分布式计算

03-云计算

Weixiong Rao 饶卫雄
Tongji University 同济大学软件学院
2023 秋季
wxrao@tongji.edu.cn

内容概述

- **计算规模的问题**
  - 可扩展性的必要性及现有情况
  - Scale-up向上扩展：从PC服务器到“数据中心”
  - 经典扩展性方法的困境

- **云计算**
  - 什么是云计算
  - 云计算类型
  - 什么应用适合于云计算
  - Virtualization虚拟化：云计算的动力
  - 云计算的难点

现代分布式系统的数量规模(用户和对象)

- Facebook: 10亿活跃用户
- Google: 每天处理12亿次请求, 访问270亿内容
- YouTube: 每天观看的视频数 >20亿次
- Flickr: >60亿照片
- 截止2016/08/18, **微信和WeChat合并月活跃用户数达8.06亿，** 同比增长34％

数据量有多大呢？

- 现代分布式系统使用海量的数据:
  - ‘Avatar’电影渲染数据 >1 PB的存储空间
  - eBay 包括 >6.5PB的用户数据
  - Google早在2008年每天就是产生20PB的数据
  - Google 现在已经着手设计 1 EB的存储系统
  - NSA Utah数据中心据说有5ZB的存储系统 (!)

National Security Agency
美国国家安全局

- 1ZB的数据到底有多少？
  - 1,000,000,000,000,000,000,000 bytes (21个零)
  - 如果用1TB的硬盘累计起来有25,400 km 的高度
    - 上海到纽约的距离是多少？**14500 km**

计算能力的情况是怎样？

- 一台机器不可能处理那么多的数据
  - 那么就利用多个计算机!
- 现代分布式系统需要多少台计算机？
  - Facebook：> 60,000 服务器
  - Akamai 在71个国家有95,000 服务器
  - Intel 在97个数据中心有 ~100,000 台服务器
  - Microsoft 在2008年有至少 200,000 服务器
  - Google 据说有 >1百万台服务器, 目前规划1千万台服务器

你来构建下一代的Google和微信系统

- 你该怎么构建下一代的Google呢？
  - ... 如何下载和存储10亿次Web页面和图片?
  - ... 如何快速找到包括输入关键字(例如: 上海 同济大学)的Web页面?
  - ... 如何找到一个给定查询最相关的页面?
  - ... 如何每天响应12亿次的查询?
- 你该怎么构建下一代的WebChat微信平台呢？
  - ... 如何存储5亿个用户的profile (画像)数据？
  - ... 如何确保所有的profile数据都没有丢失呢?
  - ... 如何找到你潜在的朋友呢?

内容概述

- 计算规模导致的问题
  - 可扩展性的必要性及现有情况The need for scalability; scale of current services
  - Scale-up: 向上扩展：从PC服务器到“数据中心” Scaling up: From PCs to data centers
  - 经典扩展性方法的困境 Problems with 'classical' scaling techniques
- 云计算 Cloud computing
  - 什么是云计算 What is cloud computing?
  - 云计算类型What kinds of clouds exist today?
  - 什么应用适合于云计算What kinds of applications run on the cloud?
  - Virtualization虚拟化：云计算的动力Virtualization: How clouds work 'under the hood‘
  - 云计算的难点Some cloud computing challenges

Scaling up 向上扩展

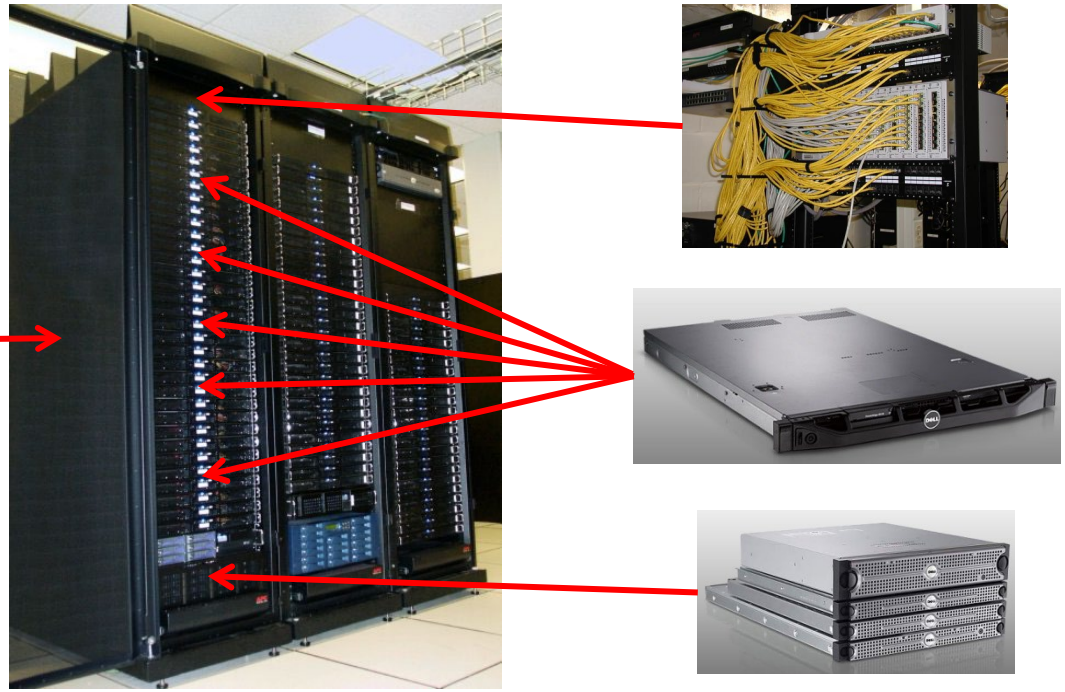
PC个人计算机

Server 服务器

Cluster 计算机集群

- 如果一台计算机的能力不够怎么办? What if one computer is not enough?
  - 买!买!买!一个更大的计算机 Buy a bigger (server-class) computer
- 如果世界上最大的服务器还不够? What if the biggest computer is not enough?
  - 买!买!买!一个更多多多的计算机Buy many computers

集群计算机Clusters



机架

网络交换机 switch
(计算节点之间的互联
及与其他机架的互联) Network
switch (connects nodes with
each other and with other racks)

节点/刀片nodes/blades
(通常是完全相同的配置)
Many nodes/blades
(often identical)

存储设备Storage device(s)

- 集群计算机的特征: Characteristics of a cluster:
  - 多个很类似的机器互联，且位于同一个物理空间(数据中心) Many similar machines, close interconnection (same room?)
  - 标准化的硬件 (机架、刀片) Often special, standardized hardware (racks, blades)
  - 通常是一个组织/公司拥有和使用Usually owned and used by a single organization

用电及制冷Power and cooling

- 集群计算机需要很多的用电Clusters need lots of power
  - 举例: 140 瓦/计算机 Example: 140 Watts per server
  - 考虑32台服务器的机架: 4.5千瓦(专用的用电系统!) Rack with 32 servers: 4.5kW (needs special power supply!)
  - 大多数用电直接转化为热Most of this power is converted into heat

- 大型集群系统需要制冷系统! Large clusters need massive cooling
  - 4.5千瓦等同于3个加热器: 4.5kW is about 3 space heaters
  - 那仅仅是一个机架! And that's just one rack!

Scaling up 向上扩展

PC个人
计算机

Server
服务器

Cluster
计算机集群

Data
Center
计算中心

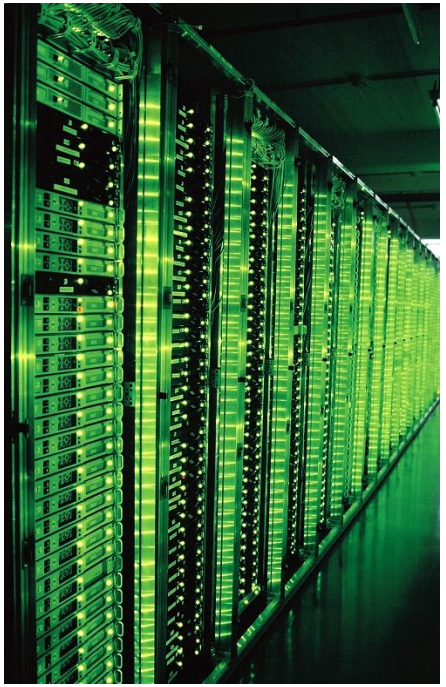
- 如果集群计算机太大了,难以使用普通的办公环境(用电和制冷问题)? What if your cluster is too big (hot, power hungry) to fit into your office building?
  - 构建一个专门用于集群计算机的建筑物 Build a separate building for the cluster
  - 有更多的制冷和供电系统Building can have lots of cooling and power
  - 数据中心 Result: Data center

真实的数据中心长什么呢？ What does a data center look like?

Google在美国Dalles, Oregon的数据中心

- 可以想象成一个有足球场大小的计算机？ A warehouse-sized computer
  - 一个数据中心可以轻松的容纳10,000机架、每个机架有100个核，总计1,000,000核 A single data center can easily contain 10,000 racks with 100 cores in each rack (1,000,000 cores total)

数据中心里面长什么样？
What's in a data center?



- 成千上万个机架Hundreds or thousands of racks

数据中心里面长什么样？ What's in a data center?

