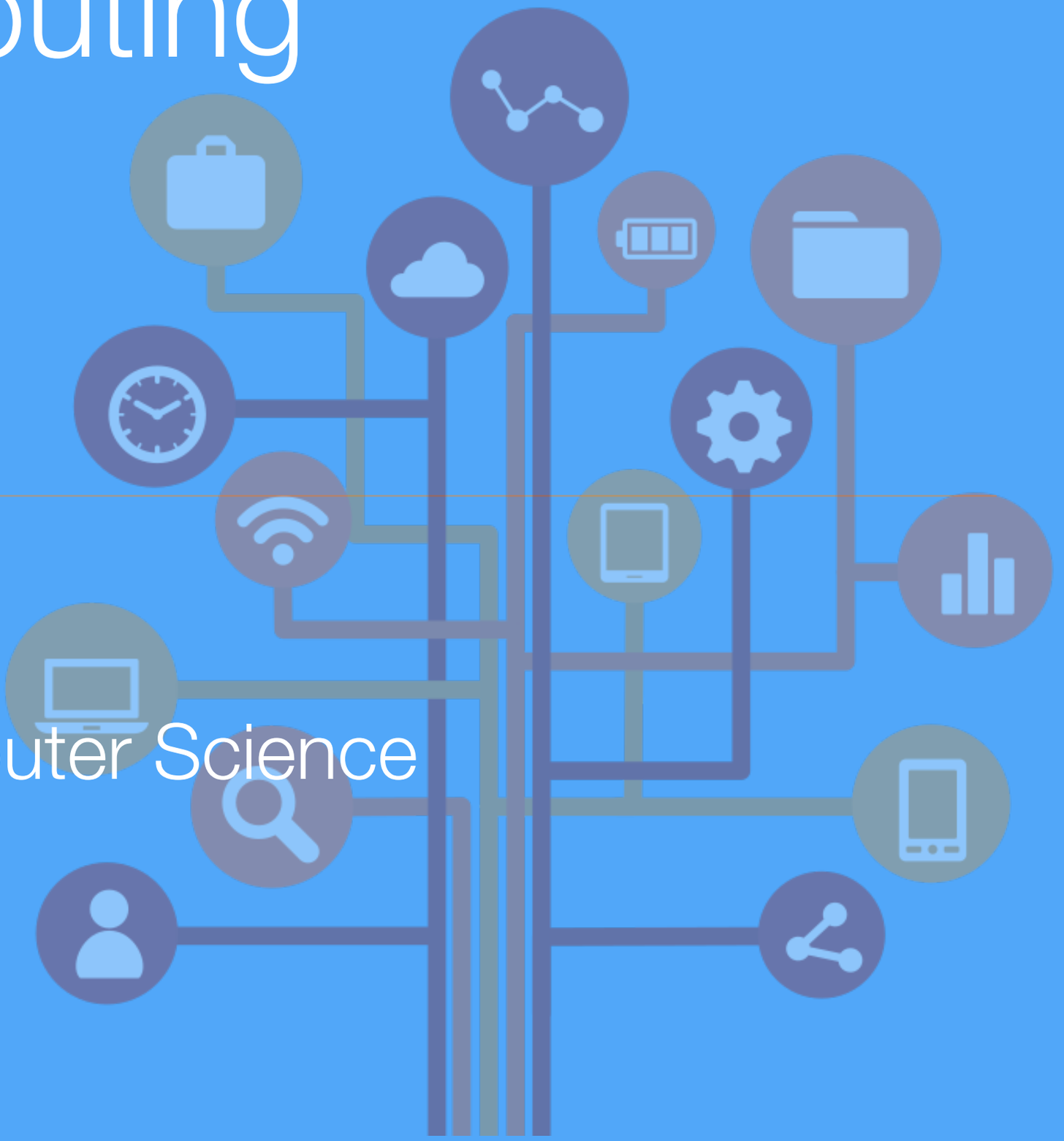
Cloud Computing
Fundamentals



Hong Xu
Department of Computer Science

Spring 2020

香港城市大學
City University of Hong Kong

Outline

- Real-world examples of the cloud
- Definitions of cloud computing
- Key cloud concepts and characteristics
- Deployment scenarios
- Service models

Cloud: Massive Scale

- Facebook [GigaOM, 2012]
  - 30K in 2009 -> 60K in 2010 -> 180K in 2012
- Microsoft [DC knowledge]
  - > 1 million, 2013
- AWS EC2 [Randy Bias, 2009]
  - 40K, 8 cores per machine
- Google [DC knowledge]
  - > 900K, 2013

