ChatGPT 专题研究之二：算力的提升是人工智能发展的基石

相关研究：

1、《内需提振+创新驱动，半导体产业结构性复苏在望》 2023.01.30

2、《探寻ChatGPT的能力圈，及“破圈”成长之路》 2023.02.08

行业评级：增持

近十二个月行业表现

%	1 个月	3 个月	12 个月
相对收益	-2.66	-11.2	-8.7
绝对收益	-3.2	-2.5	-19.9

注：相对收益与沪深 300 相比

分析师：王文瑞

证书编号：S0500523010001

Tel：(8621) 50293694

Email：wangwr2@xcsc.com

地址：上海市浦东新区银城路 88 号中国人寿金融中心 10 楼湘财证券研究所

核心要点：

- **算力的提升是人工智能发展的基石**

ChatGPT 的面世将人工智能大模型这一概念引入更多人的视野，大模型具有“可以实现众多场景通用、泛化和规模化复制”等特性，将成为 AI 规模应用的重要途径。人工智能大模型优化演进以人工智能数据中心的建设、CPU、AI 计算加速芯片、数据中心存储等硬件的性能优化为基础。

- **人工智能计算中心建设快速推进，CPU、GPU、闪存等半导体硬件的性能优化是数据中心“计算提速”的基石**

服务器在数据中心硬件成本中的占比约为 70%，CPU、GPU、存储三个模块的芯片成本在更高性能的服务器中成本占比达 50%-80%。IDC 发布的报告显示，2021 年我国 AI 服务器支出规模同比增长 44.5%，是近年来我国投资的重点方向，国内 AI 服务器的渗透率呈稳步提升趋势，会带动底层硬件产品的需求稳定增长。我们预计 2023/2024/2025 应用 AI 服务器领域的 X86 CPU 芯片出货量为 72.55 万片/90.05 万片/108.61 万片；2023 年国内市场规模预计为 43.53 亿美元，年同比增长约为 34.4%。2023/2024/2025 年国内 AI 加速芯片市场规模为 43.37/58.72/76.42 亿美元，市场规模年同比增速分别为 36.3%/35.4%/30.1%。

数据中心存储领域，全闪存存储将取代传统硬盘存储成为数据中心存储主力，优化数据存储容量及应用效率；同时降低数据中心能耗，实现绿色节能。

- **投资建议**

2023 年数字化建设在多领域稳步推进，为产业链发展带来新动能，预期提振多种半导体硬件的市场需求，带动大数据中心的建设加速。建议关注数字经济发展为传感器、CPU、GPU 等领域带来的需求增量。汽车智能化渗透率提升及出口增长驱动板块需求稳步增长，建议关注车规级 MOSFET 及 SIC 功率器件的国产化进程。中长期，人工智能技术的落地商用将持续增多，人工智能技术的发展以算力资源的扩充、CPU&GPU 处理速度、存储器及接口芯片等半导体硬件性能的提升为基础，建议关注 Chiplet 及先进封装，新型存储等先进技术的发展。建议持续关注半导体行业，维持行业增持评级。

- **风险提示**

新产品商用化进程不及预期；市场需求不振；技术研发不及预期；宏观政策变化不及预期。

1 算力的提升是人工智能发展的基石

ChatGPT 的面世也将人工智能大模型这一概念引入更多人的视野，大模型具有“可以实现众多场景通用、泛化和规模化复制”等特性，将成为 AI 规模应用的重要途径。人工智能语言大模型是一种具有复杂结构和大量参数的深度学习模型，这些参数允许模型更好地捕捉和记忆训练数据，减少了对于数据标注的依赖”的特性。根据华为的研究报告显示人工智能大模型的优势在于，随着超大模型的的与训练语言模型的不断提高，模型的“语言理解”基线能力得到提升；在大模型落地应用于其他行业时，只需根据已有的基础模型结合特定行业的领域数据进行调整，即可生成某个领域的相关模型，且得到良好的精度和性能。ChatGPT 就是自然语言应用程序中较为成熟的大模型之一；2021 年以来，国内也陆续发布了一系列大模型，如华为与鹏程实验室联合发布的“鹏程.盘古”系列超大规模预训练稠密模型，中科院自动化所发布了全球首个三模态大模型“紫东.太初”，北京智源人工智能研究院发布的“悟道 2.0”稀疏模型等。

