



Research and
Development Center

如何看待 AI 算力芯片与服务器的弹性？

2023 年 3 月 31 日

证券研究报告

行业点评

行业专题研究（普通）

电子

投资评级 看好

上次评级 看好

莫文宇 电子行业首席分析师
执业编号：S1500522090001
联系电话：13437172818
邮箱：mowenyu@cindasc.com

韩宇杰 联系人
邮箱：hanzijie@cindasc.com

信达证券股份有限公司
CINDASECURITIES CO., LTD
北京市西城区闹市口大街9号院1号楼
邮编：100031

如何看待 AI 算力芯片与服务器的弹性？

2023 年 3 月 31 日

本期内容提要：

- **Transformer 模型是参数量可以无限增长的通用模型，可以处理长序列的输入、输出。**Transformer 模型是一种基于自注意力机制的深度机器学习模型，相较于传统 AI 模型如循环神经网络（RNN）和卷积神经网络（CNN），它在处理序列数据时具有更高的并行性和可扩展性。其中，自注意力机制使得模型能够捕捉序列中长距离依赖关系，同时避免了 RNN 中的梯度消失或爆炸问题。Transformer 模型的参数量之所以会随着数据量和任务复杂度无限增长，是因为它可以通过堆叠更多的层或增加隐藏层宽度来提高模型性能，从而适应更复杂的数据和任务；在传统 CNN/RNN 模型中，增加网络参数量会提高模型的拟合能力，但过多的参数容易导致过拟合现象。因此，长期来看大模型领域或许会出现没有上限的“军备竞赛”。
- **训练：**根据英伟达测算，在 4096 颗 A100 GPU 互联的情况下，训练 10000 亿参数的模型约需要 3-4 周时间，而在 4096 颗 H100 GPU 互联的情况下，时间约需要 1 周左右。考虑到 1 周是研发阶段可承受的周期，因此我们认为 4000 颗左右 H100 将是一个大型云计算企业每年支持 30-50 个 5000 亿参数模型项目训练所需要的芯片数量。考虑到目前大模型参数量已经步入万亿级别，芯片需求量可能会继续上升。若以英伟达 HGX H100-8 计算集群为单位（1024 Nvidia H100 GPUs），总共需要 4-8 个计算集群，考虑到售价为 4900 万美金/集群，则单个支持 GPT 训练的云计算中心至少投入 1.96-3.92 亿美金用于服务器的购买。
- **推理：**相比训练，模型的部署是更大的市场。我们认为应用对于服务器的最大承载能力要求在于过峰，即满足 PCU 时刻的访问需求。根据 Statista，Steam 2021 年月活用户约为 1.32 亿人左右，而 Steam 公布服务器最高访问人数截至 2021 年 12 月 31 日为 2738 万人，最高 817 万人同时在线进行游戏。因此我们合理推断一款被高强度访问的应用，最高在线人数约为月活用户的 6.19%-20.74%。假设在国内 8 亿用户体量下，最高并发访问人数可高达 4952 万人-16592 万人。根据 10 DGX H100 服务器一秒内可支持 300 用户的并发访问，若响应时长延缓至 5 秒及以上，最高可支持 1000 人以上的并发访问。据此测算，国内由 AI 带动的服务器需求量约为 5 万台-16.6 万台，芯片为 50-166 万颗。假设全球活跃用户体量为国内用户体量的 3 倍，则全球服务器需求量为 15-50 万台，芯片需求量为 150-500 万颗。
- **投资建议：**ChatGPT 的快速渗透侧面反映出大模型在通用任务上所展现的强大能力，也为人工智能产业未来的投资指引了方向。在 AIGC 驱动新一轮产业革命的趋势下，我们认为 AI 算力产业链上下游均有投资机会。
- **建议关注：**海光信息、寒武纪-U、龙芯中科、澜起科技、工业富联、紫光国微、中芯国际、华虹半导体、华大九天、概伦电子、芯原股份、兴森科技、通富微电、华正新材、方邦股份、长电科技、海康威视、大华股份、晶晨股份、瑞芯微、全志科技。
- **风险提示：**技术迭代不及预期；制裁风险；本文对服务器、芯片需求量等测算是基于一定前提假设，存在假设条件不成立、市场发展不及预期等因素导致测算结果偏差。

