RapidIO, PCI Express and Gigabit Ethernet Comparison

Pros and Cons of Using These Interconnects in Embedded Systems

RapidIO®

The Embedded Fabric Choice

Revision 03, May 2005

Agenda

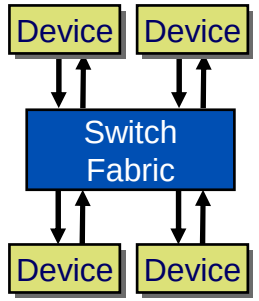
- Interconnect Requirements & Market Focus
- Architecture and Protocol
- System Level Considerations
- Conclusions

Interconnects Issues In Embedded Systems

- Desire for higher performance
  - Interconnects often bottlenecks
- Lower cost (Non-recurrent costs, capital expense, operating costs)
  - Standards-based development
- Modularity & Reuse
  - Standard interfaces promote reuse across platforms and over time
- Common Components
  - Standards reduce components and complexity
- Distributed Processing
  - Interconnects key to performance
- More Communication Standards and Interworking
  - Alphabet soup and getting worse!
- System-wide Interconnects
  - Backplane and line card (chip-to-chip)

Interconnect Trends

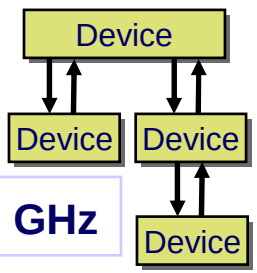
2nd Generation Point-to-Point
- Packet switched
- PHY: SERDES differential
- Lowest pin count



≥ 10 GHz

Example: PCI Ex / S-RIO

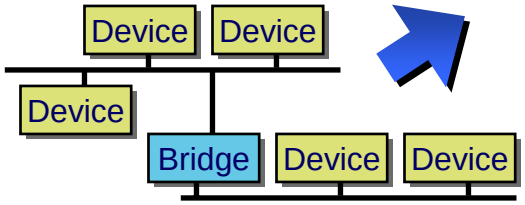
1st Generation Point-to-Point
- Packet switched
- PHY: Source-sync differential
- Lower pin count



≤ 3 GHz

Example: HT / P-RIO

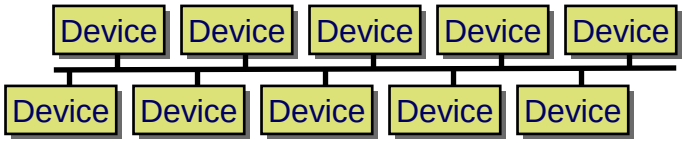
Hierarchical Bus
- Bridged Hierarchy
- Broadcast
- PHY: Single-ended



Example: PCI / PCI-X

≤ 133MHz

Shared Bus
- Single segment
- Broadcast
- PHY: Single-ended
- Highest pin count



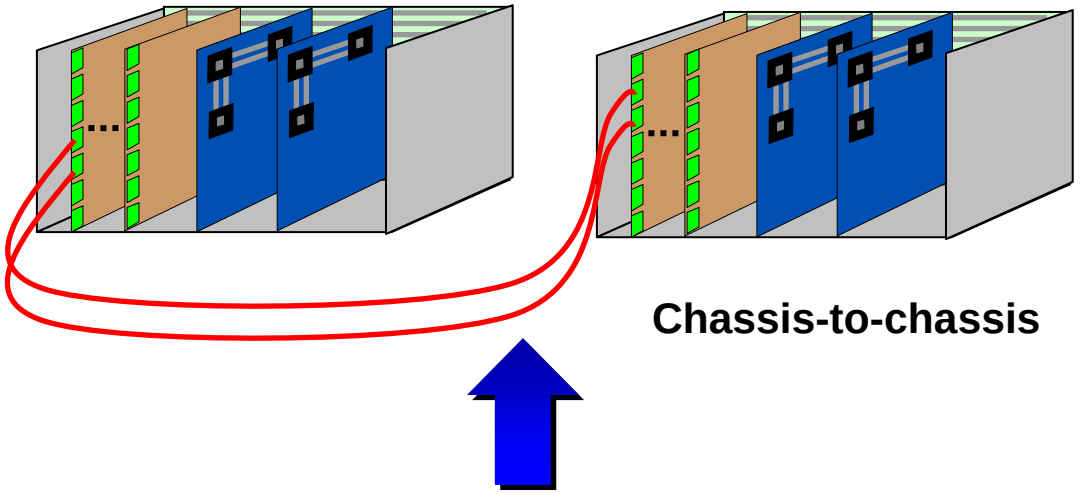
Example: VME

≤ 66MHz

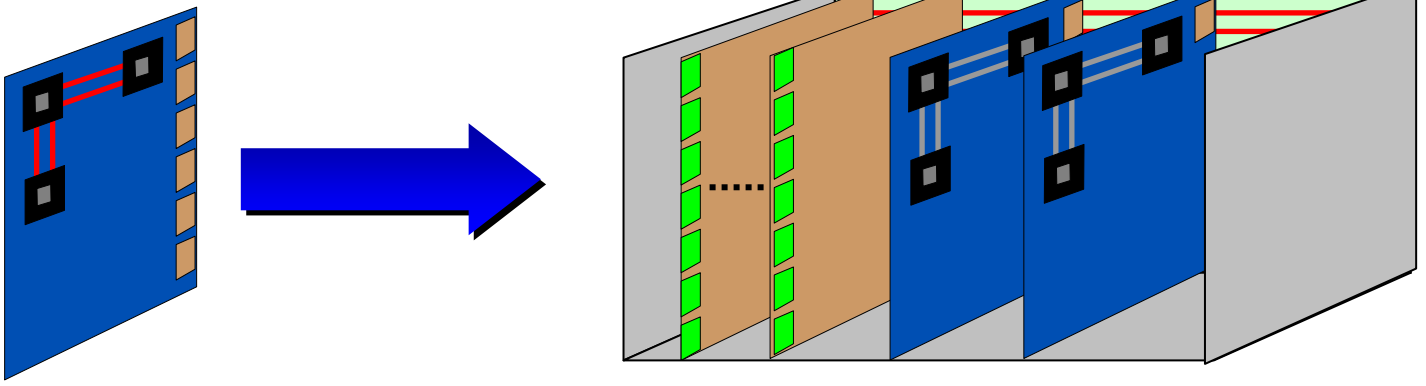
Performance

Interconnect Roles

- Interconnect roles
  - Chip-to-chip
  - Board-to-board (Backplane)
  - Chassis-to-chassis



