Most Important Charts & Diagrams

A Solutions Architect Associate Study Guide

AWS Support Plans Chart

Via https://aws.amazon.com/premiumsupport/plans/

Basic Support is included for all AWS customers and includes:

Customer Service & Communities - 24x7 access to customer service, documentation, whitepapers, and support forums.

AWS Trusted Advisor - Access to the 7 core Trusted Advisor checks and guidance to provision your resources following best practices to increase performance and improve security.

AWS Personal Health Dashboard - A personalized view of the health of AWS services, and alerts when your resources are impacted.

	Developer	**Business**	**Enterprise**
	Recommended if you are experimenting or testing in AWS.	*Recommended if you have production workloads in AWS.*	*Recommended if you have business and/or mission critical workloads in AWS.*
AWS Trusted Advisor Best Practice Checks	7 Core checks	Full set of checks	Full set of checks
Enhanced Technical Support	Business hours** email access to Cloud Support Associates	24x7 phone, email, and chat access to Cloud Support Engineers	24x7 phone, email, and chat access to Cloud Support Engineers
	Unlimited cases / 1 primary contact	Unlimited cases / unlimited contacts (IAM supported)	Unlimited cases / unlimited contacts (IAM supported)
Case Severity / Response Times*	General guidance: < 24 business hours** System impaired: < 12 business hours**	General guidance: < 24 hours System impaired: < 12 hours Production system impaired: < 4 hours Production system down: < 1 hour	General guidance: < 24 hours System impaired: < 12 hours Production system impaired: < 4 hours Production system down: < 1 hour Business-critical system down: < 15 minutes
Architectural Guidance	General	Contextual to your use-cases	Consultative review and guidance based on your applications
Programmatic Case Management		AWS Support API	AWS Support API

Third-Party Software Support		Interoperability & configuration guidance and troubleshooting	Interoperability & configuration guidance and troubleshooting
Proactive Programs		Access to Infrastructure Event Management for additional fee.	Infrastructure Event Management Well-Architected Reviews Operations Reviews Technical Account Manager (TAM) coordinates access to programs and other AWS experts as needed.
Technical Account Management			Designated Technical Account Manager (TAM) to proactively monitor your environment and assist with optimization.
Training			Access to online self-paced labs
Account Assistance			Concierge Support Team
Pricing	Greater of $29 / month*** - or - 3% of monthly AWS usage See pricing detail and example.	Greater of $100 / month*** - or - 10% of monthly AWS usage for the first $0–$10K 7% of monthly AWS usage from $10K–$80K 5% of monthly AWS usage from $80K–$250K 3% of monthly AWS usage over $250K See pricing detail and example.	Greater of $15,000 - or - 10% of monthly AWS usage for the first $0–$150K 7% of monthly AWS usage from $150K–$500K 5% of monthly AWS usage from $500K–$1M 3% of monthly AWS usage over $1M See pricing detail and example.

*We will make every reasonable effort to respond to your initial request within the corresponding timeframes.

**Business hours are generally defined as 8:00 AM to 6:00 PM in the customer country as set in My Account console, excluding holidays and weekends. These times may vary in countries with multiple time zones.

*** Plans are subject to a 30 day minimum term.

S3 Storage Classes Chart

Via https://aws.amazon.com/s3/storage-classes/

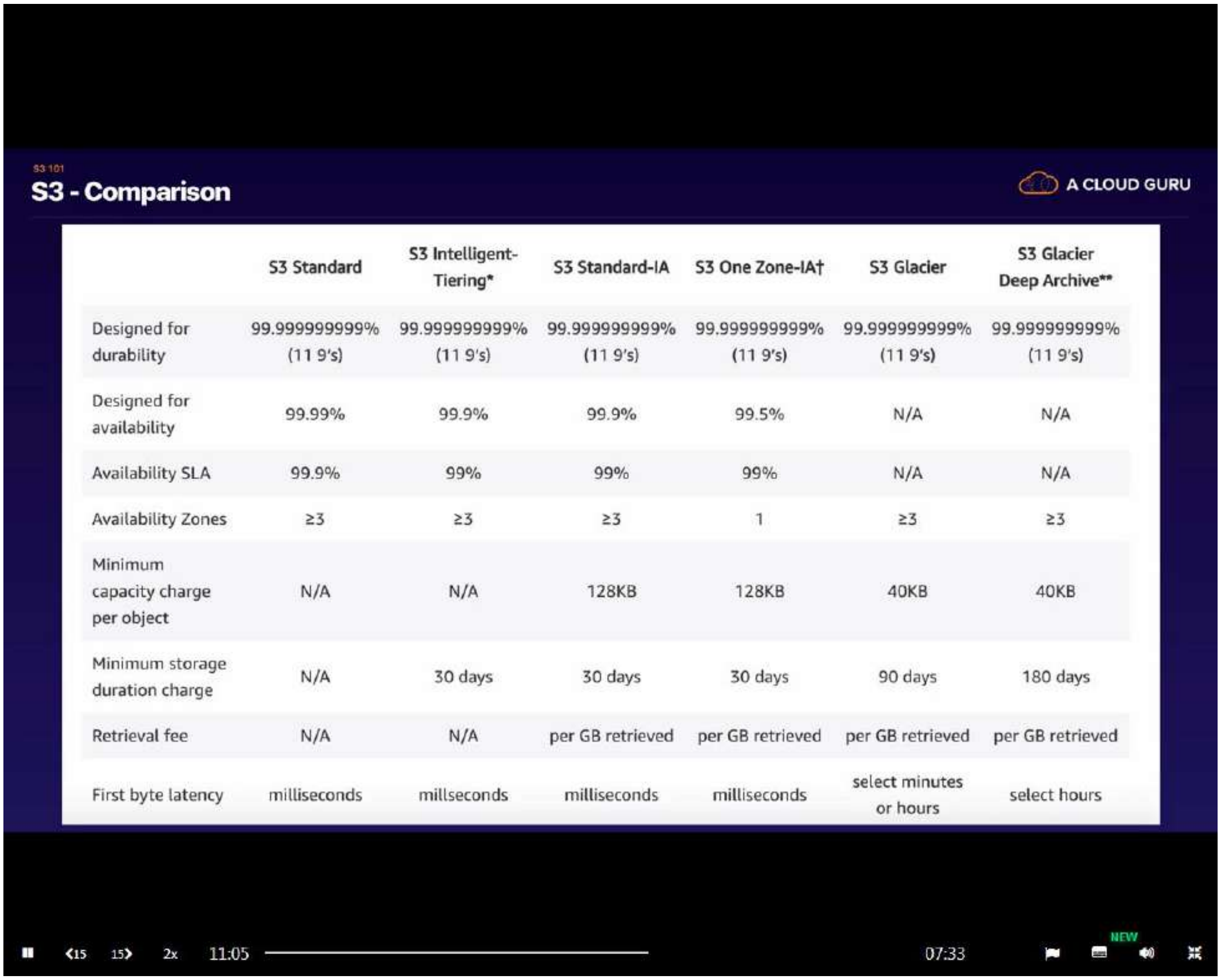
Performance across the S3 Storage Classes

	S3 Standard	**S3 Intelligent-Tiering***	**S3 Standard-IA**	**S3 One Zone-IA†**	**S3 Glacier**	**S3 Glacier Deep Archive**
Designed for durability	99.999999999 % (11 9’s)	99.999999999% (11 9’s)	99.999999999% (11 9’s)	99.999999999% (11 9’s)	99.999999999% (11 9’s)	99.999999999% (11 9’s)
Designed for availability	99.99%	99.9%	99.9%	99.5%	99.99%	99.99%
Availability SLA	99.9%	99%	99%	99%	99.9%	99.9%
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved
First byte latency	milliseconds	milliseconds	milliseconds	milliseconds	select minutes or hours	select hours
Storage type	Object	Object	Object	Object	Object	Object
Lifecycle transitions	Yes	Yes	Yes	Yes	Yes	Yes

† Because S3 One Zone-IA stores data in a single AWS Availability Zone, data stored in this storage class will be lost in the event of Availability Zone destruction.

* S3 Intelligent-Tiering charges a small tiering fee and has a minimum eligible object size of 128KB for auto-tiering. Smaller objects may be stored but will always be charged at the Frequent Access tier rates. See the Amazon S3 Pricing for more information.

S3 Storage Classes (continued)



S3 101

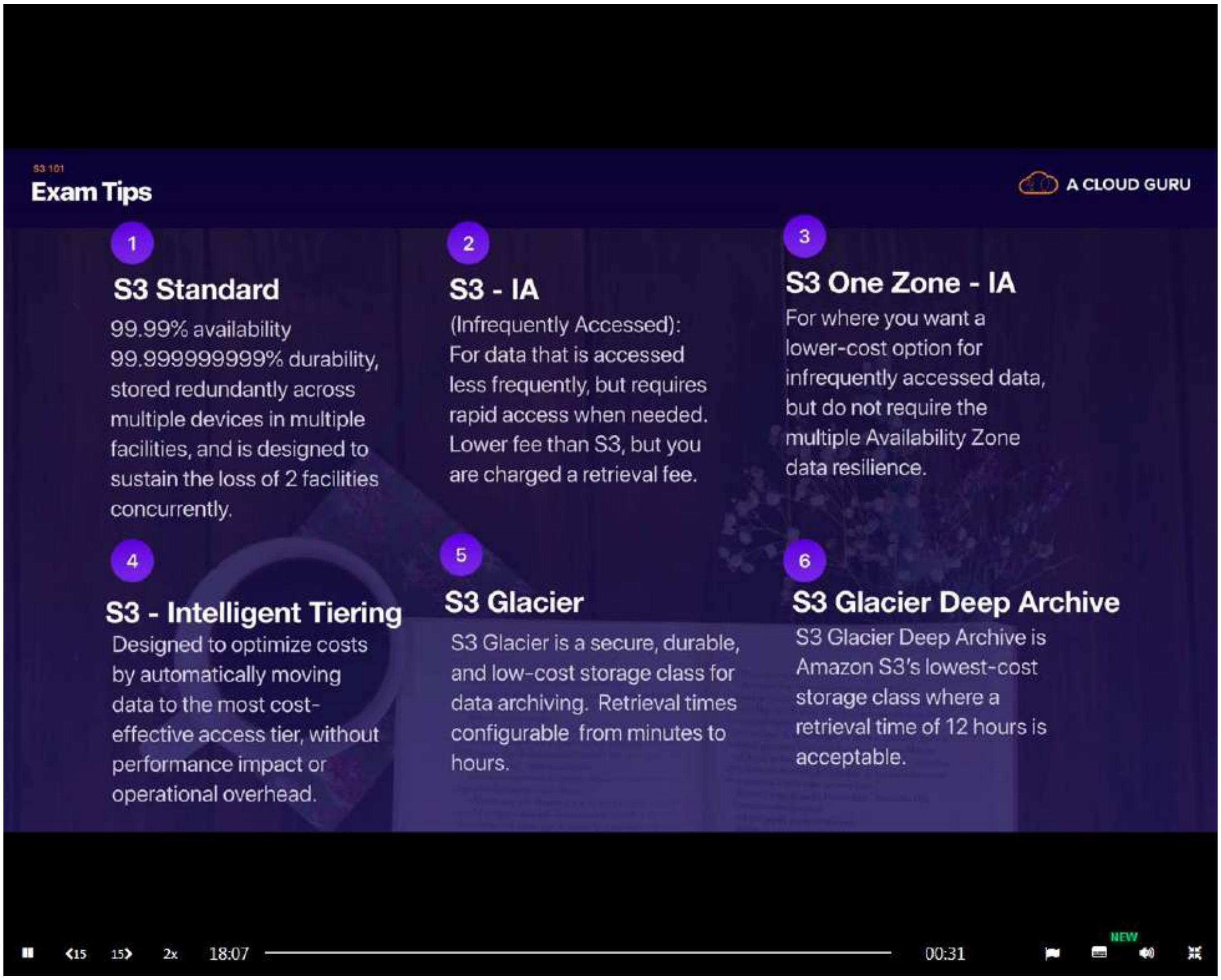
S3 - Comparison

A CLOUD GURU

	S3 Standard	S3 Intelligent-Tiering*	S3 Standard-IA	S3 One Zone-IA†	S3 Glacier	S3 Glacier Deep Archive**
Designed for durability	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)	99.999999999% (11 9's)
Designed for availability	99.99%	99.9%	99.9%	99.5%	N/A	N/A
Availability SLA	99.9%	99%	99%	99%	N/A	N/A
Availability Zones	≥3	≥3	≥3	1	≥3	≥3
Minimum capacity charge per object	N/A	N/A	128KB	128KB	40KB	40KB
Minimum storage duration charge	N/A	30 days	30 days	30 days	90 days	180 days
Retrieval fee	N/A	N/A	per GB retrieved	per GB retrieved	per GB retrieved	per GB retrieved
First byte latency	milliseconds	millseconds	milliseconds	milliseconds	select minutes or hours	select hours

The Above Diagram is from ACG

S3 Storage Classes (continued)



Exam Tips

1

S3 Standard

99.99% availability
99.999999999% durability, stored redundantly across multiple devices in multiple facilities, and is designed to sustain the loss of 2 facilities concurrently.

2

S3 - IA

(Infrequently Accessed):
For data that is accessed less frequently, but requires rapid access when needed. Lower fee than S3, but you are charged a retrieval fee.

3

S3 One Zone - IA

For where you want a lower-cost option for infrequently accessed data, but do not require the multiple Availability Zone data resilience.

4

S3 - Intelligent Tiering

Designed to optimize costs by automatically moving data to the most cost-effective access tier, without performance impact or operational overhead.

5

S3 Glacier

S3 Glacier is a secure, durable, and low-cost storage class for data archiving. Retrieval times configurable from minutes to hours.

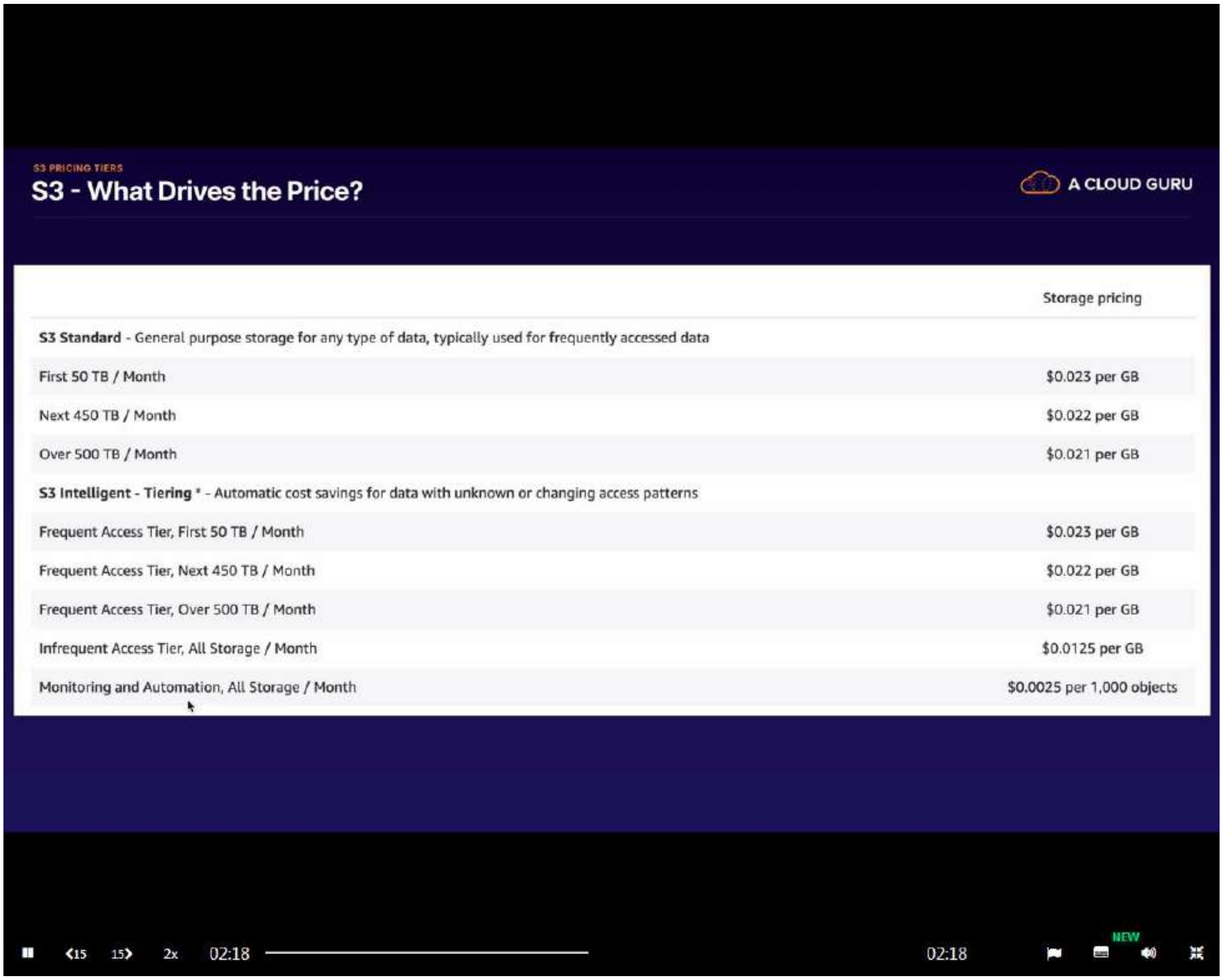
6

S3 Glacier Deep Archive

S3 Glacier Deep Archive is Amazon S3's lowest-cost storage class where a retrieval time of 12 hours is acceptable.

The Above Information is from ACG

S3 Storage Classes (continued)



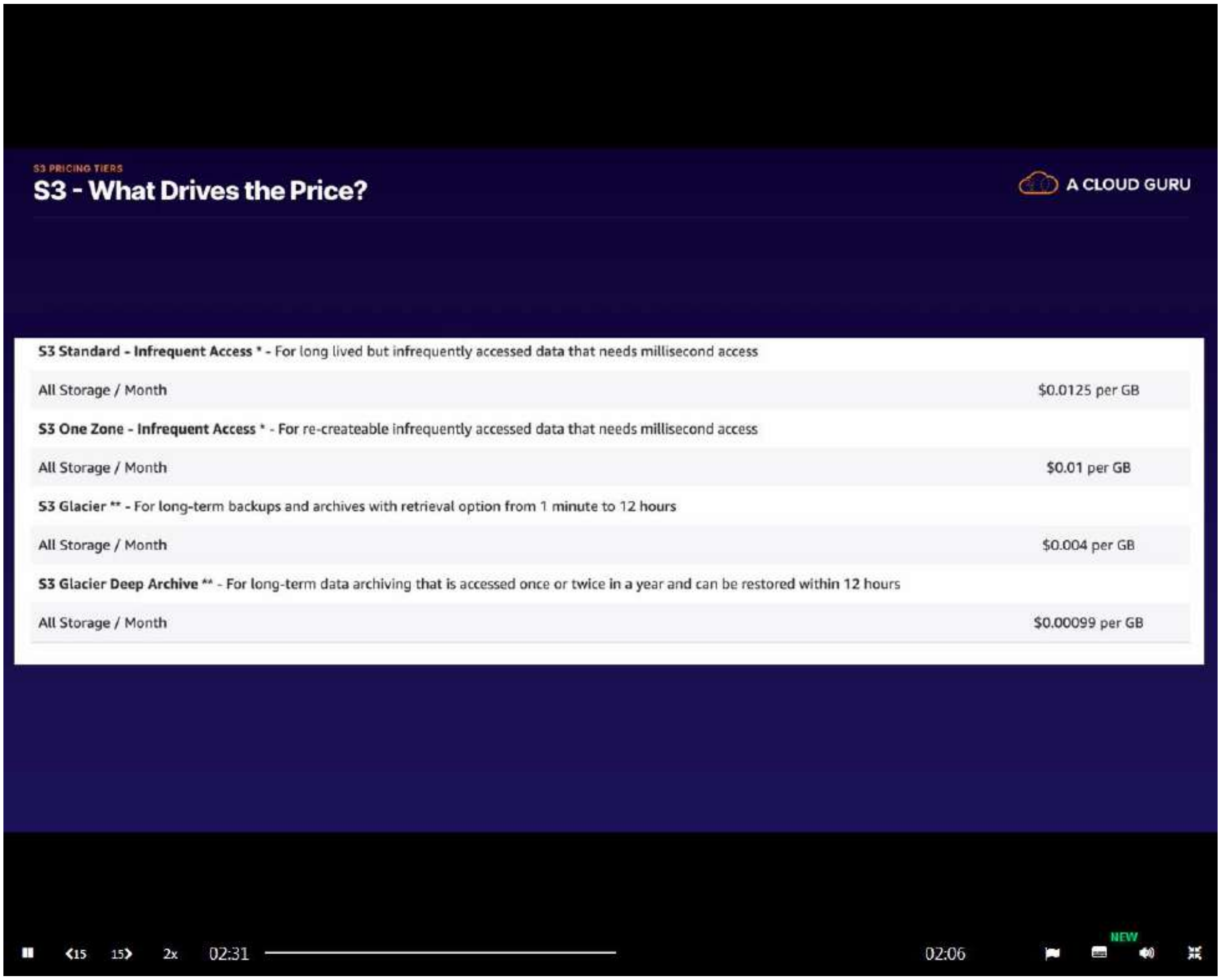
S3 PRICING TIERS

S3 - What Drives the Price?

	Storage pricing
S3 Standard - General purpose storage for any type of data, typically used for frequently accessed data	
First 50 TB / Month	$0.023 per GB
Next 450 TB / Month	$0.022 per GB
Over 500 TB / Month	$0.021 per GB
S3 Intelligent - Tiering * - Automatic cost savings for data with unknown or changing access patterns	
Frequent Access Tier, First 50 TB / Month	$0.023 per GB
Frequent Access Tier, Next 450 TB / Month	$0.022 per GB
Frequent Access Tier, Over 500 TB / Month	$0.021 per GB
Infrequent Access Tier, All Storage / Month	$0.0125 per GB
Monitoring and Automation, All Storage / Month	$0.0025 per 1,000 objects

The Above Diagram is from ACG

S3 Storage Classes (continued)



S3 PRICING TIERS

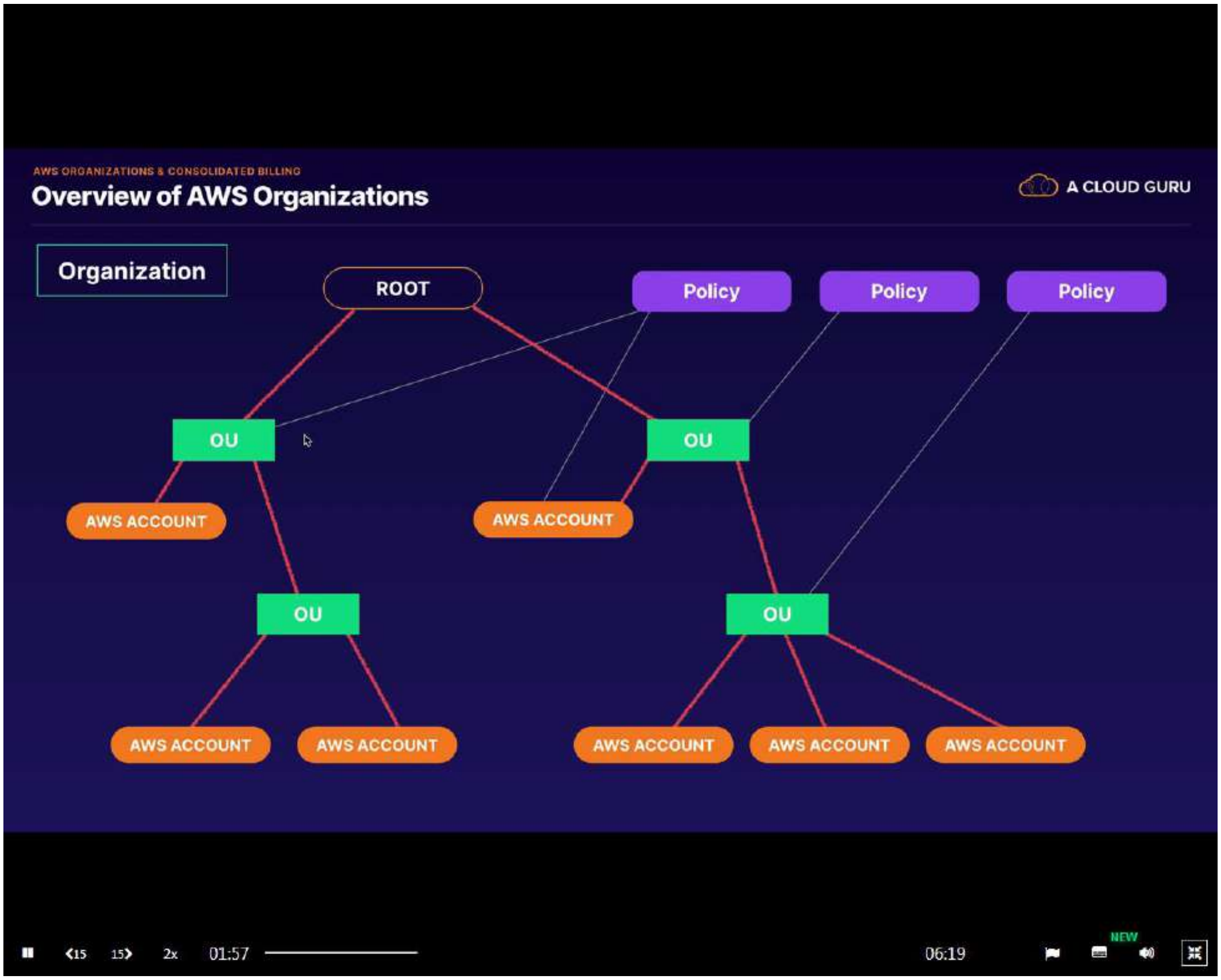
S3 - What Drives the Price?

A CLOUD GURU

S3 Standard - Infrequent Access * - For long lived but infrequently accessed data that needs millisecond access	
All Storage / Month	$0.0125 per GB
S3 One Zone - Infrequent Access * - For re-createable infrequently accessed data that needs millisecond access	
All Storage / Month	$0.01 per GB
S3 Glacier ** - For long-term backups and archives with retrieval option from 1 minute to 12 hours	
All Storage / Month	$0.004 per GB
S3 Glacier Deep Archive ** - For long-term data archiving that is accessed once or twice in a year and can be restored within 12 hours	
All Storage / Month	$0.00099 per GB

The Above Diagram is from ACG

AWS Organizations

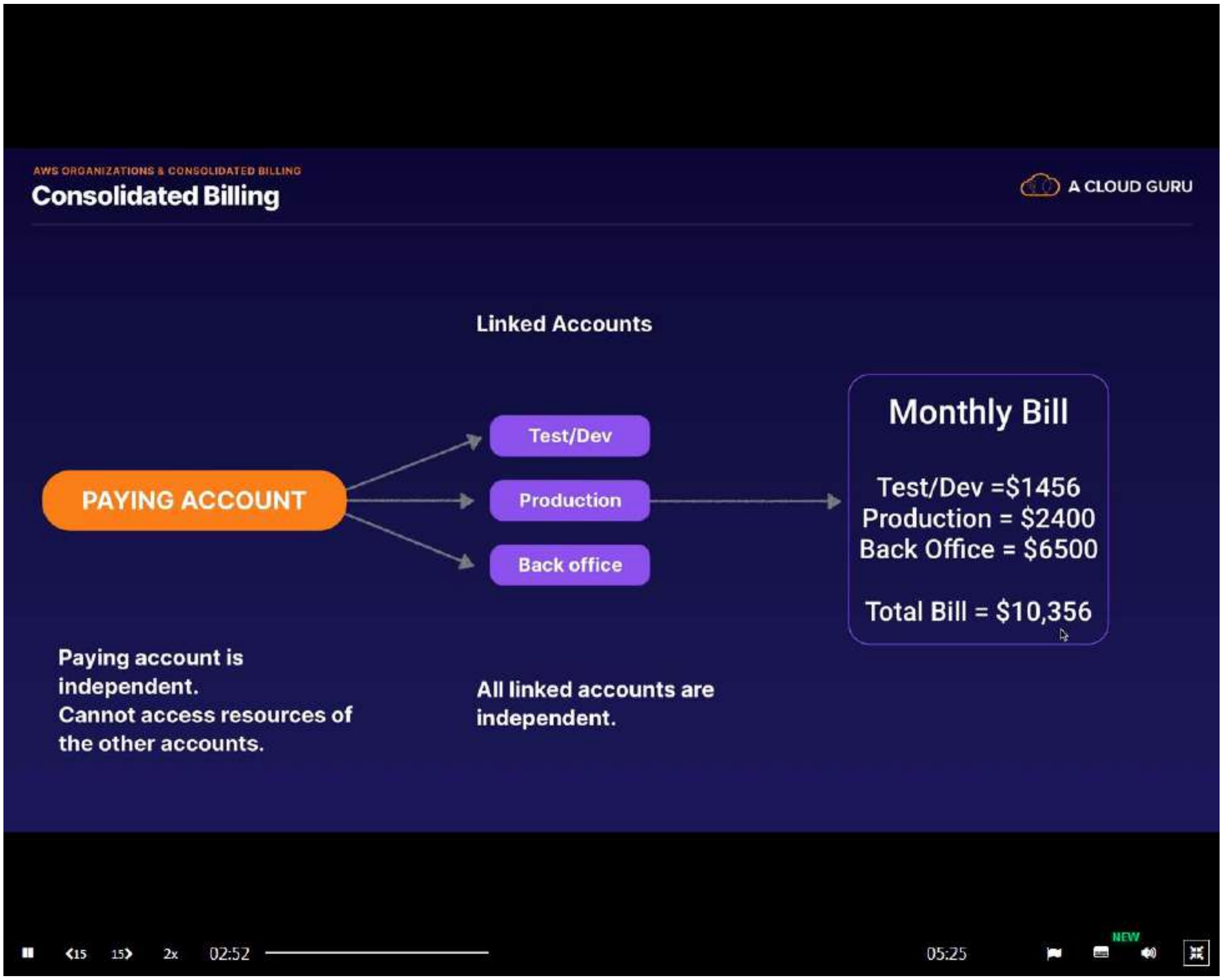


AWS ORGANIZATIONS & CONSOLIDATED BILLING

Overview of AWS Organizations

The Above Diagram is from ACG

AWS Organizations & Consolidated Billing



AWS ORGANIZATIONS & CONSOLIDATED BILLING

Consolidated Billing

Linked Accounts

PAYING ACCOUNT

Test/Dev

Production

Back office

Monthly Bill

Test/Dev =$1456
Production = $2400
Back Office = $6500

Total Bill = $10,356

Paying account is independent.
Cannot access resources of the other accounts.

