AWS

Northern Virginia is the first region → new features come first here, redundancy a bit worse(might go down for maintenance)

---------- IAM & S3----------

Granular permission setting, identity federation(which means using the same account credentials for google to access AWS), password rotation, PCI DSS compliance for credit cards

Users, Groups, Policies(JSON file), Roles
MFA → Multi Factor Authentication Google Authenticator
Everything is created global here.
Programmatic Access for using AWS CLI or Python scripts
* → Wildcard star allows anything, allows the rest.
Can also set password policies here(upper,lowercase etc.)

Roles are for services to talk to each other → like connecting an EC2 to a Bucket

New users have no permissions at first.

Billing Alarm

CloudWatch: Billing → Create Alarm Static or Anomaly Case → SNS Topic(Simple Notification) → put email → enter Alarm name etc. → check email and subscribe to this topic → done.

Simple Storage Service S3

Safe, Object Storage(movies,files,pictures etc. not code!), 0 - 5TBs storage

Names must be unique globally. → https://bucketName.s3.amazonaws.com (it creates a web address DNS name with the bucket name) → 200 if successfully uploaded the files.

Object → Key : Value : VersionId : Metadata(comments) : Access Control List & Torrents

Power User ---> can access all the AWS services except the management of groups and users within an IAM

S3 buckets have put limits that are changeable from 100 to 1000s.
S3 1 zone IA is cheaper than only IA
1 zone IA ---> availability 99.50
Multipart upload API ---> needs to be turned on for large projects.

100 ---> default S3 bucket numbers per account
s3 has read after write consistency for new put objects.

S3 Virtual Hosted Style URL ---> https://my-bucket.s3.us-west-2.amazonaws.com/fastpuppy.csv
Virtual style puts your bucket name 1st, s3 2nd, and the region 3rd. Path style puts s3 1st and your bucket as a sub domain. Legacy Global endpoint has no region. S3 static hosting can be your own domain or your bucket name 1st, s3-website 2nd, followed by the region. AWS are in the process of phasing out Path style, and support for Legacy Global Endpoint format is limited and discouraged. However it is still useful to be able to recognize them should they show up in logs.
https://docs.aws.amazon.com/AmazonS3/latest/dev/VirtualHosting.html

S3 Data Consistency
Read after Write for PUTs of new objects
eventual consistency for overwrite PUTs and DELETEs(takes time)

S3 consistency is for ---> Overwriting Put operations and delete operations

Tiered Storage Classes

1- S3 Standard - 99.9 availability, durability
2- S3 Infrequent Access → infrequent access lower fee but there is a retrieval fee
3- S3 One Zone Infrequent Access(RRS) → even lower cost infrequent access in one zone only
4- S3 Intelligent Tiering → users ML to change storage tiers
5- Glacier → Data Archive minutes to hours retrieval
6- Glacier Deep Archive → Lowest cost but 12 hours retrieval

Lifecycle Management → change tiers
Encryption
MFA Delete Gate for dummies
Access Control Lists and Bucket policies
Transfer Acceleration → uses CloudFront to accelerate file transfer with a fee.
Not a Virtual Machine or suitable for hosting a database.
You can change tiers for individual objects.

S3 Pricing Tiers

S3 Standard → 0.0023 per GB
S3 Intelligent Tiering → it is better if you do not have many objects.
Assess the needs for the access.
S3 Glacier Deep Archive → 0.00009 $ per GB

Bucket Security

a) Encryption in Transit → HTTPS, SSL/TLS
b) Encryption at Rest → Server Side(Amazon encrypts it) or Client Side
   1 - S3 Managed Keys SSE - S3
   2- AWS Managed Keys SSE - KMS (Key Management Service)
   3- Server Side Encryption Keys provided by the customer SSE - C

Versioning with S3

Once enabled can not be disabled, only suspended.
Backup all the versions of an object.
MFA Delete capability to prevent accidental deleting.
Permissions change back to private with new versions.
Delete Marker is a new version.

Lifecycle Rules

Next to permissions → click Management → add lifecycle rule
You can manage storage tiers and set expiration rules for objects.
Manual Intelligent Tiering .
Can be used with versioning.

AWS Organization and Consolidated Billing

More that you use less that you pay Principle, bill all the other sub organizations in the company in a single bill to negotiate billing discounts with Amazon. → Volume Pricing Discount
Aggregate accounts, user groups and also change policies for Organizational Units.
Always use an MFA, Paying account should only be for billing.

Sharing S3 Buckets across accounts

1- Use bucket policies and IAM. Programmatic Access Only Entire Bucket

2- Object access, Programmatic Access Only

3- Cross-Account IAM Roles works for both Programmatic and AWS Console.

Create Role → give policies → give name id etc. → give the link to another user.--> users can now switch roles.

Cross - Region Replication

Management → Replication

Delete markers will not be replicated.

Past files will not be replicated, only new ones.

Versioning must be enabled on both buckets

Transfer Acceleration(CloudFront Edge Network)

Use a distinct url to first upload to Edge Network and then transfer that to a bucket.

From sydney to london for example it makes sense to upload to the edge locations first.

CloudFront (Amazons' CDN)

CDN: is used to distribute web content to users using distributed servers.

If you do not have CDN the user will pull the content all the way from your original server.

User → Query Edge Location → cache the content for 48 days for all users. → Dynamics, Static, Streaming content. (both for web content and media streaming)

You can also write to the edge location.

TTL → Time to live

You can not clear cached content without a fee.

1 - RTMP → Media Streaming

2- Web Distribution → Web Content

Signed URLs for paying users only.

Create Invalidation → take content off of CloudFront, a fee is paid

Snowball

Physical disk for data transfer of petabytes cheaper 1/5th 50 or 80 TB or 100TB(Snowball Edge → Compute and Storage)

Snowmobile → 100 PB Truck

It just imports large amounts of data to S3.

Storage Gateway

Connect your metal database on your datacenter to AWS datacenter to migrate data.

1 - File Gateway → migrated files are S3 objects now

2- Volume Gateway → Storing Virtual hard disc drives in AWS up to 16 TB for stored volumes, cache volumes only store most frequently accessed dataset not the entire datasets.

3- Tape Gateway → get rid of existing tapes migrate to AWS

Athena vs. Macie

Athena : use SQL to query S3 files. Per query/ per TB scanned fee.

No need to do ETL it is serverless

Can be used to generate reports and run reports on click streams data.(like bids)

Macie: It uses NLP and ML to find Personal Identification sensitive data in S3, good for PCI-DSS(credit card data regulation) compliance and against identity theft. Has dashboards, reporting and alarms etc.

Summary

"Least User Privilege"

----------EC2----------

Little web server that are scalable

Pricing Structures

1- On Demand: fixed rate by hour no commitment good for development
2- Reserved for 1 -3 years: Standard of Convertible Reserved Instances or Scheduled Reserved
3- Spot Instances: like a stock market for instances
4- Dedicated Hosts: physically reserved servers. Used for regulatory purposes like for Gov or Oracle

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

Types of EC2 Instances

Choose Amazon Machine Image → Choose instance type → do config → choose storage → Security Group how you enable ports and http flow (::/0 → open to whole world) → create key Public/Private Key pair.

