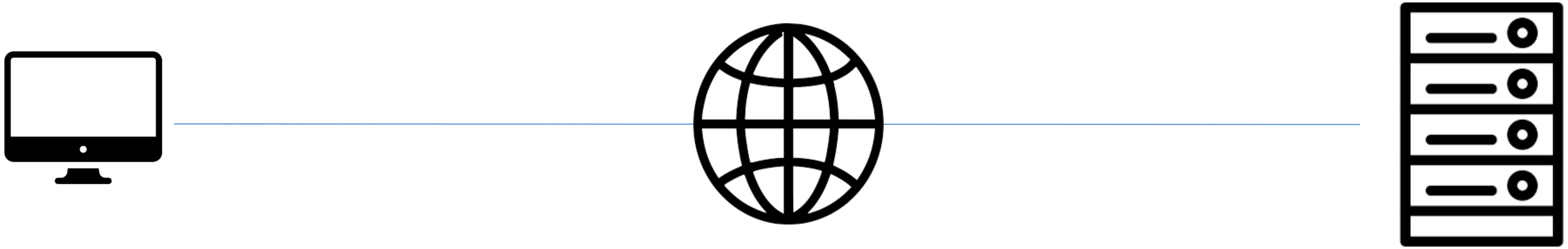
What is Cloud Computing

How websites work



network

Client

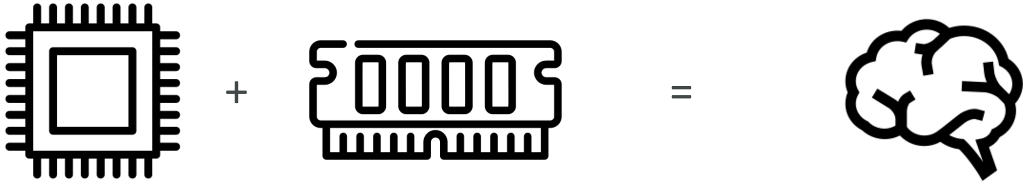
Server

Clients have IP addresses

Servers have IP addresses

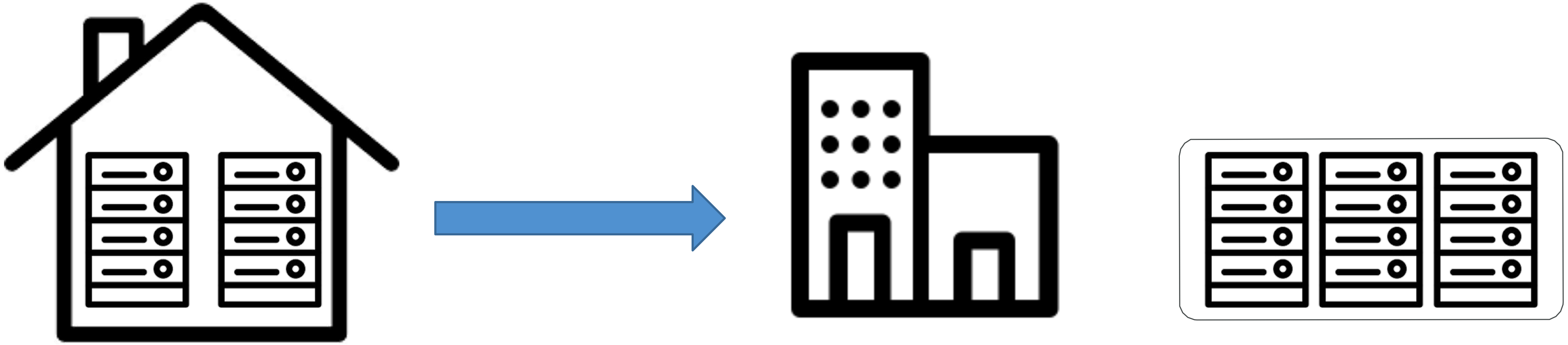
What is a server composed of?

- Compute: CPU
- Memory: RAM



- Storage: Data
- Database: Store data in a structured way
- Network: Routers, switch, DNS server

Traditionally, how to build infrastructure



Home or Garage

Office

Data center

Problems with traditional IT approach

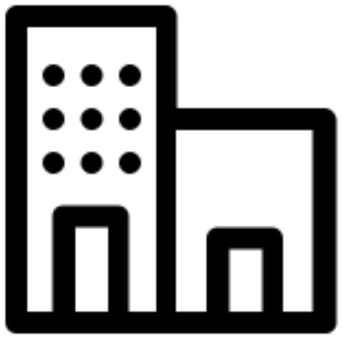
- Pay for the rent for the data center
- Pay for power supply, cooling, maintenance
- Adding and replacing hardware takes time
- Scaling is limited
- Hire 24/7 team to monitor the infrastructure
- How to deal with disasters? (earthquake, power shutdown, fire…)

- Can we externalize all this?

