KubeCon CloudNativeCon

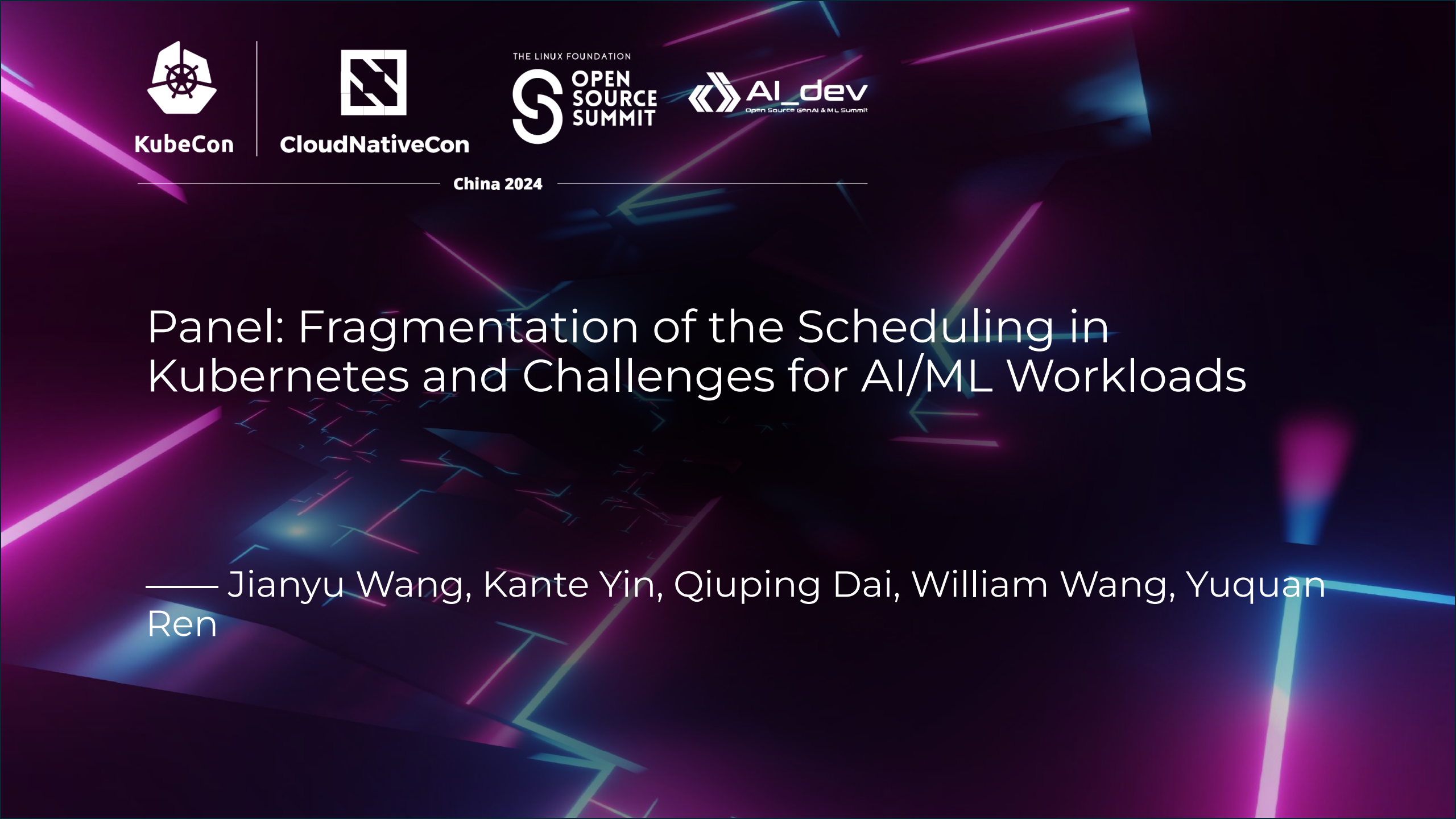
THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024



