LEARNING AWS

A CLOUD GURU TRAINING SESSIONS

Miscellaneous Notes

- AWS US East (N. Virginia)
  - This region is the oldest and it is where most of the new services are deployed first for consumption by the public
  - is a region that tends to go down at least once a year.
- AWS Support Levels
  - Basic plan
  - Developer plan
  - Business plan

AWS Global Infrastructure (2019)

- Region
  - A geographical area that may contain 2 or more availability zones
  - 24 regions globally
- Availability zone
  - An AZ is a facility similar to a data center
  - Have computers, SAN, redundant power, redundant backbone to internet.
  - Facilities which are close together are counted as being in the same availability zone
  - 72 availability zones globally
- Edge Location
  - These are endpoints for AWS which are used for caching content. Examples are
    - AWS **CloudFront** which is Amazon's CDN
  - There are over150 Edge Locations globally

Core AWS Services for Architect Cert

- AWS global infrastructure
  - Region > Availability Zone > Edge Logcation
- Compute
  - EC2, Lambda
- Storage
  - EFS, S3
- Databases
  - RDS, Dynamo
- Network & Content Delivery
  - Knowing this area inside-out is key to passing the exam
  - Know VPC & Route 53
  - **Cloudfront** (amazon's CDN)
- Security, Identity & Compliance
  - Knowing this area inside-out is key to passing the exam
  - Know IAM very well

IAM 101 WITH AWS

- Centralized control of your AWS account
- Shared access to your AWS account
- Granular permissions
- Identity federation (AD, FB, LinkedIn, etc.)
- MFA
- Provide temporary access for users/devices and services where necessary
- Allows setup of password rotation policies
- Integrates with many different AWS services
- Supports PCI DSS compliance
- IAM is the most important service in AWS because it controls access to all other AWS services.
- Power User Access – allows access to all AWS services except the management of Groups and Users within IAM
- Read this white paper: https://aws.amazon.com/whitepapers/

AWS IAM – Key Terminology

- Users
  - Represents end users (i.e. people in an org)
  - Accounts for individual users
  - Can have permissions assigned to them
- Groups
  - Can have permissions assigned to them
  - Users can be members of a group
  - Users who are members of a group will inherit the permissions of said group
- Roles
  - Are assigned to AWS resources
- Policies
  - JSON documents that describe permissions that a User/Group/Role is able to do

IAM – KEY CONCEPTS

- **IMPORTANT**:
  - ***Activate MFA on the root account.*** Keep your root account QR code in a secure location to be able to reactivate in case you lose your phone.
  - ***Operations on IAM are global.*** We cannot create a user for a specific region or availability zone.
- You can create users of 2 types
  - Programmatic Access – to access AWS services programmatically (for Devs)
  - Console Access – to access AWS web-based console to manage your AWS account settings
- **Users** – Represent the users of AWS services. They can be granted **actions** to **resources** (not recommended)
- **Groups** – contain users and are governed by **Policies**
- **Policy** – determines what **ACTIONS** can be taken on what **RESOURCES**
- **Roles** – Determines a secure way to grant permissions to entities we trust

IAM – Create a billing alarm

- A billing alarm helps keep watch over an AWS account.
- **CloudWatch** – it is the AWS service that can be used to create billing alarms
- **SNS Topic** – it is an AWS service called "Simple Notification Service" (SNS) that allows services to send messages to other services.

AWS SERVICE – S3

- S3 – Simple Storage Service
- S3 allows you to create folders (AWS calls them **Buckets**) where you can store objects (files)
- Files can have 0 bytes to 5 Terabytes
- By default, an account can have 100 buckets
- **Buckets** have a globally unique web address
- Uploading Large Objects: https://docs.aws.amazon.com/AmazonS3/latest/dev/UploadingObjects.html
- When uploading files to a bucket location you, if it was successful, you will get an **HTTP 200** response code.
- Each object (file) has certain attributes
  - Key – the name of the file. It must be unique within a **bucket**
  - Value – the content of the file (bytes)
  - Version – the version of the file. Important when you are uploading objects with the same key to the bucket location.
  - Metadata – additional data about the S3 object that is stored in a bucket
  - Subresources
    - Access Control Lists
    - Torrents

AWS SERVICE – S3

- Data Consistency
  - Read after Write consistency for 1ST PUT of a new objects
  - Eventual Consistency for overwrite PUTs and DELETEs for existing object (can take some time to propagate)
- S3 guarantees
  - Built for 99.99% availability – 52.6 seconds in a year of downtime
  - Amazon guarantees 99.9% - 8.76 hours in a year of downtime
  - Amazon guarantees 99.999999999% of Durability –
- Features
  - Tiered storage
  - Lifecycle management
  - Versioning
  - Encryption at rest
  - MFA Delete
  - Security with ACL and Bucket Policies

AWS SERVICES – S3

- Storage Classes (Tiers)
  - **S3 standard**
    - 99.99% availability
    - 99.999999999% durability (11 x 9's)
    - Stored redundantly across multiple devices in multiple facilities
    - Designed to sustain the loss of 2 facilities concurrently
  - **S3 Infrequent Access** (IA) – Cheaper than S3 Standard but data retrieval fees apply
    - Availabilty ??
  - **S3 One Zone IA** – lower cost option when IA applies but data resiliency is not require
    - Availability is 99.50%
  - **S3 Intelligent Tier** – Designed to optimize storage costs by automatically moving data to the most cost effective tier without performance impact or operational overhead
  - **S3 Glacier** – Cheaper option than previous. Competitive with on-prem data storage. Retrieval times are configurable from minutes to hours.
  - **S3 Glacier Deep Archive** – Lowest cost storage class for file storage cases where a retrieval time of 12 hours is acceptable.
  - Read through the S3 FAQ's before taking the exam

S3 – Encryption for Buckets

- By default newly created buckets are private
- Access Control can be set for a bucket with
  - Bucket Policies
  - Access Control Lists
- S3 buckets can log access to it
  - Logs can be sent to another bucket on the same account
  - Logs can be sent to a bucket owned by another account
- Buckets can be Encrypted
  - In transit – use HTTPS with ssl/tsl
  - At rest – encrypting data that is stored
    - Server side – AWS encrypts the object when it is stored
      - S3 managed keys – AWS manages the encryption keys for you with server side encryption (**SSE-S3**)
      - AWS key management service – you and AWS manage the encryption keys (**SSE-KMS**)
      - Customer provided the Keys (**SSE-C**)
    - Client side – you encrypt the object before you upload it to AWS S3 bucket

S3 - VERSIONING

- Stores all versions of an object including writes and even deletes
- Removing a delete marker will re-enable the file that was deleted. It is like a soft delete.
- Can be used as a backup tool
- Once enabled it cannot be removed from the bucket. It can only be suspended. If you need to remove you'll need to
- Integrates with LifeCycle rules
- Can activate MFA deletion for an additional layer of security

S3 – Lifecycle Management

- Automates moving objects between the different storage tiers
- Can be used in conjunction with versioning (versioning does not have to be turned on to use lifecycle management)
- Can be applied to current versions or previous versions

S3 – CROSS-REGION REPLICATION

- Versioning must be enabled in both the source and destination buckets
- Regions have to be unique when using cross-region replication.
- Same-Region replication is possible (as of recent in 2019?) to a bucket in the same region – bucket names have to be unique.
- Files in an existing bucket are not replicated automatically. Only files that are added to a bucket after **CRR** is turned on are replicated.
- All subsequent updated files will be replicated automatically
- Delete markers are not replicated
- Deleting individual versions are not replicated

S3 – TRANSFER ACCELERATION

- USES THE **CLOUDFRONT EDGE NETWORK** TO SPEED UP YOUR UPLOADS TO S3 BUCKETS
- YOU CAN UPLOAD TO CFE LOCATION WHICH THEN IT WILL TRANSFER TO AN S3 BUCKET
- USE THE TESTING TOOL TO MEASURE TIMES AND FIGURE OUT BETTER EDGE LOCATIONS TO USE FOR YOUR APPLICATION
  - HTTP://S3-ACCELERATE-SPEEDTEST.S3-ACCELERATE.AMAZONAWS.COM/EN/ACCELERATE-SPEED-COMPARSION.HTML

AWS CLOUDFRONT

- Edge Location: An AWS network location where content is cached. This is physically separate from a AWS Region or AWS Availability Zone (AZ)
- Origin: is the location where the original content resides. It can be an S3 bucket, an EC2 instance, and ELB or Route 53.
- Distribution: is the name given to a CDN which is a collection of Edge Locations
- Types of CF
  - Web Distribution – used for websites
  - RTMP – Used for media streaming. RTMP is Adobe's Real-Time Messaging Protocol
- Edge Locations are not just read-only. You can write to them too.
- Objects are cached at an Edge Location based on a Time-To-Live (TTL) value
- You can clear (invalidate) cached objects before their TTL value. However, you *will be charged extra* for that action

AWS – CLOUDFRONT LAB

- CloudFront is a GLOBAL service
- Access to Cached Content can be restricted to signed urls/cookies
- To activate, you need to create/define a Cloudfront Distribution
- This is not covered by AWS' Free Tier – you will incur in some costs
- To delete a Cloudfront Distribution you need to first mark it as "disabled" then it can be deleted

SNOWBALL

- It is a service to help move massive amounts of data
  - from on-prem to AWS
  - From AWS to on-prem
- Need to issue a request. AWS sends you a device that you need to activate and connect to your local network
- SNOWBALL - It is secure, full chain of custody, 256 bit encryption. Can transfer 50TB or 80TB. AWS performs erasure
- SNOWBALL EDGE – 100TB. It is like having a mini-AWS at your disposal. It can be clustered with other SE.
- SNOWMOBILE – 100PB. It is a truck that has serveral snowball edges networked
- Used snowball when you have
  - 44 Mbit/s to internet and you need to transfer more than 2TB of data
  - 100 Mbit/s to internet and you need to transfer more than 5TB of data
  - 1000 Mbit/s to internet and you need to transfer more than 60TB of data

AWS – STORAGE GATEWAY

- Used to move data into AWS. It is a service that connects on-prem software appliance to AWS cloud
- It can be a virtual (download a VM image) or physical device
- Options
  - **File Gateway** – files are stored as objects on s3 buckets
    - Ideal for existing operations with large bodies of documents
  - **Volume Gateway** – iSCSI block protocol - copies of data in HD for
    - stored volumes – all data on the HD will be copied/backed-up to an s3 bucket
    - cached volumes – let's you use s3 as your primary data storage while retaining frequently used data on-prem storage infrastructure
  - **Tape Gateway** – a virtual tape library
    - Allows to leverage your existing tape infra and moving your backups to cloud
    - Used for archival purposes
    - You can use Glacier service to reduce costs of archival of infrequently used data

