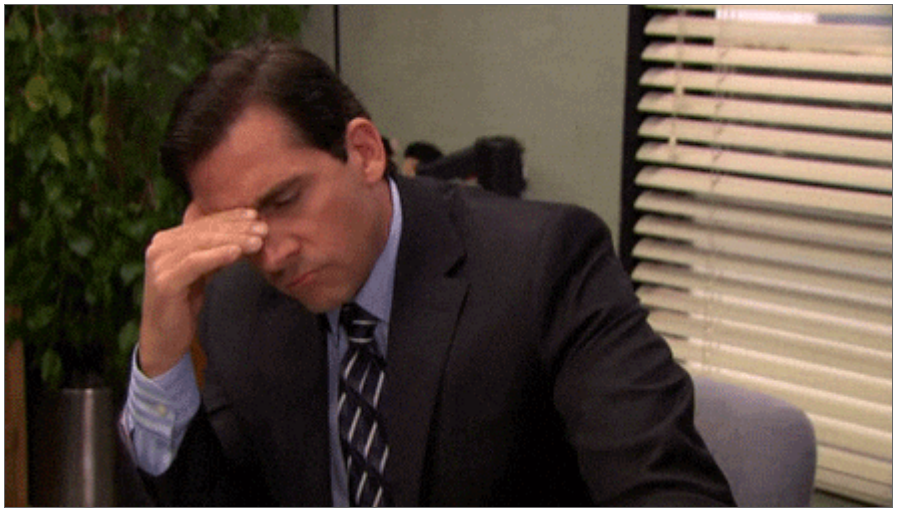
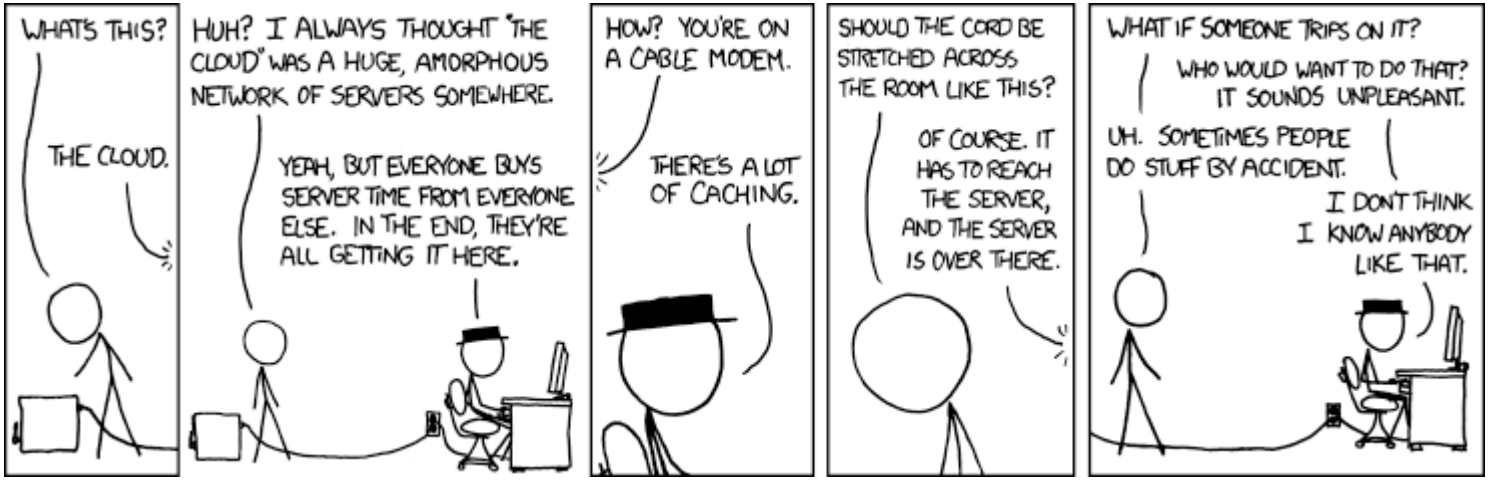
INTRO TO CLOUD COMPUTING



SDD - Data Engineering
https://supaerodatascience.github.io/deep-learning/

WHAT IS THE CLOUD ?




There is no cloud
It's just someone else's computer

But it's a bit bigger...

(Facebook's data center & server racks)

datacenters

Google Cloud Platform datacenters locations

The cloud is a real physical place - accessed over the internet - where a service is performed for you or where your stuff is stored. Your stuff is stored in the cloud, not on your device because the cloud is not on any device; the cloud lives in datacenters. A program running on your device accesses the cloud over the internet. The cloud is infinite, accessible from anywhere, at any time

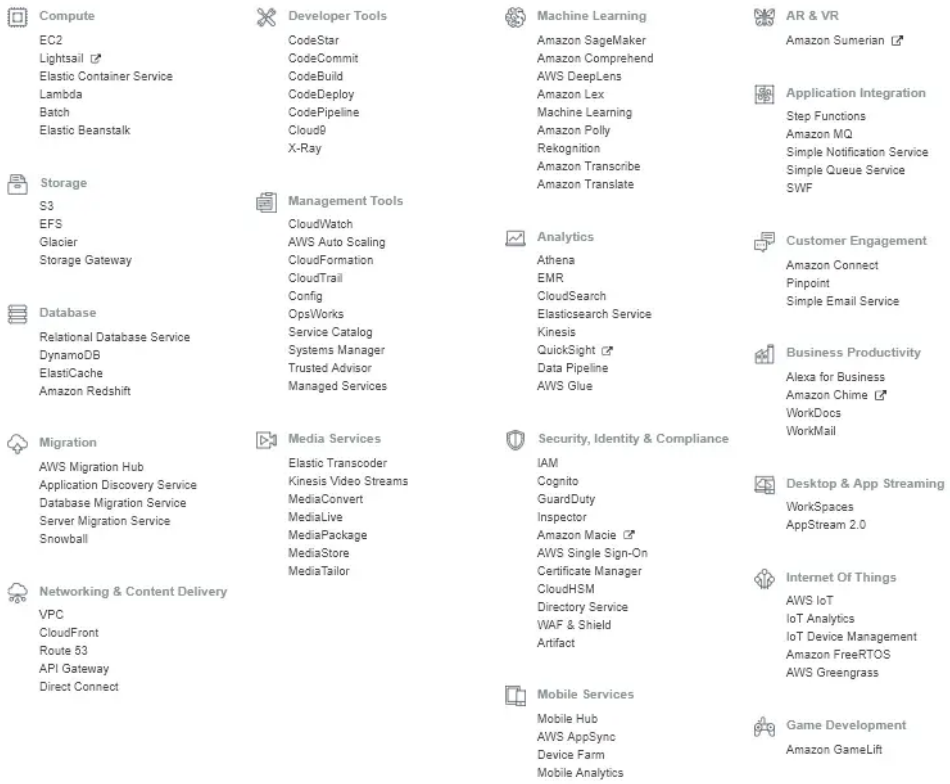
Todd Hoff in "Explain the Cloud like I'm 10"

WHAT ABOUT "CLOUD COMPUTING" ?

For us the cloud is a set of *cloud providers* renting *cloud services*
which become increasingly "abstracted" from the hardware they run on...

SERVICES ?

- "Renting a server" ... (this is pure "cloud computing")
- "Replicated & Secure storage space" ...
- "Autoscaling deployment of a microservice" ...



(a portion of aws services)

HOW IS IT POSSIBLE ?

The magic of... virtualization !

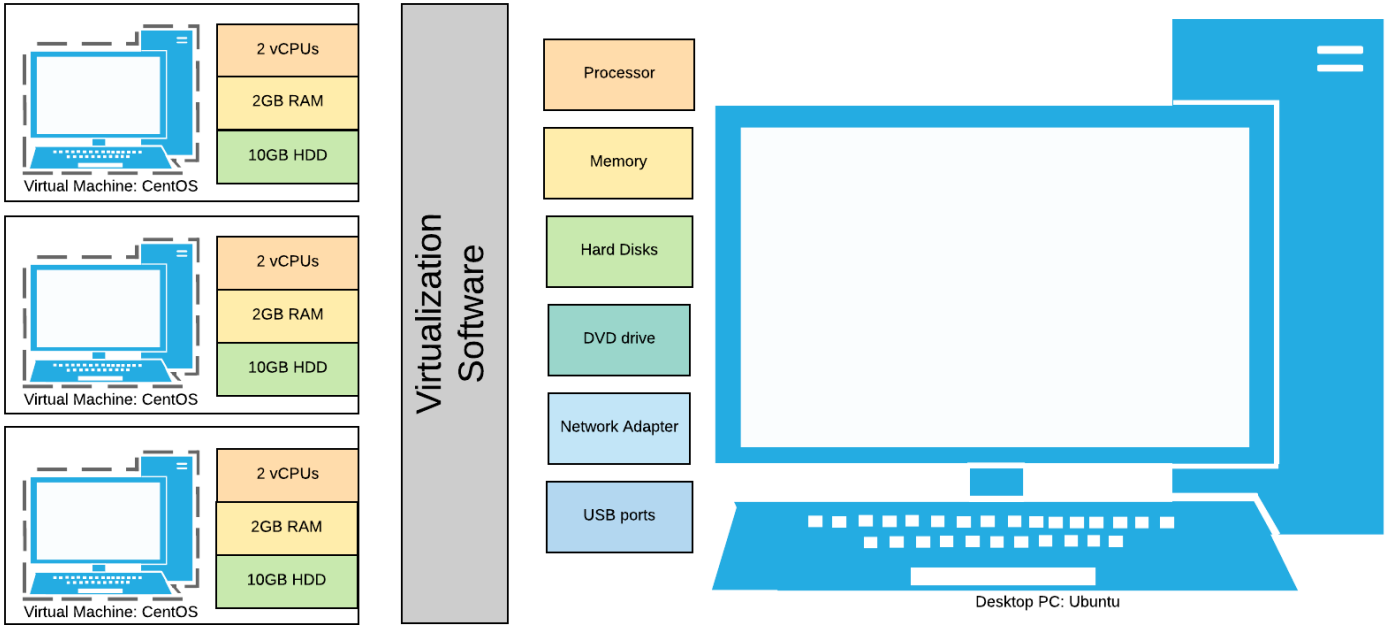
VIRTUALIZATION ?

> *In computing, virtualization refers to the act of creating a virtual (rather than actual) version of something, including virtual computer hardware platforms, storage devices, and computer network resources.*

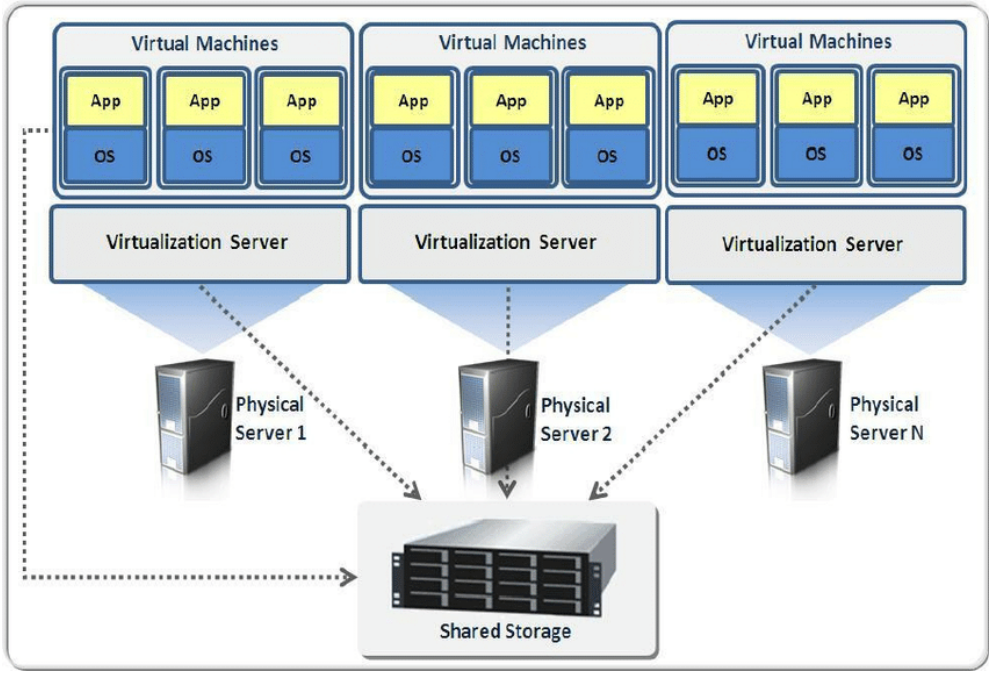
Wikipedia

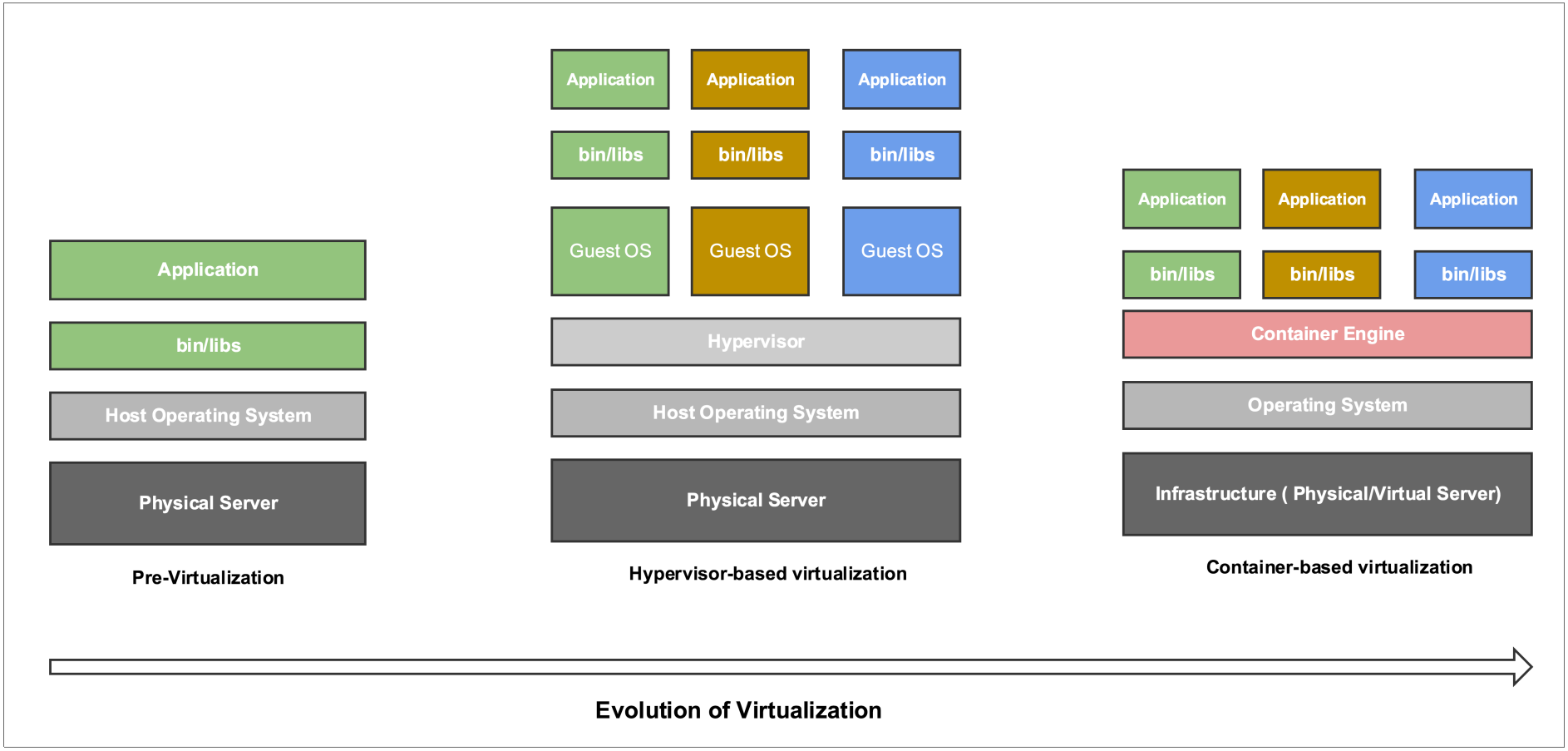
> *Basically we are running software on "abstract hardware" which is a "portion" of a real computer ("bare metal")*