Chassis-to-chassis

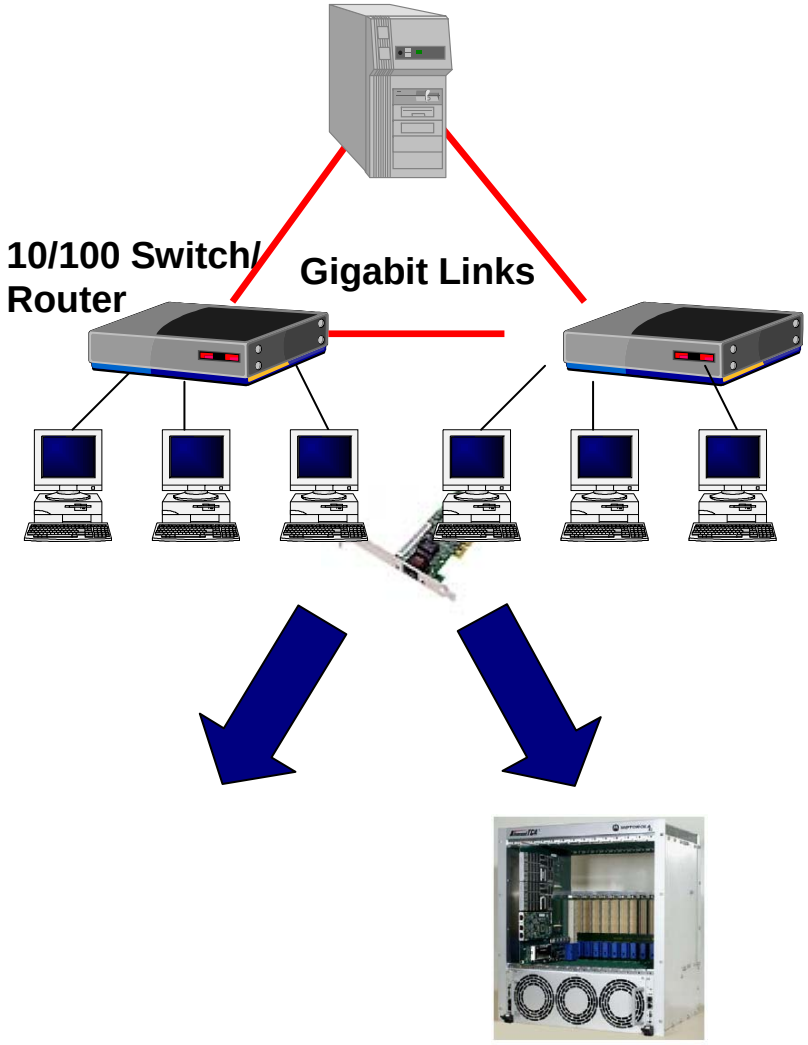


Chip-to-chip

Board-to-Board

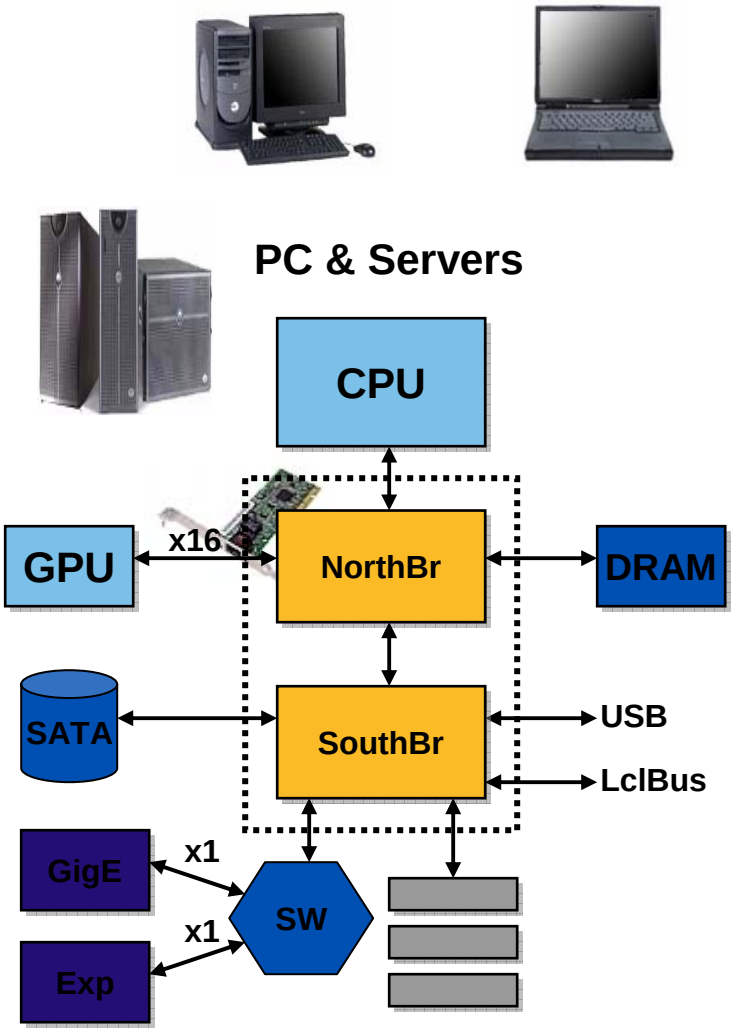
Market Focus: Gigabit Ethernet

- Successor to 100Mbps Ethernet
  - 10Ge in the wings
- First revision standard completed in 1998
  - Copper in 1999
  - Various MAC-to-PHY Standards defined
- Initial application in WAN
  - Aggregation
    - High performance switches , routers and serves in LAN backbones
  - Later WAN to workstations, PCs and laptops
- Positioning
  - Well known
  - “Safe Choice”



Market Focus: PCI Express

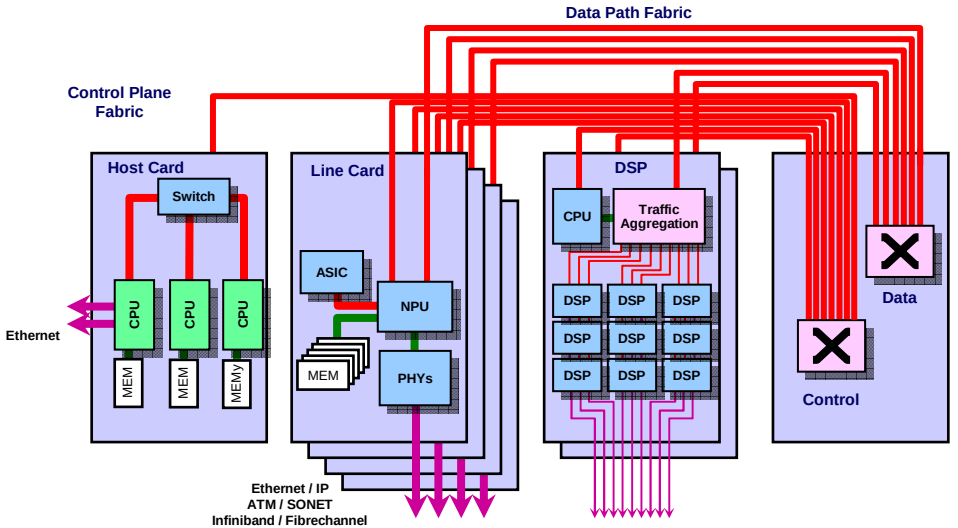
- Successor to PCI 2.3/PCI-X
  - Driven by Intel and PCISIG
  - Fully SW/firmware backward compatible to PCI
- First revision standard completed in 2002
  - Rollout this year in PC infrastructure
    - x86 chipsets, video cards, NICs
  - x1, x4, x8 and x16 most common
- Initial application in PCs and Servers
  - Follow-on to AGP8x for 3D graphics HW
  - GigE, 10GE NICs
  - Storage (RAID, FC), PCI bridges
- Positioning
  - Interconnect for PC and Servers space
  - Embedded if suitable



Market Focus: RapidIO

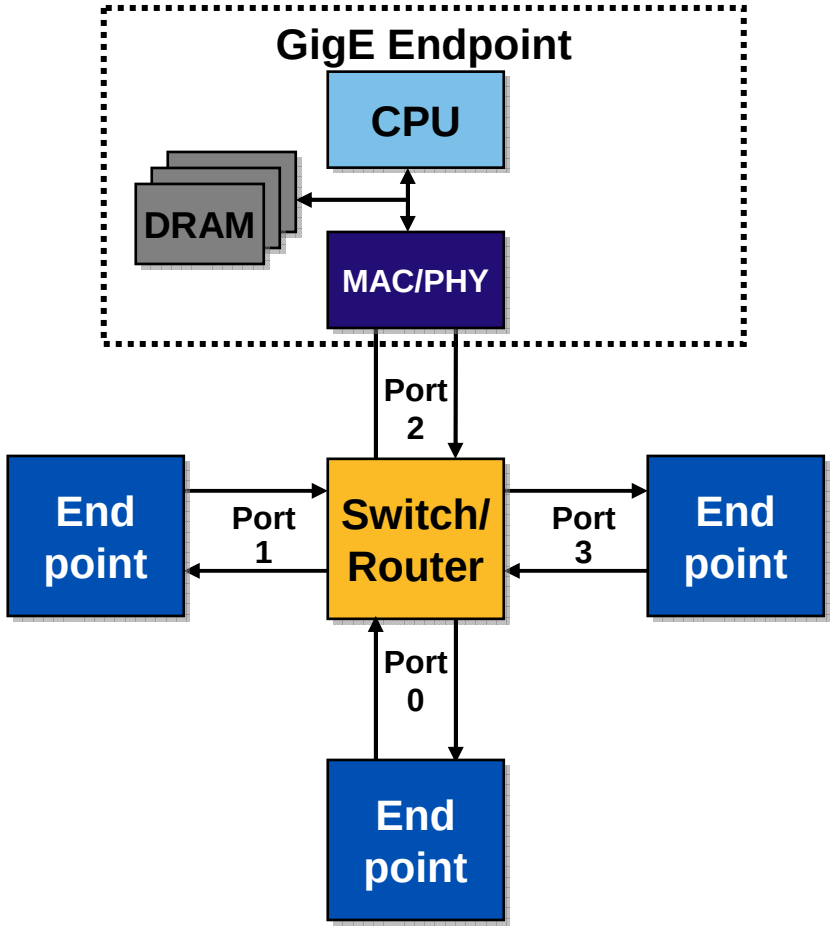
- Initially a processor interconnect
- First revision standard completed in 1999
  - Rollout in processors, bridges and switches
  - Parallel 8-bit RapidIO @ 500 MHz applied clock
- Initial application in embedded systems
  - Compute, defense, networking & telecom line cards
  - CPU I/O, Line-card aggregation, backplane
  - Serial PHY allowed expansion to data plane
    - Flow control, encapsulation, streams
- Positioning
  - Best solution for embedded systems
  - Suitable for both control and data plane

