英伟达推出新一代计算架构，关注 AI 算力及应用部署

——电子行业跟踪报告

强于大市(维持)

2024 年 03 月 22 日

行业核心观点：

北京时间 3 月 19 日凌晨，英伟达创始人黄仁勋在 2024 年英伟达 GTC 大会现场发表演讲，发布了新一代计算架构 Blackwell 及系列芯片产品，并展示部分领域应用的进展。基于 Blackwell 架构的 AI 芯片及相关硬件设备的算力性能提升明显，同时围绕 CUDA GPU 生态，英伟达积极推进 AI 应用端部署，推动 AI 算力及应用产业链的发展。

投资要点：

英伟达发布 Blackwell 系列 GPU，多方面升级提高算力：英伟达发布了新一代计算架构 Blackwell，以及采用 Blackwell 架构的 GPU B200 及 GB200，在 FP4 精度下，Blackwell 架构的 AI 计算性能达到 Hopper 架构的 5 倍。Blackwell 系列 GPU 采用台积电 4NP 工艺，集成了 2080 亿颗晶体管，且升级了 Transformer 引擎、NVLink 等以提升算力。相对上一代 Hopper 架构，Blackwell 架构的集群化计算降低了能源消耗及所需的 GPU 数量，有望降低计算成本。过去，在 90 天内训练一个 1.8 万亿参数的 MoE 架构 GPT 模型，需要 8000 个 Hopper 架构 GPU，15 兆瓦功率；如今，在 Blackwell 架构下进行训练，同样 90 天时间的情况下只需要 2000 个 GPU，以及 1/4 的能源消耗。

围绕 CUDA GPU 生态，英伟达积极推动 AI 应用部署：1）大模型领域，在 CUDA GPU 基础上推出企业级生成式 AI 服务，进一步推动模型本地部署；英伟达 NIM 是英伟达推理微服务的代表产品，在英伟达大型 CUDA 安装基础上工作，企业可使用这些微服务在自己的平台上创建和部署自定义应用程序，使开发人员能够将部署时间从几周缩短到几分钟；**2）芯片制造领域，**光刻计算库 cuLitho 通过生成式 AI 算法将工作流速度提升 2 倍，并已投入使用，随着 EDA 巨头新思科技将该技术集成到其软件工具中，cuLitho 也可能会渗透到其他芯片设计厂商；**3）MR 领域，**英伟达与苹果强强联合，将 Omniverse 平台引入 Vision Pro，让开发者在工业元宇宙里利用空间计算进行作业。

投资建议：英伟达发布新一代计算架构及芯片产品，积极推动 AI 应用部署，建议关注 AI 算力及应用产业链的投资机遇。**1）AI 算力领域，**英伟达引领 AI 芯片技术创新，算力产业链上下游厂商充分受益，建议关注上游 HBM、先进封装等细分优质赛道；同时国内政策引导及 AI 产业浪潮有望加速国内 AI 算力自主可控进程，建议关注国产算力产业链的龙头公司；**2）AI 应用部署方面，**英伟达展示大模型、芯片制造及 MR 等领域的应用，积极推动 AI 赋能千行百业，AI 应用的部署有望提升企业生产力，建议关注前瞻布局 AI 应用领域的优质公司

