Hardware and Software Technologies for Cloud Computing

Ozalp Babaoglu

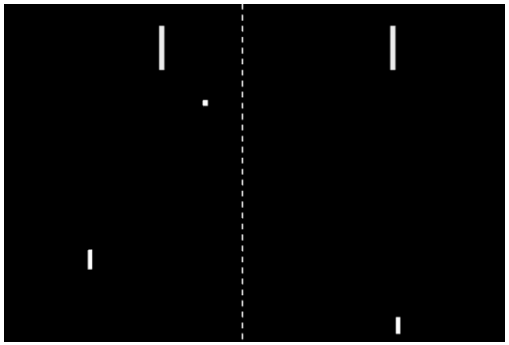
Computing

- Server farms built out of rack mount servers

Computing

- Early data centers were built out of commodity servers using commodity processors (x86 CPUs)
- Since then, many cloud providers have started building their data centers out of custom built servers for economic reasons with custom designed processors for performance reasons
- These servers often augment generic processors with Graphical Processing Units (GPUs)
- GPUs, originally designed for video games, are capable of performing certain scientific computations extremely fast because they can operate in parallel

Video Games 1970s and 80s



Video Games Today

Computing
GPU Farms

- For graphical rendering and scientific computing

Computing
Special Processing Units

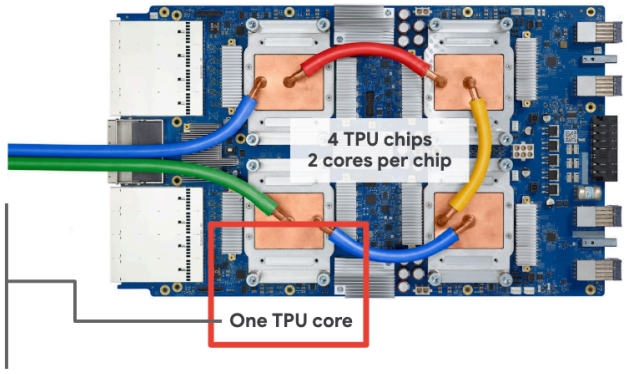
- With growing interest in Artificial Intelligence, cloud providers have started augmenting their servers with processing units specifically designed to speed up AI and HPC tasks
- Tensor Processing Unit (TPU) is an AI accelerator developed by Google specifically for neural network machine learning, particularly using Google's own TensorFlow software
- Provide acceleration for
  - AI Training
  - AI Inference
  - Advanced HPC

Computing
Tensor Processing Unit

TPU v3-8
420 teraflops
128 GB RAM
8 cores

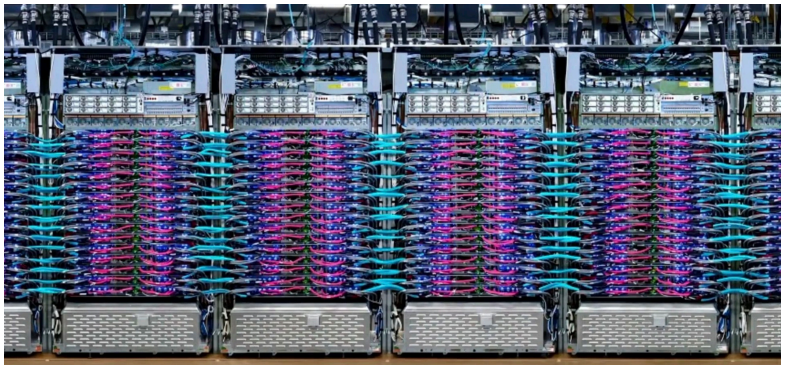
MXU
Matrix Multiply Unit
128x128 bfloat16 matrices

VPU
Vector Processing Unit
float32, int32



Computing
TPU Farms

- Google Cloud TPU v3 POD, 100+ PetaFlops (100×10^{15} Flops) 32 TB (32×10^{12} Bytes)

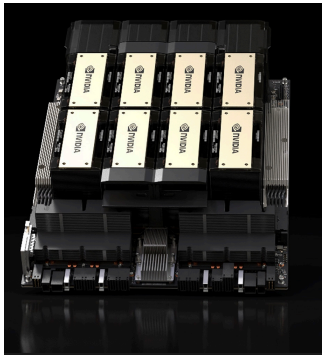


Computing
TPU Farms

- 100 PetaFlops = 100×10^{15}
  = 100 million billion Floating-point Operations Per Second
- 32 TeraBytes = $32 \times 10^{12} = 32 \times 10^{6} \times 10^{6}$
  = 32 million MegaBytes or 32 trillion Bytes

