# Platform IO DMA Transaction Acceleration

**ICS/CACHES**

Steen Larsen (steen.larsen@intel.com)
Ben Lee (benl@eecs.oregonstate.edu)
June 4 2011

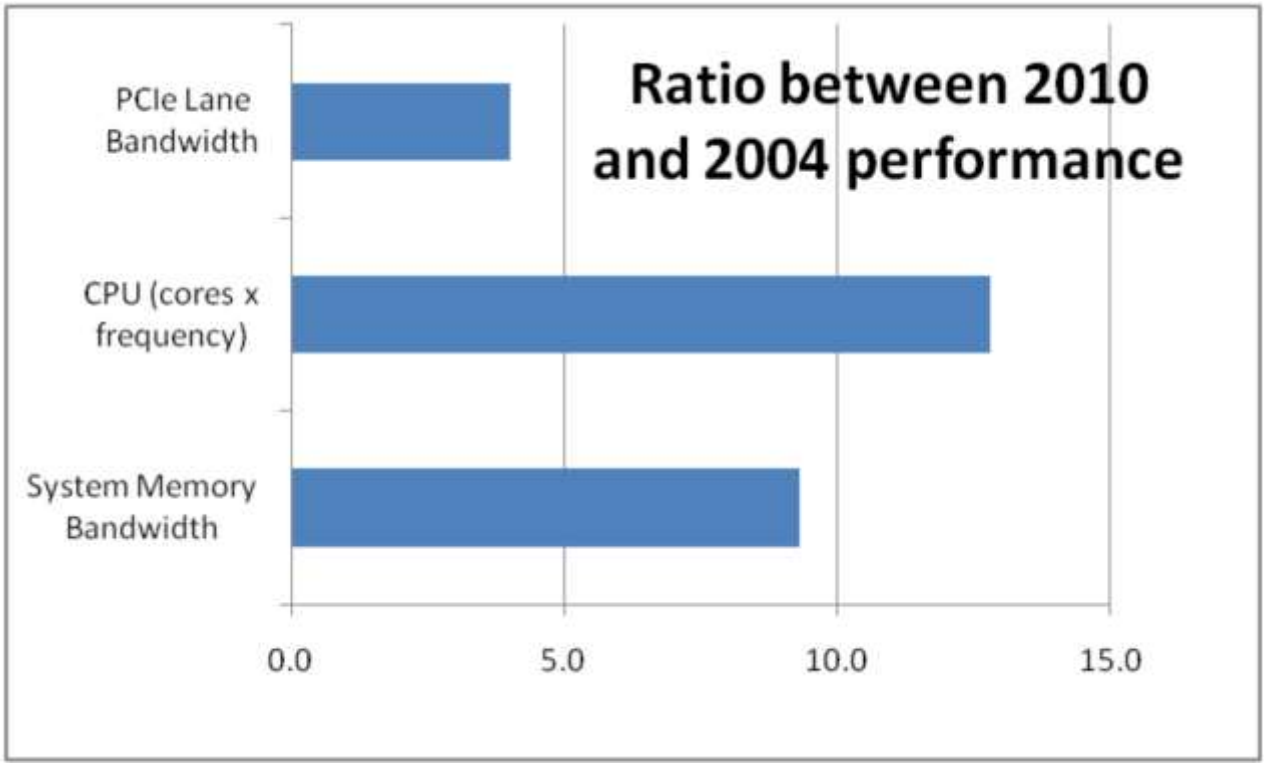
# Outline

- Introduction & Motivation
- Background
- Proposal
- Experiments & Analysis
- Related & Future work

