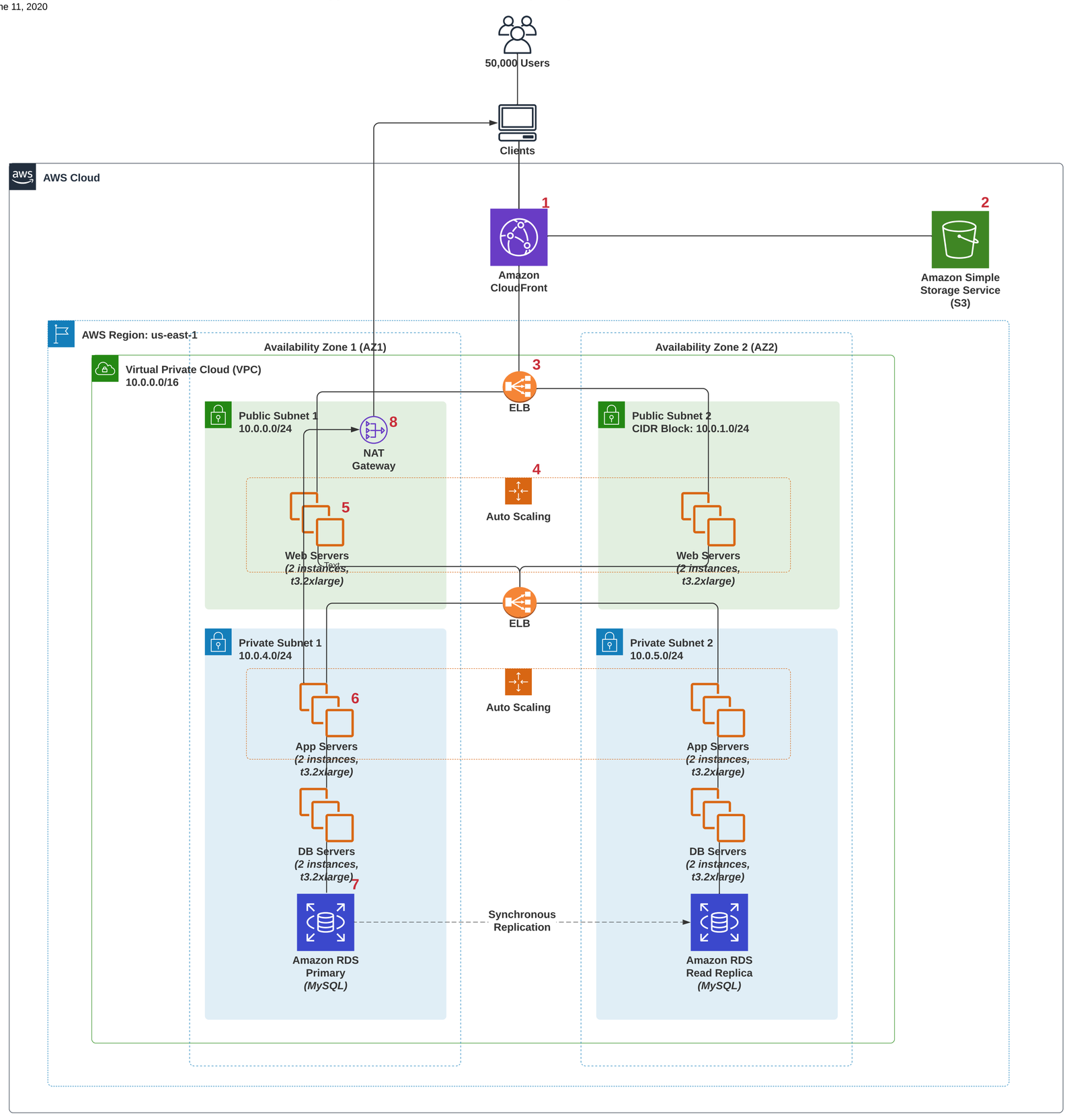
Building a Performant, Available and Cost-Effective Social Media Application in AWS with 50,000 Single-Region Users

Hou Chia | June 11, 2020



System Elements

1. Amazon CloudFront is being used for two purposes in this architecture. First, Amazon CloudFront serves as a reverse proxy server that directs client's HTTP requests to the backend (i.e., the ELB in this case). Second, it is being used as a Content Delivery Network (CDN). Amazon CloudFront consists of a distribution of servers across the globe that deliver content to users using the fastest route possible based on the user's location. Amazon CloudFront leverages a global network of 216 Points of Presence to deliver content to end users with reduced latency.

If a user requests content that is not yet available at an Edge Location, Amazon CloudFront retrieves the content from the origin - which in this case is the S3 bucket and stores it at an Edge Location. The next time the same content is requested, it's already cached and can be served immediately. To further improve the performance of the application, we can increase the cache expiration time for any media assets to ensure they are served from an Edge Location. CloudFront also enables content to be secure.

2. S3 is a global AWS managed service that offers highly durable storage. Data stored in S3 is never lost or compromised, as S3 offers 11 9's availability. S3 also scales automatically. S3 is optimized for media files, which makes it ideal for a social media application.

3. The Elastic Load Balancer (ELB) distributes incoming HTTP requests among multiple Amazon Elastic Cloud Compute (EC2) instances across two Availability Zones (AZs).

4. Auto Scaling helps make our web servers hosted on EC2 instances elastic and highly available, as it adds and scales down servers servers as needed based on changes in compute demand, currently set at XX%. The AWS EC2 instances in both AZs are treated as a logical grouping and follow the same set of automatic scaling and management rules.

5. Web Servers, which mainly serve static assets via HTTP, are deployed on Amazon EC2 instaces.

6. Application Servers, which serve dynamic content are deployed on Amazon EC2 instaces.

7. Amazon RDS streamlines the process of setting up, operating, and scaling a relational database in the AWS cloud. Amazon RDS is a managed service, which means Amazon is responsible for executing any compute, memory, network and hardware-related tasks for our database, including periodic upgrades.

We implement a primary RDS instance in AZ1 an a Read Replica in AZ2. The Read Replica is asynchronously updated when the primary database changes. The Read Replica reduces read-only traffic to the primary database, thereby freeing up the latter to perform other operations. During disaster recovery, the Read Replica can be promoted to become the primary database. Both instances are IOPS SSD provisioned to ensure high performance.

8. A network address translation (NAT) gateway enables the app server instances in the private subnet to connect to the internet (e.g., software updates or connections to other AWS services), but prevent the internet from initiating a connection with app servers. In this architecture, since the NAT gateway only exists in a single AZ per the assignment requirement, the system will lose connection to the internet completely if AZ1 or the public subnet 1 fails.