- 网络设备Massive networking

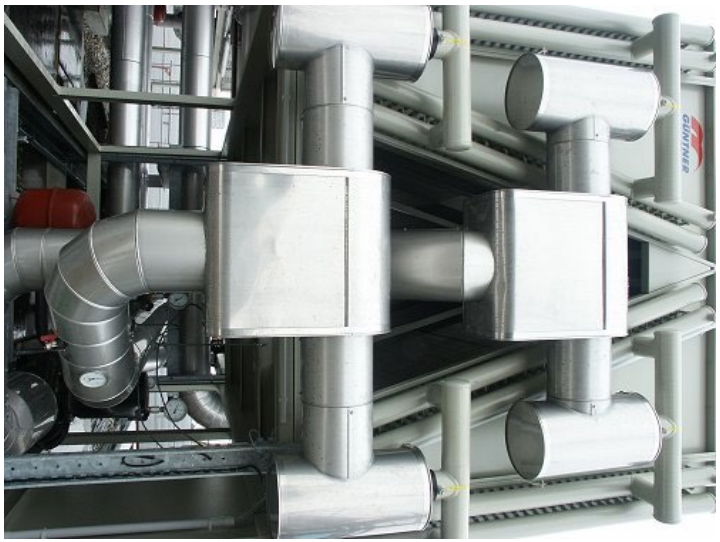
数据中心里面长什么样？ What's in a data center?



Source: 1&1

- 紧急备用电源Emergency power supplies

数据中心里面长什么样？What's in a data center?



- 大型制冷系统Massive cooling

用电及成本是关键 Energy matters!

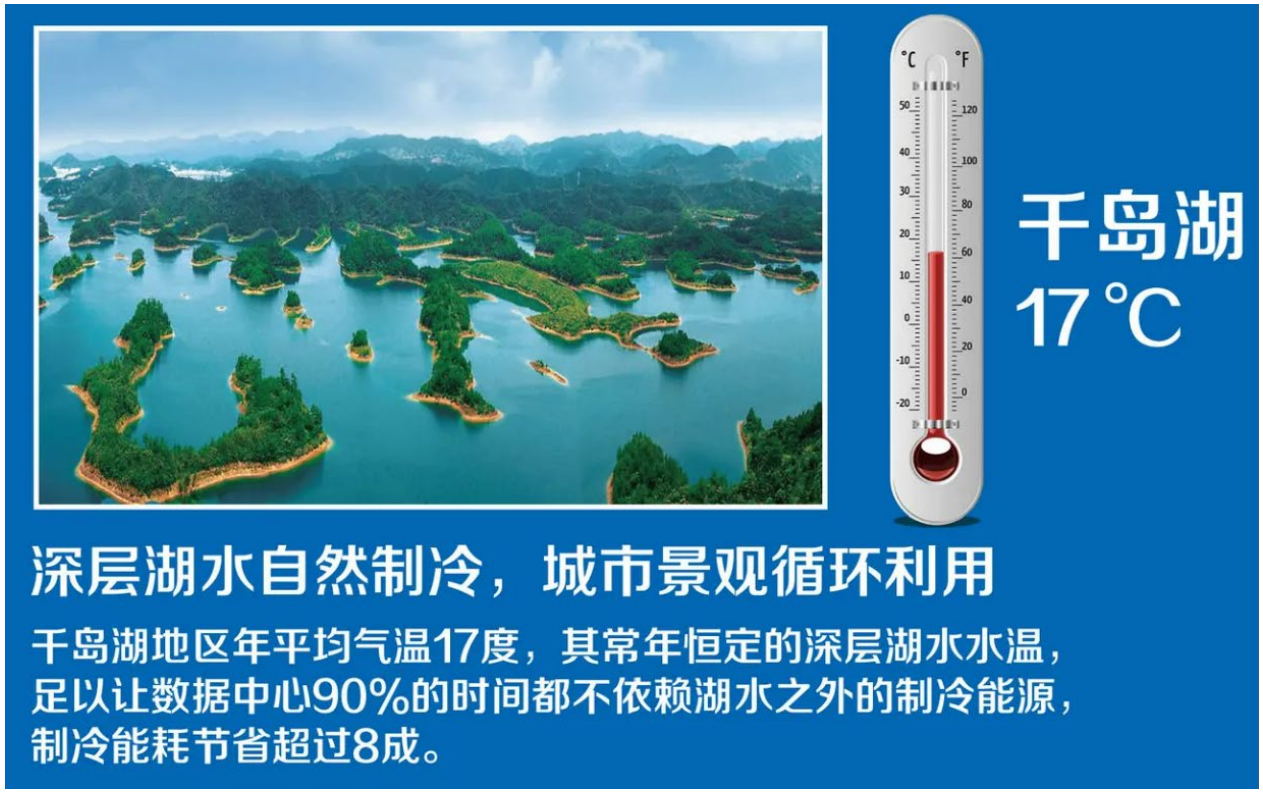
公司	服务器数量	用电	成本
eBay	16K	~$0.6*10^5$ MWh	~$3.7M/yr
Akamai	40K	~$1.7*10^5$ MWh	~$10M/yr
Rackspace	50K	~$2*10^5$ MWh	~$12M/yr
Microsoft	>200K	>$6*10^5$ MWh	>$36M/yr
Google	>500K	>$6.3*10^5$ MWh	>$38M/yr
USA (2006)	10.9M	$610*10^5$ MWh	$4.5B/yr

Source: Qureshi et al., SIGCOMM 2009

- 数据中心用电量非常大Data centers consume a lot of energy
  - 省钱之道: 把数据中心建在电厂附件Makes sense to build them near sources of cheap electricity
  - 举例:千瓦电价: 3.6ct 位于Idaho (电厂附近), 10ct 位于California (远距离输电), 18ct位于Hawaii (必须用船运输) Example: Price per KWh is 3.6ct in Idaho (near hydroelectric power), 10ct in California (long distance transmission), 18ct in Hawaii (must ship fuel)
  - 电→热, 制冷是个大问题Most of this is converted into heat → Cooling is a big issue!

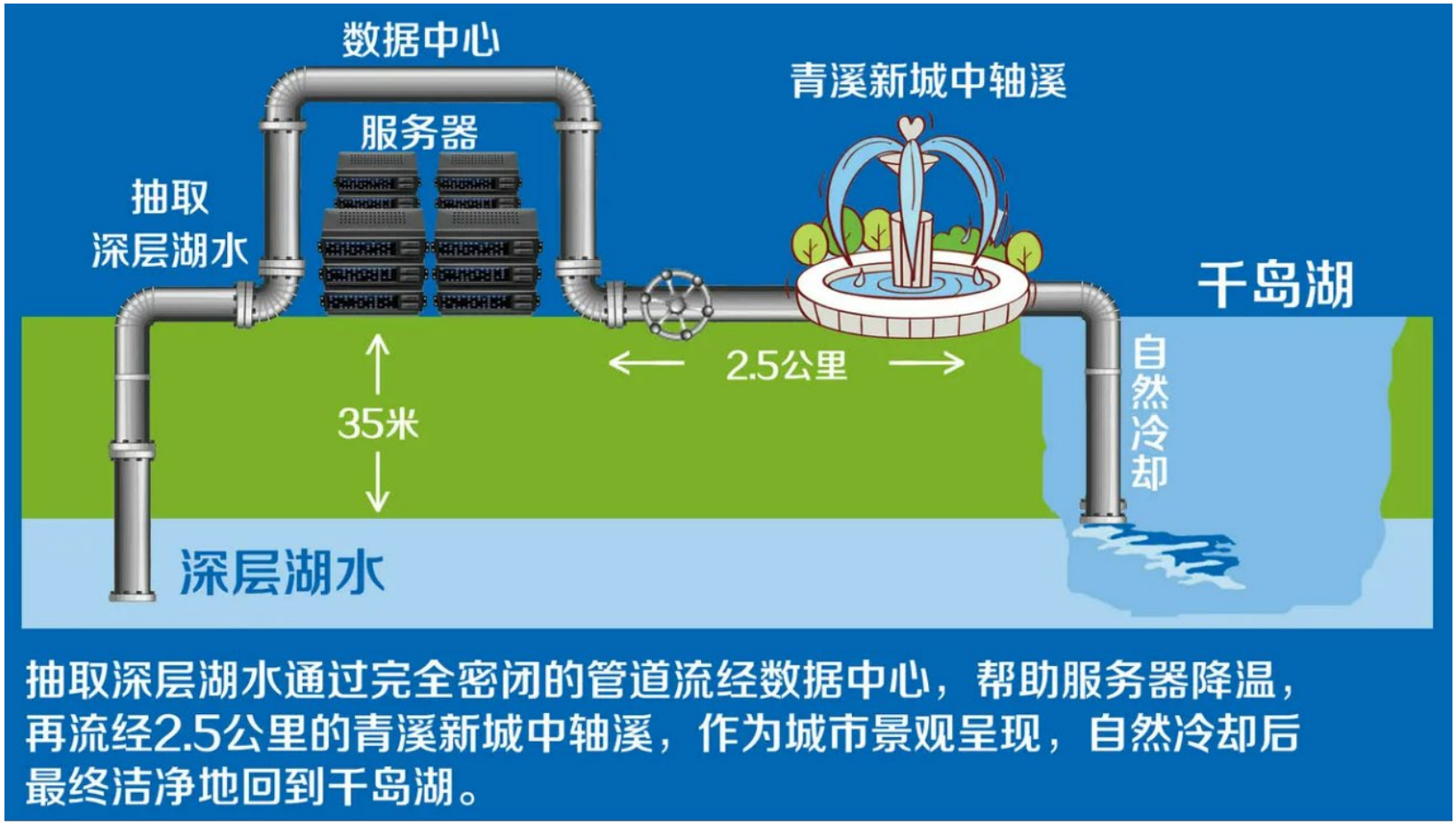
阿里巴巴千岛湖数据中心

- 早在2015年该数据中心投入使用



https://zhuanlan.zhihu.com/p/20215874

阿里巴巴千岛湖数据中心



阿里巴巴千岛湖数据中心

冷机房▲

▲屋顶冷却塔

▲屋顶光伏太阳能

与 240V 直流电源和柴油发电机互备冗余，保证供电稳定可靠

阿里巴巴千岛湖数据中心

◀数据中心微模块(ADM)

