GPU 集群互联解决方案

EZMAX GPU Cluster Interconnect Solution

北京万邦迪通科技有限公司

EZMAX 品牌 | 算力中心网络基础设施提供商

目 录

一、方案背景与需求分析

1.1 为什么 GPU 集群需要专用互联网络?

现代 AI 训练和 HPC 高性能计算场景中，GPU 之间的数据交换量远超传统通用计算场景。千亿参数大模型一次训练迭代可能触发 PB 级的梯度同步流量，若网络性能不足，GPU 将长时间处于等待数据的状态，资源利用率急剧下降。

1.2 核心互联需求分层

需求层次	具体指标	EZMAX 对应方案
带宽需求	25G ~ 100G 单链路	NETI710 系列网卡 + 25G/100G 光模块
延迟需求	端到端 < 5μs（RDMA）	芯片硬件卸载 RoCEv2
传输效率	零拷贝，避免 CPU 介入	GPUDirect RDMA + SR-IOV
网络规模	支持 256+ 节点横向扩展	Spine-Leaf 可扩展架构
运维效率	统一管控、故障快速定位	EZMAX 兼容性中心 + 技术支持

1.3 典型业务场景对网络的诉求

业务场景	流量特征	核心网络诉求
大模型训练（LLM）	大量跨 GPU 梯度同步，南北向+东西向并重	高带宽、低延迟、无丢包
分布式推理	请求级实时调度，多节点协同推理	快速收敛、高密度端口
HPC 科学计算	MPI 全互连通信，强耦合计算节点	超低延迟、全对分带宽
分布式存储	大块顺序读写，数据修复跨节点	高吞吐、大象流承载
图神经网络训练	节点邻居聚合，通信模式动态变化	多路径 ECMP、无阻塞网络

二、RDMA & GPUDirect 核心技术解读

2.1 RDMA 技术原理

RDMA（Remote Direct Memory Access，远程直接内存访问）是一种绕过操作系统内核、直接在两台服务器内存之间传输数据的技术。相比传统 TCP/IP 协议栈，RDMA 可将端到端延迟从 50~100μs 降低至 3~5μs，同时将 CPU 占用率降低 90% 以上。

对比维度	传统 TCP/IP	RDMA (RoCEv2)
协议栈开销	内核协议栈处理，数据需多次拷贝	内核旁路，零拷贝直接内存访问
端到端延迟	50~100 μs	3~5 μs（降低 90%+）
CPU 占用	5%~15% CPU 参与网络处理	< 1%，几乎零 CPU 占用
吞吐量	受限于 CPU 性能	硬件线速，接近物理极限
适用场景	通用计算、存储备份	AI 训练、HPC、分布式存储

2.2 RoCEv2 无损网络实现

RoCEv2（RDMA over Converged Ethernet v2）是 EZMAX 方案采用的核心 RDMA 协议，运行在标准以太网上，无需专用 InfiniBand 交换机，大幅降低采购与运维成本。RoCEv2 的关键使能技术包括：

- **PFC（Priority Flow Control）**基于优先级的流量控制，智算中心通常开启 priority 3，以太链路零丢包
- **ECN（Explicit Congestion Notification）**显式拥塞通知，智能感知网络拥塞，提前通知发送端降速而非丢包
- **DCQCN（Data Center Quantized Congestion Notification）**数据中心量化拥塞通知，RoCEv2 标准配置算法
- **无损队列配置**交换机需配置 ECN 阈值、PFC 水线，通常建议 ETS 带宽分配给 RDMA 流量专用优先级

2.3 GPUDirect RDMA

GPUDirect RDMA 是 NVIDIA 提出的技术标准，允许 GPU 直接与 RDMA 网卡进行数据交换，绕过 CPU 和系统内存，实现真正的零拷贝数据传输。EZMAX NETI710 系列网卡经过完整驱动适配，全面支持 GPUDirect RDMA。

▶ GPUDirect RDMA 工作原理

Step 1：GPU A 通过 NVLink 将数据写入自身 HBM 显存

Step 2：CUDA 驱动调用 NVIDIA Peer-to-Peer (P2P) API

Step 3：NETI710 网卡通过 PCIe 直接读取 GPU HBM，绕过系统内存

Step 4：网卡硬件将数据封装为 RoCEv2 报文，发送至 GPU B

Step 5：GPU B 网卡接收后直接写入本地 GPU HBM，无需 CPU 介入

关键收益：端到端路径减少 50%+，带宽利用率提升 30%+，训练效率提升 15%~25%

2.4 SR-IOV 与虚拟化支持

EZMAX NETI710 系列支持 SR-IOV（Single Root I/O Virtualization），单张物理网卡可虚拟出最多 64 个 VF（Virtual Function），每个 VF 具备独立的数据路径，适用于 Kubernetes 裸金属 GPU 调度、虚拟化 GPU 分片、容器化 AI 推理等场景。

