O'REILLY®

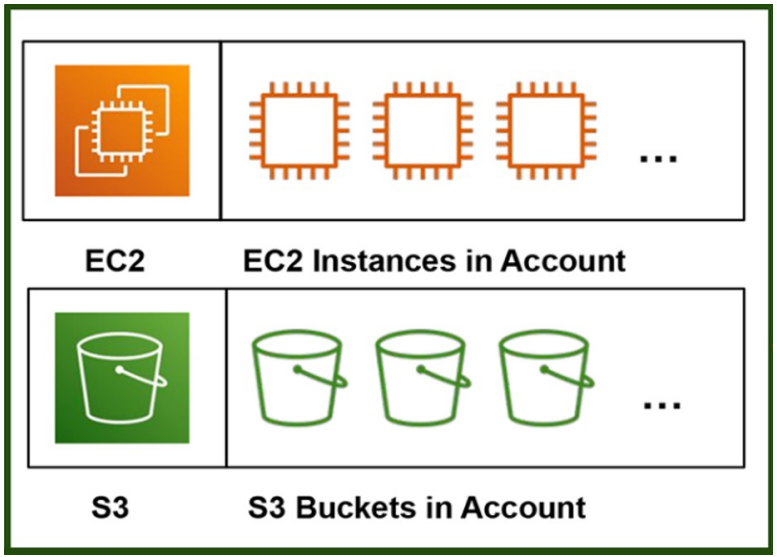
AWS Core Architecture Concepts

What We Will Cover

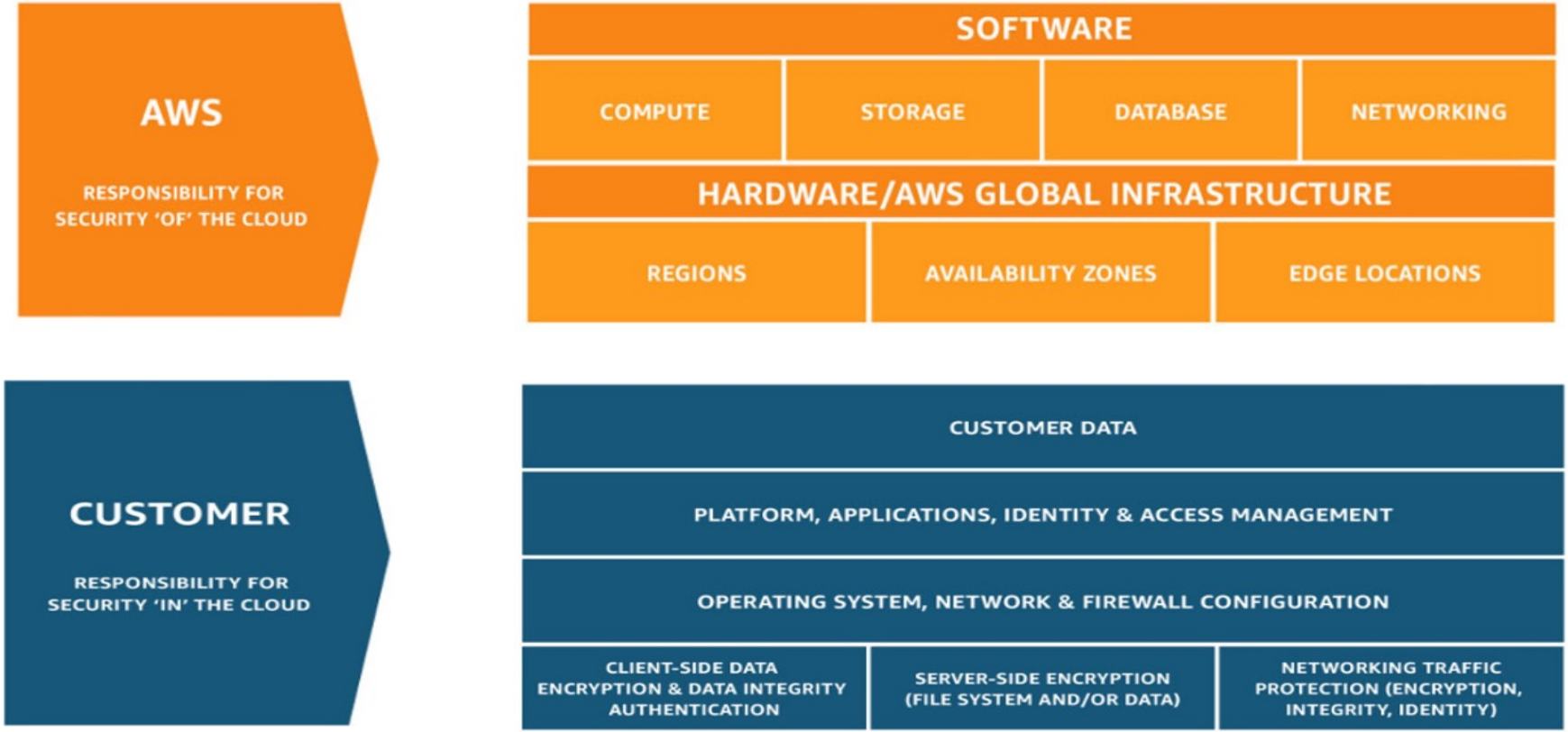
- Shared Responsibility Model
- Account Security and IAM
- Regions and Availability Zones
- VPC Architecture
- AWS Organizations
- Containers and EC2 Instances
- Load Balancing and Auto Scaling
- Storage Options
- RDS Databases

AWS Account

- Each AWS account is a container.
- Cloud services from > 26 AWS Regions hosted in a container.
- One “root” admin account - full access can’t be removed.



AWS Shared Responsibility Model



The Root Account

- When you first sign up with Amazon Web services, the first account created in your AWS account is a Root administrator account.
- The Root account’s powers can’t be curtailed.
- Full unrestricted access to all AWS account resources.
    - Store root account credentials in a safe location.
    - Enable MFA on the root account.

Root Account Tasks

- Set up Account billing
- Submit a Reverse DNS request
- Transfer Domain registration
- Change Root account details
- Change AWS support plan
- Create a CloudFront key pair
- Close AWS account
- View billing tax invoices

Identity and Access Mamagement

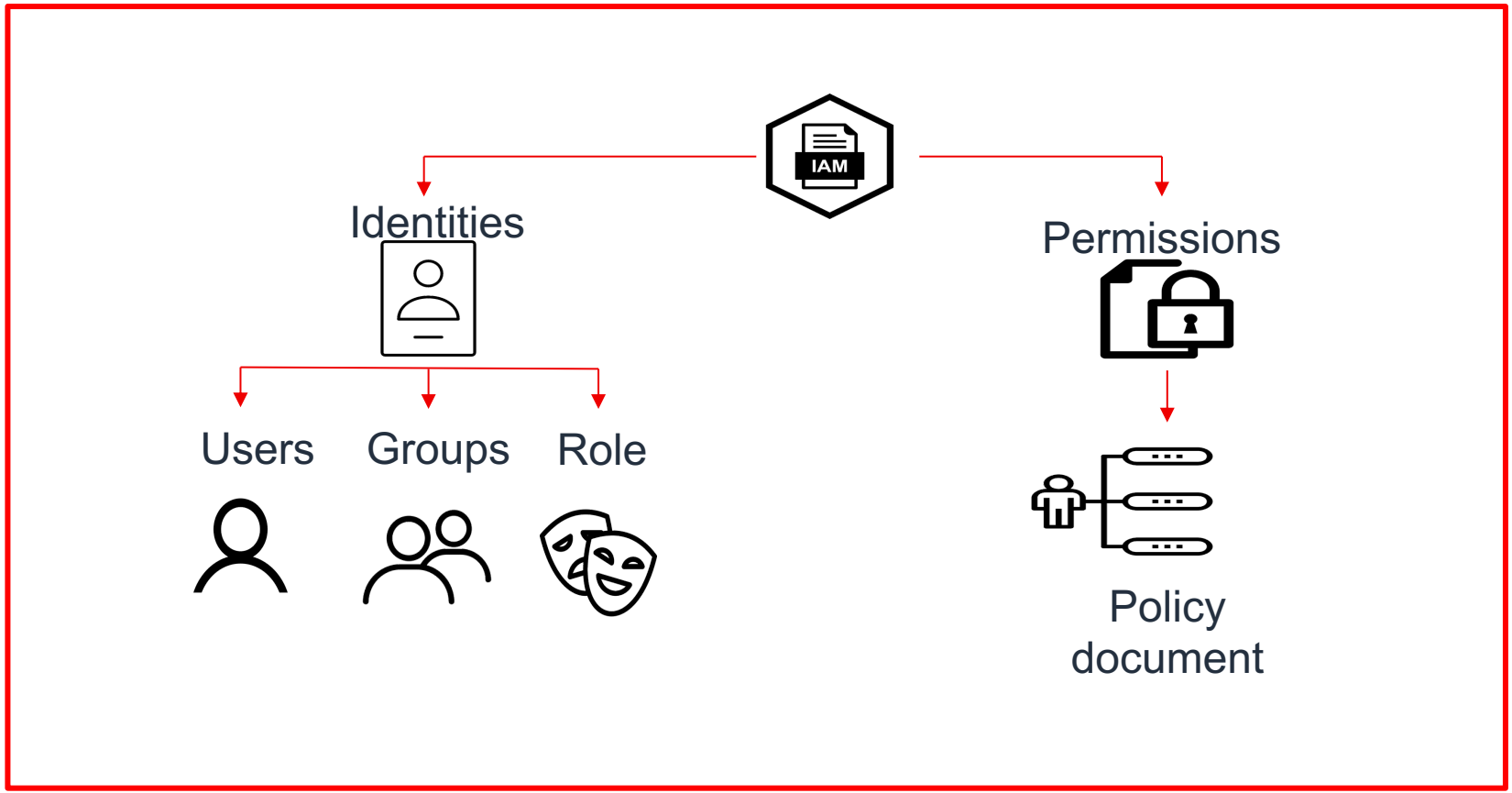
IAM Identities

IAM User

IAM Group

IAM Roles

IAM policies are attached to IAM identities


IAM
Identities
Users
Groups
Role
Permissions
Policy
document

Authentication and Authorization

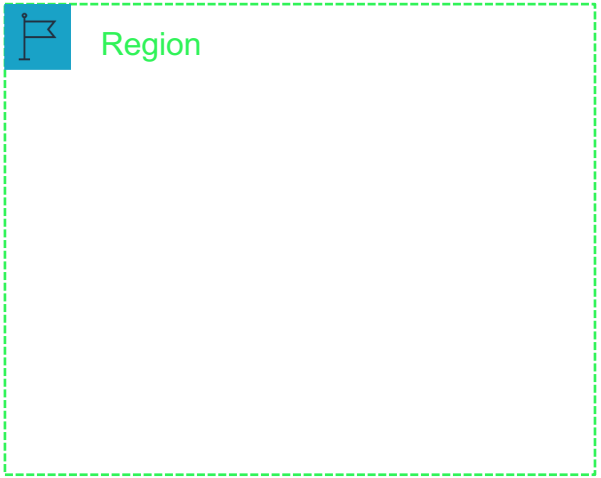
1. Requests are implicitly denied.
2. Next, the policy request is evaluated; if there is an explicit denial, the request is denied.
3. No explicit deny, but an explicit allow the request is allowed.
4. No explicit allow or deny permission was found; the default implicit deny is maintained.

AWS Regions

Each region contains at least 3 availability zones

AWS Regions

- Areas of the world where Amazon offers AWS cloud services.
- Each region is a geographical location.
- Each region is mostly independent and isolated.
- AWS regional services are not replicated to other regions by default.

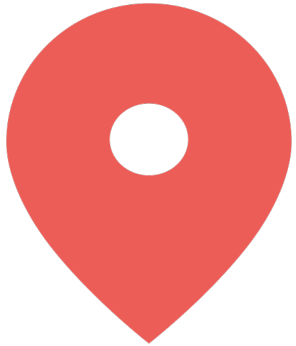


Control and Data Planes

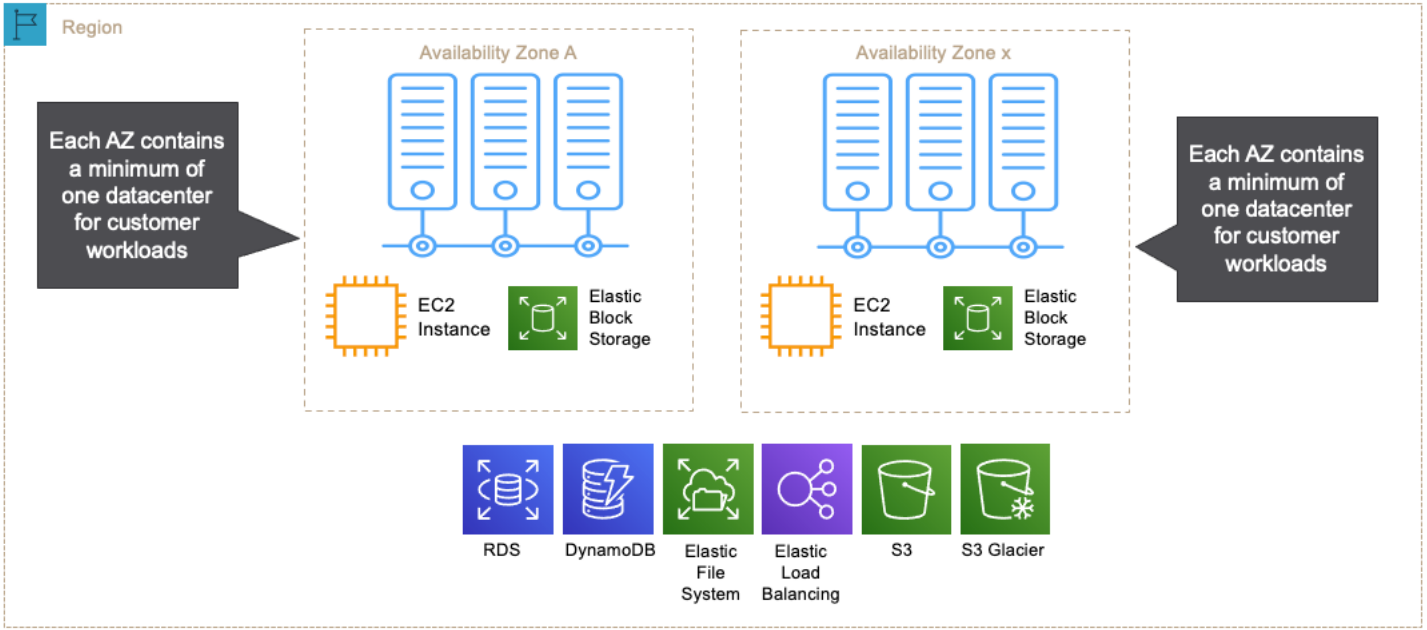
- AWS services are designed using control and data planes.
- Control planes manage administrative API calls (Create, read, write, update, delete)
- Control plane tasks: Launch an EC2 instance.
- Data plane tasks: Where the EC2 instance is deployed (AZ).
- Data plane access to resources is not dependent on the control plane.

Control Plane Locations

- AWS Organizations (us-east-1).
- AWS Account Management (us-east-1).
- Route 53 Application Recovery Controller (us-west-2).
- AWS Network Manager (us-west-2).
- Route 53 Private DNS (us-east-1).



Regional Services



Choosing an AWS Region

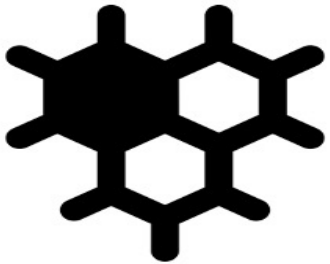
Latency to on-premise location

Costs are different per Region

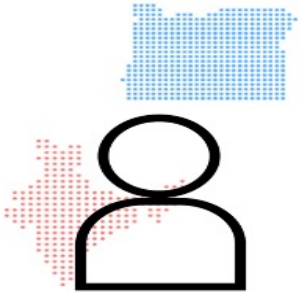
Features are different per Region

Compliance rules and regulations

Availability Zone Design

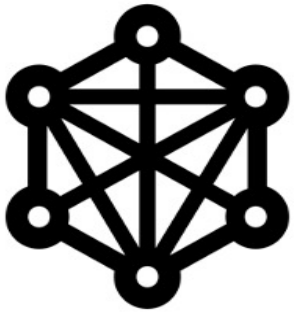


Designed as independent failure zones.



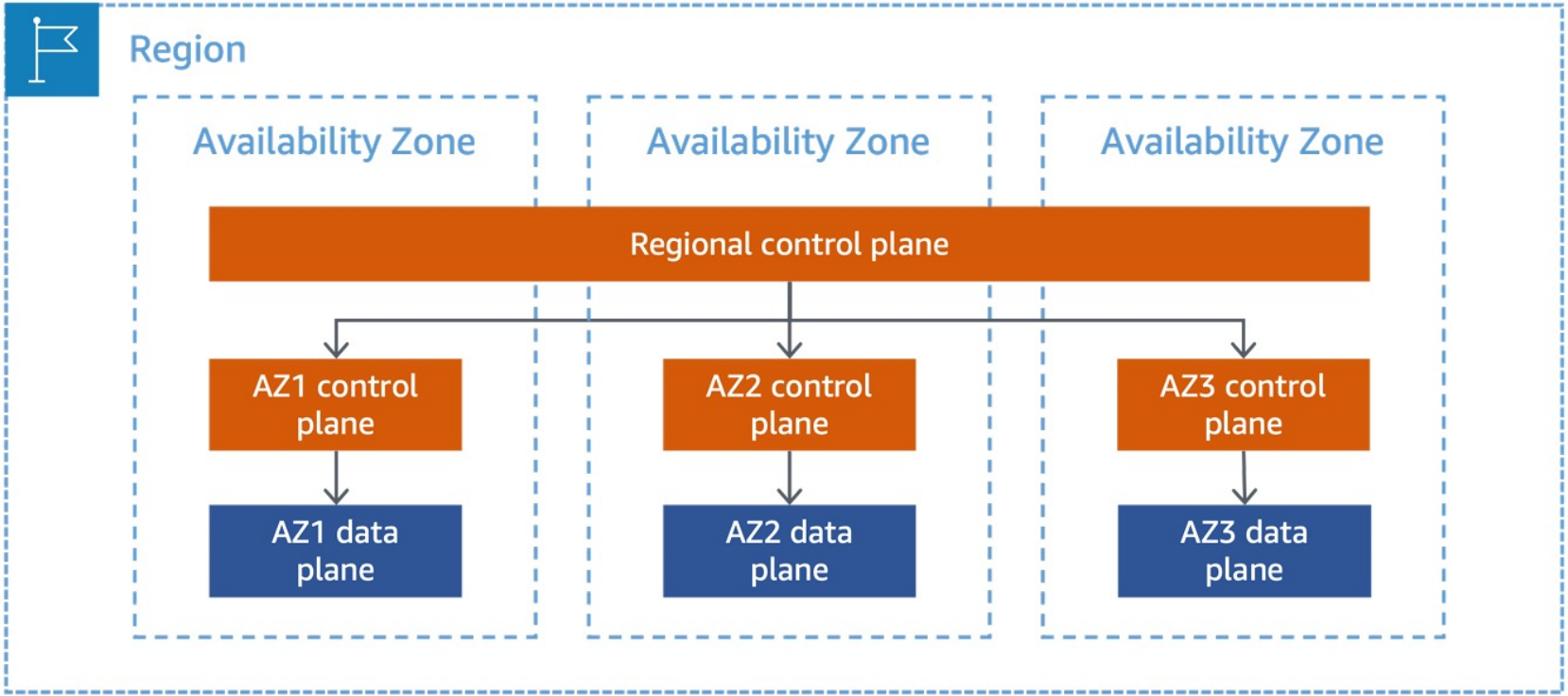
AWS account has access to all regions and associated AZ’s.

