

Computer Architecture

计算机体系结构

Lecture 12. Data Center as a Computer

第十二讲、数据中心即计算机

Chao Li, PhD.

李超 博士

SJTU-SE346, Spring 2019

Review

- Link/Channel/Buffer/Switch
- Blocking network, direct network, centralized network
- Latency estimation
- Network switch and switching strategy
- Bus, crossbar, array, ring, mesh, torus, tree, omega, butterfly, hypercube...

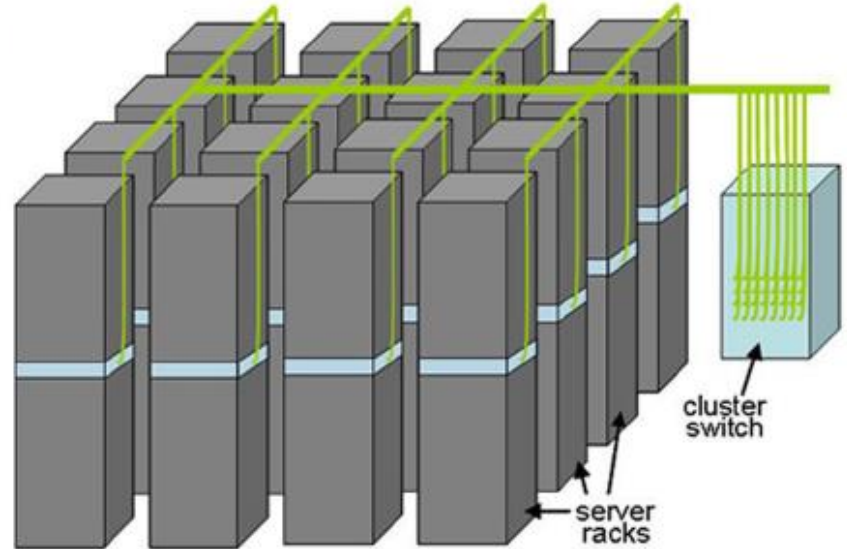
Outlines

- **Data Center Infrastructure**
- Key Design Considerations

What is a Data Center

- Key functions of a data center
 - Data processing (computation)
 - Concurrent processing of many requests (request-level parallelism)
 - Data storage
 - Large-scale, highly dependable
 - Data communication
 - High bisection bandwidth
- Key components of a data center
 - ICT equipment
 - Power system
 - Cooling system
 - Other supporting modules
 - Lighting, security, ...

Layers in a Data Center



- Basic unit: 1U
 - 1U = 4.445 cm
- Full size rack
 - 42U

- Layers in a WSC:
 - Server/Node Level; Rack/Cluster Level; Facility Level (Data Center Level)

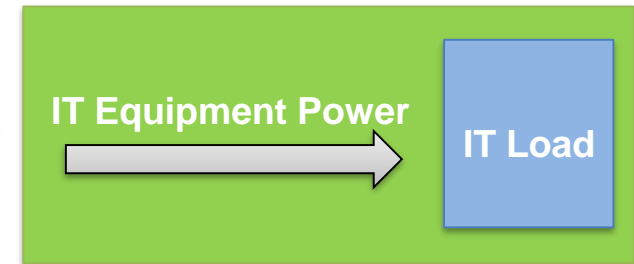

Data Center Operation Efficiency

- Power Usage effectiveness (PUE)
 - Provides a top-level view of how efficient a data center is operating



$$PUE = \frac{\text{Total Facility Power}}{\text{IT Equipment Power}}$$

Total Facility Power



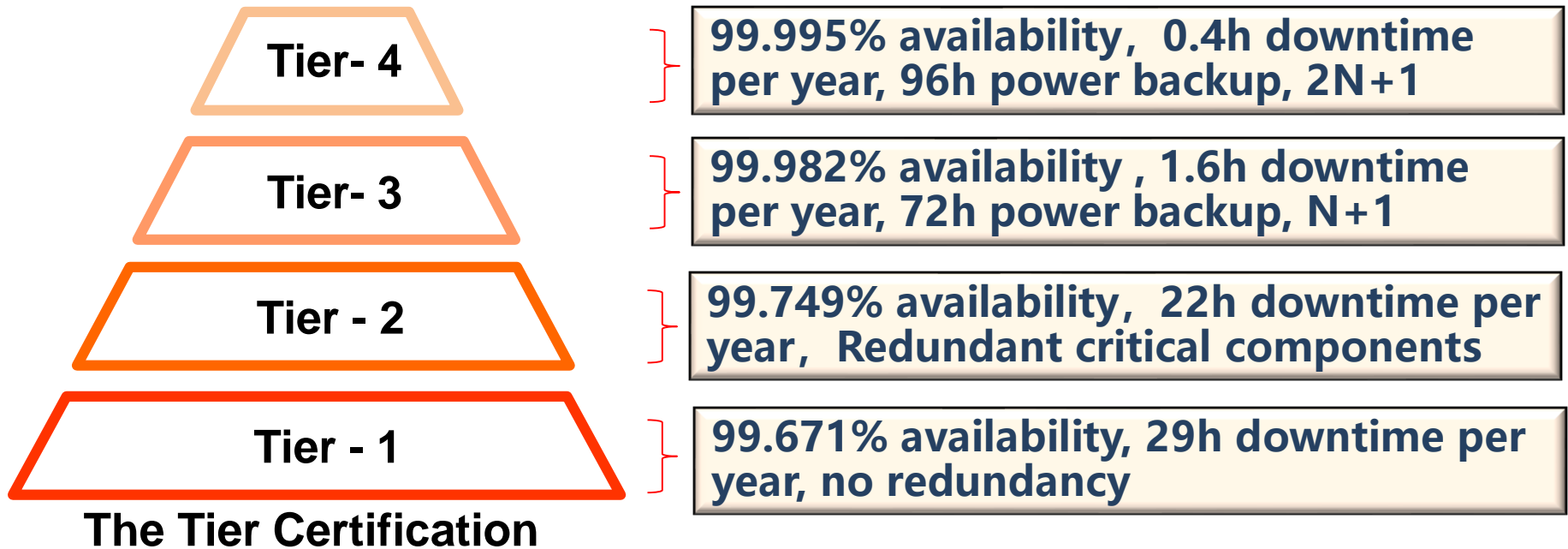
- More metrics from “The Green Grid” ...
 - WUE: water usage effectiveness (L/kWh).
 - CUE: carbon usage effectiveness (kgCO₂eq/kWh)

$$WUE = \frac{\text{Annual water usage}}{\text{IT Equipment Energy}}$$

$$CUE = \frac{\text{Total Carbon Emissions}}{\text{IT Equipment Energy}}$$

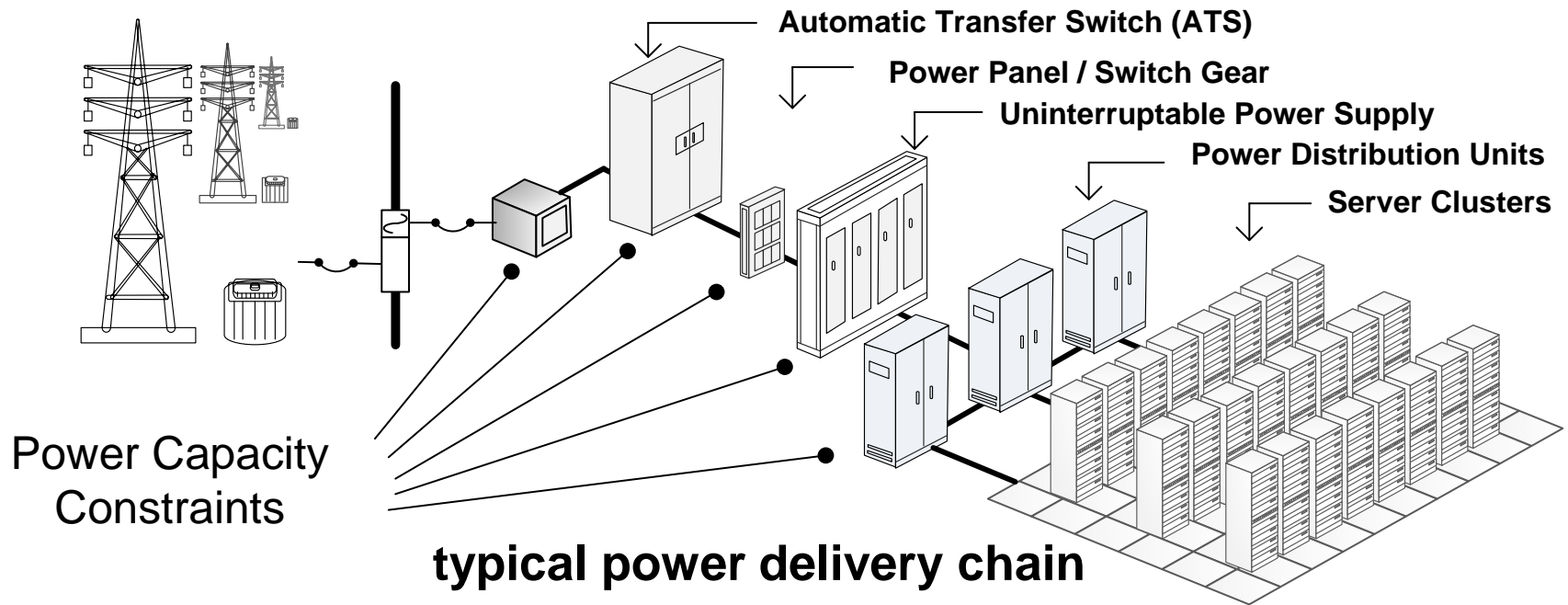
Data Center Tier Standard

Uptime's Tier Standard levels describe the availability of data processing from the hardware at a location