EC2 - 101

- Elastic Compute Cloud
- Is a web service for resizable compute in the cloud
- Reduces the time it takes to obtain and reboot servers
- Allows for quick scaling (up/down) of compute requirements
- Virtual Machines in the cloud
- Run on Xen and Nitro Hypervisor software
- Pricing
  - Pay as you go
  - Pay for what you use
  - Pay less as you use more even as capacity is reserved
  - Tiers:
    - **R**eserved – capacity reservation for 1 yr or 3 yr
    - **O**n-demand
    - **D**edicated Host – exclusive for an account
    - **S**pot – bid your price for available capacity at lower rate

EC2 Tier – On Demand

- Pay a fixed rate
- Pricing is by the hour or by the second
- No commitments

EC2 Tier – Reserved Instances

- Contracts for 1 yr or 3 yr with deeply discounted prices from the on-demand tier
- Standard Reserved Instances
  - Up to 75% discount.
  - The more you pay upfront or the longer the contract the more you save
  - Once the contract starts you cannot change the computing choices you make
- Convertible Reserved Instances
  - Up to 45% discount
  - Allows you to change the computing choices within the contract to equal or greater value
- Scheduled Reserved Instances
  - These instances allow you to use compute capacity during a specific time window (fraction of day, week, a month)

EC2 Tier – Spot & Dedicated Host Pricing

- Spot
  - Price moves up/down depending on Amazon's demand and available capacity
  - Designed for apps that have flexible start/stop lifecycle
  - Ideal for Apps that are only feasible at very low compute prices
  - Accommodates urgent computing needs and large amount of additional capacity
  - If the system terminates your work load you will not be charged for fraction of hour. If you terminate the instance, you will be charged for whatever fractions you used.
- Dedicated Host
  - When regulation does not allow for sharing compute environment (i.e multitenant virtualization)
  - When software licensing does not allow multitenant virtualization

EC2 INSTANCE TYPES

- F - FPGA
- I - IOPS
- G - GRAPHICS
- H – HIGH DISK I/O
- T – CHEAP T2 MICRO
- D – DENSE STORAGE
- R - RAM
- M- MAIN CHOICE (GEN)
- C - COMPUTE
- P – PROC PIX/BIT COIN
- X – EXTREME MEMORY
- Z – EXTREME MEM & CPU
- A – ARM BASED LOADS
- U – BARE METAL

Family	Speciality	Use case
F1	Field Programmable Gate Array	Genomics research, financial analytics, real-time video processing, big data etc
I3	High Speed Storage	NoSQL DBs, Data Warehousing etc
G3	Graphics Intensive	Video Encoding/ 3D Application Streaming
H1	High Disk Throughput	MapReduce-based workloads, distributed file systems such as HDFS and MapR-FS
T3	Lowest Cost, General Purpose	Web Servers/Small DBs
D2	Dense Storage	Fileservers/Data Warehousing/Hadoop
R5	Memory Optimized	Memory Intensive Apps/DBs
M5	General Purpose	Application Servers
C5	Compute Optimized	CPU Intensive Apps/DBs
P3	Graphics/General Purpose GPU	Machine Learning, Bit Coin Mining etc
X1	Memory Optimized	SAP HANA/Apache Spark etc
Z1D	High compute capacity and a high memory footprint.	Ideal for electronic design automation (EDA) and certain relational database workloads with high per-core licensing costs.
A1	Arm-based workloads	Scale-out workloads such as web servers
U-6tb1	Bare Metal	Bare metal capabilities that eliminate virtualization overhead

EC2 – LAB NOTES

- Creating a simple Linux / Apache HTTP service
  - Spin up an EC2 instance
    - Select a type (eligible for free tier)
    - Take all defaults
    - Storage
      - You can add additional storage space by adding volumes (on top of OS requirements)
      - Root device volumes can be encrypted (it was not always the case)
    - Termination Protection
      - Tick the box for warning for accidental termination. It is turned off by default
    - Provide some tag attributes
    - Generate keys pair (pub/private)
    - Ssh to the new instance with IP address (user is ec2-user)
    - Do "sudo su" to be root
    - Install apache httpd service
    - Setup a basic index.html file
    - Turn on httpd service
    - Voila!

EXAM TIPS:

- Termination Protection is turned "ON" by default
- On EBS-backed instance, the default action is for the root EBS volume to be deleted when instance is terminated
- EBS root volumes can now be encrypted
- You can encrypt root volumes with 3rd party tools like bitlocker
- You can add more volumes and encrypt those as well

SECURITY GROUPS

- All in-bound traffic is blocked by default
- All out-bound traffic is allowed
- Changes to security groups take place immediately
- You can have any number of EC2 instances in a security group
- You can have multiple security groups attached to an EC2 instance
- Sec Groups are **STATEFUL**
  - When opening ports, it opens for both in-bound and out-bound traffic
- Network ACL are stateless
  - When opening ports, you have to specifically declare the in-bound and out-bound rules
- If you create a rule allowing traffic in, automatically there will be a rule created that allows traffic back out
- IP addresses cannot be specifically blocked with security groups
- Network ACL's allow you to define blocked IP addresses

EBS 101 – Elastic Block Store

- Provides persistent block storage volumes for use with EC2 instances
- Think of it as Hard Drive in the cloud for a given virtual machine
- This service is based on Block Storage (EFS and FSx are also based on Block Storage)
- Read this to learn about optimizing EBS volume performance
- Each EBS is automatically replicated within it's AZ to increase performance. This done to reduce latency between the virtual mother-board and the virtual HDD
  - Durability
  - Availability
  - Prevent single point of failure
- There is a restriction of 50:1 ratio between IOPS and Volume Size
- There are 5 types of EBS
  - General purpose SSD
  - Provisioned IOPS SSD
  - Throughput Optimized HDD
  - Cold HDD
  - EBS Magnetic

ELASTIC BLOCK STORAGE – TYPES

G.P. T.E.C.

Solid State Drives (SSD)

- **General purpose SSD**
  - Use Case: Most workloads
  - Balances prices v performance
  - Size: 1GB – 16TB
  - Max IOPS 16,000
  - API name gp2
- **Provisioned IOPS SSD**
  - Use Case: Databases
  - Highest performance for mission critical cases
  - Size: 4GB – 16TB
  - Max IOPS 64,000
  - API name io1

Hard Disk Drives (Magnetic)

- **Throughput Optimized HDD**
  - Use Case: Big Data/datawarehouse
  - Low cost freq. access high throughput workloads
  - API Name: st1
  - Size: 500 GB – 16 TB
  - Max IOPS: 500
- **EBS Magnetic**
  - Use Case: infrequently accessed data and you are not using GLACIER
  - Previous Generation HDD
  - API Name: Standard
  - Size: 1 GB – 1 TB
  - Max IOPS: 40 – 200
- **Cold HDD**
  - Use case: File Servers
  - Lowest Cost HDD. Less frequently accessed workloads
  - API Name: sc1
  - Size: 500 GB – 16 TB
  - Max IOPS: 250

Elastic Block Store – Exam Tips

- Volumes exist on EBS. EBS can be thought of as virtual HDD attached to a VM
- Creating an EC2 instance will automatically create a root volume
- Snapshots (a "picture" of a volume) are stored in s3
- Snapshots are a point-in-time copy of a volume
- Snapshots are incremental. Subsequent "pictures" of the same volume will only copy DIFF blocks between the original picture and the subsequent pictures.
- The first snapshot takes the longest to create.
- In a production case, you should stop the EC2 instance before creating a snapshot. This will ensure that the snapshot is pristine.
- Snapshots can be created on a running instance.
- You can create an **AMI** (Amazon Machine Image) from either volumes or snapshots
- Volume sizes and type can be changed on the fly. You may need to access the specific OS to ensure that the added volumes are visible to the OS.
- **AEV** - You cannot decrease the size of a volume, only increase.
- EBS Volumes will ALWAYS be on the same AZ as the EC2 instance that owns them
- To move an EC2 instance from one AZ to another, you need to first create a SNAPSHOT of it's Root volume, then create an AMI off of it and then use AMI to launch the EC2 instance in another AZ.
- To move an EC2 instance from one region to another, you need to snapshot > AMI and copy the AMI to the desired region. Then you can launch the EC2 instance in the desired Region/AZ.
- You can perform actions on snapshots with AWS API, CLI or AWS Console

AMI TYPES – EBS vs INSTANCE STORE

Differences between EBS and Instance Store

- All AMI's are categorized by the type of backing of the volume of the root device (where the **OS** is installed)
- **EBS** - backed
  - The root device is a volume created from an EBS snapshot
  - It persists through time after the EC2 instance is stopped
  - If the Hypervisor running an EBS-backed VM stops working you can simply re-start it
- **Instance Store** – backed
  - The root device is a volume created from a template stored in Amazon S3
  - IS-backed EC2 instances can only be rebooted or terminated. Cannot be stopped
  - If the Hypervisor running an IS-backed VM crashes you will lose all of your data/info on that instance

S. O. L. A. R. AMI

- Selection of **AMI** can be based on
  - **S**torage of the root device volume type
    - EBS backed volumes
    - Instance Store, also known as Ephemeral (temporary) Storage
  - **O**perating System
  - **L**aunch Permissions
  - **A**rchitecture (32 bit or 64 bit)
  - **R**egion or AZ

Encrypted Root Device Volumes and Snapshots

- Snapshots of encrypted volumes are encrypted automatically
- Volumes restored from encrypted snapshots are encrypted automatically
- You can share snapshots only if they are unencrypted
- Snapshots can be shared with other AWS accounts or made public
- You can now encrypt volumes upon **creation** of the EC2 instance

To encrypt a volume after it has been created:

- Create a snapshot of the unencrypted volume
- Create a copy of the snapshot and select an encryption option
- Create an AMI from the encrypted snapshot
- Launch new EC2 instance off of the encrypted AMI

CLOUDWATCH 101

MONITORS VIRTUAL THINGS LIKE:

- COMPUTE
  - EC2 INSTANCES
  - AUTOSCALING GROUPS
  - ELASTIC LOAD BALANCERS
  - ROUTE53 HEALTHCHECKS
- STORAGE & CONTENT DELIVERY
  - EBS VOLUMES
  - STORAGE GATEWAYS
  - CLOUDFRONT

MONITORS HOST-LEVEL EVENTS LIKE:

- CPU
- NETWORK
- DISK
- STATUS CHECK
  - SYSTEM – UNDERLYING HYPERVISOR (PHYSICAL COMPUTER)
  - INSTANCE – EC2 INSTANCE (VM)

EXAM TIPS:

- **CLOUDWATCH** IS USED FOR MONITORING PERFORMANCE
- IT MONITORS MOST OF AWS INFRA AS WELL AS VMS AND APPS RUNNING ON AWS INFRA
- **CLOUDWATCH** WITH EC2 WILL MONITOR EVENTS EVERY 5 MINUTES
- WITH DETAILED MONITORING YOU CAN EVALUATE IN 1 MINUTE INTERVALS
- YOU CAN CREATE CLOUDWATCH ALARMS THAT CAN TRIGGER NOTIFICATIONS
- **CLOUDWATCH** IS ABOUT PERFORMANCE (HOW ARE THE HOST MACHINE AND VM MACHINE OPERATING).
- **DO NOT CONFUSE** WITH **CLOUDTRAIL**, WHICH MONITORS AUDITING (WHO IS DOING WHAT ON WHICH RESOURCE)
- YOU CAN CREATE DASHBOARDS TO VISUALIZE:
  - ALARMS – TO NOTIFY OF THRESHOLDS ARE HIT
  - EVENTS – TO RESPOND TO STATE CHANGES IN AWS RESOURCES
  - LOGS – AGGREGATE, MONITOR AND STORE LOGS

AWS COMMAND LINE INTERFACE (CLI)

- You can interact with AWS from anywhere in the world just by using the command line
- You will need to setup a user with Programmatic Credentials in IAM
- Commands are not in the exam but it is useful to know basic CLI commands. For example
  - **aws configure** – to setup your credentials on an EC2 instance. However this is not recommended since if your EC2 instance is compromised, then your credentials are too. Use roles to avoid these issues.
  - **aws s3 ls** – this lists all of the S3 buckets
  - **aws s3 mb** <bucket_name> – this allows you to create a new bucket. The bucket_name must be unique across all existing bucket names in Amazon S3

IAM - ROLES

- Roles are more secure than storing access keys and secret access key on individual EC2 instances
- Roles are easier to manage (i.e. imagine you need to modify access keys on 1000 EC2 instances)
- Roles can be assigned to an EC2 instance even after it is created with either the web console app or CLI tool
- Roles are universal. You can use them in any region

Using Bootstrap Scripts

- Useful tool to add programmatic behaviors when launching an EC2 instance
- These require a role with sys-admin privilege is attached to the instance especially if you need to run AWS CLI commands during the bootstrapping process

EC2 INSTANCE METADATA

- This contains specific information about a given EC2 instance
- To get to this information you need to execute a curl command to
  - http://169.254.169.254/latest/meta-data
- To get the data of the bootstrap script, you need to execute a curl command to
  - http://169.254.169.254/latest/user-data

Elastic File System (EFS)

- EFS
  - Supports the Network File System v4 (NFSv4) protocol
  - You pay only for the storage you use (no pre provisioning is required)
  - Can scale up to petabytes
  - Con support thousands of concurrent connections
  - Data is stored across multiple AZ within a region
  - Provides read after write consistency
  - Microsoft Windows is not supported
- **Not In ACG Lecture**
  - Configuring Client Access on EFS
    - Configure File System Policy
    - Configure Access Points
  - **Need to discuss this with the colleagues preparing for the exam**

EC2 Placement Groups

- Relates to the placement of a set of EC2 instances within a AWS availability zone
- Clustered Placement Group
  - EC2 instances are placed as close together as possible
  - Ideal if you are looking to have low network latency, high through put or both
  - Cannot span multiple AZ
  - AWS recommends homogeneous instances
- Spread Placement Group
  - Each EC2 instance is created on a separate hardware rack with separate power and ping
  - Ideal when you have to ensure that EC2 instances will not be susceptible to hardware rack failure within the availability zone
  - Can span AZ within a region
  - There is max of 7 EC2 instances per AZ
- Partitioned Placement Group
  - Logical segments (partitions) are created and EC2 instances are added to each segment
  - Latency/throughput is low within a partition, but susceptible to hardware failure
  - Hardware failures of one partition do not affect instances running in other partitions
  - Ideal for implementations like Hadoop (HDFS), Hbase and Cassandra
  - Can span AZ within a region

Placement Group Restrictions:

- The name of placement group has to be unique within an AWS account
- Only certain EC2 instance types can be launched within a placement group
  - Compute Optimized
  - GPU
  - Memory Optimized
  - Storage Optimized
- Groups cannot be merged
- You cannot move an existing instance into a placement group. You have to create a AMI from an existing instance and then launch that AMI into the Placement Group

DATABASES 101

Understand differences between Relational databases (RDBMS) and NO-SQL DBs

- RDBMS flavors available at AWS
  - MS-SQL
  - Oracle
  - MySQL
  - PostgreSQL
  - Aurora
  - MariaDB
- NO-SQL flavors available at AWS
  - DynamoDB (JSON doc object store / key-value pairs)
- Understand architecture configuration and which you would use for what scenario
  - Multi-AZ – For disaster recovery
  - Read Replicas – For performance

Understand differences between OLTP and OLAP

- DB solutions for OLTP
  - See all SQL & NO-SQL solutions
- DB solutions for OLAP
  - Redshift
  - Used for Business Intelligence or Datawarehousing

Understand when to use Caching Mechanism

- AWS offers ElastiCache Service with 2 options
  - Memcached
  - Redis Cache
- Used to speed up the performance of existing databases

RDS – Relational Database Services Lab

Exam tips

- RDS runs on virtual machines
- You cannot login to the OS hosting RDS instances (only AWS can)
- Patching of the OS and DB is Amazon's responsibility (not yours). When you provision your own EC2 instances patching **is** your responsibility.
- RDS is not serverless (there is one exception)
- Aurora Serverless DB is the only serverless offering in RDS
- Learn about Amazon's Athena Service (WTF?)

RDS – BACKUPS, MULTI-AZ & READ REPLICAS

- RDS Backups
  - Automated Backups
    - Allow to recover DB at any point in time within a retention period
    - Retention period could be between 1 and 35 days
    - It takes a full daily snapshot and store transaction logs during the day
    - When a recovery is needed, AWS will take the most recent snapshot and apply all transaction logs that occurred from the time the snapshot was taken
    - This allows a recovery point to the most recent second within the retention period
    - Are enabled by default
    - Backup data is stored in S3 equal to the size of your DB
    - You may experience some latency during the backup window
    - When RDS instance is deleted auto backups are deleted as well

- RDS Backups
  - DB Snapshots
    - These are use-initiated
    - These are stored even after you delete the RDS instance
  - Restoring RDS instances from backups
    - It results in a brand new RDS instance with a new DNS endpoint

- ENCRYPTION At Rest
  - Supported for MS-SQL, Oracle, MySQL, MariaDB, PostgreSQL and Aurora
  - Encryption is done via Amazon's Key management Service (KMS)
  - Once encryption at rest is turned on also backups, read replicas and snapshots are encrypted

RDS – BACKUPS, MULTI-AZ & READ REPLICAS

- RDS Multi-AZ
  - This is for disaster recovery purposes only (not for performance – see Read Replicas)
  - Allows you to have an exact copy of your DB in another AZ
  - AWS handles the replication for you
  - In the event of planned DB maintenance, DB instance failure or AZ failure, Amazon RDS will automatically fail over to the stand-by so that DB operations can resume quickly without admin intervention
  - Multi-AZ is available for MS-SQL, PostgreSQL, MySQL, Oracle, and MariaDB (Aurora is by design resilient to disasters)
- Read Replicas
  - Allows you to have read-only replicas of your primary database
  - Uses asynchronous replication from primary to the replica
  - Effective in situations where an application has read-intensive workloads
  - This is available for MySQL, Oracle, PostgreSQL, MariaDB and Aurora
  - Used for scaling an application (not for Disaster Recovery – DR)
  - You can have up to 5 read-replica copies of your DB
  - You can have read-replicas of read-replicas (latency may be an issue in this scenario)
  - Each RR will have its own DNS end point
  - RR can have multi-AZ
  - You can create RR of multi-AZ databases
  - RR can be promoted to be a primary DB (this breaks replication)
  - RR can be in a different region

REDSHIFT – OLAP DB

FEATURES

- USED FOR BUSINESS INTELLIGENCE AND DATA WAREHOUSING
- FAST, POWERFUL, FULLY MANAGED
- PETABYTE SCALE
- AS LOW AS $0.25 X HOUR. $1000 PER YR PER 1TB
- CHEAPER THAN MOST OTHER DATA WAREHOUSING SOLUTIONS

- **PRICING:** CHARGES ARE BASED ON
  - COMPUTE-NODE-HOURS
  - BACKUP
  - DATA TRANSFERS WITHIN A VPC
- WHEN IN MULTI-NODE, LEADER NODE DOES NOT INCUR IN CHARGES.

