计算机组织与系统结构

仓储级计算系统

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

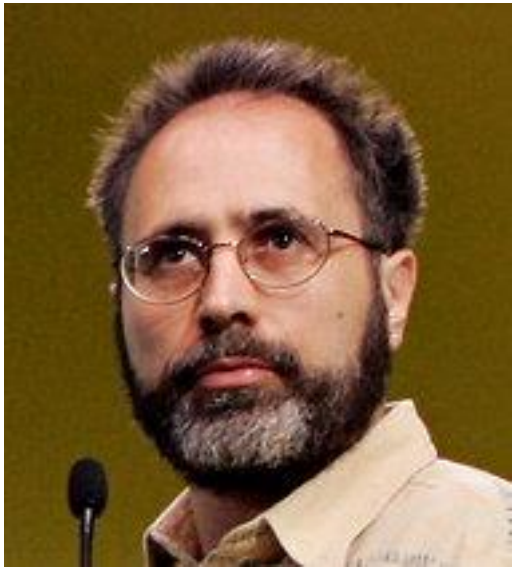
（第十六讲）

程 旭

2021.1.7

2011

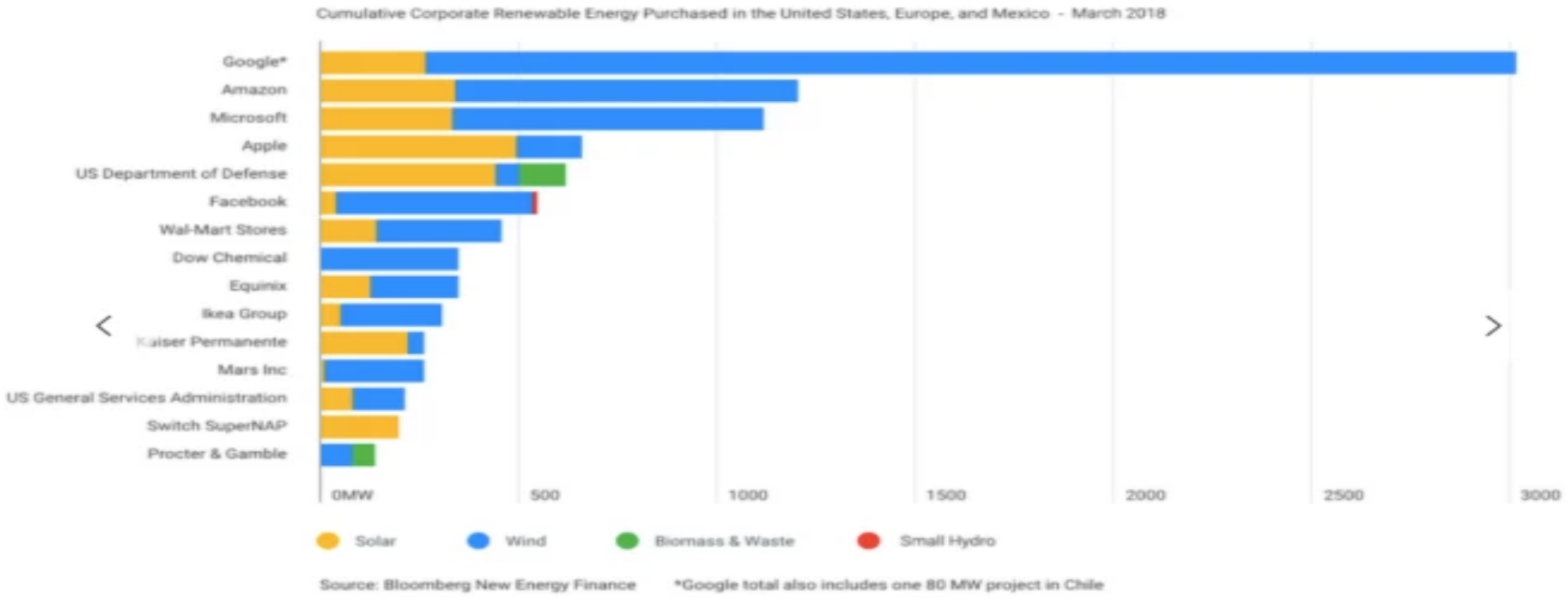
- Google disclosed that it continuously uses enough electricity to power 200,000 homes, but it says that in doing so, it also makes the planet greener.
- Search cost per day (per person) same as running a 60-watt bulb for 3 hours



Urs Hölzle, Google SVP

www.nytimes.com/2011/09/09/technology/google-details-and-defends-its-use-of-electricity.html