Connect to open terminal
ssh chrome extension for google chrome if you use windows
apt-get == yum
if you install apache web server you will be able to serve html.index files
IOPS → Input Outputs per seconds
You can encrypt the root device now → comes up in exams
always turn on termination protection
EBS(Elastic Block Store) backed instances when turned off the root will be deleted.
You can encrypt unencrypted volumes.

Security Groups

everytime you change rules it will go into production immediately
you can manage your exposure of ports from here(Inbound and Outbound rules)
if you allow one form of traffic in it will also allow outbound → Stateful
You can't block individual ports and ip addresses from here, everything is blocked by default
you allow one by one 0.0.0.0 or ::/0 allows all ips (block from Network Access Control Lists)
At first all inbound is blocked but all outbound is allowed.
You can't set deny rules.

EBS Elastic Block Store

Basically a hard disk drive for EC2 instances.

1. General Purpose SSD(gp2) -->16000 IOPS most work loads "gp2"
2. Provisioned IOPS SSD → For DBs 64000 IOPS "io1"
3. Throughput Optimised Hard Disk Drive → Big Data and Data Warehouses "st1" 500 IOPS
4. Cold Hard Disk Drive → File Servers "sc1" low cost IA 250
5. Magnetic Drive → Infrequent access "Standard" for archiving if you do not use Glacier.

EBS Volumes & Snapshots

Volume will be in the same zone as the ec2 instance CPU and Hard disk being close to each other.
You can resize the storage volumes. You can promote to a better volume on the fly!
If you want to move to another availability zone → Create a snapshot of your volumes. → launch in another by making an image of that snapshot.
Volumes exist on EBS which is the virtual hard disk and snapshots exist on S3, snapshots are incremental point in time copies.
You can create AMI from snapshots to launch a new instance to another region.

AMI Types (Amazon Machine Image)

EBS vs Instance Store(ephemeral storage)
AMI is like the .ova file in virtualbox but provided
EBS → EBS backed Amazon's own volumes. you can tell AWS to not delete root by default when terminating
Instance Store → ephemeral created from the snapshots in the S3i you will lose all your data if host fails
There are also community AMIs

ENI(Elastic Network Interface), EN(Enhanced Networking) vs EFA(Elastic Fabric Adapter)

These are different network cards.
ENI → Virtual Network Card low budget you can use multiple ones for different scenarios, production, management, test etc.
EN → provides high performance networking capabilities uses Single Root I/O virtualization SR-IOV no additional charges increasing I/O performance choose ENA for higher performance 100 Gbps reliable and high throughput
EFA → You can attach this to EC2 to accelerate high performance computing for ML purposes. Only on linux you bypass the OS kernel for fast operations. "OS Bypass"

Encrypted Root Device Volumes & Snapshots

You can encrypt from the start(comes up on exams)
You can also encrypt non secure instances formerly created.
create snapshot → copy and encrypt snapshot→ create AMI → launch new EC2 instance
you can only share snapshots if it is encrypted

CloudWatch

Monitor performance variety of stuff also good for billing alarms

- CPU
- Network
- Disk

Status Check
AWS CloudTrail → records the activities of users like a CCTV big brother not performance! good for auditing
"sshing into something" → good phrase!

AWS CLI

need programmatic access download access key id .csv file
aws s3 ls
aws configure → enter .csv credentials
aws s3 mb s3://bucketName
cd ~ → home directory
cd .aws → hidden credentials file
nano credentials

Roles

You do not have to give access keys for the root
cd .aws file does not exist it is best practice to use roles
roles are easier to manage. you can assign roles after creating instances and they are universal across all regions.

Boot Strap Scripts

Automatisation for running scripts while booting up EC2 instances
#! → shebang needed at the beginning
#!/bin/bash
yum update -y
yum install httpd
service httpd start
chkconfig httpd on
cd /var/www/html
echo " <html> <h1> Hello </h1> </html>"
> index.html
aws s3 mb s3://bucket
aws s3 cp index.html s3://bucket

sudo su → gives root privileges
curl → to get data from the public ip metadata and user data

Elastic File System

Elastic Block Storage is static but Elastic File System is flexible and will grow according to need.
Under S3
you mount the file system onto an EC2 instance.
supports NFSv4 protocol(Network file system)
only pay for the storage used. can scale up to petabytes
read after write consistency

AWS FSx and FSx for Lustre

for Windows native applications, you can migrate to AWS, SQL Web Server, Workspaces etc.
can utilize windows server message block based file services. if SMB then choose FSx.
EFS is for linux only.
Lustre file system is able to process ***hundreds of gigabytes per second*** for high compute intensive workloads like machine learning, financial processing and video processing. You can also use S3 for this.

EC2 Placement Groups

1. Clustered: All the EC2 instances are in a single availability zone, good for low latency and high performance.
2. Spread: Distributed and failsafe and in different availability zones.
3. Partitioned: similar to spread but have multiple instances on the same rack.used for HDFS, HBase and Cassandra Clusters.
   You can also migrate existing instances to the placement groups using CLI.

AWS WAF(Web Application Firewall)

Allows control access tıo your content.
Outermost layer content watchdog, can see query string parameter in an url
A lot more secure than typical firewalls.
You can deny ips and countries(for example embargoes)
You can block sql injection attacks, you can set rules for headers
You can set special regex based rules.
You can block cross-site scripting attacks.
Either this or Network ACLs are used to block malicious traffic.

Quiz

Good job!

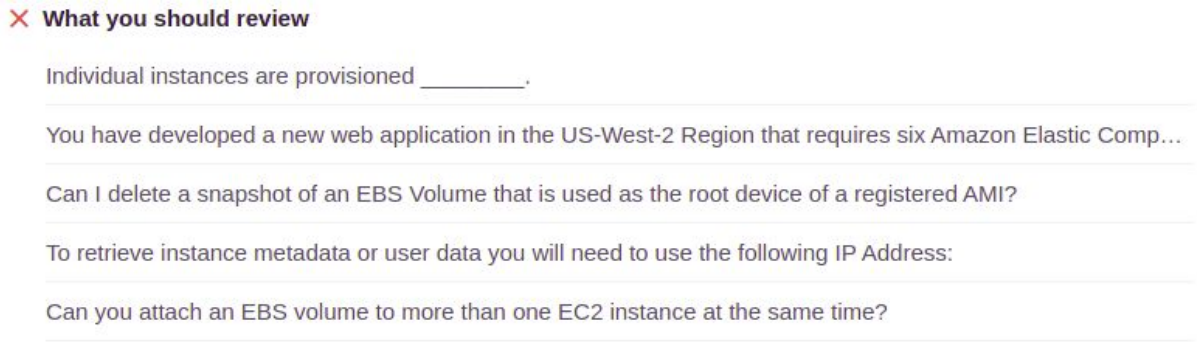
Spread placement groups have a specific limitation that you can only have a maximum of 7 running instances per Availability Zone and therefore this is the only correct option. Deploying instances in a single Availability Zone is unique to Cluster Placement Groups only and therefore is not correct. The last two remaining options are common to all placement group types and so are not specific to Spread Placement Groups.

Good job!

The use of encryption at rest is default requirement for many industry compliance certifications. Using AWS managed keys to provide EBS encryption at rest is a relatively painless and reliable way to protect assets and demonstrate your professionalism in any commercial situation.

