PCIe Fundamentals

Architecture, Topology, Bus

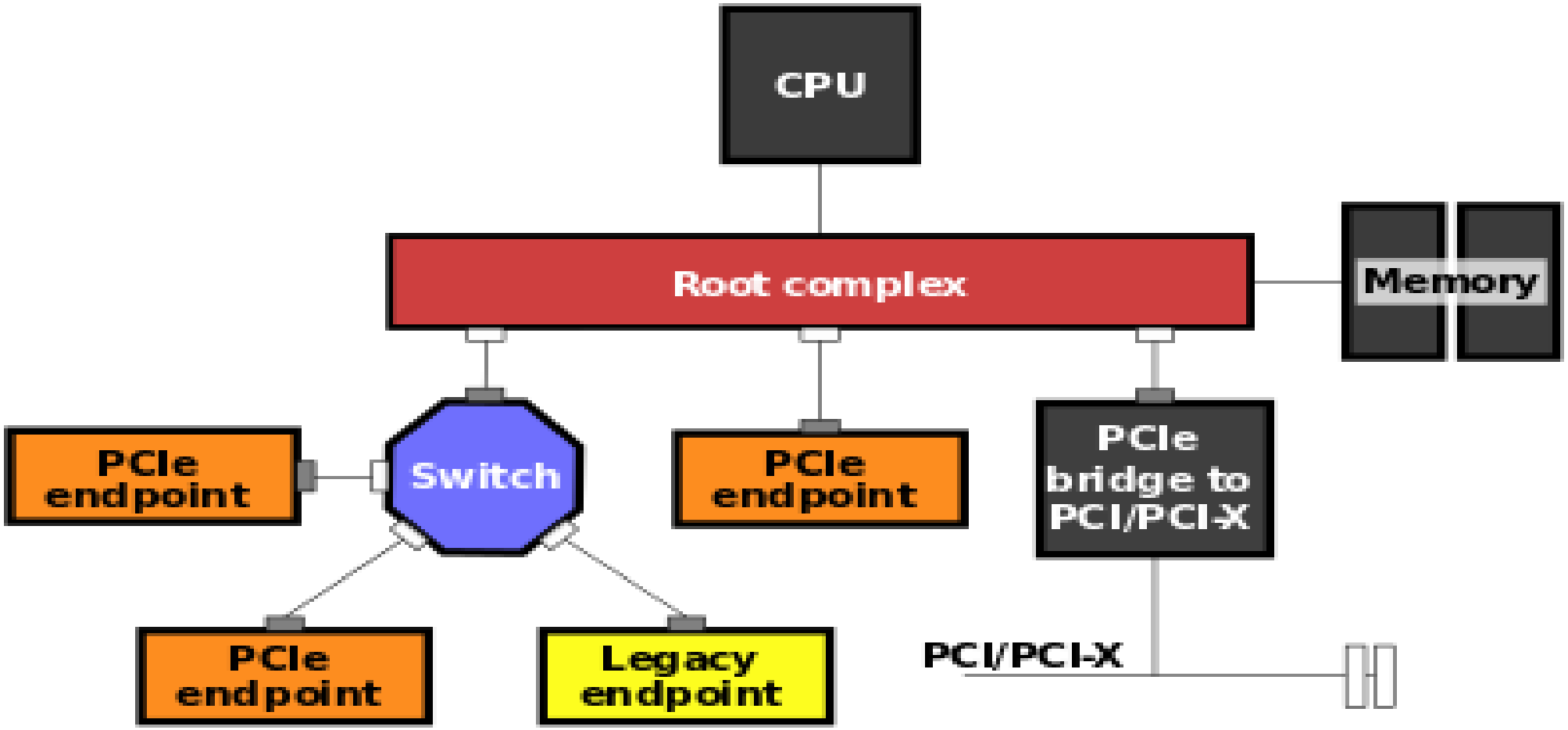
History

- PCI (1992 / 1993)
  - 32-bit / 33 MHz – 133 MB/sec
  - 64-bit / 66 MHz – 533 MB/sec
- PCI-X (1999)
  - Upto 1066 MB/sec with 64 bit / 133 MHz
- PCI Express (2002)
  - x 1 – upto 1 GB/sec in each direction
  - x16 – upto 16 GB/sec in each direction

PCIe

- High speed Serial Bus Standard used to connect peripheral device
- Provides Lower Latency and Higher data throughput
- Each PCIe Link has direct / dedicated point to point connection
- Lower pin count and smaller foot print
- Has more detailed error detection and reporting mechanism
- Packet based transmission protocol
- Has Scalable link widths (1x, 2x, 4x, 8x, 16x, 32x)
- Has Scalable Link Speeds (2.5, 5.0 and 8.0 GT/sec)

