西南证券
SOUTHWEST SECURITIES

计算机行业2024年中期投资策略

把握AI商业化进展，聚焦结构亮点

www.swsc.com.cn

西南证券研究发展中心
计算机研究团队
2024年7月

核心观点

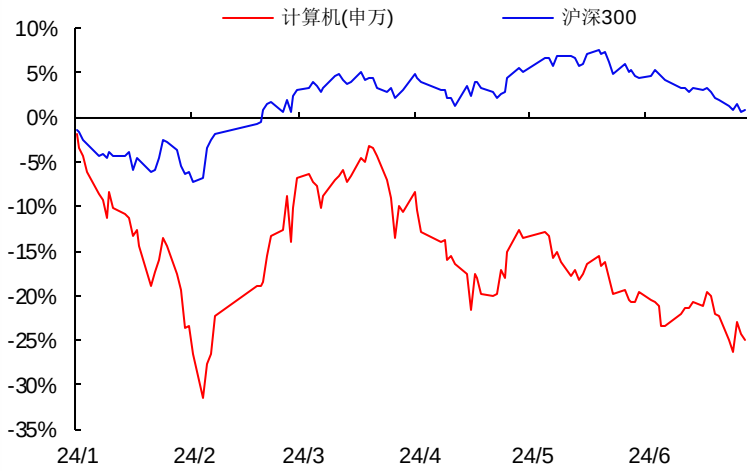
- **底部特征显现，有望在H2迎来边际缓和。**总体来看，当前计算机公司的估值水平、机构持仓水平皆回落至低位区间；叠加海外流动性中期放松预期，AI技术周期与科技政策周期共振，行业估值具备向上拔升的潜力。
- **利润修复的基础条件已经具备。1）供给端已逐步出清：**计算机公司企业2023年纷纷开始重视内部改革和提质增效，对人员招聘实施严格措施，整体人员和薪酬已得到有效控制；**2）需求端部分领域曙光初现：**计算机具备典型后周期属性，结构性机会下半年可以重点关注。部分领域在政策支持、资金配套下，景气度已开始回暖，“收入-成本”的正向剪刀差或将扩大，利润弹性逐步得到释放。
- 从方向选择看，**1）AI仍是明显的产业趋势，**算力是兑现度较高的一环，端侧AI（PC、手机、智驾）可以持续关注落地和商业化进展。**2）关注有明确政策牵引、且已带动招投标落地的领域，**包括能源IT、大交通（低空+车路云）等。**3）关注预期有政策变革的方向，**包括财税IT、信创、数据要素、医疗IT等。**4）关注中长期景气度，收并购浪潮或将启动的工业软件。**
- **即将进入半年报披露密集期，短期围绕基本面和业绩兑现进行布局胜率更高，建议辅以自下而上的角度布局绩优股。**
- **相关标的：**1）算力：海光信息、中科曙光、寒武纪、神州数码、工业富联、浪潮信息等；2）能源IT：科远智慧、国能日新、朗新集团等；3）“大交通”：莱斯信息、千方科技等；4）税改：税友股份、博思软件、中科江南等；5）信创：纳思达、金山办公、顶点软件等；6）数据要素：国新健康、拓尔思等；7）工业软件：华大九天、宝信软件、索辰科技等；8）绩优股：道通科技、锐明技术、新国都等。
- **风险提示：**国际博弈加剧；下游需求不及预期；原材料价格上涨；板块政策发生重大变化；研发进度不及预期；行业竞争加剧等。

目 录

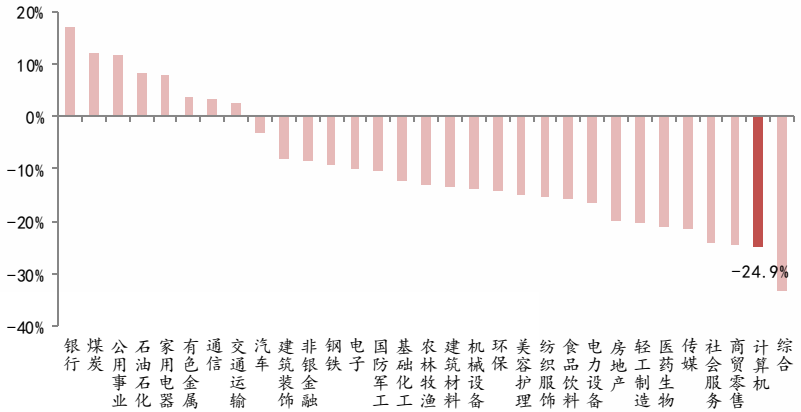
1.1 行情复盘：业绩承压拖累板块下行

- 2024年春节后，受益于Sora文生视频大模型的发布及政府工作报告首提“低空经济”等催化，计算机行业跟随市场开启一轮反弹；从3月下旬开始，指数受板块内公司一季度业绩表现不佳拖累，开启震荡下行。
- 截至6月30日，申万计算机指数年初至今下跌24.9%，跑输沪深300指数约25.8个百分点，居于全行业30/31。
- 板块内涨幅前三分别为莱斯信息（+72.4%）、荣科科技（+49.9%）、淳中科技（+43.8%）；跌幅前三分别为*ST银江（-79.6%）、ST英飞拓（-77.8%）、*ST左江（-75.9%）。

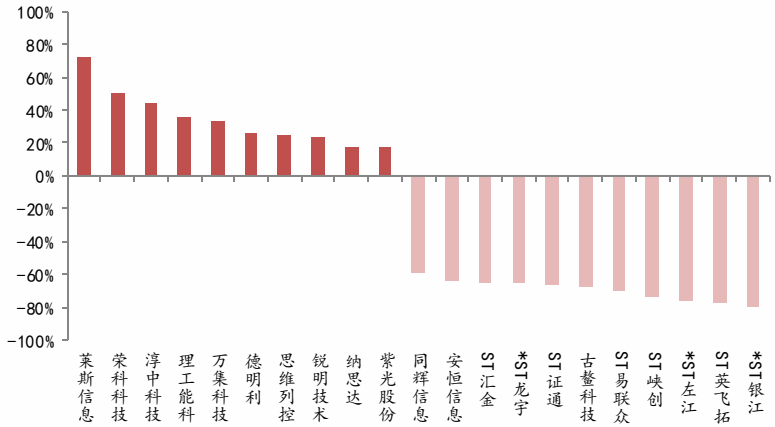
计算机指数相对沪深300走势



年初至今申万一级指数涨跌幅



计算机行业涨跌幅前十个股

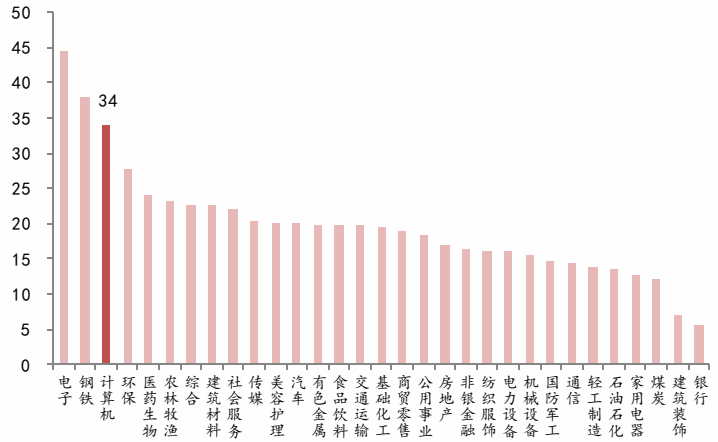


数据来源：Wind，西南证券整理

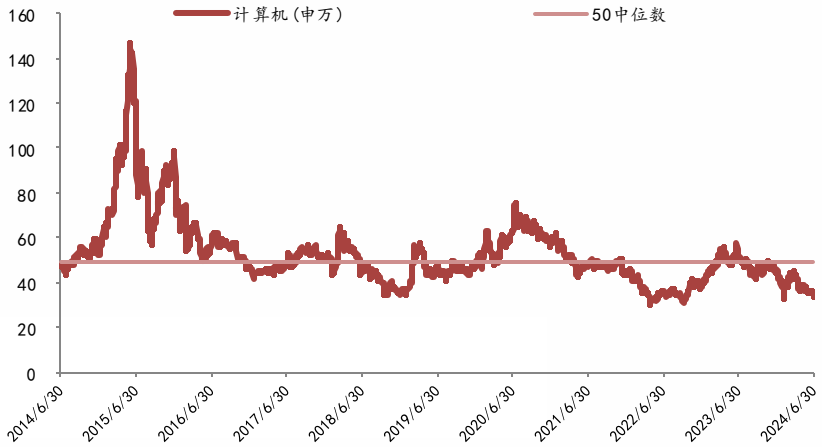
1.1 行情复盘：估值低于中枢水平，持仓比例回到低位

- **横向看**：截至2024年6月30日，计算机行业PE(TTM，整体法，剔除负值)为34倍，在申万一级行业中处于较高水平，市场仍持续认可其成长性。
- **纵向看**：计算机指数PE过去十年中位数水平49倍，经过近几个月回调，当前计算机PE大幅低于过去十年中位数，具备一定的配置价值。
- 从公募基金持仓看，2024Q1计算机行业基金重仓股合计市值为725.0亿元，环比下滑30.3%；占行业总市值比例为2.78%，环比下滑0.78pp，配置比例回至低位。

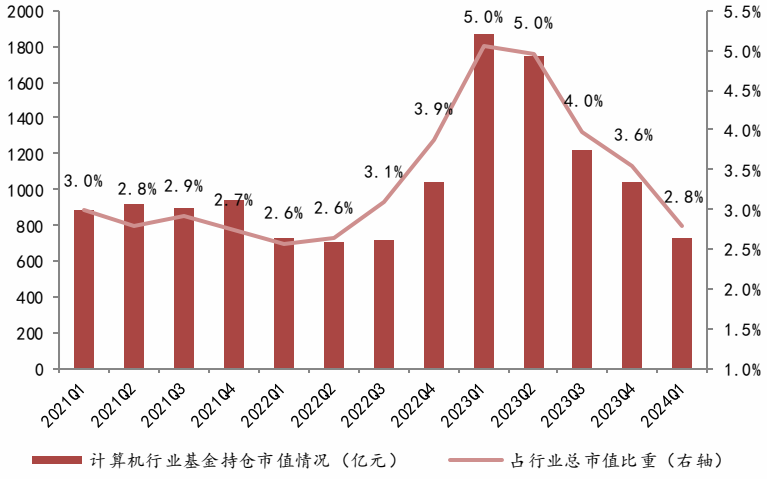
申万一级行业市盈率（TTM整体法）



申万计算机市盈率（TTM整体法）



计算机行业基金持仓市值情况



数据来源：Wind，西南证券整理

1.2 24Q1收入回顾：板块复苏分化，龙头公司收入韧性更强

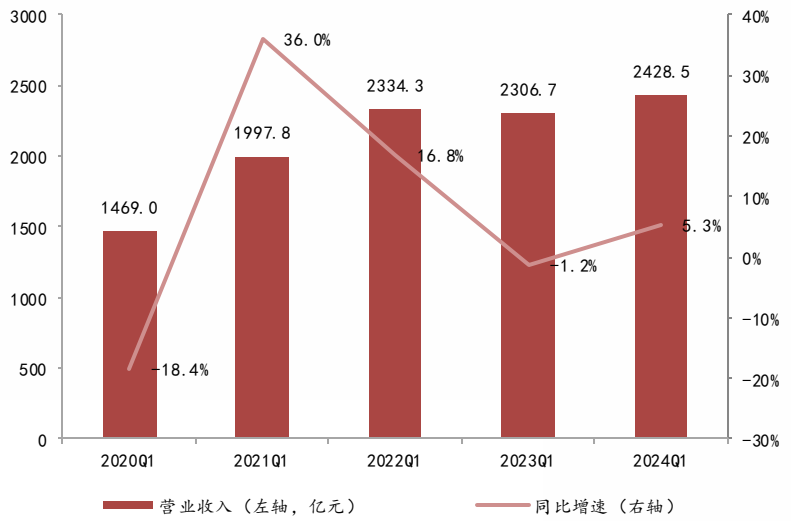
□ 2024Q1营收端小幅增长，主要系龙头公司带动

2024Q1，计算机板块整体实现营收2428.5亿元，同比增长5.3%，增速相较23Q1（同比-1.2%）显著改善。

从中位数角度看，2024Q1营收增速的中位数为2.5%，低于整体法，表明一季度龙头公司收入端增长韧性更强。

从增速的分布看，2024Q1营收增速大于50%的有26家，小于0%的有141家，增速区间同环比均有下移，板块复苏分化较大。

计算机板块2020Q1-2024Q1营收及增速



板块2024Q1收入增速情况

	20Q1	21Q1	22Q1	23Q1	24Q1
营收增速（整体法）	-18.4%	36.0%	16.8%	-1.2%	**5.3%**
营收增速（中位数法）	-5.0%	25.7%	8.9%	6.8%	**2.5%**

