云原生批量计算平台 Volcano

Cloud Native Batch Computing

王雷博 Volcano技术负责人

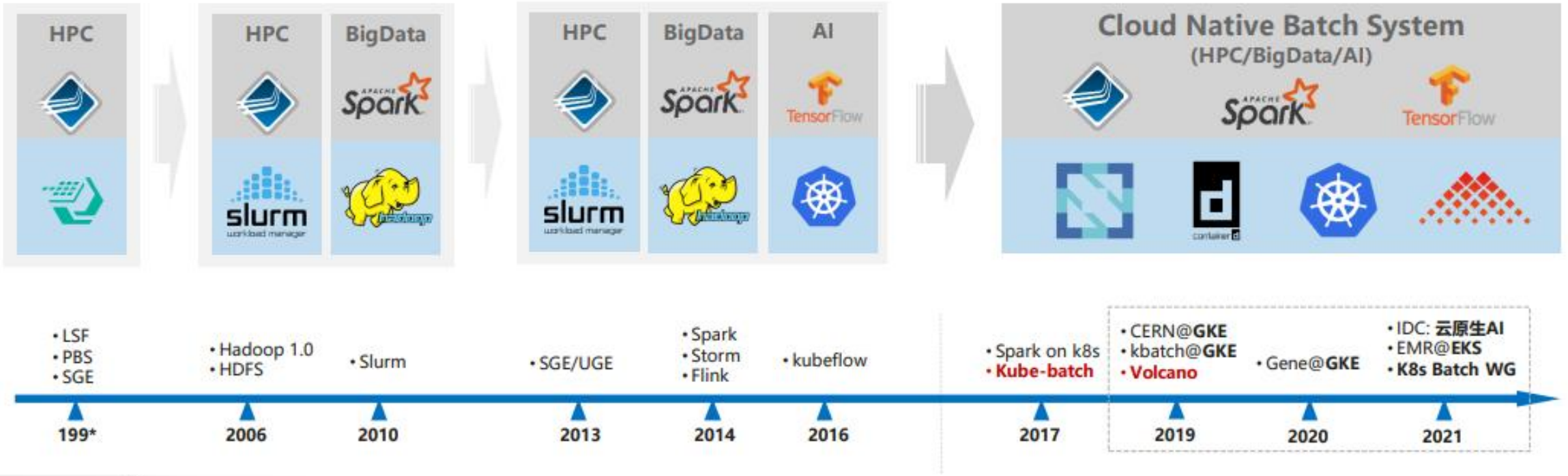
01

Volcano项目介绍

云原生批量计算的发展历程

随着业务场景的不断丰富，批量计算也由传统的HPC逐渐扩展到大数据、AI等多种场景；但各个领域独立发展，呈现出**生态割裂，技术栈不兼容，资源使用率低**等问题，严重影响批量计算的进一步发展。