Embedded, Networking and Telecom

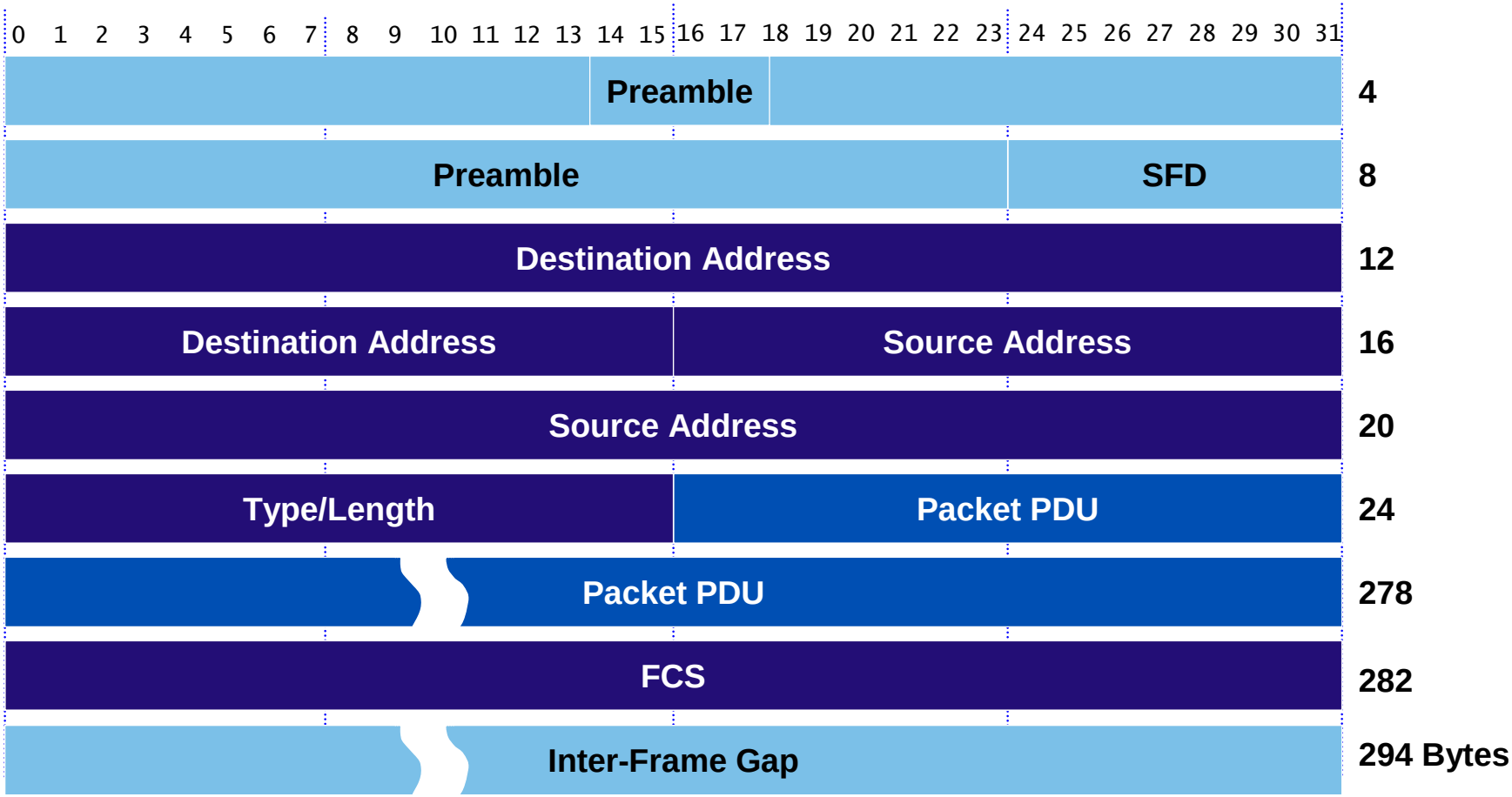


Gigabit Ethernet Overview

- WAN scale interconnect
  - Box-to-box, board-to-board, chip-to-chip, backplane
  - Connect world's computers
  - Physical layer defined for LAN-scale interconnection
    - Closet to computer
    - 100+ m distance
- Extensibility
  - Layered OSI Architecture
- Point-to-point packetized architecture
  - Variable packet size
  - High header overhead
  - 46-1500 byte packet L2 PDU
  - Up to 9000 byte jumbo frames



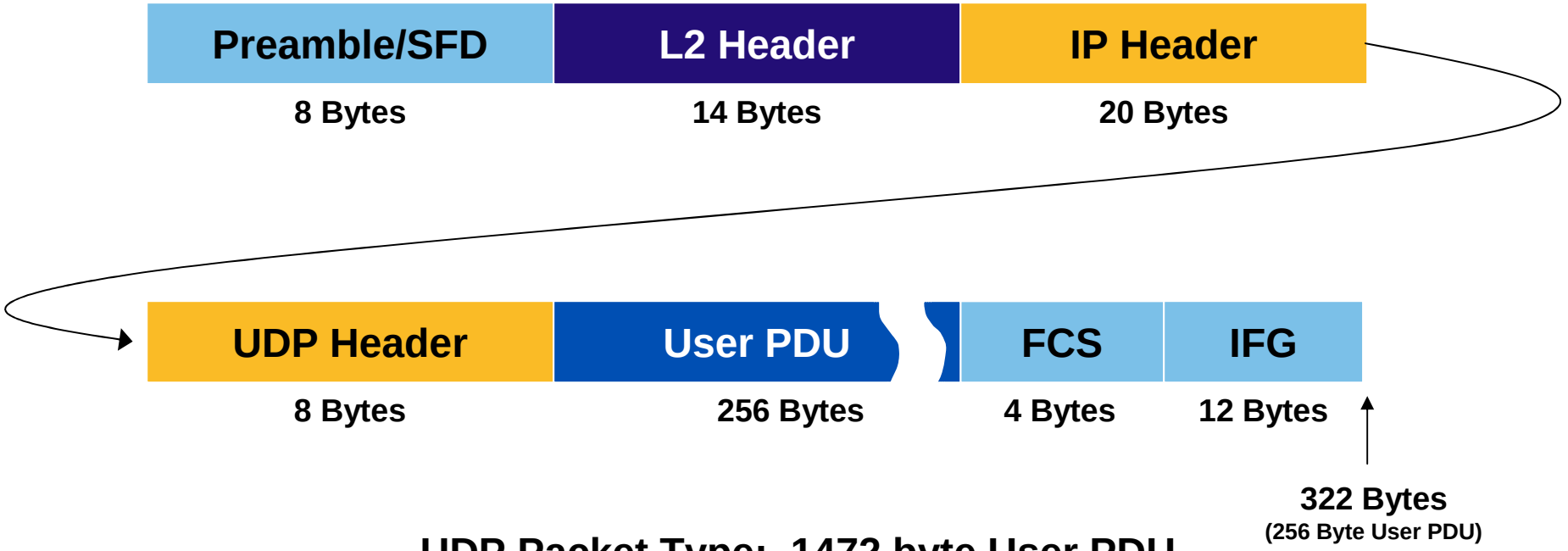
Ethernet Protocol: Layer 2



Layer 2 Packet Type: 1500 Byte Max Packet PDU

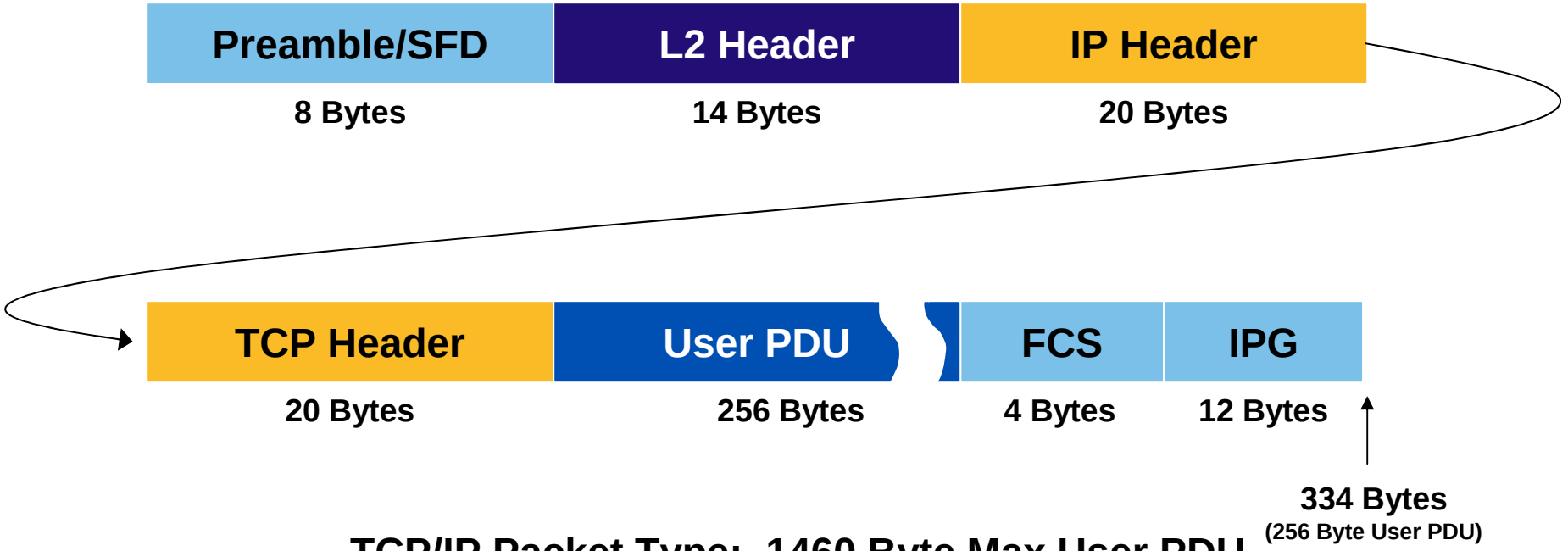
Total = 294 Bytes
(256 Byte PDU)

