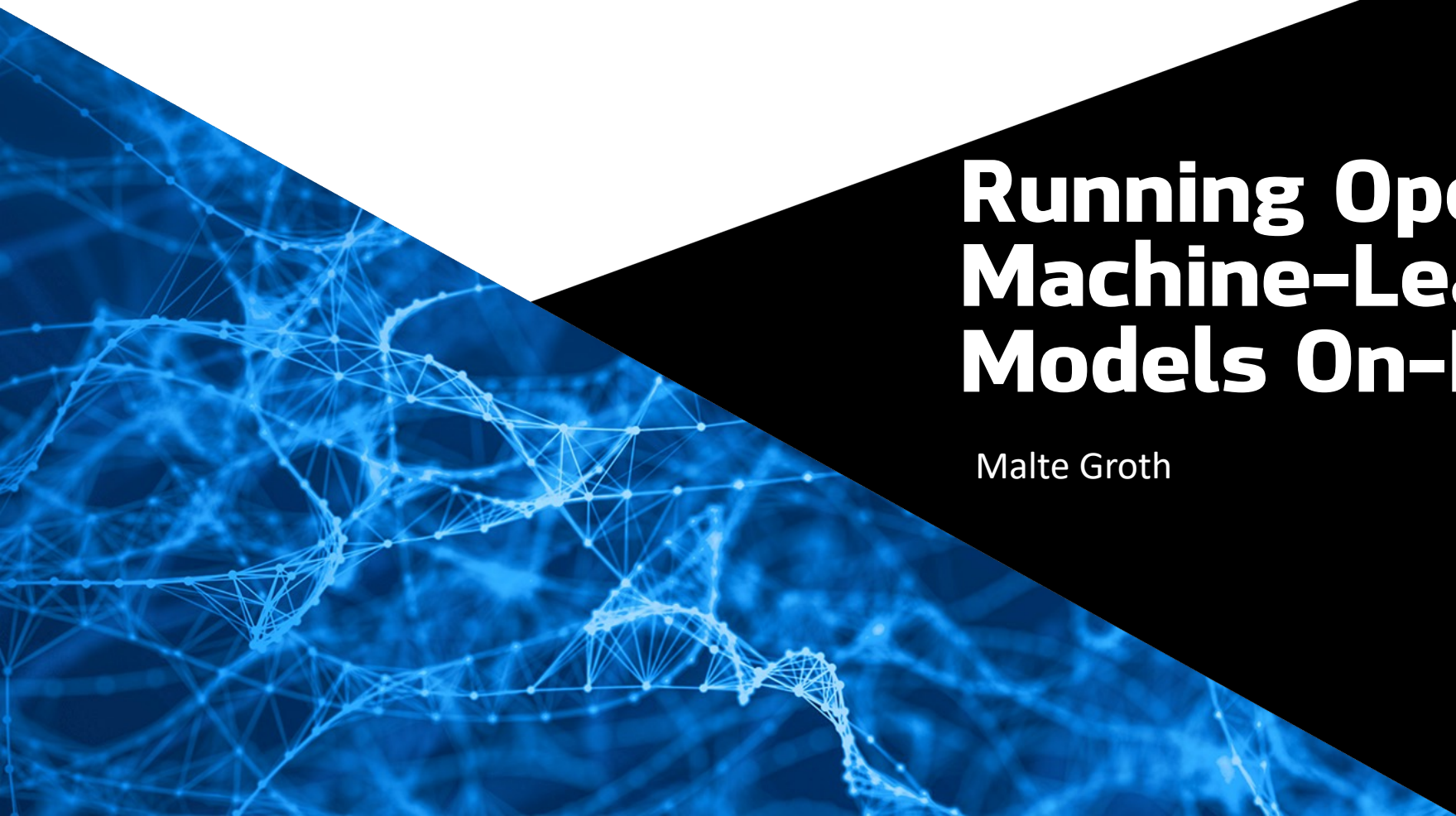
DEEPSHORE

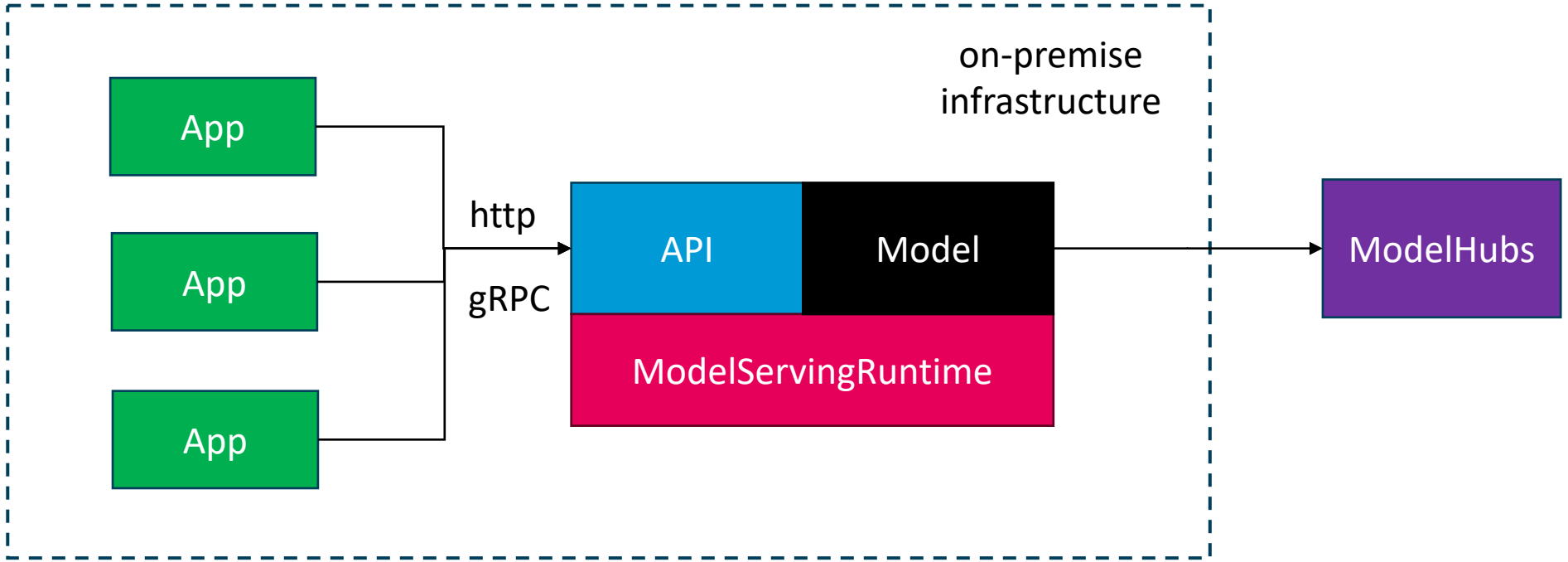


Running Open-Source Machine-Learning Models On-Premise

Malte Groth

The title of the talk in one diagram

“Providing access to Open-Source Machine-learning Models via self-hosted Web-Services for Inference”



Running Open-Source Machine-Learning Models On-Premise

Open-Source Machine-Learning Models

... are models available under an Open-Source License (e.g. Apache 2.0)

Sources: HuggingFace, torch.hub, Github

Do not confuse Open-Source with Openness: Open-Source models differ in terms of transparency, reproducibility and quality control.

Advantages of Open-Source Machine-Learning Models

- Transparency and Reliability
- Availability
- Adaptability
- Performance
- Autonomy (Avoiding Vendor-Lock-In)
- Cost-Saving

