**Notes for Cloud Practitioner**

**General knowledge**

Multi tenancy - Multiple virtual machines being able to run on the same underlying hardware

Scalability – Ability of a system to accommodate a larger workload

Elasticity – The ability for a scalable system to do auto-scaling based on the load

Agility – Faster access to IT resources

AWS has multiple **Regions,** each containing 2-6 **Availability Zones**, each of which contains data centres

There are also **Edge locations** which are used for content delivery and to provide lower latency for users in countries without AWS regions

You would choose a region based on:

* Compliance with regulations
* Proximity and latency
* Feature availability
* Pricing

There are three ways to connect with AWS:

* Management console – online, point and click
* CLI – API calls through a terminal
* SDK – Software development kit inside a programming language (boto3 in Python)

AWS has a shared responsibility model

* Security of the cloud, hardware – Responsibility of AWS
* Security in the cloud, application updates, software – Responsibility of customer

Pillars of a well architected cloud framework:

* Operational excellence
* Performance efficiency
* Security
* Cost optimisation
* Reliability
* Sustainability

AWS Well-Architected Tool can help you review your infrastructure against these pillars

Characteristics of cloud computing:

* On-demand self service
* Broad network access
* Multi tenancy and resource pooling
* Rapid elasticity and scalability
* Measured service, pay as you go

Advantages of cloud computing:

* Capital ex for operational ex
* Benefit from massive economies of scale
* Stop guessing capacity and workload
* Speed and agility
* No need for data centres
* Go global in minutes

**EC2**

EC2 provides on demand computing capacity

EC2 Instance types:

* General purpose
* Memory optimised – for memory intensive apps
* Storage optimised – for working with large volumes of data
* Compute optimised – for computationally expensive apps

EC2 pricing:

* On demand
* Savings plans – lower cost
* Reserved instances – reserve for 1 or 3 years for lower cost
* Spot instances – discount for unused instances by AWS
* Dedicated instances – private instance for compliance

EC2 has auto-scaling functionality

* Predictive scaling – Predicts the workload and scales appropriately
* Dynamic scaling – Scale in accordance with workload

EC2 has security groups to control network traffic. They only have allow rules and are stateful, as they remember past packets that have entered. By default, they block all inbound traffic and allow all outbound traffic.

**Elastic Block Store (EBS)** allows you to attach data storage volumes to EC2 instances so that they can retain data. This lets you avoid ephemeral (temporary) local storage that is lost on instance termination. EBS volumes can also be set to be deleted on instance termination. EBS volumes are bound to Availability Zones.

You can take snapshots of EBS volumes which records the data at the time of the snapshot. This snapshot is available region wide.

**Elastic File System (EFS)** is another file system that can be connected to EC2 instances. This system can have multiple instances reading from it and is a region level resource. This automatically scales with the storage needed and has a cheaper infrequent access tier.

**EC2 Instance Stores** are also available for high performance storage. This is ephemeral only, and useful for caching and temporary data.

You can create **Amazon Machine Images (AMIs)** for EC2 instances that allow you to customise the software and OS of an instance, and then build instances from that image. This provides much faster bootup times

With **EC2 Image Builder,** you can automate the creation of AMIs, where they can be maintained and tested within a pipeline

**FSx** is a 3rd party file management system, which is fully managed. Available for:

* Windows File Server
* Lustre – for high performance computing

**Elastic Load Balancer** manages workload by splitting it up across multiple instances and can react to failing and overloaded instances by diverting the workload to healthy instances. This provides greater fault tolerance. There are 3 kinds of load balancers

* Application load balancer – for http/https
* Network Load balancer – for **ultra-high performance** and millions of connections
* Classic Load Balancer – old generation, retired

**Auto scaling groups** help to scale instances to match the workload, ensure we have a minimum and maximum number of instances and replace unhealthy instances

* You can apply dynamic scaling to respond to changing demand, such as with a CloudWatch alarm monitoring CPU usage
* You can also apply predictive scaling which uses ML to predict future traffic

Vertical scaling is when you increase an instance size, such as from t2.micro to t2.large, and horizontal scaling is adding more instances to handle the workload

**IAM**

**IAM** allows you to create additional **users** and give them the permissions they need. This has global scope and by default, users have no permissions. The best practise is to only give users the permissions they need (**least privileged principle**). They can also be put into **groups**, where the group can be configured to grant the users specific permissions via policies.

