

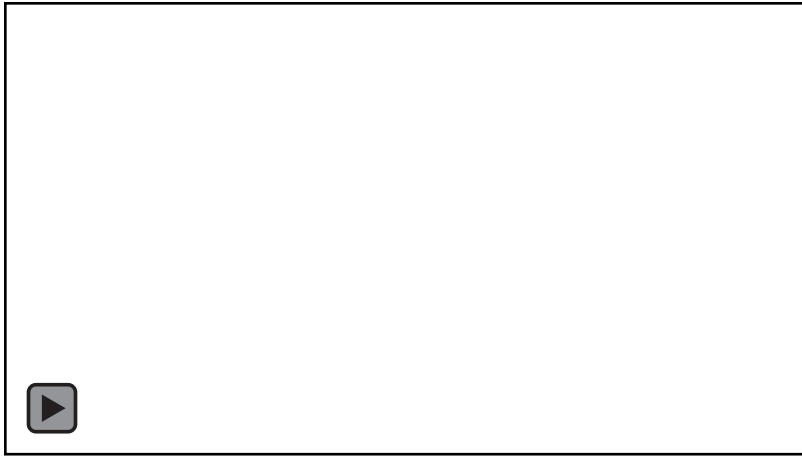
P51: High Performance Networking

Lecture 6: Low Latency Devices

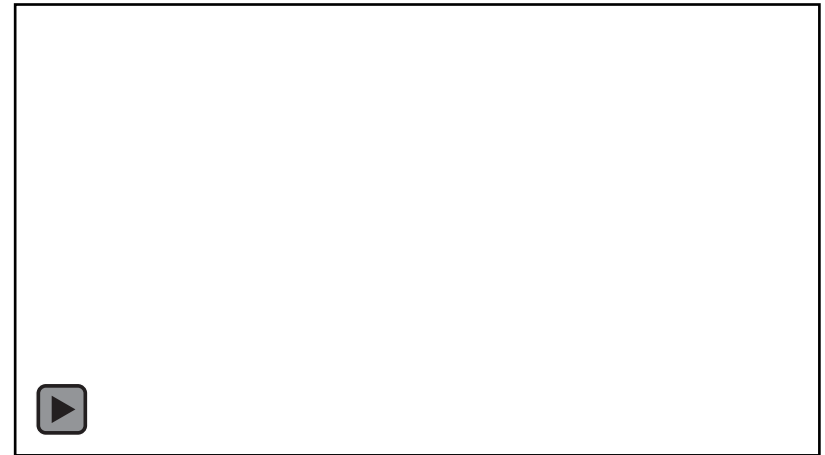
Latency matters

Latency effect on facial recognition

Source: Glimpse project, MIT, 2014



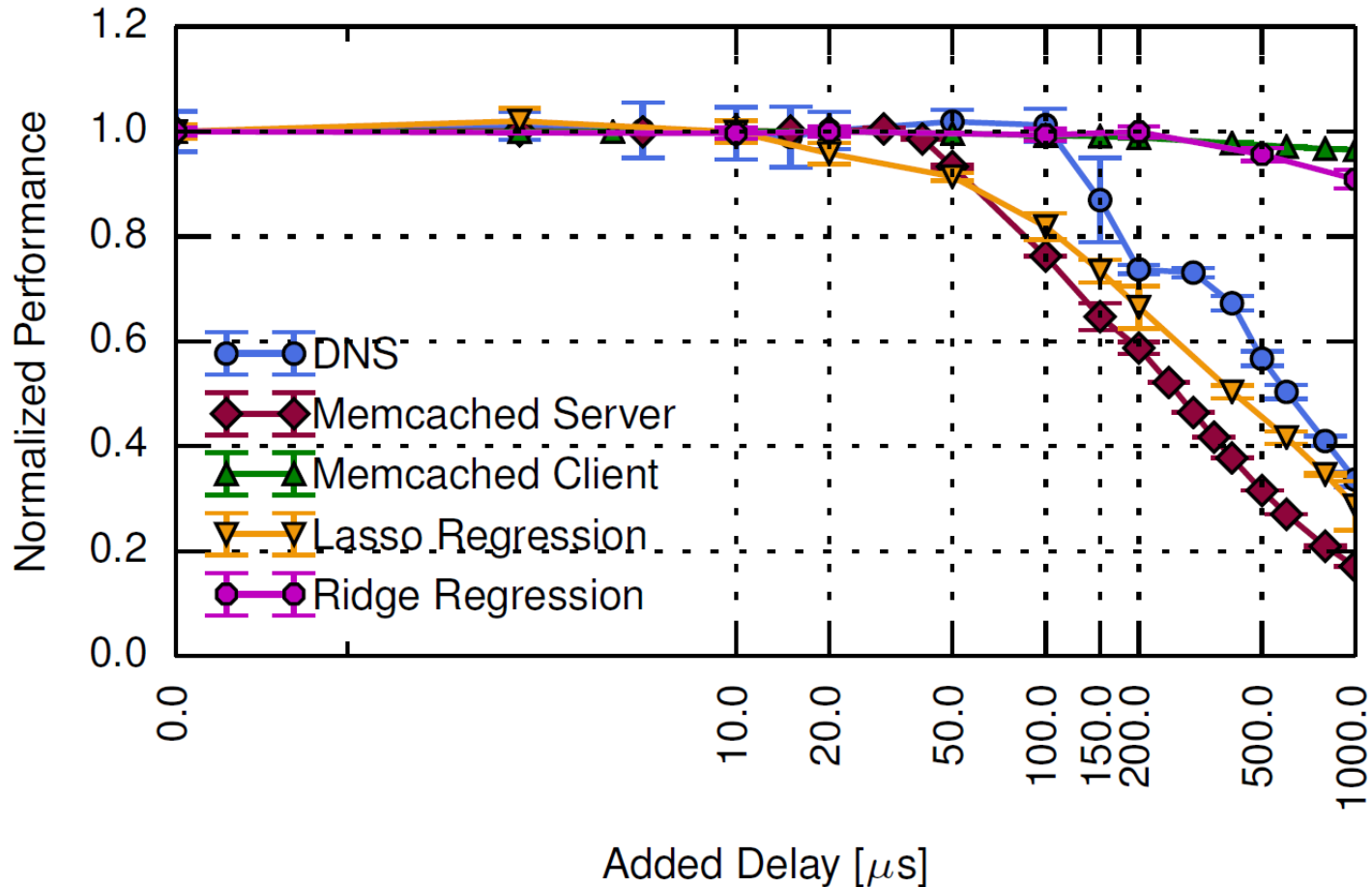
Remote Processing



Local Processing

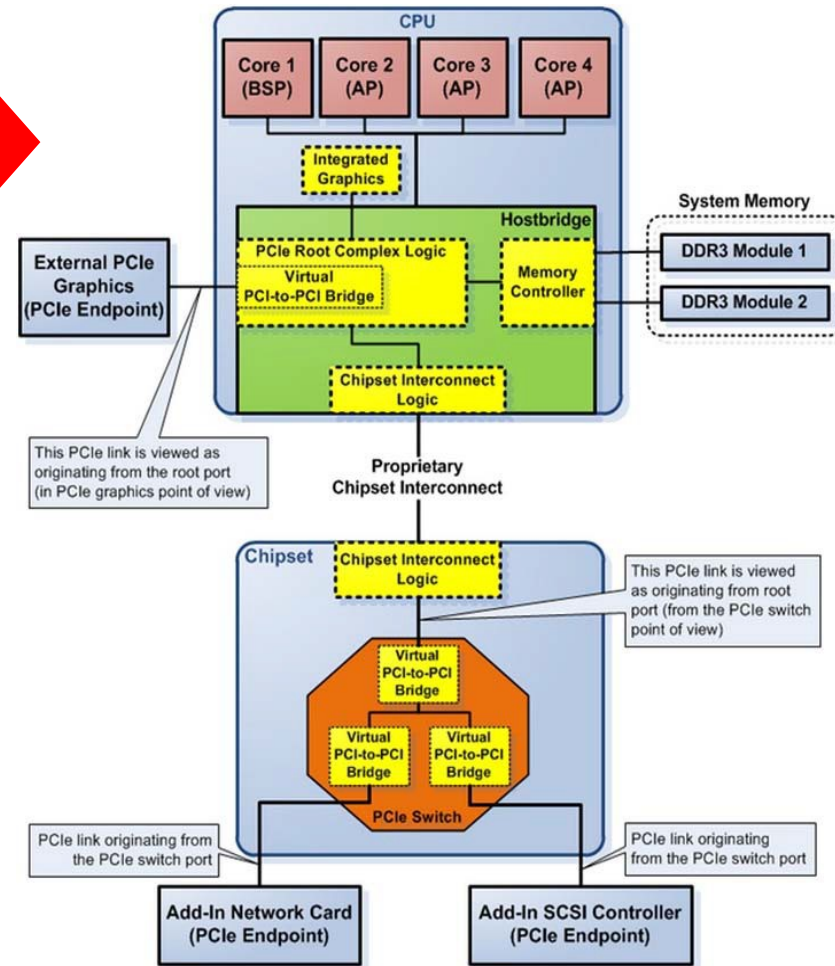
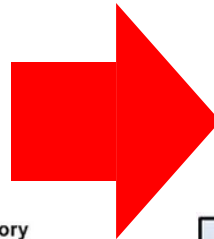
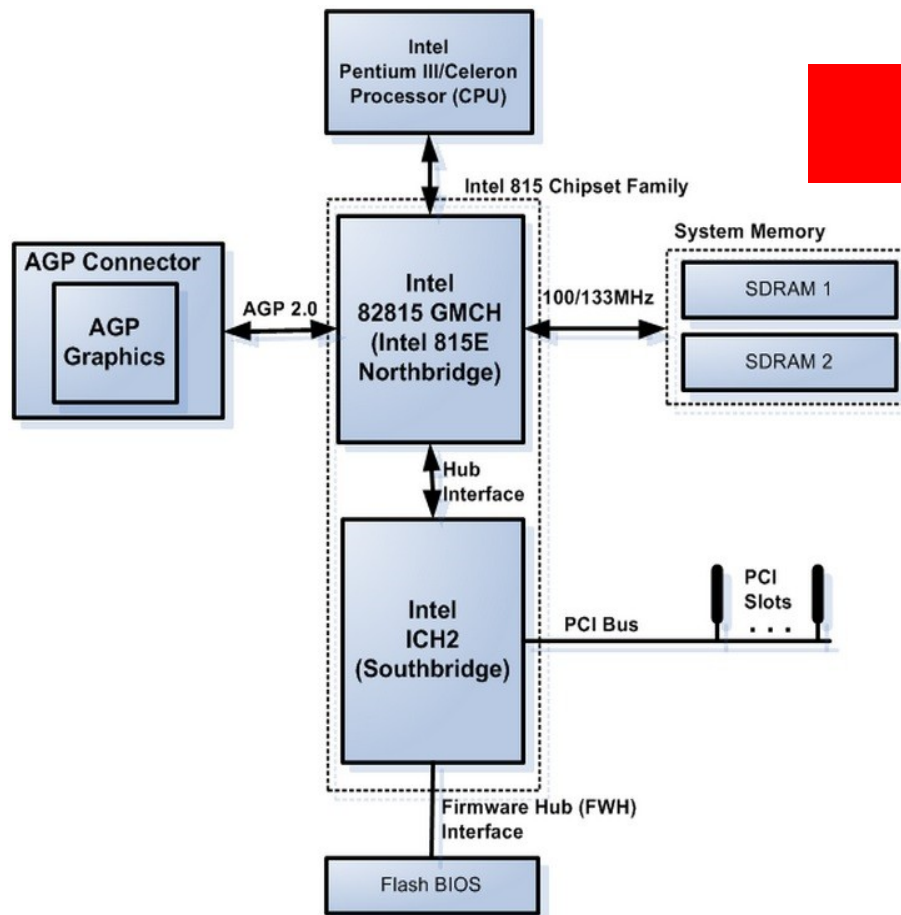
- “being fast really matters...half a second delay caused a 20% drop in traffic and it killed user satisfaction” - Marissa Mayer @ Web 2.0 (2008)
- “A millisecond decrease in a trade delay may boost a high-speed firm's earnings by about 100 million per year” – SAP, 2012

Latency within the data centre matters



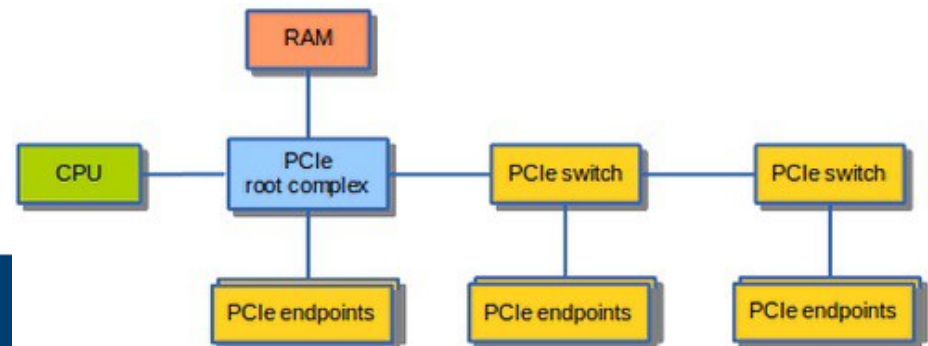
DMA

Host architecture



Interconnecting components

- **Need interconnections between**
 - CPU, memory, storage, network, I/O controllers
- **Shared Bus: shared communication channel**
 - A set of parallel wires for data and synchronization of data transfer
 - Can become a bottleneck
- **Performance limited by physical factors**
 - Wire length, number of connections
- **More recent alternative: high-speed serial connections with switches**
 - Like networks



I/O System Characteristics

- **Performance measures**

- Latency (response time)
- Throughput (bandwidth)
- Desktops & embedded systems
 - Mainly interested in response time & diversity of devices
- Servers
 - Mainly interested in throughput & expandability of devices

- **Reliability**

- Particularly for storage devices (fault avoidance, fault tolerance, fault forecasting)

I/O Management and strategies

- **I/O is mediated by the OS**
 - Multiple programs share I/O resources
 - Need protection and scheduling
 - I/O causes asynchronous interrupts
 - Same mechanism as exceptions
 - I/O programming is fiddly
 - OS provides abstractions to programs

Strategies characterize **the *amount of work*** done by the CPU in the I/O operation:

