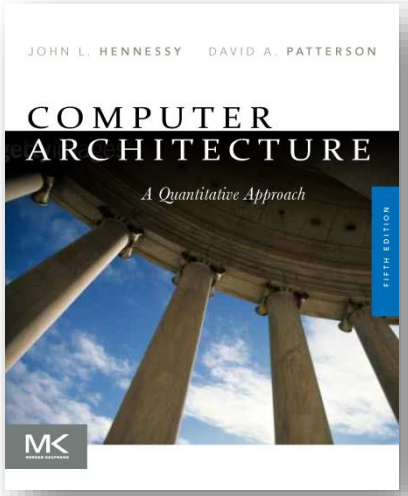


Chapter 6

Warehouse-Scale Computers (WSC) to Exploit Request-Level and Data-Level Parallelism

Adapted by Prof. Gheith Abandah

Contents

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 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
- Location counts
  - Real estate, power cost; Internet, end-user, and workforce availability
- Computing efficiently at mostly low utilization
- Scale and its opportunities and problems
  - Can afford to build customized systems since WSC require volume purchase, bulk discounts
  - Frequent failures

Failures in new 2400-server cluster

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).

Contents

Programming Models and Workloads

- Batch processing framework: MapReduce which has the Hadoop open-source implementation
  - **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