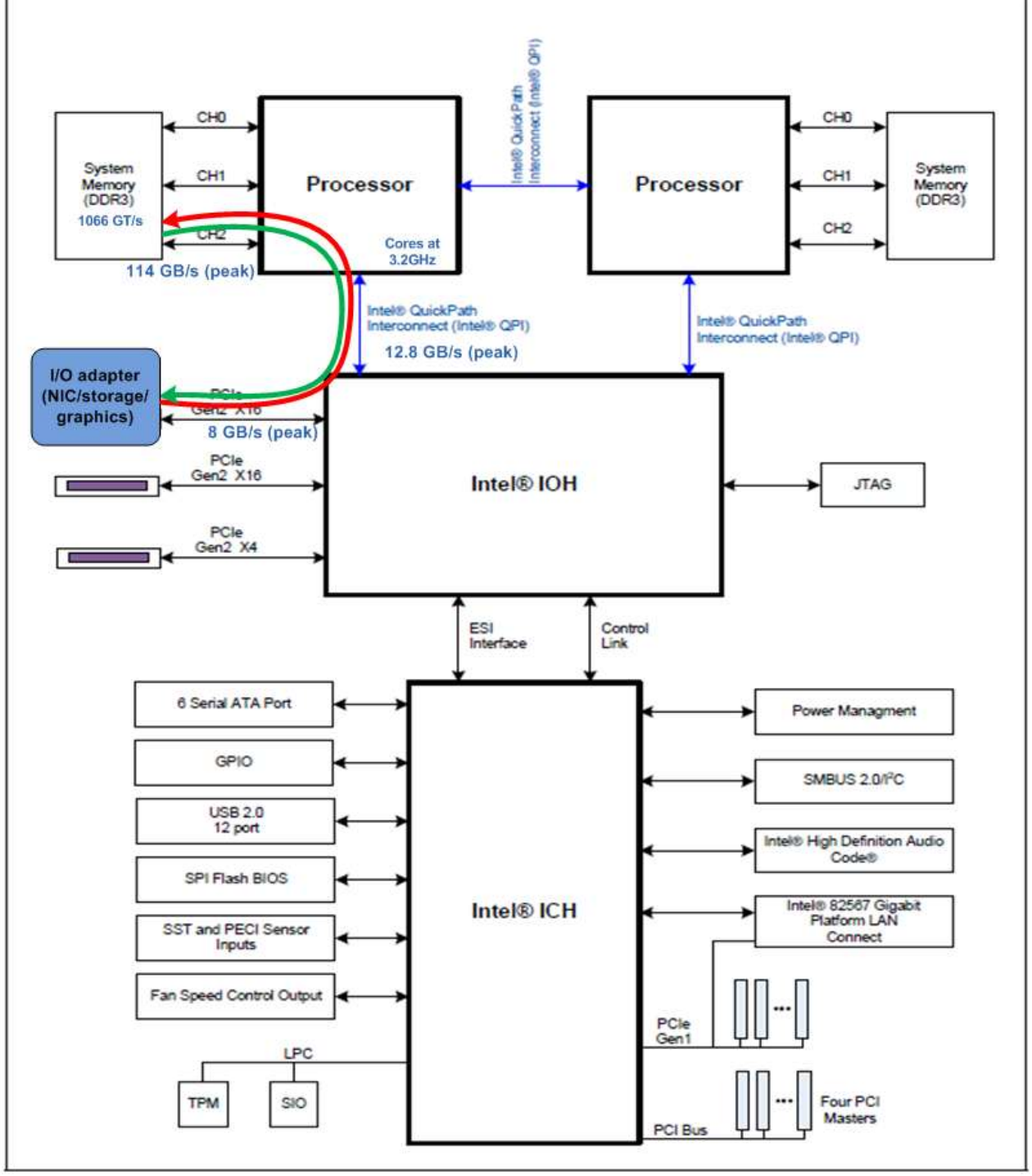
# 10,000 foot view of IO

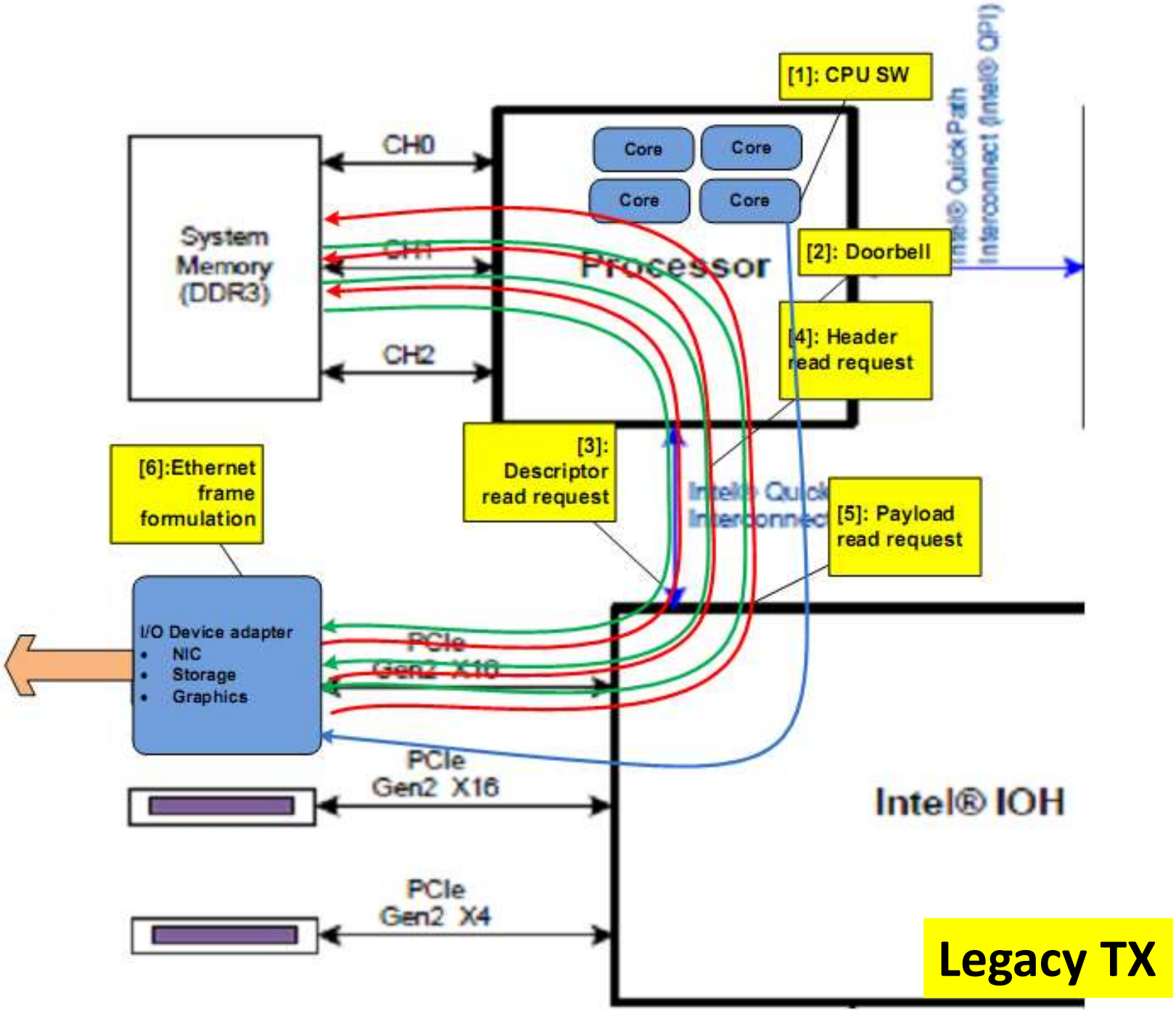
IO growth is not matching CPU and memory bandwidth growth.