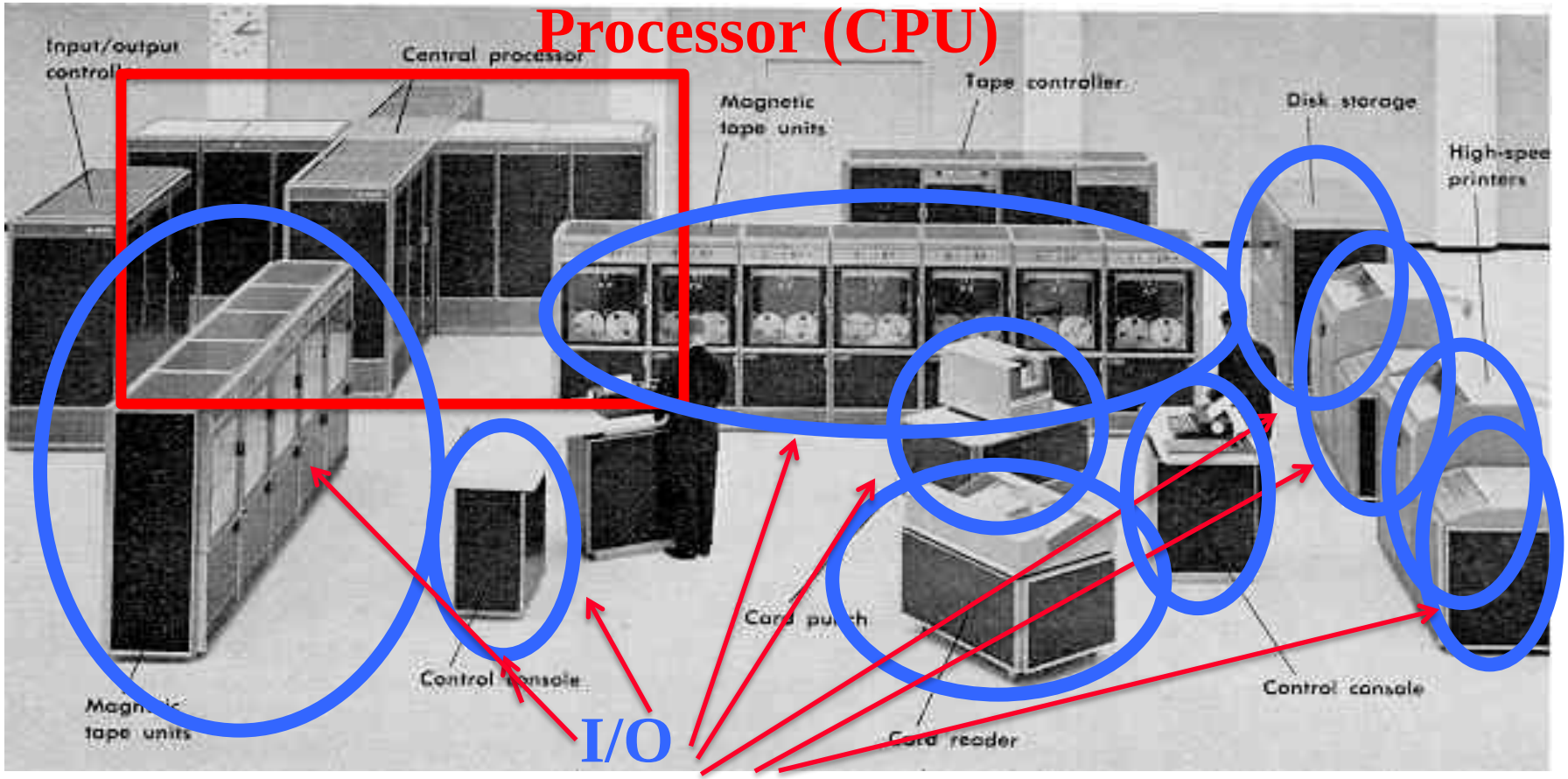
2018



techcrunch.com/2018/04/04/google-matches-100-percent-of-its-power-consumption-with-renewables/

- **Google: “Over the course of 2017, across the globe, for every kilowatt-hour of electricity we consumed, we purchased a kilowatt-hour of renewable energy from a wind or solar farm that was built specifically for Google. This makes us the first public Cloud, and company of our size, to have achieved this feat”**

Computer Eras: Mainframe 1950s-6



"Big Iron": IBM, UNIVAC, … build $1M computers for businesses ➔ COBOL, Fortran, timesharing OS

Minicomputer Eras: 1970s

Using integrated circuits, Digital, HP… build $10k computers for labs, universities ➔ C, UNIX OS

PC Era: Mid 1980s - Mid 2000s

Using microprocessors, Apple, IBM, … build $1k computer for 1 person ➔ Basic, Java, Windows OS

PostPC Era: Late 2000s - ??

Personal Mobile Devices (PMD): Relying on wireless networking, Apple, Huawei, … build $500 smartphone and tablet computers for individuals
➔ Objective C, Java, Android OS + iOS

Cloud Computing: Using Local Area Networks, BAT，Amazon, Google, … build $200M **Warehouse Scale Computers** with 100,000 servers for Internet Services for PMDs

➔ MapReduce, Ruby on Rails

Why Cloud Computing Now?

- "The Web Space Race": Build-out of extremely large datacenters (10,000's of *commodity* PCs)
  - Build-out driven by growth in demand (more users)
  - ⇒ Infrastructure software and Operational expertise
- Discovered economy of scale: 5-7x cheaper than provisioning a medium-sized (1000 servers) facility
- More pervasive broadband Internet so can access remote computers efficiently
- Commoditization of HW & SW
  - Standardized software stacks

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

Introduction

- Important design factors for WSC:
  - 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