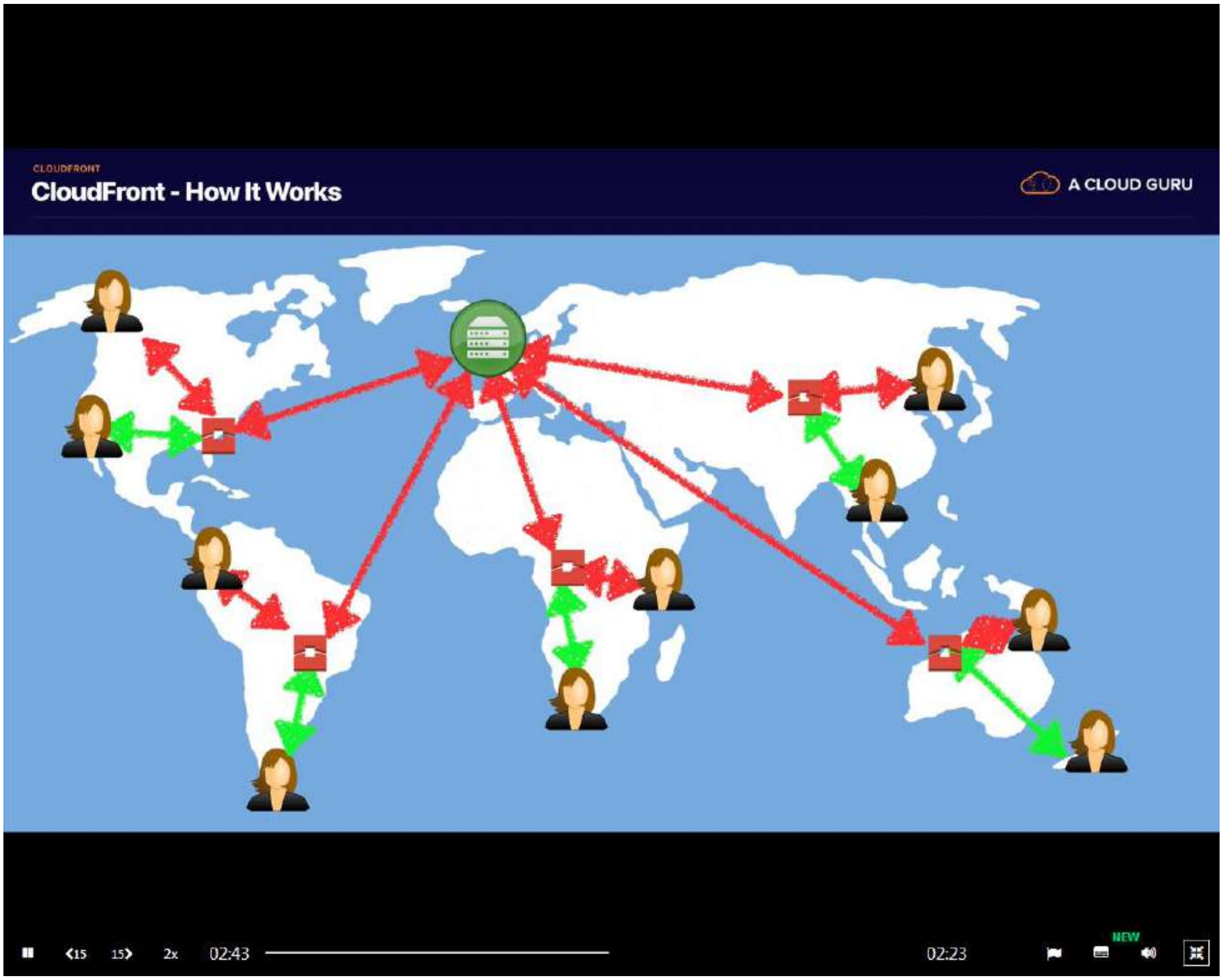
All linked accounts are independent.

The Above Diagram is from ACG

CloudFront Diagram

*The squares represent Edge Locations

*The green rack icon represents the origin of the file to be distributed



CLOUDFRONT

CloudFront - How It Works

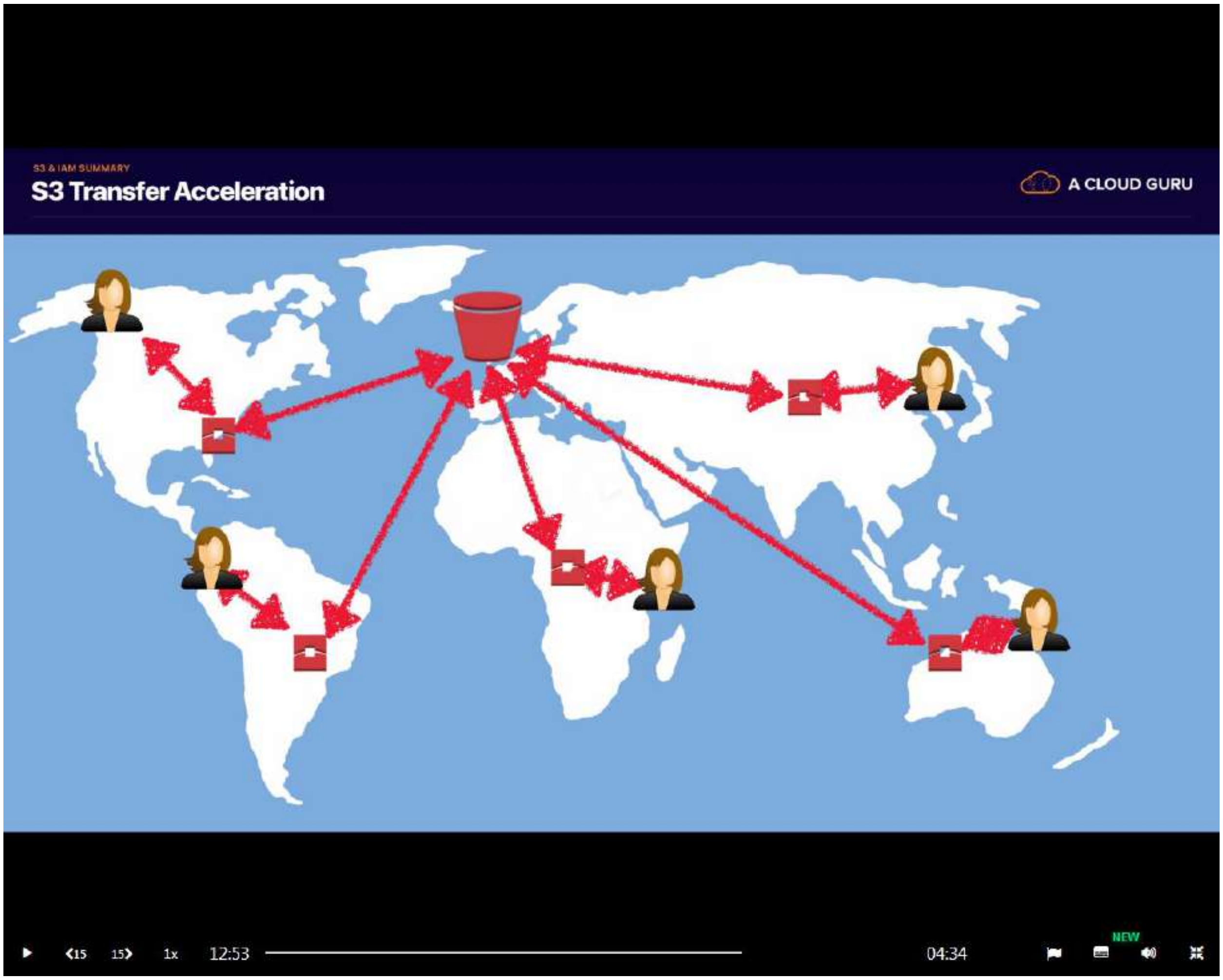
A CLOUD GURU

The above diagram is from ACG

S3 Transfer Acceleration Diagram

*The squares represent Edge Locations

*The red bucket icon represents S3



The above diagram is from ACG

EBS Volume Types Chart

Via https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/ebs-volume-types.html

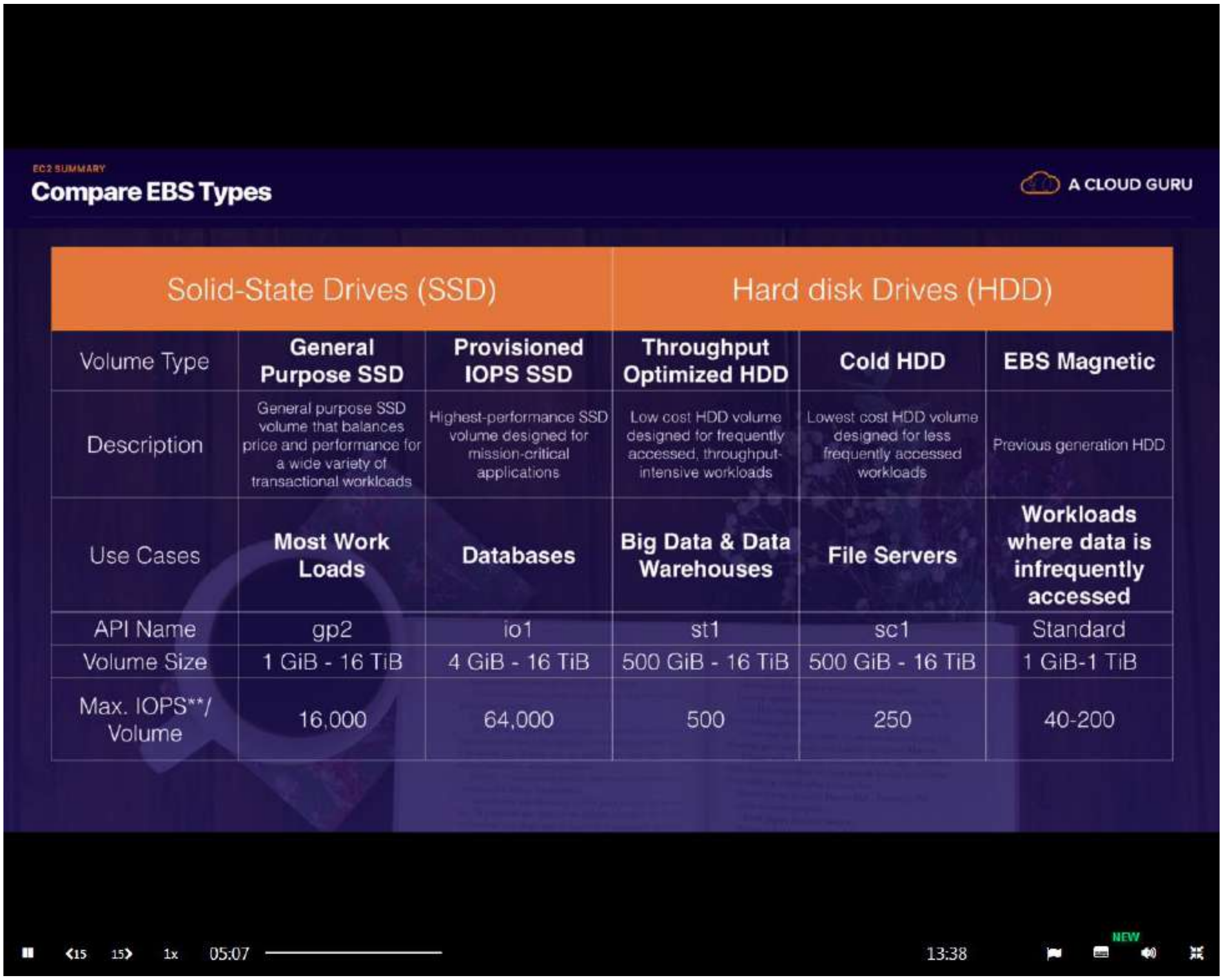
Volume characteristics

The following table describes the use cases and performance characteristics for each volume type. The default volume type is General Purpose SSD (`gp2`).

	Solid-state drives (SSD)		**Hard disk drives (HDD)**	
Volume type	General Purpose SSD (`gp2`)	Provisioned IOPS SSD (`io1`)	Throughput Optimized HDD (`st1`)	Cold HDD (`sc1`)
Description	General purpose SSD volume that balances price and performance for a wide variety of workloads	Highest-performance SSD volume for mission-critical low-latency or high-throughput workloads	Low-cost HDD volume designed for frequently accessed, throughput-intensive workloads	Lowest cost HDD volume designed for less frequently accessed workloads
Use cases	• Recommended for most workloads • System boot volumes • Virtual desktops • Low-latency interactive apps • Development and test environments	• Critical business applications that require sustained IOPS performance, or more than 16,000 IOPS or 250 MiB/s of throughput per volume • Large database workloads, such as: • MongoDB • Cassandra • Microsoft SQL Server • MySQL • PostgreSQL • Oracle	• Streaming workloads requiring consistent, fast throughput at a low price • Big data • Data warehouses • Log processing • Cannot be a boot	• Throughput-oriented storage for large volumes of data that is infrequently accessed • Scenarios where the lowest storage cost is important • Cannot be a boot volume

			volume	
API name	`gp2`	`io1`	`st1`	`sc1`
Volume size	1 GiB - 16 TiB	4 GiB - 16 TiB	500 GiB - 16 TiB	500 GiB - 16 TiB
Max IOPS per volume	16,000 (16 KiB I/O) *	64,000 (16 KiB I/O) †	500 (1 MiB I/O)	250 (1 MiB I/O)
Max throughput per volume	250 MiB/s *	1,000 MiB/s †	500 MiB/s	250 MiB/s
Max IOPS per instance ††	80,000	80,000	80,000	80,000
Max throughput per instance ††	2,375 MB/s	2,375 MB/s	2,375 MB/s	2,375 MB/s
Dominant performance attribute	IOPS	IOPS	MiB/s	MiB/s

EBS Volume Types (continued)



Compare EBS Types

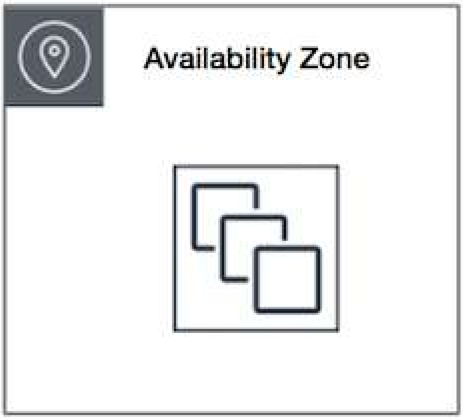
	Solid-State Drives (SSD)		Hard disk Drives (HDD)		
Volume Type	**General Purpose SSD**	**Provisioned IOPS SSD**	**Throughput Optimized HDD**	**Cold HDD**	**EBS Magnetic**
Description	General purpose SSD volume that balances price and performance for a wide variety of transactional workloads	Highest-performance SSD volume designed for mission-critical applications	Low cost HDD volume designed for frequently accessed, throughput-intensive workloads	Lowest cost HDD volume designed for less frequently accessed workloads	Previous generation HDD
Use Cases	**Most Work Loads**	**Databases**	**Big Data & Data Warehouses**	**File Servers**	**Workloads where data is infrequently accessed**
API Name	gp2	io1	st1	sc1	Standard
Volume Size	1 GiB - 16 TiB	4 GiB - 16 TiB	500 GiB - 16 TiB	500 GiB - 16 TiB	1 GiB-1 TiB
Max. IOPS**/ Volume	16,000	64,000	500	250	40-200

The above chart is from ACG

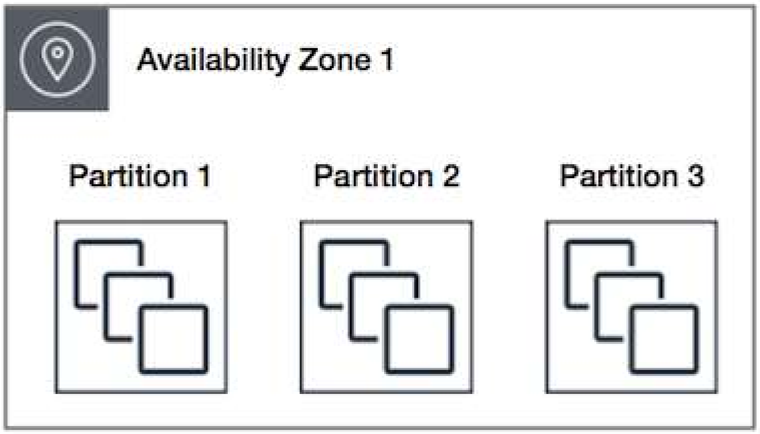
Placement Group Diagrams

via https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/placement-groups.html#placement-groups-cluster

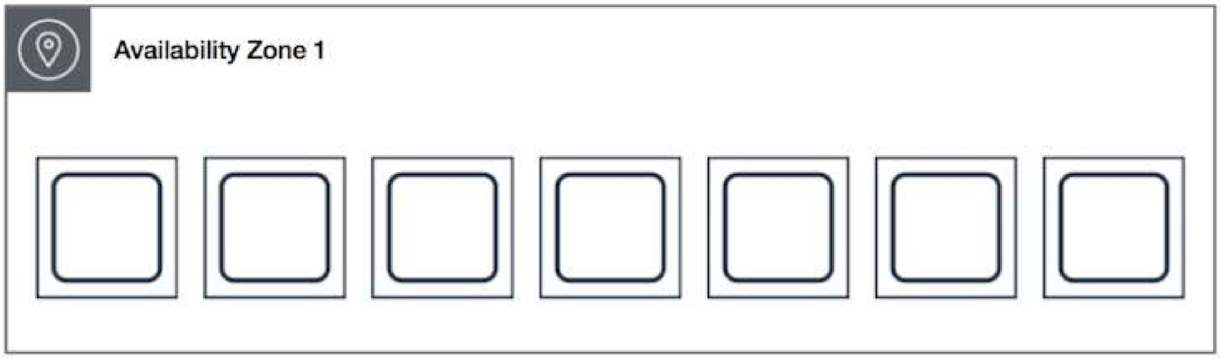
Cluster Placement Group



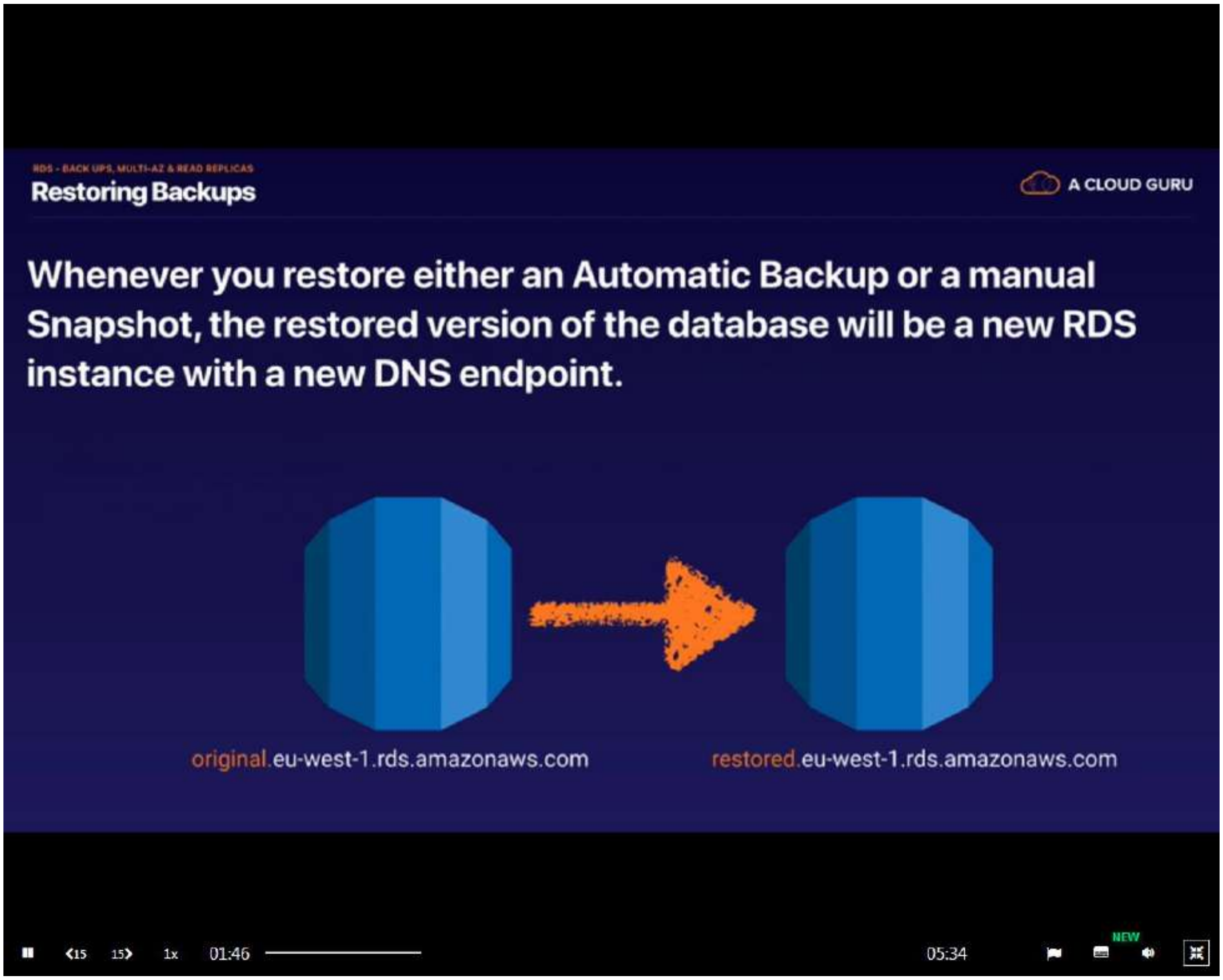
Partition Placement Group



Spread Placement Group



RDS – Restoring Backups



RDS - BACK UPS, MULTI-AZ & READ REPLICAS

Restoring Backups

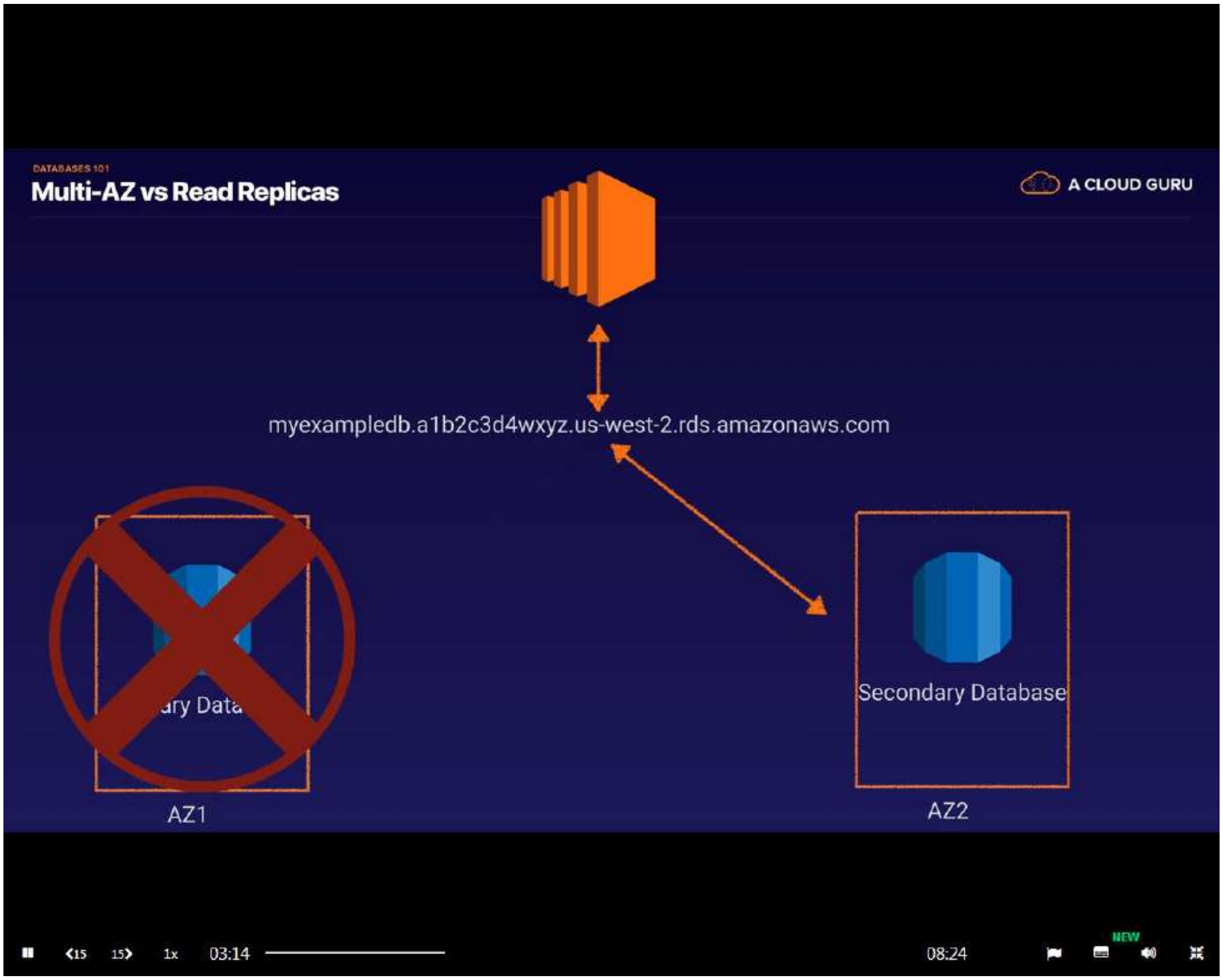
Whenever you restore either an Automatic Backup or a manual Snapshot, the restored version of the database will be a new RDS instance with a new DNS endpoint.

