

# CS 744: DATACENTER AS A COMPUTER

Shivaram Venkataraman

Spring 2025

# ANNOUNCEMENTS

- Assignments
  - Assignment zero is due!
  - Form groups for Assignment I on Piazza
- Class format
  - Review
  - Lecture
  - Discussion

# Applications

Machine Learning

SQL

Streaming

Graph

Computational Engines

Scalable Storage Systems

Resource Management



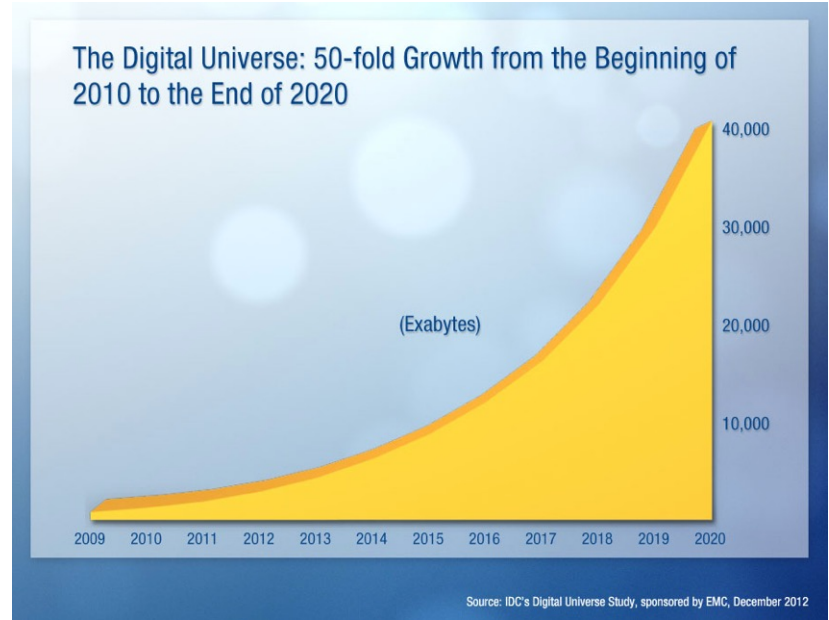
Datacenter Architecture



# OUTLINE

- Hardware Trends
- Datacenter design
- WSC workloads
