AWS

EC2 ASG ELB

Marvin Chen

EC2 – Introduction to EC2

Elastic Compute Cloud (EC2) is a **highly configurable server**. EC2 is **resizable compute capacity**. It takes **minutes** to launch new instances.

高度定制化，大小可调节，分钟级部署的 **计算资源**。

Anything and everything on AWS uses EC2 Instance underneath.

多数 aws **服**务的背后是由 ec2 **支撑的**

Choose your OS **via**
Amazon Machine Image (AMI)

Choose you **Instance Type**

t2.nano
$0.0065/hour ($4.75/month)
1 vCPU 0.5GB Mem

C4.8xlarge
$1.591/hour ($1161.43/month)
36 vCPU 60GB Mem 10 Gigabit performance

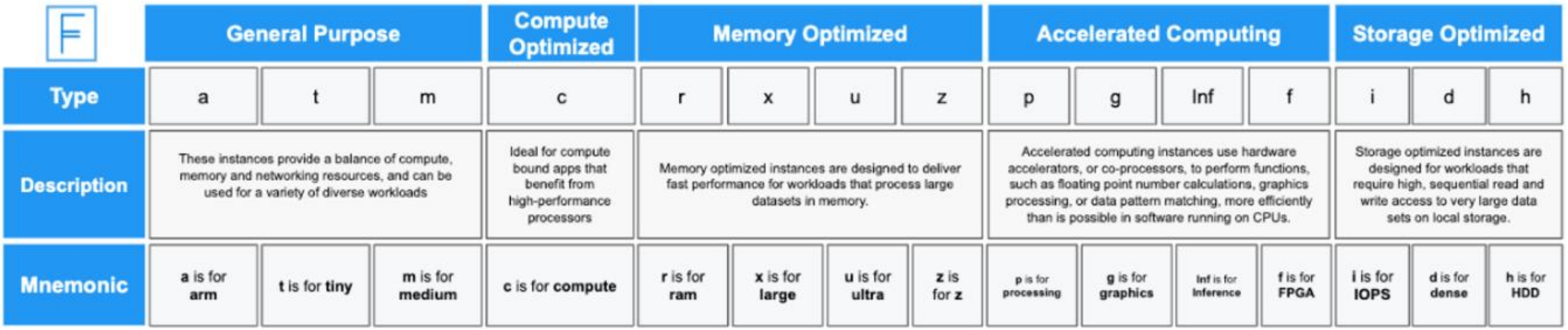
Add Storage **(EBS, EFS)**

SSD HDD Virtual Magnetic Tape Multiple Volumes

Configure your Instance

Security Groups, Key Pairs, UserData, IAM Roles, Placement Groups

EC2 – Instance Types and Usage



General Purpose	Balance of compute, memory and networking resources. 表现均衡	Webservers Code repositories
Compute Optimized	Ideal for compute bound applications that benefit from high performance processor 计算优化	Scientific modeling Dedicated gaming
Memory Optimized	Fast performance for workloads that process large data sets in memory. 内存优化	In-memory caches In-memory databases
Accelerated Optimized	Hardware accelerators, or co-processors 硬件加速	Machine learning, Speech recognition
Storage Optimized	High, sequential read and write access to very large data sets on local storage 存储优化	NoSQL datawarehosing

EC2 – Instance Sizes

EC2 Instance Sizes **generally double** in price and key attributes

Name	vCPU	RAM (GIB)	On-Demand per hour	On-Demand per month
t2.small	1	12	$0.023	$16.79
t2.medium	2	24	$0.0464	$33.87
t2.large	2	36	$0.0928	$67.74
t2.xlarge	4	54	$0.1856	$135.48

EC2 – User Data

You can provide an EC2 with **UserData** which is a script that will be automatically run when launching an EC2 instance. You could install package, apply updates or anything you like.

预配置脚本

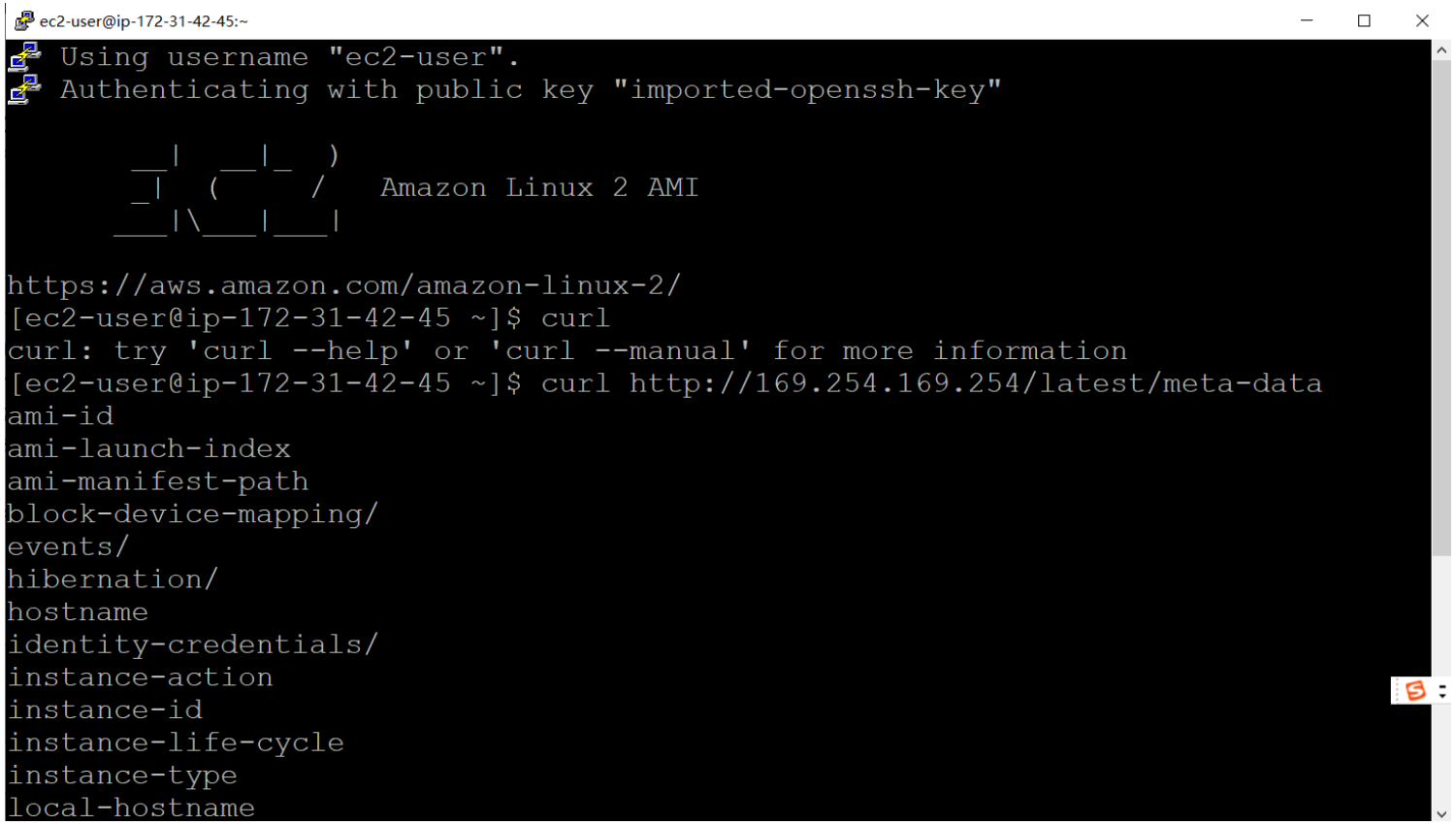
```
#!/bin/bash
yum -y install httpd
systemctl enable httpd
systemctl start httpd
echo '<html><h1>Hello From Your Web Server!</h1></html>' > /var/www/html/index.html
```

EC2 – Meta Data

From within your EC2 instance you can access information about the EC2 via a special url endpoint at 169.254.169.254. You would SSH into your EC2 instance and can use CURL command:

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

Combine metadata with user data scripts to perform all sorts of advanced AWS staging automation.



