

Tổng quan về Điện toán Đám mây – Cloud Computing

- **GV: Đỗ Oanh Cường**
- **Bộ môn Hệ thống thông tin, Khoa CNTT.**
- **Email: cuongdo@tlu.edu.vn**
- **Website: www.cuongdo.info**

QUY ĐỊNH MÔN HỌC

- ĐỔI TÊN TRÊN ZOOM: TÊN LỚP _ TÊN SV
 - VÍ DỤ: 62THNB_NGUYỄN VĂN TOÀN

Đề cương môn học

- 1- **Tên môn học:** Nhập môn điện Toán Đám Mây
- 2- **Bộ môn phụ trách môn học:** Hệ thống thông tin
- 3- **Mã số môn học:** CSE393
- 4- **Số tín chỉ:** 3 tín chỉ (30 giờ lý thuyết + 15 giờ bài tập trên lớp)

Mô tả môn học

- Môn học cung cấp cho sinh viên kiến thức lý thuyết và thực tiễn về các chủ đề căn bản liên quan đến công nghệ điện toán đám mây. Giúp sinh viên tìm hiểu và phân biệt được các mô hình dịch vụ đám mây khác nhau (IaaS, PaaS, SaaS và BPaaS).
- Sinh viên cũng được giao một số chủ đề kĩ thuật nhỏ, chia nhóm tìm hiểu và thuyết trình trong về những chủ đề này, các chủ đề đòi hỏi sinh viên hiểu và biết cách vận dụng các kiến thức về lập trình trên máy tính cá nhân áp dụng vào đám mây ra sao.

Đánh giá môn học

TT	Các hình thức đánh giá	Trọng số
1	QT= Bài tập, chuyên cần, xây dựng bài, Kiểm tra (3 Bài tập mỗi BT 10%, Chuyên cần 15%, Xây dựng bài 15%, Kiểm tra giữa kỳ 40%)	0.4
2	THM=Thi hết môn (trắc nghiệm, 60 phút, 60 câu)	0.6
	Điểm môn học = ĐQT x 0.4 + THM x 0.6	

What is Cloud Computing?

Informal: computing with large datacenters

What is Cloud Computing?

~~Informal: computing with large datacenters~~

Our focus: **computing as a utility**

- » Outsourced to a third party or internal org

Types of Cloud Services

Infrastructure as a Service (IaaS): VMs, disks

Platform as a Service (PaaS): Web, MapReduce

Software as a Service (SaaS): Email, GitHub

Public vs private clouds:

Shared across arbitrary orgs/customers
vs internal to one organization

Câu hỏi tìm hiểu

- Tìm hiểu về VMWare: công ty? giải pháp?
- Các dịch vụ tương đương VMWare?
- Nêu các tên dịch vụ SaaS triển khai tại Việt Nam

Example

AWS Lambda functions-as-a-service

- » Runs functions in a Linux container on events
- » Used for web apps, IoT apps, stream processing, highly parallel MapReduce and video encoding



Câu hỏi tìm hiểu: Amazon Web Services (AWS)

- Khái niệm AWS
- Các sản phẩm dịch vụ
- Môi nhóm ghi 1 sản phẩm/ dịch vụ AWS mà em biết

