KUBERNETES
COMMUNITY DAYS DALIAN

Volcano加速金融行业大数据分析平台云原生化改造的应用实践

汪 洋，华为云

Volcano 社区核心贡献者

大数据平台云原生面临的挑战

KUBERNETES COMMUNITY DAYS DALIAN

传统大数据平台云原生化改造成为必然趋势

大数据分析、人工智能等批量计算场景深度应用于金融场景

大数据、AI等批量计算场景云原生化面临的挑战

作业管理缺失

- Pod级别调度，无法感知上层应用
- 缺少作业概念、缺少完善的生命周期的管理
- 缺少任务依赖、作业依赖支持

调度策略局限

- 不支持Gang-scheduling、Fair-share scheduling
- 不支持多场景的Resource reservation，backfill
- 不支持CPU/IO topology based scheduling

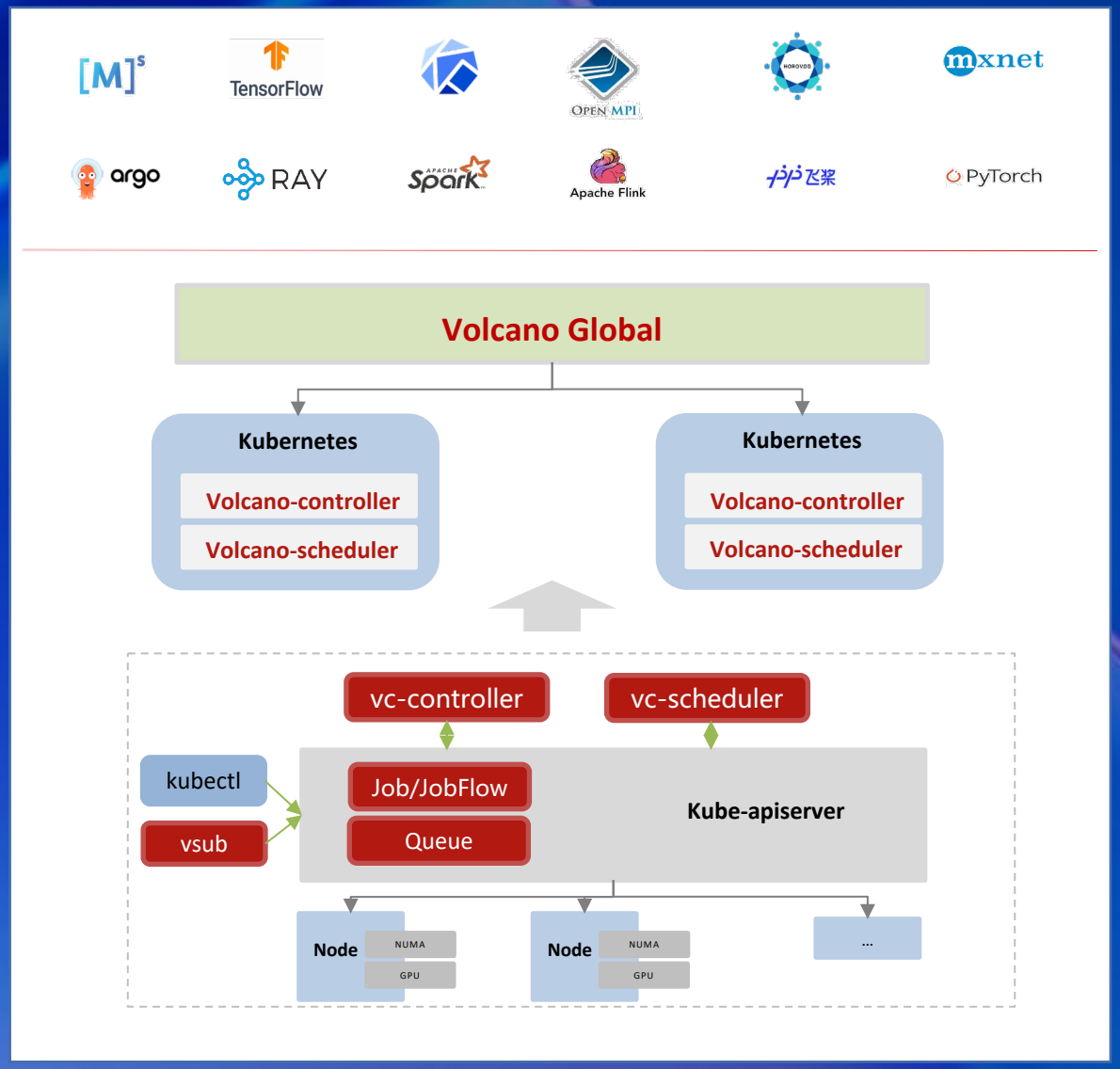
领域框架支持不足

- 1:1的operator部署运维复杂
- 不同框架对作业管理、并行计算等要求不通
- 计算密集，资源波动大，需要高级调度能力

资源规划复用、异构计算支持不足

- 缺少队列概念
- 不支持集群资源的动态规划以及资源复用
- 对异构资源支持不足

Volcano 架构



KUBERNETES COMMUNITY DAYS DALIAN

项目概况：

- 业界首个云原生批量计算平台
- 2019年6月开源，2020年进入CNCF，目前是CNCF孵化级项目
- **2.9k** star，**500+** 全球贡献者
- **50+** 企业生产落地