Hardware Virtualization: a Desktop Virtualization Example



Hardware visualisation: Server Example

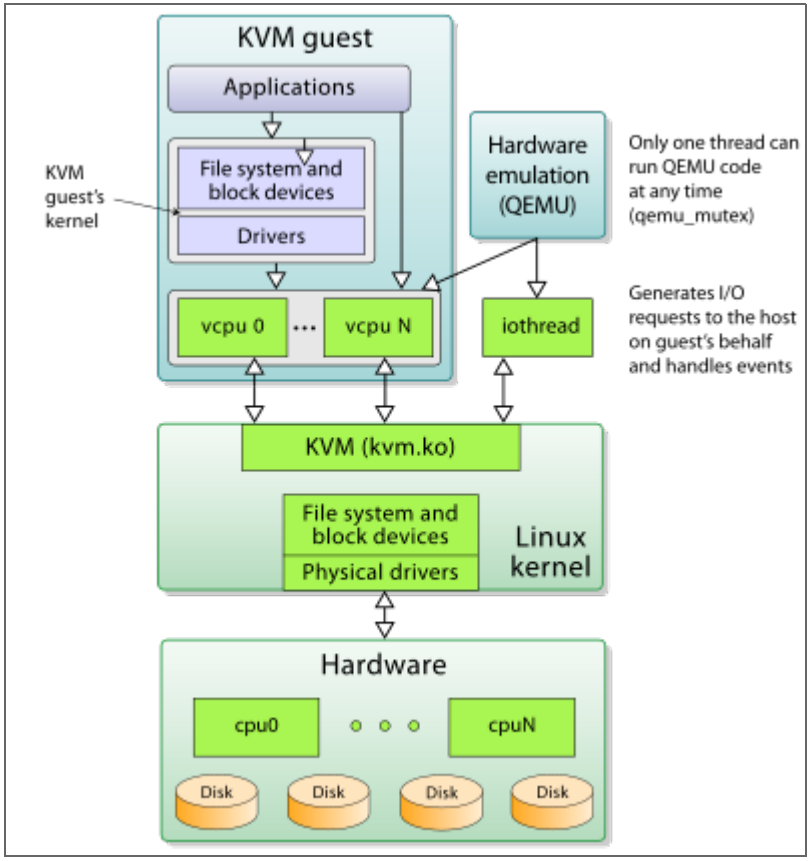



Application
bin/libs
Host Operating System
Physical Server
Pre-Virtualization
Application
Application
Application
bin/libs
bin/libs
bin/libs
Guest OS
Guest OS
Guest OS
Hypervisor
Host Operating System
Physical Server
Hypervisor-based virtualization
Application
Application
Application
bin/libs
bin/libs
bin/libs
Container Engine
Operating System
Infrastructure (Physical/Virtual Server)
Container-based virtualization
Evolution of Virtualization

DEFINITIONS

Hypervisor : A program for creating and running virtual machines.
Virtual Machine: The emulated equivalent of a computer system that runs on top of another system
Containers: Isolated environments that share the same underlying OS & resources

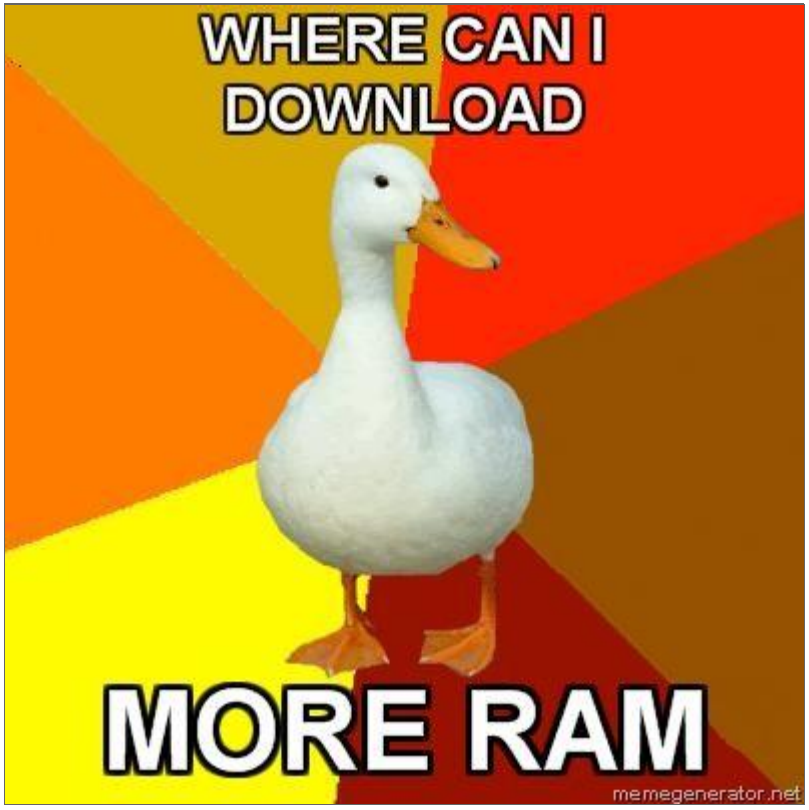
HYPERVISOR : KVM EXAMPLE (KERNEL VIRTUAL MACHINE)



NESTED HYPERVISORS : GOOGLE COMPUTE ENGINE

gce

CONSEQUENCE



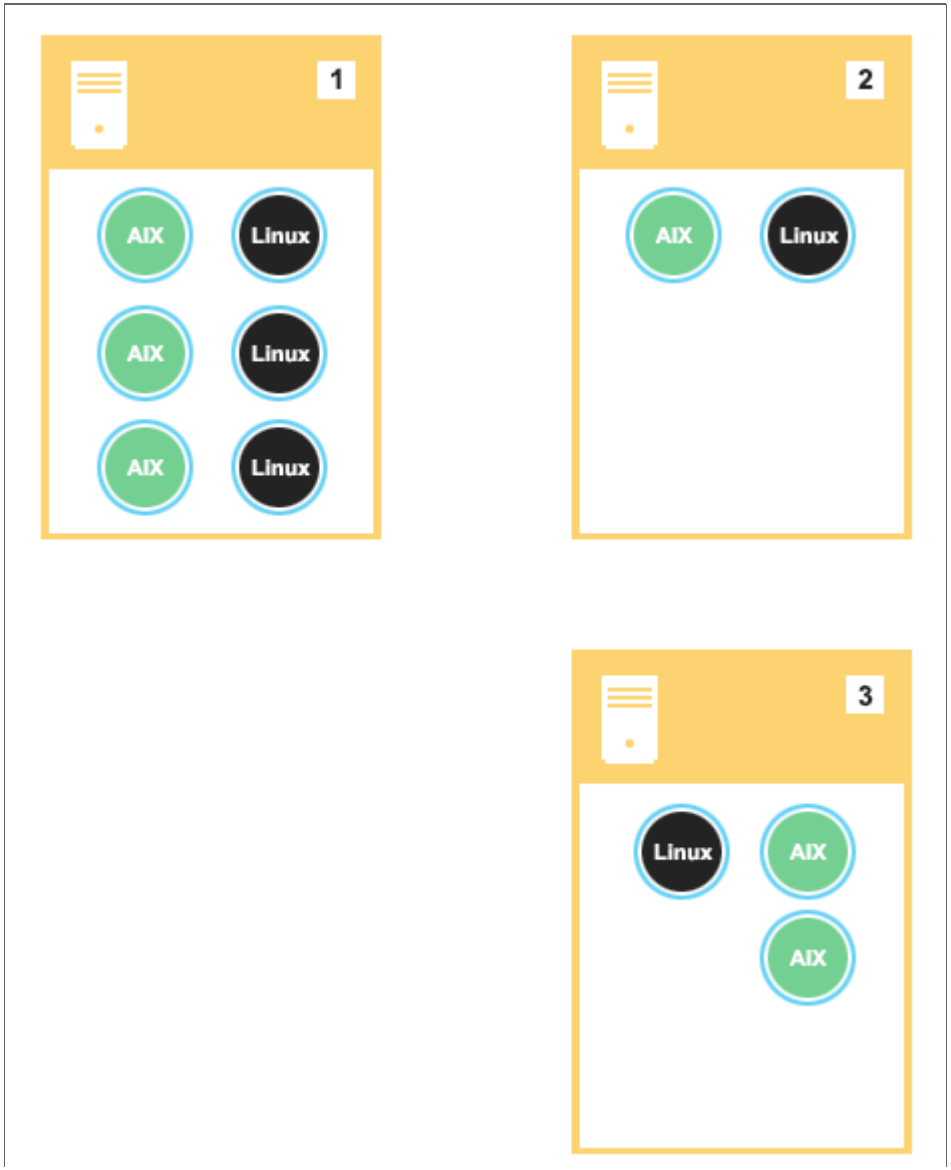
Any sufficiently advanced technology is indistinguishable from magic.

Clarke Third Law

HARDWARE ABSTRACTION

- Hardware Abstraction ("download more RAM")
- Fine-grained resource allocation / sharing
- Decouple maintenance of hardware from maintenance of software

RELIABILITY, SECURITY...



WHERE DOES IT COME FROM ?

Once upon a time...

Amazon (the e-commerce store) has "scaling" issues

So basically Amazon became very good at *running* scalable infrastructure as *services*

- For themselves...
- ... but also for other partners (target)

And that infrastructure is often there to answer peak load...

2002-2003; The idea

> *Building an infrastructure that is completely standardized, completely automated, and relied extensively on web services for things like storage*

http://blog.b3k.us/2009/01/25/ec2-origins.html

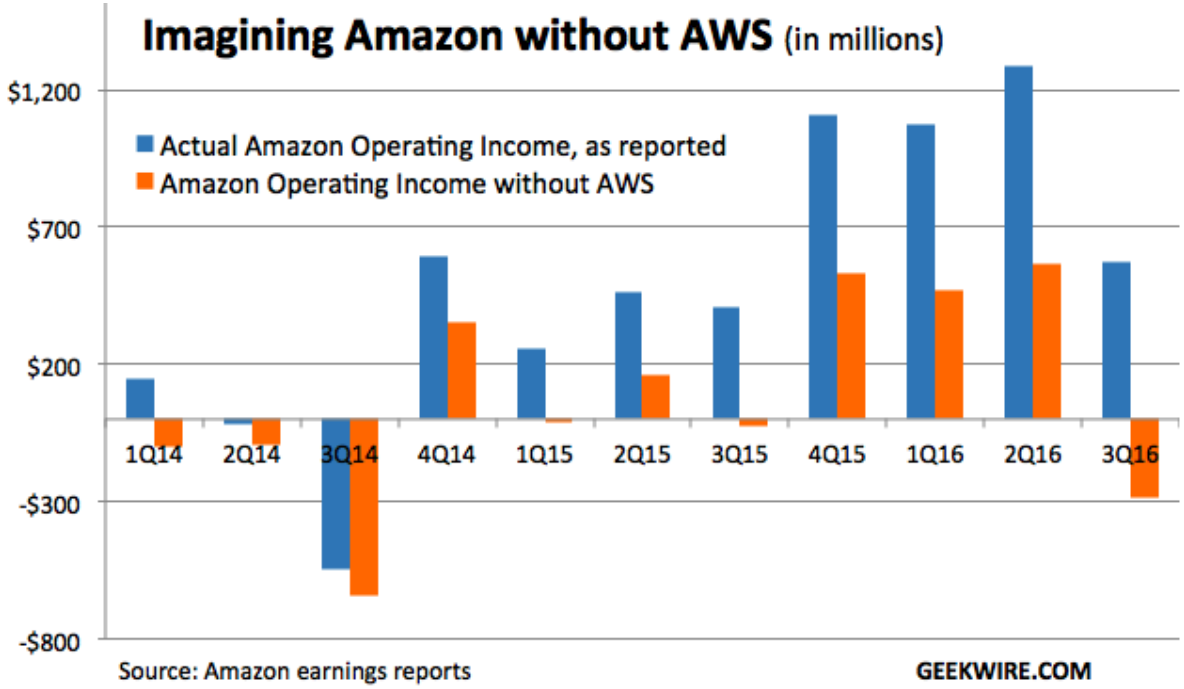
Let's sell it !

