



BITS Pilani

Cloud Computing

SEWP ZG527

Agenda



- ❖ Virtualization Recap
- ❖ Infrastructure as a Service
 - ❖ What is IaaS
 - ❖ Introduce AWS
 - ❖ AWS Reference Model
 - ❖ AWS Compute
 - ❖ AWS Storage
 - ❖ AWS Network
 - ❖ AWS Case Study - Netflix



BITS Pilani

Pilani | Dubai | Goa | Hyderabad



Recap



What is Virtualization?

Virtualization Defined



Virtualization is a computer architecture technology by which multiple virtual machines (VMs) are multiplexed in the same hardware machine.



Virtualization allows multiple operating system instances to run concurrently on a single computer



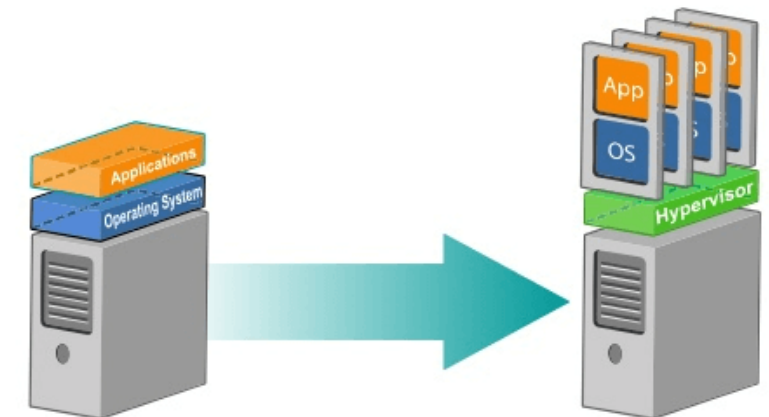
Instead of purchasing and maintaining an entire computer for one application, each application can be given its own operating system, and all those operating systems can reside on a single piece of hardware.



Virtualization allows an operator to control a guest operating system's use of CPU, memory, storage, and other resources, so each guest receives only the resources that it needs.

Key Terms:

- VM → Virtual Machine
- VMM → Virtual Machine Monitor
- Hypervisor → VMM
- Multiplexed → Many or several
- Host → System where the VMM resides
- Guest → Virtual Machines created



Virtualization Comparison

	Full Virtualization with Binary Translation	Hardware Assisted Virtualization	OS Assisted Virtualization / Paravirtualization
Technique	Binary Translation and Direct Execution	Exit to Root Mode on Privileged Instructions	Hypercalls
Guest Modification / Compatibility	Unmodified Guest OS Excellent compatibility	Unmodified Guest OS Excellent compatibility	Guest OS codified to issue Hypercalls so it can't run on Native Hardware or other Hypervisors Poor compatibility; Not available on Windows OSes
Performance	Good	Fair Current performance lags Binary Translation virtualization on various workloads but will improve over time	Better in certain cases
Used By	VMware, Microsoft, Parallels	VMware, Microsoft, Parallels, Xen	VMware, Xen
Guest OS Hypervisor Independent?	Yes	Yes	XenLinux runs only on Xen Hypervisor VMI-Linux is Hypervisor agnostic

Virtualization

Virtualization

Hardware

- Full
 - Bare-Metal
 - Hosted
- Partial
- Para

Network

- Internal Network Virtualization
- External Network Virtualization

Storage

- Block Virtualization
- File Virtualization

Memory

- Application Level Integration
- OS Level Integration

Software

- OS Level
- Application
- Service

Data

- Database

Desktop

- Virtual desktop infrastructure
- Hosted Virtual Desktop



BITS Pilani

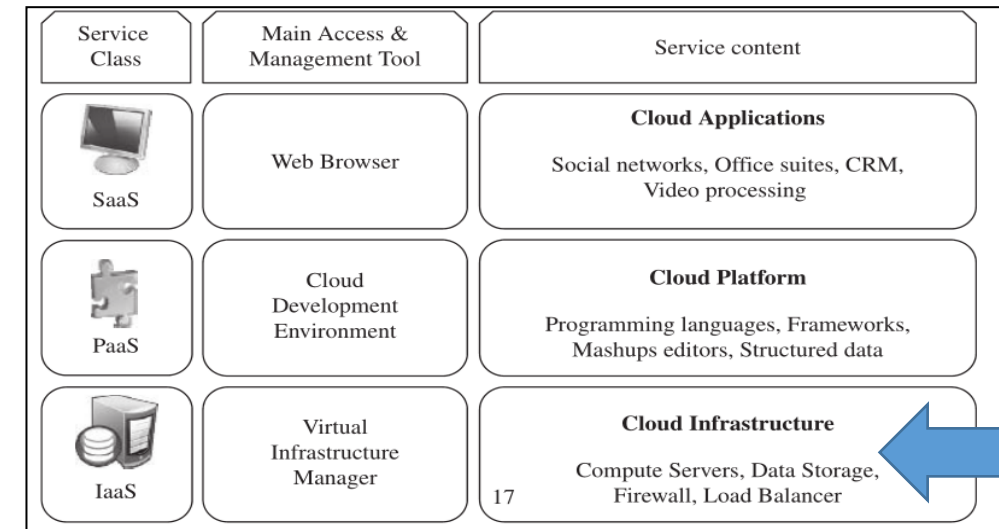
Pilani | Dubai | Goa | Hyderabad

Infrastructure as a Service



What is IaaS?

- The capability provided to the consumer is to provision processing, storage, networks, and other fundamental computing resources.
- The consumer is able to deploy and run arbitrary software, which can include operating systems and applications.
- The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed applications, and possibly limited control of select networking components (e.g., host firewalls).
- Offering virtualized resources (computation, storage, and communication) on demand is known as Infrastructure as a Service (IaaS).
- Infrastructure services are considered to be the bottom layer of cloud computing systems .
- Ex : **Amazon EC2** : Elastic Compute Cloud, Eucalyptus, GoGrid, Rackspace Cloud



Infrastructure as a Service

Why IaaS



Capability

IaaS provides the following

- Servers- compute, machines
- Storage
- Network
- Operating system

Characteristics

Resources are distributed as a service

- Allows dynamic scaling (1...10....100.....)
- Has a variable costs-
- Generally includes multiple-users on a single piece of hardware. (**multi-tenancy**)

Enabler : Virtualization Technology

- ✓ Manageability and Interoperability
- ✓ Availability and Reliability
- ✓ Scalability and Elasticity

Models

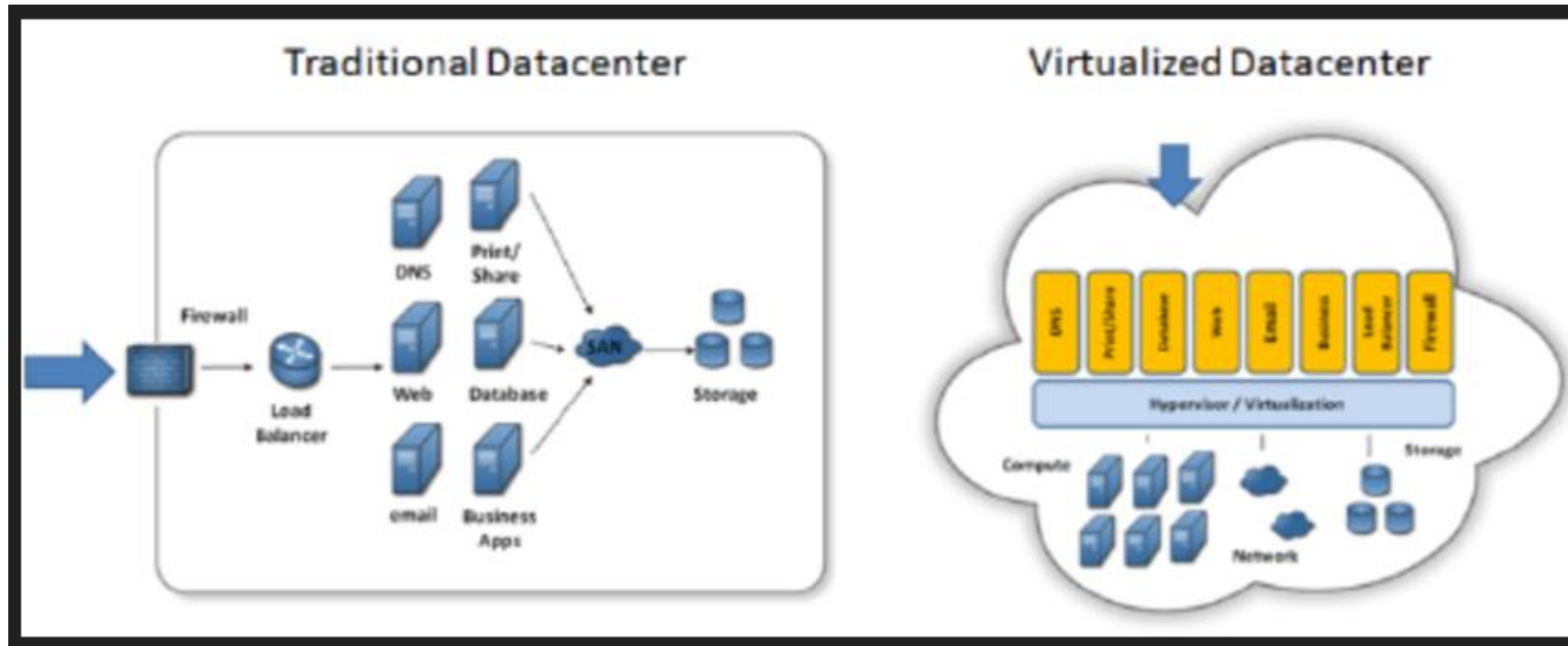
IaaS can be obtained as

- (1) Public or
- (2) Private infrastructure or
- (3) combination of both

Benefit

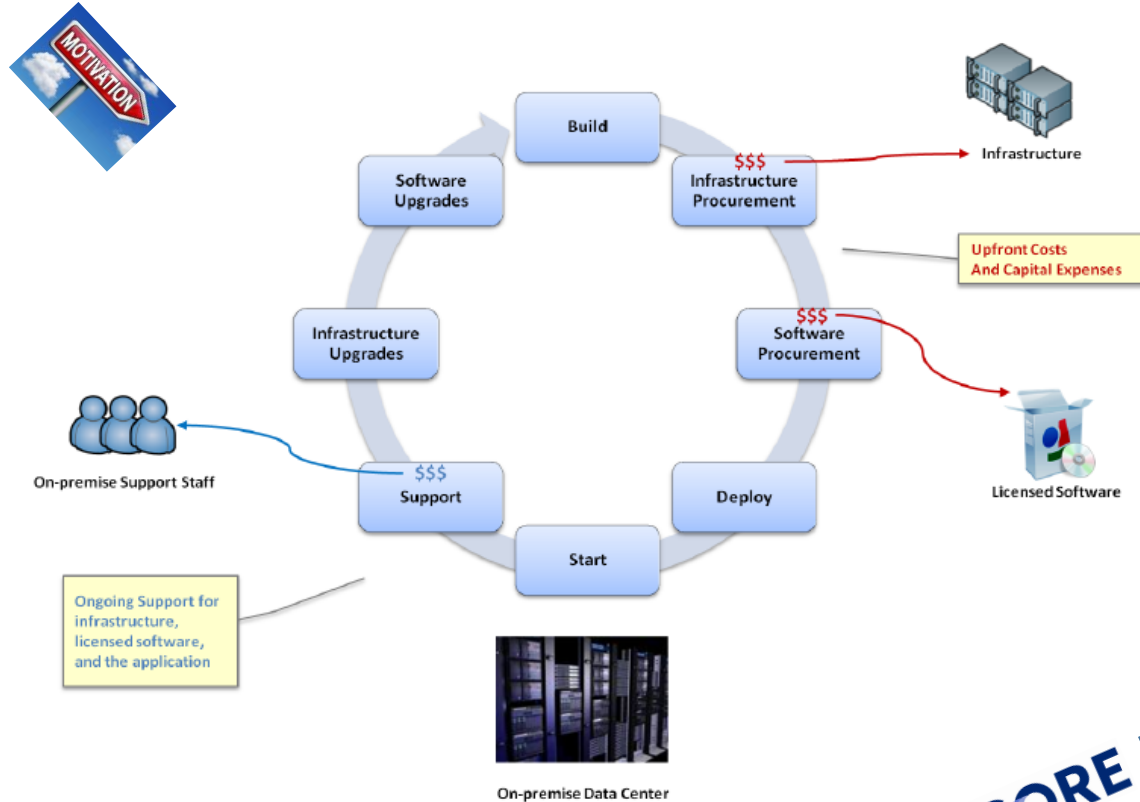
The user instead of **purchasing**- servers, software, **data center space** or network equipment, **rent** those resources as a fully **outsourced** service on-demand model.

Visualization for a New Paradigm



- IAAS , is unlike traditional IT infrastructure.
- End users will not have access to the physical machines.
- Each Hardware resource is provided as a service rather than a physical device
- Flexible/Secure/ Scalable direct connectivity.
- SLA with 99.9% uptime
- Pay as you go policy.
- Payment terms applicable for, resource, bandwidth used, data transferred etc

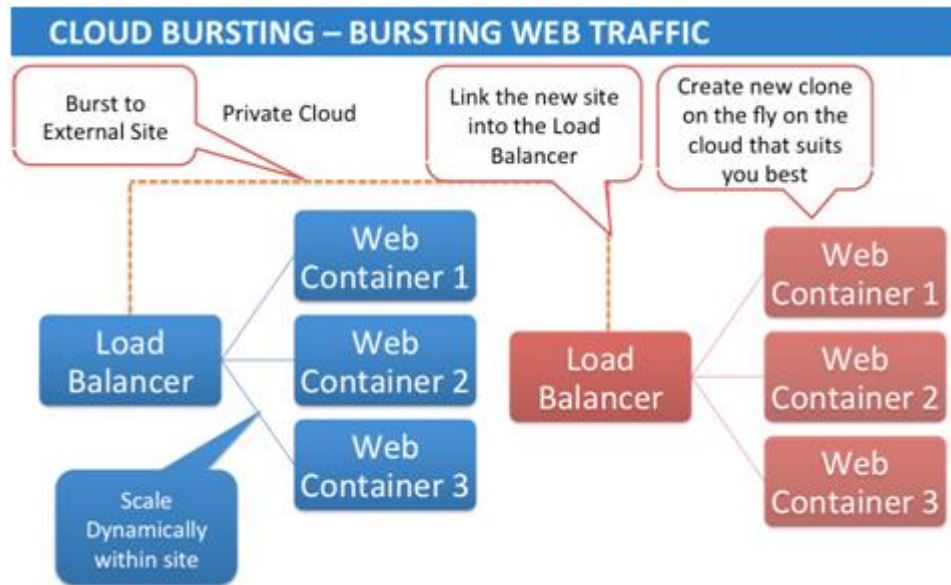
IaaS - Motivations



- No upfront cost
- Multi-tier auto provisioning within hours
- Scalability, services/resources are provided on demand
- Flexible/Secure direct connectivity
- SLA with 99.9% uptime



IaaS – Key Terms



•**Telemetry:** The process of automatic measurement & transmission of data within the cloud ecosystem to a centralized monitoring unit.