云原生技术以其**丰富的生态**和**灵活的扩展性**受到各个社区及厂商的青睐，并以云原生技术为基础构建统一的批量计算系统，**提升资源使用率**



云原生批量计算面临的挑战

作业管理

- 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

领域框架支持

- 对领域计算框架支持不足，e.g. mpi, tensorflow, mxnet, pytorch等
- 1:1的operator部署，运维复杂

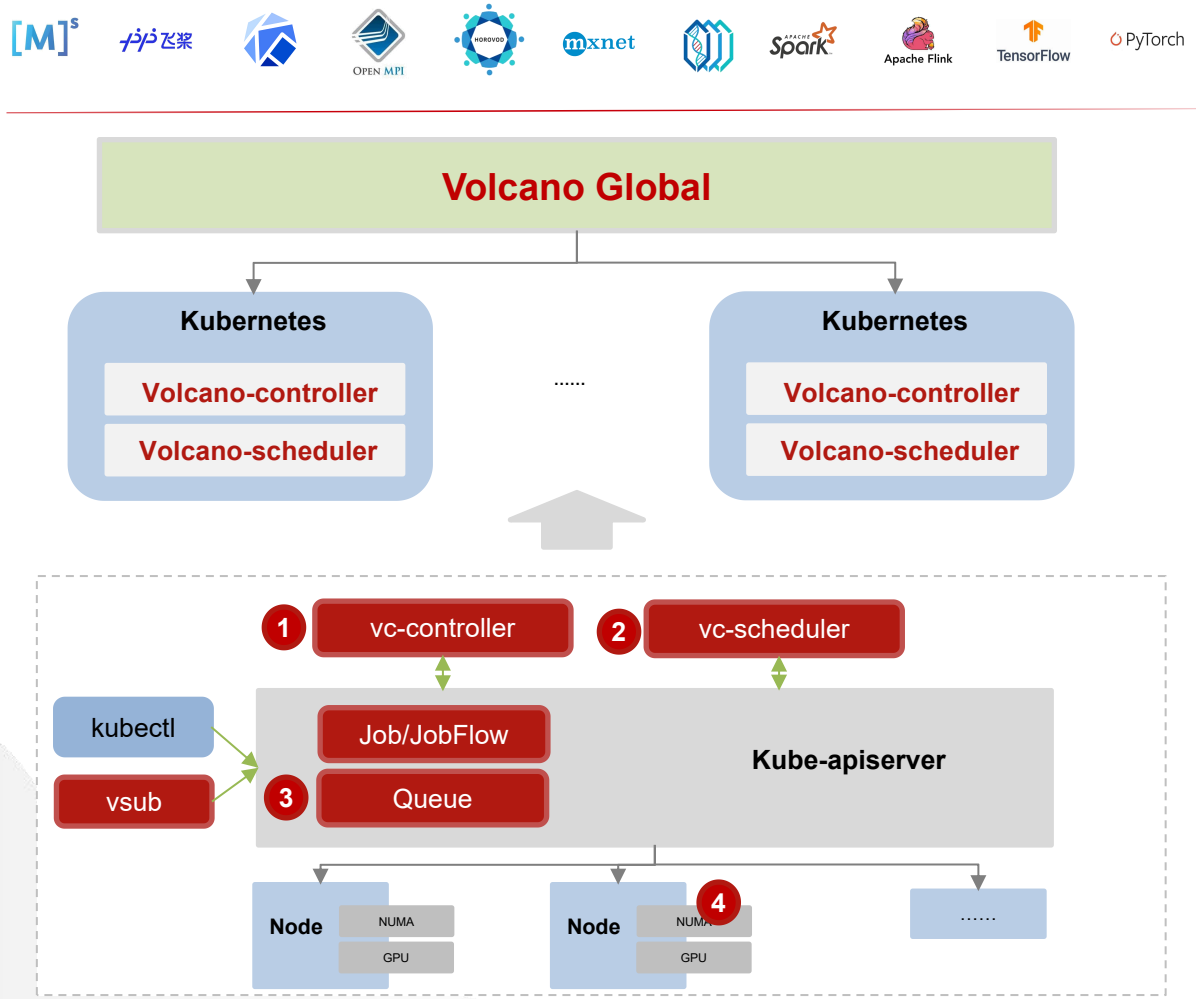
调度和性能

- 缺少Job为base的调度策略, e.g. job ordering, job priority, job preemption, job fair-share
- 缺少高级调度策略, e.g. task-topology, IO-Awareness, backfill
- 性能无法满足batch需求，e.g. throughput, roundtrip

资源共享、异构计算

- 缺少作业队列的概念
- 缺少灵活的集群资源规划，提供多租户场景下资源的公平使用以及高效复用
- 缺少对异构资源的深度支持

Volcano 架构



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-operator, Flink-operator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。

Volcano 社区

VOLCANO

- 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

VOLCANO CLOUD NATIVE COMPUTING FOUNDATION THE LINUX FOUNDATION 开源软件学园

作业管理：Job

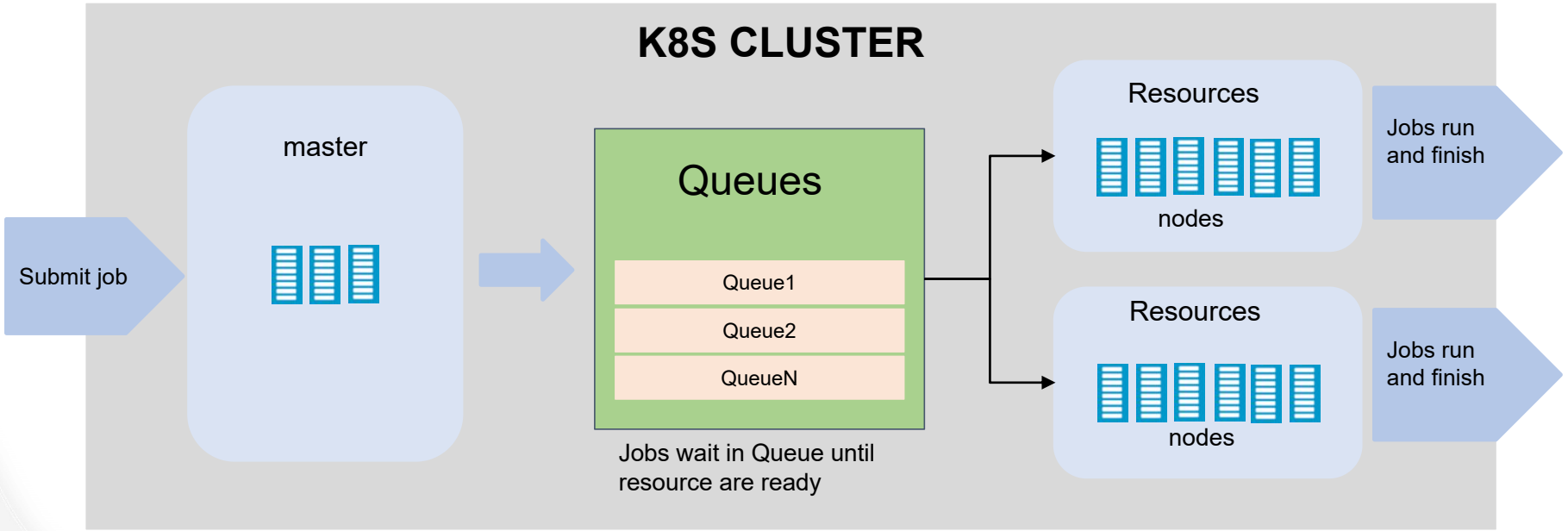
Volcano Job:

- 统一类型接口，支持业界主流作业类型，如mpi, pytorch, tensorflow, mxnet等
- 细粒度作业生命周期管理
- 易扩展的作业插件机制
  - Env
  - Svc
  - Ssh
  - Tensorflow
- Coordinate with Scheduler
- 提供作业依赖支持

```yaml
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
```

