KFNative™ 硬件方案文档

撰写日期：2026-08-15 作者：周舟

第 1 章 硬件设计目标与总体方案

1.1 硬件目标与量化指标

KFNative™ 硬件平台需支撑《KFNative™ 自主架构设计文档》第 1.3 节确立的架构目标，落为硬件级指标：

指标	目标值	说明
单 GPU 数据面带宽	Gen5 x16 64 GB/s（每方向）起步，向 Gen6 x16 约 128 GB/s 演进	PCIe 协议背景
SSD 单盘带宽	Gen5 x4 实测约 13.7–13.8 GB/s（Kioxia CM7 / Samsung PM1743）；Gen6 x16 约 100 GB/s（设计推算示例）	白皮书 5.1.5 实测规格 + 设计推算
节点内 GPU 数量	4–8 GPU（对标标准 4U/8GPU 服务器）	设计推算
节点内 SSD 容量	每 GPU 配对 4–8 块 U.2/EDSFF NVMe，单节点 32–64 TB 起	设计推算
节点间扩展	支持多节点经 Extended PCIe 组成 Fabric（2–16 节点）	架构目标
可靠性	不依赖双端口 SSD，跨节点纠删码	架构决策 4.2

1.2 GPU/SSD 配对模型

配对是 KFNative 的空间布局核心。硬件上采用"每 **GPU** 独占 **lane** 组 **+** 就近 **SSD** 组"模型：

- **1:1 配对（推理密集型）**：1 片 GPU 配 1 组 SSD（如 4×U.2 Gen5 x4），组内共享一个 Switch 下游端口组，数据通路确定性最强；
- **1:N 配对（容量密集型）**：1 片 GPU 配 2–4 组 SSD，经 Switch 扇出，容量优先；
- **N:1 配对（训练梯度暂存）**：多片 GPU 共享一组 SSD 作为梯度暂存位（Ring 归约中转），仅在训练场景由调度器动态建立。

配对关系不是物理固定连线，而是由计算层经 Fabric Manager 在共享资源池上动态建立，硬件只需保证"任意 GPU 到任意 SSD 的可达性与带宽隔离"（详见第 2 章拓扑）。

1.3 形态选择：加速板卡 + 标准服务器整机

采用两级形态，兼顾灵活性（板卡级）与规模化（整机级）：

形态	定位	内容
KFNative PCIe 加速板卡	单卡实现 GPU↔SSD 紧耦合的最小闭环	板载 PCIe Switch、Retimer、NVMe 接口（U.2/EDSFF/MCIO）、电源树、时钟、管理通道；以 Gen5 x16 上行接入主机（原理图见《KFNative PCIe 板卡原理图》与配套说明文档）
KFNative 整机	面向数据中心的对称节点	标准 4U/8GPU 服务器机箱 + 多张加速板卡 + CPU/内存（DDR5）+ 管理（BMC）+ 节点间扩展接口；兼容 DELL/HPE/Supermicro/Lenovo 生态

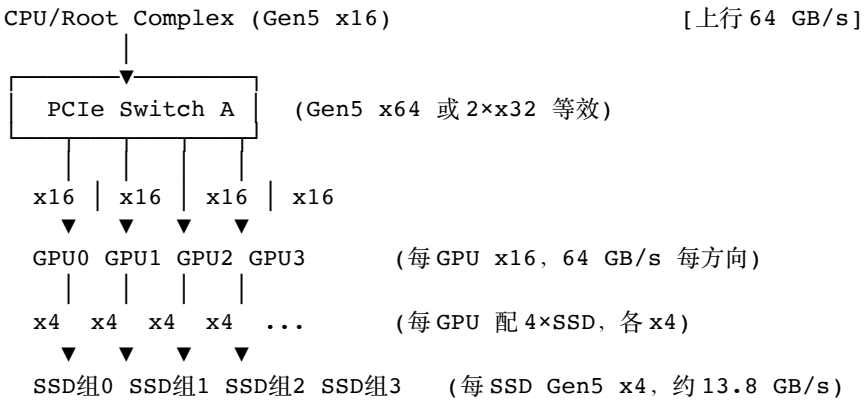
选择理由：板卡形态便于在现有服务器上增量部署（低进入门槛）；整机形态提供完整对称架构与跨节点纠删码所需的资源冗余。两级形态共用同一套板卡设计，降低研发成本。