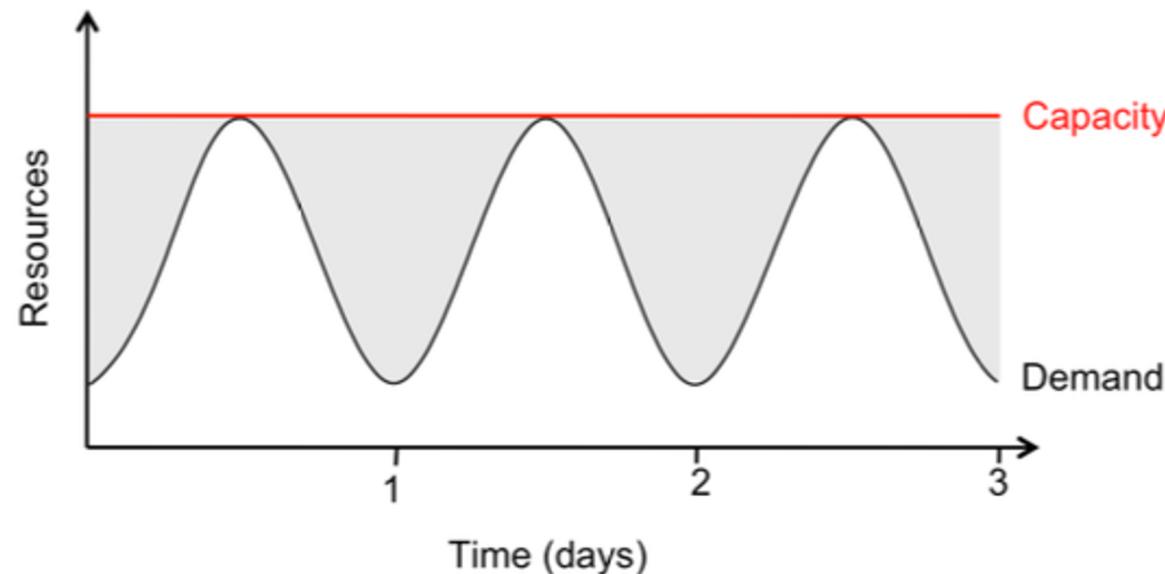
Câu hỏi thảo luận

- Tính chi phí để triển khai 1 server web – cấu hình ram 4G ssd 120G – chip 2 nhân trong 1 năm
- Tính chi phí triển khai server tương tự trên 1 hệ thống cloud mà em biết – mỗi nhóm tìm 1 hệ thống cloud

Cloud Economics: For Users

Pay-as-you-go (usage-based) pricing:

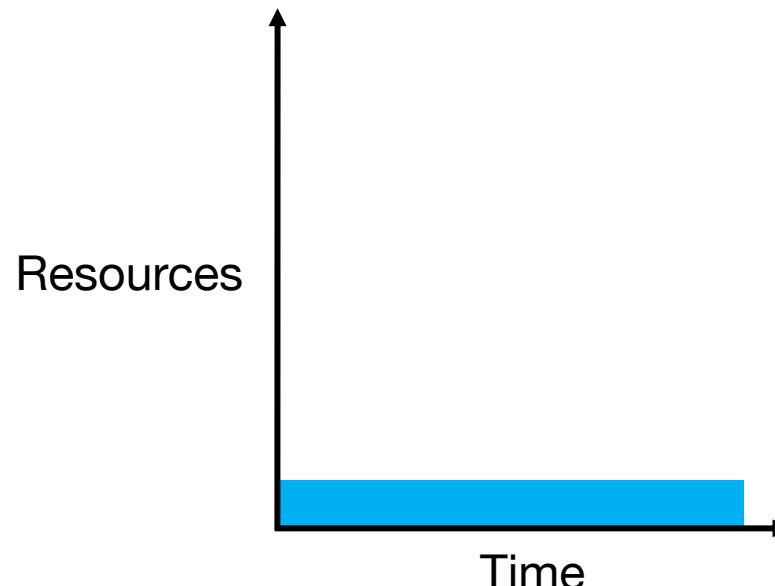
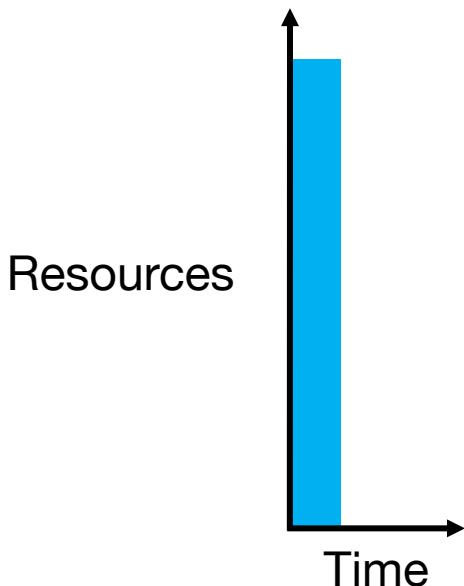
- » Most services charge per minute, per byte, etc
- » No minimum or up-front fee
- » Helpful when apps have *variable utilization*



Cloud Economics: For Users

Elasticity:

- » Using 1000 servers for 1 hour costs the same as 1 server for 1000 hours
- » Same price to get a result faster!