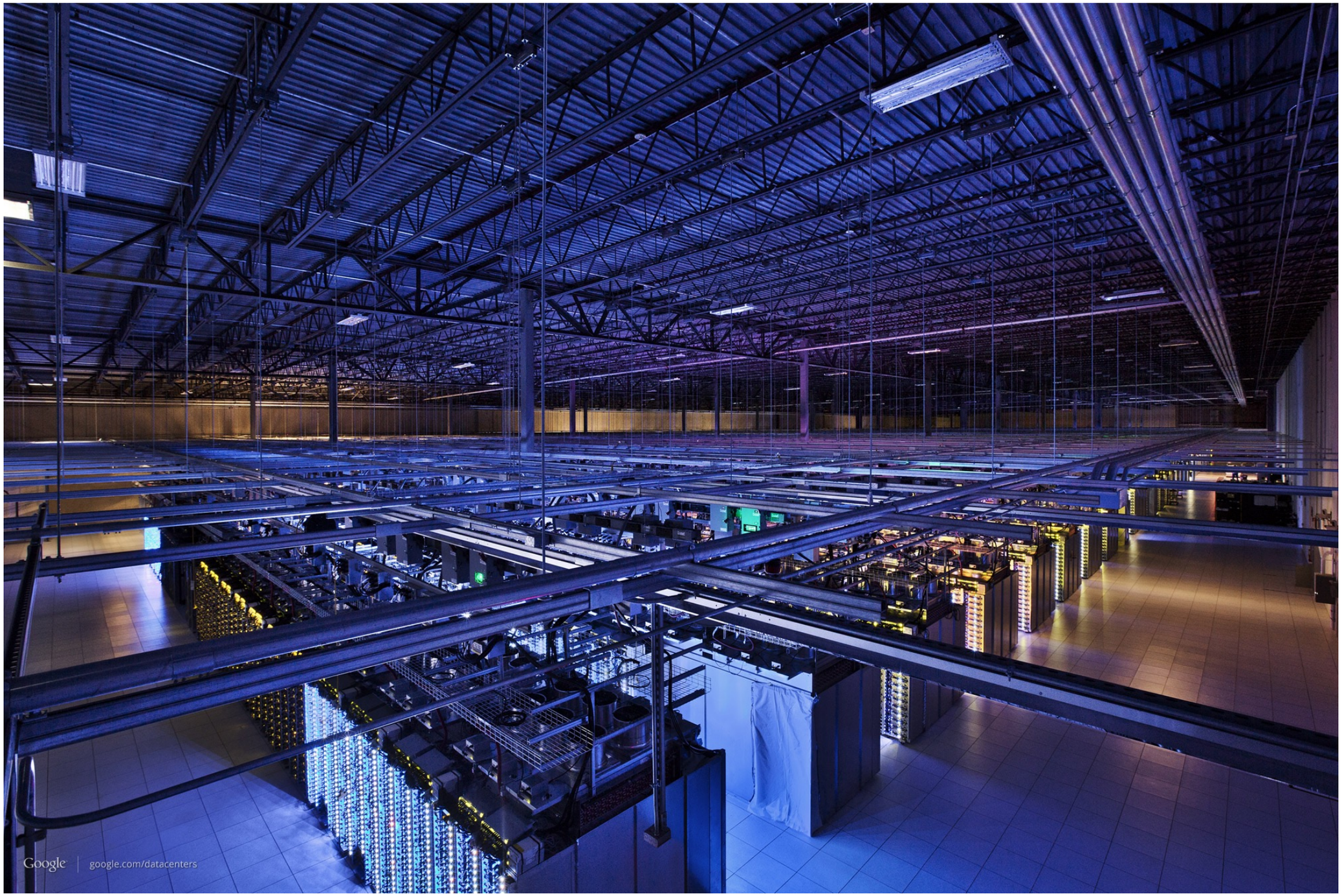
Data center: outside

Copyright: Google

Data center: outside

Copyright: Google

Data center: inside



Copyright: Google

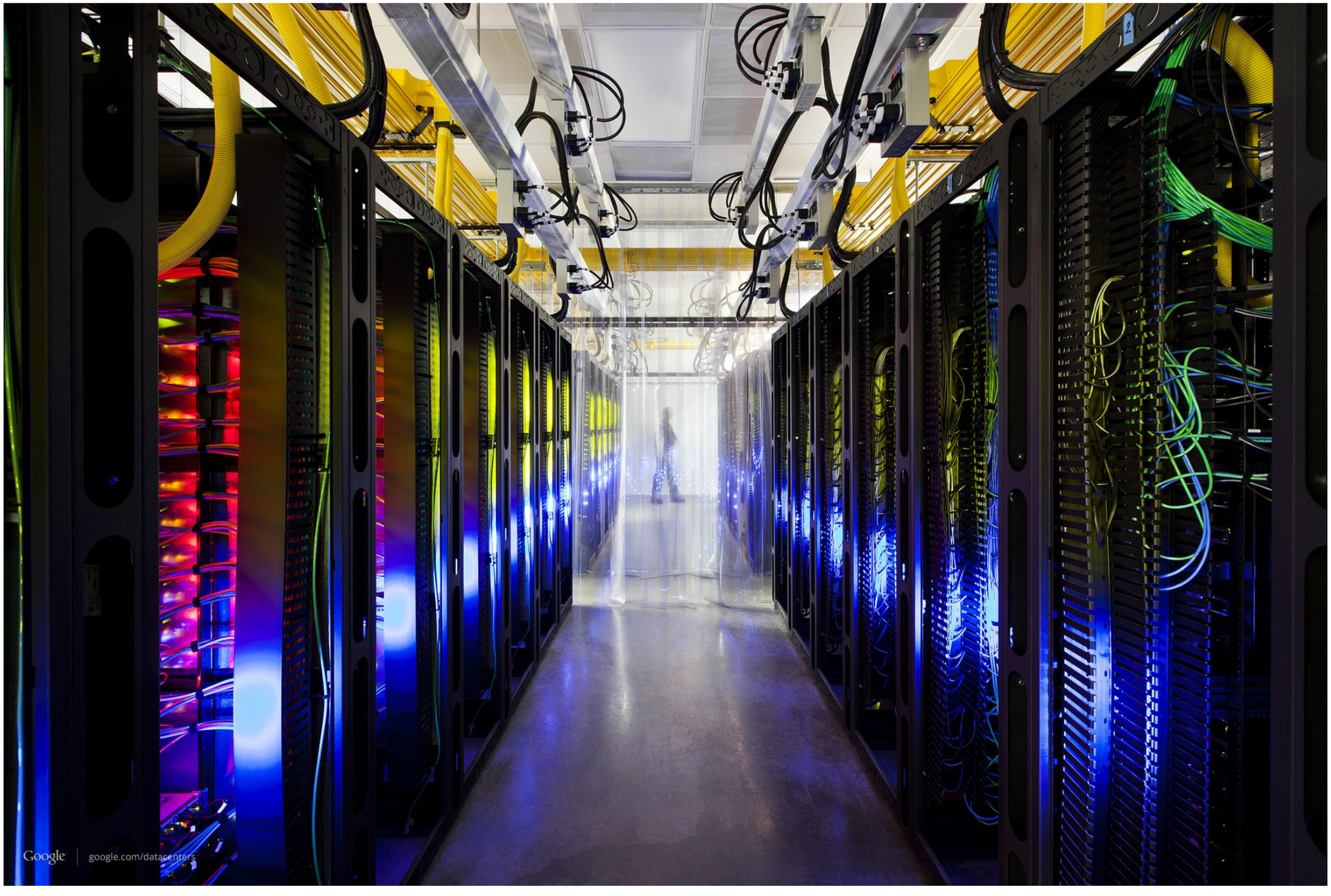
Server racks

Photo credit: Google