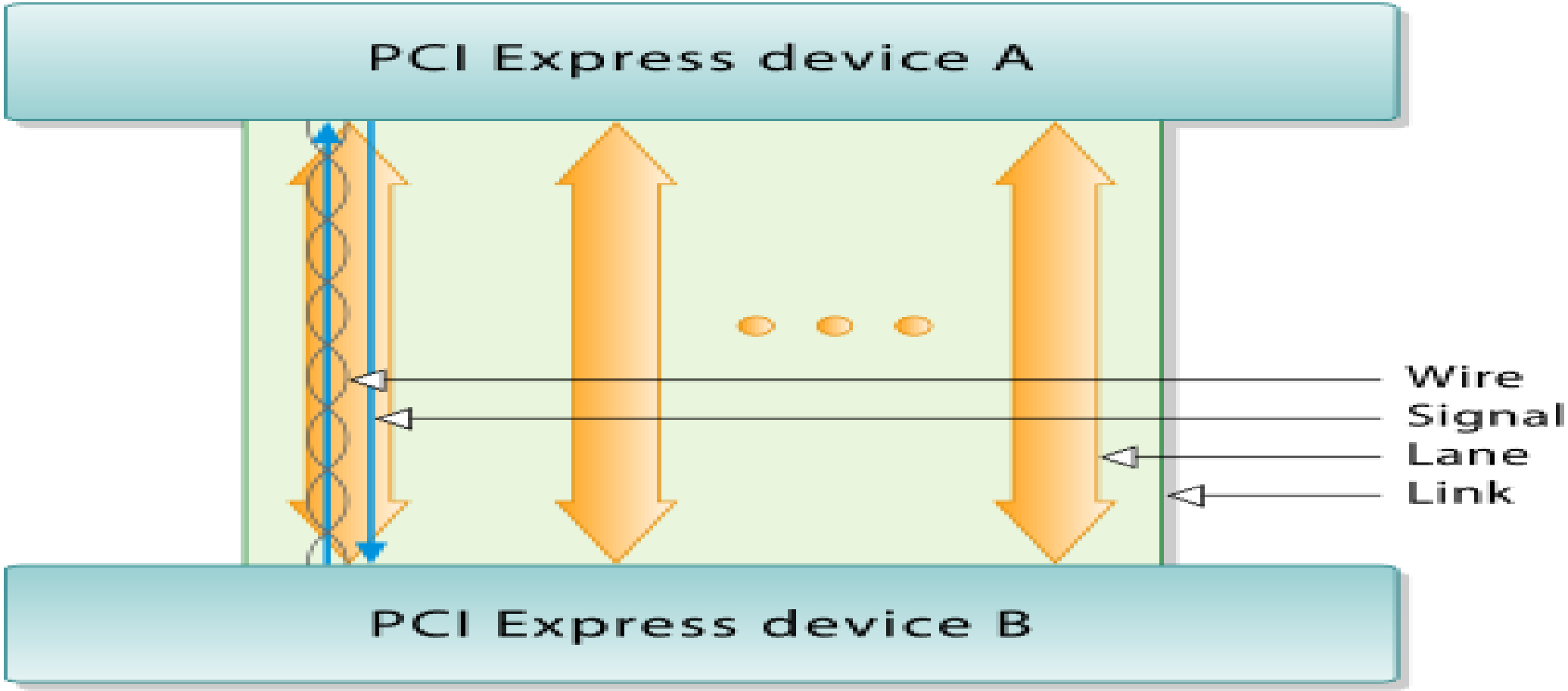
PCI Express Topology



Architecture

- Based on Point-to-Point Topology
- Separate Serial link connects every device to root complex (host)
- Supports full-duplex communication between any two endpoints
- Communication encapsulated in packets
- Transaction layer handles packetizing and de-packetizing
- Types of packets
  - Data packets
  - Status Packets

PCIe Terminology



PCIe Lane Negotiation

- PCIe link between two devices can be one to 32 lanes
- In multi lane link, packet data is stripped across lanes
- Lane count is automatically negotiated during device initialization
- Also can be restricted by endpoint
- Using this, single lane express device can be used in multi-lane slot
- The initialization cycle auto negotiates highest mutual lane count
- Links also can be down dynamically and use fewer lanes as well

PCI Express Throughput

Version	x1	x2	x4	x8	X16
PCIe 1.x	0.25	0.5	1.0	2.0	4.0
PCIe 2.x	0.5	1.0	2.0	4.0	8.0
PCIe 3.x	1.0	2.0	4.0	8.0	16.0
PCIe 4.x	2.0	4.0	8.0	16.0	32.0