Cloud Economics: For Providers

Economies of scale:

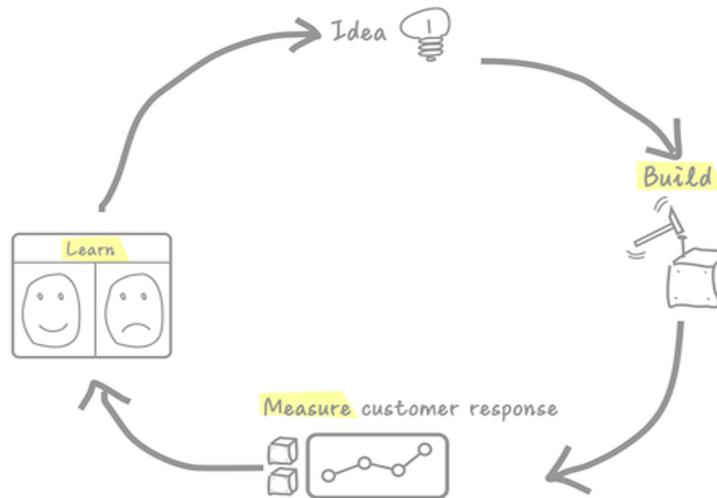
- » Purchasing, powering & managing machines at scale gives lower per-unit costs than customers'
- » Tradeoff: fast growth vs efficiency
- » Tradeoff: flexibility vs cost



Cloud Economics: For Providers

Speed of iteration:

- » Software as a service means fast time-to-market, updates, and detailed monitoring/feedback
- » Compare to speed of iteration with ordinary software distribution



Questions

- Nêu 5 dịch vụ Cloud tại Việt Nam
- Lý do các dịch vụ Cloud có thể phục vụ hàng triệu người 1 lúc?
- Lợi ích của việc sử dụng Cloud?