Computing
NVIDIA

- NVIDIA historically has been a producer of special-purpose graphical hardware for the gaming industry
- Recently, it has become the supplier of the most-advanced accelerator chips for Artificial Intelligence
- Their NVIDIA H200 Tensor Core GPU supercharges generative AI and HPC workloads



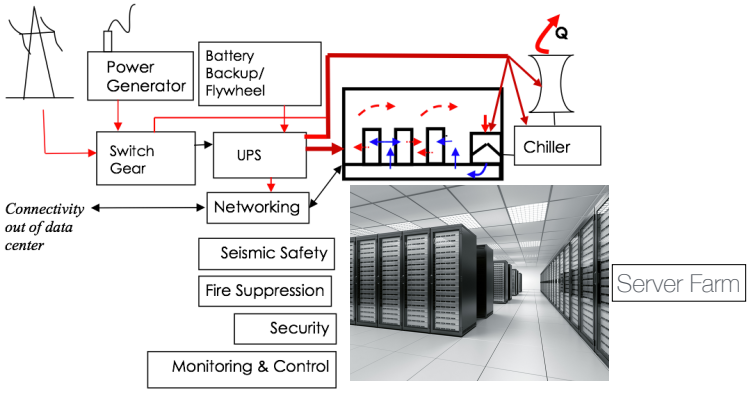
Computing
NVIDIA

- The recent surge in generative AI has transformed NVIDIA into the fourth company after Apple, Microsoft and Alphabet with a market capitalization over $2 trillion

Data Center Components

- Server farms are a central component of a data center
- A modern data center has many other components that are essential for its functioning
  - Power generation and distribution
  - Heat dissipation
  - Networking
  - Physical safety and security
  - Monitoring and control

Data Center Components

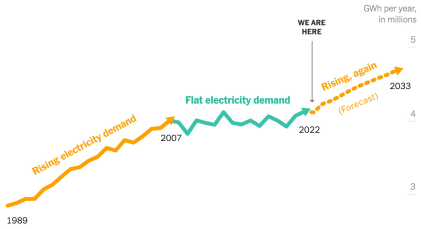


Power Consumption

- In 2010, data centers globally used about 194TWh (TeraWatt-hours) of electricity — about as much power as Iran used that year
- By 2018 that figure had increased to 205TWh (which is about 1% of all electricity consumed that year worldwide)
- While this represents a 6% increase in power usage, in the same period the amount of computing done in data centers had increased by 550%
- Today, data centers consume about 2% of electricity worldwide
- This could rise to 8% of the global total by 2030

Power Consumption

- Recent NYT article:



A New Surge in Power Use Is Threatening U.S. Climate Goals

A boom in data centers and factories is straining electric grids and propping up fossil fuels.

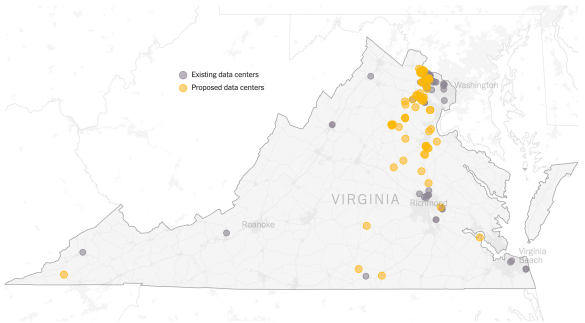
By Brad Plumer and Nadja Popovich March 13, 2024

Power Consumption

- Among the reasons for this trend are
  - The growth of remote work
  - Frenzied expansion of data centers due to video streaming and online shopping
  - The rise of artificial intelligence
- By 2030, electricity demand at U.S. data centers is expected to triple

Power Consumption

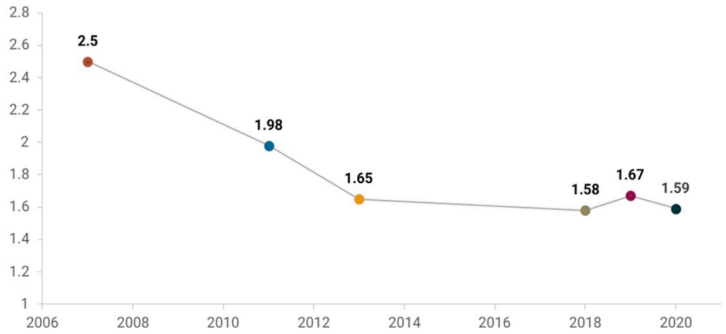
- According to the same article, at least 75 new data centers have opened in Virginia since 2019