•**Cloudbursting:** The process of off-loading tasks to the cloud during times when demand exceeds capacity and the most compute resources are needed.

•**Resource pooling:** **Pooling** is a resource management term that refers to the grouping together of resources (compute(cpu), network(bandwidth), storage) for the purposes of *maximizing usage* and/or *minimizing risk* to the users.

•**Multi-tenant computing:** Multi-tenancy is an architecture in which a single instance of a software application serves multiple customers. Each customer is called a tenant. Tenants may be given the ability to customize some parts of the application, such as color of the user interface ([UI](#)) or [business rules](#), but they cannot customize the application's [code](#).

•**Hypervisor:** Software which enables virtualization.

Pros & Cons of IaaS

IaaS helps

1. Where demand is very **volatile**- encountering **spikes and troughs**.
2. For new enterprise without **capital to invest in hardware** or entrepreneurs starting on a shoestring budget.
3. Where the enterprise is growing rapidly and scaling hardware would be problematic.
4. For specific line of business, trial or temporary infrastructural needs
5. When you need computing power on the go, turn to IaaS.

IaaS Negates

- Where regulatory **compliance** makes the offshoring or outsourcing of data storage and processing difficult
- Where the **highest levels of performance** are required, and on premise or dedicated hosted infrastructure has the capacity to meet the organization's needs



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introducing Amazon Web Service



What is AWS



Amazon Web Services (AWS) is the world's most comprehensive and broadly adopted cloud platform, offering over 200 fully featured services from data centers globally. Millions of customers—including the fastest-growing startups, largest enterprises, and leading government agencies—are using AWS to lower costs, become more agile, and innovate faster.



Global Infrastructure: AWS serves over one million active customers in more than 190 countries, and it continues to expand its global infrastructure

Security: All AWS customers benefit from data center and network architectures built to satisfy the requirements of the most security-sensitive organizations.

- Application building blocks
- Stable APIs
- Proven Amazon infrastructure
- Focus on innovation and creativity
- Long-term investment



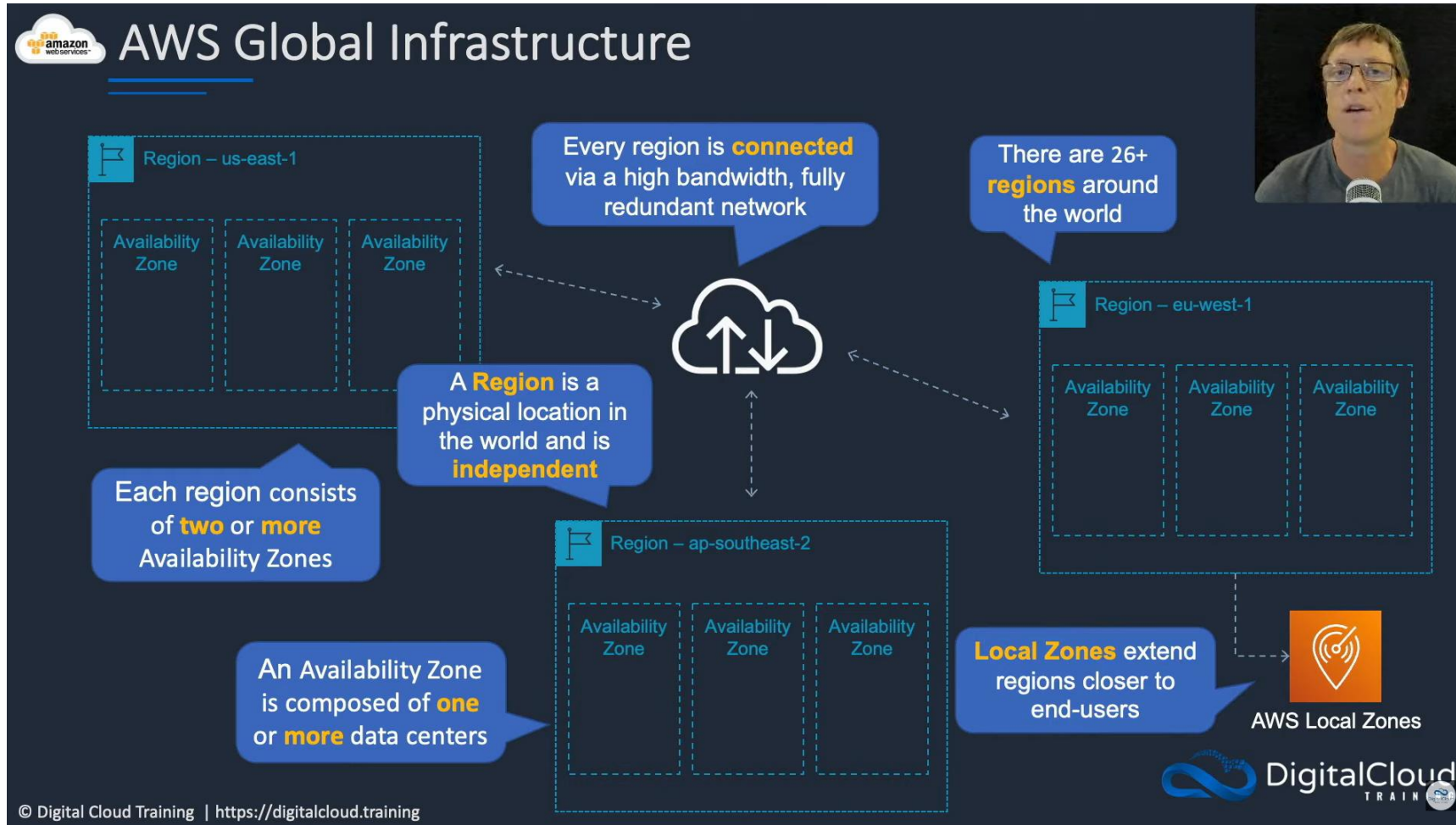
AWS Understanding Service Offering

- AWS operates **state-of-the-art, highly available data centers**. Although rare, failures can occur that affect the availability of instances that are in the same location.
- If you **host all of your instances** in a **single location** that is affected by a failure, **none of your instances would be available**.
- Amazon EC2 is hosted in multiple locations world-wide. These locations can be further classified as **AWS Regions**, **Availability Zones**, **Local Zones**, **AWS Outposts**, and **Wavelength Zones**.



Availability Zones located in **AWS Regions** consist of one or more **discrete data centers**, each of which has **redundant power**, **networking**, and **connectivity**, and is housed in **separate facilities**. Each AZ has multiple internet connections and power connections to multiple grids.

AWS Global Infrastructure



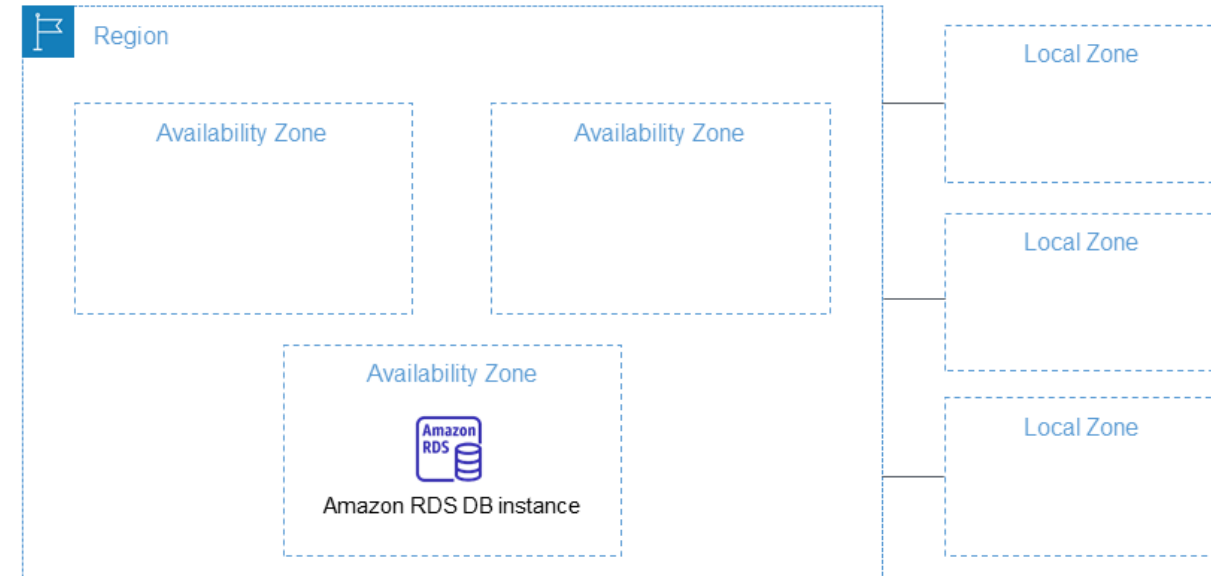
AWS Global Infrastructure

- Services provided via Region, Availability Zone, Local Zone, Wavelength Zone & Outposts
- **Region**: Denotes a Physical location across the globe. Currently there are 32 regions.
- **Availability Zone**: Physical location within a region where the data centers are built. Currently there are 102+ AZ's
- **Local Zone**: Consider this as an extension of an AZ closer to the user's location.
- **Wavelength Zone**: Creates an AWS infrastructure on the edge or wireless network of telcos.

AWS Outposts: Creates an AWS infrastructure on the premise of the customer. Provides a seamless hybrid cloud experience

AWS Regions

- AWS provides a **highly available technology infrastructure platform** with multiple **locations worldwide**. These locations are composed of **regions** and **Availability Zones**. Each **region** is a **separate geographic area**.
- It is important to remember that each AWS Region is completely independent. Any Amazon service you initiate (for example, creating database instances or listing available database instances) runs only in your current default AWS Region.
- The default AWS Region can be changed in the console, or by setting the `AWS_DEFAULT_REGION` environment variable. Or it can be overridden by using the `--region` parameter with the AWS Command Line Interface (AWS CLI).
- Each AWS Region is designed to be isolated from the other AWS Regions. This design achieves the greatest possible fault tolerance and stability.

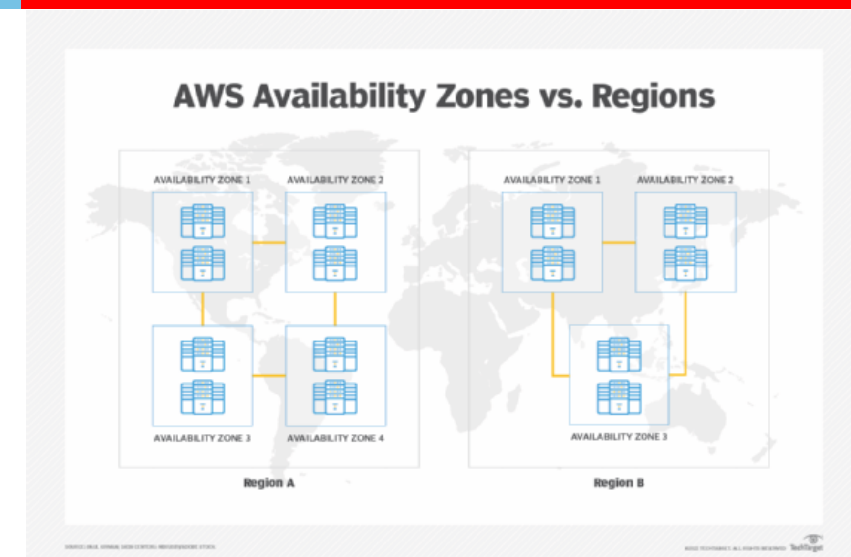


Availability Zones located in **AWS Regions** consist of one or more **discrete data centers**, each of which has **redundant power**, **networking**, and **connectivity**, and is housed in **separate facilities**.

Each AZ has multiple internet connections and power connections to multiple grids.

AWS Availability Zones

- Each Region has multiple, isolated locations known as **Availability Zones**. Each **Availability Zone** is also **isolated**, but the **Availability Zones** in a **region** are connected through **low-latency links**.
- Availability Zones** are **physically separated** within a typical metropolitan region and are located in lower-risk flood plains (specific flood zone categorization varies by region). In addition to using a **discrete uninterruptible power supply (UPS)** and on-site backup generators, they are each fed via **different grids** from **independent utilities** (when available) to reduce single points of failure further.
- Availability Zones** are all **redundantly connected** to multiple tier-1 transit providers. By placing resources in **separate Availability Zones**, you can protect your website or application from a **service disruption** impacting a single location.
- The code for Availability Zone is its Region code followed by a letter identifier. For example, us-east-1a.



- If you **distribute** your instances across **multiple Availability Zones** and one instance fails, you can design your application so that an instance in another Availability Zone can handle requests.