> **本章结论**：硬件以"每 GPU 独占 lane + 就近 SSD"实现配对，以"加速板卡 + 标准整机"两级形态覆盖灵活部署与规模化；量化指标以 PCIe Gen5/Gen6 协议规格为基准，容量/规模类参数标注设计推算。

第 2 章 PCIe Fabric 拓扑设计

2.1 单节点拓扑：Switch 汇聚 + GPU/SSD 对等接入

单节点以 PCIe Switch 为汇聚点，GPU 与 SSD 对等接入，形成对称拓扑。以 4 GPU + 16 SSD 节点为例（设计推算）：



lane 预算（Gen5，4GPU+16SSD）：上行 x16（64 GB/s）+ 4×GPU x16（64 GB/s 每方向）+ 16×SSD x4 = 64 lanes 下行数据面。合计下行 4×16 + 16×4 = 128 lanes，需 2 颗 x64 Switch 或以 x32/x16 组合，视 Switch 产品能力确定（设计推算：Broadcom PEX89xxx 家族覆盖 x16/x32/x64 档位，见第 4 章 BOM）。

要点：GPU 与 SSD 的 lane 在 Switch 上以"虚拟端口组"方式隔离，保证配对单元间流量不串扰；GPU↔SSD 数据流经 Switch 单跳转发，不经过 CPU Root Complex，实现 P2P。

2.2 链路预算与 Retimer 布局

PCIe Gen5 32 GT/s（128b/130b 编码）对通道插入损耗敏感，Gen6 64 GT/s（PAM-4）更甚。链路预算按"总损耗 = PCB 走线损耗 + 连接器损耗"核算：

段	损耗预算（Gen5，设计推算）	处理
PCB 走线（≤10 英寸）	≤ 6–8 dB @ 16 GHz	控制走线长度/阻抗（85Ω 差分）
连接器（U.2/EDSFF/MCIO/金手指）	1.5–3 dB/个	选用低插损连接器
线缆（跨板/跨节点）	约 4.5 dB/m（Gen5 延长线缆，白皮书 5.1.4）	长度 ≤ 0.5–1 m

Retimer 放置规则（设计推算）：

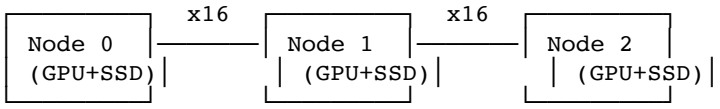
- 板内 GPU/SSD 走线短（≤10 英寸）→ 不需要 Retimer；

- 跨板/跨背板长链路（>10 英寸或含 2 个以上连接器）→ 每段链路插入 1 颗 Retimer；

- 节点间扩展链路 → 每方向 1–2 颗 Retimer（视线缆长度），协议感知 Retimer 同时承担均衡与时钟重建。

2.3 多节点扩展：Extended PCIe Fabric

多节点通过 PCIe 扩展口互联，构成 Fabric（对称优化架构）：



- **扩展方式**：每节点保留 1–2 个 Gen5 x16 扩展端口，经 Retimer + 线缆连接相邻节点；拓扑可为环（训练 Ring 场景）、网格或树（管理/大数据场景），由 Fabric Manager 配置（详见《Lattice Fabric 互联算法设计文档》）。

- **拓扑选择的权衡**：环拓扑 lane 利用率高、布线简单，适合 Ring AllReduce；网格拓扑容错性更强，但 lane 占用翻倍。生产首版建议环 + 可选旁路链路（设计推算）。

- **与 CXL 的关系**：若未来引入 CXL 内存池化，扩展口可复用 Flex Bus 协商 CXL 模式；首版保持纯 PCIe，降低复杂度。

> **本章结论**：单节点以 Switch 汇聚实现 GPU/SSD 对等 P2P，Retimer 按链路预算布置保障 Gen5/Gen6 信号完整性；多节点经扩展口组成环形/网格 Fabric，为训练 Ring 与容错重路由提供物理基础。

第 3 章 供电、散热与机械结构

3.1 功耗预算

以单张加速板卡（1× Switch + 2× Retimer + 8× SSD 接口 + 管理）与 4U 整机两级估算（**均为设计推算，依据典型器件 TDP 量级**）：

