CLOUD COMPUTING SYSTEMS

Lecture 1

Nuno Preguiça

(nuno.preguica_at_fct.unl.pt)

OUTLINE

Definitions and types

Motivation for cloud computing

Examples

A glimpse inside the infrastructure

OUTLINE

Definitions and types

Motivation for cloud computing

Examples

A glimpse inside the infrastructure

What is cloud computing?

Cloud computing

- The practice of using a network of remote servers hosted on the Internet to store, manage, and process data, rather than a local server or a personal computer.

  - [Oxford dictionary]

- **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.

  - [NIST]

What is cloud computing? (cont.)

Cloud computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services.

[M. Armbrust, et. al. A View of Cloud Computing. CACM 2010]

Cloud Computing, the long-held dream of computing as a utility […]

[M. Armbrust, et. al. A View of Cloud Computing. CACM 2010]

- As with other utilities (water, electricity, etc.), it is available for users to use them when needed, in the amount needed, and paying what is used.

WHAT IS CLOUD COMPUTING? (CONT.)

Cloud computing refers to both the applications delivered as services over the Internet and the hardware and systems software in the data centers that provide those services.

This course
Broad overview of cloud computing, including hardware and software.
Focus on (public) cloud platforms services and how to use these services to create efficient, scalable and available applications.

010]

Clo lity

[…

010]

- As with other utilities (water, electricity, etc.), it is available for users to use them when needed, in the amount needed, and paying what is used.

TYPES OF CLOUD SERVICES: PUBLIC VS. PRIVATE

Public cloud

- The cloud **infrastructure** is **available to the public** (both individuals and organizations).
- The infrastructure is managed by the provider (public or private) and is deployed in the "provider domains".

Private cloud

- The cloud **infrastructure** is **used exclusively by a single organization** (or consortium of organizations).
- Infrastructure ownership, management and operation may be carried out by the organization itself, through third parties, or a combination of both; the infrastructure may be deployed in-house, or in a remote location.

Types of cloud services: IaaS vs. PaaS vs. SaaS

Infrastructure as a Service (IaaS)

- Provides processing, storage, networks, and other fundamental computing resources where the consumer is able to deploy and run arbitrary software. [NIST]
- E.g.:
  - VMs, disks, private virtual networks.
  - IaaS public providers: Amazon (EC2); Google (Compute Engine); Microsoft (Azure); etc.
  - Software for private IaaS: OpenStack (open software), vCloud (VMware), etc.

Types of cloud services: IaaS vs. PaaS vs. SaaS

Platform as a Service (PaaS)

- The capability provided to the consumer is to deploy onto the cloud infrastructure [...] applications created using programming languages, libraries, services, and tools supported by the provider. The consumer does not manage or control the underlying cloud infrastructure including network, servers, operating systems, or storage, but has control over the deployed applications and possibly configuration settings for the application-hosting environment. [NIST]
- E.g.:
  - Web: Google App Engine, Azure App server, Heroku
  - Computing: Azure Hdinsight (Spark, Map-reduce, etc.), Amazon EMR (Spark, Hadoop, etc.)

Types of cloud services: IaaS vs. PaaS vs. SaaS

Software as a Service (SaaS)

- The capability provided to the consumer is to use the provider's applications running on a cloud infrastructure. [NIST]
- E.g.:
  - Google Gmail, Microsoft Outlook.com, Google Drive, Dropbox, Google Docs, Microsoft Office 365, etc.
  - Databases (CosmosDB, DynamoDB, etc.), Coordination/queueing, etc.

The line between the different types of services is not well defined. E.g: map-reduce service: is it a PaaS or a SaaS ?

OUTLINE

Definitions and types

Motivation for cloud computing

Examples

A glimpse inside the infrastructure

WHY CLOUD COMPUTING?

Computing as a utility

- Ready to be used when needed.

Pay-as-you-go pricing

- No minimum or up-front payment.
- Pay what you use.

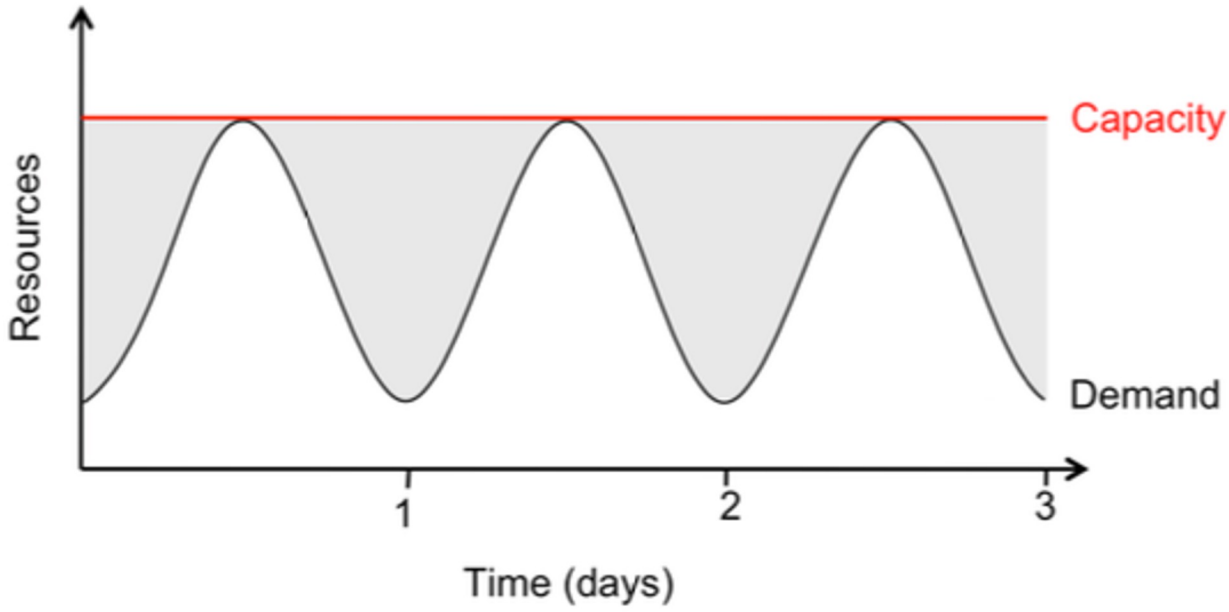
Use what you need

- Use what you need when you need.

WHY CLOUD COMPUTING? FOR USERS/DEVELOPERS (1)

Helpful for apps with variable utilization.

- Only pay what you use + use what you need.
- Allow to use and pay the resources needed in each moment.
- No need to for overprovision.



WHY CLOUD COMPUTING? FOR USERS/DEVELOPERS (2)

Helpful for deploying new apps.

