

智算中心

网络架构设计

AI Infrastructure Network Architecture

基于 EZMAX 全栈网络产品的高性能智算网络解决方案

版本 V1.0 | 2026 年 5 月

目录

- 一、概述
- 二、智算中心网络架构原则
- 三、网络分层架构设计
- 四、关键技术选型
- 五、网络拓扑设计
- 六、带宽与性能规划
- 七、可靠性设计
- 八、典型组网方案
- 九、产品配置推荐
- 十、实施建议

一、概述

1.1 背景与挑战

随着大语言模型(LLM)、生成式 AI(AIGC)等技术的快速发展，智算中心已成为数字经济的基础设施。智算中心网络面临着前所未有的挑战：

- 千亿参数大模型训练需要数千至数万张 GPU 协同工作，GPU 之间的通信带宽成为性能瓶颈
- 分布式训练中的集合通信（AllReduce、Broadcast 等）对网络延迟极度敏感
- 东西向流量占比超过 80%，传统数据中心网络架构无法满足需求
- 算力利用率直接决定模型训练效率和成本，网络设计至关重要

1.2 设计目标

目标维度	具体要求	关键指标
高带宽	满足 GPU 集群高速互联需求	100G/200G/400G 接入
低延迟	最小化通信开销	端到端 < 5 μ s, RDMA 零拷贝
无损网络	避免丢包重传影响训练	PFC、ECN 拥塞控制
高密度	最大化单机柜算力密度	MPO 高密度布线
可扩展	支持按需弹性扩容	模块化架构设计
自主可控	满足合规要求	国产芯片/操作系统

二、智算中心网络架构原则

2.1 Spine-Leaf 架构

智算中心网络采用 Spine-Leaf 两层 Clos 架构，相比传统三层架构，具有以下优势：

- 扁平化：任意两节点间只有 2 跳转发路径
- 等价路径：任意源到目的有 N 条等价路径，负载均衡无拥塞
- 弹性扩展：按需增加 Leaf 或 Spine 交换机即可扩容
- 高密度：每台 Leaf 交换机连接所有 Spine 交换机

2.2 全互联拓扑

GPU 服务器与存储设备采用全互联设计，确保任意 GPU 可与其他 GPU 直接通信，支撑分布式训练的无扭转发模式。

2.3 分离转发与存储网络

采用两套独立网络平面：

- 前端网络(InfiniBand/RoCE)：承载 HPC/AI 计算通信
- 存储网络(NVMe-oF/iSCSI)：承载数据读写流量

三、网络分层架构设计

3.1 网络分层概述

层次	功能定位	典型设备	带宽规格
接入层 (Leaf)	服务器接入	25G/100G 交换机	100G 上行/25-100G 下行
汇聚层 (Spine)	高速互联	100G/400G 交换机	400G 上行/400G 下行
核心层 (Core)	跨域互联	400G/800G 路由器	骨干链路互联
管理网络	带外管理	1G 交换机	管理/监控平面

3.2 接入层设计

接入层是智算网络的关键，直接影响 GPU 服务器的性能发挥：

- 每台 GPU 服务器配置 2-4 个 100G/200G 网卡，组成网卡绑定
- 每台服务器通过独立链路连接 2 台不同的 Leaf 交换机
- 采用自适应路由(Adaptive Routing)优化多路径负载
- 启用 RDMA 和 RoCEv2，硬件卸载降低 CPU 开销

3.3 汇聚层设计

汇聚层承担服务器间的快速互联：

- 所有 Leaf 交换机全互联到所有 Spine 交换机
- 采用 400G 端口，提供超大带宽容量
- 配置 PFC(优先级流量控制)和 ECN(显式拥塞通知)

