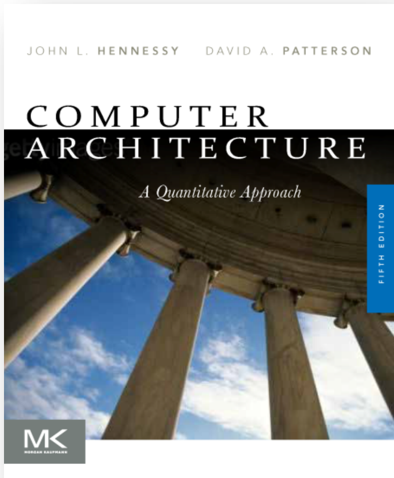


Chapter 6

Warehouse-Scale Computers to Exploit Request-Level and Data-Level Parallelism:

Introduction

- Warehouse-scale computer (WSC)
  - Provides Internet services
    - Search, social networking, online maps, video sharing, online shopping, email, cloud computing, etc.
  - Differences with 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

Introduction

- Important design factors shared with servers:
  - Cost-performance
    - Small savings add up
  - Energy efficiency
    - Affects power distribution and cooling
    - Work per joule
  - Dependability via redundancy
  - Network I/O
  - Interactive and batch processing workloads

Introduction

- Important design factors not shared with servers:
  - 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

Introduction

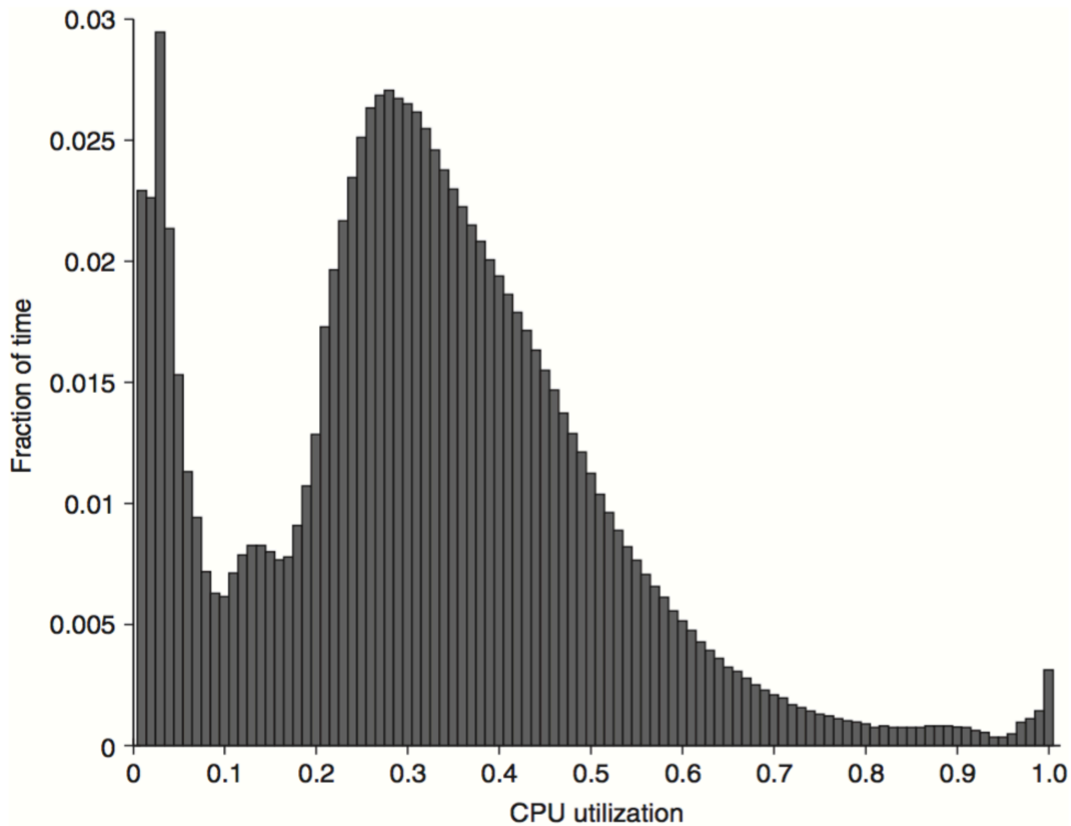


Figure 6.3 Average CPU utilization of more than 5000 servers during a 6-month period at Google. Servers are rarely completely idle or fully utilized, instead operating most of the time at between 10% and 50% of their maximum utilization. (From Figure 1 in Barroso and Hölzle [2007].) The column the third 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.

