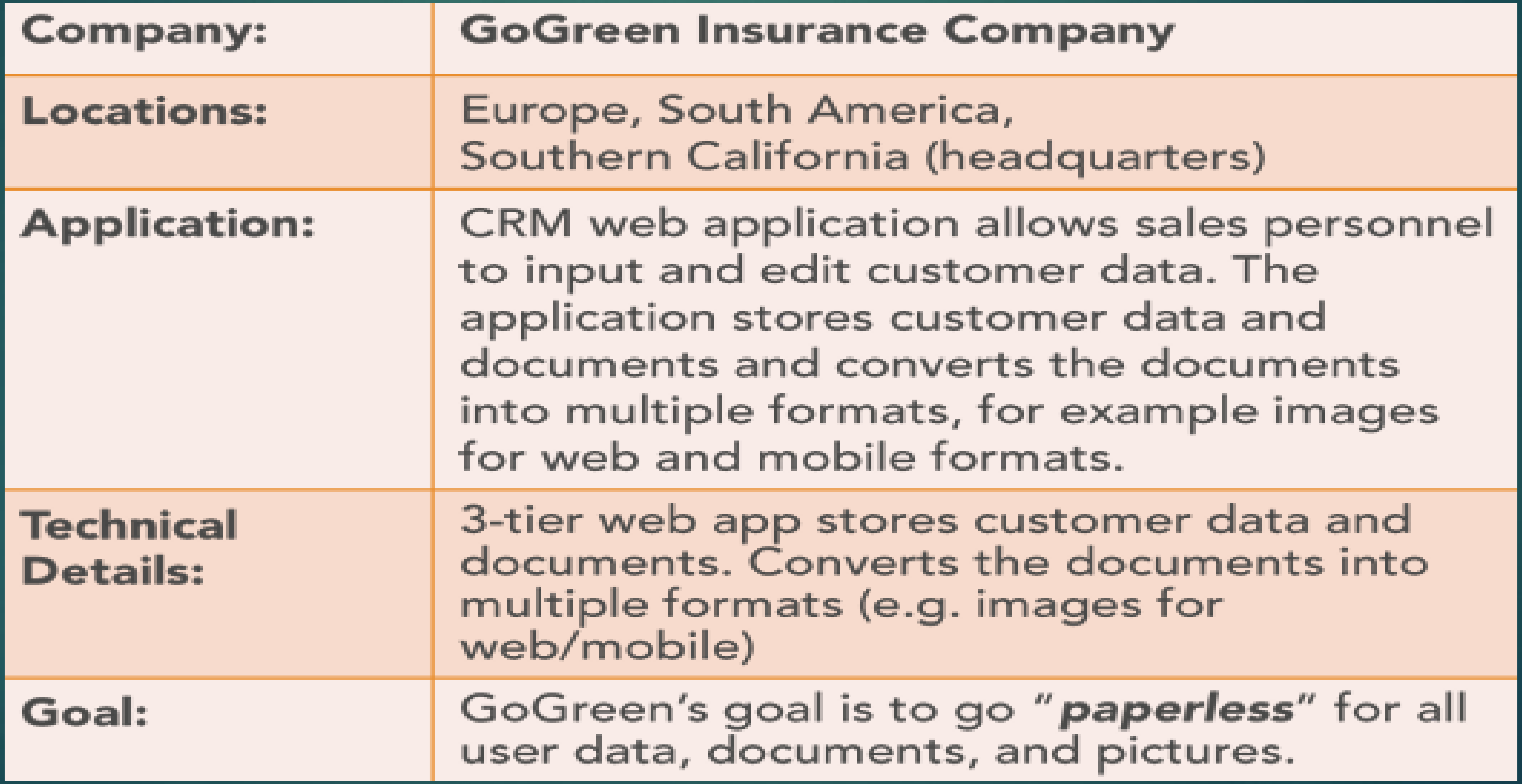
Cloud Infrastructure Services

Student Name: V. Sudheer
TP Number: TP088624
Module Name: CLOUD INFRASTRUCTURE AND SERVICES
Module Lecturer: Assoc. Prof. Dr. Muhammad Ehsan Rana
Module Code: 092025-MER
University: Asia Pacific University (APU)

COMPANY BACKGROUND



Company:	**GoGreen Insurance Company**
Locations:	Europe, South America, Southern California (headquarters)
Application:	CRM web application allows sales personnel to input and edit customer data. The application stores customer data and documents and converts the documents into multiple formats, for example images for web and mobile formats.
Technical Details:	3-tier web app stores customer data and documents. Converts the documents into multiple formats (e.g. images for web/mobile)
Goal:	GoGreen's goal is to go "***paperless***" for all user data, documents, and pictures.

