DaoCloud

# 给 GPU 刷火箭

主讲人：曾祥龙

为 数 字 世 界 寻 找 全 局 最 优 解

Powered By DaoCloud

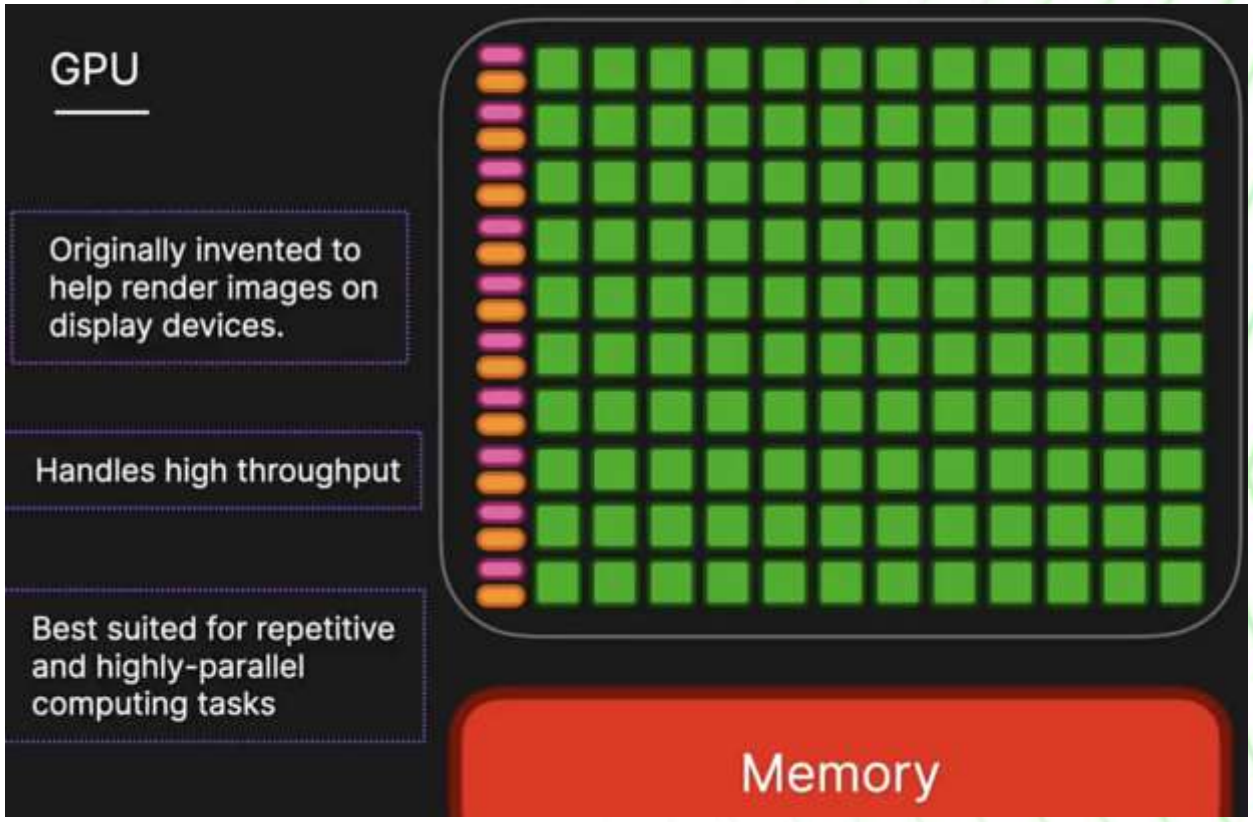
Content

# 目录

# Part 01

## 背景知识速览

# GPU vs CPU

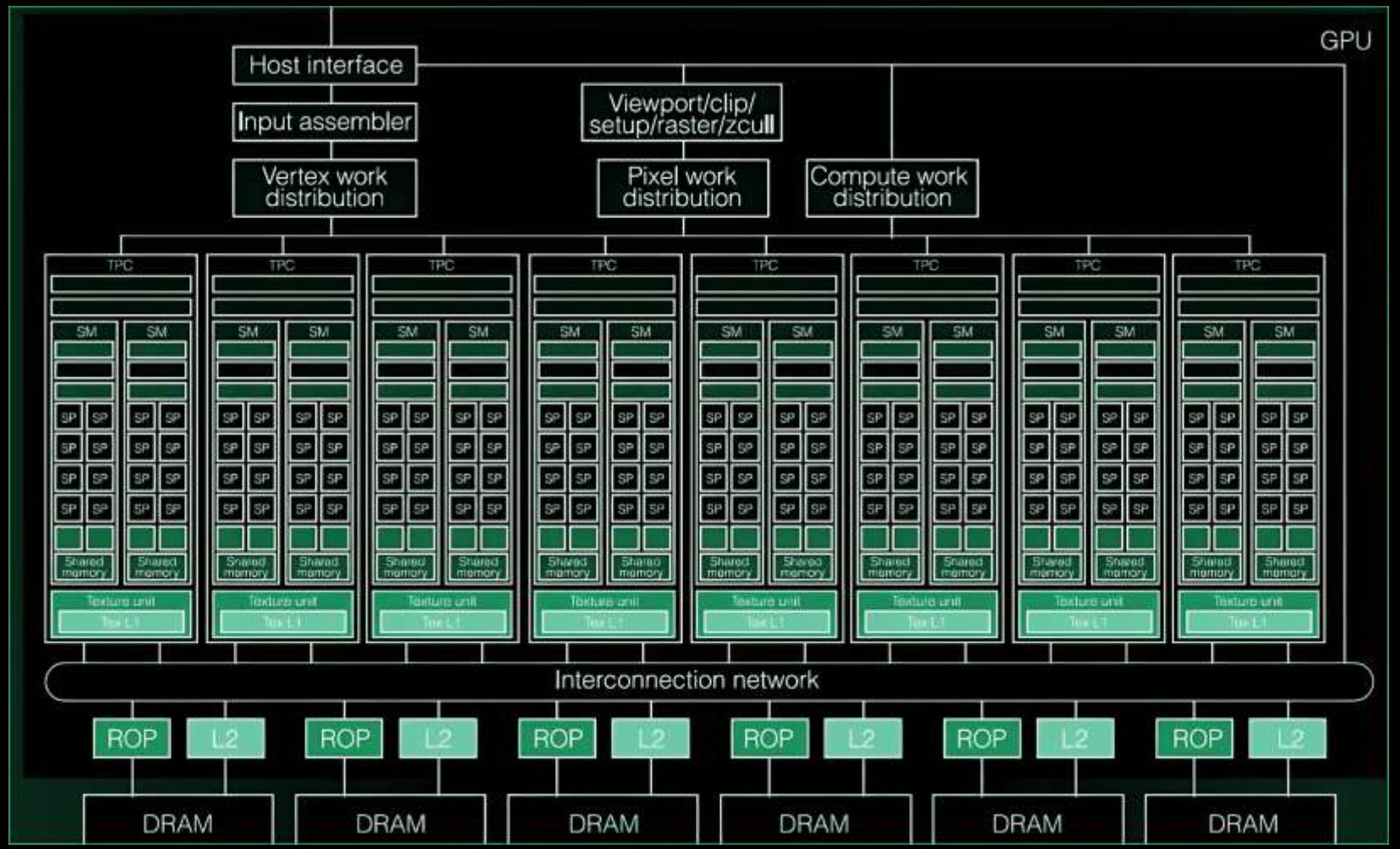



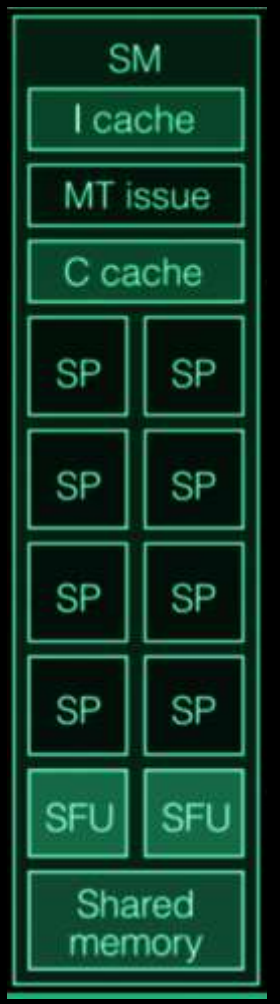
# 对比

| | CPU | GPU |
| --- | --- | --- |
| **功能** | 处理计算机主要处理功能的通用组件 | 非常适合并行计算的专用组件 |
| **处理** | 串行运行进程 | 并行运行进程 |
| **设计** | 更少的但更强大的核心 | 更多核心（性能低于CPU核心） |
| **强调** | 低延迟 | 高吞吐量 |

