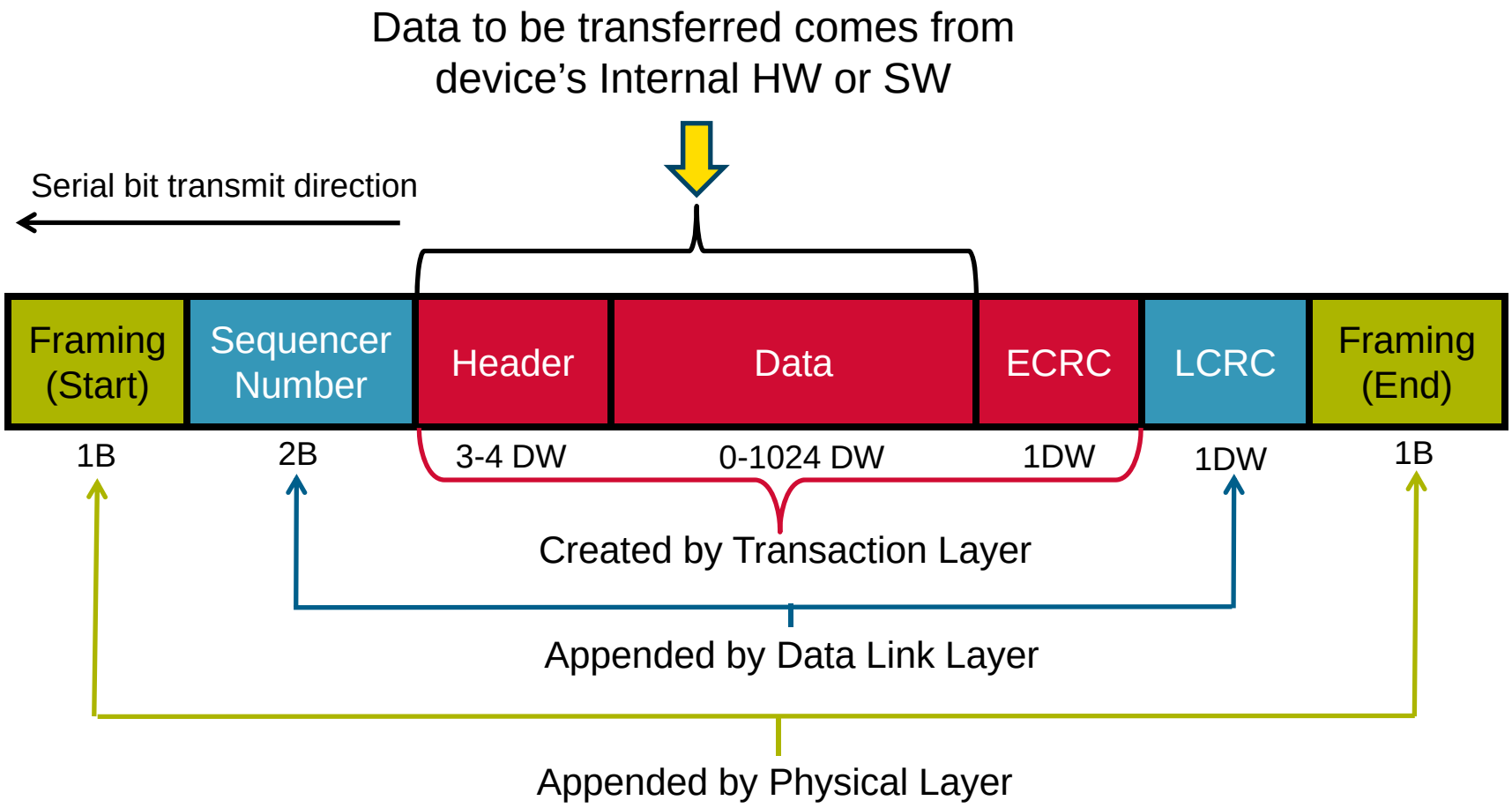
- Packet-based split transaction protocol
- Provides R/W logical transactions to software
- 4 basic transaction types: memory, I/O, configuration and message
- 32-bit and 64-bit memory addressing
- Three routing methods
  - Address routing (memory and I/O)
  - ID routing (configuration)
  - Implicit routing (messages)
- Transactions are carried by Transaction Layer Packets (TLPs)