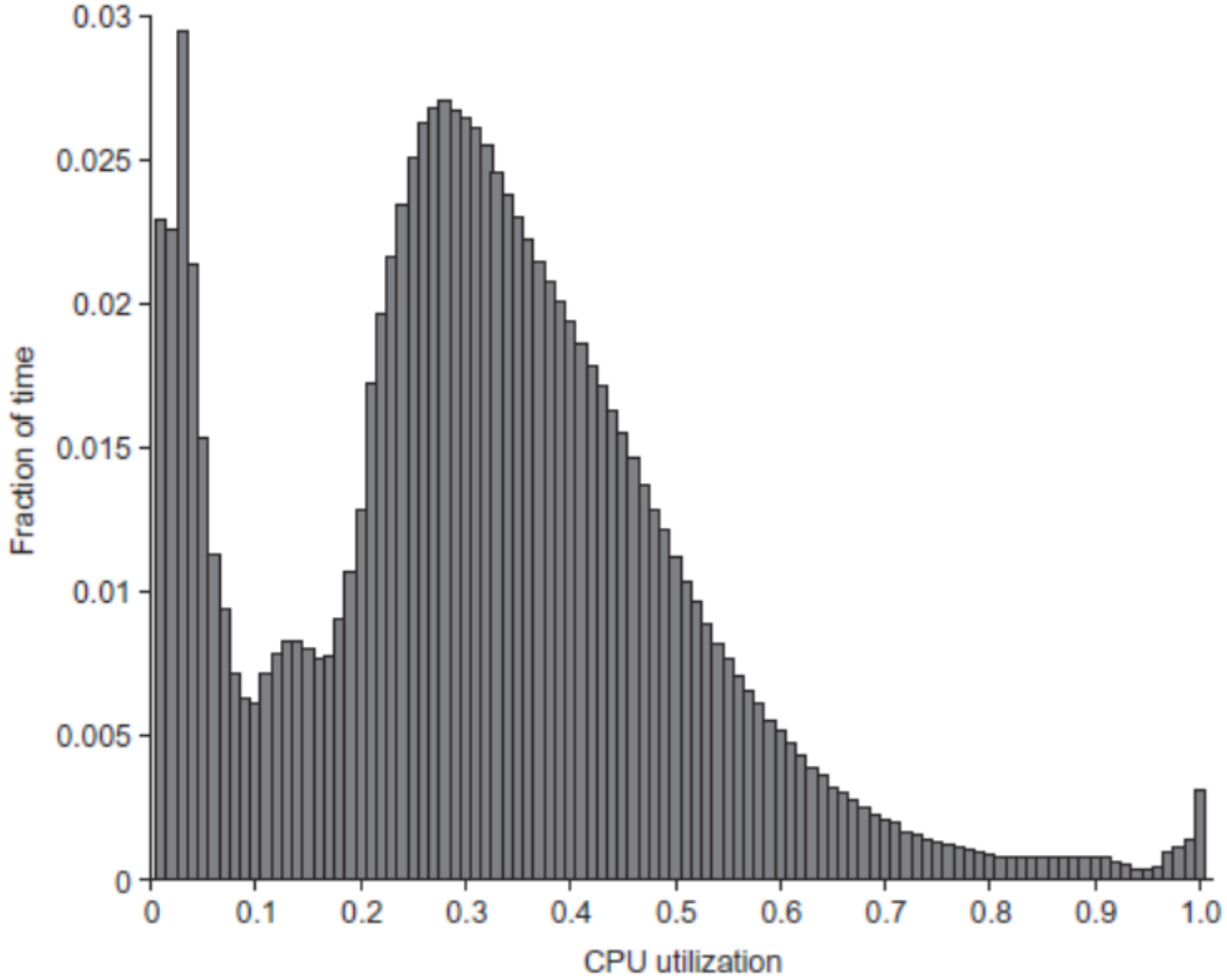
Programming Models and Workloads

- **Availability:**
  - Each node is required to report back to the master node periodically with a list of completed tasks.
  - If a node does not report back by the deadline, the master node deems the node dead and reassigns the node's work to other nodes
  - Use replicas of data across different servers
  - Use relaxed consistency:
    - No need for all replicas to always agree
- **File systems: Google File System (GFS) and Colossus**
- **Databases: Dynamo and BigTable**