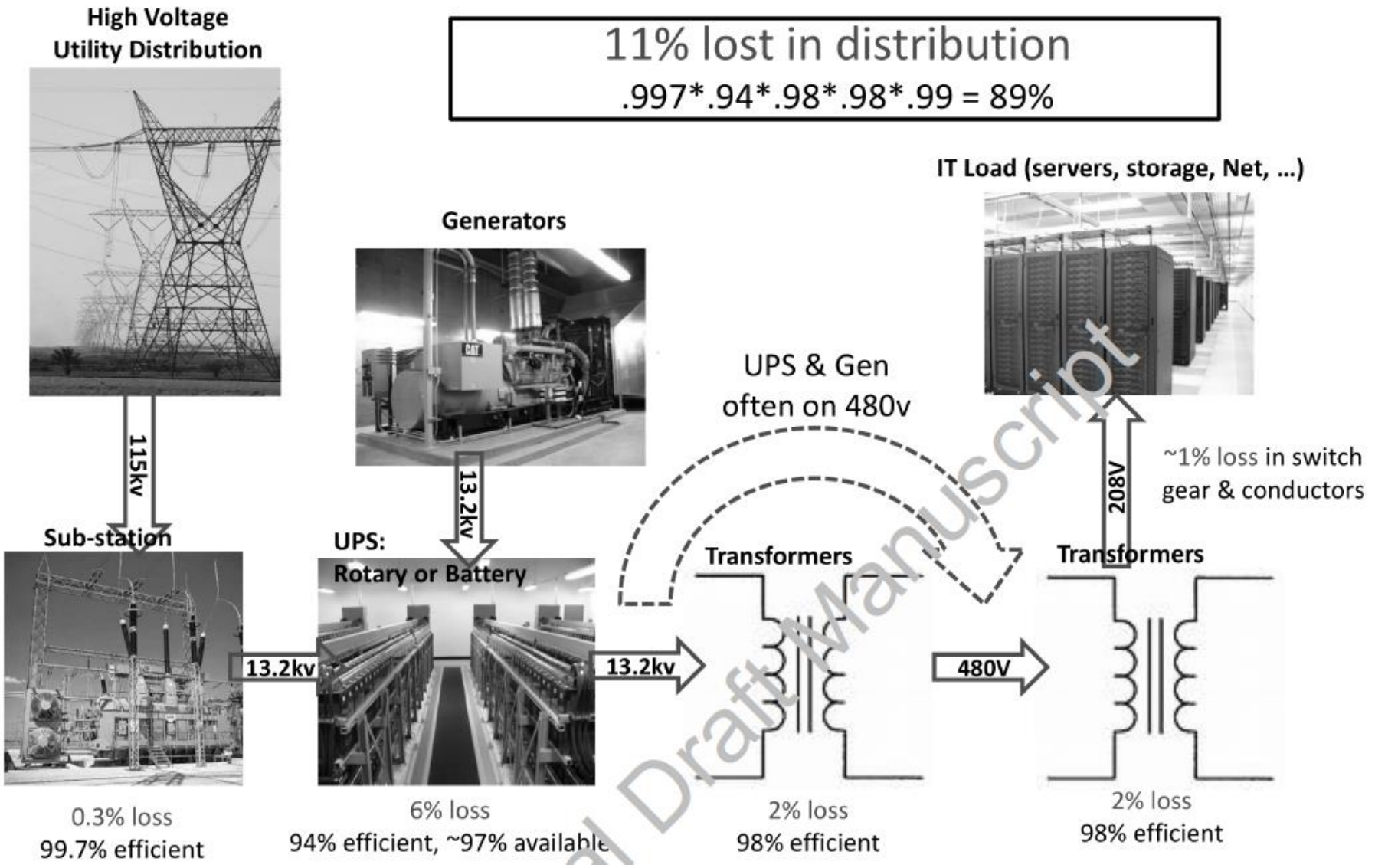
WSC Characteristics

- 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

Power Distribution



Amazon Sites

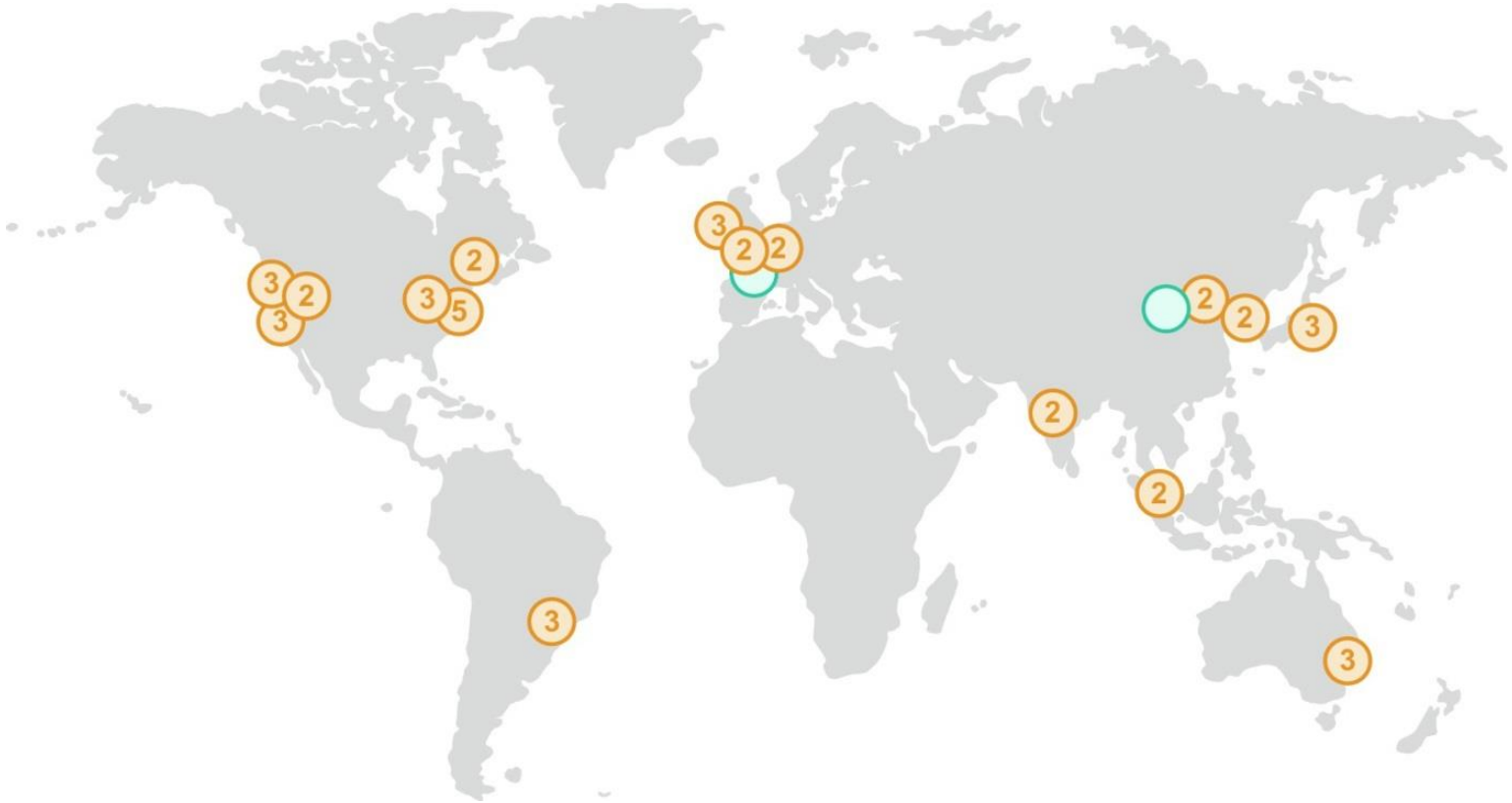


Figure 6.18 In 2017 AWS had 16 sites (“regions”), with two more opening soon. Most sites have two to three *availability zones*, which are located nearby but are unlikely to be affected by the same natural disaster or power outage, if one were to occur. (The number of availability zones are listed inside 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/.

Google Sites

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/.

