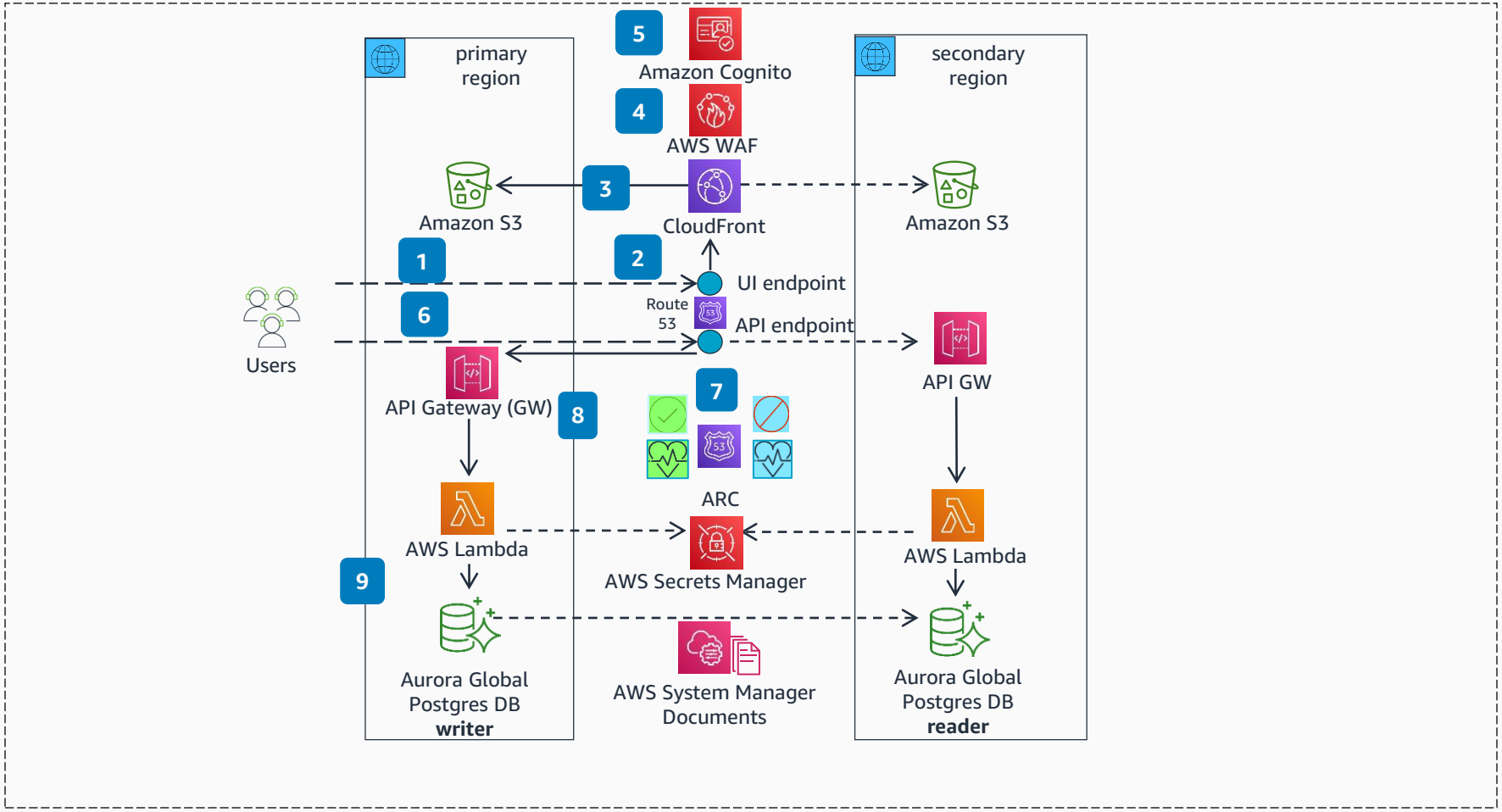
Guidance for Cross Region Failover and Graceful Failback on AWS

Application Running in Primary Region

The guidance details the regional failover and observability setup required to enable automated/manual failover and failback procedures.



1. The user opens the browser and enters the UI domain name system (DNS) endpoint hosted on **Amazon Route 53**.
2. The request is routed to the **Amazon CloudFront** instance. The data plane for **CloudFront** is globally available.
3. **CloudFront** delivers static content stored in **Amazon Simple Storage Service** (Amazon S3) buckets in the primary region (or, if it is not available, the secondary region in failover mode).
4. **CloudFront** is protected by **AWS WAF**, which is configured with standard rules to protect against common web exploits.
5. The UI is authenticated by an **Amazon Cognito** user pool configured in the primary region. **Amazon Cognito** is a regional service. If there is degradation or an outage in the **Amazon Cognito** service in the primary region, this may impact application failover.
6. The **Amazon Route 53** DNS-hosted zone is powered by **Route 53** Application Recovery Controller (ARC), which contains an ARC control for the primary and secondary regions. The ARC controls the respective health checks, which power the respective DNS records in a **Route 53**-hosted zone. Initially, the primary region ARC control is turned on and the secondary region ARC control is turned off.
7. The primary region health check becomes healthy and the secondary region health check becomes unhealthy. Consequently, the **Route 53** API DNS endpoint resolves to the API endpoint in the primary region.
8. The API endpoint in the primary region delegates the calls to corresponding **AWS Lambda** functions running in the primary region.
9. The application uses an **Amazon Aurora** global database to store application transaction data. Initially, the primary **Aurora** database cluster is configured as the writer cluster, and the secondary **Aurora** database cluster is configured as a reader cluster.

