DEEPSHORE

Inference Services on Kubernetes

Malte Groth

About me

- Cloud Technology Evangelist at Deepshore
- since 2019 at Deepshore
- founded the Analytics & ML Team
- Collaborator at MML@IKIM Essen
- main subjects:
  - Cloud Technology & Cloud Architecture
  - Life-Cycle-Management & Operations
  - Automation in Kubernetes
  - MLOps

About Deepshore

- OMI: associated partner / technical rollout partner
- Kubernetes Certified Service Provider (KCSP)
- Member of Cloud Native Computing Foundation
- Member of Innovation Park Artificial Intelligence (IPAI)
- Main subjects:
  - Cloud Technology
  - Distributed Systems
  - R&D in the field of AI, e.g.
    - AI optimized operations
    - Anomaly detection based on IoT data
    - Digital assistance for document management

Quelle: www.deepshore.de

Goal of this presentation

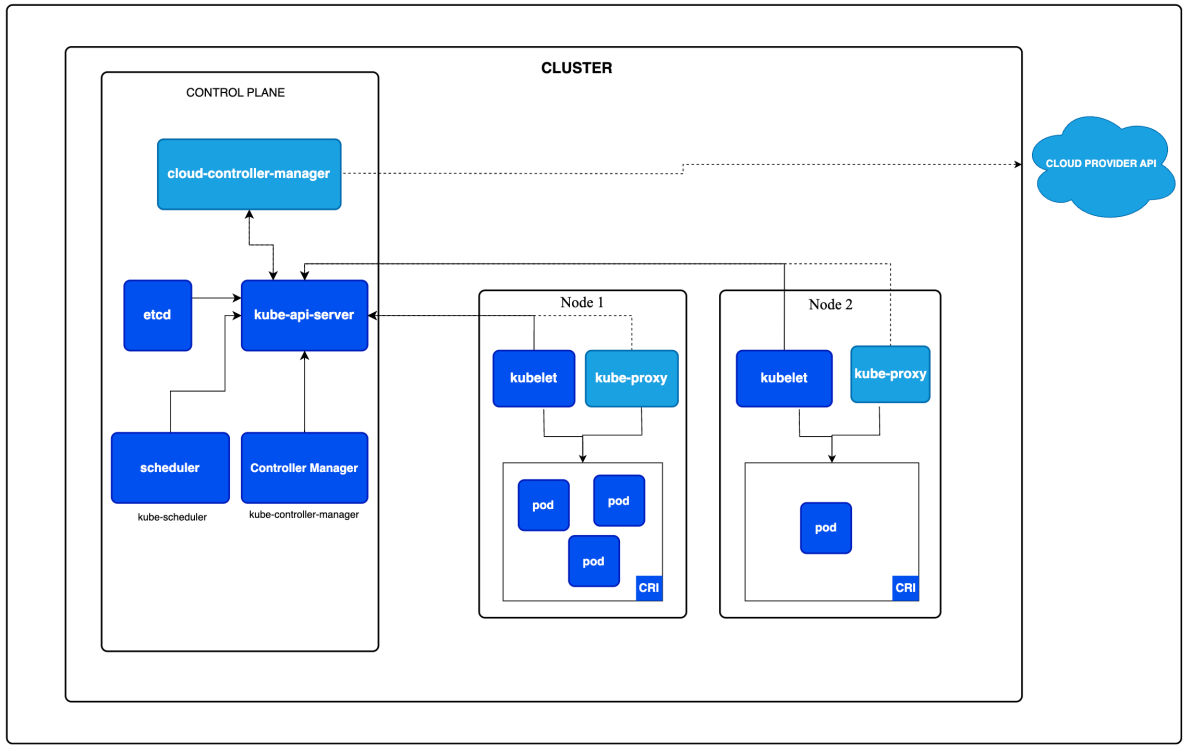
Show how we can serve AI models in consideration of

- High Availability
- Reliability
- BUT ALSO: Efficiency and Usability

What is Kubernetes?

„Kubernetes is a …

- open-source
- container orchestration system
- for automating
- software deployment,
- scaling
- and management.“



Quelle: www.kubernetes.io

Why Kubernetes?

