Good morning!

CS 744: DATACENTER AS A COMPUTER

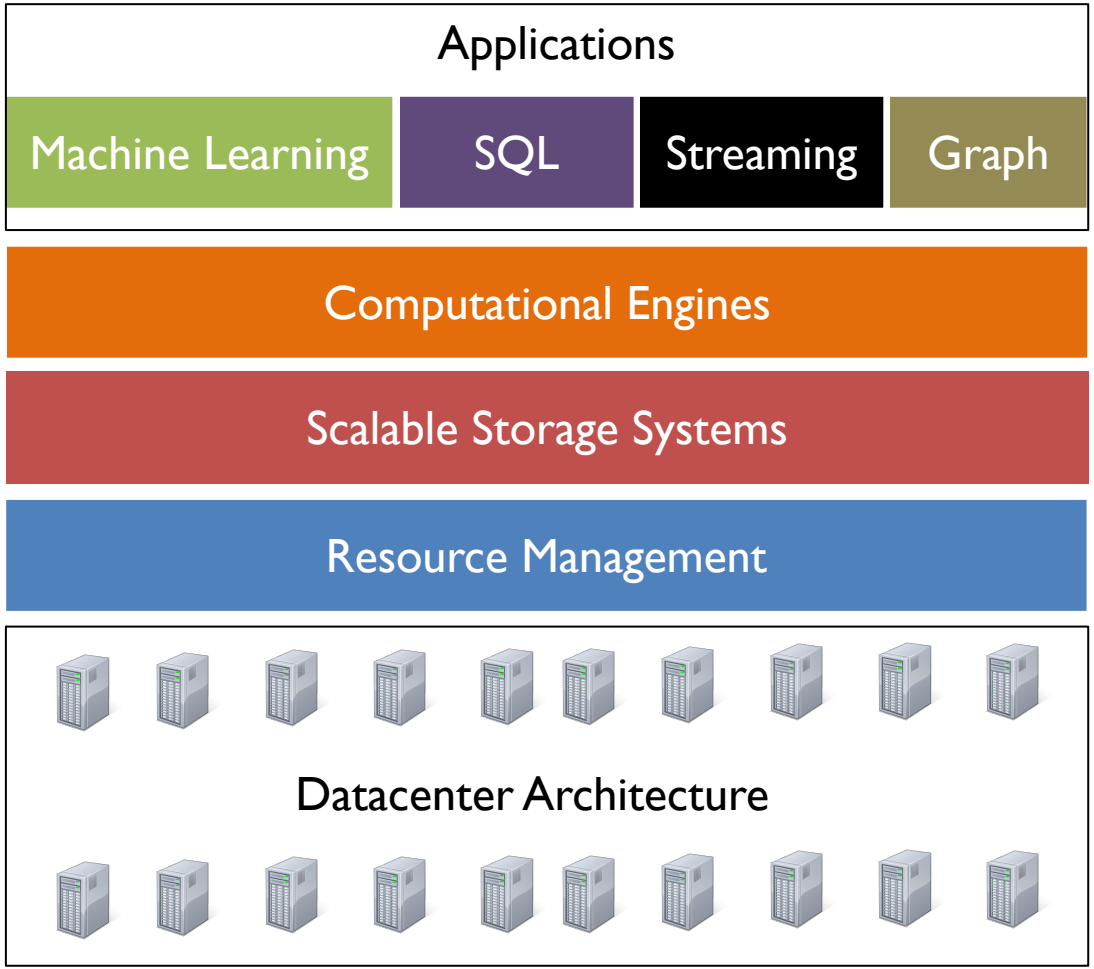
Shivaram Venkataraman

Fall 2021

ANNOUNCEMENTS

- Assignments
  - Assignment zero is due!
  - Form groups for Assignment 1 on Piazza

- Class format
  - Review
  - Lecture
  - Discussion



What is the hardware and what are its properties

OUTLINE

- Hardware Trends
- Datacenter design
- WSC workloads
- Discussion

WHY IS ONE MACHINE NOT ENOUGH?