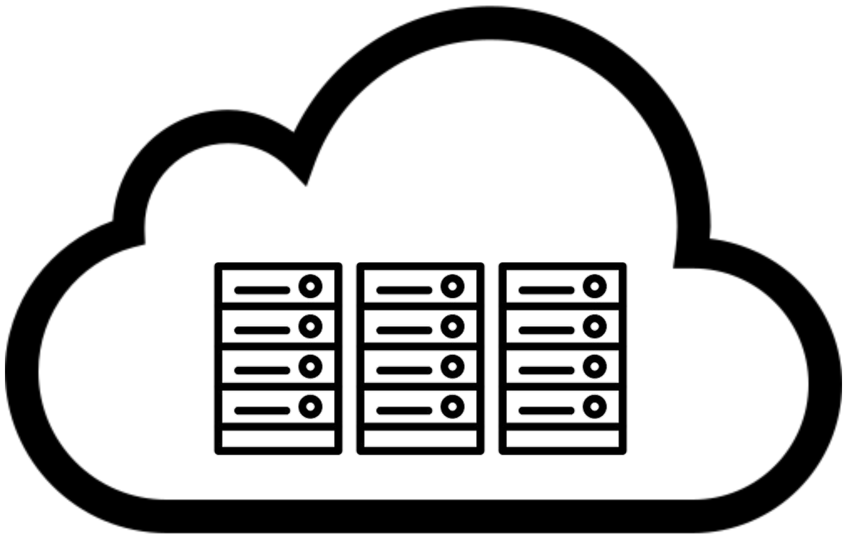
What is Cloud Computing?

- Cloud computing is the on-demand delivery of compute power, database storage, applications, and other IT resources
- Through a cloud services platform with pay-as-you-go pricing
- You can provision exactly the right type and size of computing resources you need
- You can access as many resources as you need, almost instantly
- Simple way to access servers, storage, databases and a set of application services

- Amazon Web Services owns and maintains the network-connected hardware required for these application services, while you provision and use what you need via a web application.



Office



The Cloud

You’ve been using some Cloud services

Gmail

- E-mail cloud service
- Pay for ONLY your emails stored (no infrastructure, etc.)

Dropbox

- Cloud Storage Service
- Originally built on AWS

Netflix

- Built on AWS
- Video on Demand

The Deployment Models of the Cloud

Private Cloud:

- Cloud services used by a single organization, not exposed to the public.
- Complete control
- Security for sensitive applications
- Meet specific business needs

Public Cloud:

- Cloud resources owned and operated by a third-party cloud service provider delivered over the Internet.
- Six Advantages of Cloud Computing

Hybrid Cloud:

- Keep some servers on premises and extend some capabilities to the Cloud
- Control over sensitive assets in your private infrastructure
- Flexibility and cost-effectiveness of the public cloud

The Five Characteristics of Cloud Computing

- On-demand self service:
  - Users can provision resources and use them without human interaction from the service provider
- Broad network access:
  - Resources available over the network, and can be accessed by diverse client platforms
- Multi-tenancy and resource pooling:
  - Multiple customers can share the same infrastructure and applications with security and privacy
  - Multiple customers are serviced from the same physical resources
- Rapid elasticity and scalability:
  - Automatically and quickly acquire and dispose resources when needed
  - Quickly and easily scale based on demand
- Measured service:
  - Usage is measured, users pay correctly for what they have used

Six Advantages of Cloud Computing

- Trade capital expense (CAPEX) for operational expense (OPEX)
    - Pay On-Demand: don't own hardware
    - Reduced Total Cost of Ownership (TCO) & Operational Expense (OPEX)
- Benefit from massive economies of scale
    - Prices are reduced as AWS is more efficient due to large scale
- Stop guessing capacity
    - Scale based on actual measured usage
- Increase speed and agility
- Stop spending money running and maintaining data centers
- Go global in minutes: leverage the AWS global infrastructure

Problems solved by the Cloud

- Flexibility: change resource types when needed
- Cost-Effectiveness: pay as you go, for what you use
- Scalability: accommodate larger loads by making hardware stronger or adding additional nodes
- Elasticity: ability to scale out and scale-in when needed
- High-availability and fault-tolerance: build across data centers
- Agility: rapidly develop, test and launch software applications

Types of Cloud Computing

- Infrastructure as a Service (IaaS)
  - Provide building blocks for cloud IT
  - Provides networking, computers, data storage space
  - Highest level of flexibility
  - Easy parallel with traditional on-premises IT
- Platform as a Service (PaaS)
  - Removes the need for your organization to manage the underlying infrastructure
  - Focus on the deployment and management of your applications
- Software as a Service (SaaS)
  - Completed product that is run and managed by the service provider