Example p434

Example Calculate the availability of a service running on the 2400 servers in Figure 6.1. Unlike a service in a real WSC, in this example the service cannot tolerate hardware or software failures. Assume that the time to reboot software is 5 minutes and the time to repair hardware is 1 hour.

Approx. number events in 1st year	Cause	Consequence
1 or 2	Power utility failures	Lose power to whole WSC; doesn't bring down WSC if UPS and generators work (generators work about 99% of time).
4	Cluster upgrades	Planned outage to upgrade infrastructure, many times for evolving networking needs such as recabling, to switch firmware upgrades, and so on. There are about 9 planned cluster outages for every unplanned outage.
1000s	Hard-drive failures	2% to 10% annual disk failure rate [Pinheiro 2007]
	Slow disks	Still operate, but run 10x to 20x more slowly
	Bad memories	One uncorrectable DRAM error per year [Schroeder et al. 2009]
	Misconfigured machines	Configuration led to ~30% of service disruptions [Barroso and Hölzle 2009]
	Flaky machines	1% of servers reboot more than once a week [Barroso and Hölzle 2009]
5000	Individual server crashes	Machine reboot, usually takes about 5 minutes

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 [2010].)

Example p434

Answer We can estimate service availability by calculating the time of outages due to failures of each component. We'll conservatively take the lowest number in each category in Figure 6.1 and split the 1000 outages evenly between four components. We ignore slow disks—the fifth component of the 1000 outages—since they hurt performance but not availability, and power utility failures, since the uninterruptible power supply (UPS) system hides 99% of them.

$$\text{Hours Outage}_{\text{service}} = (4 + 250 + 250 + 250) \times 1 \text{ hour} + (250 + 5000) \times 5 \text{ minutes}$$
$$= 754 + 438 = 1192 \text{ hours}$$

Since there are 365×24 or 8760 hours in a year, availability is:

$$\text{Availability}_{\text{system}} = \frac{(8760 - 1192)}{8760} = \frac{7568}{8760} = 86\%$$

That is, without software redundancy to mask the many outages, a service on those 2400 servers would be down on average one day a week, or *zero* nines of availability!

Prgrm'g Models and Workloads

- Batch processing framework: MapReduce
  - **Map:** applies a programmer-supplied function to each logical input record
    - Runs on thousands of computers
    - Provides new set of key-value pairs as intermediate values
  - **Reduce:** collapses values using another programmer-supplied function

Prgrm’g Models and Workloads

- Example:
  - **map (String key, String value)**:
    - **// key: document name**
    - **// value: document contents**
    - **for each word w in value**
      - **EmitIntermediate(w,”1”); // Produce list of all words**
  - **reduce (String key, Iterator values):**
    - **// key: a word**
    - **// value: a list of counts**
    - **int result = 0;**
    - **for each v in values:**
      - **result += ParseInt(v); // get integer from key-value pair**
    - **Emit(AsString(result));**

Prgrm'g Models and Workloads

- MapReduce runtime environment schedules map and reduce task to WSC nodes

- Availability:
  - Use replicas of data across different servers
  - Use relaxed consistency:
    - No need for all replicas to always agree

- Workload demands
  - Often vary considerably

Computer Architecture of WSC

- WSC often use a hierarchy of networks for interconnection
- Each 19” rack holds 48 1U servers connected to a rack switch
- Rack switches are uplinked to switch higher in hierarchy
  - Uplink has 48 / n times lower bandwidth, where n = # of uplink ports
    - “Oversubscription”
  - Goal is to maximize locality of communication relative to the rack