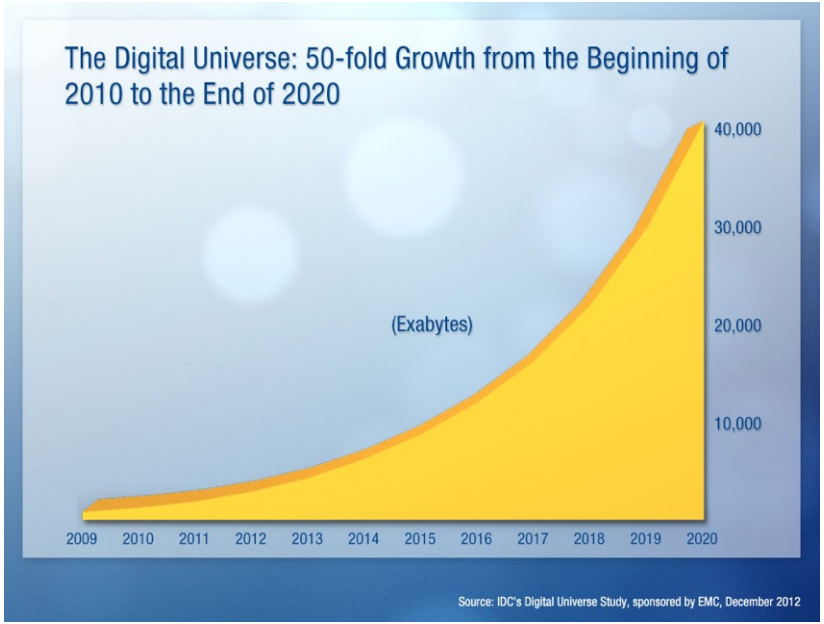
More data – doesn't fit in 1 machine

ML / video encoding
↳ parallelism when you use > 1 machine

Single point of failure

Applications might need specific hardware

Compute / Serving capacity

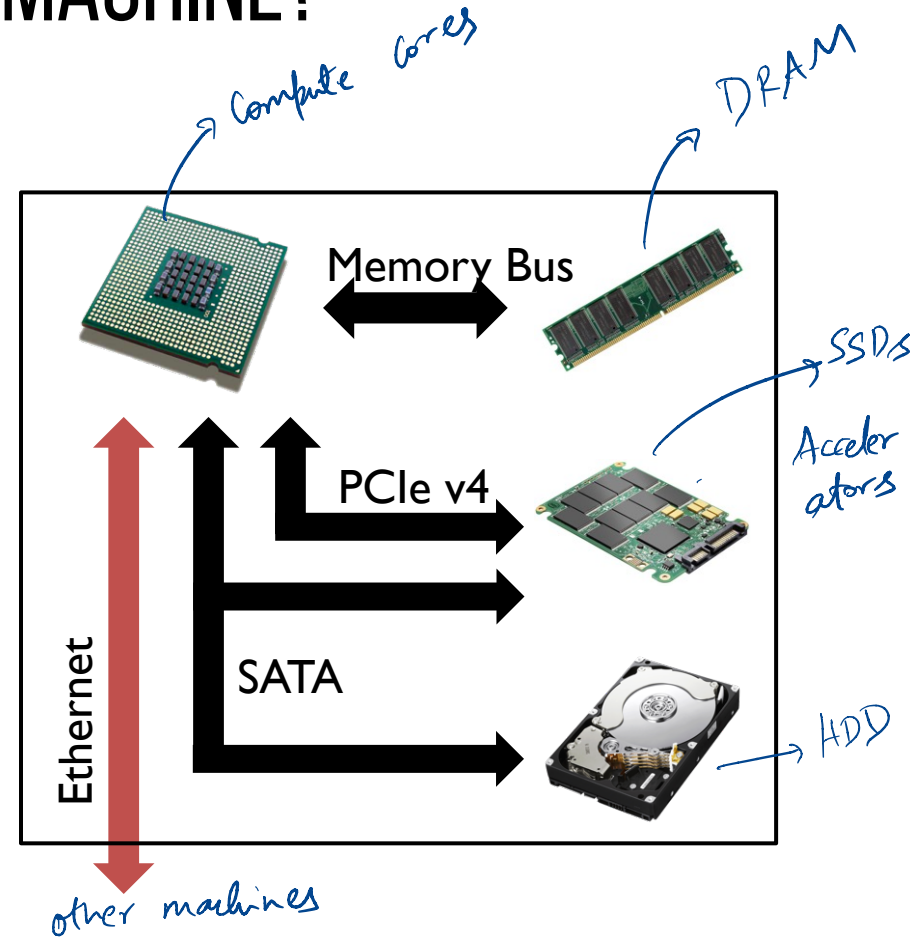


WHAT'S IN A MACHINE?

Interconnected compute and storage

Newer Hardware

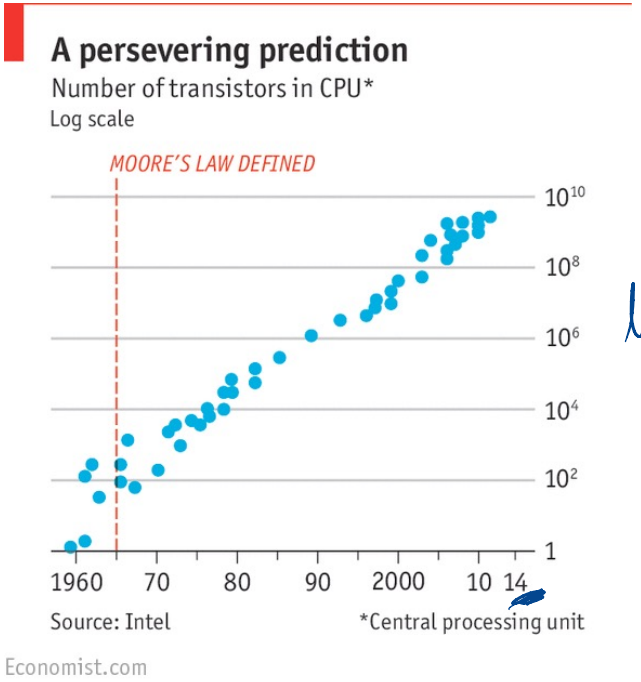
- GPUs, FPGAs
- RDMA, NVlink



SCALE UP: MAKE MORE POWERFUL MACHINES

Moore's law

- Stated by Intel founder Gordon Moore
- Number of transistors on microchip double every 2 years
- Today "closer to 2.5 years" Intel CEO Brian Krzanich



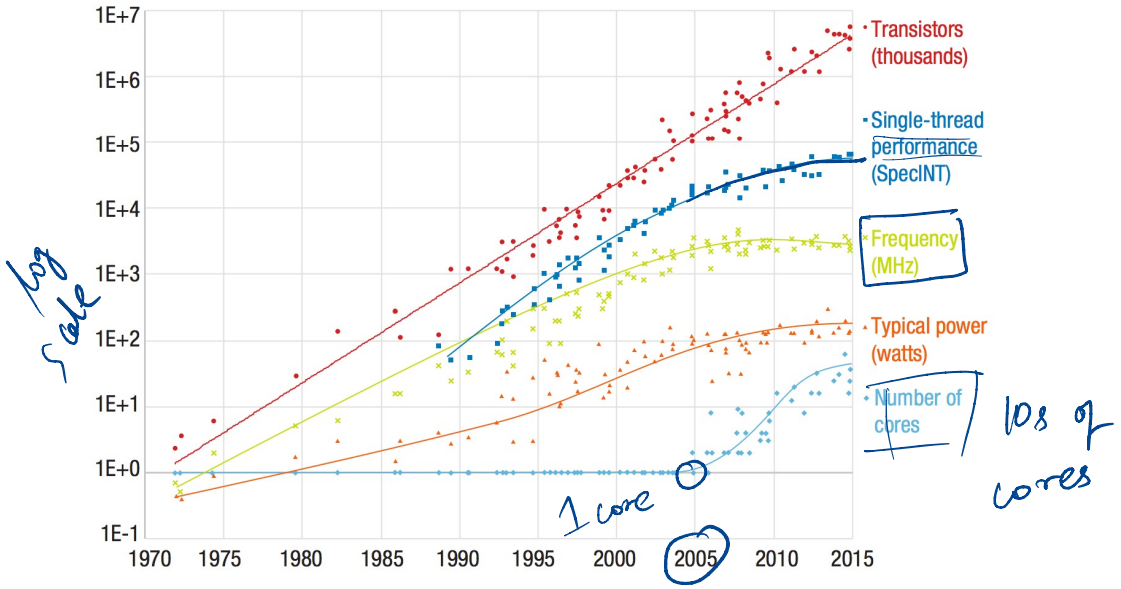
log scale

DENNARD SCALING IS THE PROBLEM

Suggested that power requirements are proportional to the area for transistors

- Both voltage and current being proportional to length
- Stated in 1974 by Robert H. Dennard (DRAM inventor)