# GPU架构分析

流式多处理器 SM




每个 SM 下面包括各种缓存和更小的处理单元（SP）
SP：流式处理器

# NVISION 08 大会上现场演示的GPU和CPU对比测试

# GPU发展概述

1999年

NVIDIA 发布 GeForce 256首次提出GPU概念

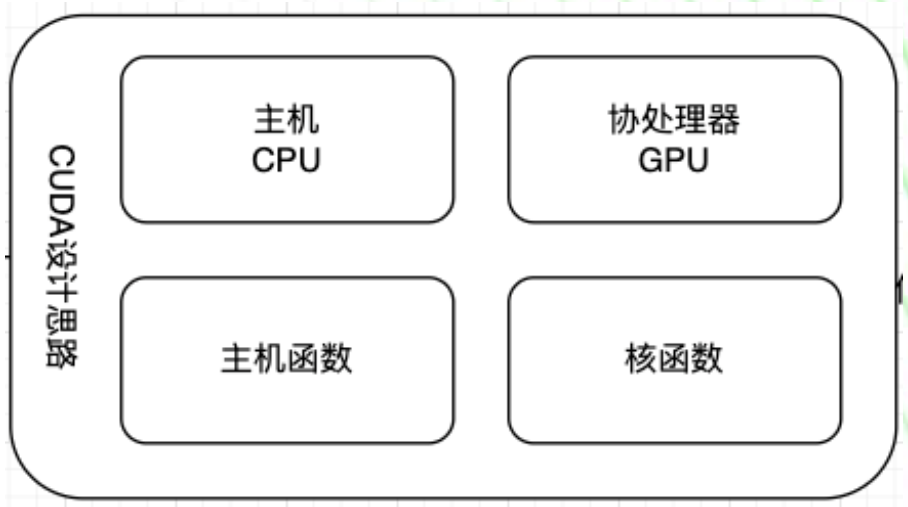
2003年

NVIDIA 推出新产品GeForce 8800 GTX，首次提出GPGPU（通用图形处理器）概念，意指GPU在非图形处理领域也支持更通用广泛的计算

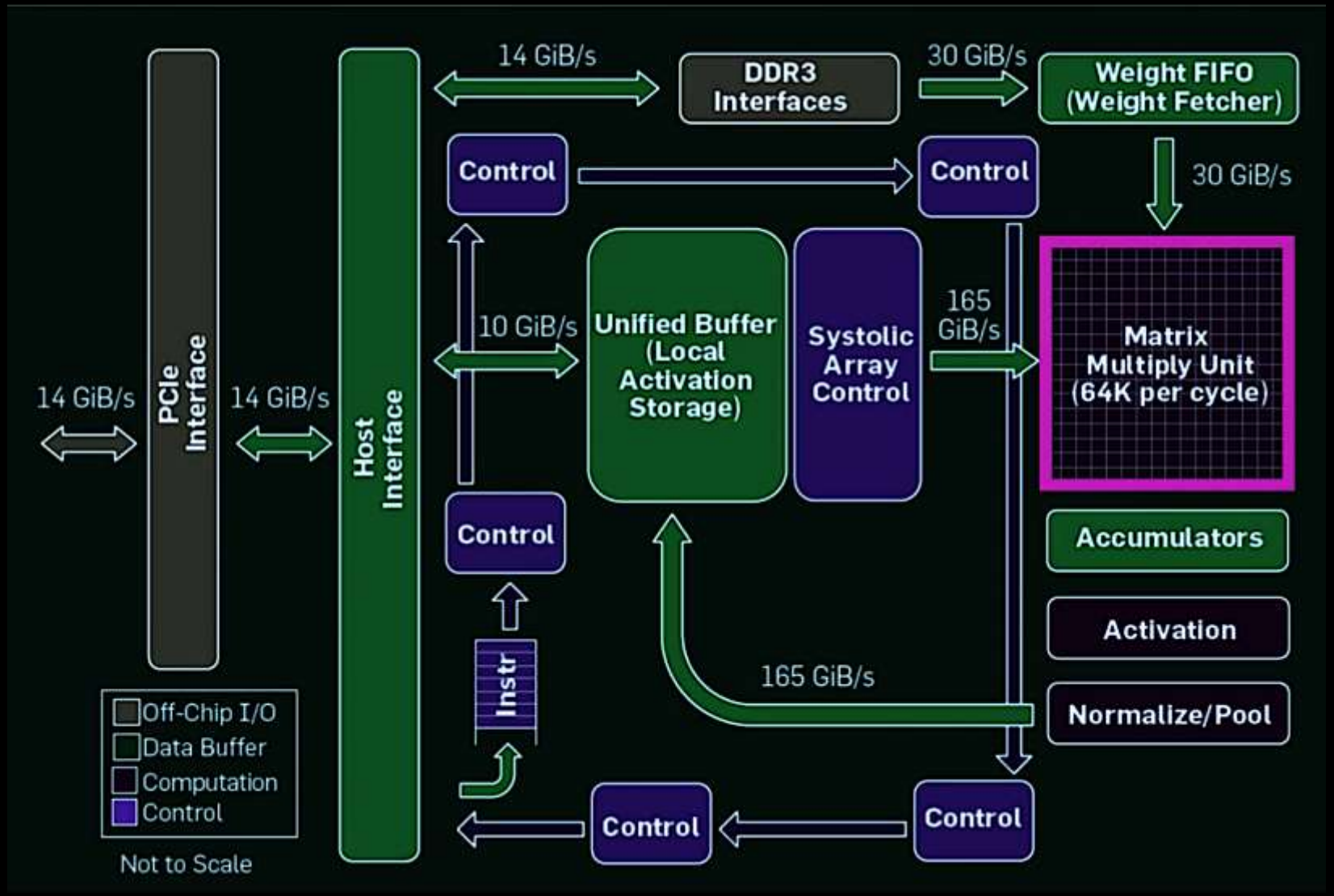
2007年

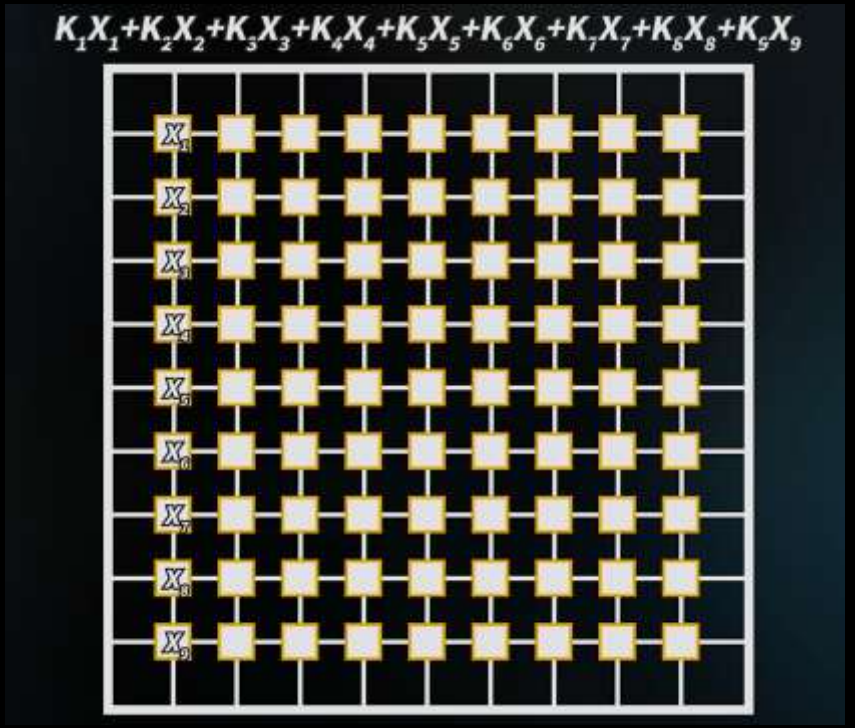
NVIDIA进CUDA

随后，AMD也推出了OpenCL架构

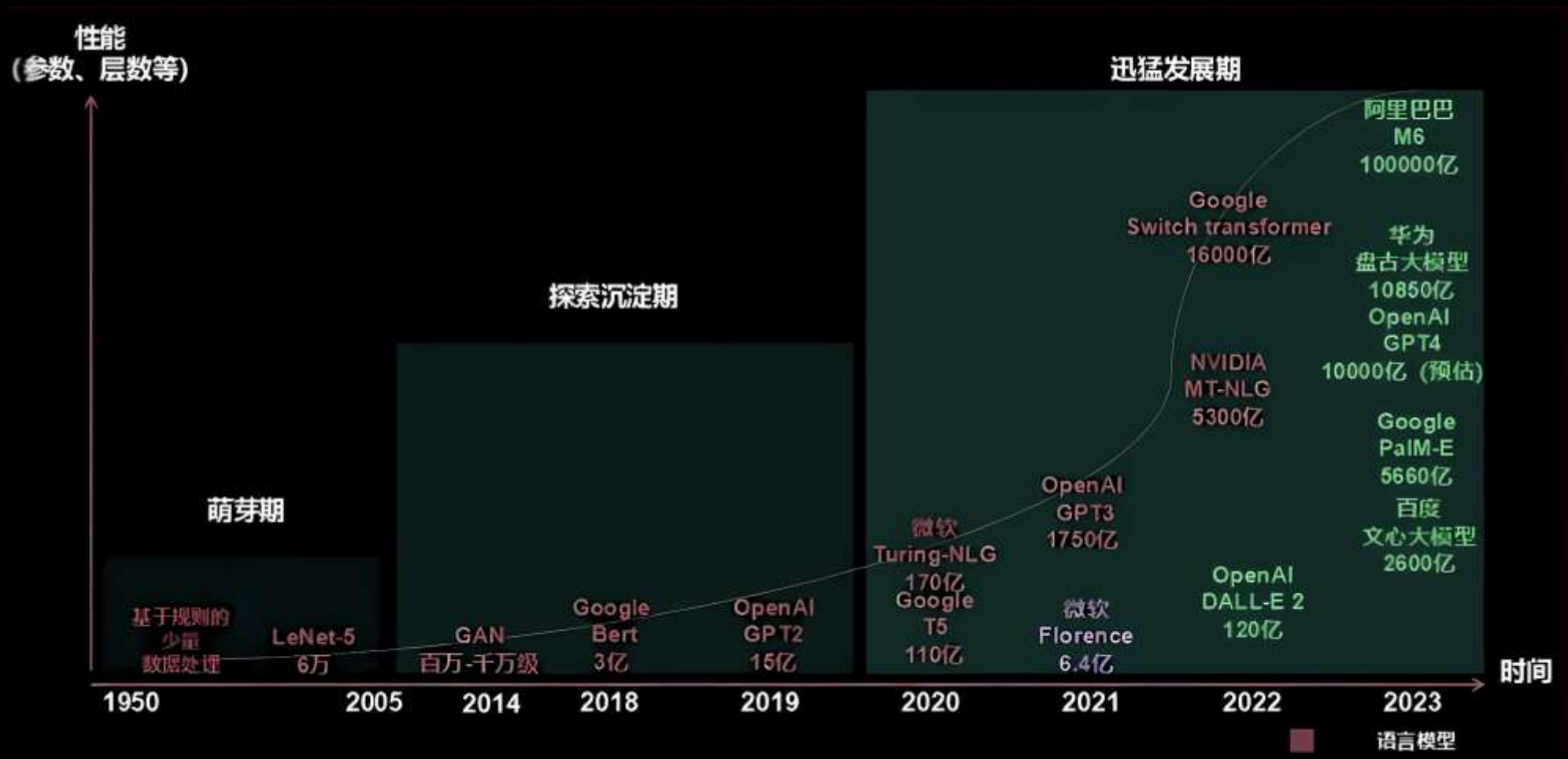


# Google TPU



# 大模型快速发展



# 英伟达的护城河

# Part 02

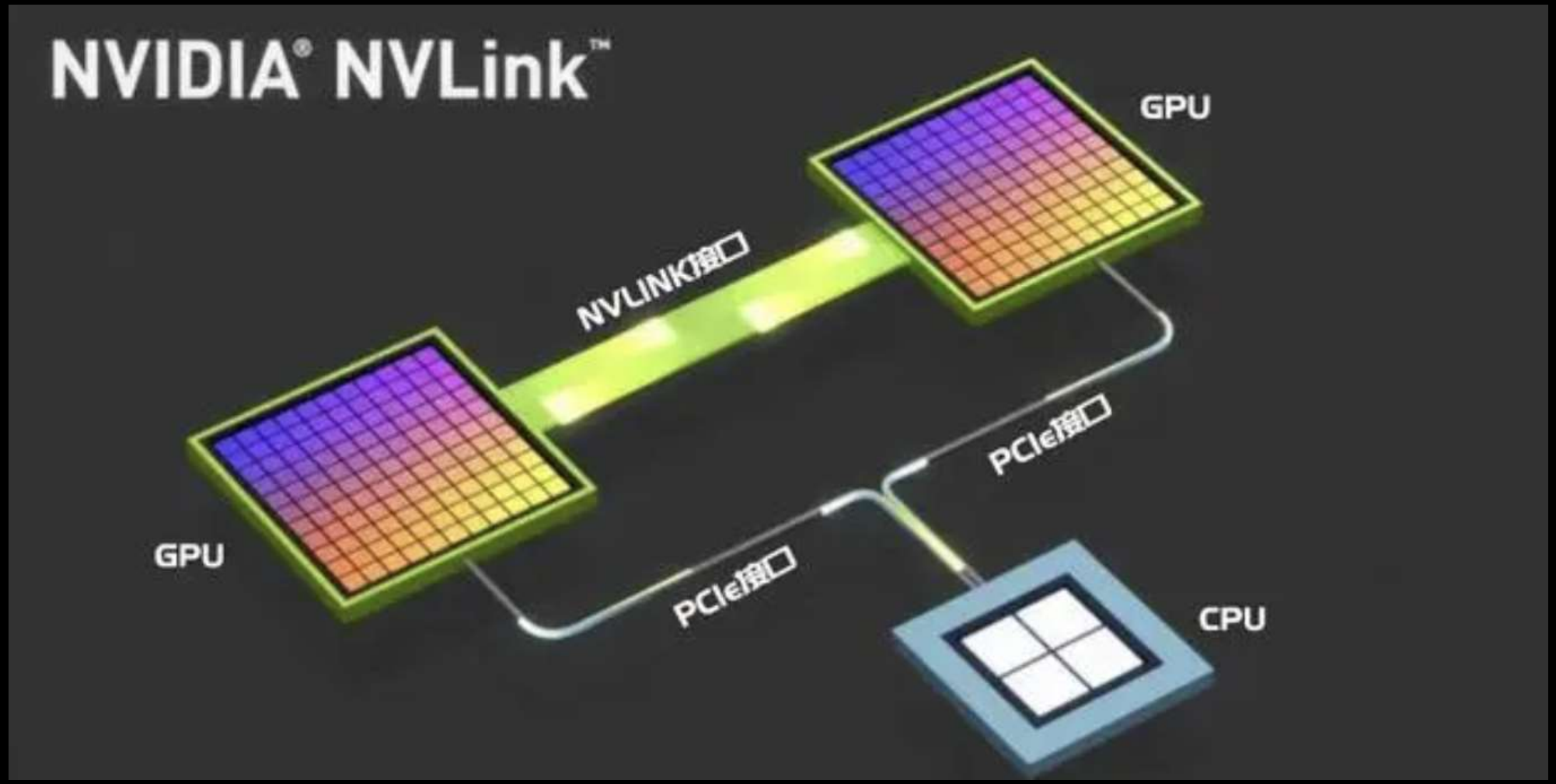
