KubeCon

CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024

KubeCon

CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024

通过 Volcano 增强的智能基础设施优化LLM工作流程

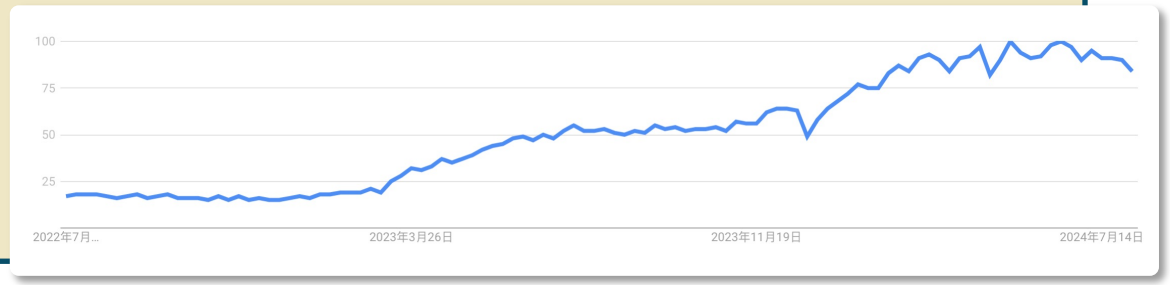
李鑫, Qihoo360

常旭征, Huawei Cloud Technologies Co., LTD

目录

背景

LLM关键词趋势

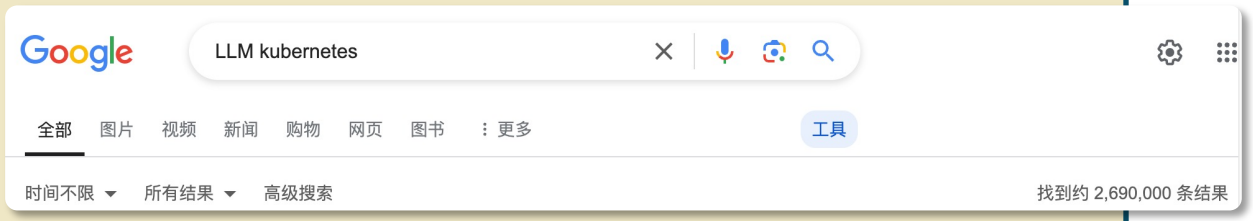


OpenAI博客文章

Scaling Kubernetes to 2,500 nodes 2018

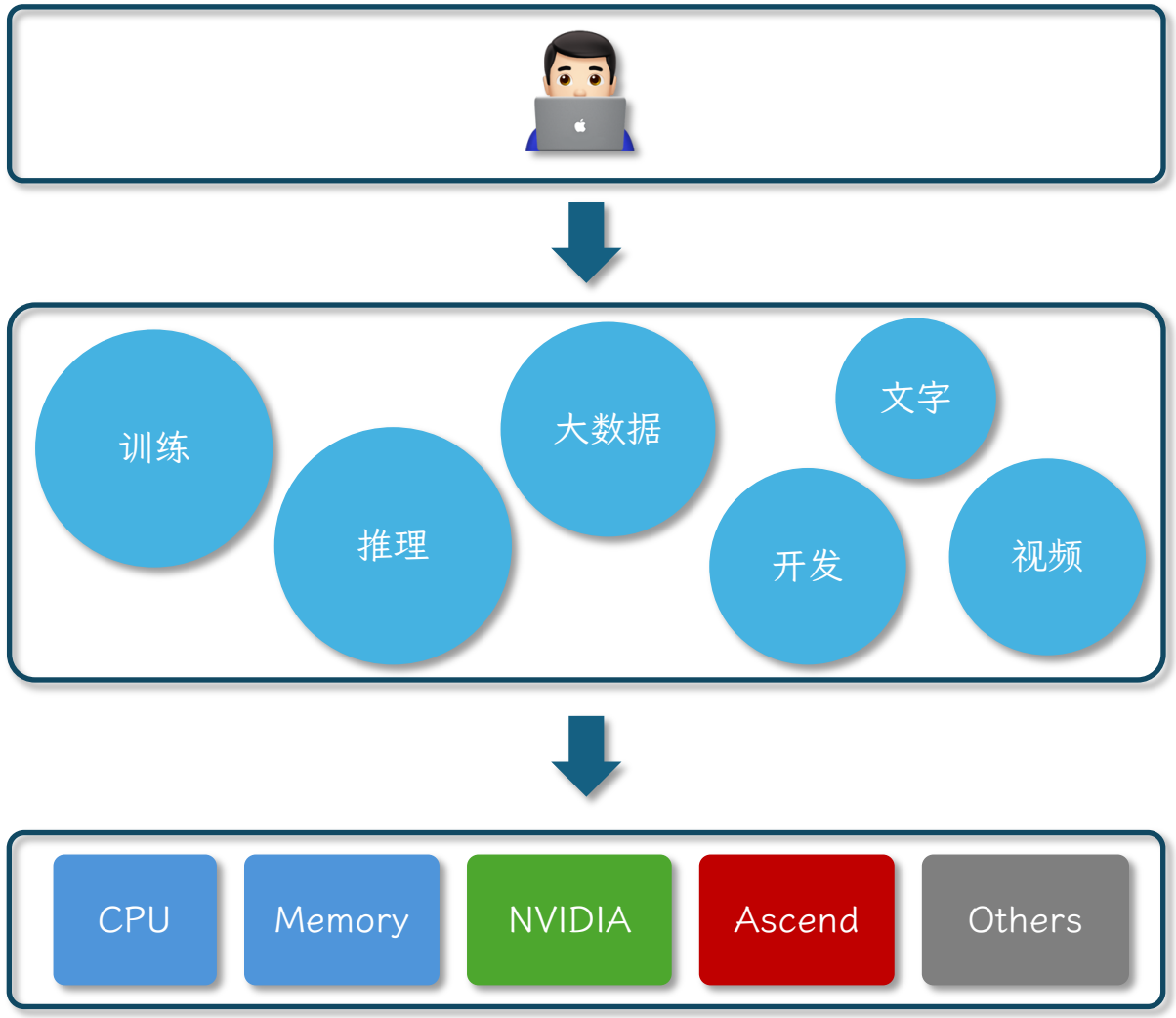
Scaling Kubernetes to 7,500 nodes 2021

Google搜索结果



- 从 2023 年开始，LLM 受到的关注越来越多
- 使用 Kubernetes 的 LLM 基础设施越来越多
- kubernetes 对于 LLM 的支持越来越好