Ethernet Protocol: UDP w/Priority Tagging



UDP Packet Type: 1472 byte User PDU

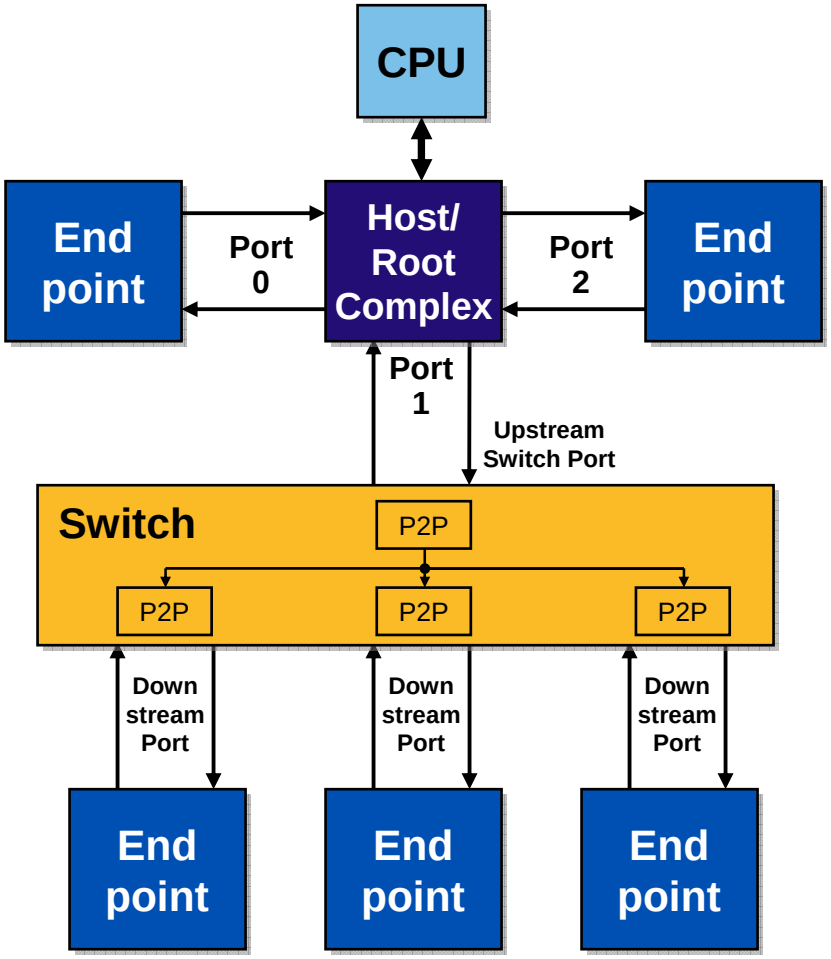
Ethernet Protocol: TCP/IP



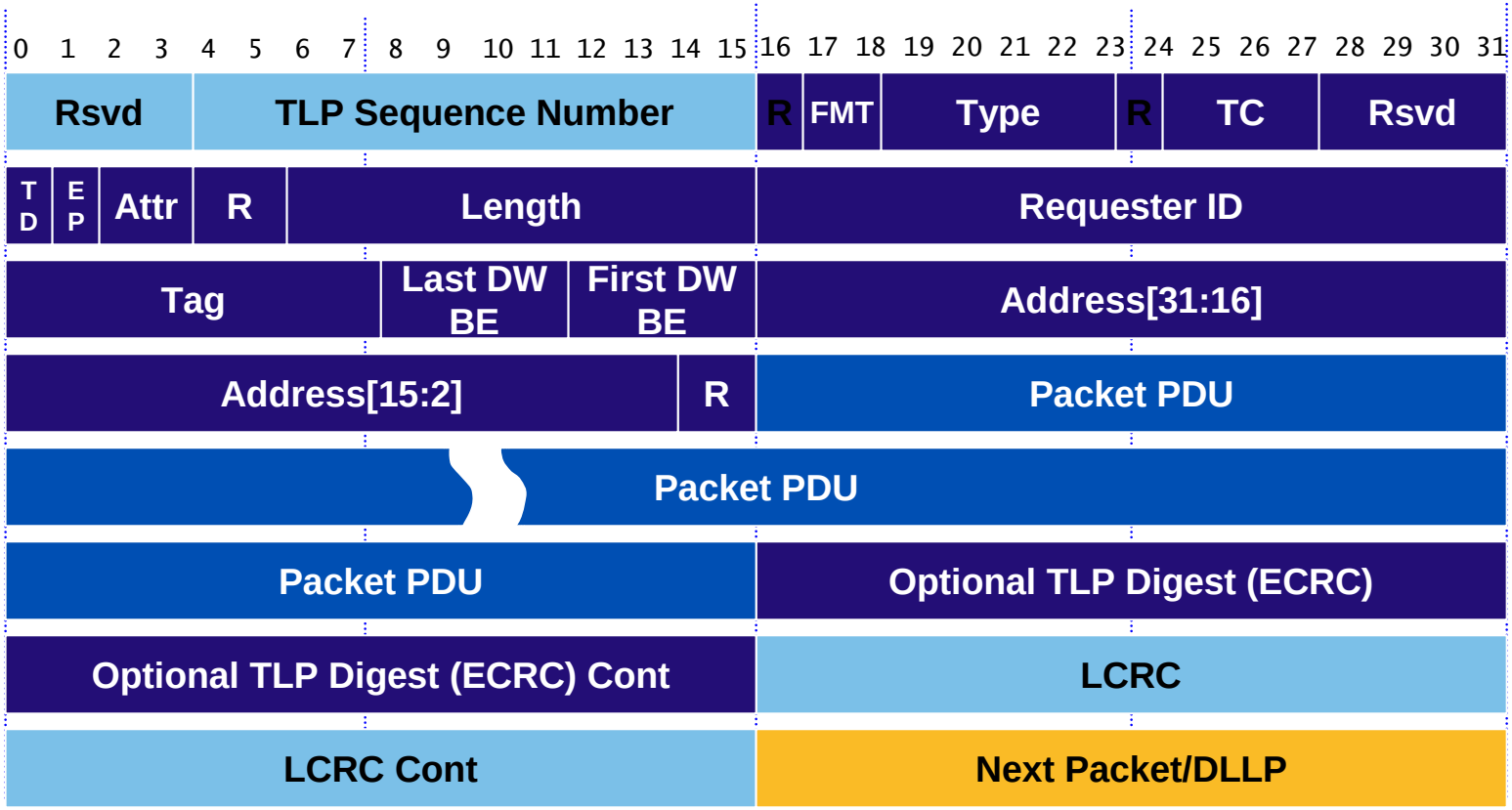
TCP/IP Packet Type: 1460 Byte Max User PDU

PCI Express Overview

- Chassis-scale interconnect
  - Chip-to-chip, Board-to-board via connector or cabling
  - Required legacy PCI compatibility
  - Physical layer defined for board + connector
    - ~40-50 cm + 2 connectors
- Extensibility
  - Layered architecture
- Point-to-point packetized architecture
  - Relatively low overhead
  - Variable size packets
  - 128-4096 byte PDU



PCI Express Protocol

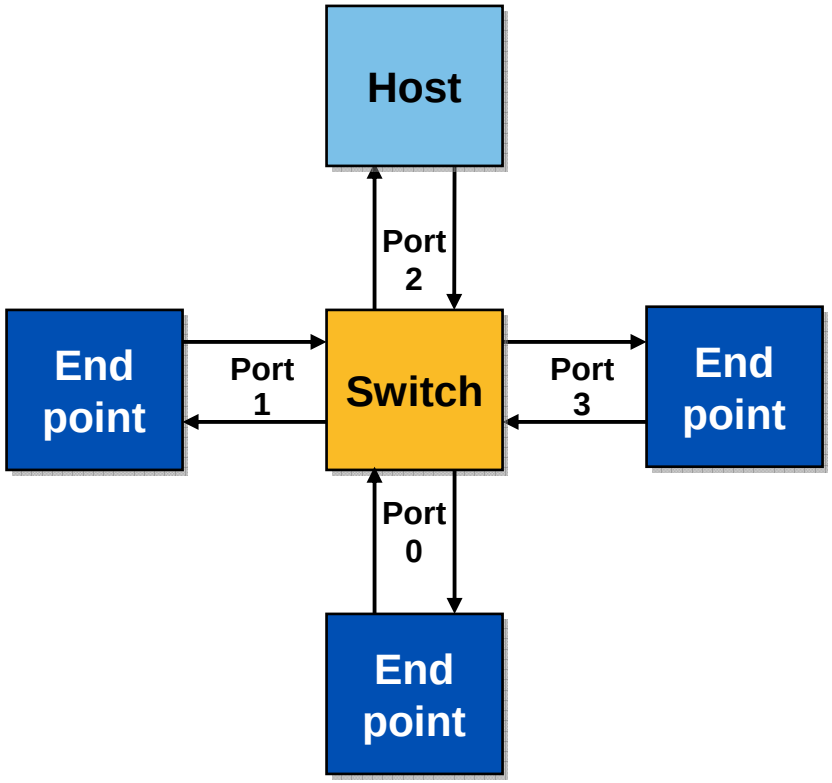


Memory Write: 4096 Byte Max Packet PDU

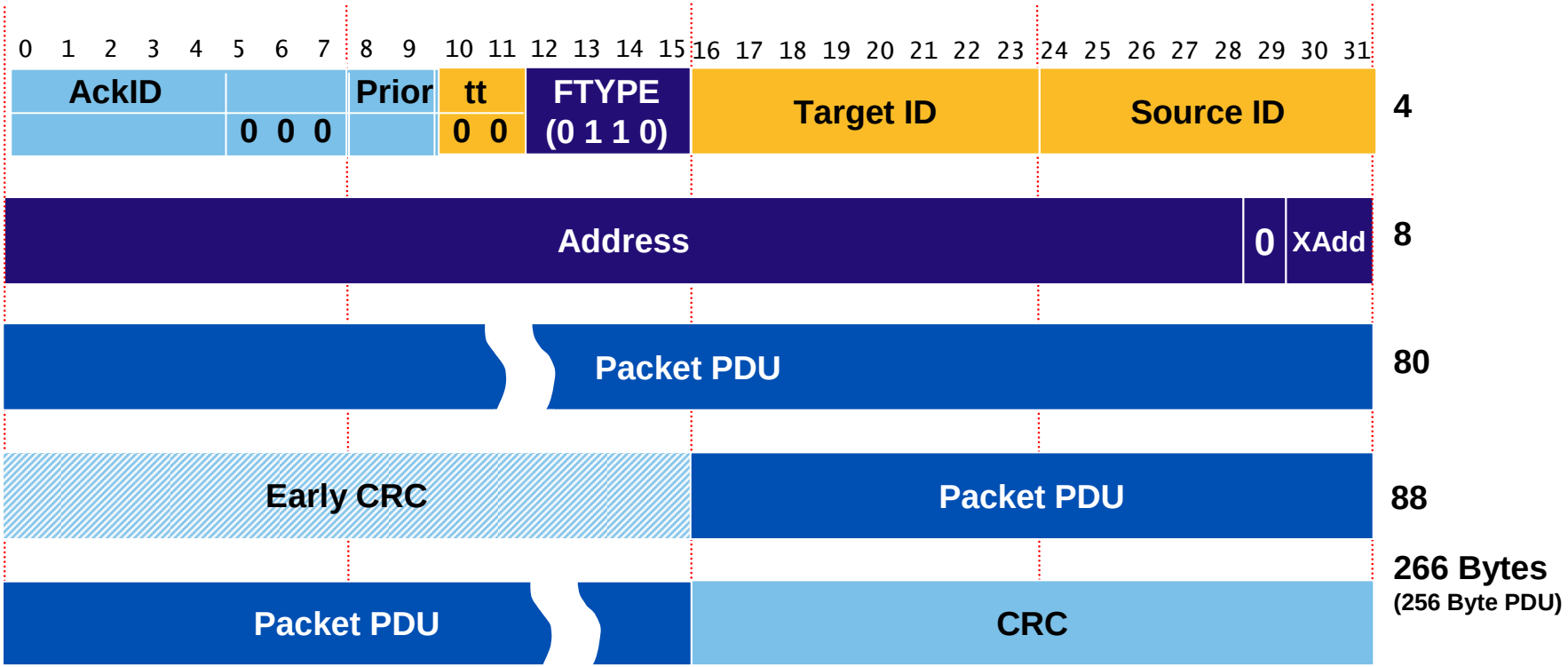
Total = 278 Bytes
(256 Byte PDU)