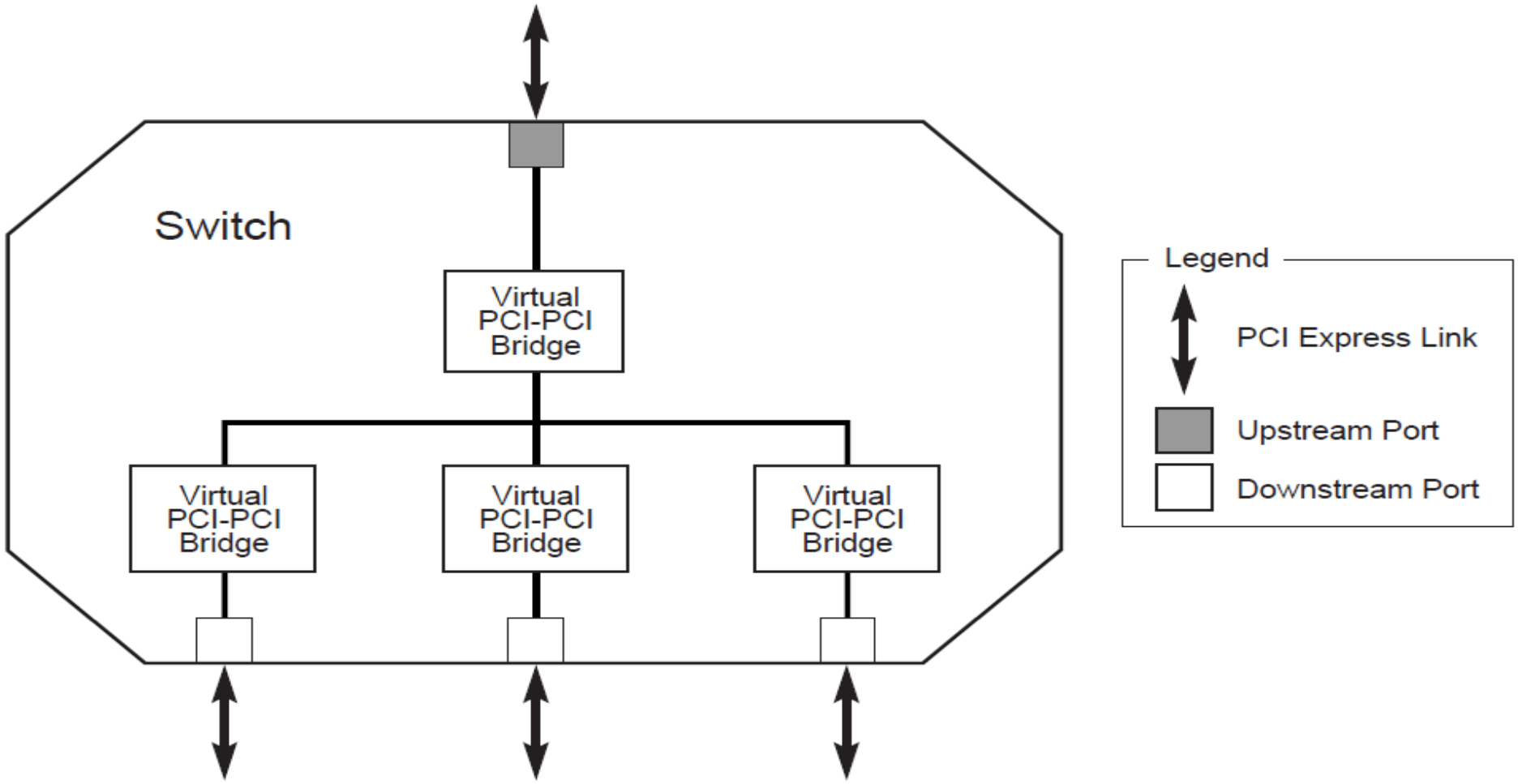
Root Complex

- Root of an hierarchy that connects the CPU / Memory sub system
- Supports one or more PCI Express Ports
- Each interface defines a separate hierarchy domain
- Splits a packet into smaller size packets
- Generates Configuration / IO Requests as a requestor
- Also generates Locked Requests as requestor

End Point

- Must be a function with a Type 00h
- Must supports Configuration Requests as Completor
- Must not depend on OS allocation of I/O resources
- Message Signaled Interrupt (MSI), if interrupt is requested
- 64 bit addressing must be supported, if prefetch bit is set
- 32 bit addressing must be supported, if prefetch bit is not set
- BAR Minimum memory address range is 128 bytes
- Should appear in one of domains

Switch


Switch
Virtual PCI-PCI Bridge
Virtual PCI-PCI Bridge
Virtual PCI-PCI Bridge
Virtual PCI-PCI Bridge
Legend
PCI Express Link
Upstream Port
Downstream Port

Switch

- Appears to be two or more logical PCI-to-PCI bridges
- Address based routing except when in Multicast
- Forwards all type of Transaction layer packets b/w set of ports
- Arbitration will happen in round robin or weighted round robin
- Does not split the packet into smaller packets
- Switch port supports flow control specification

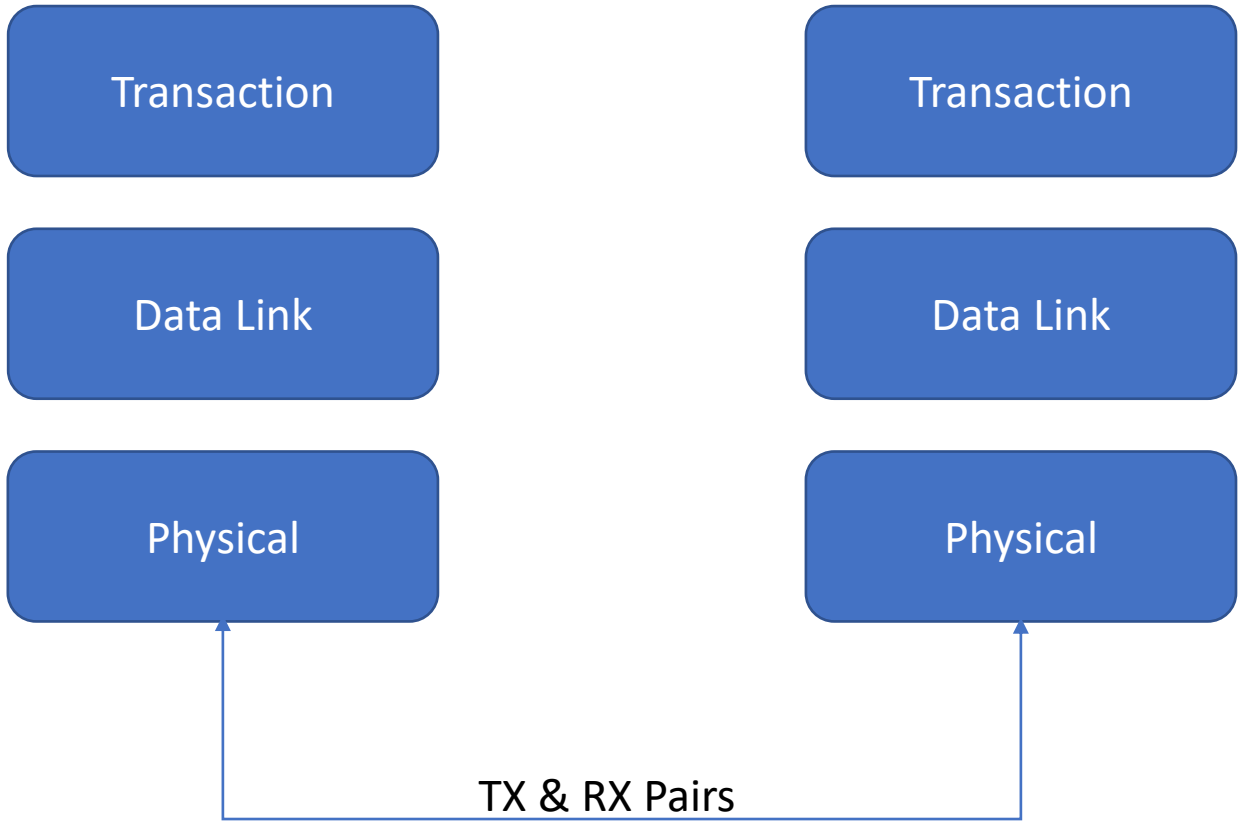
PCIe to PCI/PCI-X Bridge

- Provides a connection between PCIe and PCI/PCI-X

PCI Compatible Model

- Has the ability to enumerate & configure PCIe similar to PCI
- Boots the existing Operating System without modifications
- Existing I/O Device Drivers can be used without modifications
- Configure / Enable PCIe functionality by PCI configurations