Business Context and Problem Statement

Go-Green Insurance operates a legacy hosting environment that no longer meets performance or reliability needs.
The goal is to design a modern cloud architecture delivering high availability, 21,000 IOPS database performance, and secure storage without changing the DB schema.

Current Infrastructure Challenges

Key Challenges

1. Limited Availability

No Multi-AZ or DR redundancy.

2. Performance Bottlenecks

Legacy DB fails to hit 21k IOPS targets.

3. Maintenance Issues

Risky patching; restrictive legacy schema.

4. Backup Gaps

Cannot reliably meet 4-hour RPO.

5. Document Storage

Decentralized, unsecure file management

CURRENT ARCHITECTURE

Tier	Resource Count	vCPUs per Node	Memory per Node	Storage / Service	Operating System
Web Tier	6 VMs	2	4 GB	Apache Tomcat / PHP	RHEL 7.5
Application Tier	5 VMs	4	16 GB	Java SRE 7	RHEL 7.5
Database Tier	2 VMs	8	32 GB	5.5 TB / MySQL 5.7.22	RHEL 7.5

Design–Network

Document the VPC solution.

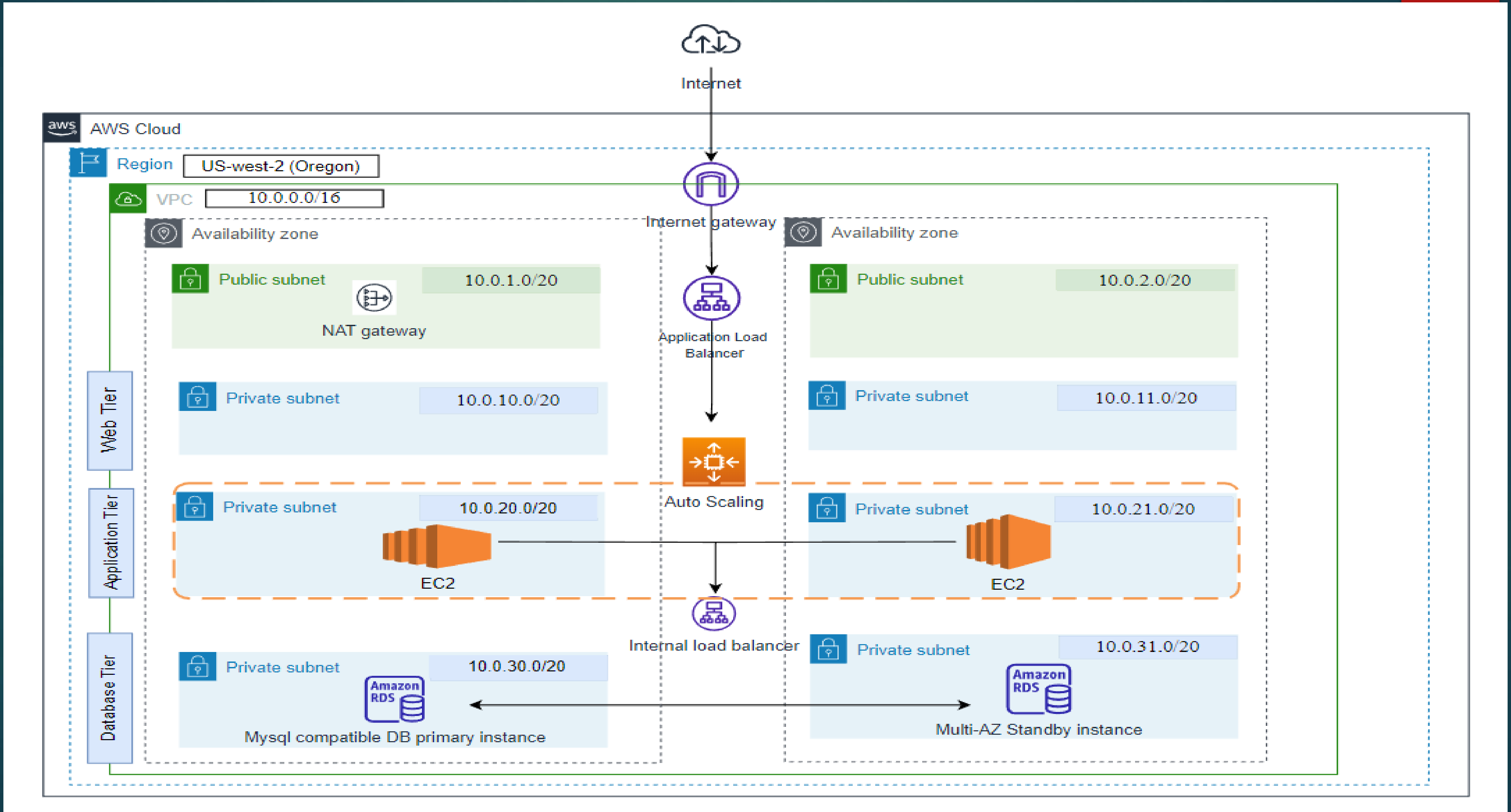
VPC	Region	Purpose	Subnets	AZs	CIDR Range
1	us-west-2 (Oregon)	Primary	Public(2), Private (2-Web, 2-App, 2-DB)	2	10.0.0.0/16
2	us-east-1 (N. Virginia)	Disaster Recovery	Public(2), Private (2-Web, 2-App, 2-DB)	2	10.1.0.0/16