铝合金预制框架

精密的契合结构

整机柜服务器(AliRack)▶

服务器上架密度和传统

机柜相比提升了30%

服务器空间硬盘容量增加一倍

◀PCIE固态硬盘(AliFlash)

吞吐量、IOPS提升5-10倍

延迟下降70%以上

广泛使用
阿里巴巴
定制硬件

阿里巴巴千岛湖数据中心

- 全景

阿里巴巴千岛湖数据中心，以低碳、绿色、节能、环保、生态为主题，是浙江省内单体建设规模最大，新技术应用最多的数据中心。这个数据中心是全球最节水的数据中心之一，预计年均PUE可达到1.3,是亚热带地区最节能的数据中心之一; 年平均WUE 0.197。

https://developer.aliyun.com/article/124793

阿里巴巴千岛湖数据中心

- 冷机房

阿里巴巴千岛湖数据中心是国内首个采用自然水制冷技术的数据中心，空调系统采用两路进水，湖水和冷冻水，可以实现同时或单独运行。湖水经过物理净化后，通过密闭管道流经每层为服务器降温，之后直接供市政景观用水，实现了资源最佳利用。

阿里巴巴千岛湖数据中心

- 柴油发电机房

阿里巴巴千岛湖数据中心配备多路市政供电，可实现主备自动切换。此外，数据中心还配置多台大型高压柴油发电机组作为应急备用，多重供电保障能够确保数据中心持续可用。

阿里巴巴千岛湖数据中心

- IT单元模块

千岛湖数据中心融入了阿里巴巴大量新的设计理念，模块设计便是其中之一。IT模块与建筑、变压器系统、制冷系统一一对应，但又分别部署，实现就近引电。网络出口多物理链路与其他数据中心互联互通，互为主备，能够实现毫秒级的在线自动切换，用户完全无感知。

阿里巴巴千岛湖数据中心

- 电力电池室

千岛湖数据中心采用一路市电 一路240V直流的供电方式，综合供电效率达到97%以上，供电可靠性接近Tier 4最高等级。为了节约电缆的使用量，减少电能损失，千岛湖数据中心的供电系统，直接深入IT模块，分散式供电，即每个IT模块与变压器系统一一对应，从相邻楼层就近取电。

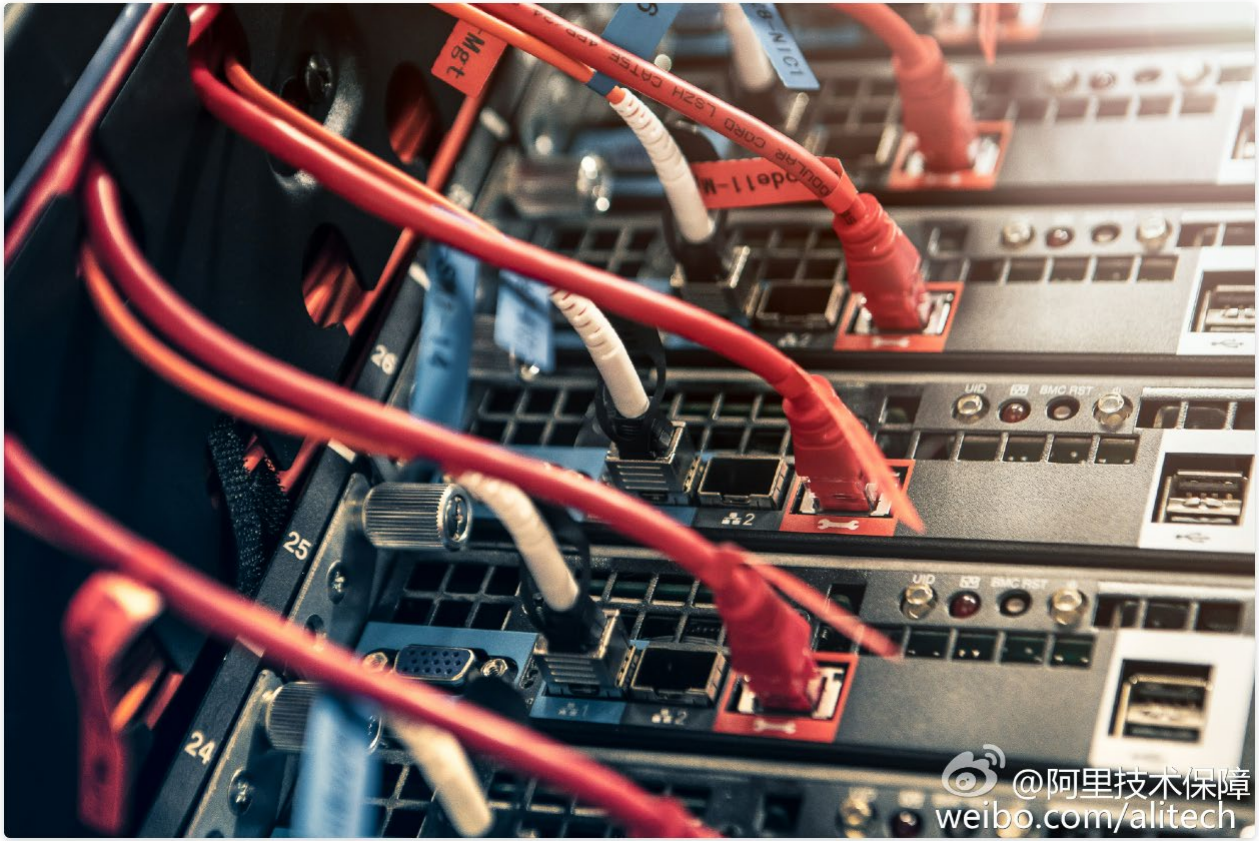
阿里巴巴千岛湖数据中心

- 阿里巴巴微模块ADM

千岛湖数据中心机房内设计部署了多组阿里巴巴自研定制的微模块ADM (Alibaba Data Center Module)，可以实现更快的交付效率，电力系统的效率高达97%，制冷系统节能20%以上。

阿里巴巴千岛湖数据中心

- AliRack 阿里整机柜服务器



AliRack是阿里自研的整机柜服务器，专门针对云计算和大数据的业务需求定制，单日物理部署可达5000台，服务器上架密度提升了30%，集成电源与散热系统能减少10%的能耗。

阿里巴巴千岛湖数据中心

- 屋顶太阳能

阿里巴巴千岛湖数据中心的屋顶部署了光伏太阳能发电系统，能够直接为服务器和网络设备供电，与240V直流电源和柴油发电互备冗余，保证IT设备供电稳定可靠。

阿里巴巴千岛湖数据中心

- 屋顶冷却塔

数据中心屋顶的冷却塔系统确保在温度比较低的季节或湖水不可用时，可以使用自然冷技术为服务器制冷。

Scaling up 向上扩展

PC个人
计算机

Server
服务器

Cluster
计算机集群

Data Center
计算中心

Networks of
Data Center
计算中心网络

- 如果一个数据中心还不够，该怎么办？What if even a data center is not big enough?
  - 那就建更多的数据中心…Build additional data centers
  - 在哪里建？要建多少？Where? How many?

Google的数据中心网络

http://www.google.com/about/datacenters/inside/locations/index.html

美洲地区

南卡罗来纳州伯克利县
爱荷华州康瑟尔布拉夫斯
佐治亚州道格拉斯县
亚拉巴马州杰克逊县
北卡罗莱纳州勒努瓦
俄克拉何马州梅斯县
田纳西州蒙哥马利县
智利基利库拉
俄勒冈州达尔斯

亚洲

台湾彰化县
新加坡

欧洲

爱尔兰都柏林
荷兰埃姆斯港
芬兰哈米纳
比利时圣吉斯兰

- 数据中心分布在世界各地Data centers are often globally distributed
- Why?
  - 靠近用户的物理所在地Need to be close to users (physics!)
  - 资源廉价(例如: 用电) Cheaper resources
  - 故障保护Protection against failures
  - 政策Policy…

阿里云全球基础设施

- 全球 30个地理区域内运营着89个可用区，未来还有更多。

华东 1、
华东 2、
华北 1 、
华北 2、
华北3、
华北5、
华南 1

提供多线 BGP 骨干网线路网络能力覆盖中国内地，提供高速稳定的中国内地访问

现有趋势: 模块化数据中心
Trend: Modular data center

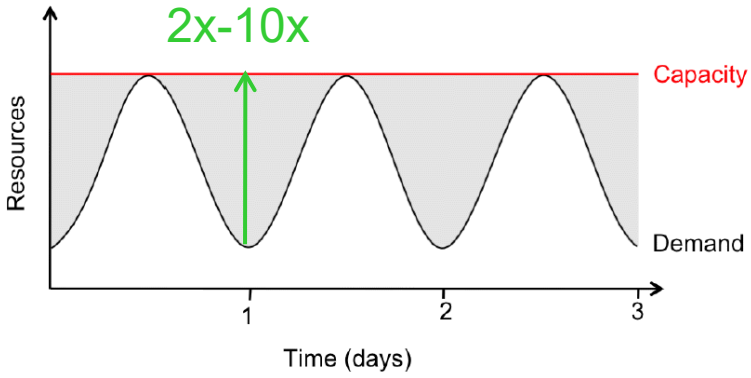
- 需要更多计算能力? Need more capacity?
- 再多部署一个黑盒子吧! Just deploy another container!
- 结果呢？ Consequence…