Announcing Amazon Elastic Compute Cloud (Amazon EC2) - beta

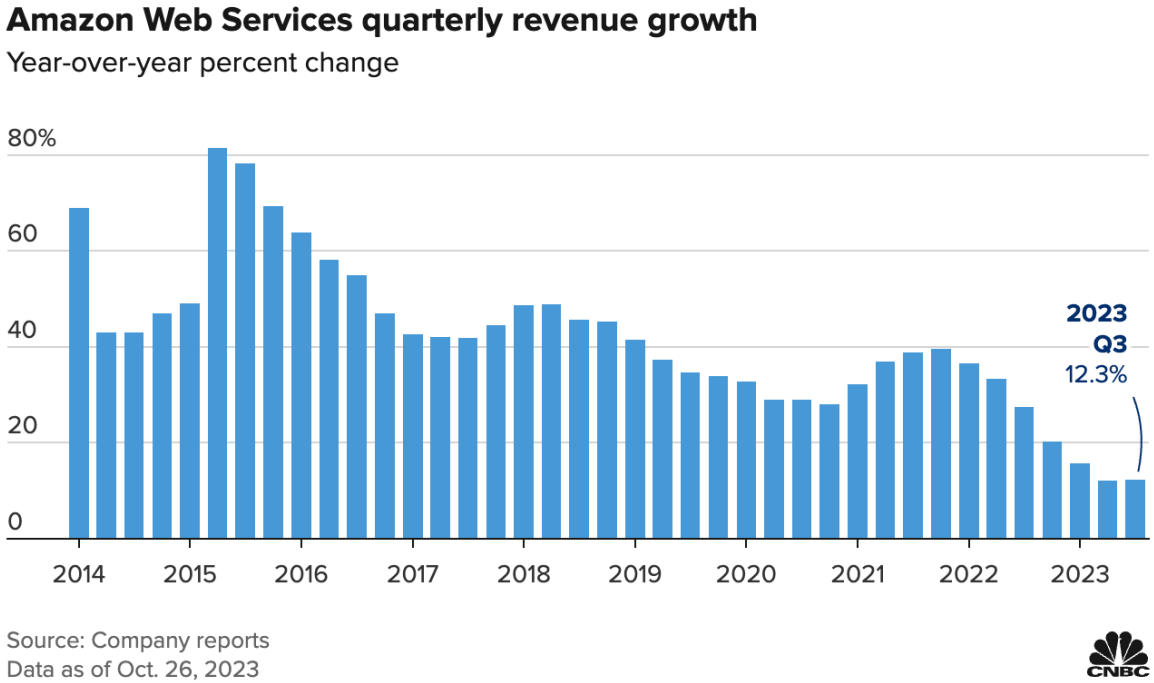
Posted On: Aug 24, 2006

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers. Just as Amazon Simple Storage Service (Amazon S3) enables storage in the cloud, Amazon EC2 enables "compute" in the cloud. Amazon EC2's simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use.

HOW DOES AMAZON CAN OFFER FREE SHIPPING TO EVERYBODY

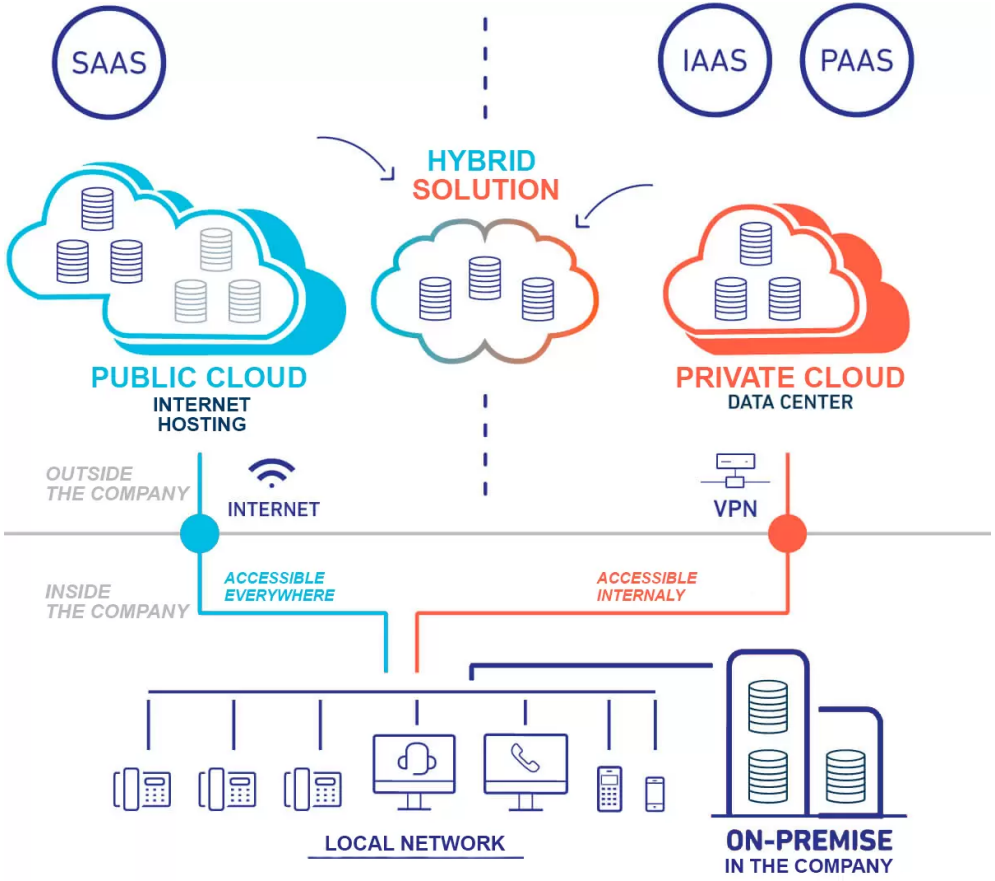


HOW DOES AMAZON CAN OFFER FREE SHIPPING TO EVERYBODY

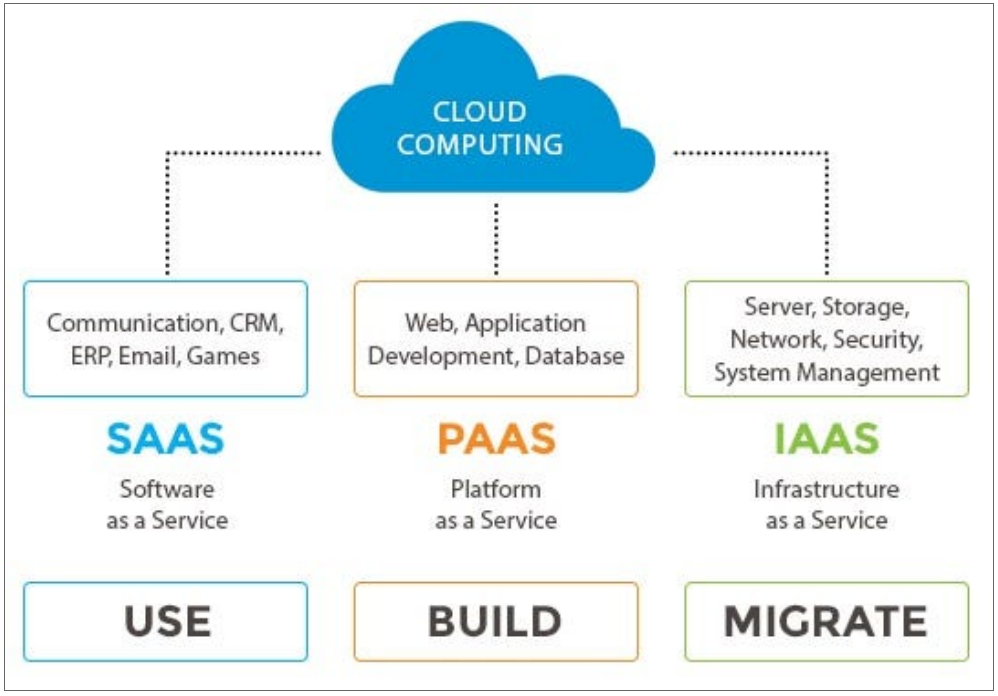


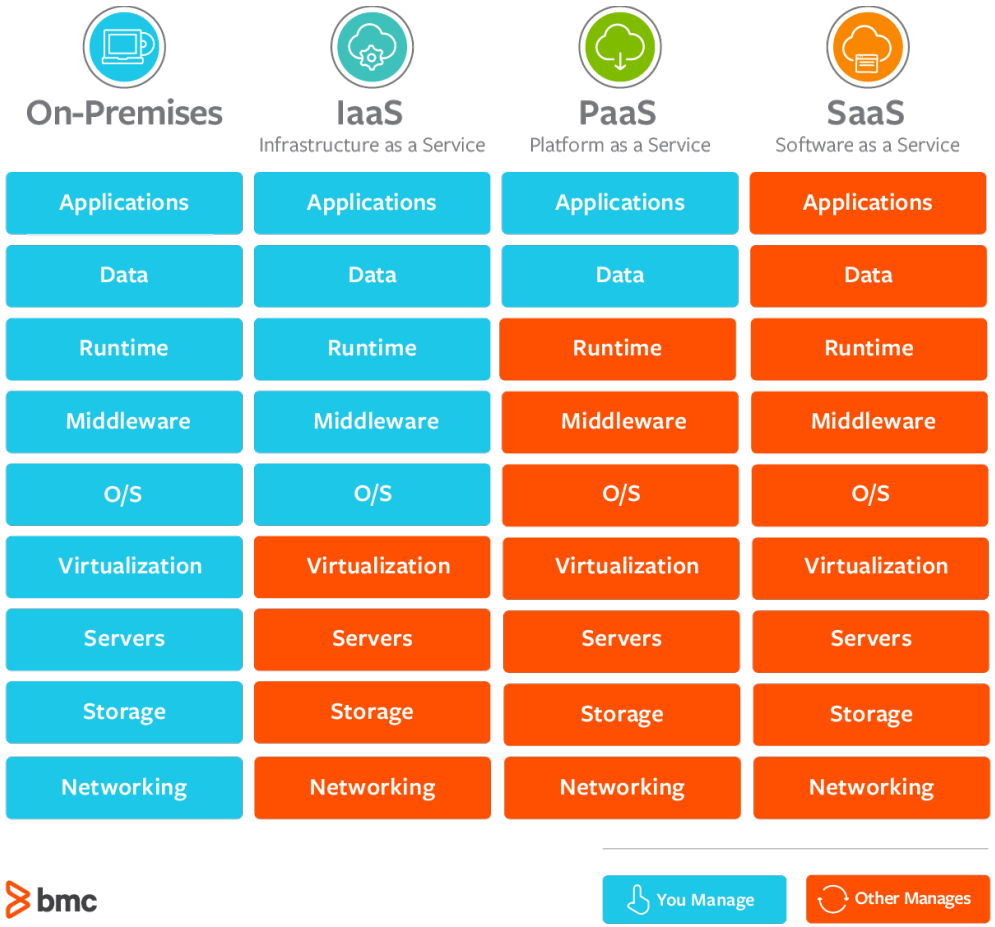
THE MANY LAYERS OF CLOUD COMPUTING

Hybrid Cloud ? Private Cloud ? Public Cloud ?



Cloud providers are offering services with increasing layers of abstraction...




On-Premises
IaaS
Infrastructure as a Service
PaaS
Platform as a Service
SaaS
Software as a Service
Applications
Data
Runtime
Middleware
O/S
Virtualization
Servers
Storage
Networking
bmc
You Manage
Other Manages

EXAMPLES

- Renting a server with hard drive and storing data
- Using data storage service like google cloud storage without managing the infrastructure
- Using google drive

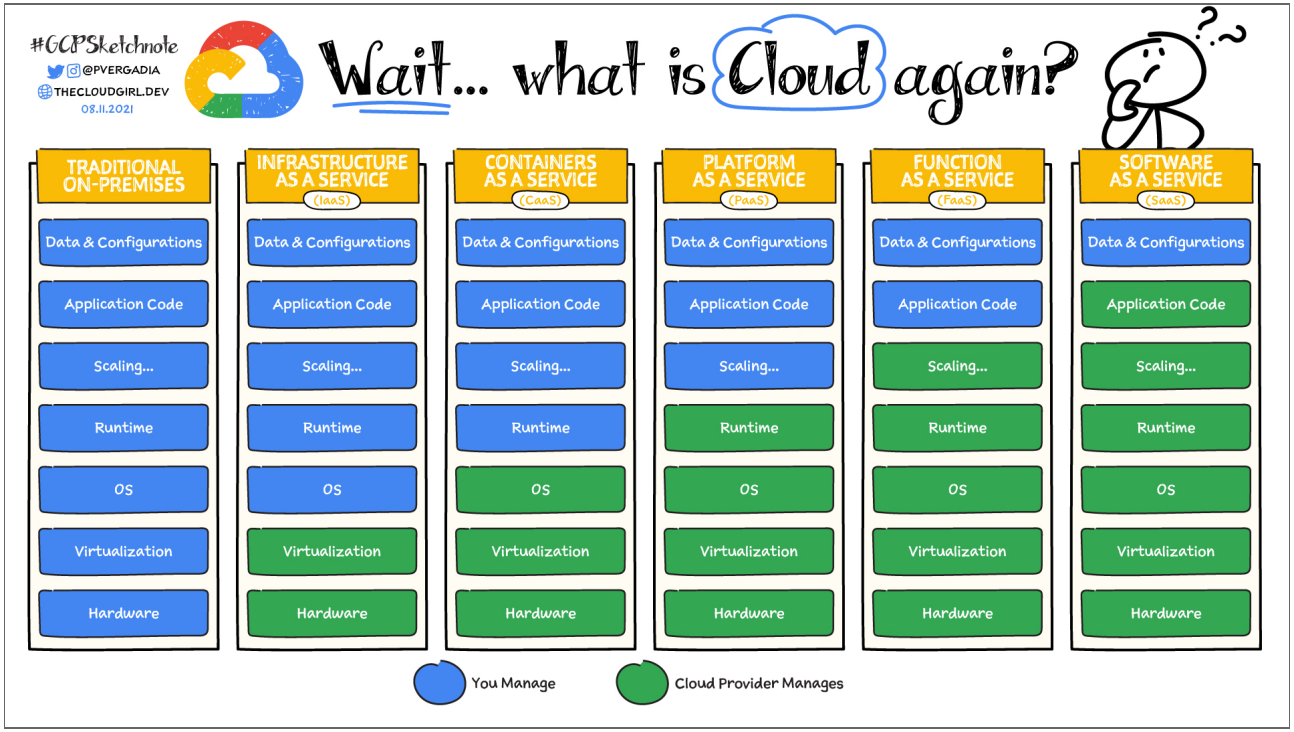
EXAMPLES

- Renting a server with hard drive and storing data **IaaS**
- Using data storage service like google cloud storage without managing the infrastructure **PaaS**
- Using Dropbox **SaaS**