- Discussion

# WHY IS ONE MACHINE NOT ENOUGH?

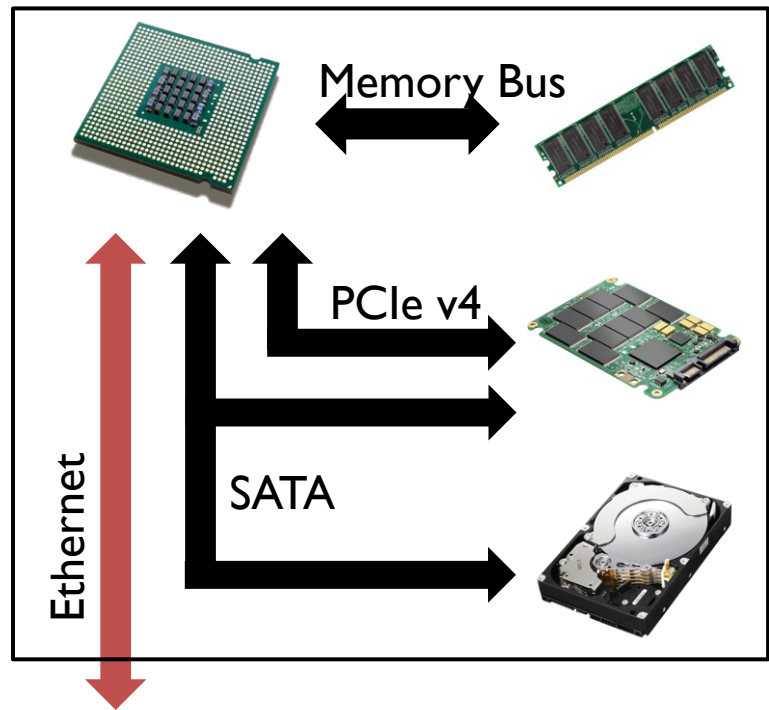


# WHAT'S IN A MACHINE?

Interconnected compute and storage

Newer Hardware

- GPUs, FPGAs
- RDMA, NVlink



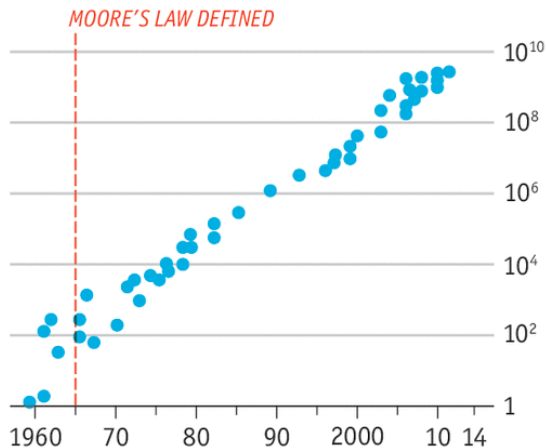
# SCALE UP: MAKE MORE POWERFUL MACHINES

## 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

### A persevering prediction

Number of transistors in CPU\*  
Log scale



Source: Intel

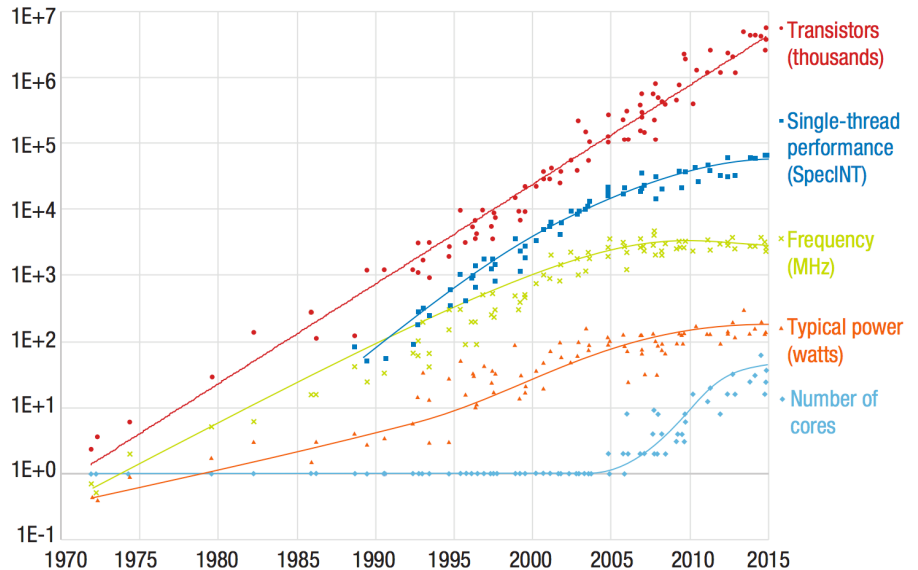