模块	单件典型功耗	数量	小计
PCIe Switch（Gen5 x64 级）	20–35 W	1	20–35 W
Retimer（Gen5 x16）	8–12 W	2	16–24 W
NVMe SSD（Gen5，满负载）	10–15 W	8	80–120 W
管理/时钟/辅助	5–10 W	–	5–10 W
板卡合计			121–189 W
4× GPU（商用 300W 级）	300 W	4	1200 W
CPU/内存/主板	–	–	300–500 W
板卡×4 + 扩展/散热	–	–	500–800 W
4U 整机合计			2000–2700 W

功耗以 Gen5 器件估算；Gen6 SSD/Switch 功耗可能上浮，需预留 15–20% 裕量（设计推算）。

3.2 电源树设计

板卡电源树（输入 12V 为主，设计推算）：

```
PCIe 金手指 12V ──► 输入滤波/热插拔保护
                    │
                    ├─► 12V → Switch 主电源域（含 POL 1.0/0.9V 核电压）
                    ├─► 12V → Retimer 电源域（0.9V 核 + 1.8V AUX）
                    ├─► 12V → SSD 电源域（12V→5V/3.3V，每盘独立负载开关 + 去耦）
                    └─► 12V → 管理域（BMC/CPLD 3.3V AUX，常供电）
```

- **热插拔保护**：输入侧放置 eFuse/热插拔控制器，支持 SSD 热插拔（U.2/U.3 热插拔规范，白皮书 5.3.4）；
- **去耦**：每 SSD 接口独立滤波电容组，隔离盘间开关噪声；
- **上电时序**：管理域 → Switch 核 → Retimer → SSD，由 CPLD 控制时序并上报 BMC；

- **整机电源**：标准冗余 PSU（2N，如 3000W 级 ×2，设计推算），保障 2000–2700 W 负载并支持纠删码重建突发。

3.3 散热设计

区域	散热方式	要点
GPU	整机风冷/液冷（按服务器标准）	商用 GPU 散热模块随卡
Switch/Retimer	主动散热片 + 系统风道	功耗集中，需独立风道与温度监控
SSD 盘阵	前置盘位风道 + 盘位导流	高密度 NVMe 对温度敏感，EDSFF/U.2 盘位设计参考白皮书 5.1.3 形态因子
板卡	被动散热 + 机箱风道	板卡厚度/散热片高度需满足标准 PCIe 槽位

散热预算按整机 2700 W 的 80% 转化为热量、进风 25°C 计算风量需求（设计推算：约 200–300 CFM 系统风量，以热仿真校核）。

3.4 机械结构约束

- **板卡**：标准全高全长（FHHL/FHFL）或定制长度，Gen5 金手指；关键信号走线控制在 10 英寸内；板卡厚度满足 U.2/EDSFF 接口安装；
- **整机**：4U 高度容纳 8×U.2/EDSFF 前置盘位 + 4 张加速板卡 + 2×GPU 电源域；节点间扩展口（x16 线缆）置于后部；
- **热插拔**：SSD 前置热插拔；板卡整卡支持维护级插拔（需主机支持 S4/S5 下操作）。

本章结论：硬件功耗以典型器件 TDP 量级估算（板卡约 121–189 W、整机约 2.0–2.7 kW），电源树以 12V 输入分级转换并配热插拔保护与 CPLD 时序控制，散热以风道+主动散热片为主，机械按标准 4U/板卡形态约束设计；所有功耗/风量数值为设计推算，需热仿真与实测校核。