Computer Architecture of WSC

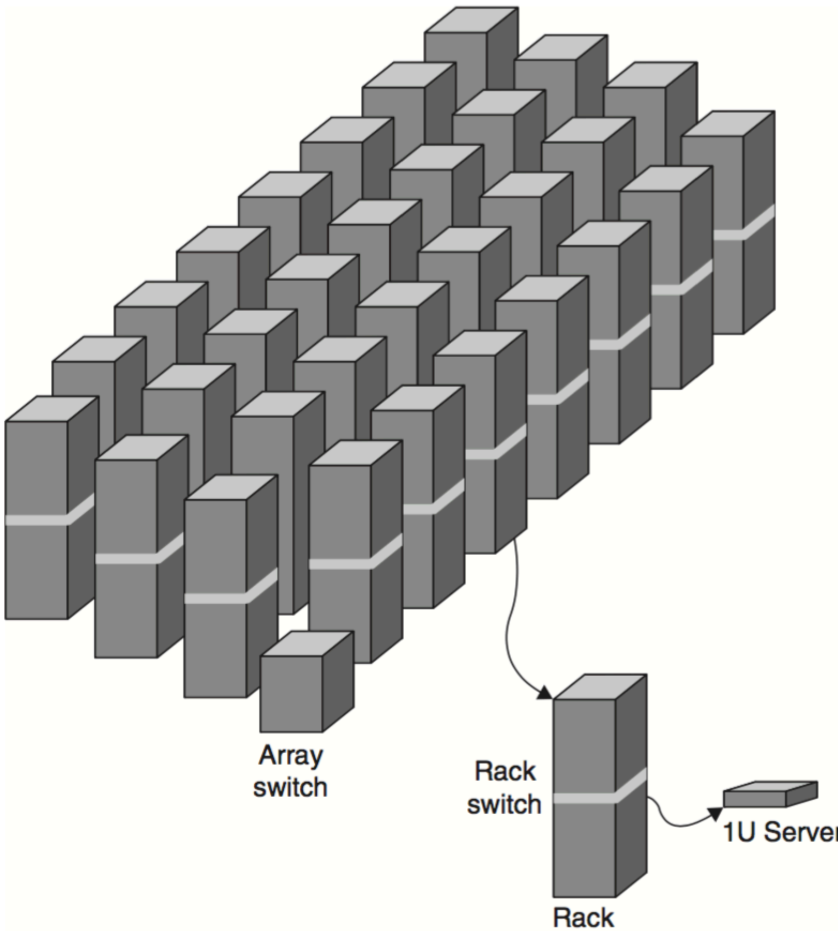


Figure 6.5 Hierarchy of switches in a WSC. (Based on Figure 1.2 of Barroso and Hölzle [2009].)