- 部署 DCQCN 算法实现 RoCE 无损传输

四、关键技术选型

4.1 RDMA 技术

技术方案	协议	延迟	生态
InfiniBand	IB	<1 μ s	NVIDIA 独占, 封闭生态
RoCEv2	RDMA over Converged Ethernet	<5 μ s	开放生态, 兼容以太网
iWARP	Internet Wide Area RDMA Protocol	5-10 μ s	复杂, 兼容性一般

推荐：智算中心采用 RoCEv2 方案，兼顾高性能与开放生态。EZMAX N20H/N30F 系列网卡原生支持 RoCEv2。

4.2 拥塞控制机制

机制	原理	应用场景
PFC (Priority Flow Control)	链路层 Pause 机制，基于优先级暂停流量	构建无损物理链路
ECN (Explicit Congestion Notification)	IP 层拥塞通知，标记而非丢弃	端到端拥塞信号
DCQCN (Data Center Quantized Congestion Notification)	RoCE 专用拥塞控制，结合 PFC+ECN	智算中心 RoCE 网络
CC (Credit-based)	基于信用的流量控制	InfiniBand 网络

4.3 GPU Direct RDMA

GPU Direct RDMA 允许数据直接从 GPU 显存经 RDMA 网络传输，绕过 CPU 和系统内存，显著降低延迟和 CPU 开销：

- 传统路径：GPU→CPU 内存→网卡→网络→对端网卡→CPU 内存→GPU (多次拷贝)
- GDR 路径：GPU 显存→RDMA 网卡→网络→RDMA 网卡→GPU 显存 (零拷贝)
- EZMAX N30F 智算网卡完整支持 GPU Direct RDMA

五、网络拓扑设计

5.1 典型 Clos 拓扑

标准 3-stage Clos 拓扑设计：

参数	小型集群(64 GPU)	中型集群(256 GPU)	大型集群(1024 GPU)
Leaf 交换机(每服务器连接)	2 台	2 台	4 台
Leaf 端口数	32-64 口	64-128 口	128-256 口
Spine 交换机数量	4-8 台	8-16 台	16-32 台
收敛比	1:1(无收敛)	2:1	4:1
服务器带宽	200G	200G	400G

5.2 路由器设计

跨 Pod/跨区域通信通过路由网络互联：

- 东西向流量：Pod 内 Spine-Leaf 转发，延迟 < 1 μ s
- 南北向流量：经 Border Switch 访问外网
- 跨域通信：通过核心路由器互联

六、带宽与性能规划

6.1 带宽计算

网络平面	带宽需求	推荐方案
AI 计算网络	GPU 互联带宽×集群规模	100G/200G/400G RoCE
存储网络	存储吞吐量×IO 并发	25G/100G NVMe-oF
管理网络	设备数量×管理带宽	1G 带外管理
HBM/参数存储	模型参数量/训练周期	100G/200G 高速通道

6.2 性能指标

指标	目标值	说明
端到端延迟	< 5 μ s	服务器网卡到网卡
网络吞吐	线速转发	满载无丢包

RDMA 带宽利用率	>95%	有效带宽接近物理带宽
集合通信效率	>90%	AllReduce 等操作效率
故障收敛时间	<100ms	故障检测+路由收敛

七、可靠性设计

7.1 链路冗余

- 每台 GPU 服务器通过 2 条以上独立链路连接不同 Leaf 交换机
- 每台 Leaf 交换机到 Spine 全互联，单链路故障不影响通信
- 关键设备配置双电源、双控制平面

7.2 网络级容错

- BFD 双向转发检测，快速发现链路故障(<50ms)
- FRR 快速重路由，故障时毫秒级切换路径
- MLAG 多机链路聚合，跨设备链路冗余

7.3 拥塞保护

- PFC 实现链路级无损传输
- ECN 端到端拥塞通知
- DCQCN 智能拥塞控制
- QoS 优先级调度，关键流量优先

八、典型组网方案

8.1 小型智算集群 (8-64 GPU)