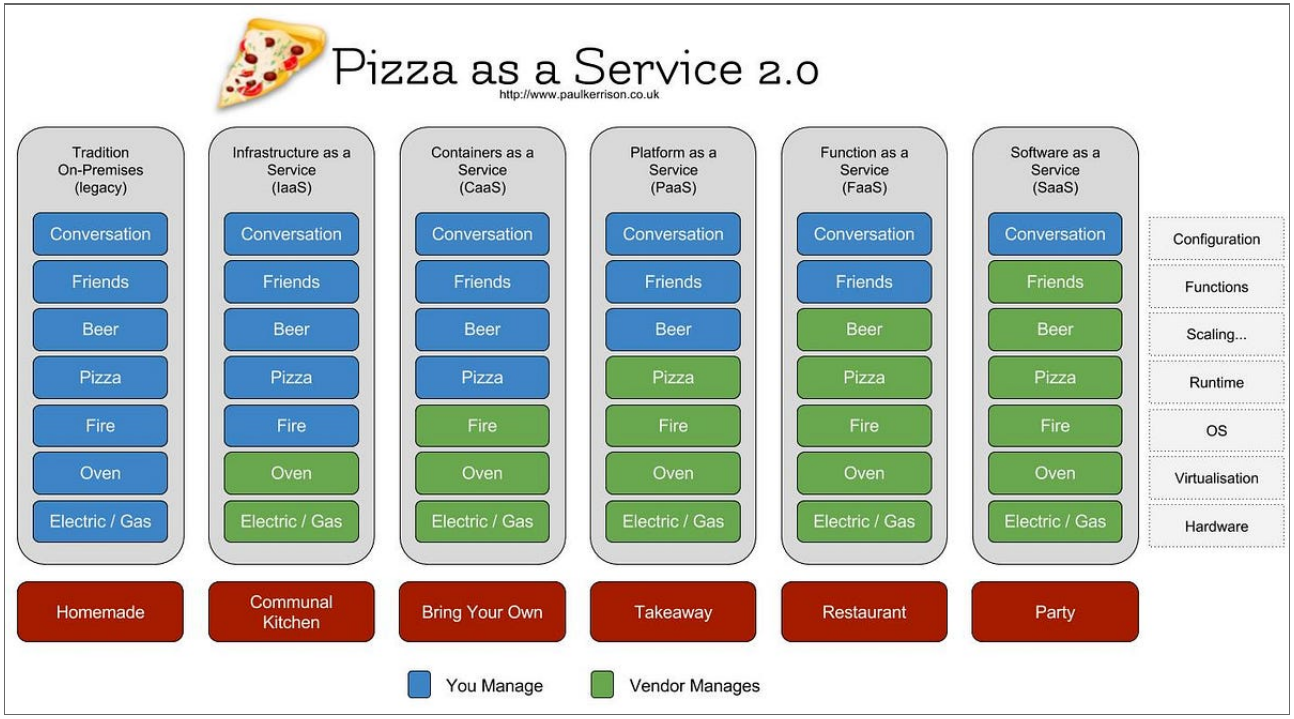
EXAMPLES

- Renting a GPU farm to deploy your Large Language Model and serve it **IaaS**
- Using the HuggingFace API to serve predictions from your model **PaaS**
- Using ChatGPT **SaaS**

IT GETS HARDER

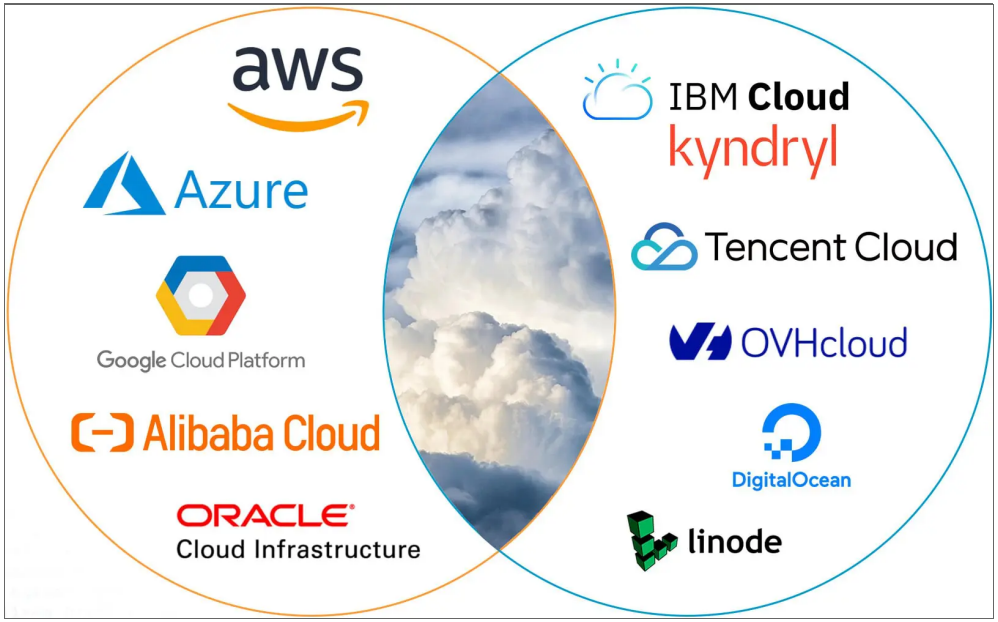


USEFUL ANALOGY

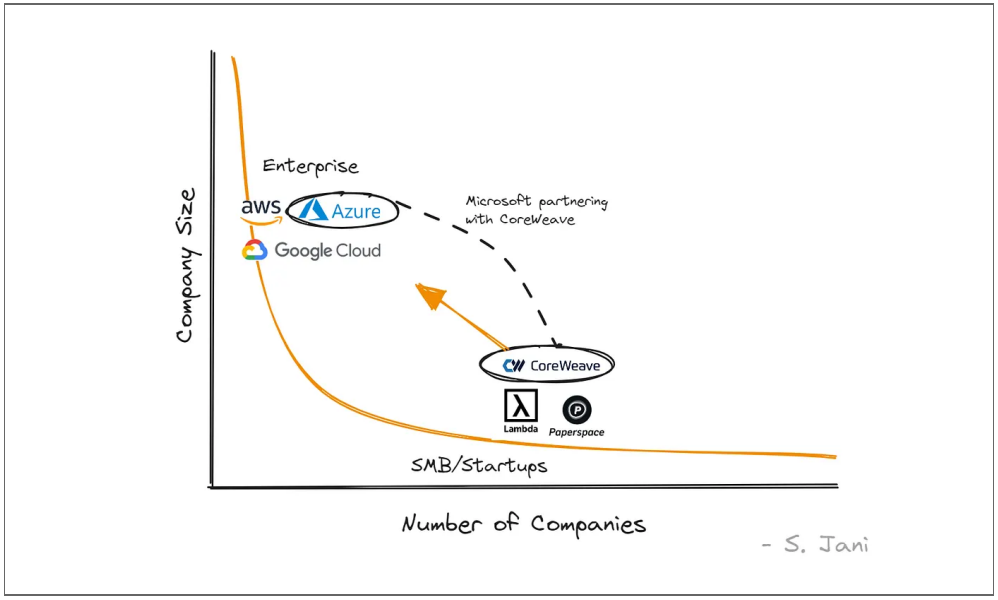


PUBLIC CLOUD PROVIDERS

MAJOR CLOUD PROVIDERS



AI CLOUD PROVIDERS



- https://www.paperspace.com/core
- https://lambdalabs.com/
- https://huggingface.co/hardware

FRENCH CLOUD PROVIDERS 🐓🧀🐸🇫🇷

OVH went public in 2021
Scaleway is leading the charge for AI in France (& Europe)
Outscale is focusing on SecNumCloud
BleuCloud is CapGemini x Orange

FRENCH CLOUD PROVIDERS 🐓🧀🐸🇫🇷

- Cloud Act !
- SecNumCloud : ANSSI's security qualification for cloud providers handling sensitive French data

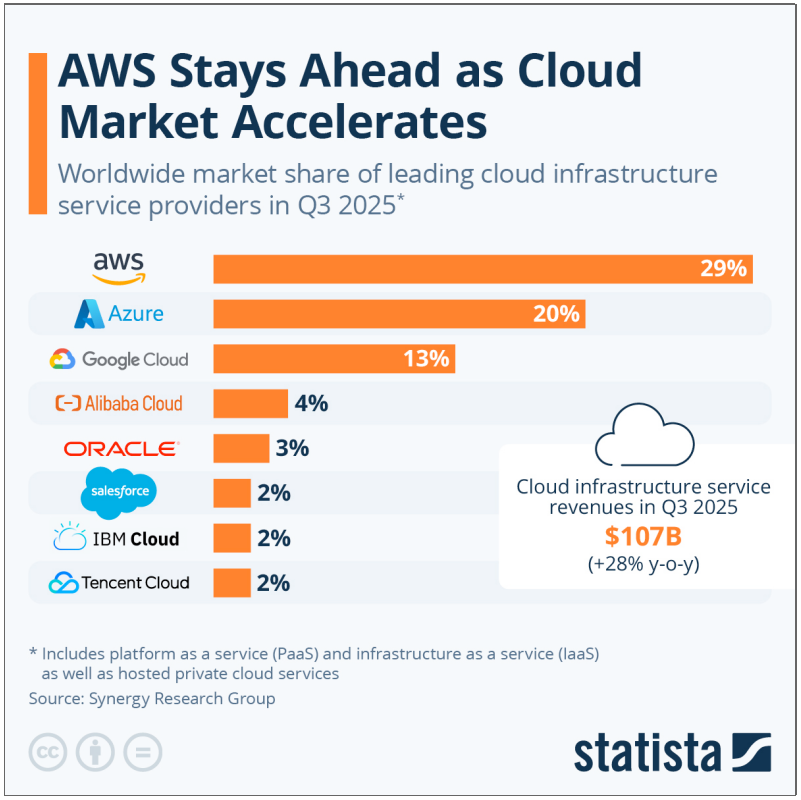
Project	Partners	Tech	Status
Bleu	Orange + Capgemini (100% French ownership), Microsoft as tech partner (not shareholder)	Azure + M365	Commercial operations launched 2024. First clients include EDF, Dassault Aviation.
S3NS.	Thales (majority) Google Cloud (minority shareholder, <24%)	GCP	SecNumCloud 3.2 obtained December 2024. Offers IaaS/PaaS/CaaS

🇪🇺 GAIA-X : Cloud Federation in Europe

https://www.data-infrastructure.eu/GAIAX/

https://www.contexte.com/article/tech/gaia-x-souverainete-cloud_150712.html

CLOUD MARKET SHARE (WORLD)



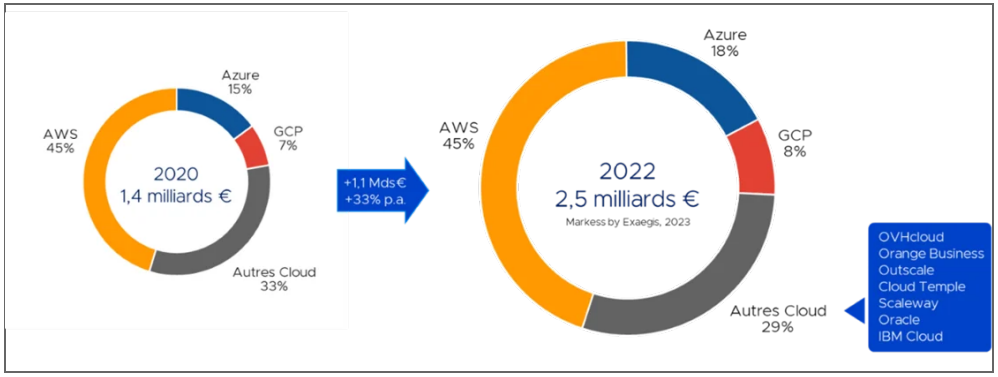
source, 2025

CLOUD MARKET SHARE (EUROPE)

europan_cloud

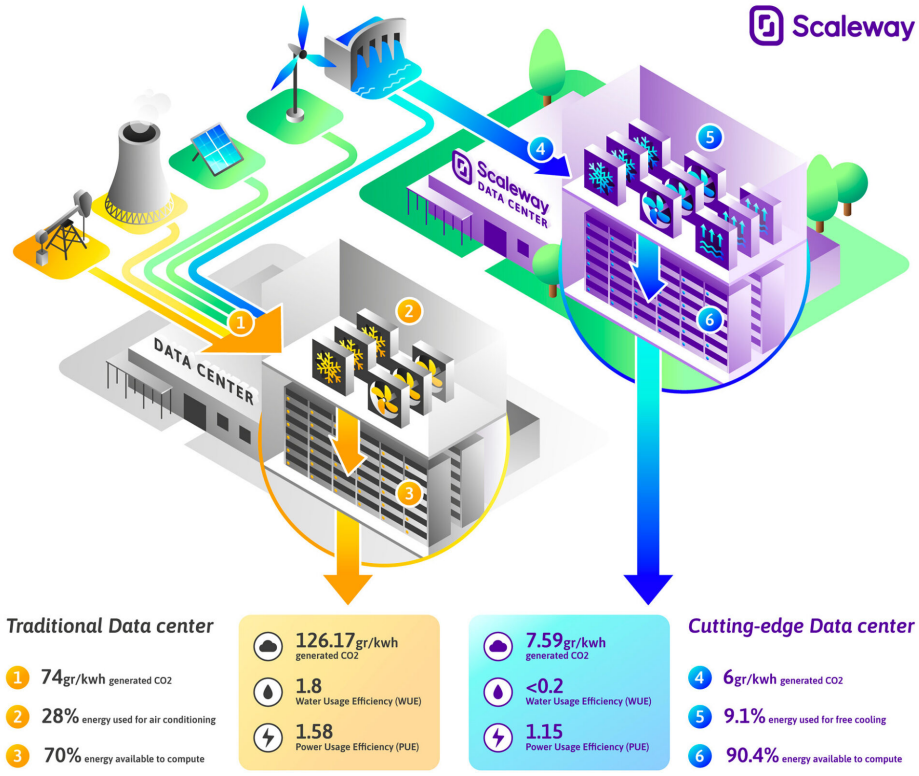
source, 2025

CLOUD MARKET SHARE (FRANCE)



