
KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023


KubeCon | CloudNativeCon


OPEN SOURCE SUMMIT

China 2023

基于Kubernetes+RoCEv2构建大规模AI训练集群实践

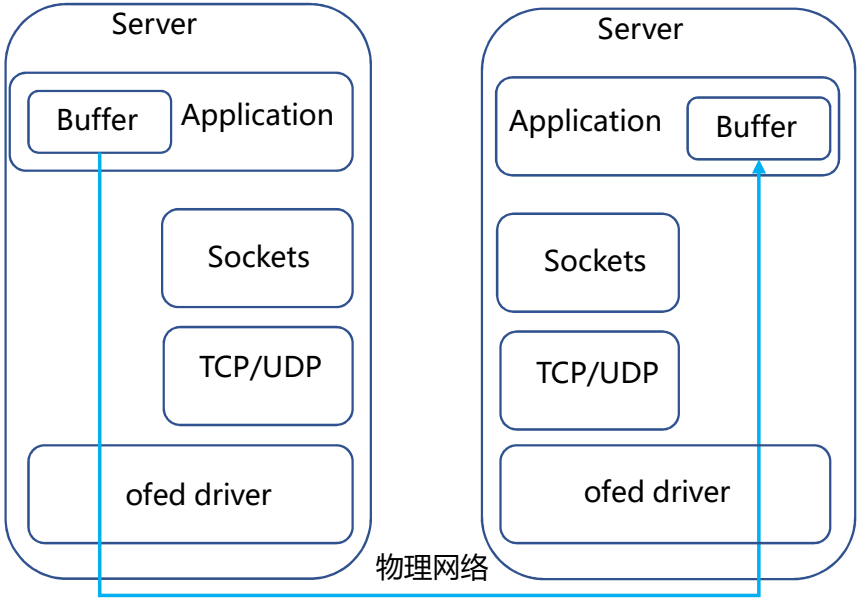
王德奎　浪潮信息

目录

大规模AI训练基础设施面临的网络问题

1. 大规模AI训练对算力显存需求巨大，单节点资源不足，依赖节点间高速互联
2. “多打一”问题导致网络拥塞，需要为分布式训练任务构建无损网络
3. IB网络与RoCE网络通信机制不同，AI基础设施适配问题
4. GPU服务器搭载多张GPU卡以及多张高性能网卡，带来的资源调度和通信问题

RDMA（Remote Direct Memory Access）



RDMA 网络数据流

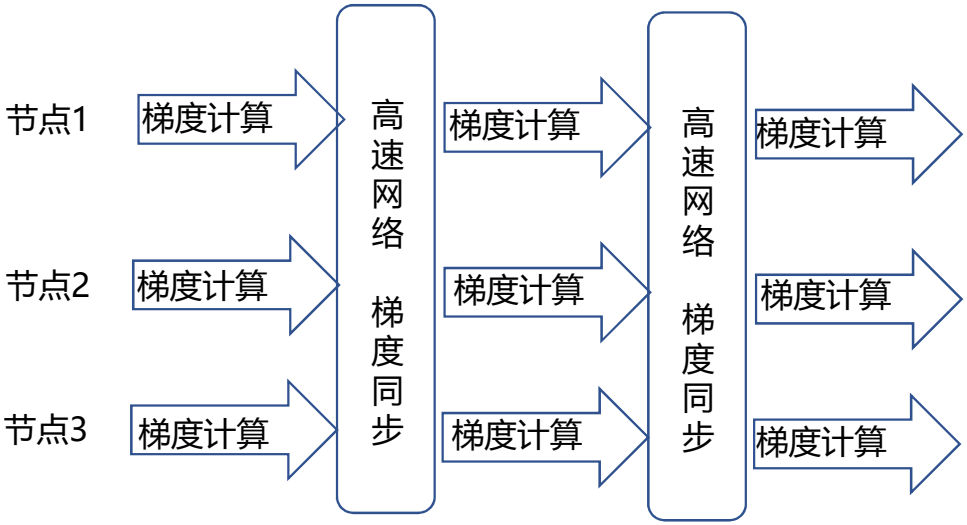
	Infiniband	RoCEv2
End-to-end delay	2us	5us
Flow Control Mechanism	Credit-based flow control mechanism	PFC/ECN，DCQN
Forwarding Mode	Forwarding based on Local ID	IP-based Forwarding
Load Balancing Mode	Packet-by-Packet Adaptive Routing	ECMP Routing
Recovery	Self-Healing Interconnect Enhancement for Intelligent Datacenters	Route Convergence
Network Configuration	Zero configuration through UFM	Manual Configuration

参考：https://www.naddod.com/blog/infiniband-vs-roce-v2-which-is-best-network-architecture-for-ai-computing-center

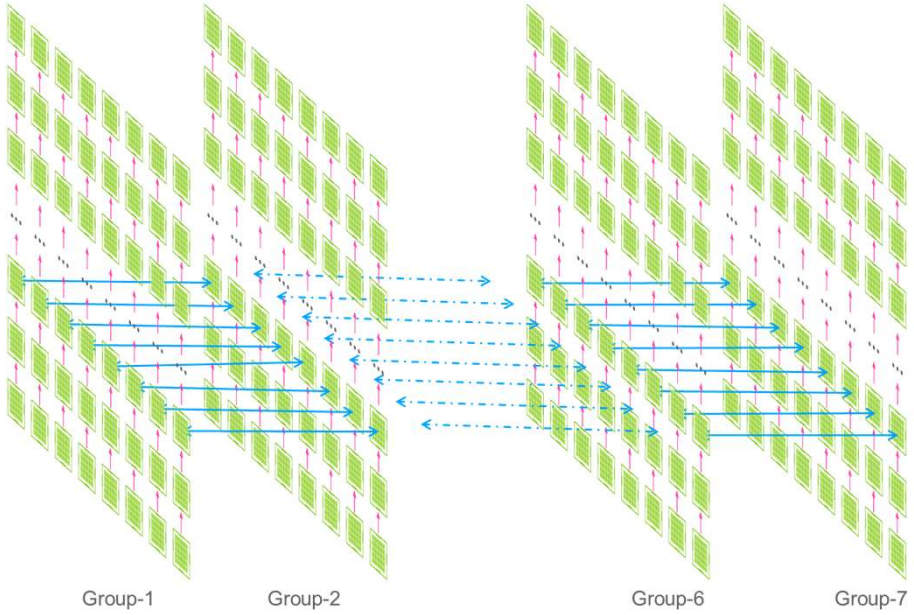
Infiniband VS RoCEv2

- Infiniband协议,一整套完整的链路层到传输层规范，无法复用已有以太网设备，需要购买全部Infiniband设备
- ROCE协议,基于以太网层的协议,可以复用已有的以太网设备,其中RoCEv2基于UDP实现
- iWARP协议, 相比于RoCE 和IB具有更好的可靠性，但会耗费很多内存资源，基于TCP实现

AI训练对高速网络需求



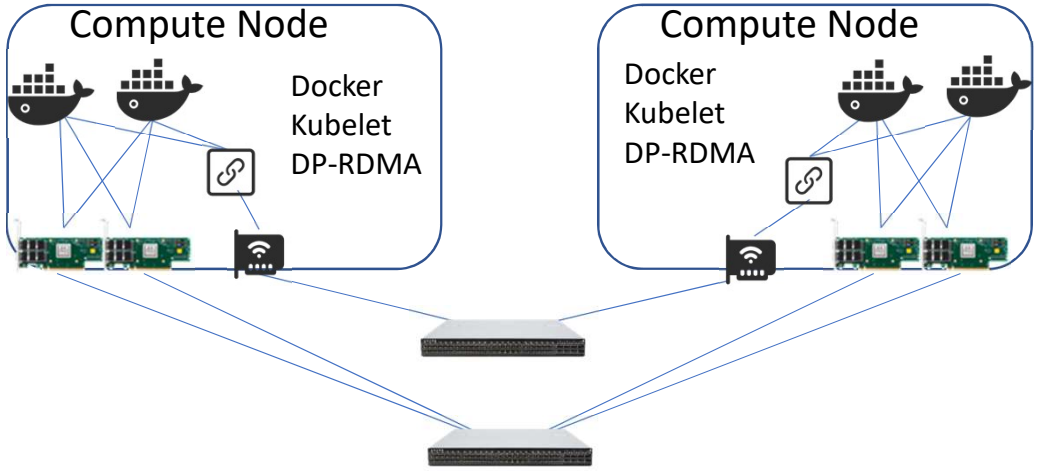
多机任务基于高速网络进行梯度同步



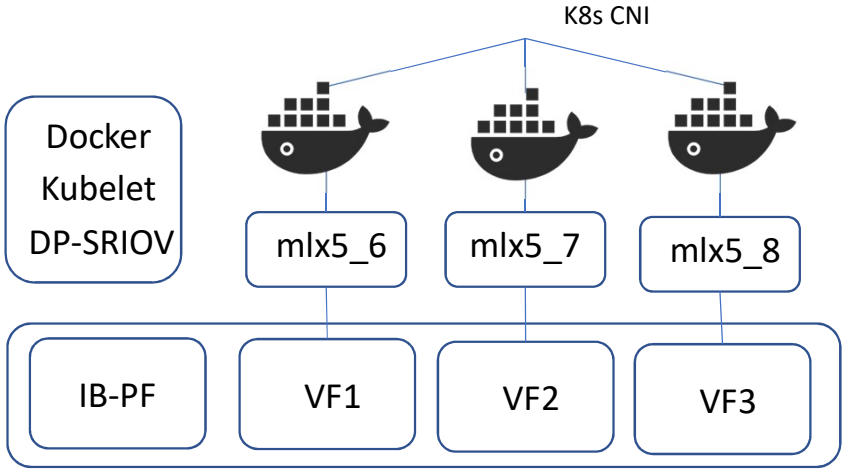
3D并行依赖高速网络进行梯度同步

- 数据并行：节点完成本地梯度计算后，节点间进行梯度同步，对带宽需求大，需要将模型梯度同步到任务中的每个GPU
- 大模型数据并行、流水线并行，同样依赖高速网络，对机间通信的需求更大
- 128台A100服务器共计1024个A100卡，3D并行运行GPT3模型，节点间流水线并行带宽需求12GB/s,每次通信量0.1GB，每次通信约0.16s，节点间数据并行带宽需求27.4 GB/s，每次通信数据量44GB，每次通信约32s

