昇腾AI芯片与英伟达GPU的技术对比

2025年5月20日　2 分钟阅读

#GPU #昇腾 #Huawei

> *昇腾AI芯片与英伟达GPU的技术对比*

以下是华为昇腾AI芯片与英伟达GPU的技术对比：

各项指标对标

1. 核心架构与算力对比

技术维度	昇腾910D / 昇腾920 (2025)	英伟达H100 / H200	性能对比
架构设计	达芬奇3.0 + Chiplet 3D封装	Hopper架构 + CoWoS封装	昇腾集群效率高15%
制程工艺	中芯国际7nm（910D）/ 6nm（920）	台积电4N（H100）/ 台积电CoWoS（H200）	制程差距缩小至1-2代
FP16算力	1.2 PFLOPS (910D) / 2.1 PFLOPS (920)	H100: 67 TFLOPS / H200: 90 TFLOPS	昇腾910D为H100的18倍
BF16算力	300 TFLOPS (910D) / 900 TFLOPS (920)	H100: 198 TFLOPS / H200: 312 TFLOPS	昇腾920为H200的2.9倍
互联带宽	光子互连4TB/s	NVLink 4.0 (900GB/s)	昇腾带宽高4.4倍
延迟	纳秒级（光子互连）	微秒级（电信号）	昇腾延迟降低99%

目录

文章信息

字数

阅读时间

发布时间

更新时间

标签

#GPU #昇腾 #Huawei

2. 能效与散热对比

技术维度	昇腾910D	英伟达H100	能效优势
功耗	310W	700W	昇腾功耗降低55%
能效比	2.1 TFLOP/W（FP16）	1.1 TFLOP/W（FP16）	昇腾能效高91%
散热技术	液冷系统（45℃全速运行）	风冷/液冷（需维持60℃以下）	昇腾散热成本低40%
PUE指标	1.1（数据中心）	1.5（传统风冷）	昇腾节能36%

3. 集群性能对比

技术维度	昇腾CM384集群（384颗910D）	英伟达GB200 NVL72集群（72颗H200）	昇腾优势
BF16集群算力	300 PFLOPS	175 PFLOPS	昇腾算力高71%
内存容量	12TB（HBM3）	3.3TB（HBM3）	昇腾内存高3.6倍
训练稳定性	28天连续运行（GPT-4级模型）	15天（需中断维护）	昇腾稳定性高87%
部署成本	5000万元/集群	1.2亿元/集群	昇腾成本低58%

4. 软件生态与行业应用

技术维度	昇腾生态	英伟达生态	昇腾进展
开发工具	CANN 7.0（对标CUDA）	CUDA 12.0	适配率70% vs 100%
框架支持	MindSpore / PyTorch / TensorFlow	TensorFlow / PyTorch	昇腾覆盖90%主流框架
行业案例	160+大模型（盘古、DeepSeek等）	GPT-4 / LLaMA	昇腾国内市占率80%
开发者迁移成本	单模型迁移平均3人天	CUDA原生支持	昇腾效率提升30%

5. 供应链与成本

技术维度	昇腾910D	英伟达H100	昇腾优势
单卡售价	14.5万元	24万元	价格低40%
国产化率	85%（芯片+工具链）	美国技术占比100%	完全规避制裁限制
代工厂	中芯国际（7nm/6nm）	台积电（4N工艺）	自主可控
HBM供应商	长鑫存储（HBM3，良率80%）	三星/海力士	打破韩国垄断

总结

通过表格可见，昇腾在**算力密度**（光子互连）、**能效比**（2.1 TFLOP/W）、**国产化率**（85%）和**成本**（价格仅为英伟达60%）等关键指标上形成显著优势，其技术路径已从追赶转向局部领先。

昇腾超大规模MoE模型推理部署技术

https://gitcode.com/ascend-tribe/ascend-inference-cluster

1. 综述部分（Overview）

时间：5月19日

内容：

全面介绍昇腾超大规模MoE模型推理部署方案的整体架构与核心技术。

重点包括推理框架优化、数学到物理的极致实现（如FlashComm）、多流并发、MLA加法优化、创新算子等技术亮点。

2. 昇腾超大规模MoE模型推理负载均衡技术（OmniPlacement）

时间：5月20日

内容：

深入讲解在大EP（Expert Parallelism）场景下，如何实现与昇腾硬件高度亲和的极致负载均衡。

涉及主要组件的技术实现原理，确保大规模推理任务高效分配与调度。

3. 投机推理技术（FusionSpec）

时间：5月21日

内容：

面向高吞吐场景，将原本用于小批量的MTP（Multi-Token Prediction）技术，基于昇腾平台进行适配和创新。

分享创新量化技术，如以Int8达到FP8精度，提升推理效率与精度。

4. 通信优化技术（FlashComm）

时间：5月22日

内容：

聚焦模型侧通信优化，包含AllReduce优化、存换传优化、大EP四流水并行等多项前沿技术。

目标是降低通信瓶颈，实现推理过程中的极致并行和吞吐。

5. 昇腾亲和硬件感知创新算子（OptimizedOps）

时间：5月23日

内容：

展示基于昇腾硬件深度定制的大量创新算子实现。

这些算子为模型和框架提供极致性能支撑，是上述所有优化的基础。

总结

本系列报告不仅覆盖了从架构到底层算子的全链路优化思路，还展示了昇腾平台在大模型推理场景下的独特优势和技术突破。相关代码也将陆续开源，推动整个生态发展。

分享这篇文章