It's the End of DRAM As We Know It

Philip Levis
Stanford University
IETF ANRW July 2023

Some Apologies

- I used to research networks
  - Low-power wireless: 802.15.4, LoRa
  - Internet of Things/sensor networks: CTP, LowPAN interoperability
  - IETF: RPL (RFC6550), Trickle (RFC6206), MRHOF (RFC6719)
  - Security: TLS-RaR
  - Congestion control: Pantheon (hi Francis!)
  - Video streaming: Puffer (hi Francis!)
- I haven't in a few years! So I don't have a lot to say about networks.
- I'm instead going to talk about a looming challenge for computing applications and systems. This will affect networks, what they can do, and how applications will use them. I think it's interesting. I hope you do too!

The Summary

- Processing and network speeds will continue to improve for a while
- RAM price (per bit) won't go down for at least 10 years, probably longer
- RAM performance (latency, throughput) is also flat
- This divergence means computers will be very different in 10 years
- And so will their applications

- It's the end of DRAM as we know it

Outline

- What's happening with scaling: $/bit
- What's happening with signaling: latency and throughput
- Three kinds of memory
- A twist: Compute Express Link (CXL)

Outline

- What's happening with scaling: $/bit
- What's happening with signaling: latency and throughput
- Three kinds of memory
- A twist: Compute Express Link (CXL)

DRAM Trends

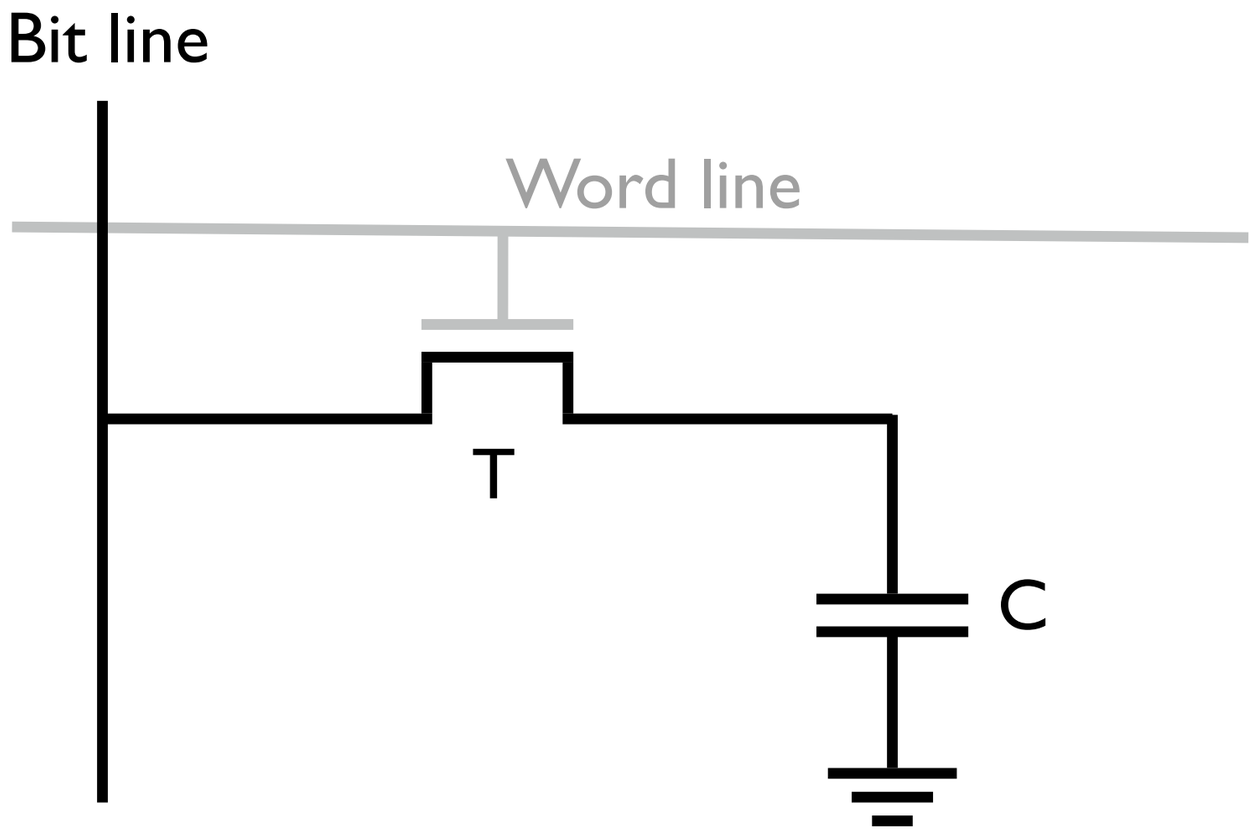
Version	Year	Throughput	Latency	$/GByte
DDR	1998	3.2GBps	134ns	$78
DDR2	2003	8.5GBps	122ns	$9
DDR3	2007	17GBps	79ns	$3
DDR4	2014	26GBps	74ns	$2
DDR5	2020	57GBps	72ns	$3
DDR6	2025?	104?GBps	?	?

DRAM Trends

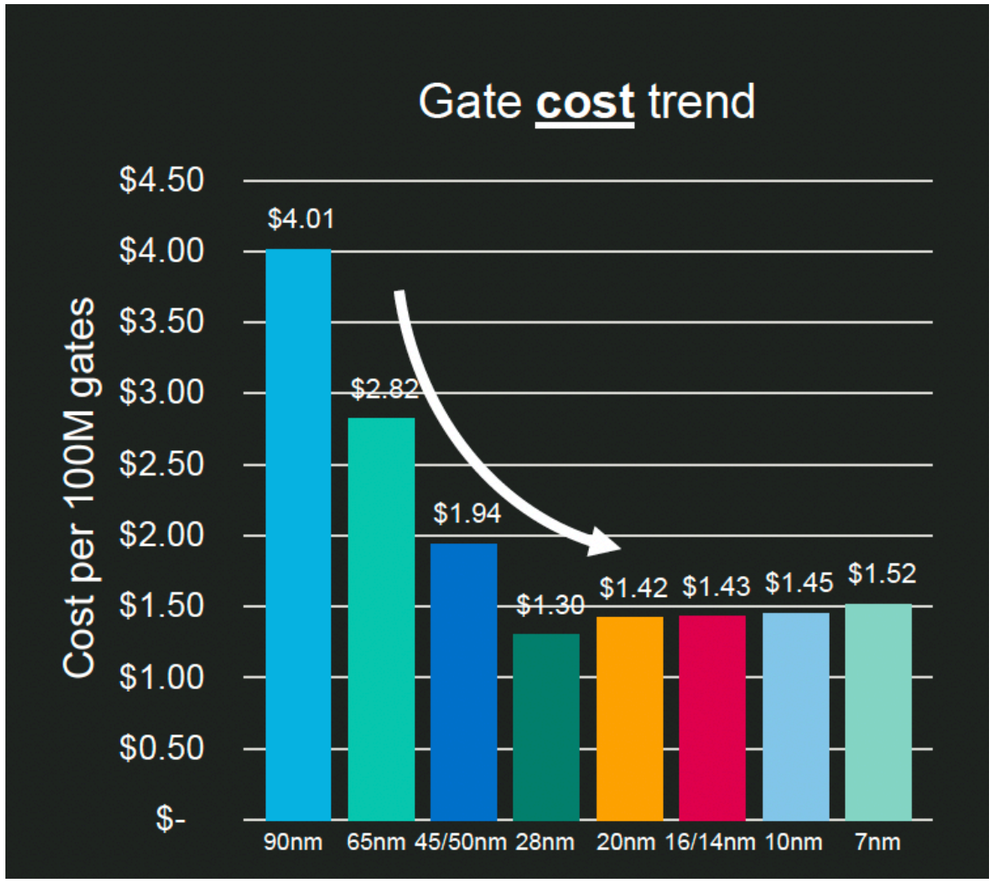
Version	Year	Throughput	Latency	$/GByte
DDR	1998	3.2GBps	134ns	$78
DDR2	2003	8.5GBps	122ns	$9
DDR3	2007	17GBps	79ns	$3
DDR4	2014	26GBps	74ns	$2
DDR5	2020	57GBps	72ns	$3
DDR6	2025?	104?GBps	?	?