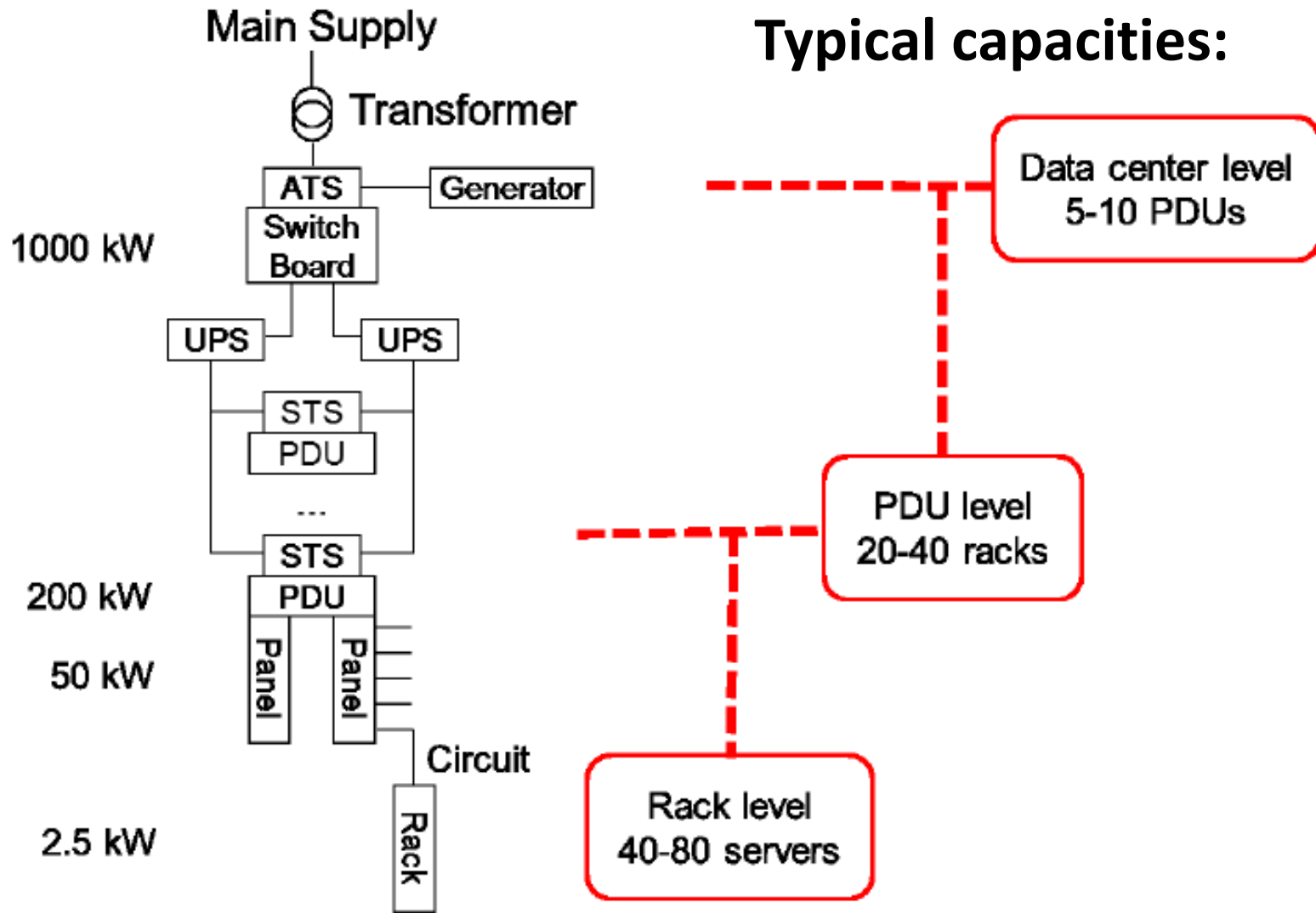
UptimeInstitute®

Datacenter Infrastructure – Power System

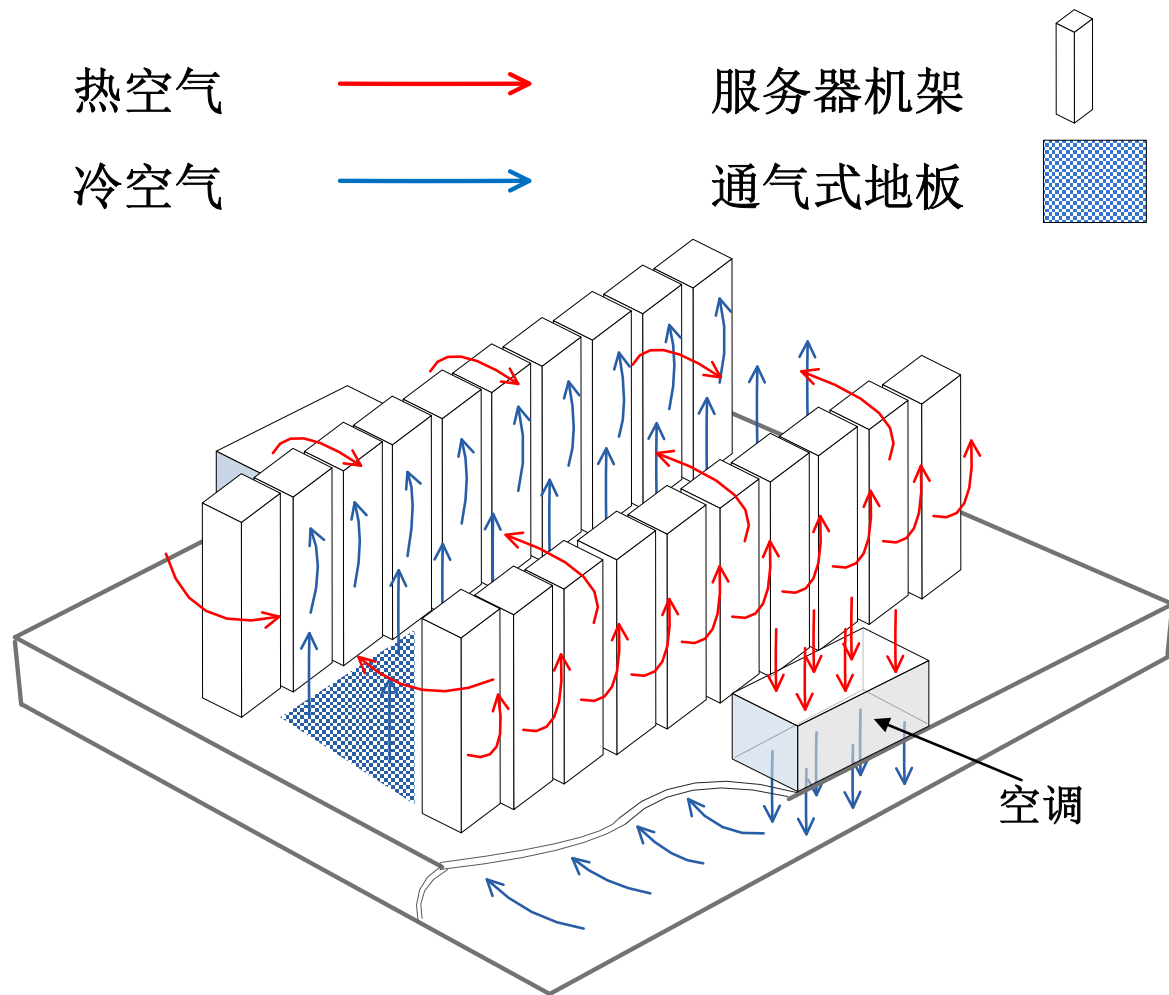
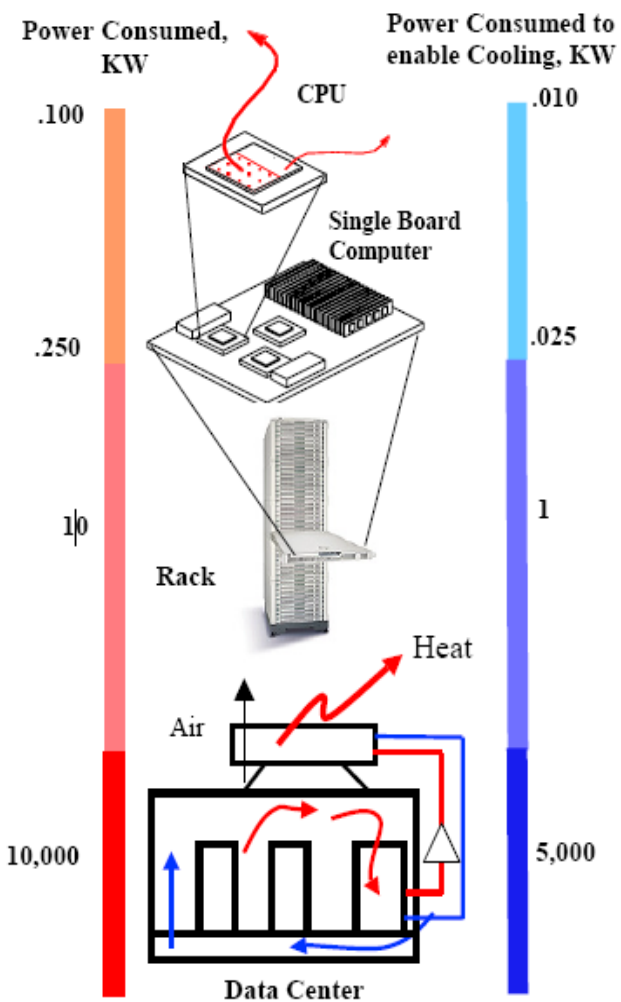