IAM has **Multi-factor Authentication (MFA**) which provides additional security to accounts. This is important to add for the **root user** which has permission to perform any action. You generally want to avoid using the root user account, create an admin account with the root user and switch to that instead.

IAM can also have **roles**, which are permissions that can be applied to services such as EC2 or Lambda.

**S3**

**Simple Storage Service (S3)** allows scalable object based storage with very high durability. There are different tiers:

* Standard for standard access
* Infrequent access, suitable for backups and long term storage
* S3 Glacier, for data archival purposes
* Intelligent tiering, which sorts objects automatically

Every file is stored as a single object. Objects can be set to be moved between different storage types with **lifecycle rules**.

S3 buckets are region bound and are private by default.

S3 can be used to host static websites.

IAM policies and bucket policies can be applied to restrict access to S3 buckets

You can perform cross region and same region replication to reuse the objects in a S3 bucket.

You can add versioning to track updates to objects and roll them back to previous versions.

S3 offers different amounts of encryption:

* No encryption
* Server-side encryption – file is encrypted after it is received by the server
* Client-side – User encrypts the file before uploading it

**AWS Snow Family** devices for data migration into or out of S3:

* Snowcone – Small portable computing device, used where Snowball Edge is too large for the space. Online and offline, less than 20 petabytes
* Snowball – Larger device for higher capacity transfer. 50 petabytes +
* Snowmobile – A truck used to transfer huge volumes of data, exabytes

These are physical, offline devices used to perform data migration. This is ideal for when there is:

* Limited connectivity
* Limited Bandwidth
* High network cost
* Poor connection stability
* Shared bandwidth

The snow family is useful for edge locations which have limited internet access.

**OpsHub** is a piece of software used to manage snow family devices.

Hybrid cloud – infrastructure that is part on-premises and part on the cloud

**AWS Storage Gateway** – Gives on premises data centres access to cloud storage on S3 (hybrid cloud)

**RDS**

AWS offers database services for relational and NoSQL databases

Benefits of AWS databases:

* Quick provisioning, high availability, vertical and horizontal scaling
* Automated backup and restore
* Operating system is patched by AWS
* Monitoring and alerts

If you were to deploy a database on EC2, it would not be managed by AWS, and you would miss out on additional benefits that RDS provides

Aurora is a proprietary technology from AWS

Supports PostgreSQL and MySQL

Aurora is cloud optimised, so it is much faster and more efficient and MySQL and PostgreSQL engines

You can create read replicas to scale to the read workload of your database

You can have it available in multiple availability zones, so you trigger failover in case of AZ outage, where the database will be replicated to another AZ.

You can also have multi-region deployments for disaster recovery

**ElastiCache** provides an in-memory database with high performance and low latency, helps reduce workload for reads

**DynamoDB** is a managed service for NoSQL databases

It is a serverless database, so we do not provision servers and it scales to massive workloads, millions of request per seconds with very low latency

Has an infrequent access tier

It’s a key value database

DynamoDB has global tables so that it can still have low-latency across multiple regions

DynamoDB Accelerator (DAX) is an in-memory cache specifically for DynamoDB

Up to a 10x performance improvement

DAX is only for DynamoDB, ElastiCache can be used for most database types

**Redshift** is a data warehousing service using for **online analytical processing (OLAP)**. Not used for transaction processing OLTP

Columnar storage of data, as opposed to row based

Has a SQL interface to make queries

**Amazon EMR** (elastic map reduce)

Used to create Hadoop clusters, for big data

Clusters can be made of hundreds of EC2 instances

**Amazon Athena** is a serverless query service to perform analytics on S3 objects

Uses SQL for queries

Amazon QuickSight is a serverless service used to create dashboards for business intelligence