目录

训练与部署：我们如何测算 AI 芯片及服务器的弹性	4
训练：迈入万亿参数时代，千卡互联、超级计算是大势所趋	4
部署：强 AI 助手有望实现大规模渗透，服务器市场注入新活力	6
投资建议：趋势确立，下注未来	7
风险因素	9

表目录

表 1：汽车及细分行业近年销量增速情况（%）	5
------------------------------	---

图目录

图 1：传统 NLP 模型存在梯度下降/爆炸问题，无法解决长序列的处理	4
图 2：英伟达 H100 训练卡对 LLM 的训练时长函数曲线	5
图 3：2022Q4 全球云计算服务商市场份额	6
图 4：英伟达 10 DGX H100 瞬时支持的 300 人的并发访问	7
图 5：英伟达过去 5 个季度收入构成（百万美元）	7
图 6：全球服务器市场产值及出货量预测	7

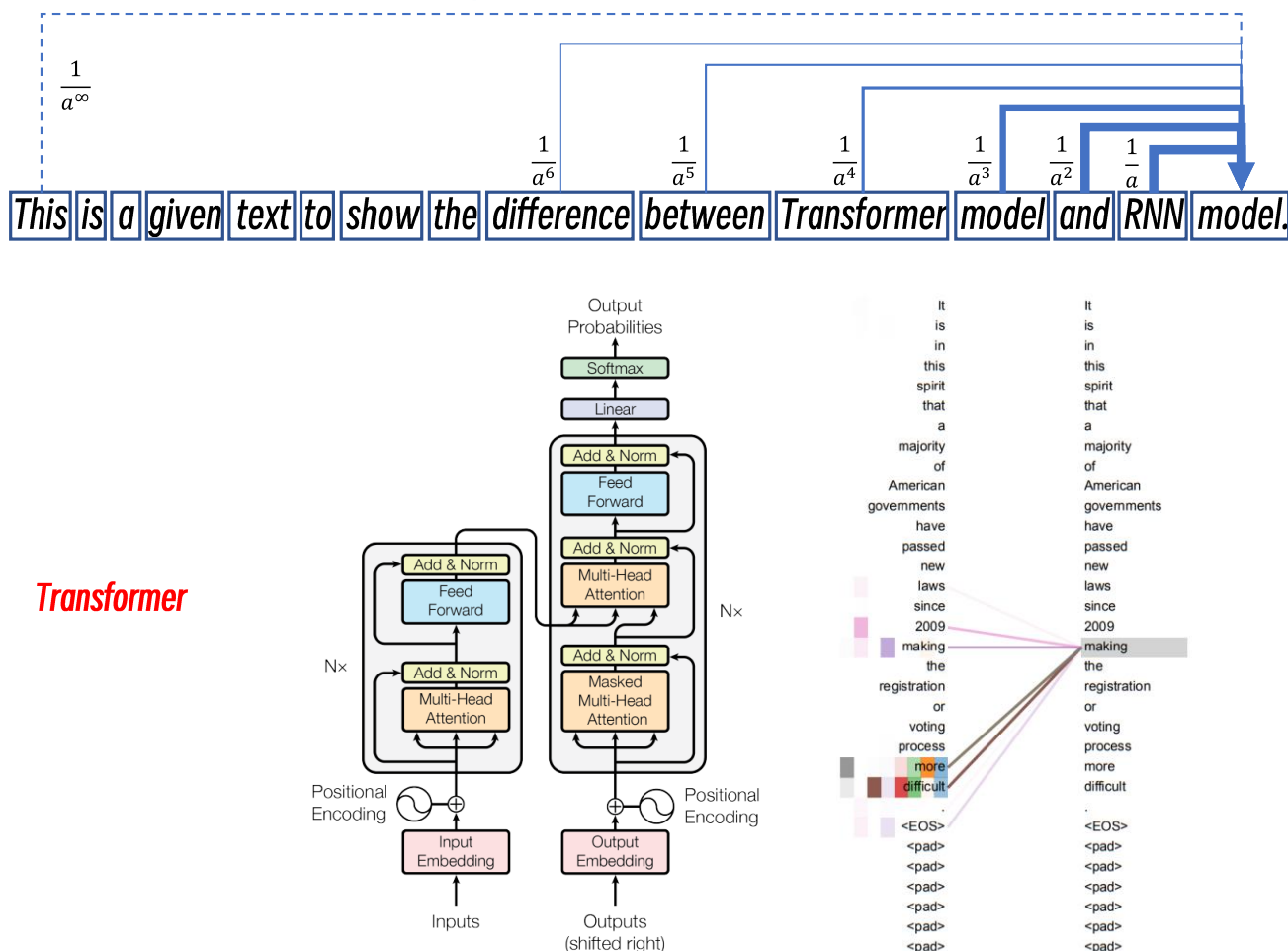
训练与部署：我们如何测算 AI 芯片及服务器的弹性

训练：迈入万亿参数时代，千卡互联、超级计算是大势所趋

Transformer 模型是参数量可以无限增长的通用模型，可以处理长序列的输入、输出。Transformer 模型是一种基于自注意力机制的深度学习模型，相较于传统 AI 模型如循环神经网络（RNN）和卷积神经网络（CNN），它在处理序列数据时具有更高的并行性和可扩展性。其中，自注意力机制使得模型能够捕捉序列中长距离依赖关系，同时避免了 RNN 中的梯度消失或爆炸问题。Transformer 模型的参数量之所以会随着数据量和任务复杂度无限增长，是因为它可以通过堆叠更多的层或增加隐藏层宽度来提高模型性能，从而适应更复杂的数据和任务；在传统 CNN/RNN 模型中，增加网络参数量会提高模型的拟合能力，但过多的参数容易导致过拟合现象。这意味着模型可能在训练集上表现良好，但在测试集或实际应用中的泛化能力较差。