RapidIO Overview

- Chassis scale interconnect
  - Chip-to-chip, Board-to-board via connector or cabling
  - Physical layer defined for backplane interconnection
    - ~80-100 cm + 2 connectors (Serial)
- Extensibility
  - Layered architecture
- Point-to-point packetized architecture
  - Low overhead
  - Variable packet size
  - Maximum 256 byte PDU
  - SAR support for 4 K-byte messages

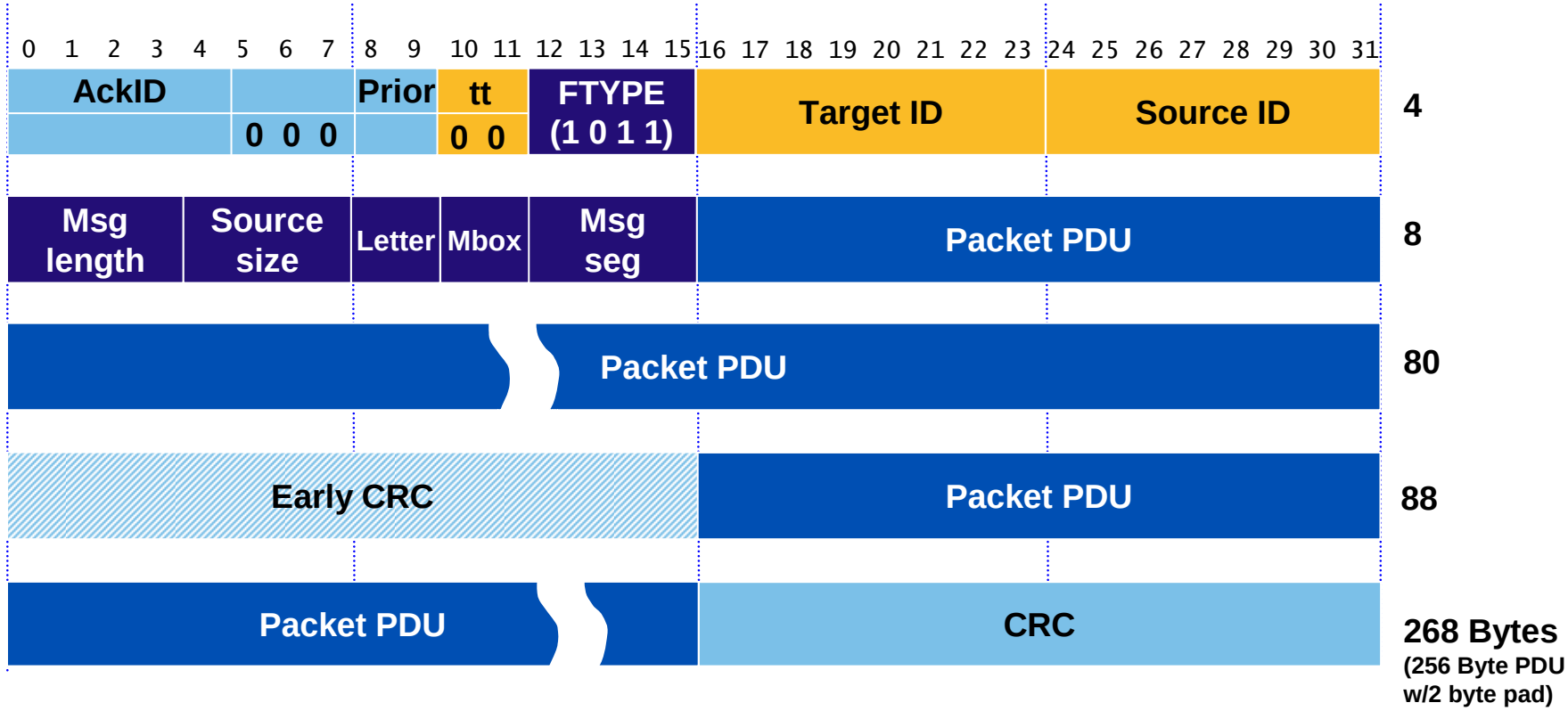


RapidIO Packet Format: SWRITE



SWRITE Packet Type: 256 Byte Max Packet PDU

RapidIO Packet Format: Message



Message Packet Type: 256 Byte Max Packet PDU, 4K w/SAR

Logical Layer Comparison

	GigE	PCI Express	RapidIO
Memory-mapped R/W	No	Read/Write Configuration	Read/Write Atomics Configuration
Write w/Response Support	No	No	Yes
Address Size	N/A	32, 64-bits	34, 50, 66-bits
Globally Shared Memory	No	No	Yes
Messaging Support	No	Interrupts and Event Signaling	Up to 4K Messages with HW SAR support, Doorbells
Datagram Support	Up to 1500 byte user payloads	No	HW SAR up to 64Kbyte user payloads

Transport Layer Comparison

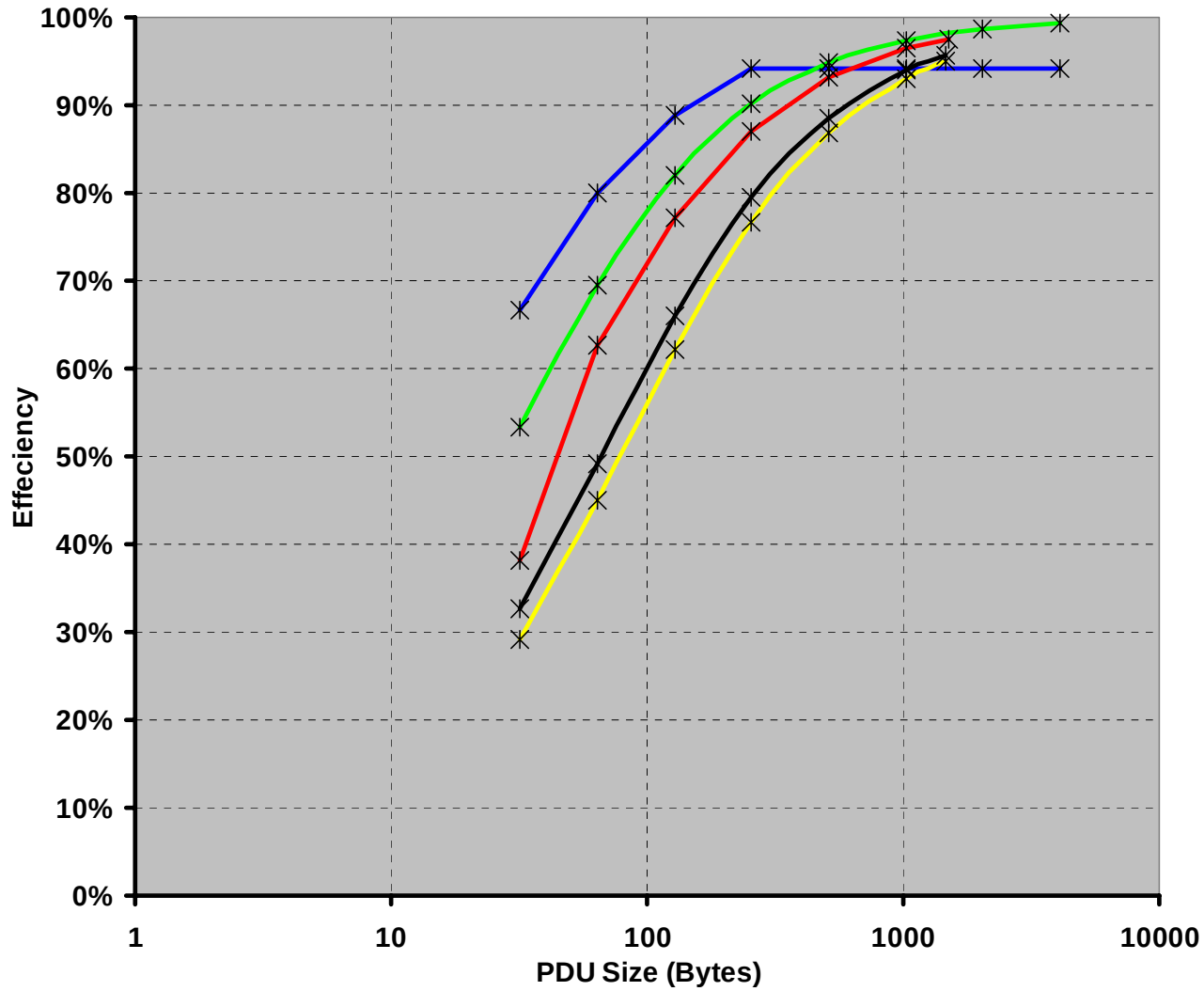
	GigE	PCI Express	RapidIO
Topologies	Any	Tree	Any
Number of endpoints	2^{48} (L2) 2^{32} (IPv4) 2^{128} (IPv6)	Large (Address-based)	2^{8} (Small) 2^{16} (Large)
What fields must switches modify?	L2: None IP: TTL, MAC, FCS	TLP, Seq Num, LCRC	AckID
Multicast	Yes	Yes (Message only)	Yes
Delivery	L2: Best Effort IP: Guaranteed	Guaranteed	Guranteed, Best Effort

Physical Layer Comparison

	GigE	PCI Express	Parallel RapidIO	Serial RapidIO
Tx/Rx Signal Pairs	4x†	1x, 2x, 4x, 8x, 12x, 16x, 32x	10 bits 19 bits	1x, 4x, 8x*, 16x
Channel	100 m cat5	~40-50 cm + 2 connectors	~50-80 cm + 2 connectors	~80-100 cm + 2 connectors
Data rate	10, 100, 1000 Mbps	2.5 GBaud	500-2000 MHz	1.25, 2.5, 3.125 GBaud
Signaling	4D-PAM5 MLS	Proprietary AC Coupled	LVDS	XAUI AC Coupled
Clocking	Embedded	Embedded	Clock + Data	Embedded
Latency	Highest	Higher	Lowest	Next Lowest