DocumentDB is the equivalent of Aurora for MongoDB, a NoSQL database

Fully managed, highly available and scales storage automatically

Amazon Neptune is a fully managed graph database

An example of a graph dataset is a social network, so highly connected datasets

Can store billions of relations and can query graphs with low latency

QLDB – Quantum ledger database

Used to recording financial transactions and review history of all changes made to application database

Immutable – no entry can be removed or modified, verified cryptographically

Different with Amazon Managed Blockchain is that QLDB is not decentralised

Amazon Managed Blockchain allows you to build applications with multiple parties making transactions, without a central authority

Can join public networks or make your own private one

Compatible with Hyperledger Fabric and Ethereum

AWS Glue is a managed extract, transform and load (ETL) service

Used to prepare data for analytics

DMS – Database Migration Service, migrate databases securely to AWS

Support homogenous or heterogenous migrations (same or different type of database engines)

**Other compute services**

Docker is used to package apps into containers so that they can run on any operating system

ECS – Elastic container service, can be used to launch containers on AWS

You need to provision and maintain the EC2 instances

Fargate is also used to launch containers, but it is serverless

You do not need to provision any instances

ECR – Elastic container registry, used to store container images

AWS Lambda – serverless virtual functions intended for short run times

Runs on demand, pay as you use

Can use CloudWatch Events to schedule a lambda function

Amazon API Gateway – service for building a serverless API

Supports security, user authentication, API throttling, keys and monitoring

AWS Batch – Fully managed batch processing at any scale

Batch job is a job with a start and an end, not continuous

Provisions the right amount of the resource you need for you batch job

Batch jobs are defined as Docker images

Lightsail – Gives you access to virtual servers, databases and networking

Lower cost

Simpler alternative to EC2, RDS and more

For people with less cloud experience

No auto-scaling, and limited AWS integrations

**Deployments and Infrastructure**

CloudFormation is a declarative way of outlining AWS Infrastructure

Allows you to provision infrastructure as code rather than manually

Costs can be seen easily by looking at the CloudFormation template

Can reuse existing templates and templates other people have created

AWS Cloud Development Kit (CDK) is a way to define cloud infrastructure using a programming language

The code is compiled into a YAML to be used with CloudFormation

Beanstalk – developer centric way of deploying apps on AWS

Uses many different services, EC2, ASG, RDS

Developers are only responsible for the application code

CodeDeploy – Deploy code for you application automatically

Works with EC2 instances

CodeCommit – A repository for code, using git-based repositories

Works with git tools

CodeBuild – Code building service in the cloud

Compile source code, run tests and produce packages

Fully managed and serverless, secure

CodePipeline – Continuous delivery service for each commit

CI/CD pipeline

Fast delivery and rapid updates

CodeArtifact – Software packages have code dependencies

Storing and retrieves these dependencies is called artifact management

CodeArtifact can manage these dependencies for you

CodeStar – Unified UI to manage software development activities in one place

Helps you implement CI/CD practices

Cloud9 – is a cloud IDE, writing, running, and debugging code

It’s in the browser, so you can write code anywhere

Allows for collaboration

AWS Systems Manager (SSM) – Helps you manage your EC2 and on-premises systems

Hybrid cloud service

Patching automation for compliance

Run commands for fleet of servers

Store configs with the parameter store

SSM Session manager – Gives you access to an EC2 instance directly via SSH. SSH does not need to be enabled in the security group but the EC2 does need the correct IAM role.

OpsWork – gives you managed Chef & Puppet in the cloud, which are applications that help you perform server configuration and repetitive actions.

**Global Infrastructure**

**Route 53** is a managed Domain name system (DNS)

DNS directs clients to servers through URLS

Routing policies:

* Simple routing policy – no health checks, directs you towards the server
* Weight routing policy – will direct traffic according to assigned server weighting, effectively load balancing
* Latency routing policy – Servers are routed to users based on proximity/latency to the user
* Failover routing policy – Route 53 will do health checks, and redirect to healthy instances if one fails (failover)

CloudFront – content delivery network that utilises edge locations to cache and deliver content with low latency. This helps improve user experience globally

Good for static content that must be available anywhere.

