MODULE 5: Adding a Compute Layer using Amazon EC2

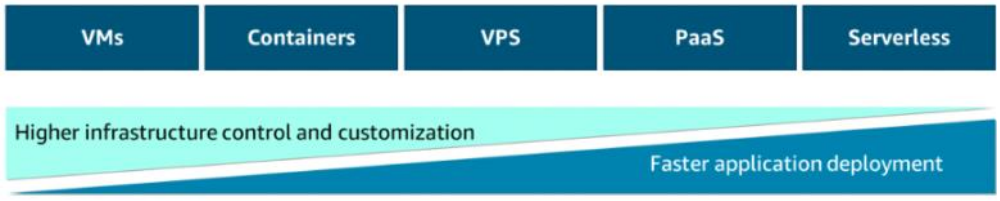
Thursday, October 9, 2025 9:25 AM

I. AWS Ruhtime compute choices:

AWS offers different compute services to meet the needs of different user's cases.

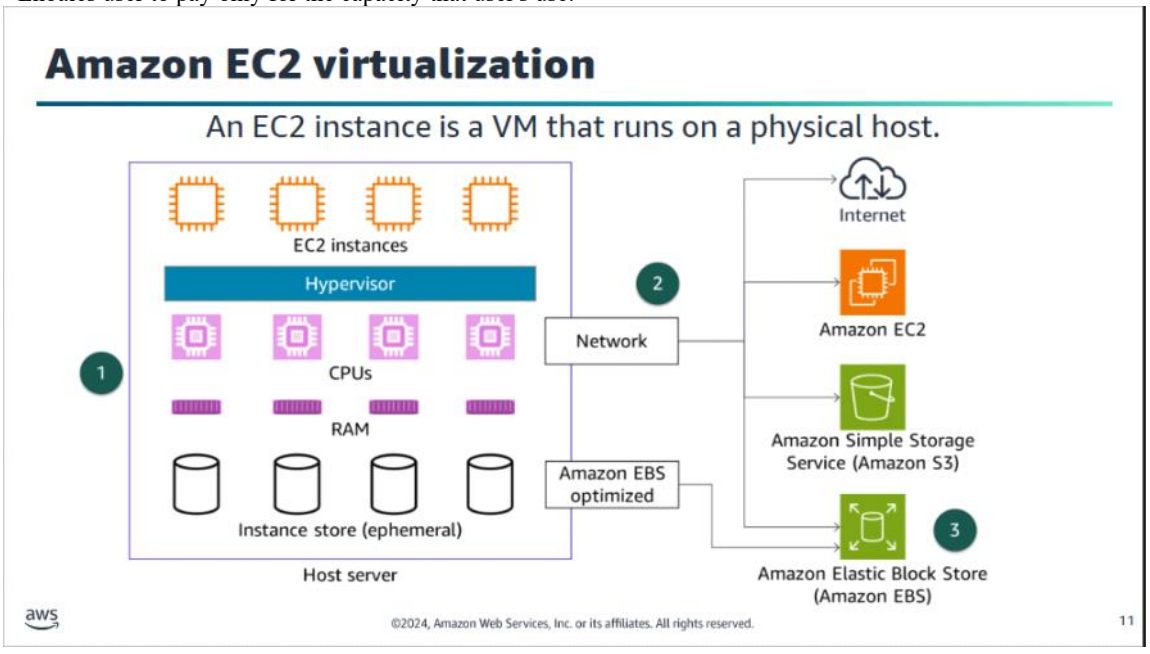
- Virtual Machine (VMs): Amazon Elastic Compute Cloud (Amazon EC2)
- Containers:
  + Amazon Elastic Container Service (Amazon ECS)
  + Amazon Elastic Kubemetes Service (Amazon EKS)
- Virtual Private Servers (VPS): Amazon lightsail
- Platform as a Service (PaaS): Amazon Elastic Beanstalk
- Serverless: AWS Lambda and AWS Fargate