Panel: Fragmentation of the Scheduling in Kubernetes and Challenges for AI/ML Workloads

—— Jianyu Wang, Kante Yin, Qiuping Dai, William Wang, Yuquan Ren

Panelists

Qiuping Dai
Product Manager
Independent
@qiupingdai

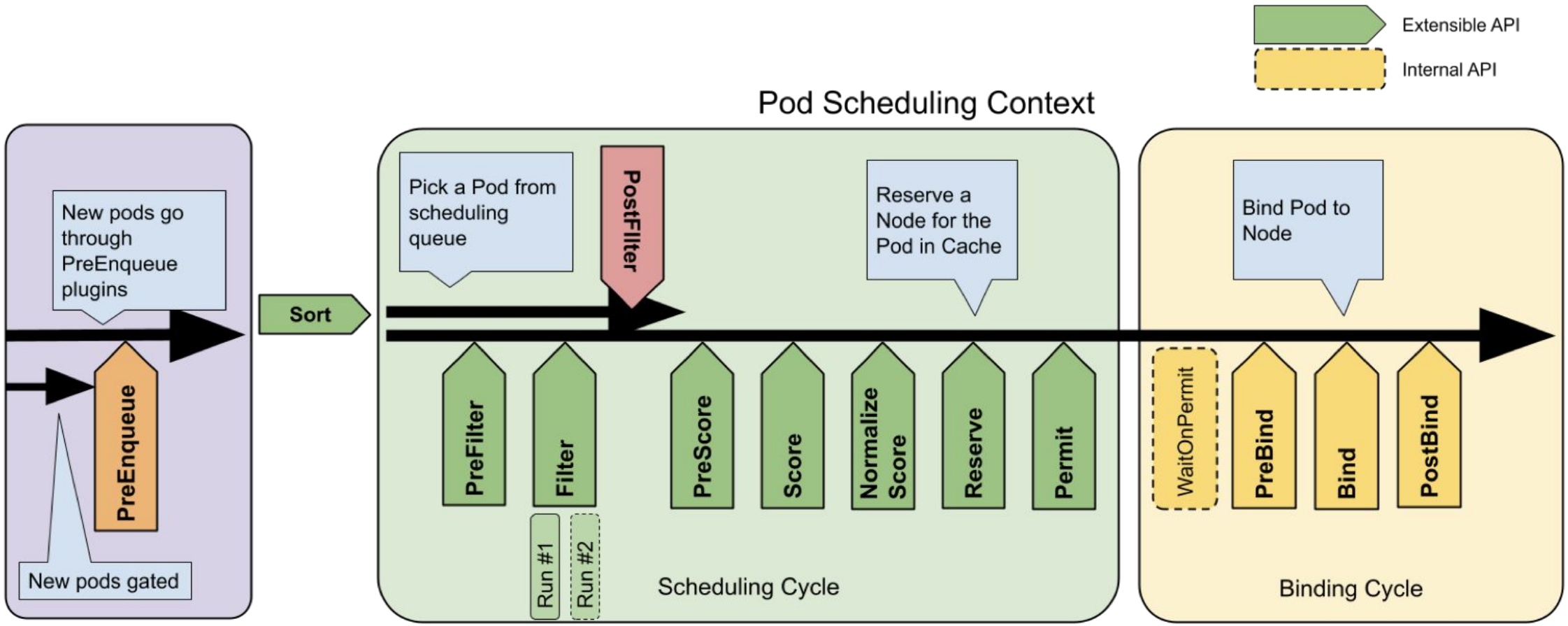
William Wang
Volcano TechLead
Huawei Cloud
@william-wang

