英伟达 H20 GPU：带宽优先的“非典型”利器

H20 是英伟达（NVIDIA）历史上最具设计巧思的产品之一。它是为了在特定出口管制政策下，依然能为大模型提供强大支撑而诞生的“特供版”Hopper 架构 GPU。

如果说 H100 是一个肌肉发达的拳击手，那么 H20 就是一个力量受限但拥有极速反应和超粗血管的马拉松运动员。在 AI 推理领域，它展现出了惊人的逆袭能力。

1. 核心设计哲学：算力换带宽

为了符合合规要求，英伟达对 H20 采取了非常极端的“刀法”。它并没有削减显存，而是精准地砍掉了计算核心（CUDA Core 和 Tensor Core）的数量。

- 算力（FLOPS）大幅削减：H20 的 FP16 算力约为 296 TFLOPS。作为对比，H100 的算力接近 2000 TFLOPS。这意味着单看单卡的计算爆发力，H20 只有 H100 的四分之一左右。
- 血管（带宽）完整保留：令人惊叹的是，H20 的显存带宽高达 4.0 TB/s，这甚至比标准版的 H100（3.35 TB/s）还要快。
- 胃口（显存）甚至更大：H20 配备了 96GB HBM3 显存，比 H100 的 80GB 还要多出 16GB。

2. 为什么 H20 在推理中“逆袭”了？

对于搞 AI 训练和推理的人来说，有一个残酷的现实：大语言模型（LLM）的性能瓶颈往往不在于计算速度，而在于显存带宽。

在大模型生成文字（Token）时，每一颗 Token 的产生都需要将整个模型的参数从显存“搬运”到计算核心。这个过程被称为 Memory-bound（受限于内存）。

- 推理奇迹：由于 H20 搬运数据的速度（4.0 TB/s）极快，在跑 Llama 3 这样的大模型推理时，它生成的首字延迟（Latency）非常低。在中小批量（Batch Size）任务中，H20 的实际推理速度甚至能比 H100 快 20% 左右。
- 显存优势：多出的 16GB 显存允许它容纳更长的上下文（KV Cache），这意味着它可以处理更长的对话而不必频繁切分任务。

3. 集群通信：满血的 NVLink 带宽

H20 的另一个“杀手锏”是它完整保留了 900 GB/s 的第四代 NVLink 互联能力。

在构建大规模算力集群时，显卡之间的通信效率决定了集群的上限。

- 大兵团作战：因为 NVLink 是满血的，你可以将数千张 H20 像连接 H100 一样紧密地组合在一起。
- 训练补救：虽然单卡算力弱，但通过增加显卡数量，并利用 900 GB/s 的高速公路进行同步，H20 集群依然可以完成大规模模型的微调（Fine-tuning）任务，且效率远高于没有高速互联的游戏显卡集群。

4. 技术规格对比表

指标	H20 (特供版)	H100 (旗舰版)	对推理的影响
显存容量	96GB HBM3	80GB HBM3	H20 能装下更长的上下文
显存带宽	4.0 TB/s	3.35 TB/s	H20 读写模型参数更快
FP16 算力	296 TFLOPS	1979 TFLOPS	H100 在纯计算任务上更强
NVLink 带宽	900 GB/s	900 GB/s	两者在大规模协同上表现一致
TDP (功耗)	400W	700W	H20 更省电，散热要求更低

5. 专家总结：H20 的生存之道

H20 是一张“为推理而生”的显卡。它完美避开了出口管制的算力红线，却在 AI 业务最核心的“带宽”和“显存”上给足了料。

适用场景：

1. 大规模大模型部署：如果你需要支撑千万级用户的实时对话（如 Llama 3 部署），H20 的性价比和响应速度是极高的。
2. 长文本应用：处理长文档、长代码解析，96GB 显存带来的 KV Cache 优势明显。
3. 大模型微调：在拥有满血 NVLink 的前提下，通过 8 卡或多机联动，A800/H800 的用户可以无缝迁移到 H20 架构。