A CLOUD GURU
original.eu-west-1.rds.amazonaws.com
restored.eu-west-1.rds.amazonaws.com
‹15
15›
1x
01:46
05:34
NEW

The above diagram is from ACG

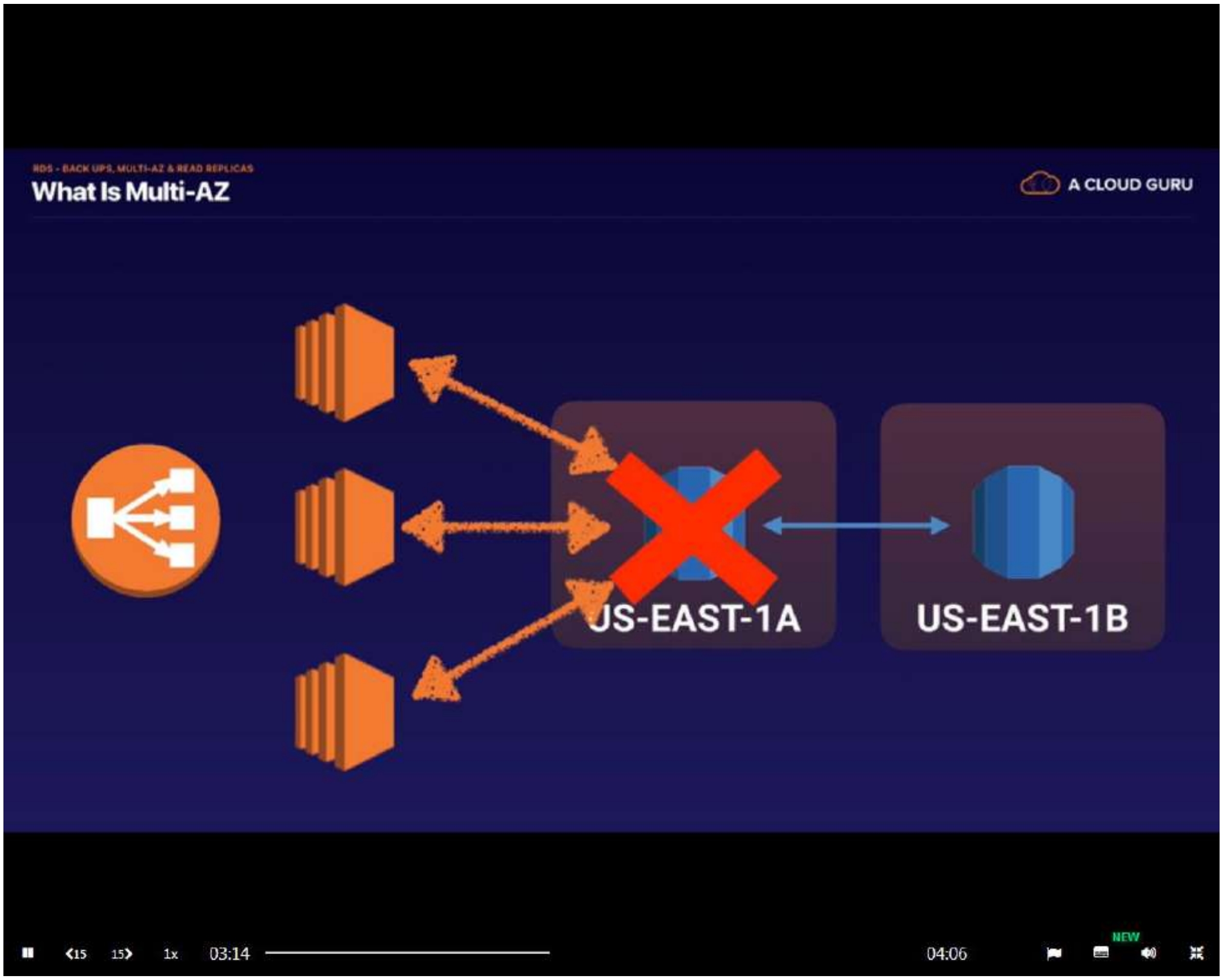
Multi-AZ

Multi-AZ keeps an exact copy of your production database in another AZ (for disaster recovery only). When you write to your primary AWS automatically synchronizes the changes to the standby database. In the event of failure, RDS will failover to the standby instance automatically. You keep the same DNS Endpoint (Amazon updates the IP address to point from one RDS instance to the other and Amazon does that for you so you don't have to do it)



The above diagram is from ACG

Multi-AZ (continued)



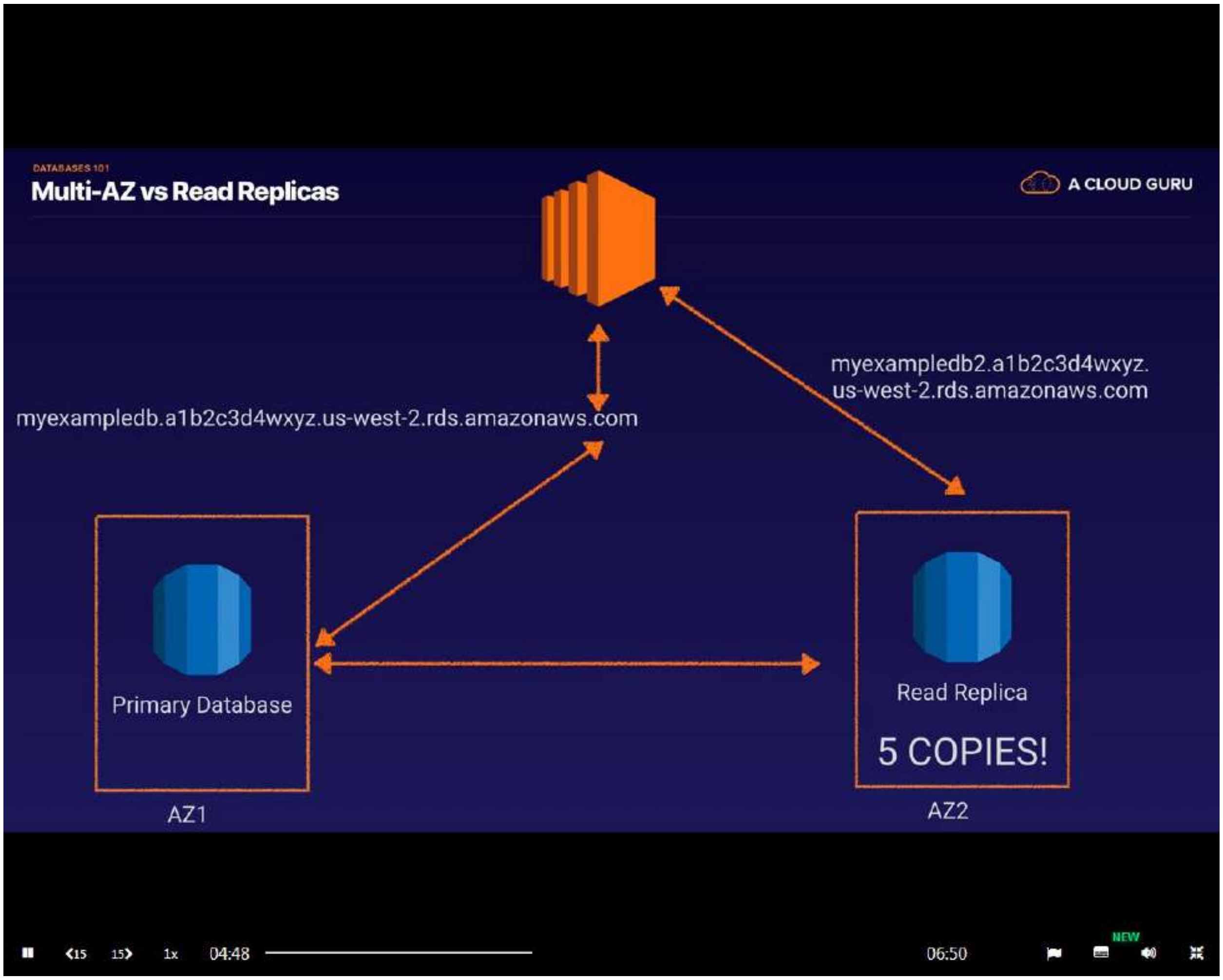
RDS - BACK UPS, MULTI-AZ & READ REPLICAS

What Is Multi-AZ

The above chart is from ACG

Read Replicas

Allow you to have a read only copy of your production database, for increased performance.



DATABASES 101

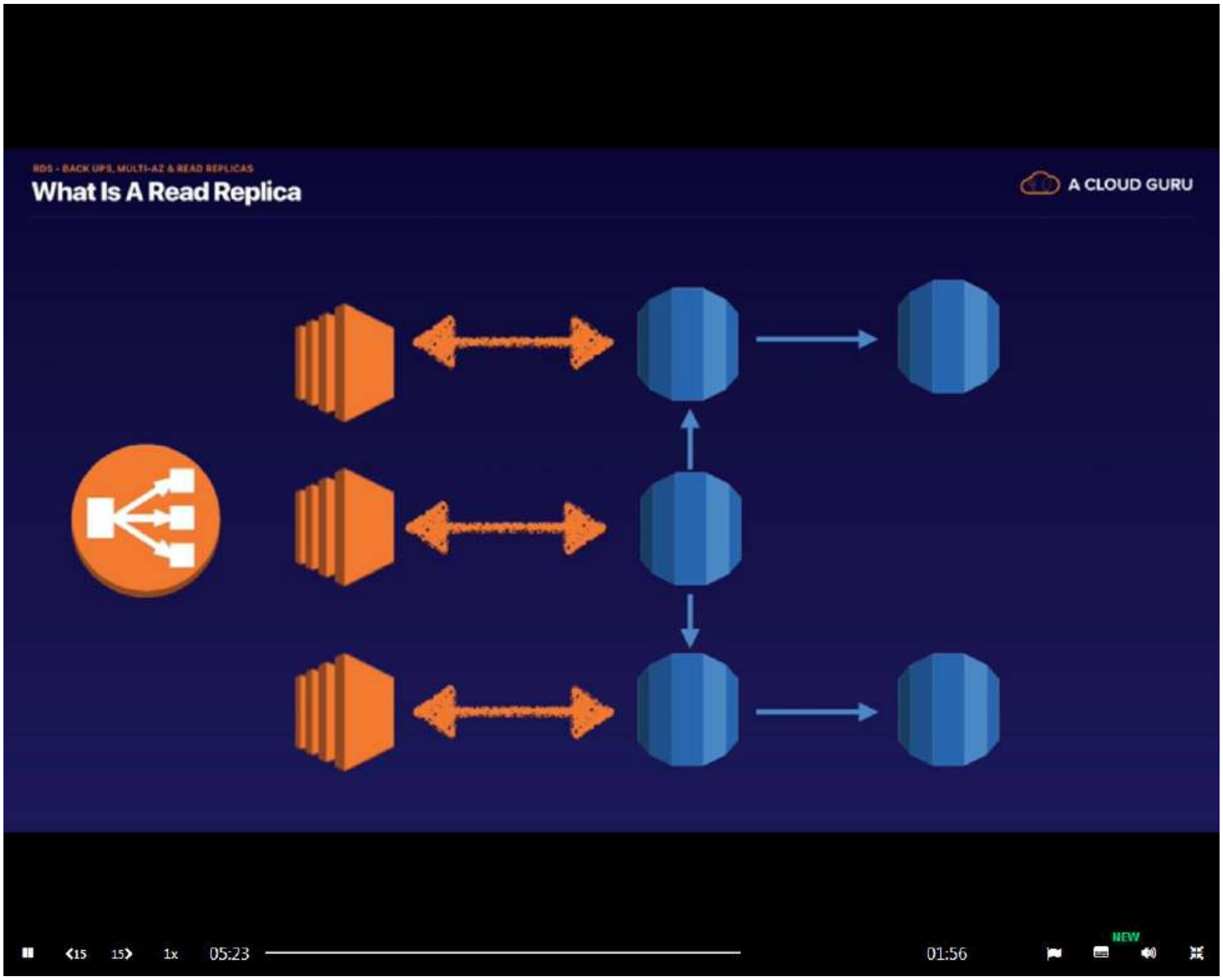
Multi-AZ vs Read Replicas

A CLOUD GURU
myexampledb2.a1b2c3d4wxyz.
us-west-2.rds.amazonaws.com
myexampledb.a1b2c3d4wxyz.us-west-2.rds.amazonaws.com
Primary Database
AZ1
Read Replica
5 COPIES!
AZ2
15
15
1x
04:48
06:50
NEW

The above diagram is from ACG

Read Replicas (continued)

Read Replicas are replicated **asynchronously**. It is possible to architect EC2 instances to read from different read replicas but only write to a single database.

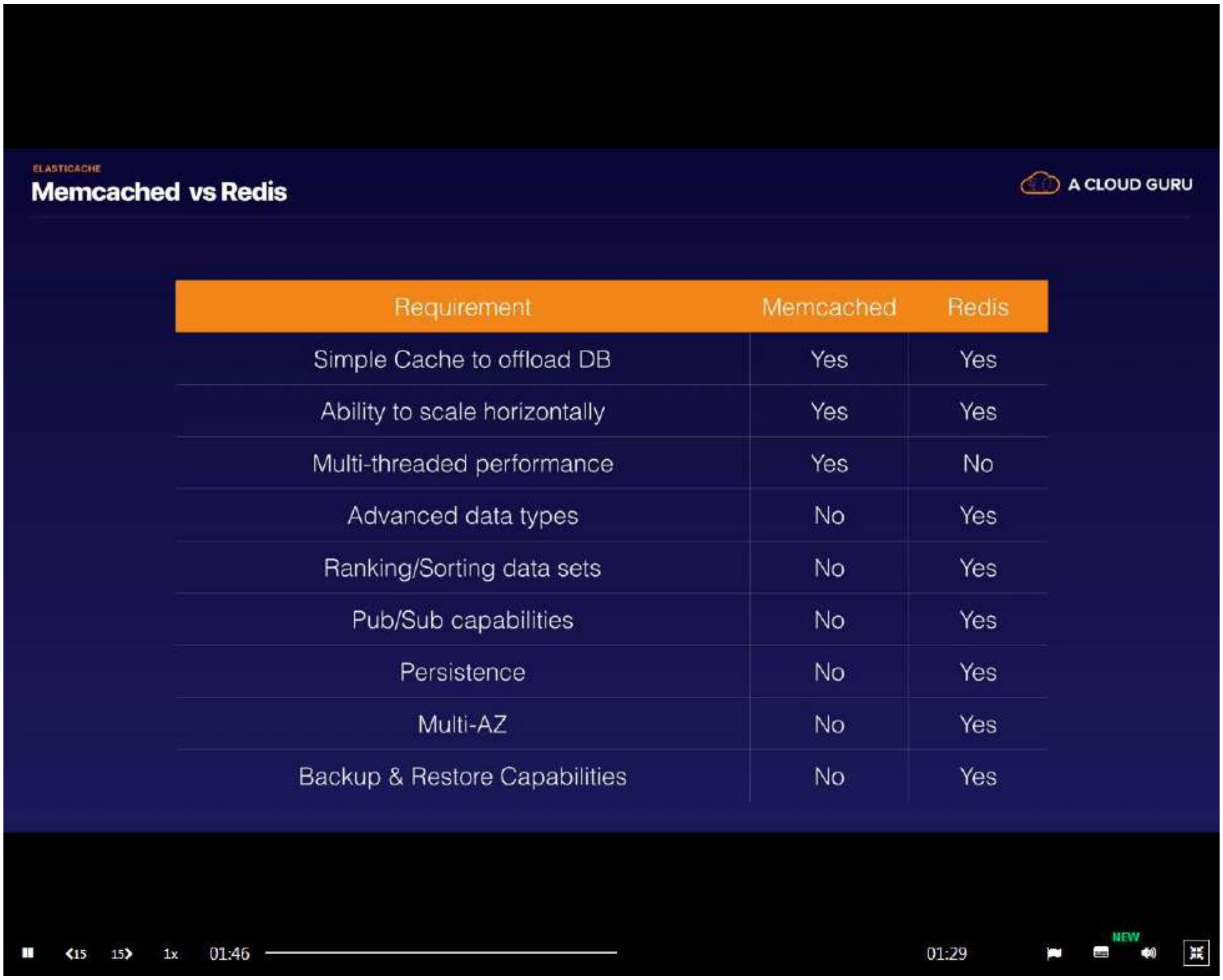


RDS - BACK UPS, MULTI-AZ & READ REPLICAS

What Is A Read Replica

The above diagram is from ACG

Elasticache – Memcached vs Redis Chart



ELASTICACHE

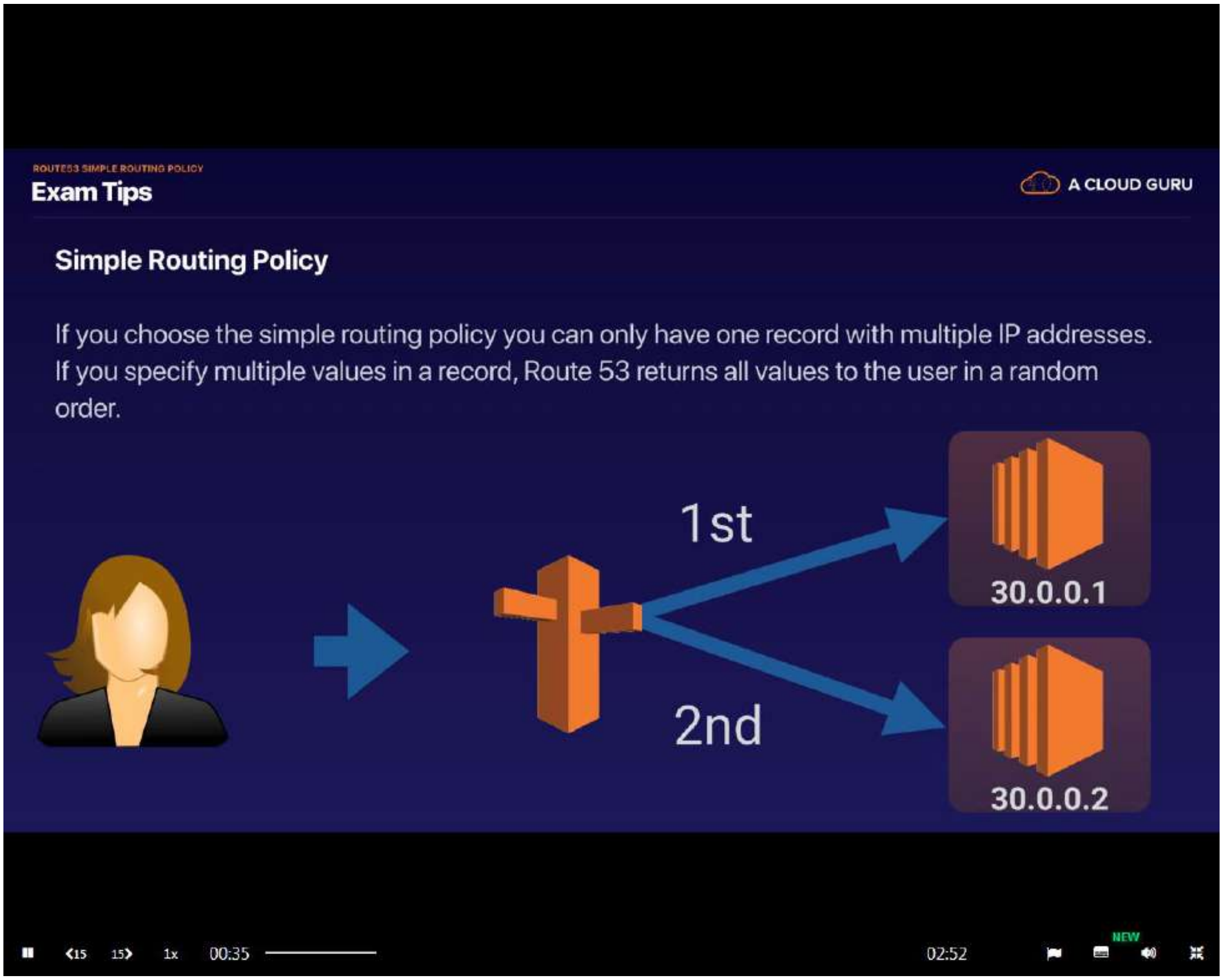
Memcached vs Redis

Requirement	Memcached	Redis
Simple Cache to offload DB	Yes	Yes
Ability to scale horizontally	Yes	Yes
Multi-threaded performance	Yes	No
Advanced data types	No	Yes
Ranking/Sorting data sets	No	Yes
Pub/Sub capabilities	No	Yes
Persistence	No	Yes
Multi-AZ	No	Yes
Backup & Restore Capabilities	No	Yes

The above chart is from ACG

DNS/Route53 Diagrams

Simple Routing Policy



ROUTE53 SIMPLE ROUTING POLICY

Exam Tips

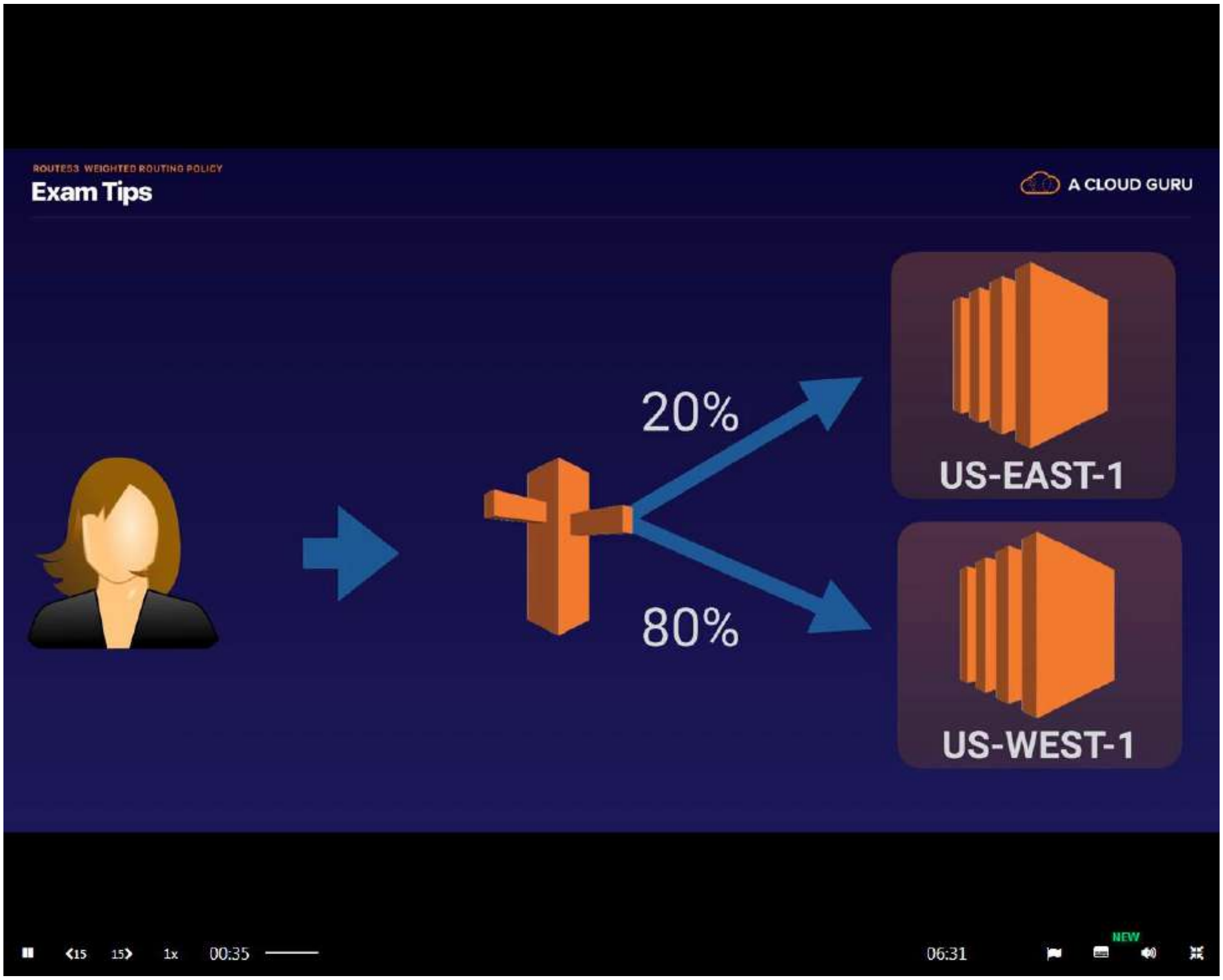
A CLOUD GURU

Simple Routing Policy

If you choose the simple routing policy you can only have one record with multiple IP addresses. If you specify multiple values in a record, Route 53 returns all values to the user in a random order.

The above diagram is from ACG

Weighted Routing Policy

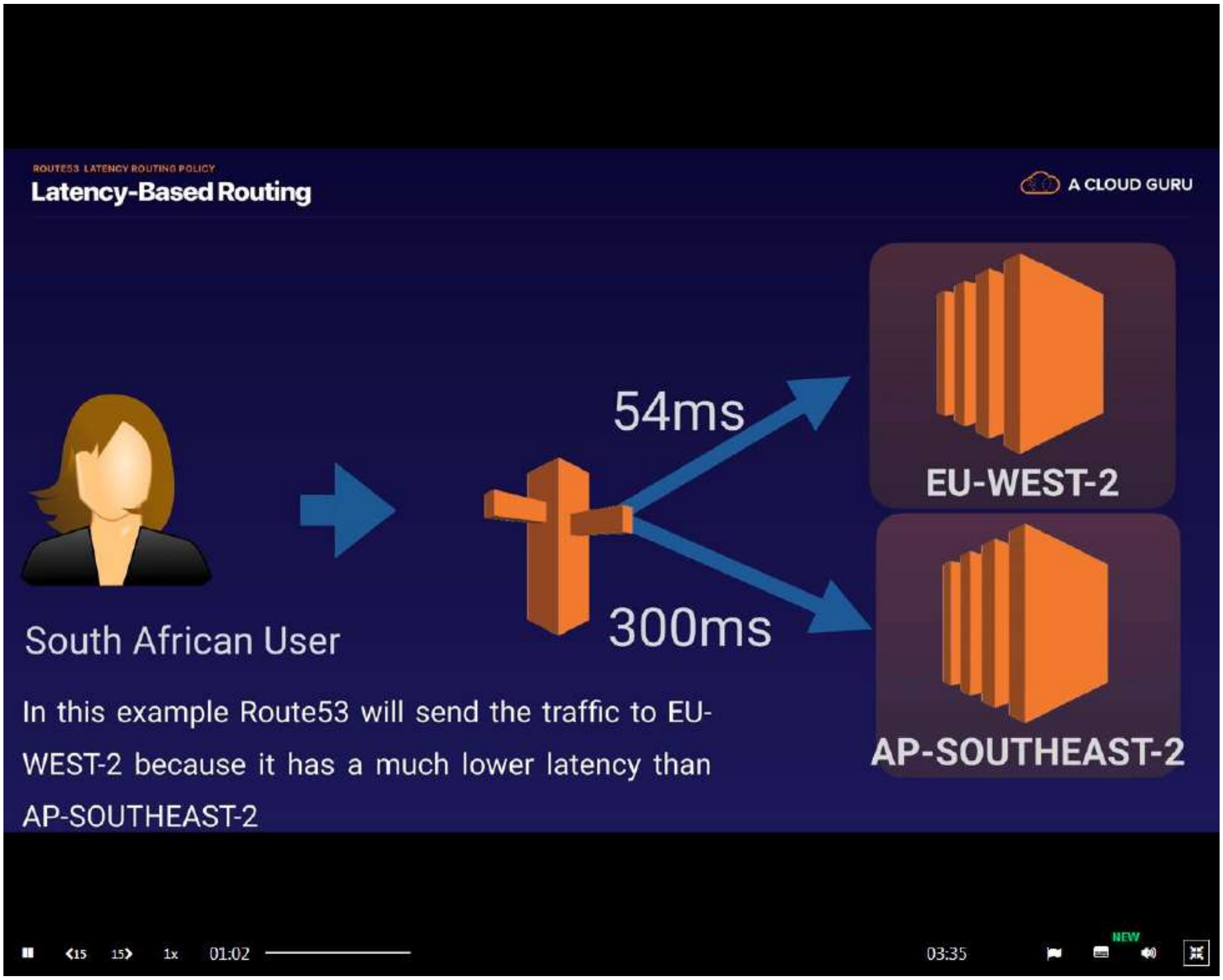


ROUTE53 WEIGHTED ROUTING POLICY

Exam Tips

The above diagram is from ACG

Latency-Based Routing



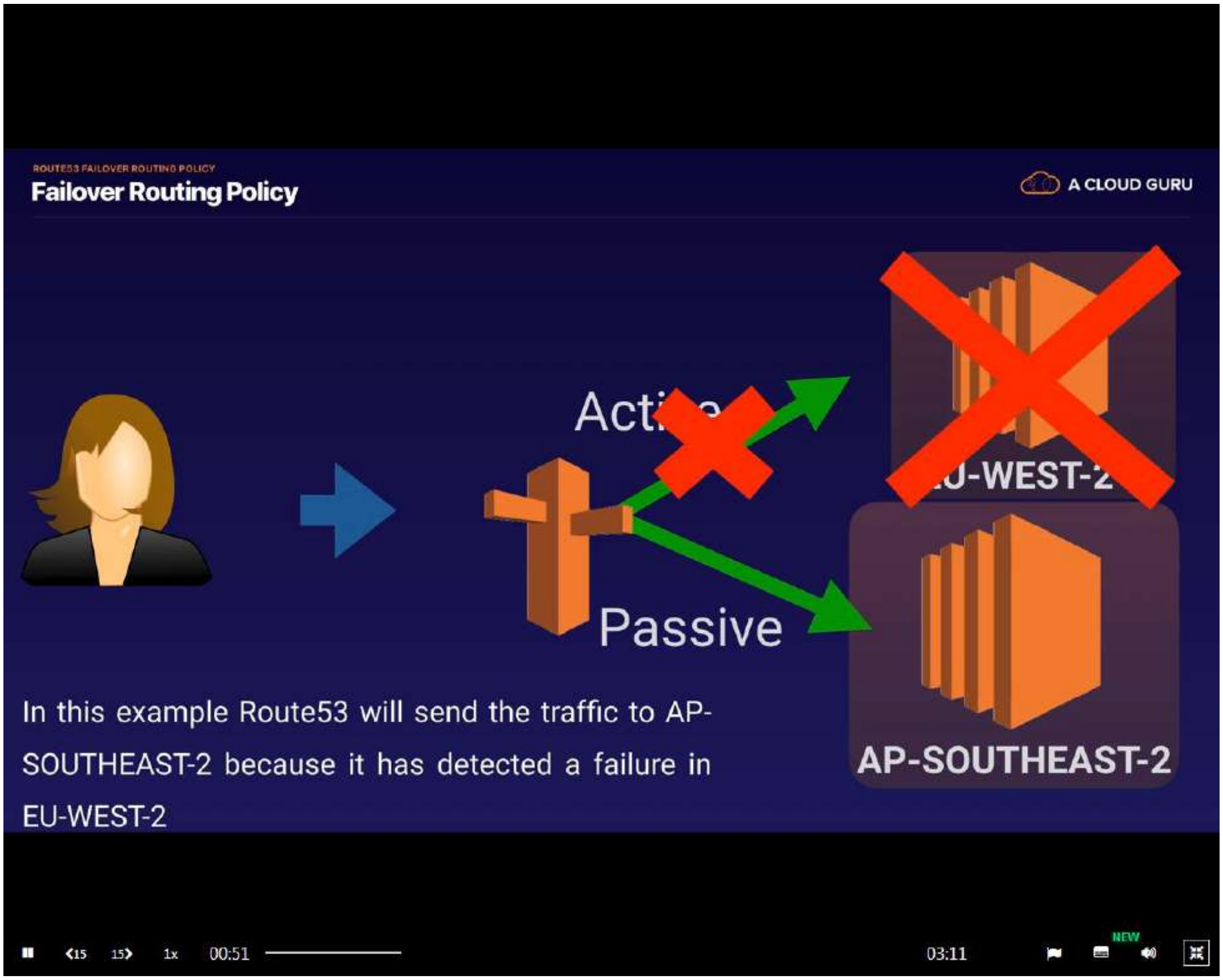
ROUTE53 LATENCY ROUTING POLICY

Latency-Based Routing

In this example Route53 will send the traffic to EU-WEST-2 because it has a much lower latency than AP-SOUTHEAST-2