	大于50%	30%-50%	10%-30%	0%-10%	小于0%
2023Q1营收yoy	**39**	33	71	69	**120**
2023Q4营收yoy	**37**	28	69	54	**144**
2024Q1营收yoy	**26**	20	67	78	**141**

数据来源：Wind，西南证券整理

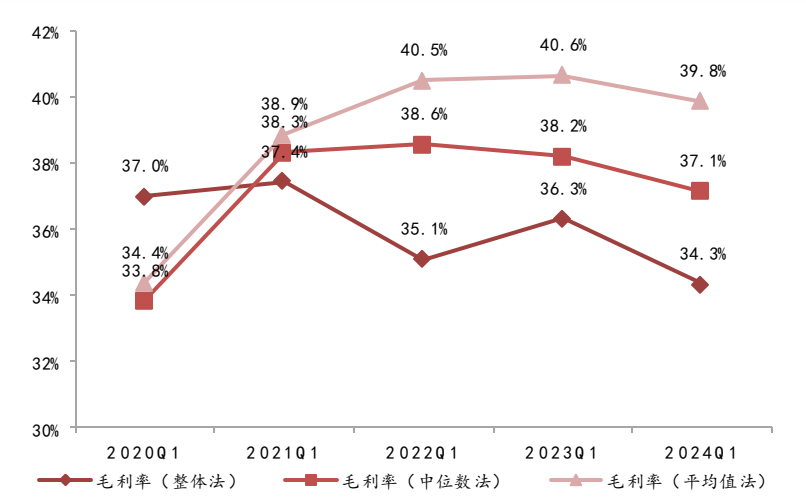
1.2 24Q1成本费用回顾：毛利端有所承压，研发继续加码

□ 行业内卷导致毛利率承压，企业继续加码研发抢占AI时代高地

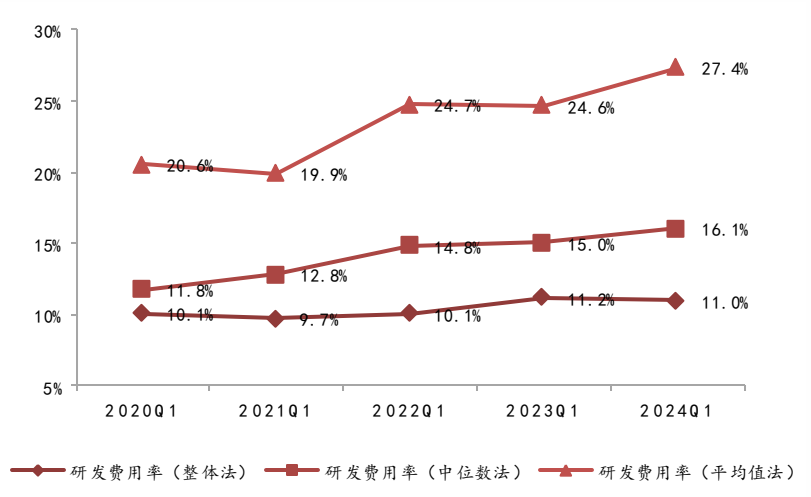
2024Q1，行业整体毛利率为34.3%，同比下滑2.0pp，中位数法和平均值法下的毛利率与整体法差距较大，均呈下降趋势，我们认为主要系近两年下游客户IT开支收紧，信息系统建设项目内卷严重所致。

2024Q1，计算机板块整体研发费用达266.4亿元，同比增长3.1%；研发费用率（整体法）/（中位数法）/（平均值法）分别为11.0%、16.1%、27.4%，其中平均值法和中位数法下研发费用率提升明显，表明中小企业在收入承压的情况下继续加码研发投入，力求借助本轮技术变革抢占新一轮战略高地。

计算机板块2020Q1-2024Q1毛利率情况



计算机板块2020Q1-2024Q1研发费用率情况

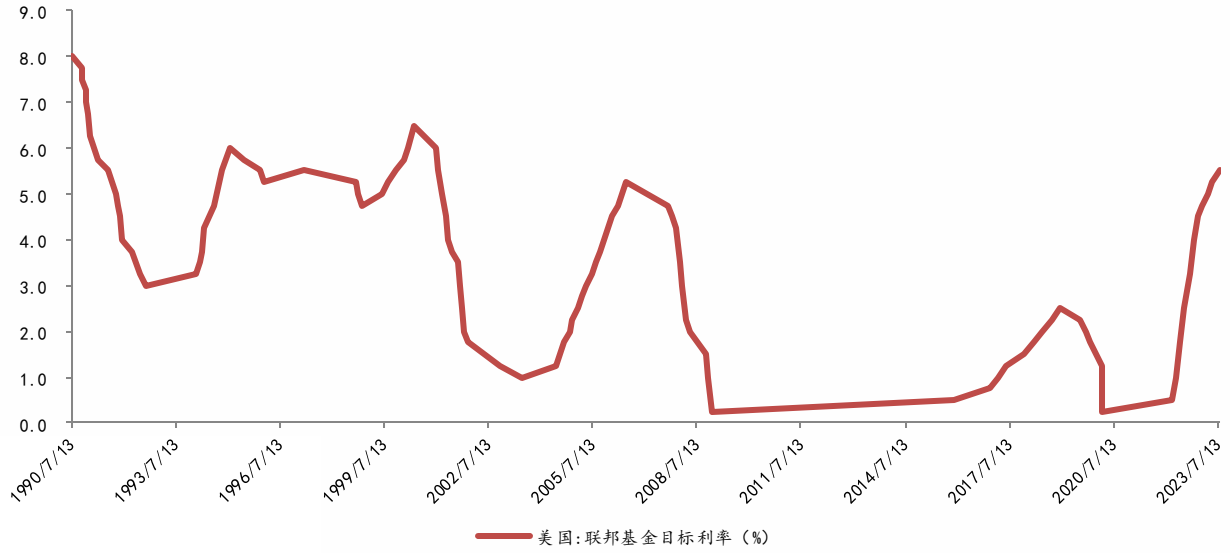


数据来源：Wind，西南证券整理

1.3 美联储降息预期升温，宏观压制要素逐步清退

- **行业走势与流动性环境强相关。**成长股的久期更长，对内外流动性环境更加敏感，计算机公司作为成长赛道主力军，呈现明显的利率敏感型资产特征。
- **欧元区已开始降息。**6月6日，欧洲央行降息25个基点，将三大利率分别降至4.25%、3.75%、4.50%，为历时接近两年的加息周期划上句号，也为2019年以来首次降息，是G7成员国中第二个降息的央行。
- **美国部分数据转弱。**当前美国仍处于历史最长的加息周期当中，但近期公布的ISM制造业指数、ADP就业人数等均不及市场预期，显示下半年美国经济或将有所转弱。根据芝加哥商品交易所的“美联储观察工具”数据，美国到9月累计降息25基点的概率为70.8%。

1990年来美国联邦基金目标利率走势



数据来源：Wind，西南证券整理

1.3 科技支持政策加码，科创资金活水注入

- 新“国九条”出台强调分红与公司质地，计算机行业中小市值公司占比较高，短期市场情绪受影响较大；但随之出台的“科企16条”、“科八条”等，**对科技企业的融资、收并购等提出支持性举措；同时大基金三期、科技再贷款等为企业提供关键资金支持。**

- 4月19日，证监会制定了《资本市场服务科技企业高水平发展的十六项措施》，从上市融资、并购重组、债券发行、私募投资等全方位提出支持性举措，进一步健全资本市场功能，优化资源配置，更大力度支持科技企业高水平发展，促进新质生产力发展。
- 5月31日，求是网发布总书记发言稿《发展新质生产力是推动高质量发展的内在要求和重要着力点》，提出大力推进科技创新，以科技创新推动产业创新等五大关键点。
- 6月19日，证监会发布《关于深化科创板改革服务科技创新和新质生产力发展的八条措施》（以下简称《八条措施》），特别强调优先支持新产业新业态新技术领域突破关键核心技术的“硬科技”企业在科创板上市，更大力度支持并购重组。
- 6月25日，中共中央政治局常委丁薛祥首次作为中央科技委员会主任出席全国科技大会并发言，强调“要以新型举国体制推进科技创新，找准重大攻关任务，凝聚力量协同攻坚，夯实基础研究根基，加快关键核心技术攻关”。

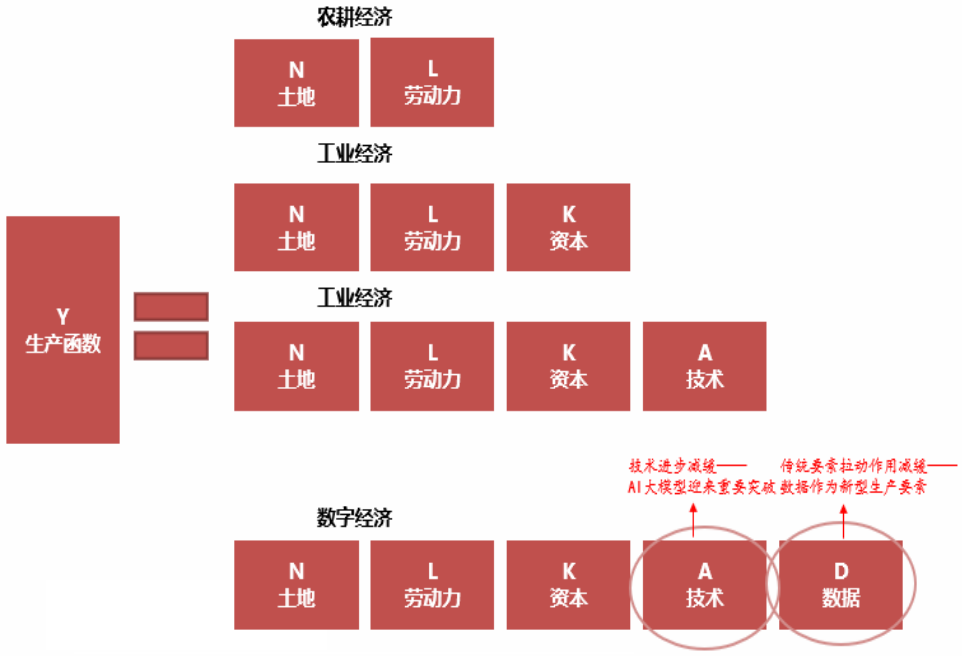
近期科技领域加大资金投入

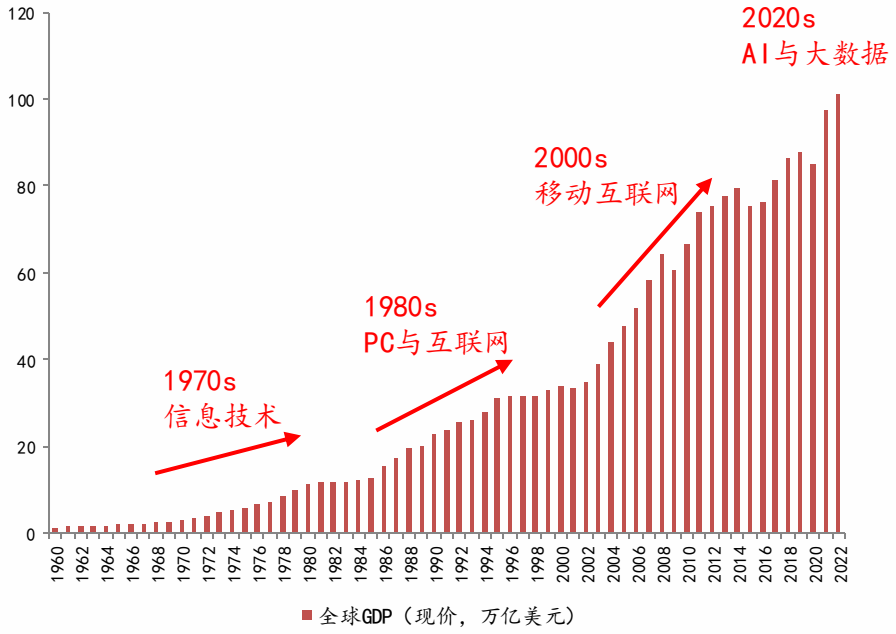
时间	事件	主要内容
5月13日	万亿超长期特别国债发行	为系统解决强国建设、民族复兴进程中一些重大项目建设的资金问题，从今年开始拟连续几年发行超长期特别国债，专项用于国家重大战略实施和重点领域安全能力建设，今年先发行1万亿元。
5月24日	大基金三期成立	注册资本达3440亿元，超过第一期的987.2亿元和第二期的2041.5亿元
6月14日	科技再贷款加速落地	首笔科技创新贷款已发放，首批有近7000家符合条件的企业；央行和科技部正在组织开展第二批32万余家科技型企业的创新积分评价，后续其他贷款将陆续投放