- **Polling**
- **Interrupt Driven**
- **Direct Memory Access**

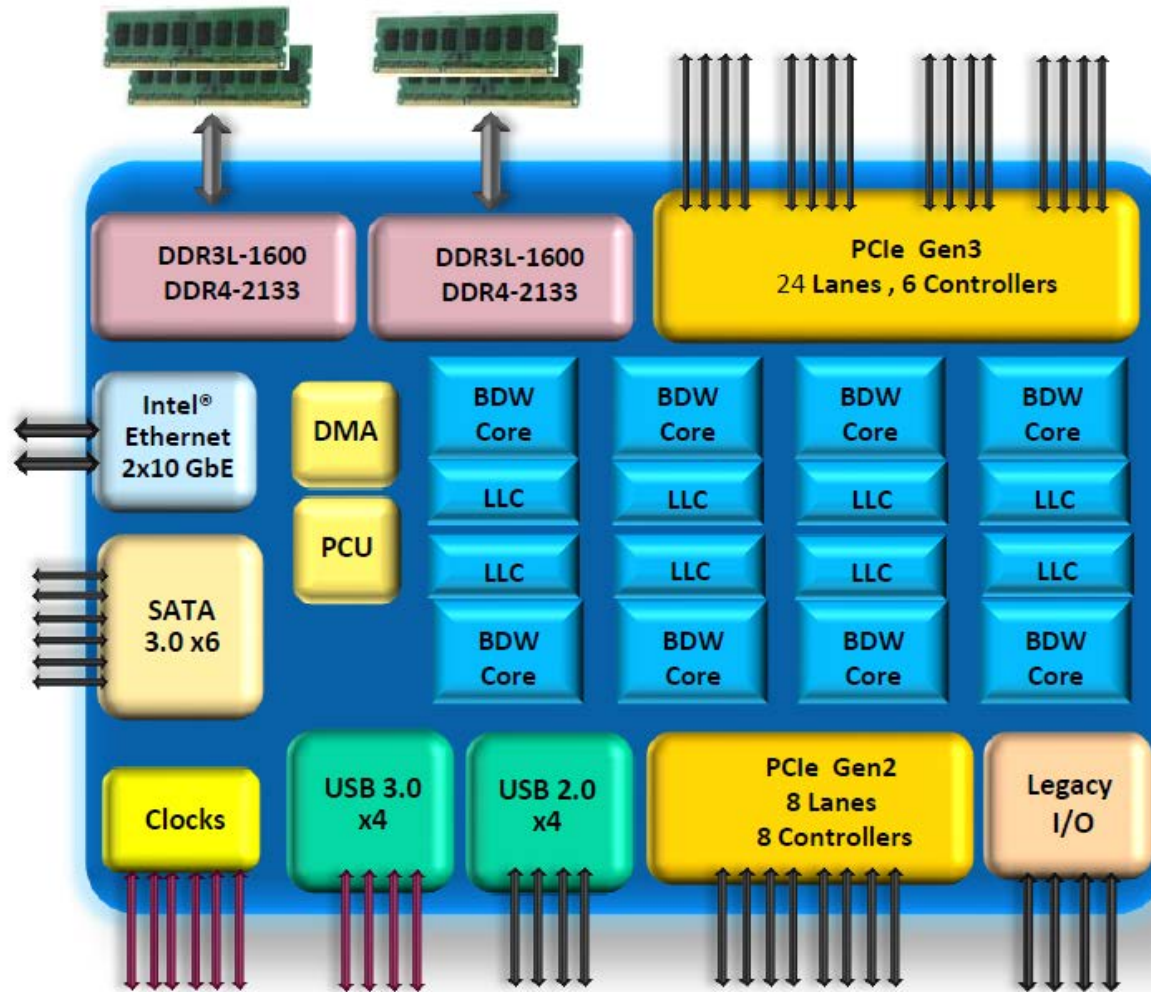
The I/O Access Problem

- Question: how to transfer data from I/O devices to memory (RAM)?
- Trivial solution:
 - Processor individually reads or writes every word
 - Transferred to/from I/O through an internal register to memory
- Problems:
 - Extremely inefficient – can occupy a processor for 1000's of cycles
 - Pollute cache

DMA

- DMA – Direct Memory Access
- A modern solution to the I/O access problem
- The peripheral I/O can issue read/write commands directly to the memory
 - Through the main memory controller
 - The processor does not need to execute any operation
- Write: The processor is notified when a transaction is completed (interrupt)
- Read: The processor issues a signal to the I/O when the data is ready in memory

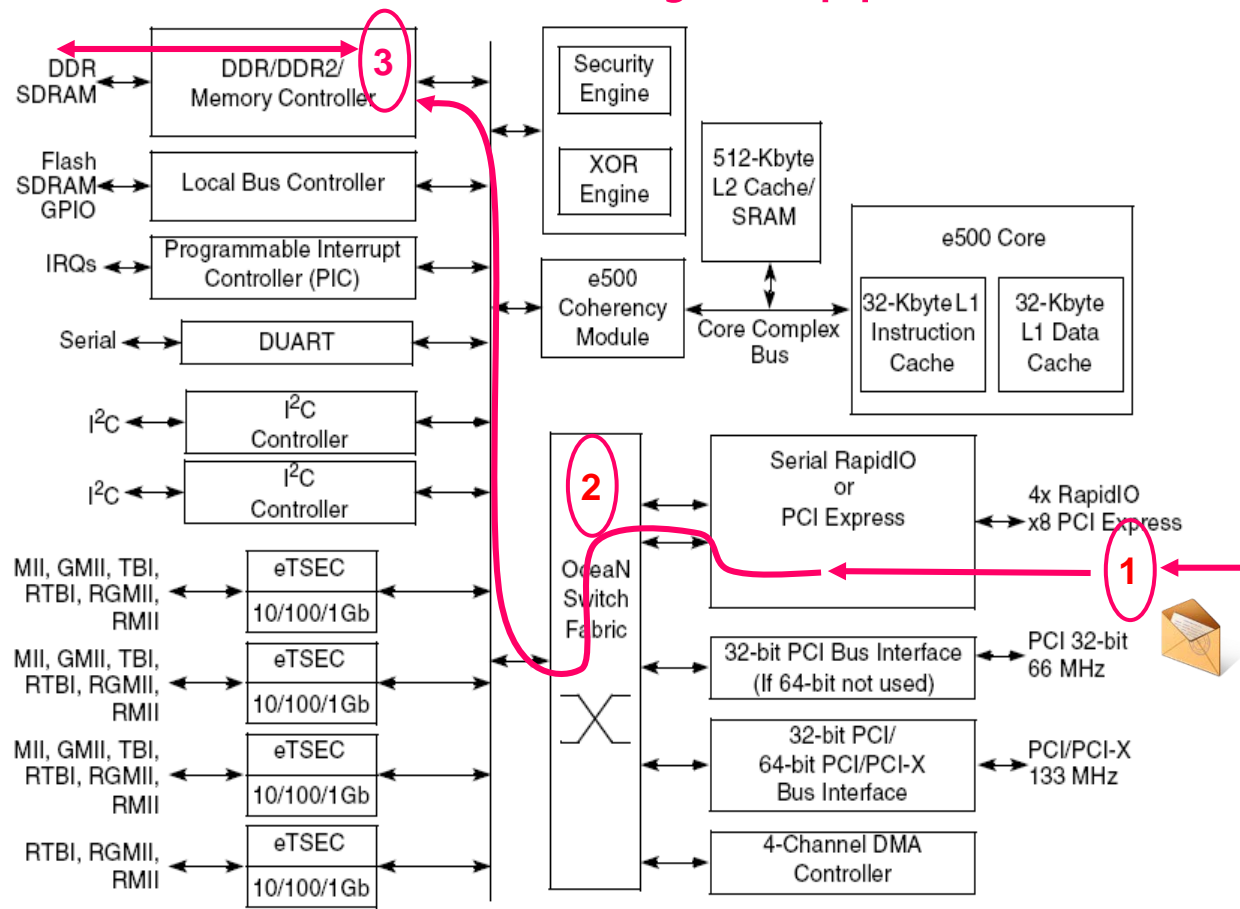
Example – Intel Xeon D



Example (Legacy Embedded Processor)

1. Message arrives on I/O interface.
Message is decoded to Mem read/write.
Address is converted to internal address.
2. Mem Read/Write command goes through the switch to the internal bus and memory controller.
3. Memory controller executes the command to the DRAM.
Returns data if required in the same manner.

