

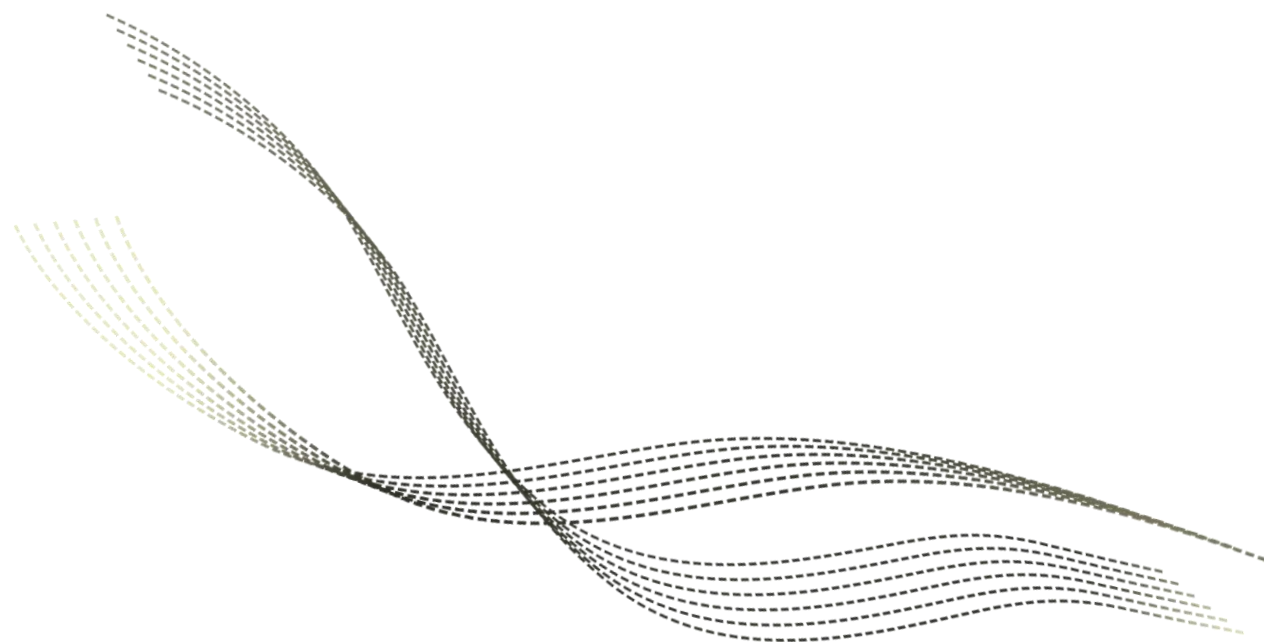
# 高级计算机体系结构

Advanced Computer Architecture

数据中心即计算机

---

沈明华



# 目录

CONTENTS

01

数据中心基础设施

02

关键设计因素

# **PART 01**

## **数据中心基础设施**

## ■ 背景

## □ 数据中心

## — 数据中心的核心功能

- 提供数据处理(计算)服务
  - 并发处理非常多的请求(以请求级并行的方式request-level parallelism)
- 提供数据存储服务
  - 大规模, 高可靠的数据存储
- 提供数据传输服务
  - 高带宽

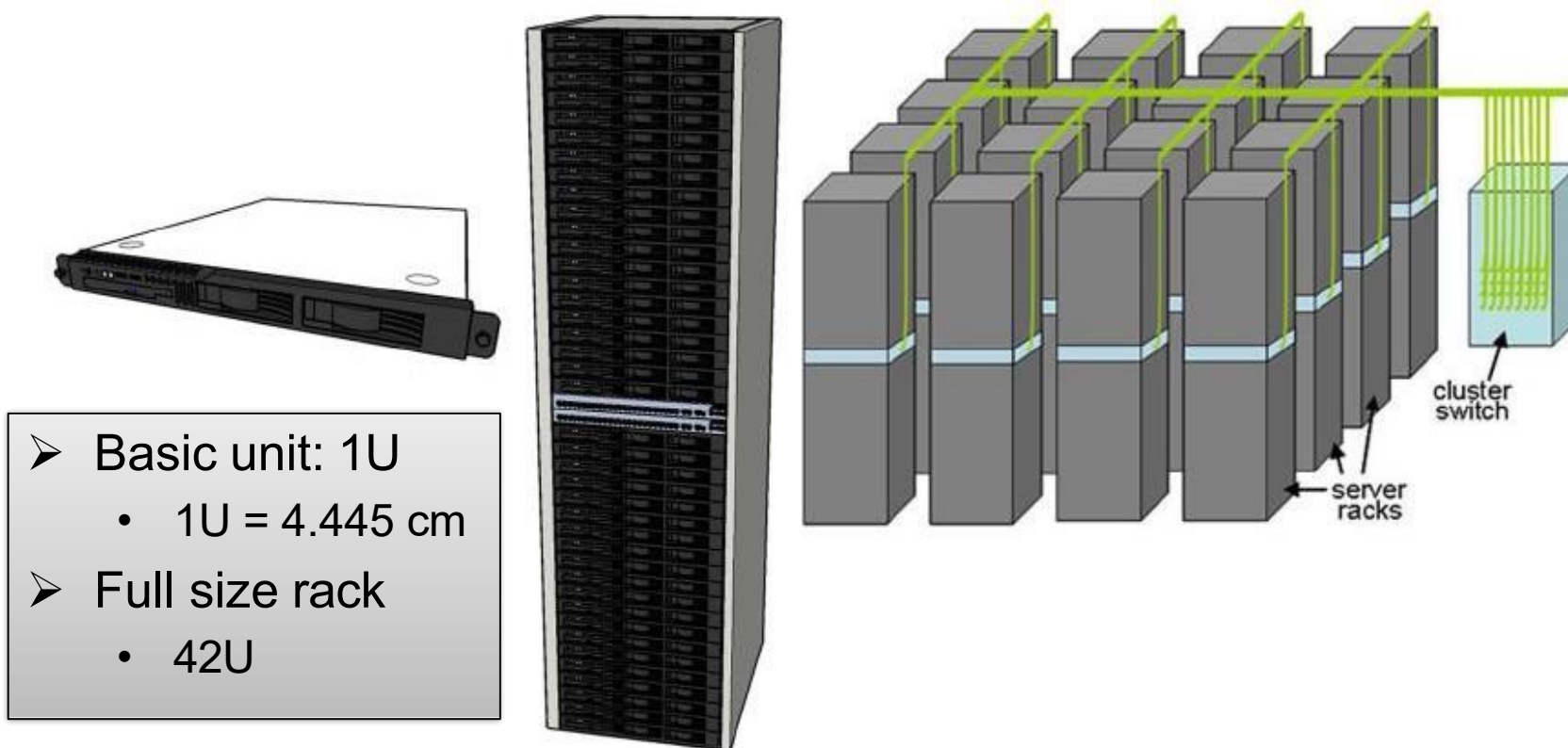
## — 数据中心的组成设备

- ICT设备、电力系统、冷却系统
- 其他支持类设备，如灯光、安全等

# ■ 数据中心层次结构

## □ WSC(warehouse scale computer)划分层次

- Server/Node层次; Rack/Cluster层次; Facility(数据中心层次)



