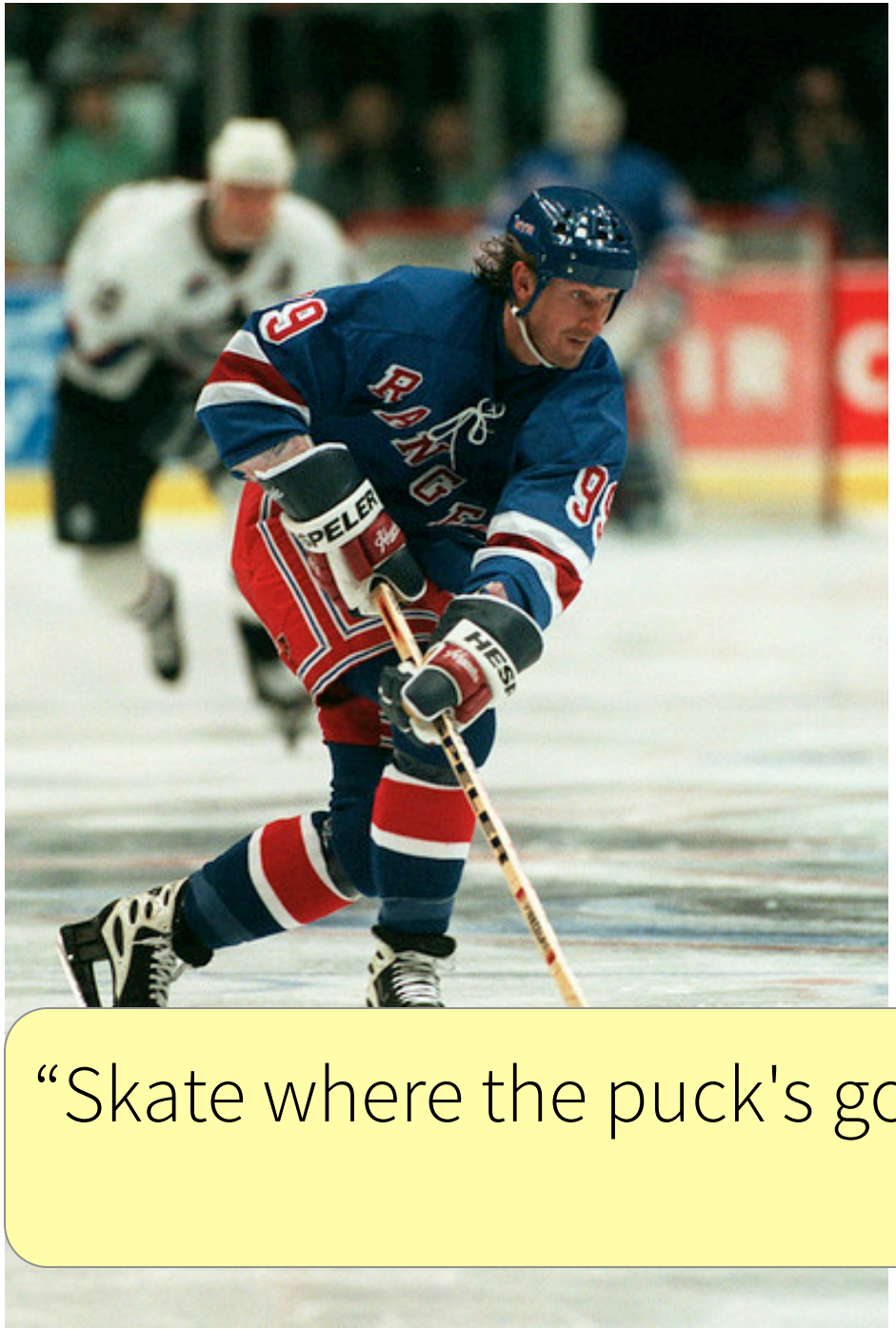
CS 294-110: Technology Trends

August 31, 2015

Ion Stoica and Ali Ghodsi

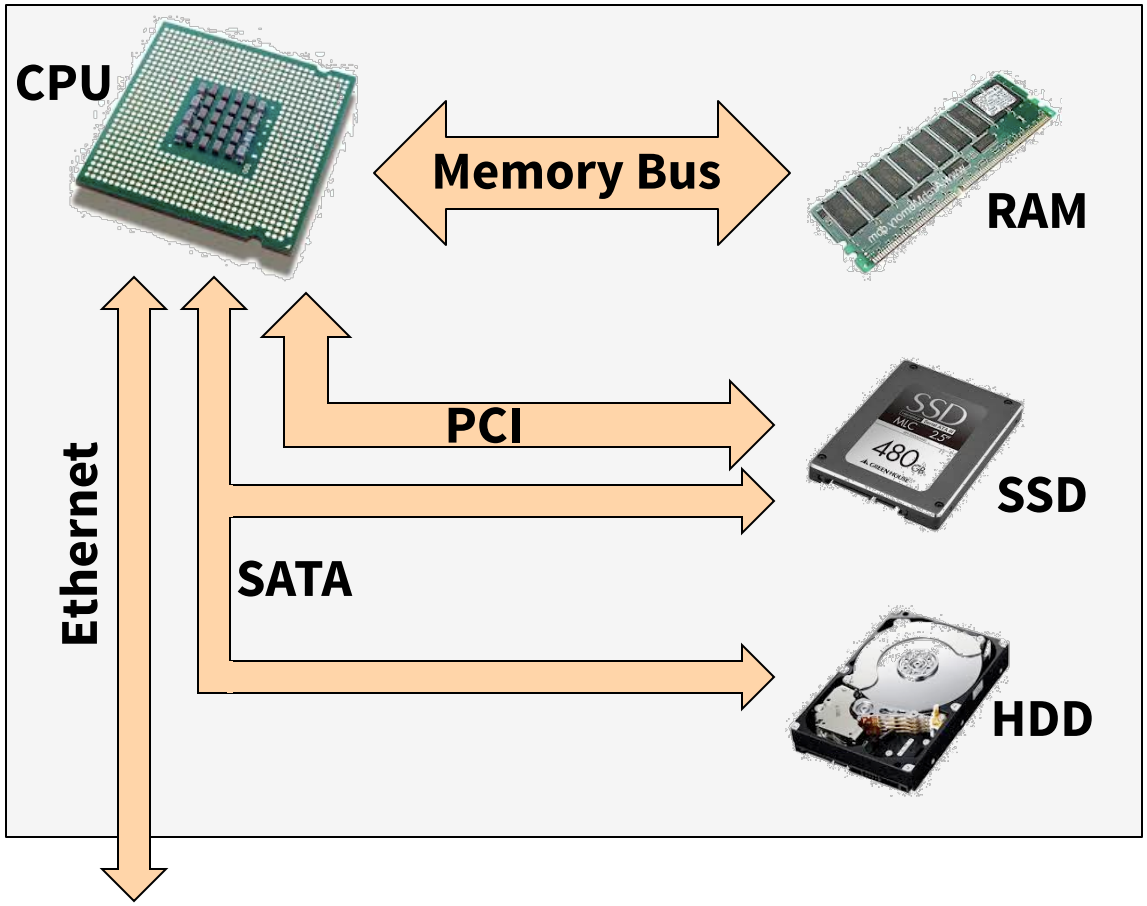
(http://www.cs.berkeley.edu/~istoica/classes/cs294/15/)



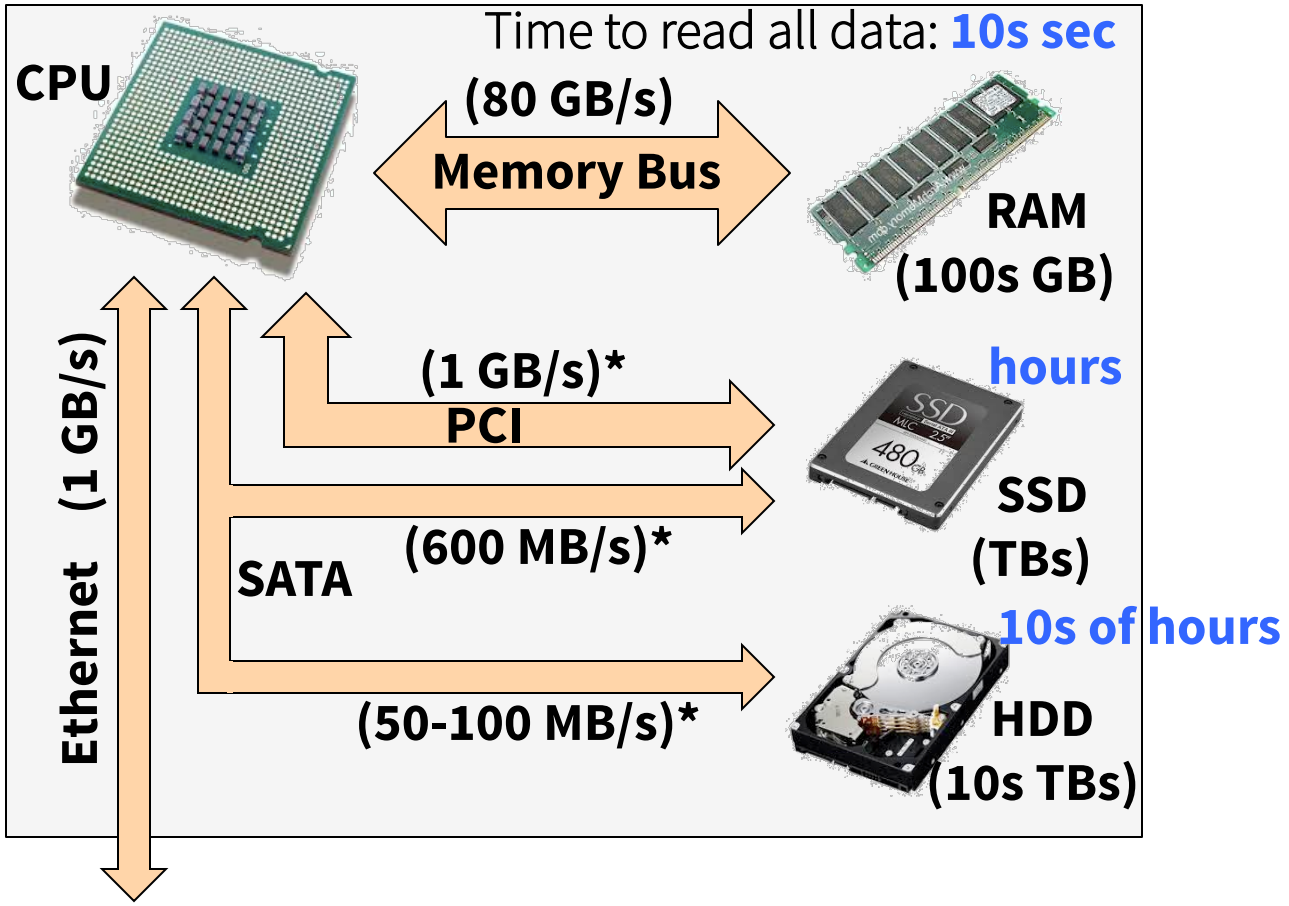
“Skate where the puck's going, not where it's been”

