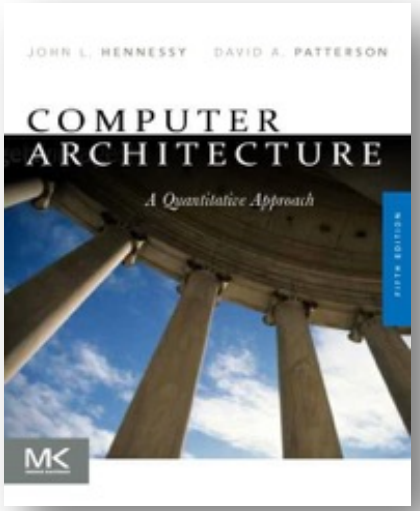


Warehouse-Scale Computers to Exploit Request-Level and Data-Level Parallelism

Based on the Notes of Computer Architecture

A Quantitative Approach, Fifth Edition

And Shangai Jiao Tong University Notes

- The datacenter is the computer
  - Luiz Andre Barroso, Google

2020 Eckert-Mauchly Award

https://www.barroso.org/

Contents

1. Introduction to WS - Data Scale Architectures
2. Cloud computing – basic conceptrs
3. Request-level parallelism and WSCs (Warehouse Scale Computers).
4. Programming models – MapReduce
5. WSC architecture
6. Energy-proportional systems
7. Scientific applications and WSC
8. Cloud computing
9. Ultra-SCALE Systems
    - Fog/Edge

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
  - 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 - Failures

- Outages and anomalies with the approximate frequencies of occurrences of a cluster with 2400 servers

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

Introduction

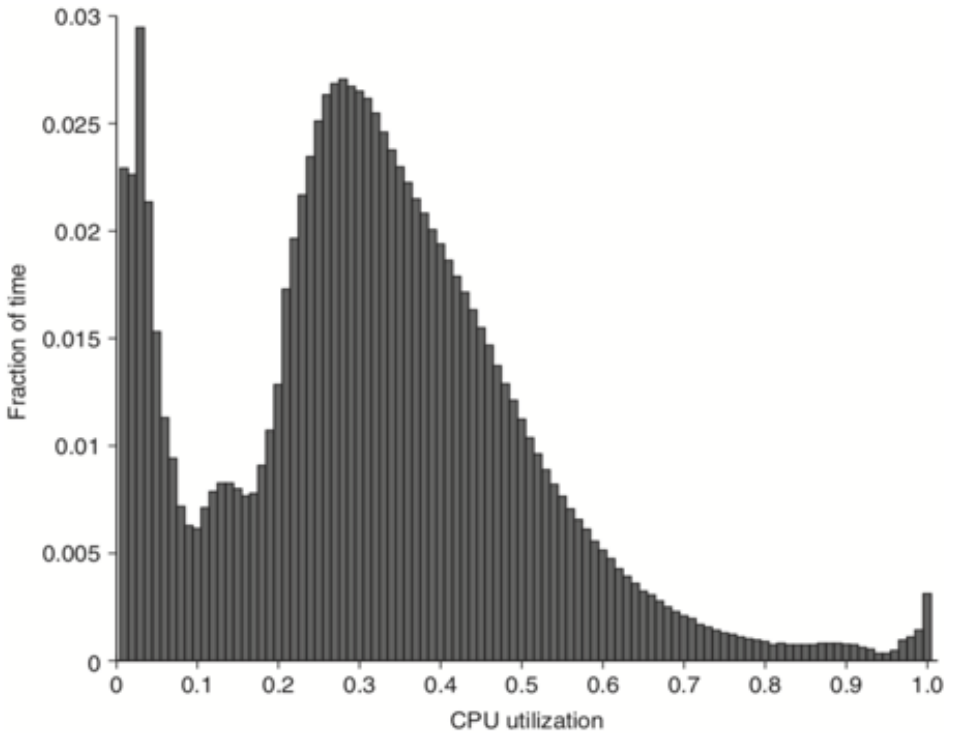
- The WSC goal is to make the hardware/software in the warehouse act like a single computer that typically runs a variety of applications.
- The largest cost in a conventional datacenter is the people to maintain it, whereas, in a well-designed WSC the server hardware is the greatest cost, and people costs shift from the topmost to nearly irrelevant.

Features of WSC

- To reduce operational costs of availability, all WSCs use automated monitoring software so that one operator can be responsible for more than 1000 servers.
-
- Programming frameworks rely upon internal software services for their success
  - AWS Infrastructure, Google FileSystem, BigTable, Dynamo…

Features of WSC

- The workload demands of public interactive services all vary considerably
  - even a popular global service such as Google search varies by a factor of two depending on the time of day
- It is more important for servers in a WSC to perform well while doing little than to just to perform efficiently at their peak, as they rarely operate at their peak.



Average CPU utilization of more than 5000 servers during a 6-month period at Google.

WSC hardware and software must cope with variability in load based on user demand and in performance and dependability due to the vagaries of hardware at this scale

Network-centric computing

- Information processing can be done more efficiently on large farms of computing and storage systems accessible via the Internet.
  - Grid computing – initiated by the National Labs in the early 1990s; targeted primarily at scientific computing
  - Utility computing – initiated in 2005-2006 by IT companies and targeted at enterprise computing.

- The focus of utility computing is on the business model for providing computing services; it often requires a cloud-like infrastructure.

- Cloud computing is a path to utility computing embraced by major IT companies including: Amazon, HP, IBM, Microsoft, Oracle, and others.

Network-centric content

- Content: any type or volume of media, be it static or dynamic, monolithic or modular, live or stored, produced by aggregation, or mixed.

- The “Future Internet” will be content-centric; the creation and consumption of audio and visual content is likely to transform the Internet to support increased quality in terms of resolution, frame rate, color depth, stereoscopic information.

Evolution of concepts and technologies

- The web and the semantic web - expected to support composition of services. The web is dominated by unstructured or semi-structured data, while the semantic web advocates inclusion of sematic content in web pages.
- The Grid - initiated in the early 1990s by National Laboratories and Universities; used primarily for applications in the area of science and engineering.
- Peer-to-peer systems
- Computer clouds
- UltraScale Systems

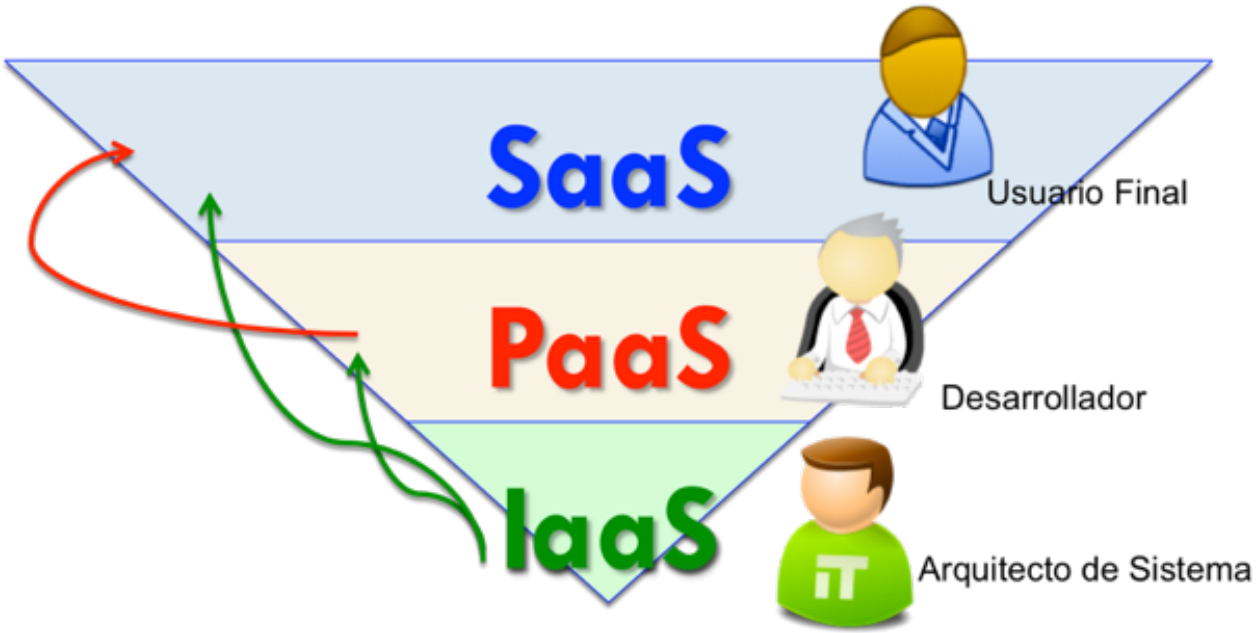
Cloud computing (Well known features

- Uses Internet technologies to offer scalable and elastic services. The term “elastic computing refers to the ability of *dynamically acquiring computing resources* and supporting a variable workload.
- The resources used for these services can be metered and the *users can be charged only for the resources they used*.
- The maintenance and security are ensured by service providers.
- The service providers can operate more efficiently due to specialization and centralization.

Cloud computing (cont'd)

- Lower costs for the cloud service provider are past to the cloud users.
- Data is stored:
  - closer to the site where it is used.
  - in a device and in a location-independent manner.
- The data storage strategy can increases reliability, as well as security and lower communication costs

Cloud Computing (model) Visibility (in spanglish)



Visit: http://prezi.com/i0sretldeyk7/computacion-en-la-nube-y-sus-implicaciones-para-la-industria-del-software-en-colombia/

Types of clouds

- Public Cloud - the infrastructure is made available to the general public or a large industry group and is owned by the organization selling cloud services.
- Private Cloud - infrastructure operated solely for an organization.
- Community Cloud - the infrastructure is shared by several organizations and supports a specific community that has shared.
- Hybrid Cloud - composition of two or more clouds (public, private, or community) bound by standardized technology that enables data and application portability.

The “good” about cloud computing

- Resources such as CPU cycles, storage, network bandwidth are shared.
- When multiple applications share a system their peak demands for resources are not synchronized thus, *multiplexing leads to a higher resource utilization*.
- Resources can be aggregated to support data-intensive applications.
- Data sharing facilitates collaborative activities. Many applications require multiple types of analysis of shared data sets and multiple decisions carried out by groups scattered around the globe.

More “good” about cloud computing

- Eliminate the initial investment costs for a private computing infrastructure and the maintenance and operation costs.
- Cost reduction: concentration of resources creates the opportunity to pay as you go for computing.
- Elasticity: the ability to accommodate workloads with very large peak-to-average ratios.
- User convenience: virtualization allows users to operate in familiar environments rather than in idiosyncratic ones.

Why cloud computing could be successful when other paradigms have failed?

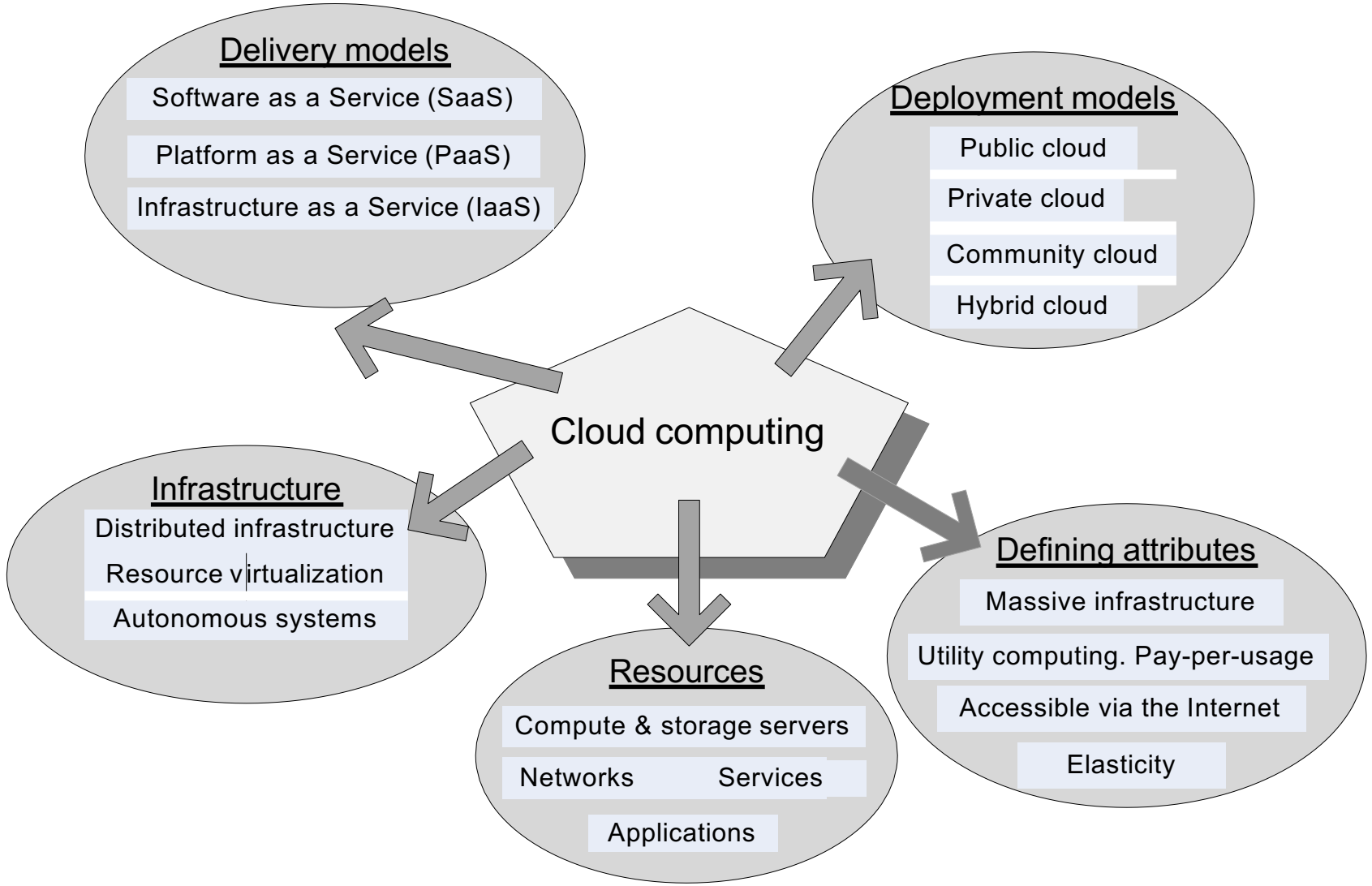
- It is in a better position to exploit recent advances in software, networking, storage, and processor technologies promoted by the same companies who provide cloud services.
- It is focused on enterprise computing; its adoption by industrial organizations, financial institutions, government, and so on could have a huge impact on the economy.
- A cloud consists of a homogeneous set of hardware and software resources.
- The resources are in a single administrative domain (AD). Security, resource management, fault-tolerance, and quality of service are less challenging than in a heterogeneous environment with resources in multiple ADs.

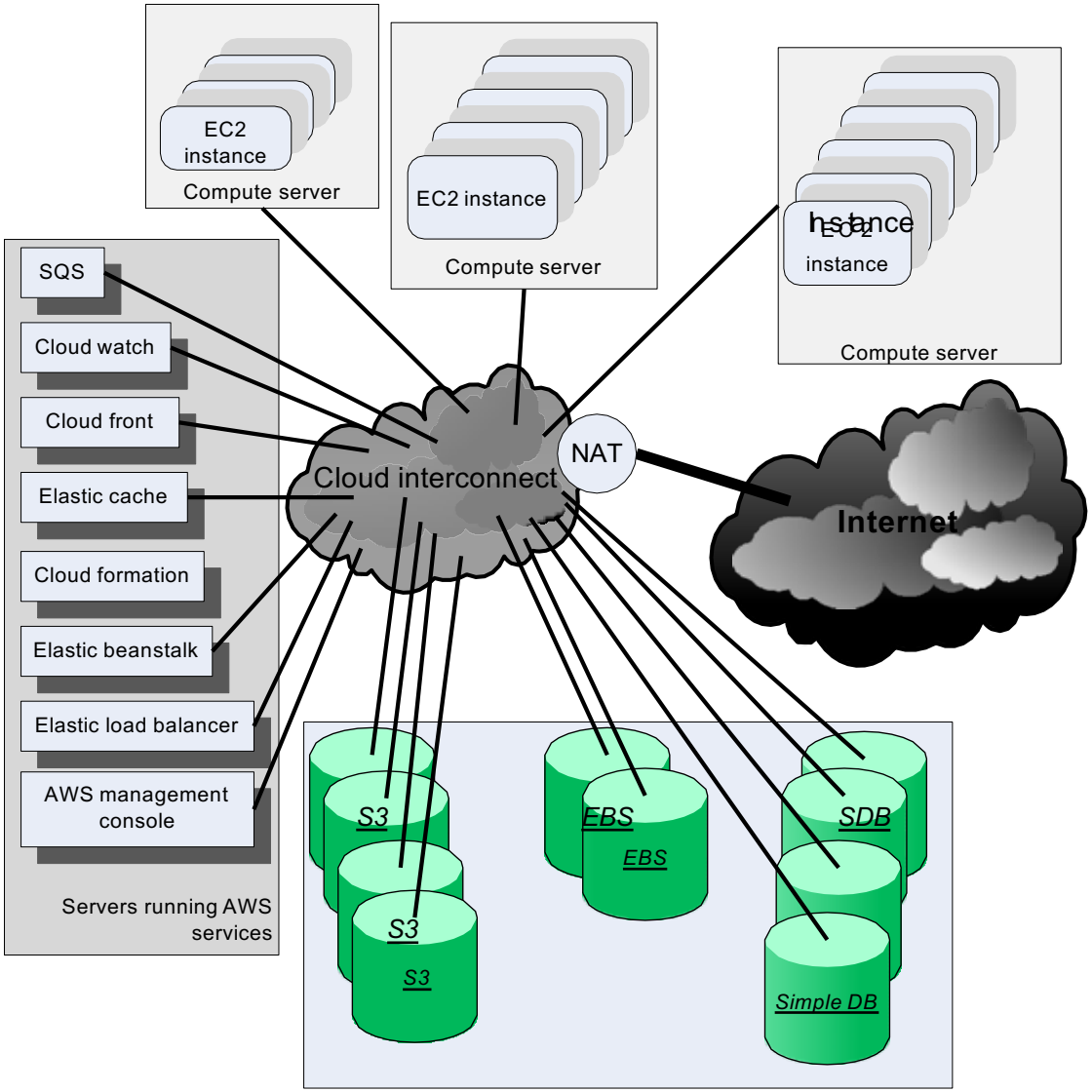
Challenges for cloud computing

- Availability of service; what happens when the service provider cannot deliver?
- Diversity of services, data organization, user interfaces available at different service providers limit user mobility; once a customer is hooked to one provider it is hard to move to another. Standardization efforts at NIST!
- Data confidentiality and auditability, a serious problem.
- Data transfer bottleneck; many applications are data-intensive.

More challenges

- Performance unpredictability, one of the consequences of resource sharing.
  - How to use resource virtualization and performance isolation for QoS guarantees?
  - How to support elasticity, the ability to scale up and down quickly?
- Resource management; is self-organization and self-management a solution?
- Security and confidentiality; major concern.
- Addressing these challenges provides good research opportunities!!


