

# Adding a Compute Layer Using Amazon EC2

COMPUTE CHOICES · AMIs · INSTANCE TYPES · STORAGE · USER DATA & CONFIG · PRICING MODELS · WELL-ARCHITECTED

STUDY & REVIEW

## 01 MUST KNOW

if you remember five things, remember these

- 01 Launch = a stack of choices:** provisioning an instance means specifying an **AMI**, an **instance type**, a **key pair** (SSH/RDP), **network** placement in a **VPC**, a **security group** (firewall rules), **storage**, an optional **IAM role** passed via an *instance profile*, and optional **user data**. Every instance launches inside a VPC.
- 02 AMI = the launch template:** an AMI supplies a **template for the root volume** (guest OS + software), **launch permissions** (who can use it), and **block device mappings** (which volumes to attach). Its value is **repeatability, reusability, recoverability**; use the **HVM** virtualization type for best performance. Sources: Quick Start, My AMIs, AWS Marketplace, Community (at your own risk).
- 03 Instance type = the sizing:** a type fixes **vCPU, memory, storage, and network performance**. Read c7gn.xlarge as family c · generation 7 · processor/capabilities gn · size xlarge. Six categories: **general purpose, compute-, storage-, memory-optimized, accelerated computing, HPC**. Prefer the **newest generation** — better price-to-performance.
- 04 Right storage for the job:** a **root** volume must be **instance store** or **SSD-backed EBS**. **Instance store** is ephemeral (lost on stop/terminate); **Amazon EBS** is persistent block storage that outlives the instance. Share data across **multiple Linux** instances with **Amazon EFS** (NFS); across **multiple Windows** instances with **Amazon FSx for Windows File Server** (SMB/NTFS).
- 05 Blend the pricing models:** cover **steady-state** work with **Reserved Instances** or **Savings Plans** (1- or 3-year commitment), absorb **spiky** demand with **On-Demand**, and run **fault-tolerant, stateless** work on **Spot** (spare capacity, ~2-minute interruption notice). **Dedicated Hosts/Instances** provide single-tenant hardware for licensing and compliance.

## 02 KEY TERMS

| TERM                              | DEFINITION                                                                                                                            |
|-----------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|
| <b>Amazon EC2</b>                 | Service providing resizable virtual servers (instances) in the cloud, provisioned in minutes; you pay only for the capacity you use   |
| <b>EC2 instance</b>               | A virtual machine that runs on a physical host in an Availability Zone, on top of an AWS-maintained hypervisor layer                  |
| <b>Hypervisor</b>                 | The operating-platform layer AWS maintains that gives each instance access to the host's physical CPU, memory, and storage            |
| <b>Amazon Machine Image (AMI)</b> | The template Amazon EC2 uses to launch an instance: root-volume template, launch permissions, and block device mappings               |
| <b>Instance type</b>              | A defined configuration of vCPU, memory, storage, and network performance; 270+ types across six category families                    |
| <b>Instance store</b>             | Temporary (ephemeral) block storage physically attached to the host; data is lost when the instance stops or terminates               |
| <b>Amazon EBS</b>                 | Network-attached persistent block storage in one Availability Zone; persists independently of the instance and snapshots to Amazon S3 |
| <b>EBS-optimized instance</b>     | An instance with a dedicated network connection to its EBS volumes that minimizes I/O contention (425–14,000 Mbps)                    |
| <b>Amazon EFS</b>                 | Fully managed, elastic shared file system for Linux instances using NFS; scales to petabytes and mounts across Availability Zones     |

| TERM                                      | DEFINITION                                                                                                                                                                                          |
|-------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Amazon FSx for Windows File Server</b> | Fully managed shared file system for Windows instances: native NTFS, SMB protocol, Active Directory integration, SSD-backed                                                                         |
| <b>User data</b>                          | A script (shell or cloud-init) passed at launch that runs once with root/Administrator privileges to bootstrap the instance                                                                         |
| <b>Instance metadata</b>                  | Data about a running instance (ID, IPs, hostname, security groups, Region, AZ), read in-instance at <a href="http://169.254.169.254/latest/meta-data/">http://169.254.169.254/latest/meta-data/</a> |

### 03 PROVISIONING AN EC2 INSTANCE

the main launch components, in order

**01 AMI** — the launch template → **02 Instance type** — CPU/memory/storage/network → **03 Key pair** — SSH/RDP access → **04 Network** — VPC, subnet, public IP/DNS → **05 Security group** — firewall port rules → **06 Storage** — instance store or EBS + extra volumes → **07 IAM role** — via instance profile, for API calls → **08 User data** — optional launch automation

### 04 INSTANCE FAMILIES & THE WORKLOADS THEY FIT

newest generation = best price-to-performance

| CATEGORY                     | TYPICAL WORKLOADS                                              | EXAMPLES               |
|------------------------------|----------------------------------------------------------------|------------------------|
| <b>General purpose</b>       | Web/app servers, enterprise apps, gaming, dev/test             | M7, M6, M5, T4, T3, T2 |
| <b>Compute optimized</b>     | Batch processing, distributed analytics, HPC                   | C7, C6, C5, C4         |
| <b>Storage optimized</b>     | High-performance databases, real-time analytics, transactional | I4, I3, D2, D3, H1     |
| <b>Memory optimized</b>      | In-memory caches, high-performance databases, big data         | R7, R6, R5, X2, X1     |
| <b>Accelerated computing</b> | Machine learning/AI, HPC, graphics processing                  | P5, P4, Inf2, G5, F1   |
| <b>HPC optimized</b>         | Deep learning, compute-intensive HPC simulations               | Hpc7, Hpc6             |

