架构设计 I：近存 MoE (Near-Memory MoE)

1. 解决路由危机

正如 671B 模拟所示，将数据搬运到专家处（传统 MoE 模式）在超大规模下是致命的。 HLPO 翻转了这一范式：**将计算搬运到数据处**。

2. 冷切换原理 (The Cold Switching Principle)

- **热切换 (传统)**：
    1. 从显存读取输入。
    2. 检查路由表。
    3. 通过 NVLink/PCIe 发送数据到专家 GPU。
    4. 专家计算。
    5. 发回结果。
    - **结果**：巨大的延迟 + 热量。
- **冷切换 (HLPO)**：
    1. **HCA** 预先计算“静态路由图谱” (Static Routing Map，确定需要哪个专家)。
    2. 数据通过预设的光学/低阻路径直接流向本地专家模块。
    3. 无“搜索”，无“等待”。
    - **结果**：流动的连续性。

3. 影响

这种架构消除了 MoE 模型中的“跨芯片”惩罚，使得分布式的 671B 模型运行起来就像是在单颗巨大的统一芯片上一样。