\*Central processing unit

# 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

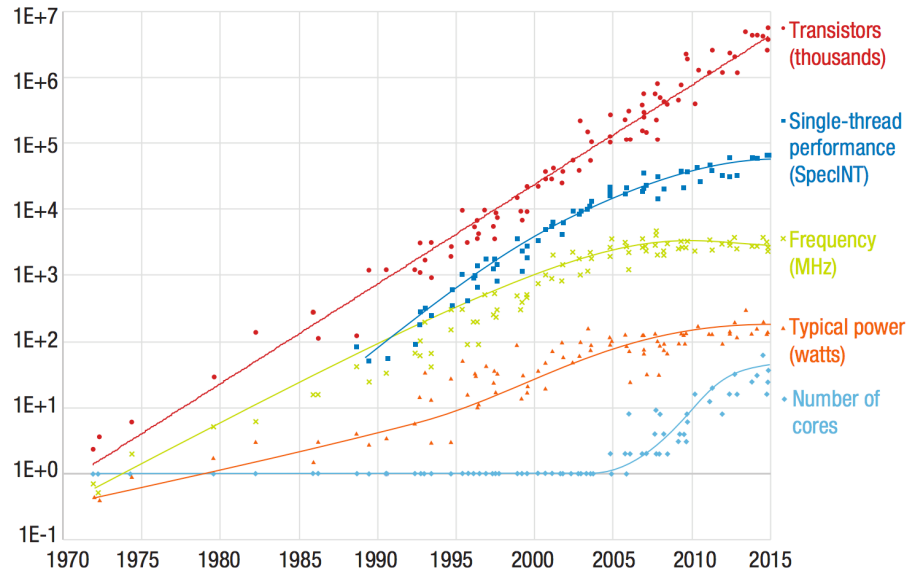


“Adapting to Thrive in a New Economy of Memory Abundance,” Bresniker et al

# DENNARD SCALING IS THE PROBLEM

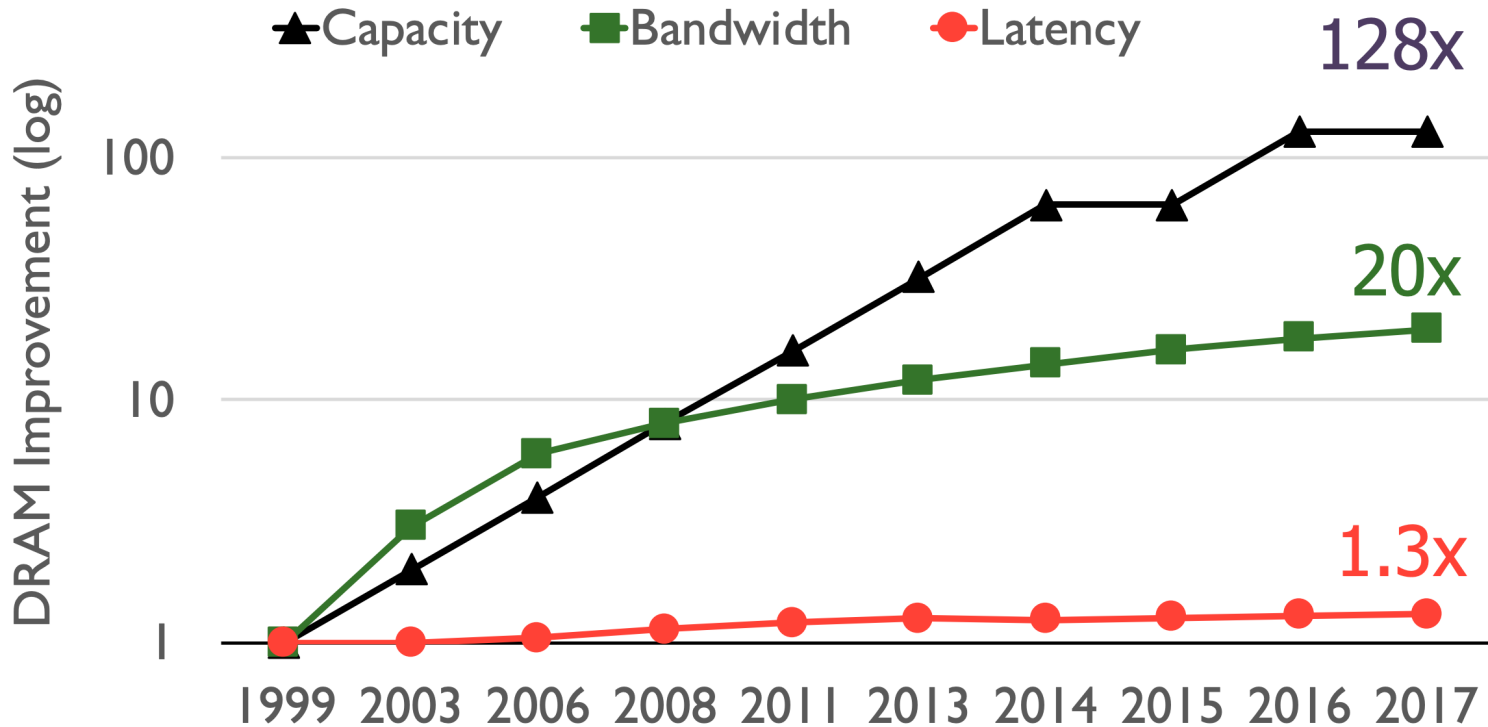
Performance per-core is stalled

Number of cores is increasing

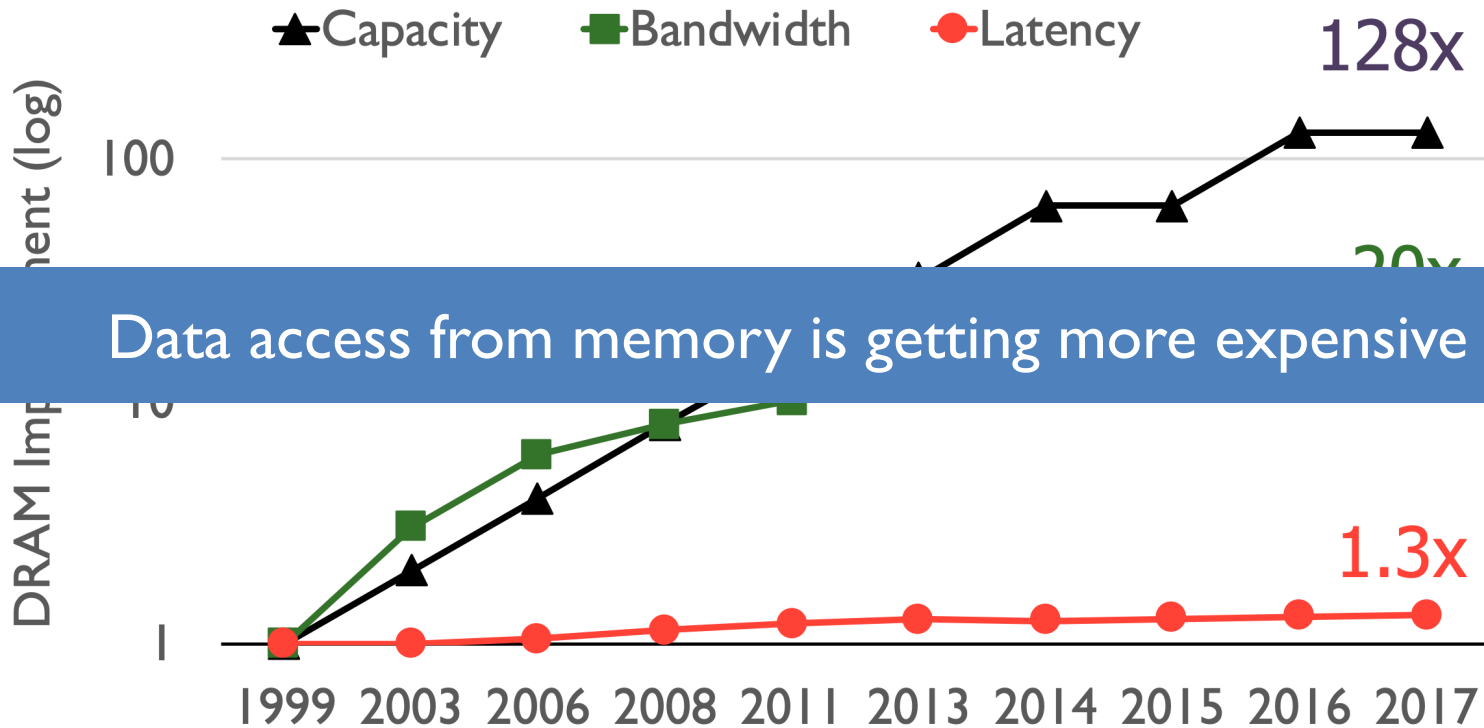


“Adapting to Thrive in a New Economy of Memory Abundance,” Bresnaker et al