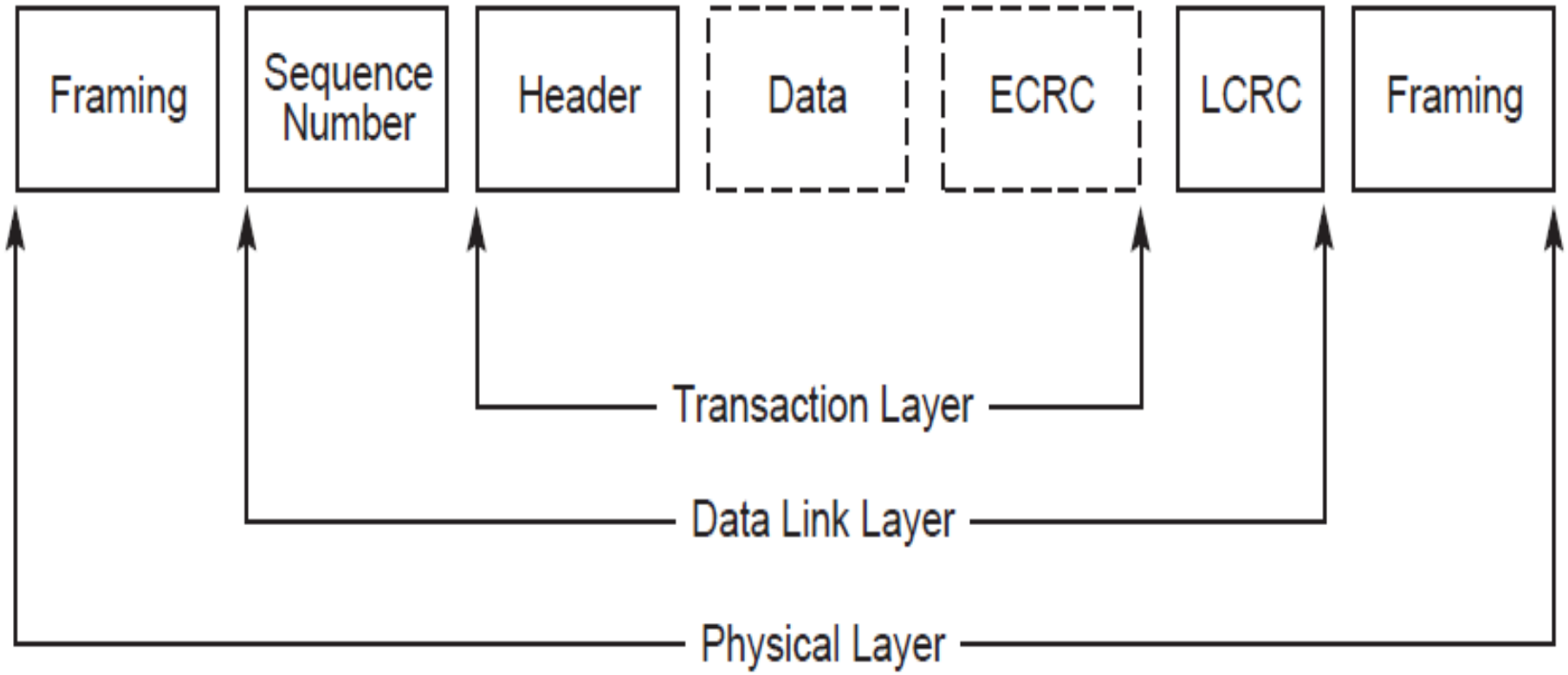
PCIe Layering



PCIe Layering

- Uses packets to communicate data between components
- Packets are formed in Transaction, Data Link layer on transmission
- At receiving side, reverse process occurs and packets get transform
- Transmitted packet flow through other layers, they get extended

PCIe Packet Flow



Transaction Layer

- Upper layer of PCIe architecture
- Assembles and Disassembles Transaction Layer Packets (TLP)
- TLPs used to communicate transactions, such as read & write
- Responsible for managing credit based flow control
- Each pkt unique identifier that enables response pkt to correct originator
- Supports four address space
  - Memory, I/O, Configuration and Message
- Message space supports prior side-band signals,
  - Interrupts, Power Management requests

Transaction Layer Services

- Process of generating and receiving TLPs
- Exchanges flow control information with other side of link
- Responsible for SW / HW initiated power management
- Stores the Link Configuration information generated by CPU
- Stores the Link Capabilities generated by Physical Layer
- Packet Generation and Processing Services
- Flow Control Services
- Power Management Services

Data Link Layer

- Middle layer in the PCIe Architecture / Stack
- Intermediate between transaction and physical layer
- Link Management and data integrity, error detection & correction
- Accepts TLPs, calculates and applies
  - data protection code, TLP Sequence Number
- Checks the integrity of received TLPs
- On TLP error, requests retransmission of TLPs until
  - correctly received or Link is determined to have failed
- Generates & Consumes packets used for Link Management Functions
- These packets are referred as Data Link Layer Packets (DLLP)

Data Link Layer Services

- Responsible for reliable information exchange with opposite link
- Initialization and Power Management Services
- Active / Reset / Disconnected / Power Managed states
- Accept power state requests and convey to physical layer
- Data Protection, Error Checking and Retry services

Physical Layer

- Includes all the hardware circuity for interface operation
- Also logical functions like interface initialization and maintenance
- Converts information from DLL into an serialized format
- Serialized format into an information for DLL
- Supports future performance enhancements via speed upgrades

