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 Amazon EC2 Auto Scaling to maintain the health and availability of your fleet. You can also use the dynamic and predictive scaling features of Amazon 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 EC2 Image Builder

EC2 Image Builder simplifies the building, testing, and deployment of Virtual Machine and container images for use on AWS or on-premises.

Keeping Virtual Machine and container images up-to-date can be time consuming, resource intensive, and error-prone. Currently, customers either manually update and snapshot VMs or have teams that build automation scripts to maintain images.

Image Builder significantly reduces the effort of keeping images up-to-date and secure by providing a simple graphical interface, built-in automation, and AWS-provided security settings. With Image Builder, there are no manual steps for updating an image nor do you have to build your own automation pipeline.

Image Builder is offered at no cost, other than the cost of the underlying AWS resources used to create, store, and share the images.

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 App Runner

AWS App Runner is a fully managed service that makes it easy for developers to quickly deploy containerized web applications and APIs, at scale and with no prior infrastructure experience required. Start with your source code or a container image. App Runner automatically builds and deploys the web application and load balances traffic with encryption. App Runner also scales up or down automatically to meet your traffic needs. With App Runner, rather than thinking about servers or scaling, you have more time to focus on your applications.

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 runs 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.

AWS Wavelength

AWS Wavelength is an AWS Infrastructure offering optimized for mobile edge computing applications. Wavelength Zones are AWS infrastructure deployments that embed AWS compute and storage services within communications service providers’ (CSP) datacenters at the edge of the 5G network, so application traffic from 5G devices can reach application servers running in Wavelength Zones without leaving the telecommunications network. This avoids the latency that would result from application traffic having to traverse multiple hops across the Internet to reach their destination, enabling customers to take full advantage of the latency and bandwidth benefits offered by modern 5G networks.