算力的提升是人工智能大模型优化演进的核心因素之一，以人工智能数据中心的建设、CPU、AI 计算加速芯片、数据中心存储等硬件的性能优化为基础。专题一中我们对于 ChatGPT 的局限性及“破圈成长路径”进行梳理分析后发现：数据库更新频率的提高、数据广度的扩展及质量的提升，模型的持续优化等是推动以 ChatGPT 为代表的人工智能大模型持续成长的重要因素，而数据库的扩展、人工智能模型的优化都对算力资源及算力成本提出了更高的要求。华为发布的研究报告显示大规模预训练模型的参数量持续增多，需要的算力也从 TFLOPS（1TFLOPS=一万亿次浮点运算/秒）增加至 PFLOPS（1PFLOPS=一千万亿次浮点运算/秒）级别，多数企业面临算力不足和算力成本昂贵两大难题，当前算力成本占据企业开发成本的 15%-25%。人工智能计算中心的建设、高效运营和可持续发展可以有效推进国内人工智能产业的发展。

2 数据中心建设的基础—半导体硬件

人工智能计算中心（AICC）是专注于 AI 计算的新型城市基础设施，为人工智能模型的训练和部署提供了大量丰富的计算资源，相较于传统数据中心在硬件配置上通常需要高性能 CPU、AI 计算加速芯片，高速存储、大内存和高带宽，以支持复杂的人工智能模型。

数据中心的投资总成本分为投资成本和运营成本两部分， IT 设备购置、服务器、数据存储设备等硬件设施的资金支出占比较高，其中服务器在数据中

心硬件成本中的占比约为 70%。服务器则主要由 CPU,GPU,PCB,DRAM,SSD,BMC（基板管理控制器）。根据 IDC 研究显示，CPU、GPU、存储三个模块的芯片成本在基础型服务器中占比约为 30%；在更高性能的服务器中，芯片成本占比达 50%-80%。

图 1 先进计算技术产业体系架构

资料来源：《中国信通院》，湘财证券研究所整理

2.1 数据中心建设带动 CPU、GPU 需求上行

根据艾瑞咨询统计数据显示，2021 年国内专用 AI 服务器的渗透率为 6%-7%；IDC 发布的报告显示，2021 年我国 AI 服务器支出规模同比增长 44.5%，是近年来我国投资的重点方向，国内 AI 服务器的渗透率呈稳步提升趋势，会带动底层硬件产品的需求稳定增长。

根据 IDC 信息显示，国内的 X86 服务器以双路服务器为主，2016 年至 2020 年占比均高于 80%,IDC 统计显示 2021 年我国 X86 服务器出货量为 375.1 万台，年同比增长约 9.1%，预计 2022 年销量为 408.4 万台，同比增长 8.9%，2023 年出货量增速约为 9%；2021-2025 年均复合增长率为 8.8%。同等平台且同一级别条件下的服务器，服务器上搭载的 CPU 路数与多线程性能正相关，服务器路数越多、服务器计算性能越强，我们认为 2 路和 4 路服务器市场占比会缓慢上升。我们结合 IDC、海光信息的数据资料假设：

- 2023/2024/2025 年双路服务器占比为 89.3%/90%/90%；4 路服务器占比为 4.6%/5.3%/5.3%。
- 2023/2024/2025 我国 AI 服务器的渗透率分别为 8%/9%/10%。

根据 X86 服务器的出货量、服务器路数分布情况，我们预计 2023 年国内

市场 X86 服务器出货量为 906.91 万片，年同比增长约 9.2%;2024 年出货量将增长至 1000.5 万片，年同比增速为 10.3%，2025 年出货量预计为 1086.1 万片，年同比增速约为 8.6%。其中 2023/2024/2025 应用 AI 服务器领域的 X86 CPU 芯片出货量为 72.55 万片/90.05 万片/108.61 万片。