– *Walter Gretzky*

Typical Server Node



Typical Server Node

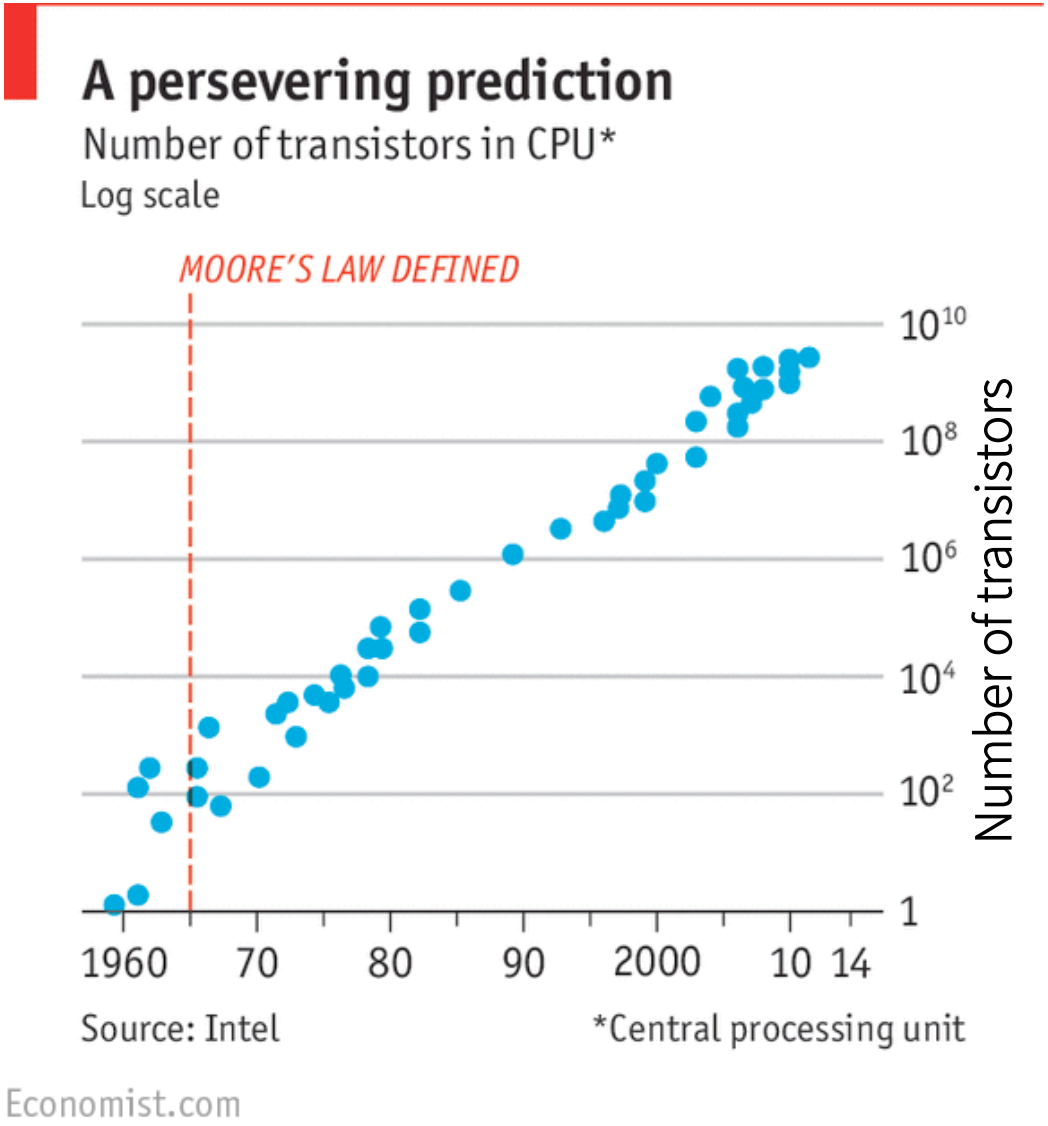


* multiple channels

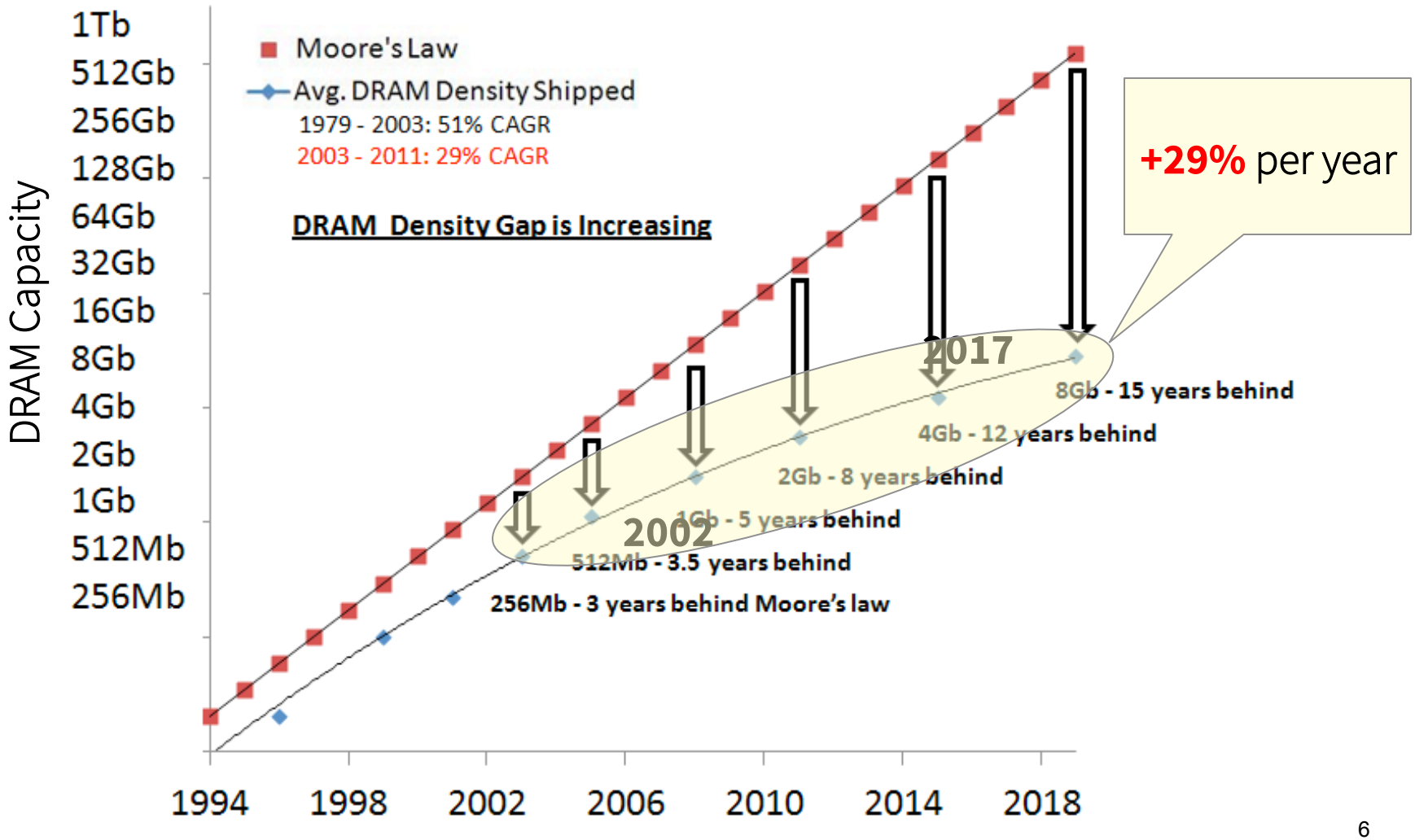
Moore's Law Slowing Down

Stated 50 years ago by Gordon Moore

- Number of transistors on microchip double every 2 years
- Today "closer to 2.5 years" - Brian Krzanich