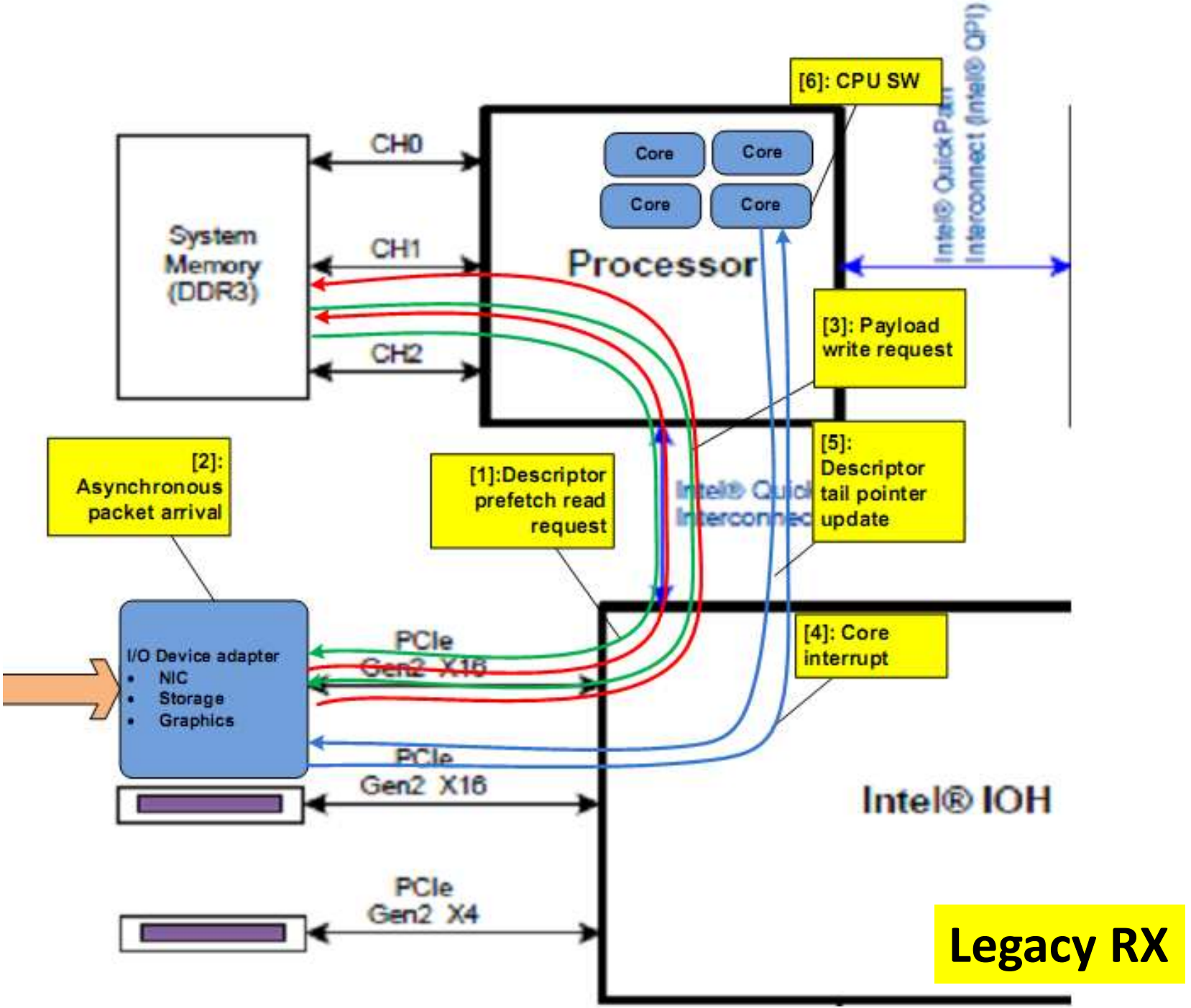
- Multi-core processors (CMP, SMT)
- NUMA



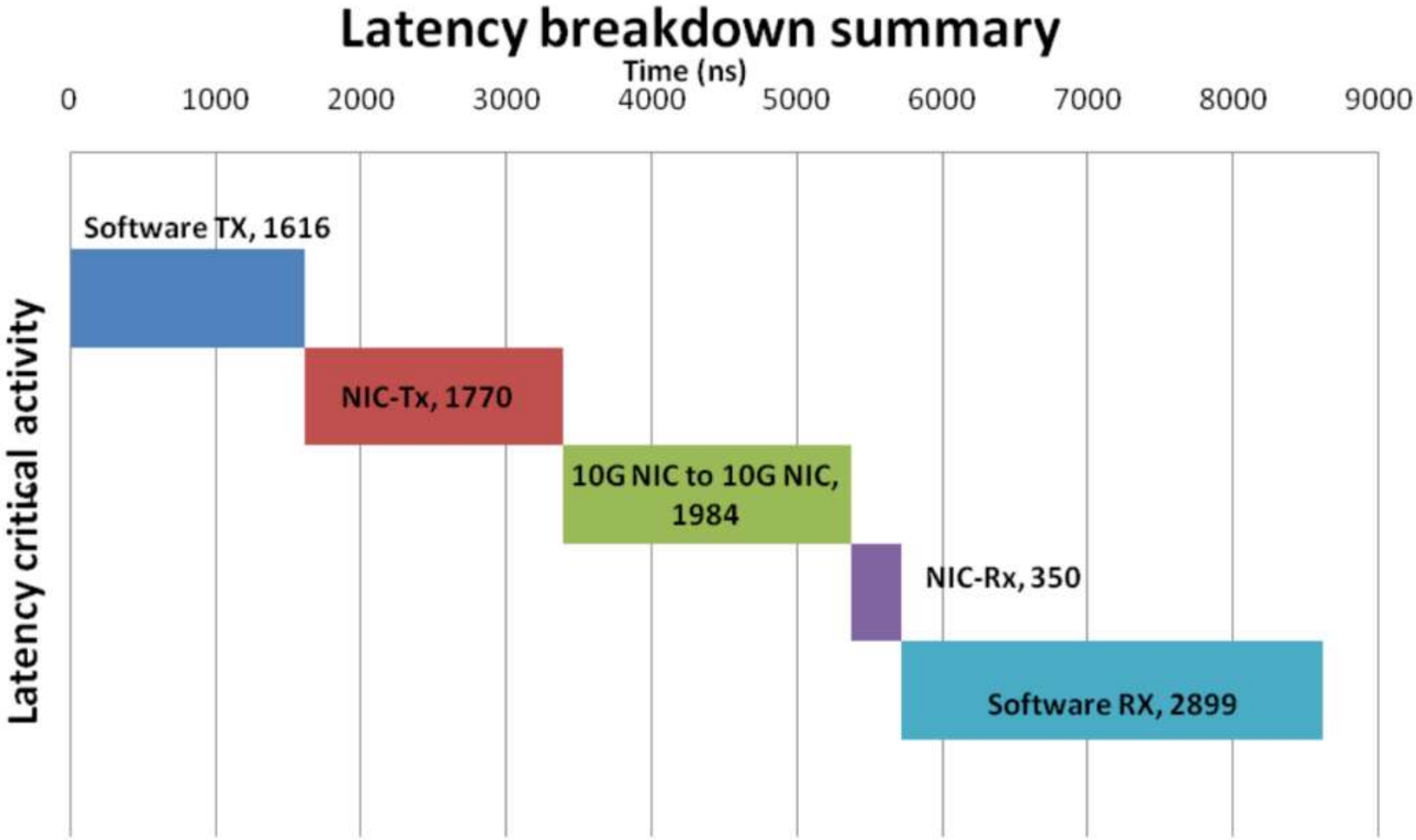
# Typical platform configuration and IO interface