The above diagram is from ACG

Failover Routing Policy



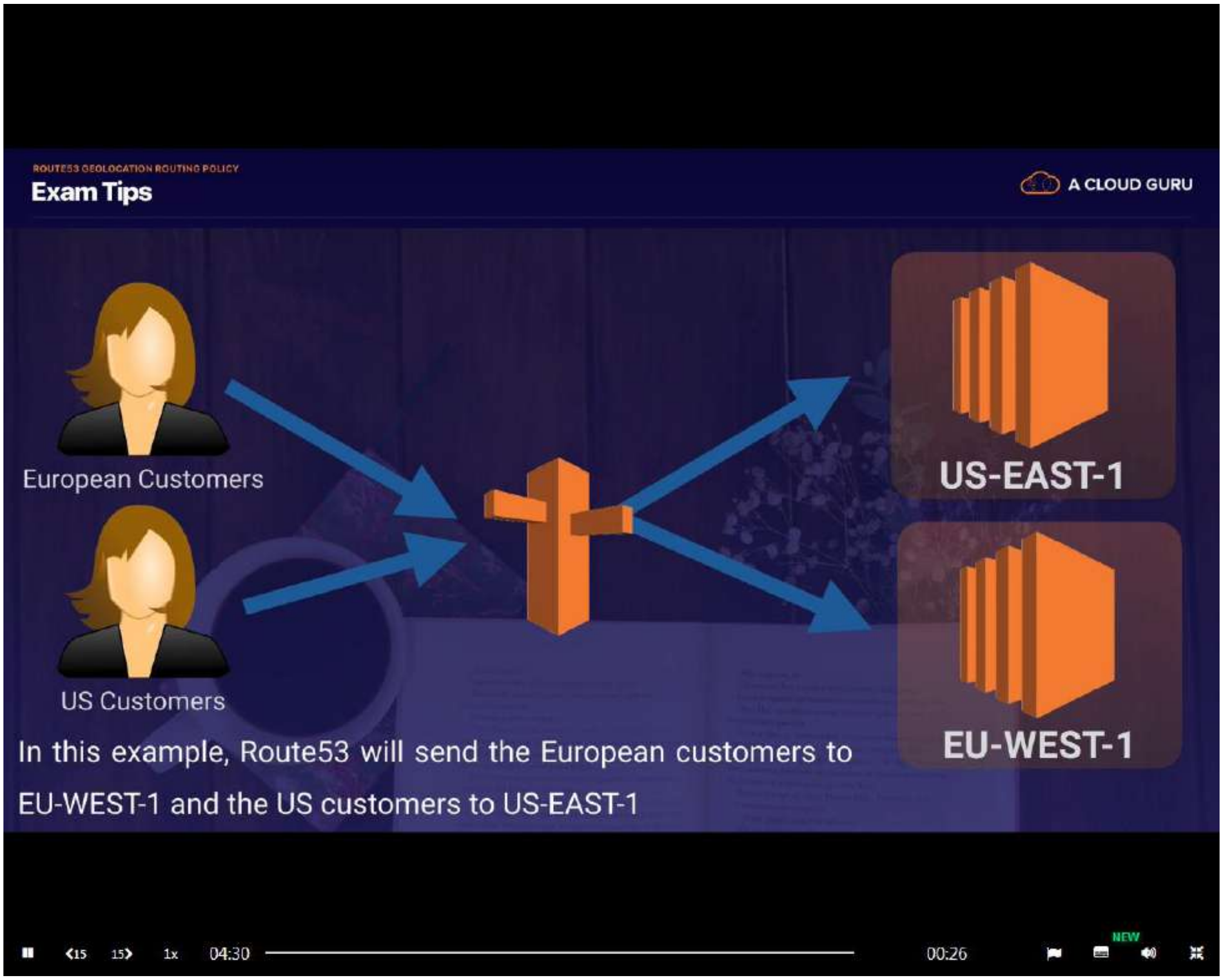
ROUTE53 FAILOVER ROUTING POLICY

Failover Routing Policy

In this example Route53 will send the traffic to AP-SOUTHEAST-2 because it has detected a failure in EU-WEST-2

The above diagram is from ACG

Geolocation Routing Policy



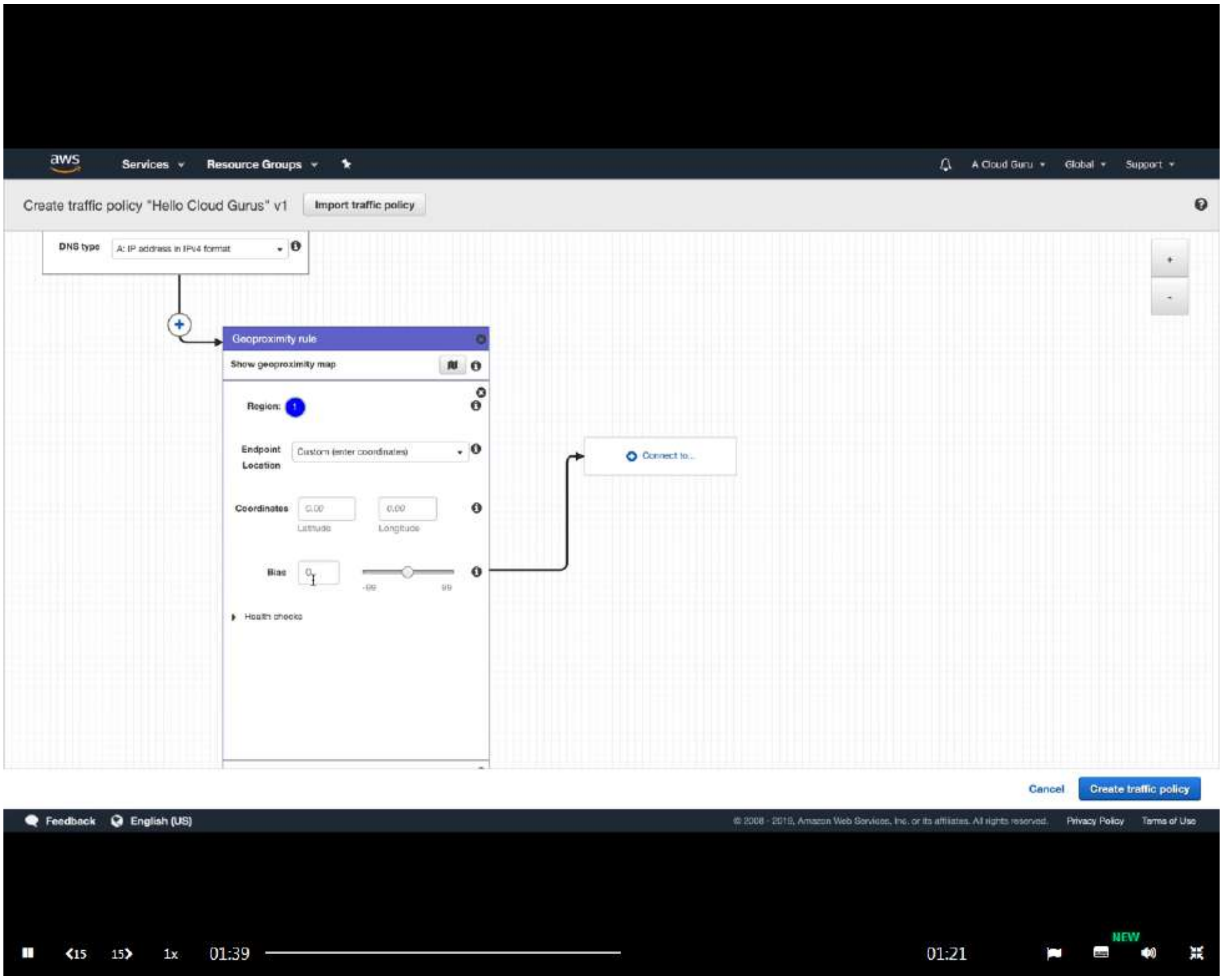
ROUTE53 GEOLOCATION ROUTING POLICY

Exam Tips

In this example, Route53 will send the European customers to EU-WEST-1 and the US customers to US-EAST-1

The above diagram is from ACG

Geoproximity Routing (Traffic Flow only mode)



Create traffic policy "Hello Cloud Gurus" v1

DNS type: A: IP address in IPv4 format

Geoproximity rule

Region: 1

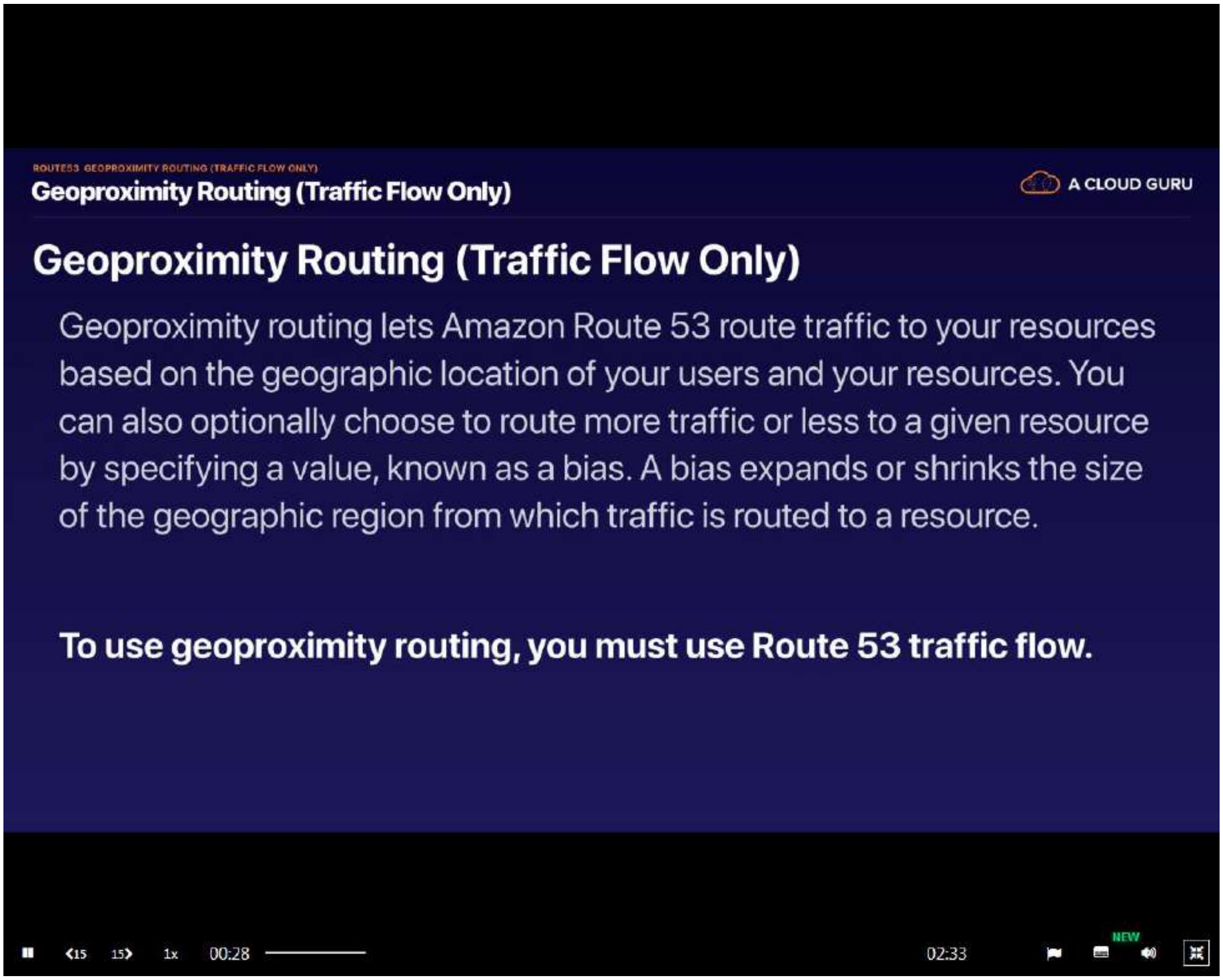
Endpoint Location: Custom (enter coordinates)

Coordinates: 0.00 (Latitude) 0.00 (Longitude)

Bias: 0

The above diagram is from ACG

Geoproximity Routing (continued)



ROUTE53 GEOPROXIMITY ROUTING (TRAFFIC FLOW ONLY)

Geoproximity Routing (Traffic Flow Only)

A CLOUD GURU

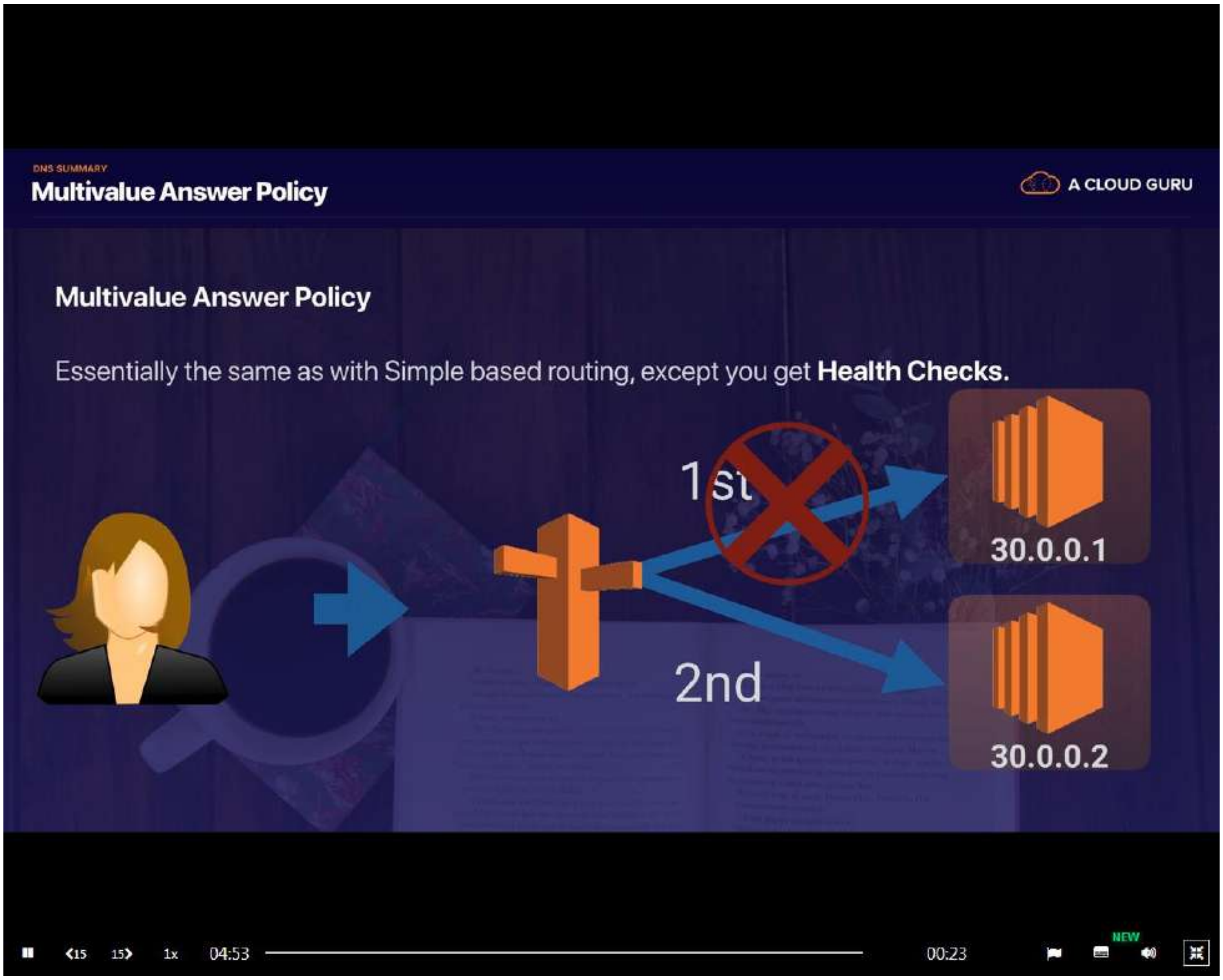
Geoproximity Routing (Traffic Flow Only)

Geoproximity routing lets Amazon Route 53 route traffic to your resources based on the geographic location of your users and your resources. You can also optionally choose to route more traffic or less to a given resource by specifying a value, known as a bias. A bias expands or shrinks the size of the geographic region from which traffic is routed to a resource.

To use geoproximity routing, you must use Route 53 traffic flow.

The above information is from ACG

Multivalue Answer Policy



DNS SUMMARY

Multivalue Answer Policy

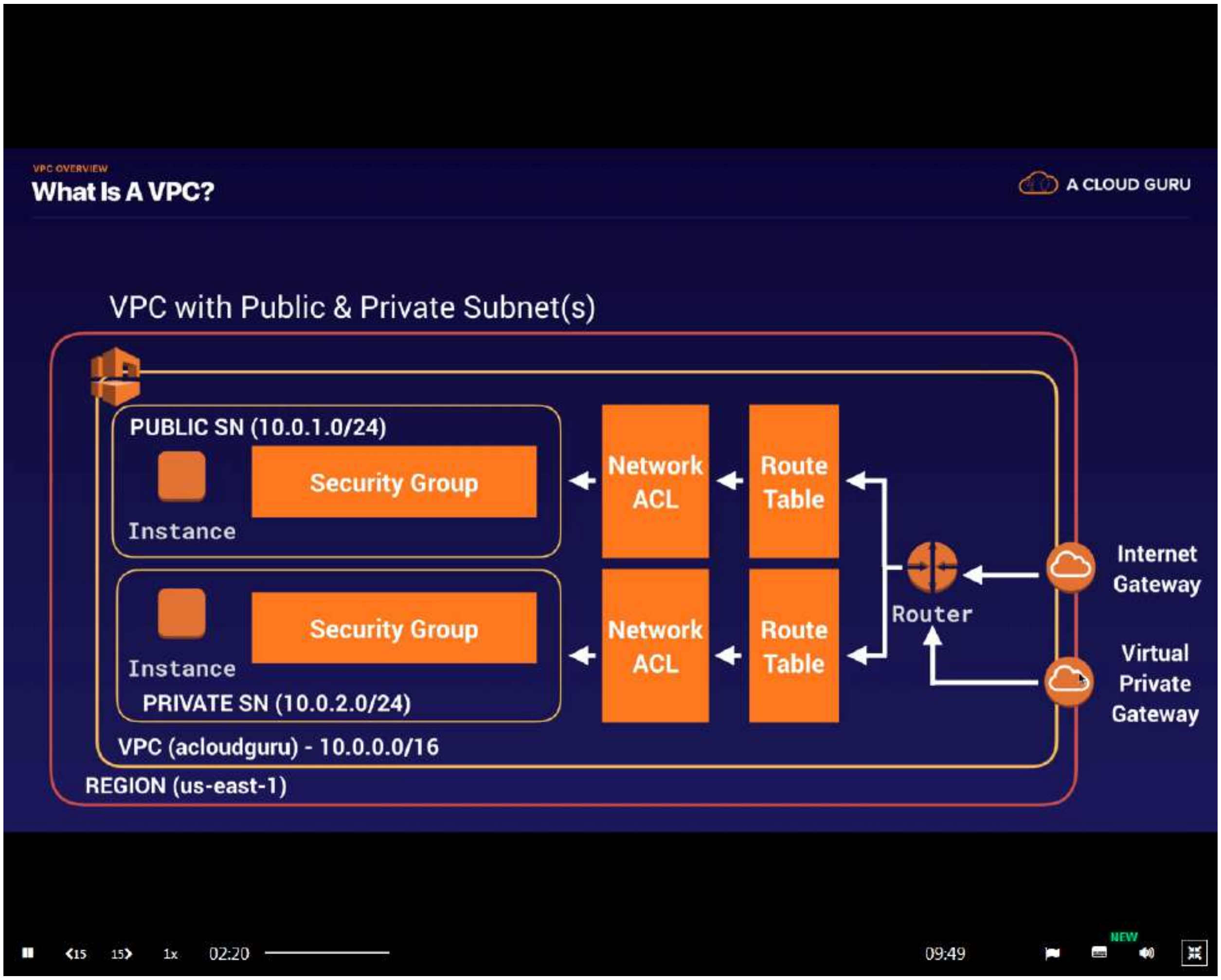
Multivalue Answer Policy

Essentially the same as with Simple based routing, except you get **Health Checks.**

The above diagram is from ACG

VPC Diagram

When you create a custom VPC a Route Table, Network ACL, and Security Group are created by default. Then it is up to you to create subnets, create an internet Gateway to be attached to your VPC, create instances, etc.



VPC OVERVIEW

What Is A VPC?

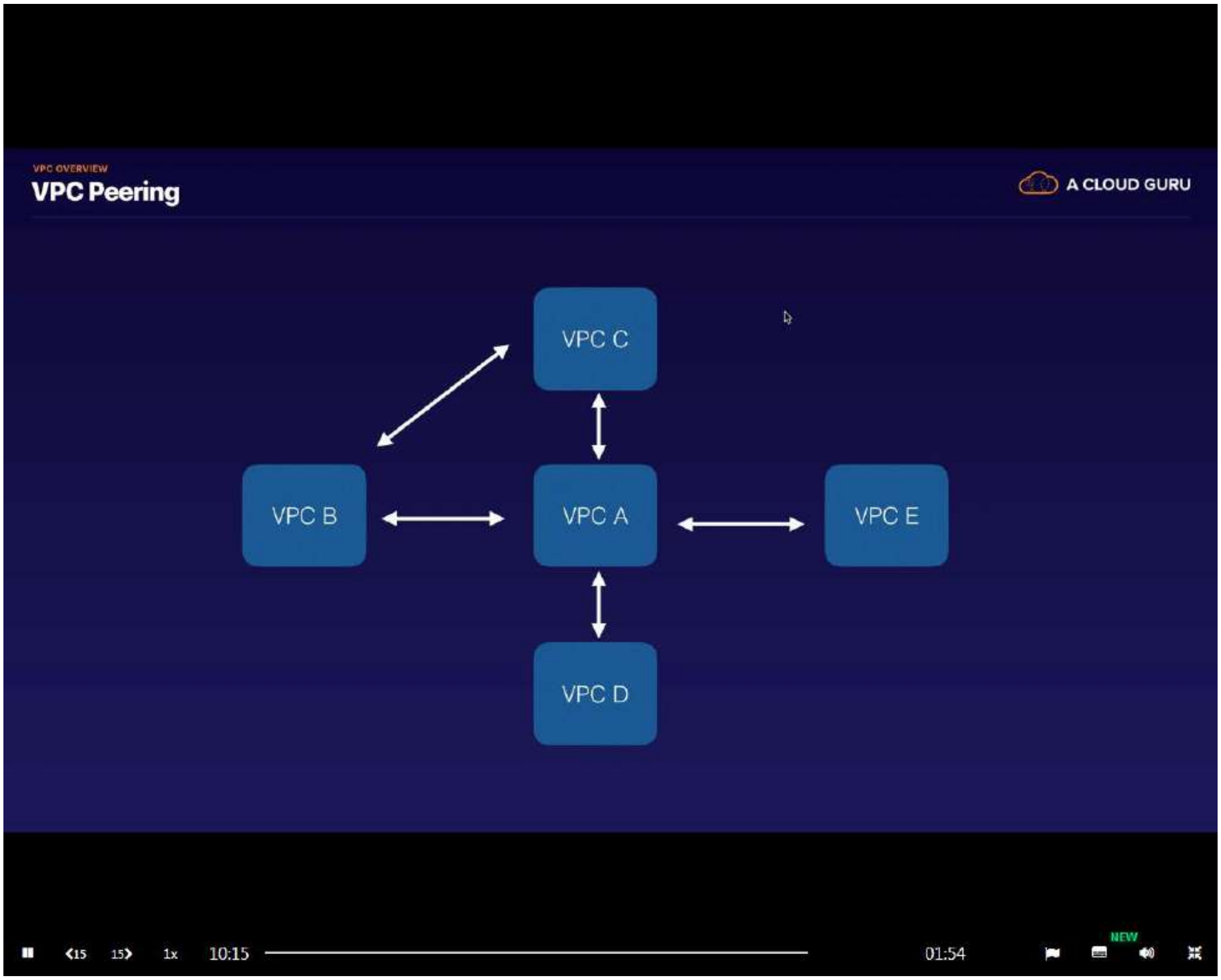
A CLOUD GURU

VPC with Public & Private Subnet(s)

The above diagram is from ACG

VPC Peering

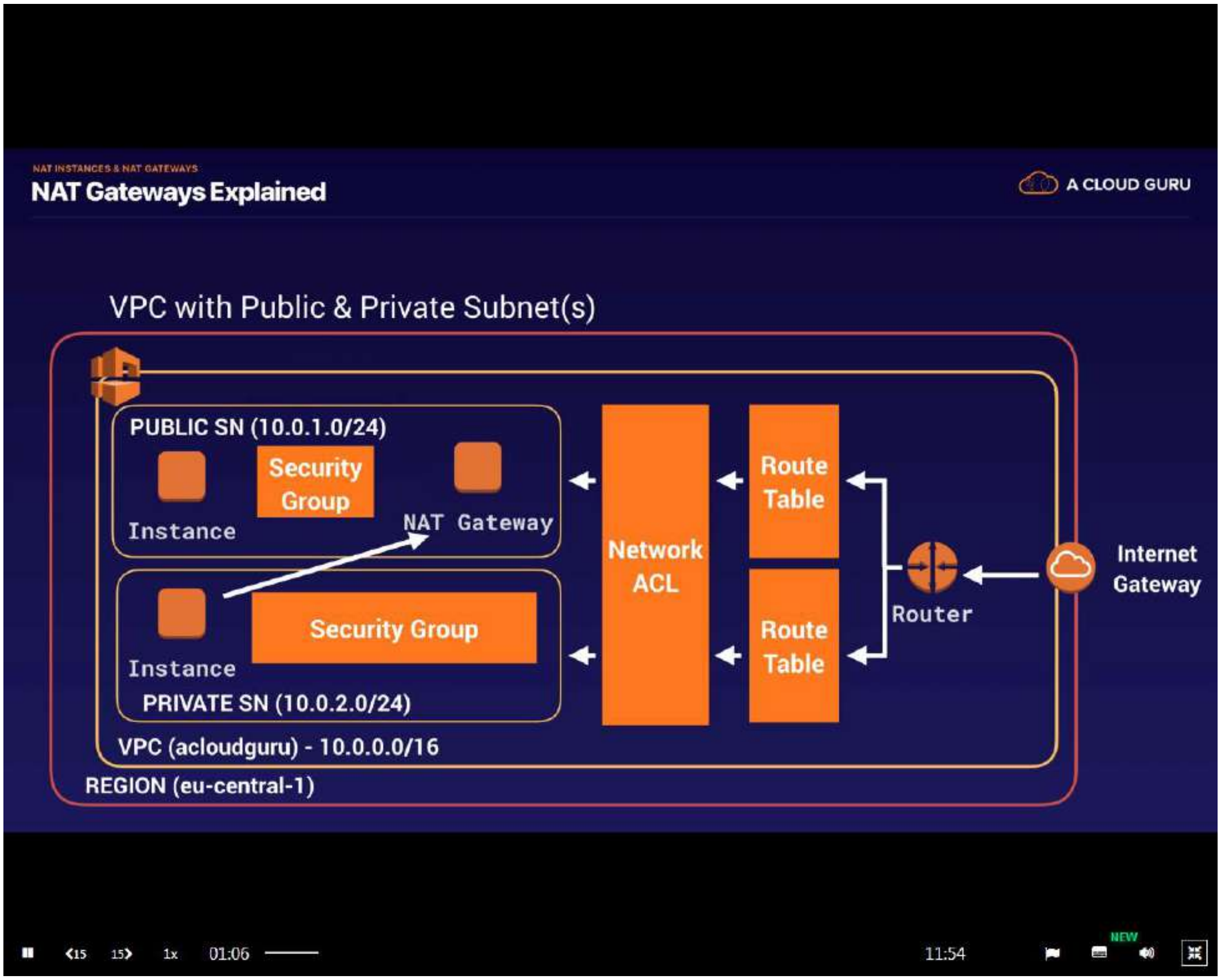
No Transitive Peering. You cannot peer through one VPC to another. You have to establish a new peering relationship. So if VPC B wants to talk to VPC C you have to establish a connection between the two.