- ATS: basically a fast, mechanical switch
- STS: basically a superfast, electronic switch
- UPS: basically a battery with control interfaces
- PDU: a power converter and distribution unit

Datacenter Infrastructure – Power System



Datacenter Infrastructure – Cooling System



Datacenter Infrastructure – Cooling System

- **CRAC: Computer room air conditioning**
- **COP系数 (Coefficient of Performance)**
 - The ratio of the heat removed to input work



Guess the typical COP value in data centers?

- A. 0.5 – 1.0**
- B. 1.0 – 1.5**
- C. 5.0 – 10**
- D. 10 – 15**

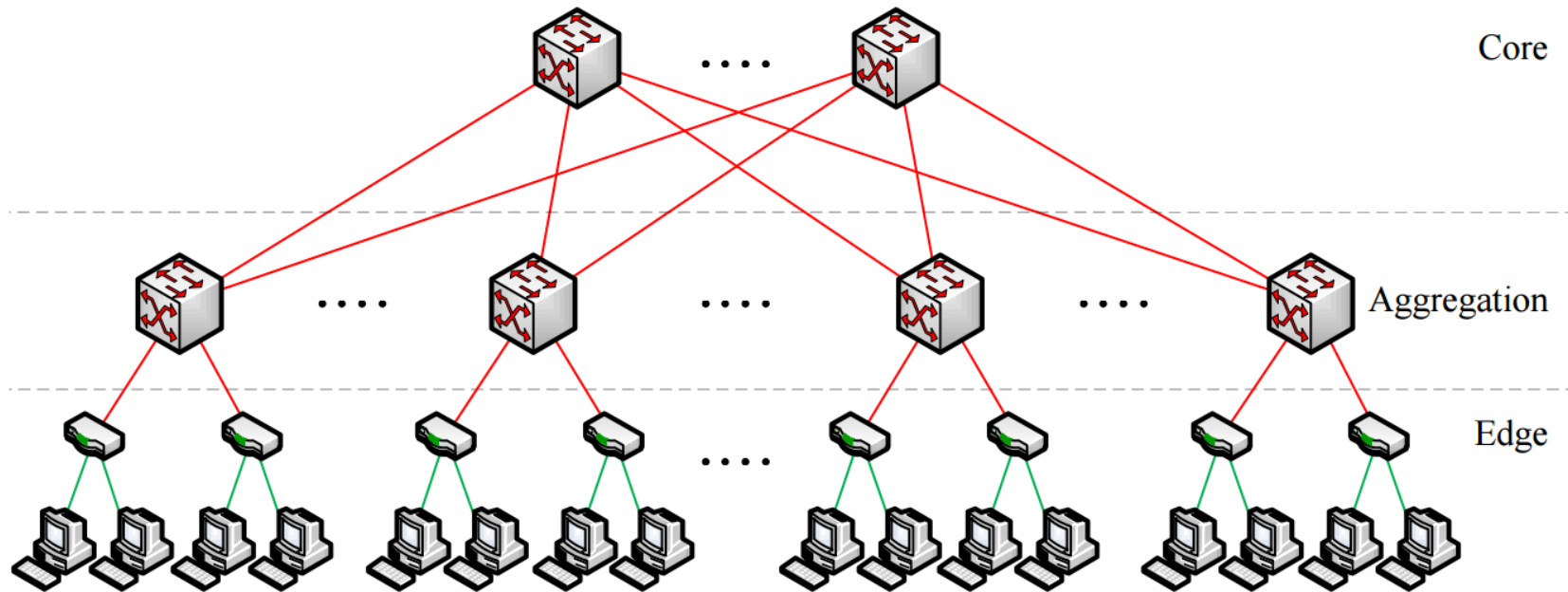


warehouse-scale
computer

cooling
towers

power substation

Datacenter Infrastructure – ICT System

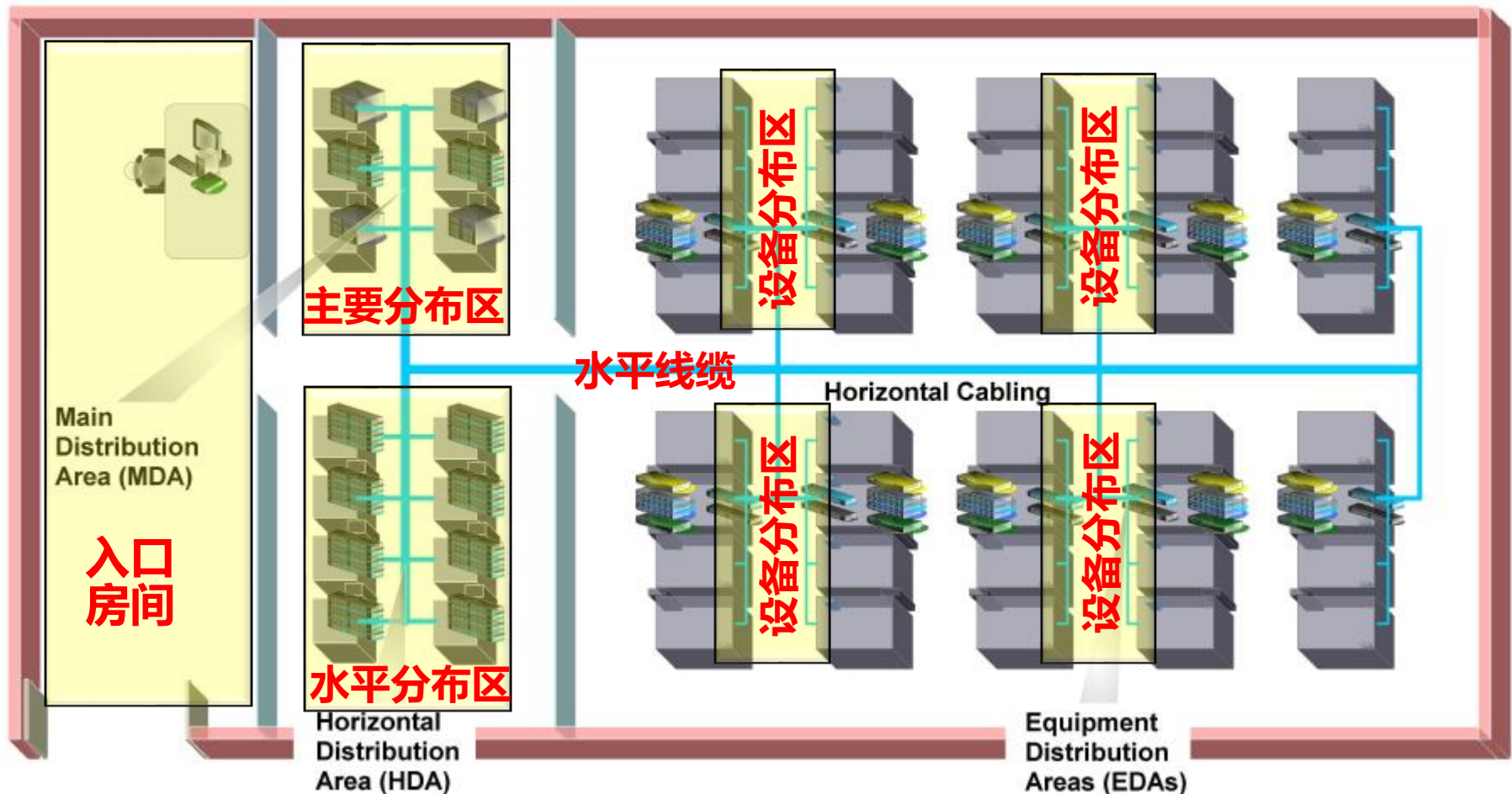


- Racks connected by switches and cables
- Types of switches
 - top-of-rack (TOR)
 - end-of-row (EOR)

	TOR1G	TOR10G	EOR
GbE Ports	48	0	0
10GbE Ports	4	24	128
Power (W)	200	200	11,500
Size (RU)	1	1	33