Delivery models
Software as a Service (SaaS)
Platform as a Service (PaaS)
Infrastructure as a Service (IaaS)
Deployment models
Public cloud
Private cloud
Community cloud
Hybrid cloud
Cloud computing
Infrastructure
Distributed infrastructure
Resource virtualization
Autonomous systems
Resources
Compute & storage servers
Networks
Services
Applications
Defining attributes
Massive infrastructure
Utility computing. Pay-per-usage
Accessible via the Internet
Elasticity


EC2 instance
Compute server
EC2 instance
Compute server
EC2 instance
Compute server
SQS
Cloud watch
Cloud front
Elastic cache
Cloud formation
Elastic beanstalk
Elastic load balancer
AWS management console
Servers running AWS services
Cloud interconnect
NAT
Internet
S3
S3
S3
EBS
EBS
SDB
Simple DB

Warehouse-scale computer (WSC)

- WSCs provide Internet services
  - Search and Email,
  - Social networking,
  - Online maps,
  - Video sharing,
  - Online shopping,
  - Cloud computing, etc.
- WSCs versus HPC “clusters”:
  1. Clusters have higher performance processors and network
  2. Clusters emphasize *thread-level parallelism*, WSCs emphasize *request-level parallelism*
- WSCs versus datacenters:
  1. Datacenters consolidate different machines and software into one location
  2. Datacenters emphasize virtual machines and hardware heterogeneity in order to serve varied customers

Request-level parallelism

- Popular Internet services like Google search
- Hundreds or thousands of requests per second
  - Requests are largely independent
  - Mostly involve read-only databases
  - Producer-consumer sharing
  - Rarely involve read-write data sharing or synchronization across requests
  - Computations easily partitioned within a request and across different requests
- However, it is observed in HPC as a Service model
  - SCALAC Model
  - PRACE Model

WSCs design factors

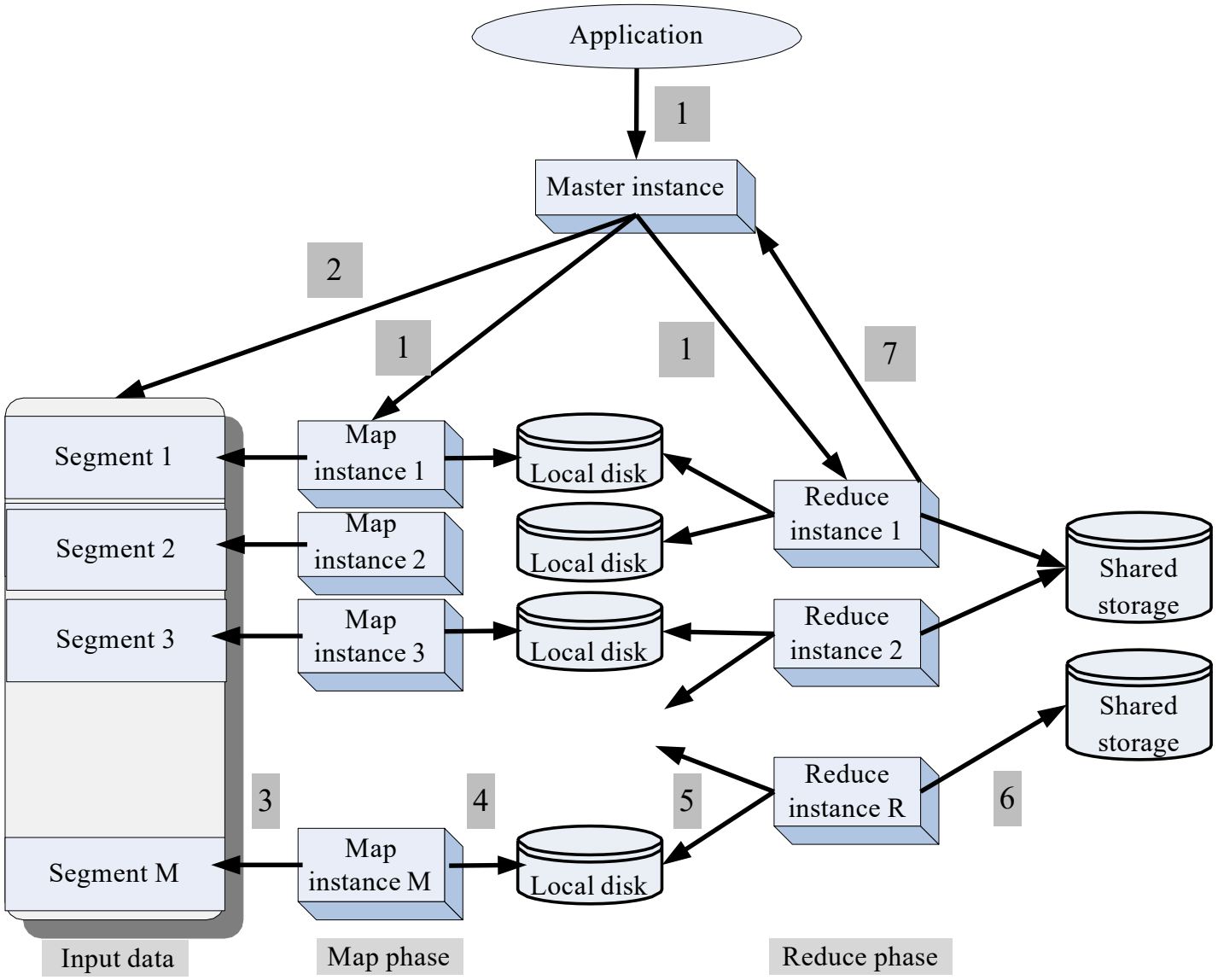
1. Cost-performance → small savings add up
2. Energy efficiency → Affects power distribution and cooling
3. Operational costs count → Power consumption is a primary, not secondary, constraint when designing system
4. Dependability via redundancy
5. Network I/O
6. Interactive and batch processing workloads
7. Ample computational parallelism is not important → Most jobs are totally independent
8. Scale and its opportunities and problems → Can afford to build customized systems since WSC require volume purchase

Programming models

- 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

MapReduce philosophy

1. An application starts:
   - A master instance;
   - M worker instances for the *Map phase,* and later
   - R worker instances for the *Reduce phase*.
2. The *master* instance partitions the input data in M *segments*.
3. A *map* instance reads its input data segment and processers the data.
4. The results of the processing are stored on the local disks of the servers where the map instances run.
5. When all *map* instances have finished processing their data the R *reduce* instances read the results of the first phase and merges the partial results.
6. The final results are written by the *reduce* instances to a shared storage server.
7. The *master* instance monitors the reduce instances and when all of them report task completion the application is terminated.


Application
1
Master instance
2
1
1
7
Segment 1
Segment 2
Segment 3
Segment M
Map instance 1
Map instance 2
Map instance 3
Map instance M
Local disk
Local disk
Local disk
Local disk
Reduce instance 1
Reduce instance 2
Reduce instance R
Shared storage
Shared storage
3
4
5
6
Input data
Map phase
Reduce phase

MapReduce 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));

MapReduce jobs at Google

	Aug-04	Mar-06	Sep-07	Sep-09
Number of MapReduce jobs	29,000	171,000	2,217,000	3,467,000
Average completion time (seconds)	634	874	395	475
Server years used	217	2002	11,081	25,562
Input data read (terabytes)	3288	52,254	403,152	544,130
Intermediate data (terabytes)	758	6743	34,774	90,120
Output data written (terabytes)	193	2970	14,018	57,520
Average number of servers per job	157	268	394	488

Figure 6.2 Annual MapReduce usage at Google over time. Over five years the number of MapReduce jobs increased by a factor of 100 and the average number of servers per job increased by a factor of 3. In the last two years the increases were factors of 1.6 and 1.2, respectively [Dean 2009]. Figure 6.16 on page 459 estimates that running the 2009 workload on Amazon's cloud computing service EC2 would cost $133M.

	Aug-04	Mar-06	Sep-07	Sep-09
Average completion time (hours)	0.15	0.21	0.10	0.11
Average number of servers per job	157	268	394	488
Cost per hour of EC2 High-CPU XL instance	$0.68	$0.68	$0.68	$0.68
Average EC2 cost per MapReduce job	$16.35	$38.47	$25.56	$38.07
Average number of EBS I/O requests (millions)	2.34	5.80	3.26	3.19
EBS cost per million I/O requests	$0.10	$0.10	$0.10	$0.10
Average EBS I/O cost per MapReduce job	$0.23	$0.58	$0.33	$0.32
Average total cost per MapReduce job	$16.58	$39.05	$25.89	$38.39
Annual number of MapReduce jobs	29,000	171,000	2,217,000	3,467,000
Total cost of MapReduce jobs on EC2/EBS	$480,910	$6,678,011	$57,394,985	$133,107,414

Figure 6.16 Estimated cost if you ran the Google MapReduce workload (Figure 6.2) using 2011 prices for AWS ECS and EBS (Figure 6.15). Since we are using 2011 prices, these estimates are less accurate for earlier years than for the more recent ones.

MapReduce runtime environment

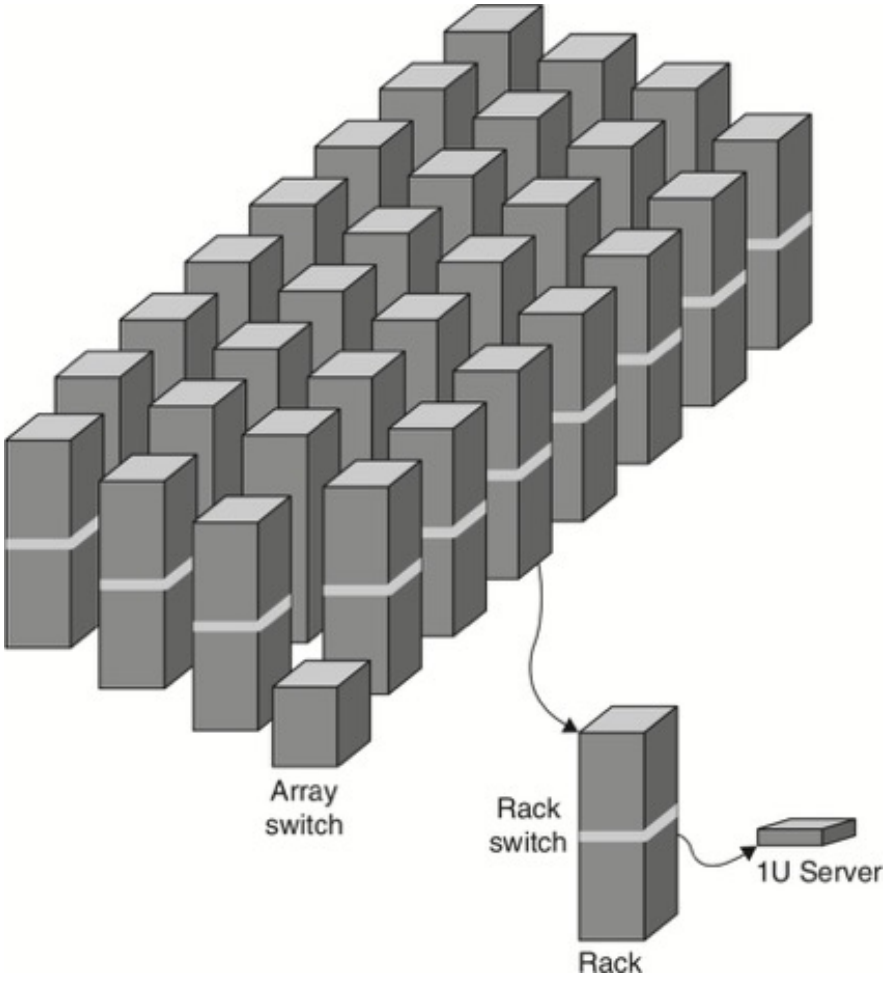
- Schedules map and reduce task to WSC nodes
- Availability:
  - Use replicas of data across different servers
  - Use relaxed consistency → No need for all replicas to always agree
- Workload demands
  - Often vary considerably

WSC architecture

- Hierarchy of networks for interconnection
- Each 19” rack holds 48 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

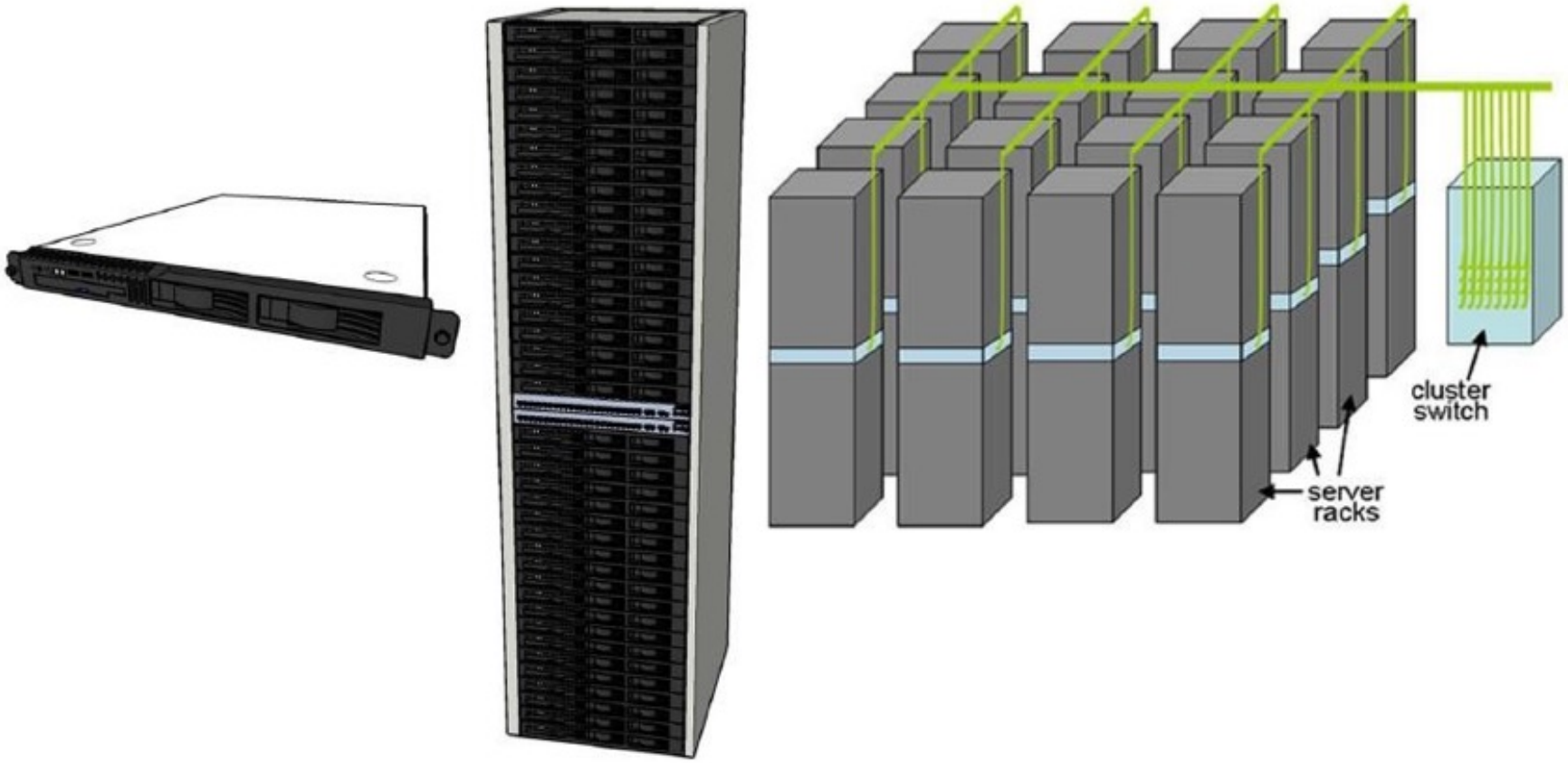
- WSCs often use a hierarchy of networks for interconnection



Computer Architecture of WSC

- The 19-inch (48.26-cm) rack is the standard framework to hold servers
- Each 19-inch rack holds 48 1U servers connected to a rack switch
  - Bandwidth within the rack is the same for each server
  - It does not matter where the software places the sender and the receiver
- 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

WSC organization



Storage

- Storage options:
  - Use disks inside the servers, or
  - Network attached storage (NAS) through Infiniband
    - NAS solution is generally more expensive per terabyte of storage, but it provides many features, including reliability
  - WSCs generally rely on local disks
  - Google File System (GFS) uses local disks and maintains at least three replicas

WSC Memory Hierarchy

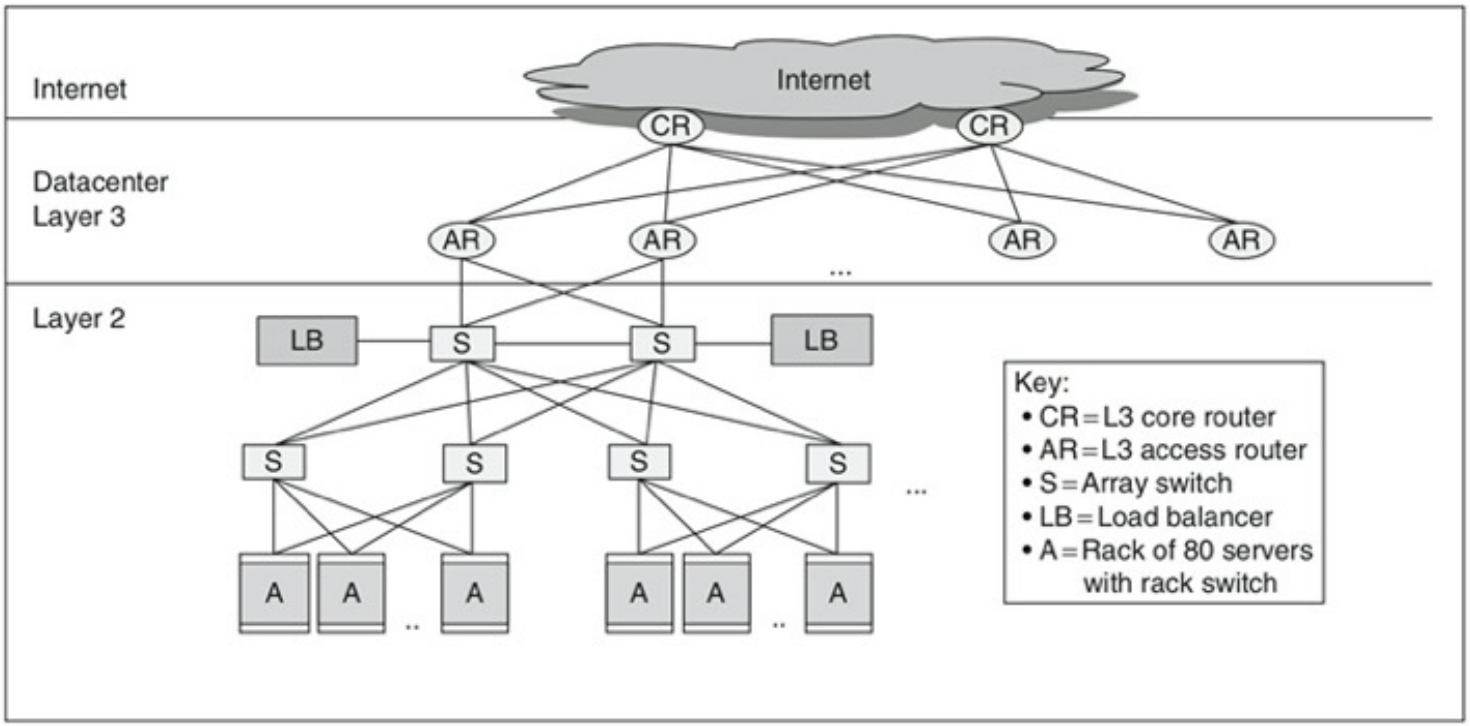
- Servers can access DRAM and disks on other servers using a NUMA-style interface
  - Each server has 16GB of memory
  - Every pair of racks includes one rack switch and holds 80 2U servers
  - The array switch can handle 30 racks