# ■ 数据中心节能

---

## □ 能效比(PUE, power usage effectiveness)

- 提供一个关于数据中心能效的总览, 即有多少能源真正用于计算
- 越接近1能效越好

$$PUE = \frac{\text{数据中心总能耗}}{\text{IT设备功率}}$$

## □ 绿色数据中心的评价指标

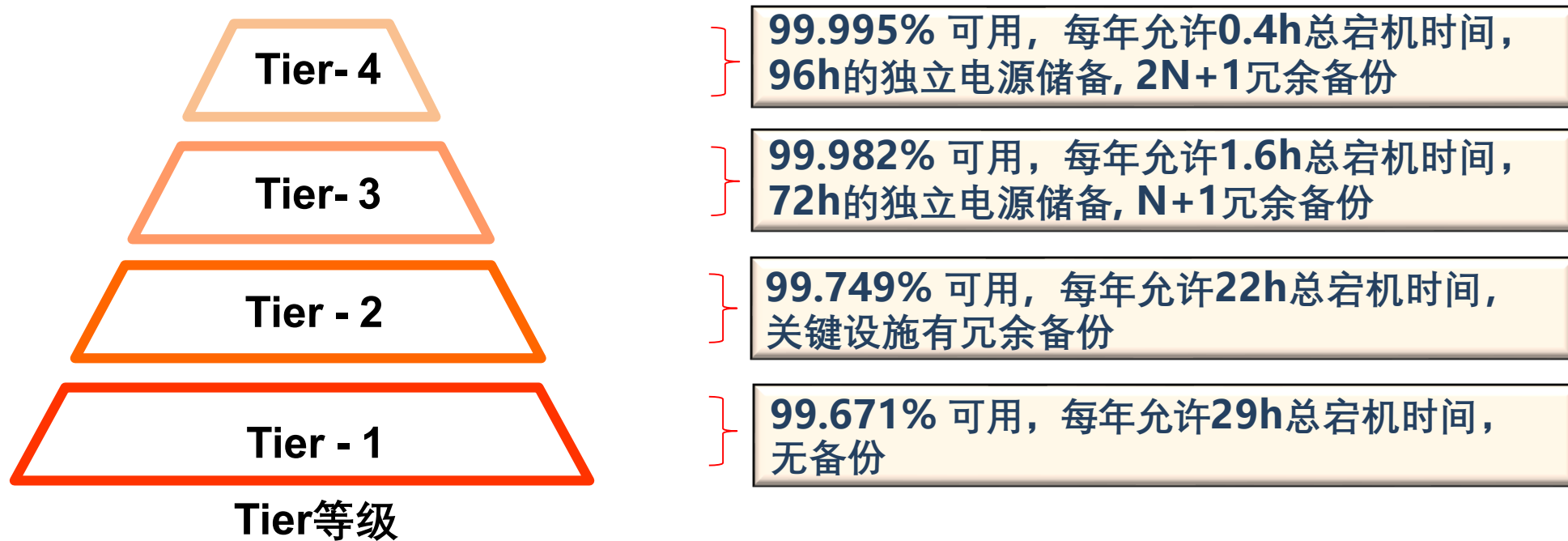
- WUE: water usage effectiveness(单位L/kWh)
- CUE: carbon usage effectiveness(单位kgCO<sub>2</sub>/kWh)

