COMPARISON OF OPEN SOURCE SERVERLESS PLATFORMS

AKSHAT HARDIKBHAI SHAH - ashah004
RUTANSH MANISHKUMAR SUTHAR - rsuth004

GOALS

Evaluating OpenFaas and KNative under Serverless ML Workload

PRESENTATION OUTLINE

1. Describe Workload
2. Overview and Setup of OpenFaaS & KNative
3. Performance Comparisions and Results
4. Important Links

WORKLOAD SETUP

WORKLOAD SETUP

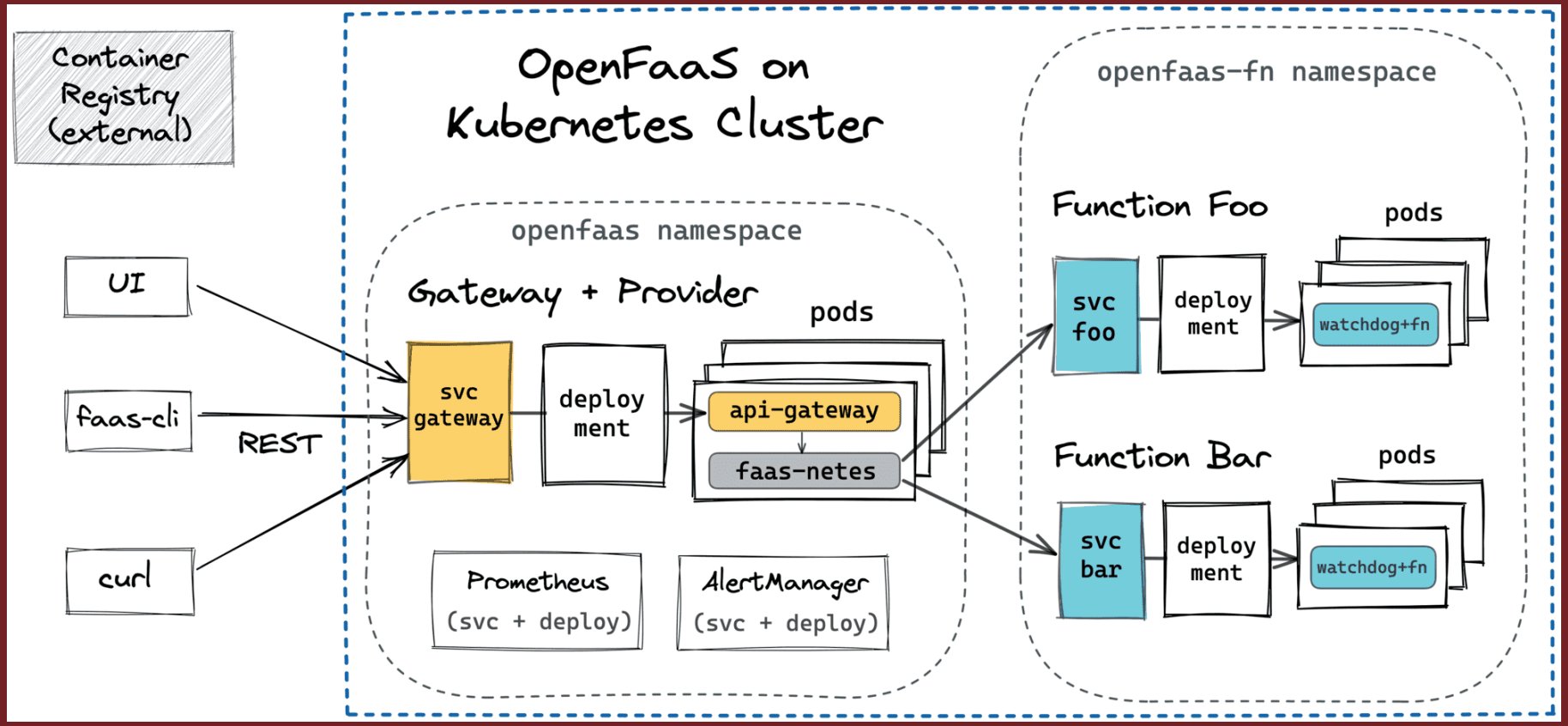
- *Input / Output : A Simple Image is the input, and it returns* JSON response with top 5 predictions and confidence scores.
- *Type : Python Based Serverless Function Running with Flask.*
- Goal : Image classification using a pre-trained MobileNetV2 model from TensorFlow.
- Steps
  - Image Pre Processing
  - Image Classification
  - Input Output Handling
  - Error Handling
- Uses Docker, Python, TensorFlow, MobileNetV2 (Pretrained Weights), Flask, JSON

OVERVIEW OF OPENFAAS

OPENFAAS

- The Community Edition is Open Source
- Lightweight and Easy to Deploy
- Supports Multiple Orchestrators
- Because it is the Community Edition, it:
- Lacks Features like Scale to Zero.
- Limited number of functions, replicas (autoscaling) and metrics.