Programming 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

Programming Models and Workloads



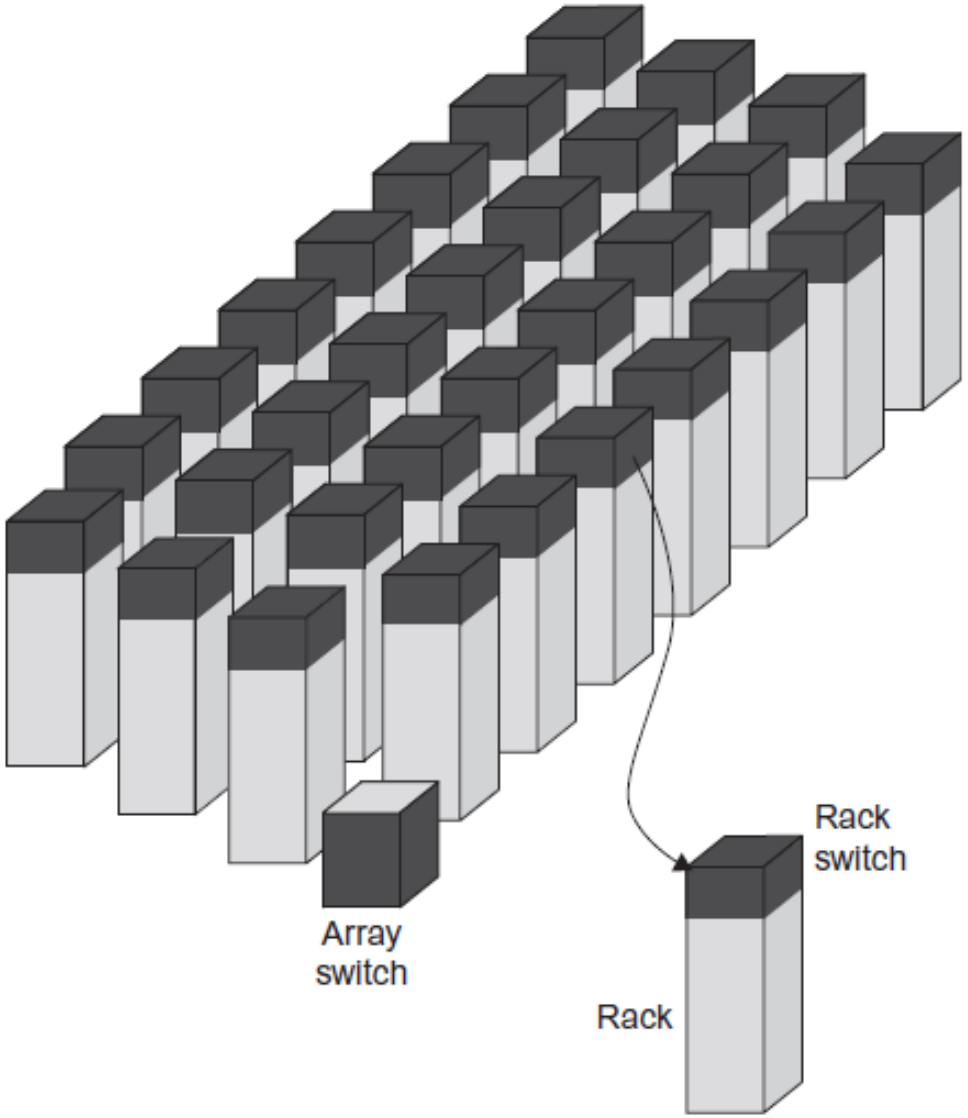
Contents

Computer Architecture of WSC

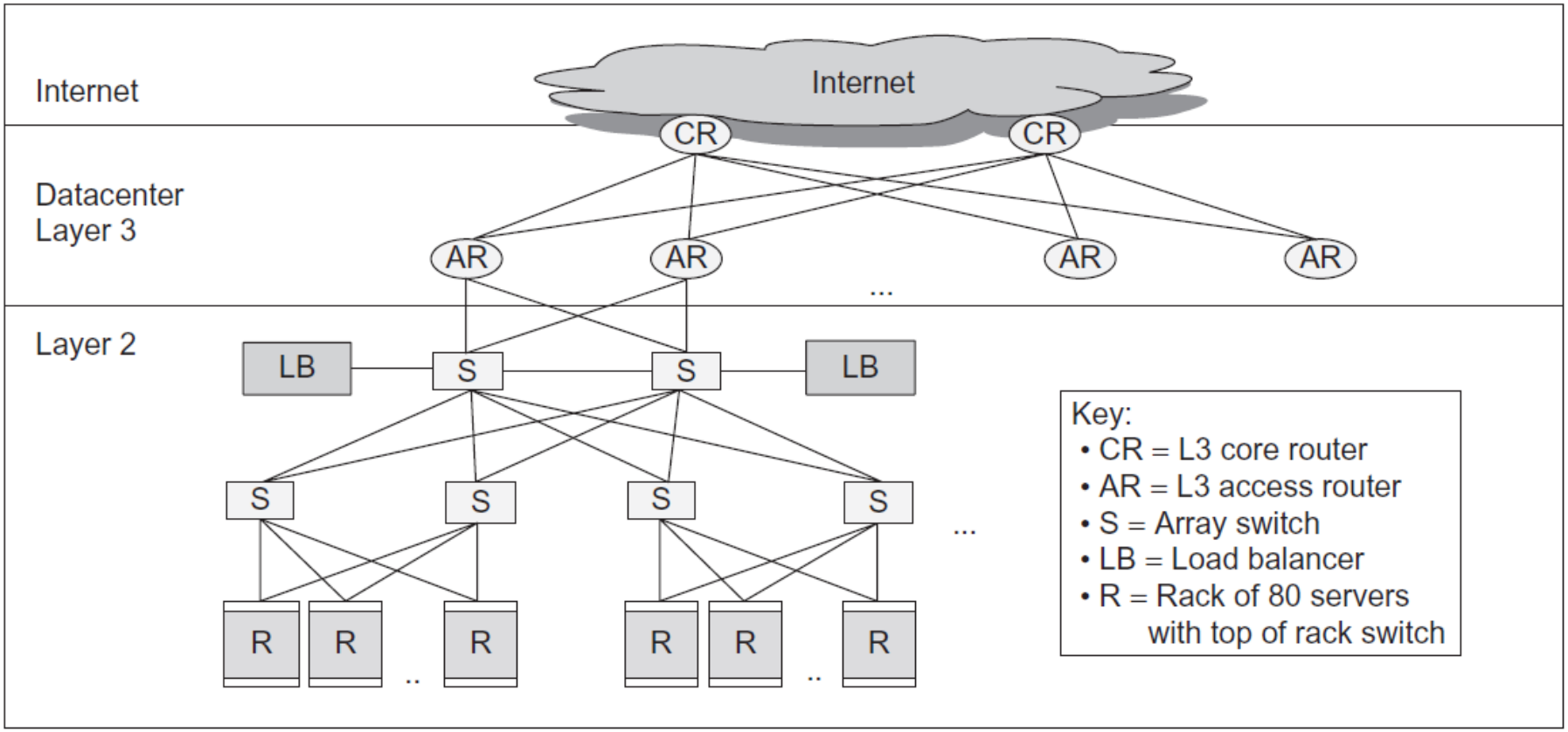
- WSC often use a hierarchy of networks for interconnection, 50,000–100,000 servers
- Each 19” rack holds 48 1U servers connected to a Top of Rack (ToR) switch
  - 1U = 1.76 inch
  - Cabinet dimensions 48 cm x 150 cm
  - ToR has 4-16 up links and 48 down links.
- ToRs are uplinked to switch higher in hierarchy
  - Uplink has 6-24X times lower bandwidth
  - Goal is to maximize locality of communication relative to the rack