现状



x3000

x6000/月

x1000

- 3000+不同部门用户，6000+任务每月
- 10+集群，1000+节点
- 任务类型复杂性。训练，推理，开发。
- 资源：单任务1-200实例，单实例CPU：1c-200c，GPU：1-8，内存20G-2T
- 功能：ssh免密，pod间打通
- 运行：所有实例同时调度
- 运行时长复杂性。时计天计月计共存。
- 计算资源复杂性。CPU，GPU，NPU等
- 网络环境复杂性。以太网、IB，RoCE

存在问题

故障

效率

易用性

故障

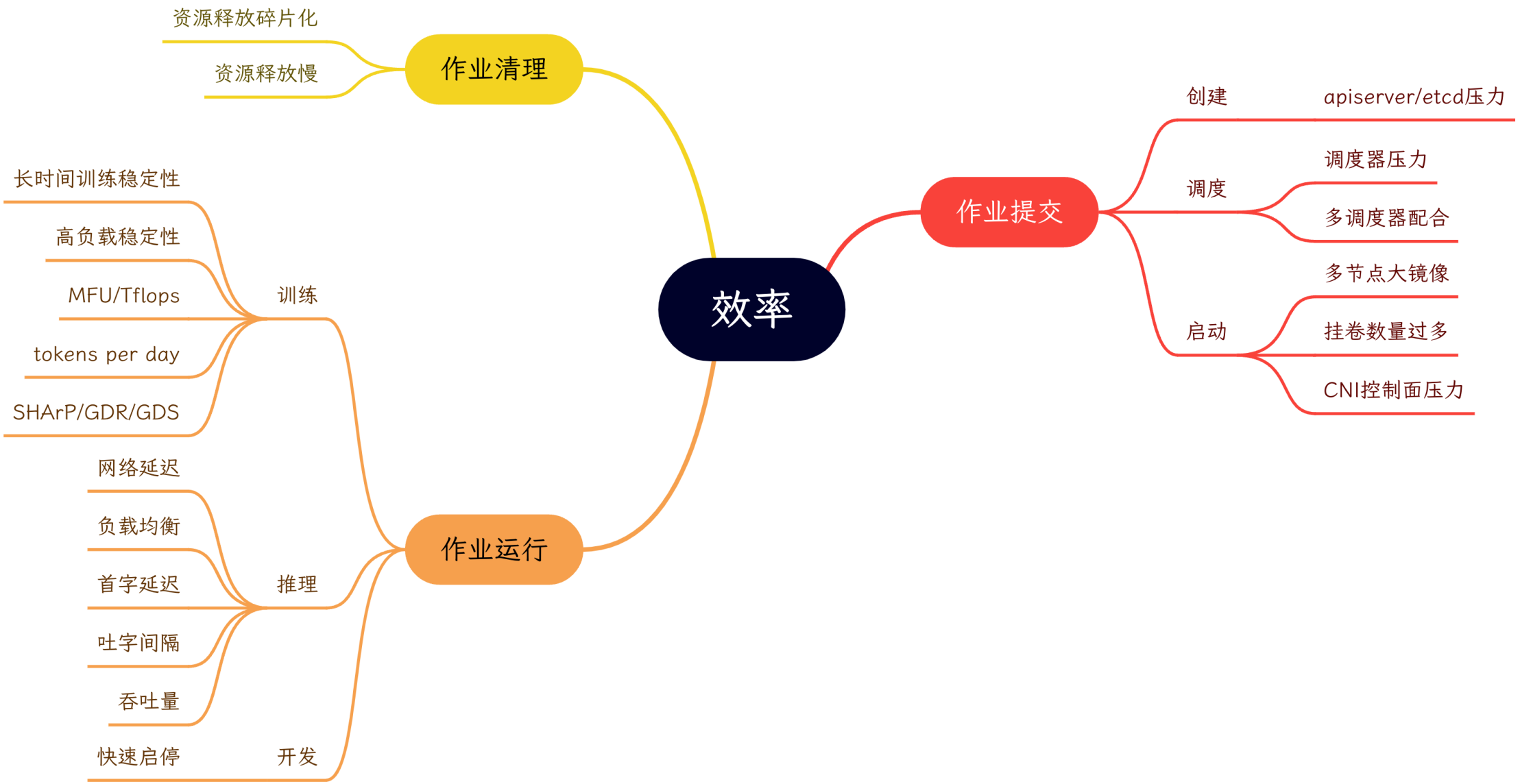
- 掉卡
- ECC 错误
- 显卡故障
- 网卡故障
- 机房掉电
- 误操作
- NAS异常
- 集群故障
- NVLINK故障
- P2P故障
- 散热故障
- ……

The Llama 3 Herd of Models

During a 54-day snapshot period of pre-training, we experienced a total of 466 job interruptions. Of these, 47 were planned interruptions due to automated maintenance operations such as firmware upgrades or operator-initiated operations like configuration or dataset updates. The remaining 419 were unexpected interruptions, which are classified in Table 5. Approximately 78% of the unexpected interruptions are attributed to confirmed hardware issues, such as GPU or host component failures, or suspected hardware-related issues like silent data corruption and unplanned individual host maintenance events. GPU issues are the largest category, accounting for 58.7% of all unexpected issues. Despite the large number of failures, significant manual intervention was required only three times during this period, with the rest of issues handled by automation.

https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

效率


效率
作业清理
资源释放碎片化
资源释放慢
作业运行
训练
长时间训练稳定性
高负载稳定性
MFU/Tflops
tokens per day
SHArP/GDR/GDS
推理
网络延迟
负载均衡
首字延迟
吐字间隔
吞吐量
开发
快速启停
作业提交
创建
apiserver/etcd压力
调度
调度器压力
多调度器配合
启动
多节点大镜像
挂卷数量过多
CNI控制面压力

