**EC2 (Elastic compute cloud)**

Ec2 is just like a server instance which allows us to pay only for capacity that we actually use.

This is cost effective. There are 4 instance types :

**Ondemand :** Allows us to pay fixed rate by hour or by second with no commitment.

**Reserved :** Provides us with a capacity reservation for 1 or two years. There are two types, one is scheduled another is standard. If we need scheduled reserved instance for weekend we can schedule the reservation and rest of the time it can be on demand. Standard reservation is the simple reservation for a specific amount of time.

**Spot :** This is just like stock market. Enables us to bid whatever price we want for an instance capacity. If our application has flexible start and end time.

**Scheduled :** Purchase instances that are always available on the specified recurring schedule, for a one-year term.

**Capacity Reservations** – Reserve capacity for your EC2 instances in a specific availability Zone for any duration.

**Single tenant physical instance**

**Dedicated hosts:** This is useful for regulatory requirements that may not support multitenant virtualization. This is great for licensing , like if we have Oracle or mysql license , we can go for this dedicated host option. It supports ondemand and reserved pricing.

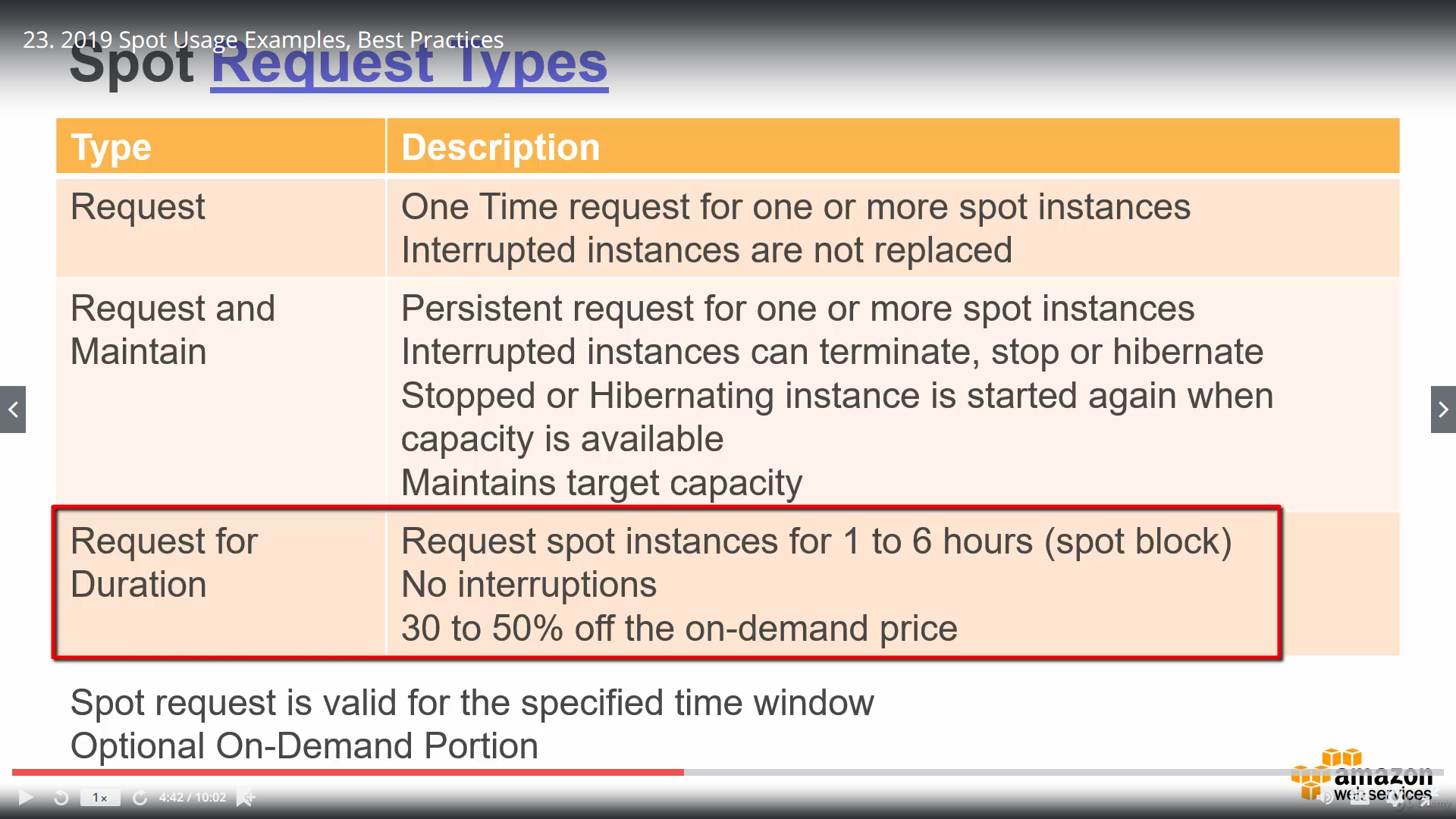
**Dedicated Instances** – Pay, by the hour, for instances that run on single-tenant hardware. It supports ondemand , resrved and spot pricing.

**Baremetal instances** : -- For specialized workload , legacy workload, licensing restrictions per host billing.Bring your own virtualization (VMware for example). Ondemand, supports reserved and spot pricing.

Note : Apart from EC2 all other services like S3, EBS all ran on multitenant servers.

**Advantage and disadvantages of spot pricing ?**

Spot instance per hour pricing is much lesser than ondemand price. But when the price goes higher rate the instance may stop with 2 minuet prior notice. There are different request types we van mention for spot pricing.





**Which instance family can have credit cpu utilization ?**

The burstable instance family like t2 and t3. There are two metric CPUCreditUsage and CPUCreditbalance used by t2 and t3.

**What is EBS (Elastic block store )?**

EBS is a virtual disk in cloud. Amazon EBS allows us to store volumes and attach them to EC2 instance. once attached we can run database, install other application on top of EBs. EBS volumes are placed in specific availability zone where they are automatically replicated to protect us from the failure of single component.

**Different EBS types**

General Purpose SSD(GP2) :3 IOPS per GB with upto 10k IOPS.

Provisioned IOP SSD(101) : Designed for I/O intensive applications such as large relational or Nosql database. Use if we need more than 10k IOPs.

Throughput optimized HDD(ST1) : Used for bigdata, warehousing, log processing. This drive is not bootable. Just like our D drive in os.

Cold HDD(Sc1) : This is comparatively lower cost than ST1, , normally used for file server. This is also not bootable.

Magnetic HDD : This is legacy harddisk, but can be bootable.

**What is security group ?**

A *security group* acts as a virtual firewall for your instance to control inbound and outbound traffic. When you launch an instance in a VPC, you can assign up to five security groups to the instance.