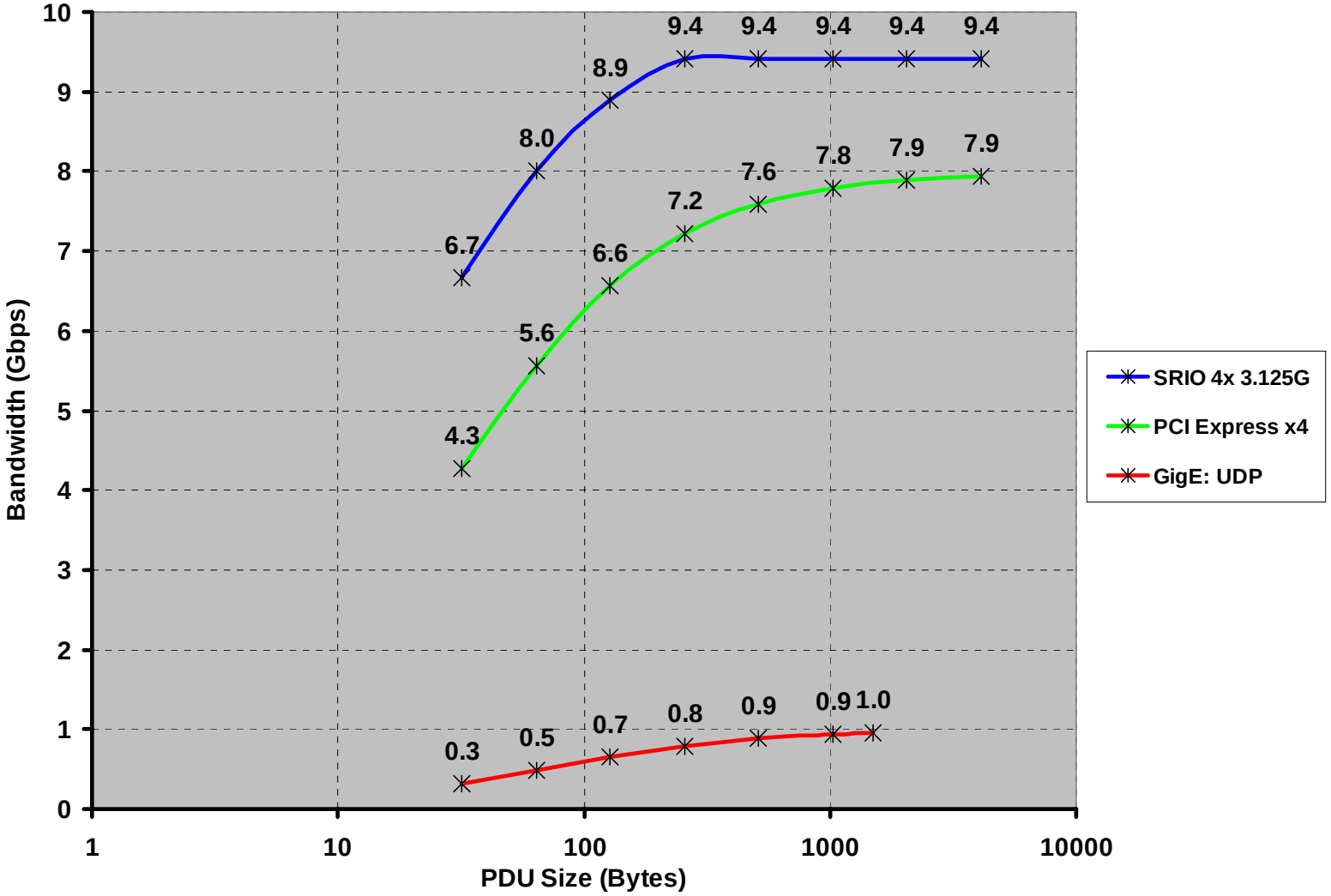
†Tx,Rx are on same wire pairs

Protocol Efficiency



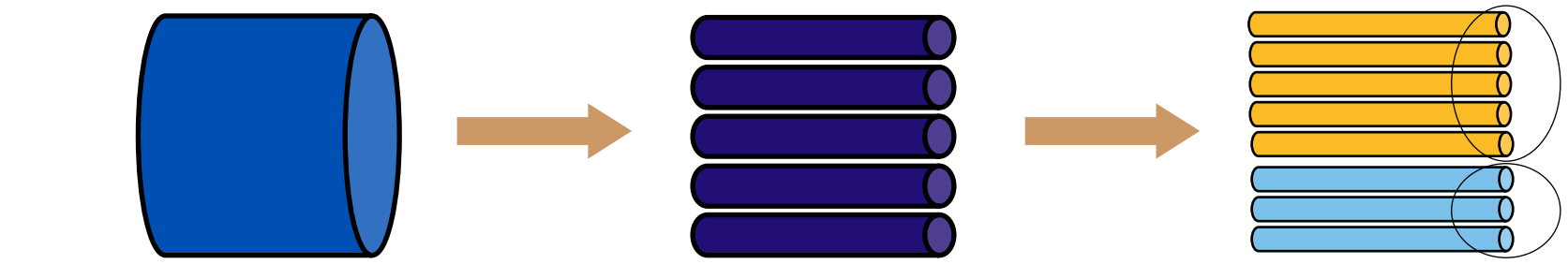
NOTE: SWrite for RapidIO, MWr for PCI Express. Includes header and ACK overhead

Effective Bandwidth



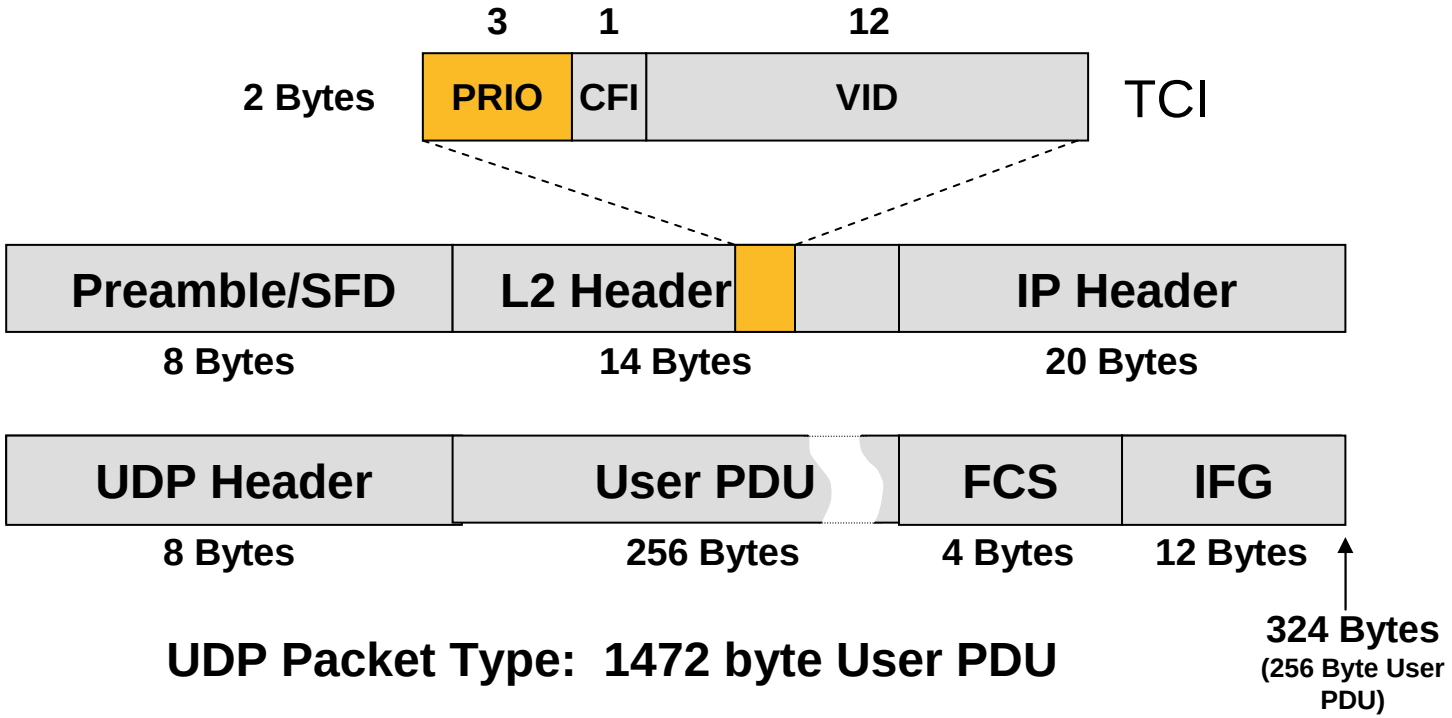
Quality-of-Service (QoS) Dependencies

- QoS depends on proper hooks across the interconnect fabric
  - Hierarchical Flow Control
    - Addresses short, medium and long-term congestion events
    - Link and end-to-end
  - Ability to define many streams of traffic
    - Often defined as a logical sequence of transactions between two endpoints
  - Ability to differentiate classes of traffic among streams
  - Ability to reserve and allocate bandwidth to streams and classes