Microsoft Sites

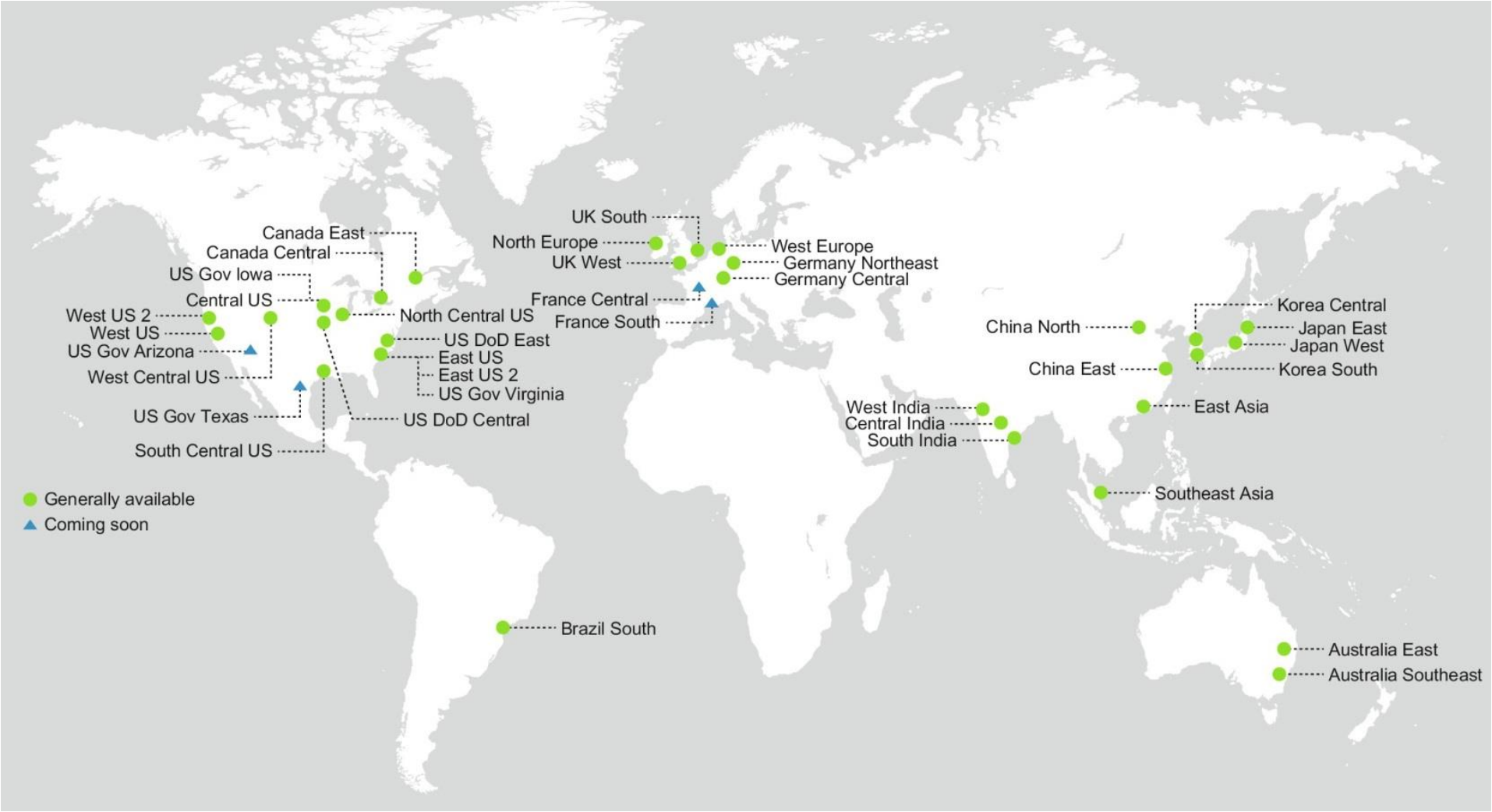


Figure 6.20 In 2017 Microsoft had 34 sites, with four more opening soon. https://azure.microsoft.com/en-us/regions/.

Outages and Anomalies

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 nine planned cluster outages for every unplanned outage.
1000s	Hard-drive failures	2%–10% annual disk failure rate (Pinheiro et al., 2007)
	Slow disks	Still operate, but run 10× to 20× 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; typically takes about 5 min (caused by problems in software or hardware).

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.

Google’s Data Center（2001）

Google’s Oregon WSC (2014)

Programing 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

Programing 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));

Programing Models and Workloads

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

- **File systems: GFS and Colossus**

- **Databases: Dynamo and BigTable**

Programing Models and Workloads

- MapReduce runtime environment schedules map and reduce task to WSC nodes
  - Workload demands often vary considerably
  - Scheduler assigns tasks based on completion of prior tasks
  - Tail latency/execution time variability: single slow task can hold up large MapReduce job
  - Runtime libraries replicate tasks near end of job