Can I delete a snapshot of an EBS Volume that is used as the root device of a registered AMI?
→ NO!

Metadata sadece bu ip adressten alınabilir.



last one is yes

----------DATABASES----------

Types of Relational DBs available in AWS:

1. SQL Server
2. Oracle
3. MySQL Server
4. PostgresSQL
5. Amazon Aurora
6. MariaDB

Multi AZ → for disaster recovery
Read Replicas → for performance

EC2 connects to DBs using a connection string

Non-Relational Databases:
JSON files documents like MongoDB

Data Warehousing
Used for business intelligence. Cognos, Jaspersoft,SQL Server Reporting Services, Oracle Hyperion, SAP NetWeaver.
Used to pull very large and complex dataset usually used by management to query data for performance evaluation et cetera.
Online Transaction Processing vs Online Analytics Processing(OLTP vs OLAP)
OLTP for a single transactions, OLAP for evaluating metrics of a campaign in a region for example uses large amounts of rows.
RedShift → it iis AWS solution for this.
Elasticache → in-memory cache in the cloud scalable. uses fast managed in memory caches to retrieve data faster compared to disk based databases. Cache most used queries from the database. It uses “REDIS” and “Memcached” engines.

RDS Databases → used for OLTP, runs on virtual machines, you can not ssh into them, database patching is amazon’s responsibility, it is not SERVERLESS(except Aurora Serverless)
DynamoDB → NoSQL
RedShift is used for OLAP purposes.
Elasticache is used for speeding up performance of existing databases(frequent identical queries) using Redis and Memcached engines.

RDS back-ups and Multi AZs(Availability Zones)

1. Automated Backups if allowed stored in S3 free of charge
2. Snapshots are done manually
   for both the DNS will change(the endpoint)
   Encryption is supported using AWS KMS

If you have Multi AZ in a failure or maintenance Amazon will change the data flow to backup
Read replica is used to improve performance for reads. Uses Async replication. For performance improvement either you use Read replicas or Elasticache.
Can have read replicas of read replicas(5 is the limit)
Automated backups must be turned on!
You can also promote them to be their own database. → this will break the replication
You can also create an Aurora Read replica
You can change Multi AZ by using failover reboot.

DynamoDB
Both document and key-value data models are available, low latency highly performant
Great for mobile,web,gaming,ad-tech and IoT

- stored on SSD storage
- spread across 3 distinct data centres
- Eventual Consistent Reads by default → takes one second to sync
- Strongly Consistent Read → if you do msec operations you use this.

RedShift

from cents to 1000 dollars for petabytes per year, most economic data warehousing solution

- Good for OLAP operations like my weekly reporting for Vallteri
- Single Node
- Multi-Node → Leader and Compute Nodes x 128
- Similar to Clickhouse columnar database
- Massive Parallel Processing
- by default backups are open, 3 snapshots and 3 copies across multiple AZs, 1 to 35 days retention period
- priced for compute node hours.you will not be charged for leader nodes.
- SSL encrypted
- Only available for 1 AZ.

AuroraDB (not RDS! a hybrid of RDS)

MySQL and PostgreSQL combined Amazon's own DB engine.
Loss tolerant, fault tolerant, 3 to 5 times more performant. 15 Aurora replicas or 5 MySQL replicas.Backups and snapshots are available.
Serverless AuroraDB is used for infrequent, unpredictable or intermittent workloads. → choose this for low cost highly performant and flexible scenarios.
6 copies of data
can share snapshots with AWS accounts.

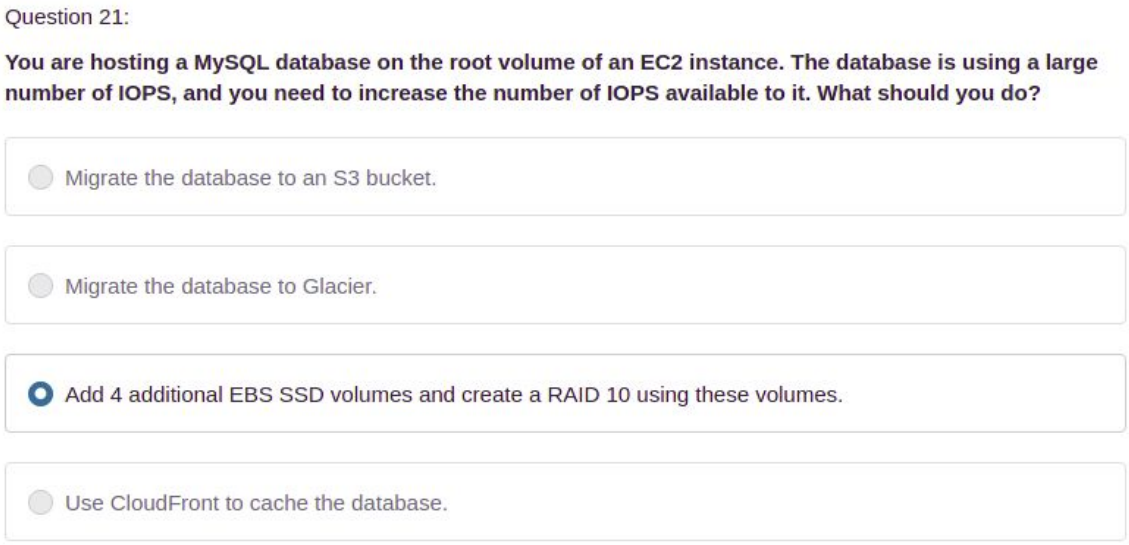
Elasticache

In Memory Caching for fast retrieval of frequently queried data.

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

Redis is more advanced and Memcached is more simple.

QUIZ



to increase IOPS for a database just add new volumes

----------DNS----------

Route 53 → first interstate was route 66 and DNS is on port 53
DNS converts domains names to IPs, IPv4 is running out(2 to the power of 32) so we have IPv6 too
.com → top level domain name
.com.tr → .tr is top level and .com is second level domain name
IANA → controls top level domain names like .gov, .tr et cetera
Each domain name must be unique and it should be registered in a central WhoIS database.
Domain registrars like Amazon, GoDaddy, Netlify etc. assign unique domain names under a top level domain.
NS Record → Name Server Record, top level domain servers use this to lookup the domain from the authoritative DNS records server.
SOA Record → Start of Authority Record, DNS starts from here
A Record → type of DNS record. means Address used to translate the name of the address to the actual ip.
TTL → Time to Live, the length of time the resolving server or the local machine caches the DNS record. the lower the ttl the faster the changes upon DNS propagate through the network. Default is 48 hours, so that can cause problems.
CName → Canonical Name, you can resolve different domain names to the same ip like m.site.com and mobile.site.com → 0.0.0.1. CNames can't be used with naked domain names like example.com, they can only be used with alias and A record names.
Alias Records → you can map one DNS to another like www.ex.com → ex123.amazonaws.com
Exam Tips for DNS:

- ELB's do not have predefined IPv4 addresses, you always get a DNS name to resolve them
- Difference between Alias Record(is used when using a naked domain name or an apex record name) and a CName, always choose Alias Record
- Types of Records: SOA, NS, A, CNames, MX, PTR ?? gotta learn these better

You can register a domain name easily with AWS. It can take 3 days to register.