Server: inside

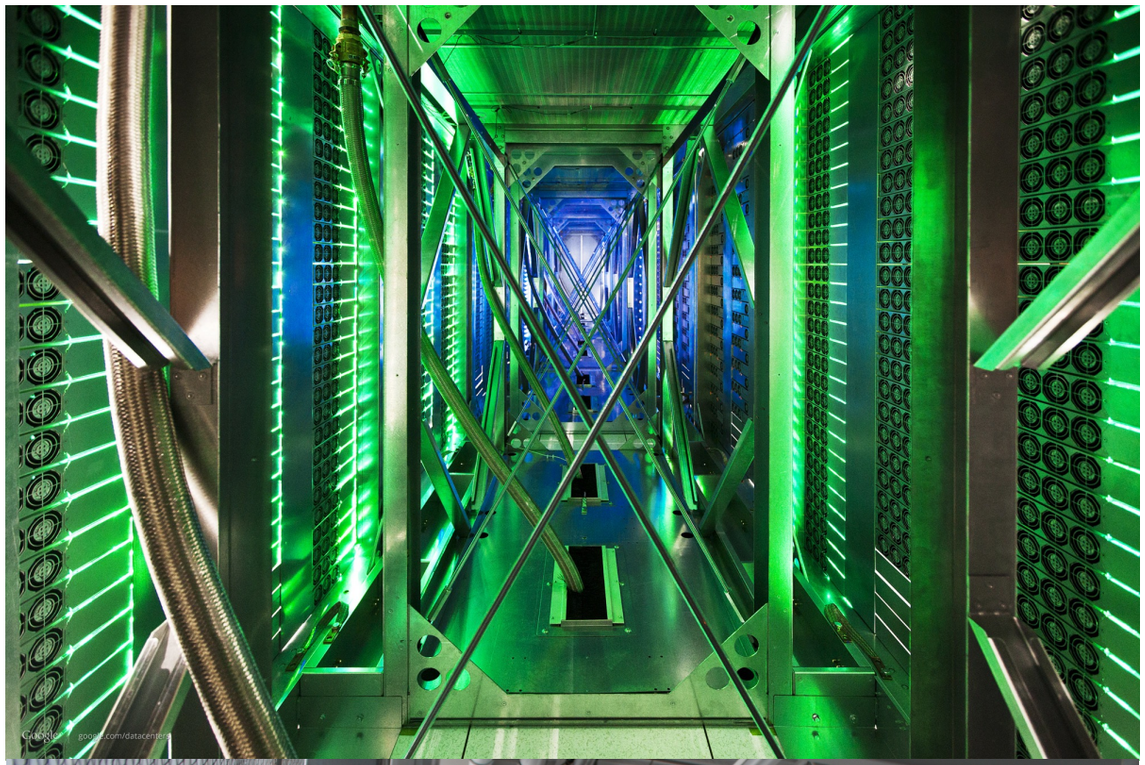
Copyright: xsnet

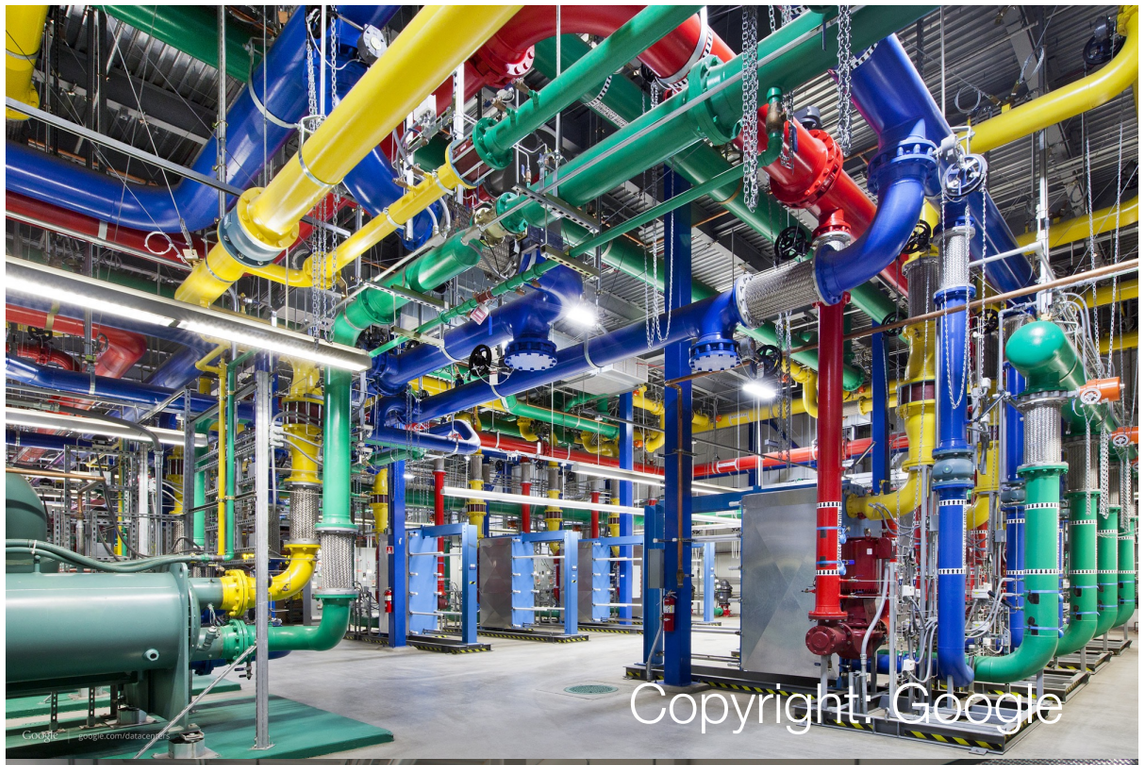
Network room



Copyright: Google

Cooling





Cloud providers

Cloud-based services (U.S.)

Cloud-based services (China)