Memory Mapped Access



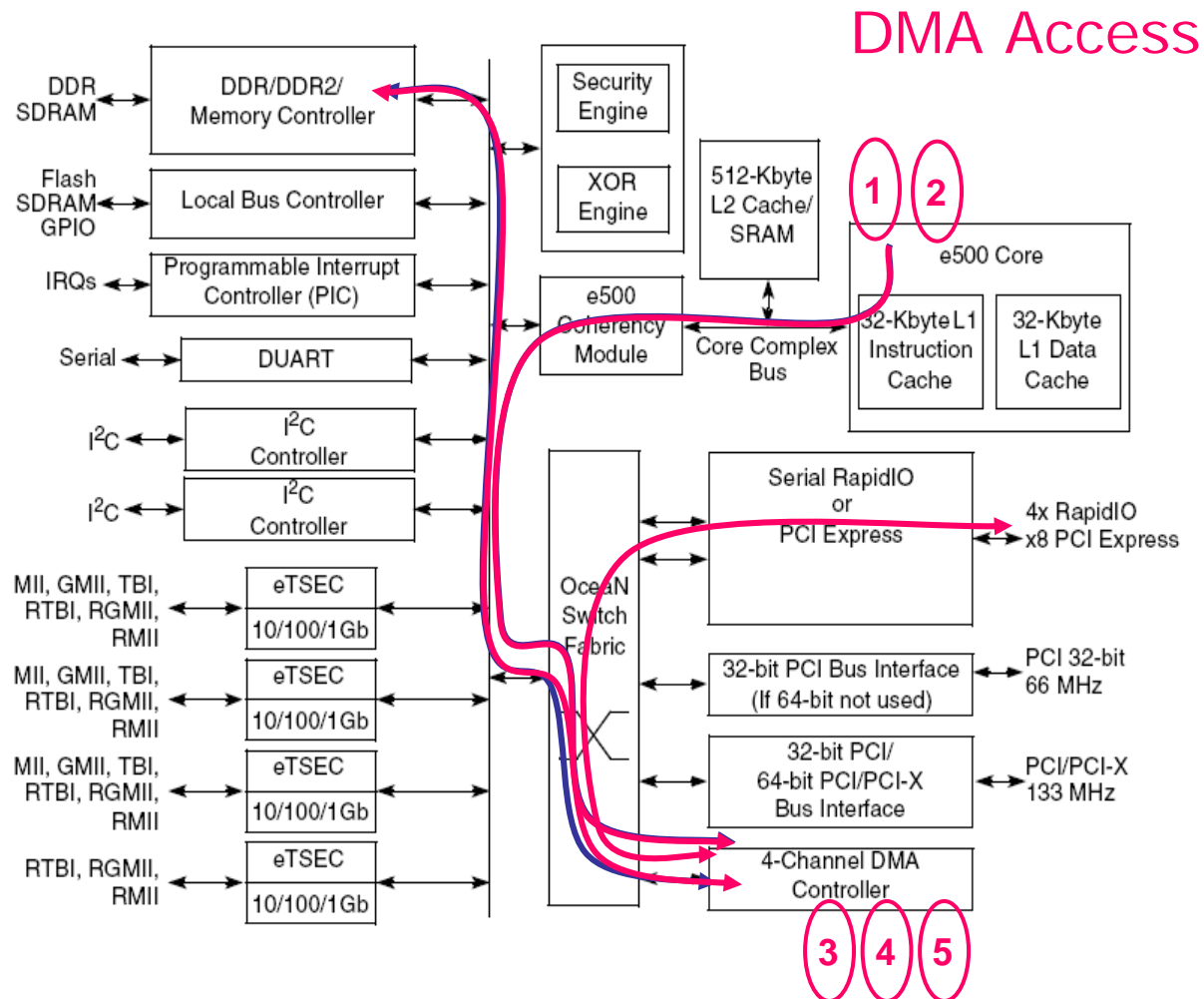
DMA

- DMA accesses are usually handled in *buffers*
 - Single word/block is typically inefficient
- The processors assigns the peripheral unit the buffers in advance
- The buffers are typically handled by *buffer descriptors*
 - Pointer to the buffer in the memory
 - May point to the next buffer as well
 - Indicates buffer status: owner, valid etc.
 - May include additional buffer properties as well

Example (Legacy Embedded Processor)

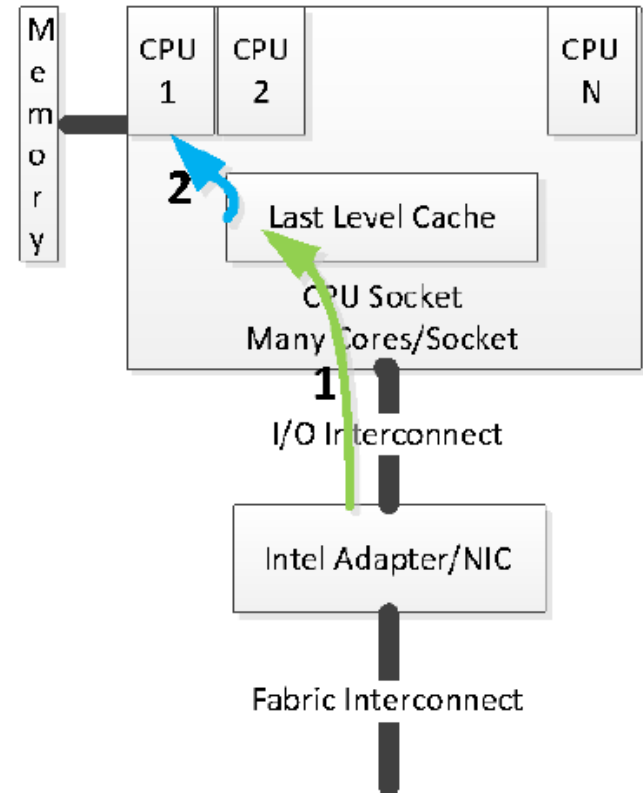
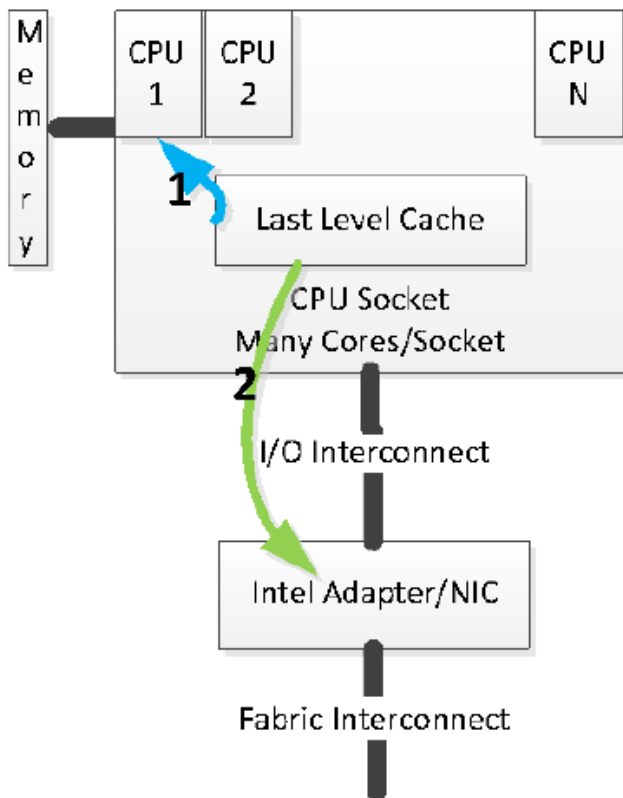
Transfers blocks of data between external interfaces and local address space

1. A transfer is started by SW writing to DMA engine configuration registers
2. SW Polls DMA channel state to idle and sets trigger
3. DMA engine fetches a descriptor from memory
4. DMA engine reads block of data from source
5. DMA engine writes data to destination



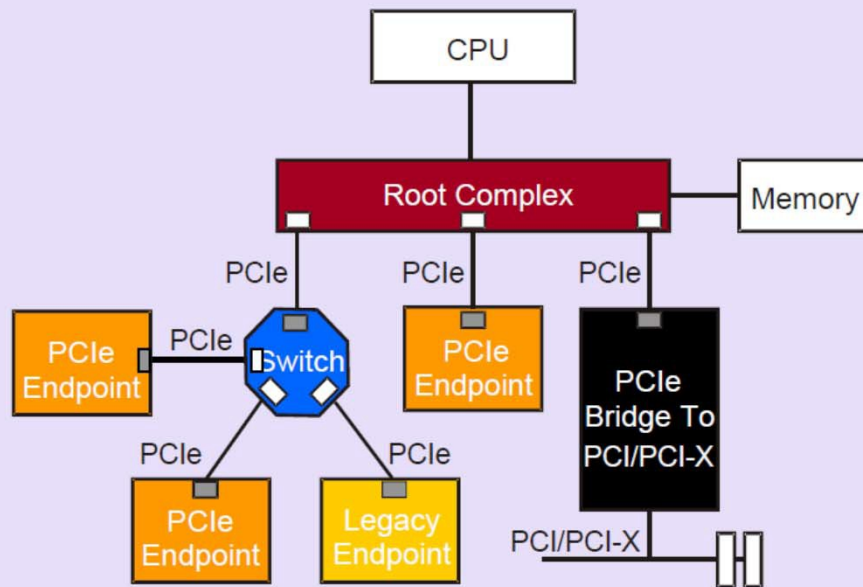
Intel Data Direct I/O (DDIO)