On-premises	Infrastructure as a Service (IaaS)	Platform as a Service (PaaS)	Software as a Service (SaaS)
🟦 Applications	🟦 Applications	🟦 Applications	🟧 Applications
🟦 Data	🟦 Data	🟦 Data	🟧 Data
🟦 Runtime	🟦 Runtime	🟧 Runtime	🟧 Runtime
🟦 Middleware	🟦 Middleware	🟧 Middleware	🟧 Middleware
🟦 O/S	🟦 O/S	🟧 O/S	🟧 O/S
🟦 Virtualization	🟧 Virtualization	🟧 Virtualization	🟧 Virtualization
🟦 Servers	🟧 Servers	🟧 Servers	🟧 Servers
🟦 Storage	🟧 Storage	🟧 Storage	🟧 Storage
🟦 Networking	🟧 Networking	🟧 Networking	🟧 Networking

🟦 Managed by you

🟧 Managed by others

Example of Cloud Computing Types

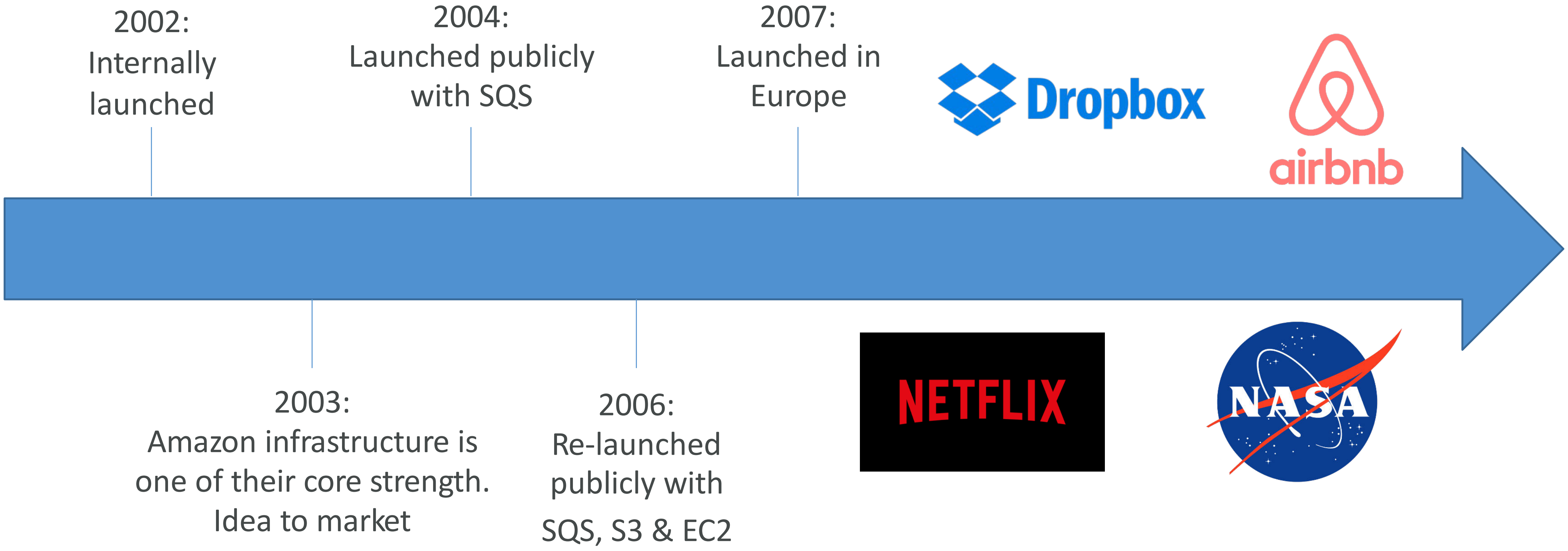
- Infrastructure as a Service:
  - Amazon EC2 (on AWS)
  - GCP, Azure, Rackspace, Digital Ocean, Linode
- Platform as a Service:
  - Elastic Beanstalk (on AWS)
  - Heroku, Google App Engine (GCP), Windows Azure (Microsoft)
- Software as a Service:
  - Many AWS services (ex: Rekognition for Machine Learning)
  - Google Apps (Gmail), Dropbox, Zoom

Pricing of the Cloud – Quick Overview

AWS has 3 pricing fundamentals, following the pay-as-you-go pricing model

- Compute:
  - Pay for compute time
- Storage:
  - Pay for data stored in the Cloud
- Data transfer OUT of the Cloud:
  - Data transfer IN is free