* We can add multiple security group and delete security group from in one instance.
* default security group cannot be deleted by the user, but we can edit the default group.
* (Inbound rules only) The source of the traffic and the source port or port range. The source can be another security group, an IPv4 or IPv6 CIDR block, or a single IPv4 or IPv6 address.
* (Outbound rules only) The destination for the traffic and the destination port or port range. The destination can be another security group, an IPv4 or IPv6 CIDR block, a single IPv4 or IPv6 address, or a prefix list ID ( A service is identified by a prefix list—the name and ID of a service for a Region).
* All inbound traffic are blocked by default.
* All outbound traffic is allowed by default
* Changes to security groups take effect immediately.
* We can have any number of ec2 instance in a security group.
* Security groups are state full means if we create an inbound rule allowing traffic in , that traffic is automatically allowed back out again.
* We cannot block specific ip address, instead use network access control list.
* We can allow rules but cannot define rules, means if we have multiple security group

consolidated allow list is applicable on the instance.

**Relation between region and availability zone**

Amazon EC2 is hosted in multiple locations world-wide. These locations are composed of Regions and Availability Zones. Each *Region* is a separate geographic area. Each Region has multiple, isolated locations known as *Availability Zones*. Amazon EC2 provides you the ability to place resources, such as instances, and data in multiple locations. Resources aren't replicated across Regions unless you do so specifically.

Each Region is completely independent. Each Availability Zone is isolated, but the Availability Zones in a Region are connected through low-latency links. The following diagram illustrates the relationship between Regions and Availability Zones.


    Regions and Availability Zones
   

**Can we keep our instance and volume in separate availability zone ?**

Inuse volume which are connected to a EC2 instance should have same availability as EC2.

Avaialble volume which are not connected with any EC2 instance can be created in any availability zone.

**Can I move a EC2 volume from one availability zone to another zone ?**

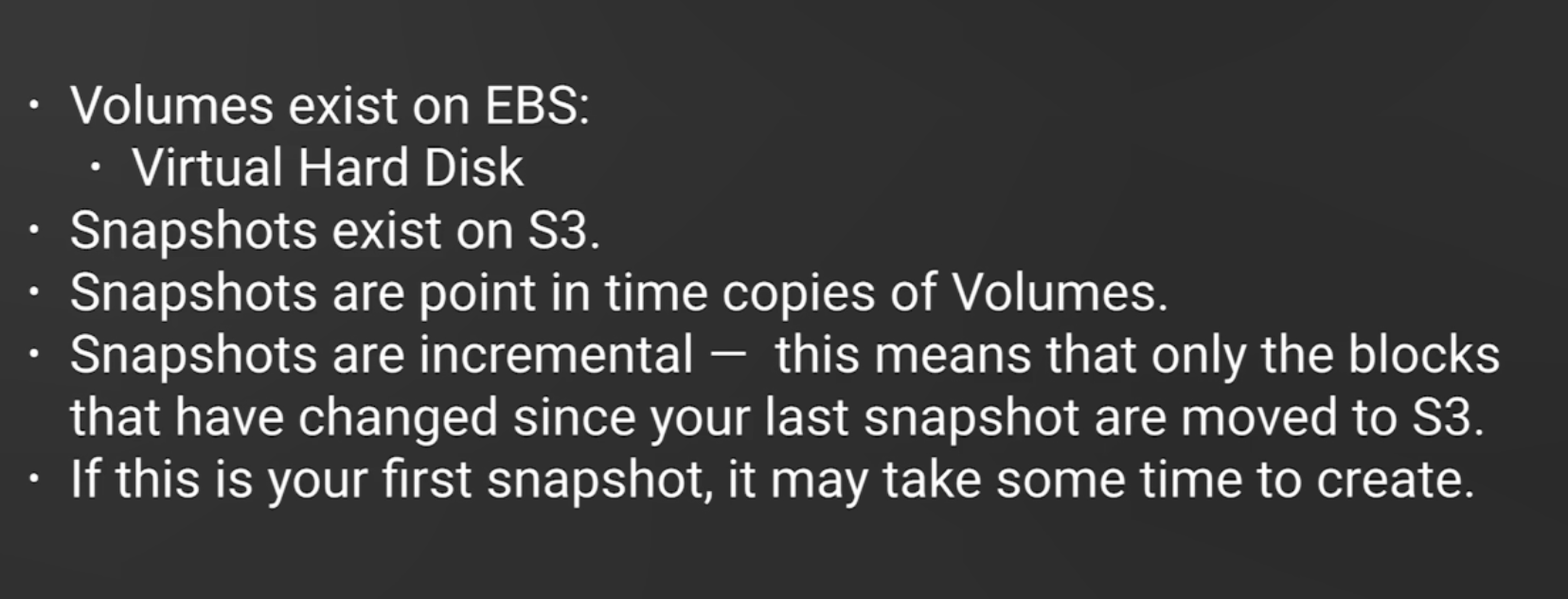
No we cannot modify availability zone of an ebs volume. For that we need to create a snapshot of the volume, then create a new volume from the snapshot in different availability zone.