- Data is written and read directly to/from the last level cache



PCIe - introduction

- PCIe is a *serial point-to-point interconnect* between two devices
- Implements *packet based protocol (TLPs)* for information transfer
- *Scalable performance* based on # of signal Lanes implemented on the PCIe interconnect
- Supports *credit-based* point-to-point flow control (not end-to-end)



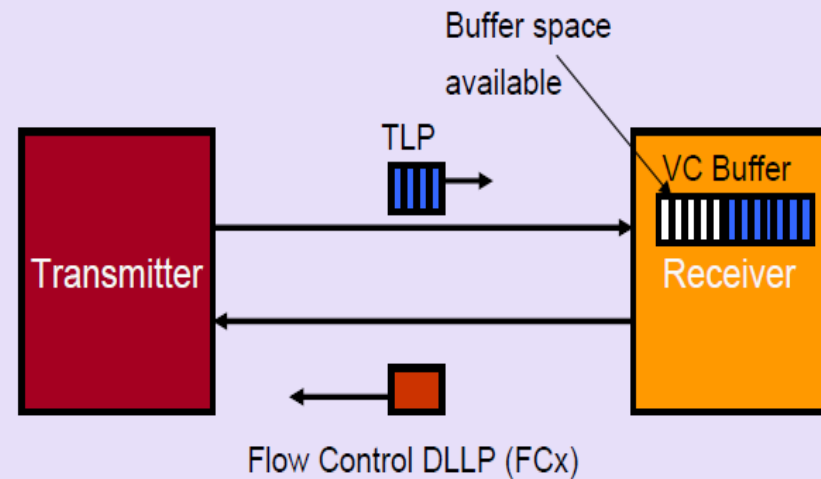
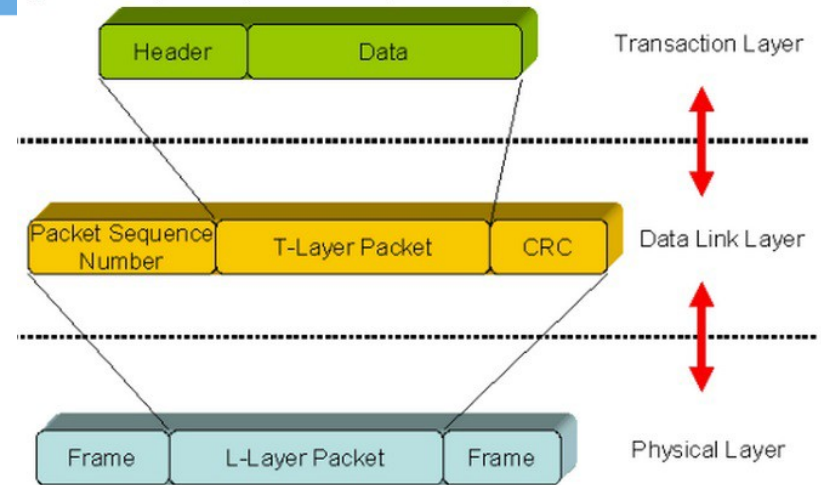
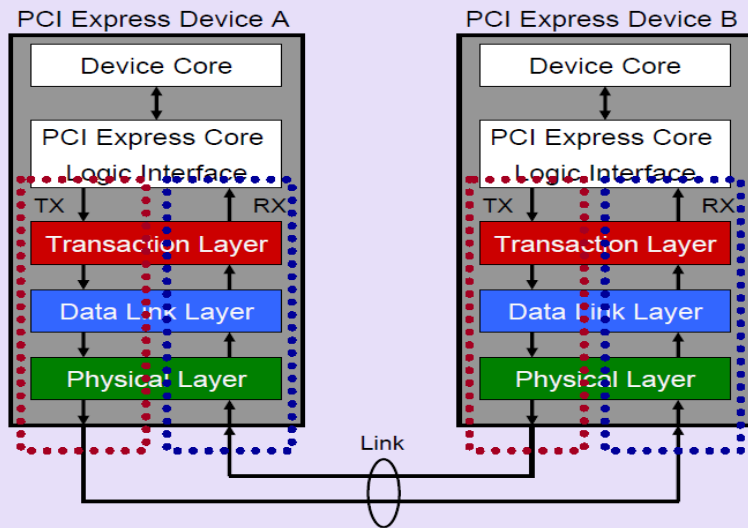
Provides:

- Processor independence & buffered isolation
- Bus mastering
- Plug and Play operation

PCIe transaction types

- Memory Read or Memory Write. Used to transfer data from or to a memory mapped location
- I/O Read or I/O Write. Used to transfer data from or to an I/O location
- Configuration Read or Configuration Write. Used to discover device capabilities, program features, and check status in the 4KB PCI Express configuration space.
- Messages. Handled like posted writes. Used for event signaling and general purpose messaging.

PCIe architecture



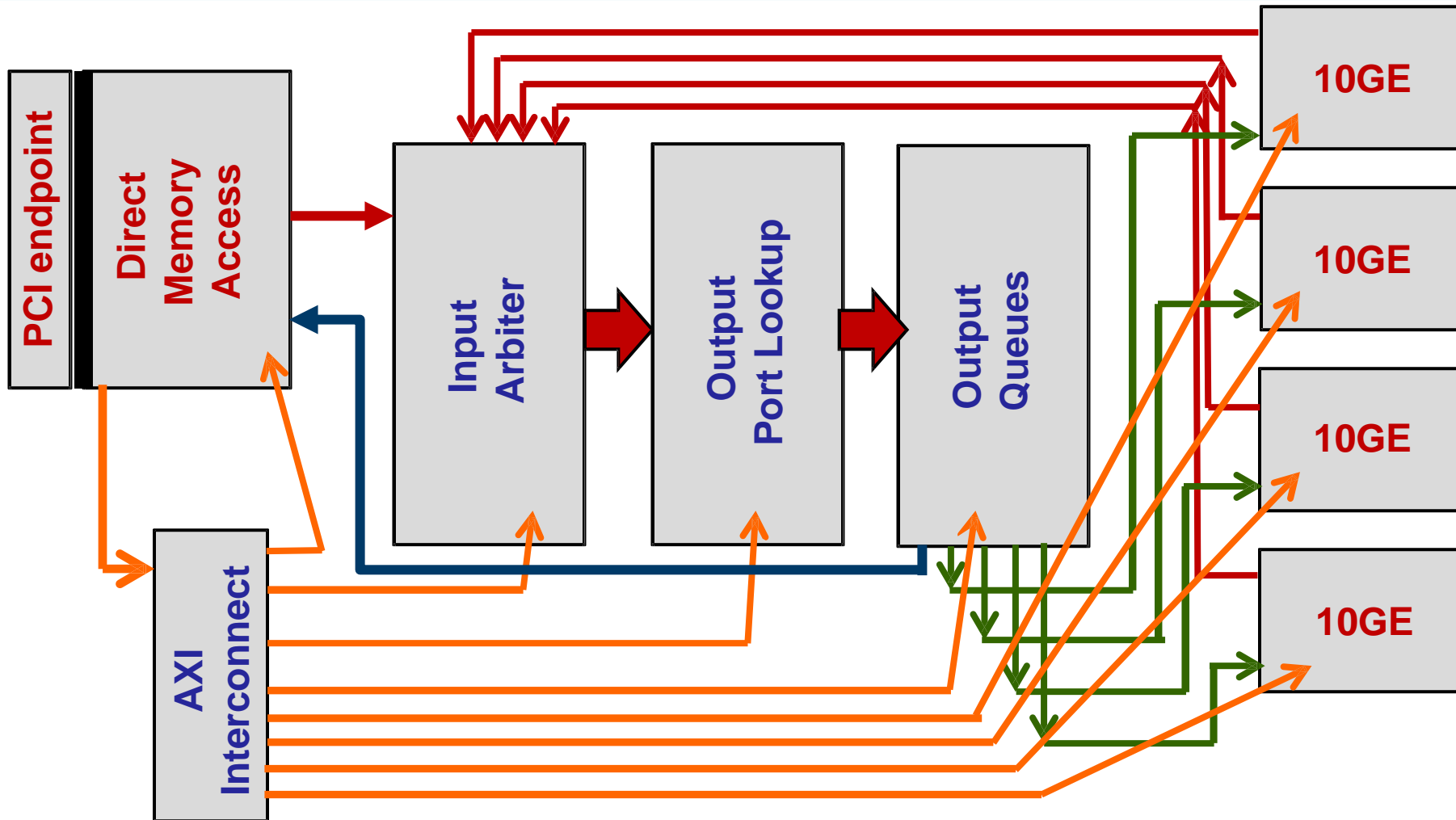
Interrupt Model

PCI Express supports three interrupt reporting mechanisms:

1. Message Signaled Interrupts (MSI)
 - interrupt the CPU by writing to a specific address in memory with a payload of 1 DW
2. Message Signaled Interrupts - X (MSI-X)
 - MSI-X is an extension to MSI, allows targeting individual interrupts to different processors
3. INTx Emulation
 - four physical interrupt signals INTA-INTD are messages upstream
 - ultimately be routed to the system interrupt controller

NetFPGA Reference Projects

Host system



Processing Overheads

- Processing in the kernel takes a lot of time...