支付宝™
ALIPAY

滴滴出行
滴滴一下美好出行

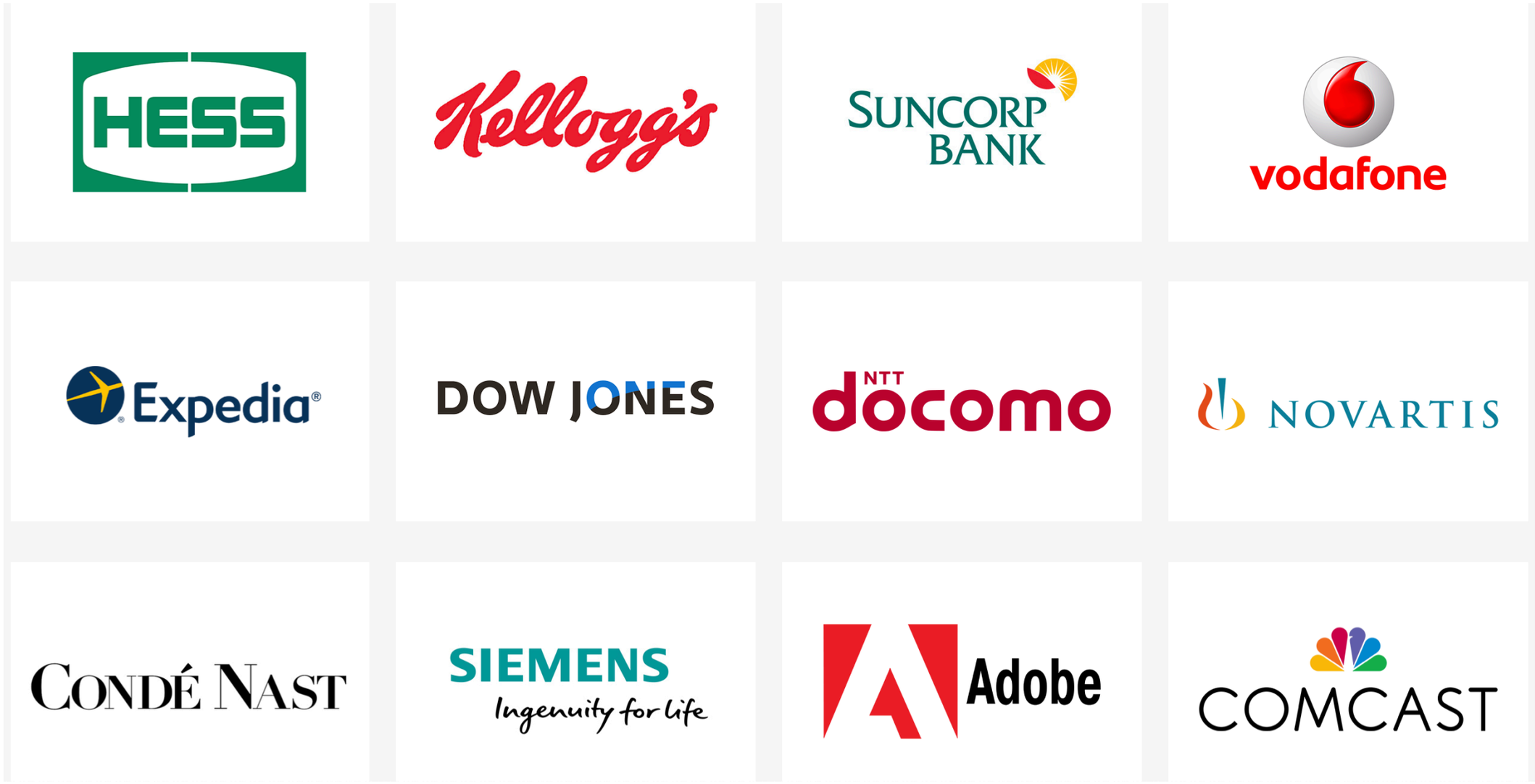
大众点评
dianping.com

美团
meituan.com

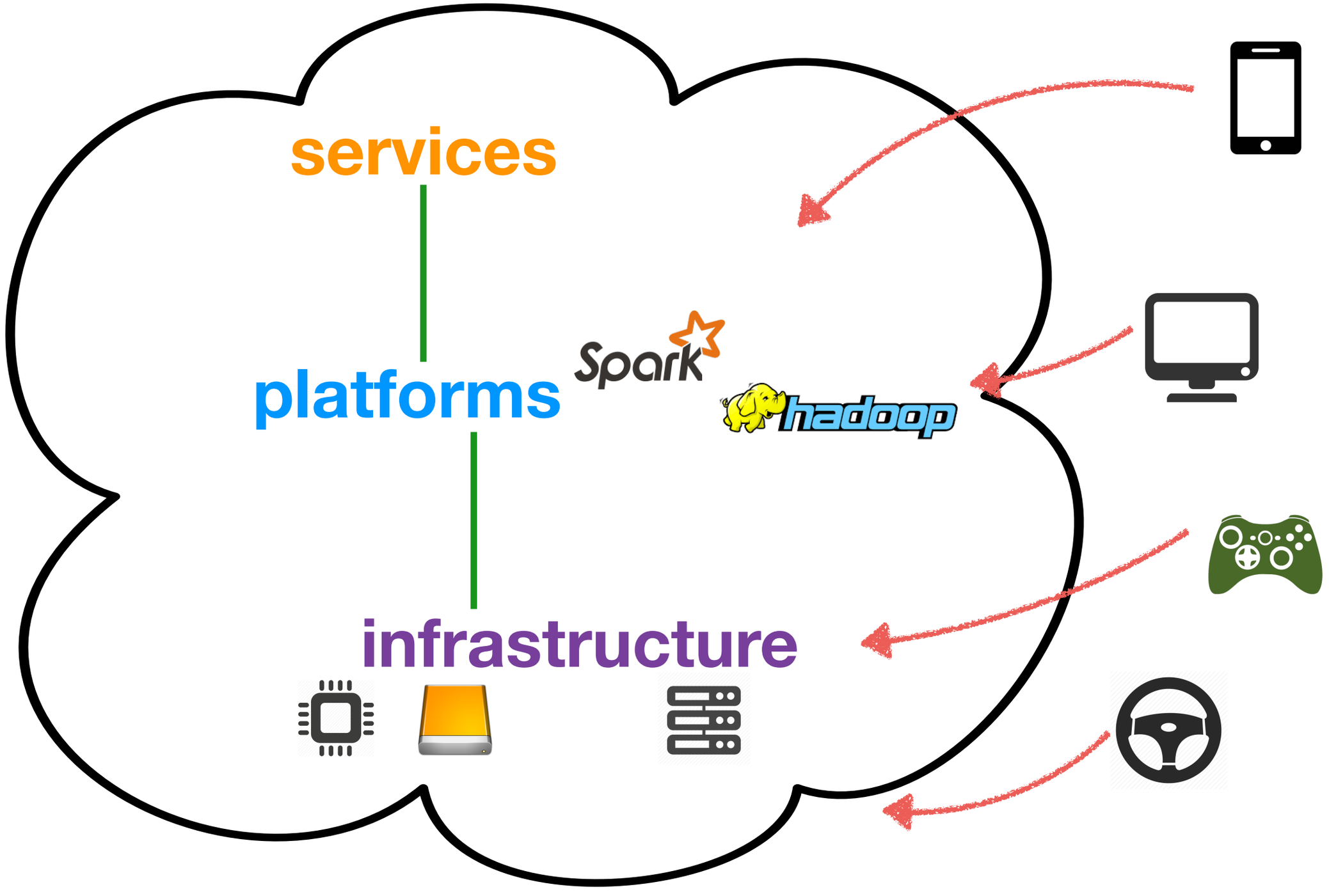
腾讯视频
V.QQ.COM

Organizations that use cloud

- Enterprise, government, education, etc.



So what is a cloud?



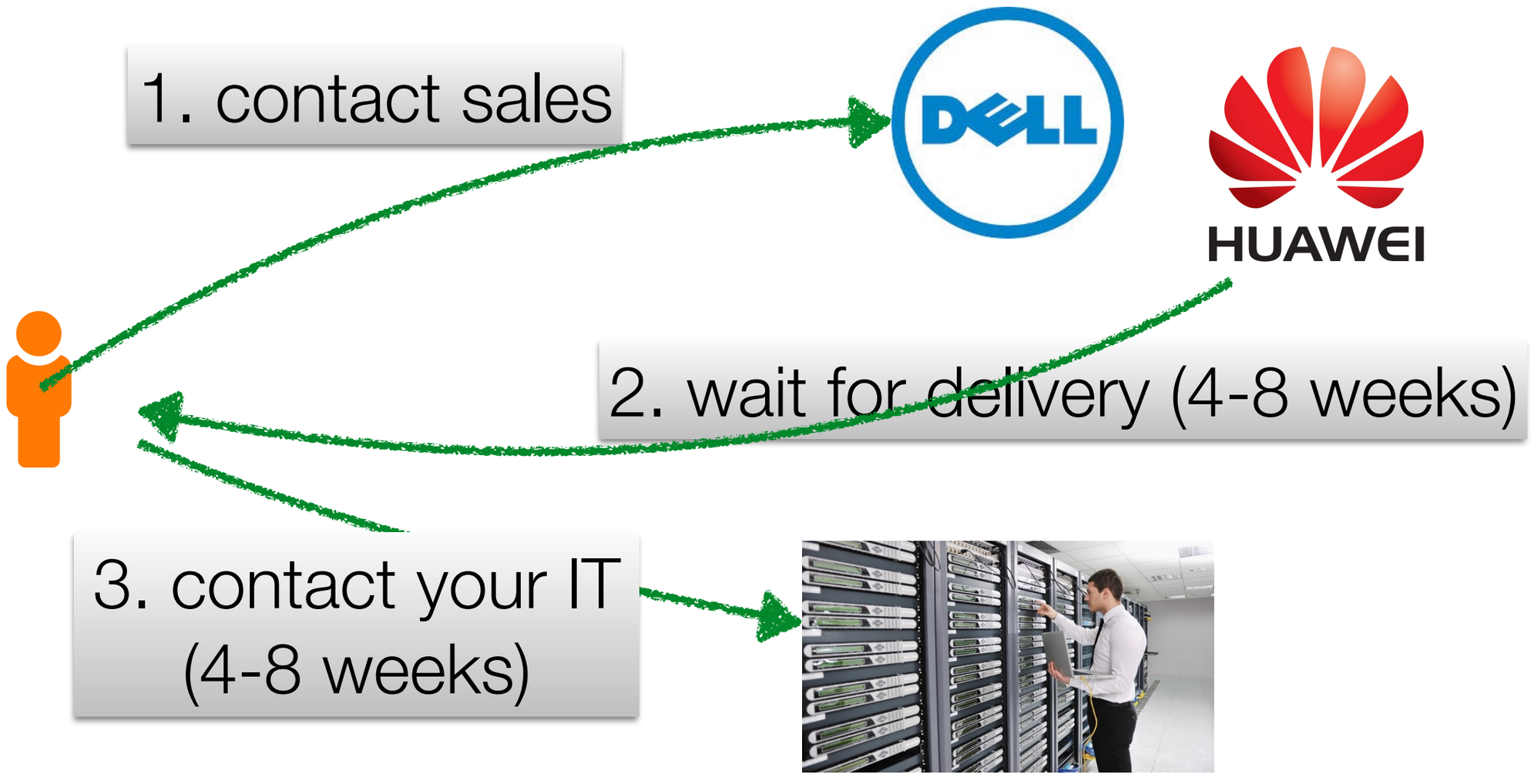
A definition

Cloud computing is a model for enabling ubiquitous, 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.

National Institute of Standards and Technology (NIST), U.S. Department of Commerce

On-demand self-service

- Suppose you run a start-up, and need 20 servers for it
- Traditionally:



On-demand self-service

- With cloud computing:

On-demand self-service

- A consumer can unilaterally provision computing capabilities, such as servers and network storage, as needed automatically without requiring human interaction with each service provider.

- Cloud computing makes the underlying technology, beyond the user device, almost invisible

- Advantages for consumers: flexible, minimal overhead, quick and easy

Broad network access

- Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., smart phones, tablets, laptops, and workstations).
- Advantages for consumers: “Always-on” experience, like utilities (electricity)