# MEMORY TRENDS



# MEMORY TAKEAWAY

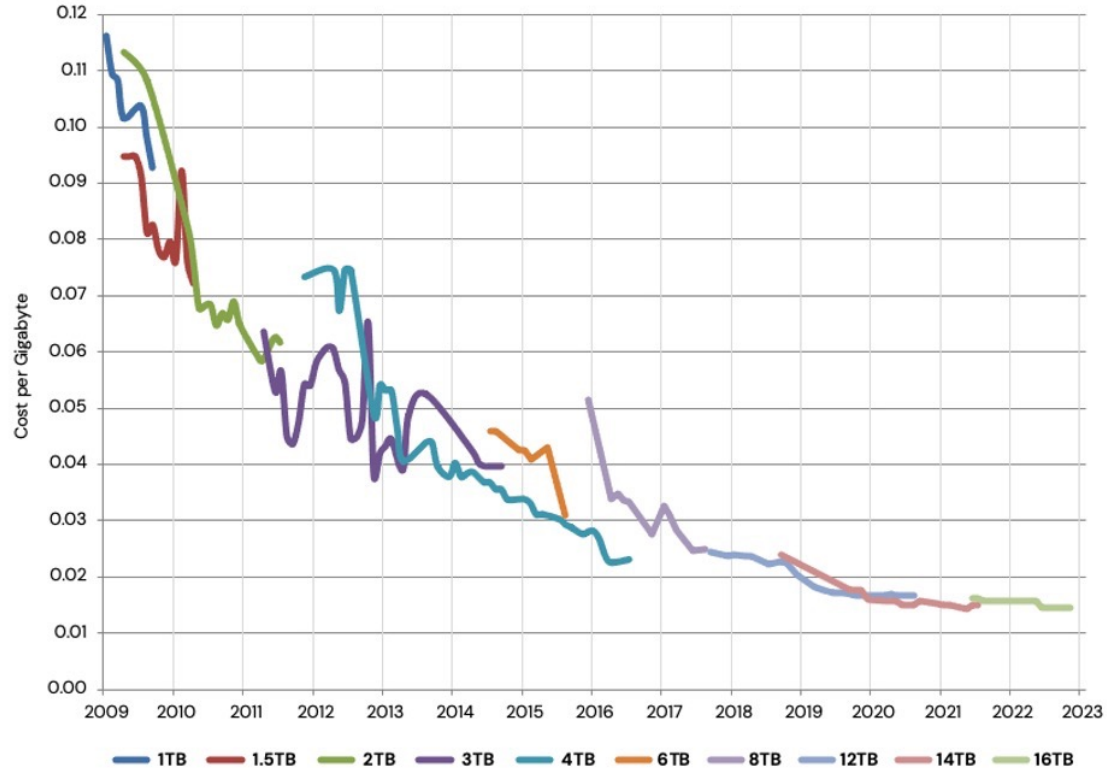


# HDD CAPACITY

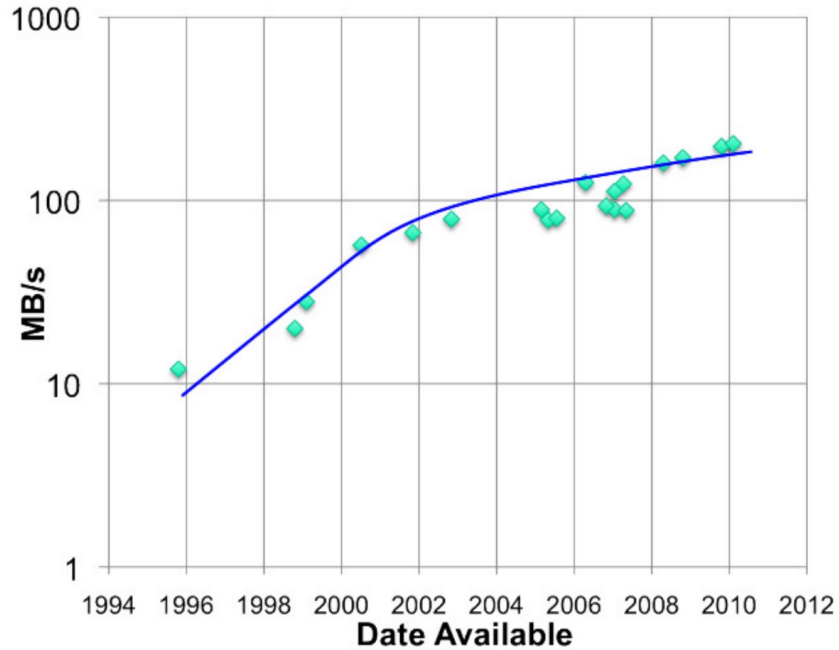
BackBlaze

### Backblaze Average Cost per Gigabyte by Drive Size Over Time

Drive sales grouped by drive size and month to compute average cost per month



# HDD BANDWIDTH



Disk bandwidth is not growing

Figure 4: Maximum sustained bandwidth trend

# SSDS

## Performance:

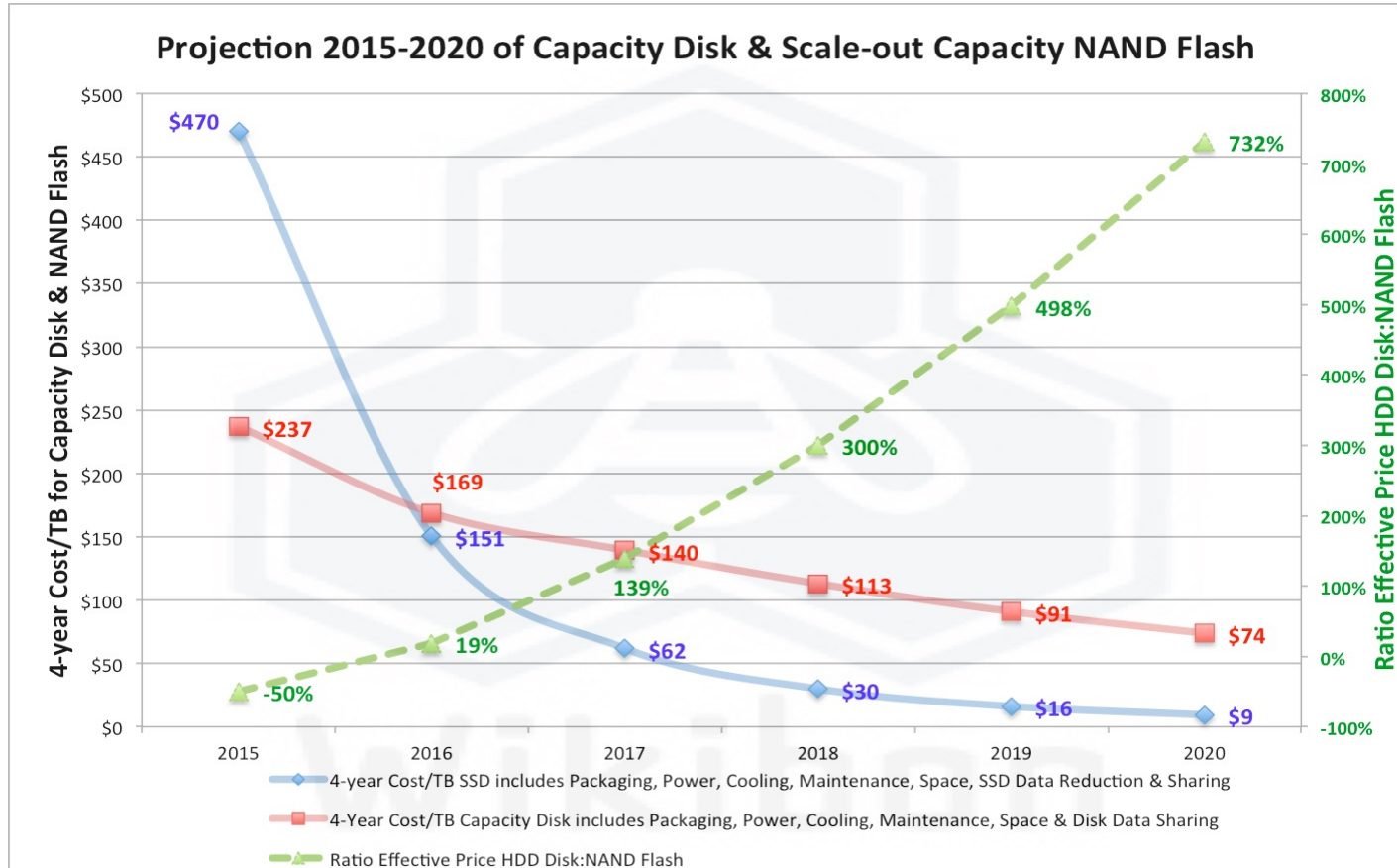
- Reads: 10-25us latency
- Write: 200us latency
- Erase: 1,5 ms