Component	Time [us]
Driver RX	0.60
Ethernet & IPv4 RX	0.19
TCP RX	0.53
Socket Enqueue	0.06
TCP TX	0.70
IPv4 & Ethernet TX	0.06
Driver TX	0.43

Source: Yasukata *et al.* “StackMap: Low-Latency Networking with the OS Stack and Dedicated NICs”, Usenix ATC 2016

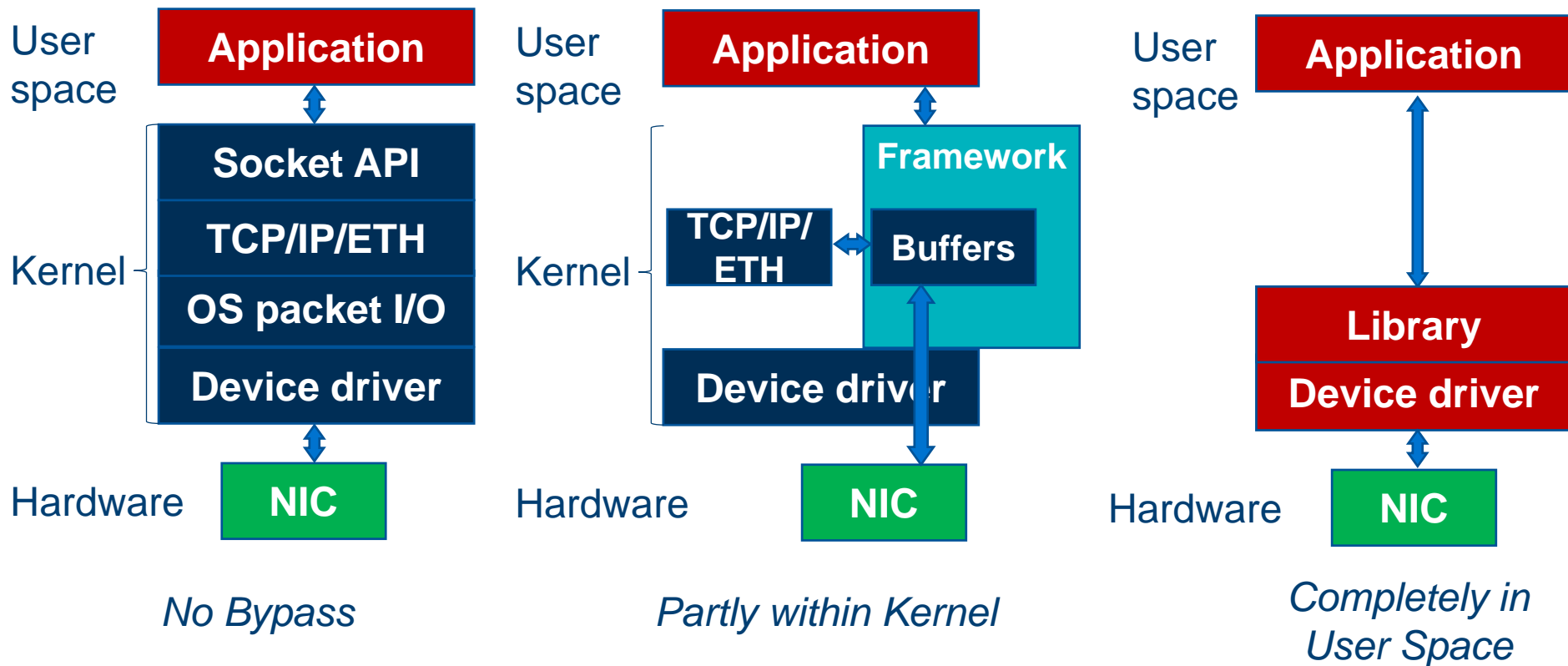
Processing Overheads

- Processing in the kernel takes a lot of time...
- Order of microseconds (~2-4us on Xeon E5-v4)
- $\times 10$ the time through a switch
- Solution: don't go through the kernel!

Kernel Bypass

- The Kernel is slow – lets bypass the Kernel!
- There are many ways to achieve kernel bypass
- Some examples:
 - Device drivers:
 - Customized kernel device driver. E.g. Netmap forks standard Intel drivers with extensions to map I/O memory into userspace.
 - Custom hardware and use bespoke device drivers for the specialized hardware.
 - Userspace library: anything from basic I/O to the entire TCP/IP stack

Kernel Bypass - Examples



DPDK

- DPDK is a popular set of **libraries and drivers** for fast packet processing.
- Originally designed for Intel processors
 - Now running also on ARM and Power CPUs
- Runs mostly in Linux User space.
- Main libraries: multicore framework, huge page memory, ring buffers, poll-mode drivers (networking, crypto etc)
- It is ***not*** a networking stack

DPDK

- Usage examples:
 - Send and receive packets within minimum number of CPU cycles
 - E.g. less than 80 cycles
 - Fast packet capture algorithms
 - Running third-party stacks
- Some projects demonstrated 100's of millions packets per seconds
 - But with limited functionality
 - E.g. as a software switch / router

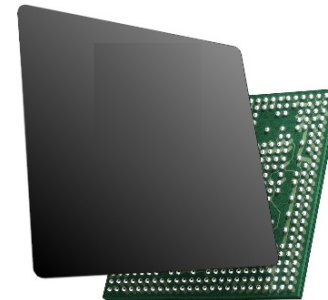
What are the challenges ahead?