Memory Capacity



Memory Price/Byte Evolution

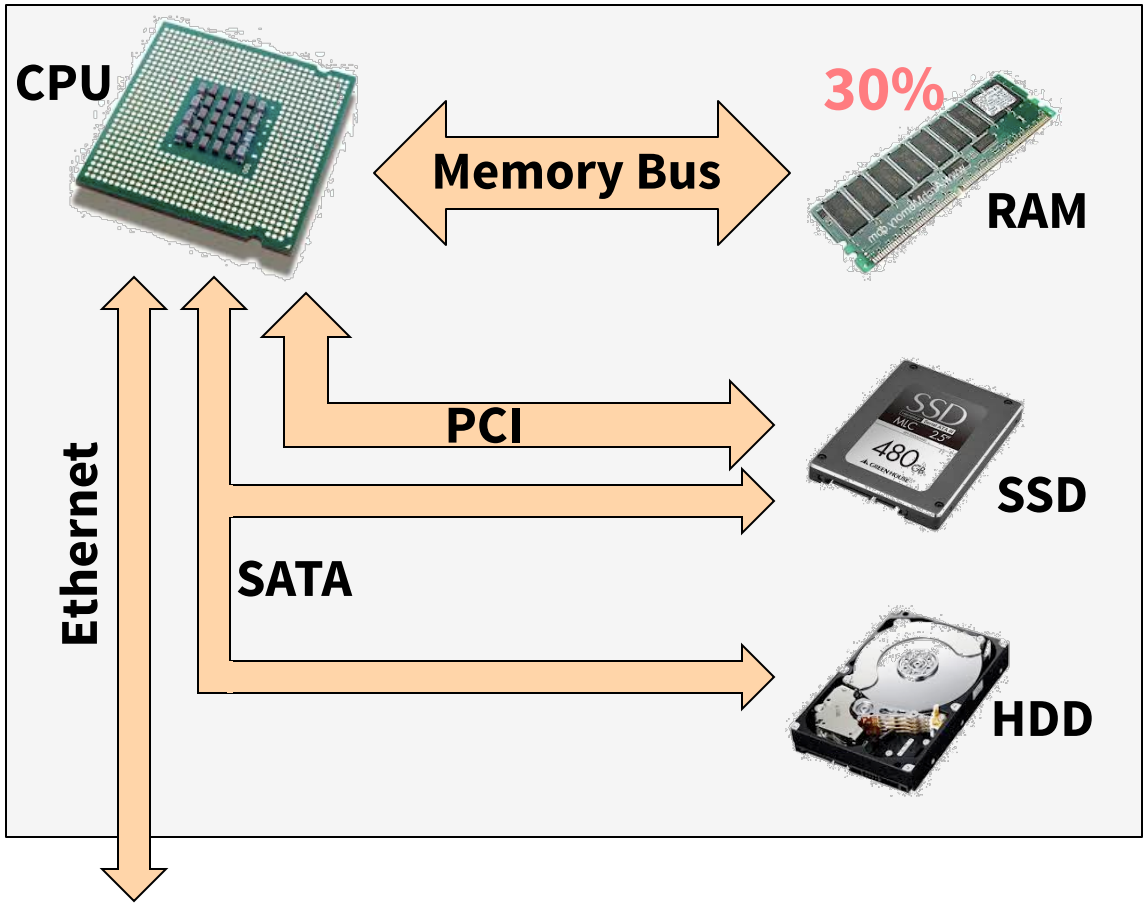
1990-2000: -54% per year

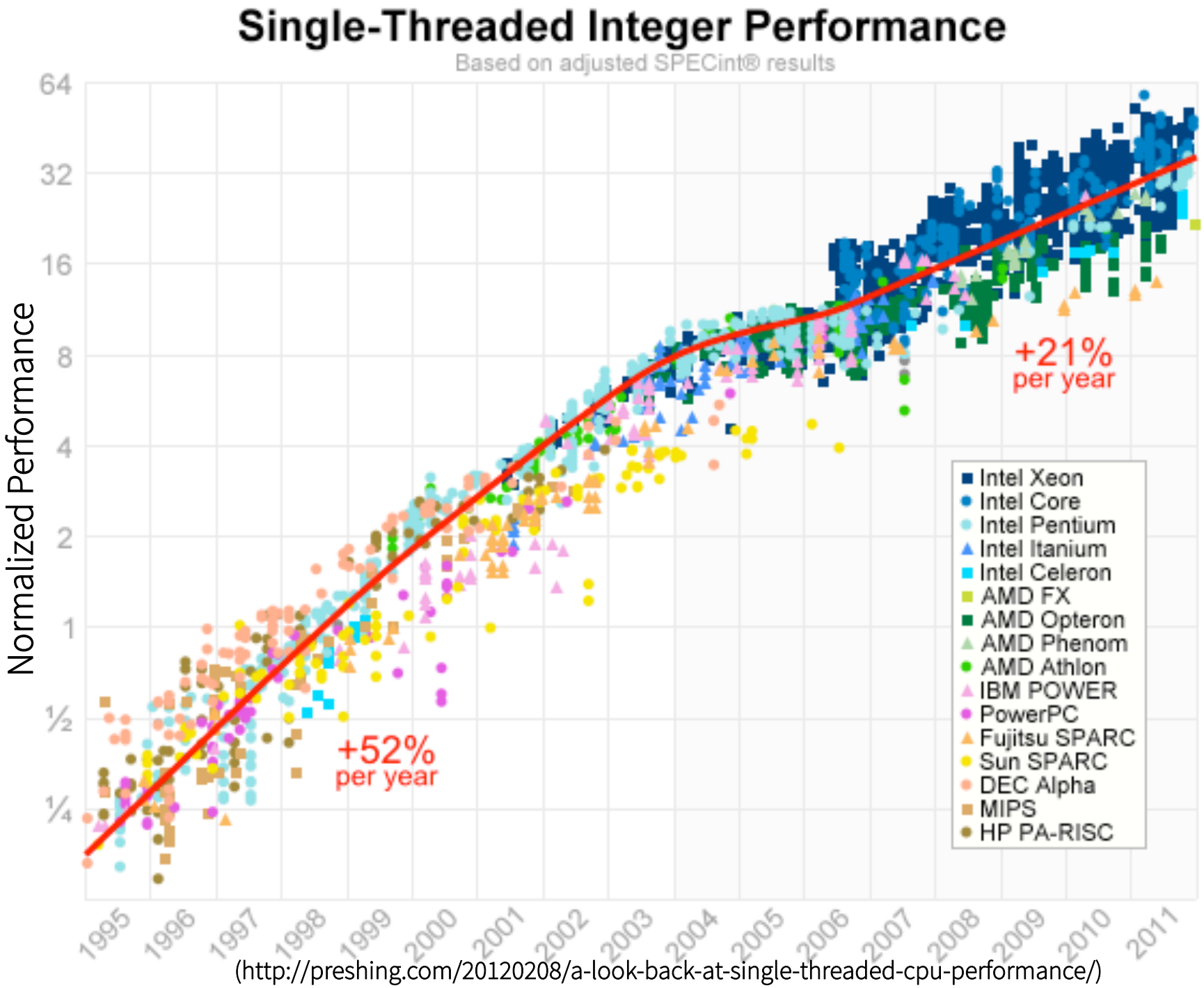
2000-2010: -51% per year

2010-2015: -32% per year

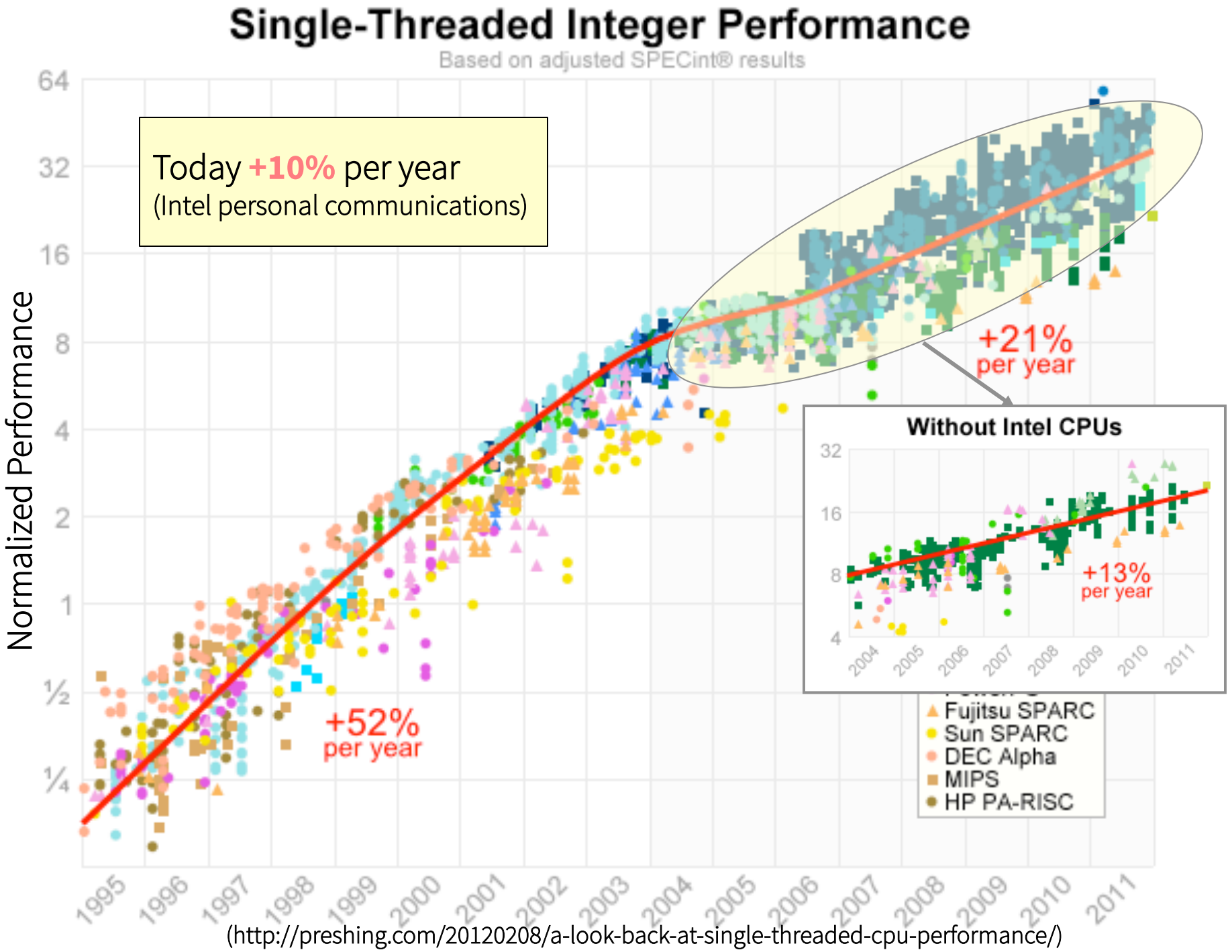
(http://www.jcmit.com/memoryprice.htm)

Typical Server Node



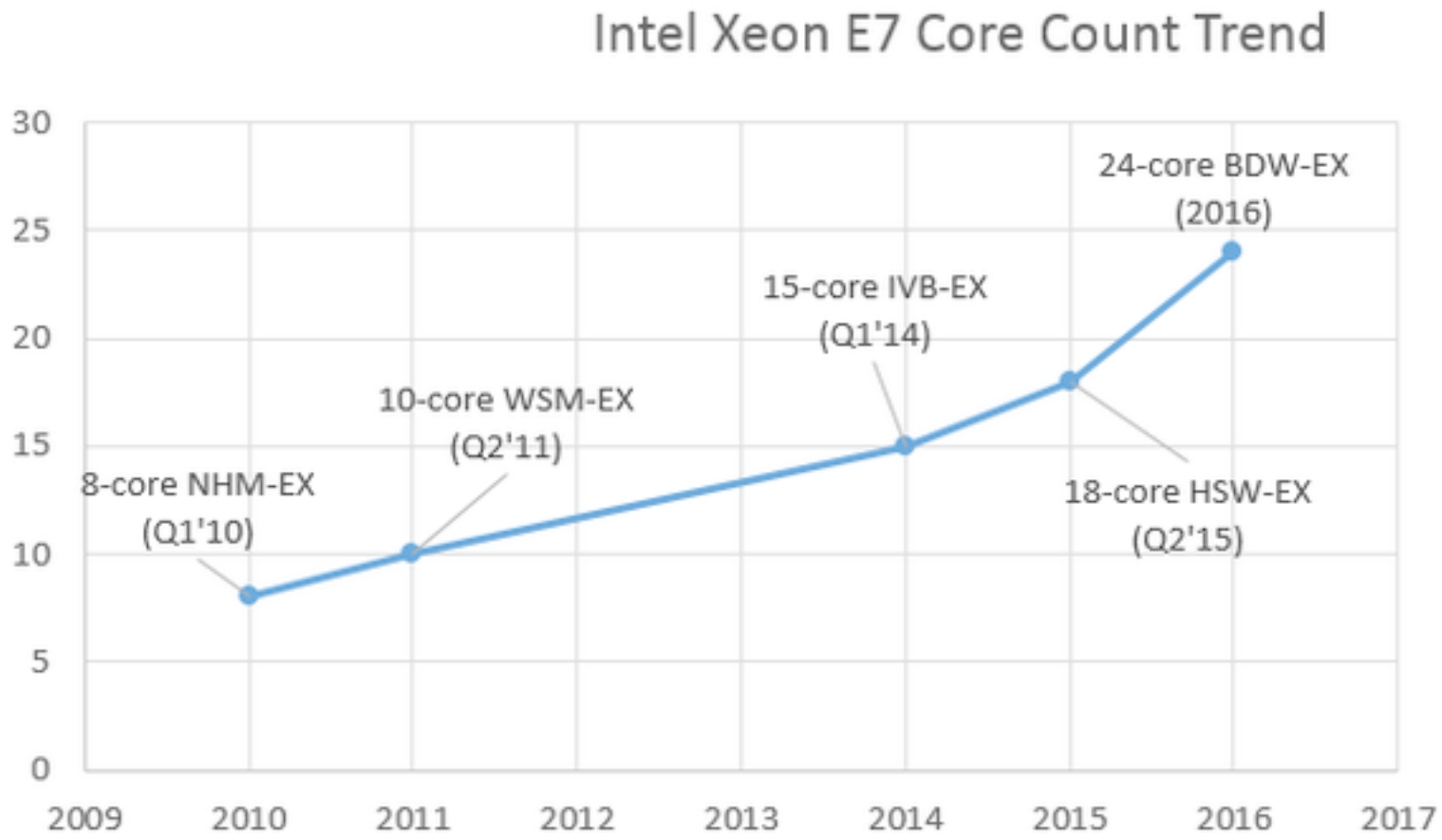


(http://preshing.com/20120208/a-look-back-at-single-threaded-cpu-performance/)



(http://preshing.com/20120208/a-look-back-at-single-threaded-cpu-performance/)

Number of cores: +18-20% per year



(Source: http://www.fool.com/investing/general/2015/06/22/1-huge-innovation-intel-corp-could-bring-to-future.aspx)