Data transfer charges for outbound network traffic across AZ’s.

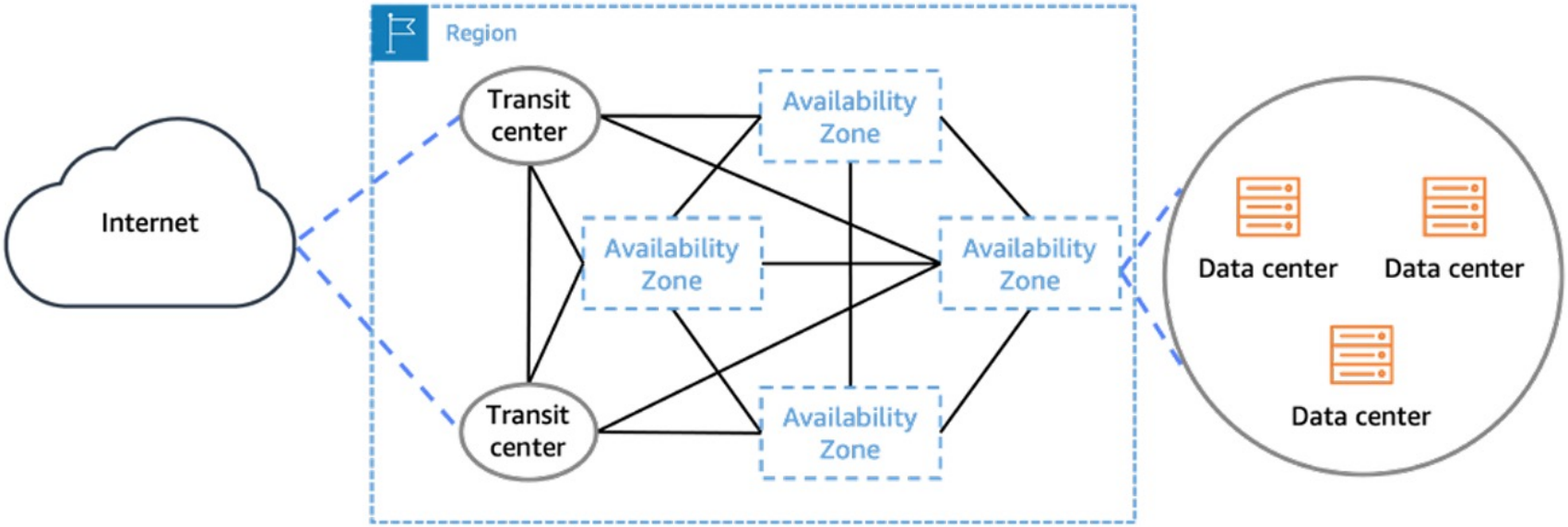


Redundant Tier-1 transit private fiber connections.

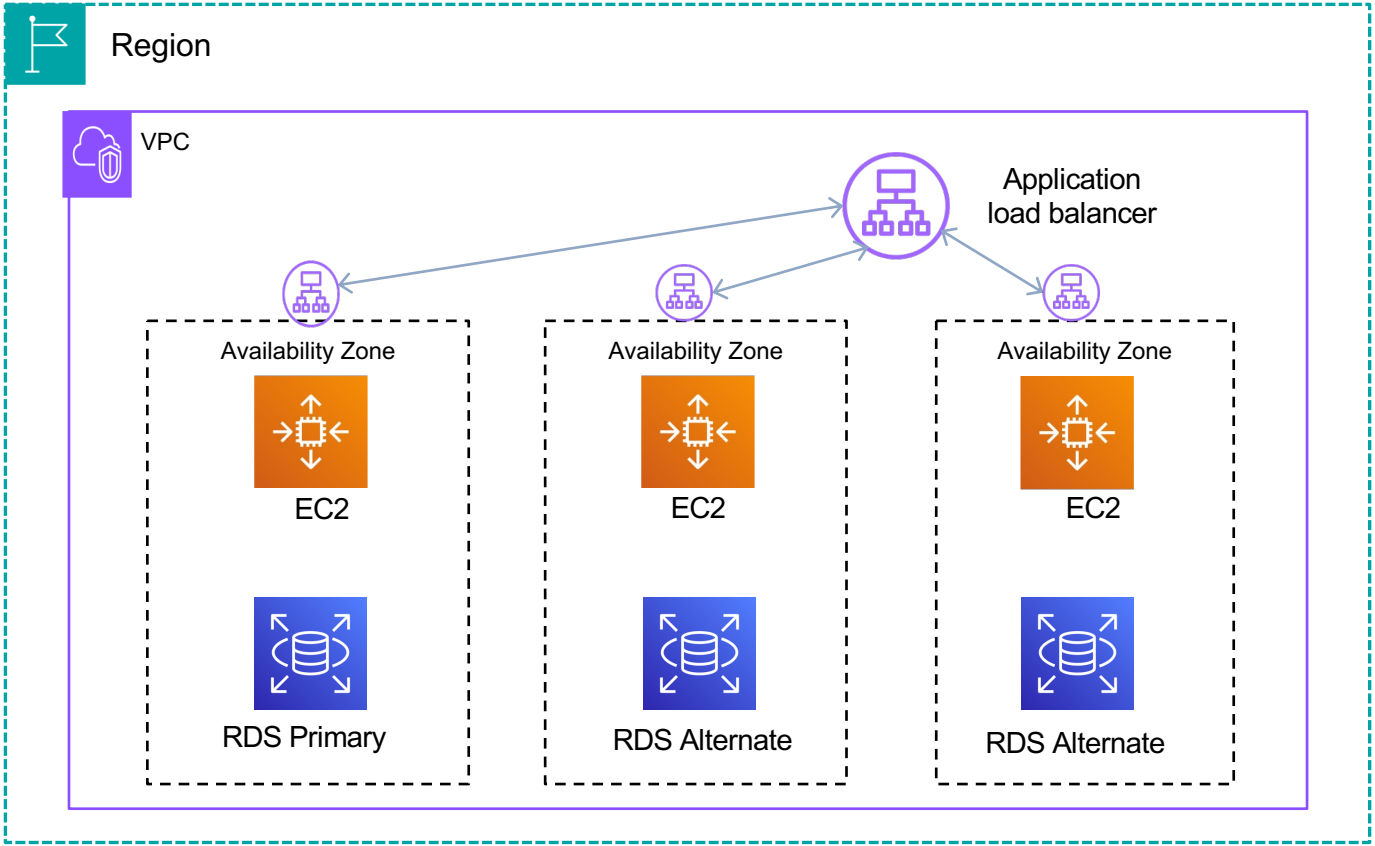
Availability Zone Independence



Availability Zone Architecture

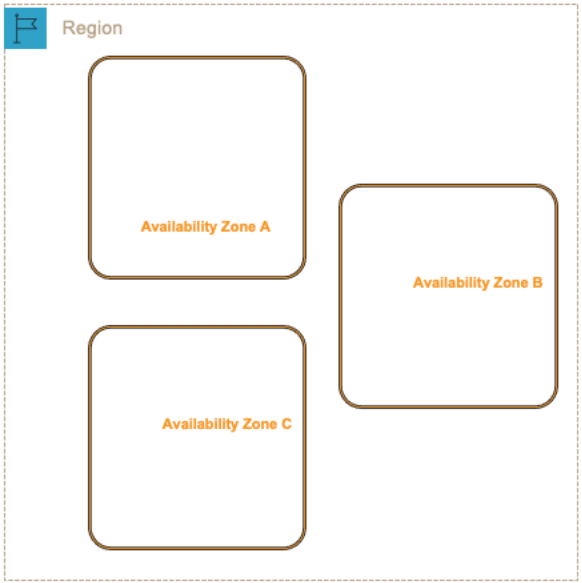


Multi-AZ Design



Availability Zones Summary

- Each availability zone contains at least one data center for customer workloads.
- Many availability zones contain multiple data centers for customer workloads.
- Each availability zone has inexpensive low-latency network connectivity to the other availability zones in the same region.
- Three AZs per region.



Availability Zones In Operation

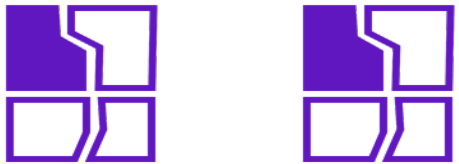
- EC2 instances are deployed in multiple subnets in multiple availability zones.
- ELB targets EC2 instances in multiple availability zones.
- EC2 Auto Scaling can scale instances in multiple availability zones.
- RDS solutions are deployed in multiple availability zones.
- Amazon Aurora is replicated across multiple availability zones.
- DynamoDB tables are replicated across multiple availability zones.

Single Availability Zone

- No recovery or failover when disaster happens in a single datacenter.
- No high availability for EC2 instances.
- No failover in single datacenter.
- All AWS regions have at least 3 availability zones.

Multiple Availability Zones

- Better high-availability design options.
- Designing applications hosted across AZs provides HA options.
- Load balancing (ELB) supports EC2 instances in multiple availability zones).
- Auto-scaling supports multiple AZs.



Protecting the Application Perimeter

AWS Shield Standard

Protects against DDOS attacks.

AWS Shield Advanced

Enhanced protection.

AWS WAF

Protection with rules.

Filtering rule

VCP Architecture

What’s a VPC?

- Networking layer at AWS Cloud.
- Logically isolated to your AWS account.
- EC2 instances are hosted on virtual private cloud subnets.

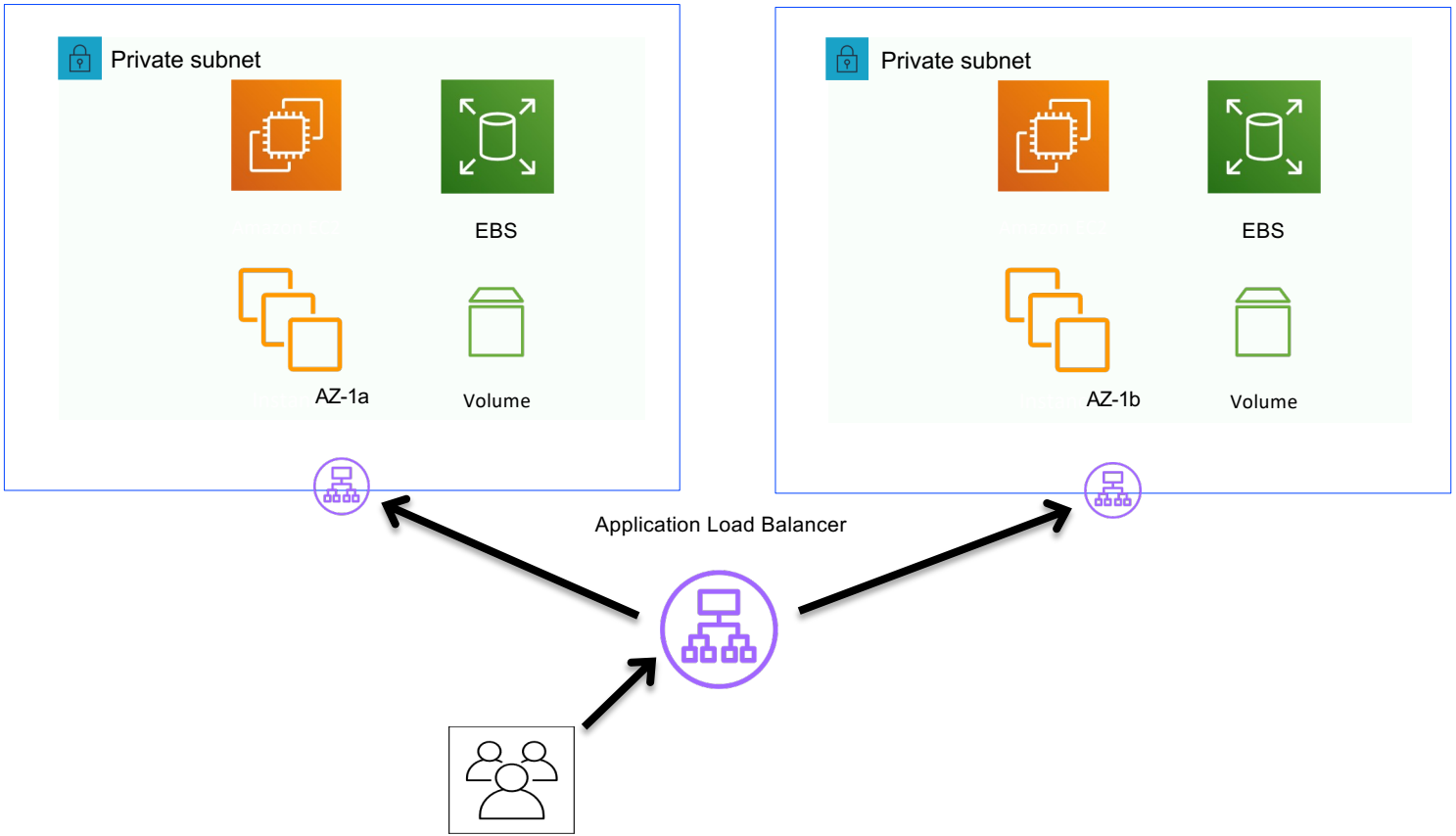
VPC Requirements

- Application stack isolation.
- Production from non-production.
- Multi-tenant / single-tenant isolation.
- Business unit design.
- Shared corporate services.

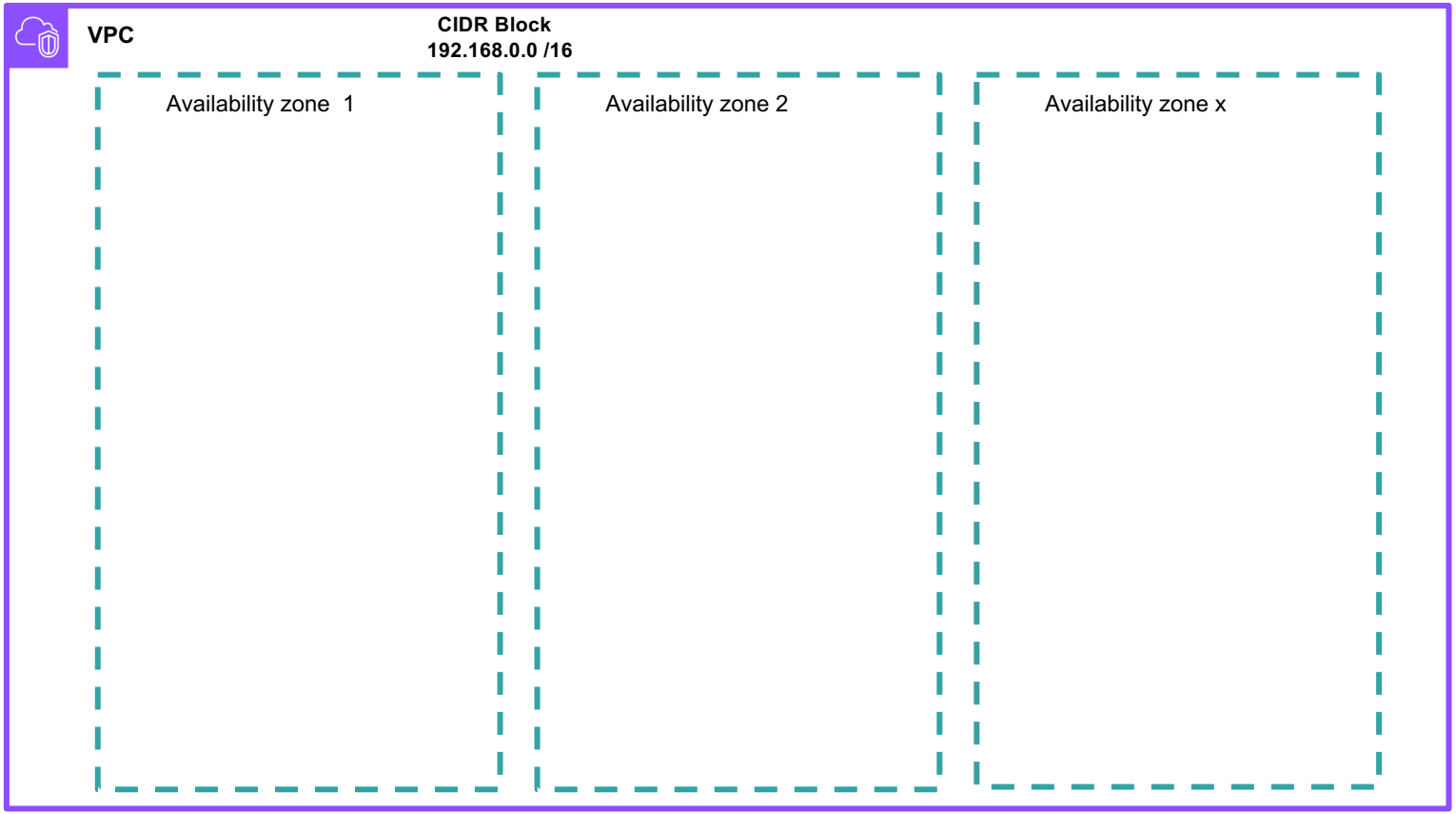
Functional Design

- VPCs can span all availability zones of the AWS region.
- Subnets are hosted within the selected availability zone.
- Route table entries control subnet traffic.
- NACLs allow or deny subnet traffic.