source, 2021

CLOUD COMPUTING & ENVIRONMENT



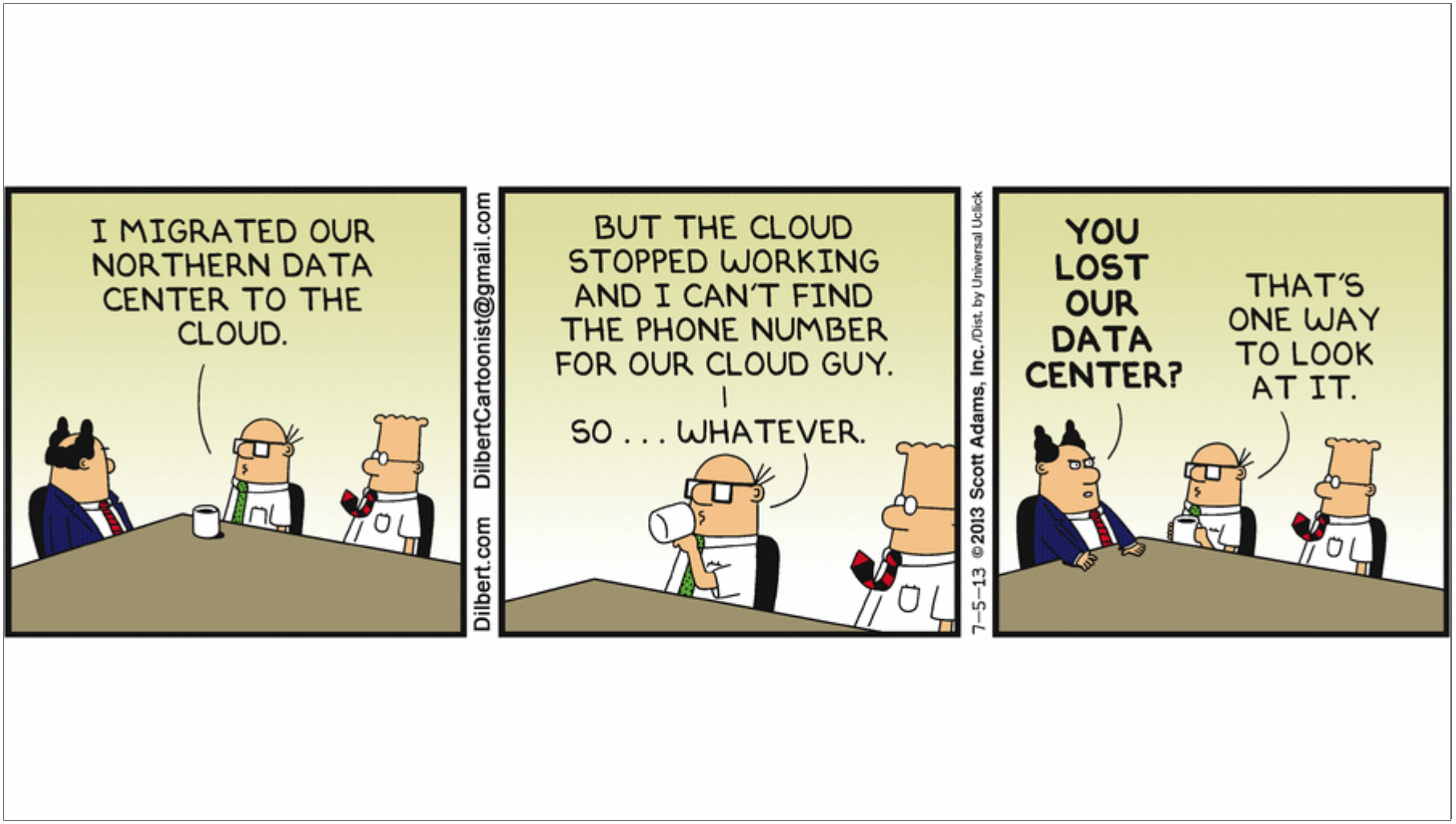
I am not competent to say anything about this. Some sources

- The Shift Project : **https://theshiftproject.org/article/deployer-la-sobriete-numerique-rapport-shift/**
- Scaleway : **https://www.scaleway.com/fr/leadership-environnemental/**
- Google : **https://cloud.google.com/sustainability**
- Earth.org : **https://earth.org/environmental-impact-of-cloud-computing/**

On "Artificial Intelligence" & sustainability, entry points

- **Power Hungry Processing: Watts Driving the Cost of AI Deployment?**
- **The Environmental Impacts of AI -- Primer**

"USING" THE CLOUD



CLOUD COMPUTING: A TECHNICAL *EVOLUTION*

- More Virtualization
- More API
- More Managed Services

CLOUD COMPUTING: A USAGE REVOLUTION

AUTONOMY : ACCESS TO COMPUTING POWER

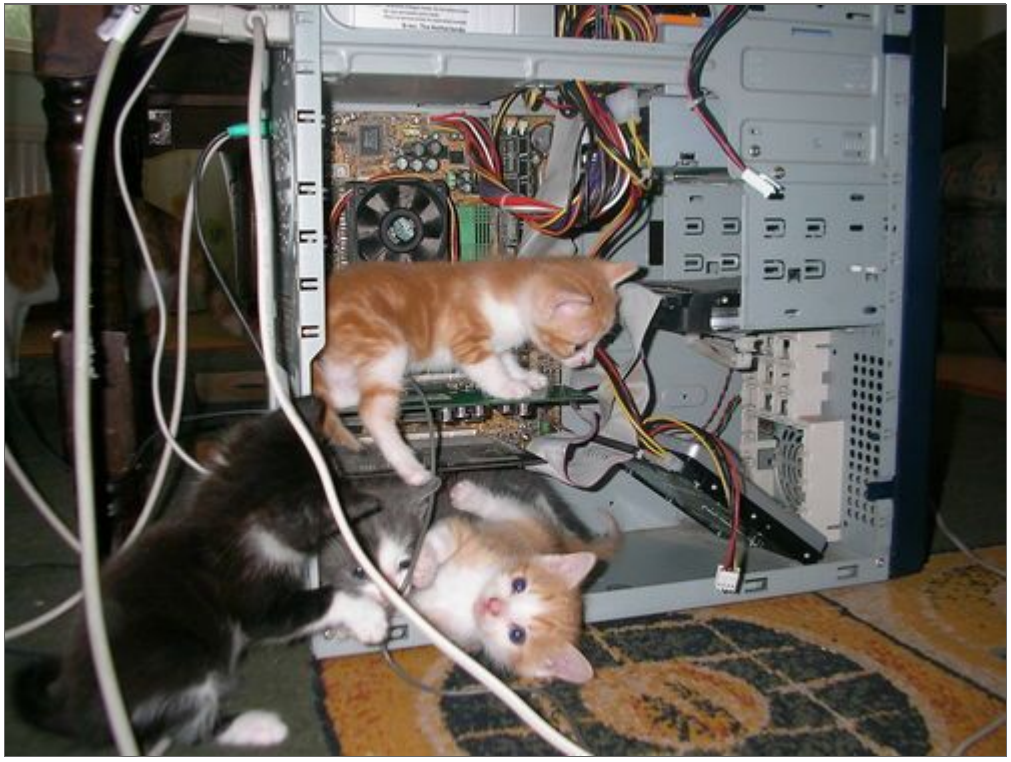
- Outsourcing infra, maintenance, security, development of new services
- Pay-per-use with "Infinitely scalable" infrastructure
- "No need to plan out" infrastructure
- Enabling innovation
- Power in the hands of developpers/builders

CHANGING THE WAY WE INTERACT WITH HARDWARE

We interact with cloud providers using APIs...

```
gcloud compute --project=deeplearningsps
instances create ${INSTANCE_NAME} \
    --zone=${ZONE} \
    --machine-type=n1-standard-8 \
    --scopes=default,storage-rw,compute-rw \
    --maintenance-policy=TERMINATE \
    --image-family=ubuntu-1804-lts \
    --image-project=ubuntu-os-cloud \
    --boot-disk-size=200GB \
    --boot-disk-type=pd-standard \
    --accelerator=type=nvidia-tesla-
p100,count=1 \
    --metadata-from-file startup-
script=startup_script.sh
```

BEFORE...



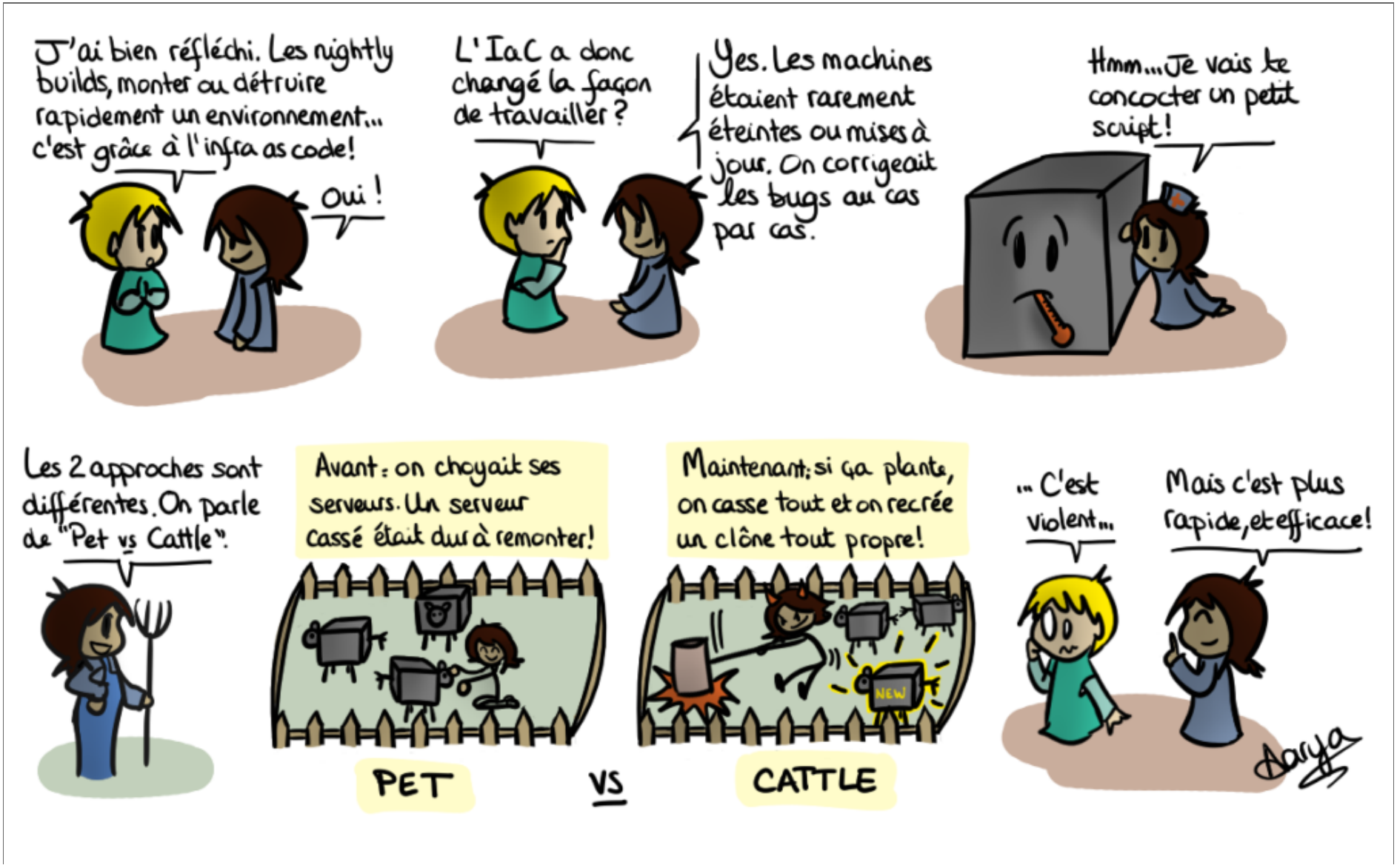
AFTER...

```
resources:
- name: vm-created-by-deployment-manager
  type: compute.v1.instance
  properties:
    zone: us-central1-a
    machineType: zones/us-central1-a/
machineTypes/n1-standard-1
    disks:
    - deviceName: boot
      type: PERSISTENT
      boot: true
      autoDelete: true
      initializeParams:
        sourceImage: projects/debian-cloud/
global/images/family/debian-9
    networkInterfaces:
    - network: global/networks/default
```

INFRASTRUCTURE AS CODE

- Infra is now managed via text files
- Data is securely stored on storage
- So we store code + urls on git... and everything is reproducible !
- We use automated deployment tools (terraform, gcp deployment manager...)

PET VS CATTLE



LET'S DISCUSS

Is using cloud computing less expensive ?

- 👍 Depend on your {normal / peak} utilization
- 👍 Access to latest hardware without investment
- 👎 Fully utilized hardware is more expensive on the cloud
- 👎 CLOUD HYGIENE !

- Watch for unused services / storage
- Shutdown machines when not used
- Services stack up...

Is using cloud computing more secure / safer ?

- 👍 The best engineers in the world working on it
- 👍 Secure regions / private cloud...
- 👎 Your data somewhere in some datacenter...
- 👎 "Dependency" towards your cloud provider...

CLOUD NATIVE CULTURE EXAMPLE : OBJECT STORAGE

BLOCK STORAGE VS OBJECT STORAGE

Aspect	Block Storage	Object Storage
Model	Disk partitions, file systems	Buckets and objects
Access	Mount as drive, POSIX	HTTP API (REST)
Use case	Databases, OS disks	Data lakes, ML datasets
Scalability	Harder to scale	Built with scale in mind (abstraction over the storage location)

OBJECT STORAGE MODEL

- **Bucket**: Container for objects (like a top-level folder)
- **Object**: File + metadata (stored as key-value)
- **Key**: Object path (e.g., `data/train/image_001.jpg`)

No real hierarchy, just a flat namespace with "/" in keys

PERMISSIONS: POSIX VS OBJECT STORAGE

POSIX (files)	Object Storage
User/Group/Other	IAM policies (users, groups, service accounts)
rwx bits	Fine-grained: read, write, delete, list, admin
Per-file only	Per-bucket or per-object
Local to machine	Managed centrally (cloud IAM)

Object storage enables **sharing across teams/projects** without managing OS users

WHY OBJECT STORAGE FOR DATA SCIENCE?