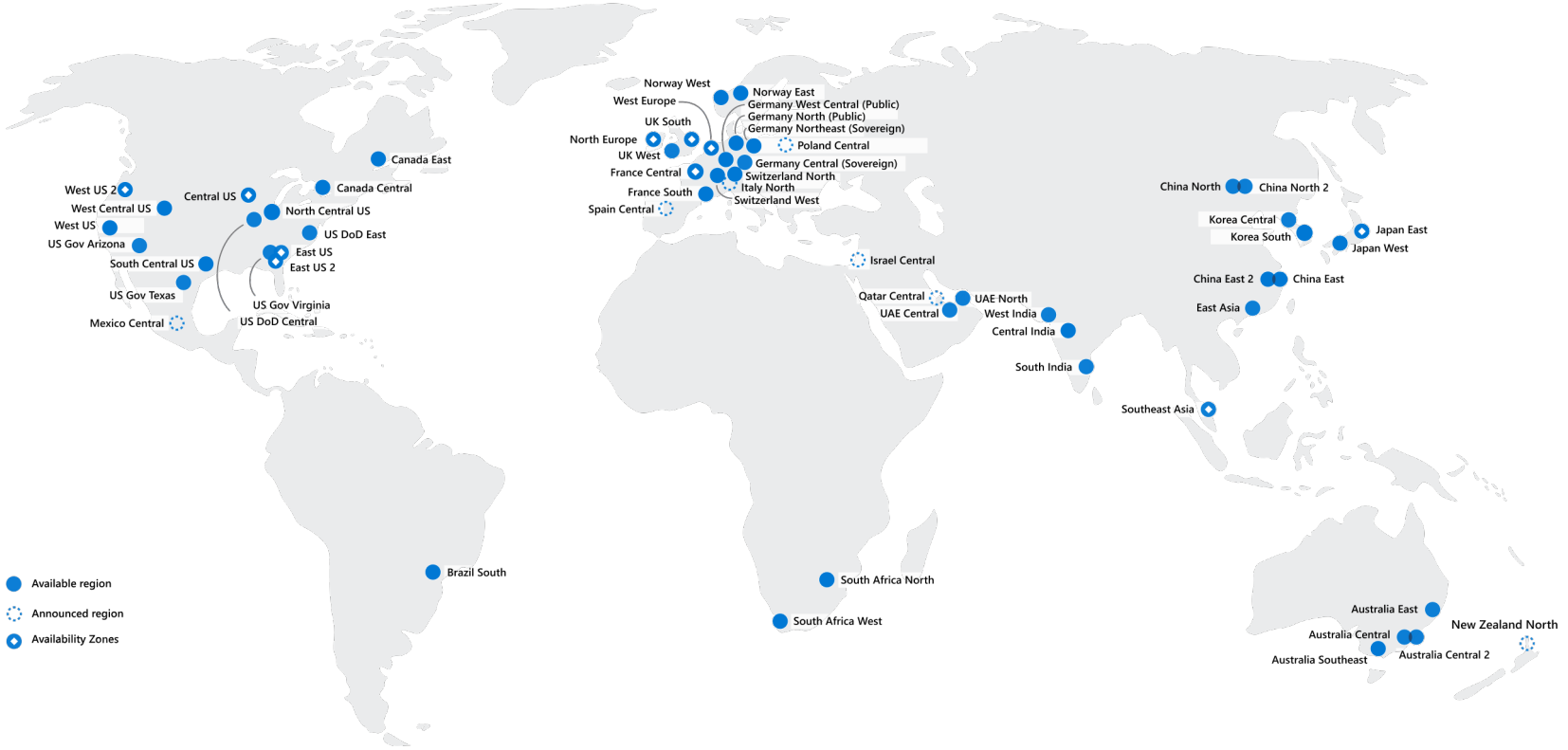
- Infrastructure ready for deploying new apps.

- Allow to deploy an application immediately.
  - No need to deploy new servers before start, etc.
- Allow to start small and scale when needed (elasticity).

WHY CLOUD COMPUTING? FOR USERS/DEVELOPERS (3)

Helpful for deploying global apps.

- Infrastructure distributed globally – increasing number of DCs and point-of-presence.
- Allow to provide low latency.

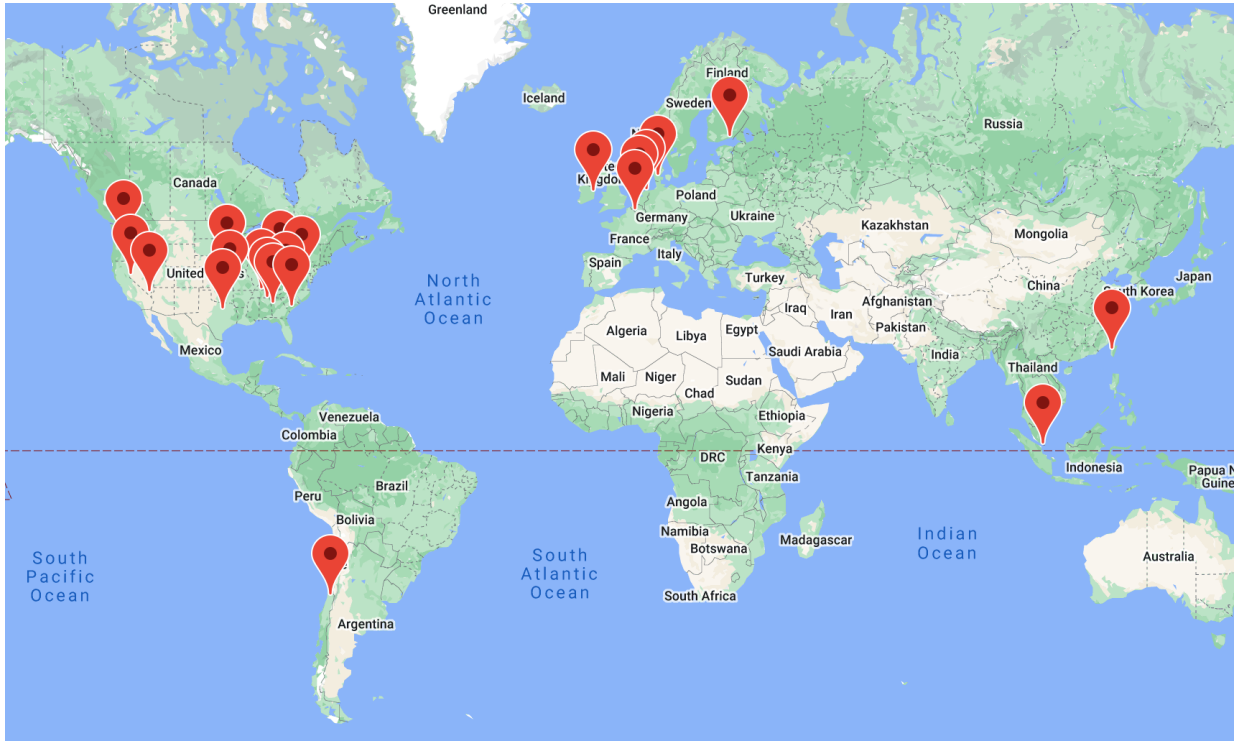


Azure locations: https://azure.microsoft.com/en-us/global-infrastructure/geographies/ (accessed: Sep, 2020)

WHY CLOUD COMPUTING? FOR USERS/DEVELOPERS (3.1)

Helpful for deploying global apps.