表 1 AI 服务器用 CPU 芯片出货量预测

	2022E	2023E	2024E	2025E
X86 服务器出货量（万台）	408	445	484	525
1 路服务器市场份额占比	6.30%	6.00%	5%	5%
1 路服务器出货量（万台）	25.7	26.7	24.19	26.26
2 路服务器市场份额占比	89.00%	89.3%	90%	90%
2 路服务器出货量（万台）	363.12	397.39	433	470.05
4 路服务器市场份额占比	4.60%	4.60%	5.30%	5.30%
4 路服务器出货量（万台）	18.77	20.47	25.64	27.84
8 路服务器市场份额占比	0.10%	0.10%	0.20%	0.20%
8 路服务器出货量（万台）	0.41	0.45	0.97	1.05
X86 CPU 芯片出货量（万片）	830.28	906.91	1,000.50	1,086.11
X86 CPU 芯片出货量同比		9.2%	10.3%	8.6%
AI 服务器市场渗透率	6.5%	8%	9%	10%
AI 服务器用 CPU 芯片（万片）	53.97	72.55	90.05	108.61

资料来源：IDC、海光信息、艾瑞咨询、湘财证券研究所

基于 2021 年国内市场服务器的市场规模及出货量（250.9 亿美元，391.1 万台），结合 CPU 占服务器成本的 1/3-1/2，则 X86 CPU 单价分别为 5000 美元—6300 美元。由于 AI 计算中心通常使用高性能的处理器，我们假设 AI 服务器用 CPU 芯片价格为 6000 美元/片。则 2023 年国内市场规模预计为 43.53 亿美元，年同比增长约为 34.4%。

表 2 AI 服务器用 CPU 市场规模预测

	2023E	2024E	2025E
AI 服务器用 CPU 芯片（万片）	72.55	90.05	108.61
CPU 市场规模（6300 美元/片）（亿）	43.53	54.03	65.17

资料来源：IDC、海光信息、湘财证券研究所

AI 服务器计算加速芯片的配置数量呈增长趋势。IDC 统计数据显示，2017 年平均每台 AI 服务器配置 4.31 个 GPU,2018 年提高为 5.1 个，到 2019 年就提高至 8.02。

我们假设：

- 2023 年平均每台服务器配备 8.02 个 AI 芯片，随后逐年增加 1 片。
- AI 计算加速芯片单价为 1500 美元，2023 至 2025 年价格年调减 3%。

则 2023/2024/2025 年国内 AI 加速芯片的出货量为 289.1/403.6/552.8 万片，出货量年同比增长率约为 36.3%/39.6%/37%。2023/2024/2025 年的 AI 加速芯片市场规模为 43.37/58.72/76.42 亿美元，年同比增速分别为 36.3%/35.4%/30.1%。IDC 预测，2026 年中国加速计算服务器市场将达到 103.4 亿美元。

表 3 AI 计算加速芯片销量及市场规模预测

	2023E	2024E	2025E
服务器年出货量（万台）	452	498	553
AI 服务器渗透率	8%	9.0%	10%
AI 服务器出货量	36.1	44.8	55.3
AI 芯片搭载量/台	8	9	10
AI 芯片出货量（万片）	289.1	403.6	552.8
出货量同比增长	36.3%	39.6%	37.0%
AI 加速器芯片单价（美元）	1500	1455	1382.25
AI 加速器市场规模（亿美元）	43.37	58.72	76.42
市场规模年均增速	36.3%	35.4%	30.1%

资料来源：芯智讯、湘财证券研究所

服务器的 AI 计算加速芯片由 GPU（GPGPU）、FPGA、ASIC（专用集成电路）这三类产品构成。其中 FPGA 的特性为现场可编程性，用户可根据自身的需求进行芯片的功能配置，具有使用灵活性强，产品开发周期较短，小批量使用成本相对较低，系统扩展性强且在并行运算加速方面表现出色；但其规模效应不及 ASIC（专用芯片），因而常年被用作 ASIC 的小批量替代品。适配于发展迅速、技术迭代尚处于高速阶段、计算任务灵活多变的数据中心业务需求，现阶段已被业界广泛用作计算密集型任务的加速卡。GPGPU 作为运算协处理器，针对不同应用领域的需求，具备更高的浮点运算精度和性能，在能效比、同构数据处理量、大规模部署下的综合成本及通用性等方面更具备优势；被广泛用于商业计算和大数据处理。