AWS Local Zones

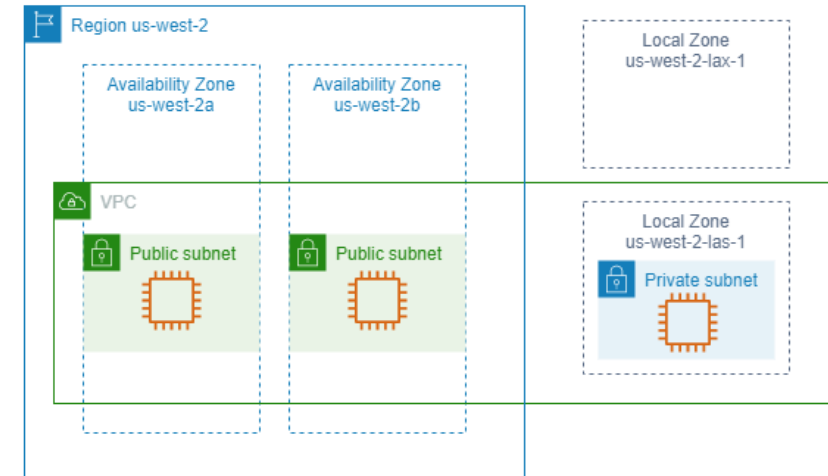
A **Local Zone** is an extension of an **AWS Region** in geographic proximity to your users.

Local Zones have their own connections to the internet and support AWS Direct Connect, so that resources created in a Local Zone can serve local users with low-latency communications.

The code for a Local Zone is its Region code followed by an identifier that indicates its physical location. For example, `us-west-2-lax-1` in Los Angeles.

The **VPC** spans the **Availability Zones** and one of the Local Zones. Each **zone** in the **VPC** has one **subnet**, and each **subnet has an instance**.

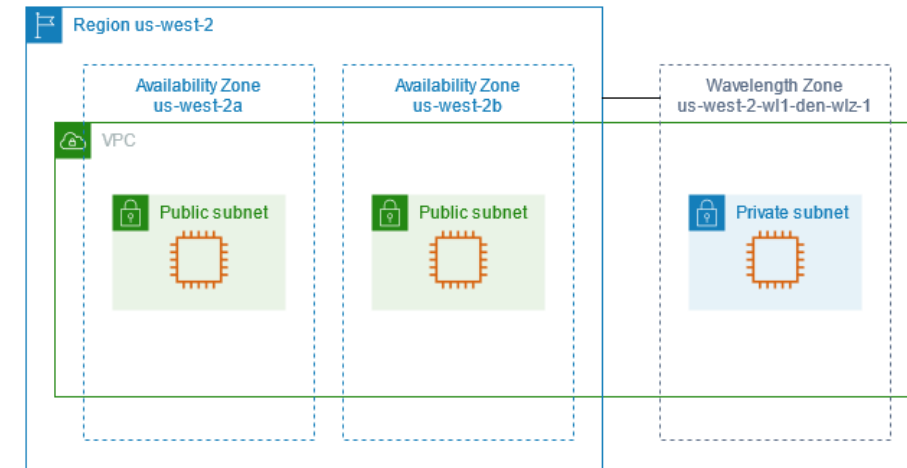
When you **launch an instance**, you can specify a **subnet that is in a Local Zone**. You also allocate an IP address from a network border group. A network border group is a unique set of Availability Zones, Local Zones, or Wavelength Zones from which AWS advertises IP addresses, for example, `us-west-2-lax-1a`.



- Some **AWS resources** might not be available in all **Regions**. Make sure that you can create the resources that you need in the desired Regions or Local Zones before launching an instance in a specific Local Zone
- Before you can **specify a Local Zone** for a **resource** or service, you must **opt in** to Local Zones.

AWS Wavelength Zones

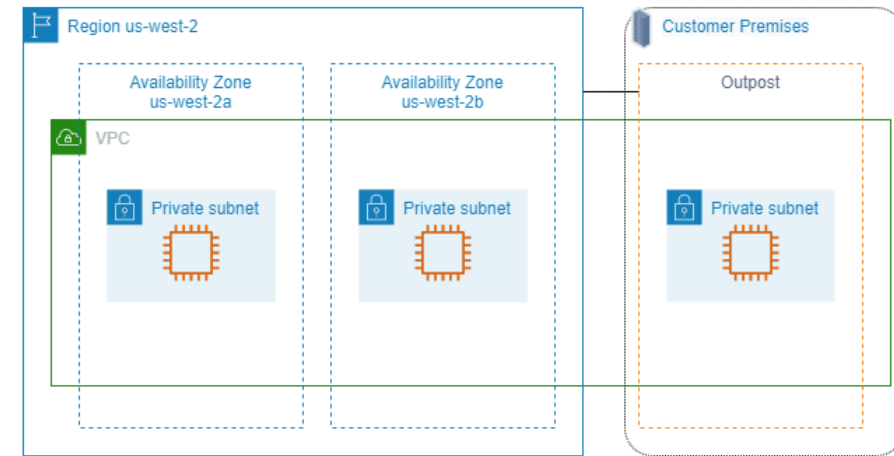
- **Wavelength Zones** are **AWS infrastructure deployments** that embed AWS compute and storage services within **telecommunications providers' data centers** at the edge of the **5G network**, so application traffic can reach application servers running in Wavelength Zones without leaving the mobile providers' network.
- This **prevents the latency** that would result from multiple hops to the internet and enables customers to take **full advantage of 5G networks**. Wavelength Zones extend **AWS to the 5G edge**, delivering a consistent developer experience across **multiple 5G networks around the world**. Wavelength Zones also allow developers to **build** the next generation of **ultra-low latency applications** using the same **familiar AWS services**, APIs, tools, and functionality they already use today.
- Processing at the network edge can help avoid transmitting large volumes of data over the network provider's infrastructure, and offload processing from mobile device hardware.



This enables new classes of compute-intensive, latency sensitive applications. For example, a **fleet of autonomous cars** interacting with road sensors to prevent crashes, smart industrial robots assessing and reacting to plant conditions in a dangerous manufacturing environment, or retailers serving **personalized promotions** to shoppers' **mobile phones** in real time as they pass **product displays**.

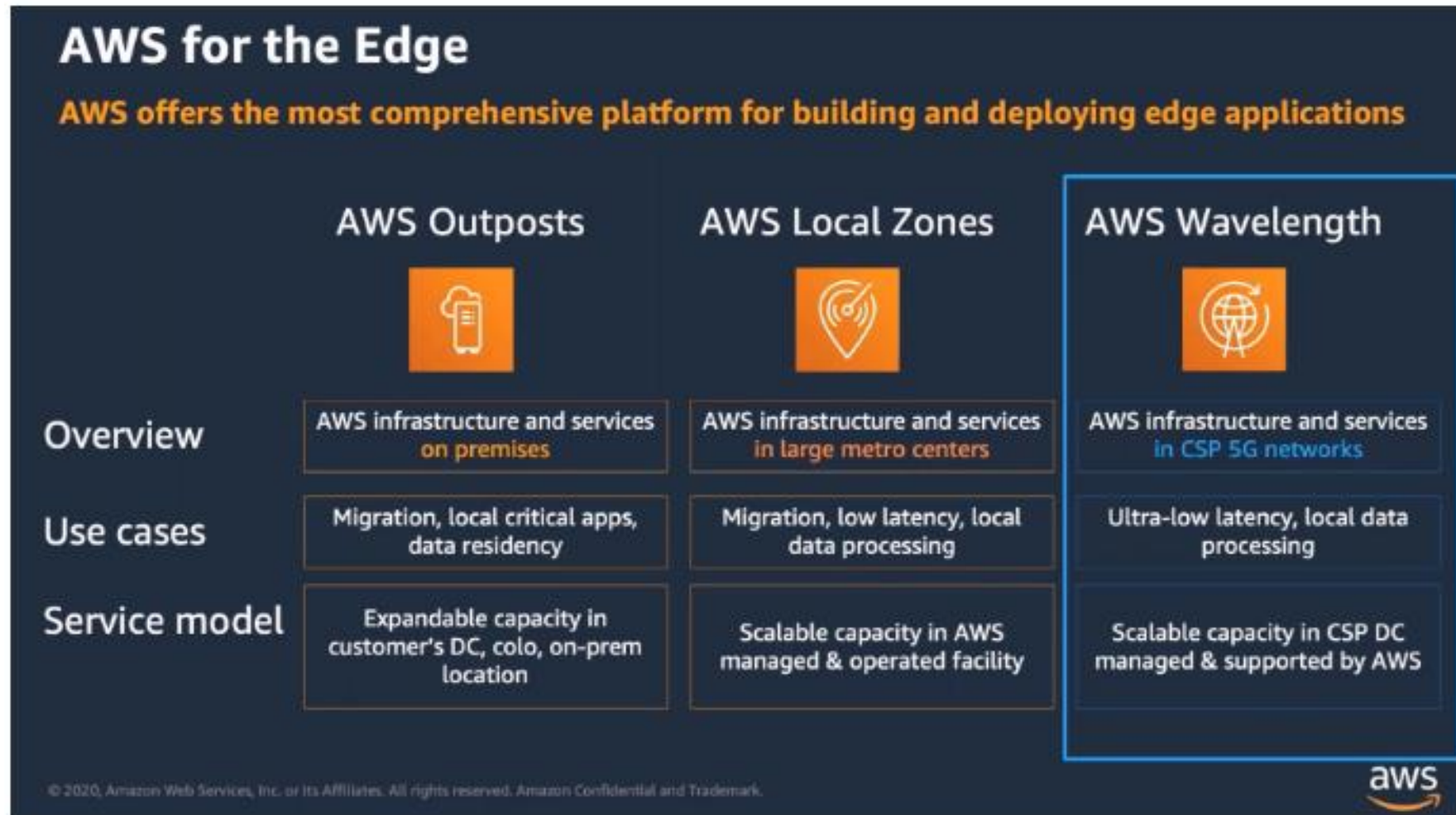
AWS Outposts

- **AWS Outposts** is a fully **managed service** that extends **AWS infrastructure**, services, APIs, and tools to **customer premises**. By providing **local access** to **AWS managed infrastructure**, AWS Outposts enables customers to build and run **applications on premises using the same programming interfaces as in AWS Regions**, while **using local compute and storage resources for lower latency** and local data processing needs.
- **AWS operates, monitors, and manages this capacity** as part of an **AWS Region**. You can create subnets on your Outpost and specify them when you create AWS resources. Instances in Outpost subnets communicate with other instances in the AWS Region using private IP addresses, all within the same VPC.
- The following diagram illustrates the AWS Region us-west-2, two of its Availability Zones, and an Outpost. The VPC spans the Availability Zones and the Outpost. The Outpost is in an on-premises customer data center. Each zone in the VPC has one subnet, and each subnet has an instance.



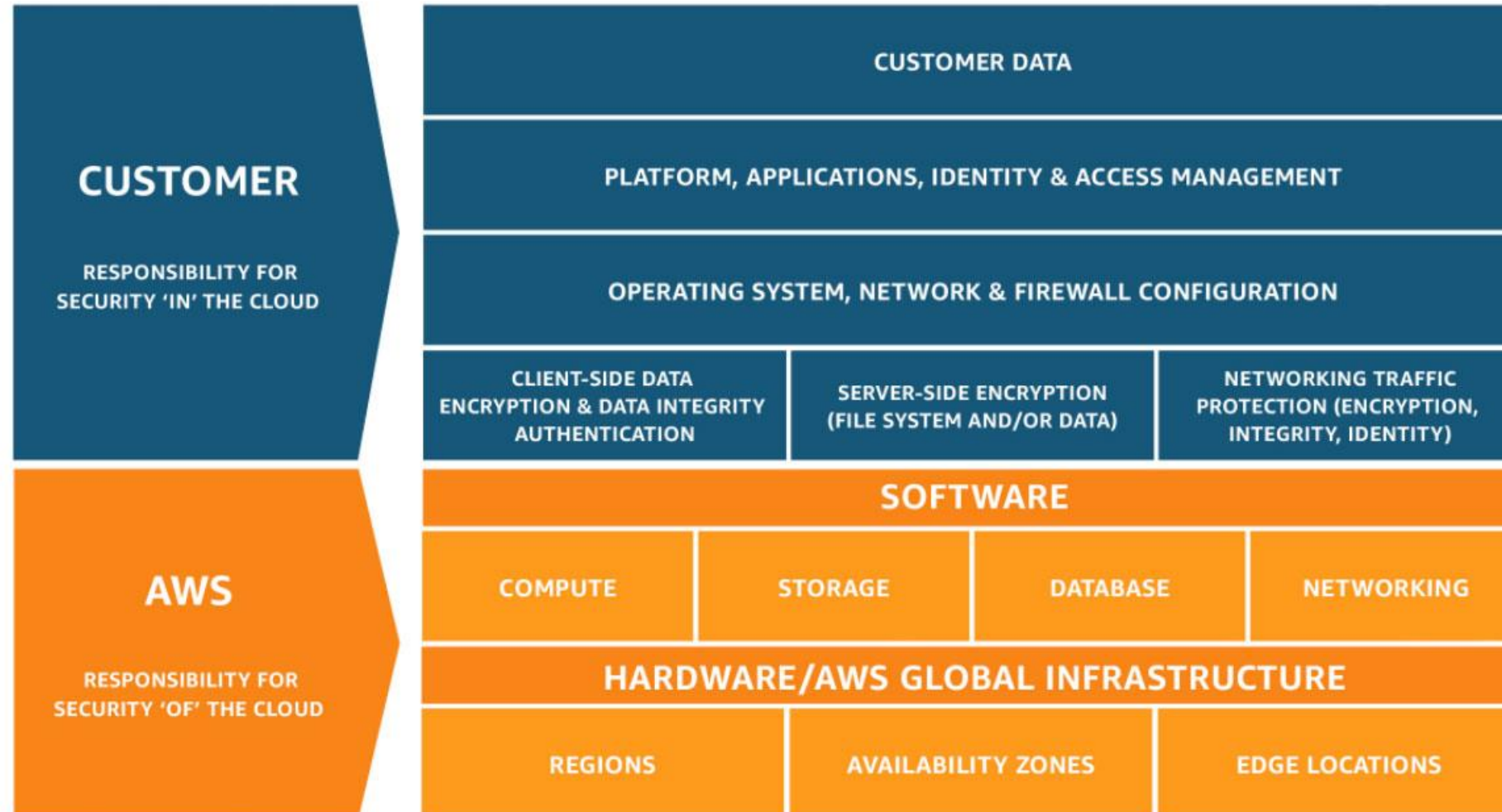
- This enables new classes of compute-intensive, latency sensitive applications. For example, a fleet of autonomous cars interacting with road sensors to prevent crashes, smart industrial robots assessing and reacting to plant conditions in a dangerous manufacturing environment, or retailers serving personalized promotions to shoppers' mobile phones in real time as they pass product displays.

