COMP5349– Cloud Computing Week 13: Review and Exam Info

Dr. Ying Zhou

The University of Sydney

Table of Contents

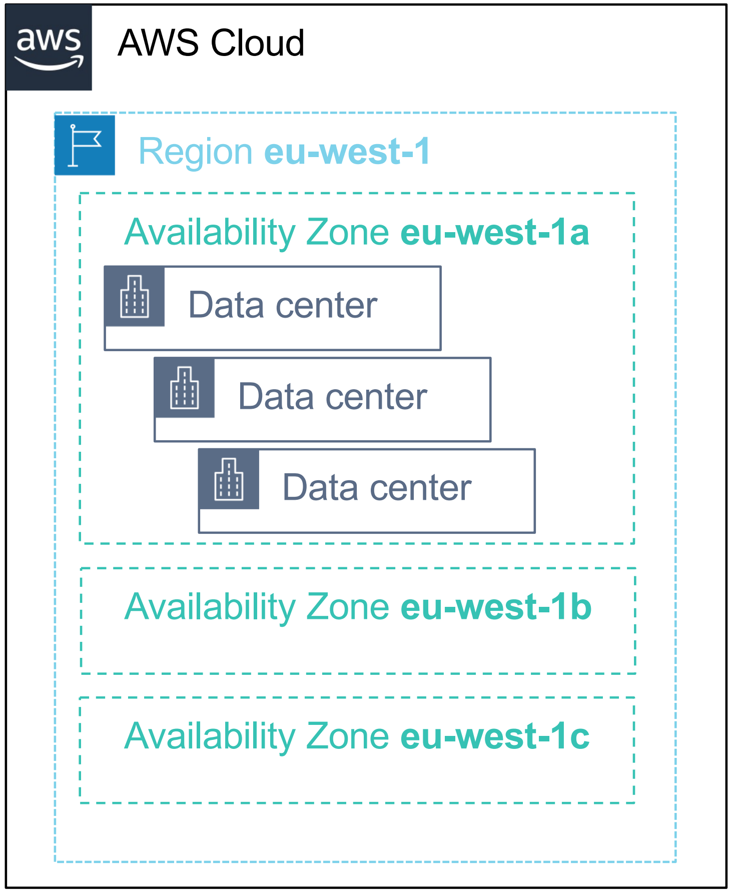
01 Course Review

02 Exam Information

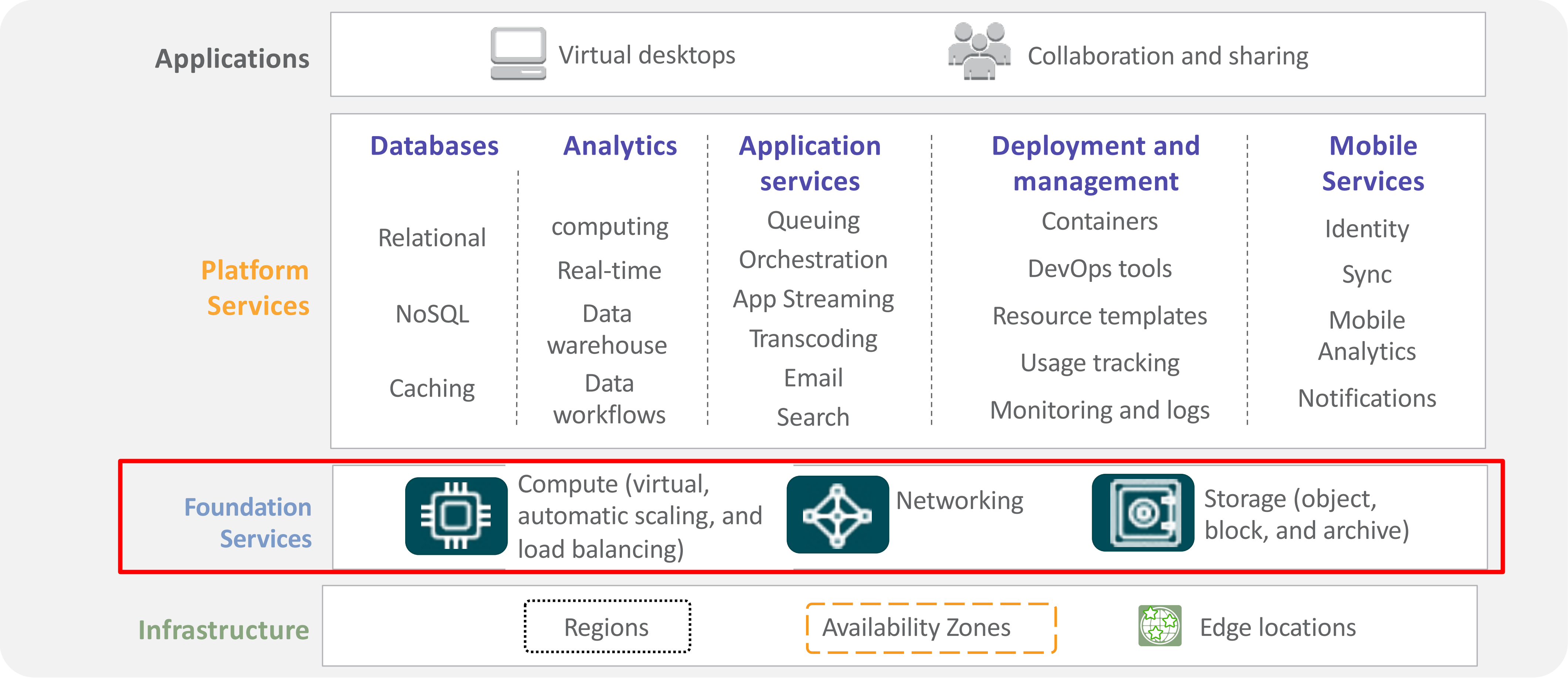
Course Review

AWS Global Infrastructure

- An **AWS Region** is a geographical area.
- A Region typically consists of two or more **Availability Zones**.
- Each **Availability Zone** is a fully isolated partition of the AWS infrastructure.
    - Availability Zones consist of discrete **data centers**
- A data center typically has 50,000 to 80,000 physical servers.


aws
AWS Cloud
Region **eu-west-1**
Availability Zone **eu-west-1a**
Data center
Data center
Data center
Availability Zone **eu-west-1b**
Availability Zone **eu-west-1c**

AWS foundational services



Applications

Virtual desktops

Collaboration and sharing

Platform Services

Databases

- Relational
- NoSQL
- Caching

Analytics

- computing
- Real-time
- Data warehouse
- Data workflows

Application services

- Queuing
- Orchestration
- App Streaming
- Transcoding
- Email
- Search

Deployment and management

- Containers
- DevOps tools
- Resource templates
- Usage tracking
- Monitoring and logs

Mobile Services

- Identity
- Sync
- Mobile Analytics
- Notifications

Foundation Services

Compute (virtual, automatic scaling, and load balancing)

Networking

Storage (object, block, and archive)

Infrastructure

Regions

Availability Zones

Edge locations

AWS Global Services

- A few AWS Services operate across the all regions, you don’t need to specify region, or az when create or access them
  - **IAM**
  - Route53
  - CloudFront

AWS regional service

- Some AWS Services operates at the region level
  - Resources are deployed at a particular region
- Examples:
  - S3
  - VPC
  - RDS (multi-AZ)
  - CloudFormation
  - Secretes Manager
  - ALB, ASG
  - SQS
  - SNS