## Steady state, when SSD full

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

Lifetime: 100,000-1 million writes per page

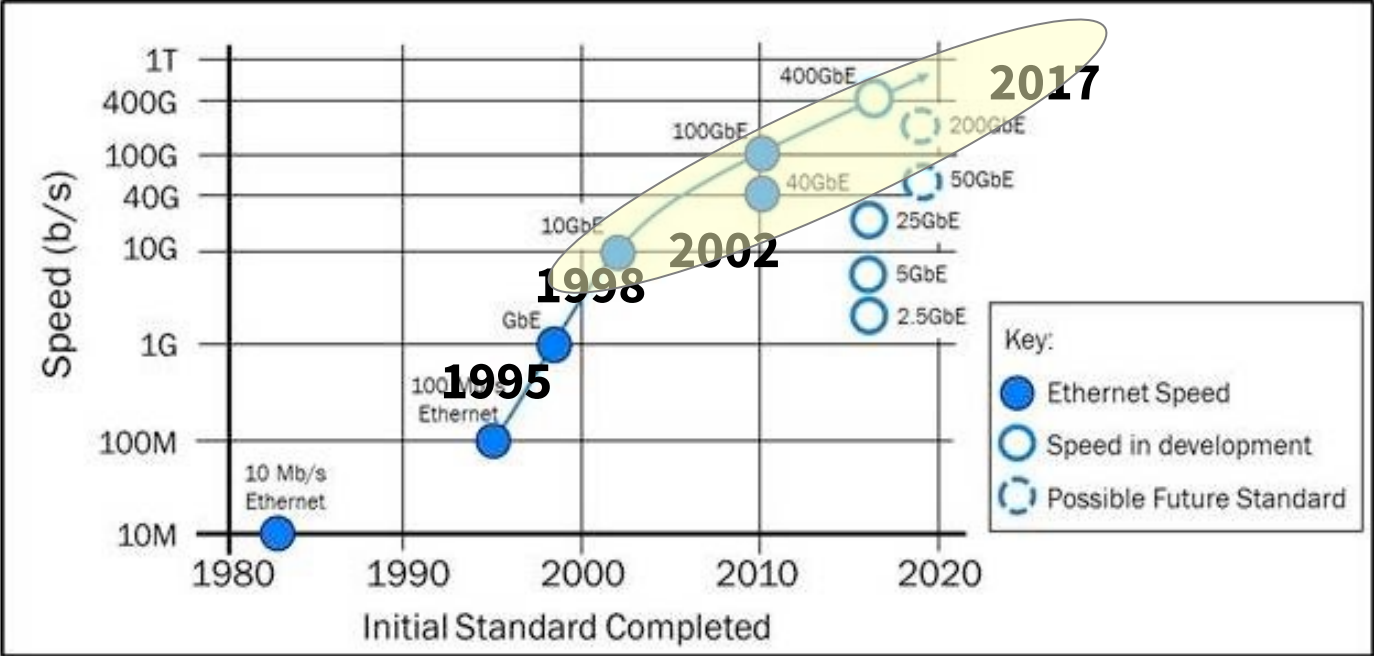
# SSD VS HDD COST



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

# ETHERNET BANDWIDTH

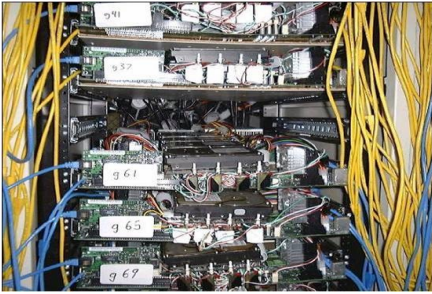
Growing **33-40%** per year !



# AMAZON EC2 (2019)

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

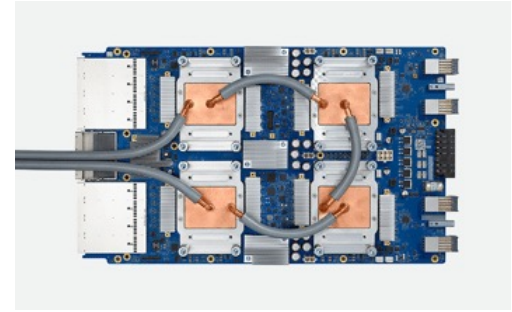
# HARDWARE EVOLUTION



Commodity CPUs  
Lots of disks  
Low bandwidth network  
(2001 Google)



GPUs – Graphics Cards  
Lots of parallelism  
Bigger power footprint  
Expensive!  
(~2010)



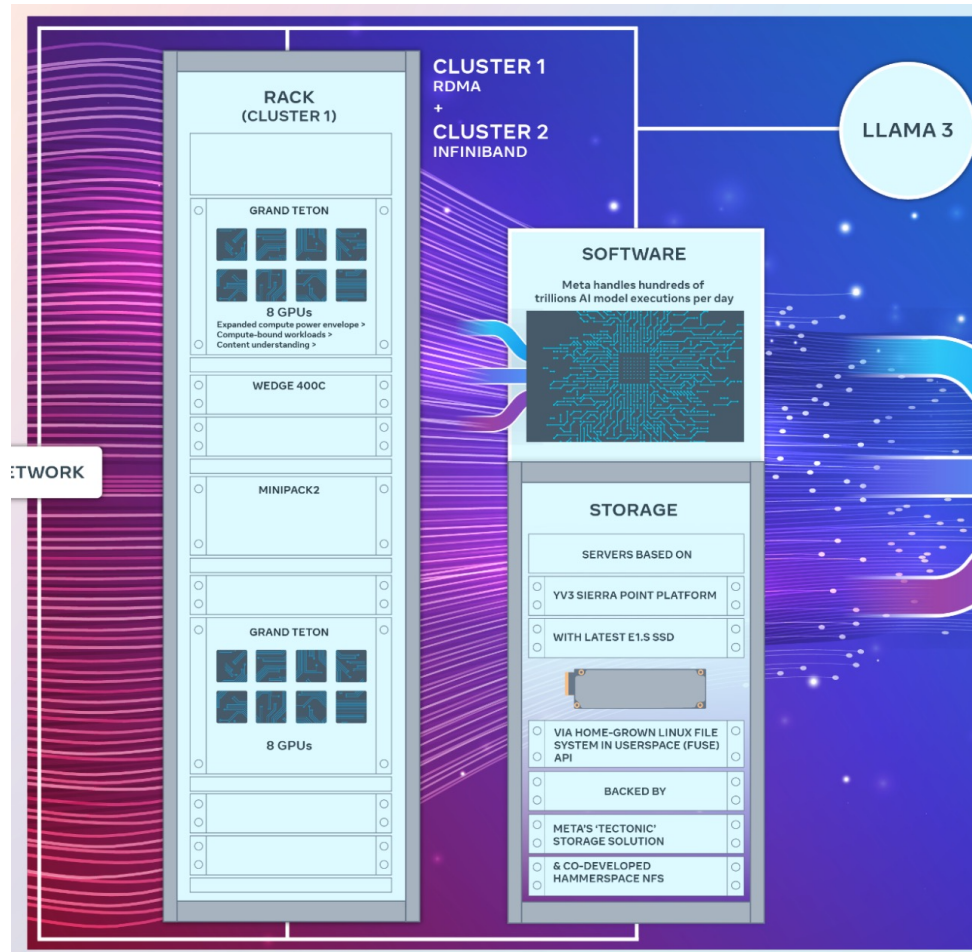
TPUs, FPGAs, ASICs  
ML specialized hardware  
(~2020)