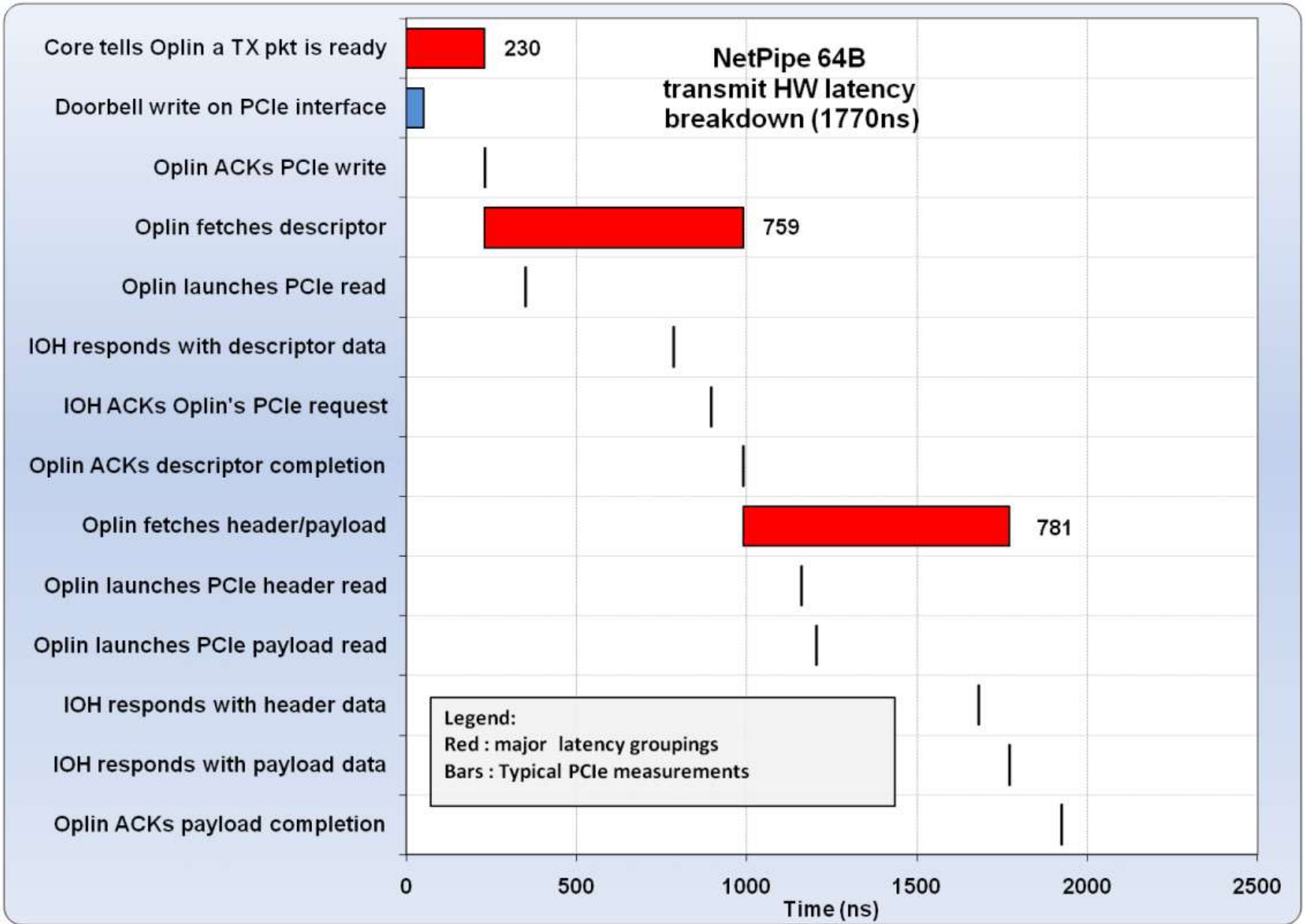

[1]: CPU SW
CH0
CH1
CH2
System Memory (DDR3)
Core
Core
Core
Core
Processor
[2]: Doorbell
[4]: Header read request
[3]: Descriptor read request
[5]: Payload read request
[6]:Ethernet frame formulation
I/O Device adapter
• NIC
• Storage
• Graphics
PCIe Gen2 X16
PCIe Gen2 X16
PCIe Gen2 X4
Intel® IOH
Legacy TX


[6]: CPU SW
Intel® QuickPath Interconnect (Intel® QPI)
CH0
CH1
CH2
System Memory (DDR3)
Core
Core
Core
Core
Processor
[3]: Payload write request
[2]: Asynchronous packet arrival
[1]:Descriptor prefetch read request
[5]: Descriptor tail pointer update
[4]: Core interrupt
I/O Device adapter
• NIC
• Storage
• Graphics
PCIe Gen2 X16
PCIe Gen2 X16
PCIe Gen2 X4
Intel® IOH
Legacy RX