EC2 – Lab

Launch EC2 in Tokyo Region | Introduction to Amazon EC2 on **qwiklabs**

Task 1: Launch Your Amazon EC2 Instance [30 mins]

SSH connect, then check metadata

If you are windows user, please use puttygen to turn pem key to ppk key

MAC and Linux user, can directly following the instruction from EC2 console to connect

User data

```
#!/bin/bash
yum -y install httpd
systemctl enable httpd
systemctl start httpd
echo '<html><h1>Hello From Your Web
Server!</h1></html>' > /var/www/html/index.html
```

service httpd status
curl http://169.254.169.254/latest/meta-data

EC2 – Pricing Model

按需实例 竞价实例
预留实例 专用主机

On-Demand Least Commitment

- low cost and flexible
- only pay per hour
- short-term, spiky, unpredictable workloads
- cannot be interrupted
- For first time apps

Spot upto 90% Biggest Savings

- request spare computing capacity
- flexible start and end times
- Can handle interruptions (server randomly stopping and starting)
- For non-critical background jobs

Reserved upto 75% off Best Long-term

- steady state or predictable usage
- commit to EC2 over a 1 or 3 year term
- Can resell unused reserved instances

Dedicated Most Expensive

- Dedicated servers
- Can be on-demand or reserved (upto 70% off)
- When you need a guarantee of isolate hardware (enterprise requirements)

EC2 – Pricing Model

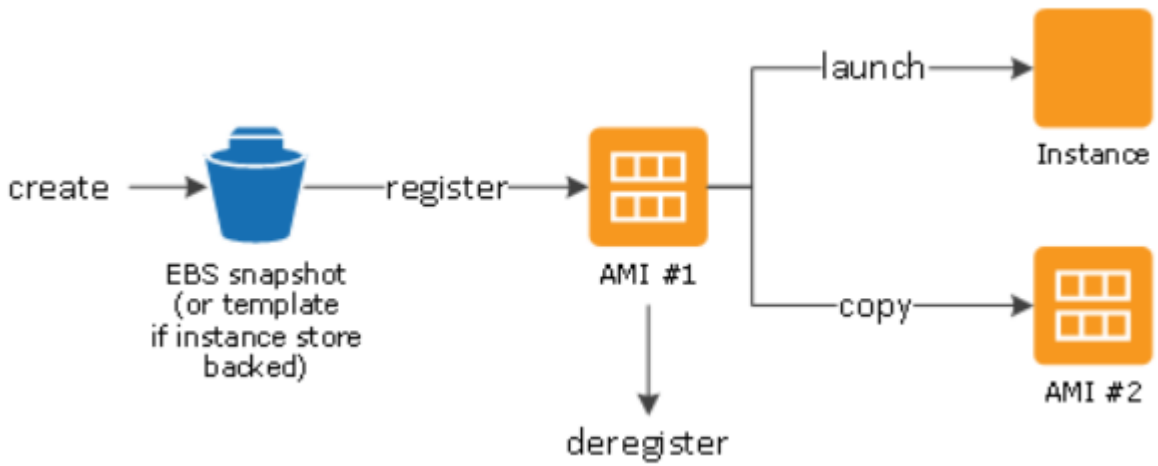
按需实例 竞价实例
预留实例 专用主机

EC2 – AMI

Amazon Machine Image (**AMI**) provides the information required to launch an instance. You can turn your EC2 instances into AMIs so you can create copies of your servers (系统镜像)

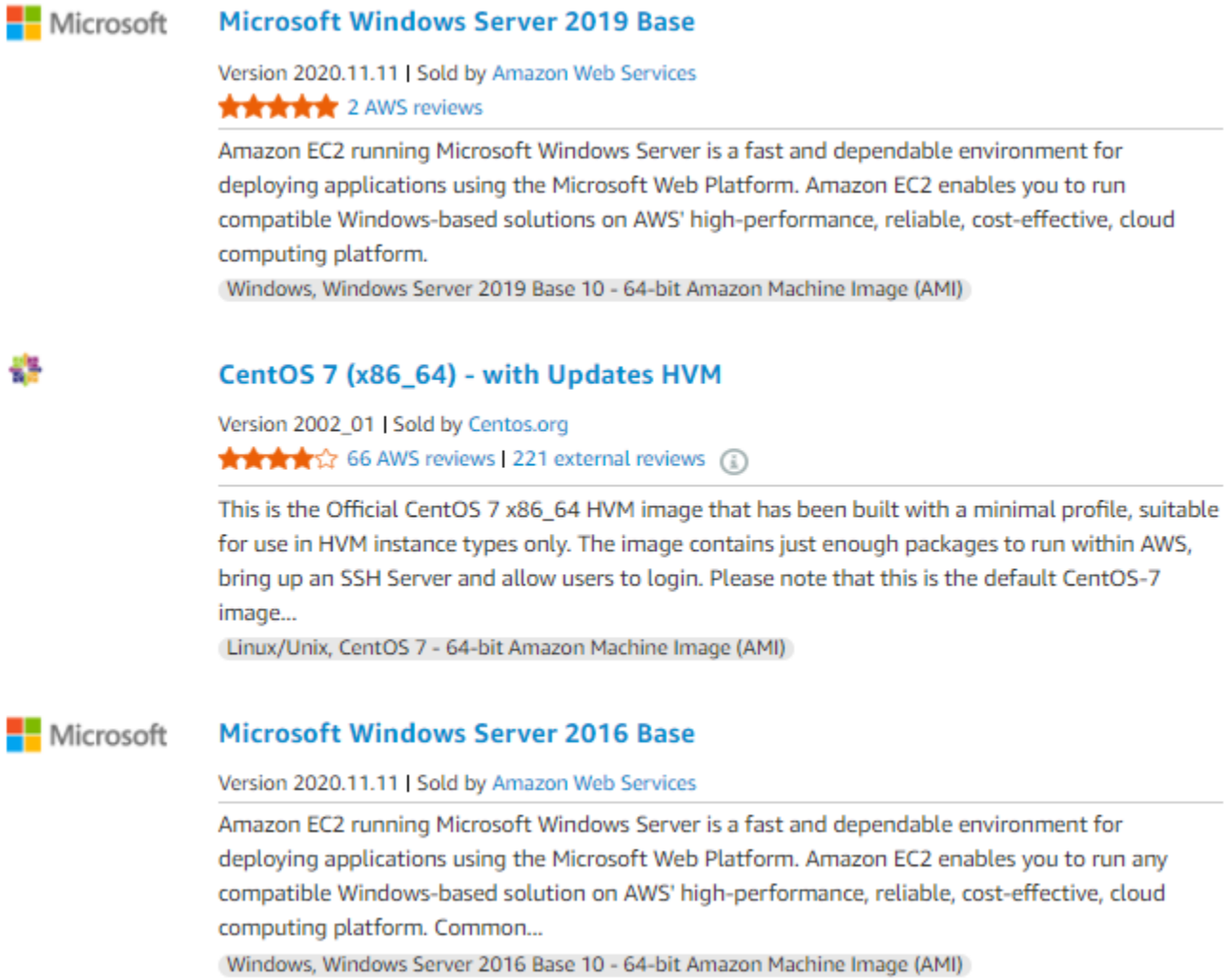
An AMI holds the following information: （AMIs are **Region** Specific）

- A template for the root volume for the instance (EBS Snapshot or instance Store template) eg. An operating system, an application server, and applications. 盘卷信息
- Launch permissions that control which AWS accounts can use the AMI to launch instances. 权限信息
- A block device mapping that specifies the volumes to attach to the instance when it's launched. 块设备信息



EC2 – AWS Marketplace

The AWS Marketplace lets you purchase subscriptions to vendor maintained AMIs.



Marketplace

可购买三方 AMI

EC2 – AMI Summary

- **Amazon Machine Image (AMI)** provides the information required to launch an instance.
- AMIs are region specific, if you need to use an AMI in another region you can copy an AMI into the destination region.
- You can create an AMI from an existing EC2 instance that's either **running** or **stopped**.
- **Community AMI** are free AMIs maintained by the community.
- **AWS Marketplace** free or paid subscription AMIs maintained by vendors.
- AMIs have an AMI ID. The same AMI eg. (Amazon Linux 2) will vary in both AMI ID and options eg. Architecture options in different regions
- An AMI holds the following information:
  - A template for the root volume for the instance (EBS Snapshot or Instance Store template) eg. An operating system, an application server, and applications.
  - Launch permissions that control which AWS accounts can use the AMI to launch instances.
  - A block device mapping that specifies the volumes to attach to the instance when it's launched.

Exams Samples

A company's application is running on Amazon EC2 instances in a single Region in the event of a disaster a solutions architect needs to ensure that the resources can also be deployed to a second Region.

Which combination of actions should the solutions architect take to accomplish this-? (Select TWO)

A. Detach a volume on an EC2 instance and copy it to Amazon S3
B. Launch a new EC2 instance from an Amazon Machine image (AMI) in a new Region
C. Launch a new EC2 instance in a new Region and copy a voiume from Amazon S3 to the new instance
D. Copy an Amazon Machine Image (AMI) of an EC2 instance and specify a different Region for the destination
E. Copy an Amazon Elastic Block Store (Amazon EBS) volume from Amazon S3 and launch an EC2 instance in the destination Region using that EBS volume

跨 region 使用 AMI 需要先将 AMI 拷贝过去，再使用

Exams Samples

A company's application is running on Amazon EC2 instances in a single Region in the event of a disaster a solutions architect needs to ensure that the resources can also be deployed to a second Region.