数据来源：Wind，西南证券整理

1.3 从宏观维度认识新质生产力——AI与大数据

- 当前阶段，**新质生产力成为推动高质量发展的内在要求和重要着力点。**
- **数据作为新型生产要素，不仅能够直接参与生产，还能够促进其他生产要素的投入并赋能其他要素，产生乘数效应。**
- **本轮AIGC大模型的突破，是第三次工业革命以来新一轮的通用技术创新，有望解放劳动力，实现对千行百业的赋能。**



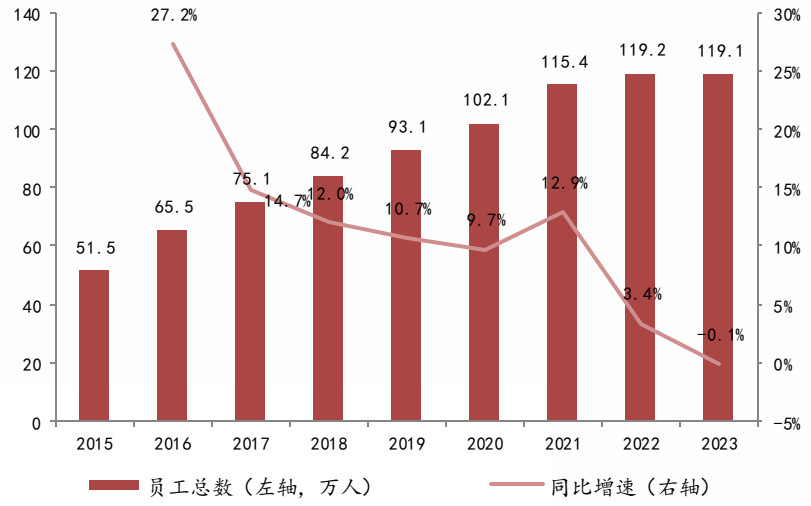


数据来源：Worldbank，西南证券整理

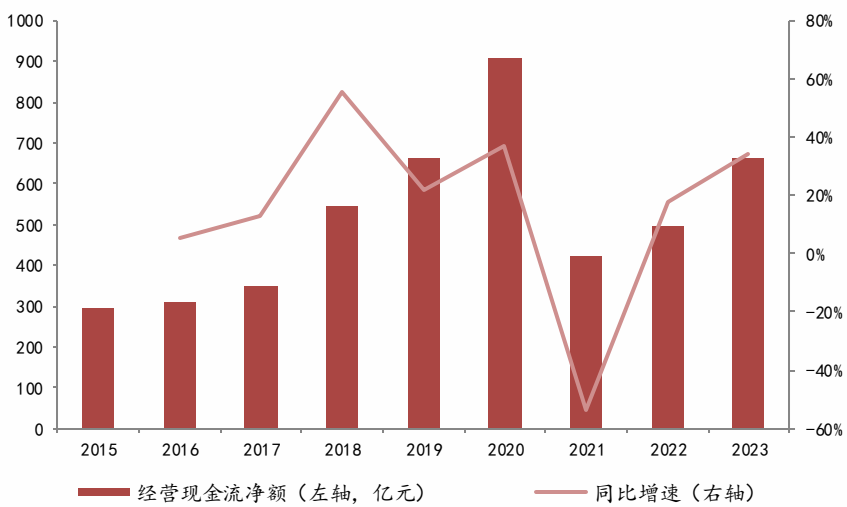
1.3 行业内生经营提质，关注需求复苏带来的业绩弹性

- 计算机行业后周期属性明显，但伴随企业进行人员及费用管控，利润修复的基础条件已经具备。
- **行业供给正出清，人效拐点或将到来。**2023年板块内公司员工人数逐步调整，后续期间费用率有望得到有效控制。**经营现金流净额大幅增长，经营质量有所改善。**2023年板块内公司经营现金流净额为662.9亿元，同比增长33.9%，回款情况逐步好转。
- **部分领域订单有所恢复，Q3起有望逐季报表兑现。**分板块看，年初至今部分领域订单已初现修复迹象，例如电力IT、医疗IT、工业软件等，考虑项目周期，预期将于Q3/Q4逐季开始兑现进报表。

计算机板块员工总数情况



计算机板块经营现金流净额情况



数据来源：工信部，Wind，西南证券整理

1.3 核心观点

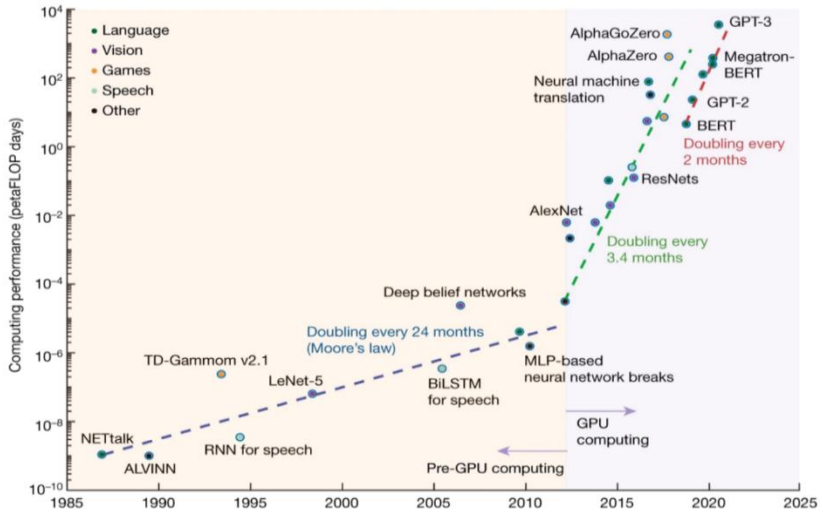
- **底部特征显现。**总体来看，当前计算机公司的估值水平、机构持仓水平皆回落至低位区间；叠加海外流动性中期放松预期，AI技术周期与科技政策周期共振，行业估值具备向上拔升的潜力。
- **利润修复的基础条件已经具备。1）供给端已逐步出清：**计算机公司企业2023年纷纷开始重视内部改革和提质增效，整体人员和薪酬已得到有效控制；**2）需求端，部分领域曙光初现：**计算机具备典型后周期属性，结构性机会可以。部分领域在政策支持、资金配套下，景气度已开始回暖，“收入-成本”的正向剪刀差或将扩大，利润弹性逐步得到释放。
- 从方向选择看，**1）AI仍是明显的产业趋势，**算力是兑现度较高的一环，端侧AI（PC、手机、智驾）可以持续关注落地和商业化进展。**2）关注有明确政策牵引、且已带动招投标落地的领域，**包括能源IT、大交通（低空+车路云）等。**3）关注预期有政策变革的方向，**包括财税IT、信创、数据要素、医疗IT等。**4）关注中长期景气度，收并购浪潮或将启动的工业软件。**
- 考虑到目前宏观环境，**叠加即将进入半年报披露密集期，短期围绕基本面和业绩兑现进行布局胜率更高，建议辅以自下而上的角度布局绩优股。**
- 相关标的：1）算力：海光信息、中科曙光、寒武纪、神州数码、工业富联、浪潮信息等；2）能源IT：科远智慧、国能日新、朗新集团等；3）“大交通”：莱斯信息、千方科技等；4）税改：税友股份、博思软件、中科江南等；5）信创：纳思达、金山办公、顶点软件等；6）数据要素：国新健康、拓尔思等；7）工业软件：华大九天、宝信软件、索辰科技等；8）绩优股：道通科技、锐明技术、新国都等。

目 录

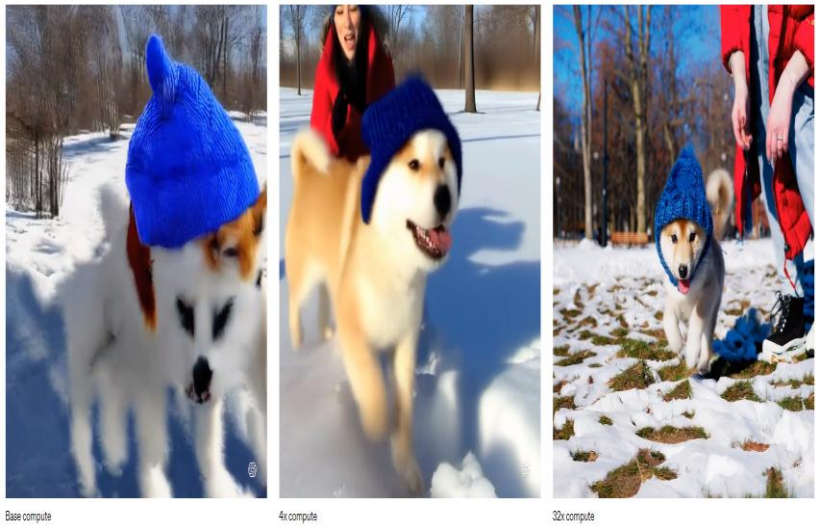
2.1 Scaling Law持续有效，算力需求未至天花板

- **算力是大模型取得更高性能的核心：**随着参数量与数据量的同步扩大，模型效果会出现“涌现”。从目前已经发布的大模型和头部厂商的研发方向来看，“大力出奇迹”仍是主要优化方向之一，Scaling Law的上限还远未达到。
- **多模态大模型对于算力的需求更高：**多模态大模型需要接受文字、图像、语音等不同类型的数据处理，涉及到的非结构化数据较多，算法亦更为复杂，在训练和推理阶段相较文本类的LLM，消耗的算力更多。伴随OpenAI、谷歌、百度等纷纷投入多模态大模型的研发，后续有望持续带动算力的需求增长。