Routing Policies:

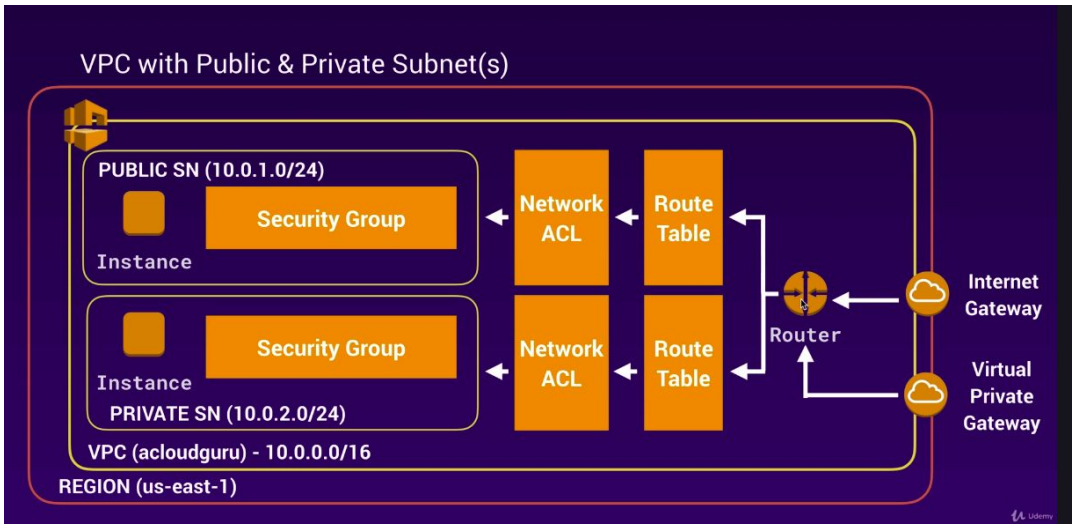
1. Simple Routing: you only have 1 DNS → 2-3 IPs returns randomly the IPs according to the TTL(or you can flush your DNS to see other IPs)
2. Weighted: You can split traffic %90 US-East-1 and %10 US-West-1 for example.You can assign weights to multiple IPs assigned to 1 DNS
3. Latency-based: Routing according to lowest latency region
4. Failover: Active/Passive when primary goes down the passive is activated.
5. Geolocation: You route your incoming DNS queries according to regionality.
6. Geoproximity(Traffic Flow only): Traffic Flow must be used. You can set bias against regions and use specific X and Y coordinates.
7. Multivalue Answer: Exactly same as the simple routing 1 DNS → 2-3 IPs but you can also return only the health check passed IPs.

Health Checks

You can set health checks for EC2 instances to create robust highly available web sites. If it fails it will be removed from Route 53 and you can also trigger an SNS notification. This monitors the health of your endpoints. If you have ephemeral IP addresses you will have to update every time.
#VPN changes the geolocation of your sent DNS queries.
https://docs.aws.amazon.com/Route53/latest/DeveloperGuide/dns-failover-complex-configs.html
The default limit is 50 DNS names but you can increase them by contacting support.
If you have 10 web servers serving the same content the most resilient way to route them is using multi value routing.

----------VPC----------

Virtual Private Cloud → Logical Datacenter in AWS, 1 subnet = 1 availability zone, no transitive peering, security groups are stateful by network acl's are stateless(meaning they can deny ips), a vpc is → Virtual Private Gateways + Route Tables + Network ACLs + Subnets + Security Groups



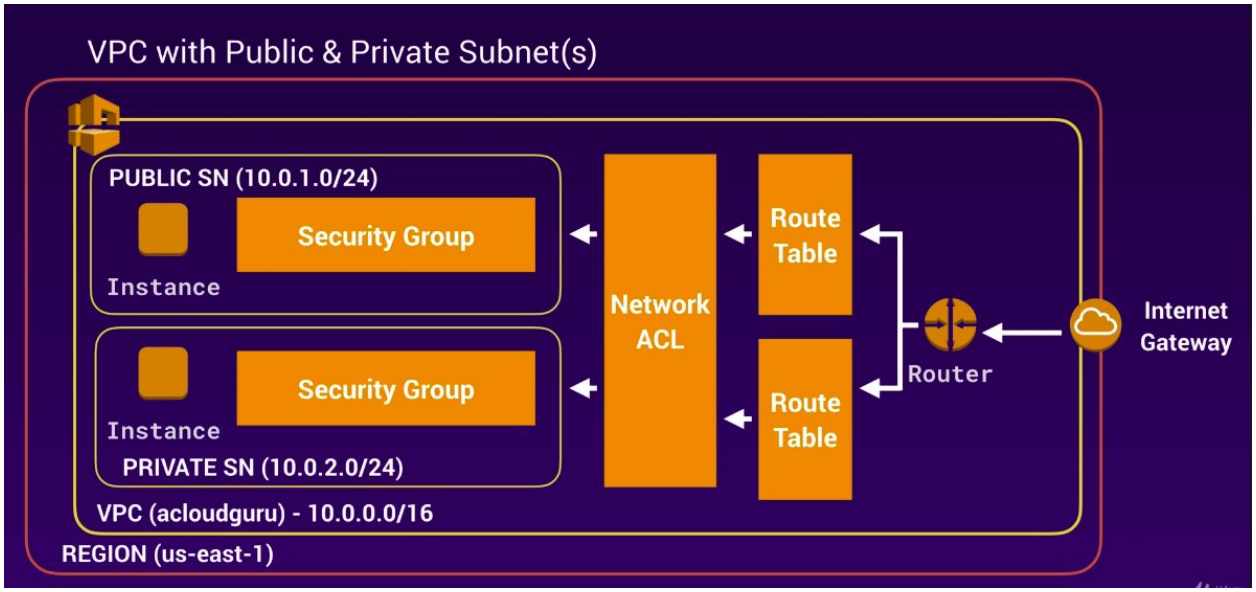
IPs available for only private networks

https://cidr.xyz/ → subnet ip creator
subnet → https://study-ccna.com/subnetting-explained/
In a corporate setting subnets are used to give different networks to different organizational units.
routing table → https://geek-university.com/ccna/routing-table-explained/
peering → a connection between instances in the vpc
YOU SHOULD BE ABLE TO CREATE A CUSTOM VPC TO PASS THE EXAM!

IPv4 CIDR → It is just a scheme replacing old ABC naming convention to 10.0.0.0/16
Tenancy options → dedicated allows you to separate your own VPC to a hardware, costly
You have to create subnets separately after creating VPC, also set one private and set others public manually. Also create an internet gateway. Only one internet gateway per VPC
Also there has to be a public and a private route table. Also we need 2 EC2 instances for the public and private subnet.

- When you create a VPC → a default route table, Network access control list and a default security group is created as well.
- No default subnets are created as well as an internet gateway.