- KNOW THE DIFFERENCE BETWEEN OLTP/OLAP
- CONFIGURATION
  - SINGLE-NODE
    - 160 GB STORAGE SIZE
  - MULTI-NODE
    - 1 LEADER NODE
    - N COMPUTE NODES
    - UP TO 128 COMPUTE NODES
- USES ADVANCE COMPRESSION BY COLUMN BECAUSE SIMILARITY IN THE DATA TYPE
- USES LESS SPACE THAN TRADITIONAL OLTP DBS
- USES MASSIVE PARALLELISM
- EASY TO SCALE OUT BY ADDING NODES
- BACKUPS ARE ENABLED BY DEFAULT BY 1 DAY RETENTION (UP TO 35)
  - UP TO 3 COPIES STORED IN S3
  - ASYNCHRONOUSLY REPLICATES SNAPSHOTS TO ANOTHER REGION FOR DR PURPOSES
- SECURITY
  - ENCRYPTED IN TRANSIT WITH SSL
  - ENCRYPTED AT-REST AES-256
  - KEY MANAGEMENT THROUGH
    - HSM (YOU MANAGE)
    - AWS KMS (AWS MANAGES)
- AVAILABILITY
  - 24/7 WITHIN ONE AZ (NOT MULTI AZ)
  - YOU CAN MOVE BACKUPS TO ANOTHER AZ AND RESTORE THERE FOR DR PURPOSES

AWS AURORA

- It is Amazon's proprietary SQL database compatible with MySQL and PostgreSQL
- 2 copies of your data is contained in each AZ with a minimum of 3 AZ's (6 copies of your data)
- It is only available in regions with 3 or more AZs
- You can share Aurora snapshots with other AWS accounts
- Offers 2 types of replicas
  - Aurora Replicas
  - MySQL replicas
- Offers automated failover is only available with Aurora replicas
- Automated backup is turned on by default

ELASTICACHE

- It is a service to assist with web app performance improvement
- It comes with 2 options
  - MemcacheD
    - Very simple, high-performing option
    - Lacks some features
  - REDIS Cache
    - Offers more features than MemcacheD
    - Complex objects
    - Available in multiple AZ
    - You can do backups and restore
    - And much more - - -> see pic

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

AWS – Route 53 – A domain name service (DNS 101)

- Elastic Load Balancer service (ELBs) do not have pre-defined IPv4 addresses
- You resolve to ELBs using a DNS name
- Understand the difference between
  - CNAME
  - Alias Record
- On the exam, when given the choice between a CNAME or an Alias Record, always pick the Alias Record (it'd be nice to know why tho)
- Understand Common DNS Types
  - SOA Records – start of authority record
  - NS Records – Name Server – indicates which DNS server is authoritative for a domain
  - A Records – most basic DNS records. Points a domain/sub-domain name to an IPv4 address
  - CNAMEs – canonical name record
  - MX Records – electronic mail
  - PTR Records – allow lookup of DNS records given an IP address

AWS – Route 53, Register a Domain Name

- You can buy domain names from Amazon Route 53
- It can take up to 3 days to register and propagate through the internet

AWS – Route 53, Routing Options

- Simple Routing (round robin)
- Weighted Routing (%-based routing)
- Latency-based Routing (ping-based routing)
- Failover Routing (active/passive routing)
- Geolocation Routing (region based)
- Geo-proximity Routing (Traffic Flow Only)
- Multi-value Answer Routing (simple w/ health check)

- Resources:
  - https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/ResourceRecordTypes.html

AWS – Route 53, Routing Options

- Simple Routing
  - As the name indicates, it is the most basic routing option
  - You can have a single DNS record with multiple IP addresses associated to it
  - When a browser hits the DNS record and if multiple IP addresses have been defined, Route 53 will return a random IP address from the list

AWS – ROUTE 53, ROUTING OPTIONS

- Weighted Routing Policy
  - R53 uses a % with which to split the traffic coming in to a given DNS
  - The % splits have to add to 100

- Helathchecks (this is for all routing policies)
  - You can set health checks on individual record sets
  - If a record fails the health check, it will be removed from R53 until it passes the healthcheck
  - You can set SNS (simple notification service) alerts to notify you that a Health Check failed

AWS – Route 53, Routing Options

- Latency-based Routing Policy
  - Route 53 detects the latency for a specific request
  - It then routes the user to the AWS that is experiencing the least latency from the perspective of the user
  - If you are in Japan and hit a DNS record and there is low latency to Sydney than to US-West, the traffic will be routed to Sydney

AWS – Route 53, Routing Options

- Fail-over Routing Policy
  - You need to configure the passive and active IP addresses
  - You also need to define Health Checks for the EC2 instances behind your IP addresses
  - Route 53 detects a request to a hosted domain
  - All requests will go to the Active IP address as long as it has a passing health-check
  - In case the health-check fails, the R53 will re-route all requests to the passive IP address registered in the routing policy

AWS – Route 53, Routing Options

- Geo-Location Routing Policy
  - You define which are the geographical locations that a web request should go to
  - If you define that users in Europe should go to IP address "A" and USA users should go to IP address "B" R53 will forward requests according to settings
  - **Questions:**
    - What happens when users come in from regions not defined?
    - What happens when one of the regions fails health-checks?

AWS – Route 53, Routing Options

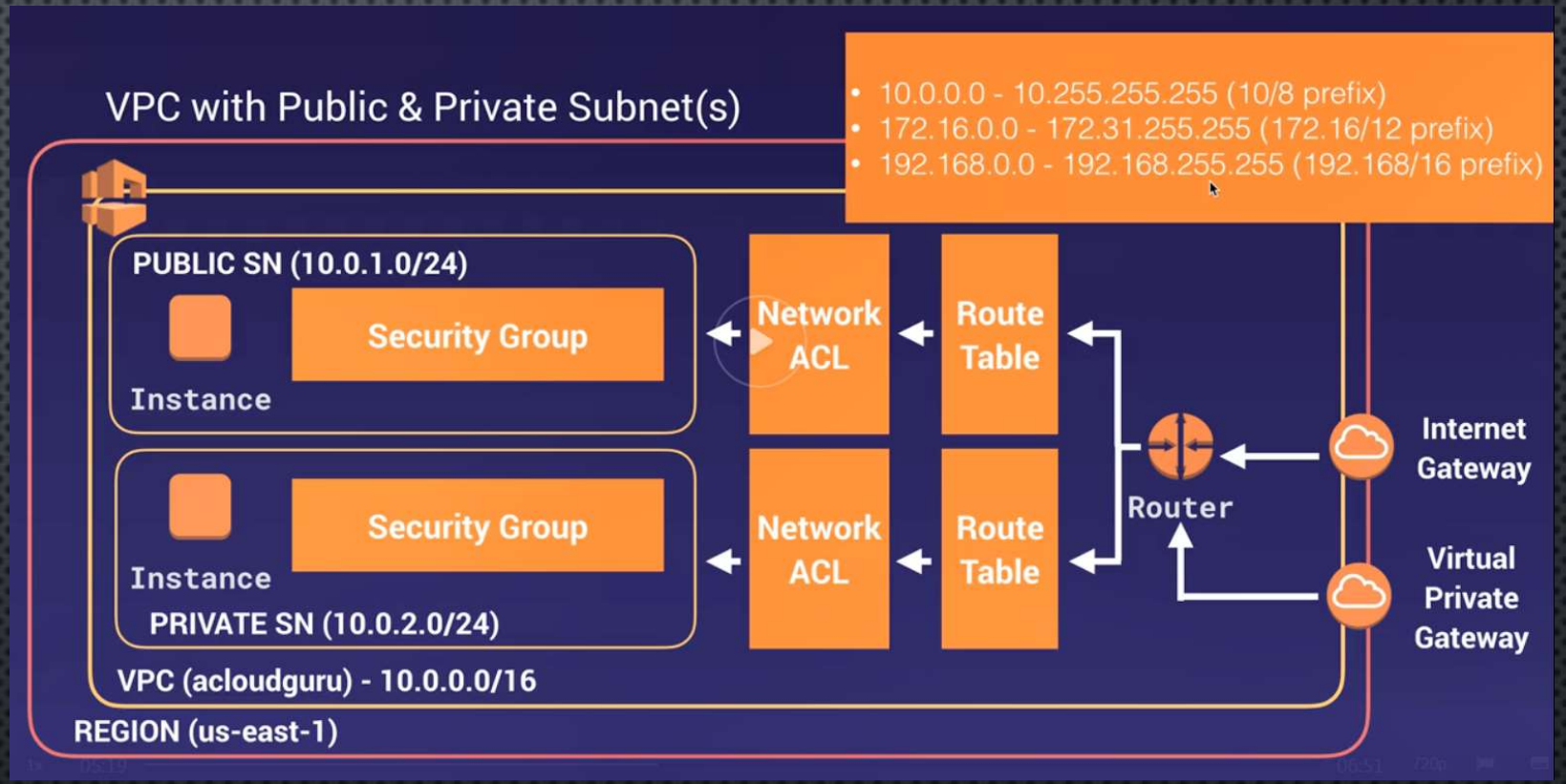
- Geo-Proximity Routing Policy
  - Lets traffic to your AWS resources based on
    - Geographic location of your user (lat/lon)
    - Geographic location of your resources (lat/lon)
    - You can adjust (increase or decrease) traffic flow based on a bias
  - To use this feature, you must use R53 Traffic Flow tool

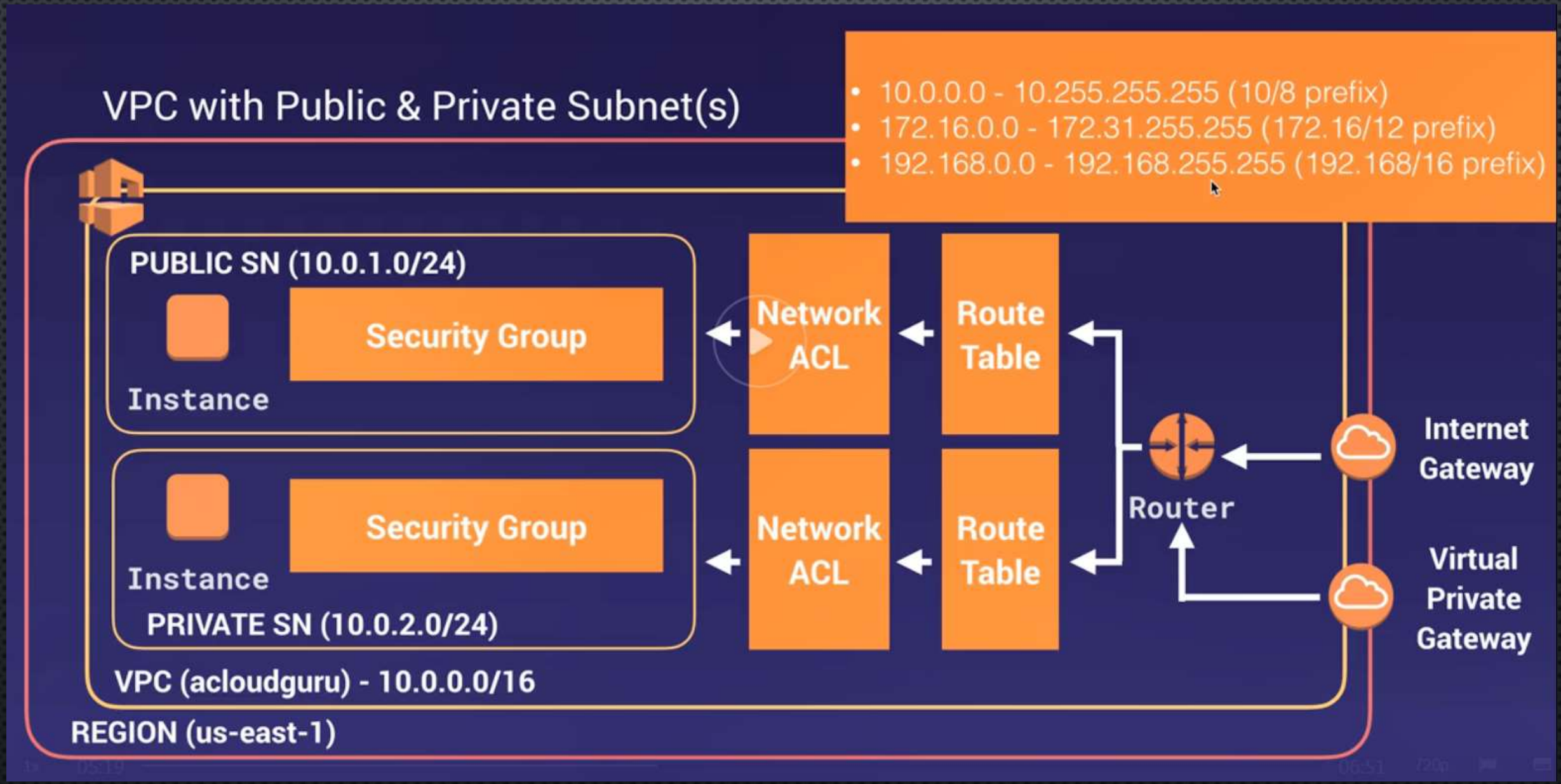
AWS – Route 53, Routing Options

- Multi-Value Answer Routing Policy
  - This is basically simple routing but for each IP address defined as an independent record
  - Health-checks need to be enabled for this to work