风险因素：AI 应用发展不及预期；AI 需求不及预期；算力建设进程不及预期；市场竞争加剧。

行业相对沪深 300 指数表现

数据来源：聚源，万联证券研究所

相关研究

台积电继续扩大先进封装产能，关注本周英伟达 GTC 大会

行业巨头持续加码 AIPC，存储产业营收有望保持增长

加快发展新质生产力，推动高水平科技自立自强

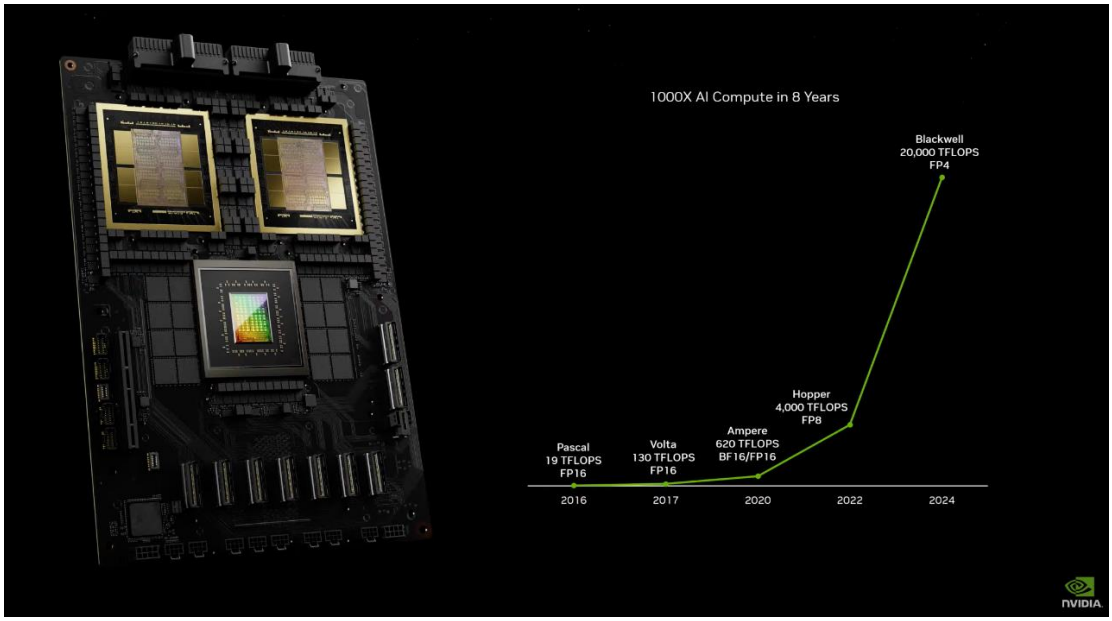
分析师： 夏清莹
执业证书编号：S0270520050001
电话：075583223620
邮箱：xiaqy1@wlzq.com.cn
研究助理： 陈达
电话：13122771895
邮箱：chenda@wlzq.com.cn

正文目录

1 英伟达发布 Blackwell 系列 GPU，多方面升级提高算力

英伟达2024 GTC大会发布新一代计算架构及芯片产品，算力达到上一代产品的5倍。北京时间3月19日凌晨，英伟达创始人黄仁勋在2024年英伟达GTC大会现场发表演讲，并发布了新一代计算架构Blackwell，以及采用Blackwell架构的GPU，分为B200和GB200产品系列，后者集成了1个Grace CPU和2个B200 GPU；其中B200 GPU拥有2080亿个晶体管，并以10TBps的互联速度将两块小芯片合在一起，大幅提高处理能力，提供高达20petaflops的FP4吞吐量；而GB200 GPU通过900GB/秒的超低功耗芯片连接方式，将两个B200 GPU连接到1个Grace CPU上。在FP4精度下，Blackwell架构的AI计算性能达到Hopper架构的5倍，经过8年时间的发展，英伟达AI算力实现了1000倍的增长。

图表1：GB200 同过去架构的 AI 芯片性能对比



资料来源：机器之心，万联证券研究所

Blackwell GPU在晶体管数量、Transformer引擎、NVLink方面均有所提升。1）晶体管承载方面，Blackwell GPU采用的台积电定制工艺从上一代4N升级至4NP，采用统一内存架构及双芯配置，将2个受光刻模板（reticle）限制的GPU die通过10TB/s芯片间接口连成一个统一GPU，集成了2080亿颗晶体管，共有192GB HBM3e内存、8TB/s显存带宽；**2）Transformer引擎方面，**Blackwell GPU搭载第二代Transformer引擎，采用新的微张量扩展支持和集成到英伟达TensorRT-LLM和NeMo Megatron框架中的先进动态范围管理算法，使Blackwell具备在FP4精度的AI推理能力，可支持2倍的计算和模型规模，能在将性能和效率翻倍的同时保持混合专家模型的高精度；**3）互联方面，**Blackwell GPU采用第五代NVLink，新一代NVLink为每个GPU提供1.8TB/s双向带宽，支持多达576个GPU间的无缝高速通信，适用于复杂大语言模型；**4）其次，**还有RAS引擎、安全AI、解压缩引擎等方面的优势。