} Cost is flat

DRAM is a Transistor

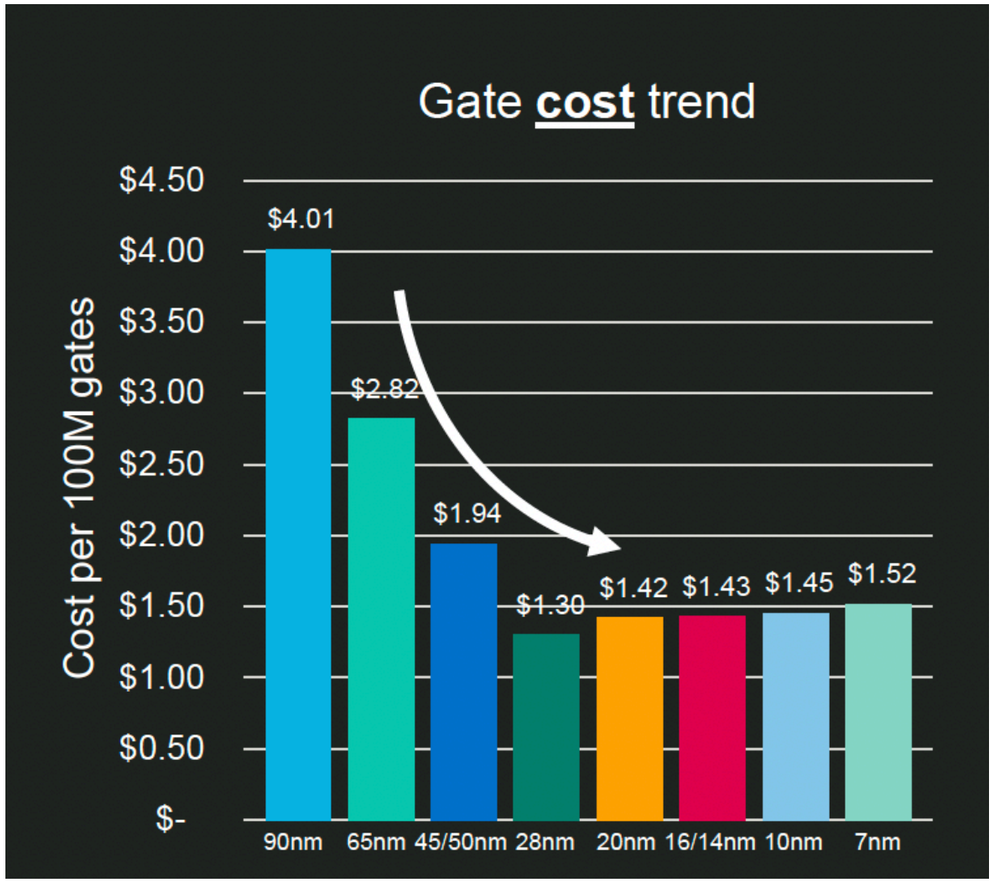


End of Scaling



Marvell Investor Day 2020 presentation, slide 43

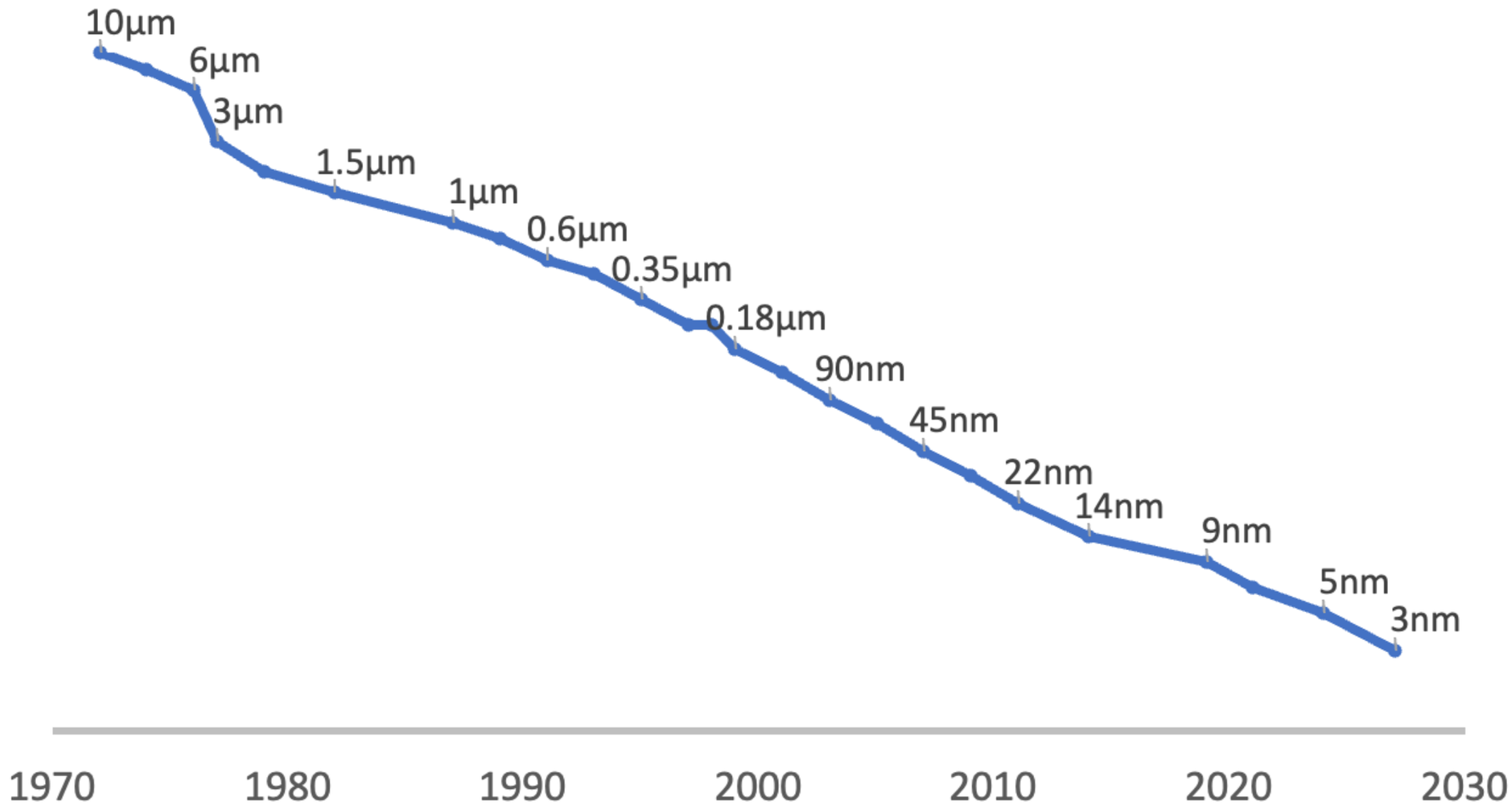
End of Scaling



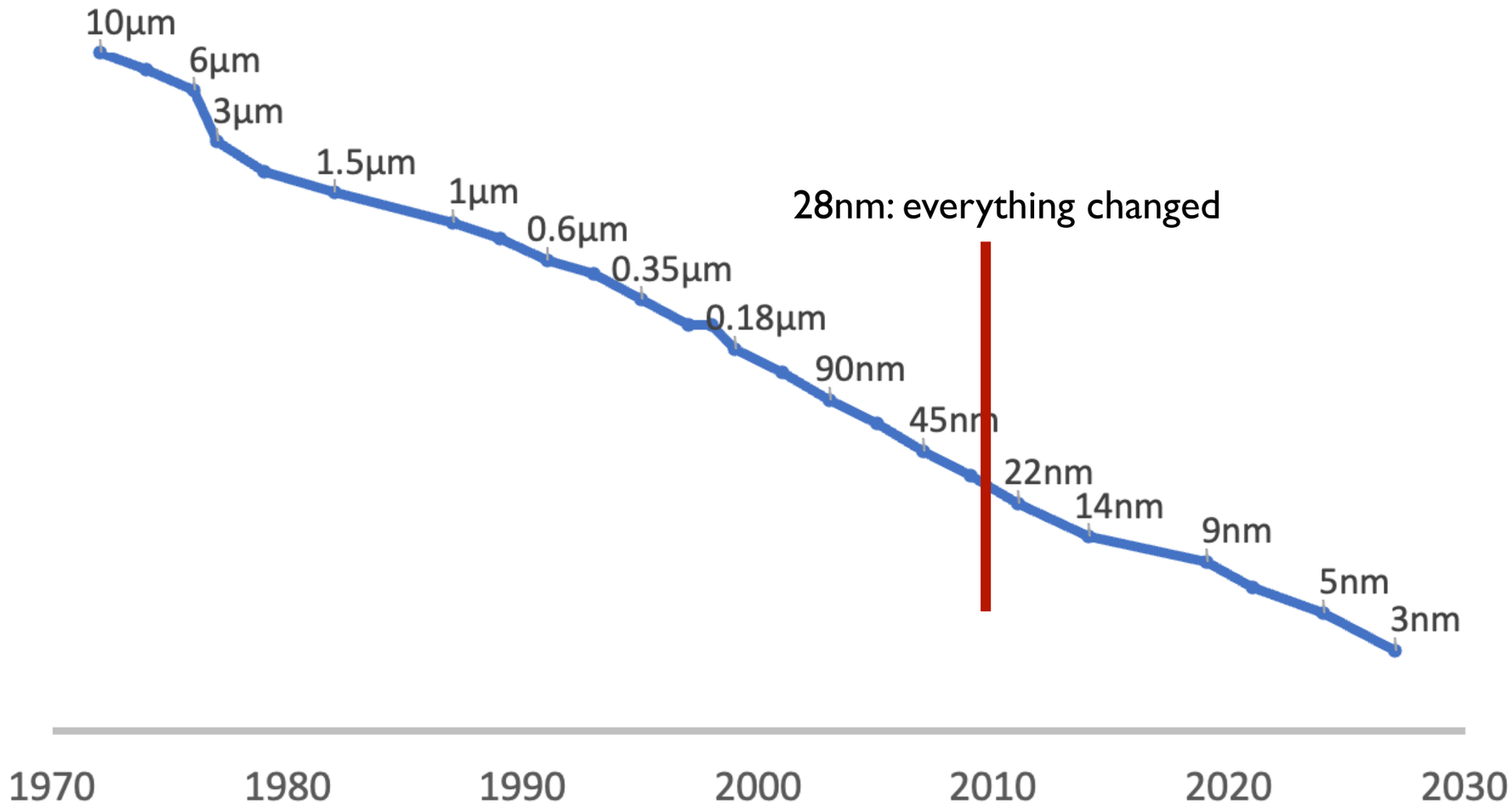
- For a long time, smaller meant cheaper
- No longer
- We can continue to make smaller transistors for a little while, but transistor costs are flat

Marvell Investor Day 2020 presentation, slide 43

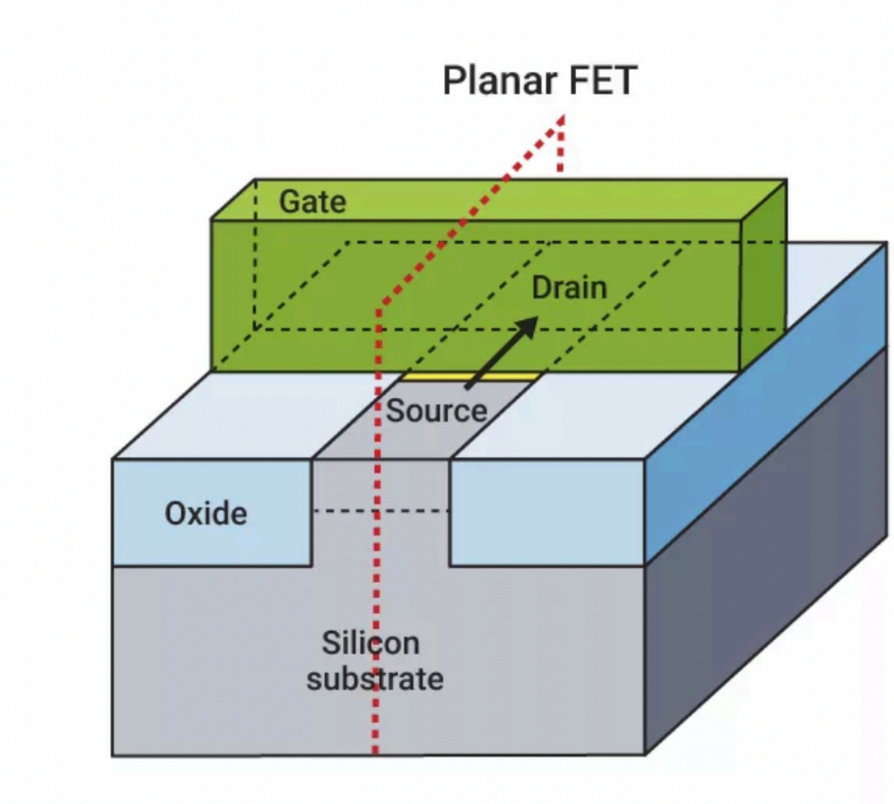
Process Node

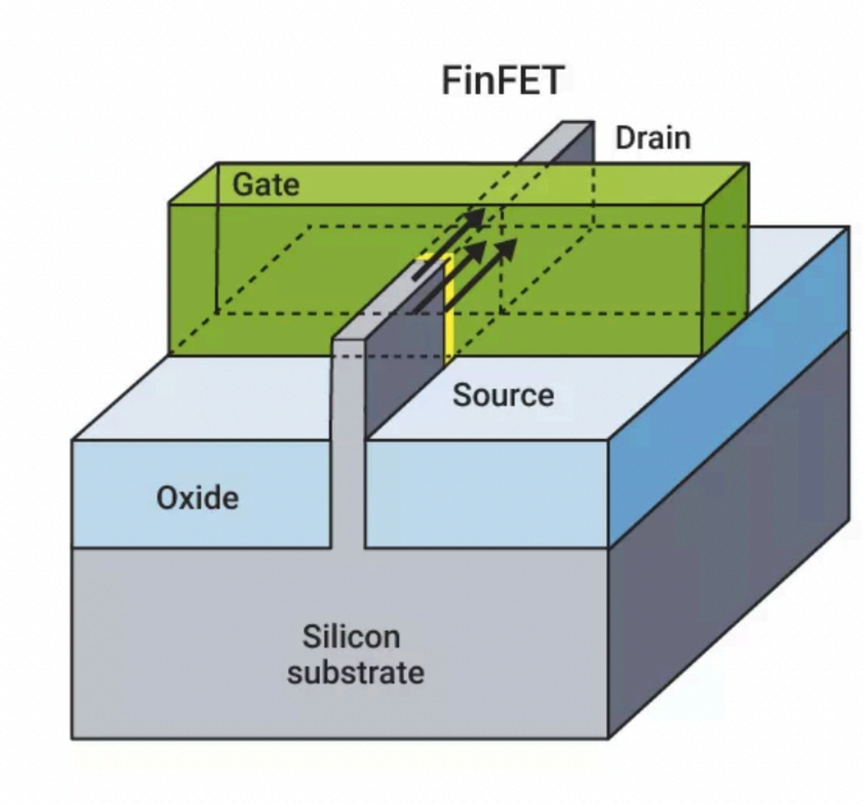


Process Node

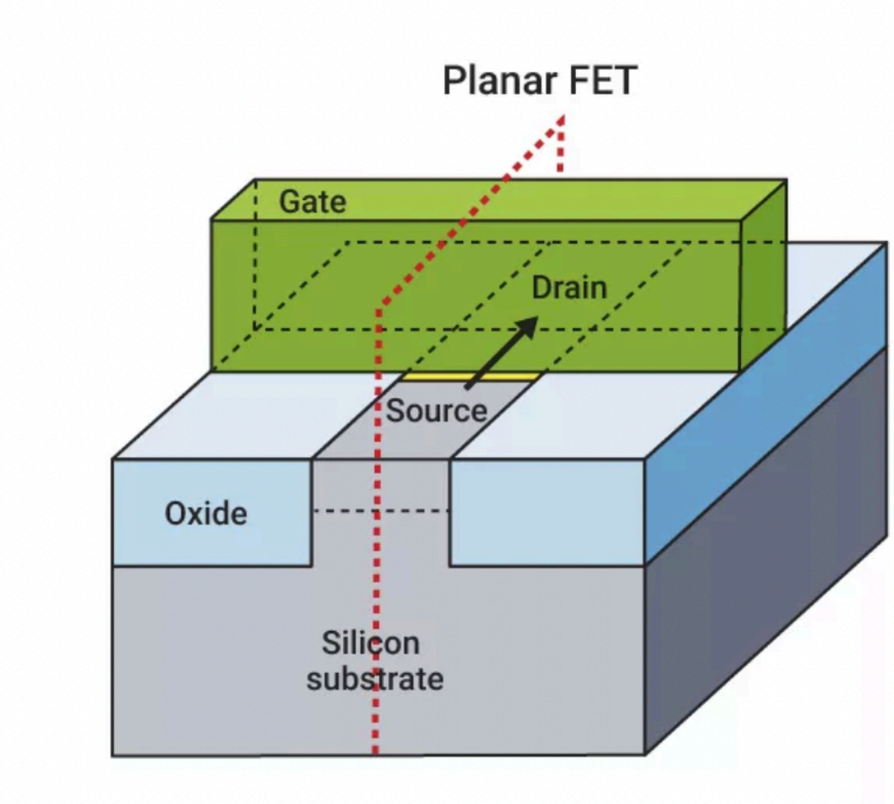


Why?

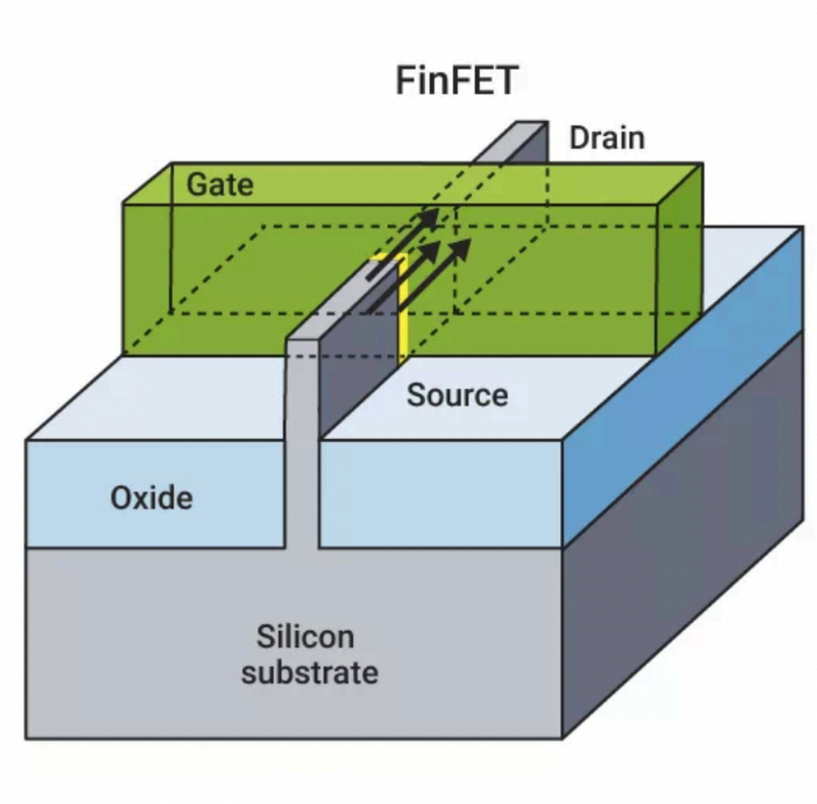

Planar FET
Gate
Drain
Source
Oxide
Silicon
substrate


FinFET
Drain
Gate
Source
Oxide
Silicon
substrate

Why?