OPENFAAS


Container Registry (external)
OpenFaaS on Kubernetes Cluster
openfaas-fn namespace
openfaas namespace
Gateway + Provider
pods
UI
faas-cli
REST
curl
svc gateway
deploy ment
api-gateway
faas-netes
Prometheus (svc + deploy)
AlertManager (svc + deploy)
Function Foo
pods
svc foo
deploy ment
watchdog+fn
Function Bar
pods
svc bar
deploy ment
watchdog+fn

OPENFAAS SETUP

- Infrastructure:
  - *CloudLab Environment: 3 Nodes (1 Control Plane and 2 Worker).*
  - *Kubernetes Cluster: Setup using kubeadm and CRI-O container runtime.*
- OpenFaas Installation:
  - *Installed OpenFaaS using Arkade.*
  - *Arkade also installed Grafana and Prometheus.*
- Docker
- ML Workload Deployment
- WRK Benchmarking

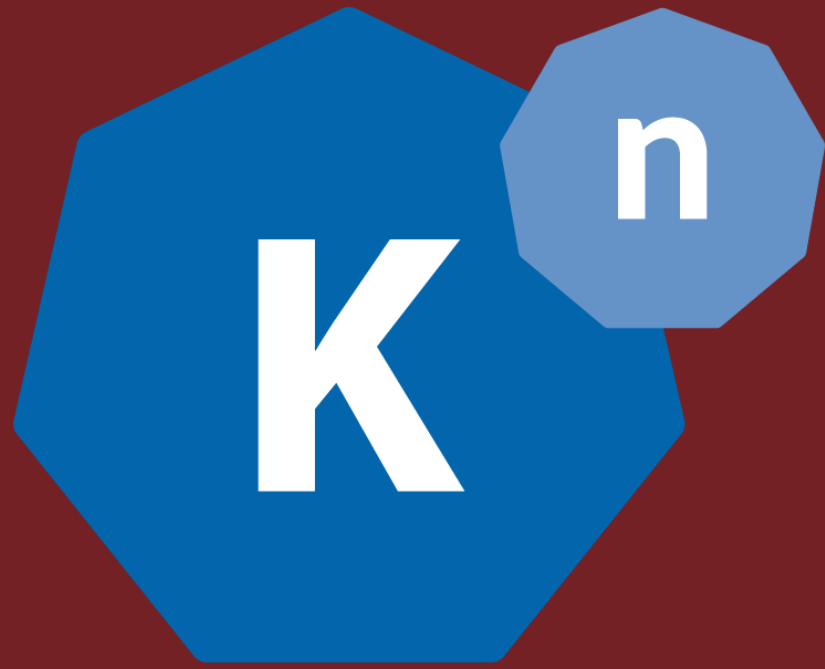
OPENFAAS

- OpenFaaS Orchestrates the functions in the openfaas-fn namespace. It uses its own autoscaler and that will create another pod in this namespace per replica. But the community version autoscaler limits to 5 replicas that is also based on the requests per second.
- It has its own openfaas namespace to host all the pods it need to function as is shown.
- They are all created and linked automatically when installing openfaas with the specific flags. We just need to expose the required services.

```
rsuthar@node0:~$ kubectl get pods -n openfaas-fn | grep image-classifier-new*
image-classifier-new-77f95ddbf5-5v7rg   1/1     Running   0          65m
rsuthar@node0:~$
```

```
rsuthar@node0:~$ kubectl get pods -n openfaas
NAME                                READY   STATUS    RESTARTS      AGE
alertmanager-5948f75c9d-48p26       1/1     Running   0             17h
gateway-7fcc449479-twp6p            2/2     Running   2 (17h ago)   17h
grafana-5974ccbc87-pmrh7            1/1     Running   0             16h
metrics-server-8445b4c4b8-422vj     1/1     Running   0             155m
nats-75958fd77b-9zpsj               1/1     Running   0             17h
prometheus-7d4874c5-hgcs6           1/1     Running   0             39m
queue-worker-5f977b86db-clbnz       1/1     Running   0             17h
```

