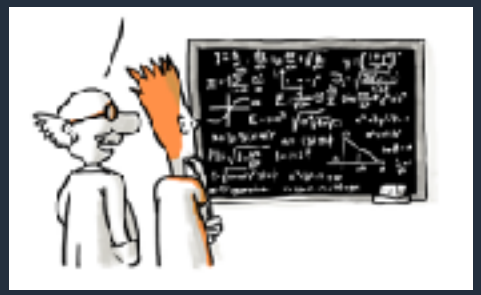
DSC516: Cloud Computing

Part II: Cloud Building Blocks

Module 3: Cloud Computing Infrastructure

University of Cyprus
Department of Computer Science

Summary of previous lectures



- From mainframes to Cloud Computing
- Distributed Computing Architecture and Models
- Cloud Computing definition and models
- The Economics of Cloud Computing
- The Computing Landscape: from Cloud to Fog and Edge Computing

Lecture 5

Data Centers and Warehouse Scale Computers

University of Cyprus
Department of Computer Science

Readings

- Chapters 1, 3, 4,5, 6 **The data center as a Computer. An Introduction to the Design of Warehouse-Scale Machines**. Barroso, L. A., Holzle, U. & P. Raganathan (2018). Third Edition. In Synthesis Lectures on Computer Architecture (Vol. 2, Issue 1). Morgan & Claypool Publishers.

Data Centers and Warehouse Scale Computers

Data Center Basics

University of Cyprus
Department of Computer Science

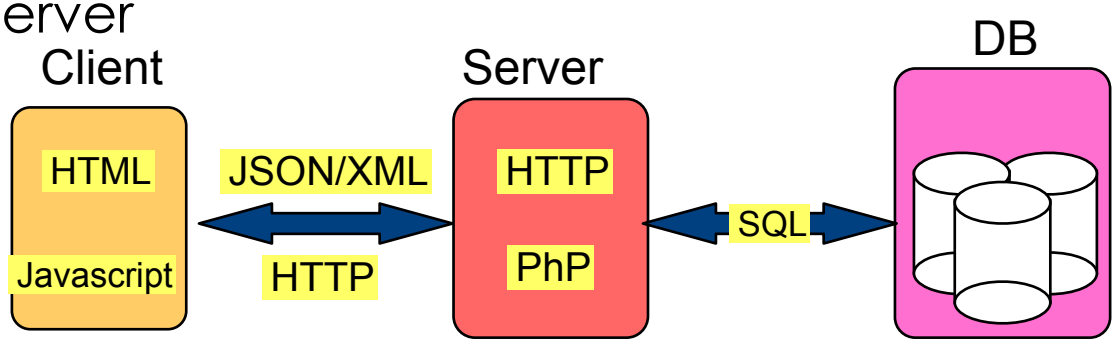
Learning objectives

- Understand and explain the basic characteristics of **Warehouse Scale Computers**.
- Understand and explain the differences between **WSCs** and **Datacenters**.
- Understand and explain the **software infrastructure building blocks** of WSC.
- Understand and describe main characteristics of **WSC buildings** and their **power provision**.
- Understand and explain the **basic power, cooling components** of a modern datacenter.
- Understand and explain the concept of **energy efficiency** in datacenters, and associated mechanisms.
- Understand, and explain the cost structure of running a modern datacenter, and the concept of TCO (**total cost of ownership**).
- Explore, understand and explain concepts and techniques for energy-efficiency in datacenters.

A Paradigm Shift

- Traditional programming model: Algorithms + Data Structures = Programs

- Three-tier Web programming model: Client-Web/Application-Database Server



- Cloud Computing application development signifies a departure from the traditional programming model where a program runs on a single machine or from the three-tier Web programming model where applications are split between client-application server and database.

A Paradigm Shift

- In **warehouse-scale computing:**
  - the program is an **Internet service**, which may consist of **tens or more individual programs** that **interact** to **implement complex end-user services**.
    - These programs might be implemented and maintained by different teams of engineers, perhaps across organizational, geographic, and company boundaries.
  - The computing platform consists of **thousands of individual computing nodes** with their corresponding **networking** and **storage** subsystems, **power distribution** and **conditioning** equipment, and extensive **cooling** systems. The enclosure for these systems is a building structure.

Datacenters vs WSC

- **Datacenters**: buildings w co-located multiple servers & communication gear:
  - Large number of **small**- or **medium-sized** applications, each running on a dedicated hardware infrastructure that is **de-coupled** and **protected** from other systems in the same facility.
  - Host hardware and software for **multiple organizational units** or **different companies**.
  - Different computing systems within a DC have little in common in terms of hardware, software, or maintenance infrastructure; tend not to communicate with each other at all.
- **WSCs** currently power the services offered by companies such as GAFA:
  - belong to a **single organization**, use a **relatively homogeneous hardware and system software** platform, and share a **common systems management layer**.
  - Application, middleware, and system software is built **in-house**.
  - Run a **smaller number of very large applications** (or **Internet services**), and the **common resource management infrastructure** allows significant deployment flexibility.

WSC Requirements

- **High Availability** (at least 99.99% uptime - about **an hour of downtime per year**).
- **Fault-free operation** possible but extremely expensive.
  - WSC workloads must be designed to gracefully tolerate large numbers of component faults with little or no impact on service level performance and availability
- **Cost-Efficiency:** a primary metric of interest in the design of WSCs (why?)
  - Defined to account for all significant components of cost:
    - hosting-facility Capital and Operational expenses (power provisioning and energy costs)
    - hardware
    - software
    - management personnel
    - repairs

WSC Technical Challenges

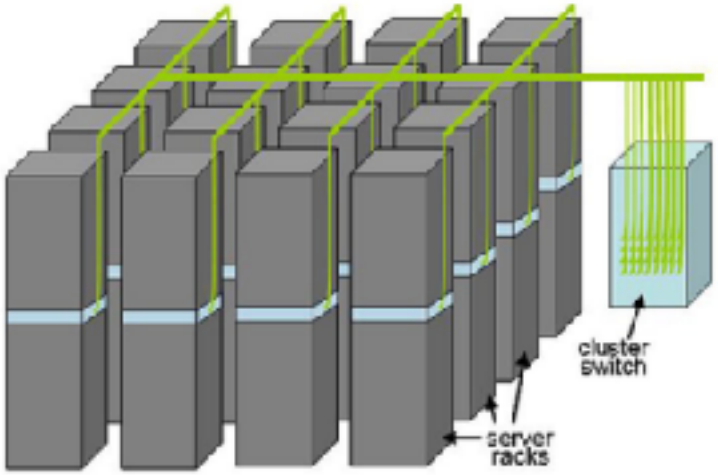
- Design, implementation, deployment driven by **new** and **rapidly evolving workloads**.
- **Efficient experimentation with** and **simulation of WSC**, is difficult because of size: need for new techniques to guide design decisions.
- **Fault behavior** and **power / energy considerations** have a more significant impact in the design of WSCs.
- **Additional layer of complexity** beyond systems consisting of individual servers or small groups of servers.
- Maintaining **high programmer productivity**, while having to deal with: larger scale of the application domain, deeper and less homogeneous storage hierarchy, higher fault rates, and possibly higher performance variability.

Niche or Wider Relevance?

- Problems that today's large Internet services face will soon be meaningful to a much larger constituency because many organizations will soon be able to afford **similarly sized computers at a much lower cost**.
- A single rack of servers may soon have as many or more hardware threads than many of today's datacenters (2018).
  - Eg: a rack with 40 servers, each with four 16-core dual-threaded CPUs, would contain more than 4000 hardware threads.
  - Such systems will arguably be affordable to a very large number of organizations within just a few years, while exhibiting some of the scale, architectural organization, and fault behavior of today's WSCs.
  - Experience building these systems is useful in understanding the design issues and programming challenges for the next-generation cloud computing platform.

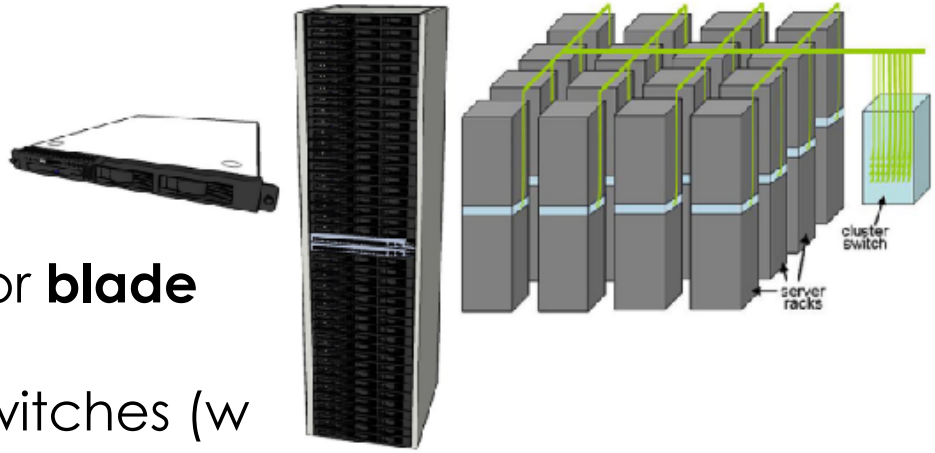
Blade
Κόβος
1U ή
πτερύγιο

Rack
Ικρίωμα



Cluster
Συστοιχία

Architectural Overview



- Low-end servers, typically in a **1U** or **blade enclosure** format.
- Interconnected using rack-level switches (w 40 Gbps or 100 Gbps link).
- Each rack-level switch has a number of uplink connections to one or more cluster-level (or data center-level) Ethernet switches; spanning potentially more than 10K servers.
- In the case of a blade enclosure, there is an additional first level of networking aggregation within the enclosure where multiple processing blades connect to a small number of networking blades through an I/O bus such as PCIe.
- More recently, WSCs have featured additional compute hardware building blocks, including **GPUs** and custom **accelerators**.
  - Similar to servers, these are connected through custom or industry-standard interconnects at the rack (or multi-rack pod) levels, leading up to the data center network.

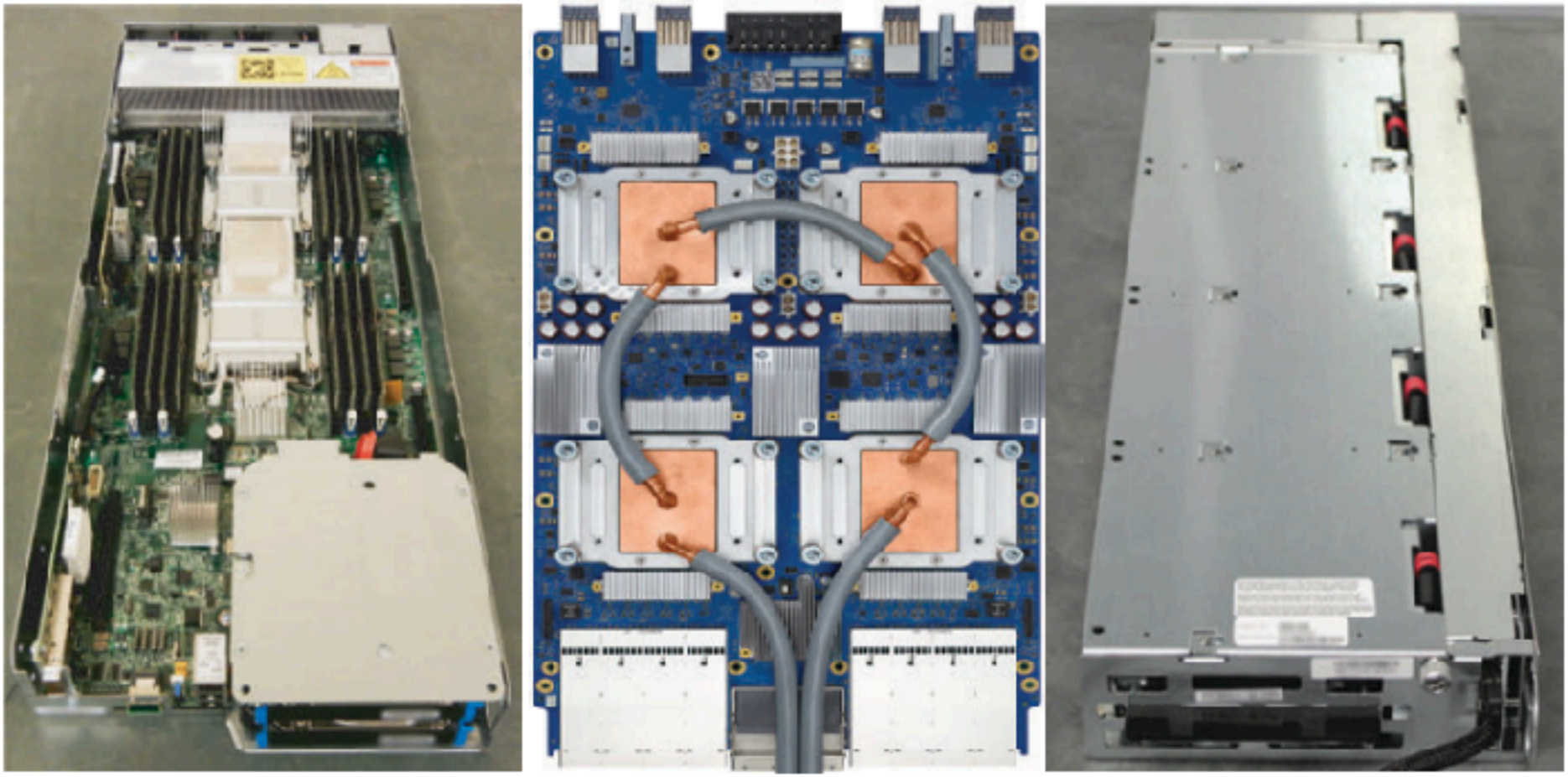


Figure 1.1: Example hardware building blocks for WSCs. Left to right: (a) a server board, (b) an accelerator board (Google's Tensor Processing Unit [TPU]), and (c) a disk tray.

Buildings and Infrastructure

- **WSC building design decisions** can dramatically influence the availability and uptime of the WSC.
- WSCs often consume more power than thousands of individual households and need **elaborate designs**, which comprise
  - holistic and hierarchical power delivery
  - backup and redundancy
  - end-to-end cooling solutions
- The building design, delivery of input energy, and subsequent removal of waste heat:
  - drive a **significant fraction of data center costs** proportional to the amount of power delivered
  - have implications on: **design, performance, availability** service level objectives (**SLOs**)

Energy and Power Usage concerns

- Energy-related costs have become an important component of the total cost of ownership of WSC.

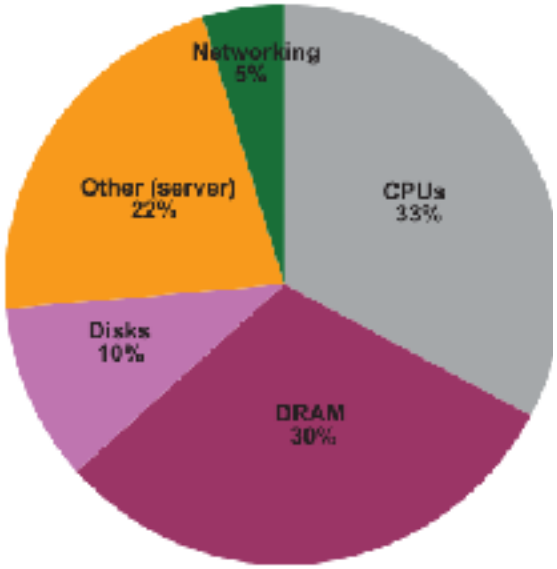


FIGURE 1.4: Approximate distribution of peak power usage by hardware datacenters (circa 2007).

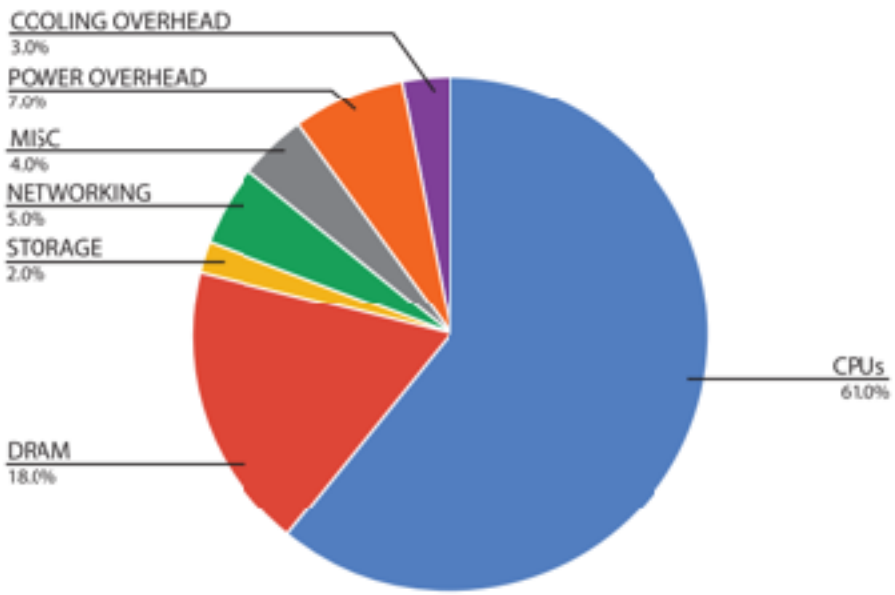


Figure 1.3: Approximate distribution of peak power usage by hardware subsystem in a modern data center using late 2017 generation servers. The figure assumes two-socket x86 servers and 12 DIMMs per server, and an average utilization of 80%.

Faults and Failures

- WSC-hosted internet services software must tolerate relatively high component fault rates.
  - Disk drives can exhibit annualized failure rates **> 4%.**
- Reports of **1.2-16** average **server-level restarts** per year.
- An application running across 1000s of machines may need to react to failure conditions **on an hourly basis**.

Data Centers and Warehouse Scale Computers

Data Center Server Hardware

University of Cyprus
Department of Computer Science

Clusters [συστοιχίες Η/Υ]

- Collections of commodity servers that work together on a single problem

NAME THE MAIN ADVANTAGES OFFERED BY CLUSTERS AS THE SOLUTION OF CHOICE FOR DATA CENTERS

Why clusters?

- Absolute scalability. A successful network service must scale to support a substantial fraction of the world's population.
- Cost and performance
  - No alternative to clusters can match the required scale
  - Hardware cost is typically dwarfed by bandwidth and operational costs.
- Independent components. Users expect 24-hour service from systems that consist of thousands of hardware and software components.
  - Transient hardware failures and software faults due to rapid system evolution are inevitable
  - Clusters simplify the problem by providing **(largely) independent faults**.
- Incremental scalability. Clusters should allow for scaling services as needed to account for the uncertainty and expense of growing a service.
  - three-year depreciation lifetime (χρόνος απόσβεσης) - should generally be replaced only when they no longer justify their rack space compared to new nodes.
  - A unit of rack space should quadruple in computing power over three years [Moore's law]. Actual increases appear to be faster due to improvements in packaging and disk size.

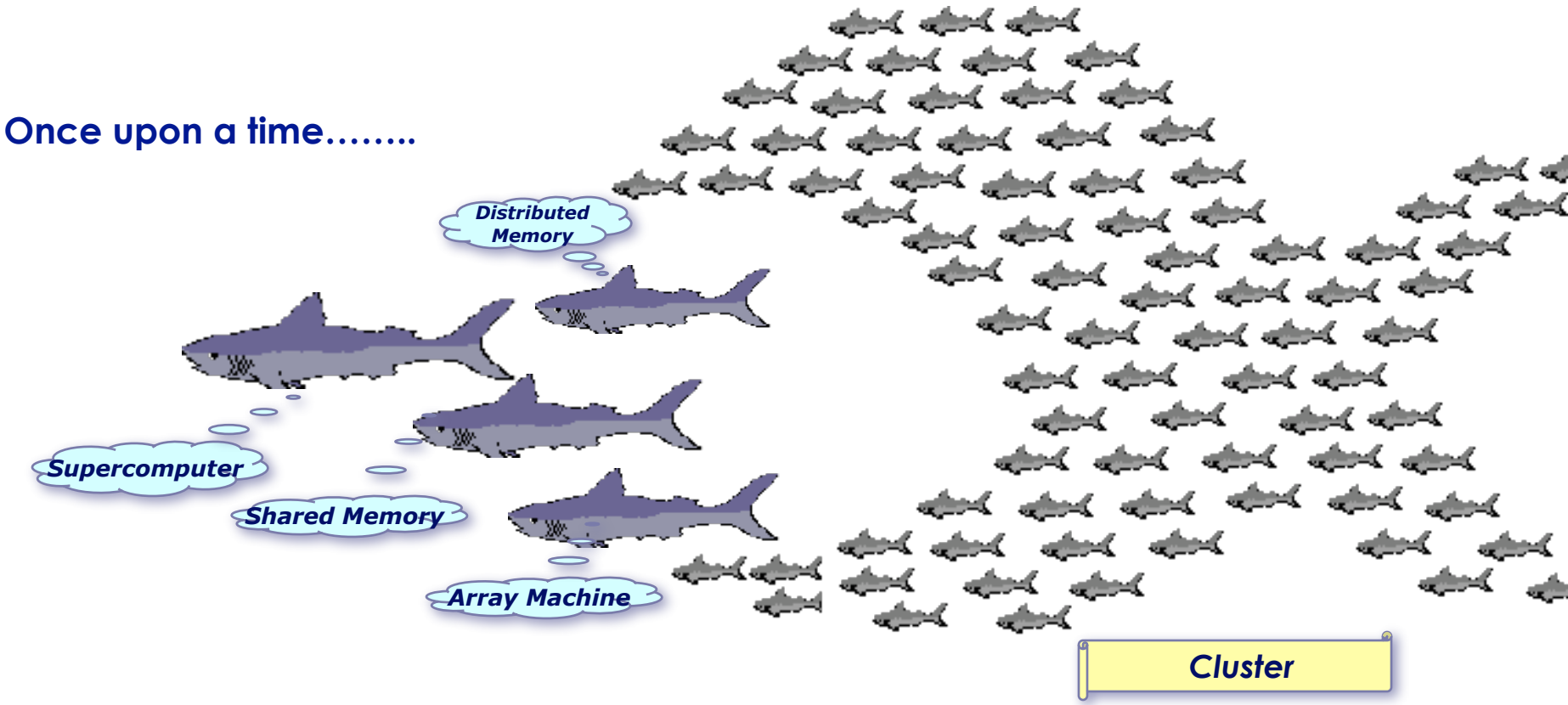
Hardware Building Blocks

Server Hardware

Server Hardware

- Clusters of mid-range servers are the preferred building blocks for WSCs today, because of **cost-efficiency**.
  - Studies with **benchmarks** showed difference in cost-efficiency over a factor of 4 in favor of lower-end servers than more expensive, high-end servers (circa 2009).
  - With CPU core count increase, most VM/task instances can comfortably fit into a two-socket server
- The more interesting discussions now: between mid-range server nodes and extremely low end (so-called "**wimpy**") servers.