Which combination of actions should the solutions architect take to accomplish this-? (Select TWO)

A. Detach a volume on an EC2 instance and copy it to Amazon S3
B. Launch a new EC2 instance from an Amazon Machine image (AMI) in a new Region
C. Launch a new EC2 instance in a new Region and copy a voiume from Amazon S3 to the new instance
D. Copy an Amazon Machine Image (AMI) of an EC2 instance and specify a different Region for the destination
E. Copy an Amazon Elastic Block Store (Amazon EBS) volume from Amazon S3 and launch an EC2 instance in the destination Region using that EBS volume

跨 region 使用 AMI 需要先将 AMI 拷贝过去，再使用

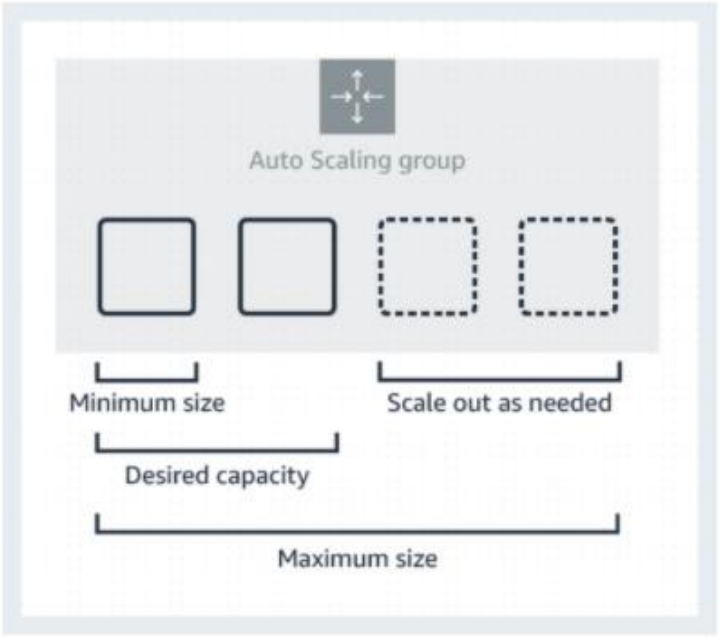
EC2 – Auto Scaling Groups

EC2 – Auto Scaling Groups

Auto Scaling Groups (**ASG**) contains a collection of EC2 instances that are treated as a group for the purposes of automatic scaling and management. **自动伸缩一组 EC2 实例**

Automatic scaling can occur via:

1. **Capacity Settings**（比如 cpu 使用率过高，扩展）
2. **Health Check Replacements**（比如 某个 instance 宕机，替换）
3. **Scaling Policies** （其他特别的要求或规则，比如娱乐网站工作时间，收缩）

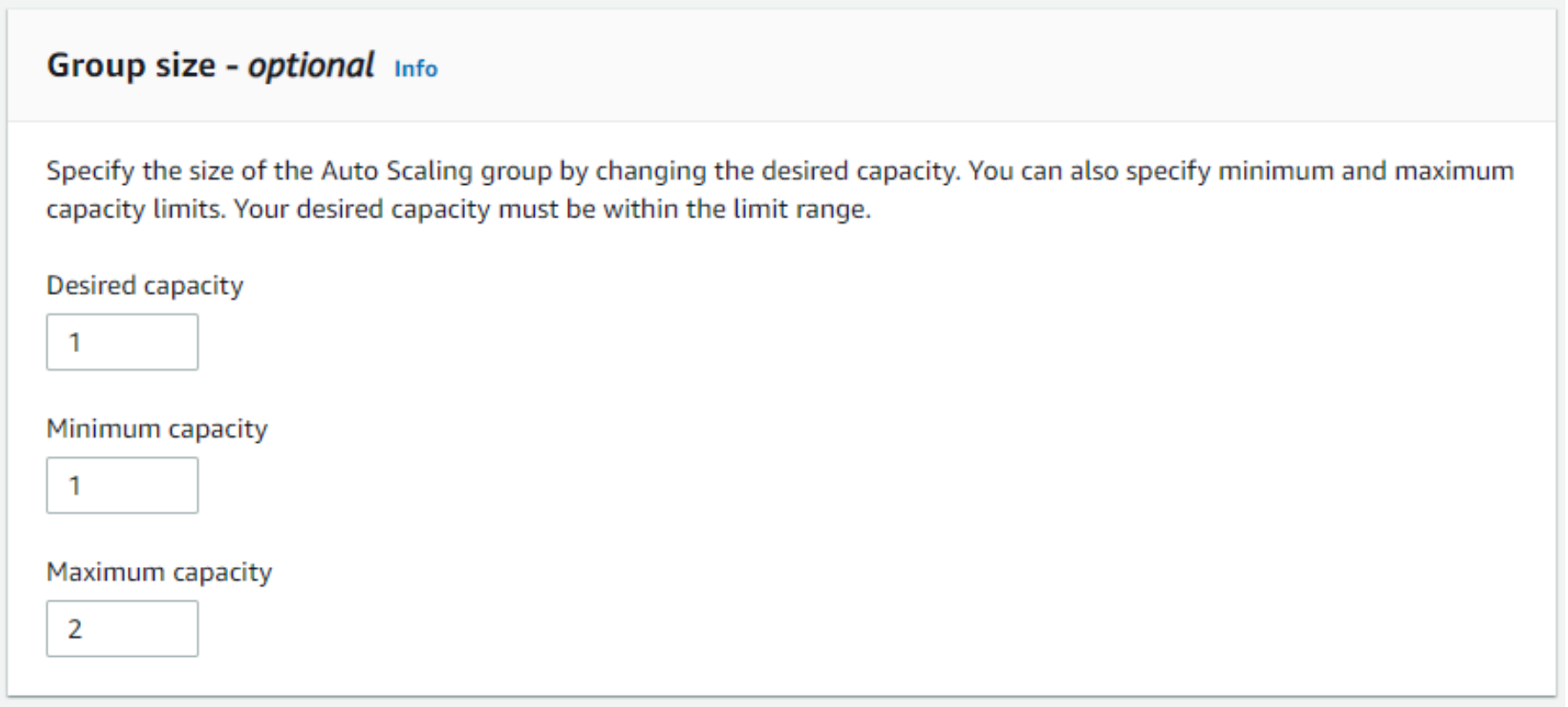


EC2 – ASG Capacity Setting

The size of an Auto Scaling Group is based on **Min**, **Max** and **Desired Capacity**.

- **Min** is how many EC2 instances should at least be running.
- **Max** is number EC2 instances allowed to be running.
- **Desired Capacity** is how many EC2 instances you want to ideally run.

ASG will always launch instances to meet minimum capacity.



EC2 – ASG Health Check Replacements

EC2 Health Check Type

ASG will perform a health check on EC2 instances to determine if there is a software or hardware issue. This based on the **EC2 Status Checks**. If an instance is considered unhealthy. ASG will terminate and launch a new instance.