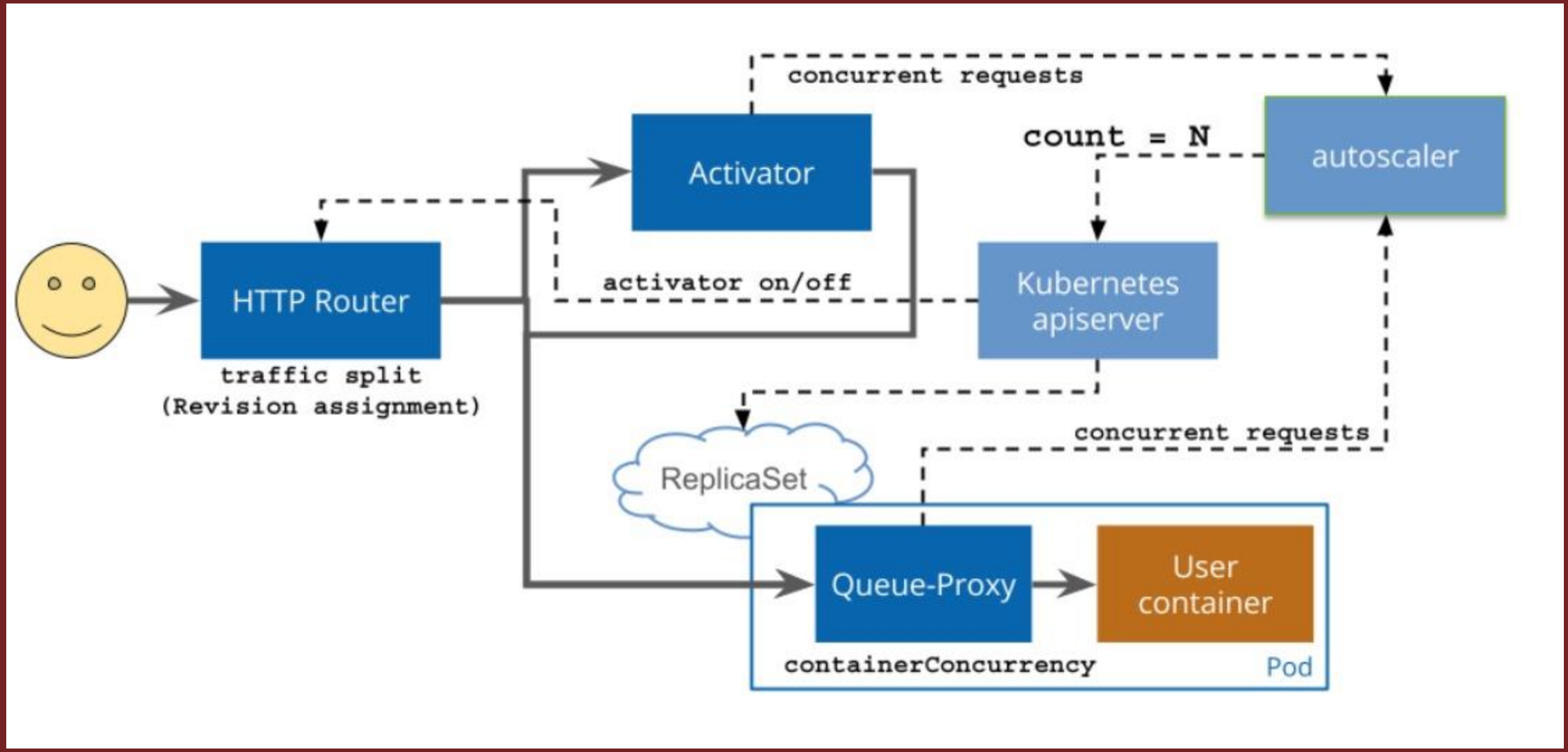
OVERVIEW OF KNATIVE



FEATURES OF KNATIVE

- Autoscaling – Automatically scales workloads up/down, including scale-to-zero.
- Traffic Management – Supports traffic splitting, blue-green, and canary deployments.
- Event-driven Architecture – Seamless event production, routing, and consumption.
- Pluggable Networking – Works with Istio, Contour, Kourier, etc.
- Portability – Runs on any Kubernetes cluster, cloud or on-premises.
- Developer-friendly – Simplifies deployment with minimal configuration.
- Flexible Observability – Supports logging, monitoring, and tracing tools.