PCI Express Transaction Types

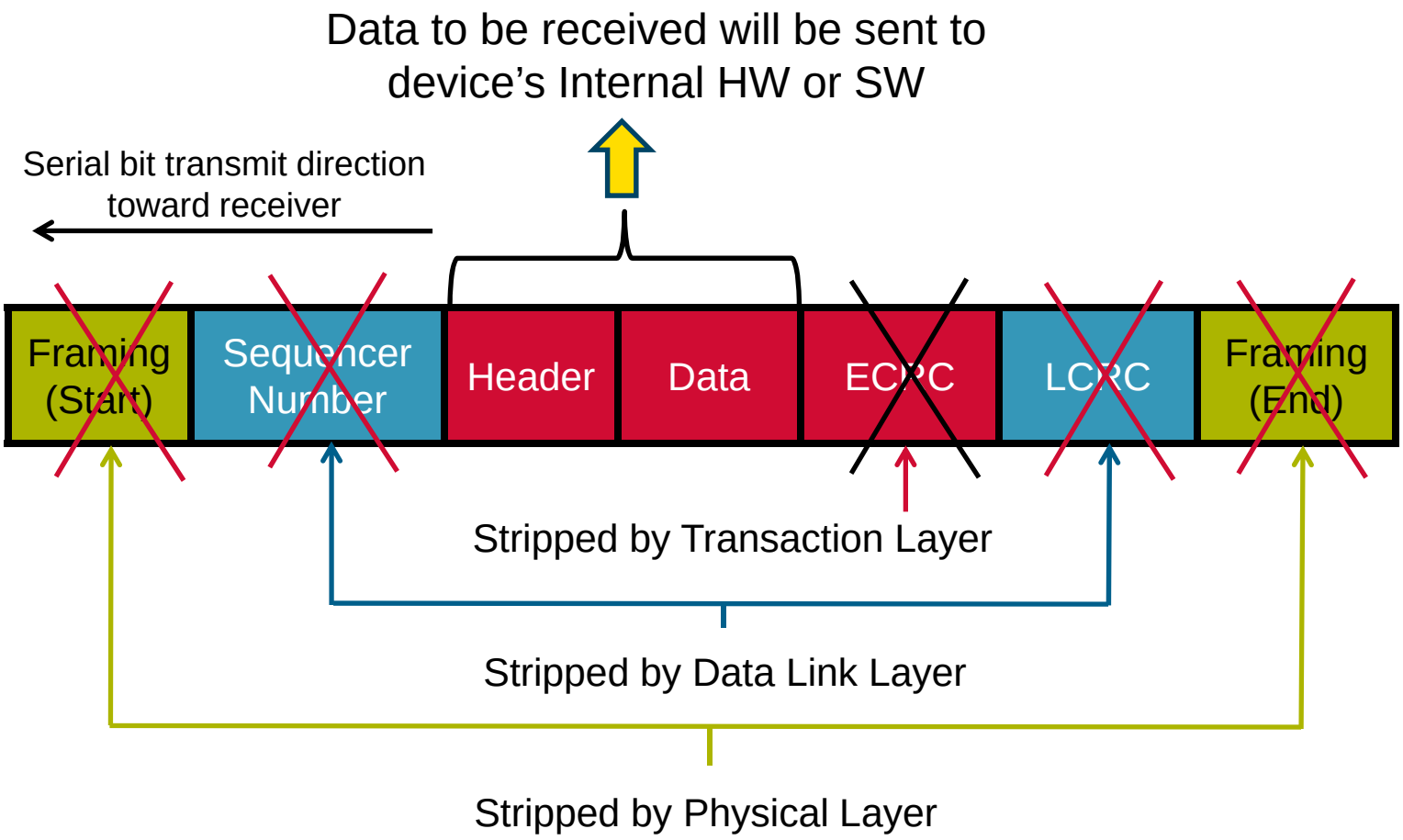
Transaction			Completion	
Request Type	Request TLP	Non-Posted or Posted	Required	Packet Type
Memory Read	MRd	Non-Posted	Yes	CplD, Cpl (error)
Memory Write	MWr	Posted	NO	
Memory Read Lock	MRdLk	Non-Posted	Yes	CplD, Cpl (error)
IO Read	IORd	Non-Posted	Yes	CplD, Cpl (error)
IO Write	IOWr	Non-Posted	Yes	Cpl
Configuration Read	CfgRd0, CfgRd1	Non-Posted	Yes	CplD, Cpl (error)
Configuration Write	CfgWr0, CfgWr1	Non-Posted	Yes	Cpl
Message w/o Data	Msg	Posted		
Message w/Data	MsgD	Posted		

