谁是中国 AI 算力资源“自如”

2024 年 08 月 17 日

市场回顾

本周（8.12-8.16）本周沪深 300 指数上涨 0.42%，中小板指数下跌 0.88%，创业板指数下跌 0.26%，计算机（中信）板块上涨 0.11%。板块个股涨幅前五名分别为：优博讯、安博通、雷柏科技、GQY 视讯、朗科科技；跌幅前五名分别为：任子行、威创股份、*ST 博信、汇金股份、彩讯股份。

行业要闻

- 中国信通院：中国信息通信研究院云计算与大数据研究所云应用与服务团队组织开展研究，编写《中国企业级 SaaS 产业发展研究报告（2024 年）》报告。
- 华为：华夏银行股份有限公司与华为技术有限公司在深圳举行战略合作签约仪式。

公司动态

- 萤石网络：8 月 12 日消息，公司董事长兼总经理蒋海青先生以集中竞价方式增持公司 243,415 股份，增股后占持股 2,085,815 股，直接持有公司股份的 0.2649%，蒋海青先生增持计划已实施完成
- 艾融软件：8 月 15 日消息，持股 5%以上股东孟庆有持有 25,681,690 股，占公司股份总数 12.2079%，拟通过大宗交易方式减持股份不高于 6,000,000 股，不高于公司股份总数 2.8521%

本周观点

- “自如式”算力服务能够有效解决大模型训练、部署、应用环节的痛点，或成为全新范式。作为“自如式”算力服务的领导者，英伟达于 GTC2024 大会提出的“AI Foundry”具备可随时随地部署、可使用行业标准 API 进行开发、满足特定领域的模型需求、优化的推理引擎、支持企业级 AI、支持多领域 AI 模型等优势，能够显著降低千行百业 AI 应用落地的门槛；同时，阿里云与国内几家头部大模型厂商合作，通过在集群架构、功耗散热、资源利用、网络通信、模型算法的综合优化，在模型训练和推理上实现了显著的效能提升，国内“自如式”算力服务渗透率有望进入快速增长的拐点，国内智算中心如火如荼建设提供蓝海市场，建议关注慧辰股份、恒为科技、浪潮信息、中科曙光、星网锐捷、网宿科技等具备 AI 模型调优与算力调优相结合能力的行业龙头。

风险提示

- 政策落地不及预期；行业竞争加剧。

推荐 维持评级

分析师 吕伟
执业证书：S0100521110003
邮箱：lvwei_yj@mszq.com

相关研究

1.计算机周报 20240810：科技内需为王：再次强调自主可控是确定主线-2024/08/10
2.计算机行业事件点评：政策新方向：算力与电力协同-2024/08/06
3.计算机周报 20240804：科技内需为王向工业软件与汽车基础软硬件演绎-2024/08/04
4.计算机行业点评：网络身份认证打开蓝海市场-2024/07/28
5.计算机周报 20240727：科技内需为王，信创风云再起-2024/07/27

目录

1 本周观点 3
1.1 “自如式”算力服务有望解决 AI 应用落地的“最后一公里”难题 3
1.2 英伟达 AI Foundry 引领“自如式”算力服务发展 4
1.3 “自如式”算力服务有望进入快速渗透期 7
1.4 国内 AI 模型与算力协同优化标杆项目持续落地 10
1.5 投资建议 14
2 行业新闻 15
3 公司新闻 16
4 本周市场回顾 17
5 风险提示 19
附录 20
插图目录 21
表格目录 21

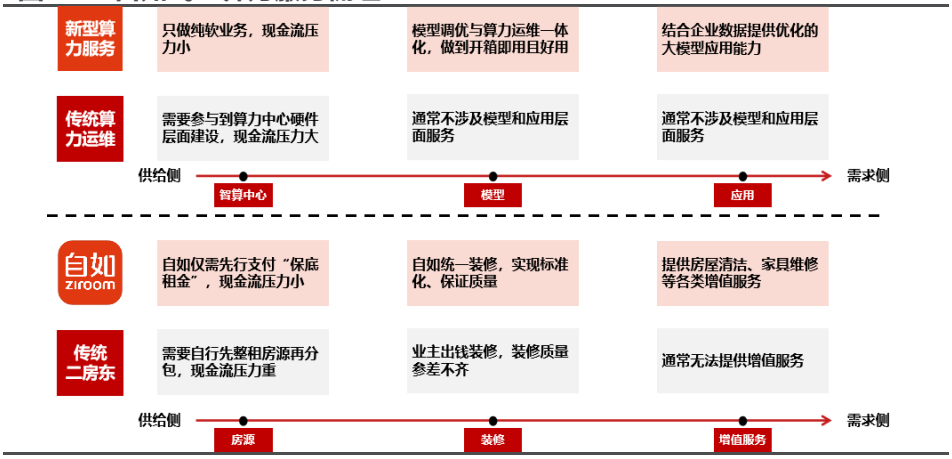
1 本周观点

1.1 “自如式”算力服务有望解决 AI 应用落地的“最后一公里”难题

当前从算力平台建设到模型应用部署面临诸多挑战，新型算力服务呼之欲出。据华为 ICT 服务与软件微信公众号，区别于传统计算业务，新兴的智能算力平台需要从规划、建设、集成、模型训练到推理的落地，整个过程是一个复杂的系统工程，需要包括：1）大规模集群、软硬一体强耦合的复杂交付，大幅提升了算力平台的设计与实施难度，以及成本、高能耗等挑战；2）模型训练底层机制，理论上决定了训练中断是不可避免，如何稳定训练的时长，故障快速恢复也是重点考虑的问题；3）新兴技术领域，各类软硬件技术都在快速迭代，客户模型训练和应用开发过程中，对底层软硬件的适配调优及专业人才获取上也面临巨大的挑战。

“自如式”算力服务能够高效整合算力资源并灵活满足客户在不同层次的需求，解决 AI 应用落地的“最后一公里”难题。我们认为租房服务龙头自如成功的核心是专注赚装修和增值服务的钱，而不是像传统二房东一样赚信息差带来的房租差价，因此自如平台的房屋采取统一装修的标准化运营，并提供如房屋清洁、家具维修等各类增值服务；如果把智算中心比作房源，传统的算力运维就像传统二房东一样通过重资产的模式参与智算中心建设，而对后续应用开发的模型调优、应用开发等环节渗透不足，而新型算力服务将业务重点着手于模型调优与算力运维一体化的纯软业务，同时具备在应用侧赋能企业定制开发的高业务扩展性。