AWS Zonal services

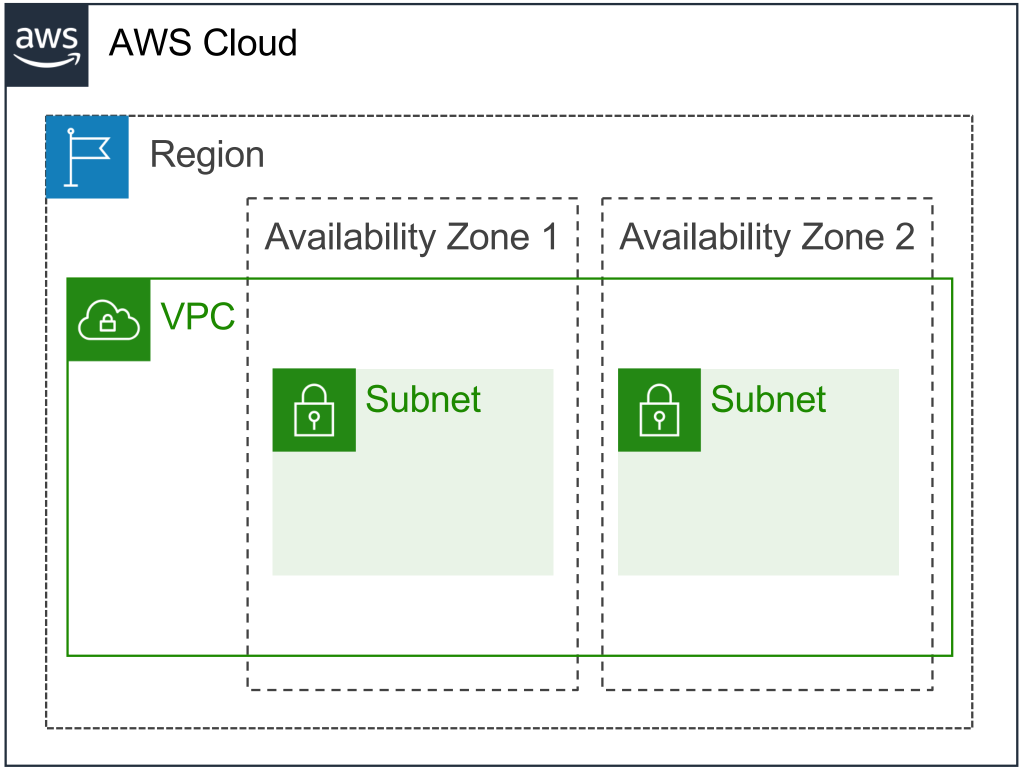
- Some services operate at AZ level
  - You must specify the AZ or equivalent when deploy them
- Examples
  - EC2
  - EBS
  - Single AZ RDS
  - Subnet

Network isolation and access control

- Some services requires users to control network access, may need IP address and security management. Usually those services are placed ***inside*** a VPC
  - EC2, databases and all related services like ALB, ASG, etc
- Other services are fully managed by AWS and does not require user to manage the network access are placed ***outsid*e** VPC
  - S3, SNS, SQS, DynamoDB, Secretes Manager, IAM, CloudFormation

Amazon Virtual Private Cloud

- VPCs:
  - Logically isolated from other VPCs
  - Dedicated to an AWS account
  - Belong to a single AWS Region and can span multiple Availability Zone
  - Each VPC has a CIDR range
  - Different VPCs can have overlapping CIDR range
- Subnets:
  - Range of IP addresses that divide a VPC
  - Belong to a single Availability Zone
  - Classified as public or private



Route table

- **Default Traffic Control:** Every VPC automatically comes with a "main route table."
  - This route table acts as the default traffic director for all subnets within the VPC, unless a subnet is explicitly associated with a custom route table.
  - By default, the main route table contains a local route, allowing communication within the VPC itself.
- A VPC can have many other route tables
  - They can be associated with different subnets
- Route table controls traffic at IP level
- Each subnet should only have one route table