# META AI CLUSTER

Network: RDMA / Infiniband  
(400 Gbps)

Integrate power, control,  
compute, and fabric interfaces  
into a single chassis

Storage: high capacity E1.S SSD



# TRENDS SUMMARY

CPU speed per core is flat

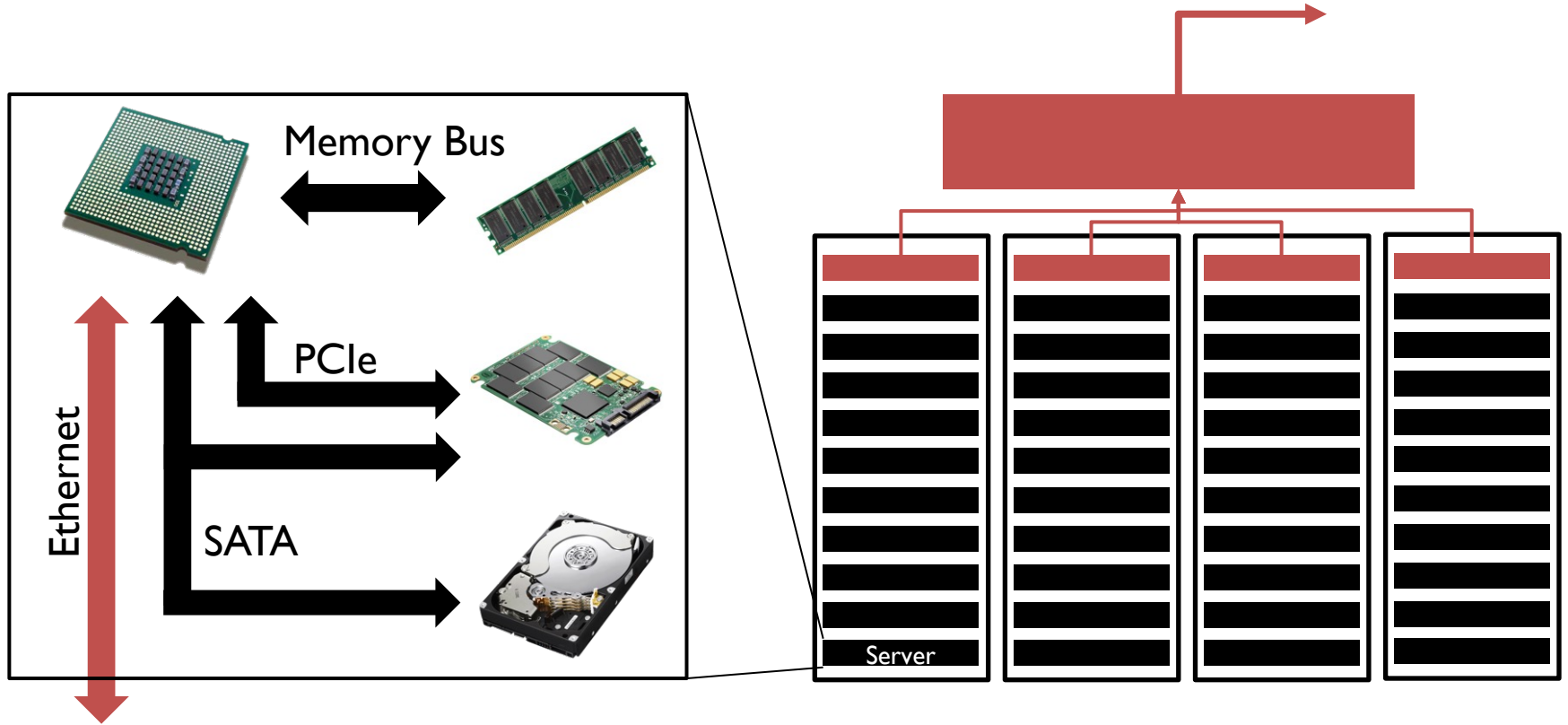
Memory bandwidth growing slower than capacity