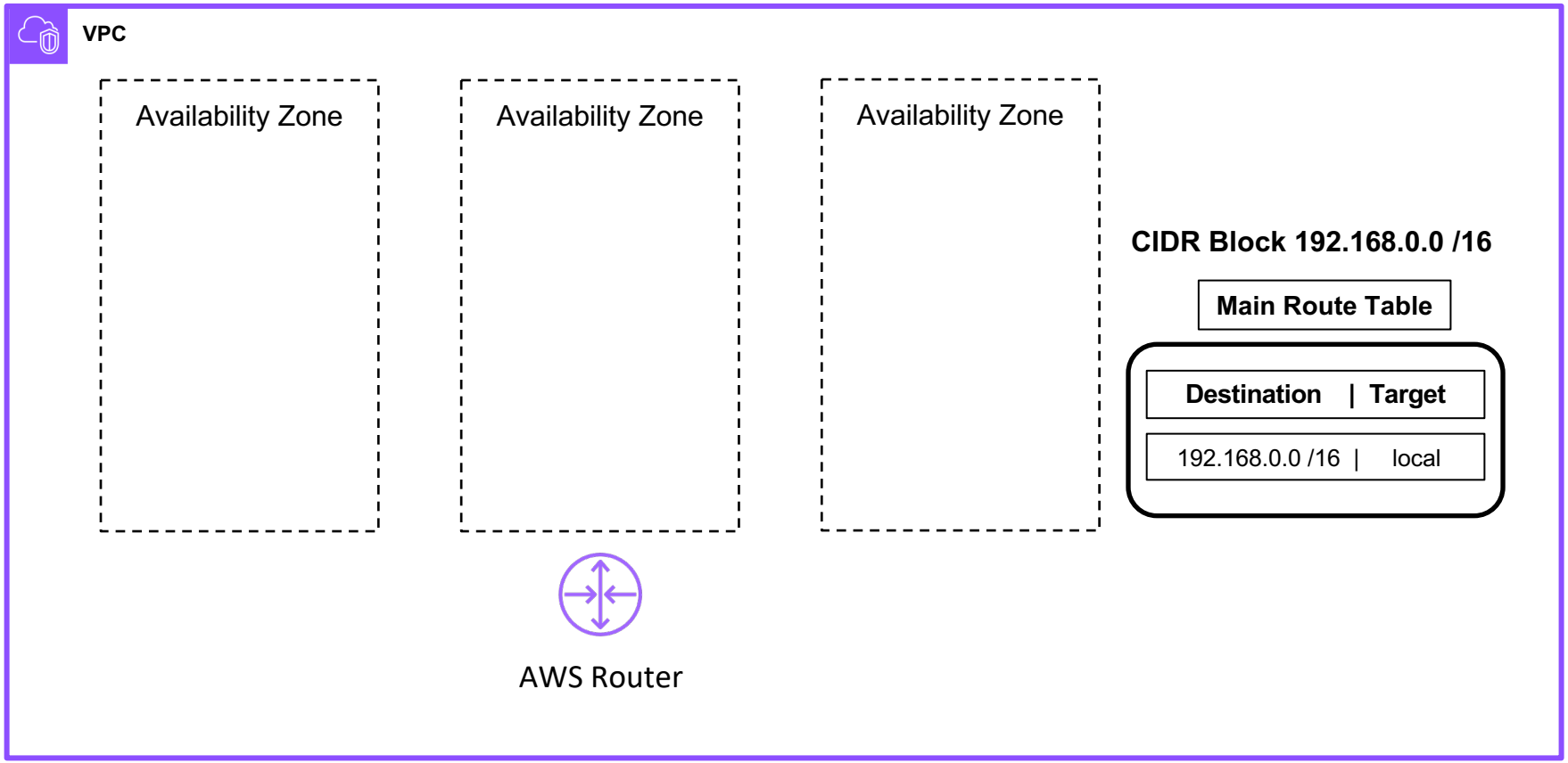
Failover Possibilities



VPCs Span All Regional AZs



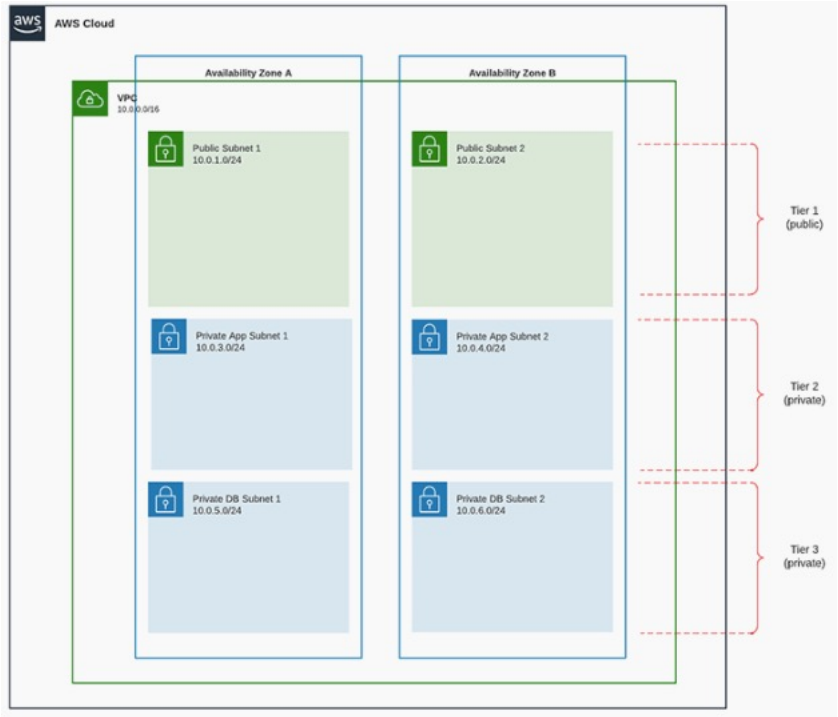
VPC Components



Subnets and IP Addresss

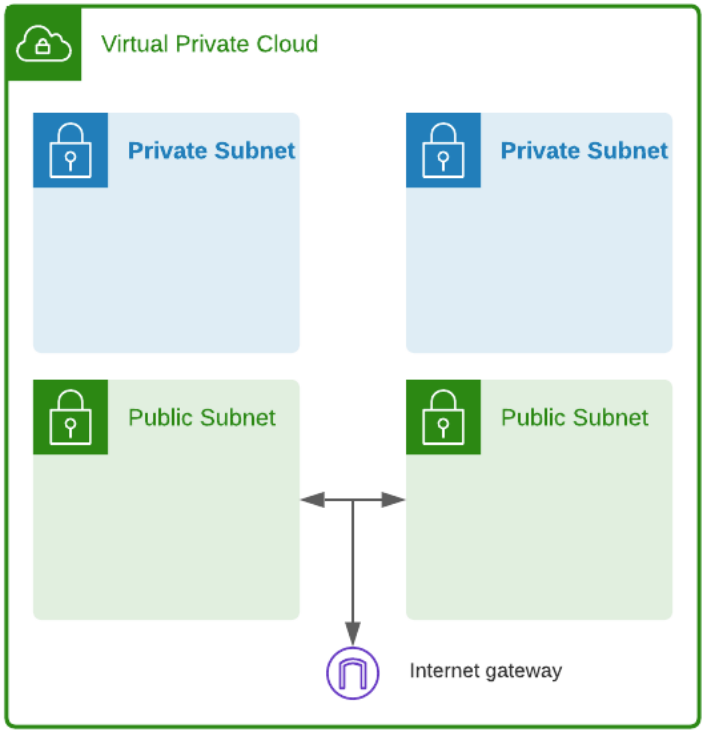
Subnets

- EC2 instances are hosted on VPC subnets.
- Public subnets can be used for resources/services that need direct Internet access (IGW, ELB, NAT Services).
- Private subnets host resources that don’t directly connect to the Internet (e.g., Web and Application servers and RDS instances).



Subnet Rules

- Public or private subnets can be created in availability zones.
- Subnets cannot span across multiple availability zones.
- A subnet that doesn’t directly route to the Internet gateway is private.
- A subnet that directly routes to an Internet gateway is public.

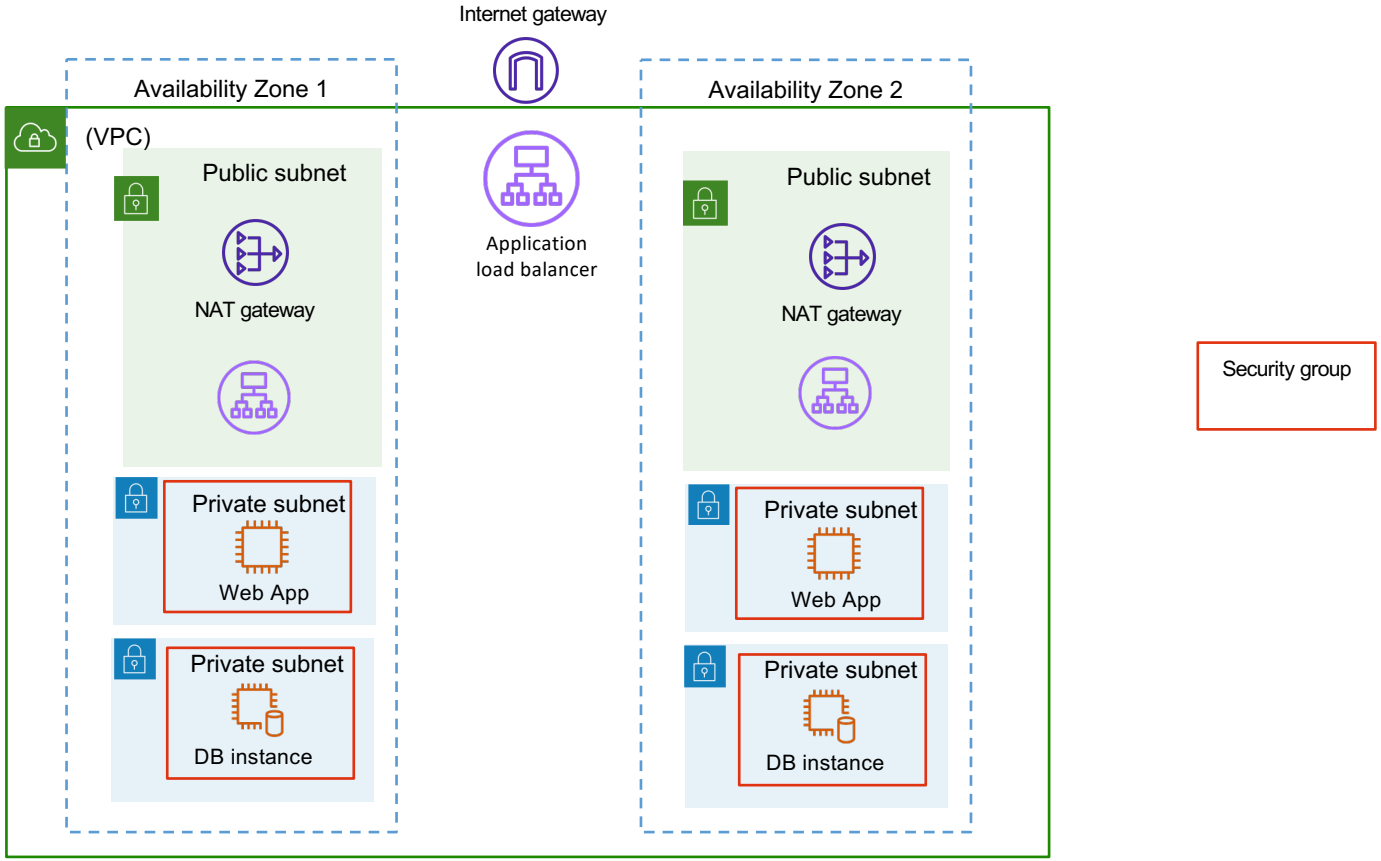


IP Addresses

- Each EC2 instance is assigned a private DNS hostname associated with its IP address.
- IPv4 is the default and required address, while IPv6 is optional.

- EC2 instance address types:
  - Private / Public IPv4
  - Static public IPv4
  - Elastic IP address (IPv4)
  - Public IPv6

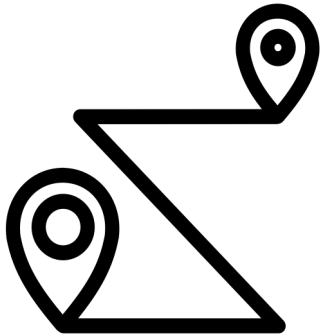
Public - Private Subnets



Route Tables

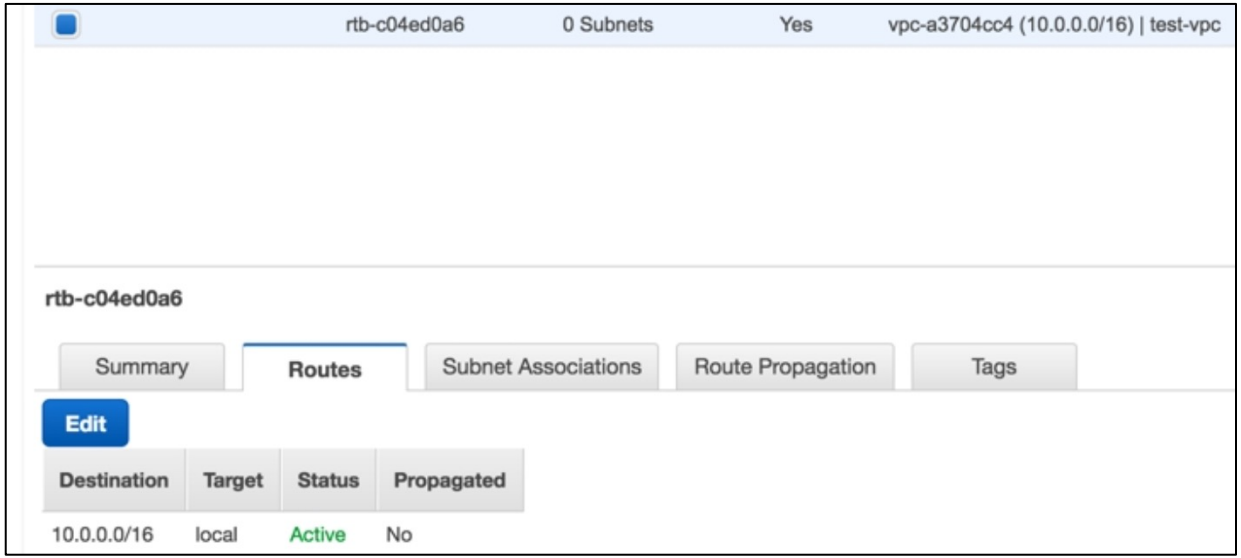
Route Tables

- External traffic patterns are defined by adding additional route table entries.
- Each route specifies a destination and a target.
- Route tables connect subnets to an Internet gateway (IGW), a VPN gateway (VGW), or a NAT gateway.
- When you create a VPC, a default route table is created.

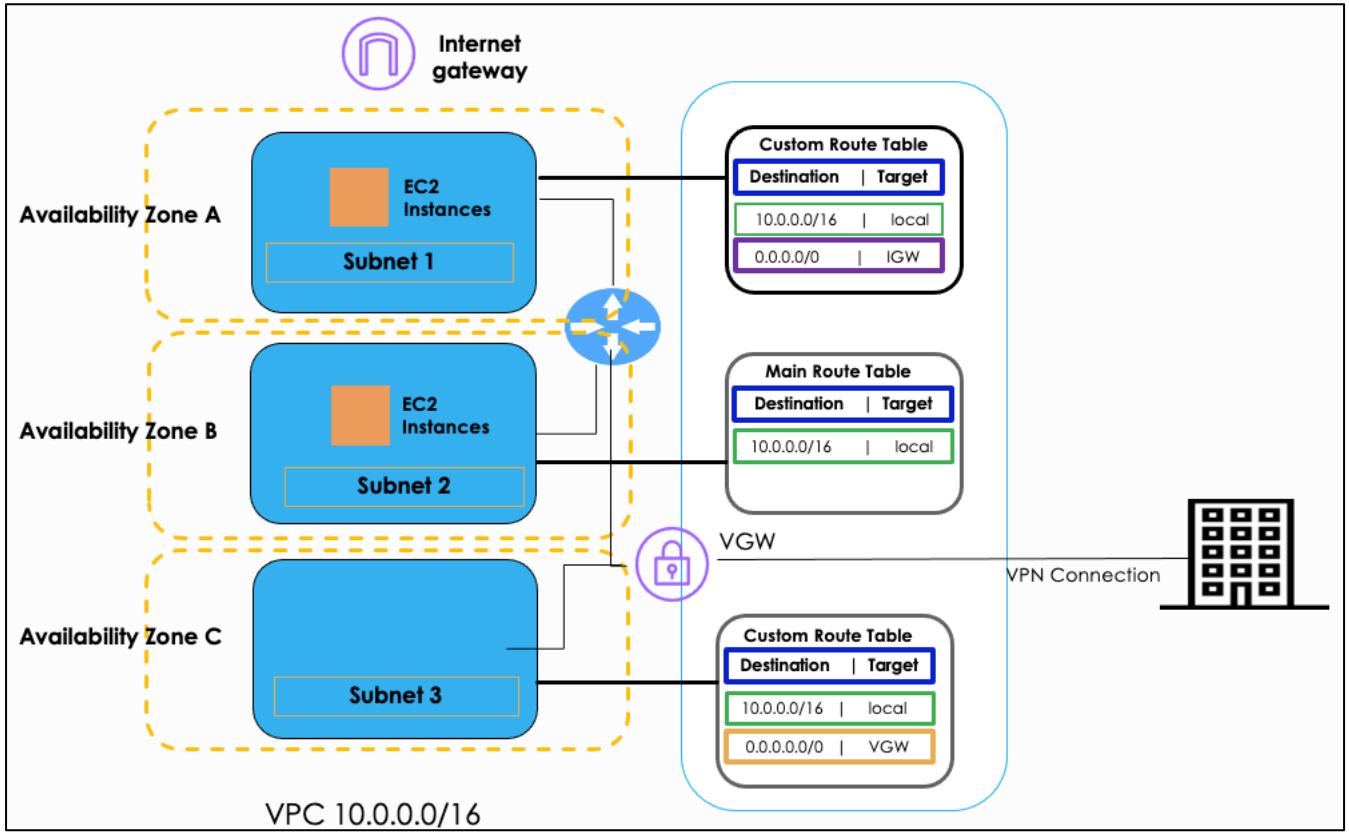


The Default Route Table

- Each new subnet is associated with the default route table, which is implicitly attached to subnets not explicitly associated with a custom route table.
- Each custom route table is assigned to a custom subnet.