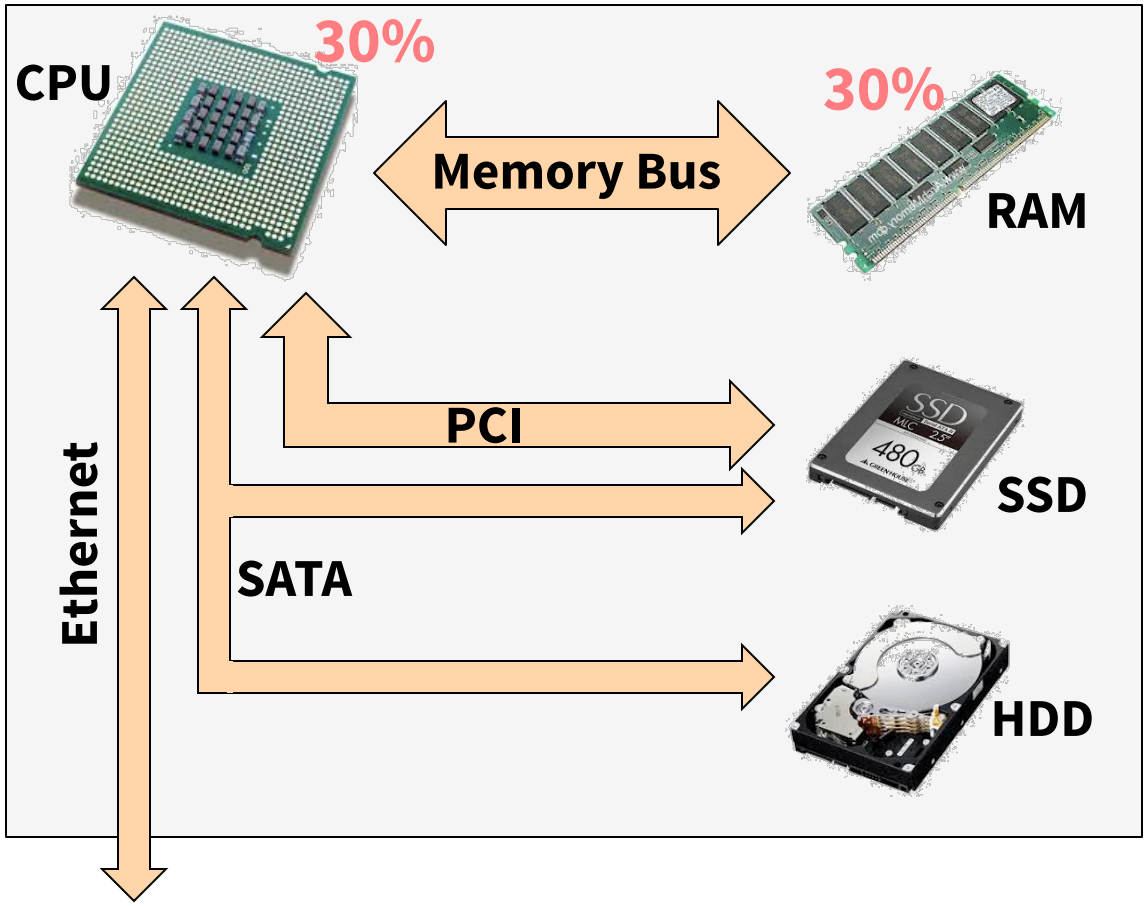
CPU Performance Improvement

Number of cores: +18-20%

Per core performance: +10%

Aggregate improvement: +30-32%

Typical Server Node



SSDs

Performance:

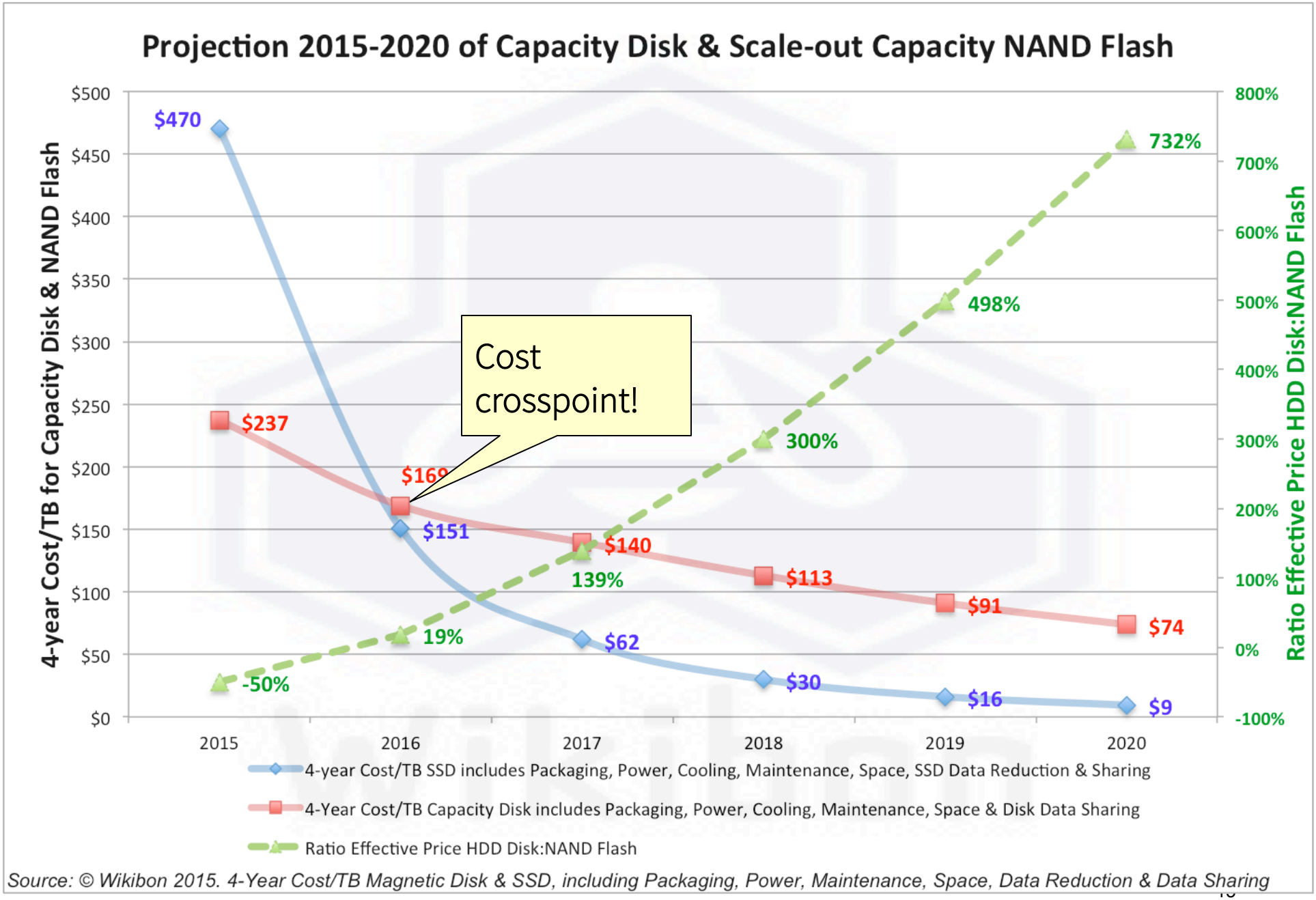
- Reads: 25us latency
- Write: 200us latency
- Erase: 1,5 ms

Steady state, when SSD full

- One erase every 64 or 128 reads (depending on page size)

Lifetime: 100,000-1 million writes per page

Rule of thumb: writes 10x more expensive than reads, and erases 10x more expensive than writes



Source: © Wikibon 2015. 4-Year Cost/TB Magnetic Disk & SSD, including Packaging, Power, Maintenance, Space, Data Reduction & Data Sharing