内容概述

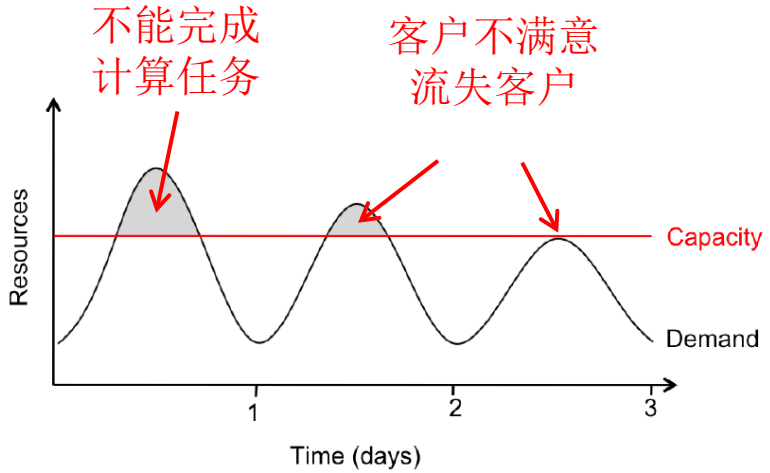
- **计算规模导致的问题**
  - 可扩展性的必要性及现有情况 
  - Scale-up: 向上扩展：从PC服务器到“数据中心” 
  - 经典扩展性方法的困境 

- **云计算**
  - 什么是效用计算、云计算
  - 云计算类型
  - 什么应用适合于云计算
  - Virtualization虚拟化：云计算的动力
  - 云计算的难点

困境#1: 惑 Problem #1: Difficult to dimension



根据负载峰值进行计算能力供应
Provisioning for the peak load



负载峰值之下进行计算能力供应
Provisioning below the peak

- 困境: 计算负载变化很大 Problem: Load can vary considerably
  - **峰值负载** 操过**均值**达2-10倍Peak load can exceed average load by factor 2x-10x [Why?]
  - 但是: 没有人以少于峰值设计服务器/数据中心/网络的计算能力But: Few users deliberately provision for less than the peak
  - 结果是: 服务器的利用率仅仅是 ~5%-0%!! Result: Server utilization in existing data centers ~5%-20%!!
  - 资源浪费、流失客户! Dilemma: Waste resources or lose customers!

困境 #2：贵 Problem #2: Expensive

- 硬件投资成本高, 需要很多很多$$$ Need to invest many $$$ in hardware
  - 小集群就高达$100,000: Even a small cluster can easily cost $100,000
  - 微软最近投资的一个数据中心费用达 $499 Microsoft recently invested $499 million in a single data center
- 专业知识Need expertise
  - 规划建设一个大型的数据中心非常的费心费力费钱 Planning and setting up a large cluster is highly nontrivial
  - 集群管理需要专业的软件等等 Cluster may require special software, etc.
- 维护成本Need maintenance
  - 需要系统管理员来进行更新硬件、安装和升级软件、维护用户账户、日志/故障分析…Someone needs to replace faulty hardware, install software upgrades, maintain user accounts, ...

困境 #2: 难上难下Problem #3: Difficult to scale

- 向上扩展：难 Scaling up is difficult
  - 订购更多的新机器、安装、(与现有集群)集成，往往花费好几周的时间Need to order new machines, install them, integrate with existing cluster - can take weeks
  - 大规模的扩展，往往需要重新设计，涉及到全新的存储系统、网络构建，甚至新的设备大楼 Large scaling factors may require major redesign, e.g., new storage system, new interconnect, new building (!)
- 向下扩展：难 Scaling down is difficult
  - 过剩的硬件怎么处理? What to do with superfluous hardware?
  - 服务器闲置状态的功耗是峰值情况下的60% → 即使系统没有任何计算任务，还是会消电 Server idle power is about 60% of peak → Energy is consumed even when no work is being done
  - 还有其他固定成本， 比如重建费用 Many fixed costs, such as construction

小结: 大规模分布式计算 Recap: Computing at scale

- 现代的计算机应用需要大量的处理能力和数据 Modern applications require huge amounts of processing and data
  - 往往涉及到PB级数据、数百万用户和10亿的对象 Measured in petabytes, millions of users, billions of objects
  - 需要专业硬件、算法和工具进行处理 Need special hardware, algorithms, tools to work at this scale
- 集群和数据中心可以提供所需的计算资源 Clusters and data centers can provide the resources we need
  - 主要区别: 规模(房间大小 vs. 足球场大小) Main difference: Scale (room-sized vs. building-sized)
  - 专用硬件: 电力设施和制冷系统亦是重要问题 Special hardware; power and cooling are big concerns
- 集群和数据中心非最佳方案 Clusters and data centers are not perfect
  - 惑、贵、难 Difficult to dimension; expensive; difficult to scale

内容概述

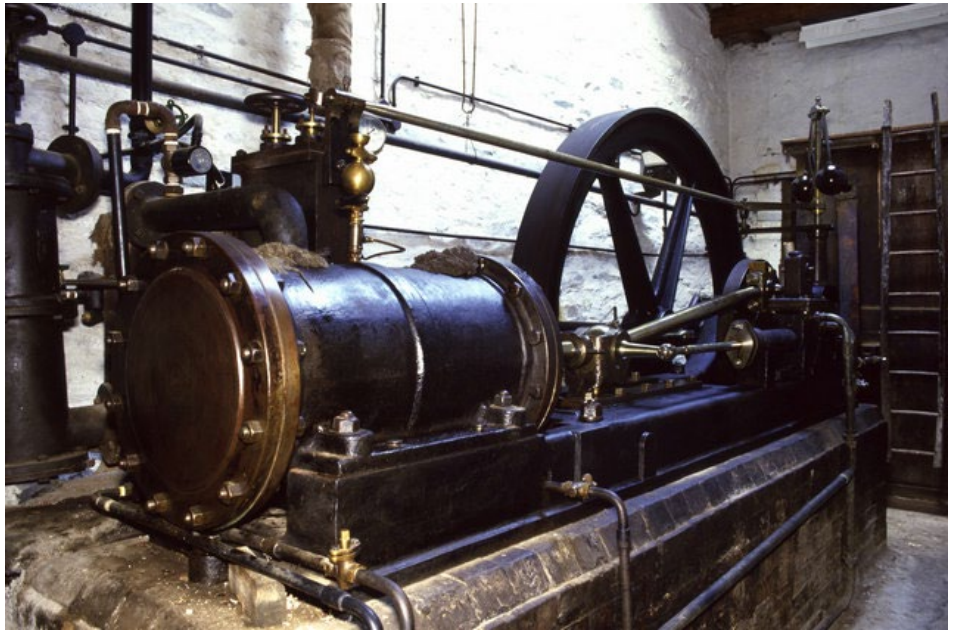
- **计算规模导致的问题**
  - 可扩展性的必要性及现有情况
  - Scale-up: 向上扩展：从PC服务器到“数据中心”
  - 经典扩展性方法的困境

- **云计算**
  - 什么是云计算
  - 云计算类型
  - 什么应用适合于云计算
  - Virtualization虚拟化：云计算的动力
  - 云计算的难点

NEXT

电厂

Waterwheel at the Neuhausen ob Eck Open-Air Museum



Steam engine at Stott Park Bobbin Mill

- 任何单位均有对应的供电源
  - 与集群计算类似：需要持续的升级改造,专业人员进行维护，难以向上/向下扩展..

电厂的升级The power plant analogy




- 趋势: 大型集中式电厂, 提供更强的发电能力... It used to be that everyone had their own power source
  - Challenges are similar to the cluster: Needs large up-front investment, expertise to operate, difficult to scale up/down...

用电计量模型Metered usage model

- 电网将各电厂连接起来，为用户进行输电Power plants are connected to customers by a network
- 电表记录用户的实际用电，用户则根据电表记录的用电进行付费 Usage is metered, and everyone (basically) pays only for what they actually use

供电模型与计算模型Why is this a good thing?

	供电模型Electricity	计算平台Computing
规模经济 Economies of scale	单一大型电厂的运行效费比远比多个小厂来得更高Cheaper to run one big power plant than many small ones	一个大型数据中心的运行效费比远比多个小型数据中心来得高！Cheaper to run one big data center than many small ones
多路复用Statistical multiplexing	高利用率! High utilization!	高利用率High utilization!
投资成本No up-front commitment	无需投资电厂费用 现收现付模型No investment in generator; pay-as-you-go model	用户无需数据中心的建设费用 现收现付计费模型No investment in data center; pay-as-you-go model
可扩展性 Scalability	按需供电(>千瓦); 秒级响应 Thousands of kilowatts available on demand; add more within seconds	按需计算；秒级响应Thousands of computers available on demand; add more within seconds

什么是云计算What is cloud computing?

The interesting thing about Cloud Computing is that we've redefined Cloud Computing to include everything that we already do.... I don't understand what we would do differently in the light of Cloud Computing other than change the wording of some of our ads.

Larry Ellison, quoted in the Wall Street Journal, September 26, 2008

A lot of people are jumping on the [cloud] bandwagon, but I have not heard two people say the same thing about it. There are multiple definitions out there of "the cloud".