- Infrastructure distributed globally – increasing number of DCs and point-of-presence.
- Allow to provide low latency.

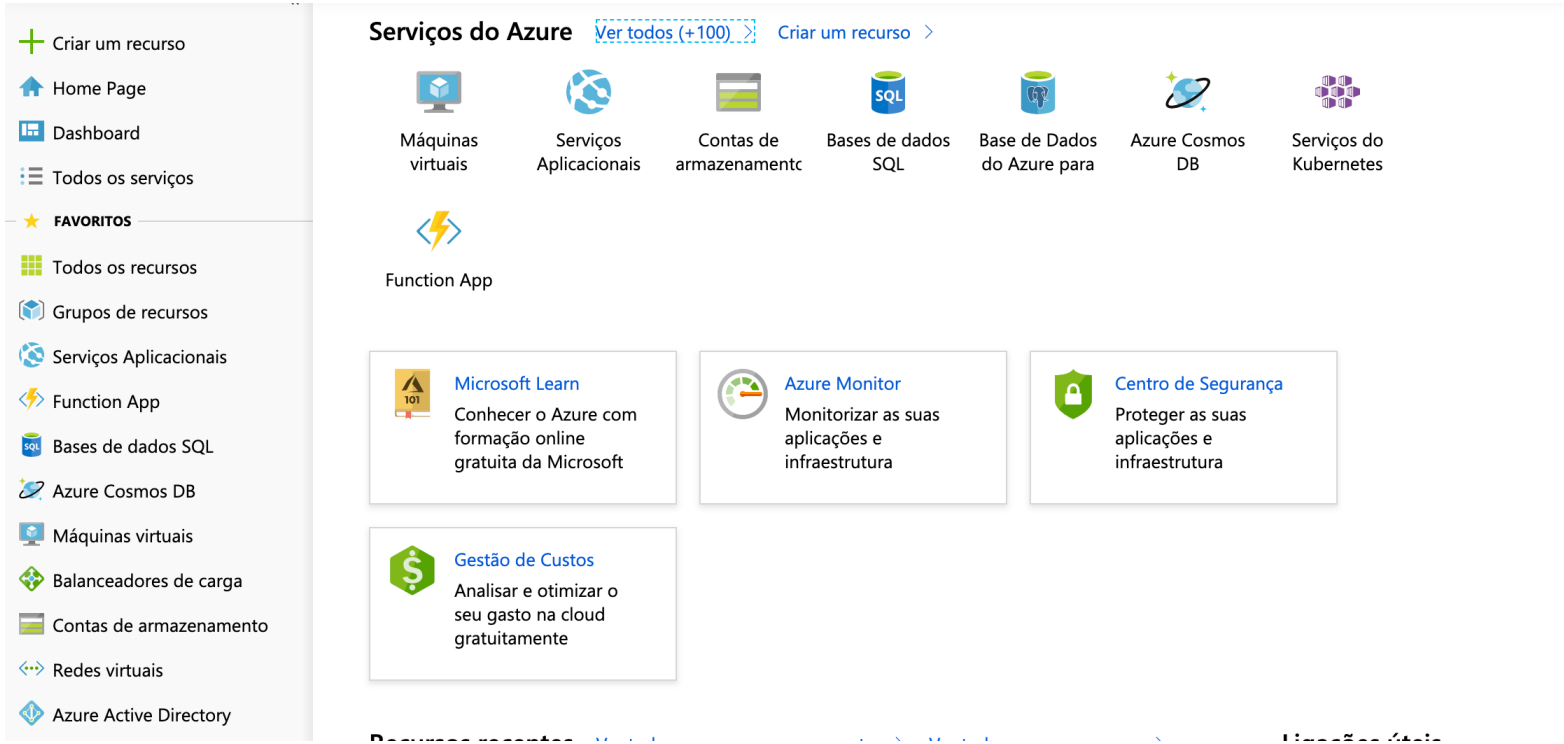


Google data centers (accessed: Sep, 2022)
: https://www.google.com/about/datacenters/locations/

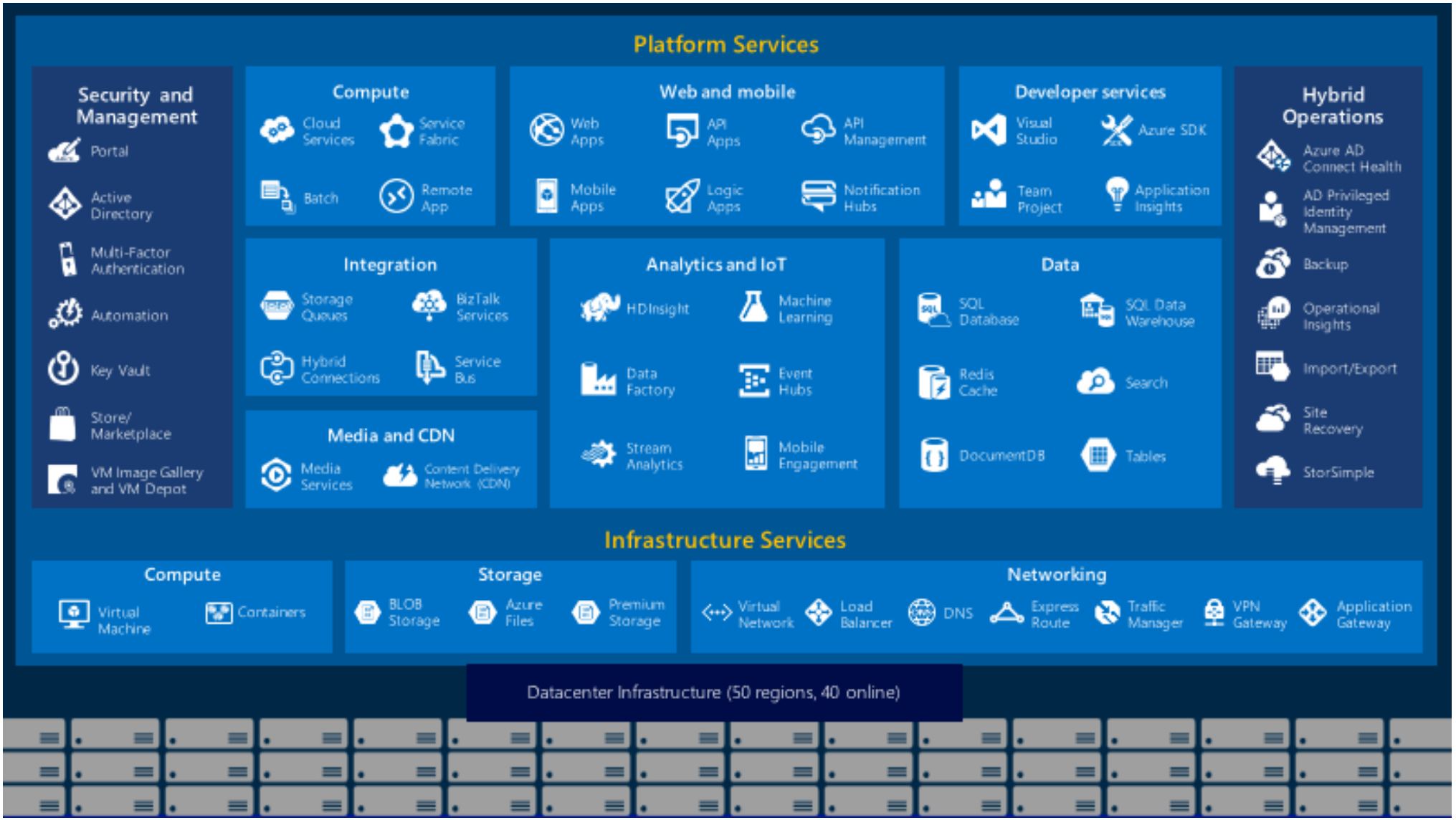
WHY CLOUD COMPUTING? THE CLOUD ECOSYSTEM