Data Center Efficiency

- Measured in Power Usage Effectiveness (PUE)
- PUE = (Total Facility Energy) / (Computing Equipment Energy)
- PUE is always greater than 1, and the bigger the PUE, the less efficient the data center
- Ideally, PUE should be as close as possible to 1.0 but values 1.3 or less are considered very good
- With PUE = 1.3, for every Watt-hour of energy that goes into computing, 0.3 Watt-hour of energy goes to non-computing (cooling, lighting, surveillance)

Data Center Efficiency
PUE Over the Years

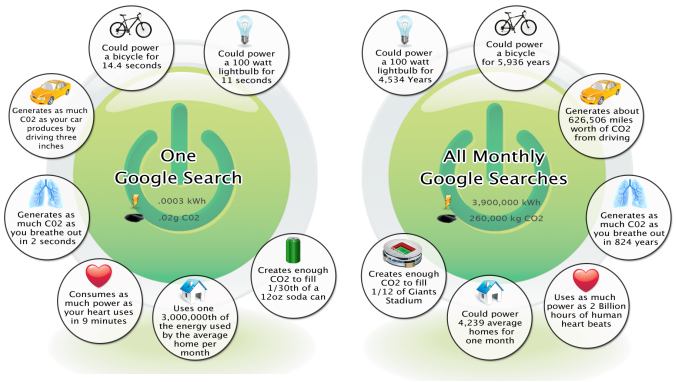


- The significant progress that was made until 2013 has been flattening out recently

Power Consumption and Carbon Emissions

- In April 2018, the music video for Despacito set an Internet record when it became the first video to hit *five billion views* on YouTube
- In the process, Despacito reached a less celebrated milestone: it consumed as much energy as 40,000 U.S. homes use in a year
- For comparison, an average Google search query consumes about 0.3Wh of energy, generating roughly 0.02g of carbon dioxide
- In one month, Google processes roughly 13 billion search queries
- In one month, Google search queries are responsible for emitting 260 metric tons of carbon dioxide into the atmosphere

Power Consumption and Carbon Emissions



Power Consumption and Carbon Emissions

- The U.S. is home to 3 million data centers, or roughly one for every 100 Americans
- Data centers contribute 0.3% to global carbon emissions
- The ICT sector as a whole contributes over 2%
- All major cloud providers have made commitments to achieving carbon neutrality at some near future either by switching to renewable energy sources or by paying for “carbon credits”

Renewable Energy — Solar

Renewable Energy — Wind

AWS

- Amazon claims to have reached a 50% share of renewable energy in 2018 and has a sustainability timeline for the future of their cloud operations
- Amazon's goal to reach 100% net carbon neutrality by 2040 however, seems to have vanished from its official statements

Microsoft Azure

- In 2012, Microsoft established an internal "tax" on its assumed carbon footprint to encourage operations to go carbon neutral
- The money has since been used for environmental projects with negative carbon footprints, very much like the way air travel compensations work
- Microsoft is on track to cutting operational carbon emissions by 75% by 2030
- https://vimeo.com/172088946

Google Cloud Platform

- Google started their plan to have a carbon neutral footprint back in 2009
- Since then, the company has reached its goal to effectively power all of their operations (including Google Cloud Platform) by renewable energy
- Since it is not possible for Google itself to produce renewable energy where it is needed, Google buys renewable energy from the same grid that their data centers are using

The Big Three (plus Alibaba)

Cloud provider	Estimated annual electricity consumption (TWh)	Percentage Renewable	Annual Carbon Debt (metric tons CO23)
Amazon Web Service	20	50%	5,150,000
Microsoft Azure	10	100%	150,000
Google Cloud	20	100%	300,000
Alibaba Cloud	3	10%	1,624,500

Datacenter Heat Dissipation

- Most of the (electrical) energy consumed in a data center is converted to heat which must be dissipated
- Heat dissipation technologies use air or some liquid (often water) for cooling
- Large data centers are often located in cold climates (Finland) or close to sources of natural cold water (Oregon)