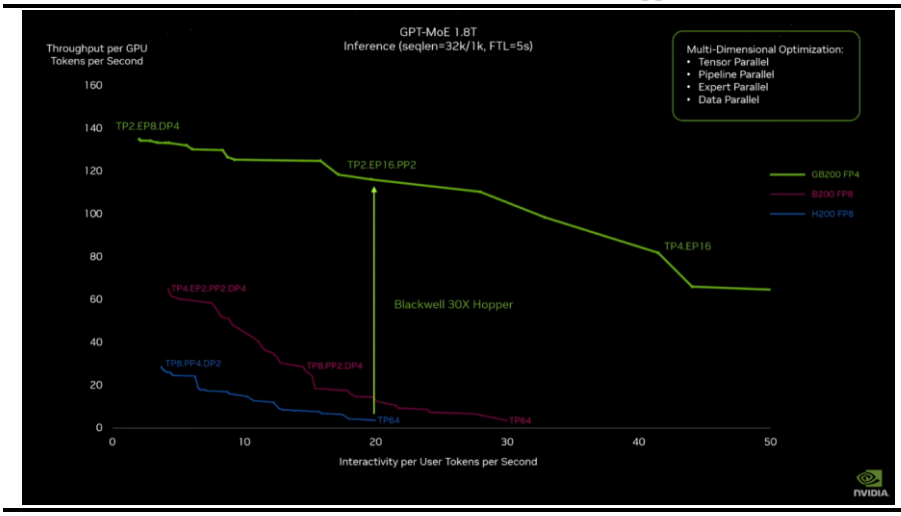
图表2：Blackwell 系列 GPU 与 H100 对比

	GB200	B200	B100	H100
Memory Clock	8Gbps HBM3E	8Gbps HBM3E	8Gbps HBM3E	5.23Gbps HBM3
Memory Bandwidth	2x8TB/sec	8TB/sec	8TB/sec	3.35TB/sec
VRAM	384GB	192GB	192GB	80GB
Interconnects	2xNVLink 5(1800GB/sec)	NVLink 5(1800GB/sec)	NVLink 5(1800GB/sec)	NVLink 4(900GB/sec)
GPU	2xBlackwell GPU	Blackwell GPU	Blackwell GPU	GH100
GPU Transistor Count	416B	208B	208B	80B
TDP	2700W	1000W	700W	700W
Manufacturing Process	TSMC 4NP	TSMC 4NP	TSMC 4NP	TSMC 4N
Architecture	Grace+Blackwell	Blackwell	Blackwell	Hopper

资料来源：全球半导体观察，Trendforce，万联证券研究所

Blackwell架构的集群化计算相对降低了能源消耗及所需的GPU数量，有望降低计算成本。1）超级计算机的配置方面，36颗NVIDIA Grace CPU和72块Blackwell GPU通过第五代NVLink连接成一台超级计算机DGX GB200，而8个或以上的DGX GB200系统将构建成DGX SuperPOD，这些系统通过NVIDIA Quantum InfiniBand进行网络连接，可扩展到数万个GB200超级芯片。DGX GB200 SuperPod采用新型高效液冷机架规模架构，标准配置可在FP4精度下提供11.5 Exaflops算力和240TB高速内存，还支持增加额外的机架扩展性能。**2）实践测试方面，**在具有1750亿个参数的GPT-3 LLM基准测试中，GB200的性能是H100的7倍，并且训练速度是H100的4倍，用于大模型推理的速度是上代的30倍。过去，在90天内训练一个1.8万亿参数的MoE架构GPT模型，需要8000个Hopper架构GPU，15兆瓦功率；如今，在Blackwell架构下进行训练，同样90天时间的情况下只需要2000个GPU，以及1/4的能源消耗。