Storage

- 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 relicas

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

Computer Architecture of WSC

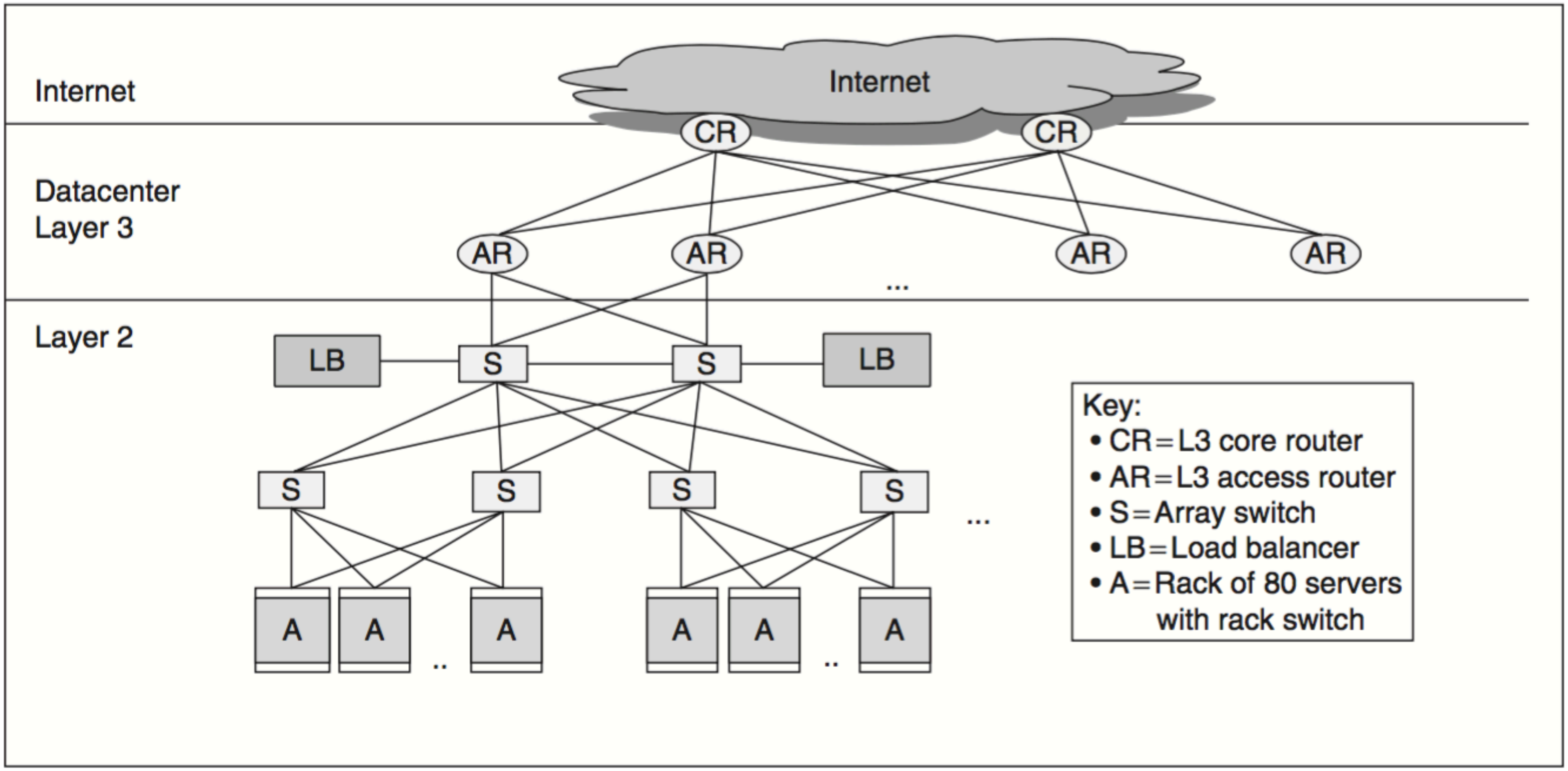


Figure 6.8 The Layer 3 network used to link arrays together and to the Internet [Greenberg et al. 2009]. Some WSCs use a separate *border router* to connect the Internet to the datacenter Layer 3 switches.

WSC Memory Hierarchy

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

	Local	Rack	Array
DRAM latency (microseconds)	0.1	100	300
Disk latency (microseconds)	10,000	11,000	12,000
DRAM bandwidth (MB/sec)	20,000	100	10
Disk bandwidth (MB/sec)	200	100	10
DRAM capacity (GB)	16	1,040	31,200
Disk capacity (GB)	2000	160,000	4,800,000