- The availability zones are randomized, Us-east-1a in one user does not equal to the other ones us-east-1a.
- Amazon has reserved 5 IP addresses for subnets
- Security groups cant span VPCs
- 1 internet gateway per VPC



NAT Instances and Gateways

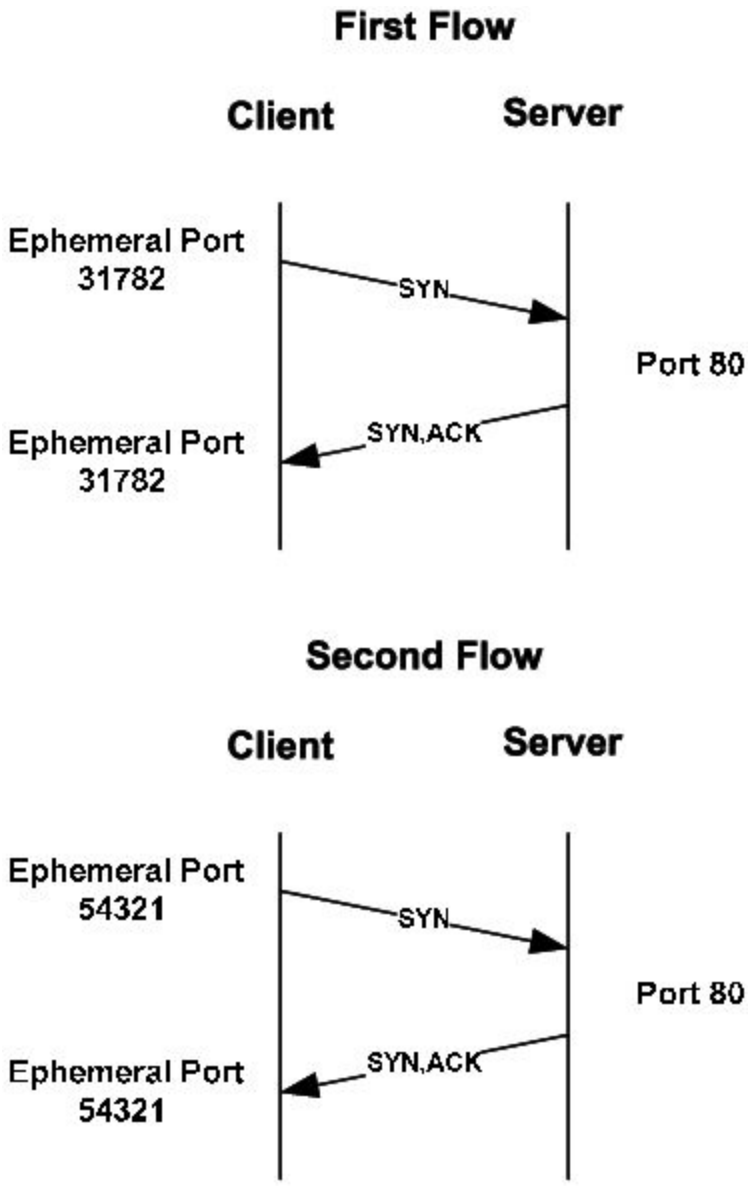
NAT instance is a single EC2 instance to expose the private subnet to the public subnet and NAT gateway is the scalable version NAT instances are deprecated over NAT gateways. Disable source checking. After creating a route in the route table you can now update your private subnet.

0.0.0.0/0 → opens the route to the whole internet.

- When you create a NAT instance disable source destination check
- NAT instances must be in a public subnet.
- There must be a route out of the private subnet.
- Always behind a security group
- NAT gateways are for 1 AZ only. Keep this in mind for high availability architecture so create a NAT gateway for each AZ.

Network ACLs vs Security Groups

ephemeral ports → An **ephemeral port** is a short-lived transport protocol **port** for Internet Protocol (IP) communications. **Ephemeral ports** are allocated automatically from a predefined range by the IP stack software.



Ephemeral ports are short lived and changed after transport.

AWS ports rules are evaluated in increasing order → rule 100 → rule 101 → rule 102

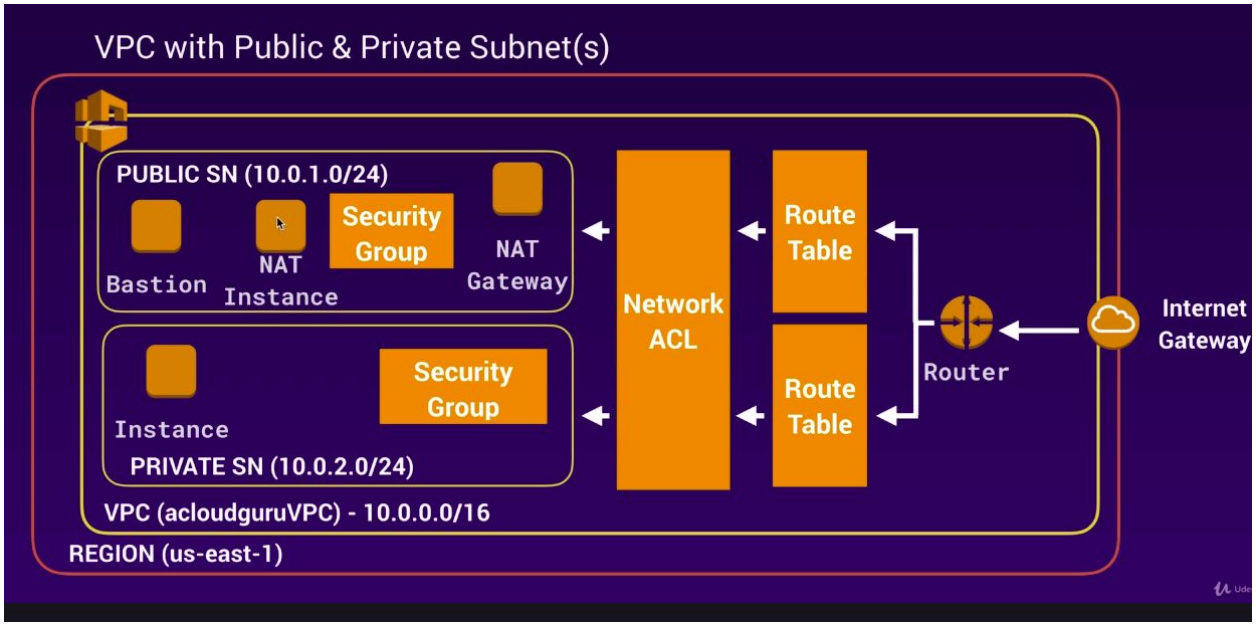
- With ACLs you can set inbound-outbound rules.
- VPC by default allows all traffic, ACL by default denies all traffic.
- Each subnet should have an ACL or it allows everything
- For blocking IP addresses choose ACL instead of Security Groups.
- ACLs are stateless, you have to manually enable/disable inbound outbound rules security groups update themselves.
- When creating a load balancer you need at least two subnets.

VPC FlowLogs(Using CloudWatch)

You can look at the flow log of the network traffic. You can do this for VPCs , subnets and overall traffic and put the logs in an S3 bucket. You cannot do this across accounts you can not change IAM afterwards.