Hierarchy of Switches in a WSC

- The Array Switch 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 addressable memory chips and FPGAs



Hierarchy of Switches in a WSC



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 replicas**

WSC Memory Hierarchy

- **Example: 2 racks have 80 servers with one switch, the array is 30 racks**

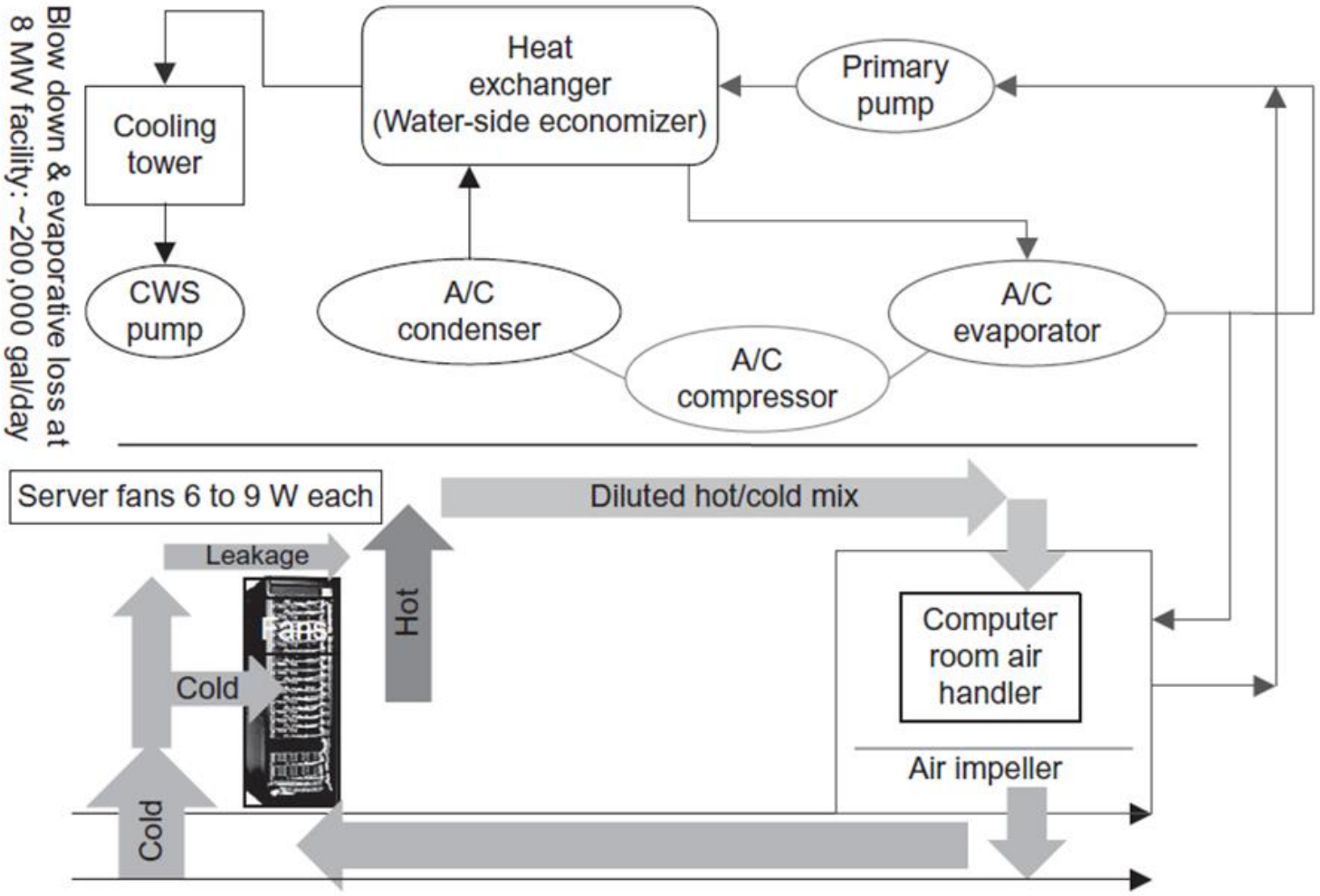
	Local	Rack	Array
DRAM latency (μs)	0.1	300	500
Flash latency (μs)	100	400	600
Disk latency (μs)	10,000	11,000	12,000
DRAM bandwidth (MB/s)	20,000	100	10
Flash bandwidth (MB/s)	1000	100	10
Disk bandwidth (MB/s)	200	100	10
DRAM capacity (GB)	16	1024	31,200
Flash capacity (GB)	128	20,000	600,000
Disk capacity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

