Data Center Fundamentals: The Datacenter as a Computer

George Porter
CSE 124
Feb 10, 2017

Includes material from (1) Barroso, Clidaras, and Hölzle, as well as (2) Evrard (Michigan), used with permission and/or via a Creative Commons License
Announcements

• Readings:
  – Barroso book (chs 1-5)
  – Brewer paper (linked off course schedule)

• Today:
  – Datacenter basics
SDSC Data center tour

- Interest list will be posted soon
- Please sign up if interested
Host Virtualization

- Multiple virtual machines on one physical machine
- Applications run unmodified as on real machine
- VM can migrate from one computer to another
VMM Virtual Switches
The storage hierarchy

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
Latency, bandwidth, and capacity
Performance of flash

Latency (us)
Ops/sec
$/GB
# Hardware comparisons

**TABLE 3.1:** Server hardware configuration, benchmark results, and derived (unofficial) cost-efficiency metrics for a large SMP server and a low-end, PC-class server.

<table>
<thead>
<tr>
<th></th>
<th>HP INTEGRITY SUPERDOME-ITANIUM2</th>
<th>HP PROLIANT ML350 G5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>64 sockets, 128 cores</td>
<td>1 socket, quad-core,</td>
</tr>
<tr>
<td></td>
<td>(dual-threaded), 1.6 GHz</td>
<td>2.66 GHz X5355 CPU,</td>
</tr>
<tr>
<td></td>
<td>Itanium2, 12 MB last-level cache</td>
<td>8 MB last-level cache</td>
</tr>
<tr>
<td>Memory</td>
<td>2,048 GB</td>
<td>24 GB</td>
</tr>
<tr>
<td>Disk storage</td>
<td>320,974 GB, 7,056 drives</td>
<td>3,961 GB, 105 drives</td>
</tr>
<tr>
<td>TPC-C price/performance</td>
<td>$2.93/tpmC</td>
<td>$0.73/tpmC</td>
</tr>
<tr>
<td>price/performance</td>
<td>$1.28/transactions per minute</td>
<td>$0.10/transactions</td>
</tr>
<tr>
<td>(server HW only)</td>
<td></td>
<td>per minute</td>
</tr>
<tr>
<td>Price/performance</td>
<td>$2.39/transactions per minute</td>
<td>$0.12/transactions</td>
</tr>
<tr>
<td>(server HW only)</td>
<td></td>
<td>per minute</td>
</tr>
<tr>
<td>(no discounts)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Big computer vs. lots-of-small-computers

**FIGURE 3.2:** Performance advantage of a cluster built with high-end server nodes (128-core SMP) over a cluster with the same number of processor cores built with low-end server nodes (four-core SMP), for clusters of varying size.

Barroso and Holzle, *The Data Center as a Computer: An introduction to the Design of Warehouse-Scale Machines*
Big computer vs. lots-of-small-computers

FIGURE

over a cluster with the same number of processor cores built with low-end server nodes (four core SMP), for clusters of varying size.

Barroso and Holzle, The Data Center as a Computer: An introduction to the Design of Warehouse-Scale Machines
Heat management

CRAC unit
rack
rack
rack
rack
CRAC unit

Liquid supply

floor tiles

Regents of the University of Michigan
Quantifying energy-efficiency: PUE

- PUE = Power Usage Effectiveness
- Simply compares
  - Power used for computing
  - Total power used

- Historically cooling was a huge source of power
  - E.g., 1 watt of computing meant 1 Watt of cooling!

\[
PUE = \frac{\text{Facility Power}}{\text{Computing Equipment power}}
\]

\[
\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \left(\frac{1}{\text{PUE}}\right) \times \left(\frac{1}{\text{SPUE}}\right) \times \left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)
\]

EQUATION 5.1: Breaking an energy efficiency metric into three components: a facility term (a), a server energy conversion term (b), and the efficiency of the electronic components in performing the computation per se (c).
FIGURE 5.1: LBNL survey of the power usage efficiency of 24 datacenters, 2007 (Greenberg et al.) [41].
LBNL PUE Survey (2013)

