UC Berkeley
Teaching Professor
Dan Garcia

# CS61C

## Great Ideas in Computer Architecture

(a.k.a. Machine Structures)

UC Berkeley
Professor
Bora Nikolić

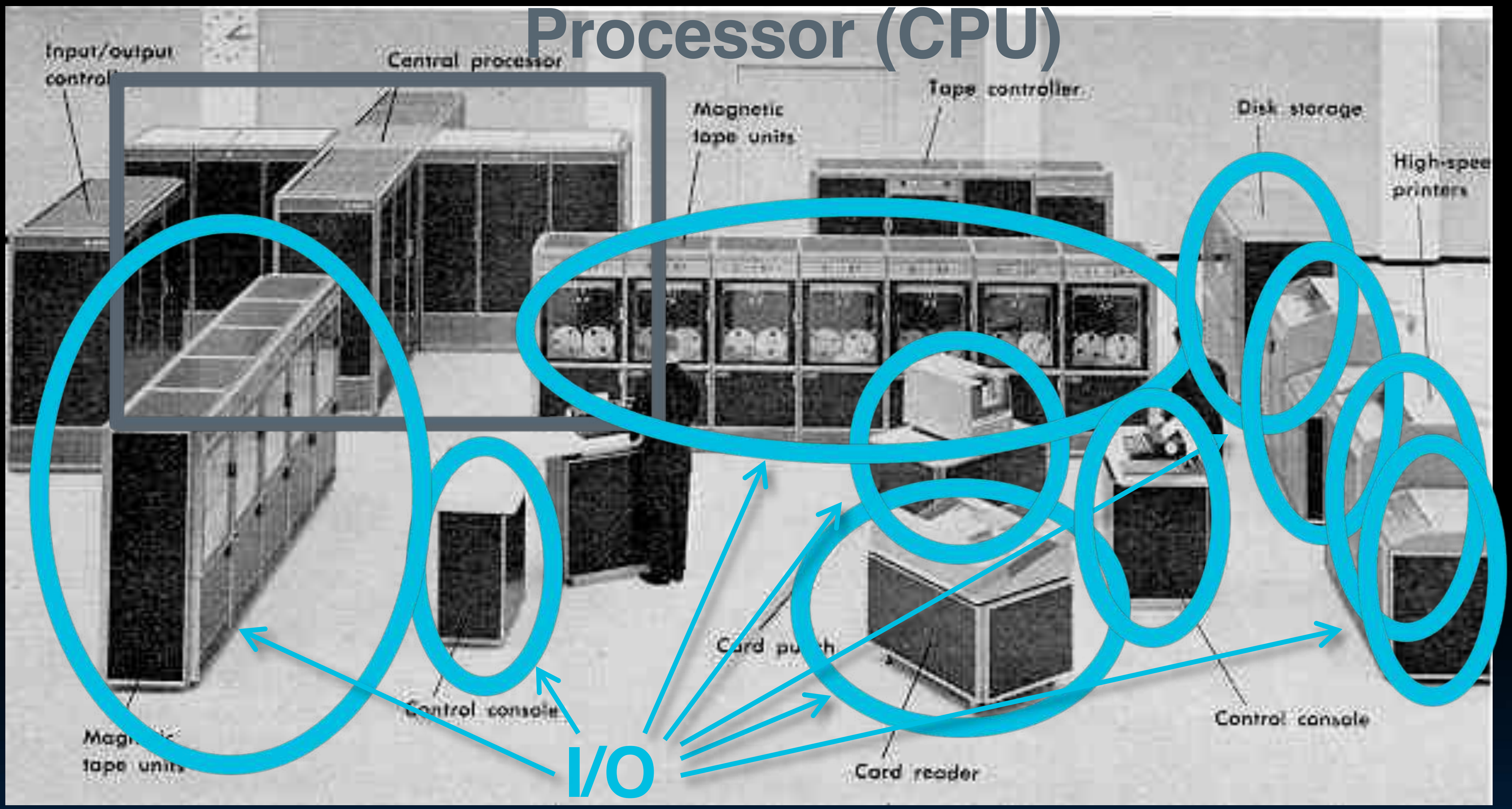
# Datacenters & Cloud Computing

# Eras of Computer Hardware

# Review … where are we?

- Great Ideas in Computer Architecture
  - ✓ Layers of Representation/Interpretation
  - ✓ Moore's Law
  - ✓ Principle of Locality/Memory Hierarchy
  - Parallelism
  - Performance Measurement and Improvement
  - Dependability via Redundancy