Main VPC route table	
Destination	Target
10.0.0.0/16	local

Public and Private Subnet

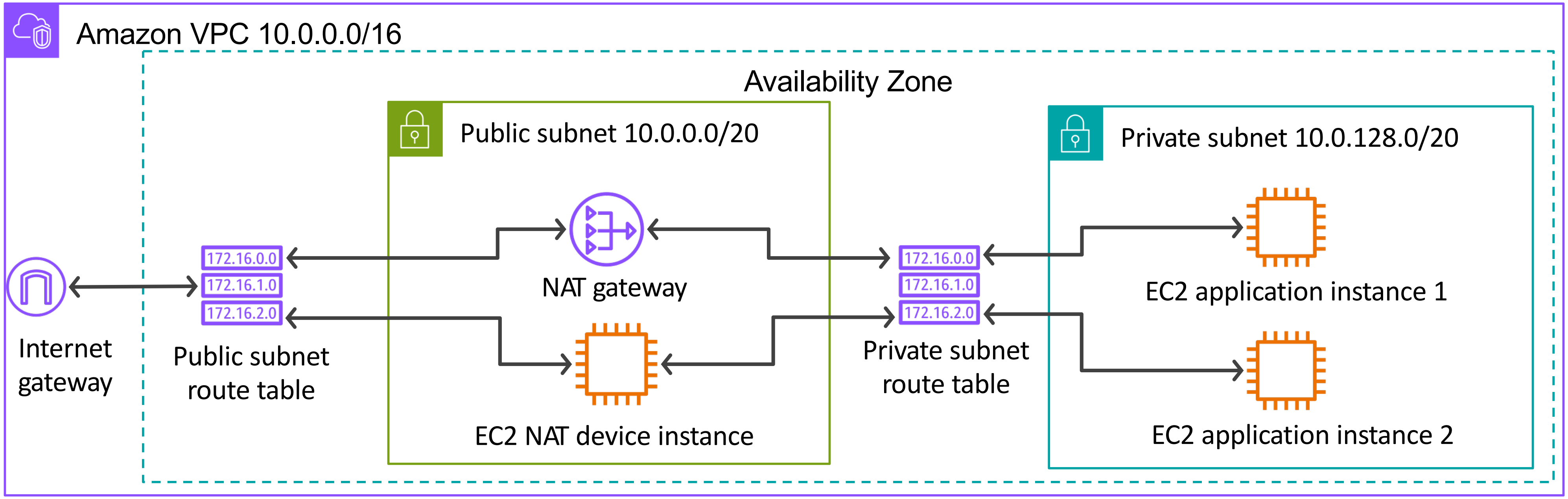
- Public subnet is reachable from the Internet
  - Two way traffic
- Private subnet is not reachable from the Internet, but may access Internet
  - One way traffic
- The accessibility is controlled by customized route tables

Public subnet route table	
Destination	Target
10.0.0.0/16	local
0.0.0.0/0	Internet gateway ID