图1：“自如式”算力服务梳理

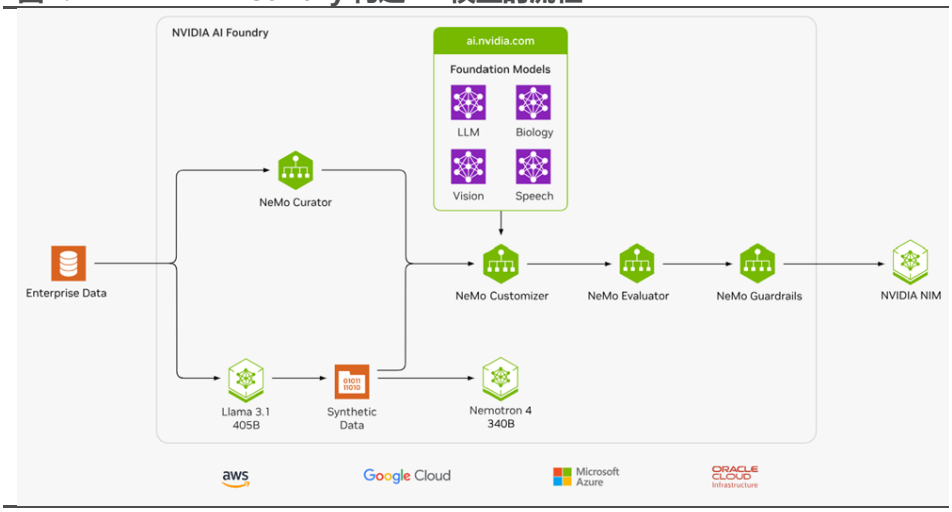


资料来源：民生证券研究院绘制

1.2 英伟达 AI Foundry 引领“自如式”算力服务发展

NVIDIA AI Foundry 是一项使企业能够使用数据、加速计算和软件工具创建并部署自定义模型的服务，用于使用企业数据和特定领域的知识构建自定义生成式 AI 模型。如同台积电代工生产其它企业设计的芯片一样，NVIDIA AI Foundry 使企业能够开发自己的 AI 模型：芯片代工可提供最先进的晶体管技术、制程、大型晶圆代工厂、专业知识以及包含第三方工具和资料库提供商的多元化生态系统；而英伟达 AI Foundry 包括英伟达创建的 AI 模型，如 Nemotron 和 Edify 等流行的开源基础模型，用于定制模型的 NVIDIA NeMo™软件，以及由英伟达 AI 专家构建和支持的英伟达 DGX Cloud 专用计算能力。 英伟达 AI Foundry 的输出形式是一个 NVIDIA NIM™ 推理微服务，包括自定义模型、经过优化的引擎和标准 API——可随时随地部署。

图2：NVIDIA AI Foundry 构建 AI 模型的流程



资料来源：NVIDIA 官网，民生证券研究院

NVIDIA AI Foundry 通过提供模型调优与算力调优的一站式服务，赋能 AI 走入千行百业。NVIDIA NIM 作为 NVIDIA AI Enterprise 的一部分，使得企业可以在云、数据中心、工作站和 PC 上运行 AI 模型。NVIDIA AI Foundry 使用企业数据以及合成生成的数据来增强和改变预训练基础模型中包含的一般知识。一旦模型经过定制、评估等流程后，就会作为 NIM 推理微服务输出。NIM 微服务通过打包算法、系统和运行时优化并添加行业标准 API 来简化 AI 模型部署流程。这使得开发者无需大量定制或专业知识就能够将 NIM 集成到其现有应用程序和基础设施中。借助 NIM，企业可以优化其 AI 基础架构，以更大限度地提高效率和成本效益，有助于提高性能和可扩展性，同时降低硬件和运营成本。总体而言 NVIDIA NIM 具有以下几点核心优势：

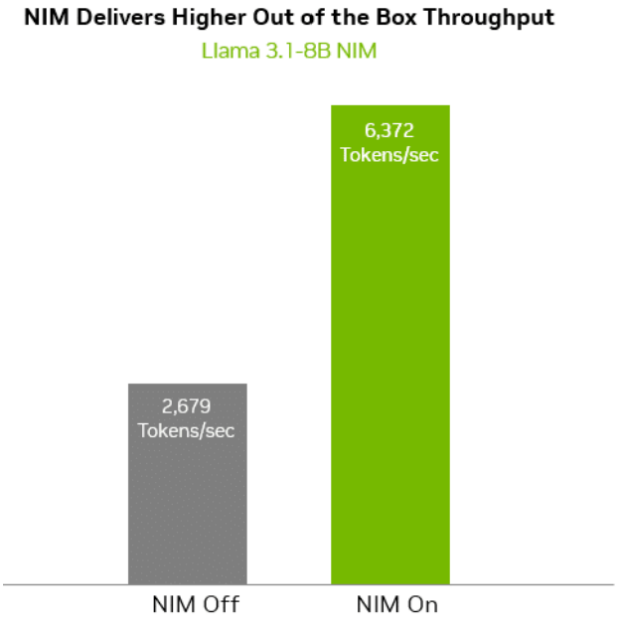
1）可随时随地部署：NIM 专为可移植性和可控性而构建，支持跨各种基础设施（从本地工作站到云再到本地数据中心）进行模型部署。其中包括 NVIDIA DGX、 NVIDIA DGX 云、 NVIDIA 认证系统、 NVIDIA RTX 工作站和 PC；

2）可使用行业标准 API 进行开发：开发者可以通过符合每个领域行业标准的 API 访问 AI 模型，从而简化 AI 应用的开发；

3）满足特定领域的模型需求：NIM 还通过几个关键功能满足了对特定领域解决方案和优化性能的需求。包含特定于领域的 NVIDIA CUDA 库，以及为语言、语音、视频处理、医疗健康等各个领域量身定制的专用代码；

4）优化推理引擎：NIM 针对每个模型和硬件设置利用经过优化的推理引擎，在加速基础设施上提供尽可能低的延迟。降低了在扩展推理工作负载时运行推理工作负载的成本，改善了最终用户体验。