# Computer Eras: Mainframe 1950s-60s




"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

Garcia, Nikolić

# 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, Nokia, … build $500 smartphone and tablet computers for individuals
➔ Objective C, Swift, Java, Android OS + iOS

Cloud Computing: Using Local Area Networks, Amazon, Google, … build $200M Warehouse Scale Computers with 100,000 servers for Internet Services for PMDs
➔ MapReduce, Ruby on Rails


Garcia, Nikolić

# Warehouse Scale Computers

# 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

# November 2020 AWS Instances & Prices

| Instance | Per Hour | $ Ratio to Small | EC2 Compute Unit (integer) | Virtual Cores (vCPU) | Memory (GiB) | Disk (GiB) |
|---|---|---|---|---|---|---|
| Standard Small (t3.small) | $0.021 | 1 | Variable | 2 | 2 | EBS |
| Standard Large (t3.large) | $0.083 | 4 | Variable | 2 | 8 | EBS |
| Standard 2x Extra Large (t3.2xlarge) | $0.333 | 16 | Variable | 8 | 32 | EBS |
| High-Mem Large (r5.large) | $0.126 | 6 | 10 | 2 | 16 | EBS |
| High-Mem Double Xlarge (r5.2xlarge) | $0.504 | 24 | 37 | 8 | 64 | EBS |
| High-Mem 24x Large (r5.24xlarge) | $6.048 | 288 | 337 | 96 | 768 | EBS |
| High-CPU Large (c5.large) | $0.085 | 4 | 10 | 2 | 4 | EBS |
| High-CPU 18x Large (c5.18xlarge) | $3.060 | 146 | 281 | 72 | 144 | EBS |

- Closest computer in WSC example is Standard 2X Extra Large
- At these low rates, Amazon EC2 can make money! (even utilized 50% time)
- EBS = Elastic Block Store (SSD=$0.10/GB-month, HDD=$0.045/GB-month)
- Each also comes with dedicated attached SSD if you choose & pay for that

# Warehouse Scale Computers

- Massive scale datacenters: 10,000 to 100,000 servers + networks to connect them together
  - Emphasize cost-efficiency
  - Attention to power: distribution and cooling
- (relatively) homogeneous hardware/software
- Offer very large applications (Internet services): search, social networking, video sharing
- Very highly available: < 1 hour down/year
  - Must cope with failures common at scale
- "…WSCs are no less worthy of the expertise of computer systems architects than any other class of machines"

  – Barroso and Hoelzle 2009

# Design Goals of a WSC

- Unique to Warehouse-scale
  - Ample parallelism:
    - Batch apps: large number independent data sets with independent processing.
    - Also known as Data-Level Parallelism
  - Scale and its Opportunities/Problems
    - Relatively small number of these make design cost expensive and difficult to amortize
    - But price breaks are possible from purchases of very large numbers of commodity servers
    - Must also prepare for high # of component failures
  - Operational Costs Count:
    - Cost of equipment purchases << cost of ownership