The power of the masses



(by Christophe Jacquet)

Design considerations

- The **key design considerations** that determine server's form factor and functionalities are:
  - **CPU**: CPU power, often quantified by the **thermal design power**, or TDP; number of CPU sockets and NUMA topology; CPU selection (core count, core and uncore frequency, cache sizes, and number of inter-socket coherency links).
  - **Memory**: Number of memory channels, number of DIMMs per channel, and DIMM types supported (such as RDIMM, LRDIMM, and so on).
  - **Plug-in IO cards**: Number of PCIe cards needed for SSD, NIC, and accelerators; form factors; PCIe bandwidth and power, and so on.
  - **Tray-level power and cooling**, and **device management** and **security options**: Voltage regulators, cooling options (liquid versus air-cooled), board management controller (BMC), root-of-trust security, and so on.
  - **Mechanical design**: how are individual components assembled - server form-factors (width, height, depth), front or rear access for serviceability.

WHICH FACTORS NEED TO BE CONSIDERED TO MAKE THESE DECISIONS?

Workload performance

Total cost of ownership (TCO)

Need for flexibility

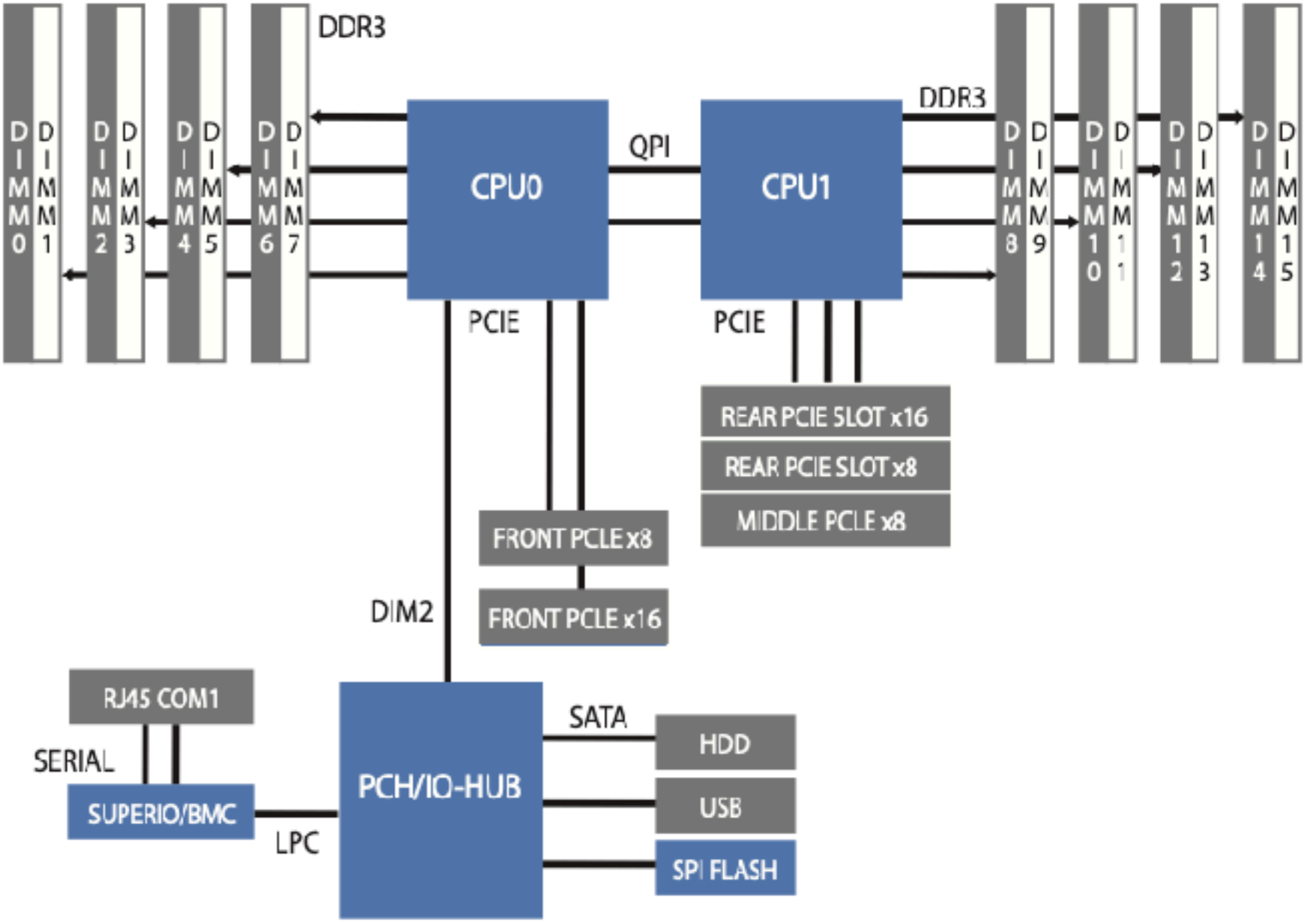
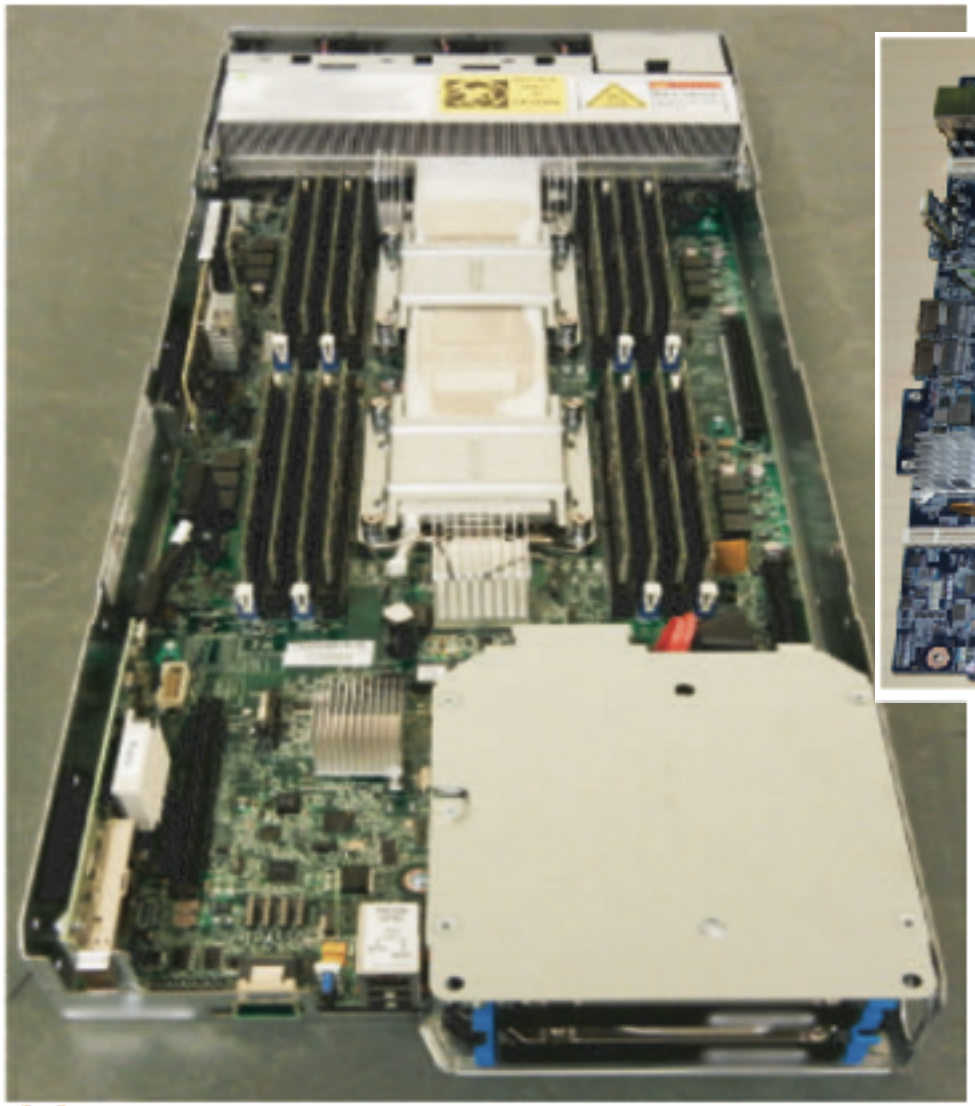


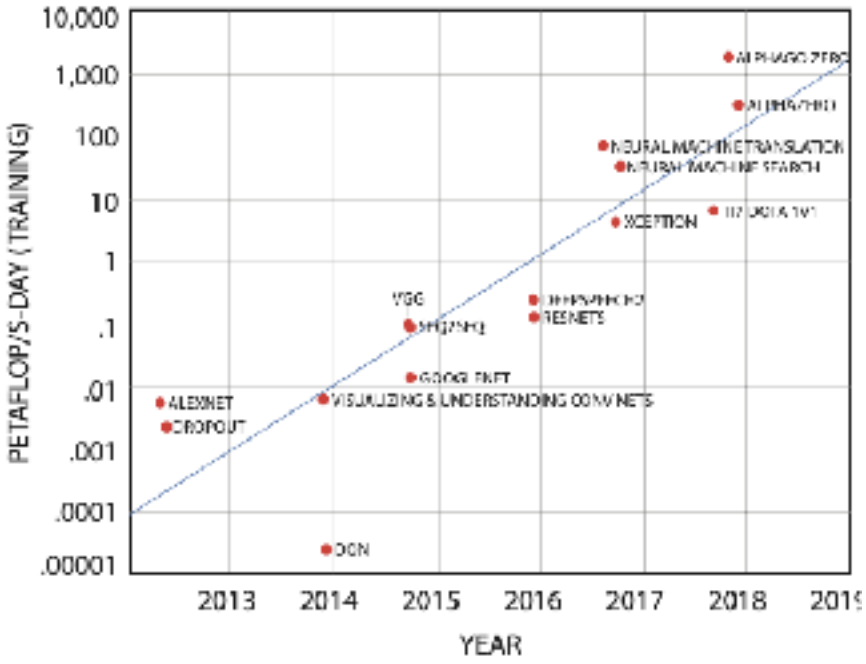
Figure 3.1: Block diagram of a server.

Server trays (examples)



Accelerators

- Growth of general-purpose computing significantly slowed down with a **doubling rate** (2018) exceeding **4 or more years** (vs. 18–24 months expected from Moore's Law).
- However: since 2013, AI training compute requirements have **doubled every 3.5 months**.



Training compute requirements for models over time [OA18].

Accelerators

- To satisfy the growing compute needs for deep learning, WSCs deploy:
  - Graphics Processing Units (TPU)
  - Tensor Processing Units (TPU by Google)
  - FPGA-based accelerators (project Catapult, Microsoft)
  - Other specialized accelerator hardware
- GPUs/TPUs are configured with a CPU host, connected to a PCIe-attached accelerator tray with multiple GPUs.
  - GPUs within the tray are connected using high-bandwidth interconnects.

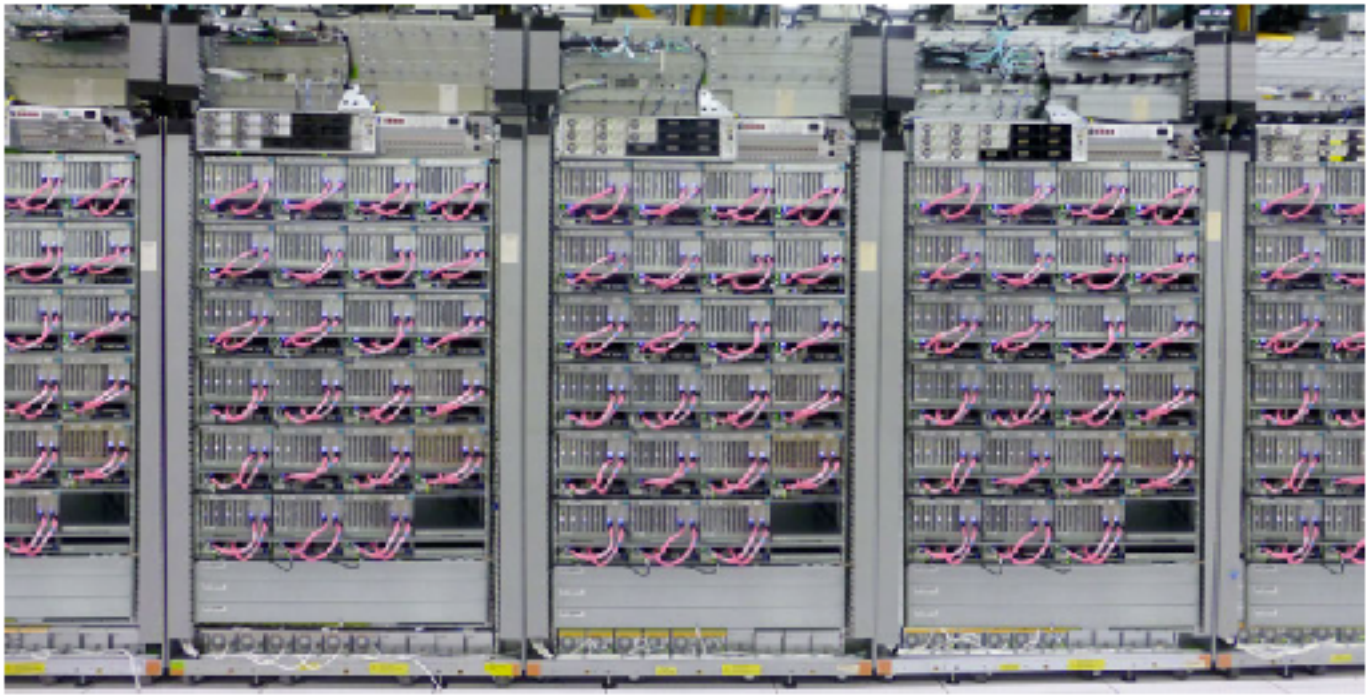


Figure 3.7: Interconnected GPUs for training.

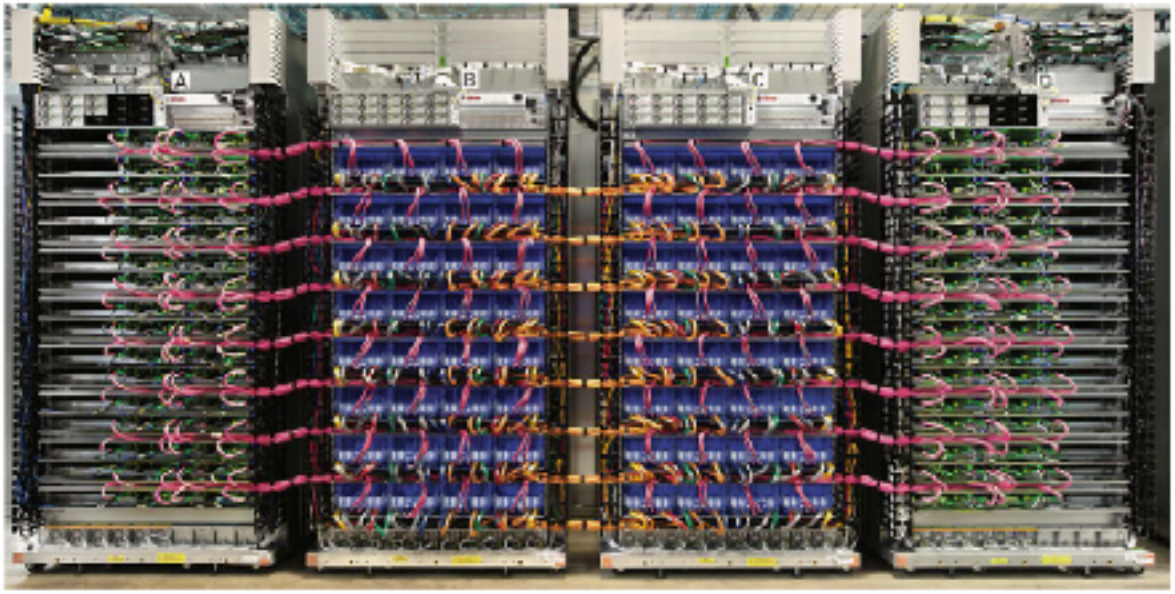


Figure 3.9: Four-rack TPUv2 pod.

Rack

- The physical structure that holds tens of servers together.
- Handle shared power infrastructure: power delivery, battery backup, and power conversion (such as AC to 48V DC).
- Width and depth of racks vary across WSCs: some are classic 19-in wide, 48-in deep racks, while others can be wider or shallower.

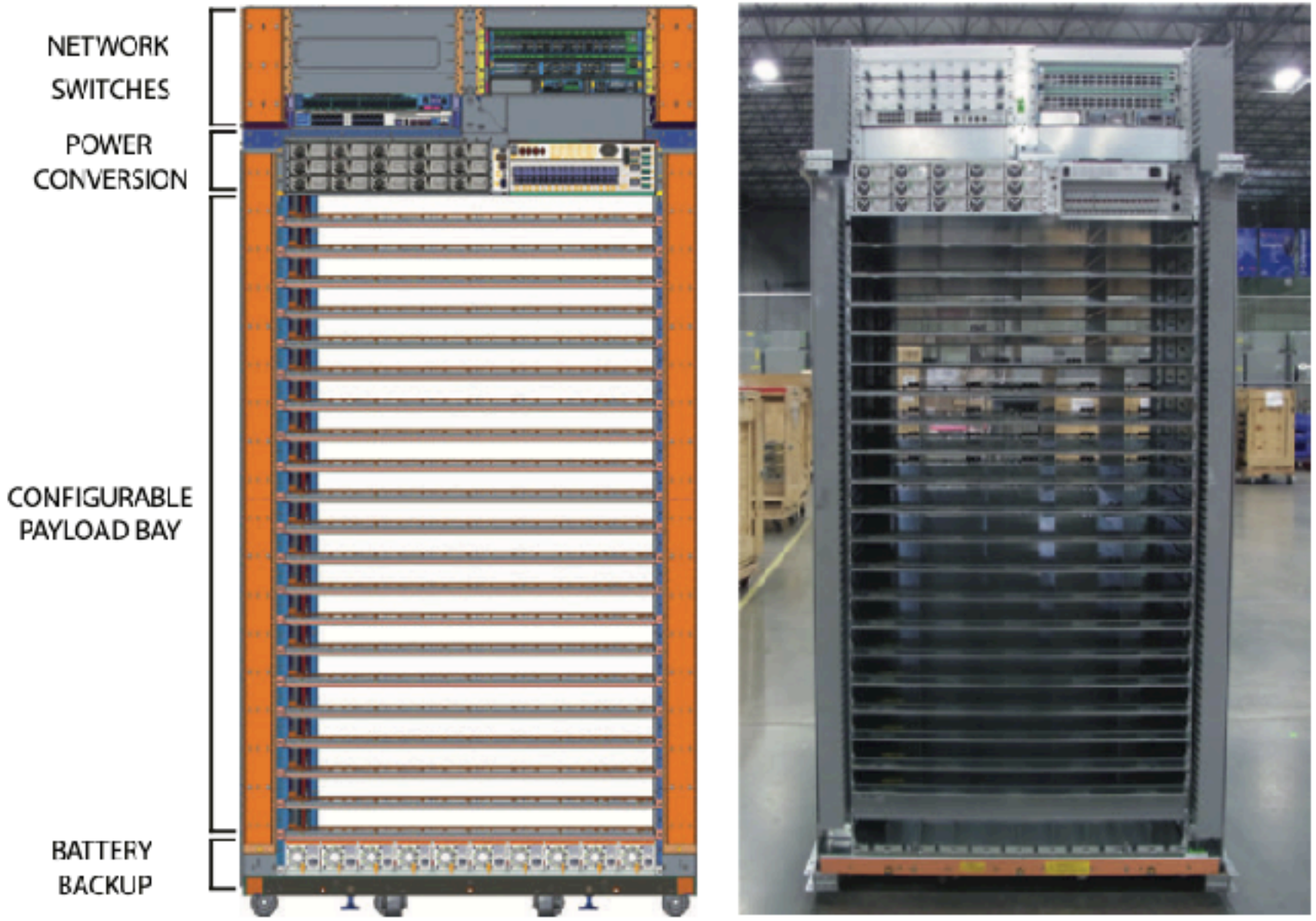


Figure 3.3: Machine racks like this support servers, storage, and networking equipment in Google's data centers.

