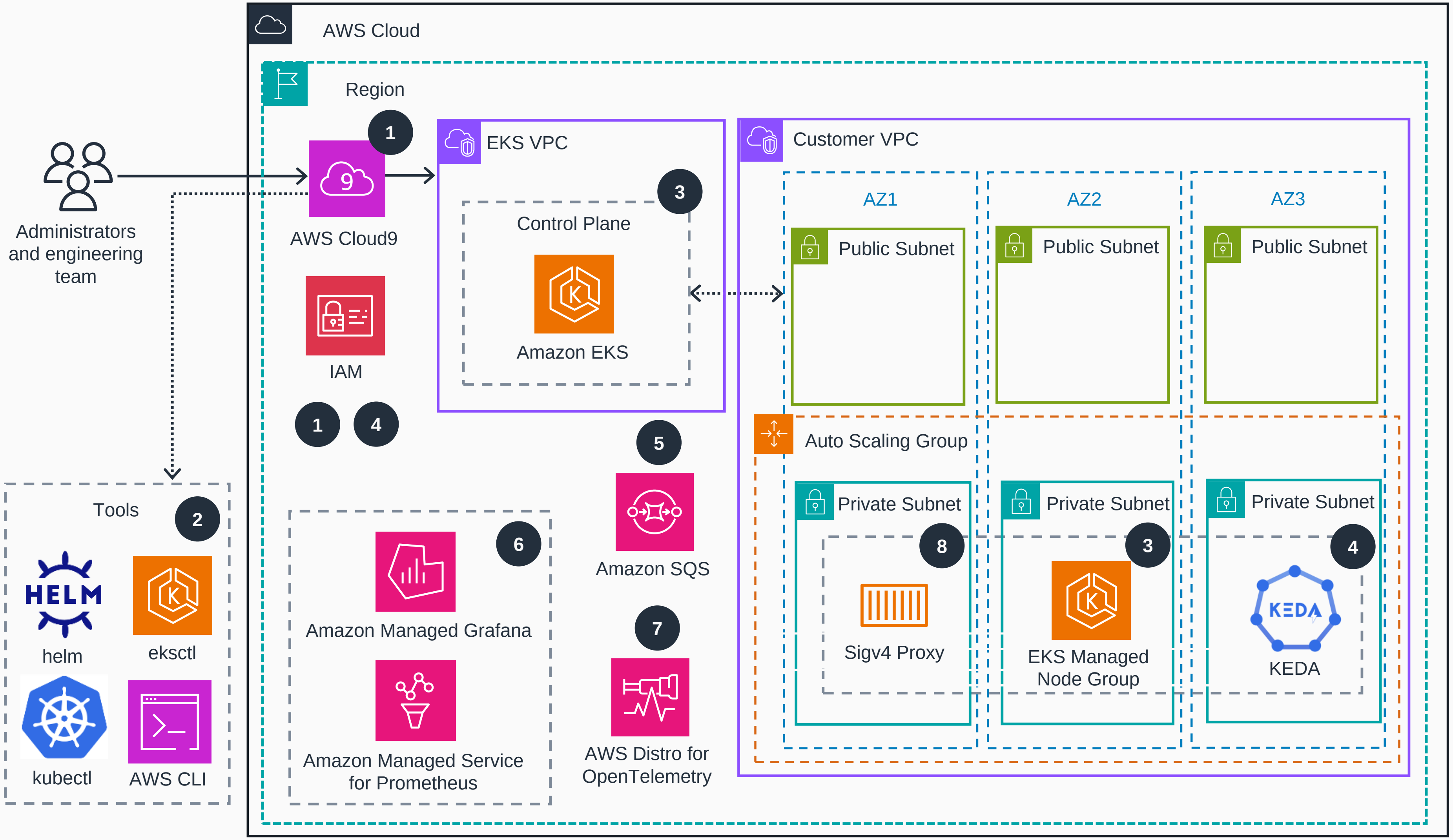
Guidance for Event-Driven Application Autoscaling with KEDA on Amazon EKS – EKS Cluster

This architecture diagram shows how to deploy KEDA on Amazon EKS clusters to improve auto scaling, performance, and cost efficiency.



1. Set up an **AWS Cloud9** environment with **AWS Identity and Access Management (IAM)** permissions.

2. Install helm, eksctl, kubectl, and **AWS Command Line Interface (CLI)** in **AWS Cloud9**.

3. **Amazon Elastic Kubernetes Service (Amazon EKS)** cluster and **EKS** managed node groups are launched through **AWS Cloud9**.

4. KEDA is deployed with the required **IAM** role for service account (IRSA).

5. Deploy **Amazon Simple Queue Service (Amazon SQS)** to decouple communication between applications and attach a policy on KEDA IRSA to access **Amazon SQS**.

6. Create **Amazon Managed Service for Prometheus** and optionally, **Amazon Managed Grafana**.

7. Configure **AWS Distro for OpenTelemetry** to send application metrics to **Amazon Managed Service for Prometheus**, deployed with the required **IAM** IRSA.

