aws academy

CSCI 5902 Adv. Cloud Architecting
Fall 2023
Instructor: Dr. Lu Yang

Modules 4 Adding a Compute Layer (Sections 7 - 8)
Oct 6, 2023

Housekeeping items and feedback

1. Start recording
2. Change your emails to real names on AWS Academy Cloud Architecting course
3. Questions from the last lecture:
   - Is instance store encrypted by default?

   The data on NVMe instance storage is encrypted by default using an XTS-AES-256 block cipher implemented in a hardware module on the instance. The encryption keys are generated using the hardware module and are unique to each NVMe instance storage device. All encryption keys are destroyed when the instance is stopped or terminated and cannot be recovered. You cannot disable this encryption and you cannot provide your own encryption key.
   https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/data-protection.html
   https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/instance-store-volumes.html

Module 4: Adding a Compute Layer

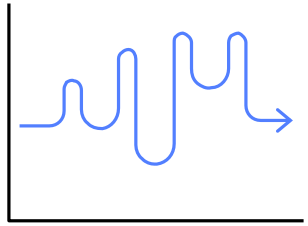
Section 7: Amazon EC2 pricing options

aws academy

Amazon EC2 pricing options (1 of 2)

On-Demand Instances

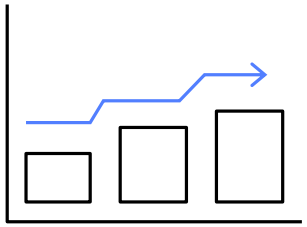
Pay for compute capacity
by the second or by the hour with no long-term commitments.



Spiky workloads,
workload experimentation

(Scheduled) Reserved Instances

Make a 1-year or 3-year commitment and receive a significant discount off on-demand prices.

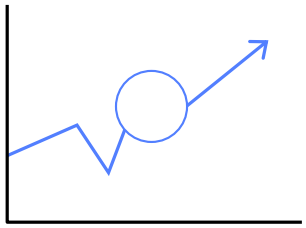


Committed and
steady-state workloads

Amazon EC2 pricing options (2 of 2)

Spot Instances

Spare Amazon EC2 capacity at substantial savings off On-Demand Instance prices.

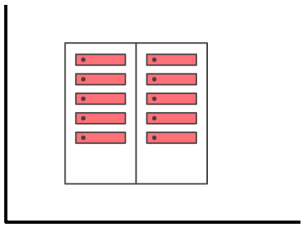


Fault-tolerant, flexible, stateless workloads

The average frequency of interruption across all regions and instance types is <5%

Dedicated Hosts

Physical server with Amazon EC2 instance capacity fully dedicated for your use.



Workloads that require the use of your own software licenses or single tenancy to meet compliance requirements

Amazon EC2 dedicated options

Amazon EC2 dedicated options provide EC2 instance capacity on physical servers that are dedicated for your use (single-tenant hardware).

Dedicated Instances

- Per-instance billing
- Automatic instance placement
- Benefit – Isolates the hosts that run your instances

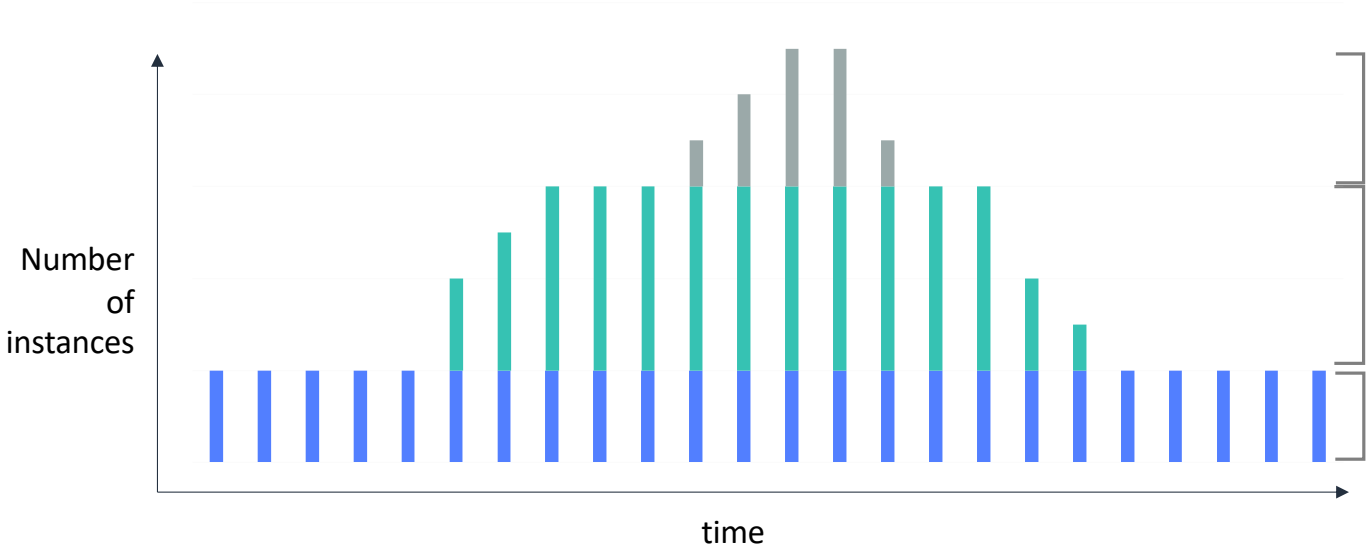
Which one is more expensive?