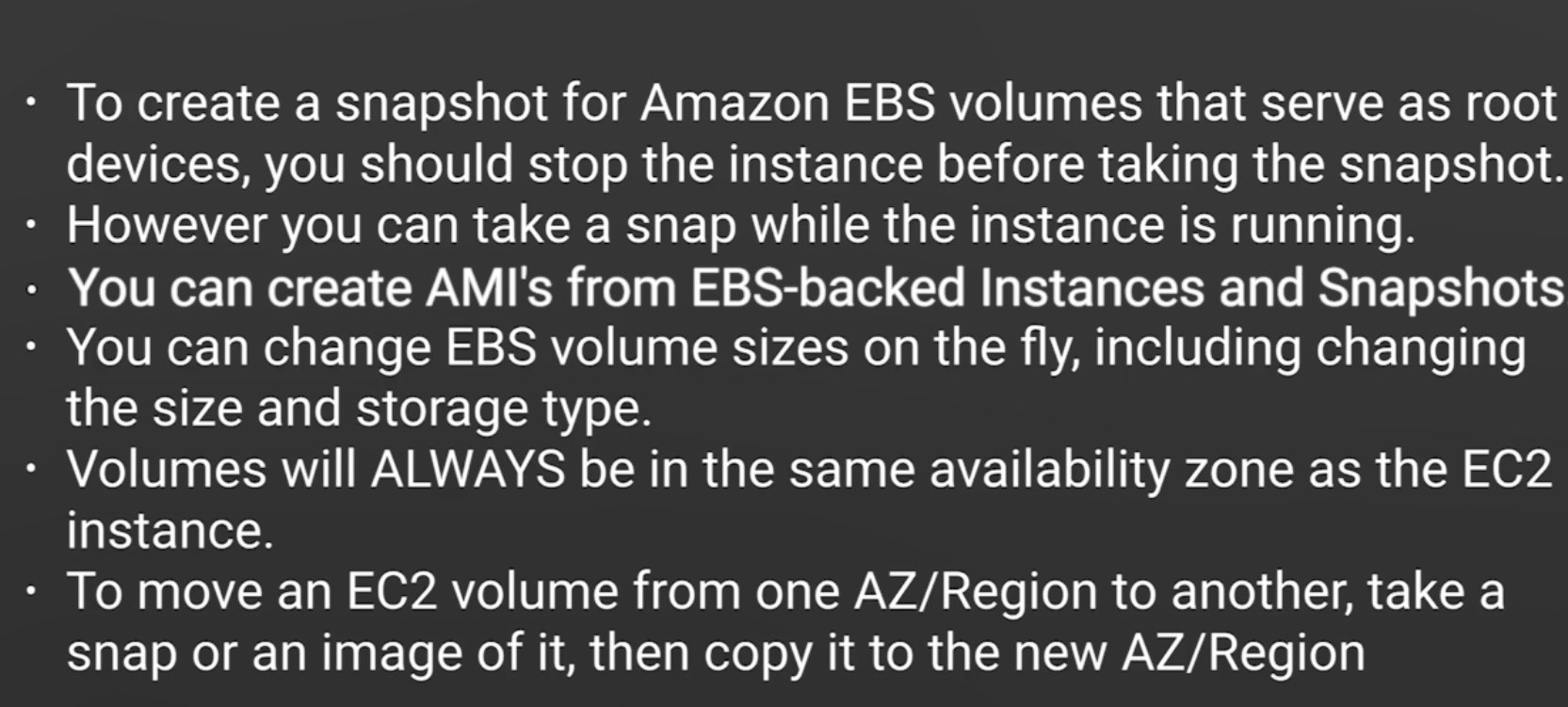
**Can we copy ec2 volume to another region ?**

No like availability zone, create a snapshot for the volume, and select the action copy to create same volume in another region.

We can create an AMI or amazon image from the snapshot. From this Ami we can launch an instance.

**Rules for snapshot,volume**



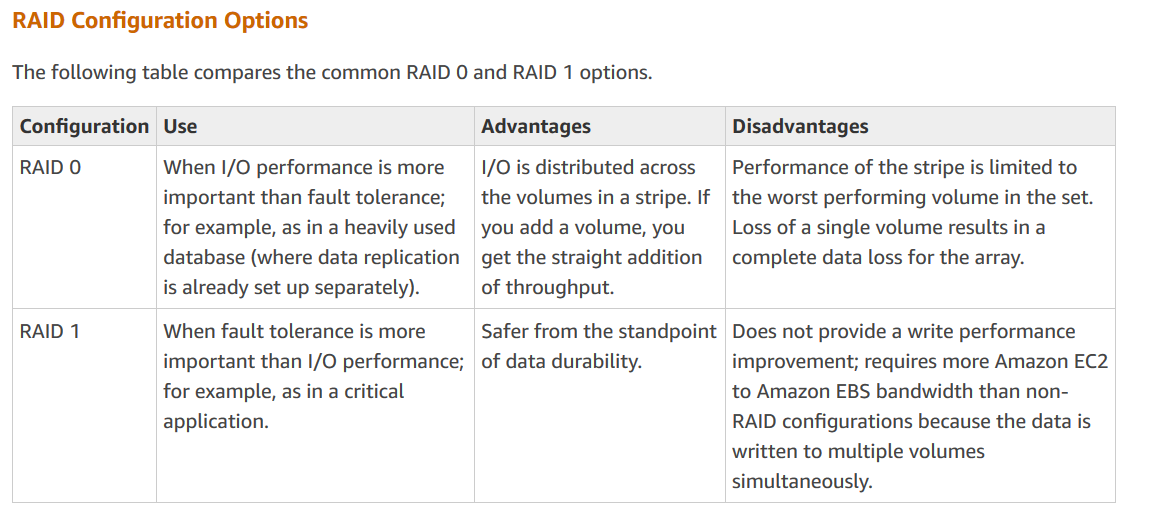


**What is RAID ?**

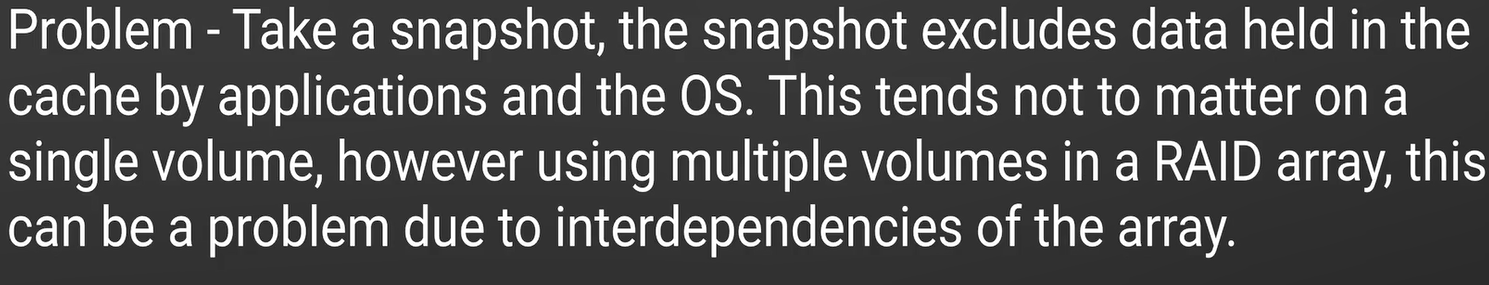
RAID is redundant array of independent disks.