基于IB网络构建AI训练集群



方案一 IB网卡透传Pod

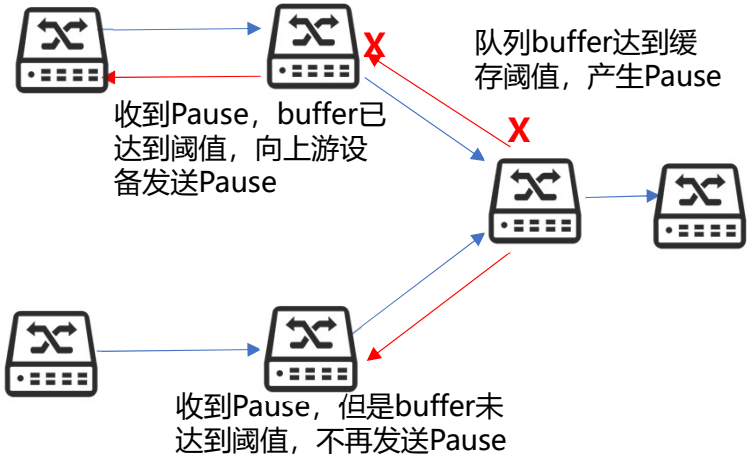


方案二 IB卡sriov

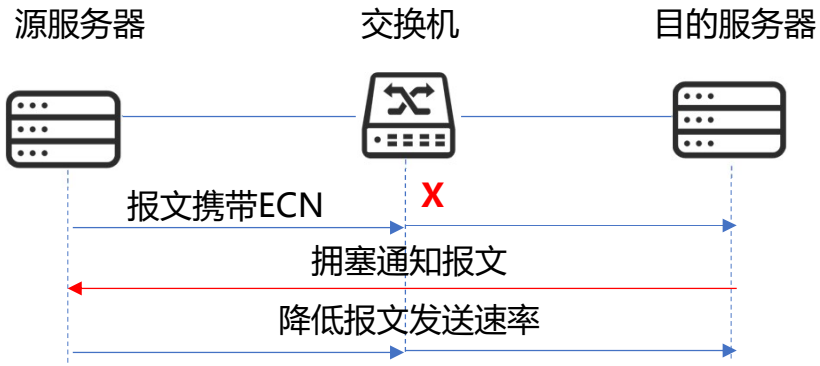
- Infiniband 基于OpenSM、 LID完成寻址与通信，不依赖传统的网络软件栈, 也可以购买UFM进行网络管理
- 基于Kubernetes+Infiniband网卡构建AI集群，只需要解决RDMA通信过程中的元数据交换问题
- 支持1W+节点集群

参考: https://github.com/Mellanox/k8s-rdma-shared-dev-plugin.git
https://docs.nvidia.com/networking/pages/releaseview.action?pageId=18481842

基于PFC+ECN构建无损以太网络



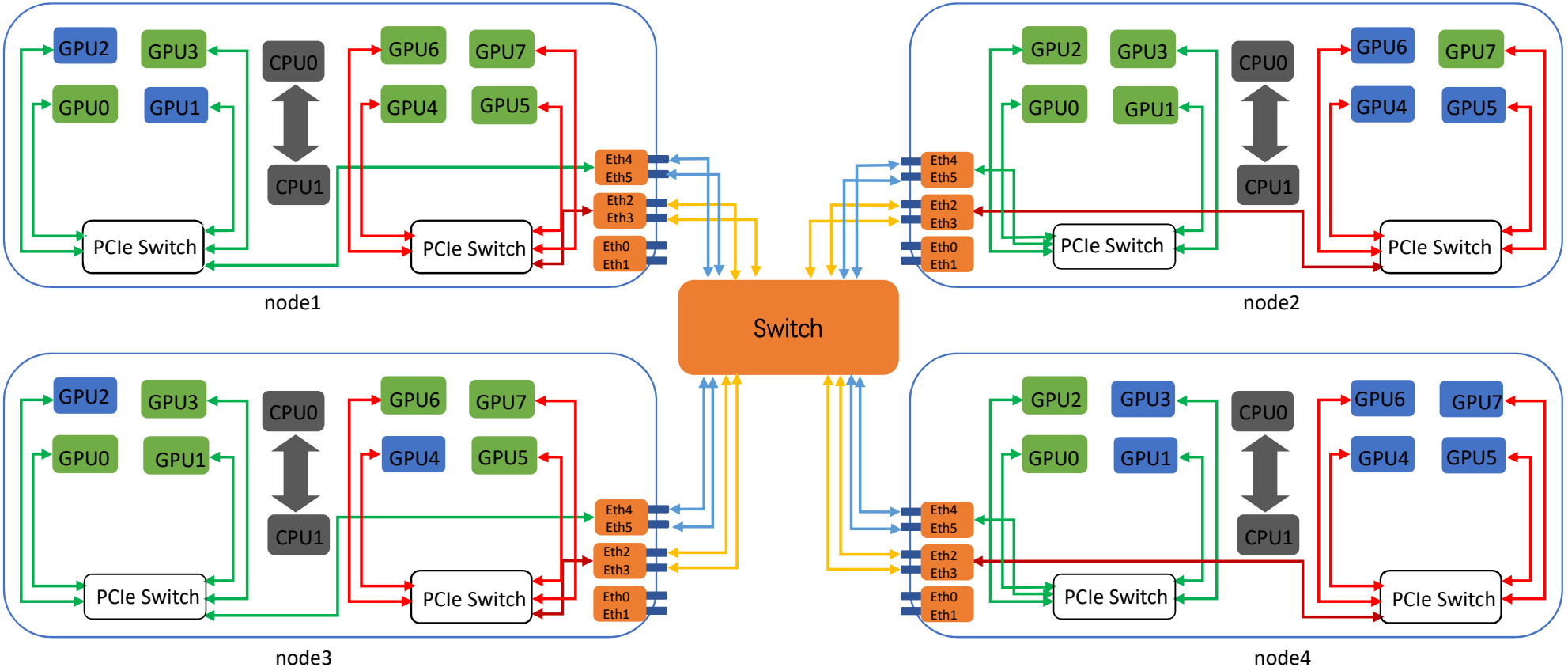
基于优先级的流量控制(Priority-based Flow Control)



显式拥塞通知(Explicit Congestion Notification)

- 交换机侧控制
  - PFC在数据链路层，基于报文-队列优先级，在交换机入口侧进行拥塞控制
  - ECN在网络层，基于数据包头中的标识位，在交换机出口侧进行拥塞控制
- 主机容器侧控制
  - K8s Pod基于Linux、OFED驱动进行拥塞控制