大模型对于算力的需求每2个月翻倍



Sora在不同算力支持下的视频生成效果

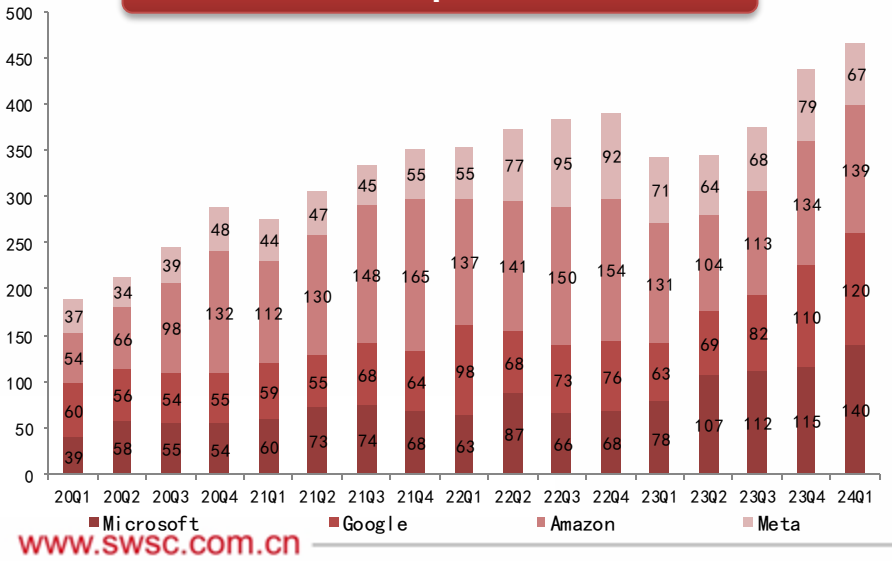


数据来源：《Scaling Laws for Neural Language Models》，Nature，西南证券整理

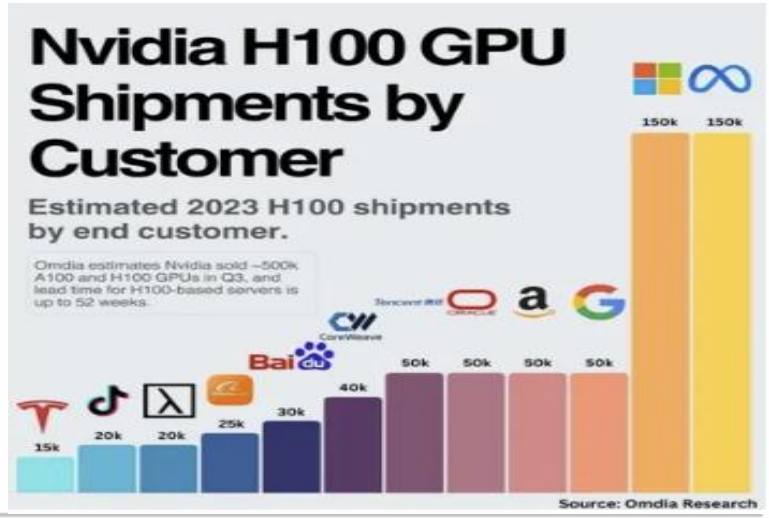
2.1 海外大厂Capex指引乐观，算力景气度持续验证

- **Microsoft**：24Q1资本开支140亿美元（包含31.5亿融资租赁），同比+79%，环比+22%，预期资本支出将逐季增加。
- **Google**：24Q1资本开支120亿美元，同比+91%，环比+9%，预期每季度资本开支保持在Q1水平或以上。
- **Amazon**：24Q1资本开支139亿美元，同比+6%，环比+4%，预期后续三个季度资本支出比24Q1更高。
- **Meta**：24Q1资本开支67亿美元（包含3亿融资租赁），同比-6%，环比-15%，上调24年全年资本支出至350-400亿美元（此前为300-370亿美元）。

海外互联网大厂Capex情况（亿美元）



2023年H100客户采购量

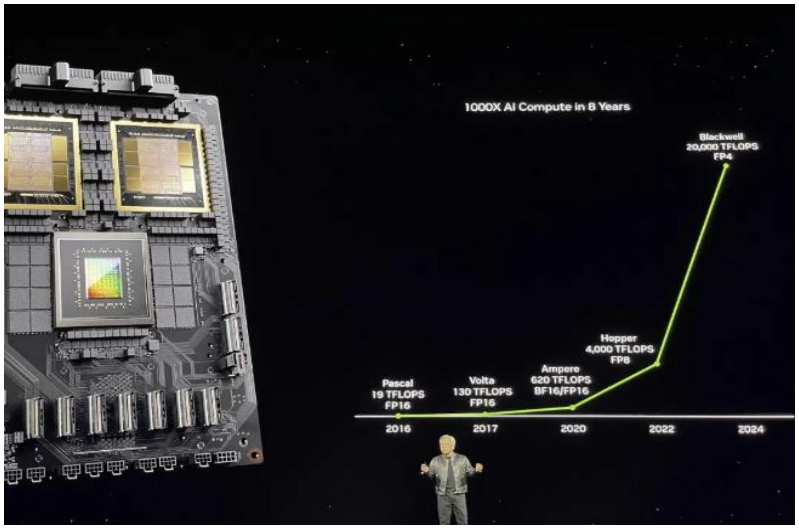


数据来源：各公司财报，Omidia，西南证券整理

2.1 英伟达业绩超预期，架构快速迭代

- **英伟达开启“Tick-Tock”一年一迭代的更新周期**
- 2023年11月，英伟达发布人工智能芯片H200，针对超大规模的大模型训练和推理进行显存升级，首次采用HBM3e，显存容量达到141GB，显存带宽达到4.8TB/s。
- 2024年3月，英伟达发布全新Blackwell架构的芯片系列，除了工艺迭代、HBM升级、双die设计外，还内置第二代Transformer引擎，采用“降精度、提性能”的方式，首度引入FP4新格式，峰值算力达到40PFLOPS。
- 2024年6月，英伟达表示将开启一年一更新的迭代周期，将在25/26/27年分别推出Blackwell Ultra、Rubin、Rubin Ultra架构，其中在2026年推出的下一代AI芯片平台Rubin，预计将采用HBM4和3nm工艺。

8年间英伟达单卡性能提升1000倍



B系列与H系列芯片性能对比

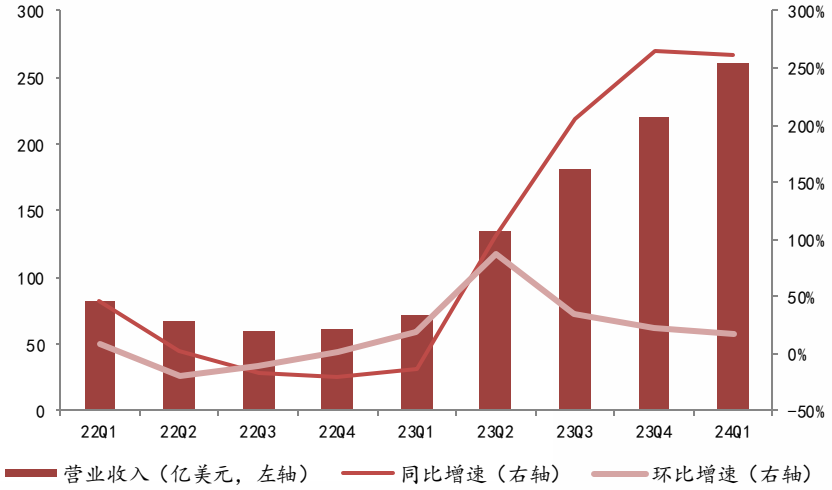
Blackwell GPU		
FP8	20 PFLOPS	2.5X Hopper
NEW FP6	20 PFLOPS	2.5X
NEW FP4	**40 PFLOPS**	**5X**
HBM Model Size	740B param	6X
HBM Bandwidth	34T param/sec	5X
NVLINK All-Reduce with SHARP	7.2 TB/s	4X

数据来源：英伟达，半导体行业观察，The Next Platform，西南证券整理

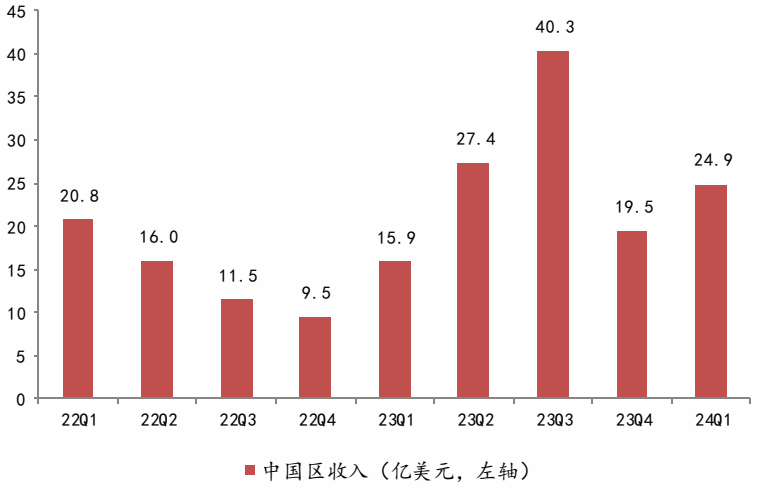
2.1 英伟达业绩超预期，架构快速迭代

- **英伟达Q1业绩超预期，后续指引仍然乐观**
  - 英伟达24Q1实现收入260亿美元，同比增长262%，环比增长19%，高于市场246亿美元的指引；其中数据中心业务实现收入226亿美元，同比增长427%，环比增长23%，同样超出市场预期的221亿美元。
  - 公司预期24Q2实现收入280亿美元（上下浮动2%），强于市场预期的268亿美元。
  - 公司表示，当前B系列芯片已经全面投入生产，预计将在Q2发货、Q3加速、Q4投放至数据中心，消除市场对于H、B代际切换的担忧。
  - 根据SemiAnalysis，英伟达有望在未来几个月交付超过100万颗H20芯片，考虑单颗芯片售价1.2-1.3万美元，2024年全年或可贡献超过120亿美元营收。

英伟达总收入情况



英伟达中国大陆地区收入情况

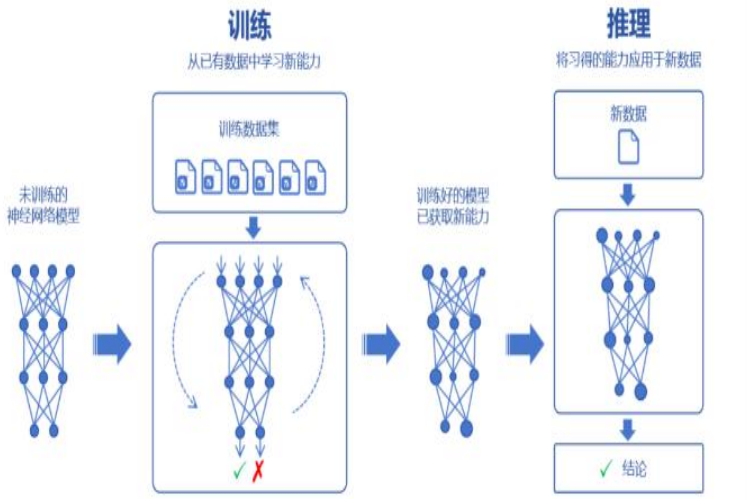


数据来源：英伟达，Semianalysis，西南证券整理

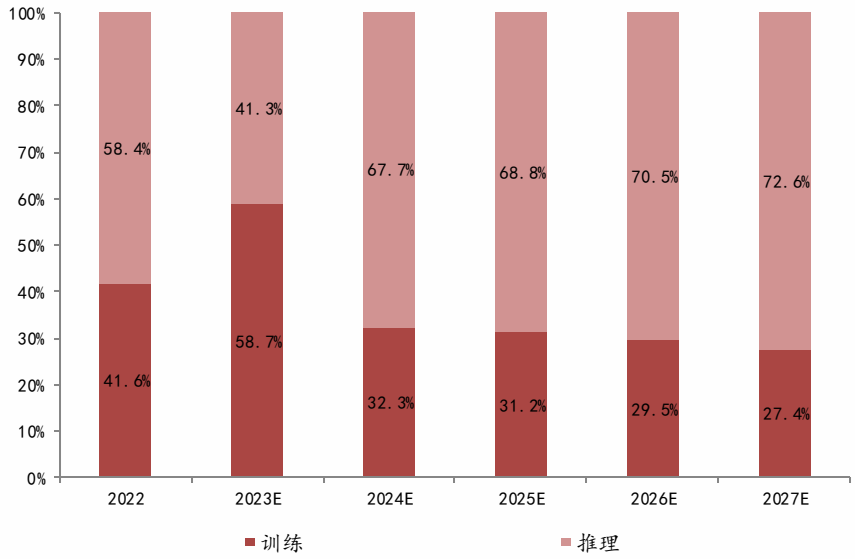
2.1 算力需求结构向推理倾斜

- 伴随模型的逐步成熟和应用的落地投产，算力需求的重心向推理转移。
- 根据英伟达，2023年大约40%的数据中心收入来自AI推理。
- 根据IDC、浪潮信息联合发布的《2023-2024年中国人工智能计算力发展评估报告》，2023年国内AI服务器工作负载中训练:推理的占比约为6:4；到2027年，用于推理的工作负载将达到72.6%。
- 幻方DeepSeek-V2发布创新性MLA架构，开启国内大模型API降价浪潮；而推理成本的降低有望进一步吸引开发者加入促进生态的繁荣，拉动大模型推理侧的需求。

