DaoCloud

教你成为算力网络背后的掌舵人

蓝维洲

2024.11.26

为 数 字 世 界 寻 找 全 局 最 优 解

Powered By DaoCloud

内容预览

- GPU 互联的两种网络技术：RDMA 和 NVLINK
- PTD-P 和 AI 网络流量特点
- 算力网络拓扑设计
- 提升 AI 网络性能
- Kubernetes 下的 CNI：提供 AI 负载的 RDMA 通信设备
- RoCE 网络可观测性

Part 01

GPU 互联的两种网络技术
RDMA 和 NVLINK

GPU 高速互联技术

NVIDIA

InfiniBand Trade Association

VS

RoCE

Ultra Ethernet Consortium

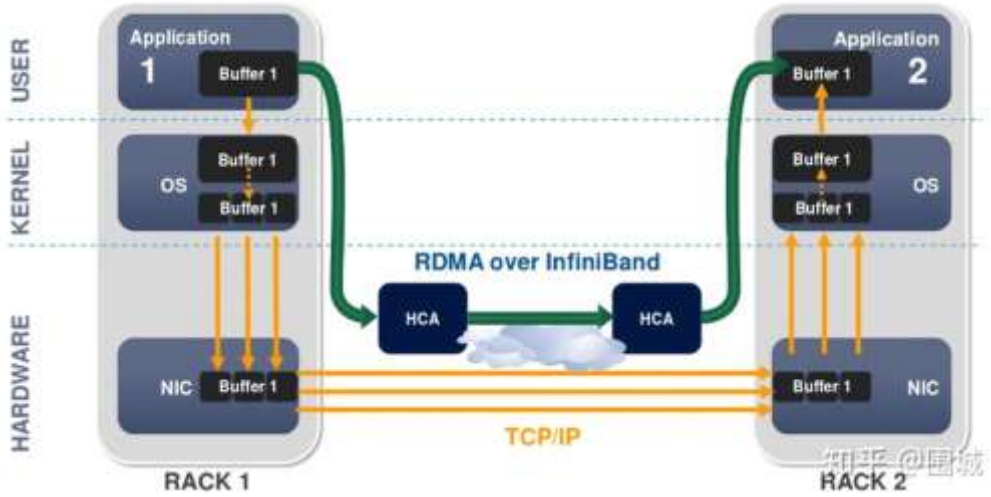
UA-Link ONLINE STORE

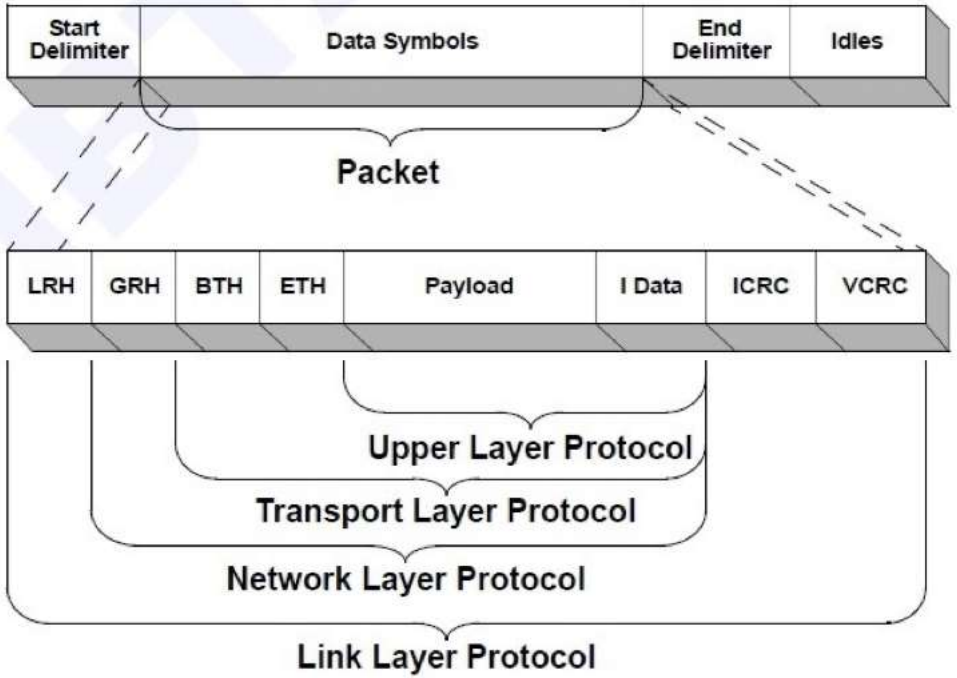
Scale Out 网络：RDMA

一种由网卡硬件直接读写内存的网络传输协议

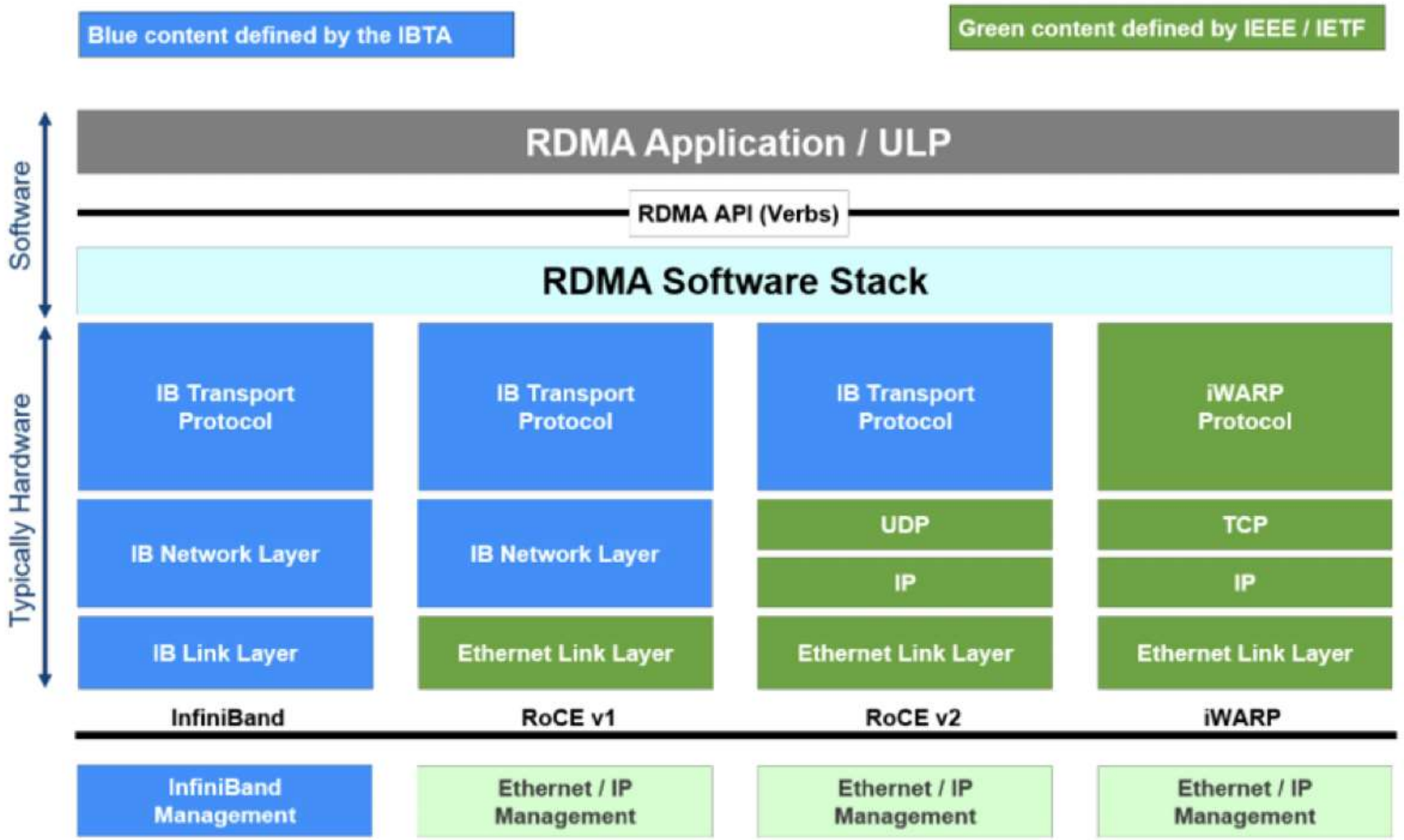
优势：

- 网络转发实现网卡 offload，无需 CPU
- 数据零拷贝
- Linux 内核旁路，无需内核态和用户态上下文切换
- 消息基于事务
  数据被处理为离散消息而不是流，消除了应用程序将流切割为不同消息/事务的需求





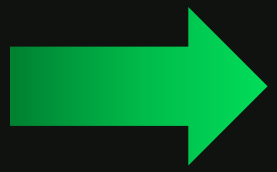
RDMA 技术的实现



RDMA 需要无损网络运行环境

2% 丢包率

GPU 集群 RDMA 吞吐量降为0



无损网络

不产生丢包的网络

RoCE 无损网络

- 以太网交换机 PFC 和 ECN

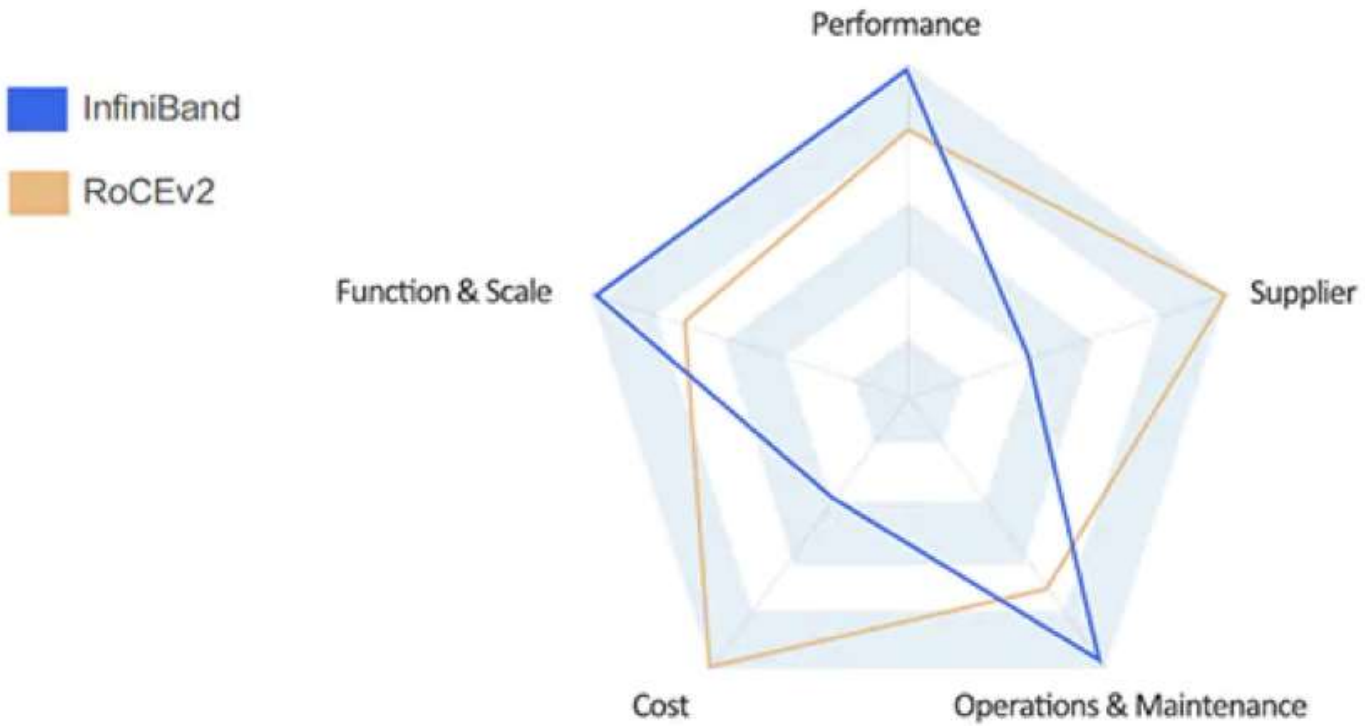
Infiniband 无损网络

- 信用机制的缓存流控机制
- 将物理链路划分为多个虚拟通道
- 暂停帧，实施数据的暂停发送

RoCEv2 or InfiniBand ?