图3：NVIDIA NIM 减少了 Llama 3.1 模型的推理延迟、更快地生成 token

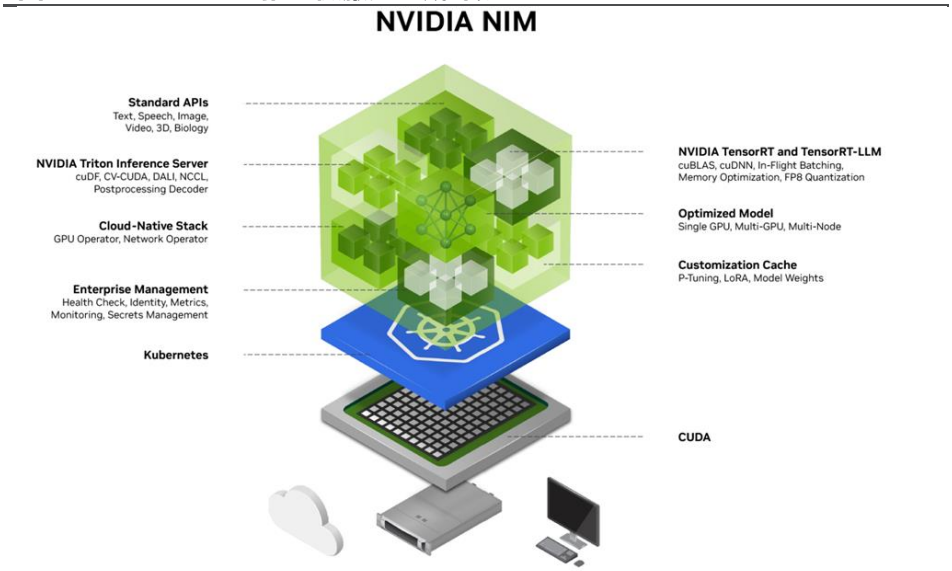


资料来源：NVIDIA 官网，民生证券研究院

5）支持企业级 AI：作为 NVIDIA AI Enterprise 的一部分，NIM 采用企业级基础容器构建，通过功能分支、严格的验证、通过服务级别协议提供的企业级支持以及针对 CVE 的定期安全更新，为企业 AI 软件提供坚实的基础；

6）支持多领域 AI 模型：包括 NVIDIA AI 基础模型 和 NVIDIA 合作伙伴提供的定制 AI 模型。NIM 支持多领域的 AI 应用，包括 大型语言模型 (LLM)、视觉语言模型 (VLM)，以及用于语音、图像、视频、3D、药物研发、医学成像等的模型。

图4：NVIDIA NIM 推理微服务组成部分



资料来源：NVIDIA 官网，民生证券研究院

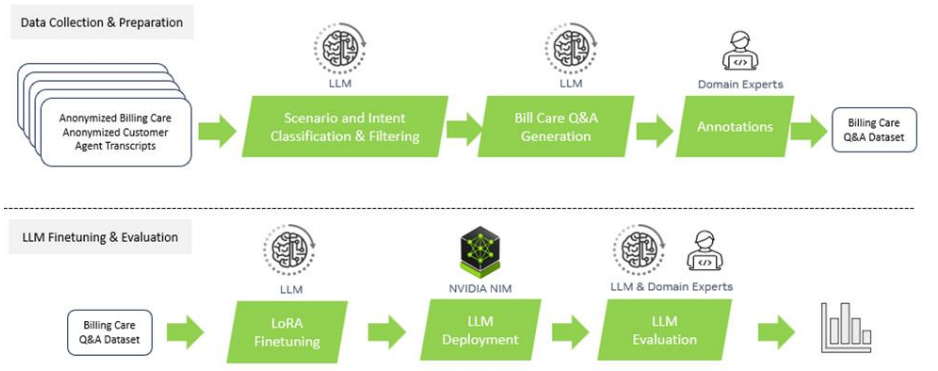
Amdocs 使用英伟达 AI Foundry，通过模型微调与算力调优的功能实现了 AI 精度提高 30%的同时运营成本和使用延迟大幅度降低。Amdocs 使用 Llama2-7b-chat、Llama2-13b-chat 和 Mixtral-8x7b LLM 来增强具有意向分类和账单问答功能的客户服务聊天机器人，Amdocs 设计了带有说明的提示，其中包括目标账单（原始 XML 格式的连续一到两个计费月）和问题。通过 AI Foundry，Amdocs 实现了显著的 AI 性能提升：

1）AI Foundry 基于 NVIDIA Triton 推理服务器和使用 TensorRT-LLM 对 NVIDIA GPU 上的 LLM 推断进行优化，结合 PEFT、LoRA、多种模型混用的方法实现 AI 相应准确度提升 30%；

2）通过对数据格式的整理，tokens 消耗降低了 40%-60%；

3）通过讲 Llama2-13b 型号部署在一个 GPU 上，而 Mixtral-8x7B 部署在两个 GPU 上的算力调度方式使得 AI 均推理速度比领先的最先进的托管 LLM 服务快 4-6 倍。

图5：Amdocs 使用 AI Foundry 打造聊天机器人



资料来源：NVIDIA 官网，民生证券研究院

1.3 “自如式”算力服务有望进入快速渗透期

目前，我国正在大力推进智算中心和智算网络的建设，以进一步完善算力产业链上的基础设施。智能计算中心作为大规模智能计算集群的所在地，不仅具备处理网络数据的能力，还能成功地将计算任务更智能地并行化处理，以高效能源为动力，可以根据不同地区的计算能力需求进行灵活调度和分配。这些措施有效地提升了我国的算力水平，为数字经济的发展提供了强有力的支撑。