PCI Express TLP (Transaction Layer Packet) Assembly

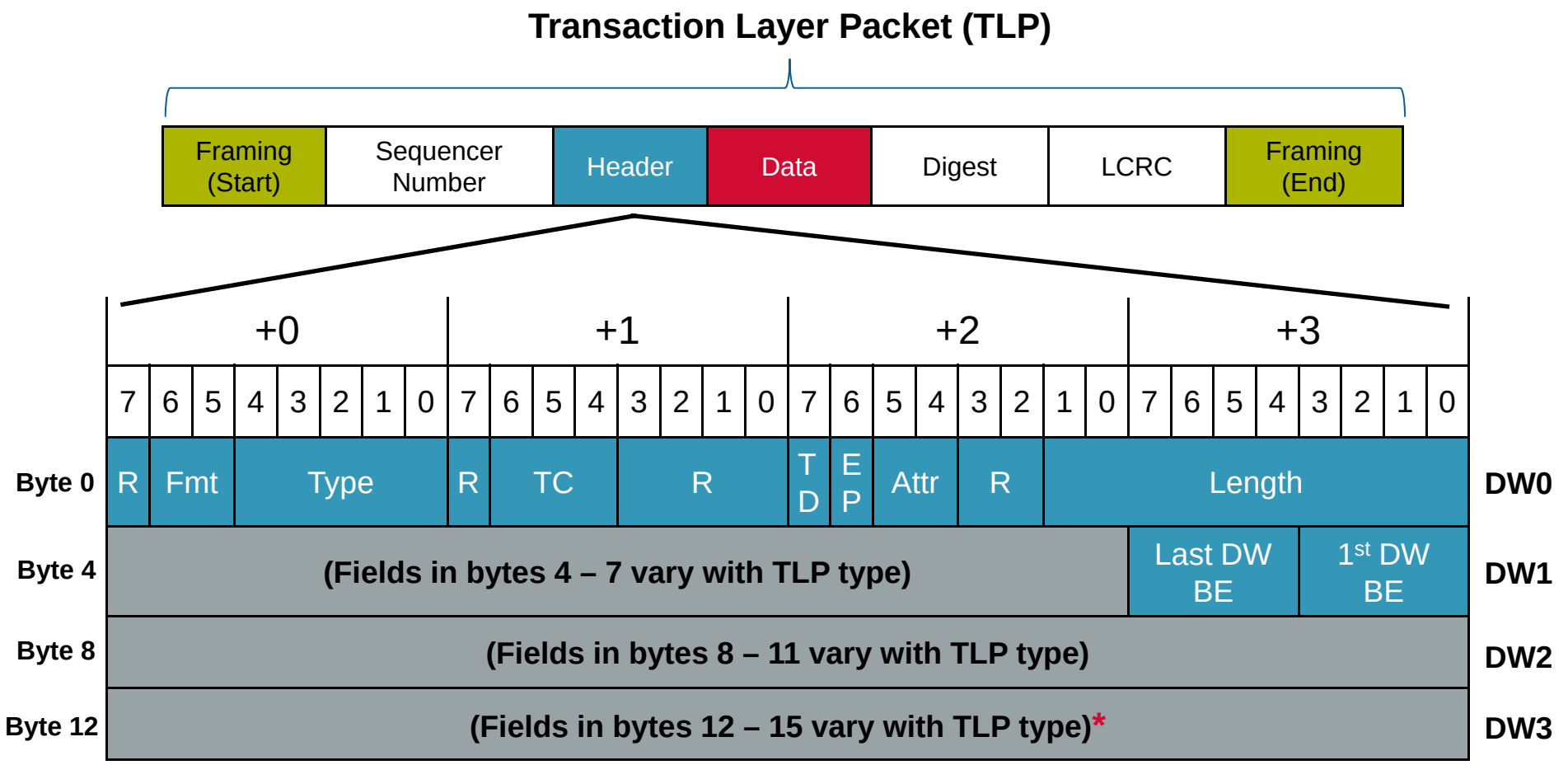


freescale semiconductor

PCI Express TLP Disassembly



PCI Express TLP Generic Header Fields



**** Only applicable for 4 DW TLP headers***

freescale semiconductor

PCI Express TLP Fmt[1:0] and Type Encoding

Table 2-3: Fmt[1:0] and Type[4:0] Field Encodings

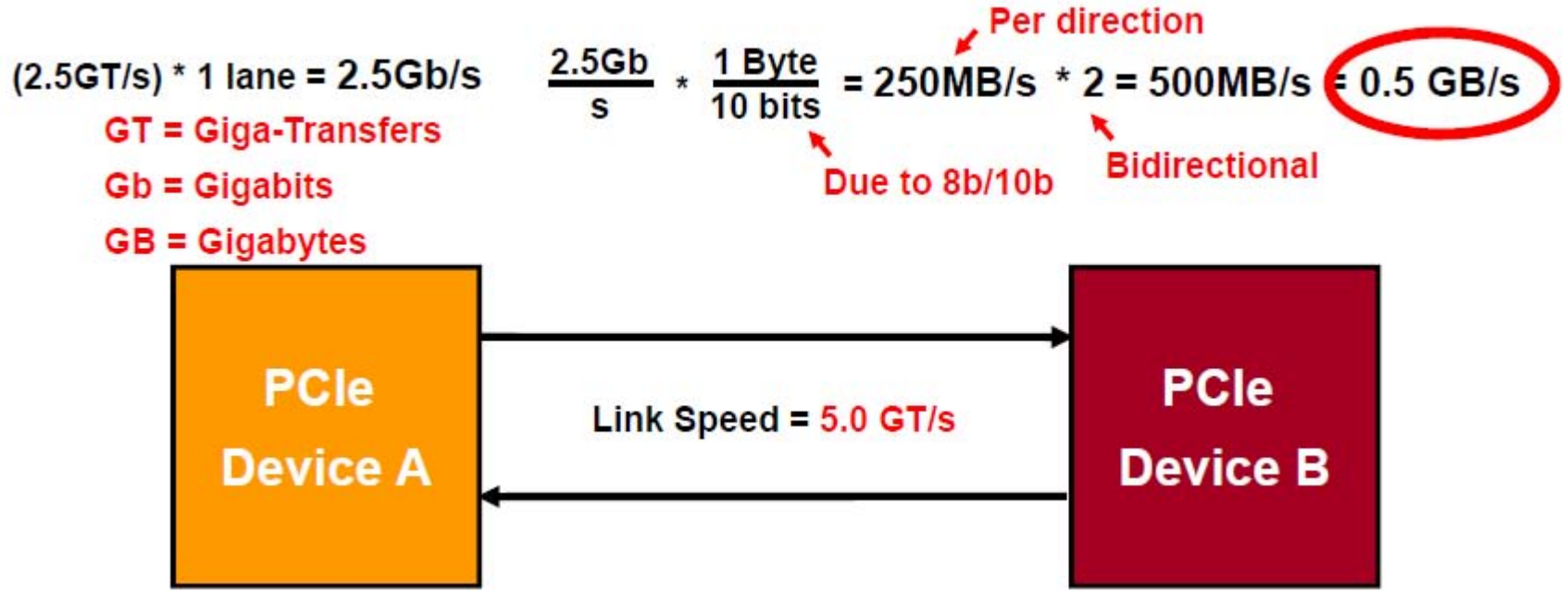
TLP Type	Fmt [1:0][2]	Type [4:0]	Description
MRd	00 01	0 0000	Memory Read Request
MRdLk	00 01	0 0001	Memory Read Request-Locked
MWr	10 11	0 0000	Memory Write Request
IORd	00	0 0010	I/O Read Request
IOWr	10	0 0010	I/O Write Request
CfgRd0	00	0 0100	Configuration Read Type 0
CfgWr0	10	0 0100	Configuration Write Type 0
CfgRd1	00	0 0101	Configuration Read Type 1
CfgWr1	10	0 0101	Configuration Write Type 1
Msg	01	1 0r$_2$r$_1$r$_0$	Message Request – The sub-field r[2:0] specifies the Message routing mechanism (see Table 2-11).
MsgD	11	1 0r$_2$r$_1$r$_0$	Message Request with data payload – The sub-field r[2:0] specifies the Message routing mechanism (see Table 2-11).
Cpl	00	0 1010	Completion without Data – Used for I/O and Configuration Write Completions and Read Completions (I/O, Configuration, or Memory) with Completion Status other than Successful Completion.
CplD	10	0 1010	Completion with Data – Used for Memory, I/O, and Configuration Read Completions.
CplLk	00	0 1011	Completion for Locked Memory Read without Data – Used only in error case.
CplDLk	10	0 1011	Completion for Locked Memory Read – otherwise like CplD.
			All encodings not shown above are Reserved.