In practical business scenarios
RoCEv2 is a good solution
while InfiniBand is an excellent solution

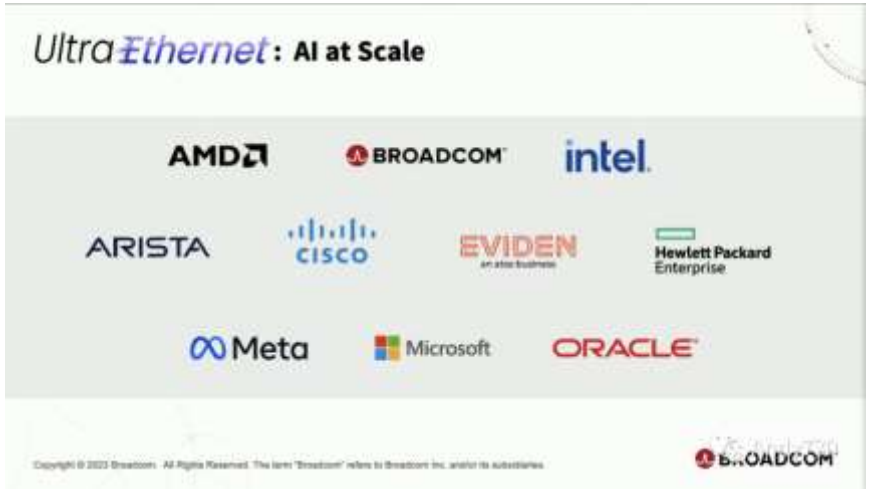
RoCEv2 or InfiniBand 综合比较

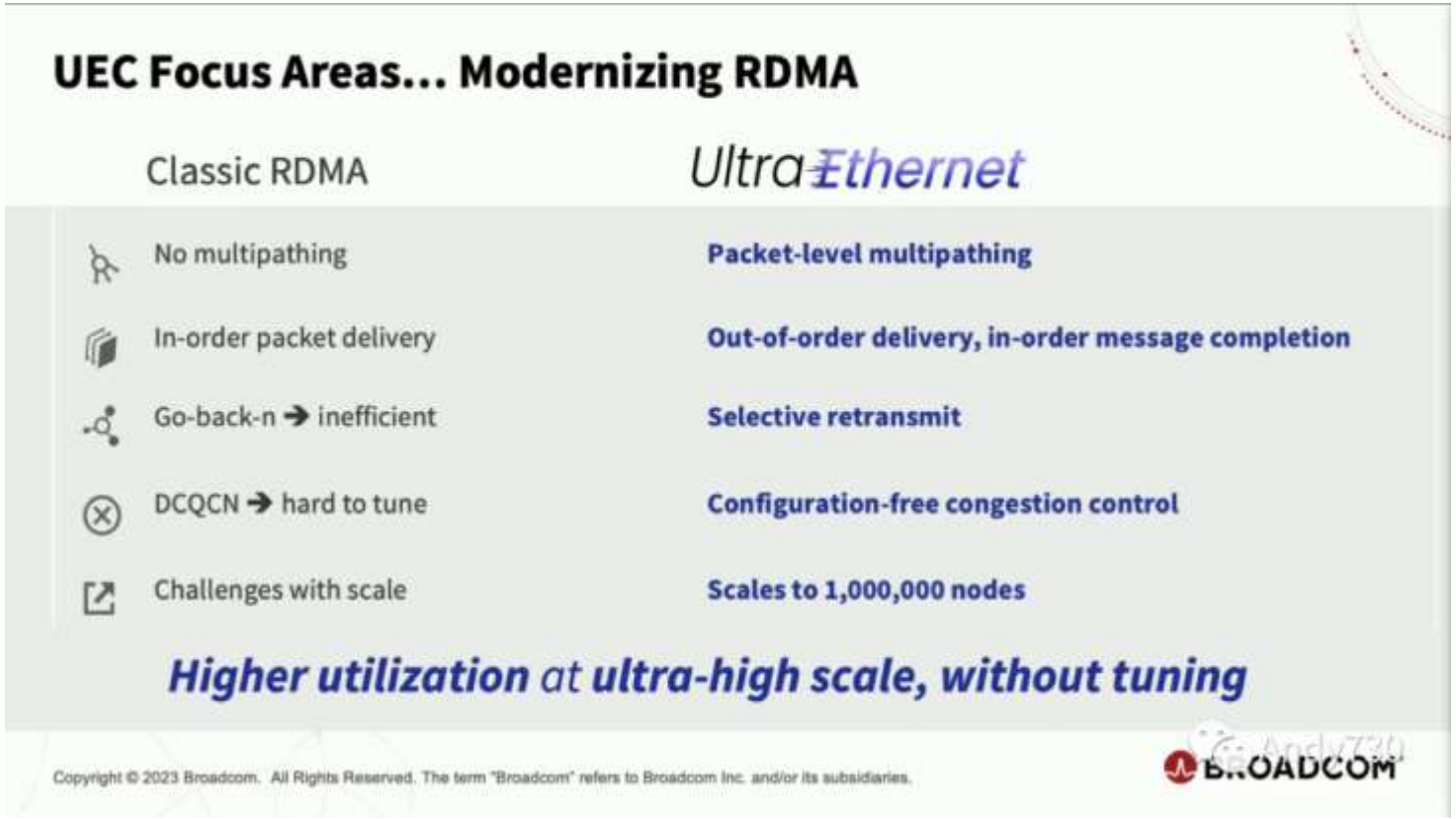


- **规模**
  Infiniband 有效支持上万个GPU集群；RoCE 有效支持数千个 GPU 集群，更大规模的网络需要进行精细的调优
- **运维**
  Infiniband 网络支持成熟的观测运维
- **硬件成本**
  Infiniband 网络设备的成本较高
- **供应商**
  Nvidia 几乎垄断 Infiniband 网络设备；RoCE 供应商选择多样
- **性能**
  在大规模网络场景下，拥塞控制导致网络带宽利用率表现存在差异

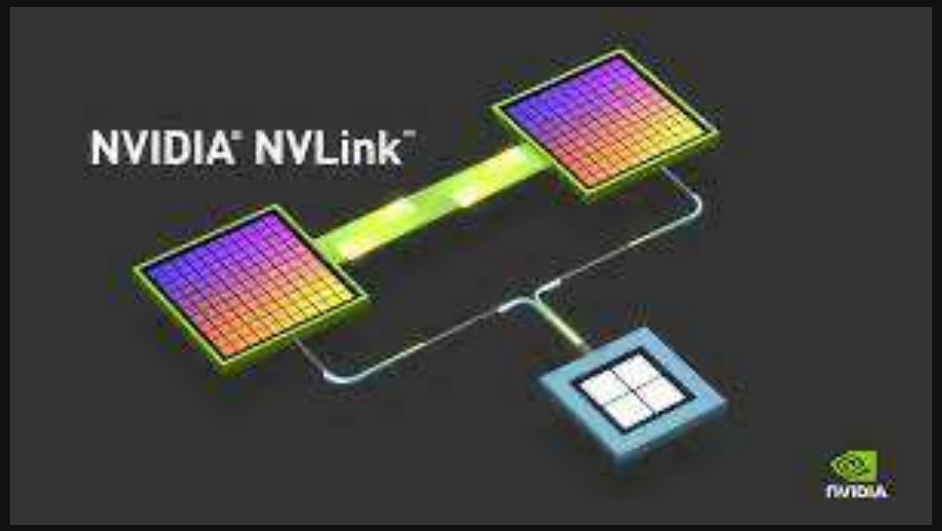
业界新趋势： 超级以太网联盟 UEC（Ultra Ethernet Consortium）

在以太网上使用新的传输协议，来取代（兼容） RoCE，实现更加高效的网络，希望打破 nvidia 的 Infiniband 垄断。当前在协议制定阶段，预计会在 2025 年开始陆续出现相关产品





