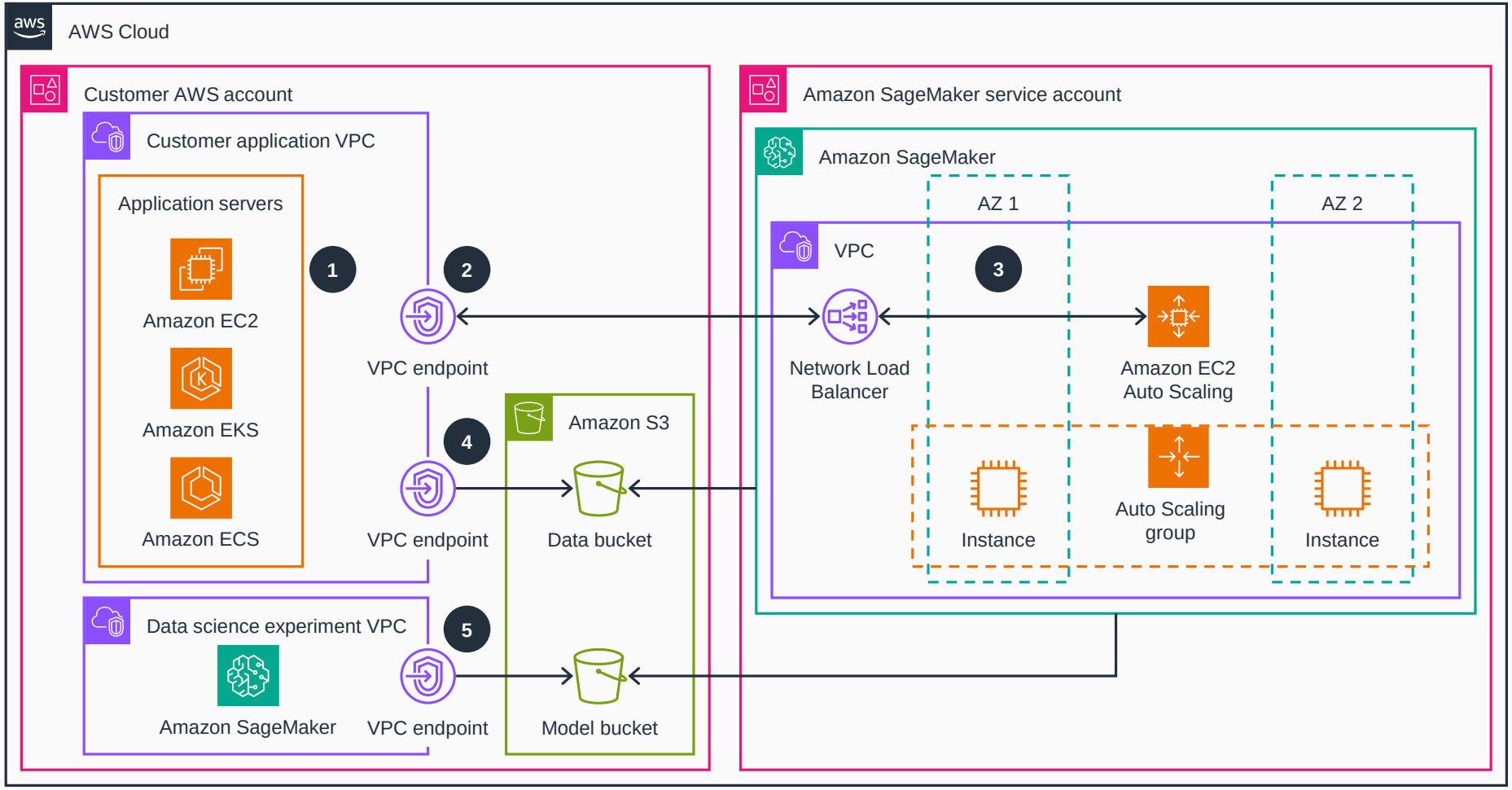
Guidance for Low-Latency, High-Throughput Model Inference Using Amazon SageMaker

This architecture diagram shows how to use Amazon SageMaker to deploy and host machine learning (ML) models while supporting low-latency, high-throughput workloads, such as programmatic advertising and real-time bidding (RTB).



1. A consumer application is deployed within a virtual private cloud (VPC) in your AWS account, using **Amazon Virtual Private Cloud (Amazon VPC)**. This application can be hosted on **Amazon Elastic Compute Cloud (Amazon EC2)** instances or as containers running on either **Amazon Elastic Kubernetes Service (Amazon EKS)** or **Amazon Elastic Container Service (Amazon ECS)**.

2. The consumer application connects to Amazon SageMaker Real-Time inference servers using VPC endpoints powered by AWS PrivateLink. This means that all API calls happen over the private network of AWS and not the public internet, minimizing the latency of the invocations and improving your security posture.

3. The inference requests are routed through a Network Load Balancer to the **SageMaker** real-time inference servers. These servers are hosted across multiple Availability Zones (AZs) within an Amazon EC2 Auto Scaling group. This allows the model inference infrastructure to be elastic and highly available. **SageMaker** real-time inferences provide a choice of **Amazon EC2** instance types. These include **Amazon EC2** Inf1 instances based on **AWS Inferentia**, high-performance machine learning (ML) inference chips designed and built by AWS, and GPU instances, such as **Amazon EC2** G4dn. Multiple hosting options, including shadow testing and an inference recommendation feature in the managed service, reduce operational burden and accelerates time to value.

4. Consumer applications and batch applications use **Amazon Simple Storage Service (Amazon S3)** to store and retrieve data and use it for offline ML training and experiments. Access to **Amazon S3** from the VPC is secured through **PrivateLink**.

5. Data scientists use **SageMaker** to experiment, build, and train the ML model. Once the model is ready, it is saved in **Amazon S3** for the model inference task to load. Access to **Amazon S3** from the VPC is again secured through **PrivateLink**.

Reviewed for technical accuracy March 13, 2024

AWS Reference Architecture