关键特性：

1. **统一的作业管理**

   提供完善作业生命周期管理，统一支持几乎所有主流的计算框架，如Pytorch, MPI, Horovod, Tensorflow、Spark等。

2. **丰富的高阶调度策略**

   公平调度、任务拓扑调度、基于SLA调度、作业抢占、回填、弹性调度、混部等。

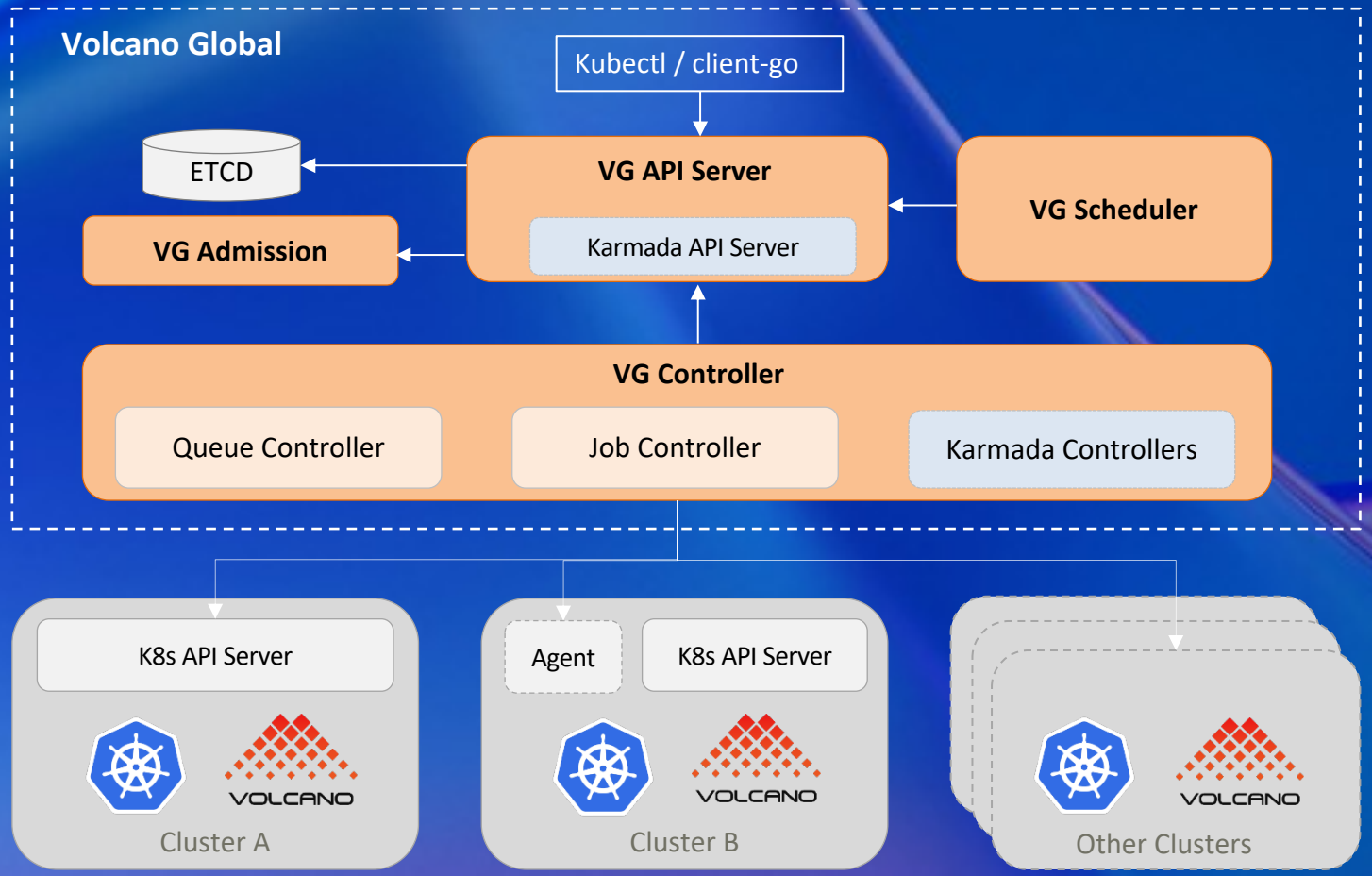
3. **细粒度的资源管理**

   提供作业队列，队列资源预留、队列容量管理、多租户的动态资源共享。

4. **性能优化和异构资源管理**

   调度性能优化，并结合 Kubernetes 提供扩展性、吞吐、网络、运行时的多项优化，异构硬件支持x86, Arm, GPU, 昇腾，昆仑等。

VolcanoGlobal 架构



多中心

低成本

无绑定

关键特性:

- 开箱即用的多集群管理功能
- 分级调度，保证调度性能
- 多租户公平调度
- 成本感知

Volcano 使用方法

Volcano部署命令如下，详情参见项目主页：https://github.com/volcano-sh/volcano

kubectl apply -f https://raw.githubusercontent.com/volcano-sh/volcano/master/installer/volcano-development.yaml