Broken since 2005

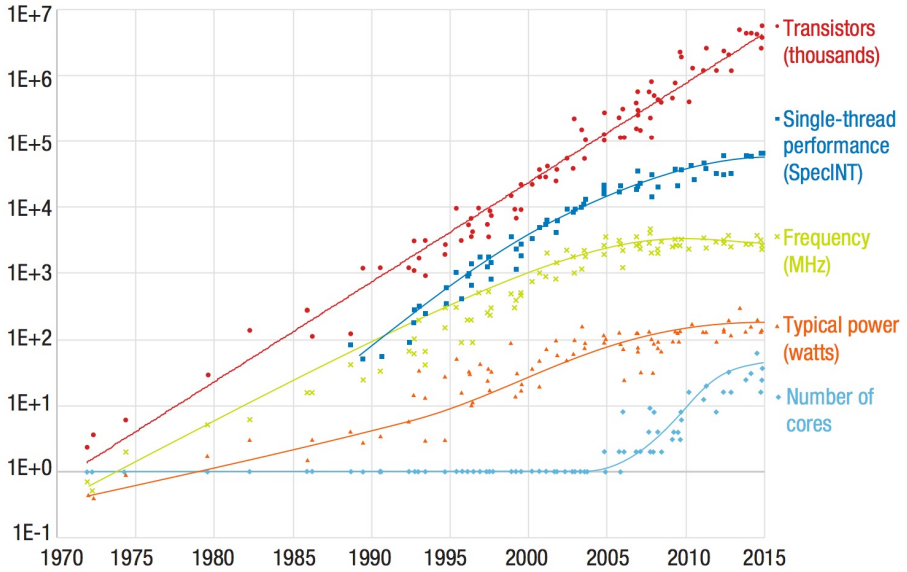


"Adapting to Thrive in a New Economy of Memory Abundance," Bresniker et al

DENNARD SCALING IS THE PROBLEM

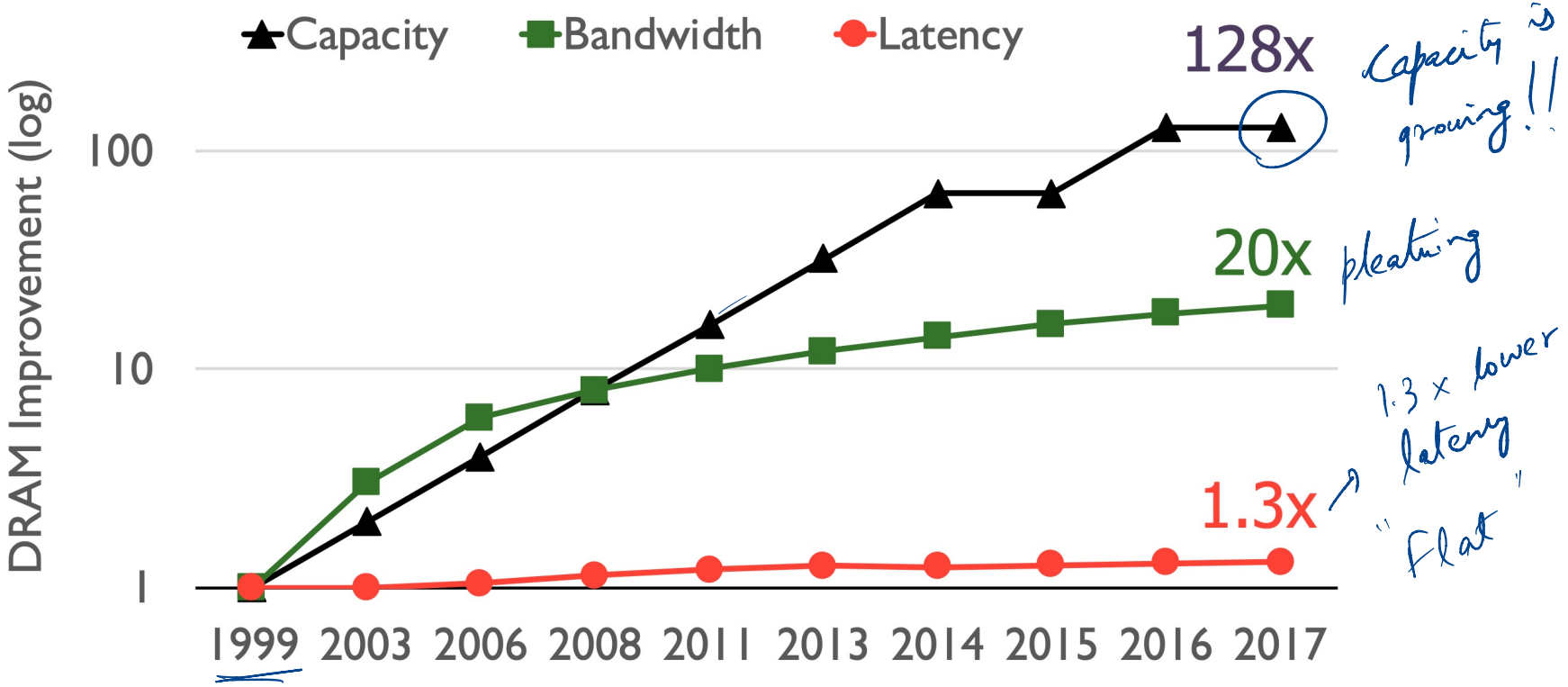
Performance per-core is stalled

Number of cores is increasing

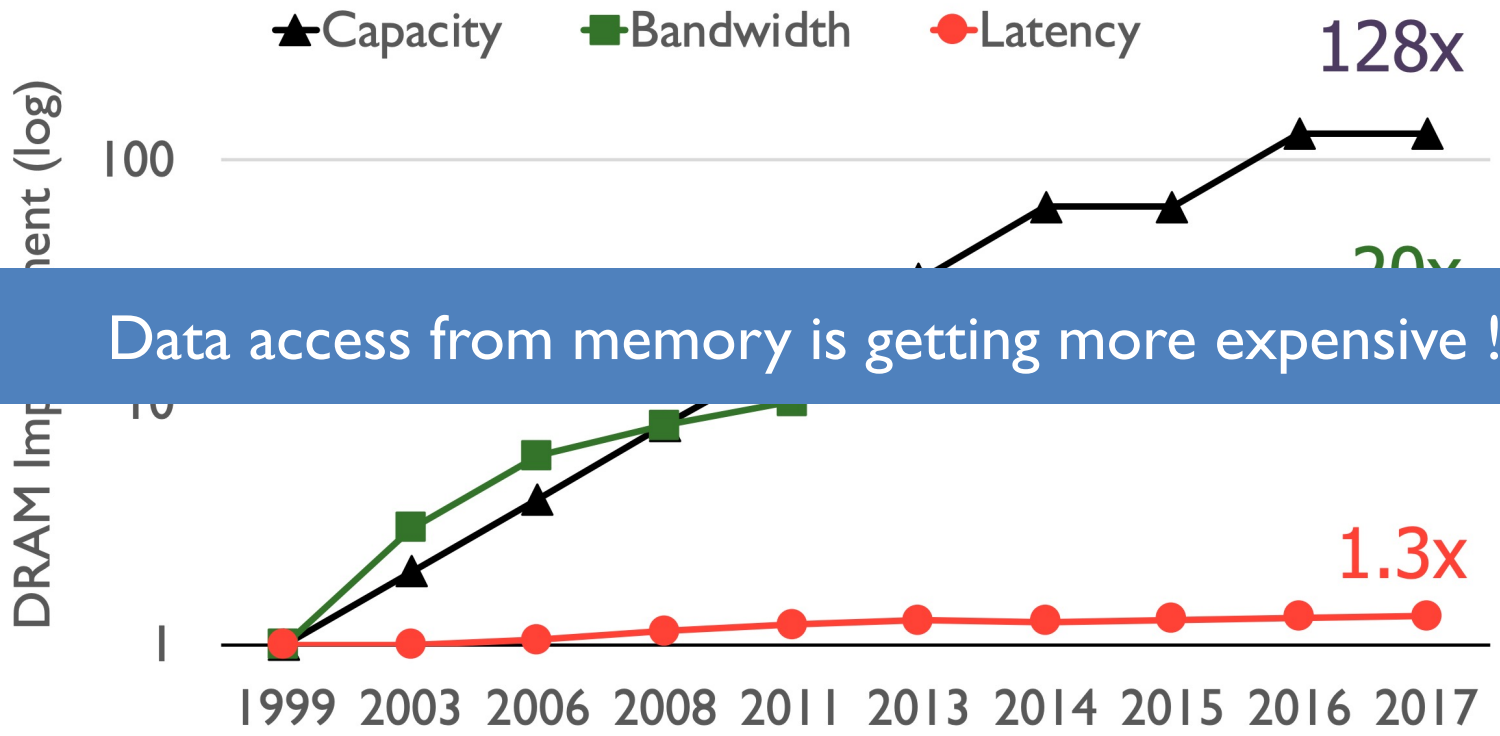


"Adapting to Thrive in a New Economy of Memory Abundance," Bresniker et al