Dedicated Hosts

- Per-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
- Benefit – Enables you to use your server-bound software licenses and address compliance requirements

Amazon EC2 cost optimization guideline

To optimize the cost of Amazon EC2 instances, combine the available purchase options.



Demonstration: How to launch Spot Instances in AWS

https://www.youtube.com/watch?v=2ludC91LeMc (7:41 - end)

Section 7 key takeaways

- Amazon EC2 pricing models include On-Demand Instances, Reserved Instances, Spot Instances, and Dedicated Hosts
- Per-second billing is available only for On-Demand Instances, Reserved Instances, and Spot Instances that run Amazon Linux or Ubuntu
- Use a combination of Reserved Instances, On-Demand Instances, and Spot Instances to optimize Amazon EC2 compute costs

Module 4: Adding a Compute Layer

Section 8: Amazon EC2 considerations

aws academy

Placement groups

Placement groups enable you to control where instances run in an Availability Zone.

- They influence where a group of interdependent instances run –
  - Increase network performance between them
  - Reduce correlated or simultaneous failure
- Placement strategies –
  - Cluster
  - Partition
  - Spread
- Limitations –
  - An instance can be launched in only one placement group at a time
  - Instances with a tenancy of *host* cannot be launched in a placement group
  - Only certain types of instances can be launched in a placement group (compute optimized, GPU, memory optimized, storage optimized)
  - You cannot merge placement groups
  - You can move an existing instance into a placement group. The instance must be stopped and moved. You can only move or remove an instance using AWS CLI or AWS SDK, but you cannot do this via the console.

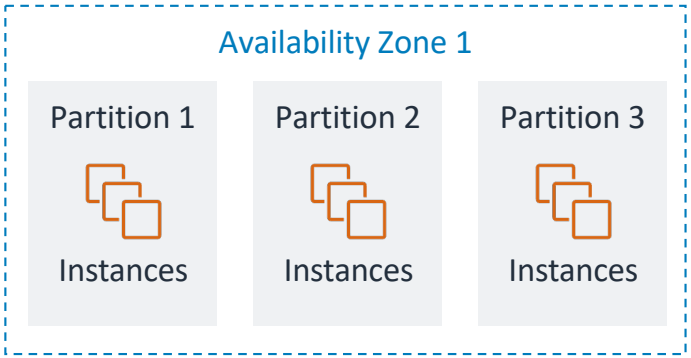
Cluster placement group

Cluster placement groups provide low-latency and high packet-per-second network performance between instances in the same Availability Zone and same rack.