- High Availability:
  - K8s clusters consists of multiple nodes
  - K8s controllers enable service (pod) replication
  - K8s controllers provide self-healing mechanisms

- Resource Efficiency:
  - K8s comes with a powerful scheduling

Link: https://dzone.com/articles/kubernetes-advantages-and-disadvantages

KServe: Model Inference platform on Kubernetes

KServe

- offers (auto)scaling, e.g.
  - if number of requests increases
  - if there is no load at all (scale to zero)
- standardized inference protocol across ML frameworks
- simplifies model deployment

Link: https://kserve.github.io/website/0.11/

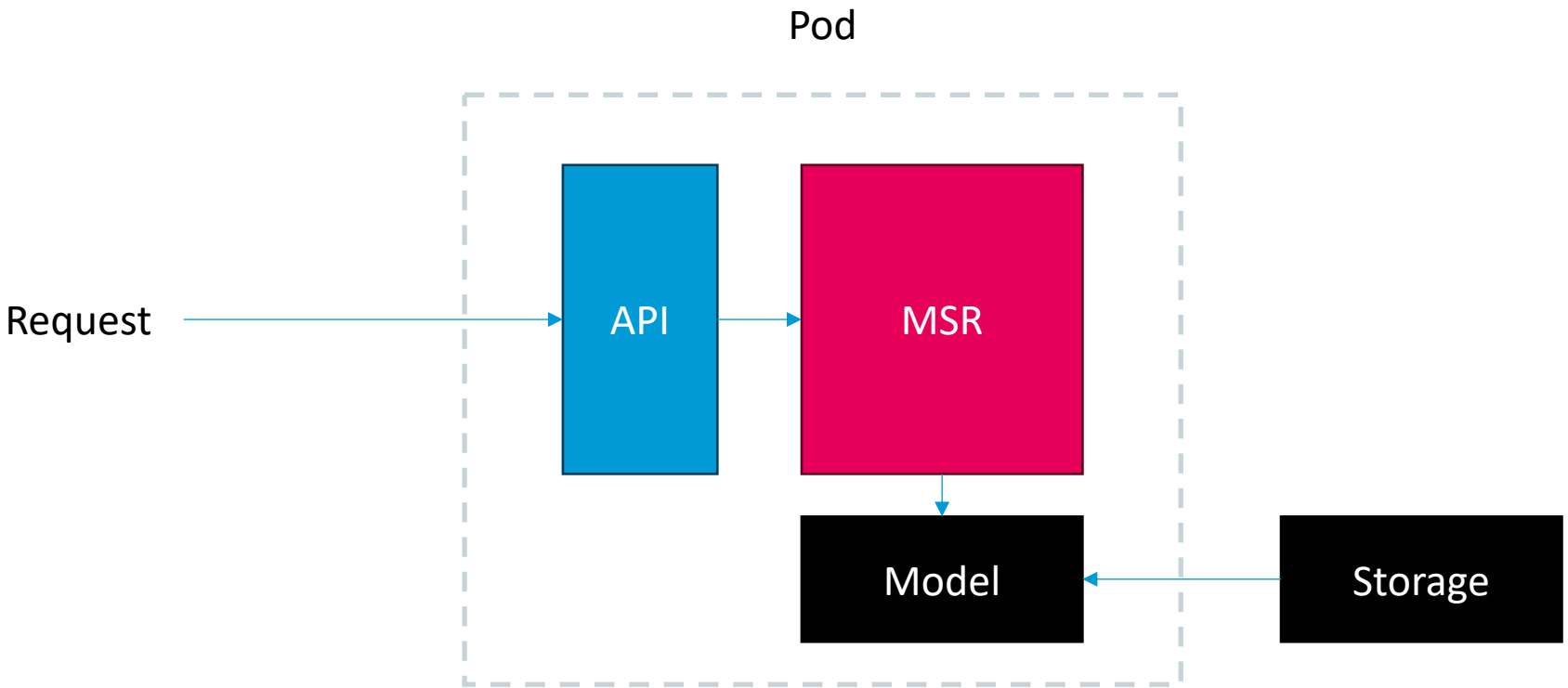
InferenceService

InferenceServices

- provide Inference APIs out-of-the-box
- support multiple ML frameworks/Model Serving Runtimes
- support for obtaining models from different storage locations
- provide Autoscaling, incl. Scale-To-Zero