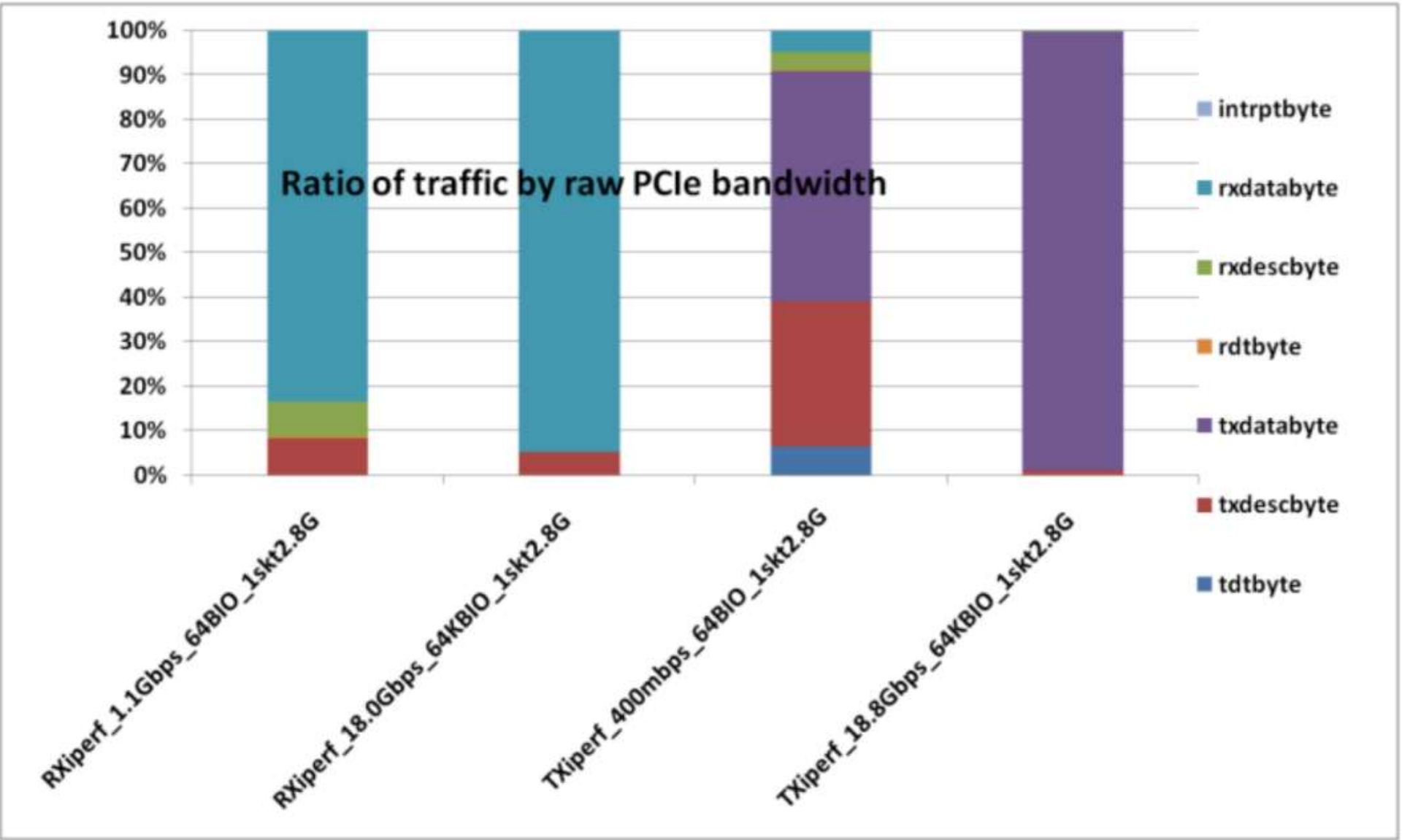
# Critical path latency (10GbE 64B)