$$WUE = \frac{\text{数据中心平均用水量}}{\text{IT设备功率}}$$

$$CUE = \frac{\text{数据中心总碳排放}}{\text{IT设备功率}}$$

# ■ 数据中心等级认证体系

## □ Uptime的Tier体系评价数据中心的等级

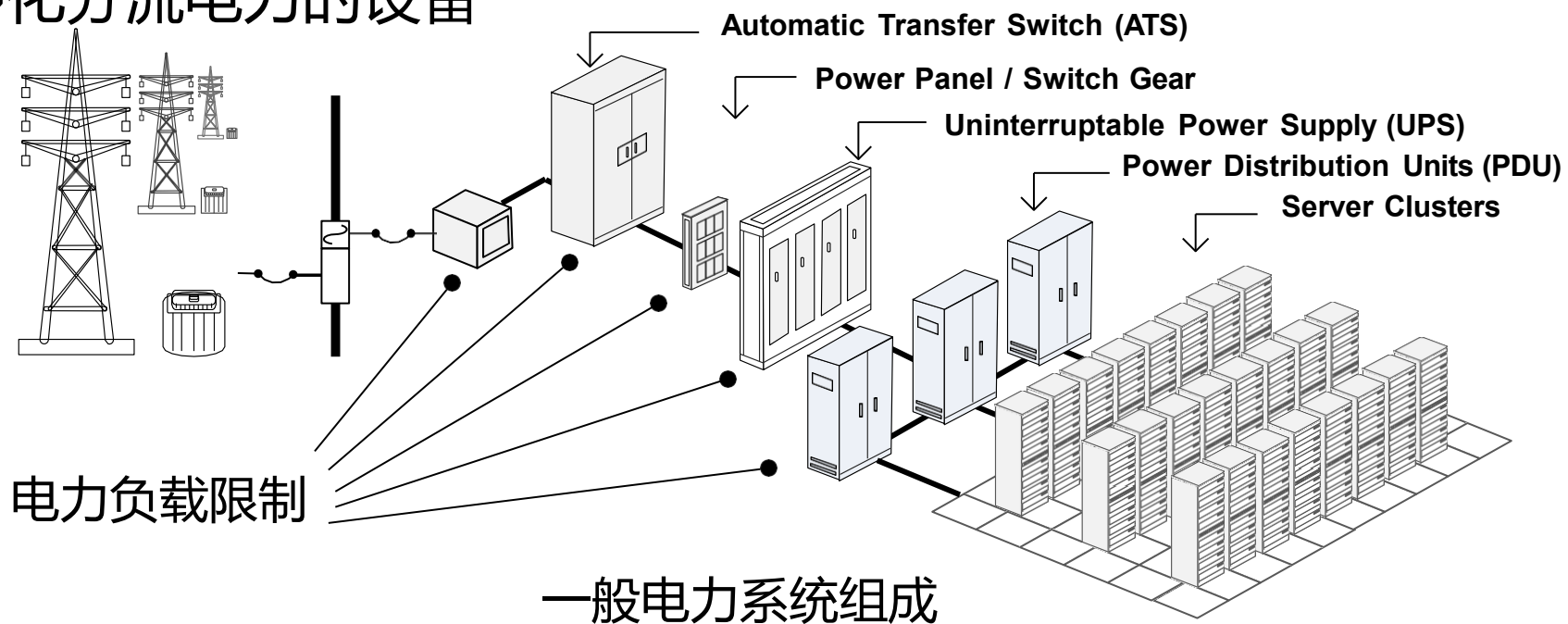


UptimeInstitute®

# ■ 数据中心基础设施

## □ 电力系统组成

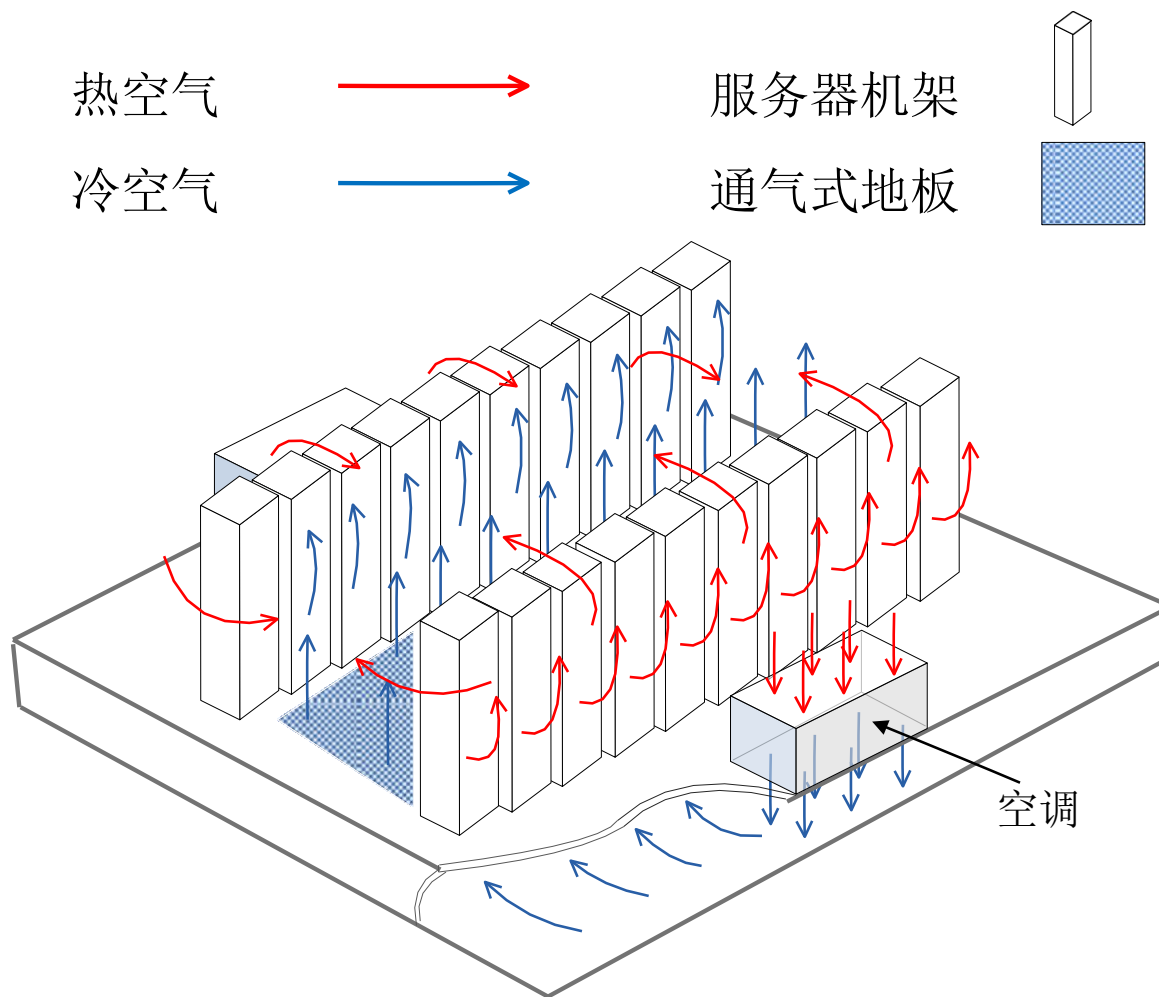
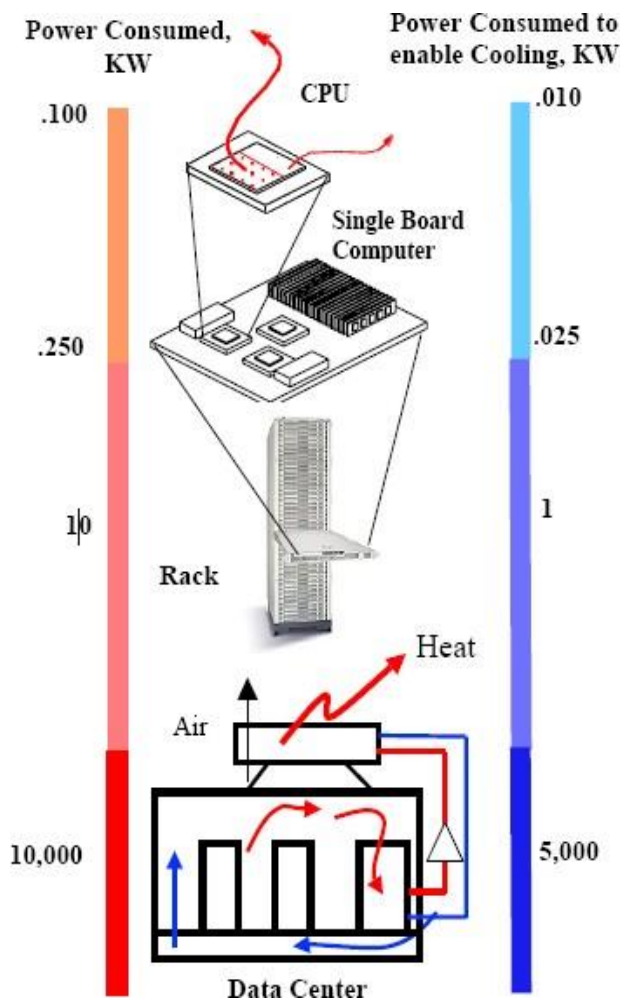
- ATS：一个快速的机械开关，人工控制的电闸
- STS：一个超快速的电子开关，自动控制的电闸
- UPS：通常是带控制接口的一组电池，防止电力波动烧坏电子设备
- PDU：转化分流电力的设备





# ■ 数据中心基础设施

## □ 冷却系统



# ■ 数据中心基础设施

---

## □ 冷却系统

- 制冷系数(COP系数, coefficient of performance)
  - 表示制冷设备输入单位能源, 能提供的制冷量



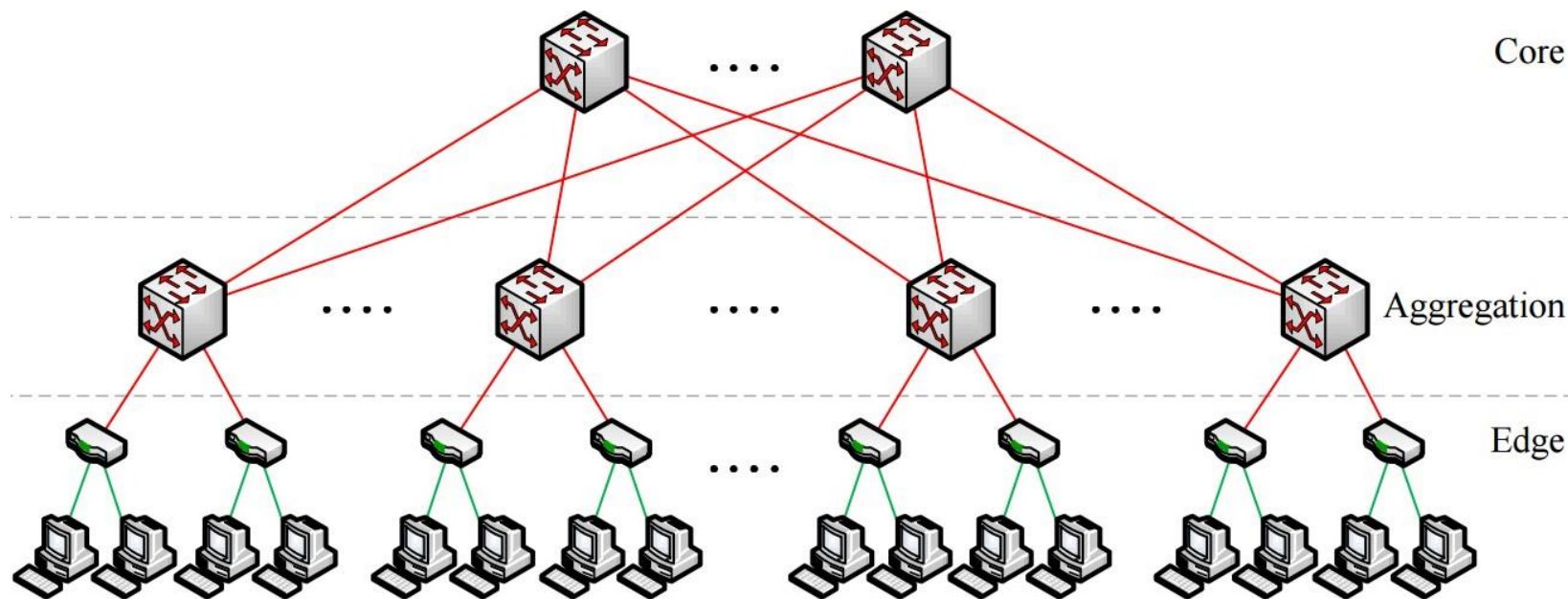
# ■ 数据中心基础设施

## □ ICT系统

### – 两种互联交换机

- 在机架顶部(TOR, top of rack)
  - 聚合机架上的服务器
- 在一行机架末尾(EOR, end-of-row)
  - 聚合多个机架的服务器