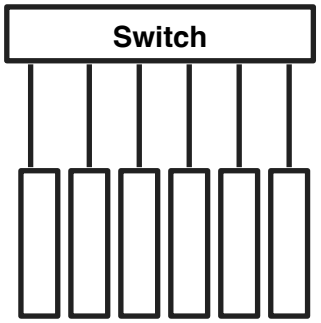
Hardware Building Blocks

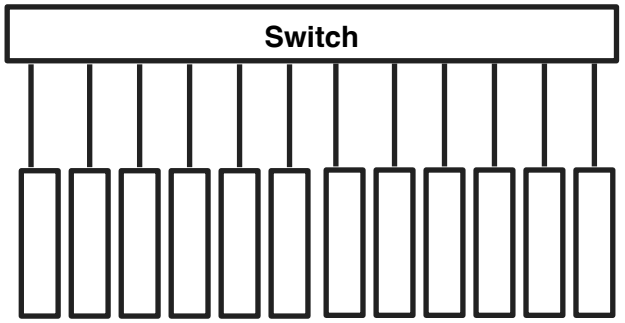
Data Center Networking

University of Cyprus
Department of Computer Science

Cluster networking

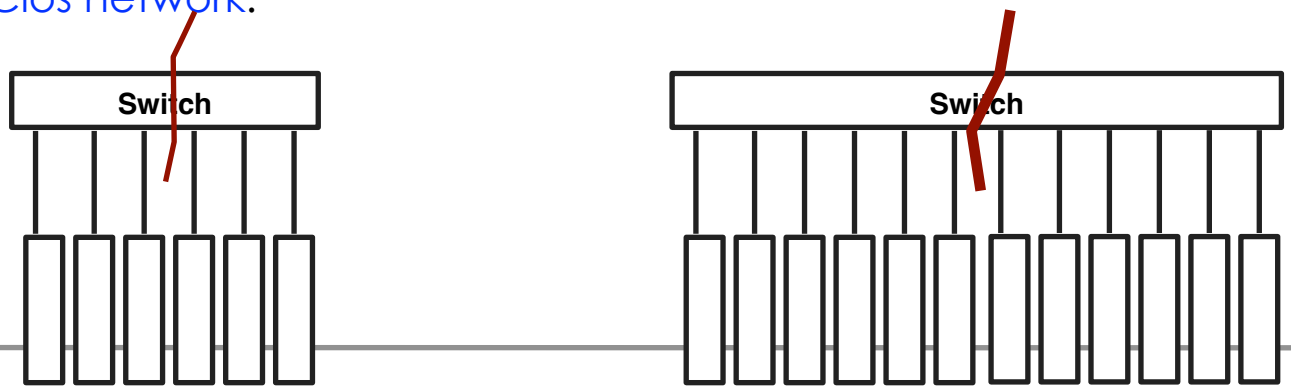
- Networking has no straightforward horizontal scaling solution
  - Doubling ***leaf* bandwidth** is easy
  - With twice as many servers, we'll have twice as many network ports and thus twice as much bandwidth.
- Is this good or bad?





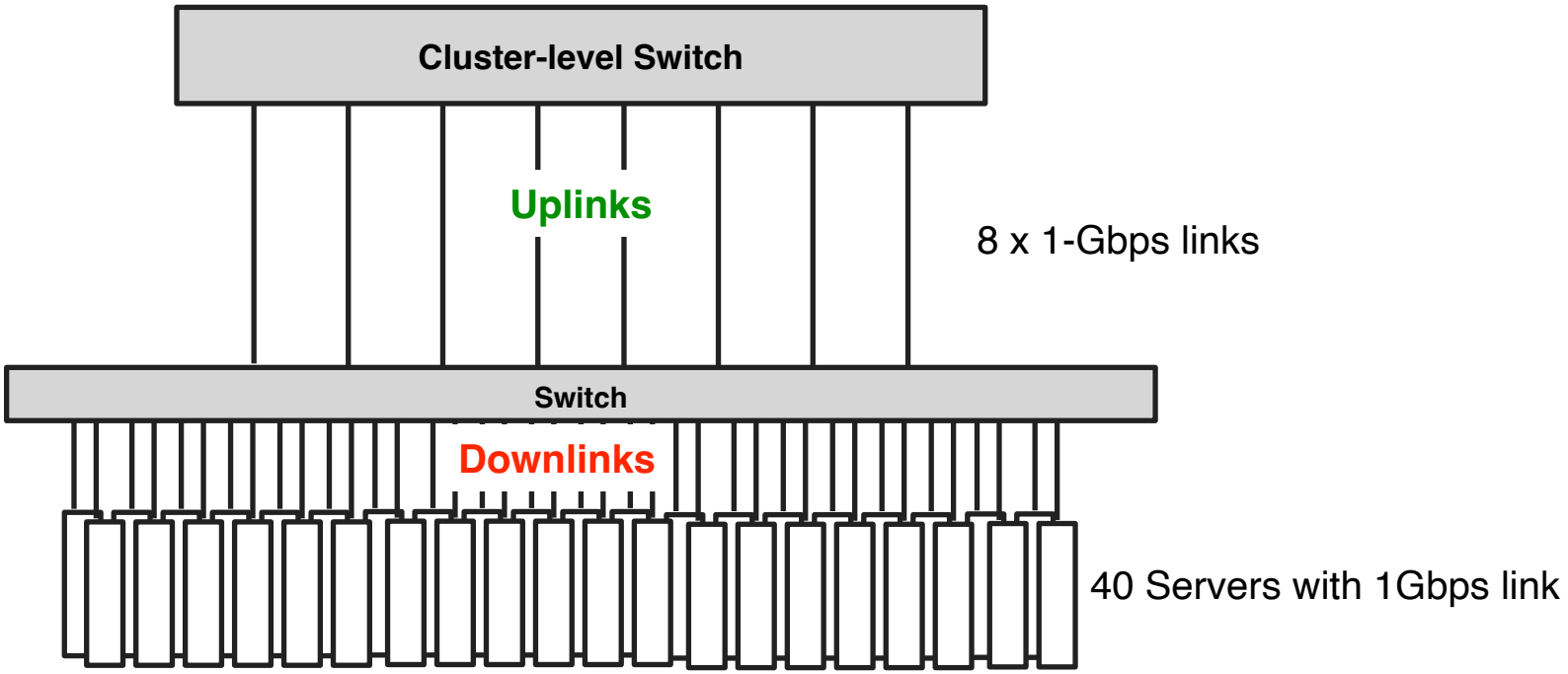
Cluster networking

- If every server needs to talk to every other server, we need to double **bisection bandwidth**: the **bandwidth across the narrowest line that equally divides the cluster into two parts**.
  - Using bisection bandwidth to characterize network capacity is common since *randomly communicating processors must send about half the bits across the "middle" of the network*.
- Doubling bisection bandwidth is **difficult**
  - We can't just buy (or make) an arbitrarily large switch - typical silicon switch chip can support a bisection bandwidth of about **1 Tbps** (**16x 40 Gbps ports**)
  - Switch chips are pin- and power-limited in size: no chips can do 10 Tbps (2018).
  - We can build larger switches by cascading these switch chips, typically in the form of a fat tree or Clos network.



Rack networking

- Often convenient to connect the network cables at the top of the rack.
- **Top-Of the-Rack** switches:
  - Bandwidth within a rack of servers tends to have a homogeneous profile.
  - E.g., Google's Jupiter network uses TOR switches with **64x 40 Gbps** ports,
- TOR ports split between:
  - **downlinks** that connect rack servers to the TOR
  - **uplinks** that connect the TOR to the rest of the WSC network fabric (inter-rack communication)
- **Oversubscription Ratio:** the ratio between number of downlinks and uplinks.
  - determines how much the intra-rack fabric is over-provisioned with respect to the data center fabric (typical values **between 5 and 10**)

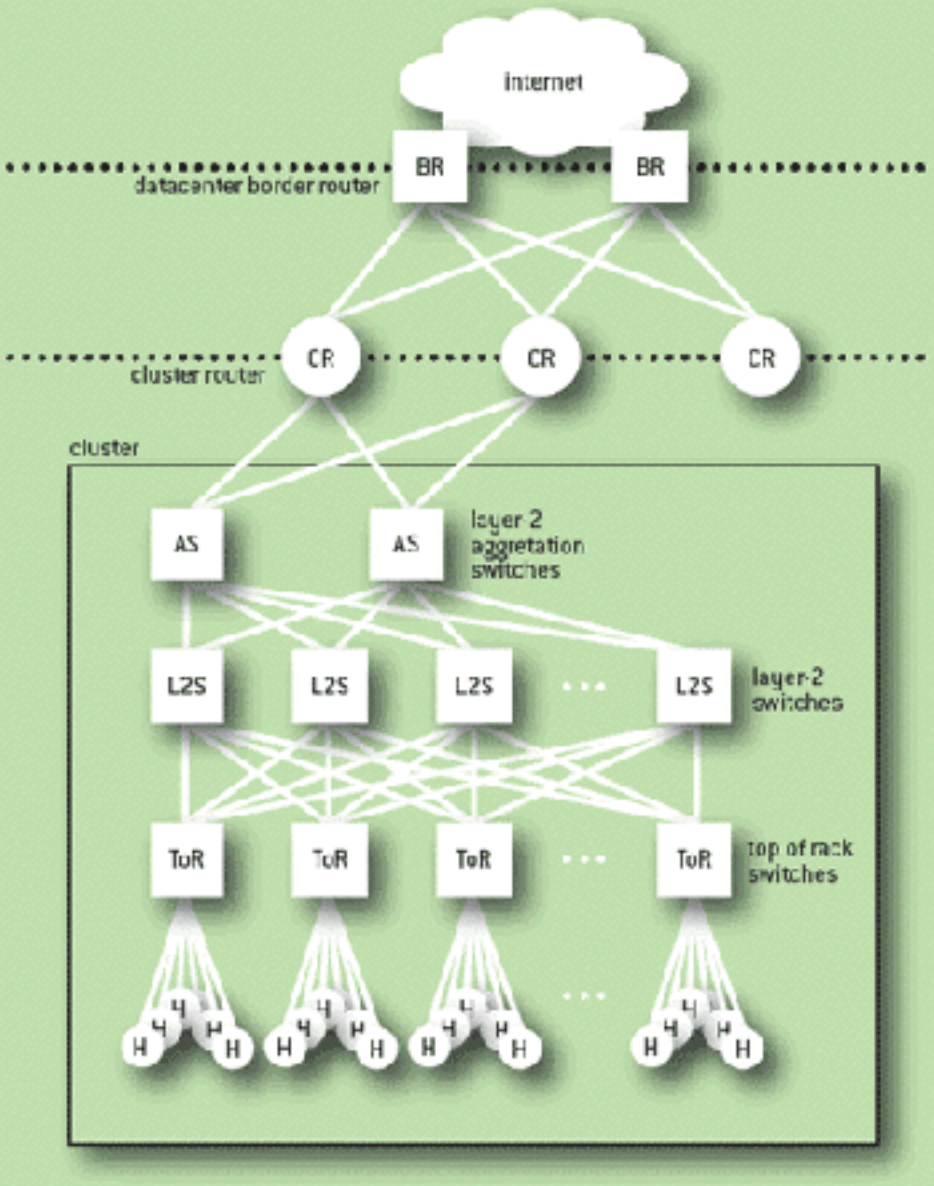

Cluster-level Switch
Uplinks
8 x 1-Gbps links
Switch
Downlinks
40 Servers with 1Gbps link

Network Fabric Design

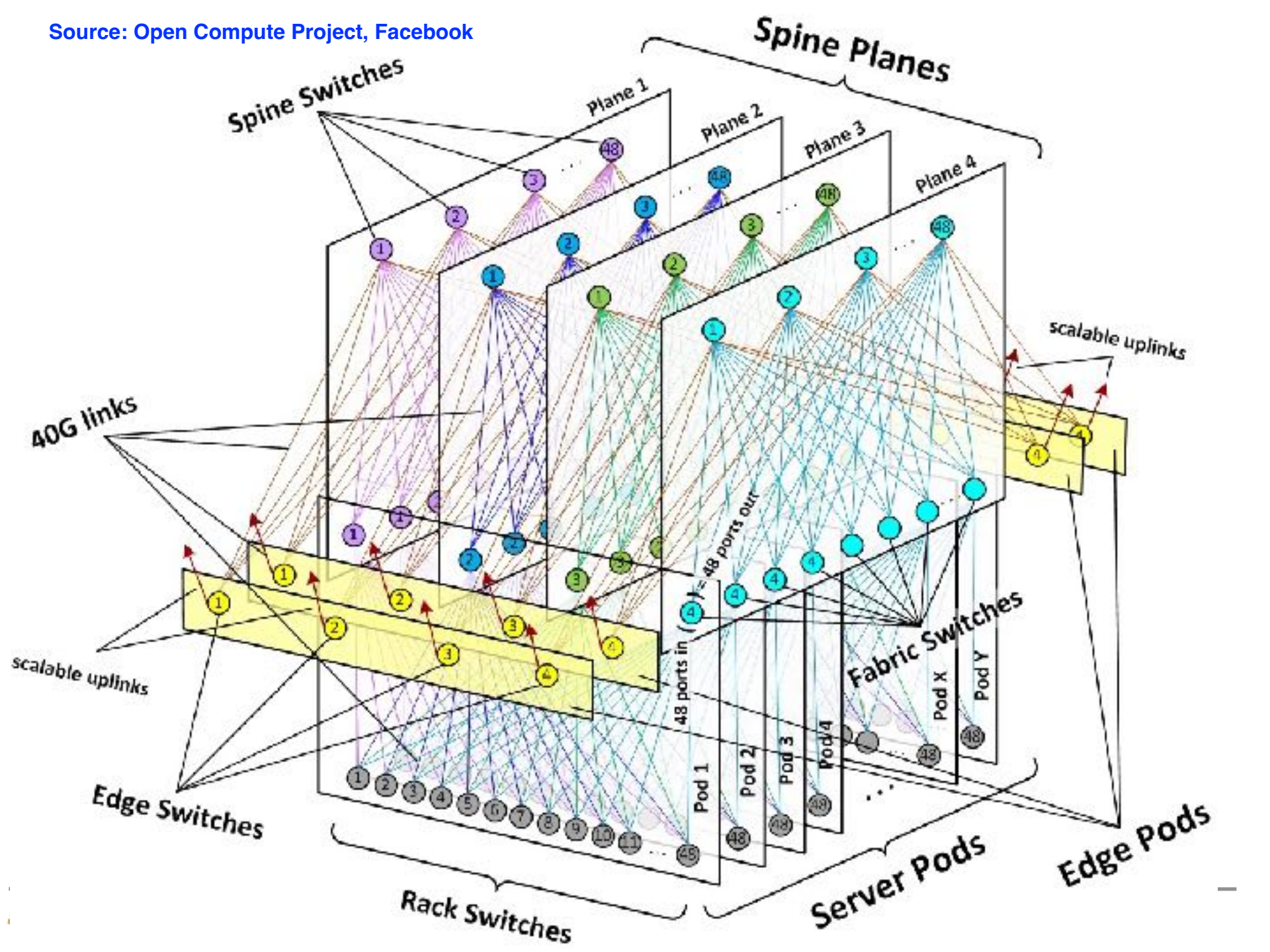
- Trade-off between **speed**, **scale**, and **cost**.
- TOR switches: Low cost, commodity components w **40-100 Gbps** ethernet switches with up to **48 ports** (**<$10/Gbps per server** to connect a single rack).
- **Network switches** with high-port counts (needed to tie together WSC clusters) **> 10x** more expensive.
  - A switch that has 10 times the bi-section bandwidth of a TOR switch, costs about **100 times as much**.
- Because of this cost discontinuity, WSC networking fabric is often organized as a two-level hierarchy.

FIGURE 2

A Conventional Tree-Like Datacenter Network Topology



Source: Open Compute Project, Facebook



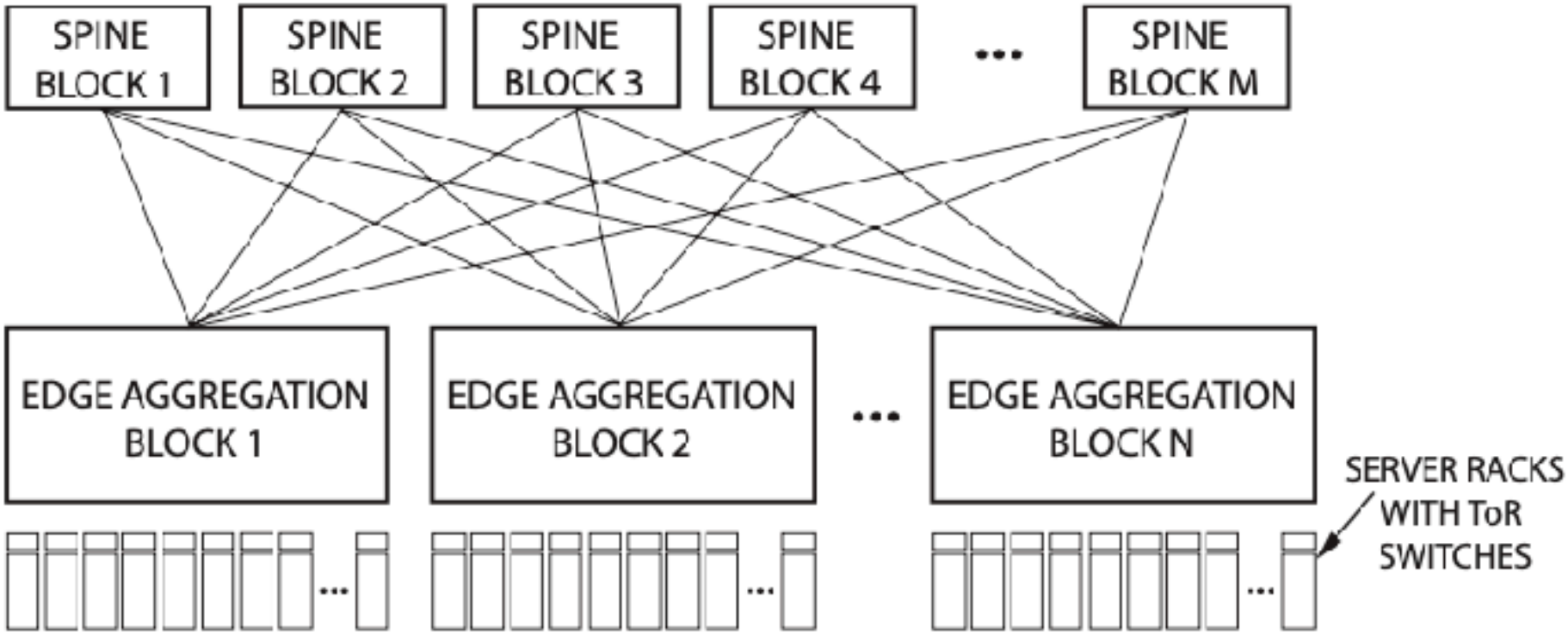


Figure 3.12: Sample three-stage fat tree topology. With appropriate scheduling this tree can deliver the same throughput as a single-stage crossbar switch.

Programming implications

- Programmers must:
  - be aware of the relatively scarce cluster-level bandwidth resources and
  - try to exploit rack-level networking **locality.**
- This complicates software development and possibly impacting resource utilization.

Alternative solutions

- Spending more money on the interconnect fabric:
  - **Infiniband** interconnects typically **scale to a few thousand ports** but can cost **$500–$2,000 per port**.
  - Larger-scale Ethernet fabrics, at a cost of at least **hundreds of dollars** per server.
- Lower-cost fabrics can be formed from commodity Ethernet switches by building "fat tree" Clos networks.
- How much to spend on networking vs. spending the equivalent amount on buying more servers or storage is an *application-specific question* that has no single correct answer.
- However, for now, assume that **intra-rack connectivity is often cheaper than inter rack connectivity.**

Hardware Building Blocks

Storage

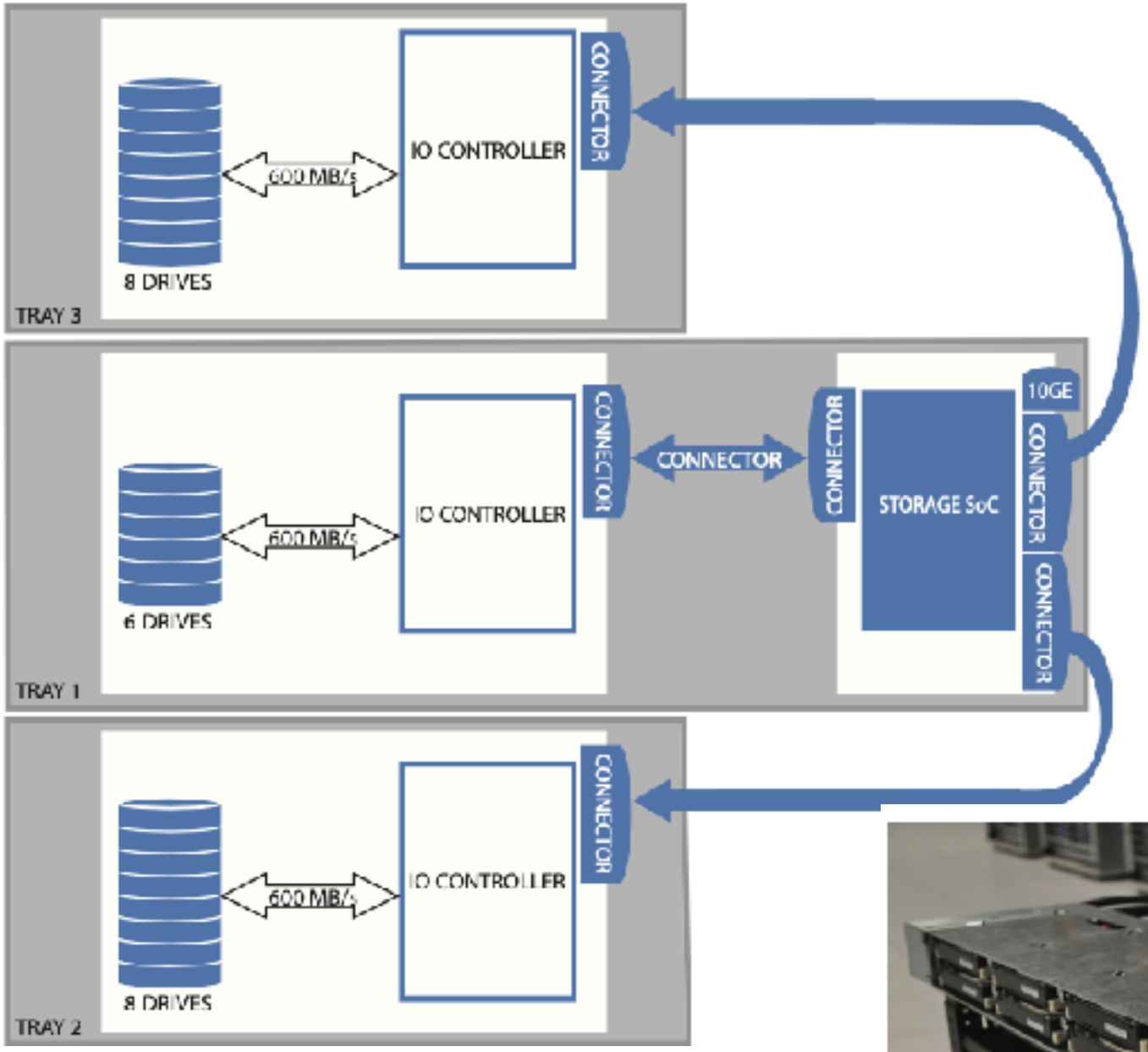
WSC Data Categories

Data manipulated by WSC workloads tends to fall into two categories:

- Data **private** to individual running tasks:
  - tends to reside in local DRAM or disk
  - rarely replicated
  - its management is simplified by virtue of its single user semantics.
- Data that is part of **shared state** of distributed workload
  - must be much more durable
  - accessed by a large number of clients
  - requires a much more sophisticated distributed storage system

Storage Building Blocks

- Disks and Flash SSDs are the building blocks of today's WSC storage system.
- Disk drives:
  - Connected directly to each individual server - **Directly Attached Storage** (DAS) and managed by a **global filesystem** (like GFS, GCS) or
  - Part of a Network Attached Storage (**NAS**) device connected directly to cluster-level switch.