	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

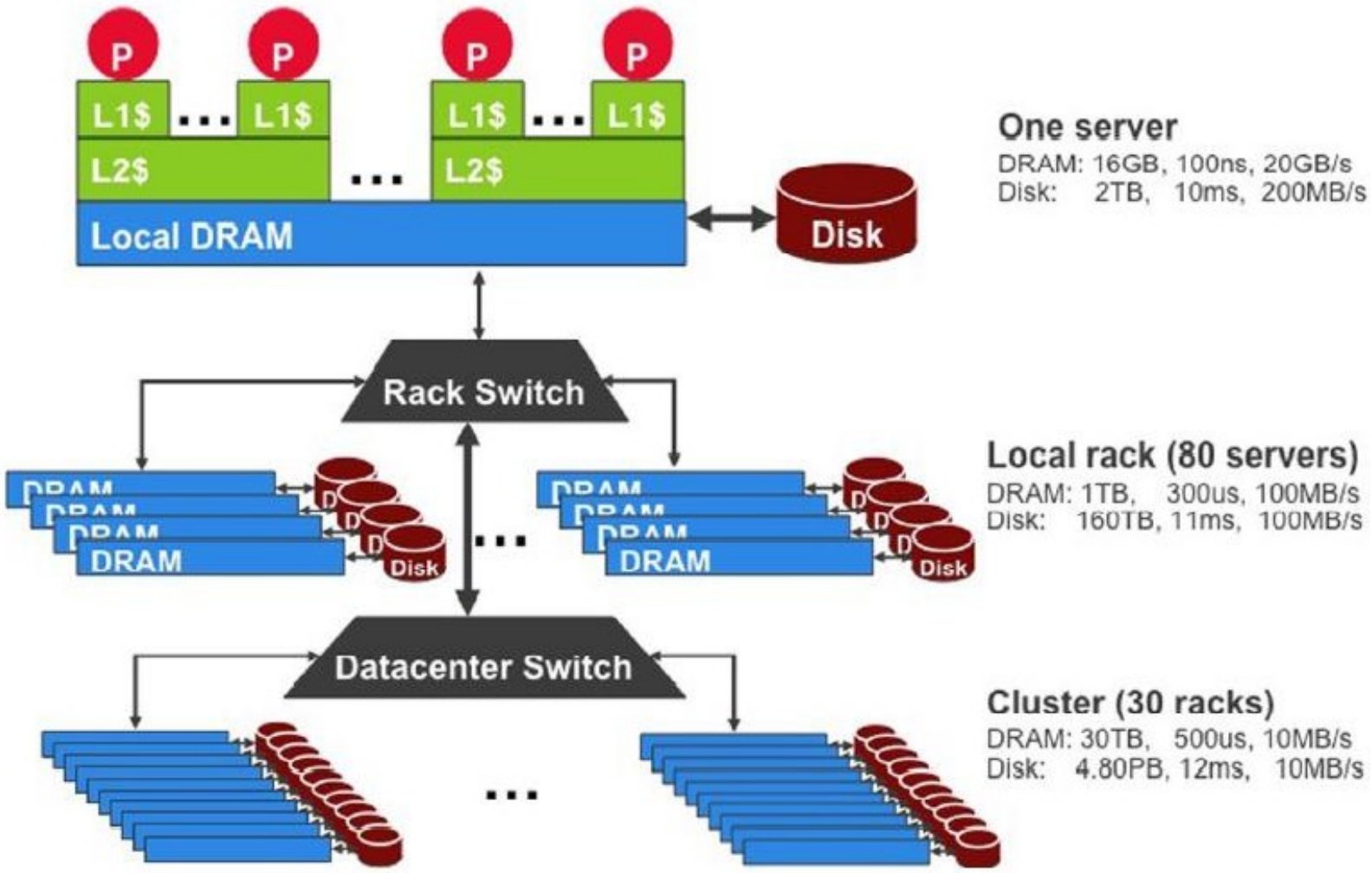
WSC Memory Hierarchy (2)

- Network overhead increases latency from local DRAM to rack DRAM and array DRAM, but still have more than 10X better latency than local disk

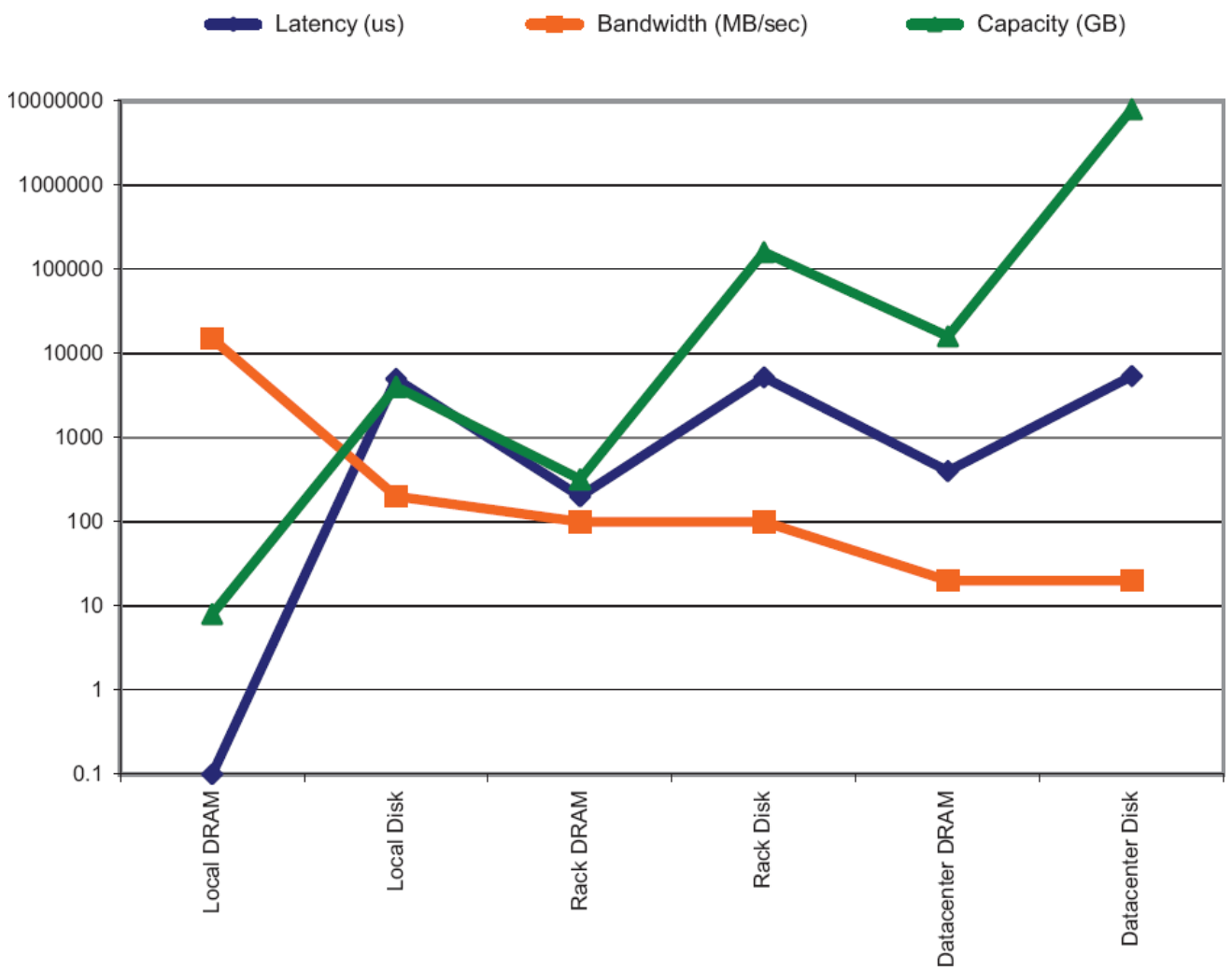
- WSC needs 20 arrays to reach 50000 servers, one more level of the networking hierarchy is added.



WSC storage hierarchy



Storage latency, bandwidth, capacity

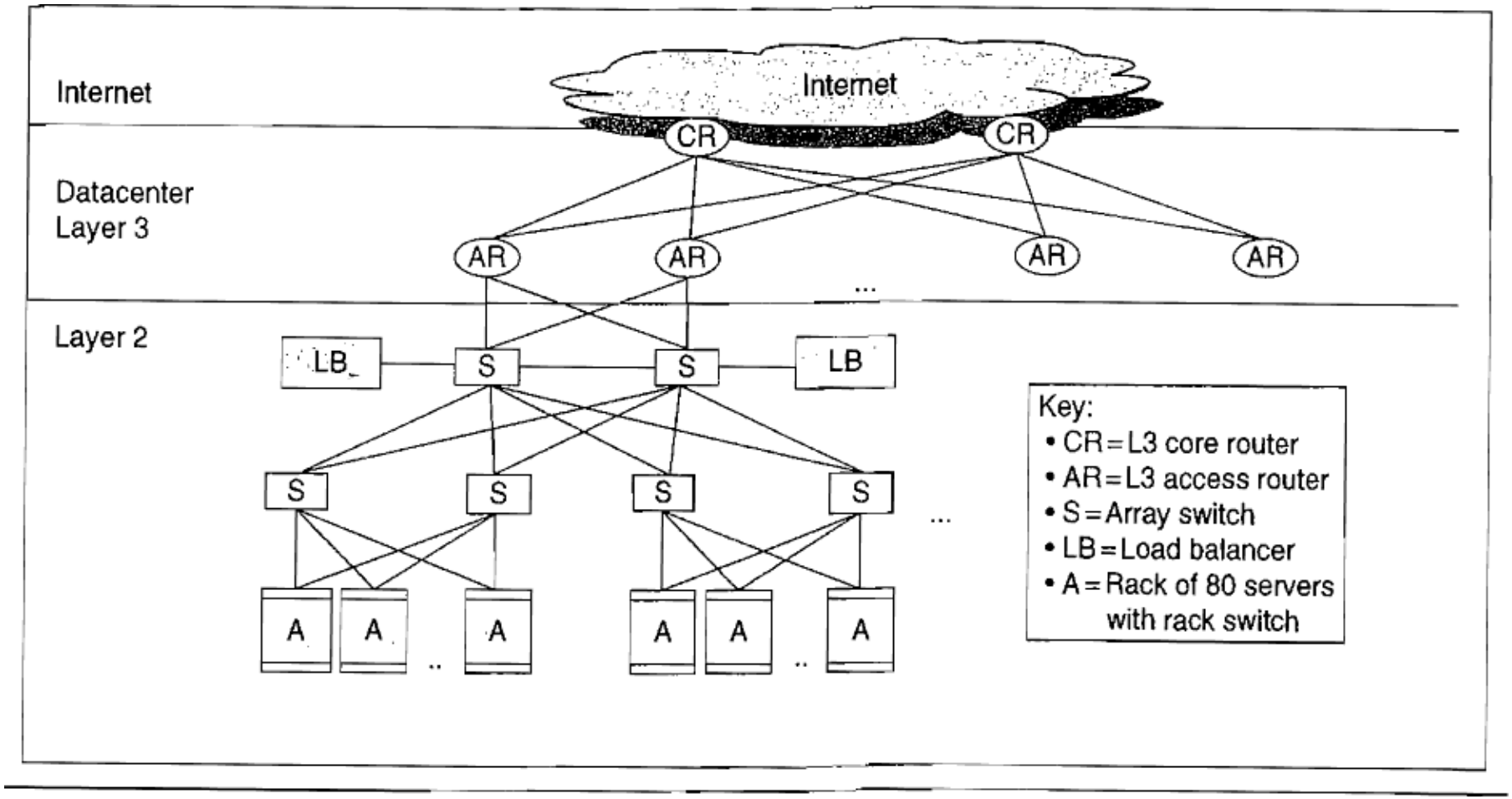


Array switch

- Switch that connects an array of racks
- Should have 10 X the bisection bandwidth of rack switch
- The cost of *n*-port switch grows as n^2
- Often utilize content addressable memory chips and FPGAs (Field Programmable Gate Arrays)

- And Also in other hybrid computing platforms
  - GPU/Accelerators envioroments

Remember Cloud networking infrastructure



Server hardware configuration

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

- TPC-C is an on-line transaction processing benchmark http://www.tpc.org/tpcc/

Communication intensity

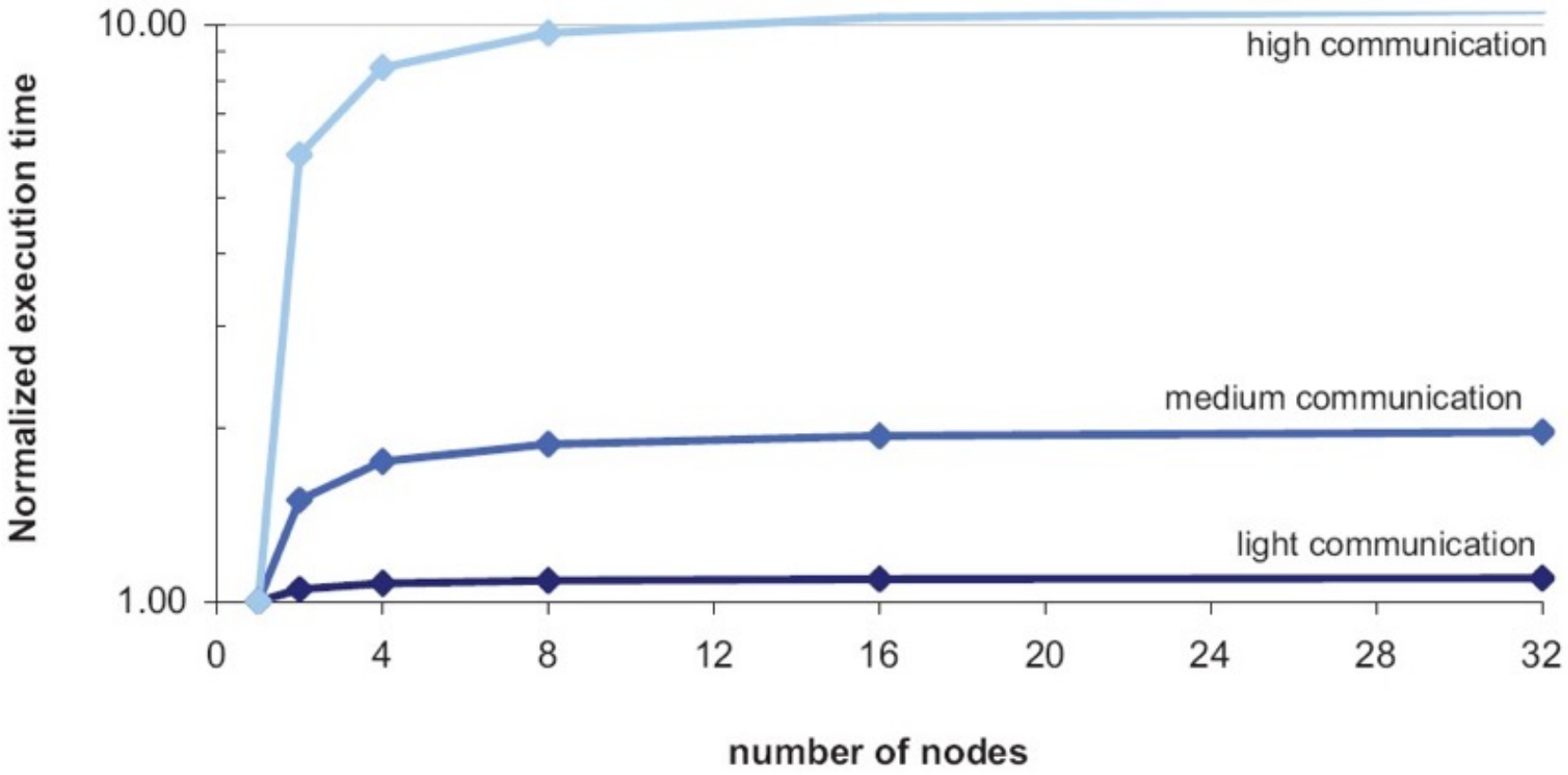


FIGURE 3.1: Execution time of parallel tasks as the number of SMP nodes increases for three levels of communication intensity. Execution time is normalized to the single node case and plotted in logarithmic scale.

Performance advantage

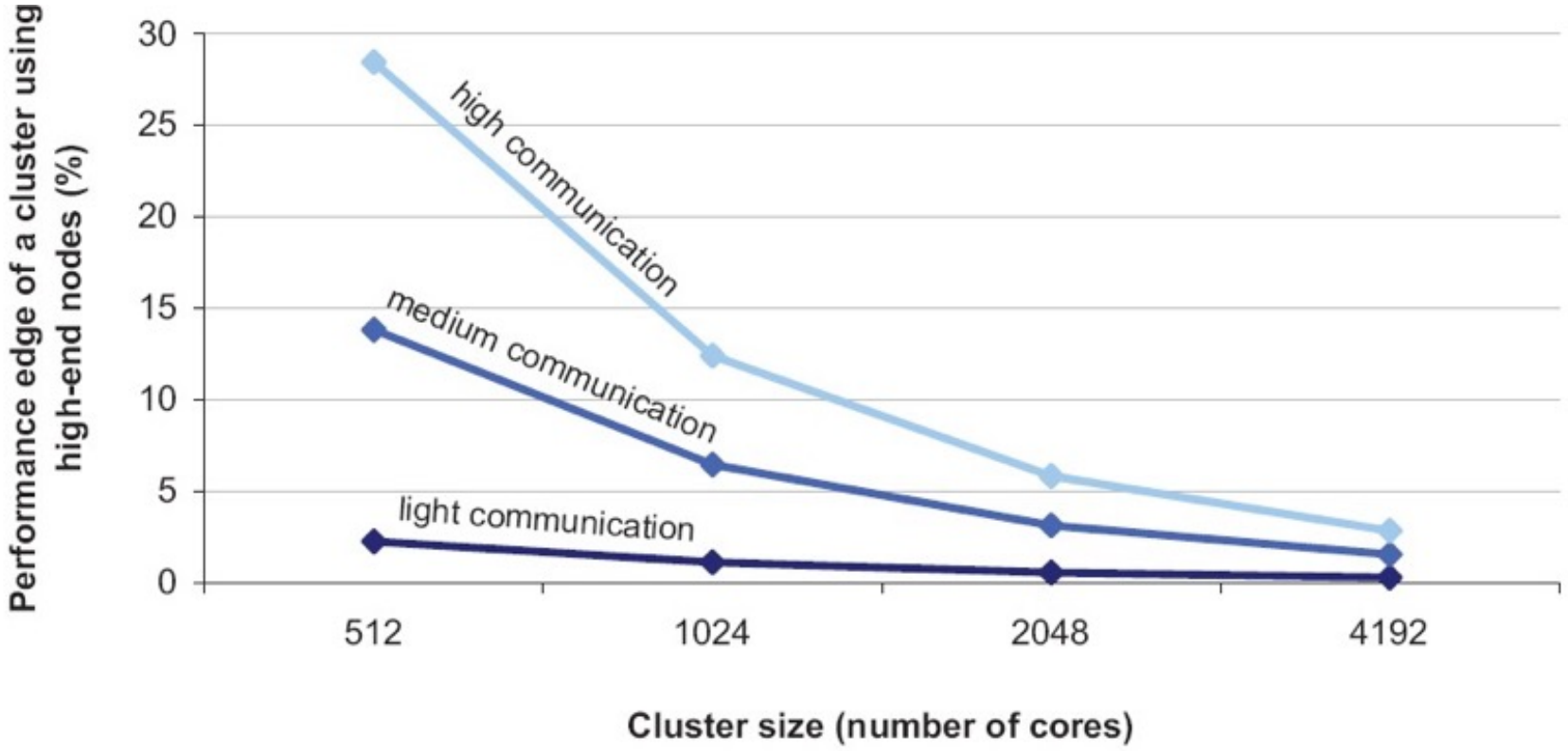


FIGURE 3.2: Performance advantage of a cluster built with high-end server nodes (128-core SMP) over a cluster with the same number of processor cores built with low-end server nodes (four-core SMP), for clusters of varying size.