易用性

开发过程

- 巨量数据传输
- 环境依赖
- 环境保存
- 多种 IDE 集成
- tensorboard，
- grafana
- 可观测性优化

调度策略优化

- 多部门资源分配
- 独占资源/公共资源
- 任务抢占
- 任务排队
- gang调度策略
- binpack调度策略

多种任务类型

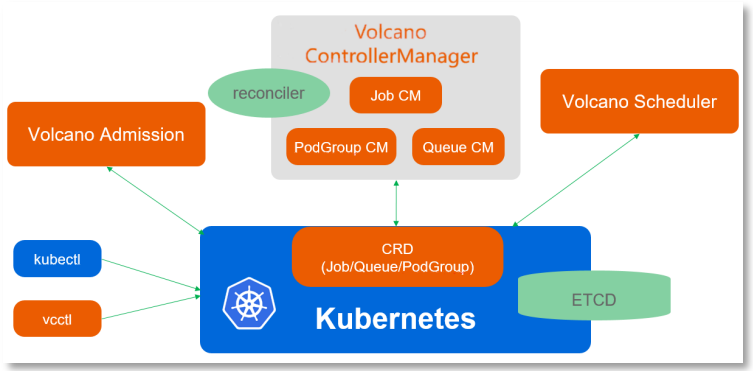
- Megatron-LM
- DeepSpeed
- opensora
- 分布式训练任务
- LLM 任务
- 多模态任务
- 数据处理

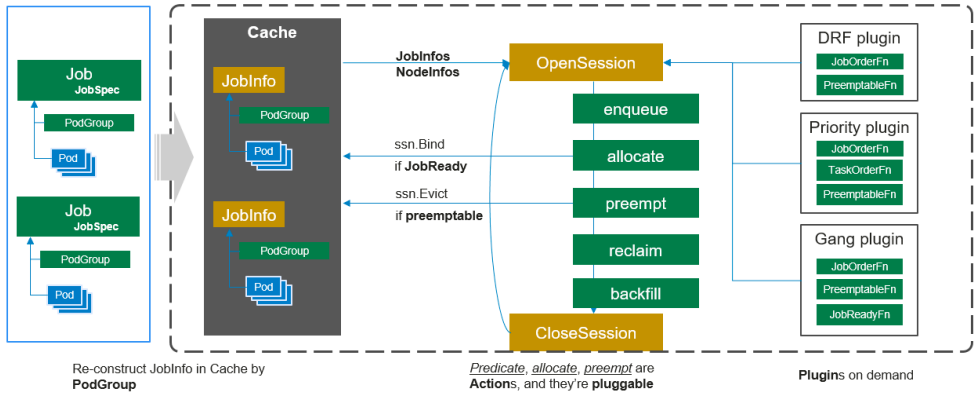
多种硬件

- 单机单卡，单机多卡，多机多卡任务
- NVIDIA
- Ascend
- 纯 CPU 任务
- RoCE/IB
- GPU 切分

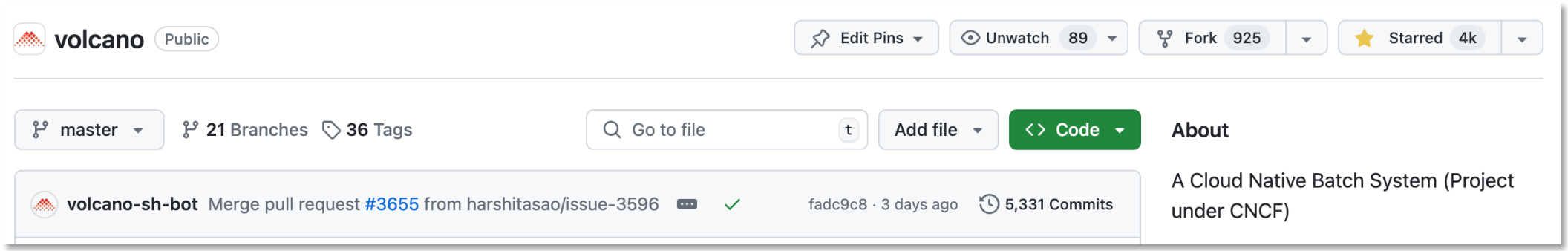
如何解决

Volcano介绍



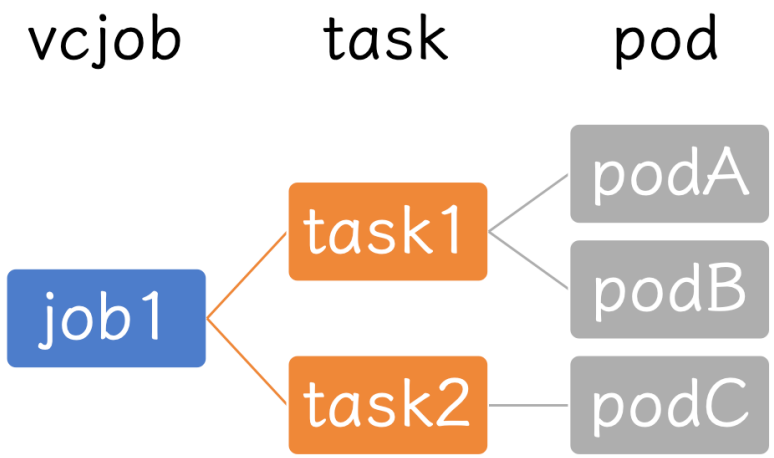


- volcano 项目是华为开源，CNCF 孵化项目。
- 依照 AI 作业特性，建立新的资源抽象。弥补 Kubernetes 原生不足。支持queue，支持一个 job 内多 task，更好的支持批量作业
- 支持多种计算资源，包括但不限于CPU、GPU、NPU。
- 支持多种训练框架，tensorflow，pytorch，飞桨等
- 支持多种调度策略，并且支持多种调度策略组合，比较重要的有gang，priority，binpack，drf等。并且兼容 kubernetes 原生调度策略，例如镜像感知。