Andy Isherwood, quoted in ZDnet News, December 11, 2008

到底什么是云计算？So what is it, really?

- 根据美国国家标准与技术协会(NIST):

  Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

- 基本特征Essential characteristics:
  - On-demand self service 按需的自服务机制
  - Broad network access 宽带网络访问
  - Resource pooling资源池(如: 网络, 服务器, 存储, 应用和服务等)
  - Rapid elasticity 快速回弹性
  - Measured service可度量的服务

其他常用术语 Other terms you may have heard

- Web 万维网
  - Internet因特网的信息共享模型 The Internet's information sharing model
  - 部分的web services运行在云上，但非全部 Some web services run on clouds, but not all

- Internet因特网
  - 用于连接网络的网络A network of networks.
  - 用于Web万维网; 并将多数云连接到客户 Used by the web; connects (most) clouds to their customers

内容概述

- **计算规模导致的问题**
  - 可扩展性的必要性及现有情况
  - Scale-up: 向上扩展：从PC服务器到“数据中心”
  - 经典扩展性方法的困境

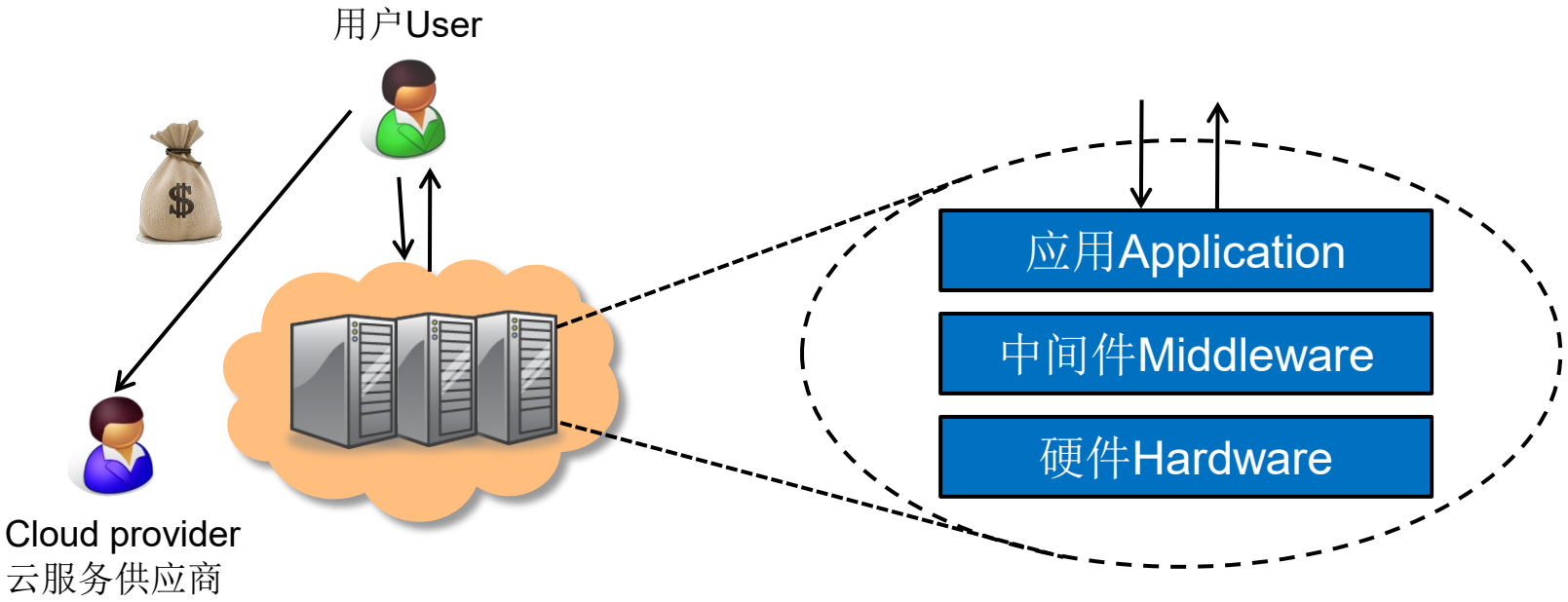
- **云计算**
  - 什么是云计算
  - 云计算类型
  - 什么应用适合于云计算
  - Virtualization虚拟化：云计算的动力
  - 云计算的难点

NEXT

云世界里一切皆为服务

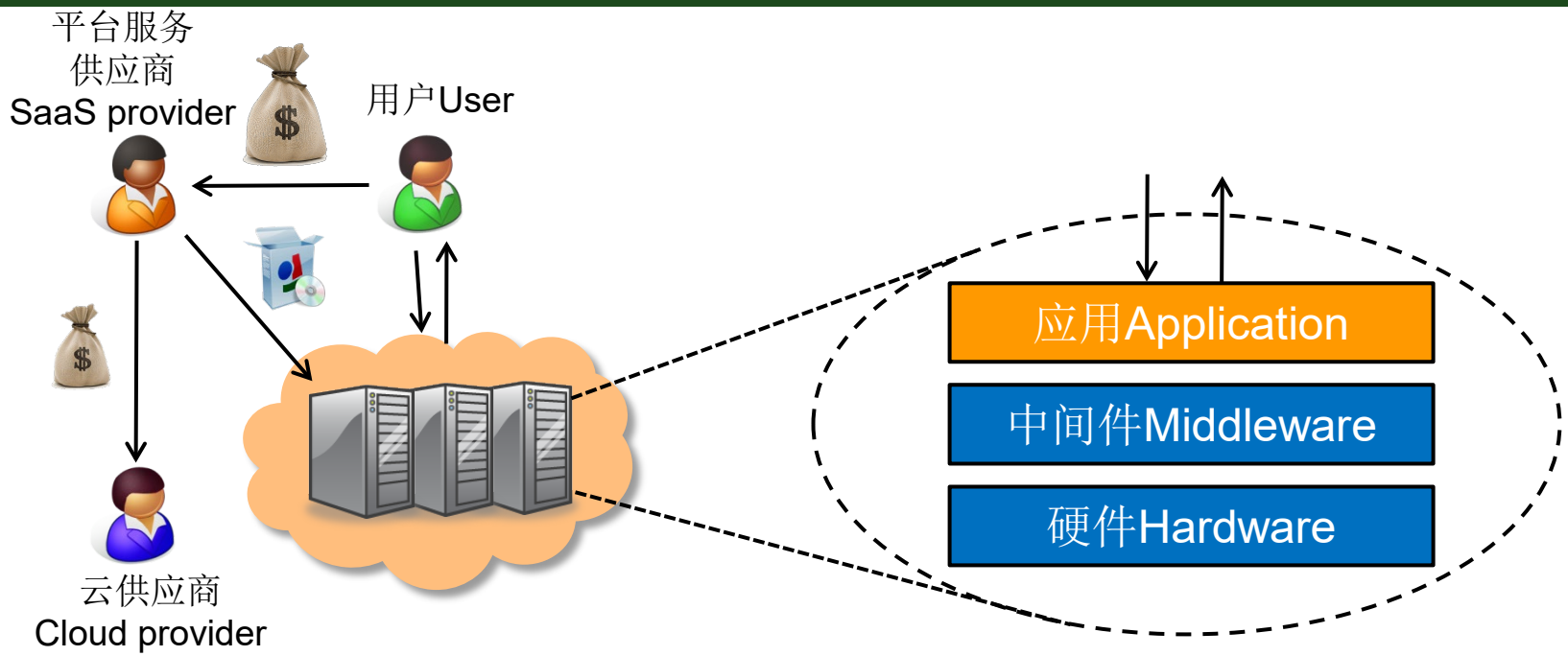
- 云可以提供什么类型的服务? What kind of service does the cloud provide?
  - 是完整的应用还是仅仅特定的资源? Does it offer an entire application, or just resources?
  - 如果仅仅是资源，是什么类型/级别的抽象? If resources, what kind / level of abstraction?
- 三种类型的云服务: Three types commonly distinguished:
  - Software as a service (SaaS) 软件服务
    - 餐馆Restaurant. Prepares & serves entire meal, does the dishes, ...
  - Platform as a service (PaaS)平台服务
    - 外卖Take-out food. Prepares meal, but does not serve it.
  - Infrastructure as a service (IaaS)基础设施服务
    - 食品杂货店Grocery store. Provides raw ingredients.
  - XaaS: 尽管有其他类型的服务，但均不常用
    - 前端桌面应用Desktop 、后端系统Backend 、通信Communication 、网络、监控...