SSD, NVMe replacing HDDs

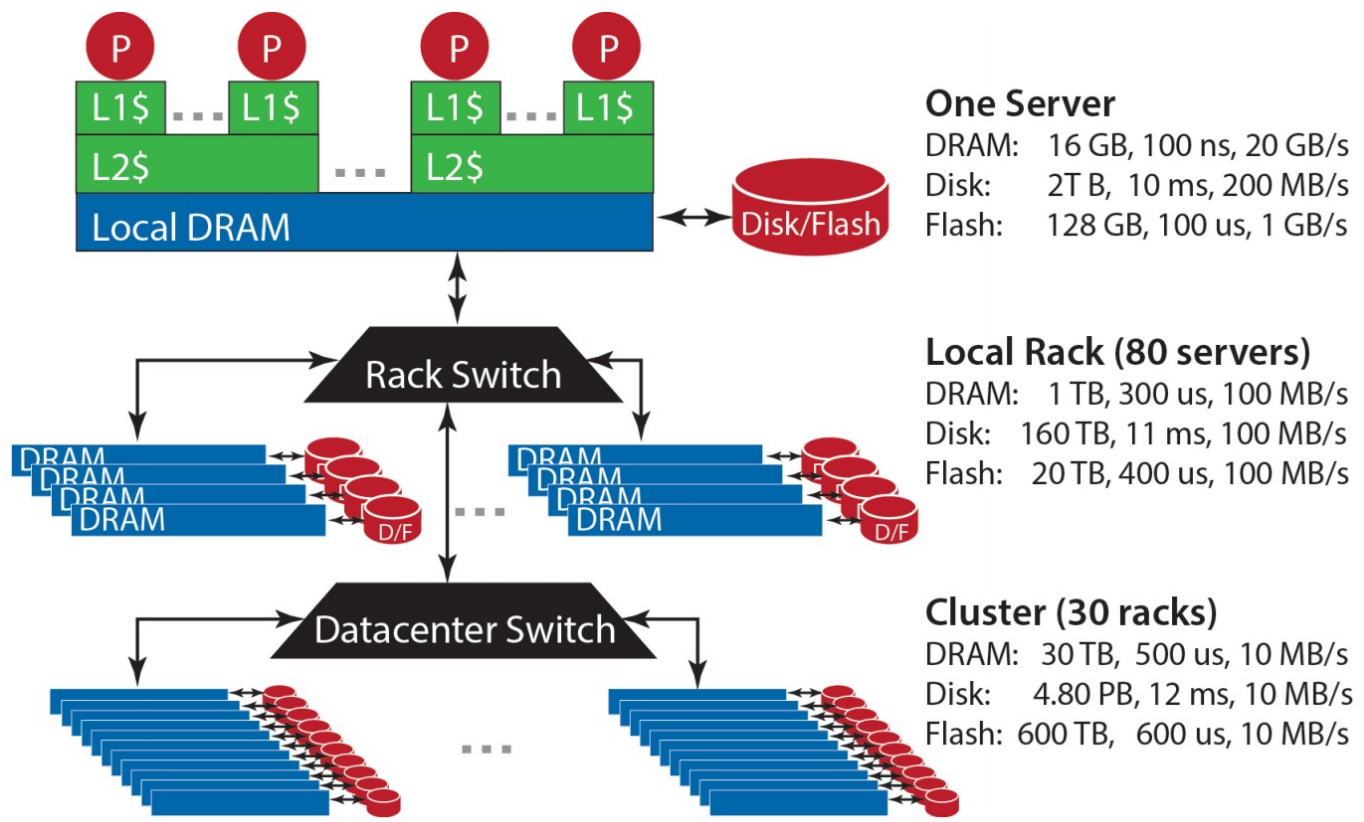
Ethernet bandwidth growing

New accelerators

# SCALE OUT: DATACENTER ARCHITECTURE



# STORAGE HIERARCHY (DC AS A COMPUTER V2)



## 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

## 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

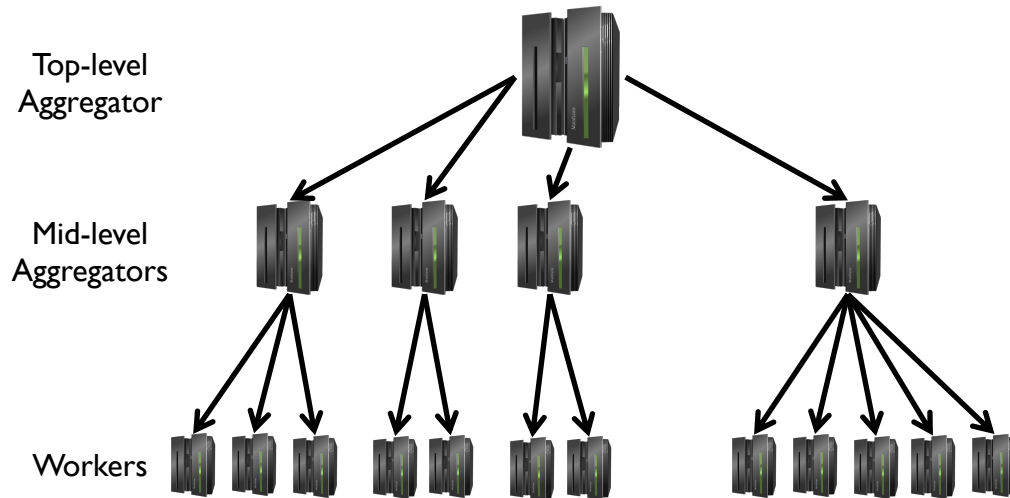
Reliability

Storage Hierarchy

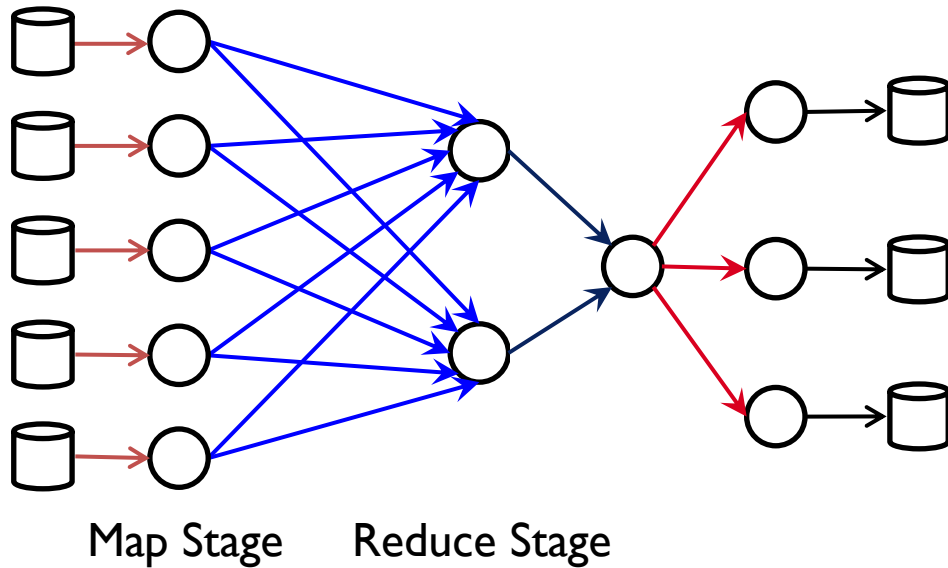
Workload Diversity

Single organization

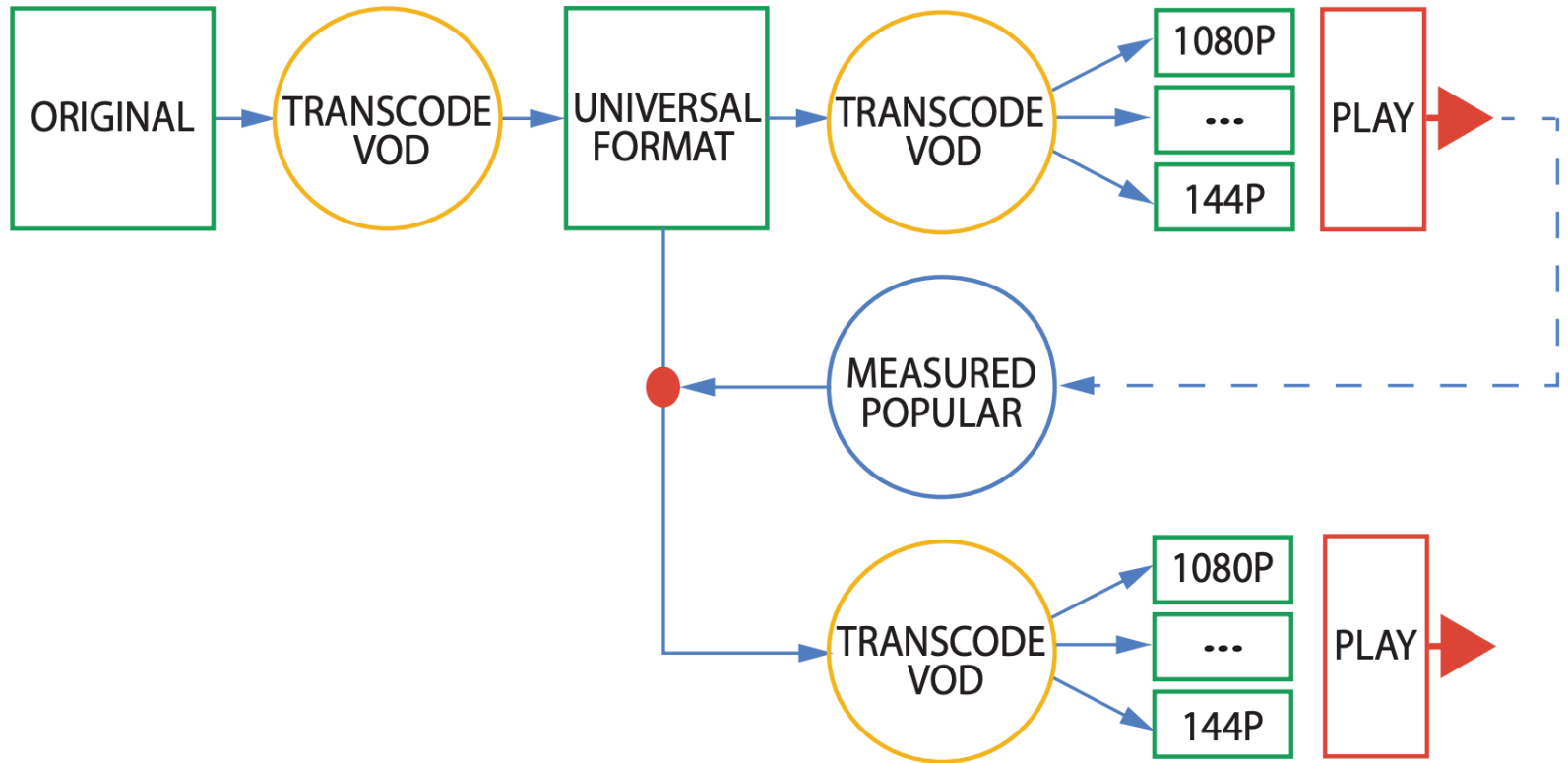