训练与推理的原理



2022-2027年中国AI服务器工作负载及预测



数据来源：英伟达，幻方，IDC，浪潮信息，西南证券整理

2.1 推理场景用量提升，关注ASIC份额提升机遇

- **算力开支高成本，倒逼科技巨头加速自研。**近年来，海内外各大互联网厂商纷纷自研芯片，虽然综合性能，尤其是算力和带宽，较英伟达仍有一定差距，但更多面向特定领域任务进行优化，在功耗和成本方面优势明显。
- **伴随推理场景的铺开，AISC市占率有望快速提升。**根据Marvell，其2023年数据中心TAM为210亿美元，其中加速定制计算为66亿美元；预计2028年加速定制计算TAM将达到429亿美元，CAGR为45%。

各大厂自研AI芯片和CPU产品情况

AI芯片布局

厂商	Google	Amazon		Meta	Tesla	Microsoft
芯片	Cloud TPU v5e	AWS Inferentia 2	AWS Trainium	MTIA v1	Dojo D1	Azure Maia 100
推出时间	2023年8月29日	2019年	2022年	2023年5月18日	2021年	2023年11月16日
代际	第五代	第二代	第一代	第一代	第一代	第一代
工艺&制程	最多允许256个芯片互连，总带宽超过400 Tb/s，INT8性能达到100 petaOps	5nm	5 nanometer	7nm；内部内存可以从128MB扩展到128GB	7nm；500 亿个晶体管；400W TDP	5nm；1050亿个晶体管
用途	专用于大中型训练与推理；支持Google cloud和聊天机器人Bard等应用产品	推理芯片	训练芯片	推理芯片	训练芯片；用于搭建Tesla Dojo超算平台，以支持自动驾驶和机器人业务	专门用于云端训练和推理；支持Microsoft Azure OpenAI服务和Microsoft Copilot(Bing Chat)等应用产品

CPU芯片布局

厂商	Google		Amazon	Microsoft
芯片名称	Cypress	Maple	Graviton 3	Azure Cobalt 100
推出时间	预计2025年部署上线	预计2025年部署上线	2021年	2023年11月16日
代际	第一代	第一代	第三代	第一代
工艺&制程	5nm；基于Arm	5nm；基于Arm	5nm；550亿个晶体管	5nm；基于Arm；128核
用途	自研用于运营数据中心的 Plan A 处理器	自研用于运营数据中心的 Plan B 处理器	云原生通用处理器	旨在用于执行常规计算任务，如为微软Teams提供动力；暂时没有销售计划，更倾向于供内部使用

数据来源：The information，Marvell，各公司官网，西南证券整理

2.1 国产芯片补缺

- **各省市印发算力建设规划：**算力已上升为国家战略资源之一，多地印发算力基础设施建设规划，提出实现算力调度和交易，构建公共算力平台。
  - **上海：**到2025年，本市数据中心算力超过18000PFLOPS（FP32），新建数据中心绿色算力占比超过10%，集聚区新建大型数据中心综合PUE降至1.25以内，绿色低碳等极达到4A级以上。
  - **深圳：**建设城市级智能算力平台，整合深圳市算力资源，建设城市级算力统筹调度平台，实现“算力一网化、统筹一体化、调度一站式”
  - **北京：**按照集约高效原则，分别在海淀区、朝阳区建设北京人工智能公共算力中心、北京数字经济算力中心。在人工智能产业聚集区新建或改建升级一批人工智能商业化算力中心，加强国产芯片部署应用，推动自主可控软硬件算力生态建设。

省市	时间	文件	要求
河北	2023.01	《加快建设数字河北行动方案》	到2023年底，数据中心在营标准机柜95万架，算力总规模20EFLOPS；到2025年，数据中心在营标准机柜165万架，算力总规为35EFLOPS
贵州	2023.3	《面向全国的算力保障基地建设规划》	到2023年，标准机架25万架，大型/超大型数据中心数量20个，算力总规模2EFLOPS；到2025年，标准机架80万架，大型/超大型数据中心数量26个，算力总规模10EFLOPS；
上海	2023.04	《上海市推进算力资源同意调度指导意见》	到2023年底，可调度智能算力达到1000PFLOPS（FP16）以上；到2025年，数据中心算力超过18000PFLOPS（FP32）
北京	2023.05	《北京市加快建设具有全球影响力的人工智能创新策源地实施方案（2023-2025年）》	按照集约高效原则，分别在海淀区、朝阳区建设北京人工智能公共算力中心、北京数字经济算力中心。在人工智能产业聚集区新建或改建升级一批人工智能商业化算力中心，加强国产芯片部署应用
深圳	2023.05	《深圳市加快推动人工智能高质量发展高水平应用行动方案（2023-2024年）》	建设城市级智能算力平台，打造大湾区智能算力枢纽，建设企业级智能算力平台

数据来源：各政府官网，西南证券整理

2.1 国产芯片补缺

- 当前国内海光信息、寒武纪、华为、燧原、壁仞等一二线厂商推出的推理端AI芯片已基本达到可用状态；训练端芯片虽然较H100性能还有显著差距，但已在国家智算中心得到广泛应用，并且已与多数互联网大厂展开适配合作，生态成熟度有望进一步追赶。

公司	海光	寒武纪	华为	百度	阿里	壁仞科技	燧原科技	天数智芯	摩尔线程
GPU	深算一号	MLU370-X8	昇腾910	昆仑芯R200	含光800（NPU）	BR100	云燧T20&i20	天垓100	春晓
工艺	7nm	7nm	7nm	7nm	12nm	7nm	12nm	7nm	7nm
算力	4096 core（64CUs） 1.5 GHz（FP64） 1.7GHz（FP32） 国内唯一同时支持全精度和半精度训练的加速计算芯片	96 TFLOPS@FP16 96 TFLOPS@BF16 256 TOPS @ INT8	320TFLOPS@FP16 640 TOPS@INT8	128TFLOPS@FP16 256TOPS@INT8	78563 IPS	1024TFLOPS@BF16 支持FP32、BF16、FP16、INT8等主流数据精度	32 TFLOPS@FP32 128TFLOPS@TF32 128TFLOPS@FP16/BF16 256TOPS/INT8	37 TFPLOPS@FP32 147TFLOPS@FP16/BF16 支持INT32 INT15多精度混合训练	14.4 TFLOPS@FP32
功耗	350W TDP	75W TDP	310W TDP	120W TDP	-	550W TDP	300W TDP&150W TDP	250W 板级功耗	255W TDP
接口	PCIe4.0 × 16 lane xGMI × 2, 最高184 GB/s	PCIe4.0 × 16 lane	集成HCCS、PCIe 4.0、RoCE v2	PCIe4.0 × 16 lane	-	PCIe5.0 × 16 lane CXL2.0	PCIe4.0 × 16 lane 300GB/s片间互联	PCIe4.0 × 16 lane 共享64GB/s 主控双向带宽 共享64GB/s 片间互联带宽	PCIe5.0 × 16 lane
用途	云端训练&推理	云端训练&推理	云端训练	云边端通用	云端推理	深度学习、云端通用计算	云端训练	云端训练	云端训练&推理，终端图形渲染

数据来源：华为，寒武纪，壁仞科技等公司官网，芯东西，西南证券整理

2.1 国产芯片补缺

- **昇腾处理器训练推理均有布局**
- 昇腾310：2018年发布，主要用于推理，主打高能效，在典型配置下可以输出16TOPS@INT8, 8TFLOPS@FP16，功耗仅为8W。
- 昇腾910：2019年发布，主要用于训练，主打高算力，支持云边端全栈场景应用。在典型配置下可以输出320TFLOPS@FP16，640 TOPS@INT8，功耗为310W。
- 以上两颗芯片均采用华为自研的达芬奇架构，通过独创的16*16*16*16的3D Cube设计，每时钟周期可以进行4096个16位半精度浮点MAC运算，整个AI Core可以看成是一个相对简化的微处理器架构，包含计算、存储、控制单元，并含有指令流水线的设计，是华为面向AI计算专门设计的NPU架构。
- **关注昇腾新款芯片迭代进度。**

昇腾310、910关键参数

关键特性

Architecture	• HUAWEI Da Vinci
Computing Engine	• 3D Cube
Performance	• 16 TOPS@INT8 and 8 TOPS@FP16
Max Power	• 8W
Process	• 12nm FFC

关键特性

Architecture	• HUAWEI Da Vinci
Computing Engine	• 3D Cube
Performance	• 320 TFLOPS @FP16 and 640 TOPS @INT8
Max Power	• 310W
Process	• N7+

昇腾芯片实物图

数据来源：海思官网，《昇腾计算产业发展白皮书》，新华网，集成电路IC，西南证券整理

2.1 国产芯片补缺

- **与主流海外训练芯片对比：**

- 昇腾910在INT8、FP16等算力精度下的性能表现超过V100，可对标A100，暂不支持FP32单精度及FP64双精度算力格式。
- 在互联方面，昇腾处理器集成了HCCS、PCIE 4.0、RoCE v2接口，不如英伟达但应与AMD等相当，华为7月最新发布了万卡集群测试，6月支持单机8000卡商用，计划年底或者明年初超过16000张卡。当前，1750亿的大模型，半天就能训练完成。

公司	华为	华为	英伟达	英伟达	英伟达	AMD	Google
GPU	昇腾910	昇腾310	V100	A100	H100	MI100	TPU V4
工艺	7nm	12nm	12nm	7nm	5nm	7nm	7nm
算力	320 TFLOPS@FP16 640 TOPS@INT8	8 TFLOPS@FP16 16 TOPS@INT8	7 TFLOPS@FP64 14 TFLOPS@FP32 62 TOPS @INT8	9.7 TFLOPS@FP64 19.5 TFLOPS@FP32 312 TFLOPS@FP16 624 TOPS @INT8	26 TFLOPS@FP64 51 TFLOPS@FP32 756.5 TFLOPS@FP16 1513 TOPS @INT8	11.5 TFLOPS@FP64 23.1 TFLOPS@FP32 184.6 TFLOPS@FP16 184.6 TOPS @INT8	275 TFLOPS@BF16 275 TOPS@INT8
功耗	310W TDP	8W TDP	250W TDP	300W TDP	300-350W TDP	300W TDP	300W TDP
互联带宽	-	-	NV Link 300GB/S PCIE 32GB/S	NV Link 600GB/S PCIE 4.0 64GB/S	NV Link 600GB/S PCIE 5.0 128GB/S	PCIE 4.0 64GB/S	300GB/S

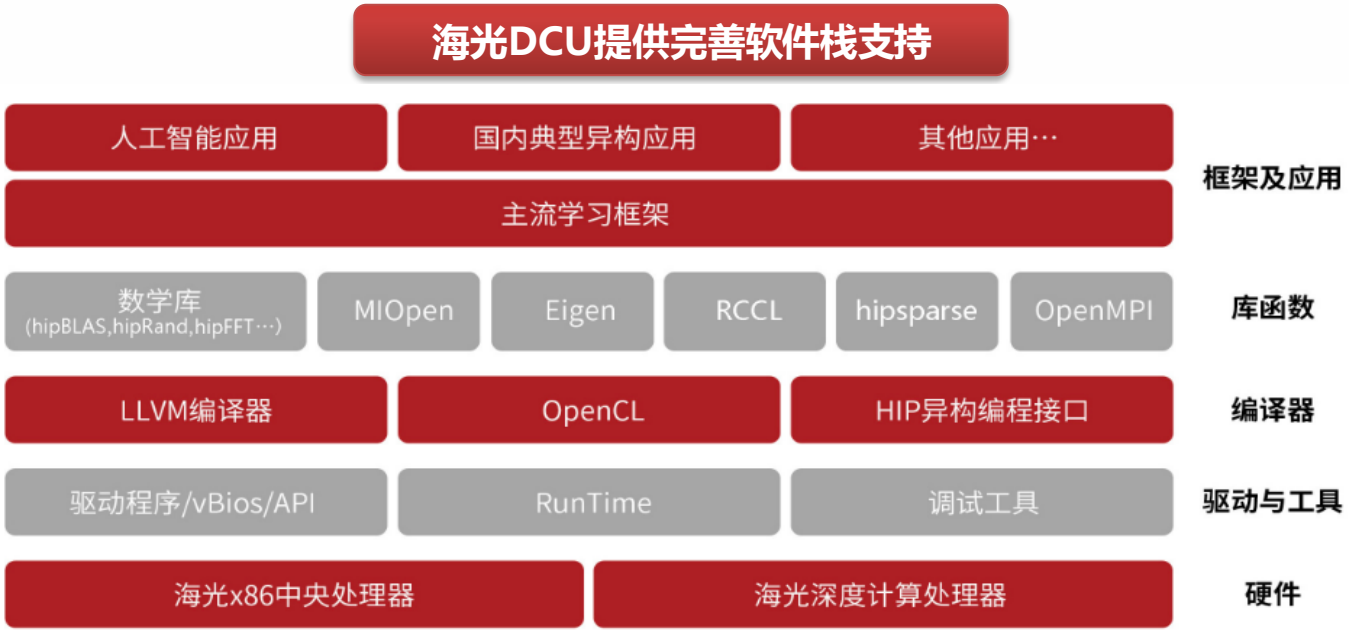
数据来源：海思官网，英伟达官网，AMD官网，谷歌官网，西南证券整理

2.1 国产芯片补缺

数据来源：昇腾社区官网，西南证券整理

2.1 国产芯片补缺

- **海光DCU在落地、生态、造血等方面具备多重优势，深算二号性能大幅提升有望脱颖而出。**
  - 海光DCU是国内唯一同时支持全精度和半精度训练的加速计算芯片。
  - 公司深算二号在2023年9月发布，性能比起上一代产品翻倍提升，并且兼容“类CUDA”环境，软硬件生态丰富，当前已在各地智算中心落地应用，对部分英伟达产品实现了良好替代，有望深度受益于下游AI算力的需求爆发。
  - 深度参与“东数西算”先进计算中心、“新基建”智算中心的建设，与大股东中科曙光深入协同，打造国产算力底座。
- **关注海光DCU新款芯片迭代进度，训练性能有望明显提升。**