### 05 WHICH STORAGE, AND WHEN

root vs. single-instance data vs. shared file system

| STORAGE OPTION                   | ROOT | SINGLE-INSTANCE DATA            | SHARED FILE SYSTEM                     |
|----------------------------------|------|---------------------------------|----------------------------------------|
| <b>Instance store</b>            | Yes  | Yes (that instance only)        | No — ephemeral, lost on stop/terminate |
| <b>Amazon EBS (SSD for root)</b> | Yes  | Yes (one at a time; detachable) | No — persistent block storage          |
| <b>Amazon EFS</b>                | No   | No                              | Yes — Linux, NFS                       |
| <b>Amazon FSx for Windows</b>    | No   | No                              | Yes — Windows, SMB/NTFS                |

### 06 PURCHASING OPTIONS — MATCH THE MODEL TO THE WORKLOAD

combine models to optimize cost

| MODEL                            | HOW YOU PAY / COMMIT                                                      | RECOMMENDED FOR                                                   |
|----------------------------------|---------------------------------------------------------------------------|-------------------------------------------------------------------|
| <b>On-Demand</b>                 | Per second or per hour; no commitment                                     | Spiky, short-term, or experimentation workloads                   |
| <b>Reserved</b>                  | 1- or 3-year commitment; significant discount off On-Demand               | Committed, steady-state workloads                                 |
| <b>Savings Plans</b>             | \$/hour commitment (1 or 3 yr); Reserved-level discount, more flexibility | All EC2 workloads; those wanting flexibility with committed usage |
| <b>Spot</b>                      | Spare capacity at substantial savings; 2-minute interruption notice       | Fault-tolerant, flexible, stateless workloads                     |
| <b>Dedicated Hosts/Instances</b> | Single-tenant physical hardware (per-host or per-instance)                | Compliance, regulatory, or server-bound (BYOL) licenses           |

- × “Instance store keeps my data safe.” — It is **ephemeral**: data survives a **reboot** but is **lost** when the instance is **stopped or terminated**. Only Amazon EBS (or EFS/FSx/S3) persists data independently of the instance.
- × “You can stop any instance to pause billing.” — An **instance store-backed** instance can only be **rebooted or terminated**, never stopped — and you **can't change its instance type**. Only **EBS-backed** instances stop, start, and change type.
- × “Use PV virtualization for speed.” — For best performance choose **HVM**; PV boots differently and can't use certain hardware extensions. HVM is the recommended default.
- × “EFS works for Windows, or just use S3 to share.” — **Amazon EFS is Linux/NFS**; **FSx for Windows File Server is Windows/SMB**. Amazon S3 is an **object** store (whole-file overwrites), workable but not ideal for high-throughput file sharing.
- × “Bigger instance = better.” — Over-provisioning wastes money and can hide performance problems. **Slightly under-spec, then resize**; within a family the **newest generation** usually wins on price-to-performance.
- × “Spot Instances are guaranteed capacity.” — AWS can reclaim them with a **2-minute** notice — use only for **fault-tolerant, flexible, stateless** work. Per-second billing (60-second minimum) applies only to **Amazon Linux/Ubuntu**; other OSes bill per hour.
- × “An AMI captures the whole instance.” — It does **not** carry the **security group, IAM role, or instance type** — those are separate settings you must specify each time you launch from the AMI.
- × “User data re-runs on every boot.” — It runs **only at first launch**. To re-run it on a running instance you must **stop, edit** the data, **delete** the `config_scripts_user` semaphore file, then restart — skip the delete and the new script won't run.

## 08 SELF-QUIZ

answers are upside-down below

- Q1** When you provision an EC2 instance you specify eight main items — name as many as you can (AMI and instance type are two).
- Q2** What three pieces of information does an AMI provide for launching an instance?
- Q3** Name the three benefits of using AMIs — all begin with the letter R.
- Q4** For best performance, which of the two AMI virtualization types should you choose?
- Q5** Decode each part of the instance type name `c7gn.xlarge`.

- Q6** Which two storage options can hold an EC2 root (boot) volume?
- Q7** Which service gives a shared file system to multiple Linux instances, and which does the same for Windows instances?
- Q8** Which two EBS volume types can serve as boot volumes, and which two cannot?
- Q9** What inbound TCP port must a mount target's security group allow so instances can reach an EFS file system over NFS?
- Q10** Sample exam: a workload runs  $\geq 1$  year uninterrupted in one Region, steady except seasonal spikes that increase instance size within the same family — which pricing option is lowest cost?

ANSWER KEY (rotate the page)

**Q1** AMI, instance type, key pair, network/PC placement (with public IP/DNS), security group, storage, IAM role (via an instance profile), and user data **Q2** A template for the root volume (OS + software), launch permissions, and block device mappings **Q3** Repeatable, reusable, and recoverability **Q4** HVM (Hardware Virtual Machine) **Q5** c = family · 7 = generation · g = AWS Graviton processor · n = network and EBS optimized · xlarge = size **Q6** Instance store or an SSD-backed Amazon EBS volume **Q7** Amazon EFS for Linux (NFS); Amazon FSx for Windows File Server (SMB/NFS) **Q8** gp2 and io1 (SSD) can be boot volumes; st1 and sc1 (HDD) cannot **Q9** TCP port 2049 **Q10** EC2 Instance Savings Plans — locks to one instance family in a Region for the largest discount while still allowing size changes