II. Compute service category differentiators



III. Amazon EC2 and its virtualization

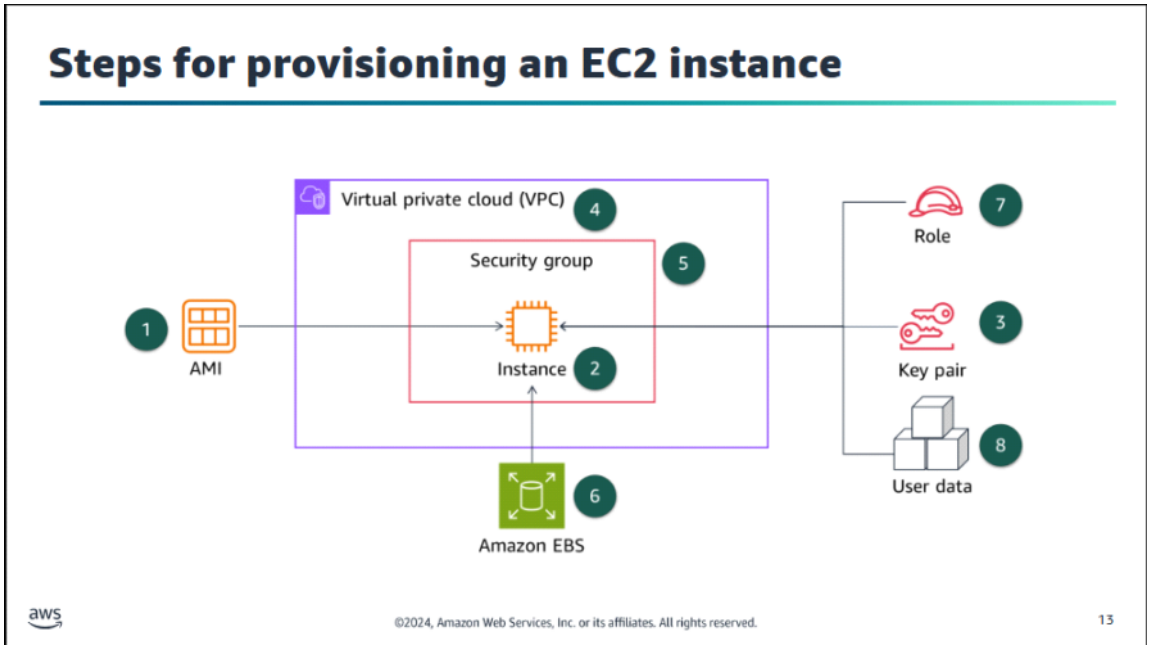
- Provides VMs (servers) in the cloud
- Provisions servers in minutes
- Can automatically scale capcity up or down as needed
- Enbales user to pay only for the capacity that user's use.



IV. Amazon EC2 Use Cases

- Complete control of user's computing resources, including operating system and processor type.
- Options for optimizing user's compute costs:
  + On-Demand Instances, Reseved Instances, and Spot Instances
  + Saving Plans
- Ability to run any type of workload:
  + Simple websites
  + Enterprise locations
  + Generative AI applications

V. Steps for Provisioning an EC2 instance



1. **AMI** – Choose an Amazon Machine Image (the system template).
2. **Instance Type** – Pick hardware specs (CPU, memory, etc.).
3. **Key Pair** – Create or select key for secure login (SSH/RDP).
4. **VPC** – Choose network and IP settings.
5. **Security Group** – Set firewall rules for traffic.
6. **Storage (EBS)** – Attach storage volumes.
7. **IAM Role** – Give instance permissions to access AWS services.
8. **User Data** – Add startup scripts for automation.

VI. Amazon Machine Image (AMI)

- An AMI provides the information that is needed to launch an instance including:
  + A template for the root volume: Contains the guest operating system (OS) and perhaps other installed software
  + Launch permissions: Controls who can access the AMI
  + Block device mappings: Specifies any storage volumes to attach the instance
- AMI benefits:
  + **Reapeatability**: An AMI can be used repeatedly to launch instances with efficiency and precision
  + **Reusability**: Instances launched from the same AMI are identically configured.
  + **Recoverability**:
    * Users can create an AMI from a configured instance as a restorable backup
    * Users can replace a failed instance by launching a new instance from the same AMI
- Choosing an AMI based on following:
  + Region
  + Operating System
  + Storage type of the root device
  + Architecture
  + Virtualization type: For best performance, use an AMI with a Hardware Virtual Machine (HVM) virtualization type.