## 给 GPU 刷火箭的N种方式

Content

# 目录

# 给GPU刷火箭——单机篇

# 给GPU刷火箭：从PCIe 到 NVLINK



# PCIe NVLink性能对比

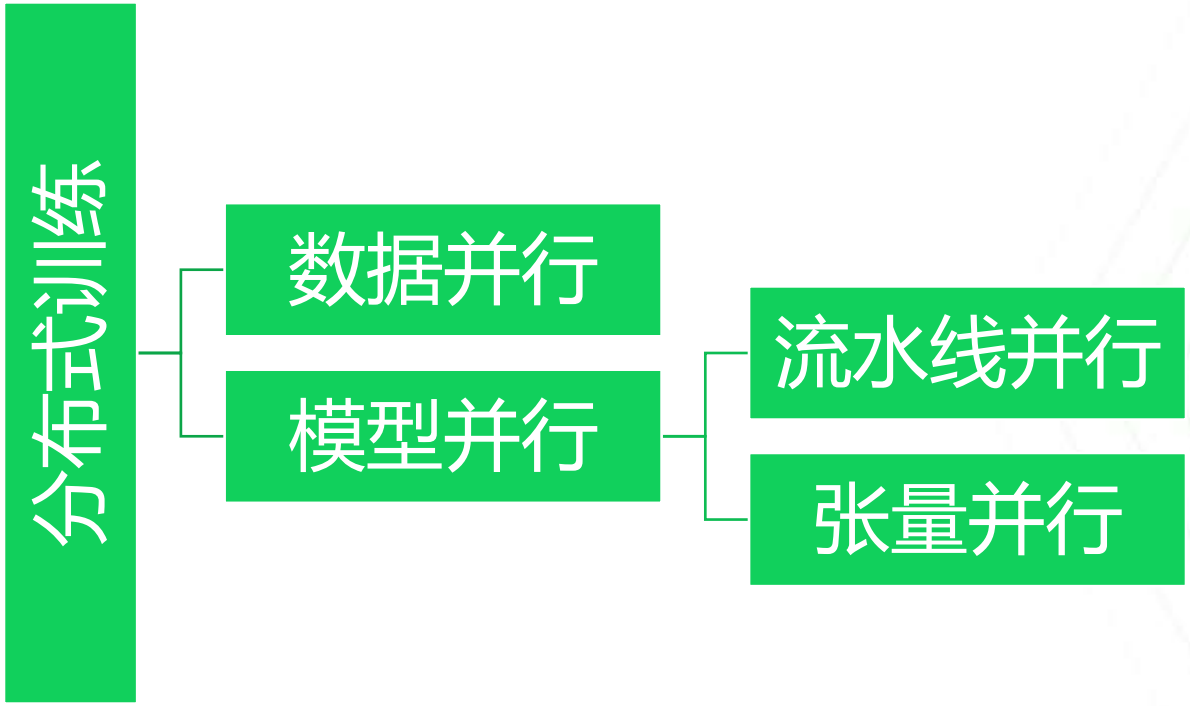
| 类型 | GPU | 通道数 | 双向互联带宽 |
| --- | --- | --- | --- |
| PCIE互联 | A100 | PCIE 4.0 x16 | 2GBx16x2=64GB/s |
| | H100 | PCIE 5.0 x16 | 4GBx16x2=128GB/s |
| NVLink互联 | A100 | 每个GPU链路Nvlink x12 | 25GBx12x2=600GB/s |
| | H100 | 每个GPU链路Nvlink x18 | 25GBx18x2=900GB/s |

# 给GPU刷火箭——智算中心集群篇

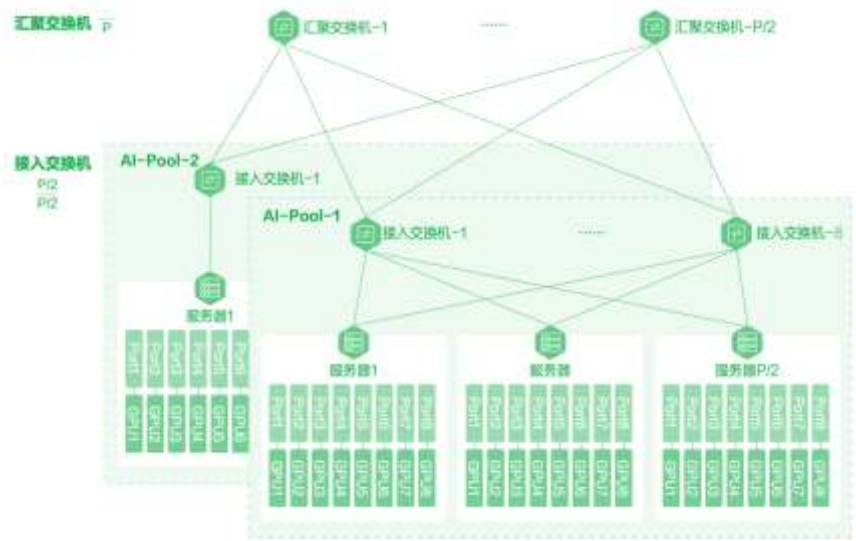
# 大模型业务全流程


数据&模型算法