KNATIVE


concurrent requests
count = N
autoscaler
Activator
HTTP Router
activator on/off
Kubernetes
apiserver
traffic split
(Revision assignment)
concurrent requests
ReplicaSet
Queue-Proxy
User
container
containerConcurrency
Pod

```
root@node0:/users/ashah004# kubectl get all -n metallb-system
NAME                              READY   STATUS    RESTARTS   AGE
pod/controller-586fdc8ccf-q9zvj   1/1     Running   0          10d
pod/speaker-47rcz                 1/1     Running   0          10d
pod/speaker-w5brr                 1/1     Running   0          10d

NAME                      DESIRED   CURRENT   READY   UP-TO-DATE   AVAILABLE   NODE SELECTOR            AGE
daemonset.apps/speaker    2         2         2       2            2           kubernetes.io/os=linux   10d

NAME                         READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/controller   1/1     1            1           10d

NAME                                    DESIRED   CURRENT   READY   AGE
replicaset.apps/controller-586fdc8ccf   1         1         1       10d
root@node0:/users/ashah004#
```

```
root@node0:/users/ashah004# kubectl get all -n istio-system
NAME                                    READY   STATUS    RESTARTS   AGE
pod/istio-ingressgateway-5d9657df6-2vtrk   1/1     Running   0          10d
pod/istio-ingressgateway-5d9657df6-6ptw8   1/1     Running   0          10d
pod/istio-ingressgateway-5d9657df6-xml22   1/1     Running   0          10d
pod/istiod-ddcf4fdd9-7dkbp              1/1     Running   0          10d
pod/istiod-ddcf4fdd9-886hp              1/1     Running   0          10d
pod/istiod-ddcf4fdd9-9khw5              1/1     Running   0          10d

NAME                           TYPE           CLUSTER-IP      EXTERNAL-IP       PORT(S)                                      AGE
service/istio-ingressgateway   LoadBalancer   10.110.9.133    130.127.133.200   15021:31830/TCP,80:31009/TCP,443:30772/TCP   10d
service/istiod                 ClusterIP      10.98.212.56    <none>            15010/TCP,15012/TCP,443/TCP,15014/TCP        10d
service/knative-local-gateway  ClusterIP      10.110.113.29   <none>            80/TCP,443/TCP                               10d

NAME                                READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/istio-ingressgateway   3/3     3            3           10d
deployment.apps/istiod              3/3     3            3           10d

NAME                                          DESIRED   CURRENT   READY   AGE
replicaset.apps/istio-ingressgateway-5d9657df6   3         3         3       10d
replicaset.apps/istiod-ddcf4fdd9              3         3         3       10d

NAME                                        REFERENCE          TARGETS              MINPODS   MAXPODS   REPLICAS   AGE
horizontalpodautoscaler.autoscaling/istiod  Deployment/istiod  cpu: <unknown>/60%   3         10        3          10d
root@node0:/users/ashah004#
```

```
root@node0:/users/ashah004# kubectl get all -n knative-serving
NAME                                        READY   STATUS    RESTARTS      AGE
pod/activator-cc64979b9-6s4j2               2/2     Running   9 (44m ago)   2d
pod/autoscaler-6b8db7c449-98cvj             1/1     Running   0             10d
pod/controller-775f8576cc-xvrcn             1/1     Running   0             10d
pod/istio-webhook-76957d65bb-gdw4q          1/1     Running   0             10d
pod/net-istio-controller-75c76d7475-cwbtx   1/1     Running   0             10d
pod/net-istio-webhook-6c8c5986d-97smq       2/2     Running   0             2d1h
pod/networking-istio-5c4976c565-n4975       1/1     Running   0             10d
pod/webhook-694c5d68b7-zfd25                1/1     Running   0             10d

NAME                                TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)                                  AGE
service/activator-service           ClusterIP   10.99.73.2       <none>        9090/TCP,8008/TCP,80/TCP,81/TCP,443/TCP  10d
service/autoscaler                  ClusterIP   10.108.91.192    <none>        9090/TCP,8008/TCP,8080/TCP               10d
service/autoscaler-bucket-00-of-01  ClusterIP   10.104.47.242    <none>        8080/TCP                                 10d
service/controller                  ClusterIP   10.103.37.54     <none>        9090/TCP,8008/TCP                        10d