Resource pooling

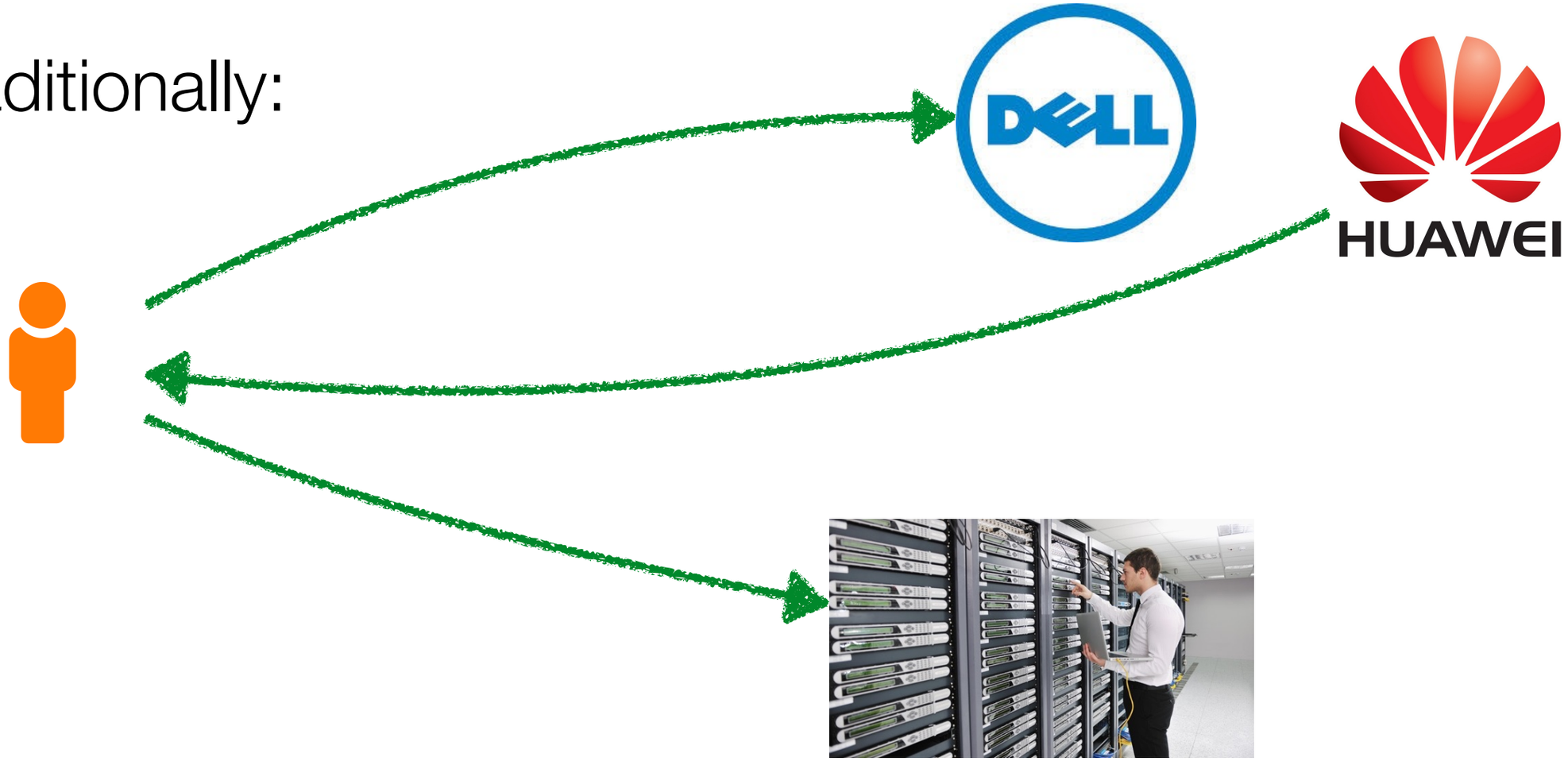
- The provider's resources are pooled to serve consumers using a multi-tenant model, with different physical and virtual resources dynamically allocated according to consumer demand.

- Purchasing, powering, & managing machines at scale gives lower per-unit costs than customers

- Advantage for providers: efficiency in utilization

Rapid elasticity

- Resources can be rapidly and elastically scaled up and down.
- Suppose your business grows and needs 40 servers now
- Traditionally:



Rapid elasticity

- With cloud computing

- Advantage for consumers: flexible, quick and easy

Measured service

- Resource usage can be monitored, controlled, and reported, providing transparency for both the provider and consumer
- **Pay-as-you-go, pay only for what you use**
  - Per minute, per byte, etc.
  - Mo minimum or upfront fee

Measured service

- EC2 example:
  - https://aws.amazon.com/ec2/pricing/
  - Spot market for pre-emptible machines, much cheaper (use them for your assignments)

Common cloud applications

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

Cloud software stack

Web Server
Java, PHP, JS, …

Cache
memcached, TAO, …

Operational Stores
SQL, Spanner, Dynamo, Cassandra, BigTable, …

Other Services
model serving, search, workflow systems, …

Message Bus
Kafka, Kinesis, …

Analytics UIs
Tableau, FBLearner, …

Analytics Engines
MapReduce, Spark, BigQuery, Pregel, …

Metadata
Hive, AWS Catalog, …

Coordination
Chubby, ZK, …

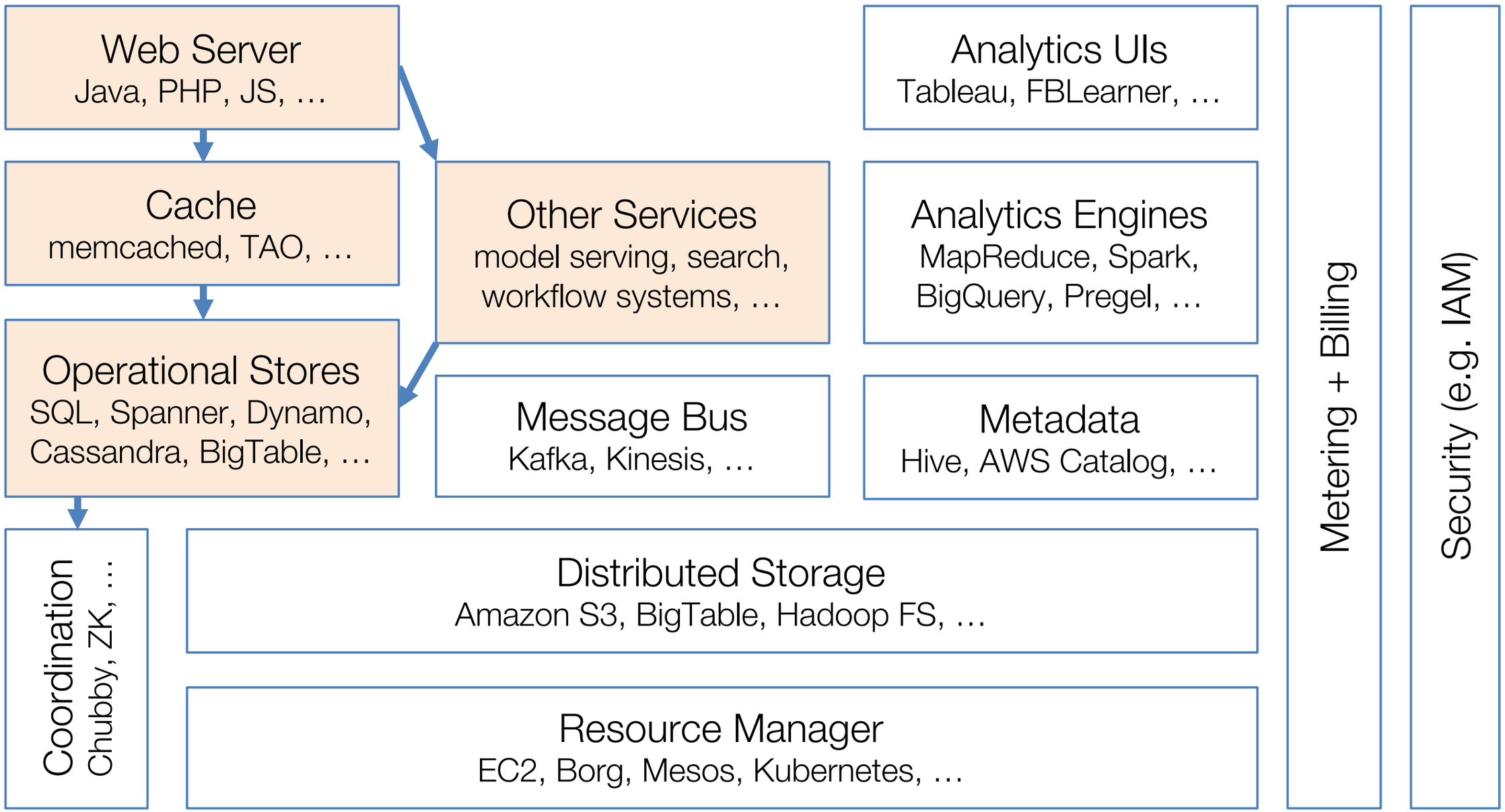
Distributed Storage
Amazon S3, BigTable, Hadoop FS, …

