CS295: Modern Systems Warehouse Scale Computers

2019 Spring
Slides are from
CS 152 Computer Architecture and Engineering, by
Dr. George Michelogiannakis
and
CS252 Graduate Computer Architecture, by
Krste Asanovic
Warehouse-scale computers (WSCs)

- Provides Internet services
  - Search, social networking, online maps, video sharing, online shopping, email, cloud computing, etc.

- Differences with high-performance computing (HPC) “clusters”:
  - Clusters have higher performance processors and network
  - Clusters emphasize thread-level parallelism, WSCs emphasize request-level parallelism

- Differences with datacenters:
  - Datacenters consolidate different machines and software into one location
  - Datacenters emphasize virtual machines and hardware heterogeneity in order to serve varied customers
WSC Characteristics

- Ample computational parallelism is not important
  - Most jobs are totally independent
  - “Request-level parallelism”
- Operational costs count
  - Power consumption is a primary, not secondary, constraint when designing system
- Scale and its opportunities and problems
  - Can afford to build customized systems since WSC require volume purchase
- Location counts
  - Real estate, power cost; Internet, end-user, and workforce availability
- Computing efficiently at low utilization
- Scale and the opportunities/problems associated with scale
  - Unique challenges: custom hardware, failures
  - Unique opportunities: bulk discounts
Efficiency and Cost of WSC

- **Location of WSC**
  - Proximity to Internet backbones, electricity cost, property tax rates, low risk from earthquakes, floods, and hurricanes
Figure 6.18 In 2017 AWS had 16 sites (“regions”), with two more opening soon. Most sites have two to three availability zones, which are located nearby but are unlikely to be affected by the same natural disaster or power outage, if one were to occur. (The number of availability zones are listed inside each circle on the map.) These 16 sites or regions collectively have 42 availability zones. Each availability zone has one or more WSCs. https://aws.amazon.com/about-aws/global-infrastructure/.
Figure 6.19 In 2017 Google had 15 sites. In the Americas: Berkeley County, South Carolina; Council Bluffs, Iowa; Douglas County, Georgia; Jackson County, Alabama; Lenoir, North Carolina; Mayes County, Oklahoma; Montgomery County, Tennessee; Quilicura, Chile; and The Dalles, Oregon. In Asia: Changhua County, Taiwan; Singapore. In Europe: Dublin, Ireland; Eemshaven, Netherlands; Hamina, Finland; St. Ghislain, Belgium. 
https://www.google.com/about/datacenters/inside/locations/.
Figure 6.20 In 2017 Microsoft had 34 sites, with four more opening soon. https://azure.microsoft.com/en-us/regions/.
Components

- Apart from computers & network switches, you need:
  - Power infrastructure: voltage converters and regulators, generators and UPSs, ...
  - Cooling infrastructure: A/C, cooling towers, heat exchangers, air impellers, ...

- Everything is co-designed!
Example: MS Quincy Datacenter

- 470k sq feet (10 football fields)
- Next to a hydro-electric generation plant
  - At up to 40 MegaWatts, $0.02/kWh is better than $0.15/kWh 😊
  - That’s equal to the power consumption of 30,000 homes
Example: MS Chicago Datacenter

[K. Vaid, Microsoft Global Foundation Services, 2010]

- $500M+ investment
- 1.5 million man-hours-of-labor
- 3000 construction related jobs
- 3400 tons of steel
- 707,000 sq ft
- 190 miles of conduit
- 60 MW Total Critical Power
- 2400 tons of copper
- 7.5 miles of chilled water piping
- 26,000 cubic yards of concrete
Power Distribution

11% lost in distribution
\[ 0.997 \times 0.94 \times 0.98 \times 0.98 \times 0.99 = 89\% \]

Generators

IT Load (servers, storage, Net, ...)

~1% loss in switch gear & conductors

0.3% loss
99.7% efficient

6% loss
94% efficient, ~97% available

2% loss
98% efficient

2% loss
98% efficient

Sub-station

High Voltage Utility Distribution

13.2kV

13.2kV

13.2kV

13.2kV

480V

208V

28V

UPS: Rotary or Battery

Transformers

© 2019, Elsevier Inc. All rights Reserved
Infrastructure and Costs of WSC

- Cooling system also uses water (evaporation and spills)
  - E.g. 70,000 to 200,000 gallons per day for an 8 MW facility

- Power cost breakdown:
  - Chillers: 30-50% of the power used by the IT equipment
  - Air conditioning: 10-20% of the IT power, mostly due to fans

- How many servers can a WSC support?
  - Each server:
    - “Nameplate power rating” gives maximum power consumption
    - To get actual, measure power under actual workloads
  - Oversubscribe cumulative server power by 40%, but monitor power closely
Infrastructure and Costs of WSC