PCI Express Gen2 vs. Gen1

- Gen2 base spec was publicly released on Jan. 15, 2007
- Higher speed (5.0 GT/s) supported:
  - Selectable de-emphasis levels
  - Selectable transmitter voltage swing
- Runtime bandwidth changes
  - Power savings and flexible bandwidth
  - Software notification of changes
- IO virtualization support
  - Access control services
  - Function level reset
- Other new features
  - Completion timeout control
  - Modified compliance pattern for testing

PCI Express Gen2 vs. Gen1 (continued)

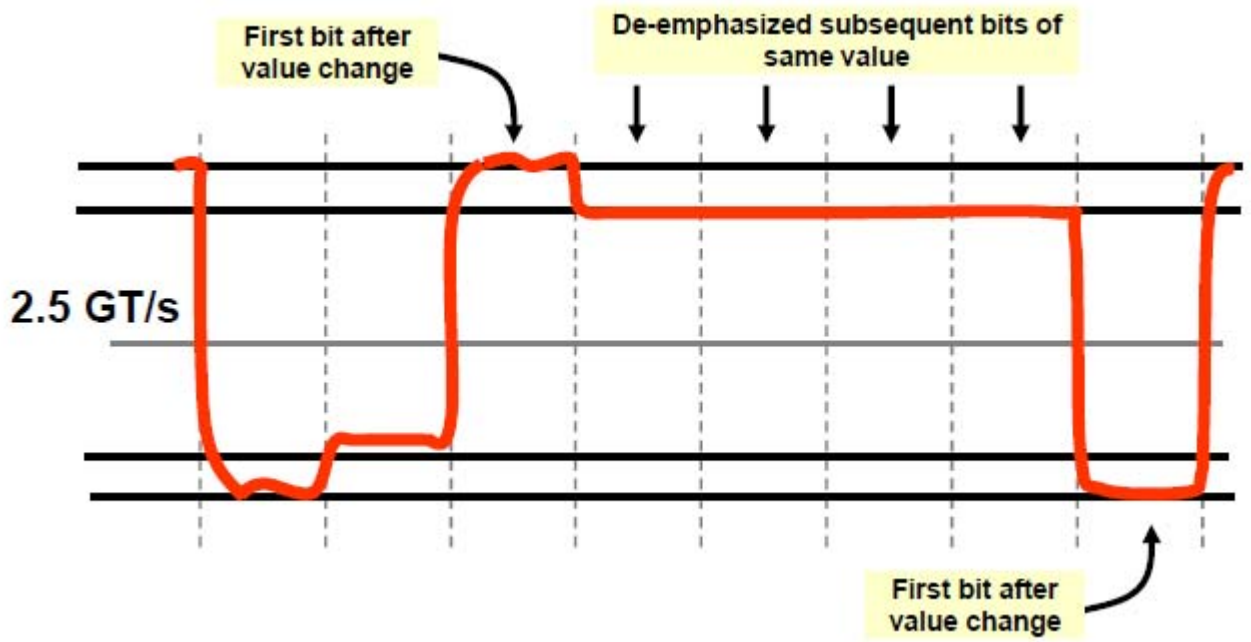
- Improved performance, reduced pin count
- Bandwidth based on 8b/10b encoding, aggregate bandwidth will be:
  - Gen1: 0.5 GB/s per lane
  - Gen2: 1.0 GB/s per lane