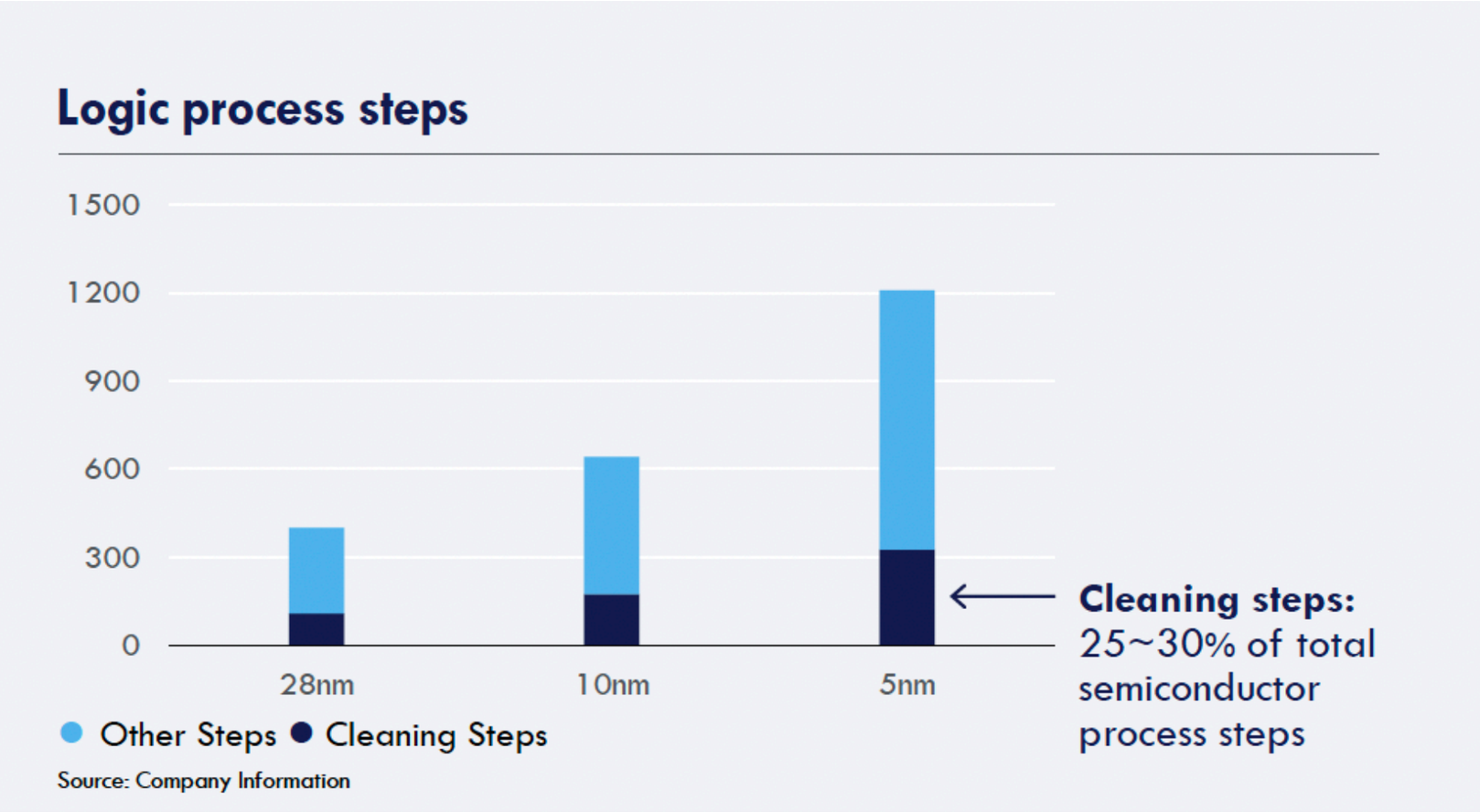


>28nm



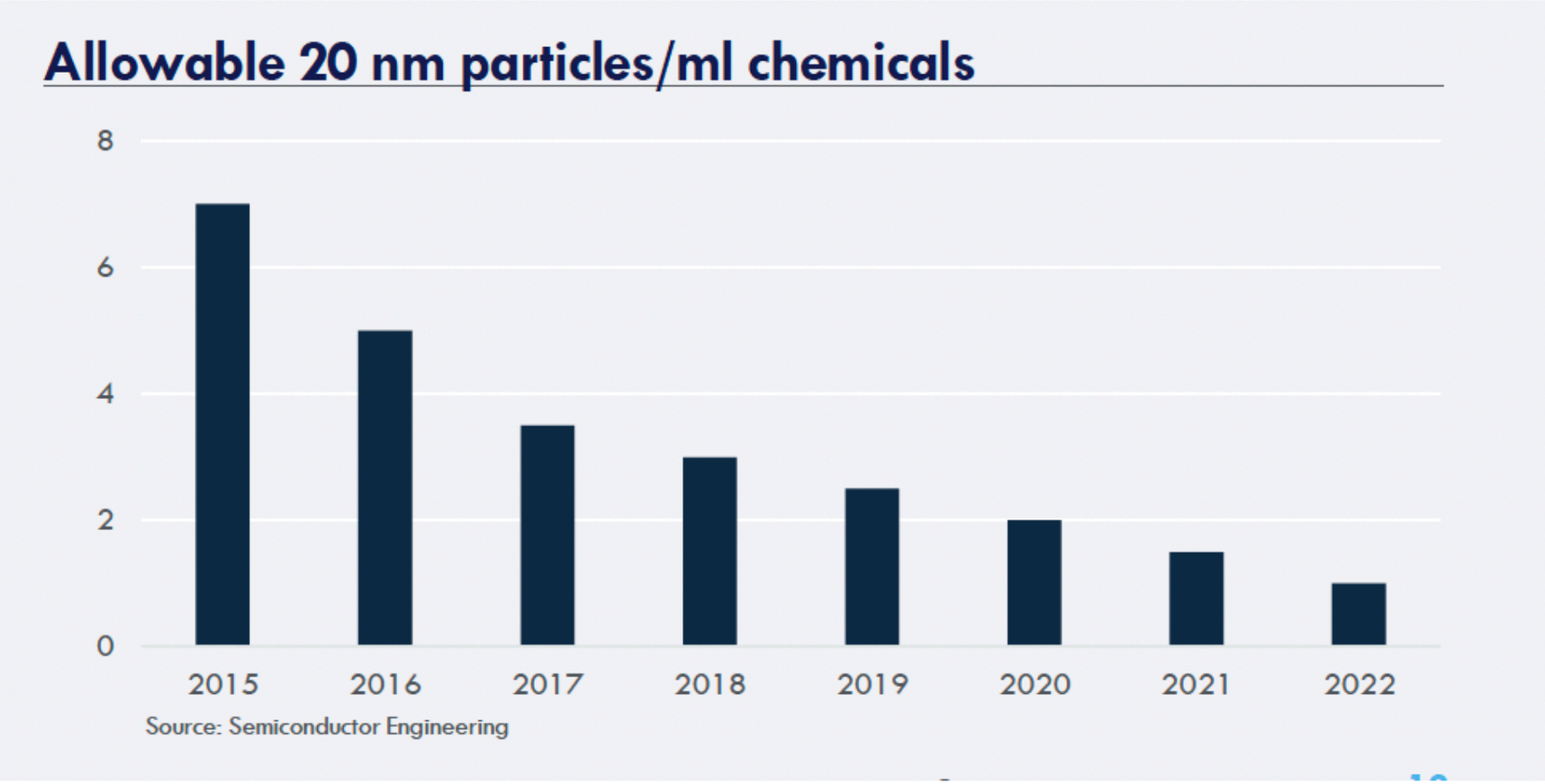
28nm - 5nm

It's Harder and Harder



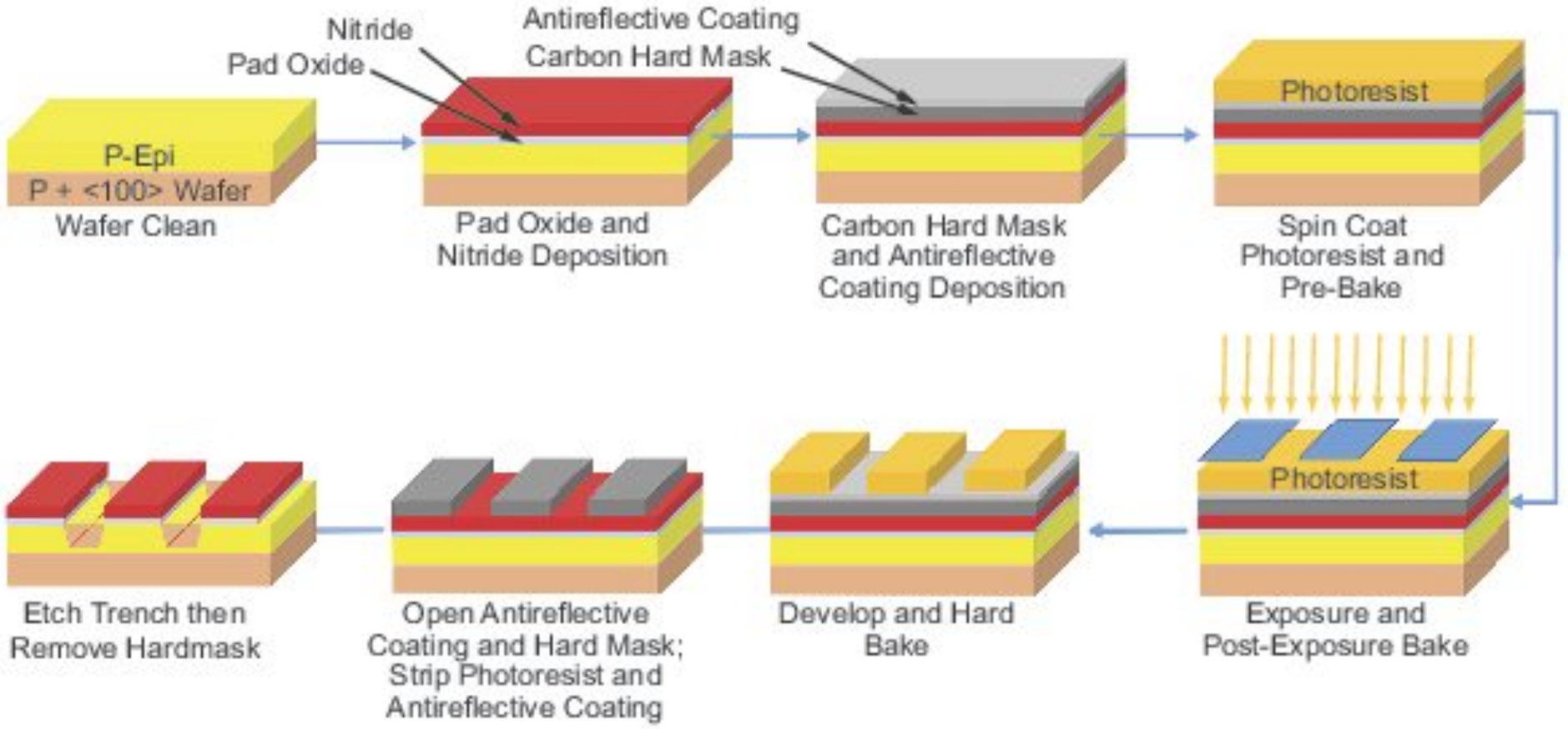
CMC Investors Presentation, 2021

In More Ways Than One



CMC Investors Presentation, 2021

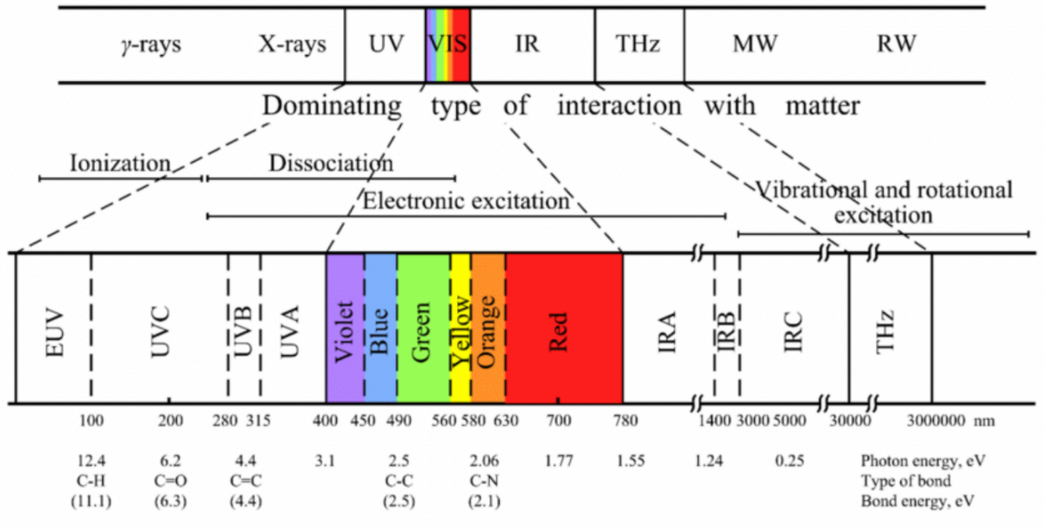
Lithography



https://www.newport.com/n/photolithography-overview

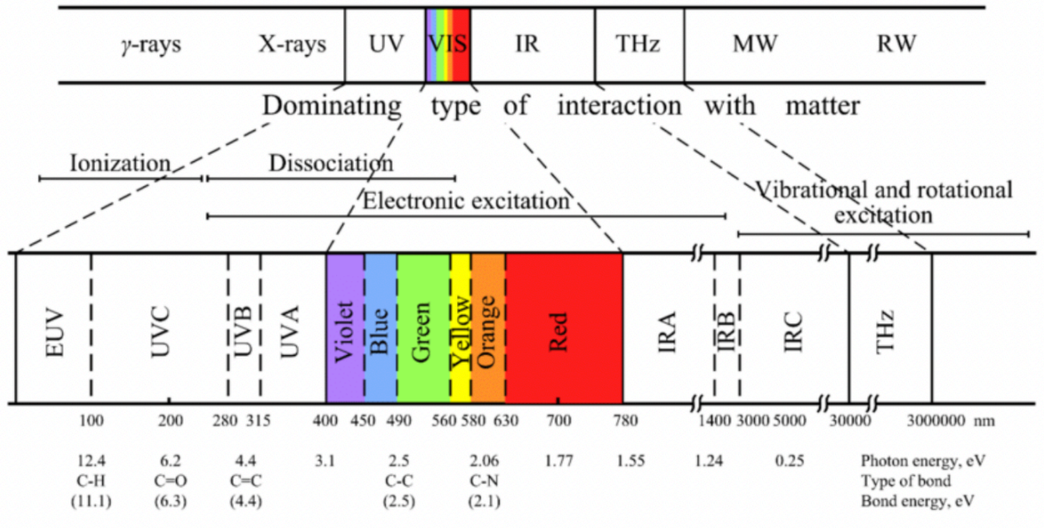
Patterning at < 28nm

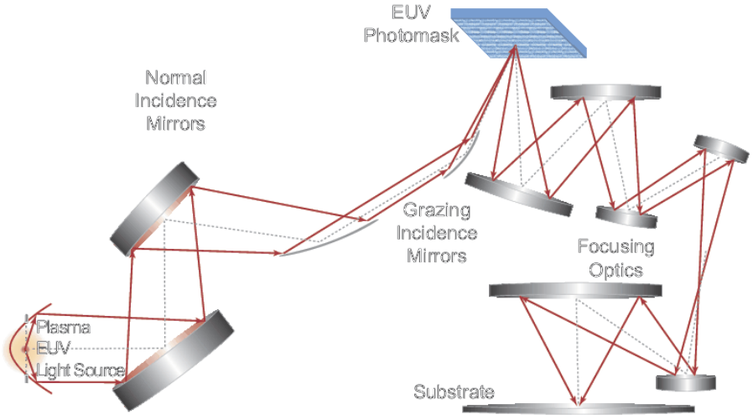
- Lithography at < 28nm requires extreme ultraviolet (EUV) light (10-100nm wavelengths)

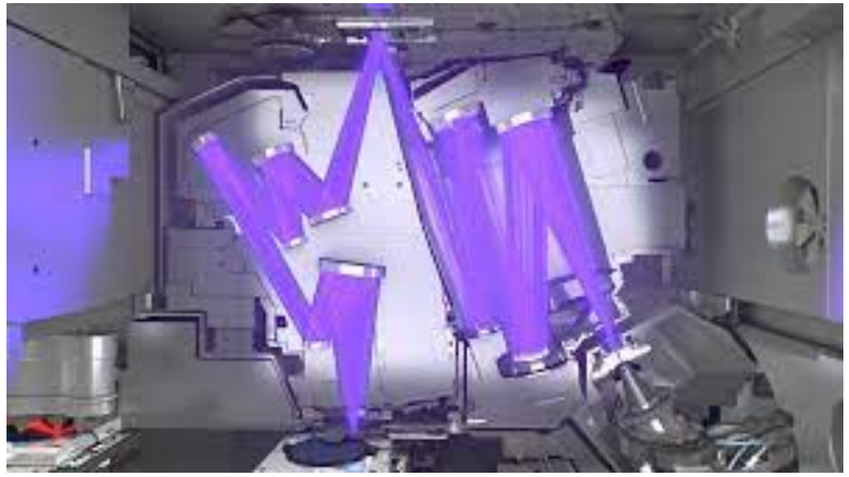


Patterning at < 28nm

- Lithography at < 28nm requires extreme ultraviolet (EUV) light (10-100nm wavelengths)
- Problem: glass, air, and almost everything absorb EUV
  - Have to operate in vacuum
  - No lenses
  - Mirrors

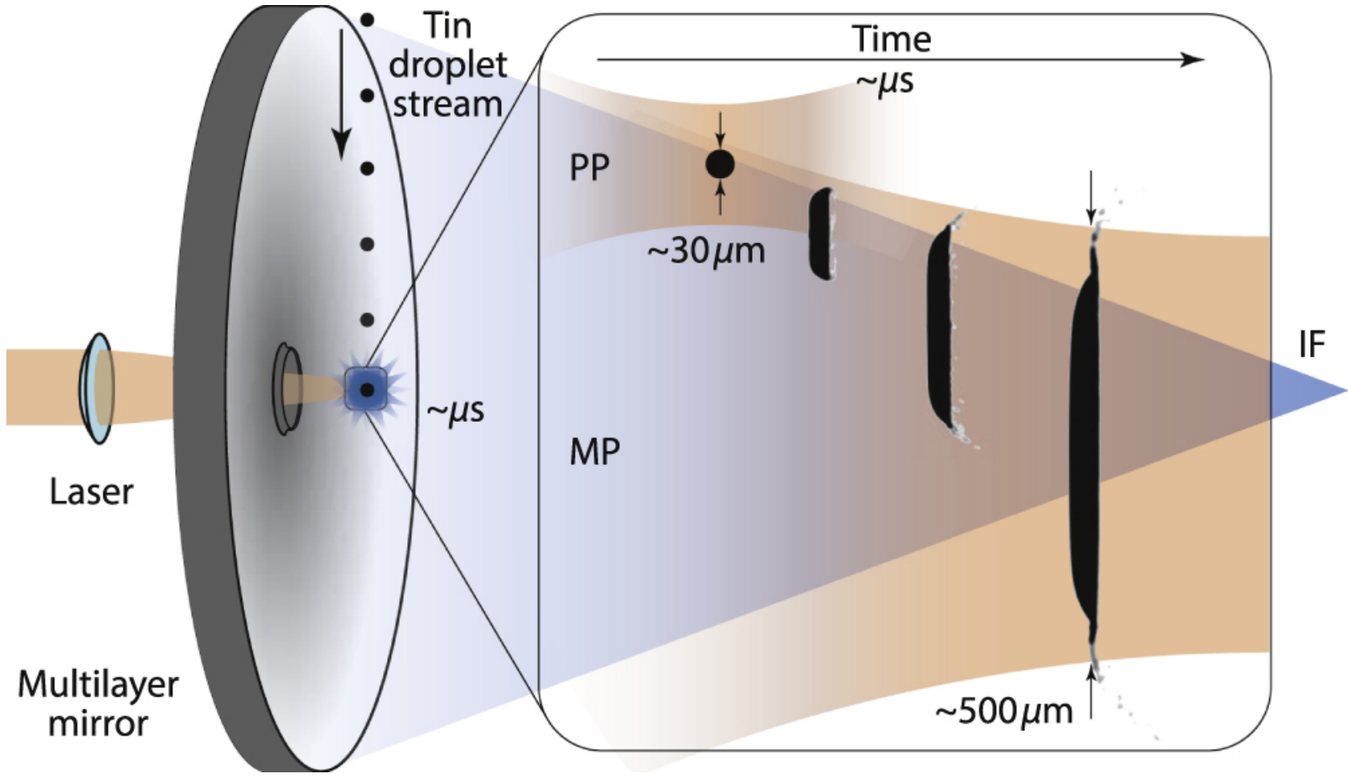






Generating EUV

- Prepulse (PP) hits tiny droplets of tin
- Droplets spreads into a disc
- Disc irradiated by main pulse (MP)
- Tin plasma produces 13.5nm light
- 50,000 droplets/second



EUV Lithography

- Took decades of research, tens of billions of dollars of research
- Many thought it was impossible
- "There is no plan B"
- We can continue to make smaller transistors, but the per-transistor manufacturing cost is flat

What About CPUs?