Physical Layer Services

- Interface initialization, maintenance control and status tracking
- Symbol and special ordered set generation
- Symbol transmission and alignment
  - Transmission circuits
  - Reception Circuits
  - Elastic Buffer at receiving side

Inter Layer Interface

- Transaction / Data Link Interface
  - Byte or Multi-Byte data sent across the link
  - TLP framing information for the received byte
  - Requested power state for the link
  - Link Status Information
- Data Link / Physical Interface
  - Byte or Multi-Byte wide data received from PCIe Link
  - TLP and DLLP framing information for data
  - Indication of errors detected by the physical layer
  - Connection status information

Transaction Types, Address Space

Address Space	Transaction Type	Usage
Memory	Read / Write	Transfer data to/from memory mapped location
I/O	Read / Write	Transfer data to/from an IO mapped location
Configuration	Read / Write	Device Function Config Setup
Message	Baseline	From event signaling mechanism to general purpose messaging

Memory Transactions

- Memory transaction are given below,
  - Read Request / Completion
  - Write request
  - AtomicOp Request / Completion
- Two different address formats
  - Short Address format : 32 bit address
  - Long Address format : 64 bit address

I/O Transactions

- Supports I/O space for compatible with legacy devices
- Future this may be deprecated the use of I/O space
- The following are I/O transaction types
  - Read Request / completion
  - Write Request / Completion
- Address formats
  - Short address format : 32 bit address

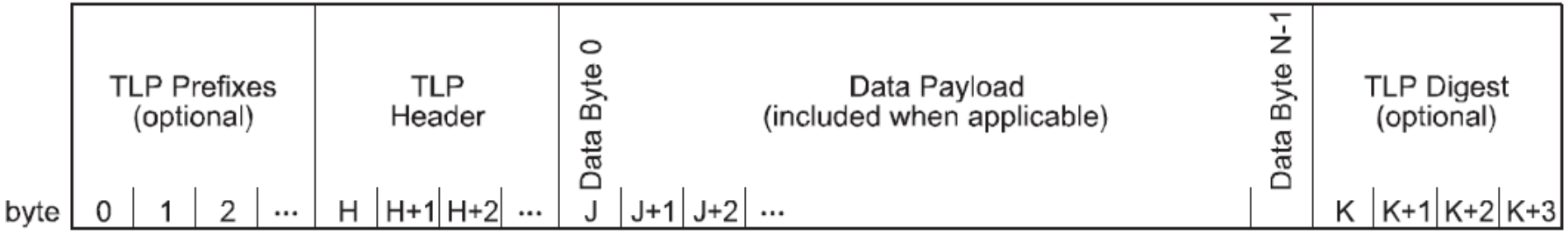
Configuration Transaction

- Used to access the configuration registers of function in devices
- Configuration transaction types
  - Read Request / Completion
  - Write Request / Completion

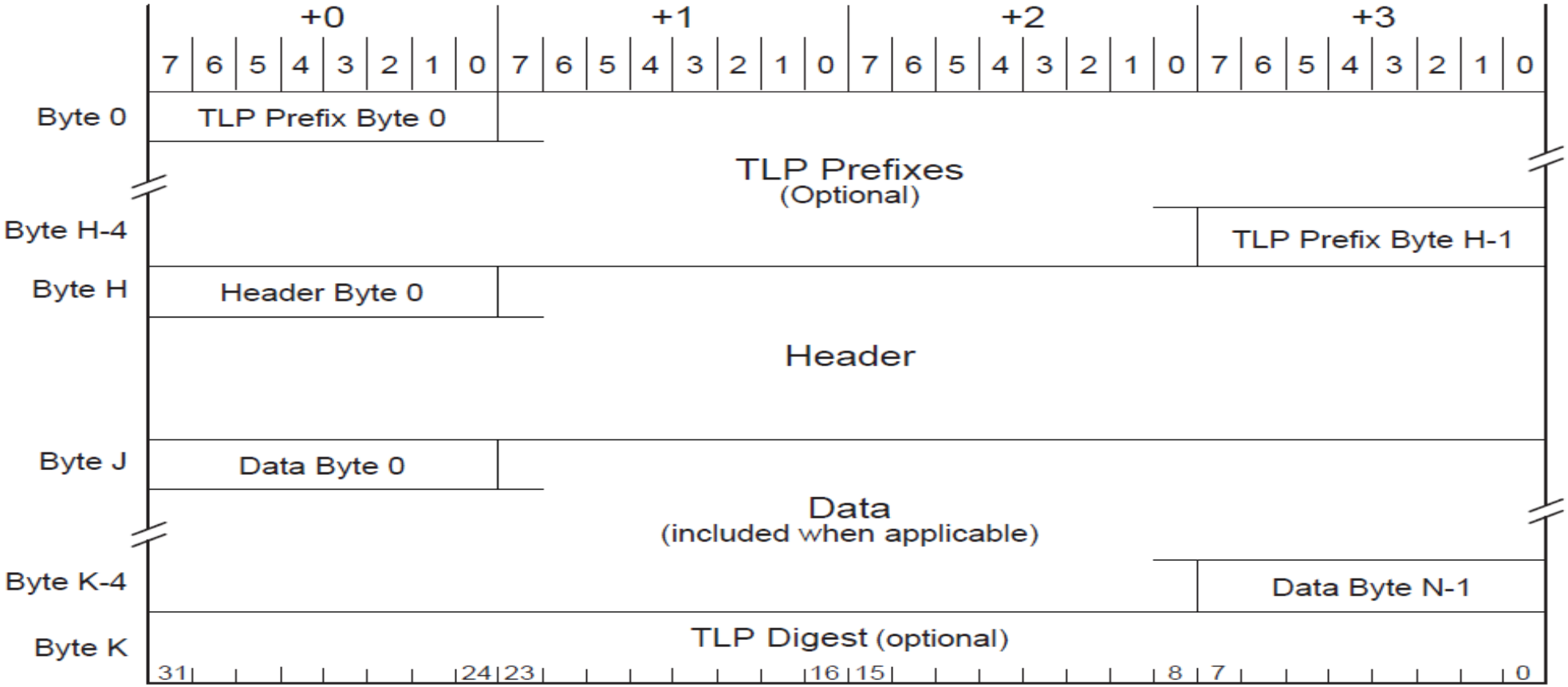
Message Transactions

- Used to support in-band communication of events between devices
- Vendor defined messages using specified message codes
- Vendor defined messages are more specific to the platform