Helpful for deploying new apps.

- Ecosystem of cloud platforms is growing every day, with additional services being added.
- Easier to deploy new applications.
  - Developers can focus on implementing the business logic.



WHY CLOUD COMPUTING? THE CLOUD ECOSYSTEM



Why Cloud Computing? For providers (1)

Amazon (and others) had a problem

- Need to scale to support Christmas & Black Friday sales…
- … but did not require the same power in the remaining of the year.
- In 2010, estimates of average server utilization in data centers ranged from 5% to 20%.

- Is it possible to sell the unused computing power to other users?
- … and sell the services developed for running internal services?
- ... and build a business around this model?

Why Cloud Computing? For providers (2)

Economies of scale:

- Purchasing, powering & managing machines at scale gives lower per-unit costs than customers'
- Software/services can be reused

Why Cloud Computing? For providers (3)

Some challenges

- How do you avoid having many customers spiking at the same time?
- How to encourage customers to use resources when demand is low?

- E.g.: Amazon Spot instances – much lower price (market-based, up to 90% discount), but instances can be terminated if needed

CLOUD COMPUTING: SOME DATES

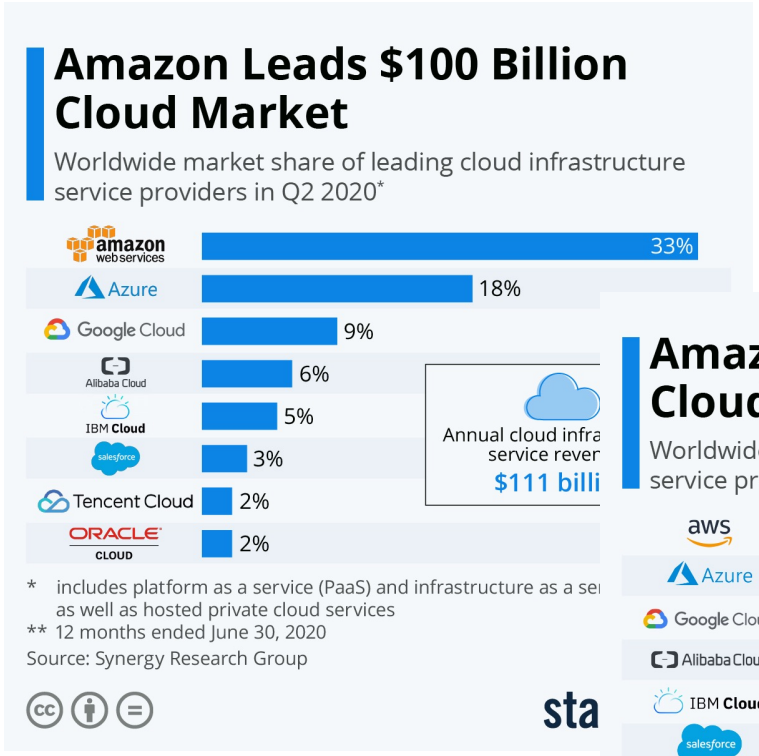
Amazon create Amazon Web Services subsidiary and introduced EC2 in 2006.

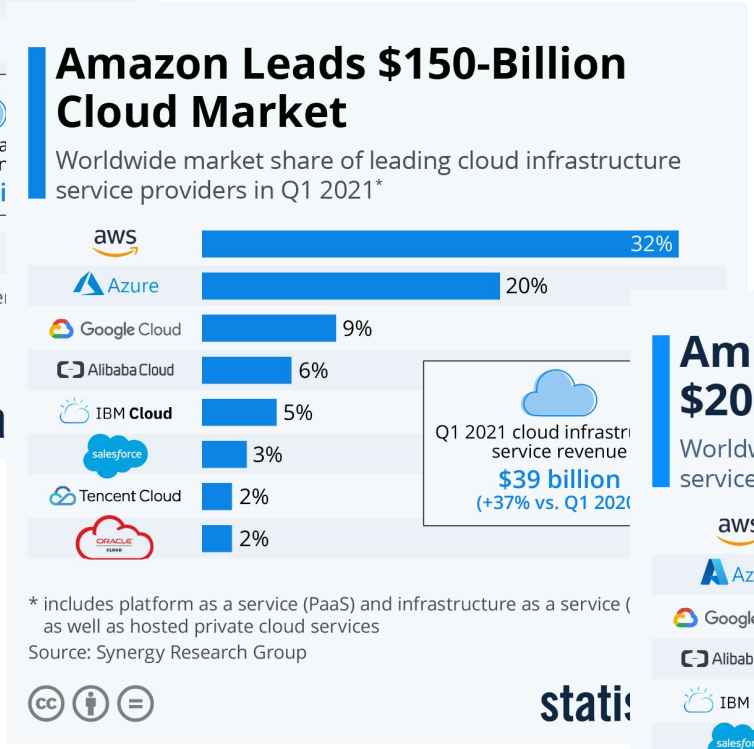
Google introduced Google App Engine in 2008.

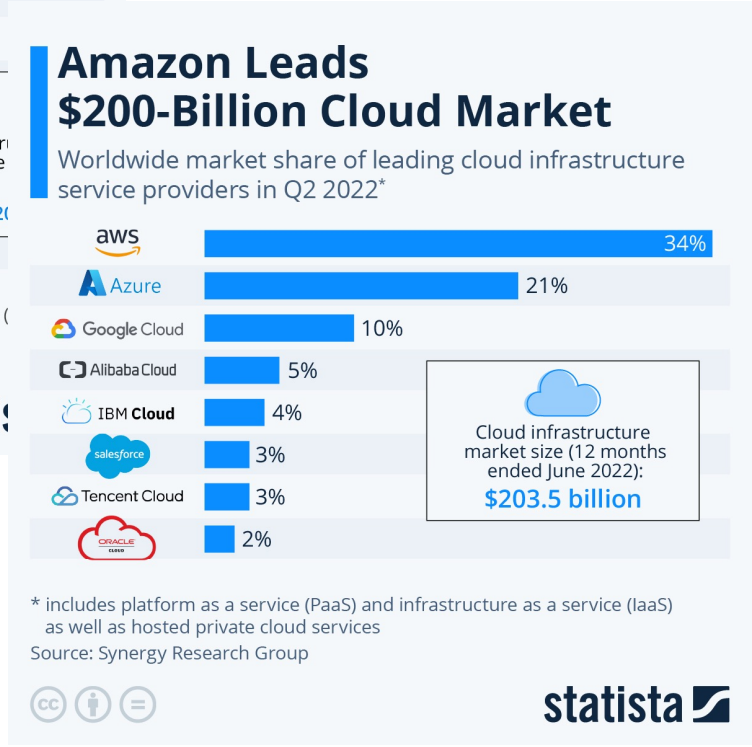
Azure was announced in 2008 and launched in 2010.