Scientific applications and WSC

- Applications in computational science and engineering typically exhibit fine-grain parallelism.
- Do not perform well on WSCs!!
- WSC are designed to support request-level rather than tread-level parallelism

- However, there are interesting exemples
  - French Grid5000 Project www.grid5000.fr

Failure rates and WSC reliability

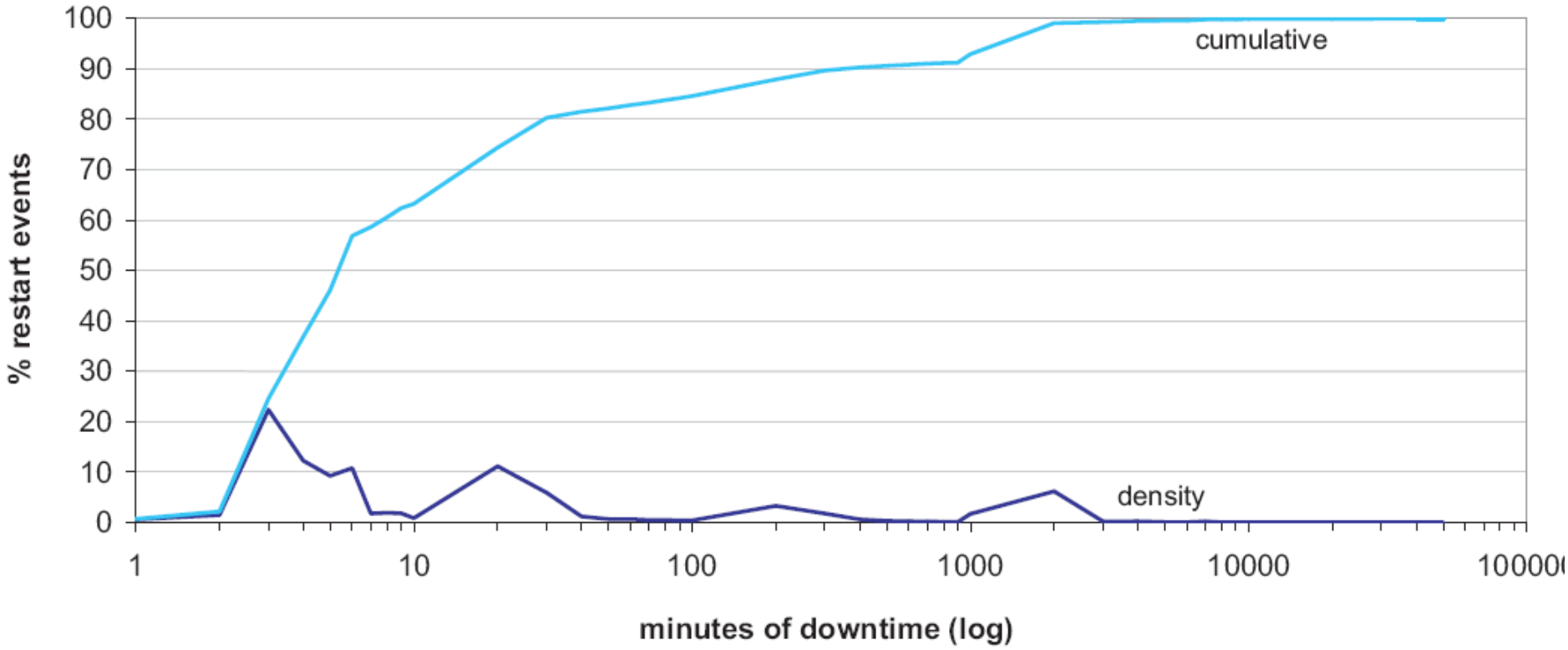
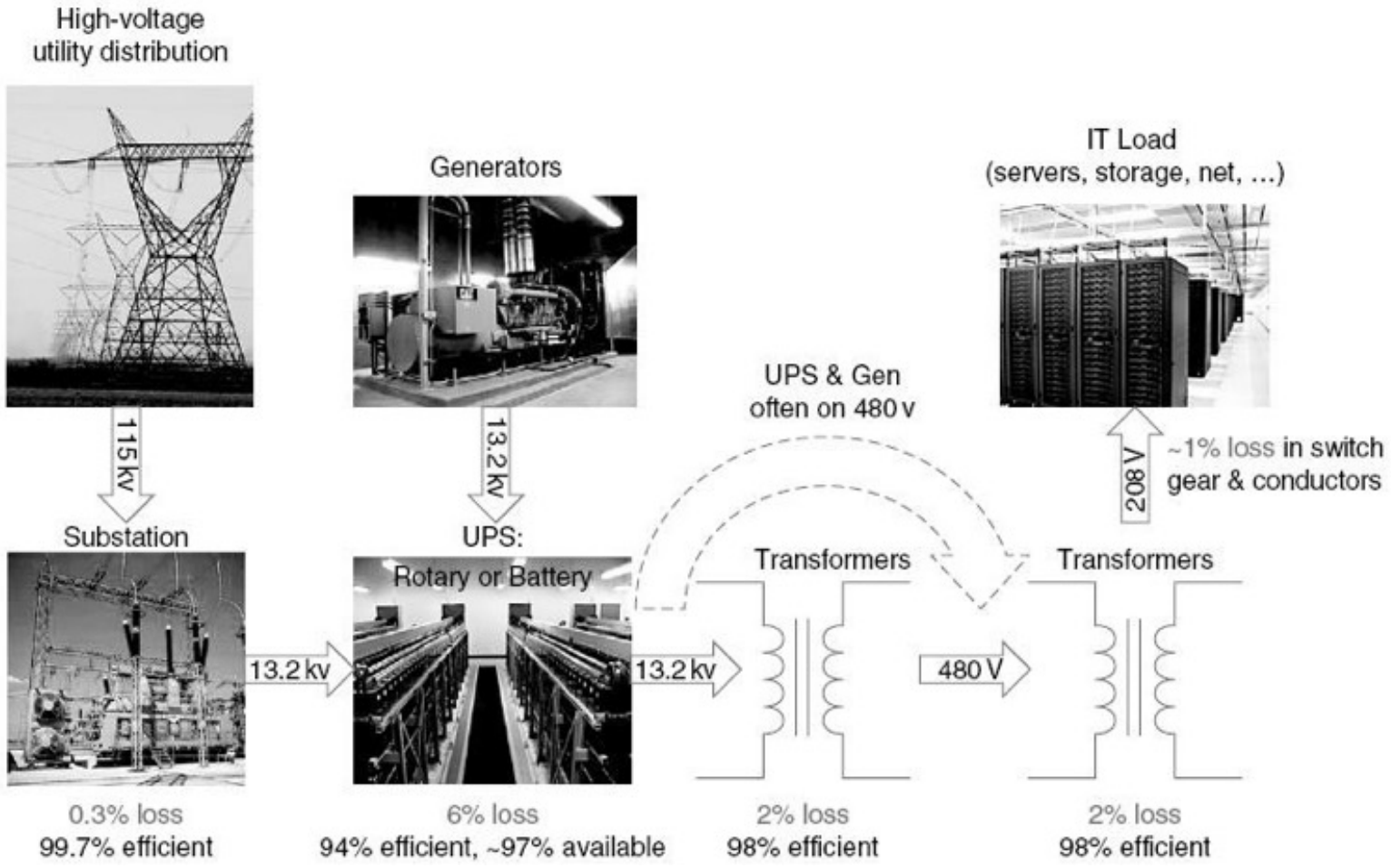


FIGURE 7.3: Distribution of machine downtime, observed at Google over 6 months. The average annualized restart rate across all machines is 4.2, corresponding to a mean time between restarts of just less than 3 months.

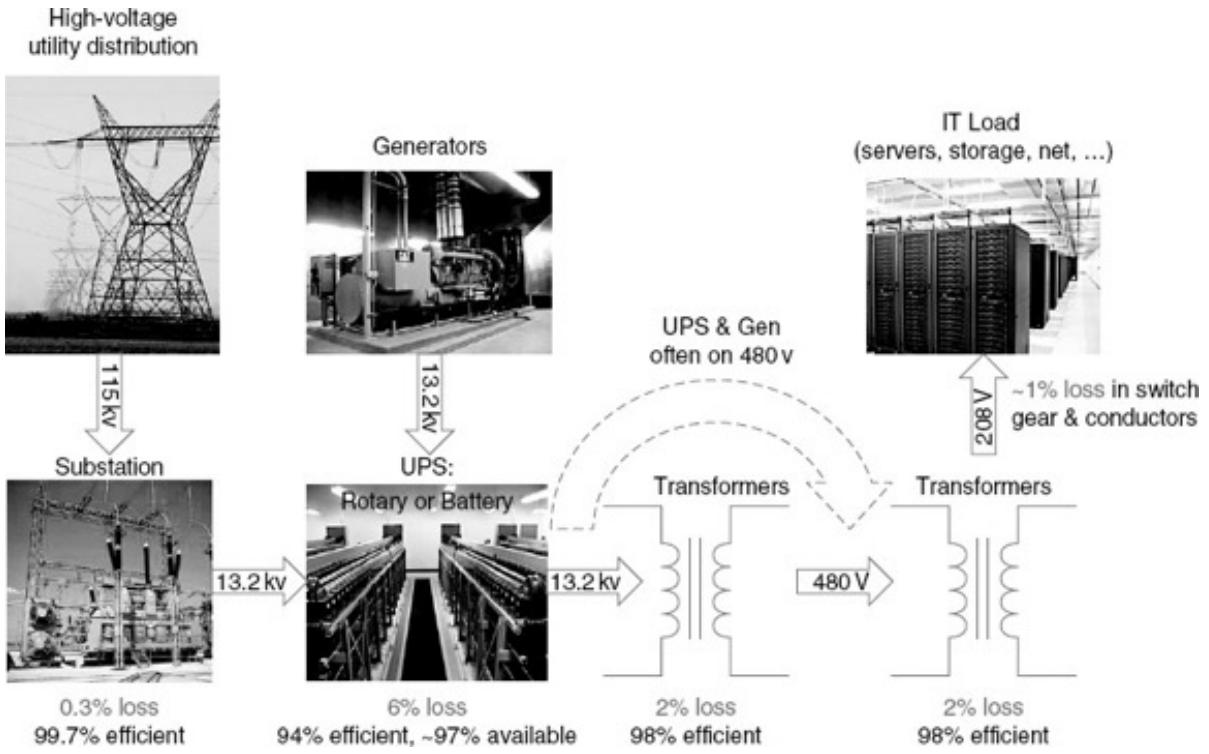
WSC infrastructure

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



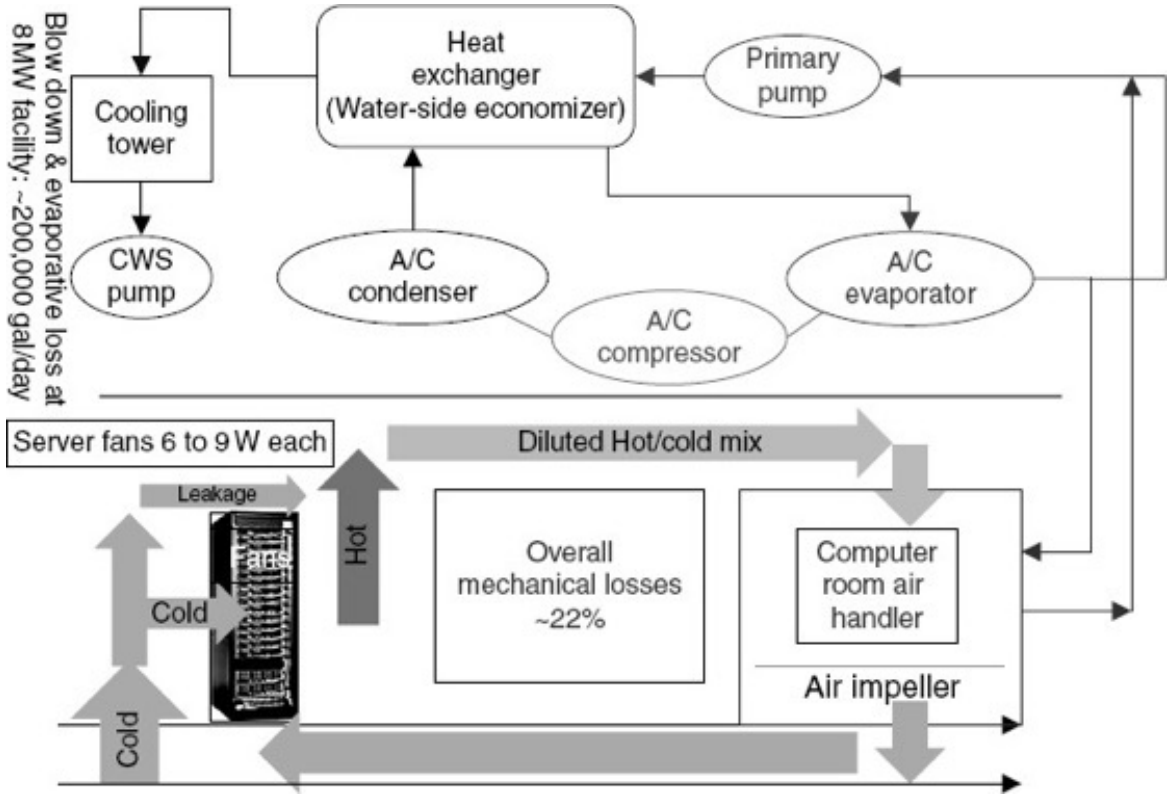
Infrastructure and Costs of WSC

- Power distribution
  - UPS: uninterruptible power supply. Located in a separate room from the IT equipment
    - Power conditioning, holding the electrical load when switching



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
  - Air conditioning used to cool server room using chilled water
  - 18 C – 22 C
    - Keep temperature higher (closer to 22 C)
  - Fans push warm air past a set of coils filled with cold water and a pump moves the warmed water to the external chillers to be cooled down.

The In-House Anecdote:
SC3UIS-PTG Fishbowl (BAD) Exemple

WSC cooling

- 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

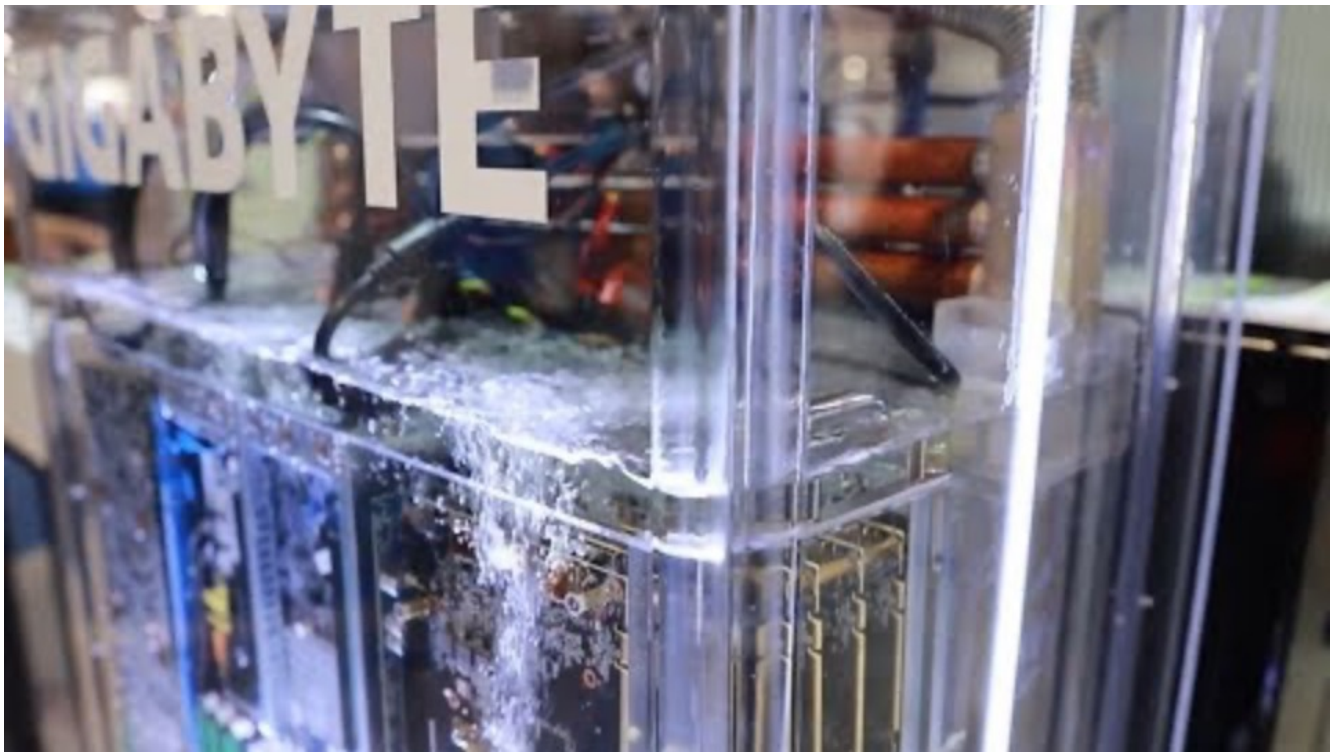
WSC efficiency

- Power Utilization Effectiveness (PUE) → Total facility power / IT equipment power. Median PUE on 2006 study was 1.69
- Performance
  - Latency → important metric because it is seen by users
  - Service Level Agreements (SLAs) E.g. 99% of requests be below 100 ms
  - 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%

Infrastructure and Costs of WSC

- Cooling towers can also be used
  - Leverage the colder outside air to cool the water before it is sent to the chillers
  - Minimum temperature is "wet bulb temperature"
  - Warm water flows over a large surface in the tower, transferring heat to the outside air via evaporation and thereby cooling the water. This technique is called airside economization.
  - An alternative is use cold water instead of cold air.
    - Google's WSC in Belgium uses a water-to-water intercooler that takes cold water from an industrial canal to chill
  - Weather Environment also should be used
    - FrigID Data weather Exchange (LIG – Project)



Two Phase inmersion liquid cooling by
https://www.gigabyte.com/Solutions/Cooling/immersion-cooling

Infrastructure and Costs of WSC

- Airflow is carefully planned. Efficient designs preserve the temperature of the cool air by reducing the chances of it mixing with hot air.