AWS for the Edge



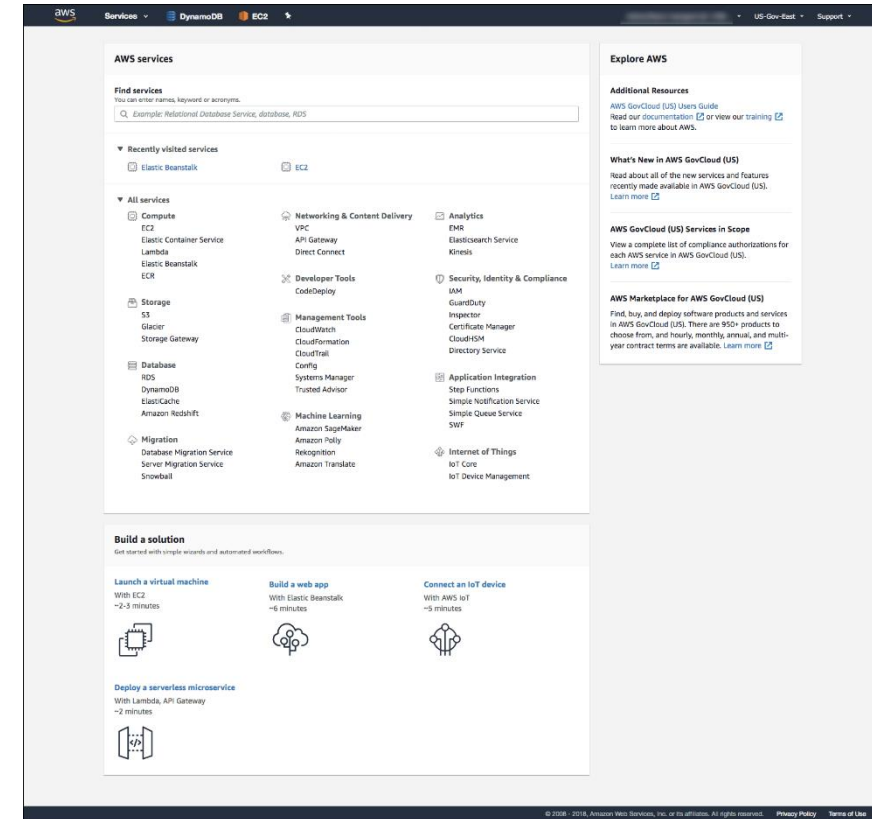
AWS edge computing infrastructure and service platforms. Source: AWS

AWS & Customer



Using AWS - Connecting

- **AWS Management Console**: is a web application for managing AWS Cloud services. The console provides an intuitive user interface for performing many tasks. Each service has its own console, which can be accessed from the AWS Management Console. The console also provides information about the account and billing.
- **AWS Command Line Interface (CLI)** is a unified tool used to manage AWS Cloud services. With just one tool to download and configure, you can control multiple services from the command line and automate them through scripts.
- **The AWS Software Development Kits (SDKs)** provide an application programming interface (API) that interacts with the web services that fundamentally make up the AWS platform. The SDKs provide support for many different programming languages and platforms to allow you to work with your preferred language.



Using AWS - Video

greatlearning
Power Ahead



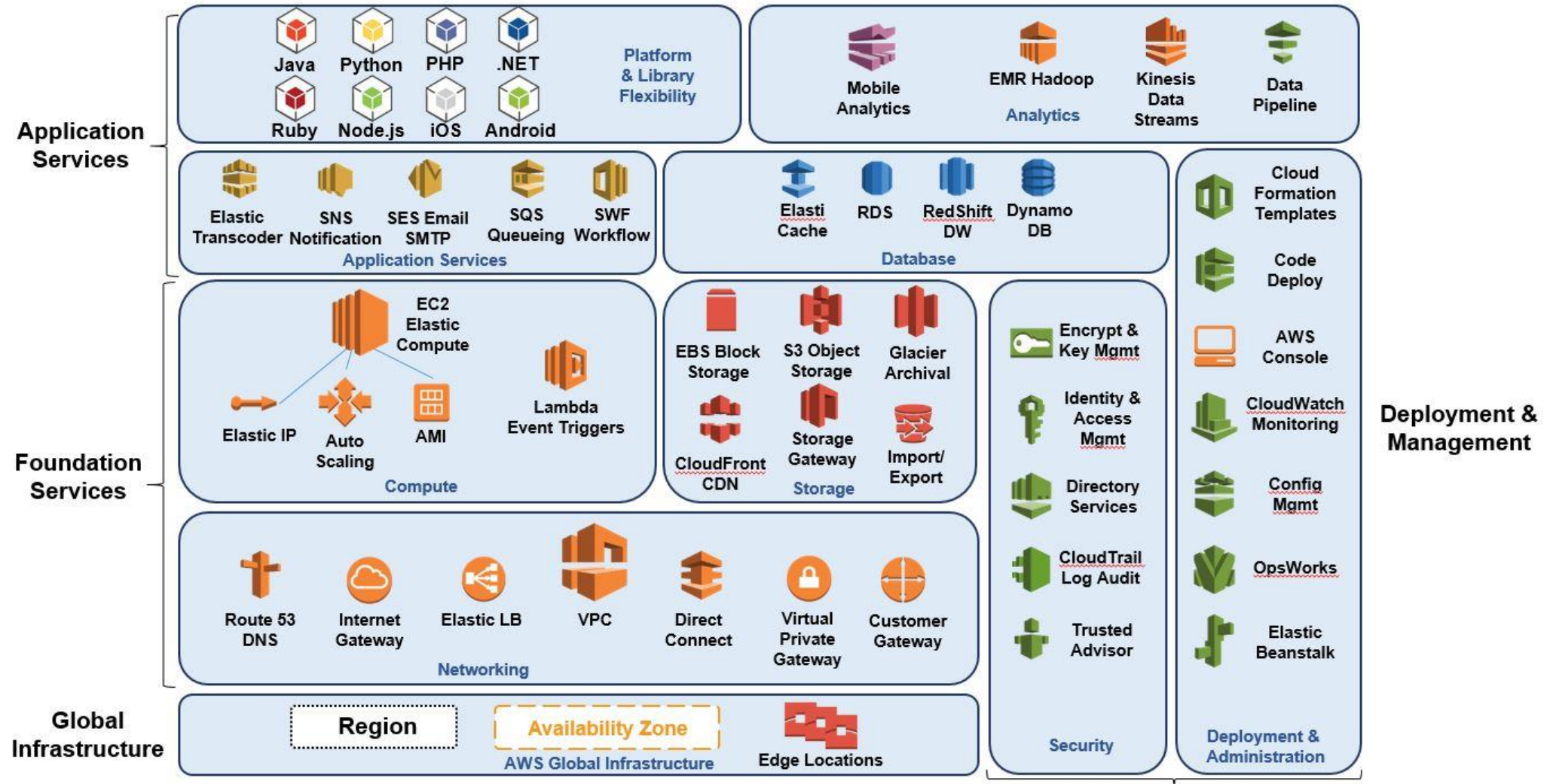
AWS Management Console: is a web application for managing AWS Cloud services.

The console provides an intuitive user interface for performing many tasks.

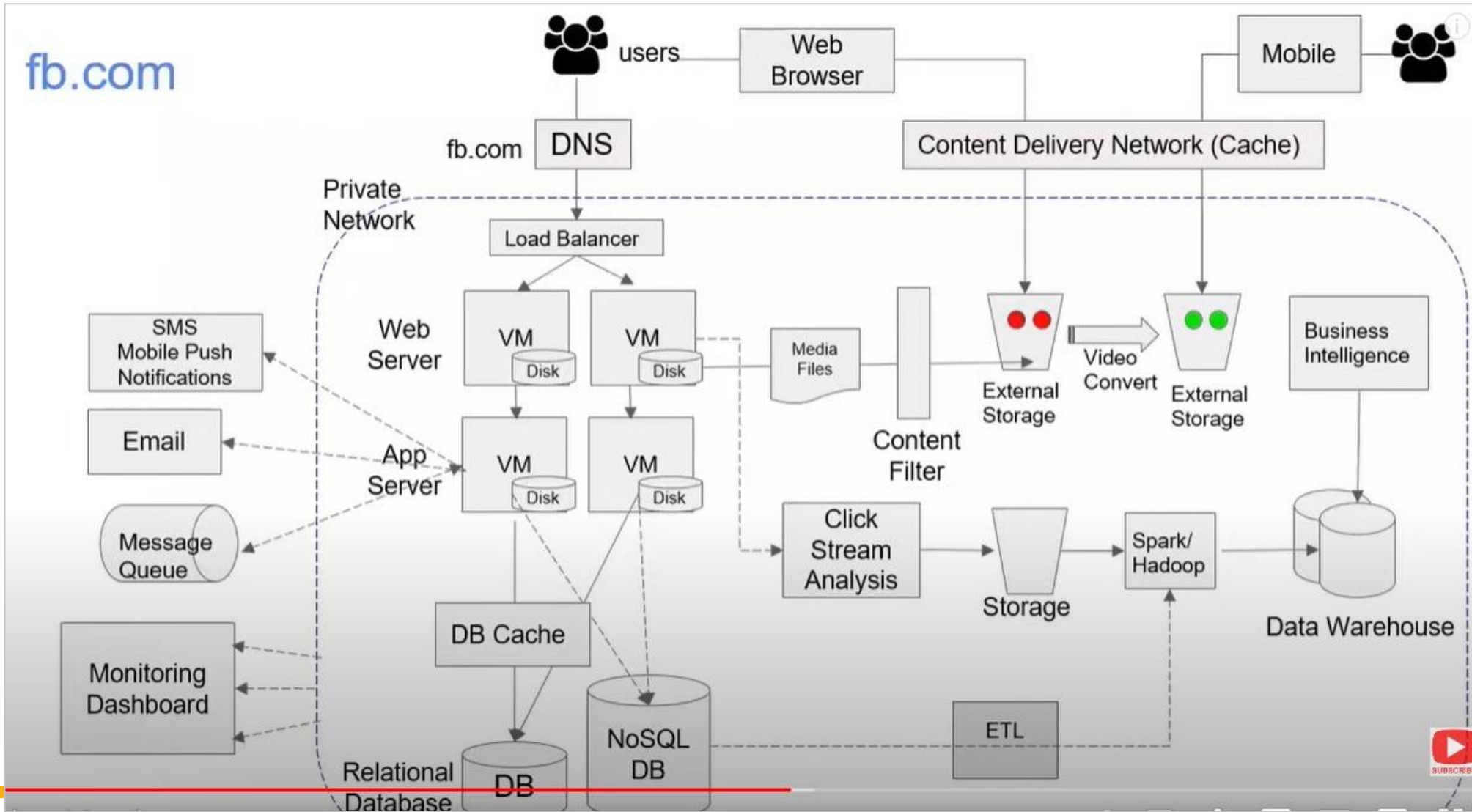
Each service has its own console, which can be accessed from the AWS Management Console.

The console also provides information about the account and billing.

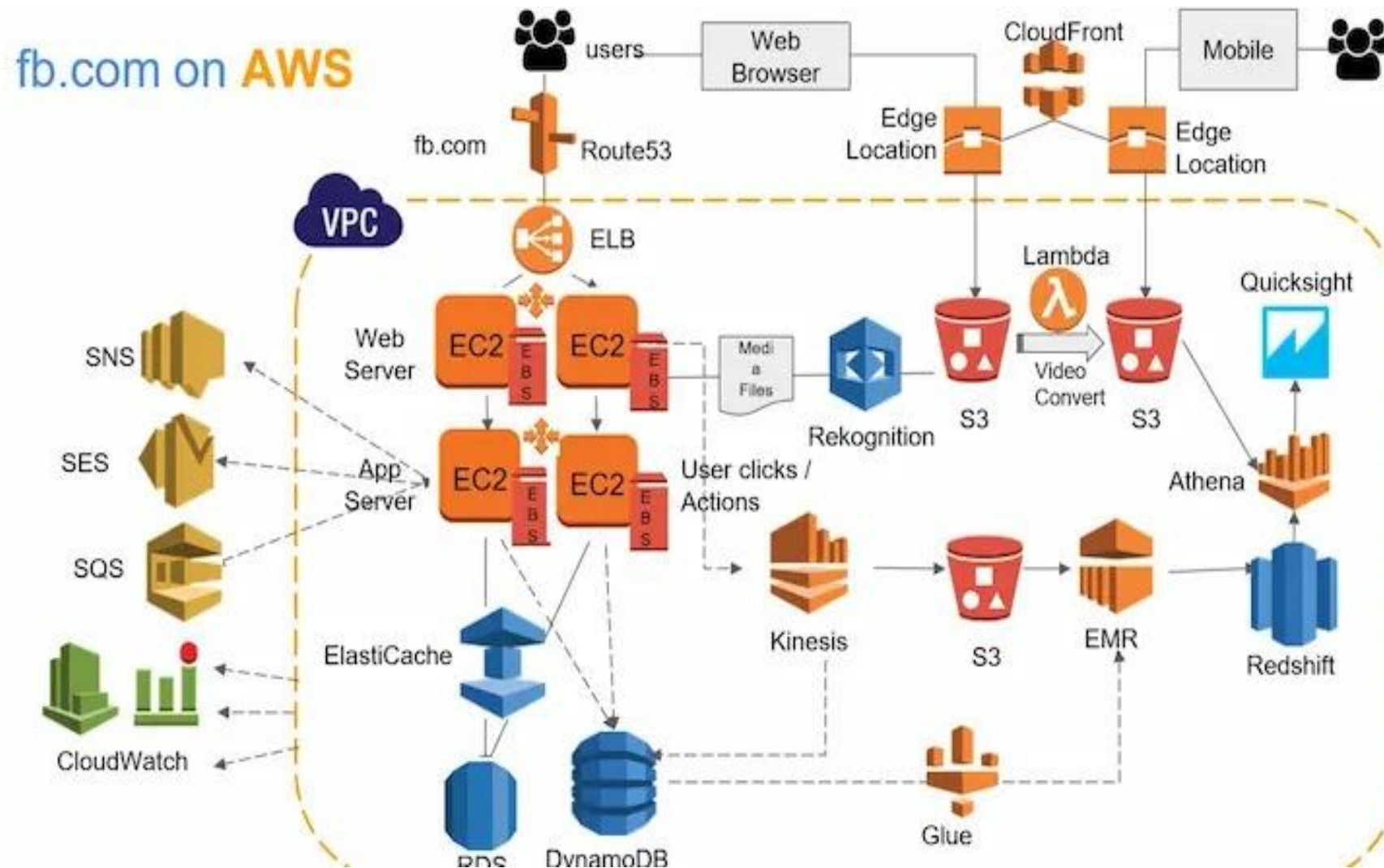
AWS Reference Model



Design of Social Media Application



Design of Social Media Application with AWS Services





BITS Pilani

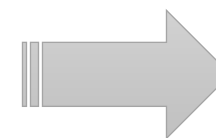
Pilani | Dubai | Goa | Hyderabad



AWS Elastic Compute Cloud EC2



Amazon EC2



- Virtual Compute Cloud
- Root-level System Access
- Elastic Capacity
- Management API
- Scale in Minutes
- Multiple Instance Sizes
- Network Security Model