TLP Packet Format



TLP Packet Format



TLP Packet Header

- Format of the packet
- Type of the packet
- Length of associated data
- Transaction Descriptor
- Address / Routing information
- Byte enables
- Message encoding
- Completion Status

Common TLP Header

	+0								+1								+2								+3							
	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0
Byte 0 >	Fmt			Type					R	TC			R	Attr	R	TH	TD	EP	Attr		AT		Length									

TLP Header Details

Bit Fields	Description
Fmt	Format
Type	TLP Packet Type
TC	Traffic Class
TH	Presence of TLP Prefix (if present)
Attr	Attributes
TD	TLP Digest Present
EP	TLP is poisoned
Length	Data Length notation
AT	Address Translation

Packet format & Type Encoding

TLP Type	Format	Type	Description
MRd	000 / 001	0 0000	Memory Read Request
MRdLk	000 / 001	0 0001	Memory Read Request Locked
MWr	010	0 0000	Memory Write Request
IORd	000	0 0010	I/O Read Request
IOWr	010	0 0010	I/O Write Request
CfgRd0	000	0 0100	Configuration Read Type 0
CfgWr0	010	0 0100	Configuration Write Type 0
CfgRd1	000	0 0101	Configuration Read Type 1
CfgWr1	010	0 0101	Configuration Write Type 1
Msg	001	1 0 r2 r1 r0	Message Request
MsgD	011	1 0 r2 r1 r0	Message Request with Data

Packet format & Type Encoding

TLP Type	Format	Type	Description
Cpl	000	0 1010	Completion without data
CplD	010	0 1010	Completion with data
CplLk	000	0 1011	Completion for Locked Memory Read without data
CplDLk	010	0 1011	Completion for Locked Memory Read
FetchAdd	010 / 011	0 1100	Fetch and Add AtomicOp Request
Swap	010 / 011	0 1101	Unconditional Swap AtomicOp Request
CAS	010 / 011	0 1110	Compare and Swap AtomicOp Request
LPrfx	100	0 L3 L2 L1 L0	Local TLP Prefix
Eprfx	100	1 E3 E2 E1 E0	End-End TLP Prefix

TLP Length Encoding

Length [9:0]	TLP Data Payload Size
00 0000 0001	1 DW
00 0000 0010	2 DW
……	
11 1111 1111	1023 DW
00 0000 0000	1024 DW

Address Formats

	+0								+1								+2								+3							
	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0
Byte 0 >	0	Fmt x		1	Type				R	TC			R	Attr	R	TH	TD	EP	Attr		AT				Length							
Byte 4 >	{Fields in bytes 4 through 7 depend on type of Request}																															
Byte 8 >	Address[63:32]																															
Byte 12 >	Address[31:2]																														PH	

Figure 2-7: 64-bit Address Routing

	+0								+1								+2								+3							
	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0
Byte 0 >	0	Fmt x		0	Type				R	TC			R	Attr	R	TH	TD	EP	Attr		AT				Length							
Byte 4 >	{Fields in bytes 4 through 7 depend on type of Request}																															
Byte 8 >	Address[31:2]																														PH	

Figure 2-8: 32-bit Address Routing

Address Translation

AT Coding	Description
00	Default / Untranslated
01	Translation Request
10	Translated
11	Reserved

Routing

	+0								+1								+2								+3							
	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0
Byte 0 >	0	Fmt x	1	Type					R	TC			R	Attr	R	TH	TD	EP	Attr		AT		Length									
Byte 4 >	{Fields in bytes 4 through 7 depend on type of TLP}																															
Byte 8 >	Bus Number								Device Number / Function Number (with ARI)					Function Number			{Fields in bytes 10 and 11 depend on TLP type}															
Byte 12 >	{Fields in bytes 12 through 15 depend on type of TLP}																															

ID Routing with 4 DW Header

	+0								+1								+2								+3							
	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0	7	6	5	4	3	2	1	0
Byte 0 >	0	Fmt x	0	Type					R	TC			R	Attr	R	TH	TD	EP	Attr		AT		Length									
Byte 4 >	{Fields in bytes 4 through 7 depend on type of TLP}																															
Byte 8 >	Bus Number								Device Number / Function Number (with ARI)					Function Number			{Fields in bytes 10 and 11 depend on TLP type}															