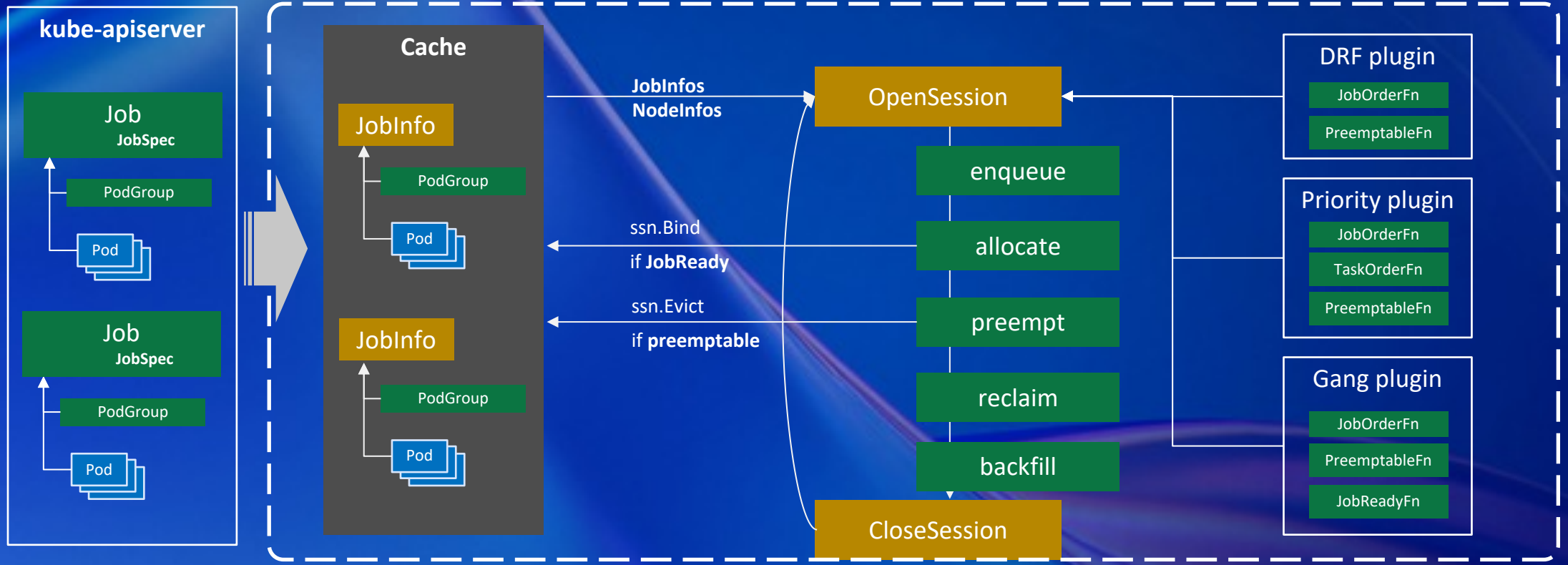
volcano-scheduler-configmap 示例

```
actions:   “enqueue, reclaim, preempt, allocate, backfill"
tiers:
- plugins:
  - name: priority
  - name: gang
    enablePreemptable: false
  - name: conformance
- plugins:
  - name: overcommit
  - name: drf
    enablePreemptable: false
  - name: predicates
  - name: proportion
  - name: nodeorder
  - name: binpack
```

vcjob 示例

```
apiVersion: batch.volcano.sh/v1alpha1
kind: Job
metadata:
  name: mpi-job
  labels:
    "volcano.sh/job-type": "MPI"
spec:
  # minimum number of pods need to be started
  minAvailable: 3
  schedulerName: volcano
  plugins:
    # job level ssh trust
    ssh: []
    # define network relevant info for running,
    # hosts, headless services etc.
    svc: []
  # restart who job if any pod get evicted
  policies:
    - event: PodEvicted
      action: RestartJob
  tasks:
    - replicas: 1
      name: mpimaster
      # Mark whole job completed when mpiexec completed
      policies:
        - event: TaskCompleted
          action: CompleteJob
```