表1：国内各地 AI 智算中心建设情况

合作方	中心名称	地域	算力	建设情况
中科曙光	青岛“海之心”人工智能计算中心	山东省青岛市崂山区	混合精度	建设中
	长沙人工智能创新中心硅立方	湖南省长沙市	混合精度	已建成
	万达开先进计算中心	四川省达州市	混合精度	已建成
	宜昌先进计算中心一期	湖北省宜昌市	混合精度	已建成
	芜湖一体化智算中心	安徽省芜湖市	混合精度	建设中
	合肥先进计算中心“巢湖明月”	安徽省合肥市	混合精度	已建成
华为昇腾	北京昇腾人工智能计算中心	北京市门头沟区	一期 100PFLOPS(短期 500P，远期 1000P)	已建成
	中原人工智能计算中心	河南省郑州市	计划 100PFLOPS	已建成
	未来人工智能计算中心	陕西省西安市雁塔区	一期规划 300P FLPOS FP16	已建成
	成都人工智能计算中心	四川省成都市	300P FLOPS(最终 1000P)	已建成
	武汉人工智能计算中心	湖北省武汉市东湖区	100P FLOPS	已建成
	重庆人工智能计算中心	重庆科学城	一期 400PFLOPS	已建成
	长沙人工智能计算中心	湖南省长沙市高新区	200P FLOPS(25 年达 1000P)	已建成
	横琴人工智能超算中心	广州省珠海市横琴区	1.14EopS(完全建成 4EopS)	已建成
	杭州人工智能计算中心	浙江省杭州市	40P FLOPS(后期 100P)	已建成
	南京鲲鹏·昇腾人工智能计算中心	江苏省南京市	800P OpS	已建成
	济南人工智能计算中心	山东省济南市	/	已建成
	青岛人工智能计算中心	山东省青岛市	100P FLOPS	已建成
	天津人工智能计算中心	天津市河北区	300P FLOPS	已建成
	河北人工智能计算中心	河南省廊坊经济开发区	100P FLOPS	已建成
	大连人工智能计算中心	辽宁省大连市	100P FLOPS	建设中
	沈阳人工智能计算中心	辽宁省沈阳市	100P FLOPS(后期 300P)	已建成
	中国-东盟人工智能计算中心	广西省南宁市	一期建设 40P 训练系统和 1.4P 推理系统	已建成
	深圳人工智能融合赋能中心	广东省深圳市龙岗区	接入 20 余万路视频资源	已建成
	广州人工智能公共算力中心	广东省广州市	99P FLOPS	已建成
百度	阳泉智算中心	陕西省阳泉市	计划 100PFLOPS	已建成
	盐城智算中心	盐城市盐南高新区	200P	已建成
	百度沈阳元宇宙智算中心	沈阳市皇姑区	一期 200P 总体 500P	建设中
华海智汇(华为控股子公司)	合肥人工智能计算中心	安徽合肥	100P FLOPS FP16	已建成

商汤科技	商汤人工智能计算中心	上海自贸区临港新片区	同时接入 850 万路视频	已建成
腾讯	腾讯长三角人工智能超算中心	上海市松江区	1400P FLOPS	部分建成
	智慧产业长三角(合肥)智算中心	合肥高新区	/	已建成
上海交通大学等	太湖量子智算中心	江苏省无锡市	/	已建成
浪潮、寒武纪	南京智能计算中心	江苏省南京市	800P OpS	已建成
浪潮信息	淮海智算中心	安徽省宿州市	300P FLOPS	建设中
	青田元宇宙智算中心	浙江省青田县	每秒算力性能超 10 亿亿次	建设中
	克拉玛依智算中心	新疆克拉玛依市	机柜数量超 1 万个	已建成
寒武纪	昆山智算中心	江苏省昆山市	建成后峰值智能算力不低于 500POPS	建设中
阿里云	阿里云张北超级智算中心	河北省张家口市张北县	12EFLOPS	已建成
	阿里云华东智算中心	上海市金山区	在建	建设中
	阿里云乌兰察布超级智算中心	内蒙古乌兰察布市	3EFLOPS	部分建成
中科曙光、华为	长沙 5A 级智算中心	湖南省长沙市	建成后算力规模可达 1024P	已建成
宁数科创	宁波人工智能超算中心	宁波市高新区	-期 100P(FP16)+5P(FP64)二期 300P(FP16)+15P(FP64)	部分建成
福州市电子信息集团	福州智能计算中心	福建省福州新区	一期 105P，总体 400P	已建成
浪潮云、中国电信武汉分公司	武昌智算中心	武汉市武昌区	建成后 100P	建设中
/	哈尔滨人工智能先进计算中心	哈尔滨	100P	已建成
/	北京数字经济算力中心	北京	建成后 1000P 以上	建设中
/	郑州人工智能计算中心	河南省郑州市	一期 2000P,二期 10000P.建成后 30000P	建设中
/	石景山智算中心	北京北重科技文化产业园	建成后 610P	建设中
/	北京七星园数字经济产业智算中心	北京市丰台区	建成后 2300P	建设中
/	华南数谷智算中心	韶关市武江区	一期 16000P	建设中
/	博大数据深圳前海智算中心	深圳	一期 40000P	已建成
/	中国移动智算中心(青岛)	青岛	建成后 825PFlops,参数网络带宽达 200GB	建设中

资料来源：《智能算力产业发展白皮书》，民生证券研究院整理

头部大模型厂商已经开始逐渐落地软硬协同优化提升实践，AI 算力资源的模型调优与算力调优协同解决方案或进入快速渗透的拐点。据阿里研究院，以国内几家头部大模型厂商的创新实践为例（节选），通过在集群架构、功耗散热、资源利用、网络通信、模型算法的综合优化，在模型训练和推理上实现了显著的效

能提升：

1）阿里云的 HPN（高性能网络）通过创新的非堆叠双 ToR（顶部接入交换机）设计和双平面架构，有效提升了大型语言模型训练的吞吐量和可靠性。在功耗散热方面，通过优化的蒸汽腔散热器提高了冷却效率，使得 51.2Tbps 单芯片交换机能在全功率下稳定运行。HPN（高性能网络）部署在生产环境中超过八个月，显著提高了大模型训练的网络性能，数据传输效率提升 14.9%；

2）字节跳动的 MegaScale 系统采用全栈方法，从算法系统共同设计到 3D 并行通信重叠，显著提升了模型训练的效率和稳定性。通过混合并行策略和深度优化的数据流水线，MegaScale 在 12288 个 GPU 的算力集群上训练 175B 参数的模型时，实现了 55.2%的 MFU（模型浮点运算利用率），比业界同尺寸模型的训练效率提升 34%；

3）月之暗面的 Mooncake 平台通过 KVCache（键值缓存）中心化的调度策略，优化了大型语言模型服务的吞吐量和响应速度。在长上下文场景下，与基线方法相比，Mooncake 平台在模拟场景中实现了高达 525%的吞吐量增加，同时在真实工作负载下使模型推理能力提升 75%；

4）深度求索的 DeepSeek-V2 模型引入了 MLA（多头潜在注意力）和 DeepSeekMoE（DeepSeek 混合专家）架构，通过经济高效的训练和推理，显著减少了键值缓存需求，提高了生成吞吐量。在激活参数数量相同的情况下，DeepSeek-V2 与前代相比节省了 42.5%的训练成本，模型推理能力提高 5.76 倍。