易用-vcjob

更好的批处理作业



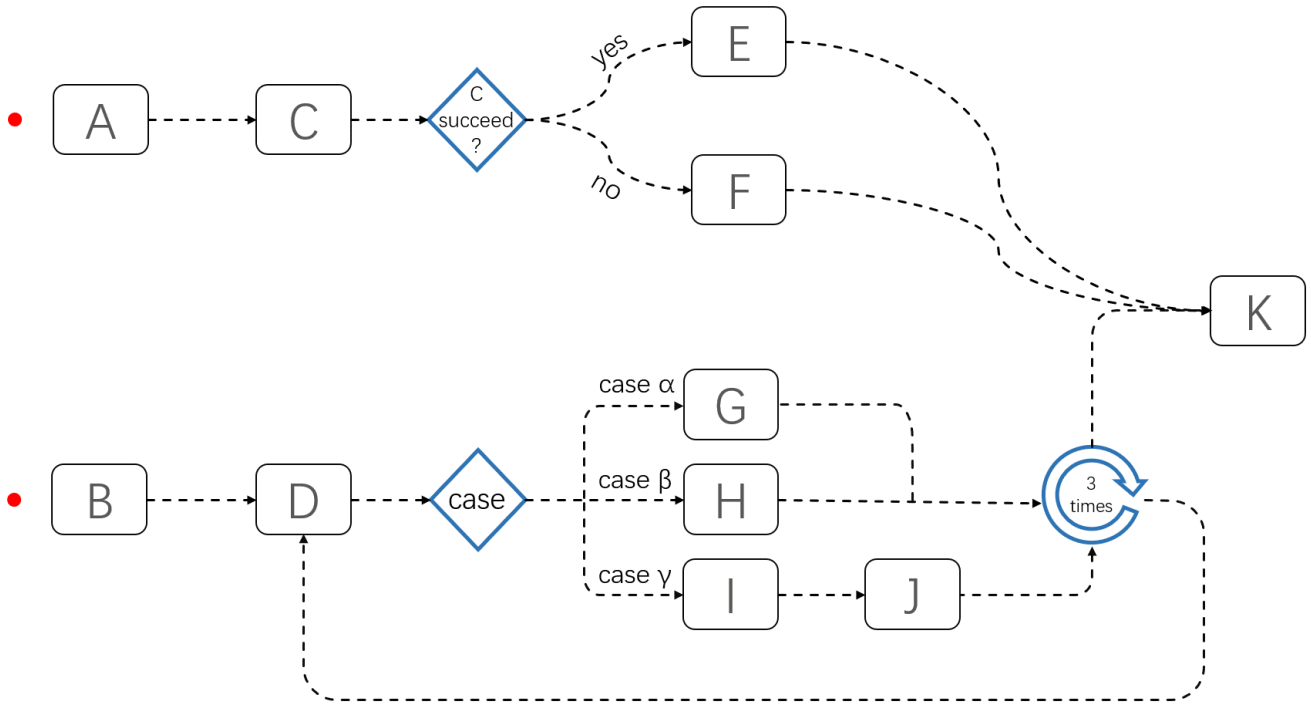
- 一个 vcjob 包含多个 task，每个 task 是不同的角色，一个 task 包含多个 pod
- task1 和 task2 可类比与 tensorflow 的 ps/worker

插件

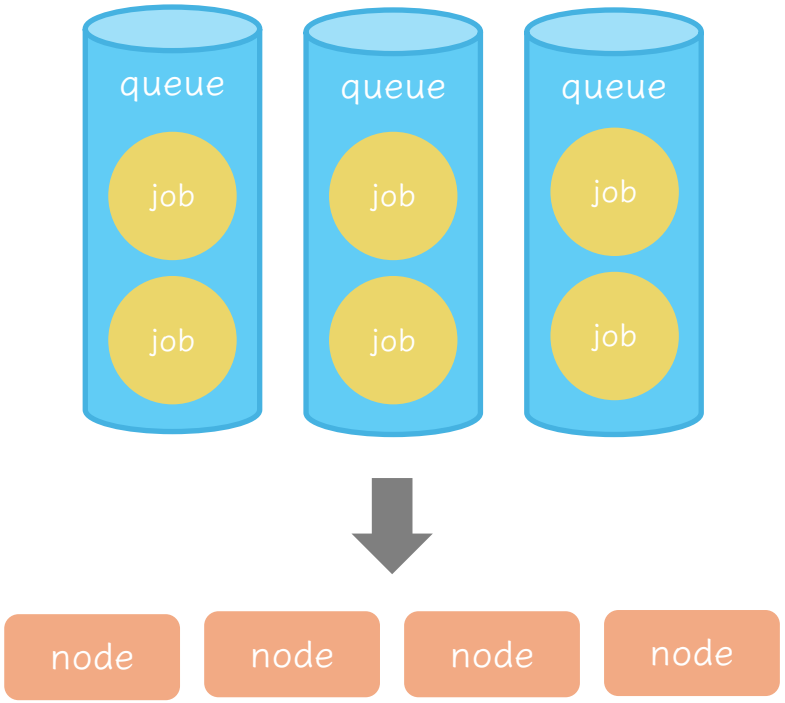
- ssh：pod ssh 免密
- env：创建 pod 索引的环境变量
- svc：为 vcjob 创建 svc 和 networkpolicy
- Pytorch：开启 svc 插件，打开端口，在 pod 中创建 》ytorch 使用的环境变量
- Mpi：强制开启 svc、ssh，打开端口
- Tensorflow：开启 svc 插件，打开端口，在 pod中 创建 Tensorflow 使用的环境变量