- Cooling system also uses water (evaporation and spills)
  - E.g. 70,000 to 200,000 gallons per day for an 8 MW facility

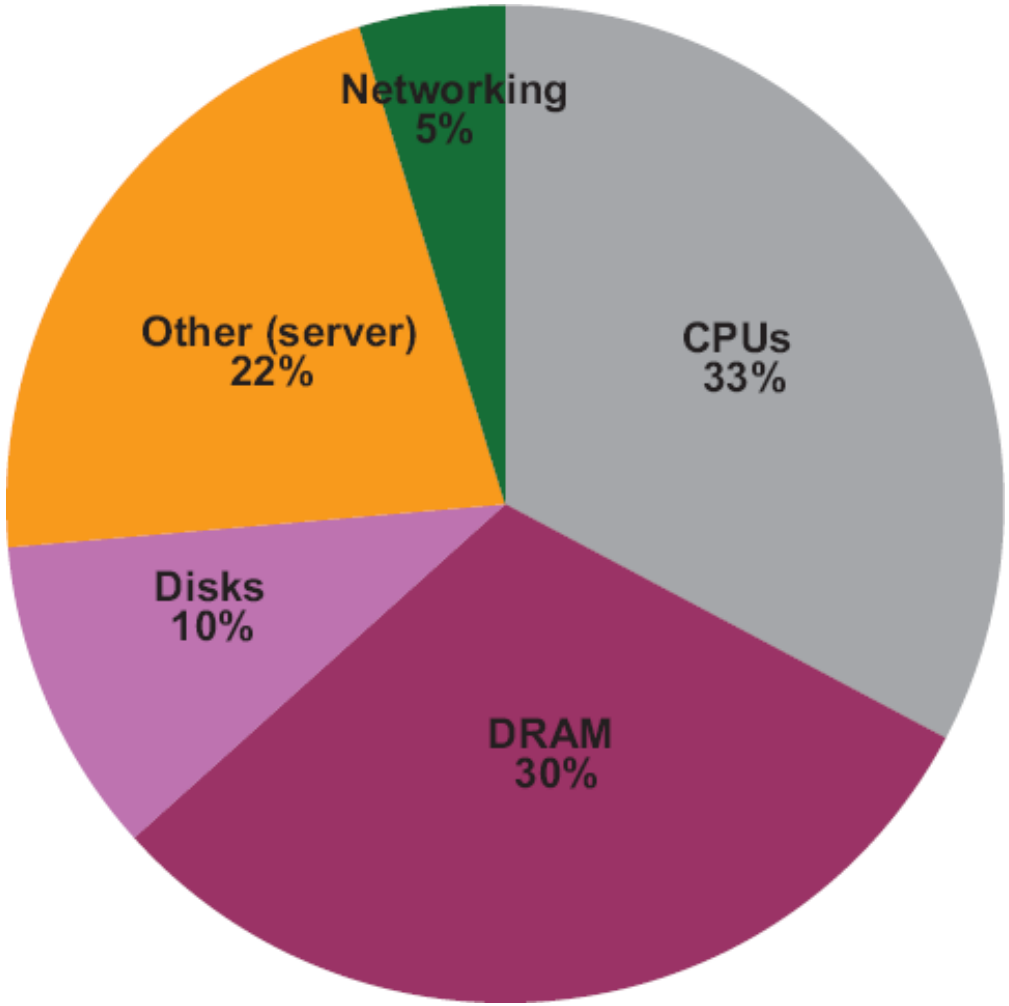
- Power cost breakdown:
  - Chillers: 30-50 % of the power used by the IT equipment
  - Air conditioning: 10-20 % of the IT power, mostly due to fans

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

Power utilization



(Remember) Measuring Efficiency of a WSC

- Power Utilization Effectiveness (PEU)
  - = Total facility power / IT equipment power
  - Median PUE on 2006 study was 1.69
  - The bigger the PUE, the less efficient the WSC

- Performance
  - Latency is important metric because it is seen by users
  - Cutting system response time 30% shaved the time of an inter- action by 70%.

Measuring Efficiency of a WSC

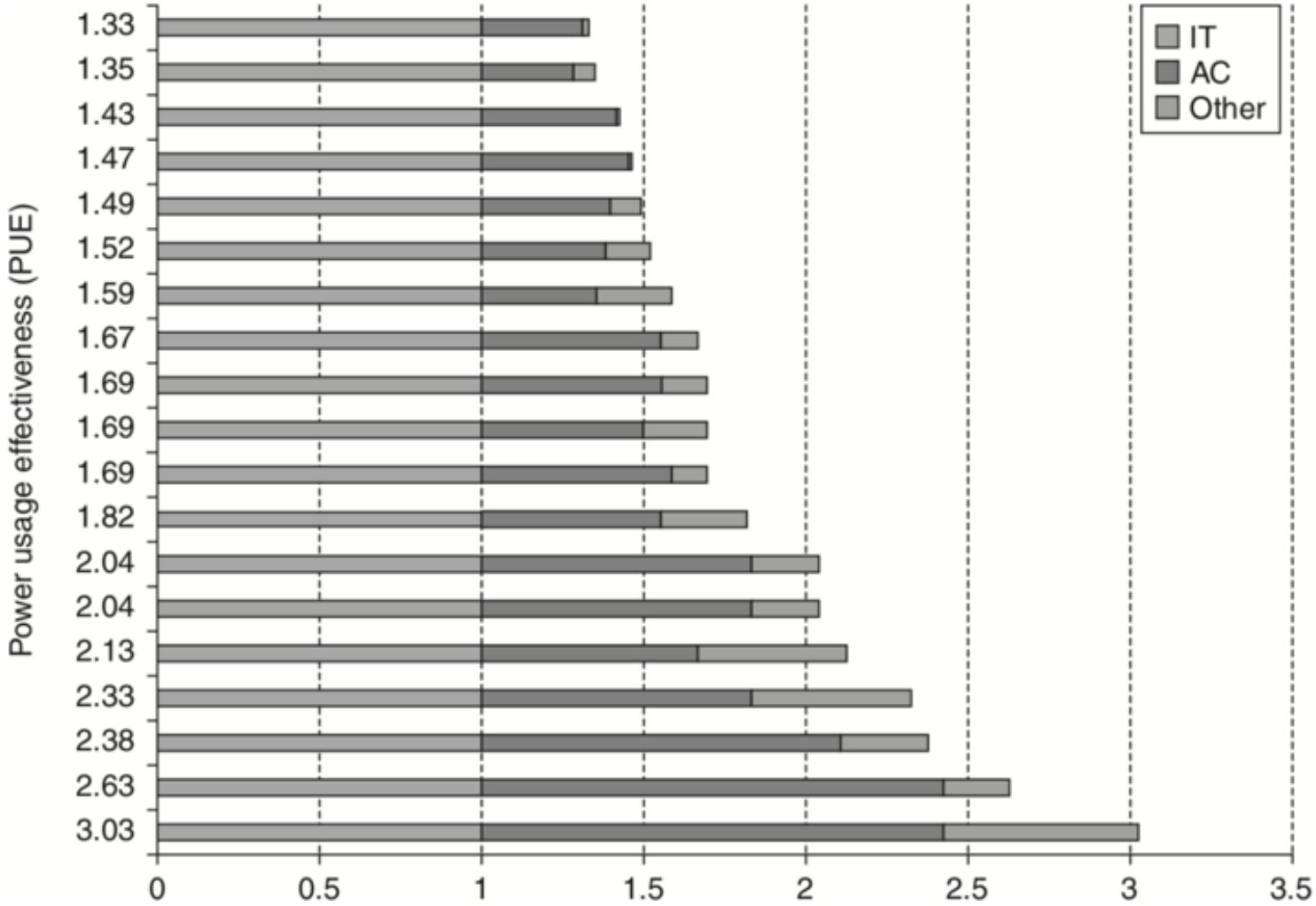
- Bing study: users will use search less as response time increases

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%

- A high percentage of requests be below a latency threshold rather just offer a target for the average latency.

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

Measuring Efficiency of a WSC



Cost of a WSC

- Capital expenditures (CAPEX)
  - Cost to build a WSC
- Operational expenditures (OPEX)
  - Cost to operate a WSC
    - Energy
    - Locations
    - Maintenance
    - Manware (Human Ressources)

Energy efficiency

$$\text{Efficiency} = \frac{\text{Computation}}{\text{Total Energy}} = \underbrace{\left(\frac{1}{\text{PUE}}\right)}_{\text{(a)}} \times \underbrace{\left(\frac{1}{\text{SPUE}}\right)}_{\text{(b)}} \times \underbrace{\left(\frac{\text{Computation}}{\text{Total Energy to Electronic Components}}\right)}_{\text{(c)}}$$

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).

Power utilization effectiveness (PUE)

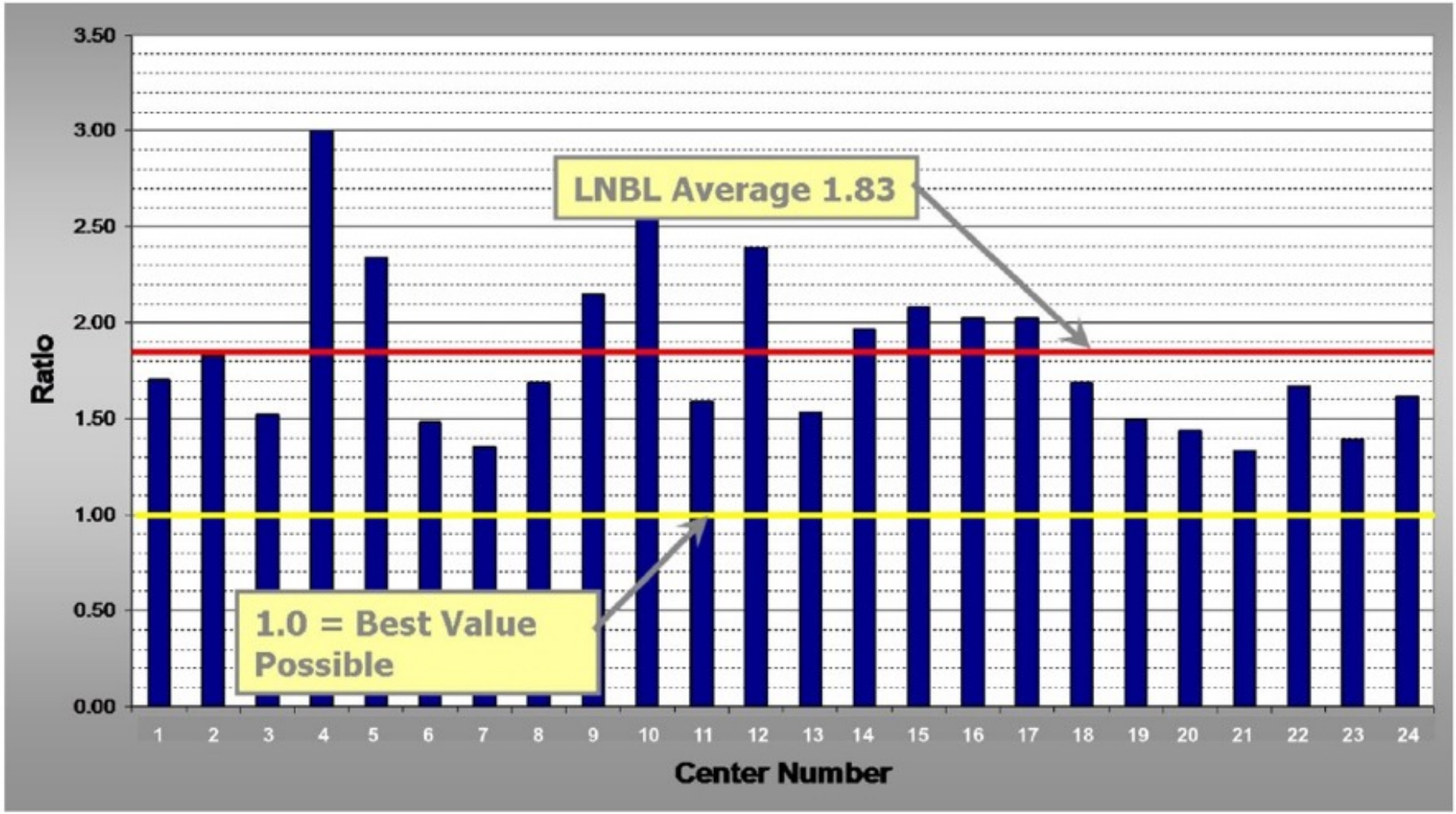


FIGURE 5.1: LBNL survey of the power usage efficiency of 24 datacenters, 2007 (Greenberg et al.)

Problem

- In the US the cost of electricity is in the range $0.03 – 0.15 per kWh.
- Assuming
  - Critical load = 8 MW
  - PUE = 1.45
  - Average power usage = 80% (0.8)
- What is the impact of hourly server cots for the two extremes?

Solution

- The average power usage is

  8 x 0.8 x 1.45 = 9.29MW

  - The hourly costs for the low range is

    9.29 MW x $0.03 = $205,000

  - The hourly costs for the high range is

    9.29 M x $0.15 = $1,015,000

A benchmark

FIGURE 5.3: An example benchmark result for SPECpower_ssj2008; energy efficiency is indicated by bars, whereas power consumption is indicated by the line. Both are plotted for a range of utilization levels, with the average metric corresponding to the vertical dark line. The system has a single-chip 2.83 GHz quad-core Intel Xeon processor, 4 GB of DRAM, and one 7.2 k RPM 3.5″ SATA disk drive.

Energy efficiency

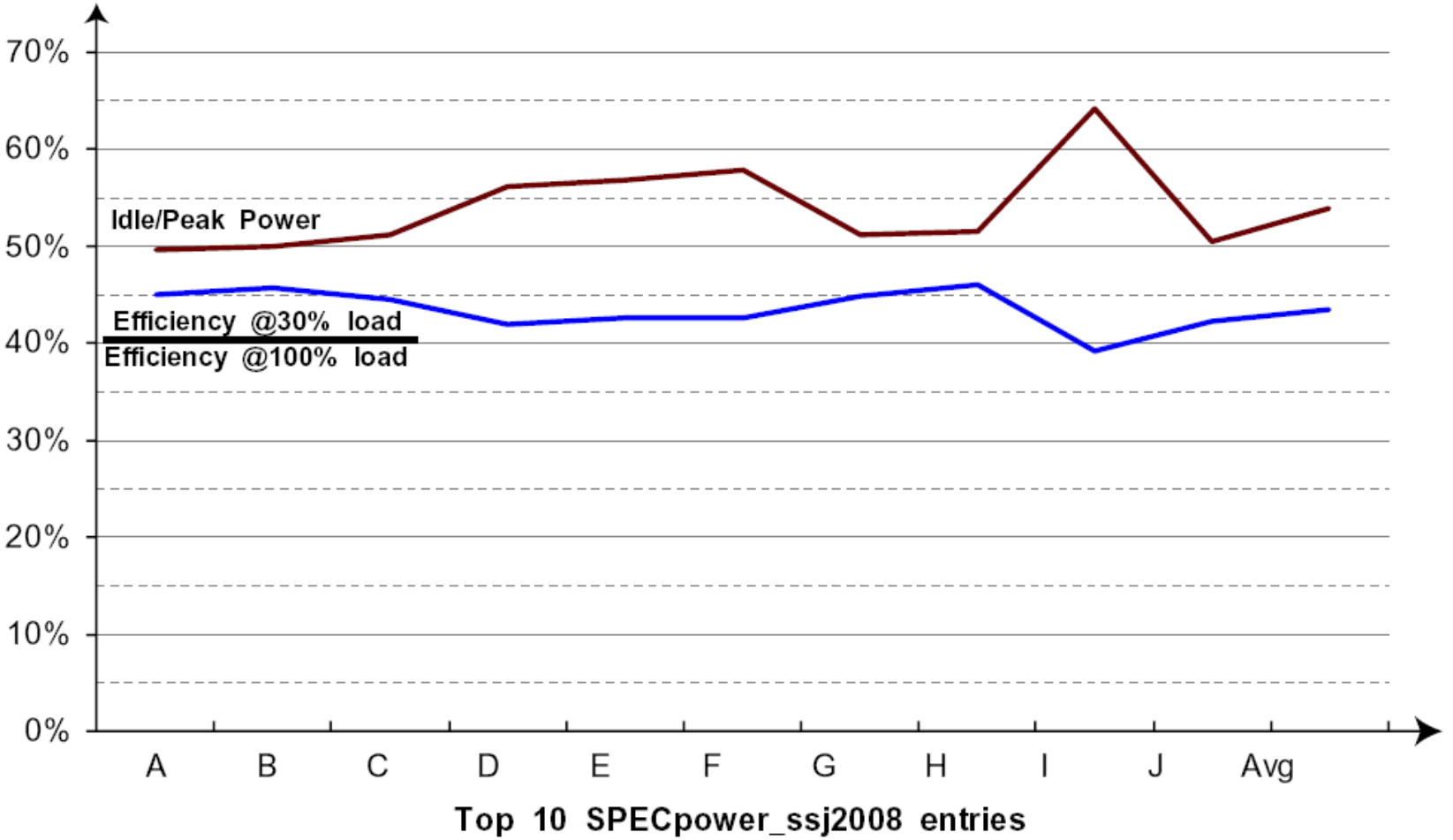


FIGURE 5.4: Idle/peak power and energy efficiency at 30% load (relative to the 100% load efficiency) of the top 10 SPECpower_ssj2008 entries. (data from mid 2008)