   The primary **Aurora** database cluster automatically replicates data to the secondary **Aurora** database cluster, which enables the application to run from the secondary region (using the data replicated from primary database cluster to secondary database cluster).

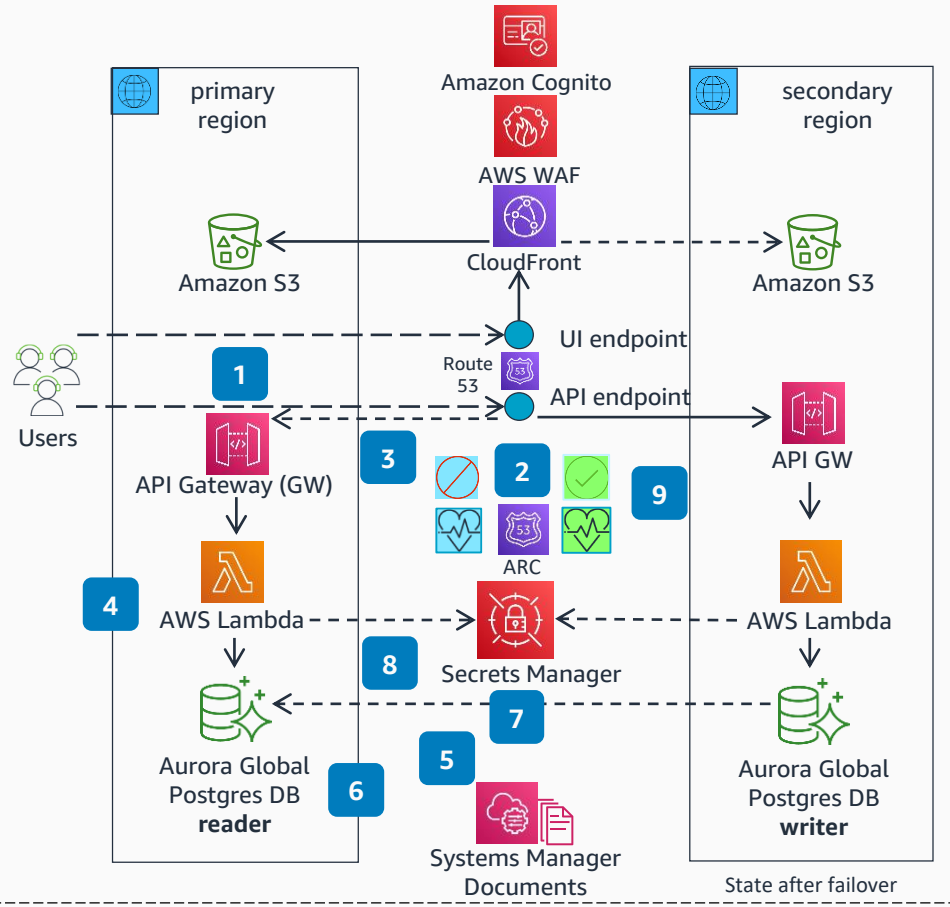
Reviewed for technical accuracy November 23, 2022

AWS Reference Architecture

Guidance for Cross Region Failover and Graceful Failback on AWS

Cross Region Failover

The guidance details the regional failover and observability setup required to enable automated/manual failover and failback procedures.



1. The user clicks the “Failover” button in the UI, which invokes the failover API endpoint hosted on Route 53.
2. The **Amazon Route 53** DNS-hosted zone routes to the API endpoint in the primary region based on the state of the **Route 53** ARC controls.
3. The primary API endpoint is invoked.
4. The primary API endpoint delegates the invocation to the corresponding **AWS Lambda** function running in the primary region.
5. The **Lambda** function calls the failover runbook, automated as an **AWS Systems Manager** Document. The runbook automates the three steps involved with the failover process.
6. The runbook first fails over the **Amazon Aurora** global database from the primary to secondary regions, making the database cluster in the secondary region the writer cluster and the database cluster in the primary region the reader cluster.
7. After failover, the **Aurora** global cluster is configured to automatically replicate data from the database cluster in the secondary region (writer) to the database cluster in the primary region (reader).
8. The runbook updates the database secret in **AWS Secrets Manager** with the database endpoint of the **Aurora** database cluster in the secondary region so that **Lambda** functions use the new database endpoint to interact with the database.
9. The runbook flips the **Route 53** ARC controls, turning the ARC control for the primary region off and turning ARC control for the secondary region on. As a result, the secondary region health check becomes healthy and the primary region health check becomes unhealthy. **Route 53** API DNS endpoint resolves to the API endpoint in the secondary region. This completes the failover.

   The application will be live in the secondary region, routing API traffic to the secondary region. It invokes Lambda functions in the secondary region, which interacts with the **Aurora** database cluster in the secondary region.

Reviewed for technical accuracy November 23, 2022

AWS Reference Architecture