WSC Memory Hierarchy

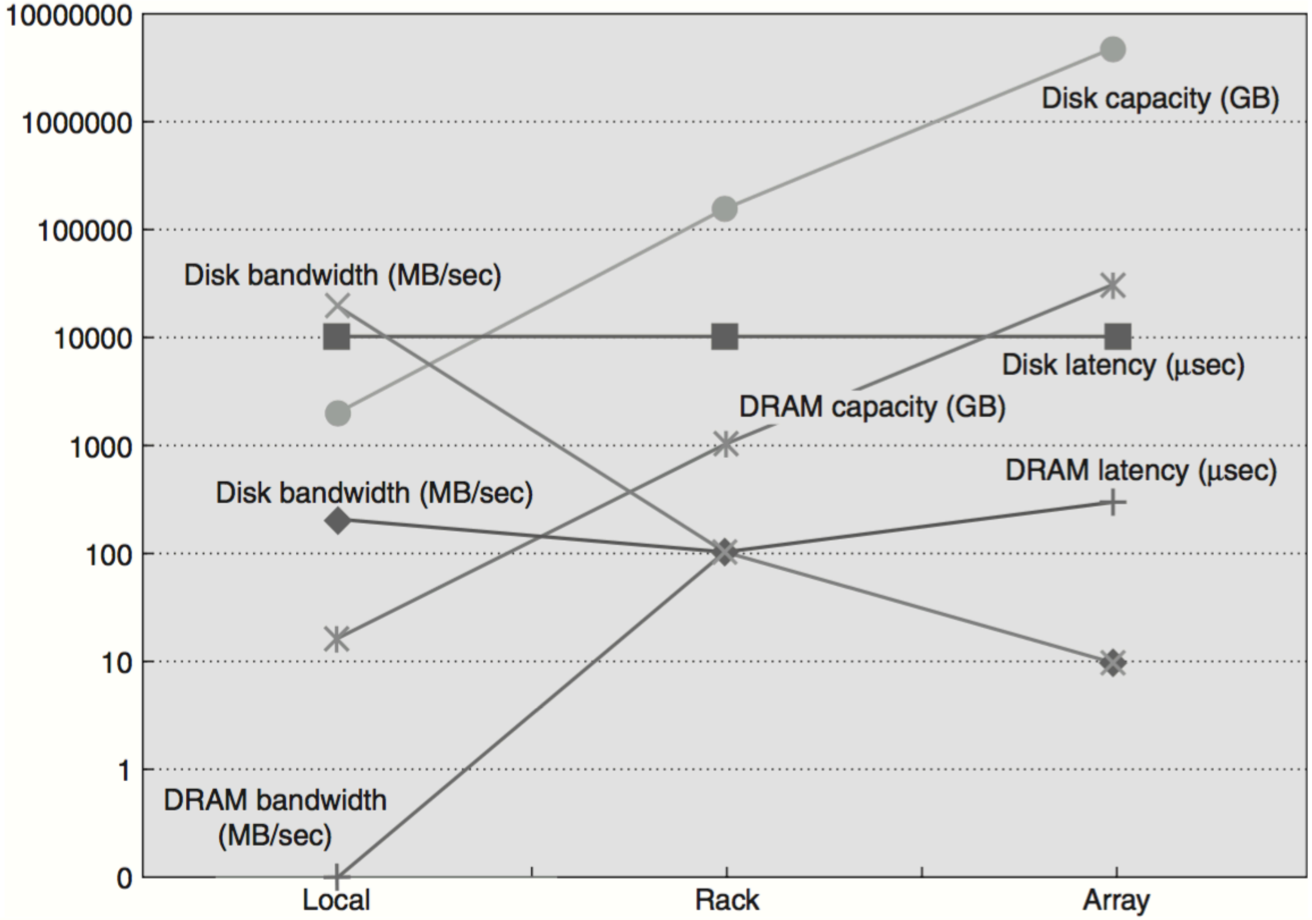


Figure 6.7 **Graph of latency, bandwidth, and capacity of the memory hierarchy of a WSC for data in Figure 6.6 [Barroso and Hölzle 2009].**

WSC Memory Hierarchy

Example What is the average memory latency assuming that 90% of accesses are local to the server, 9% are outside the server but within the rack, and 1% are outside the rack but within the array?

Answer The average memory access time is

$$(90\% \times 0.1) + (9\% \times 100) + (1\% \times 300) = 0.09 + 9 + 3 = 12.09 \text{ microseconds}$$

or a factor of more than 120 slowdown versus 100% local accesses. Clearly, locality of access within a server is vital for WSC performance.

WSC Memory Hierarchy

Example How long does it take to transfer 1000 MB between disks within the server, between servers in the rack, and between servers in different racks in the array? How much faster is it to transfer 1000 MB between DRAM in the three cases?

Answer A 1000 MB transfer between disks takes:

Within server = 1000/200 = 5 seconds

Within rack = 1000/100 = 10 seconds

Within array = 1000/10 = 100 seconds

A memory-to-memory block transfer takes

Within server = 1000/20000 = 0.05 seconds

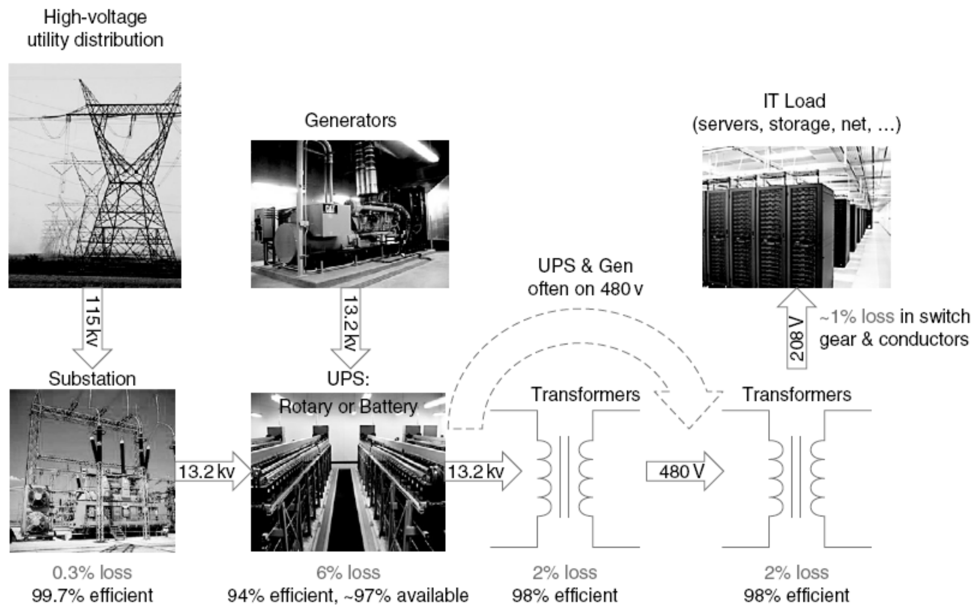
Within rack = 1000/100 = 10 seconds

Within array = 1000/10 = 100 seconds

Thus, for block transfers outside a single server, it doesn't even matter whether the data are in memory or on disk since the rack switch and array switch are the bottlenecks. These performance limits affect the design of WSC software and inspire the need for higher performance switches (see Section 6.6).