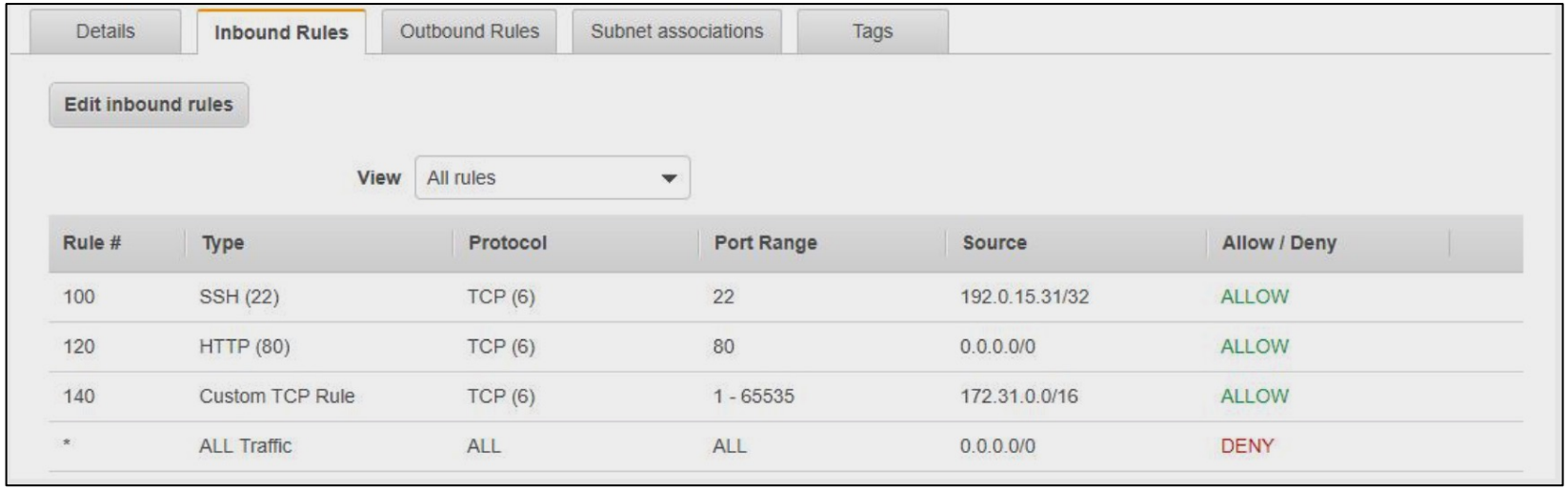
Route Table Entries



Network Access Control Lists

Network Access Control List

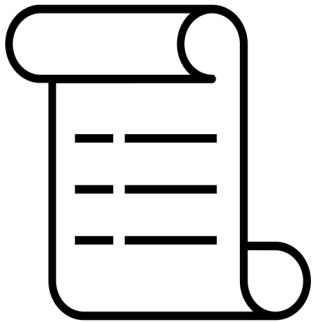
- Nacls control traffic in and out of the subnet.
- Each subnet must be associated with a Nacl.
- Nacl rules are defined as stateless.
- Rules are evaluated in order.

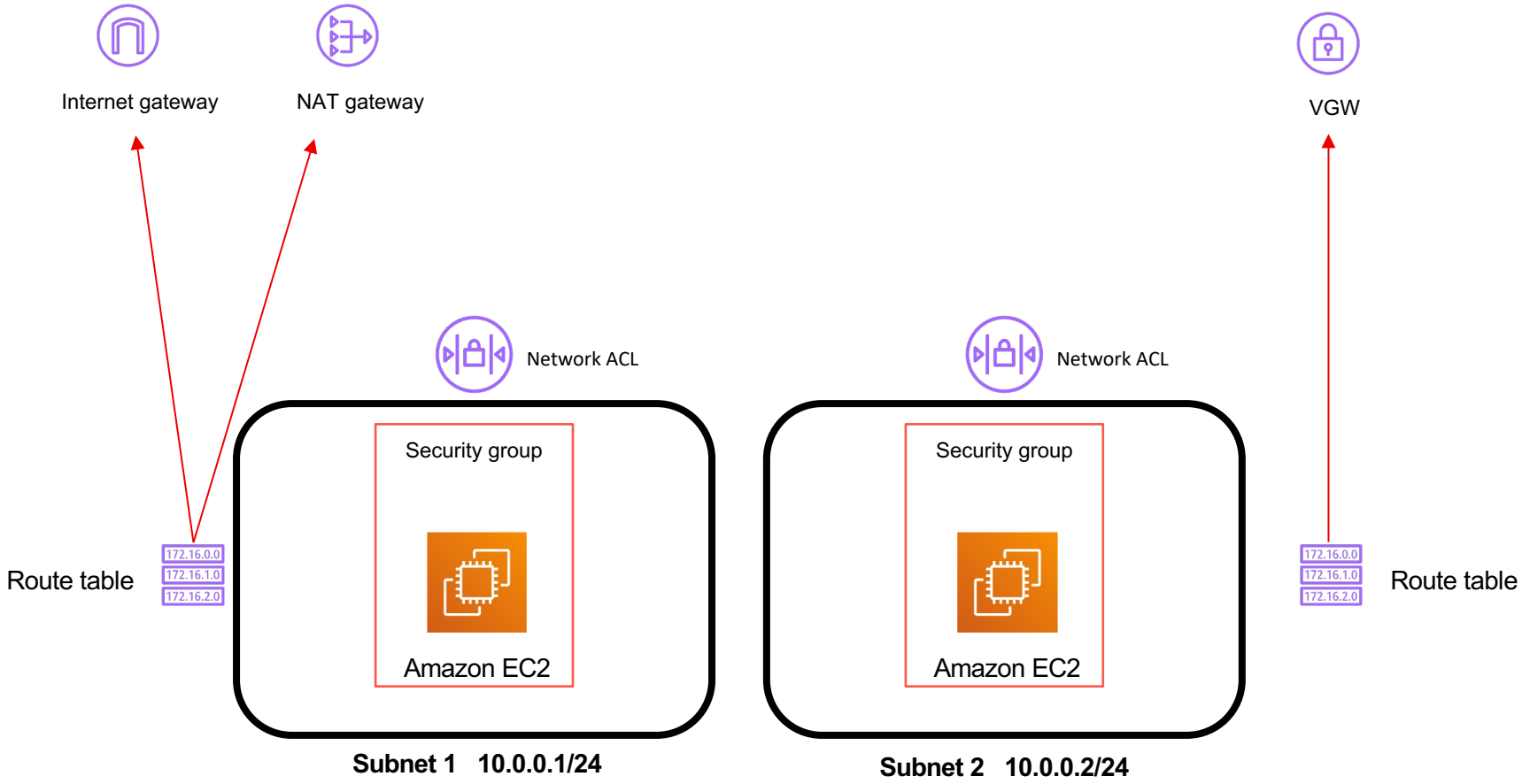


Rule #	Type	Protocol	Port Range	Source	Allow / Deny
100	SSH (22)	TCP (6)	22	192.0.15.31/32	ALLOW
120	HTTP (80)	TCP (6)	80	0.0.0.0/0	ALLOW
140	Custom TCP Rule	TCP (6)	1 - 65535	172.31.0.0/16	ALLOW
*	ALL Traffic	ALL	ALL	0.0.0.0/0	DENY

Network ACL Rules

- Inbound Rule - Allow or deny the specified traffic pattern
- Outbound Rule - Allow or deny the specified traffic pattern
- Inbound and outbound rules are separate entities.
- Use Nacls for blocking IP address ranges.



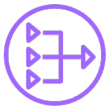

Internet gateway
NAT gateway
VGW
Network ACL
Network ACL
Security group
Security group
Amazon EC2
Amazon EC2
Route table
172.16.0.0
172.16.1.0
172.16.2.0
172.16.0.0
172.16.1.0
172.16.2.0
Route table
Subnet 1 10.0.0.1/24
Subnet 2 10.0.0.2/24

Gateways

AWS Gateways

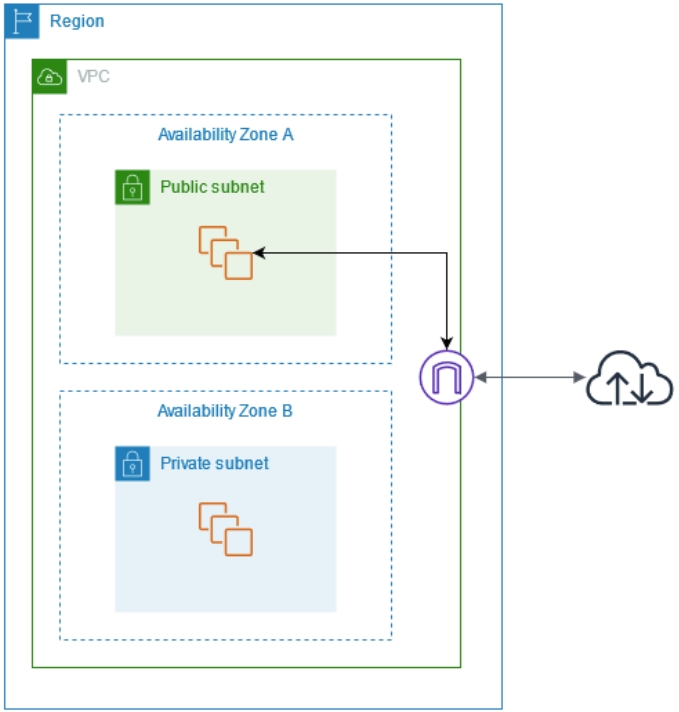
- Internet Gateway (IGW) – VPC side of the public Internet.
- VGW - VPN endpoint on AWS side of private connection.
- NAT Gateway - Private IPv4 instance indirect Internet access.
- Egress-only Internet Gateway – Public IPv6 protection from Internet access.
- Transit Gateway – Custom customer private routing paths.





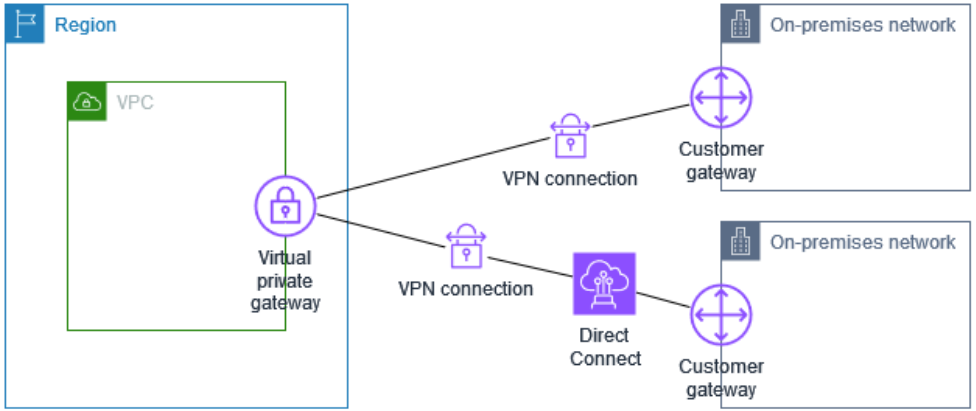

Internet Gateway

- An internet gateway allows communication between a VPC and the internet.
- It supports IPv4 and IPv6 traffic.



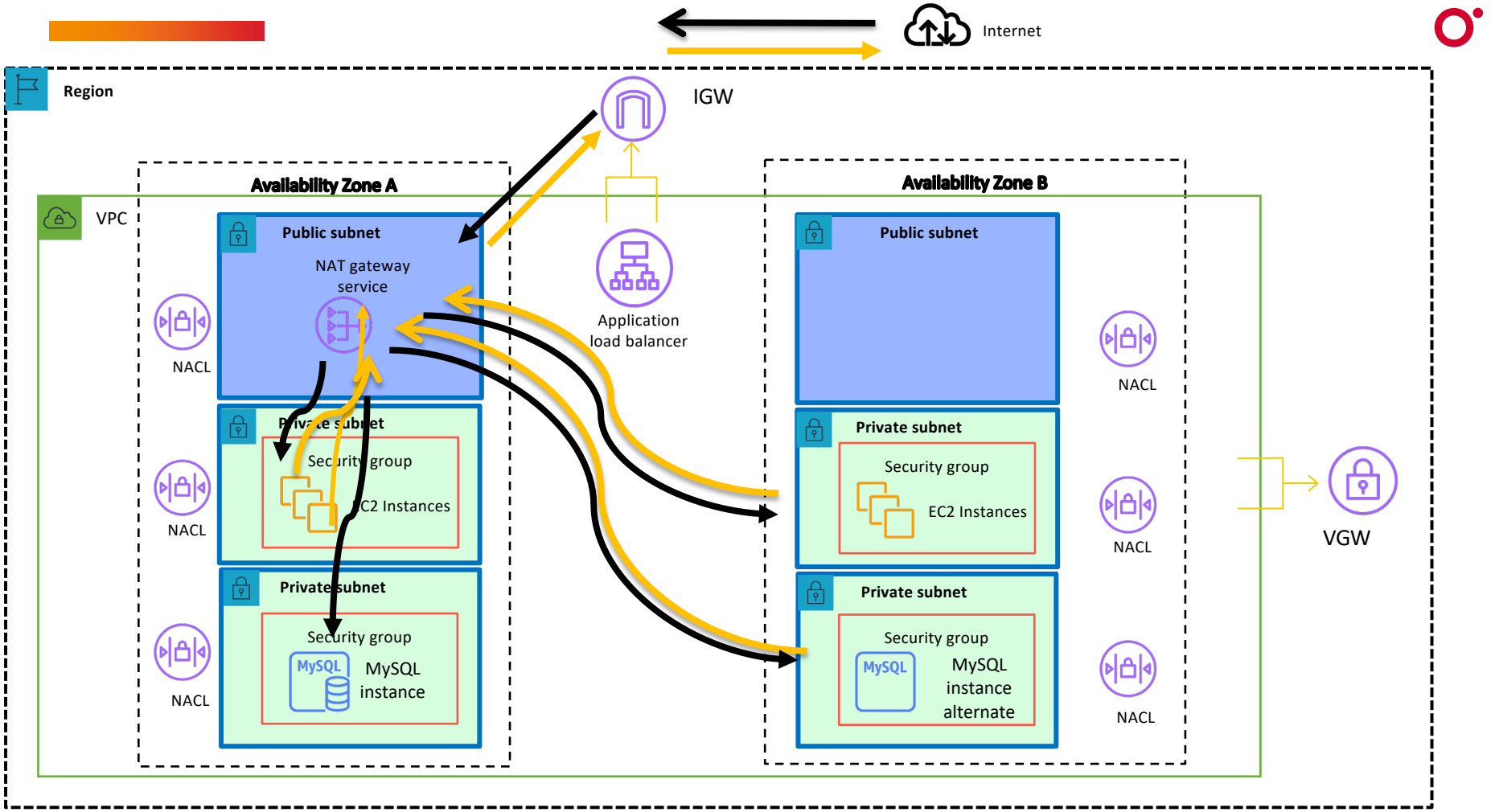
Virtual Private Gateway

- The VPN endpoint on the Amazon side of your Site-to Site VPN connection is attached to a single VPC.

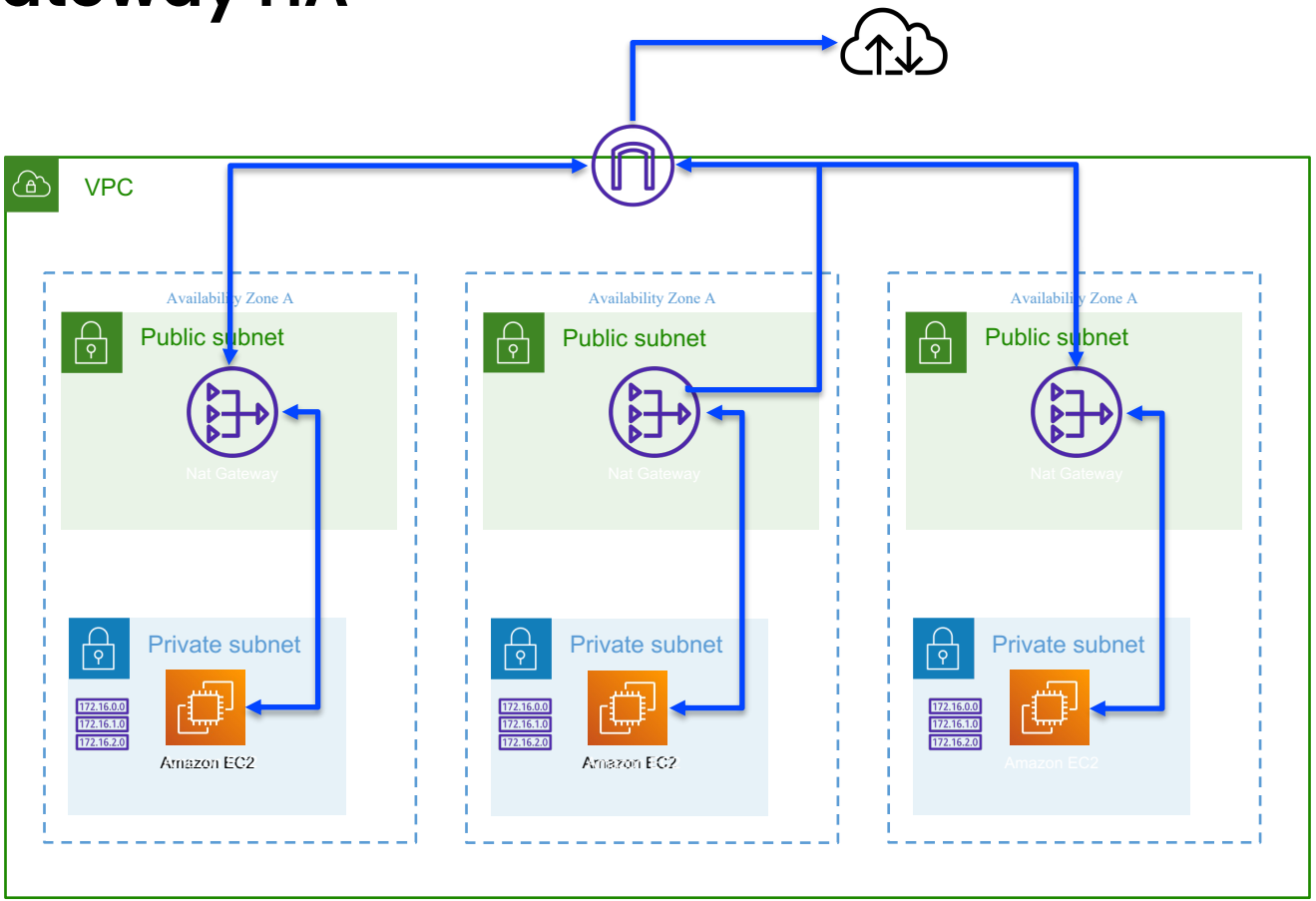


NAT Gateway Services

- NAT service enables instances in private subnets to connect to the Internet to get updates.
- Traffic requests are forwarded to the NAT service hosted in the public subnet.
- Internet response is sent back to the instance that made the request
- NAT Options:
    - NAT gateway service provided by AWS
    - NAT instance – compute instance created with a NAT AMI