AWS - VPC

- Think of a VPC as a logical datacenter in AWS
- A VPC consists of
  - Internet Gateways & Virtual Private Gateways
  - Route Tables
  - Network ACLs
  - Subnets
  - Security Groups
- 1 Subnet = 1 AZ, meaning subnets cannot span over multiple AZs
- 1 AZ can have multiple subnets
- Security Groups are Stateful
  - You can add allow & deny rules
  - If you define allow or deny, you don't need to define the return rule
- Network ACLs are Stateless
  - You can add allow & deny rules
  - Opening in-bound port does not automatically add out-bound traffic. You have to add manually
- No transitive peering is allowed
  - Given: Vpc a < - > Vpc b & vpc b < - > vpc d
  - Then no traffic can go from VPC A to VPC D unless you define peering between VPC A and VPC D




VPC with Public & Private Subnet(s)
• 10.0.0.0 - 10.255.255.255 (10/8 prefix)
• 172.16.0.0 - 172.31.255.255 (172.16/12 prefix)
• 192.168.0.0 - 192.168.255.255 (192.168/16 prefix)
PUBLIC SN (10.0.1.0/24)
Instance
Security Group
Network ACL
Route Table
Router
Internet Gateway
Virtual Private Gateway
Instance
Security Group
Network ACL
Route Table
PRIVATE SN (10.0.2.0/24)
VPC (acloudguru) - 10.0.0.0/16
REGION (us-east-1)

VPC – Exam Tips

- When you create a VPC, the following are created by default
  - Default Route Table
  - Network ACL
  - Default Security Group
- When you create a VPC, you have to define
  - Subnets
  - Internet Gateway
- Remember
  - For 2 different AWS accounts the same AZ name could refer to different, physical datacenters. The AZs are randomized when assigned.
  - Amazon always reserves 5 ip addresses within your subnets
    - x.x.x.0 – network address
    - x.x.x.1 – reserved for VPC router
    - x.x.x.2 – reserved for mapping to the AWS-provider DNS
    - x.x.x.3 – future use (not really disclosed why)
    - x.x.x.255 – standard broadcast address
- You can only have 1 internet gateway per VPC
- Security Groups cannot span VPCs

VPC – NAT INSTANCES VS NAT GATEWAYS

- **RESOURCE:** https://docs.aws.amazon.com/vpc/latest/userguide/VPC_NAT_Instance.html
- **NAT INSTANCE TIPS**
  - THESE ARE NOT POPULAR ANY MORE
    - NOW WE HAVE **NAT GATEWAYS**
  - MUST PROVIDE AND EC2 INSTANCE BE IN A PUBLIC SUBNET AND **DISABLE SOURCE/DESTINATION CHECKS OF THAT INSTANCE**
  - THERE MUST BE A ROUTE (0.0.0.0/0) FROM THE PRIVATE SUBNET TO THE NAT INSTANCE FOR IT TO WORK
  - THE SIZE/RESOURCES OF THE NAT INSTANCE WILL DICTATE ITS ABILITY TO SUPPORT TRAFFIC
  - IF YOUR NAT INSTANCE IS BOTTLENECKING, YOU CAN INCREASE THE INSTANCE SIZE
  - YOU CAN CREATE HIGH-AVAILABILITY USING AUTO SCALING GROUPS, MULTIPLE SUBNETS IN DIFFERENT AZS AND A SCRIPT TO AUTO-FAIL-OVER, BUT THIS IS VERY DIFFICULT TO SCRIPT AND MAINTAIN
  - NAT INSTANCES SHOULD BE WITHIN A SECURITY GROUP (TYPICALLY A WEBDMZ SG)
  - YOU ARE RESPONSIBLE FOR PATCHING THE OS OF YOUR NAT INSTANCE
- **NAT GAWTEWAYS TIPS**
  - NAT-GW ARE REDUNDANT WITHIN AN AZ
  - PREFERRED ENTERPRISE WAY TO ACCESS INSTANCES IN PRIVATE SUBNETS
  - SPEED RATES START AT 5 GB/S AND SCALES UP TO 45 GB/S
  - NO NEED TO PATCH THE OS
  - NO NEED TO ASSOCIATE THEM TO A SECURITY GROUP
  - AUTOMATICALLY GET ASSIGNED AN IP ADDRESS
  - REMEMBER TO UPDATE YOUR DEFAULT VPC ROUTE TABLE AND ATTACH TRAFFIC TO YOUR NAT GW
  - NO NEED TO DISABLE SOURCE/DESTINATION CHECKS
  - MULTIPLE AZS CAN SHARE A SINGLE NAT GW BUT THIS CREATES A SINGLE POINT OF FAILURE CONCERN
  - IF YOU HAVE RESOURCES IN PRIVATE SUBNETS IN MULTIPLE AZS THAT NEED INTERNET ACCESS, CONFIGURE A NAT GW FOR EACH OF THE AZ AND UPDATE THE DEFAULT ROUT TABLE TO ALLOW 0.0.0.0/0 TRAFFIC

VPC – NETWORK ACL vs SECURITY GROUPS

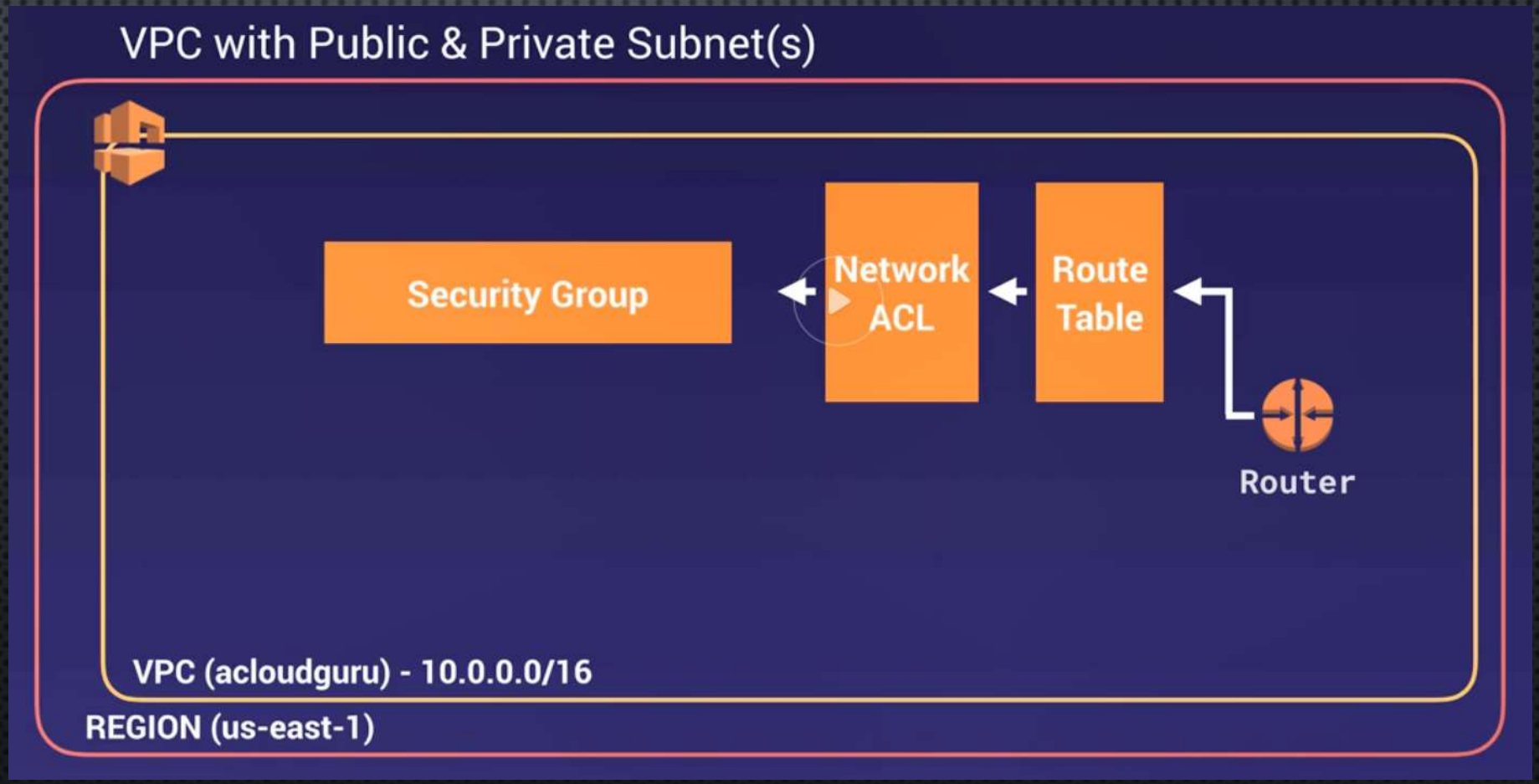
- Traffic Network ACL rules are evaluated **BEFORE** traffic rules in Security Groups
- Upon creation of a VPC, a default NACL is created
- A default NACL allows all in-bound/out-bound traffic by default
- You can create custom NACL
- By default, custom NACL denies all in-bound/out-bound traffic until you add specific in-bound/out-bound rules
- Each subnet in your VPC must be associated to a NACL
- When a subnet is created, it will be assigned to the default NACL. You can change that association later to a custom NACL
- To block specific IP addresses use a NACL rule deny access to that IP. You cannot use Security Groups to block IP addresses
- A NACL can be associated to multiple subnets, but a subnet can only be associated to 1 NACL at a time
- When you associate a subnet to a NACL, the previous NACL association will be removed
- A NACL contains a list of allow/deny rules that are ordered sequentially. Starting from the lowest number, the sequence number indicates the order in which the rule will be evaluated by networking HW/SW.
- NACLs have separate in-bound/out-bound rules which can either allow or deny traffic
- NACLs are stateless. Responses to allowed in-bound traffic are subject to the rules for out-bound traffic (and vice versa)
    - **Note**: with Security Groups (which are stateful) that consideration is not need. Allowed in-bound traffic does not require to define the allowed out-bound traffic.