- 支持 VMware vSphere、Linux KVM、Microsoft Hyper-V 等主流虚拟化平台；
- VF 级别的独立带宽保证，每个 VF 可配置 QoS 策略；
- 与 GPUDirect RDMA 协同，支持 VF 直通（VFIO）场景下的 GPU-NIC 协同加速。

三、网络架构设计

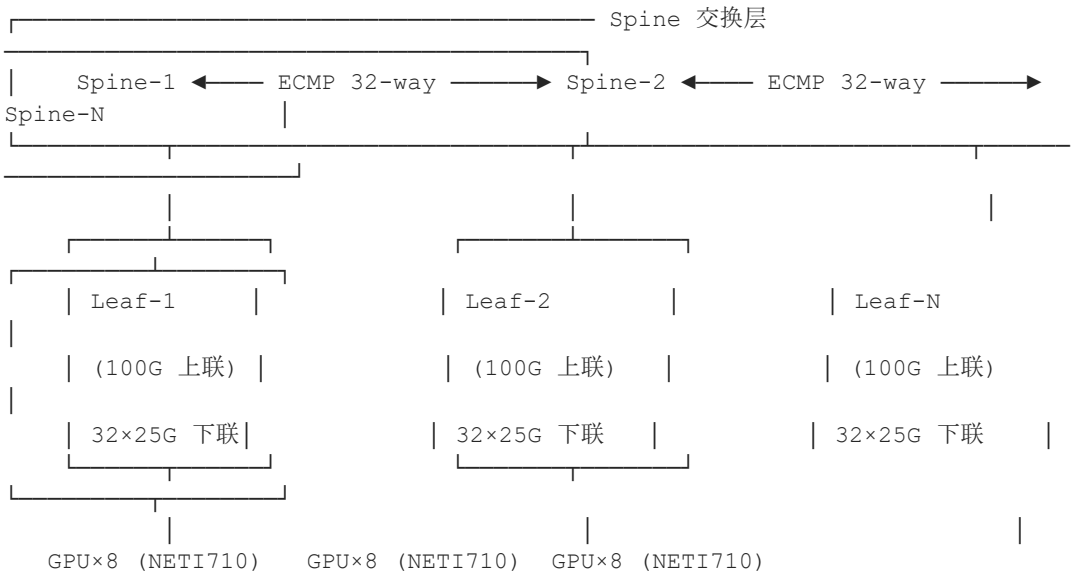
3.1 架构选型：胖树（Fat-Tree）拓扑

GPU 集群互联推荐采用经典的两层 Fat-Tree（胖树）拓扑，又称 Spine-Leaf 架构。该拓扑具有以下优势：

- 任意两个 GPU 节点之间的路径数固定（2 条），保证对分带宽一致；
- 水平扩展：增加 Leaf 或 Spine 交换机即可平滑扩展集群规模；
- 高可用：任意单设备故障不影响跨节点通信；

- 运维友好：布线规范、故障域小。

3.2 两层 Fat-Tree 架构详解



说明：每台 GPU 服务器配备 1~2 张 EZMAX NETI710-2CP / 4CP 网卡，
经 25G SFP28 DAC 直连到所在机柜的 Leaf 交换机。

3.3 关键网络参数配置建议

配置项	推荐值	说明
Spine 交换机型号	支持 100G/400G 端口的主流交换机	如 Cisco Nexus 9000 / Huawei CloudEngine 系列
Leaf 上联带宽	收敛比 4:1（32 × 25G 下行 / 4 × 100G 上行）	根据业务量可调整至 2:1
ECMP 路径数	≥ 4 路	提升带宽利用率 + 链路冗余
MTU	9216 Bytes（Jumbo Frame）	大块数据传输效率最优
PFC 优先级	Priority 3 启用 PFC	建议 RDMA 流量独占该优先级
ECN 阈值	交换机 ECN WRED 阈值动态调整	参考 NIC 端配置，建议 70%~80% 队列深度
RoCE 版本	RoCEv2（强制）	三层可路由，支持跨 POD 部署
拥塞控制算法	DCQCN（发送端+接收端协同）	ECN 触发，适配高密度 GPU 集群

