Amazon EC2

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It is designed to make web-scale computing easier for developers.

The Amazon EC2 simple web service interface allows you to obtain and configure capacity with minimal friction. It provides you with complete control of your computing resources and lets you run on Amazon's proven computing environment. Amazon EC2 reduces the time required to obtain and boot new server instances (called Amazon EC2 instances) to minutes, allowing you to quickly scale capacity, both up and down, as your computing requirements change. Amazon EC2 changes the economics of computing by allowing you to pay only for capacity that you actually use. Amazon EC2 provides developers and system administrators the tools to build failure resilient applications and isolate themselves from common failure scenarios.

Instance Types

Amazon EC2 passes on to you the financial benefits of Amazon's scale. You pay a very low rate for the compute capacity you actually consume. See Amazon EC2 Instance Purchasing Options for a more detailed description.

- **On-Demand Instances**—With On-Demand instances, you pay for compute capacity by the hour with no long-term commitments. You can increase or decrease your compute capacity depending on the demands of your application and only pay the specified hourly rate for the instances you use. The use of On-Demand instances frees you from the costs and complexities of planning, purchasing, and maintaining hardware and transforms what are commonly large fixed costs into much smaller variable costs. On-Demand instances also remove the need to buy "safety net" capacity to handle periodic traffic spikes.
- **Reserved Instances**—Reserved Instances provide you with a significant discount (up to 75%) compared to On-Demand instance pricing. You have the flexibility to

change families, operating system types, and tenancies while benefitting from Reserved Instance pricing when you use Convertible Reserved Instances.

- **Spot Instances**—Spot Instances allow you to bid on spare Amazon EC2 computing capacity. Since Spot instances are often available at a discount compared to On-Demand pricing, you can significantly reduce the cost of running your applications, grow your application's compute capacity and throughput for the same budget, and enable new types of cloud computing applications.

Amazon EC2 Auto Scaling

Amazon EC2 Auto Scaling helps you maintain application availability and allows you to automatically add or remove EC2 instances according to conditions you define. You can use the fleet management features of EC2 Auto Scaling to maintain the health and availability of your fleet. You can also use the dynamic and predictive scaling features of EC2 Auto Scaling to add or remove EC2 instances. Dynamic scaling responds to changing demand and predictive scaling automatically schedules the right number of EC2 instances based on predicted demand. Dynamic scaling and predictive scaling can be used together to scale faster.

Amazon Elastic Container Registry

Amazon Elastic Container Registry (ECR) is a fully-managed Docker container registry that makes it easy for developers to store, manage, and deploy Docker container images. Amazon ECR is integrated with Amazon Elastic Container Service (Amazon ECS), simplifying your development to production workflow. Amazon ECR eliminates the need to operate your own container repositories or worry about scaling the underlying infrastructure. Amazon ECR hosts your images in a highly available and scalable architecture, allowing you to reliably deploy containers for your applications. Integration with AWS Identity and Access Management (IAM) provides resource-level control of each repository. With Amazon ECR, there are no upfront fees or commitments. You pay only for the amount of data you store in your repositories and data transferred to the Internet.

Amazon Elastic Container Service

Amazon Elastic Container Service (Amazon ECS) is a highly scalable, high-performance container orchestration service that supports Docker containers and allows you to easily run and scale containerized applications on AWS. Amazon ECS eliminates the need for you to install and operate your own container orchestration software, manage and scale a cluster of virtual machines, or schedule containers on those virtual machines.

With simple API calls, you can launch and stop Docker-enabled applications, query the complete state of your application, and access many familiar features such as IAM roles, security groups, load balancers, Amazon CloudWatch Events, AWS CloudFormation templates, and AWS CloudTrail logs.

Amazon Elastic Container Service for Kubernetes

Amazon Elastic Container Service for Kubernetes (Amazon EKS) makes it easy to deploy, manage, and scale containerized applications using Kubernetes on AWS.

Amazon EKS runs the Kubernetes management infrastructure for you across multiple AWS availability zones to eliminate a single point of failure. Amazon EKS is certified Kubernetes conformant so you can use existing tooling and plugins from partners and the Kubernetes community. Applications running on any standard Kubernetes environment are fully compatible and can be easily migrated to Amazon EKS.

Amazon Lightsail

Amazon Lightsail is designed to be the easiest way to launch and manage a virtual private server with AWS. Lightsail plans include everything you need to jumpstart your project – a virtual machine, SSD-based storage, data transfer, DNS management, and a static IP address – for a low, predictable price.

AWS Batch

AWS Batch enables developers, scientists, and engineers to easily and efficiently run hundreds of thousands of batch computing jobs on AWS. AWS Batch dynamically provisions the optimal quantity and type of compute resources (e.g., CPU or memory-optimized instances) based on the volume and specific resource requirements of the batch jobs submitted. With AWS Batch, there is no need to install and manage batch computing software or server clusters that you use to run your jobs, allowing you to

focus on analyzing results and solving problems. AWS Batch plans, schedules, and executes your batch computing workloads across the full range of AWS compute services and features, such as Amazon EC2 and Spot Instances.

AWS Elastic Beanstalk

AWS Elastic Beanstalk is an easy-to-use service for deploying and scaling web applications and services developed with Java, .NET, PHP, Node.js, Python, Ruby, Go, and Docker on familiar servers such as Apache, Nginx, Passenger, and Internet Information Services (IIS).

You can simply upload your code, and AWS Elastic Beanstalk automatically handles the deployment, from capacity provisioning, load balancing, and auto scaling to application health monitoring. At the same time, you retain full control over the AWS resources powering your application and can access the underlying resources at any time.