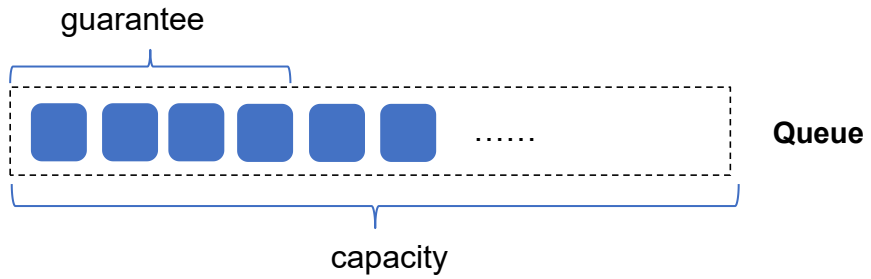
资源共享： Queue

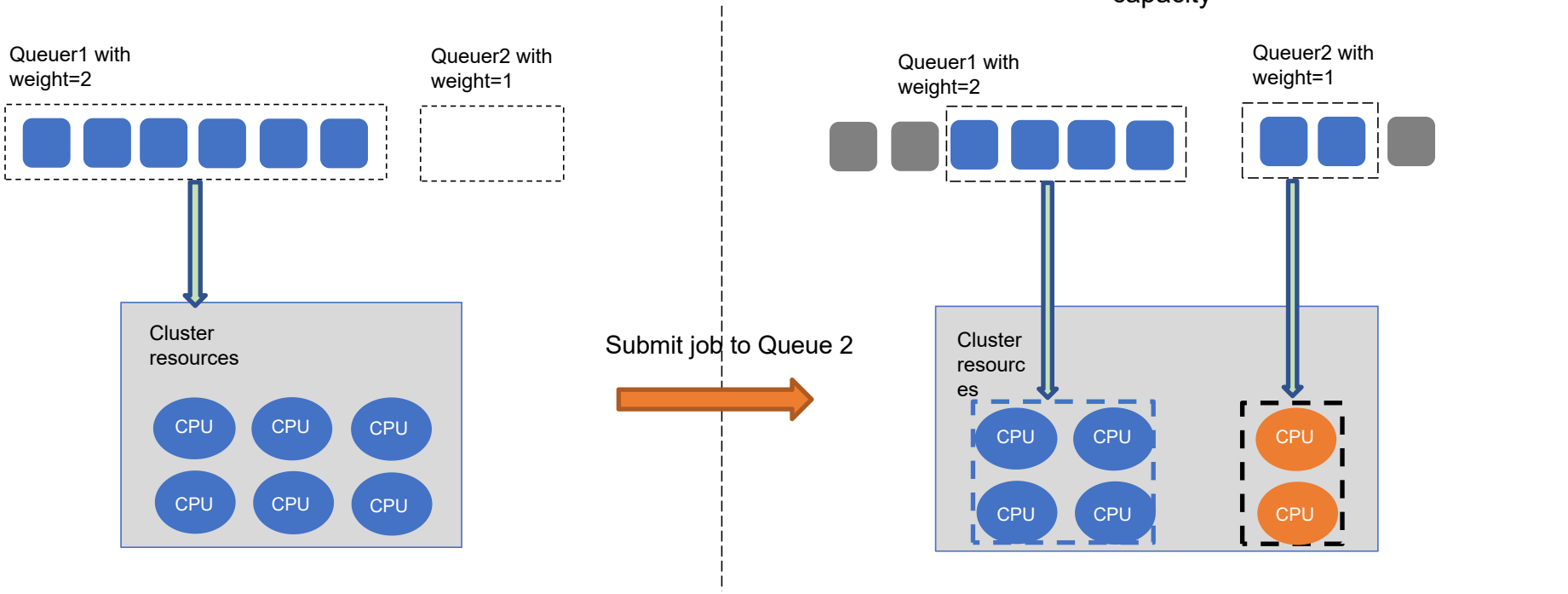
- 集群级别资源对象，与用户/namespace解耦
- 可用于租户/资源池之间共享资源
- 支持每个队列独立配置Policy，如 FIFO, fair share, priority, SLA等



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

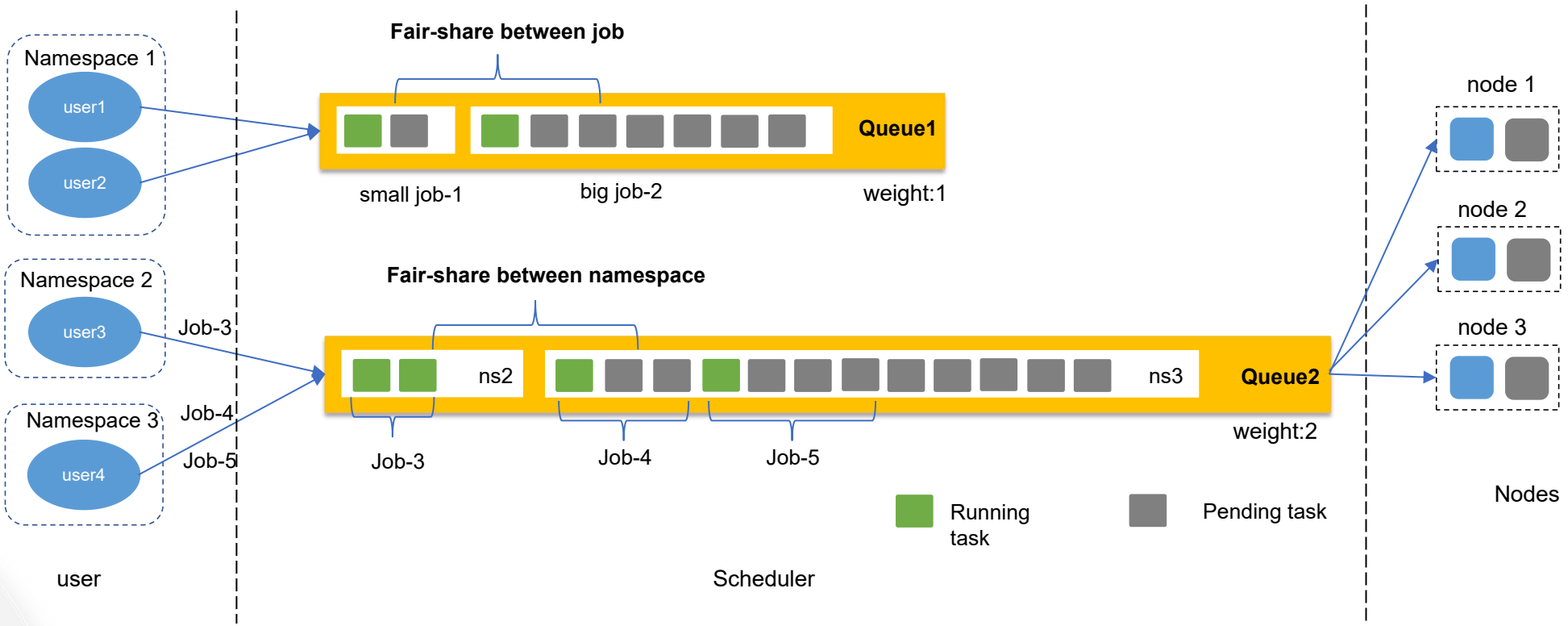




Queue2 is empty. Q1 can borrow resources from Queue2.

Queue2 has workload, it will reclaim resources from Queue1.