IDC 研究报告中显示，2021 年中国加速服务器市场中 GPU 服务器市场份额 90%，占据主导地位。NPU、FPGA 和 ASIC 加速服务器的市场份额约为 11.6%，但市场规模增速接近 43.8%。IDC 预计，GPU 仍将是 AI 计算加速芯片市场上的主流产品，2025 年市场份额占比接近 80%。假设 2023/2024/2025 年 GPU 在该领域的市场份额占比分为别 88%/85%/80%,则 GPU 加速器市场规模 2022 至 2025 年均复合增速为 29.64%。

Frost&sullivan 数据显示，2020 年应用于服务器加速器领域的 FPGA 芯片中国销售额约占中国 FPGA 芯片市场份额的 10.7%。2022 年全国 FPGA 销

售额为 208.8 亿元，预计 2025 年国内 FPGA 的销售额为 332.2 亿元，假设 2022 年应用于数据加速器领域的 FPGA 销售额占比为 10%，2025 年比例 13%。则 2025 年国内该领域的 FPGA 市场规模为 49.8 亿元(约为 7.7 亿美元)。预计 2022 至 2025 年该领域的 FPGA 年均复合增速为 41.8%。

表 4 各类 AI 计算加速芯片市场规模预期

	2023E	2024E	2025E	CAGR
AI 加速器市场规模（亿美元）	43.37	58.72	76.42	33.92%
GPU 计算加速器市场规模（亿美元）	38.166	49.91	61.1	29.64%
FPGA 计算加速器（亿美元，汇率取 6.5）	3.21	5.27	7.7	41.80%
其余计算加速器（亿美元）	2.0	3.5	7.6	96%

资料来源：IDC，Frost&sullivan、湘财证券研究所

2.2 全闪存+分布式存储，助推数据中心存储“量速”齐升

内存和存储（硬盘）也是影响数据中心算力提升的重要因素。内存是服务器系统运行程序和数据的主要工作区域，若内存的容量不足则会导致系统频繁地读写磁盘，导致性能降低。存储则是系统数据存储的主要区域，若存储容量不足，则系统会因磁盘空间不足导致系统性能下降，同时系统的稳定性和可靠性也会受到影响。

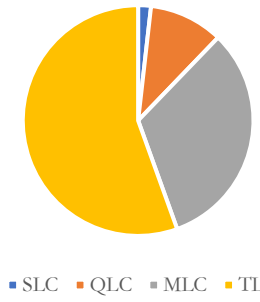
随着 5G、云计算、AI、高性能数据分析等新技术、新应用的发展落地，数据存储需求快速提升，同时非结构化数据的占比也持续提升（如视频，语音，图片，文件等，容量正在从 PB 到 EB 级跨越），对于存储系统性能提出了更高的要求。受存储技术发展和下游需求的驱动，存储行业的发展趋势为：

（1）**全闪存存储取代传统硬盘存储，3D NAND 堆叠层数、存储单元数位量持续增加；**闪存存储与机械磁盘存储相比，具有响应时间短、可用性高、能耗和占用空间低的优势，其中 NAND Flash 是全球市场大容量非易失性存储的主流方案。NAND FLASH 产品的技术路线为：

- 提高存储单元的可存储数位（bit）量，NAND Flash 根据其每个存储单元存储的数据位数/存储密度分为：SLC NAND（1 位/存储单元），MLC NAND（2 位/存储单元），TLC NAND（3 位/单元），QLC NAND（4 位/存储单元）。随着数据存储量的快速提升，TLC 颗粒已经成为企业级 SSD 主流选择，QLC SSD 也已登上市场舞台。
- 提升 NAND Flash 的堆叠层数，由传统的平面结构（2D NAND）演进为多层垂直堆叠结构（3D NAND），提升了 NAND 容量的同时