Yuquan Ren
Godel-Scheduler Maintainer
Bytedance
@NickrenREN

Jianyu Wang
Koordinator Member
Alibaba Cloud
@jianyuwang

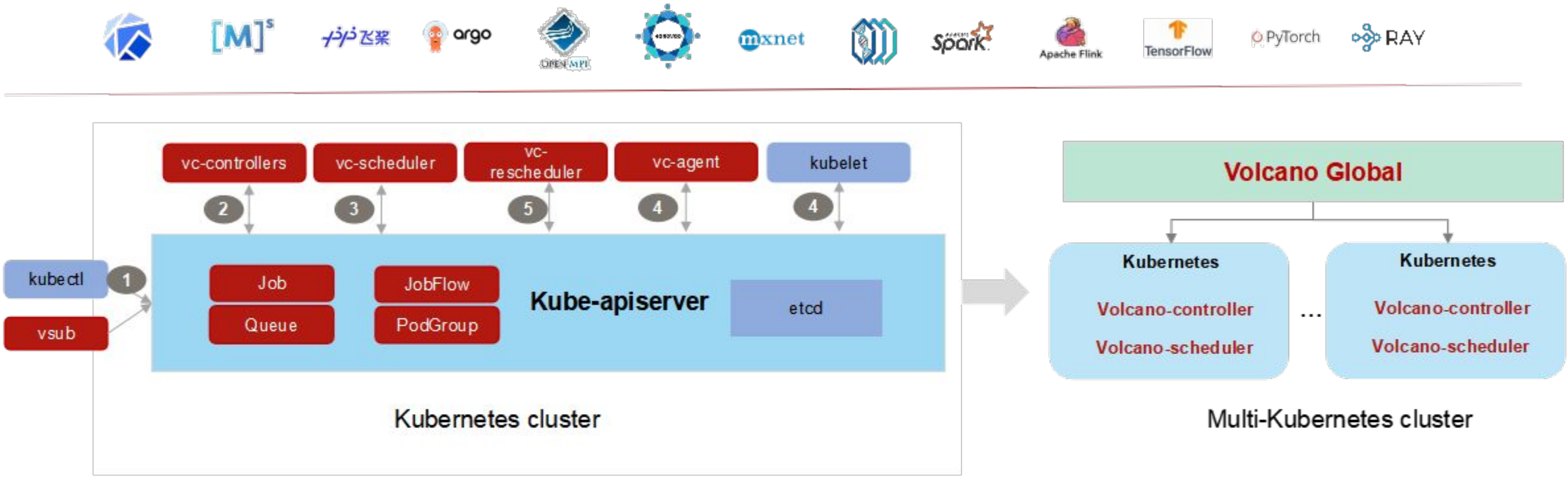
Kante Yin
SIG-Scheduling Maintainer
DaoCloud
@kerthcet