组件	配置	数量
GPU 服务器	8×A100/H100 + N30F 网卡	1-8 台
Leaf 交换机	64×100G + 8×400G 上行	2 台
Spine 交换机	32×400G	2 台
光模块	100G QSFP28 SR4/LR4	按需

布线	MPO-24 预端接 / DAC 铜缆	按需
----	---------------------	----

8.2 中型智算集群 (64-256 GPU)

组件	配置	数量
GPU 服务器	8×A100/H100 + N20H 网卡(100G 双口)	8-32 台
Leaf 交换机	128×100G + 32×400G 上行	4-8 台
Spine 交换机	64×400G	4-8 台
核心交换机	128×400G	2 台(跨域互联)
存储系统	NVMe 全闪存储 + 100G 网络	按需

8.3 大型智算集群 (256+ GPU)

大型集群采用多 Pod 架构，每个 Pod 内独立 Spine-Leaf，Pod 间通过核心网络互联：

- 每个 Pod 容纳 64-256 GPU，采用标准 Spine-Leaf 架构
- Pod 间核心网络提供跨域高速通道
- 全局存储与调度系统统一管理
- 建议采用 N30F 200G 网卡获得极致性能

九、产品配置推荐

9.1 网卡选型

场景	推荐网卡	关键特性	备注
小型集群	N30F (200G×2)	PCIe 5.0x16, GPU Direct RDMA	AI 训练首选
中型集群	N20H (100G×2)	PCIe 4.0x16, RoCEv2, <5μs	高性价比
大型集群	N30F (200G×2)	多 Pod 互联优化	按需扩展
通用接入	NETI710-4CP (10G×4)	Intel XL710, SR-IOV	入门级

9.2 光模块选型

距离	推荐光模块	光纤类型	适用场景
≤100m	100G QSFP28 SR4 (850nm)	OM4 MMF	机柜内/跨柜

≤500m	100G QSFP28 PSM4 (1310nm)	SMF	中等距离
≤2km	100G QSFP28 LR4 (1310nm)	SMF	数据中心间
≤10km	100G QSFP28 ER4 (1550nm)	SMF	长距离/城域
200G 连接	200G QSFP56 DR4/FR4	SMF/MMF	下一代集群

9.3 布线方案

连接类型	推荐方案	距离	优势
服务器-交换机(短)	DAC 铜缆	≤3m	最低延迟, 最低成本
服务器-交换机(中)	AOC 有源光缆	≤100m	轻量化, 即插即用
Leaf-Spine 主干	MPO-12/24 光跳线	≤300m	高密度, 快速部署
跨机相互联	LC-SC 单模跳线	≤500m	灵活, 易维护

十、实施建议

10.1 规划阶段

- 明确业务需求: AI 训练/推理/混合负载
- 评估集群规模: GPU 数量、扩展计划
- 制定带宽预算: 计算收敛比和端口需求
- 选择技术路线: RoCEv2 为主, 兼顾信创

10.2 部署阶段

- 网络设备上架: 按拓扑图连接 Spine-Leaf
- 光模块安装: 确认波长度数和光纤类型匹配
- 布线敷设: MPO 预端接+DAC 混合方案
- 网卡配置: 启用 RDMA、设置 PFC/ECN

10.3 调优阶段

- 网络性能测试: iperf3、perftest 验证带宽延迟
- RDMA 功能验证: 确认 RoCEv2 正常工作

- 拥塞控制调优：调整 DCQCN 参数
- GPU Direct 测试：验证端到端 GDR 路径

10.4 运维建议

- 建立监控体系：带宽利用率、延迟、丢包率
- 定期健康检查：光功率、端口状态、温度
- 容量规划：预留扩展端口和带宽余量
- 备件管理：关键光模块和 DAC 线缆储备

联系我们获取更多技术支持

EZMAX 算力中心网络基础设施 | 智算中心网络架构设计 V1.0

本方案基于 2026 年 5 月产品技术资料，如有更新恕不另行通知