The above diagram is from ACG

NAT Gateway Diagram

A Nat Gateway (as well as a NAT instance) is used to provide internet traffic to EC2 instances in a private subnet. A NAT Gateway is redundant. A Nat instance is an individual EC2 instance located behind a security group.

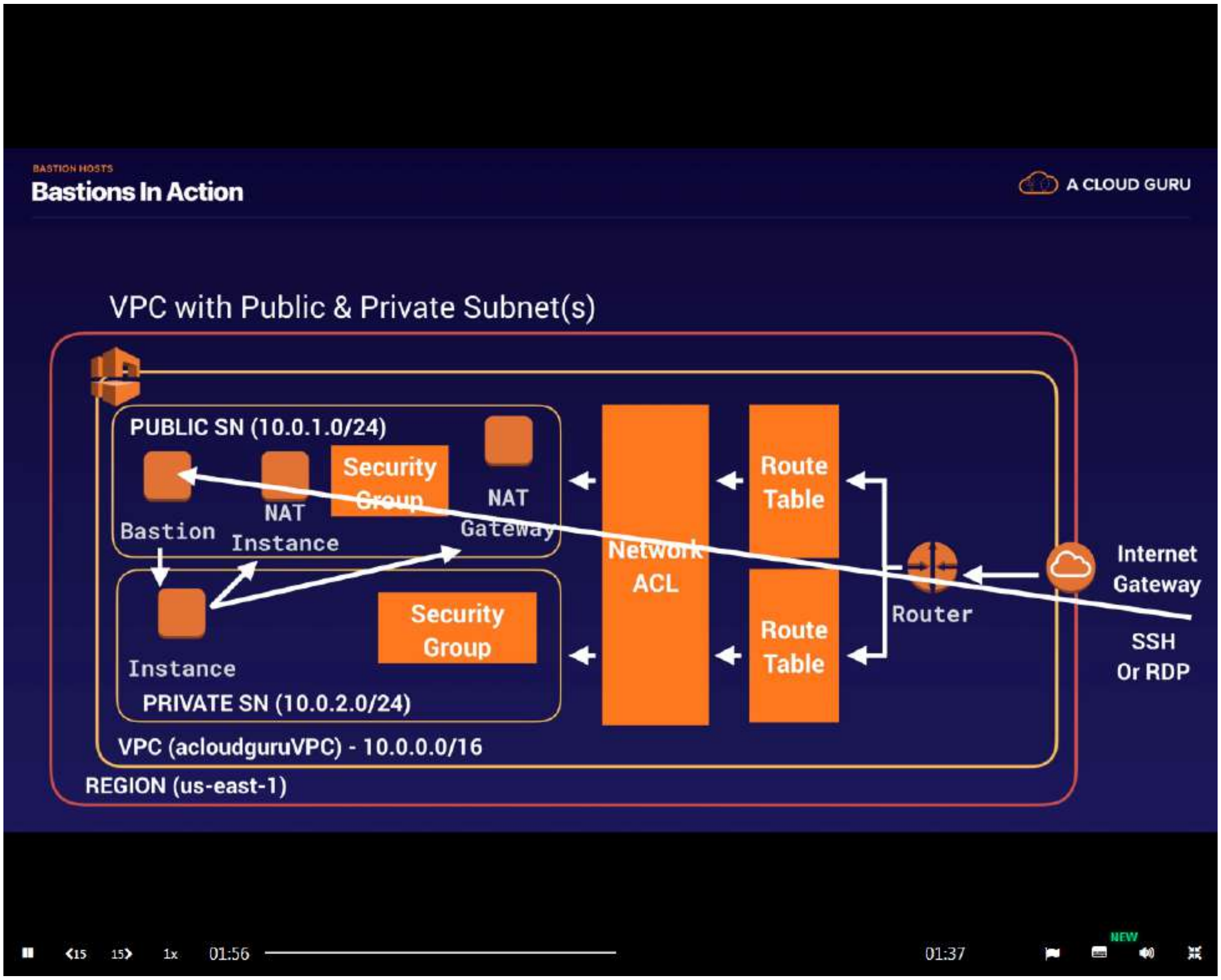


The above diagram is from ACG

Bastion Host Diagram

A Bastion is used to SSH or RDP into an instance in your private subnet.

Whereas, A Nat Gateway (as well as a NAT instance) is used to provide internet traffic to EC2 instances in a private subnet. A NAT Gateway is redundant. A Nat instance is an individual EC2 instance located **behind** a security group.



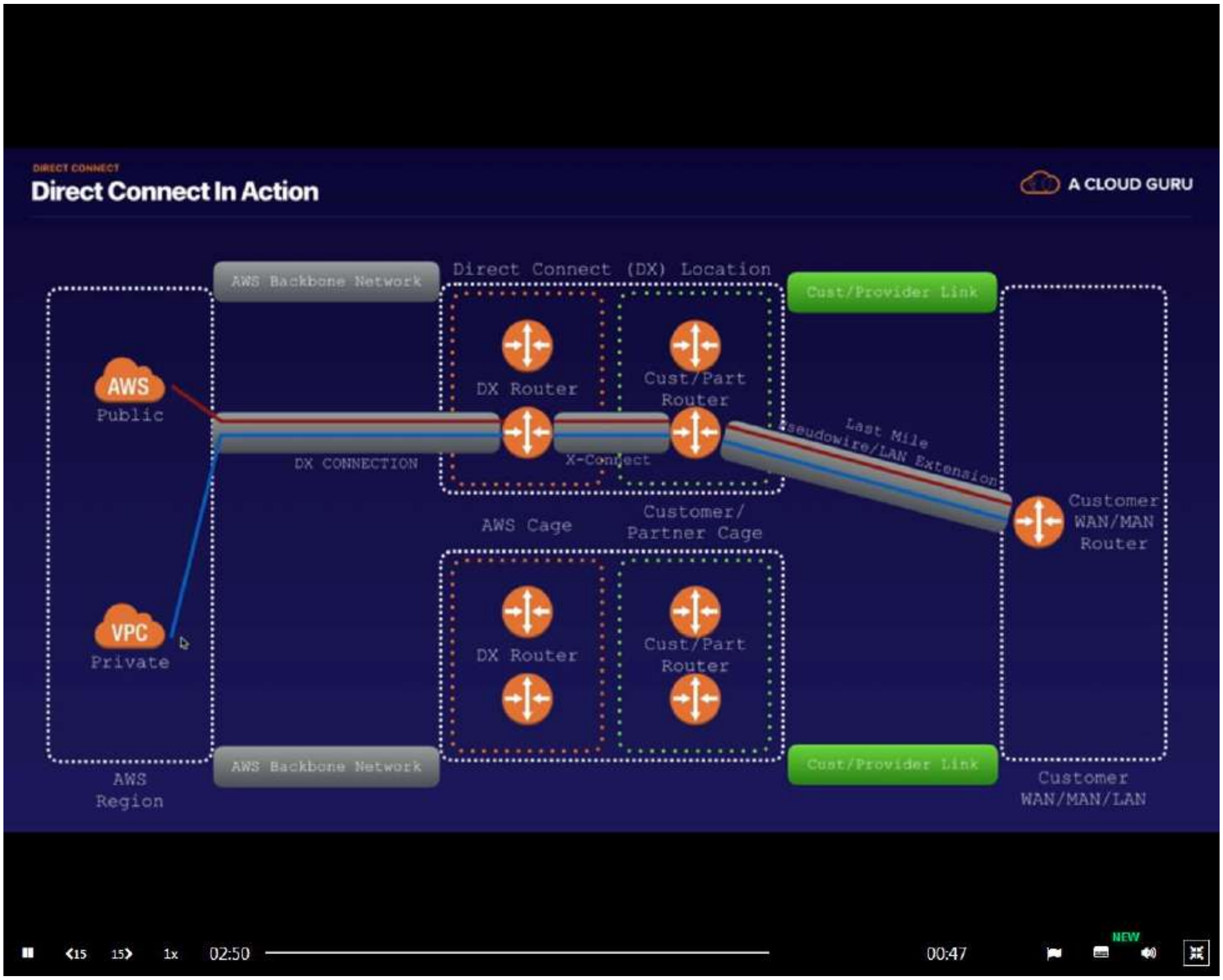
The above diagram is from ACG

Direct Connect

-Directly connects your data center to AWS

-Useful for high throughput workloads (i.e. lots of network traffic)

-Stable Reliable Secure Connection

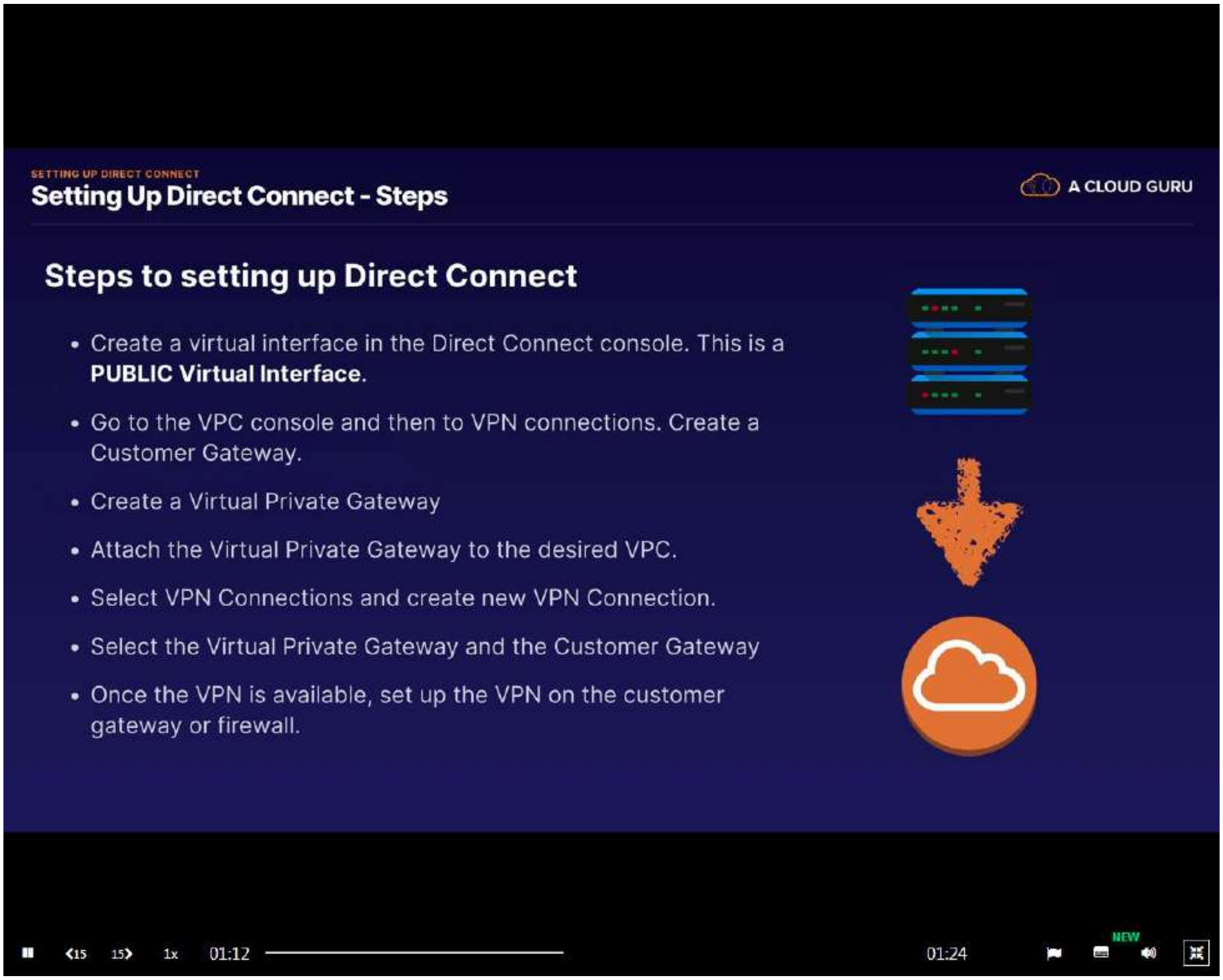


DIRECT CONNECT

Direct Connect In Action

The above diagram is from ACG

Direct Connect Setup Steps



SETTING UP DIRECT CONNECT

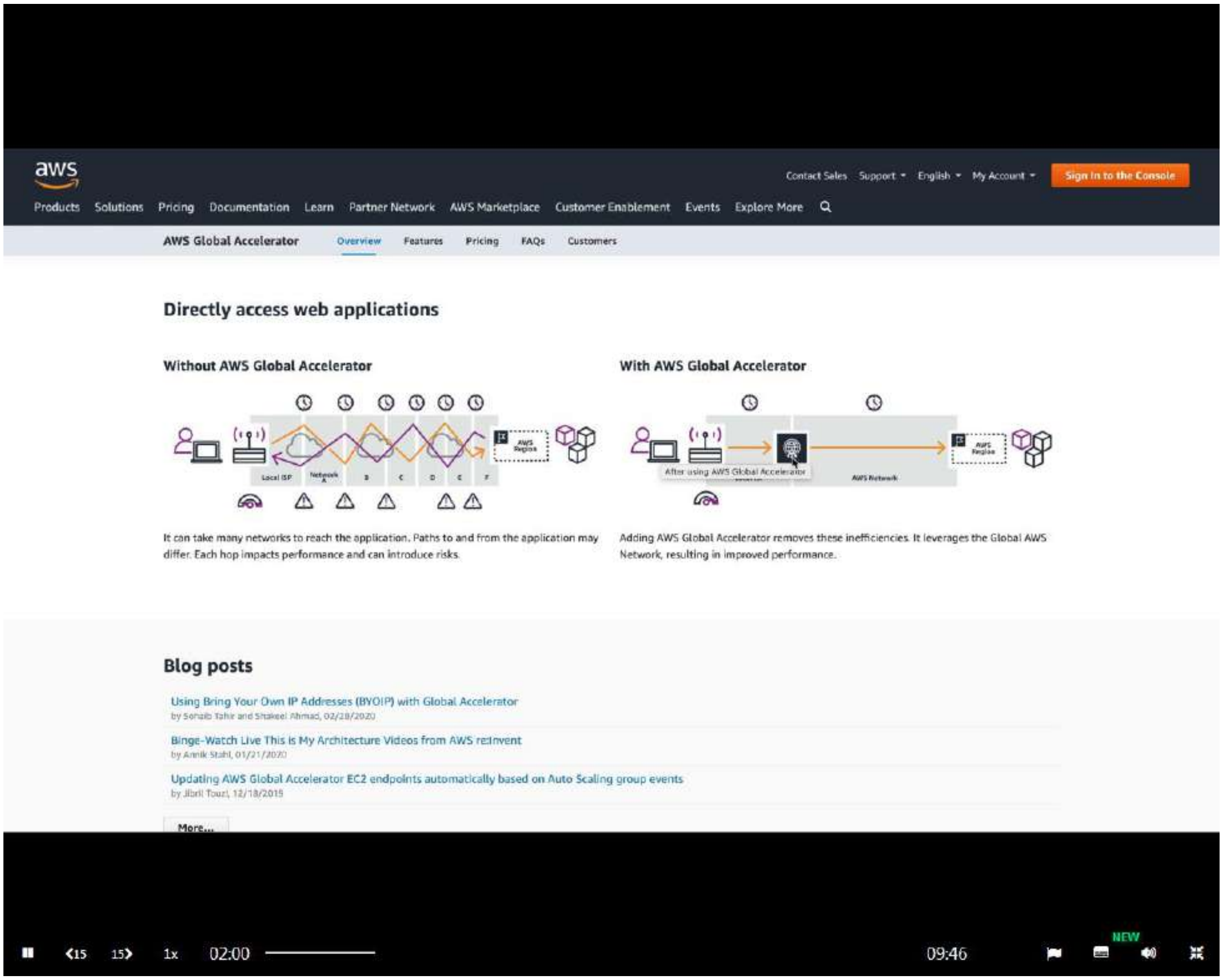
Setting Up Direct Connect - Steps

Steps to setting up Direct Connect

- Create a virtual interface in the Direct Connect console. This is a **PUBLIC Virtual Interface**.
- Go to the VPC console and then to VPN connections. Create a Customer Gateway.
- Create a Virtual Private Gateway
- Attach the Virtual Private Gateway to the desired VPC.
- Select VPN Connections and create new VPN Connection.
- Select the Virtual Private Gateway and the Customer Gateway
- Once the VPN is available, set up the VPN on the customer gateway or firewall.

The above information is from ACG

Global Accelerator



Directly access web applications

Without AWS Global Accelerator

It can take many networks to reach the application. Paths to and from the application may differ. Each hop impacts performance and can introduce risks.

With AWS Global Accelerator

Adding AWS Global Accelerator removes these inefficiencies. It leverages the Global AWS Network, resulting in improved performance.

Blog posts

Using Bring Your Own IP Addresses (BYOIP) with Global Accelerator
by Sohaib Tahir and Shakeel Ahmad, 02/28/2020

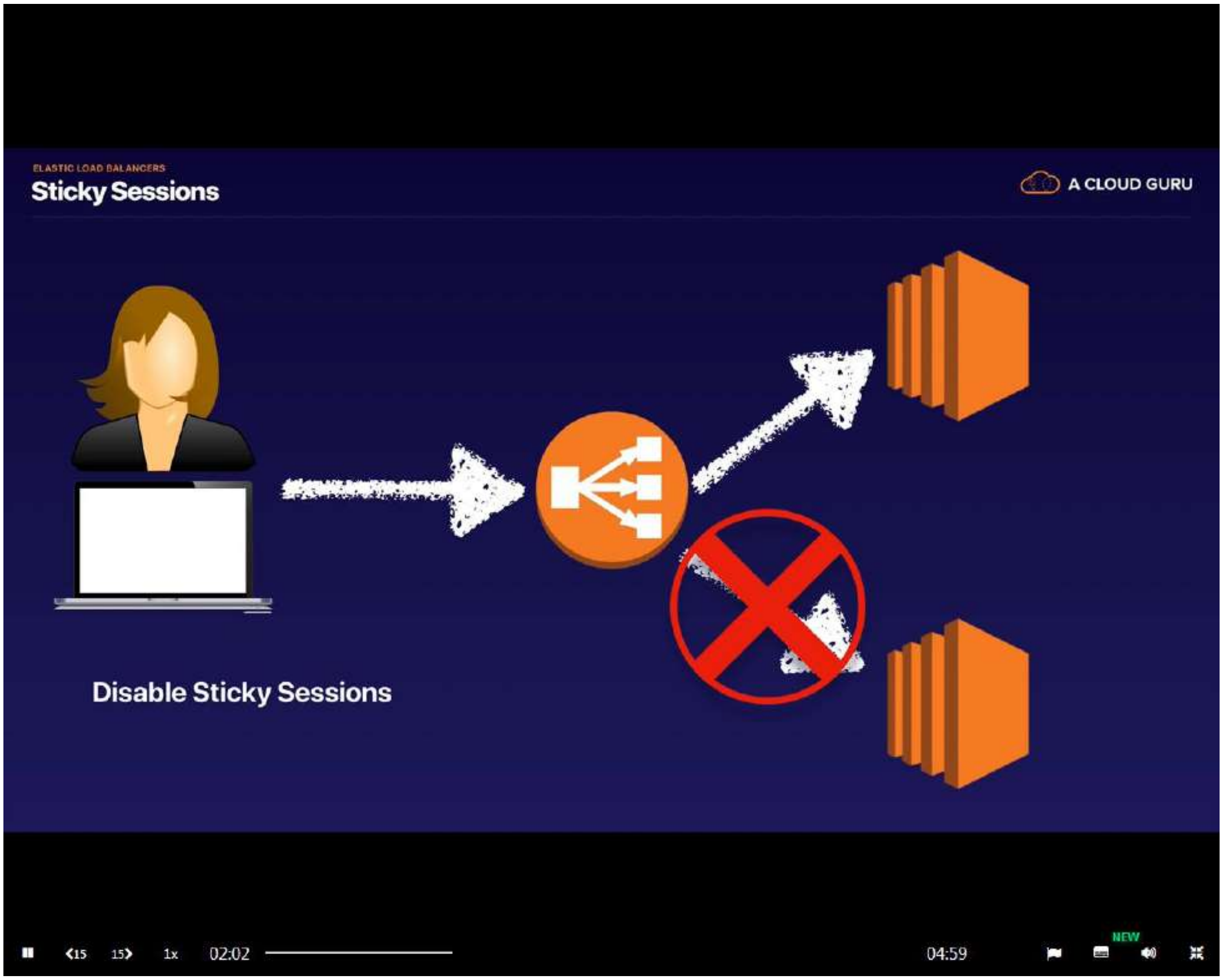
Binge-Watch Live This is My Architecture Videos from AWS re:Invent
by Annik Stahl, 01/21/2020

Updating AWS Global Accelerator EC2 endpoints automatically based on Auto Scaling group events
by Jibril Touzi, 12/18/2019

The above diagram is from ACG

Sticky Sessions

Enabling and Disabling sticky sessions given the scenario



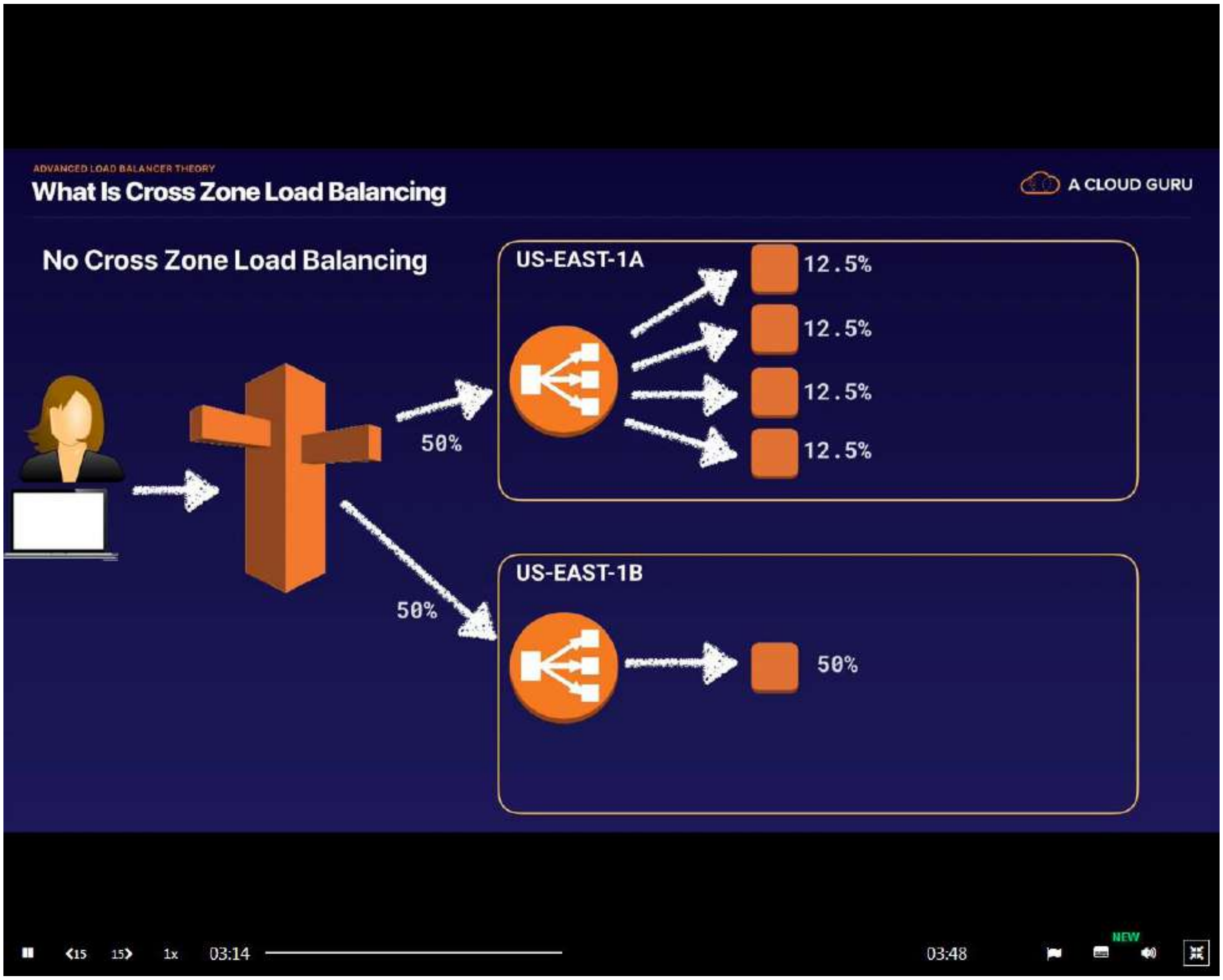
ELASTIC LOAD BALANCERS

Sticky Sessions

Disable Sticky Sessions

The above diagram is from ACG

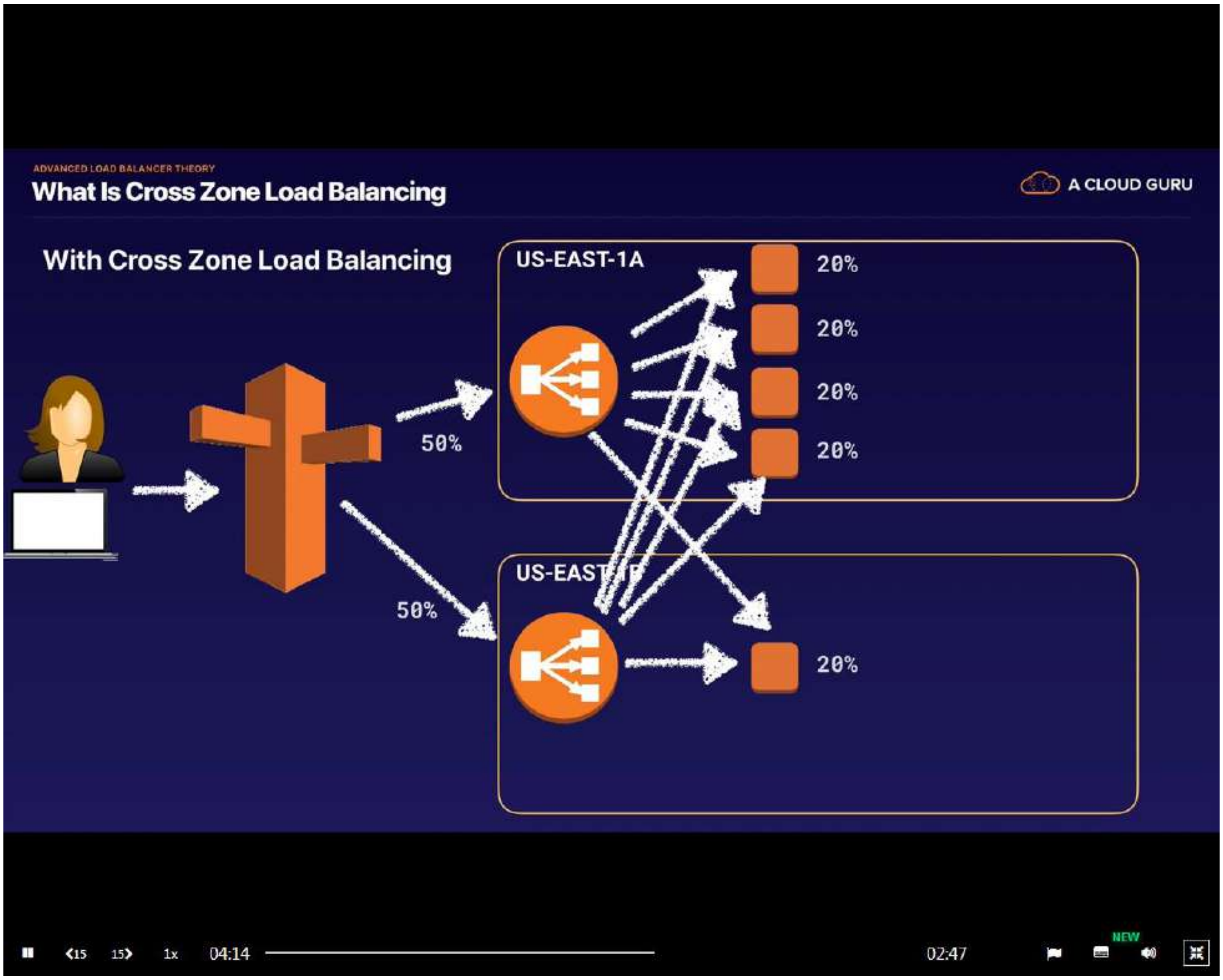
Cross Zone Load Balancing **NOT ENABLED**