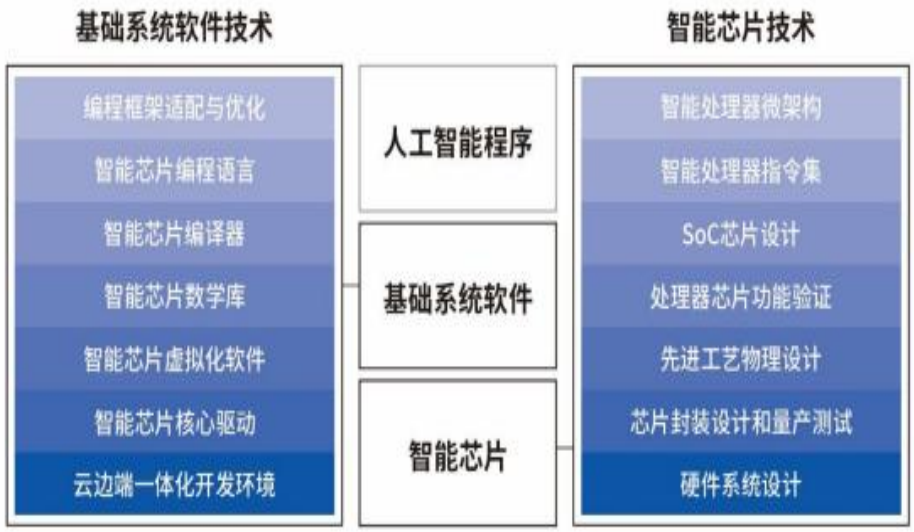
数据来源：海光信息，西南证券整理

2.1 国产芯片补缺

- **寒武纪云边端业务线协同发力，竞争力不断加强。**
  - 云端产品：当前寒武纪已推出三代云端智能加速卡，包括训练芯片思元290和推理芯片思元100、思元270、思元370；其中思元370在数家头部互联网企业完成视觉、语音、图文识别、自然语言处理等场景下的适配工作，已于2022年开始批量出货实现收入突破，并进一步在金融、运营商等行业实现落地。
  - 边缘产品：思元220边缘加速卡，自发布以来累计销量突破百万片。
  - 终端产品：1A、1H、1M系列智能处理器IP已集成于超过1亿台智能手机及其他智能终端设备。
- **公司新一代智能处理器微架构和指令集正在研发中，对大模型训练推理等场景进行重点优化。**

寒武纪思元370核心优势

寒武纪核心技术框架结构

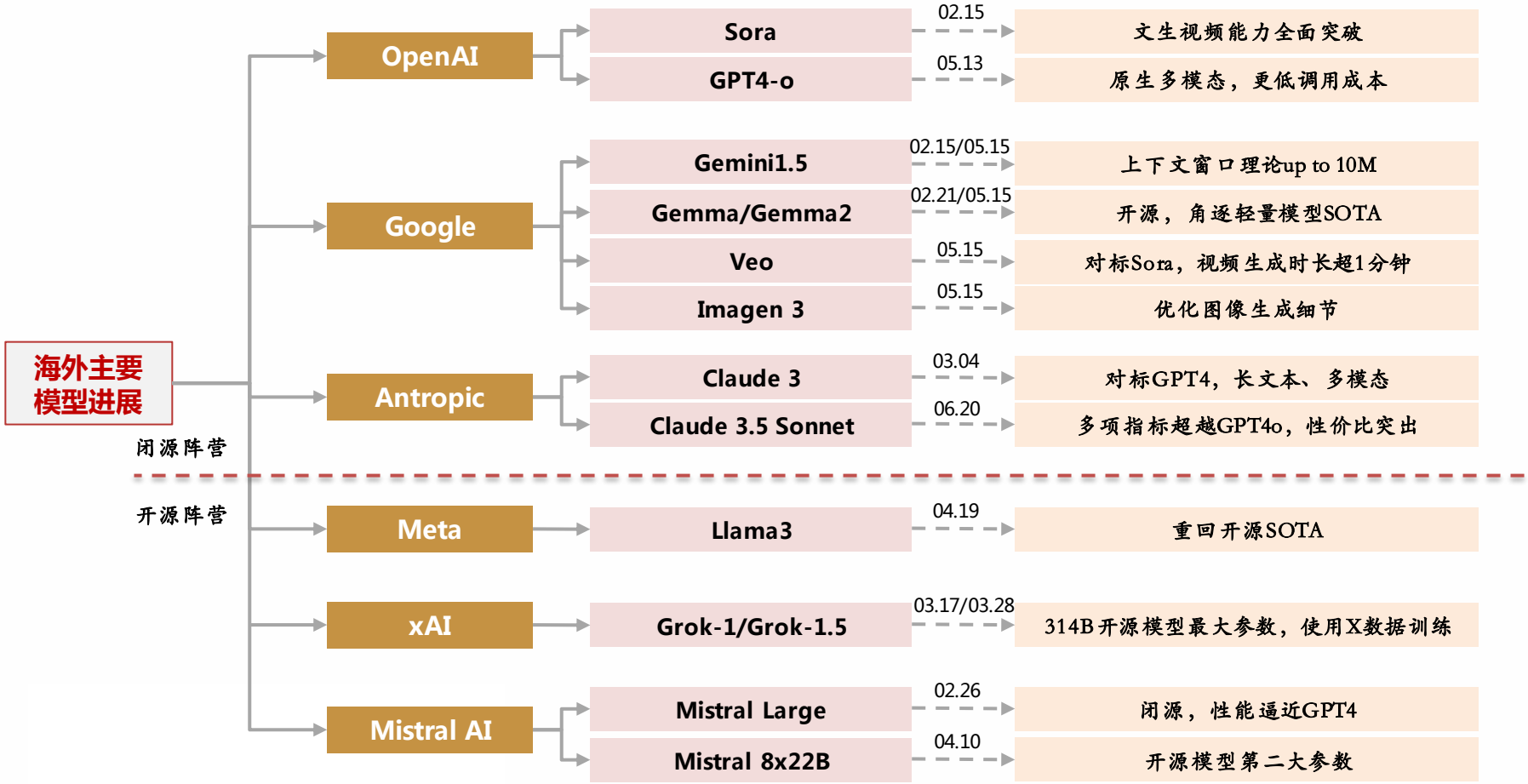


数据来源：寒武纪，西南证券整理

2.2 模型能力上限不断突破

➢ 2024年以来，海外OpenAI、Google、Antropic等厂商引领，模型能力上限取得进一步突破。

2024年以来海外大模型厂商主要进展



数据来源：OpenAI，谷歌，Antropic，MistralAI，Meta，xAI等，西南证券整理

2.2 国内模型能力亦不断追赶

➢ 2024年以来，阿里、百度、月之暗面等厂商百花齐放，奋起直追，国内大模型性能已经逼近国际先进水平。

数据来源：OpenAI，谷歌，Antropic，MistralAI，Meta，xAI等，西南证券整理

2.2 国产模型发展加快，优化方向逐步清晰

- OpenAI宣布自7月9日起停止非支持国家和地区的API服务，有望加速国内模型发展。
- 根据UC Berkeley发布的Chatbot Arena大模型榜单，OpenAI、Antropic、Google系列模型的综合能力依然名列前茅（该榜单由用户盲测客观评选，被Sam Altman多次引用，相对更具权威性和客观性），但零一万物的Yi-Large-Preview已经进入前十名。
- **开源社区上，国内大模型已取得长足突破。**根据Hugging Face6月末最新的开源大模型表排行榜，Qwen2-72B-Instruct以43.02的综合评分位于榜单第一，大幅领先排名第二的Llama-70B的35.13分。并且Qwen系列的另两个变体也占据了第三和第十的位置，外加排名第七的零一万物Yi-1.5-34B，中国的开源模型在整体上占主导地位。
- 在现有架构体系下，当前海内外大模型纷纷向**更强的多模态能力，更长的上下文，更低的训推成本**三大方向进行优化，以打开应用使能的想象空间。

LMSYS Chatbot Arena大模型排行榜

Rank* (UB)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1287	+4/-3	55826	OpenAI	Proprietary	2023/10
2	Claude 3.5 Sonnet	1272	+4/-4	23748	Anthropic	Proprietary	2024/4
2	Gemini Advanced 0514	1267	+4/-4	42364	Google	Proprietary	Online
3	Gemini 1.5 Pro API 0514	1263	+3/-4	48795	Google	Proprietary	2023/11
4	Gemini 1.5 Pro API 0409 Preview	1258	+3/-3	55576	Google	Proprietary	2023/11
4	GPT-4 Turbo 2024-04-09	1257	+3/-3	72076	OpenAI	Proprietary	2023/12
7	GPT-4-1106-preview	1251	+3/-3	86254	OpenAI	Proprietary	2023/4
7	Claude 3 Opus	1248	+3/-2	142537	Anthropic	Proprietary	2023/8
7	GPT-4-0125-preview	1246	+3/-3	79527	OpenAI	Proprietary	2023/12
9	Yi-Large-preview	1240	+4/-4	47212	01 AI	Proprietary	Unknown
11	Gemini 1.5 Flash API 0514	1228	+4/-4	42766	Google	Proprietary	2023/11
12	Gemma 2 27B it	1217	+5/-6	11407	Google	Gemma license	2024/6
12	Yi-Large	1217	+6/-5	11968	01 AI	Proprietary	Unknown

Hugging face开源模型排行榜

T	Model	Average	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU-PRO
	Qwen/Qwen2-72B-Instruct	42.49	79.89	57.48	35.12	16.33	17.17	48.92
	meta-llama/Meta-Llama-3-70B-Instruct	36.18	80.99	50.19	23.34	4.92	10.92	46.74
	Qwen/Qwen2-72B	35.13	38.24	51.86	29.15	19.24	19.73	52.56
	mistralai/Mixtral-8x22B-Instruct-v0.1	33.89	71.84	44.11	18.73	16.44	13.49	38.7
	HuggingFaceH4/zephyr-orpo-141b-A35b-v0.1	33.77	65.11	47.5	18.35	17.11	14.72	39.85
	microsoft/Phi-3-medium-4k-instruct	32.67	64.23	49.38	16.99	11.52	13.05	40.84
	01-ai/Yi-1.5-34B-Chat	32.63	60.67	44.26	23.34	15.32	13.06	39.12
	CohereForAI/c4ai-command-r-plus	30.86	76.64	39.92	7.55	7.38	20.42	33.24
	internlm/internlm2_5-7b-chat	30.46	61.4	57.67	8.31	10.63	14.35	30.42
	Qwen/Qwen1.5-110B	29.56	34.22	44.28	23.04	13.65	13.71	48.45
	abacusai/Smaug-72B-v0.1	29.56	51.7	42.42	17.75	9.62	15.39	40.46