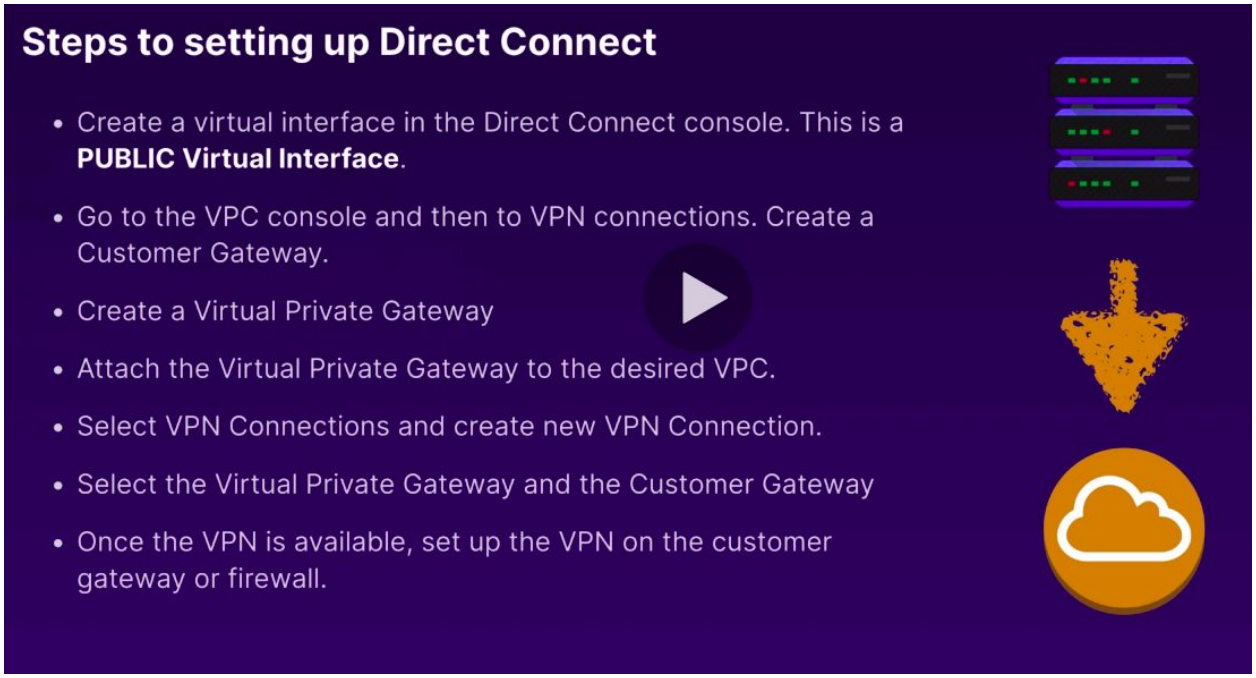
Bastion Host

DMZ → demilitarized zone is a public subnet. When you SSH or RDP into your private subnet instance Bastion is used it is like a jumpbox.



Direct Connect

You connect your office environment or DB directly to AWS. You literally connect it physically. Secure and can handle high throughput.



This has been asked in the exam

Global Accelerators

Basically this chooses the optimal endpoint according to clients needs and endpoint health, optimizes the performance. Accelerator has its own IP & DNS address. Supports TCP and UDP. Network Load Balancers, Application Load Balancers, EC2 instances or Elastic IP addresses are ENDPOINTS. You can group endpoints across same or different AZs and you can assign them weights or let the accelerator to the optimisation of the network flow for you.
Two static IP addresses are assigned to each accelerator.

VPC Endpoints

VPC Endpoints allow traffic flow inside the VPC and allows the services inside VPC to talk with each other without leaving AWS network, no availability and bandwidth constraints inside the VPC.

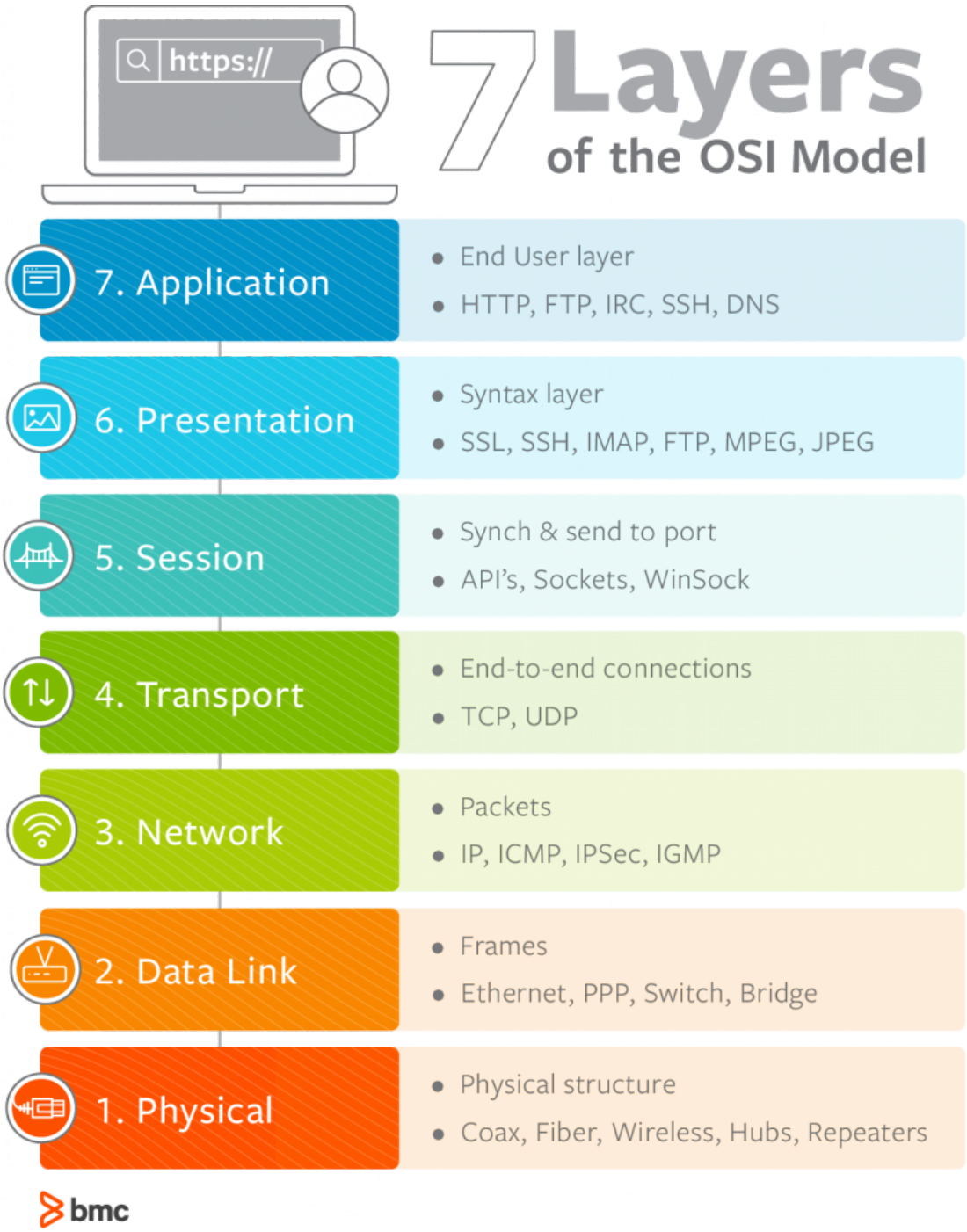
1. Interface Endpoints → allows you to stay in AWS network.
2. Gateway Endpoints → supported for S3 and DynamoDB

Quiz

----------HIGH AVAILABILITY ARCHITECTURE----------

Elastic Load Balancer:

1. **Application Load Balancer:** load balances HTTP/S traffic, OSI Layer 7(application), advanced request routing
2. **Network Load Balancer:** load balances TCP traffic,OSI Layer 4(connection) better than http for latency can handle millions of requests per sec. Used for extreme performance! but they are expensive
3. **Classic Load Balancer:** Legacy, HTTP/S but can also be used for TCP, can do sticky sessions and X-forwarded. Cheaper, when the app stops responding returns 504(gateway timeout) error.(web server or database has errors not the balancer)



OSI LAYERS(Just a refresher)

X-forwarded-For Header: when you use a Load Balancer you might not get the original users IP so you use X-forwarded-for header to get it.