Ethernet switch models. Prices vary

Datacenter Infrastructure – ICT System



- Hierarchical network topology
 - Entrance facility + MDA + HAD + EDA

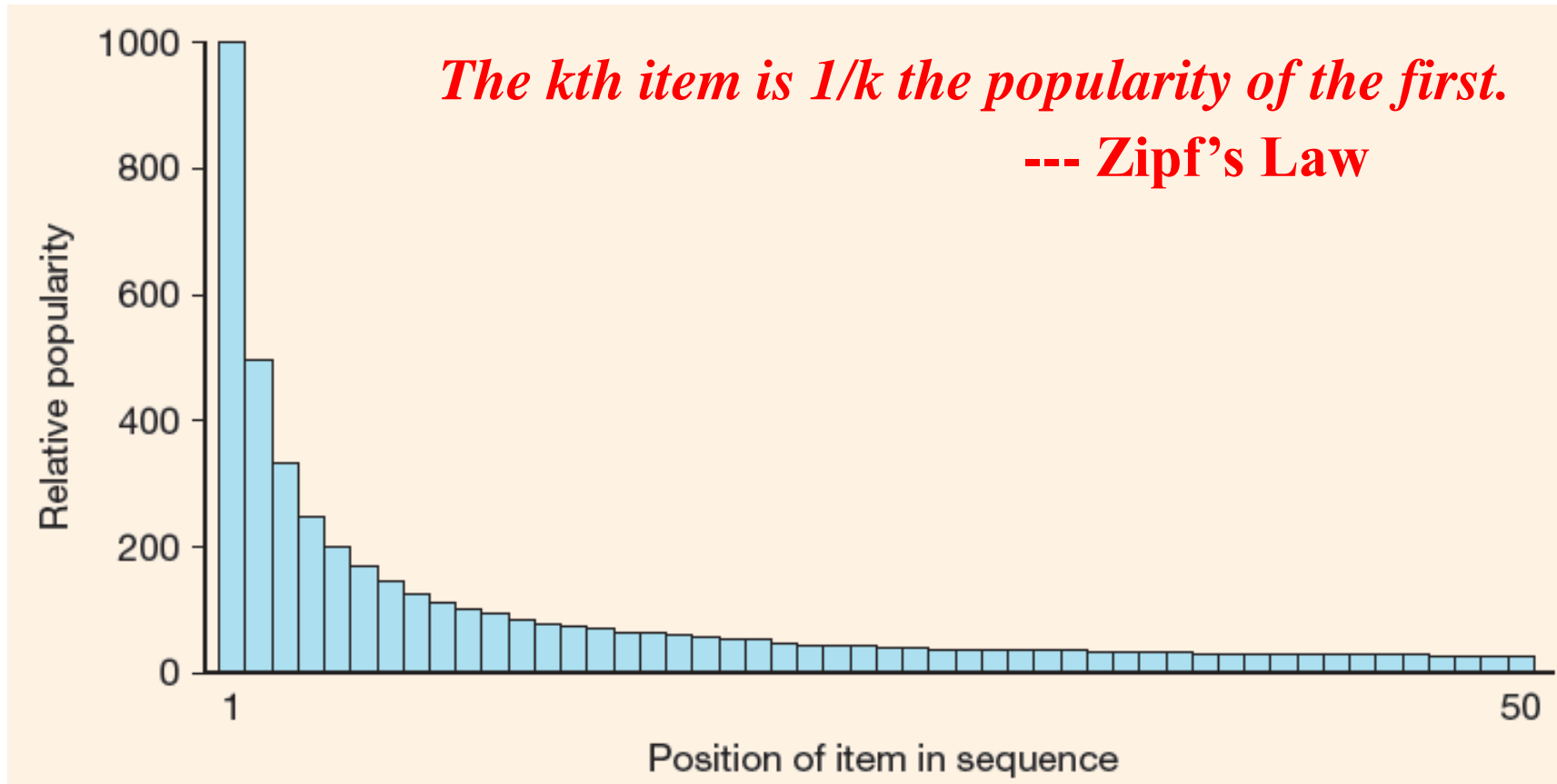
Outlines

- Data Center Infrastructure
- **Key Design Considerations**

Design Considerations of WSC

- The design and operation of WSC can be challenging
 - Performance
 - Energy Efficiency
 - Dependability/Availability
 - Networking
 - ...
- Service-Level Agreement
 - How the service should be provided
 - Agreement between the service provider and user
- Service-Level Objectives
 - a key element of a service level agreement (SLA)
 - measuring the performance of the service provide