GPU 资源碎片化导致通信链路复杂化

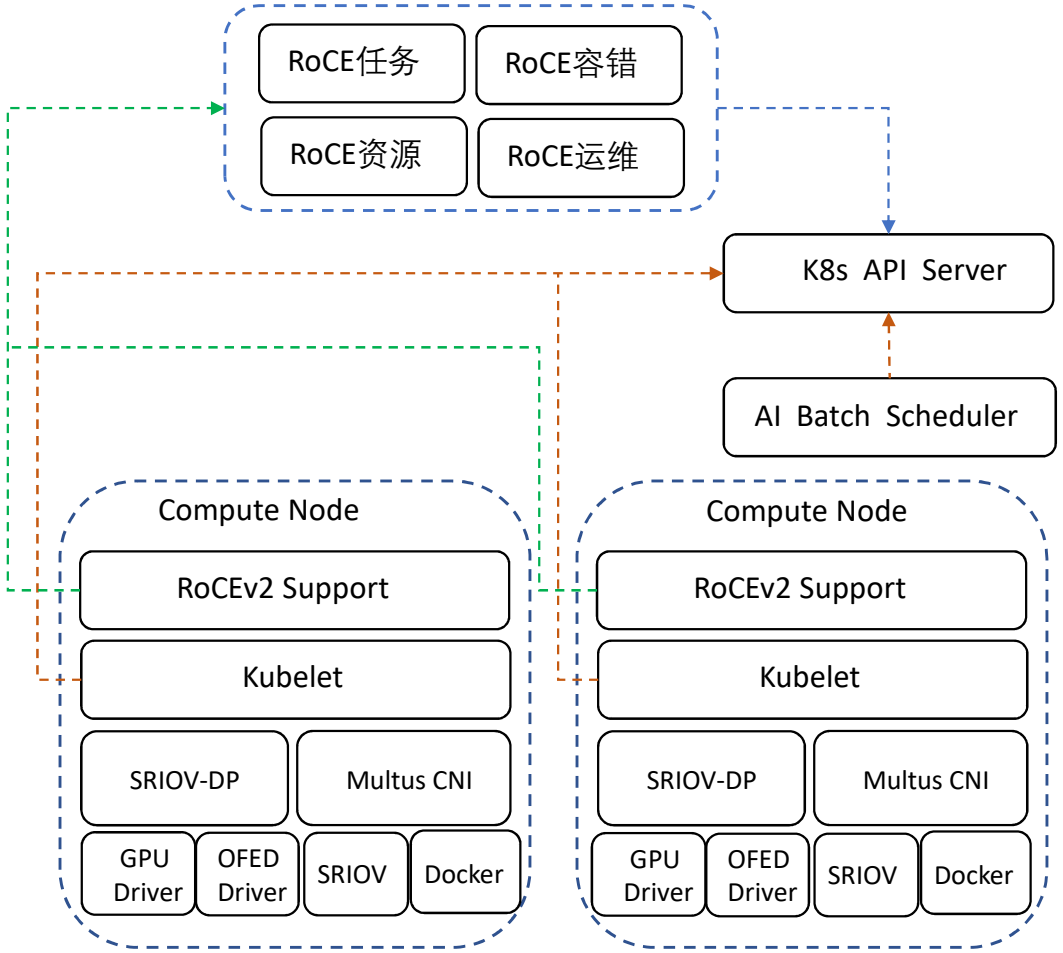


- 经过多轮次的GPU分配与回收，导致集群GPU分布碎片化，空闲GPU卡位置无序化
- GPU无序化影响多机训练任务使用的RoCE网卡，通信方式与物理机场景不同

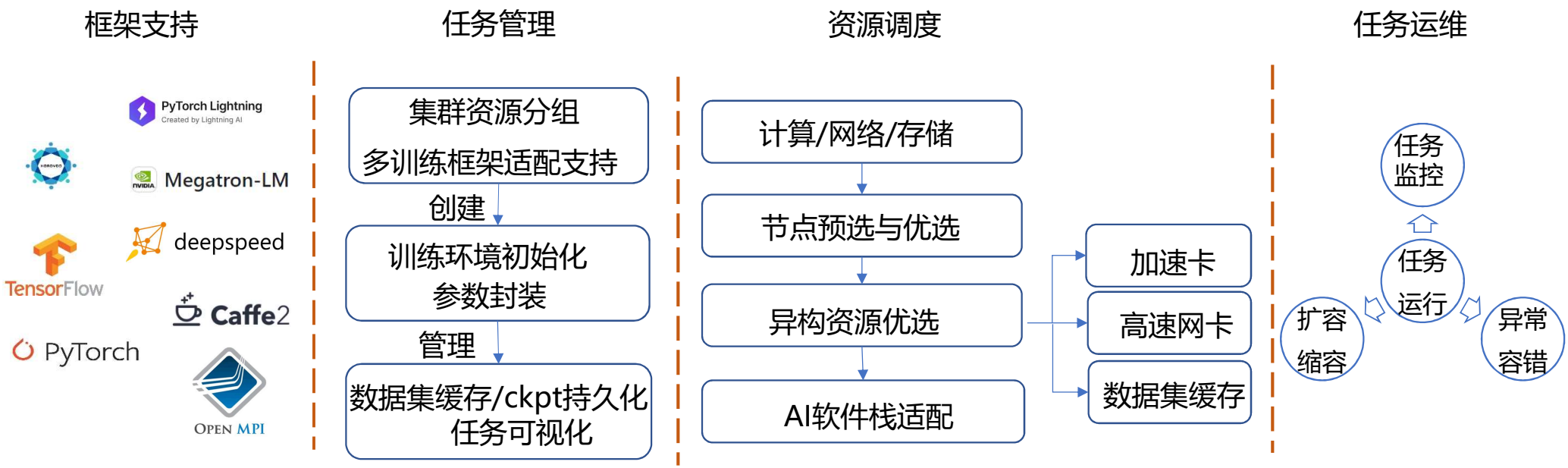
目录

RoCEv2方案软件架构

- 资源管理
  - RoCE资源管理分配、基于RoCE网络资源调度
  - 一类K8s资源表示多类VF网卡
  - PF/VF 网络流量监控，网卡异常告警, 任务容错
  - 不同网络类型节点资源调度
- 网络管理
  - 基于Calico构建业务网络，多VF作为计算网络
  - 跨子网通信管理、路由管理
- 开源组件适配
  - SRIOV-DP支持多张RoCE网卡VF管理分配、异常PF监测
  - Multus CNI支持多VF管理