EC2 – ASG Scaling Policies

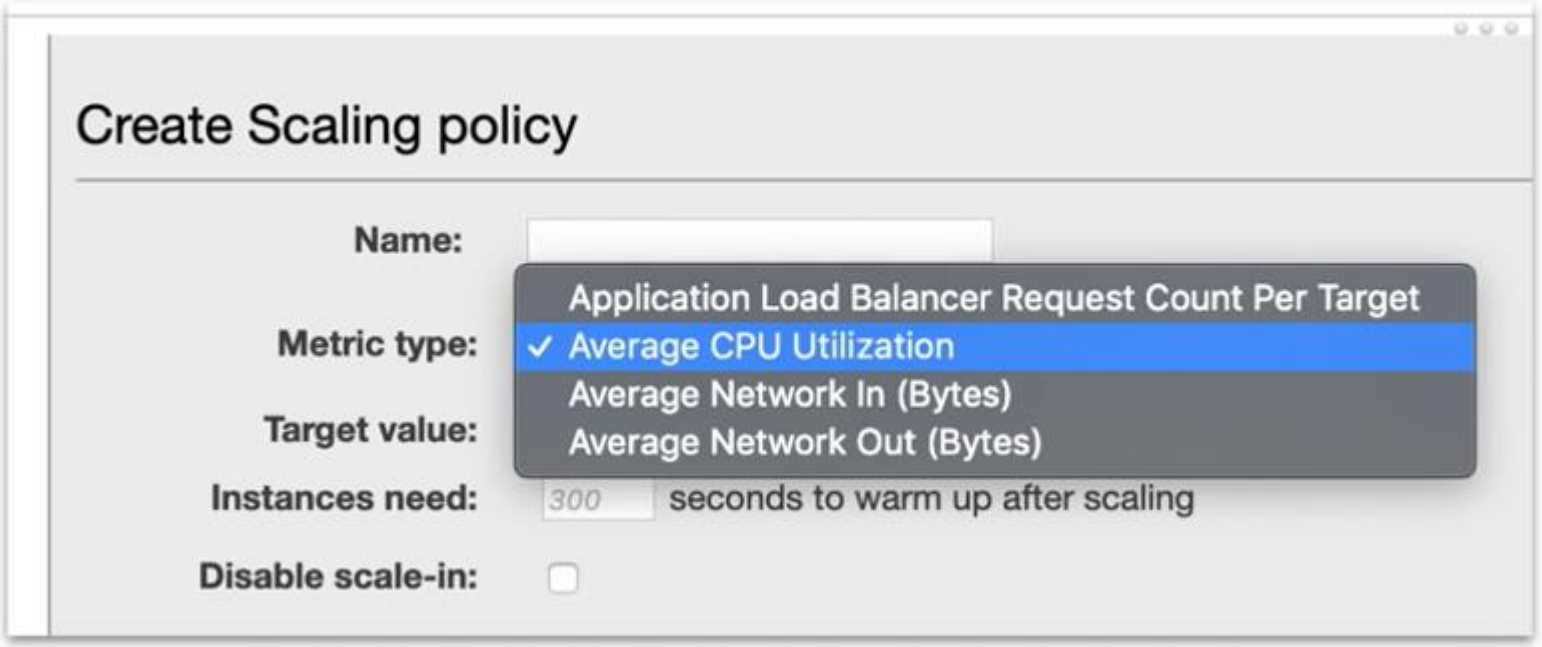
Scaling Out: Adding More Instances

Scaling In: Removing Instances

Target Tracking Scaling Policy

Maintains a specific metric at a target value. 类似定速巡航

eg. If Average CPU Utilization exceeds 75% then add another server.



EC2 – ASG Scaling Policies

Simple Scaling Policy

Scales when an alarm is breached.

Not recommended, legacy scaling policy.

如果在考试的备选项里出现，基本都是备选项

Exams Samples

An application runs on Amazon EC2 instances across multiple Availability Zones.

The instances run in an Amazon EC2 Auto Scaling group behind an Application Load Balancer. The application performs best when the **CPU utilization** of the EC2 instances is at or near 40%.

What should a solutions architect do to maintain the desired performance across all instances m the group?

A. Use a simple scaling policy to dynamically scale the Auto Scaling group
B. Use a target tracking policy to dynamically scale the Auto Scaling group
C. Use an AWS Lambda function to update the desired Auto Scaling group capacity
D. Use scheduled scaling actions to scale up and scale down the Auto Scaling group

Exams Samples

An application runs on Amazon EC2 instances across multiple Availability Zones.

The instances run in an Amazon EC2 Auto Scaling group behind an Application Load Balancer. The application performs best when the **CPU utilization** of the EC2 instances is at or near 40%.

What should a solutions architect do to maintain the desired performance across all instances m the group?

A. Use a simple scaling policy to dynamically scale the Auto Scaling group
B. Use a target tracking policy to dynamically scale the Auto Scaling group
C. Use an AWS Lambda function to update the desired Auto Scaling group capacity
D. Use scheduled scaling actions to scale up and scale down the Auto Scaling group

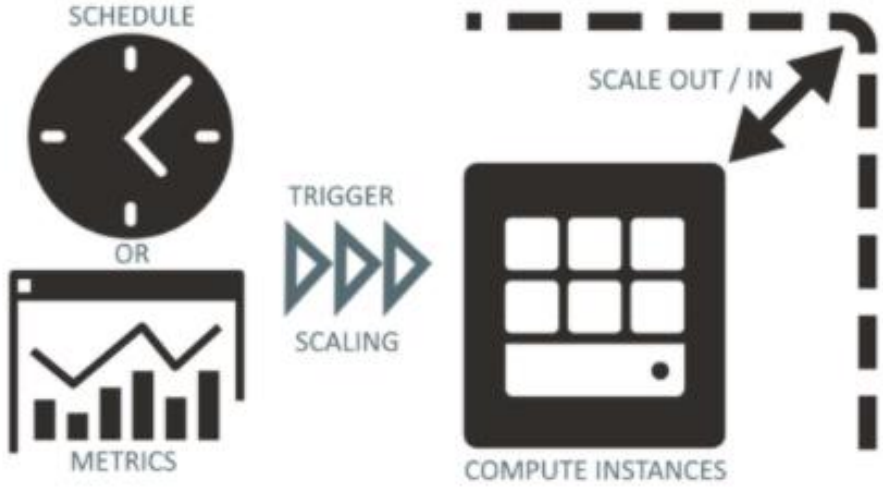
EC2 – ASG Scheduled Scaling Policies

Scaling based on a schedule allows you to set your own **scaling schedule for predictable load changes**.

定时伸缩

For example, every week the traffic to your web application starts to increase on Wednesday, remains high on Thursday, and starts to decrease on Friday.
比如你的网站流量，周三开始加，周四维持峰值，周五开始减少

You can configure Application Auto Scaling to increase capacity on Wednesday and decrease capacity on Friday.
你可以相对应的设置伸缩



Exams Samples

A company's application runs on Amazon EC2 instances behind an Application Load Balancer (ALB).

The instances run in an Amazon EC2 Auto Scaling group across multiple Availability Zones. **On the first day of every month at midnight the application becomes much slower when the month-end financial calculation batch executes.** This causes the CPU utilization of the EC2 instances to immediately peak to 100% which disrupts the application.
因为月结，每个月的第一天的半夜，系统超慢。

What should a solutions architect recommend to ensure the application is able to handle the workload and avoid downtime?

A. Configure an Amazon CloudFront distribution in front of the ALB
B. Configure an EC2 Auto Scaling simple scaling policy based on CPU utilization
C. Configure an EC2 Auto Scaling **scheduled scaling policy** based on the monthly schedule.
D. Configure Amazon ElastiCache to remove some of the workload from the EC2 instances

Exams Samples

A company's application runs on Amazon EC2 instances behind an Application Load Balancer (ALB).

The instances run in an Amazon EC2 Auto Scaling group across multiple Availability Zones. **On the first day of every month at midnight the application becomes much slower when the month-end financial calculation batch executes.** This causes the CPU utilization of the EC2 instances to immediately peak to 100% which disrupts the application.
因为月结，每个月的第一天的半夜，系统超慢。

What should a solutions architect recommend to ensure the application is able to handle the workload and avoid downtime?

A. Configure an Amazon CloudFront distribution in front of the ALB
B. Configure an EC2 Auto Scaling simple scaling policy based on CPU utilization
C. Configure an EC2 Auto Scaling **scheduled scaling policy** based on the monthly schedule.
D. Configure Amazon ElastiCache to remove some of the workload from the EC2 instances

Exams Samples

A gaming company has multiple Amazon EC2 instances in a single Availability Zone for its multiplayer game that communicates with users on Layer 4.

The chief technology officer (CTO) wants to make the architecture **highly available** and **cost- effective**. What should a solutions architect do to meet these requirements? (Select TWO.)

A. Increase the number of EC2 instances.
B. Decrease the number of EC2 instances
C. Configure a Network Load Balancer in front of the EC2 instances.
D. Configure an Application Load Balancer in front of the EC2 instances
E. Configure an Auto Scaling group to add or remove instances in multiple Availability Zones automatically.

Exams Samples

A gaming company has multiple Amazon EC2 instances in a single Availability Zone for its multiplayer game that communicates with users on Layer 4.

The chief technology officer (CTO) wants to make the architecture **highly available** and **cost- effective**. What should a solutions architect do to meet these requirements? (Select TWO.)

A. Increase the number of EC2 instances.
B. Decrease the number of EC2 instances
C. Configure a Network Load Balancer in front of the EC2 instances.
D. Configure an Application Load Balancer in front of the EC2 instances
E. Configure an Auto Scaling group to add or remove instances in multiple Availability Zones automatically.

Exams Samples

A solutions architect must create a highly available bastion host architecture.

The solution needs to be **resilient (可迅速恢复的，有弹性的)** within a single AWS Region and should require only minimal effort to maintain.

What should the solutions architect do to meet these requirements?