ADVANCED LOAD BALANCER THEORY
What Is Cross Zone Load Balancing
A CLOUD GURU
No Cross Zone Load Balancing
US-EAST-1A
50%
12.5%
12.5%
12.5%
12.5%
US-EAST-1B
50%
50%
03:14
03:48
NEW

The above diagram is from ACG

Cross Zone Load Balancing **ENABLED**

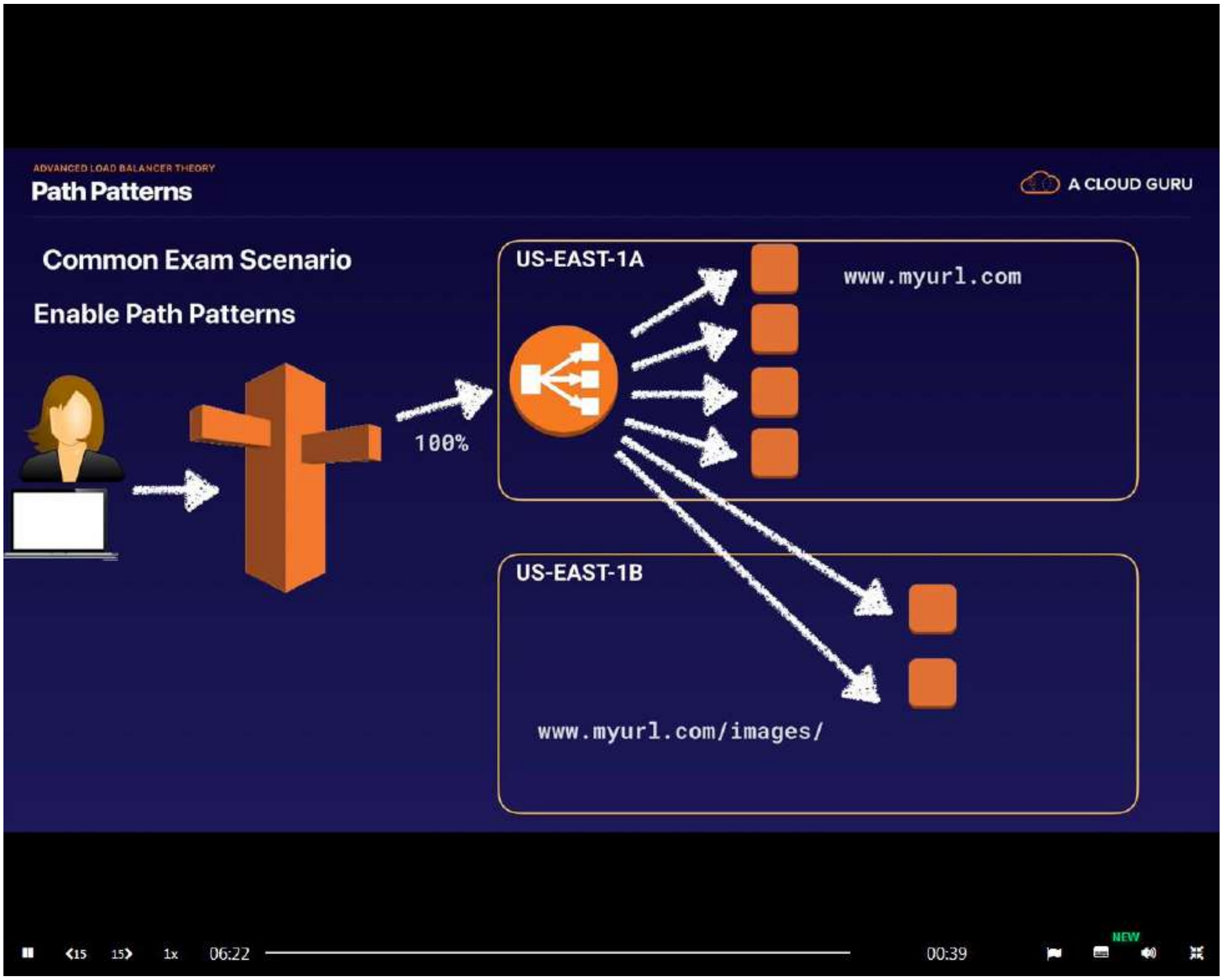

ADVANCED LOAD BALANCER THEORY
What Is Cross Zone Load Balancing
A CLOUD GURU
With Cross Zone Load Balancing
US-EAST-1A
20%
20%
20%
20%
50%
50%
US-EAST-1B
20%
04:14
02:47
NEW
1x

The above diagram is from ACG

Path Patterns

Example: Sending the Normal Path myurl.com to the 4 web servers in US-EAST-A1

And sending the myurl.com/images traffic to your media instances located in a separate availability zone

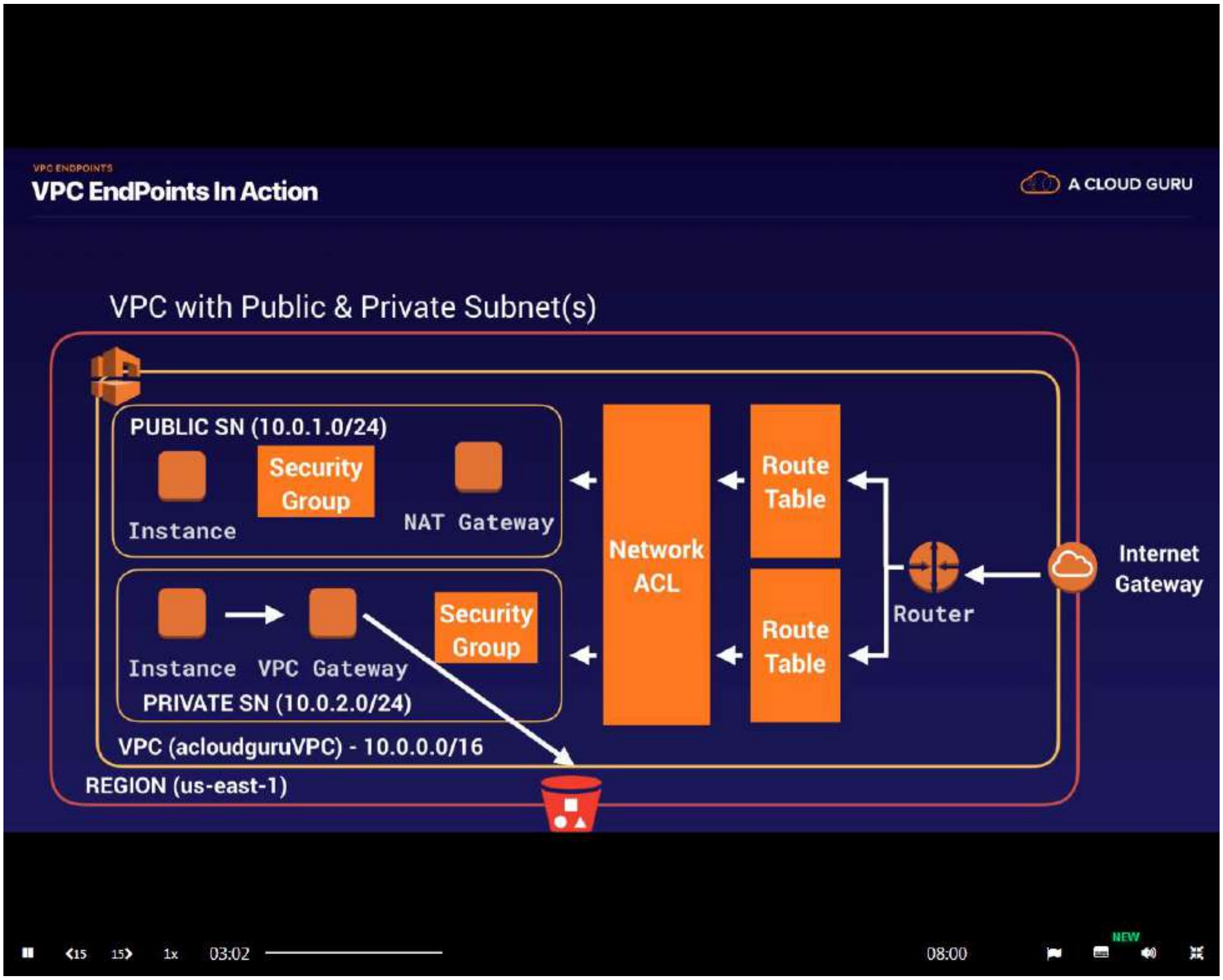


The above diagram is from ACG

VPC Gateway Endpoint Diagram

*The bucket represents S3

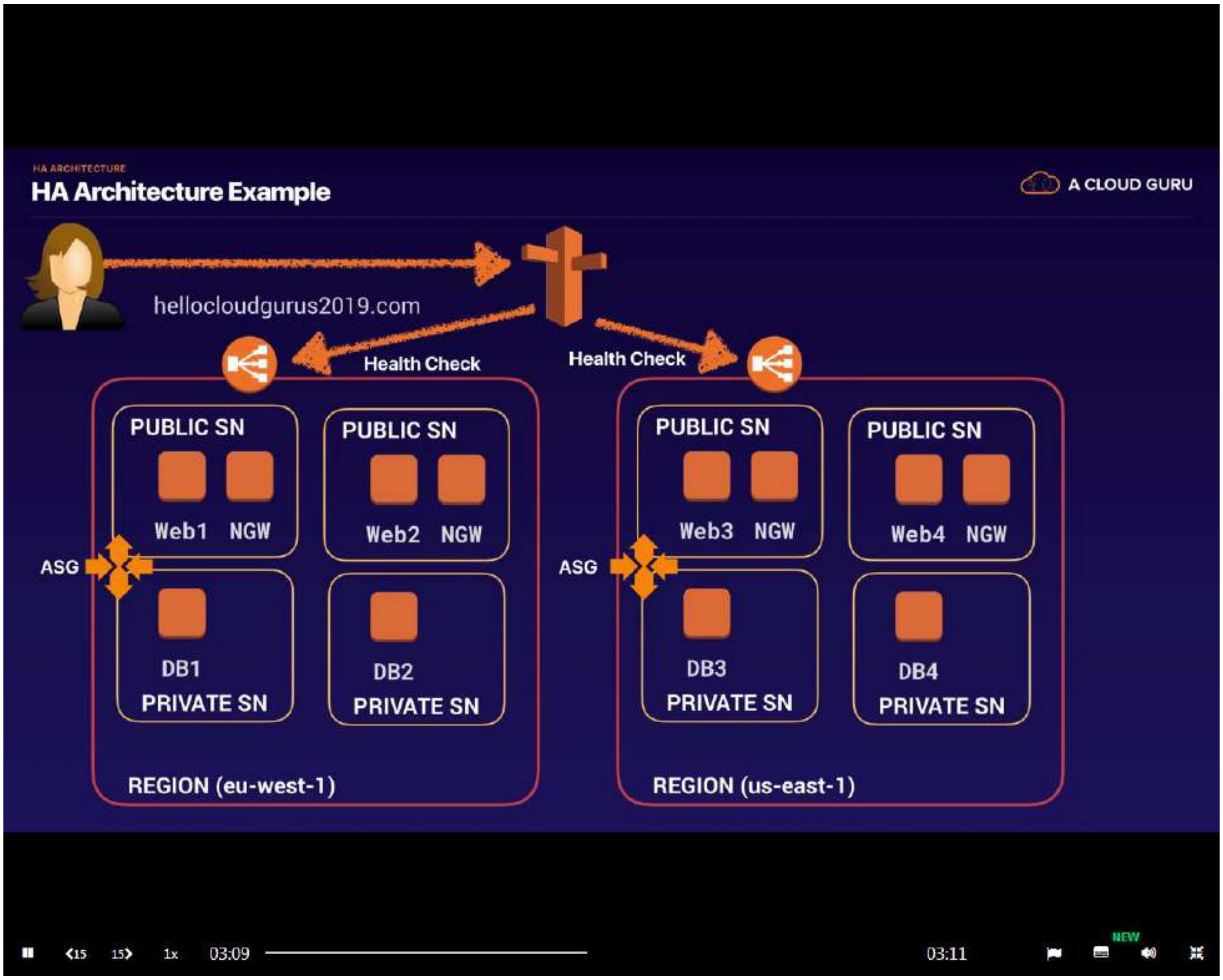
Our instance sends files to the VPC Gateway and that Gateway is going to send the file to our S3 bucket and it will not leave the Amazon network.



The above diagram is from ACG

HA Architecture Example

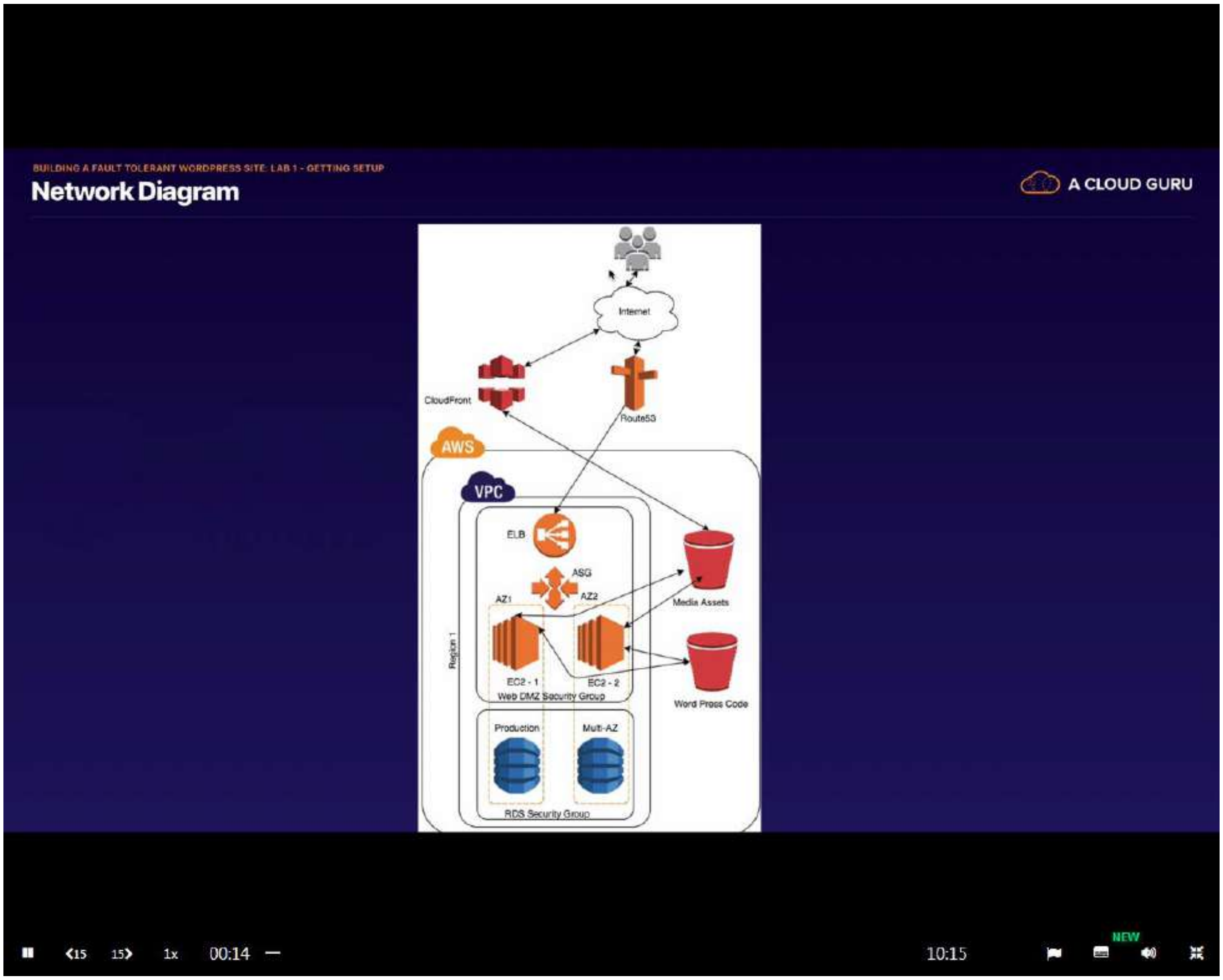
If one of the regions goes down or an AZ goes down then you've got failover. You can failover from one region to another or one AZ to another.



The above diagram is from ACG

Building a fault tolerant WordPress Site

Example: Here we have a user browsing the internet to our Route53 domain name which will connect up to an elastic load balancer. We have some EC2 instances behind an autoscaling group which are going to be in separate AZ's. We have RDS instances that are multi-AZ. We have two S3 buckets, one for our media and one for our code. And we serve our pictures from our wordpress site through cloudfront.



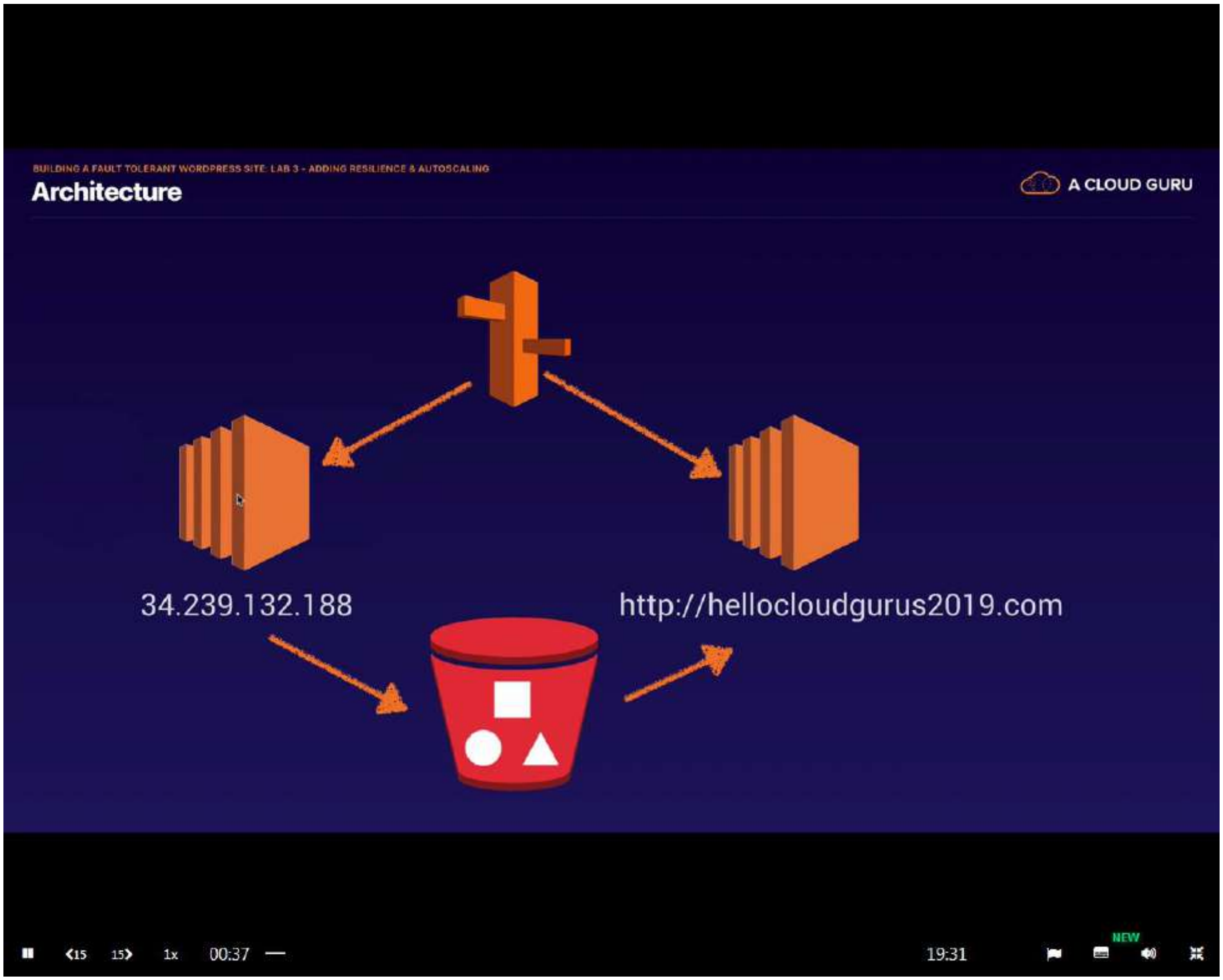
BUILDING A FAULT TOLERANT WORDPRESS SITE: LAB 1 - GETTING SETUP

Network Diagram

The above diagram is from ACG

Fault Tolerant Wordpress Site Network Diagram

On the left is the IP of our writer node. Our blog writers will navigate directly to the IP address on the left side of the diagram. That EC2 instance will push any changes to S3. The Fleet of EC2 instances on the right will be pulling the S3 bucket every minute looking for changes. Our site visitors will visit the domain and Route53 will send them to the fleet of EC2 instances on the right side of the diagram, so it is just sending them to our read nodes only. (We will also have the writer node instances on the right be an autoscaling group situated behind an ELB)



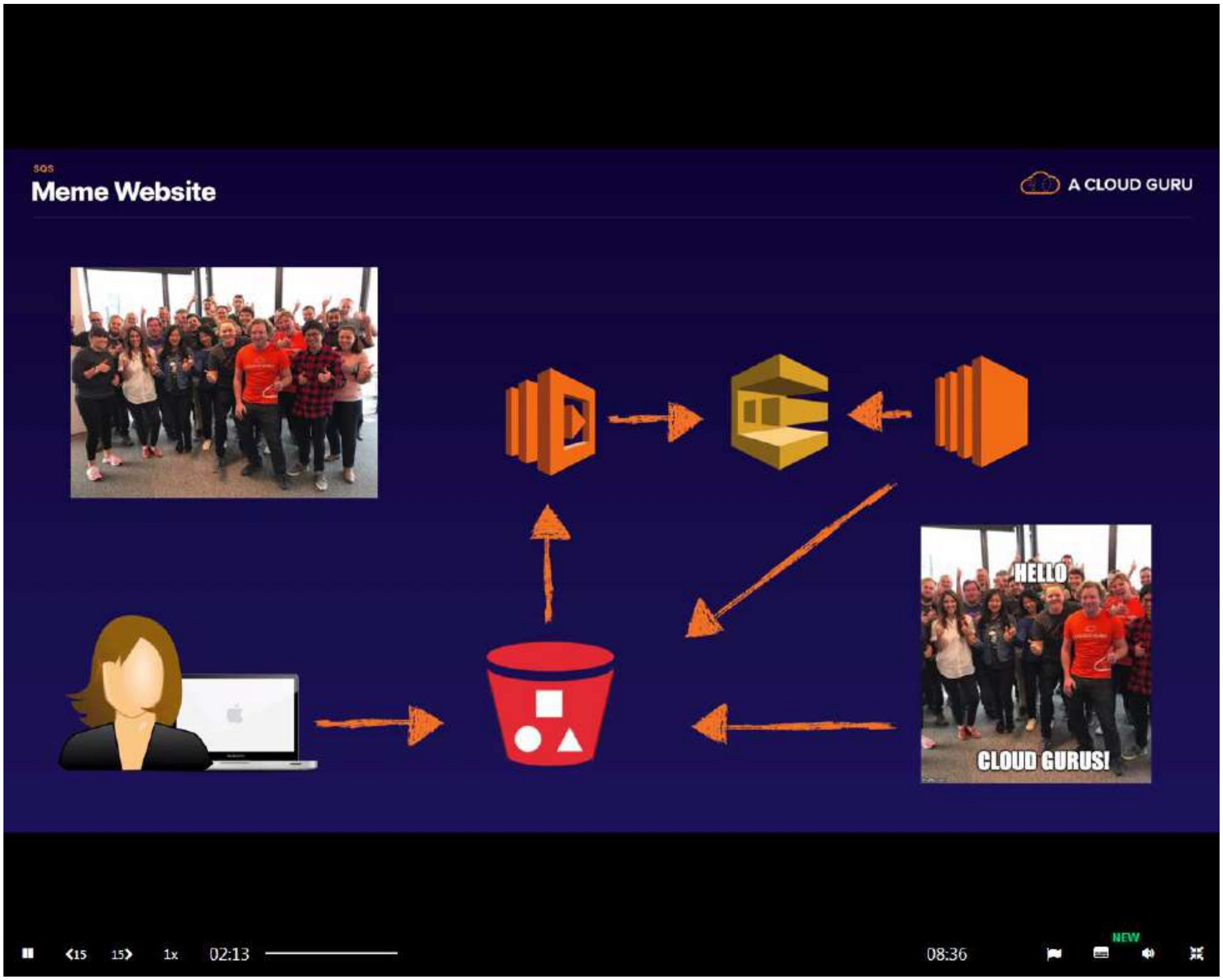
BUILDING A FAULT TOLERANT WORDPRESS SITE: LAB 3 - ADDING RESILIENCE & AUTOSCALING

Architecture

The above diagram is from ACG

SQS

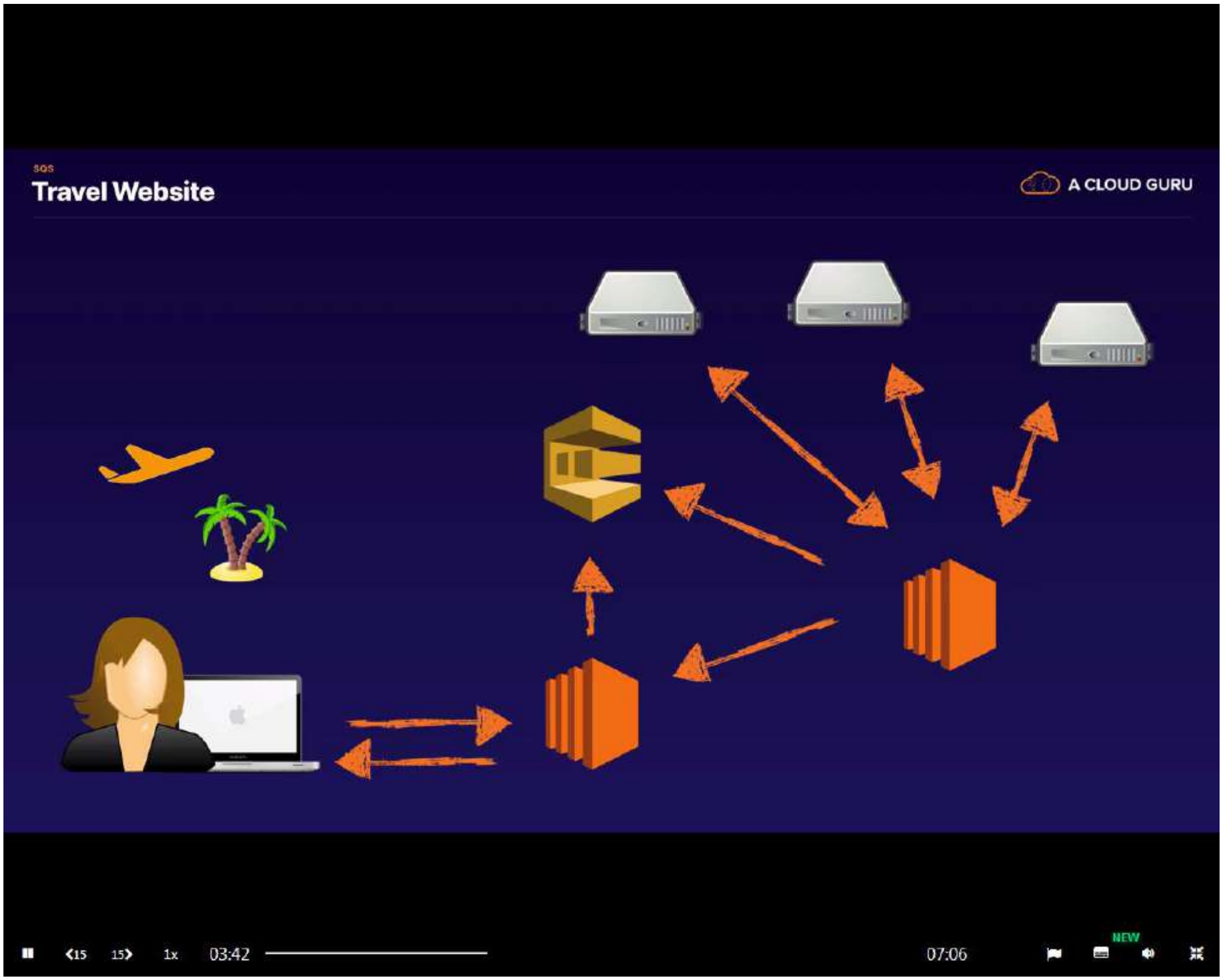
User uploads a photo to s3 which triggers a lambda function which will take the image and write text over it. It will store that text in SQS. Then a fleet of EC2 instances will pull the message queue for work and it will then create the meme and store it in S3. If an EC2 instance fails while creating the meme the message will become available again in the queue and another instance will create the meme. So SQS is storing the message independently so that even if an EC2 instance can’t process the message, another EC2 instance will come along and take the message.