图 1：传统 NLP 模型存在梯度下降/爆炸问题，无法解决长序列的处理

RNN：存在梯度下降/爆炸问题



资料来源：Attention Is All You Need (Ashish Vaswani 等)，信达证券研发中心

训练的过程就是不断调整参数矩阵的过程，参数矩阵体量越大，消耗的数据和算力越多。AI 模型的训练可以理解为一个根据输入数据（如文字、图像、声音等）进行预测的过程。在这个过程中，模型通过不断调整其内部参数，试图让预测结果尽可能接近真实的目标。参数矩阵是 AI 模型中用于存储权重和偏置的矩阵。权重和偏置是模型的关键参数，它们在模型的训练过程中被不断更新，以使模型在训练数据上获得更好的预测性能。在训练过程中，每次迭代时，模型都会通过计算损失函数（衡量模型预测值与真实值之间的差距）的

梯度，然后根据这个梯度调整参数矩阵中的数值，使损失函数最小化。对参数矩阵的操作主要包括以下几个方面：

- **初始化：**在训练开始时，参数矩阵中的权重和偏置通常会被随机初始化，这样可以保证模型从一个比较广泛的参数空间开始学习。
- **前向传播：**根据输入数据和当前参数矩阵，模型计算输出预测值。这个过程中涉及到矩阵乘法和激活函数等操作。
- **反向传播：**计算损失函数关于参数矩阵的梯度，这个梯度表示了损失函数在当前参数下的变化方向和程度。
- **更新参数：**根据梯度和设定的学习率，调整参数矩阵中的数值。这个过程通常使用优化算法（如随机梯度下降、Adam 等）完成。