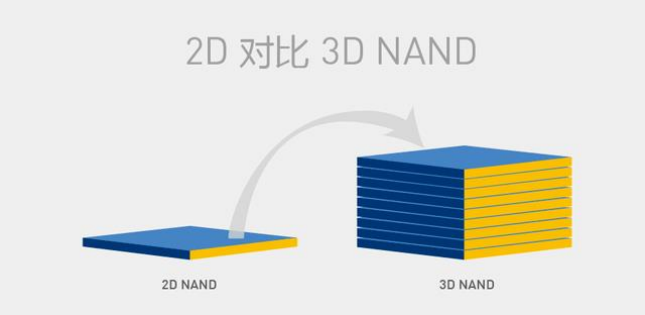
实现了更低功耗、更佳的耐用性。目前 3D NAND 已进入 200 层+时代，镁光、SK 海力士及长江存储都发布了 200+层 NAND 存储。

图 2 2020 年 NANDS 闪存细分类型结构占比



资料来源：Trendforce、湘财证券研究所

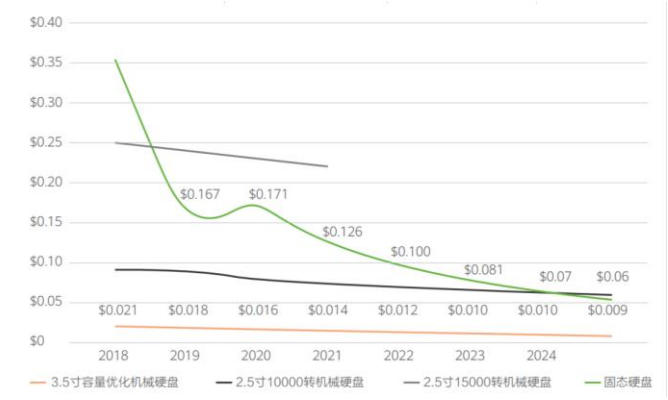
图 3 2D NAND VS 3D NAND



资料来源：智能计算芯世界、湘财证券研究所

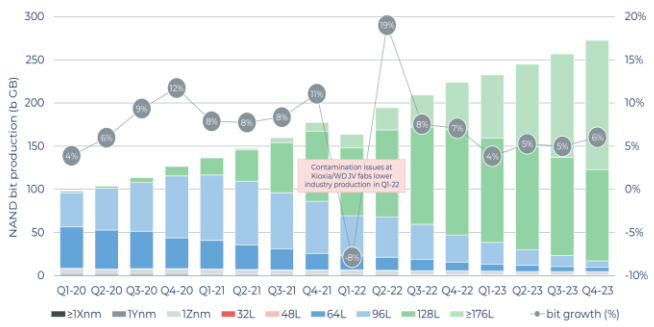
SSD 单盘价格随着堆叠层数的增加和出货量的提升将逐步降低。根据 IDC 预测，2025 年 SSD 单位容量价格将低于 10K 转速的 HDD。性能优势叠加成本的下滑，Trendforc 预计 2023 年服务器领域存储容量总需求增长率约为 44.6%。智能手机、数据中心等下游需求端的数据存储容量持续上行，多层数&大容量 3D NAND 产品市场份额稳步提升。

图 4 IDC 预测 SSD 单位容量价格走势



资料来源：华为技术、湘财证券研究所

图 5 NAND FLASH 各类产品出货量预期



资料来源：Trendforce、湘财证券研究所

（2）**分布式存储携手多样化数据加速引擎，提升数据应用效率**：人工智能大模型中的数据多以文本、图像、音频、视频、社交媒体帖子、电子邮件等形式存在，为非结构化数据；非结构化数据通常具有大量的不确定性和复杂性，在数据处理和存储层面都有更高的要求；数据处理方面需要通过自然语言处理将非结构化数据转化为结构化数据；存储方面，非结构化数据需要占据更大的存储空间。OpenAI 采用了分布式存储来存储和处理 ChatGPT 模型的数据，分布式存储是一种将数据分散存储在多台计算机节点上的存储方式。在分布式存储系统中，多个存储节点协同工作，将数据划分成多个块并存储在不同