Activity profile

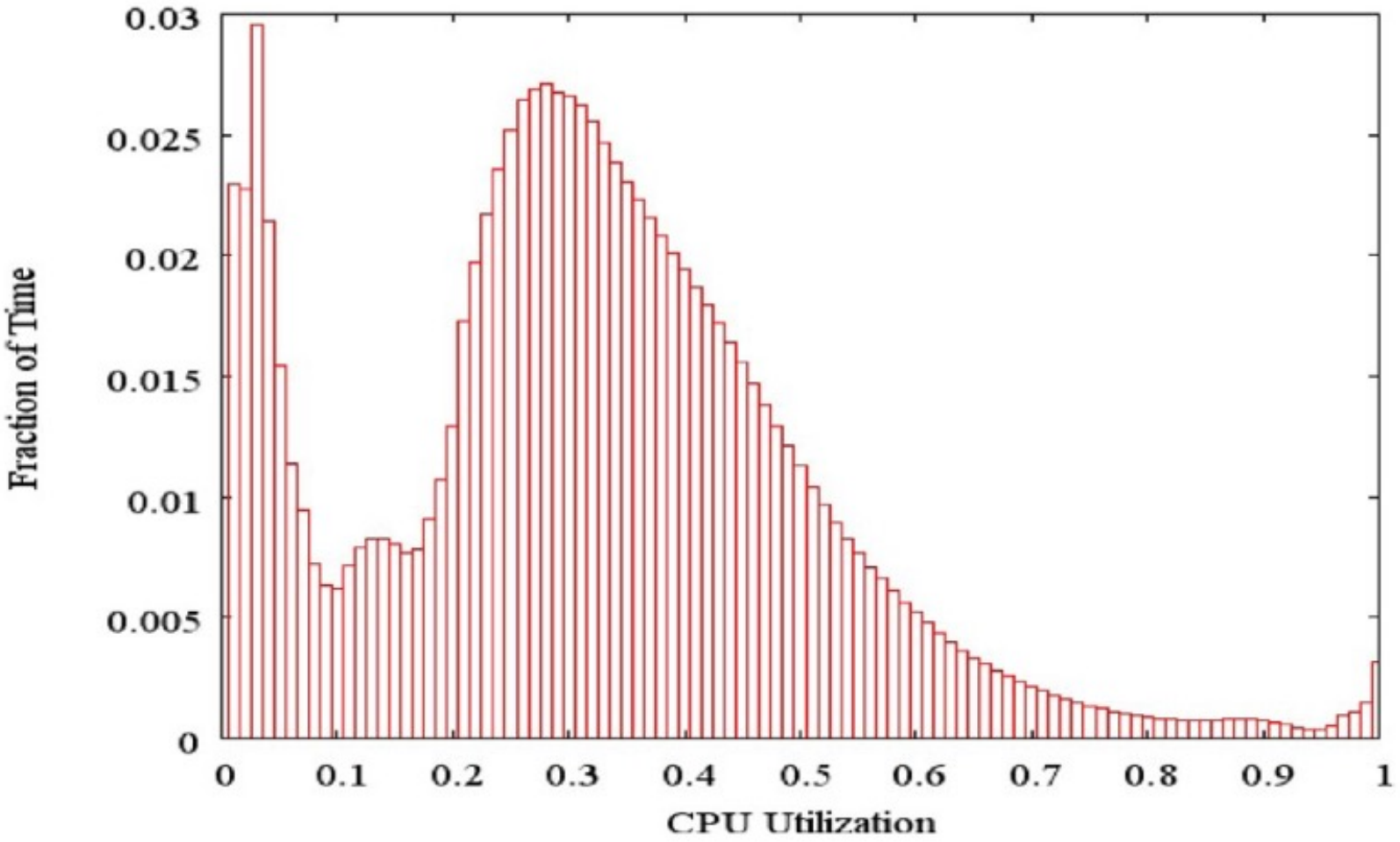


FIGURE 5.5: Activity profile of a sample of 5,000 Google servers over a period of 6 months.

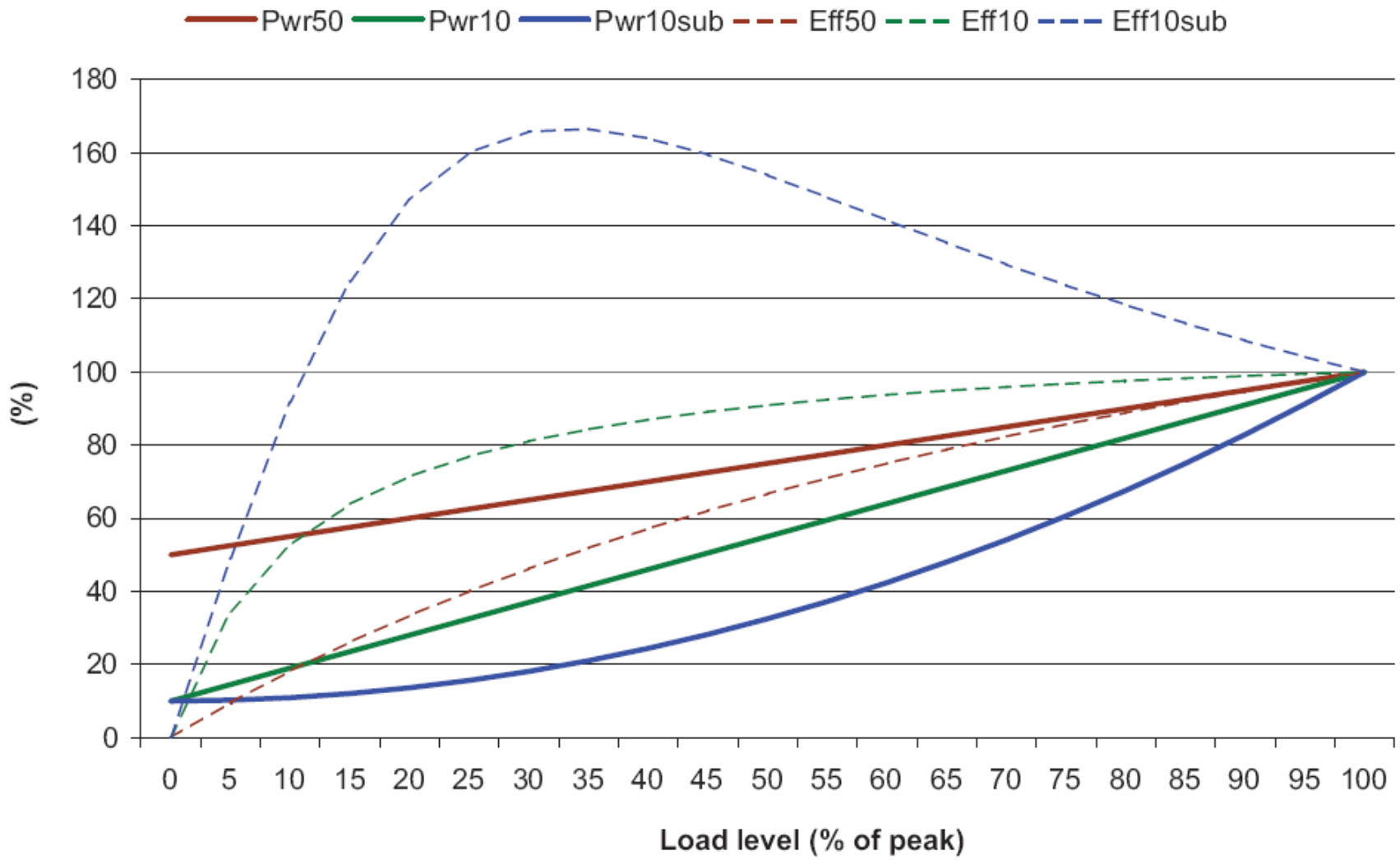


FIGURE 5.6: Power and corresponding power efficiency of three hypothetical systems: a typical server with idle power at 50% of peak (Pwr50 and Eff50), a more energy-proportional server with idle power at 10% of peak (Pwr10 and Eff10), and a sublinearly energy-proportional server with idle power at 10% of peak.

Humans as energy proportional systems

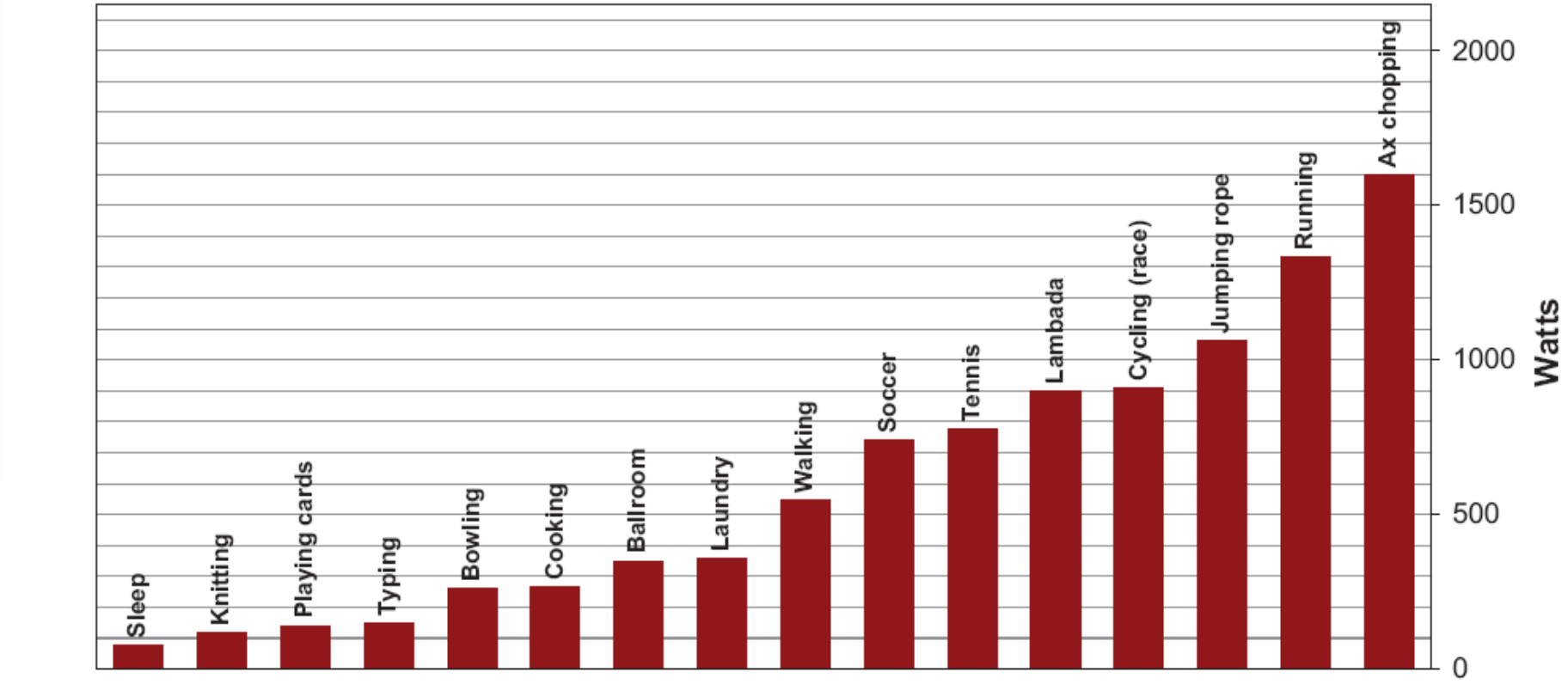


FIGURE 5.7: Human energy usage vs. activity levels (adult male) [52].

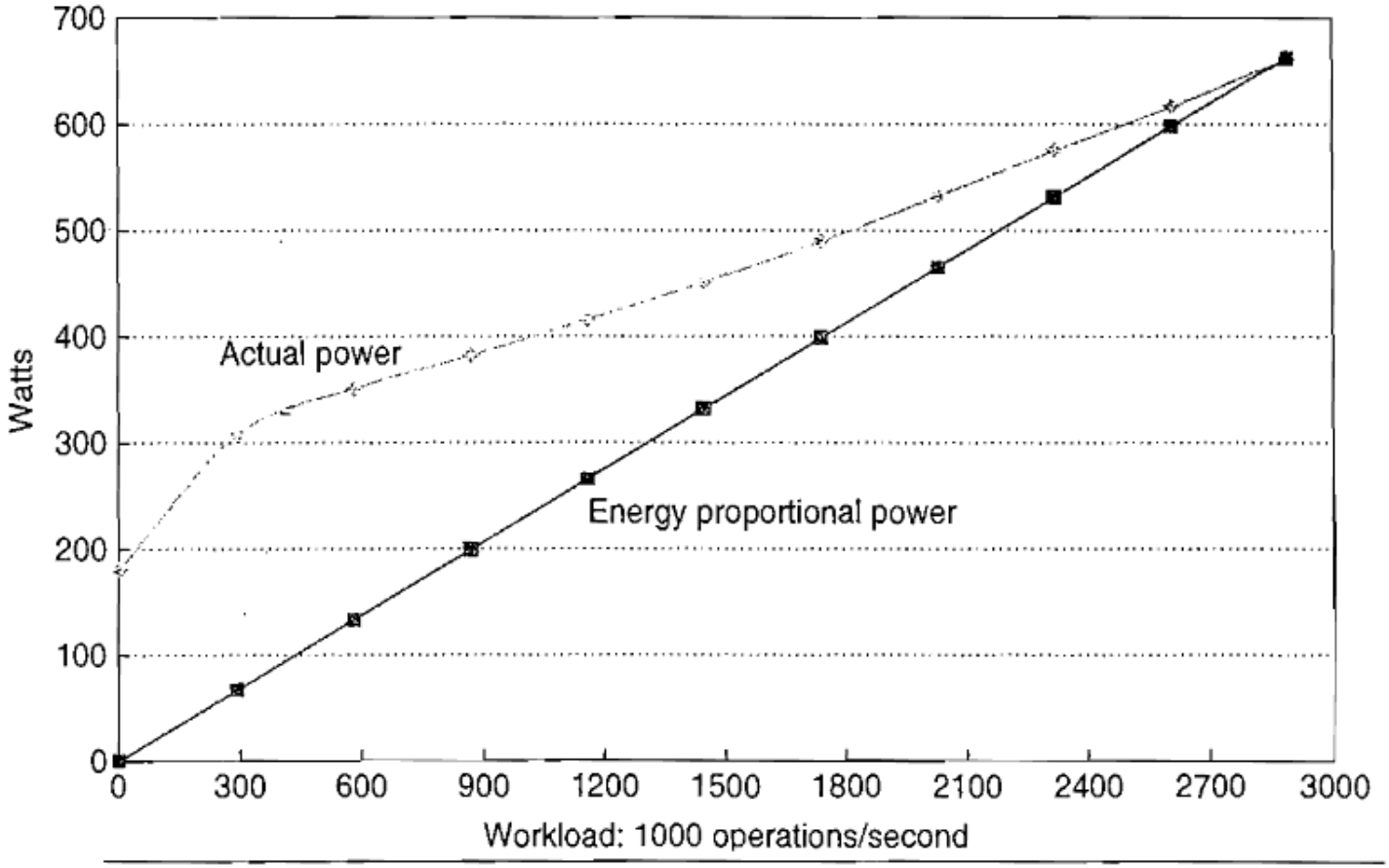


Figure 6.18 The best SPECpower results as of July 2010 versus the ideal energy proportional behavior. The system was the HP ProLiant SL2x170z G6, which uses a cluster of four dual-socket Intel Xeon L5640s with each socket having six cores running at 2.27 GHz. The system had 64 GB of DRAM and a tiny 60 GB SSD for secondary storage. (The fact that main memory is larger than disk capacity suggests that this system was tailored to this benchmark.) The software used was IBM Java Virtual Machine version 9 and Windows Server 2008, Enterprise Edition.

Server power usage vs load

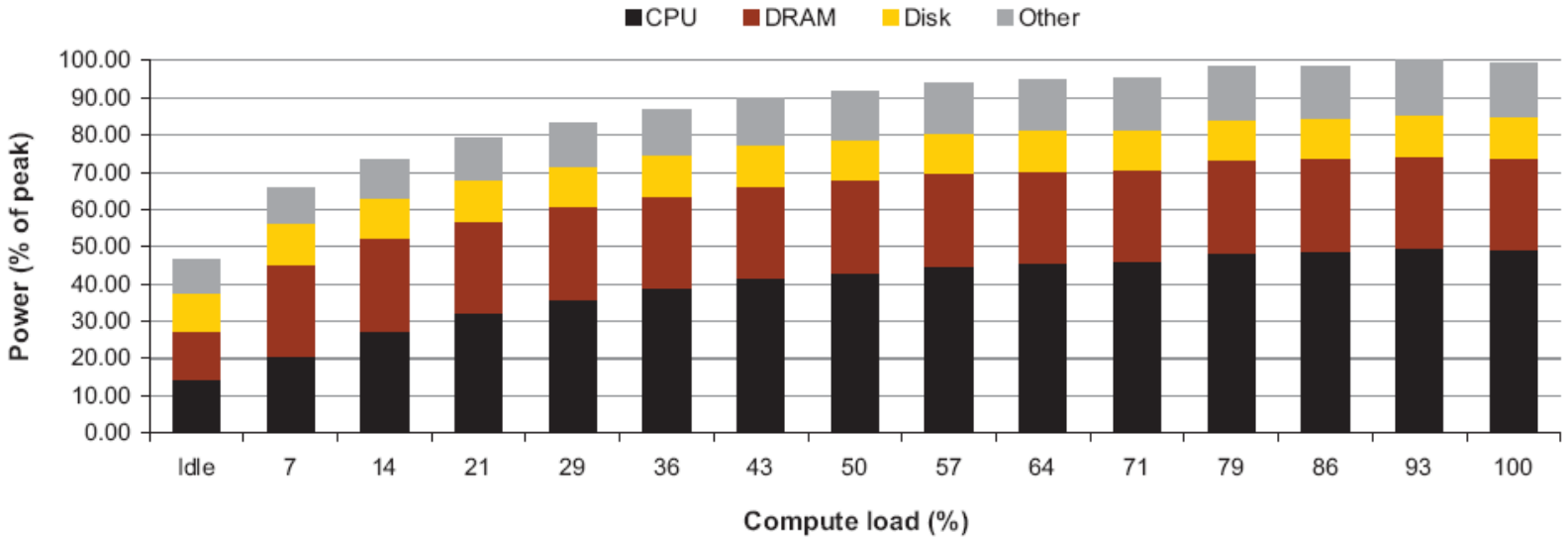


FIGURE 5.8: Subsystem power usage in an ×86 server as the compute load varies from idle to full usage.