service/istio-webhook               ClusterIP   10.100.231.165   <none>        9090/TCP,8008/TCP,443/TCP                10d
service/net-istio-webhook           ClusterIP   10.99.170.65     <none>        9090/TCP,8008/TCP,443/TCP                10d
service/webhook                     ClusterIP   10.102.1.146     <none>        9090/TCP,8008/TCP,443/TCP                10d

NAME                                READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/activator           1/1     1            1           10d
deployment.apps/autoscaler          1/1     1            1           10d
deployment.apps/controller          1/1     1            1           10d
deployment.apps/istio-webhook       1/1     1            1           10d
deployment.apps/net-istio-controller   1/1     1            1           10d
deployment.apps/net-istio-webhook   1/1     1            1           10d
deployment.apps/networking-istio    1/1     1            1           10d
deployment.apps/webhook             1/1     1            1           10d

NAME                                          DESIRED   CURRENT   READY   AGE
replicaset.apps/activator-cc64979b9           1         1         1       2d
replicaset.apps/activator-fd687d67            0         0         0       10d
replicaset.apps/autoscaler-6b8db7c449         1         1         1       10d
replicaset.apps/controller-775f8576cc         1         1         1       10d
replicaset.apps/istio-webhook-76957d65bb      1         1         1       10d
replicaset.apps/net-istio-controller-75c76d7475   1         1         1       10d
replicaset.apps/net-istio-webhook-5dd7568545      0         0         0       2d3h
replicaset.apps/net-istio-webhook-675dd68fdf      0         0         0       2d2h
replicaset.apps/net-istio-webhook-6c8c5986d       1         1         1       2d1h
replicaset.apps/net-istio-webhook-796db555d6      0         0         0       10d
replicaset.apps/networking-istio-5c4976c565       1         1         1       10d
replicaset.apps/webhook-694c5d68b7            1         1         1       10d

NAME                                                  REFERENCE              TARGETS              MINPODS   MAXPODS   REPLICAS   AGE
horizontalpodautoscaler.autoscaling/activator         Deployment/activator   cpu: <unknown>/100%  1         20        1          10d
horizontalpodautoscaler.autoscaling/webhook           Deployment/webhook     cpu: <unknown>/100%  1         5         1          10d
```

```
root@node0:/users/ashah004# kubectl get all -n knative-eventing
NAME                                        READY   STATUS    RESTARTS   AGE
pod/eventing-controller-57cd9b767f-xqv7x    1/1     Running   0          23h
pod/eventing-webhook-c7fd5c458-mrzx7        1/1     Running   0          23h
pod/job-sink-57975ccb69-85qkb               1/1     Running   0          23h
pod/kafka-broker-receiver-c7997d4d9-7w2vf   1/1     Running   0          20h
pod/kafka-channel-receiver-6769d4885b-wtjkk 1/1     Running   0          20h
pod/kafka-controller-5b969c7d5d-m2jrs       1/1     Running   0          20h
pod/kafka-webhook-eventing-6fb4564895-c5fcl 1/1     Running   0          20h

NAME                              TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)                                 AGE
service/eventing-webhook          ClusterIP   10.96.112.101    <none>        443/TCP                                 23h
service/job-sink                  ClusterIP   10.105.189.208   <none>        80/TCP,443/TCP,9092/TCP                 23h
service/kafka-broker-ingress      ClusterIP   10.106.7.239     <none>        80/TCP,443/TCP,8080/TCP,8443/TCP,9090/TCP   20h
service/kafka-channel-ingress     ClusterIP   10.102.14.57     <none>        80/TCP,443/TCP,8080/TCP,9090/TCP        20h
service/kafka-webhook-eventing    ClusterIP   10.108.219.193   <none>        443/TCP,9090/TCP                        20h

NAME                                    READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/eventing-controller     1/1     1            1           23h
deployment.apps/eventing-webhook        1/1     1            1           23h
deployment.apps/job-sink                1/1     1            1           23h
deployment.apps/kafka-broker-receiver   1/1     1            1           20h
deployment.apps/kafka-channel-receiver  1/1     1            1           20h