Nộp bài tập gửi về: cuongdo@tlu.edu.vn

Other Interesting Features

Spot market for preemptible machines

Wide geographic access for disaster recovery
and speed of access

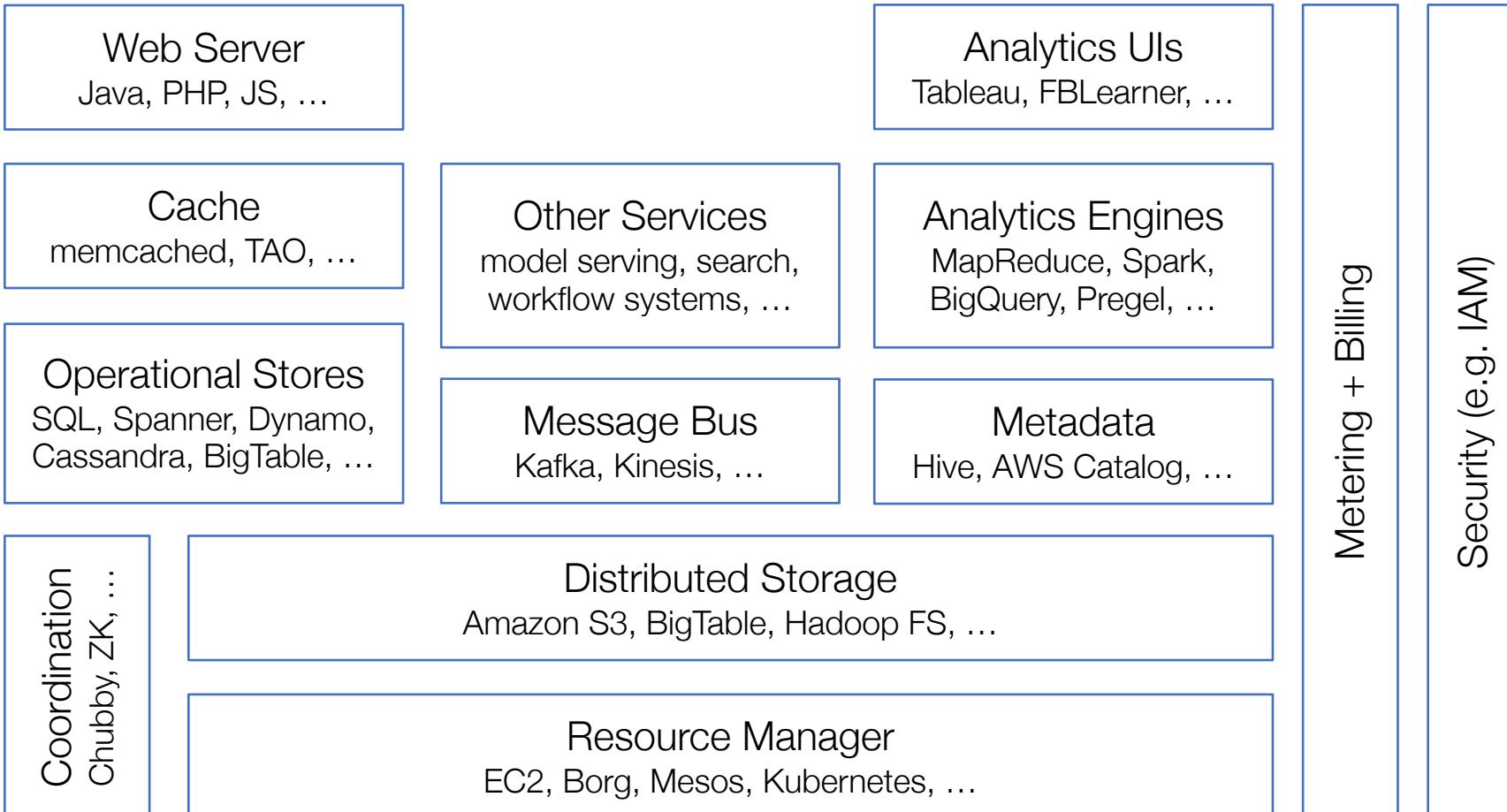
Ability to quickly try exotic hardware

Ability to A/B test anything

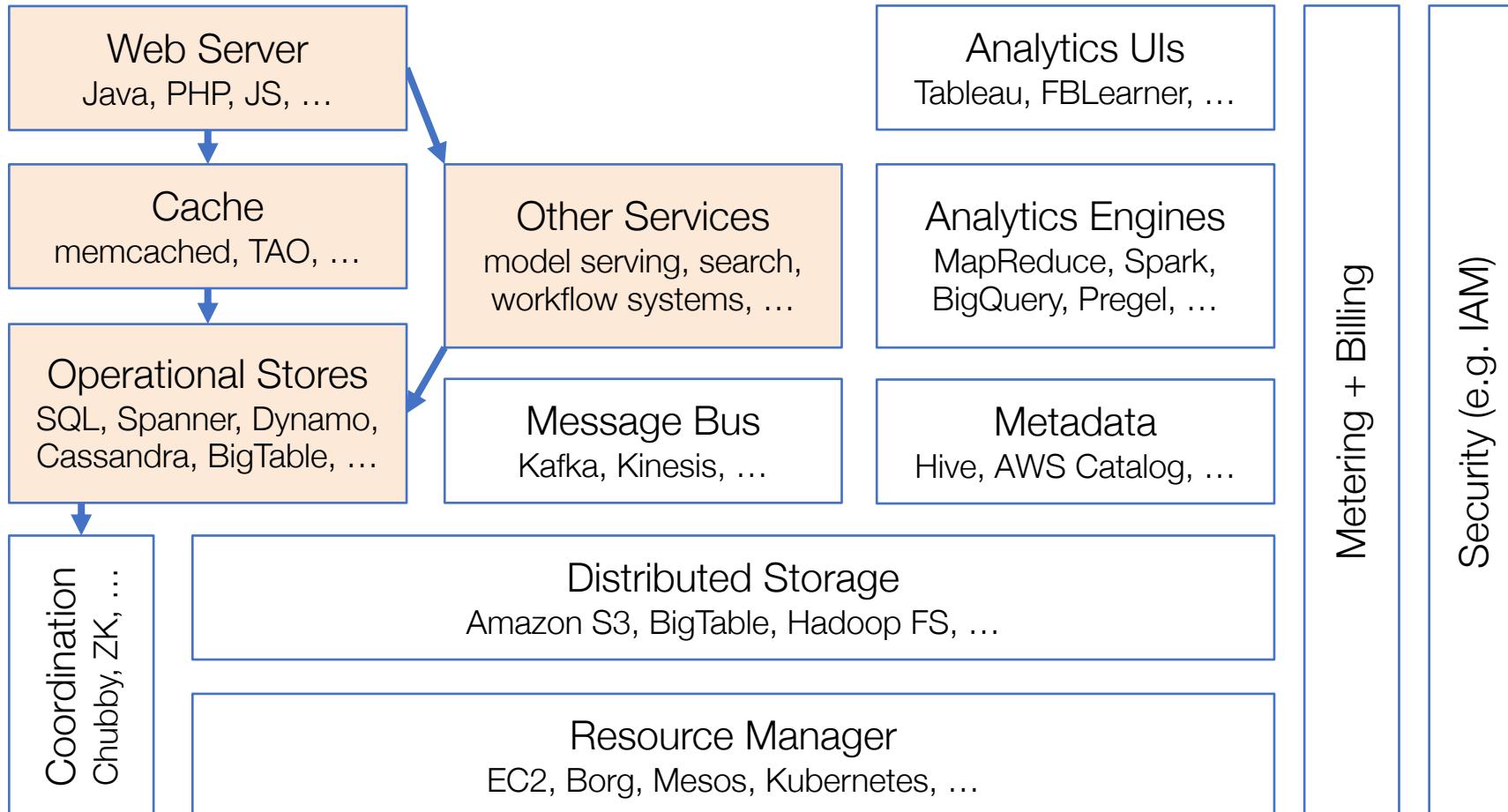
Common Cloud Applications

1. Web and mobile applications
2. Data analytics (MapReduce, SQL, ML, etc)
3. Stream processing
4. Batch computation (HPC, video, etc)