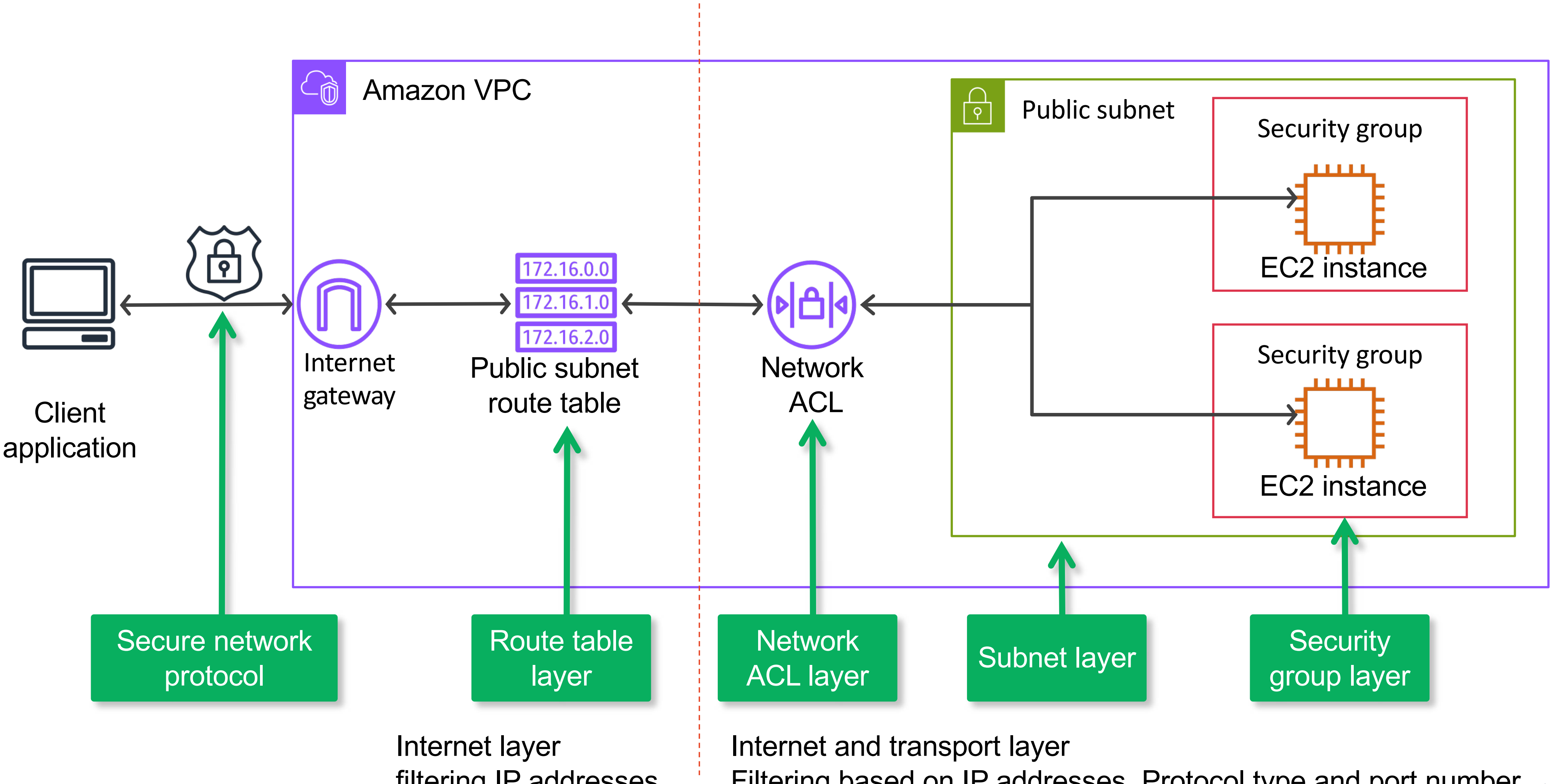
Private subnet route table	
Destination	Target
10.0.0.0/16	local
0.0.0.0/0	NAT gateway ID

Internet Gateway and NAT Gateway

- Internet Gateway is a VPC component
  - There is one per VPC
- NAT Gateway is a subnet component
  - It needs to be placed inside a specific subnet (public subnet)


Amazon VPC 10.0.0.0/16
Availability Zone
Public subnet 10.0.0.0/20
Private subnet 10.0.128.0/20
172.16.0.0
172.16.1.0
172.16.2.0
NAT gateway
EC2 application instance 1
172.16.0.0
172.16.1.0
172.16.2.0
Internet gateway
Public subnet route table
Private subnet route table
EC2 NAT device instance
EC2 application instance 2

Security layers of defense



Internet layer
filtering IP addresses

Internet and transport layer
Filtering based on IP addresses, Protocol type and port number

Network Access Control Lists (ACL)

- One Network ACL per subnet
- The Default network ACLs allow all inbound and outbound IPv4 traffic.
- A network ACL has separate inbound and outbound rules, and each rule can either allow or deny traffic.
- ACL rules are stateless
- To enable request-response based traffic, we always need to specify a pair of inbound and outbound rules