# WORKLOAD: PARTITION-AGGREGATE



# WORKLOAD: SCHOLAR SIMILARITY



# VIDEO ENCODING



# MACHINE LEARNING

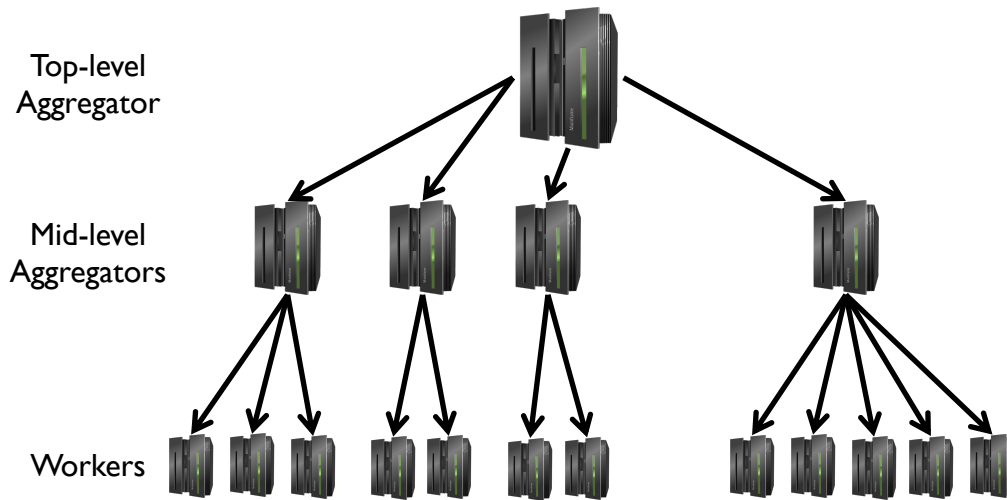
**Table 2.1:** Six production applications plus ResNet benchmark. The fourth column is the total number of operations (not execution rate) that training takes to converge.

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



Scale-up vs Scale-out



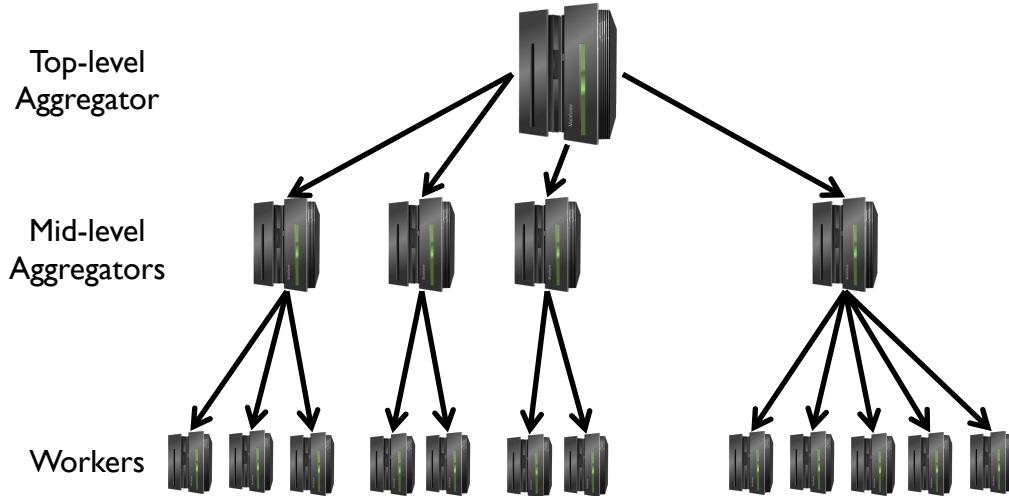
# DISCUSSION

Scale-up vs Scale-out

# DISCUSSION

<https://forms.gle/3AwAz6qSCwneqgLN9>



# NEXT STEPS

Next class: Storage Systems

Assignment 1 out Tuesday.

Submit groups before that!