CPU Utilization is Usually Low

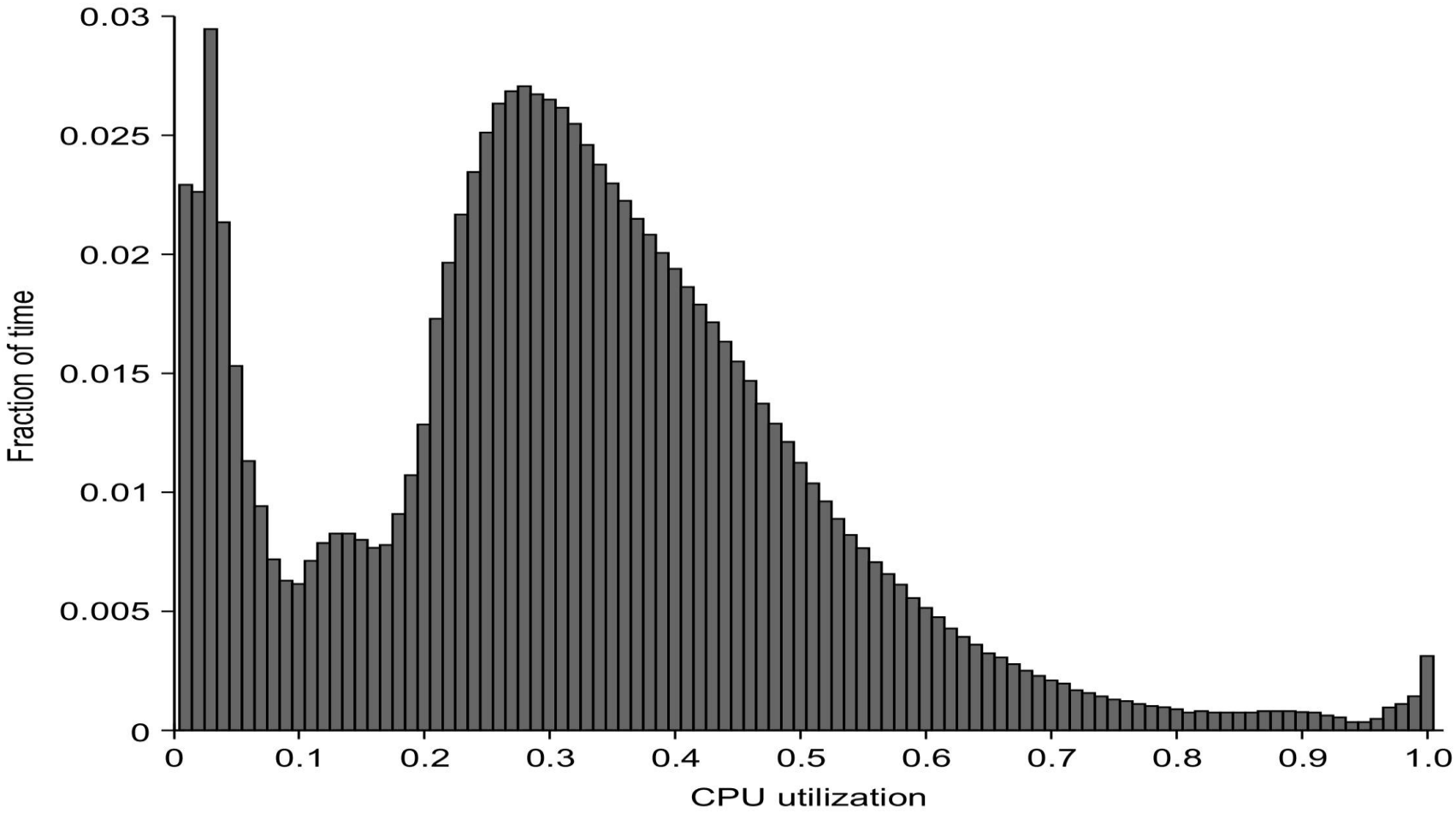


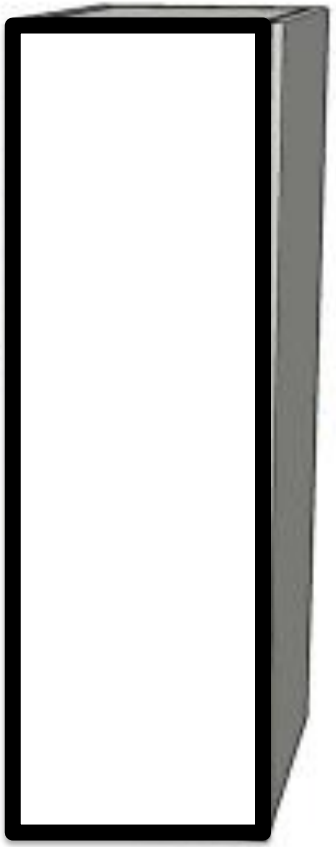
Figure 6.3 Average CPU utilization of more than 5000 servers during a 6-month period at Google. Servers are rarely completely idle or fully utilized, instead operating most of the time at between 10% and 50% of their maximum utilization. The third column from the right in Figure 6.4 calculates percentages 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.

Computer Architecture of WSC

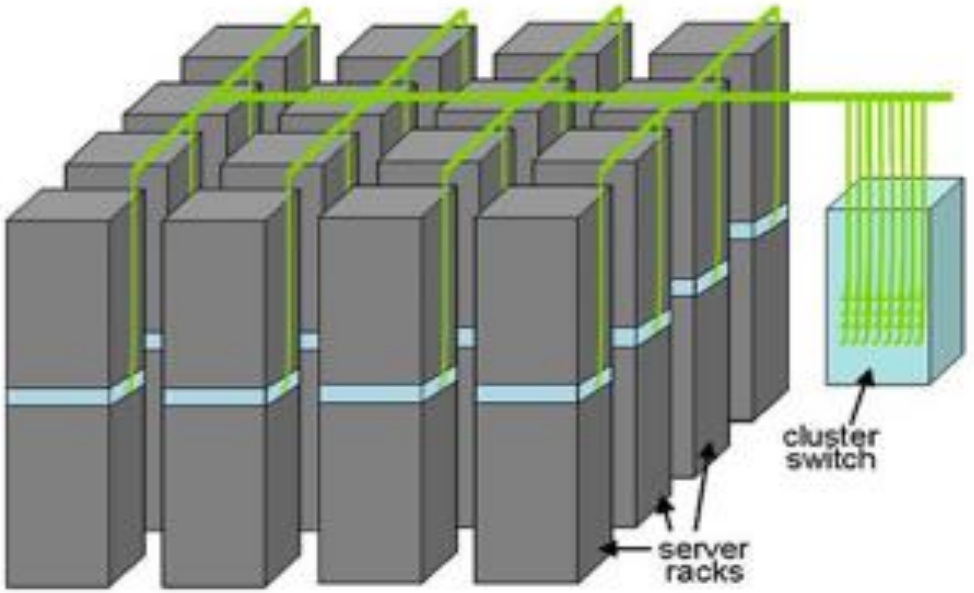
- WSC often use a hierarchy of networks for interconnection
- Each 19” rack holds 48 1U servers connected to a rack switch
- Rack switches are uplinked to switch higher in hierarchy
  - Uplink has 6-24X times lower bandwidthGoal is to maximize locality of communication relative to the rack

Equipment Inside a WSC

Server (in rack format): 1 ¾ inches high "1U", x 19 inches x 16-20 inches: 8 cores, 16 GB DRAM, 4x1 TB disk



7 foot Rack: 40-80 servers + Ethernet local area network (1-10 Gbps) switch in middle ("rack switch")



Array (aka cluster): 16-32 server racks + larger local area network switch ("array switch") 10X faster ➔ cost 100X: cost f(N^2)

Google WSC Rack

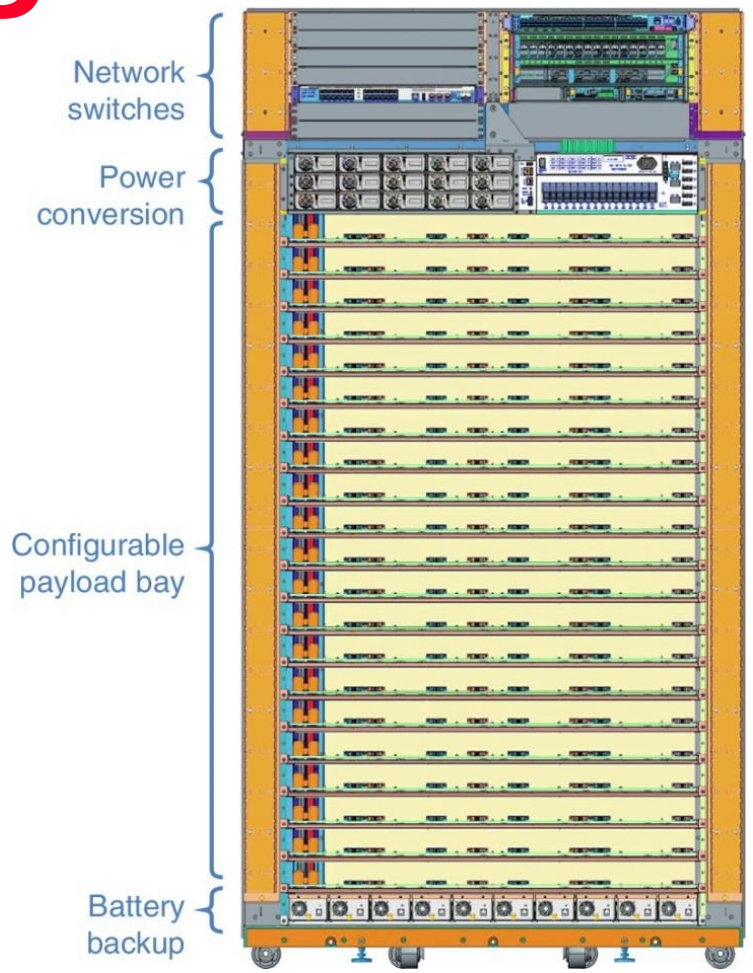


Figure 6.30 A Google rack for its WSC. Its dimensions are about 7 ft high, 4 ft wide, and 2 ft deep (2 m × 1.2 m × 0.5 m). The Top of Rack switches are indeed at the top of this rack. Next comes the power converter that converts from 240 V AC to 48 V DC for the servers in the rack using a bus bar at 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.

