While the new architecture scales infinitely, we are noticing a latency penalty on the first request after a period of inactivity, often referred to as a 'cold start.' To mitigate this, we are using provisioned concurrency for critical paths. The application logic is broken into ephemeral, stateless functions that have a maximum execution time of 15 minutes. We rely heavily on managed services for state (like DynamoDB) since the compute environment is destroyed immediately after execution.