- Determining the maximum server capacity
  - Nameplate power rating: maximum power that a server can draw
  - Better approach: measure under various workloads
  - Oversubscribe by 40%

- Typical power usage by component:
  - Processors: 42%
  - DRAM: 12%
  - Disks: 14%
  - Networking: 5%
  - Cooling: 15%
  - Power overhead: 8%
  - Miscellaneous: 4%
Power Utilization Effectiveness (PEU)

\[ n = \frac{\text{Total facility power}}{\text{IT equipment power}} \]

**Continuous PUE improvement**
Average PUE for all data centers

![Graph showing PUE improvement over years]

**Figure 6.11** Average power utilization efficiency (PUE) of the 15 Google WSCs between 2008 and 2017. The spiking line is the quarterly average PUE, and the straighter line is the trailing 12-month average PUE. For Q4 2016, the averages were 1.11 and 1.12, respectively.
Performance, Latency

- Latency is important metric because it is seen by users
- Bing study: users will use search less as response time increases
- Service Level Objectives (SLOs)/Service Level Agreements (SLAs)
  - E.g. 99% of requests be below 100 ms

<table>
<thead>
<tr>
<th>Server delay (ms)</th>
<th>Increased time to next click (ms)</th>
<th>Queries/ user</th>
<th>Any clicks/ user</th>
<th>User satisfaction</th>
<th>Revenue/ user</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>200</td>
<td>500</td>
<td>—</td>
<td>—0.3%</td>
<td>—0.4%</td>
<td>—</td>
</tr>
<tr>
<td>500</td>
<td>1200</td>
<td>—</td>
<td>—1.0%</td>
<td>—0.9%</td>
<td>—1.2%</td>
</tr>
<tr>
<td>1000</td>
<td>1900</td>
<td>—0.7%</td>
<td>—1.9%</td>
<td>—1.6%</td>
<td>—2.8%</td>
</tr>
<tr>
<td>2000</td>
<td>3100</td>
<td>—1.8%</td>
<td>—4.4%</td>
<td>—3.8%</td>
<td>—4.3%</td>
</tr>
</tbody>
</table>
Larger clusters \rightarrow more prone to high tail latency
# Outages and Anomalies

<table>
<thead>
<tr>
<th>Approx. number events in 1st year</th>
<th>Cause</th>
<th>Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 or 2</td>
<td>Power utility failures</td>
<td>Lose power to whole WSC; doesn’t bring down WSC if UPS and generators work (generators work about 99% of time).</td>
</tr>
<tr>
<td>4</td>
<td>Cluster upgrades</td>
<td>Planned outage to upgrade infrastructure, many times for evolving networking needs such as recabling, to switch firmware upgrades, and so on. There are about nine planned cluster outages for every unplanned outage.</td>
</tr>
<tr>
<td>1000s</td>
<td>Hard-drive failures</td>
<td>2%-10% annual disk failure rate (Pinheiro et al., 2007)</td>
</tr>
<tr>
<td></td>
<td>Slow disks</td>
<td>Still operate, but run 10× to 20× more slowly</td>
</tr>
<tr>
<td></td>
<td>Bad memories</td>
<td>One uncorrectable DRAM error per year (Schroeder et al., 2009)</td>
</tr>
<tr>
<td></td>
<td>Misconfigured machines</td>
<td>Configuration led to ~30% of service disruptions (Barroso and Hölzle, 2009)</td>
</tr>
<tr>
<td></td>
<td>Flaky machines</td>
<td>1% of servers reboot more than once a week (Barroso and Hölzle, 2009)</td>
</tr>
<tr>
<td>5000</td>
<td>Individual server crashes</td>
<td>Machine reboot; typically takes about 5 min (caused by problems in software or hardware).</td>
</tr>
</tbody>
</table>

Figure 6.1 List of outages and anomalies with the approximate frequencies of occurrences in the first year of a new cluster of 2400 servers. We label what Google calls a cluster an array; see Figure 6.5. Based on Barroso, L.A., 2010. Warehouse Scale Computing [keynote address]. In: Proceedings of ACM SIGMOD, June 8–10, 2010, Indianapolis, IN.
Sources of Outages

Computed from 41 benchmarked data centers

- UPS system failure: 29%
- Accidental/human error: 12%
- Water, heat or CRAC failure: 15%
- Weather related: 10%
- Generator failure: 5%
- IT equipment failure: 5%
- Other: 5%