Subnet Name	VPC	Type	AZ	Subnet Address
Public-sub-A	1	Public	us-west-2a	10.0.1.0/20
public-sub-B	1	Public	us-west-2b	10.0.2.0/20
web-1-private-A	1	Private	us-west-2a	10.0.10.0/20
web-2-private-B	1	Private	us-west-2b	10.0.11.0/20
App-1-private-A	1	Private	us-west-2a	10.0.20.0/20
App-2-private-B	1	Private	us-west-2b	10.0.21.0/20
DB-1-private-A	1	Private	us-west-2a	10.0.30.0/20
DB-2-private-B	1	Private	us-west-2b	10.0.31.0/20

Secondary (us-east-1) — Disaster Recovery-standby footprint

Subnet Name	VPC	Type	AZ	Subnet Address
Public-sub-A	1	Public	us-east-1a	10.0.1.0/20
public-sub-B	1	Public	us-east-1b	10.0.2.0/20
web-1-private-A	1	Private	us-east-1a	10.0.10.0/20
web-2-private-B	1	Private	us-east-1b	10.0.11.0/20
App-1-private-A	1	Private	us-east-1a	10.0.20.0/20
App-2-private-B	1	Private	us-east-1b	10.0.21.0/20
DB-1-private-A	1	Private	us-east-1a	10.0.30.0/20
DB-2-private-B	1	Private	us-east-1b	10.0.31.0/20

Proposed VPC Architecture Diagram



Design – Security

Security Group (SG)	SG Name	Rule (Allowed Port)	Source
ELB Load Balancer	SG-External-ALB-public	TCP 80, 443(HTTPS)	0.0.0.0/0
Web Tier	SG-web-Tier	TCP 8080(HTTP)	SG-External-ALB-public
App Tier	SG-APP-Tier	TCP 8080	SG-Web Tier Security Group
Database Tier	SG-DB-Tier	TCP 3306 (MySQL)	SG-App Tier Security Group

Other Security Options	Justification
AWS IAM Roles for EC2	Allows instances to securely access S3 buckets for document conversion without hard-coding credentials.
Amazon WAF (Web Application Firewall)	Integrated with the ALB to protect against common web exploits like SQL injection and Cross-Site Scripting (XSS).
Network isolation using private subnets	Prevents direct internet access to application and database tiers
AWS Shield Standard	Provides automatic protection against Distributed Denial of Service (DDoS) attacks for the public-facing ALB.
Multi-AZ deployment	Improves availability and fault tolerance
VPC Flow Logs	Enables the Cloud Team to monitor and analyze network traffic for troubleshooting and security audits.