freescale semiconductor

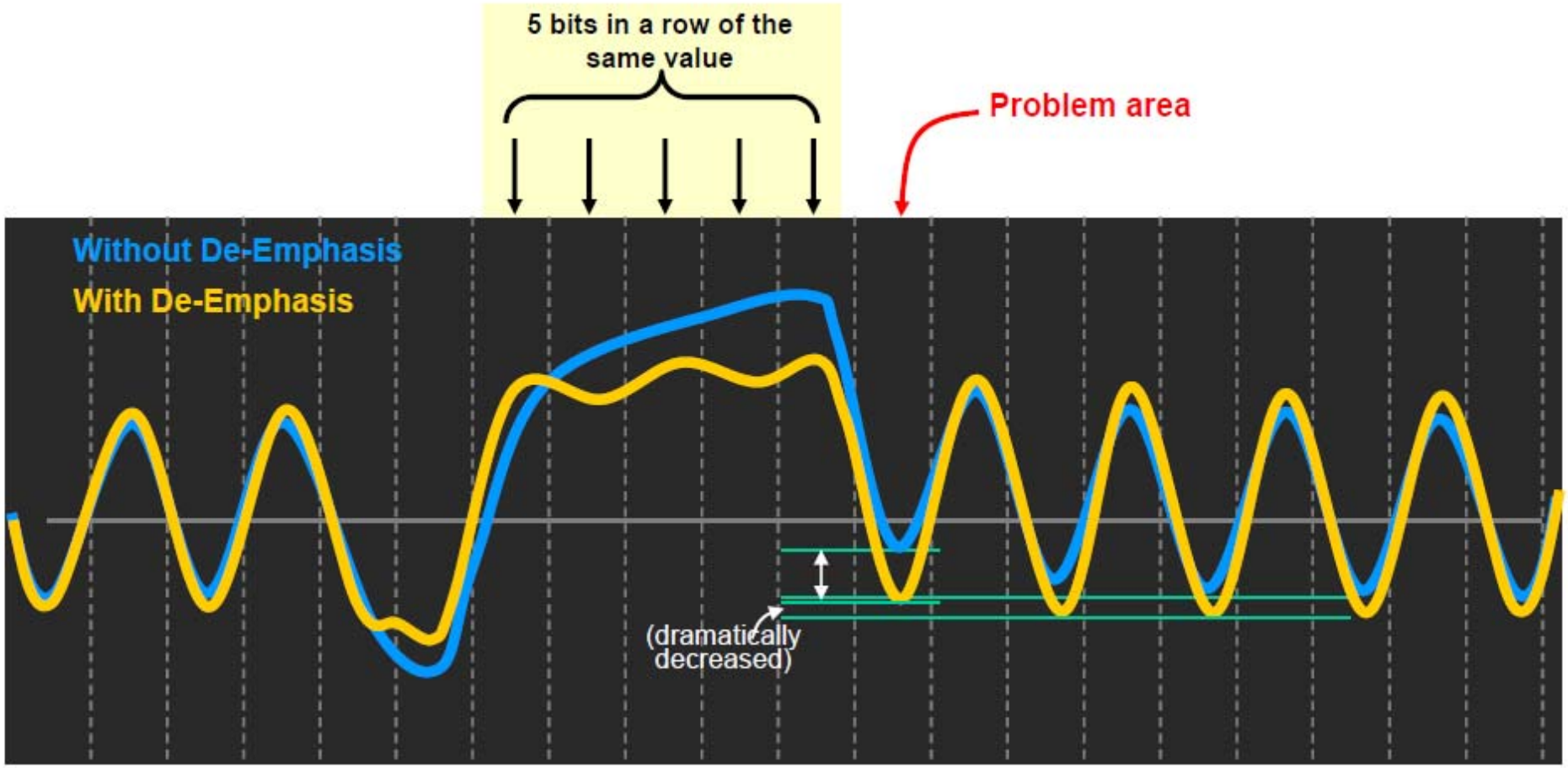
Brief De-Emphasis Review

- Higher frequency means greater sensitivity to jitter
- De-emphasis helps by reducing transmitter power during repeated bits
- Goals:
  - Reduced data-dependent jitter
  - Alleviate ISI (inter-symbol interference)



freescale semiconductor

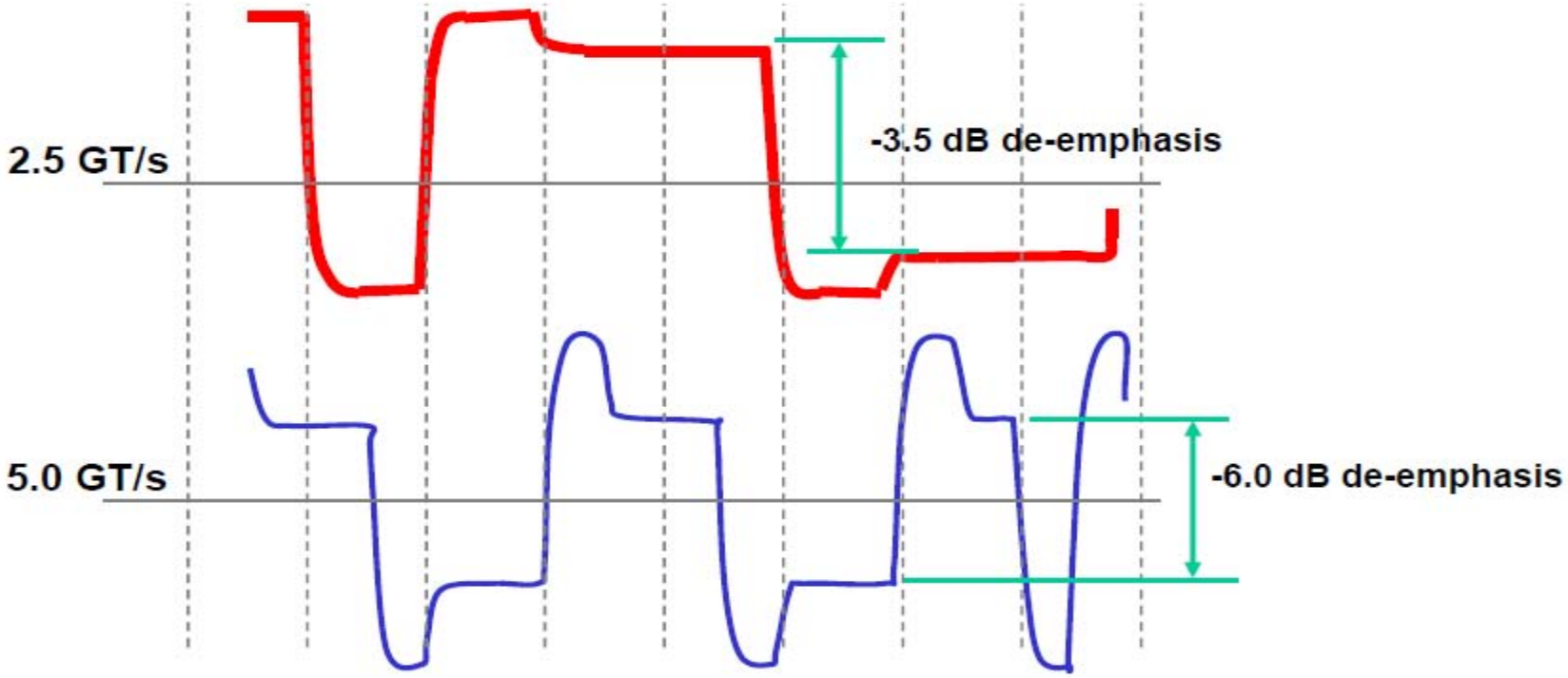
ISI (Inter-Symbol Interference)



freescale semiconductor

Selectable De-Emphasis

- Gen1 de-emphasis was always -3.5dB
- Gen2 de-emphasis is selectable:
  - At 2.5 GT/s: -3.5 dB
  - At 5.0 GT/s: -6.0 dB (-3.5 dB optional)