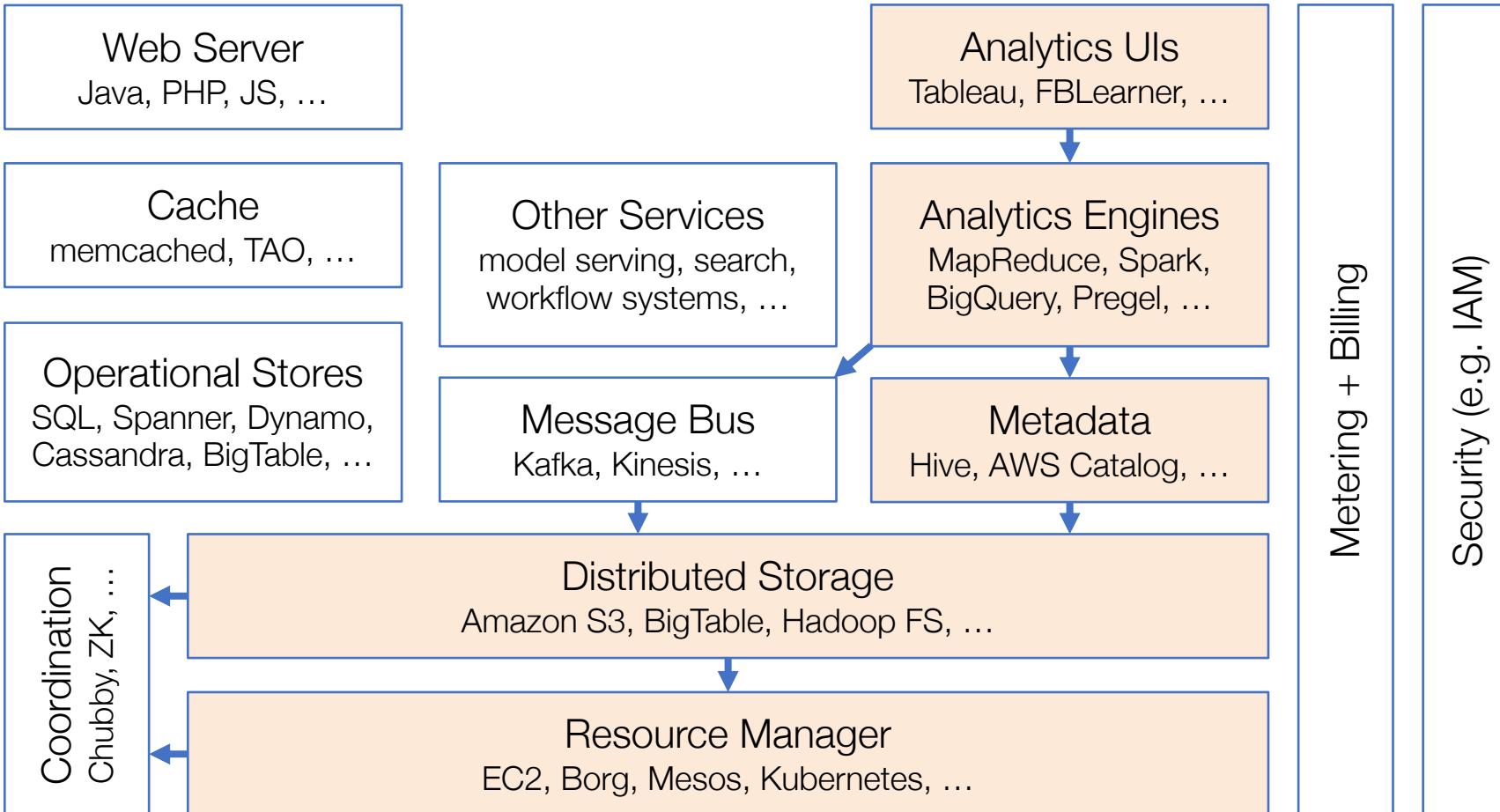
Cloud Software Stack



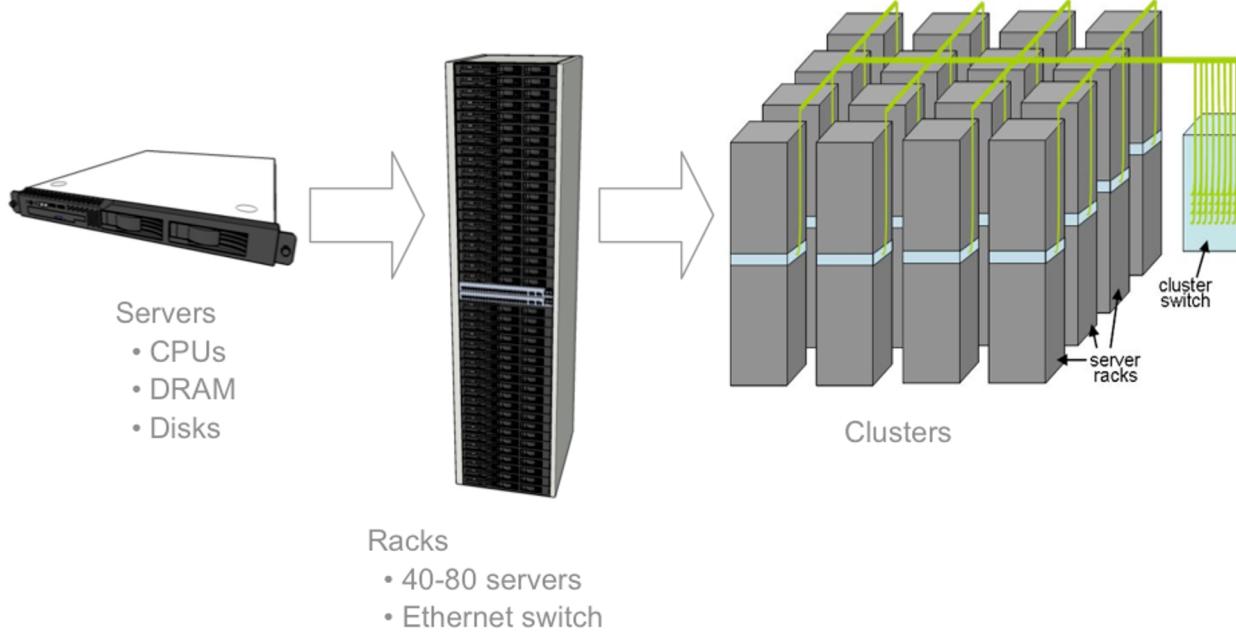
Example: Web Application



Example: Analytics Warehouse



Datacenter Hardware



Rows of rack-mounted servers