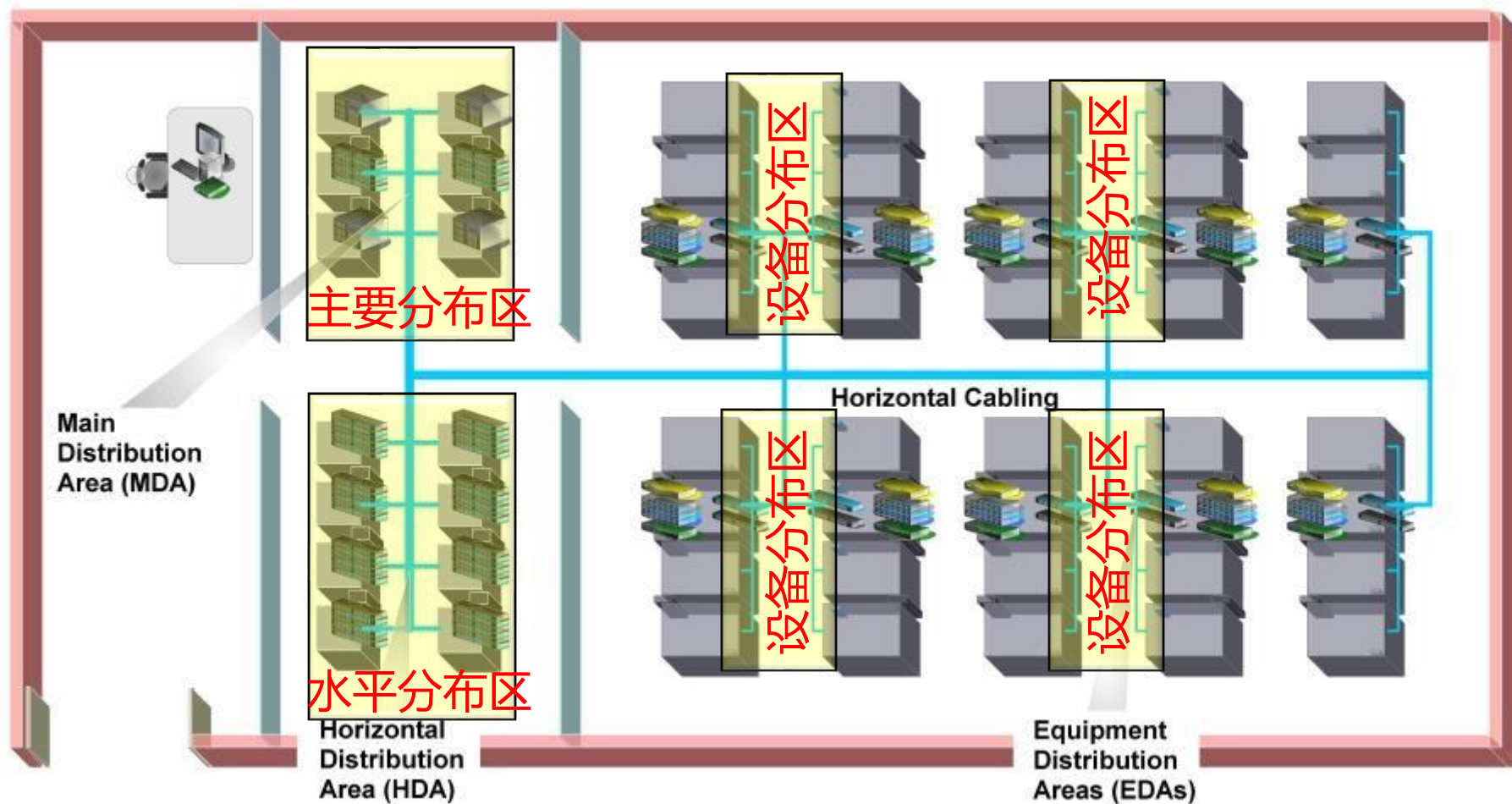
	TOR1G	TOR10G	EOR
GbE Ports	48	0	0
10GbE Ports	4	24	128
Power (W)	200	200	11,500
Size (RU)	1	1	33



# ■ 数据中心基础设施

## □ ICT系统

### — 分层网络拓扑



# **PART 02**

## **关键设计因素**

# ■ 背景

---

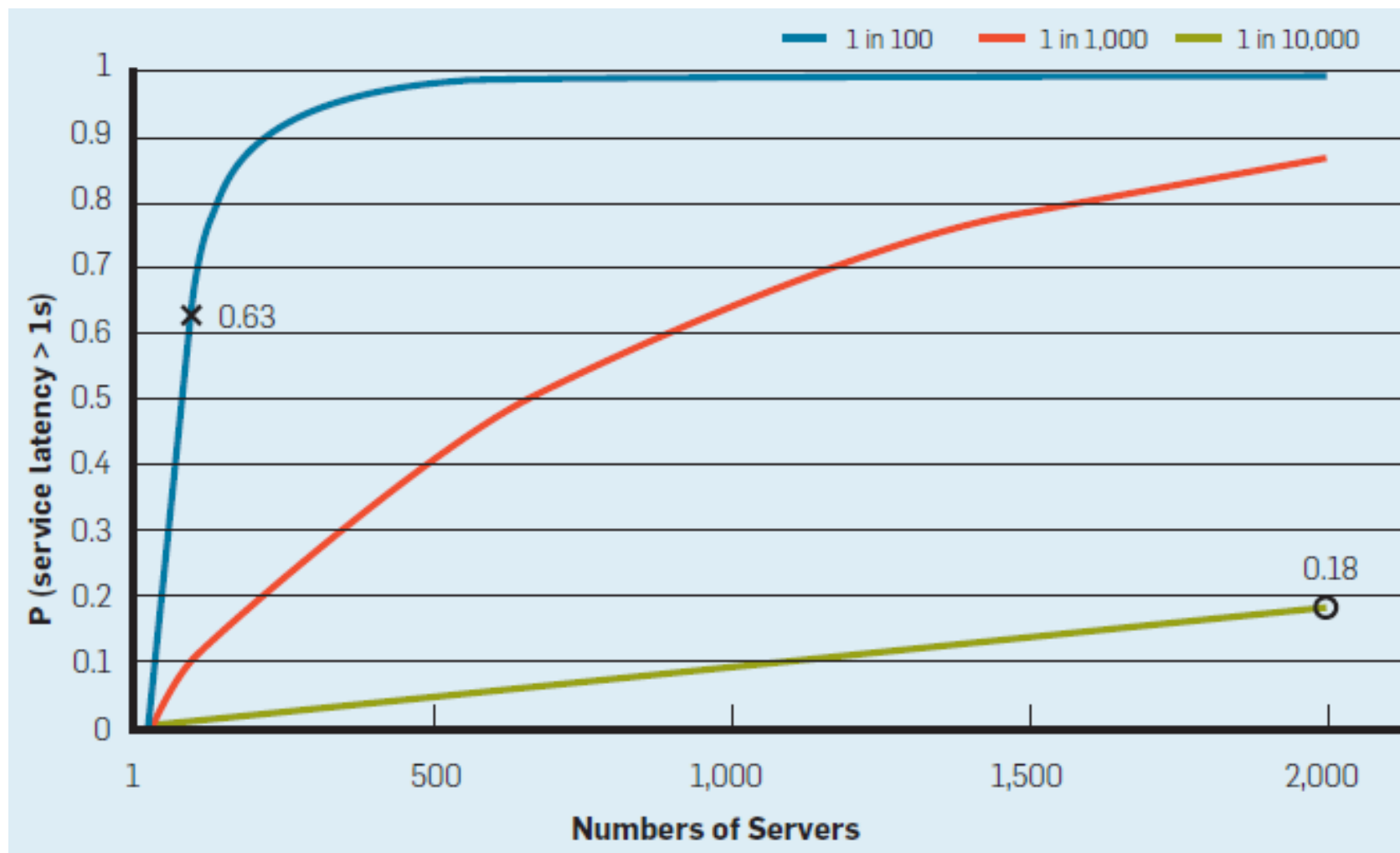
## □ 设计和运营一台仓库级计算机(WSC)十分具有挑战性

- 需要考虑性能、能效、独立性/可用性、网络等等因素
- 本节介绍
  - 降低服务延迟小概率事件的发生
  - 提高数据中心算力利用率
  - 满足数据中心电力供给最优策略
  - 模块化数据中心方案