易用-jobflow

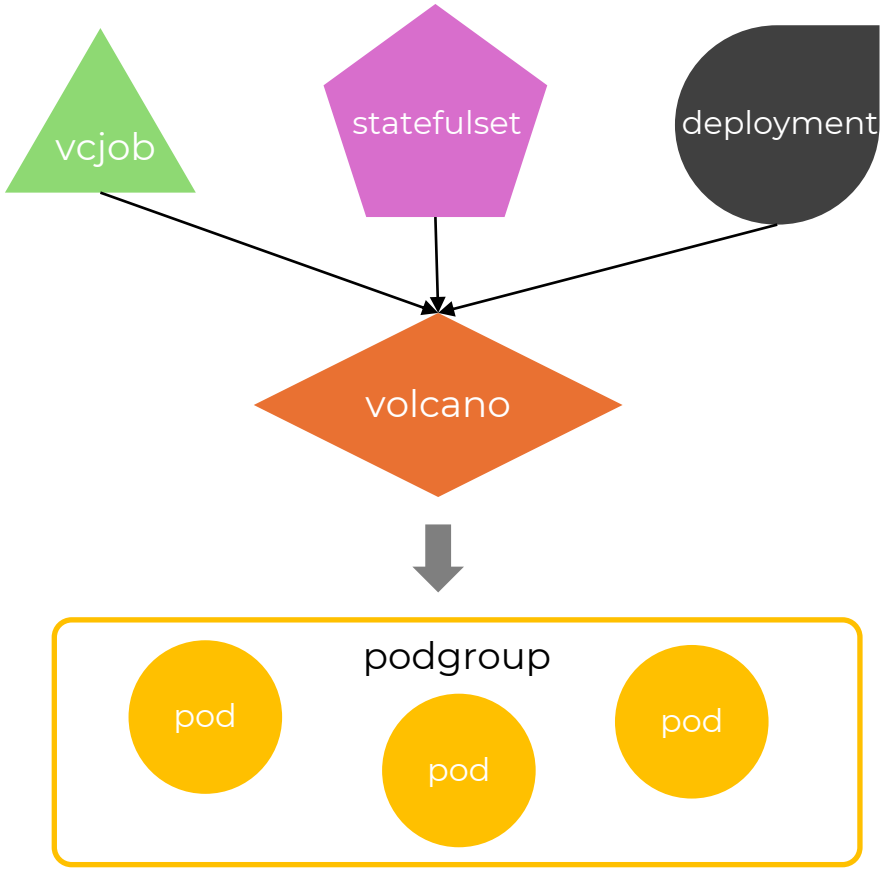
- 支持 vcjob 作业流
- 一种轻量级的 argo workflow 替代
- 支持多种运行条件(ongoing)



易用-queue & podgroup

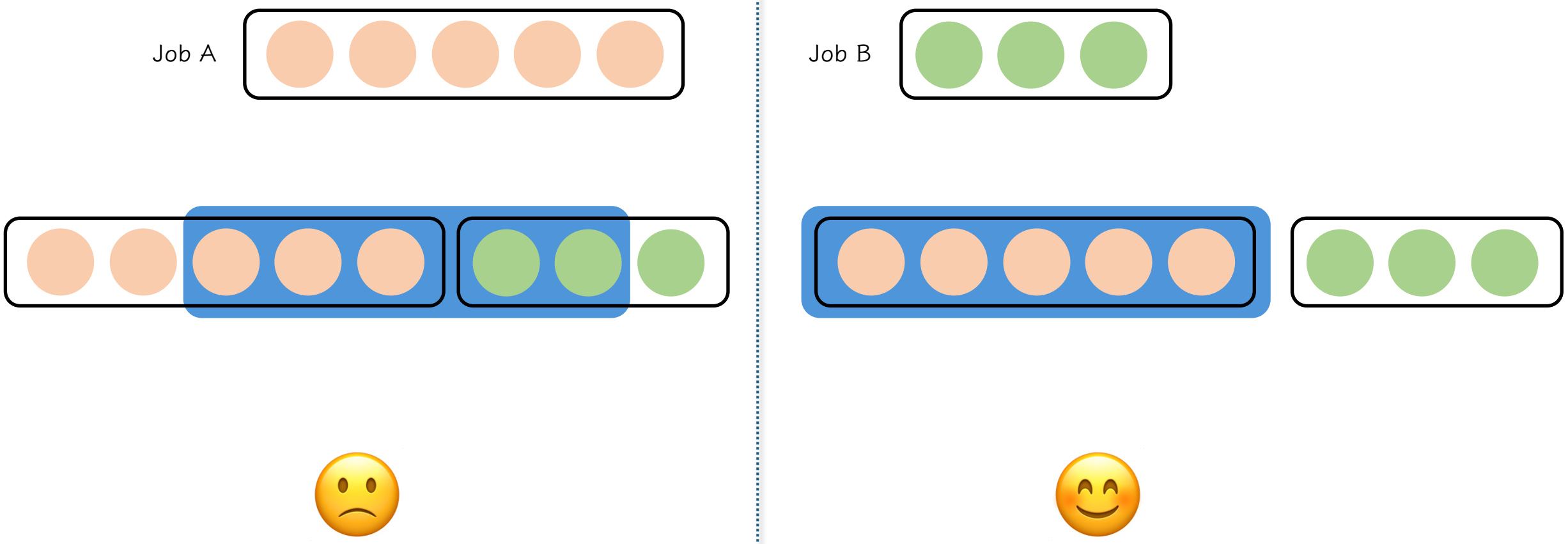


- 多队列支持
- 队列内作业抢占
- 队列间队列抢占
- 队列配额
- 按权重划分资源
- 层级队列(ongoing)