Cloudfront integrates with WAF and Shield to protect from DDoS attacks.

S3 Transfer Acceleration – increases transfer speed from a client to a S3 bucket by using an edge location in the middle for a faster transfer speed

AWS Global accelerator – Improves performance and availability of global application by using the AWS internal network to optimise the route to your application. Different from CloudFront as it does not cache content

AWS Outposts – Server racks that let you host AWS infrastructure within on-premises systems

Gives you greater control over the security and residency of your cloud data

AWS Wavelength – Infrastructure deployments at the edge of 5G telecommunication networks

Lets you serve applications through 5G networks, giving users an ultra-low latency

AWS Local Zones – Place compute, storage and database applications closer to end users, for example extend to Chicago from us-east-1 in North Virginia

CloudFront is different because it’s globally available

Different types of architecture

(Difficulty to implement increases as you go down):

* Single-region – Architecture is only a single region. Can be in one AZ or multi-AZ for higher availability
* Multi-region – Architecture extends to multiple regions. More globally available
* Multi-region Active-passive – Some instances in the architecture are active so they can be written to, some are passive so can only be read from. Much better global read latency
* Multi-region Active-Active – All instances can be read from or written to globally, very good read and write latency

**Cloud Integrations**

AWS provides services to let applications communicate with each other

SQS – Simple Queue Service, producers can send messages to a queue, which can be read from through polls from consumers

Used to help decouple applications and keep them separate by communicating through SQS

Kinesis – Real time big data streaming

Used to manage, collect, and process and analyse real-time streaming data

SNS – Secure Notification Service

Publishers can send messages to a SNS topic

Multiple subscribers can subscribe to the SNS topic to receive the message

MQ – managed message broker service for RabbitMQ and ActiveMQ

SQS and SNS are proprietary protocols, created by AWS

If you were migrating an app to the cloud, it may use other protocols such as MQTT, AMQP, etc. Instead of re-engineering to use SNS/SQS, you would use MQ

CloudWatch Metrics – enables monitoring and access to metrics of applications

Can configure alarms to trigger notifications for any metric

CloudWatch Logs can collect logs from a variety of different services

This allows for real-time monitoring of your logs, and the retention can be monitored

EventBridge (formerly CloudWatch Events)

Can be used to schedule cron jobs, or to form an event pattern so it reacts to a service doing something

CloudTrail – Shows history of events and API calls made within your AWS account

All activity is saved into log files

AWS X-Ray – Gives you visual analysis of applications and can show you where they are failing, used for debugging and troubleshooting performance

Understand dependencies and find errors

Amazon CodeGuru – ML powered service for automated code reviews and performance recommendations

CodeGuru profiler helps you understand the runtime behaviour of your application

Service Health Dashboard – shows you if AWS services are functional in every region

Personal Health Dashboard – provides alerts and guidance when AWS is experiencing events that may impact you

Shows health of AWS services relevant to you

**VPC / Networking**

**VPC** – Virtual private cloud, private network to deploy your resources

Region-bound

VPC can be partitioned into **subnets**, which is AZ-bound

The public subnet is accessible from the internet

Private subnet is not accessible from the internet

**Internet Gateways** helps the VPC instances connect with the internet

Public subnets are connected to the internet gateway

**NAT gateways** (AWS managed) and **NAT instances** (self-managed) allow private subnets to connect to the internet privately

**Network ACL** is a firewall that controls traffic to a VPC subnet

Can have allow and deny rules for certain IP addresses

Stateless, does not remember previous traffic

Allows all traffic by default

**Security groups** are also a firewall that controls traffic to an EC2 instance

Only allow rules

Can be IP addresses or other security groups

Stateful, remembers entering traffic

Denies inbound but allows outbound by default

**VPC Flow** logs captures information about IP traffic going into your interfaces

Monitor connectivity issues

**VPC Peering** lets you connect two VPCs together through AWS

Behave as part of the same network after peering

VPC Endpoints lets you connect to AWS services over a private network rather than publicly