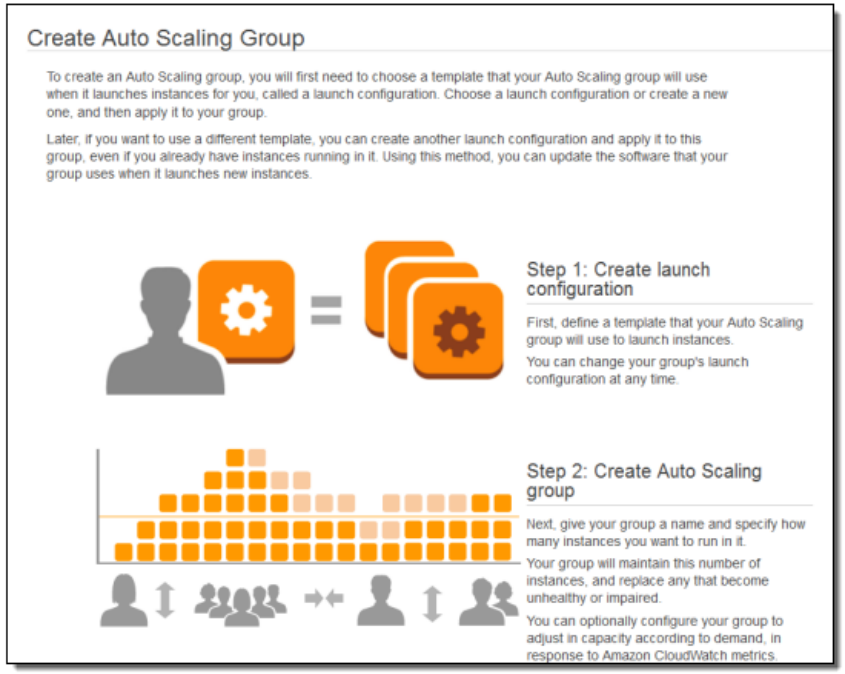
A. Create a Network Load Balancer backed by an Auto Scaling group with a UDP listener.
B. Create a Network Load Balancer backed by a Spot Fleet with instances in a group with instances in a partition placement group.
C. Create a Network Load Balancer backed by the existing serves in different Availability Zones as the target.
D. Create a Network Load Balancer backed by an Auto Scaling with instances in multiple Availability zones as the target

EC2 – ASG Launch Configuration

A launch configuration is an instance configuration template that an **Auto Scaling group uses to launch EC2 instances**.

A launch Configuration is the same process as Launching an EC2 instance except you are saving that configuration to Launch an Instance for later.

Launch Configurations cannot be edited, when you need update your Launch Configuration you create a new one or clone the existing configuration and then manually associate that new Launch Configuration



ASG – Lab

Introduction to Amazon EC2 Auto Scaling on **qwiklabs [20 mins]**

Task 1: Create a Launch Template
Task 2: Create an Auto Scaling Group
Task 3: Verify your Auto Scaling group
Task 3: Test Auto Scaling

EC2 – ASG Summary

- An ASG is a collection of EC2 instances grouped for scaling and management
- Scaling Out is when add servers
- Scaling In is when you remove servers
- Scaling Up is when you increase the size of an instance (eg. Updating Launch Configuration with larger size)
- Size of an ASG is based on a **Min**, **Max** and **Desired Capacity**
- **Target Scaling policy** scales based on when a target value for a metric is breached eg. Average CPU utilization > 75%
- **Simple Scaling policy** triggers a scaling when an alarm is breached
- **Scaling Policy with Steps** is the new version of Simple Scaling policy and allows you to create steps based on alarm values. (you choose scaling metrics and threshold values for the CloudWatch alarms that trigger the scaling process. You also define how your Auto Scaling group should be scaled when a threshold is in breach for a specified number of evaluation periods.)
- Desired Capacity is how many EC2 instances you want to ideally run
- An ASG will always launch instances to meet minimum capacity
- Health checks determine the current state of an instance in the ASG
- Health checks can be run against either an ELB or the EC2 instances
- When an Autoscaling launches a new instance it uses a Launch Configuration which holds the configuration values for that new instance eg. AMI, Instance Type, Role
- Launch Configurations cannot be edited and must be cloned or a new one created
- Launch Configurations must be manually updated in by editing the Auto Scaling settings

ELB – Elastic Load Balancer

ELB – Elastic Load Balancer

Distributes incoming application traffic across multiple targets, such as Amazon EC2 instances, containers, IP addresses, and Lambda functions.
流量分发

Load Balancers can be physical hardware or virtual software that accepts incoming traffic, and then distributes the traffic to multiple targets. They can **balance** the load via different rules. These rules vary based on types of load balancers.
负载均衡

Elastic Load Balancer (ELB) is the AWS solution for load balancing traffic, and there are 3 types available:

Type		Comment
Application Load Balancer	ALB	HTTP/HTTPS
Network Load Balancer	NLB	TCP/UDP
Classic Load Balancer	CLB	Legacy

Gateway Load Balancer link **2020-Nov**

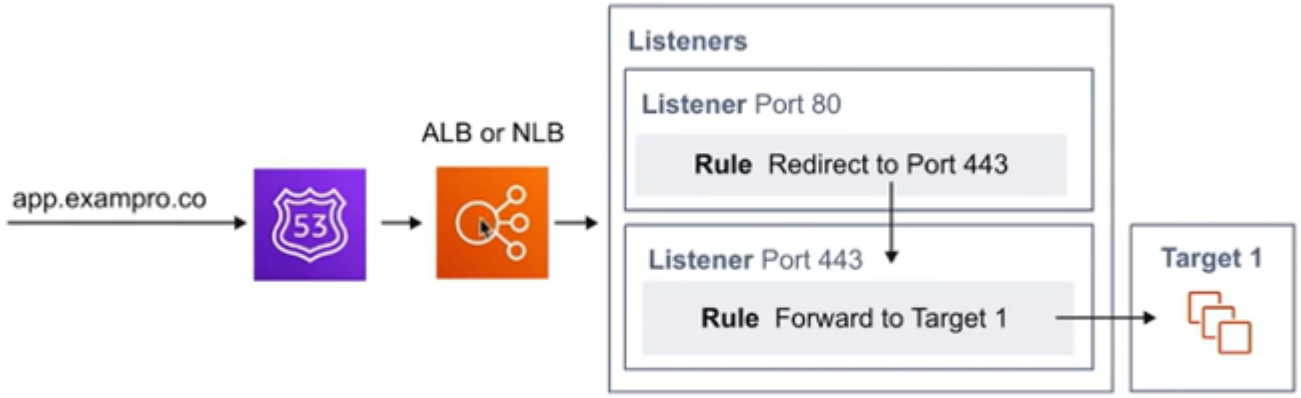
ELB – The Rules of Traffic

Listeners
Incoming traffic is evaluated against listeners. Listeners evaluate any traffic that is matches the Listener's port. For Classic Load Balancer, EC2 instances are directly registered to the Load Balancer.

Rules (Not available for Classic Load Balancer)
Listeners will then invoke rules to decide what to do with the traffic. Generally the next step is to forward traffic to a Target Group.

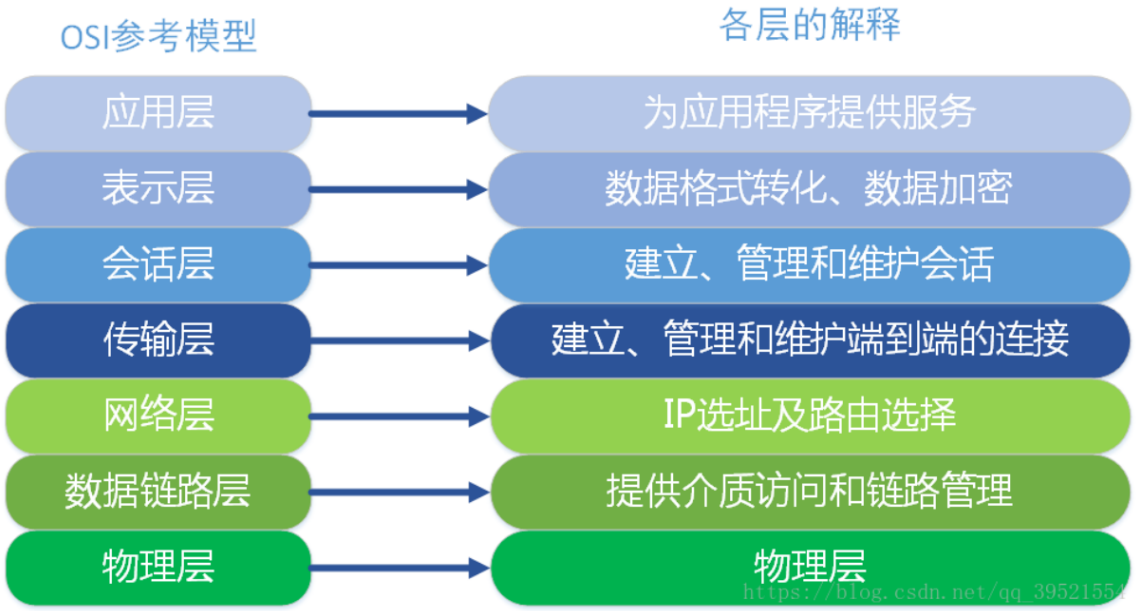
Target Groups (Not available for Classic Load Balancer)
EC2 instances are registered as targets to a Target Group.