# ■ 讨论：The Tail at Scale

□ 随着服务器规模的增加，出现秒级响应的概率会增加

若用户调用100台计算机提供服务，每台机器有1%的概率延迟响应，那么响应超过1秒的概率为  
 $0.63 \approx (1 - 0.99^{100})$



# ■ 讨论：The Tail at Scale

---

- 一种简单的策略消除这种随机事件，降低秒级响应出现概率
  - 同时发射多个相同的请求，仅采用最快响应的结果

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

BY JEFFREY DEAN AND LUIZ ANDRÉ BARROSO

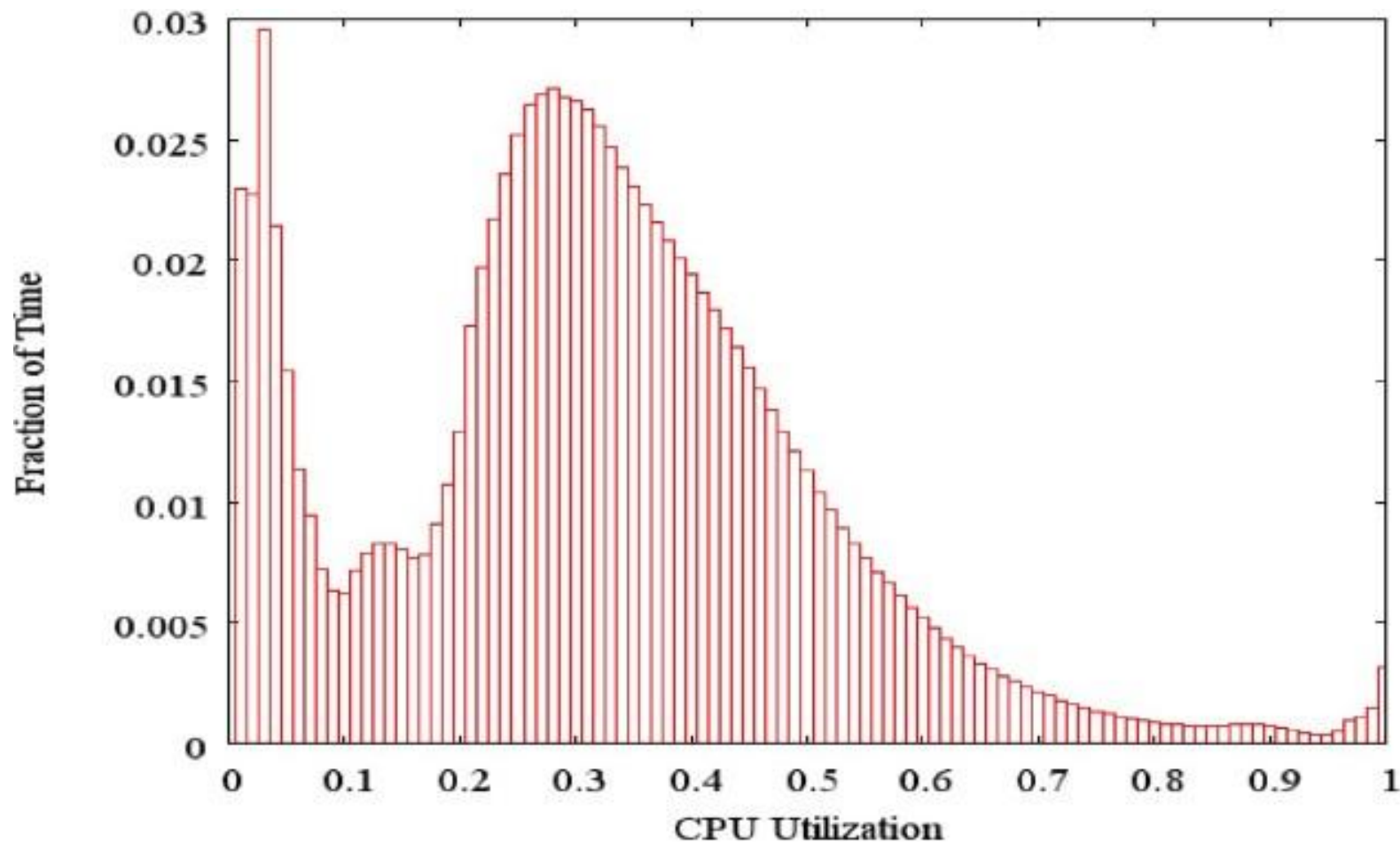
## The Tail at Scale



# ■ 讨论：利用率

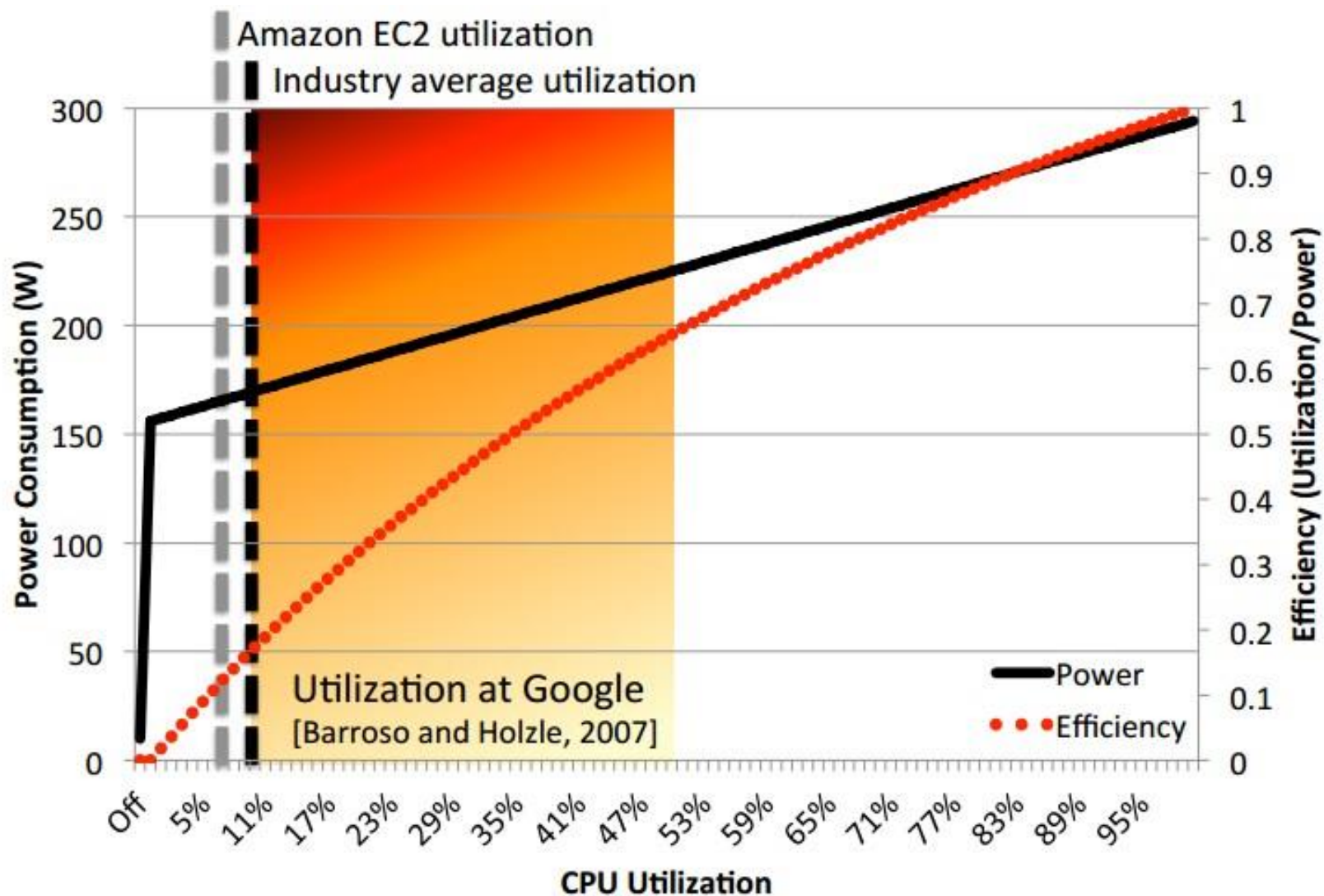