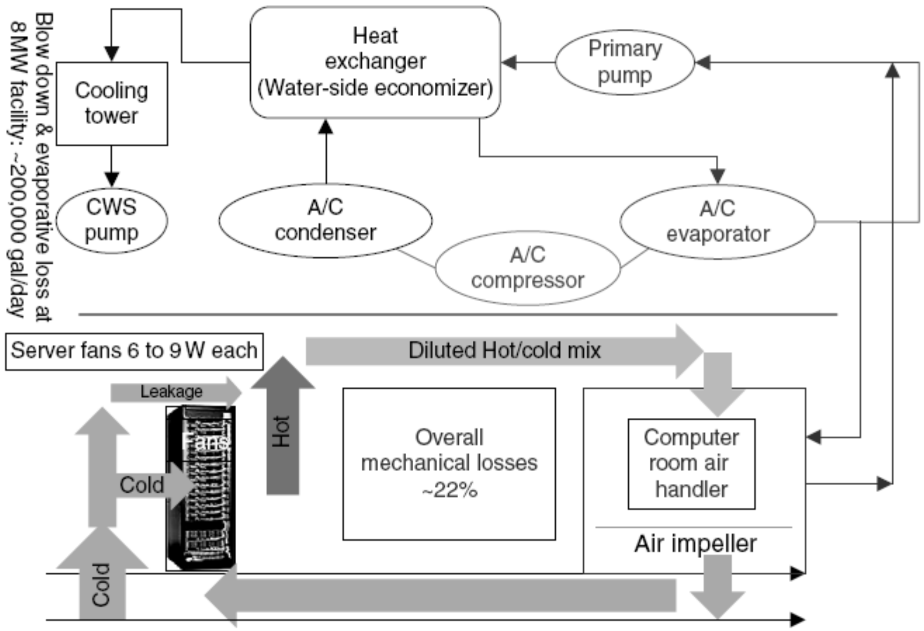
Infrastructure and Costs of WSC

- **Location of WSC**
  - **Proximity to Internet backbones, electricity cost, property tax rates, low risk from earthquakes, floods, and hurricanes**
- **Power distribution**



Infrastructure and Costs of WSC

- **Cooling**
  - **Air conditioning used to cool server room**
  - **64 F – 71 F**
    - **Keep temperature higher (closer to 71 F)**
  - **Cooling towers can also be used**
    - **Minimum temperature is "wet bulb temperature"**



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

Breaking down power usage inside the IT equipment itself, Barroso and Hölzle [2009] reported the following for a Google WSC deployed in 2007:

- 33% of power for processors
- 30% for DRAM
- 10% for disks
- 5% for networking
- 22% for other reasons (inside the server)

Measuring Efficiency of a WSC

- Power Utilization Effectiveness (PUE)
  - = Total facility power / IT equipment power
  - Median PUE on 2006 study was 1.69
- Performance
  - Latency is important metric because it is seen by users

Power utilization effectiveness

- Power Utilization Effectiveness (PUE)