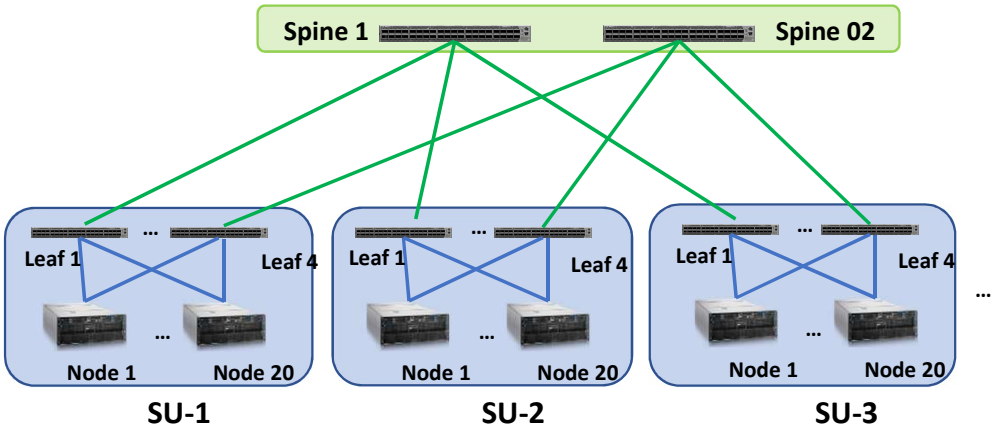


训练任务管理

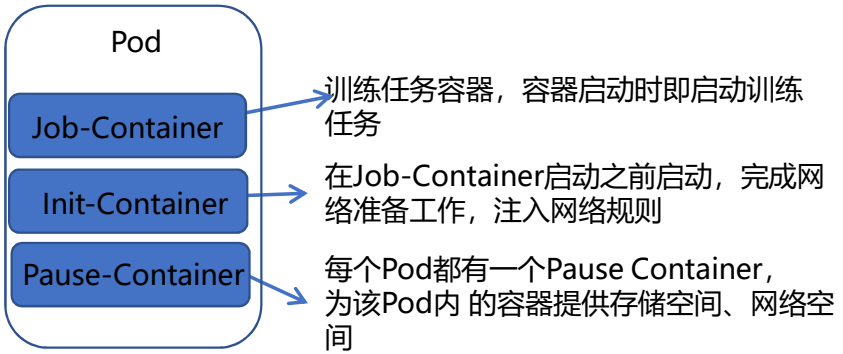


训练平台训练任务管理

RoCEv2方案物理网络架构



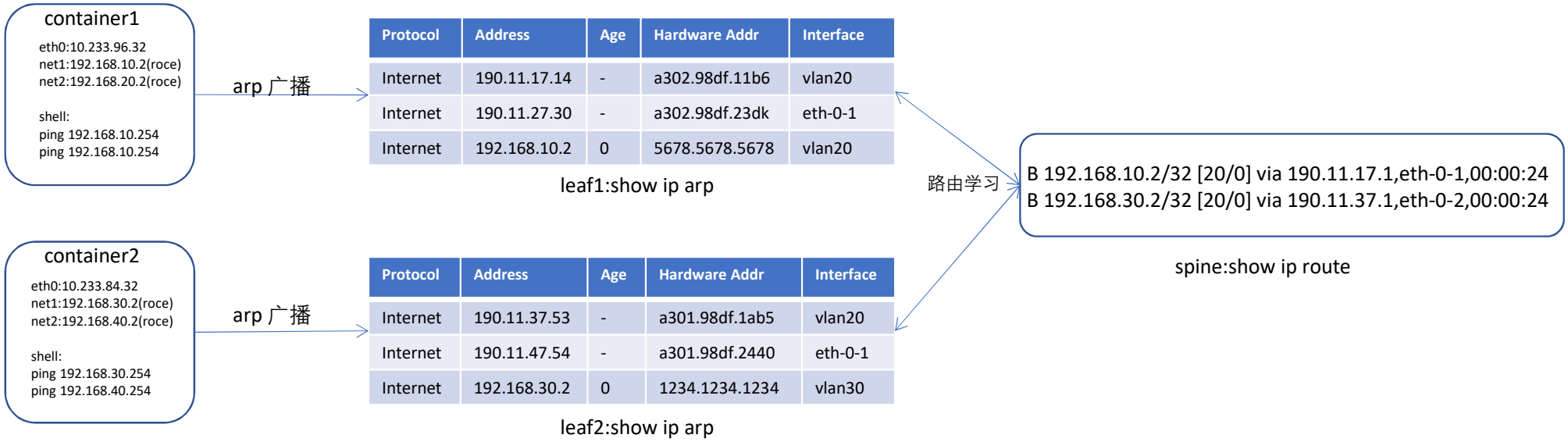
物理网络拓扑



网络路由自动化注入

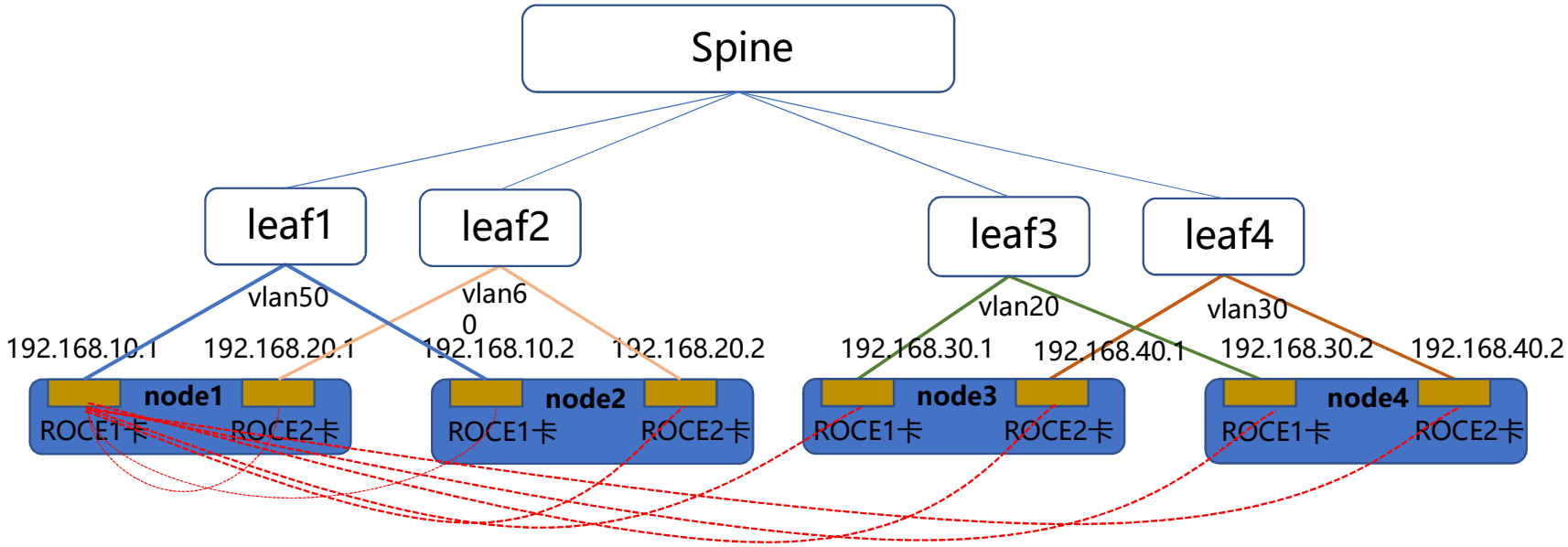
- Spine-Leaf组网，Spine交换机、Leaf交换机支持横向扩展
- 不同RoCE网卡划分不同VLAN
- 在物理交换机分配子网信息以及网关，容器中VF网卡使用物理子网和路由

容器内进行物理网络行为模拟



- 与虚拟机/物理机启动不同，容器秒级启动，导致容器内任务启动时，leaf/spine交换机未学习到路由表项,需要容器主动去上报arp信息
- 容器快速销毁，导致物理网络IP可能会被复用，但是交换机表项尚未刷新，配置IP递增使用
- 调整交换机arp表项老化时间，增加网络连通性监测
- 调整交换机arp表项容量配置

任意RoCE网卡之间点对点通信



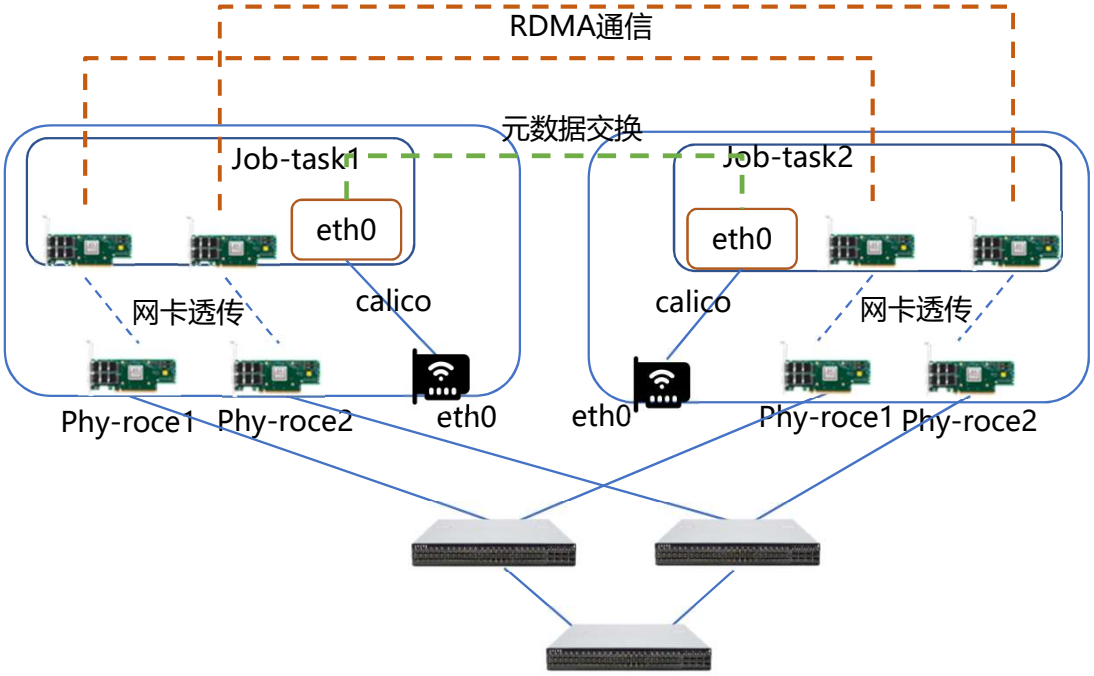
以node1节点的RoCE1卡为例，该RoCE网卡的VF可以与集群中任意节点的任意VF进行通信

其他注意事项

- GPU节点SRIOV虚拟化，导致GPU P2P异常，影响任务正常运行
- RoCE网卡SRIOV虚拟化后，多个VF RDMA流量共享问题
- 基于MacVlan网络方案，导致gid index递增问题
- 容器内自动识别到PF、全部VF，nccl无法正确选择VF
- RoCE网卡最大VF数量问题

RoCEv2大模型支持方案

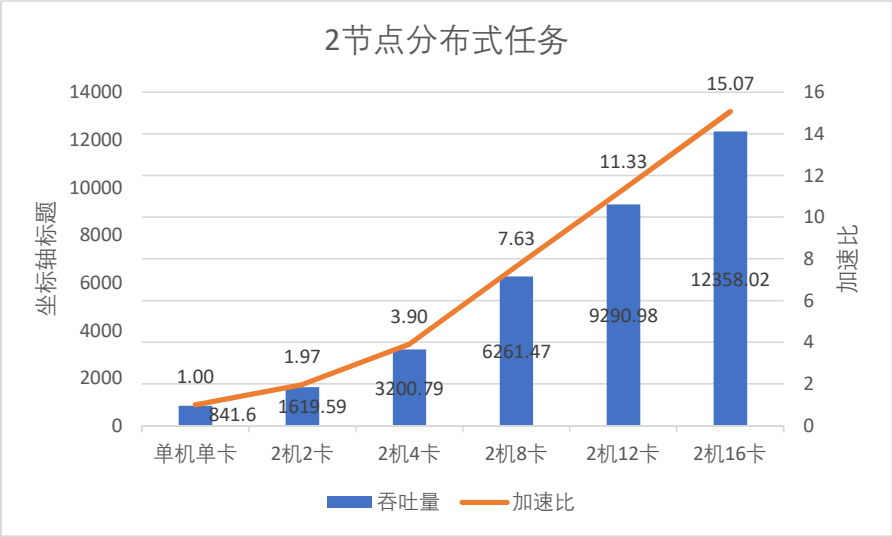
- 大模型训练场景，节点GPU资源被独占，不需要GPU资源灵活分配，GPU碎片化问题较少
- 基于Calico 构建元数据交换网络，基于物理RoCE网卡构建RDMA通讯网络
- multus cni,sriov-dp支持分配RoCE网卡PF
- 大模型训练任务充分利用nccl通信优化特性,例如PXN