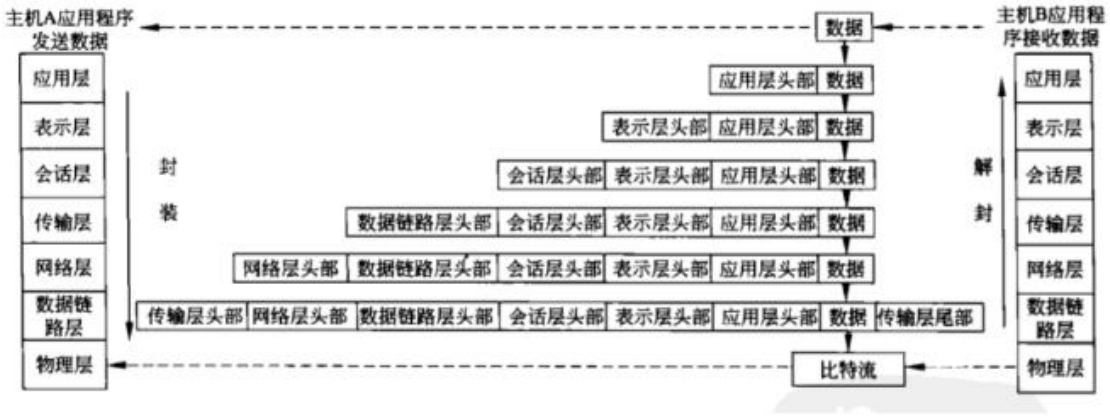
For ALB or NLB traffic is sent to the Listeners. When the port matches it then checks the rules what to do. The rules will forward the traffic to a Target Group. The target group will evenly distribute the traffic to instances registered to that target group.



网络七层

OSI（Open System Interconnect），即开放式系统互联。一般都叫OSI参考模型，是ISO（国际标准化组织）组织在1985年研究的网络互连模型。





ELB – Application Load Balancer (ALB)

ALB are designed to balance **HTTP** and **HTTPS** traffic.

They **operate at Layer 7** (of the OSI Model)
7 层负载均衡

ALB has a feature called Request Routing which allows you to add routing rules to your listeners based on the HTTP protocol.

Web Application Firewall (**WAF**) can be attached to ALB.

Great for Web Applications.

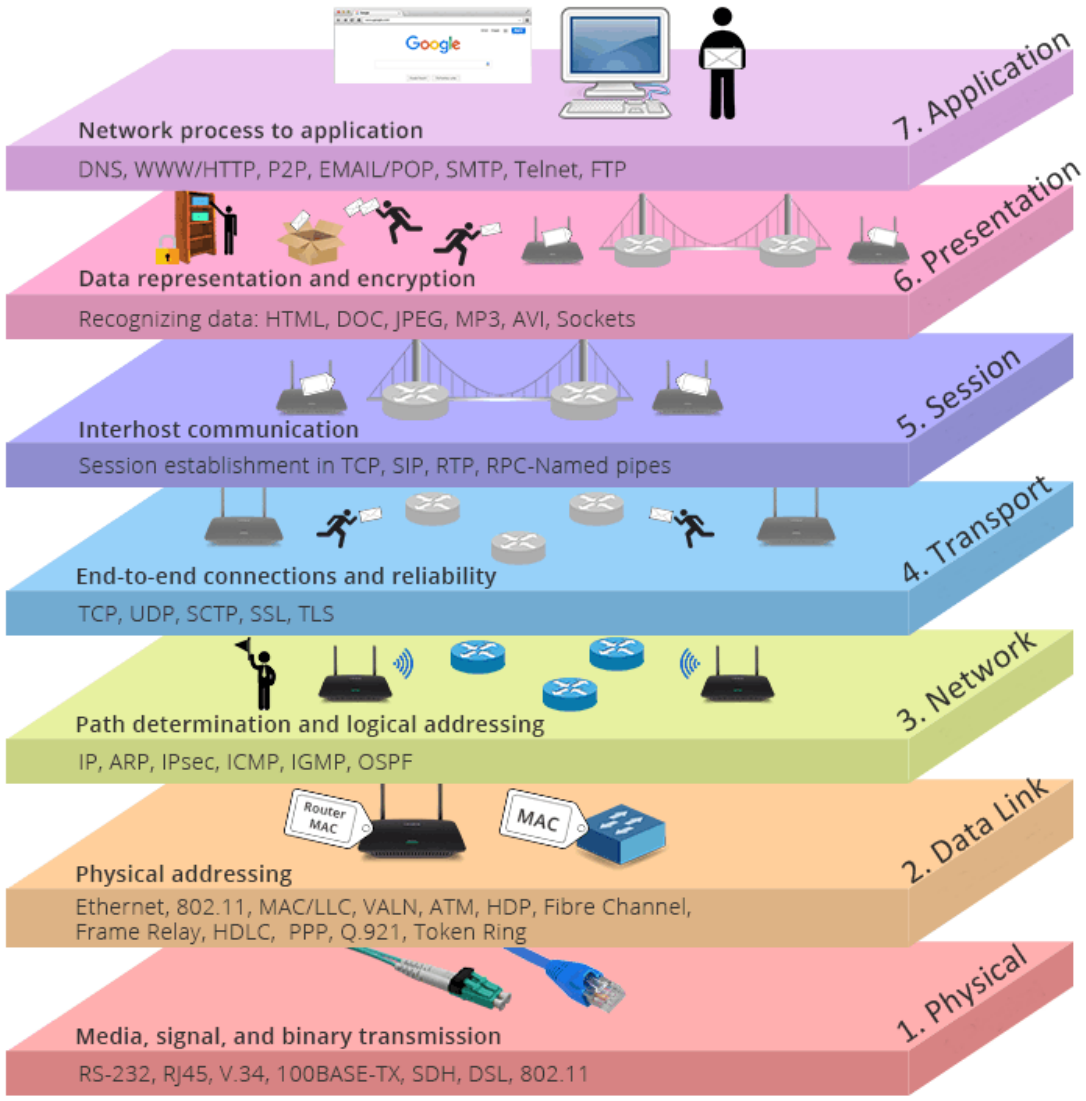


Figure 1: seven layers of the OSI model.

Exams Samples

A web application is deployed in the AWS Cloud It consists of a two-tier architecture that includes ***a*** **web** layer and a database layer.

The web server is vulnerable to cross-site scripting (**XSS**) attacks.

What should a solutions architect do to remediate the vulnerability?

A. Create a Classic Load Balancer. Put the web layer behind the load balancer and enable AWS WAF.
B. Create a Network Load Balancer. Put the web layer behind the load balancer and enable AWS WAF.
C. Create an Application Load Balancer. Put the web layer behind the load balancer and enable AWS WAF.
D. Create an Application Load Balancer. Put the web layer behind the load balancer and use AWS Shield Standard.

Exams Samples

A web application is deployed in the AWS Cloud It consists of a two-tier architecture that includes ***a*** **web** layer and a database layer.

The web server is vulnerable to cross-site scripting (**XSS**) attacks.

What should a solutions architect do to remediate the vulnerability?

A. Create a Classic Load Balancer. Put the web layer behind the load balancer and enable AWS WAF.
B. Create a Network Load Balancer. Put the web layer behind the load balancer and enable AWS WAF.
C. Create an Application Load Balancer. Put the web layer behind the load balancer and enable AWS WAF.
D. Create an Application Load Balancer. Put the web layer behind the load balancer and use AWS Shield Standard.

ELB – Network Load Balancer (NLB)

NLB are designed to balance **TCP/UDP** traffic.

They **operate at Layer 4** (of the OSI Model)
4 层负载均衡

Can handle millions of requests per second while still maintaining extremely low latency.