EC2 Introduction

- Amazon EC2 is **AWS primary web service** that provides **resizable compute capacity** in the **cloud**.
- **Compute** refers to the amount of **computational power required to fulfill your workload**.
- Amazon EC2 allows you to acquire compute through the **launching** of **virtual servers** called **instances**.
- When you launch an **instance**, you can make use of the compute as you wish, just as you would with an on-premises server.
- **Users pay** for the **computing power** of the instance. Charged per hour while the instance is running. When you **stop the instance**, you are **no longer charged**.



Amazon EC2



- Virtual Compute Cloud
- Root-level System Access
- Elastic Capacity
- Management API
- Scale in Minutes
- Multiple Instance Sizes
- Network Security Model

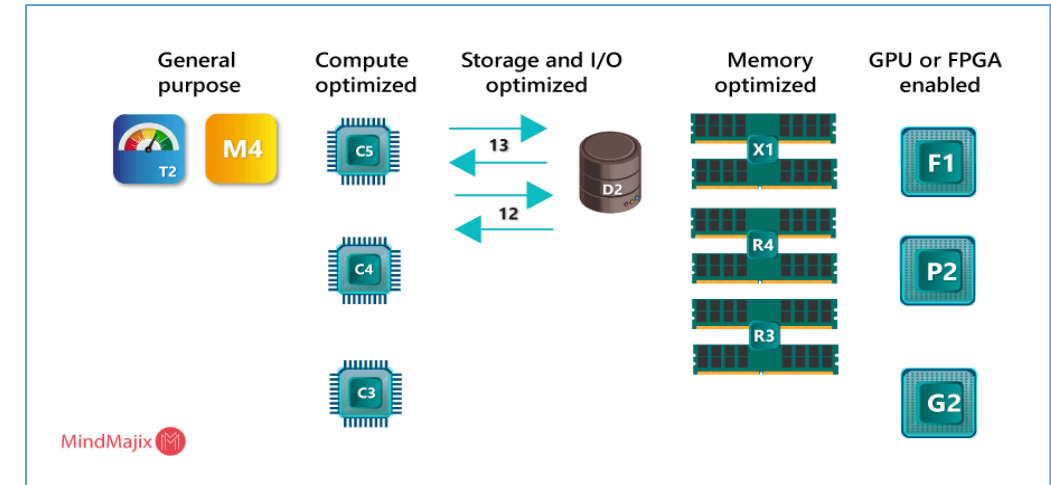
NETFLIX



VOLKSWAGEN
GROUP

EC2 - Instance Type & AMI

- There are two concepts that are key to launching instances on AWS:
- (1) **Instance Type**: The amount of virtual hardware dedicated to the instance and
- (2) **AMI**: The software loaded on the instance.
 - AMI → Amazon Machine Image
- Note
 - Instance Type is similar to the processor*
 - AMI is similar to the OS*



aws **EC2 instance types**

	General Purpose		Compute Optimized	Memory Optimized		Accelerated Computing	Storage Optimized		
Type	t2	m5	c5	r4	x1e	p3	h1	i3	d2
Description	Burstable, good for changing workloads	Balanced, good for consistent workloads	High ratio of compute to memory	Good for in-memory databases	Good for full in-memory applications	Good for graphics processing and other GPU uses	HDD backed, balance of compute and memory	SDD backed, balance of compute and memory	Highest disk ratio
Mnemonic	t is for tiny or turbo	m is for main or happy medium	c is for compute	r is for RAM	x is for xtreme	p is for pictures	h is for HDD	i is for IOPS	d is for dense

ParkMyCloud

EC2 - Instance Type

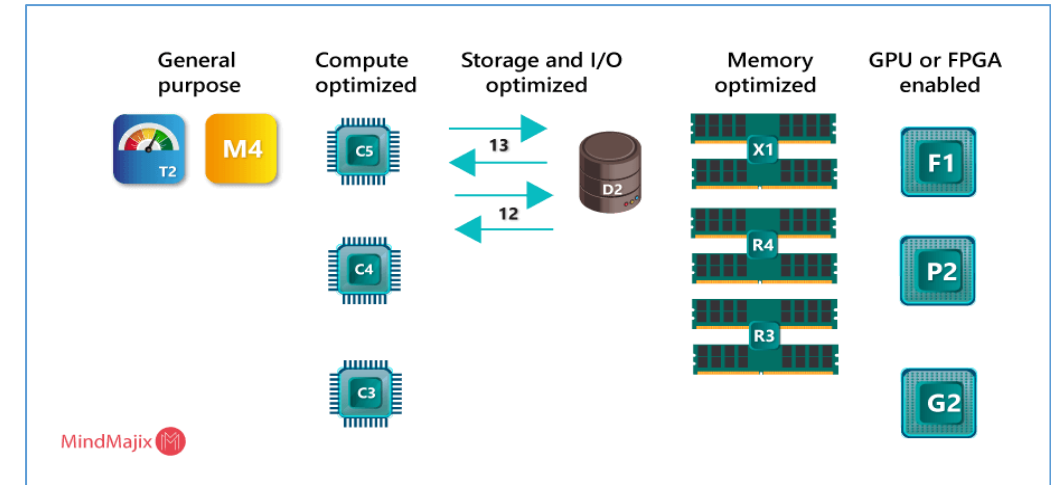
- The **instance type** defines the **virtual hardware** supporting an Amazon EC2 instance.
- There are dozens of instance types available, varying in the following dimensions:
 - Virtual CPUs (vCPUs)
 - Memory
 - Storage (size and type)
- Network performance **Instance types** are **grouped into families** based on the **ratio of these** values to **each other**.
- For instance, the **m4 family** provides a balance of **compute, memory, and network** resources, and it is a good choice for many applications.
- Within each family there are several choices that scale up linearly in size.
- Note that the ratio of vCPUs to memory is constant as the sizes scale linearly.

Instance Family	Instance Type(s)
» General Purpose (M3)	» M3.medium, M3.large, M3.xlarge, M3.2xlarge
» Compute Optimized (C3)	» C3.large, C3.xlarge, C3.2xlarge, C3.4xlarge, C3.8xlarge
» Memory Optimized (R3)	» R3.large, R3.xlarge, R3.2xlarge, R3.4xlarge, R3.8xlarge
» Storage Optimized (I2, HS1)	» I2.xlarge, I2.2xlarge, I2.4xlarge, I2.8xlarge, HS1.8xlarge
» GPU (G2)	» G2.2xlarge
» Micro (T1, M1)	» T1.micro, M1.small

Instance type	EBS only	NVME EBS	Instance store	Placement group	Enhanced networking
M6i	Yes	Yes	No	Yes	ENA EFA
M6id	No	Yes	NVMe	Yes	ENA EFA
M6idn	No	Yes	NVMe	Yes	ENA EFA
M6in	Yes	Yes	No	Yes	ENA EFA
M7a	Yes	Yes	No	Yes	ENA EFA
M7g	Yes	Yes	No	Yes	ENA EFA
M7gd	No	Yes	NVMe	Yes	ENA EFA
M7i	Yes	Yes	No	Yes	ENA EFA
M7i-flex	Yes	Yes	No	Yes	ENA
Mac1	Yes	Yes	No	Yes	ENA
Mac2	Yes	Yes	No	Yes	ENA
T2	Yes	No	No	Yes	Not supported
T3	Yes	Yes	No	Yes	ENA
T3a	Yes	Yes	No	Yes	ENA
T4g	Yes	Yes	No	Yes	ENA

EC2 - Instance Type

- Another variable to consider when choosing an instance type is **network performance**.
- For most instance types, AWS publishes a relative measure of **network performance**: *low*, *moderate*, or *high*.
- Some instance types specify a network performance of **10 Gbps**.
- The network performance increases within a family as the instance type grows.
- For workloads which require low latency, AWS provides enhanced networking support.

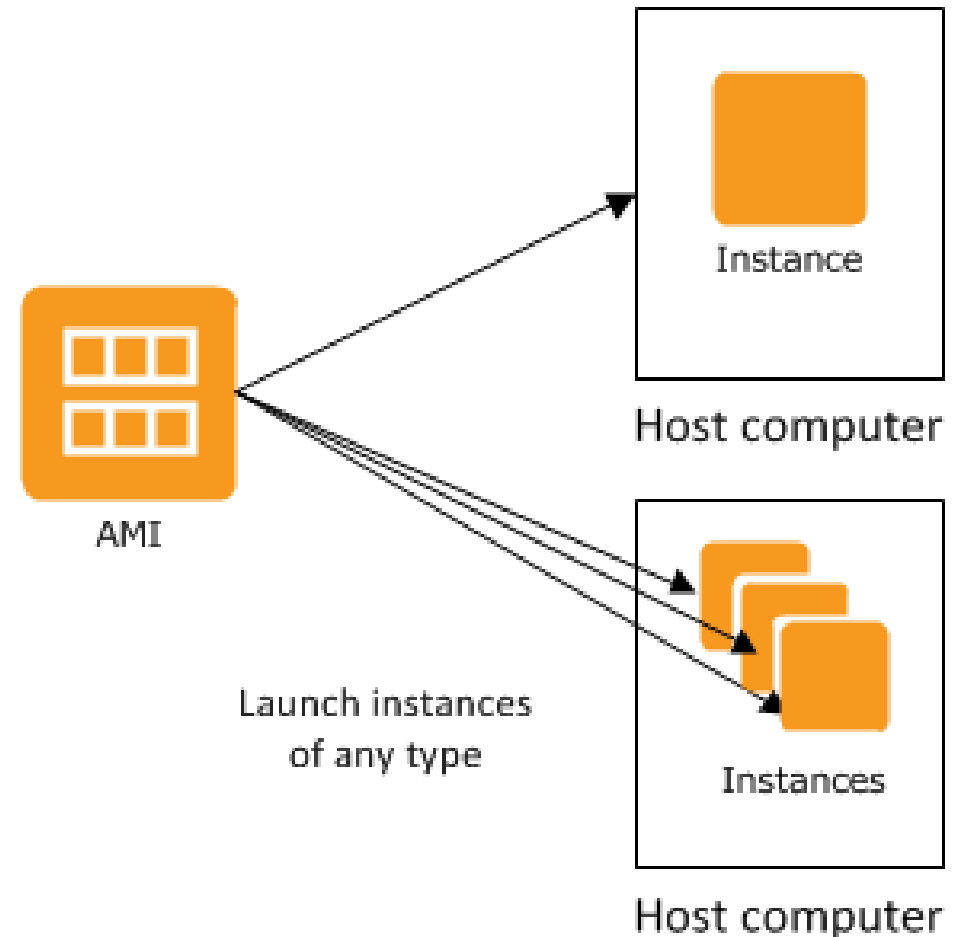


aws EC2 instance types									
	General Purpose		Compute Optimized	Memory Optimized		Accelerated Computing	Storage Optimized		
Type	t2	m5	c5	r4	x1e	p3	h1	i3	d2
Description	Burstable, good for changing workloads	Balanced, good for consistent workloads	High ratio of compute to memory	Good for in-memory databases	Good for full in-memory applications	Good for graphics processing and other GPU uses	HDD backed, balance of compute and memory	SDD backed, balance of compute and memory	Highest disk ratio
Mnemonic	t is for tiny or turbo	m is for main or happy medium	c is for compute	r is for RAM	x is for xtreme	p is for pictures	h is for HDD	i is for IOPS	d is for dense

ParkMyCloud

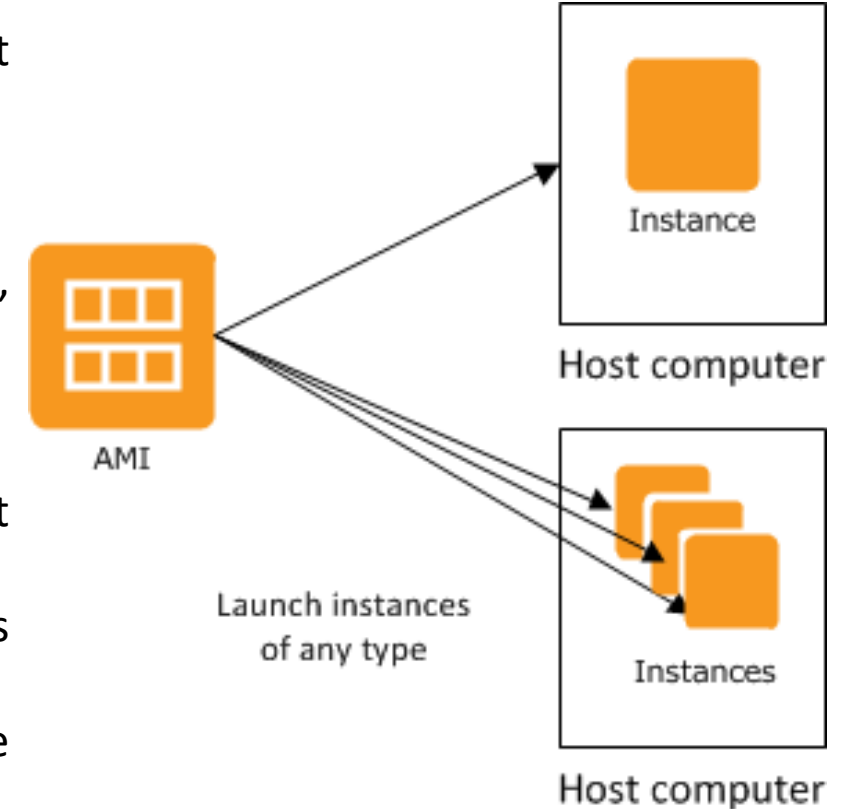
EC2 - Amazon Machine Image (AMI)

- The **Amazon Machine Image (AMI)** defines the **initial software** that will be on an instance when it is launched.
- An **AMI defines** every aspect of the **software state at instance launch**, including:
 - The Operating System (OS) and its configuration
 - The initial state of any patches
 - Application or system software
- All AMIs are based on x86 OSs, either Linux or Windows.