Datacenter — Cooling Exterior

Datacenter — Cooling Interior

Cloud Storage Technologies

- Magnetic hard disks

Parameter	1956	2016
Capacity	3.75 MB	10 TB
Average access time	≈ 600 msec	2.5–10 ms
Density	200 bits/sq. inch	1.3 TB sq. inch
Average life span	≈ 2000 hours/MTBF	≈ 22, 500 hours/MTBF
Price	$9,200/MB	$0.032/GB
Weight	910 Kg	62 g
Physical volume	1.9 m^3	34 cm^3

- While density, price, and capacity have improved by astonishing factors (hundreds of million to one), access time and reliability have improved by much more modest factors
- Solid-state disks have become economic alternatives to magnetic hard disks while improving access time and reliability

Cloud Storage Technologies

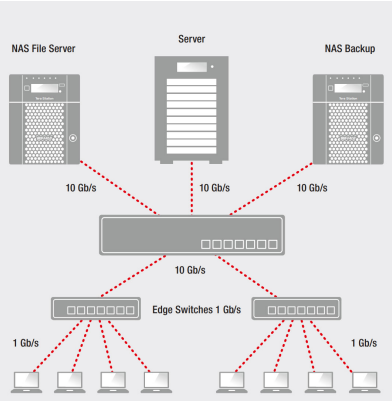
- Direct-Attached Storage (DAS)
- Network-Attached Storage (NAS)
- Storage Area Networks (SAN)
- Redundant Arrays of Inexpensive Disks (RAID)

Cloud Storage Technologies Direct-Attached Storage

- DAS is storage that is directly attached to a single computer
- It is not networked thus cannot be easily accessed by other devices
- Each DAS device is managed separately

Cloud Storage Technologies Network-Attached Storage

- NAS is a *file-level* computer data storage server connected to a computer network providing data access to a heterogeneous group of clients
- NAS can be an important part of an Infrastructure-as-a-Service



Cloud Storage Technologies
Network-Attached Storage

- NAS benefits:
  - Scale-out capacity: Adding more storage capacity to NAS is as easy as adding more hard disks
  - Performance: Because NAS is dedicated to serving files, it removes the responsibility of file serving from other networked devices
  - Easy setup: NAS architectures are often delivered as appliances preinstalled with a streamlined operating system, greatly reducing the setup time
  - Accessibility: Every networked device has access to NAS
  - Fault tolerance: NAS can be formatted to support replicated disks

Cloud Storage Technologies
Storage Area Networks

- SAN is a computer network which provides access to consolidated, *block-level* data storage
- SANs are primarily used to access storage devices that appear to the operating system as direct-attached storage
- SAN is an extension of DAS to a specialized local-area network, such as fibre channel

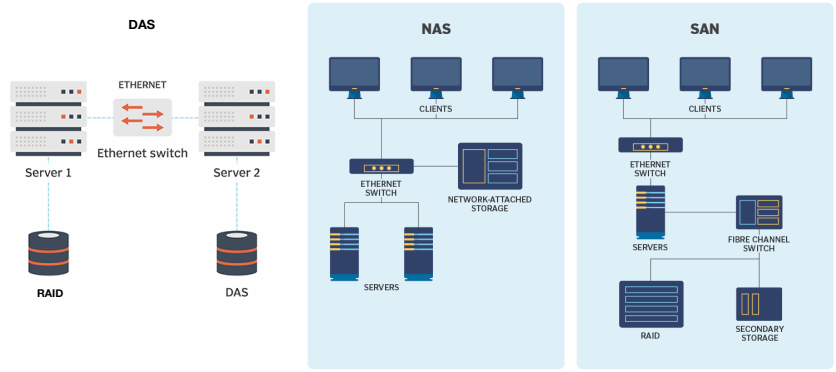
Cloud Storage Technologies
Redundant Arrays of Inexpensive

- RAID is a storage technology that combines several physical hard disks to create a logical drive with better performance and reliability than individual units
- It *increases speed* of storing and accessing data while *preventing loss* or corruption of data
- The manner in which data is distributed, organized and managed across multiple disks in an array defines the RAID level 0-7, 10
- Each level has a different fault tolerance, data redundancy, and performance properties, and the choice depends on requirements or goals as well as cost