MEMORY TRENDS



MEMORY TAKEAWAY

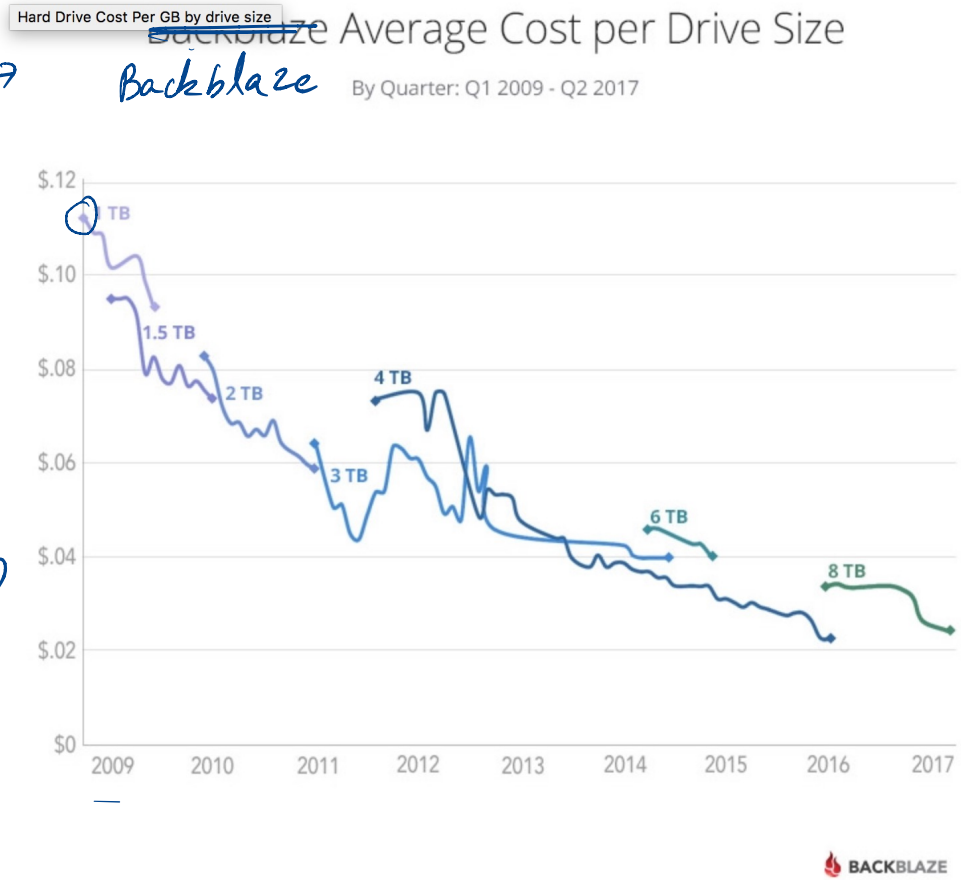


Data access from memory is getting more expensive !

HDD CAPACITY

Cloud backup

Price per GB



HDD BANDWIDTH

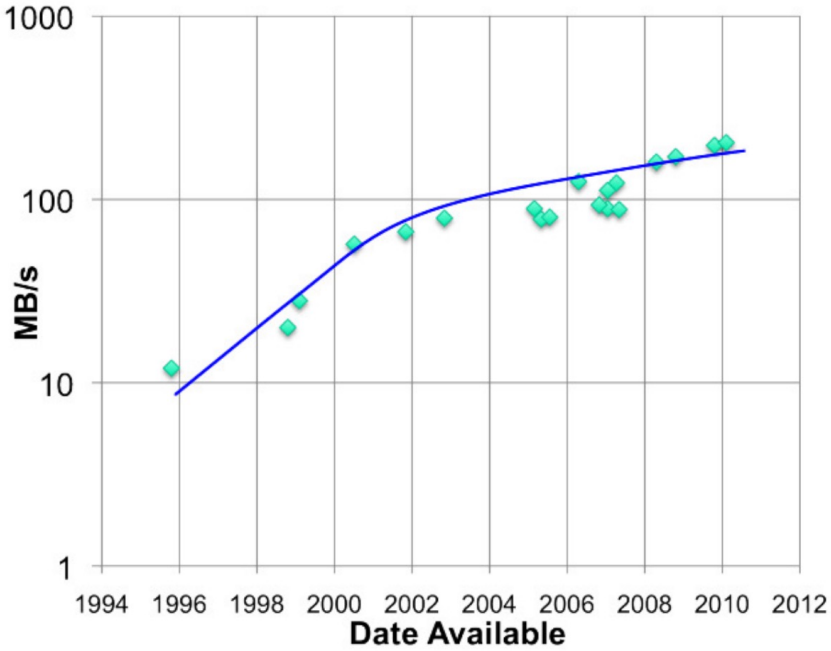


Figure 4: Maximum sustained bandwidth trend

Disk bandwidth is not growing

SSDS

Performance:

- Reads: 25us latency ~ 10us
- Write: 200us latency
- Erase: 1,5 ms

Steady state, when SSD full

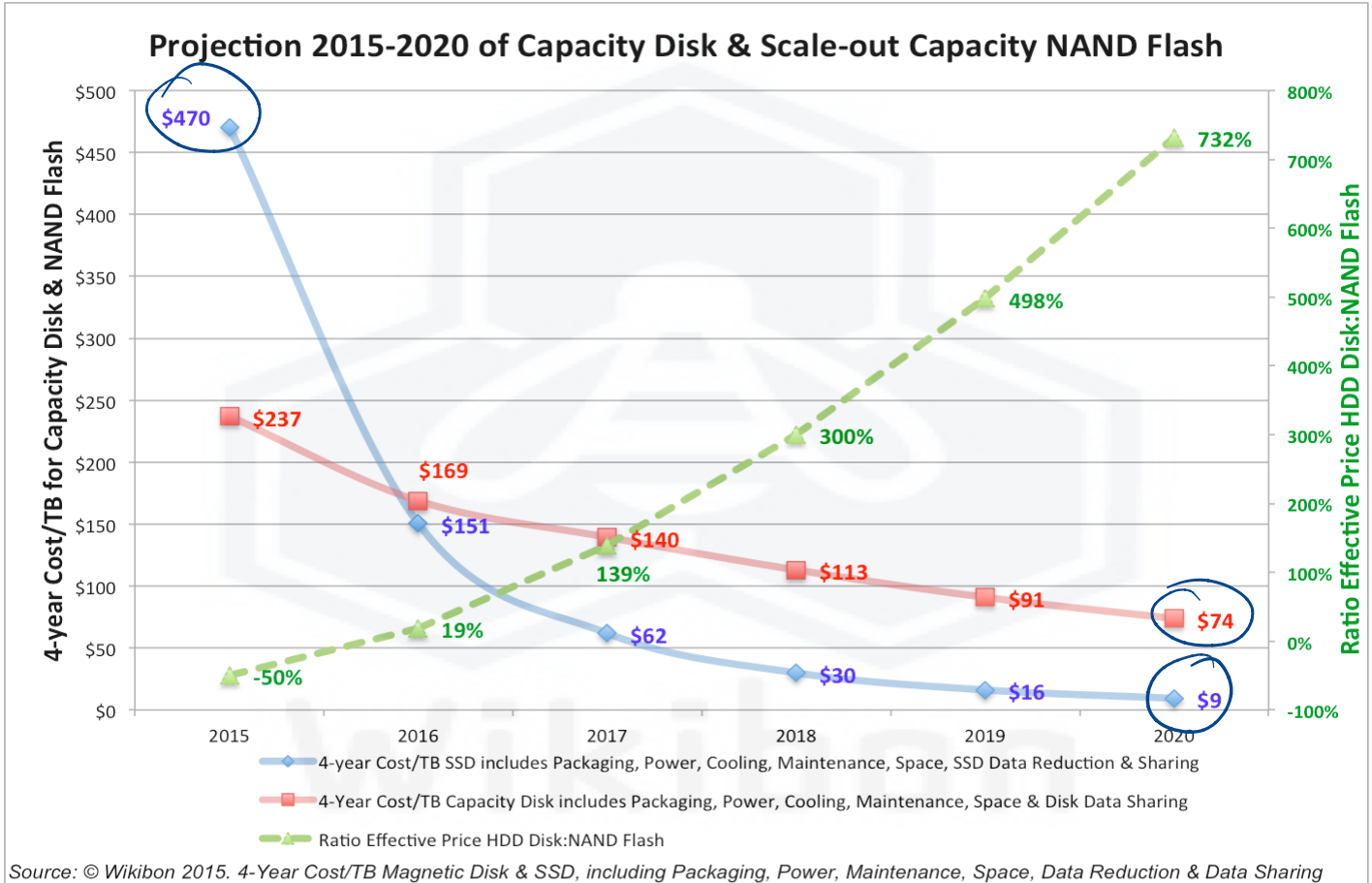
- One erase every 64 or 128 reads (depending on page size)

Lifetime: 100,000-1 million writes per page

Tape/ HDDs are used for archival storage

SSD VS HDD COST

SSDs are widespread

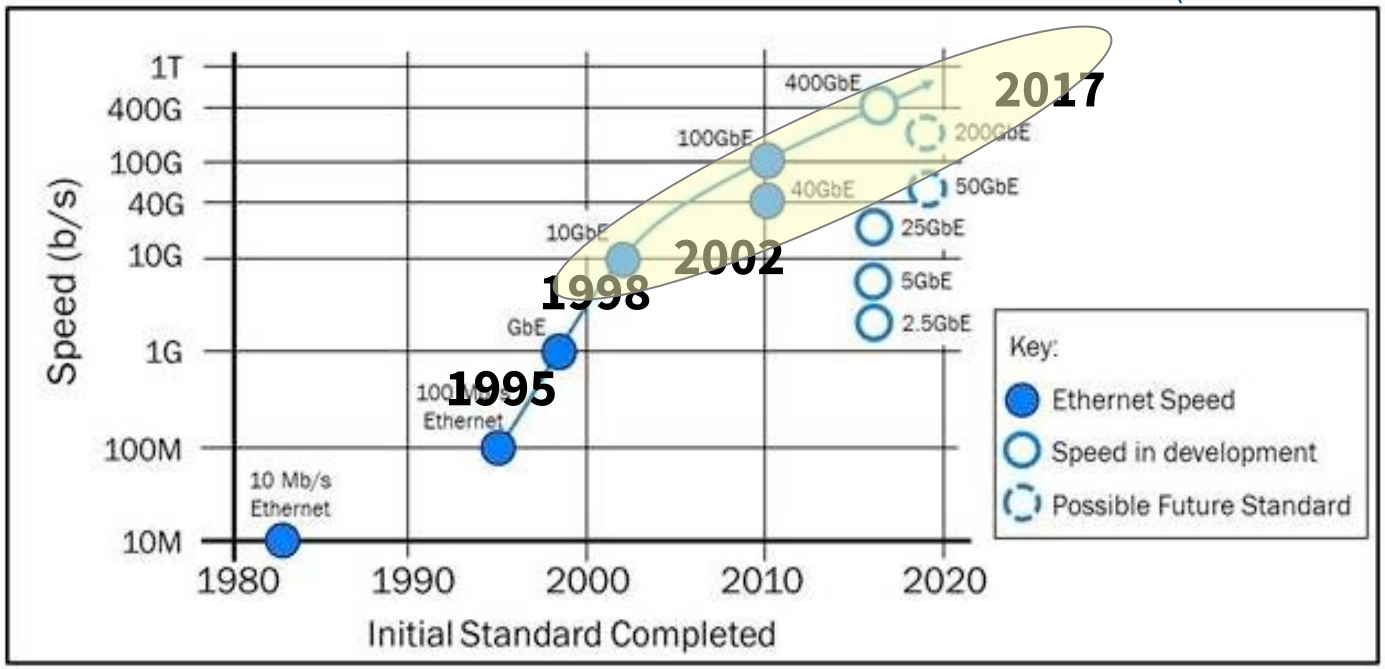


Source: © Wikibon 2015. 4-Year Cost/TB Magnetic Disk & SSD, including Packaging, Power, Maintenance, Space, Data Reduction & Data Sharing

ETHERNET BANDWIDTH

Growing 33-40% per year !

40 Gbps or 100 Gbps Ethernet



AMAZON EC2 (2019)

