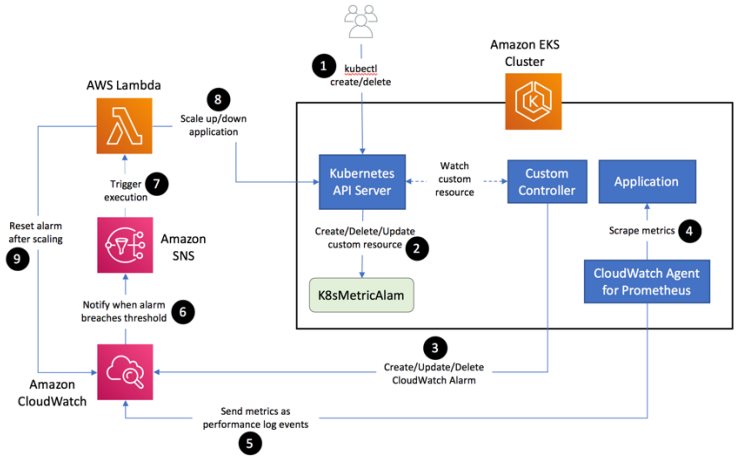
EKS Auto Scaling

30/04/2023

1. EKS auto scaling based on custom Prometheus metrics and CloudWatch Container Insights
https://aws.amazon.com/blogs/containers/autoscaling-amazon-eks-services-based-on-custom-prometheus-metrics-using-cloudwatch-container-insights/

- Custom controller in Kubernetes
- CloudWatch => Lambda => Kubernetes API Server => Custom Controller



2. Kubernetes HPA custom metrics
https://kubernetes.io/docs/tasks/run-application/horizontal-pod-autoscale/#support-for-custom-metrics

- Pod metric, Container metric
- Custom metric
- External metrics – provider adapter

3. EKS Fargate auto scaling based on custom metrics
https://aws.amazon.com/blogs/containers/autoscaling-eks-on-fargate-with-custom-metrics/

- Prometheus adapter queries and expose metrics for Kubernetes custom metrics
- Adapter => External metrics => HPA

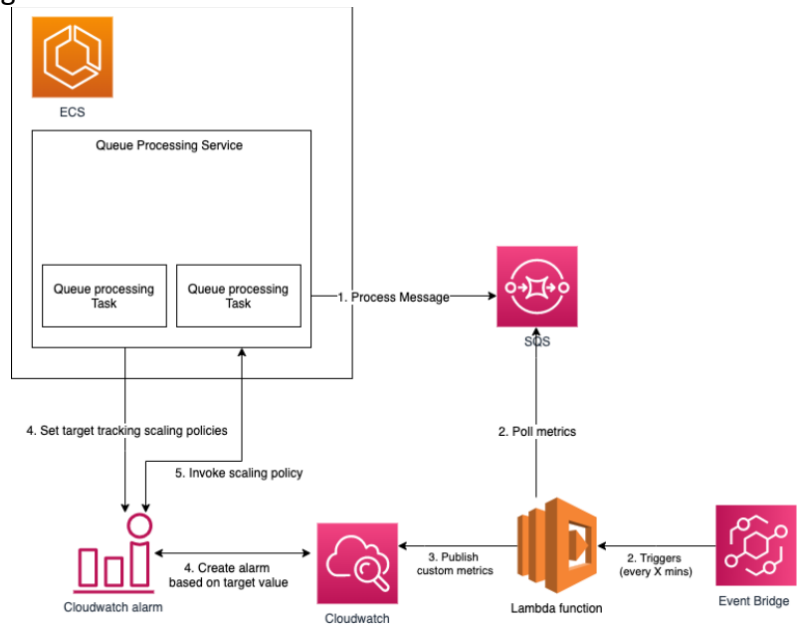
4. EKS scaling based on CloudWatch metrics (CPU, SQS Length, etc.)
https://aws.amazon.com/blogs/compute/scaling-kubernetes-deployments-with-amazon-cloudwatch-metrics/

- HPA with external metrics (SQS length from CloudWatch)
- Need CloudWatch metrics adapter
- CW metrics => Adapter => External metrics => HPA

5. ECS Fargate scaling based on custom metrics

https://aws.amazon.com/blogs/containers/amazon-elastic-container-service-ecs-auto-scaling-using-custom-metrics/