Contents

Infrastructure and Costs of WSC

- Cooling and power distribution and are the majority of the construction costs of a WSC.
- Cooling
  - Air conditioning used to cool server room
  - 64 F – 71 F
  - Cooling towers can also be used
- 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

Cooling



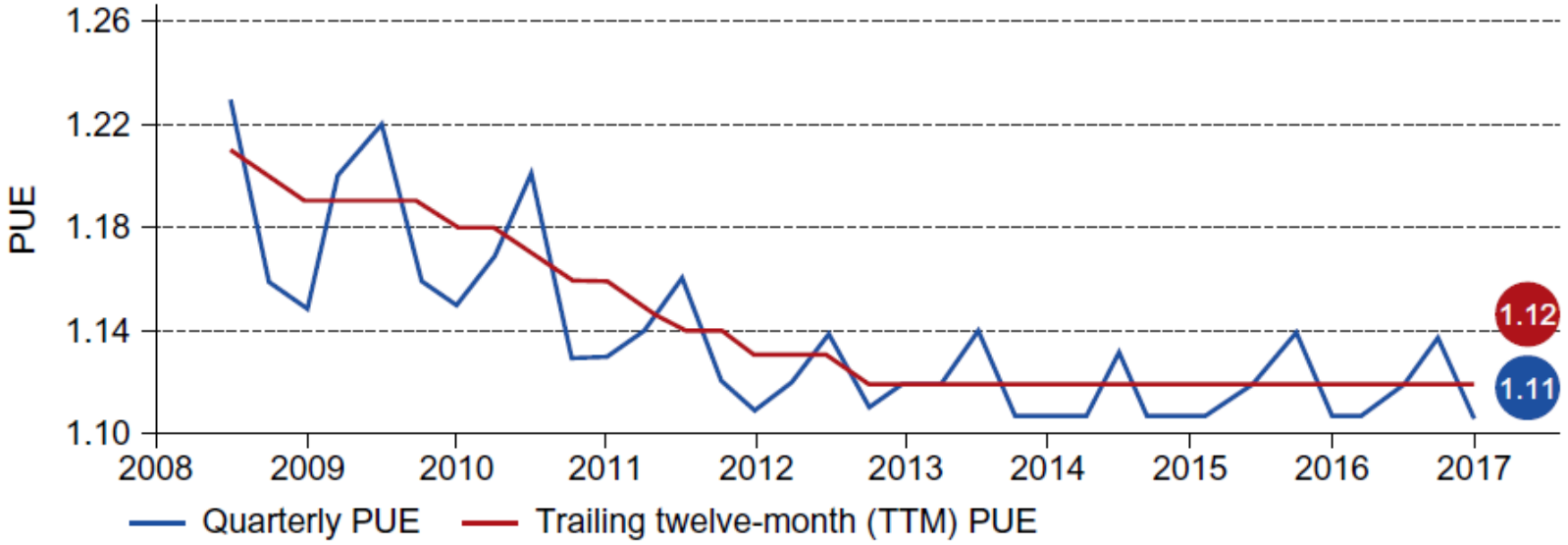
Electric Power

- **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%

Measuring Efficiency of a WSC

- **Power Utilization Effectiveness**
  - **PUE = Total facility power / IT equipment power**
  - **Median PUE on 2006 study was 1.69**
  - **Average PUE of the 15 Google WSCs between 2008 and 2017:**



Measuring Efficiency of a WSC

- Performance
  - Latency is important because it is seen by users
  - Bing study: users will use search less as response time increases

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%

  - Service Level Objectives (SLOs)/Service Level Agreements (SLAs)
    - E.g. 99% of requests be below 100 ms

Cost of a WSC

- Capital expenditures (CAPEX)
  - Cost to build a WSC
  - $9 to 13/watt for the building, power, and cooling
  - CAPEX Example:
    - 8-MW facility $88 million
    - 46,000 servers $67 million
    - Networking $13 million
    - Total $168 million
- Operational expenditures (OPEX)
  - Cost to operate a WSC
  - OPEX Example:
    - Monthly power use $475,000
    - Monthly people salaries and benefits $85,000

