英伟达 B300 GPU：AI 推理与长文本时代的“终极核弹”

NVIDIA B300（正式名称为 Blackwell Ultra）是英伟达在 2025 年推出的旗舰级数据中心 GPU。它是 Blackwell
架构的完全体，专门针对大规模推理（Reasoning）、长文本处理和
万亿参数模型训练进行了极限强化。

如果说 B200 是 Blackwell 的首秀，那么 B300 则是将各项物理参数推向了当前的半导体极限。

1. 核心架构：Blackwell Ultra 的全面进化

B300 沿用了 B200
的双芯片封装（Dual-die）设计，但在制造工艺和核心调度上进行了深度优化。

- 晶体管规模：集成 2080 亿个晶体管，采用台积电定制的 4NP 工艺。
- 统一计算域：两个芯片通过每秒 10 TB/s
  的极速链路互联，对软件层呈现为一颗拥有超强算力的单体 GPU。
- 计算密度提升：相比 B200，B300 的 FP4 密集算力提升了约 1.5 倍。

2. 存储革命：突破“内存墙”

B300 最显著的提升在于显存，这使其成为处理长文本和复杂推理任务的利器。

- 288GB HBM3e 显存：采用全新的 12 层（12-high）堆叠技术，容量从 B200 的 192GB 暴涨 50%。
- 8 TB/s 显存带宽：作为对比，H100 仅为 3.35
  TB/s。这极大地缓解了在推理大模型时数据读取的瓶颈。
- 实战意义：单台 8 卡 HGX B300 服务器的总显存高达 2.3 TB，能够支持更大 Batch Size 的推理，或在单机上跑通原本需要多机集群才能承载的超大型模型。

3. 计算规格与精度：FP4 时代的巅峰

B300 进一步挖掘了 第二代 Transformer Engine 的潜力，特别强化了低精度计算。

- FP4 Tensor Core：单卡 Dense（密集）算力可达 15 PetaFLOPS（Sparse
  模式下翻倍），性能较 B200 提升显著。
- Attention 性能翻倍：针对 Transformer 模型中最耗资源的
  Attention（注意力机制）层，B300 提供了 2 倍于 B200 的加速。
- 精度降维：通过 FP4 精度，B300
  能够以更低的能耗维持极高的模型准确度，大幅降低了单位 Token 的推理成本。

4. B300 vs B200 参数横向对比

关键指标	B200 (Blackwell)	B300 (Blackwell Ultra)	提升表现
显存容量	192GB HBM3e	288GB HBM3e	+50% 空间，更适合长文本
显存带宽	8 TB/s	8 TB/s	维持顶级数据交换速度
FP4 密集算力	10 PetaFLOPS	15 PetaFLOPS	+50% 计算效能
Attention 加速	1.0x 基准	2.0x 提升	核心算法效率质变
最大功耗 (TDP)	1000W - 1200W	1400W	散热要求提高，推崇液冷

5. 集群与系统级方案：GB300 NVL72

B300 并不单兵作战，它通常以 GB300 NVL72 的整机柜形式部署。

- 超级芯片：将 Grace CPU 与 B300 GPU 通过 NVLink-C2C 深度绑定，实现 900GB/s 的内存一致性互联。
- NVLink 5.0：每张卡提供 1.8 TB/s 的互联带宽。在 NVL72 集群中，72 张 GPU 形成一个拥有 20 TB 显存 的“超级 GPU”。
- 联网升级：标配 ConnectX-8 网卡，支持 800G/1.6T 网络，确保卡间通信不再成为万卡集群的瓶颈。

6. 专家总结：为什么 B300 是 2025 年的焦点？

B300 的发布标志着英伟达从“算力提供商”转型为“推理产能工厂”。

1. 推理成本终结者：凭借 FP4 和海量带宽，它能以极低的延迟处理海量并发，是搜索、视频生成、智能体（Agents）的首选。
2. 长文本救星：288GB 显存彻底释放了模型对长上下文的理解力。
3. 行业门槛：高达 1400W 的功耗使得全液冷（Liquid-to-Chip）正式成为顶级数据中心的入场券。