Datacenter: 50 – 200K of servers, 10 – 100MW

Often organized as few and mostly independent clusters

Datacenter Example



Datacenter HW: Compute

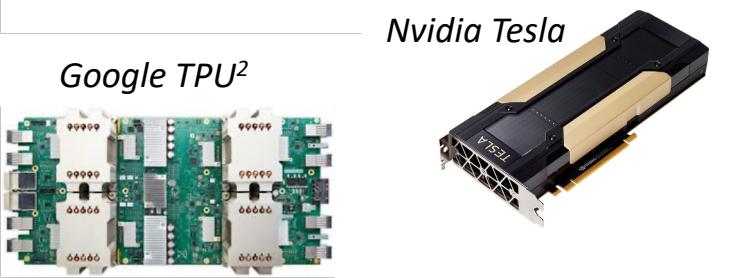
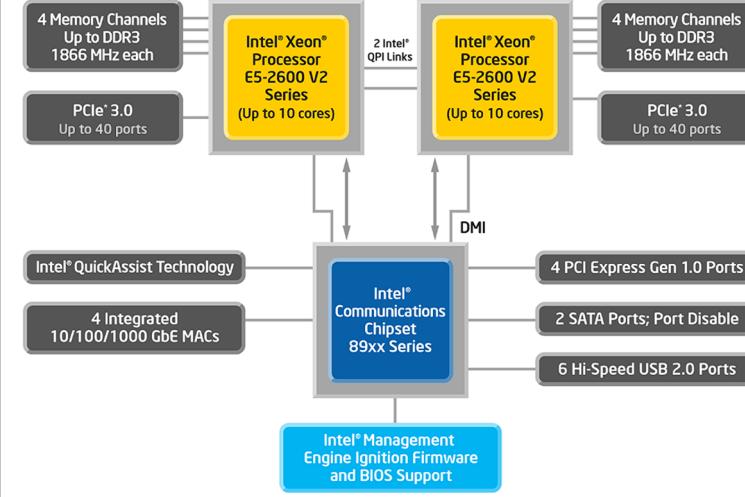
The basics