- Instances are placed in the same high-bisection bandwidth segment of the network
- Provides per-flow throughput limit of up to 10 Gbps for TCP/IP traffic
- Recommended for applications that benefit from low network latency, high network throughput, or both
- AWS recommend homogenous instances within cluster
- Best practice – Launch all instances in a single request

Partition placement group

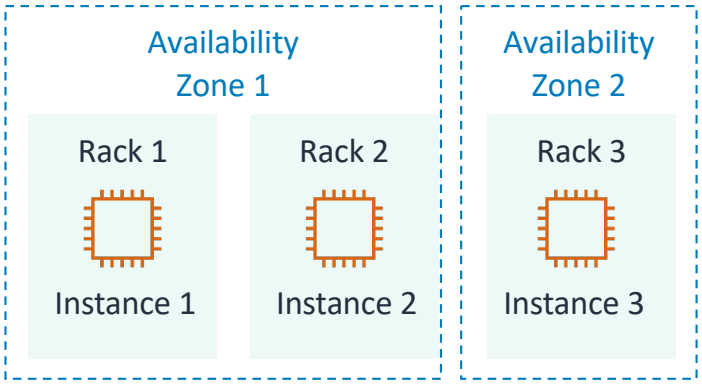
A partition placement group spreads instances across logical partitions to reduce the likelihood of correlated hardware failure.



- Each partition has its own set of racks (network and power source)
- Each rack has its own network and power source
- Partitions can be in multiple Availability Zones
- They are recommended for large distributed and replicated workloads

Spread placement group

Spread placement groups place instances across distinct physical racks to reduce correlated hardware failure.



- Each rack has its own network and power source
- Group can span multiple Availability Zones
- Each EC2 has its own hardware
- They are recommended for applications that have a small number of critical instances that should be kept separate from each other

Professional level (1/2)

- EC2 included metrics
  - Disk I/O: Read and Write only for **instance store**
    How about EBS? Use CloudWatch (https://www.datadoghq.com/blog/ec2-monitoring/#disk-io-metrics)
    CloudWatch offers a set of EBS disk I/O metrics within the EC2 namespace, but these are only available for C5 and M5 instance types. For all other instance types, disk I/O for EBS volumes must be monitored via CloudWatch's EBS metrics.
  - Network: Network In and Out
  - CPU: CPU Utilization + Credit Usage and Balance
  - Memory usage is not included
    - Customized the CloudWatch metric
- Available CloudWatch metrics for EC2 (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/viewing_metrics_with_cloudwatch.html#ec2-cloudwatch-metrics)
  - Instance metrics, CPU credit metrics, Dedicated Host metrics, Amazon EBS metrics for Nitro-based instances, Status check metrics, Traffic mirroring metrics, Auto Scaling group metrics, Amazon EC2 metric dimensions, Amazon EC2 usage metrics

Professional level (2/2)

- Instance recovery (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ec2-instance-recover.html#)
  - System check based

  If an instance becomes unreachable because of an underlying hardware failure or a problem that requires AWS involvement to repair, the instance is automatically recovered. A recovered instance is **identical** to the original instance, including
    - instance ID
    - private IP addresses
    - Elastic IP addresses
    - all instance metadata
    - public IPv4 address
    - placement group
  - CloudWatch based (https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/UsingAlarmActions.html#AddingRecoverActions)

  The following problems can cause system status checks to fail:
    - Loss of network connectivity
    - Loss of system power
    - Software issues on the physical host
    - Hardware issues on the physical host that impact network reachability

Module 4: Adding a Compute Layer

Module wrap-up

aws academy

Module summary

In summary, in this module, you learned how to:

- Identify how Amazon Elastic Compute Cloud (Amazon EC2) can be used in an architecture
- Explain the value of using Amazon Machine Images (AMIs) to accelerate the creation and repeatability of infrastructure
- Differentiate between the EC2 instance types
- Recognize how to configure Amazon EC2 instances with user data
- Recognize storage solutions for Amazon EC2
- Describe EC2 pricing options
- Determine the placement group given an architectural consideration
- Launch an Amazon EC2 instance

A real-world use case – DeepSense S3 management

Use case:

I have set up the users in our AWS project. I will be transferring data to S3 for the users soon. How do I organize the user S3 structure? How do I restrict their access to their own buckets?

Solutions:

- For our S3 bucket structure, please make it like:

ds-projects

|- OGEN (project name)

|- OnDeck

|.....

- Then, grant user (e.g., Jay) the access to the OGEN folder by setting bucket policies. We try not to make folder names using a user's id because the same project may have several users who work on it. And a project could be extended and later students could work on it again.
- Demo: https://www.youtube.com/watch?v=hTb94p_b9YY