Kube-Scheduler Overview



https://kubernetes.io/docs/concepts/scheduling-eviction/scheduling-framework/

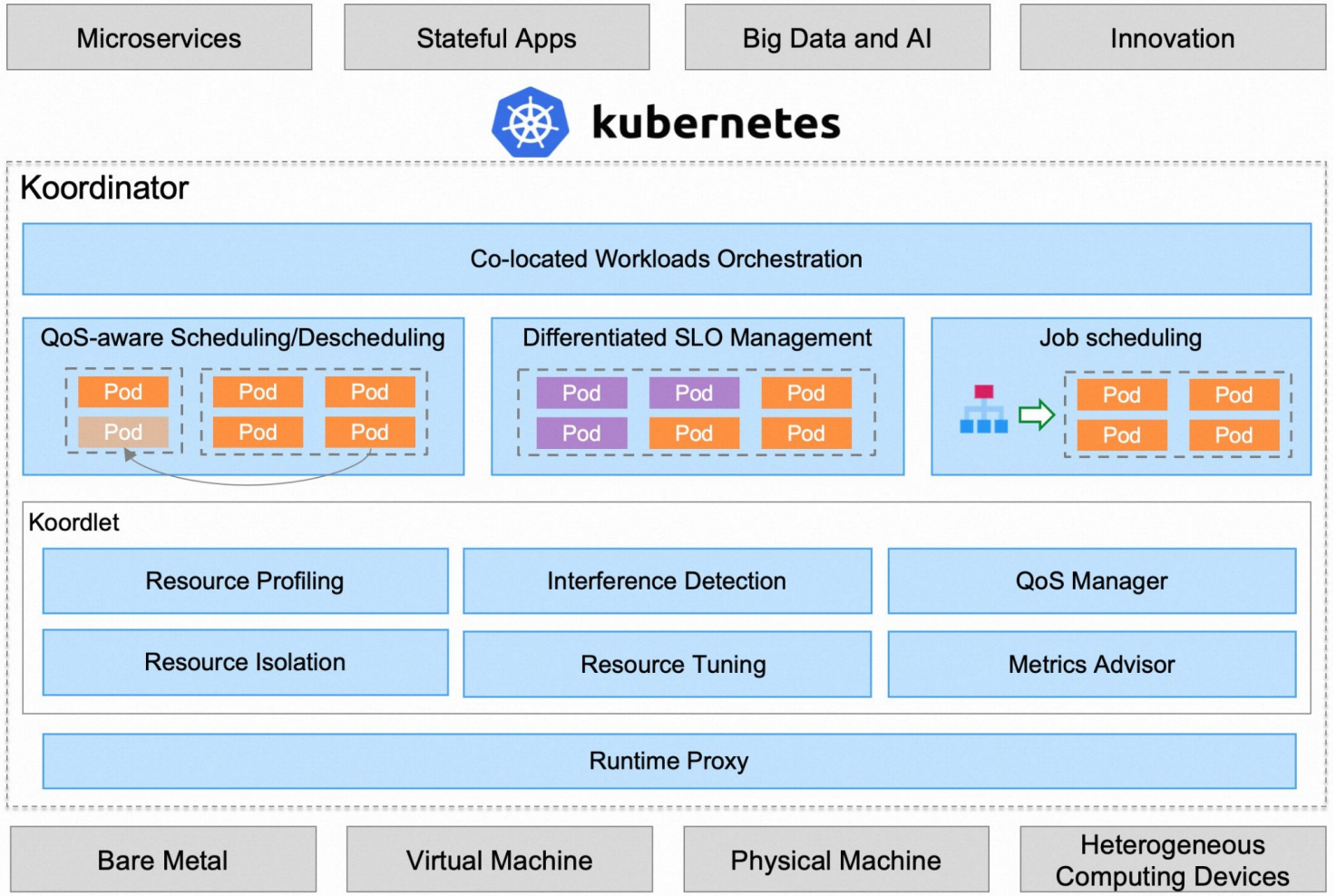
Volcano Overview



- Generic batch APIs for AI and BigData, e.g. PodGroup, Job, Jobflow, Queue, etc.
- Rich advanced scheduling policy, e.g. fair-share, topology scheduling, SLA, preempt, backfill.
- Job lifecycle management, e.g. multiple pod template, error handling, unified support mainstream framework such as Pytorch, MPI, etc.
- Resource management enhancement, e.g. dynamic resource share, heterogeneous device support like GPU, NPU etc.
- Performance tuning， e.g. scalability, throughput, network, container runtime.