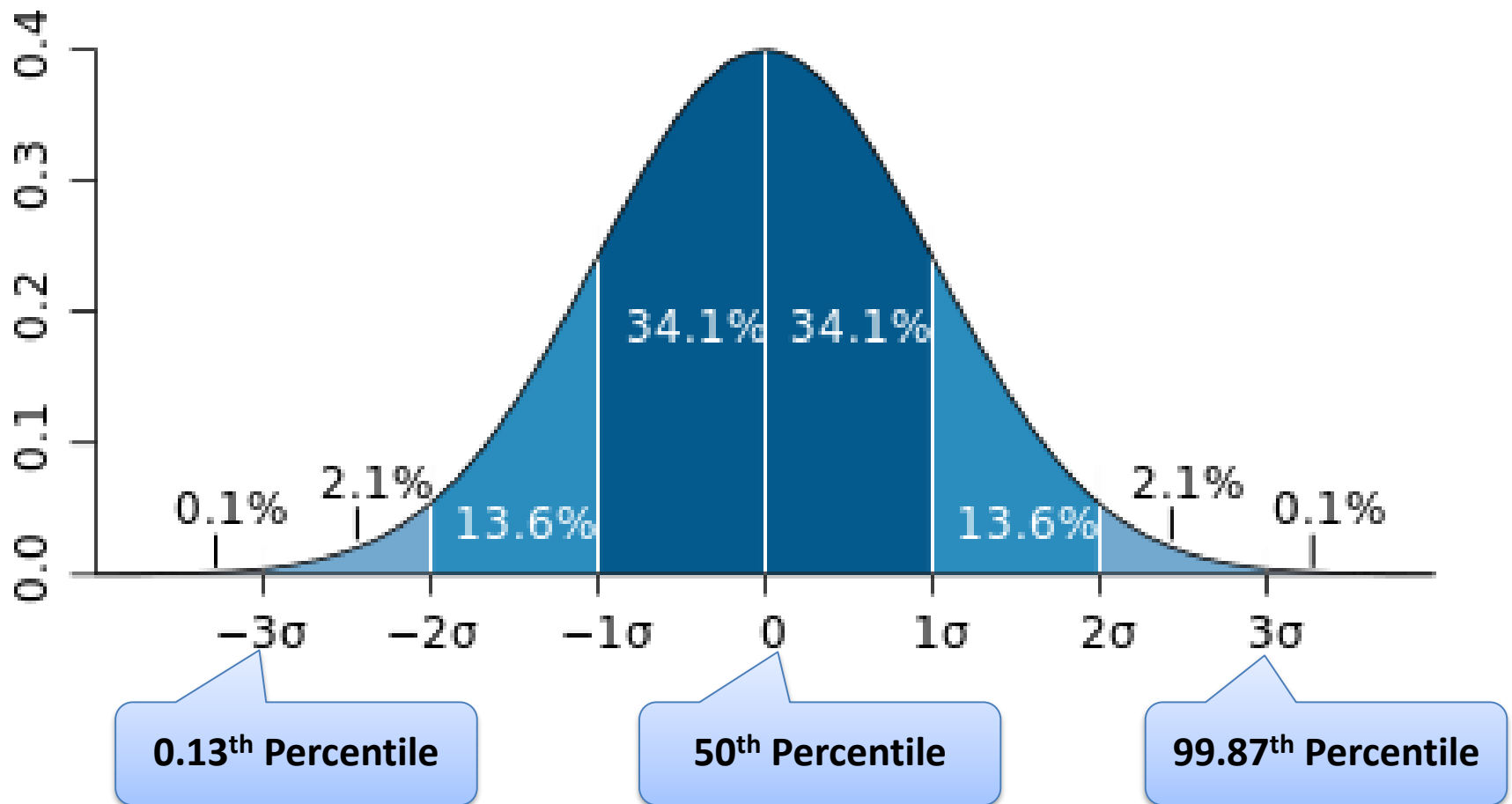
Discussion: The Tail at Scale



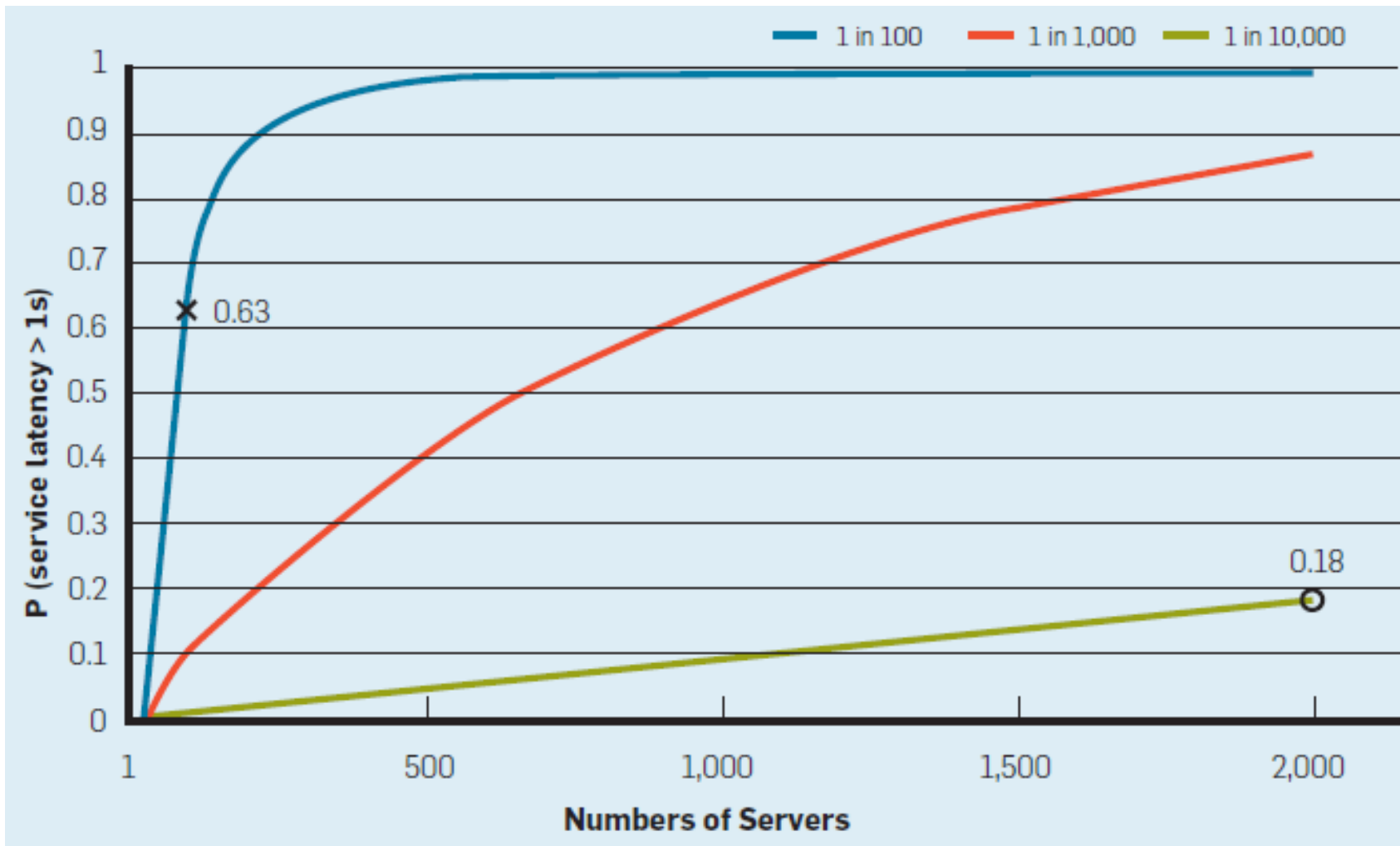
Zipf's law, showing decrease in popularity of items within an ordered sequence

正态分布中的百分位数举例

- The n -th Percentile of a set of data
 - The value at which $n\%$ of the data is below it.



Discussion: The Tail at Scale



Probability of one-second service-level response time as the system scales and frequency of server-level high-latency outliers varies.

Discussion: The Tail at Scale

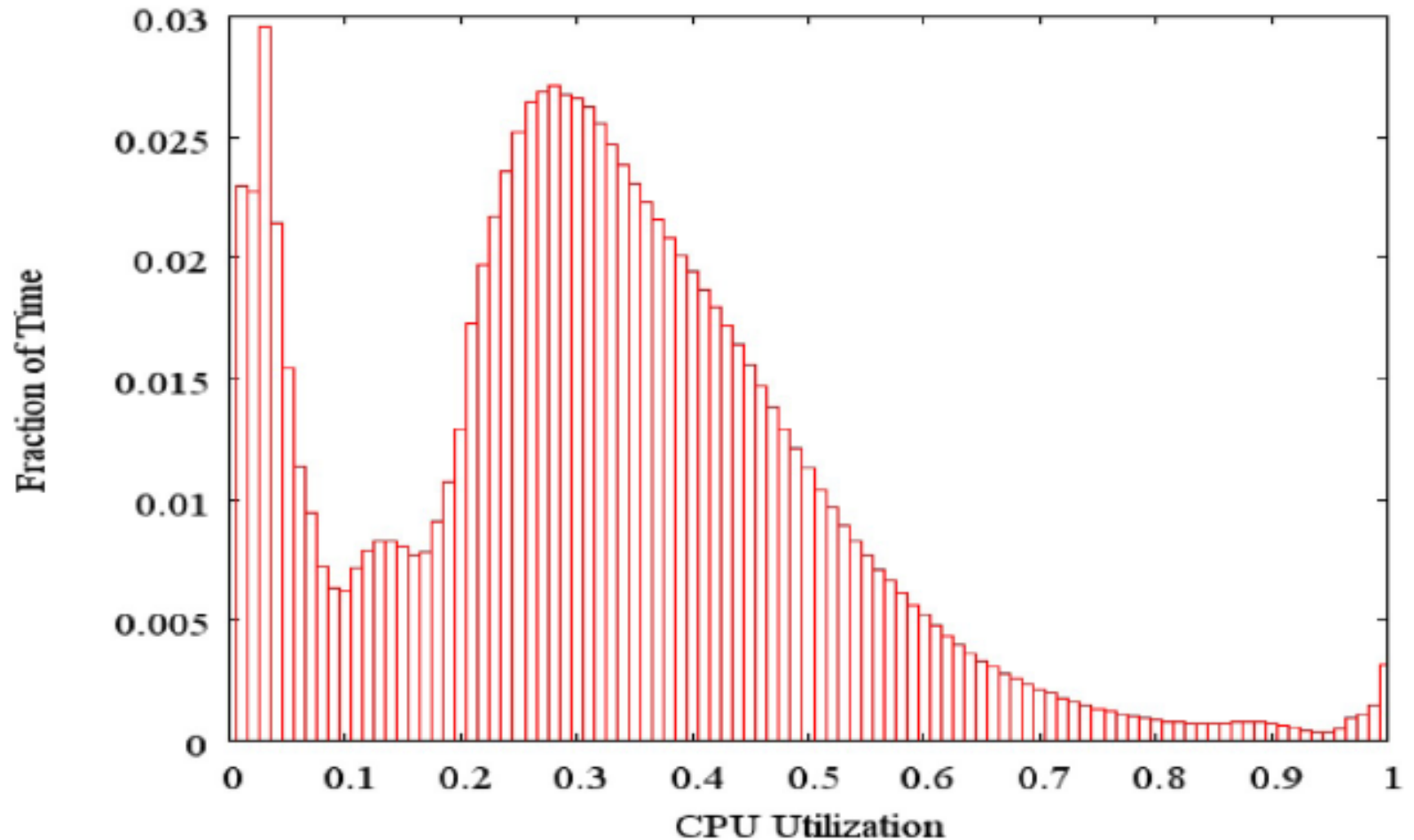
- A simple way to curb latency variability is to issue the same request to multiple replicas and use the results from whichever replica responds first.

Software techniques that tolerate latency variability are vital to building responsive large-scale Web services.