- Price per core is going down slightly
- Cores are getting faster (better accelerators, etc.)
- Design is a higher fraction of CPU costs
- You can use transistors more efficiently
- Chiplet designs are reducing design costs

Year	Processor	Cores	Transistors	Cost	$/core
2019	Rome	64	40T	$6950	$109
2022	Milan	64	26T	$8800	$138
2022	Genoa	96	90T	$10,625	$110
2024	Bergamo	128	82T	$11,900	$92

Prices from http://en.wikipedia.com/wiki/Epyc

Price Per Bit

- The price per bit of DRAM is not going down soon
  - It's the cost of manufacturing transistors
  - Unlike CPUs, there isn't flexibility/design space
- Much cheaper RAM will require new materials
  - Nothing on the roadmap for anyone: don't expect anything in the next 10 years
  - Micron representative: "We've tried everything in the periodic table"
  - Optane was a 2x, but not good enough for the market now, it might return
  - Long shot: multi-layer DRAM, multiple transistors with one lithographic pass (like flash)
- Processors and networks still have some runway
  - NICs are a few nodes behind
  - Still a lot of room for acceleration in CPUs (e.g., Sapphire Rapids accelerators)

Outline

- What's happening with scaling: $/bit
- What's happening with signaling: latency and throughput
- Three kinds of memory
- A twist: Compute Express Link (CXL)

DRAM Trends

Version	Year	Throughput	Latency	$/GByte
DDR	1998	3.2GBps	134ns	$78
DDR2	2003	8.5GBps	122ns	$9
DDR3	2007	17GBps	79ns	$3
DDR4	2014	26GBps	74ns	$2
DDR5	2020	57GBps	72ns	$3
DDR6	2025?	104?Gbps	?	?

DRAM Trends

Version	Year	Throughput	Latency	$/GByte
DDR	1998	3.2GBps	134ns	$78
DDR2	2003	8.5GBps	122ns	$9
DDR3	2007	17GBps	79ns	$3
DDR4	2014	26GBps	74ns	Latency is flat
DDR5	2020	57GBps	72ns	$3
DDR6	2025?	104?Gbps	?	?

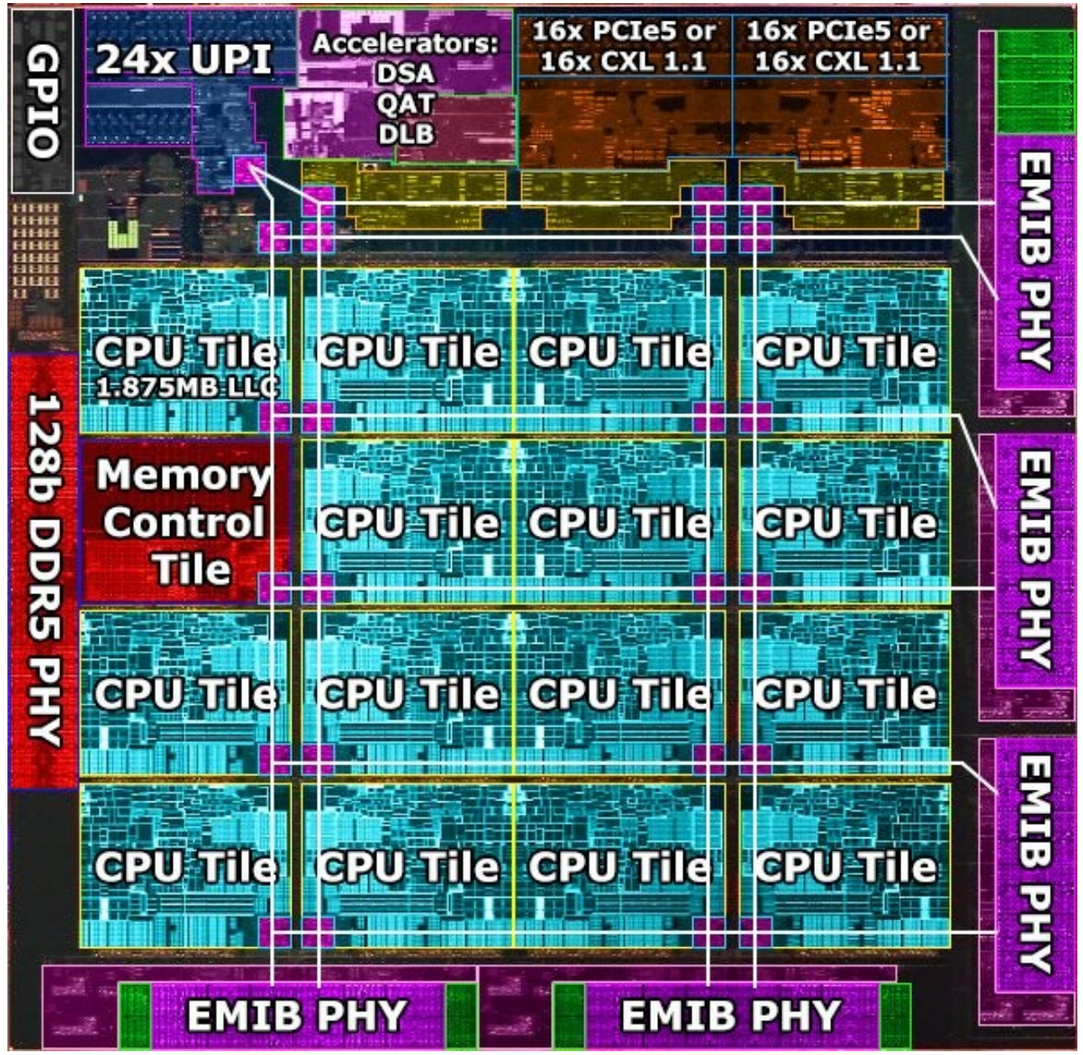
CPU Observed Latency

- Observed CPU latency is increasing
- Interconnect and coherence
- Bigger caches, fewer misses

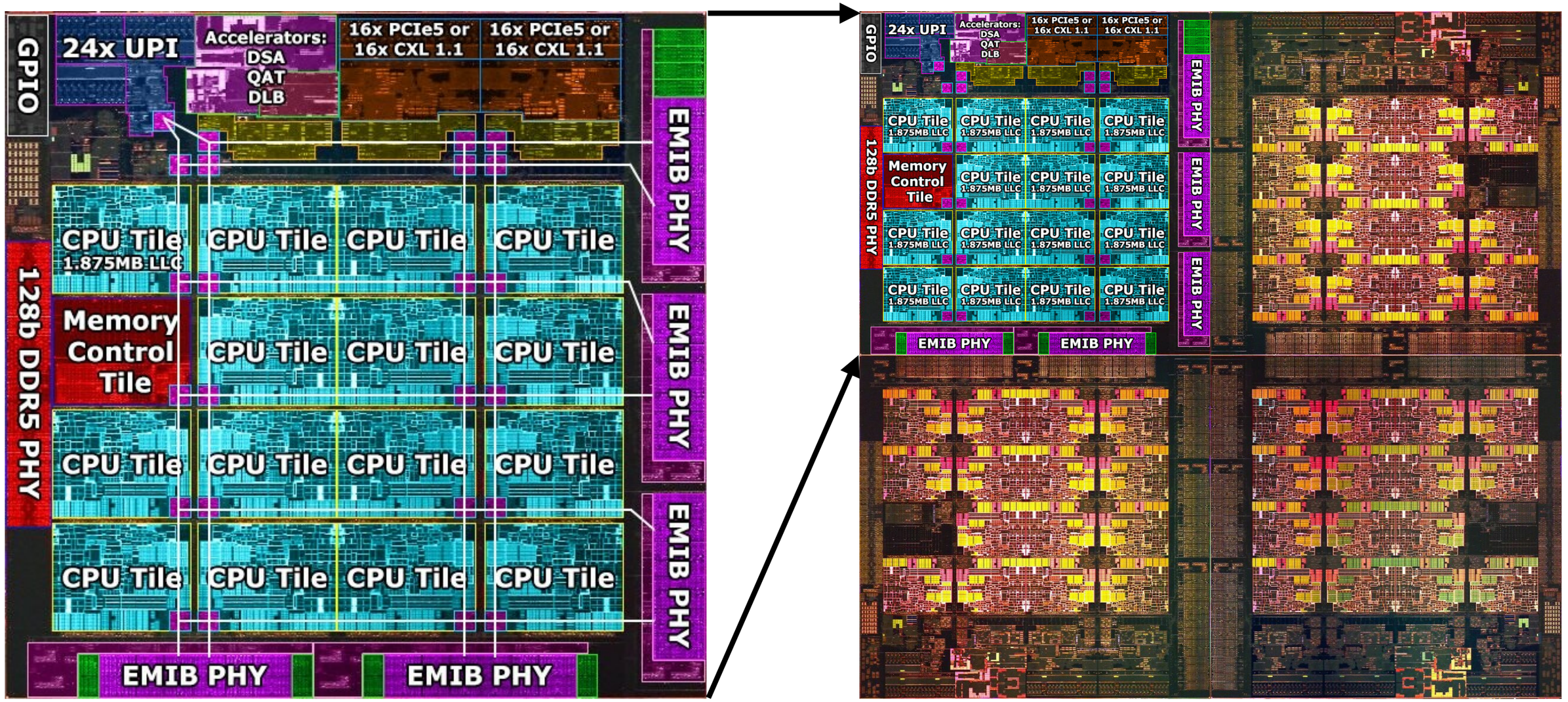
CPU	Arch	Cores	Latency
Xeon 8160T	Skylake	24	87ns
Xeon 8272CL	Cascade Lake	26	124ns
Xeon 8370C	Ice Lake	32	117ns
Xeon 8480	Sapphire Rapids	56	142ns

Data from https://chipsandcheese.com/memory-latency-data/

Sappire Rapids Die

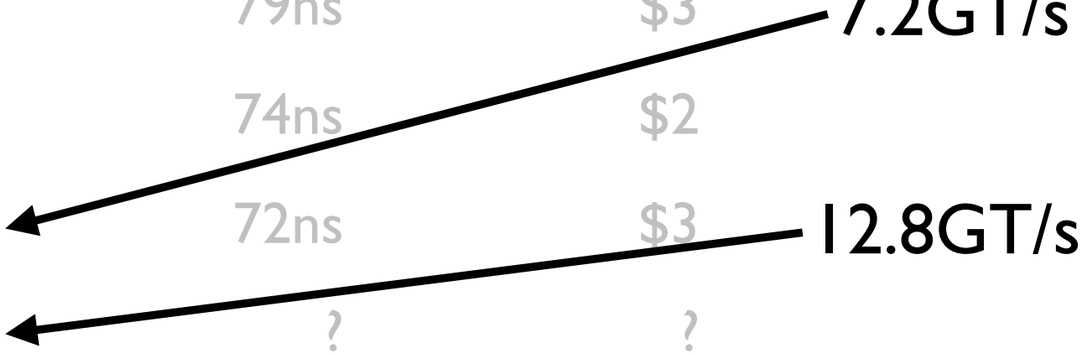


Sappire Rapids Chip



DDR Throughput (per DIMM)

Version	Year	Throughput	Latency	$/GByte
DDR	1998	3.2GBps	134ns	$78
DDR2	2003	8.5GBps	122ns	$9
DDR3	2007	17GBps	79ns	$3
DDR4	2014	26GBps	74ns	$2
DDR5	2020	57GBps	72ns	$3
DDR6	2025?	104?Gbps	?	?



Signaling Limits

- DRAM data lines are single-ended
- Single-ended data lines have a signaling limit of ~9.6Gbps
- DDR6 tries to push this further by adding buffering (increases latency)

Signaling and Latency

- DDR latency is not going down
  - System memory latency is going up due to more complex caches
- DDR is reaching its signaling limits (9.6Gbps/pin)
  - Can maybe go to 12.8Gbps with buffering, increasing latency
- One possible escape hatch: go from single-ended to differential signaling
  - Completely new DRAM designs
  - No current JEDEC plans
  - Some historical concerns

It's the End of DRAM As We Know It

- Cost per bit of DRAM isn't going down in the medium term
- DRAM latency isn't going down either
- DRAM bandwidth doesn't have much room left
  - Unless next DDR uses differential signaling, this is >5 years out, so 8 years out before you can buy it

Outline

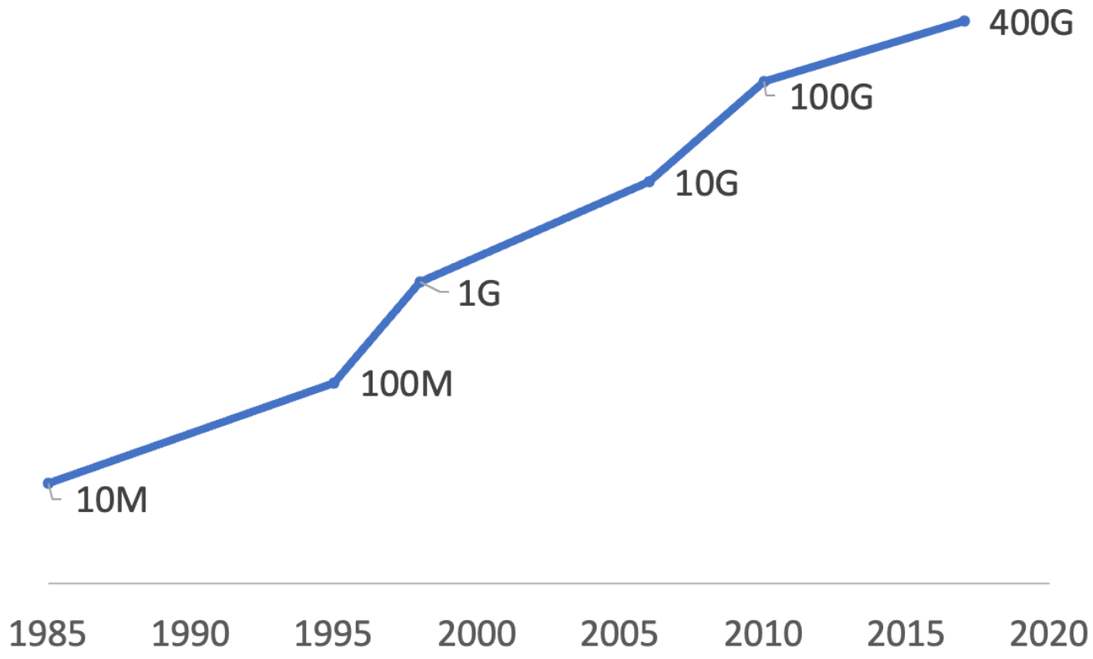
- What's happening with scaling: $/bit
- What's happening with signaling: latency and throughput
- Three kinds of memory
- A twist: Compute Express Link (CXL)

Performance Diverges

- RAM capacity, latency, and throughput are flat
- Everything else continues to improve

