

Computing (Servers)

EC2 VIRTUALIZATION · INSTANCE TYPES · AMIs & STORAGE · NETWORKING · USER DATA & METADATA · LIFECYCLE · STUDY & REVIEW
· PURCHASING OPTIONS

01 MUST KNOW

if you remember five things, remember these

- 01 EC2 is virtual machines.** Instances run as VMs on an **AWS-managed hypervisor** inside Availability Zones; each gets **vCPUs** and **RAM**. **Instance store** is *ephemeral* — it survives a reboot but is **lost on stop or terminate**; **Amazon EBS** is *persistent* block storage that survives a stop.
- 02 The AMI is the template.** Every instance launches from an **Amazon Machine Image** (root-volume OS/apps + launch permissions + block device mappings). Determining the AMI is the **first launch step**; then instance type, network placement/IP, block storage, a security group, a key pair, and optional user data.
- 03 Security groups are stateful, instance-level firewalls.** Every instance has at least one; they filter by **port range**, **IP/CIDR range**, or **source security group**. VPC groups control **inbound and outbound** traffic and can be modified while the instance runs. A source of `0.0.0.0/0` means anywhere.
- 04 State drives billing and what survives.** Billing starts at **Running**. A **reboot** keeps the same public IP and billing hour; a **stop** → **start** assigns a **new** public IP and starts a **new** billing hour. Instance store is lost on stop, EBS is retained, and an **Elastic IP** is static.
- 05 Use an IAM role, never stored keys.** Attach an **instance profile** (a container for an IAM role) for **temporary, auto-rotated** credentials — never store permanent access keys on an instance. Metadata lives at `169.254.169.254`; user data runs at first boot via **cloud-init** (Linux) or **EC2Launch** (Windows).

02 KEY TERMS

TERM	DEFINITION
Amazon EC2	Web service providing resizable virtual servers (instances) that run on AWS-managed hypervisors in Availability Zones
Instance type	A named bundle of virtual hardware (vCPU, memory, storage, network), organized as category → family → type (e.g., <code>t3.nano</code>)
AMI (Amazon Machine Image)	The template EC2 uses to launch an instance: a root-volume image, launch permissions, and block device mappings
Instance store	Ephemeral block storage physically attached to the host; survives a reboot but is lost on stop or terminate
Amazon EBS	Persistent block-storage volumes used for the boot disk and data; the volume survives an instance stop
EBS-optimized instance	An instance that minimizes I/O contention between EBS and other network traffic for better performance
Security group	A stateful virtual firewall applied at the instance level, allowing or blocking traffic by port, IP/CIDR, or source group
Elastic IP address	A static, public IPv4 address you allocate from AWS and can reassign between instances at any time
Instance profile	A container for an IAM role attached to an instance, granting temporary, automatically rotated credentials
User data	A launch-time script (shell, batch, or PowerShell) run by cloud-init (Linux) or EC2Launch (Windows) to automate configuration
Instance metadata	Data about a running instance, retrieved from the link-local service at <code>169.254.169.254/latest/meta-data/</code>

TERM	DEFINITION
Key pair	Public/private keys for login: SSH on Linux (port 22) or decrypting the RDP password on Windows (port 3389)

03 PROVISIONING AN INSTANCE

the launch is layered on a template

01 Choose an **AMI** → 02 Select **instance type** & size → 03 Set **network placement** — VPC/subnet + optional public IP → 04 Attach **block storage** — instance store or EBS → 05 Configure a **security group** → 06 Specify a **key pair** → 07
Optional: **user data** + instance profile

04 EC2 PURCHASING OPTIONS

match the pricing model to the usage pattern

OPTION	BENEFIT	BEST FOR
On-Demand	Lowest upfront cost, most flexibility; no commitment	Short-term, spiky, or unpredictable workloads; dev/test
Reserved (RI)	Up to 75% off + a capacity reservation; 1–3 yr term	Steady-state, predictable usage needing reserved capacity
Scheduled RI	Capacity that recurs daily/weekly/monthly; 1 yr term	Recurring jobs run on a schedule for a finite time
Spot	70–90% savings on spare capacity; 2-min notice	Fault-tolerant, urgent, or large batch workloads
Dedicated Host	Physical server fully dedicated to you; placement control	Bringing per-socket / per-core / per-VM licenses
Dedicated Instance	Runs on hardware dedicated to a single account	Regulatory rules requiring physical isolation
Bare metal (i3.metal)	Direct access to the physical host's hardware	Non-virtualized workloads; features like Intel VT-x

