

Service Name

Amazon Elastic Compute Cloud (EC2)

Service Overview

Amazon Elastic Compute Cloud (Amazon EC2) is a web service that provides secure, resizable compute capacity in the cloud. It enables users to launch and manage virtual servers, called EC2 instances, in the AWS Cloud with complete control over the operating system and applications. EC2 offers various instance types with different balances of compute, memory, storage, and networking resources to meet diverse workload requirements. Users can configure security, networking, and storage for their instances, scale capacity up or down based on demand, and pay only for the resources they use. EC2 provides the foundation for many cloud computing solutions and serves as a core building block for AWS infrastructure.

Key Use Cases

- Web servers and application hosting: Running web applications, APIs, and microservices with scalable compute capacity
- High-performance computing (HPC): Scientific computing, financial modeling, batch processing, and simulation workloads requiring powerful processors
- Enterprise applications: Business-critical applications, databases, ERP systems, and development environments with customizable configurations
- Big data and analytics: Data processing, machine learning training, real-time analytics, and distributed computing workloads
- Gaming and media: Game servers, video encoding, rendering workloads, and content delivery applications
- Disaster recovery and backup: Creating backup systems, disaster recovery environments, and failover infrastructure

Features

- **Instance Types:** Over 750 instance types across 7 categories (General Purpose, Compute Optimized, Memory Optimized, Storage Optimized, Accelerated Computing, High Performance Computing, Previous Generation) with various processor, memory, storage, and networking configurations
- **Amazon Machine Images (AMIs):** Pre-configured templates containing operating systems, applications, and software configurations for quick instance deployment
- **Elastic Block Store (EBS):** High-performance block storage volumes that persist independently of instance lifecycle, with multiple volume types and encryption capabilities

- **Auto Scaling:** Automatically adjusts the number of instances based on demand, maintaining application availability while optimizing costs
- **Security Groups:** Virtual firewalls that control inbound and outbound traffic to instances at the protocol and port access level
- **Elastic Load Balancing:** Distributes incoming traffic across multiple instances to ensure high availability and fault tolerance
- **Spot Instances:** Access to spare EC2 capacity at up to 90% discount compared to On-Demand prices for fault-tolerant and flexible workloads
- **Enhanced Networking:** High packet-per-second performance, low latency, and high throughput networking capabilities with SR-IOV and Elastic Fabric Adapter support

Value Proposition

EC2 provides unmatched flexibility and control over computing resources in the cloud, allowing customers to choose from the broadest selection of processors (Intel, AMD, AWS Graviton, NVIDIA) and instance types to optimize performance and cost for specific workloads. The service offers multiple purchasing options including On-Demand, Reserved Instances, Savings Plans, and Spot Instances, enabling up to 90% cost savings:

- EC2's global infrastructure spans multiple regions and availability zones, ensuring high availability and fault tolerance
- Integration with 200+ AWS services provides a comprehensive cloud ecosystem, while features like Auto Scaling and Elastic Load Balancing ensure applications remain available and performant under varying loads
- The pay-as-you-go pricing model eliminates upfront investments and capacity planning guesswork

Service Integration

EC2 integrates seamlessly with core AWS services to provide comprehensive cloud solutions. Amazon VPC provides isolated networking environments with custom IP ranges, subnets, and security controls:

- Amazon EBS offers persistent block storage with multiple performance tiers and encryption
- Amazon CloudWatch monitors instance performance, health, and resource utilization with custom metrics and alarms
- AWS IAM manages access control and permissions for EC2 resources
- Elastic Load Balancing distributes traffic across instances for high availability
- Amazon Route 53 provides DNS services for domain management
- AWS Systems Manager enables instance management, patching, and configuration at scale

- Integration with Amazon S3 provides object storage for backups and data archiving, while Amazon RDS offers managed database services that work with EC2-hosted applications

Onboarding and Offboarding

Customers can get started with EC2 through the AWS Management Console using the launch instance wizard, which provides default configurations for quick deployment. The process involves selecting an Amazon Machine Image (AMI), choosing an instance type, configuring security groups and key pairs, and setting up storage:

- AWS offers a Free Tier that includes 750 hours per month of t2.micro instances for eligible customers
- The getting started tutorial guides users through launching, connecting to, and managing their first instance
- Multiple access methods are available including the console, AWS CLI, SDKs, CloudFormation, and APIs
- AWS provides extensive documentation, tutorials, and hands-on labs to help customers learn EC2 fundamentals
- For production workloads, customers should consider launch templates for consistent deployments and Auto Scaling for high availability

Getting Started Guide:

https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html

Data Removal

To remove data when offboarding from EC2, customers must terminate instances, which permanently deletes all data on instance store volumes and detaches EBS volumes. EBS volumes can be separately deleted to remove persistent data. Snapshots stored in Amazon S3 should be explicitly deleted. AMIs and associated snapshots created by customers need manual deletion. Key pairs can be deleted from the console. Security groups and other network resources should be cleaned up. AWS provides CloudFormation templates to automate resource cleanup and ensure complete removal of infrastructure and associated data.