图6：模型与芯片协同优化，促进大模型算力高效供给

资料来源：阿里研究院微信公众号，民生证券研究院

1.4 国内 AI 模型与算力协同优化标杆项目持续落地

1.4.1 慧辰股份：联合中科信控、棱镜数聚发布算力服务管理平台赋能 AI+

HCR 联合中科信控与棱镜数聚，融合三方优势（AI 应用建设、企业级平台、智算建设运营）联合打造的融合算力管理服务平台，通过全新的服务平台解决智算运营精细化的需求。据慧辰股份微信公众号，融合算力管理服务平台以融合多元异构算力资源为基础，支持通用算力与智算算力的融合管理（智算设备支持英伟达、华为等多家设备生态），并以全生命周期、细粒度的算力运维服务功能提升运维效率，服务智算的运维运营者。另一方面，客户端侧，平台预置大量多样化 AI 模型镜像资源，以及 AI 应用的全流程开发部署工具，助力用户 AI 研发快捷高效。

图7：慧辰股份融合算力管理服务平台

资料来源：慧辰股份微信公众号，民生证券研究院

慧辰股份“融合算力服务管理平台”基于精细化管理运维的思路，在架构、任务资源模式与全生命期运维产品/计费设计方面，融合多种机制，提升业务管理便捷性与运行效益，如支持基于 Slurm 的资源编排模式，可满足 HPC 超算模式用户计算的资源需求。针对当前大量英伟达企业级智算设备环境，深度融合英伟达企业级 AI 架构特性与高性能镜像资源，可大幅提升算力设备计算效率和管理深度。平台的相关设计，在有效降低运维成本的同时，扩大服务的客户群与 ARPU 值，切实帮助智算运营企业提升预期收益。

1.4.2 恒为科技：智算可视化解决方案已经落地恒为智云•前海智算中心

恒为科技基于自身优势推出智算可视化系统，已部署于恒为智云•前海智算中心。2024 年 8 月 14 日，恒为科技建设运营的“恒为智云•前海智算中心”正式入驻博大数据并点亮运行。公司正式发布的“智算可视化解决方案”已部署于“恒为智云•前海智算中心”，在实际运营中发挥重要作用，为智算中心提供资源调度、运维管理、数字孪生、训推可视化、集群测试工具等一体化服务，并基于“恒为智云•前海智算中心”进行集群优化、模型移植和调优服务的成功实践。

恒为科技推出“恒为智云”第三方运维运营的业务品牌，“恒为智云•前海智算中心”首批算力 300P 于 8 月 14 日在深圳正式点亮。“恒为智云•前海智算中心”是广东省第一个由民企投资、民企运营的华为昇腾集群，也是第一个支持深圳上海两地实现大带宽低延时专网打通计算资源的华为昇腾集群，面向大湾区，服务珠三角、辐射全中国，为各行各业的数字化转型和智算升级提供算力服务。“恒为智云•前海智算中心”的正式投入运营，象征着国产智算领域的又一个进展。作为科技与产业的深度融合体，“恒为智云•前海智算中心”将以其强大的算力支撑，为各行各业注入强劲的创新动力，加速人工智能在各行业中的深度融合与应用，赋能行业转型升级。

图8：恒为科技智算服务业务

资料来源：恒为科技官网，民生证券研究院

1.4.3 浪潮信息：以开放创新的全向 Scale 应对大模型 Scaling Law

浪潮信息在 2024 开放计算中国峰会上提出开放算力模组 OCM 规范，旨在建立基于处理器的标准化算力模组单元，通过统一不同处理器算力单元对外高速互连、管理协议、供电接口等，实现服务器主板平台的深度解耦和模块化设计，兼容不同架构的多代处理器芯片，方便客户根据人工智能、云计算、大数据等多样化应用场景，灵活、快速匹配最适合的算力平台，推动算力产业高质量快速发展。

图9：浪潮信息 OCM 架构

资料来源：浪潮服务器微信公众号，民生证券研究院

浪潮 OCM 规范在算力、管理、基础设施等方面提出更好解决方案：

在算力方面，智算中心需要同时应对两个方向的扩展，分别是强算力支持、一机多芯、多元多模的单机系统 Scale up 要求和大规模 AI 组网、高带宽、资源池化的大规模化扩展 Scale out 要求，以开放加速模组和开放网络实现算力的 Scale。UBB2.0 开放标准支持更高算力规格的加速卡、可以实现更大的 OAM domain 互联，未来可以支持 8000+ 张加速卡 Scale up，突破大模型 All to All 通信过程中的互联瓶颈。同时，大模型的发展需要更大规模的算力系统，浪潮信息开放网络交换机可实现 16000+个计算节点 10 万+加速卡的 Scale out 组网，满足加速卡之间的互联通信需求，带宽利用率高达 95%+。

在管理方面，依托于开源社区的开源固件平台，构建原生解耦架构提升可扩展性，建立统一标准的接口规范，支持用户对于自主模块进行定制化，实现标准接口规范下的异步、自主定制迭代，以满足智算时代的算力迭代需求。

在基础设施方面，采用开放标准、开放生态构建的数据中心基础设施，能更好地匹配智算时代多元、异构算力的扩展和迭代速度，进而支撑上层智能应用的进一步普及。以浪潮信息为例，基于开放标准推出的液冷冷板组件，支撑单机系统内 GPU 和 CPU 核心算力原件 Scale up 扩展；推出模块化、标准接口的 120kw 机柜，兼容液冷、风冷场景，支撑柜内更大的部署需求；推出基于开放标准的预制化集装箱数据中心，大幅压缩建设周期，其扩展性很好的满足了 AI 算力系统的 Scale 需要。

1.4.4 中科曙光：上线“模型仓库” 打造智算中心，让 AI 成为“最强生产力倍增器”

中科曙光正式上线“曙光 AI 模型仓库”，开放百余款模型源码下载，创新“算法分享×算力支持”模式，以下载即用、快速适配、稳定高效的使用体验，为千行百业开发者搭建起易用高效的大模型平台。在算法分享方面，曙光 AI 模型仓库涵盖 100+关键场景、100+款精调最优模型，集成了图片分类、物体检测、语义分割、超分等典型场景网络模型，覆盖 90%以上的 NLP 大模型及 10%以上的 CV 大模型。在算力支持方面，基于曙光智算平台的澎湃算力，低延时、快生成，为用户提供从开发到验证的一站式服务，未来还将实现在线调用功能，助力开发者更高效构建应用。截至目前，曙光 AI 模型仓库模型下载量超 2000 次，镜像下载量超 5000 次，模型从下载到使用最快可达 10 分钟以内。