TRAY 3
8 DRIVES
600 MB/s
IO CONTROLLER
CONNECTOR
TRAY 1
6 DRIVES
600 MB/s
IO CONTROLLER
CONNECTOR
CONNECTOR
CONNECTOR
STORAGE SoC
10GE
CONNECTOR
CONNECTOR
TRAY 2
8 DRIVES
600 MB/s
IO CONTROLLER
CONNECTOR

HD requirements

- Current hard drives designed for enterprise servers and not specifically for WSC.
- "Cloud" disks should aim for:
  - Higher I/O per second (IOPS): typically limited by seeks
  - Higher capacity
  - Lower tail latency when used in WSC
  - Meeting security requirements
  - Lower TCO

Network-attached Storage (NAS)

- Simpler to deploy
- Extra reliability through replication / error correction capabilities within the appliance
- More expensive
- Single point of failure in case NAS machine fails
- Smaller read bandwidth

Distributed File Systems

- Provide unstructured and structured APIs for application developers (BigTable, Dynamo, Spanner)
- Lower hardware cost - cheaper disks
- Lower networking fabric utilization
- Requires a fault-tolerant F/S at cluster level
- Extra reliability through replication across different machines: more bandwidth to complete write operations
- Keeps data available even after loss of entire server or rack
- May allow higher read bandwidth since same data can be sourced from multiple replicas
- Enables distributed software to exploit data locality
- Higher fault rates of cheap disks - can be mitigated by distributed F/S replication

Unstructured Storage

- E.g. Google File System (GFS) / Colossus / Google Cloud Storage offer:
  - simple file-like abstraction
  - high throughput for thousands of concurrent readers/writers
  - robust performance under high hardware failures rates
- System architecture: **primary server** (master), handles metadata operations, and **thousands of chunkserver** (secondary) processes running on every server with a disk drive, manage the data chunks on those drives.
- **Fault tolerance** provided by replication across machines.
  - Cross-machine **replication** allows system to tolerate machine and network failures and enables **fast recovery**, since replicas for a given disk or machine can be spread across thousands of other machines.
  - Colossus and Google Cloud Storage support replication with more **space-efficient Reed-Solomon codes**, which reduce the space overhead of replication by **roughly a factor of two** over simple replication for the same level of availability.
- **High availability** and **fast recovery** achieved by distributing file chunks across the whole cluster in such a way that a small number of correlated failures is extremely unlikely to lead to data loss

Structured Storage

- Structured distributed storage systems offer database-like functionality, with structured and indexed datasets, and support small updates and complex queries: Google's Bigtable, Amazon's DynamoDB.
- These systems **sacrifice** features like **richness** of **schema representation** and **strong consistency**, in favor of higher performance and massive availability:
  - Bigtable, presents a **simple multi-dimensional sorted map** consisting of row keys (strings) associated with multiple values organized in columns, forming a distributed sparse table space. Column values are associated with timestamps in order to support versioning and time-series.
  - Eventual consistency in Bigtable and DynamoDB shifts the burden of resolving temporary inconsistencies to the applications using these systems.
- Second-generation structured storage systems such as Megastore and Spanner address application developer concerns, providing:
  - richer schemas
  - SQL-like functionality
  - simpler, stronger consistency models

Interplay of Storage & Networking

- Success of WSC distributed storage systems partially attributed to evolution of data center networking fabrics.
- Gap between networking and disk performance is so wide that **disk locality is no longer relevant** in **intra-data center computations**.
  - Dramatic simplifications in the design of distributed disk-based storage systems
  - Disk utilisation improvement: any disk can be utilised by any task, regardless of relative locality
- Single **enterprise flash device** can achieve well over **100x the operations throughput of a disk drive**; one server machine with multiple flash SSDs could easily saturate a single 40 Gb/s network port even within a rack:
  - Stretching DC fabric bisection bandwidth
  - Requiring more CPU resources in storage nodes to process storage operations at such high rates.
  - Looking ahead, rapid improvements in WSC network bandwidth and latency will likely match flash SSD performance and reduce the importance of flash locality.
- Emerging **non-volatile memory** (NVM) has the potential to provide even higher bandwidth and **sub-microsecond access latency:**
  - bridging the gap between today's DRAM and flash SSDs
  - presenting even bigger challenge for WSC networking.

Design Considerations

- Computer architects seek to find balance in WSC design so that important workloads are properly served.
- Smart programmers restructure algorithms to better match more inexpensive design alternatives.
- Find solutions by software-hardware co-design, while keeping machines **not too complex to program**.
- Most cost-efficient and balanced configuration for hardware may be a match with the **combined resource requirements of multiple workloads** and not necessarily a perfect fit for any one workload
- Provided there is reasonable amount of connectivity within a WSC, effort should be put on creating software systems that can flexibly utilize resources in remote servers:
  - effective use of remote disk drives may require that the networking bandwidth to a server be equal or higher to the combined peak bandwidth of all disk drives locally connected to the server.
- **Workload churn** challenge: the software base may evolve so fast that a server design choice becomes suboptimal during its lifetime (typically three to four years).
  - Data center facility lifetime **spans several server lifetimes**, or more than a decade or so.

CAN YOU DESIGN THE MEMORY HIERARCHY OF A WSC?

Storage hierarchy

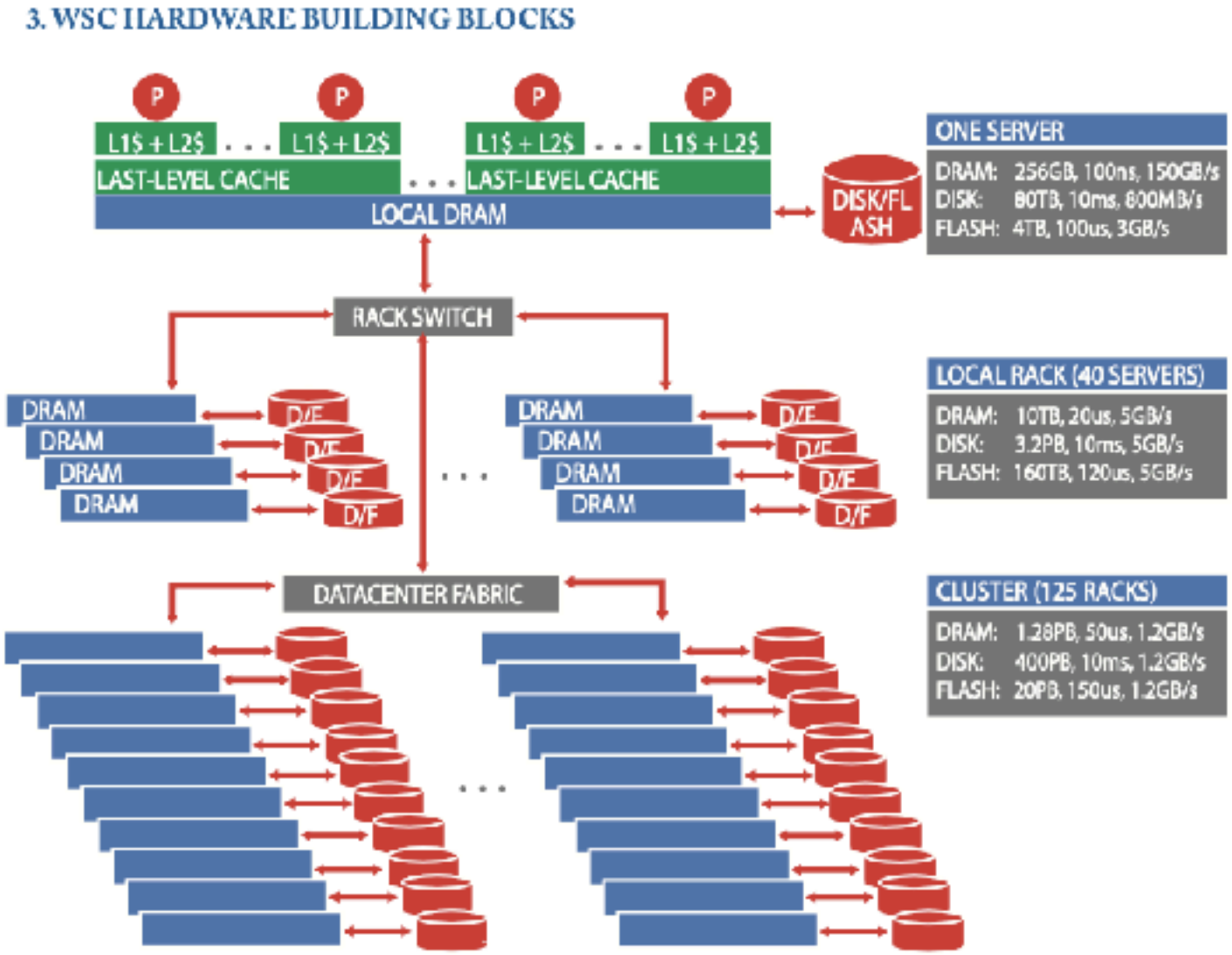


Figure 3.15: Storage hierarchy of a WSC.

Quantifying WSC Latency, Bandwidth, Capacity

- Assume a system with **5,000 servers**, each with **256 GB** of DRAM, one **4 TB** SSD, and **eight 10 TB** disk drives.
- Each group of 40 servers is connected through a **40-Gbps** link to a rack-level switch that has an **additional 10-Gbps uplink bandwidth per machine** for connecting the rack to the cluster-level switch (an **oversubscription factor of 4**).
- Network latency numbers assume TCP/IP transport
- Networking bandwidth values assume that each server behind an oversubscribed set of uplinks is using its fair share of the available cluster-level bandwidth.
- For disk latencies and transfer rates, assume typical commodity disk drive (SATA)

WSC Latency, Bandwidth, Capacity

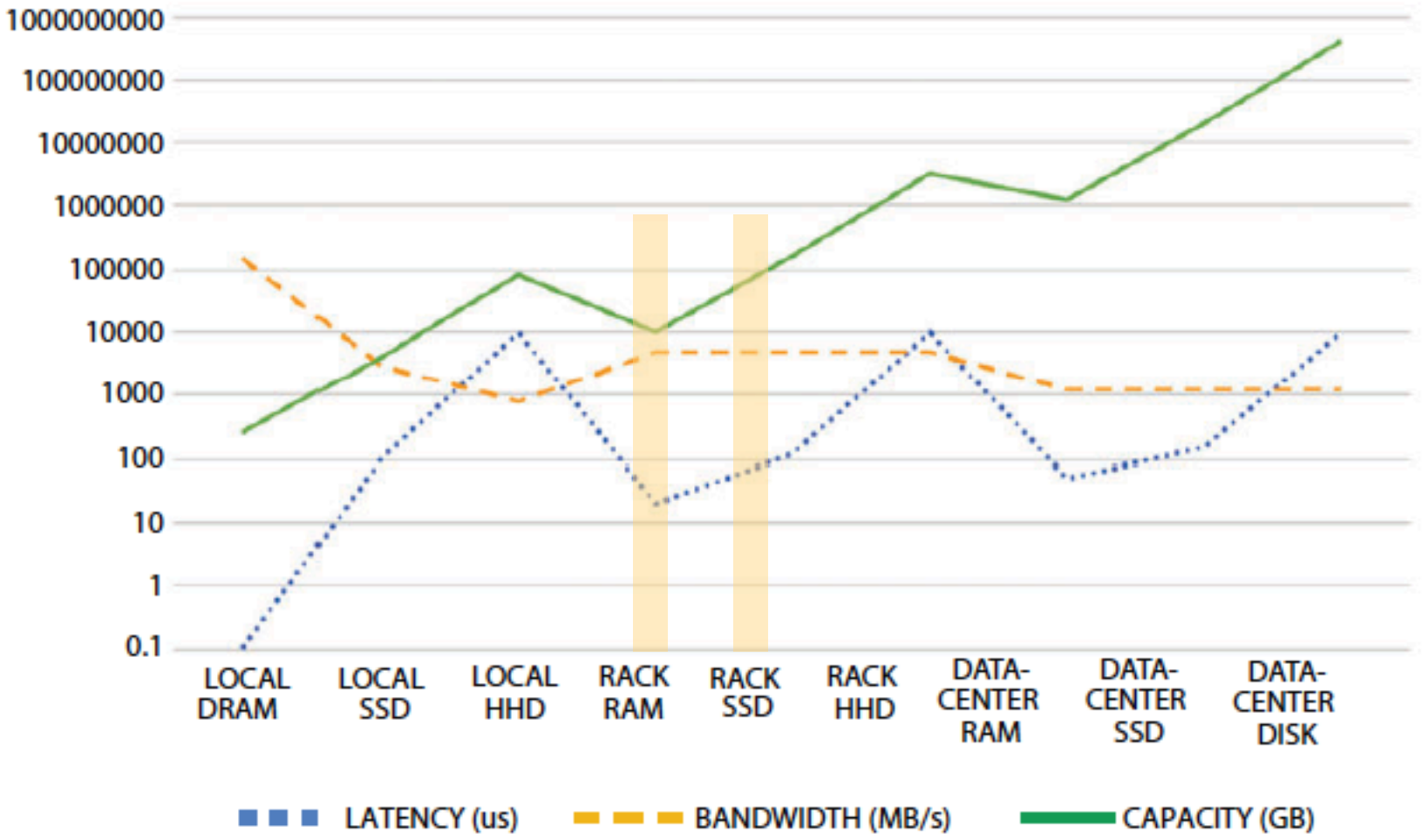


Figure 3.16: Latency, bandwidth, and capacity of WSC storage hierarchy levels.

2018

WSC Latency, Bandwidth, Capacity

- A large application that requires **many more servers than can fit on a single rack** must deal effectively with large discrepancies in **latency**, **bandwidth**, and **capacity**, which make it more difficult to program a WSC.
  - Key challenge for WSC architects: smooth out these discrepancies in a cost-efficient manner.
  - Key challenge for software architects: build cluster infrastructure and services that hide WSC complexity from application developers
- Hardware evolution offers solutions and new tradeoffs

Flash-based SSDs

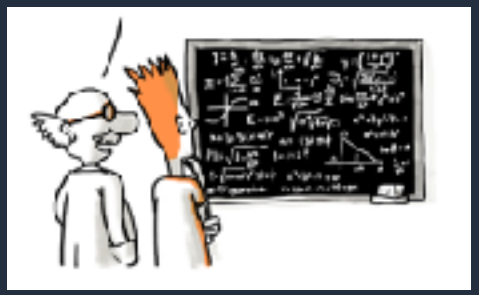
- Bridge the cost & performance gap between DRAM & disks.
- Flash's most appealing characteristic: performance **under random read operations**, is nearly **three orders of magnitude** better than HDD.
- Flash's performance is so high that it becomes a challenge to use it effectively in distributed storage systems since it demands:
  - much higher bandwidth from the WSC fabric
  - microsecond performance support from the hardware/software stack
- In the worst case, **writes to flash can be several orders of magnitude slower than reads**, and **garbage collection** can further increase write amplification and tail latency.

WHAT IS GARBAGE COLLECTION IN FLASH STORAGE?

NVM and fast SSD

- **Non-volatile memories (NVM)** and **fast SSD** products add another tier between today's DRAM and flash/storage hierarchy.
- NVM has the potential to provide **cheaper and more scalable alternatives to DRAM**, which is fast approaching its scaling bottleneck.
- NVM presents challenges for WSC architects who now have to consider **data placement, prefetching, and migration over multiple memory/storage tiers**.
- NVM and flash present new performance and efficiency challenges and opportunities, as traditional system design and software optimizations lack support for their microsecond (□s)-scale latencies.
- A new set of hardware and software technologies are needed to provide a simple programming model to achieve high performance

Summary of previous lecture



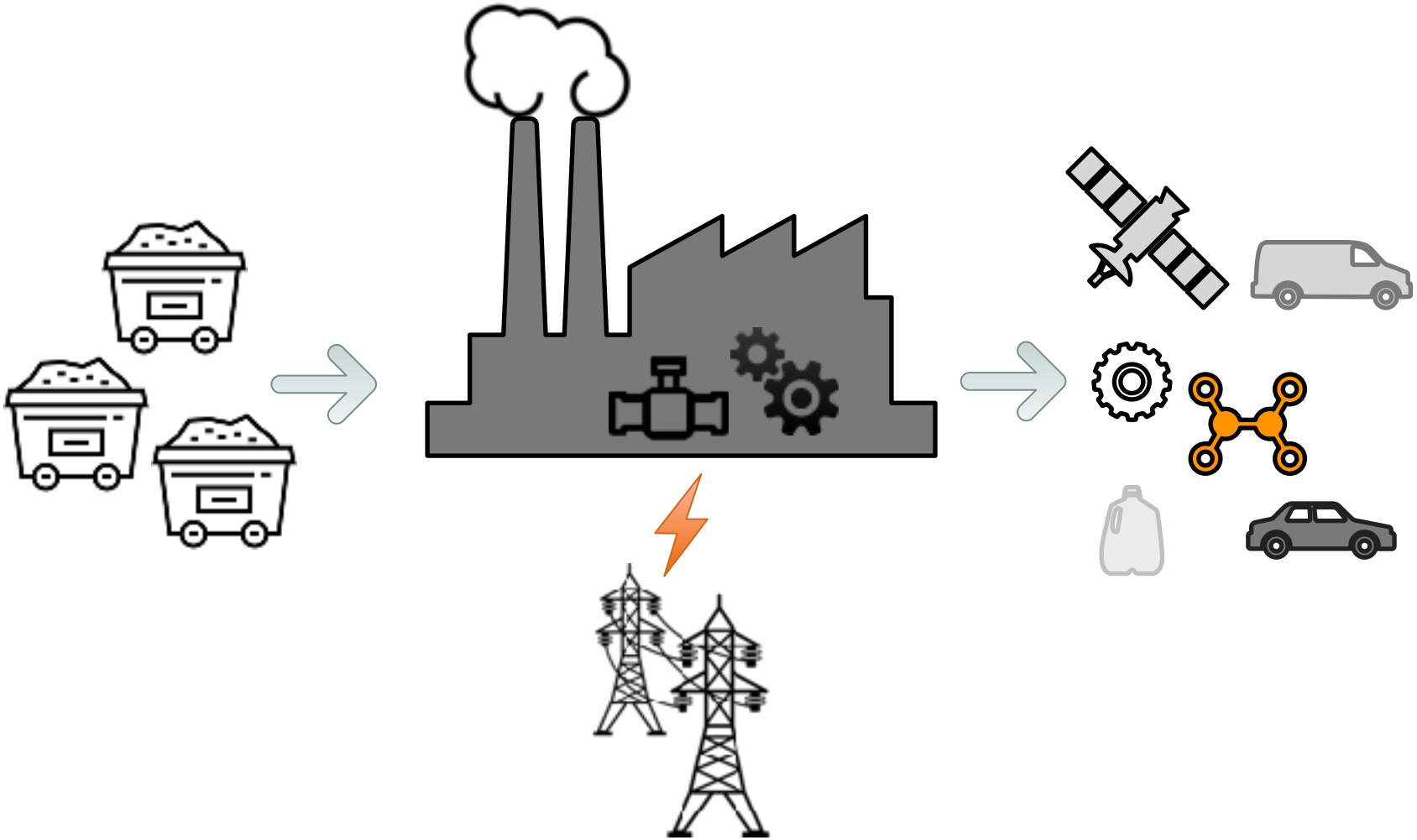
- Explained why public cloud infrastructures rely on Warehouse scale computers for deploying and running their services.
- Reviewed the key requirements for large-scale public cloud infrastructures (WSC): high availability, cost-efficiency, fault-free operation, and the implications thereof.
- Examined the IT architecture of WSC, and the key characteristics of WSC clusters.
- Introduced terms like *Form Factor, Thermal Design Power, Total Cost of Ownership, TOR Switch, Bisection Bandwidth, Network Fabric, Network Attached Storage (NAS), Oversubscription Ratio of Intra to Inter-rack networking*
- Analyzed the key design considerations that determine the Form Factor of WSC servers
- Reviewed key percentages regarding energy usage of IT and other components of a WSC, and how these can drive decisions on workload management.
- Examined the structure of WSC network fabric and their main components.
- Discussed the concept of storage hierarchy, the components of the storage hierarchy of a WSC, and their key performance characteristics.
- Reviewed the basic functionality of distributed file systems, unstructured and structured storage of WSC.
- Discussed the pros and cons of Flash Storage vs Hard Drives and the implications of flash storage on networking performance requirements.
- Explored the main parts of a WSC's building infrastructure and their role and requirements.