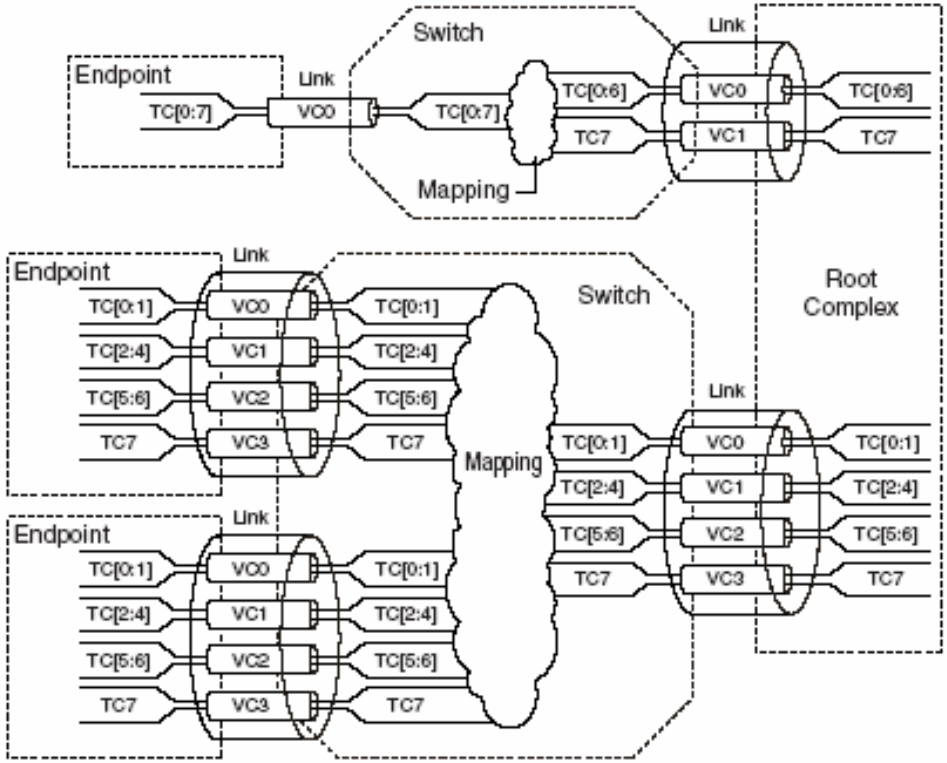
QoS Comparison: Gigabit Ethernet

- No universal QoS standard
- Some Layer 2+ switches support Priority Tagging (802.1d/q)
  - Eight classes
- Increasing number of routers support MPLS at L3



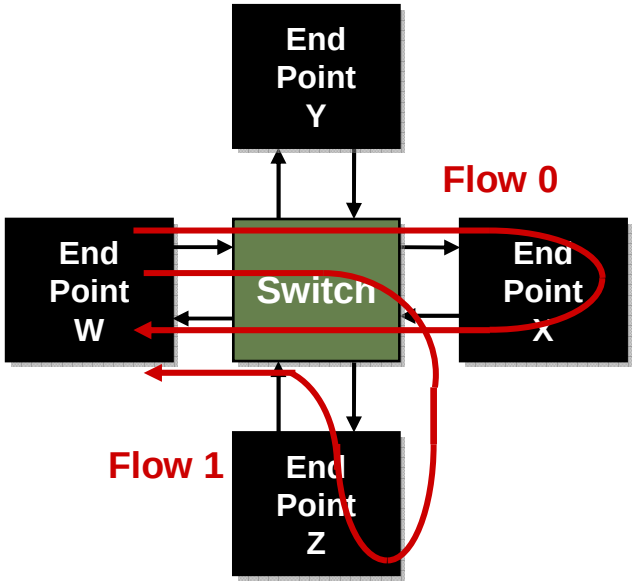
QoS Comparison: PCI Express

- 8 Traffic Classes (TC)
  - No ordering between TCs
- 8 Virtual Channel (VC)
  - Separate buffer resources per VC
  - TCs are mapped onto VCs
    - TC to VC mapping per port
      - No VC field in TLP
- Flexible arbitration
  - Arbitrary, RR, WRR
- Most implementations support only a single TC/VC



QoS Comparison: RapidIO

- All implementations must support 3 prioritized flows
  - No ordering between flows
  - Allows shared buffer pool across flows
- Switches required to provide some improved service
  - Extent of improvement is implementation dependant
- Supports carrier-grade level QoS
  - Support for 1000s of flows, hundreds of traffic classes
  - End-to-end traffic management



Flow Control Comparison

Gigabit Ethernet

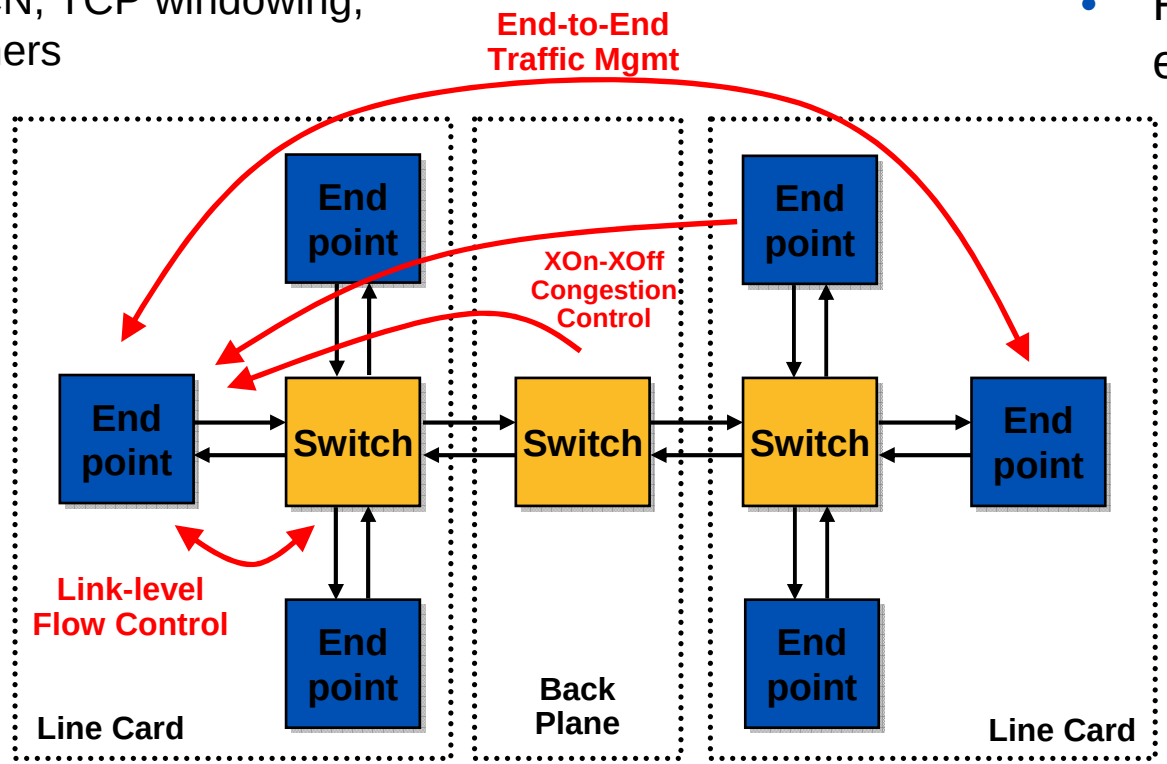
- Link-to-link flow control
  - PAUSE frames
- L3+ end-to-end flow control
  - ECN, TCP windowing, others

PCI Express

- Link-to-link flow control

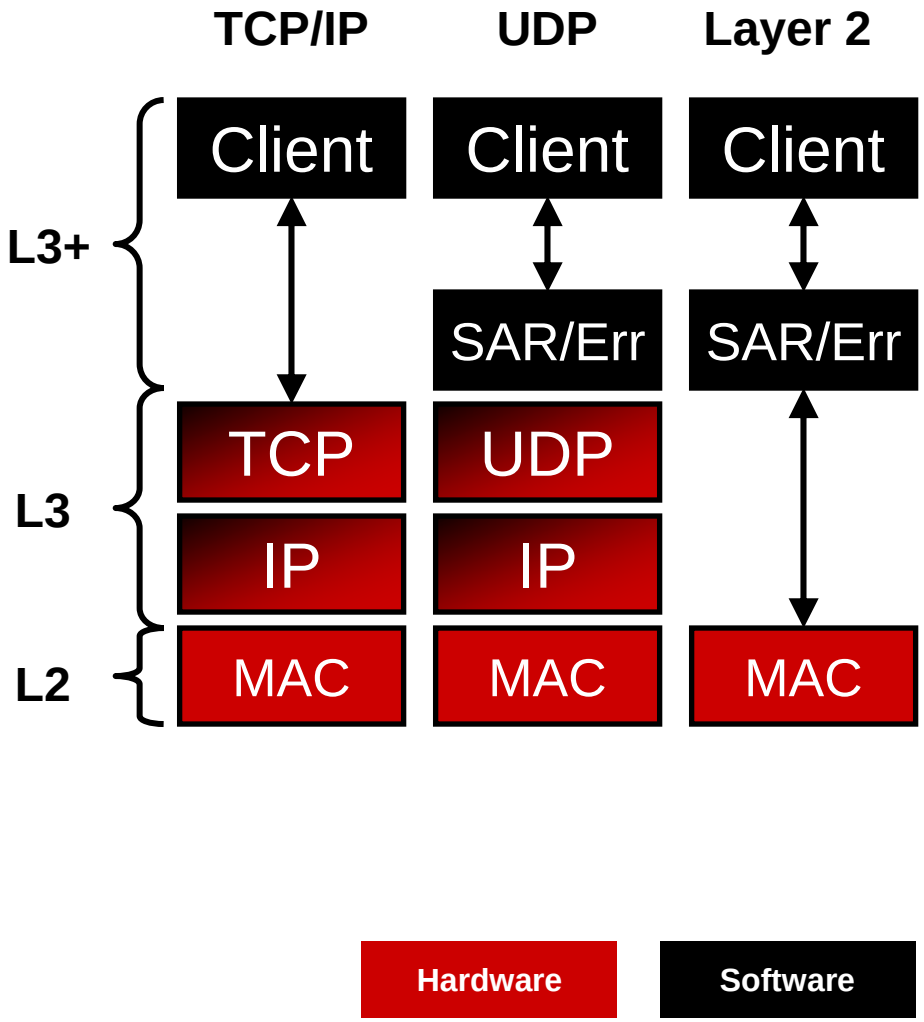
RapidIO

- Link-to-link flow control
- Congestion control
  - XON, XOFF
- Fine-grained end-to-end flow control
  - Data Streaming Logical Layer