图10：中科曙光上线模型仓库

资料来源：中国上市公司协会微信公众号，民生证券研究院

中科曙光基于技术经验积累与生态资源，于 2021 年打造出业内标杆—5A 级智算中心，紧密契合时代与产业的发展诉求。中心凭借“开放、融合、绿色、普惠、服务”等显著特点，实现了包括算力供给、算法优化、数据服务及行业应用在内的全场景智能计算服务，并依托开放多元的产业生态，进一步驱动 AI 与各行各业从点到面的融合，为需求侧带来更易用、更通用、更经济、更节能、更省心的算力应用体验。目前，曙光 5A 级智算中心已在合肥、长沙、北京、青岛等多地投运，赋能金融、通信、能源、工业、医疗、科研等诸多领域，驱动海量生态应用落地，为区域和行业的数智发展提供了强有力的支撑。

1.5 投资建议

“自如式”算力服务能够有效解决大模型训练、部署、应用环节的痛点，或成为全新范式。作为“自如式”算力服务的领导者，英伟达于 GTC2024 大会提出的“AI Foundry”具备可随时随地部署、可使用行业标准 API 进行开发、满足特定领域的模型需求、优化的推理引擎、支持企业级 AI、支持多领域 AI 模型等优势，能够显著降低千行百业 AI 应用落地的门槛；同时，阿里云与国内几家头部大模型厂商合作，通过在集群架构、功耗散热、资源利用、网络通信、模型算法的综合优化，在模型训练和推理上实现了显著的效能提升，国内“自如式”算力服务渗透率有望进入快速增长的拐点，国内智算中心如火如荼建设提供蓝海市场，建议关注慧辰股份、恒为科技、浪潮信息、中科曙光、星网锐捷、网宿科技等具备 AI 模型调优与算力调优相结合能力的行业龙头。

2 行业新闻

中国信通院：推进大模型赋能网络安全

8 月 12 日消息，在第十二届互联网安全大会上中央网络安全和信息化委员会办公室副主任、国家互联网信息办公室副主任王京涛介绍，截至目前，我国已经完成备案并上线、能为公众提供服务的生成式人工智能服务大模型达 180 多个，注册用户数已突破 5.64 亿。近年来，我国人工智能发展取得显著成效。一方面，初步构建了较为全面的人工智能技术产业体系，相关企业超过 4500 家，产业规模持续扩大；另一方面，人工智能与实体经济融合不断深化，人工智能应用加速探索，建成 2500 多个数字化车间和智能工厂，经过人工智能改造，研发周期平均缩短 20%，生产效率提升 35%。与会专家认为，人工智能技术的广泛应用正推动各行各业转型升级，并为数字安全行业发展注入强劲动力。要利用大模型重塑安全体系，护航数字经济稳健发展。

华为：华夏银行与华为签署战略合作协议

8 月 13 日消息，华夏银行股份有限公司（以下简称“华夏银行”）与华为技术有限公司（以下简称“华为”）在深圳举行战略合作签约仪式。华夏银行党委书记、董事长李民吉，华为常务董事、华为云 CEO 张平安出席活动并见证签约。华夏银行与华为双方代表分别签署战略合作协议。根据协议，华夏银行与华为将围绕金融行业信息科技、数字化人才培养等方面探索有效合作路径。在云原生、智慧数据、分布式架构演进、新型基础设施建设、创新科技等领域，双方将建立长期、稳定、互惠、互利的全面合作关系，共谱华曲，再续新章。

中国信通院：中国信通院发布《中国企业级 SaaS 产业发展研究报告（2024 年)》

8 月 14 日消息，中国信息通信研究院（简称“中国信通院”）云计算与大数据研究所云应用与服务团队组织开展研究，编写《中国企业级 SaaS 产业发展研究报告（2024 年)》报告。报告从中国企业级 SaaS 产业的发展背景出发，深入剖析中国企业级 SaaS 产业七大发展态势，结合 AIGC、出海等最新热点进行分析，同时梳理产业当前面临的六大挑战，并从政策、技术、监管、合规、人才、标准等方面提出产业发展建议。报告还通过行业应用与典型案例展示了 SaaS 在不同领域的解决方案和实际应用成效。此报告的核心观点为：我国 SaaS 产业历经二十余年发展，目前正处于成长变革阶段；我国企业级 SaaS 产业市场规模不断增加，细分赛道龙头显现；AI 赋能加之海外市场探索，企业级 SaaS 产业发展迎来新机遇。

中国信通院：中国信通院发布《数据交易场所发展指数研究报告（2024 年)》

8 月 16 日消息， 为助推数据交易场所高质量发展，繁荣场内数据交易市场，中国信息通信研究院围绕“以评促统、以评促建、以评促进”，撰写了《数据交易场所发展指数研究报告（2024 年)》。报告紧扣数据要素价值化主线，深入分析国内外数据交易场所发展现状和趋势，系统剖析我国数据交易场所发展面临的机遇和挑战，并在贵阳大数据交易所的大力支持下，遵循国家政策导向、实践导向、发展需求导向和问题导向，从构建规范高效的数据交易场所入手，围绕发展环境、基础支撑、市场交易、生态构建、辐射影响五个维度，研究建立了数据交易场所发展指数体系 1.0，希望能为各级数据管理机构和产业界相关方推进数据交易高质量发展提供有价值的参考。此报告的核心观点为：我国数据交易场所发展提速；数据交易场所机遇与挑战并存；以评促进引领数据交易场所改革创新。

3 公司新闻

萤石网络：8 月 12 日消息，公司董事长兼总经理蒋海青先生以集中竞价方式增持公司 243,415 股份，增股后占持股 2,085,815 股，直接持有公司股份的 0.2649%，蒋海青先生增持计划已实施完成

智迪科技：8 月 12 日消息，持股 5%以上股东智控投资持有 5,280,000 股，占公司总股本比例 6.60% ，拟通过集中竞价方式减持公司股份不超过 800,000 股，即不超过公司总股本的 1.00%

税友股份：8 月 12 日消息，公司已完成工商变更登记

正元地信：8 月 12 日消息，公司收到董事、副总经理侯凤辰先生的书面辞职报告。因退休，申请辞去公司第二届董事会董事、副总经理职务。辞职后，侯凤辰先生不再担任公司任何职务

卫宁健康：8 月 13 日消息，公司收到了由国家知识产权局颁发的三项《发明专利证书》，发明名称为：一种肺功能训练装置及训练数据处理方法，5G 应用领域中一种考虑用户正负隐式反馈关联的医疗资讯推荐方法，5G 应用领域中一种基于异构双层网络的医疗资讯推荐方法

山石网科：8 月 13 日消息，限制性股票激励计划，公司向激励对象授予限制性股票总计 992.00 万股，占目前公司股本总额 18,023.0255 万股的 5.50%