SSDs vs. HDDs

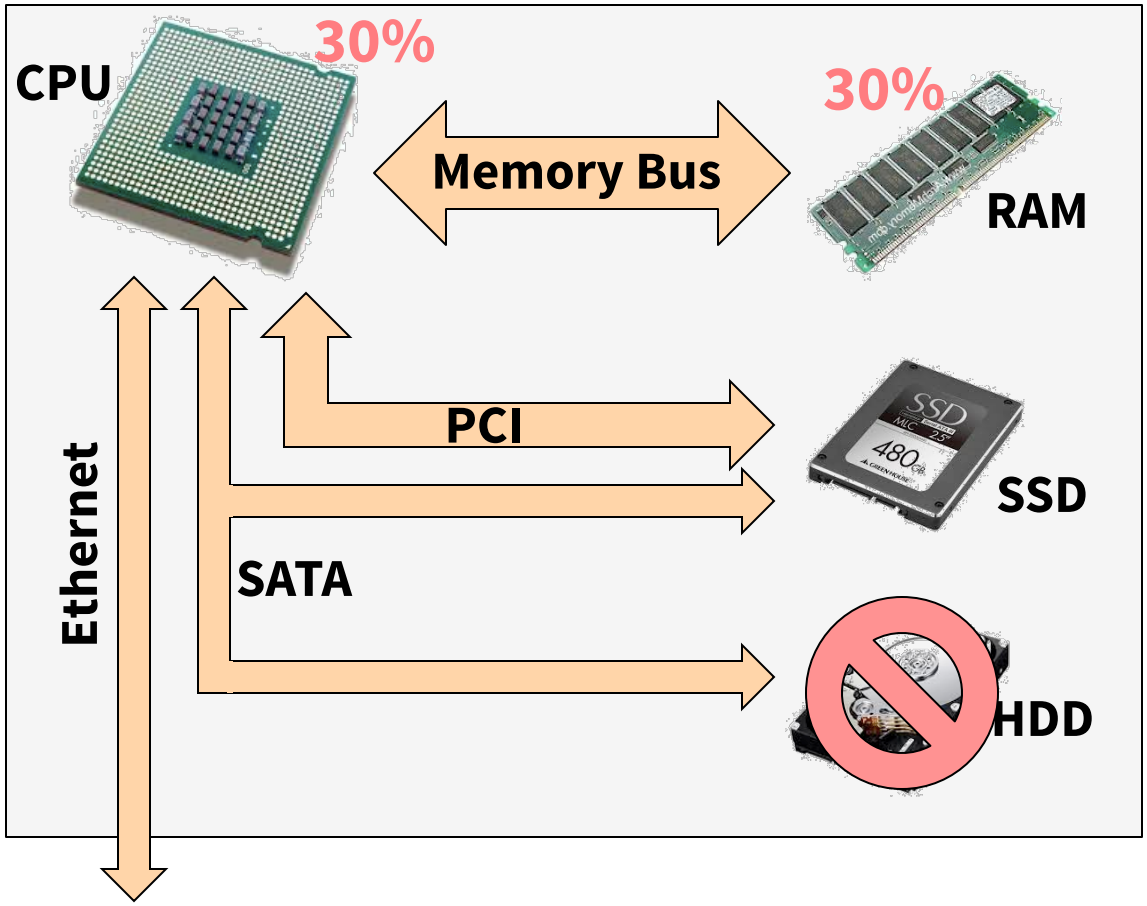
SSDs will soon become cheaper than HDDs

Transition from HDDs to SSDs will accelerate

- Already most instances in AWS have SSDs
- Digital Ocean instances are SSD only

Going forward we can assume SSD only clusters

Typical Server Node

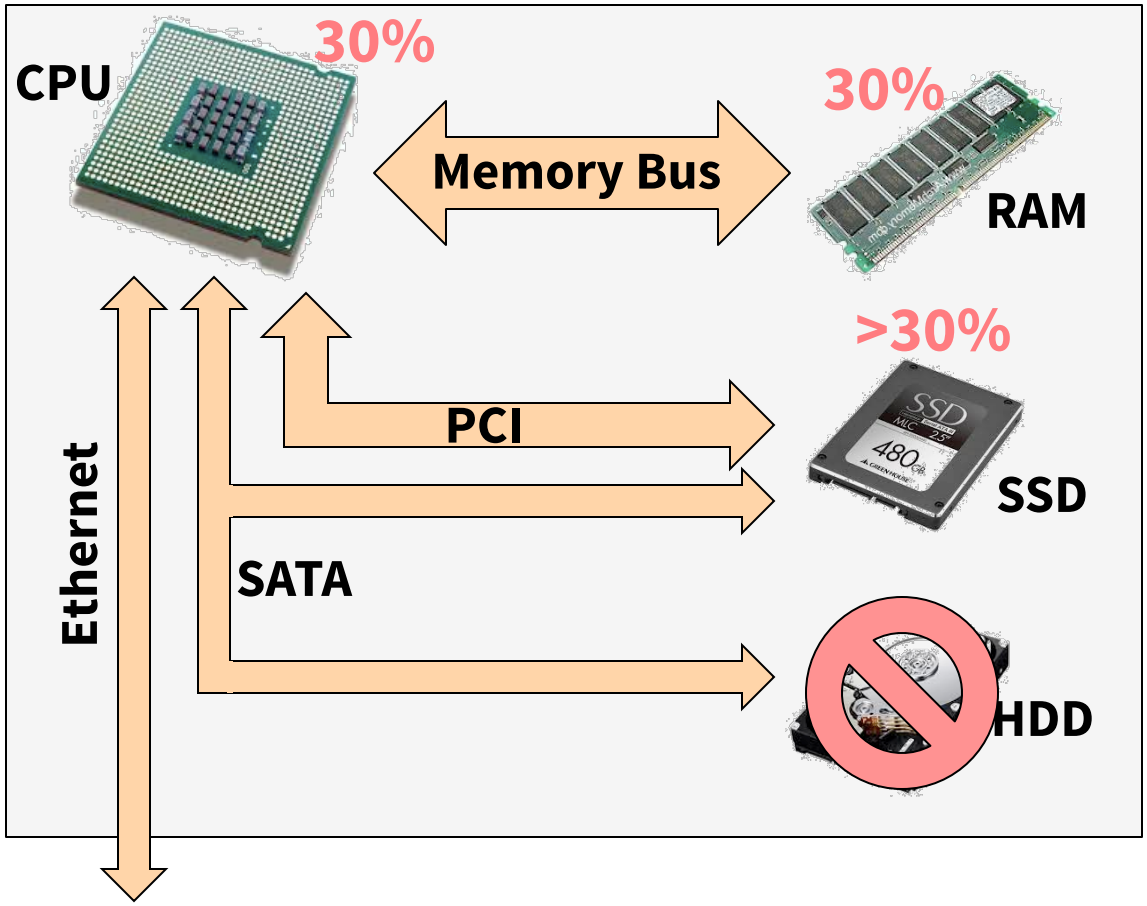


SSD Capacity

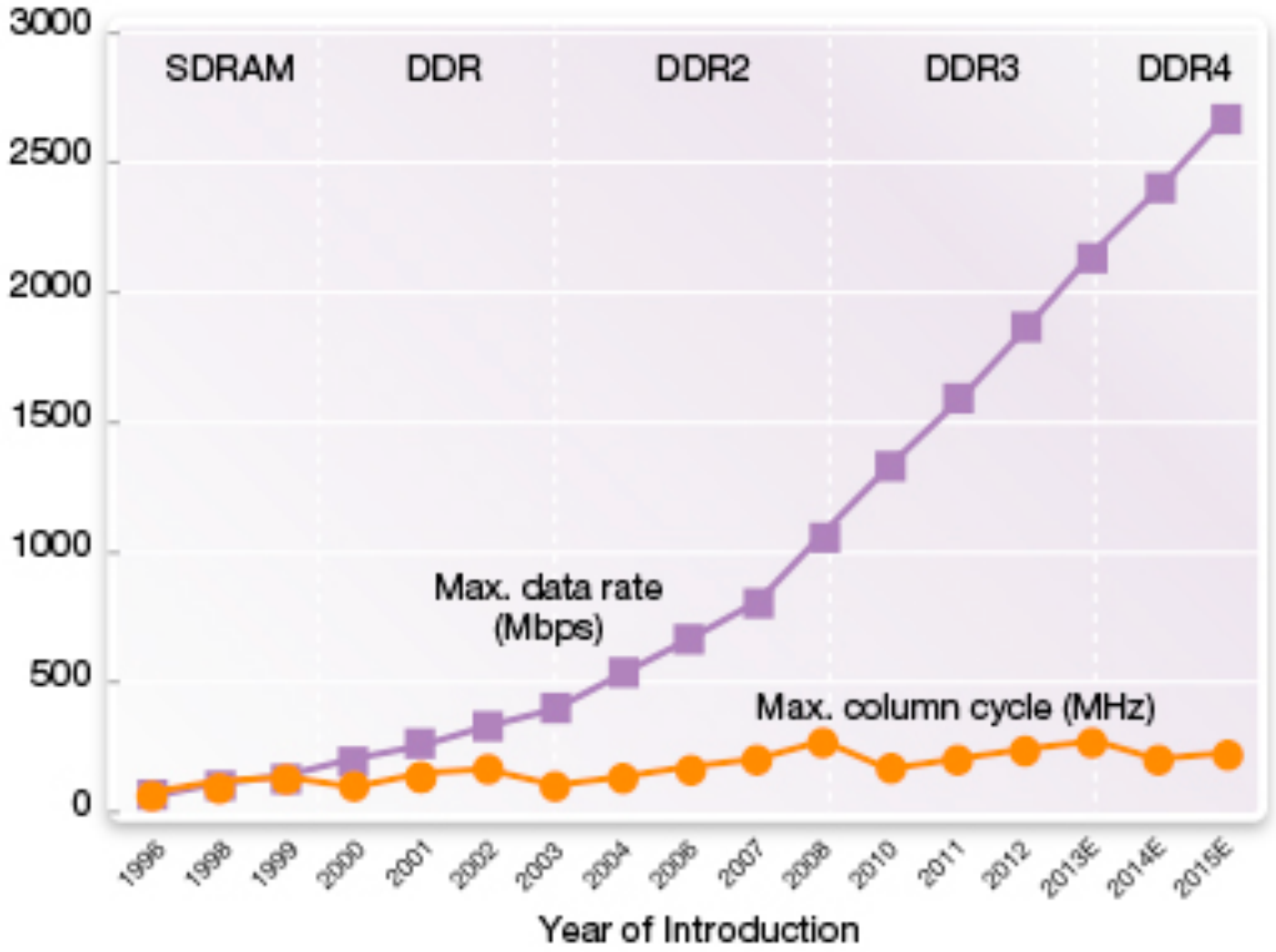
Leverage Moore's law