VPC – STEPS TO CREATE VPC WITH PUB/PRIVATE SUBNETS

- Create VPC (10.0.0.0/16 – 65K+ ip addresses)
  - Default SG
  - Default NACL
  - Default RT
- Create Internet Gateway
  - Attach IG to VPC
- Create a subnets
  - Public (10.0.1.0/24 – 256 ip addr)
  - Private (10.0.2.0/24 – 256 ip addr)
- Create custom RT for public-exposed traffic
  - In bound traffic
    - Allow ports 22, 80, and 443 (SSH, HTTP, HTTPS)
  - Out bound traffic
    - Allow all (0.0.0.0/0)

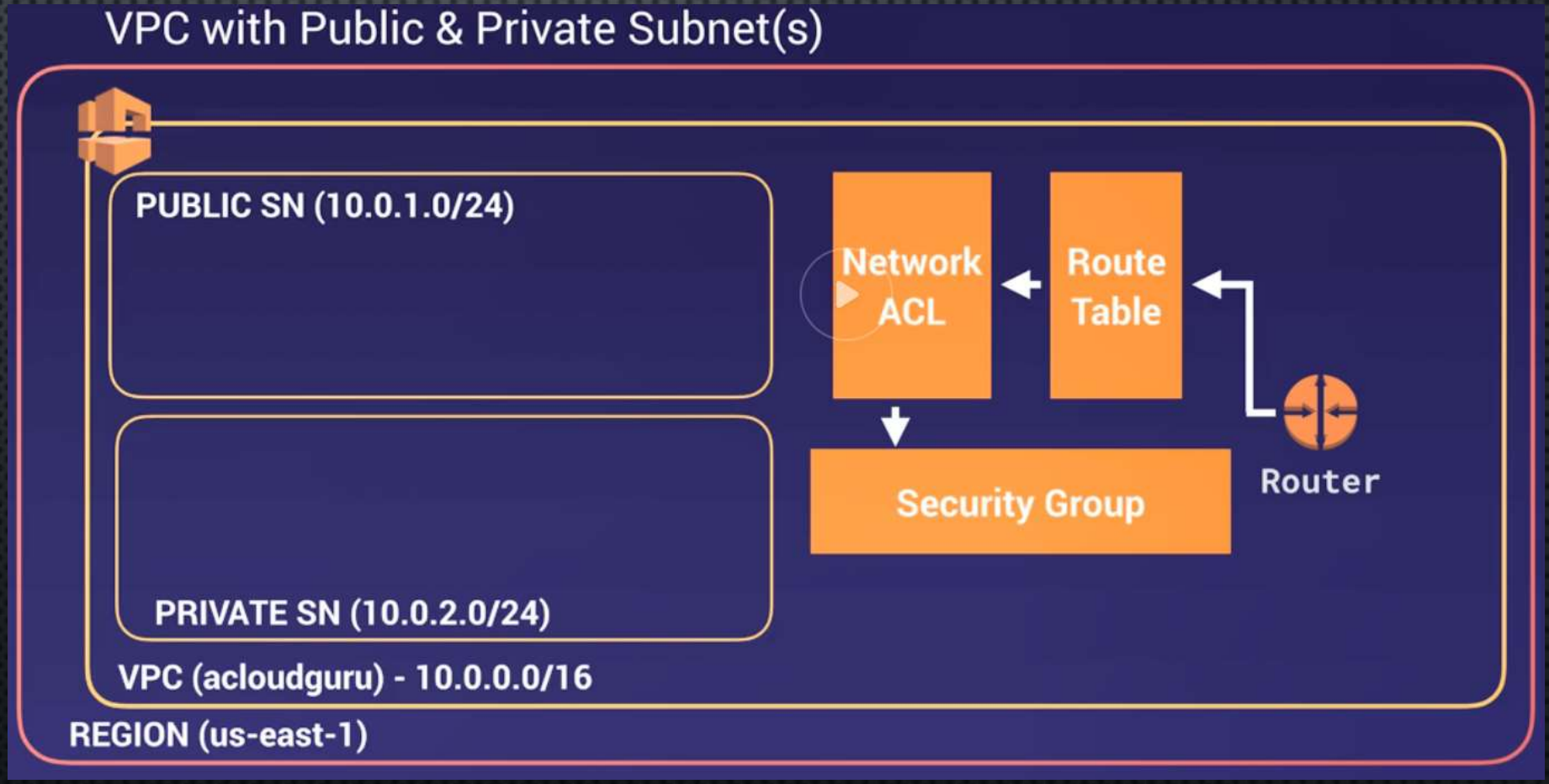
VPC – STEPS TO CREATE VPC WITH PUB/PRIVATE SUBNETS

- Select a Region
- Create VPC in that region
  - CIDR 10.0.0.0/16 for biggest address possible – 65K+ ip addresses
  - Give a name
- By default, the following are created
  - Default vpc Routing Table
  - Default vpc NACL
  - Default vpc Security Group



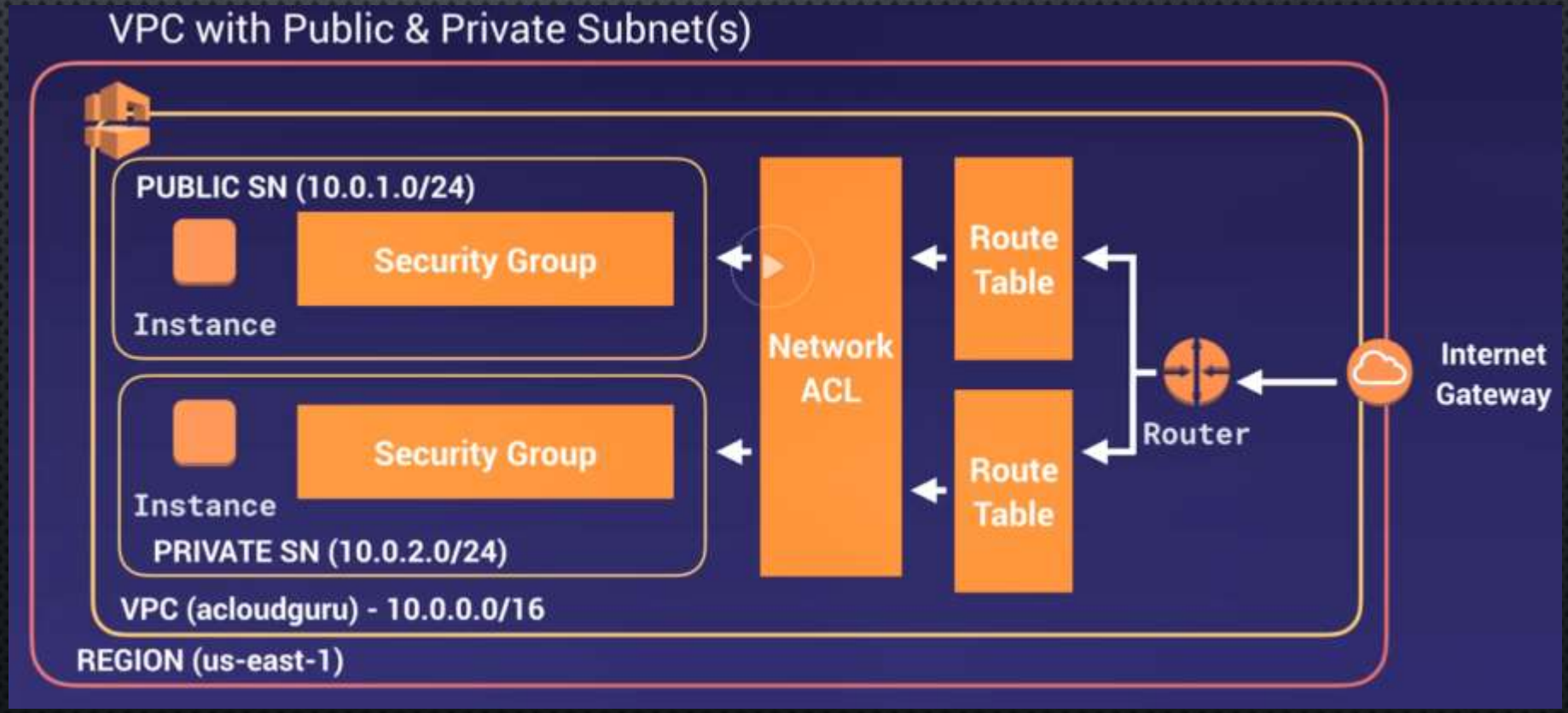
VPC – STEPS TO CREATE VPC WITH PUB/PRIVATE SUBNETS

- Create **Public** Subnet
  - CIDR 10.0.1.0/24 – 256 IP addresses
  - Amazon reserves 5 of those for their use cases
  - Select AZ for subnet
  - For name, consider using "<CIDR> – <AZ>"
  - Set the subnet with auto-assign public IP addresses
- Create **Private** Subnet
  - Use a different AZ
  - CIDR 10.0.2.0/24
- Associate all subnets to custom VPC



VPC – STEPS TO CREATE VPC WITH PUB/PRIVATE SUBNETS

- Add Internet Gateway and attach it with the VPC
- Configure the Main Route table
  - Always keep it PRIVATE without access to the internet
- Create a new Public Route Table and give it access to the internet
  - Add a route for IPv4 to the Internet Gateway associated to the VPC: 0.0.0.0/0 (all traffic)
  - If needed add a route for IPv6 (::/0)
- Associate the Public subnet to this route table
- Provision an EC2 Instance within our Custom VPC with the Public subnet.
- Create a WebDMZ Security Group to firewall the publicly exposed EC2 instances
- Provision an EC2 instance within Custom VPC with the Private subnet
- Use the default Security group for the VPC