四、端到端产品配置清单

EZMAX 提供 GPU 集群互联全栈产品，从服务器网卡到机柜布线全覆盖，一站式交付，零兼容风险。

产品类别	推荐型号	规格	在集群中的角色
服务器网卡	NETI710-2CP	10G 双口 SFP+，PCIe 3.0 x4，支持 RoCEv2	GPU 服务器接入网卡（管理/存储网络）
服务器网卡	NETI710-4CP	10G 四口 SFP+，PCIe 3.0 x8，支持 SR-IOV 64VF	高密度 GPU 节点（AI 训练节点首选）
高速光模块	25G SFP28	25GBASE-SR，100m OM4，低功耗 < 1W	Leaf-Server 互联（中距离）
高速铜缆 DAC	25G SFP28 DAC	直连铜缆，1~3m，首选机柜内短距	Leaf-Server 互联（首选低成本方案）
高速光模块	100G QSFP28	100GBASE-SR4，100m OM4（规划中）	Spine-Leaf 骨干互联
高密度布线	MPO-12 光跳线	OM4/MPO-12，0.5~10m，支持并行光传输	机柜内高密度汇聚
高密度布线	MPO-24 光跳线	OM4/MPO-24，高密度，支持 100G × 4 路并行	Leaf 到 Spine 互联
网络控制器	芯片系列	自主研发，支持 RoCEv2 RDMA，硬件卸载	核心芯片

4.1 典型集群规模配置示例（64-GPU 集群）

组件	数量	规格	带宽
GPU 服务器	8 台 × 8 GPU = 64 GPU	每台 8 × NVIDIA A100/H100	每台 200G（8 × 25G）
NETI710-4CP	8 张（每台 1 张）	10G 四口 → 4 × 25G 端口	每卡 40G 汇聚
Leaf 交换机	2 台	32 × 100G 上联 + 32 × 25G 下联	上联 3.2T，128 × 25G 下行
100G 光模块	64 只（每台 8 只）	QSFP28 SR4，MPO 接口	Leaf 上联骨干
25G DAC/光模块	256 只	SFP28 DAC（机柜内）/ 光模块（跨柜）	GPU 服务器接入
MPO-12 跳线	按需	OM4，0.5~5m	机柜内汇聚

▶ 为什么优先选择 DAC 而非光模块?

✔ 成本优势：DAC 价格为同速率光模块的 10%~20%，64-GPU 集群可节省布线成本 60%+

✔ 功耗优势：DAC 无需光收发器，功耗为 0（纯无源），降低机柜 PDU 负载

✔ 延迟优势：铜缆传输延迟 < 0.1μs/m，比光纤低 30%，适合机柜内 1~3m 短距互联

✔ 推荐策略：机柜内（< 3m）→ DAC；跨机柜（> 3m）→ 25G SFP28 光模块

五、集群规模与性能规划

5.1 规模分级与带宽规划

集群规模	GPU 数量	网络架构	NETI710 配置	总接入带宽
小规模	8~32 GPU	单组 Leaf（2 × Spine）	NETI710-2CP × N	20G ~ 80G
中规模	64~128 GPU	2 组 Leaf + 2 × Spine	NETI710-4CP × N	160G ~ 320G
大规模	256+ GPU	4+ 组 Leaf + 4 × Spine	NETI710-4CP × N	640G+
超大规模	1024+ GPU	多 POD 互联，VXLAN	混合组网	按需扩展

5.2 性能基线指标

性能指标	目标值	说明
RoCEv2 端到端延迟	< 3 μs（PINGPONG）	两端 GPU Direct RDMA，实测值
GPU 间梯度同步带宽	≥ 9.4 Gbps（25G 链路）	实测 GPUDirect P2P 带宽
AllReduce 集合通信性能	接近线速（> 95%）	NCCL 实测，单次 32GB 数据
网络收敛时间（故障）	< 50 ms（BFD + ECMP）	链路故障后流量自动切换
对分带宽保证	100%（两层 Fat-Tree）	任意两个节点间无阻塞

Jumbo Frame MTU	支持 9216 Bytes	大象流传输效率最优
网卡功耗	≤ 15W（NETI710-4CP）	行业领先的低功耗设计

六、配置调优指南

6.1 网卡侧调优（EZMAX NETI710）

调优项	推荐配置	预期收益
RoCEv2 使能	ethtool -K eth0 roce on	开启 RDMA 功能
RX/TX 队列数	ethtool -L eth0 combined 32	高并发性能提升
RSS 队列	ethtool -X eth0 equal 32	多核 CPU 负载均衡
中断合并	ethtool -C eth0 rx-usecs 10 rx-frames 1	低延迟优先
PCIe 带宽协商	确保 PCIe Gen3 x8（NETI710-4CP）	避免 PCIe 瓶颈
MTU	ip link set eth0 mtu 9216	大块数据传输效率
VF 虚拟化（可选）	echo 64 > /sys/class/net/eth0/device/sriov_numvfs	64 VF 分片