- EventBridge => Lambda => custom metric => ECS task



6. Prometheus adapter to auto scale applications running on EKS
https://aws.amazon.com/blogs/mt/automated-scaling-of-applications-running-on-eks-using-custom-metric-collected-by-amazon-prometheus-using-prometheus-adapter/

- Prometheus => Prometheus Adapter => Kubernetes Custom Metrics => HPA

7. Proactive auto scaling Kubernetes keda metrics
https://aws.amazon.com/blogs/mt/proactive-autoscaling-kubernetes-workloads-keda-metrics-ingested-into-aws-amp/

8. Kubernetes metrics server
https://kubernetes.io/docs/tasks/debug/debug-cluster/resource-metrics-pipeline/

9. Kubernetes client Java and Python (lambda update deployment)
https://github.com/aws-samples/k8s-cloudwatch-operator/tree/main/cloudwatch-lambda
https://github.com/kubernetes-client/python/blob/master/kubernetes/docs/AppsV1Api.md
https://javadoc.io/static/io.kubernetes/client-java/12.0.0/io/kubernetes/client/util/generic/GenericKubernetesApi.html
https://github.com/aws-samples/k8s-cloudwatch-operator/blob/main/cloudwatch-lambda/src/main/java/com/amazonwebservices/blogs/containers/CloudWatchAlarmHandler.java

10. Kubernetes Production chapter 13, page 381

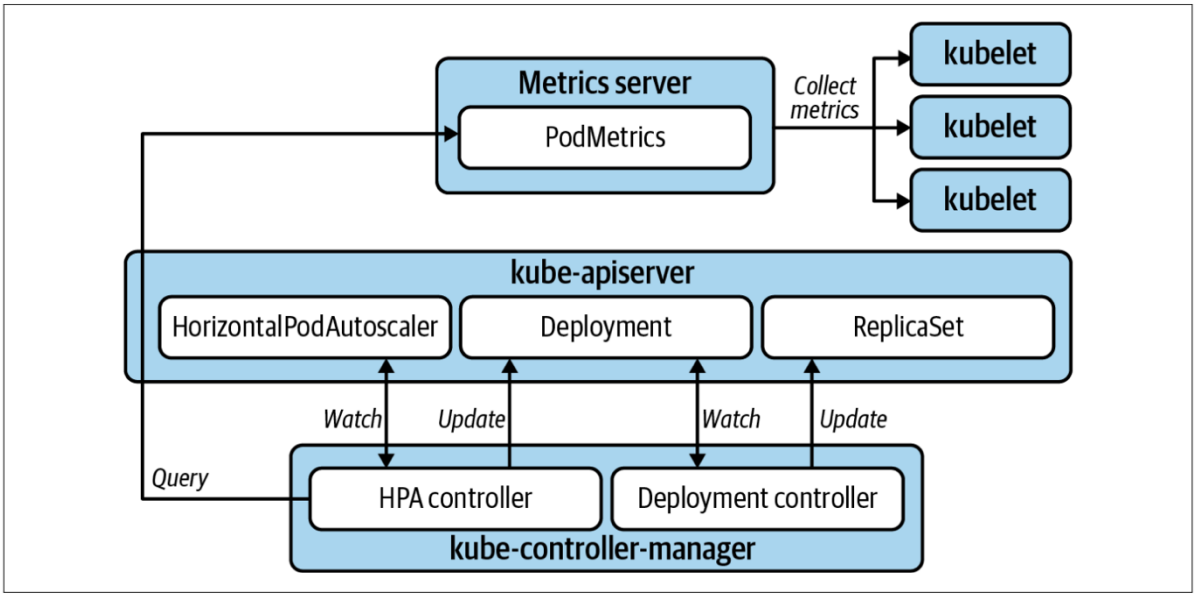


Figure 13-1. Horizontal Pod autoscaling.

Kubernetes Networking Basics

https://kubernetes.io/docs/concepts/cluster-administration/networking/
https://github.com/kubernetes/design-proposals-archive/blob/main/network/networking.md
https://www.redhat.com/sysadmin/kubernetes-pods-communicate-nodes
https://mayankshah.dev/blog/demystifying-kube-proxy/
https://www.redhat.com/sysadmin/kubernetes-pod-network-communications
https://docs.aws.amazon.com/eks/latest/userguide/cni-increase-ip-addresses.html
https://kubernetes.io/docs/reference/networking/virtual-ips/