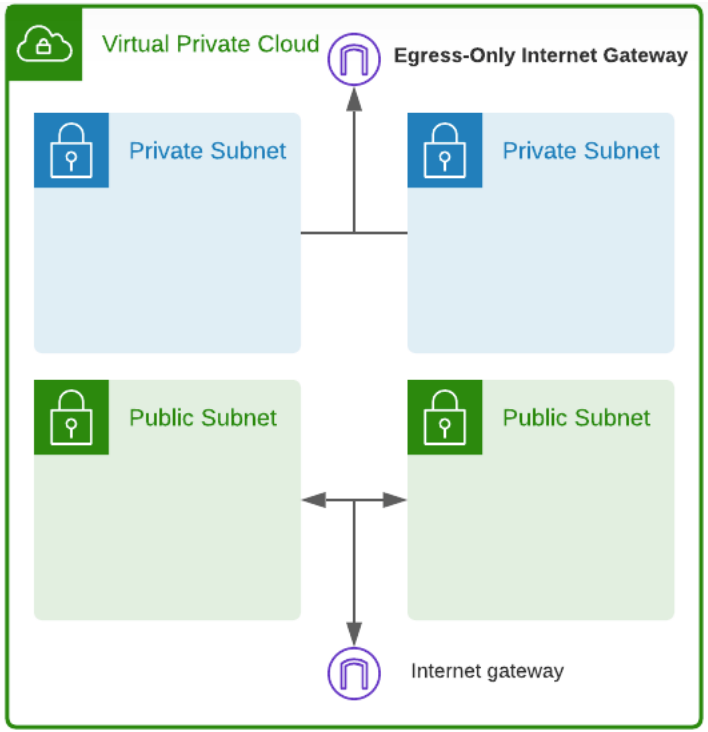

Internet
Region
IGW
Availability Zone A
Availability Zone B
VPC
Public subnet
NAT gateway service
Application load balancer
NACL
Private subnet
Security group
EC2 Instances
Private subnet
Security group
MySQL
MySQL instance
Public subnet
Private subnet
Security group
EC2 Instances
Private subnet
Security group
MySQL
MySQL instance alternate
NACL
VGW

NAT Gateway HA



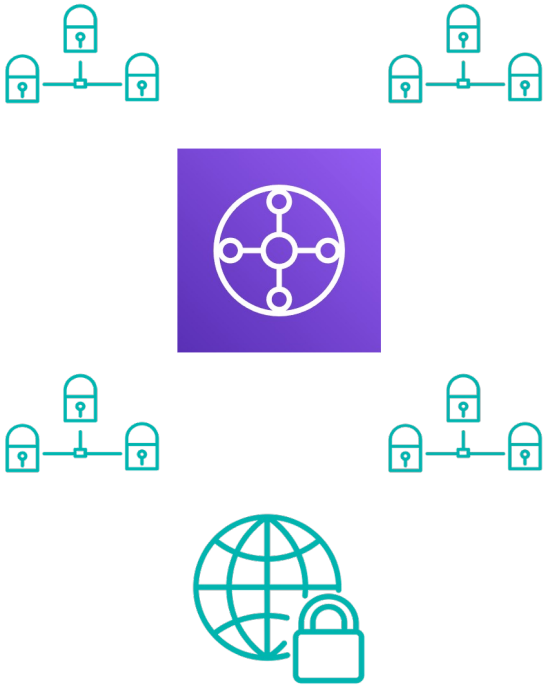
Egress-only Internet Gateway

- Outbound Internet access for IPv6 EC2 Instances.
- Stops direct inbound access.
- Forwards request to the Internet and send back the response.
- Egress-Only Internet Gateway **instead** of NAT gateway for IPv6 instances.

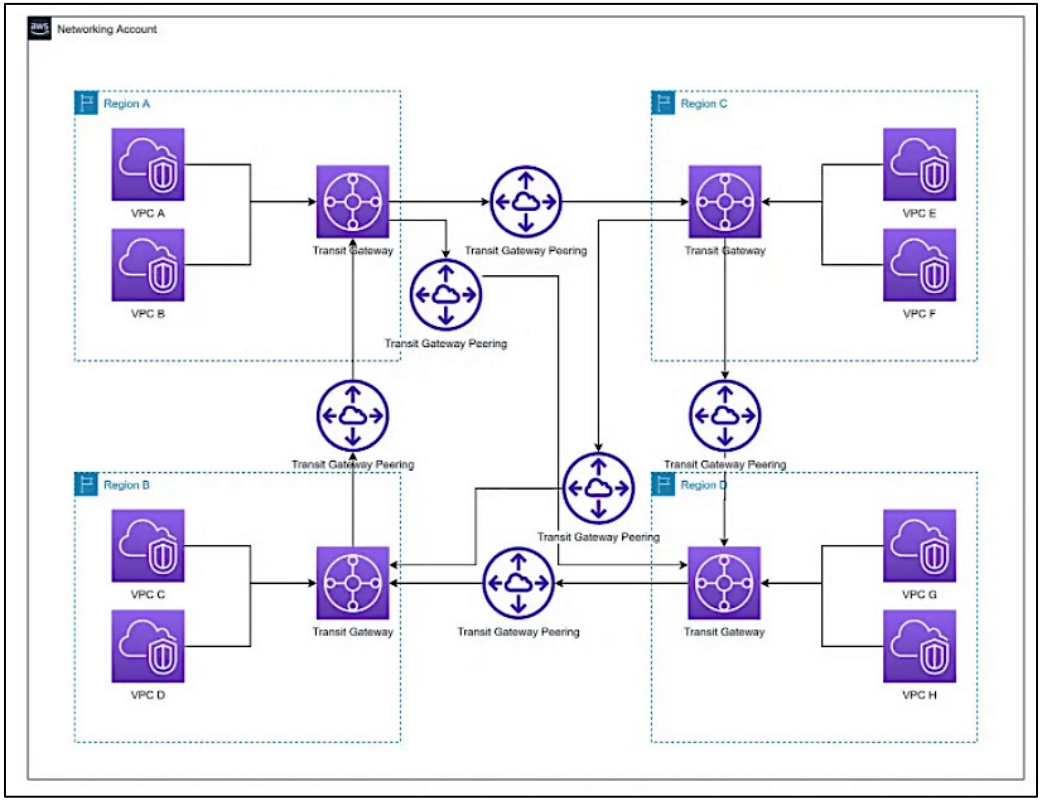


Transit Gateway

- VPCs connect to the Transit gateway using a hub and spoke model.
- Transit gateway traffic remains on AWS's private network.
- Supports dynamic and static routing between VPCs and VPN, Direct Connect connections.
- Custom transit gateway route tables allow you to segment and isolate your networks.
- Connect AWS regions using Transit Gateway Peering connections.

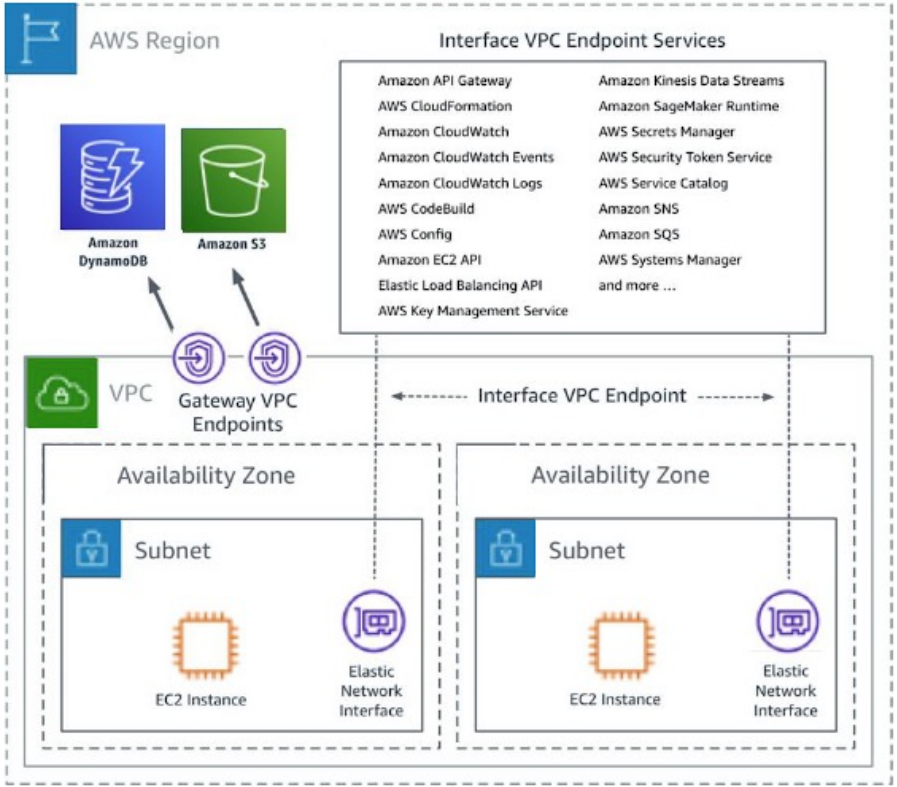


Transit Gateway Architecture



Endpoints

VPC Endpoints

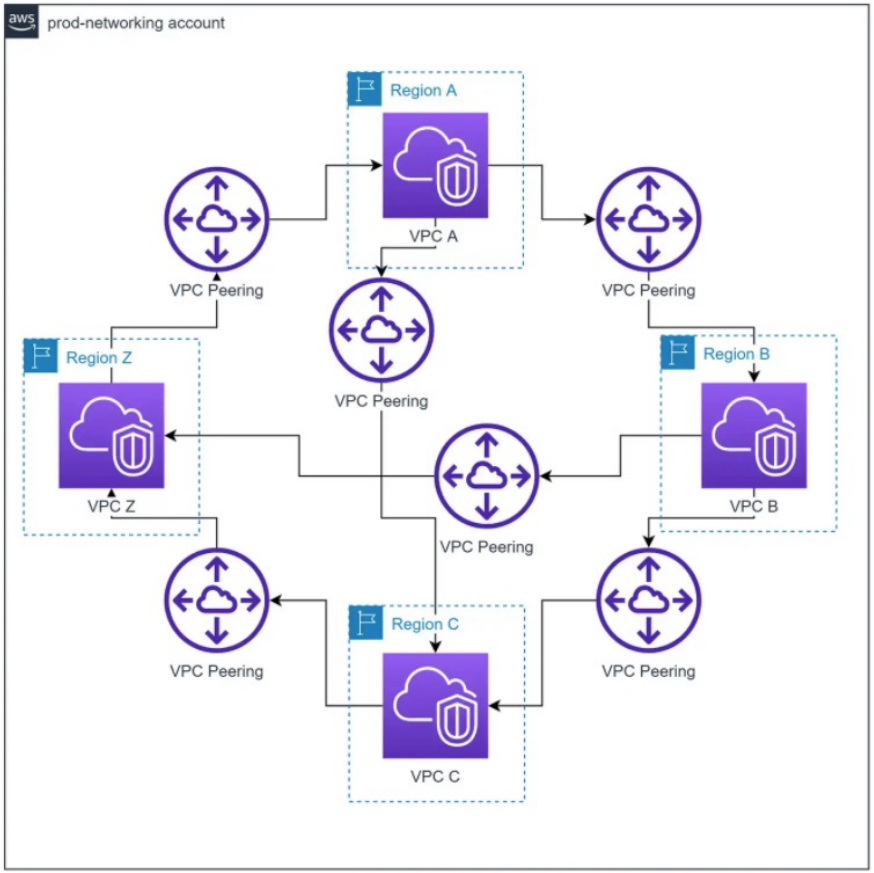


Endpoint Services

- Provides private connectivity between VPCs, AWS services, and on-premises applications securely on the Amazon network
- Connect services across different accounts and VPCs to simplify the network architecture significantly.
- Create endpoints to AWS or Marketplace service in subnets.
- Applications in a VPC can securely access PrivateLink endpoints across regions using Inter-Region VPC Peering.

Peering VPC’s

- Networking connection between two VPCs.
- Peer your VPCs or between other account holders' VPCs using a private IP address.
- Peering is a one-to-one relationship.
- Peering connections are not transitive.
- CIDR blocks can’t overlap in a peering relationship.
- Peering connections can be created between VPCs in the same or different regions.

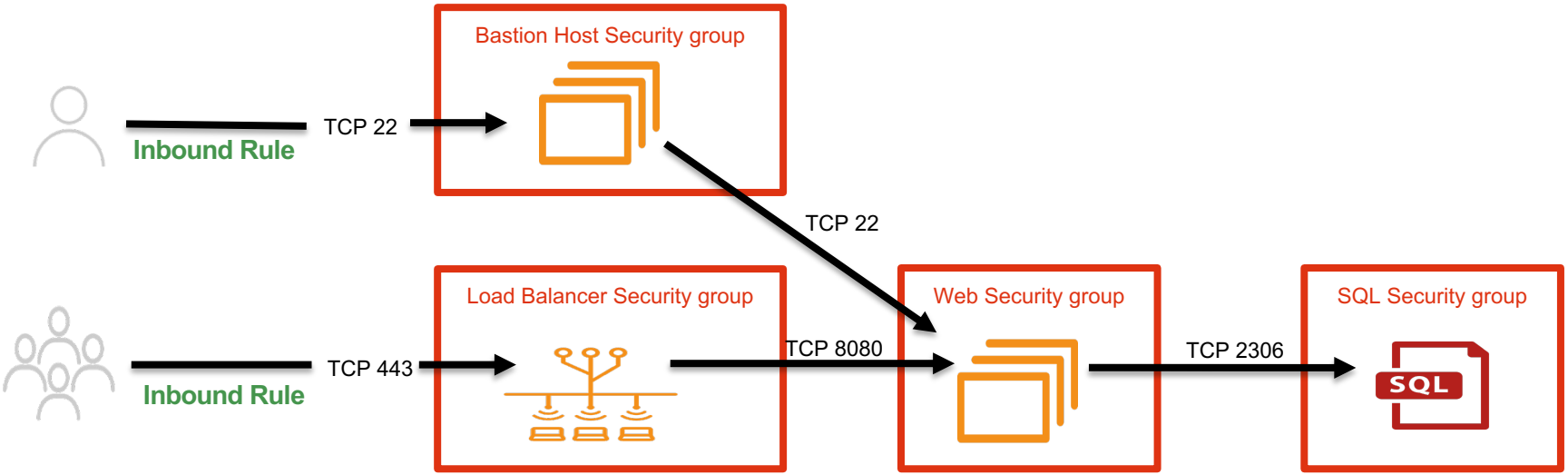

prod-networking account
Region A
VPC A
VPC Peering
VPC Peering
VPC Peering
Region Z
VPC Z
Region B
VPC B
VPC Peering
VPC Peering
VPC Peering
Region C
VPC C
VPC Peering

Security Groups

Security Groups

- Required EC2 instance firewall allow rules.
- Explicit deny rules can’t be specified.
- Inbound rules define the source of the traffic, and the destination port, port range, or security group.
- Any TCP protocol that is defined with a standard port number.
- Requests made inbound are allowed outbound.
- A security group can be a source or destination for other security groups.