捷安高科：8 月 14 日消息，公司已完成工商变更登记

艾融软件：8 月 15 日消息，持股 5%以上股东孟庆有持有 25,681,690 股，占公司股份总数 12.2079%, 拟通过大宗交易方式减持股份不高于 6,000,000 股，不高于公司股份总数 2.8521%

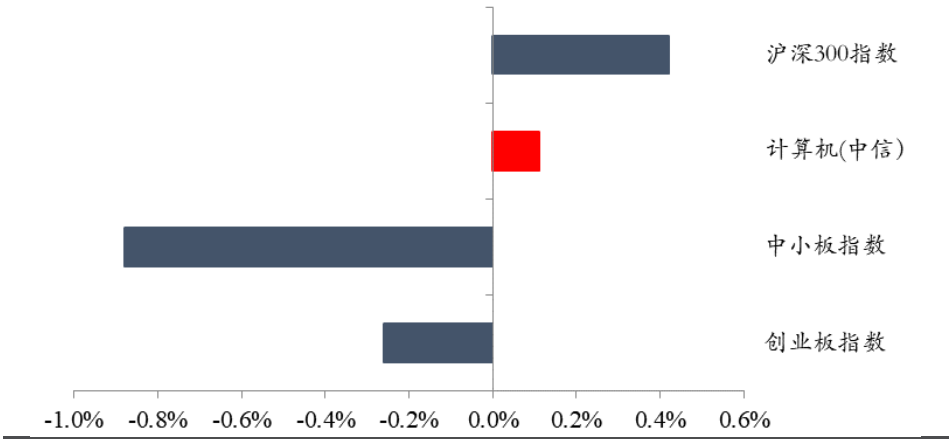
太极股份：8 月 15 日消息，公司已完成工商变更登记

安居宝：8 月 16 日消息，高管张焕清女士持有本公司股份 92,900 股，占本公司总股本比例 0.0166%，拟通过集中竞价方式减持公司股份合计不超过 23,225 股，占本公司总股本比例 0.0041%

4 本周市场回顾

本周（8.12-8.16）本周沪深 300 指数上涨 0.42%，中小板指数下跌 0.88%，创业板指数下跌 0.26%，计算机（中信）板块上涨 0.11%。板块个股涨幅前五名分别为：优博讯、安博通、雷柏科技、GQY 视讯、朗科科技；跌幅前五名分别为：任子行、威创股份、*ST 博信、汇金股份、彩讯股份。

图11：计算机板块本周表现

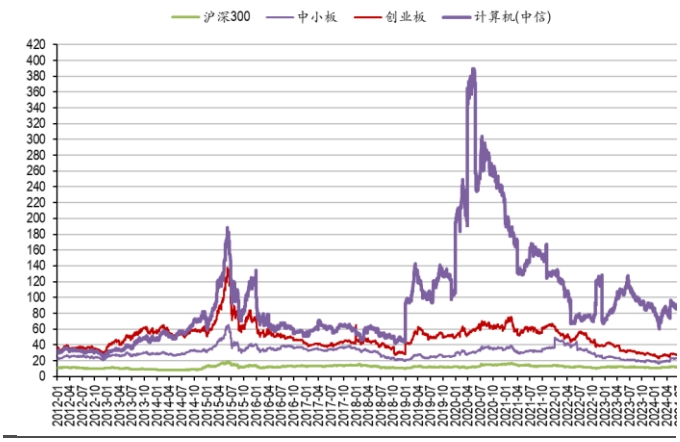


资料来源：iFinD，民生证券研究院

图12：计算机板块指数历史走势

资料来源：iFinD，民生证券研究院

图13：计算机板块历史市盈率



资料来源：iFinD，民生证券研究院

表2：本周计算机板块个股涨幅前五名

证券代码	证券简称	周涨跌幅(%)	收盘价（元）	周最低价（元）	周最高价（元）
300531.SZ	优博讯	18.31%	11.63	9.44	12.35
688168.SH	安博通	17.46%	29.20	24.50	30.18
002577.SZ	雷柏科技	16.67%	13.16	11.03	13.74
300076.SZ	GQY 视讯	16.07%	3.90	3.21	4.42
300042.SZ	朗科科技	15.95%	19.70	16.20	21.80

资料来源：iFinD, 民生证券研究院；(涨幅区间为 2024 年 8 月 12 日至 2024 年 8 月 16 日)

表3：本周计算机板块个股跌幅前五名

证券代码	证券简称	周涨跌幅(%)	收盘价（元）	周最低价（元）	周最高价（元）
300311.SZ	任子行	-24.43%	3.00	2.86	3.18
002308.SZ	威创股份	-22.41%	0.45	0.45	0.55
600083.SH	*ST 博信	-11.54%	2.07	2.06	2.28
300368.SZ	汇金股份	-10.80%	2.56	2.54	2.85
300634.SZ	彩讯股份	-10.50%	13.81	13.46	14.43

资料来源：iFinD, 民生证券研究院；(跌幅区间为 2024 年 8 月 12 日至 2024 年 8 月 16 日)

表4：计算机行业重点关注个股

证券代码	证券简称	股价（元）	周涨跌幅	2021EPS	2022EPS	2023EPS	2021PE	2022PE	2023PE	PB
002230.SZ	科大讯飞	34.46	-4.01%	0.70	0.24	0.28	49	144	123	4.8
600570.SH	恒生电子	16.40	-3.36%	1.01	0.57	0.75	16	29	22	3.9
000977.SZ	浪潮信息	33.09	-0.03%	1.38	1.39	1.18	24	24	28	2.7
300170.SZ	汉得信息	6.09	3.57%	0.22	0.49	-0.03	28	12	/	1.2
300454.SZ	深信服	47.11	-0.88%	0.67	0.47	0.47	70	100	100	2.4
300451.SZ	创业慧康	3.37	2.74%	0.27	0.03	0.02	12	112	169	1.1
300253.SZ	卫宁健康	5.49	1.67%	0.18	0.05	0.17	31	108	33	2.1
002368.SZ	太极股份	15.21	-2.25%	0.64	0.65	0.61	24	23	25	1.8
300212.SZ	易华录	16.06	-4.18%	-0.25	0.02	-2.83	/	923	/	3.4
002410.SZ	广联达	9.92	-5.88%	0.56	0.82	0.07	18	12	141	2.8
002153.SZ	石基信息	5.23	-0.76%	-0.32	-0.37	-0.04	/	/	/	1.9
600588.SH	用友网络	8.68	-2.91%	0.22	0.06	-0.29	39	145	-30	3.1
002912.SZ	中新赛克	15.98	1.33%	0.33	-0.71	0.67	48	/	24	1.7
300365.SZ	恒华科技	4.75	5.09%	0.10	-0.37	0.03	48	/	158	1.4
300523.SZ	辰安科技	15.46	-1.34%	-0.68	0.03	0.34	-23	515	45	2.5
603039.SH	泛微网络	27.69	-0.47%	1.20	0.86	0.69	23	32	40	3.5
002376.SZ	新北洋	5.41	3.84%	0.22	-0.04	0.03	25	/	183	1.2
603660.SH	苏州科达	5.06	2.85%	0.13	-1.18	-0.54	40	/	/	3.1
002439.SZ	启明星辰	13.29	-5.07%	0.93	0.67	0.79	14	20	17	1.4