Data Centers and Warehouse Scale Computers

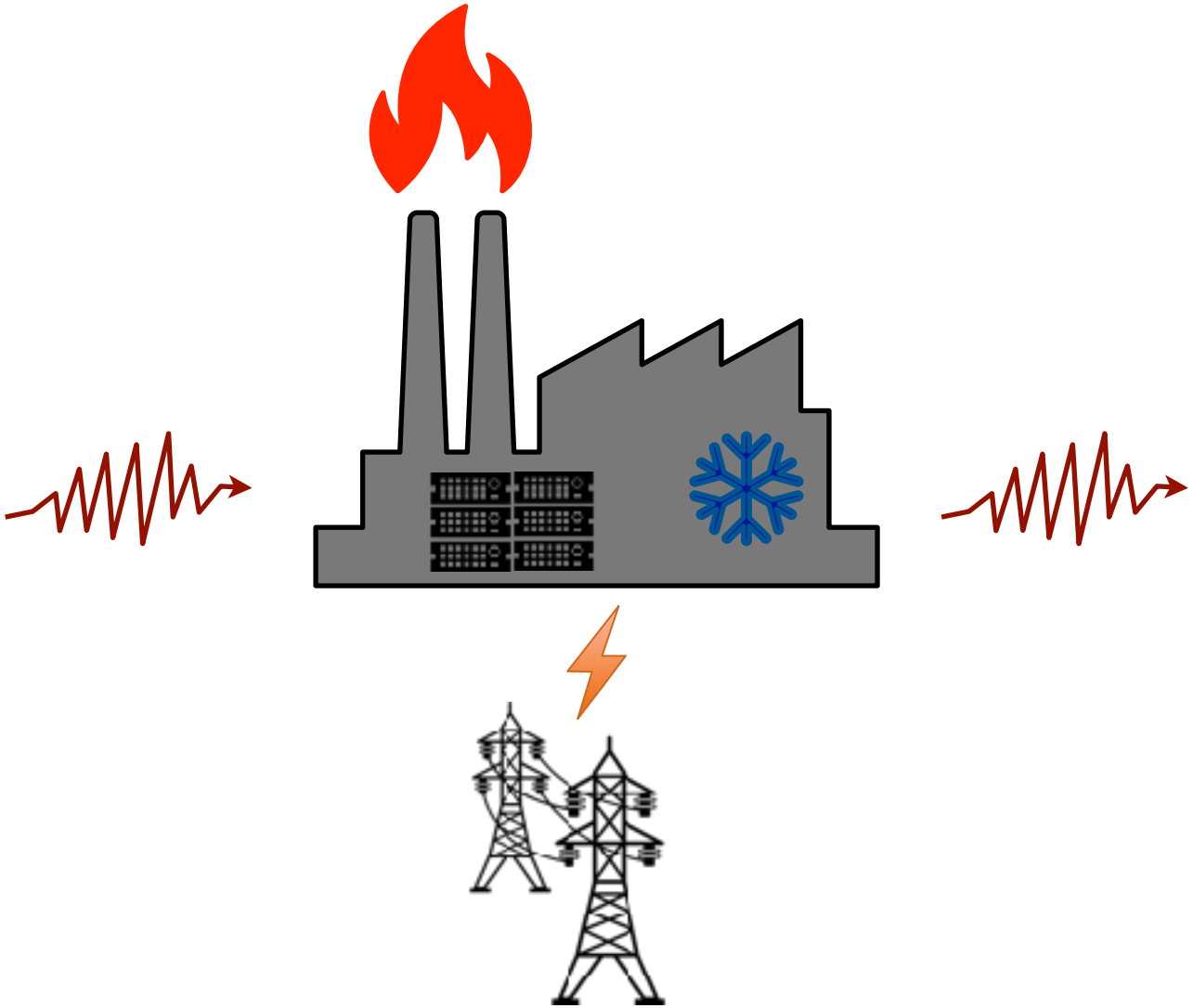
Building Infrastructure

University of Cyprus
Department of Computer Science

Industrial Buildings' Function



Datacenter Buildings



Datacenter Buildings

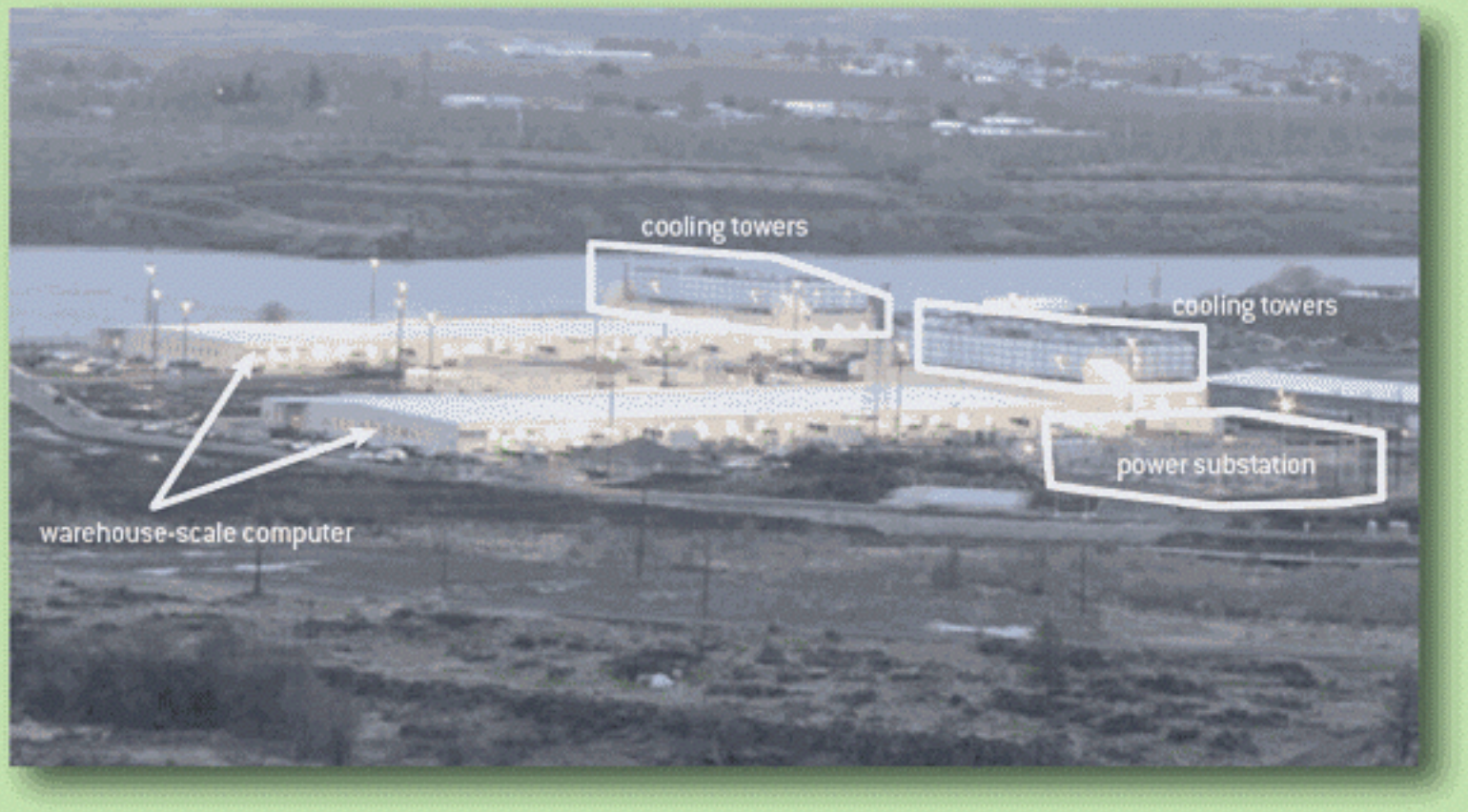
- By classic definitions, there is **little work produced at the data center**.
- Other than some **departing photons**, all of the energy consumed is converted into heat.
- The **delivery of input energy** and **subsequent removal of waste heat** are at the heart of the data center's design and drive the vast majority of *non-computing costs* -
  - in range of **$10–20 per watt**, but can vary considerably depending on size, location, and design.

DC Building Components

- **Mechanical yard** or central utility building, hosting cooling towers and chillers.
- **Electrical yard**: hosts generators, power distribution centers.
- **Main server hall**: hosts IT equipment, organised into hot and cold aisles.
- **Networking areas**: inter-cluster, campus-level, facility management, long-haul connectivity.
  - Additional physical security and high-availability features to ensure increased reliability
- Buildings follow established codes for fire-resistance, non-combustible construction, safety, secure access, cameras etc.

FIGURE 1

An Example Data Center and Warehouse-Scale Computer



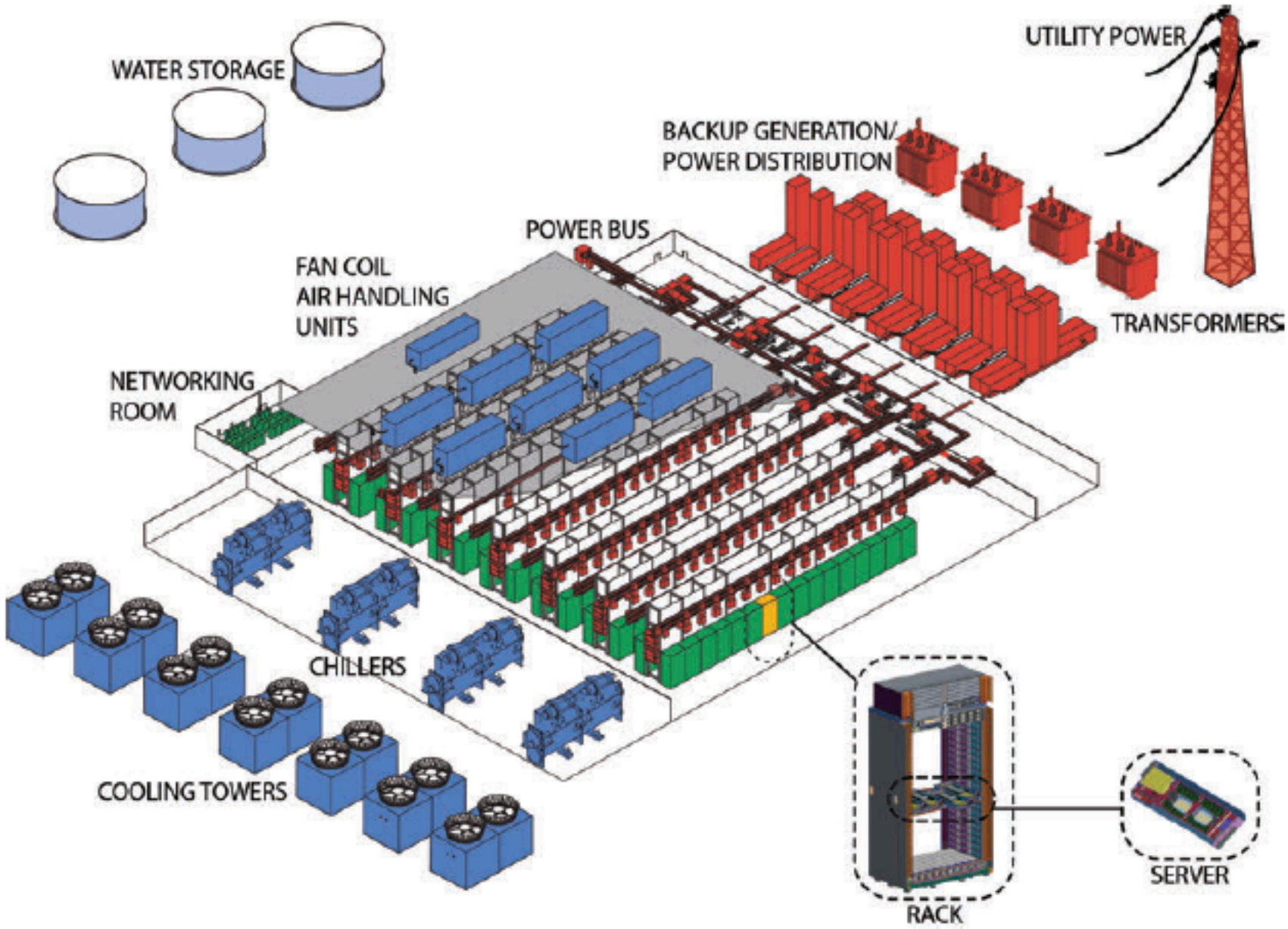
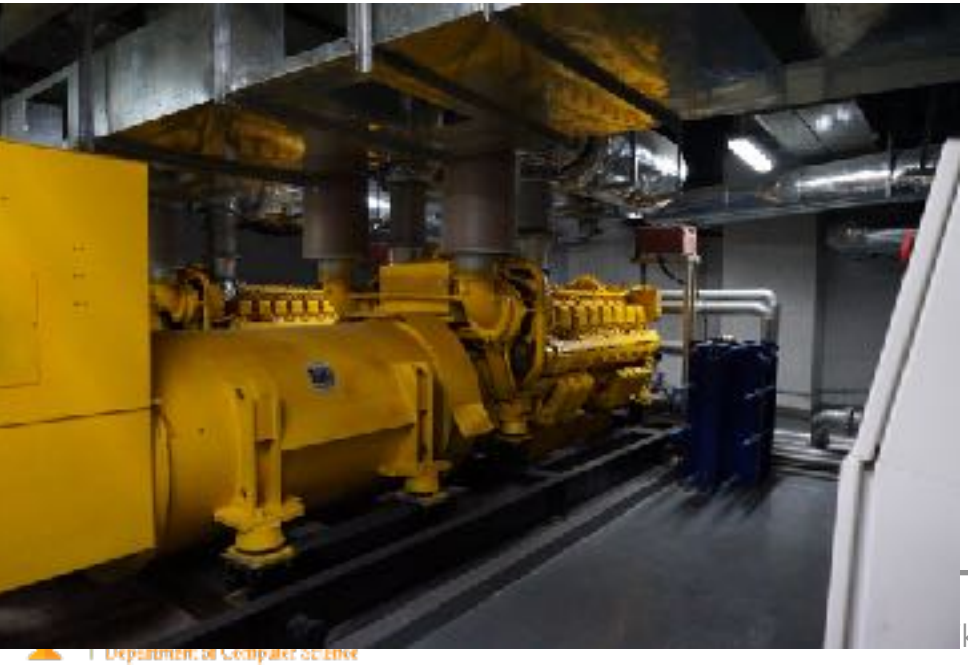
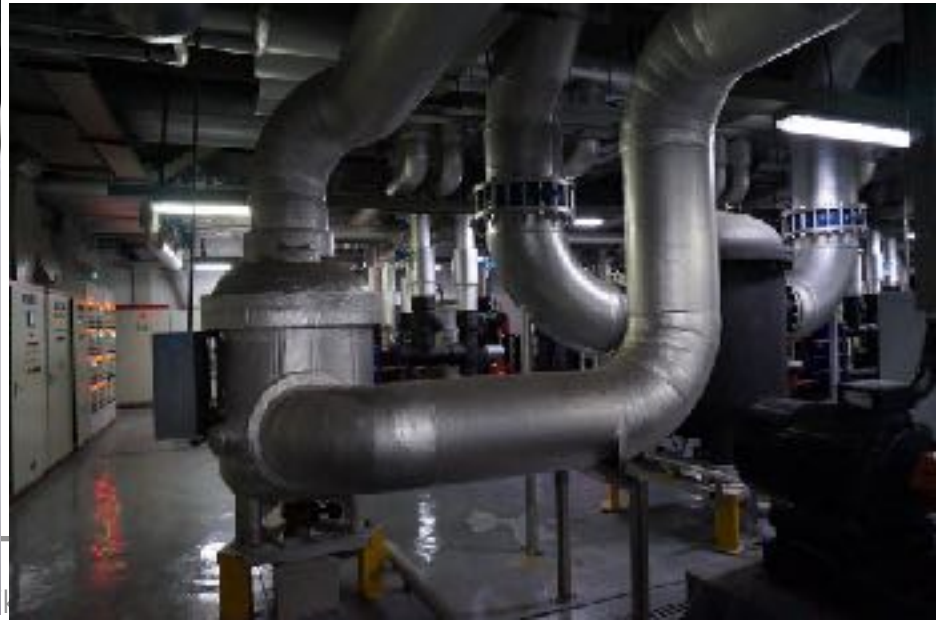


Figure 4.4: The main components of a typical data center.


UPS设备间
UPS Room





Data Centers and Warehouse Scale Computers

Tier Classification System

4-Tier Data Center Classification

- Standardized ranking system that indicates the reliability of data center infrastructure.
- Classification ranks facilities from 1 to 4, with 1 being the worst and 4 the best-performing level.
- Loosely based on **redundancy** built into the DC infrastructure for:
  - Uninterruptible power supply (UPS)
  - Power distribution
  - Backup generators
  - Cooling delivery
- Reliability goes up with higher levels
- Tier 4 is not always a better option than a data center with a lower rating: Each tier fits different business needs, so tiers 3 or 4 (the most expensive options) are often an over-investment.

Uptime Institute®

WHAT IS REDUNDANCY?

"the duplication of critical components or functions of a system with the intention of increasing reliability of the system, usually in the case of a backup or fail-safe"

WHY IS IT IMPORTANT?

WHAT IS THE COST OF DOWNTIME?

Hourly Cost of Downtime now exceeds $300,000 for 91% of SME and large enterprises.

Overall, 44% of mid-sized and large enterprise survey respondents reported that **a single hour of downtime**, can potentially cost their businesses over one million ($1 million).

catastrophic outage that interrupts a major business transaction or occurs during peak business hours can exceed millions of dollars per minute.

ITIC Annual Hourly Cost of Downtime survey, 2022

University of Cyprus
Department of Computer Science

Data Center Redundancy

- Not a "one-size-fits-all" endeavor.
- Building a redundant architecture is increasingly expensive as more components are added.
- Redundancy measures are characterized as: N, N+1, N+2, 2N and 2N+1.
- **N** refers to the **minimum capacity** needed to **power** or **cool** a data center at full IT load:
  - Does not include any redundancy: susceptible to single points of failure.
  - E.g. if a DC requires **4 UPS units** to operate at full capacity, then **N = 4**.

N+1 Redundancy

- Adds one independent backup component—a UPS, HVAC system or generator—to the N architecture to support a failure or allow a single machine to be serviced.
- N+1 systems:
  - Provide a minimal level of resiliency:
    - When one system is offline, the extra component takes over its load.
  - Are **not fully redundant** and can still fail because they run on common circuitry or feeds at one or more points rather than two completely separate feeds.
- Data centers with **N+1 redundancy** typically ensure that **a UPS system is always available:**
  - N+1 is for the number of UPS modules required to handle adequate supply of power for essential connected systems, plus one more.

2N Redundancy

- **2N systems** contain double the amount of equipment needed that **run separately** with **no single points of failure**.
- Offer a fully redundant system that can be easily maintained on a regular basis without losing any power to subsequent systems.
- In the event of an extended power outage, a 2N system will still keep things up and running.
- Some data centers offer **2N+1**, which is actually double the amount needed plus an extra piece of equipment as well.

4-Tier Datacenter Classification

- **Tier 1**: A single path for power and cooling, and no backup components.
  - Expected uptime: **99.671%** per year.
- **Tier 2**: A single path for power and cooling, and some redundant and backup components.
  - Expected uptime: **99.741%** per year.
- **Tier 3**: Multiple paths for power and cooling, and redundant systems that allow the staff to work on the setup without taking it offline.
  - Expected uptime: **99.982%** per year.
- **Tier 4**: Completely fault-tolerant data center with redundancy for every component.
  - Expected uptime of **99.995%** per year.
- Each tier includes the requirements of the lower rankings.

Tier I

- **Basic capacity level** with infrastructure to support IT for an office setting and beyond. The requirements for a Tier I facility include:
  - An uninterruptible power supply (UPS) for power sags, outages, and spikes.
  - An area for IT systems.
  - Dedicated cooling equipment that runs outside office hours.
  - An engine generator for power outages.
- Protects against disruptions from human error, but not unexpected failure or outage.
- Redundant equipment includes chillers, pumps, UPS modules, and engine generators.
- Facility will have to **shut down completely for preventive maintenance** and **repairs** - failure to do so increases the risk of unplanned disruptions and severe consequences from system failure.

Source: https://uptimeinstitute.com/tiers

Tier II

- Redundant capacity components for power and cooling that provide better maintenance opportunities and safety against disruptions:
  - Engine generators; Energy storage; Chillers; Cooling units; UPS modules; Pumps; Heat rejection equipment; Fuel tanks; Fuel cells.
- Tier II distribution path serves a critical environment, and the components can be removed without shutting it down.
- Unexpected shutdown of a Tier II data center will affect the system.

Source: https://uptimeinstitute.com/tiers

Tier III

- Concurrently maintainable with:
  - redundant components
  - redundant distribution paths.
- Unlike Tier I and Tier II, Tier III facilities require no shutdowns when equipment needs maintenance or replacement.
- The components of Tier III are added to Tier II components so that any part can be shut down without impacting IT operation.

Source: https://uptimeinstitute.com/tiers

Tier IV

- Several independent and physically isolated systems that act as redundant capacity components and distribution paths.
- Environment will not be affected by a disruption from planned and unplanned events.
- If redundant components or distribution paths are **shut down for maintenance**, the environment may experience a **higher risk** of disruption if a failure occurs.
- Tier IV facilities add fault tolerance to the Tier III topology:
  - When a piece of equipment fails, or there is an interruption in the distribution path, IT operations will not be affected.
  - All of the IT equipment must have a fault-tolerant power design to be compatible.
  - Continuous cooling to make the environment stable.