支持将其他 Kubernetes 工作负载转换为 podgroup，使用 volcano 调度

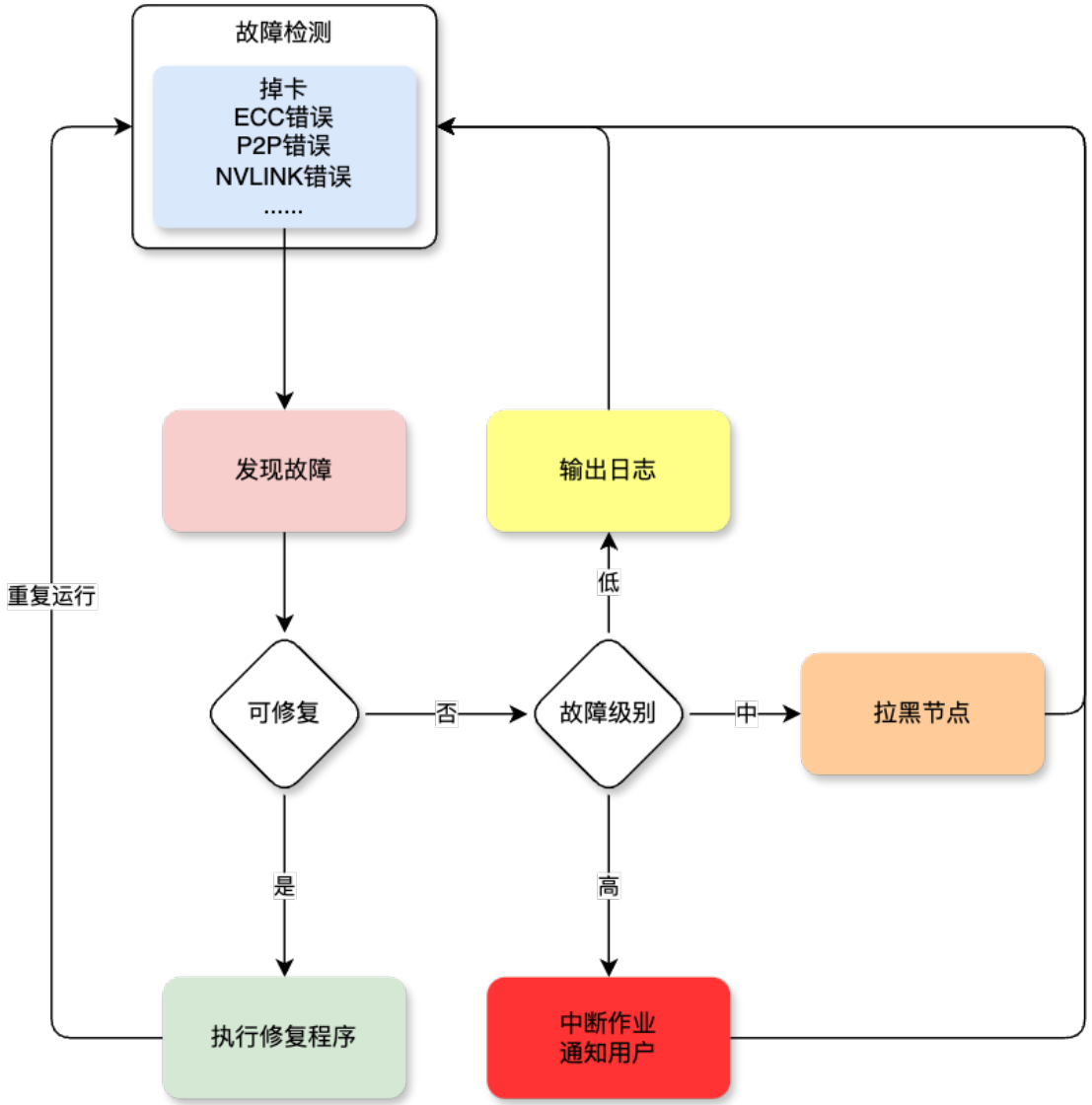
易用-调度策略-gang



故障处理-检测&修复

软硬件故障不可避免

- 自动检测
- 尝试修复

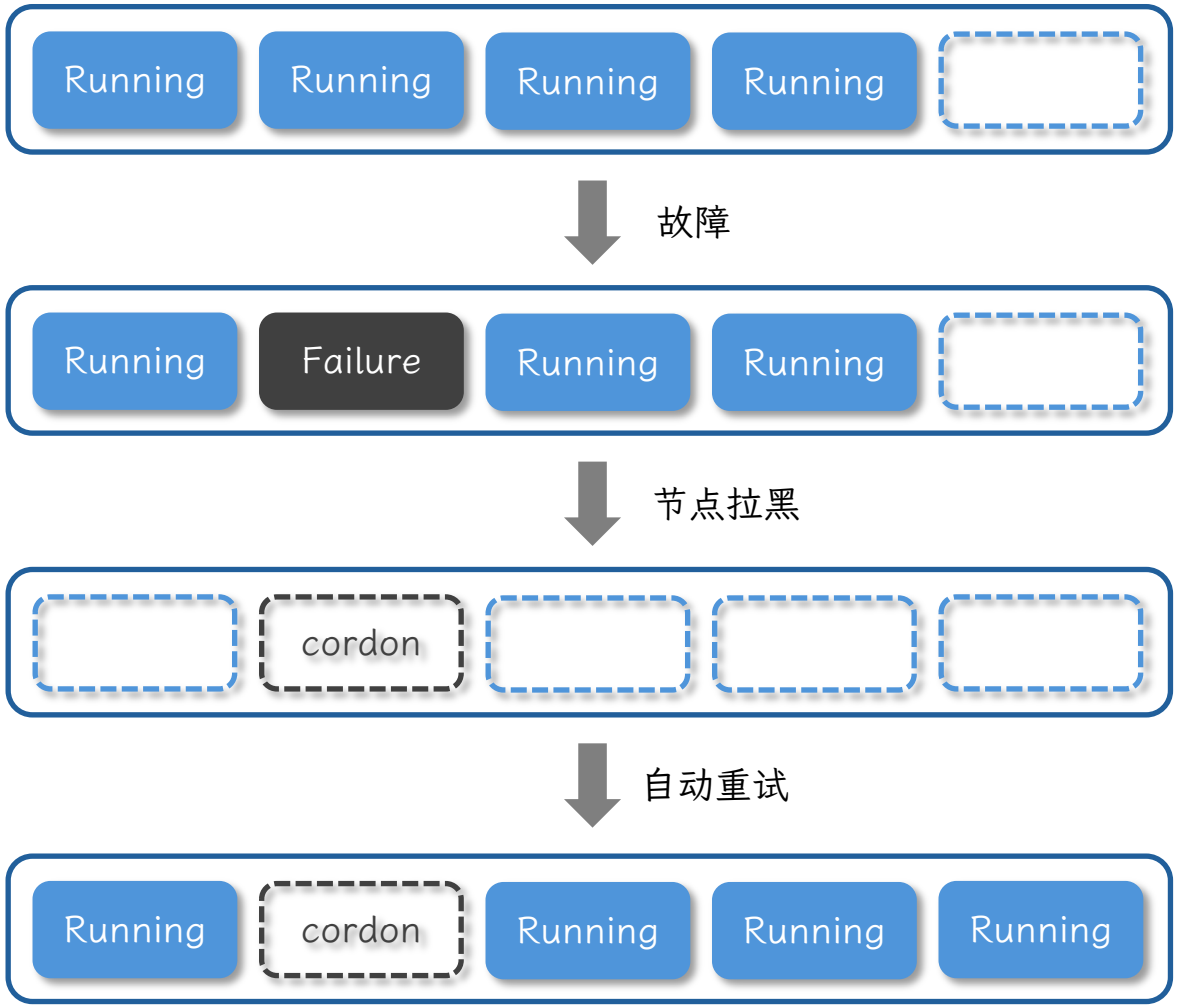


故障处理-作业重试

作业自动重试

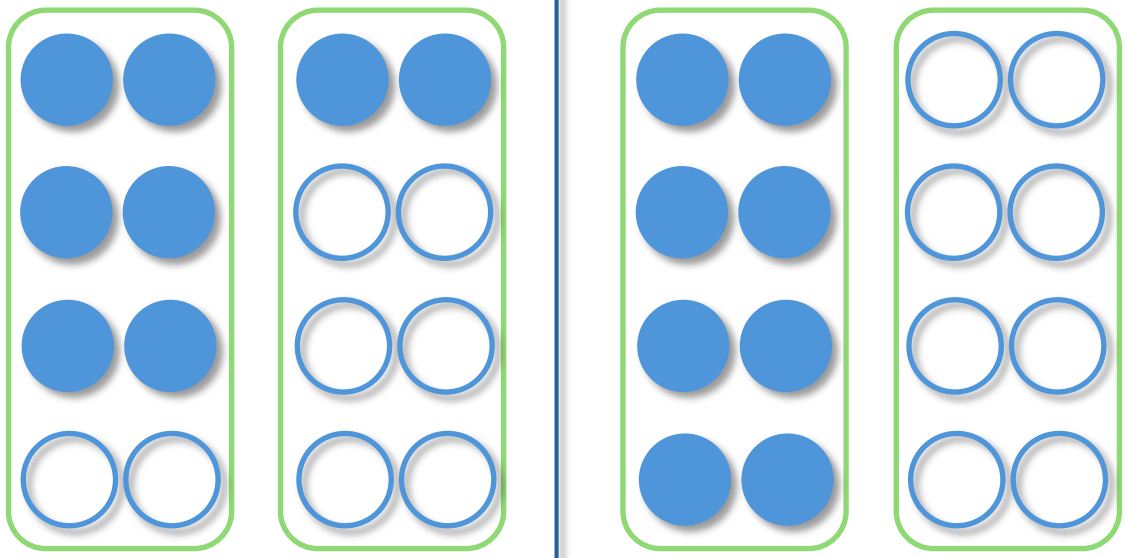
```yaml
apiVersion: batch.volcano.sh/v1alpha1
kind: Job
metadata:
  name: vcjob-demo
spec:
  maxRetry: 3
  minAvailable: 128
  queue: default
  schedulerName: volcano
  tasks:
  - maxRetry: 3
    minAvailable: 1
    name: master
    ...
```

```yaml
policies:
- action: CompleteJob
  event: TaskCompleted
- action: TerminateJob
  event: PodFailed
- action: TerminateJob
  event: PodEvicted
```

