信息技术行业动态点评

英伟达发布AI超级计算机，智能算力浪潮汹涌

2023 年 05 月 31 日

【事项】

◆ 5 月 29 日，英伟达在当日的 COMPUTEX 大会推出了 NVIDIA DGX GH200 人工智能（AI）超级计算机，这款计算机由 NVIDIA GH200 Grace Hopper 超级芯片和 NVIDIA NVLink Switch System 提供支持。DGX GH200 集成了 256 颗 GH200 Grace Hopper 超级芯片，拥有 144TB 共享内存，可以为大型生成式人工智能模型以及其他应用提供高达 1exaflop 的计算能力。英伟达 CEO 黄仁勋在大会上宣布，公司的 Grace Hopper 超级芯片现已全面投产。

黄仁勋表示，DGX GH200 AI 超级计算机集成了英伟达最先进的加速计算和网络技术，以拓展人工智能的前沿领域。同时，英伟达正致力于使 DGX GH200 在今年年底上市。谷歌云、Meta 和微软将会是首批有望获得 DGX GH200 访问权的公司。

东方财富证券 Eastmoney Securities

挖掘价值 投资成长

强于大市（维持）

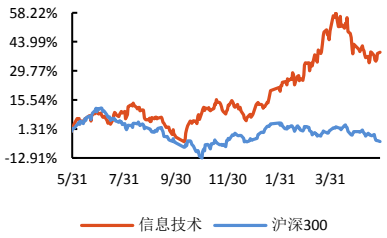
东方财富证券研究所

证券分析师：方科

证书编号：S1160522040001

电话：021-23586361

相对指数表现



相关研究

《华为完成首批 MetaERP 切换，ERP 国产替代砥砺前行》

2023.05.25

《机器人——强 AI 的承载体》

2023.03.27

《AIGC 再下一城，数字虚拟人迎来大爆发》

2023.03.20

《内容监管——AI 生态良好发展的先决条件》

2023.03.07

《大安全成为新基调，信创、自动驾驶开启新起点》

2023.01.17

【评论】

◆ 自去年年底 OpenAI 发布 ChatGPT 以来，生成式人工智能逐渐确定成为新趋势，而创建文本、图像、视频等内容需要通过超强算力来实现，算力已经成为 AI 的刚需，芯片巨头英伟达生产的人工智能芯片在该领域至关重要。此前，英伟达在 AI 训练端先后推出了 V100、A100、H100 三款芯片，以及为了满足美国标准，向中国大陆销售的 A100 和 H100 的带宽缩减版产品 A800 和 H800。

DGX GH200 人工智能超级计算平台是英伟达针对最高端的人工智能和高性能计算工作负载而设计的系统和参考架构，目前的 DGX A100 系统只能将八个 A100 GPU 联合起来作为一个单元，考虑到生成型人工智能的爆炸式增长，英伟达的客户迫切需要更大、更强大的系统，DGX GH200 就是为了提供最大的吞吐量和可扩展性而设计的，它通过使用英伟达的定制 NVLink Switch 芯片来避免标准集群连接选项（如 InfiniBand 和以太网）的限制。通过 256 块超级芯片组成的 DGX GH200 显然有着超越前代产品 DGX A100 的计算能力。而且，英伟达也正在打造基于 DGX GH200 的大型 AI 超级计算机 NVIDIA Helios，其中采用 4 个 DGX GH200 系统、1024 颗 Grace Hopper 超级芯片，每个都将与英伟达 Quantum-2 InfiniBand 网络连接，带宽高达 400Gb/s，预计于今年年底上线。

芯片巨头的算力迭代极其迅速，说明下游云厂商以及企业侧对于生成式 AI 技术具备强烈需求，相关算力板块（包括 GPU、服务器、光模块、数据中心等）有望具备较大业绩弹性。