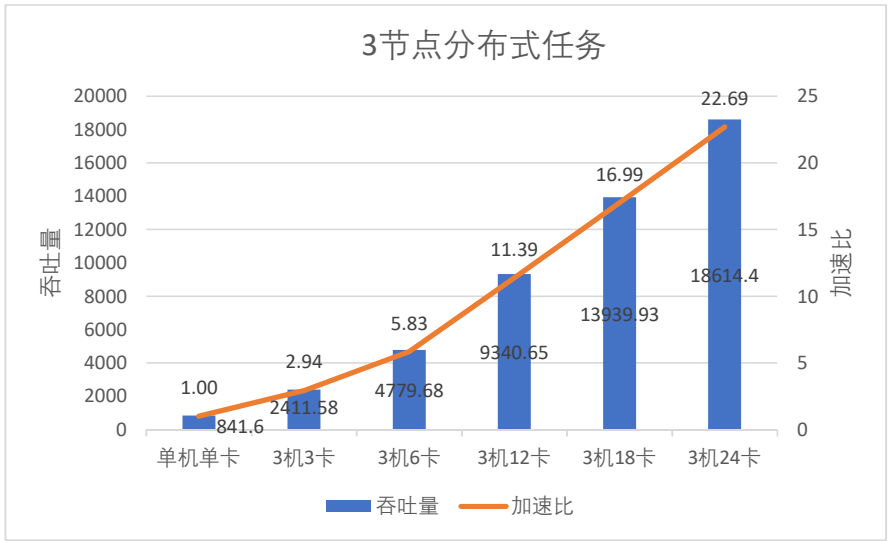


Pod加载Calico网与物理RoCE网卡方案

目录

多机多卡任务测试





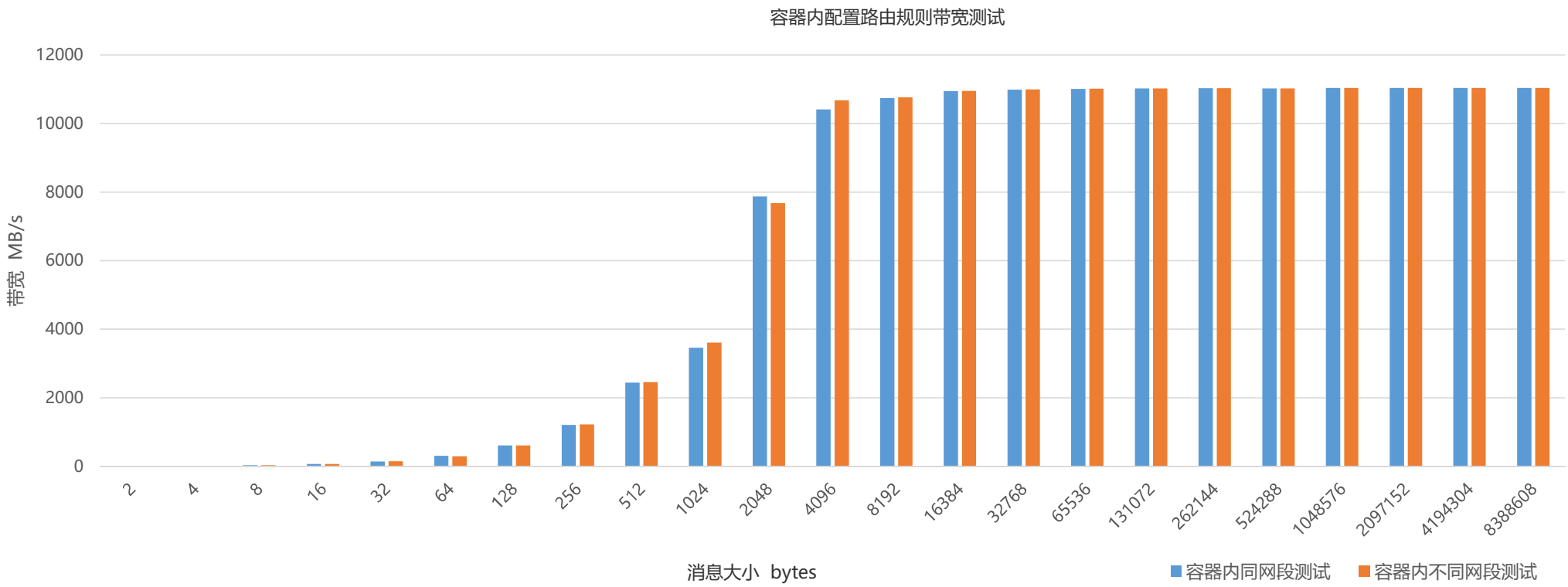
测试主机信息：

NF5468M5
CPU：Intel(R) Xeon(R) Gold 6230R CPU @ 2.10GHz
GPU：A100-PCIE-40GB
IB：Mellanox Technologies MT27800 Family 100Gb
GPU driver：450.102.04
IB Driver：5.4-1.0.3.0

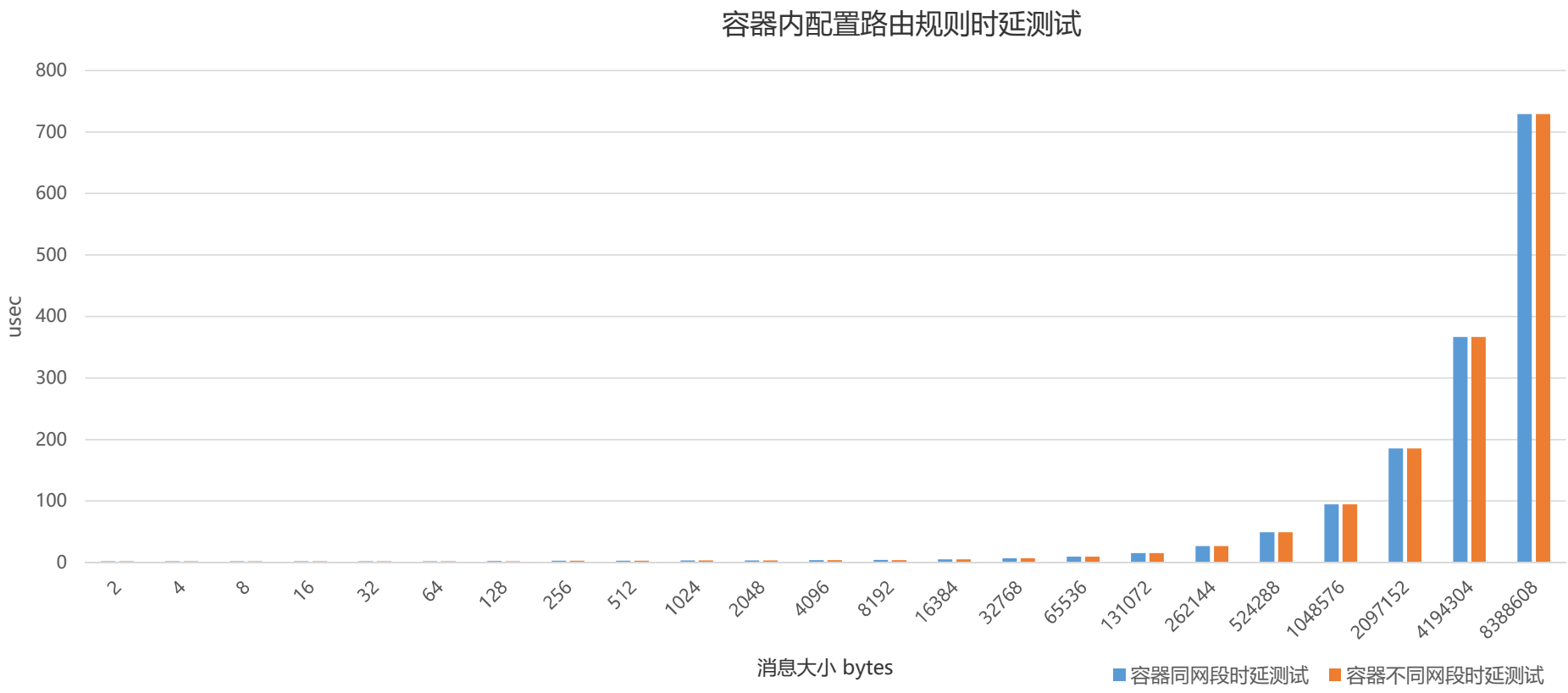
测试软件信息：

CUDA：11.0
NCCL：2.12.6
Tensorflow：1.15.3+nv
Tensorflow-cnn-benchmark,imagenet(synthetic),resnet50,bs=256,iter=500