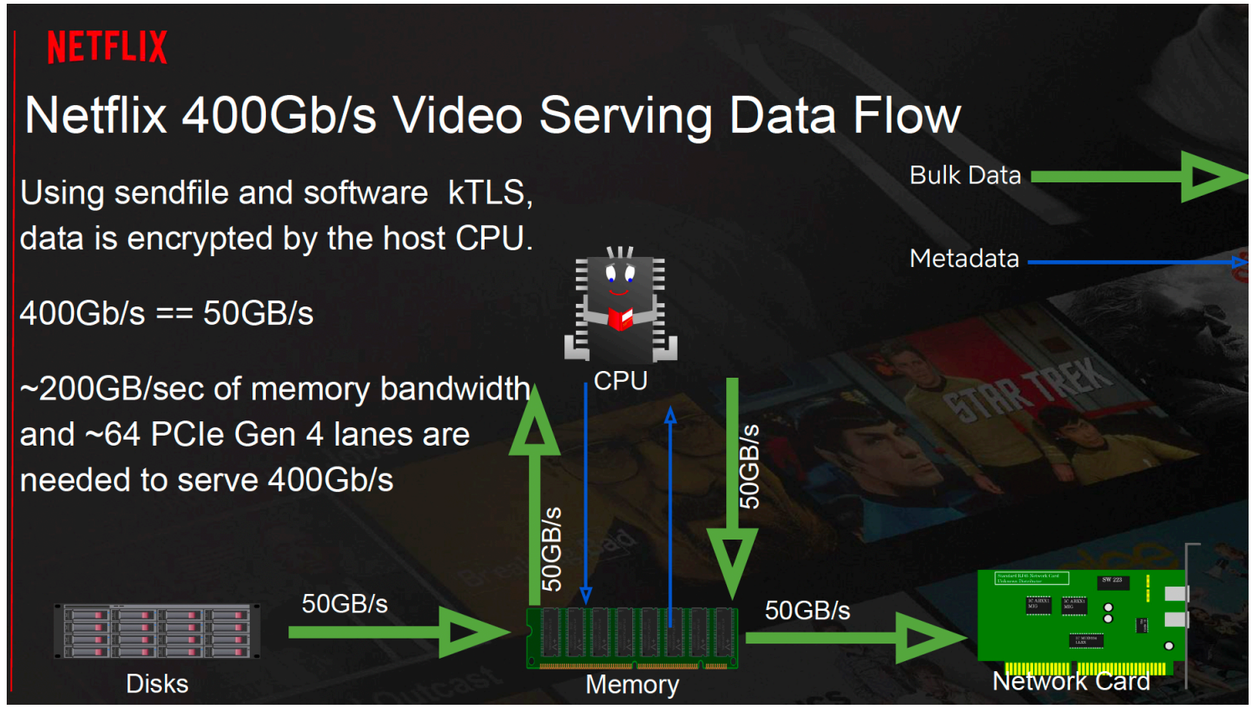
NIC Speeds

- NIC speeds: 200G today, 400G soon
- At 400G, a 4kB packet is 80ns
  - Less than a single cache miss
- At 400G, a 64B packet is 1.25ns
- 400Gbit is 10% of server memory bandwidth
  - Echoing packets is 20% of bandwidth (write into RAM on reception, read from RAM for transmit)



More Complex Pipelines: Netflix

- 4x memory amplification
  - DMA data from disk to memory
  - Read data from memory into CPU
  - Write encrypted data to memory
  - DMA data from memory to NIC
- What about more operations?
  - Compression
  - Serialization
  - Comment from engineer at Google: "We architect our pipeline so data enters the cache only once."



https://papers.freebsd.org/2021/eurobsdcon/gallatin-netflix-freebsd-400gbps/

Accelerators

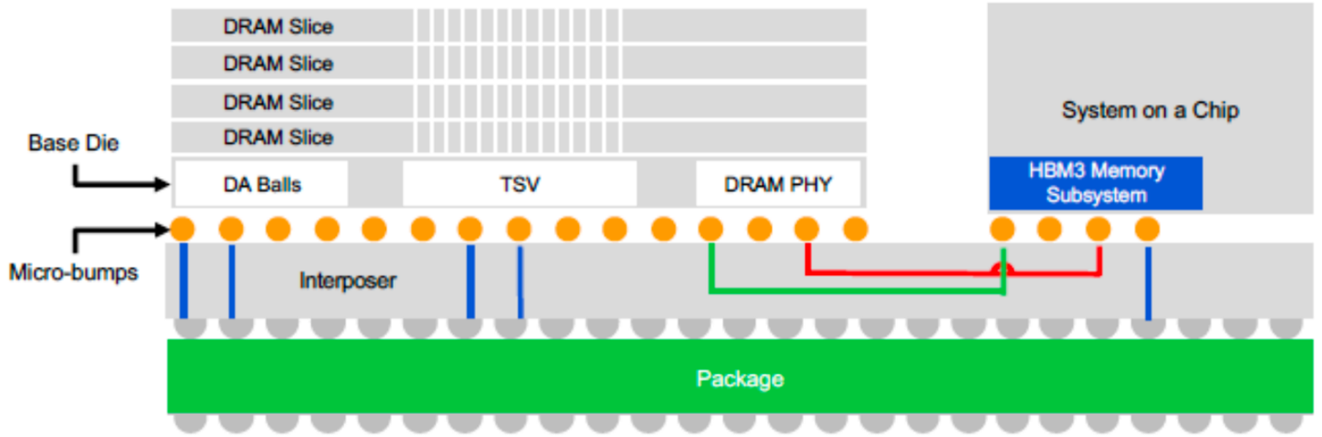
- Processors need to continue to improve performance and efficiency
- Answer: computational accelerators
- Intel Sapphire Rapids
  - DSA: Data Streaming Accelerator
  - QAT: Quick Assist Technology
  - AMX: Advanced Matrix Extensions
  - IAA: In-memory Analytics Accelerator

The Problem With Accelerators

- Performing a computation faster means it reads and writes memory faster
- In terms of silicon, computation is cheap and can be parallelized
  - E.g., GPUs, TPUs, other accelerators
- At some point you can't feed the accelerator fast enough: DDR doesn't have enough bandwidth

HBM: High Bandwidth Memory

- Basically, DDR with a lot more data lines
  - DDR5: 64 data line; HBM3: 1024 data lines (16 64-bit-wide channels)
  - Latency is higher: ~300ns
- Problem: you can't run 1024 copper data lines on a PCB
  - Thinnest traces are 0.152mm, 0.152mm x 1024 = 155.8mm with no spacing between them!
- Solution: do it in silicon, with an interposer
  - Think of it like a PCB done with silicon lithography
  - Have to package processor and memory
  - Expensive! And lots of failures



Coming to a CPU Near You

Intel® Xeon® CPU Max Series

Maximize bandwidth with the Intel® Xeon® CPU Max Series, the only x86-based processor with high-bandwidth memory (HBM). Architected to supercharge the Intel® Xeon® platform with HBM, Intel® Max Series CPUs deliver up to 4.8x better performance compared to competition on real-world workloads[1], such as modeling, artificial intelligence, deep learning, high performance computing (HPC) and data analytics.

Product brief: Intel® Xeon® CPU Max Series ›

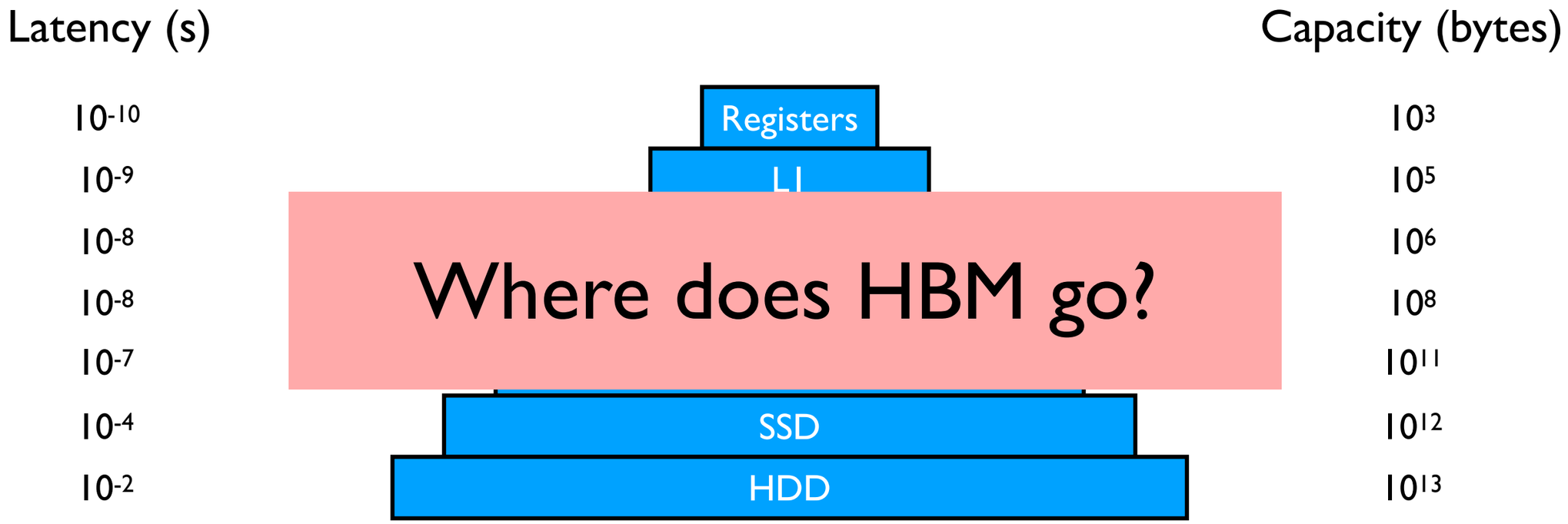
Overview | Products

SKU Number	Cores	Base (GHz)	All Core Turbo (GHz)	Max Turbo (GHz)	Cache (MB)	TDP (Watts)	Maximum Scalability	DDR5 Memory Speed	UPI Links Enabled	Default DSA Devices	Default QAT Devices	Default DLB Devices	Default IAA Devices	Intel SGX Enclave Capacity (Per Processor)	Recommended Customer Pricing (RCP) in $ US Dollars	Intel® On Demand Capable	Die Chop
HPC OPTIMIZED (Intel® Xeon® CPU Max Series)																	
9480	56	1.9	2.6	3.5	112.5	350	2S	4800	4	4	0	0	0	512GB	$12,980		XCC
9470	52	2	2.7	3.5	105	350	2S	4800	4	4	0	0	0	512GB	$11,590		XCC
9468	48	2.1	2.6	3.5	105	350	2S	4800	4	4	0	0	0	512GB	$9,900		XCC
9460	40	2.2	2.7	3.5	97.5	350	2S	4800	3	4	0	0	0	128GB	$8,750		XCC
9462	32	2.7	3.1	3.5	75	350	2S	4800	3	4	0	0	0	128GB	$7,995		XCC

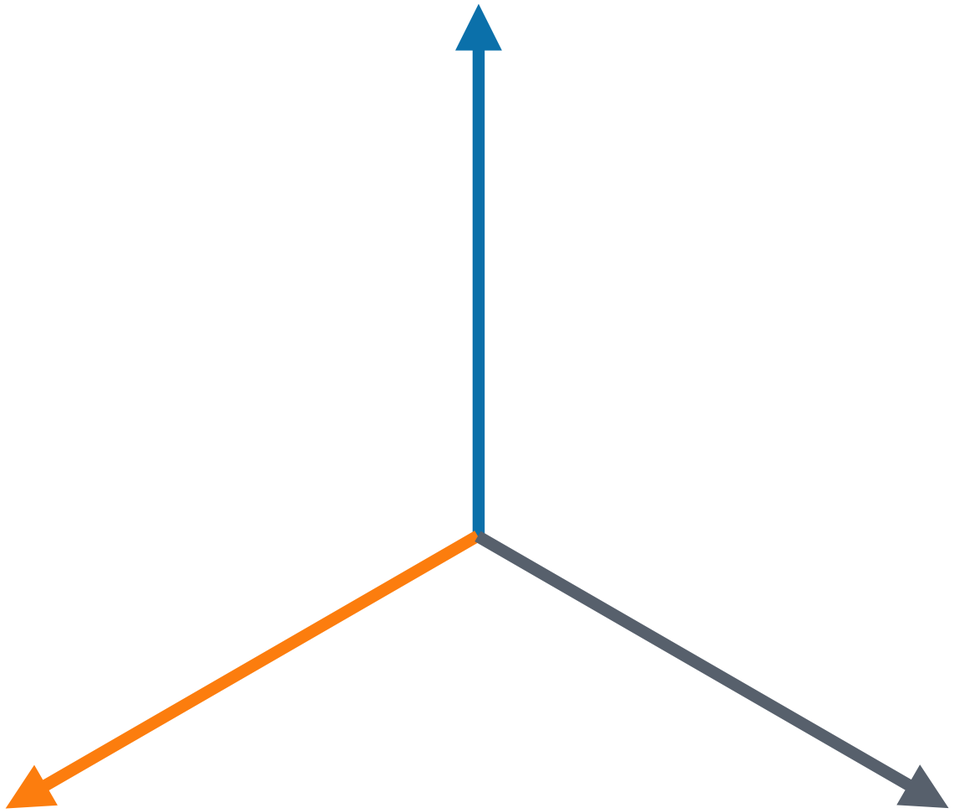
Memory Hierarchy

Latency (s)		Capacity (bytes)
10^{-10}	Registers	10^{3}
10^{-9}	L1	10^{5}
10^{-8}	L2	10^{6}
10^{-8}	L3	10^{8}
10^{-7}	DRAM	10^{11}
10^{-4}	SSD	10^{12}
10^{-2}	HDD	10^{13}

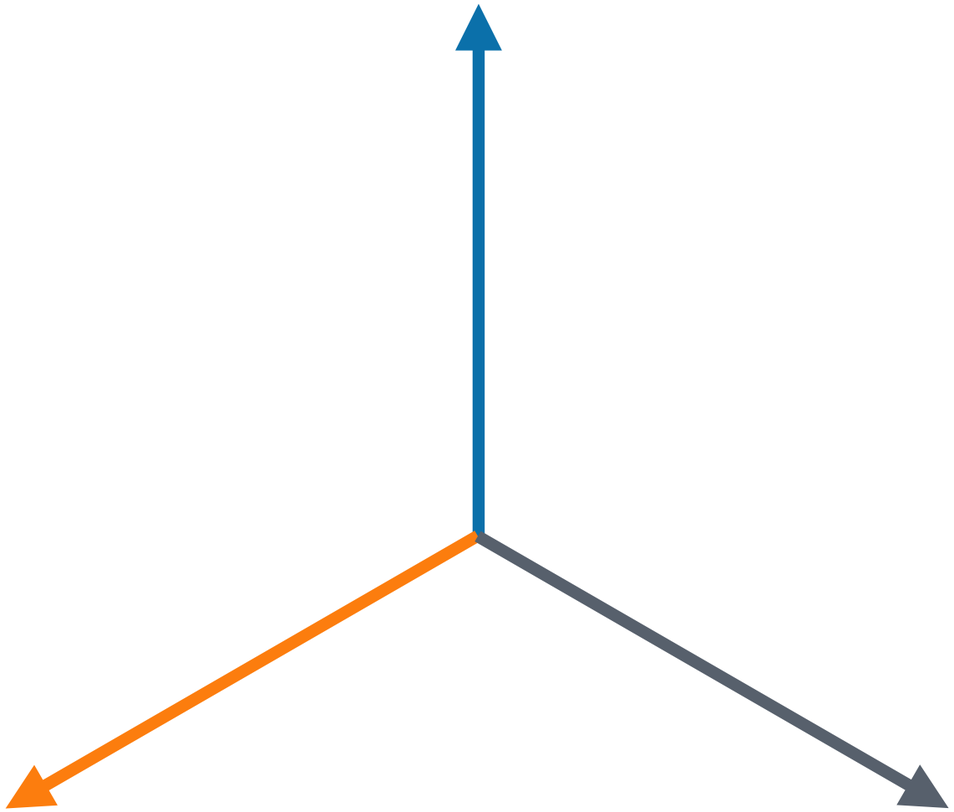
Memory Hierarchy



Three Types of Memory



Three Types of Memory



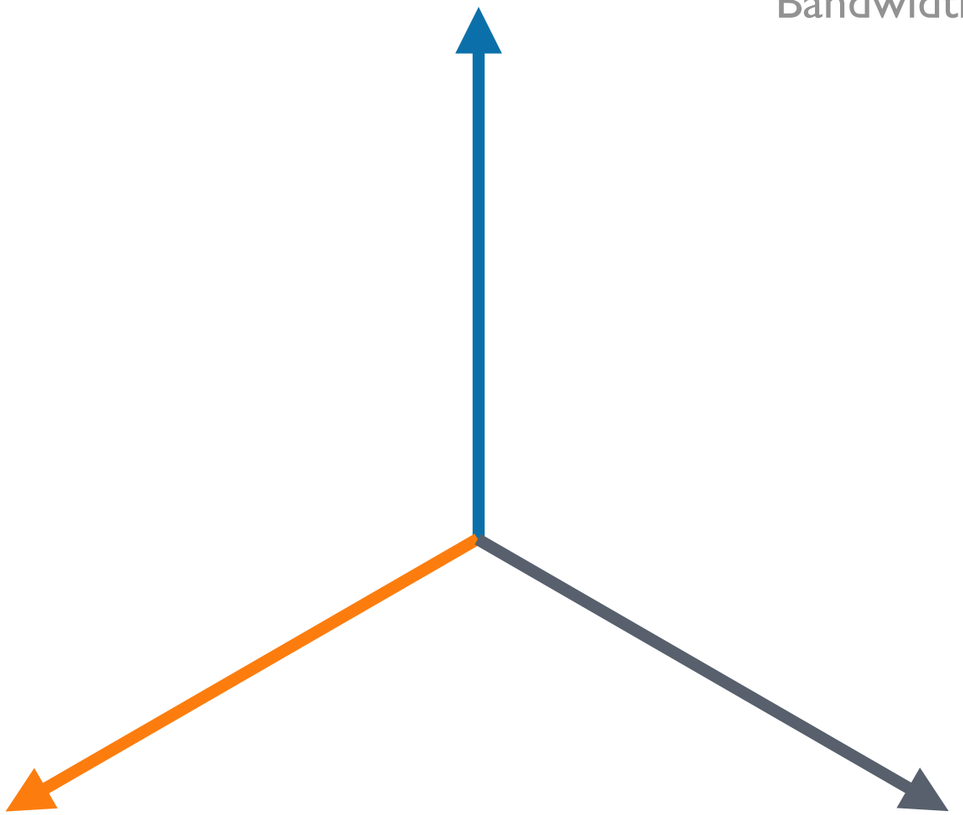
Size: 100s GB
Latency: 100ns
Bandwidth: 50GB/s