6.2 交换机侧调优（参考）

- 全局开启 RoCEv2 兼容模式（PFC + ECN）；
- 配置 DCQCN 拥塞控制参数：g 扳_高速（gauge 高速）、Alpha 增益等；
- ETS（Enhanced Transmission Selection）配置：为 RDMA 流量预留带宽；
- QCN（Quantized Congestion Notification）参数调优（部分厂商支持）；
- 开启 LLDP（Link Layer Discovery Protocol），配合 EZMAX 网卡拓扑发现。

6.3 NCCL 测试验证

完成网络配置后，建议使用 NVIDIA NCCL-Tests 进行全链路验证，EZMAX 提供经过测试的 NCCL 兼容性配置参数。推荐测试项：

- NCCL Bandwidth Test（单卡对单卡，全互连）；
- NCCL AllReduce Test（多卡集合通信性能）；
- NCCL Broadcast Test（大模型梯度同步场景）；
- GPUDirect RDMA P2P Bandwidth Test（Nvidia peer_memory 驱动验证）。

七、典型应用场景与客户收益

场景 A：AI 训练集群（百卡级）

百卡 GPU 训练集群，跨节点 AllReduce 频繁

- 痛点：传统网络 AllReduce 慢，GPU 利用率低至 60%~70%
- EZMAX 方案：RoCEv2 + GPUDirect RDMA，GPU 利用率提升至 90%+
- ROI：训练周期缩短 20%~40%，电费节省显著

场景 B：HPC 高性能计算集群

分子动力学、气候模拟等 HPC 场景，强耦合 MPI 通信

- 痛点：MPI 全互连带宽不足，计算节点等待时间长
- EZMAX 方案：25G/100G 全对分带宽，Fat-Tree 无阻塞架构
- ROI：计算效率提升 30%+，同等算力节省 25% 集群节点

场景 C：分布式推理服务

实时推理请求跨 GPU 分发，低延迟优先

- 痛点：推理时延受网络限制，SLA 难以保证
- EZMAX 方案：RDMA 支撑超低延迟推理分发，SLA < 50ms
- ROI：推理吞吐量提升 2~3 倍，降低单次推理成本

7.1 客户收益总览

维度	客户收益	量化指标
训练效率	GPU 利用率从 65% 提升至 92%+	利用率 +27pp
模型迭代	千亿参数模型单次训练周期缩短	缩短 20%~40%
运维成本	统一供应商，统一支持响应	MTTR 降低 70%+
采购成本	一站式采购，降低管理复杂度	TCO 节省 15%~30%
扩展能力	架构支持平滑横向扩展至 256+ GPU	集群规模无上限

八、项目实施路径

阶段	周期	主要工作	里程碑交付物
Phase 1 需求对接	1~2 周	业务需求梳理、流量模型分析、网络现状调研	需求调研报告
Phase 2 方案设计	2~3 周	拓扑规划、产品选型、配置方案出图	详细设计方案（含拓扑图）
Phase 3 POC 验证	3~4 周	样机部署、NCCL 测试、RDMA 性能基准测试	POC 测试报告 + 性能基线
Phase 4 小批量部署	2~3 周	首批服务器部署、配置调优、全量测试	小批量验收报告
Phase 5 规模上线	按需	全量设备安装、网络调试、业务割接	竣工报告 + 运维手册
Phase 6 运维支持	持续	7×24 支持、远程运维、现场巡检	月度健康度报告

▶ EZMAX 增值服务

✔ 免费 POC 支持：提供样机借用 + 技术工程师现场支持，协助完成性能验证

✔ 兼容性认证：出具 EZMAX × 目标服务器 / 交换机兼容性认证报告

✔ 交钥匙交付：提供从网络规划到验收全程技术支持，无需客户额外投入专业网络人员

✔ 定期回访：每季度提供集群健康度巡检，提前识别潜在风险

九、关于 EZMAX

9.1 品牌定位

EZMAX 是北京万邦迪通科技有限公司旗下专注于算力中心网络基础设施的品牌。以网络控制器为核心，EZMAX 提供覆盖 10G/25G/100G 的全系列网卡、光模块与高密度布线产品，致力于成为智算中心建设者的首选网络基础设施供应商。

9.2 核心优势

- 全栈自主：芯片 → 网卡整机 → 方案交付，完全自主可控；

- 极致性能：端到端 < 5μs RDMA，支撑千亿参数大模型训练；
- 极简采购：一站式采购网卡 + 光模块 + 布线，零兼容风险；
- 贴身服务：从 POC 到交付全程技术支持，7×24 响应保障。

9.3 联系我们

获取 GPU 集群互联方案支持

样卡申请：访问官网「联系我们」→「样卡申请」，2 个工作日内技术团队响应

定制方案：提供集群规模、GPU 型号、业务场景定制化方案设计与报价