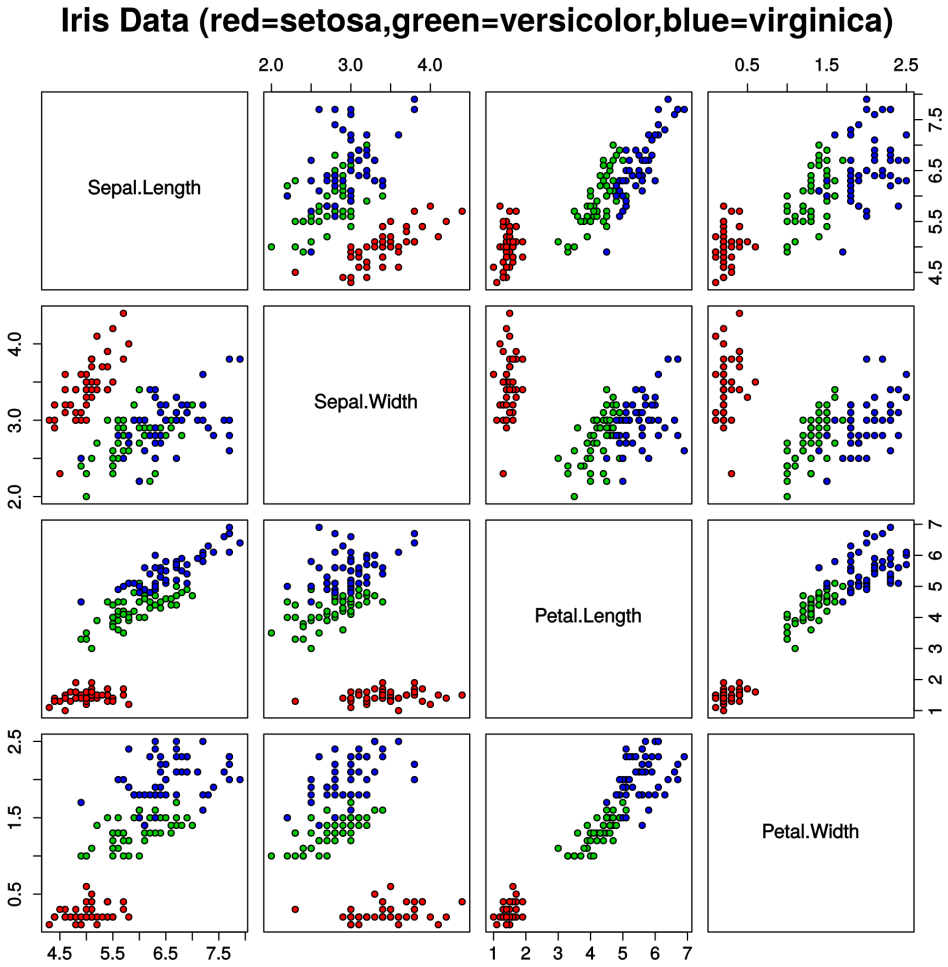
"Since your model is being deployed as an InferenceService, not a raw Kubernetes Service, you just need to provide the storage location of the model and it gets some super powers out of the box 🚀."

InferenceService



Setup Hands-On Demo

- k8s cluster by minikube
- KServe is already installed
- the model classifies species of iris
  (see: https://en.wikipedia.org/wiki/Iris_flower_data_set)
- the model was already uploaded to MinIO



Quelle: www.wikipedia.de

Inference API (data plane protocol)

- Versions:
  - v1 (REST)
  - v2 (REST, grpc)
- v2 corresponds to OIP: https://github.com/open-inference/open-inference-protocol

Model Serving Runtimes

MSR: Applications providing/serving models efficiently and quickly

Kserve supports

- TorchServe (Link: https://pytorch.org/serve/server.html)
- Triton Inference Server (Link: https://github.com/triton-inference-server/server)
- MLServer (Link: https://mlserver.readthedocs.io/en/latest/)
- Custom MSR

Model Storage

Models can be provided by

- S3
- Azure Blob Storage
- URL
- PVC

Autoscaling

Autoscaling

- Scales to zero if there is no request (given minReplicas is 0)
- Scales up to maxReplicas if more replicas are needed

Summary

- Kubernetes + KServe support High Availability (multiple nodes, pod replication, self-healing)
- Kubernetes + KServe help to utilize resources efficiently (scheduling, autoscaling)
- KServe makes serving models very easy --> Usability
- KServe has features that improve reliability (versioning, monitoring etc.)

Outlook

- Perform autoscaling on GPUs
- Show model versioning in action
- Increase reliability by making use of GitOps
- Integration of FHIR resources in the Kubernetes-API
- ...

Thank you for your attention

Contact: malte.groth@deepshore.de