Multi-core CPU servers
1 & 2 sockets

What's new

GPUs
FPGAs
Custom accelerators (AI)

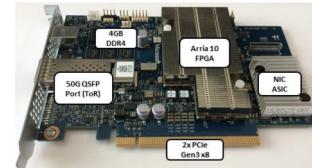
2-socket server



Google TPU²



Microsoft Catapult



Hardware Heterogeneity

Standard Systems	I Web	III Database	IV Hadoop	V Haystack	VI Feed
CPU	High 2 x E5-2670	High 2 x E5-2660	High 2 x E5-2660	Low 1 x E5-2660	High 2 x E5-2660
Memory	Low 16GB	High 144GB	Medium 64GB	Med-Hi 96GB	High 144GB
Disk	Low 250GB	High IOPS 3.2 TB Flash	High 15 x 4TB SATA	High 30 x 4TB SATA	Medium 2TB SATA + 1.6TB Flash
Services	Web, Chat	Database	Hadoop	Photos, Video	Multifeed, Search, Ads

[Facebook server configurations]

Custom-design servers

Configurations optimized for major app classes

Few configurations to allow reuse across many apps

Roughly constant power budget per volume

Datacenter HW: Storage

The basics

Disk trays

SSD & NVM Flash

NVMe Flash



JBOD disk array

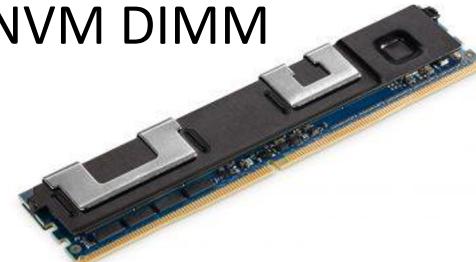


What's new

Non-volatile memories

New archival storage (e.g., glass)

NVM DIMM



Distributed with compute or NAS sys

Remote storage access for many use cases (why?)

Datacenter HW: Networking

The basics

10, 25, and 40GbE NICs

40 to 100GbE switches

Clos topologies

40GbE Switch



What's new

Software defined networking

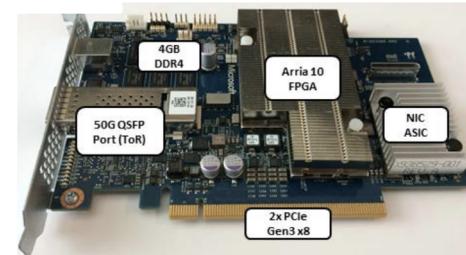
Smart NICs

FPGAs

Smart NIC



Microsoft Catapult



Useful Latency Numbers

Initial list from Jeff Dean, Google

L1 cache reference	0.5 ns
Branch mispredict	5 ns
L3 cache reference	20 ns
Mutex lock/unlock	25 ns
Main memory reference	100 ns
Compress 1K bytes with Snappy	3,000 ns
Send 2K bytes over 10Ge	2,000 ns
Read 1 MB sequentially from memory	100,000 ns
Read 4KB from NVMe Flash	50,000 ns
Round trip within same datacenter	500,000 ns
Disk seek	10,000,000 ns
Read 1 MB sequentially from disk	20,000,000 ns
Send packet CA → Europe → CA	150,000,000 ns

Useful Throughput Numbers

DDR4 channel bandwidth	20 GB/sec
PCIe gen3 x16 channel	15 GB/sec
NVMe Flash bandwidth	2GB/sec
GbE link bandwidth Gbps	10 – 100
Disk bandwidth Gbps	6
NVMe Flash 4KB IOPS	500K – 1M
Disk 4K IOPS – 200	100

Performance Metrics

Throughput

Requests per second

Concurrent users

Gbytes/sec processed

...