Latency: DDR

Three Types of Memory

Capacity: Flash

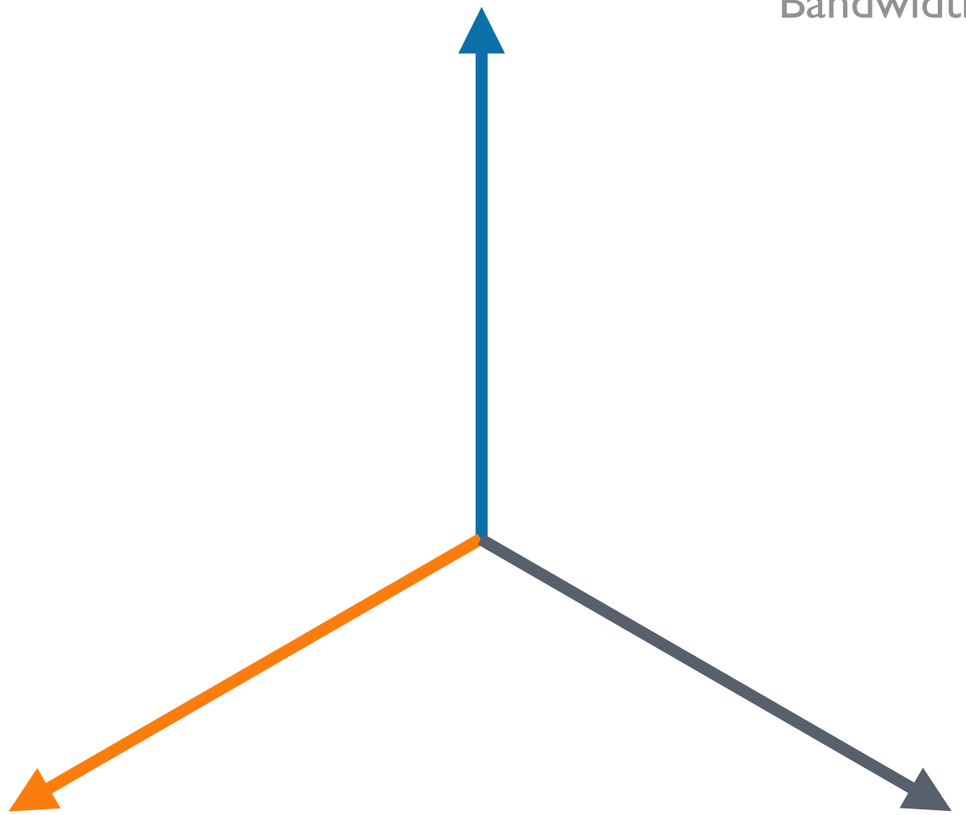
Size: TBs
Latency: 100µs
Bandwidth: 10s GB/s



Size: 100s GB
Latency: 100ns
Bandwidth: 50GB/s

Latency: DDR

Three Types of Memory



Future Computers

- They will have a mixture of memories: latency, capacity, and bandwidth
- Each memory is tied to a particular compute element
  - Disks/capacity: CPUs, or perhaps to the IPU/DPU
  - Latency: CPUs
  - Bandwidth: CPUs, TPUs, GPUs
- Networks are increasingly going to transmit data for bandwidth memory/ accelerated processing: input for models, etc.
- What does a networking stack look like when data can be loaded into a SmartNIC DRAM before offloading to a GPU's HBM?
- What happens when NICs need HBM?

Outline

- What's happening with scaling: $/bit
- What's happening with signaling: latency and throughput
- Three kinds of memory
- A twist: Compute Express Link (CXL)

PCIe is a problem

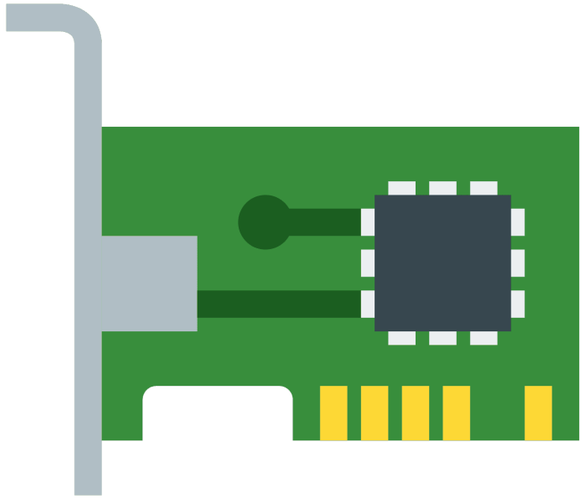
- High latency
  - Minimum PCIe latency is ~800ns
  - At 400Gbps, 800ns is 40kB
    - ➡ *27 1.5kB MTUs*
    - ➡ *10 4kB MTUs (common in datacenters)*
    - ➡ *5 8kB jumbo frames*
- High throughput
  - 16 PCIe Gen5 lanes is 480Gbps
    - ➡ *Overheads mean this isn't enough to drive 400Gbps; NVIDIA Bluefield 3 uses 32 Gen5 lanes*
- Just a data bus
  - Can read/write but memories on two sides of the bus are independent
  - Checking if a NIC has a packet requires reading across bus

Compute Express Link (CXL)

- Replacement for PCIe
  - Same physical layer, signaling, form factor, etc.: you can plug in a CXL or PCIe card, either/both will work
- Lower latency
  - Simplifies protocols to bring minimum latency down to 200ns
- Cache coherent
  - Allows two devices connected on the bus to use a MESI cache coherence protocol
  - A read from one device can be put in its cache, it can read it from the cache later
  - The other device, when it writes, invalidates the cache line