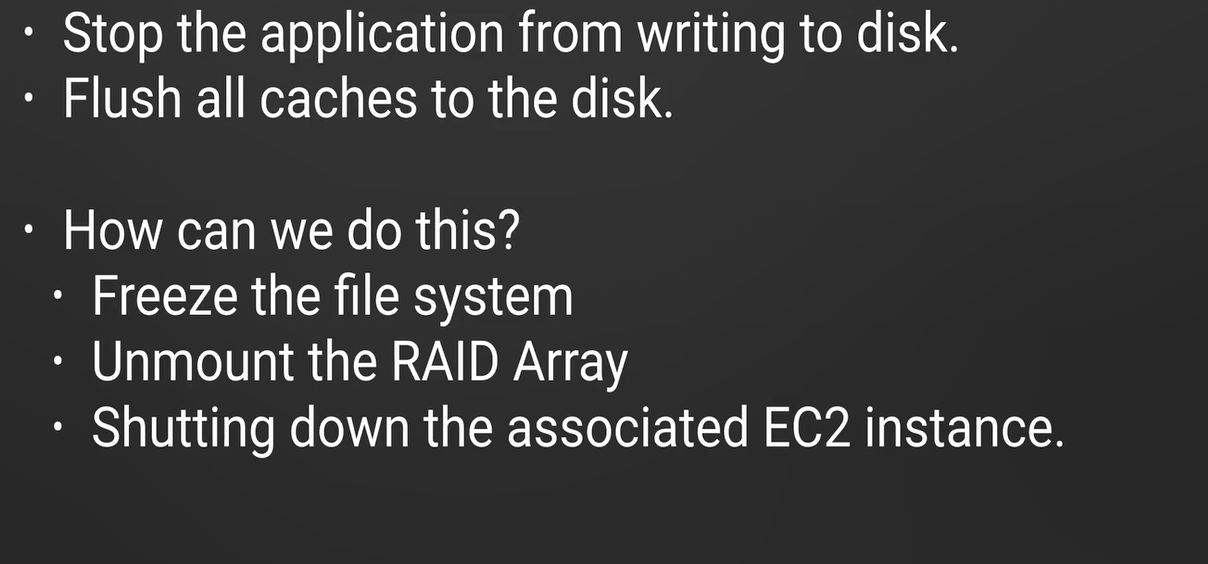
**Why use RAID ?**

When the amount of disk space is not enough, then we can use array of disk to increase the space. Suppose for an EC2 instance we are using one root volume and 3 more ebs volume each is having 10 gb diskspace. But some software we need to install is more than 10 gb, This time we can create a array of disk, that will be 30gb, and this array of disk or RAID will act as a single disk of 30gb, so diskspace and IOPs can be increased. **Follow the video as it is not a free service**



**Problem of taking snapshots of RAID ?**



**Solution :**

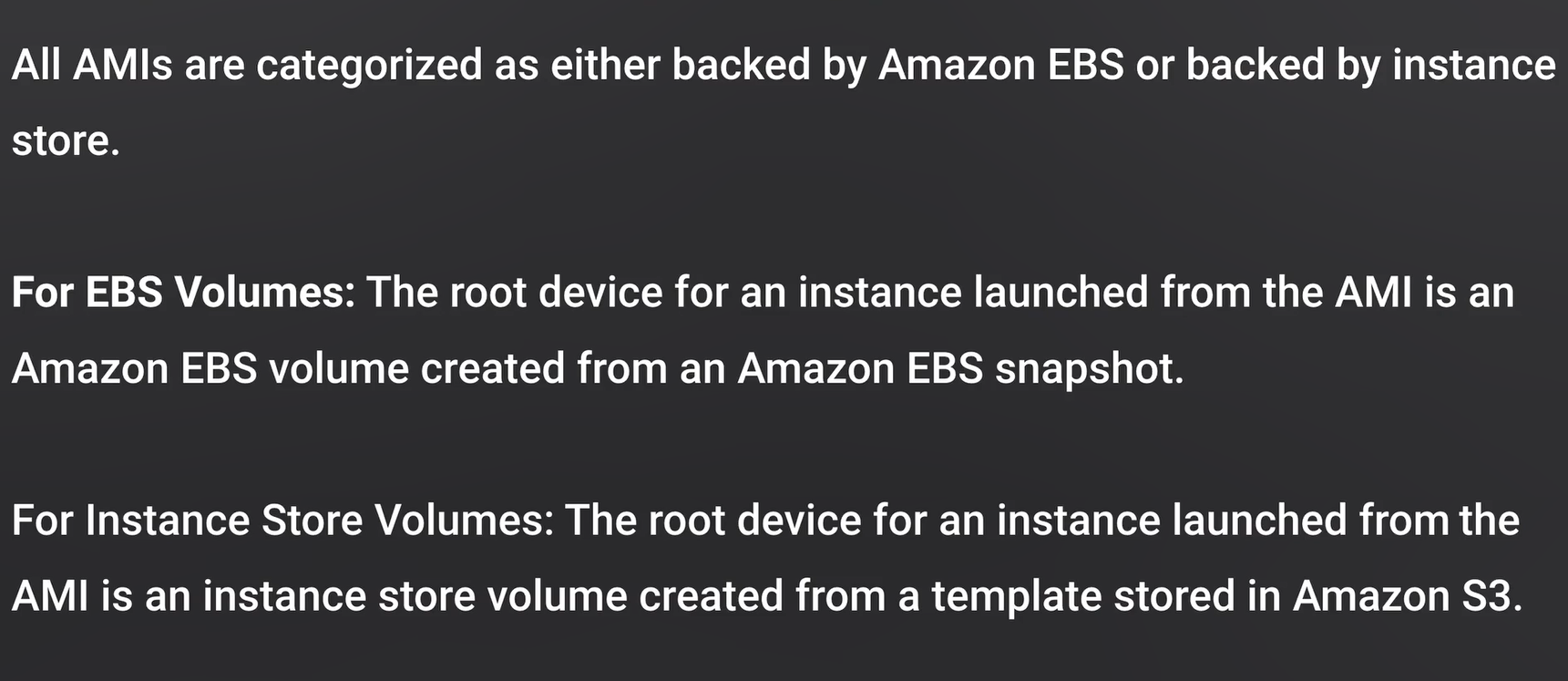
**Why aws recommend us to stop the instance and take snapshots ?**