Table 1: Summary of power/energy data

<table>
<thead>
<tr>
<th>Data center IT devices (W/device)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume server</td>
<td>235</td>
</tr>
<tr>
<td>Midrange server</td>
<td>450</td>
</tr>
<tr>
<td>External HDD spindle</td>
<td>26</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data center PUE</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Server closet</td>
<td>2.5</td>
</tr>
<tr>
<td>Server room</td>
<td>2.1</td>
</tr>
<tr>
<td>Localized</td>
<td>2</td>
</tr>
<tr>
<td>Mid-tier</td>
<td>2</td>
</tr>
<tr>
<td>Enterprise-class</td>
<td>1.5</td>
</tr>
<tr>
<td>Cloud</td>
<td>1.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Network data transmission (μJ/bit)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Wired</td>
<td>100</td>
</tr>
<tr>
<td>Wi-Fi</td>
<td>100</td>
</tr>
<tr>
<td>Cellular (3G/4G)</td>
<td>450</td>
</tr>
</tbody>
</table>
Breakdown of data center overheads

**FIGURE 5.2:** Breakdown of datacenter energy overheads (ASHRAE).

Barroso and Holzle, *The Data Center as a Computer: An introduction to the Design of Warehouse-Scale Machines*
Numbers from James Hamilton (MSFT, Amazon)
Limits of lowering PUE?

- Google’s “Chiller-less” data center in Belgium
- Most of the year it is cool enough to not need cooling
- What about on hot days?
  - Shed load to other data centers!
Power-proportional Computing
Power-proportional humans

**FIGURE 5.7:** Human energy usage vs. activity levels (adult male) [52].
Web-service load fluctuations

**FIGURE 2.2:** Example of daily traffic fluctuation for a search service in one datacenter; $x$-axis is a 24-h period and the $y$-axis is traffic measured in queries per second.
Do different components scale similarly?

**FIGURE 5.8:** Subsystem power usage in an x86 server as the compute load varies from idle to full usage.

Barroso and Holzle, *The Data Center as a Computer: An introduction to the Design of Warehouse-Scale Machines*
Improving efficiency

* datacenter-level (PUE)
  – use passive cooling (NCSA Blue Waters, UC Berkeley CRTF)
  – avoid UPS devices (lower Tier level)
  – avoid AC/DC conversion
  – use green power (Icelandic model)

* rack-level (SPUE)
  – avoid local power supplies
  – more efficient voltage regulation

* processor/process-level
  – speed-variable processors
  – embedded, low-power processors (memristors?)
  – smarter parallelization / distribution of work
FIGURE 5.5: Activity profile of a sample of 5,000 Google servers over a period of 6 months.
What about “power saving” features on modern computers?
FIGURE 5.6: Power and corresponding power efficiency of three hypothetical systems: a typical server with idle power at 50% of peak (Pwr50 and Eff50), a more energy-proportional server with idle power at 10% of peak (Pwr10 and Eff10), and a sublinearly energy-proportional server with idle power at 10% of peak.
Power supply efficiency and SPUE

FIGURE 5.9: A survey of conversion efficiency of several server power supplies (by Ton and Fortenbury [84]).
Oversubscribing data center power?

**FIGURE 5.10:** Cumulative distribution of the time that groups of machines spend at or below a given power level (power level is normalized to the maximum peak aggregate power for the corresponding grouping) (Fan et al. [27]).
Decreasing role of voltage scaling

**FIGURE 5.11:** Power vs. compute load for an x86 server at three voltage-frequency levels and the corresponding energy savings.
The connection between data center networks and energy
Traditional DC Topology

- **Core**
  - Layer-3 router

- **Aggregation**
  - Layer-2/3 switch

- **Access**
  - Layer-2 switch
  - Servers

- **Internet**

Data Center
Layer 2 Pods w/L3 Backbone

- **Internet**
- **DC-Layer 3**
- **DC-Layer 2**

Key:
- CR = Core Router (L3)
- AR = Access Router (L3)
- S = Ethernet Switch (L2)
- A = Rack of app. servers

~ 1,000 servers/pod == IP subnet
Capacity Bottlenecks

• Discussion: Implications for energy efficiency?

• Recall: 

\[
\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \left( \frac{1}{\text{PUE}} \right) \times \left( \frac{1}{\text{SPUE}} \right) \times \left( \frac{\text{Computation}}{\text{Total Energy to Electronic Components}} \right)
\]
Tree-based network topologies

Can’t buy sufficiently fast core switches!

100,000 x 10 Gb/s = 1 Pb/s
Folded-Clos multi-rooted Trees

With k-port switches, can support $k^3/4$ hosts

Al Fares, et al., Sigcomm’08

10 Gb/s
Switches

10 Gb/s
servers
Multi-rooted Trees

• Benefits:
  – No more bandwidth bottlenecks
  – All switches can be same speed
    • Don’t need expensive/fast switches at root

• Downsides:
  – Bandwidth achieved only if traffic evenly spread across all the possible paths
  – Network itself uses quite a bit of power
    • But how much?