# Eg., Google's Oregon WSC



Garcia, Nikolić

# Containers in WSCs

Inside WSC

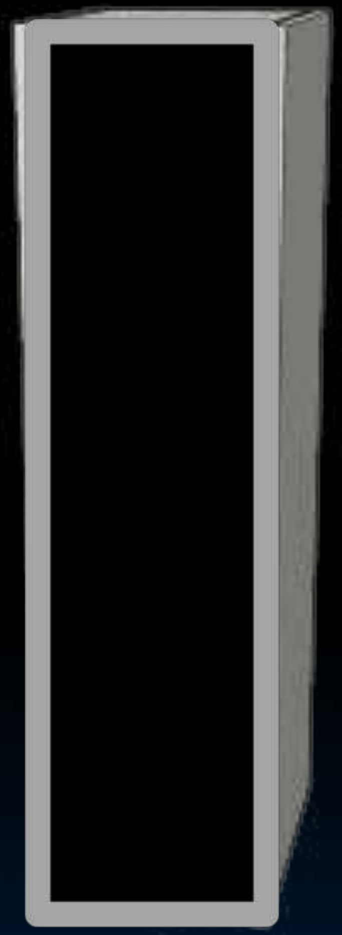
Inside Container



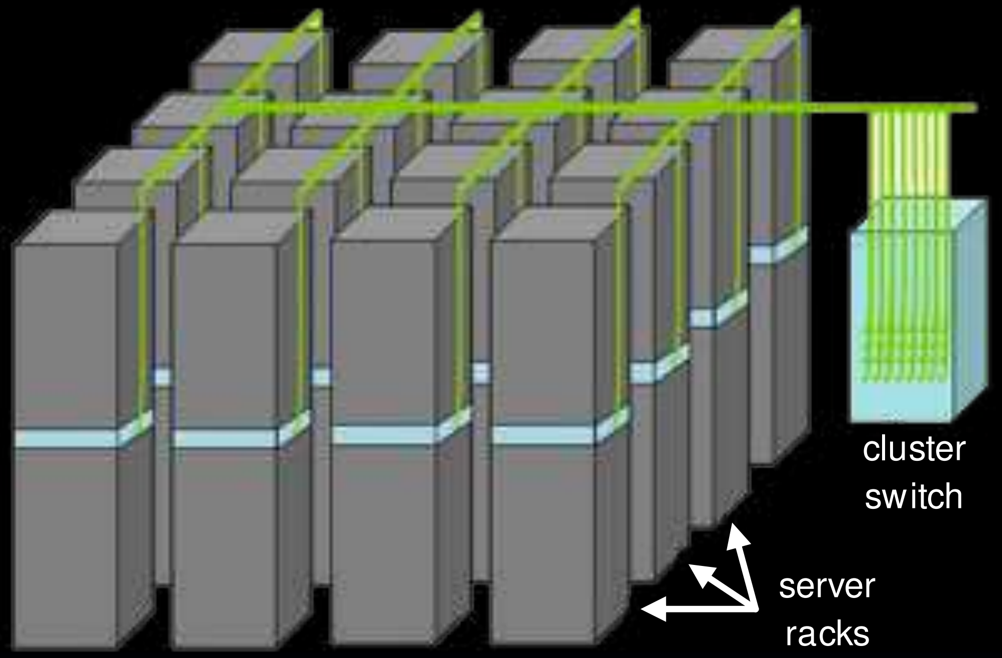
Garcia, Nikolić

# 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")



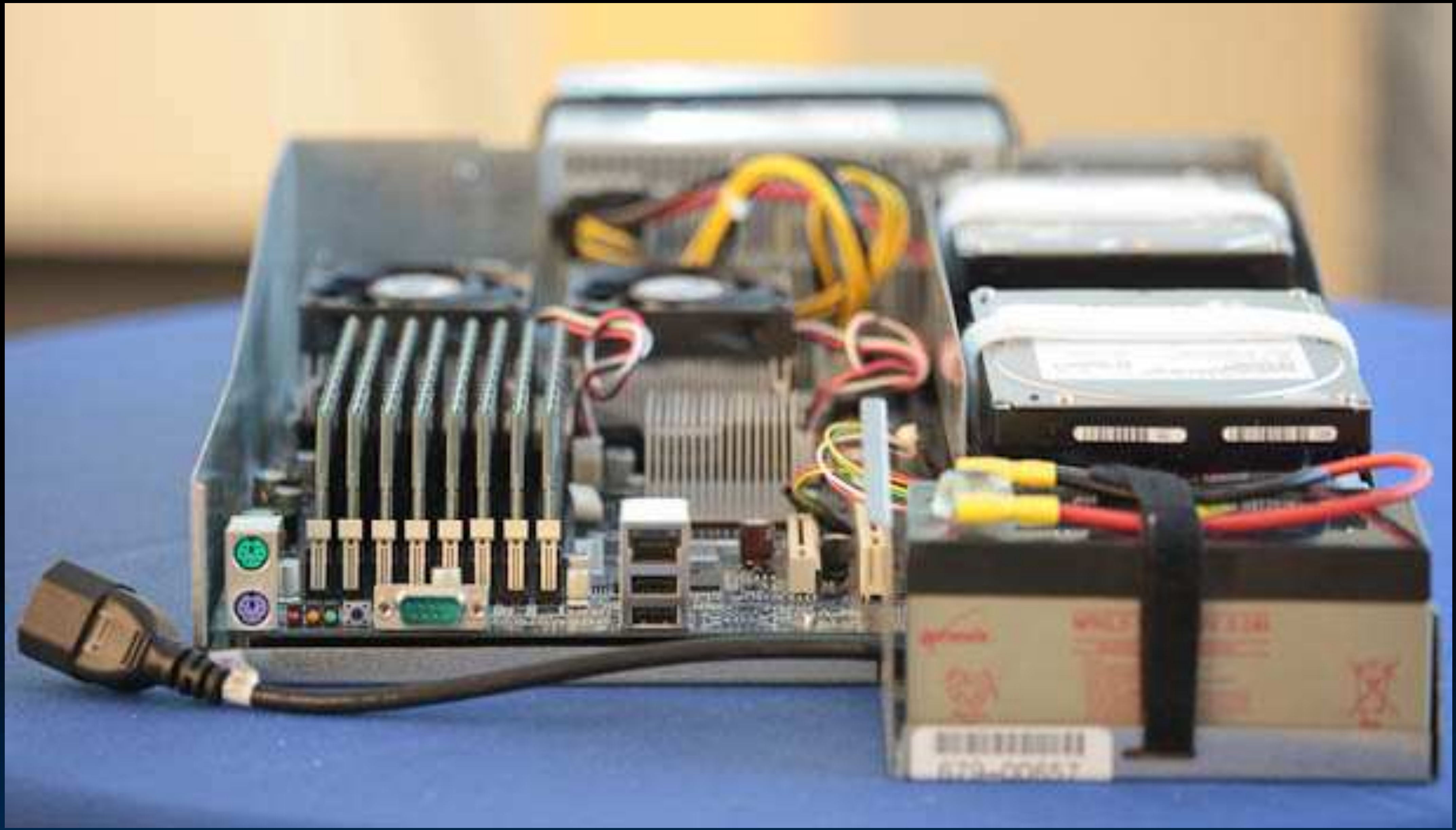

Array (aka cluster):
16-32 server racks + larger local area network switch ("array switch") 10X faster → cost 100X: cost f($N^2$)


Garcia, Nikolić

# Server, Rack, Array

# Google Server Internals

# Google Server Internals



# Defining Performance

- What does it mean to say X is faster than Y?

- 2009 Ferrari 599 GTB
  - 2 passengers, 11.1 secs for quarter mile (call it 10sec)
- 2009 Type D school bus
  - 54 passengers, quarter mile time? (let's guess 1 min)
- Response Time or Latency
  - time between start and completion of a task
  - E.g., time to move vehicle ¼ mile
- Throughput or Bandwidth
  - total amount of work in a given time
  - E.g., passenger-miles in 1 hour



Garcia, Nikolić

# Coping with Performance in Array

Lower latency to DRAM in another server than local disk

Higher bandwidth to local disk than to DRAM in another server

| | Local | Rack | Array |
|---|---|---|---|
| Racks | -- | 1 | 30 |
| Servers | 1 | 80 | 2400 |
| Cores (Processors) | 8 | 640 | 19,200 |
| DRAM Capacity (GB) | 16 | 1,280 | 38,400 |
| Disk Capacity (TB) | 4 | 320 | 9,600 |
| 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 |