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 , …)

丰富的调度策略

- 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
- ……

02

Volcano典型场景及用户案例

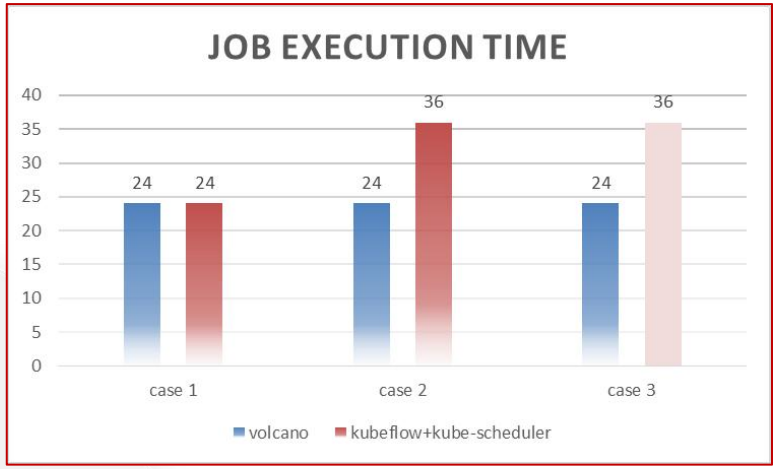
场景：分布式训练、大数据

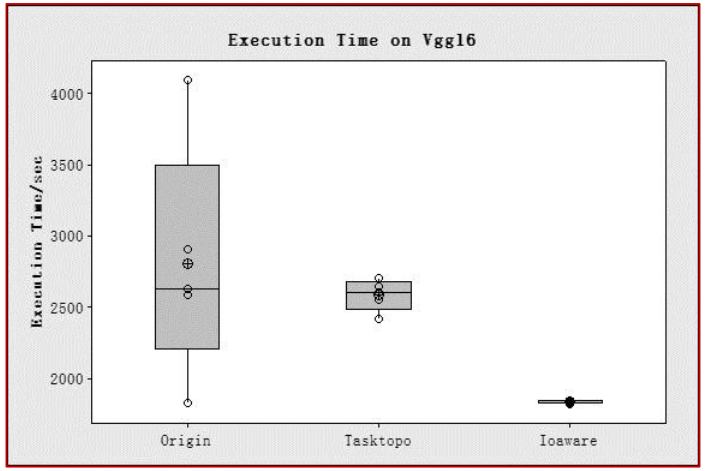
AI 分布式训练场景

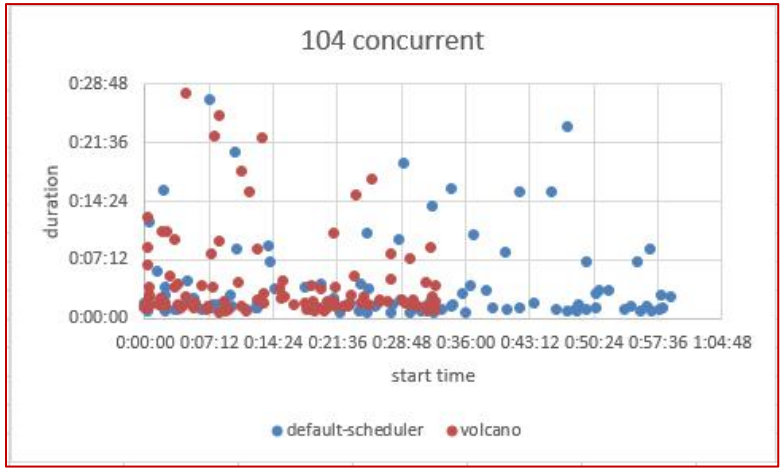
- **Gang-Scheduling**：解决分布式训练ps-worker忙等、死锁问题，性能提升30%
- **Task-topology/ IO aware scheduling**：最大程度降低传输时延，针对IO密集型应用，性能提升31%

大数据场景

- **minResource**： 解决高并发场景Spark driver和executor资源竞争问题，合理规划并行度，性能提升39.9%





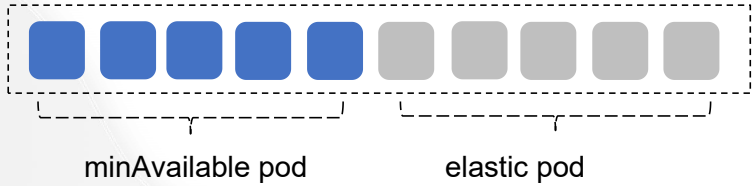


分布式训练场景

场景：弹性调度

- **弹性作业**

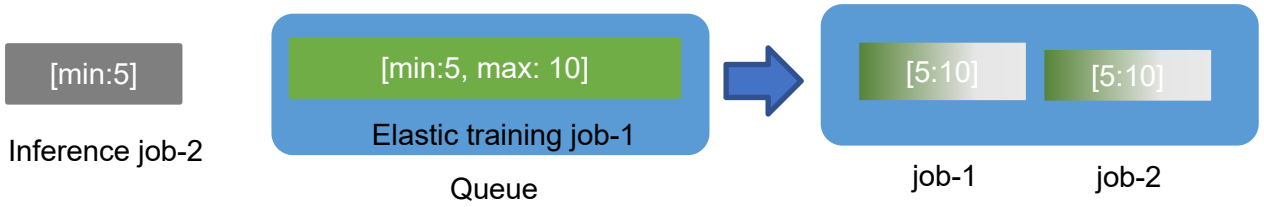
```
apiVersion: batch.volcano.sh/v1alpha1
kind: Job
metadata:
  name: test-job
spec:
  minAvailable: 5 #min
  tasks:
    - replicas: 10  #max
      template:
        spec:
          containers:
            - image: train_script
              resources:
                  nvidia.com/gpu: 1
... ...
```



