Elastic Cloud Compute – EC2

EC2 instances	Virtual computing environments
Amazon Machine Images (AMIs)	Preconfigured templates for EC2 instances
Instance types	Various configurations of CPU, memory, storage, and networking capacity for your instances
key pairs	Secure login information for your instances using key pairs (public-private keys where private is kept by user)
Instance store volumes	Storage volumes for temporary data that’s deleted when you stop or terminate your instance,
Elastic Block Store (EBS)	Persistent storage volumes for data
Regions and Availability Zones	Multiple physical locations for your resources, such as instances and EBS volumes
Security Groups	A firewall to specify the protocols, ports, and source IP ranges that can reach your instances
Elastic IP addresses	Static IP addresses,

Elastic Cloud Compute – EC2 (cont)

tags	can be created and assigned to EC2 resources
Virtual private clouds (VPCs)	Virtual networks that are logically isolated from the rest of the AWS cloud, and can optionally connect to on-premises network

EC2 Monitoring

CloudWatch provides monitoring for EC2 instances

Status monitoring helps quickly determine whether EC2 has detected any problems that might prevent instances from running applications.

Status monitoring includes. System Status checks – indicate issues with the underlying hardware. Instance Status checks – indicate issues with the underlying instance.

Elastic Load Balancer

Managed load balancing service and scales automatically

distributes incoming application traffic across multiple EC2 instances

is distributed system that is fault tolerant and actively monitored by AWS scales it as per the demand

are engineered to not be a single point of failure

supports routing traffic to instances in multiple AZs in the same region

performs Health Checks to route traffic only to the healthy instances

support Listeners with HTTP, HTTPS, SSL, TCP protocols

Elastic Load Balancer (cont)

has an associated IPv4 and dual stack DNS name

can offload the work of encryption and decryption (SSL termination) so that the EC2 instances can focus on their main work

supports Cross Zone load balancing to help route traffic evenly across all EC2 instances regardless of the AZs they reside in

to help identify the IP address of a client

supports Proxy Protocol header for TCP/SSL connections

supports X-Forward headers for HTTP/HTTPS connections

supports Stick Sessions (session affinity) to bind a user’s session to a specific application instance,

supports Connection draining to help complete the in-flight requests in case an instance is deregistered

For High Availability, it is recommended to attach one subnet per AZ for at least two AZs, even if the instances are in a single subnet.

supports Static/Elastic IP (NLB only)

VPC now supports IPV6.

HTTPS listener does not support Client Side Certificate

For SSL termination at backend instances or support for Client Side Certificate use TCP for connections from the client to the ELB, use the SSL protocol for connections from the ELB to the back-end application, and deploy certificates on the back-end instances handling requests

Uses Server Name Indication to supports multiple SSL certificates

Auto Scaling

ensures correct number of EC2 instances are always running to handle the load by scaling up or down automatically as demand changes

attempts to distribute instances evenly between the AZs that are enabled for the Auto Scaling group

performs checks either using EC2 status checks or can use ELB health checks to determine the health of an instance and terminates the instance if unhealthy, to launch a new instance

can be scaled using manual scaling, scheduled scaling or demand based scaling

cooldown period helps ensure instances are not launched or terminated before the previous scaling activity takes effect to allow the newly launched instances to start handling traffic and reduce load

cannot span multiple regions.

Amazon Machine Image – AMI

Template from which EC2 instances can be launched quickly

Does NOT span across regions, and needs to be copied

Can be shared with other specific AWS accounts or made public

Instance Types

T2 instances are Burstable Performance Instances that provide a baseline level of CPU performance with the ability to burst above the baseline.

T2 instances accumulate CPU Credits when they are idle, and consume CPU Credits when they are active.

Instance Types (cont)

T2 Unlimited Instances can sustain high CPU performance for as long as a workload needs it at an additional cost.

R for applications needing more RAM or Memory

C for applications needing more Compute

M for applications needing more Medium or Moderate performance on both Memory and CPU

I for applications needing more IOPS

G for applications needing more GPU

Placement Group

Cluster Placement Group:
1.provide low latency, High-Performance Computing via 10Gbps network
2.is a logical grouping on instances within a Single AZ
3.don’t span availability zones, can span multiple subnets but subnets must be in the same AZ
can span across peered VPCs for the same Availability Zones
4.An existing instance can be moved to a placement group, or moved from one placement group to another, or removed from a placement group, given it is in the stopped state.
5.for capacity errors, stop and start the instances in the placement group
6.use homogenous instance types which support enhanced networking and launch all the instances at once
Spread Placement Groups:
1.is a group of instances that are each placed on distinct underlying hardware i.e. each instance on a distinct rack across AZ

Placement Group (cont)

2.recommended for applications that have a small number of critical instances that should be kept separate from each other.
3.reduces the risk of simultaneous failures that might occur when instances share the same underlying hardware.
Partition Placement Groups:
1.is a group of instances spread across partitions i.e. group of instances spread across racks across AZs
2.reduces the likelihood of correlated hardware failures for the application.
3.can be used to spread deployment of large distributed and replicated workloads, such as HDFS, HBase, and Cassandra, across distinct hardware

Application Load Balancer

supports HTTP and HTTPS (Secure HTTP) protocols

supports HTTP/2, which is enabled natively. Clients that support HTTP/2 can connect over TLS

supports WebSockets and Secure WebSockets natively

supports Request tracing, by default.