Design – Encryption

Requirement	Solution
Encryption option for data at rest	Use of AWS Key Management Service (KMS) for all storage services. This includes EBS volume encryption, RDS database encryption, and S3 bucket encryption using SSE-KMS.
Encryption option for data in transit	Transport Layer Security (TLS 1.2/1.3) is implemented for all communications, with ACM certificates securing HTTPS traffic at the Application Load Balancer and SSL/TLS encryption protecting database connections and internal tier communications.

Two Step Cloud Migration: PHASE 1: Legacy → AWS (x86 / AMD)

Tier	AMI	Tag	Type	Size	Justification	# of Instances
Web	Amazon Linux 2	Key: Name Value: web-tier-phase-1	EC2	m7a.large (AMD EPYC)	Handles HTTP requests efficiently, cost-effective, supports Auto Scaling	4 instances Auto-scale
App	Amazon Linux 2	Key: Name Value: app-tier-phase-1	EC2	m7a.2Xlarge	Supports document processing and business logic with higher memory	4 instances Auto-scale
DB	N/A (Managed)	N/A	Amazon RDS (MySQL)	db.r6a.2Xlarge + io2	High IOPS, Multi-AZ support, fully managed, meets performance and availability requirements	1 primary + 1 standby

Two Step Cloud Migration: PHASE 2: AMD → AWS GRAVITON (ARM)

Tier	AMI	Tag	Type	Size	Justification	# of Instances
Web	Amazon Linux 2023.9 (ARM64)	Key: Name Value: web-tier-phase-2	EC2	m7g.Xlarge (AMD EPYC)	delivers improved price-performance compared to x86-based instances while maintaining predictable performance under sustained web traffic.	4 instances Auto-scale
App	Amazon Linux 2023.9 (ARM64)	Key: Name Value: app-tier-phase-2	EC2	m7g.2Xlarge Or r7g.2Xlarge	predictable CPU utilization and increased memory capacity, ensuring application performance remains within the defined CPU and memory utilization thresholds during peak workloads.	5 instances Auto-scale
DB	N/A (Managed)	N/A	Amazon RDS (MySQL)	db.r7g.2xlarge + io2	High IOPS, Multi-AZ support, fully managed, meets performance and availability requirements	1 primary + 1 standby

Design: Recovery Point Objective

How would you achieve a Recovery Point Objective (RPO) of four hours?

- A four-hour Recovery Point Objective (RPO) is met using AWS Backup and RDS Automated Backups. RDS transaction logs are captured every 5 minutes, while EC2 incremental snapshots are taken every four hours, all replicated to the us-east-1 disaster recovery region.
- For unstructured data, S3 versioning and cross-region replication ensure continuous synchronization. Combined, these strategies guarantee that no more than four hours of data is lost during a regional failure, fully meeting the RPO requirement.

Design: Document Storage

Storage / Archive Option	Detail
S3 Standard	Designed for frequent, millisecond access to active data (0–90 days). Offers high durability and availability with no retrieval fees.
S3 Standard-IA	Optimized for data accessed less frequently (90–180 days), providing a 40% cost reduction compared to Standard, with the same millisecond retrieval.
S3 Glacier Deep Archive	Lowest-cost archival storage for long-term retention (180 days to 5+ years) at $0.00099/GB, with retrieval times ranging from hours to 48 hours.

Design: Web Tier

Requirement	Solution
Architecture must be flexible and handle any peak in traffic or performance.	m7a.large instances provide **non-burstable, predictable CPU.** Auto Scaling Group ensures horizontal scaling. ALB distributes load evenly and supports sudden spikes
The overall acceptable incoming network bandwidth is between 300 Mbps and 750 Mbps.	Each m7a.2xlarge supports multi-Gbps networking. Aggregate bandwidth across 4–6 instances easily exceeds 750 Mbps ALB scales automatically without bandwidth constraints.
Application administrators want to be notified by email if there are more than 100 “400 HTTP errors” per minute in the application.	ALB access logs + CloudWatch metrics Set up a CloudWatch Alarm with an SNS email notification for 400 errors exceeding threshold>100.