模型训练&微调
模型验证&推理部署
数据处理
模型训练
模型微调
模型验证
开源数据
数据预处理
向量数据库
混合精度
梯度检查
梯度累积
全参微调
下游任务
测评标准
benchmarks
模型算法
分布式并行
低参微调
推理与智能体
LLM模型架构
多模态—切皆 tokens
训练集群稳定性
指令微调
量化压缩
推理加速
Agent 智能体
智算中心

# 为什么需要智算集群——模型分布式训练的策略



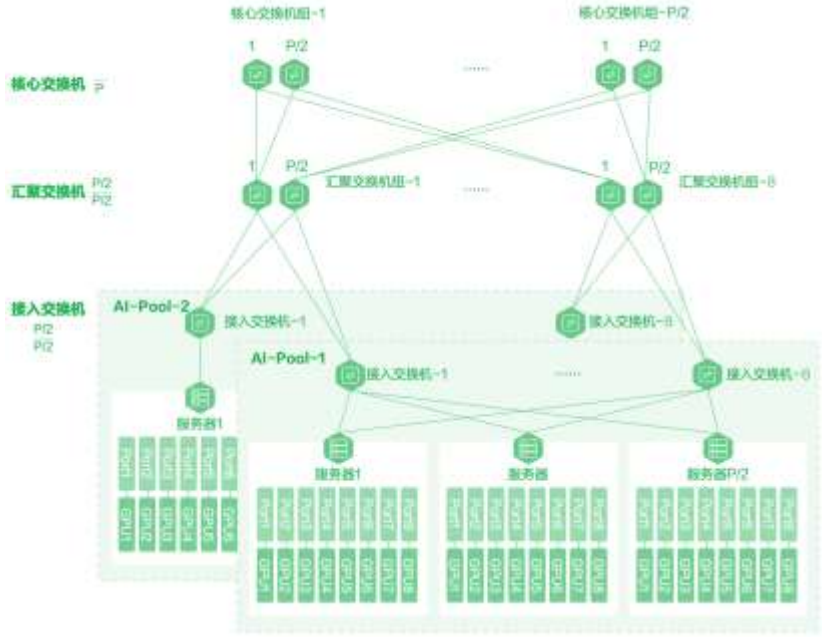
# 智算中心网络架构设计 – 千卡规模 • 万卡规模





### 网络交换设备要求

- 51.2 Tbps 交换容量
- 适用于二/三层 Clos 架构组网
- 64 口 400 GE 数据中心级非框交换机
- 支持 RoCE V2 协议
- 满足高密度服务器无收敛接入




### 千卡 规模

- 1024卡 / 128 台 8 卡 GPU 服务器
- 单机配置 8 张 400G CX 系列网卡
- 8 台 Leaf 交换机为一个组，共 32 台
- 16 台 Spine 交换机连接所有 Leaf