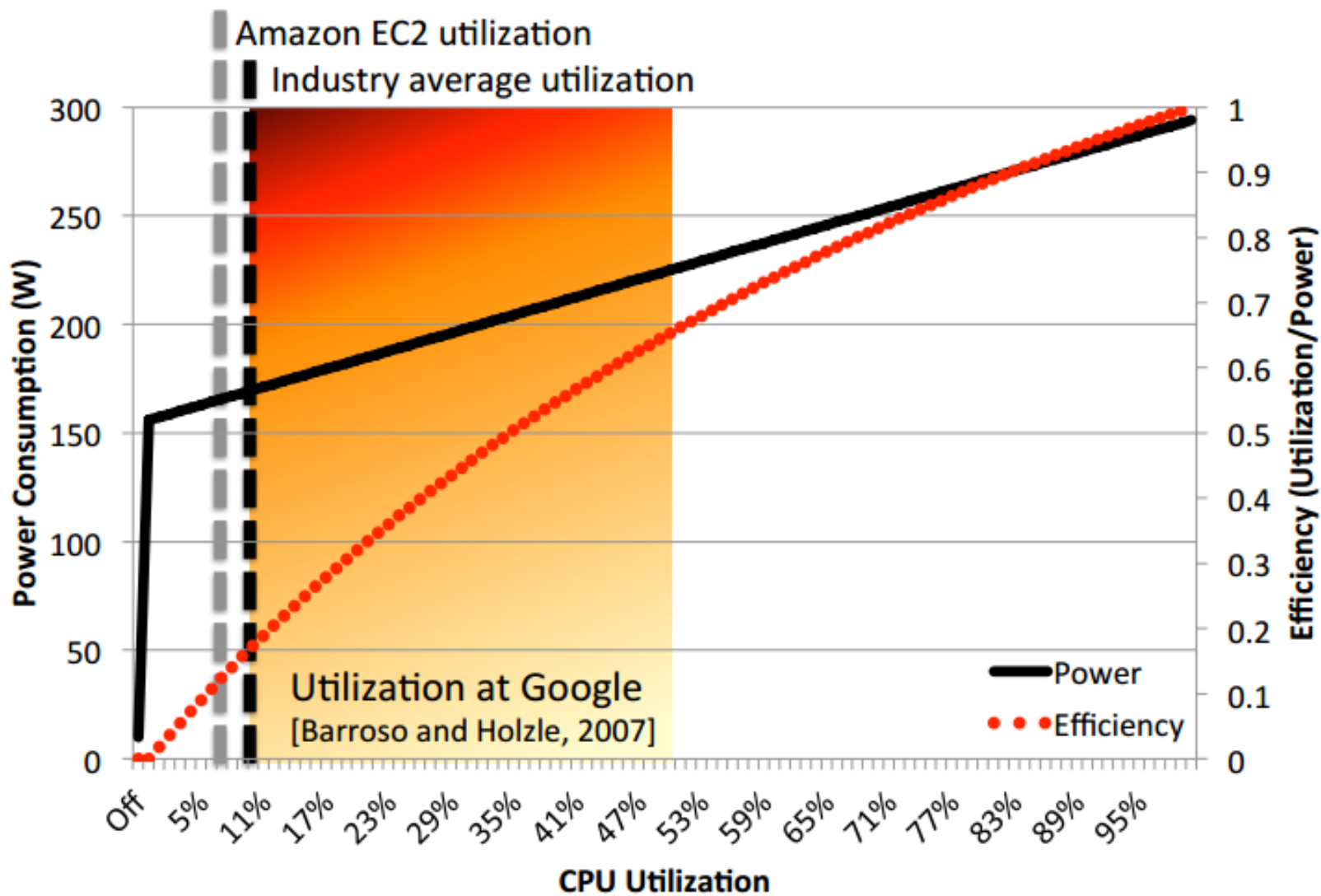
BY JEFFREY DEAN AND LUIZ ANDRÉ BARROSO

The Tail at Scale

Discussion: Utilization Matters

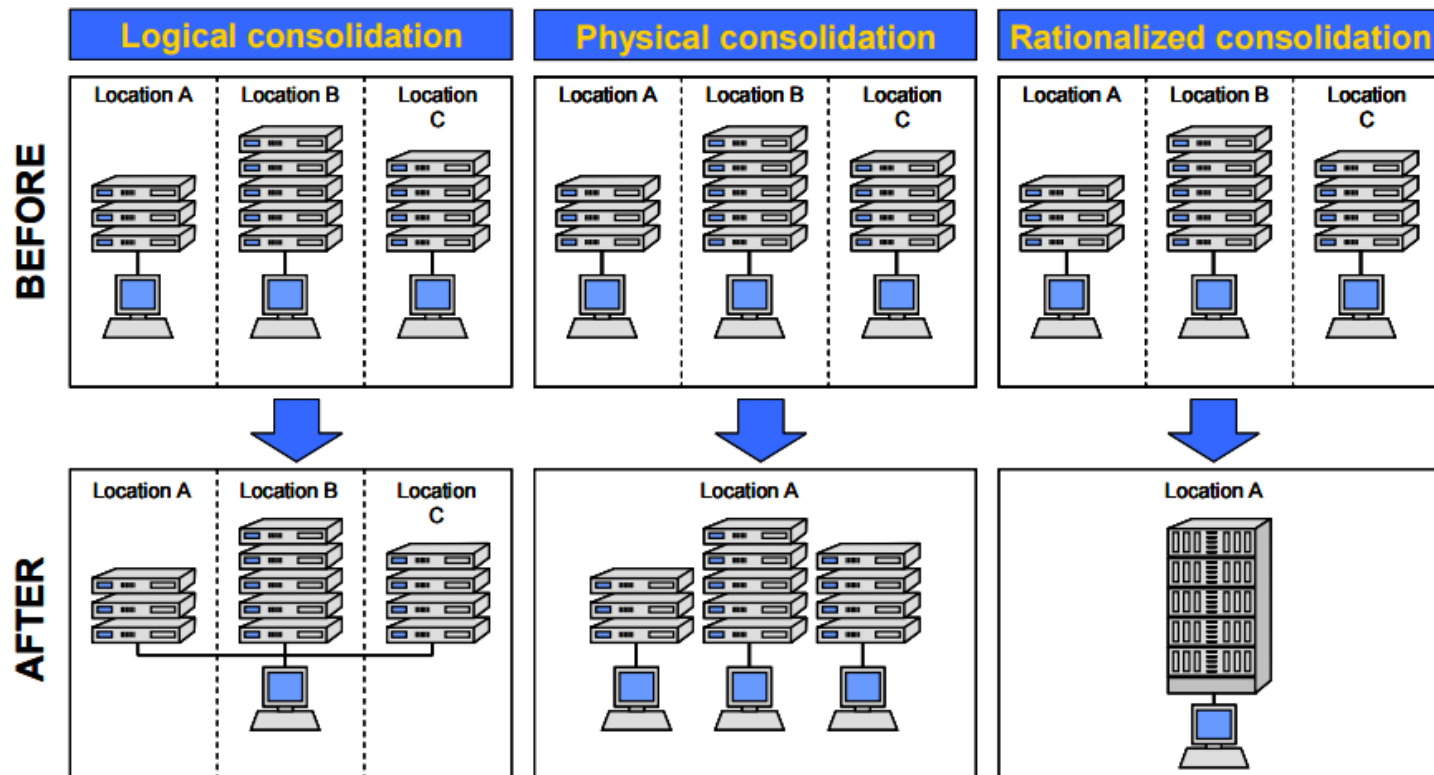


Discussion: Utilization Matters



Discussion: Utilization Matters

- **Workload consolidation improves utilization**
 - Using virtual machine techniques
 - Minimize online physical servers



Discussion: Power Provisioning Problem

- Nameplate Power

- rated capacity; rated power
- the maximum power demand

DC OUTPUT	+3.3V	+5V	+12V1	+12V2	+12V3	-12V	+5Vsb
PEAK(A)	30	30	18	20	10	1.0	2.5
CONTINUITY(A)	20	25	12	12	6	1.0	2.0

MAX. POWER: 191W (3.3V), 360W (5V), 550W (12V)

警告: 内部有危险电压, 请勿打开电源供应器盖子, 请选择正确的输入电压.

HAZARDOUS VOLTAGE INSIDE! DO NOT OPEN POWER SUPPLY COVER! SELECT THE RIGHT INPUT VOLTAGE!

CE, FCC, RoHS, and other certification logos are present.

Nameplate Example

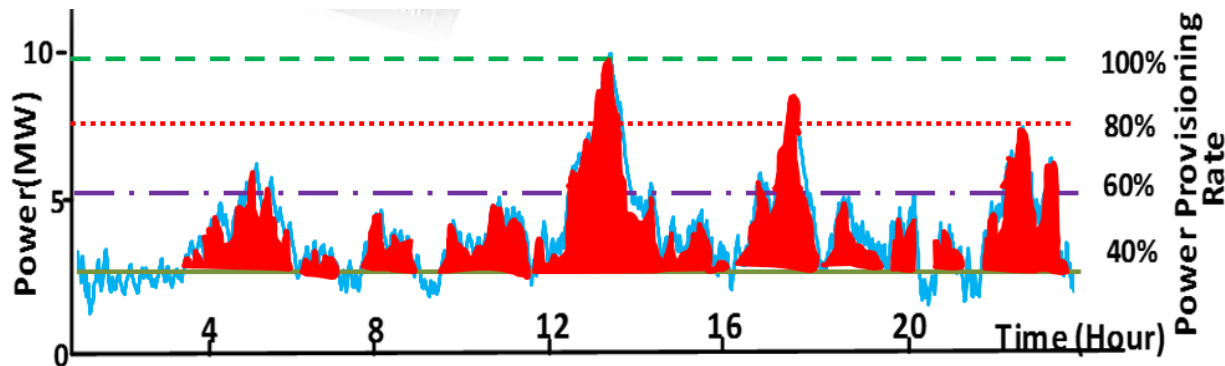
- Over-provisioning (超额供给)

- total server nameplate power < data center power capacity
- nameplate power + worst-case numbers for a safety margin
- conservative design that leads to low power utilization

- Over-subscribing (超额认购)

- total server nameplate power > data center power capacity
- one can not run all the servers at peak speed at the same time
- a.k.a **under-provisioning** (of power capacity)