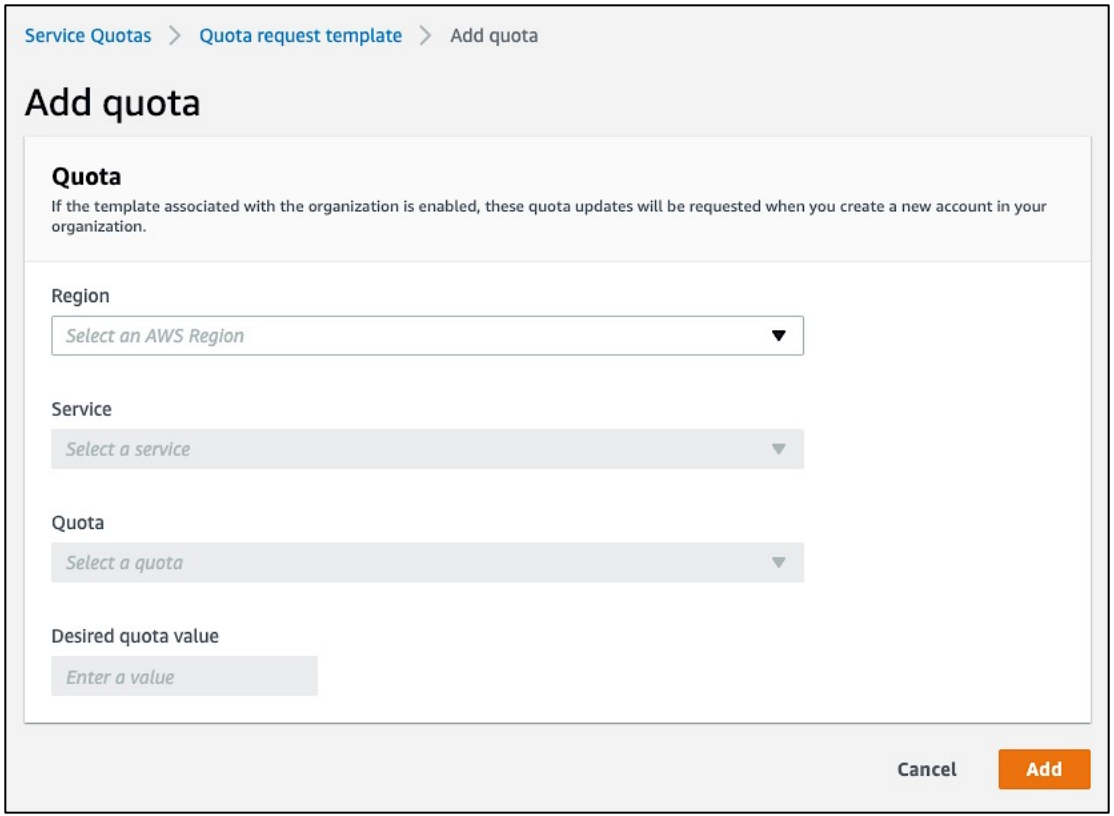
Security Group Design



Service Quotas

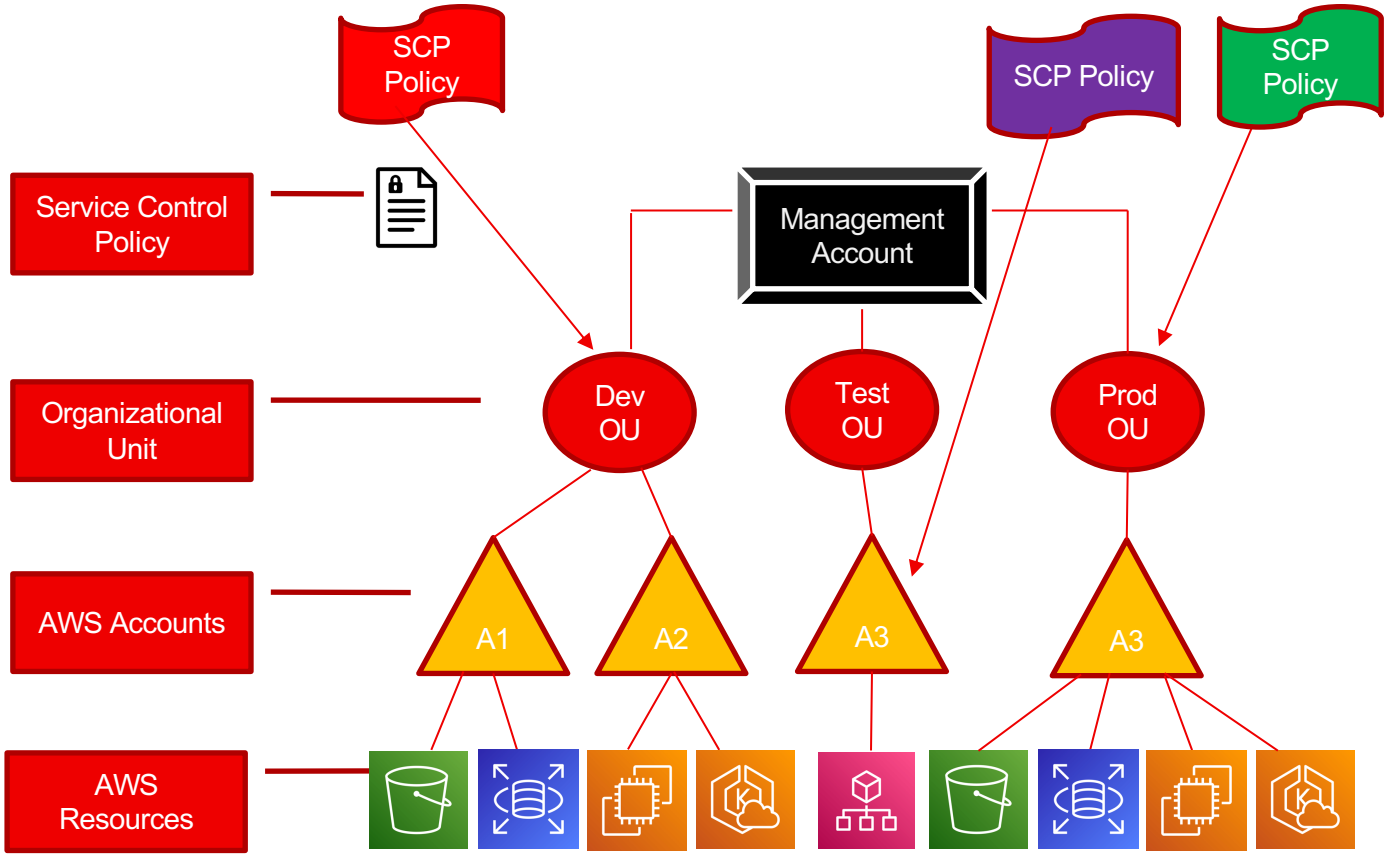
Service Quotas

- Each AWS account has default quotas, formerly referred to as limits, for each AWS service.
- Most quotas are region-specific.
- Customers can request increases for some quotas, and other quotas cannot be increased.
- Service Quotas helps you manage your quotas for many AWS services.
- Customers can request quota increases from the Service Quotas console.


Service Quotas
Quota request template
Add quota
Add quota
Quota
If the template associated with the organization is enabled, these quota updates will be requested when you create a new account in your organization.
Region
Select an AWS Region
Service
Select a service
Quota
Select a quota
Desired quota value
Enter a value
Cancel
Add

AWS Organizations

AWS Organizations



Containers

Containers at AWS

Elastic Container Service

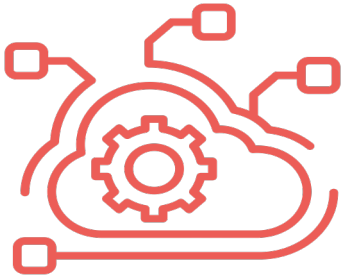
Run Docker containers:
Applications, Micro-services.

Amazon Elastic Kubernetes Service

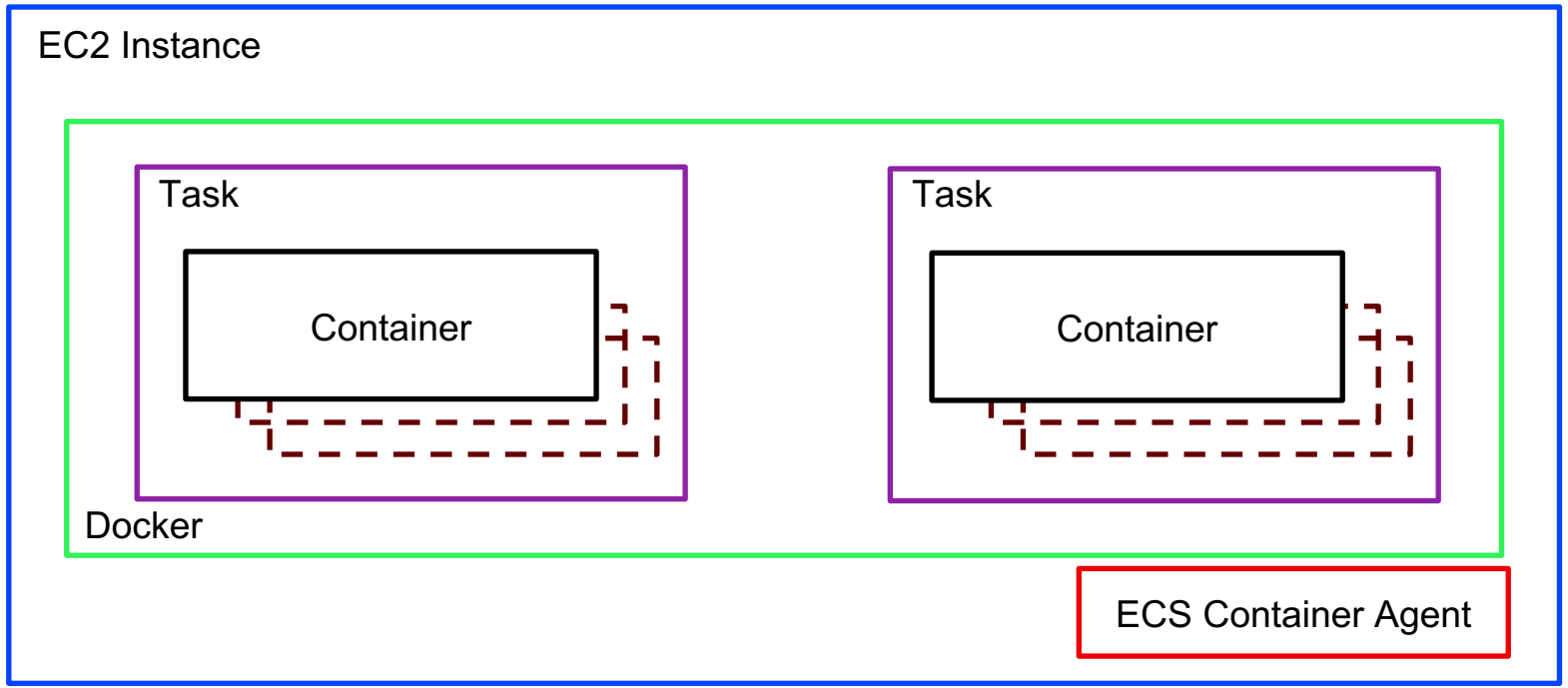
Manage containers with Kubernetes.

AWS Fargate

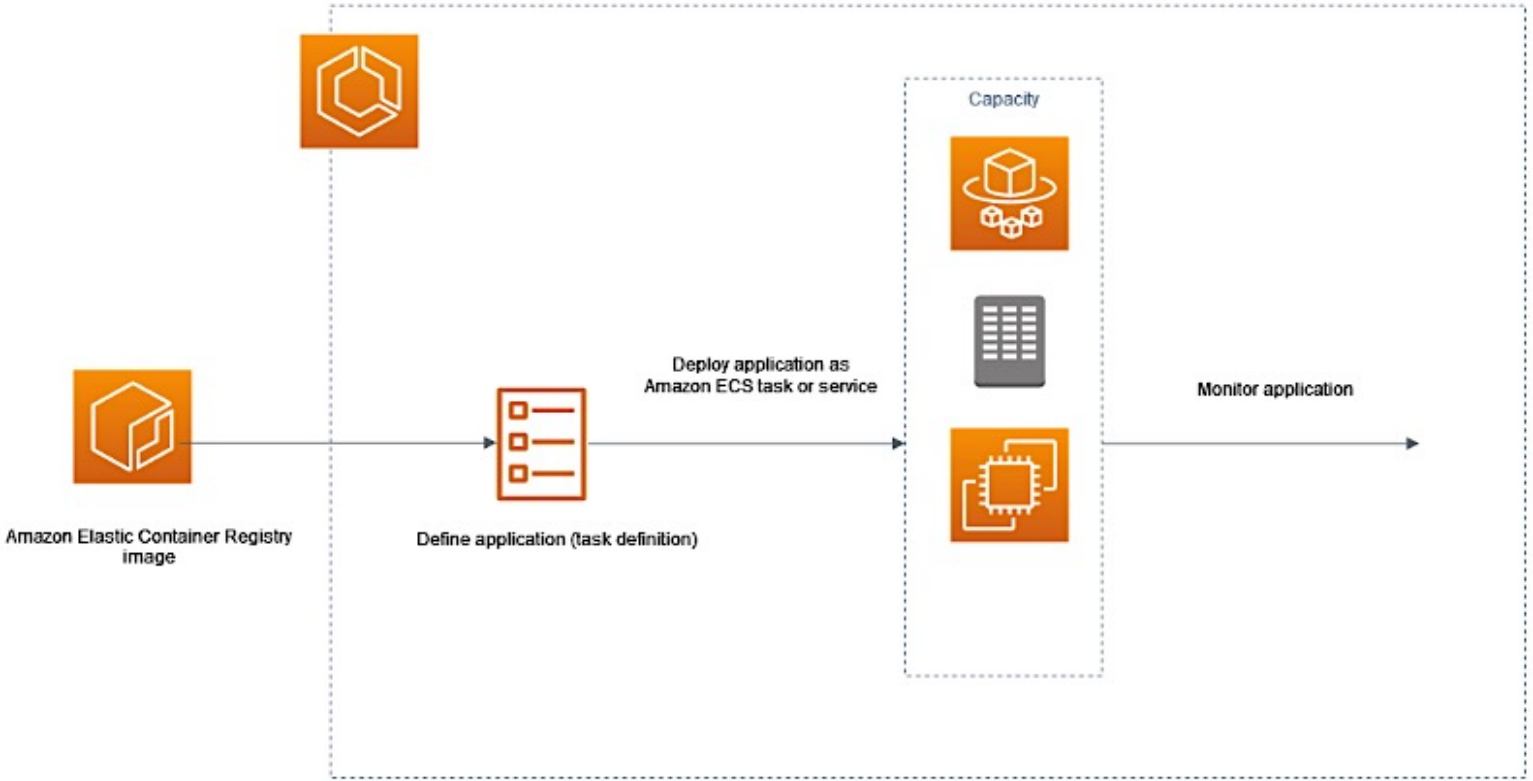
Management service for Docker or Kubernetes containers.



Task Definition Logic



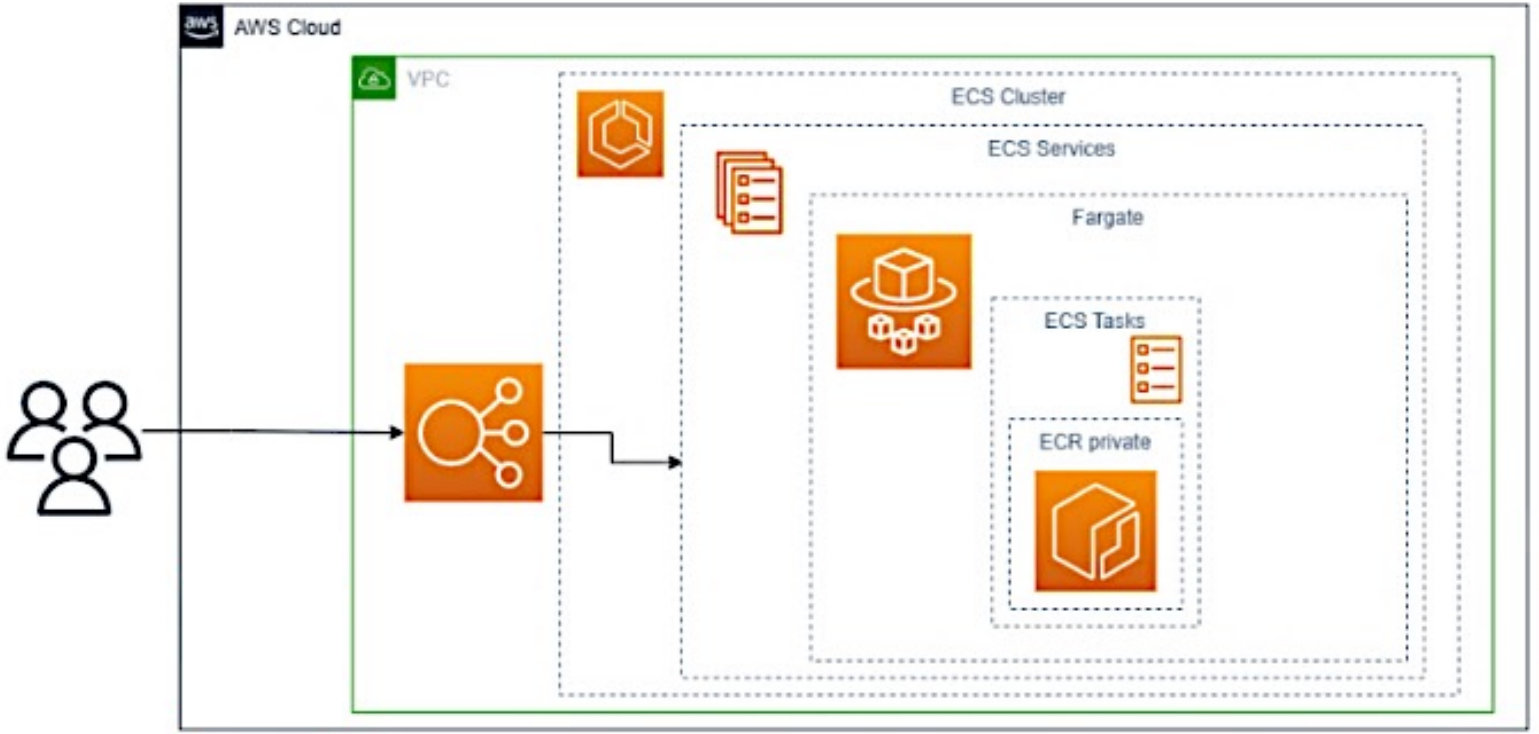
Amazon ECS (Elastic Container Service)



Container Orchestration with Fargate

- Run your applications in containers and manage EC2 instances and clusters with Fargate.
- Specify the container image, memory, and CPU requirements using an ECS task definition.
- Fargate schedules the containers' orchestration, management and placement throughout the cluster, providing a highly available operation environment.

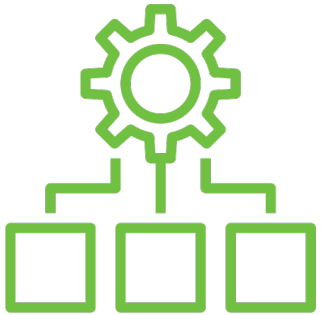
AWS Fargate


AWS Cloud
VPC
ECS Cluster
ECS Services
Fargate
ECS Tasks
ECR private

EC2 Instances

The Nitro System

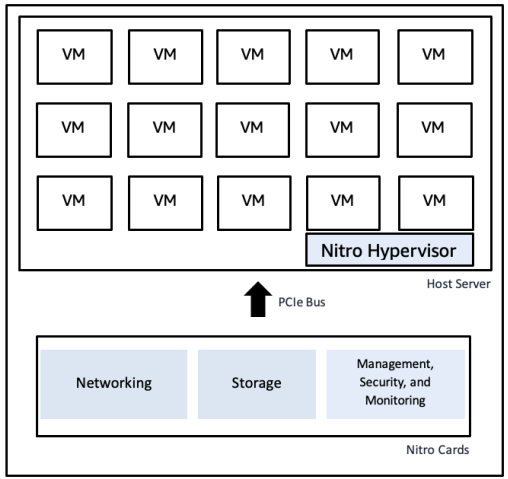
- The Nitro hypervisor is replacing XEN.
- Custom hardware chipsets now perform hypervisor tasks.
- Networking, storage, and encryption duties offloaded.
- The C5 instance was the first to be hosted by the Nitro hypervisor.



The Nitro System

The Nitro System enables the virtualization system to run nearly all the functionality on the Nitro cards rather than on the host's system mainboard.

- The Nitro Controller - H/W root of trust.
- Nitro Cards for I/O process (VPC, EBS, Local NVME storage PCI cards)
- Nitro Security Chip - Secure boot.

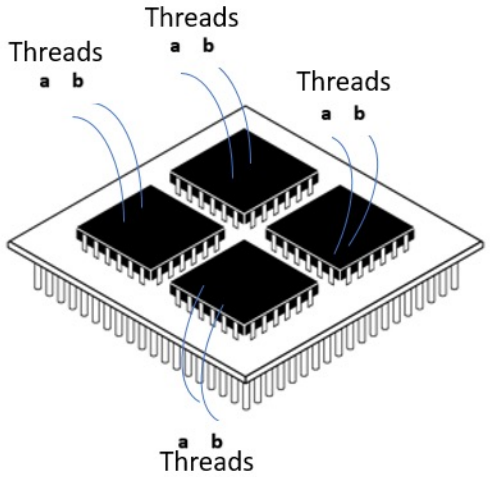


EC2 Instance FYI

- EC2 instances are members of compute families.
- The first letter of each instance's name is the instance family, which describes the resources allocated to it.
- EC2 resources (vCPUs, memory, storage, network bandwidth) are assigned to your account and are never shared with other AWS customers.

What’s a vCPU?

- AWS defines the amount of CPU power assigned to each instance as a virtual CPU (vCPU).
- vCPUs are software cores that “thread” to the physical CPU core on the host bare metal server.
- Hyper-threading associates two virtual threads to each physical core, working in a multitasking operating mode.



4 Core CPU / 8 Threads

Compute-optimized

- Compute-optimized instances are designed for batch processing workloads, media transcoding, high-performance applications, or Web servers.
- The C5 architecture takes advantage of the Nitro system with enhanced networking.