场景：弹性作业和并行作业混部，提高资源利用率

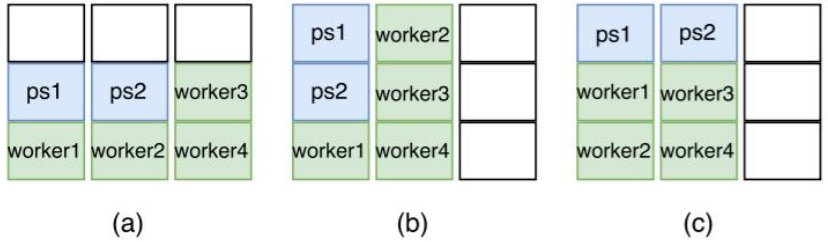
样例：

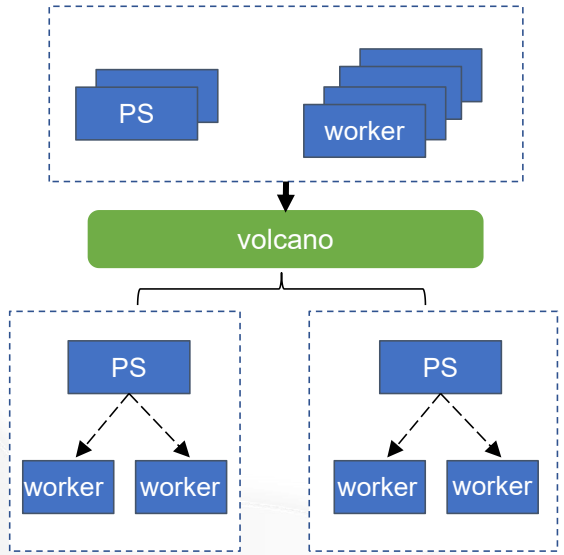
- 推理作业 job-2（高优先级）抢占弹性训练作业job-1（低优先级）的资源，保证SLA.
- 空闲资源增多后，弹性训练作业job-1调度更多Pod



场景：作业拓扑感知调度

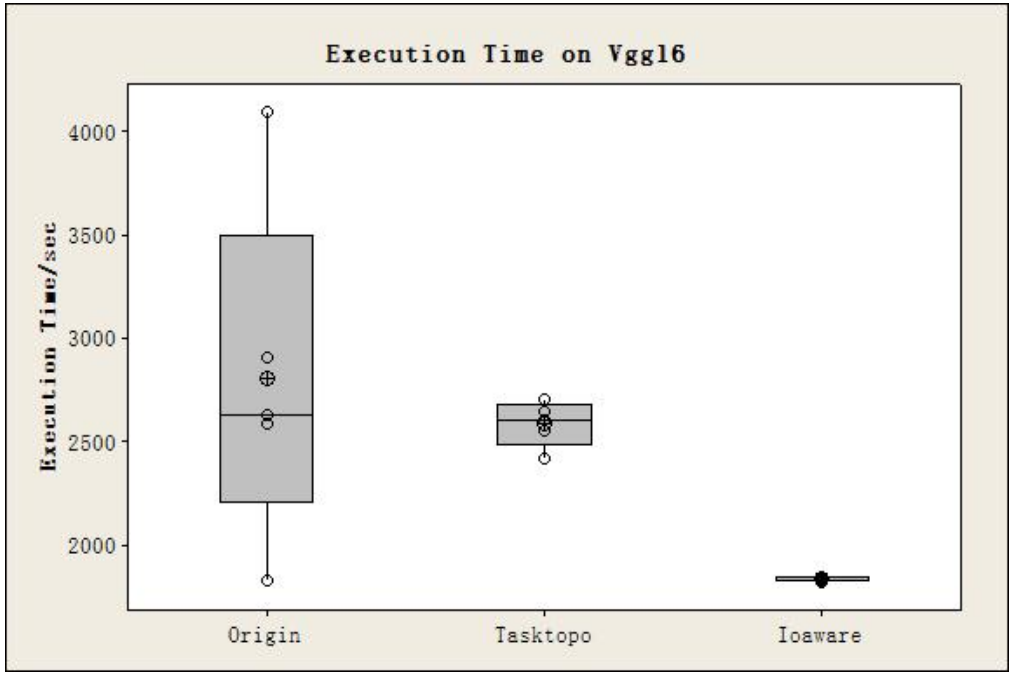
- The execution time of 3 jobs in total; 2ps + 4workers for each job in k8s



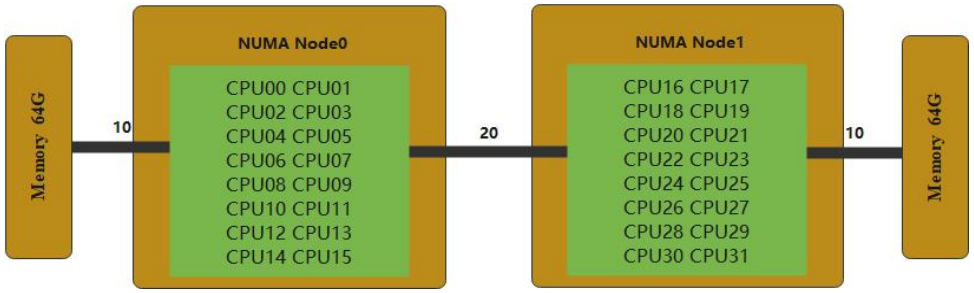


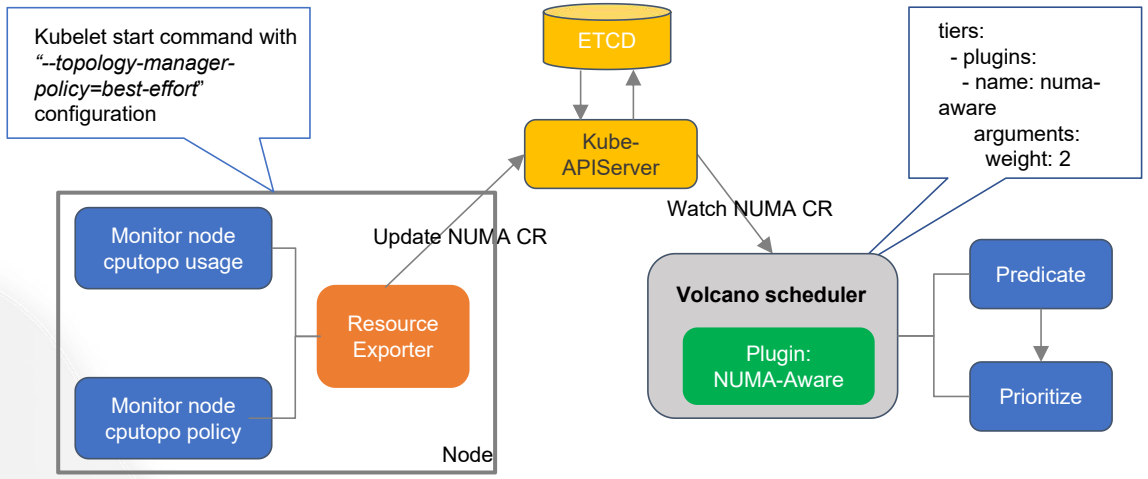
Task-topology scheduling

- IO-Awareness: minimized the maximum of communication cost between any two nodes.