VII. Instance store-backed versus Amazon EBS-backed AMI

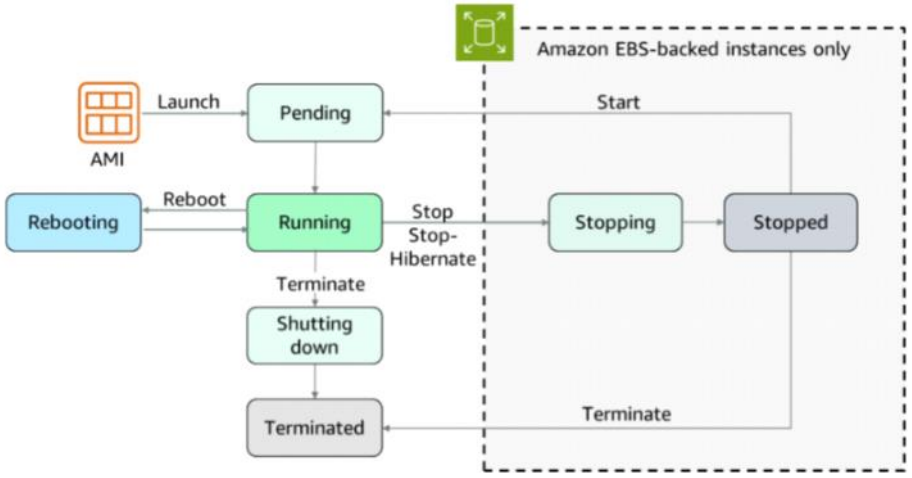
Instance store-backed versus Amazon EBS-backed AMI

Characteristic	Amazon EBS-Backed Instance	Instance Store-Backed Instance
Boot time for the instance	Boots faster	Takes longer to boot
Maximum size of root device	16 TiB	10 GiB
Ability to stop the instance	Can stop the instance	Cannot be in a stopped state; instances are running or terminated
Ability to change the instance type	Can change the instance type by stopping instance	Can't change the instance type because the instance can't be stopped
Instance charges	You are charged for instance usage, EBS volume usage, and storing your AMI as an EBS snapshot	You are charged for instance usage and storing your AMI in Amazon S3
Use case	Persistent storage	Temporary storage

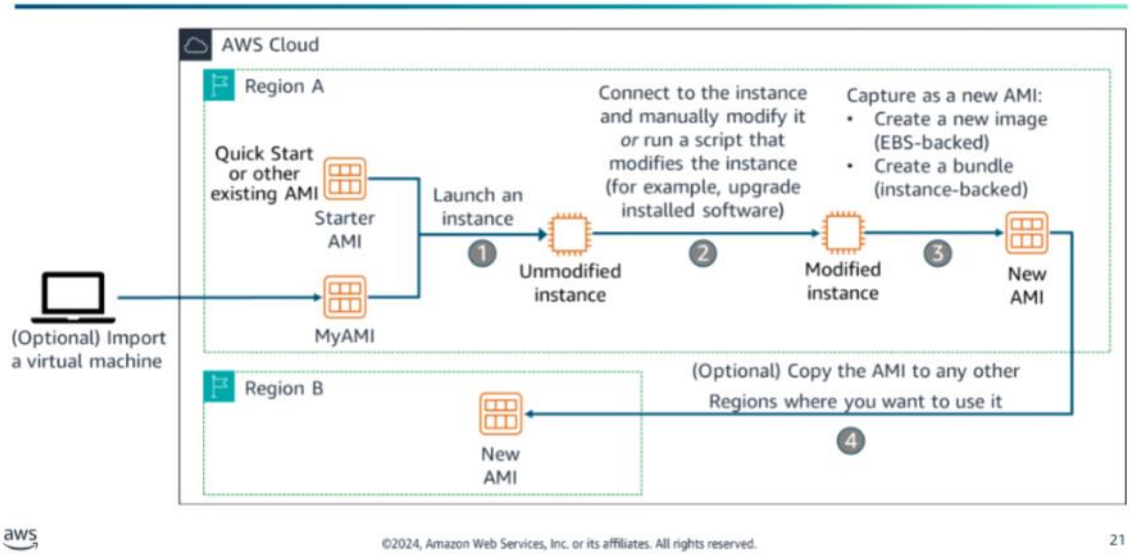
aws ©2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