通过反复进行以上操作，AI 模型在训练过程中不断调整参数矩阵，以提高在训练数据上的预测性能。

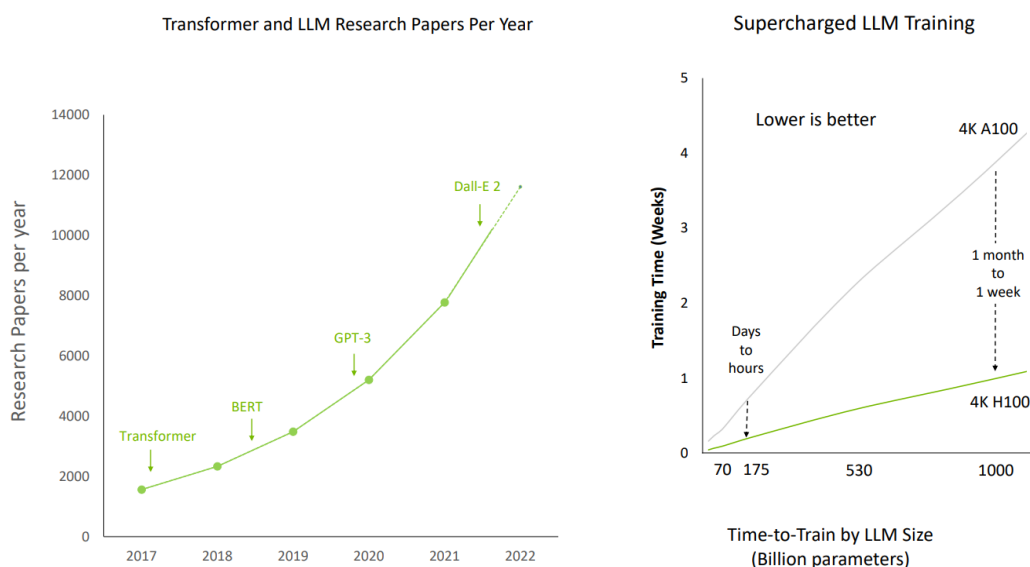
表 1：汽车及细分行业近年销量增速情况（%）

Model Name	n _{params}	n _{layers}	d _{model}	n _{heads}	d _{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}

资料来源：Language Models are Few-Shot Learners (Tom B. Brown 等)，信达证券研发中心

根据英伟达测算，在 4096 颗 A100 GPU 互联的情况下，训练 10000 亿参数的模型约需要 3-4 周时间，而在 4096 颗 H100 GPU 互联的情况下，时间约需要 1 周左右。考虑到 1 周是研发阶段可承受的周期，因此我们认为 4000 颗左右 H100 将是一个大型云计算企业每年支持 30-50 个 5000 亿参数模型项目训练所需要的芯片数量。考虑到目前大模型参数量已经步入万亿级别，芯片需求量可能会继续上升。