## □ CPU利用率出现频率

- 数据中心CPU利用率大部分在30%左右



# ■ 讨论：利用率

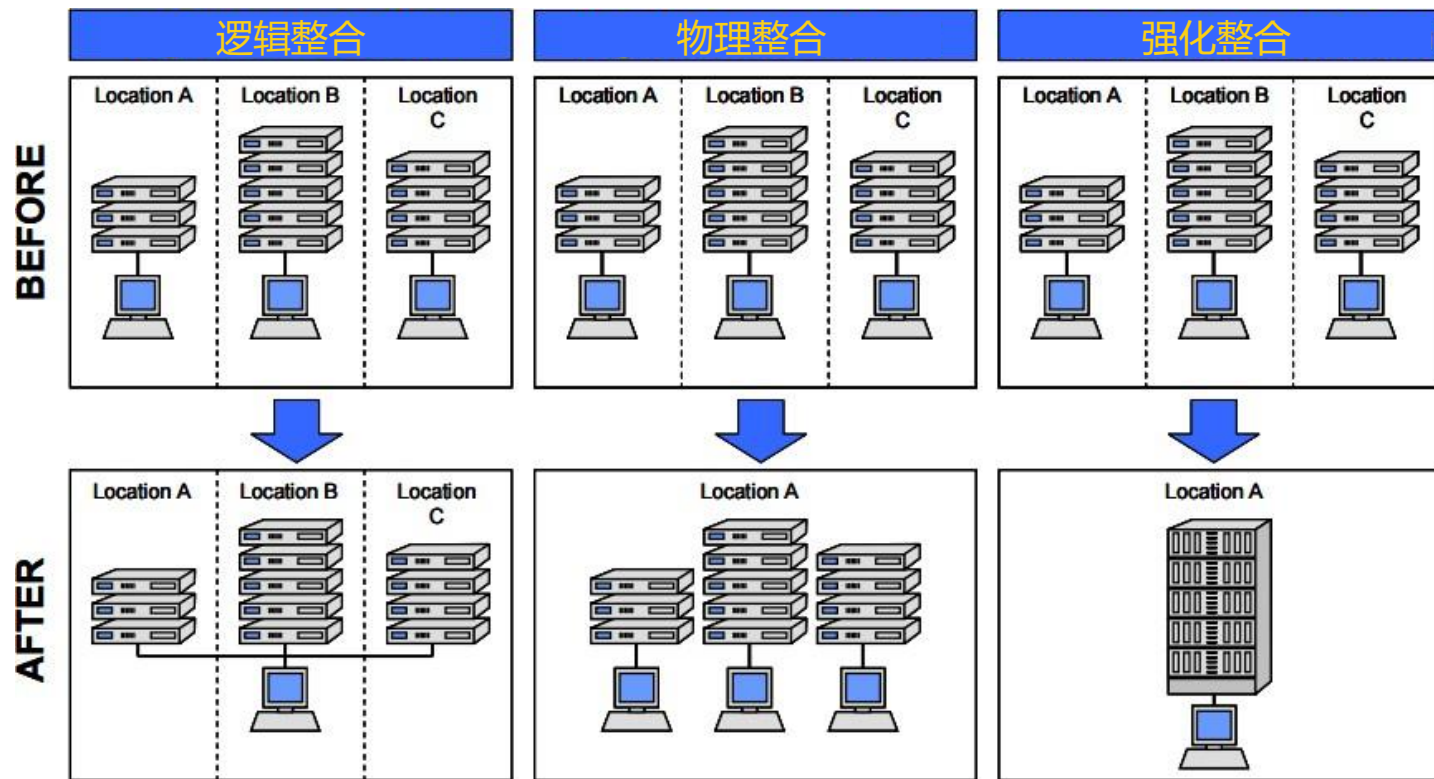
## □ CPU利用率和能效的关系



# ■ 讨论：利用率

## □ 整合工作负载提高系统利用率

- 利用虚拟机技术，一台物理机虚拟成多台提供给用户使用
- 最小化在线的物理服务器数量，整合多处服务器资源以提高利用率



# ■ 讨论：电力供给问题

---

## □ 超额供给(over-provisioning)

- 服务器额定功率 < 数据中心设计功率
- 过于保守的设计导致能源利用率低

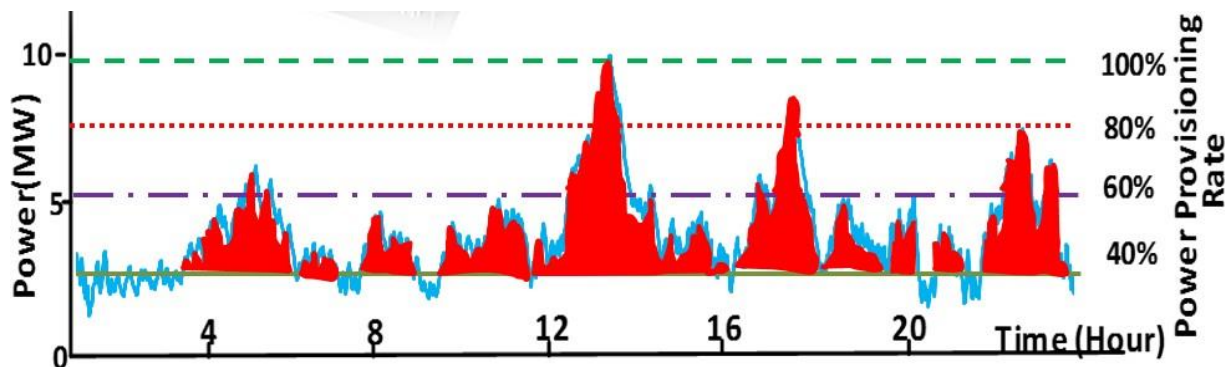
## □ 超额认购(over-subscribing)