3D technologies will help outpace Moore's law

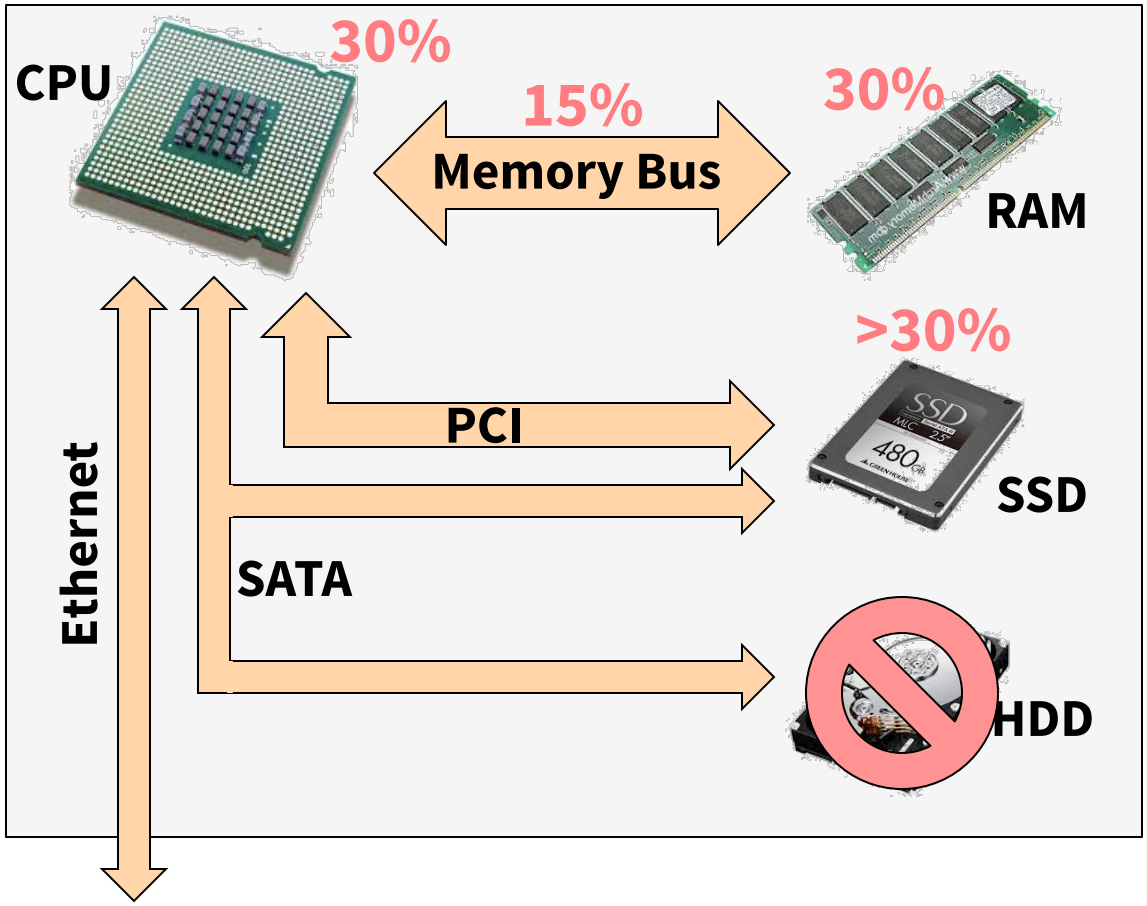
Typical Server Node



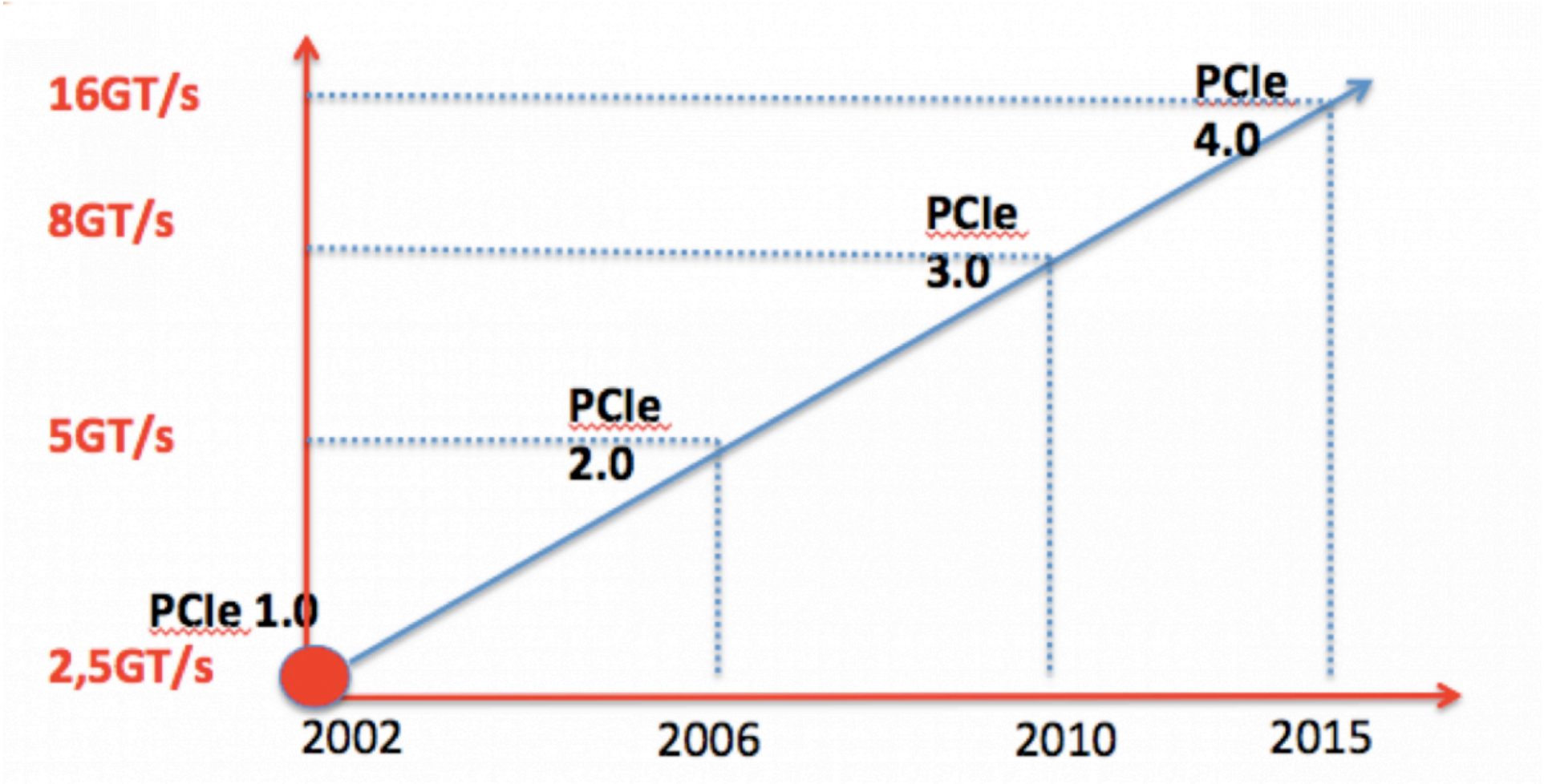
Memory Bus: +15% per year



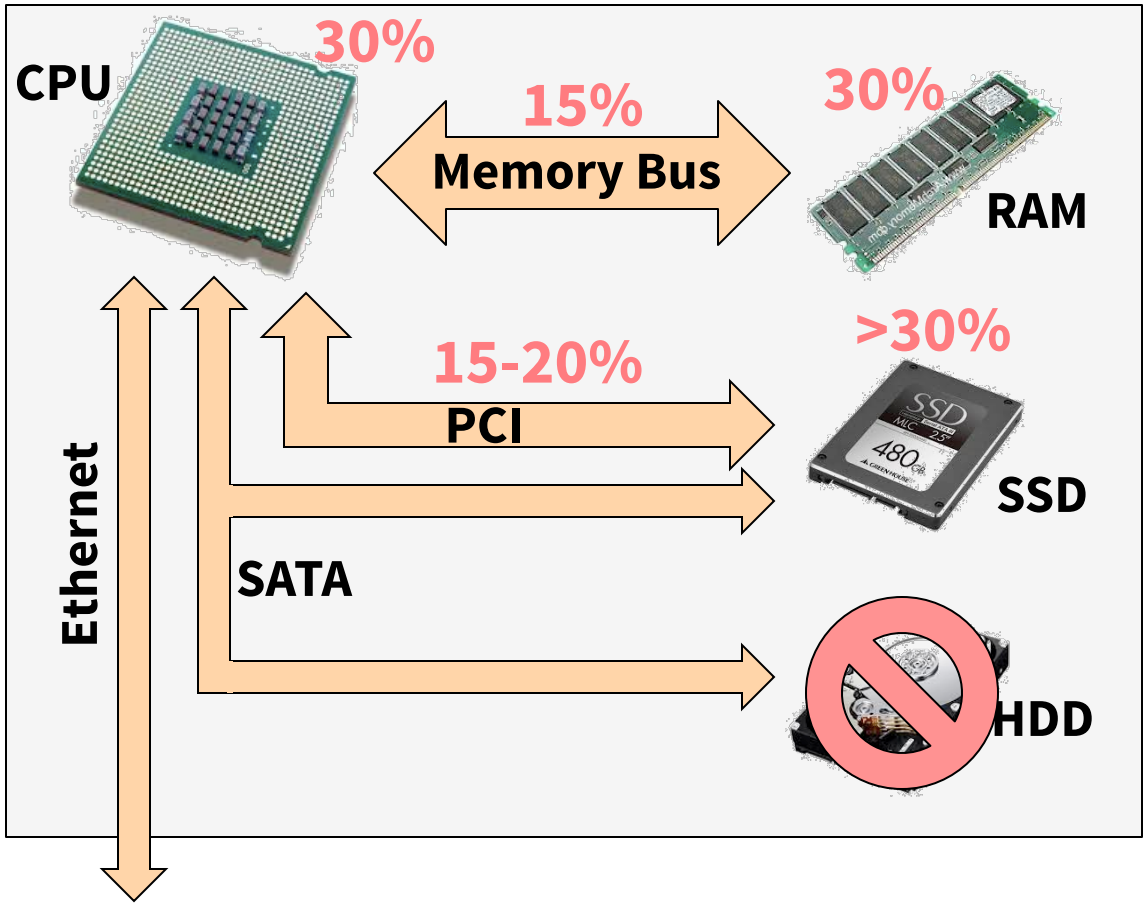
Typical Server Node



PCI Bandwidth: 15-20% per Year



Typical Server Node



SATA

2003: 1.5Gbps (SATA 1)

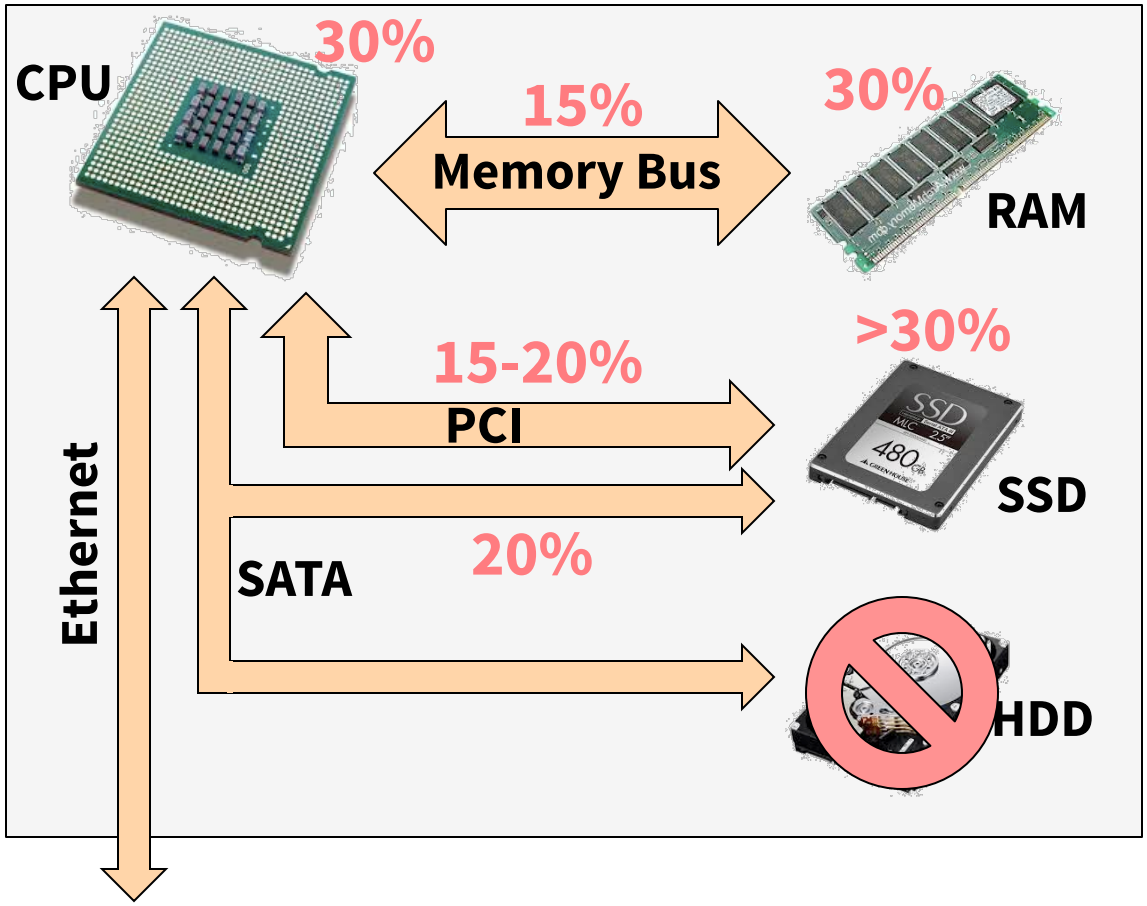
2004: 3Gbps (SATA 2)

2008: 6Gbps (SATA 3)

2013: 16Gbps (SATA 3.2)