数据来源：Hugging Face，LMSYS Chatbot Arena，西南证券整理

2.2 大模型优化方向——原生多模态

- 传统的多模态基础模型，通常为每种模态采用特定的「编码器」或「解码器」，将不同的模态分离开，跨模态信息融合能力受到限制。
- Google率先开启原生多模态的探索，区别于传统的多模态“后融合”的训练方式，Gemini在设计时原生支持多模态，从一开始便同时对多模态的数据同时进行预训练，对于文字、图像、视频、音频、代码的理解推理效果进一步提升。
- OpenAI在GPT4o上弃用了原来拼接Whisper、DALLE-E、GPT等多个模型的方式，采用了端到端多模态架构，在视觉、音频理解方面大幅升级，并实现了极致流畅的用户体验。
- Meta在5月份发布了Chameleon变色龙多模态大模型，同样采用“前融合”（early-fusion）方法，从一开始就将所有模态投影到共享的表示空间中。

传统多模态架构示意

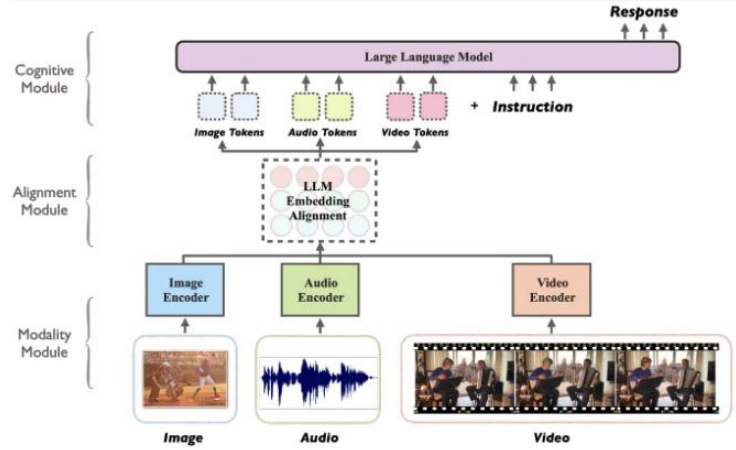
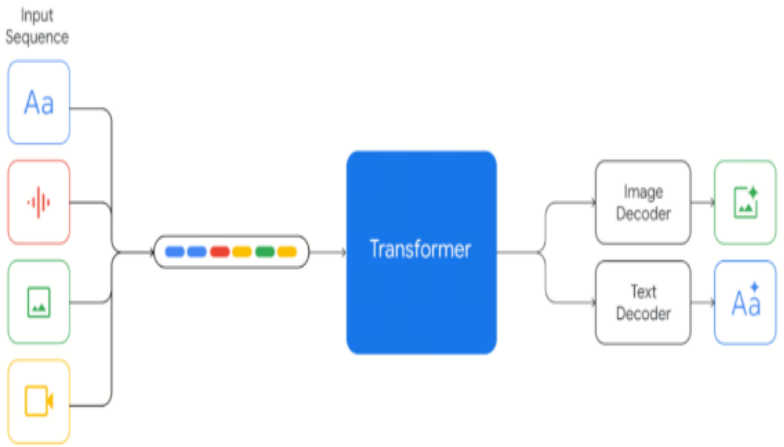


Figure 1: An overview of MACAW-LLM model architecture.

Gemini基于多模态原生思路设计

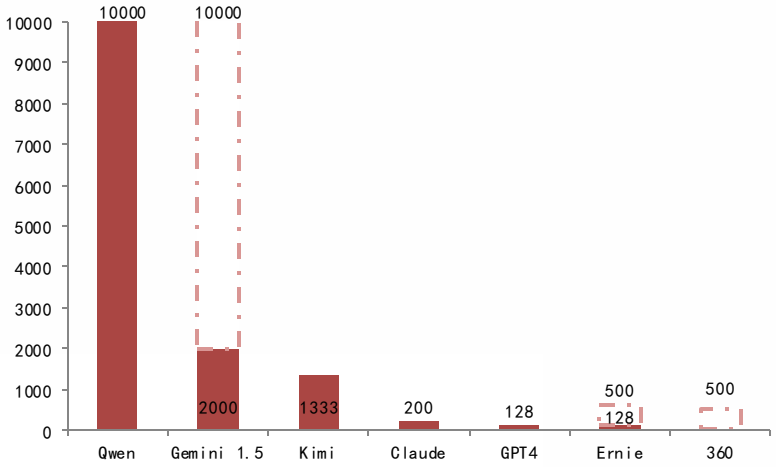


数据来源：腾讯《MACAW-LLM: MULTI-MODAL LANGUAGE MODELING WITH IMAGE, AUDIO, VIDEO, AND TEXT INTEGRATION》DeepMind，西南证券整理

2.2 大模型优化方向——长文本

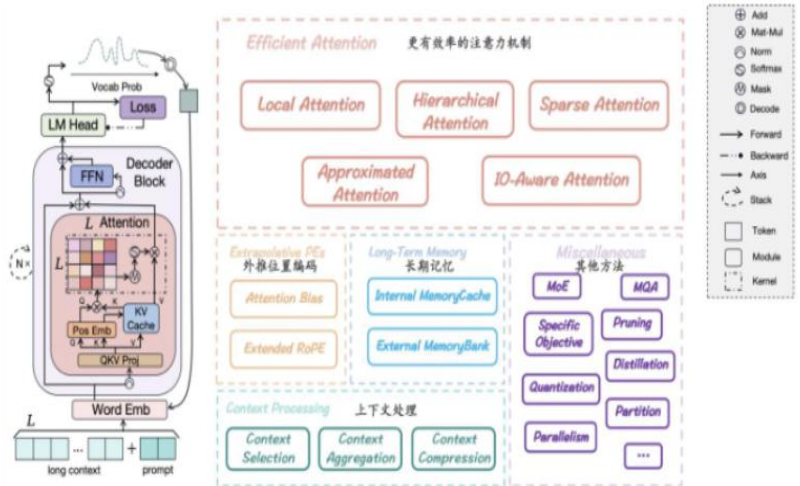
- 2023年年末，各家大模型主流上下文窗口还在32Ktokens量级，仅有GPT4 Turbo达到128K、Claude3达到200K量级。
- 而截至2024年H1，主流模型基本均支持128K量级，头部厂商卷入1M以上量级。2024年2月，谷歌Gemini1.5将上下文长度扩展至1M量级，并宣称最高理论可达到10M（尚未公开开放）；3月Kimi、阿里先后宣布支持200万文字（对应约1333Ktokens）、10Mtokens上下文窗口。
- 当前实现长上下文窗口主要有“内生”和“外挂”两种优化方向。“内生”主要通过改进Transformer架构中的各个模块：1）注意力机制优化，如Sparse Attention等；2）长期记忆力机制；3）外推位置编码，如扩展RopE；4）上下文预/后处理：压缩、聚合等；5）其他：MoE、目标函数优化、权重压缩等。“外挂”RAG是当前算力和内存限制下比较“取巧”和简单的方法，在B端场景比较适合，有其存在的必要性。

各大模型上下文窗口（千tokens）



注：虚线表示尚未公开开放

长文本问题的解决方案

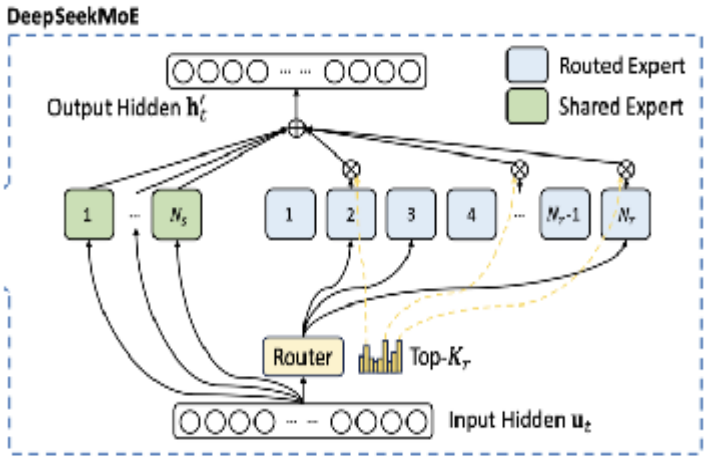


数据来源：《Advancing Transformer Architecture in Long-Context Large Language Models: A Comprehensive Survey》，西南证券整理

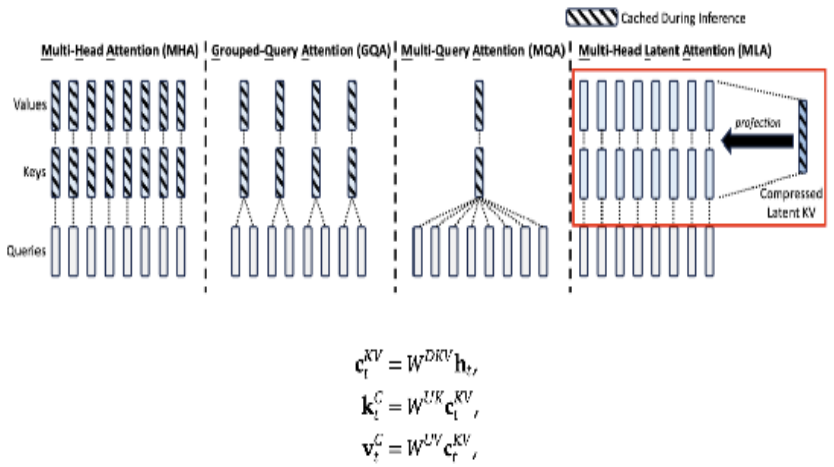
2.2 大模型优化方向——降本

- 5月6日，DeepSeek发布最新模型DeepSeek-V2，百万tokens仅需1元，拉开国内大模型降价序幕；随后，智谱、百度、字节、腾讯、讯飞等纷纷加入降价浪潮。
- 大模型降价背后是训练&推理成本的下降，是模型工程进步的必然趋势，此前OpenAI就伴随新模型的推出进行过多次调价。**1）头部模型厂商纷纷从稠密架构切换至MoE架构：**2023年4月GPT4发布，外界猜测其使用16个专家的MoE架构参数达到1.8T；2023年12月，Mistral开源8×7B的MoE模型，在多项测评逼近或超过1750亿参数的GPT3.5水平，再次引发业界关注；2024年以来，Google Gemini1.5pro开启国内外模型稠密到稀疏MoE的切换浪潮，幻方、xAI、阿里、Minimax、商汤、阶跃星辰等纷纷在新一代模型里使用并改进MoE。**2）围绕KV Cache压缩进行改进：**主要方向包括优化注意力机制、减少模型层数、输入的token数等；当前主要聚焦注意力机制的优化，例如DeepSeek引入MLA，在大幅降低训推成本的同时，保持较强性能。

DeepSeek进一步引入细粒度专家和共享专家



MLA与其他注意力机制的比较



数据来源：《DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model》，西南证券整理

2.2 大模型优化方向——降本

- 自有算力基础设施的互联网厂商降价幅度更大，百川、Minimax、月之暗面等初创厂商尚未跟进；降价幅度最大、甚至免费的，一般是轻量级模型。
- 降价有望加速**“吸引开发者-应用落地-数据反馈-模型能力迭代”**的生态飞轮。

厂商	大模型版本	降价前（元/百万tokens）		降价后（元/百万tokens）		备注
		输入	输出	输入	输出	
OpenAI	GPT4o			36	109	
	GPT4-turbo			73	218	
	GPT4			218	436	
深度求索	DeepSeek-V2			1	2	定价即低价
智谱	GLM4/GLM4V			100	100	
	GLM-3 Turbo	5	5	1	1	降价80%
字节	Doubao-pro-32K			0.8	2	定价即低价
	Doubao-pro-128K			5	9	
	Doubao-lite-4K/32K			0.3	0.6	
	Doubao-lite-128K			0.8	1	
阿里	Qwen-Max	120	120	40	120	输入降价67%，输出不变
	Qwen-Long	20	20	0.5	2	输入降价97%，输出降价90%
	Qwen-Plus	20	230	4	16	输入降价80%，输出降价40%
	Qwen-Turbo	8	8	2	6	输入降价75%，输出降价25%
百度	ERNIE Speed-8K/128K	4	8	免费	免费	免费轻量级版本 旗舰版价格不变
	ERNIE Lite 8K/128K	3	6	免费	免费	
	ERNIE 4.0	120	120	-	-	
	ERNIE 3.5	48	96	-	-	
腾讯	混元-Lite	8	8	免费	免费	免费轻量级版本
	混元-Standard	10	10	4.5	5	输入降价55%，输出降价50%
	混元-Standard-256K	120	120	15	60	输入降价87.5%，输出降价50%
	混元-pro	100	100	30	100	输入降价70%，输出不变
讯飞	Spark Lite	18	18	免费	免费	免费轻量级版本 旗舰版价格不变
	Spark pro	21-30	21-30	-	-	
	Spark 3.5 Max	21-30	21-30	-	-	