05 INSTANCE CATEGORIES

category → family → type

CATEGORY	EMPHASIZES
General purpose	Balanced compute, memory, and networking (e.g., the T3 family)
Compute optimized	Extra CPU / processing power
Memory optimized	Extra RAM for memory-intensive apps (e.g., x1e.32xlarge)
Storage optimized	High-capacity, high-throughput local storage
Accelerated computing	Extra I/O and specialized hardware acceleration

06 INSTANCE STATE → WHAT SURVIVES

billing, storage, and IPs by state

ASPECT	RUNNING	STOPPED	TERMINATED
Billing	Charged	Not charged	Not charged
Instance store	Retained	Lost	Lost
EBS volumes	Retained	Retained	Deleted if DeleteOnTermination set
Public IP	Retained	Released	Released
Elastic IP	Retained	Retained	Disassociated

- × “Rebooting changes the public IP.” — No. A **reboot keeps the same public IP** and the same billing hour; only a **stop → start** assigns a new public IP and starts a new billing hour.
- × “Instance-store data survives a stop.” — It is **ephemeral**: it survives a reboot but is **lost on stop or terminate** and cannot be recovered. Only EBS persists across a stop.
- × “Security groups filter at the subnet boundary.” — They act at the **instance level**, not the network entry point — instances in one subnet can carry different groups — and they are **stateful**.
- × “You can turn on hibernation for a running instance.” — Hibernation **must be enabled at launch**; you cannot enable it on an existing running or stopped instance, and the EBS root volume must be **encrypted**.
- × “Any instance can be stopped.” — An **instance-store-backed** instance **cannot be stopped — only terminated**. Only instances with an **EBS root volume** can be stopped and started again.
- × “Store AWS keys on the instance so apps can call AWS.” — Never store permanent keys on an instance. Attach an **IAM role via an instance profile** for temporary, auto-rotated credentials.
- × “gp2 is just a slower io1.” — gp2 (General purpose SSD) is the **default** for broad transactional workloads; io1 (Provisioned IOPS SSD) is the **highest-performance** choice for I/O-intensive work — different jobs, not tiers.
- × “A Reserved Instance is a physical server you're given.” — An RI is a **billing capacity reservation, not a physical instance**. The physical-server options are the **Dedicated Host** and **Dedicated Instance**.

08 SELF-QUIZ

answers are upside-down below

Q1 Which storage type is lost when an instance is stopped but survives a reboot?

Q2 What is the first step when launching an EC2 instance?

Q3 At what IP address does every instance reach its metadata service?

Q4 A reboot versus a stop-and-start — in which case does an instance receive a new public IP?

Q5 What object attaches an IAM role to an EC2 instance?

Q6 Which EBS volume type is the default for EC2 instances?

Q7 Which purchasing option gives 70–90% savings on spare capacity but can be reclaimed with a 2-minute notice?

Q8 Which service runs user-data scripts at launch on Linux, and which runs them on Windows?

Q9 In the shared responsibility model, who patches the guest OS on your EC2 instance?

Q10 Which instance state stops billing yet keeps the EBS volumes and the Elastic IP?

ANSWER KEY (rotate the page)
Q1 instance store (ephemeral storage) **Q2** determine / select the AMI **Q3** 169.254.169.254 **Q4** when it is stopped and started (a reboot keeps the same public IP) **Q5** an instance profile **Q6** General purpose SSD (gp2) **Q7** Spot Instances **Q8** cloud-init on Linux; EC2Launch on Windows **Q9** the customer (security in the cloud) — AWS has no OS-layer visibility **Q10** Stopped