AWS Cloud History



2002: Internally launched

2003: Amazon infrastructure is one of their core strength. Idea to market

2004: Launched publicly with SQS

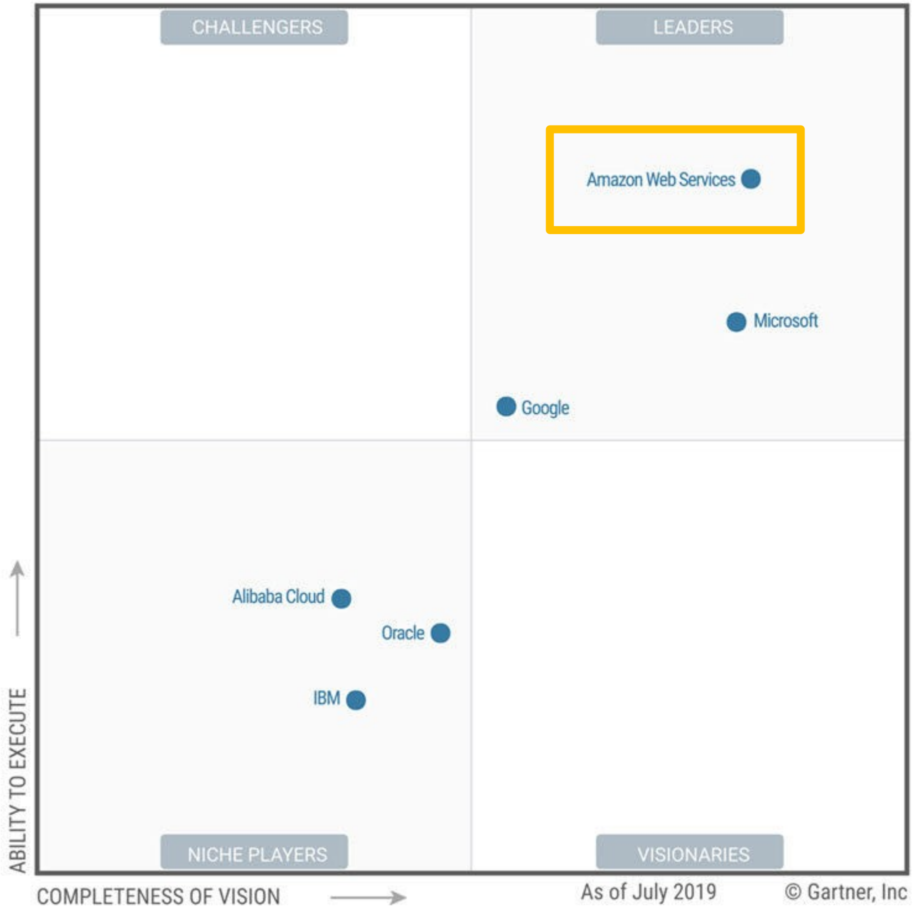
2006: Re-launched publicly with SQS, S3 & EC2

2007: Launched in Europe

AWS Cloud Number Facts

- In 2019, AWS had $35.02 billion in annual revenue
- AWS accounts for 47% of the market in 2019 (Microsoft is 2nd with 22%)
- Pioneer and Leader of the AWS Cloud Market for the 9th consecutive year
- Over 1,000,000 active users

Figure 1. Magic Quadrant for Cloud Infrastructure as a Service, Worldwide



Source: Gartner (July 2019)

Gartner Magic Quadrant

AWS Cloud Use Cases

- AWS enables you to build sophisticated, scalable applications
- Applicable to a diverse set of industries
- Use cases include
  - Enterprise IT, Backup & Storage, Big Data analytics
  - Website hosting, Mobile & Social Apps
  - Gaming