软件服务Software as a Service (SaaS)



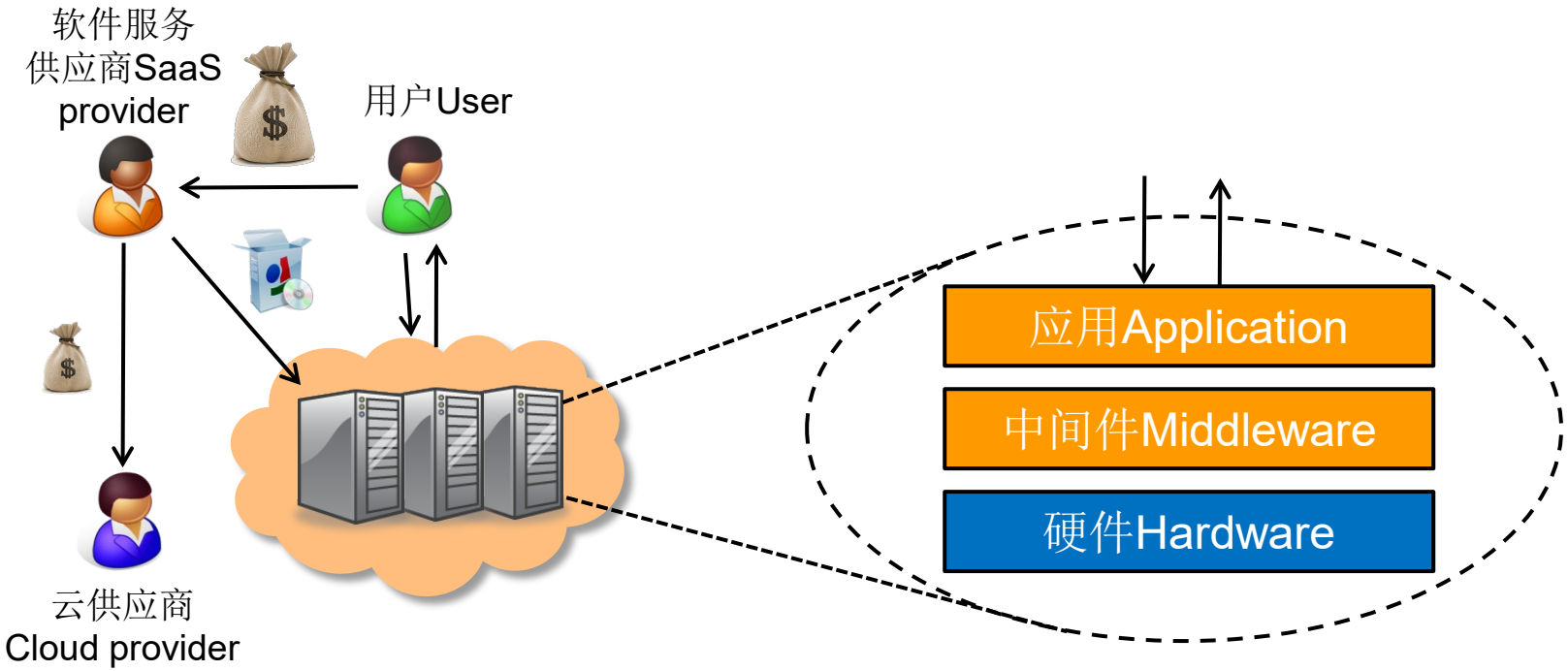
- 云端提供完整的应用解决方案Cloud provides an entire application
  - Word、Excel、CRM、日历等软件应用...
  - 用户付费给云端服务供应商Customer pays cloud provider
  - 例如: Google Apps, Salesforce.com

平台服务Platform as a Service (PaaS)



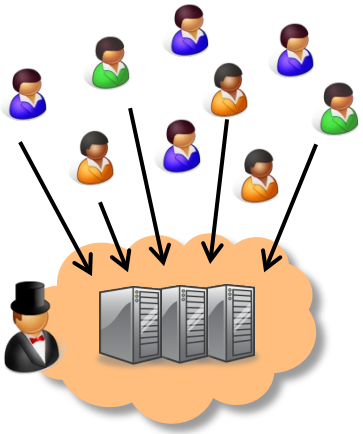
- 云端提供中间件/基础设施服务Cloud provides middleware/infrastructure
  - 中间件/基础设施服务举例 Microsoft Common Language Runtime (CLR)
  - 客户付费给SaaS供应商; SaaS供应商付费给云端供应商用于支付基础设设施的费用Customer pays SaaS provider for the service; SaaS provider pays the cloud for the infrastructure
  - Example: Windows Azure, Google App Engine

基础设施服务Infrastructure as a Service (IaaS)

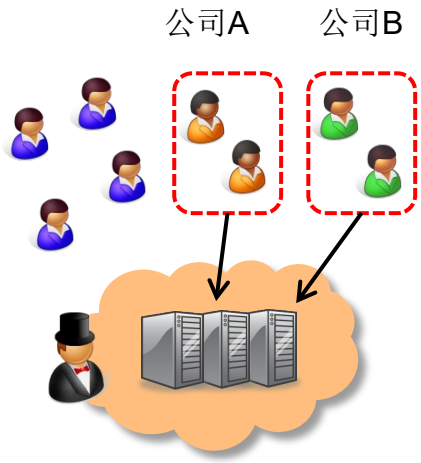


- Cloud云端提供最基础的计算资源Cloud provides raw computing resources
  - 虚拟机、刀片服务器、硬盘... Virtual machine, blade server, hard disk, ...
  - 用户付费给SaaS供应商以支付服务费用; SaaS供应商则付费给云端以支付资源使用的费用 Customer pays SaaS provider for the service; SaaS provider pays the cloud for the resources
  - Example: Amazon Web Services, Rackspace Cloud, GoGrid

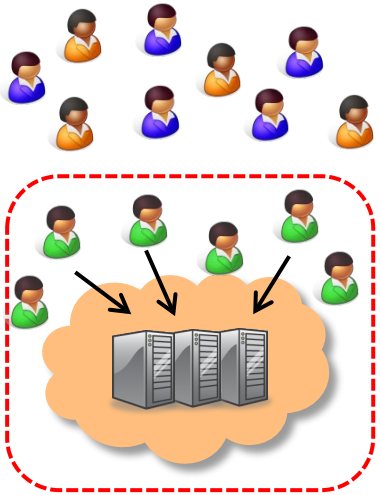
私有云、混合云和社区云
Private/hybrid/community clouds



共有云Public



社区云Community



私有云Private

- 谁可以成为云端客户? Who can become a customer of the cloud?
  - 共有云Public cloud：商业服务;几乎是对任何人开放Commercial service; open to (almost) anyone. 例如: Amazon AWS, Microsoft Azure, Google App Engine, 阿里云
  - 社区(行业)云Community cloud：多个类似组织共享的云Shared by several similar organizations.例如Google's "Gov Cloud"
  - 私有云Private cloud：仅仅在一个特定组织内部用户使用Shared within a single organization. 例如: 某大型公司的内部数据中心

本课程的关注内容 Focus of this class →

还可以称之为云么？ Is this a 'real' cloud? →

内容概述

- 计算规模导致的问题Computing at scale
  - 可扩展性的必要性及现有情况
  - Scale-up: 向上扩展：从PC服务器到“数据中心”
  - 经典扩展性方法的困境
- 云计算Cloud computing
  - 什么是云计算What is cloud computing?
  - 云计算类型What kinds of clouds exist today?
  - 什么应用适合于云计算What kinds of applications run on the cloud?
  - Virtualization虚拟化：云计算的动力How clouds work 'under the hood'
  - 云计算的难点Some cloud computing challenges

NEXT

云应用示例Examples of cloud applications

- 应用托管Application hosting
- 备份、存储Backup and Storage
- 内容分发Content delivery
- 电子商务E-commerce
- 高性能计算High-performance computing
- 搜索引擎Search engines
- Web托管Web hosting

http://aws.amazon.com/solutions/case-studies/washington-post/

案例研究Case study: The Washington Post

- 2008年3月19日: 希拉里·克林顿的官方行程安排正式对外公开 Hillary Clinton's official White House schedule released to the public
  - 17,481页的非搜索、低质量的PDF文件17,481 pages of non-searchable, low-quality PDF
  - 这些内容对记者而言非常有用, 但是需要几百人力小时的工作Very interesting to journalists, but would have required hundreds of man-hours to evaluate
  - 华盛顿邮报的高级工程师Peter Harkins: 可以做的更好、更快吗？ Peter Harkins, Senior Engineer at The Washington Post: Can we make that data available more quickly, ideally within the same news cycle?
  - 测试不同的Optical Character Recognition (OCR)程序；估计所需的速度和时间 Tested various Optical Character Recognition (OCR) programs; estimated required speed
  - 启动200EC2 instances; 整个项目仅仅在9个小时就完成、仅需要1,407小时的VM时间 ($144.62) Launched 200 EC2 instances; project was completed within nine hours (!) using 1,407 hours of VM time ($144.62)
  - 在希拉里·克林顿的官方行程安排正式对外公开之后的26小时系统就上线！ Results available on the web only 26 hours after the release

http://animoto.com/blog/company/amazon-com-ceo-jeff-bezos-on-animoto/

案例研究Case study:

- Animoto: 用户可以
  通过个人照片和音乐
  创建视频Animoto: Lets users
  create videos from
  their own photos/music
  - 自动编辑照片、对齐音乐,
  使得照片、音乐视觉效果更佳Auto-edits photo
  music, so it "looks good"

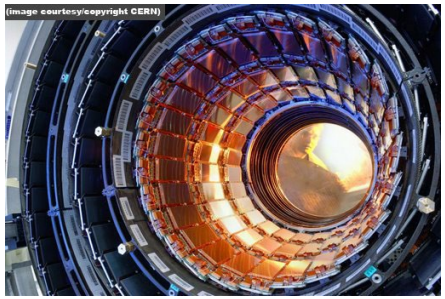
Source: Jeff Bezos' talk at Stanford on 4/19/08

- 使用Amazon EC2+S3+SQS
- 在2008年4月中旬公开发布Facebook应用Released a Facebook app in mid-April 2008
  - 在3天内多达750,000用户签到More than 750,000 people signed up within 3 days
  - EC2使用率从50台机器迅速上升至3,500 (x70增长!) EC2 usage went from 50 machines to 3,500 (x70 scalability!)