* **Endpoint Gateway** lets you connect to S3 and DynamoDB
* **Endpoint Interface** is used for other services

AWS **PrivateLink** – Allows you connect a service within a VPC to other VPCs

Requires a **network load balancer** and **elastic network interface**

To connect an on-premises data center to a VPC, you can use:

* **Site to site VPN**, connect an on-premises VPN to AWS
* **Direct Connect,** between on-premises and AWS

For a Site to site VPN, the on-premises data centre requires a **customer gateway** and the VPC requires a **virtual private gateway**

**AWS ClientVPN** – Lets you connect from your computer to a private network on AWS

Can also connect to on-premises if site to site VPN is implemented

**Transit Gateway** – transitive peering between thousands of VPCs and connections between all other types of other connections

**Security and Compliance**

DDoS – distributed denial of service, servers are overwhelmed by requests

To protect from DDoS attacks, you can use:

* AWS Shield standard – protects against attacks for no additional costs
* AWS Shield Advanced – premium protection against sophisticated attacks with access to a response team
* AWS WAF – Filter requests based on rules
* CloudFront and Route 53 – higher availability with edge network
* Using auto scaling to scale to a higher number of requests

You can perform penetration testing for certain types of security threats on the cloud

Two types of encryption, encryption at rest and encryption in transit

AWS KMS – Key management service, encrypted keys managed by AWS

CloudHSM – Like KMS but you manage the encryption keys, AWS manages hardware only

Certificate Manager – Lets you provision, managed and deploy SSL or TLS certificates, encryption for websites

Secrets Manager – For storing secret keys, capability to force rotation of keys

Keys are encrypted using KMS

Integrates with RDS

Artifact – Gives customers access to AWS compliance docs and agreements

GuardDuty – Performs intelligent threat discovery to protect your AWS account

Uses ML algorithms to detect anomalies and give you notifications

Works across multiple accounts

Can protect against cryptocurrency attacks

Inspector – Automated security assessments, helps improve security and compliance, only for EC2 instances and container infrastructure

Config – Helps with auditing and recording compliance by recording configs and changes over time

Macie – Uses ML and pattern matching to alert you to sensitive data such as Personal identifiable information in AWS

Security Hub – Central security tool to manage security across several accounts

And automate security checks

Detective – Helps you investigate security findings and find the root cause of security issues

Root user has permission to do everything, including:

* Change account settings
* Close AWS account
* Change AWS support plan
* Register as a seller in the reserved instance marketplace

**Machine Learning**

Rekognition – Identify object, text and scenes in images and videos using ML

Transcribe – Automatically convert speech into text

Uses deep learning process called ASR

Polly – Convert text into speech

Translate – Natural and accurate language translation

Lex – Automatic speech recognition to understand natural language, used for chatbots

Amazon Connect – Virtual contact center

Comprehend – Used for natural language processing in text

SageMaker – Service to build ML models, train and tune it based on given data

ML Model can be deployed on new data once it is built

Forecast – Uses ML to deliver highly accurate forecasts

Kendra – Uses ML to search documents to extract answers

Natural language search capabilities

Personalize – ML service to build apps with real-time personalised recommendations

Textract – extract text, handwriting and data from scanned documents with ML

**Account management and billing**

AWS Organisations lets you manage multiple AWS accounts

Gives you access to consolidated billing and pricing benefits from aggregated usage

Can use API to automate account creation

Can centrally manage all account permissions with service control policies

Service control policies can whitelist or blacklist IAM actions

AWS Control Tower – Easy way to set up and govern a secure and compliant multi-account AWS environment

Will automatically set up organisations for you

AWS Pricing models:

* Pay as you go – pay for what you use
* Save when you reserve – minimise risks, manage budgets predictably and comply with long-term requirements
* Pay less by using more – volume-based discounts
* Pay less as AWS grows

Free services and free tier

IAM, VPC, Consolidated billing, Elastic Beanstalk, CloudFormation and Auto scaling groups

Savings Plan – commit a certain amount of money per hour for 1 or 3 years

Easiest way to set up long term commitments on AWS

Includes:

* EC2 Savings Plan
* Compute Savings Plan for EC2, Lambda and Fargate and for flexibility in EC2 instance types
* Machine Learning Savings Plan with SageMaker

AWS Compute Optimiser – reduce costs and improve performance by recommending optimal AWS resources for your workloads

Billing and costing tools:

* AWS pricing calculator can help you estimate costs of your infrastructure
* Billing dashboard – shows cost appreciated and forecasts
* Free tier dashboard – usage of free tier services so far

Cost Allocation Tags – used to track AWS costs on a detailed level

Cost and usage reports – most comprehensive data of cost and usage available, including for each account and IAM user

Cost explorer – visualise and understand your costs in real time, forecast up to 12 months ahead

CloudWatch has billing alarms that warn you of spending, But not as powerful as AWS budgets

AWS Budgets - create budgets and send alarms when costs exceed your budget

AWS Trusted Advisor – High level account assessment which provides recommendations based on 5 categories:

* Cost optimisation
* Performance
* Security
* Fault tolerance
* Service limits

Checks for basic and developer support plan:

* S3 Bucket permissions
* Security groups
* IAM Use
* MFA on root account
* EBS public snapshots
* RDS public snapshots
* Service limits

Checks for business and enterprise support plan:

* Cloudwatch alarms when reaching limits
* Programmatic access with AWS support API
* Full checks available on the 5 categories

AWS support plans:

Basic support plan:

* Customer service and communities
* 7 core checks from trusted advisor
* Personal health dashboard

Developer plan:

* Business hours access to cloud support associates
* Response times between 12-24 hours

Business support plan:

* Full trusted advisor checks and API access
* 24/7 phone email and chat access to cloud support engineers
* Faster response times for production emergencies

Enterprise on-ramp support plan:

* Access to technical account managerss
* Support team for best practices

Enterprise support plan:

* Designated TAM
* Faster response time when business critical system is down

Data transfer into S3 is free

For Linux EC2 instances you pay every second

**Advanced Identity**

AWS STS (Security Token Service) – Enables you to create temporary, limited credentials to access AWS resources

Cognito – Identity for web and mobile application users

Instead of creating an IAM user, you create a user in Cognito

Microsoft Active Directory (AD) is found on any windows server

Database of user accounts, computers, printers etc.

Centralised security and permissions management

Can be extended onto AWS using AWS Directory services

AWS IAM Identity Center – One login to access all AWS accounts centrally in your AWS Organisation

**Other services (rarely appear in exam)**

WorkSpaces – Managed Desktop as a Service to easily provision Windows or Linux Desktops

AppStream 2.0 – Desktop Application Streaming service

Delivery to any computer

Sumerian – Used to create VR/AR and 3D applications

IoT Core – IoT is internet of things, IoT Core lets you connect IoT devices to the cloud

Amazon Elastic Transcoder – convert media files stored in s3 to file formats required by playback devices

AppSync – Store and sync data across mobile and web apps in real-time

Uses GraphQL

Amplify – Set of tools and services that helps you develop and deploy scalable full stack web and mobile applications

Device Farm – Service that tests your web and mobile apps against real desktop browsers, mobile devices and tablets

AWS Backup – Centrally manage and automate backups across AWS services

Disaster Recovery Strategies (higher is lower cost):

* Backup and restore – create backups and restore applications from them
* Pilot light – minimal setup of applications available outside of main server
* Warm standby – Full version of app available
* Multi-site – Multiple full versions of app available

Elastic Disaster Recovery (DRS) – Quickly and easily recover physical, virtual and cloud-based servers into AWS in case of disaster

DataSync – Move large amounts of data from on-premises to AWS, tasks are incremental

Application Discovery Service – Plan your migration project by gathering information about your on-premises data centers

Application Migration Service – converts your servers to run natively on AWS

Fault Injection Simulator – Create disruptive events for testing (Chaos Engineering)

Step Functions – Build visual workflow to orchestrate Lambda functions

Ground Station – Fully managed service for managing satellite communications

Pinpoint – 2 way marketing communication service

Supports email, SMS, push, voice and in-app messaging