Resource Manager
EC2, Borg, Mesos, Kubernetes, …

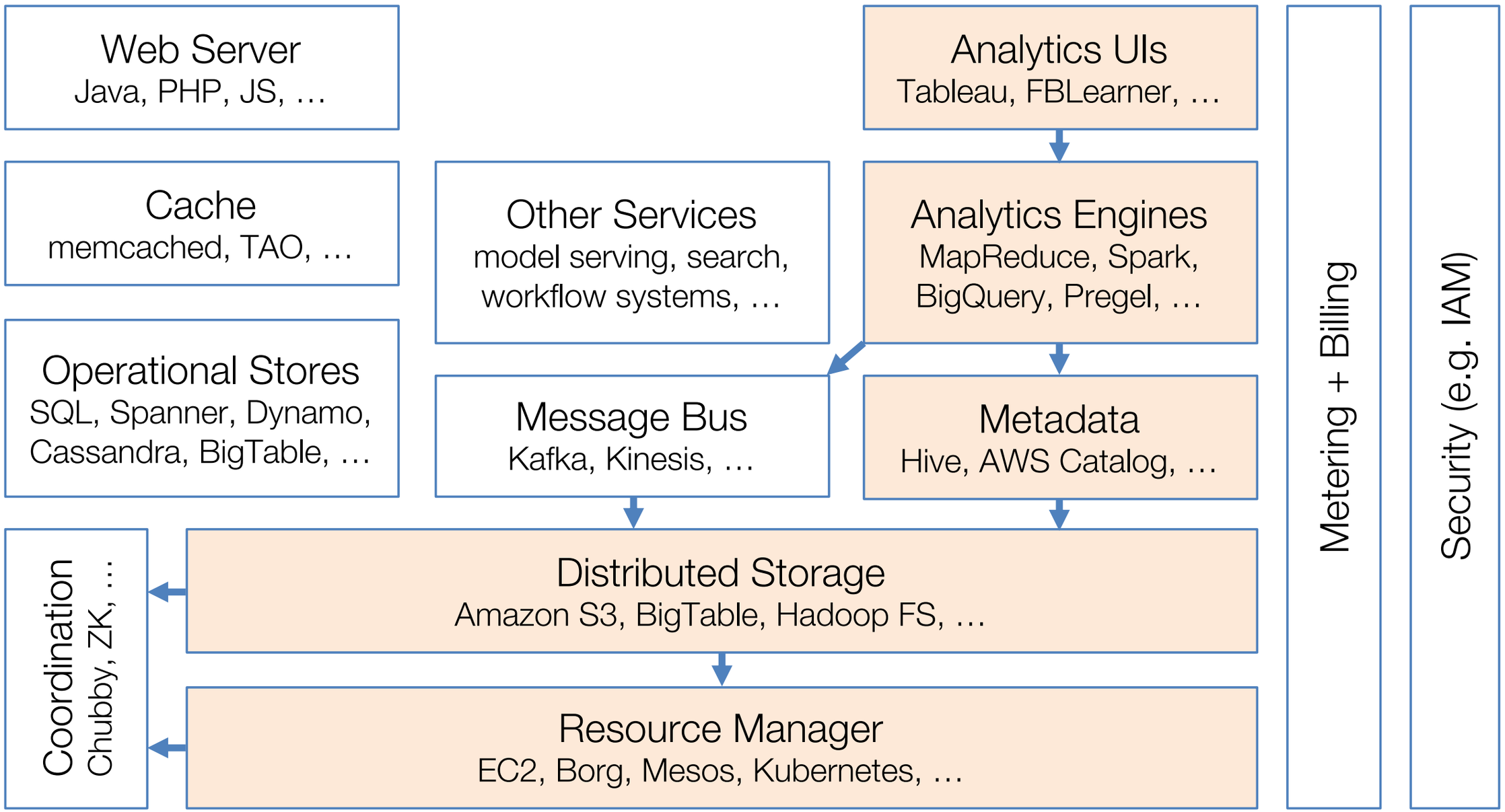
Metering + Billing

Security (e.g. IAM)

Example: Web app



Example: Data analytics

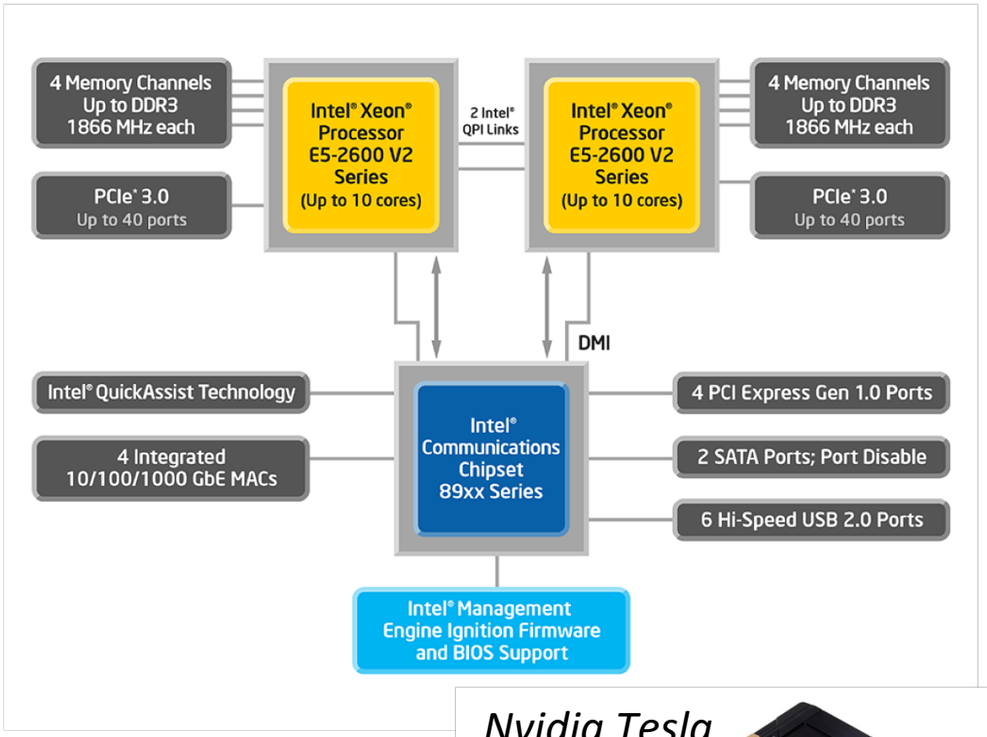


Datacenter HW: Compute

The basics

- Multi-core CPU servers
- 1 & 2 sockets

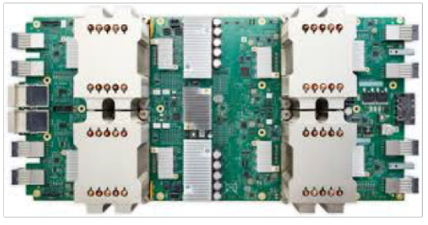
2-socket server



What's new

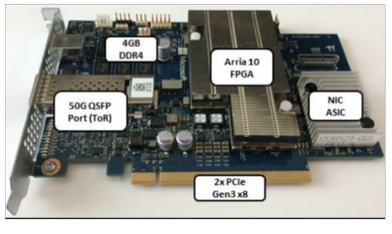
- GPUs
- FPGAs
- Custom accelerators (AI)