场景：CPU拓扑感知调度





NUMA-Aware scheduling

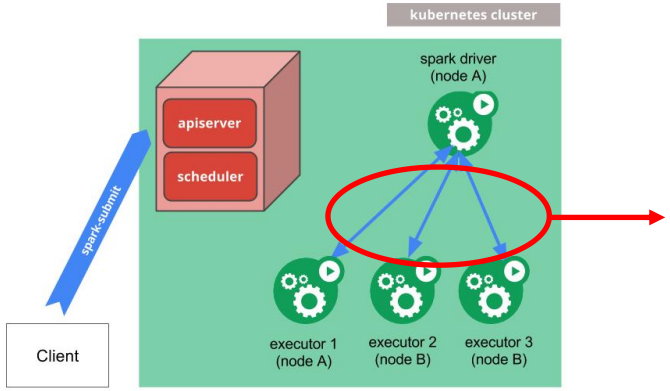
Numa感知

支持感知CPU Numa拓扑，并提供对Numa的调度

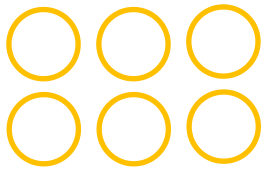
场景

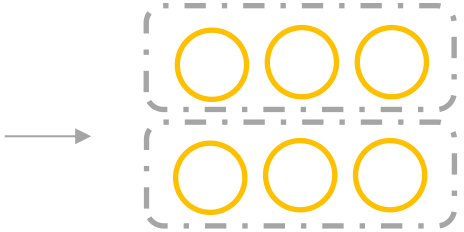
适用于科学计算，转码、渲染、大数据离线分析等计算密集型业务场景

场景：为Spark提供批量调度

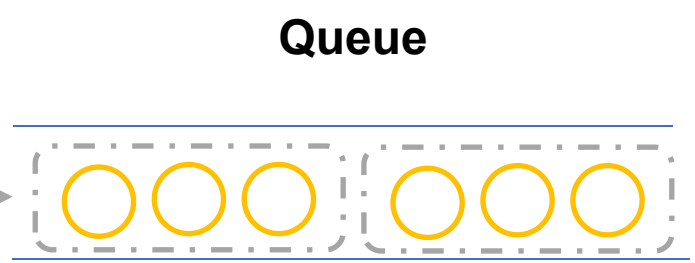


More k8s native resource support





- Schedule based on **Pod Group**
- Minimal Resource reservation（CPU/MEM）
- Job priority