第 4 章 BOM 级器件选型建议

4.1 关键器件选型表

以下选型基于"标准件 + 生态成熟度"原则，规格与实测数据引用白皮书第 5 章，具体型号为**设计推算/建议级**，需在采购时按最新产品手册核验：

器件类别	建议规格/型号方向	关键参数	选型依据	替代方向
PCIe Switch	Broadcom PEX89xxx 家族（Gen5）	x16/x32/x64 lane；集成 PEA 协议事件分析	白皮书 5.1.4 将 PEX89000 作为 Gen5 Switch 成熟方案；生态成熟	Microchip/ Marvell PCIe Switch（Gen5）
Retimer	协议感知 Gen5 Retimer（如 Astera Labs / Parade 方案）	32 GT/s，支持均衡/时钟重建	白皮书 5.1.4 Gen5 Retimer 卡为信号增强标准件	集成 Retimer 的 Switch（若未来产品支持）
NVMe SSD	Kioxia CM7-R / Samsung PM1743（Gen5 U.2/E3.S）	顺序读实测 13.7–13.8 GB/s；CM7 支持 SR-IOV/CMB/多流；容量 7.68–30.72 TB	白皮书 5.1.5 实测规格；企业级特性（SR-IOV/ CMB）匹配驱动设计	Kioxia CD8（读密集 1 DWPD，成本优先）
GPU	商用 GPU（Commodity，非旗舰）	PCIe x16；TDP 约 300 W 级	架构"去特化、无 NVLink 依赖"决策	任意厂商 x16 商用 GPU
电源	冗余 PSU（整机）+ 板级 POL/eFuse	整机 2×3000 W 级（设计推算）；板级 12V→0.9/1.8/3.3/5V	功耗预算（第 3 章）	按实测功耗调整
时钟	板级 PCIe 参考时钟（100 MHz）	支持 SRIS（独立参考时钟）以简化布线	白皮书 5.3.4 测试要点含 SRIS；Gen5 支持独立参考时钟	共用参考时钟（SRNS）方案

管理控制器	BMC（标准服务器 BMC）+ 板级 CPLD	IPMI/Redfish；板卡管理/上电时序/传感器	标准服务器生态；配合驱动固件管理通道	简化版：仅 CPLD + I2C
连接器/线缆	U.2/U.3（SFF-TA-1001）/ EDSFF（E1.S/ E3.S）/ MCIO	三模/热插拔；85Ω；Gen5 延长线缆约 4.5 dB/m	白皮书 5.1.3/5.1.4 形态与测试要点	按盘位形态选择

4.2 选型理由与权衡

1. **Switch 选 PEX89xxx 家族**：白皮书测试章节明确该系列是 Gen5 Switch 的成熟选择，且内嵌 PEA 引擎便于线上协议监控；lane 档位覆盖单节点 128 lane 需求。
2. **SSD 选 Gen5 U.2/E3.S 企业盘**：实测 13.7–13.8 GB/s 远高于 Gen4 时代，接近 Gen5 x4 上限；支持 SR-IOV/CMB 可直接服务于驱动固件设计的队列/缓存机制（见《KFNative 驱动固件设计文档》）；SR-IOV 可支持 K8s 多租户直接分配盘。
3. **Retimer 独立部署而非依赖 Switch 集成**：首版保守起见独立 Retimer 保证跨板/跨节点链路裕量；待集成 Retimer 的 Switch 成熟后再收敛 BOM。
4. **GPU 不绑定厂商**：满足 x16 PCIe 即可，配合 P2P 能力（GPUDirect/P2P）实现 GPU↔SSD 直通，符合"任意厂商服务器 + 商用 GPU"主张。

4.3 BOM 成本量级估算（设计推算）

以单张加速板卡（8×SSD 接口 + Switch + 2×Retimer）为口径：

项目	量级（设计推算）
PCIe Switch	100–200 USD
Retimer ×2	40–80 USD
NVMe SSD ×8（Gen5 7.68–15.36 TB）	8×150–400 USD
电源/时钟/管理/连接器	100–200 USD
PCB/散热/结构	100–200 USD
板卡合计（不含 SSD）	约 350–700 USD
含 SSD 合计	约 1,550–3,900 USD

成本量级基于 2024–2026 市场公开价格趋势推算，仅作预算参考，非报价依据。

风险与替代：Gen6 SSD/Switch 尚未大规模量产时，首版以 Gen5 交付，架构（lane 分配/电源/散热）预留 Gen6 升级裕量；若 Gen6 延迟，Gen5 x4 SSD 堆叠（更多盘位）仍可逼近带宽目标。

本章结论：BOM 全部采用标准生态组件（Broadcom 系 Switch、协议感知 Retimer、Kioxia/Samsung Gen5 企业 SSD、任意商用 GPU、标准 BMC/PSU），以生态成熟度和实测规格为选型依据；成本与功耗为设计推算，采购前需按最新产品手册核验。