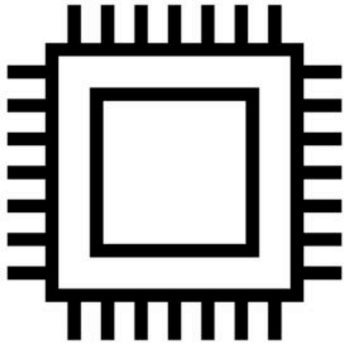
PCIe Example

NIC memory

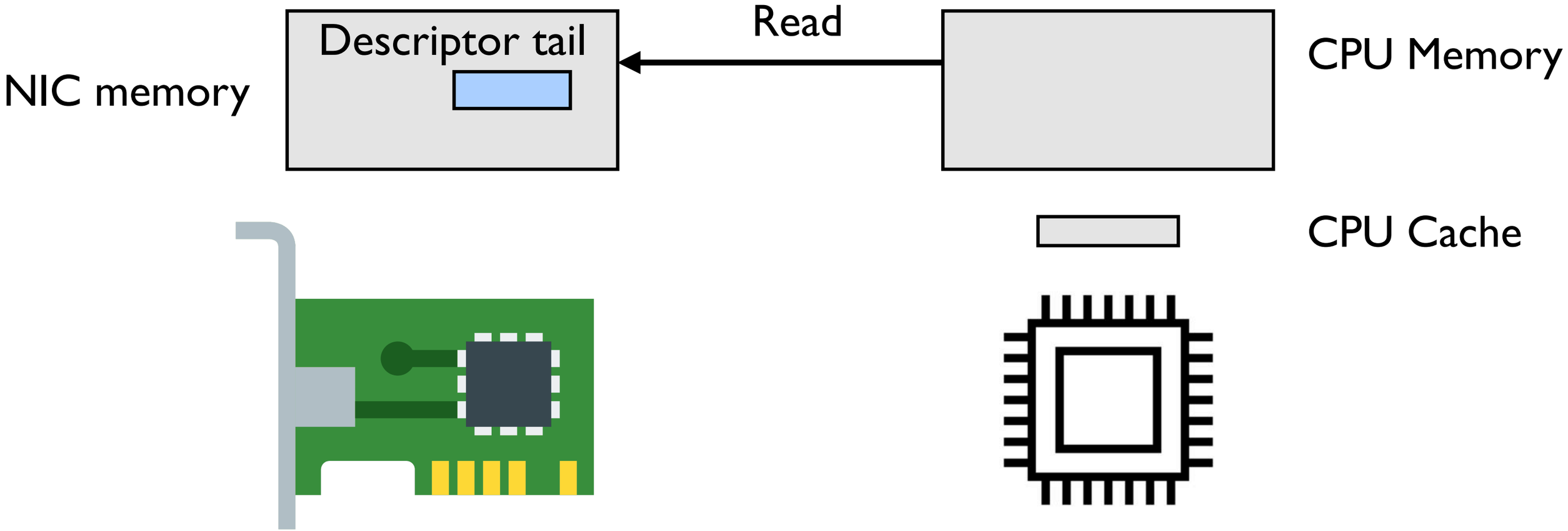


CPU Memory

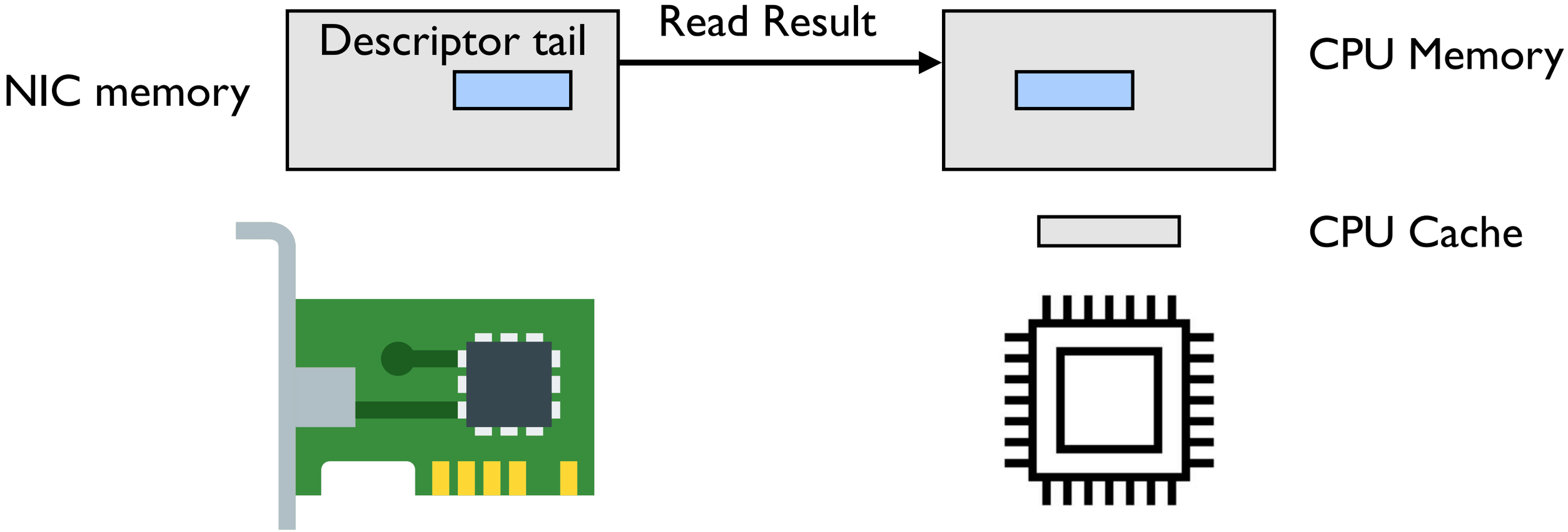
CPU Cache



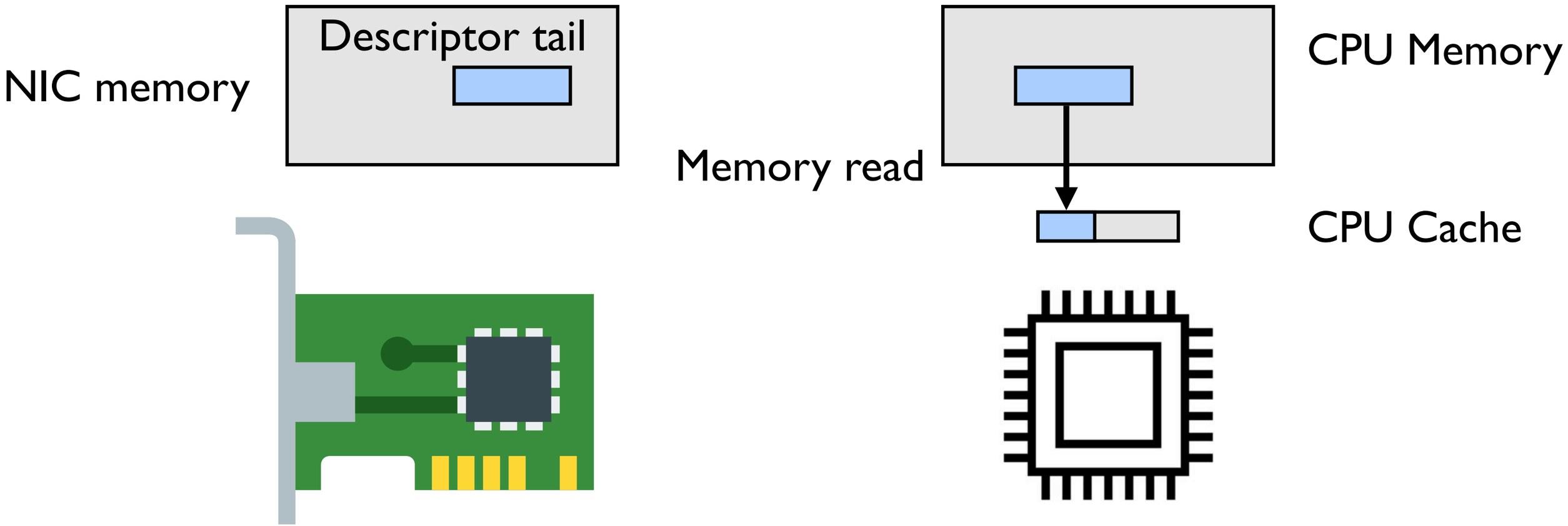
PCIe Example



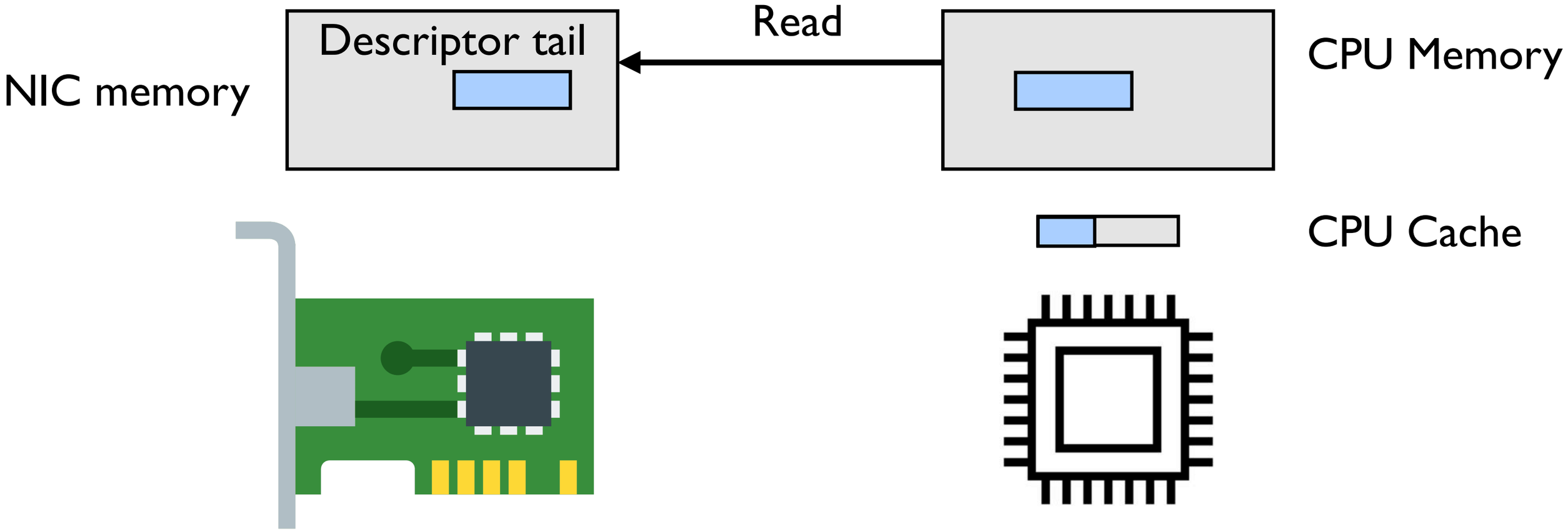
PCIe Example



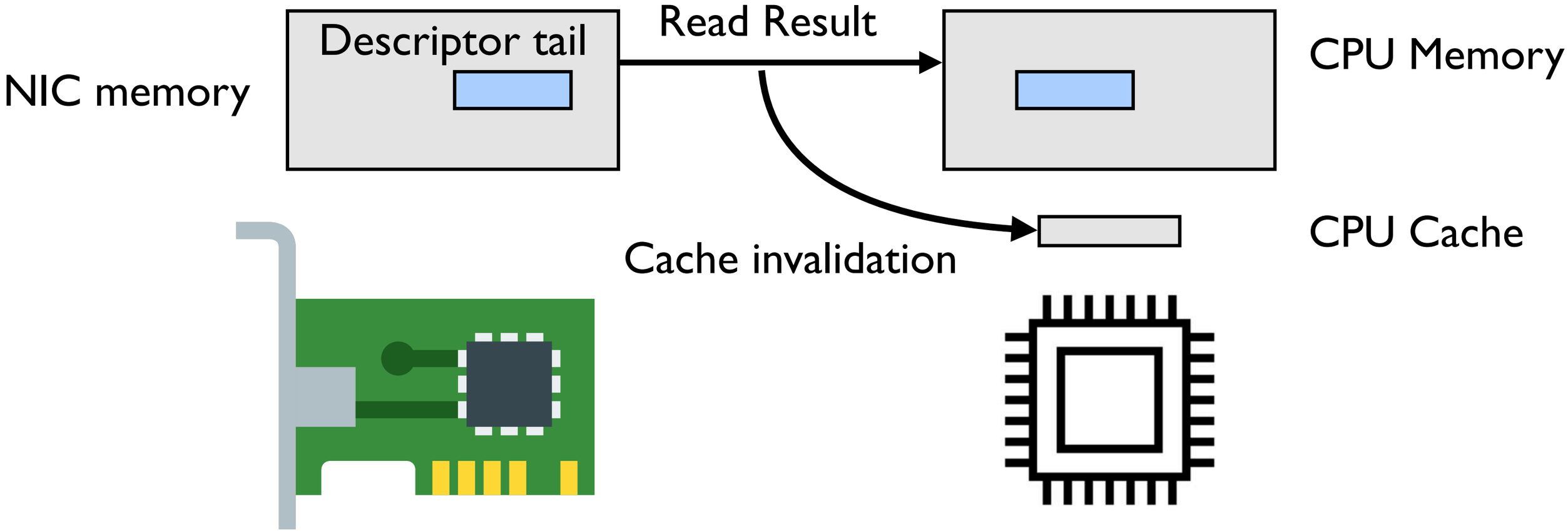
PCIe Example



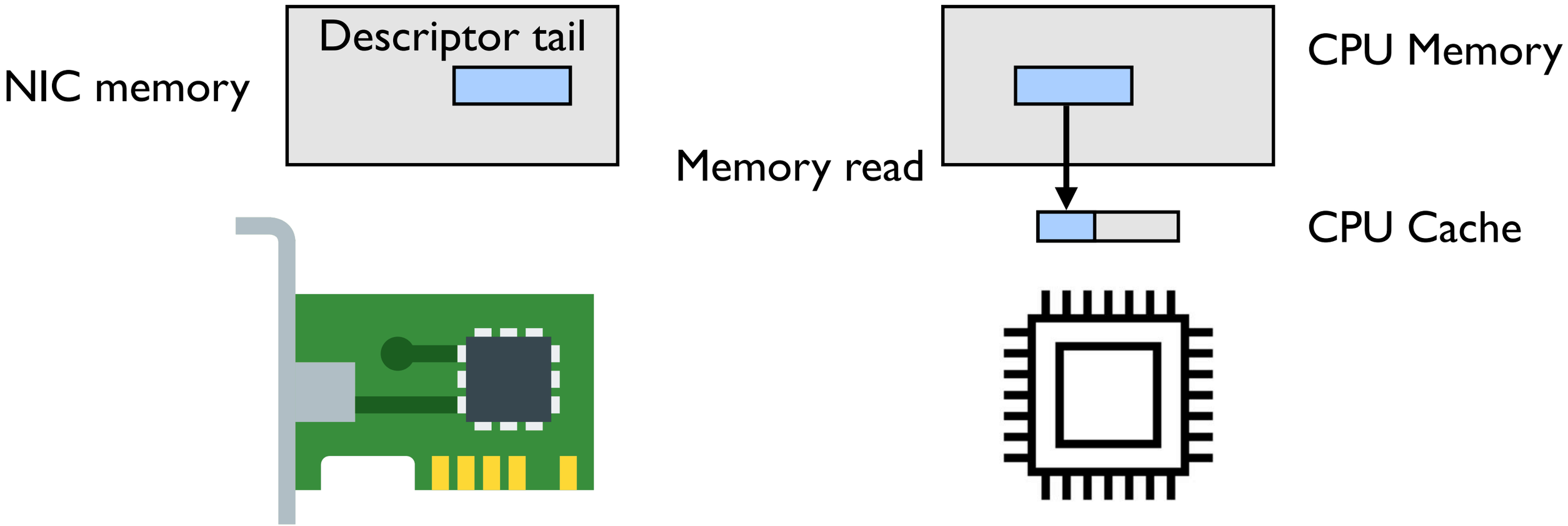
PCIe Example



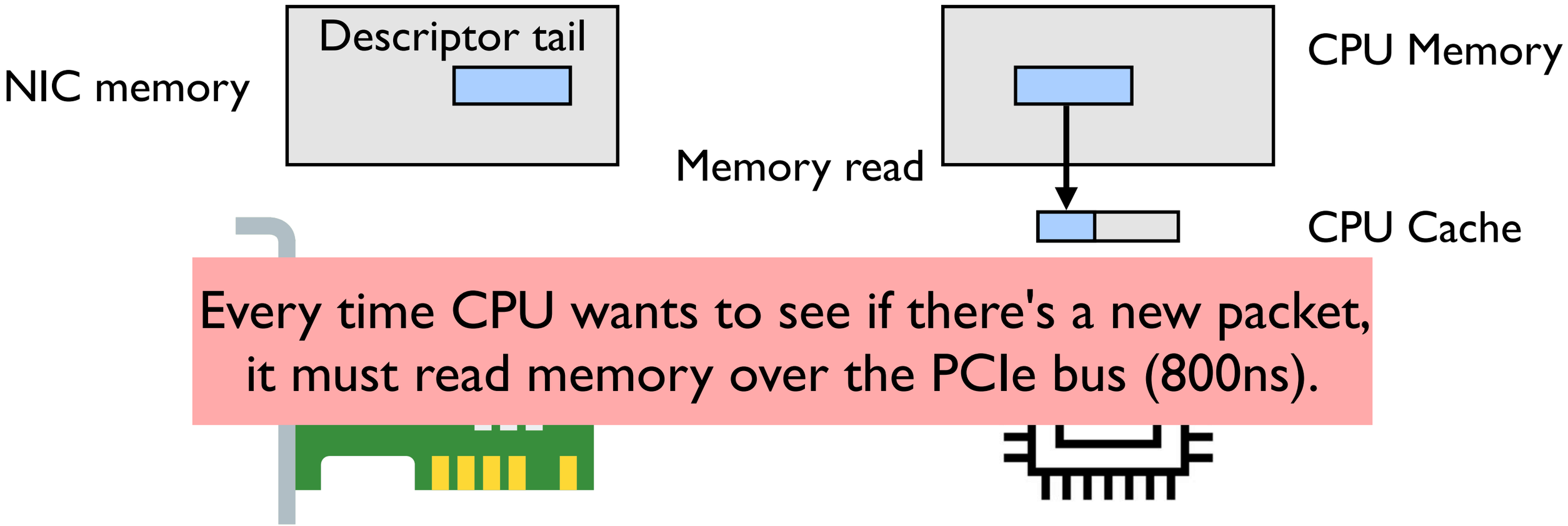
PCIe Example



PCIe Example



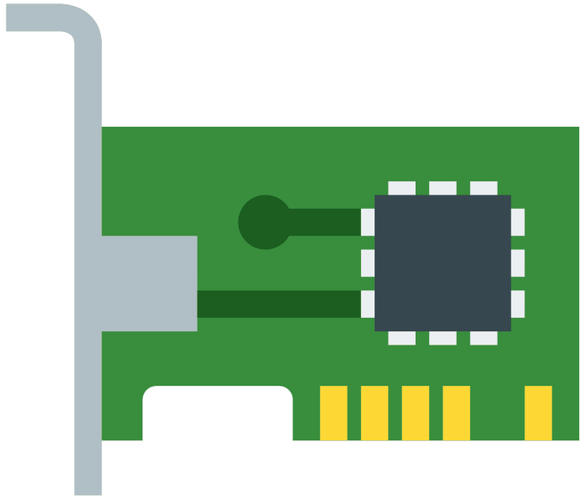
PCIe Example

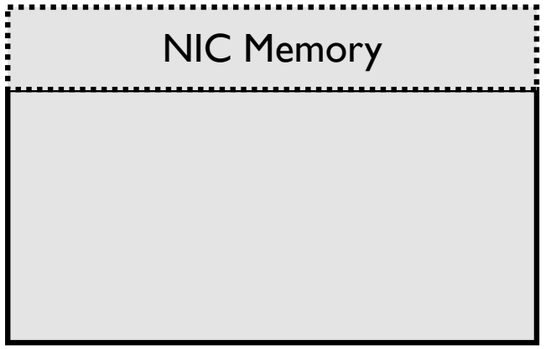


Every time CPU wants to see if there's a new packet, it must read memory over the PCIe bus (800ns).