ID Routing with 3 DW Header

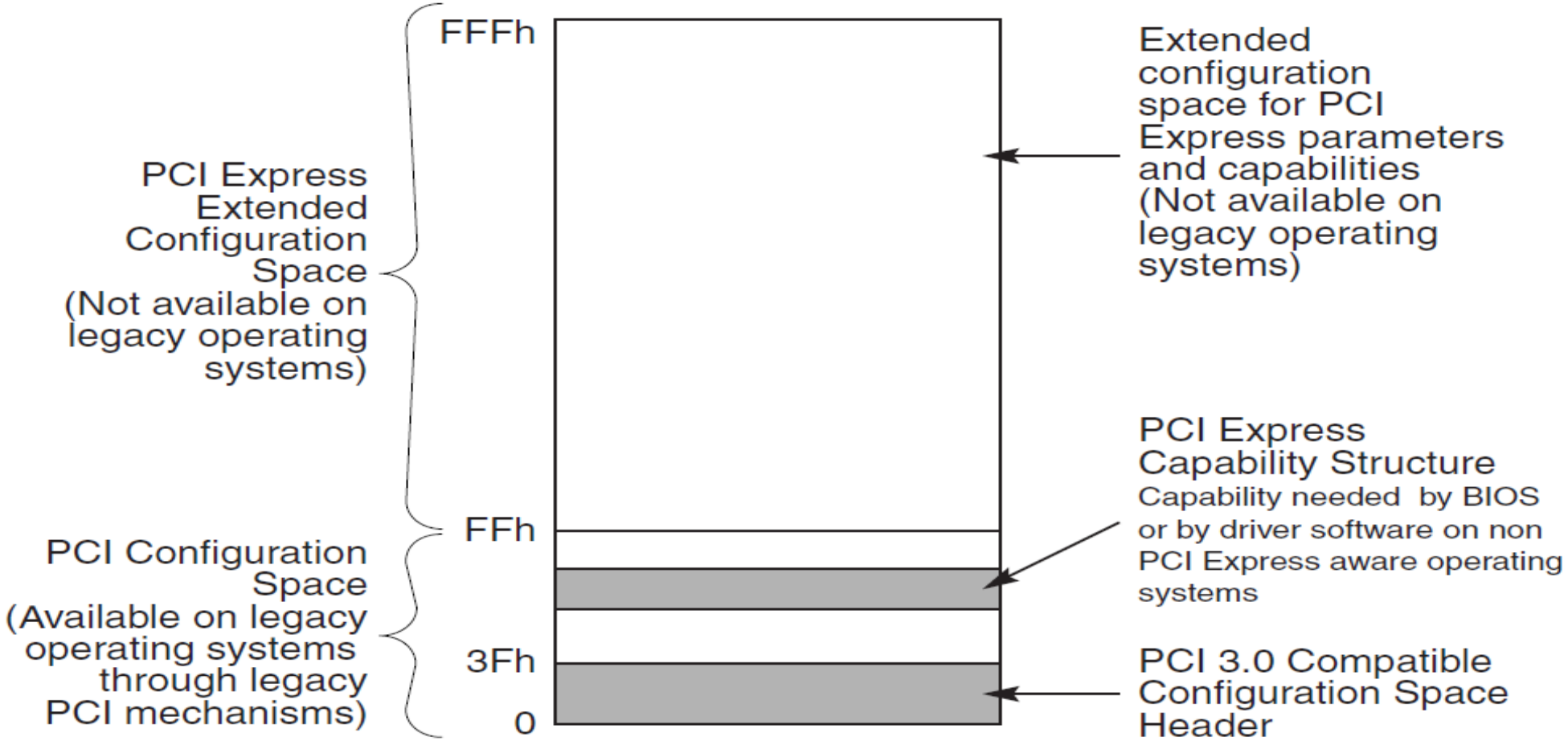
TLP Read Request

	31	30 29	28 27 26	25 24	23	22 21 20	19 18 17 16	15	14	13 12	11 10	9 8 7 6 5 4 3 2 1 0
DW 0	R 0	Fmt 0x0	Type 0x00		R 0	TC 0	R 0	TD 0	EP 0	Attr 0	R 0	Length 0x001
DW 1	Requester ID 0x0000							Tag 0x0c			Last BE 0x0	1st BE 0xf
DW 2	Address [31:2] 0x3f6bfc10											R 0