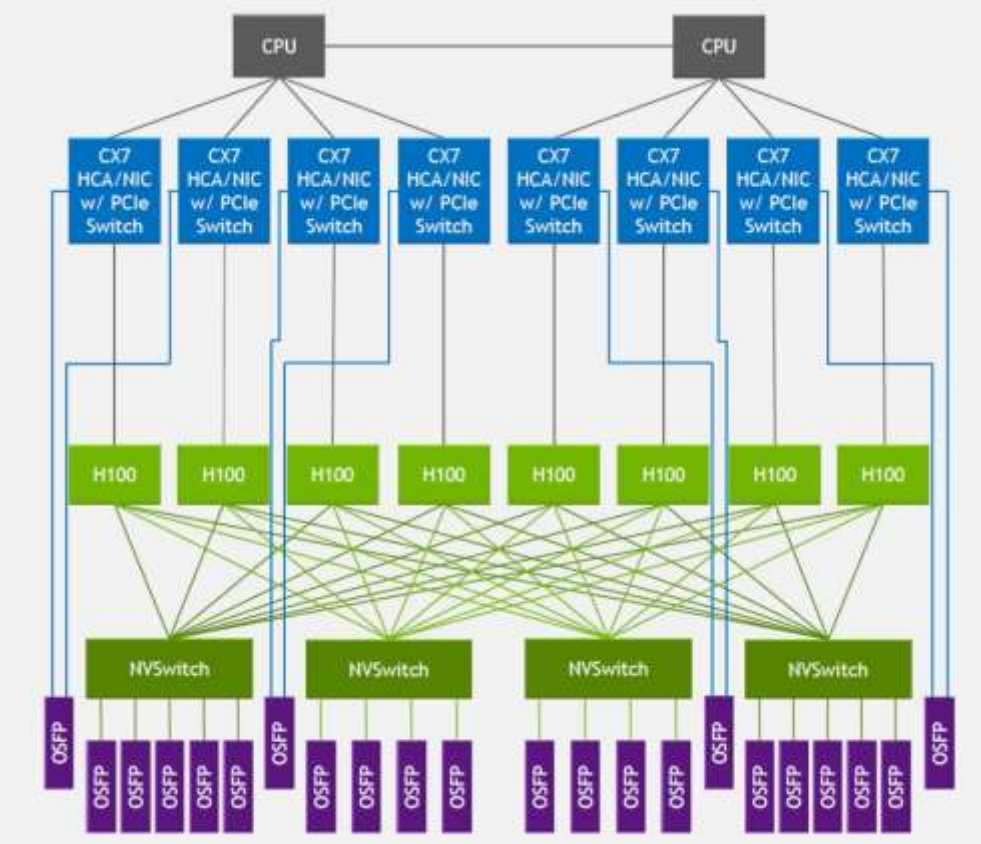
Scale Up 网络：NVLINK

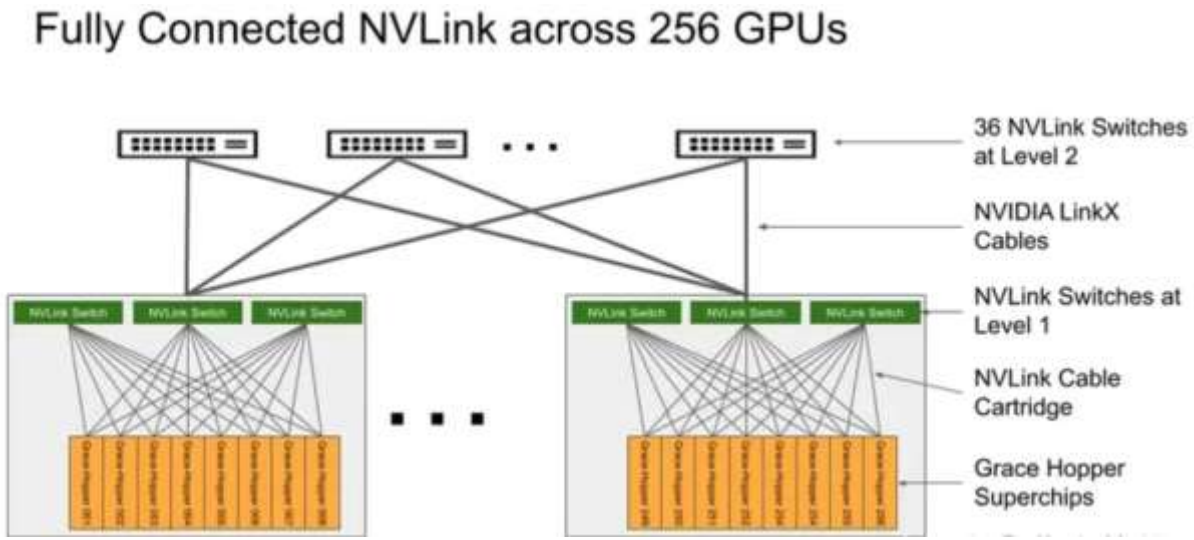


解决传统 PCIE 的通信问题：

- 更高带宽和更低延时
  nvlink 5.0 提供最大 50GB/s x 18 = 1.8 TB/s 的双向带宽
  16 通道 pcie 5 提供最大 128 GB/s 双向带宽
- 支持统一虚拟内存寻址
  nvlink 提供统一的虚拟内存地址空间（Unified Memory）。在多 GPU 系统中，通过 NVLink，多个 GPU 可以共享彼此的显存地址。
  而传统的 PCIE 不支持 GPU 之间的统一虚拟地址空间，需要通过主机内存的拷贝，帮助 GPU 之间同步数据

