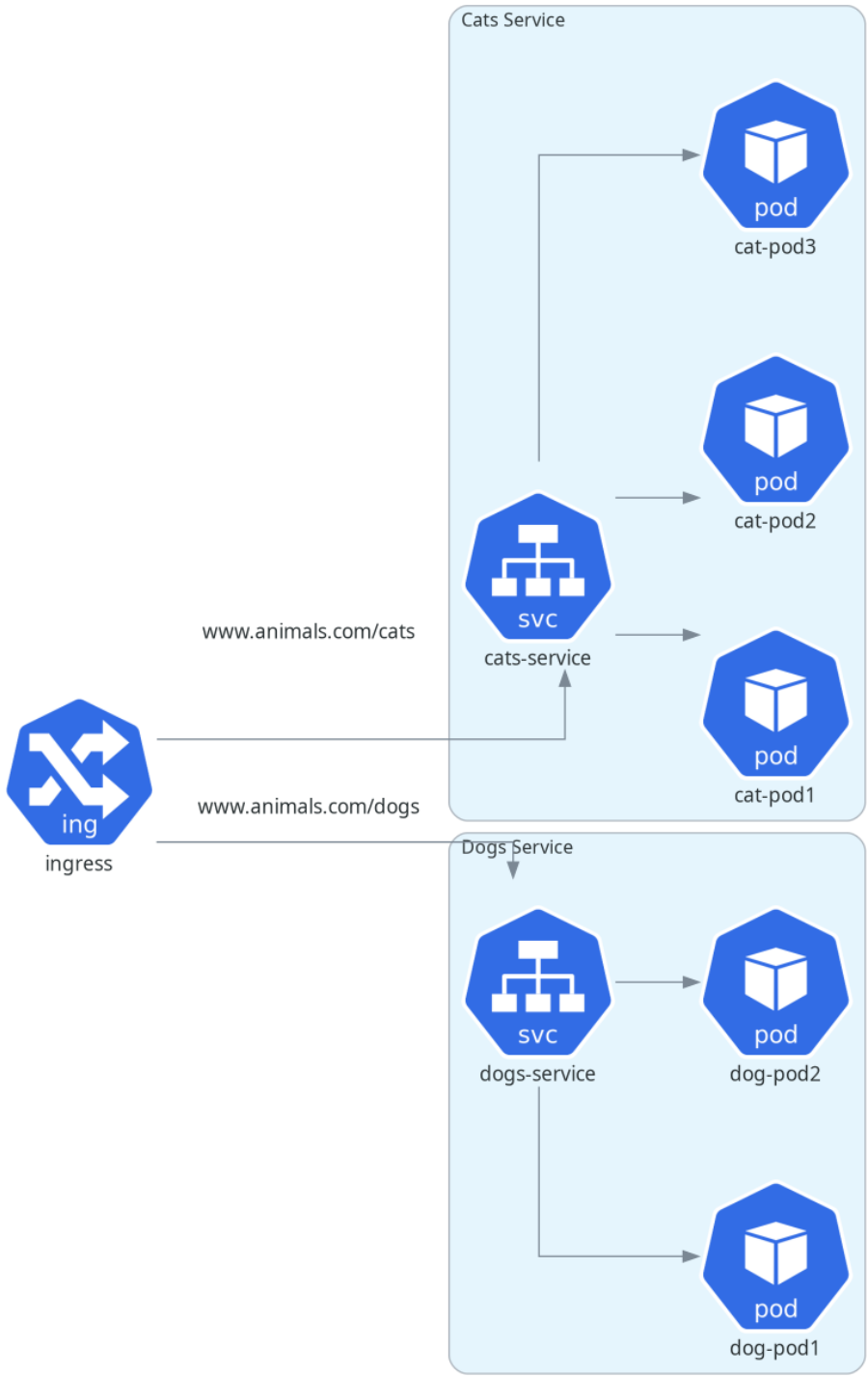
Multiple, auto-scaling services routed through Ingress

In this example we deploy two services, as in the multiple service ingress, but this time we will scale the pods using Horizontal Pod Autoscalers. Traffic will be routed to the services using an Ingress and two temporary pods will be used to generate load for testing purposes.