Of particular interest in Medicine: Transparency, Reliability and Adaptability

Running Open-Source Machine-Learning Models On-Premise

On-Premise

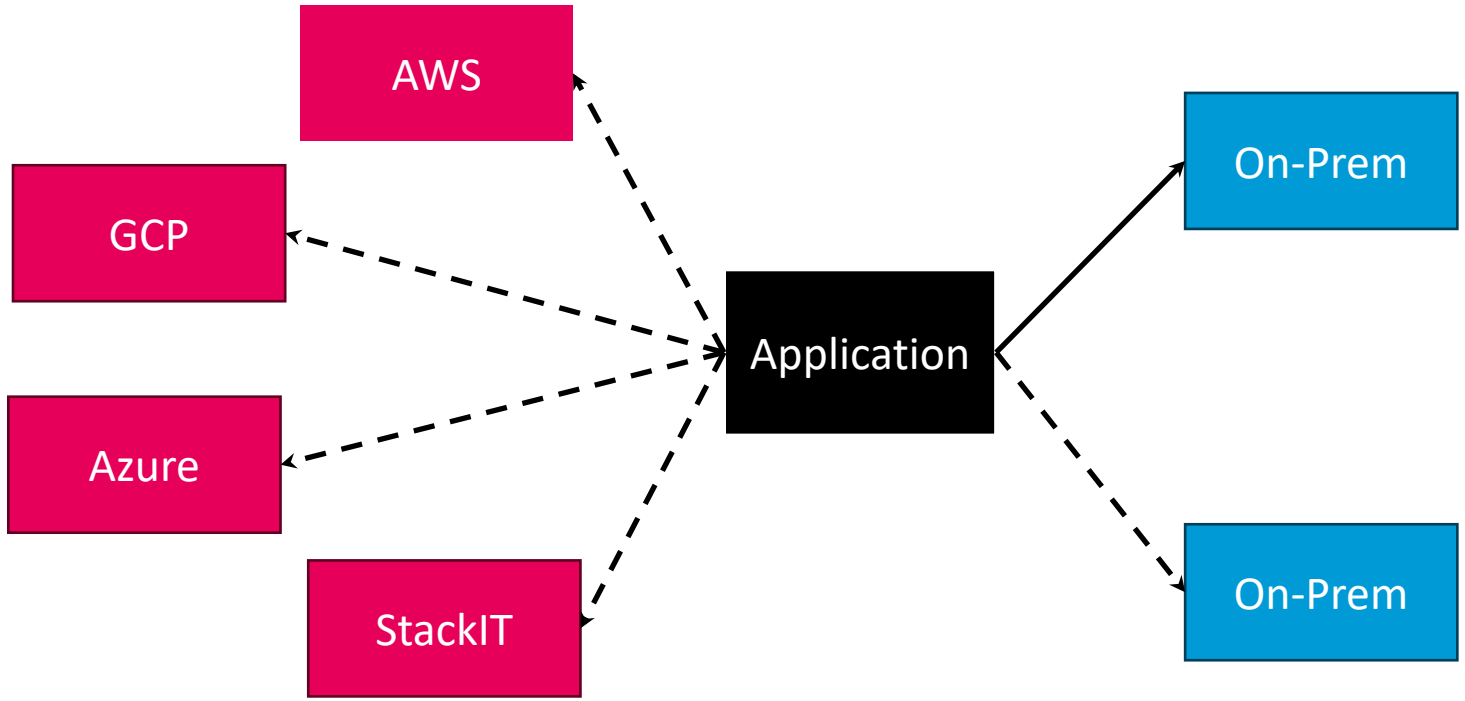
- Deployments and operations are located within physical premises of a company/organization
- Opposite of the cloud
- Full control over IT infrastructure, data and applications

Motivation for running Models on-premise

- Data Security and Privacy
- Latency and Performance
- Cost-Saving
- Offline Access
- Control over intellectual property
- Flexibility → especially true if following a …

Cloud-agnostic deployment strategy

Designing your Applications, tools and services in a way so they can migrate seamlessly between multiple cloud platforms and on-premises.