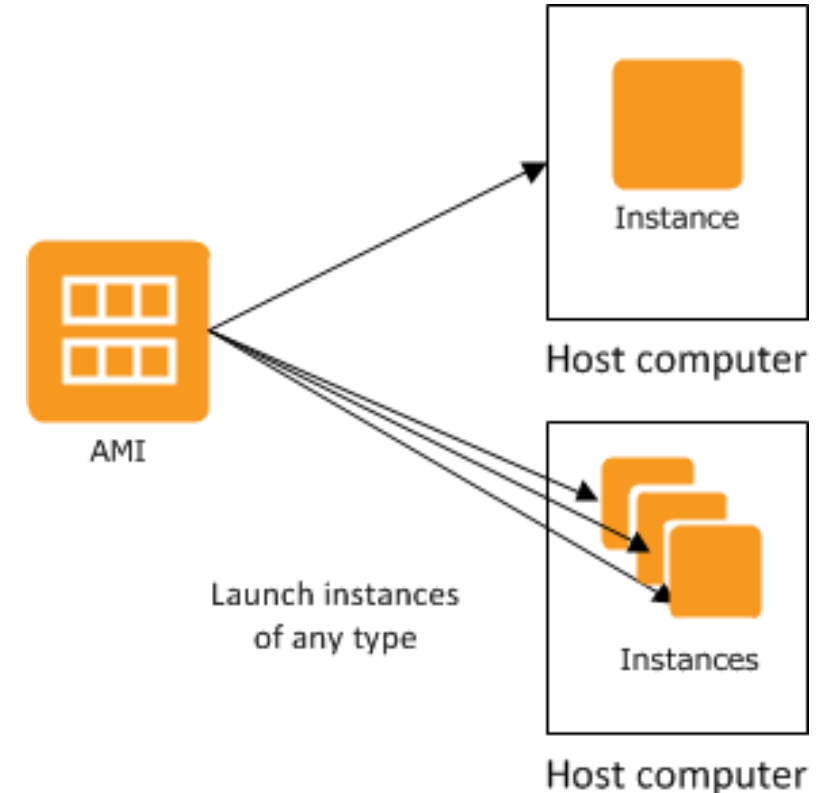
EC2 - AMI Types

- **Published by AWS**— AWS publishes AMIs with versions of many different OSs, both Linux and Windows.
- These include multiple distributions of Linux (including Ubuntu, Red Hat, and Amazon's own distribution) and Windows 2008 and Windows 2012.
- Launching an instance based on one of these AMIs will result in the default OS settings, similar to installing an OS from the standard OS ISO image. As with any OS installation, you should immediately apply all appropriate patches upon launch.



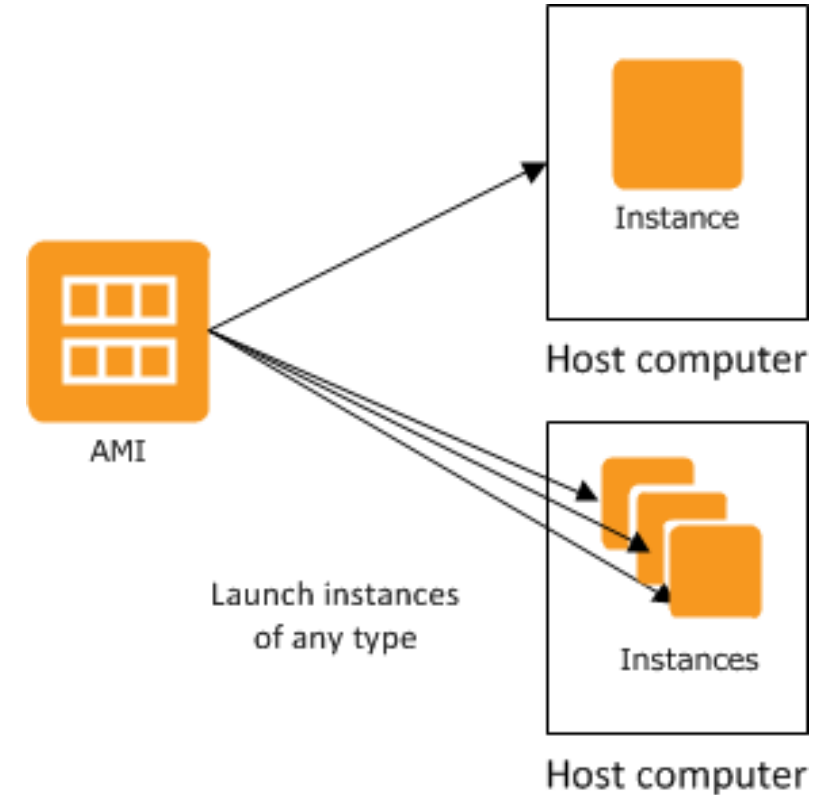
EC2 - AMI Types

- **The AWS Marketplace**— AWS Marketplace is an **online store** that helps customers find, buy, and immediately start using the software and services that run on Amazon EC2.
- Many AWS partners have made their software available in the AWS Marketplace.
- This provides two benefits:
 - The customer **does not need to install** the software, and
 - The **license agreement is appropriate** for the cloud.
- Instances launched from an AWS Marketplace AMI incur the standard hourly cost of the instance type plus an additional per-hour charge for the additional software (some open-source AWS Marketplace packages have no additional software charge).



EC2 - AMI Types

- **Generated from Existing Instances**— An AMI can be created from an **existing Amazon EC2 instance**.
- This is a very **common source** of AMIs. Customers **launch an instance** from a published AMI, and then the **instance is configured to meet all the customer's corporate standards** for updates, management, security, and so on.
- **An AMI is then generated** from the configured instance and used to **generate all instances** of that OS.
- In this way, all new instances **follow the corporate standard** and it is more difficult for individual projects to launch non-conforming instances.



EC2 - AMI Types

- **Uploaded Virtual Servers**— Using AWS VM Import/ Export service, customers can create images from various virtualization formats, including raw, VHD, VMDK, and OVA. The current list of supported OSs (Linux and Windows) can be found in the AWS documentation. It is incumbent on the customers to remain compliant.
- **VMDK** → Virtual Machine Disk : is a file format that describes containers for virtual hard disk drives to be used in virtual machines like VMware Workstation or VirtualBox.
- **VHD** → Virtual Hard Disk: is a file format which represents a **virtual hard disk** drive (HDD). It may contain what is found on a physical HDD, such as disk partitions and a file system, which in turn can contain files and folders. It is typically used as the hard disk of a virtual machine.
- **OVA** → Open Virtual Appliance/Application Format: is merely a single **file** distribution of the same **file** package, stored in the TAR format



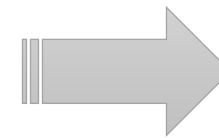
BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Connecting to EC2 Instance



Amazon EC2



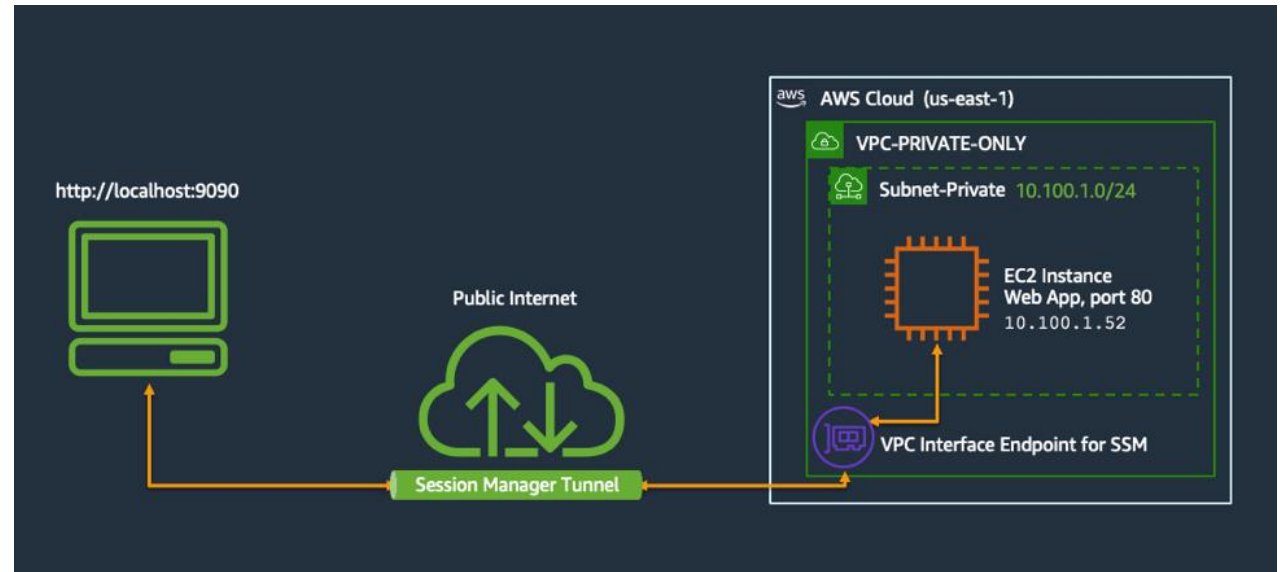
- Virtual Compute Cloud
- Root-level System Access
- Elastic Capacity
- Management API
- Scale in Minutes
- Multiple Instance Sizes
- Network Security Model

Creating an EC2 Instance - Video

- Amazon **EC2** uses **public-key cryptography** to encrypt and decrypt login information.
- **Public-key cryptography** uses a **public key** to **encrypt** a piece of data and an **associated private key** to decrypt the data.
- These two keys together are called a **key pair**.
- **Key pairs** can be created through the **AWS Management Console**, CLI, or API, or customers can upload their own key pairs.
- **AWS stores the public key**, and the **private key** is kept by the **customer**.
- The **private key** is essential to acquiring **secure access** to an **instance** for the **first time**.

EC2 – Accessing over Web

- There are several ways that an instance may be addressed over the web upon creation:
- **Public Domain Name System (DNS) Name**— When you launch an instance, AWS creates a DNS name that can be used to access the instance. This DNS name is generated automatically and cannot be specified by the customer.
- **Public IP**— A launched instance may also have a public IP address assigned. This IP address is assigned from the addresses reserved by AWS and cannot be specified. This IP address is unique on the Internet, persists only while the instance is running, and cannot be transferred to another instance.
- **Elastic IP**— An elastic IP address is a static address unique on the Internet that you reserve independently and associate with an Amazon EC2 instance.

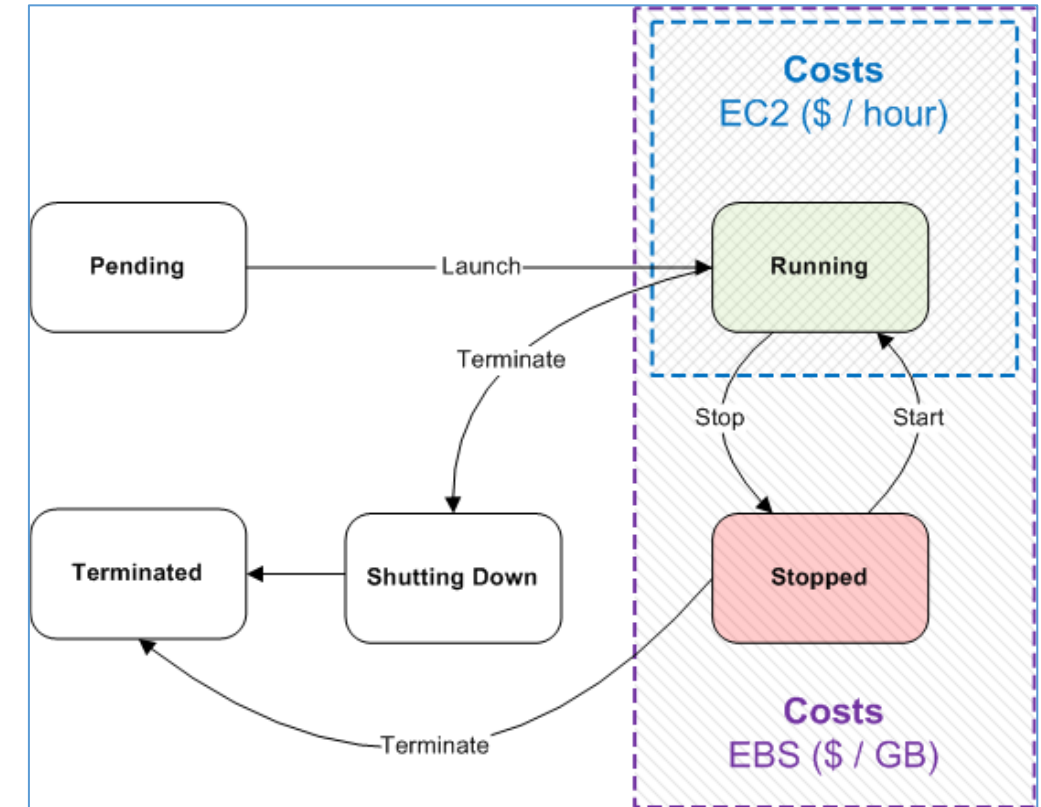


EC2 – Lifecycle

- Amazon EC2 has several features and services that facilitate the management of Amazon EC2 instances over their entire lifecycle.
- Launching
- Bootstrapping: The process of providing code to be run on an instance at launch is called bootstrapping.