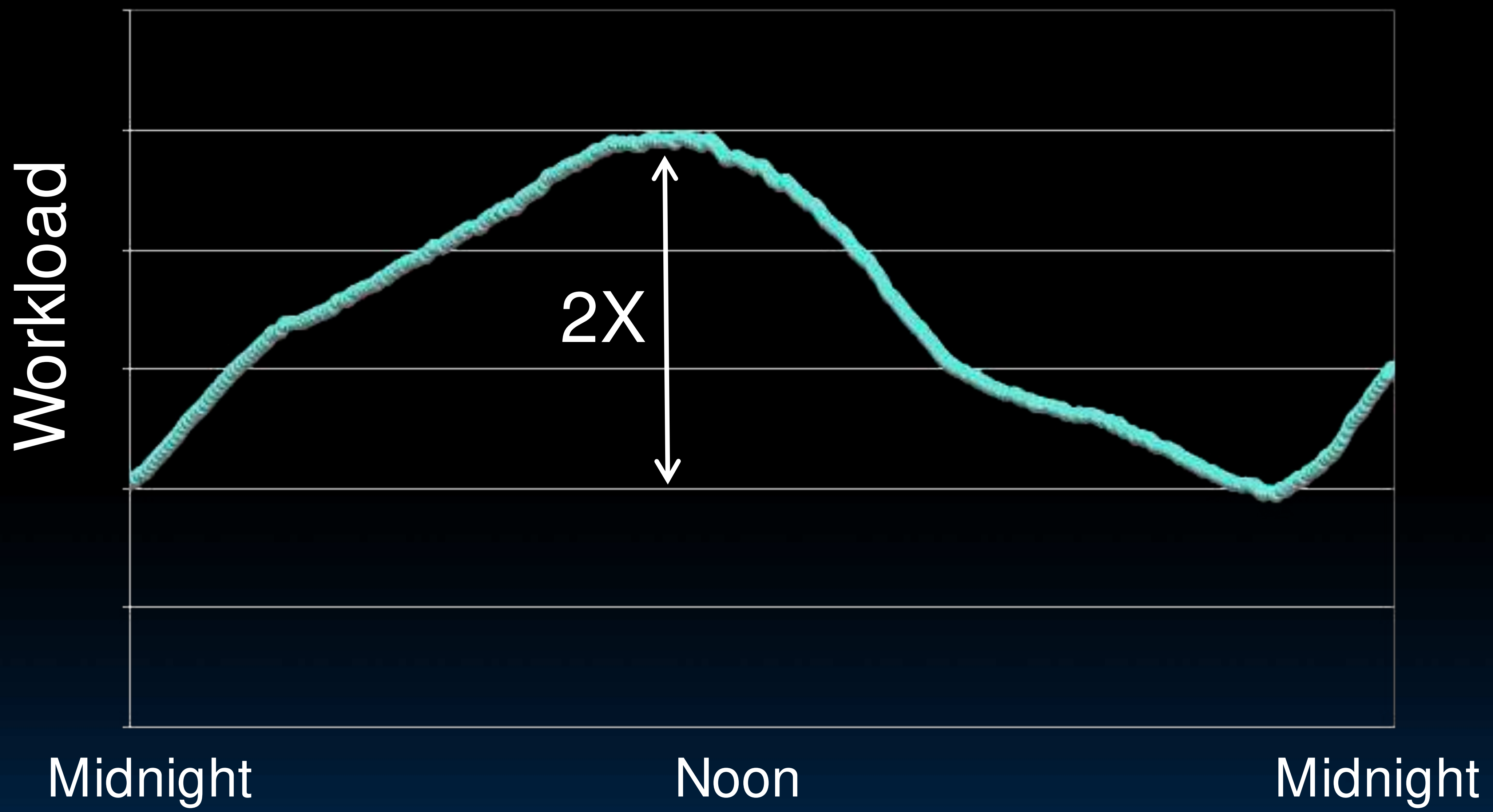
☑



Garcia, Nikolić

# Power Usage Effectiveness (PUE)

# Coping with Workload Variation

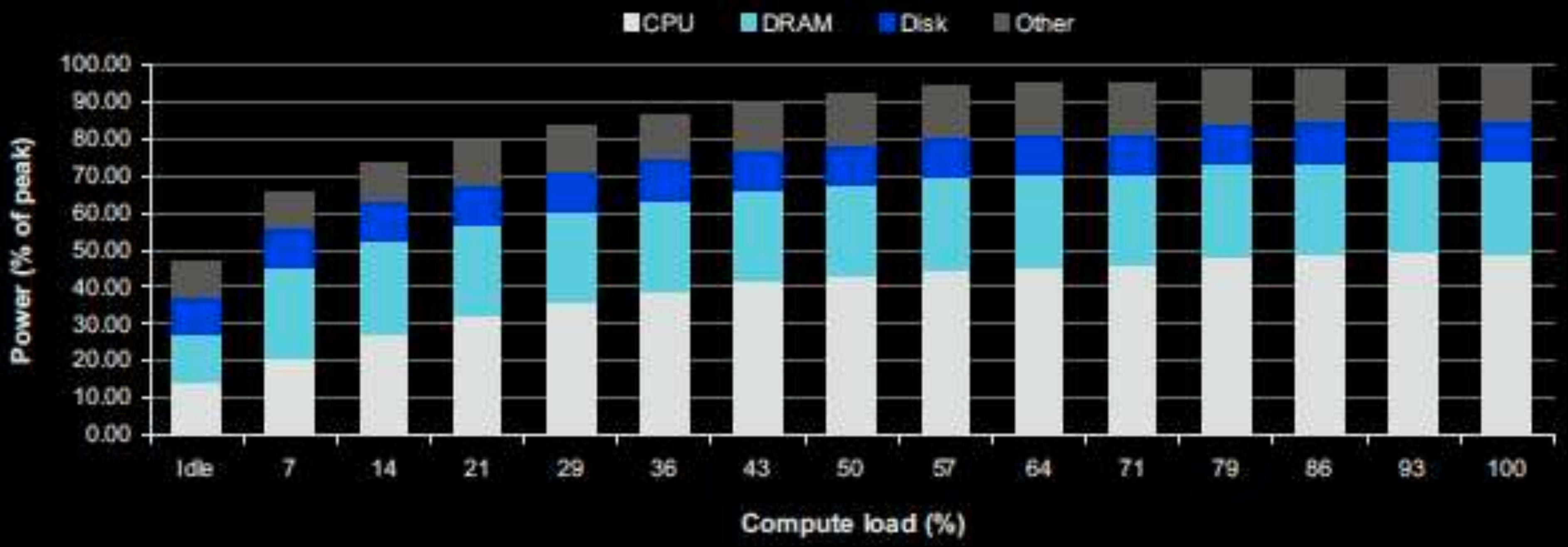



- Online service: Peak usage 2X off-peak

# Impact of latency, bandwidth, failure, varying workload on WSC software?

- WSC Software must take care where it places data within an array to get good performance
- WSC Software must cope with failures gracefully
- WSC Software must scale up and down gracefully in response to varying demand
- More elaborate hierarchy of memories, failure tolerance, workload accommodation makes WSC software development more challenging than software for single computer