Array of Racks

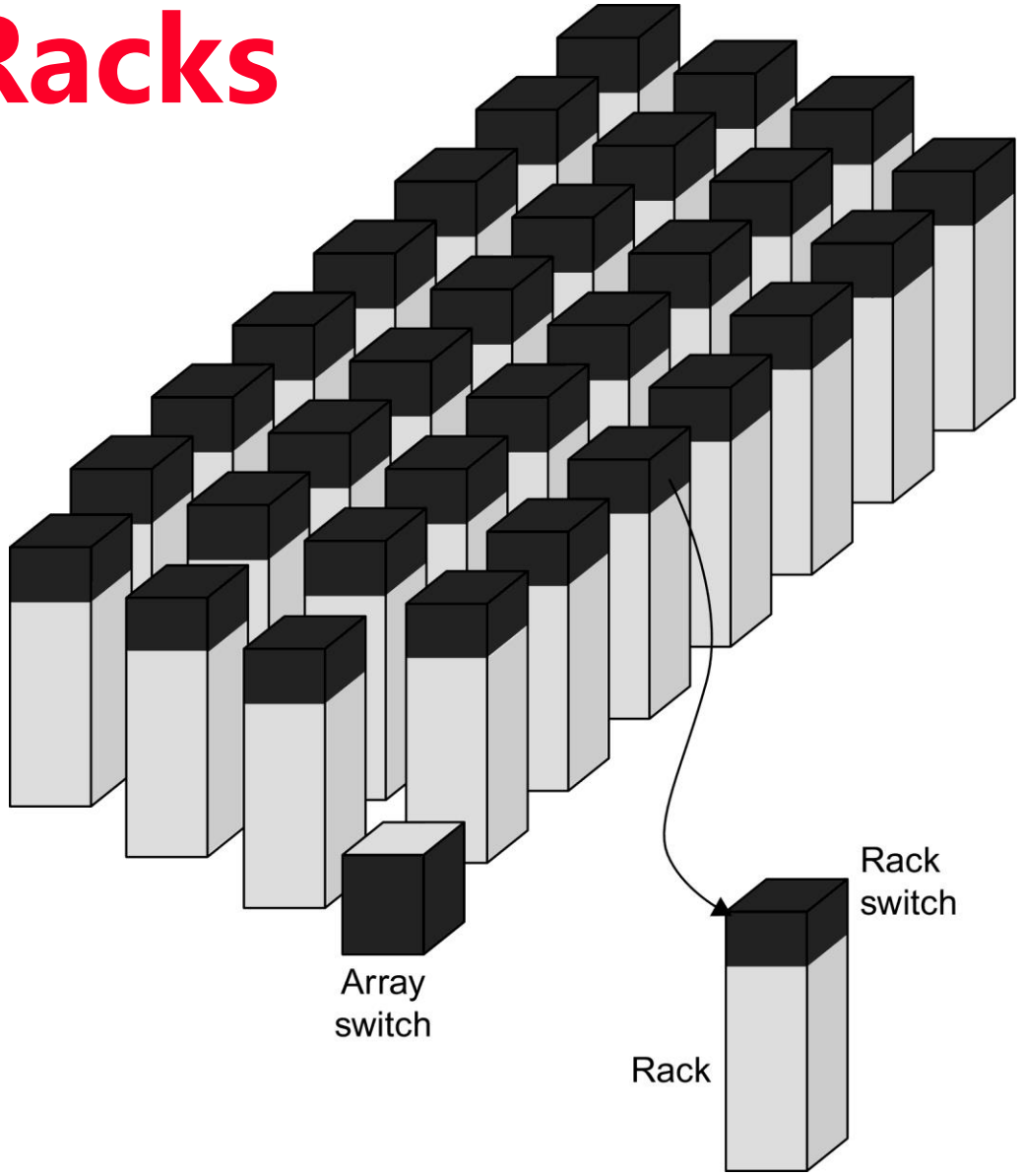


Figure 6.5 Hierarchy of switches in a WSC. Based on Figure 1.1 in Barroso, L.A., Clidaras, J., Hölzle, U., 2013. The datacenter as a computer: an introduction to the design of warehouse-scale machines. Synth. Lect. Comput. Architect. 8 (3), 1–154.

WSC Memory Hierarchy

- **Servers can access DRAM and disks on other servers using a NUMA-style interface**
  - Lower latency to DRAM in another server than local disk
  - Higher bandwidth to local disk than to DRAM in another server

	Local	Rack	Array
DRAM latency (μs)	0.1	300	500
Flash latency (μs)	100	400	600
Disk latency (μs)	10,000	11,000	12,000
DRAM bandwidth (MB/s)	20,000	100	10
Flash bandwidth (MB/s)	1000	100	10
Disk bandwidth (MB/s)	200	100	10
DRAM capacity (GB)	16	1024	31,200
Flash capacity (GB)	128	20,000	600,000
Disk capacity (GB)	2000	160,000	4,800,000

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 (PUE)