- **Cheap**: ~$0.02/GB/month (vs$0.10+ for disk)
- **Accessible**: HTTP API from anywhere
- **Scalable**: Petabytes without infrastructure management
- **Shareable**: Easy to share datasets across teams/machines

OBJECT STORAGE PROVIDERS

Provider	Service	URI Format
AWS	S3	`s3://bucket/key`
GCP	GCS	`gs://bucket/key`

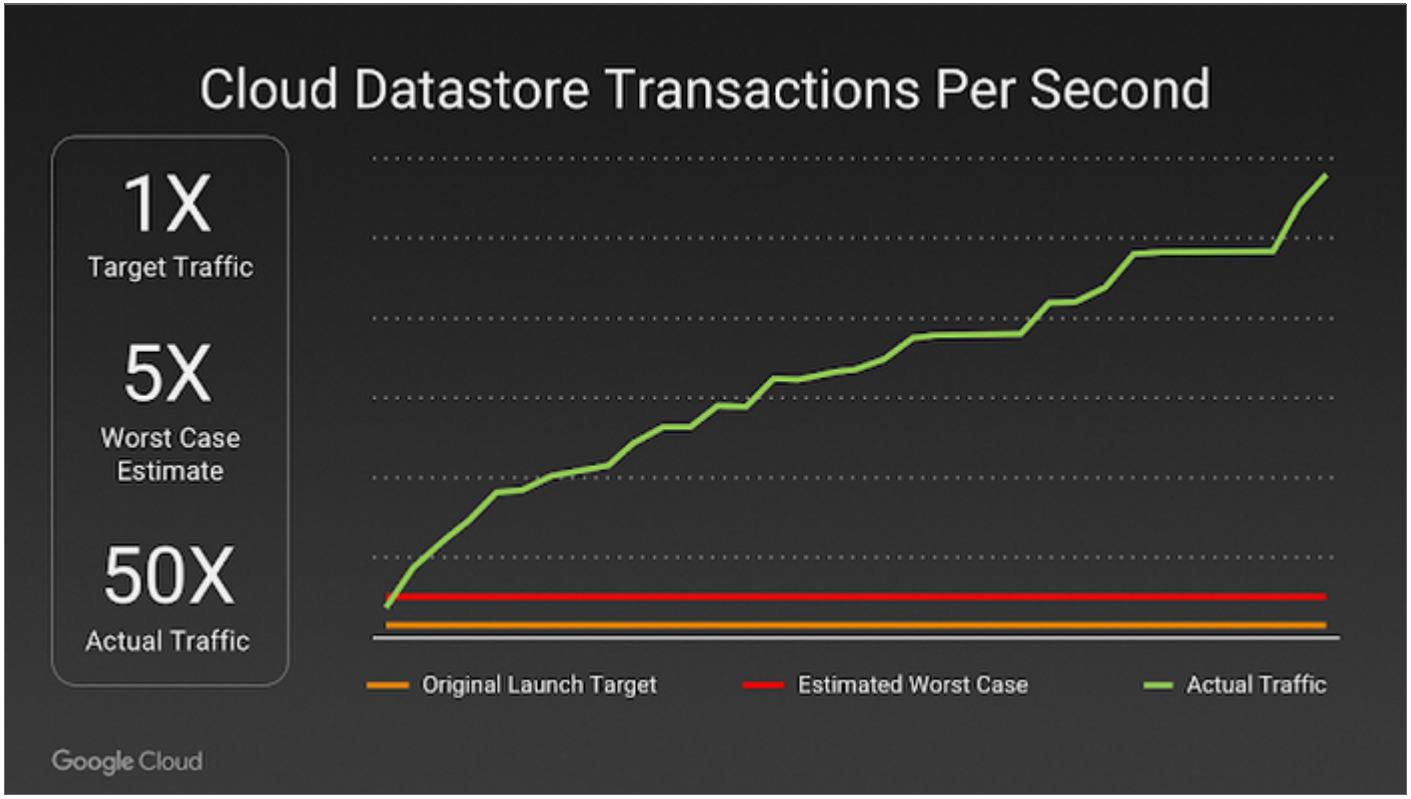
Same concepts, similar APIs, different CLIs

CLOUD USAGE, SOME ANECDOTES

BIG TECH PUBLIC CLOUD BILLS

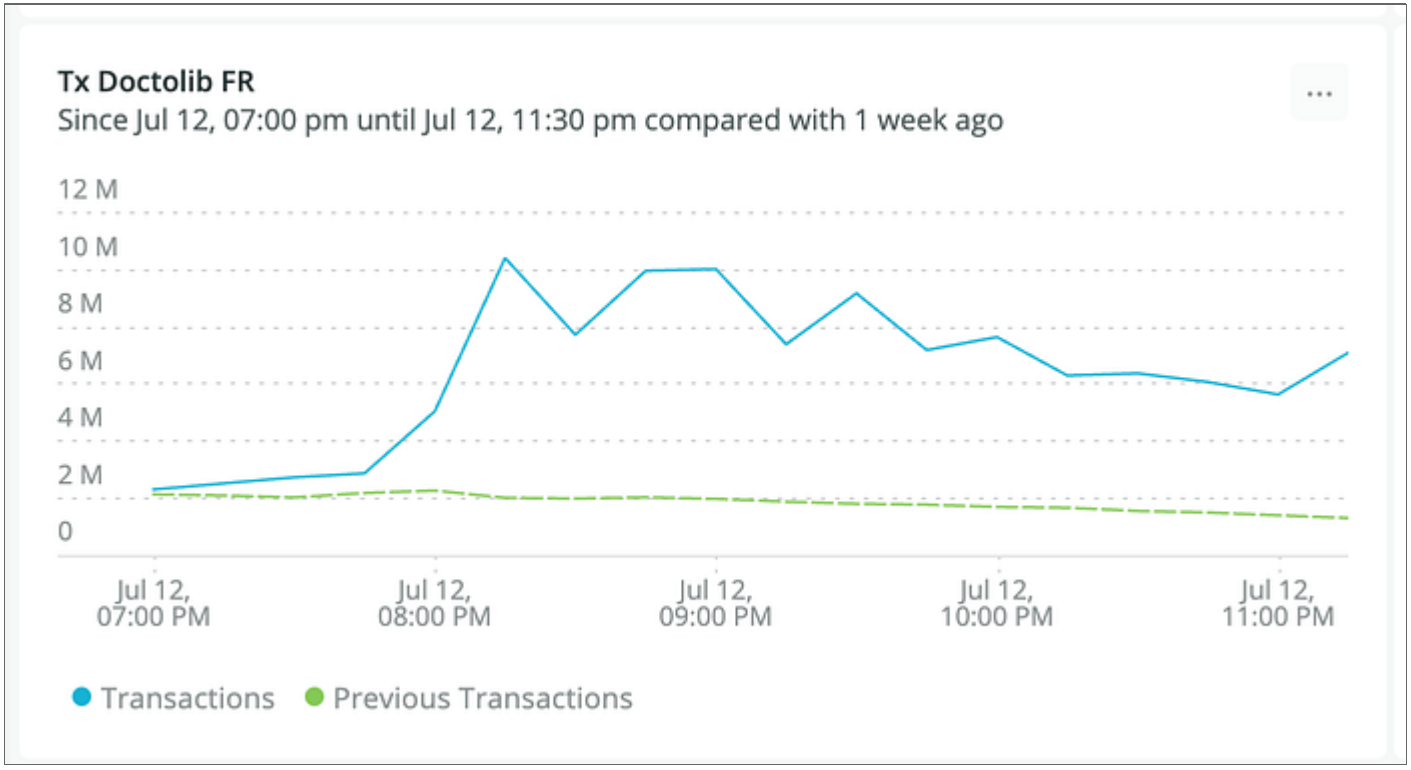
- Apple in 2019 **350m$ on AWS / year**
- Spotify in 2018 **150m$ on GCP / year**
- Lyft in 2019 **100m$ on AWS / year**

POKEMON GO LAUNCH (2016)



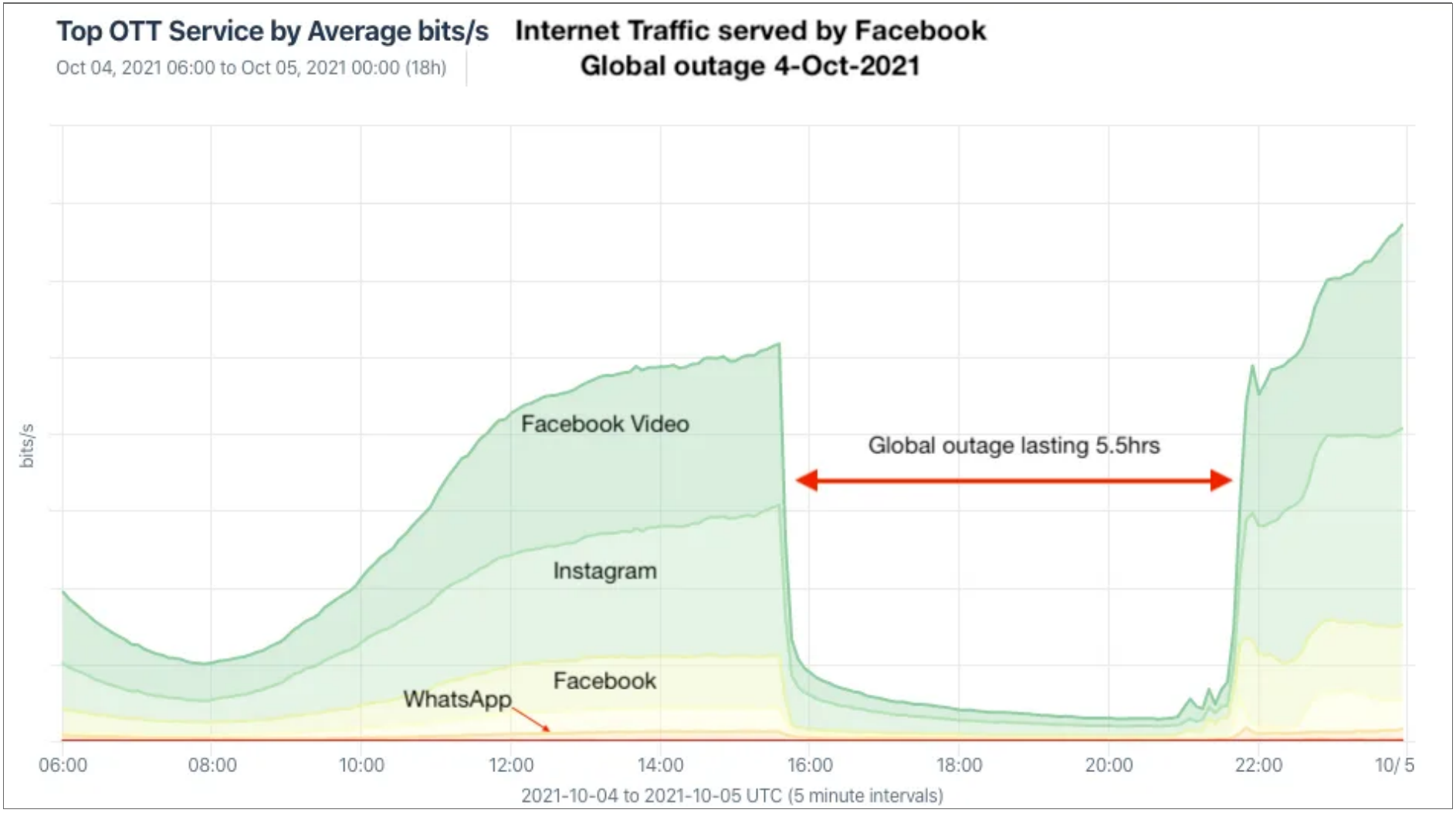
source

DOCTOLIB (2021)



source

FACEBOOK OCTOBER 2021 FAILURE



https://blog.cloudflare.com/october-2021-facebook-outage/

AWS US-EAST-1 FAILURE (2022)

13 June 2023: AWS. The largest AWS region (us-east-1) degraded heavily for 3 hours, impacting 104 AWS services. A joke says that when us-east-1 sneezes the whole world feels it, and this was true: Fortnite matchmaking stopped working, McDonalds and Burger King food orders via apps couldn't be made, and customers of services like Slack, Vercel, Zapier and many more all felt the impact. (incident details). We did a deepdive into this incident earlier in AWS's us-east-1 outage.

https://aws.amazon.com/message/12721/

LINKS

http://highscalability.com
http://highscalability.com/all-time-favorites
Netflix: What happens when you press play - 2017
Mind boggling statistics on Amazon Prime Day

CLOUD COMPUTING & AI

ALL ABOUT THAT SCALE

This was in 2022,

BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

> *Training BLOOM took about 3.5 months to complete and consumed 1,082,990 compute hours. Training was conducted on 48 nodes, each having 8 NVIDIA A100 80GB GPUs (a total of 384 GPUs);*

ALL ABOUT THAT SCALE

This is 2024,

Network topology. Our RoCE-based AI cluster comprises 24K GPUs[5] connected by a three-layer Clos network (Lee et al., 2024). At the bottom layer, each rack hosts 16 GPUs split between two servers and connected by a single Minipack2 top-of-the-rack (ToR) switch. In the middle layer, 192 such racks are connected by Cluster Switches to form a pod of 3,072 GPUs with full bisection bandwidth, ensuring no oversubscription. At the top layer, eight such pods within the same datacenter building are connected via Aggregation Switches to form a cluster of 24K GPUs. However, network connectivity at the aggregation layer does not maintain full bisection bandwidth and instead has an oversubscription ratio of 1:7. Our model parallelism methods (see Section 3.3.2) and training job scheduler (Choudhury et al., 2024) are all optimized to be aware of network topology, aiming to minimize network communication across pods.

