

# 交换芯片解决什么？

---

第二期 · PCIe 交换网络与 CPU / GPU 兼容

## 第二期

01 端口与数据路径

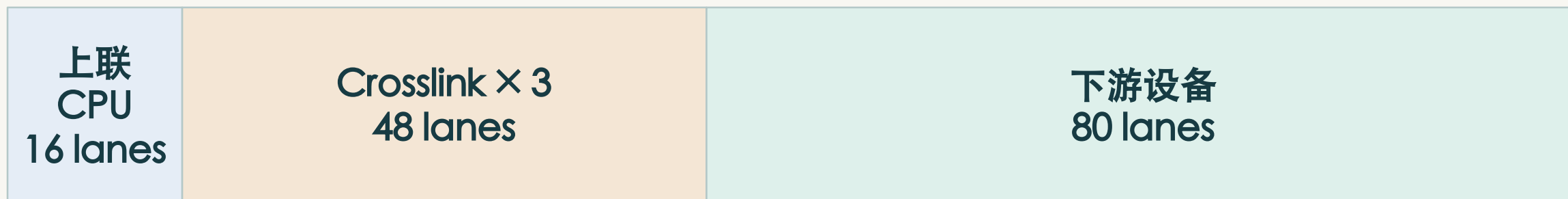
02 共享带宽与资源管理

03 驱动支持与系统验收

器件组合提出可能性，系统证据决定范围

# 144 lanes 怎样分给整机?

教学假设：每颗 switch 的上联和 crosslink 均为  $\times 16$



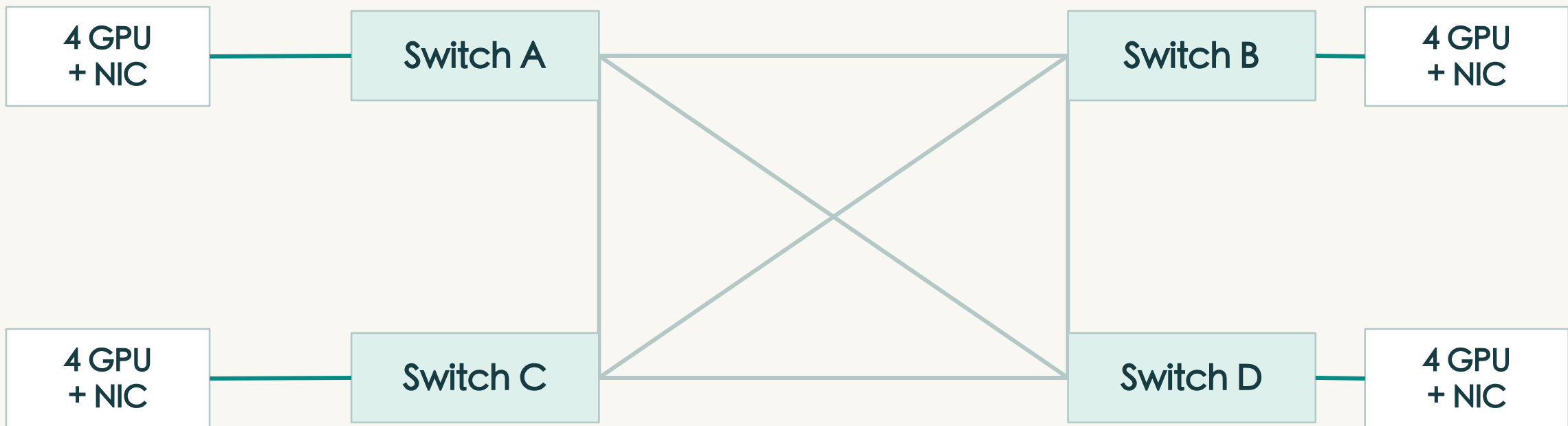
$$144 - 16 - 3 \times 16 = 80$$

80 lanes  $\rightarrow$  5 个  $\times 16$  设备端口

预算必须同时计入 CPU、crosslink、GPU、NIC 和 SSD

# 四颗 switch：逻辑 full mesh

基于原文的条件化教学图；每条 crosslink 假设为  $\times 16$

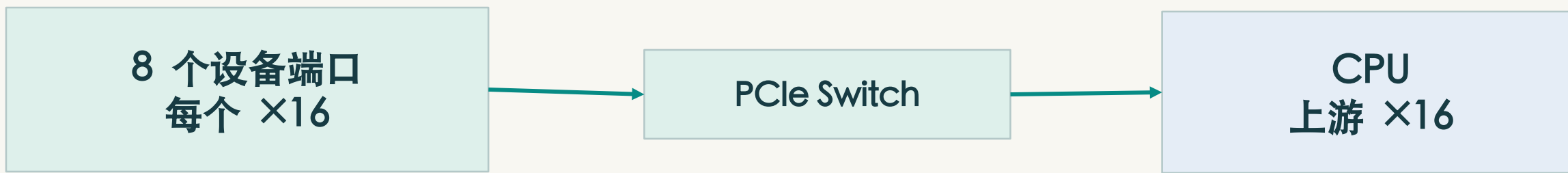


四个交换芯片间的逻辑 full mesh；主机上联省略

**full mesh 描述连接关系；不自动证明路由与无阻塞**

# 多端口，不等于多倍上游带宽

教学假设：8 个下游  $\times 16$ ，共享 1 个同代际上游  $\times 16$



主机方向：8:1 带宽超卖（同代际假设）

局部 P2P 可留在交换芯片内；仍受共享切面影响

共享切面、出口竞争与内部交换能力要分别测量

# 同一主机与多主机，不是一种管理模型

地址不冲突只是条件之一，不能推出跨 RC 自动可达

| 模式             | 资源管理特点         | 必须补充的证据          |
|----------------|----------------|------------------|
| 透明 fan-out     | 一个主机管理 PCIe 层级 | 枚举、BAR、设备驱动      |
| 多主机 / NTB      | 独立主机之间建立窗口等机制  | 地址域、权限、请求与返回路径   |
| 专用 PCIe fabric | 依具体芯片与管理软件组织   | 路由、归属、隔离、恢复与支持矩阵 |

跨域访问需要资源归属、路由、隔离和恢复设计

# 国产 CPU + 海外 GPU：逐层核验

---

产地不是 ISA；有 PCIe 插槽也不是完整兼容承诺

ISA / ABI

具体指令集、系统调用与二进制接口

OS / 固件

受支持发行版、内核、枚举与 MMIO

GPU 驱动

明确 SKU、主机平台和驱动版本

用户态软件

运行时、数学库、编译器与框架

多卡通信