案例研究Other examples

- DreamWorks使用 Cerelink云服务进行动画电影的渲染DreamWorks is using the Cerelink cloud to render animation movies
  - *Shrek Forever After*
  - *How to Train your Dragon*

- CERN使用“science cloud”处理实验数据CERN is working on a "science cloud" to process experimental data



- Virgin atlantic 在Amazon AWS上托管新第一代的旅行门户网站Virgin atlantic is hosting their new travel portal on Amazon AWS

万事皆云？Is the cloud good for everything?

- 否！ No
- 医疗记录Example: Processing medical records
  - HIPAA (Health Insurance Portability and Accountability Act)隐私和安全规定 privacy and security rule
- 财务信息Processing financial information
  - Sarbanes-Oxley法令 Sarbanes-Oxley act
- 你愿意把个人医疗记录数据放到云端处理么？ Would you put your medical data on the cloud?
  - 会 /不会？ Why / why not?

小结：云应用Recap：Cloud applications

- 云的确对很多情况有用... Clouds are good for many things...
  - 消耗大量计算、存储和带宽的应用Applications that involve large amounts of computation, storage, bandwidth
  - 特别是需要快速提供大量计算资源的场景 (华盛顿邮报) 或者负载变化较大的情况(TicketLeap) Especially when lots of resources are needed quickly (Washington Post example) or load varies rapidly (TicketLeap example)
- ... 但很显然并非万能良药! but not for all things

内容概述

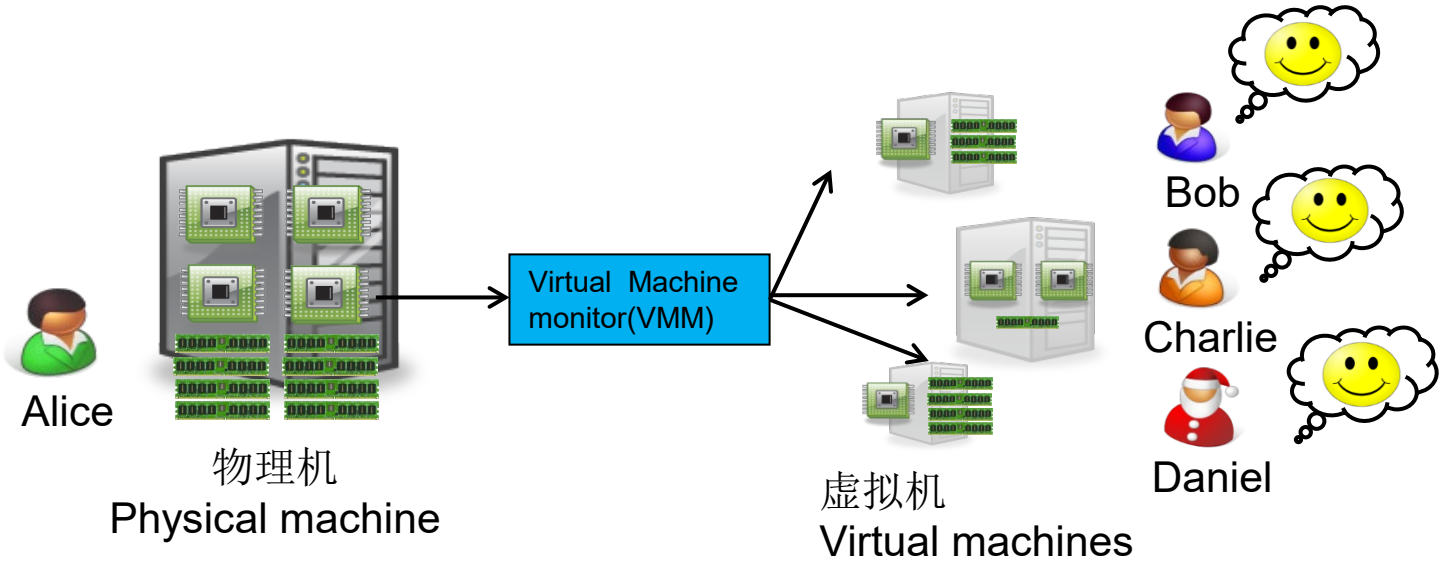
- **计算规模导致的问题**
  - 可扩展性的必要性及现有情况
  - Scale-up: 向上扩展：从PC服务器到“数据中心”
  - 经典扩展性方法的困境

- **云计算**
  - 什么是云计算
  - 云计算类型
  - 什么应用适合于云计算
  - Virtualization虚拟化：云计算的动力 NEXT
  - 云计算的难点

虚拟化是什么？What is virtualization?

- 假设Alice有一台4CPUs和8GB内存的物理机，同时还有3个用户: Suppose Alice has a machine with 4 CPUs and 8 GB of memory, and three customers:
  - 其中Bob希望有一台1 CPU和3GB内存的机器Bob wants a machine with 1 CPU and 3GB of memory
  - Charlie希望有一台2 CPUs和1GB内存的机器Charlie wants 2 CPUs and 1GB of memory
  - 而Daniel则希望有一台1 CPU和4GB内存的机器Daniel wants 1 CPU and 4GB of memory
- 那么，Alice该如何处理? What should Alice do?
  - 另外再购买3台物理机吗？

虚拟化是什么？What is virtualization?

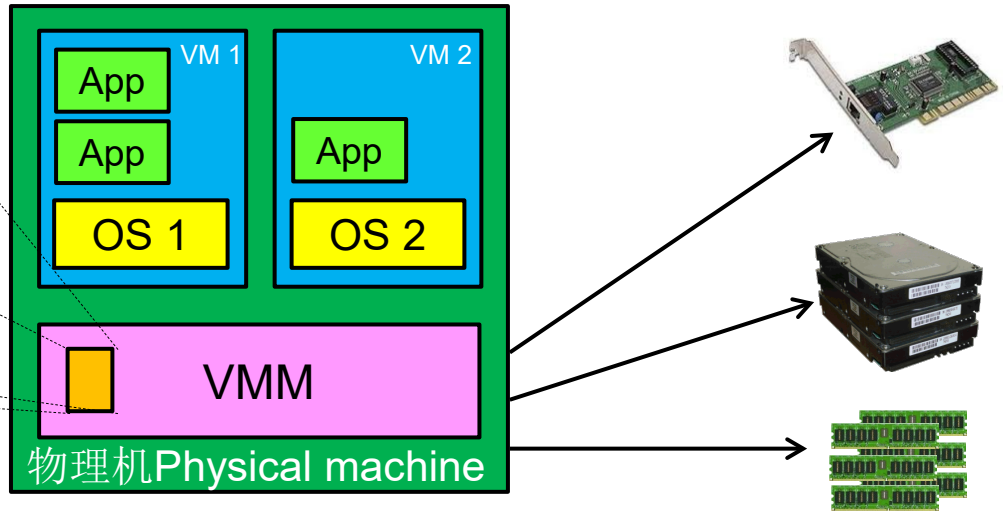


- Alice可以为每个用户出售virtual machine (VM)虚拟机以满足其各自的需求Alice can sell each customer a virtual machine (VM) with the requested resources
  - 从用户的角度上看，每个人均好像拥有一个独立的物理机一样(isolation隔离性) From each customer's perspective, it appears as if they had a physical machine all by themselves (isolation)

什么是虚拟化？How does it work?

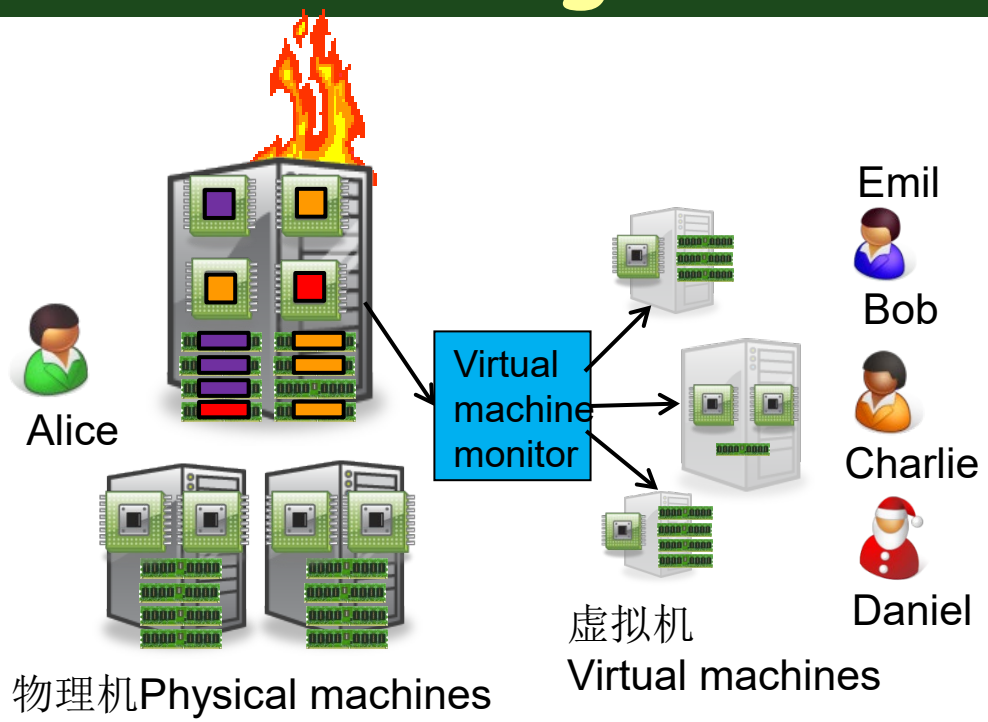
虚拟机	虚拟资源 Virtual	物理资源 Physical
1	0-99	0-99
1	299-399	100-199
2	0-99	300-399
2	200-299	500-599
2	600-699	400-499

