

学习 笔记

GPU 卡间互联 技术大起底

PCIe · NVLink · Infinity Fabric · CXL · UALink ·
光互连

作者：两年砍柴

公众号：两年砍柴



原文封面图 · 制图人：两年砍柴

从「超级工厂」到「超级城市群」

- ◆ 单张 GPU 是一个「超级工厂」
- ◆ 多张 GPU 集群如「超级城市群」
- ◆ 互联 = 工厂之间的高速公路网
- ◆ 互联决定整体集群效率

⚡ 核心判断

选对互联方案，和选对 GPU 型号一样重要。

01 • GPU互联在整机系统中的三层位置



第一层： CPU↔GPU PCIe Gen5 x16 ≈ 128 GB/s	CPU通过PCIe总线连接GPU，负责数据加载、模型下发、监控管理。带宽最低，仅适合任务调度和少量数据搬运。
第二层：GPU↔GPU（节点内） NVLink 5 ≈ 1,800 GB/s · xGMI ≈ 896 GB/s	本讲核心：同一服务器内GPU直连，统一地址空间，张量并行/流水线并行在此层完成。
第三层：节点↔节点（集群间） InfiniBand NDR ≈ 400 Gb/s · RoCE	跨服务器通过以太网/IB网络连接，数据并行梯度同步在此层。带宽低一个数量级，延迟更高，跨节点通信应尽量压缩。
核心判断：第二层（节点内GPU互联）带宽是第一层的14倍、第三层的36倍——大模型训练的性能胜负手就在这一层	

GPU卡间互联技术详解 - 01 互联背景与全景

图：GPU 互联在整机系统的三层位置

带宽

一次能传多少数据

GB/s · TB/s · 决定吞吐

延迟

一来一回要多久

纳秒级 → 微秒级

拓扑

数据走哪条路径

直连 GPU↔GPU / 经 CPU 中转

不同场景的关注点

- ◆ 数据并行训练：主要看梯度同步效率
- ◆ 张量并行 / 流水并行：非常依赖低延迟高带宽
- ◆ 大模型推理：关注跨卡激活值、KV Cache 传输
- ◆ HPC：通信性能往往直接决定整体吞吐

通用、便宜、兼容性极强

- ◆ 几乎所有 GPU / CPU / SSD / 网卡都走 PCIe
- ◆ PCIe 4.0 x16 $\approx$ 64 GB/s
- ◆ PCIe 5.0 x16 $\approx$ 128 GB/s
- ◆ PCIe 6.0 x16 $\approx$ 256 GB/s
- ◆ 哪里都能去，什么车都能跑

⚡ 关键判断

带宽比 NVLink 5 慢 14 倍，且 PCIe 非全互联

02 • PCIe 基线：通用总线够用吗？



PCIe 各代 x16 双向带宽

代际	x16双向带宽	典型应用
Gen3 (2010)	~32 GB/s	~2020年前GPU
Gen4 (2017)	~64 GB/s	A100/H100底座
Gen5 (2022)	~128 GB/s	B200/GH200底座
Gen6 (2027?)	~256 GB/s	规范草案中

注：Gen6 规划中（2027预计），仍约256 GB/s，与NVLink5差距仍达7倍

与GPU专用互联的差距

PCIe Gen5 x16
128
GB/s

vs

NVLink 5 / GPU
1,800
GB/s

14倍带宽差距
且PCIe为树形/星形拓扑，非全互联

PCIe在GPU系统中的正确角色

[正确角色]

不是GPU间数据通路——PCIe用于CPU-GPU数据加载、模型下发、NVLink配置管理。把训练流量跑在PCIe上会成为瓶颈。

[为何不可替代]

专用互联不可替代——NVLink/xGMI用私有SerDes，无PCIe协议开销、支持点对点全互联和原子操作，通用总线做不到。

来源：PCIe-SIG Gen5/Gen6 规格书 · NVIDIA H100/B200 Datasheet · 带宽为x16双向

GPU卡间互联技术详解 · 02 PCIe与传统网络基线

图：PCIe 各代 x16 双向带宽

✓ 优势

- ◆ 标准开放、兼容性极强
- ◆ 成本很低
- ◆ 可混合不同厂商 GPU / 加速卡 / 存储
- ◆ 灵活度高

✗ 劣势

- ◆ 延迟较高（微秒级）
- ◆ 多 GPU 通信常需 CPU/Switch 中转
- ◆ 跨 Socket 通信性能可下降 30%-50%
- ◆ AI 大模型带宽需求指数级增长

