41891 Cloud Computing Infrastructure

Assessment 2: Group Major Project

Aden Northcote 13960570
Christopher Atchison 13911870
Rory O'Hara 13919749

Contents

1 Background

SmartV is a popular online platform providing a wide range of video tools to end-users. Services are provided to consumers over the internet and include video streaming and delivery, video searching, video editing, video transcoding and adaptation.

The SmartV platform is well used and is currently experiencing significant growth in user-base. This trend has prompted a transition from on-premises infrastructure to the cloud with the aim of future-proofing the company's scalability and cost-competitiveness.

2 Business Requirements

The operating model of the SmartV platform relies heavily on ready access to large amounts of storage, compute, and networking capability. This is especially true given the large amounts of high resolution video data the platform is expected to process and store. The business requirements captured from this understanding are presented here in two sections: the non-functional requirements relating to SmartV's operations and the functional requirements relating to the cloud infrastructure specifically.

2.1 Operational Requirements

The requirements outlined here represent non-functional and operational functions of the cloud infrastructure solution.

Scalability

Scalability represents a primary business objective for SmartV's transition to cloud-based infrastructure, and underlies most of the business requirements outlined in this document. The deployment of scalable infrastructure provides SmartV the ability to react to fluctuations in platform usage while minimising any costs associated with under-utilised infrastructure.

Availability

The SmartV platform's customer-facing model requires services to be highly available to end-users, ensuring uninterrupted access to videos and features. Services need to be available even during peak usage times or unforeseen spikes in demand.

Reliability

The cloud infrastructure used to enable SmartV must also prioritise high availability, with resilient architecture that minimises downtime and ensures continuous service delivery.

Security

A high level of security is required to protect user data and the SmartV platform itself. All data hosted by the cloud infrastructure should be encrypted and subject to strict access control and authentication. Both ingress and egress traffic should be restricted with additional monitoring across the network and infrastructure enabling incident detection and response.

2.2 Infrastructure Requirements

The functional requirements of the cloud infrastructure itself form the core project, with each item listed here built on the operational requirements outlined above.

Storage

Effective storage management is a key requirement of SmartV's infrastructure, with the service housing massive amounts of multi-media data for its users. All stored data also needs to be safely and regularly replicated for back-ups and disaster recovery.

Compute

The computational intensity of video based workloads necessitates significant amounts of domain-specific compute availability in the form of graphics processing unit (GPU) access, as well as the additional components required to support them.

Network

High throughput and reliable networking is key to effective delivery of SmartV services.

Infrastructure Management

Deployed cloud infrastructure should be fully managed and configurable in anticipation of changes in future requirements, deployment of new services, and regular updates and maintenance.

3 Cloud Architecture and Design

This system will be deployed using Infrastructure as a Service (IaaS). It is assumed that SmartV brings their own software and platform to be placed on top of the infrastructure, and deployed in a narrow global region, in this case the Asia Pacific region.