Designing new network devices

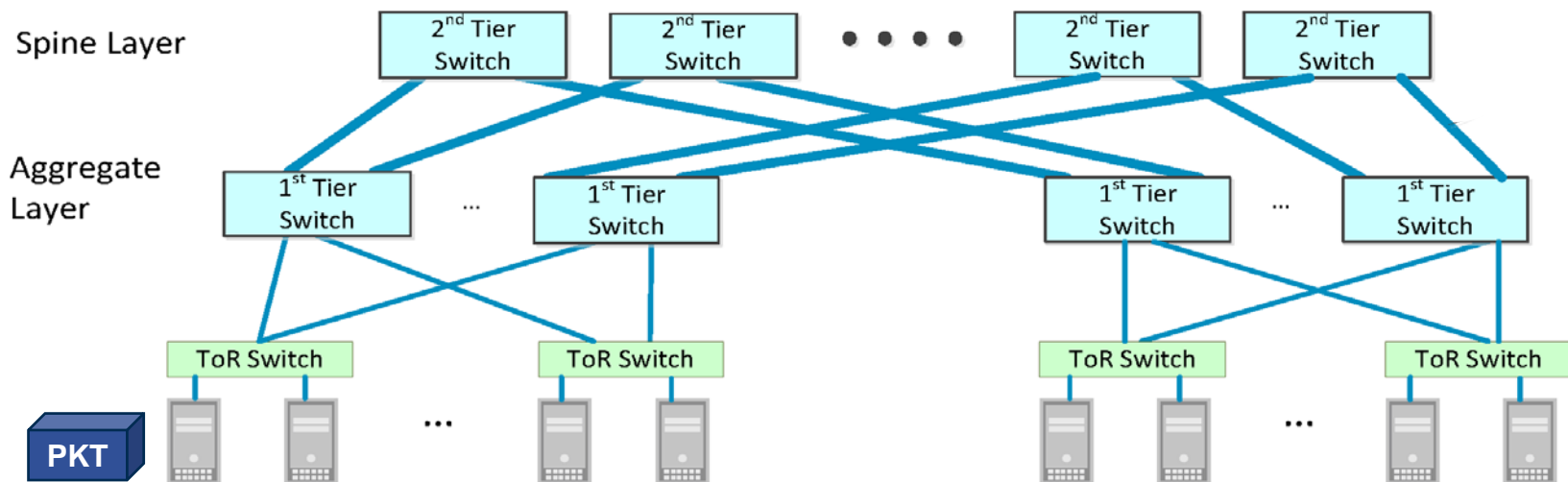
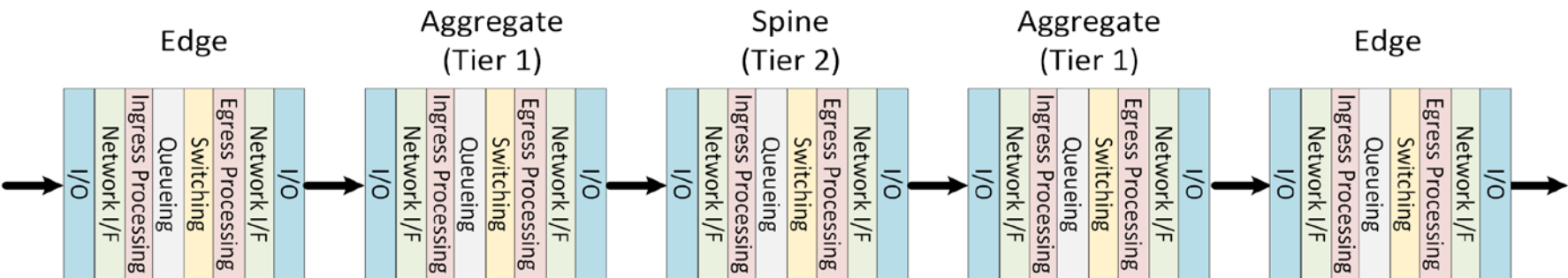
- A decade ago: *“Can we implement this feature?”*
- Today: *“Is this feature worth implementing, given the design constraints?”*

The resource wall

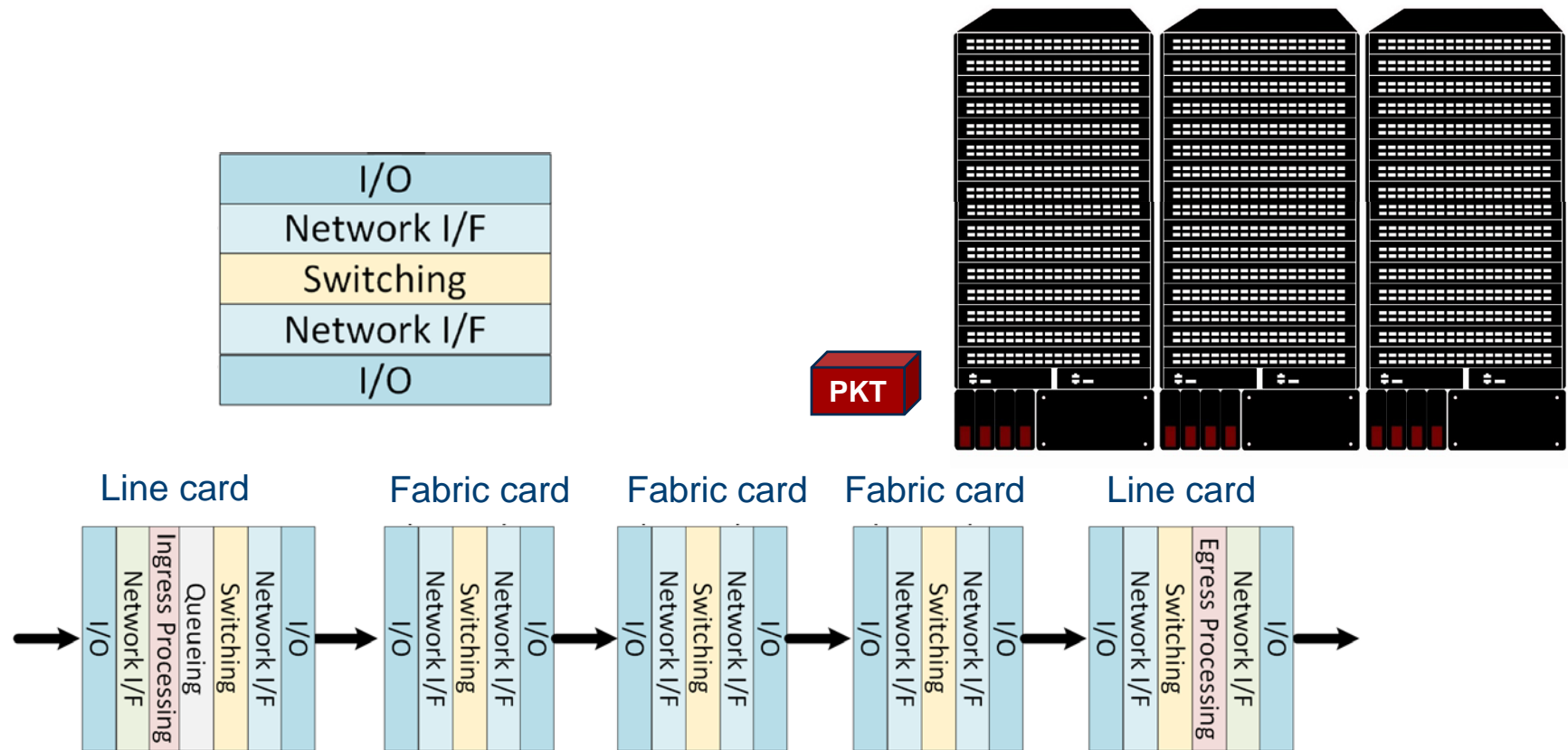
- Network silicon die > 20 Billion transistors (2019)
- Limited by:
 - Power density
 - Die size
 - Manufacturing feasibility