资料来源：iFinD, 民生证券研究院；(注：股价为 2024 年 8 月 16 日收盘价)

5 风险提示

1）政策落地不及预期：目前国产软硬件在产品性能和生态上都尚且不及国外巨头，但受益于国产化政策推动市场份额连年提升，若后续国产化支持政策落地进度不及预期，可能会导致国产软硬件推进进度变慢，影响公司业绩增长前景。

2）行业竞争加剧：目前国产软硬件尚未呈现出清晰的格局，芯片、数据库、操作系统等行业仍处于高度竞争状态，若后续行业竞争加剧，可能会影响公司的毛利率水平，进而影响相关公司的盈利能力。

附录

表5：计算机行业限售股解禁情况汇总

公司代码	公司简称	解禁日期	解禁数量(万股)	解禁市值(万元)	总股本（万股）	流通 A 股（万股）
688479.SH	友车科技	2024-11-11	263.80	4,028.23	14,431.74	6,087.42
300348.SZ	长亮科技	2024-11-07	68.44	481.78	80,505.89	63,127.99
002180.SZ	纳思达	2024-11-04	6,727.17	155,061.29	141,650.97	137,640.90
300348.SZ	长亮科技	2024-11-04	488.72	3,440.55	80,505.89	63,059.55
833030.BJ	立方控股	2024-11-04	6,694.05	88,026.77	9,224.32	9,224.32
300167.SZ	*ST 迪威	2024-11-01	289.50	489.26	36,055.00	35,722.75
839493.BJ	并行科技	2024-11-01	1,450.76	55,694.68	5,823.00	4,049.32
430564.BJ	天润科技	2024-10-31	13.80	140.48	7,452.90	3,392.05
688152.SH	麒麟信安	2024-10-28	86.51	2,818.65	7,873.86	3,105.86
688291.SH	金橙子	2024-10-28	128.33	1,942.97	10,266.67	3,366.67
002405.SZ	四维图新	2024-10-21	2,667.30	19,124.54	237,775.03	236,184.93
600728.SH	佳都科技	2024-10-21	619.54	2,304.69	214,323.03	214,033.19
688244.SH	永信至诚	2024-10-21	128.22	2,917.09	10,223.42	4,948.50
301085.SZ	亚康股份	2024-10-18	4,377.38	172,819.15	8,677.57	8,315.80
688031.SH	星环科技	2024-10-18	120.84	3,568.48	12,084.21	9,367.87
002987.SZ	京北方	2024-10-14	17.64	164.40	61,790.76	60,064.86
300377.SZ	赢时胜	2024-10-14	217.40	1,223.96	75,107.51	64,476.72
300743.SZ	天地数码	2024-10-14	2.51	30.09	15,347.91	12,962.38
600131.SH	国网信通	2024-10-14	159.11	2,665.13	120,175.90	119,590.51
688657.SH	浩辰软件	2024-10-10	2,471.39	78,491.31	6,551.43	3,945.06
300659.SZ	中孚信息	2024-10-08	3,485.16	41,926.50	26,039.24	19,037.75
300846.SZ	首都在线	2024-09-23	3,363.93	31,351.84	50,046.22	39,025.78
603383.SH	顶点软件	2024-09-23	18.58	532.57	20,543.64	20,379.48
688561.SH	奇安信	2024-09-23	14,956.16	344,440.46	68,517.24	51,005.87
301378.SZ	通达海	2024-09-20	820.97	16,977.63	9,660.00	5,207.18
300743.SZ	天地数码	2024-09-18	22.27	267.49	15,347.91	12,959.88
688316.SH	青云科技	2024-09-18	1,185.55	31,962.34	4,779.13	4,779.13
688695.SH	中创股份	2024-09-13	115.15	3,376.19	8,505.14	1,913.66

资料来源：iFinD，民生证券研究院（数据截至 2024 年 8 月 16 日）

插图目录

图 1： “自如式”算力服务梳理 3
图 2： NVIDIA AI Foundry 构建 AI 模型的流程 4
图 3： NVIDIA NIM 减少了 Llama 3.1 模型的推理延迟、更快地生成 token 5
图 4： NVIDIA NIM 推理微服务组成部分 6
图 5： Amdocs 使用 AI Foundry 打造聊天机器人 6
图 6： 模型与芯片协同优化，促进大模型算力高效供给 9
图 7： 慧辰股份融合算力管理服务平台 10
图 8： 恒为科技智算服务业务 11
图 9： 浪潮信息 OCM 架构 12
图 10： 中科曙光上线模型仓库 13
图 11： 计算机板块本周表现 17
图 12： 计算机板块指数历史走势 17
图 13： 计算机板块历史市盈率 17

表格目录

表 1： 国内各地 AI 智算中心建设情况 7
表 2： 本周计算机板块个股涨幅前五名 18
表 3： 本周计算机板块个股跌幅前五名 18
表 4： 计算机行业重点关注个股 18
表 5： 计算机行业限售股解禁情况汇总 20

分析师承诺

本报告署名分析师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注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、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。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

民生证券股份有限公司（以下简称“本公司”）具有中国证监会许可的证券投资咨询业务资格。

本报告仅供本公司境内客户使用。本公司不会因接收人收到本报告而视其为客户。本报告仅为参考之用，并不构成对客户的投资建议，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，客户应当充分考虑自身特定状况，不应单纯依靠本报告所载的内容而取代个人的独立判断。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而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。

民生证券研究院：

上海：上海市浦东新区浦明路 8 号财富金融广场 1 幢 5F； 200120

北京：北京市东城区建国门内大街 28 号民生金融中心 A 座 18 层； 100005

深圳：广东省深圳市福田区益田路 6001 号太平金融大厦 32 层 05 单元； 518026