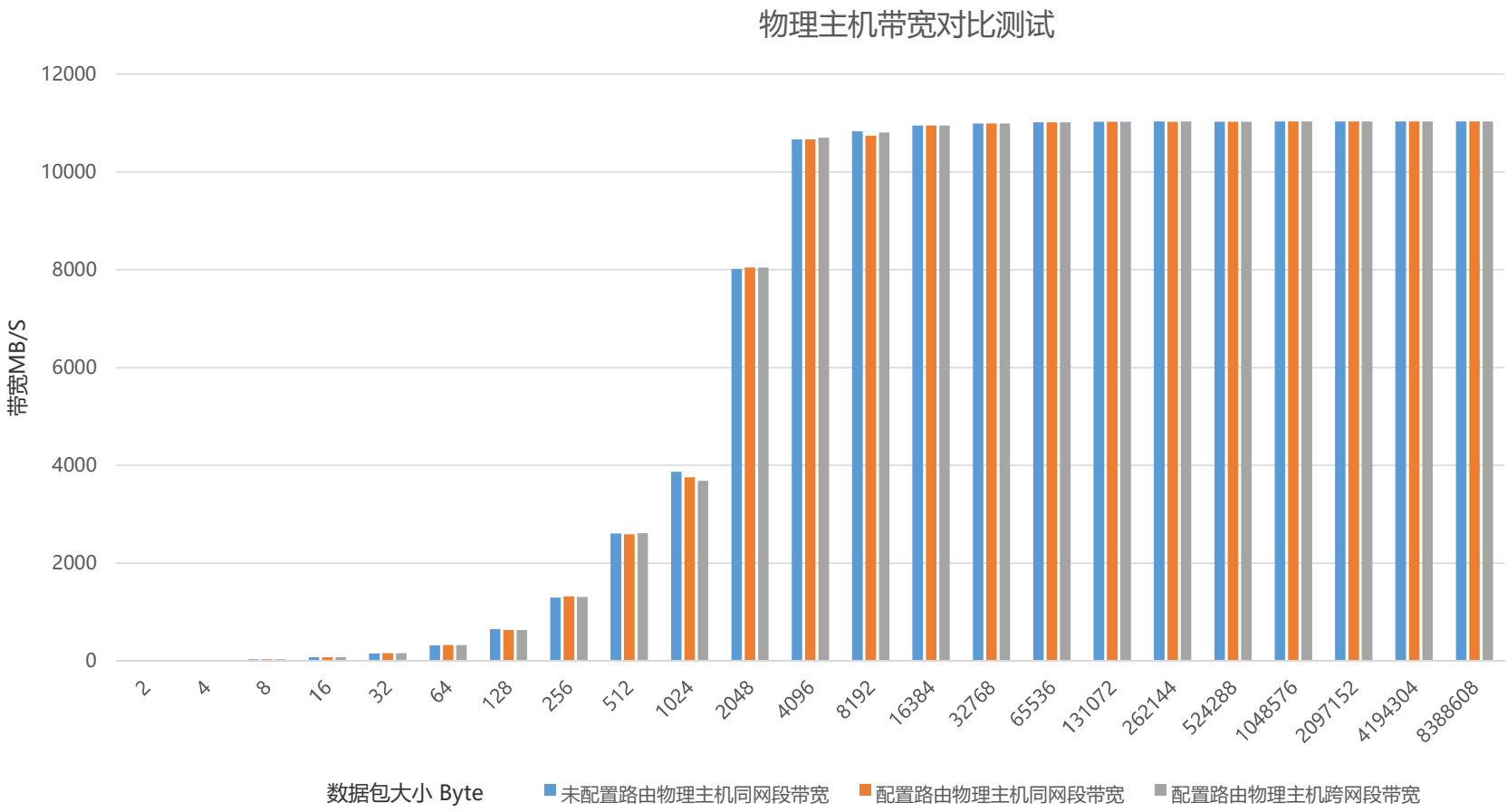
容器内RDMA打流带宽测试



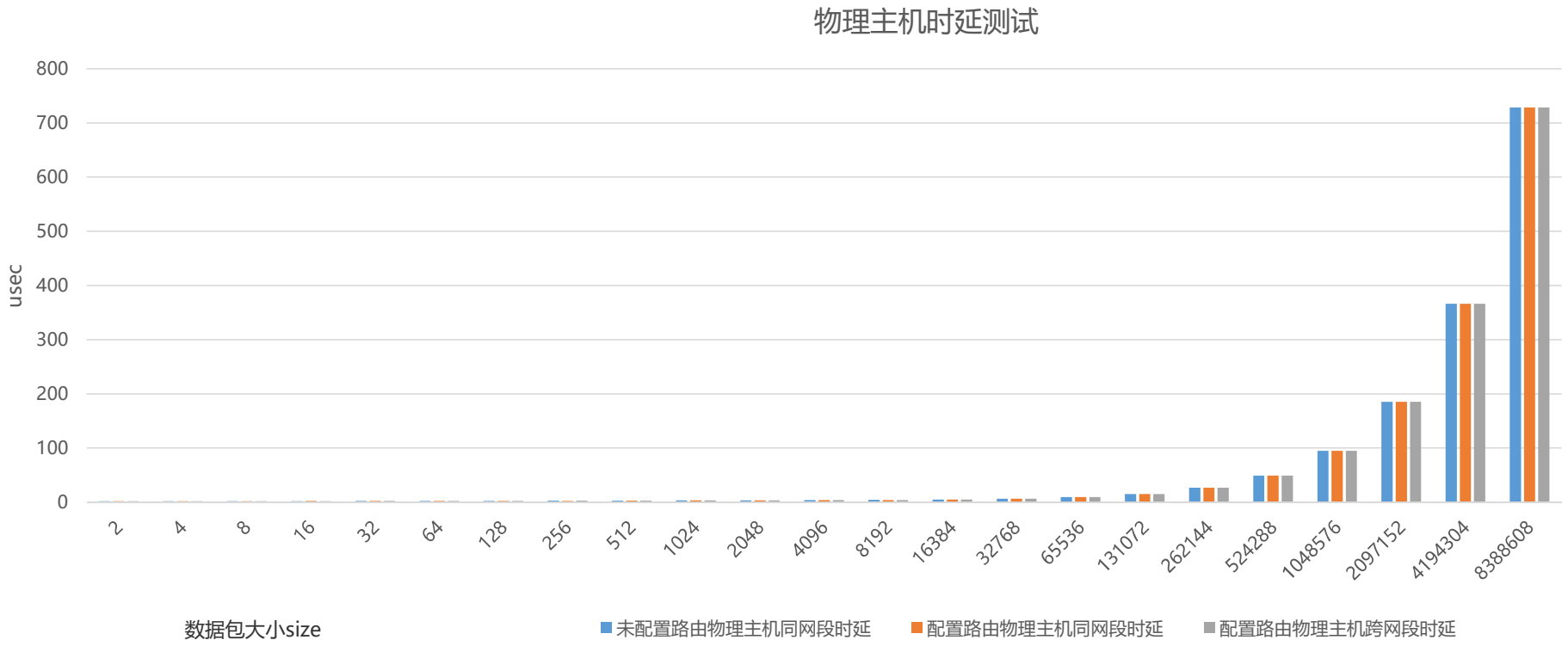
容器内RDMA打流时延测试



物理机RDMA打流带宽测试



物理机RDMA打流时延测试




KubeCon | CloudNativeCon

OPEN SOURCE SUMMIT

China 2023

AIStation 人工智能开发平台

王超　浪潮信息

大模型：生成式AI的核心技术

创新层出不穷

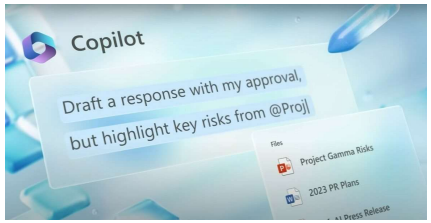
Open Source

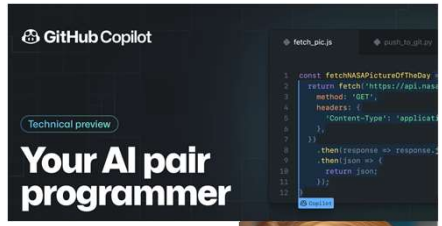
stability.ai, mosaicML, EleutherAI, neptune.ai, Meta, aws, 北京智源 BAAI, 百川智能 BAICHUAN AI, 智谱·AI, Hugging Face, cerebras, together.ai, databricks

Closed Source

OpenAI, AI21 studio, Google, co:here, Formic, ANTHROP\C, Google DeepMind, NVIDIA, Baidu 百度, NAVER, Mistral AI, STELLARIS AI, MINIMAX, YUAN

