Overall system high availability system design can be illustrated here.

Key design considerations are

1. Route 53 which by design is a highly available and scalable service. It uses a global network of DNS servers to handle large amounts of traffic and provide low-latency responses. Most importantly, it is able to failover between regions without interruptions.
2. Next we have an active/active setup in two different regions. This is to ensure seamless failover if one region goes down. We do not waste any time trying to failover or start up EC2 instances.
3. Application load balancer distributed to incoming request load to different EC2 instances that uses AWS Auto-scaling to scale out based on current load.
4. Monolight applications are deployed into EC2 instances.
5. Application also requires external services such as S3, DynamoDb, Kms and others. These services are configured to support 2 availability zones. This will prevent a single service failure from bringing down the entire cluster. So S3 will need to have multi zone setup, KMS need to have a multi-zone setup and the same applies to DynamoDB.
6. Applications often need to pass data between components and these data can sometimes be critical. For example, when a user makes a payment, this information must be retained even in the advent of a system failure. This is where AWS SQS comes into the picture. Applications need to be architected to be resilient and fault tolerant too. One of the key aspects is to embrace common cloud architecture patterns. Queue is a medium to persist data or messages between applications components.
7. While a system can be designed to be highly available, sometimes it cannot escape downtime for example when an OS patches gone wrong. To address this we need AWS backup so we can restore the system to a healthy point.
8. As for monitoring and alerting, we make full use of CloudWatch