Google TPU²



Nvidia Tesla

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

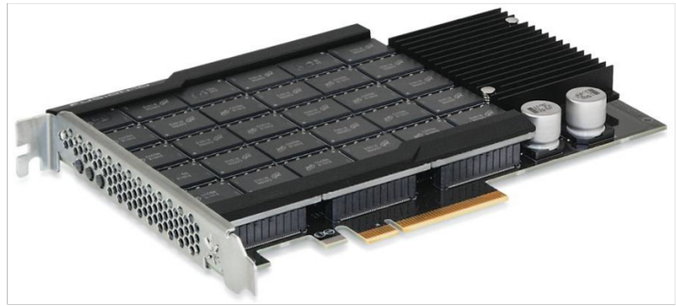
- Configurations optimized for major app classes
- Few configurations to allow reuse across many apps
- Roughly constant power budget per volume

Datacenter HW: Storage

The basics

- Disk trays
- SSD & NVM Flash

NVMe Flash

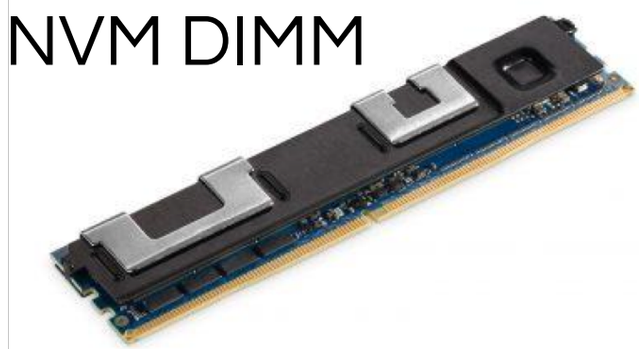


JBOD disk array

What's new

- Non-volatile memories
- New archival storage (e.g., glass)

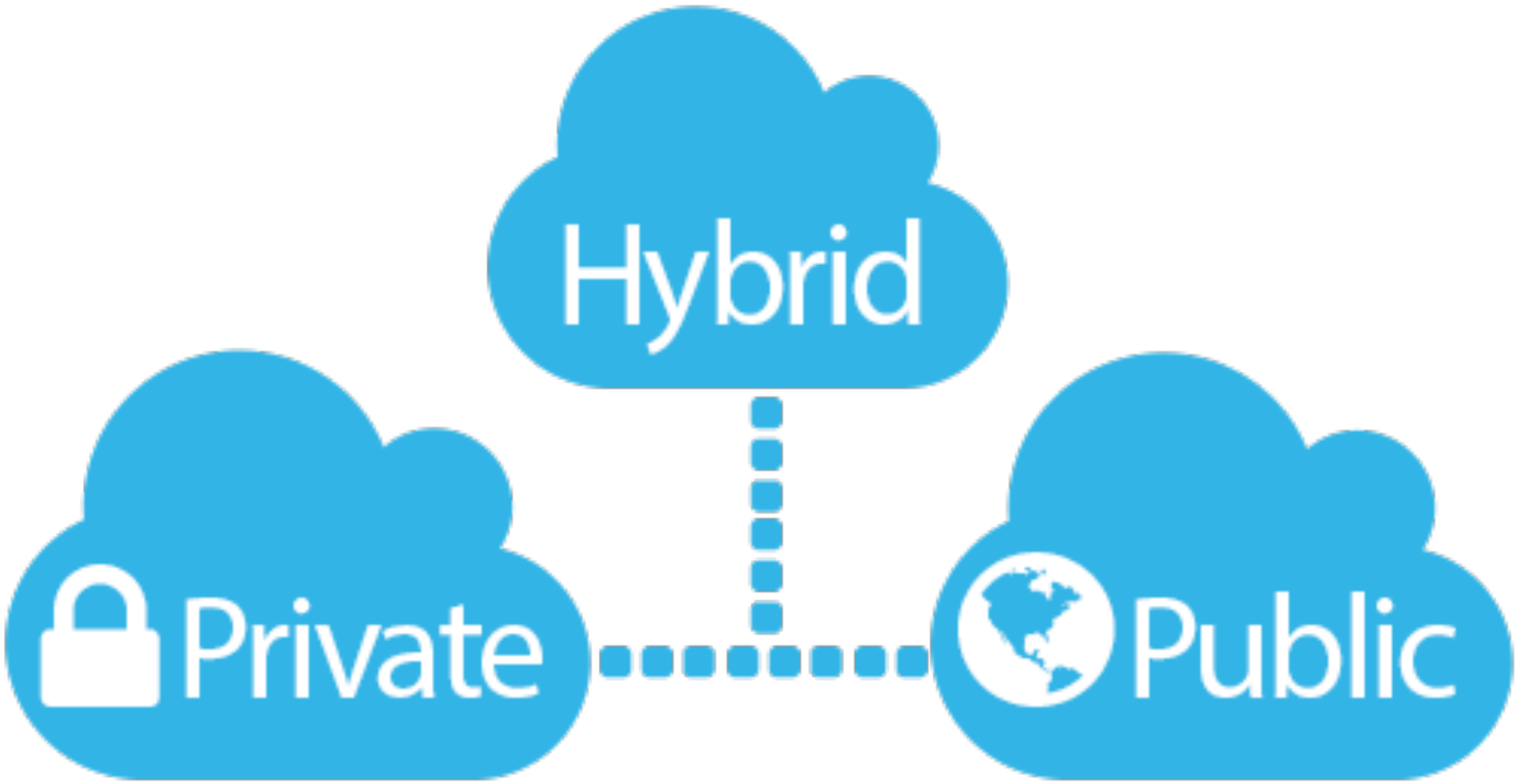
NVM DIMM



Distributed with compute or NAS systems

- Remote storage access for many use cases (why?)

Cloud deployment models



Cloud deployment models

- Public cloud
  - Providers let clients access the cloud via Internet
  - Made available to the general public

Cloud deployment models

- Public cloud
  - Multitenant virtualization, global-scale infrastructure
  - Functions and pricing vary

Copyright: Google

Cloud deployment models

- Private cloud
  - The cloud is used solely by an organization (e.g. CityU, HSBC)
  - May reside in-house or off-premise

Cloud deployment models

- Private cloud
  - Secure, dedicated infrastructure with the benefits of on-demand provisioning
  - Not burdened by network bandwidth and availability issues and security threats associated with public clouds.
  - Greater control, security, and resilience.

Cloud deployment models

- Hybrid cloud
  - Composed of multiple clouds (private, public, etc.) that remain independent entities, but interoperate using standard or proprietary protocols
  - Banks, hospitals, government

Service models

- Infrastructure-as-a-Service (IaaS)
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)

IaaS

- Providers give you the computing infrastructure made available as a service, in the form of virtual machines (VM)
- Providers manage a large pool of resources with virtualization
- Customers “rent” these physical resources to customize their own infrastructure
- You operate OS and software by yourself

IaaS use case

- Netflix rents thousands of servers, terabytes of storage from Amazon Web Services (AWS)
- Develop and deploy specialized software for transcoding, storage, streaming, analytics, etc. on top of it
- Is able to support tens of millions of connected devices, used by 40+ million users from 40+ countries

PaaS

- Providers give you a software platform, or middleware, where applications run
- You develop and maintain and deploy your own software on top of the platform
- The hardware needed for running the software is automatically managed by the platform. You can’t explicitly ask for resources.

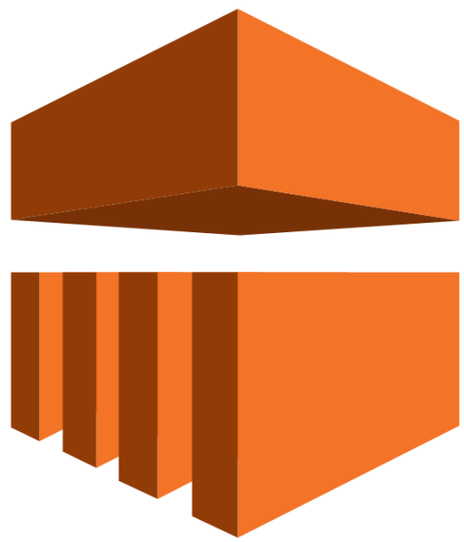
PaaS

- You have automatic scalability, without having to respond to request load increase/decrease

Web runtime

deploy web apps and services
e.g. node.js, .Net, Go, etc.