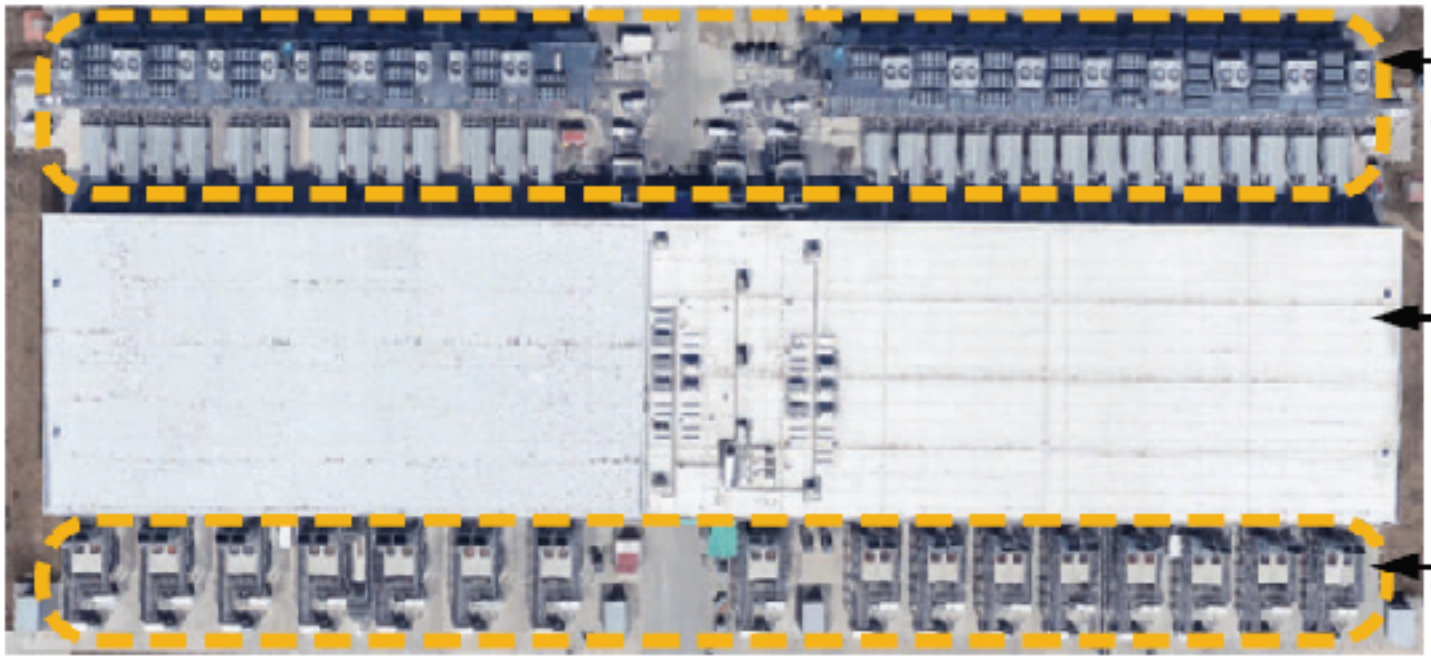
Comparison of Tiers

PARAMETERS	TIER 1	TIER 2	TIER 3	TIER 4
Uptime guarantee	99.671%	99.741%	99.982%	99.995%
Downtime per year	<28.8 hours	<22 hours	<1.6 hours	<26.3 minutes
Component redundancy	None	Partial power and cooling redundancy (partial N+1)	Full N+1	Fault tolerant (2N or 2N+1)
Concurrently maintainable	No	No	Partially	Yes
Price	$	$$	$$$	$$$$
Compartmentalization	No	No	No	Yes
Staffing	None	1 shift	1+ shift	24/7/365
Typical customer	Small companies and start-ups with simple requirements	SMBs	Growing and large businesses	Government entities and large enterprises
The main reason why companies select this tier	The most affordable data center tier	A good cost-to-performance ratio	A fine line between high performance and affordability	A fault-tolerant facility ideal for consistently high levels of traffic or processing demands

DC Building Size and Power

- DC Building **sizes** vary and commonly described in terms of:
  - **Floor area for IT equipment**
  - **Critical power:** maximum power that can be continuously supplied (to the IT infrastructure)
  - **2/3 of US DC** take up less than **464 m²** (5000 ft²) with less than 1MW of critical power.
  - Some DC are multi-story with critical power exceeding **100 MW**.


BUILDING 1
BUILDING 2
CENTRAL UTILITY
BUILDING &
WATER TANK
BUILDING 3
SUBSTATION 1
SUBSTATION 2


MECHANICAL YARD
CHILLERS, COOLING TOWERS, ...
MAIN SERVER HALL
MACHINE ROWS, NETWORK, OPERATION AREAS, ...
ELECTRICAL YARD
TRANSFORMERS, GENERATORS, ...

Data Center Basics

Power Systems

University of Cyprus
Department of Computer Science

Power Systems

- Power enters first at a utility substation which transforms **high voltage** (typically 110 kV and above) to **medium voltage** (typically less than 50 kV).
- **Medium voltage** is used for site-level distribution to the **primary distribution centers** (also known as **unit substations**), which include the primary switchgear and **medium-to-low** voltage transformers (typically below 1,000 V).
- From here, the power enters the building with the low-voltage lines going to the **uninterruptible power supply** (UPS) systems. The UPS switchgear also **takes a second feed at the same voltage** from a set of **diesel generator**s that cut in when utility power fails.
- The outputs of the UPS system are routed to the data center floor where they are connected to **Power Distribution Units (PDUs)**.

Power distribution

Figure 1.3: Power distribution, Council Bluffs, Iowa, U.S.

UPS

- The UPS typically combines three functions:
  - A **transfer switch** that chooses the active power input (either utility power or generator power). After a power failure, the transfer switch senses when the generator has started and is ready to provide power; typically, a generator takes **10–15 s** to start and assume the full rated load.
  - Some form of **energy storage** (electrical, chemical, or mechanical) to bridge the time between the utility failure and the availability of generator power.
  - It **conditions the incoming power feed**, removing voltage spikes or sags, or harmonic distortions in the AC feed. This conditioning can be accomplished via "double conversion" (AC-DC-AC).
- UPS systems take up a sizeable amount of space, they are usually housed in a room separate from the data center floor.
  - Typical UPS capacities range from **hundreds of kilowatts** up to **two megawatts** or more, depending on the power needs of the equipment.

Power Distribution Units (PDUS)

- PDUs resemble breaker panels in residential houses but can also incorporate transformers for final voltage adjustments.
- They take a larger input feed and break it into many smaller circuits that distribute power to the actual servers on the floor.
- Each circuit is protected by its own breaker, so a short in a server or power supply will trip only the breaker for that circuit, not the entire PDU or even the UPS.

Figure 2.1 Vertical Rack PDU Configuration

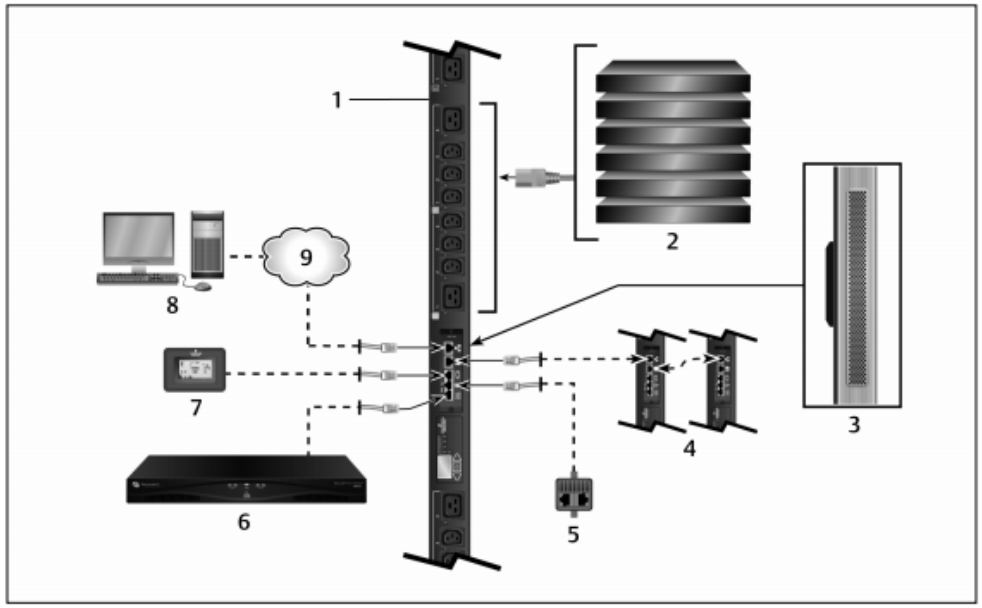


Table 2.1 Vertical rack PDU Configuraton Descriptions

ITEM	DESCRIPTION	ITEM	DESCRIPTION
1	Vertical rack PDU	6	Serial appliance
2	Connected equipment	7	RPC basic display module (BDM)
3	Case ventilation, both sides (Optional)	8	Monitoring station
4	Rack PDU array	9	Network connection (10 MB/100 MB/1 GB)
5	Sensors—integrated and modular		

Data Center Basics

Cooling Systems

Cooling Datacenters

- Purpose: Remove the heat generated by the DC equipment
- Historically, data centers have **consumed** twice as much energy as needed to power the servers, but when best practices are employed this overhead shrinks to **10–20%**.
- Key energy saving techniques in cooling systems:
  - free-cooling (further boosted by raising the target inlet temperature of servers)
  - well-managed air flow
  - high-efficiency power distribution
  - UPS components

Datacenter Cooling Systems

- To this end, a cooling system must employ some **hierarchy of loops**, each circulating a cold medium that warms up via some form of heat exchange and is somehow cooled again.
- An **open loop** replaces the outgoing warm medium with a cool supply from the outside, so that each cycle through the loop uses new material.
- A **closed loop** recirculates a separate medium, continuously transferring heat to either another loop via a heat exchanger or to the environment.
- All systems of loops must eventually transfer heat to the outside environment.

Fresh-air Cooling

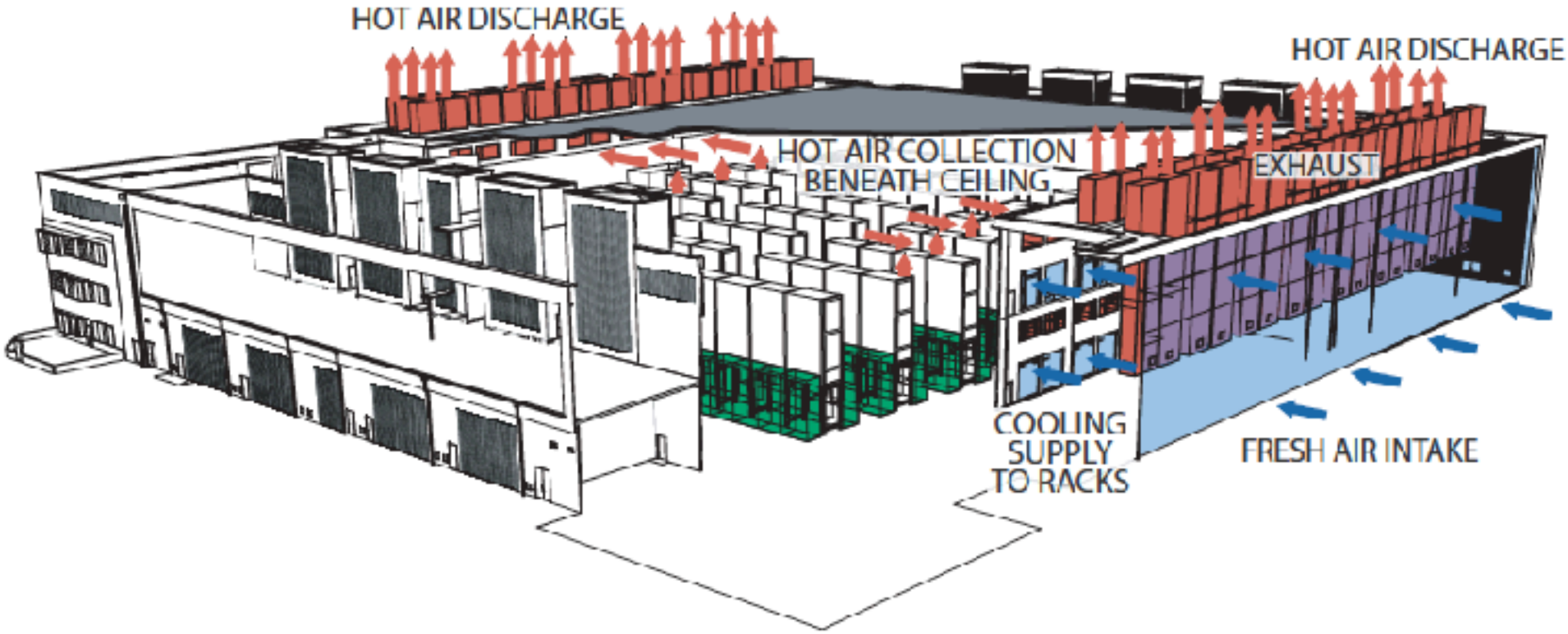


Figure 4.8: Airflow schematic of an air-economized data center.

CRAC Cooling (closed loop)

- Isolate and remove heat from the servers and transport it to a heat exchanger.

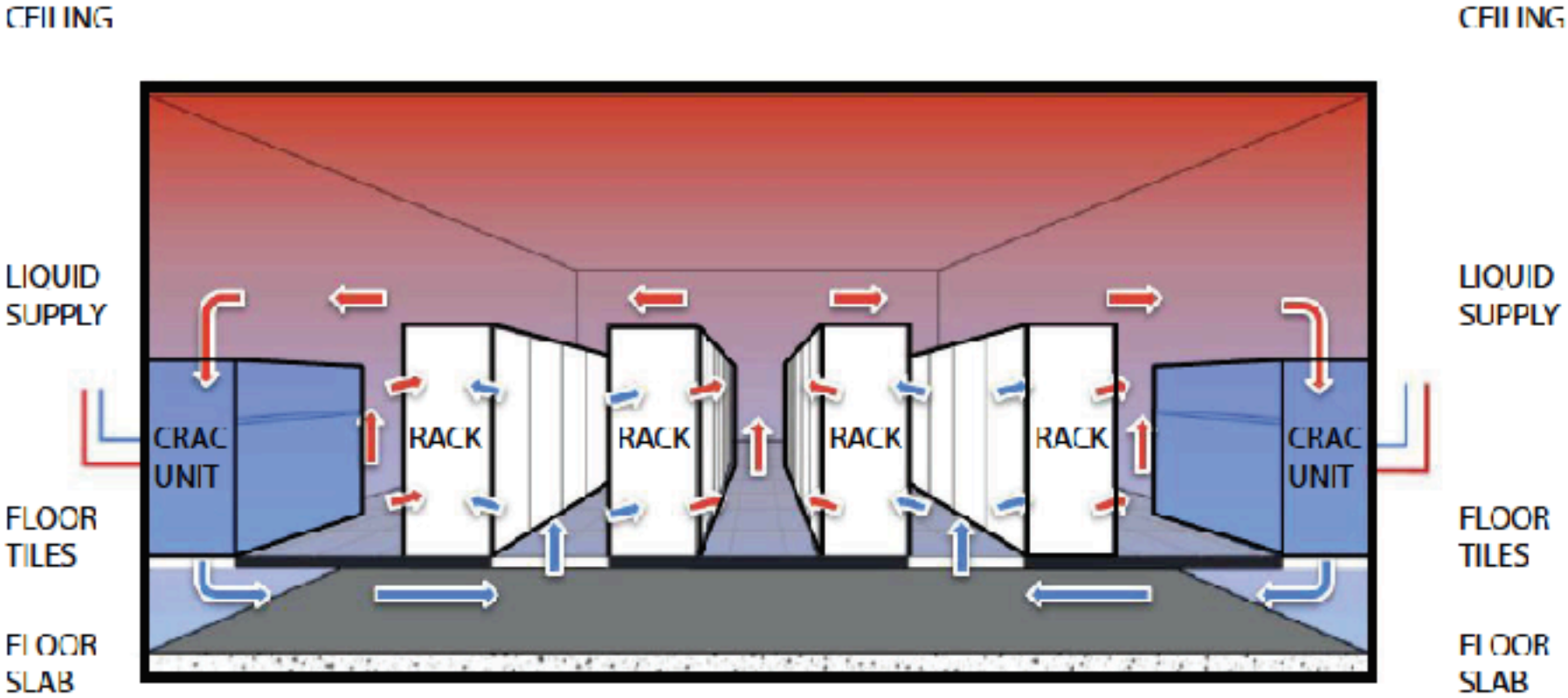


Figure 4.9: Raised floor data center with hot-cold aisle setup (image courtesy of DLB Associates [Dyc06]).

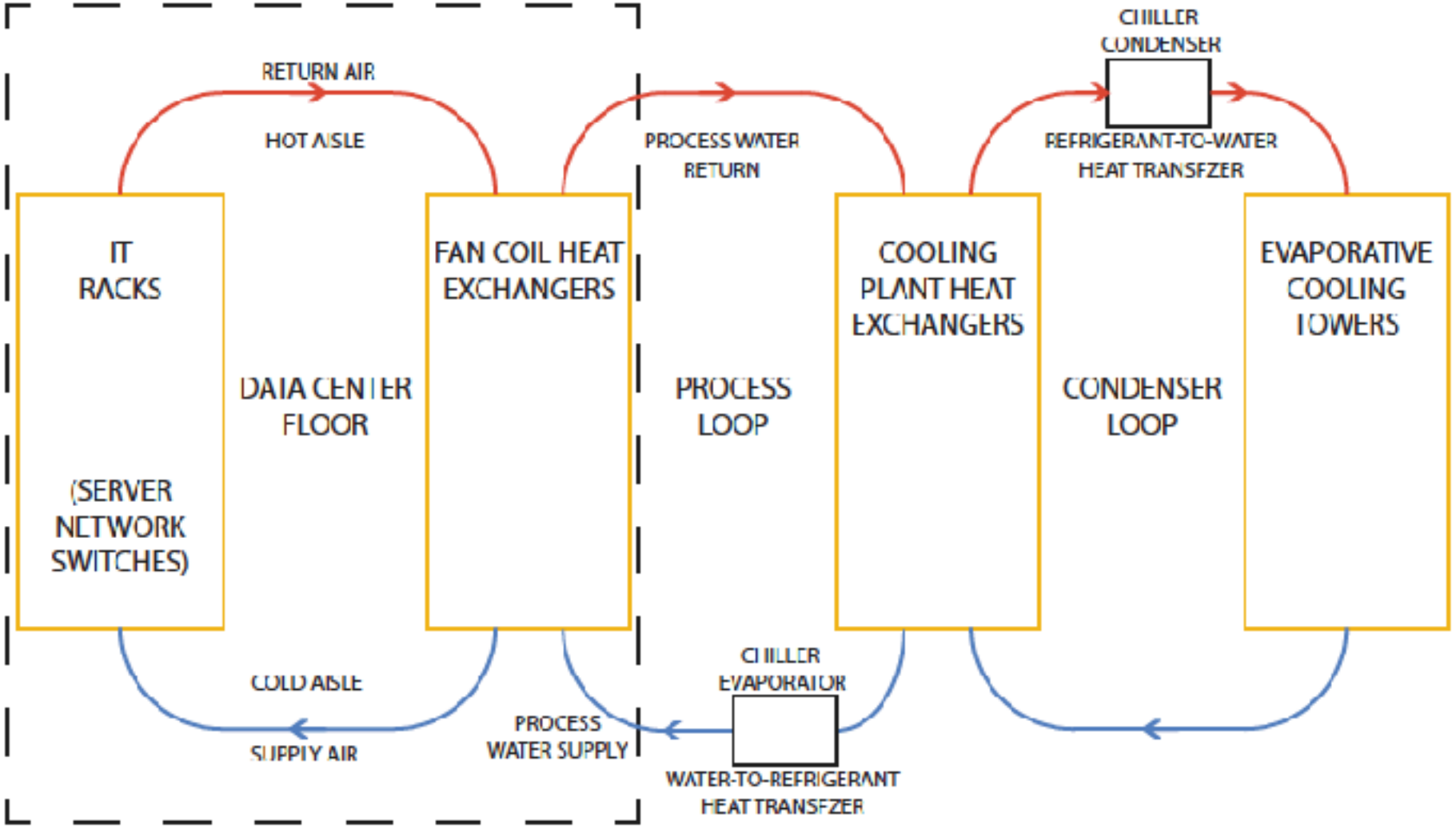


Figure 4.10: Three-loop data center cooling system. (Note that in favorable weather conditions, the entire data center heat load can be removed by evaporative cooling of the condenser water; the chiller evaporator and chiller condenser heat transfer steps then become unnecessary.)

Datacenter cooling

Figure 1.4: Data center cooling, Douglas County, Georgia, U.S.

Figure 4.11: Water-cooled centrifugal chiller.

Cooling towers

Power for Cooling

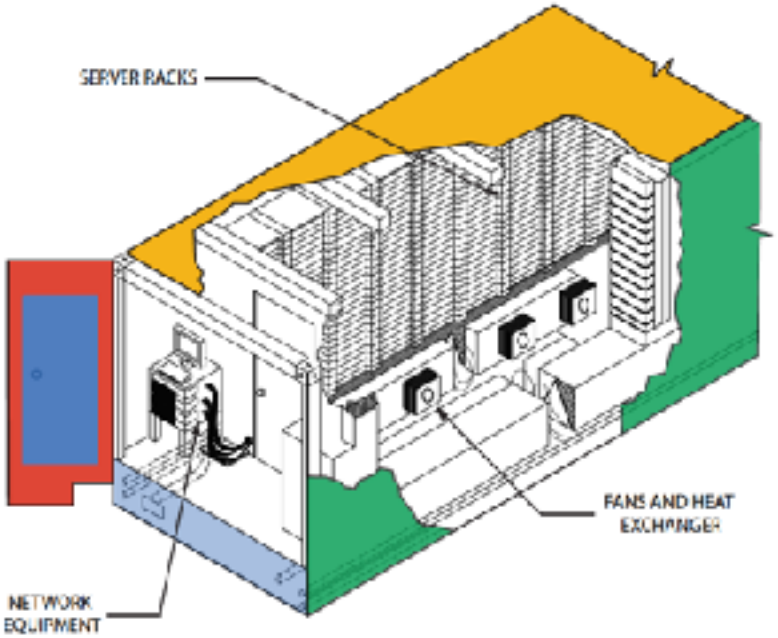
- Generators (and sometimes UPS units) provide backup power for most mechanical cooling equipment
  - DC may overheat in a matter of minutes without cooling.
- In a typical data center, chillers and pumps can add **40% or more to the critical load** supported by generators, significantly adding to the overall construction cost.