# IO transmit breakdown (10GbE 64B)

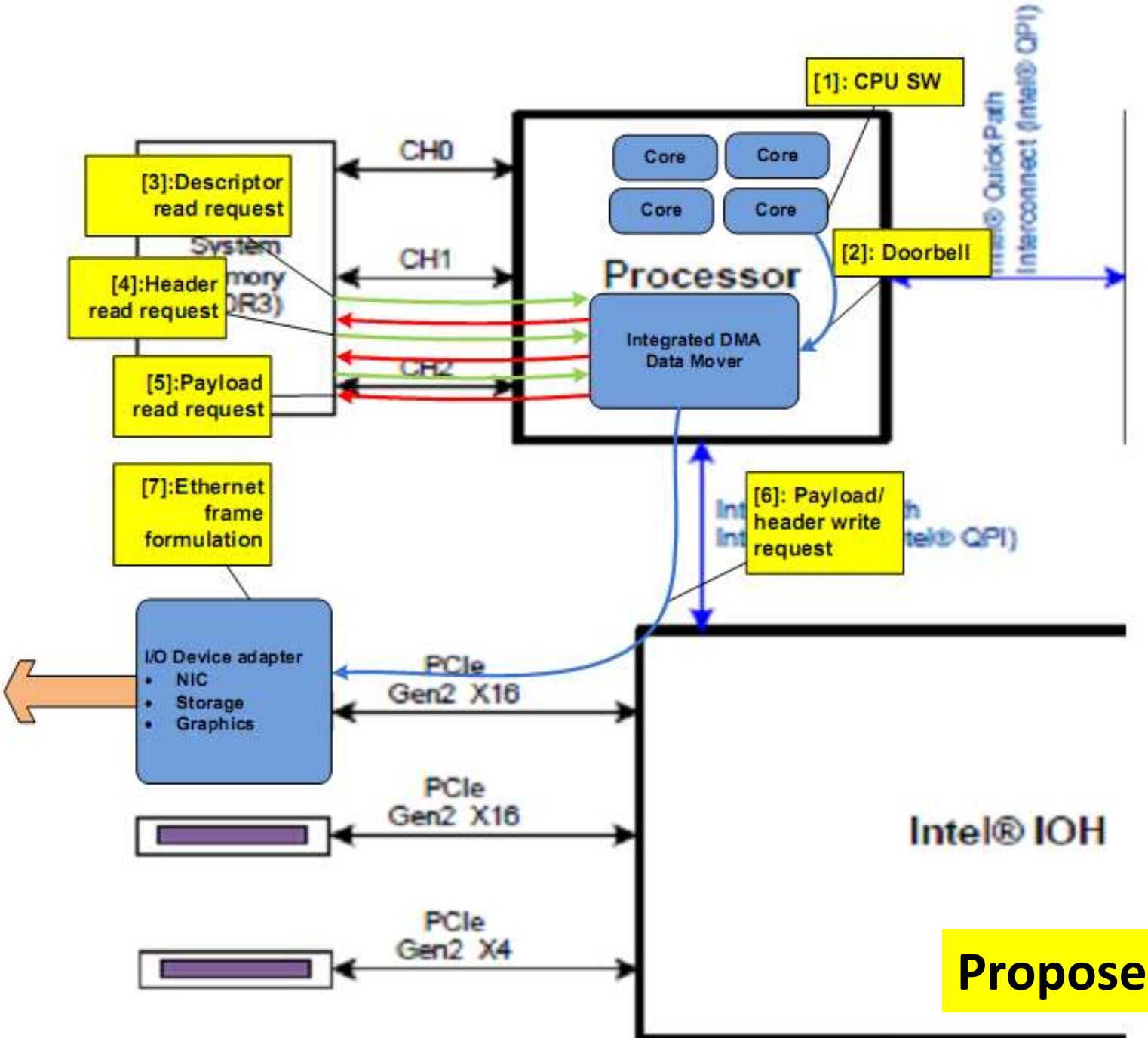


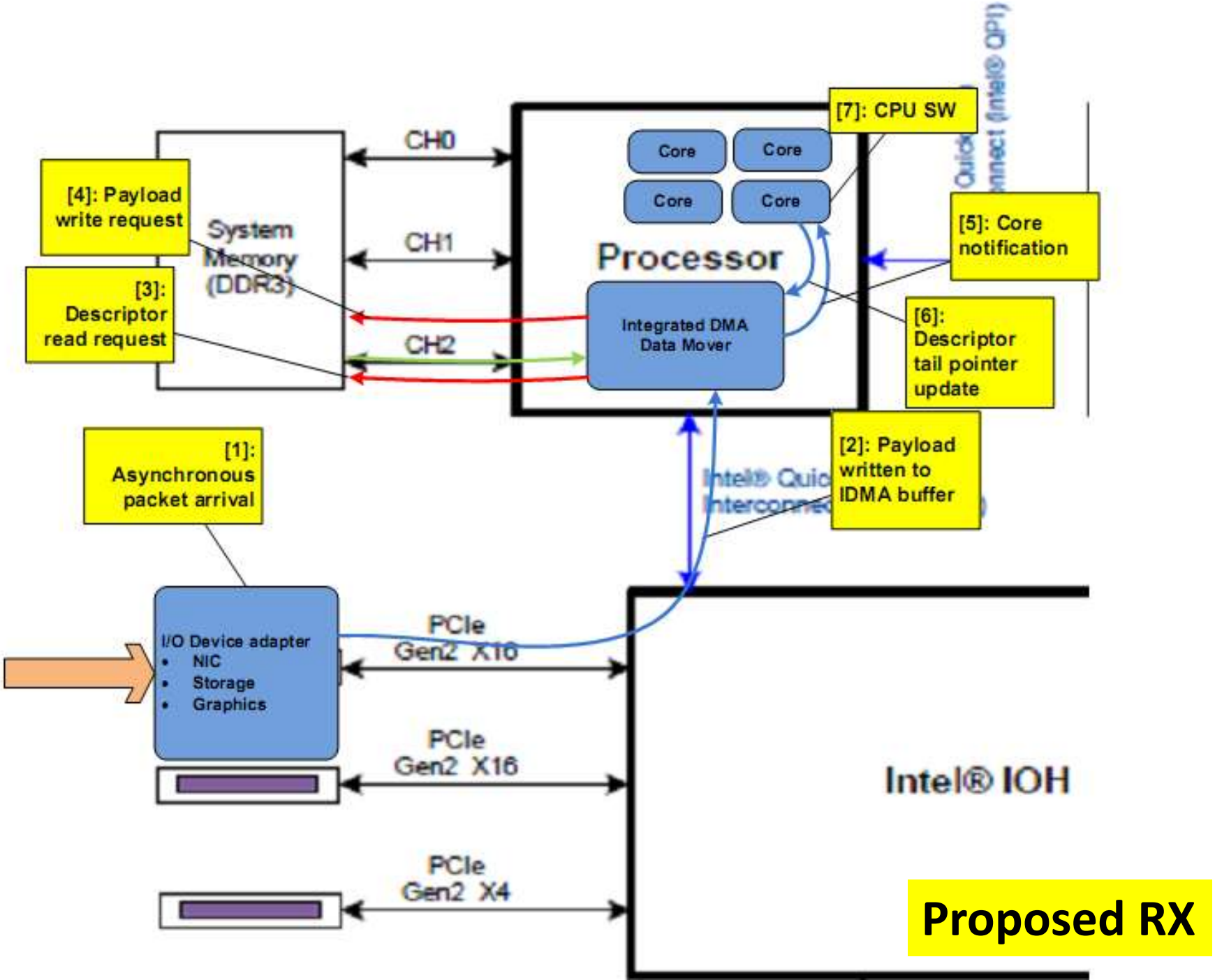
# PCIe bandwidth utilization



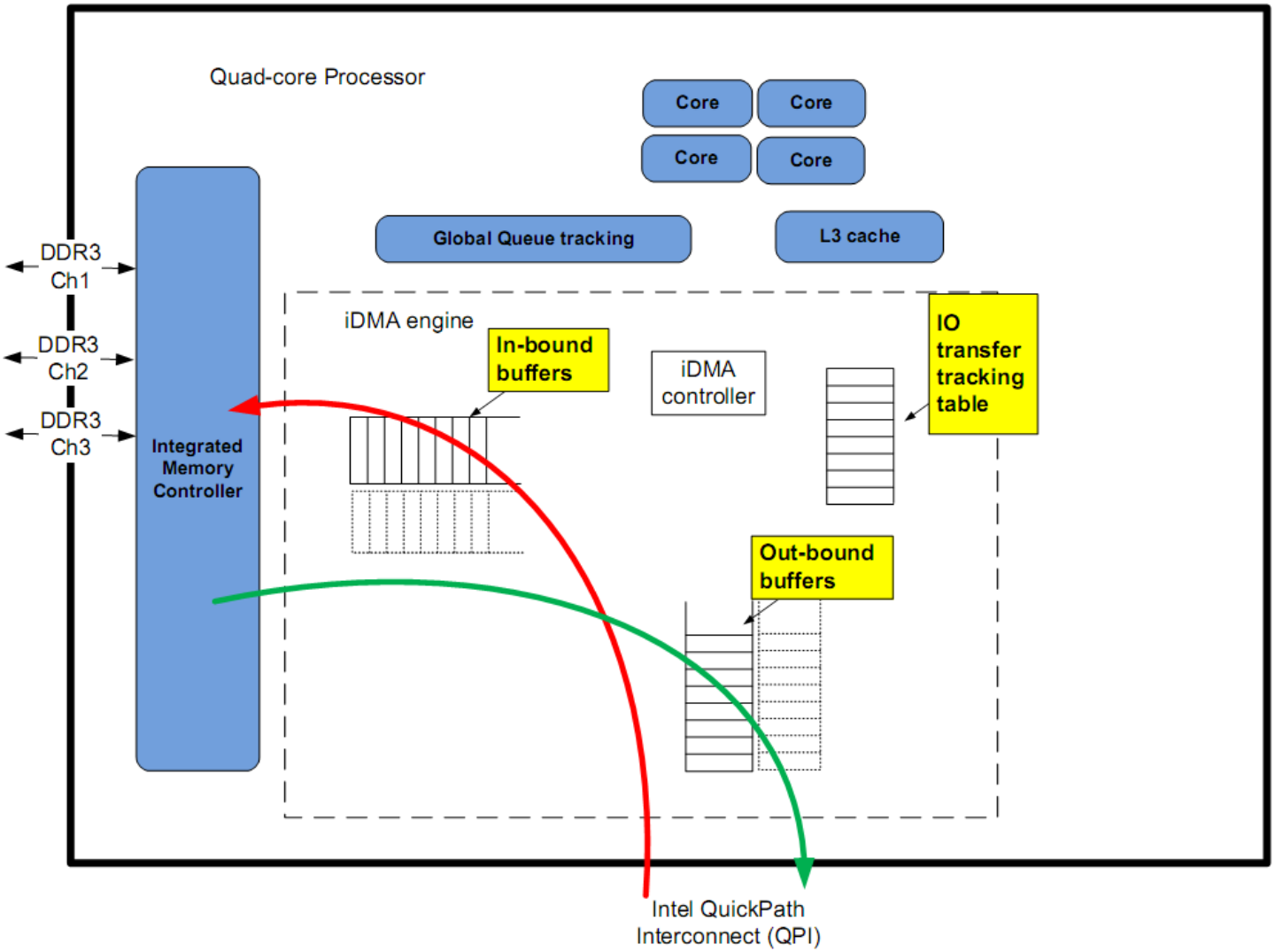
# Basic proposal claims

| Factor | Measurement unit | Descriptor DMA | iDMA | Estimated Improvement | Comment/justification |
|---|---|---|---|---|---|
| Latency | microseconds to send a TCP/IP message between two systems | 8.8 | 7.38 | 16% | Descriptors are no longer latency critical |
| Bandwidth-per-pin | Gbps per serial lane link | 2.5 | 2.67 | 17% | Descriptors no longer consume chip-to-chip bandwidth |


[1]: CPU SW
[2]: Doorbell
[3]:Descriptor read request
[4]:Header read request