⚡ 关键判断

PCIe 仅仅是「能用的方案」，并不是「好用的方案」。
通用且便宜，但带宽与延迟成为大模型训练瓶颈。

2014 年推出 · 专有高速互联

- ◆ 极宽专用物理通道
- ◆ 把多张 GPU 焊成「超级巨无霸」
- ◆ 支持缓存一致性
- ◆ 多 GPU 共享统一内存空间

三层架构

- ◆ 物理层：SerDes · PAM4
- ◆ 链路层：包聚合 · CRC 校验
- ◆ 语义层：load/store · 内存语义

03 · NVLink协议原理：物理层到语义层



物理层 SerDes

每link一对差分线，PAM4编码。NVLink4单lane 50 GT/s，NVLink5达100 GT/s。每GPU 18个link (NVL5)，双向聚合1,800 GB/s。NVLink6翻倍至36 link。

链路层 包聚合

多link bond为一条逻辑通道，硬件重传、CRC校验、流量控制。无PCIe TLP/ECRC开销，协议栈更薄，延迟比PCIe低一个数量级。

语义层 内存操作

支持load/store（直接读远端显存）、atomic（原子加/比较交换）、DMA（大块拷贝）、一致性（NVLink-C2C时cache-coherent）。GPU看到的是统一物理地址空间。

与PCIe的本质区别：PCIe是I/O协议（DMA描述符+TLP），NVLink是内存语义（load/store直访）
前者GPU要主动发DMA请求，后者CPU/GPU可像访问本地内存一样访问远端显存——这是张量并行能跑起来的基础

GPU卡间互联技术详解 · 03 NVIDIA NVLink阵营

图：NVLink 协议原理 · 物理层 / 链路 / 语义层

NVLink 1.0

2016 · Pascal

160 GB/s

NVLink 3.0

2020 · A100

600 GB/s

NVLink 4.0

2022 · H100

900 GB/s

NVLink 5.0

2024 · B200

1.8 TB/s

⚡ 1.8 TB/s 是什么概念？

约为 PCIe 5.0 x16 带宽（128 GB/s）的 14 倍以上。



图：NVLink 代际对照（NVLink 3 600 → NVLink 4 900 → NVLink 5 1,800 → NVLink 6 3,600 GB/s）

「交通枢纽」

- ◆ 8 卡 mesh 升级为全互联 crossbar
- ◆ 所有 GPU 接同一 Switch 矩阵
- ◆ 任意两卡一跳直达
- ◆ 第四代 NVSwitch：144 端口
- ◆ 无阻塞交换容量 14.4 TB/s

NVL72 机柜拓扑

72 GPU + 36 Grace CPU，整柜聚合 130 TB/s

03 • NVSwitch：把点对点变成全互联



无Switch：点对点mesh的局限

- 8卡直连：每张GPU只有少量邻居link
- 远距离GPU对需多跳中继，延迟翻倍
- 规模上限：8卡后link数不够连满
- AMD xGMI走这条路线（见第四章）

NVSwitch：全互联crossbar (NVIDIA路线)

- 所有GPU接到同一Switch矩阵，任意两卡一跳直达
- NVSwitch 5 (NVL72): 9个tray级Switch
- 整机聚合带宽: 130 TB/s (NVL72)
- NVL576: 多Switch级联，域扩大到576 GPU
- 代价: Switch芯片贵、功耗高、整柜液冷

NVL72 机柜拓扑

- 72个Blackwell GPU + 36个Grace CPU，装在一个液冷机柜内
- 每GPU 18条NVLink5上联到NVSwitch矩阵，任意两GPU一跳直达
- 整柜130 TB/s双向聚合带宽，对外通过NVLink光纤连到下一个NVL72
- 这就是NVIDIA的“机架即服务器”理念：72 GPU作为一台超大机器运行

GPU卡间互联技术详解 · 03 NVIDIA NVLink阵营

图：无 Switch mesh vs NVSwitch 全互联

从 chiplet 到 GPU 互联

- ◆ 起源：chiplet 设计的必然产物
- ◆ die 间高速总线互连
- ◆ 设计哲学：die 直连 mesh
- ◆ 对外接口 xGMI（扩展 GMI）
- ◆ $xGMI \approx 896 \text{ GB/s}$
- ◆ 成本 / 功耗优于 NVLink

⚡ AMD vs NVIDIA

mesh 直连 vs 全互联 Switch；MI350X 约 896 GB/s

04 • AMD Infinity Fabric: 从chiplet到GPU互联



起源：chiplet设计的必然产物

与NVLink路线的根本差异

- AMD把CPU/GPU拆成多个小die (chiplet)
- die间必须用高速互连连起来——这就是Infinity Fabric(IF)
- IF最初在CPU多die间使用 (EPYC Rome/Zen)
- 延伸到GPU：CDNA架构 (MI300系列)
- 对外GPU间接口叫**xGMI (扩展GMI)**
- 设计哲学：die间直连mesh，不做crossbar Switch