Volcano 内部机制

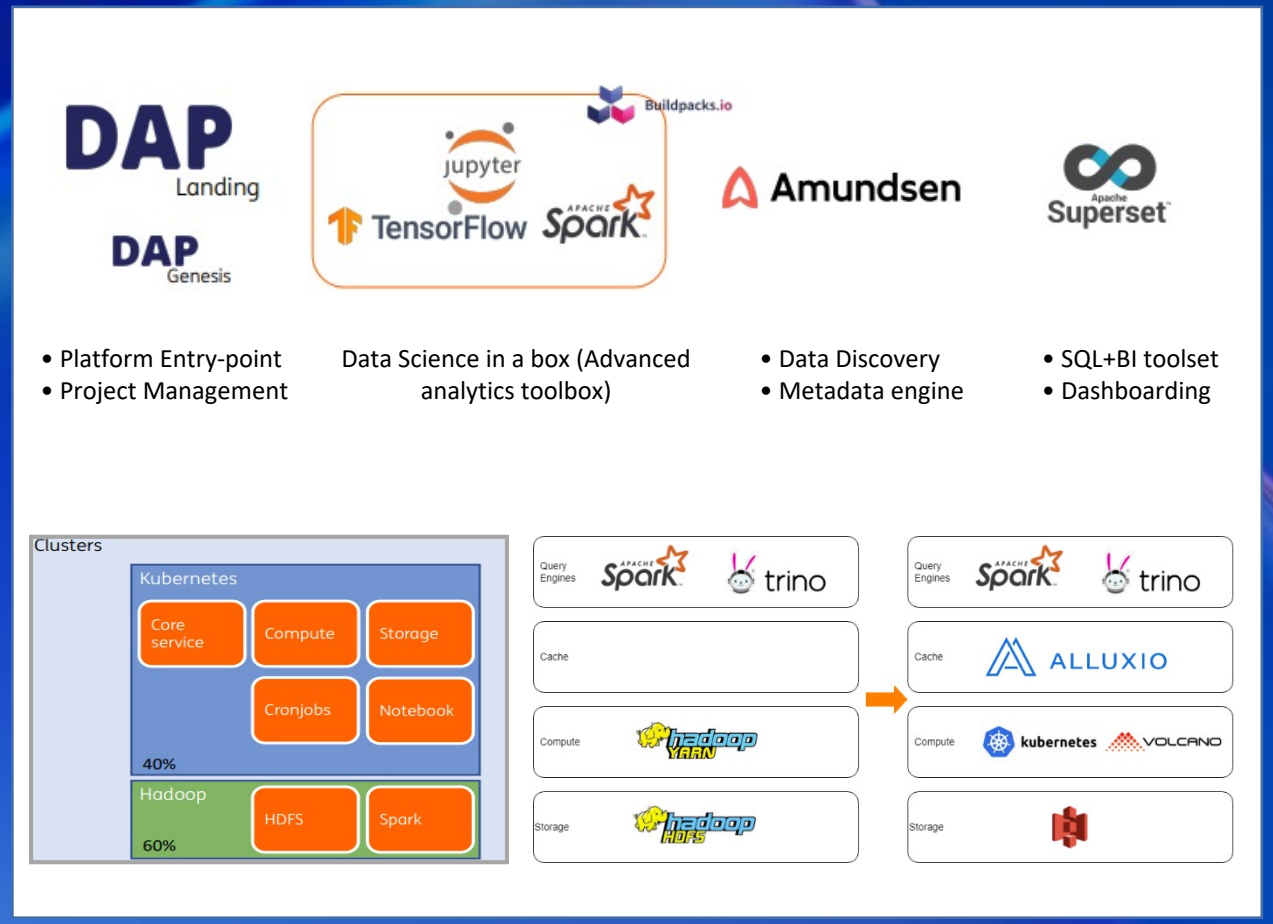


Re-construct JobInfo in Cache by **PodGroup**

Predicate, *allocate*, *preempt* are **Action**s, and they're **pluggable**

Plugins on demand

用户案例：荷兰ING银行大数据平台云原生化改造



Information reference： https://volcano.sh/en/blog/ing_case-en/

业务场景:

- *ING荷兰国际集团（International Netherlands Groups）为全球排名前列的资产管理公司，服务遍及40多个国家，核心业务是银行、保险及资产管理等。引入云原生基础设施，打造新一代大数据分析自助平台。*

客户诉求:

- 交互式服务、常驻服务、离线分析业务**统一平台调度**；
- **Job级别**的调度管理，包括生命周期、依赖关系等；
- 支持业界**主流计算框架**，如Spark、TensorFlow等；
- 多用户公平分配资源，快速响应高优先级作业

解决方案:

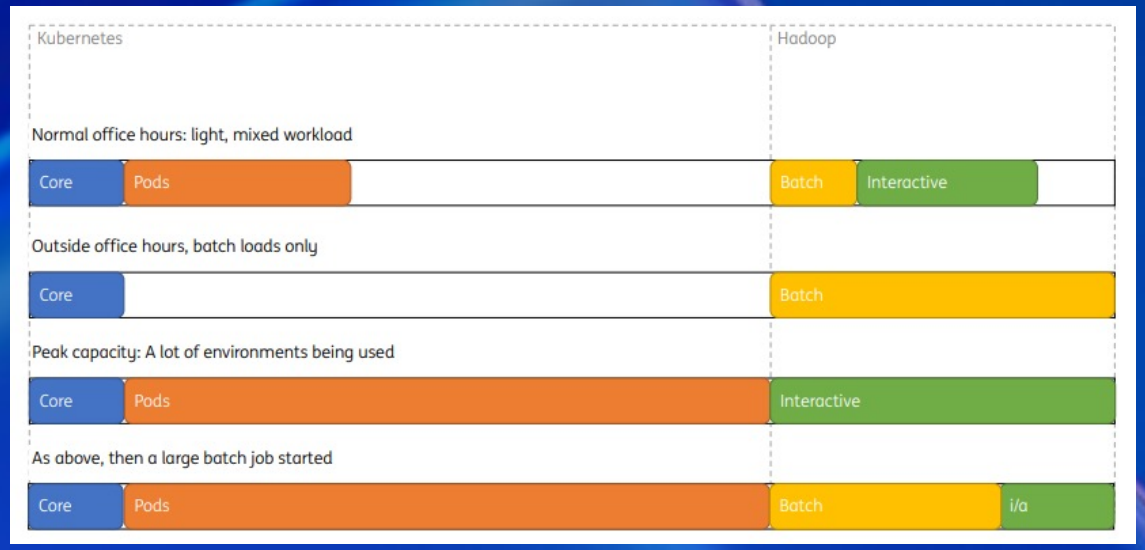
- K8s + Volcano 统一调度所有工作负载；
- Queue动态资源共享，DRF、优先级抢占

用户收益：

- 大数据作业从**Yarn平滑迁移至K8s**；
- 云原生DAP平台服务于**17个国家/地区**，**1100**用户，年增长率**8.1%**；
- DAP平台运行项目**450+**