  $$\text{PUE} = \frac{\text{Total facility power}}{\text{IT equipment power}}$$

- PUE
  - always >1
  - ideal =1

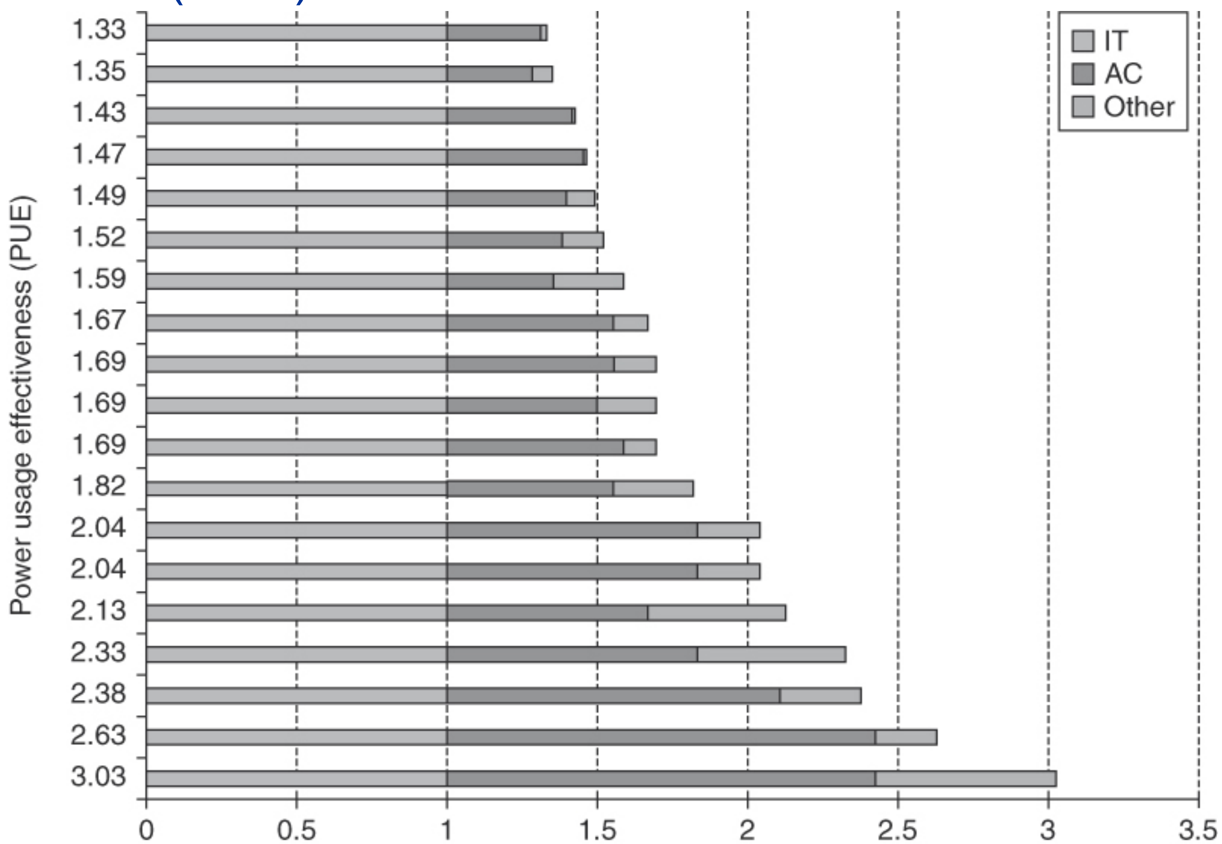


Figure 6.11 Power utilization efficiency of 19 datacenters in 2006 [Greenberg et al. 2006]. The power for air conditioning (AC) and other uses (such as power distribution) is normalized to the power for the IT equipment in calculating the PUE. Thus, power for IT equipment must be 1.0 and AC varies from about 0.30 to 1.40 times the power of the IT equipment. Power for "other" varies from about 0.05 to 0.60 of the IT equipment. Median = 1.69