Tradeoffs

- Complexity, efficiency, and cost.
- **Fresh air cooling** can be very efficient but does not work in all climates, requires filtering of airborne particulates, and can introduce complex control problems.
- **Two-loop systems:** easy to implement, relatively inexpensive to construct, and offer isolation from external contamination, but typically have lower operational efficiency.
- **Three-loop system:** the most expensive to construct and has moderately complex controls, but offers contaminant protection and good efficiency.

Container-based Datacenters

- Server racks placed inside a container (typically 20 or 40 ft long) and integrate heat exchange and power distribution into the container as well.
- Provide all the functions of a typical data center room (racks, CRACs, PDU, cabling, lighting) in a small package.



Underwater Datacenters

Project Natick@Microsoft:

- A sealed container on the ocean floor could provide ways to improve the overall reliability of datacenters. On land, corrosion from oxygen and humidity, temperature fluctuations and bumps and jostles from people who replace broken components are all variables that can contribute to equipment failure.
  - The Northern Isles deployment confirmed their hypothesis.
- The proven reliability of underwater datacenters can be useful in deploying and operating tactical and critical datacenters anywhere in the world (near cost lines)
- "We are populating the globe with edge devices, large and small, to learn how to make datacenters reliable enough not to need human touch is a dream of ours."

William Chappell, vice president of mission systems for Azure.

Data Centers and Warehouse Scale Computers

Energy and Power Efficiency

University of Cyprus
Department of Computer Science

Energy Efficiency & Mobile Computing

- A major technology driver in the mobile and embedded computing:
  - extend battery life
  - reducing peak power because thermal constraints began to limit:
    - further CPU performance improvements
    - packaging density in small devices.

Energy Management in WSC

- Goal: **Reduce** all energy-related costs, including:
  - capital expenses
  - operating expenses
  - environmental impact

WSC Energy Efficiency

- Measures the **energy used to run a particular workload** (e.g. sort a petabyte of data)
  - It is hard to benchmark WSCs this way (why?).
  - Even though such benchmarks have been contemplated, they haven't yet been widely used.
- **Energy efficiency** is seen as the product of three factors we can **independently measure** and **optimize**:

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

PUE: Power Usage Effectiveness

Efficiency factors

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

A. **Facility efficiency**

B. **Server power conversion efficiency**

C. Server's **architectural efficiency**

Power Usage Effectiveness (PUE)

- Reflects quality of **data center building infrastructure** itself
- Captures the ratio of **total building power** to **IT power** (aka **critical power**).

PUE = (Facility power) / (IT Equipment power)

- Easily measured by adding electrical meters to the lines powering the various parts of a data center, determining how much power is used by chillers and UPS.
- Average PUEs: **1.13** (WSC), **1.6–2.35** (traditional DC) (2016).
- Most common improvements implemented:
  - Cold and hot aisle containment
  - Increased cold aisle temperature


27.2 °C
28.9
17.6
FLIR


28.1 °C
27.8
18.0
FLIR


27.7 °C
28.9
17.6
FLIR


21.5 °C
28.9
17.6
FLIR

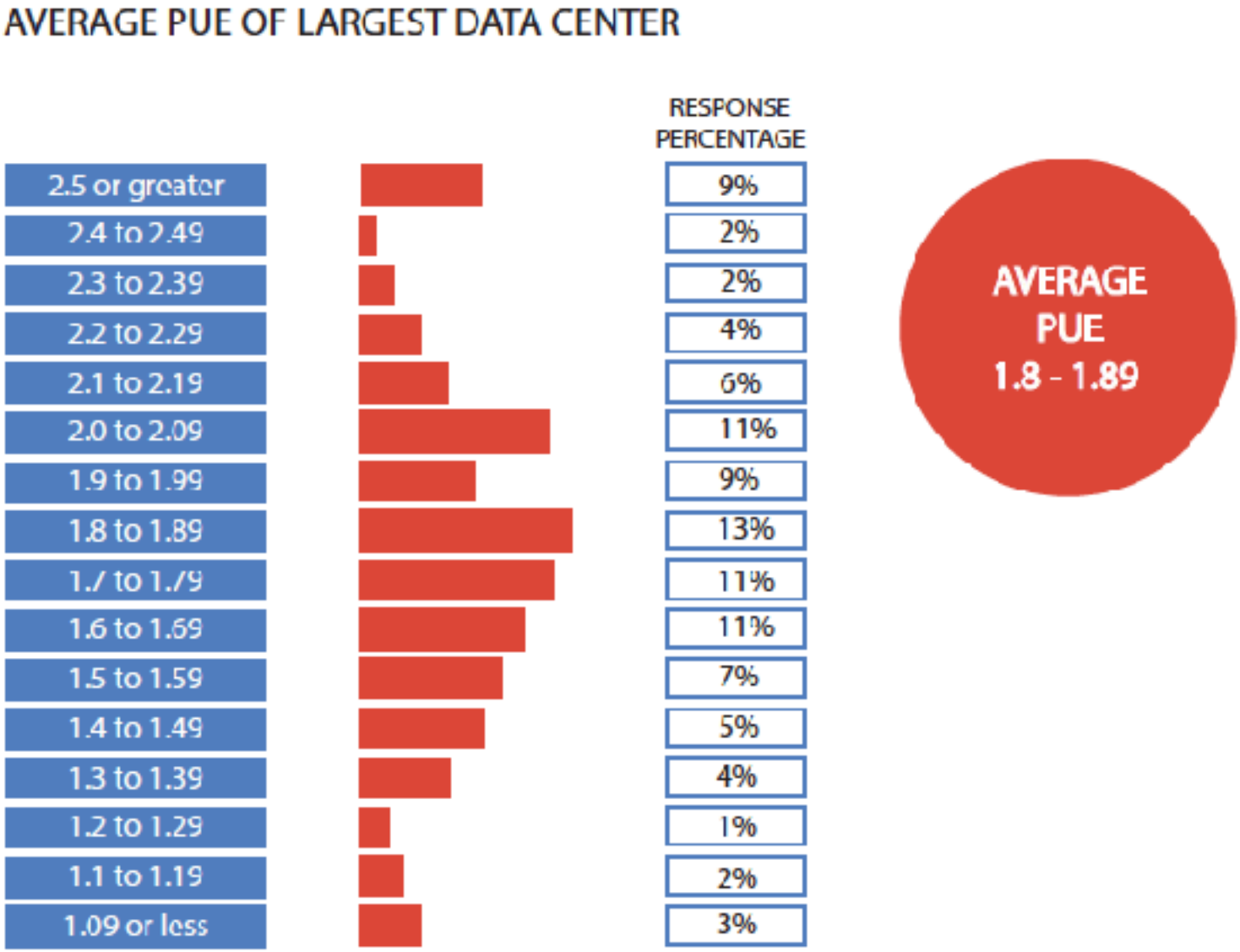


Figure 5.1: Uptime Institute survey of PUE for 1100+ data centers. This detailed data is based on a 2012 study [UpI12] but the trends are qualitatively similar to more recent studies (e.g., 2016 LBNL study [She+16]).

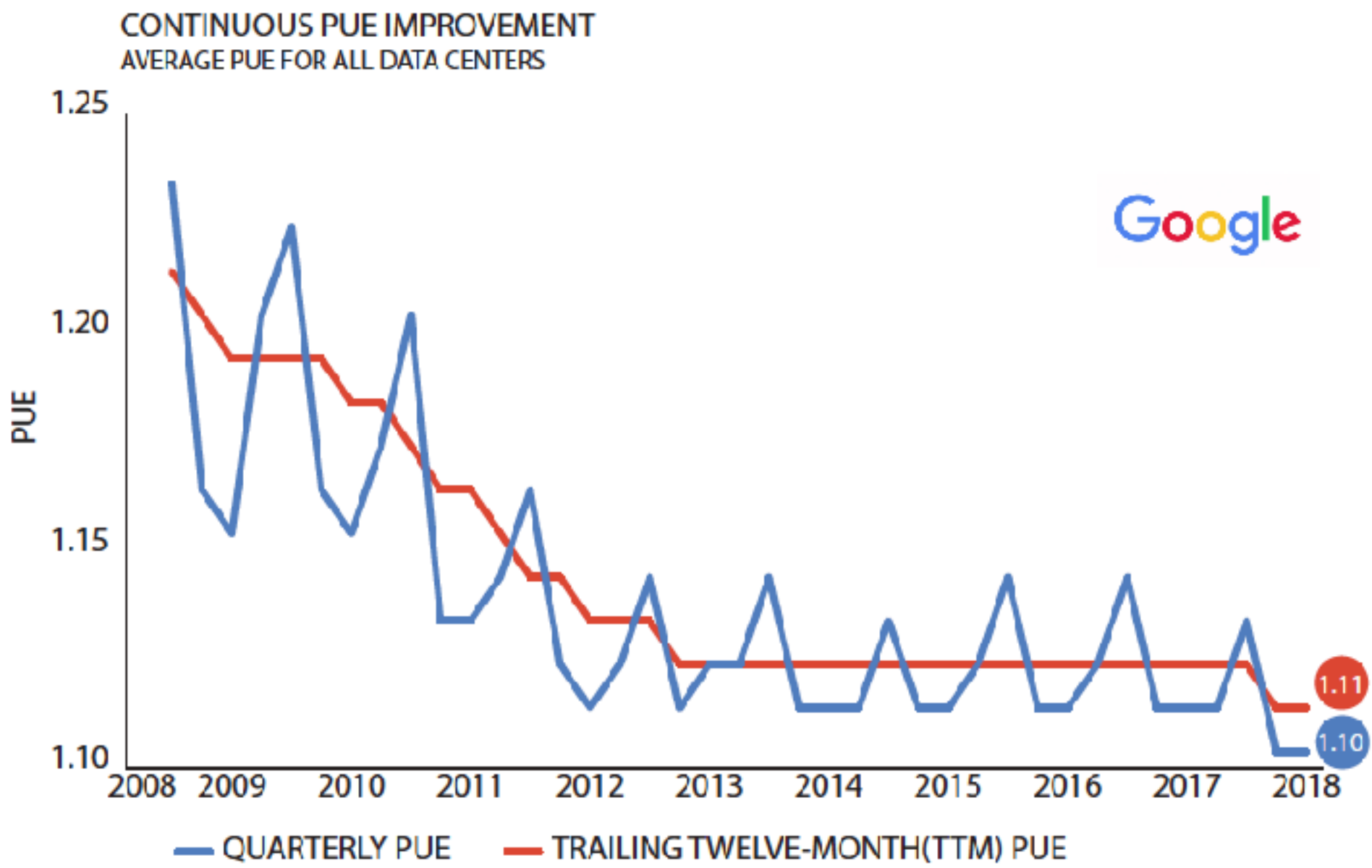


Figure 5.2: PUE data for all large-scale Google data centers over time [GDCa].

Issues with PUE

- Published PUEs in marketing documents show best-case values or values measured under optimal conditions that **aren't real.**
  - Typically, PUE values provided without details fall into this category.
- Different PUE measurements include **different overheads:** eg some account for losses in primary substation transformers, in wires feeding racks from PDUs, others don't
- Instantaneous PUEs differ from average PUEs. Over the course of a day or a year, a facility's PUE can **vary considerably**.
- Some PUE values have higher error bars because they're based on **infrequent manual readings**, or on **coarsely placed meters** that force some PUE terms to be **estimated** instead of measured.

Addressing PUE issues

- In practice, PUE values **should be measured in real time:**
  - better picture of diurnal and seasonal variations
  - allows the operator to react to unusual readings during day-to-day operations
- Data center owners and operators should adhere to **Green Grid guidelines** in measurements and reporting, and be transparent about the methods used in arriving at their results.

HOW CAN WE IMPROVE PUE?

University of Cyprus
Department of Computer Science

Sources of Efficiency loss in DC

- **Power transformation** steps (high- to medium-voltage and mid- to low voltage: losses typically below **0.5%** for each step.
- **UPS**:
  - Conventional double-conversion UPSs cause the most electrical loss - efficiencies of 88–94%.
  - Rotary UPSs (flywheels) and high-efficiency UPSs can reach efficiencies of about 97%.
- Final transformation in the **PDUs**: additional **0.5% loss.**
- **Cables feeding low-voltage power** (110 or 220 V) to the racks can be quite long: **1–3%** loss
- **Cooling overhead: cooling losses are three times greater than power losses**, presenting the most promising target for efficiency improvements:
  - If all cooling losses were eliminated, PUE would drop to 1.26, whereas a zero-loss UPS system would yield a PUE of only 1.8.
  - Typically, the worse a facility's PUE is, the higher the % of the total loss comes from the cooling system

Energy Loss Breakdown in typical DC

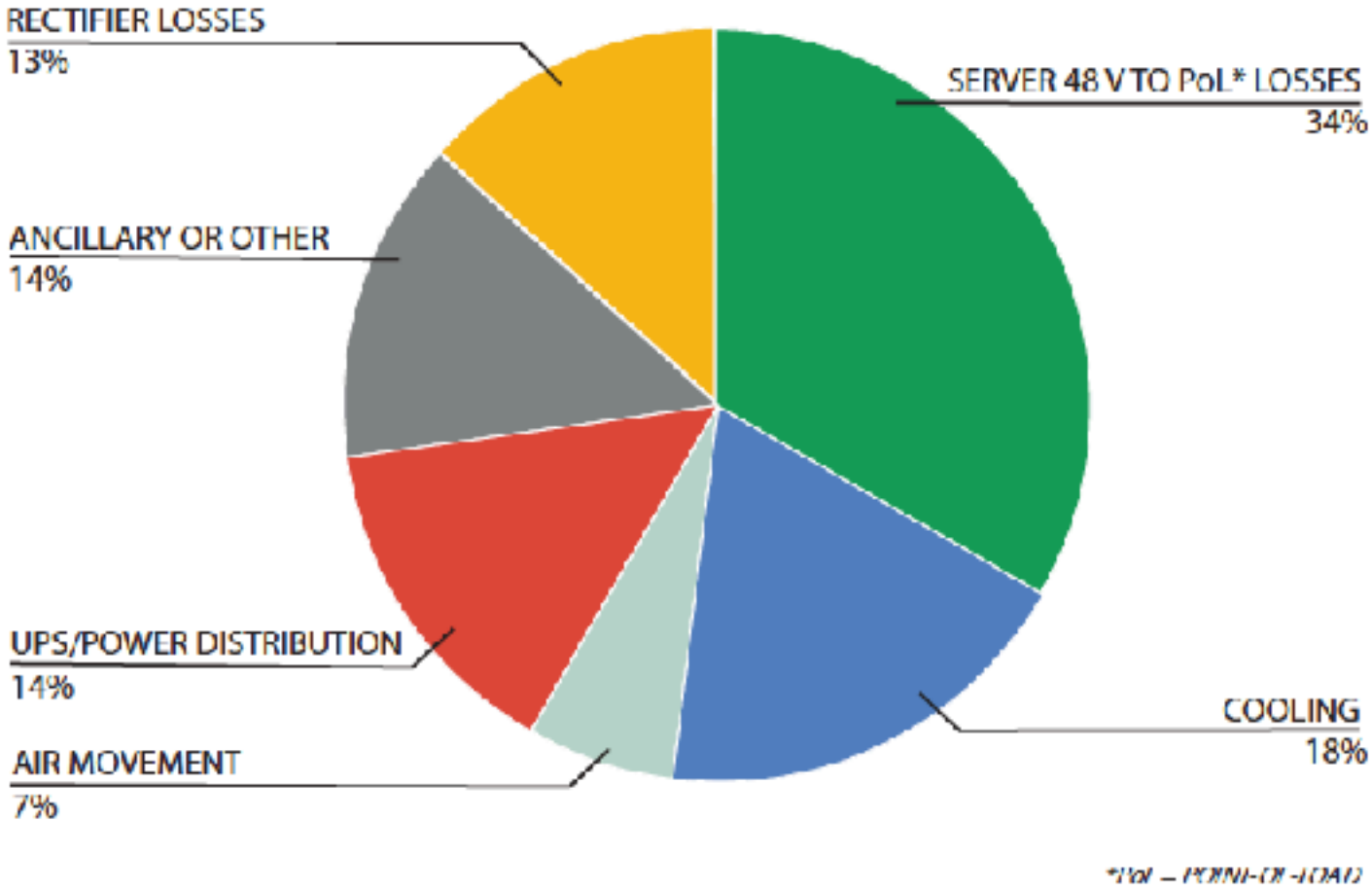


Figure 5.3: A representative end to end breakdown of energy losses in a typical datacenter. Note that this breakdown does not include losses of up to a few percent due to server fans or electrical resistance on server boards.

Improving DC Energy Efficiency

- **Careful air flow handling**: **Isolate** hot air exhausted by servers from cold air, and keep the **path to the cooling coil short** so that little energy is spent moving cold or hot air long distances.
- **Elevated temperatures**: Keep the cold aisle at **25–30°C** rather than **18–20°C**. Higher temperatures make it much easier to cool data centers efficiently.
- **Free cooling**: In most moderate climates, free cooling can eliminate the majority of chiller runtime or eliminate chillers altogether.
- **Better power system architecture**: UPS and power distribution losses can often be greatly reduced by selecting higher-efficiency gear.
- **Machine learning**: Apply novel machine learning techniques to discover non-intuitive techniques for controlling data center infrastructure to further reduce cooling requirements.

Why ML?

- The energy for running the cooling infrastructure has a **nonlinear relationship** with many system parameters and environmental factors, such as:
  - the total system load
  - the total number of chillers operating, and
  - the outside wind speed.
- Difficult to intuit the relationship between these variables and total cooling power.
- Large amount of data is being collected regularly from a network of sensors used to operate the control loop for data center cooling:
  - suggests that **machine learning** and **artificial intelligence** could be used to find additional PUE efficiencies

Energy and Power Efficiency

Power Efficiency beyond the Facility

University of Cyprus
Department of Computer Science

Server Power Usage Effectiveness

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

- **Server PUE** accounts for overheads **inside servers** or **other IT equipment**

SPUE = total server input power / useful power

Useful power: *consumed by the electronic components directly involved in the computation: motherboard, disks, CPUs, DRAM, I/O cards, etc.*

Overheads: *in server power supply, voltage regulator modules (VRMs), cooling fans*

- State-of-the-art SPUE is **1.11** or **less.**

True or Total PUE: TPUE = PUE x SPUE

Accurate assessment of the end-to-end electromechanical efficiency of a WSC.

- A decade ago **TPUE > 3.2** for the average data center: **For every productive watt**, at least another **2.2 W** were consumed.
- Modern facility with an average **PUE of 1.11** as well as an average **SPUE of 1.11** achieves a **TPUE of 1.23** (one order of magnitude reduction in overhead)

SUPPOSE THE TRUE PUE OF A DATA CENTER IS 1.23.

BY HOW MUCH IS THE TOTAL ENERGY EFFICIENCY GOING TO BE IMPROVED IF YOU ELIMINATE ALL ELECTROMECHANICAL OVERHEADS?

University of Cyprus
Department of Computer Science

Energy Efficiency of Computing

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

Energy Efficiency of Computing

- **What is it?**
  - How much of the electricity delivered to electronic components is actually **translated into useful work**.
- **Why does it matter?**
  - In a state-of-the-art facility, electromechanical components have a **limited potential for improvement**
  - Energy efficiency of computing has **doubled approximately every 1.5 years** in the last half century.
  - These rates have declined due to CMOS scaling challenges, but are still able to outpace any electromechanical efficiency improvements.

Energy Efficiency of Computing

- **How do we measure it?**

Energy consumed to produce a **certain result**

- **How do we compare alternatives fairly?**
- Benchmarks:
  - **Green 500**: ranks the energy efficiency of the world's top supercomputers using LINPACK.
  - **Joulesort**: server-level benchmark measures the total system energy to perform an out-of-core sort and derives a metric that enables the comparison of systems ranging from embedded devices to supercomputers.
  - **SPECpower**: focuses on server-class systems and computes the performance-to-power ratio of a system running a typical business application on an enterprise Java platform.
  - **Emerald** and **SPC-2/E** measure storage servers under different kinds of request activity and report ratios of transaction throughput per watt.

Server Energy Efficiency

- Factors affecting server energy efficiency:
  - **Server architecture**: same application can consume different amounts of power on different architectures.
  - **Software performance tuning:** An application can consume more or less of a server's capacity depending on its tuning.
  - Utilization: under **low levels of utilization**, computing systems tend to be significantly **more inefficient** than when they are exercised at maximum utilization.

With decreasing load, **system power** decreases much **more slowly** than does **performance** (transactions/sec)

Energy efficiency at 30% load is 30% lower than at 100% load.

When system is **idle**, it is still consuming just under 60 W, which is **16% of the peak power consumption** of the server.