Contents

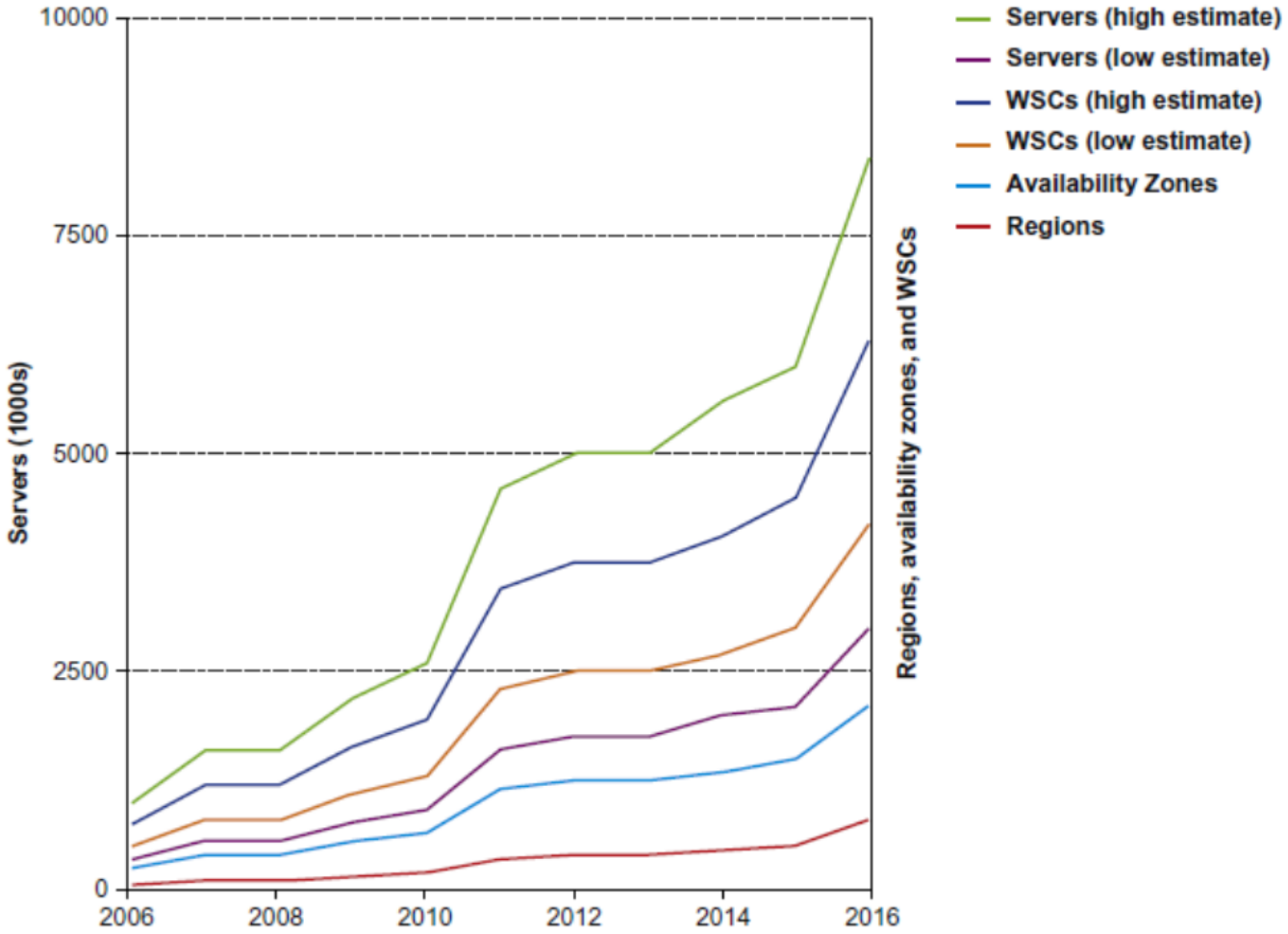
- Introduction
- Programming Models and Workloads for WSC
- Computer Architecture of WSC
- The Efficiency and Cost of WSC
- Cloud Computing: The Return of Utility Computing
- Putting It All Together: A Google WSC
- Fallacies and Pitfalls

Cloud Computing

- Amazon, Google and Microsoft build WSC to provide could services
- WSC are better data centers
  - 5.7 reduction in storage costs
  - 7.1 reduction in administrative costs
  - 7.3 reduction in networking costs
- Amazon Web Services
  - Virtual Machines: Linux/Xen
  - Low cost
  - Open source software
  - Initially no guarantee of service
  - No contract

Cloud Computing

- **Cloud Computing Growth**



Contents

- Introduction
- Programming Models and Workloads for WSC
- Computer Architecture of WSC
- The Efficiency and Cost of WSC
- Cloud Computing: The Return of Utility Computing
- Putting It All Together: A Google WSC
- Fallacies and Pitfalls