### 存算 分离

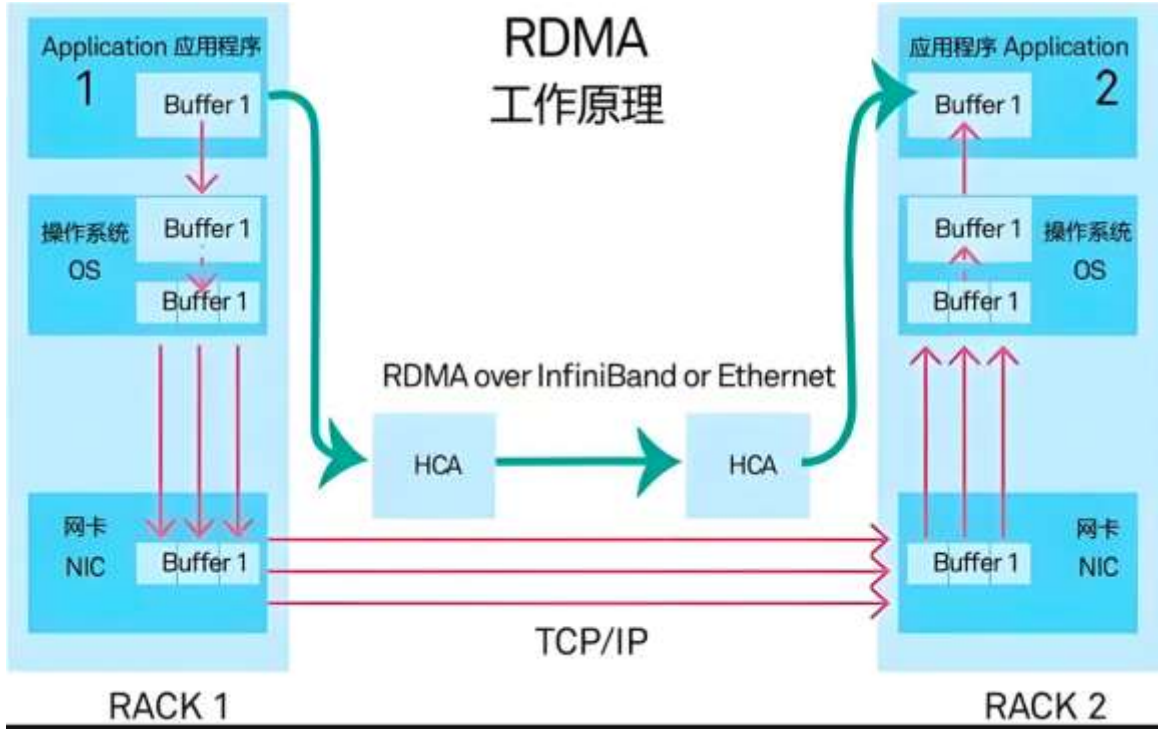
- 存储采用双口 200G CX 系列网卡
- Leaf 组 “8+1” 配置计算存储 Leaf
- 以 4:1 收敛比配置计算存储 Spine
- GPFS 分布式存储按需配置存储 Leaf
- 存储系统支持 NVIDIA GDS 方案
- 存储 Leaf 并网接入计算存储 Spine

### 万卡 规模

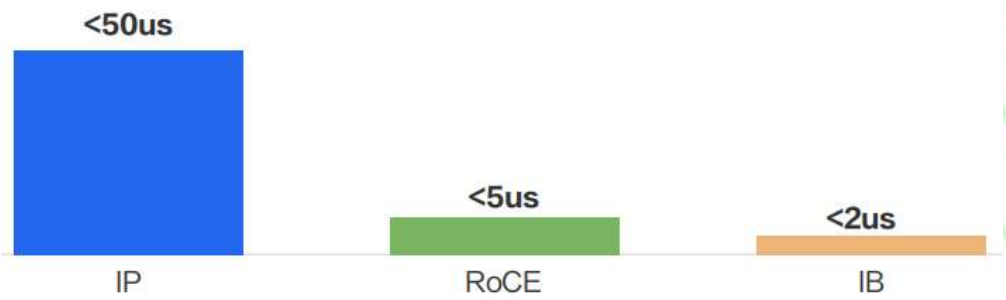
- 8192 卡 / 1024 台 8 卡 GPU 服务器
- 单机配置 8 张 400G CX 系列网卡
- 8 台 Leaf 交换机为一个组，共 256 台
- 256 台 Spine 交换机分组连接所有 Leaf
- 128台 Core 交换机分组连接所有 Spine

# 智算中心-低延时的RDMA网络

**RDMA技术**-降低多机多卡间端到端通信时延。RDMA可以绕过操作系统内核，让一台主机可以直接访问另外一台主机的内存。实现RDMA的方式有InfiniBand、RoCEv1、RoCEv2、iWARP四种。 其中RoCEv1技术已经被淘汰，iWARP 使用较少。当前智算中心的RDMA技术主要采用的方案为 InfiniBand 和 RoCEv2 两种。




IB和RoCEv2 与传统IP的端到端时延在实验室的测试数据显示，绕过内核协议栈后，应用层的端到端时延可以从 **50us（TCP/IP），降低到 5us（RoCE）或 2us（IB）**。