Google compute engine was introduced in 2012.

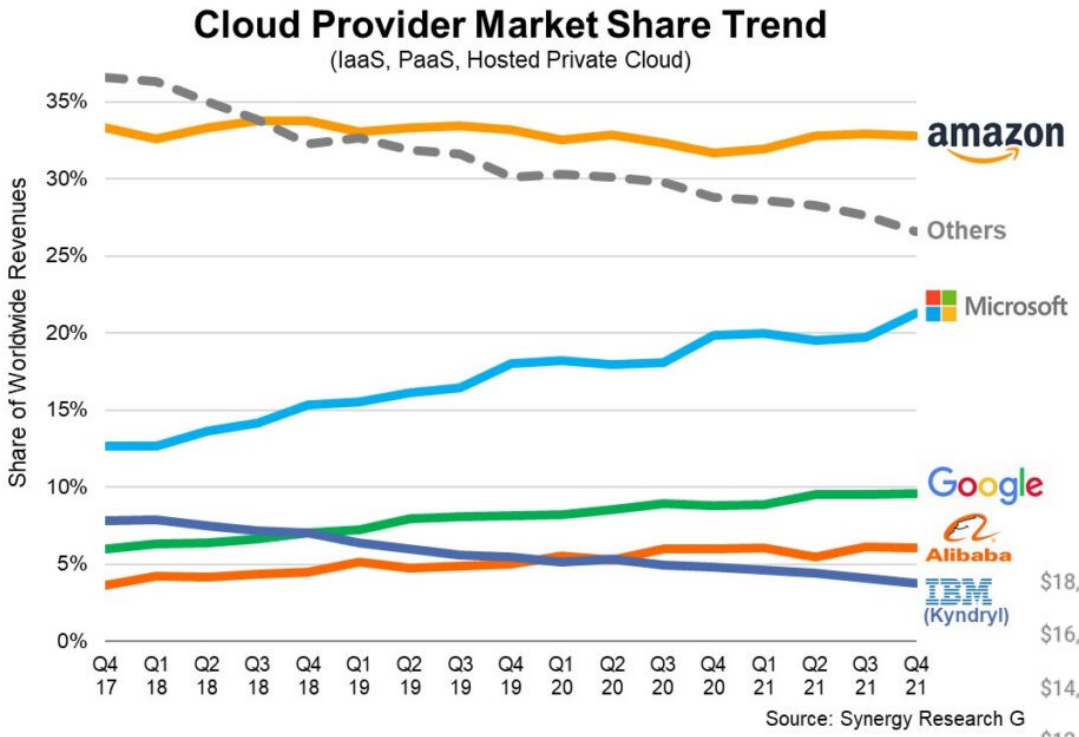
Cloud computing: Where is Europe?