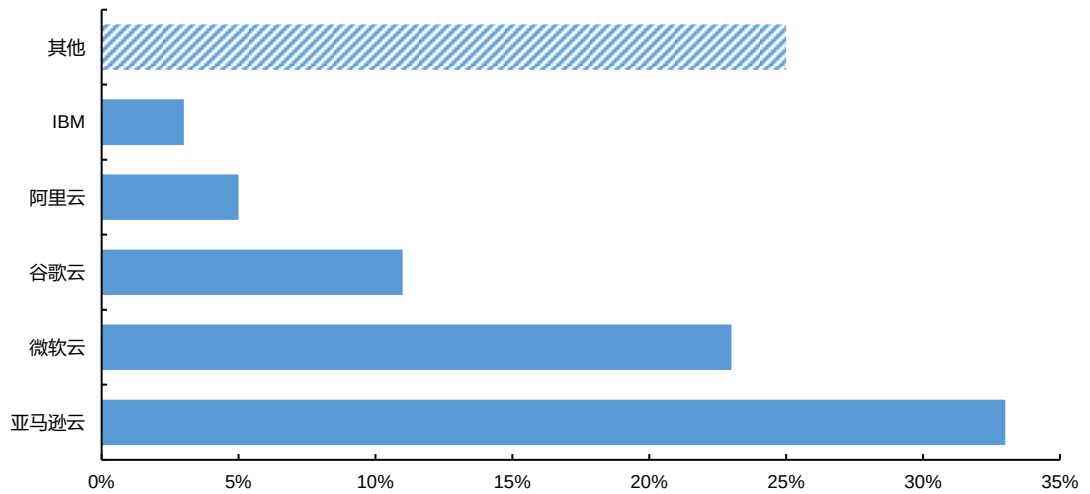
图 2：英伟达 H100 训练卡对 LLM 的训练时长函数曲线



资料来源：COLFAX，信达证券研发中心

若以英伟达 HGX H100-8 计算集群为单位（1024 Nvidia H100 GPUs），总共需要 4-8 个左右集群，考虑到售价为 4900 万美金/集群，则单个支持 GPT 训练的云计算中心至少投入 1.96-3.92 亿美金用于服务器的购买。

图 3：2022Q4 全球云计算服务商市场份额



资料来源：CRN，信达证券研发中心

部署：强 AI 助手有望实现大规模渗透，服务器市场注入新活力

部署即将大模型应用到服务器端，向消费者提供服务的过程，要讨论部署对服务器市场的拉动，我们必须先理清三个名词的含义和他们之间的对应关系：

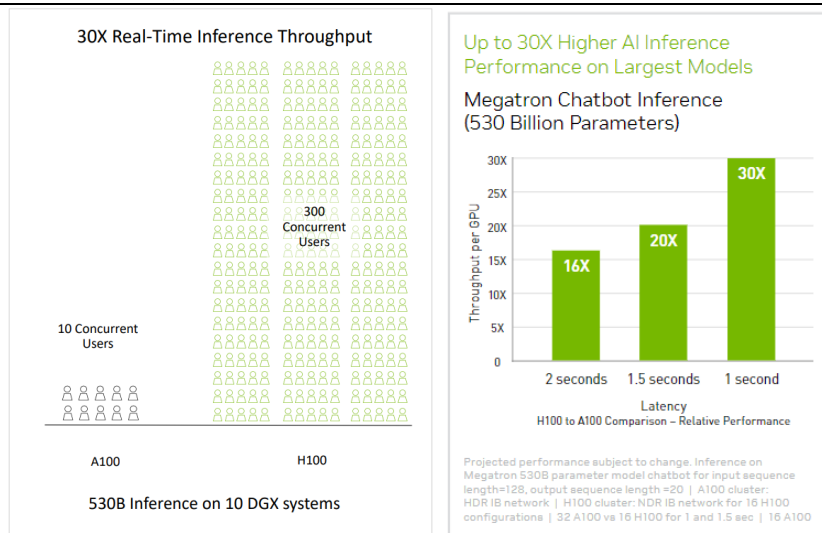
- **活跃用户数量**：指的是在任意一个时间段内活跃的用户总数，这个数字可以反映一个网站或应用的流行程度。若这个时间段以月为单位，则称为月活用户数量（Monthly Active Users，简称 MAU）。
- **最高并发访问人数（Peak Concurrent Users，简称 PCU）**：表示在某一特定时刻，服务器能够同时承载的最大访问人数。这个指标可以用来衡量服务器的处理能力和承载压力。
- **服务器压力（Server Load）**：指的是服务器在处理用户请求时所承受的负荷，它通常受到并发访问人数、请求处理速度、网络带宽等多种因素的影响。服务器压力过大可能导致服务器性能下降、访问速度减缓甚至宕机。维持合理的服务器压力对于保障用户体验至关重要。

对于服务器的需求数量来说，它取决于最高并发访问时的服务器压力，而最高访问数量又与活跃用户数量挂钩，但远小于总用户体量。月活用户数量（MAU）和服务器最高并发访问人数（PCU）之间的相关关系确实存在，但这两者并非完全线性关系。尽管月活用户数量能够反映一个网站或应用的流行程度，但并发访问人数受到很多其他因素的影响，例如用户访问行为、访问时间段以及地域分布等。