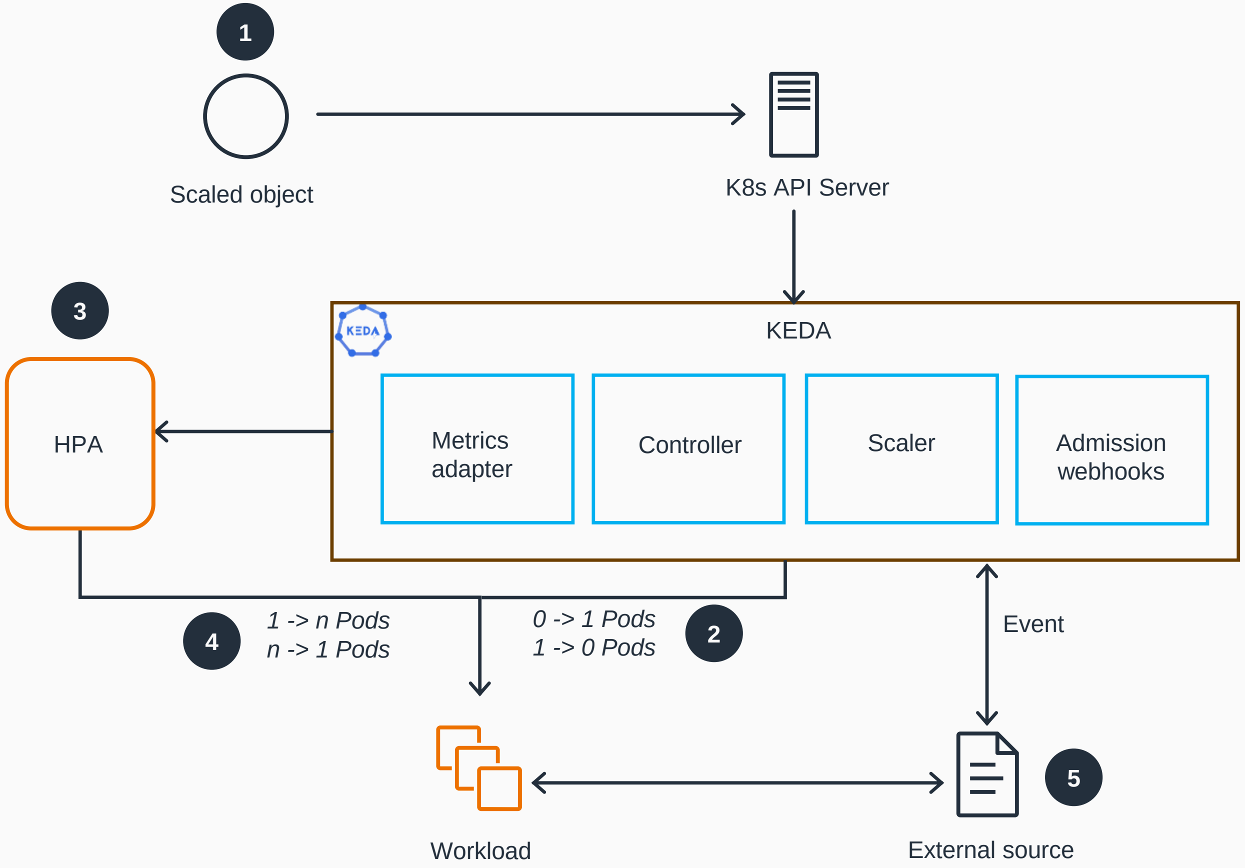
8. Configure the Sigv4 proxy pod to authenticate KEDA with **Amazon Managed Service for Prometheus**, deployed with the required **IAM** IRSA.

Reviewed for technical accuracy May 24, 2024

AWS Reference Architecture

Guidance for Event-Driven Application Autoscaling with KEDA on Amazon EKS – KEDA Overview

This architecture diagram shows an overview of how KEDA components work in conjunction with the Kubernetes Horizontal Pod Autoscaler (HPA) and external event sources.



1 The scaled object is a CustomResourceDefinition (CRD) to configure the event source, deployment to be scaled, and scaling behavior.

2 KEDA activates and deactivates Kubernetes deployments to scale to and from zero on no events. This is one of the primary roles of the keda-operator container that runs when you install KEDA.

3 KEDA feeds custom metrics for Kubernetes Horizontal Pod Autoscaling (HPA) to scale from one to the required amount of pods.