WEB TIER

PARAMETER	EXISTING SETUP	PROPOSED SETUP
INSTANCE TYPE	6VMs(each 2 vcpus, 4gb memory)	4 instance of **m7a.large** (each 2 vcpus, 8gb memory
Total vcpus	6 * 2= 12 vcpus	4*2 = 8vcpu’s
Total Memory	6 * 4 gb = 24gb	4 * 8 gb = 32gb
Memory utilization (%)	18 gb (75% of 24 gb)	18 gb (56% of 32gb)

Design: Application Tier

Requirement	Solution
Architecture must be flexible and handle any peak in traffic or performance.	m7a.2xlarge provides sustained CPU for Java workloads EC2 Auto Scaling responds to sustained load.
Overall memory and CPU utilization should not go above 80% and 75% respectively or below 30% for either.	Increased capacity reduces baseline utilization from ~90% to ~50% **CloudWatch alarms:** Scale out if CPU > 75% or Memory > 80% Scale in if CPU < 30% for sustained duration
Internet access is required for patching and updates without exposing the servers.	Instances remain in private subnets(web,app,db) **NAT Gateway** in public subnet provides outbound internet access No inbound internet connectivity → servers remain protected Security Group allows traffic only from Web Tier No direct internet exposure

APPLICATION TIER

PARAMETER	EXISTING SETUP	PROPOSED SETUP
INSTANCE TYPE	5VMs(each 4vcpus, 16gb memory)	5 instance of **m7a.2xlarge** (each 8 vcpus, 32gb memory
Total vcpus	5 * 4= 20 vcpus	5*8 = 40vcpu's
Total Memory	5 * 16 gb = 80gb	5 * 32 gb = 160gb
Memory utilization (%)	CPU: 18vcpus (90% of 20) Memory: 72gb (90% of 80gb)	CPU: 18 vcpus (45% of 40) Memory: 72gb (45% of 160gb)

Design: Database Tier

Requirement	Solution
Database needs consistent storage performance at 21,000 IOPS.	Use **Amazon RDS with Provisioned IOPS (io2) EBS storage**. Provisioned IOPS ensures predictable and consistent high performance suitable for transactional workloads requiring 21,000 IOPS.
High availability is a requirement.	Enable **Amazon RDS Multi-AZ deployment**. This automatically maintains a synchronous standby replica in a different Availability Zone, providing automatic failover and minimal downtime during failures.
No change to the database schema can be made at this time.	Use a **managed relational database engine (Amazon RDS e.g., MySQL or PostgreSQL)**. RDS allows lift-and-shift migration without schema changes, ensuring application compatibility and reducing migration risk.

DATABASE TIER

PARAMETER	EXISTING SETUP	PROPOSED SETUP
INSTANCE TYPE	2VMs (each 8VCPUs, 32GB memory)	2 instance of **db.r6a.2Xlarge + io2** (each 8 vcpus, 32gb memory

Design – Additional Services

1. **Amazon CloudWatch**
   To monitor CPU, memory, and HTTP error metrics across all tiers and trigger alarms for scaling or alerts.