supports containerized applications. Using Dynamic port mapping, ECS can select an unused port when scheduling a task and register the task with a target group using this port.

supports Sticky Sessions (Session Affinity) using load balancer generated cookies, to route requests from the same client to the same target

supports SSL termination, to decrypt the request on ALB before sending it to the underlying targets.

supports layer 7 specific features like X-Forwarded-For headers to help determine the actual client IP, port and protocol

Application Load Balancer (cont)

automatically scales its request handling capacity in response to incoming application traffic.

supports hybrid load balancing, to route traffic to instances in VPC and an on-premises location

provides High Availability, by allowing more than one AZ to be specified

integrates with ACM to provision and bind a SSL/TLS certificate to the load balancer thereby making the entire SSL offload process very easy

supports multiple certificates for the same domain to a secure listener

supports IPv6 addressing, for an Internet facing load balancer

supports Cross-zone load balancing, and cannot be disabled.

supports Security Groups to control the traffic allowed to and from the load balancer.

provides Access Logs, to record all requests sent the load balancer, and store the logs in S3 for later analysis in compressed format

provides Delete Protection, to prevent the ALB from accidental deletion

supports Connection Idle Timeout – ALB maintains two connections for each request one with the Client (front end) and one with the target instance (back end). If no data has been sent or received by the time that the idle timeout period elapses, ALB closes the front-end connection

integrates with CloudWatch to provide metrics such as request counts, error counts, error types, and request latency

Application Load Balancer (cont)

integrates with AWS WAF, a web application firewall that helps protect web applications from attacks by allowing rules configuration based on IP addresses, HTTP headers, and custom URI strings

integrates with CloudTrail to receive a history of ALB API calls made on the AWS account

Back-end server authentication is NOT supported
Does not provide Static, Elastic IP addresses

Instance Purchasing Option

On-Demand Instances:
1.pay for instances and compute capacity that you use by the hour
2.no long-term commitments or up-front payments
Reserved Instances:
1.provides lower hourly running costs by providing a billing discount
2.capacity reservation is applied to instances
3.suited if consistent, heavy, predictable usage
4.provides benefits with Consolidate Billing
5.can be modified to switch Availability Zones or the instance size within the same instance type, given the instance size footprint (Normalization factor) remains the same
6.pay for the entire term regardless of the usage
7.is not a physical instance that is launched, but rather a billing discount applied to the use of On-Demand Instances
Scheduled Reserved Instances:
1.enable capacity reservations purchase that recurs on a daily, weekly, or monthly basis, with a specified start time and duration, for a one-year term.

Instance Purchasing Option (cont)

2.Charges are incurred for the time that the instances are scheduled, even if they are not used
3.good choice for workloads that do not run continuously, but do run on a regular schedule
Spot Instances:
1.cost-effective choice but does NOT guarantee availability
2.applications flexible in the timing when they can run and also able to handle interruption by storing the state externally
3.provides a two-minute warning if the instance is to be terminated to save any unsaved work
4.Spot blocks can also be launched with a required duration, which are not interrupted due to changes in the Spot price
5.Spot Fleet is a collection, or fleet, of Spot Instances, and optionally On-Demand Instances, which attempts to launch the number of Spot and On-Demand Instances to meet the specified target capacity
Dedicated Instances:
1.is a tenancy option that enables instances to run in VPC on hardware that's isolated, dedicated to a single customer
Dedicated Host:
1.is a physical server with EC2 instance capacity fully dedicated to your use
2.Light, Medium, and Heavy Utilization Reserved Instances are no longer available for purchase and were part of the Previous Generation AWS EC2 purchasing model

Enhanced Networking

results in higher bandwidth, higher packet per second (PPS) performance, lower latency, consistency, scalability, and lower jitter

supported using Single Root – I/O Virtualization (SR-IOV) only on supported instance types

is supported only with a VPC (not EC2 Classic), HVM virtualization type and available by default on Amazon AMI but can be installed on other AMIs as well

Network Load Balancer

handles volatile workloads and scale to millions of requests per second, without the need of pre-warming

offers extremely low latencies for latency-sensitive applications.

provides static IP/Elastic IP addresses for the load balancer

allows registering targets by IP address, including targets outside the VPC (on-premises) for the load balancer.

supports containerized applications. Using Dynamic port mapping, ECS can select an unused port when scheduling a task and register the task with a target group using this port.

monitors the health of its registered targets and routes the traffic only to healthy targets

enable cross-zone loading balancing only after creating the NLB

Network Load Balancer (cont)

preserves client side source IP allowing the back-end to see client IP address. Target groups can be created with target type as instance ID or IP address. If targets registered by instance ID, the source IP addresses of the clients are preserved and provided to the applications. If register targets registered by IP address, the source IP addresses are the private IP addresses of the load balancer nodes.

supports both network and application target health checks.

supports long-lived TCP connections ideal for WebSocket type of applications

supports Zonal Isolation, which is designed for application architectures in a single zone and can be enabled in a single AZ to support architectures that require zonal isolation

Does not support stick sessions

AWS Auto Scaling & ELB

Auto Scaling & ELB can be used for High Availability and Redundancy by spanning Auto Scaling groups across multiple AZs within a region and then setting up ELB to distribute incoming traffic across those AZs

With Auto Scaling, use ELB health check with the instances to ensure that traffic is routed only to the healthy instances