- = Total facility power / IT equipment power

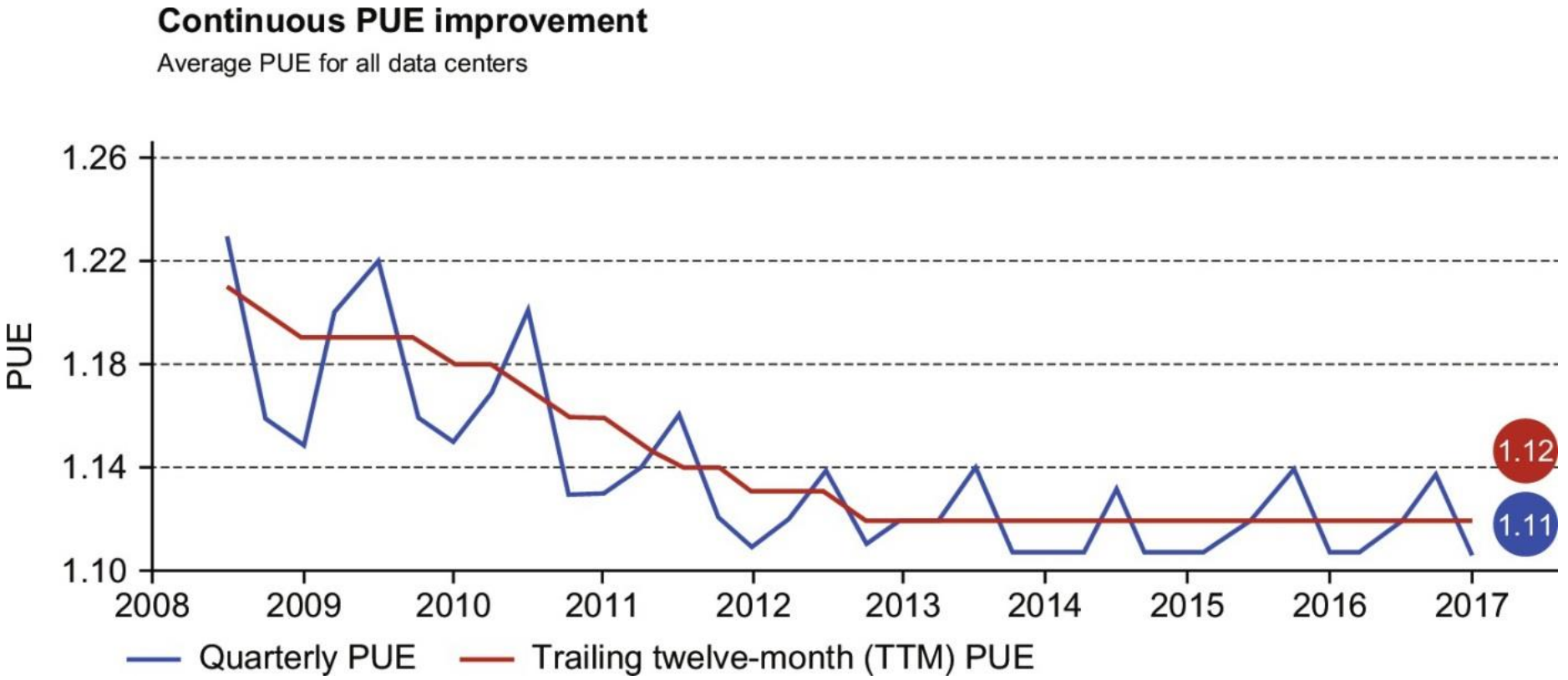


Figure 6.11 Average power utilization efficiency (PUE) of the 15 Google WSCs between 2008 and 2017. The spiking line is the quarterly average PUE, and the straighter line is the trailing 12-month average PUE. For Q4 2016, the averages were 1.11 and 1.12, respectively.

Power Usage Effectiveness

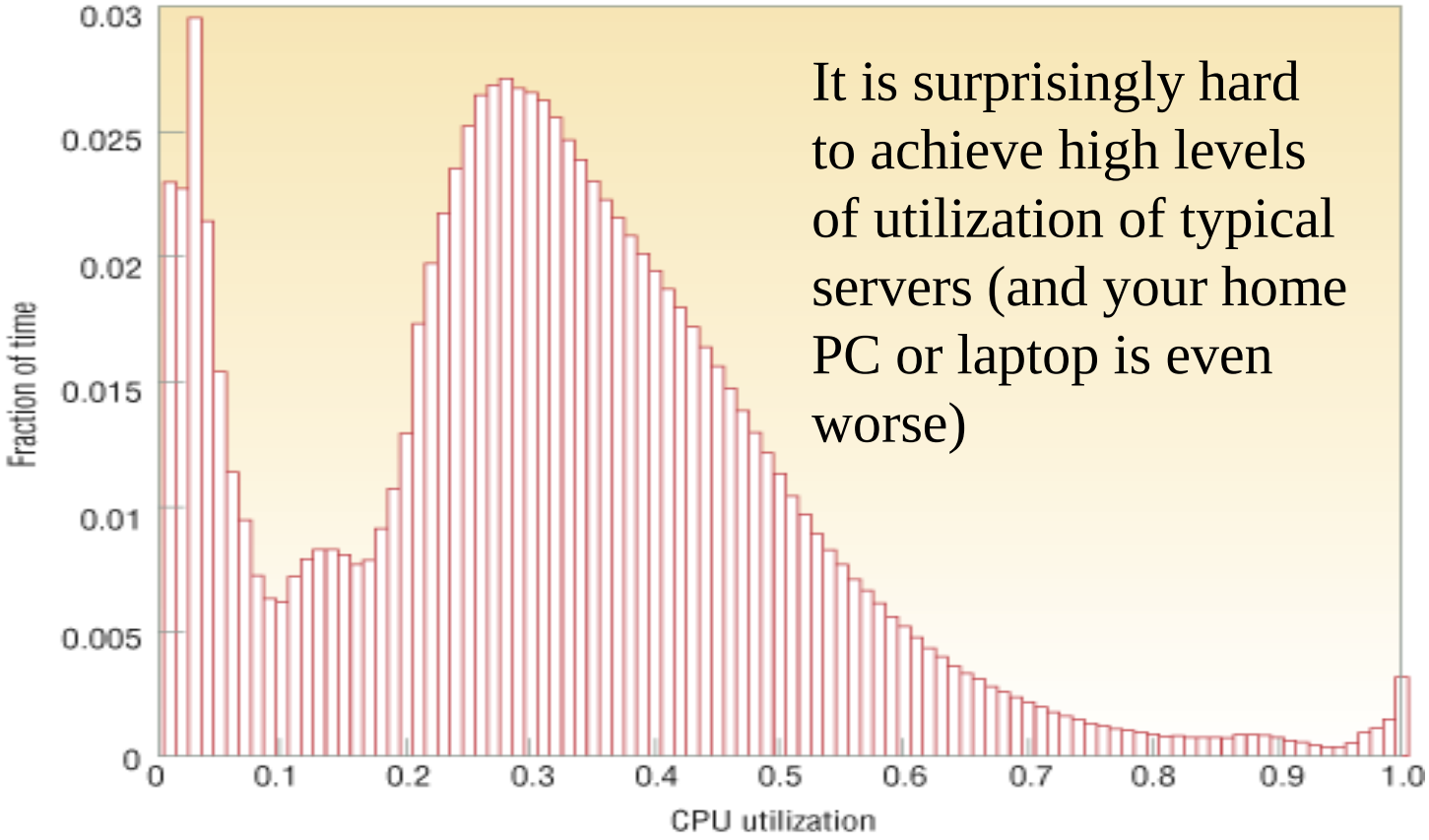
- Energy efficiency
  - Primary concern in the design of WSC
  - Important component of the total cost of ownership
- Power Usage Effectiveness (PUE):

$$\frac{\text{Total Building Power}}{\text{IT equipment Power}}$$

  - Power efficiency measure for WSC
  - Not considering efficiency of servers, networking
  - Perfection = 1.0