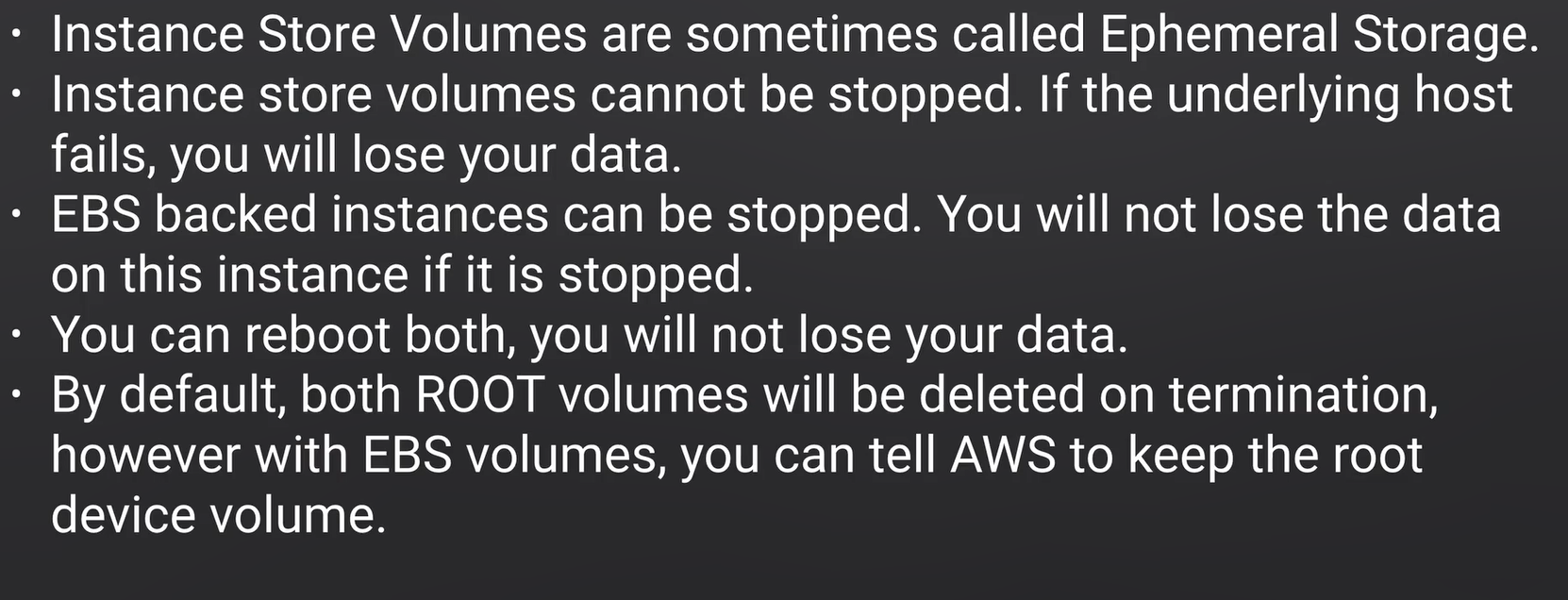
To avoid inconsistency. By the time we create a snapshot some more data can be created in the instance and the corresponding volumes.

**Can we share an encrypted snapshot with another aws account ?**

No , only unencrypted snapshots can be shared between two aws account, because in that case sse keys are unware by the other aws account.

**Difference between instance store and EBS volume ?**





**What is Elastic load balancer ?**

Elastic Load Balancing supports the following types of load balancers: Application Load Balancers, Network Load Balancers, and Classic Load Balancers. Amazon ECS services can use either type of load balancer. Application Load Balancers are used to route HTTP/HTTPS (or Layer 7) traffic. Network Load Balancers and Classic Load Balancers are used to route TCP (or Layer 4) traffic.

**Application load balancer :**

Best suited for http and https traffic, works at application layer (7th layer in OSI model).

**Network load balancer :**

These are best suited for load balancing of TCP traffic where extreme performance is needed. This balancer operates at layer 4(Connection layer).

**Classic load balancer :**

You can load balance http/https applications in layer 7. It has layer specific feature like x-forwarded sticky session. It can also works at layer 4 as well. This is legacy load balancer.

**Advantages of application load balancer ?**

1. Rules based mapping , it is very useful for microservices where each instance group having different services.
2. SSL offloading and performance is little faster than old classic load balancer.

architecture of application and network load balancer components


                The components of a basic Application Load Balancer
            

**What is target group ?**

Each target group is used to route requests to one or more registered targets. When you create each listener rule, you specify a target group and conditions. When a rule condition is met, traffic is forwarded to the corresponding target group. You can create different target groups for different types of requests. For example, create one target group for general requests and other target groups for requests to the microservices for your application. You define health check settings for your load balancer on a per target group basis. Each target group uses the default health check settings,

Different types of target is acceptable

instance

The targets are specified by instance ID.

ip

The targets are IP addresses. You can't specify publicly routable IP addresses rather use CIDR range

lambda

The target is a Lambda function.

**What is x-forwarded-for header ?**

The X-Forwarded-For (XFF) header is a de-facto standard header for identifying the originating IP address of a client connecting to a web server through an HTTP proxy or a load balancer. When traffic is intercepted between clients and servers, server access logs contain the IP address of the proxy or load balancer only. To see the original IP address of the client, the X-Forwarded-For request header is used.

Syntax : X-Forwarded-For: <client>, <proxy1>, <proxy2>

<client> is the client address

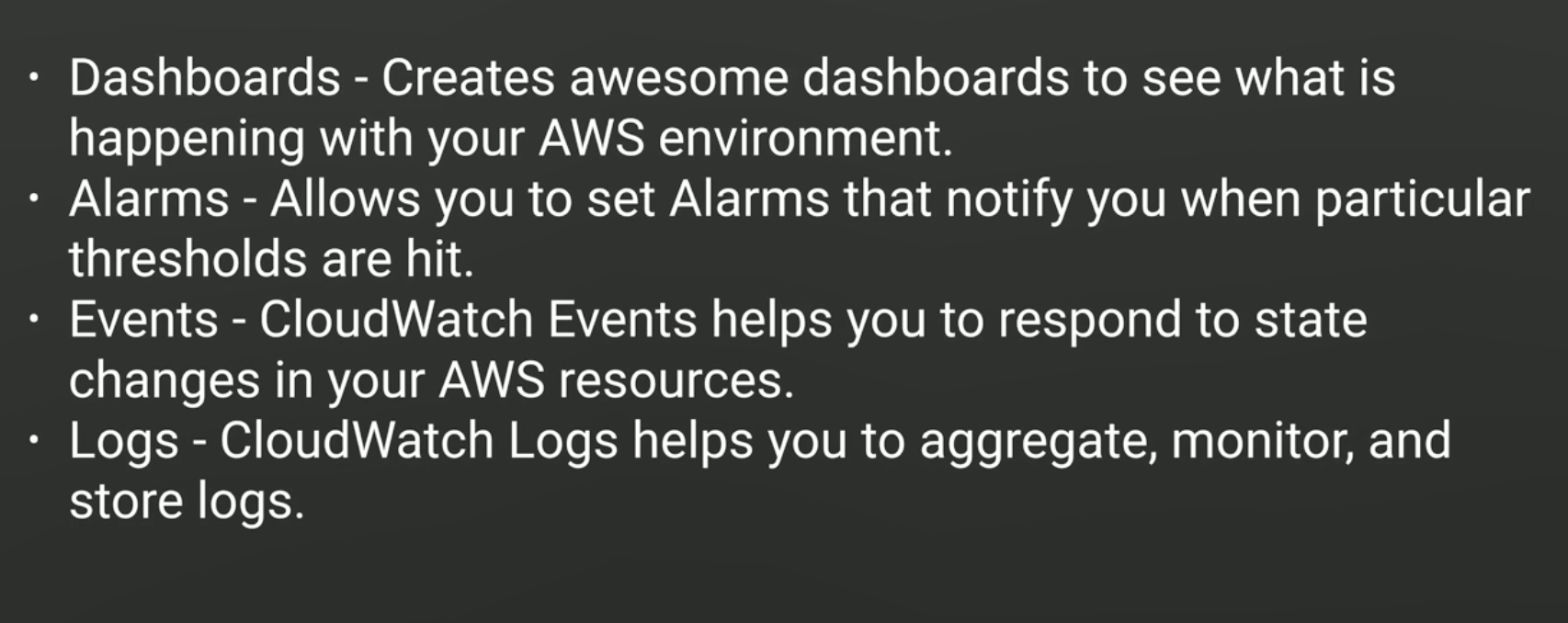
proxy1,proxy2 are the address of different load balancer or proxy server.