deployment.apps/kafka-controller        1/1     1            1           20h
deployment.apps/kafka-webhook-eventing  1/1     1            1           20h
deployment.apps/pingsource-mt-adapter   0/0     0            0           23h

NAME                                              DESIRED   CURRENT   READY   AGE
replicaset.apps/eventing-controller-57cd9b767f    1         1         1       23h
replicaset.apps/eventing-webhook-c7fd5c458        1         1         1       23h
replicaset.apps/job-sink-57975ccb69               1         1         1       23h
replicaset.apps/kafka-broker-receiver-c7997d4d9   1         1         1       20h
replicaset.apps/kafka-channel-receiver-6769d4885b 1         1         1       20h
replicaset.apps/kafka-controller-5b969c7d5d       1         1         1       20h
replicaset.apps/kafka-webhook-eventing-6fb4564895 1         1         1       20h
replicaset.apps/pingsource-mt-adapter-646cfcfccd  0         0         0       23h

NAME                                    READY   AGE
statefulset.apps/kafka-broker-dispatcher   0/0     20h
statefulset.apps/kafka-channel-dispatcher  0/0     20h

NAME                                                        REFERENCE                      TARGETS              MINPODS   MAXPODS   REPLICAS   AGE
horizontalpodautoscaler.autoscaling/eventing-webhook   Deployment/eventing-webhook   cpu: <unknown>/100%   1         5         1          23h
```

```
root@node0:/users/ashah004# kubectl get all -n default
NAME                                                     READY   STATUS    RESTARTS      AGE
pod/alertmanager-prometheus-kube-prometheus-alertmanager-0   2/2     Running   0             5d7h
pod/prometheus-grafana-54d864bf96-fckgj                  3/3     Running   0             5d7h
pod/prometheus-kube-prometheus-operator-dd4b85cc8-stlr8  1/1     Running   0             5d7h
pod/prometheus-kube-state-metrics-55c78bd9d5-cqz2q       1/1     Running   2 (20h ago)   5d7h
pod/prometheus-prometheus-kube-prometheus-prometheus-0   2/2     Running   0             5d7h
pod/prometheus-prometheus-node-exporter-9421r            1/1     Running   0             5d7h
pod/prometheus-prometheus-node-exporter-qxwt2            1/1     Running   0             5d7h
pod/prometheus-prometheus-node-exporter-xk2lp            1/1     Running   0             5d7h

NAME                                              TYPE        CLUSTER-IP       EXTERNAL-IP   PORT(S)                      AGE
service/alertmanager-operated                     ClusterIP   None             <none>        9093/TCP,9094/TCP,9094/UDP   9d
service/kubernetes                                ClusterIP   10.96.0.1        <none>        443/TCP                      10d
service/prometheus-grafana                        ClusterIP   10.98.219.32     <none>        80/TCP                       9d
service/prometheus-kube-prometheus-alertmanager   ClusterIP   10.100.184.248   <none>        9093/TCP,8080/TCP            9d
service/prometheus-kube-prometheus-operator       ClusterIP   10.108.126.190   <none>        443/TCP                      9d
service/prometheus-kube-prometheus-prometheus     ClusterIP   10.107.165.95    <none>        9090/TCP,8080/TCP            9d
service/prometheus-kube-state-metrics             ClusterIP   10.105.131.2     <none>        8080/TCP                     9d
service/prometheus-operated                       ClusterIP   None             <none>        9090/TCP                     9d
service/prometheus-prometheus-node-exporter       ClusterIP   10.105.144.29    <none>        9100/TCP                     9d

NAME                                                DESIRED   CURRENT   READY   UP-TO-DATE   AVAILABLE   NODE SELECTOR            AGE
daemonset.apps/prometheus-prometheus-node-exporter  3         3         3       3            3           kubernetes.io/os=linux   9d

NAME                                                READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/prometheus-grafana                  1/1     1            1           9d
deployment.apps/prometheus-kube-prometheus-operator 1/1     1            1           9d
deployment.apps/prometheus-kube-state-metrics       1/1     1            1           9d

NAME                                                      DESIRED   CURRENT   READY   AGE
