英伟达 H800 GPU：算力满血的“孤胆英雄”

H800 是英伟达（NVIDIA）在 2023
年推出的特供型号。它是一款极具戏剧性的产品：它拥有与顶级旗舰 H100
几乎完全一致的“大脑”（计算核心），却被精准地切断了部分“沟通渠道”（互联带宽）。

它是大模型预训练（Pre-training）和大规模推理场景中，除了 H100 之外最强悍的工具。

1. 核心架构：满血 Hopper 算力的传承

H800 最核心的卖点在于：它的单卡计算性能几乎没有“缩水”。

- Transformer Engine (FP8 精度)：与 H100 一样，H800 完整搭载了针对大模型优化的 Transformer 引擎。它支持 FP8
  混合精度计算，在训练万亿参数级别的模型时，算力表现极其惊人。
- 算力指标：
- FP8 算力：高达 3,026 TFLOPS（稀疏模式）。
- FP16 算力：高达 1,513 TFLOPS（稀疏模式）。

- 意义：在单卡微调（Fine-tuning）或中小规模训练任务中，你几乎感受不到 H800 与 H100 之间的性能差距。

2. 核心限制：被腰斩的“社交能力”

为了符合出口管制政策，英伟达对 H800 的 NVLink（卡间互联）进行了精准打击。

- NVLink 带宽降级：
- H100：双向互联带宽为 900 GB/s。
- H800：双向互联带宽被限制在 400 GB/s。

物理意义：H800 的计算核心处理数据的速度极快，但当 8 张卡或者数千张卡需要通过 NVLink
交换模型梯度信息时，400 GB/s 的带宽会成为瓶颈。这就像一群天才（H800
核心）坐在一起工作，但他们之间的电话线被降级成了窄带。

专家解读：这意味着 H800 在单机（8 卡）环境下的效率依然极高，但在进行万卡级别的超大规模集群预训练时，集群的整体扩展效率（Scaling Efficiency）会比 H100 低 15%-25%。

3. 存储性能：80GB HBM3 高速通路

H800 的存储规格与 H100 基本保持一致，确保了它在推理场景下的统治力。

- 显存容量：80GB。
- 显存带宽：
- SXM 版本：约为 3.35 TB/s（与 H100 相同）。
- PCIe 版本：约为 2 TB/s（采用 HBM2e）。

优势：得益于 Hopper 架构带来的极高带宽，H800 在处理像 Llama 3 这样的大模型推理时，响应速度极快，是目前高性能推理服务器的核心首选。

4. H800 vs A100/A800：代差级的碾压

很多工程师会问，如果都有 400 GB/s 的带宽限制，为什么不选更便宜的 A800？

维度	A800 (Ampere)	H800 (Hopper)
FP8 支持	不支持	完全支持 (Transformer Engine)
LLM 推理速度	1x 基准	最高提升 30 倍
LLM 训练速度	1x 基准	最高提升 9 倍
显存带宽	2 TB/s	3.35 TB/s (SXM版)

结论：H800 是专门为 Transformer 架构 优化的。即便互联带宽相同，H800 处理 LLM 的效率依然远超 A100 系列。最近大火的 DeepSeek R1 模型，其初始训练正是主要基于 H800 集群完成的。

5. 专家总结：H800 的最佳实战姿态

H800 是一款“单兵作战极强，大兵团协同稍逊”的战神。

1. 大规模微调与小规模预训练：它是目前除了 H100 之外的最佳选择，算力溢出带来的收益远超带宽缩水带来的损耗。
2. 高性能推理服务：由于其满血的显存带宽和 FP8 支持，它是目前市面上最强的大模型推理卡。
3. 科研计算的取舍：需要注意，H800 极大削减了双精度（FP64）算力（仅为 H100 的约 1/30），如果你是做气象模拟或流体物理计算，请避开 H800，选择 A100/A800。