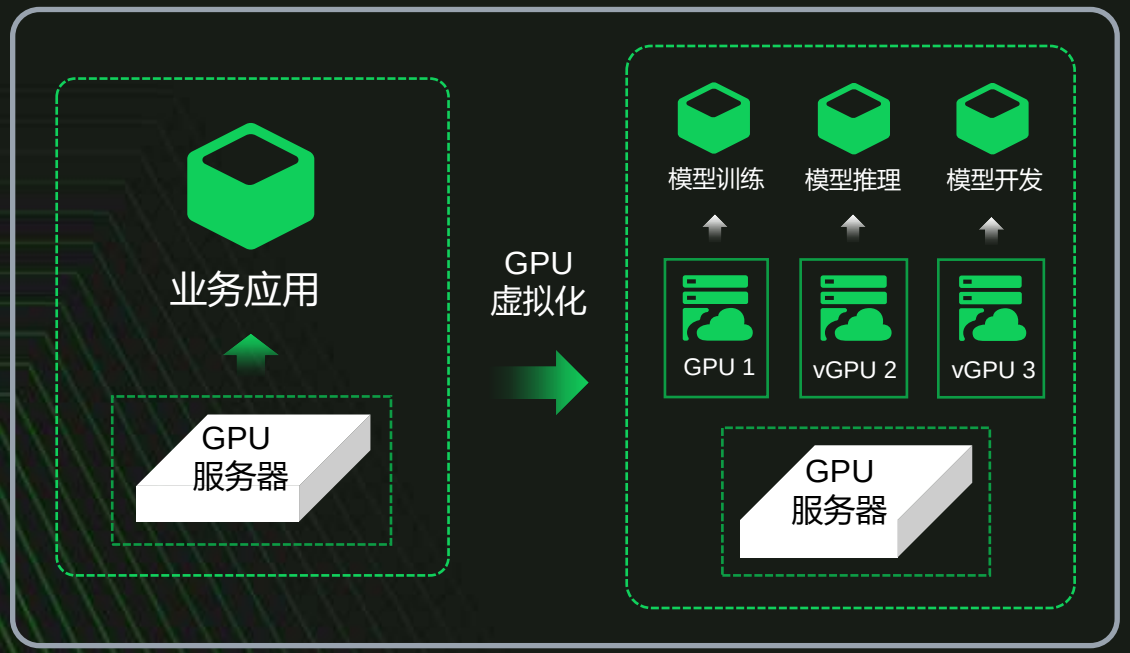


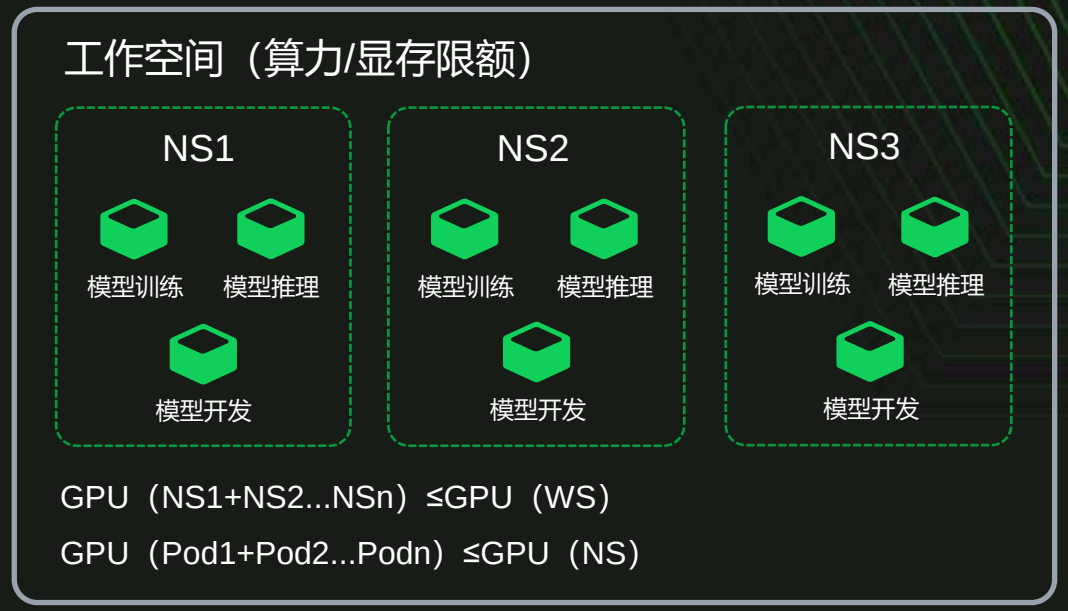

不同技术的端到端通信时延

# GPU池化共享优化

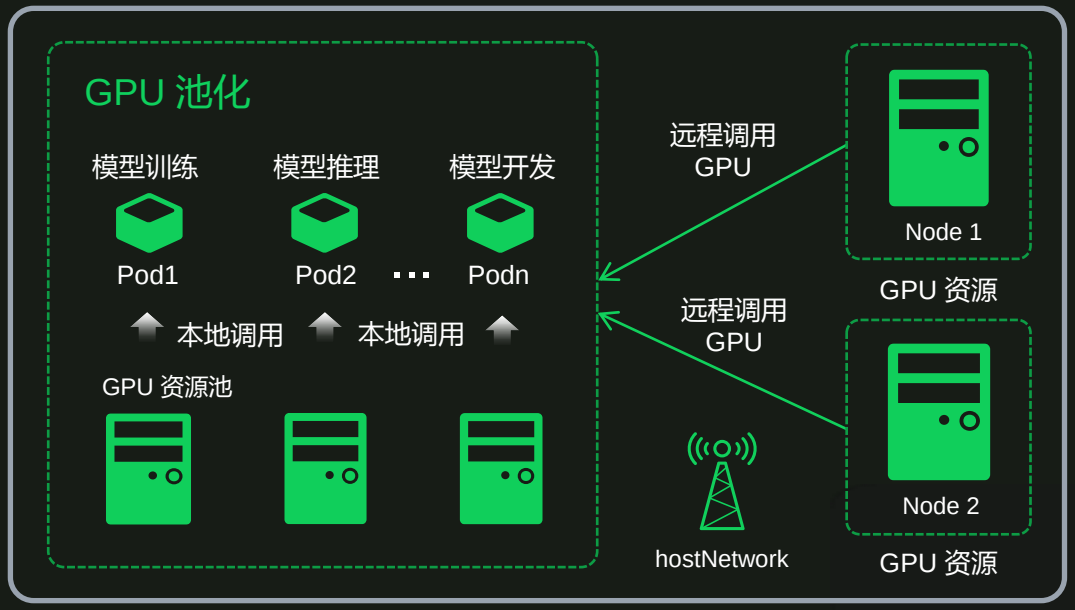
通过网络远程调用 GPU/vGPU 资源进行加速，本地无需 GPU 卡。让单个任务可以使用更多的 GPU 资源而无需关心单机的 GPU 数量。同时部署应用时可按照 1% 的算力颗粒度和 1MB 显存颗粒度极致压榨 GPU 资源，有效提高资源利用率，避免资源浪费。

## GPU 虚拟化



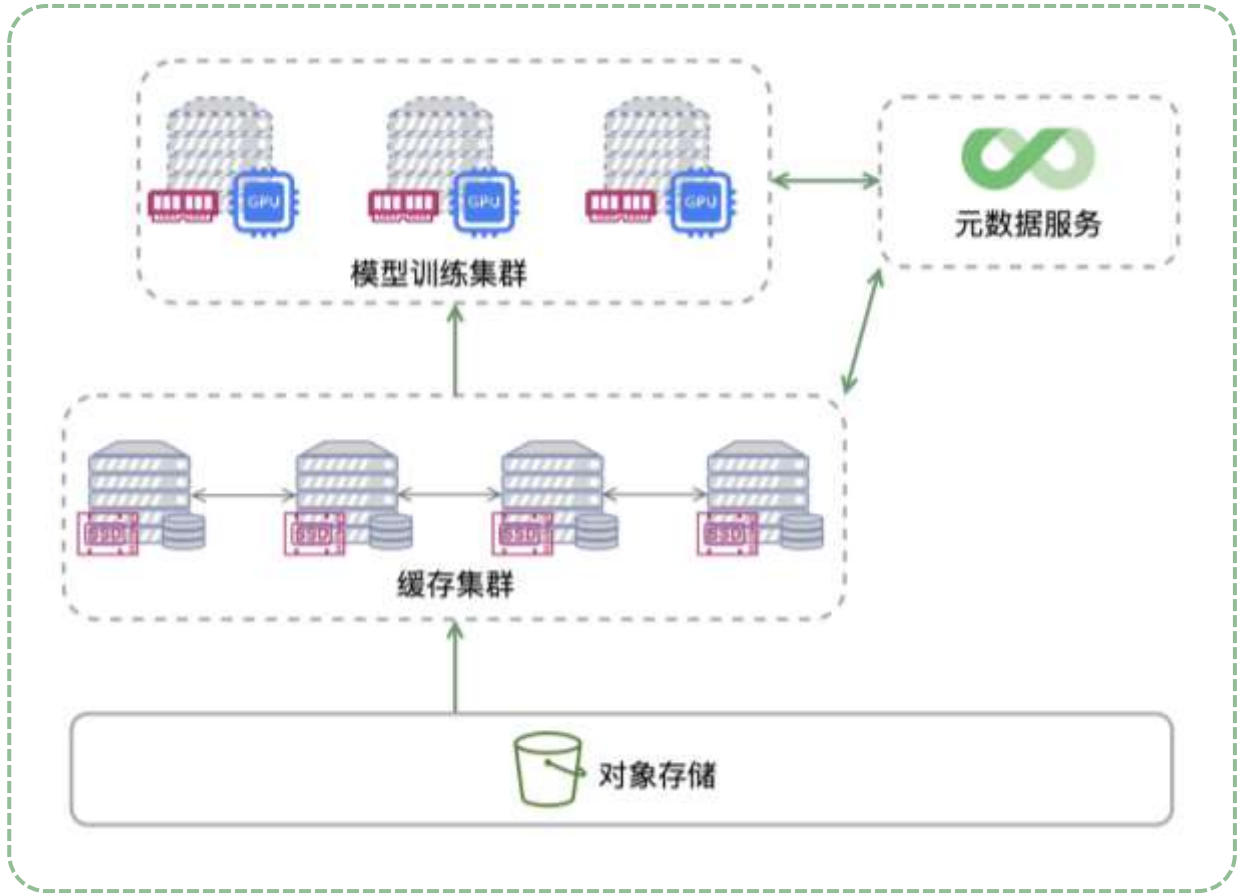

## GPU 远程调用



# 高性能近存方案——HwameiStor

HwameiStor 是一款 Kubernetes 原生的容器附加存储 (CAS) 解决方案，将 HDD、SSD 和 NvMe 磁盘形成 **本地存储资源池** 进行统一管理，使用 CSI 架构提供分布式的本地数据卷服务，为 **有状态的云原生应用** 或组件提供数据持久化能力。 HwameiStor 是一个 CNCF Sandbox 项目。