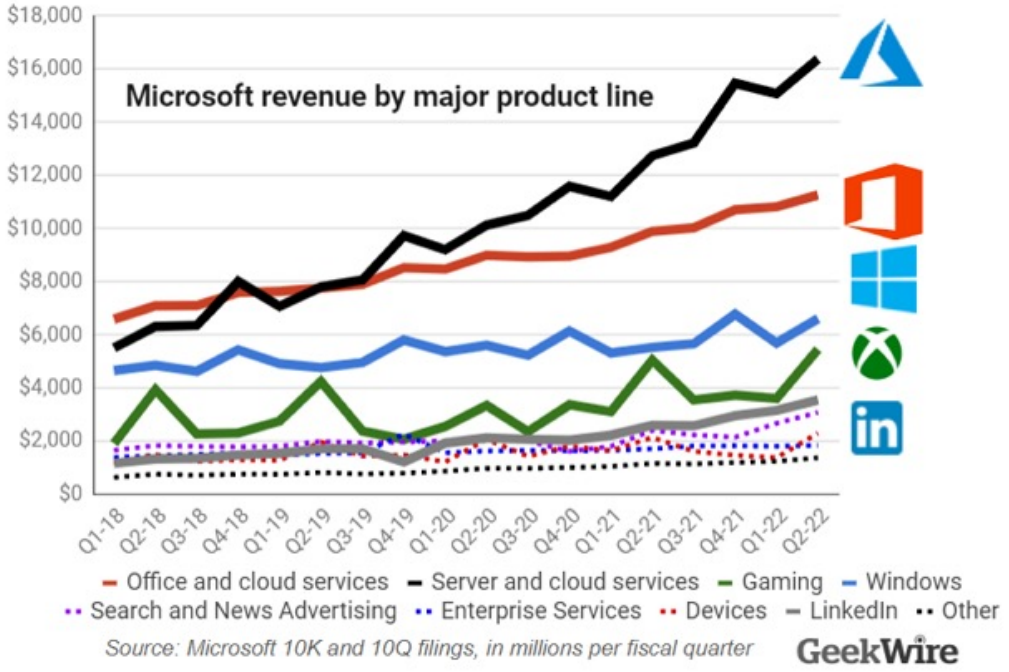






The Importance of Cloud computing





OUTLINE

Definitions and types

Motivation for cloud computing

Examples

A glimpse inside the infrastructure

Common Cloud Applications

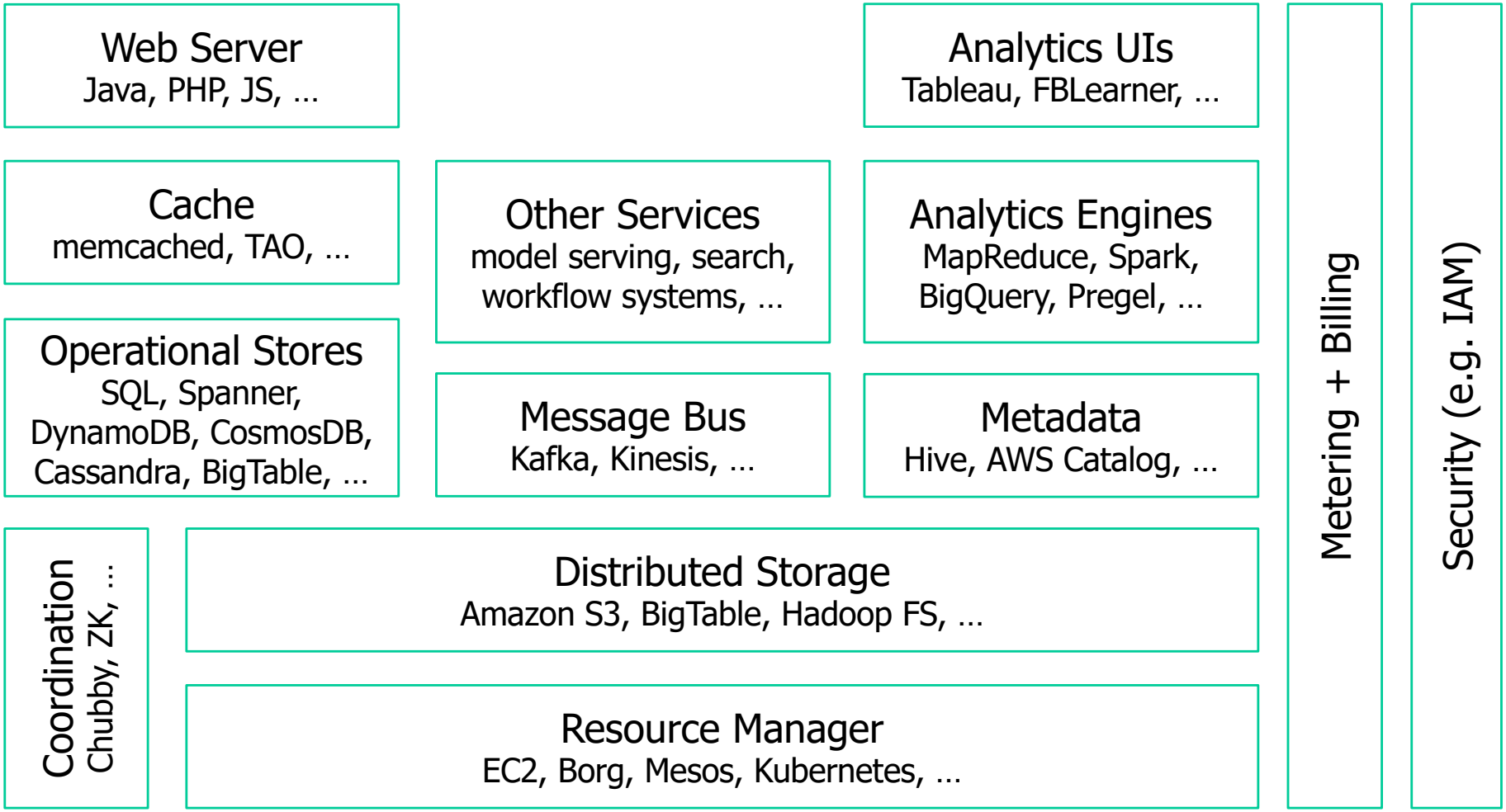
Web and mobile applications.

Data analytics (Spark, MapReduce, SQL, ML, etc.).

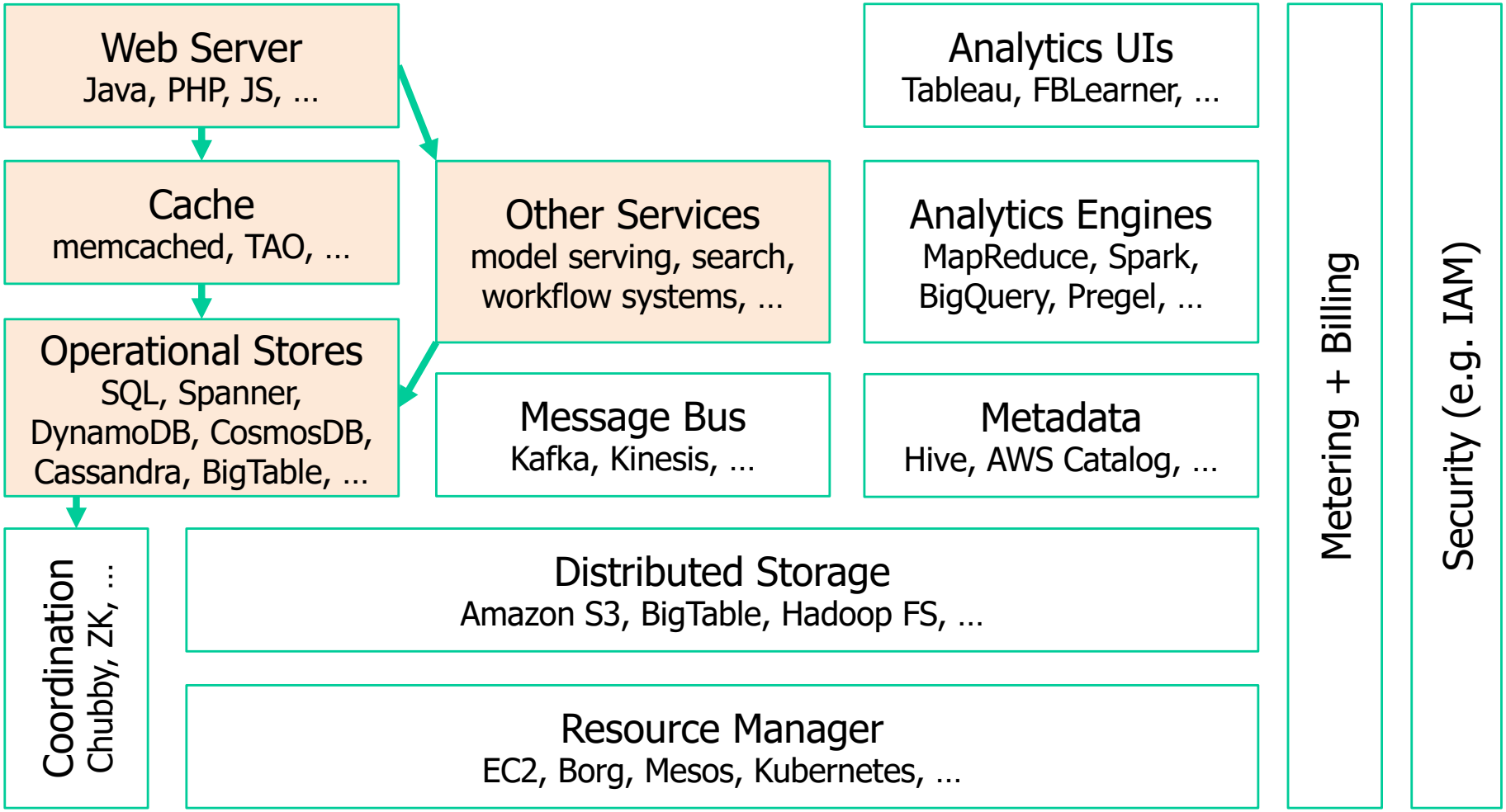
Stream processing.

Batch computation (HPC, video, etc).