Backup/Restore and Disaster Recovery

EC2 provides comprehensive backup and disaster recovery capabilities through EBS snapshots, AMIs, and AWS Backup integration. EBS snapshots create point-in-time backups stored in Amazon S3 with 99.999999999% durability. AMIs capture complete instance configurations including operating system and applications. AWS Backup offers centralized backup management across EC2 instances with automated scheduling, lifecycle management, and cross-region replication. Instance store volumes require

application-level backup to durable storage like S3 or EBS. Multi-AZ deployments and Auto Scaling groups provide built-in fault tolerance and automatic recovery capabilities.

Backup Capabilities

EC2 offers native backup capabilities through EBS snapshots and AMIs for creating point-in-time backups. AWS Backup provides centralized backup management for EC2 instances with automated scheduling, retention policies, and compliance reporting:

- Backup jobs can be configured at the instance level with custom backup plans
- Cross-region backup replication is supported for disaster recovery scenarios
- AWS Backup enables item-level search and recovery from backups, allowing restoration of specific files without full instance recovery

Disaster Recovery

EC2 supports disaster recovery through multi-region deployments, cross-AZ redundancy, and integration with AWS Elastic Disaster Recovery. Customers can replicate instances and data across regions for geographic separation:

- Auto Scaling groups automatically replace failed instances
- EBS snapshots can be copied across regions for data recovery
- AWS Elastic Disaster Recovery provides continuous replication and fast recovery of physical, virtual, and cloud-based servers
- CloudFormation templates enable consistent infrastructure recreation in disaster scenarios

Recovery Objectives

EC2 helps customers achieve RPO and RTO objectives through multiple mechanisms. EBS snapshots can be taken as frequently as every hour for low RPO requirements. AMIs enable rapid instance recreation for improved RTO. Auto Scaling groups provide automatic instance replacement within minutes. Cross-region replication and AWS Elastic Disaster Recovery support enterprise RPO/RTO requirements. Reserved capacity and placement groups ensure resources are available during recovery scenarios. Automated backup scheduling through AWS Backup maintains consistent recovery points across the infrastructure.

Service Constraints

Service Quotas: <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-resource-limits.html>

FAQ: <https://aws.amazon.com/ec2/faqs/>

Technical Requirements

Service Documentation:

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/concepts.html>

User Guide: <https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/>

API Reference: <https://docs.aws.amazon.com/AWSEC2/latest/APIReference/>

Pricing Overview

Please see the AWS UK G Cloud 15 Pricing Document accompanying this service in the Digital Marketplace.

Public Pricing: <https://aws.amazon.com/ec2/pricing/>

Cost Optimization

EC2 offers multiple unique cost optimization options including Spot Instances providing up to 90% savings for fault-tolerant workloads, Savings Plans offering up to 72% discount for consistent usage commitments, and Reserved Instances for predictable workloads. AWS Graviton-based instances deliver up to 40% better price performance:

- Flexible instance types like M7i-flex provide optimized pricing for general-purpose workloads
- EC2 Instance Type Finder and AWS Compute Optimizer recommend optimal instance types
- Auto Scaling automatically adjusts capacity to match demand
- Hibernation allows pausing instances while preserving memory state
- Dedicated Hosts enable license mobility for cost savings
- Mixed instance types in Auto Scaling groups optimize costs while maintaining performance

Savings Considerations

Key cost savings considerations include analyzing workload patterns to choose appropriate purchasing options - use On-Demand for variable workloads, Reserved Instances for steady-state usage, and Spot Instances for fault-tolerant applications. Right-sizing instances based on actual usage prevents over-provisioning:

- Implementing Auto Scaling ensures paying only for needed capacity
- Choosing the optimal instance family and generation impacts price performance
- AWS Graviton instances offer significant cost advantages
- Storage optimization through appropriate EBS volume types and snapshot lifecycle management reduces costs

- Using placement groups and Enhanced Networking features can improve efficiency without additional charges
- Regular cost monitoring through Cost Explorer and Trusted Advisor identifies optimization opportunities