Good afternoon!

CS 744: DATACENTER AS A COMPUTER

Shivaram Venkataraman
Fall 2022
ANNOUNCEMENTS

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

- Class format
  - Review
  - Lecture
  - Discussion
Scalable Storage Systems

Datacenter Architecture

Resource Management

Computational Engines

Applications

Machine Learning  SQL  Streaming  Graph

Hardware

Properties
OUTLINE

- Hardware Trends
- Datacenter design
- WSC workloads
- Discussion
WHY IS ONE MACHINE NOT ENOUGH?

→ Too slow
  → no parallelism
→ Storage limit
→ Cost of creating a “giant” machine
→ Availability, reliability
→ Compute / network limitations
  → too many users / requests
WHAT'S IN A MACHINE?

Interconnected compute and storage

Newer Hardware
- GPUs, FPGAs
- RDMA, NVlink
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
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

Capacity has grown a lot, but not as fast as the increase in capacity.

- Capacity: 128x increase
- Bandwidth: 20x increase
- Latency: 1.3x increase

In the past 20 years, DRAM capacity has improved significantly, but latency has not kept pace with bandwidth improvements.
MEMORY TAKEAWAY

Data access from memory is getting more expensive!

Graph showing the growth of memory capacity, bandwidth, and latency from 1999 to 2017.
HDD CAPACITY

BackBlaze

Average Cost per Drive Size
By Quarter: Q1 2009 - Q2 2017

HDD Storage is getting cheaper!
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

wear out
SSDs are widely used in data centers.
ETHERNET BANDWIDTH

Growing 33-40% per year!
AMAZON EC2 (2019)

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

SSD standard
TRENDS SUMMARY

CPU speed per core is flat
Memory bandwidth growing slower than capacity
SSD, NVMe replacing HDDs
Ethernet bandwidth growing
SCALE OUT: DATACENTER ARCHITECTURE

- Memory Bus
- Ethernet
- SATA
- PCIe

Internet

20-40 servers

Global switch / cross rack

Racks
Data movement is a primary concern.

Data

STORAGE HIERARCHY (DC AS A COMPUTER V2)

One Server
- DRAM: 16 GB, 100 ns, 20 GB/s
- Disk: 2 TB, 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: 4.8 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

Workload Diversity

Storage Hierarchy

Single organization

⇒ control software stack on every machine
WORKLOAD: PARTITION-AGGREGATE

BigData

latency → sensitive

Busty

Web Index →

Top-level Aggregator

Mid-level Aggregators

Workers

Shard it or partition it
WORKLOAD: SCHOLAR SIMILARITY

- Process
- Large corpus of data
- Not sensitive
- Large latency
- Throughput
- Map Stage
- Reduce Stage

Network traffic
VIDEO ENCODING

- Youtube
- Netflix

- Original -> Transcode VOD
  - Universal Format
  - Transcode VOD
  - Measured Popular
    - 1080P
    - 144P
  - PLAY

- Different clients want different resolutions

- Compute intensive
  - Lot of parallelism from video fragments
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.

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

https://forms.gle/L6deUFjTpfaDDjm6A

Scale-up vs Scale-out

Submission
- Online only!
- Either type responses or attach a picture
- Each student should submit!
- Link on piazza

BigData

Top-level Aggregator

Mid-level Aggregators

Workers
DISCUSSION

Scale-up vs Scale-out

- Scale up
  - extremely low latency
    - avoid network hops
  - Consistency
    - multiple threads
    - network partitions
  - LSTMs / irregular (graphs)
- Scale out
  - incrementally add resources
  - fault tolerance
  - Scale up works better
DISCUSSION

BigData

why are they good
→ work per aggregator is bounded

why are they not good
→ more hops → more latency
Next class: Storage Systems

Assignment 1 out Thursday.
Submit groups before that!