replicaset.apps/prometheus-grafana-54d864bf96             1         1         1       9d
replicaset.apps/prometheus-kube-prometheus-operator-dd4b85cc8   1         1         1       9d
replicaset.apps/prometheus-kube-state-metrics-55c78bd9d5  1         1         1       9d

NAME                                                    READY   AGE
statefulset.apps/alertmanager-prometheus-kube-prometheus-alertmanager   1/1     9d
statefulset.apps/prometheus-prometheus-kube-prometheus-prometheus       1/1     9d
```

```
root@node0:/users/ashah004# kubectl get all -n demo
NAME                                          TYPE           CLUSTER-IP       EXTERNAL-IP                                                  PORT(S)                                          AG
service/logger-service                        ExternalName   <none>           knative-local-gateway.istio-system.svc.cluster.local         80/TCP                                           15
service/logger-service-00001                  ClusterIP      10.111.93.28     <none>                                                       80/TCP,443/TCP                                   15
service/logger-service-00001-private          ClusterIP      10.106.115.121   <none>                                                       80/TCP,443/TCP,9090/TCP,9091/TCP,8022/TCP,8012/TCP   15
service/ml-image-classifier                   ExternalName   <none>           knative-local-gateway.istio-system.svc.cluster.local         80/TCP                                           20
service/ml-image-classifier-00001             ClusterIP      10.105.28.242    <none>                                                       80/TCP,443/TCP                                   20
service/ml-image-classifier-00001-private     ClusterIP      10.110.251.78    <none>                                                       80/TCP,443/TCP,9090/TCP,9091/TCP,8022/TCP,8012/TCP   20

NAME                                                READY   UP-TO-DATE   AVAILABLE   AGE
deployment.apps/logger-service-00001-deployment     0/0     0            0           15h
deployment.apps/ml-image-classifier-00001-deployment   0/0     0            0           20h

NAME                                                           DESIRED   CURRENT   READY   AGE
replicaset.apps/logger-service-00001-deployment-847d5f4b8b     0         0         0       15h
replicaset.apps/ml-image-classifier-00001-deployment-65bd9c5df5   0         0         0       20h

NAME                                  URL   AGE   READY   REASON
broker.eventing.knative.dev/default         20h

NAME                                        BROKER    SUBSCRIBER_URI   AGE   READY   REASON
trigger.eventing.knative.dev/log-ml-predictions   default                    20h

NAME                                              LATESTCREATED              LATESTREADY                READY   REASON
configuration.serving.knative.dev/logger-service        logger-service-00001       logger-service-00001       True
configuration.serving.knative.dev/ml-image-classifier   ml-image-classifier-00001   ml-image-classifier-00001   True

NAME                                                CONFIG NAME          GENERATION   READY   REASON   ACTUAL REPLICAS   DESIRED REPLICAS
revision.serving.knative.dev/logger-service-00001       logger-service       1            True             0                 0
revision.serving.knative.dev/ml-image-classifier-00001   ml-image-classifier   1            True             0                 0

NAME                                        URL                                                        READY   REASON
route.serving.knative.dev/logger-service        http://logger-service.demo.130.127.133.200.xip.io          True
route.serving.knative.dev/ml-image-classifier   http://ml-image-classifier.demo.130.127.133.200.xip.io     True

NAME                                          URL                                                        LATESTCREATED              LATESTREADY                READY   REASON
service.serving.knative.dev/logger-service        http://logger-service.demo.130.127.133.200.xip.io          logger-service-00001       logger-service-00001       True
service.serving.knative.dev/ml-image-classifier   http://ml-image-classifier.demo.130.127.133.200.xip.io     ml-image-classifier-00001   ml-image-classifier-00001   True
root@node0:/users/ashah004#
```