- 服务器额定功率 > 数据中心设计功率
- 不能让所有服务器同时进行全功率计算
- 同电力供应不足

# ■ 讨论：电力供给问题

## □ 数据中心电力需求波动较大

- 而电力供给系统每瓦的成本在10~24\$
- 对数据中心电力需求进行削峰，可以降低电力供给系统成本

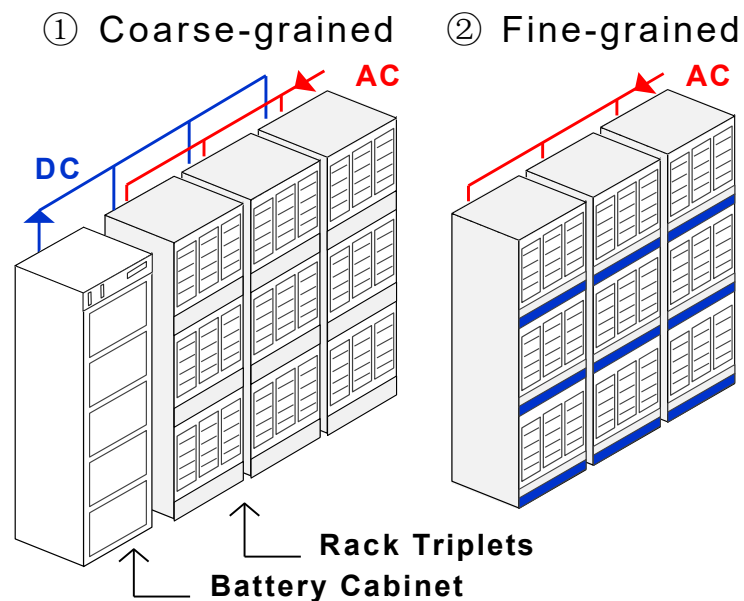


**Power mismatches due to the rare peak power demands**

# ■ 讨论：电力供给问题

- 基于电池组的削峰措施
  - 灵活按需削减电力峰值

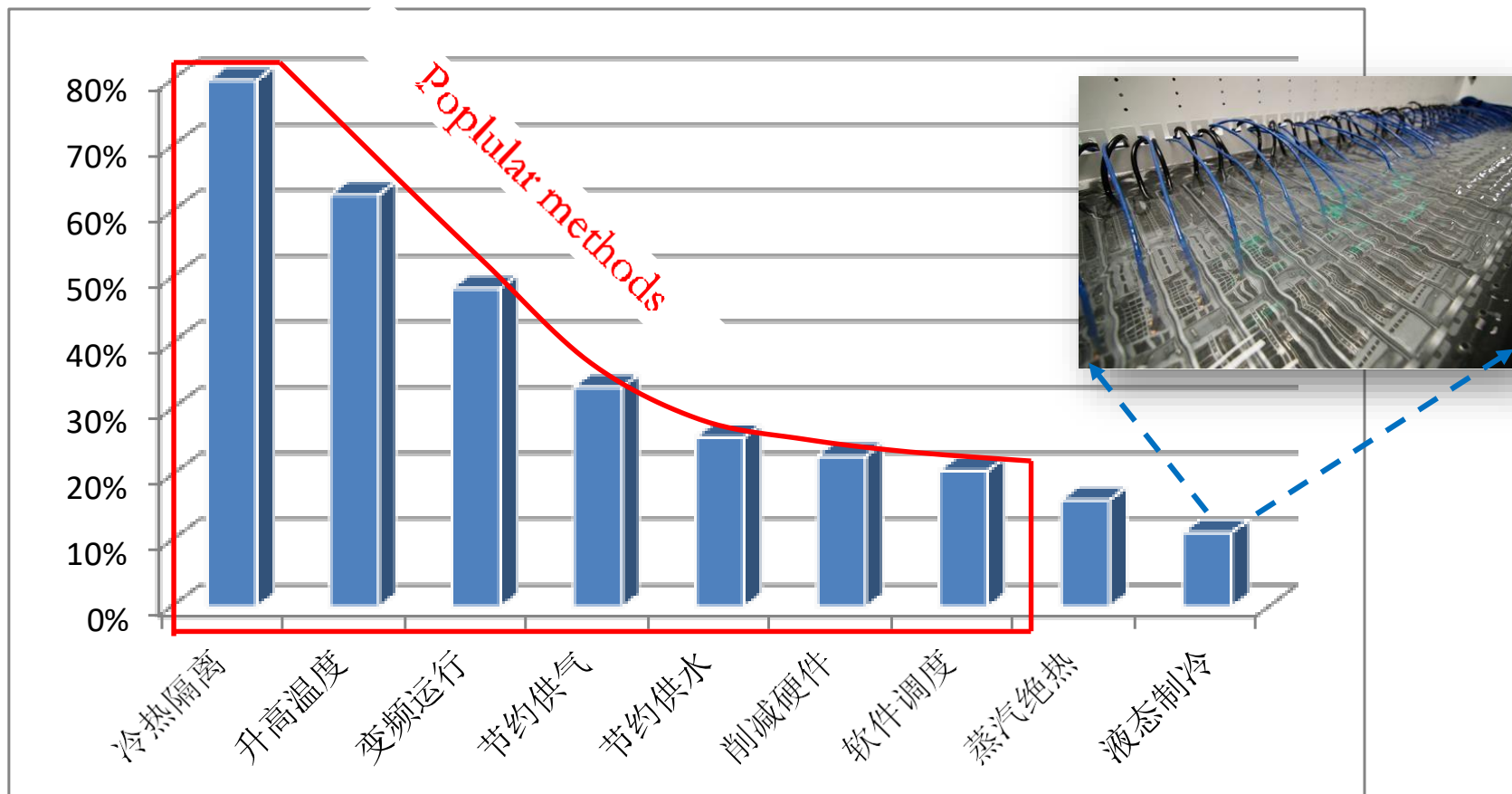
Facebook的分布式电池组





# ■ 讨论：冷却系统

## □ 冷却系统最受欢迎的措施



Based on the data from the Uptime Institute, 2014

## ■ 讨论：冷却系统

---

### □ 模块化数据中心系统(MDC, modular data center)

- 是一种便携式扩展数据中心容量的方法
- 又称为集装箱化数据中心，或便携式数据中心





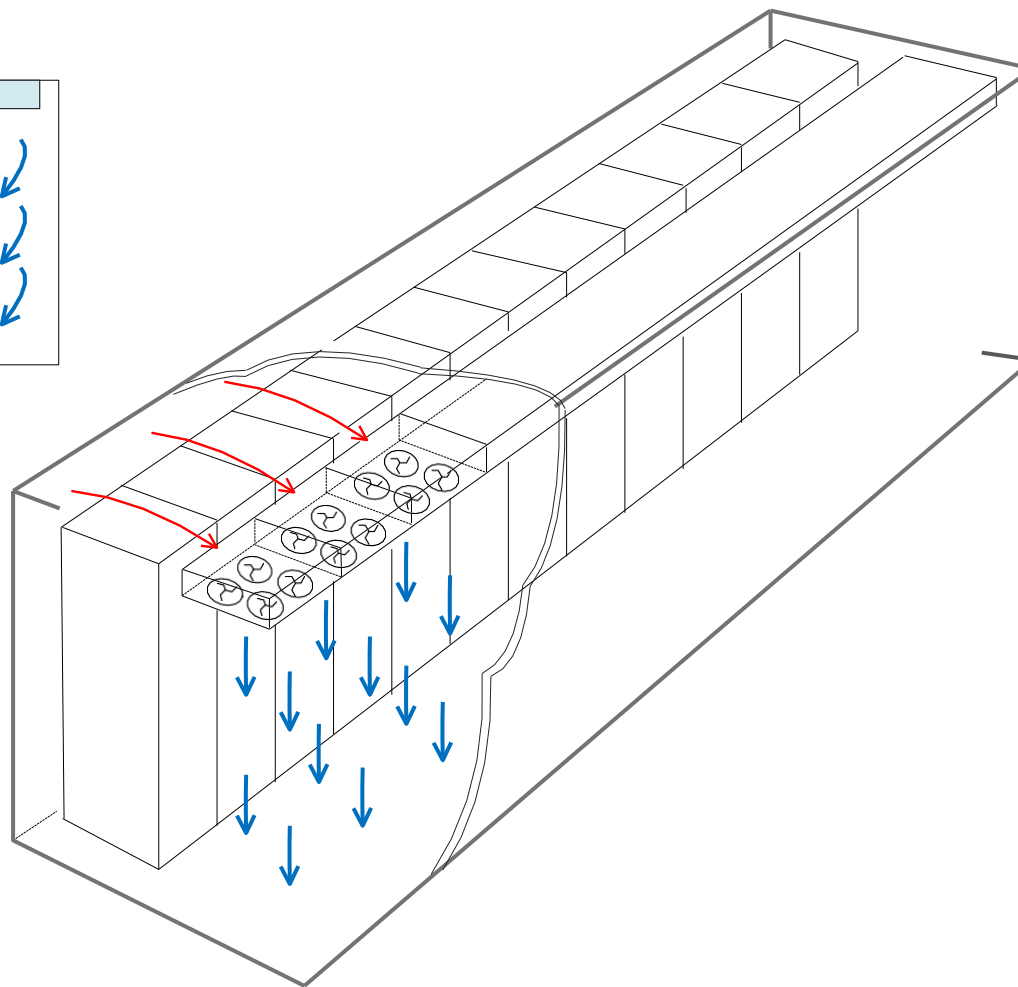
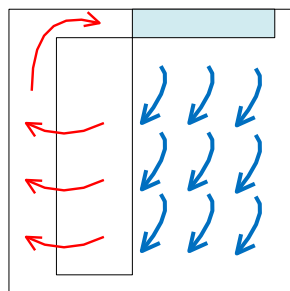
# ■ 讨论：冷却系统