Figure 6.3 Average CPU utilization of more than 5000 servers during a 6-month period at Google. Servers are rarely completely idle or fully utilized, instead operating most of the time at between 10% and 50% of their maximum utilization. The third column from the right in Figure 6.4 calculates percentages plus or minus 5% to come up with the weightings; thus 1.2% for the 90% row means that 1.2% of servers were between 85% and 95% utilized. From Figure 1 in Barroso, L.A., Hölzle, U., 2007. The case for energy-proportional computing. IEEE Comput. 40 (12), 33–37.
Figure 6.30 A Google rack for its WSC. Its dimensions are about 7 ft high, 4 ft wide, and 2 ft deep (2 m × 1.2 m × 0.5 m). The Top of Rack switches are indeed at the top of this rack. Next comes the power converter that converts from 240 V AC to 48 V DC for the servers in the rack using a bus bar at the back of the rack. Next is the 20 slots (depending on the height of the server) that can be configured for the various types of servers that can be placed in the rack. Up to four servers can be placed per tray. At the bottom of the rack are high-efficiency distributed modular DC uninterruptible power supply (UPS) batteries.
Figure 6.5 Hierarchy of switches in a WSC. Based on Figure 1.1 in Barroso, L.A., Clidaras, J., Hölzle, U., 2013. The datacenter as a computer: an introduction to the design of warehouse-scale machines. Synth. Lect. Comput. Architect. 8 (3), 1–154.
Figure 6.8 A Layer 3 network used to link arrays together and to the Internet (Greenberg et al., 2009). A load balancer monitors how busy a set of servers is and directs traffic to the less loaded ones to try to keep the servers approximately equally utilized. Another option is to use a separate border router to connect the Internet to the data center Layer 3 switches. As we will see in Section 6.6, many modern WSCs have abandoned the conventional layered networking stack of traditional switches.
Array Switch

- Switch that connects an array of racks
  - Array switch should have 10 X the bisection bandwidth of rack switch
  - Cost of $n$-port switch grows as $n^2$
  - Often utilize content addressible memory chips and FPGAs
Figure 6.31 A Clos network has three logical stages containing crossbar switches: ingress, middle, and egress. Each input to the ingress stage can go through any of the middle stages to be routed to any output of the egress stage. In this figure, the middle stages are the \( M \) Spine Blocks, and the ingress and egress stages are in the \( N \) Edge Activation Blocks. Figure 6.22 shows the changes in the Spine Blocks and the Edge Aggregation Blocks over many generations of Clos networks in Google WSCs.
Clos Network

- Circuit-switched with three layers of crossbar switches
  - Ingress, middle, egress stages
- Non-blocking if \( m \geq 2n+1 \)
  - Unused input in ingress can be connected to an unused output in egress without re-arranging existing paths

Each ingress switch has \( n \) inputs and \( m \) outputs
Each egress switch has \( m \) inputs and \( n \) outputs
Each middle switch has \( r \) inputs and outputs
Google Jupiter Clos Network

Figure 6.32 Building blocks of the Jupiter Clos network.
© 2019 Elsevier Inc. All rights reserved.
WSC Memory Hierarchy

- Servers can access DRAM and disks on other servers using a NUMA-style interface

<table>
<thead>
<tr>
<th></th>
<th>Local</th>
<th>Rack</th>
<th>Array</th>
</tr>
</thead>
<tbody>
<tr>
<td>DRAM latency (μs)</td>
<td>0.1</td>
<td>300</td>
<td>500</td>
</tr>
<tr>
<td>Flash latency (μs)</td>
<td>100</td>
<td>400</td>
<td>600</td>
</tr>
<tr>
<td>Disk latency (μs)</td>
<td>10,000</td>
<td>11,000</td>
<td>12,000</td>
</tr>
<tr>
<td>DRAM bandwidth (MB/s)</td>
<td>20,000</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Flash bandwidth (MB/s)</td>
<td>1000</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>Disk bandwidth (MB/s)</td>
<td>200</td>
<td>100</td>
<td>10</td>
</tr>
<tr>
<td>DRAM capacity (GB)</td>
<td>16</td>
<td>1024</td>
<td>31,200</td>
</tr>
<tr>
<td>Flash capacity (GB)</td>
<td>128</td>
<td>20,000</td>
<td>600,000</td>
</tr>
<tr>
<td>Disk capacity (GB)</td>
<td>2000</td>
<td>160,000</td>
<td>4,800,000</td>
</tr>
</tbody>
</table>
Storage options

- Use disks inside the servers, or
- Network attached storage through Infiniband

- WSCs generally rely on local disks
- Google File System (GFS) uses local disks and maintains at least three replicas
Cost of a WSC

- **Capital expenditures (CAPEX)**
  - Cost to build a WSC
  - $9 to 13/watt

- **Operational expenditures (OPEX)**
  - Cost to operate a WSC
Observations

- >50% of cost in buying the hardware
- ~30% costs related to power
- Networking ~10% of overall costs (including cost for servers)
TCO Breakdown (2)