X-Forwarded-For: 203.0.113.195

X-Forwarded-For: 203.0.113.195, 70.41.3.18, 150.172.238.178

for the first option only client address is mentioned in the header.

**What can we do with cloud watch ?**



**What are the different metrics in cloud watch ?**

Cpu realted,

Disk related

Network related,

Status check related.

**What is the difference between CoudWatch and cloudtrail ?**

Cloudtrial is used for monitoring the API calls made to a particular service or Application. It is primarily used to monitor API calls and is applicable for a selected services only.  
  
CloudWatch is used for logging events that happen on any particular AWS service. It is the default logging service provided by AWS and can be configured to detect Alarm conditions such as High CPU, Low Disk Space, Network Parameters etc.

**How to secure ec2 instance ?**

If we login to ec2 instance by CLI,

sudo su

cd ~

cd .aws [this is a hidden diretory]

Here we can see a credential folder and inside it all the access key and access key id are stored.

To avoid this, try to attach a role with each ec2 instance.

If we attach a role we can’t see any credentials folder inside .aws

**How to copy an S3 bucket from one region to another region ?**

sudo su

aws s3 cp –recursive s3://acloudguru-useast1/home/ec2-user

It will not work for all the regions, acloudguru-useast1 is the fullname of the bucket

aws s3 cp –recursive s3://acloudguru-useast1/home/ec2-user –region us-east1

**How to check metadata for an ec2 instace from CLI ?**

curl <http://169.254.169.254/latest/meta-data>

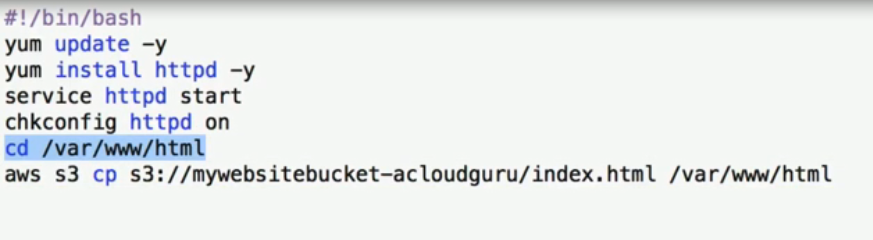
If we want to print the public-ipv4 into a html file

curl <http://169.254.169.254/latest/meta-data/public-ipv4> > publicipV4.html

user-data is the bootstartp scripts

curl <http://169.254.169.254/latest/meta-data/user-data>

**Some bootstrap script**



#!/bin/bash to denote it is a bash script

yum update –y will update the instance with latest jar

yum install httpd –y will install apache

service httpd start will start apache server

chkconfig httpd on will check the server is on every time we open CLI

cd /var/www/html is a simple directory inside apache server.

last line is going to copy index.html file from s3 bucket to /var/www/html folder inside EC2 instance.

**Where the EC2 metadata stored ?**

In Ec2 cli go to ec2 instance.

*curl* [*http://169.254.169.254/latest/meta-data*](http://169.254.169.254/latest/meta-data)

The metadatas are like public ip profile, hostname etc.

To find the user data

*curl* [*http://169.254.169.254/latest/user-data*](http://169.254.169.254/latest/user-data)

It returns the bash script written at the beginning to start the instance

**What is EC2 placement group ?**

<https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html>

When you launch a new EC2 instance, the EC2 service attempts to place the instance in such a way that all of your instances are spread out across underlying hardware to minimize correlated failures. You can use placement groups to influence the placement of a group of interdependent instances to meet the needs of your workload

**Different types of placement groups ?**

**Clustered placement group :**

A cluster placement group is a logical grouping of instances within a single Availability Zone. A placement group can span peered VPCs in the same Region.


                A cluster placement group
            

Cluster placement groups are recommended for applications that benefit from low network latency, high network throughput, or both, and if the majority of the network traffic is between the instances in the group

**Partition placement group :**

When using partition placement groups, Amazon EC2 divides each group into logical segments called partitions. Amazon EC2 ensures that each partition within a placement group has its own set of racks. Each rack has its own network and power source. No two partitions within a placement group share the same racks, allowing you to isolate the impact of hardware failure within your application.


                A partition placement group with three partitions
            

Partition placement groups can be used to deploy large distributed and replicated workloads, such as HDFS, HBase, and Cassandra, across distinct racks. When you launch instances into a partition placement group, Amazon EC2 tries to distribute the instances evenly across the number of partitions that you specify. You can also launch instances into a specific partition to have more control over where the instances are placed.

A partition placement group can have partitions in multiple Availability Zones in the same Region. A partition placement group can have a maximum of seven partitions per Availability Zone. The number of instances that can be launched into a partition placement group is limited only by the limits of your account.