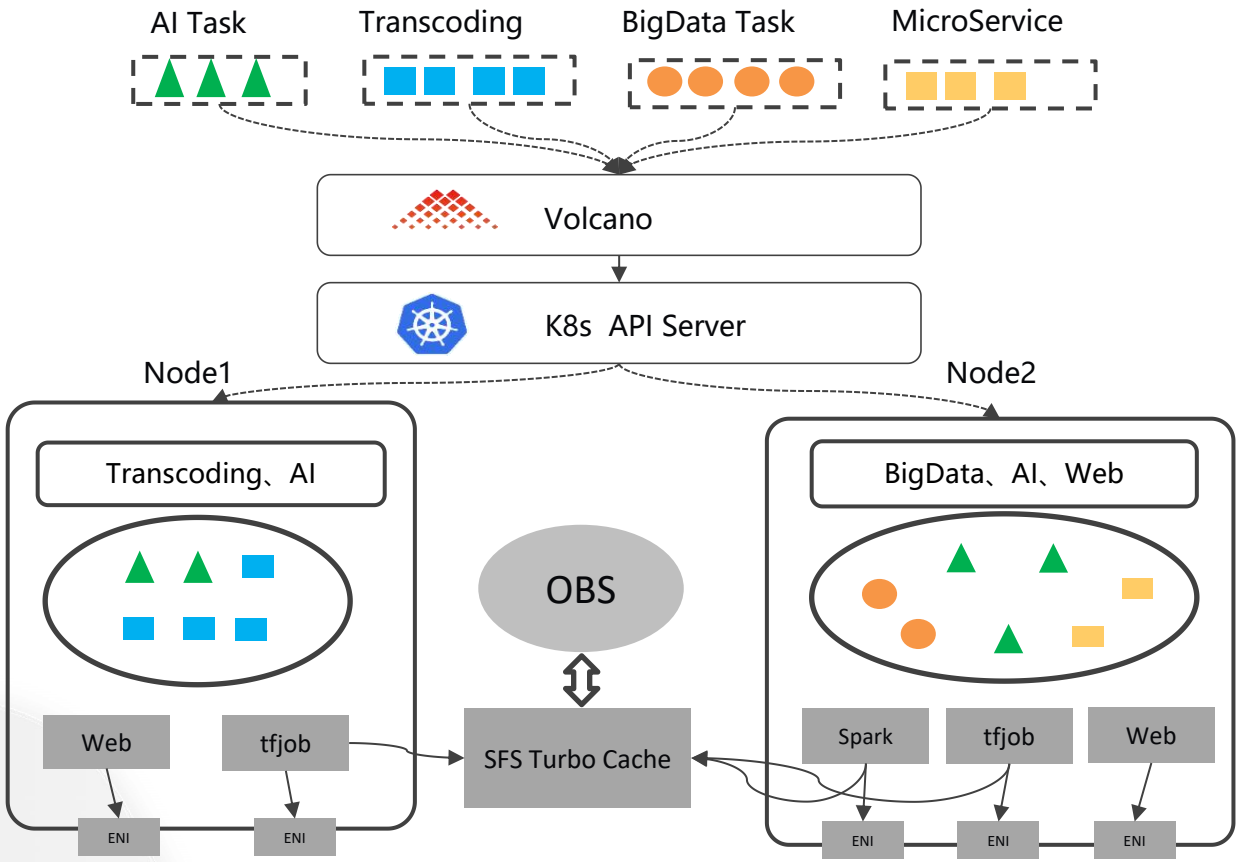


- Introduce Queue
- Multi-tenants
- Priority
- Fair-share
- Preempt
- ……

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

场景：在离线作业混部



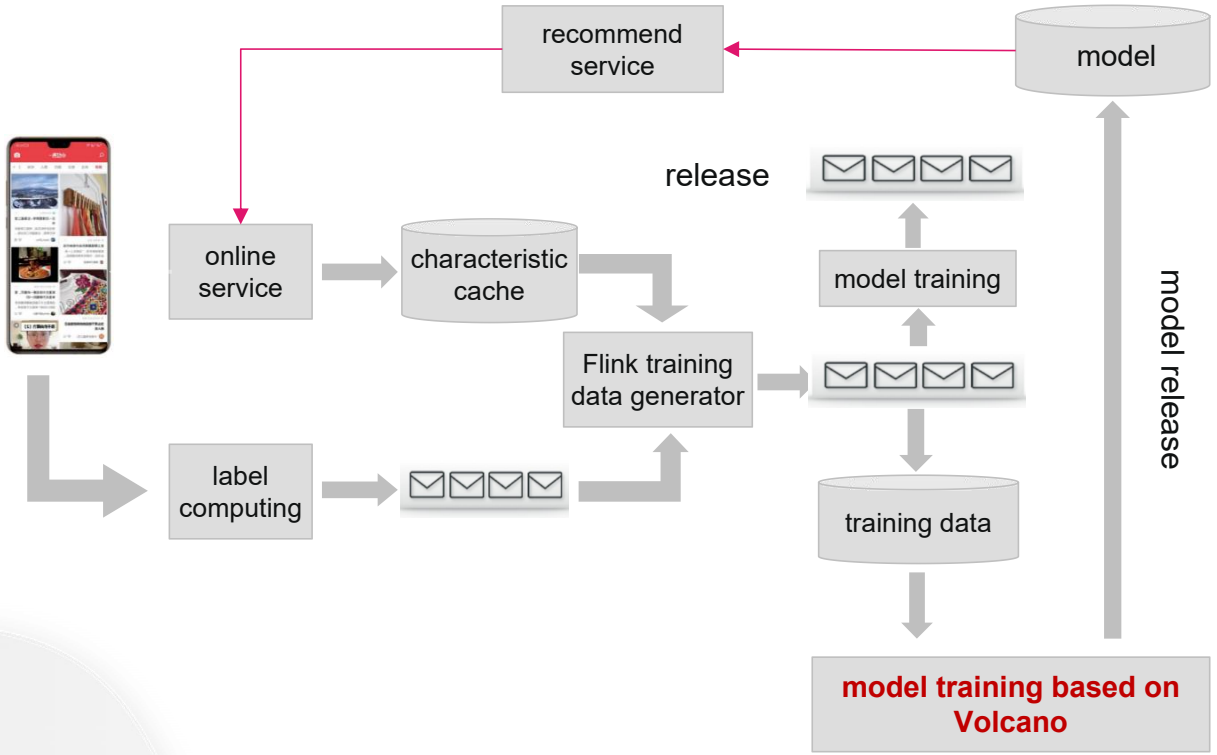
场景需求：

- 在线服务有峰谷，资源请求根据峰值设置
- 有些用户不知道其服务的资源使用情况，请求过多的资源。
- 高分配率，低利用率，CPU平均利用率小于15%

关键技术

- qos-feature：操作系统级别的QoS支持
  - qos/CPU 快速抢占
  - 具有优先级的Qos内存回收
  - qos/NET和IO监控和隔离
- kubelet：更新/更改Qos模型，支持真实业务负载上报、资源超分、作业驱逐。
- 调度器：
  - 多类型业务智能混合调度
  - 基于时间的调度策略，实现资源分时复用

案例：小红书基于Volcano优化AI训练平台



业务场景：

- 作为月活过亿的顶级社交媒体和电商平台，推荐是小红书的核心业务之一。小红书AI平台由包含在、离线训练，同时承担数十万样本分析和模型训练作业。模型生成时间在分钟量级。