Power utilization effectiveness

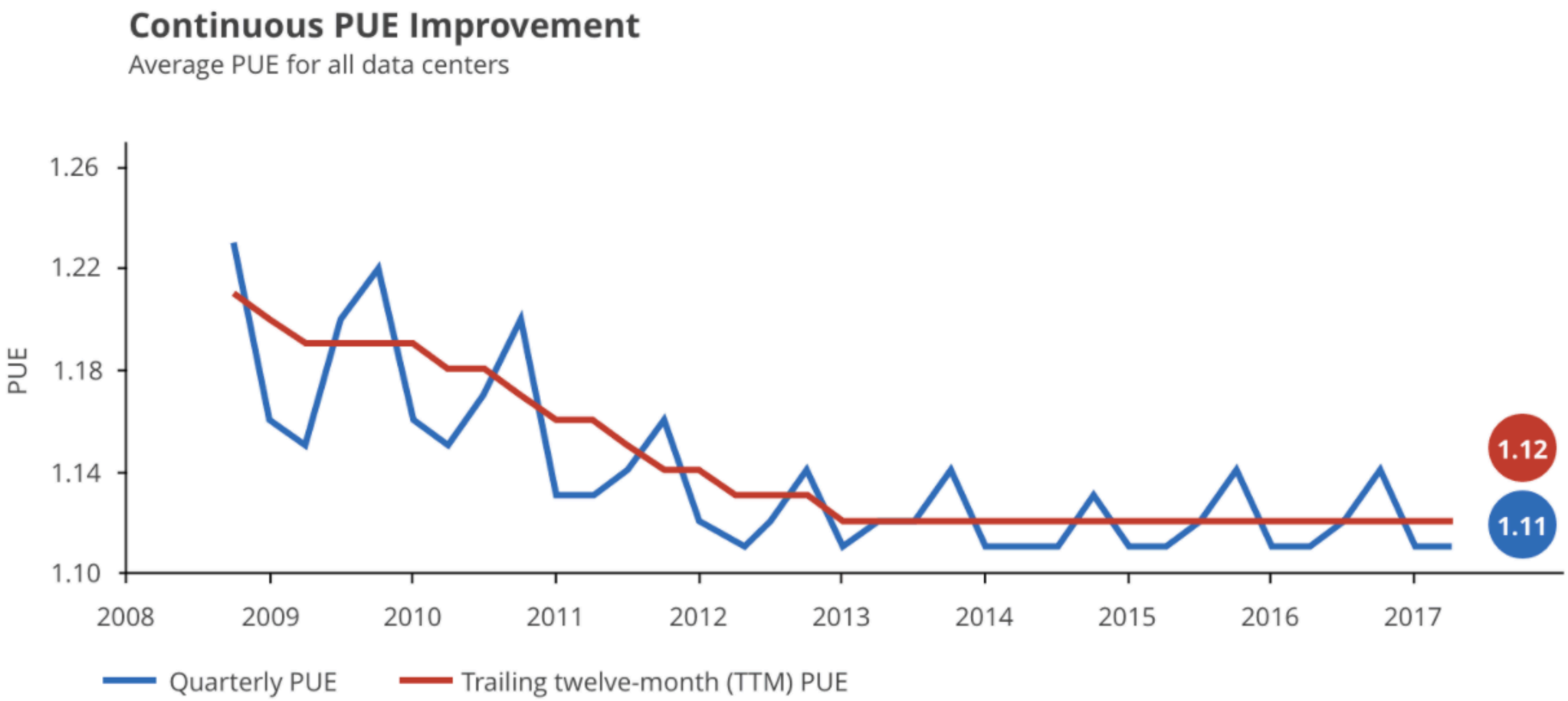


Figure 2: PUE data for all large-scale Google data centers

Source: https://www.google.com/about/datacenters/efficiency/internal/

Measuring Efficiency of a WSC

- Performance
  - 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

Measuring Efficiency of a WSC

- **Performance: Bing Study**

Server delay (ms)	Increased time to next click (ms)	Queries/ user	Any clicks/ user	User satisfaction	Revenue/ user
50	--	--	--	--	--
200	500	--	–0.3%	–0.4%	--
500	1200	--	–1.0%	–0.9%	–1.2%
1000	1900	–0.7%	–1.9%	–1.6%	–2.8%
2000	3100	–1.8%	–4.4%	–3.8%	–4.3%

Figure 6.12 Negative impact of delays at Bing search server on user behavior Schurman and Brutlag [2009].

Cost of a WSC

- Capital expenditures (CAPEX)
  - Cost to build a WSC
- Operational expenditures (OPEX)
  - Cost to operate a WSC

Cost of a WSC

Size of facility (critical load watts)	8,000,000
Average power usage (%)	80%
Power usage effectiveness	1.45
Cost of power ($/kwh)	$0.07
% Power and cooling infrastructure (% of total facility cost)	82%
CAPEX for facility (not including IT equipment)	**$88,000,000**
Number of servers	45,978
Cost/server	$1450
CAPEX for servers	**$66,700,000**
Number of rack switches	1150
Cost/rack switch	$4800
Number of array switches	22
Cost/array switch	$300,000
Number of layer 3 switches	2
Cost/layer 3 switch	$500,000
Number of border routers	2
Cost/border router	$144,800
CAPEX for networking gear	**$12,810,000**
Total CAPEX for WSC	**$167,510,000**
Server amortization time	3 years
Networking amortization time	4 years
Facilities amortization time	10 years
Annual cost of money	5%

Figure 6.13 Case study for a WSC, based on Hamilton [2010], rounded to nearest $5000. Internet bandwidth costs vary by application, so they are not included here. The remaining 18% of the CAPEX for the facility includes buying the property and the cost of construction of the building. We added people costs for security and facilities management in Figure 6.14, which were not part of the case study. Note that Hamilton's estimates were done before he joined Amazon, and they are not based on the WSC of a particular company.

Cloud Computing

- WSCs offer economies of scale that cannot be achieved with a datacenter:
  - 5.7 times reduction in storage costs
  - 7.1 times reduction in administrative costs
  - 7.3 times reduction in networking costs
  - This has given rise to cloud services such as Amazon Web Services
    - “Utility Computing”
    - Based on using open source virtual machine and operating system software