AWS Fargate

AWS Fargate is a compute engine for Amazon ECS that allows you to run containers without having to manage servers or clusters. With AWS Fargate, you no longer have to provision, configure, and scale clusters of virtual machines to run containers. This removes the need to choose server types, decide when to scale your clusters, or optimize cluster packing. AWS Fargate removes the need for you to interact with or think about servers or clusters. Fargate lets you focus on designing and building your applications instead of managing the infrastructure that runs them.

Amazon ECS has two modes: Fargate launch type and EC2 launch type. With Fargate launch type, all you have to do is package your application in containers, specify the CPU and memory requirements, define networking and IAM policies, and launch the application. EC2 launch type allows you to have server-level, more granular control over the infrastructure that runs your container applications. With EC2 launch type, you can use Amazon ECS to manage a cluster of servers and schedule placement of containers on the servers. Amazon ECS keeps track of all the CPU, memory and other resources in your cluster, and also finds the best server for a container to run on based on your specified resource requirements. You are responsible for provisioning, patching, and scaling clusters of servers. You can decide which type of server to use, which applications and how many containers to run in a cluster to optimize utilization, and

when you should add or remove servers from a cluster. EC2 launch type gives you more control of your server clusters and provides a broader range of customization options, which might be required to support some specific applications or possible compliance and government requirements.

AWS Lambda

AWS Lambda lets you run code without provisioning or managing servers. You pay only for the compute time you consume—there is no charge when your code is not running. With Lambda, you can run code for virtually any type of application or backend service—all with zero administration. Just upload your code, and Lambda takes care of everything required to run and scale your code with high availability. You can set up your code to automatically trigger from other AWS services, or you can call it directly from any web or mobile app.

AWS Serverless Application Repository

The AWS Serverless Application Repository enables you to quickly deploy code samples, components, and complete applications for common use cases such as web and mobile back-ends, event and data processing, logging, monitoring, IoT, and more. Each application is packaged with an AWS Serverless Application Model (SAM) template that defines the AWS resources used. Publicly shared applications also include a link to the application's source code. There is no additional charge to use the Serverless Application Repository - you only pay for the AWS resources used in the applications you deploy.

You can also use the Serverless Application Repository to publish your own applications and share them within your team, across your organization, or with the community at large. To share an application you've built, publish it to the AWS Serverless Application Repository.

AWS Outposts

AWS Outposts bring native AWS services, infrastructure, and operating models to virtually any data center, co-location space, or on-premises facility. You can use the same APIs, the same tools, the same hardware, and the same functionality across on-premises and the cloud to deliver a truly consistent hybrid experience. Outposts can be

used to support workloads that need to remain on-premises due to low latency or local data processing needs.

AWS Outposts come in two variants: 1) VMware Cloud on AWS Outposts allows you to use the same VMware control plane and APIs you use to run your infrastructure, 2) AWS native variant of AWS Outposts allows you to use the same exact APIs and control plane you use to run in the AWS cloud, but on-premises.

AWS Outposts infrastructure is fully managed, maintained, and supported by AWS to deliver access to the latest AWS services. Getting started is easy, you simply log into the AWS Management Console to order your Outposts servers, choosing from a wide range of compute and storage options. You can order one or more servers, or quarter, half, and full rack units.

VMware Cloud on AWS

VMware Cloud on AWS is an integrated cloud offering jointly developed by AWS and VMware delivering a highly scalable, secure and innovative service that allows organizations to seamlessly migrate and extend their on-premises VMware vSphere-based environments to the AWS Cloud running on next-generation Amazon Elastic Compute Cloud (Amazon EC2) bare metal infrastructure. VMware Cloud on AWS is ideal for enterprise IT infrastructure and operations organizations looking to migrate their on-premises vSphere-based workloads to the public cloud, consolidate and extend their data center capacities, and optimize, simplify and modernize their disaster recovery solutions. VMware Cloud on AWS is delivered, sold, and supported globally by VMware and its partners with availability in the following AWS Regions: US East (N. Virginia), US West (Oregon), Asia Pacific (Sydney), Asia Pacific (Tokyo), Europe (Frankfurt), Europe (Ireland), and Europe (London). With each release, VMware Cloud on AWS availability will expand into additional global regions.

VMware Cloud on AWS brings the broad, diverse and rich innovations of AWS services natively to the enterprise applications running on VMware's compute, storage and network virtualization platforms. This allows organizations to easily and rapidly add new innovations to their enterprise applications by natively integrating AWS infrastructure and platform capabilities such as AWS Lambda, Amazon Simple Queue Service (SQS), Amazon S3, Elastic Load Balancing, Amazon RDS, Amazon DynamoDB, Amazon Kinesis, and Amazon Redshift, among many others.

With VMware Cloud on AWS, organizations can simplify their Hybrid IT operations by using the same VMware Cloud Foundation technologies including vSphere, vSAN, NSX, and vCenter Server across their on-premises data centers and on the AWS Cloud without having to purchase any new or custom hardware, rewrite applications, or modify their operating models. The service automatically provisions infrastructure and provides full VM compatibility and workload portability between your on-premises environments and the AWS Cloud. With VMware Cloud on AWS, you can leverage AWS's breadth of services, including compute, databases, analytics, Internet of Things (IoT), security, mobile, deployment, application services, and more.

Source: https://docs.aws.amazon.com/aws-technical-content/latest/aws-overview/compute-services.html