Size	Storage	AMI
Maximum 72 vCPUs, 144 GB of memory, enhanced networking up to 25 Gbps	EBS optimized storage with dedicated bandwidth up to 4,000 Mbps	64-bit HVM AMIs that include drivers for enhanced networking and NVMe storage

Memory Optimized

- Workloads that need to process vast data sets hosted in memory

Instance Type	Hypervisor	Maximum Size	Storage	AMI
r5	Nitro	96 vCPUs, 769 GiG of memory, enhanced networking speeds up to 25 Gbps	EBS optimized storage with dedicated EBS bandwidth up to 4,000 Mbps	64-bit HVM AMIs that include drivers for enhanced networking and NVMe storage
r4	Nitro	16 vCPUs, 488 GiG of memory, enhanced networking speeds up to 25 Gbps	Local instance storage using NVMe SSD	
x1	Nitro	128 vCPUs, 1952 GiG of memory, enhanced networking speeds up to 25 Gbps		CPU configuration of both the C-state and P-state registers is possible on the x8, x16, and x32xlarge models.
x1e	Nitro	128 vCPUs, 3904 GiG of memory, enhanced networking speeds up to 10 Gbps	14,000 Mbps of EBS optimized storage bandwidth	

Storage Optimized

- Storage-optimized instances are designed for workloads that require local NVME storage for large data sets.

Instance Type	Hypervisor	Maximum Size	Processor	Features
h1	Xen	64 vCPUs, 256 GiB memory, 4 x 200 GiB of instance storage, up to 25 Gbps enhanced networking	Intel Broadwell E5-2686V4	1.15 Gb/s read-write with 2 MiB block size
d2	Xen	36 vCPUs, 244 GiB memory, 24 x 2048 GiB of instance storage,10 Gbps enhanced networking	Intel Xeon E5-2670v2	EBS optimized
i3	Nitro	36 vCPUs, 244 GiB memory, 8 X 1900 GiB of instance storage, up to 25 Gbps enhanced networking	Intel Broadwell E5-2686V4	1.4 million Write IOPS

t instances

t instances store CPU credits that are created when the CPU is idle.

When performance is needed, the banked CPU credits are used.

Credits are used when your application bursts above the instance baseline assigned.

Burst Credits

At launch, there are enough CPU credits allocated to boot the operating system and load the application.

A single CPU credit has the performance of one full CPU core running at 100 % for one minute.

The larger the t instance, the more CPU credits are earned up to a defined maximum value.

Earned credits expire after 24 hours.

EBS-optimized instances

- EBS–optimized instances have an optimized configuration stack with dedicated capacity for Amazon RDS.
- EBS–optimized instances utilize dedicated bandwidth paths to Amazon EBS storage.

NVMe

- Non-volatile memory express SSD instance flash storage volumes.
- EBS volumes are exposed as NVMe block devices on Nitro-based instances.

Instance Name	vCPUs	RAM	Local Storage	EBS-Optimized Bandwidth	Network Bandwidth
m5d.large	2	8 GiB	1 x 75 GB NVMe SSD	Up to 2.120 Gbps	Up to 10 Gbps
m5d.xlarge	4	16 GiB	1 x 150 GB NVMe SSD	Up to 2.120 Gbps	Up to 10 Gbps
m5d.2xlarge	8	32 GiB	1 x 300 GB NVMe SSD	Up to 2.120 Gbps	Up to 10 Gbps
m5d.4xlarge	16	64 GiB	2 x 300 GB NVMe SSD	2.210 Gbps	Up to 10 Gbps
m5d.12xlarge	48	192 GiB	2 x 900 GB NVMe SSD	5.0 Gbps	10 Gbps
m5d.24xlarge	96	384 GiB	4 x 900 GB NVMe SSD	10.0 Gbps	25 Gbps

Dedicated Hosts

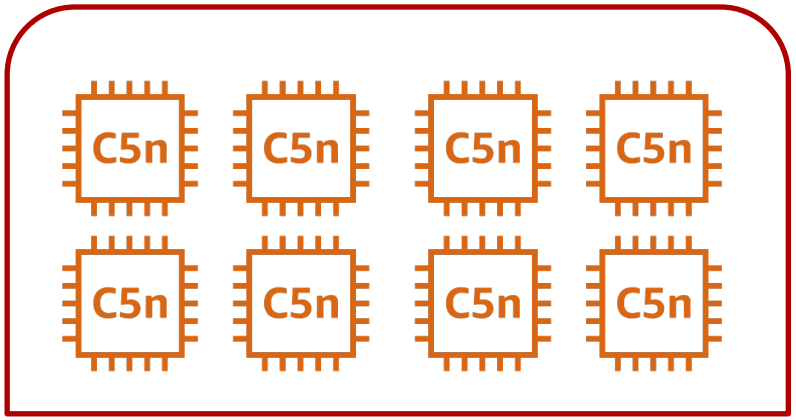
- It supports “affinity”: specifying which Dedicated Host an instance will run on after it has been stopped and restarted.
- Compliance with specific regulatory requirements that don’t permit multi-tenant virtualization.

Dedicated Instances

- Dedicated Instances are **physically isolated at the host hardware level** from other dedicated instances and instances in other AWS accounts.
- When you launch a Dedicated Instance backed by EBS volumes, the EBS volume itself doesn't run on single-tenant hardware.

Cluster Placement Group

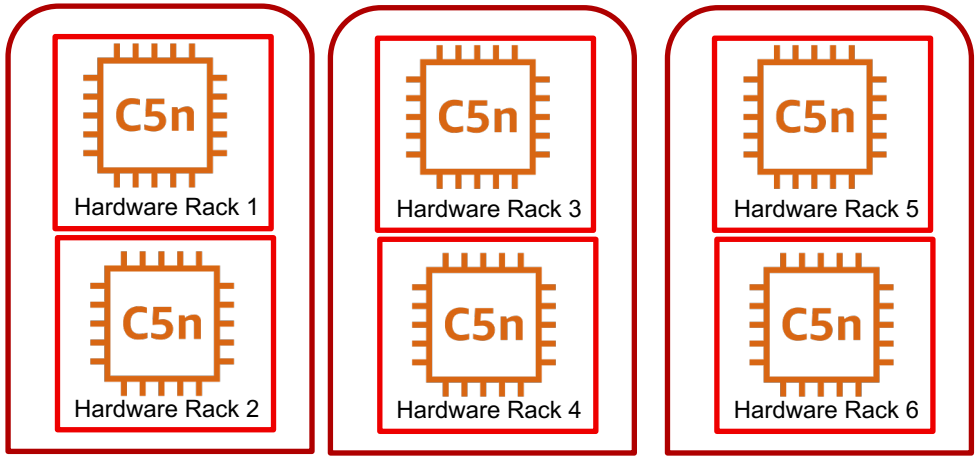
- A cluster design has the lowest latency possible within a single AZ.
- EC2 instances are placed on the same underlying hardware, which minimizes the network latency between them.
- Use case: Big Data or HPC.



Same Hardware Rack

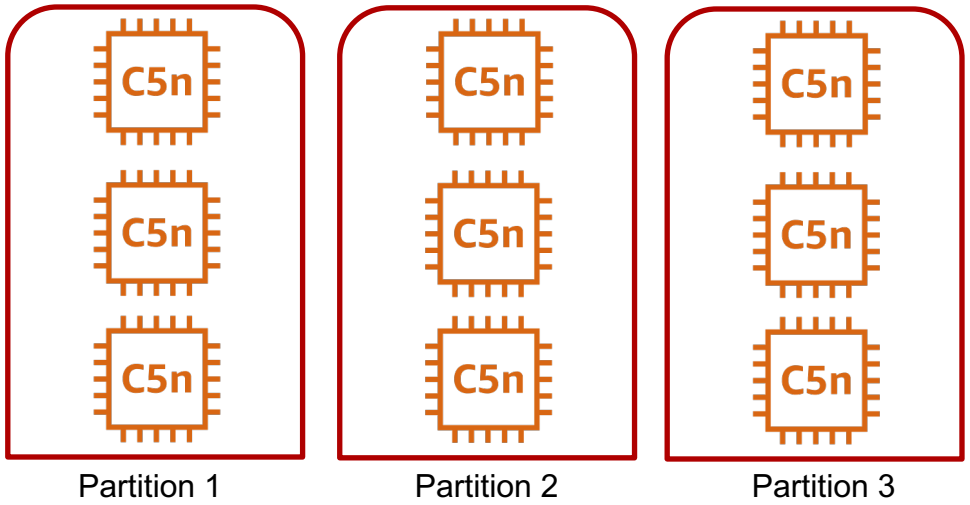
Spread Placement Group

- Spread your instances across distinct underlying hardware to reduce the risk of simultaneous failures.
- Racks have separate power sources.

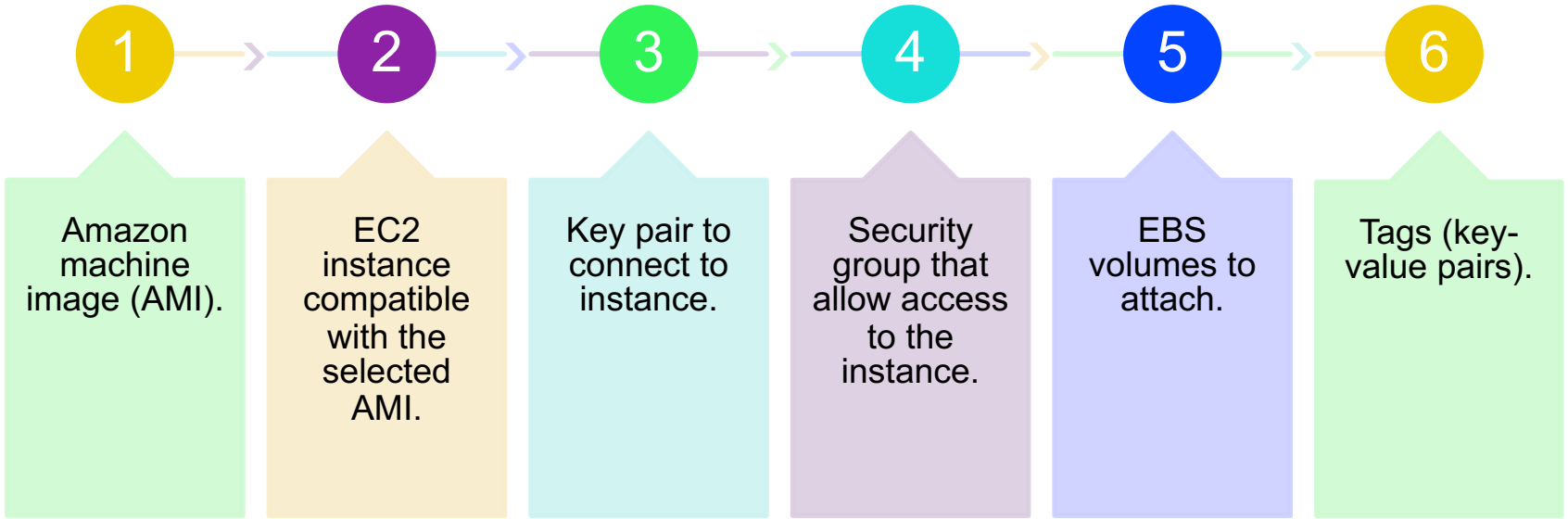


Partition Placement Group

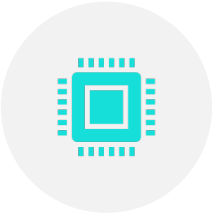
- Create multiple “partitions” of nodes with separate racks.
- No two partitions within a placement group share the same rack.
- Distributed, fault-tolerant, and low-latency workloads.



Launch Templates



EC2 Image Builder



Automate the creation, maintenance, and deployment of Linux or Windows images.

Images can be generated based on source AMIs.

Define collections of security settings that you can edit, update, and use to harden your images.

AWS-provided tests and customer tests can be run before image finalization.

EC2 Pricing Considerations

Server runtime	Instance type	Pricing model
Number of instances	Load Balancing	Auto Scaling
Detailed monitoring enabled	Elastic IP addresses	Operating system

EC2 Instances: Pricing Options

On-Demand

Pay by per hour/ second

Short-term, unpredictable workloads

Reserved Instances

Discount for 1 - 3-year commitment

Applications with consistent usage

Spot Requests

Spare AWS capacity > 90% discount

Applications with flexible start and end times

Savings Plans

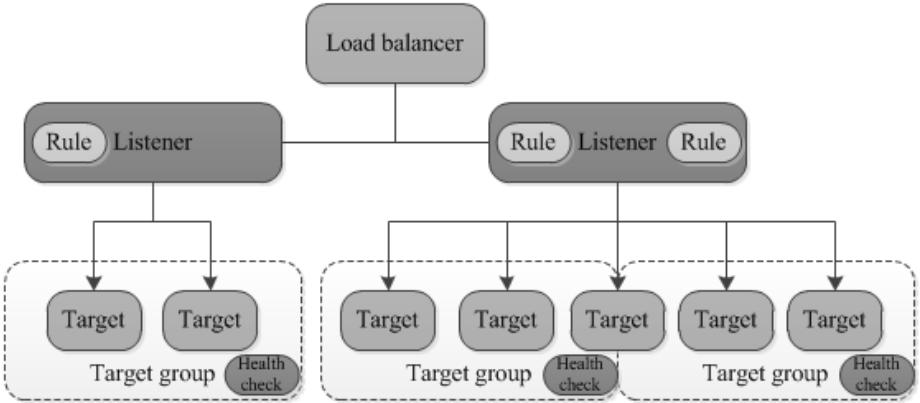
- Savings Plans provide savings of up to 72% on your AWS compute usage.
    - Compute: Applies to all Amazon EC2 instances (OS, tenancy or Region), Fargate and Lambda usage.
    - EC2 Instances.
    - Amazon Sage Maker.

- Commit to using a specific amount of compute power (measured in $/hour) for one or three years.

Load Balancing and Auto Scaling

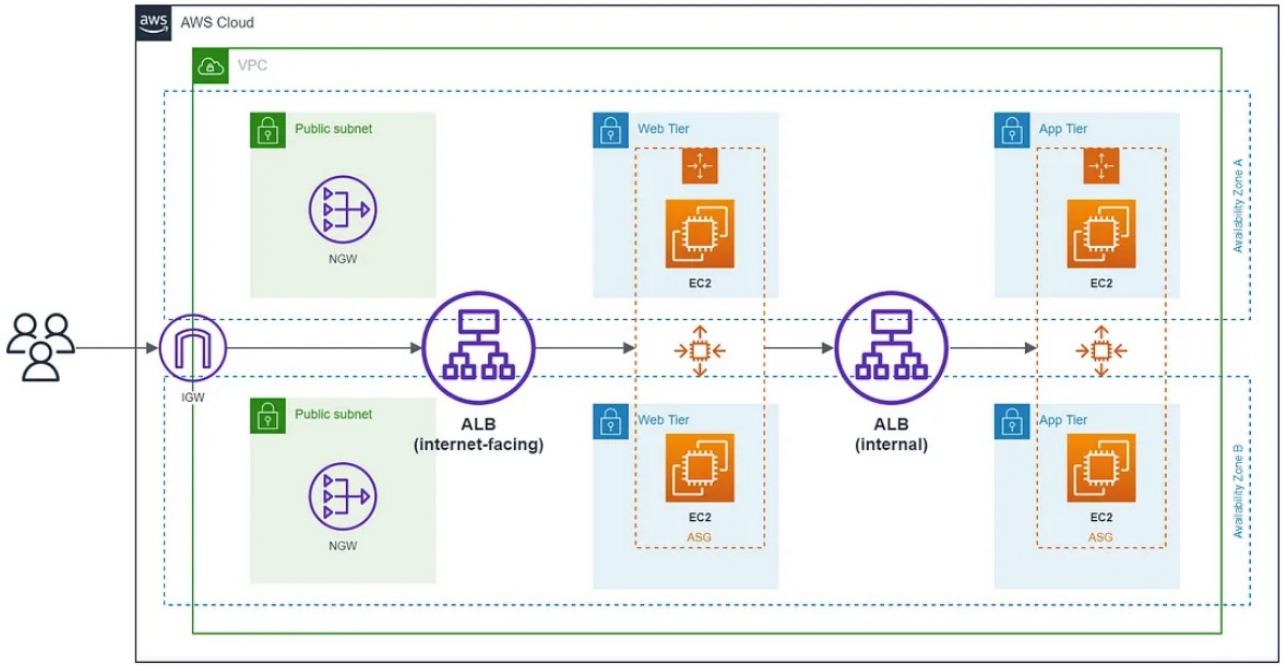
Elastic Load Balancing Service

- Distribute incoming traffic across multiple EC2 instances or containers across one or multiple availability zones.
- Network load balancer: Target groups support the TCP, UDP, TCP_UDP, and TLS protocols.
- Application load balancer: Target groups support the HTTP and HTTPS protocols.



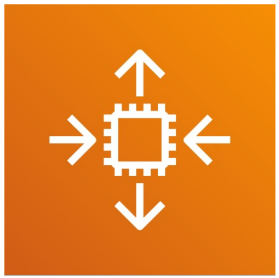
ELB Setup

- Internal / Extern
- VPC / Subnets
- Traffic (80 / 443
- Health Checks



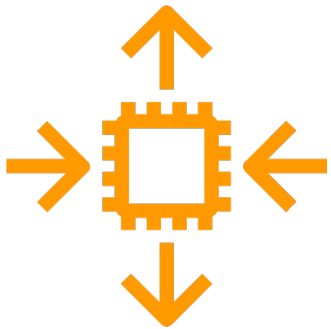
Auto Scale Concepts

- Create collections of EC2 instances with Auto-Scaling groups.
- Auto-scaling provides automatic horizontal scaling (scale-out) for your EC2 instances.
- ASGs span multiple AZs within the same AWS region and integrate with ELB and CloudWatch.



Auto Scaling Components

- Auto Scaling Groups: A managed group of EC2 instances.
- Launch Templates: A configuration template for adding EC2 instances.
- Scaling options are auto–scale group scaling options.
- CloudWatch metrics and alarms initiate scaling.



Scaling Policy Options

Scaling Policy	Details
Target tracking	Increase or decrease the capacity of the ASG based on a target value for specific cloud watch metric.
Step	Increase or decrease the capacity based on a set of scaling percentages.
Simple	Increase or decrease based on a single scaling adjustment.

Storage Options

Data Storage Options at AWS

STORAGE TYPE	DETAILS	AWS SERVICES
Persistent data Storage	Data is persistent and durable.	S3, S3 Glacier, EBS, EFS, FSx for Windows File Server.
Transient data Storage	Data is temporarily stored before being consumed.	SQS, SNS
Ephemeral data Storage	Data records are lost when system(s) are stopped.	EC2 Instance Storage, ElastiCache for Redis.

Amazon Elastic Block Store

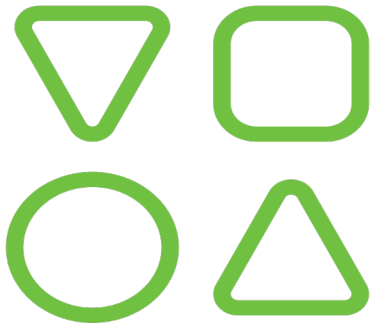
Amazon EBS

- High-performance block storage service designed for use with the Amazon Elastic Compute Cloud (EC2).
- Virtual hard drives attached to EC2 instances.
- Data volumes are persistent; no data loss when the associated instance is stopped or terminated.