应用快速落地





Auto gpt

character.ai

Jasper

大模型：研发应用的需求与挑战

KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2023

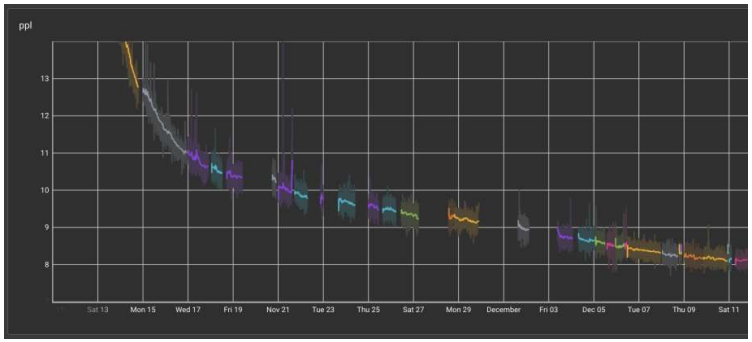
大模型实践

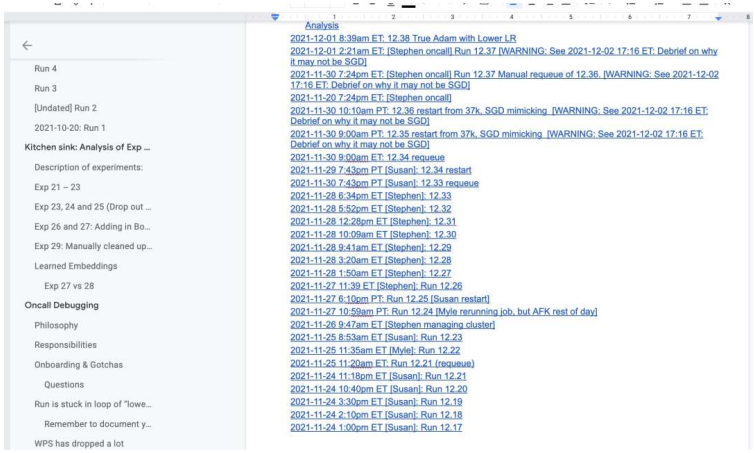
- 数千颗GPU芯片智算集群部署
- PB级数据爬取、筛选、分类……
- 千卡性能优化
- GPU失效、网卡失效管理
- 大规模计算输出不稳定、loss爆炸
- ……

服务客户实践

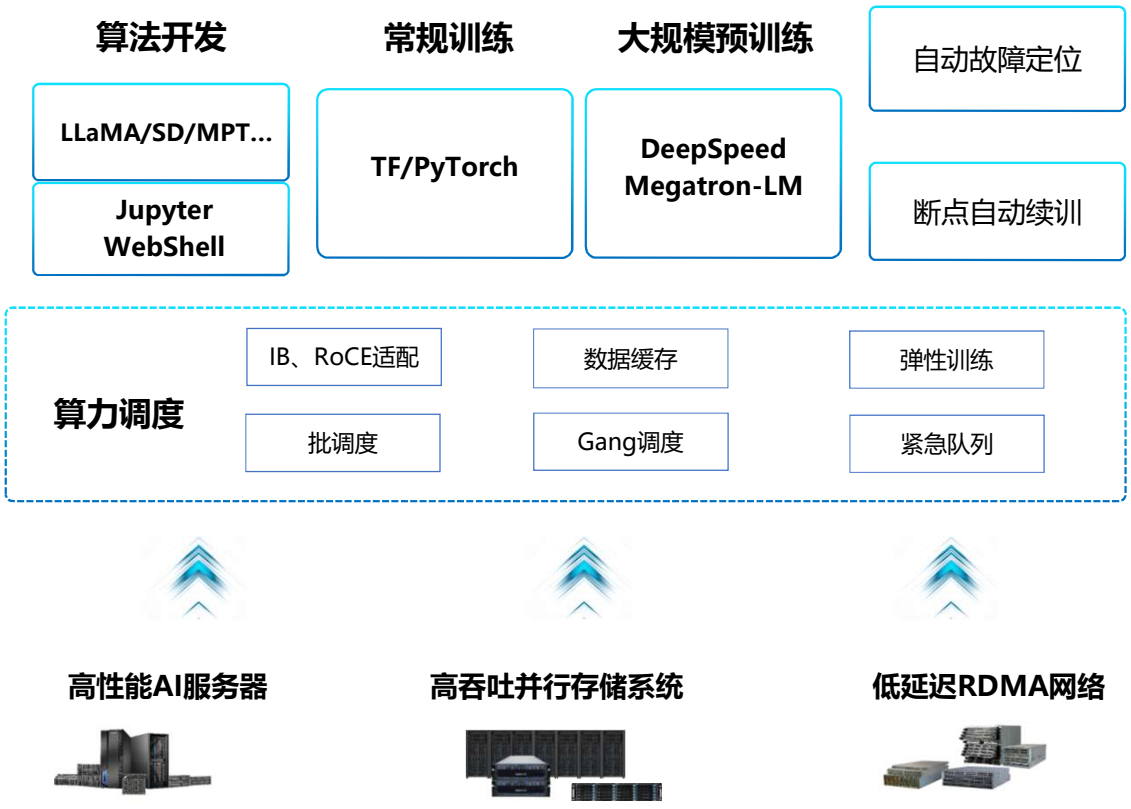
- CUDA初始化失败、GPU掉卡
- NCCL通信性能低
- GPU direct RDMA 未使能
- RoCE网络用不起来、不稳定
- 分布式任务的环境配置复杂、易出错
- 集群性能上不去
- ……

- Meta OPT-175B 模型，整个训练过程都要面对不停地重启和中断
- Meta训练日志显示两个星期的时间段内因为硬件、基础设施或实验稳定性问题而重新启动了40多次





面向大模型开发人工智能算力调度平台



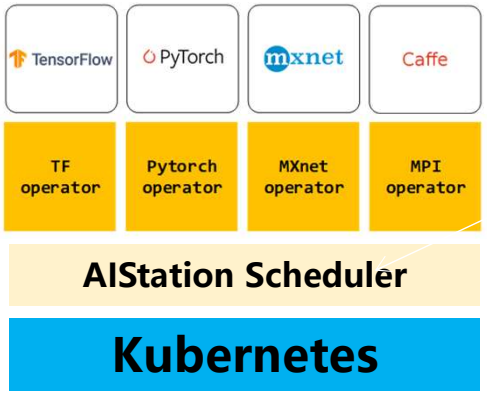
训练异常快速定位，断点自动续训

- 快速定位芯片、网卡、通讯设备异常或故障
- 全局训练暂停保持，热备算力自动弹性替换，健康节点快速CheckPoint读取，断点自动续训
- 训练全生命周期监管和异常全自动化处理，实现无人自动化

简化复杂网络适配，灵活高效使用

- 兼容IB、RoCE等复杂集群组网环境，解决开源调度版本对RoCE网络无损透传、灵活资源配比的使用难题
- 针对大规模训练场景，设备故障自动容错，保证大模型训练长时间高效、稳定运行

高效承载大规模模型训练



便捷的使用方式

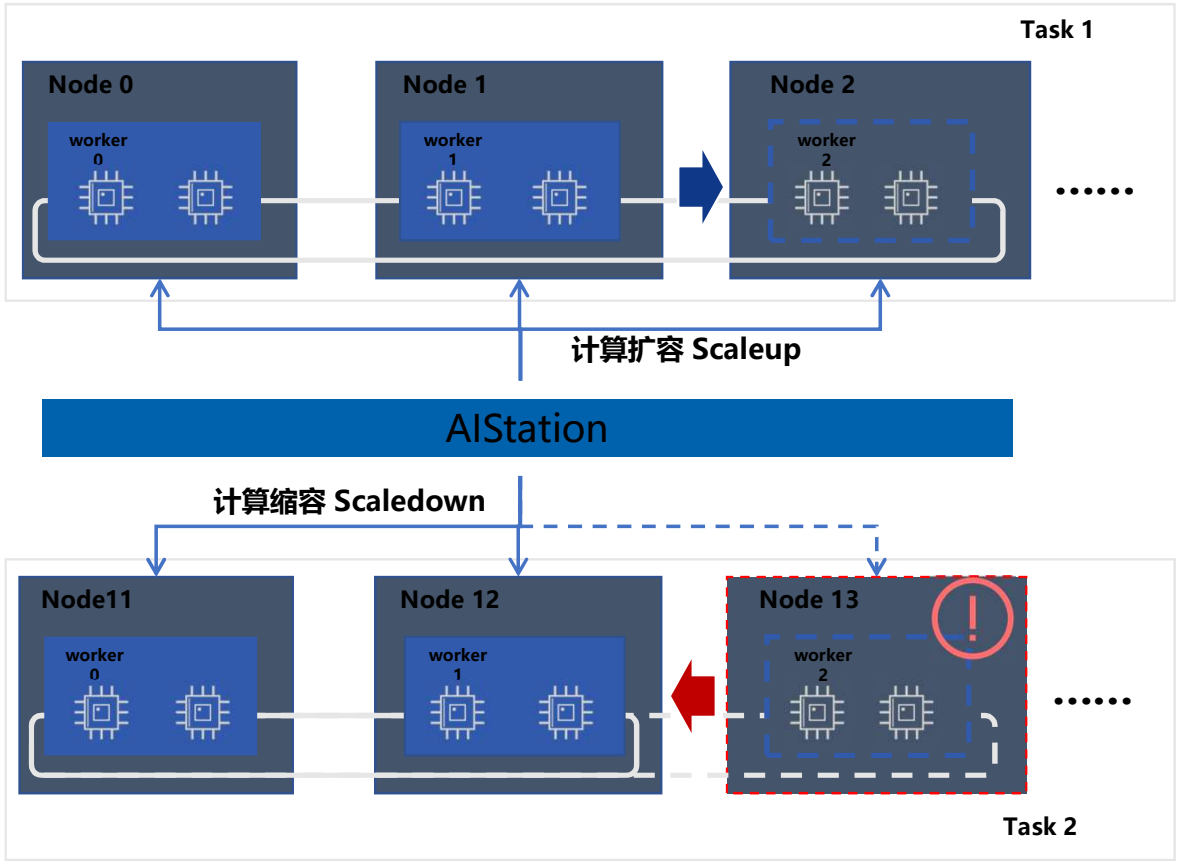
- ✓ 简单资源配置，免除分布式的网络、计算、连接方式的配置，一键launch分布式作业快速启动，针对大模型训练场景，能做到快速启动，支持Megatron-LM，DeepSpeed的训练模式等