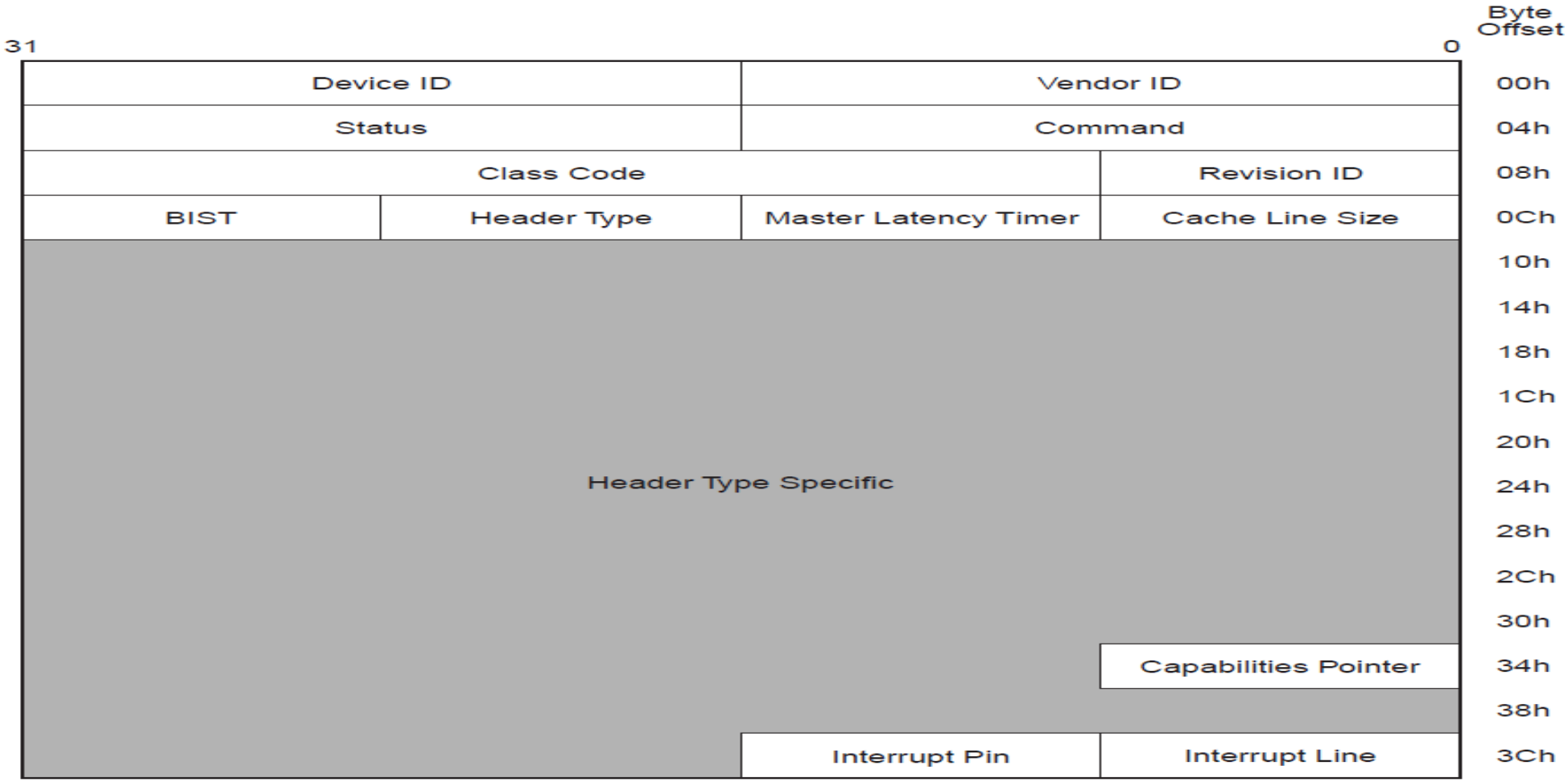
TLP Completion with data

	31 30 29 28	27 26 25 24	23	22 21 20	19 18 17 16	15	14	13 12	11 10	9 8 7 6 5 4 3 2 1 0
DW 0	R 0	Fmt 0x2	Type 0x0a	R 0	TC 0	R 0	TD 0	EP 0	Attr 0	R 0	Length 0x001
DW 1	Completer ID 0x0100					Status 0x00	BCM 0	Byte Count 0x004		
DW 2	Requester ID 0x0000					Tag 0x0c		R 0	Lower Address 0x40	
DW 3	Data DW 0 0x12345678									

PCIe Configuration Space



Type 0 / 1 Configuration Space



Identification Registers

- Device ID
  - Device Unique Identification Value
- Vendor ID
  - Vendor Unique identification value

Command Register

Bit	Register	Description
2	Bus Master Enable	Controls the PCIe Endpoint to issue memory / io req
3	Special Cycle Enable	Not applicable to PCIe
4	Memory Write and Inval	Not applicable to PCIe
5	VGA Pallete Snoop	Not applicable to PCIe
6	Parity Error Response	Logging of Master Data Parity Error
7	IDSEL	Not applicable to PCIe
8	SERR# Enable	Enables reporting of non-fatal & fatal error
9	Fast back to back transactions enable	Not applicable to PCIe
10	Interrupt Disable	Controls the ability of PCIe function generate interrupts

Status Register

Bit	Register	Description
3	Interrupt Status	Set indicates INTx emulation interrupt is pending
4	Capabilities List	Indicates presence of Extended PCIe capability list
5	66 MHz Capable	Not applicable to PCIe
7	Fast back to back enable	Not applicable to PCIe
8	Master Data Parity Error	Set by endpoint if parity error response bit in cmd reg
10:9	DEVSEL timing	Not applicable to PCIe
11	Signaled Target Abort	Set when function completes as a Completor Abort Error
12	Received Target Abort	Set when function receives completion with abort
13	Received Master Abort	Set when requester receives a completion with unsupported Request completion status
14	Signaled System Error	Set when function sends a fatal or non-fatal error message
15	Detected Parity Error	Set when function receives poisoned TLP

Registers

- Cache Line Register
- Latency Timer Register
- Interrupt Line Register
- Interrupt Pin Register
- Error Register

Type 0 Configuration Space

31			0	Byte Offset
Device ID		Vendor ID		00h
Status		Command		04h
Class Code			Revision ID	08h
BIST	Header Type	Master Latency Timer	Cache Line Size	0Ch
Base Address Registers				10h 14h 18h 1Ch 20h 24h
Cardbus CIS Pointer				28h
Subsystem ID		Subsystem Vendor ID		2Ch
Expansion ROM Base Address				30h
Reserved			Capabilities Pointer	34h
Reserved				38h
Max_Lat	Min_Gnt	Interrupt Pin	Interrupt Line	3Ch

Registers

- Base Address Registers (offset 10h – 24h)
  - Resource are mapped into memory space using BAR register
  - Supports 64 bit addressing for any BAR request prefetchable memory
  - Min memory space range requested is 128 Bytes
- Min_Gnt / Max_Lat Registers (offset 3Eh / 3Fh)
  - Does not applicable to PCIe, Hardwired to 00h

Type 1 Configuration Space

31			0	Byte Offset
Device ID		Vendor ID		00h
Status		Command		04h
Class Code			Revision ID	08h
BIST	Header Type	Primary Latency Timer	Cache Line Size	0Ch
Base Address Register 0				10h
Base Address Register 1				14h
Secondary Latency Timer	Subordinate Bus Number	Secondary Bus Number	Primary Bus Number	18h
Secondary Status		I/O Limit	I/O Base	1Ch
Memory Limit		Memory Base		20h
Prefetchable Memory Limit		Prefetchable Memory Base		24h
Prefetchable Base Upper 32 Bits				28h
Prefetchable Limit Upper 32 Bits				2Ch
I/O Limit Upper 16 Bits		I/O Base Upper 16 Bits		30h
Reserved			Capability Pointer	34h
Expansion ROM Base Address				38h
Bridge Control		Interrupt Pin	Interrupt Line	3Ch

Registers

- Base Address Register (offset 10h / 14h)
- Primary Bus Number
- Secondary Bus Number
- Secondary Latency Timer
- Secondary Status Register
- Prefetchable Memory Base / Limit
- Bridge Control Register

Web : www.neeveetech.com
E-Mail : nvhariharan@neeveetech.com
Youtube : https://www.youtube.com/user/neeveehariharan
Facebook : https://www.facebook.com/neeveetech/
Linkedin : https://www.linkedin.com/in/neeveehariharan/

Thank You.