Cloud Storage Technologies
RAID Levels

RAID LEVEL	METHOD	HARDWARE / SOFTWARE	MINIMUM # OF DISKS	COMMON USAGE	PROS	CONS
JBOD	SPANNING		2	INCREASE CAPACITY	COST-EFFECTIVE STORAGE	NO PERFORMANCE OR SECURITY BENEFITS
0	STRIPING		2	HEAVY READ OPERATIONS	HIGH PERFORMANCE (SPEED)	DATA IS LOST IF ONE DISK FAILS
1	MIRRORING		2	STANDARD APP SERVERS	FAULT TOLERANCE, HIGH READ PERFORMANCE	LAG FOR WRITE OPS, REDUCED STORAGE (BY 1/2)
5	STRIPING & PARITY		3	NORMAL FILE STORAGE & APP SERVERS	SPEED + FAULT TOLERANCE	LAG FOR WRITE OPS, REDUCED STORAGE (BY 1/3)
6	STRIPING & DOUBLE PARITY		4	LARGE FILE STORAGE & APP SERVERS	EXTRA LEVEL OF REDUNDANCY, HIGH READ PERFORMANCE	LOW WRITE PERFORMANCE, REDUCED STORAGE (BY 2/5)
10 (1+0)	STRIPING & MIRRORING		4	HIGHLY UTILIZED DATABASE SERVERS	WRITE PERFORMANCE + STRONG FAULT TOLERANCE	REDUCED STORAGE (1/2), LIMITED SCALABILITY

Cloud Storage Technologies Summary

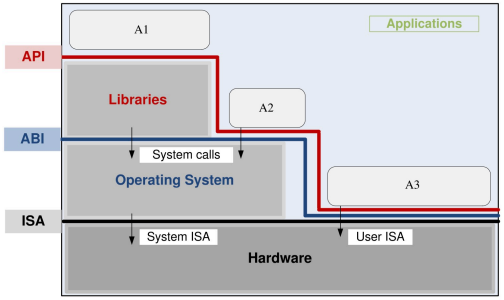


Software Technologies

- Cloud computing is based on the following software technologies
  - virtualization
  - containerization
  - scaling and fault tolerance
  - databases

Software Technologies Layering, Interfaces, Abstractions

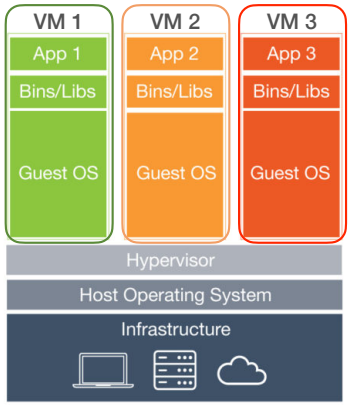
- Software components including applications, libraries, and operating system interact with the hardware via several interfaces: the Application Program Interface (API), the Application Binary Interface (ABI), and the Instruction Set Architecture (ISA)
- An application uses library functions (A1), makes system calls (A2), and executes machine instructions (A3)



Software Technologies Virtualization

- Virtualization is a technology for creating a software-based (virtual) representation of something, such as applications, servers, storage (e.g., RAID) or networks
- Processor virtualization, invented by IBM in the 1960s, creates multiple independent instances of the underlying hardware called *Virtual Machines* (VMs), running one or more operating systems
- Each VM appears to be running on the bare hardware

Software Technologies
Virtualization



- Virtualization is implemented through a software layer called a ***Hypervisor*** (or ***Virtual Machine Monitor***) running on top of the ***host operating system***
- Commercial hypervisors: ***VMware***, ***Zen***, ***Parallels***, ***VirtualBox***

Software Technologies
Virtualization

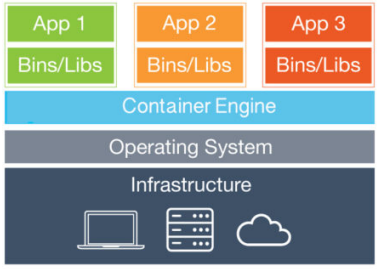
- To the cloud user, virtualization provides *isolation*, *encapsulation* and *hardware independence*
- To the cloud provider, virtualization provides *increased utilization* of its resources, resulting in cost saving and higher profits
- Isolation guarantees that multiple VMs running on the same server do not interfere
- Encapsulation allows VMs to be stopped, started and moved between servers (their state is just a file)
- Applications can be scaled-out by starting up additional VMs and scaled-in by stopping them, paving the way to achieving *elasticity*