Volcano大幅度提高大数据平台资源利用率

Kubernetes + YARN



静态划分资源池

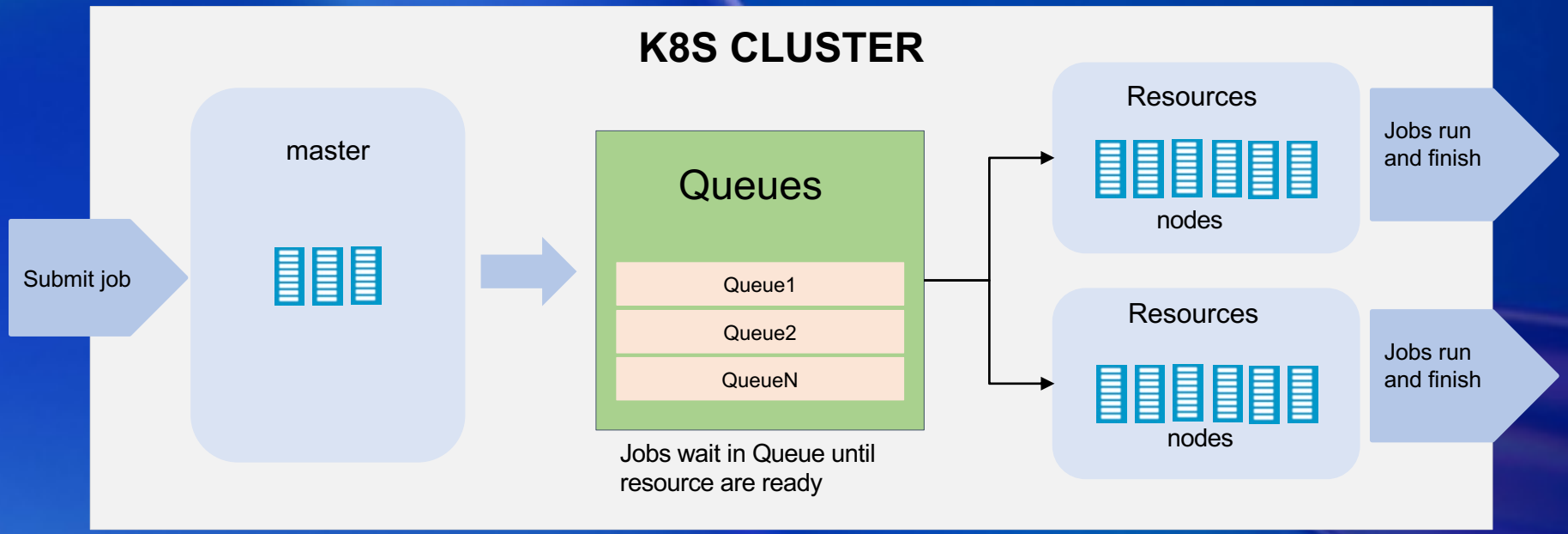
Kubernetes + Volcano

统一资源池

	Kubernetes + YARN	Kubernetes + Volcano
集群低负载场景	K8s资源池空闲，大数据业务无法使用	大数据业务可以使用集群整体空闲资源，提高整体资源利用率
集群高负载场景	通过静态划分的资源池保证大数据业务和通用业务的资源配额	通过Volcano提供的队列保证各类业务资源配额

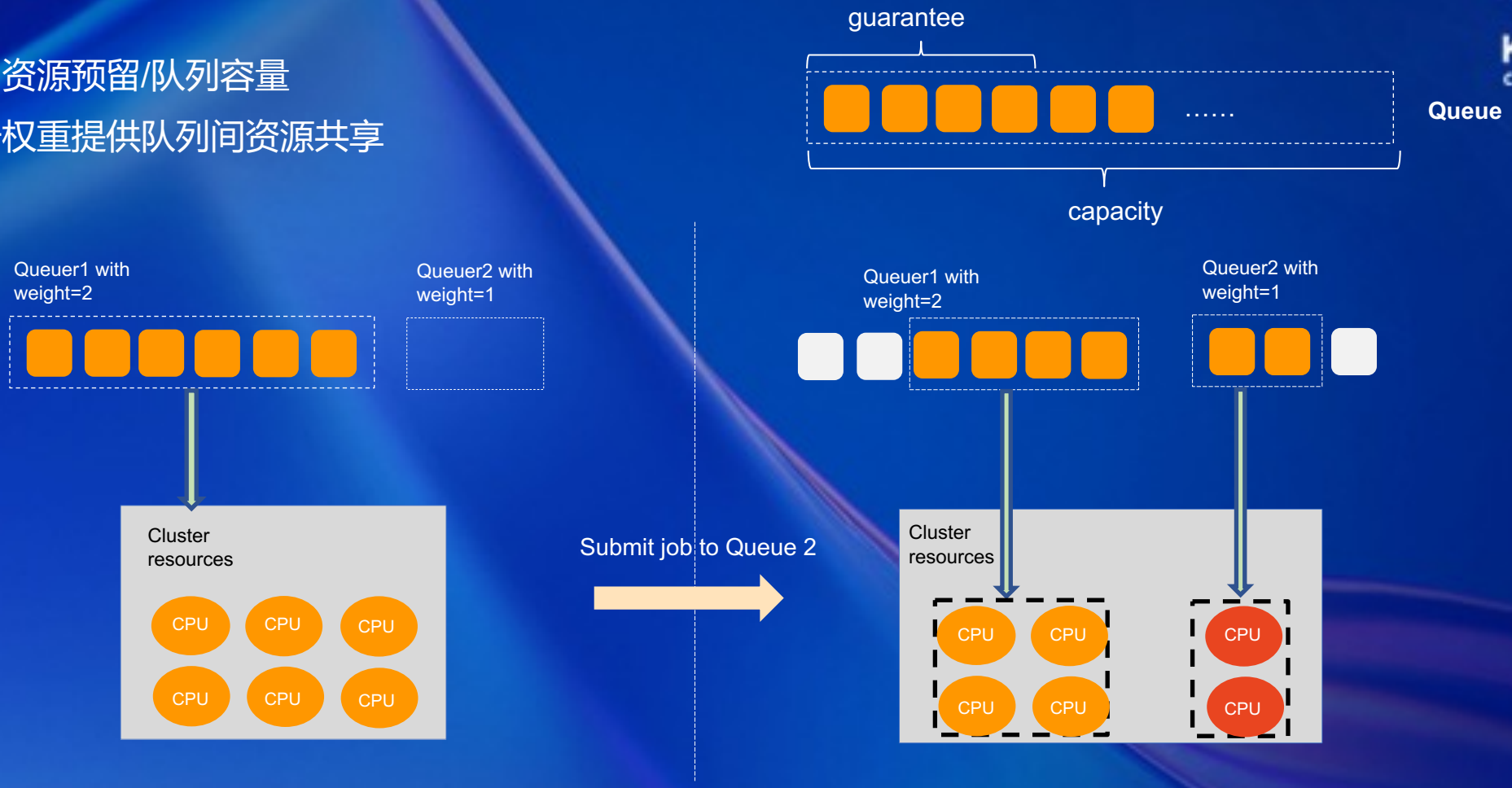
资源共享：Queue

- 集群级别资源对象，与用户/namespace解耦
- 可用于租户/资源池之间共享资源
- 支持每个队列独立配置Policy，如 FIFO, fair share, priority, SLA等