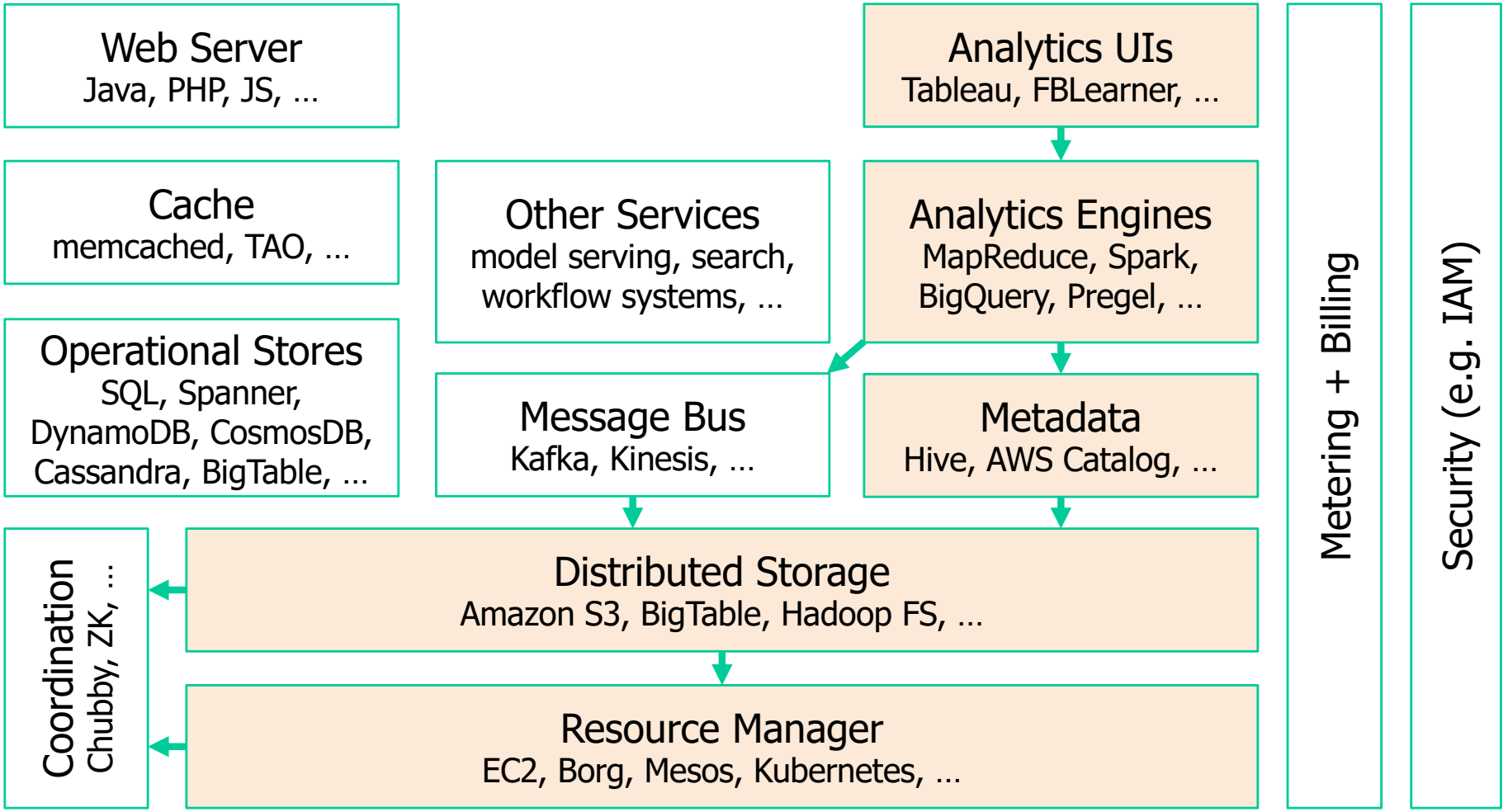
Cloud Software Stack



EXAMPLE: WEB APPLICATION



EXAMPLE: ANALYTICS WAREHOUSE



OUTLINE

Definitions and types

Motivation for cloud computing

Examples

A glimpse inside the infrastructure

BUILDING BLOCKS

Compute

Networking

Storage

BUILDING BLOCKS

Compute

- Basically a computer: processors (CPU, GPU, custom accelerators (AI), etc.), memory (RAM), channels to network and storage nodes.
- In a DC, a compute node is a server, usually mounted on a rack.

Networking

Storage

BUILDING BLOCKS

Compute

Networking

- Allows communications to flow among parties: compute nodes and, when used, external storage nodes.
- Network nodes are typically switches and routers.
- New: SDN, smart NICs, remote direct memory access.

Storage

BUILDING BLOCKS

Compute

Networking

Storage

- Stores "data" persistently in disk devices.
- It is either mounted inside a compute node, or outside.
- E.g:
  - Disk trays
  - SSD & NVM Flash
- New: non-volatile memory, novel storage technologies.

Useful Latency Numbers (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
Send 2K bytes over 10Ge	2,000 ns
Read 1 MB sequentially from memory	100,000 ns
Round trip within same datacenter	500,000 ns
Read 1 MB sequentially from SSD*	1,000,000 ns
Read 1 MB sequentially from disk	20,000,000 ns
Send packet CA → Europe → CA	150,000,000 ns

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)

Data from Christos Kozyrakis & Matei Zaharia

SOME CONSEQUENCE

Co-location is very important for efficiency.

Costly to access a single location.

Costly to coordinate at a global scale.

Need fault-tolerant mechanisms.