NVLINK 和 NVSWITCH 网络





Part 02

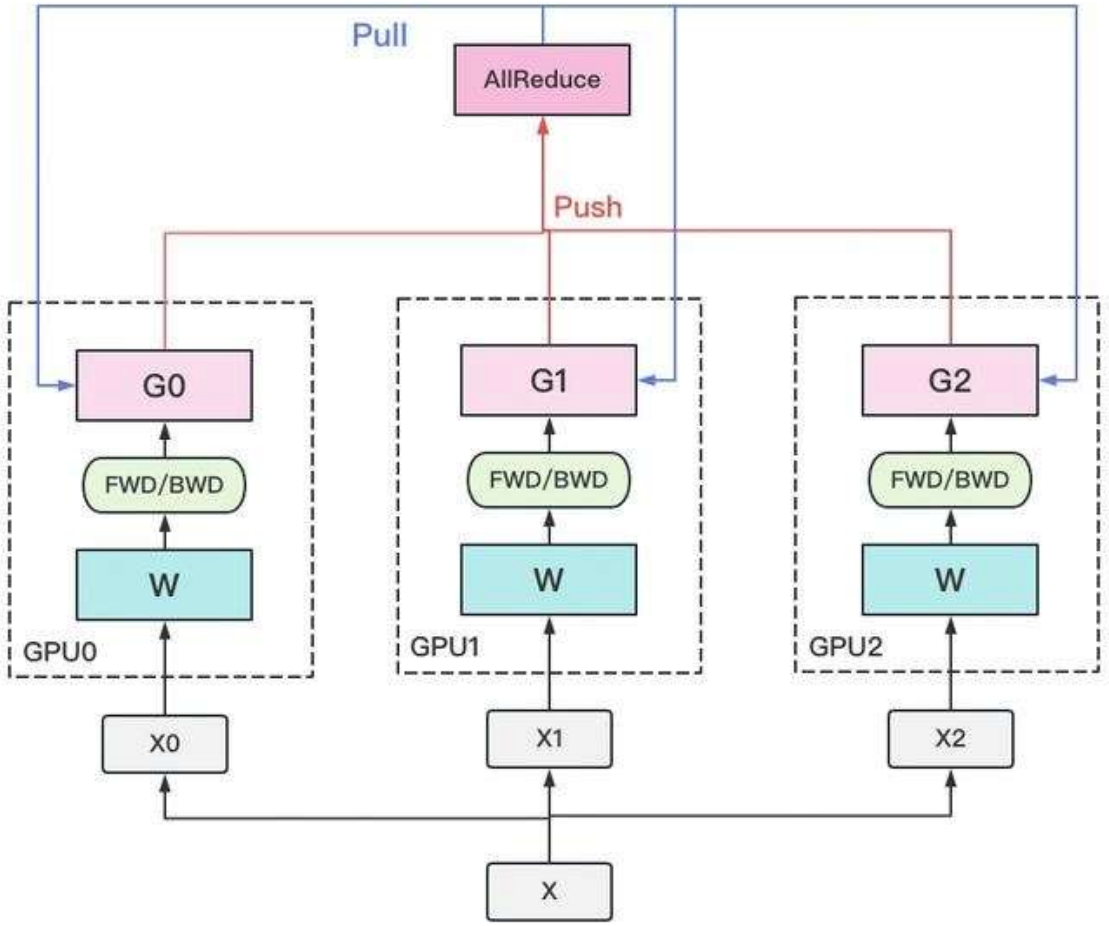
PTD-P 和 AI 网络流量特点

数据并行

DP

庞大的数据集拆分成多份，给到多份模型副本进行训练，每一轮迭代训练完成后，各个 GPU 需要把各自反向计算得到的梯度做全局同步 AllReduce

特点：AllReduce 网络流量大



流水线并行

PP

模型按照神经元，把不同的层分配到不同的 GPU 上运行，且对数据集进行微批次的拆分，以避免层间计算的顺依赖而降低 GPU 利用率

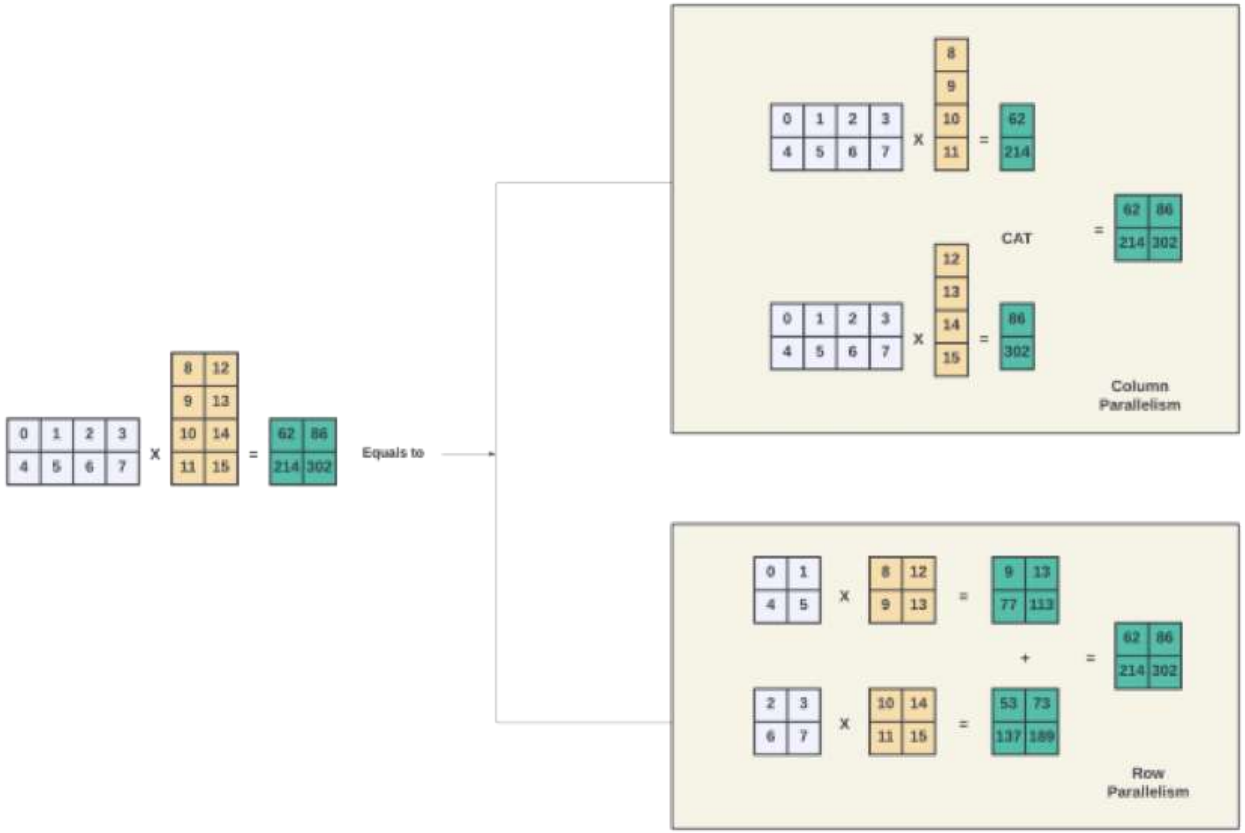
特点：point-to-point 通信，通信量相对小

张量并行

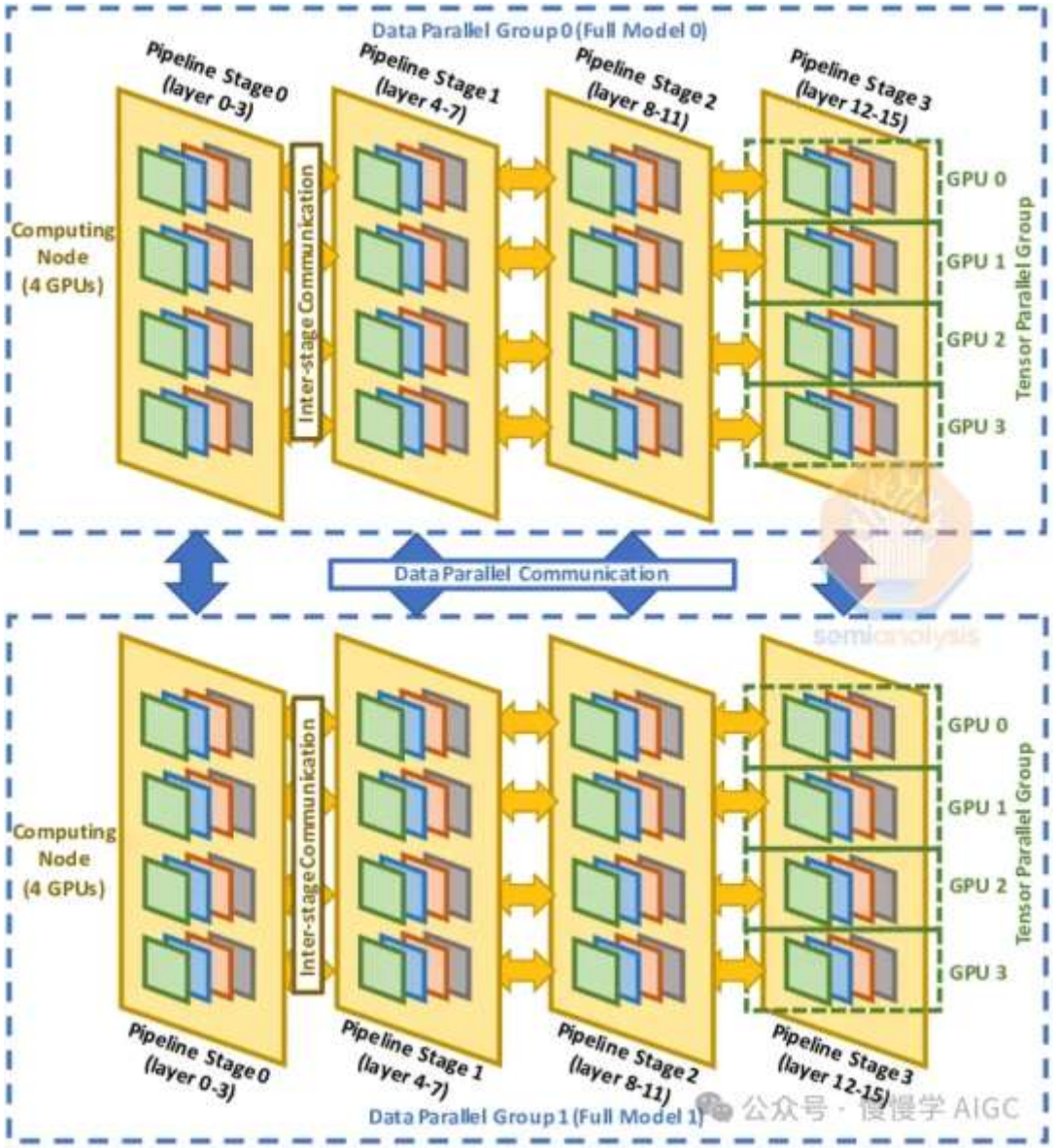
TP

层内分割，把某一个层做切分，放置到不同 GPU 之上，也可以理解为把矩阵运算分配到不同的设备之上

特点：传输的数据量巨大，且通信频繁



PTD-P



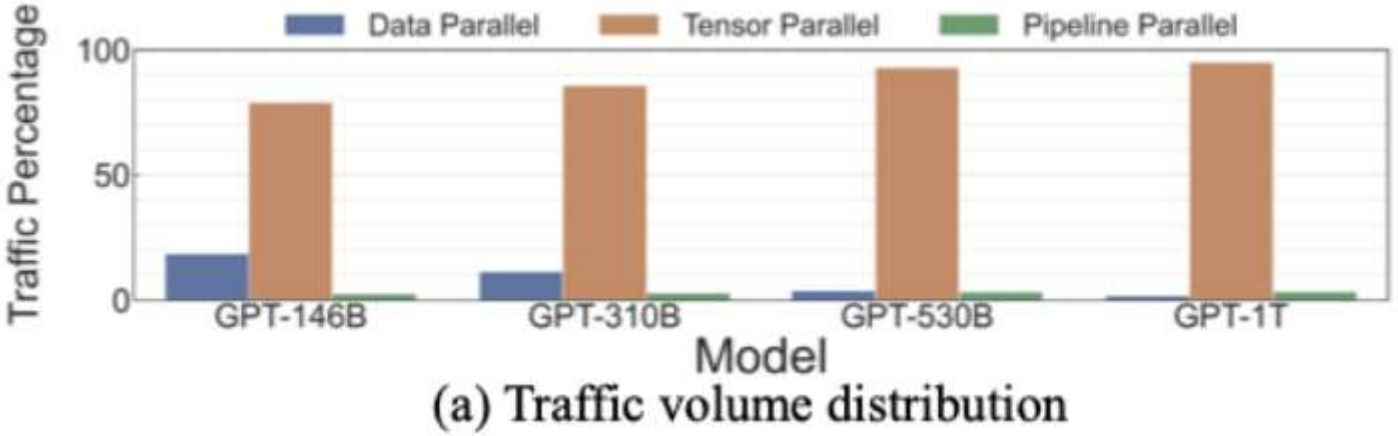
- 一个节点内 4 GPU，做张量并行
- 流水线并行策略，模型不同层拆分到水平 4 节点上
- 垂直方向上，做 2 路数据并行

AI 负载的网络流量特点

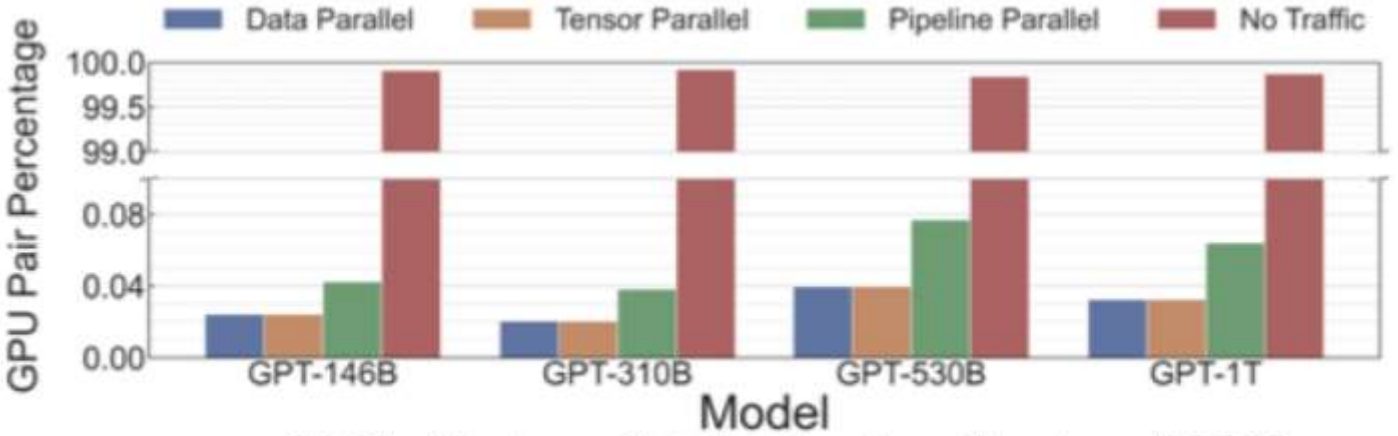
测试环境：
DGX A100
HB domain of size eight.
the default algorithm in NCCL

规律：
- TP 流量主要在 HB 域内转发，几乎不会出现在网络中
- 集群中超过 90% GPU Pairs 之间是没有网络传输的，不同 GPU pairs 之间几乎没有流量。
- 网络流量中，90%流量是 DP （AllReduce） 和 PP （point-to-point）流量