动态资源共享

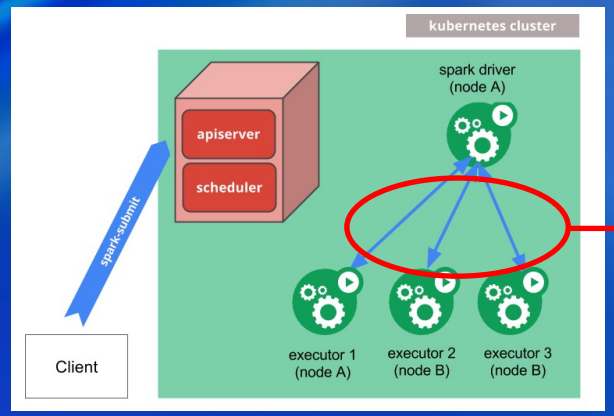
- 队列资源预留/队列容量
- 基于权重提供队列间资源共享



Queue2 is empty. Q1 can borrow resources from Queue2.

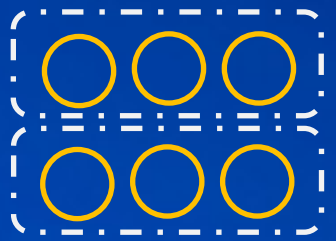
Queue2 has workload, it will reclaim resources from Queue1.

Spark首个Batch调度器



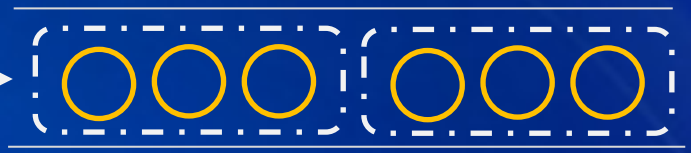
More k8s native resource support

Pod Group



- **依据PodGroup调度**
- 最小资源预留（CPU/MEM）
- 作业优先级

Queue



- 多队列
- 多租户
- 优先级
- 公平调度
- 抢占
- ……

1.	Support replicasets/job API		RESOLVED	Holden Karau
2.	Add the ability to specify a scheduler & queue		IN PROGRESS	Apache Spark
3.	Support backing off dynamic allocation increases if resources are "stuck"		OPEN	*Unassigned*
4.	Create a PodGroup with user specified minimum resources required		OPEN	*Unassigned*
5.	Support for specifiying executor/driver node selector		RESOLVED	Yikun Jiang
6.	Support the Volcano Job API		OPEN	*Unassigned*

SPARK-36057: Support volcano/alternative schedulers

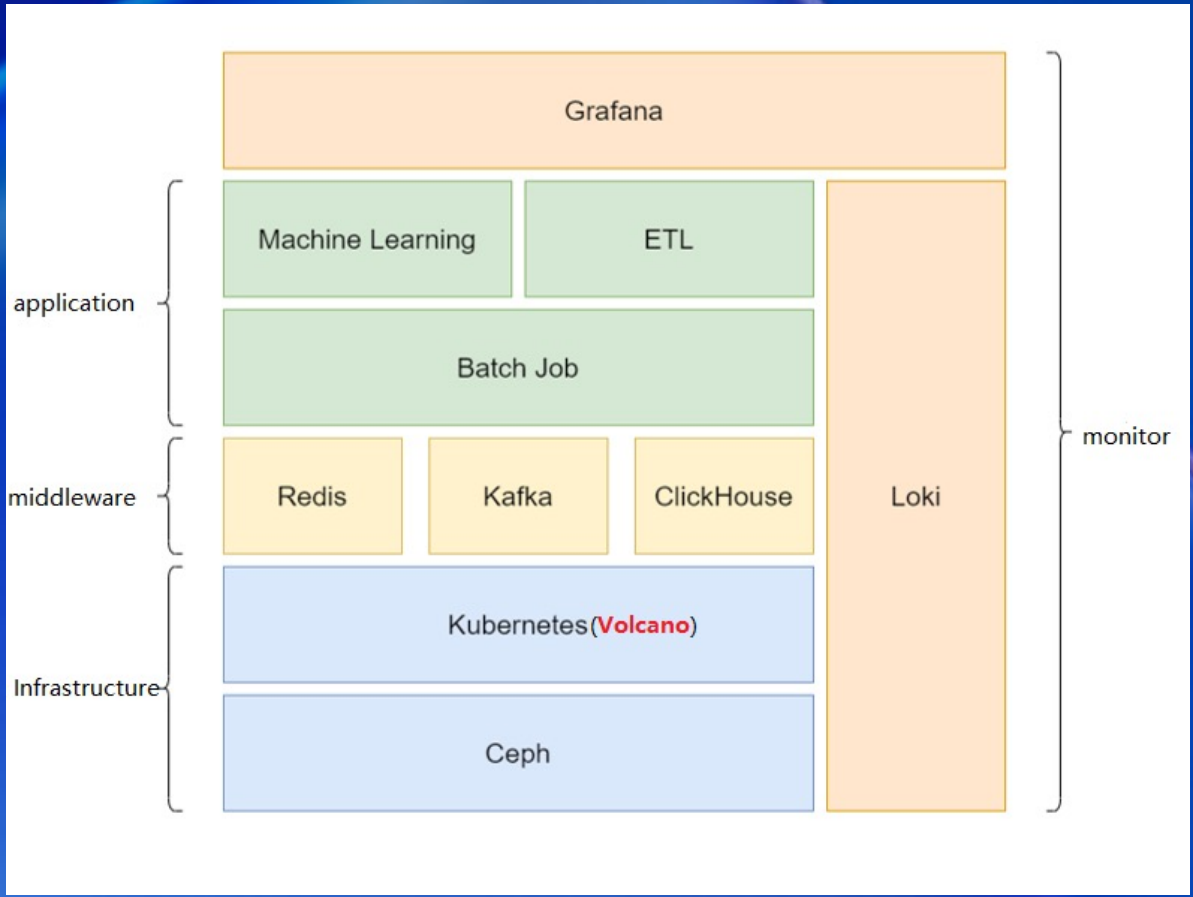
- **首个Batch调度器**
  - ✓ 2022年Volcano成为Spark on kubernetes的首个batch调度器
  - ✓ **1.5K Pod/s** 的大规模批量任务调度能力