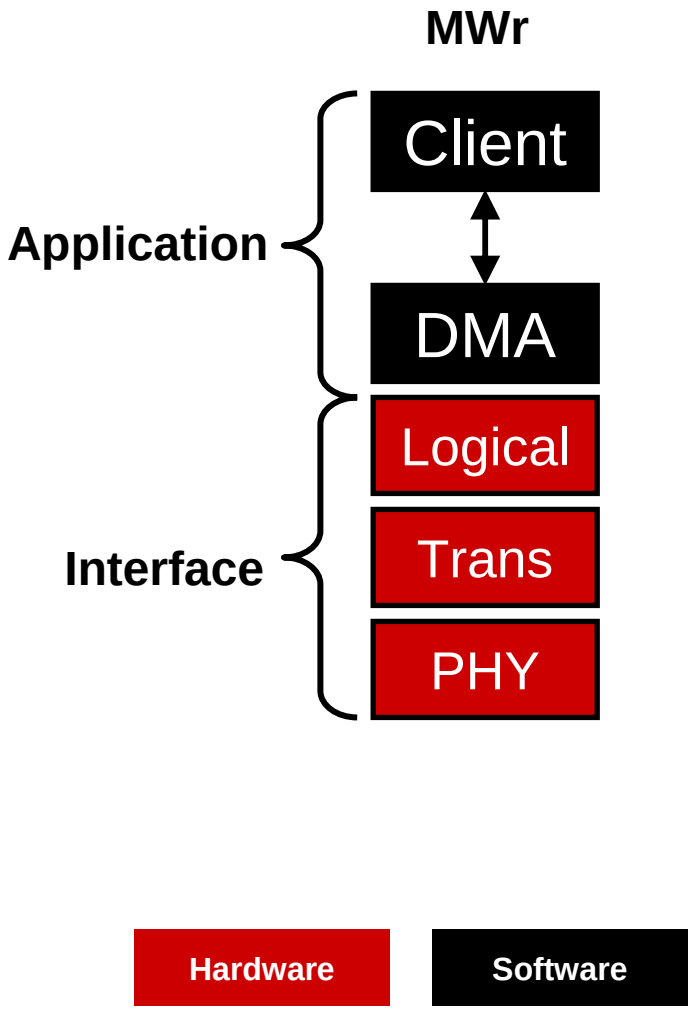
GigE Usage Models

- Control plane
  - Sockets and TCP/IP
    - Protocol encapsulation
    - SAR support
    - Guaranteed delivery
  - Custom SW stack?
- Data plane
  - Custom UDP-like stack?
  - UDP
    - Flow multiplexing via port number
    - Address-less datagrams
    - No SAR support
    - Multicast/Broadcast support
    - Best effort
  - MAC-layer with L2 switching
    - Address-less datagrams
    - No SAR support
    - Best effort



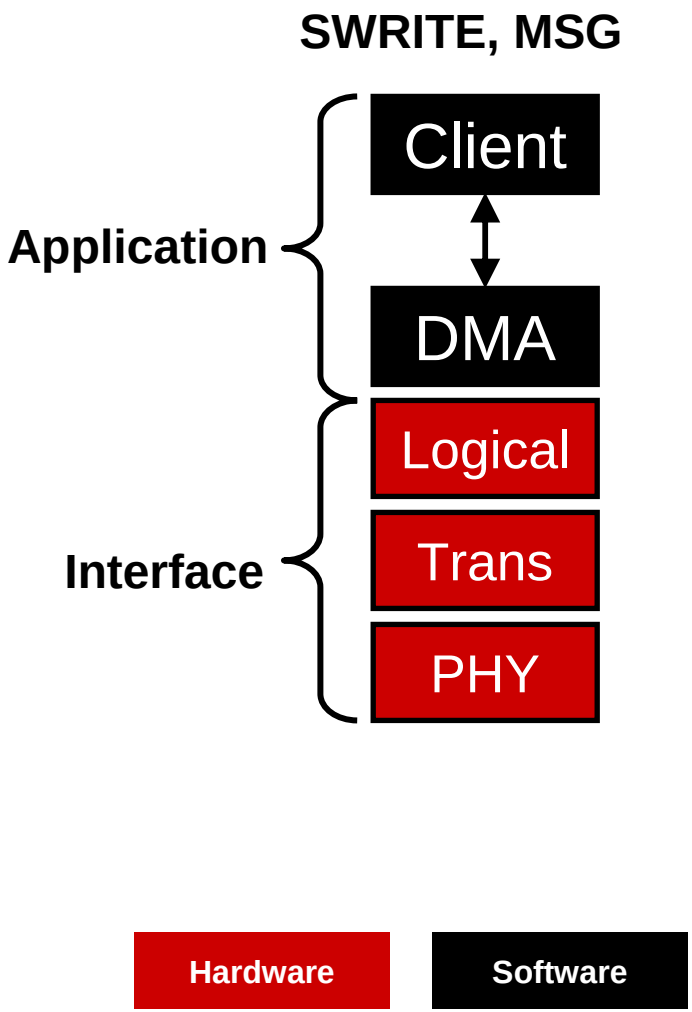
PCI Express Usage Models

- Control plane
  - CPU load/store & DMA
  - Interrupts and Event signaling
- Data plane
  - Address-based read & writes
    - DMA completion requires notification transaction
      - No Write w/Response
    - Hardware-based error recovery and ACK (lossless)



RapidIO Usage Models

- Control plane
  - Address-based reads, writes and messaging
  - CPU load/store & DMA
- Data plane
  - Address-based read & writes
    - DMA completion requires notification transaction
    - Hardware-based error recovery and ACK (lossless)
  - Messaging
    - Address-less datagrams
    - Hardware-based SAR
    - 4K Byte max user PDU size
    - Hardware-based error recovery and ACK (lossless)
  - RapidFabric
    - Address-less datagrams
    - Hardware-based SAR
    - Up to 64KB user PDU size
    - Lossy



Performance: Gigabit Ethernet

- Microsecond+ fall through latencies (~100us?)
  - Not just the hardware, data has to traverse the SW stack
- High CPU overhead
  - Rule of thumb appears to be borne out in data for TCP/IP SW overhead
    - 1 Hz of CPU per bit of throughput (per direction)
  - Wire speed achievable with GHz class processors
    - Some CPU will be left but how much depends on
      - Protocol being terminated
      - Offload features of GigE interfaces
  - Too often advanced TOE features cannot be leveraged
    - OS & SW stack support issues
  - Use of UDP or MAC/Layer 2 often involve proprietary protocols
    - Can defeat the value of off-the-shelf “standards-based” solution
- Error correction at endpoint stacks introduce latency jitter and determinism issues
- Lack of end-to-end flow control problematic for non-traffic managed systems that can’t significantly overprovision

Performance: PCI Express & RapidIO

- Latency
  - Sub-microsecond switch latencies
    - PCI Express switches must do complicated address comparisons
  - End-to-end latency
    - Lower latency than GigE since latency does not include a SW stack
- Architecture
  - PCI Express switches allow limited but not complete peer-to-peer communication
    - Multiple hosting for redundancy problematic
      - Maintenance transactions cannot move upstream or peer-to-peer
      - Proposed switches use non-transparent bridges as work around (i.e., create two separate spaces for each host)
  - PCI Express systems with multiple hosts must use switches with non-transparent bridges
    - Bridging is non-standard and implementation specific
  - RapidIO switches can be simple and orthogonal in architecture
    - Header architected to reduce logic
    - No need to recalculate CRC

Efficiency and Throughput

- Efficiency
  - RapidIO has significant advantage in header efficiency
    - Especially true when using Layer 3+ to transport user PDUs smaller than 128 bytes
  - Control plane traffic often bursty and small packet oriented
    - Effective utilization of GigE likely to be low on this basis alone
- Throughput
  - What percentage of raw interface bandwidth can be utilized?
    - Flow control mechanisms key in increasing utilization
    - Well under 50% typical for Ethernet to avoid packet loss
    - Only RapidIO has a full range of flow-control mechanisms
  - Throughput can be limited by bottlenecks unrelated to interconnect
    - Smaller packet sizes involve increasing overhead per byte of data
      - More buffer descriptor fetches
      - More header overhead
      - More interrupts
    - Memory and bridging device bottlenecks
    - Extensive protocol termination requires CPU cycles

Some Economics

- RapidIO, PCI Express and Gigabit Ethernet with some TCP/IP offload have similar underlying silicon costs
  - PCI Express logic size is larger than RapidIO
  - Aggressive TCP/IP Offload engine larger than PCI Express and RapidIO endpoints
  - GigE Copper PHY is very large (~20mm2 in 130nm)
- Leveraging Ethernet volume economics is rarely a reality
  - L2+ GigE switches suitable for aggregation and backplanes are not high volume
    - 12-16 ports, QoS features and SERDES PHYs
  - Terminating TCP/IP demands significant processor overhead
    - Dedicate processor or reduce performance and/or application features
- RapidIO has the underlying economics to offer better performance at a price competitive with PCI Express and GigE

Pros & Cons

Gigabit Ethernet

- Pro
  - Well understood and hence low risk
- Con
  - High overhead
  - High latency and latency jitter (i.e. poor determinism)
  - Significant cost jump for bandwidth above 1Gbps
  - No standard backplane SERDES PHY
  - No standard HW acceleration

PCI Express

- Pro
  - Long-term use in PC-related HW
  - Long-term role as legacy chip-to-chip interconnect
- Con
  - Unsuitable connecting more than a few devices
  - Higher protocol overhead than RapidIO
  - Features not driven by embedded requirements
  - No data plane support

RapidIO Technology

- Pro
  - Strong system interconnect solution
  - Full QoS and Flow Control features
  - Arbitrary topologies supported
  - Control and data plane support
  - Lowest overhead with minimal silicon footprint
- Con
  - Not intended for PC & Server space

Conclusion

- **RapidIO Technology** will expand its existing role as standard system fabric
  - Efficient protocol supporting both control and data plane
  - Variety of PHY speeds
  - Cost competitive underlying economics
  - Available now
- **Gigabit Ethernet** will serve a limited role as a system interconnect
  - Low performance settings where significant over provisioning is possible
- **PCI Express** will remain largely within the PC and Server space and have a limited role in the embedded space
  - Where there is an intersection with the PC & Server space
  - In places where PCI is used today
  - Rarely as fabric due to handling of large numbers of endpoints

RapidIO® www.rapidio.org