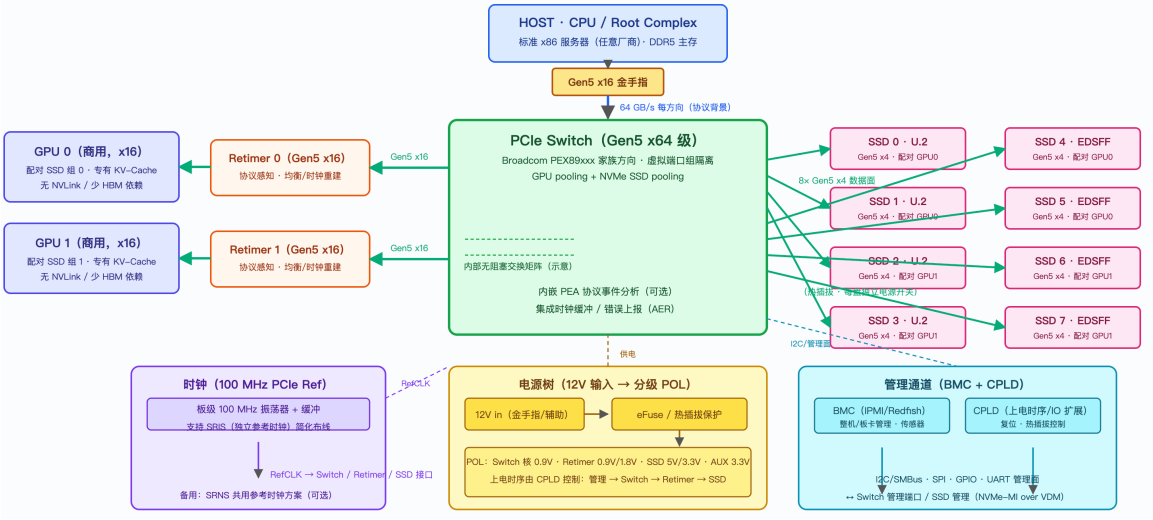
KFNative™ PCIe 板卡原理说明

撰写日期：2026年9月1日 作者：周舟

1. 原理图总体说明

本原理图（原理框图级）以单张 KFNative™ PCIe 加速板卡为对象，覆盖六大核心模块：**PCIe Switch、Retimer、SSD 接口、电源树、时钟、管理通道**，并按《硬件方案文档》第 1.2 节"GPU/SSD 配对"模型布置数据面。



图例与说明

- PCIe 数据面（Gen5；Switch↔SSD 为 P2P 路径，不经主机内存/网卡）
- 100 MHz 参考时钟（RefCLK）
- 管理面（I2C/SMBus/SPI/GPIO/UART）
- 电源（12V 输入 → POL 各 rail）

模块说明：

- HOST：任意厂商标准 x86 服务器（DELL/HPE/Supermicro/Lenovo 生态），上行 Gen5 x16 接入板卡。
- PCIe Switch：板载汇聚点，GPU 与 SSD 对等接入；虚拟端口组隔离保证配对单元流量不串扰；内嵌 PEA 可做协议监控。
- Retimer：跨板/长链路信号重建（均衡 + 时钟重建），板内短走线（≤10 in）可不布；本框图按跨背板场景预留。
- SSD 接口：U.2（SFF-8639）/ EDSFF（E1.S/E3.S）/ MCIO 均可，三模 U.3 兼容；支持热插拔与 NVMe-MI 管理；每盘 x4 Gen5。
- 电源树：12V 输入 → eFuse/热插拔保护 → 多路 POL；CPLD 控制上电时序；SSD 独立负载开关（热插拔）。
- 时钟：100 MHz 参考时钟，支持 SRIS 独立参考时钟以简化布线；SRNS 方案可选。
- 管理：BMC（IPMI/Redfish）承载整机/板卡管理，CPLD 承载上电时序/复位/热插拔控制；管理面经 I2C/SMBus 与 Switch/SSD 交互。

注：本图为首版设计推算原理框图，用于架构评审与原理图设计输入；具体器件型号、lane 分配、功耗与信号完整性需经仿真与实测校核（详见《KFNative 硬件方案文档》第 2、3、4 章）。

信号流主线：

```
HOST (Gen5 x16) → 金手指→ PCIe Switch → (左) Retimer → GPU 0/1
                                       └──→ (右) 8× SSD 接口（U.2/EDSFF，每盘 x4）
管理面（BMC/CPLD，I2C/SMBus/SPI/GPIO）
电源面（12V → eFuse → POL rails）
时钟面（100 MHz RefCLK，SRIS 可选）
```

2. 模块级说明

2.1 PCIe Switch（板载汇聚点）

- 定位：GPU 与 SSD 对等接入的 Fabric 汇聚点，实现 GPU pooling 与 NVMe SSD pooling（与白皮书三层栈第 1 层一致）。
- 选型方向：Broadcom PEX89xxx 家族（Gen5，x16/x32/x64 档位），白皮书 5.1.4 将其列为成熟 Gen5 Switch 方案（协议背景）。
- 设计要点：
- 上行 Gen5 x16（64 GB/s 每方向，协议背景）连接 HOST；
- 下行按配对模型分配：GPU x16 ×N、SSD x4 ×M，虚拟端口组隔离，保证配对单元流量不串扰；
- 支持 P2P（GPU↔SSD 直接 DMA，不经 Root Complex），对应驱动固件 DMA 通路（见驱动文档第 3 章）；
- 内嵌 PEA 协议事件分析引擎，可选用于线上监控（白皮书 5.1.4 提及 PEA）。
- **lane 预算**（4 GPU + 8 SSD 版，设计推算）：上行 x16 + GPU 2×x16 + SSD 8×x4 = 80 lanes，选用 x64 级 Switch 或双颗组合。