# Power vs. Server Utilization




- Server power usage as load varies idle to 100%
- Uses ½ peak power when idle!
- Uses ⅔ peak power when 10% utilized! 90%@50%!
- Most servers in WSC utilized 10% to 50%
- Goal should be Energy-Proportionality:
  % peak load = % peak energy


Garcia, Nikolić

# Power Usage Effectiveness

- Overall WSC Energy Efficiency: amount of computational work performed divided by the total energy used in the process
- Power Usage Effectiveness (PUE):
  Total building power / IT equipment power
  - A power efficiency measure for WSC, not including efficiency of servers, networking gear
  - 1.0 = perfection

# PUE in the Wild (2007)

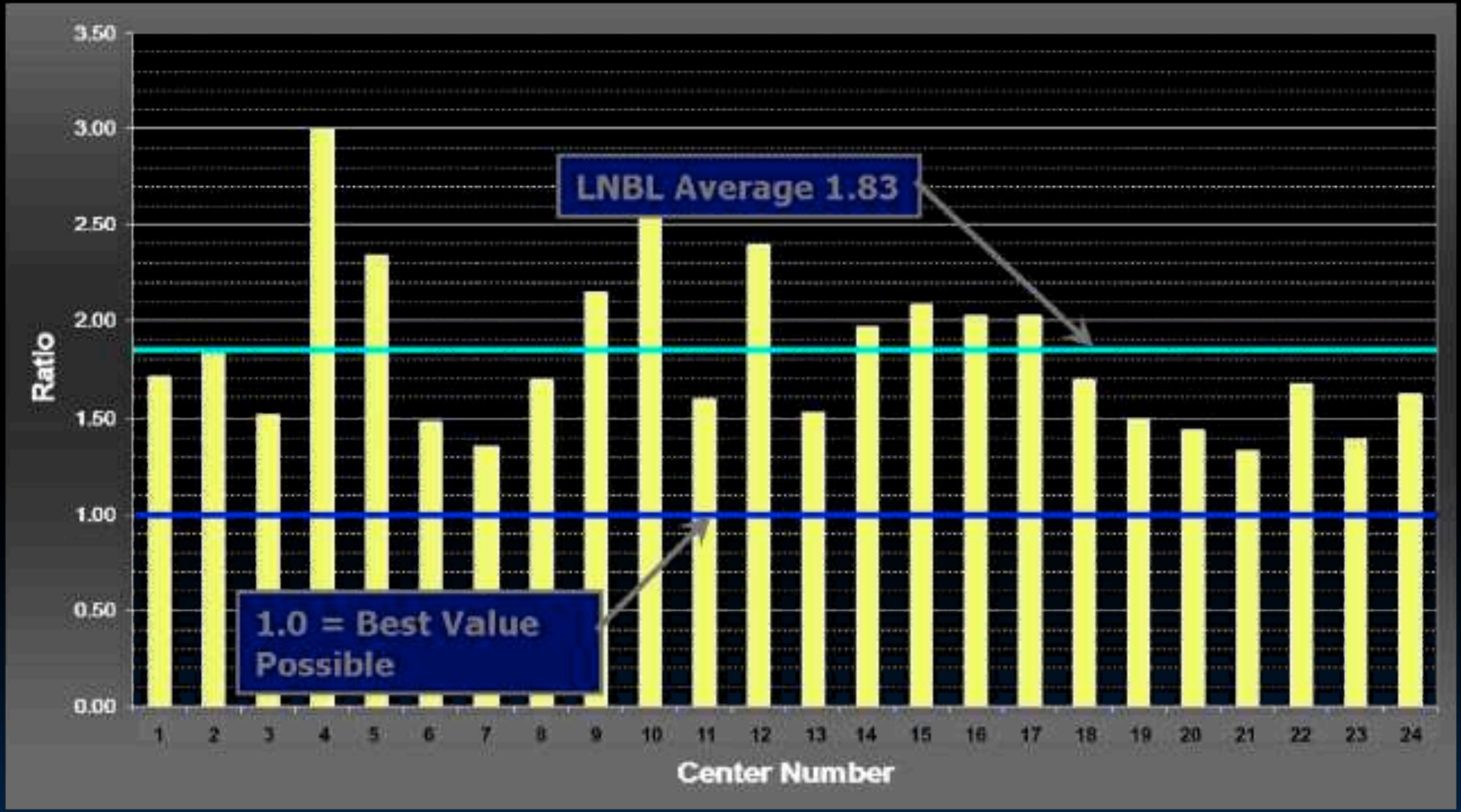
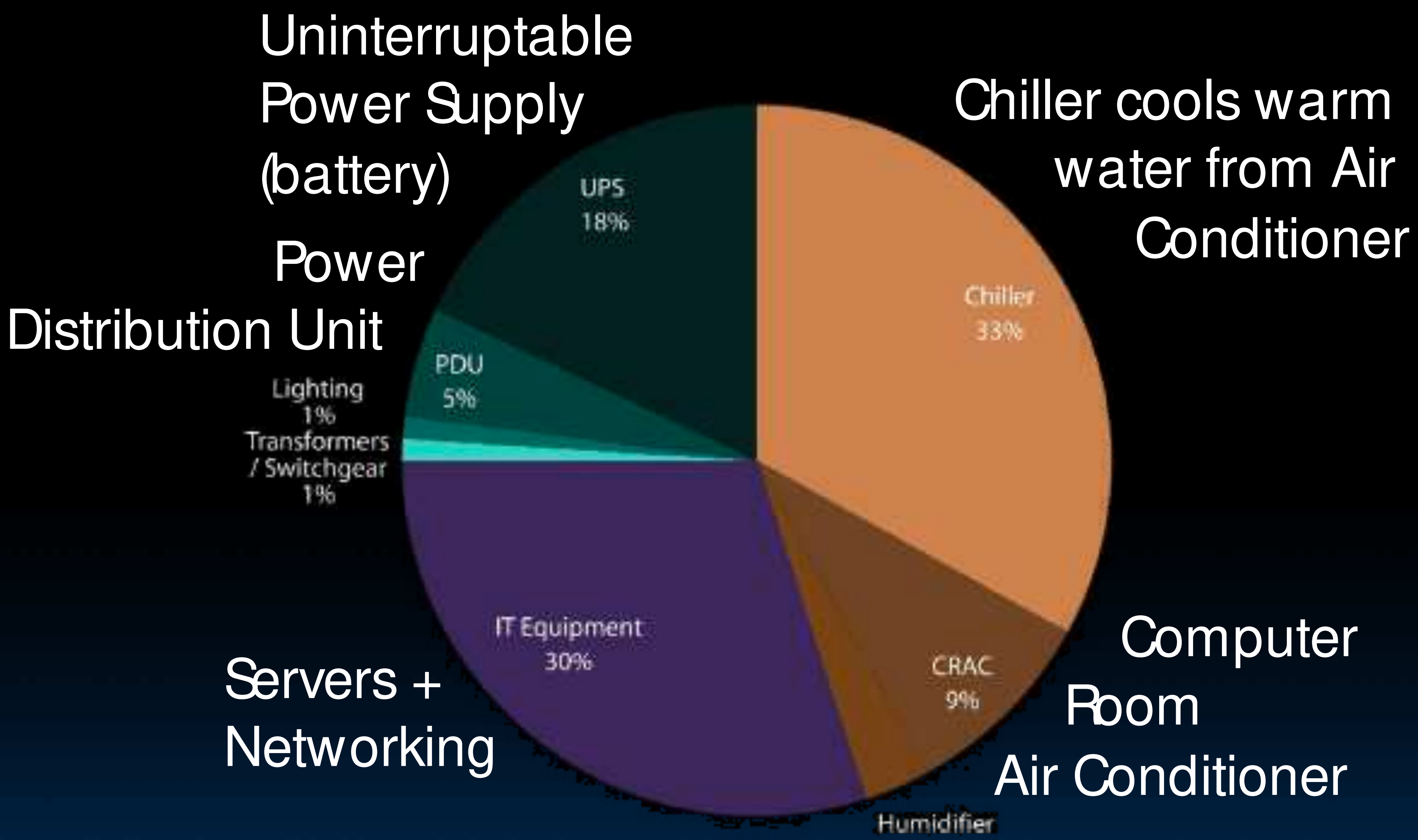



FIGURE 5.1: LBNL survey of the power usage efficiency of 24 datacenters, 2007 (Greenberg et al.)