Volume

Amazon S3 Objects

- S3 can store any data type.
  - Up to 5 TB max for a single object.
- Each object has a unique key:
  - Key = filename
  - Must be unique within each bucket.
  - Metadata describes the data.



- System metadata - AWS date, size, storage class.
- User metadata - tags specified at the time the object is created.

S3 Architecture

- S3 buckets are stored in a single region.
- An S3 bucket name is a DNS namespace (Route 53); names must be globally unique.

  https://s3-***eu-east-1***.amazonaws.com/***<bucketname>***

- S3 automatically scales to high request rates:
  - 3,500 PUT/POST/DELETE and 5,500 GET requests per second.
  - For faster read requests, use CloudFront edge locations.

S3 Durability

Standard Storage Class

- 11 9’s durability
- 4 9’s availability
- No minimum storage time.

Standard IA Storage Class

- 11 9’s durability
- 4 9’s durability
- Minimum storage time of 30 days

S3 Storage Classes

S3 Standard - No minimum storage time.

S3 Intelligent-tiering - Move to the most cost-effective tier after 30 days.

S3 Standard-IA - Min 30 days.

S3 One Zone-IA - Store in one AZ, less resilience - min 30 days.

S3 Glacier Deep Archive - Long-term retention min 180 days.

S3 on Outposts.

Same or Cross-Region Replication (CCR)

- Asynchronous replication from source bucket in AWS region to destination bucket in the same or another AWS region.
- Versioning must be enabled for both the source and destination buckets.
- Separate S3 lifecycle rules can be configured on source and destination buckets.
- Help move data closer to end-users.
- Compliance / additional durability.

S3 Encryption

- Amazon S3 and S3 Glacier automatically encrypt all object uploads to all buckets.
- Amazon S3 supports three additional encryption options.

SSE – C

The client supplies the encryption keys.

SSE – KMS

Key Management Service supplies the encryption keys.

DSSE + KMS

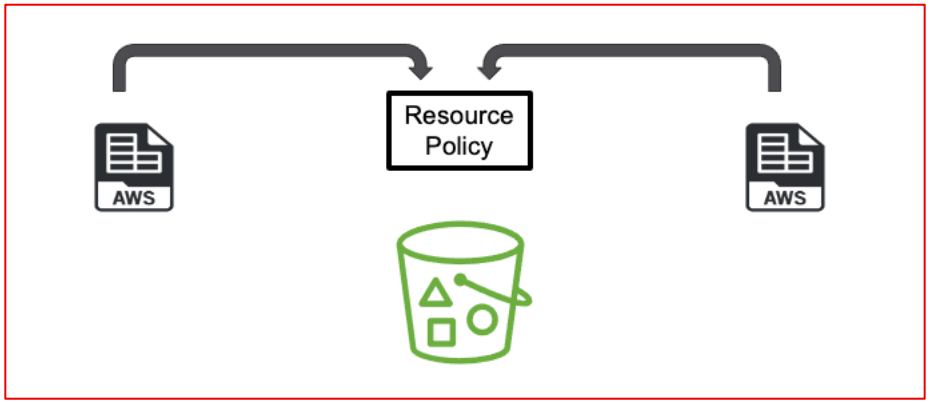
Enable dual-layer - SSE (Default SSE +KMS encryption)

Access Control Options

- Only the owner of the S3 bucket has access by default.
- IAM policies.
- Bucket policies – Resource-based policies.
- Query string authentication (Time-limited URL).
- Access auditing can be configured by configuring access log records for all requests made.

Bucket Policies

- Bucket policies can be used to grant other AWS accounts or IAM users permission to access the bucket and objects.
- S3 buckets can be accessed by different AWS accounts using a bucket policy.



S3 Glacier Storage

- Archives from 100MB up to 40TB can be uploaded to S3 Glacier.
- Archives are assigned a unique ID and are stored in vaults.
- Each AWS account can have up to 1,000 vaults.
- Enable compliance controls per vault using a vault lock policy (WORM).
- Retrieval policies control the frequency of access.

Amazon S3 Glacier

- Designed for archival storage and long-term backup at a very low cost.
- Ideal for data that is infrequently accessed.
- Retrieval times from seconds to hours are available.
- Upload data to S3 Glacier directly using the AWS Management Console, AWS SDKs, or the AWS CLI.
- Supports multipart upload, which allows the uploading of large archives in parts.

S3 Glacier Storage Classes and Retrieval Options

Instant Retrieval

- Retrieval in milliseconds.

Flexible Retrieval

- Retrieval time of minutes to hours.

Deep Archive

- Retrieval time of hours.

EFS Performance Modes

- **Automatic Scaling**: Storage capacity scales up or down automatically as files are added or removed, without manual intervention or provisioning.
- **Performance Modes:**
  - **General Purpose**: Recommended starting mode of operation.
  - **Max I/O Performance**: Optimized for high aggregate throughput and operations levels.
  - Suitable for large-scale data processing and media processing workloads.

EFS Throughput Modes

- **Bursting Throughput**: Automatically scales with the size of the file system. Suitable for workloads with sporadic read/write operations.
- **Provisioned Throughput**: Allows you to specify the throughput independent of the amount of data stored. It is beneficial for applications that require a higher sustained level of throughput.
- **Elastic Throughput:** If the application has low baseline activity but has unexpected spikes, then select elastic throughput mode.

EFS Storage Classes

- **Standard**: For frequently accessed files.
- **Infrequent Access (IA)**: Lower-cost option for files accessed less frequently. EFS automatically moves files to and from the IA storage class based on access patterns.
- **One Zone Infrequent Access:** The same operation as the Infrequent Access storage class but stored in a single availability zone.

Amazon FSx for Windows File Server

- Fully managed native Microsoft Windows file system.
- Automatically grows and shrinks as files are added and removed; no provisioning storage in advance.
- Supports SMB protocol and Windows NTFS.
- Integrates with Microsoft Active Directory deployments.

File system

FSx for Windows File Server Features

- **Native Windows File System**: Supports the SMB protocol and Windows NTFS and is integrated with Microsoft Active Directory.
- **Data Deduplication**: Optimize storage utilization by removing duplicate data blocks.
- **Highly Available and Durable**: Data is automatically replicated within multiple Availability Zones. A single AZ can also be chosen.
- **Backup and Restore**: Supports automatic daily backups and user-initiated backups.
- **Windows Authentication:** Choose a self-managed or managed Active Directory deployment for user authentication and file system access control.

FSx for Windows File Server Performance Options

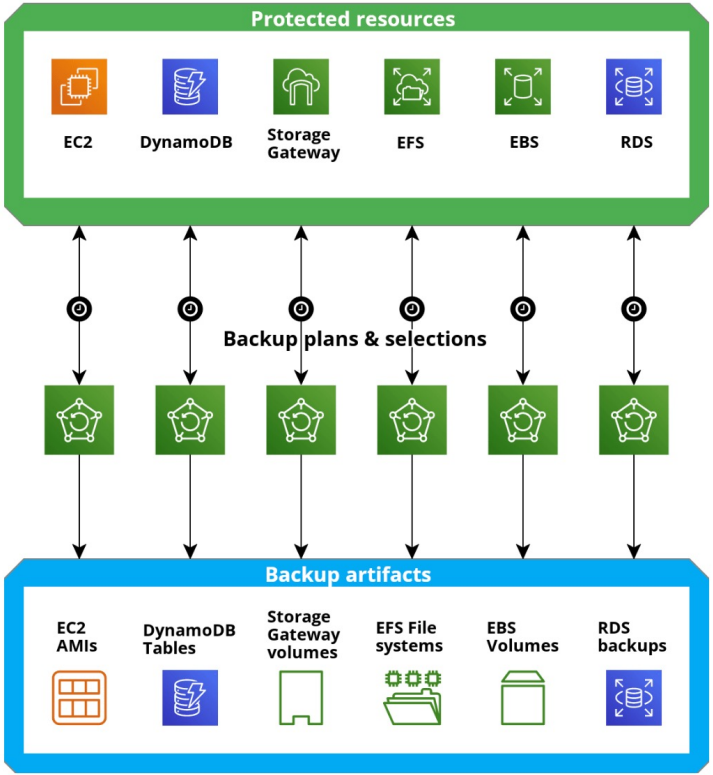
Storage Types

- **SSD Storage**: Offers high IOPS and throughput, suitable for performance-sensitive databases and media processing.
- **HDD Storage**: More cost-effective, suitable for applications such as file sharing

Throughput Capacity

- **Configurable Throughput**: Select the throughput capacity of the file system in megabytes per second (MB/s).
- **Automatic Throughput Scaling**: Some configurations support automatic scaling of throughput capacity based on your workload's needs.

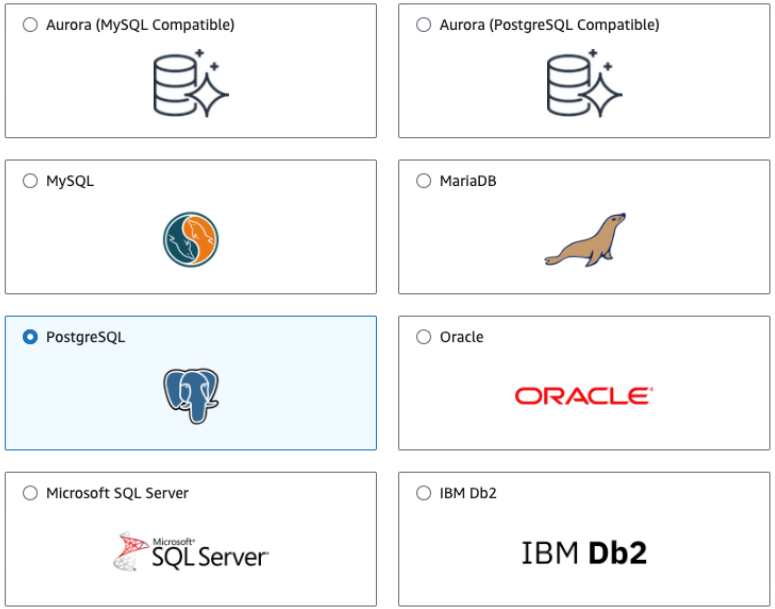
AWS Backup


Protected resources
EC2
DynamoDB
Storage Gateway
EFS
EBS
RDS
Backup plans & selections
Backup artifacts
EC2 AMIs
DynamoDB Tables
Storage Gateway volumes
EFS File systems
EBS Volumes
RDS backups

RDS Databases

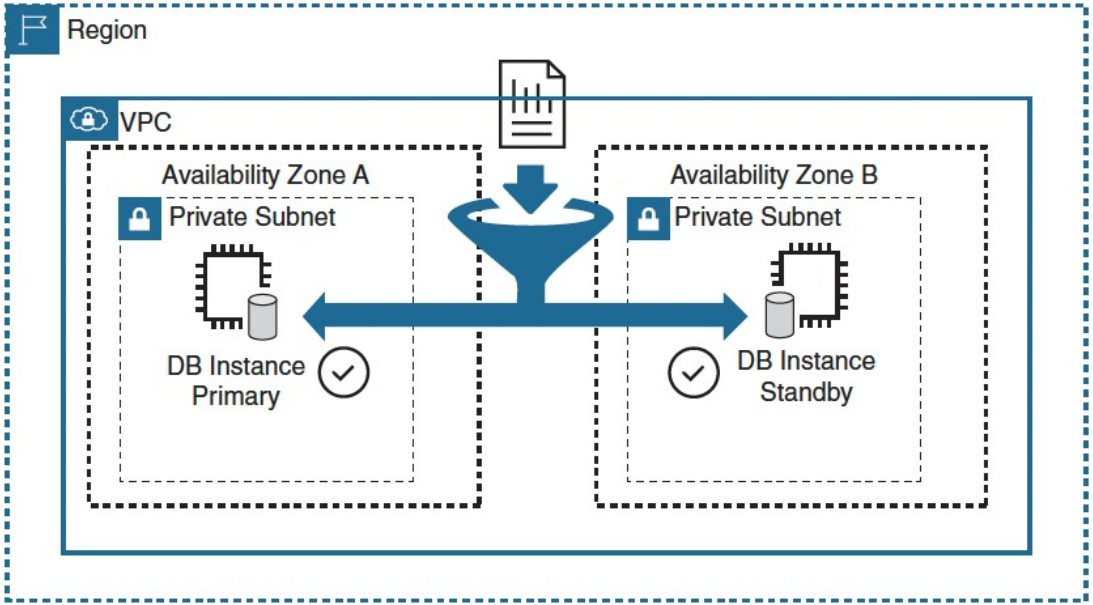
RDS Database Engines

- RDS uses AWS-managed database instances (Primary DB, One or two standby DBs) across single or multi-AZs.



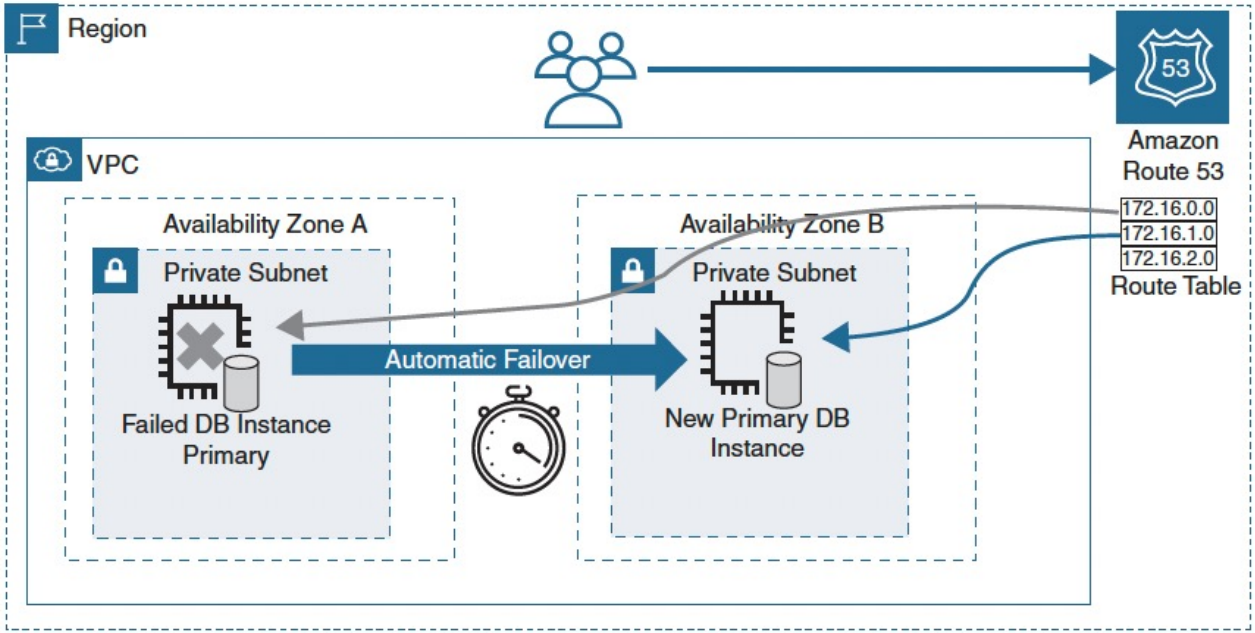
RDS Updates are Synchronous

- Data updates to primary and standby instances are carried out at the same time.



RDS Failover

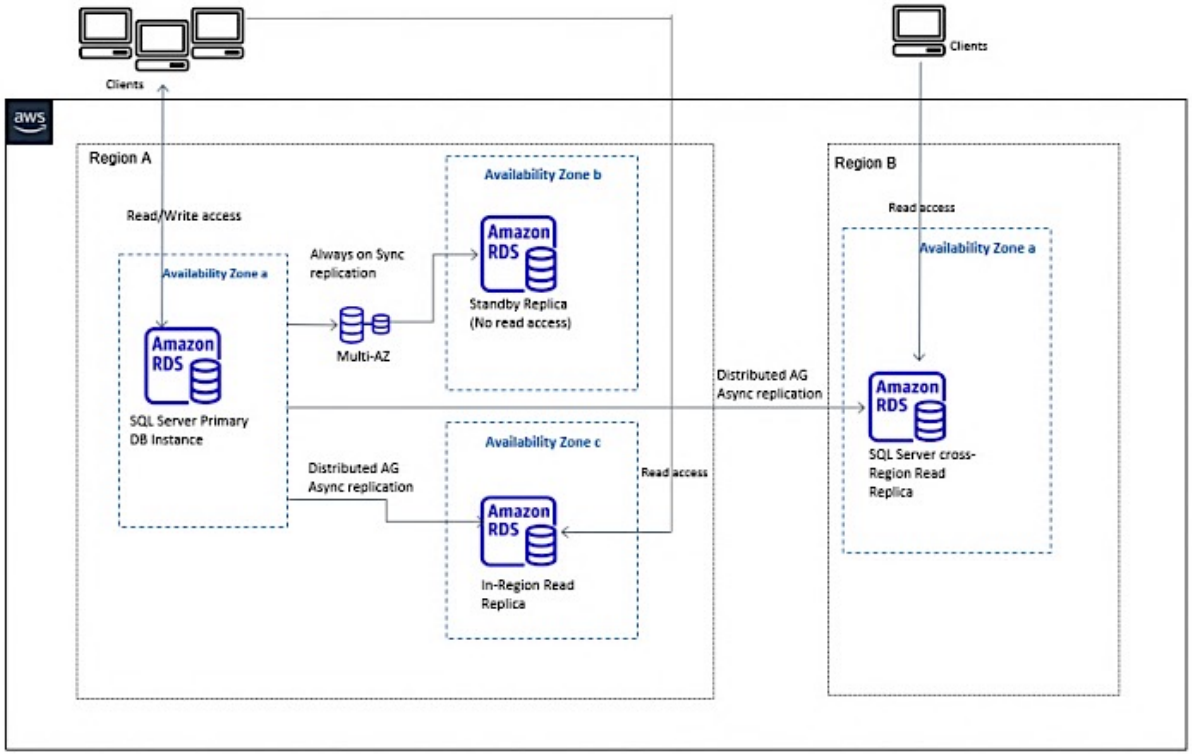
- The loss of the primary AZ or primary DB instance triggers automatic failover.



DB Read Replicas

- RDS Read Replicas
  - Amazon RDS Read Replicas provide additional performance and durability for Amazon RDS database instances.
  - Serve read traffic from multiple copies of your data, increasing aggregate read throughput.

Amazon RDS Read Replicas



What We Covered

- Shared Responsibility Model
- Account Security and IAM
- Regions and Availability Zones
- VPC Architecture
- AWS Organizations
- Containers and EC2 Instances
- Load Balancing and Auto Scaling
- Storage Options
- RDS Databases


O'REILLY®