Load Balancers & Health Checks: Cross-Zone Load Balancing and Health checks are best practice. You can create Target Groups to use load balancers for different types of audiences and their request profile and different types of apps.

You can set if else rules for your application load balancers as they are more intelligent than classic load balancers..
Load balancers have their own DNS names and you never give them an IP.

Advanced Load Balancer Theory:

1. **Sticky Sessions:** Allow you to bind a user during the session so that all the requests that are coming from that user are routed to that specific EC2 instance. Can be used directly with the Classic LBs or can be used with Application LBs using Target Group Levels. You can enable or disable them according to the EC2 traffic load.
2. **No Cross Zone Load Balancing:** When you do turn this off you can divide the traffic across cross availability zones.
3. **Path Patterns:** Based on URLs you can divide traffic to different instances like www.x.com/ → EC1 and www.x.com/url → EC2

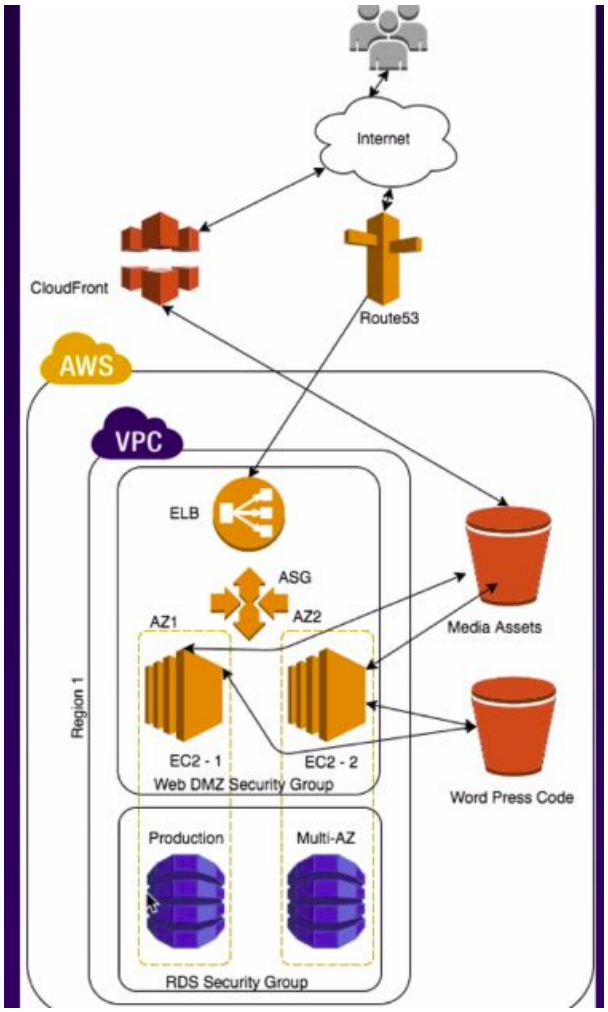
Autoscaling Theory

1. **Groups:** Web Server, Application or Database Group
2. **Configuration Templates:** Basically every group has their bootstrap scripts to create and configure the EC2 instances. Minimum, Maximum # of instances
3. **Scaling Options:**
    a. Maintain current levels at all times
    b. Scale Manually
    c. Scale on a Schedule
    d. Scale on Demand → based on CPU utilization or on loading levels
    e. Use Predictive Scaling

High Availability Architecture

- Plan for failure.
- Chaos engineering → https://netflix.github.io/chaosmonkey/
- Use multiple regions and availability zones
- for database systems RDS the multi AZ makes it secure and read replicas increase performance
- scaling out → use autoscaling
- scaling up → increase performance resources not the number of parallel instances.
- always consider the cost.
- know the different types of storage options for S3.

Real Example Wordpress Site on AWS



you can use cloudcraft.io to create these diagrams

- Your Route53 points to the load balancer
- writer and reader nodes
- crontab to take snapshots
- cloudfront and SNS propagation takes time.
- RDS has a reboot with failover option in which you will still be available in seconds, so still highly available.

CLOUDFORMATION

- has sample stack templates for developers to use like LAMP, MERN, MEAN etc.
- it also has other templates for S3, ELB etc.
- https://aws.amazon.com/quickstart/?quickstart-all.sort-by=item.additionalFields.updateDate&quickstart-all.sort-order=desc

ELASTICBEANSTALK

Elasticbeanstalk is for clueless developers who just want to use their stacks. Similar to Cloudformation.

You upload your code to here and just deploy immediately. You can use auto scaling and all the other goodies of AWS.

QUIZ
if you have distinct web servers with distinct traffic use application load balancer.
bespoke: particular
You need to do regular rereadings of this note as well as the quiz exercises.