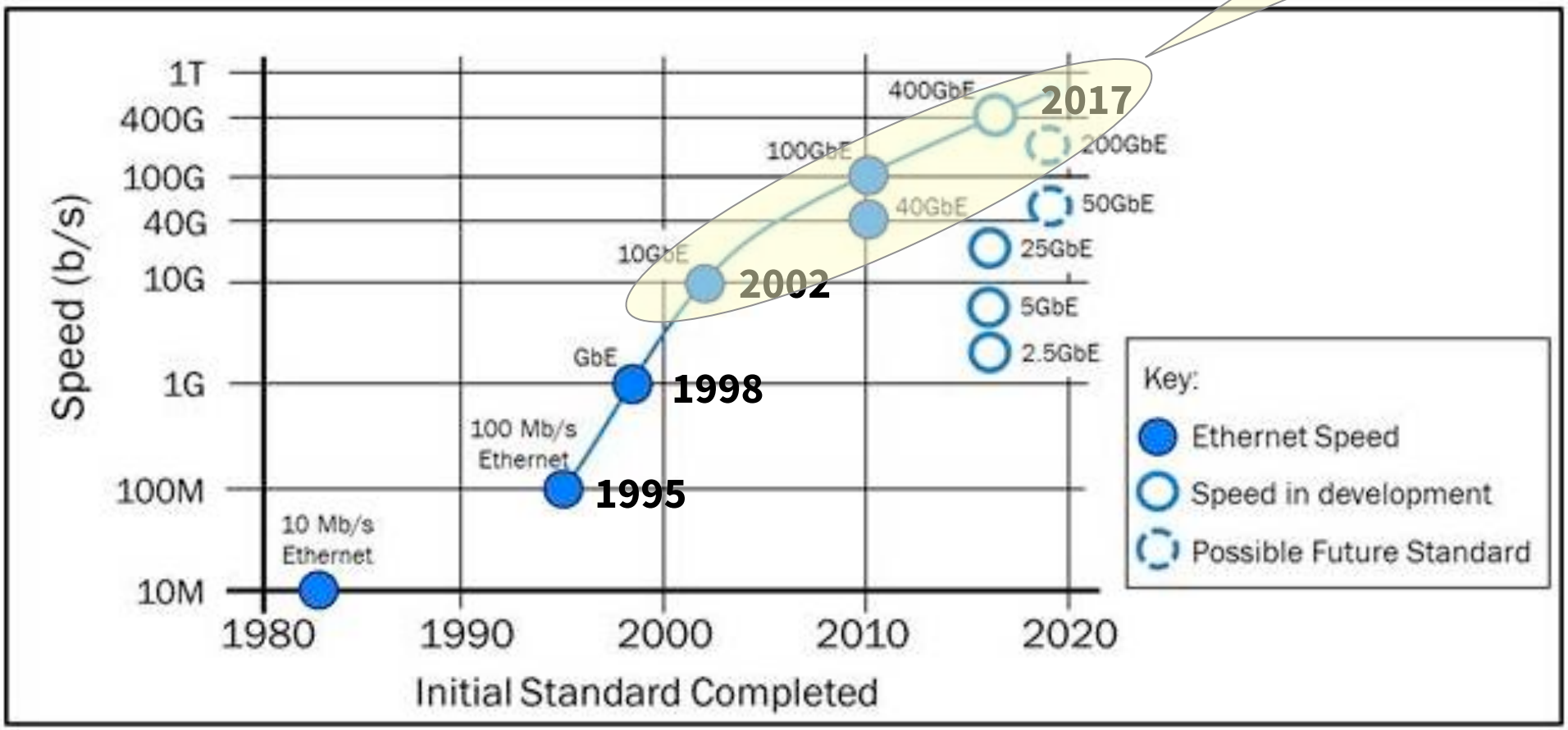
+20% per year since 2004

Typical Server Node

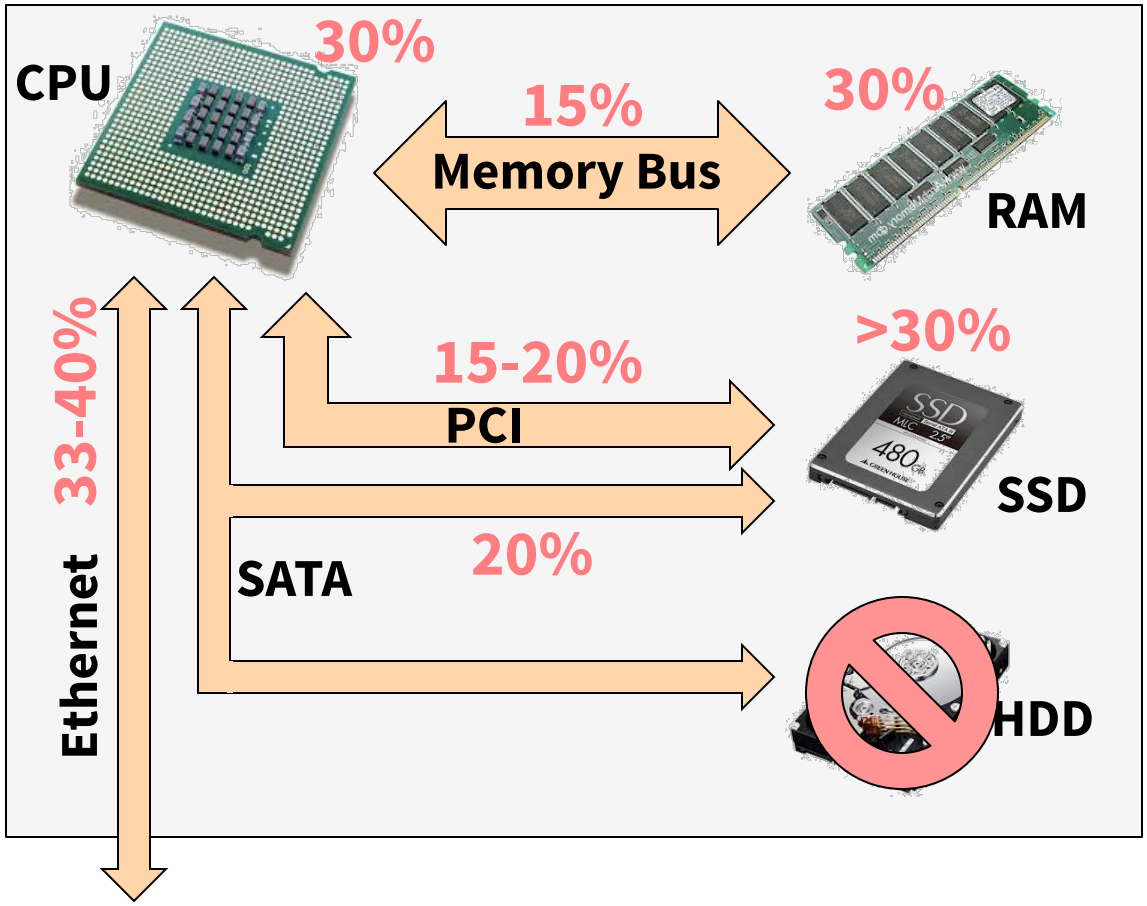


Ethernet Bandwidth

33-40% per year



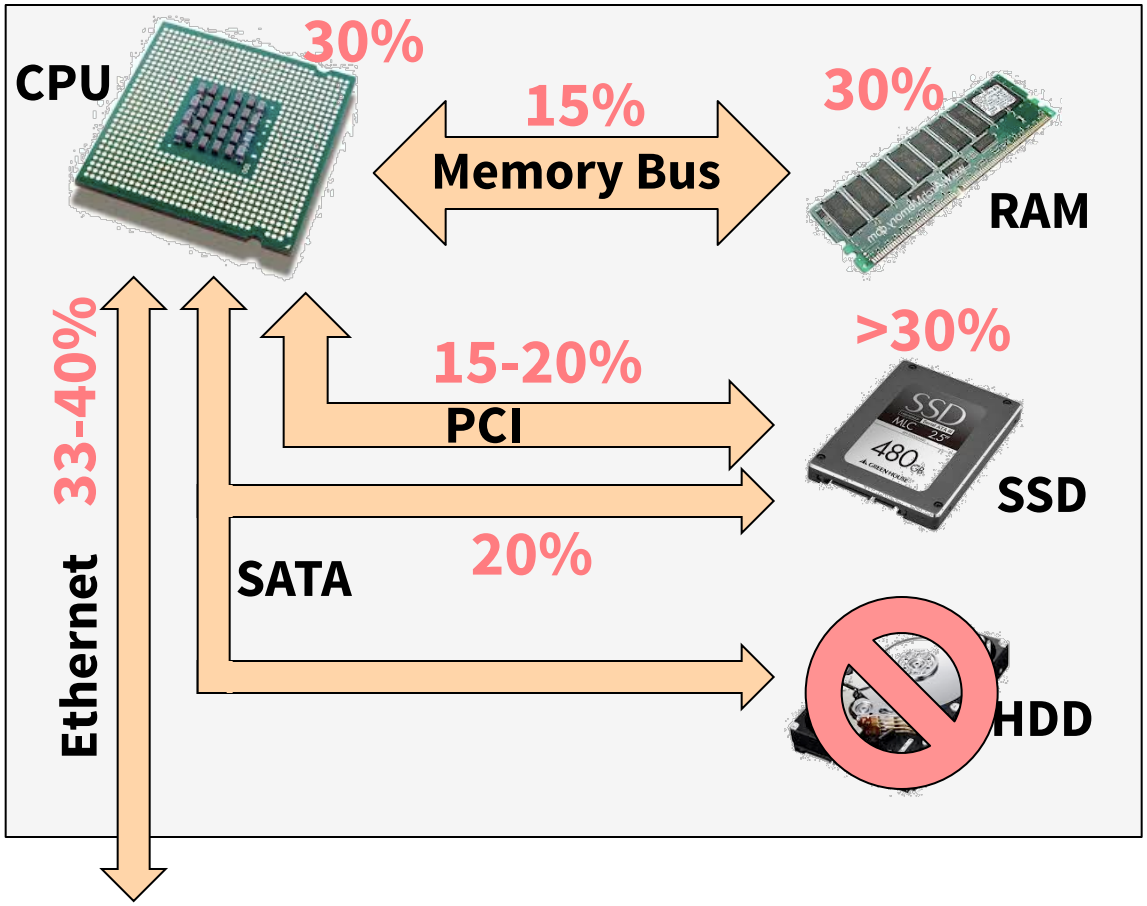
Typical Server Node



Summary so Far

Bandwidth to storage is the bottleneck

Will take longer and longer and longer to read entire data from RAM or SSD



But wait, there is more…

3D XPoint Technology

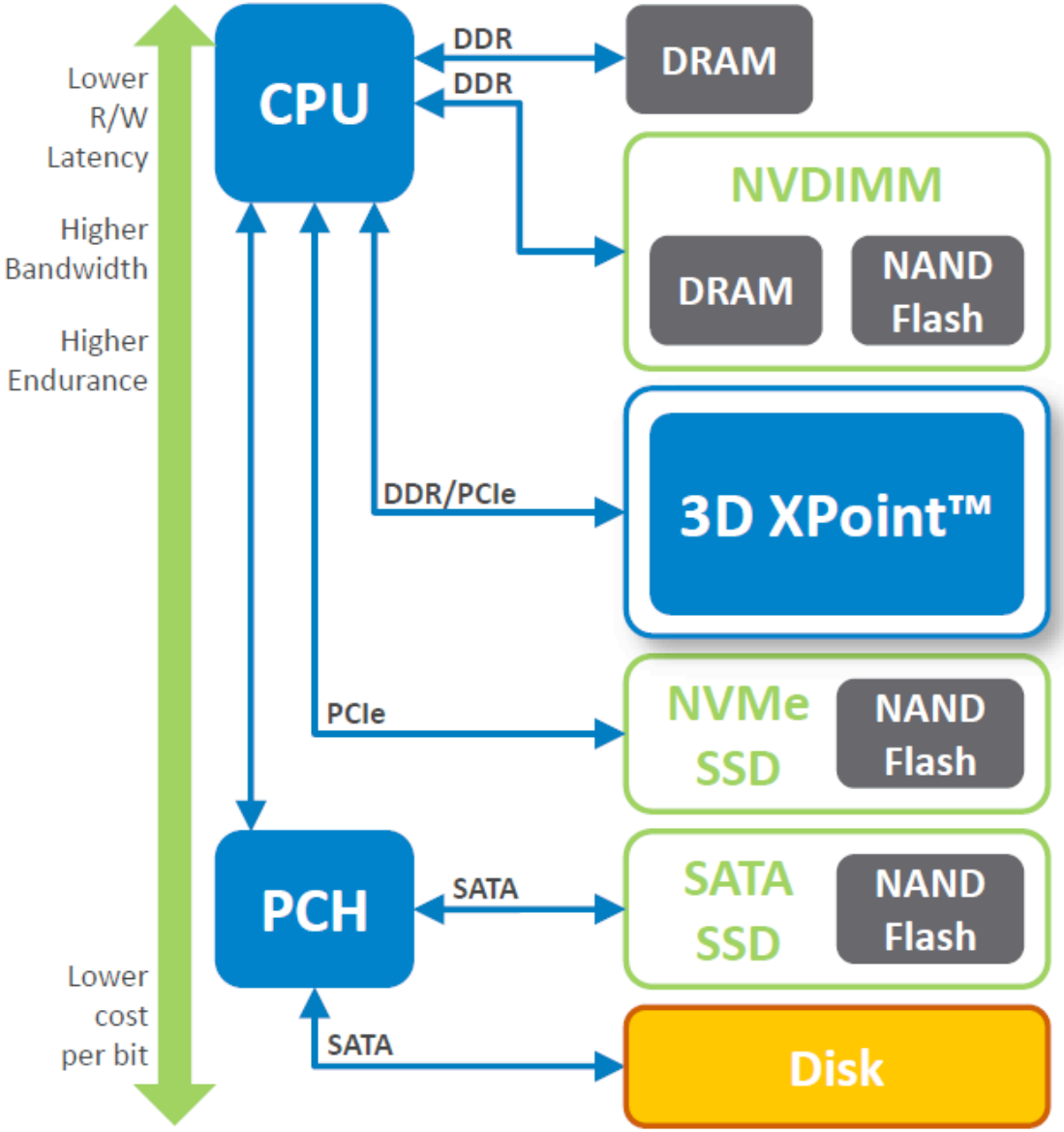
Developed by Intel and Micron

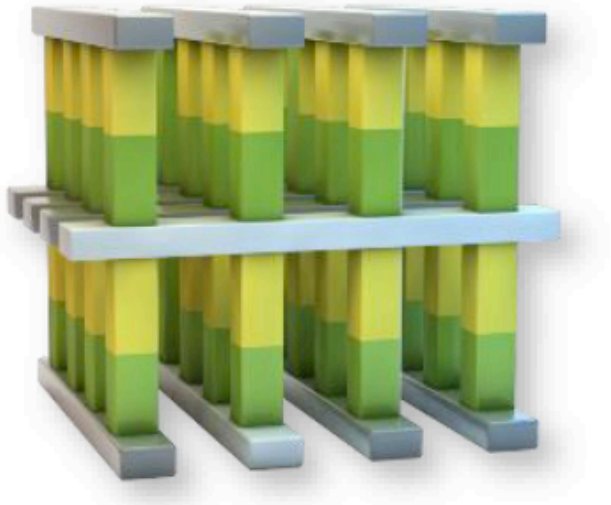
- Released last month!