21ST CENTURY FOX

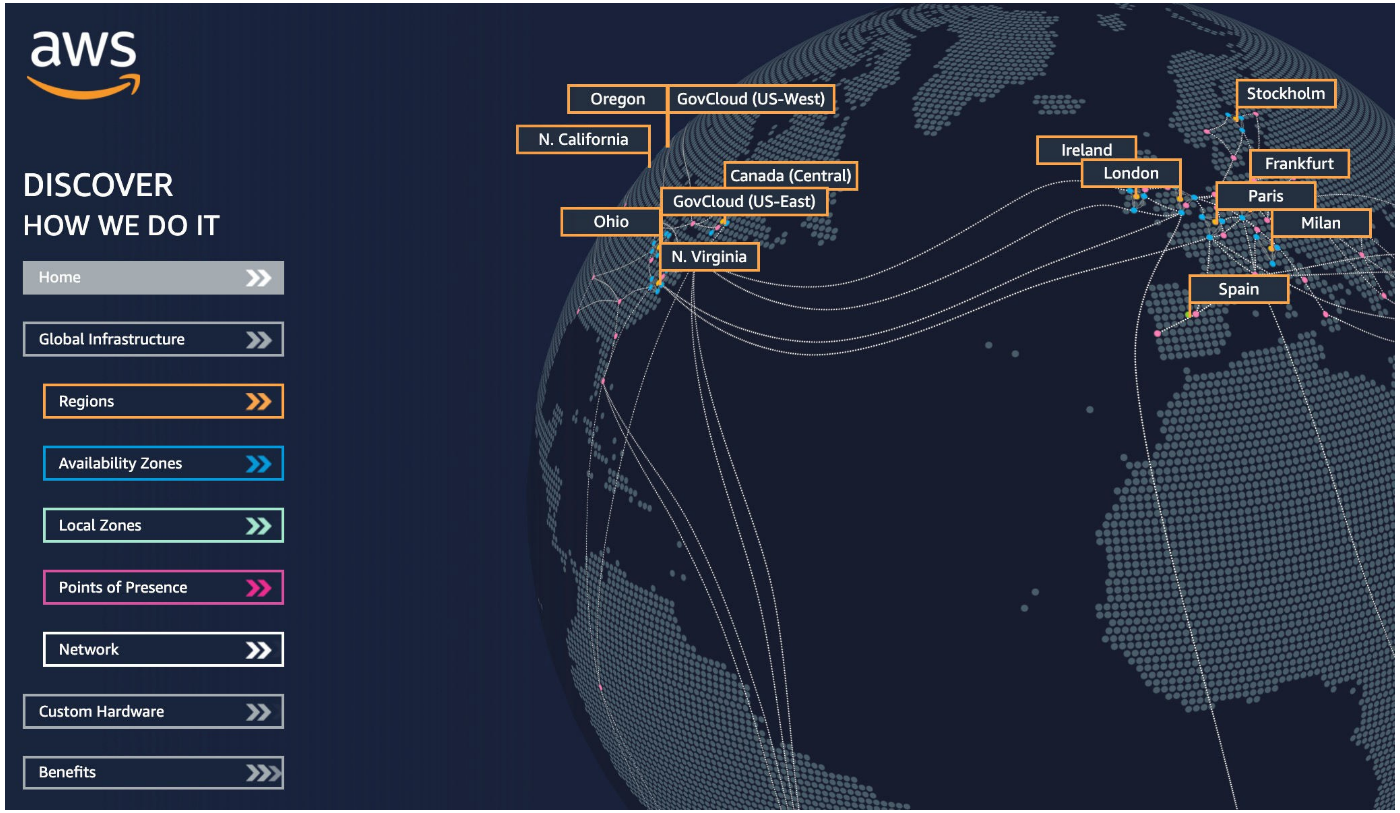
ACTIVISION

NETFLIX

AWS Global Infrastructure

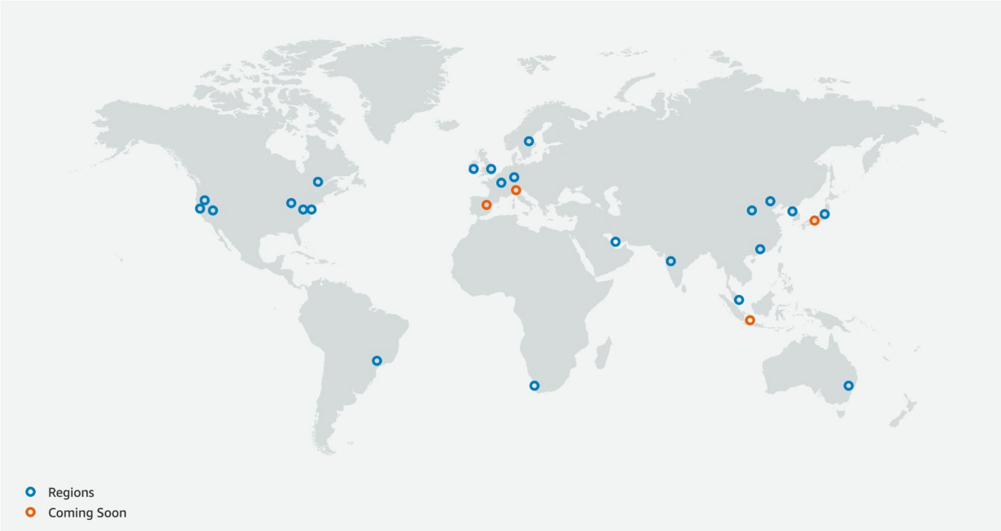
- AWS Regions
- AWS Availability Zones
- AWS Edge Locations / Points of Presence