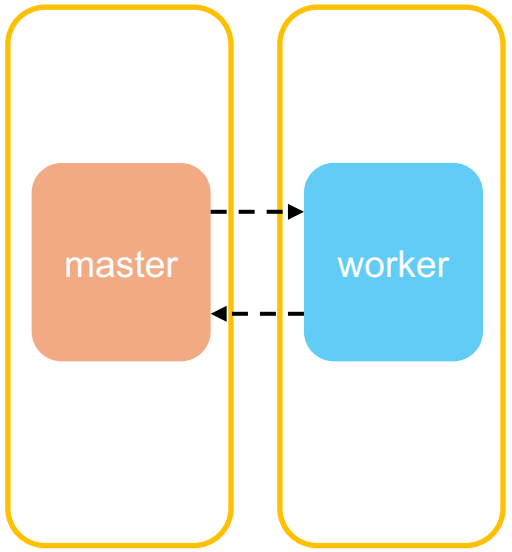


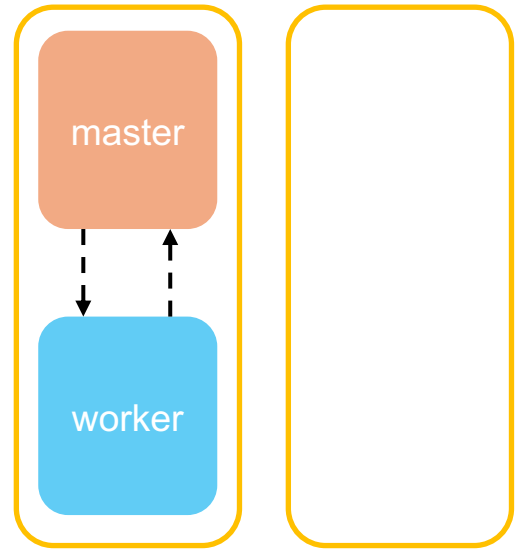
效率提升-binpack&task-topology

binpack



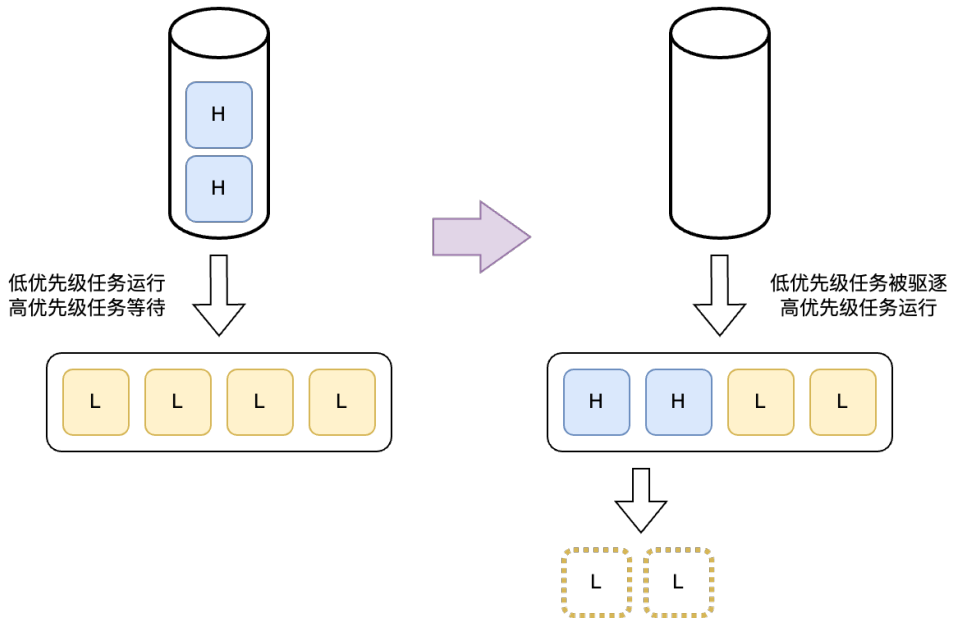
Task-topology



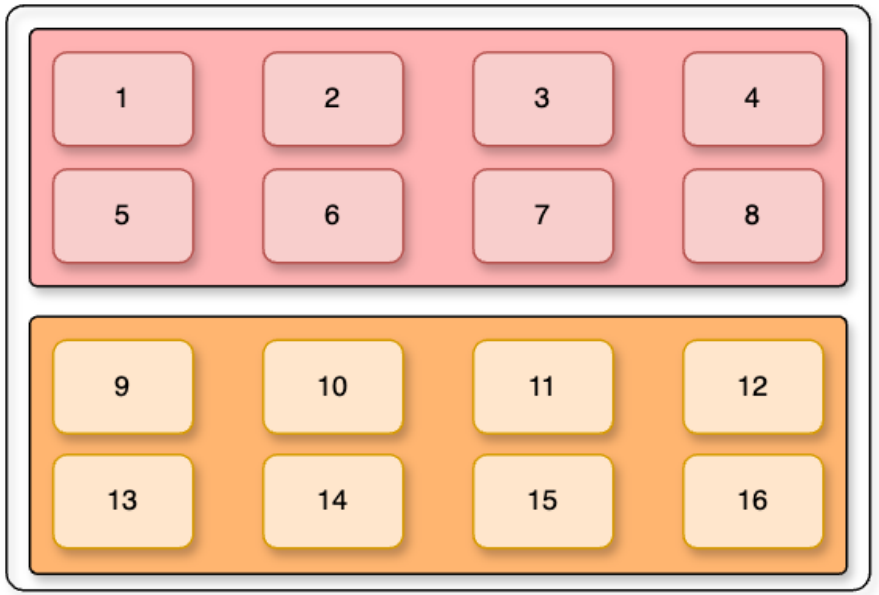


效率提升-抢占&昇腾优化

抢占



昇腾优化

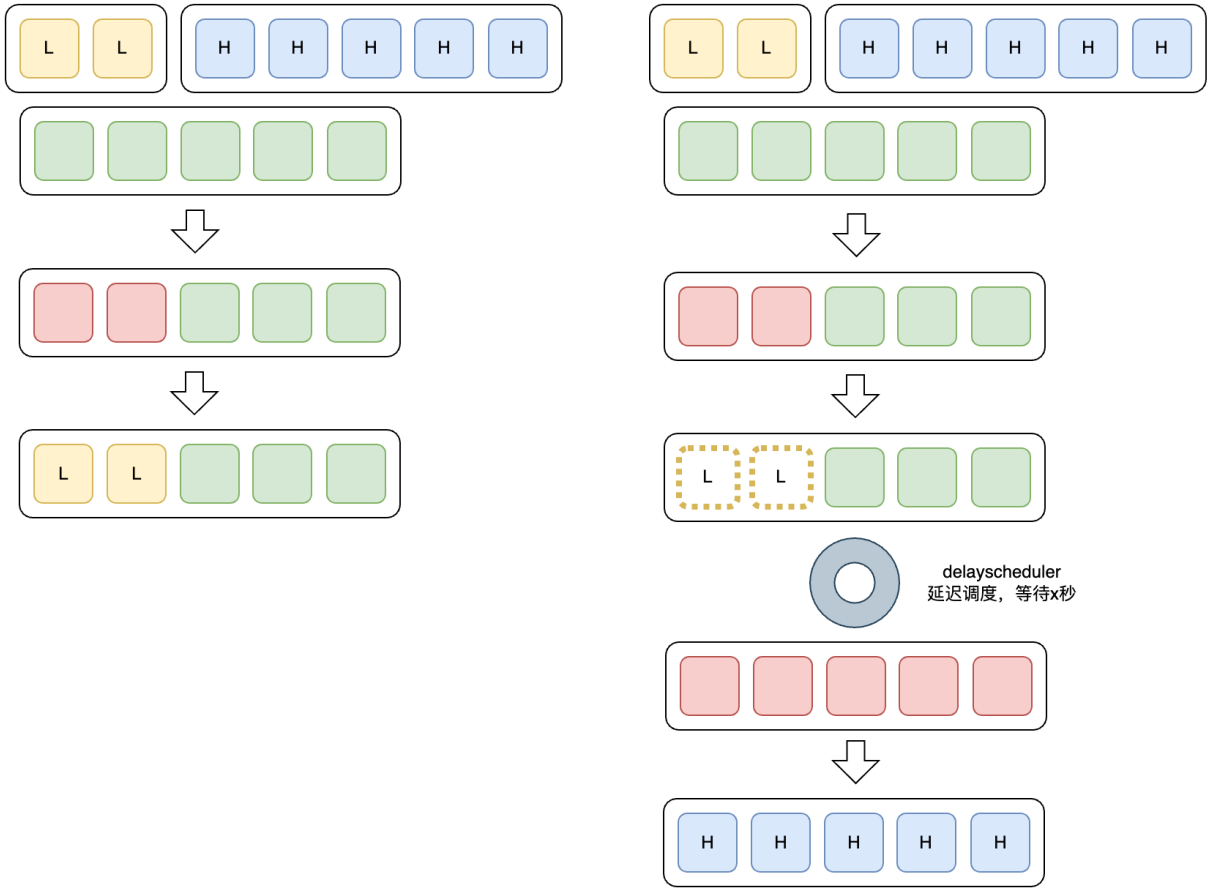


- 单节点 16 张卡
- 每八张为一个环
- 作业需要在同一环内运行

效率提升-启动&清理

延迟调度

- Kubernetes 在释放 Pod 时，每个节点释放时间不同。小任务总是先于大任务运行
- 低优先级任务等待 x 秒后再调度，给节点清理的时间




KubeCon
CloudNativeCon
THE LINUX FOUNDATION
OPEN SOURCE SUMMIT
AI_dev
Open Source GenAI & ML Summit
China 2024
谢谢！