2. AWS Global Accelerator: Routes global traffic over the AWS backbone to reduce latency for Europe and South America.

3. AWS Systems Manager (SSM): Secure shell-less access to instances.

4. Secrets Manager: Automatic rotation of database credentials.

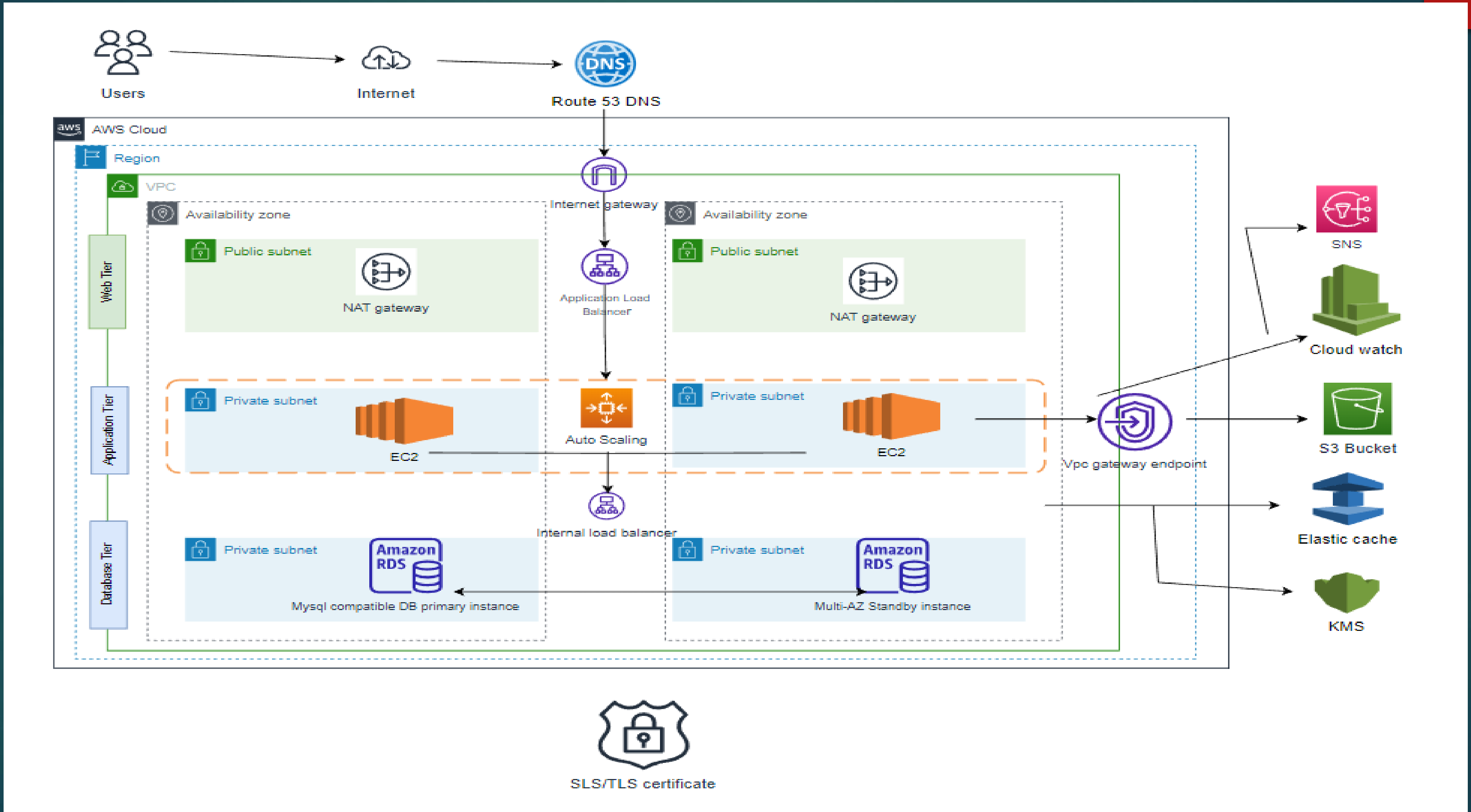
5. **Amazon SNS (Simple Notification Service)**
   Used to send email alerts to administrators when 400 HTTP errors exceed 100 per minute.

6. **Amazon S3 with Cross-Region Replication**
   Ensures document continuity and durability across primary and DR regions.

7. **Amazon RDS Snapshot Copy**
   Supports disaster recovery by preserving database backups across regions, aligned with RPO goals.

8. **AWS Auto Scaling**
   Automatically adjusts EC2 capacity in Web and Application tiers to handle performance spikes and maintain target utilization.

Proposed Architecture Diagram



GoGreen Insurance Company Cost Considerations

The proposed solution should use the most cost-conscious financial options. What are the cost considerations?

Cost-Conscious Financial Options

- **Global Acceleration vs. 3rd Region:** Using AWS Global Accelerator (~$18/month + transfer fees) is significantly cheaper than deploying a third full stack in Europe, which would cost thousands more in redundant compute and storage.

- **AMD Price Advantage:** Choosing AMD instances saves ~10% compared to Intel for the initial migration phase.

- **Storage Tiering:** Archiving documents in S3 Glacier Deep Archive reduces storage costs by over 95% compared to S3 Standard.

- **Savings Plans**: Committing to a 3-year Compute Savings Plan for the baseline fleet can reduce compute costs by up to 50%.

- **Use Auto Scaling for EC2 instances**
  Automatically adjusts server capacity based on demand, so you only pay for what you use — no overprovisioning.

- **Implement warm standby in DR region**
  Keeps disaster recovery costs low by running minimal resources that can scale up only when needed.

Thank You