freescale semiconductor

PCIe Gen2 Support on Freescale QorIQ Products – P4080 and P4040

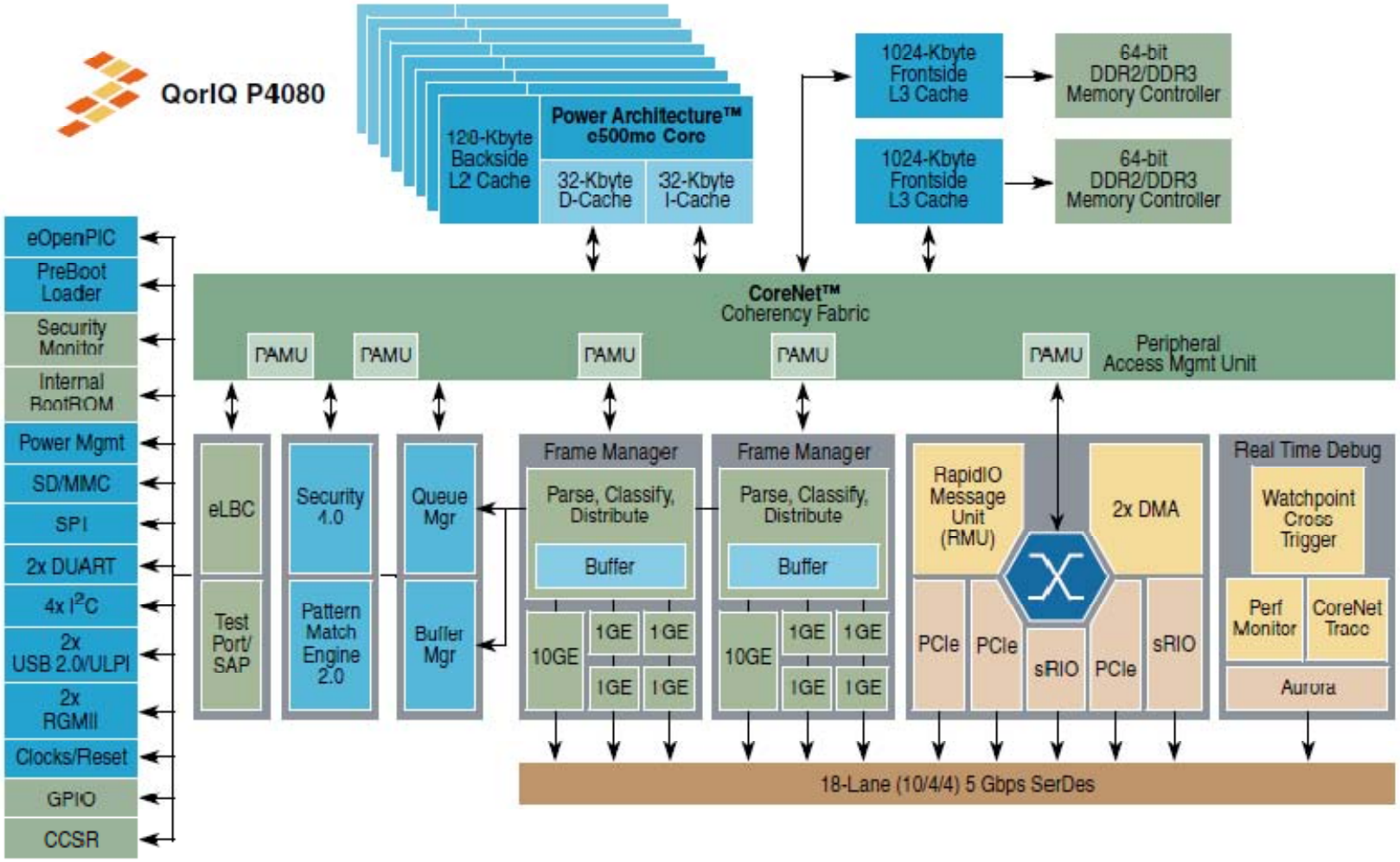


Figure 1-1. QorIQ P4080 Preliminary Block Diagram

freescale semiconductor

P4080 and P4040 PCI Express Gen2 Feature Highlights

- Three PCI express controllers:
  - Support both 2.5 GT/s and 5.0 GT/s speed
  - Support up to x4 5.0 GT/s maximum bandwidth for each controller
    - 5.0 GT/s per lane x4 lanes = 20 GT/s per direction for each controller
  - Initial flow control credit advertisement increased – for PH and PD
    - Now up to 6 posted credits (header and corresponding data credits) for inbound posted transactions, when SRIO is not used

- Other major changes:
  - Only one change in transaction layer → good news for software folks!
  - Few changes in PCIe controller registers
  - Some change in POR SerDes and clock ratio setup
  - Major change in electrical area → challenge for board guys!

Register Change #1: PEX_CONFIG – 2-bit Fields Added

- PCI Express Configuration Register (PEX_CONFIG)
  - Posted credit
    - 100: 4 posted credits (default).
    - 110: 6 posted credits, only if SRIO is not used! Good for Gen1 x8 or Gen2 x4
  - Outbound transaction address checking *against base/limit register setup*
    - 0 : Default, same as Gen1 parts, no checking.
    - 1 : Enable checking. If no hit → flag error!

Offset 0x014 Access: Read/Write

	0	12 13 15	16 17	18	26 27	28 29	30	31
R W	—	PC	—	OB_CK	—	SAC	—	SP	SCC
Reset	0 0 0 0 0 0 0 0 0 0 0 0 0	0 n n n	0 0	0	0 0 0 0 0 0 0 1 0	1	0	0	0

Figure 18-6. PCI Express Configuration Register (PEX_CONFIG)

Register Change #2: Dedicated MSI Inbound Window in RC

- Add a dedicated inbound window for MSI in RC Mode
  - Added 4 registers: PEXMSIITAR, PEXMSIIWBAR, PEXMSIIWBEAR, PEXMSIIWAR
  - Old usage model → MSI is handled by one of the P4080 core
    - Translation Address must be configured to point to one of the three registers (MSIIRA, MSIIRB, MSIIRC) in MPIC
  - New usage model → *MSI can be handled by other device*
    - *Set the translation address and target interface to route inbound MSI to a device connected to one of the SRIO port or a different PCIe port on P4080*