结论：
基于 any-to-any 的网络拓扑连接是浪费资源



(a) Traffic volume distribution



(b) Traffic type distribution for all pairs of GPUs

Part 03

算力网络拓扑设计

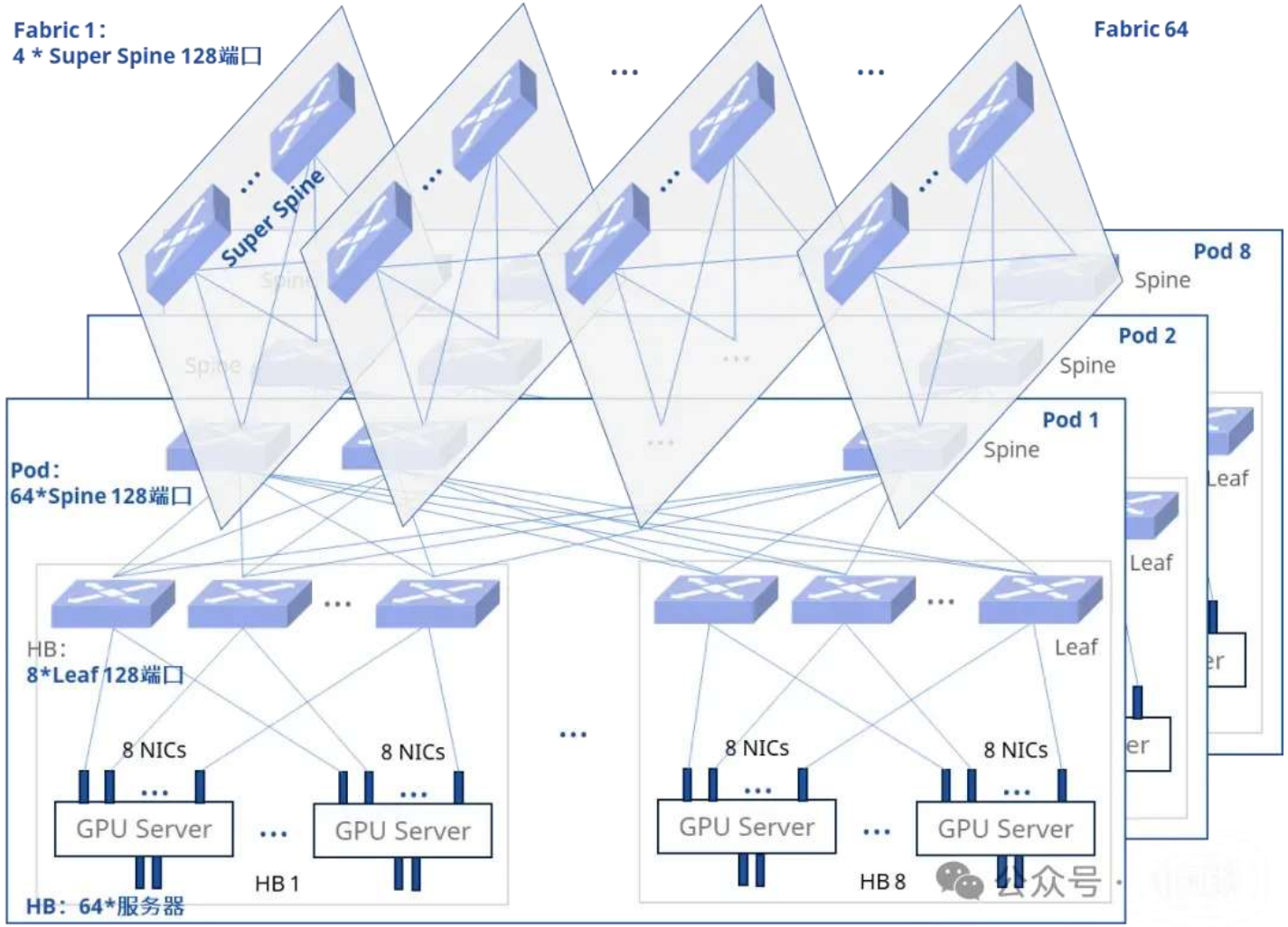
传统数据中心网络拓扑

32768 GPU 集群集群

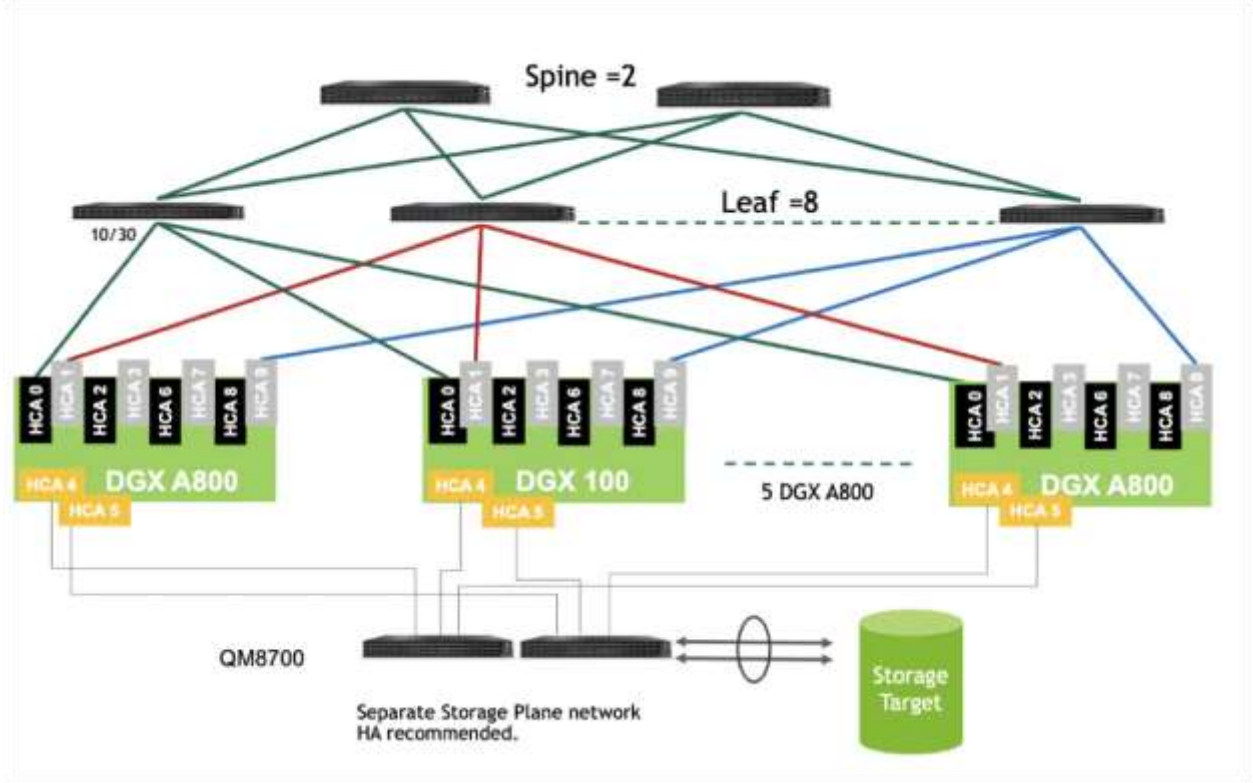
- full mesh 全互联
- 3 层 fat-tree 设计
- 总的 Switch
  256+512+512=1280
- 总的光模块数
  1280*128+32768=196608

缺点：

- 硬件成本和电力成本
- 网络设备管理成本
- 提高PFC 死锁、环路等风险
- 更多的硬件，意味更多的故障，
  提高训练中断的风险

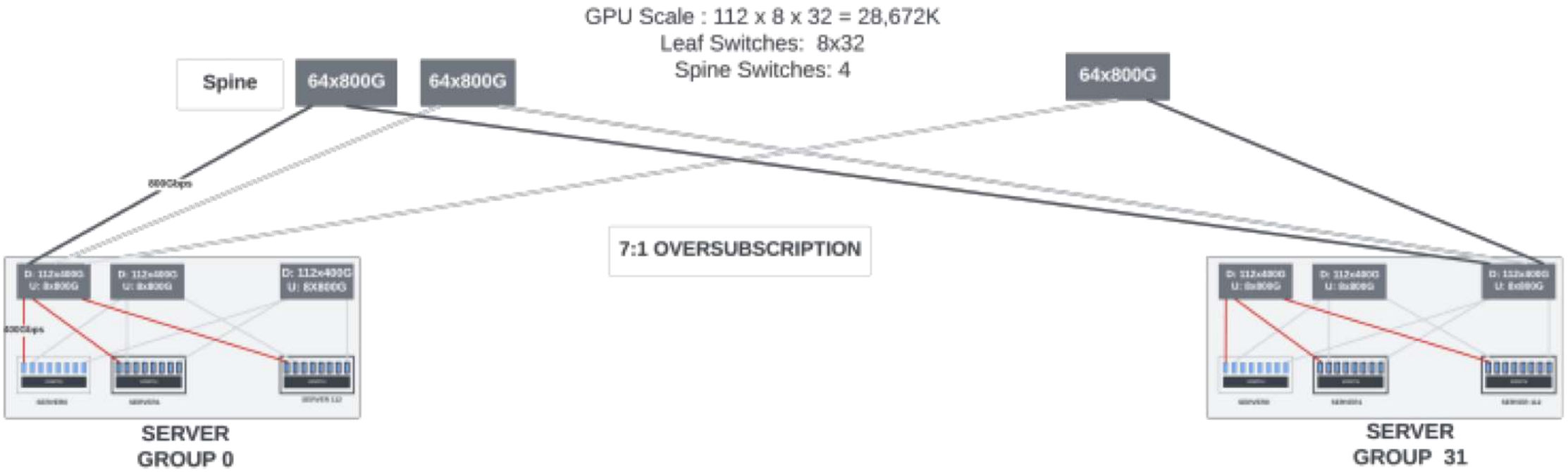


rail optimized 小规模 512 卡组网



- 前后端网络分离
- 受限于交换机端口数量，采用1:1 带宽收敛比设计，8 个 400G leaf switch （最大64口） + 2 个 800G spin switch （最大64口），两层交换机，连接最大 512 个 A800 GPU （200G网卡)
- 实现 any-to-any 连接，支持 all-reduce 操作的性能最大

3 万卡超大规模优化组网



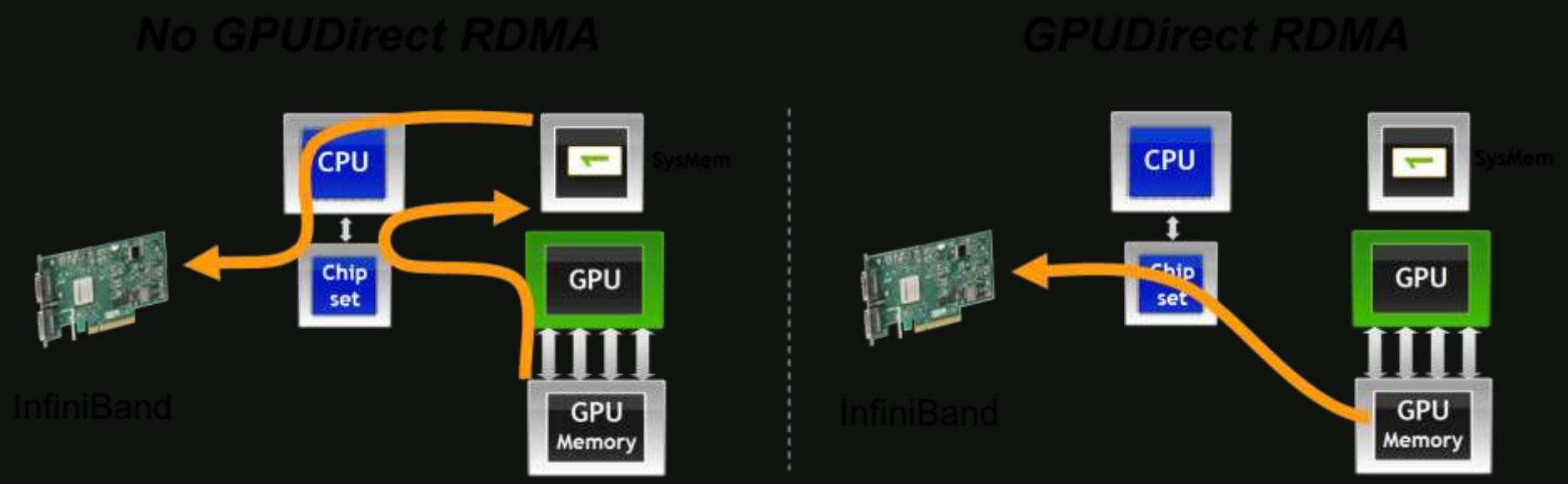
3W GPU 集群

- 实际的 spine -leaf 的流量主要是DP 数据， 采用 7:1 带宽收敛比
- 单 server group 内 有 8 个 400G 的 leaf 交换机（128 port ），连接 896 个 H100(400G网卡)
- 32 个 800G spin 交换机（64 port ） 连接 32 个 server group
- 共连接 32*896 = 2.8万GPU
- 交换机数量：256 + 32