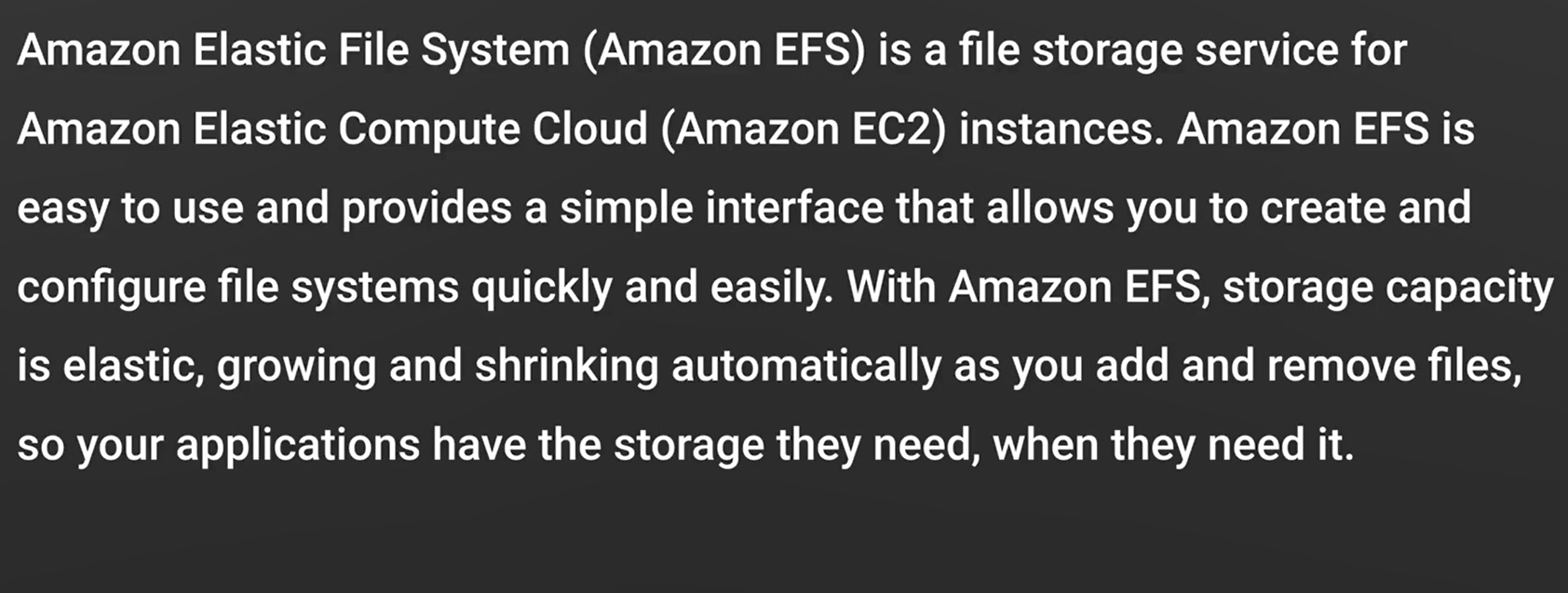
**Spread placement group :**

A spread placement group is a group of instances that are each placed on distinct racks, with each rack having its own network and power source.


                A spread placement group
            

A spread placement group can span multiple Availability Zones in the same Region. You can have a maximum of seven running instances per Availability Zone per group.

**What is EFS?**



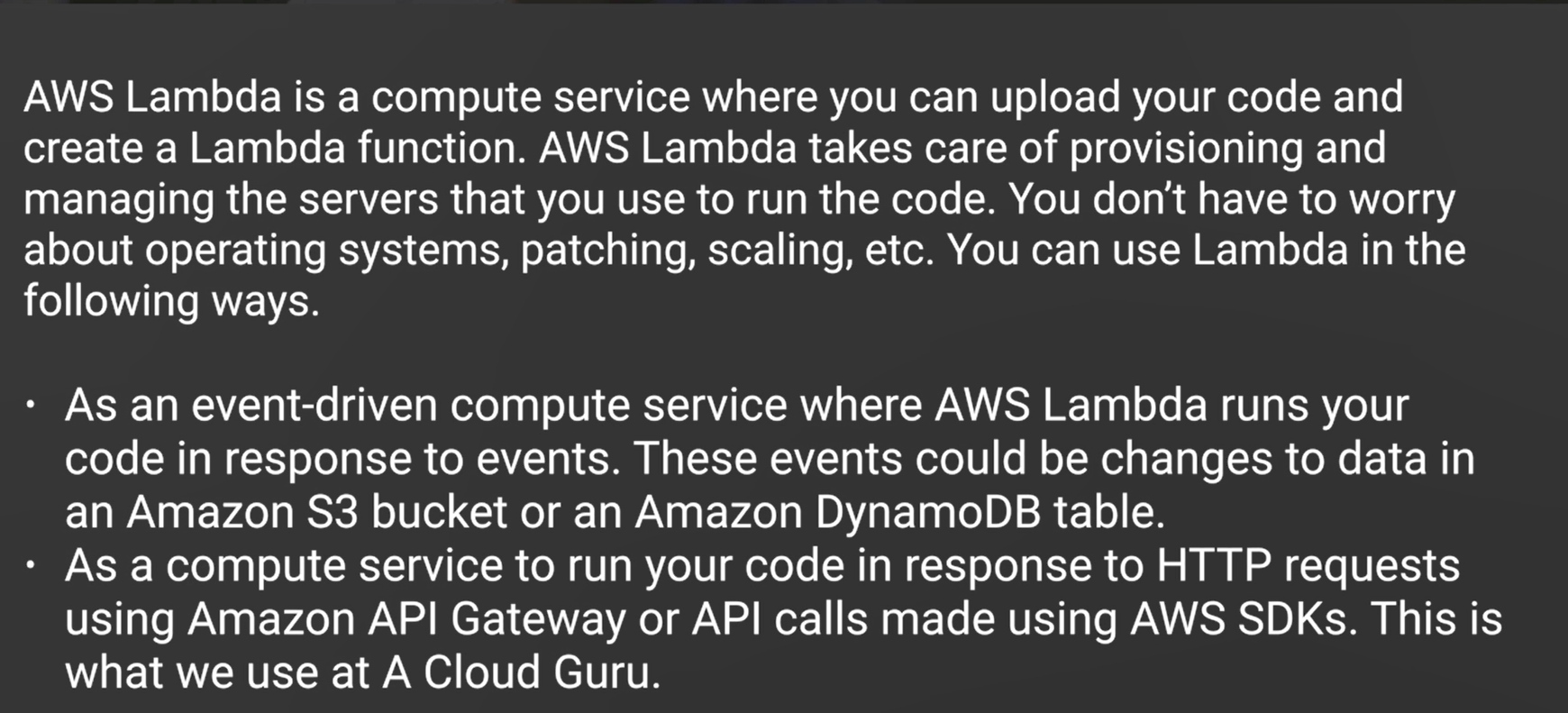
**What if EFS region and instance region does not match ?**

Make sure that your EFS and Ec2 instance region should be same. and they should use same vpc and same security group. Otherwise we can’t mount our file system into the instance servers.