根据 Statista，Steam 2021 年月活用户约为 1.32 亿人左右，而 Steam 公布服务器最高访问人数截至 2021 年 12 月 31 日为 2738 万人，最高 817 万人同时在线进行游戏。因此我们合理推断一款被高强度访问的应用，最高在线人数约为月活用户的 6.19%-20.74%。

假设在国内 8 亿用户体量下，最高并发访问人数可高达 4952 万人-16592 万人。根据 10 DGX H100 服务器一秒内可支持 300 用户的并发访问，若响应时长延缓至 5 秒及以上，最高可支持 1000 人以上的并发访问。据此测算，国内由 AI 带动的服务器需求量约为 5 万台-16.6 万台，芯片为 50-166 万颗。假设全球活跃用户体量为国内用户体量的 3 倍，则全球服务器需求量为 15-50 万台，芯片需求量为 150-500 万颗。

图 4：英伟达 10 DGX H100 瞬时支持的 300 人的并发访问

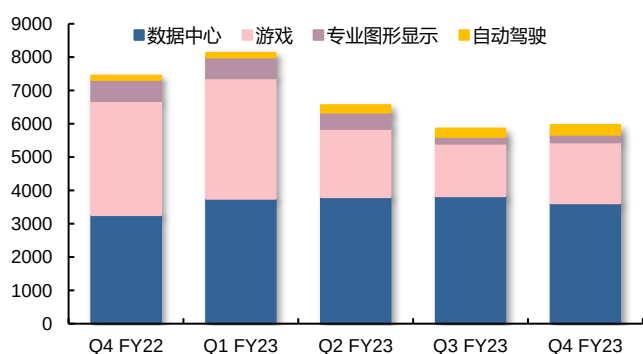


资料来源：Nvidia 官网，NVIDIA H100 Customer Deck，信达证券研发中心

投资建议：趋势确立，下注未来

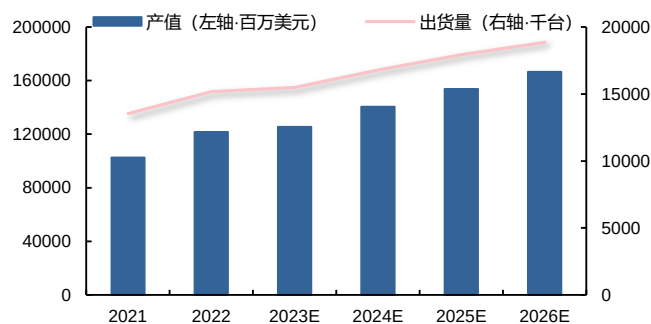
回顾英伟达业绩，可以看出在宏观经济、美国加息等因素影响下，英伟达下游的虚拟货币需求受到影响，主要体现在游戏、专业图形显示两个业务的收入规模显著下滑。而 AI 驱动的数据中心业务在过去 5 个季度保持了较为稳定的收入，侧面反映出该业务受其他不相干因素影响较小，可以作为弹性测量的基准。

图 5：英伟达过去 5 个季度收入构成（百万美元）



资料来源：英伟达官网，信达证券研发中心

图 6：全球服务器市场产值及出货量预测



资料来源：工业富联年报，信达证券研发中心

我们以英伟达 H100 及相关服务器、计算集群为依据，给出如下总结：

- **训练的市场体量：**由于模型训练市场一般在 1 周左右，因此云计算企业的 AI 服务器可以复用，单一云计算企业投入训练的 GPU/AI 芯片在 4000 颗以上，甚至高达上万颗。
- **训练的市场格局：**现代的人工智能模型大多为多种基础模型算法、特性的组合，因此 AI 训练芯片应当具备灵活的编程环境，提供完整的工具链。因此，我们认为在训练侧国内仅有少数具备体量优势、市场优势的算力芯片企业能够实现量变与质变；如寒武纪的思元系列、海思的昇腾系列、平头哥的含光系列等。
- **推理的市场体量：**相比训练，模型的部署是更大的市场。我们认为应用对于服务器的最大承载能力要求在于过峰，即满足 PCU 时刻的访问需求。考虑 AI 未来可能渗透至数十亿人的工作、生活，全球 AI 服务器需求增量可能高达 50 万台。
- **推理的市场格局：**在微架构设计上，推理芯片通常会相对简单一些，国内布局的企业更多，云计算大厂也有相关的研发和投资。这主要是因为推理任务相对于训练任务有更低的计算密集度和内存需求。此外，推理过程主要关注模型的前向传播，这涉及到