4 HPA scales the pods based on KEDA instructions.

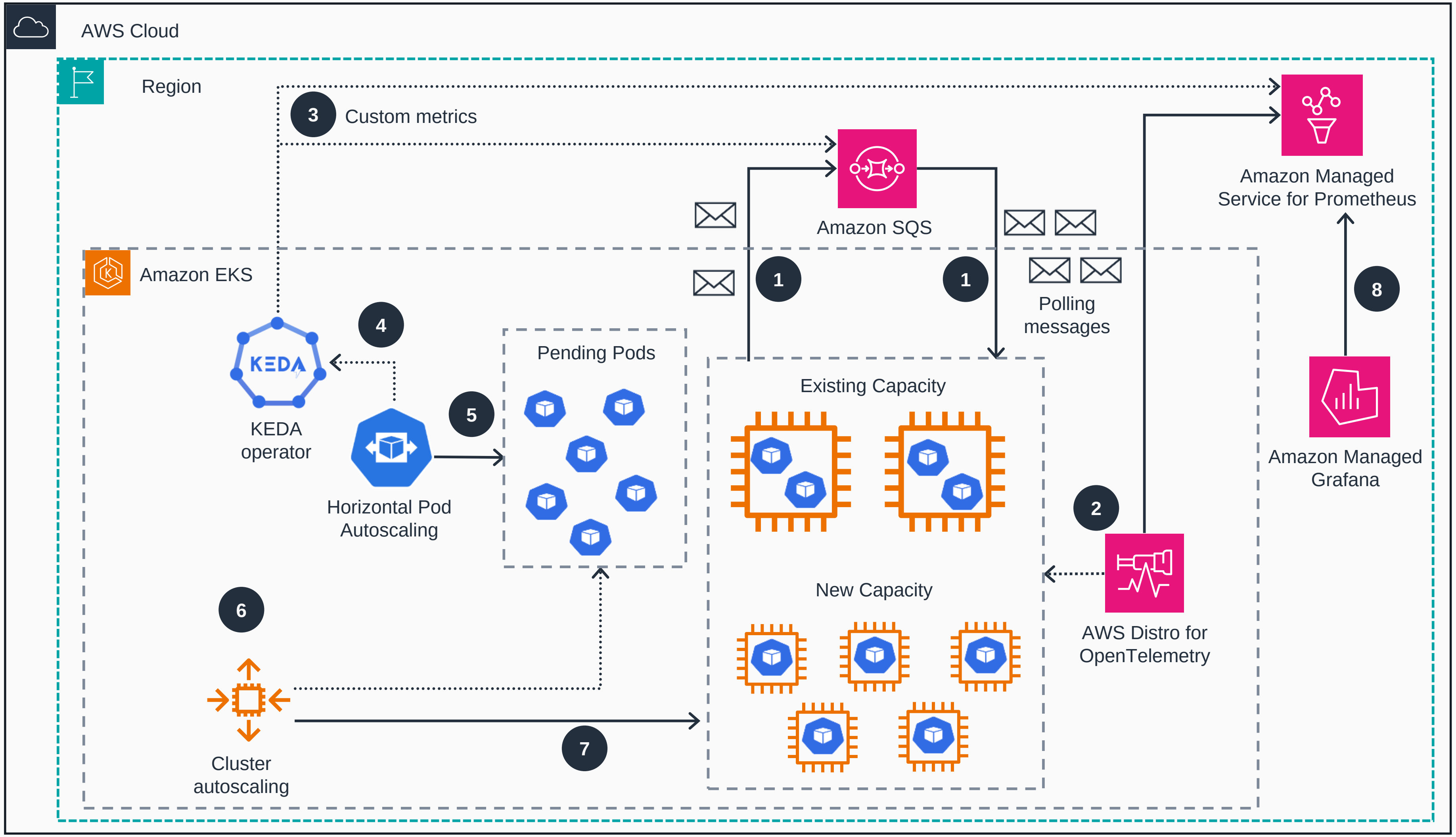
5 KEDA supports more than 60 event sources, available at: Currently available scalers for KEDA.

Reviewed for technical accuracy May 24, 2024

AWS Reference Architecture

Guidance for Event-Driven Application Autoscaling with KEDA on Amazon EKS – Scaling with KEDA

This architecture diagram shows KEDA scaling deployment pods based on custom metrics sources, such as Amazon SQS and Amazon Managed Prometheus.



1. The app uses **Amazon SQS** to decouple communication between microservices

2. **AWS Distro for OpenTelemetry** gets metrics from the application and sends them to **Amazon Managed Service for Prometheus**.

3. KEDA is configured to use **Amazon SQS** and the **Amazon Managed Service for Prometheus** scaler to get **Amazon SQS** queue length and Prometheus custom metrics.

4. KEDA (keda-operator-metrics-apiserver) exposes event data for HPA to scale.

5. HPA scales to the appropriate number of pods.

6. Cluster Autoscaling (CA) provisions the required nodes using auto scaling group. Instead of CA, you can also use Karpenter.

7. New capacity is provisioned as required.

8. You can optionally configure **Amazon Managed Grafana** to show metrics from **Amazon Managed Service for Prometheus** in a dashboard.

Reviewed for technical accuracy May 24, 2024

AWS Reference Architecture