WSC network as a Bottleneck

- Large switches are manually configured and fragile at a large scale
- It is difficult to afford more than dual redundancy in a WSC using these large switches, which limits the options for fault tolerance
- WSC network bottlenecks constrain data placement and complicate WSC software

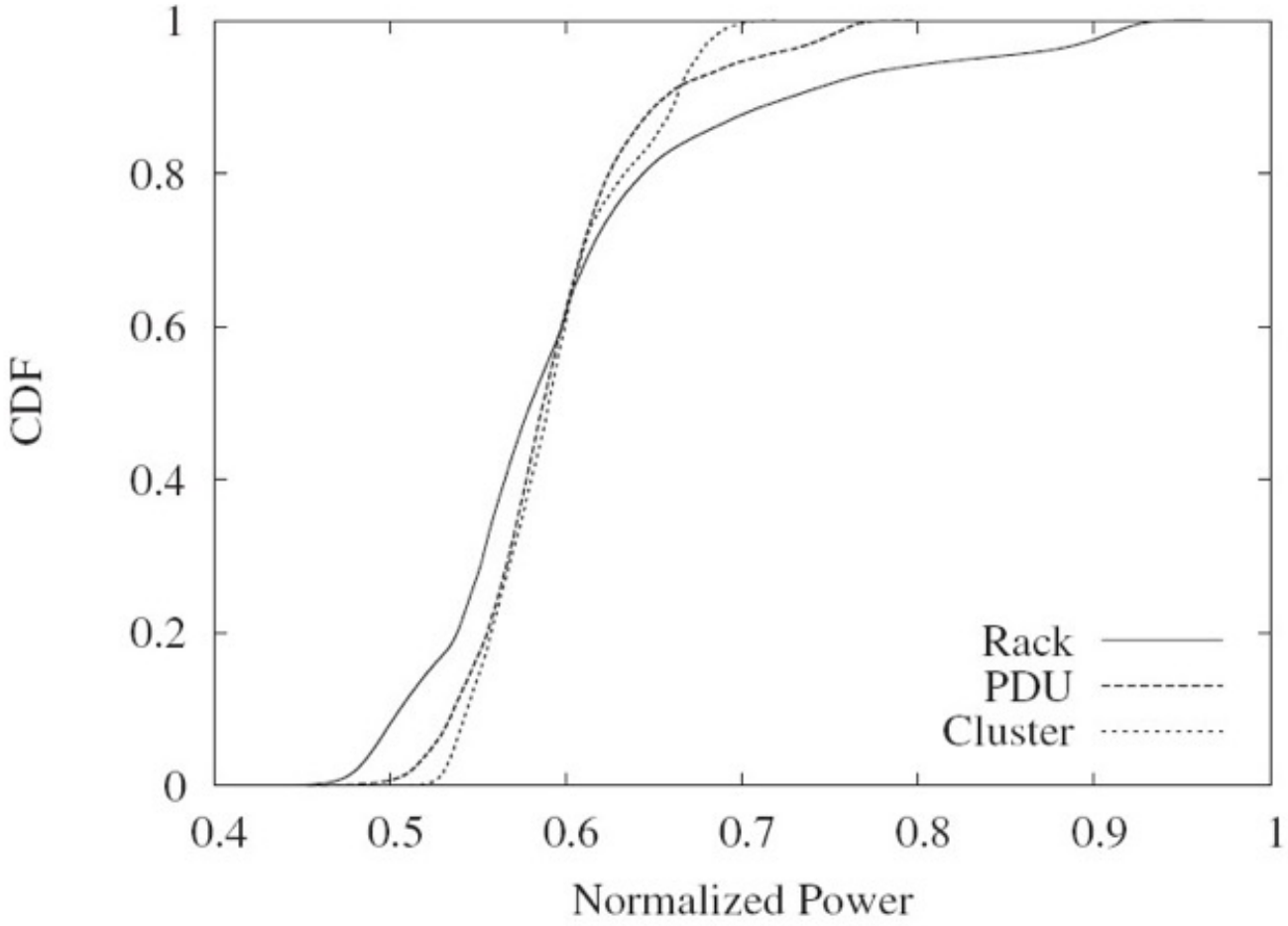


FIGURE 5.10: Cumulative distribution of the time that groups of machines spend at or below a given power level (power level is normalized to the maximum peak aggregate power for the corresponding grouping) (Fan et al. [27]).

DVFS -dynamic voltage and frequency scaling

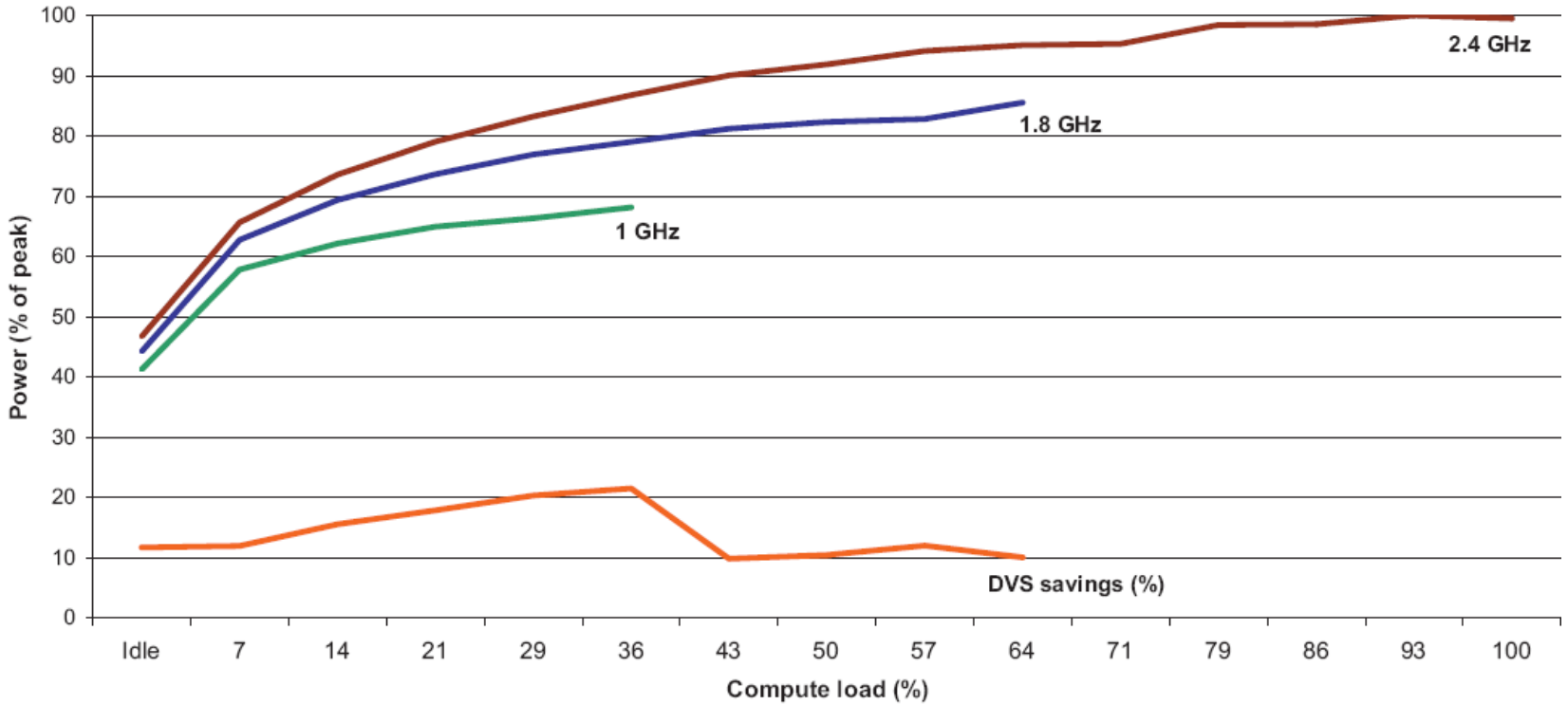
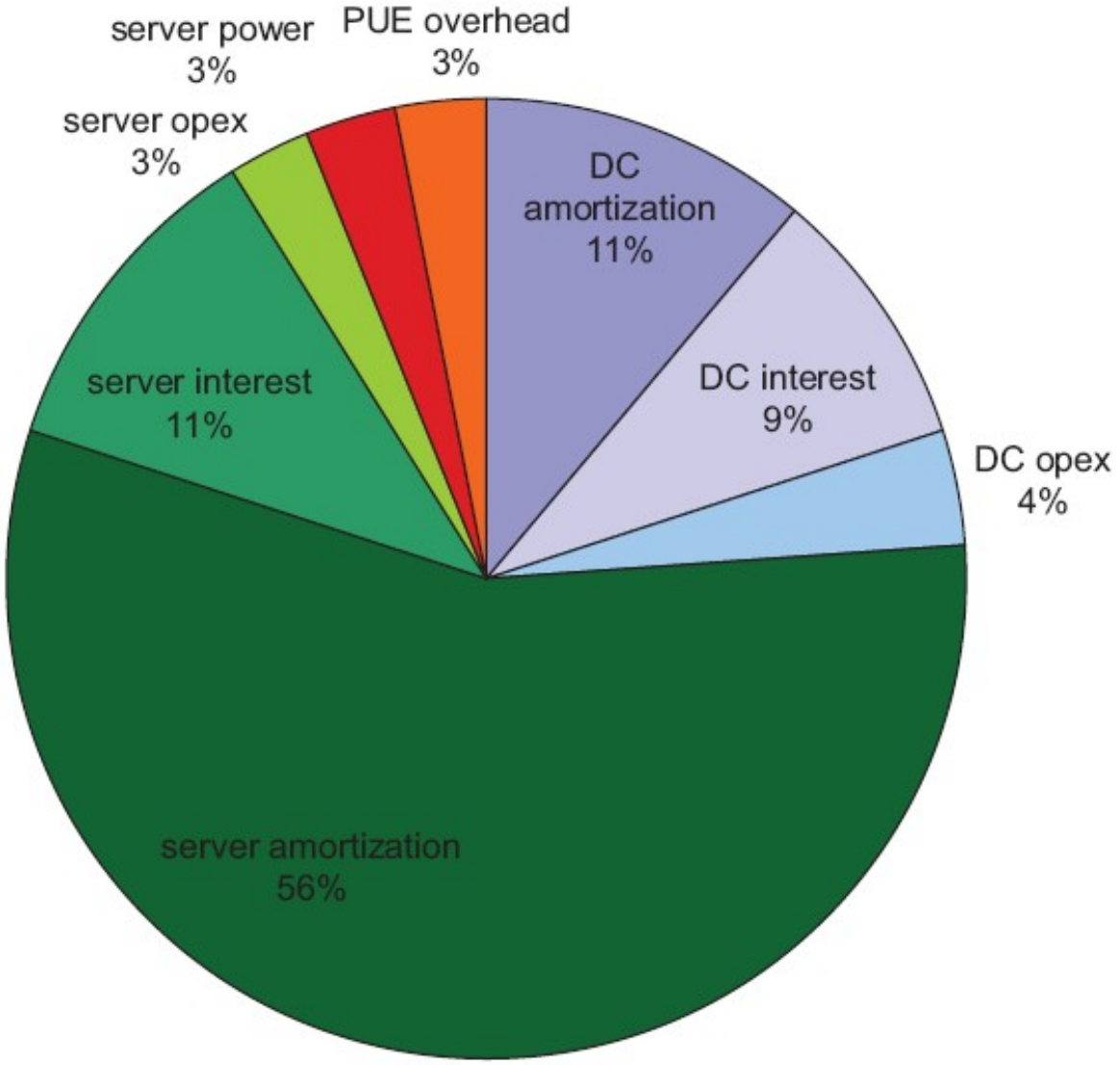


FIGURE 5.11: Power vs. compute load for an ×86 server at three voltage-frequency levels and the corresponding energy savings.

WSC costs

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

Operating cost breakdown



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.

Monthly operating costs - OPEX

Expense (% total)	Category	Monthly cost	Percent monthly cost
Amortized CAPEX (85%)	Servers	$2,000,000	53%
	Networking equipment	$290,000	8%
	Power and cooling infrastructure	$765,000	20%
	Other infrastructure	$170,000	4%
OPEX (15%)	Monthly power use	$475,000	13%
	Monthly people salaries and benefits	$85,000	2%
	Total OPEX	$3,800,000	100%

Figure 6.14 Monthly OPEX for Figure 6.13, rounded to the nearest $5000. Note that the 3-year amortization for servers means you need to purchase new servers every 3 years, whereas the facility is amortized for 10 years. Hence, the amortized capital costs for servers are about 3 times more than for the facility. People costs include 3 security guard positions continuously for 24 hours a day, 365 days a year, at $20 per hour per person, and 1 facilities person for 24 hours a day, 365 days a year, at $30 per hour. Benefits are 30% of salaries. This calculation doesn't include the cost of network bandwidth to the Internet, as it varies by application, nor vendor maintenance fees, as that varies by equipment and by negotiations.

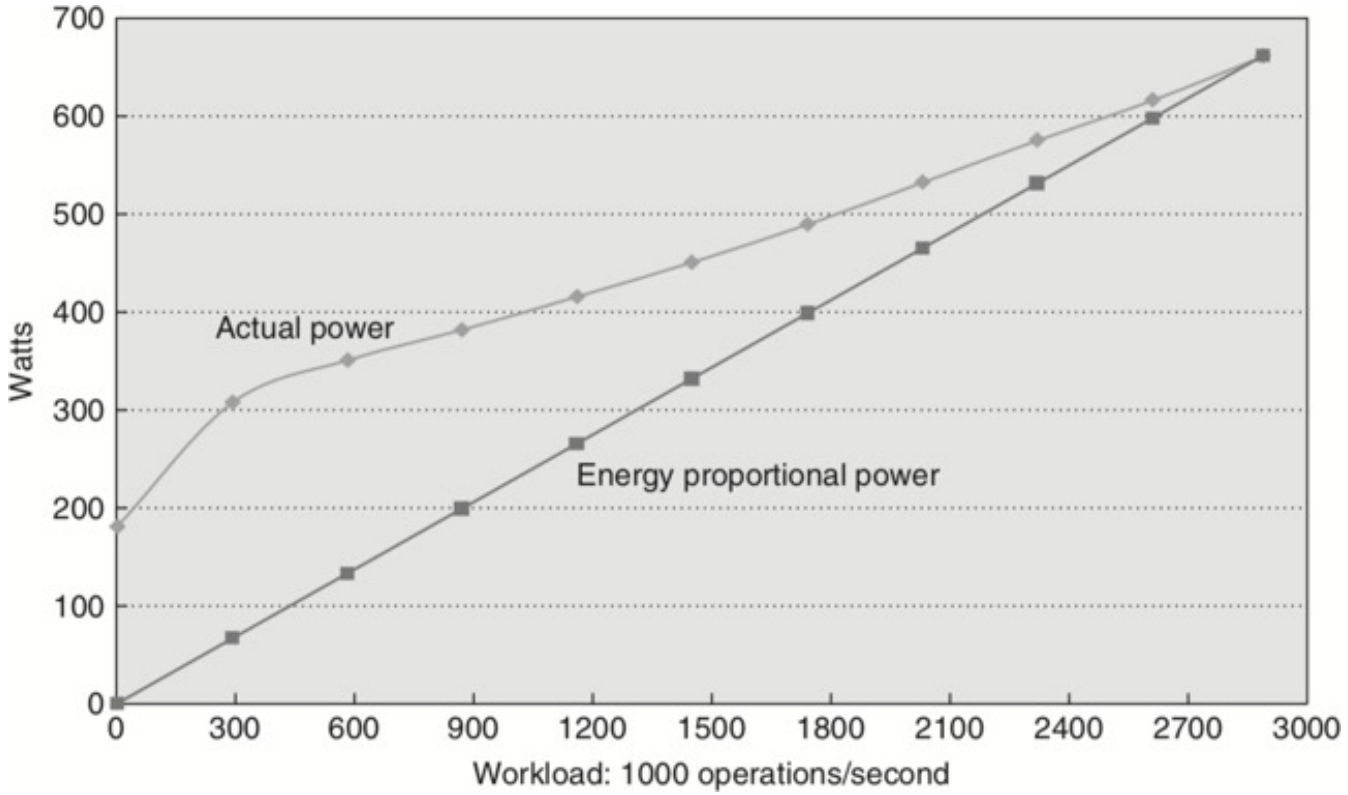
Energy Efficiency

- Another source of electrical inefficiency is the power supply *inside* the server
- Power supplies were 60%~80% efficient, thus there were greater losses inside the server
  - Supply a range of voltages to chips and disks…

A Fishbowl: May be is a good idea
(Cinvestav.mx ABACUS – Data Center)

Energy Efficiency

- Energy Proportionality – servers should consume energy in proportion to the amount of work performed



Cloud computing

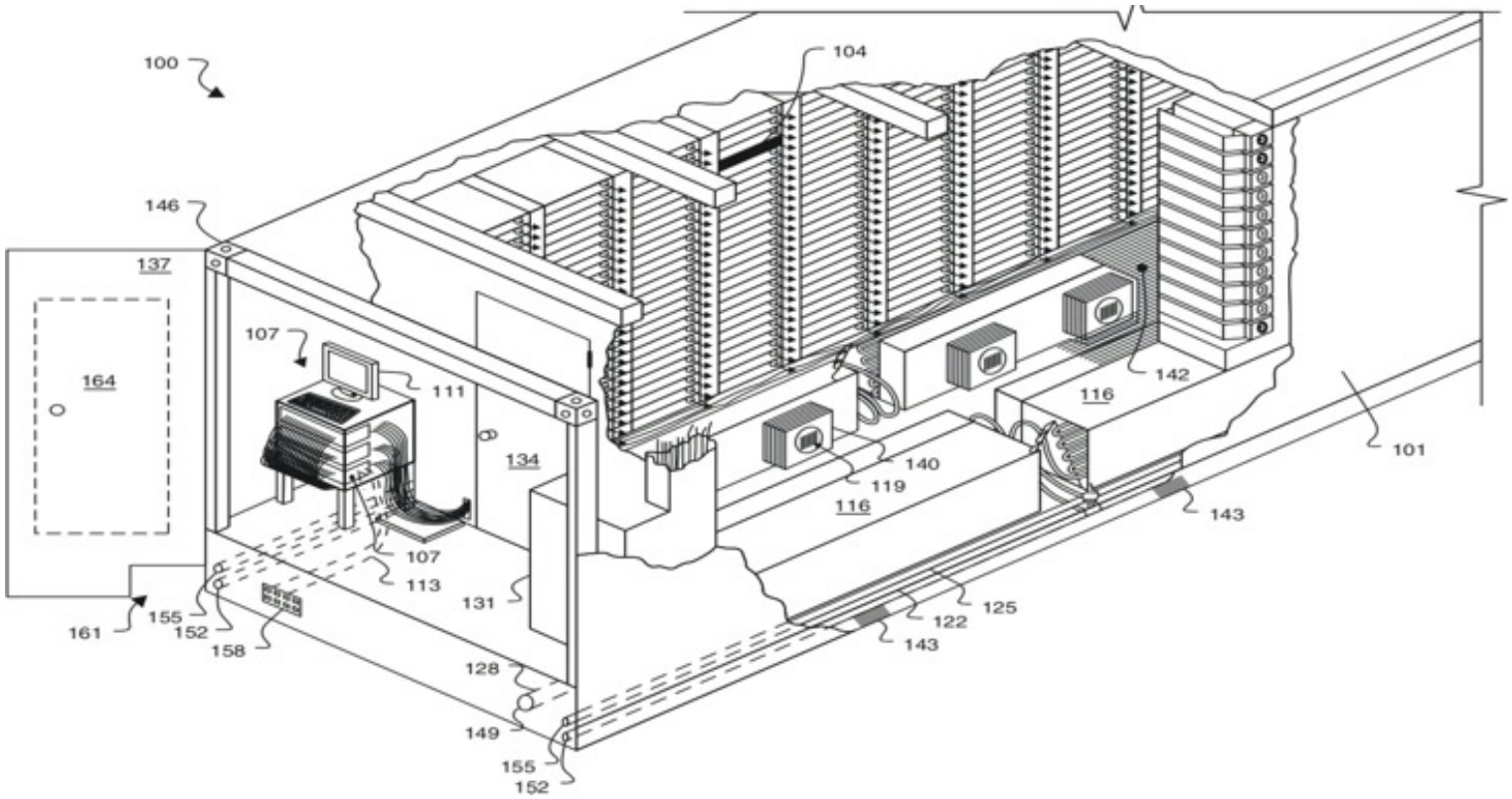
- WSCs offer economies of scale that cannot be achieved with a datacenter:
  - 5.7 times reduction in storage costs
  - 7.1 times reduction in administrative costs
  - 7.3 times reduction in networking costs
  - This has given rise to cloud services such as AWS (Amazon Web Services)
    - "Utility Computing"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

Instance	Per hour	Ratio to small	Compute units	Virtual cores	Compute units/core	Memory (GB)	Disk (GB)	Address size
Micro	$0.020	0.5–2.0	0.5–2.0	1	0.5–2.0	0.6	EBS	32/64 bit
Standard Small	$0.085	1.0	1.0	1	1.00	1.7	160	32 bit
Standard Large	$0.340	4.0	4.0	2	2.00	7.5	850	64 bit
Standard Extra Large	$0.680	8.0	8.0	4	2.00	15.0	1690	64 bit
High-Memory Extra Large	$0.500	5.9	6.5	2	3.25	17.1	420	64 bit
High-Memory Double Extra Large	$1.000	11.8	13.0	4	3.25	34.2	850	64 bit
High-Memory Quadruple Extra Large	$2.000	23.5	26.0	8	3.25	68.4	1690	64 bit
High-CPU Medium	$0.170	2.0	5.0	2	2.50	1.7	350	32 bit
High-CPU Extra Large	$0.680	8.0	20.0	8	2.50	7.0	1690	64 bit
Cluster Quadruple Extra Large	$1.600	18.8	33.5	8	4.20	23.0	1690	64 bit

Figure 6.15 Price and characteristics of on-demand EC2 instances in the United States in the Virginia region in January 2011. Micro Instances are the newest and cheapest category, and they offer short bursts of up to 2.0 compute units for just $0.02 per hour. Customers report that Micro Instances average about 0.5 compute units. Cluster-Compute Instances in the last row, which AWS identifies as dedicated dual-socket Intel Xeon X5570 servers with four cores per socket running at 2.93 GHz, offer 10 Gigabit/sec networks. They are intended for HPC applications. AWS also offers Spot Instances at much less cost, where you set the price you are willing to pay and the number of instances you are willing to run, and then AWS will run them when the spot price drops below your level. They run until you stop them or the spot price exceeds your limit. One sample during the daytime in January 2011 found that the spot price was a factor of 2.3 to 3.1 lower, depending on the instance type. AWS also offers Reserved Instances for cases where customers know they will use most of the instance for a year. You pay a yearly fee per instance and then an hourly rate that is about 30% of column 1 to use it. If you used a Reserved Instance 100% for a whole year, the average cost per hour including amortization of the annual fee would be about 65% of the rate in the first column. The server equivalent to those in Figures 6.13 and 6.14 would be a Standard Extra Large or High-CPU Extra Large Instance, which we calculated to cost $0.11 per hour.