图表3：Blackwell 系列用于大模型推理的速度是 Hopper 的 30 倍



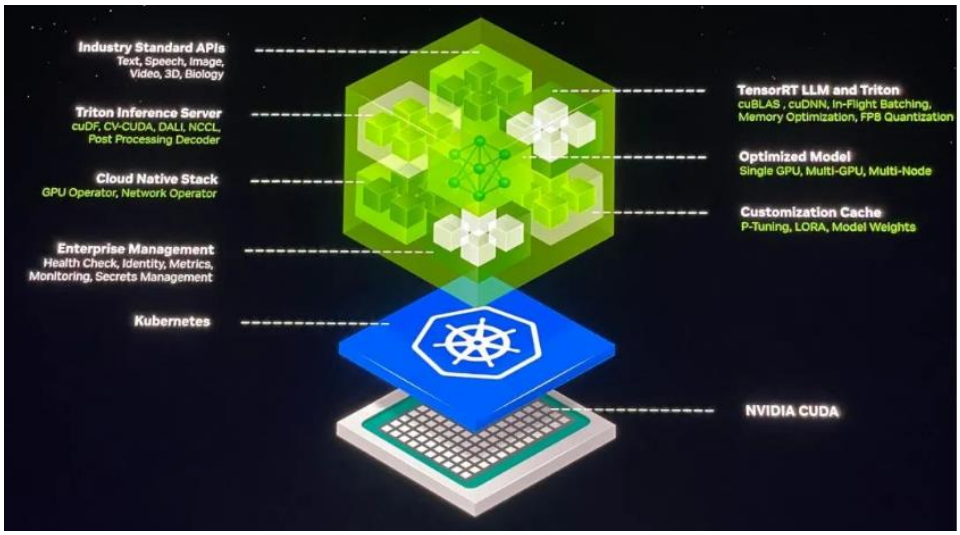
资料来源：机器之心，万联证券研究所

2 围绕 CUDA GPU 英伟达积极推动 AI 应用部署

在CUDA GPU基础上推出企业级生成式AI服务，进一步推动模型本地部署。英伟达继续扩大凭借CUDA和生成式AI生态积累的优势，推出数十个企业级生成式AI微服务，以

便开发者在英伟达CUDA GPU安装基础上创建和部署生成式AI Copilots。英伟达NIM是英伟达推理微服务的代表产品，是由英伟达的加速计算库和生成式AI模型构建的。微服务支持行业标准的API，在英伟达大型CUDA安装基础上工作，并针对新的GPU进行优化。企业可使用这些微服务在自己的平台上创建和部署自定义应用程序，同时保留对其知识产权的完全所有权和控制权。NIM微服务提供由英伟达推理软件支持的预构建生产AI容器，使开发人员能够将部署时间从几周缩短到几分钟。NIM微服务可用于部署来自英伟达、AI21、Adept、Cohere、Getty Images、Shutterstock的模型，以及来自谷歌、Hugging Face、Meta、微软、Mistral AI、Stability AI的开放模型。