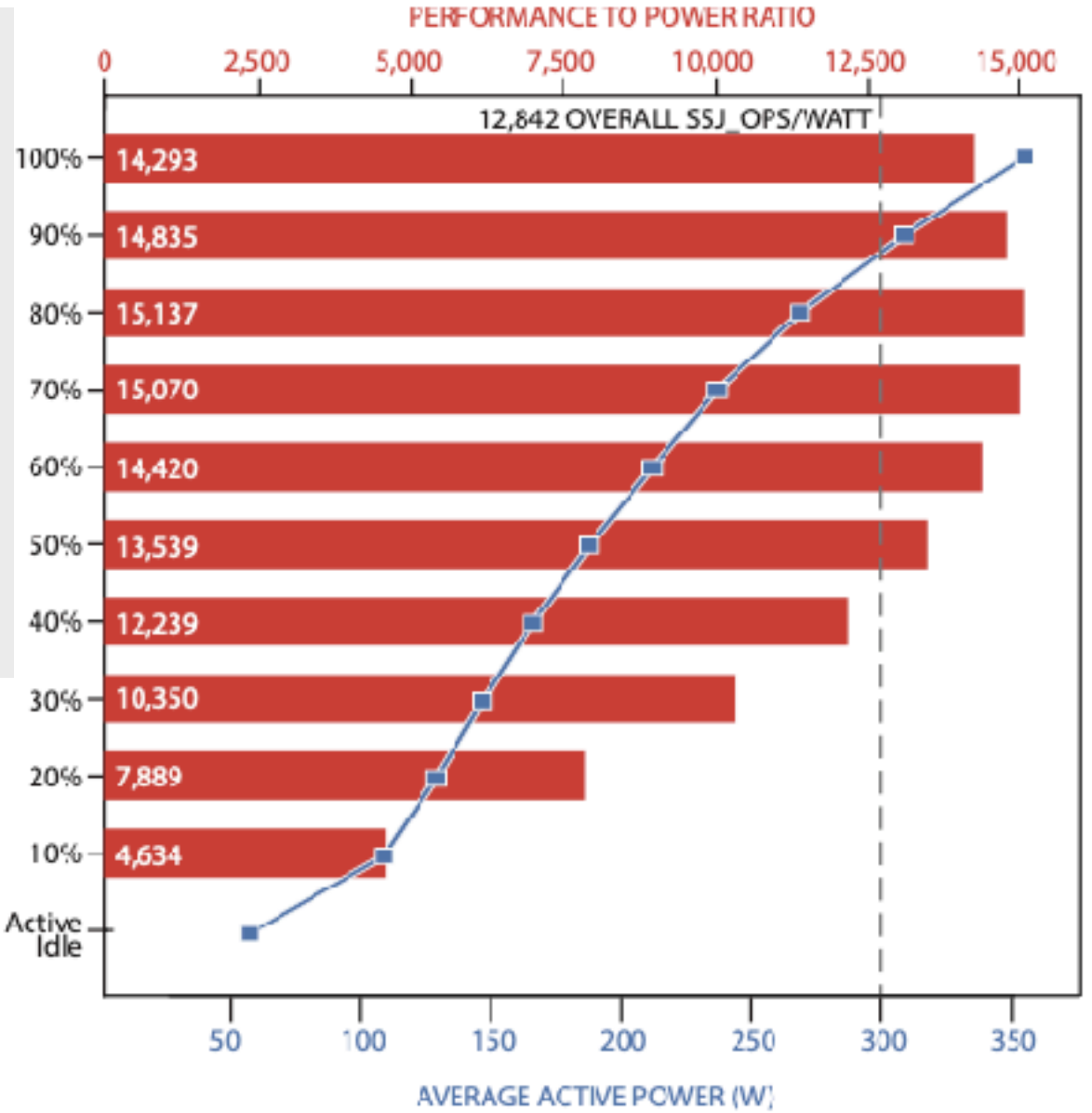


Figure 5.4: Example benchmark result for SPECpower_ssj2008. The bars indicate energy efficiency and the line indicates power consumption. Both are plotted for a range of utilization levels, with the average energy efficiency metric corresponding to the vertical dark line. The system has two 2.1 GHz 28 core Intel Xeon processors, 192 GB of DRAM, and one M.2 SATA SSD.

Efficiency in WSC: Challenges

- Typical shared WSC: relatively low average utilization.

Utilization

Efficiency in WSC: Challenges

- Typical shared WSC: relatively low average utilization.
- Typically servers spend most of their time in inefficient load region: mismatch with energy efficiency profile

Efficiency in WSC: Challenges

- Another WSC feature: Individual servers spend **little time idle**
- **Why?**
  - **Design principles** for **robust** distributed systems software: at lighter load periods, we tend to have a **lower load in multiple servers** instead of **concentrating the load in fewer servers** and **idling the remaining ones**
  - **Resilient distributed storage systems** (GFS) distribute data chunk replicas for a given file **across an entire cluster**. Thus, low traffic levels translate into lower activity for all machines instead of full idleness for a significant subset of them.
  - **Practical considerations**: Networked servers frequently perform many **small background tasks** on periodic intervals.
- Idleness can be **manufactured** by **migrating** workloads and their corresponding state to fewer machines during periods of low activity
  - This is doable when servers/apps are mostly **stateless**
  - But.. this is **more expensive** for more complex data distribution models or those with significant state and aggressive exploitation of data locality

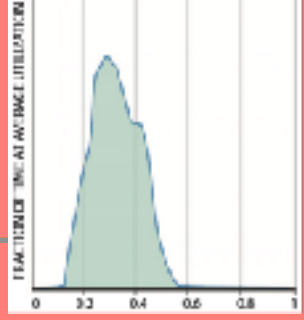
Energy and Power Efficiency

Energy-proportional computing

Energy proportionality

- **What is it?**
  - A desired design goal for computing components.
  - Energy-proportional systems **consume almost no power when idle** (particularly in active idle states) and **gradually consume more power as the activity level increases**.
- **How do we represent it?**
  - The energy proportionality of a server for a WSC can be represented as the **ratio between the energy efficiency at 30% and 100% utilizations**.
  - A perfectly proportional system will be as efficient at 30% as it is at 100%.
- **Why is it important?**
  - **Linearity** between activity and power usage would make energy efficiency uniform across the activity range.

Energy proportionality



How to translate it to energy efficiency?

- Linearity not necessarily the **optimal relationship** for energy savings (why?)
  - Servers spend relatively little time at high activity levels:
    - might be fine to decrease efficiency at high utilizations, particularly when approaching maximum utilization
- However, doing so would increase the maximum power draw of the equipment, thus increasing facility costs.

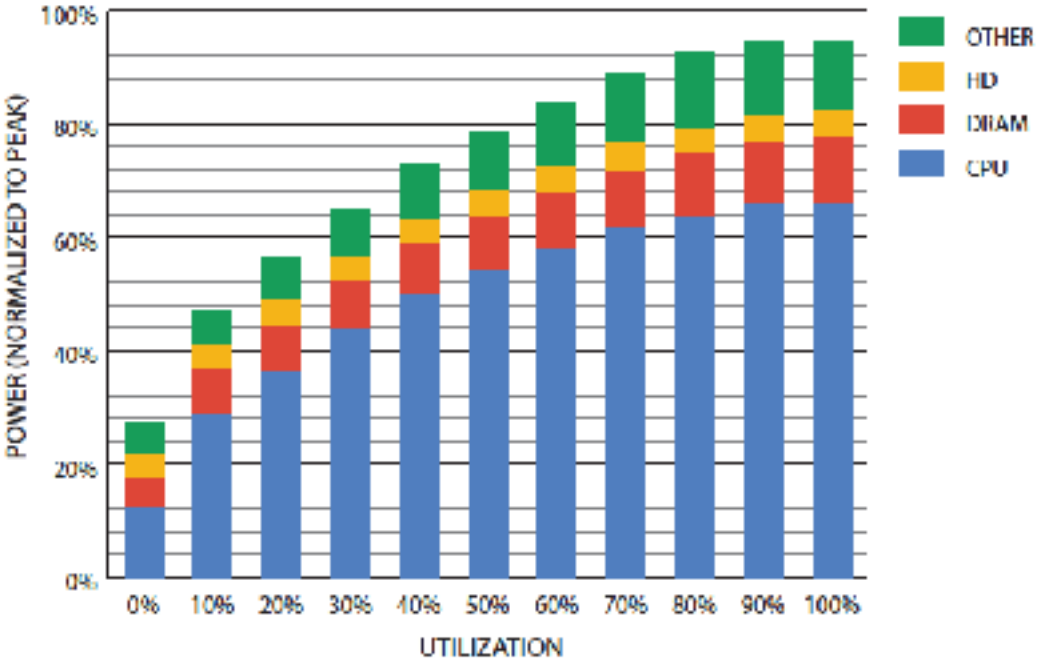
Energy proportionality gains

- Fan, Weber, Barroso (ISCA2007) study using traces of activity levels of thousands of machines over six months.
- Simulate energy savings gained from using more energy-proportional servers:
  - servers with **idle consumption at 10% of peak** instead of at **50%**
  - Baseline and energy-proportional servers have the same peak energy efficiency
- Models suggest that energy usage would be **halved** through increased energy proportionality alone .

Causes of poor proportionality

CPU is **dominant** energy consumer in servers: uses **2/3 of the energy at peak utilization** and about **40% when (active) idle**:

- Server-class CPUs have a dynamic power range that is generally greater than **3.0x** (3.5x in this example).

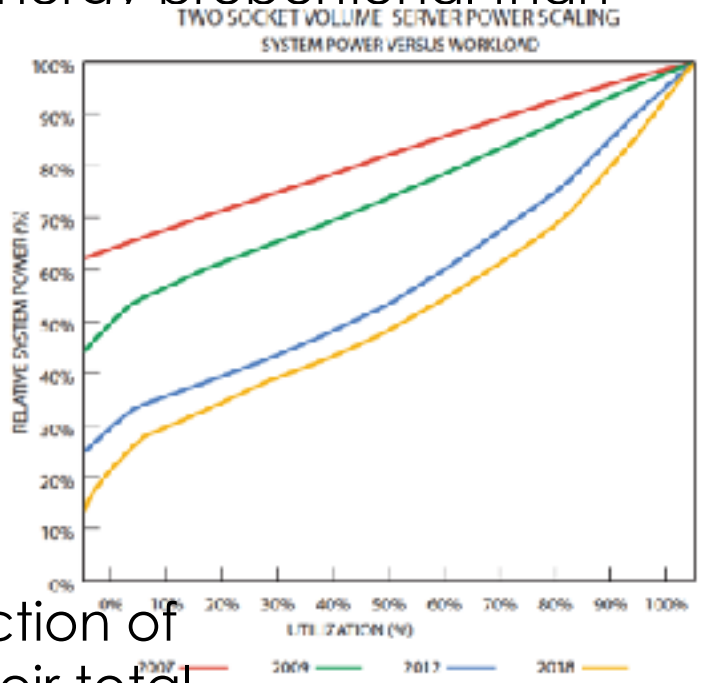


- The dynamic range for **memory** is approximately **2.0x**, for **disks** **1.3x**, and for **networking** switches less than **1.2x.**
- Energy proportionality at the system level cannot be achieved through CPU optimizations **alone**, but requires improvements across all components.

Figure 5.7: Subsystem power usage in an x86 server as the compute load varies from idle to full usage.

Energy proportionality improvement

- Processor energy proportionality has improved recently, with more recent systems being dramatically more energy proportional than their predecessor



- Greater effort is still required for DRAM, storage, and networking.

- Disk drives, for example, spend a large fraction of their energy budget (as much as 70% of their total power for high RPM drives) simply keeping the platters spinning

Low-power modes

- Long idleness intervals would make it possible to achieve higher energy proportionality by using various kinds of **low-power modes:**
- **Inactive** (sleep) modes: devices not usable while in those modes
  - Most of those techniques are a poor fit for WSC systems
    - Typically, a **sizeable latency** and **energy penalty** is incurred when **load** is **reapplied**.
    - In WSC, the inactive-to-active penalties would be paid too **frequently**.
    - Successful techniques have **very low wake-up latencies,** but the savings are not big, as these occur in low-power modes with smallest degrees of energy savings.
    - Higher savings could be achieved by restricting spin-down modes of HDDs
- **Active** modes: save energy at a performance cost while not requiring inactivity.
  - **CPU voltage-frequency scaling**: CPU remains able to execute instructions albeit at a slower rate.
  - **Useful even when** the latency and transition energy **penalties are significant**, because systems can remain active in low-energy states for as long as they remain below certain load thresholds.

Software in Energy Proportionality

- Intelligent power management and scheduling software infrastructure
- Software strategies for intelligent use of power management features:
  - Use low-overhead, low-power modes (inactive or active) in existing hardware
  - Implement power-friendly scheduling of tasks.
- Key software challenges:
- **Encapsulation:** Energy-aware mechanisms must be encapsulated in lower-level modules to minimize exposing additional infrastructure complexity to application developers.
- **Performance robustness:** Individual servers should not exhibit excessive response time variability as a result of mechanisms for power management.

Addressing Performance Variability

- Incorporating **end-to-end metrics** and **service level objective (SLO)** targets from WSC applications into power-saving decisions can greatly help overcome performance variability while moving toward energy proportionality.
  - During periods of low utilization, latency slack exists between the (higher latency) SLO targets and the currently achieved latency.
  - This slack represents power saving opportunities, as the application is running faster than needed.
- Having **end-to-end performance metrics** is critical to safely reduce the performance of the WSC in response to lower loads.
  - Combining **application-level metrics** with **fine-grained hardware actuation mechanisms**, the system is able to make overall server power more energy proportional while respecting the latency SLOs of the WSC application.

Cluster-level power efficiency

- Despite poor servers' energy proportionality cluster mgt software could:
  - **Increase utilization** of each individual server, and avoid operating servers in the region of poor energy efficiency at low loads.
  - Cluster **scheduling** software (**Borg**, **Mesos**) improve machine-level utilization through better bin-packing of disparate jobs (encapsulation), pushing servers to operate closer to their most energy efficient operating point (higher utilisation).
  - Even larger benefit of higher utilization: **reduced number of servers needed** to serve a given capacity requirement => dramatically lower TCO.

Contention-aware scheduling

- **What is it?**
  - Use **performance metrics** in scheduling and resource allocation decisions
- **Why needed?**
  - As server utilization is pushed higher, shared resource **contention** leads to **performance degradation** when using workload agnostic scheduling and with server capacity increasing due to the scaling of CPU core counts.
  - To counter interference effects, service owners tend to try gaming the system by **increasing the resource requirements of sensitive workloads** in order to ensure that their jobs will have sufficient compute capacity in the face of resource contention. This lowers server utilization and negatively impacts energy efficiency.
- **What does it achieve?**
  - Significantly **higher server utilizations** while **maintaining strict application-level performance requirements** (Bubble-Up, Heracles, Quasar.)
  - More resource sharing opportunities, increased machine utilization, and ultimately energy efficient WSCs that can sidestep poor energy proportionality.

Energy Efficiency & Specialization

- Specialized accelerators need to perform well only for a specific kind of computation: opportunity for domain-specific optimizations:
  - Relatively simple control logic for TPUs => much more energy efficient.
  - Parallelism in ML apps is easier to extract, the TPU has no need for the complicated and energy hungry control hardware found in CPUs.
- These and other design decisions for the TPU unlocked a **vast improvement in energy efficiency**.

Data Centers and Warehouse Scale Computers

Cost Modeling

Total Cost of Ownership (TCO)

- Costs split into capital expenses (Capex) and operational expenses (Opex).
- **Capex:** investments that must be made upfront and that are then depreciated over a certain timeframe (construction, purchase price of servers).
- **Opex:** recurring monthly costs of actually running the equipment, *excluding depreciation* (electricity costs, repairs and maintenance, salaries of on-site personnel, etc).

TCO = data center depreciation + data center Opex + server depreciation + server Opex

Data Center Construction Cost

- Data center construction costs vary widely.
- **How do we report construction cost?**
  - Dollars per square foot?
    - Metric does not correlate well with the primary cost driver of data center construction
  - **Dollars per watt of usable critical power**
    - All of the data center's primary components—power, cooling, and space—**roughly scale linearly with watts** (true for *larger data centers* where size-independent fixed costs are relatively small fraction of overall cost)
- **Why watts of "critical" power?**
  - Data center with 20 MW of generators may have been built in a 2N configuration and provide only **6 MW** of critical power (plus **4 MW** to power chillers).
  - Thus, if construction costs $120 million, it costs **$20/W**, not $6/W
- Typically, **~60–80%** of **total construction cost** goes toward **power and cooling**, and the remaining **20–40%** toward the general building and site construction.
- Historical costs of data center construction of Tier III facilities range from $9–$13 per watt.
- As DC construction projects increase, **costs drop** ranging from **$7–$9 per watt** (US, 2018)

Table 6.1: Range of data center construction costs expressed in U.S. dollars per watt of critical power. Critical power is defined as the peak power level that can be provisioned to IT equipment

Cost/W	Source
$12-25	Uptime Institute estimates for small- to medium-sized data centers; the lower value is for Tier I designs that are rarely built in practice [TS06].
$9-13	Dupont Fabros 2011 Form 10K report [DuP11] contains financial information suggesting the following cost for its most recent facilities (built in 2010 and 2011; see page 39 for critical load and page 76 for cost): $204M for 18.2 MW (NJ1 Phase I) => $11.23/W $116M for 13 MW (ACC6 Phase I) => $8.94/W $229M for 18.2 MW (SC1 Phase 1) => $12.56/W
$8-10	Microsoft's investment of $130M for 13.2 MW ($9.85/W) capacity expansion to its data center in Dublin, Ireland [Mic12]. Facebook is reported to have spent $210M for 28 MW ($7.50/W) at its Prineville data center [Mil12].

Building Depreciation Costs

- Monthly **depreciation** cost (or **amortization** cost) results from the initial construction expense
  - depends on the duration over which the investment is amortized (related to its expected lifetime) and the assumed interest rate:**15–20 years**.
- Under U.S. accounting rules, the value of the asset declines by a fixed amount each month. For example, if we depreciate a $12/W data center over 12 years, the depreciation cost is **$0.08/W per month**.
- If you take out a loan to finance construction at an interest rate of 8%, the associated monthly interest payments add an additional $0.13/W, for a total of $0.21/W per month.
  - Typical interest rates vary over time, but many companies use a cost of capital rate in the 7–12% range.

Calculation excluding financing	
Building cost ($)	218,000,000
Peak Consumption (W)	18,200,000
Depreciation (per W)	11.98
Number of months	144
Depreciation per month per Watt	0.08
Calculation including financing	
Building cost ($)	218,000,000
Financing at Cost (6000*(1.08^12)	548,961,085
Peak Consumption (W)	18,200,000
Depreciation (per W)	30.16
Number of months	144.00
Depreciation per month per Watt	0.21

Server Costs

- Server costs are computed similarly, except that servers have a shorter lifetime and thus are typically depreciated over **3–4 years**.
- To normalize server and data center costs, it is useful to characterize server costs per watt as well, using the server's peak real-life power consumption as the denominator.
- For example, a $4,000 server with an actual peak power consumption of 500 W costs $8/W.
- Depreciated over 4 years, the server costs $0.17/W per month.
- Financing that server at 8% annual interest adds another $0.06/W per month, for a total of $0.19/W per month.

Calculation excluding financing		
Server cost ($)	4000	
Peak Consumption (W)	500	
Depreciation (per W)	8	
Number of months	48	
Depreciation per month per Watt	0.1667	
Calculation including financing		
Server cost ($)	4000	
Financing at Cost (4000*(1.08^4)	5441.9558	
Peak Consumption (W)	500	
Depreciation (per W)	10.883912	
Number of months	48	
Depreciation per month per Watt	0.2267	

Operational Costs

- Harder to characterize because it depends heavily on operational standards as well as on the data center's size
  - larger data centers are cheaper because fixed costs are amortized better.
- Typical operational costs for multi-megawatt data centers in the U.S. range from **$0.02–$0.08/W per month, excluding the actual electricity costs**.
- Server maintenance costs vary greatly depending on server type and maintenance standards.
- In traditional IT environments, the bulk of the operational cost lies in applications; software licenses and the cost of system administrators, database administrators, network engineers, etc.
  - These vary greatly depending on the situation.

Cost of Public Clouds

- Typically apply **spot pricing** - **"pay-as-you-go."**
- Spot instances **fairly expensive**: at $0.76/hr, using one for three years at full price will cost $19,972, vs. roughly $8,000 for an owned server. (Note, however, that this does not burden the costs for other factors like utilization as discussed earlier.)
- If you need a server for an extended period, public cloud providers will lower the hourly price in exchange for a long-term commitment and an upfront fee.
- Public cloud providers compete with your in-house costs thanks to scale.

Sample Questions

- What is the typical requirement for availability of Warehouse Scale Computers running cloud services?
- What are the key factors that determine cost-efficiency of WSCs?
- Draw an architectural diagram with the main IT components of a WSC and give indicative values for the key performance characteristics of these components.
- Give the definition of the Thermal Design Power (TDP) of a CPU and the Form Factor of a server, and explain how they affect the running cost of a Cloud infrastructure.
- Name three types of accelerators found in modern data centers and explain why they have been incorporated in modern DCs.
- What is a TOR? Describe its key characteristics.
- What is the bisection bandwidth of a rack and why it is an important consideration?
- Give a definition of the oversubscription ratio of a TOR switch. Why is it an important metric?
- What percentage (approximately) of the power consumption of a cloud server is spent by the CPU?
- What is a NAS and what are its pros and cons for storing cloud data?
- How is fault-tolerance and high-availability achieved in a distributed file system?
- Draw an architectural diagram with the memory hierarchy of a WSC and give indicative values for the key performance characteristics of these components.
- Explain what is redundancy in WSC, where it is used and what is the aim of adopting redundancy measures?

Sample Questions

- Suppose you plan for the development of a Tier-3 data center with 20000 servers, adopting best practices in the design/implementation. Can you provide an estimate of how much power will be required for cooling?
- Give the equation that defines the energy efficiency of a data center and explain its components
- What is the PUE? For a PUE of 1.4, what is the percentage of power going to IT components of a data center?
- What is SPUE and which power losses contribute to its value?
- What is the "True PUE" and what does it measure?
- Suppose the True PUE of a data center is 1.23. By how much is the total energy efficiency going to be improved if you eliminate all electromechanical overheads?
- How can we calculate energy efficiency of computing by running a benchmark?
- Name and explain the three main factors affecting the energy efficiency of computing.