◆ AI 芯片是 AI 算力的根基。需求逐渐爆发，数据海量增长，大模型参数趋多，对计算性能要求愈发严格。GPU 相较于 CPU，优势在于并行计算。在大会上，黄仁勋向传统 CPU 服务器集群发起“挑战”，直言在人工智能和加速计算这一未来方向上，GPU 服务器有着更为强大的优势。随着需要大量计算能力的 AI 应用出现，GPU 将成为主角，英伟达主导了当前全球 AI GPU 市场。举例来说，训练一个 LLM 大语言模型，将需要 960 个 CPU 组成的服务器集群，这将耗费大约 1000 万美元（约合人民币 7070 万元），并消耗 11 千兆瓦时的电力。相比之下，同样以 1000 万美元的成本去组建 GPU 服务器集群，将以仅 3.2 千兆瓦时的电力消耗，训练 44 个 LLM 大模型。如果同样消耗 11 千兆瓦时的电量，那么 GPU 服务器集群能够实现 150 倍的加速，训练 150 个 LLM 大模型，且占地面积更小。而当用户仅仅想训练一个 LLM 大模型时，则只需要一个 40 万美元左右，消耗 0.13 千兆瓦时电力的 GPU 服务器即可。相比 CPU 服务器，GPU 服务器能够以 4%的成本和 1.2%的电力消耗来训练一个 LLM，这将带来巨大的成本节省。在大模型时代背景和高景气的需求带动下，GPU 将会成为算力产业链中至关重要不可或缺的一环。

建议关注：
算力芯片：景嘉微、寒武纪、海光信息（电子组覆盖）、云天励飞等
算力服务：中科曙光、浪潮信息、中国长城等
边缘算力：网宿科技、首都在线、润泽科技（未覆盖）、优刻得（未覆盖）等

【风险提示】

人工智能技术落地应用不及预期；
竞争格局恶化；
信创不及预期。

东方财富证券股份有限公司（以下简称“本公司”）具有中国证监会核准的证券投资咨询业务资格

分析师申明：

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

投资建议的评级标准：

报告中投资建议所涉及的评级分为股票评级和行业评级（另有说明的除外）。评级标准为报告发布日后3到12个月内的相对市场表现，也即：以报告发布日后的3到12个月内的公司股价（或行业指数）相对同期相关证券市场代表性指数的涨跌幅作为基准。其中：A股市场以沪深300指数为基准；新三板市场以三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）为基准；香港市场以恒生指数为基准；美国市场以标普500指数为基准。

股票评级

买入：相对同期相关证券市场代表性指数涨幅15%以上；
增持：相对同期相关证券市场代表性指数涨幅介于5%～15%之间；
中性：相对同期相关证券市场代表性指数涨幅介于-5%～5%之间；
减持：相对同期相关证券市场代表性指数涨幅介于-15%～-5%之间；
卖出：相对同期相关证券市场代表性指数跌幅15%以上。

行业评级

强于大市：相对同期相关证券市场代表性指数涨幅10%以上；
中性：相对同期相关证券市场代表性指数涨幅介于-10%～10%之间；
弱于大市：相对同期相关证券市场代表性指数跌幅10%以上。

免责声明：

本研究报告由东方财富证券股份有限公司制作及在中华人民共和国（香港和澳门特别行政区、台湾省除外）发布。

本研究报告仅供本公司的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。

本研究报告是基于本公司认为可靠的且目前已公开的信息撰写，本公司力求但不保证该信息的准确性和完整性，客户也不应该认为该信息是准确和完整的。同时，本公司不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司会适时更新我们的研究，但可能会因某些规定而无法做到。除了一些定期出版的报告之外，绝大多数研究报告是在分析师认为适当的时候不定期地发布。

在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他投资标的的邀请或向人作出邀请。

本报告中提及的投资价格和价值以及这些投资带来的收入可能会波动。过去的表现并不代表未来的表现，未来的回报也无法保证，投资者可能会损失本金。外汇汇率波动有可能对某些投资的价值或价格或来自这一投资的收入产生不良影响。

那些涉及期货、期权及其它衍生工具的交易，因其包括重大的市场风险，因此并不适合所有投资者。

在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任，投资者需自行承担风险。