Can perform Cross-Zone Balancing

Great for Multiplayer **Video Games** or When network performance is critical

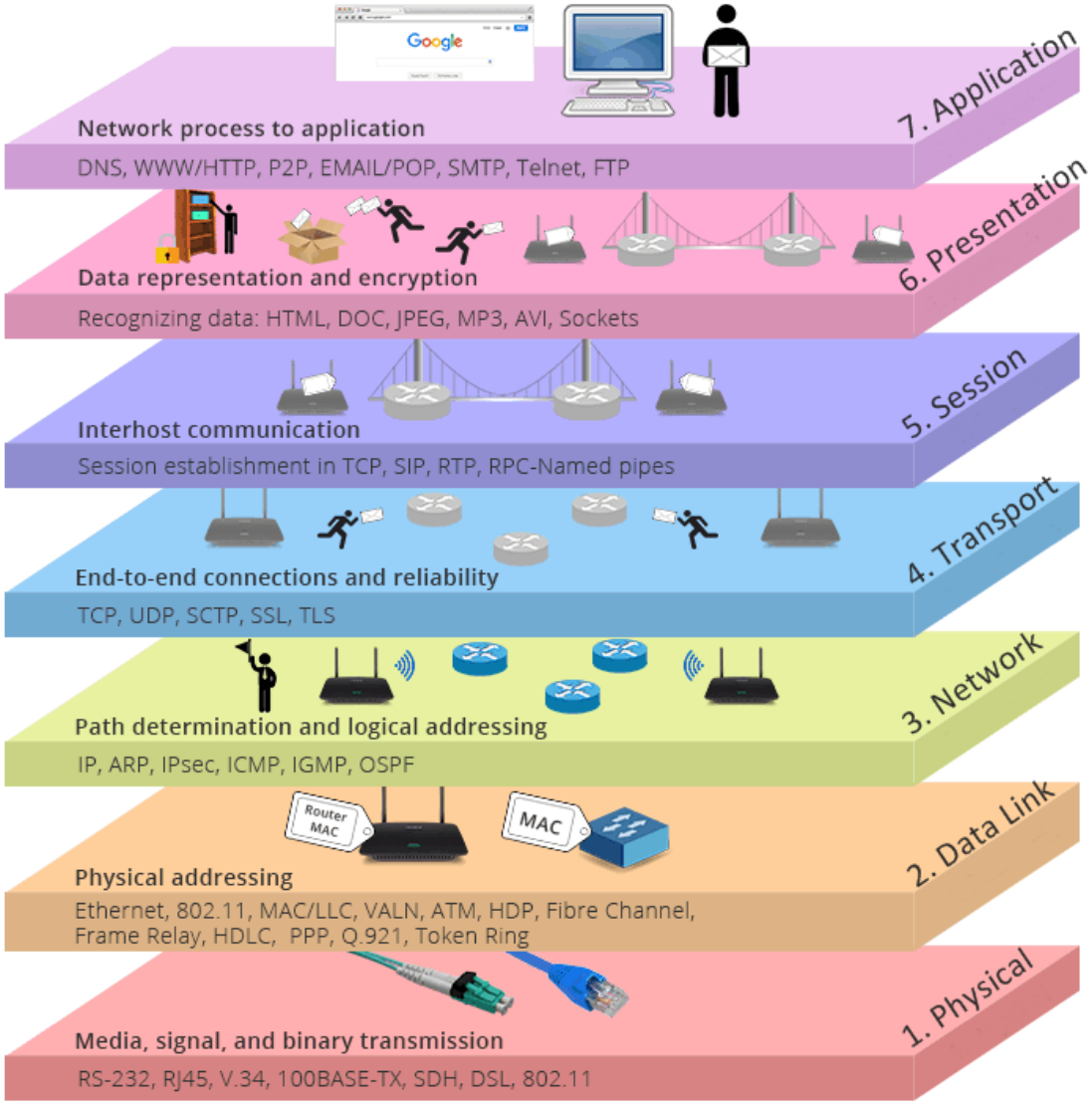


Figure 1: seven layers of the OSI model.

Exams Samples

A company is hosting multiple **websites** for several lines of business under its registered parent domain. Users accessing these websites will be routed to appropriate backend Amazon **EC2** instances based on the subdomain. The websites host static webpages, images, and server-side scripts like PHP and JavaScript.

Some of the websites experience peak access during the first two hours of business with constant usage throughout the rest of the day. A solutions architect needs to design a solution that will **automatically adjust capacity to these traffic patterns** **while** **keeping costs low**.

Which combination of AWS services or features will meet these requirements? (Choose two.)

A. AWS Batch
B. Network Load Balancer
C. Application Load Balancer
D. Amazon EC2 Auto Scaling
E. Amazon S3 website hosting

Exams Samples

A company is hosting multiple **websites** for several lines of business under its registered parent domain. Users accessing these websites will be routed to appropriate backend Amazon **EC2** instances based on the subdomain. The websites host static webpages, images, and server-side scripts like PHP and JavaScript.

Some of the websites experience peak access during the first two hours of business with constant usage throughout the rest of the day. A solutions architect needs to design a solution that will **automatically adjust capacity to these traffic patterns** **while** **keeping costs low**.

Which combination of AWS services or features will meet these requirements? (Choose two.)

A. AWS Batch
B. Network Load Balancer
C. Application Load Balancer
D. Amazon EC2 Auto Scaling
E. Amazon S3 website hosting

ELB – Sticky Sessions

Sticky Sessions is an advanced load balancing method that allows you to **bind a user's session to a specific EC2 instance**.
绑定用户会话到固定的 EC2 实例

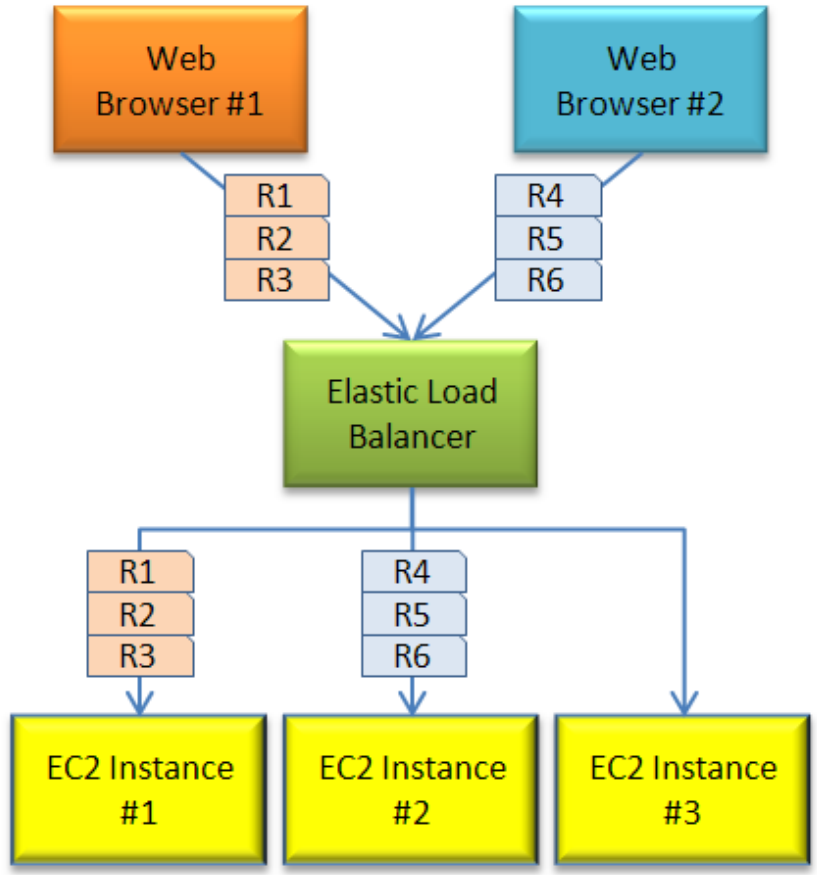
Ensure all **requests** from that session are sent to **the same instance**.

Typically **utilized** with a **Classic Load Balancer.**

Can be enabled for ALB though can only be set on a Target Group not individual EC2 instances.

Cookies are used to remember which EC2 instance.

Useful when specific **information is only stored locally on a single instance**



ELB – X-Forwarded-For (XFF) Header

If you **need the IPv4 address of a user**, check the **X-Forwared-For** header

The **X-Forwared-For (XFF)** header is a command method for identifying the **originating IP address** of a client connecting to a web server through an HTTP proxy or a load balancer.