VPC – STEPS TO CREATE VPC WITH PUB/PRIVATE SUBNETS

- Create a new security group for your assets deployed in your Private subnet
- Give it a Name and Description
- Restrict in-bound traffic that only comes from Public subnet
- In-bound traffic – allow
  - Ping – All ICMP
  - HTTP - 80
  - HTTPS - 443
  - SSH - 22
  - DB – your DB's port
- Change the security group of your private EC2 instances
- For private instances you need to provide a safe route out to the Internet. For that you can use
  - NAT Instances (EC2 instance)
  - NAT Gateways – a more modern way to provide access to internet resources

VPC Flow Logs

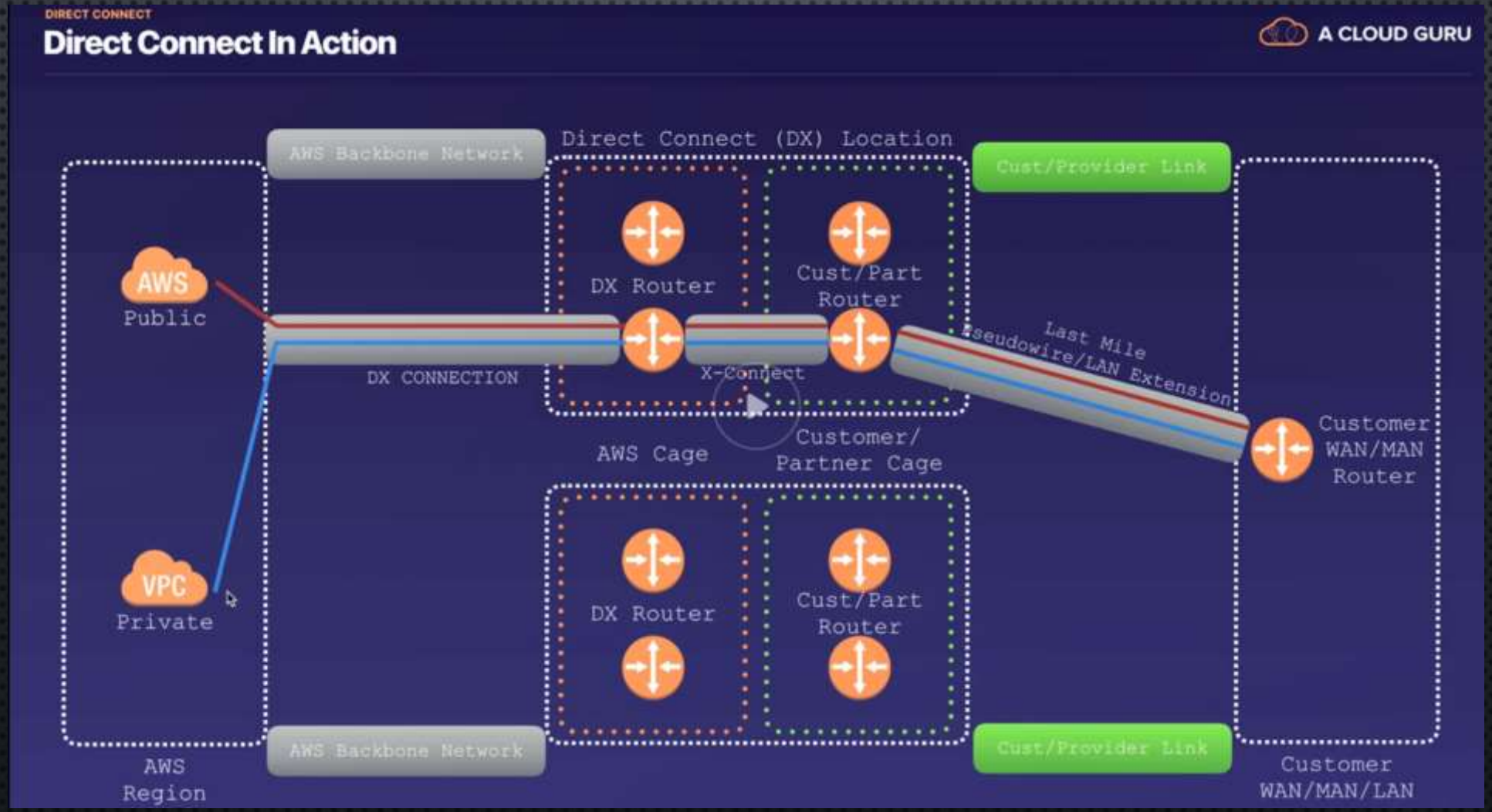
- Flow Logs:
  - Allow you to capture information about IP traffic going through network interfaces in you custom VPC
  - Flow Log data is stored in Cloud Watch or in S3 bucket
  - When using Cloud Watch you can view logs there
  - Can be created at the level of: **VPC** > **Subnet** > **network interface** (Elastic Network Interface - ENI)
- You cannot enable Flow Logs for VPCs that are peered with your VPC unless the peer-VPC is in your account.
- You cannot add Tags for a Flow Log (name it wisely)
- After you have created a flow log, you cannot change it's configuration.
  - For example: you cannot associate a different IAM role for a Flow Log after its creation
- Not all IP traffic is monitored
  - Traffic generated by instances when they contact Amazon's DNS server. However, if you use your own DNS server, then all traffic to that DNS server is logged.
  - Traffic generated by a Windows Instance for Amazon Windows license activation
  - Traffic to/from 169.254.169.254 for instance meta data
  - DHCP traffic
  - Traffic to the reserved IP address for the default VPC router

VPC – Bastion Hosts

- Is an EC2 instance that lives in a public sub net in custom VPC to allow remote access to resources in private subnets
- A NAT Gateway or NAT instance is used to provide internet traffic to EC2 instances in private subnets
- A bastion is used to securely administer EC2 instances, via SSH or RDP (windows machines). Bastions are also known as jump-boxes elsewhere.
- You cannot use a NAT gateway as a bastion host

DIRECT CONNECT

- DIRECT CONNECTION FROM YOUR DATA CENTER TO AWS
- USEFUL FOR HIGH-THROUGHPUT WORKLOADS (LOTS OF NETWORK TRAFFIC)
- USEFUL FOR ESTABLISHING A STABLE, RELIABLE, AND SECURE CONNECTION TO AWS



VPC ENDPOINTS

- Enables you to privately connect your VPC to
  - AWS supported services
  - VPC endpoint services powered by Private Link
- It does not require you to have connectivity via
  - Internet Gateway
  - NAT Device / NAT Gateway
  - VPN connection
  - AWS Direct Connect
- EC2 instances do not require a public IP address
- Traffic between your VPC and AWS services do not have to leave the AWS network
  - Example – you have an EC2 instance that needs to drop objects into an S3 bucket
- VPC Endpoints are Virtual Devices
  - Scale horizontally
  - Are redundant
  - Highly available
  - Do not impose
    - availability risks
    - Bandwidth constraints on your network traffic
- Types of VPC Endpoints
  - Interface endpoints
  - Gateway endpoints
- Currently GW Endpoints support
  - Amazon S3
  - DynamoDB

Elastic Load Balancers (ELBs)

- Application Load Balancer (Layer 7)
  - http/https
  - Intelligent
  - Advance / intelligent routing
- Network Load Balancer (layer 4)
  - Tcp traffic balancer
  - Use if you want
    - Maximum performance
    - Fixed IP address
  - Millions of requests per second
  - Highest cost
- Classic Load Balancer
  - HTTP/HTTPS
  - Not app aware
  - Sticky session (x-forwarded-for header, IPv4)
  - Lowest cost
- Errors
  - 504 errors – Gateway Error on the app at the web server or back-end, DB
- Instances monitored by ELBs are reported as
  - In service
  - Out of service
- Health checks ensure that an instance under ELB is healthy by hitting an HTTP path to that instance
- Load Balancers have their own DNS name
- You will never get an IP address for a load balancer
- SLAs Guarantee (all 3 types of ELBs)
  - Monthly availability of 99.99%
- Read the ELB FAQ:
  - https://aws.amazon.com/elasticloadbalancing/faqs/

Advanced Load Balancer Theory

- Sticky Sessions
  - Users can stick to the same EC2
  - Useful when there is application state for given user stored in the EC2
- Cross Zone LB
  - Allows an ELB to forward traffic to a different AZ other than the one it has been defined for
- Path Patterns
  - Allows an ELB to forward traffic to a different AZ based on the URL path coming in the HTTP request

HAA – High Availability Architecture

- Always design for failure
- Use multiple AZ in multiple regions whenever you can
- Know the difference between
  - RDS Multi AZ
  - RDS Read Replicas
- Know the difference between
  - Scaling Out – procuring more similar instances (launching more T2 micro EC2's)
  - Scaling Up – Increasing the resources of an instance (from T2 – > XL)
  - S3 storage classes
    - HA S3 – Standard S3, Infrequent Access S3
    - Not HA S3 – Infrequent Access 1 zone or Reduced Redundancy Storage (RRS)
- For the exam, read questions carefully and always consider the cost factor

HA – EXAM TIPS

- Load balancer types
  - Application LB
    - Layer-7 aware. These LBs know more intimate details about the application metadata
    - With these LB's you get a DNS / URL – never an IP address
  - Network LB
    - Extreme performance at the hardware level (Layer-4 aware)
    - Can hard-code IP address
    - Highest cost
  - Classic LB
    - Lowest cost
    - No "intelligent" routing
    - X-forwarded-for header tells you the IP address of the client calling the app
    - With these LB's you get a DNS / URL – never an IP address
- ELB Reporting
  - Errors
    - 504 error code – Gateway timeout
    - Error is not with the LB. It is probably with the app or the backend (App server or DB)
  - EC2's monitored by ELB present either
    - In-service
    - Out-of-service
    - You can track instance's health with Health Checks (getting a 200 HTTP code after calling GET HTTP verb)
- Read ELB FAQs
- LB Theory
  - Sticky Sessions
    - Allows users to "stick" to the same instance during their session
  - Cross-zone LB
    - Enables you to load-balance traffic between Availability Zones
  - Path Patterns
    - Enables you to load-balance traffic between different EC2 instances based on URL patterns
- CloudFormation
  - A way to script definition of a cloud environment – lots of granular control: VPC, subnets, NACL, EC2, etc.
  - Quickstart Templates – predefined CloudFormation cloud environments targeting specific tech stacks
  - Targeted to DevOps teams who want full control of their environments
- Elastic Beanstalk
  - Designed to quickly deploy and manage Applications without worrying about infra settings
  - User deploys the app and EBS takes it from there (similar to Cloud foundry)
  - Targeted to Developers who do not have/want/need to deal with infrastructure settings