Tensorflow	Pytorch	Mxnet	Paddle
ParameterServer	Distributed Data Parallel Training （DDP）	Data Parallel (Server-worker-scheduler)	ParameterServer
Mirrored			
MultiWorker Mirrored	Collective Communication		Collective
CentralStorage			
MPI	MPI	MPI	MPI

专业的底层优化

- ✓ 优化 Operator，支持tensorflow、pytorch、mxnet、caffe、paddle的原生分布式和全框架MPI分布式训练
- ✓ 优化 调度策略快速分配多机计算资源，自动启动分布式训练进程

计算资源弹性伸缩，资源动态使用



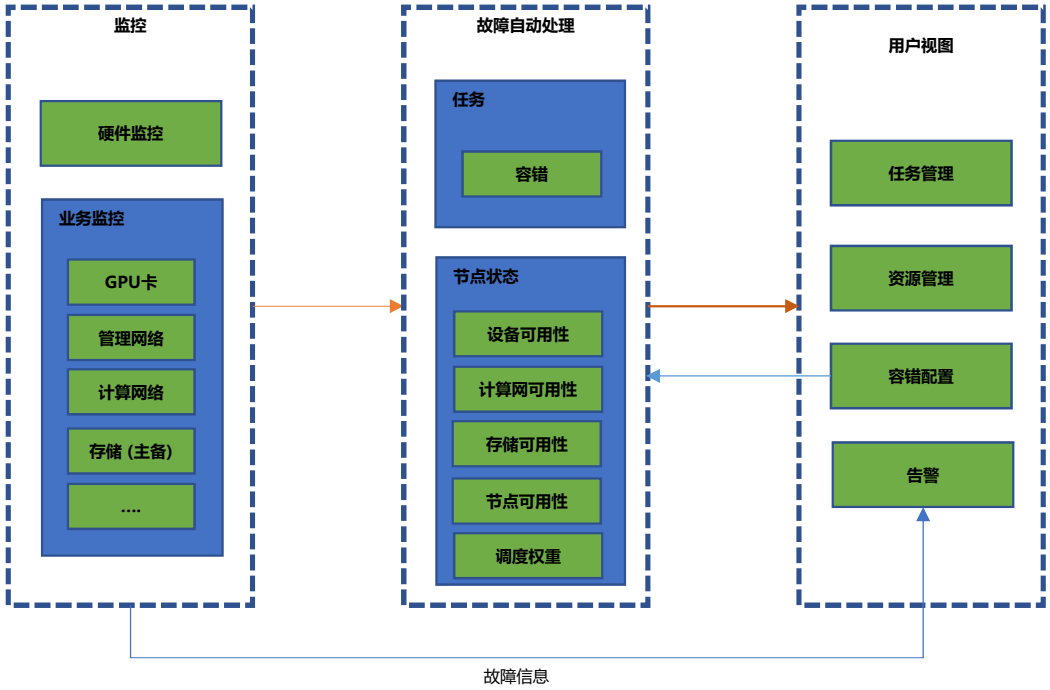
弹性训练，动态计算资源使用

- ✓ 计算资源弹性设置，训练任务动态使用底层计算资源
- ✓ 训练任务自动伸缩，按需动态使用计算资源
- ✓ 最大化利用计算资源，实现真正的资源利用率的提升

规模化训练部署的全方位保障

- ✓ 简化对资源数的评估策略，任务进行中可以根据运行情况，动态调整训练任务资源，保障巨量规模训练的时效性和可靠性
- ✓ 训练异常与资源受限时，无需算法人员干预，自动感知，训练任务自愈
- ✓ 弹性资源使用，让资源池利用更合理，如有限、限时资源使用时对如巨量模型训练对有计算资源要求的任务进行合理调度

分布式自适应系统：容错处理



监 控：能够监控业务故障类型；并且建立业务-硬件的映射关系

故 障 处 理：任务级别的容错；节点调度状态细分

用 户 视 图：报警信息；实现任务管理、和资源管理

AIStation平台业务容错能力

基本能力

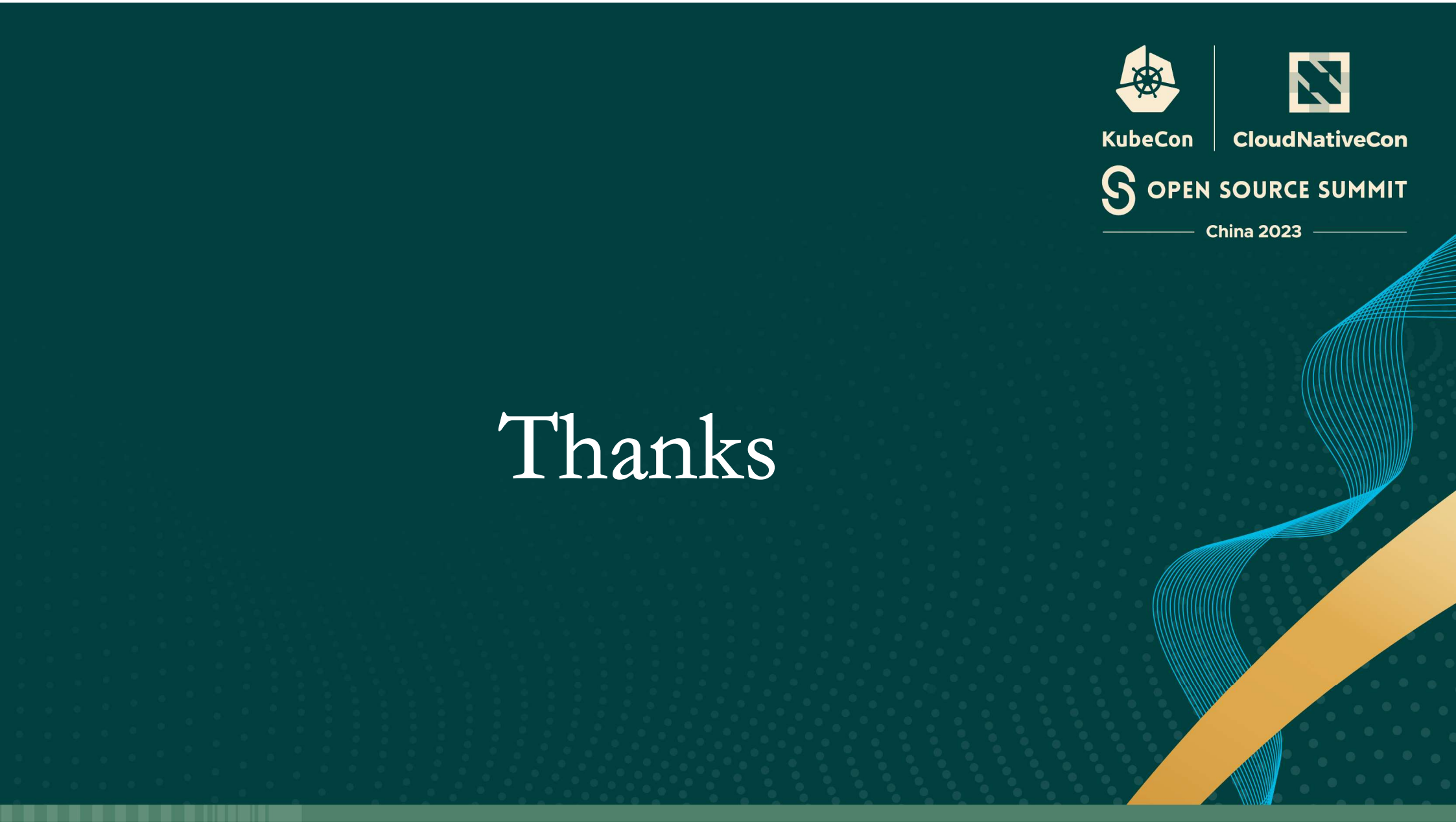
- 当训练任务异常终止时，例如worker退出、master退出
- 网络容错、GPU掉卡：掉卡节点置换处理，任务重启

对于弹性训练任务

- Master故障：重新提交训练任务
- Worker故障：框架自行处理
- 网络容错：非网卡故障，由框架处理
- GPU掉卡：掉卡节点置换处理，任务重启

弹性训练框架

- 副本数减少时，原任务自动识别异常节点，可以继续运行：当训练任务某个worker失联、宕机、异常退出、被删除后，训练任务可以较小的副本数继续运行
- 副本数增加时，原任务自动识别新增副本，继续运行训练任务


KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2023
Thanks