Security group

- Security group is a VPC component
- A VPC can have multiple security groups
- Security Groups are primarily applied to **compute resources or services running on compute resources** within the VPC,
  - EC2, RDS, ALB , Lambda
- Security Groups contain **rules that control inbound and outbound traffic** to the associated resources.
- **All Security Group rules are “allow” rules** — there are no explicit deny rules.
- Security Groups are **stateful**: if inbound traffic is allowed, the response outbound traffic is automatically allowed (and vice versa).
- It is common and recommended to specify **other Security Groups as the source or destination** in rules to allow controlled traffic between resources.

Referencing other SG in SG rules

- Security Group references is more flexible than using CIDR range as source/destination
  - Dynamic Membership
  - Simplifies Management
- Typical scenario:
  - **Web Server SG:** Allows inbound HTTP (port 80) from **anywhere** (0.0.0.0/0).
  - **App Server SG:** Allows inbound traffic on port 8080 **only from Web Server SG**.
  - **Database SG:** Allows inbound traffic on port 3306 (MySQL) **only from App Server SG**.

Identity and Access Management

Concept	What it is	Relation to Access
Account user	The original, fully privileged user of the AWS account. Can be authenticated using email/password	Has full access to everything by default; not recommended for daily use.
IAM User	An identity created within AWS account for a person or service. Can be authenticated using email/password or programmatically	Represents an entity that can be granted access via policies
Federated User	A user authenticated externally	Gains temporary credentials after external authentication; Access AWS resources by assuming roles
Group	A collection of IAM users	Helps manage permissions by attaching policies to groups; users inherit those permissions
Role	An IAM identity you can assume temporarily, often by AWS services or users	Used to grant access without long-term credentials; policies attached to roles define what the role can do
Policies	JSON documents defining permissions (allowed/denied actions)	Define what actions identities are allowed or denied to perform

IAM policies and permissions

- **Permissions in AWS** control what actions a user, group, or service can perform on AWS resources.
  - Doing anything with AWS resources requires proper permission
- A policy is a document defines a collection of permissions
- Account user has the permission to do anything to any resources belonging to the account
- All other entities get permission through policy
  - Identity based or resource based
  - There are different ways to specify permissions to allowing one resource to perform action on another resource

S3 permission

- S3 permissions can be specified as resource-based or identity based
- Public bucket means means either the bucket policy or the object ACL allows public read access.
  - No further identity-based policy is needed to read from the bucket
- Private bucket may use bucket policy to allow limited access
- We can also use identity-based policy to specify permission for S3 buckets

Allowing EC2 instance to access private S3 bucket

- An IAM role
- A policy attached to the role
- An instance profile
  - An IAM resource created to handle permissions for EC2 instances
  - It manages the temporary credential on behalf of the instance
- EC2 – instance profile – role - policy

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": [
        "s3:GetObject",
        "s3:PutObject"
      ],
      "Resource": "arn:aws:s3:::your-
bucket-name/*"
    }
  ]
}
```

Allowing EC2 instance to read from RDS?

- This is an application-level permission managed by RDS itself
- RDS instances manages its users and associated permission
  - EC2 instance need database credential to access the RDS
- Allowing EC2 instance to start/stop RDS is an AWS managed permission

```
{
  "Effect": "Allow",
  "Action": [
    "rds:StartDBInstance",
    "rds:StopDBInstance"
  ],
  "Resource": "arn:aws:rds:region:account-
id:db:your-db-instance-identifier"
}
```

Allowing S3 to publish to SNS topic

- Neither S3 nor SNS can assume role
- Such permission need to be managed as *resource-based policy*
- Attache a policy to SNS topic