The above diagram is from ACG

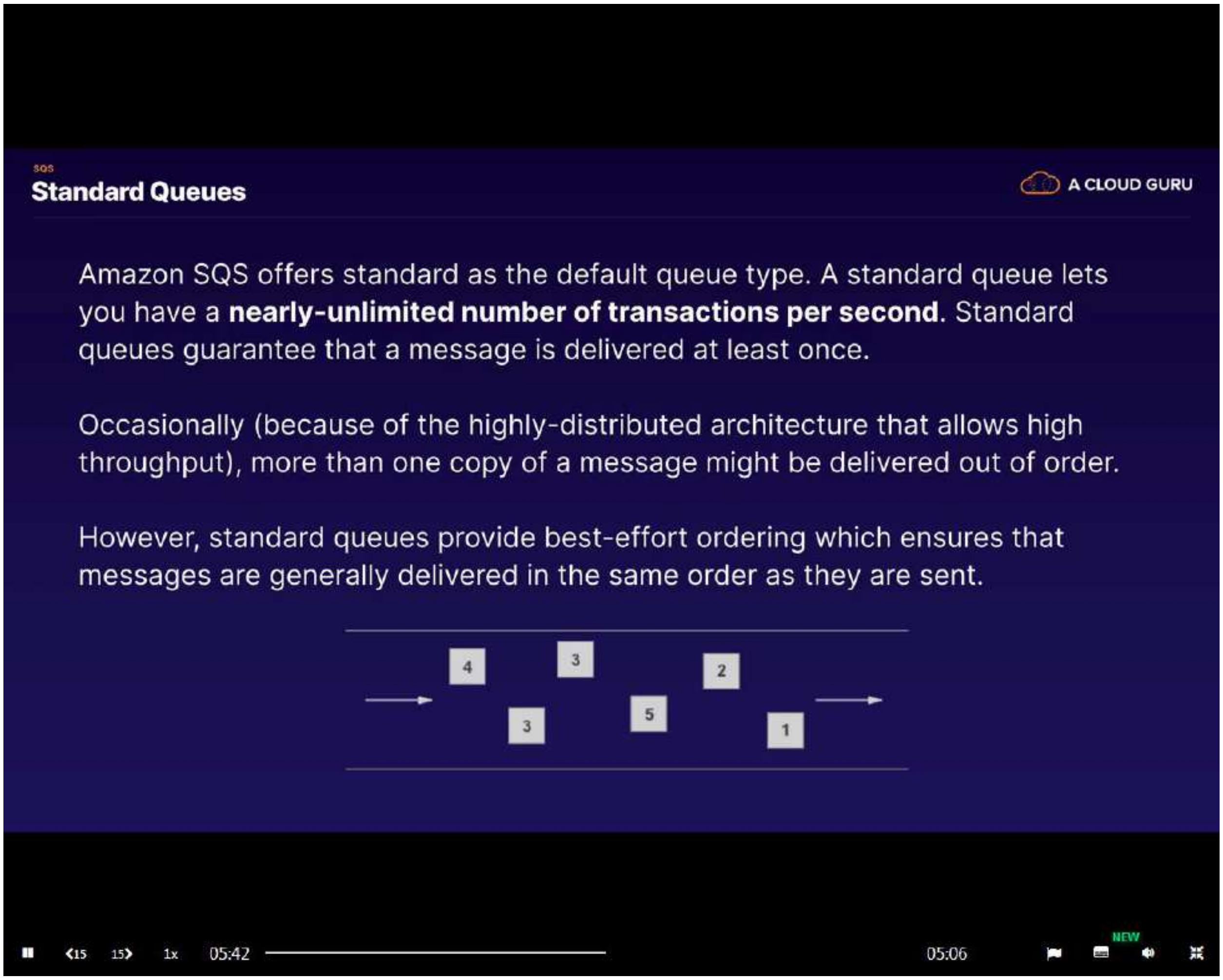
SQS (Continued)

User goes to ec2 server and says they want to go to Rome on these particular dates. The web server passes that information to an SQS queue. A fleet of application servers are configured to pull that message and look for different airlines. Once the information is retrieved it is passed back to the web server and then back to the end user. If we lose an individual ec2 instance we won't lose the information and another one will come along and poll the queue to do the work and ultimately return the result back to the end user.



The above diagram is from ACG

SQS Standard Queues



SQS

Standard Queues

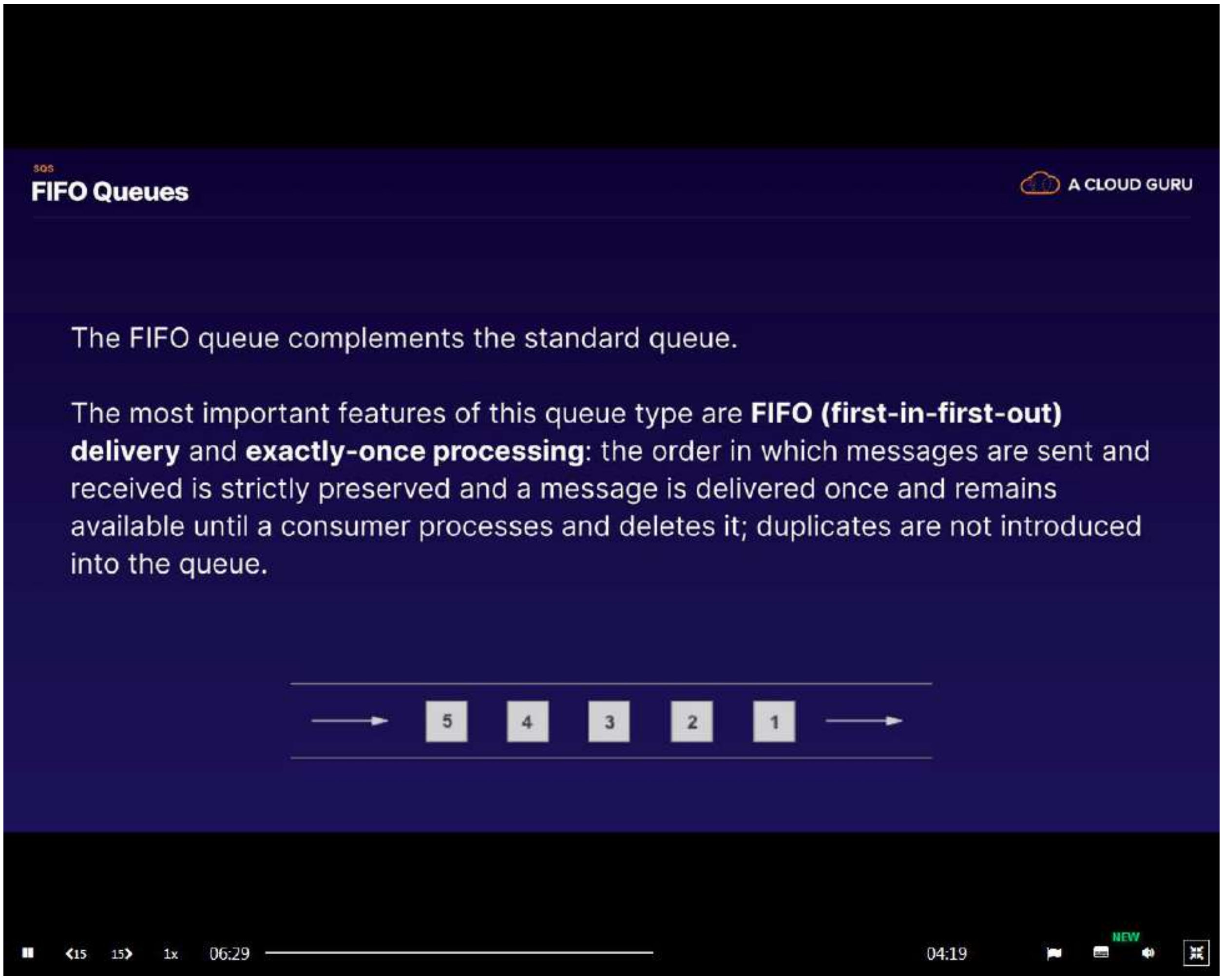
Amazon SQS offers standard as the default queue type. A standard queue lets you have a **nearly-unlimited number of transactions per second**. Standard queues guarantee that a message is delivered at least once.

Occasionally (because of the highly-distributed architecture that allows high throughput), more than one copy of a message might be delivered out of order.

However, standard queues provide best-effort ordering which ensures that messages are generally delivered in the same order as they are sent.

The above diagram is from ACG

FIFO Queues



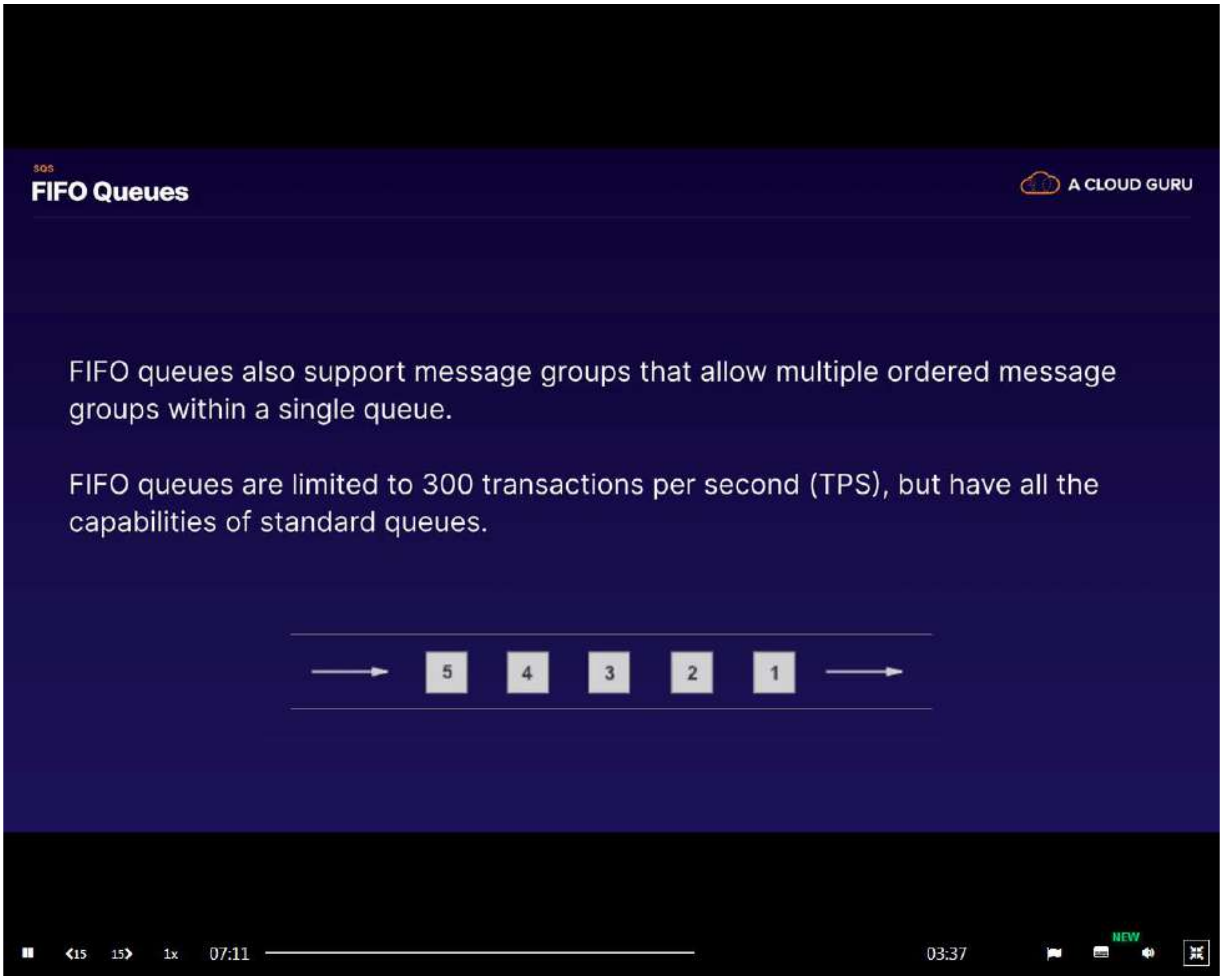
FIFO Queues

The FIFO queue complements the standard queue.

The most important features of this queue type are **FIFO (first-in-first-out) delivery** and **exactly-once processing**: the order in which messages are sent and received is strictly preserved and a message is delivered once and remains available until a consumer processes and deletes it; duplicates are not introduced into the queue.

The above diagram is from ACG

FIFO Queues



SQS

FIFO Queues

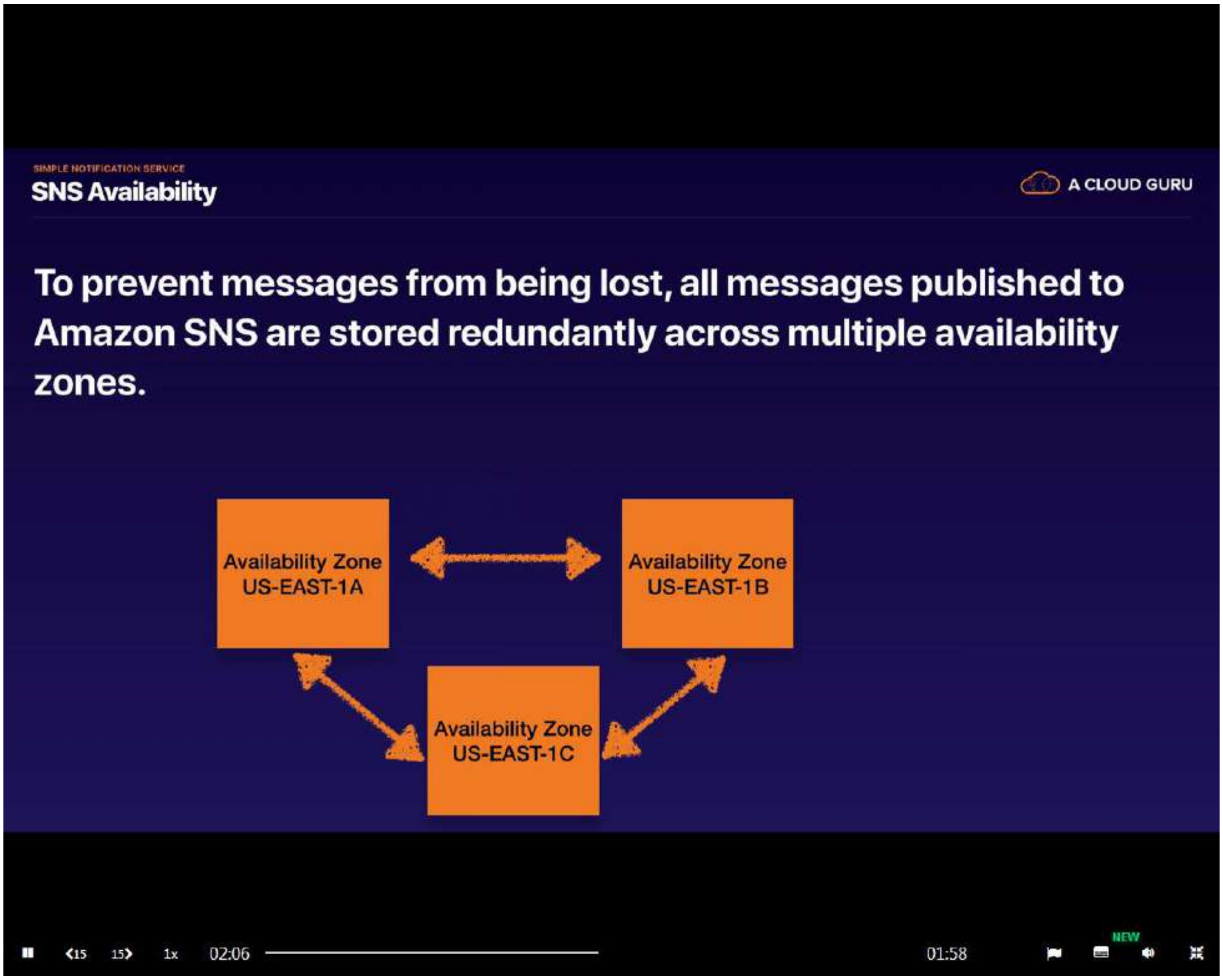
A CLOUD GURU

FIFO queues also support message groups that allow multiple ordered message groups within a single queue.

FIFO queues are limited to 300 transactions per second (TPS), but have all the capabilities of standard queues.

The above diagram is from ACG

SNS Availability



SIMPLE NOTIFICATION SERVICE

SNS Availability

To prevent messages from being lost, all messages published to Amazon SNS are stored redundantly across multiple availability zones.

Availability Zone US-EAST-1A

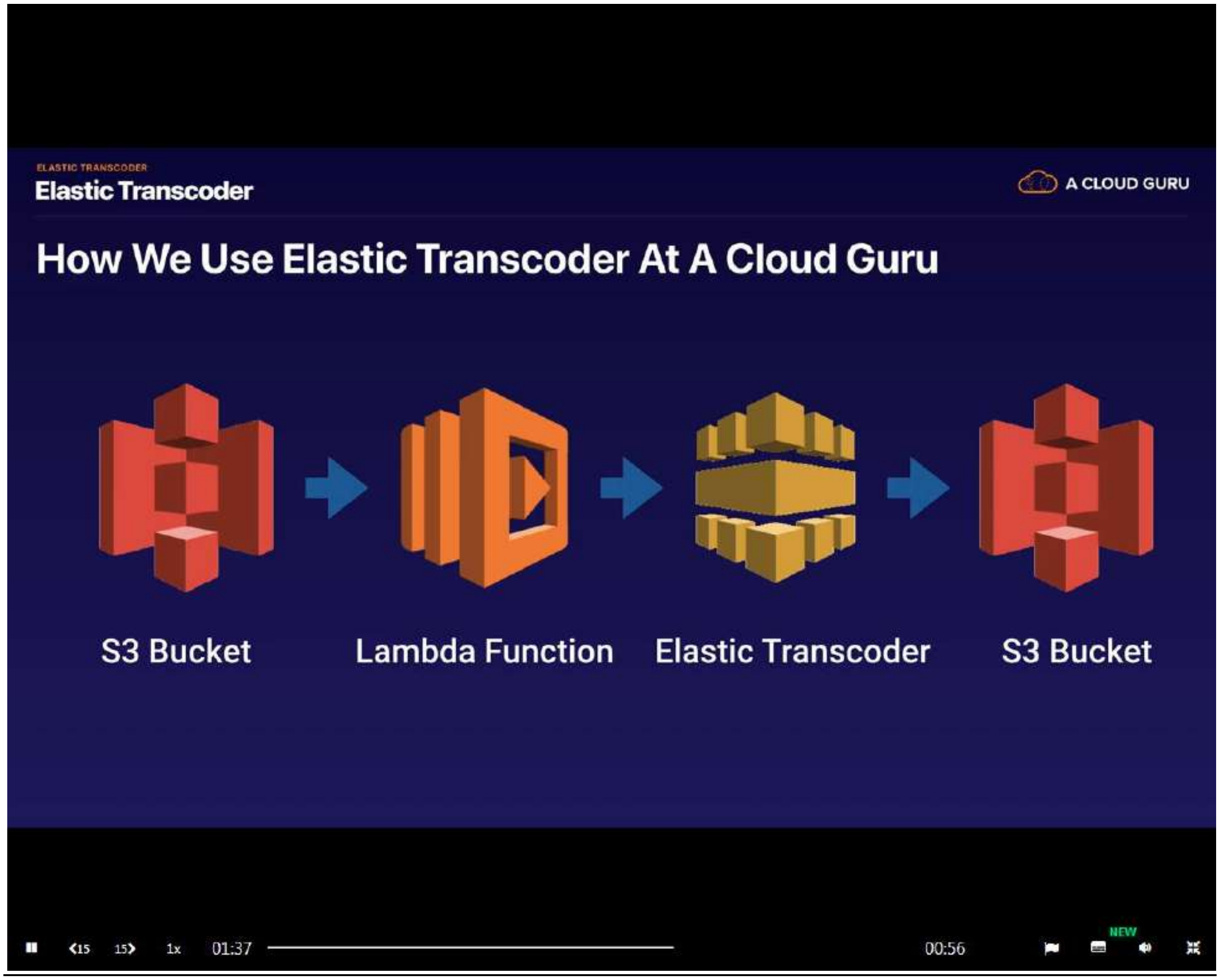
Availability Zone US-EAST-1B

Availability Zone US-EAST-1C

The above diagram is from ACG

Elastic Transcoder

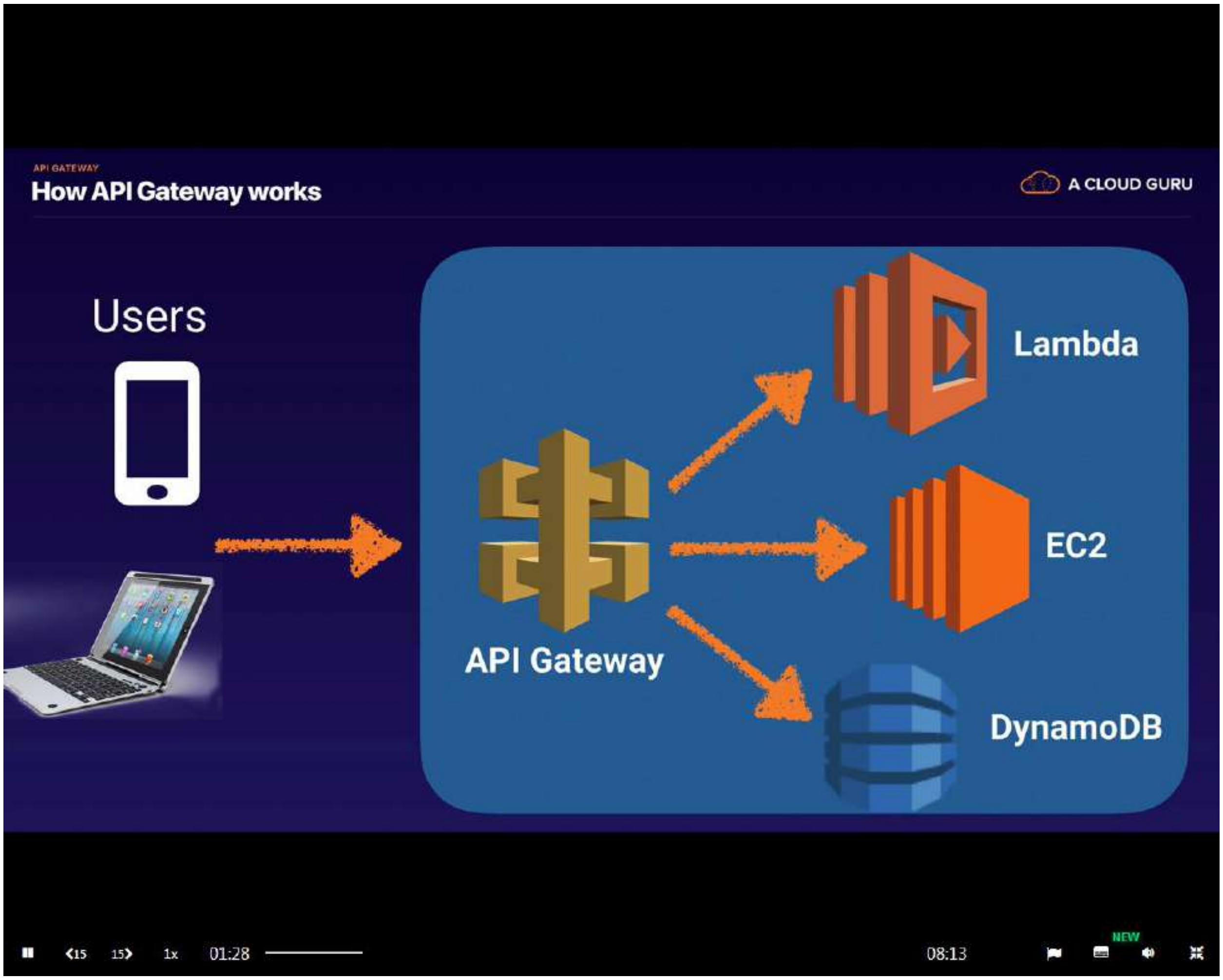
Upload a video into S3 Bucket…then a lambda function will take the metadata of the video and send it to Elastic Transcoder which will transcode the video so that it looks good on various devices and is in high resolution and then stores the transcoded video in another S3 Bucket.



The above diagram is from ACG

API Gateway

Users do a call to our API Gateway. API Gateway is the ‘front-end/front door’ to our AWS environment. That API call could be passed to Lambda, or an EC2 instance, or writing to DynamoDB, etc.



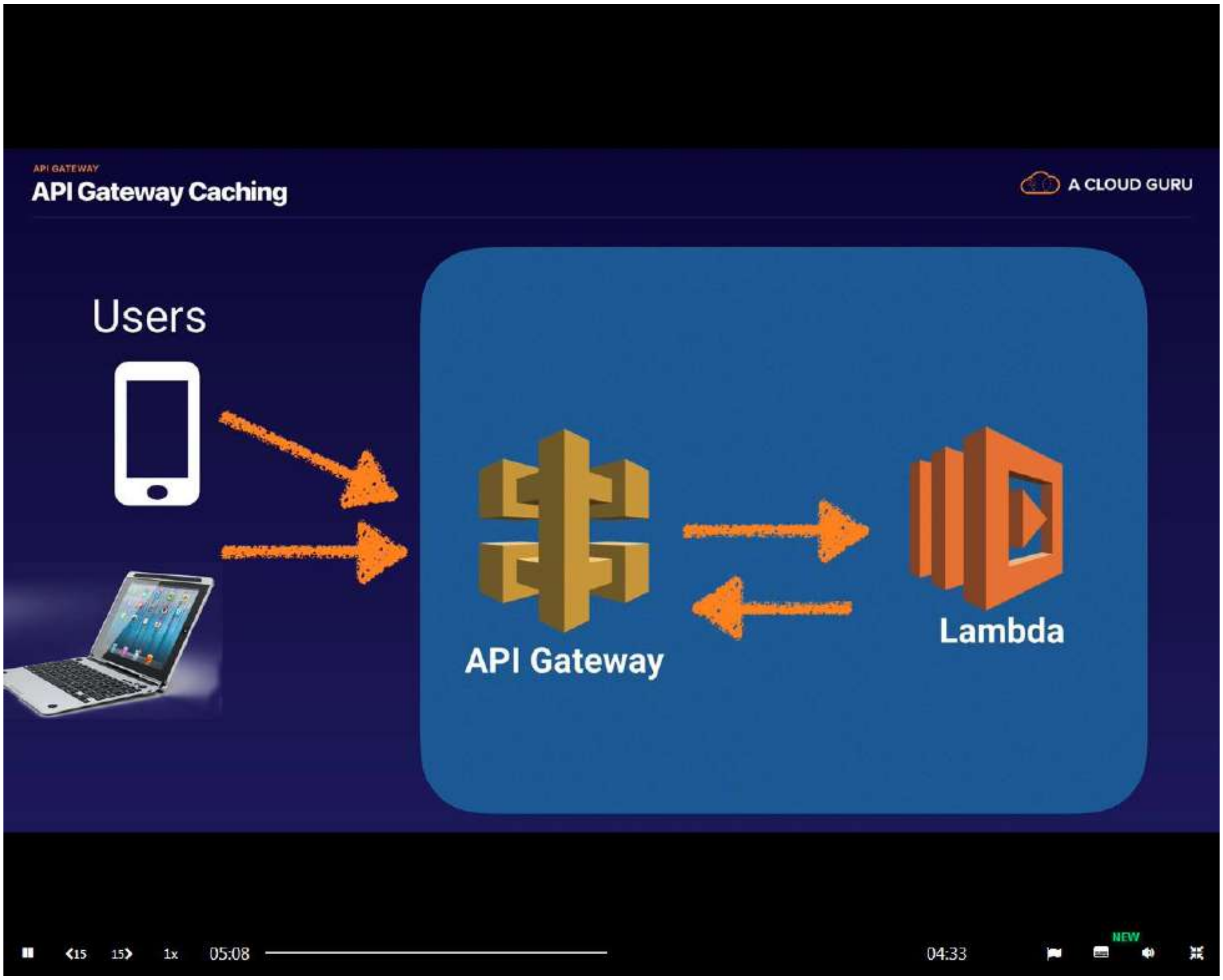
API GATEWAY

How API Gateway works

The above diagram is from ACG

API Gateway Caching

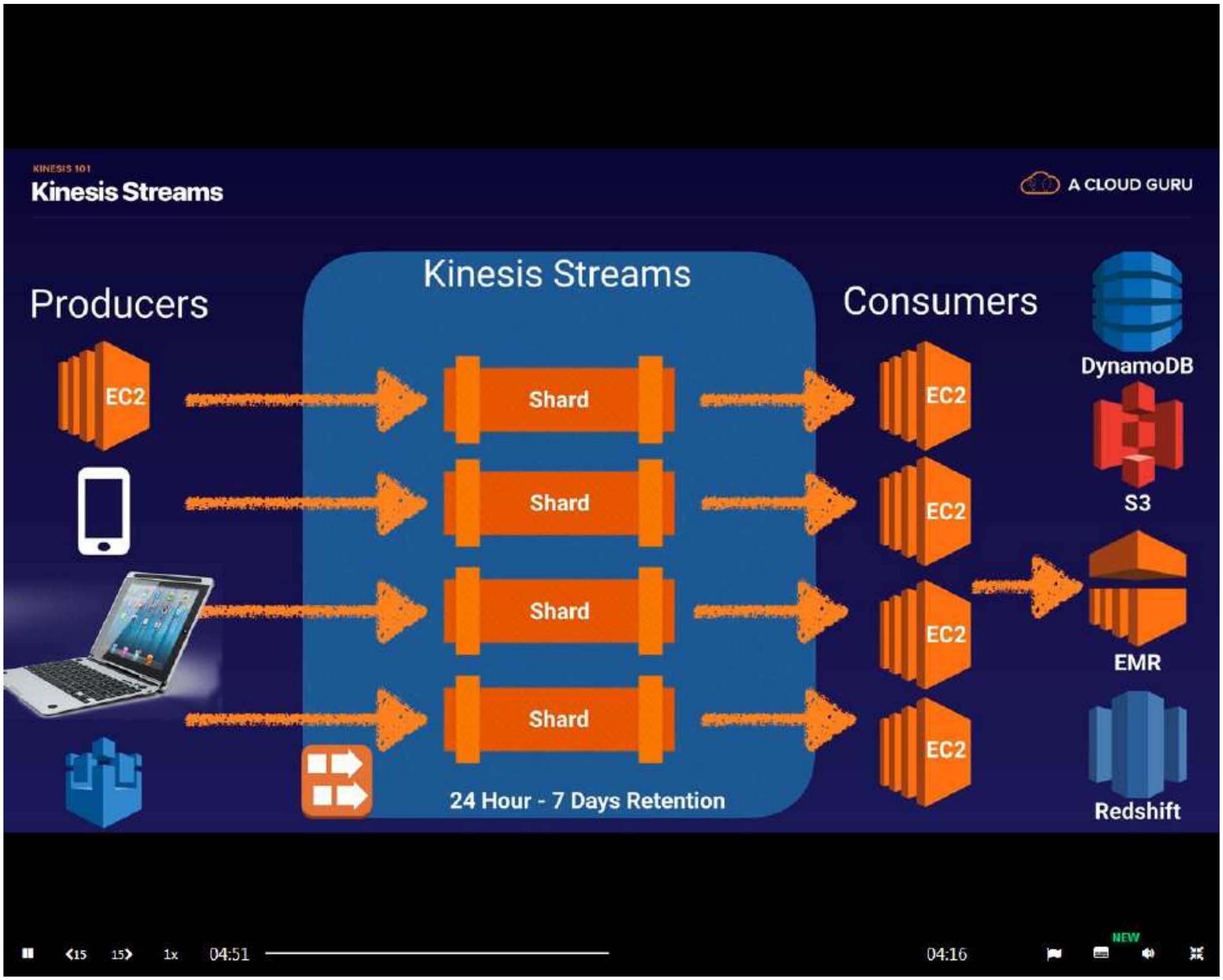
For example, User 1 makes a get request that is forwarded on to lambda and lambda returns a response. Then User 2 makes the same get request and since API Gateway has cached it, it does not go to the Lambda function.