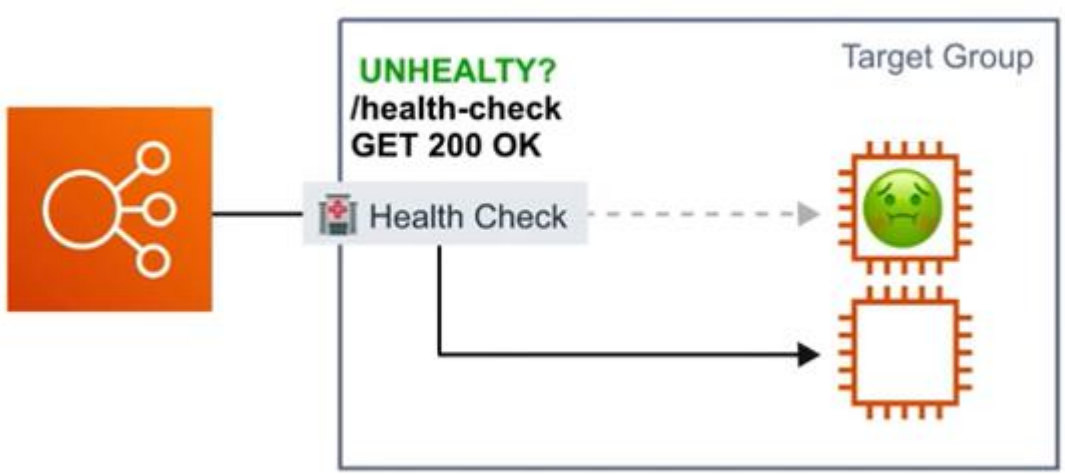
XFF 帮助确定访问源 IP

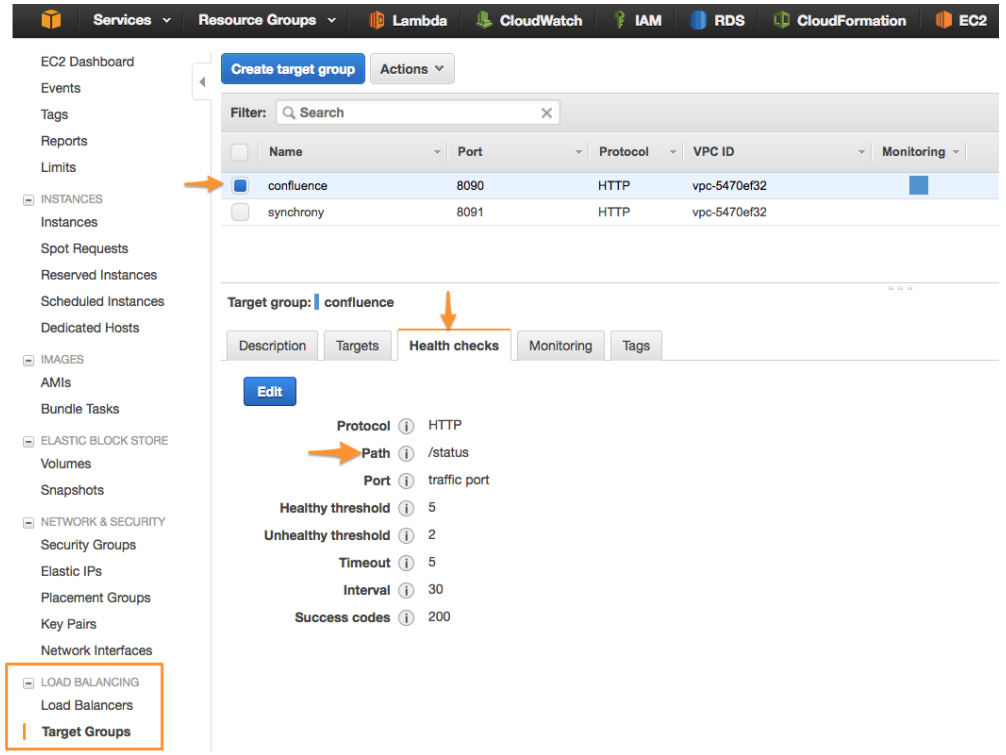
ELB – Health Checks

Instances that are monitored by the Elastic Load Balancer (ELB) report back Health Checks as **InService, or OutofService**. Health Checks communicate directly with the instance to determine its state.

ELB **does not terminate (kill) unhealthy instance**. It will just redirect traffic to healthy instances.
ELB 不会停掉故障实例，而是把流量导向健康实例。

For ALB and NLB the Heath checks are found on the **Target Group**



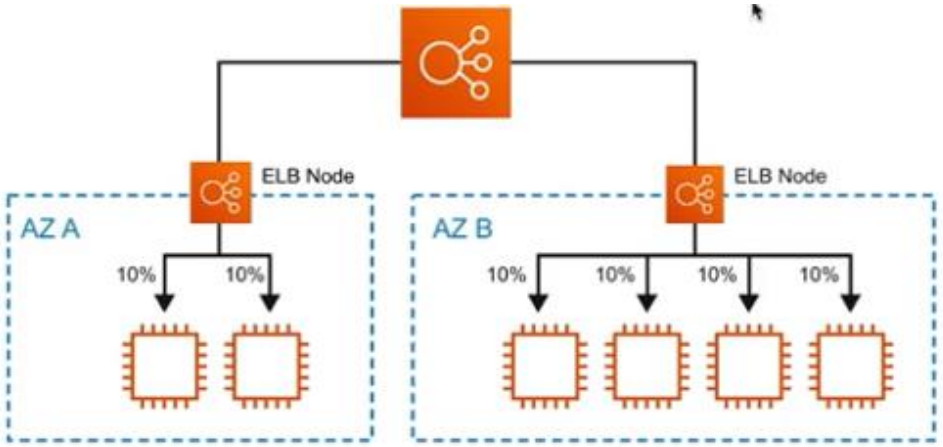


ELB – Cross-Zone Load Balancing

Only for **Classic** and **Network** Load Balancer

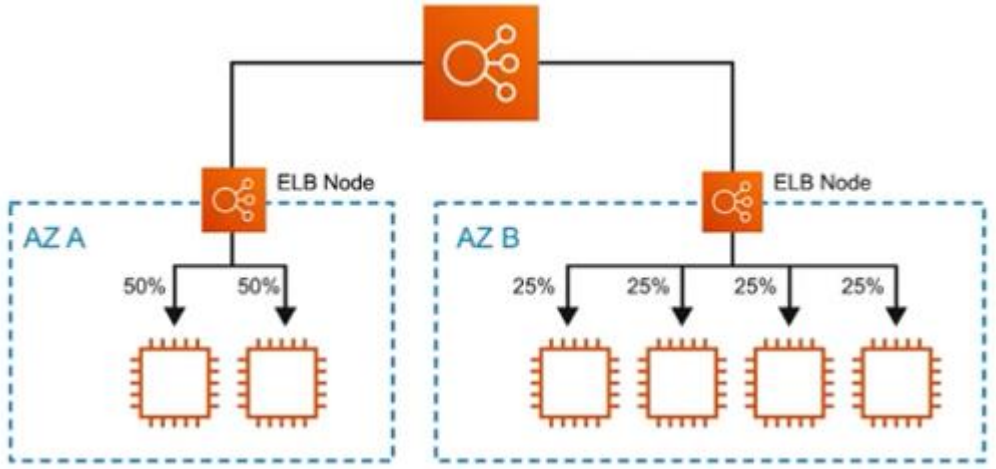
Cross-zone Load Balancing Enabled

Requests are distributed evenly across the instances in all enabled Availability Zones.
跨可用区做流量均衡



Cross-zone Load Balancing Disabled

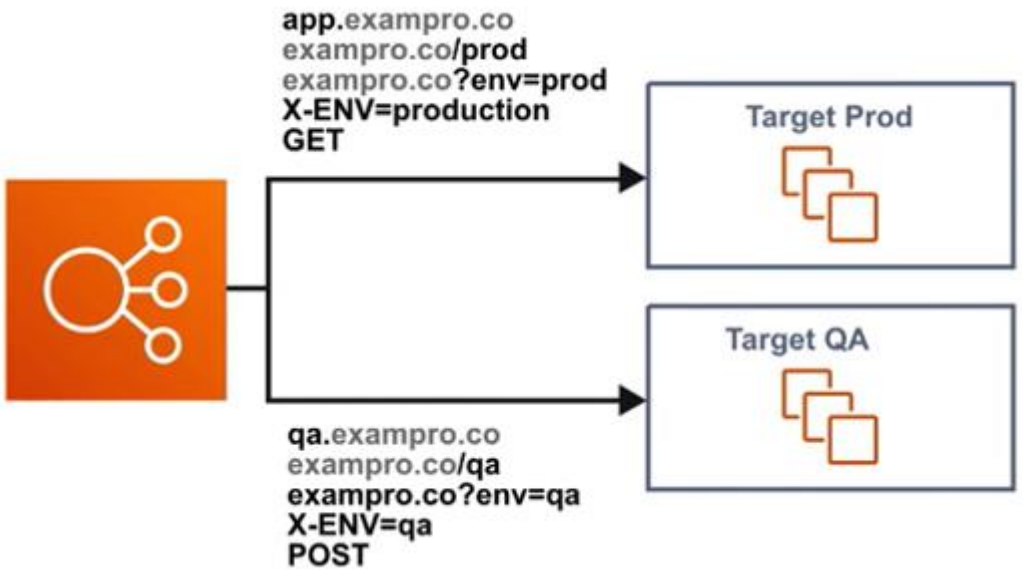
Requests are distributed evenly across the instances in **only its** Availability Zone.
在单个可用区内做流量均衡



ELB – Request Routing

Apply rules to incoming request and then **forward** or **redirect** traffic.

- Host header
- Http header
- Source IP
- Http header method
- Path
- Query string



不需要每个都去实验测试，那就无止境了

ALB lab （Optional）

Working with Elastic Load Balancing on **qwiklabs [30 mins]**

Task 1: Launch Web Servers
Task 2: Connect to Each Web Server
Task 3: Create a Load Balancer

ELB - Summary

- There are three Elastic Load Balancers: **Network**, **Application** and **Classic** Load Balancer
- A Elastic Load Balancer must have **at least two** Availability Zones
- Elastic Load Balancers **cannot go cross-region**. You must create one per region
- ALB has **Listeners, Rules** and **Target groups** to route traffic
- NLB use **Listeners** and **Target Groups** to route traffic
- CLB use **Listeners** and EC2 instances are **directly registered** as targets to CLB
- Application Load Balancer is for HTTP(s) traffic and the name implies it good for Web Applications
- Network Load Balancer is for TCP/UDP is good for high network throughput eg. Video Games
- Classic Load Balancer is legacy and its recommend to use ALB and NLB
- Use X-Forwarded-For (XFF) to get original IP of incoming traffic passing through ELB
- You can attach Web Application Firewall (WAF) to ALB but not to NLB or CLB
- You can attach Amazon Certification Manager SSL to any of the Elastic Load Balancers for SSL
- ALB has advanced Request Routing rules where you can route based on subdomain header, path and other HTTP(S) information
- Sticky Sessions can be enable for CLB or ALB and sessions are remembered via Cookie