Containers

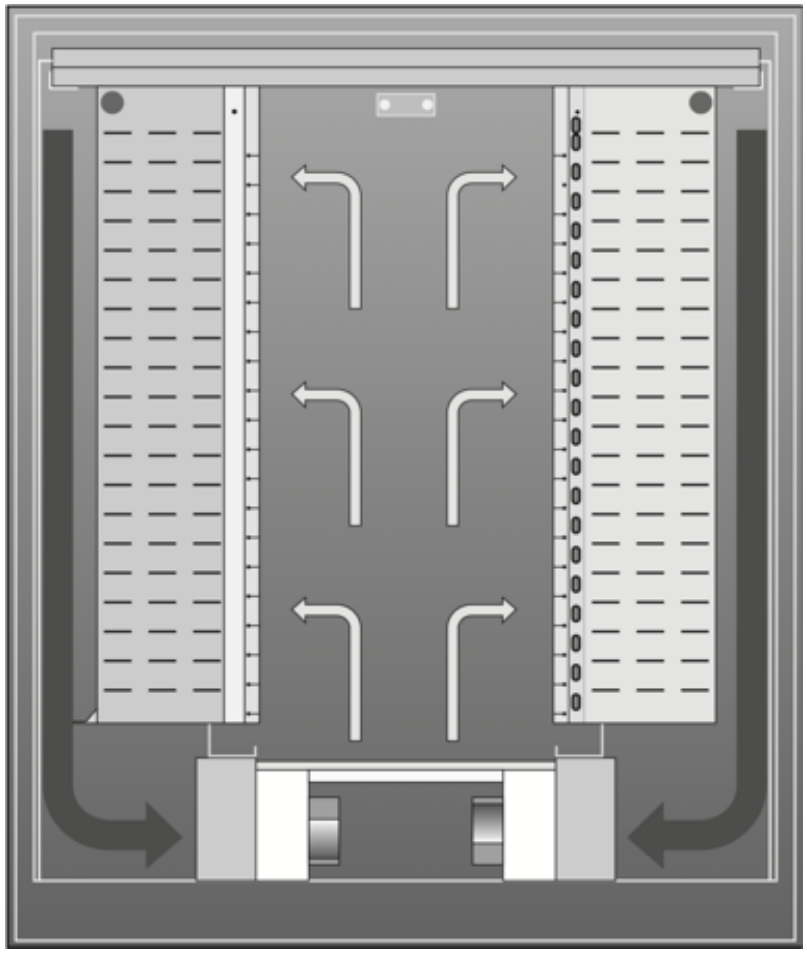
- Google built WSCs using shipping containers. The idea of building a WSC from containers is to make WSC design modular



http:// www.google.com/corporate/green/datacenters/summit.html

Cooling

- By controlling airflow to prevent hot spots, the container can run at a much higher temperature.
- The cooling is below a raised floor that blows into the aisle between the racks
- Hot air is returned from behind the racks. The restricted space of the container prevents the mixing of hot and cold air, which improves cooling efficiency



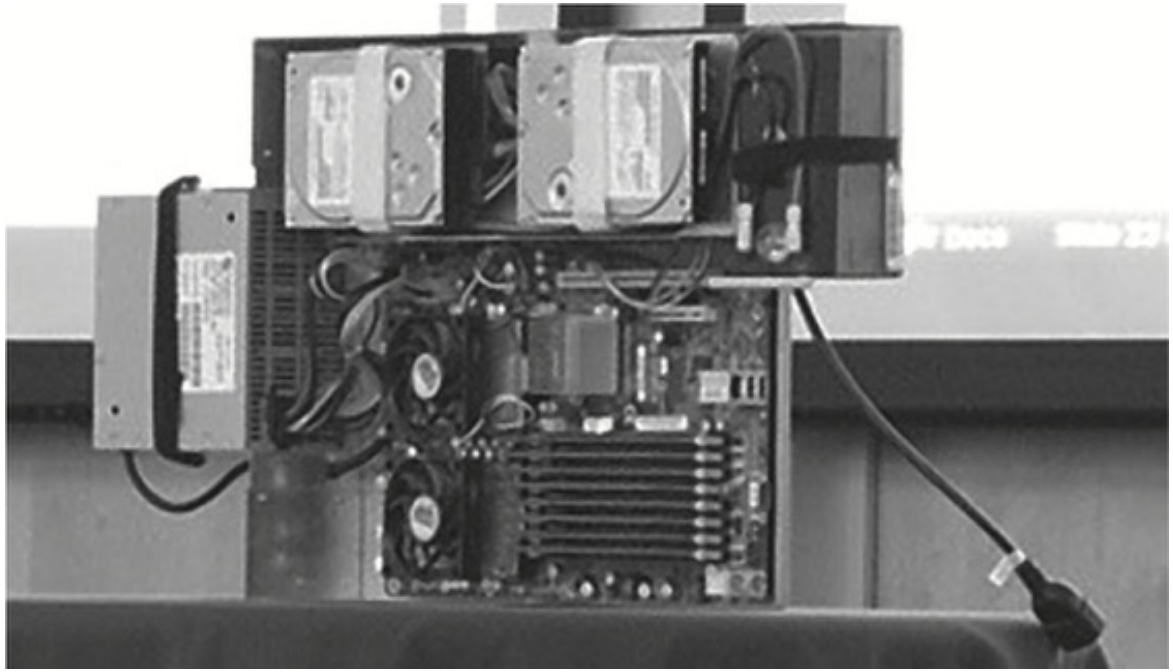
Two racks on each side of the container

Power

- Rather than have a separate battery room, each server has its own lead acid battery that is 99.99% efficient
  - Deployed incrementally with each machine, thus no money or power spent on overcapacity
- Use standard off-the-shelf UPS units to protect network switches
- DVFS was not deployed to avoid quality-of-service violation for online workloads

Server

- Two sockets
- Eight DIMMs
- Single network interface card (NIC)
- Two sata disk drives



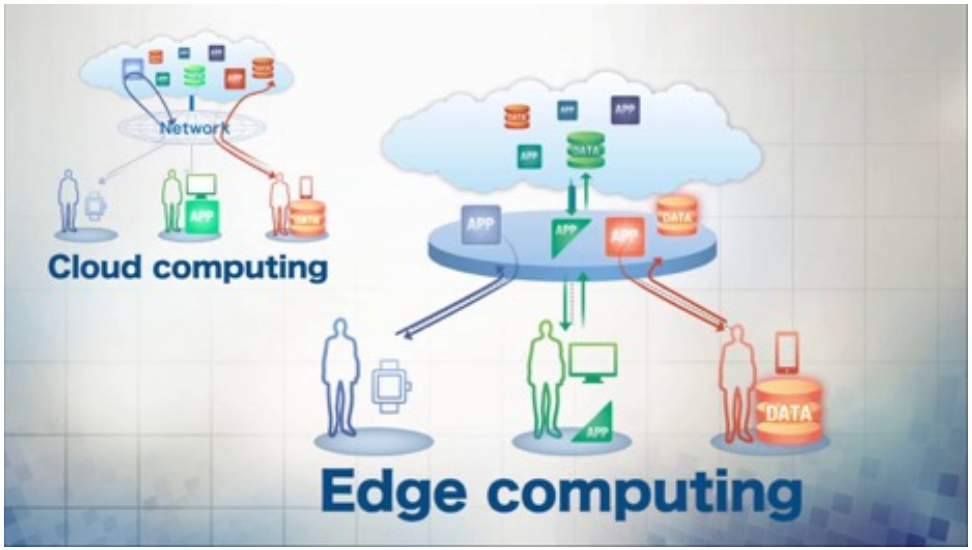
Networking

- 40000 servers are divided into three arrays of more than 10000 servers each

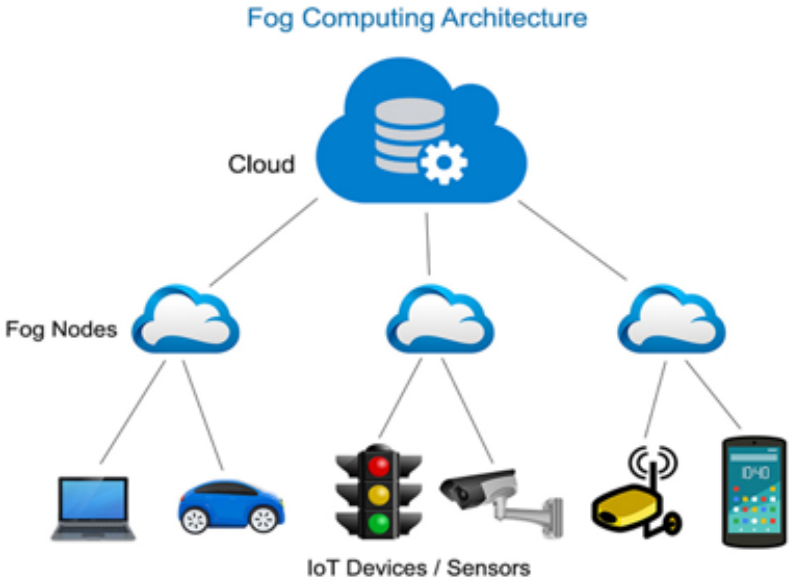
- The 48-port rack switches uses 40 ports to connect to servers, leaving 8 for uplinks to the array switches

- The number of uplink ports used per rack switch varies from a minimum of 2 to a maximum of 8
  - Applications with significant traffic demands beyond a rack tended to suffer from poor network performance

Why Ultra Scale?



Mesh network of micro data centers that process or store critical data locally



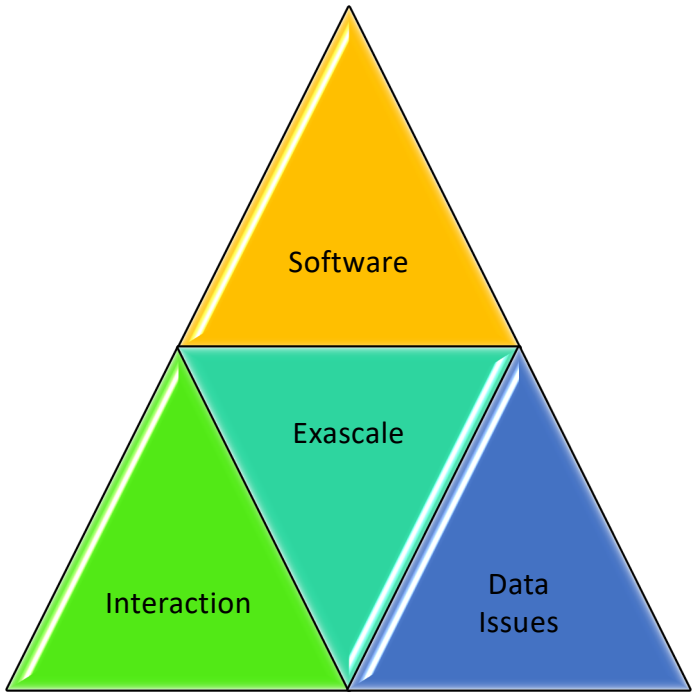
Extends Cloud computing and services to the edge of the network

NESUS cost IC1305
Network for Sustainable Ultrascale Computing

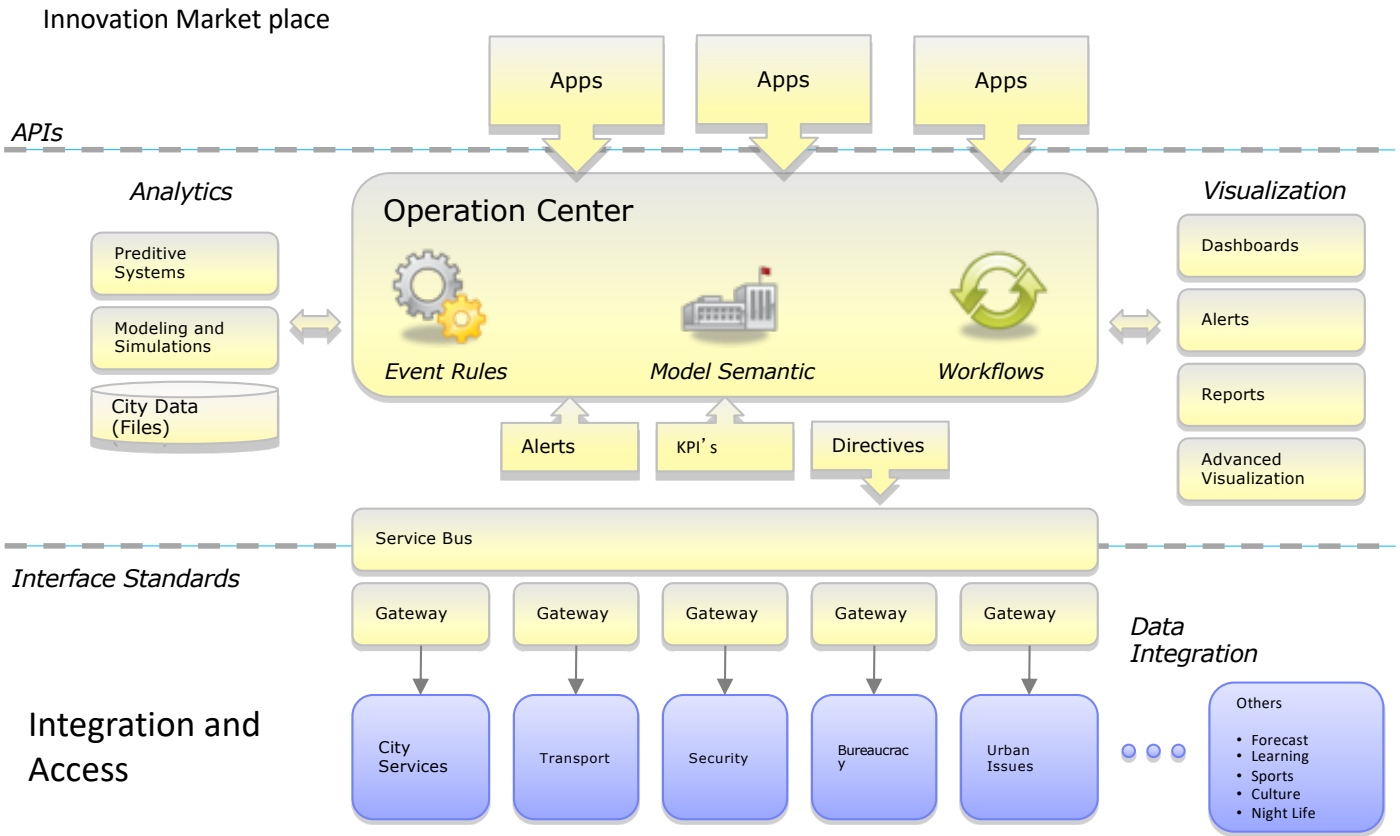
« Ultrascale systems are envisioned as large-scale complex systems joining parallel and distributed computing systems that will be two to three orders of magnitude larger that today's systems » (Carretero et al.)

Ultra Scale Systems needs Ultra Scale Software (Almost the concept)

- **Hybrid Systems (Hardware and Software Architecture)**
  - High Performance Capabilities
  - Parallelism
  - Energy Efficiency
- **Large Scale (Scalability)**
  - Latency
  - Distribution
  - Diversity (Networks, Protocols)
  - In Situ and In Transit
- **Software Quality**
  - Compatibility
  - Functionality
  - Reliability
  - Availability
  - Dependability
  - Usability



Example
Smart Citizen/City Context



CATAÏ
Collaboration France-Colombie

Monitoring and Repair

- Use monitoring software to track the health of all servers and networking gear
- Diagnostics are running all the time. When a system fails, many of the possible problems have simple automated solutions
- Failed machines are addressed in batches to amortize the cost of repair

Fallacies

1. CSPs (Cloud Service Providers) are loosing money.
   - Statistics show that the average income at AWS is
     $0.55/hour for reserved instances $0.45 for on-demand
   - Gross margins are 75% - 80%
1. Capital costs of a WSC facility are higher than for the servers it hosts.
   - The servers need to be replaced every 3-4 years
   - The facility lasts 10-15 years
   - The amortization makes a difference
3. Improved DRAMS availability and software based fault-tolerance diminish the need for ECC memory in a WSC
   - Measurements show that 1/3 of the servers experience memory errors with 22,000 correctable and 1 uncorrectable errors/year, one error is corrected every 2.5 hours
4. Turning off servers during periods of low activity improves WSC cost performance.

Universidad Industrial de Santander
CONSTRUIMOS FUTURO

Super Computación y Cálculo Científico UIS

Thanks

More information about large scale architectures at UIS:

www.sc3.uis.edu.co