Discussion: Power Provisioning Problem



Power mismatches due to the rare peak power demands

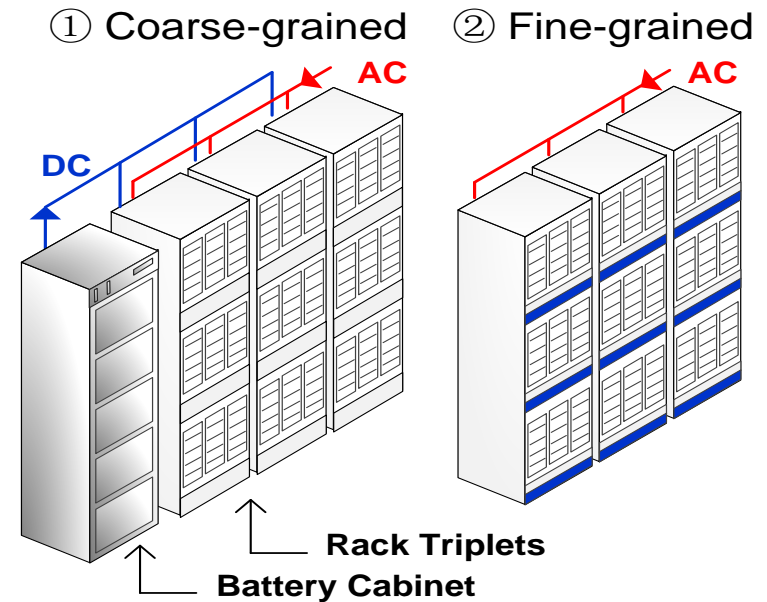
**Power system cost:
\$10-25 per watt !**

**Peak power demand
can be shaved**

- Load power capping
 - Typically a feed-back control based scheme
 - Key tuning knobs: DFS/DVFS/Power Gating
- More ways for peak power shaving

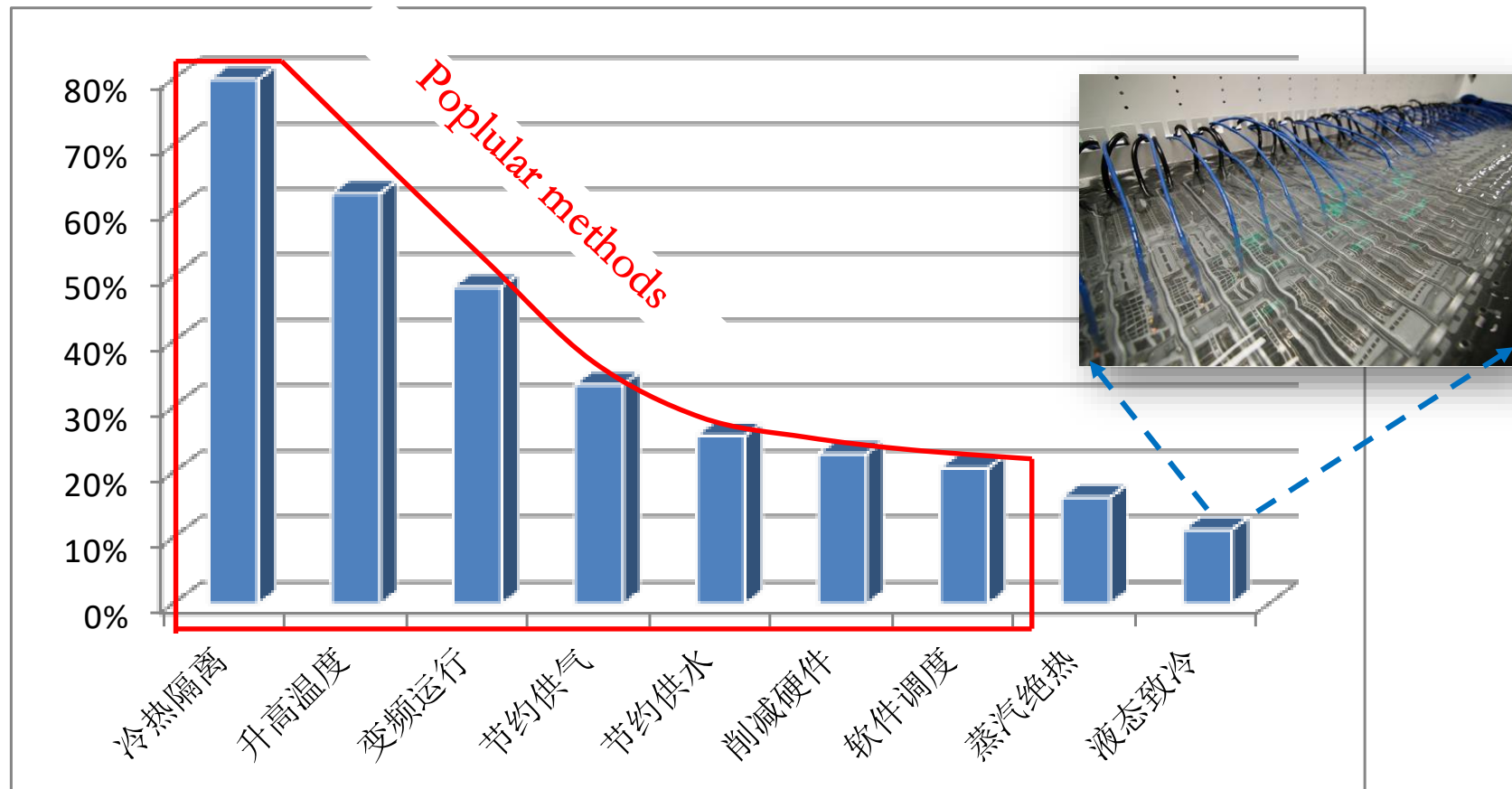
Discussion: Power Provisioning Problem

FB's Distributed Battery System



- Distributed battery and battery-based peak shaving
 - Provides the flexibility of on-demand power peak shaving
 - Other benefits?

Discussion: Cooling Approach



Based on the data from the Uptime Institute, 2014

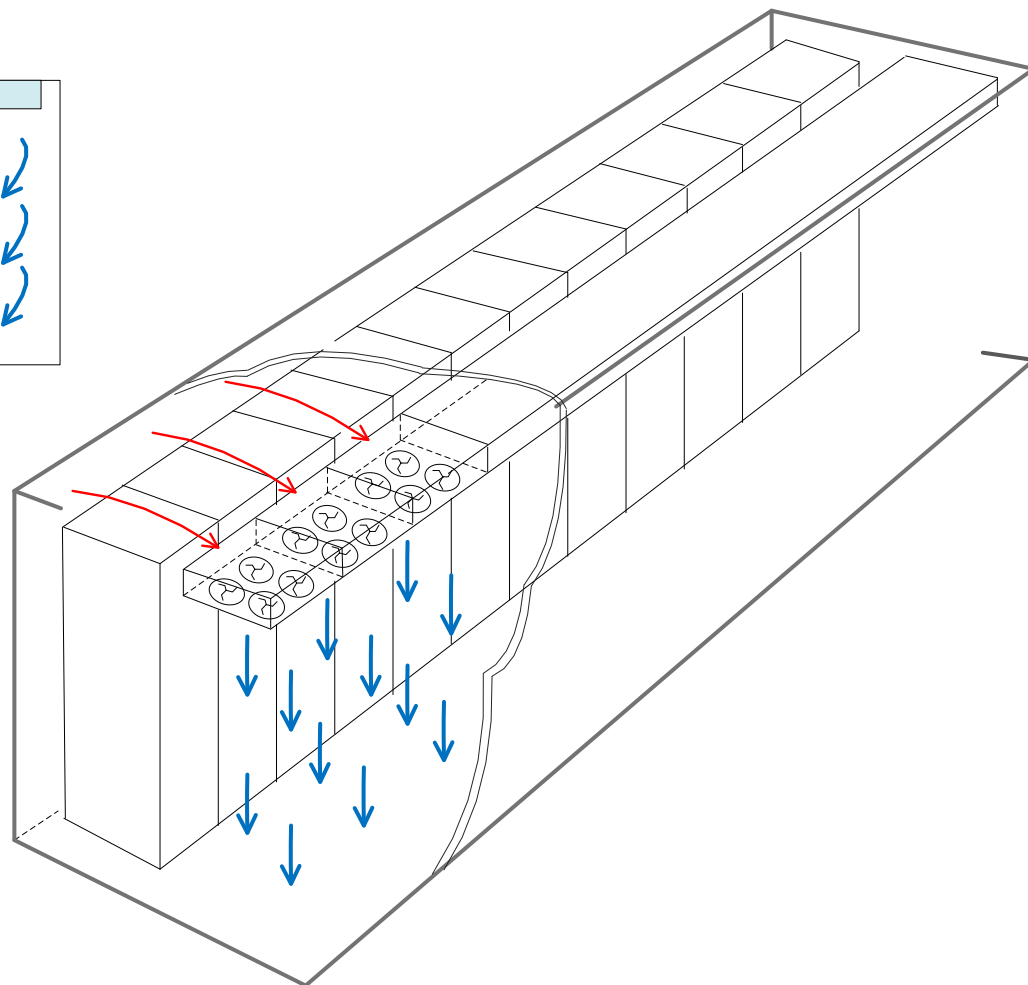
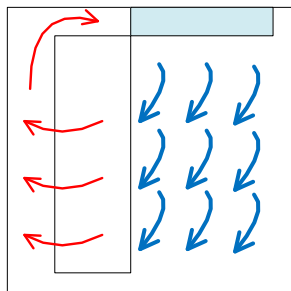
Discussion: Cooling Approach

- A modular data center (MDC) system is a portable method of deploying data center capacity.
- Similar concept:
 - containerized data center or portable data center