转换表Translation table



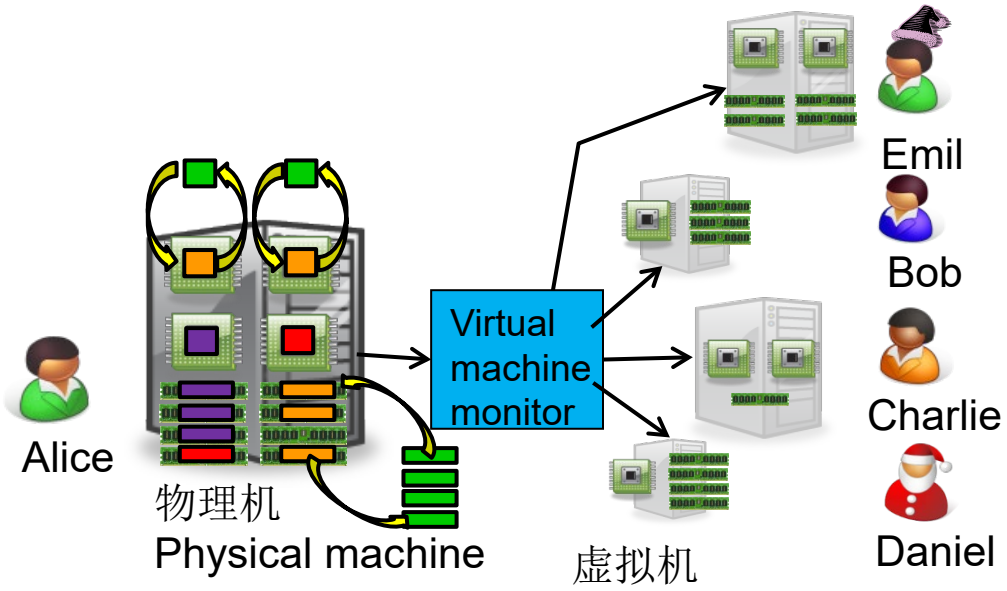
- 虚拟化资源 (CPU, memory...) Resources (CPU, memory...) are virtualized
  - VMM (“Hypervisor”) 通过转换表将请求的虚拟资源映射到物理资源 VMM ("Hypervisor") has translation tables that map requests for virtual resources to physical resources
  - 举例: VM 1 访问虚拟内存单元#323; VMM将其映射到物理内存单元123. Example: VM 1 accesses memory cell #323; VMM maps this to memory cell 123.
  - 如果OS内存有不同，VMMs将如何工作? How do VMMs differ from OS kernels?

优点：迁移Benefit：Migration



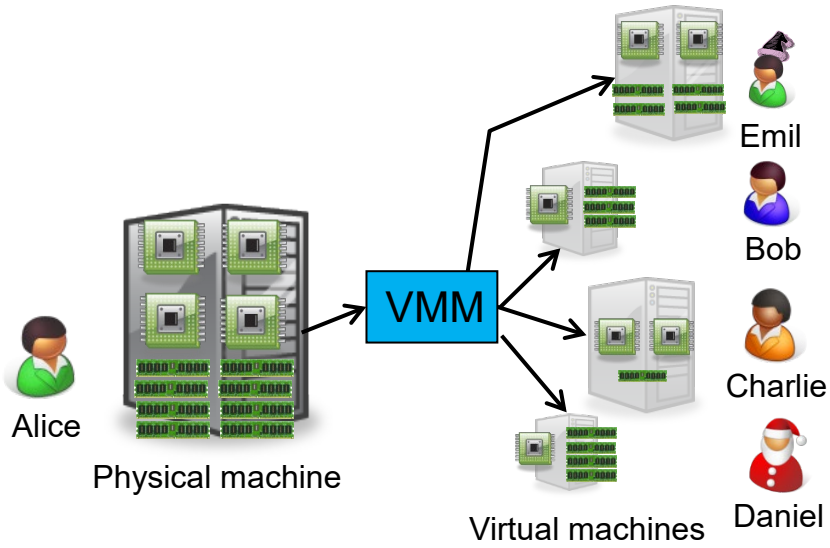
- 如何物理机关机，VMMs将如何工作？ What if the machine needs to be shut down?
  - 比如系统进行维护的情况... e.g., for maintenance, consolidation, ...
  - Alice可以将VMs迁移到其他的物理机之上，而客户则并未察觉到 Alice can migrate the VMs to different physical machines without any customers noticing

优点：分时共享Benefit: Time sharing



- 如何Alice有更多的客户该怎么办呢？What if Alice gets another customer?
  - 多个VMs可以分时共享现有的计算资源 Multiple VMs can time-share the existing resources
  - 结果: Alice有比物理虚拟机更多的虚拟CPUs和内存 (不过这些虚拟资源不能同时在线使用) Result: Alice has more virtual CPUs and virtual memory than physical resources (but not all can be active at the same time)

优点与不足并存：隔离性 Physical machine



- 优点: Emil 不能访问Charlie的数据 Good: Emil can't access Charlie's data
- 不足: 如果Emil的负载突然增加会怎样呢？ Bad: What if the load suddenly increases?
  - Emil和Charlie的VM共享CPUs, 但是Charlie突然启动了一个大计算量的任务 Example: Emil's VM shares CPUs with Charlie's VM, and Charlie suddenly starts a large compute job
  - Emil的VM性能因此下降 Emil's performance may decrease as a result
  - VMM则将Emil的软件移至其他的CPU，或者迁移至其他的物理机VMM can move Emil's software to a different CPU, or migrate it to a different machine

小结:云端虚拟化 Recap: Virtualization in the cloud

- 虚拟化给云服务商更大的灵活性Gives cloud provider a lot of flexibility
  - 提供不能能力的VMs: Can produce VMs with different capabilities
  - 根据情况迁移VMs (如：系统维护) Can migrate VMs if necessary (e.g., for maintenance)
  - 过度使用虚拟资源，则会增加系统负载 Can increase load by overcommitting resources
- 提供安全性和隔离性Provides security and isolation
  - 一个VM上的程序不能影响另外一个VM上的程序 Programs in one VM cannot influence programs in another
- 为用户提供方便Convenient for users
  - 完全控制虚拟硬件(可以按照自己的操作系统和应用程序...) Complete control over the virtual 'hardware' (can install own operating system own applications, ...)
- But: 难以预测性能But: Performance may be hard to predict
  - 同一台物理机的VMs上负载变化会影响其他VMs的性能 Load changes in other VMs on the same physical machine may affect the performance seen by the customer

内容概述

- **计算规模导致的问题**
  - 可扩展性的必要性及现有情况
  - Scale-up: 向上扩展：从PC服务器到“数据中心”
  - 经典扩展性方法的困境

- **云计算**
  - 什么是云计算
  - 云计算类型
  - 什么应用适合于云计算
  - Virtualization虚拟化：云计算的动力
  - 云计算的难点Some cloud computing challenges

NEXT

痛点 Obstacles

1. 可用性Availability
   - 如果云平台出现故障，上层应用会怎样？What happens to my business if there is an outage in the cloud?
2. 数据入闸Data lock-in
   - 如何将数据一个云平台移动到另外一个云平台? How do I move my data from one cloud to another?
3. 数据保密性和安全性Data confidentiality and auditability
   - 如何保证云平台不会泄露需要保密的数据？ How do I make sure that the cloud doesn't leak my confidential data?
   - 如何遵守HIPAA 和Sarbanes/Oxley法令? Can I comply with regulations like HIPAA and Sarbanes/Oxley?

云供应商 Service	持续时间 Duration	日期 Date
S3	6-8 hrs	7/20/08
AppEngine	5 hrs	6/17/08
Gmail	1.5 hrs	8/11/08
Azure	22 hrs	3/13/09
Intuit	36 hrs	6/16/10
EBS	>3 days	4/21/11
ECC	~2 hrs	6/30/12

云服务故障案例
Some prominent cloud outages

痛点 Obstacles

4. 数据转移瓶颈Data transfer bottlenecks
   - 如何将海量在云平台中进行转移? How do I copy large amounts of data from/to the cloud?
   - 10 TB数据从同济大学转移到复旦大学？ Example: 10 TB from Tongji U to Fudan U?
   - 可以利用AWS的导入/导出功能 Motivated Import/Export feature on AWS
5. 性能不可预测性Performance unpredictability
   - 举例: VMs的分时共享磁盘→ I/O 干扰 Example: VMs sharing the same disk → I/O interference
   - HPC计算任务需要资源调度和协调 Example: HPC tasks that require coordinated scheduling
6. 大型分布式系统的bugs (Bugs in large distributed systems)
   - 相当多的严重错误很难重现 Many errors cannot be reproduced in smaller configs

方法	时间
Internet (20Mbps)	45 days
FedEx	1 day

数据转移时间10TB

	平均性能	标准偏差
内存带宽	1.3GB/s	0.05GB/s (4%)
磁盘带宽	55MB/s	9MB/s (16%)

75台 EC2VMs的性能测试基准

内容概述

- **计算规模导致的问题**
  - 可扩展性的必要性及现有情况
  - Scale-up: 向上扩展：从PC服务器到“数据中心”
  - 经典扩展性方法的困境
- **云计算**
  - 什么是云计算
  - 云计算类型
  - 什么应用适合于云计算
  - Virtualization虚拟化：云计算的动力
  - 云计算的难点

敬请期待

下节课：

大规模编程、并发性和一致性