- https://infrastructure.aws/

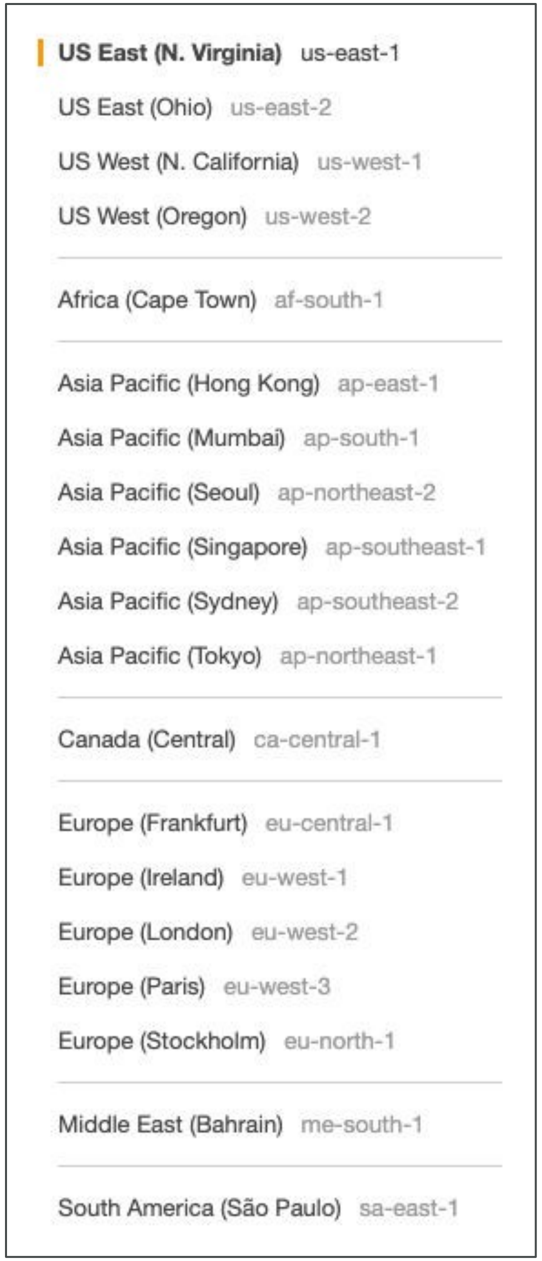


AWS Regions

- AWS has Regions all around the world
- Names can be us-east-1, eu-west-3…
- A region is a cluster of data centers
- Most AWS services are region-scoped



https://aws.amazon.com/about-aws/global-infrastructure/



US East (N. Virginia)	us-east-1
US East (Ohio)	us-east-2
US West (N. California)	us-west-1
US West (Oregon)	us-west-2
Africa (Cape Town)	af-south-1
Asia Pacific (Hong Kong)	ap-east-1
Asia Pacific (Mumbai)	ap-south-1
Asia Pacific (Seoul)	ap-northeast-2
Asia Pacific (Singapore)	ap-southeast-1
Asia Pacific (Sydney)	ap-southeast-2
Asia Pacific (Tokyo)	ap-northeast-1
Canada (Central)	ca-central-1
Europe (Frankfurt)	eu-central-1
Europe (Ireland)	eu-west-1
Europe (London)	eu-west-2
Europe (Paris)	eu-west-3
Europe (Stockholm)	eu-north-1
Middle East (Bahrain)	me-south-1
South America (São Paulo)	sa-east-1

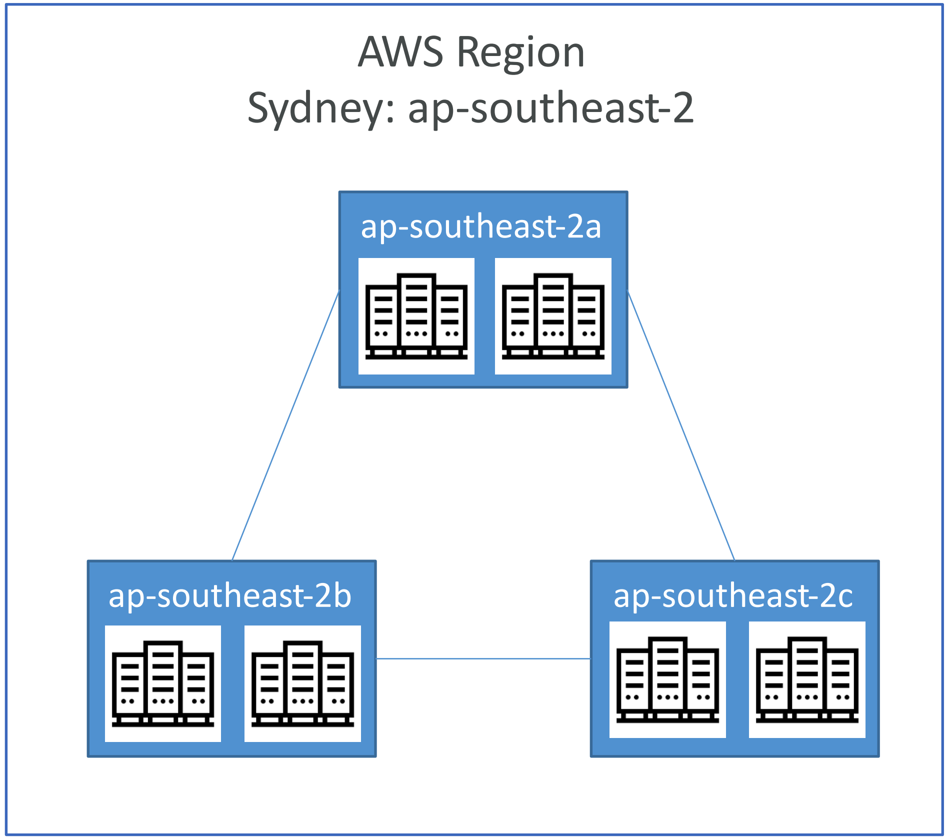
How to choose an AWS Region?

If you need to launch a new application, where should you do it?