Part 04

提升 AI 网络性能

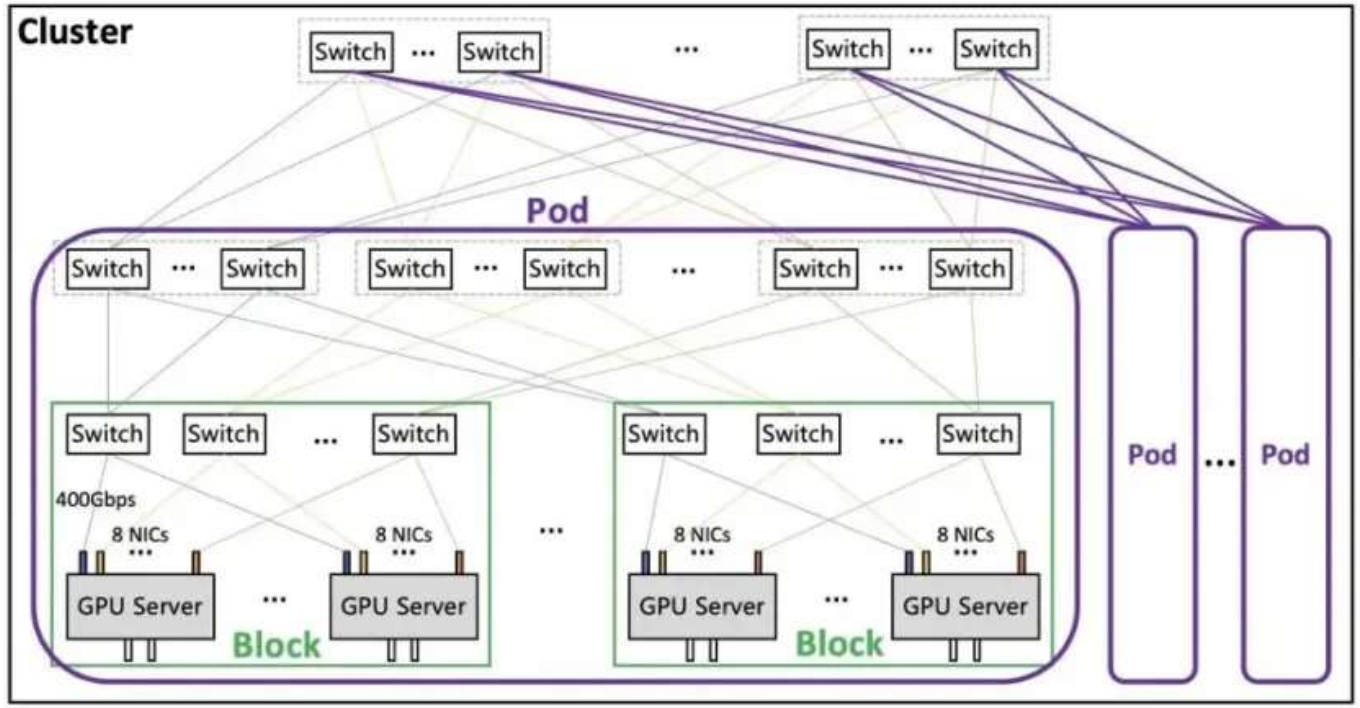
RDMA 加速 GPU 显存同步：GPU direct RDMA



GDR 允许网卡直接读写 GPU 显存，无需 CPU 和 内存的干预

- 降低延迟
- 提高吞吐量
- 降低 CPU 开销

合理的 POD 网络调度，降低网络拥塞



降低网络跳数，才能保障网络性能

- 若干 node 划分为 block，若干 block 划分为 pod，实现不同 GPU 数量规模的调度单元
- 尽量把任务调度到集中的 block 或 pod 中，减少数据传输的交换机跳数
- 能够根据网络遥测数据进行调度

RoCE 在大规模网络下的拥塞改进

ROCE，所有数据包必须按顺序到达，如果出现丢包或乱包，会导致回退N帧问题，影响传输效率

- 配置无损网络，避免丢包
  （1）PFC
  上下游交换机之间的拥塞控制机制，暂停发包
  缺点：不公平问题，会暂停任何流的转发
  （2）ECN
  为了避免触发 PFC ， ECN 实现端到端的拥塞控制，发送方通过 DCQCN 算法调整发包速率
  （3）Automatic ECN
  解决传统 ECN 的人工运维负担，实现自动化 ECN 调节

- 提高网络带宽利用率，避免出现网络部分链路过度拥塞，部分链路闲置

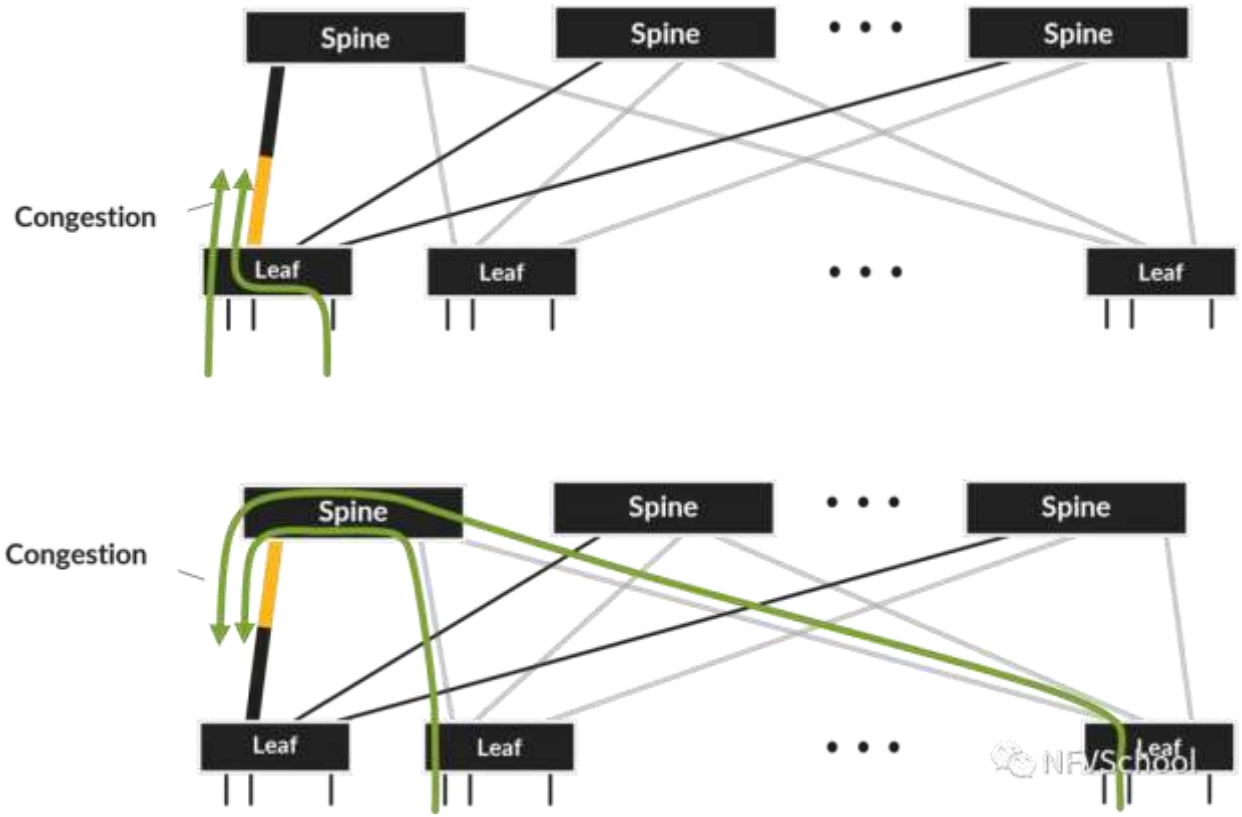
提高 RoCE 网络带宽利用率

问题

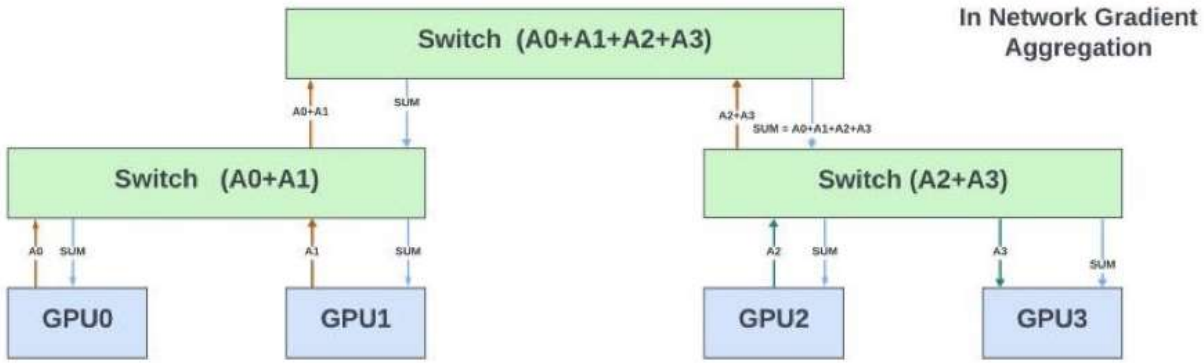
- 大象流问题
- RoCE 不支持数据包乱序，导致重传
- ECMP 基于流的负载均衡，该模型下的网络带宽利用率不足 50%

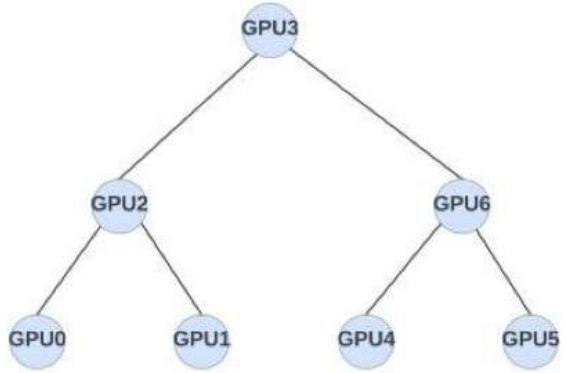
方案

- Nvidia 的自适应路由和数据包喷洒，配合 DPU 网卡的数据包乱序重组和标记，实现高达 95% 的有效带宽
- telemetry-based congestion control
  nvidia Spectrum X 以太网网络平台中，交换机会发送拥塞遥测数据，通知发送方 DPU 网卡，调节发送速率
- 其它大厂自建数据中心，自研类似的负载均衡技术，提高带宽利用率，确保 ROCE 性能