Exceptional characteristics:

- Non-volatile memory
- 1000x more resilient than SSDs
- 8-10x density of DRAM
- Performance in DRAM ballpark!

The Future: Nonvolatile Memories in Server Architecture

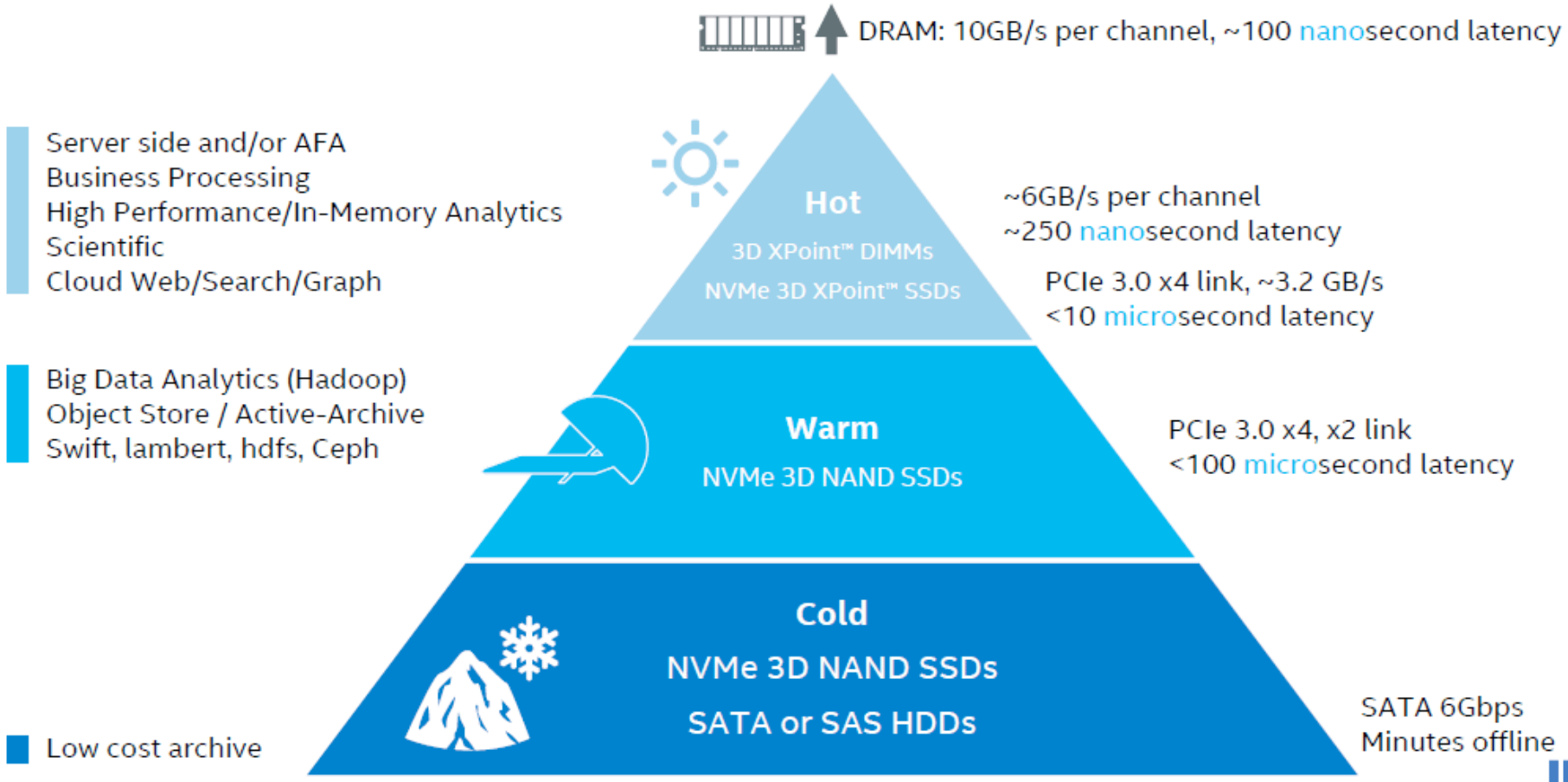




- 3D XPoint technology provides the benefit in the middle
- It is considerably faster than NAND Flash
- Performance can be realized on PCIe or DDR buses
- Lower cost per bit than DRAM while being considerably more dense

(https://www.youtube.com/watch?v=IWsjbqbkqh8)

Storage Hierarchy Tomorrow



Comparisons between memory technologies based on in-market product specifications and internal Intel specifications.

IDF15 INTEL DEVELOPER FORUM

High-Bandwidth Memory Buses

Today's DDR4 maxes out at 25.6 GB/sec

High Bandwidth Memory (HBM) led by AMD and NVIDIA

- Supports 1,024 bit-wide bus @ 125 GB/sec

Hybrid Memory Cube (HMC) consortium led by Intel

- To be release in 2016
- Claimed that 400 GB/sec possible!

Both based on stacked memory chips

- Limited capacity (won't replace DRAM), but much higher than on-chip caches
- Example use cases: GPGPUs

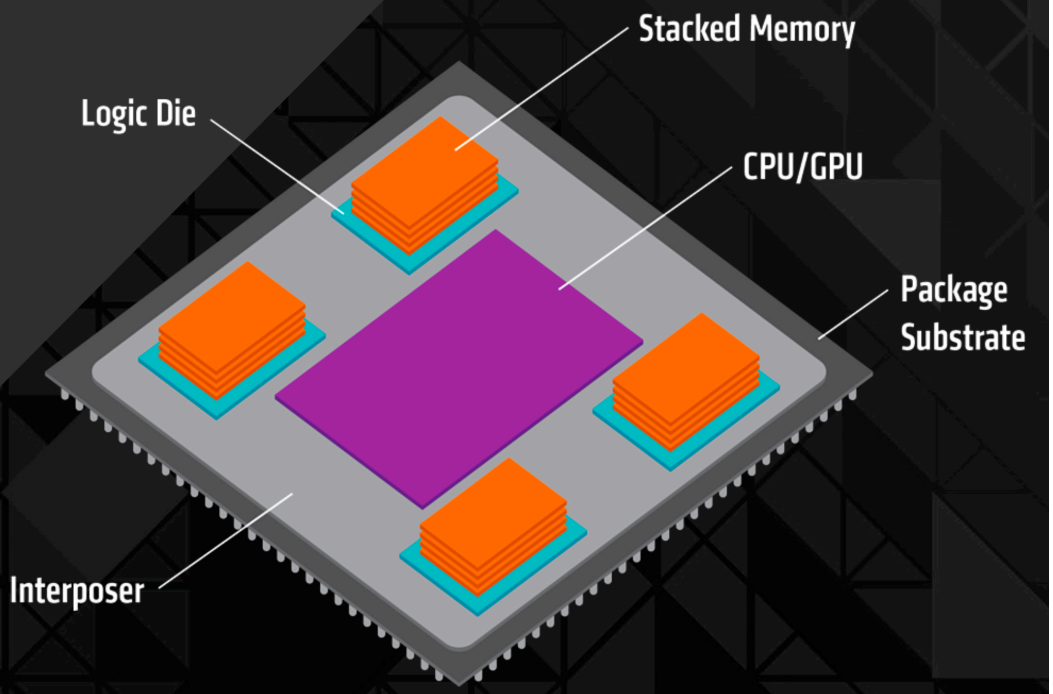
Example: HBM

THE INTERPOSER

THE NEXT STEP IN INTEGRATION

AMD

- Brings DRAM as close as possible to the logic die
- Improving proximity enables extremely wide bus widths
- Improving proximity simplifies communication and clocking
- Improving proximity greatly improves bandwidth per watt
- Allows for integration of disparate technologies such as DRAM
- AMD developed industry partnerships with ASE, Amkor & UMC to develop the first high-volume manufacturable interposer solution



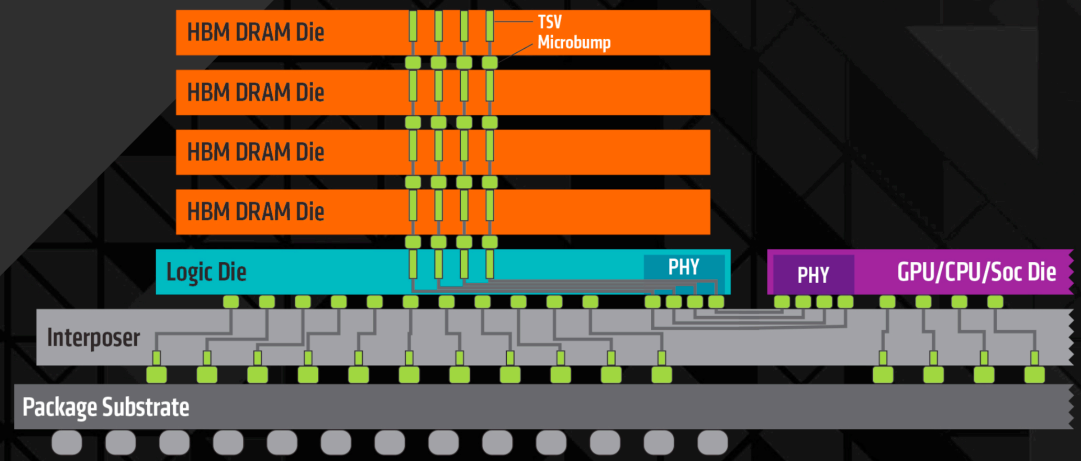
Example: HBM

HIGH-BANDWIDTH MEMORY

DRAM BUILT FOR AN INTERPOSER

AMD

- A new type of memory chip with low power consumption and an ultra-wide bus width
- Many of those chips stacked vertically like floors in a skyscraper
- New interconnects, called “through-silicon vias” (TSVs) and “µbumps”, connect one DRAM chip to the next
- TSVs and µbumps also used to connect the SoC/GPU to the interposer
- AMD and SK Hynix partnered to define and develop the first complete specification and prototype for HBM



A Second Summary

3D XPoint promises virtually unlimited memory

- Non-volatile
- Reads 2-3x slower than RAM
- Writes 2x slower than reads
- 10x higher density
- Main limit for now 6GB/sec interface

High memory bandwidth promise

- 5x increase in memory bandwidth or higher, but limited capacity so won't replace DRAM

What does this Mean?

Thoughts

For big data processing HDD are virtually dead!

- Still great for archival thought

With 3D XPoint, RAM will finally become the new disk

Gap between memory capacity and bandwidth still increasing

Thoughts

Storage hierarchy gets more and more complex:

- L1 cache
- L2 cache
- L3 cache
- RAM
- 3D XPoint based storage
- SSD
- (HDD)

Need to design software to take advantage of this hierarchy

Thoughts

Primary way to scale processing power is by adding more core

- Per core performance increase only 10-15% per year now
- HBM and HBC technologies will alleviate the bottleneck to get data to/from multi-cores, including GPUs

Moore's law is finally slowing down

Parallel computation models will become more and more important both at node and cluster levels

Thoughts

Will locality become more or less important?

New OSes that ignore disk and SSDs?

Aggressive pre-computations

- Indexes, views, etc
- Tradeoff between query latency and result availability

…