Managing Instances

- When the number of instances in your account starts to climb, it can become difficult to keep track of them.
- Tags can help you manage not just your Amazon EC2 instances, but also many of your AWS Cloud services.
- Tags are key/ value pairs you can associate with your instance or other service.
- Tags can be used to identify attributes of an instance like project, environment (dev, test, and so on), billable department, and so forth.
- You can apply up to 10 tags per instance.



Monitoring Instances

AWS offers a service called Amazon CloudWatch that provides monitoring and alerting for Amazon EC2 instances, and other AWS infrastructure.

EC2 – Tenancy Options

- There are several tenancy options for Amazon EC2 instances that can help customers achieve security and compliance goals.
- **Shared Tenancy** Shared tenancy is the default tenancy model for all Amazon EC2 instances, regardless of instance type, pricing model, and so forth. **Shared tenancy means that a single host machine may house instances from different customers.** As AWS does not use overprovisioning and fully isolates instances from other instances on the same host, this is a secure tenancy model.
- **Dedicated Instances** Dedicated Instances run on hardware that's dedicated to a single customer. As a customer runs more Dedicated Instances, more underlying hardware may be dedicated to their account. Other instances in the account (those not designated as dedicated) will run on shared tenancy and will be isolated at the hardware level from the Dedicated
- **Dedicated Host** An Amazon EC2 Dedicated Host is a physical server with Amazon EC2 instance capacity fully dedicated to a single customer's use. Dedicated Hosts can help you address licensing requirements and reduce costs by allowing you to use your existing server-bound software licenses.

EC2 – Placement Groups

- **Placement Groups** A placement group is a **logical grouping** of instances within a **single Availability Zone**. Placement groups enable **applications to participate** in a low-latency, **10 Gbps network**. To fully use this network performance for your placement group, choose an **instance type** that supports **enhanced networking** and 10 Gbps network performance.
- **Instance Stores** An instance store (sometimes referred to as **ephemeral storage**) provides **temporary block-level storage** for your instance. This storage is located on disks that are physically attached to the host computer.
- The size and type of **instance stores** available with an Amazon EC2 instance depend on the **instance type**. Can range from **no instance store** to **24 TB** instance store
- **Instance stores** are included in the cost of an **Amazon EC2 instance**, so they are a very cost-effective solution for appropriate workloads. The key aspect of instance stores is that they are temporary.
- Data in the instance store is lost when:
 - The underlying disk drive fails.
 - The instance stops (the data will persist if an instance reboots).
 - The instance terminates.

innovate

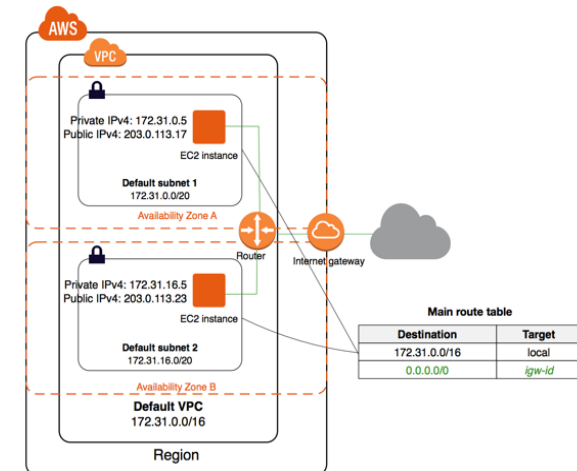
achieve

lead

BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Networking in AWS - VPC



AWS VPC



- A **virtual private cloud** (VPC) is a virtual network **dedicated to your AWS account**. It is **logically isolated** from other virtual networks in the AWS cloud.
- You can **launch your AWS resources**, such as Amazon EC2 instances, into **your VPC**. You can provision your own **logically isolated section of AWS**, similar to designing and implementing a separate **independent network** that would operate in an on-premises data center.
- You can configure your VPC; you can select its **IP address range**, **create subnets**, and **configure route tables**, network **gateways**, and **security settings**. A **subnet** is a **range of IP addresses** in your VPC. You can **launch AWS resources into a subnet that you select**.
- Use a **public subnet** for **resources that must be connected to the Internet**, and a **private subnet** for resources that **won't be connected to the Internet**. Within a **region**, you can create **multiple Amazon VPCs**, and each **Amazon VPC** is **logically isolated** even if it shares its IP address space..

uses

- ❖ Build virtual networks on the cloud
- ❖ No need for any VPN, hardware or physical DC
- ❖ Define bespoke network space like:
 - ❖ VPC with a single public subnet only
 - ❖ VPC with public and private subnets
 - ❖ VPC with public and private subnets and AWS Site-to-Site VPN access
 - ❖ VPC with a private subnet only and AWS Site-to-Site VPN access

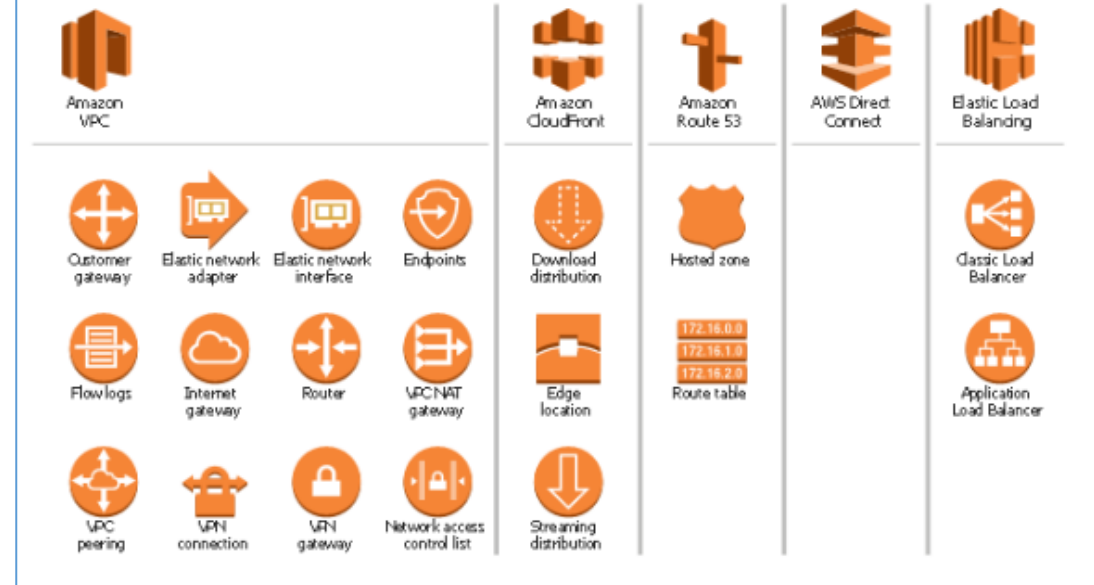
AWS VPC - Components



- An Amazon VPC consists of the following components:
 - Subnets
 - Route tables
 - Dynamic Host Configuration Protocol (DHCP) option sets
 - Security groups
 - Network Access Control Lists (ACLs)
- An Amazon VPC has the following optional components:
 - Internet Gateways (IGWs)
 - Elastic IP (EIP) addresses
 - Elastic Network Interfaces (ENIs)
 - Endpoints
 - Peering
 - Network Address Translation (NATs) instances and
 - NAT gateways Virtual Private Gateway (VPG), Customer
 - Gateways (CGWs), and Virtual Private Networks (VPNs)



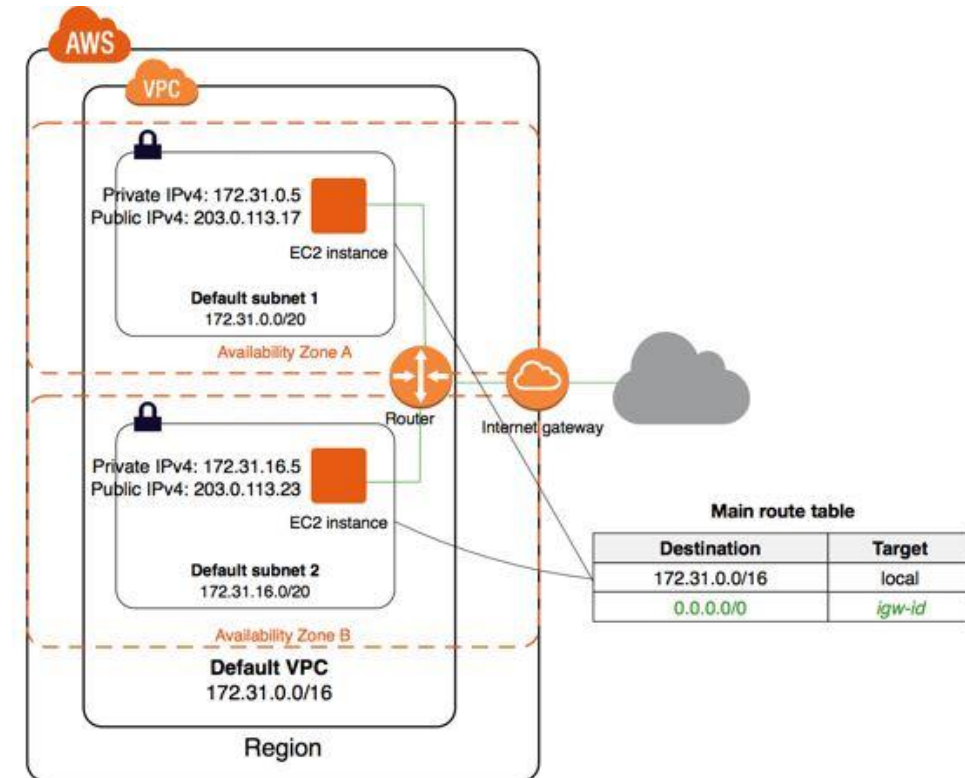
Networking & Content Delivery



AWS VPC - Functioning



- You control how the instances that you launch into a VPC access resources outside the VPC.
- Your default VPC includes an Internet gateway, and each default subnet is a public subnet.
- Each instance that you launch into a default subnet has a private IPv4 address and a public IPv4 address.
- These instances can communicate with the Internet through the Internet gateway.
- By default, each instance that you launch into a non-default subnet has a private IPv4 address, but no public IPv4 address, unless you specifically assign one at launch, or you modify the subnet's public IP address attribute.
- These instances can communicate with each other, but can't access the Internet.





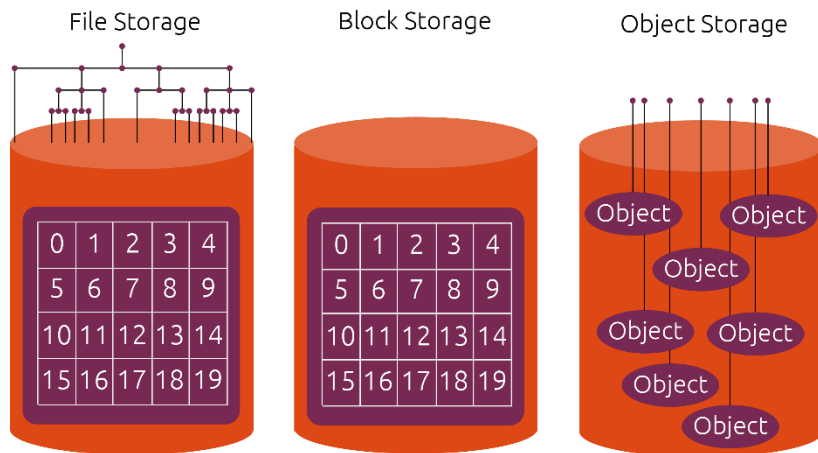
AWS Storage









AWS Cloud Storage Services



Storage Types

- **Block storage** : Operates at a lower level— the raw storage device level— and manages data as a set of numbered, fixed-size blocks.
- **File storage** : Operates at a higher level— the operating system level— and manages data as a named hierarchy of files and folders.
- *Block and file storage are often accessed over a network in the form of a Storage Area Network (SAN) for block storage, using protocols such as iSCSI or Fiber Channel, or as a Network Attached Storage (NAS) file server or “filer” for file storage.*



	Amazon Simple Storage Service (Amazon S3)	A service that provides scalable and highly durable object storage in the cloud.
	Amazon Glacier	A service that provides low-cost highly durable archive storage in the cloud.
	Amazon Elastic File System (Amazon EFS)	A service that provides scalable network file storage for Amazon EC2 instances.
	Amazon Elastic Block Store (Amazon EBS)	A service that provides block storage volumes for Amazon EC2 instances.
	Amazon EC2 Instance Storage	Temporary block storage volumes for Amazon EC2 instances.
	AWS Storage Gateway	An on-premises storage appliance that integrates with cloud storage.
	AWS Snowball	A service that transports large amounts of data to and from the cloud.
	Amazon CloudFront	A service that provides a global content delivery network (CDN).