Infiniband 在网计算和 sharp 协议





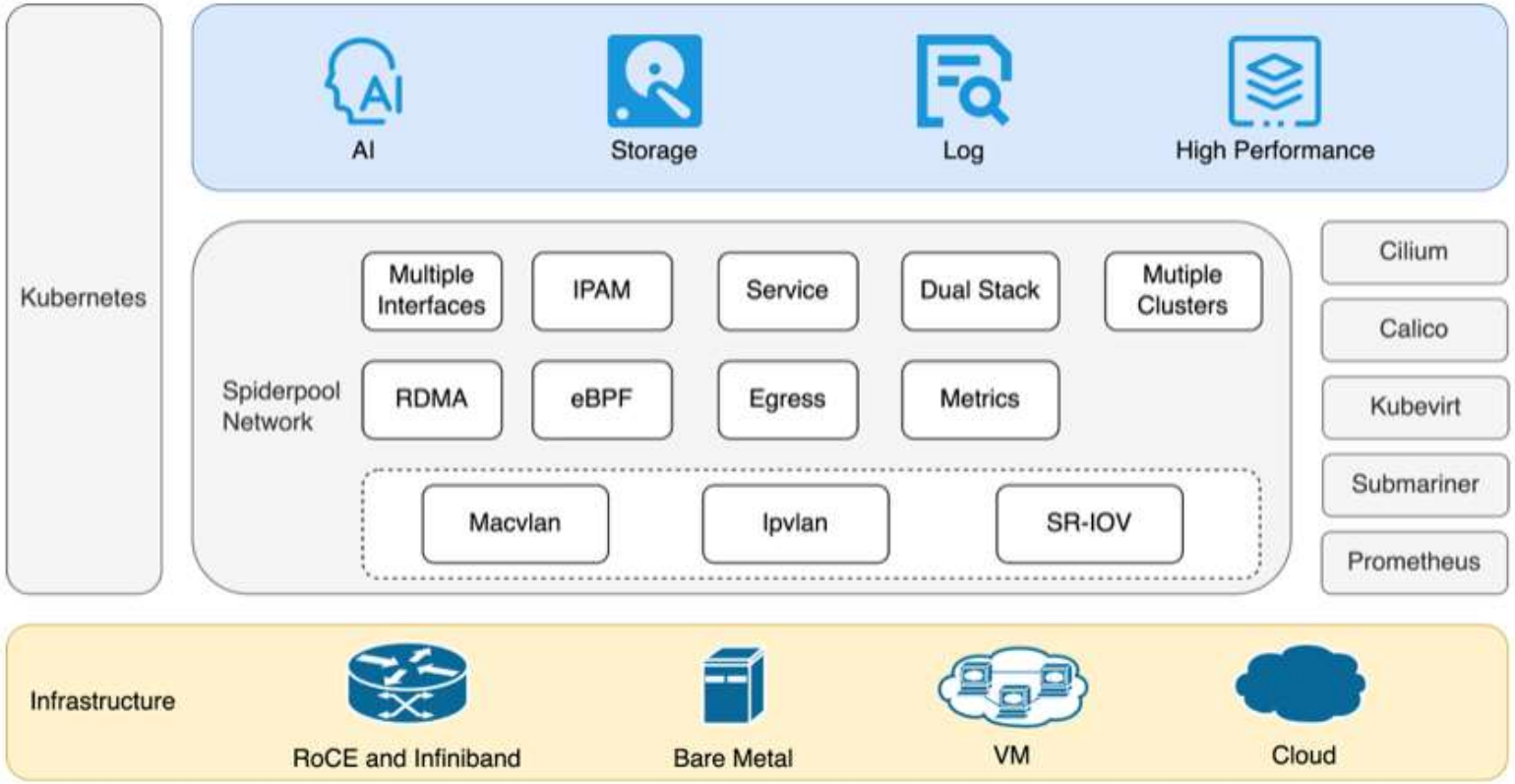
- **要求**:
  nvidia 的 Infiniband 或 nvlink 网络，CX6+ 网卡
- **优势**:
  降低 GPU 之间的网络传输次数和数据量
  加速 MPI 计算
  算力 offload
- **支持的 MPI 操作**
  allReduce / AllGather / Broadcast / ReduceScatter / ScatterGather

Part 05

Kubernetes 下的 CNI
提供 AI 负载的 RDMA 通信设备

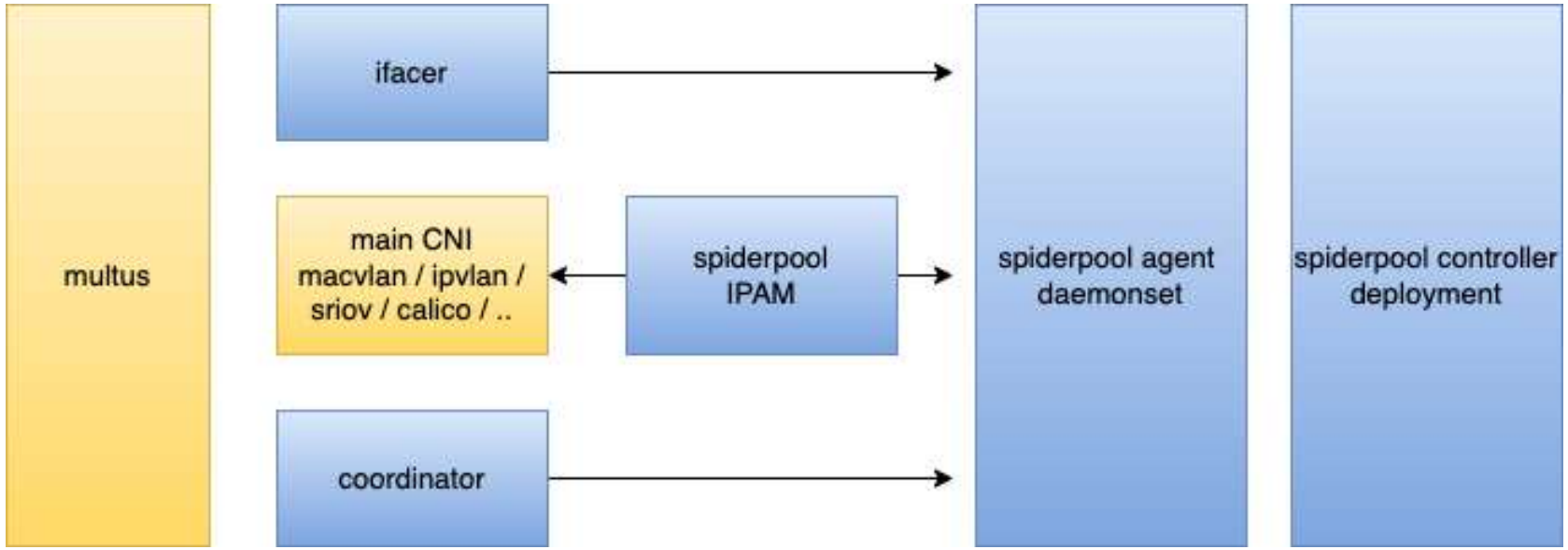
Spiderpool

作为一个 CNCF 沙盒项目，Spiderpool 是 Kubernetes 的底层和 RDMA 网络解决方案，特别适合网络 I/O 密集型和低延迟的应用场景，如存储、中间件和 AI。它可以运行在裸机、虚拟机以及公共云等多种环境中。



- First release in 2022
- CNCF Landscape Project
- 2023 信通院 “云原生技术创新”

架构



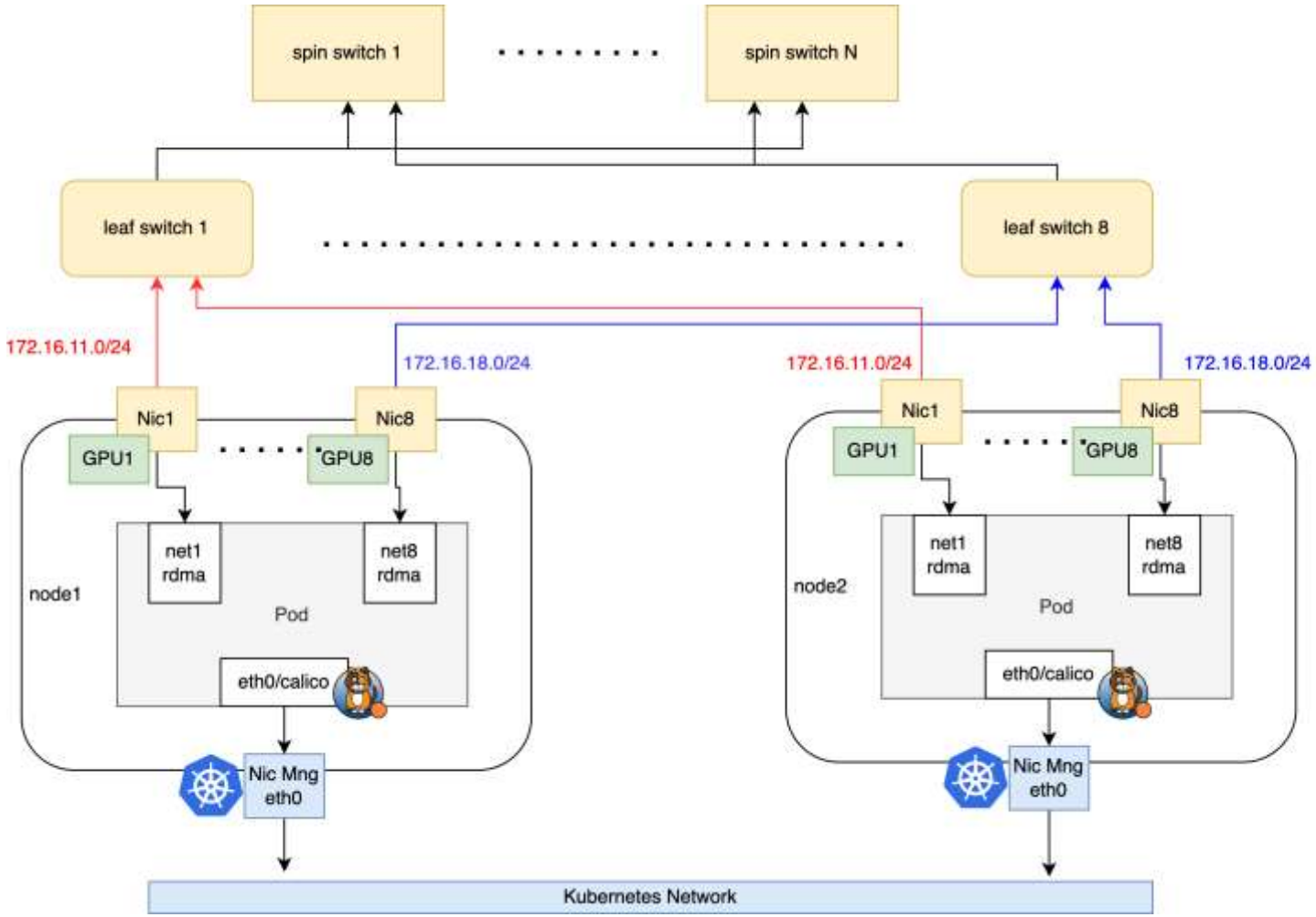
Spiderpool 整合了当前开源社区的 CNI 能力，以轻量化的架构实现容器 underlay 网络方案，其构成主要包括：

- Spiderpool agent，以 daemonset 运行在每一个节点上，只要为以下 CNI plugin 提供服务：
  1. spiderpool IPAM plugin，基于 custom resources 管理方式为 main CNI 插件分配 IP 地址
  2. coordinator meta plugin，完成多样化的工作，包括多网卡策略路由调谐、检测 IP 冲突、检测网关可达等
  3. ifacer meta plugin，在宿主机网络命名空间中创建 bond 和 vlan 虚拟接口
- Spiderpool controller，以 deployment 运行，完成 IPAM 的 GC、CR资源校验等

