**EC2**

* + - Elastic Compute Cloud
    - Provides resizable compute capacity in the cloud
    - Designed to make web-scaling computing easier for developers
    - Generally limited to 20 instances/type(on-demand/spot/reserved)
    - EC2 Compute Units(ECU) provide the relative measure of the integer processing power of an Amazon EC2 instance
    - EC2 gives you full control at the OS layer
    - Key pairs are used to securely connect to EC2 instances
      * KP consists of a public key that AWS stores, and a private key file that you store
      * For Windows AMIs the private key file is required to obtain the password used to log into your instance
      * For Linux AMIs the private key file allows you to securely SSH into your instance
    - User Data
      * Information provided by user at launch in the form of a script
      * Max size is 16KB
      * Not encrypted
      * available at <http://169.254.169.254/latest/user-data>
      * On Linux User Data can be viewed 'curl <http://169.254.169.254/latest/user-data>
    - Metadata
      * Data about your instance that you can use to configure or manage the running instance
      * Instance metadata available at <http://169.254.169.254/latest/meta-data/>
        + Trailing / at the end of meta-data is required
      * Not encrypted
      * Instance Metadata Query tool allows you to query the instance metadata without having to type out the full URL or category names
    - Billing Types
      * On Demand
        + Pay for hours used with no commitment
        + No upfront cost
        + Ideal for autoscaling groups and unpredictable workloads
        + Good for dev/test
      * Spot
        + EC2 Spot instances let you take advantage of unused EC2 capacity in the cloud
        + Up to 90% discount compared to on-demand
        + You pay for the Spot price that's in effect for the current hour for the instances you launch
        + Spot Instances receive a two-minute interruption notice when these instances are about to be reclaimed by EC2
        + Instances are not interrupted because of higher competing bids
      * Reserved
        + Purchase usage of EC2 instances in advance
        + Capacity is reserved for a term of 1 or 3 years
        + 3 Reserved Instance types: Standard, Convertible and Scheduled

Standard: commitment of 1 or 3 years, you are charged whether it's on or off

Scheduled: reserved for specific periods of time, accrue charges hourly, billed monthly increments over 1 year. Scheduled RI's match your capacity reservation to a predictable recurring schedule

RI's are ideal for:

Steady preditable usage workloads

Can sell reservations AWS marketplace if gone unused

* + - * Dedicated hosts
        + Physical servers dedicated just for you
        + Most expensive
        + May be necessary when complying with regulatory requirements
      * Dedicated instances
        + Virtualized instances on hardware solely for you
        + Uses physically dedicated EC2 servers
    - Instance Types

|  |  |
| --- | --- |
| **Category** | **Purpose/Design** |
| General Purpose | Provide a balance of compute, memory and networking resources. Can be used for a **wide variety** of diverse workloads |
| Compute Optimized | Ideal for compute bound applications that **benefit from high performance processors** |
| Accelerated Computing | Uses hardware accelerators, or co-processors to perform functions, such as floating-point number calculations |
| Storage Optimized | Provides non-volatile memory express SSD backed instance storage optimized for low latency, very high random **I/O** performance, high sequential read throughput and provide high IOPS at a low cost |