The above diagram is from ACG

Kinesis Streams

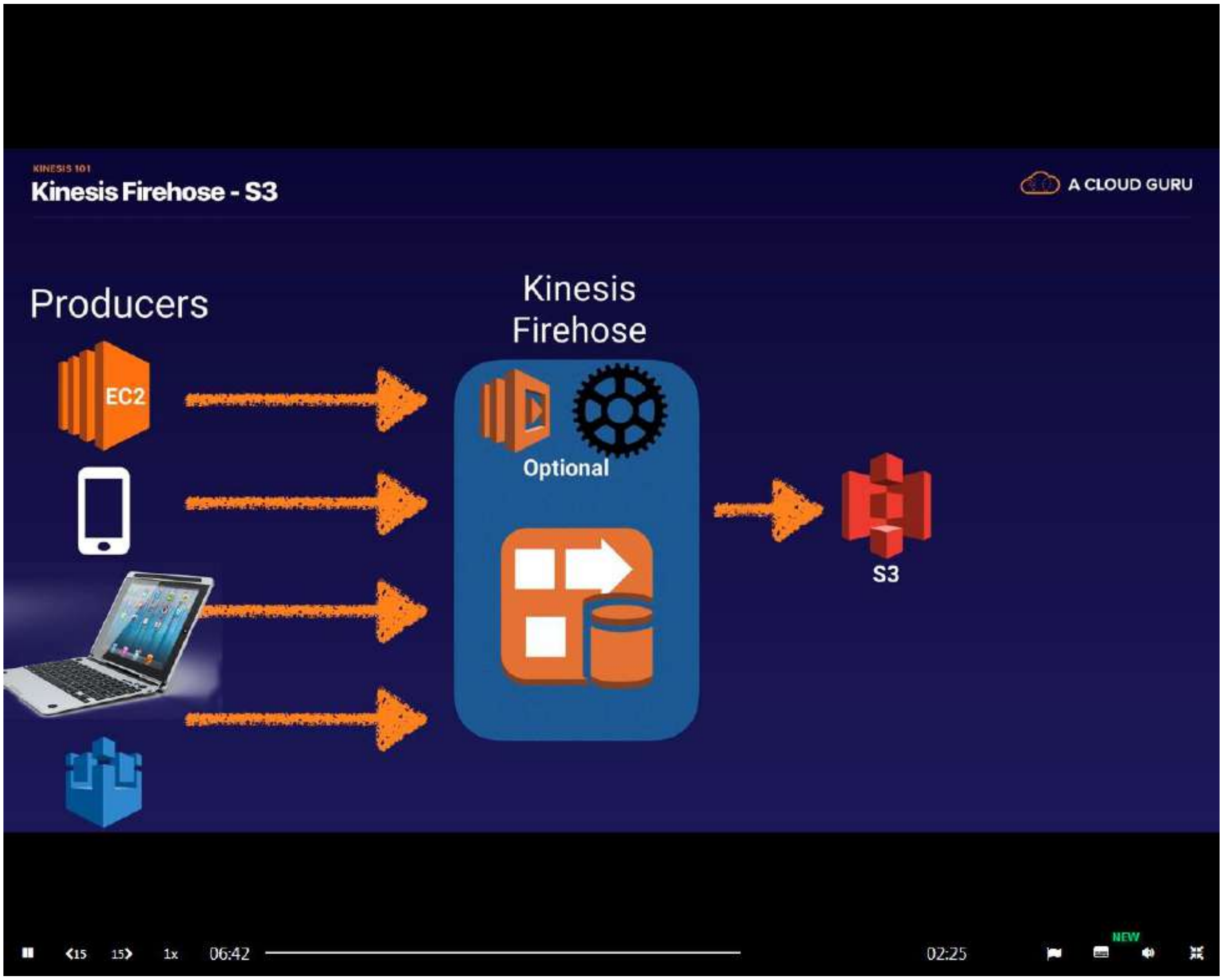
Data Producers (such as an EC2 monitoring stock prices, or perhaps IoT monitoring farm data) stream the data to kinesis and Kinesis Streams is a place to store that data coming in (by default stores for 24 hrs but can be stored for 7 days). The data is contained in 'shards'. EC2 instances (aka the data consumers) can then analyze that data located within the shards and then can store it in various places such as DynamoDB, S3, EMR, Redshift, RDS etc. Kinesis allows you to persistently store your data for 24 hrs to 7 days, while your data consumers do something with that data.



The above diagram is from ACG

Kinesis Firehose

Data Producers send the data to Kinesis Firehose (of which does NOT have persistent storage and does NOT have ‘shards’). The data has to be analyzed as it comes in. So perhaps you have lambda function within your Kinesis Firehose running a set of code for the data as it comes in and outputs it somewhere. It can output it to S3. Or it could output it to S3 and then you can import it to Redshift. Or it can output to Elasticsearch Cluster, etc.



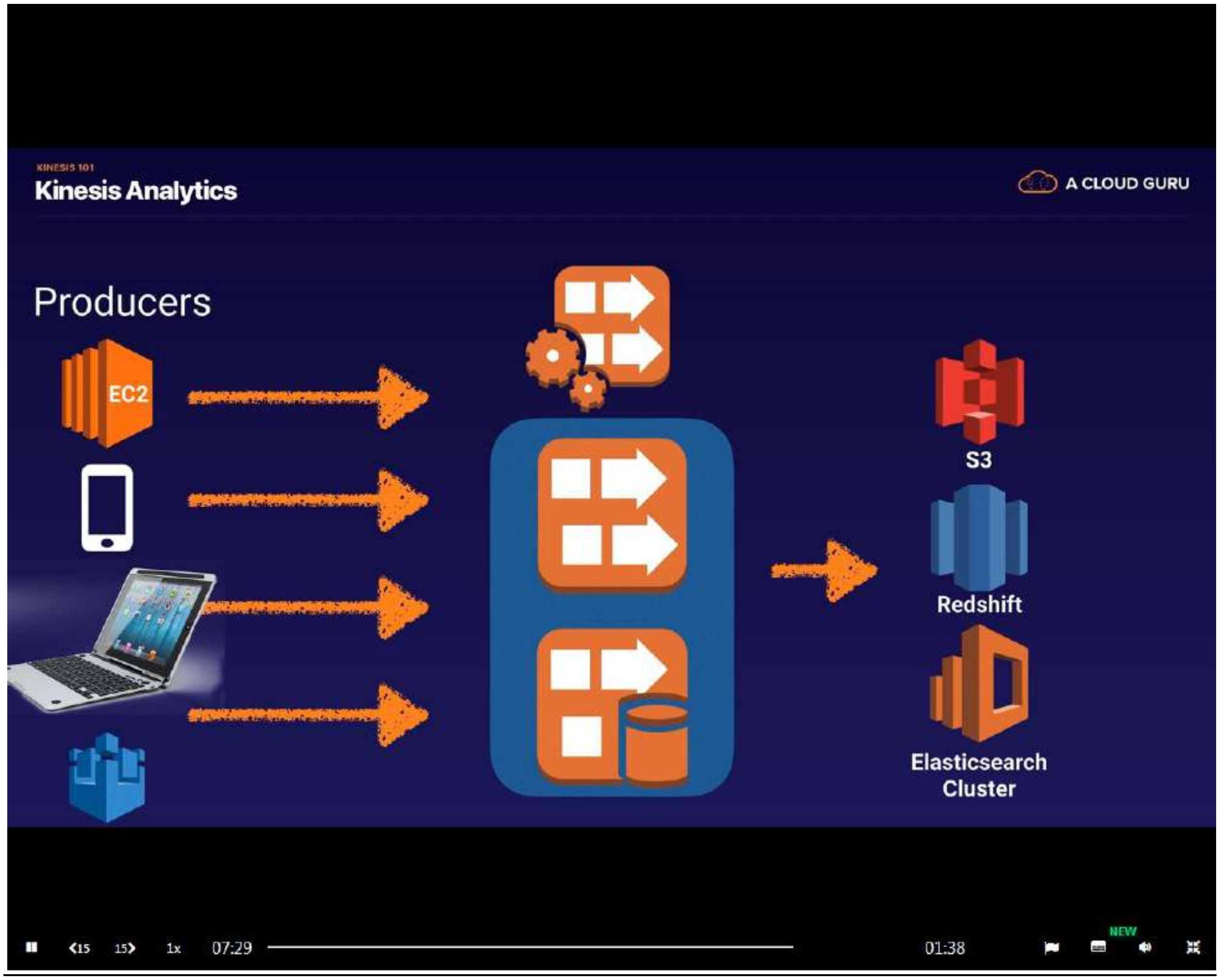
KINESIS 101

Kinesis Firehose - S3

The above diagram is from ACG

Kinesis Analytics

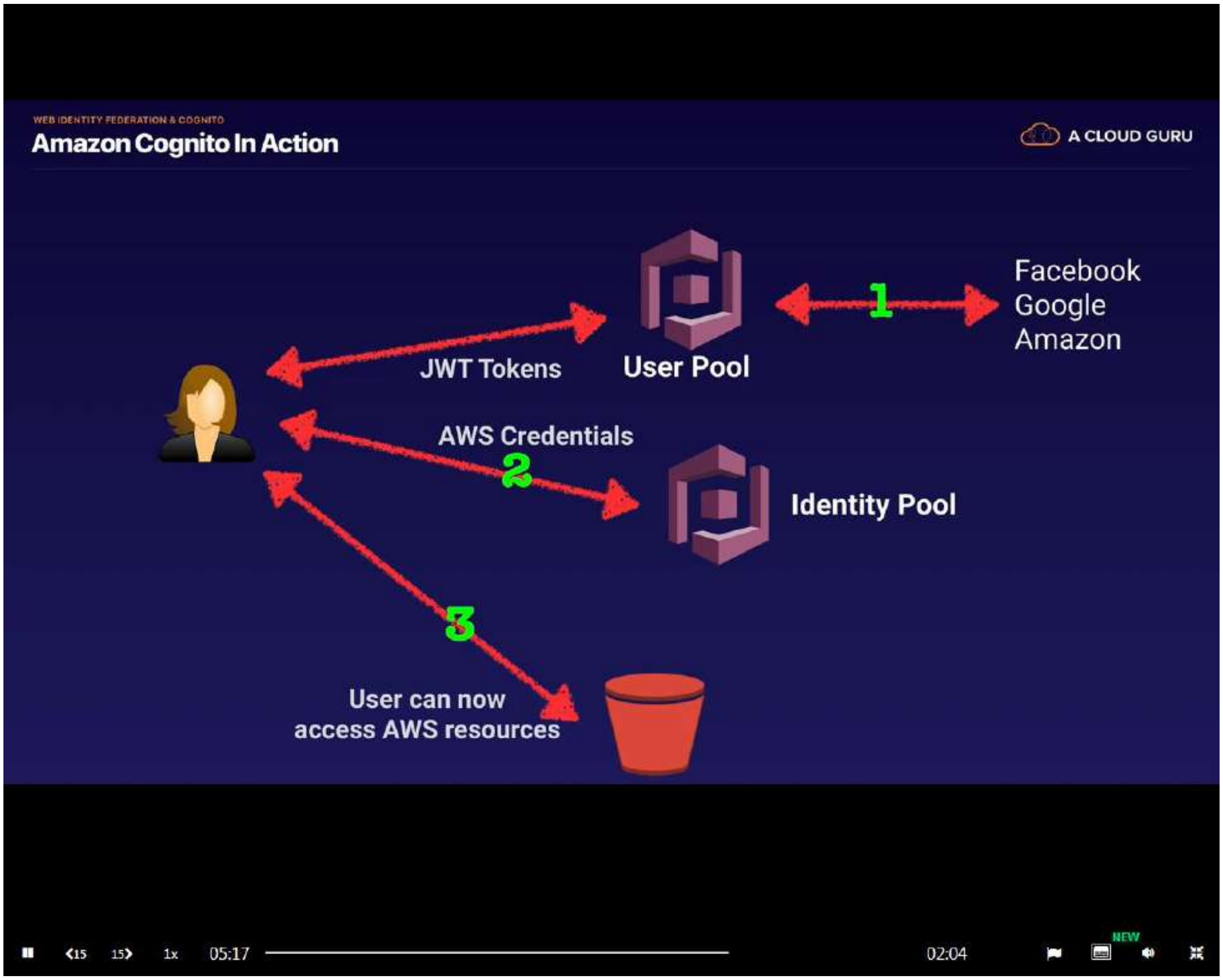
Works with Kinesis Streams, as well as with Kinesis Firehose. It can analyze the data on the fly inside either service and then it stores this data either on S3, Redshift, or Elasticsearch Cluster.



The above diagram is from ACG

Amazon Cognito

User logs in with Facebook account. Facebook authenticates her account and passes back an authentication token to Cognito, which converts it to a JWT token. The user sends the token to an identity pool and that identity pool will grant her AWS credentials in the form of an IAM role and then she will have access to AWS resources. User pools pertains to registration, authentication, usernames, and passwords. Identity pools pertains to **the actual granting of access** to use AWS resources.



WEB IDENTITY FEDERATION & COGNITO

Amazon Cognito In Action

The above diagram is from ACG

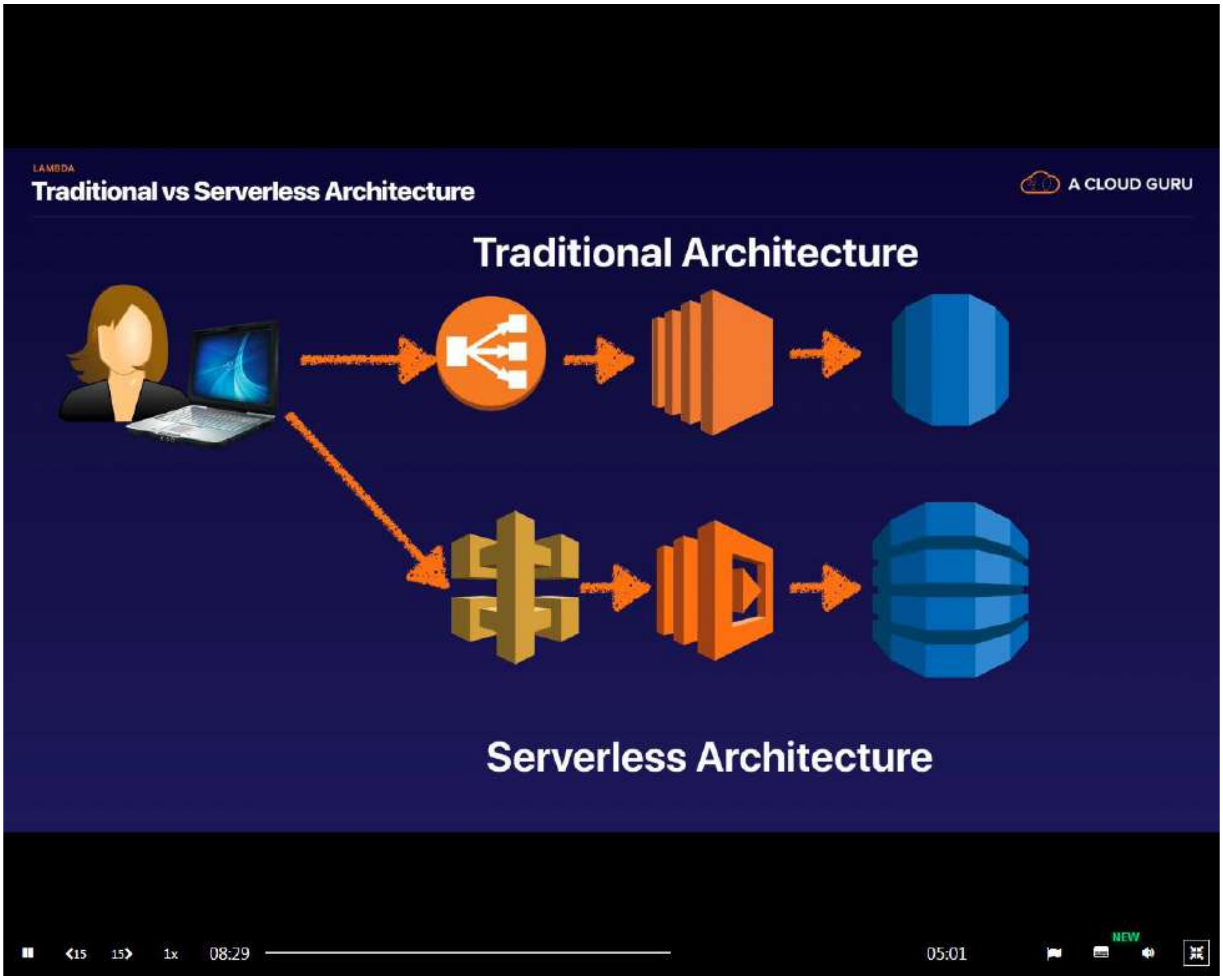
Traditional vs Serverless Architecture

Traditional

ELB ->EC2->Database Storage (RDS)

Serverless

API Gateway ->Lambda->DynamoDB

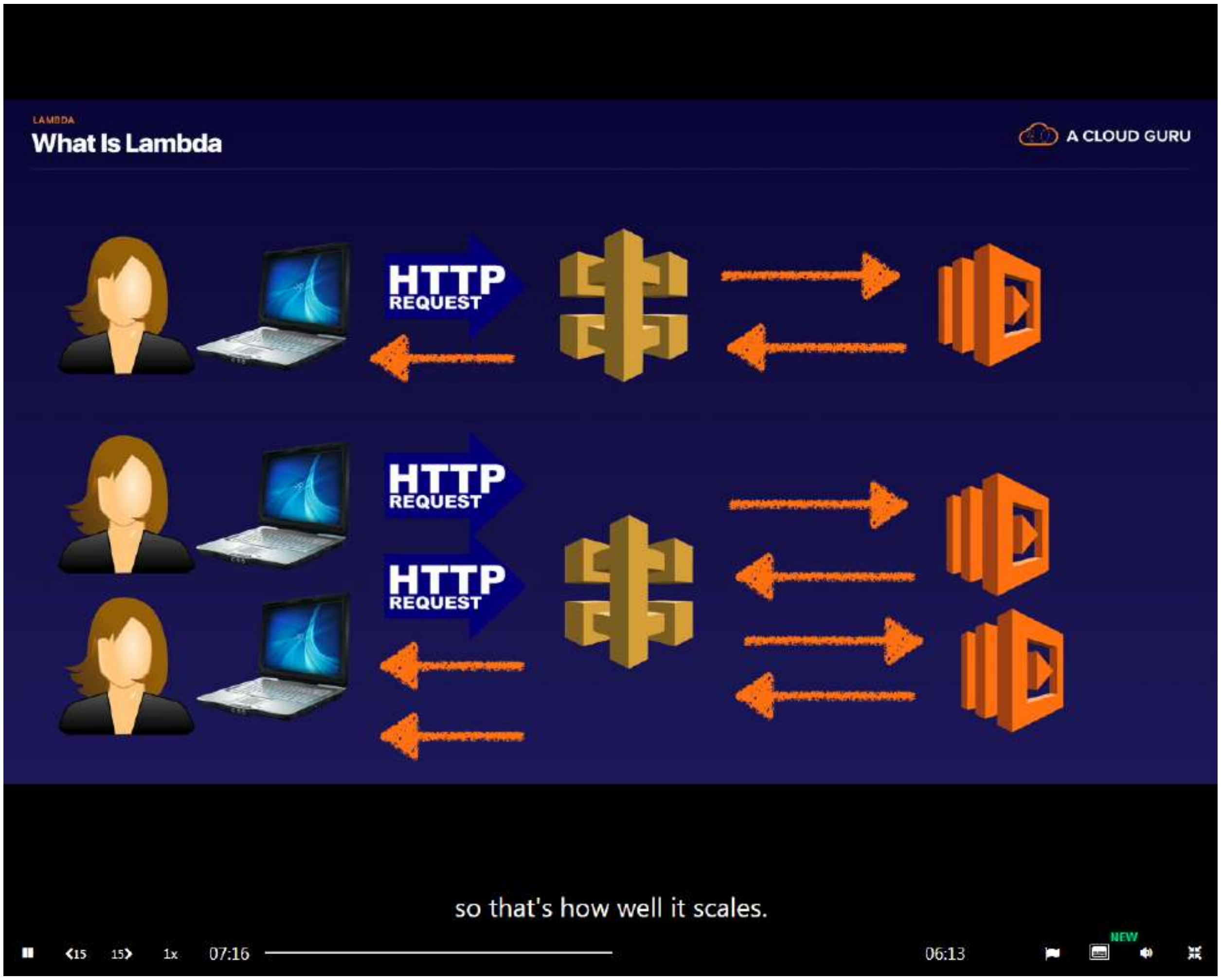


The above diagram is from ACG

Lambda

User sends http request to API Gateway. API Gateway proxies that to Lambda. Lambda will run the code in response to that HTTP request, and then send it back to API Gateway, that will send it back to the user.

If two users have two separate HTTP requests, the same lambda function will be triggered but it will be separately run, so it will be two separate lambda functions. If you have a million users hitting your API gateway at once, it will trigger a million lambda functions, so that's how well it scales.



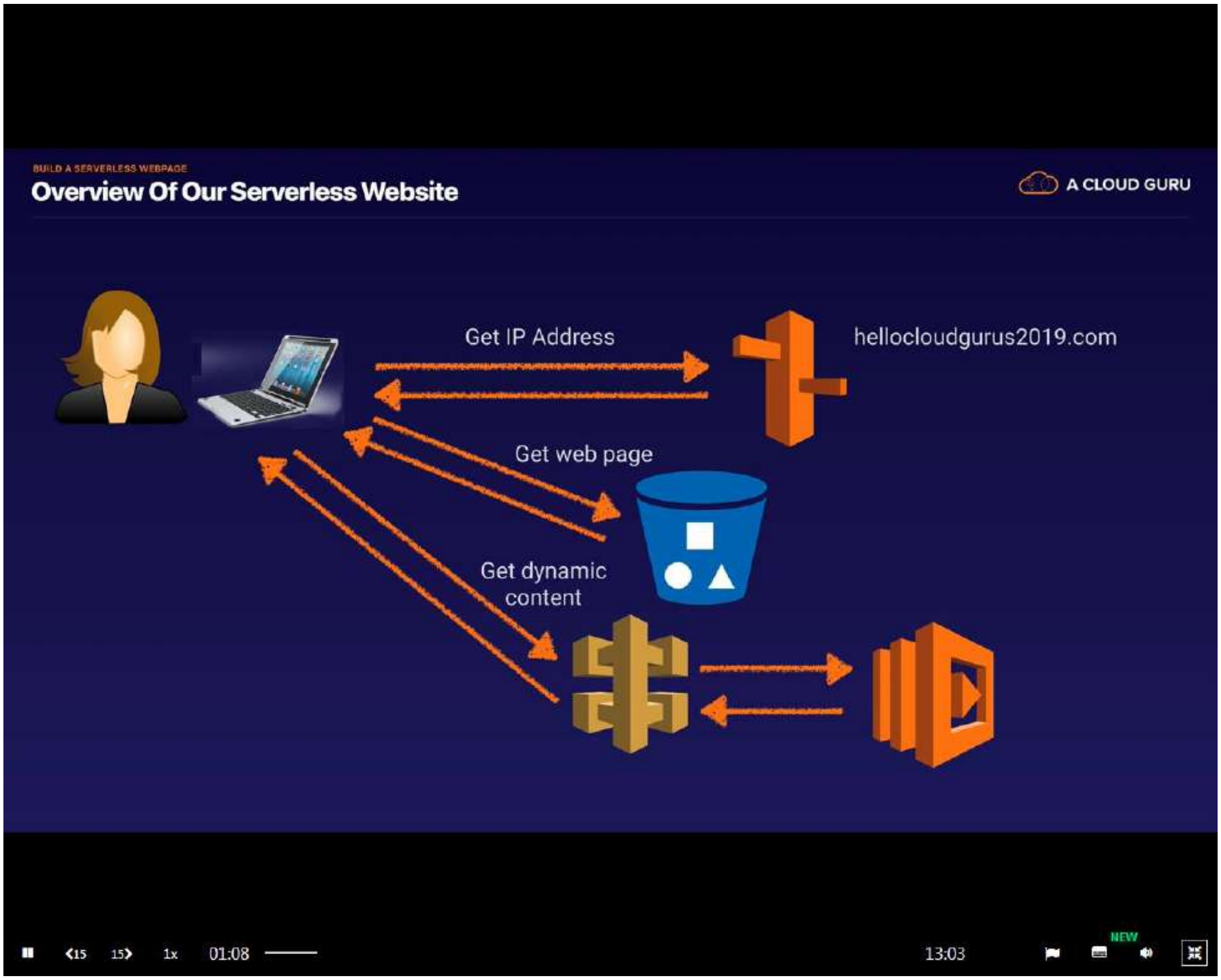
LAMBDA

What Is Lambda

so that's how well it scales.

Serverless Website with API Gateway & Lambda

User wants to go to helloucloudgurus2019.com. User is sending a query across to Route53 which will respond with the bucket address for our website. So our user goes to our S3 bucket and they are going to go to our index.html, which will show up as a static page. But it will have a button. When the button is pushed they are going to get dynamic content. Because a request is send through to API gateway which will proxy a request to a lambda function which will take the data and return a result to API gateway, which will then return a result, to our user.



BUILD A SERVERLESS WEBPAGE

Overview Of Our Serverless Website