https://dblalock.substack.com/p/2024-8-4-arxiv-roundup-llama-31-training

AI DISTRIBUTED COMPUTING

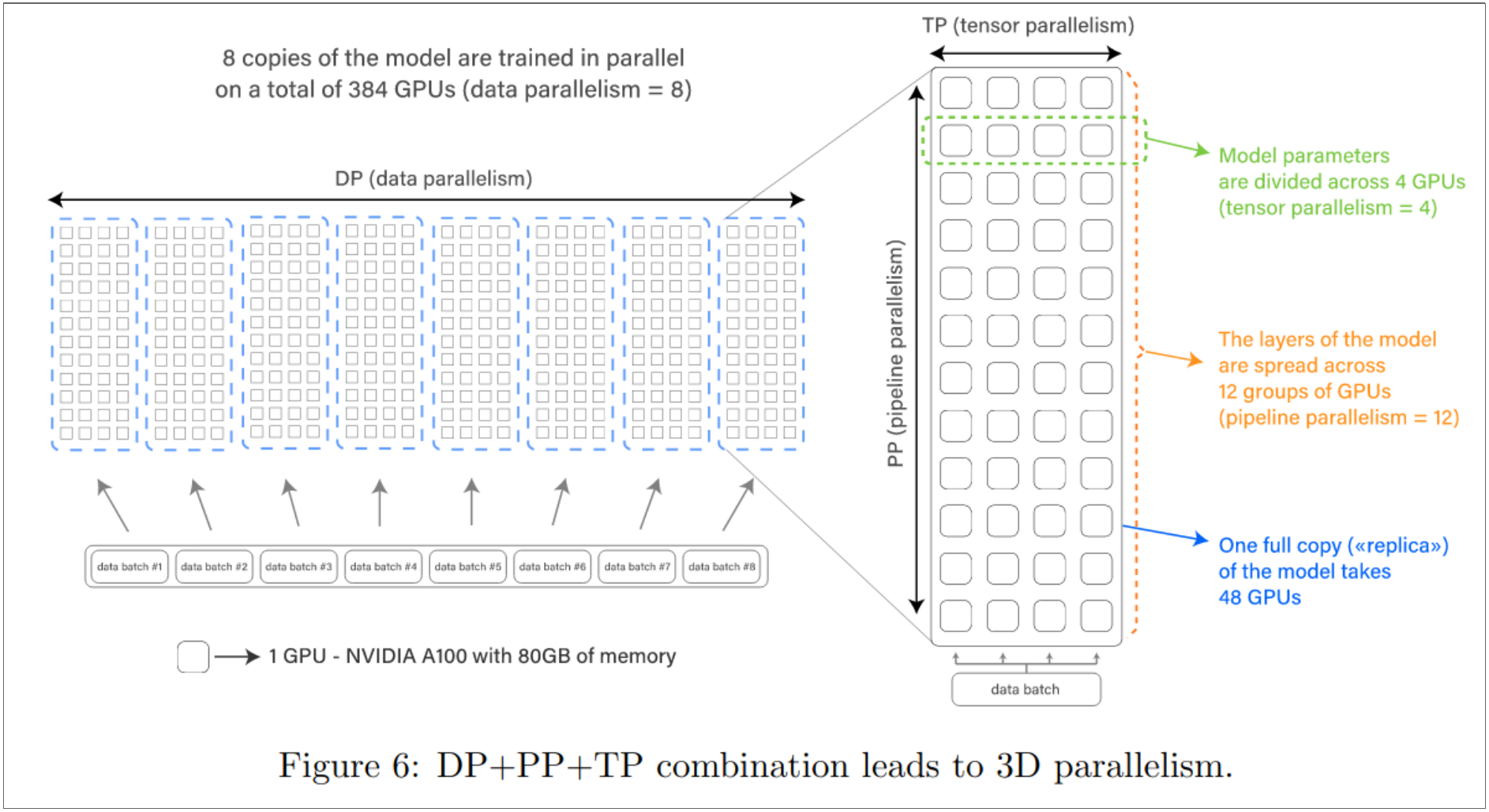


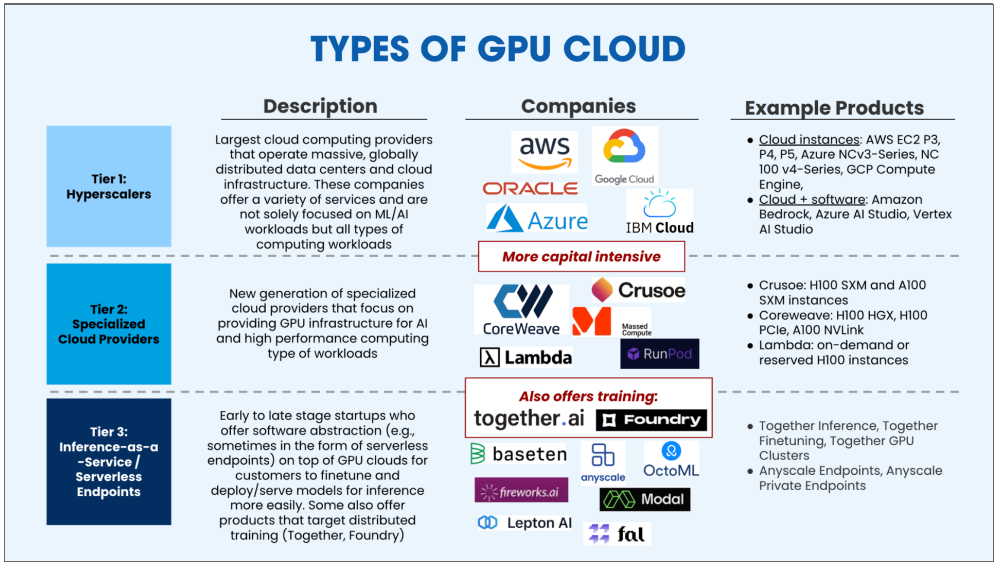
Figure 6: DP+PP+TP combination leads to 3D parallelism.

STABLE DIFFUSION

Number of A100s	Throughput (images / second)	Days to Train on MosaicML Cloud	A100-hours	Approx. Cost on MosaicML Cloud
8	128.2	258.83	49,696	$99,000
16	254.0	130.63	50,166	$100,000
32	485.7	68.33	52,470	$105,000
64	912.2	36.38	55,875	$110,000
128	1618.4	20.5	62,987	$125,000
256*	2,589.4	12.83	78,735	$160,000

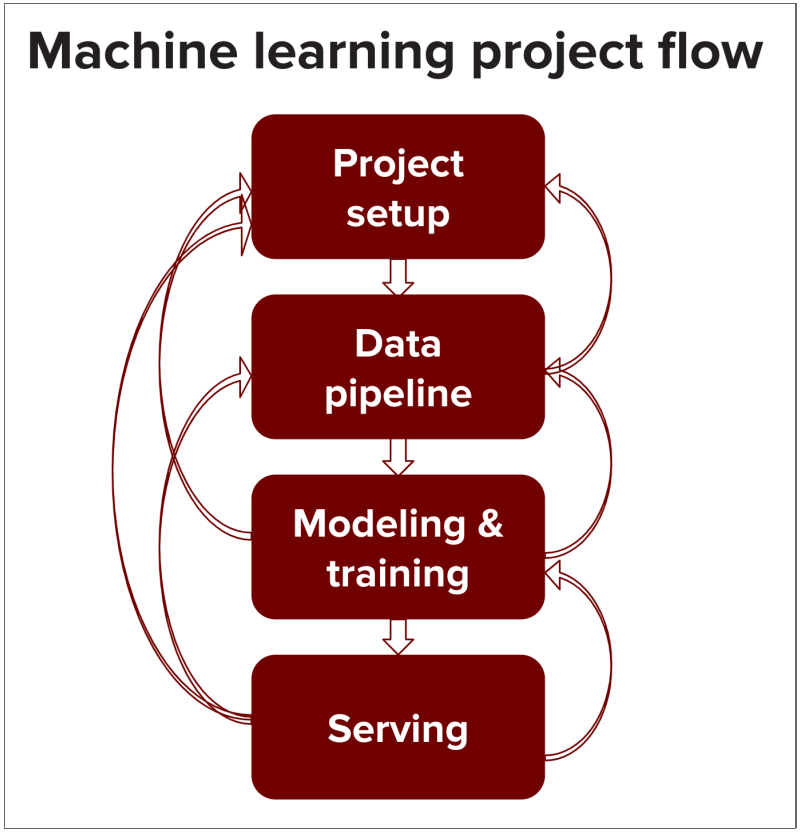
Stable Diffusion Training Times

AI CLOUD PROVIDERS

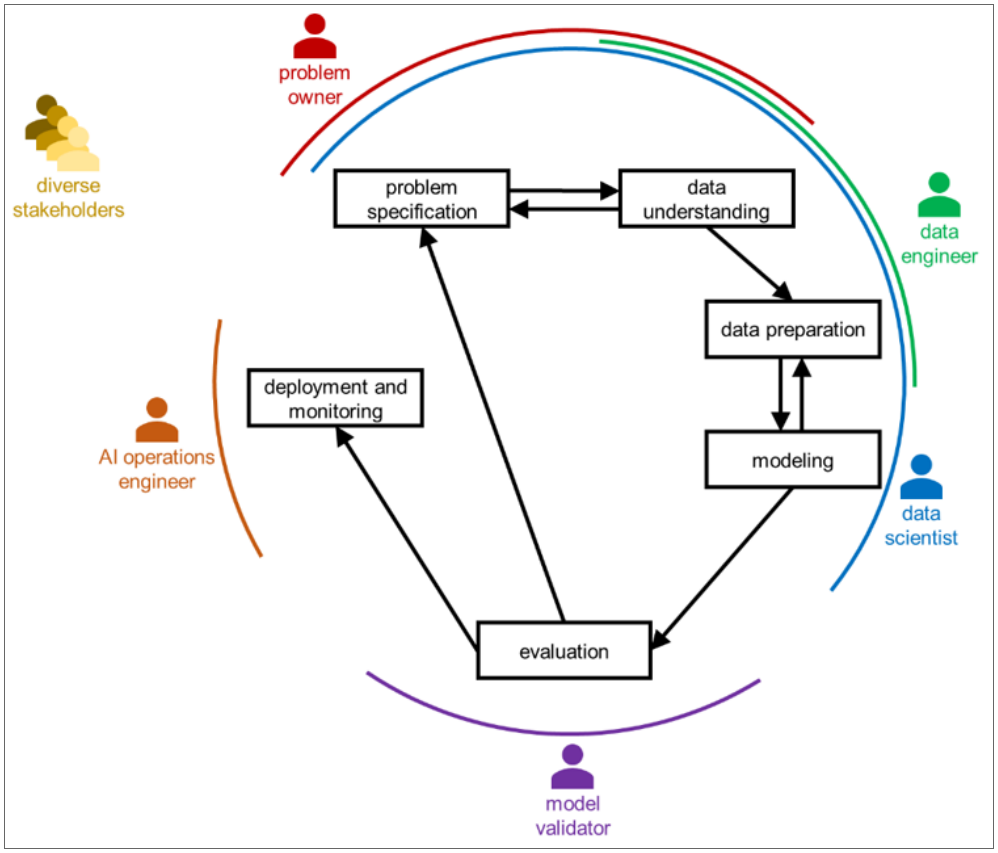


VERY QUICK INTRO TO MLOPS

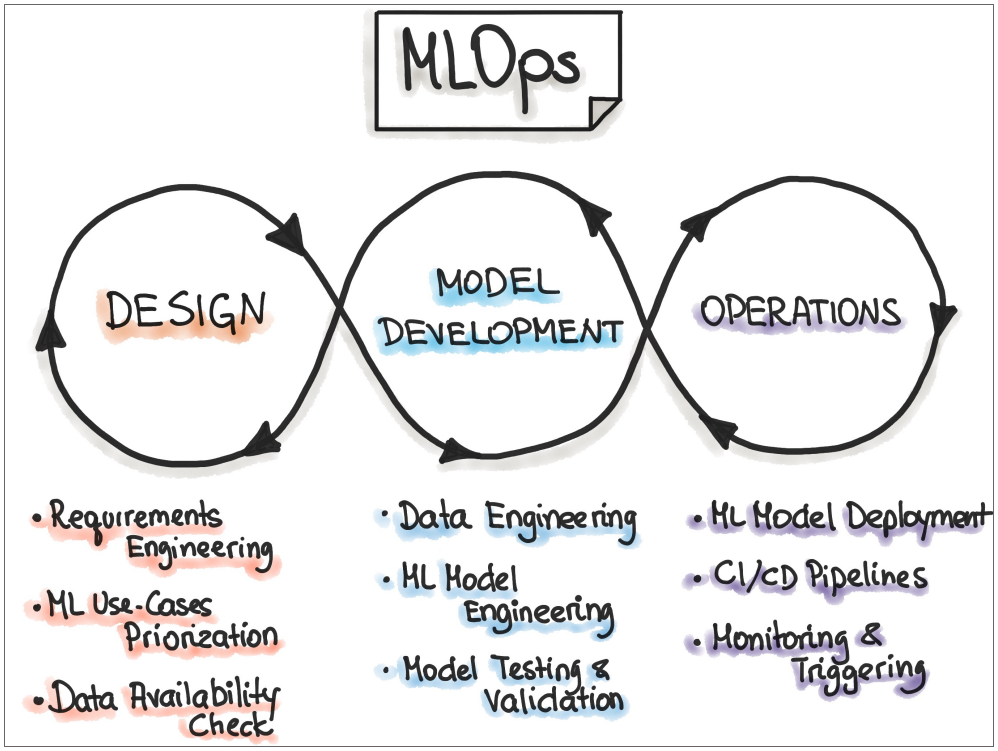
- https://huyenchip.com/machine-learning-systems-design/toc.html
- https://ml-ops.org/content/references.html


Machine learning project flow
Project setup
Data pipeline
Modeling & training
Serving

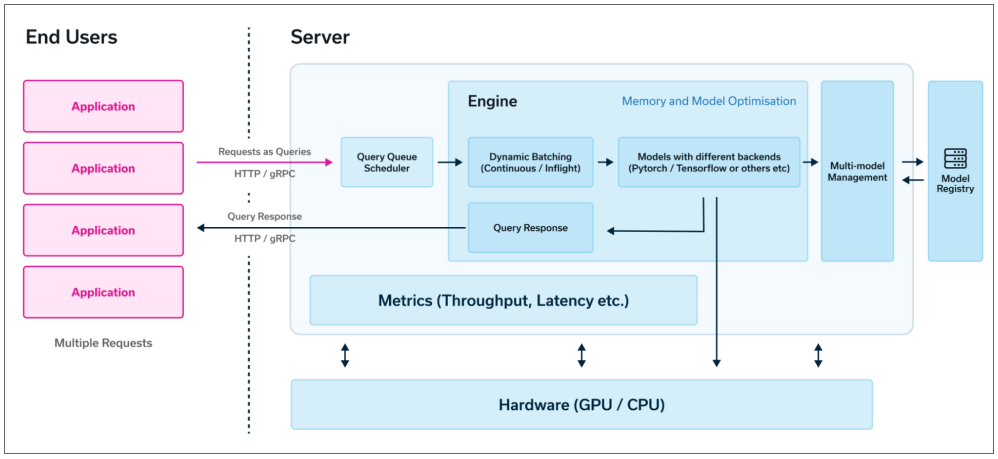
MLOps Lifecycle



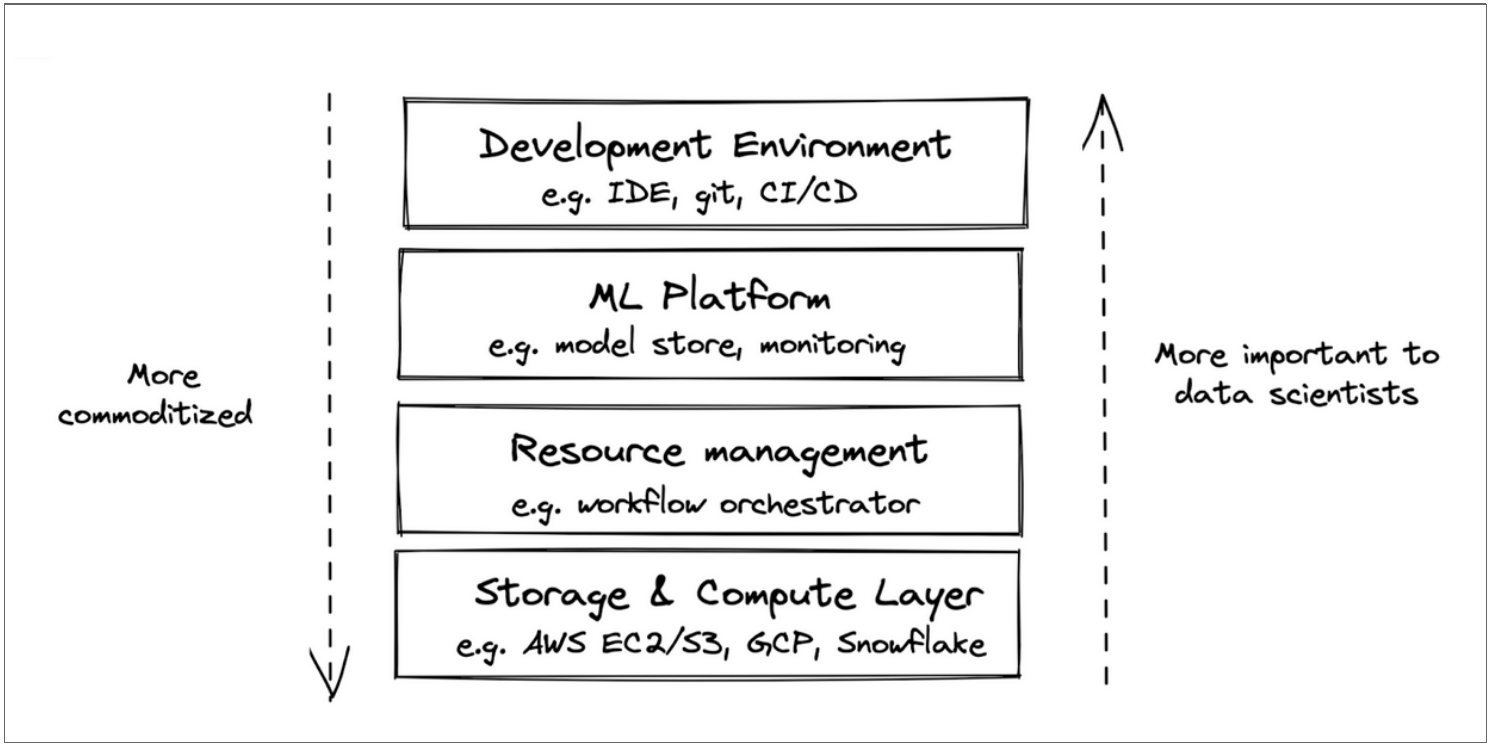
MLOps Loop


MLOps
DESIGN
MODEL DEVELOPMENT
OPERATIONS
• Requirements Engineering
• ML Use-Cases Priorization
• Data Availability Check
• Data Engineering
• ML Model Engineering
• Model Testing & Validation
• ML Model Deployment
• CI/CD Pipelines
• Monitoring & Triggering