[5]:Payload read request
[6]: Payload/ header write request
[7]:Ethernet frame formulation
CH0
CH1
CH2
Core
Core
Core
Core
Processor
Integrated DMA Data Mover
I/O Device adapter
• NIC
• Storage
• Graphics
PCIe Gen2 X16
PCIe Gen2 X16
PCIe Gen2 X4
Intel® IOH
Proposed TX

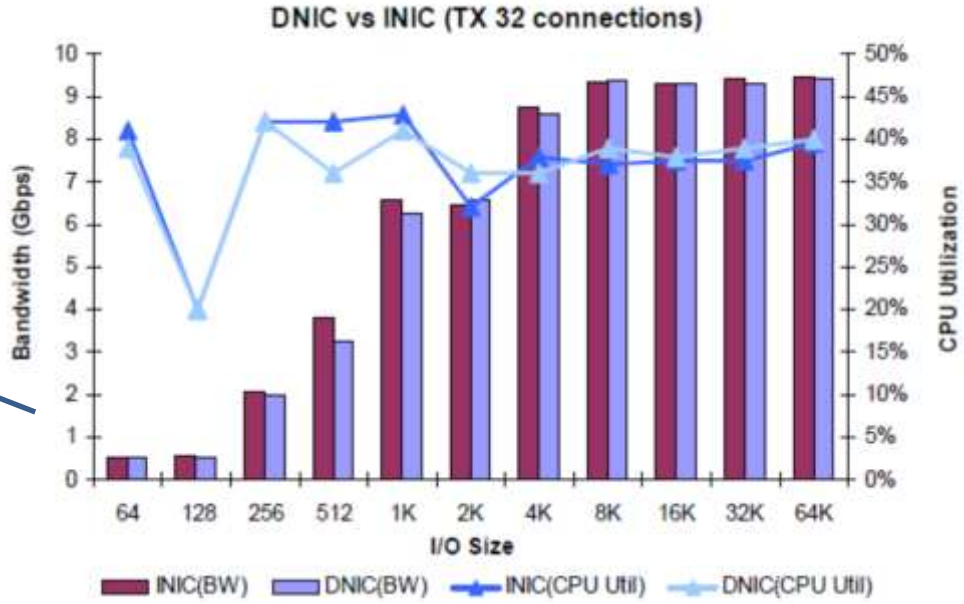

[7]: CPU SW
Core
Core
Core
Core
Processor
[5]: Core notification
[4]: Payload write request
System Memory (DDR3)
CH0
CH1
CH2
[3]: Descriptor read request
Integrated DMA Data Mover
[6]: Descriptor tail pointer update
[2]: Payload written to IDMA buffer
[1]: Asynchronous packet arrival
I/O Device adapter
• NIC
• Storage
• Graphics
PCIe Gen2 X16
PCIe Gen2 X16
PCIe Gen2 X4
Intel® IOH
Proposed RX