mount.nfs4: mount point var/html does not exist

**Please follow the video of EFS**

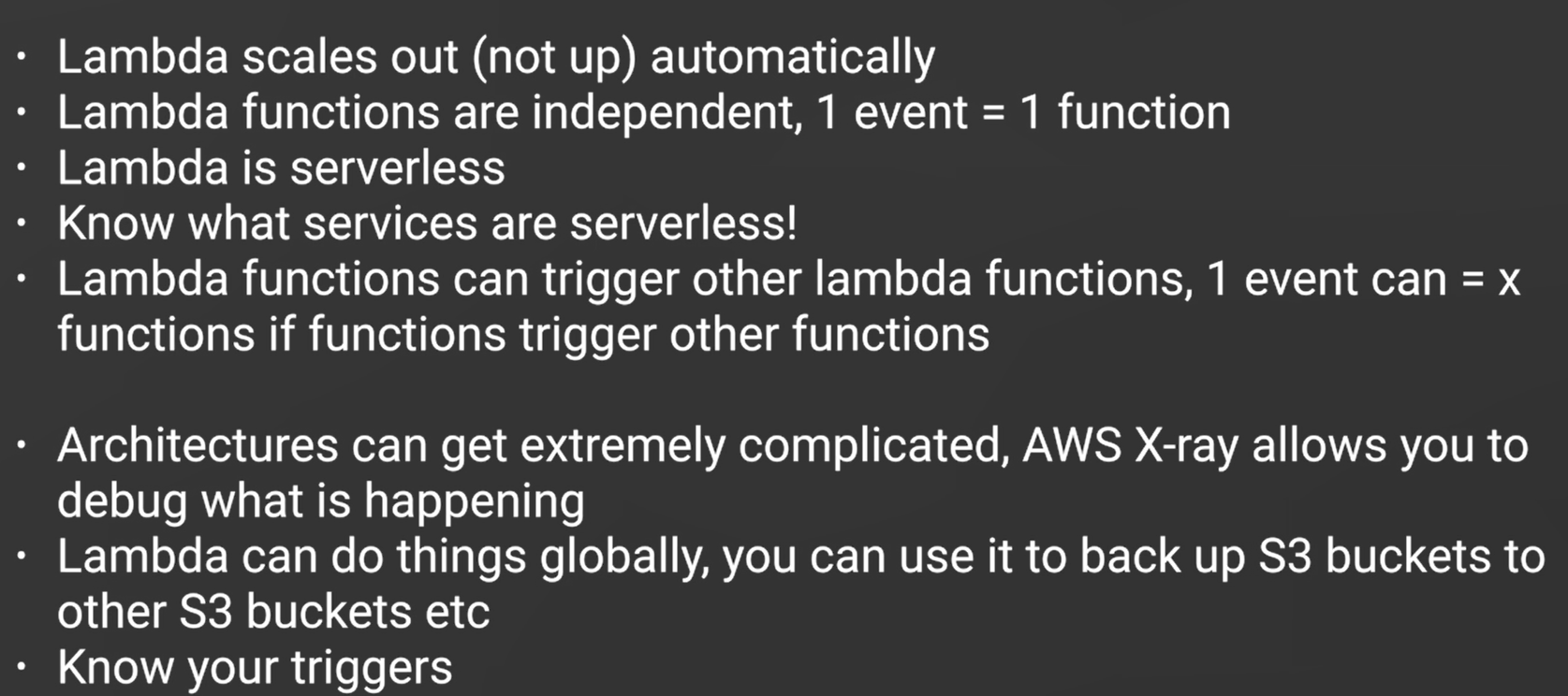
**What is lambda ?**



**Difference between scaling up and scaling out in cluster system ?**

Scaling up means to increase the performance increase the Ram, the memory size or CPu speed. Scale out means increase number of instances in the cluster.

**Properties of Lambda**



**Relation between scaling out and lambda function ?**

lambda function can scaled out by number of requests. Suppose an api gateway triggers a request to lambda, then there will be only one lambda function, if there are 4 requests from 4 users, then Lambda automatically scales out and returns 4 response from 4 different lambda function of identical code base.

**can we scale out all the triggers ?**

There are different kind of triggers in lambda like API, EC2,S3, Dynamo db. A few trigger can be scaled out. As example Ec2 can’t be scaled out as it always attached to several instances. Lambda is server less.

**What are the services in aws from where lambda can be triggered ?**

Services That Invoke Lambda Functions Synchronously

* [Elastic Load Balancing (Application Load Balancer)](https://docs.aws.amazon.com/lambda/latest/dg/services-alb.html)
* [Amazon Cognito](https://docs.aws.amazon.com/lambda/latest/dg/services-cognito.html)
* [Amazon Lex](https://docs.aws.amazon.com/lambda/latest/dg/services-lex.html)
* [Amazon Alexa](https://docs.aws.amazon.com/lambda/latest/dg/services-alexa.html)
* [Amazon API Gateway](https://docs.aws.amazon.com/lambda/latest/dg/with-on-demand-https.html)
* [Amazon CloudFront (Lambda@Edge)](https://docs.aws.amazon.com/lambda/latest/dg/lambda-edge.html)
* [Amazon Kinesis Data Firehose](https://docs.aws.amazon.com/lambda/latest/dg/services-kinesisfirehose.html)
* [AWS Step Functions](https://docs.aws.amazon.com/step-functions/latest/dg/connect-lambda.html)

Services That Invoke Lambda Functions Asynchronously

* [Amazon Simple Storage Service](https://docs.aws.amazon.com/lambda/latest/dg/with-s3.html)
* [Amazon Simple Notification Service](https://docs.aws.amazon.com/lambda/latest/dg/with-sns.html)
* [Amazon Simple Email Service](https://docs.aws.amazon.com/lambda/latest/dg/services-ses.html)
* [AWS CloudFormation](https://docs.aws.amazon.com/lambda/latest/dg/services-cloudformation.html)
* [Amazon CloudWatch Logs](https://docs.aws.amazon.com/lambda/latest/dg/services-cloudwatchlogs.html)
* [Amazon CloudWatch Events](https://docs.aws.amazon.com/lambda/latest/dg/with-scheduled-events.html)
* [AWS CodeCommit](https://docs.aws.amazon.com/lambda/latest/dg/services-codecommit.html)
* [AWS Config](https://docs.aws.amazon.com/lambda/latest/dg/services-config.html)
* [AWS IoT Events](https://docs.aws.amazon.com/lambda/latest/dg/services-iotevents.html)

EC2 cannot be a Lambda trigger and a lambda function too. Lambda is calling another lambda function not triggering event.

**Which is the policy to attach with the IAM role attached with lambda ?**

Simple microservice permission.

**Which policy needs to invoke when calling another lambda function ?**

AWSLambdaRole

**What is ENI(Elastic network interface)**

An elastic network interface (referred to as a network interface in this documentation) is a logical networking component in a VPC that represents a virtual network card. You can create and configure network interfaces in your account and attach them to instances in your VPC. Your account might also have requester-managed network interfaces, which are created and managed by AWS services to enable you to use other resources and services.

**Instance and its tenancy levels ?**

Each VPC has a related instance tenancy attribute. This attribute has the following values.

|  |  |
| --- | --- |
| **Tenancy Value** | **Description** |
| default | An instance launched into the VPC runs on shared hardware by default, unless you explicitly specify a different tenancy during instance launch. |
| dedicated | An instance launched into the VPC is a Dedicated Instance by default, unless you explicitly specify a tenancy of host during instance launch. You cannot specify a tenancy of default during instance launch. |

**Which ELB component is responsible for monitoring the Load Balancers?**

Controller service

Explanantion : Answer: (A) Elastic Load Balancing (ELB) consists of two components: the load balancers and the controller service. The load balancers monitor the traffic and handle requests that come in through the Internet. The controller service monitors the load balancers, adding and removing load balancers as needed and verifying that the load balancers are functioning properly.