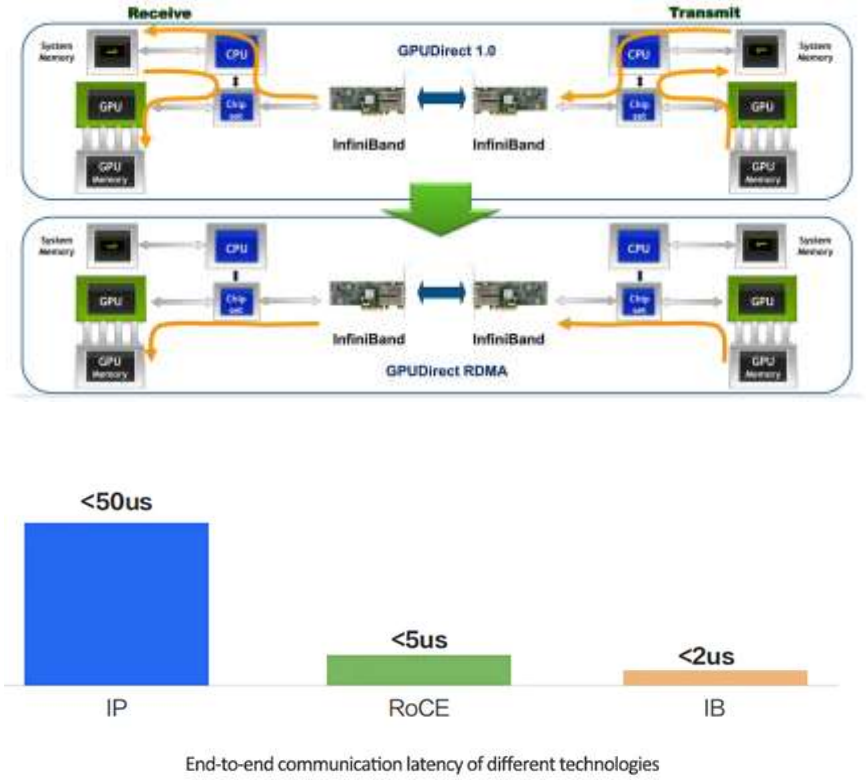
# 存储缓存调度




- **全类型数据统一存储**，非结构化数据、半结构化数据、结构化数仓统一存储
- **统一名字空间**，直接链接对象存储数据
- **多级缓存加速**：分布式缓存与本地缓存结合，最大化性能
- **数据预热管理**：按需与计划的热数据激活，灵活掌控
- **多区域数据镜像**，简化数据分发问题

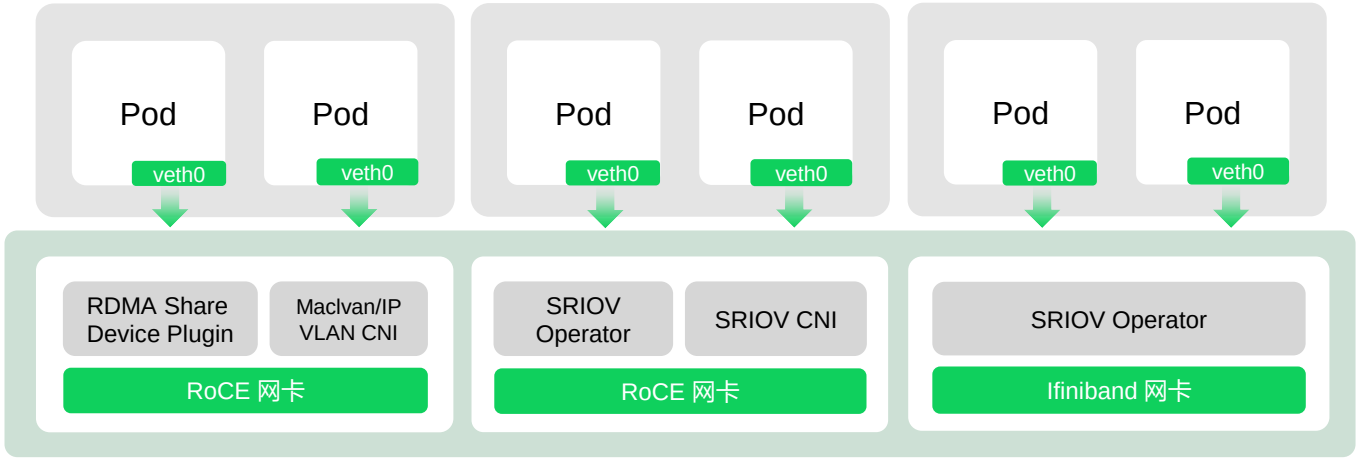
# GPU加速

GPUDirectRDMA 是一种特定于 GPU 的 RDMA 技术，允许 GPU 卡和 RDMA 网卡之间直接进行高性能的数据传输，无需通过 CPU，常应用于 HPC 和深度学习等场景，能够在数据传输过程中绕过 CPU，减少传输延迟，提高 GPU 的计算性能，尤其是在大规模数据集的训练和推理任务中。




End-to-end communication latency of different technologies

1. 基于 Macvlan/IPVLAN 使用 RoCE 网络直通
2. 基于 SR-IOV 使用 RoCE 网络直通
3. 基于 SR-IOV 使用 Infiniband 网络直通模式