Energy Proportionality

"The Case for Energy-Proportional Computing,"
Luiz André Barroso, Urs Hölzle,
IEEE Computer
December 2007



It is surprisingly hard to achieve high levels of utilization of typical servers (and your home PC or laptop is even worse)

Figure 1. Average CPU utilization of more than 5,000 servers during a six-month period. Servers are rarely completely idle and seldom operate near their maximum utilization, instead operating most of the time at between 10 and 50 percent of their maximum

Google WSC A PUE: 1.24 (2008)

- Careful air flow handling
  - Don't mix server hot air exhaust with cold air (separate warm aisle from cold aisle)
  - Short path to cooling so little energy spent moving cold or hot air long distances
  - Keeping servers inside containers helps control air flow
- Elevated cold aisle temperatures
  - 81° F（27.2 ° C ） instead of traditional 65° - 68° F（18.3-20 ° C）
  - Found reliability OK if run servers hotter
- Use of free cooling
  - Cool warm water outside by evaporation in cooling towers
  - Locate WSC in moderate climate so not too hot or too cold
- Per-server 12-V DC UPS
  - Rather than WSC wide UPS, place single battery per server board
  - Increases WSC efficiency from 90% to 99%
- Measure vs. estimate PUE, publish PUE, and improve operation

WSC power management

Goals

- Allow very high average utilization of the facility power
- Minimally disruptive to applications

Key technology: energy storage

Triple-use of UPS functionality

1. **Survive grid outages**
2. **Ride load peaks (see Govindan et al., ISCA 2011)**
3. **Make renewable energy dependable**

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

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%

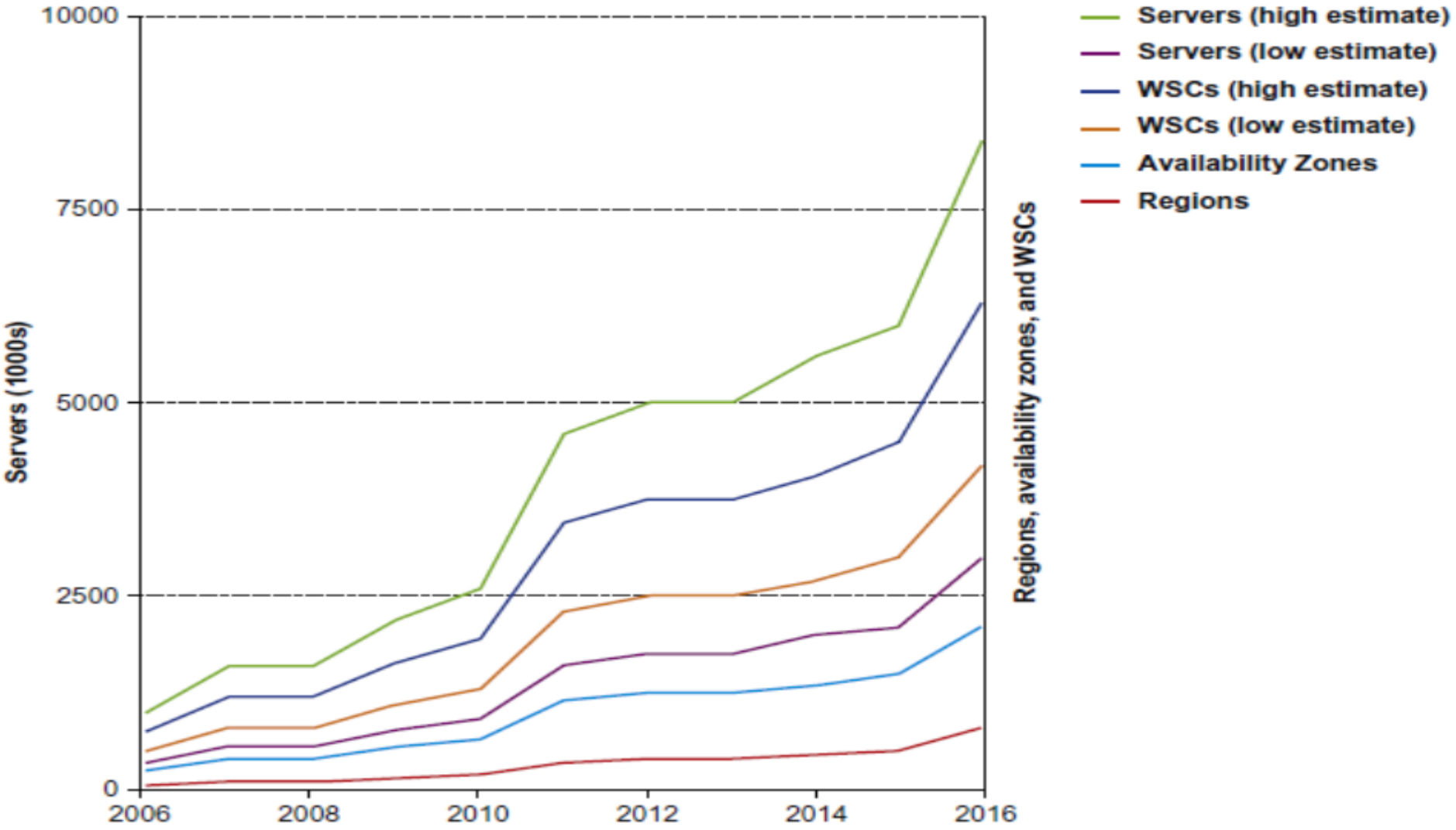
Cost of a WSC

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

Cloud Computing

- Amazon Web Services
  - Virtual Machines: Linux/Xen
  - Low cost
  - Open source software
  - Initially no guarantee of service
  - No contract

Cloud Computing Growth



What is different between DC and WSC ?

Data center

- Co-located machines that shore security, environmental requirements
- Applications -- a few binaries, running on a small number of machines
- Heterogeneous hardware and system software
- Partitioned resources, managed and scheduled separately
- Facility and computing equipment designed separately

Warehouse-scale computer

- Computing system designed to run massive Internet services
- Applications —10s of binaries running on 1000s of machines
- Homogeneous hardware and system software
- Common pool of resources managed centrally
- Integrated design of facility and computing machinery

Alan Curtis Kay
(born May 17, 1940)

- 2003: ACM Turing Award “For pioneering many of the ideas at the root of contemporary object-oriented programming languages, leading the team that developed Smalltalk, and for fundamental contributions to personal computing.”
- People who are really serious about software should make their own hardware
- 真正认真在乎软件的人，应该做自己的硬件。