https://github.com/volcano-sh/volcano

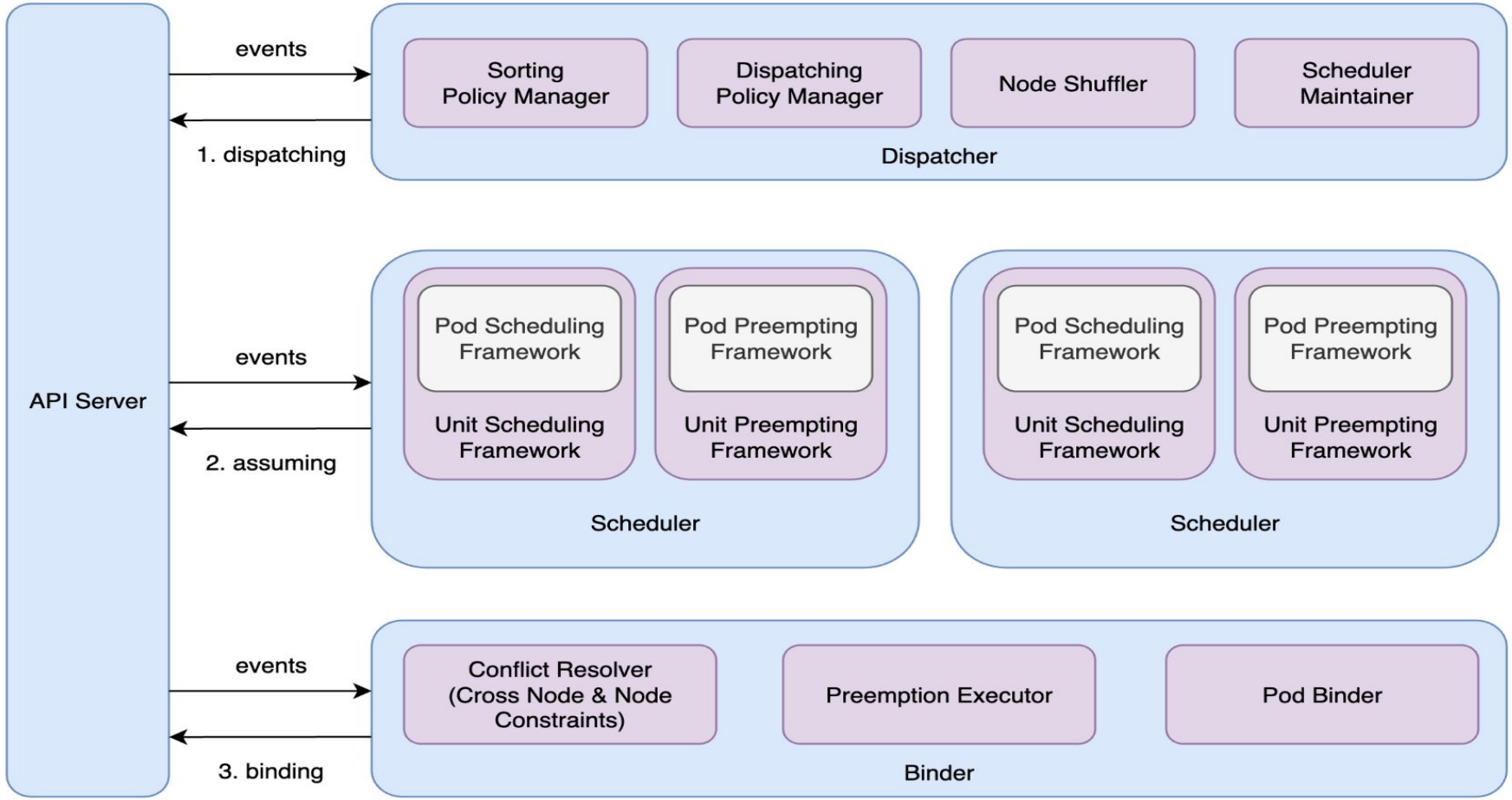
Koordinator Overview



https://github.com/koordinator-sh/koordinator

Godel Scheduler Overview

Overall Architecture



https://github.com/kubewharf/godel-scheduler

Panelists

Qiuping Dai
Product Manager
Independent
@qiupingdai

William Wang
Volcano TechLead
Huawei Cloud
@william-wang

Yuquan Ren
Godel-Scheduler Maintainer
Bytedance
@NickrenREN

Jianyu Wang
Koordinator Member
Alibaba Cloud
@jianyuwang

Kante Yin
SIG-Scheduling Maintainer
DaoCloud
@kerthcet

Thanks!

群聊: Kubernetes 调度交流群