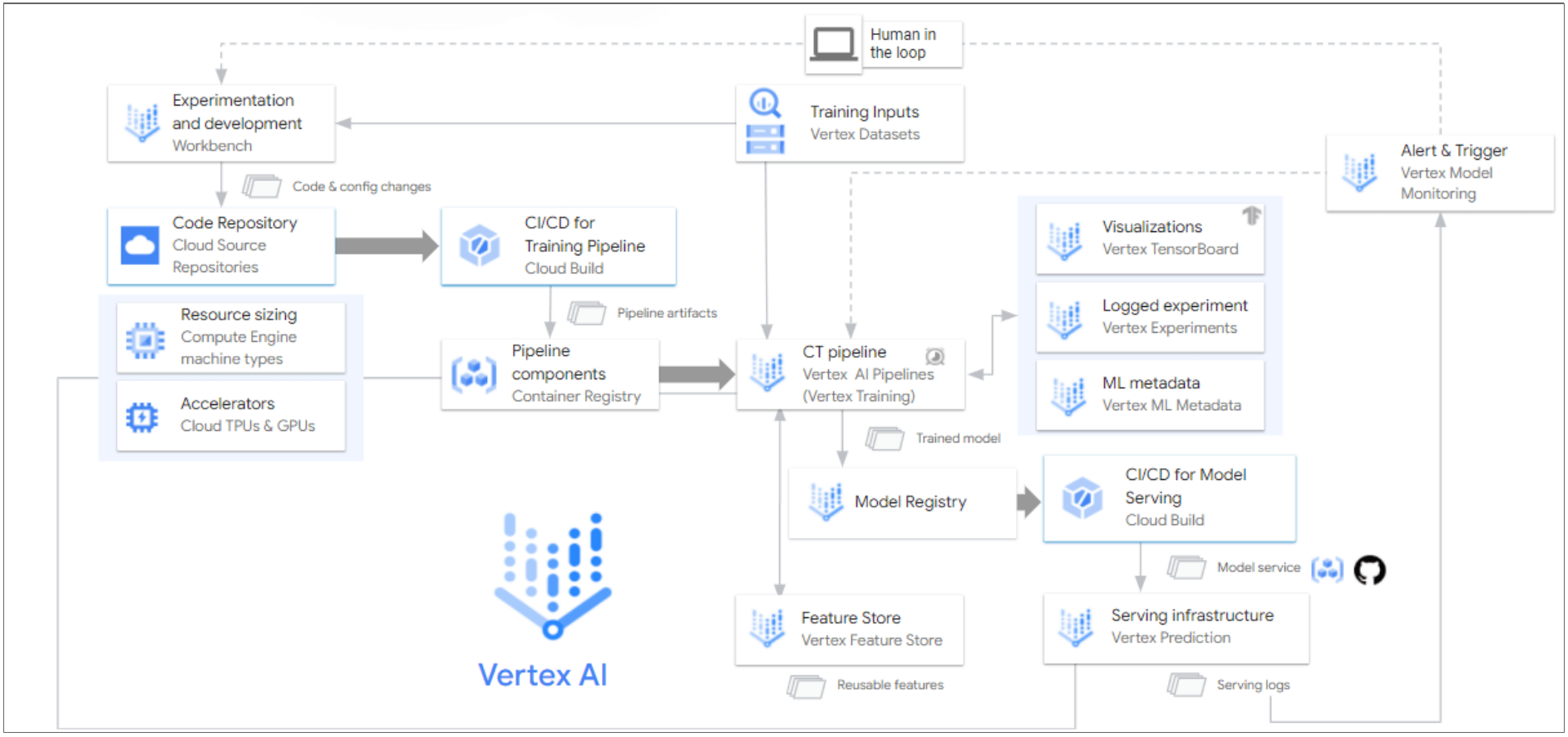
Deployment architecture



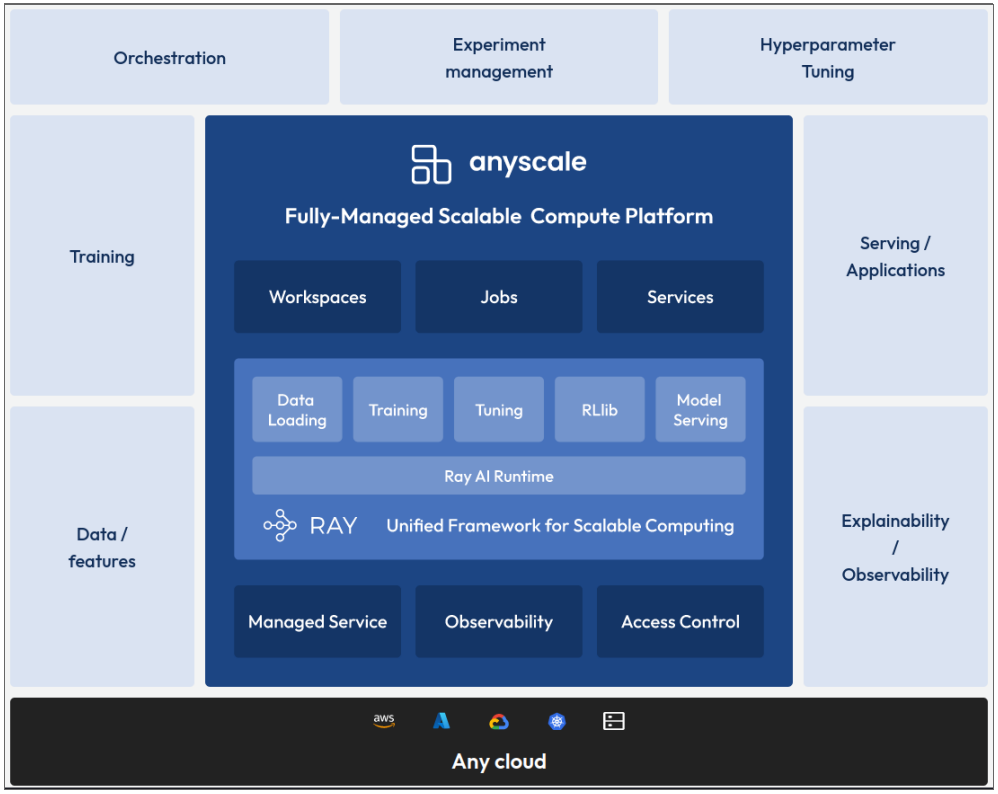
LAYERS OF "ENABLING TECHNOLOGY"



A FULL WORKFLOW



THE NEED FOR TECH



And dask !

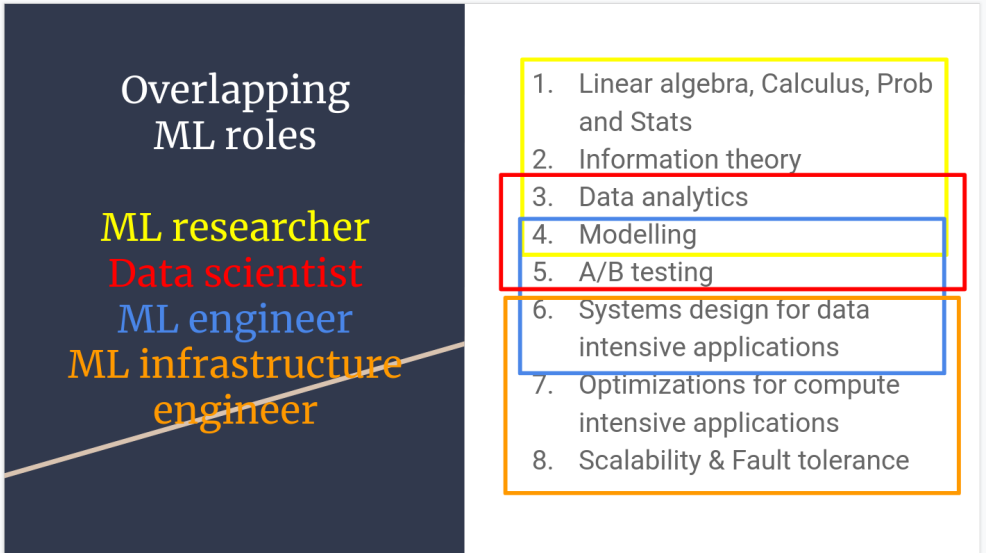
WHAT ABOUT ME ?

What does it mean for YOU ?

ML interviews

Broad spectrum of skills

1. Linear algebra, Calculus, Prob and Stats
2. Information theory
3. Data analytics
4. ML/DL Modelling
5. A/B testing
6. Systems design for data intensive applications
7. Optimizations for compute intensive applications
8. Scalability & Fault tolerance


Overlapping
ML roles
ML researcher
Data scientist
ML engineer
ML infrastructure
engineer
1. Linear algebra, Calculus, Prob and Stats
2. Information theory
3. Data analytics
4. Modelling
5. A/B testing
6. Systems design for data intensive applications
7. Optimizations for compute intensive applications
8. Scalability & Fault tolerance

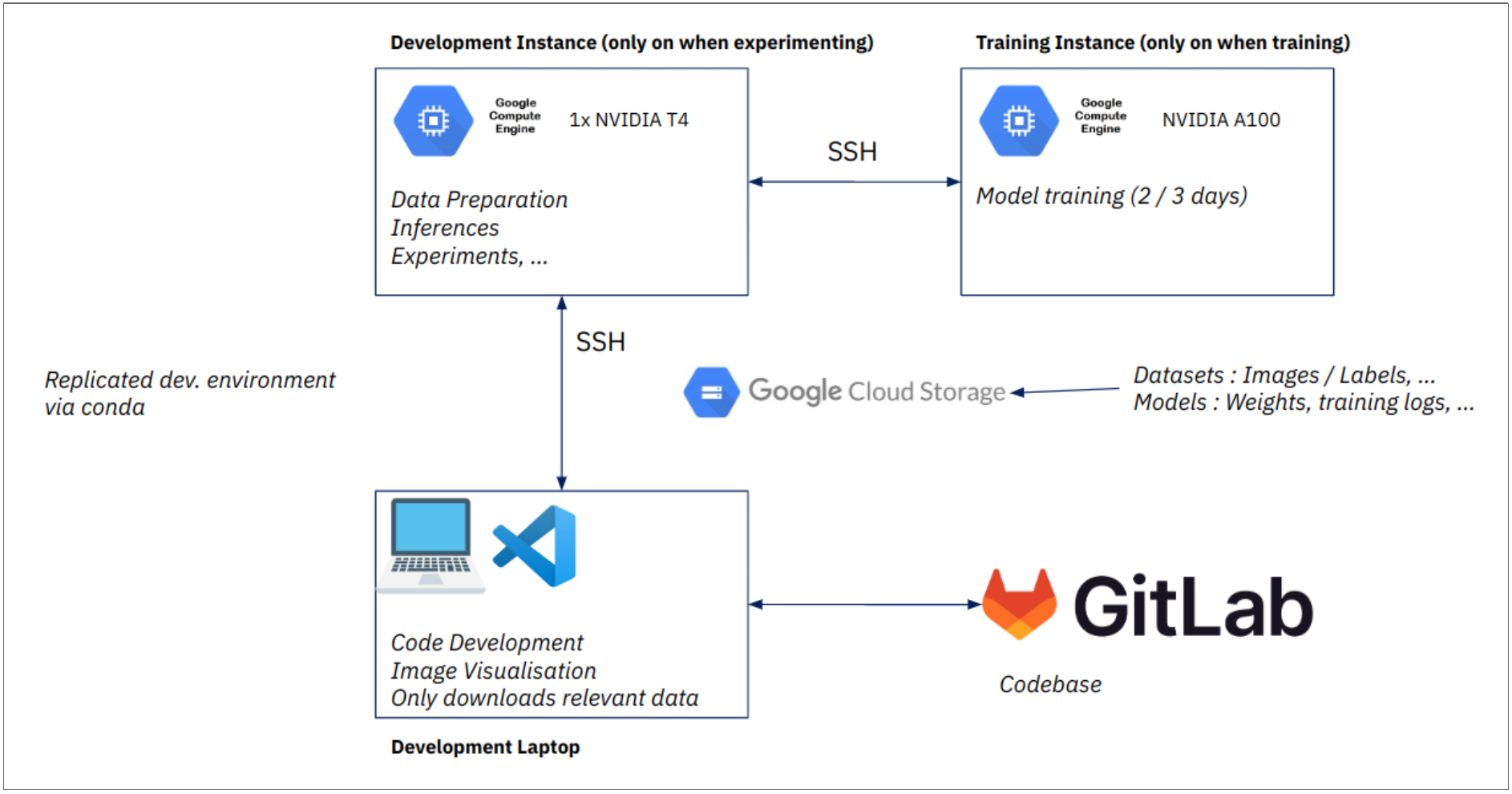
YOUR MILEAGE MAY VARY

depending on:

- Your company
- Your role

but you will "deal with" cloud computing one way or another !

PERSONAL EXPERIENCE



Speaker notes