P2P、集合通信与目标模型表现

每层都需要具体型号、版本与可复现实验

# 地址统一、内存管理、缓存一致性

三个名字相似的问题，需要三份不同证明

| 概念         | 它帮助解决什么       | 仍需关注         |
|------------|---------------|--------------|
| 统一虚拟寻址 UVA | 地址空间和指针识别     | 访问许可、远端路径    |
| 统一内存管理     | 按实现管理可访问性与迁移  | 驻留、缺页、迁移和局部性 |
| 硬件缓存一致性    | 规定各缓存副本如何保持一致 | 具体互连、范围、同步规则 |
| 显存总量聚合     | 模型可分片使用多卡容量   | 分配策略与跨卡访问代价  |

可寻址 ≠ 等速访问 ≠ 硬件缓存一致性

# RTX 5090：观察与支持范围分开

---

原文 p2 的 16 卡结果，尚缺完整复现条件

**官方明确答复（2025-03-13）：RTX 50 GeForce 不支持 GPU-GPU P2P**

**作者报告  
特定配置观察**

**官方支持  
型号 / 版本限定**

**可复现实验  
正确性 + 路径日志**

截至本轮核验未找到正式开放的新公告；不能据原文推出常规受支持能力。

**不否定作者可能的实验；不把实验推广为官方能力**

# 从“能看见”到“确实有用”

---

验证顺序决定结论的可信度

- 1 枚举：设备、链路速率、宽度、 NUMA 与地址资源
  - 2 正确性： peer 能力矩阵、读写权限与数据校验
  - 3 性能：每一对设备、并发流量与共享瓶颈
  - 4 集合通信：消息大小、算法、实际传输路径
  - 5 工作负载：质量、吞吐、尾延迟与持续稳定性
- 

拓扑图 → 正确性 → 并发 → collective → 真实任务

# 把方案结论写得可检验

明确已证明的能力、依赖条件与待补证据

| 方案声称        | 应交付的证据                         |
|-------------|--------------------------------|
| 多卡可扩展       | 完整拓扑、端口宽度、共享切面和流量矩阵            |
| 国产主机兼容      | 具体 CPU / ISA / OS / 驱动 / 用户态版本 |
| 支持 GPU P2P  | peer 矩阵、正确性、配置与实际路径日志          |
| 显存池化 / 更高性能 | 内存语义、任务基线、结果和稳定性               |
| 可以直接替换器件    | 引脚之外的固件、热设计、功能和验收记录            |

下一期：可靠、畅通与正确可见，分别由谁保证？