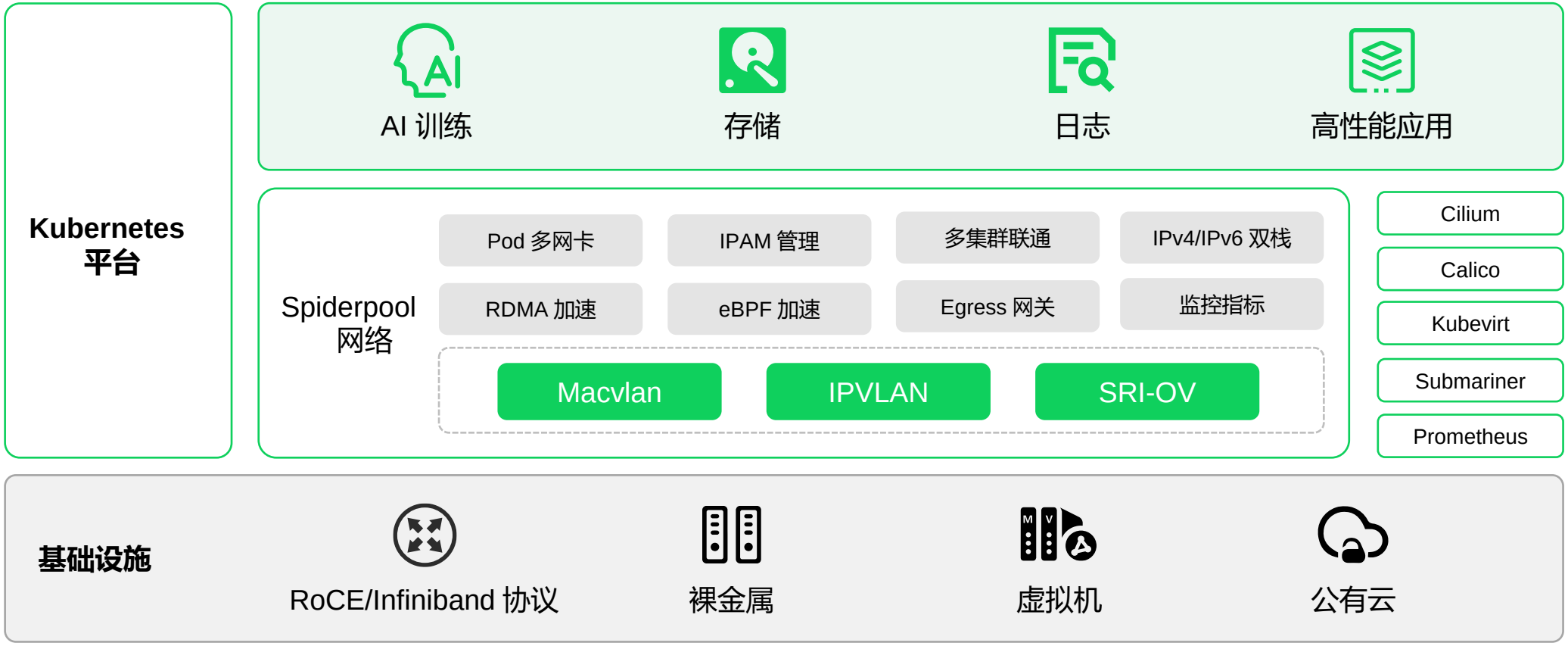

基于 Macvlan/IPVLAN 使用 RoCE 直通

基于 SRIOV 使用 RoCE 直通

基于 SRIOV 使用 Ifiniband 直通

# 网络亲和性调度

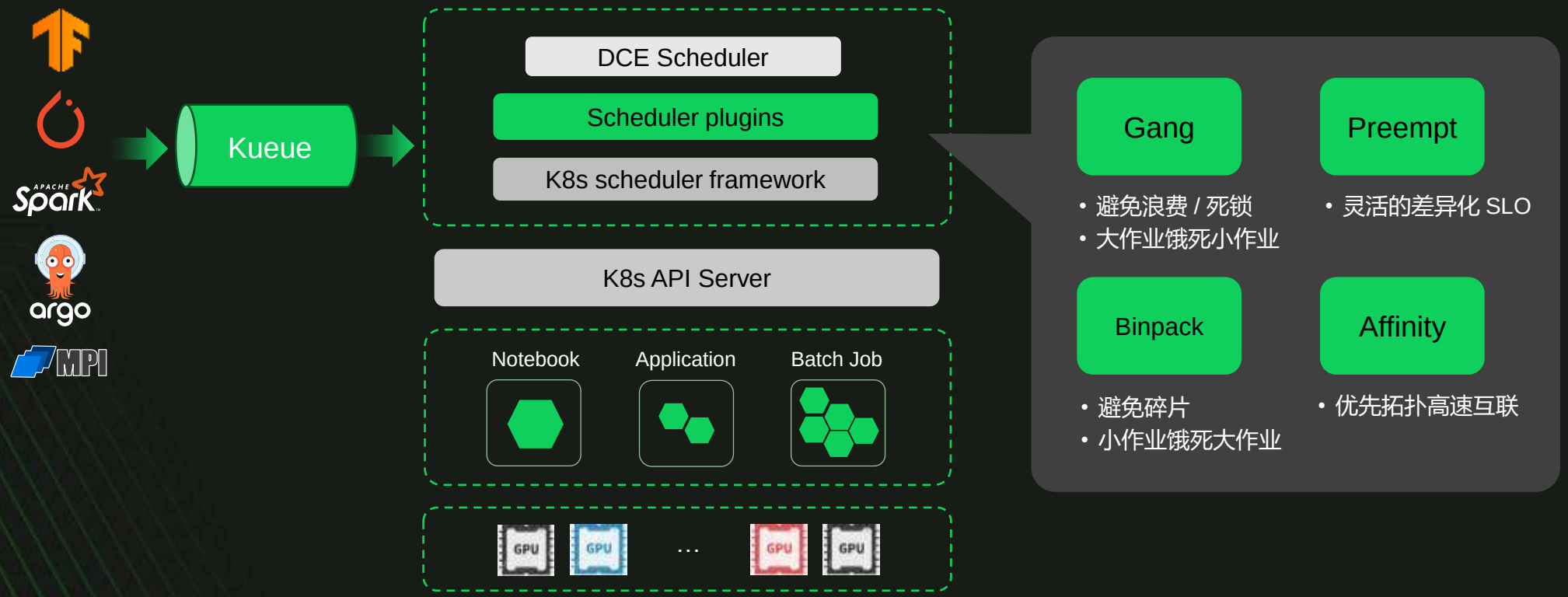
Spiderpool 是一个 Kubernetes 的 Underlay 和 RDMA 网络解决方案，它增强了 Macvlan CNI, ipvlan CNI, SR-IOV CNI 的功能，满足了各种网络需求，使得 Underlay 云原生网络方案可应用在 **裸金属、虚拟机和公有云环境** 中，可为网络 I/O 密集性、低延时应用带来优秀的网络性能，包括 **存储、中间件、AI 等应用**。 Spiderpool 是一个 CNCF Sandbox 项目。



# 算力队列调度

## 灵活的算力调度，提升 GPU 利用率

- DaoCloud 与 Google 联合发起的顶级项目 Kueue™
- 应对不同算力场景实现公平调度、亲和、组调度、紧凑等调度算法



# 模型应用套件

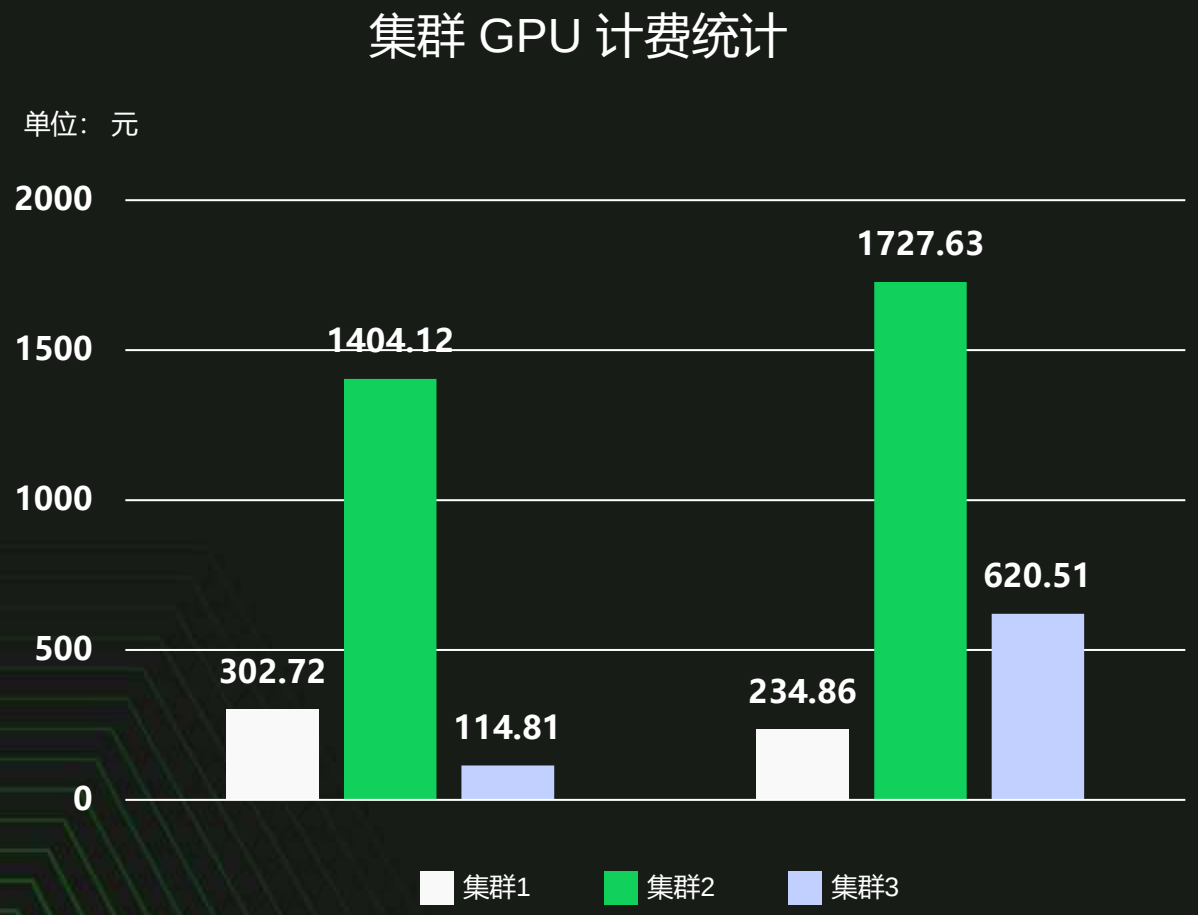

模型套件
智能应用
智能知识库
- DAK
模型流程引擎
猴子无限
模型工具
模型微调
DataTuneX
模型开发
BAIZE
分布式框架
Ray
模型训练
模型加速
Colossal
断卡重训
BAIZE
模型推理
模型推理
Triton
模型推理
vLLM
模型部署
kServe
数据管理
分布式文件缓存
JuiceFS
云原生存储
HwaimeiStor

# 资源可观测性


- **实时洞察**，内建丰富的仪表盘，直观了解 GPU 负载、内存和资源利用率，找到 GPU 使用中的瓶颈。并将模型服务指标与 GPU 指标相关联。
- 内建 **100+ 精选告警规则** 来主动识别性能和资源等问题，多种途径自动通知来低效负载，有效提高 GPU 利用率。
- 内置 **告警知识库**，知悉告警原因及后续处理方案，践行云原生运维最佳实践。
- 支持多租户自定义 **模型监控**，实现 指标采集，存储，分析，呈现全过程。

# 算力资源运营优化






支持对集群、节点、容器组、工作空间和命名空间 5 种资源类型分别进行 GPU 资源计量和计费，支持统计使用量、使用率、资源花费等内容。

支持自定义计费配置，根据不同类型自定义 GPU 单价与货币单位。可按照所选时间自动计算出集群、节点、容器组等在一段时间内的总计费用和 CPU、内存、存储、GPU 的各自使用费用。

支持通过 Excel、CSV 两种方式导出计费结果。支持计费报表中关联数据的快捷跳转，如查看同一时间段中集群下的节点计费等。

# Part 03

## 总结回顾

# 总结

给GPU刷火箭的方案概览

1、节点层面；

2、集群层面；

3、运营层面；

# Q&A

# Thanks.