一系列矩阵乘法和非线性激活函数。相比之下，训练过程不仅包括前向传播，还包括反向传播以及参数更新。

ChatGPT 的快速渗透侧面反映出大模型在通用任务上所展现的强大能力，也为人工智能产业未来的投资指引了方向。在 AIGC 驱动新一轮产业革命的趋势下，我们认为 AI 算力产业链上下游均有投资机会。建议关注以下方向

- (1) 服务器+AI: 海光信息、寒武纪-U、龙芯中科、澜起科技、工业富联、紫光国微
- (2) 上游晶圆制造+IP: 中芯国际、华虹半导体、华大九天、概伦电子、芯原股份
- (3) chiplet: 兴森科技、通富微电、华正新材、方邦股份、长电科技
- (4) 下游: 海康威视、大华股份、晶晨股份、瑞芯微、全志科技

风险因素

1. 技术迭代不及预期；
2. 制裁风险；
3. 本文对服务器、芯片需求量等测算是基于一定前提假设，存在假设条件不成立、市场发展不及预期等因素导致测算结果偏差。

研究团队简介

莫文字，毕业于美国佛罗里达大学，电子工程硕士，2012-2022 年就职于长江证券研究所，2022 年入职信达证券研发中心，任副所长、电子行业首席分析师。

韩宇杰，电子行业研究员。华中科技大学计算机科学与技术学士、香港中文大学硕士。研究方向为半导体设备、半导体材料、集成电路设计。

郭一江，电子行业研究员。本科兰州大学，研究生就读于北京大学化学专业。2020 年 8 月入职华创证券电子组，后于 2022 年 11 月加入信达证券电子组，研究方向为光学、消费电子、汽车电子等。

机构销售联系

区域	姓名	手机	邮箱
全国销售总监	韩秋月	13911026534	hanqiuyue@cindasc.com
华北区销售总监	陈明真	15601850398	chenmingzhen@cindasc.com
华北区销售副总监	阙嘉程	18506960410	quejiacheng@cindasc.com
华北区销售	祁丽媛	13051504933	qiliyuan@cindasc.com
华北区销售	陆禹舟	17687659919	luyuzhou@cindasc.com
华北区销售	魏冲	18340820155	weichong@cindasc.com
华北区销售	樊荣	15501091225	fanrong@cindasc.com
华北区销售	秘侨	18513322185	miqiao@cindasc.com
华北区销售	李佳	13552992413	lijia1@cindasc.com
华北区销售	赵岚琦	15690170171	zhaolanqi@cindasc.com
华北区销售	张澜夕	18810718214	zhanglanxi@cindasc.com
华东区销售总监	杨兴	13718803208	yangxing@cindasc.com
华东区销售副总监	吴国	15800476582	wuguo@cindasc.com
华东区销售	国鹏程	15618358383	guopengcheng@cindasc.com
华东区销售	朱尧	18702173656	zhuyao@cindasc.com
华东区销售	戴剑箫	13524484975	daijianxiao@cindasc.com
华东区销售	方威	18721118359	fangwei@cindasc.com
华东区销售	俞晓	18717938223	yuxiao@cindasc.com
华东区销售	李贤哲	15026867872	lixianzhe@cindasc.com
华东区销售	孙僮	18610826885	suntong@cindasc.com
华东区销售	贾力	15957705777	jiali@cindasc.com
华东区销售	石明杰	15261855608	shimingjie@cindasc.com
华东区销售	曹亦兴	13337798928	caoyixing@cindasc.com
华东区销售	王赫然	15942898375	wangheran@cindasc.com
华南区销售总监	王留阳	13530830620	wangliuyang@cindasc.com
华南区销售副总监	陈晨	15986679987	chenchen3@cindasc.com
华南区销售副总监	王雨霏	17727821880	wangyufei@cindasc.com
华南区销售	刘韵	13620005606	liuyun@cindasc.com
华南区销售	胡洁颖	13794480158	hujieying@cindasc.com
华南区销售	郑庆庆	13570594204	zhengqingqing@cindasc.com
华南区销售	刘莹	15152283256	liuying1@cindasc.com
华南区销售	蔡静	18300030194	caijing1@cindasc.com
华南区销售	聂振坤	15521067883	niezhenkun@cindasc.com
华南区销售	宋王飞逸	15308134748	songwangfeiyi@cindasc.com

