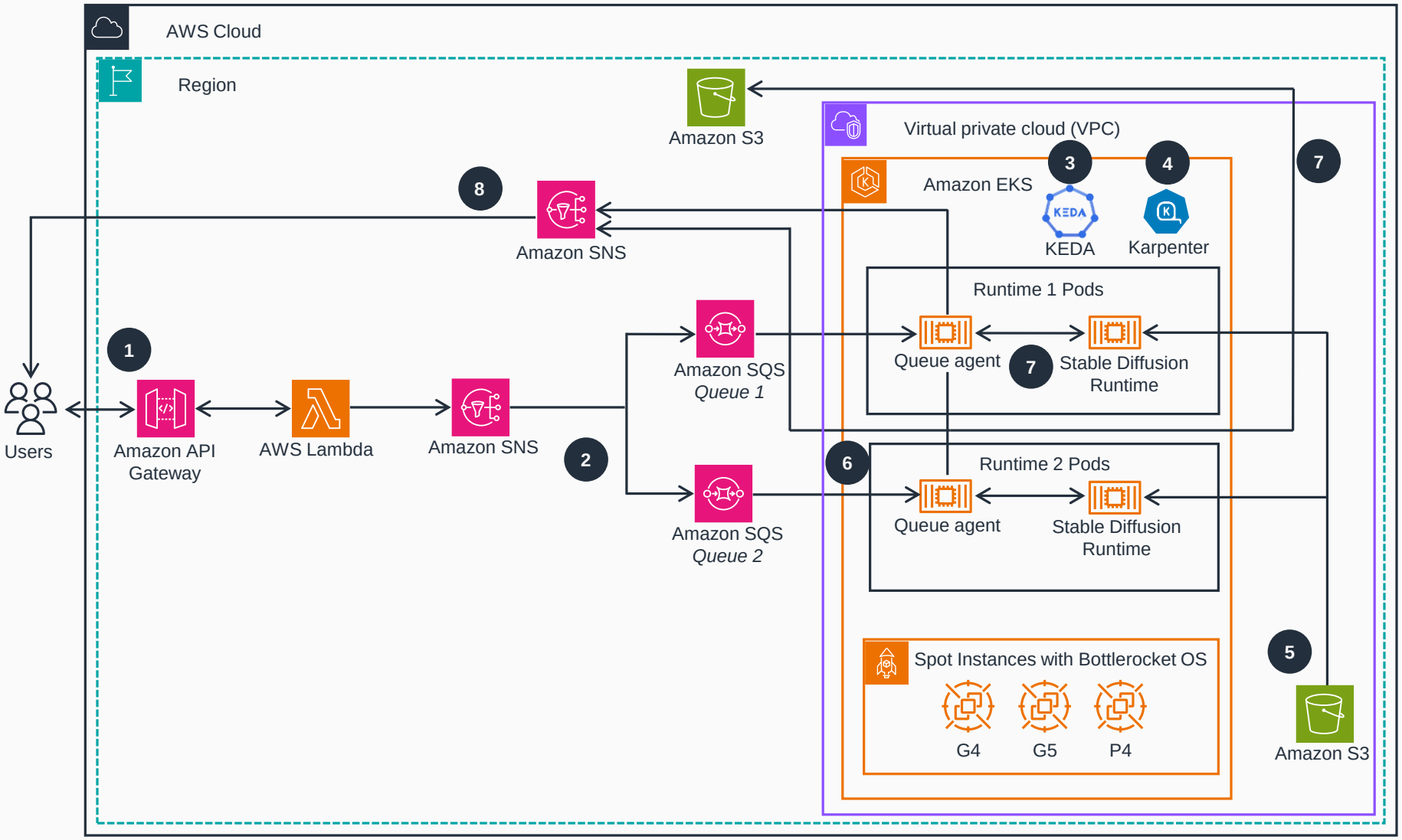
Guidance for Asynchronous Image Generation with Stable Diffusion on AWS

This architecture diagram supports asynchronous machine learning (ML) inferences with a Stable Diffusion web UI using serverless and container services running on Amazon Elastic Kubernetes Service (Amazon EKS).



1 User or application sends a prompt to **Amazon API Gateway** that acts as an endpoint for overall Guidance, including authentication. **AWS Lambda** validates the requests, publishes the requests to the designated **Amazon Simple Notification Service (Amazon SNS)** topic, and immediately returns a response.

2 **Amazon SNS** publishes the message to **Amazon Simple Queue Service (Amazon SQS)** queues. Each message contains a Stable Diffusion (SD) runtime name attribute and will be delivered to the queues with matching SD Runtime names.

3 In the **Amazon Elastic Kubernetes Service (Amazon EKS)** cluster, the previously deployed open source Kubernetes Event Driven Auto-Scaler (KEDA) scales up new pods to process the incoming messages from **Amazon SQS** queues (such as queue 1, queue 2).

4 In the **Amazon EKS** cluster, the previously deployed open source Kubernetes compute auto-scaler, Karpenter, launches new compute nodes based on **Amazon Elastic Compute Cloud (Amazon EC2)** GPU instances (such as g4, g5, and p4) to schedule pending pods. The instances use pre-cached SD Runtime images and are based on **Bottlerocket OS** for faster boot. The instances can be launched using On-Demand or Spot pricing models.

5 Stable Diffusion Runtimes load machine learning (ML) inference model files from an **Amazon Simple Storage Service (Amazon S3)** bucket through the Mountpoint for Amazon S3 CSI Driver upon runtime initialization or on-demand.

6 Queue agents (a software component created for this Guidance) receive messages from **SQS** processing queues and converts them to inputs for SD Runtime API calls.

7 Queue agents call SD Runtime APIs, receive and decode responses, and save the generated images to designated **Amazon S3** buckets.

8 Queue agents send notifications to the designated **SNS** topic. The user receives notifications from the **SNS topic** and can render images.

Reviewed for technical accuracy October 19, 2023

AWS Reference Architecture