Data analytics engine



managed Hadoop/Spark deployment

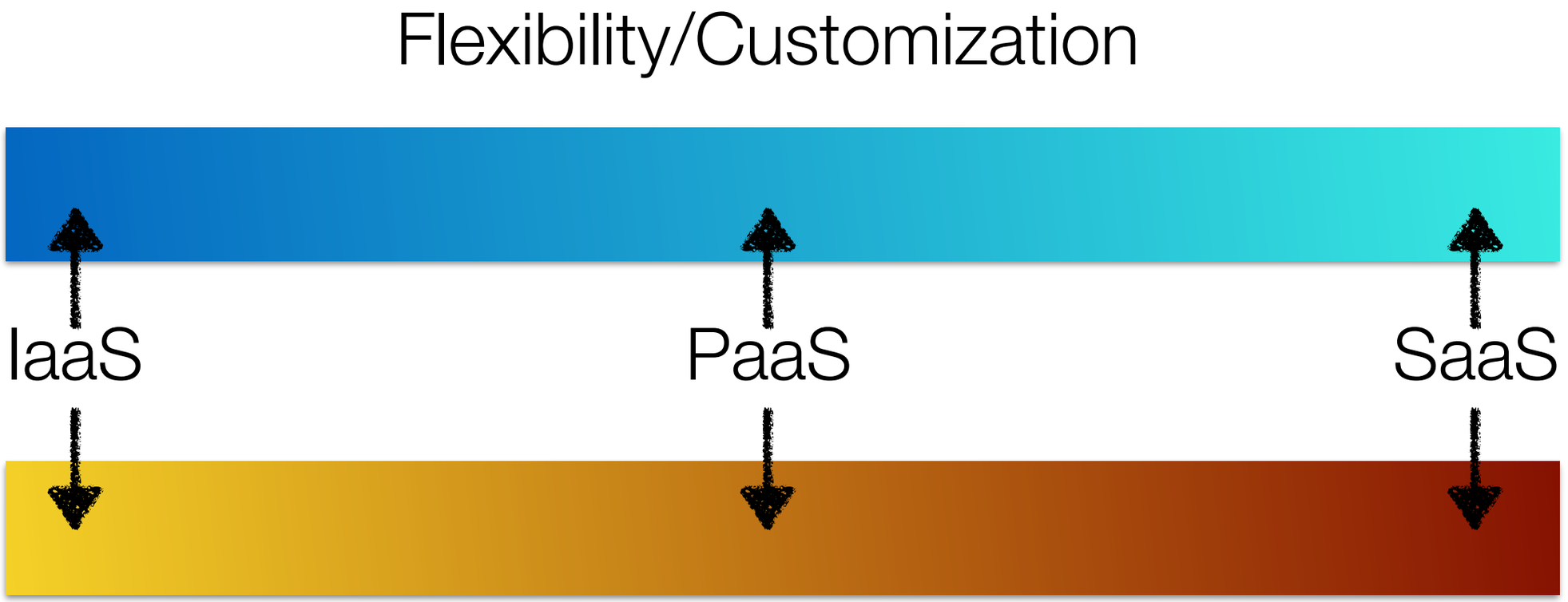
SaaS

- Providers give you a piece of software/application. They take care of updating, and maintaining it.
- You simply use the software through the Internet.

SaaS

- CityU uses Google Apps and Office 365 for staff email, calendar, etc.
- Don’t know how much they charge CityU though…

A comparison



Tradeoff between flexibility and “built-in” functionality

Issues of cloud

Total Cost of Ownership (TCO)

TCO = capital (CapEx) + operational (OpEx) expenses

Operators perspective

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

Including spares, amortized over 3 – 15 years

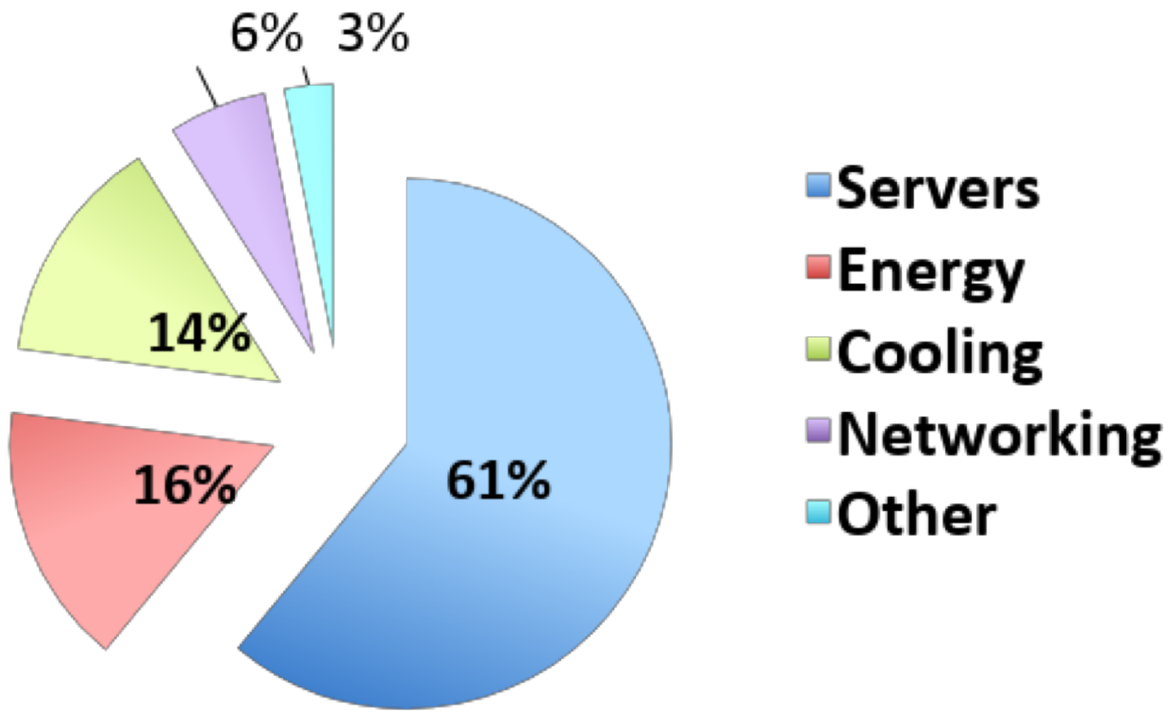
OpEx: electricity (5-7c/KWh), repairs, people, WAN, insurance, …

Users perspective

CapEx: cost of long term leases on HW and services

OpeEx: 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

Yearly Datacenter Flakiness

~0.5 overheating (power down most machines in <5 mins, ~1-2 days to recover)

~1 PDU failure (~500-1000 machines suddenly disappear, ~6 hrs to come back)

~1 rack-move (plenty of warning, ~500-1000 machines powered down, ~6 hrs)

~1 network rewiring (rolling ~5% of machines down over 2-day span)

~20 rack failures (40-80 machines instantly disappear, 1-6 hours to get back)

~5 racks go wonky (40-80 machines see 50% packet loss)

~8 network maintenances (4 might cause ~30-minute random connectivity losses)

~12 router reloads (takes out DNS and external vIPs for a couple minutes)

~3 router failures (have to immediately pull traffic for an hour)

~dozens of minor 30-second blips for dns

~1000 individual machine failures (2-4% failure rate, machines crash at least twice)

~thousands of hard drive failures (1-5% of all disks will die)

Add to these SW bugs, config errors, human errors,

...

Security and Privacy

- A losing battle…
  - Can an intruder/attacker get my data in the cloud?
  - Will the provider look at my data in the cloud?
  - Will the provider give my data to third parties?
    - https://www.apple.com/legal/transparency/
    - https://www.apple.com/legal/transparency/us.html

Credit

- Some slides are from Patrick Lee's slides for CSCI 4180 at CUHK and Matei's slides for CS349D at Stanford