1 GPUs
4, 8 ...

New – EC2 P3dn GPU Instances with 100 Gbps Networking & Local NVMe Storage

new SSD standard
& promises faster (low latency)
higher bandwidth

TRENDS SUMMARY

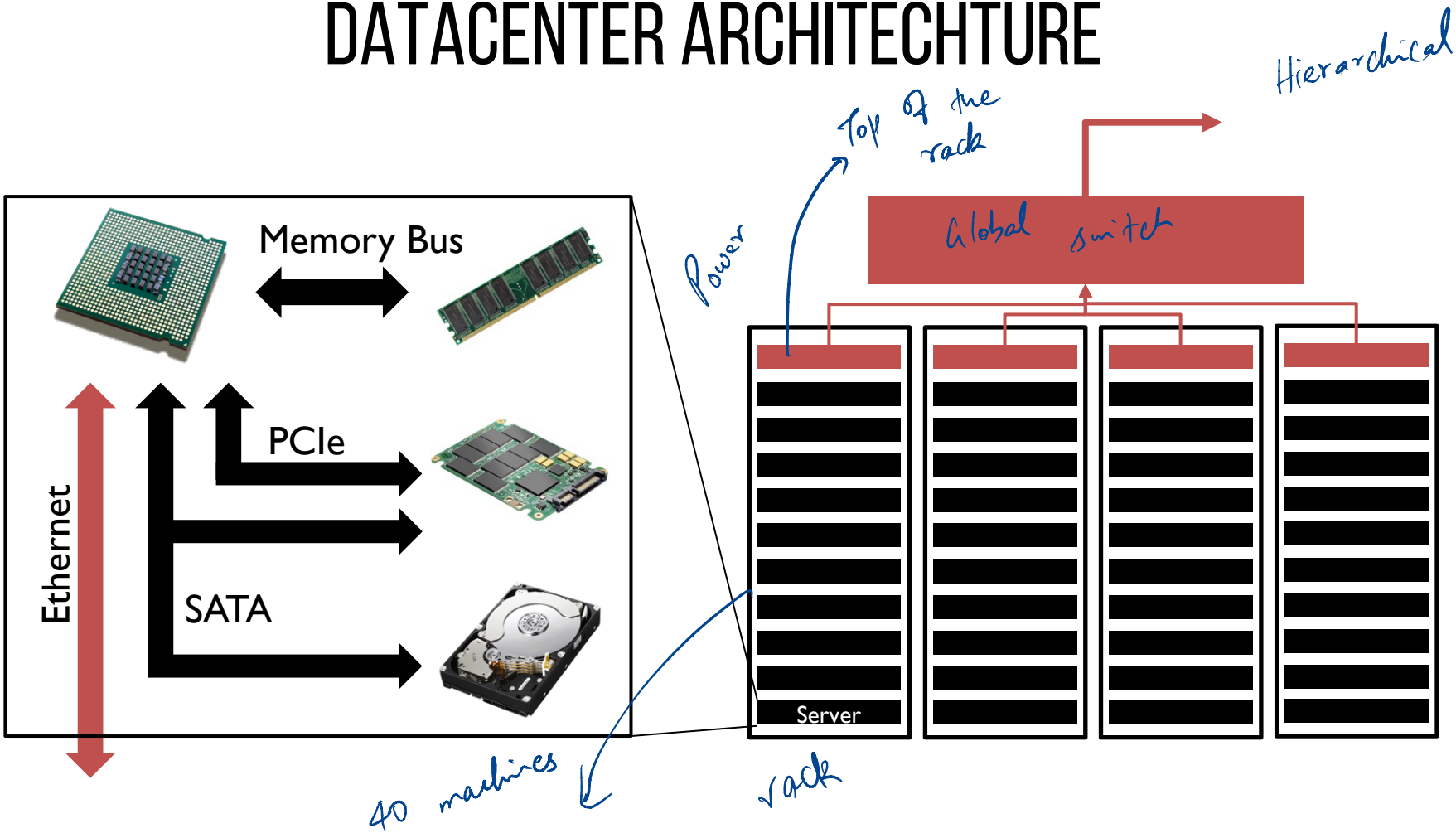
CPU speed per core is flat

Memory bandwidth growing slower than capacity

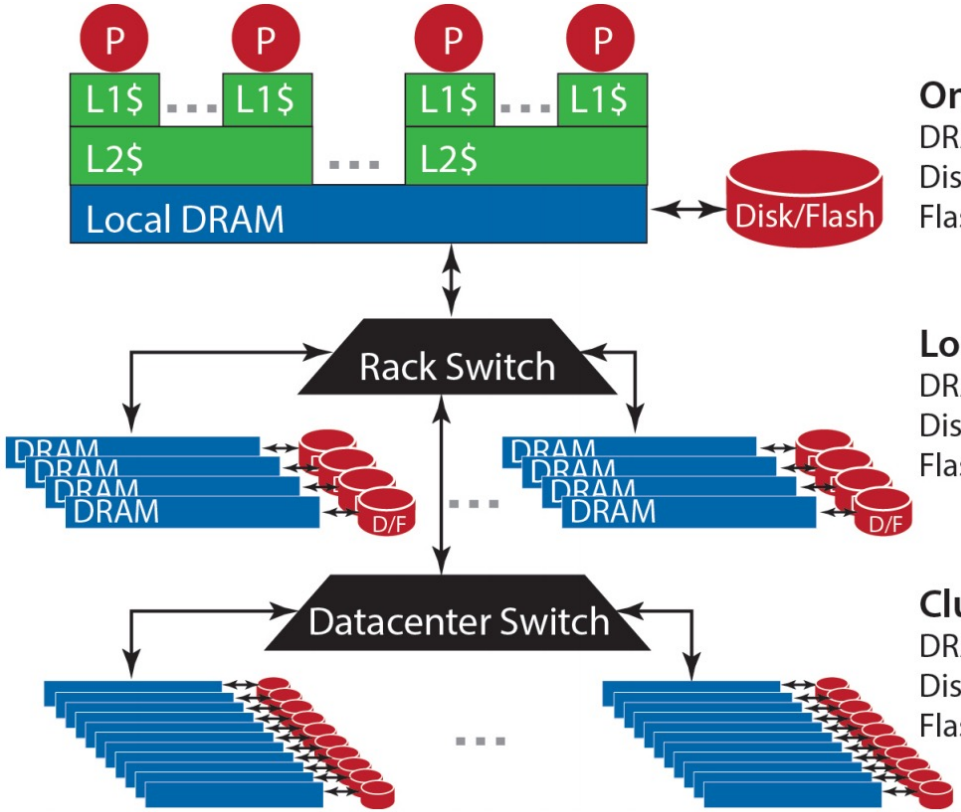
SSD, NVMe replacing HDDs

Ethernet bandwidth growing

DATACENTER ARCHITECHTURE



STORAGE HIERARCHY (DC AS A COMPUTER V2)



Capacity

latency → Bandwidth

One Server

DRAM: 16 GB, 100 ns, 20 GB/s
Disk: 2T B, 10 ms, 200 MB/s
Flash: 128 GB, 100 us, 1 GB/s

network latency, bandwidth

Local Rack (80 servers)

DRAM: 1 TB, 300 us, 100 MB/s
Disk: 160 TB, 11 ms, 100 MB/s
Flash: 20 TB, 400 us, 100 MB/s

Cluster (30 racks)

DRAM: 30 TB, 500 us, 10 MB/s
Disk: 4.80 PB, 12 ms, 10 MB/s
Flash: 600 TB, 600 us, 10 MB/s

WAREHOUSE-SCALE COMPUTERS

Single organization

Homogeneity (to some extent)

Cost efficiency at scale

- Multiplexing across applications and services
- Rent it out!