Software Technologies
Containerization

- The problem with virtualization through a hypervisor is that VMs are *heavy-weight* and can take a *long time* to fire up
- Containers are an alternative technology based on *operating-system-level* virtualization rather than *hardware virtualization*
- A container is a standard unit of software that packages up code and all its dependencies so the application can be moved quickly and reliably from one computing environment to another

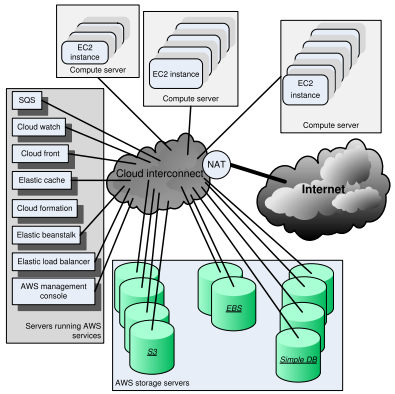
Software Technologies
Containerization

- An application running inside a container is isolated from other applications running in different containers and all applications are isolated from the underlying physical system
- Containers are light-weight, self-contained, highly portable and easier to monitor and manage
- If a VM is a *house*, a container is an *apartment*
- Examples: Docker, Kubernetes



Amazon Web Services

- AWS is the IaaS cloud offering from Amazon
- A set of services for developing and deploying cloud applications
- Services for computing, storage, monitoring and managing



AWS

- *Elastic Compute Cloud* (EC2)
  - Provides virtual machine instances
  - Ability to create EC2 instances on the fly according to demand
- *Auto Scaling*
  - Automatically adds or removes EC2 instances based on changing and predicted demand to achieve elastic provisioning and high availability
  - Scaling can be scheduled, dynamic or predictive
- *Amazon Elastic Container Service* (ECS)
  - Provides a highly scalable, fast, container management service that makes it easy to run, stop, and manage Docker containers on a cluster of EC2 instances
- *Elastic Beanstalk*
  - PaaS for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker

AWS

- *Simple Storage Service* (S3)
  - Store and retrieve any amount of data, at any time, from anywhere on the Internet
- *Simple Queue Service* (SQS)
  - Provides reliable messaging service that enables the decoupling of distributed systems and serverless applications
- *Elastic Block Store* (EBS)
  - Provides persistent block-level storage volumes for use with EC2 instances where a volume appears as a raw, unformatted and reliable physical disk

AWS

- *CloudFront*
  - A web service for content delivery
- *CloudWatch*
  - A monitoring infrastructure used by application developers, users, and system administrators to collect and track metrics important for optimizing the performance of applications
- *Virtual Private Cloud* (VPC)
  - Provides a bridge between the existing IT infrastructure of an organization and the AWS cloud; the existing infrastructure is connected securely to a set of isolated AWS compute resources

AWS EC2

- AWS offers several types of EC2 instances
  - *T2* – provide a baseline CPU performance
  - *M3 & M4* – provide a balance of compute, memory, and network resources
  - *C4* – use high performance processors and have the lowest price/compute performance
  - *R3* – are optimized for memory-intensive applications
  - *G2* – target graphics and general-purpose GPU applications
  - *I2* – are storage optimized
  - *D2* – deliver high disk throughput

AWS EC2

- Resources offered by M4, C4, and G2 instances

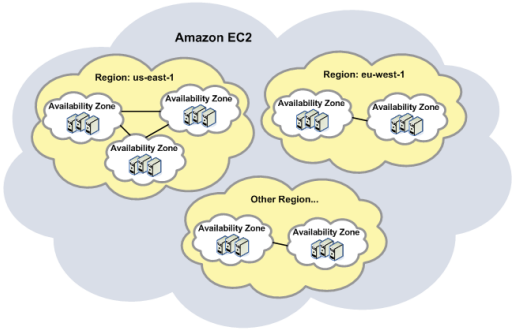
Instance type	vCPU	Memory (GiB)	EBS throughput (Mbps)	Cost ($/hour)
m4.large	2	8	450	0.12
m4.xlarge	4	16	750	0.239
m4.2xlarge	8	32	1 000	0.479
m4.4xlarge	16	64	2 000	0.958
m4.10xlarge	40	160	4 000	2.394
c4.large	2	3.75	500	0.105
c4.xlarge	4	7.5	750	0.209
c4.2xlarge	8	15	1 000	0.419
c4.4xlarge	16	30	2 000	0.838
c4.8-xlarge	36	60	4 000	1.675
g2.2xlarge	8	15	–	0.65
g2.4xlarge	32	60	–	2.60