- Compliance with data governance and legal requirements: data never leaves a region without your explicit permission
- Proximity to customers: reduced latency
- Available services within a Region: new services and new features aren’t available in every Region
- Pricing: pricing varies region to region and is transparent in the service pricing page

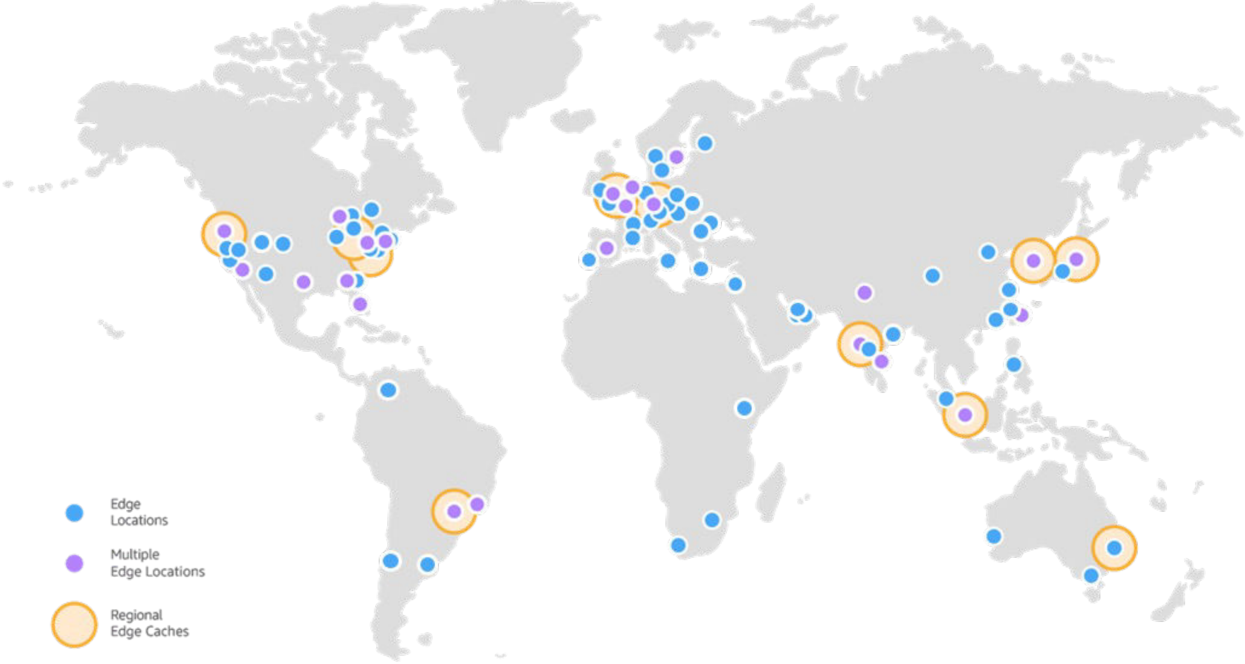
AWS Availability Zones

- Each region has many availability zones (usually 3, min is 2, max is 6). Example:
  - ap-southeast-2a
  - ap-southeast-2b
  - ap-southeast-2c
- Each availability zone (AZ) is one or more discrete data centers with redundant power, networking, and connectivity
- They're separate from each other, so that they're isolated from disasters
- They're connected with high bandwidth, ultra-low latency networking



AWS Points of Presence (Edge Locations)

- Amazon has 216 Points of Presence (205 Edge Locations & 11 Regional Caches) in 84 cities across 42 countries
- Content is delivered to end users with lower latency



https://aws.amazon.com/cloudfront/features/

Edge Locations & Regional Caches

- Edge locations are AWS data centers designed to deliver services with the lowest latency possible. Amazon has dozens of these data centers spread across the world. They’re closer to users than Regions or Availability Zones, often in major cities, so responses can be fast and snappy.
- The regional edge caches sit between the origin server and the edge POPs. If content isn’t cached in a particular edge POP, it can be retrieved from the regional edge cache without going back to the origin server.
- For example, England has one regional edge cache in London, and 11 edge POPs spread across the country. If a user in Manchester visits a site, CloudFront will first try the cache in their nearest edge POP, then the cache in the London REC; only if that doesn’t work will it go back to the origin server.