Garcia, Nikolić

# High PUE: Where Does Power Go?





Garcia, Nikolić

# Google WSC A PUE: 1.24

- 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 instead of traditional 65° - 68° F
  - 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


Garcia, Nikolić

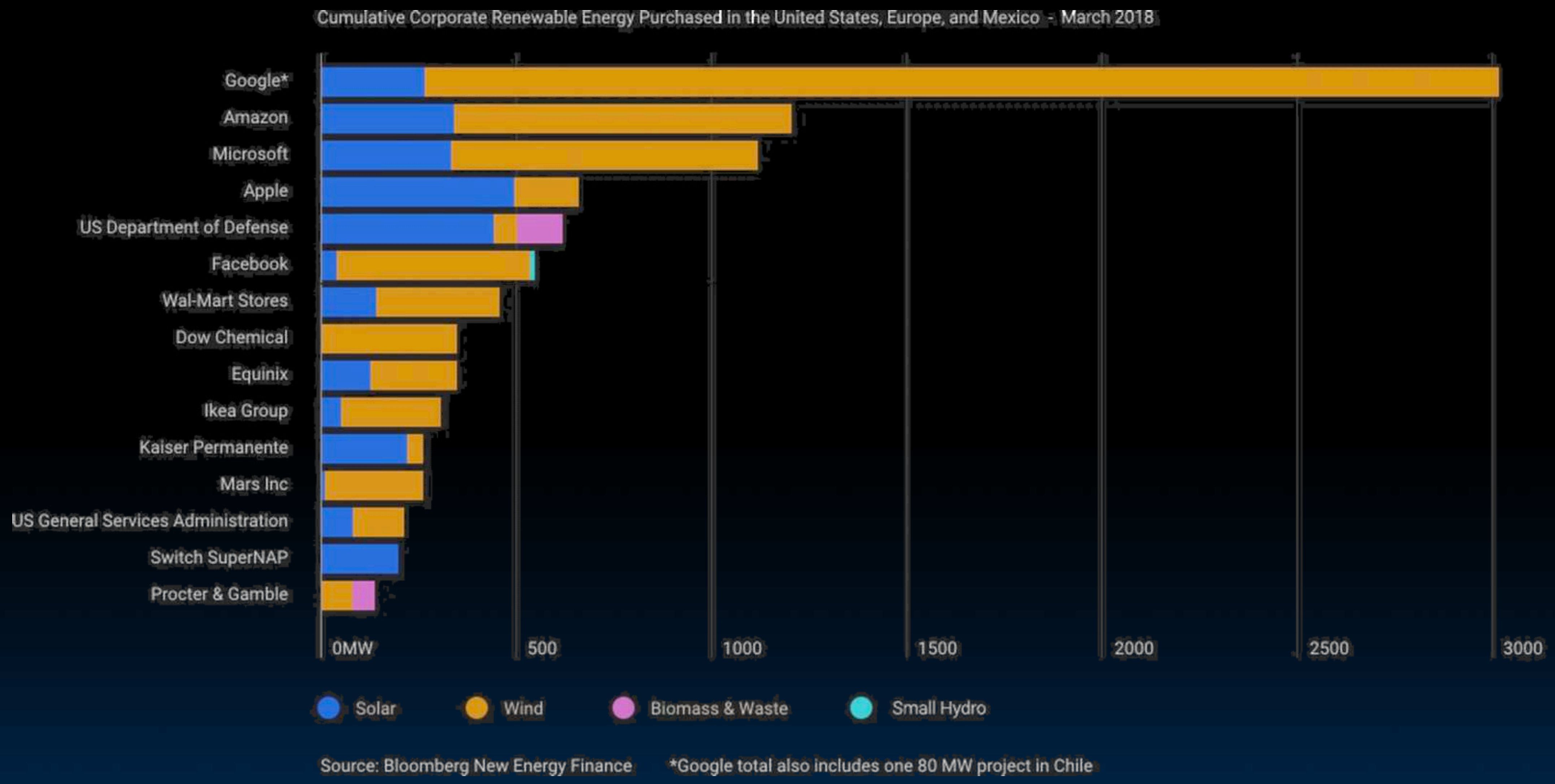
# Computing in the News

Urs Hölzle, Google SVP
Co-author of today's reading

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

# Computing in the News

Urs Hölzle, Google SVP
Co-author of today's reading



# Summary

- Parallelism is one of the Great Ideas
  - Applies at many levels of the system – from instructions to warehouse scale computers
- Post PC Era: Parallel processing, smart phone to WSC
- WSC SW must cope with failures, varying load, varying HW latency bandwidth
- WSC HW sensitive to cost, energy efficiency
- WSCs support many of the applications we have come to depend on


Garcia, Nikolić