英伟达 B200 GPU：迈向 AGI 的万亿参数算力巨塔

2024 年发布的 NVIDIA Blackwell B200 标志着 GPU
发展史上的一个重大范式转变。它不再是单纯追求单块芯片的极限，而是通过“系统级
GPU”的设计，彻底打破了单芯片制造的物理上限。

如果说 H100 是单台发动机的巅峰，那么 B200
就是一台由两台顶级引擎并联、共用一个控制系统的超级动力站。

1. 核心架构：双芯合一的“胶水”艺术

B200 最显著的特征是采用了 双芯片封装（Dual-die）
技术。由于单块芯片的尺寸已经触及了光刻机的物理极限（Reticle
Limit），英伟达选择将两块巨大的芯片封装在一起。

- 2080 亿晶体管：这是 H100（800 亿）的两倍多。
- 10 TB/s
  的芯间互联
  ：这两块芯片通过极速链路连接，在软件层面，它们被识别为一张完全统一、缓存一致的
  GPU。开发者不需要修改代码来适配双芯，系统会自动完成任务分配。
- 制程工艺：采用台积电专门定制的 4NP 工艺，实现了更高的集成度和能效比。

2. 算力革命：FP4 精度与第二代 Transformer 引擎

在 B200 上，英伟达引入了足以改变大模型运行规则的新精度：FP4（4 位浮点数）。

- 第二代 Transformer 引擎：B200 能够根据神经网络的需求，动态地在 FP8、FP6 甚至
  FP4 之间切换。
- 性能爆发：在 FP4 精度下，单张 B200 的算力达到了恐怖的 20 PetaFLOPS（2
  亿亿次运算/秒）。
- 推理效率：相比 H100，B200 在运行大模型推理时，性能提升最高可达 30
  倍。这意味着以前需要一个机柜完成的任务，现在可能只需要一张卡。

3. 显存与带宽：消除“内存墙”

为了喂饱如此恐怖的算力，B200 搭载了目前世界上最强的存储系统。

- 192GB HBM3e 显存：单卡容量几乎是 A100 的两倍多。如此巨大的空间可以让万亿参数
  的模型在更少的显卡上跑起来，减少跨机通信。
- 8 TB/s 显存带宽：这是 H100 的 2.4
  倍。在处理实时交互的大模型（如对话机器人）时，高带宽意味着更快的响应速度。

4. 全新互联：第五代 NVLink

在 B200 时代，英伟达进一步强化了“集群即 GPU”的概念。

- 1.8 TB/s 双向带宽：单张 B200 的互联带宽比 H100 翻了一倍。
- 576 张 GPU 互联：通过全新的 NVSwitch，最多支持 576 张显卡在满血带宽下无缝沟通。
- RAS 引擎：由于芯片规模巨大，B200
  加入了专门的可靠性、可用性和可维护性引擎，能够利用 AI
  预测潜在的硬件故障，确保数万张显卡组成的集群能稳定运行几周不宕机。

5. 技术规格横向对比

指标	H100 (Hopper)	B200 (Blackwell)	提升幅度
晶体管数量	800 亿	2080 亿	~2.6x
显存容量	80GB HBM3	192GB HBM3e	2.4x
显存带宽	3.35 TB/s	8 TB/s	~2.4x
FP4 算力	不支持	20 PetaFLOPS	维度跨越
FP8 算力	4 PetaFLOPS	10 PetaFLOPS	2.5x
最大功耗 (TDP)	700W	1000W - 1200W	散热挑战剧增

6. 专家总结：B200 改变了什么？

B200 的出现让 “万亿参数模型”的训练和推理
从“极少数天才的实验室产物”变成了“工业化的大规模生产”。

1. 能效比奇迹：虽然单卡功耗过千瓦，但在处理同样规模的 AI 任务时，B200 的能耗仅为 H100 的 1/25。
2. 推理成本骤降：通过 FP4 精度和巨大的带宽，大模型的运营成本（Token 成本）将迎来数量级的下降。
3. 水冷时代的到来：由于 B200
   功耗
   极高，传
   统的风冷已经很难压住它的热量。这意味着未来的算力中心将全面转向液冷方案。