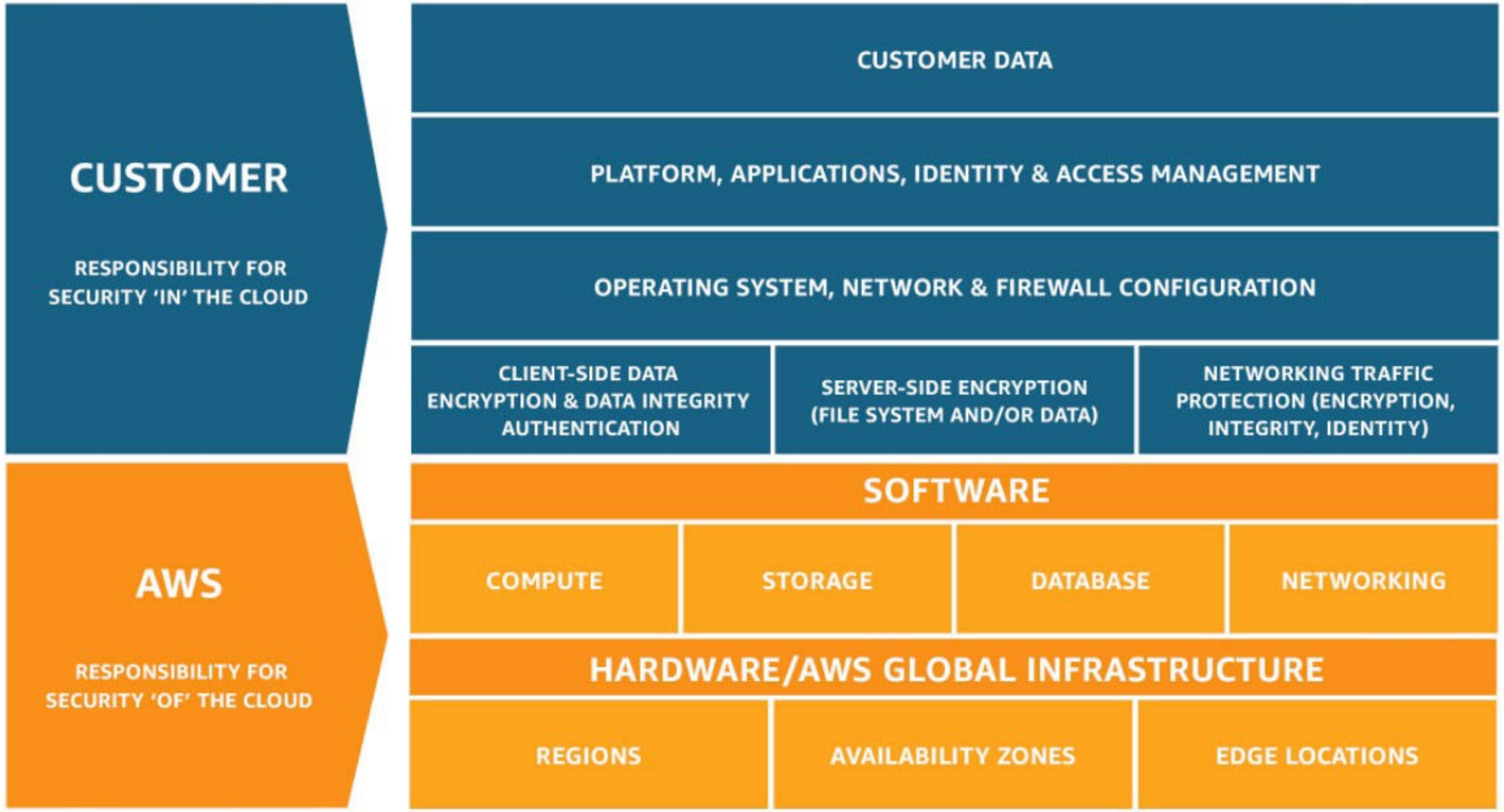
Tour of the AWS Console

- AWS has Global Services:
  - Identity and Access Management (IAM)
  - Route 53 (DNS service)
  - CloudFront (Content Delivery Network)
  - WAF (Web Application Firewall)
- Most AWS services are Region-scoped:
  - Amazon EC2 (Infrastructure as a Service)
  - Elastic Beanstalk (Platform as a Service)
  - Lambda (Function as a Service)
  - Rekognition (Software as a Service)
- Region Table: https://aws.amazon.com/about-aws/global-infrastructure/regional-product-services/?p=ngi&loc=4

Shared Responsibility Model diagram

CUSTOMER = RESPONSIBILITY FOR THE SECURITY **IN** THE CLOUD

AWS = RESPONSIBILITY FOR THE SECURITY **OF** THE CLOUD



CUSTOMER

RESPONSIBILITY FOR SECURITY 'IN' THE CLOUD

CUSTOMER DATA

PLATFORM, APPLICATIONS, IDENTITY & ACCESS MANAGEMENT

OPERATING SYSTEM, NETWORK & FIREWALL CONFIGURATION

CLIENT-SIDE DATA ENCRYPTION & DATA INTEGRITY AUTHENTICATION	SERVER-SIDE ENCRYPTION (FILE SYSTEM AND/OR DATA)	NETWORKING TRAFFIC PROTECTION (ENCRYPTION, INTEGRITY, IDENTITY)

AWS

RESPONSIBILITY FOR SECURITY 'OF' THE CLOUD

SOFTWARE

COMPUTE	STORAGE	DATABASE	NETWORKING

HARDWARE/AWS GLOBAL INFRASTRUCTURE

REGIONS	AVAILABILITY ZONES	EDGE LOCATIONS

https://aws.amazon.com/compliance/shared-responsibility-model/