```
{
    "Sid": "AllowS3ToPublish",
    "Effect": "Allow",
    "Principal": {
        "Service": "s3.amazonaws.com"
    },
    "Action": "SNS:Publish",
    "Resource":
       "arn:aws:sns:<your-region>:<your-account-id>:<your-topic-name>",
    "Condition": {
        "ArnLike": {
            "aws:SourceArn": "arn:aws:s3:::<your-bucket-name>"
        }
    }
}
```

Allow Lambda function to access S3

- Lambda function get permission through execution role
- It does not need a separate profile to be associated with the role
- Lambda function – execution role - policy

```
{
  "Effect": "Allow",
  "Action": [
    "s3:GetObject",
    "s3:ListBucket"
  ],
  "Resource": [
    "arn:aws:s3:::your-bucket-name",
    "arn:aws:s3:::your-bucket-name/*"
  ]
}
```

Allowing S3 to invoke Lambda function

- We need a resource based policy associated with the lambda function to allow it to be triggered by S3
- When this is done through console, the permission is automatically added

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Effect": "Allow",
      "Action": "lambda:InvokeFunction",
      "Principal": "s3.amazonaws.com",
      "Resource":
"arn:aws:lambda:REGION:ACCOUNT_ID:function:
MyLambdaFunction",
      "Condition": {
        "StringEquals": {
          "aws:SourceArn": "arn:aws:s3:::my-bucket"
        }
      }
    }
  ]
}
```

Compute resources

- Virtual Machine
  - EC2 powered by Nitro system
- Container
  - Docker or other container run time
- MicroVM
  - Firecracker VM running on bare metal instance to Support Lambda function

Replication and Scalability

- Theoretically unlimited copies
- Load balancer to distribute traffic
- Auto scaling to control the number of copies
- Provision unit
  - VM (Auto scaling group)
  - Task of container (ECS)
  - Pod of container (EKS)

Event Driven Architecture Services

- Compute
  - Lambda Function
- Event Router
  - SNS
  - Publisher/Subscriber model
  - Used to fan out messages/events
- Event Store
  - SQS
  - Support point-to-point communication

Simple Queue Service

- The messages are replicated in multiple queue servers to ensure high availability
- Message stays in the queue while they are processed by a consumer
  - Messages being process are hidden from other consumers
  - They will be deleted if the processing is completed within a configurable visibility time out window
  - They will become visible if the consumer does not finish processing within the timeout window or fail to delete it
- Queue types
  - Standard queue
    - At least once delivery
    - Best-effort ordering
    - Message could be processed multiple times
  - FIFO queue
    - FIFO delivery
    - Exactly once processing

Storage Services

- Simple Storage Service (S3)
  - Object based storage
  - Each S3 bucket is created within a specific region
  - The bucket name needs to be globally unique
- Elastic Block Store (EBS)
  - Block storage (like file system storage)
  - Each EBS volume is create within an AZ
  - It can only be attached to EC2 instance within the same AZ
  - It is automatically replicated within the AZ

Simple Storage Service

- Object stored in bucket will be automatically stored in different locations
  - Usually 3 physical copies are stored in different Azs
- Bucket can have version enabled to protect accidental update or delete
  - Each update creates a new version
  - Delete create a special version called delete marker
- Bucket can be replicated
  - Within the same region
  - Across different regions
  - By default, delete markers are not replicated

Cloud Database

- Hosted Database
  - RDS: fully managed by AWS
  - Single AZ, Multi-AZ deployment
  - Read replica configuration
- Cloud Native Database
  - Dynamo
  - Aurora (running as one of the DB engines of RDS)

Amazon Aurora

- Large scale relational database system
- The key innovation is the separation of db engine and storage nodes
  - DB engine runs modified MySQL or PostgreSQL and can have up to 15 readers
  - Data is always replicated 6 times in three different AZs
- It supports very large data volume (64TB as mentioned in the paper)
  - Data are partitioned into fixed sized segments called Protection Groups (PG)
  - There is no indication if the partitioning algorithm take into consideration the actual schema
- The primary DB engine is the one connecting both layers
  - Redo logs are sent from primary to DB readers and storage nodes
- The LSNs are used to ensure completeness at the the storage layer and consistency at the DB layer
  - Multiple important LSNs (VCL, CPL, VDL, etc) are established and advanced continuously