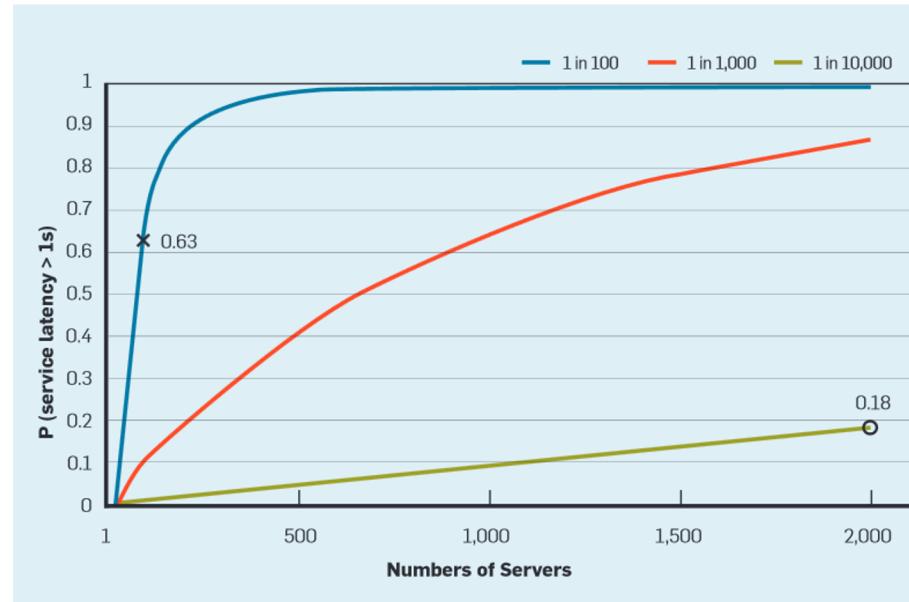
Latency

Execution time

Per request latency

Tail Latency

[Dean & Barroso,'13]



The 95th or 99th percentile request latency
End-to-end with all tiers included

Larger scale → more prone to high tail latency

Total Cost of Ownership (TCO)

$\text{TCO} = \text{capital (CapEx)} + \text{operational (OpEx) expenses}$

Operators perspective

CapEx: building, generators, A/C, compute/storage/net
HW

Including spares, amortized over 3 – 15 years

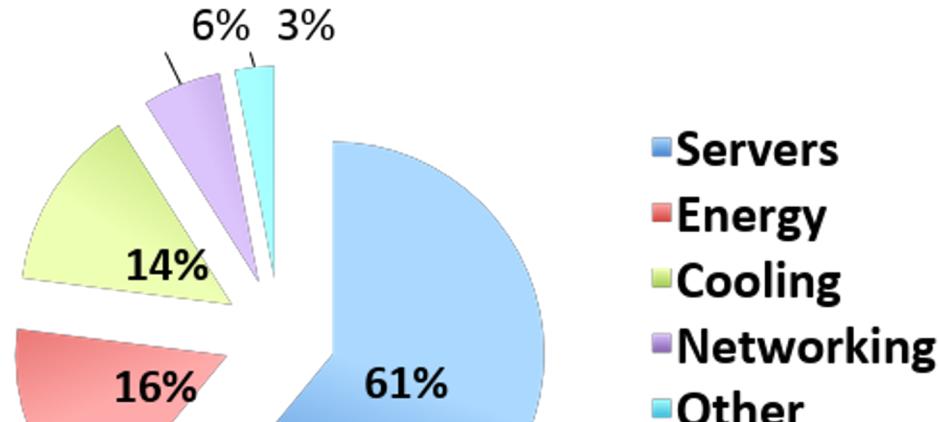
OpEx: electricity (5-7c/KWh), repairs, people, WAN,
insurance, ...

Users perspective

CapEx: cost of long term leases on HW and services

OpEx: pay per use cost on HW and services, people

Operator's TCO Example



[Source: James Hamilton]

Hardware dominates TCO, make it cheap
Must utilize it as well as possible

Reliability

Failure in time (FIT)

Failures per billion hours of operation = $10^9/\text{MTTF}$

Mean time to failure (MTTF)

Time to produce first incorrect output

Mean time to repair (MTTR)

Time to detect and repair a failure

Thảo luận

- Mỗi nhóm nêu 1 giải pháp giúp ảo hóa Desktop/ Server
- Phân tích ưu nhược điểm và thị phần các giải pháp
- Các thông tin cần có:
 1. Tên giải pháp
 2. Công ty – môi trường cài đặt: linux/ Window/ Mac?
 3. Thị phần giải pháp: bao công ty dùng? phát triển từ năm nào?
 4. Ưu điểm? nhược điểm

Lưu ý

- Các bạn vào lớp muộn đề nghị trật tự đi vào cửa sau
- Tắt/ chuyển điện thoại sang chế độ rung trong lớp học
- Các nhóm mang theo laptop đến lớp để tra cứu/ làm thảo luận tại lớp

Bài tập về nhà

- Nhóm gửi tên 3 sản phẩm mã nguồn mở triển khai hệ thống VDI