AWS SQS – Simple Queue Service

- The first AWS service – has lots of legacy
- Messaging / Queuing platform
- Decouples components of an application
- Messages are stored in a fail-safe queue
- SQS is "pull" based not "push" based. A caller needs to access the queue to either add or get messages from the Queue
- Messages can only be up to 256 KB of text in any format
  - Messages can be bigger but it uses S3 to store those – up to 2 GB
- Messages can be retrieved programmatically using the SQS API
- Auto-scaling can be triggered by number of messages in the Queue
- Messages can be kept from 1 min – 14 days – default is 4 days
- Msgs are subject to **Visibility Timeout** – a period of time (12 hours max) that a msg that is accessed by an EC2 is invisible to any other EC2 which is getting messages – this could result in the same message being delivered twice
- Long Polling – a request for a message does not return response if Q is empty or Long Poll times out
- Short Polling – A request for msg returns immediately even if Q is empty

- **Standard Queue Type**
  - Is the default type
  - Nearly unlimited number of TPS
  - Guarantee that messages are delivered at least once
  - Provide "Best-Effort" Ordering: Messages are generally delivered in the order they were sent
  - However, multiple copies of a message may be delivered out of order more than once
- **FIFO Queue Type**
  - First-in First-out
  - Up to 300 TPS
  - Complements the Standard Queue
  - Order of messages is strictly preserved
  - Message delivery is guaranteed to be only once and remains available until a consumer deletes it
  - Duplicates are not introduced into the Queue

AWS Simple Workflow Service (SWF)

- Know the difference between SWF vs SQS
- Retention
  - SQS has a retention period of 14 days
  - SWF – workflow executions can persist for up to 1 year
- Orientation of Use Case
  - SQS – offers a "Message-Oriented" API solution
  - SWF – Offers a "Task-Oriented" API solution
- Duplication
  - SQS – has flexibility of handling duplicated messages or enforcing FIFO
  - SWF – ensures that tasks are assigned only once and are never duplicated
- Tracking
  - SQS – your application needs to implement app-level tracking of messages
  - SWF – Keeps track of all tasks and events in an application
- Actors (applies only to SWF)
  - Workflow Starters – applications that can initiate an SWF workflow
  - Deciders – decide what to do next after a task finishes in a workflow, handle error state
  - Activity Workers – Carry out logic work within a step in the workflow

AWS – Simple Notification Service (SNS)

- Service that allows setting up, operate and send messages from the cloud
- Usable for Push Notifications for mobile applications
- Able to deliver messages by SQS, SMS Text or Email
- Uses Topics to associate publishers and subscribers for message distribution
- Messages published to SNS are stored redundantly across multiple AZs
- Benefits
  - Instantaneous push-based delivery
  - Simple API and simple integration with Apps
  - Flexible message delivery over multiple transport protocols
  - Inexpensive, pay-as-you-go model. No upfront costs
  - AWS console offers simple UI to manage
- SNS vs SQS
  - Both are messaging systems
  - SNS is push-based. SQS is pull-based

AWS – Elastic Transcoder

- Is a could-based, media transcoder
- Converts media files from their original source format into different formats that will play on different devices (tablets, smart phones, PCs, etc)

AWS – API GATEWAY

- API GW – is a "door" from the internet to your AWS environment (different than internet gw to EC2 instances)
- It has caching capabilities to increase performance (need to set a TTL in seconds)
- It has a low cost
- Scales automatically
- It can be throttled to prevent attacks
- Results can be logged to CloudWatch
- If your application uses multiple domains with API GW, ensure that CORS is enabled
- Remember that CORS is enforced by your browser to reduce vulnerabilities and injection attacks of malicious code

AWS – KINESIS

- It is an AWS service/platform for processing data streams
- Comes in 3 types: streams, Firehose, Analytics
- K-Streams
  - Data producers can send data to Kinesis
  - Data has persistence and is stored 24 hours (default) up to 7 days
  - Data is stored in "shards" which are logical groupings that Data Consumers (EC2) instances can access, process and store the results of the work
  - Performance of Shards
    - 5 tps for read rate up to 2 mb maximum read rate per second
    - 1000 tps for writes up to 1 mb maximum write rate per second (including partition keys)
    - Data capacity for a stream depends on the number of shards that are provisioned for it
- K-Firehose
  - Data is not persisted. It has to be processed as it is received
  - Lambda functions are good processors of in-coming data
- K-Analytics
  - Works with k-streams or k-firehose and analyzes the data on the fly
  - After analysis, it sores the data in S3, Redshift or Elasticsearch Cluster

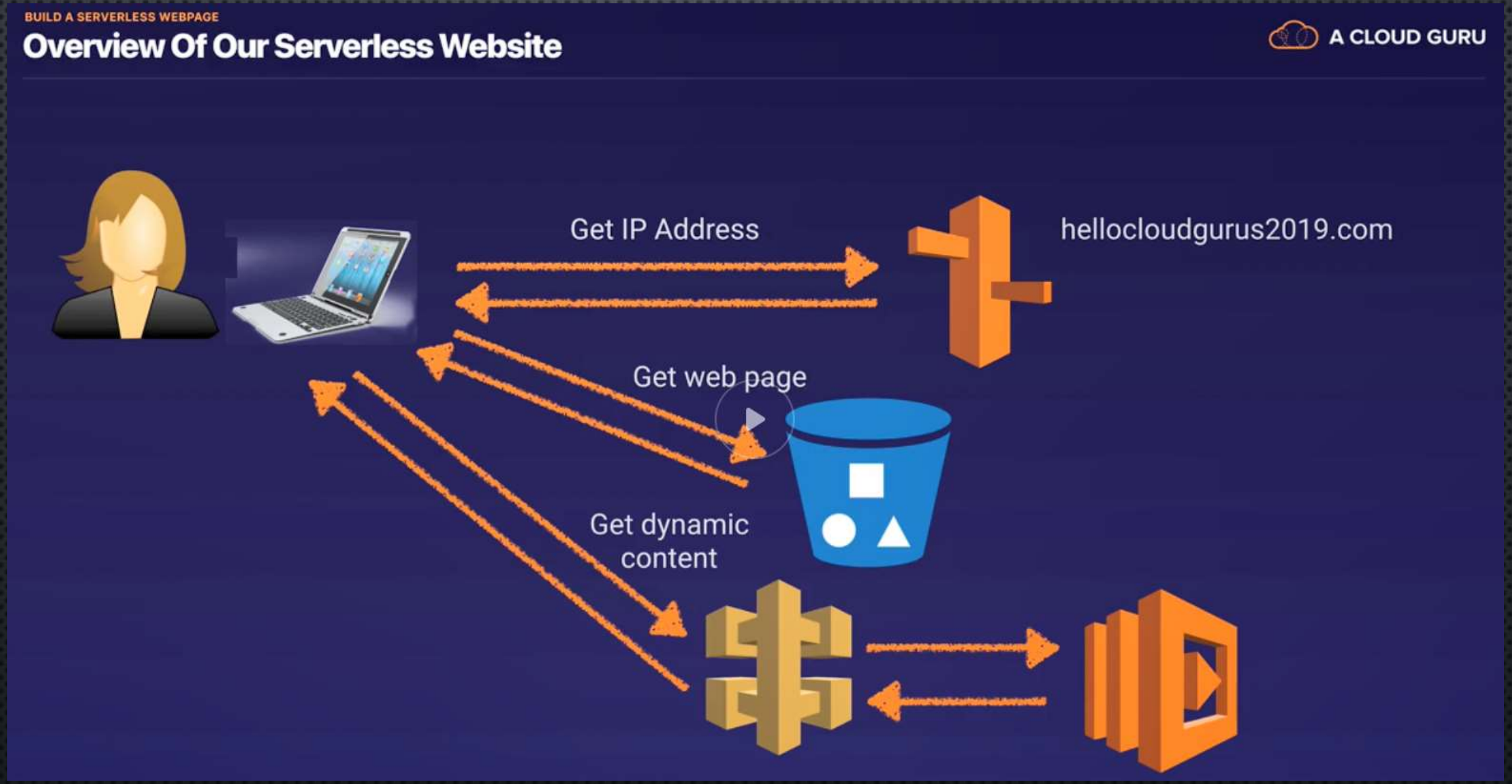
IAM – COGNITO

- AWS allows you to leverage Web Identity Federation services (i.e. google id, facebook id, amazon id, etc) to provide temporary access to AWS services for authenticated users via Amazon Cognito
- Allows sign-up and sign-in to apps
  - If not using federated IDs, then the app can use Cognito user pool
- Allows access for "guest" users
- Acts as an ID broker between web-id providers and AWS
- Successful Authentication results in the creation of JWT to grant access to services
- Synchronizes user data for multiple devices
- Recommended for all mobile applications that access AWS services
- Identity pools
  - These deal with authorization of users
  - Provide temporary access to AWS resources

AWS – Lambda Functions

- Lambda scales out (not up) automatically
- Lambda functions are independent, 1 event = 1 function
- Lambda is server-less. Understand what AWS services are server-less:
  - Aurora DB
  - DynamoDB
  - API Gateway
  - S3
  - Lambda
- Lambda functions can trigger other Lambda functions
- Draw-backs/Trade offs
  - Solutions can get really complex with lambda
  - Networks of lambda functions can be difficult to debug
  - Use AWS X-Ray service to debug server-less applications
- Lambda functions can help you trigger actions globally across AWS' global infrastructure (i.e. Back-up an S3 bucket to another one)
- Understand what are the different Triggers for Lambda functions

LAMBDA SERVERLESS WEB PAGE EXAMPLE



SERVER-LESS SUMMARY

- Remember the difference between
  - Traditional architecture
    - Manual construction or automatic (via AWS Cloudformation or AWS Elastic Beanstalk)
    - Has limitations for scalability
    - Could be costly as you scale out
  - Server-less architecture
    - Highly scalable
    - Highly available
    - Lowest cost
- Lambda Function Benefits
  - Scale out not up
  - Are independent… 1 event = 1 function
  - Functions are serverless
  - Know what services trigger (or not) Lambda functions
  - Lambda Functions can trigger other functions
  - Can do things globally within the AWS infrastructure and services (i.e. use it to back up S3 buckets)
- Lambda Function drawbacks/tradeoffs
  - Architectures can get really complex and hard to debug (Mitigation: AWS X-Ray service)