Exam Information

Final Exam Setting

- Duration
  - The exam is 2 hours
  - There are extra 10 minutes reading time
- The final exam is closed book; you can bring
  - A physical standard linguistic dictionary
- You will be given scratch paper
- The final exam has 100 points in total
- The final exam is worth 60% of your mark
- The final exam has 40% barrier
- You need to get 40 out of 100 points in the final exam to pass the unit

Final Exam Script

- Nine questions
- Question 1~4 are conceptual short answer questions
- Question 5~9 are scenario-based question
  - Each question has a scenario followed by multiple sub questions
  - All scenarios are related to some AWS services
- There is no coding questions
- You won’t be asked to read/write CloudFormation template
- There is no specific questions on Kubernetes, even though you may refer to it in some responses.

Exam Techniques

- Read each question carefully and make sure you understand it thoroughly before answering it
  - You have 10 minutes to just read the exam paper
  - Plan your time, choose the easiest questions to do first
- If you are uncertain about a question, you should attempt to answer it to the best of your ability. You may include your interpretation, or assumption as part of the answer
  - The exam invigilators are not part of the teaching team, they are not able to clarify anything

Exam Techniques

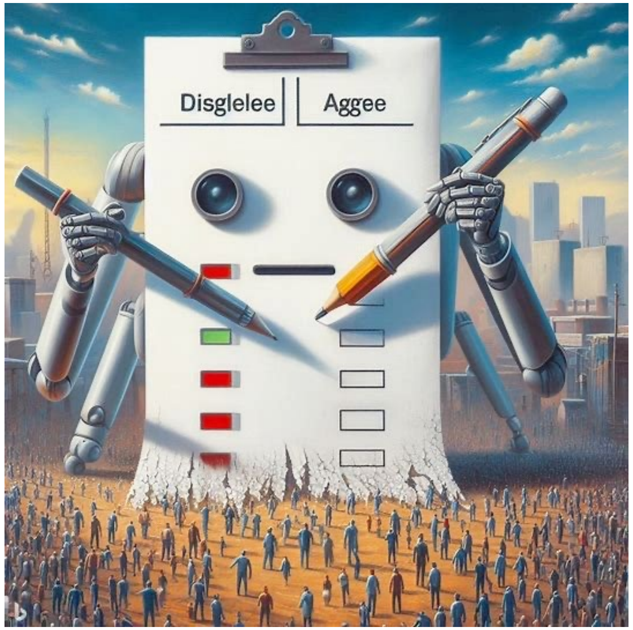
- Remember to bring your **student ID card**
- You can bring water and snacks
- Please check the university exam info web page for other permitted and not permitted items
- Visit the exam venue during the break to ensure you know how to get there
- Write **all** your answers on the exam paper, there are spaces provided after each question and at the end of the exam paper.
- Exam venues are usually equipped with many answer booklets, you can request one as scratch paper, but do not write any answer there.
- Remember to answer all questions

Replacement Exams

- “If you have a serious illness, injury or circumstances arise that affect your ability to attend or complete an exam, you may have grounds to apply for special consideration or special arrangements.”
- “If your special consideration or special arrangements application is approved, you may need to sit a replacement exam**. It's important to be aware of how this may affect you in terms of delaying certain aspects of your course**. For example, sitting an exam later **may delay when you receive your results**, which means you are not able to graduate or re-enrol for the next semester until your results are finalised.”
- Replacement exam period 1 will be held between 8 – 11 July 2025.
- Replacement exam period 2 will be held between 29 July – 1 August 2025

USS Survey Reminder

- https://student-surveys.sydney.edu.au/students/



powered by **DALL·E 3 2023**

THE UNIVERSITY OF
SYDNEY