图表4：英伟达展示在 CUDA 上部署生成式 AI 服务



资料来源：智东西，万联证券研究所

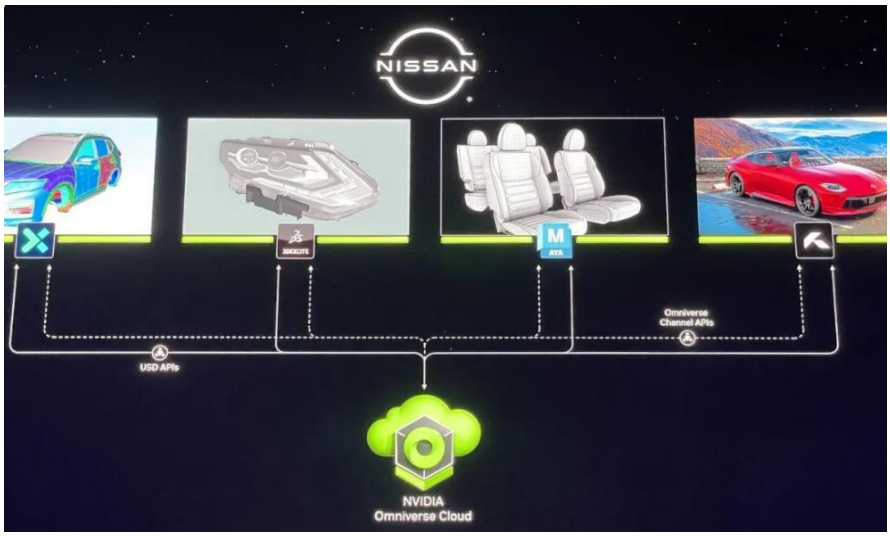
芯片制造领域，光刻计算库cuLitho通过生成式AI算法将工作流速度提升2倍，并已投入使用。英伟达在2023年GTC大会上发布了cuLitho，今年的更新是在cuLitho加速流程的基础上，通过生成式AI算法将工作流的速度又提升了2倍。在芯片制造过程中，计算光刻是计算最密集的工作负载，每年在CPU上消耗数百亿小时。相比基于CPU的方法，基于GPU加速计算光刻的库cuLitho能够改进芯片制造工艺，通过加速计算，350个英伟达H100系统可取代40,000个CPU系统，大幅提高了吞吐量，加快生产，降低成本、空间和功耗。随着EDA巨头新思科技将该技术集成到其软件工具中，cuLitho也可能会渗透到其他芯片设计厂商。

图表5：英伟达展示与 EDA 巨头新思科技的合作

资料来源：智东西，万联证券研究所

MR领域，英伟达与苹果强强联合，将Omniverse平台引入Vision Pro。本次大会中，英伟达特别宣布了与苹果在Vision Pro方面的合作，让开发者在工业元宇宙里利用空间计算进行作业。面向工业数字孪生应用，英伟达将以API形式提供Omniverse Cloud，开发人员可借助该API将交互式工业数字孪生流传输到VR头显中。通过使用API，开发者能轻松地将Omniverse的核心技术直接集成到现有的数字孪生设计与自动化软件应用中，或是集成到用于测试和验证机器人或自动驾驶汽车等自主机器的仿真工作流中。

图表6：英伟达展示 Omniverse Cloud 服务



资料来源：智东西，万联证券研究所

3 投资建议

英伟达发布新一代计算架构及芯片产品，积极推动AI应用部署，建议关注AI算力及应用产业链的投资机遇。

1）AI算力领域，英伟达引领AI芯片技术创新，算力产业链上下游厂商充分受益，建议关注上游HBM、先进封装等细分优质赛道；同时国内政策引导及AI产业浪潮有望加速国内AI算力自主可控进程，建议关注国产算力产业链的龙头公司。

2）AI应用部署方面，英伟达展示大模型、芯片制造及MR等领域的应用，积极推动AI赋能千行百业，AI应用的部署有望提升企业生产力，建议关注前瞻布局AI应用领域的优质公司。

4 风险因素

AI应用发展不及预期；AI需求不及预期；算力建设进程不及预期；市场竞争加剧。

行业投资评级

强于大市：未来6个月内行业指数相对大盘涨幅10％以上；
同步大市：未来6个月内行业指数相对大盘涨幅10％至-10％之间；
弱于大市：未来6个月内行业指数相对大盘跌幅10％以上。

公司投资评级

买入：未来6个月内公司相对大盘涨幅15％以上；
增持：未来6个月内公司相对大盘涨幅5％至15％；
观望：未来6个月内公司相对大盘涨幅-5％至5％；
卖出：未来6个月内公司相对大盘跌幅5％以上。
基准指数 ： 沪深300指数

风险提示

我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

证券分析师承诺

本人具有中国证券业协会授予的证券投资咨询执业资格并登记为证券分析师，以勤勉的执业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

免责条款

万联证券股份有限公司 研究所
上海浦东新区世纪大道 1528 号陆家嘴基金大厦
北京西城区平安里西大街 28 号中海国际中心
深圳福田区深南大道 2007 号金地中心
广州天河区珠江东路 11 号高德置地广场