Spark基于Volcano的用法


```
FEATURES="org.apache.spark.deploy.k8s.features.VolcanoFeatureStep"
~ bin/spark-submit  \
    --master k8s://https://127.0.0.1:60250  \
    --deploy-mode cluster  \
    --conf spark.executor.instances=1  \
    --conf spark.kubernetes.scheduler=volcano \
    --conf spark.kubernetes.driver.pod.featureSteps=$FEATURES \
    --conf spark.kubernetes.executor.pod.featureSteps=$FEATURES \
    --conf spark.kubernetes.scheduler.volcano.podGroupTemplateFile=/path/to/podgroup-template.yaml \
    --conf spark.kubernetes.namespace=spark \
    --conf spark.kubernetes.authenticate.driver.serviceAccountName=spark-sa  \
    --conf spark.kubernetes.container.image=spark:latest  \
    --class org.apache.spark.examples.SparkPi  \
    --name spark-pi  \
    local:///opt/spark/examples/jars/spark-examples_2.12-3.3.0-SNAPSHOT.jar
```

1. Specify custom scheduler
2. Specify custom feature step
3. Specify scheduler hints (podgroup template)

用户案例：锐天投资基于Volcano的分布式计算平台



业务场景：

- 金融投资公司，业务场景主要为策略研究开发、AI 训练与推理、大数据ETL和离线批处理任务

客户诉求：

- 要求调度系统提供公平机制，满足公司内多团队资源共享，保证各自业务的SLA
- 要求系统提供Gang-scheduling解决基本死锁问题
- 要求调度系统统一支持AI、大数据、Batch Job