Explanation

Horizontal Pod Autoscalers are used to automatically scale the number of pods in a deployment based on resouces utilization, such as CPU or memory. In the container's configuration is mandatory to add resources' requests and limits so that the HPA knows the limit with which to scale the pods. Moreover the requests and limits are used by Kubernetes to schedule the pods on the right nodes, so that the pods are not scheduled on nodes that do not have enough resources to run them. If a pod exceeds the limit is terminated.

The resulting system will look like this:



Kubernetes Ingress Routing

Usage

Requirementes and deployment

Make sure you have enabled the Ingress addon in Minikube:

```
minikube addons enable ingress
```

and the metrics server:

```
minikube addons enable metrics-server
```

You can check if the metrics server is running by executing:

```
kubectl get pods -n kube-system | grep metrics-server
```

or

```
kubectl top nodes
```

Apply the configuration files, the `-R` option will recursively search for files in the current directory:

```
kubectl apply -R -f manifests
```

Verification of the deployment

Verify the deployment by checking the pods:

```
kubectl get all
```

You should see one Pod for each service, the Ingress, the two Services and the two HPAs.

Load test

Before the test start watching the HPAs status with:

```
kubectl get hpa --watch
```

Then to execute the load test run the two scripts in the `load` directory in two separate terminals. Watch the HPAs react to the load:

```
$ kubectl get hpa --watch
NAME       REFERENCE                      TARGETS       MINPODS   MAXPODS   REPLICAS
AGE
cats-hpa   Deployment/cats-deployment   cpu: 0%/50%   1         10        1
50m
dogs-hpa   Deployment/dogs-deployment   cpu: 0%/50%   1         10        1
50m
# Beginning load test
cats-hpa   Deployment/cats-deployment   cpu: 26%/50%   1         10        1
51m
dogs-hpa   Deployment/dogs-deployment   cpu: 26%/50%   1         10        1
51m
dogs-hpa   Deployment/dogs-deployment   cpu: 86%/50%   1         10        1
52m
cats-hpa   Deployment/cats-deployment   cpu: 86%/50%   1         10        1
52m
# Load goes over the limit, pods are scaled up
dogs-hpa   Deployment/dogs-deployment   cpu: 86%/50%   1         10        2
52m
cats-hpa   Deployment/cats-deployment   cpu: 86%/50%   1         10        2
52m
cats-hpa   Deployment/cats-deployment   cpu: 53%/50%   1         10        2
53m
dogs-hpa   Deployment/dogs-deployment   cpu: 53%/50%   1         10        2
53m
# End of load test, the load goes under the limit
cats-hpa   Deployment/cats-deployment   cpu: 43%/50%   1         10        2
54m
dogs-hpa   Deployment/dogs-deployment   cpu: 43%/50%   1         10        2
54m
cats-hpa   Deployment/cats-deployment   cpu: 0%/50%    1         10        2
55m
dogs-hpa   Deployment/dogs-deployment   cpu: 0%/50%    1         10        2
55m
cats-hpa   Deployment/cats-deployment   cpu: 0%/50%    1         10        2
59m
dogs-hpa   Deployment/dogs-deployment   cpu: 0%/50%    1         10        2
59m
# After the stabilization window the pods are scaled down
cats-hpa   Deployment/cats-deployment   cpu: 0%/50%    1         10        1
60m
dogs-hpa   Deployment/dogs-deployment   cpu: 0%/50%    1         10        1
60m
```

Notes about the deployment

The replica count in the deployments configuration is removed so that the Horizontal Pod Autoscaler can manager the number of replicas. However it is not necessary to remove it: Kubernetes will spawn the number

of pods specified and then the HPA will take control of the scaling, this may not be desired and could be troublesome when an HPA is active, resulting in thrashing or flapping behavior. For more information about the behavior of the HPAs in this situation check the official documentation.