KNATIVE SETUP

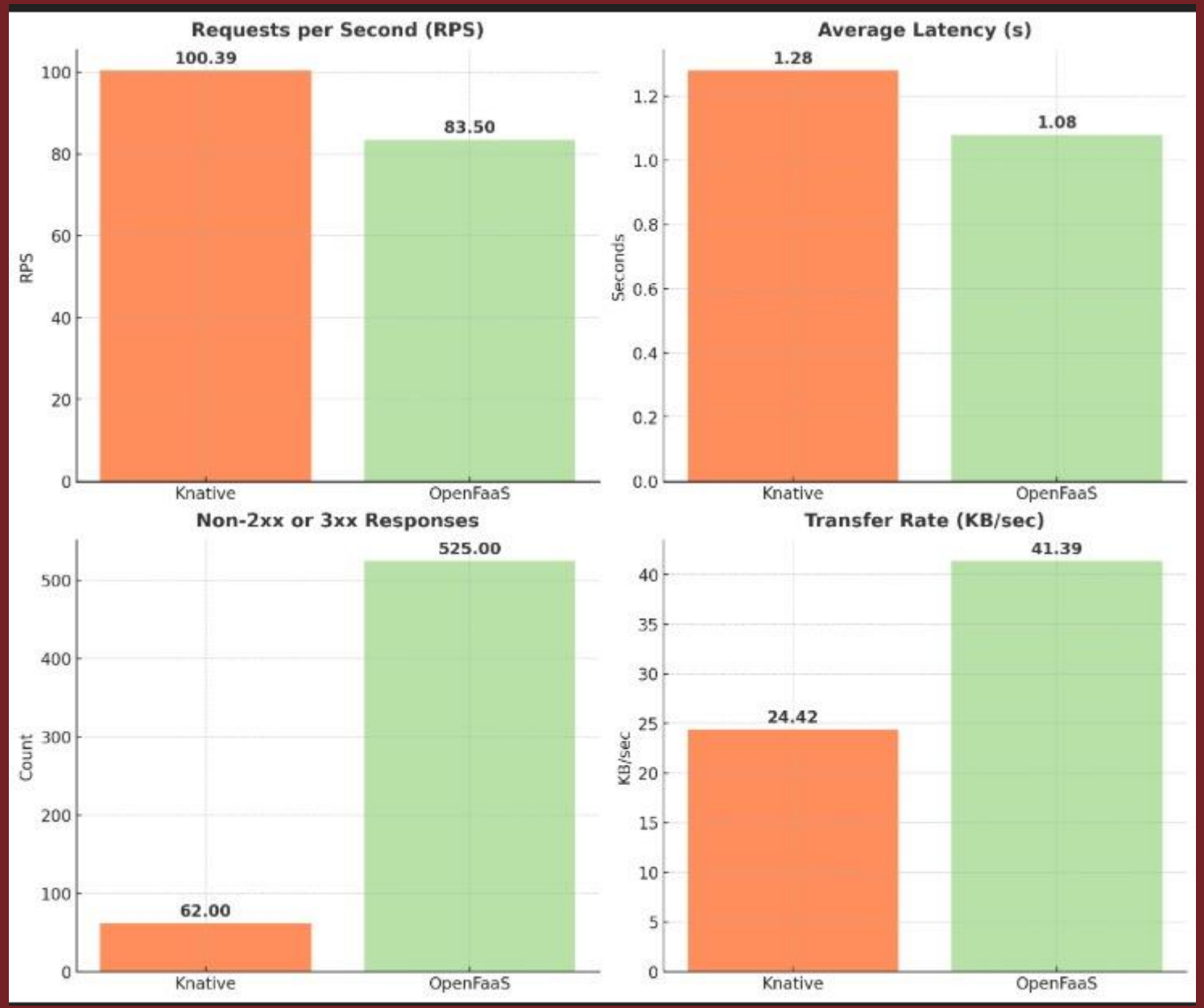
- Infrastructure:
  - *CloudLab Environment: 3 Nodes (1 Control Plane and 2 Worker).*
  - *Kubernetes Cluster: Setup using kubeadm and CRI-O container runtime.*
- MetalLB Load Balancer:
  - Deployed MetalLB to manage external traffic routing and assign IPs for services exposed to the outside world.
- Istio Service Mesh
  - Configured Istio for advanced traffic management, security, and observability between microservices.

- Knative Setup
  - Set up Knative for serverless deployment of the ML model, automating scaling, routing, and networking management.
- ML Model Deployment on Knative
  - Deployed the machine learning model on Knative, allowing automatic scaling and efficient resource management.
- Prometheus and Grafana Setup
  - *Implemented Prometheus for metrics collection and Grafana for visualizing system health and performance.*
- WRK Benchmarking
  - *Used WRK to benchmark the deployed ML model, testing performance under load and ensuring it meets required throughput and response times.*

DEMO

METRICS

- *KNative outperforms OpenFaaS in RPS, handling ~20% more requests per second (100.39 RPS vs. 83.50 RPS).*
- KNative (1.28s) has a slightly higher latency than OpenFaaS (1.08s).
- KNative had only 62 failed requests, while OpenFaaS had 525 in total of 6000 requests.
- OpenFaaS had a higher transfer rate (41.39 KB/sec) compared to KNative (24.42 KB/sec).



IMPORTANT LINKS

- Team GitHub Repo : https://github.com/RutanshS/ml-serverless-evaluation
- Google Drive : https://drive.google.com/drive/folders/1rPVVXa27XYXUZ-X5kv1xAo0S2u8n-WZ8?usp=drive_link
- Kubernetes Setup: https://devopscube.com/setup-kubernetes-cluster-kubeadm/
- OpenFaaS Documentation : https://docs.openfaas.com/
- KNative Documentation : https://knative.dev/docs/

THANK YOU

Happy To Answer Questions