A Google WSC

On-site Substation

A Google WSC

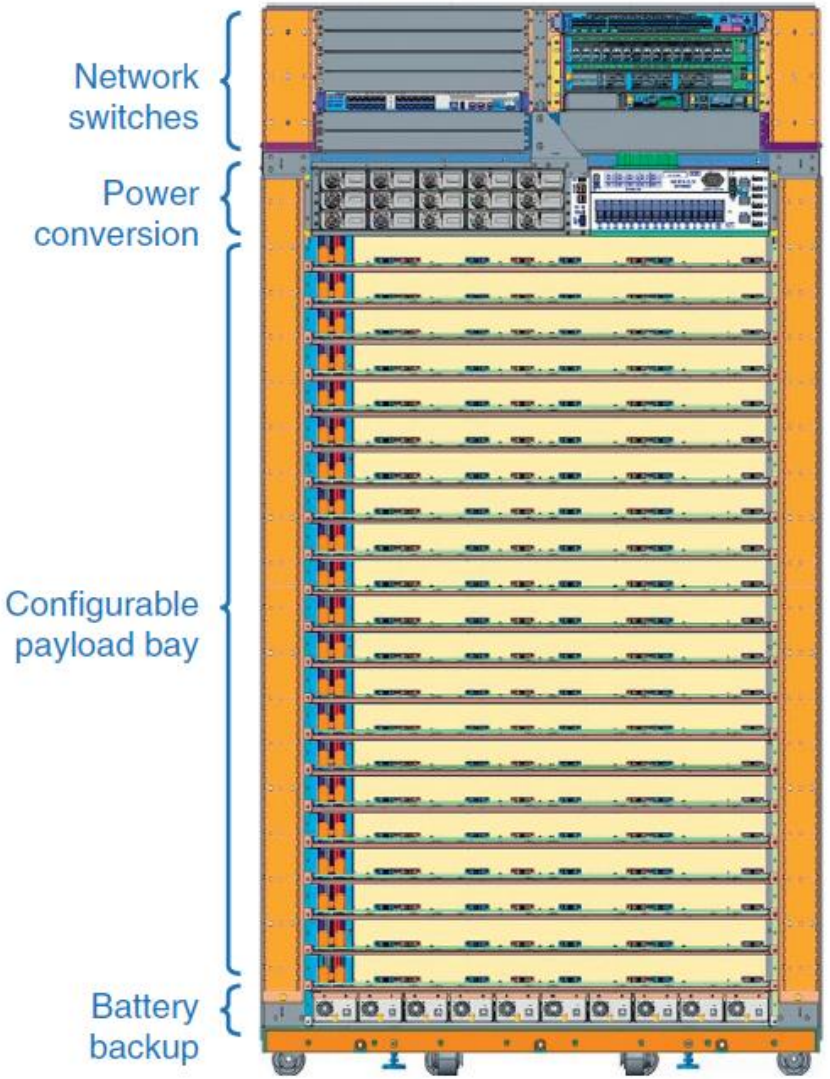
Transformers, switch gear, and generators in close proximity to a WSC

A Google WSC

Row of servers with the copper bus ducts above that distribute 400 V

A Google Rack

- Dimensions: 2 m×1.2 m×0.5m
- The switches are at the rack top
- The power converter converts from 240 V AC to 48 V DC for
- 20 slots 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 are distributed modular DC uninterruptible power supply (UPS) batteries



An Example Server

- Haswell CPUs
- 2 sockets × 18 cores × 2 threads = 72 “virtual cores”
- 2.5 MiB last level cache per core or 45 MiB
- 16 DDR3-1600 DIMMs, 256 GB
- 2 8TB SATA disks
- 10 Gbit/s NIC
- TFP of 150 W
- 4 servers can fit in one tray

Contents

- Introduction
- Programming Models and Workloads for WSC
- Computer Architecture of WSC
- The Efficiency and Cost of WSC
- Cloud Computing: The Return of Utility Computing
- Putting It All Together: A Google WSC
- Fallacies and Pitfalls

Fallacies and Pitfalls

- F: Cloud computing providers are losing money
  - AWS has a margin of 25%, Amazon retail 3%
- P: Focusing on average performance instead of 99th percentile performance
- P: Using too wimpy a processor when trying to improve WSC cost-performance
- P: Inconsistent measure of PUE by different companies
- F: Capital costs of the WSC facility are higher than for the servers that it houses

Fallacies and Pitfalls

- P: Trying to save power with inactive low power modes versus active low power modes
- F: Given improvements in DRAM dependability and the fault tolerance of WSC systems software, there is no need to spend extra for ECC memory in a WSC
- P: Coping effectively with microsecond (e.g. Flash and Ethernet) delays as opposed to nanosecond or millisecond delays
- F: Turning off hardware during periods of low activity improves the cost-performance of a WSC. No: better to use it.