----------SERVERLESS - AWS LAMBDA----------

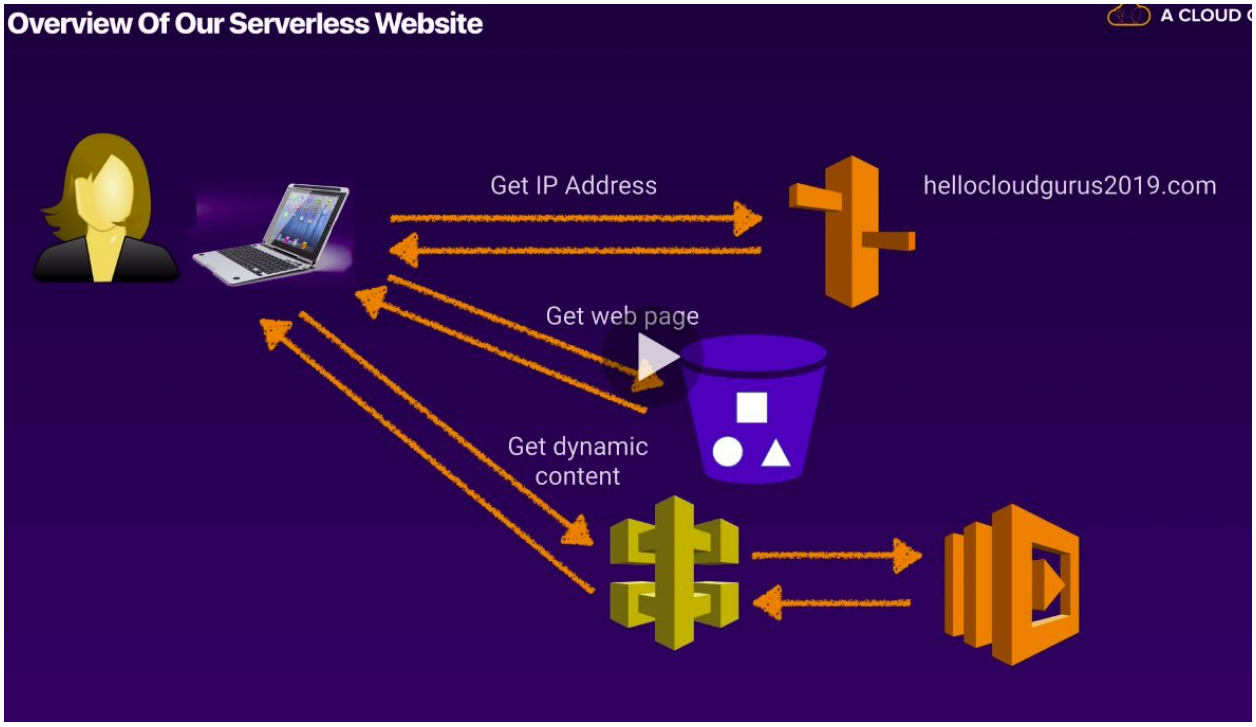
- Alexa uses Lambda

IAAS → Elastic Cloud 2
PASS → Elastic Beanstalk
Containers → Docker
Serverless → Lambda or GCP Functions this is the ultimate abstraction layer. A developer only worries about the code with lambda functions.

1. **Event Driven** → your lambda function can respond to events like S3 bucket updates or DynamoDb table updates.
2. **Compute Service** → run your code as response to http request/api calls using API gateways or AWS SDKs.

Traditional vs Serverless Architecture
Route 53 → ELB → EC2 → RDS
Route 53 → API Gateway → Lambda → DynamoDB or Aurora Serverless
API Gateway scales immediately.
supports: go, python, c#, powershell, Java, Node.js
pricing: 1st million requests free than 0.2 cents per million reqs. And also code runtime allocation.

- Scales out automatically.
- 1 event = 1 function independent functions.
- Lambda functions can trigger other lambda functions.
- X-ray service helps with debugging serverless.



so the static content is handled by bucket and when a dynamic event occurs api gateway relays it to the lambda functions. Also you can open concurrency mode on from the console.

API Gateway Triggers

- AWS IoT
- Load Balancer
- CloudWatch
- DynamoDB
- S3
- and much more

EC2, ECS and Lambda all can do multithreading.

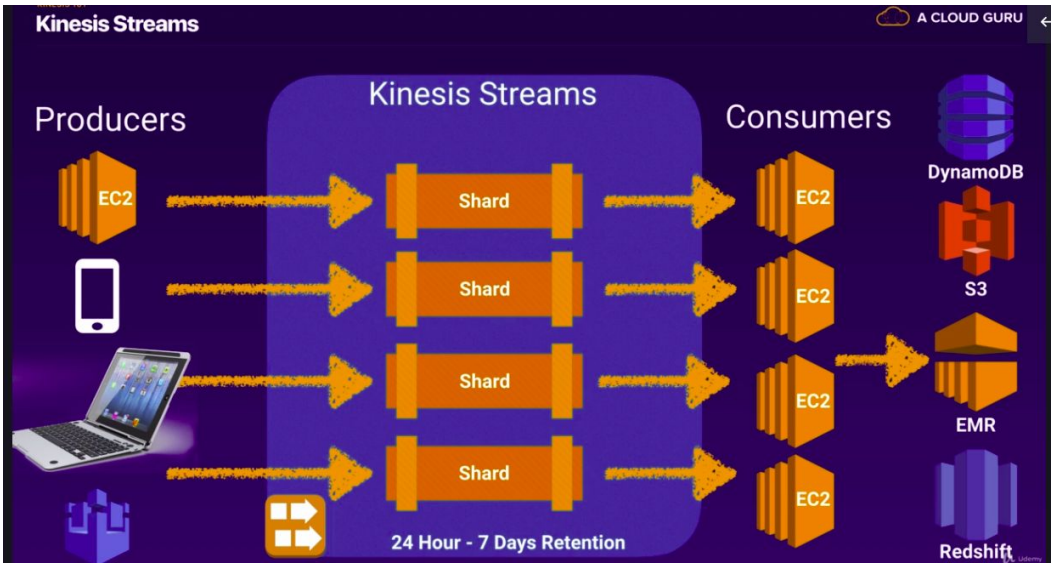
----------Extra Applications----------

- **SQS:** queues messages stores them for processing like a message broker. Allows asynchronicity and robustness, decouples microservices, it is pull based, 256 kBs of text files containing messages.Poll or Pull based.
    a. Standard Queue: at least once processing
    b. FIFO Queue: exactly once processing, no duplicates and orderly

  Visibility timeout is max 12 hours. Long and short polling for your EC2 instances.
- **Simple Workflow Service:** Coordinates tasks in applications, web service calls, scripts, human actions and executable codes. Amazon uses it in warehouses. The maximum workflow period can be 1 year even. Task-oriented API, no duplicate tasks. Automatically tracks tasks.

Workflow Actors(like Actor models in Data streams Apache): Workflow Starters, Deciders, Activity Workers.

- **SNS:** Simple Notification Service: Push notifications to google, apple etc. Also can do email and text messages. Has topics and subscribers. All SNS messages are stored in multiple AZs. No polling instantly pushes messages. Simple API, Cheap.
- **Elastic Transcoder:** Converts media to different formats, by minute transcoding fee, like transcode ios video to android for example.
- **API Gateway:** Create an API endpoint to your AWS applications. ECC or EC2(Elastic Compute Cloud), expose HTTPS endpoints to define a RESTful API, can be used with Serverless functions and DynamoDB, Low Cost, Scales effortlessly, Control usage with API key, throttle API request attacks, can have multiple versions for dev and deployment. Has a very simple console interface.
  API caching allows reducing frequent requests.
  Has same origin policy to block cross-site scripting attacks.
  If origin policy cannot be read enable CORS(cross origin resource) policy.
  You can use CloudWatch to set alarms and look at logs.
- **Kinesis:** Deals with streaming data which is data sent in kilobytes from thousands of data resources almost simultaneously like purchases from a store or like stock prices or game data or social network data or geospatial data or IoT data.
  a. **Kinesis Streams**: Data producers stream to Kinesis from 24 hours to 7 days, shards send data to EC2 instances. Shards have 5 TPS for reads up to 2 Mbps and 1000 TPS for writes up to 1 Mbps. Sum of shards is the total capacity of your stream.



  b. **Kinesis Firehose:** There is no shard storage so you have to use Lambda serverless or EC2 instances to do something with that data immediately.
  c. **Kinesis Analytics:** Helps you analyse Firehose and Streams data on the fly.

- **Cognito(Web Identity Federation):** Basically users use google facebook or amazon credentials to login to your apps. It is an identity broker. Good for mobile applications.

User Pools: Directories to manage sign-in and sign-up. Successful authentication generates a JSON Web Token(JWT). You can use **Identity Pools** to give access to AWS resources like S3 et cetera.

Cognito also uses push notifications to all their devices if the credentials change for the user. This allows synchronization.