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

AWS Simple Storage Service S3

AWS Cloud Storage Services



AWS S3



- ❖ Amazon S3 is easy-to-use object storage with a simple web service interface that you can use to store and retrieve any amount of data from anywhere on the web.
- ❖ Amazon S3 also allows you to pay only for the storage you actually use, which eliminates the capacity planning and capacity constraints associated with traditional storage.
- ❖ Amazon S3 can be used alone or in conjunction with other AWS services, and it offers a very high level of integration with many other AWS cloud services.

uses

- ❖ Backup and archive for on-premises or cloud data
- ❖ Content, media, and software storage and distribution
- ❖ Big data analytics
- ❖ Static website hosting
- ❖ Cloud-native mobile and Internet application hosting
- ❖ Disaster recovery



AWS S3



- ✓ Amazon S3 is cloud object storage. Instead of being closely associated with a server, Amazon S3 storage is independent of a server and is **accessed** over the **Internet**.
- ✓ Instead of managing data as blocks or files using SCSI, CIFS, or NFS protocols, data is managed as objects using an Application Program Interface **(API) built on standard HTTP verbs**. Each Amazon **S3 object** contains both **data and metadata**.
- ✓ Objects reside in containers called buckets, and each object is identified by a unique user-specified key (filename).
- ✓ Buckets are a simple flat folder with no file system hierarchy.
- ✓ That is, you can have multiple buckets, but you can't have a sub-bucket within a bucket. Each **bucket** can hold an unlimited number of objects.

- ❖ However, keep in mind that **Amazon S3** is not a **traditional file system** and differs in significant ways.
- ❖ In **Amazon S3**, you **GET** an object or **PUT** an **object**, operating on the **whole object** at once, instead of **incrementally updating** portions of the object as you would with a file.
- ❖ Instead of a **file system**, **Amazon S3** is **highly-durable** and **highly-scalable object storage** that is optimized for reads and is built with an intentionally minimalistic feature set.
- ❖ It provides a **simple and robust abstraction** for file storage that frees you from many underlying details that you normally do have to deal with in traditional storage.
- ❖ Amazon S3 objects are automatically replicated on multiple devices in multiple facilities within a region.



AWS S3

Objects:

Opaque data to be stored (1 byte ... 5 Gigabytes)

Authentication and access controls

Buckets:

Object container – any number of objects

100 buckets per account

Keys:

Unique object identifier within bucket

Up to 1024 bytes long

Flat object storage model

Standards-Based Interfaces:

REST and SOAP

URL-Addressability – every object has a URL

Service:

ListAllMyBuckets

Buckets:

CreateBucket

DeleteBucket

ListBucket GetBucketAccessControlPolicy

SetBucketAccessControlPolicy

GetBucketLoggingStatus

SetBucketLoggingStatus

Objects:

PutObject

PutObjectInline

GetObject GetObjectExtended

DeleteObject

GetObjectAccessControlPolicy

SetObjectAccessControlPolicy





BITS Pilani

Pilani | Dubai | Goa | Hyderabad



AWS Elastic Block Storage EBS

AWS Cloud Storage Services



AWS EBS



- ❖ Amazon EBS provides persistent block-level storage volumes for use with Amazon EC2 instances.
- ❖ Each Amazon EBS volume is automatically replicated within its Availability Zone to protect you from component failure, offering high availability and durability.
- ❖ Amazon EBS volumes are available in a variety of types that differ in performance characteristics and price.
- ❖ Multiple Amazon EBS volumes can be attached to a single Amazon EC2 instance, although a volume can only be attached to a single instance at a time.

uses

- ❖ Boot Volumes
- ❖ SQL & NoSQL Database
- ❖ Big Data workloads
- ❖ Data Warehouses
- ❖ Logging & Telemetry
- ❖ Transaction Processing

AWS EBS Types

General-Purpose SSD General-purpose SSD volumes offer cost-effective storage that is ideal for a broad range of workloads. They deliver strong performance at a moderate price point that is suitable for a wide range of workloads.

A general-purpose SSD volume can range in size from 1 GB to 16 TB and provides a baseline performance of three IOPS per gigabyte provisioned, capping at 10,000 IOPS.

They are suited for a wide range of workloads where the very highest disk performance is not critical, such as:

- System boot volumes
- Small- to medium-sized databases
- Development and test environments



AWS EBS Types

Provisioned IOPS SSD Provisioned IOPS SSD volumes are designed to meet the needs of I/ O-intensive workloads, particularly database workloads that are sensitive to storage performance and consistency in random access I/ O throughput. While they are the most expensive Amazon EBS volume type per gigabyte, they provide the highest performance of any Amazon EBS volume type in a predictable manner.

A Provisioned IOPS SSD volume can range in size from 4 GB to 16 TB. Provisioned IOPS SSD volumes provide predictable, high performance and are well suited for:

- Critical business applications that require sustained IOPS performance
- Large database workloads





AWS Elastic File Storage EFS

AWS Cloud Storage Services



AWS EFS



- ❖ EFS(Elastic file system) is a file-level storage service that basically provides a shared elastic file system with virtually unlimited scalability support
- ❖ EFS is highly available storage that can be utilized by many servers at the same time. AWS EFS is a fully managed service by amazon and it offers scalability on the fly.
- ❖ This means that the user need not worry about their increasing or decreasing workload. If the workload suddenly becomes higher then the storage will automatically scale itself and if the workload decreases then the storage will itself scale down.
- ❖ This scalability feature of EFS also provides cost benefits as you need not pay anything for the part of storage that you don't use, you only pay for what you use(Utility-based computing).

uses

- ❖ Lift-and-shift application support
- ❖ Analytics for big data
- ❖ Web server support
- ❖ Application development and testing



BBC Uses Shared File Storage to Migrate Red Button Application to the AWS Cloud



Faculty Uses Amazon EFS to Scale Innovative Machine-Learning Platform



T-Mobile Improves Customer Experience, Lowers Costs of Kubernetes Storage Using AWS



Caltech Uses Amazon EFS to Automate Academic Computing File Management



Swisstopo uses EFS to deliver fast map content to millions of users



How one Caribbean university digitally transformed and saved money by migrating to the cloud



Alpha Vertex uses EFS for financial modeling with machine learning



Atlassian uses EFS with Jira Data Center for scalable shared storage



LoanLogics kickstarts its all-in move to the cloud with AWS storage



Be Software simplifies storage and cuts costs using AWS Cloud and EFS



Butterfly Network brings medical imaging to more people using Amazon EFS



Celgene chooses Amazon EFS to store file data for all of its R&D workloads

AWS EFS vs AWS EBS vs AWS S3

Category	S3	EBS	EFS
Storage Type	Object Storage	Block Storage	File Storage
Pricing	Pay as you Use	Pay for provisioned capacity	Pay as you Use
Storage Size	Unlimited Storage	Limited storage	Unlimited Storage
Scalability	Unlimited Scalability	Increase/decrease size manually	Unlimited Scalability
Durability	Stored redundantly across multiple Azs	Stored redundantly in a Single AZ	Stored redundantly across multiple Azs
Availability	Max is 99.99% with S3 Standard	99.99%	No SLAs
Security	Supports Data at Rest and Data in Transit encryption	Supports Data at Rest and Data in Transit encryption	Supports Data at Rest and Data in Transit encryption
Back up and Restore	Use Versioning or cross-region replication	Automated Backups and Snapshots	EFS to EFS replication
Performance	Slower than EBS and EFS	Faster than S3 and EFS	Faster than S3, Slower than EBS
Accessibility	Publicly and Privately accessible	Accessible only via the attached EC2 instance	Accessible simulatenously from multiple EC2 and on-premises instance
Interface	Web Interface	File System Interface	Web and File System Interface
Use cases	Media, Entertainment, Big data analytics, backups and archives, web serving and content management	Boot volumes, transactional and NoSQL databases, data warehousing ETL	Media, Entertainment, Big data analytics, backups and archives, web serving and content management, home directories



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

AWS Glacier

AWS Cloud Storage Services



Amazon Glacier



- Amazon Glacier is an extremely low-cost storage service that provides durable, secure, and flexible storage for data archiving and online backup.
- To keep costs low, Amazon Glacier is designed for infrequently accessed data where a retrieval time of three to five hours is acceptable.
- Amazon Glacier can **store an unlimited amount of virtually any kind of data**, in any format.
- In most cases, the data stored in Amazon Glacier consists of large **TAR** (Tape Archive) or ZIP files.
- In **Amazon Glacier**, data is stored in **archives**. An **archive** can contain up to **40TB** of data, and you can have an **unlimited number of archives**.
- Each **archive** is assigned a unique archive ID at the time of creation.
- (Unlike an Amazon S3 object key, you cannot specify a user-friendly archive name.) All archives are automatically **encrypted**, and **archives are immutable**— after an archive is created, it cannot be modified.

uses

- ❖ Digital Storage.
- ❖ Scientific Data Storage.
- ❖ Healthcare information Archiving.
- ❖ Regulatory and Compliance Archiving.
- ❖ Magnetic Tape Replacement.



Rock & Roll Hall of Fame preserves rock music history and modernizes on AWS.

[Read the case study »](#)



Qube Cinema cuts costs by 80% with archival on Amazon S3 Glacier.

[Read the case study »](#)



Reuters builds easily accessible large-scale news archives on Amazon S3 Glacier.

[Read the blog »](#)



BandLab decreases costs and improves availability using Amazon S3 Glacier.

[Read the case study »](#)



Joyn readies exclusive content for audiences with Amazon S3 Intelligent-Tiering and Amazon S3 Glacier.

[Read the blog »](#)



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

AWS Case Study



NETFLIX



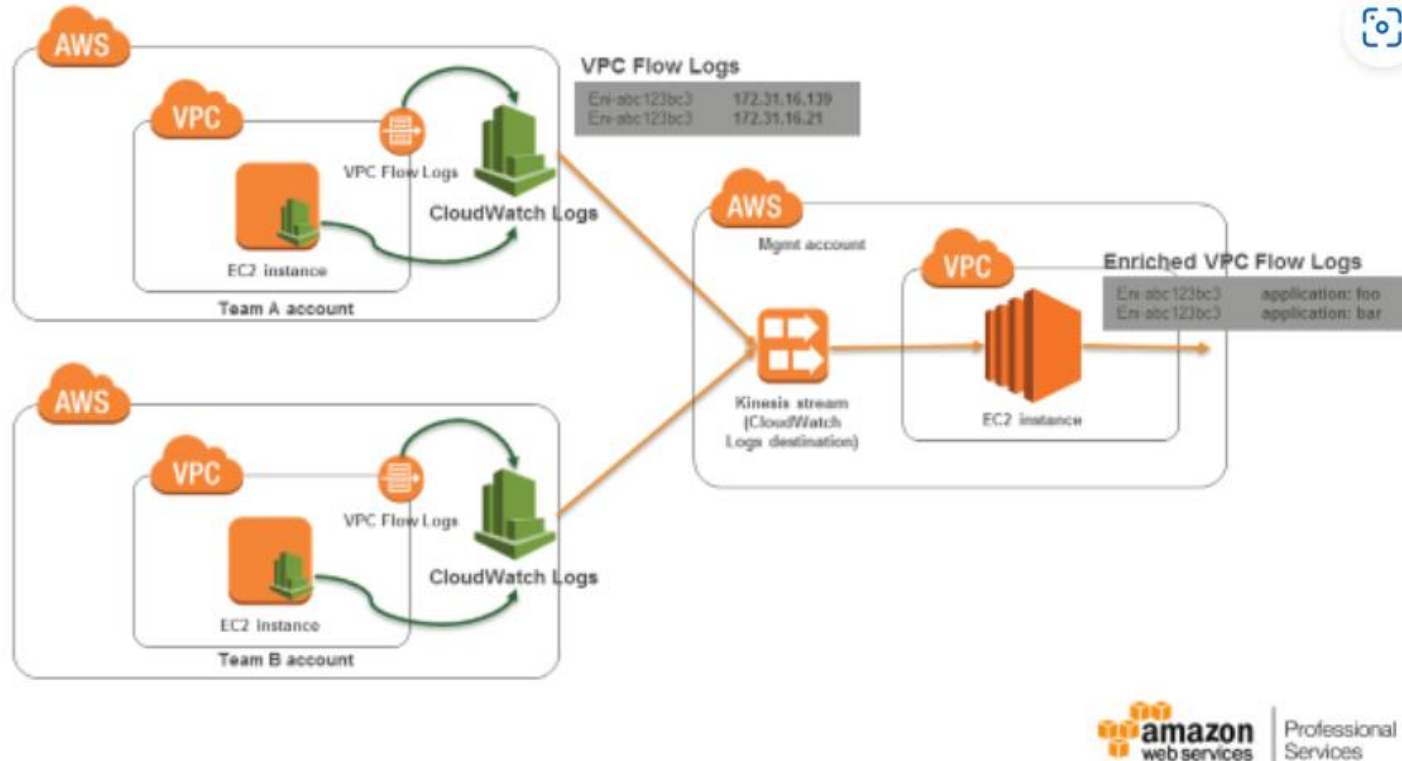
- Online content provider Netflix can support seamless global service by using Amazon Web Services (AWS). AWS enables Netflix to quickly deploy thousands of servers and terabytes of storage within minutes. Users can stream Netflix shows and movies from anywhere in the world, including on the web, on tablets, or on mobile devices such as iPhones.

[AWS re:Invent 2022 - Capacity plan optimally in the cloud \(NFX304\) - YouTube](#)

Amazon Web Services (AWS) offers hundreds of different types of Amazon Elastic Cloud Compute (EC2) options, and Netflix has a multitude of different stateful and stateless workloads that need different combinations of CPU, RAM, disk, and network.

In this 2022 AWS re:Invent session, learn how Netflix uses a service capacity modeling system to optimally consume Amazon EC2 to run a variety of uncertain workloads ranging from Cassandra databases to stateless Java applications.

NETFLIX



What happens when you need to move 89 million viewers to a different AWS region? Netflix's infrastructure, built on AWS, makes it possible to be extremely resilient, even when the company is running services in many AWS Regions simultaneously. In this episode of This is My Architecture, Coburn Watson, director of performance and reliability engineering at Netflix, walks through the company's DNS architecture—built on Amazon Route 53 and augmented with Netflix's Zuul—that allows the team to evacuate an entire region in less than 40 minutes.

[Netflix: Multi-Regional Resiliency and Amazon Route 53 - YouTube](#)

Q & A.....





BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Credits

*Hwang, Kai; Dongarra, Jack; Fox, Geoffrey C.. Distributed and Cloud Computing: From Parallel Processing to the Internet of Things (Kindle Locations 3532-3533). Elsevier Science. Kindle Edition.