# iDMA internals

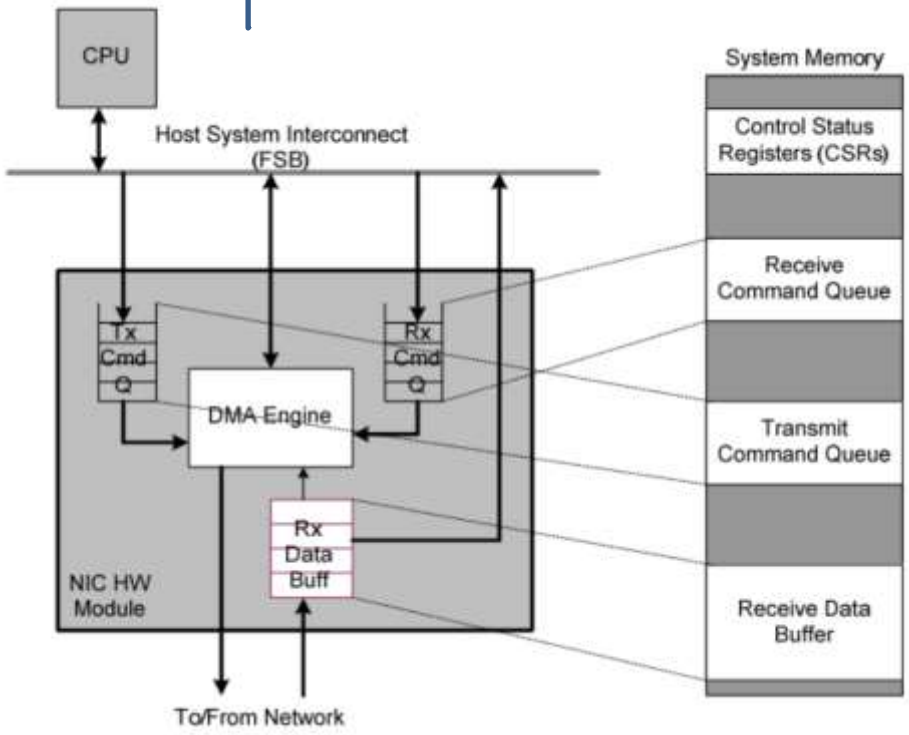


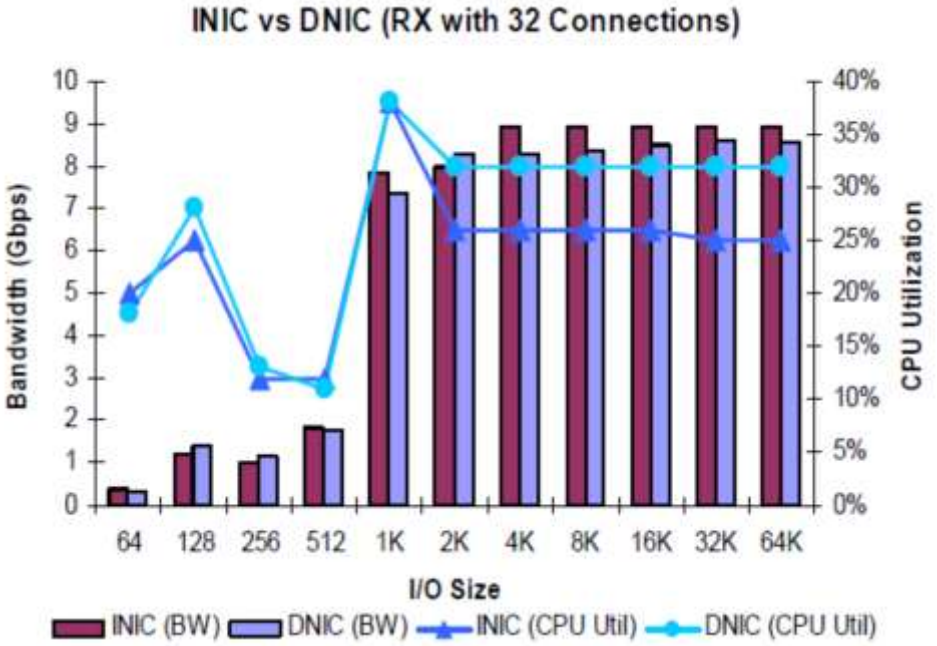
# Related work

Sun Niagara2




Memory coherent IO



| Factor | Measurement unit | Descriptor DMA | iDMA | Estimated Improvement | Comment/justification |
|---|---|---|---|---|---|
| Latency | microseconds to send a TCP/IP message between two systems | 8.8 | 7.38 | 16% | Descriptors are no longer latency critical |
| Bandwidth-per-pin | Gbps per serial lane link | 2.5 | 2.67 | 17% | Descriptors no longer consume chip-to-chip bandwidth |
| Bandwidth scalability | Not quantifiable | | | | Reduced silicon area and power |
| Power efficiency | Normalized core power (maximum) | 100% | 29% | 71% | Power reduction due to more efficient core allocation of IO |
| Quality of service | Nanoseconds to control connection priority from software perspective | 600 | 50 | 92% | Round trip latency to queuing control reduced from PCIe to system memory |
| Multiple IO complexity | Die cost reduction | 100% | <50% | >50% | Silicon, power regulation and cooling cost reduction of multiple IO controllers into a single iDMA instance |
| Security | | na | na | na | not quantifiable |

Thank you!

Questions?