- 拓扑：**mesh直连 vs NVSwitch crossbar**
- 不需要额外Switch芯片——成本更低
- 远距离GPU对需多跳，延迟比Switch一跳高
- 单GPU带宽略低：MI350X约896 GB/s vs NVLink5 1800 GB/s
- 但：每卡成本/功耗更优，生态开放
- UALink联盟核心贡献者——IF技术外溢到开放标准

一句话定位：NVIDIA用Switch买“全互联一跳直达”，AMD用mesh买“省掉Switch的成本与功耗”。两条路线在MI300X/MI350X和NVL72/NVL576上正面交锋，UALink则试图把mesh路线标准化给所有厂商。

GPU卡间互联技术详解 · 04 AMD Infinity Fabric阵营

图：mesh 直连 vs NVSwitch crossbar

云厂商联盟 · 打破 NVLink 专有锁定

- ◆ 2024-10 成立，2025-04 发布 1.0 规范
- ◆ 加速器厂商：AMD / Intel / Google TPU / Meta
- ◆ NVIDIA 不在联盟内
- ◆ 支持 1,024 加速器互联
- ◆ 单通道 200 GT/s，延迟 < 1 微秒



本质

地缘 + 商业博弈，云厂不愿被「第二来源」卡死

05 · UALink为什么出现：云厂商的联盟



两年砍柴

2024年10月成立，2025年4月发布1.0规范

加速器厂商

- AMD (技术基础)
- Intel
- Google TPU
- Meta

芯片/网络厂商

- Broadcom
- Marvell
- HPE
- NVIDIA不在联盟内

根本诉求

- 打破NVIDIA NVLink专有锁定
- 多厂商加速器可混插互联
- 自研ASIC (Google TPU等) 有标准对外接口
- 用以太网SerDes降低成本

本质：这是一场地缘+商业博弈

- 云厂商是最大买家，受单一供应商制约——议价权、供应链、地缘风险都需要“第二来源”
- Google/Meta已有自研ASIC，需要一个不依赖NVIDIA的互联标准把自家芯片和外购芯片连起来
- 历史类比：以太网之于专有FDDI、CXL之于专有MPT——开放标准最终赢

GPU卡间互联技术详解 · 05 UALink开放阵营

图：UALink 为什么出现 · 云厂联盟

Compute Express Link

- ◆ 基于 PCIe 的开放互联标准
- ◆ 主要解决内存扩展与共享
- ◆ CPU / GPU / 加速器可访问外部内存池
- ◆ 接近本地内存的访问延迟
- ◆ 延迟比 NVLink 高 50-100 ns
- ◆ 广泛的行业支持

⚡ 开放阵营

CXL 内存语义 + UALink 加速器互联 = 「开放对抗封闭」

CXL vs UALink

- ◆ CXL：以内存语义为主
 - → 解决 CPU/GPU/ 加速器内存池
 - → 底层走 PCIe 物理层
 - → 延迟略高于 NVLink
- ◆ UALink：以加速器互联为主
 - → 直接互联 1024 卡
 - → 单通道 200 GT/s
 - → 功耗为以太网 1/3-1/2
- ◆ 共同点：都是开放标准

下一代 GPU 互联方向

- ◆ 互联带宽需求将超 1 TB/s
- ◆ 端口速率将达到 224 Gbps
- ◆ 铜缆电信号 3 米内严重衰减
- ◆ 光互连从实验室走向产业化
- ◆ 壁仞 NPO（近封装光学）方案
- ◆ 支持单超节点 1,024 卡规模化扩展
- ◆ 英伟达积极布局硅光技术

光互连演进路径

① 带宽需求爆炸

GB/s → TB/s → 多 TB/s

② 铜缆达到物理上限

电信号 3 米衰减严重

③ 光互连登场

NPO · 硅光 · 1,024 卡超节点

方 案	类 比	关 键 词
PCIe	村村通公路	通用 · 便宜 · 兼容
NVLink	英伟达专属高铁	性能天花板 · 生态封闭
Infinity Fabric	AMD 挑战者高铁	mesh 直连 · 成本更优
CXL / UALink	开源高铁	开放生态 · 标准探索
光互连	未来磁悬浮	下一代 · 实验室 → 产业化

⚡ 选择建议

本质是「性能 · 成本 · 开放性」三者权衡 · 选对互联 ≈ 选对 GPU

谢谢观看

主要内容回顾

PCIe（村村通） • NVLink（英伟达专属高铁） • Infinity Fabric（挑战者）
CXL / UALink（开放生态） • 光互连（未来磁悬浮）