CXL Example

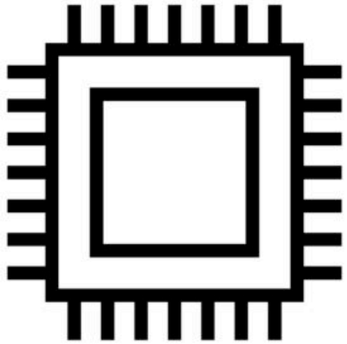
NIC memory



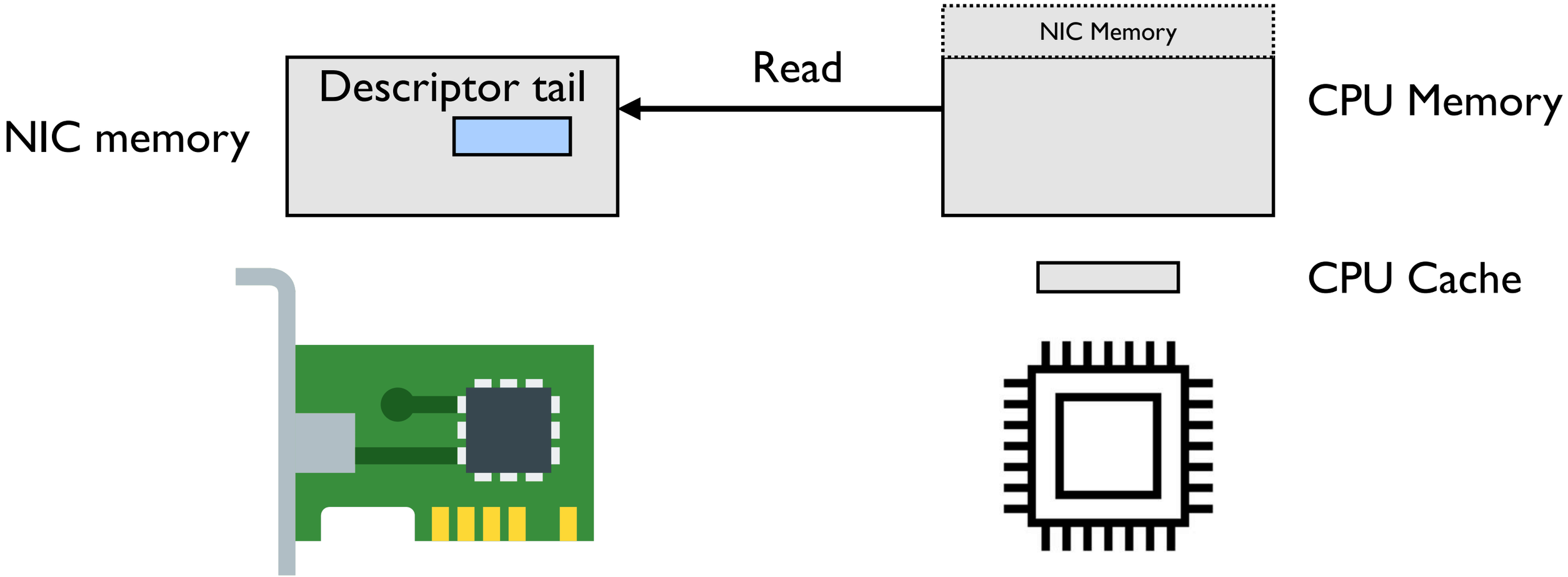


CPU Memory

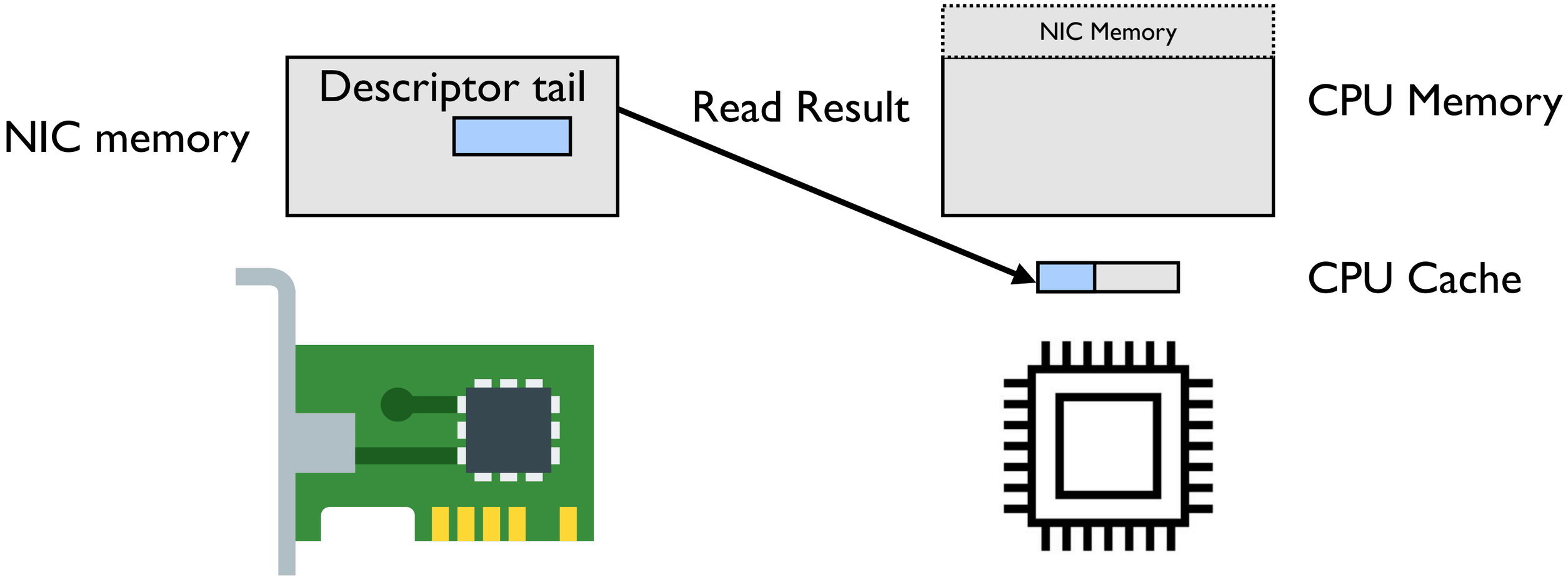
CPU Cache



CXL Example

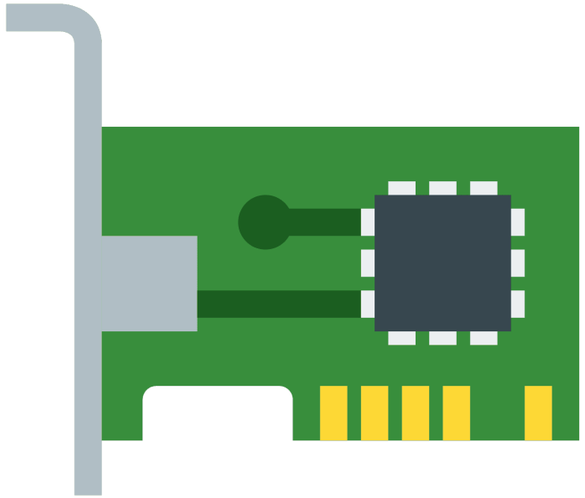

NIC Memory
Read
Descriptor tail
NIC memory
CPU Memory
CPU Cache

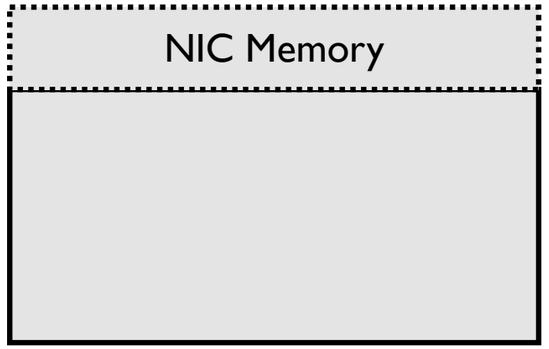
CXL Example


NIC memory
Descriptor tail
Read Result
NIC Memory
CPU Memory
CPU Cache

CXL Example

NIC memory

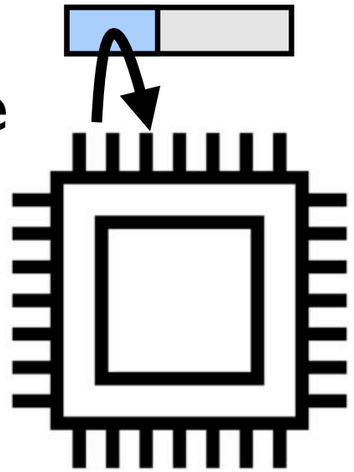




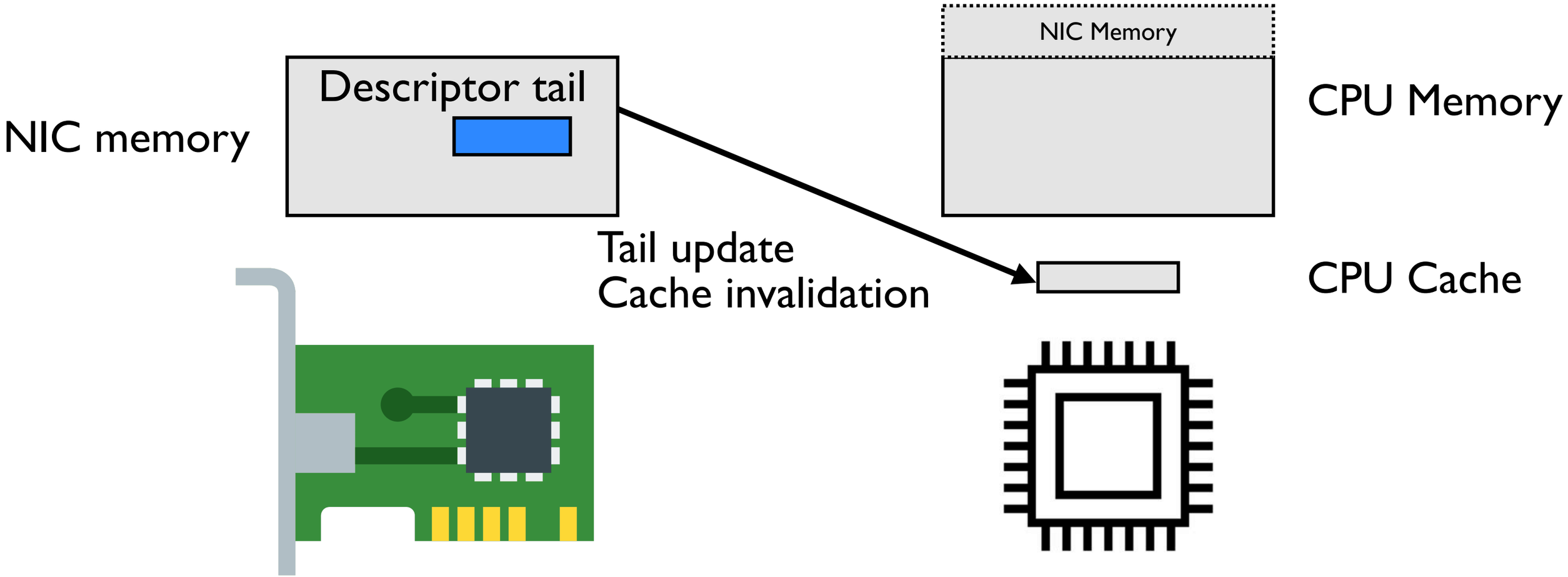
CPU Memory

CPU Cache

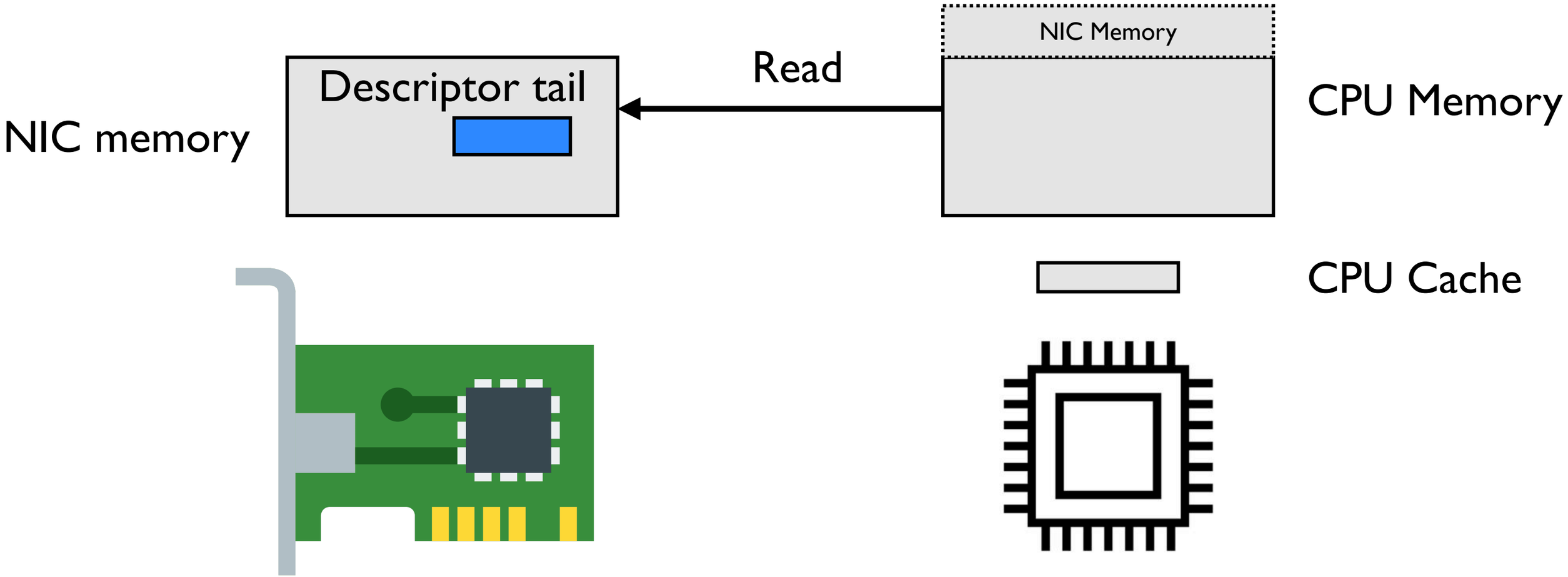
Read From Cache



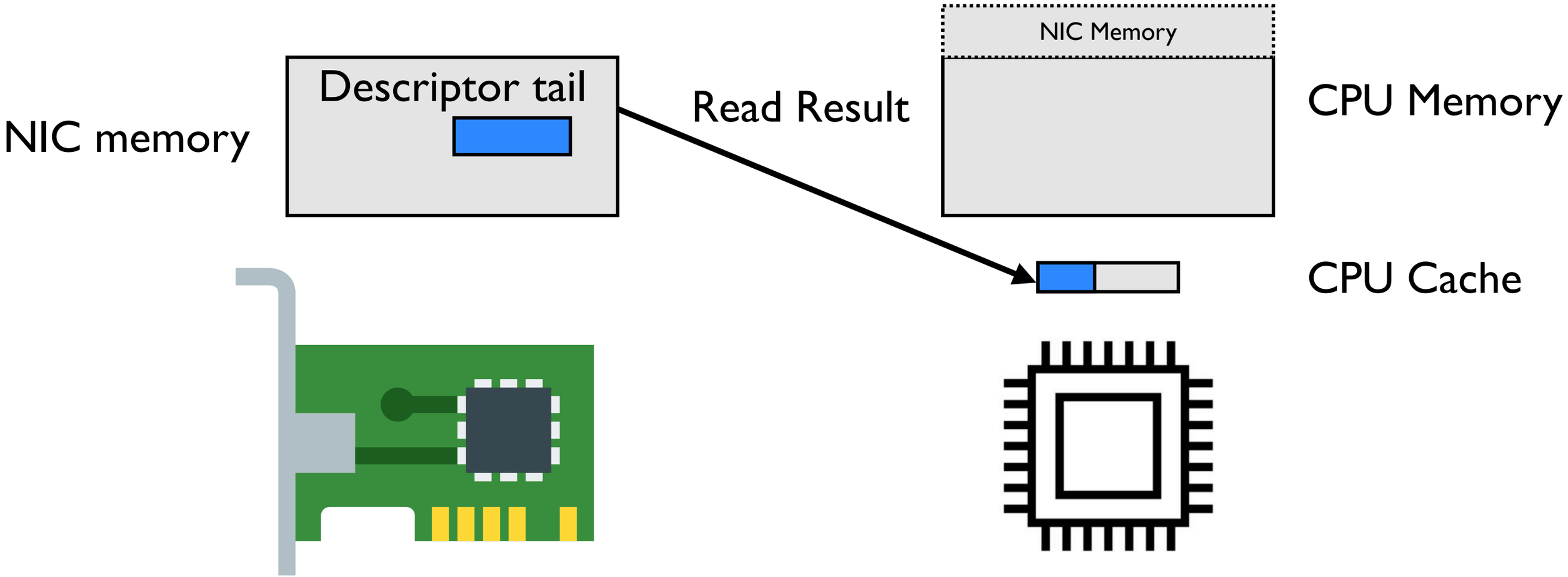
CXL Example


NIC Memory
Descriptor tail
NIC memory
CPU Memory
Tail update
Cache invalidation
CPU Cache

CXL Example


NIC Memory
Read
Descriptor tail
NIC memory
CPU Memory
CPU Cache

CXL Example

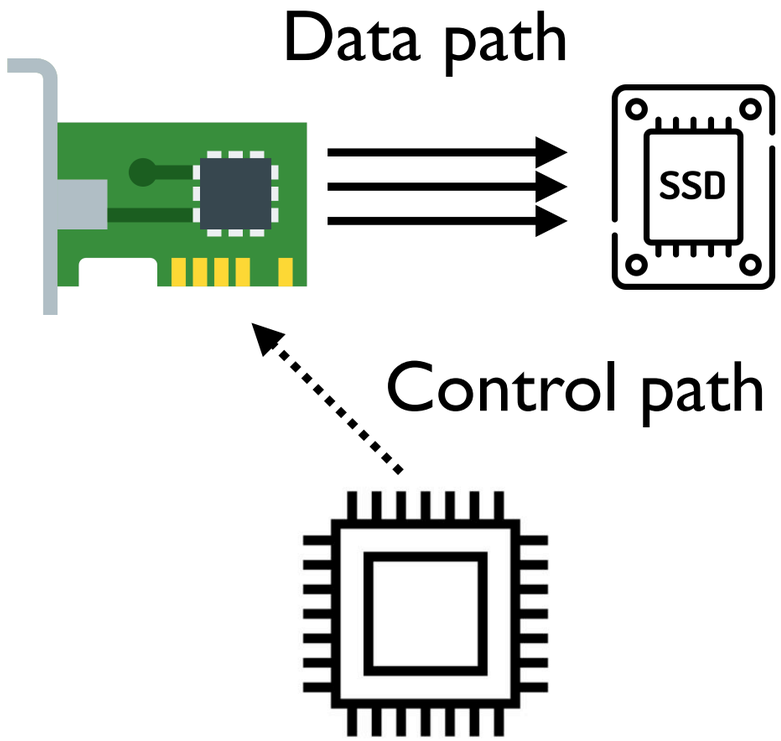

NIC memory
Descriptor tail
Read Result
NIC Memory
CPU Memory
CPU Cache

CXL Memory Devices

- CXL's lower latency means it can support *memory devices*
- Basically, away to get more memory bandwidth by using CXL lanes
  - AMD Genoa: 4.4 Tbps DDR bandwidth, 4.7Tbps PCIe bandwidth
- There's a lot of buzz around this: you can in principle pool memory
  - A single CXL memory device is attached to multiple servers
  - Servers now have an elastic pool of memory, get closer to average use
- I'm skeptical
  - Latency is too high (400ns instead of 100ns)
  - Potential cost savings in RAM are less than the cost of a CXL memory pool device
- But, CXL is definitely going to change how we interact with the NIC

CXL Possibilities

- Cache coherence allows low-cost coordination between devices
  - Polling is very efficient
  - Need to batch updates to pay write cost
- A CPU can easily look into the memory of all of its peripherals
  - Use intrinsics/copies to move data between peripherals
- Move the CPU out of the data path
  - Transfer directly between devices



Computers in 10 Years

- They will look very different
- Three kinds of memory
  - Capacity
  - Bandwidth
  - Latency
- Cache-coherent interconnect (CXL)
  - Allows tighter integration of memory across devices
  - Applications will be higher bandwidth
  - Moving large-scale machine learning models

The Summary

- Processing and network speeds will continue to improve for a while
- RAM price (per bit) won't go down for at least 10 years, probably longer
- RAM performance (latency, throughput) is also flat
- This divergence means computers will be very different in 10 years
- And so will their applications

- It's the end of DRAM as we know it

Thank you!