客户诉求：

- 训练集群计算节点数千量级；
- 推荐模型参数千亿量级；
- 单个训练作业包含ps和worker百量级；
- 需求最佳拓扑调度和性能表现；

解决方案：

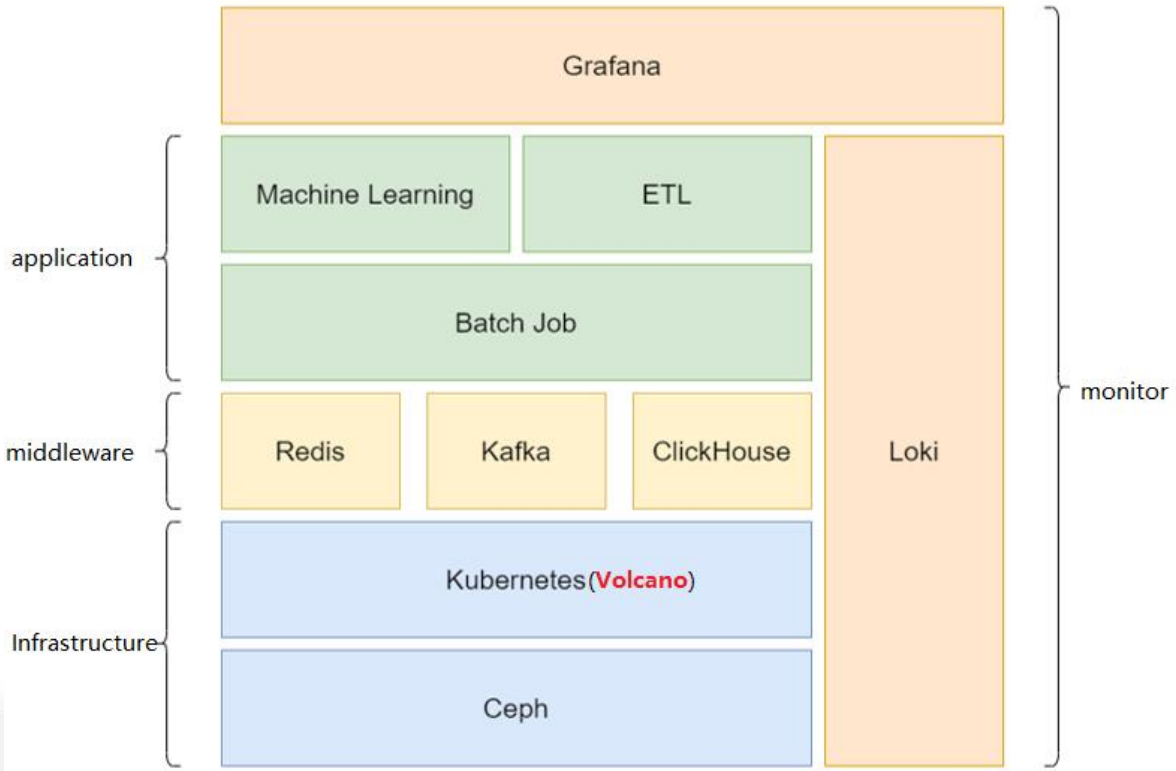
- 采用binpack和task-topology调度策略，保障最佳拓扑部署；
- 采用gang和SLA调度策略，避免资源竞争，提供防饿死机制；

方案价值：

- 应用Volcano后，总体AI训练速度提升20%；
- AI训练平台吞吐量提升20%；
- 避免大作业饿死；

VOLCANO CLOUD NATIVE COMPUTING FOUNDATION THE LINUX FOUNDATION 开源软件学园

案例：锐天基于Volcano构建大规模计算平台



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：

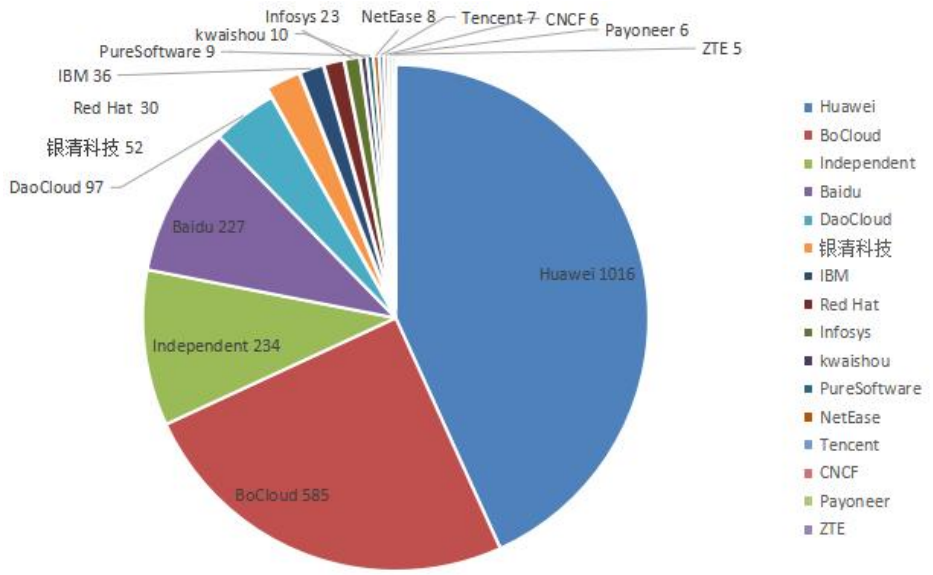
- 要求调度系统提供公平机制，满足公司内多团队资源共享，保证各自业务的SLA
- 要求系统提供Gang-scheduling解决基本死锁问题
- 要求调度系统统一支持AI、大数据、Batch Job

方案价值：

- 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
- 使用 Volcano，生产环境稳定支持30w Pod/天增长量
- 用户基于Volcano进行深度二次开发

Volcano社区

Volcano 社区年度**Top**贡献 企业、机构



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

CLOUD NATIVE COMPUTING FOUNDATION

VOLCANO CLOUD NATIVE COMPUTING FOUNDATION THE LINUX FOUNDATION 开源软件学园

社区版本发布

Volcano: A Kubernetes Native Batch System

v1.0:
- GPU Sharing
- Preempt And Reclaim Upgrade to Beta
- Job Dynamic Scale Up And Down
- Integrate With Flink Operator
- DAG Job Based On Argo

2020.7

v1.4:
- Support Multiple Volcano Scheduler
- Support CPU NUMA-Aware Scheduling
- Support Proportion Of Resource For GPU Node
- ……

2021.9

v1.5:
- Reserve Resource for Queue
- Support spark Native As Custom Scheduler
- Support Task Dependency
- Support Specified Nodes for Volcano in Cluster
- Add Tensorflow Job Plugin
- ……

2022.H1

v1.6:
- Resource management
  - Hierarchy Queue
- Inference
  - GPU virtualization
- Training
  - Elastic scheduling
- Support Job DAG scheduling
- Stability Enhancement
- Configuration hot update
- ……

2022.H2

v1.7:
- Enhanced Plugin for PyTorch Jobs
- Ray on Volcano
- Enhance Scheduling for Kubernetes long-running services
- Support multi-arch images for Volcano
- Optimize Queue Status Information
- ……

2023.H1

v1.x:
- Heterogeneous computing power enhancement
- Jobflow orchestration engine
- Hierarchy Queue
- Multi-cluster scheduling
- Data affinity scheduling
- Utilization optimization
  - Intelligent co-location
  - Efficient auto-scaling
  - Small Job Backfill
- Performance
  - Improve Throughput
  - Reduce Roundtrip

202x

VOLCANO CLOUD NATIVE COMPUTING FOUNDATION THE LINUX FOUNDATION 开源软件学园

加入社区

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