2.2 Retimer

- 定位：长链路/跨连接器信号重建，补偿 Gen5 32 GT/s 的插入损耗。
- 选型方向：协议感知 Gen5 Retimer（Astera Labs / Parade 方向），支持均衡与时钟重建（白皮书 5.1.4 将 Gen5 Retimer 卡列为信号增强标准件，协议背景）。
- 布局规则（设计推算）：
- 板内短走线（≤10 in）不布 Retimer；
- 跨背板/跨板长链路布 Retimer，每段 1 颗；
- 节点间扩展链路每方向 1–2 颗（视线缆长度与损耗预算，白皮书 5.1.4 延长线缆约 4.5 dB/m）。

2.3 SSD 接口

- 形态：U.2（SFF-8639）/ EDSFF（E1.S/E3.S）/ MCIO，支持三模 U.3（白皮书 5.1.3 形态因子，协议背景）。
- 数据面：每盘 Gen5 x4（SSD 实测顺序读 13.7–13.8 GB/s，白皮书 5.1.5 实测规格）。
- 配对：本版按 8 盘 + 2 GPU 布置（SSD 0–3 配 GPU0，SSD 4–7 配 GPU1，示意），生产版按《硬件方案文档》1.2 节配对模型调整。
- 管理与热插拔：每盘独立负载开关（热插拔，白皮书 5.3.4 测试要点）；管理经 NVMe-MI over VDM（SMBus/I2C）访问盘内管理信息；SR-IOV/CMB 特性供驱动直接使用（见驱动文档第 3 章 NVMe 通路）。

2.4 电源树

- 输入：12V（PCIe 金手指 + 板卡辅助供电），经输入滤波与 eFuse/热插拔保护；
- 分级 **POL**：Switch 核 0.9V、Retimer 0.9V/1.8V、SSD 5V/3.3V、管理 AUX 3.3V（常供电）；
- 上电时序：CPLD 控制：管理域 → Switch 核 → Retimer → SSD，时序失败上报 BMC；
- 功耗量级（设计推算）：Switch 20–35 W + Retimer 2×8–12 W + SSD 8×10–15 W + 管理 5–10 W ≈ 121–189 W，详见《硬件方案文档》3.1 节。

2.5 时钟

- 板级 100 MHz PCIe 参考时钟振荡器 + 缓冲，分发至 Switch/Retimer/SSD 接口；
- **SRIS**（独立参考时钟，Separate Refclk with SRIS architecture）为推荐模式，简化布线并支持热插拔/线缆场景（白皮书 5.3.4 测试要点含 SRIS，协议背景）；SRNS 共用参考时钟作为可选降级方案；
- 时钟抖动需满足 Gen5/Gen6 规范（设计推算：按 PCI-SIG 相关规范要求，待仿真校核）。

2.6 管理通道

- **BMC**：标准服务器 BMC，IPMI/Redfish 接口，承载板卡/整机管理、传感器（电压/温度/功耗）、固件升级；
- **CPLD**：上电时序、复位管理、热插拔控制、IO 扩展；

- **管理面拓扑**：I2C/SMBus 总线连接 Switch 管理端口与 SSD 管理（NVMe-MI over VDM）；SPI 用于固件存储访问；GPIO/UART 用于调试；
- **与驱动关系**：管理通道独立于数据面，驱动固件经管理面获取板卡健康与拓扑信息（见驱动文档第 4 章固件功能模块）。

3. 关键信号清单（设计推算，原理图输入级）

信号组	方向	说明
PETp/n ×16（上行）	HOST ↔ Switch	Gen5 32 GT/s，64 GB/s 每方向
PETp/n ×16 ×2	Switch ↔ Retimer ↔ GPU	每 GPU x16 数据面
PETp/n ×4 ×8	Switch ↔ SSD	每盘 x4，P2P DMA 路径
REFCLKp/n 100 MHz	时钟 → 各端点	SRIS 模式为主
SMBus/I2C	BMC/CPLD ↔ Switch/SSD	管理面，NVMe-MI over VDM
SPI/GPIO/UART	CPLD/BMC ↔ 板卡	固件、复位、调试
12V/5V/3.3V/1.8V/0.9V	电源树 → 各模块	POL 输出，独立去耦

4. 设计推算与待校核项

- 具体 Switch/Retimer 器件型号与 lane 档位（以最新产品手册为准）；
- SSD 盘位数与 GPU 配对比（按负载模型调整）；
- 功耗与散热（热仿真 + 实测）；
- 时钟抖动/信号完整性（SI 仿真 + 白皮书 5.3.4 测试流程）；
- Gen6（64 GT/s PAM-4）演进路径需按硬件方案第 2 章链路预算重新校核 Retimer 布局。

> **小结**：本原理框图覆盖 PCIe Switch、Retimer、SSD 接口、电源树、时钟、管理通道六大模块，数据面走"GPU↔Switch↔SSD"P2P 通路，管理面与电源面独立成域；可作为原理图设计与评审的输入，图中未标注的具体值均属设计推算，需按仿真与实测校核后转正式原理图。