分析师声明

负责本报告全部或部分内容的每一位分析师在此申明，本人具有证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师，以勤勉的职业态度，独立、客观地出具本报告；本报告所表述的所有观点准确反映了分析师本人的研究观点；本人薪酬的任何组成部分不与、不与，也将不会与本报告中的具体分析意见或观点直接或间接相关。

免责声明

信达证券股份有限公司（以下简称“信达证券”）具有中国证监会批复的证券投资咨询业务资格。本报告由信达证券制作并发布。

本报告是针对与信达证券签署服务协议的签约客户的专属研究产品，为该类客户进行投资决策时提供辅助和参考，双方对权利与义务均有严格约定。本报告仅提供给上述特定客户，并不面向公众发布。信达证券不会因接收人收到本报告而视其为本公司的当然客户。客户应当认识到有关本报告的电话、短信、邮件提示仅为研究观点的简要沟通，对本报告的参考使用须以本报告的完整版本为准。

本报告是基于信达证券认为可靠的已公开信息编制，但信达证券不保证所载信息的准确性和完整性。本报告所载的意见、评估及预测仅为本报告最初出具日的观点和判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会出现不同程度的波动，涉及证券或投资标的的历史表现不应作为日后表现的保证。在不同时期，或因使用不同假设和标准，采用不同观点和分析方法，致使信达证券发出与本报告所载意见、评估及预测不一致的研究报告，对此信达证券可不发出特别通知。

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、意见及推测仅供参考，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做出邀请。

在法律允许的情况下，信达证券或其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，并可能会为这些公司正在提供或争取提供投资银行业务服务。

本报告版权仅为信达证券所有。未经信达证券书面同意，任何机构和个人不得以任何形式翻版、复制、发布、转发或引用本报告的任何部分。若信达证券以外的机构向其客户发放本报告，则由该机构独自为此发送行为负责，信达证券对此等行为不承担任何责任。本报告同时不构成信达证券向发送本报告的机构之客户提供的投资建议。

如未经信达证券授权，私自转载或者转发本报告，所引起的一切后果及法律责任由私自转载或转发者承担。信达证券将保留随时追究其法律责任的权利。

评级说明

投资建议的比较标准	股票投资评级	行业投资评级
本报告采用的基准指数：沪深 300 指数（以下简称基准）； 时间段：报告发布之日起 6 个月内。	买入 ：股价相对强于基准 20% 以上；	看好 ：行业指数超越基准；
	增持 ：股价相对强于基准 5%~20%；	中性 ：行业指数与基准基本持平；
	持有 ：股价相对基准波动在±5%之间；	看淡 ：行业指数弱于基准。
	卖出 ：股价相对弱于基准 5% 以下。	

风险提示

证券市场是一个风险无时不在的市场。投资者在进行证券交易时存在赢利的可能，也存在亏损的风险。建议投资者应当充分深入地了解证券市场蕴含的各项风险并谨慎行事。

本报告中所述证券不一定能在所有的国家和地区向所有类型的投资者销售，投资者应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专业顾问的意见。在任何情况下，信达证券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。