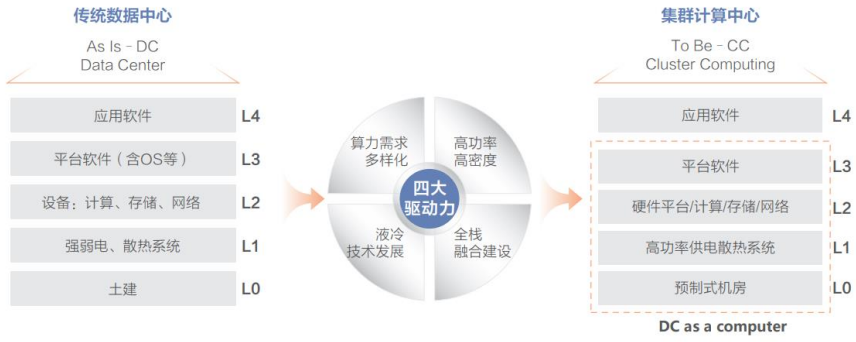
的节点上，这些节点之间通过网络连接互相通信协作，实现数据的分发、备份和恢复等功能。分布式存储具有高可用、高容错、高扩展性等特点，可为大规模数据处理提供良好的基础设施支持，通常需要大量的 DRAM 内存和 SSD 硬盘来实现高性能和高可靠性的存储和访问。

3 数据中心建设与双碳

数据中心是耗能大户，据数字能源产业智库预测，全球数据中心能耗将从 2020 年的 6700 亿度电，增长至 2025 年的 9500 亿度电，占全球用电量的比例约为 3%。数据中心的节能可以通过数据中心冷却系统的优化，数据中心供配电系统的优化，及存储硬件产品优化等方面实现。

其中数据中心供电系统包括变配电系统、备用发电机组、不间断电源系统、机柜配电系统、照明及建筑电气系统等，各部分都可以通过选用自损耗较小、能源效率高的设备，高效变压设备，高压直流系统和锂电池实现节能；如高频化、智能化、绿色节能的不间断电源系统（UPS）就已经获得了市场的认可。数据中心存储领域，华为技术研究显示，相同容量的闪存盘相较于机械硬盘能耗降低约 70%。相同条件下，全闪存数据中心的能耗则比传统数据中心能耗下降约 21%。

图 6 数据中心实现绿色节能



资料来源：《华为技术》，湘财证券研究所整理

4 投资建议

2023 年数字化建设在多领域稳步推进，为产业链发展带来新动能，预期提振多种半导体硬件的市场需求，带动大数据中心的建设加速。建议关注数字经济发展为传感器、CPU、GPU 等领域带来的需求增量。汽车智能化渗透率提升及出口增长驱动板块需求稳步增长，建议关注车规级 MOSFET 及 SIC 功率器件的国产化进程。中长期，人工智能技术的落地商用将持续增多，人工智

能技术的发展以算力资源的扩充、CPU&GPU 处理速度、存储器及接口芯片等半导体硬件性能的提升为基础，建议关注 Chiplet 及先进封装，新型存储等先进技术的发展。建议持续关注半导体行业，维持行业增持评级。

5 风险提示

新产品商用化进程不及预期；市场需求不振；技术研发不及预期；宏观政策变化不及预期。

分析师声明

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注册为证券分析师，以独立诚信、谨慎客观、勤勉尽职、公正公平准则出具本报告。本报告准确清晰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。

湘财证券投资评级体系（市场比较基准为沪深 300 指数）

买入： 未来 6-12 个月的投资收益率领先市场基准指数 15%以上；

增持： 未来 6-12 个月的投资收益率领先市场基准指数 5%至 15%；

中性： 未来 6-12 个月的投资收益率与市场基准指数的变动幅度相差-5%至 5%；

减持： 未来 6-12 个月的投资收益率落后市场基准指数 5%以上；

卖出： 未来 6-12 个月的投资收益率落后市场基准指数 15%以上。

重要声明

湘财证券股份有限公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可。

本研究报告仅供湘财证券股份有限公司的客户使用。本公司不会因接收人收到本报告而视其为客户。

本报告由湘财证券股份有限公司研究所编写，以合法地获得尽可能可靠、准确、完整的信息为基础，但对上述信息的来源、准确性及完整性不做任何保证。湘财证券研究所将随时补充、修订或更新有关信息，但未必发布。

在任何情况下，报告中的信息或所表达的意见仅供参考，并不构成所述证券买卖的出价或征价。本公司及其关联机构、雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。投资者应明白并理解投资证券及投资产品的目的和当中的风险。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎抉择。

在法律允许的情况下，我公司的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。