* + - AMI's
      * Provides information required to launch an instance
      * AMI's include:
        + A template for the root volume of the instance
        + Launch permissions that control which AWS accounts can use the AMI to launch instances
        + A block device mapping that specifies the volumes to attach to the instance when it's launched
      * AMI's are **regional**
        + You can only launch an AMI from the region in which it is stored, however you can **copy** AMI's to other regions using the console, command line or the API
      * Volumes attached to the instace are either EBS or Instance store
      * EBS provides persistent storage. EBS snapshots, which reside on S3 are used to create the volume
      * Instance store volumes are **ephemeral**(non-persistent). That means data is lost if the instance is shut down. A template stored on S3 is used to create the volume
    - IP Addresses
      * There are three types of IP addresses that can be assigned to an EC2 instance
      * Public
        + Lost when the instance is stopped
        + Used in public subnets
        + No charge
        + Associated with private IP address on the instance
        + Cannot be moved between instances
      * Private
        + Retained when the instance is stopped
        + Used in public and private subents
      * Elastic IP
        + Static Public IP address
        + You are chaged if not used
        + Associated with private IP address on the instance
    - Can be moved between instances and the ENA’sElastic Network Interfaces(ENI)
      * An elastic network interface is a logical networking component in a VPC that represents a virtual network card
      * Has the following attributes
        + Primary private IPv4 address from IPv4 address range of your VPC
        + One Elastic IP address(IPv4) per private IPv4 address
        + One public IPv4 address
        + One or more IPv6 addresses
        + One or more security groups
        + A MAC address
        + A source/destination check flag
        + A description
      * You can create and configure network interfaces in your account and attach them to instances in your VPC
      * eth0 is the primary network interface and cannot be moved or detached
      * You can add additional interfaces to EC2 instances
      * An ENI is bound to an AZ
    - Enhanced Networking Adapater(ENA)
      * Enhanced networking provides higher bandwidth, higher packet-per-second performance and consistently lower inter-instance latencies
      * Enhanced networking is enabled using an Elastic Network Adapter
      * If your packets-per-second rate appears to have reached its ceiling, you should consider moving enhanced networking because you have likely reached the upper thresholds of the VIF driver
      * AWS supports enhanced networking capabilities using SR-IOV
      * SR-IOV provides direct access to network adapters, provides higher performance and lower latency
    - Elastic Fabric Adapter(EFA)
      * Elastic Fabric Adapters are ENA's with added capabilities
      * It allso supports an important access model commonly called OS bypass
      * With EFA, High Performance Computing(HPC) applications using the Message Passing Interface(MPI) and ML applications using NVIDIA Collective Communications Library can scale to thousands of CPUs or GPUs
    - When to use
      * ENI
        + This is the basic adapter type for when you don’t have any high performance requirements
        + Can use with all ilnstance types
      * ENA
        + Good for use cases that require higher badnwidth and lower inter-instance latency  
          Supported for limited instance types(HVM only)
      * EFA
        + High performance computing
        + MPI and ML use cases
        + Tightly coupled applications
    - Placement Groups
      * Cluster
        + Cluster instances into a low-latency group in a single AZ
        + Logical goruping of instances within a single AZ
        + Cluster placement groups are recommended for applications that benefit from low network latency, high network throughput, or both
      * Spread
        + Spreads instances across underlying hardware(can span AZ's)
        + Group of instances that are each placed on distinct underlying hardware
        + Recommended for applications that have a small number of critical instances that should be kept apart from eachother
      * Partition
        + Divides each group into logical segments called partitions
        + Eeach partition within a placement group has its own set of racks
        + Used to deploy large deistributed and replicated workloads, such as HDFS, HBase,Cassandra across distinct racks
    - IAM Roles
      * More secure than storing access keys and secret access keys on EC2 instance
      * Easy to manage
      * Can be attached modified or replaced at any time
      * Only 1 IAM role can be attached to an EC2 instance at a time
      * Universal - can be used in any region
    - Bastion Hosts
      * You can configure EC2 instances as bastion hosts in order to access your VPC instances for management
      * Can use SSH or RDP protocols to connect to your bastion host
    - Monitoring
      * EC2 status checks are performed every **minute** and each returns a pass or fail status
      * If all checks, the instance returns status of **OK**
      * If one or more checks fil the overall status is **impaired**
      * You can create or delte alarms that are triggered based on the result of status checks
      * You can create CloudWatch alarms that monitor EC2 instances and automatically perform an action if the status check fails
      * Actions include:
        + Recover the instance
        + Stop the instance
        + Terminate the instance
        + Reboot the instance
      * CloudWatch Monitoring frequency
        + Standard Monitoring - 5 mins
        + Detailed Monitoring - 1 min
    - Tags
      * A tag is a label that you assign to an AWS resource
      * Used to manage AWS assets
      * Tags are arbitrary name/value pairs that you can assign to AWS assets to serve as metadata
      * Tagging strategies can be used for cost allocation, security, etc
    - Resource Groups
      * Mappings of AWS assets defined by tags
      * Create custom consoles to consolidate metrics, alarms, and config details around given tags
    - High Availability Approaches for Compute
      * Up to date AMI's are critical for rapid fail-over
      * AMIs can be copied to other regions for safety or DR staging
      * Horizontally scalable architectures are preffered because risk can spread across smaller machines versus one large machine
      * Reserved instances are the only way to guarantee that resources will be available when needed
      * Auto Scaling and ELB work together to provide automated recovery by maintaining minimum instances
      * Route 53 health checks also provide 'self-healing' redirection of traffic
    - Migration
      * AWS Server Migration Service(SMS) is an agentless service which makes it easier and faster for you to migrate thousands of on-premises workloads to AWS
      * SMS allows you to automate, schedule and track incremental replications of live server volumes, making it easier for you to coordinate large-scale server migrations
      * Replicates VMs to AWS cyncing volumes and creating periodic AMIs

Table

Description automatically generated