Kubernetes

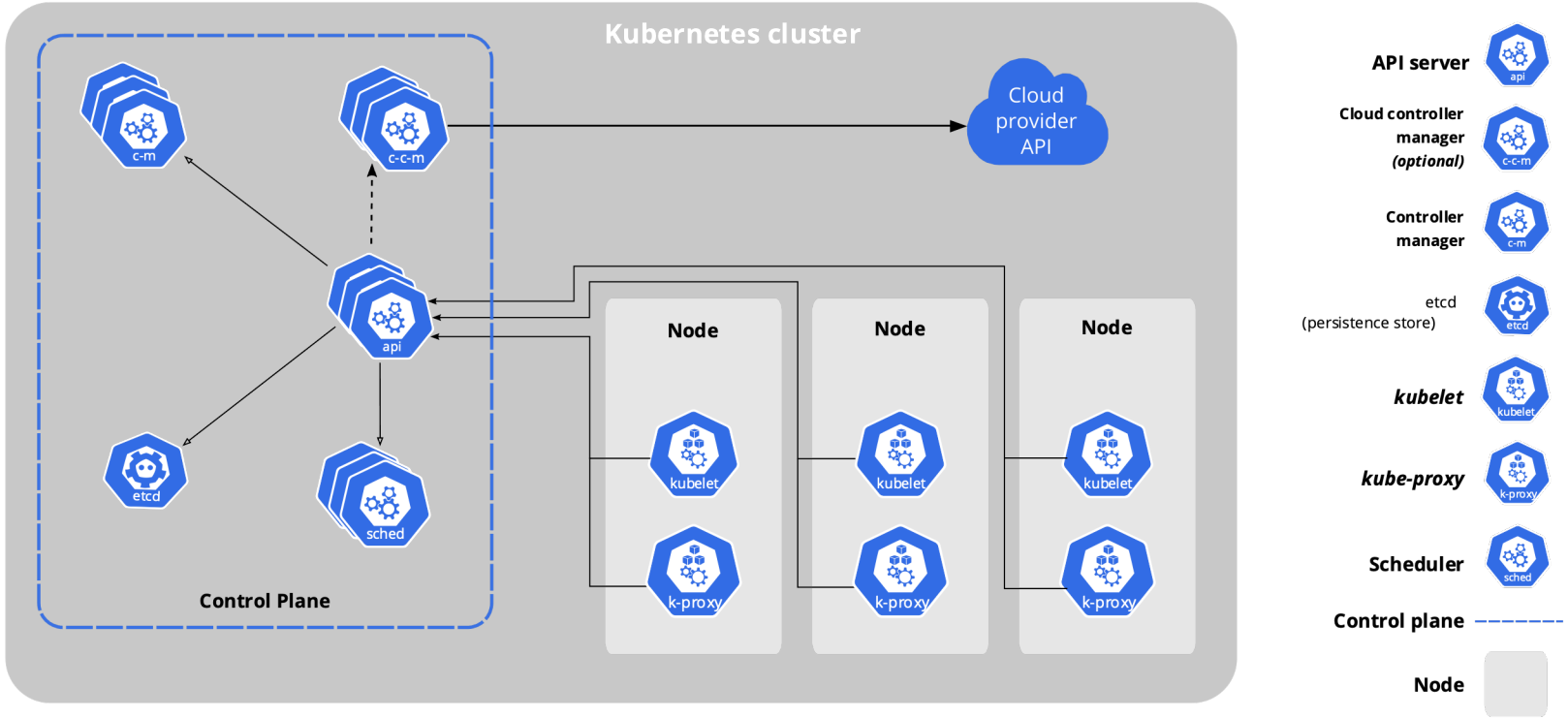
"Kubernetes (k8s) is a ...

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

And it is a key technology for implementing a cloud-agnostic deployment strategy because it

- **abstracts** the underlying infrastructure
- has a **consistent** tooling and user interface
- provides **scaling at ease**

Kubernetes



DEEPSHORE

Running Open-Source Machine-Learning Models On-Premise with Kubernetes

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

Running Open-Source Machine-Learning Models On-Premise In Practice

Example

1) Download model files from external source
2) Prepare model files
3) Store the prepared model on model storage
4) Create an InferenceService
5) InferenceService takes care of running the model serving runtime

Summary

- Open-Source Machine-Learning Models have a lot of advantages.
- There are many reasons why it may be appropriate or even necessary to deploy models on-premise.
- You can gain high flexibility by choosing a cloud-agnostic approach.
- Kubernetes is a key technology for implementing a cloud-agnostic deployment strategy.
- KServe is a inference platform suited for deploying models in production on Kubernetes.

Thank you for your attention

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

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