Many concerns

- Infrastructure
- Networking
- Storage
- Software
- Power/Energy
- Failure/Recovery
- …

SOFTWARE IMPLICATIONS

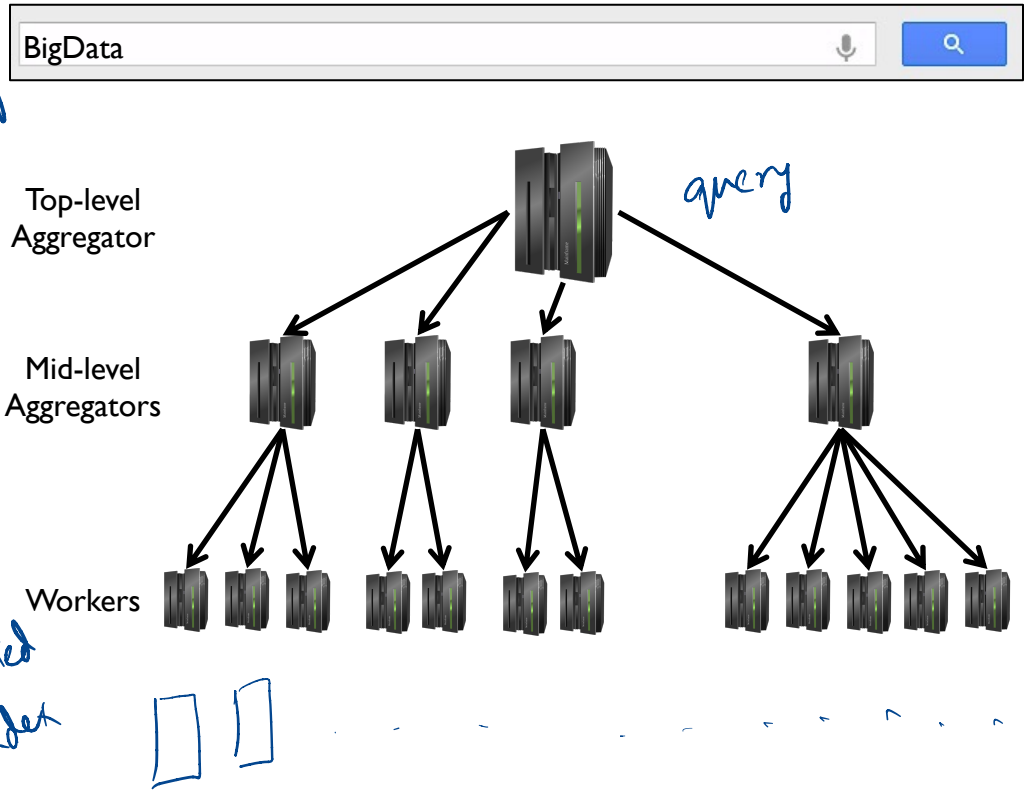
Reliability

Storage Hierarchy

Workload Diversity

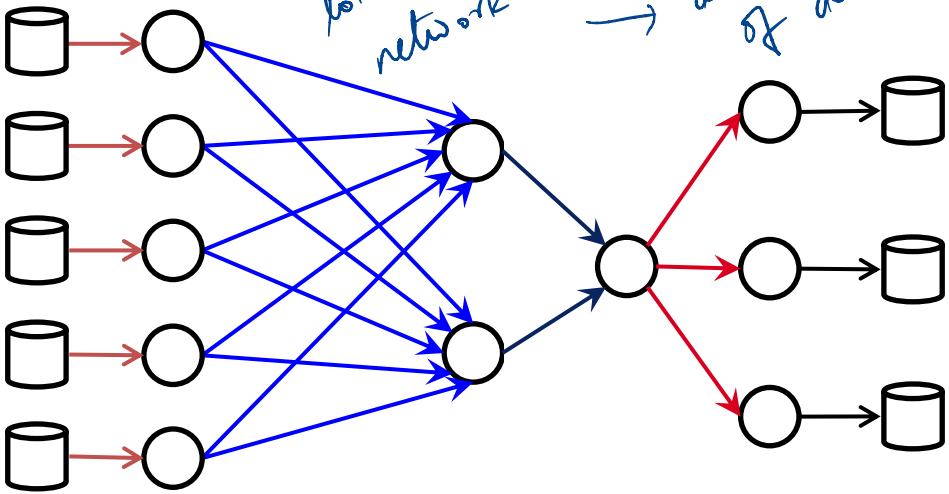
Single organization

WORKLOAD: PARTITION-AGGREGATE



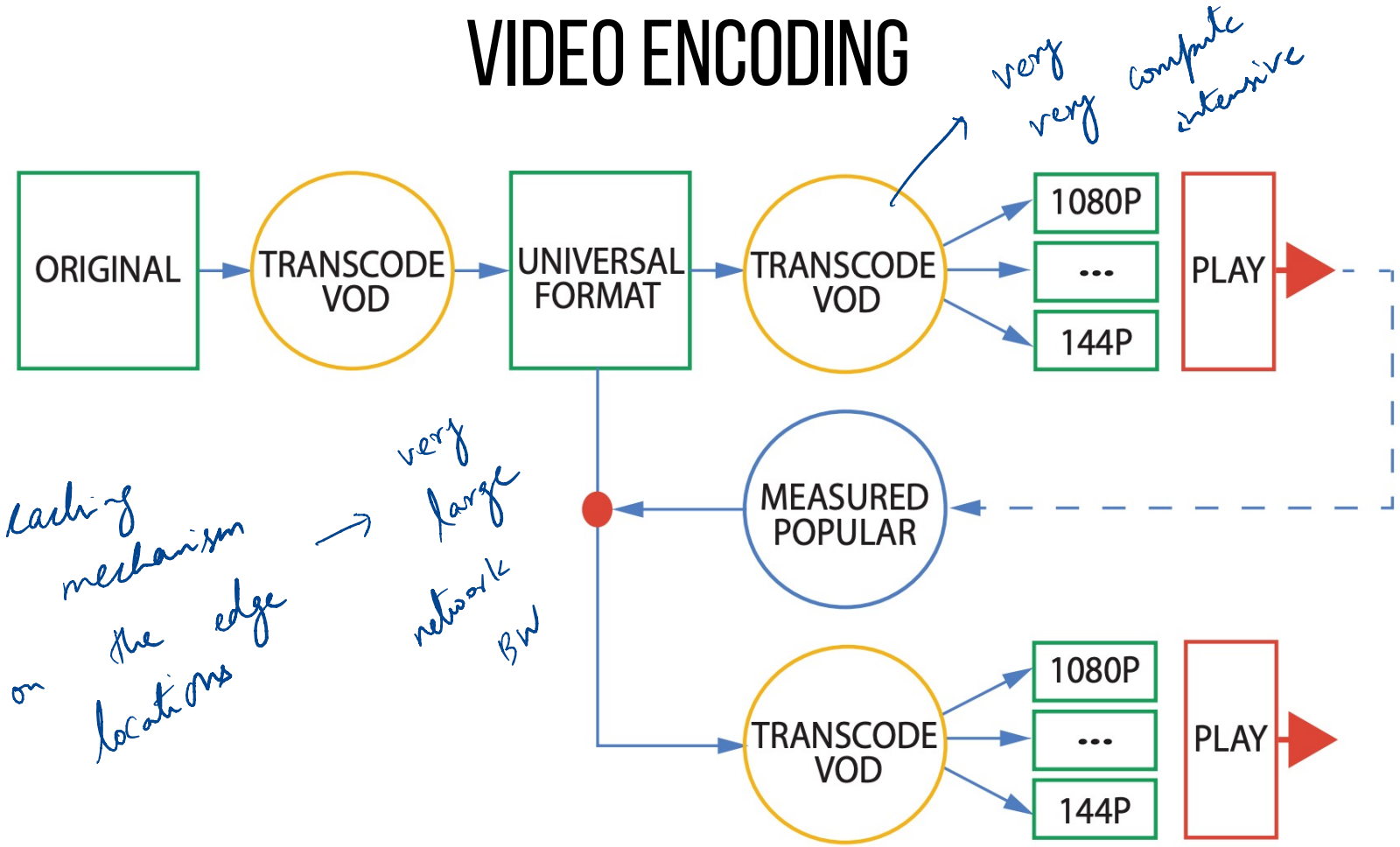
WORKLOAD: SCHOLAR SIMILARITY

Process a large number of articles (large data)