DATACENTER ORGANIZATION

Rows of rack-mounted servers

Datacenter: 50 – 200K of servers

- Often organized as few and mostly independent clusters

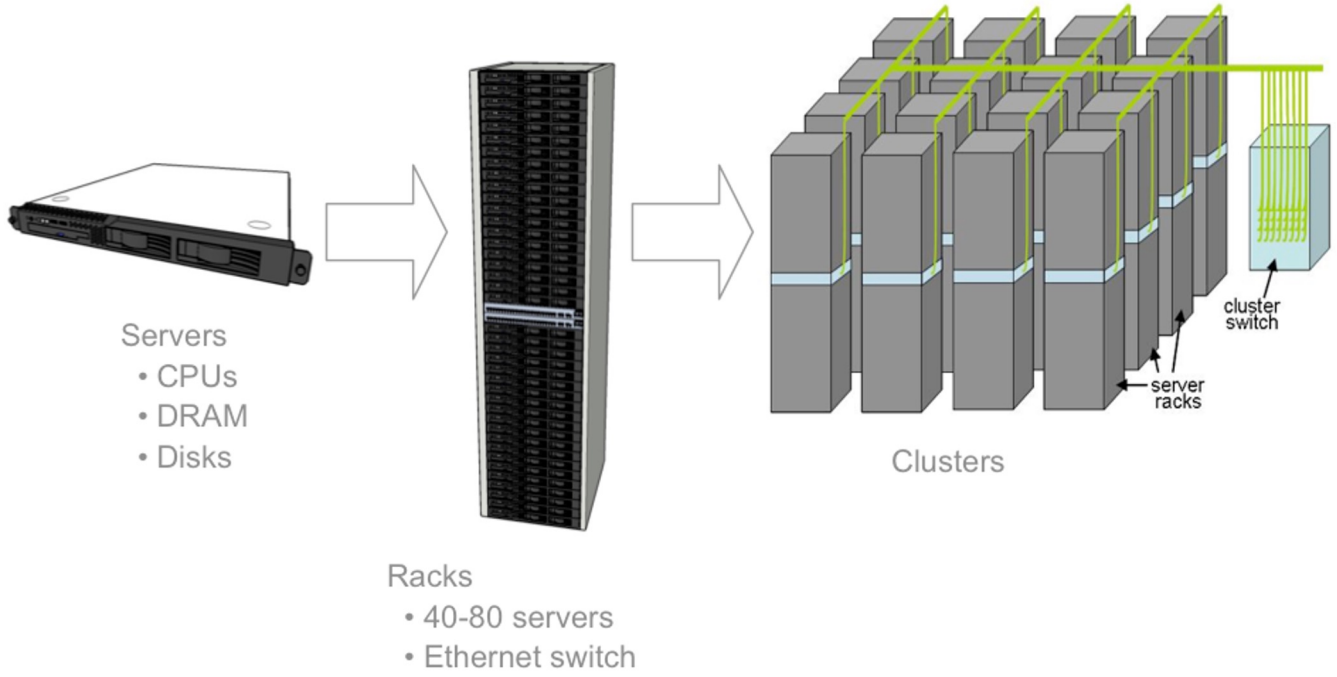
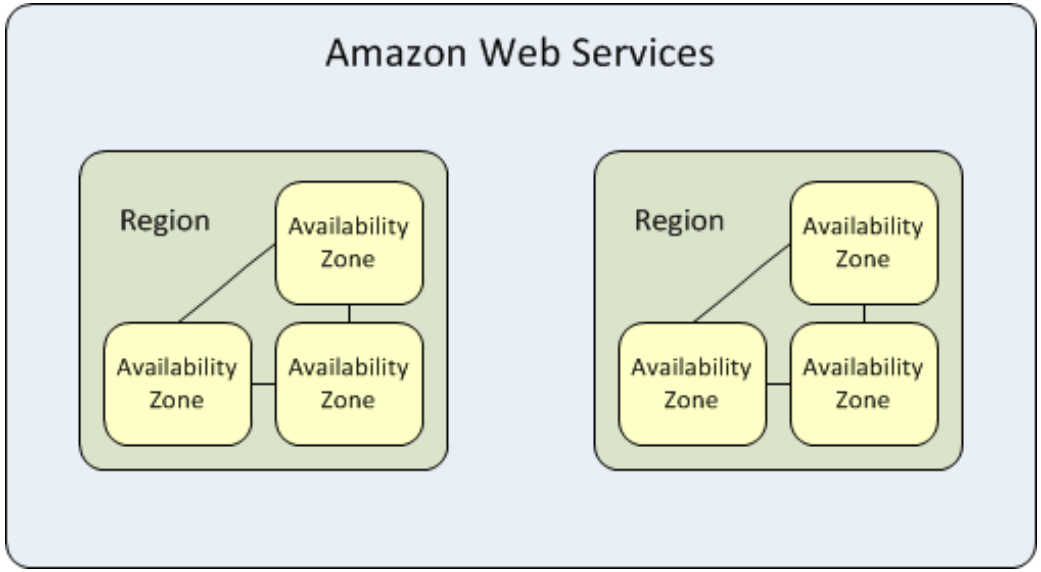


Image from Christos Kozyrakis & Matei Zaharia

REGIONS AND AVAILABILITY ZONES

Cloud infrastructure are often organized in regions and availability zones



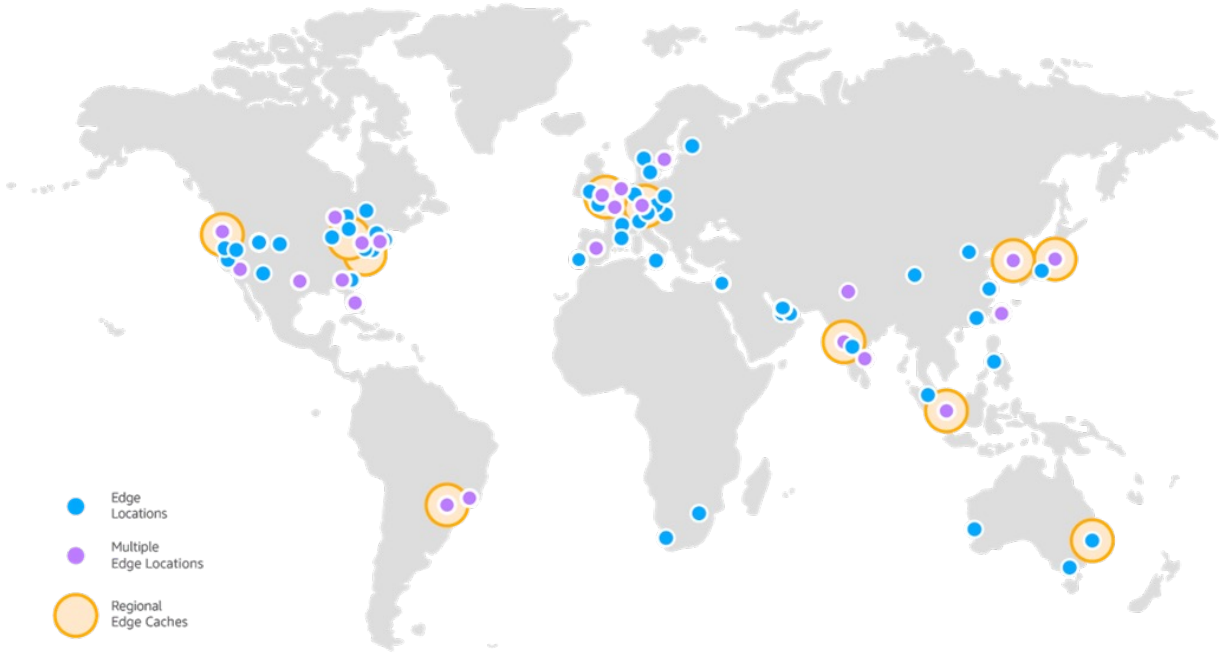
A *Region* is a separate geographic area. Each Region has multiple, isolated locations known as *Availability Zones*. Locations are connected through a dedicated regional low-latency network.

Availability Zones are physically separate locations, consisting in one or more datacenters equipped with independent power, cooling, and networking.

EDGE LOCATION

Cloud infrastructure also have edge locations.

- Programmable content-distribution networks.
- Serverless functions.



Amazon edge locations, Sep. 2019

TO KNOW MORE

M. Armbrust, et. al. A View of Cloud Computing. CACM, Apr. 2010.