RDMA CNI 方案

	Infiniband with SR-IOV	Infiniband with IPoIB	RoCE wirh SR-IOV	RoCE with ipvlan/macvlan
适用环境	bare metal	bare metal and VM	bare metal	bare metal and VM
提供 RDMA 设备	√	×	√	√
RDMA 共享模式	exclusive or shared	×	exclusive or shared	shared
DPDK	√	×	√	×
Bandwidth	√	×	√	×
兼容传统 TCP 应用	兼容 或双 CNI 网络	适用于传统以太网应用	兼容	兼容

spiderpool AI 组网



- 能够检测 GPU 分配，实现基于 PCI-E 亲和的 RDMA 设备动态分配，有效支持 GDR
- 多种 RDMA 设备接入方案，适用于 RoCE 和 Infiniband 场景
- 多网卡：多网卡分配、路由调谐
- 生产稳定、功能强大的 IPAM，支持 IPv4-only，IPv6-only，Dual-Stack
- 可与 overlay 搭配，也可作为唯一的 CNI 支撑 POD 的所有网络通信，包括 Service Access，避免搭建双 CNI

spiderpool 简化的 AI workload 部署

系统管理员创建网络资源

```yaml
apiVersion: spiderpool.spidernet.io/v2beta1
kind: SpiderIPPool
metadata:
  name: gpu1-net11
spec:
  gateway: 172.16.11.254
  subnet: 172.16.11.0/16
  ips:
    - 172.16.11.1-172.16.11.200
---
apiVersion: spiderpool.spidernet.io/v2beta1
kind: SpiderMultusConfig
metadata:
  name: gpu1-sriov
  namespace: spiderpool
  labels:
    cni.spidernet.io/rdma-resource-inject: gpu-ibsriov
spec:
  cniType: ib-sriov
  ibsriov:
    resourceName: spidernet.io/gpu1sriov
    rdmaIsolation: true
    ippools:
      ipv4: ["gpu1-net91"]
```

应用管理员书写 annotation

Pod yaml:

....

annotation:

cni.spidernet.io/rdma-resource-inject=gpu-ibsriov

最终生效，自动注入资源和网卡

```yaml
...
  annotations:
    k8s.v1.cni.cncf.io/networks: |-
        [{"name":"gpu1-sriov","namespace":"spiderpool"},
        {"name":"gpu2-sriov","namespace":"spiderpool"},
        {"name":"gpu3-sriov","namespace":"spiderpool"},
        {"name":"gpu4-sriov","namespace":"spiderpool"},
        {"name":"gpu5-sriov","namespace":"spiderpool"},
        {"name":"gpu6-sriov","namespace":"spiderpool"},
        {"name":"gpu7-sriov","namespace":"spiderpool"},
        {"name":"gpu8-sriov","namespace":"spiderpool"}]
...
    resources:
      limits:
        spidernet.io/gpu1sriov: 1
        spidernet.io/gpu2sriov: 1
        spidernet.io/gpu3sriov: 1
        spidernet.io/gpu4sriov: 1
        spidernet.io/gpu5sriov: 1
        spidernet.io/gpu6sriov: 1
        spidernet.io/gpu7sriov: 1
        spidernet.io/gpu8sriov: 1
```

Part 06

RoCE 网络可观测性

DCE 的 RDMA 网络监控

RMDA 网络可观测性的意义

展示网络带宽利用率和 RDMA 网络拥塞情况

有效追踪网络设备健康状态，助力快速恢复集群运行。例如，Llama 3 405B 训练 45 天期间，网络交换机/线缆故障 35次 ， 占据所有故障的 8.4%

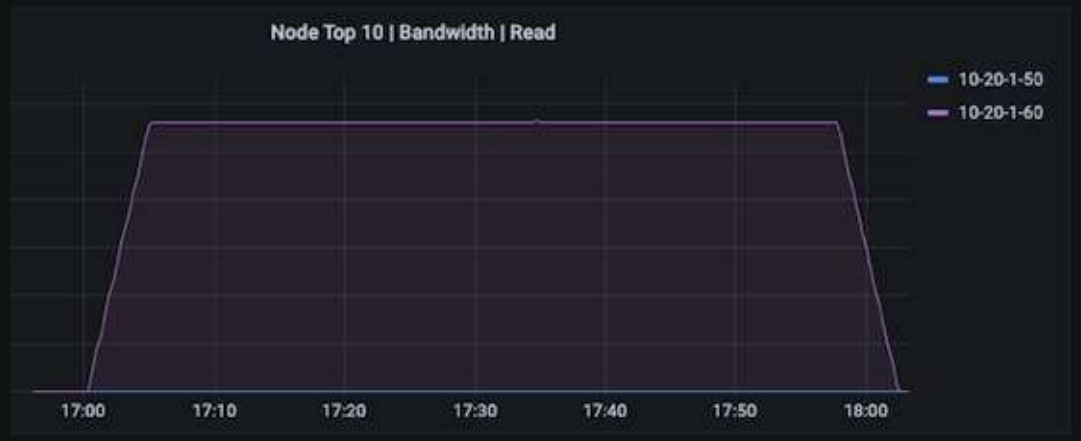
追踪 RDMA 通信吞吐量的热点 POD 和 node，实现网络拓扑、容器调度等调优

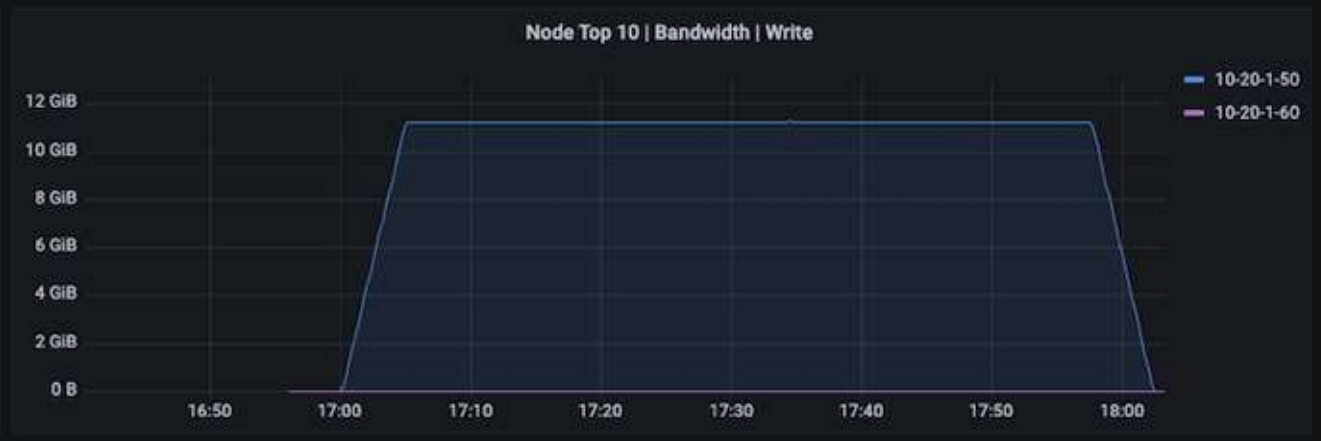
DCE 的 RDMA 网络监控，覆盖容器和交换机：

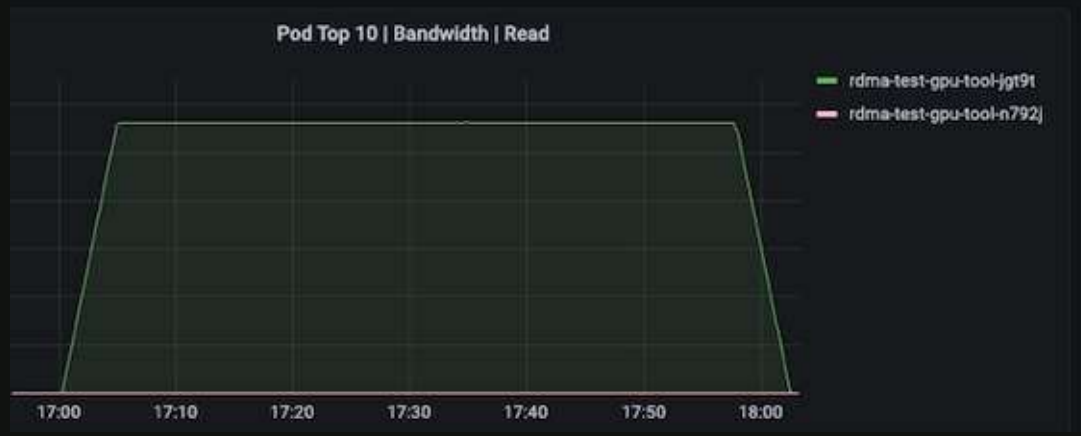
- 展示网络交换机、集群主机的拓扑
- 查看网络交换机端口状态和指标
- 网络设备健康状态监控和告警
- 集群、node、POD 、交换机等多维度的 RDMA 指标监控
- 追踪单次 AI 任务中的所有容器、交换机的指标和路径

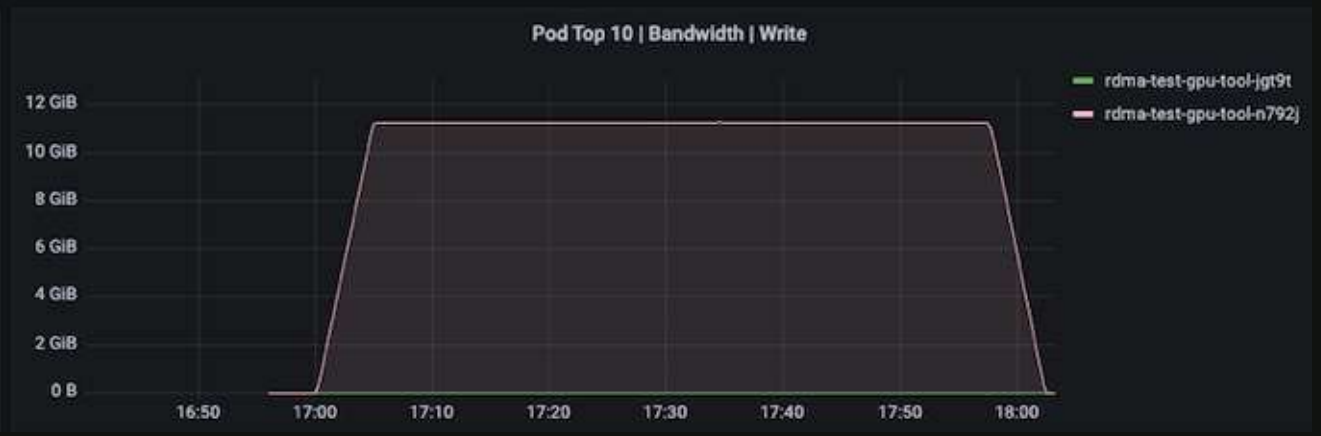
DCE AI 任务 RDMA 流量监控

DCE 集群 RDMA 流量监控









DCE 节点 RDMA 流量监控

DCE POD RDMA 流量监控

DaoCloud

回顾

回顾

- RDMA 和 NVLINK 有效支撑 GPU 之间的数据传输
- PTD-P 并行计算的特点，设计出 rail optimized 的 AI 网络拓扑
- 保障交换机中的 RDMA 传输效率是 AI 网络的关键
- Kubernetes 下容器化的 AI 负载需要 CNI 组件支持网络传输
- RDMA 网络可观测性

DaoCloud

谢谢

欢迎扫码入群讨论交流，获得课件