Not latency sensitive

VIDEO ENCODING



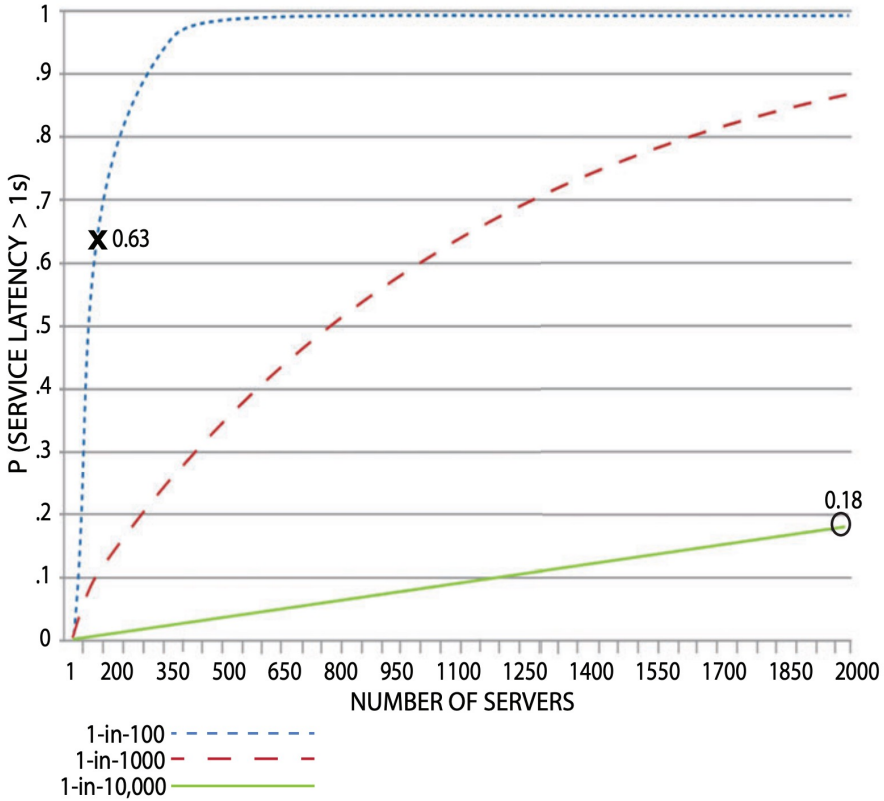
MACHINE LEARNING

Table 2.1: Six production applications plus ResNet benchmark. The fourth column is the total number of operations (not execution rate) that training takes to converge.

Type of Neural Network	Parameters (MiB)	Training			Inference
		Examples to Convergence	ExaOps to Conv	Ops per Example	Ops per Example
MLP0	225	1 trillion	353	353 Mops	118 Mops
MLP1	40	650 billion	86	133 Mops	44 Mops
LSTM0	498	1.4 billion	42	29 Gops	9.8 Gops
LSTM1	800	656 million	82	126 Gops	42 Gops
CNN0	87	1.64 billion	70	44 Gops	15 Gops
CNN1	104	204 million	7	34 Gops	11 Gops
ResNet	98	114 million	<3	23 Gops	8 Gops

DISCUSSION

https://forms.gle/nFrMPkSAWTjMcgUp6



Scale-up vs Scale-out

DISCUSSION

100 GB of memory

Scale-up vs Scale-out

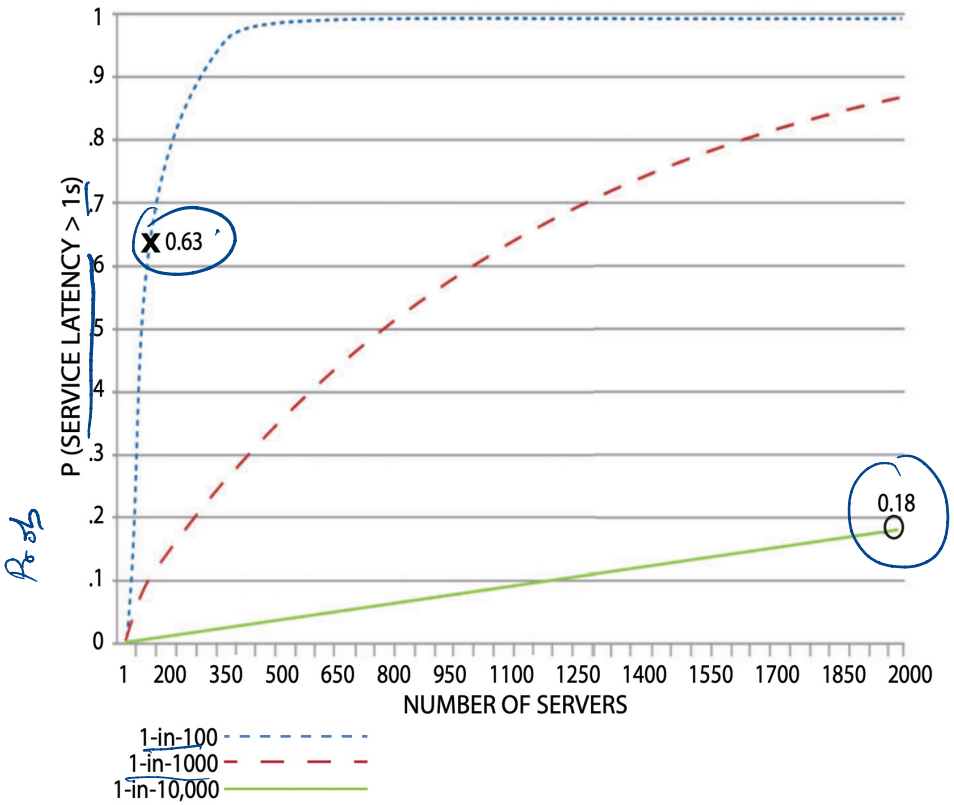
Scale up [1 big machine]

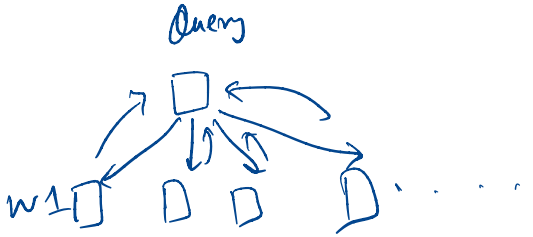
- Workloads that are hard to parallelize
- less network B/W
  ⇓
- less communication overhead
  ↳ depends on the workload

Scale-Out [10 × 10 GB of memory]

- handling very large capacity / large data
- incremental deployments
- Fault tolerance
  ↳ fail reliability slow down stragglers

DISCUSSION





1 - in - 100

⇒ w1 returns slow response 1% of time

NEXT STEPS

Next class: Storage Systems

Assignment 1 out Thursday.

Submit groups before that!