Offset 0xD00 Access: Read only

	0	7	8	11	12	31
R	—		TEA		TA	
W						
Reset	0xnnnn_nnnn					

Figure 18-25. PCI Express MSI Inbound Translation Address Register (PEXMSIITAR)

Offset 0xD10 Access: Mixed

	0	1	2	3	7	8	11	12	15	16	19	20	25	26	31
R	EN	DIEN	PF	—		TRGT		RTT		WTT		—		IWS	
W															
Reset	0	0	0	0	0 0 0 0	1	1 1 1	0	1 0 0	0	1 0 0	0 0 0 0	0	0 0 1	0 1 1 1

Figure 18-28. PCI Express MSI Inbound Window Attributes Register (PEXMSIIWAR)

freescale semiconductor

Register Change #3: PCIe Capabilities Register

- PCI Express Capabilities Register
  - The “Version” bit field = 0010b = 0x2h → this is a PCIe Gen2 device
  - But, *this doesn’t necessary mean it supports 5 GT/s Gen2 speed*
    - Check the Link Capabilities Register to confirm!

Offset 0x4E Access: Read only

	15	14	13					8	7			4	3			0
R	—		Interrupt Message Number					Slot	Device/Port Type				Version			
W																
Reset	0	0	0	0	0	0	0	0	0	*n*	0	0	0	0	1	0

Figure 18-89. PCI Express Capabilities Register

freescale semiconductor

Register Change #4: PCIe Link Capabilities Register

- PCI Express Link Capabilities Register
  - The “MAX_LINK_SP” bit field = 0010b = 0x2h
    - Yes! This PCIe Gen2 device supports both 2.5 GT/s and 5 GT/s speed
    - But, *this doesn’t necessary mean it actually operates at 5 GT/s speed*
      - Check the Link Status Register to confirm!
  - Link bandwidth notification capable (LBWN)
    - LBWN=1 indicates P4080 supports Link Bandwidth Notification status and interrupt mechanisms

Offset 0x58 Access: Read only

	31–24	23	22	21	20–18	17–15	14–12	11	10	9–4	3–0
R	Port Number	—		LBWN	—	L1_EX_LAT	L0s_EX_LAT	ASPM		MAX_LINK_W	MAX_LINK_SP
W											
Reset	0 0 0 0 0 0 0 0	0	0	1	0	0 0 1	1 1 0 1	1	0	1 0 0 0 0	0 0 1 0

Figure 18-93. PCI Express Link Capabilities Register

freescale semiconductor

Register Change #5: PCIe Link Control Register

- PCI Express Link Control Register
  - Hardware autonomous width disable (HWAWD)
    - When set, disables HW from changing the link width *for reasons other than attempting to correct unreliable link operation by reducing link width*
  - Link bandwidth management interrupt enable (LBMIE)
    - Only applicable for RC or switch downstream port
    - When set, enables interrupt generation if Link Status Register [LBMS] is set
  - Link autonomous bandwidth interrupt enable (LABIE)
    - Only applicable for RC or Switch downstream port
    - When set, enables interrupt generation if Link Status Register [LABS] is set

Offset 0x5C Access: Mixed

	15–12	11	10	9	8	7	6	5	4	3	2	1–0
R W	—	LABIE	LBMIE	HWAWD	—	EXT_SYNC	CCC	RL	LD	RCB	—	ASPM_CTL
Reset	All zeros											

Figure 18-94. PCI Express Link Control Register

Register Change #6: PCIe Link Status Register

- PCI Express Link Status Register
  - Negotiated link speed (LINK_SP)
    - 0001 : Default, 2.5 GT/s
    - 0010 : 5.0 GT/s (*final proof that the link is operating at Gen2 speed*)
  - Negotiated link width (NEG_LINK_W)
  - Link bandwidth management status (LBMS) → *for RC/Switch downstream port*
    - Set by hardware to indicate either of following occurred on downstream port without transitioning through DL_Down status:
      - A link training has completed following a write of 1b to Retrain Link bit
      - HW has changed link speed or width *to attempt to correct unreliable link*
  - Link autonomous bandwidth status (LABS) → *for RC/Switch downstream port*
    - Set by HW to indicate that HW has autonomously changed link speed or width on downstream port without transitioning through DL_Down status, *for reasons other than attempt to correct unreliable link operation*

Offset 0x5E Access: Read only

	15	14	13	12	11	10	9				4	3				0
R	LABS	LBMS	—	SCC	LT	—	NEG_LINK_W					LINK_SP				
W																
Reset	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1

Figure 18-95. PCI Express Link Status Register

freescale semiconductor

Register Change #7: PCIe Device Capabilities 2 Register

- PCI Express Device Capabilities 2 Register (***Read-Only***)
  - Completion timeout range supported (CPL_TO_RS)
    - Value reflects *completion timeout programmability* supported by this device
    - Purpose → Allows system software to modify completion timeout value
  - Completion timeout disable supported (CPL_TO_DS)
    - 1: indicates this device supports *completion timeout disabling mechanism*

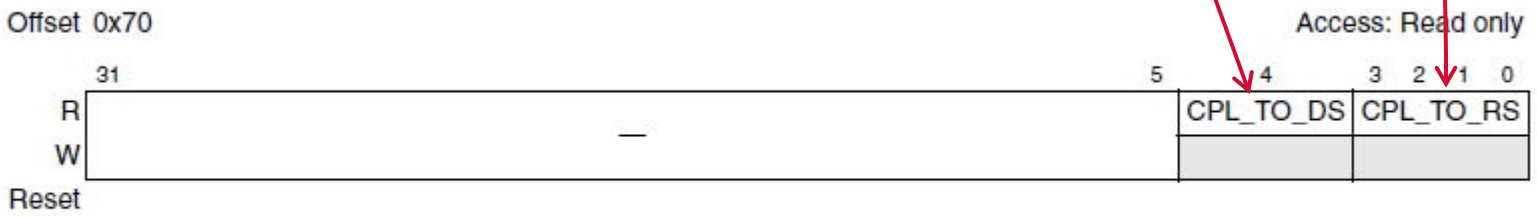


Figure 18-101. PCI Express Device Capabilities 2 Register

freescale semiconductor

Register Change #8: PCIe Device Control 2 Register

- PCI Express Device Control 2 Register
  - Completion timeout value (CPL_TO_VAL)
  - Completion timeout disable (CPL_TOD)
    - Purpose: Allows software to dynamically disable or enable completion timeout detection mechanism
    - 0 : Default, Completion timeout detection is enabled
    - 1 : Completion timeout detection is disabled