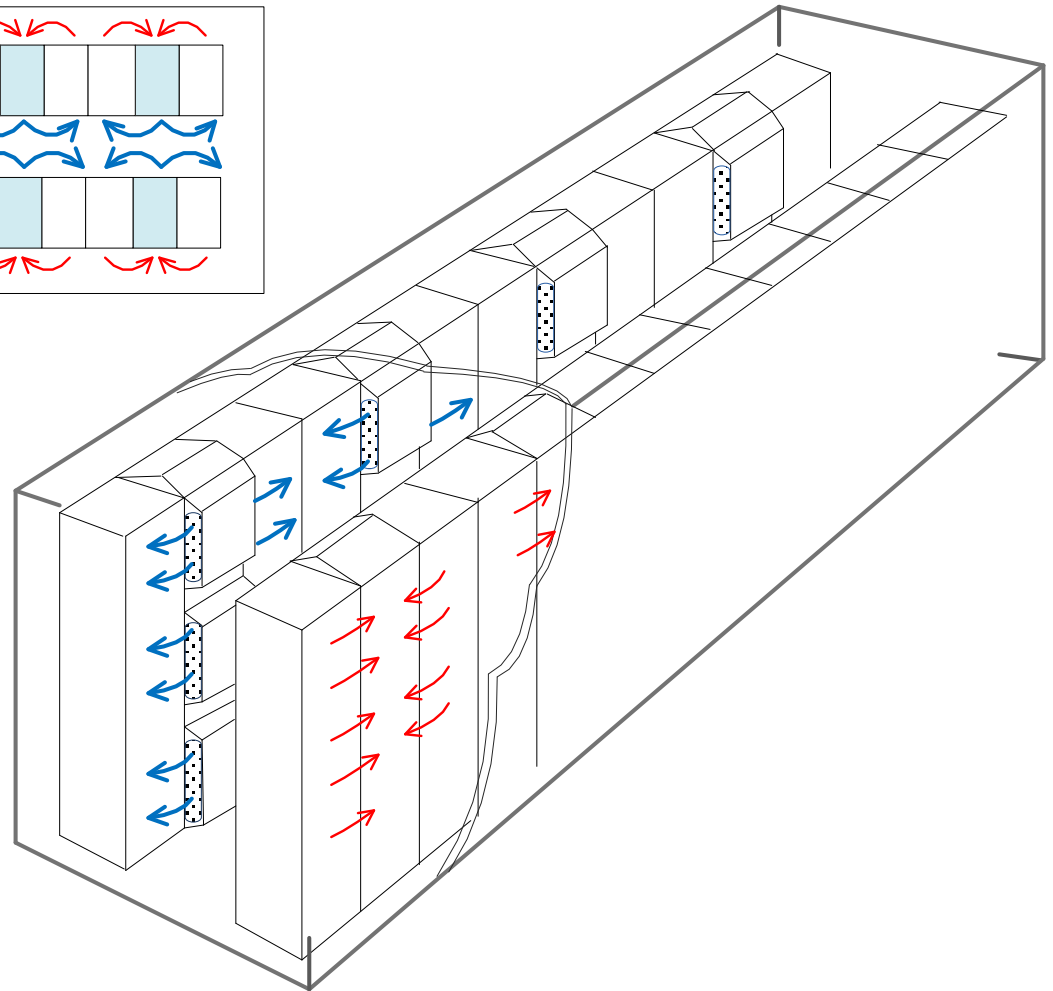
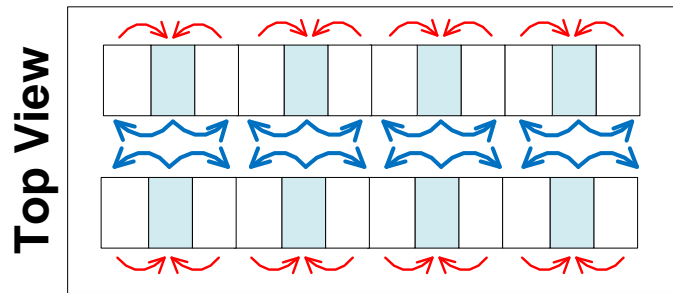
Discussion: Cooling Approach (MDC Cooling)

主视图



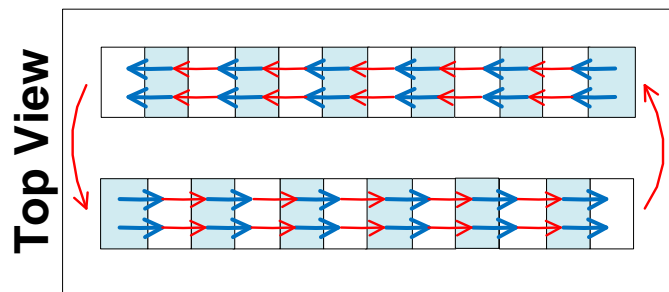
(A) 单列顶端制冷
例如惠普公司的POD数据中心

Discussion: Cooling Approach (MDC Cooling)

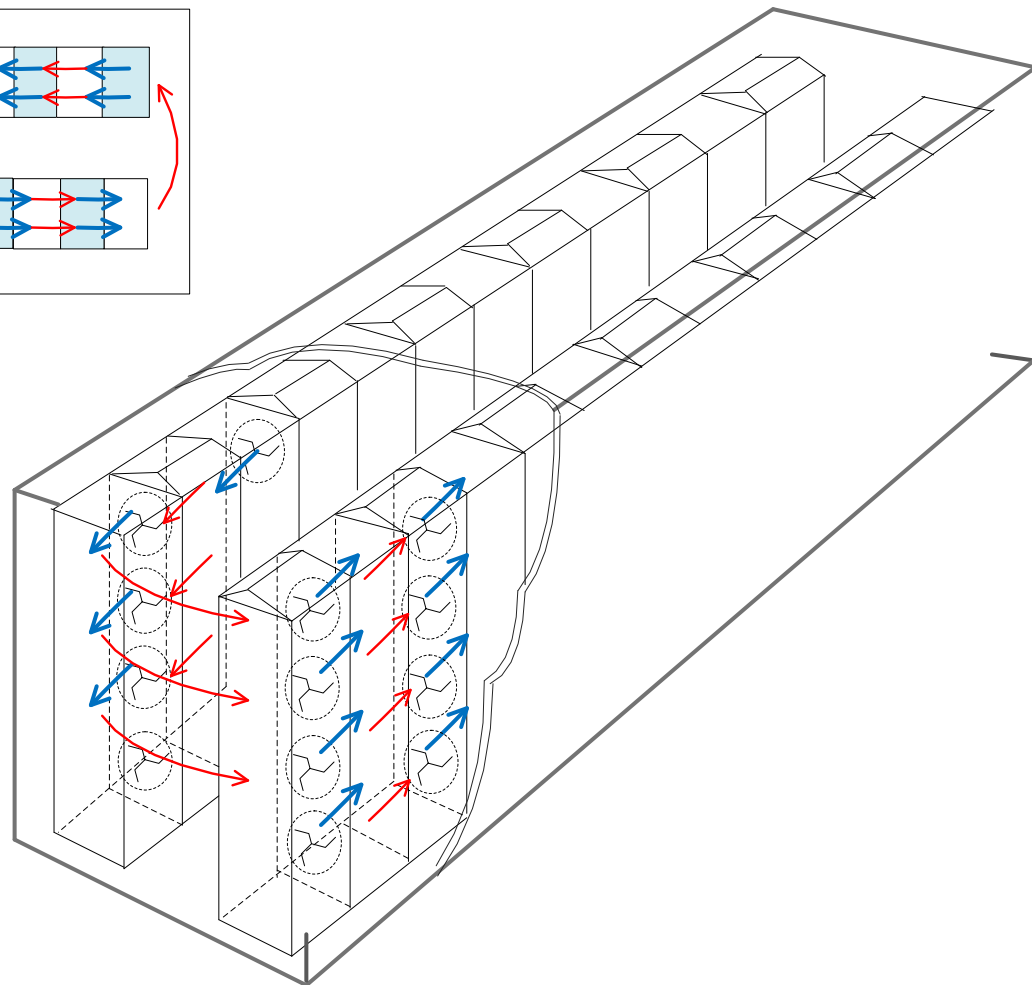


(B) 双列柜间制冷
例如Sun的模块化数据中心

Discussion: Cooling Approach (MDC Cooling)



(C) 柜间循环制冷
例如SGI公司的ICE Cube



Summary

- What is a data center
- Major metrics of data center design
- Data center infrastructure
 - Power system
 - Cooling system
 - ICT system
- The long tail concept
- Data center capacity utilization
- Types of power provisioning
- Modular data center and cooling

References

- 课本内容：J. Hennessy, D. Patterson. Computer Architecture, Fifth Edition: A Quantitative Approach.
 - Chapter 6
- 其它参考：L. Barroso et al., 《The Datacenter as a Computer》, Second Edition.

Exercises

- The key differences between a Tier-2 data center and a Tier-4 data center?
- Why conventional raised-floor cooling scheme lead to low efficiency?
- Try to explain over-provisioning, under-provisioning and over-subscription in the context of data center power management.