3.1 Platform Selection

3.2 Design Assumptions

3.3 Infrastructure Design

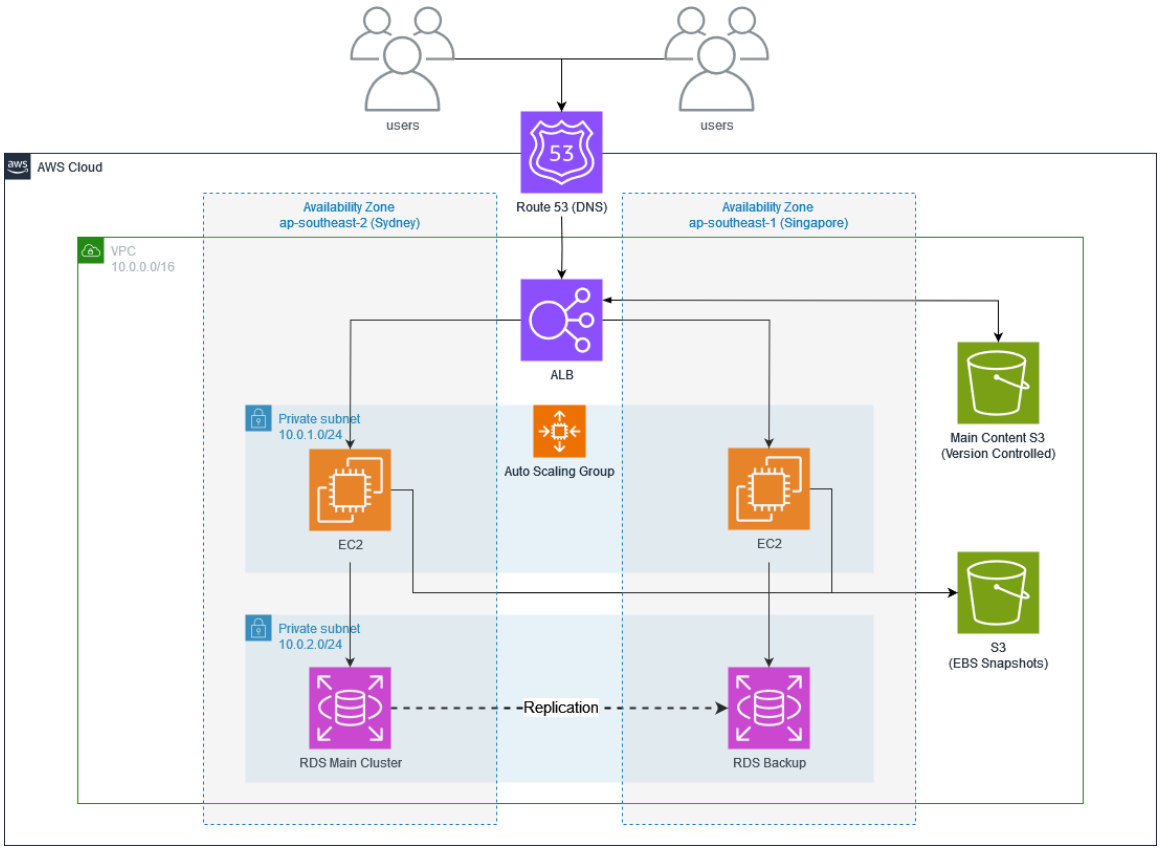

Figure 1: Cloud Infrastructure Topology

3.4 Infrastructure Components

DNS (Route 53)

Allowing global access through http, a DNS record is needed and will be achieved through an AWS Route 53 domain. This will transfer users to the ALB.

Network (ALB)

To balance load across seperate compute clusters, an application load balancer (ALB) will handle distribution of incoming requests to allow transparent scaling of compute abstracted from the user,

meaning bucekts of storage can scale infinitely.

Compute (EC2)

Using an autoscaling group of EC2 instances of type "p3.8xlarge", compute will be handled and scaled with demand. The instance types are very large to accomodate for the intense RAM, CPU, and GPU requirements video demands. A key feature of S3 is the inherent scalability, as this is abstracted away from the end user.

Content Storage (S3)

AWS Simple Storage Service (S3) allows for cheap and easy object storage through a HTTP API endpoint. This is suitable for storing the large files used by the service, and allowing avilability across the platform.

Backup Storage (S3)

Snapshots of the drives associated with compute instances, the elastic block storage (EBS) drives, will be palced daily in a seperate S3 bucket to allow disaster recovery of user's work and files retained locally on the compute instance.

User Database (RDS)

AWS Relational Database System (RDS) will be used to handle user information such as login, account details, and payment information. Using a traditional database system is suitable over S3 as it allows structure and replication.

Network (Public)

The entire cloud system will exist within a private VPC, with the only publically exposed element being the ALB accessible through the Route 53 DNS record or associated IP address.

Network (Internal)

Internally, the platform is networked through a private IPv4 subnet in CIDR block 10.0.0.0/16. Each private subnet in turn has it's own CIDR block with a netmask of 24. Compute and the RDS service reside in seperate private subnets to improve security.

Scaling (Compute)

Using EC2 autoscaling groups, AWS will launch new compute instances as demand increases on pre-existing ones, placing them behind the ALB. Retaining a minimum of 1 instance ensures constant availability. An option available is keep "warm pools" of compute instances pre-initialized but not in use, allowing near instant upscaling as demand spikes.

High Availability

High availability will be achieved through seperating cloud resources across two AWS availability zones (AZs), allowing an entire region of AWS to lose access, while maintaining service for customers.

3.5 Pricing

4 Considerations and Challenges

5 Evolution of Technology