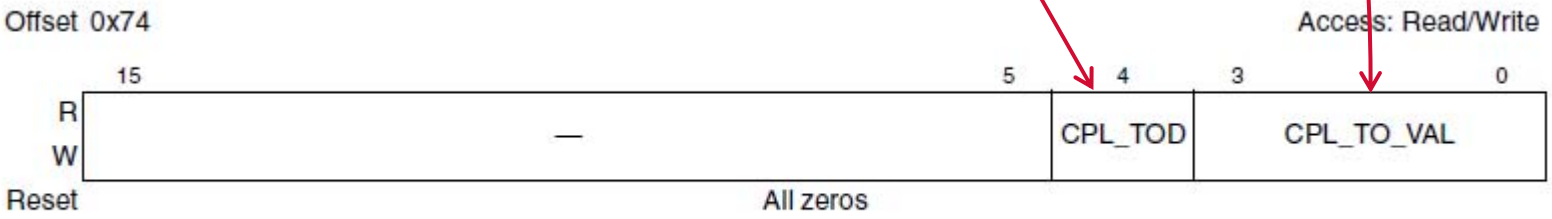


Figure 18-102. PCI Express Device Control 2 Register

Register Change #9: PCIe Link Control 2 Register

- PCI Express Link Control 2 Register
  - Target link speed (T_LS)
    - *Operation mode: used to set downstream port's link operation speed upper limit*
    - Compliance mode: used to set both upstream and downstream ports' target compliance mode speed when software uses EC to force a link into compliance mode
  - Enter compliance (EC)
    - Software can set this bit to force a link to enter compliance mode at the speed indicated in T_LS
  - Hardware autonomous speed disable (HWASD)
    - When set, *disables HW from changing link speed for device specific reason* other than attempting to correct unreliable link operation by reducing link speed
  - Selectable de-emphasis (SDE) → applicable for downstream port of RC/Switch
    - *Selects downstream port's de-emphasis level when link is operating at 5 GT/s*
    - 1 : -3.5 dB
    - 0 : -6 dB

Offset 7C — Access: Read/Write

	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0
R W	—			CDE	CSOS	EMC	TxM			SDE	HWASD	EC	T_LS			
Reset	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Figure 18-103. PCI Express Link Control 2 Register

freescale semiconductor

Register Change #10: PCIe Link Status 2 Register

- PCI Express Link Status 2 Register (***Read-Only***)
  - Current de-emphasis level (DE_LVL)
    - Applicable only when the link is operating at 5 GT/s
    - 1 : -3.5 dB
    - 0 : -6 dB

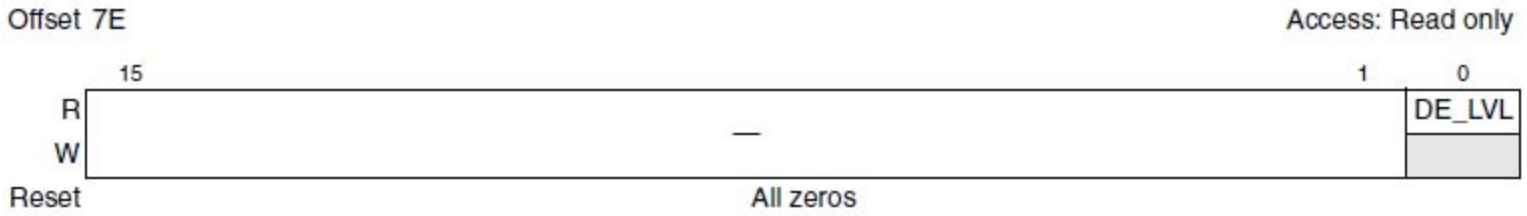


Figure 18-104. PCI Express Link Status 2 Register

freescale semiconductor

P4080 PCI Express Major Steps for Bring-up

- Brief POR pin configuration
  - Most POR configuration is stored in RCW now
  - POR configuration pins to be configured:
    - RCW source, DRAM type, flash ECC

- Configure all other proper POR values in RCW
  - HOST_AGT_B1, HOST_AGT_B2
    - set PCIe RC or EP mode for Bank 1 & 2
  - SRDS_PRTCL
    - select SerDes lane/protocol usage based on allowed mapping
  - SRDS_RATIO_Bn
    - select SerDes clock ratio for Bank 1 & 2
  - SRDS_DIV_Bn
    - select SerDes PLL clock divider for Bank 1 & 2
  - SRDS_EN
    - set to 1 to enable SerDes

P4080 PCI Express Major Steps for Bring-up (*continue*)

- Configure all SerDes registers in RCW's PBI section
  - *Please only touch those that really need to be changed and leave as much in default as possible!*
  - SerDes registers you might need to touch:
    - SRDSBnPLLCR0 → selects ref. CLK frequency other than 100 MHz
    - SRDSGR0 → selects the DDR supply voltage that the SerDes uses
    - *Other registers called out by the workaround section of the errata document*

- Configure LAW and ATMU windows
  - P4080 offers 32 LAWs
  - Each PCIe controller offers:
    - 4 ATMU outbound windows in addition to OW#0, and
    - 3 ATMU inbound windows in addition to the dedicated one for MSI

- Configure major registers required for the PCIe controller
  - Both configuration space and memory-mapped space


freescale™
semiconductor