AWS Regions, Availability Zones

- ***AWS Regions*** are physical locations around the world where Amazon clusters its ***data centers***
- A ***group of logical data centers*** constitutes an ***Availability Zone*** (AZ)
- Each Region consists of ***multiple***, ***isolated***, and ***physically separate*** AZs within a geographic area
- AZs have ***independent power***, ***cooling***, and ***physical security*** and are interconnected with high-bandwidth, low-latency networking, over ***fully redundant***, ***dedicated metro fiber***
- AWS customers can design their applications to run in multiple AZ's to achieve greater fault-tolerance

AWS Regions, Availability Zones

- Relationship among regions and availability zones



AWS Regions, Availability Zones

- 76 Availability Zones in over 24 geographic Regions around the world

AWS Regions

- https://www.cloudping.info/

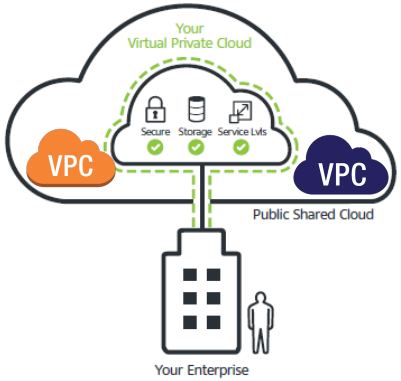
Region	Latency
US-East (Virginia)	114 ms
US East (Ohio)	125 ms
US-West (California)	173 ms
US-West (Oregon)	224 ms
Canada (Central)	125 ms
Europe (Ireland)	48 ms
Europe (London)	37 ms
Europe (Frankfurt)	26 ms
Europe (Paris)	33 ms
Europe (Stockholm)	60 ms
Middle East (Bahrain)	137 ms
Asia Pacific (Hong Kong)	318 ms
Asia Pacific (Mumbai)	139 ms
Asia Pacific (Osaka-Local)	251 ms
Asia Pacific (Seoul)	327 ms
Asia Pacific (Singapore)	268 ms
Asia Pacific (Sydney)	318 ms
Asia Pacific (Tokyo)	264 ms
South America (São Paulo)	219 ms
China (Beijing)	986 ms
China (Ningxia)	269 ms
AWS GovCloud (US-East)	120 ms
AWS GovCloud (US)	218 ms

Virtual Private Cloud

- VPC is a commercial service offered by a cloud provider that creates a virtual private cloud by provisioning a *logically isolated section* of the public cloud
- The virtual private cloud closely resembles a traditional private cloud that you would operate in your enterprise data center, with the benefits of using the scalable infrastructure of the cloud provider
- Enterprise customers are able to access the VPC over an IPsec-based virtual private network

Virtual Private Cloud

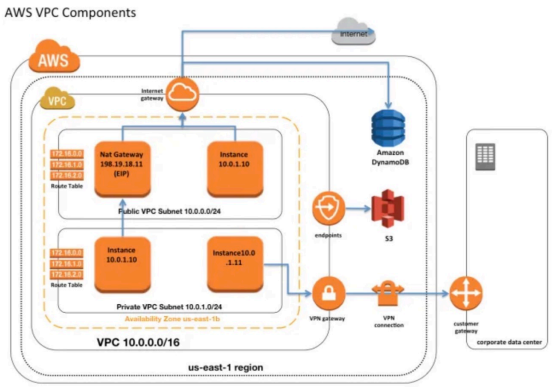
- An organization may create multiple VPCs within the public cloud, potentially all from different providers



AWS and VPC

- Unlike traditional EC2 instances which are allocated internal and external ***IP numbers by Amazon***, VPC customers can assign IP numbers ***of their choosing*** from one or more subnets
- By giving the user the option of selecting which AWS resources are ***public facing*** and ***which are not***, VPC provides a ***finer control*** over security
- VPC is Amazon's answer to growing interest in ***private and hybrid clouds***
- Potential risk of ***vendor lock-in*** since the private part is within the provider's domain

AWS and VPC



AWS Local Zones

- ***Local Zone*** location is an ***extension*** of an AWS Region where customers place resources such as compute and storage in multiple locations closer to end users and run latency sensitive applications using AWS services
- Resources are ***not replicated*** across Regions unless customers specifically choose to do so

AWS Local Zones

- Local zones allow applications to reduce latency times