## □ 例如惠普公司的POD数据中心

### — 单列顶端制冷



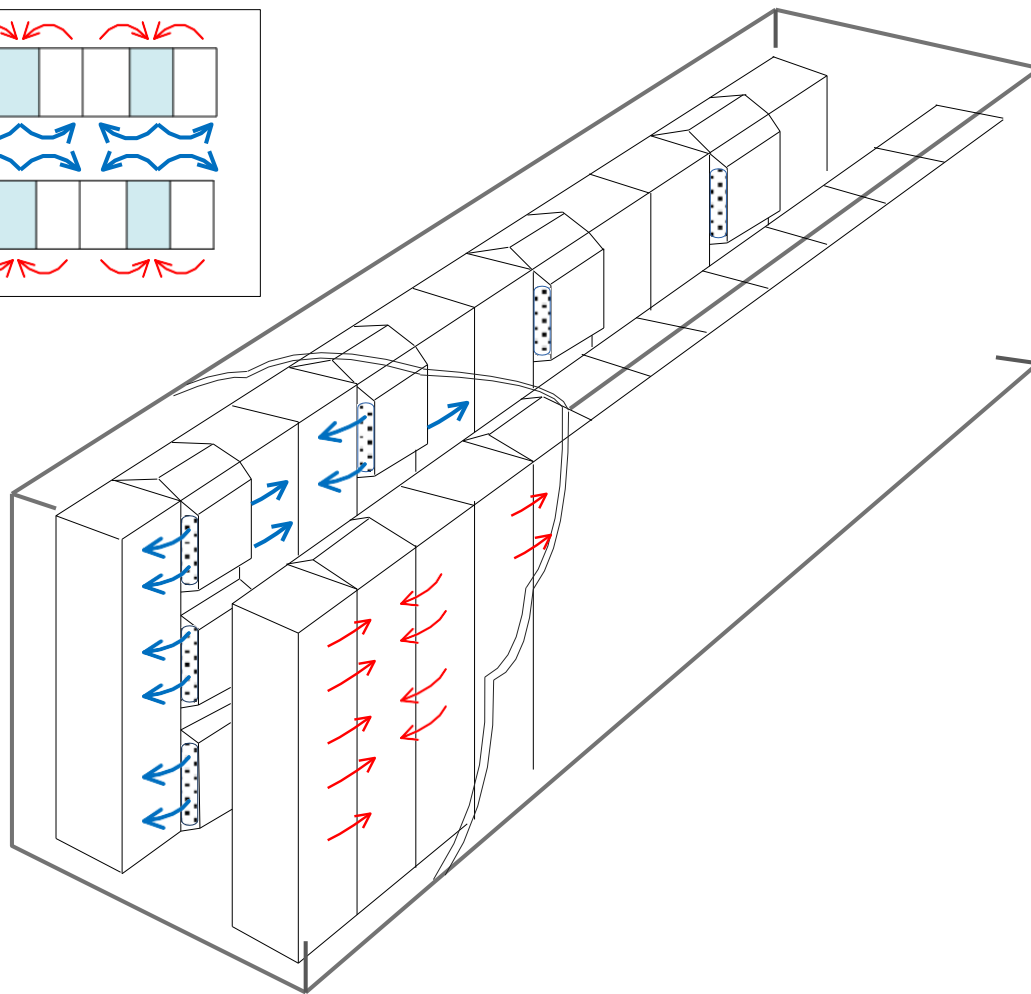
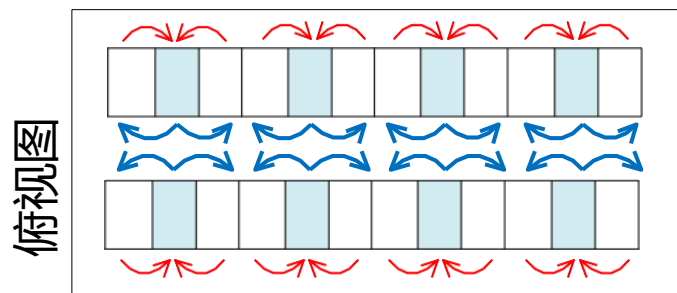
主视图



# ■ 讨论：冷却系统

## □ 例如Sun的模块化数据中心

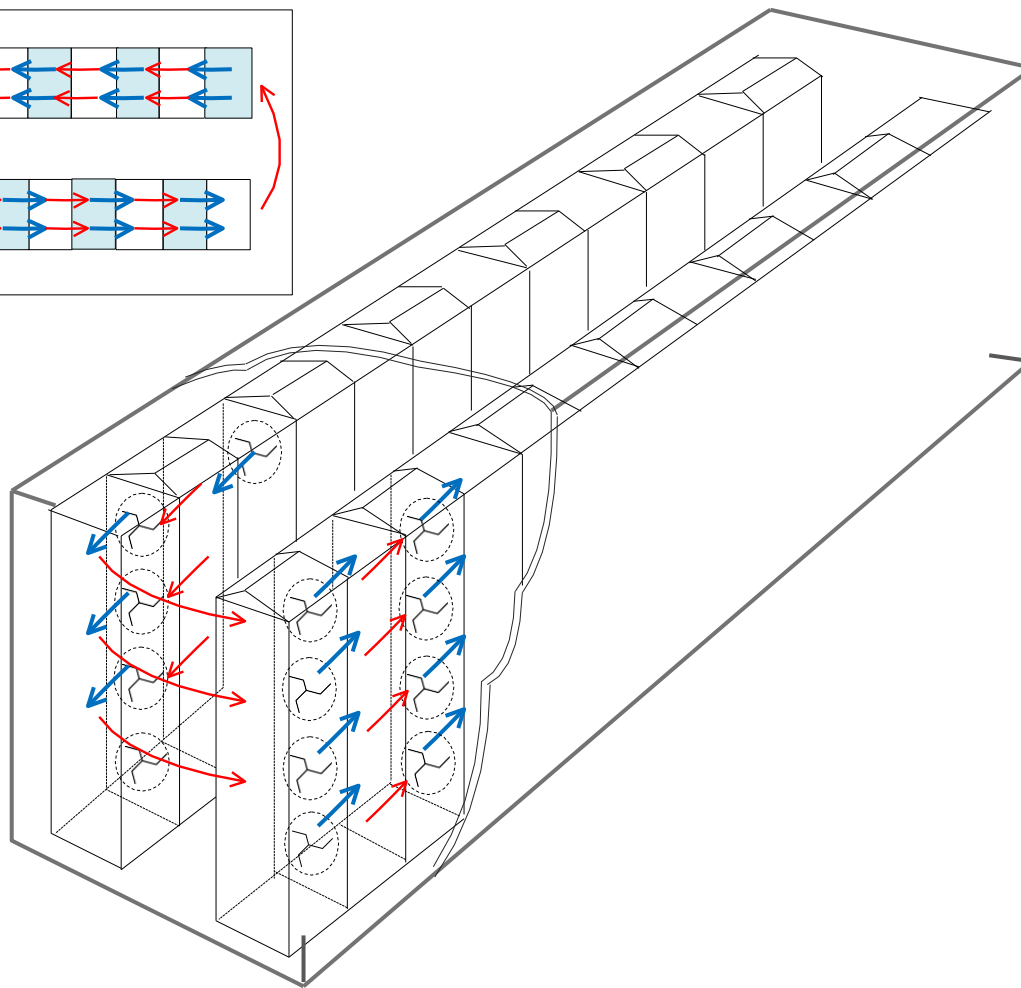
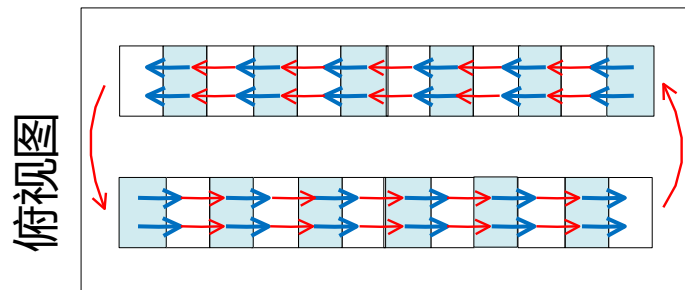
### — 双列柜间制冷



# ■ 讨论：冷却系统

## □ 例如SGI公司的ICE Cube

— 柜间循环制冷



# ■ 思考

---

## □ 开放性练习

- Tier-2数据中心合Tier-4级数据中心关键不同点在哪里
- 解释数据中心的超额供给(over-provisioning)、超额认购(over-subscribing)和缺额供给(under-provisioning)的内容

**感谢！**

---