Data centre networks

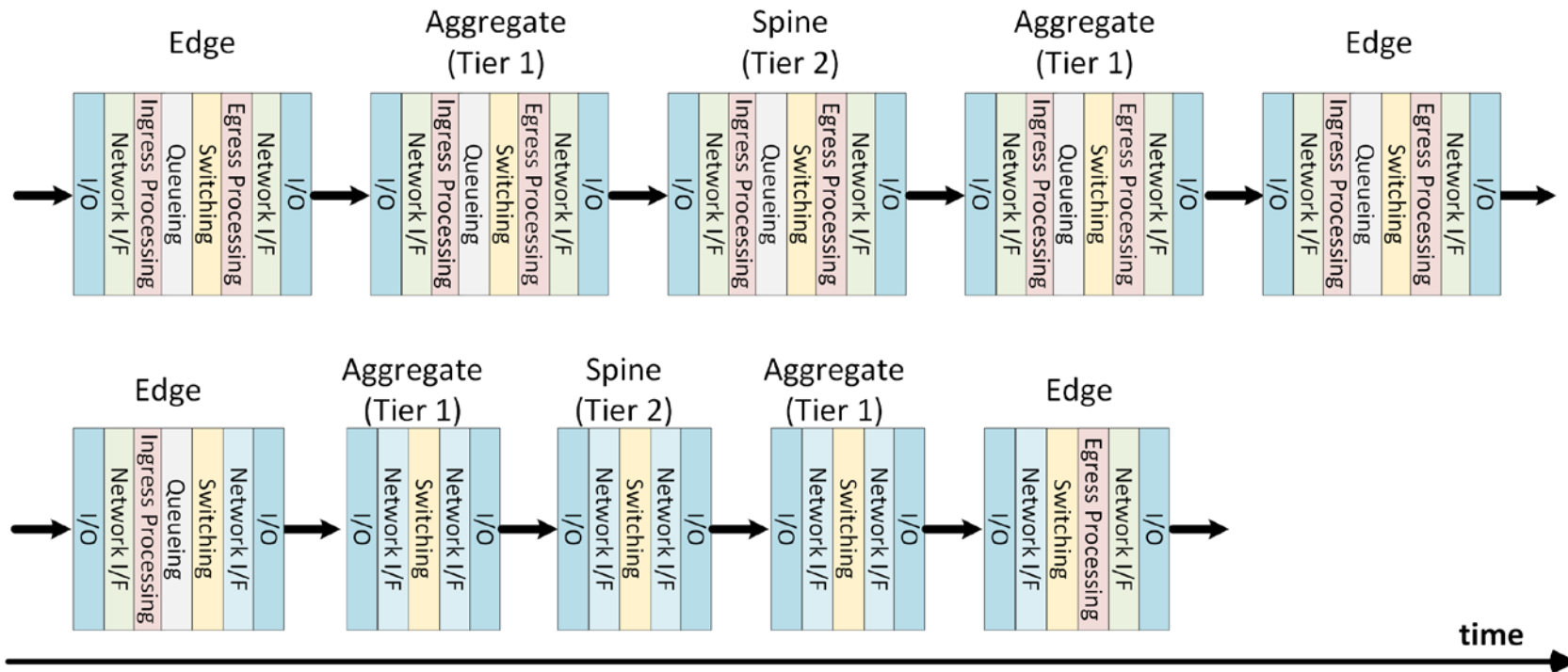


Switch system



Why waste resources? in n tier network

$$O(n \times (\text{Switching} + 2 \times \text{I/O} + 2 \times \text{NIF}) + n \times (\text{Ingress Processing} + \text{Egress Processing} + \text{Queueing}))$$



$$O(n \times (\text{Switching} + 2 \times \text{I/O} + 2 \times \text{NIF}) + 1 \times (\text{Ingress Processing} + \text{Egress Processing} + \text{Queueing}))$$

The switch pipeline

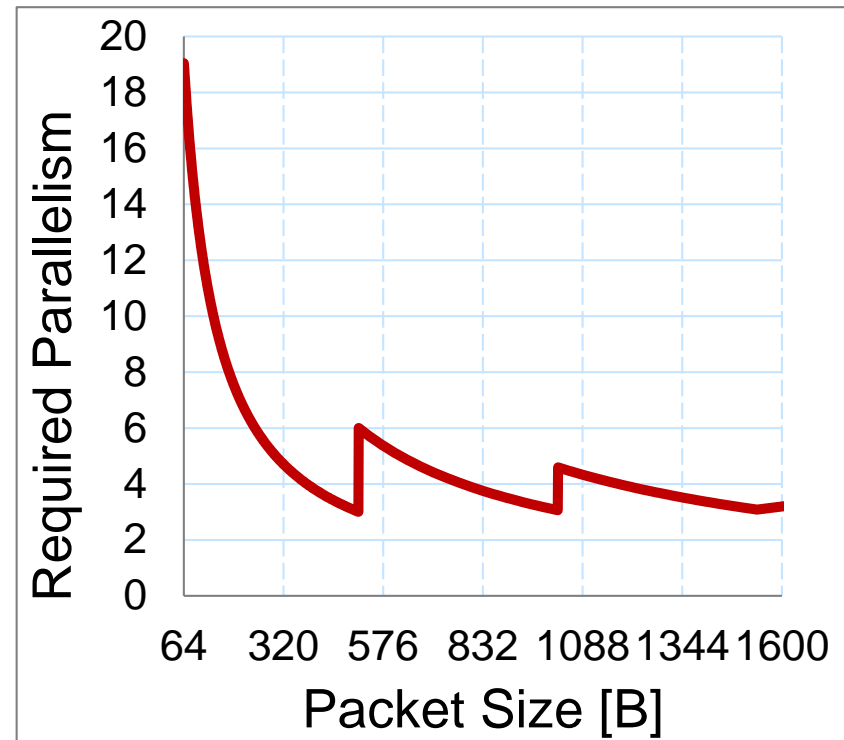
12.8Tbps Switches! 32×400GE!

Lets convert to packet rate requirements:

5800 Mpps @ 256B (100GE→38.7Mpps)

19050 Mpps @ 64B (100GE→150Mpps)

But clock rate is only ~1GHz....



The single-pipeline switch

Observation:

To support full line rate for all packet sizes, network devices need to process multiple packets each and every clock cycle.

The age of multi core has reached switching...

The switch pipeline

... Surely it has little effect on real workloads...

Bus Width	Data-path Utilization			
	TensorFlow	KMeans	PageRank	Web Server
256B	97%	97.6%	96.2%	97.3%
512B	95.8%	97.6%	92.9%	96.4%
1024B	71.1%	73.2%	68%	72.7%

Assuming VXLAN headers and 1500B NIC MTU

Using traces from Dukic et al., “Is advance knowledge of flow sizes a plausible assumption?”, NSDI’19

Do data centre networks scale?

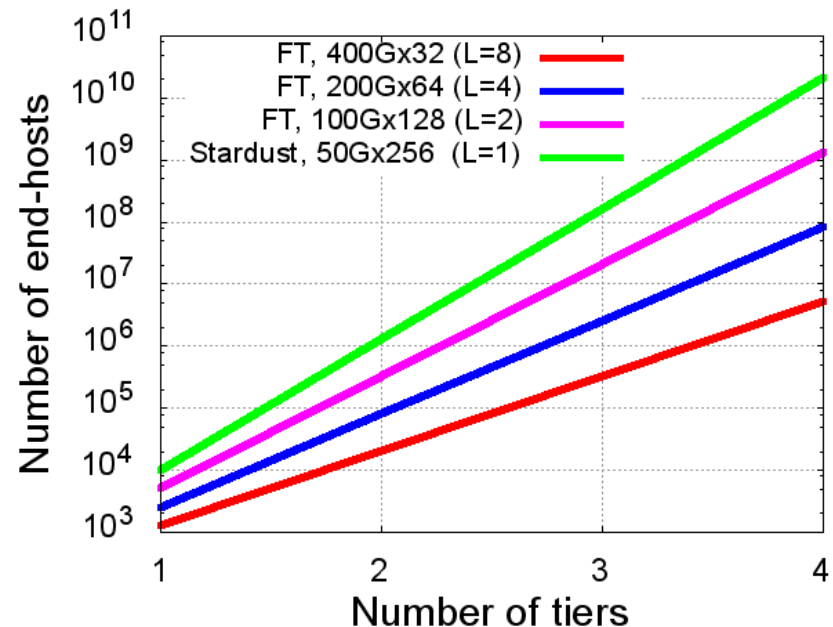
- Example: Building DC with 100K servers (2500 ToR switches)
- Option 1 – Link bundle of 1 ($L=1$):
 - 6.4Tbps Fabric Switch, 256×25G
 - Requires 2 Tiers

#fabric-switches = 1172

- Option 2 – Link bundle of 4 ($L=4$):
 - 6.4Tbps Fabric Switch, 64×100G
 - Requires 3 Tiers

#fabric-switches = 1954 ($\times 1.66$ more)

In a network of n tiers
scale is $O(L^{-n})$



Scalability limitations

- ✗ Significant resources are wasted within the network fabric
- ✗ Aggregating links into ports limits the scalability of the network
- ✗ Misaligned packet sizes require significant resource overprovisioning (or lead to performance loss)
- ✗ To support full line rate for all packet sizes, network devices need to process multiple packets each and every clock cycle.

What should we do?

My personal opinion

- Think before you code
 - Make it easier for the network: batching, large packets, ...
- Think again: do you really need this feature ?
 - Maybe there is a different way to get the same results?
 - Maybe a trade-off is worth the cost?
- Make the core of the network simple, push complexity to the edge
 - And then push complexity outside the DC
 - Why? Latency, privacy, data is too big to move, ...
 - (not always possible)

P51 Summary

- Architecture of high-performance network devices
- High throughput devices
- Low latency devices
- Programmable network devices



- This is just a little glimpse into high performance networking...