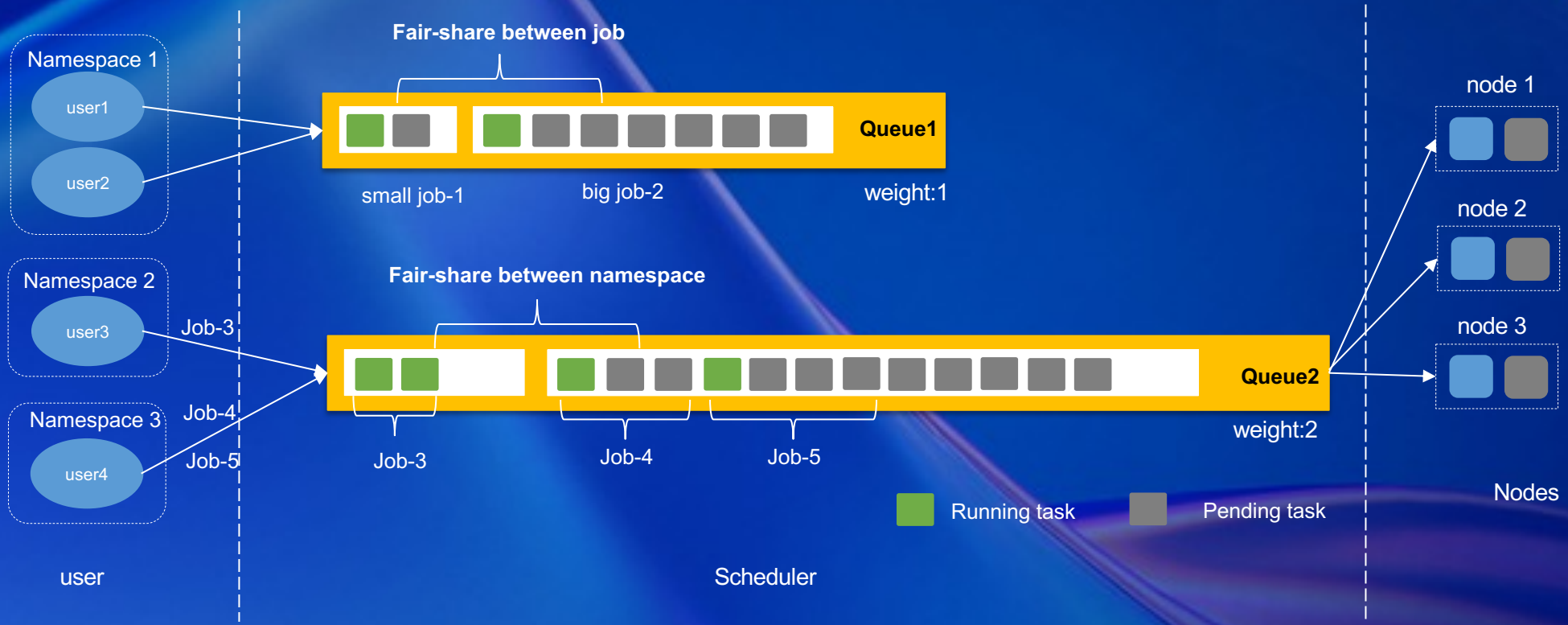
解决方案：

- Volcano 统一支持AI、数据ETL和离线Batch job
- Volcano提供的队列调度、公平调度策略，满足用户的多租户资源共享的诉求

用户收益：

- 使用 Volcano，生产环境稳定支持30w Pod/天增长量
- 基于Volcano二次开发，支持特定业务场景

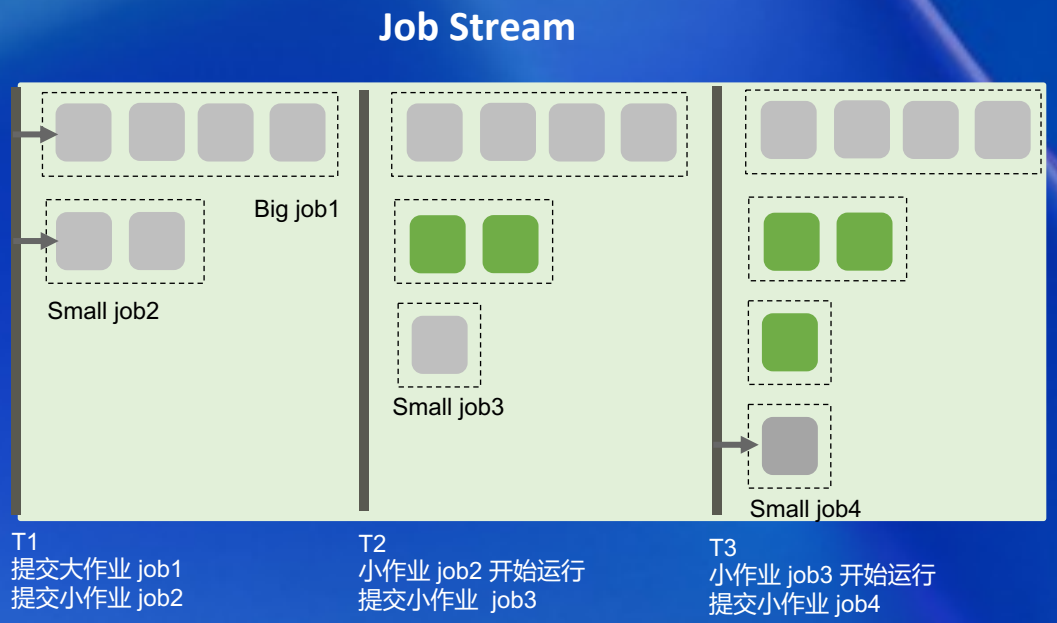
公平调度



- Job间资源共享
- namespace之间资源共享
- 队列级Policy(FIFO, Priority, Fair share , …)

SLA 避免大作业饿死

场景：大作业与小作业共存时，存在饿死问题



解决方案：通过SLA配置作业的最长等待时间，降低大作业饿死的可能性

```
apiVersion: batch.volcano.sh/v1alpha1
kind: Job
metadata:
  name: test-job
  annotations:
    sla-waiting-time:1h
spec:
  minAvailable: 5
  tasks:
    - replicas: 5
      template:
  ... ...
```

丰富的调度算法

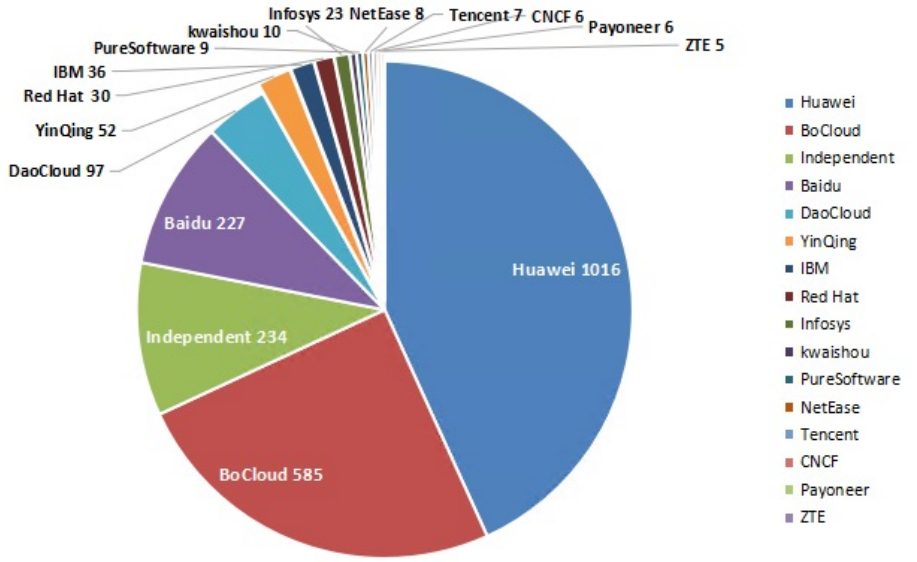
- Gang-Scheduling
- Job priority
- Job queue
- Job order
- Preemption
- backfill
- Job Fair-share
- Namespace fair-share
- Task-topology
- IO-Awareness
- Resource reservation

- SLA
- GPU sharing
- NUMA-Awareness
- HDRF
- Hierarchy Queue
- Co-location
- Elastic scheduling
- TDM
- Proportional scheduling
- ……

Volcano 社区

社区用户示例

Volcano community **Top** contribution orgination



Data from https://volcano.devstats.cncf.io

加入社区

KUBERNETES
COMMUNITY DAYS DALIAN

Website: https://volcano.sh/en/

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

Slack Channel: https://volcano-sh.slack.com/

容器魔方公众号

加入微信群