VIII. Amazon EC2 Lifecycle

Amazon EC2 instance lifecycle



IX. Creating a new AMI



X. EC2 Image Builder

EC2 Image Builder automates the creation, management, and deployment of up-to-date and compliant golden VM images:
+ Provides a graphical interface to create image-building pipelines
+ Creates and maintains Amazon EC2 AMIs and on-premises VM images
+ Produces secure, validated, and up-to-date images
+ Enforces version control

XI. EC2 instance type configuration

- An EC2 instance type defines the configuration of CPU, memory, storage, and network performance

Instance Type	vCPU	Memory	Storage	Network performance
m5d.large	2	4 GB	1x50 NVMe SSD	Up to 10 Gpbs
m5d.xlarge	4	8 GB	1x100 NVMe SSD	Up to 10 Gpbs
m5d.8xlarge	32	128 GB	2x600 NVMe SSD	10 Gbps

- Instance types named incluing components Family, Generation, Processor Family, Additional Capabilities, and Size

XII. Amazon E2 Pricing Options

- 12 months free:
  + 750 hours per month of t4g.small instance dependent on region
  + 750 hours per month of Linux, RHEL, or SLES t2.micro or r3.micro instance dependent on region
  + 750 hours of Windows t2.micro or t3.micro instance dependent on region
- Amazon EC2 Pricing Model: Amazon EC2 provides the following purchasing strategies to help users optimize their costs based on their needs:
  + **Purchase Models**: Emphasis is on providing big saving through different use cases
  + **Capacity reserved models:** Emphasis is on providing reserved instances to guarantee that user have them when they need
  + **Dedicated models:** Emphasis is on providing dedicated hardware that will help user's meet compliance and regulation requirements
- Amazon EC2 Purchase Model:
  + On-Demand:
    * Pay for compute capacity by the second or by the hour with no long-term commitments
    * Recommended use cases: Spiky workloads and Experimentation workloads
  + Reserved:
    * Make a 1-yeat or 3-year commitment and receive a significant discount off on-demand prices
    * Recommended use cases: Committed workloads and Steady-rate workloads
  + Saving Plans:
    * Same discounts as Reserved Instances with more flexibility in exchange for $/hour commitment
    * Recommended use cases: All Amazon EC2 workloads and Amazon EC2 workloads that might flexibility with committed usage

+ Amazon EC2 Spot:
  * Spare Amazon EC2 capacity at a sustainable savings off the On-Demand Instance prices
  * Recommended use cases: Fault-tolerance workloads, Flexible workloads, and Stateless workloads

- Amazon EC2 Capacity Reservations: Capacity Reservations let users reserve compute capacity for Amazon EC2 instances in a specific AZ:

On-Demand Capacity Reservations	Amazon EC2 Capacity Blocks for ML
This guarantees that user always have access to EC2 capacity when user need it, for as long as user need it	Reserve GPU instances for a future date to run any of user's machine learning (ML) workloads
Recommended use cases: + Workloads that need to meet regulatory requirements for high availability + Workloads that require capacity assurance	Recommendd use cases: + Training and fine-tuning ML models + Running experiments and building prototypes + Planning for future surges in demand for ML applications

- Amazon EC2 dedicated options: Amazon EC2 dedicated options provide EC2 instance capacity on physical servers that are dedicated for user's to use (single-tenant hardware)

Dedicated Instances	Dedicated Hosts
• Pre-instance billing • Automatic Instance implement • Isolates the hosts that tun user's instance	• Pre-host billing • Visibility of sockets, cores, and host ID • Affinity between a host and an instance • Targeted instance placement • Add capacity by using an allocation request • Lets user to use server-bound software licenses and address compliance requirement