数据来源：各公司官网，西南证券整理

2.3 大模型由云到端协同演进，倒逼硬件终端革新

- 大参数模型受到算力资源有限、高质量数据集有限、部署成本过高等限制，实际应用中涉及资源如何最优化配置的问题。伴随模型蒸馏等压缩技术方法的成熟，**大小模型开始实现由云到端的协同演进。**
  - **谷歌为了发挥旗下Android生态优势而始终志在云边端结合，推动轻量化模型升级。**谷歌于2023年5月发布PaLM 2大模型，率先推出四种不同大小的模型；后续发布的Gemini系列也同样延续了PaLM的策略，分为Ultra、Pro、Nano等多个型号，其中最小的Nano提供1.8B和3.25B两个版本，并且成功在Pixel 8 Pro和三星Galaxy S24手机上实现部署。此外，Google还于2024年2月开源了轻量级模型Gemma，并在5月更新至Gemma2，在小参数的情况下实现大幅性能提升。
  - **OpenAI当前仍延续大参数路线，但微软积极布局轻量级模型。**当前头部厂商中，OpenAI几乎是唯一只做大参数模型的厂商；而微软在23年11月Ignite大会提出SLM（Small Language Models）策略，表示SLM是LLM的重要补充，可以为AI应用提供另一类的选择，并陆续推出Phi-2、Phi-3系列模型。其中Phi-3最小提供3.8B版本，在多项测试集上评分领先LLaMA-8B。

Gemma2-9B性能大幅超过LLaMA 3-8B

Benchmark	metric	Gemma-1 2.5B	Gemma-2 2.6B	Mistral 7B	LLaMA-3 8B	Gemma-1 7B	Gemma-2 9B	Gemma-2 27B
MMLU	5-shot	42.3	**51.3**	62.5	66.6	64.4	**71.3**	75.2
ARC-C	25-shot	48.5	**55.4**	60.5	59.2	61.1	**68.4**	71.4
GSM8K	5-shot	15.1	**23.9**	39.6	45.7	51.8	**68.6**	74.0
AGIEval	3-5-shot	24.2	**30.6**	44.0†	45.9†	44.9†	**52.8**	55.1
DROP	3-shot, F1	48.5	**52.0**	63.8*	58.4	56.3	**69.4**	74.2
BBH	3-shot, CoT	35.2	**41.9**	56.0°	61.1°	59.0°	**68.2**	74.9
Winogrande	5-shot	66.8	**70.9**	78.5	76.1	79.0	**80.6**	83.7
HellaSwag	10-shot	71.7	**73.0**	**83.0**	82.0	82.3	81.9	86.4
MATH	4-shot	11.8	**15.0**	12.7	-	24.3	**36.6**	42.3
ARC-e	0-shot	73.2	**80.1**	80.5	-	81.5	**88.0**	88.6
PIQA	0-shot	77.3	**77.8**	**82.2**	-	81.2	81.7	83.2
SIQA	0-shot	49.7	**51.9**	47.0*	-	51.8	**53.4**	53.7
Boolq	0-shot	69.4	**72.5**	83.2*	-	83.2	**84.2**	84.8
TriviaQA	5-shot	53.2	**59.4**	62.5	-	63.4	**76.6**	83.7
NQ	5-shot	12.5	**16.7**	23.2	-	23.0	**29.2**	34.5
HumanEval	pass@1	**22.0**	17.7	26.2	-	32.3	**40.2**	51.8
MBPP	3-shot	29.2	**29.6**	40.2*	-	44.4	**52.4**	62.6
Average (8)		44.0	**49.9**	61.0	61.9	62.4	**70.2**	74.4
Average (all)		44.2	**48.2**	55.6	-	57.9	**64.9**	69.4

Phi-3-3.8B

	Phi-3-mini 3.8b	Phi-3-small 7b	Phi-3-medium 14b	Phi-2 2.7b	Mistral 7b	Gemma 7b	Llama-3-In 8b	Mixtral 8x7b	GPT-3.5 version 1106
MMLU (5-Shot) [HBK+21]	68.8	75.7	78.0	56.3	61.7	63.6	66.5	70.5	71.4
HellaSwag (5-Shot) [ZHB+19]	76.7	77.0	82.4	53.6	58.5	49.8	71.1	70.4	78.8
ANLI (7-Shot) [NWD+20]	52.8	58.1	55.8	42.5	47.1	48.7	57.3	55.2	58.1
GSM-8K (8-Shot; CoT) [CKB+21]	82.5	89.6	91.0	61.1	46.4	59.8	77.4	64.7	78.1
MedQA (2-Shot) [JPO+20]	53.8	65.4	69.9	40.9	50.0	49.6	60.5	62.2	63.4
AGIEval (0-Shot) [ZCG+23]	37.5	45.1	50.2	29.8	35.1	42.1	42.0	45.2	48.4
TriviaQA (5-Shot) [JCWZ17]	64.0	58.1	73.9	45.2	75.2	72.3	67.7	82.2	85.8
Arc-C (10-Shot) [CCE+18]	84.9	90.7	91.6	75.9	78.6	78.3	82.8	87.3	87.4
Arc-E (10-Shot) [CCE+18]	94.6	97.0	97.7	88.5	90.6	91.4	93.4	95.6	96.3
PIQA (5-Shot) [BZGC19]	84.2	86.9	87.9	60.2	77.7	78.1	75.7	86.0	86.6
SociQA (5-Shot) [BZGC19]	76.6	79.2	80.2	68.3	74.6	65.5	73.9	75.9	68.3

数据来源：Google，《Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone》，西南证券整理

2.3 大模型由云到端协同演进，倒逼硬件终端革新

- 2023年7月Llama开源7B、13B、70B三种模型，轻量级模型的更新明显加快。当前轻量模型受制于参数量问题，仍有通用性、鲁棒性等局限，例如Phi-3在factual knowledge方面表现有严重短板，3.8B的版本还舍弃了多语言能力等。但多种优化策略下，当前小模型的性能已有质变。
- 通过提升模型训练数据量和数据集质量（Data-Optimal）、使用计算最优缩放（Compute-optimal）、更新模型架构等方法，**可以使性能逼近甚至超越跨数量级的大模型。**
- Google强调“Compute-Optimal”策略，其在PaLM2技术报告中提出，**数据集和模型大小应该大约以1：1的比例同时缩放，以达到最佳性能**；Gemma应该延续了此项理论，7B和2B模型分别对应6T和2T tokens的训练数据，其7B模型MMLU评分达到64.3，高于LLaMA-2-13B的54.8。
- 微软更强调“Data-Optimal”的策略，微软在Phi-3技术报告中提出，**数据质量是影响模型性能的首要因素**，通过数据过滤、分阶段训练、合成数据等方式，Phi-3-3.8B的MMLU评分达到68.8，超过了Llama-3-8B的66.5。

“最优数据”策略使得小模型在参数量受限条件下，性能逼近跨数量级大模型

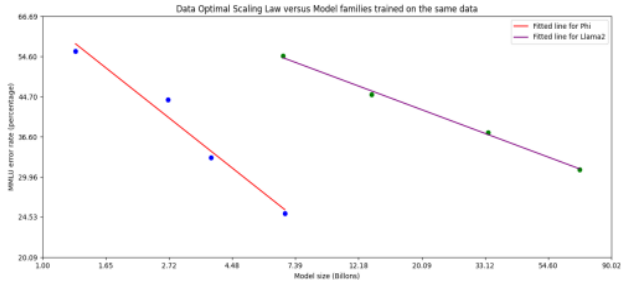


Figure 3: Scaling law close to the “Data Optimal Regime” (from left to right: phi-1.5, phi-2, phi-3-mini, phi-3-small) versus Llama-2 family of models (7B, 13B, 34B, 70B) that were trained on the same fixed data. We plot the log of MMLU error versus the log of model size.

数据来源：《Palm2 Technical Report》，商汤日日新5.0发布会，《Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone》，西南证券整理

2.3.1 AI重新定义PC处理器标准

- **AI定义硬件门槛，颠覆“安迪-比尔”定律**
  - 过去IT产业的软硬件升级换代遵循“安迪-比尔”定律（Andy gives, Bill takes away），即英特尔不断推出更高性能的CPU，而微软随之发布更占用资源的操作系统，双方交替迭代以催动下游终端消费者的更新需求。
  - 而为了适应端侧模型的计算需求，本次AIPC则是由微软来定义硬件门槛，根据Trendforce等报道，**微软计划在Windows12为Copilot+PC设定40TOPS算力和16GB内存DRAM最低要求**，或标志着IT产业进入AI定义硬件终端时代。
- 从当前已经正式发布的PC芯片看，**【提升算力】、【提高内存】、【降低功耗】成为明显趋势**，随之或带来**【指令集架构】、【异构计算】、【内存升级】等重要变化。**

各厂商芯片算力及功耗对比

	高通 X Elite	苹果M3	苹果M4	Intel Meteor Lake	Intel Lunar Lake	AMD Phoenix Point	AMD Hawk Point	AMD Strix Point
发布时间	**2023.10**	2023.10	2024.05	2023.12	**预计2024H2**	2023.04	2023.12	**预计2024H2**
NPU性能	**45 TOPS**	18TOPS	38TOPS	11TOPS	**48TOPS**	10TOPS	16TOPS	**~50TOPS**
综合性能	**75TOPS**	-	-	34TOPS	**120TOPS**	33TOPS	39TOPS	**~80TOPS**
是否满足微软定义	√	-	-	×	√	×	×	√

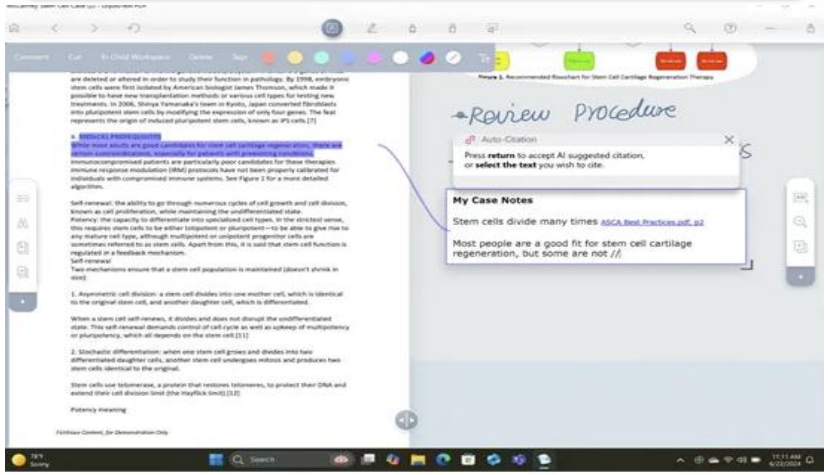
注：苹果M3或为FP16精度下数据

数据来源：微软，英特尔，高通，AMD，苹果，西南证券整理

2.3.1 Copilot+PC跨越式升级，端侧AI智能体形态初现

- **微软在2024年Build大会上推出全新的Copilot+PC，具备以下基础功能**
- 新增Copilot按键，一键呼出Copilot助手，可以访问OpenAI最新的大模型GPT-4o，可内置微软Phi-3等端侧模型，集成RAG技术，提供文档总结、文案扩写、翻译、问答等功能。
- 受益于GPT4-o的加持，**Copilot完全具备上下文感知和视觉感知能力，能够帮助用户处理屏幕上正在进行的任何工作**，发布会上展示Copilot帮助用户进行游戏指导。
- 提供离线实时翻译功能，可在非联网状态将电脑的任何音频翻译为英语字幕，并在所有应用程序的屏幕上持续实时显示。
- 内置文生图功能，可以使用Image Cocreator免费生成自定义风格的图像并进行微调，同时和Adobe官方合作，大幅优化响应速度。

一键唤醒Copilot助手



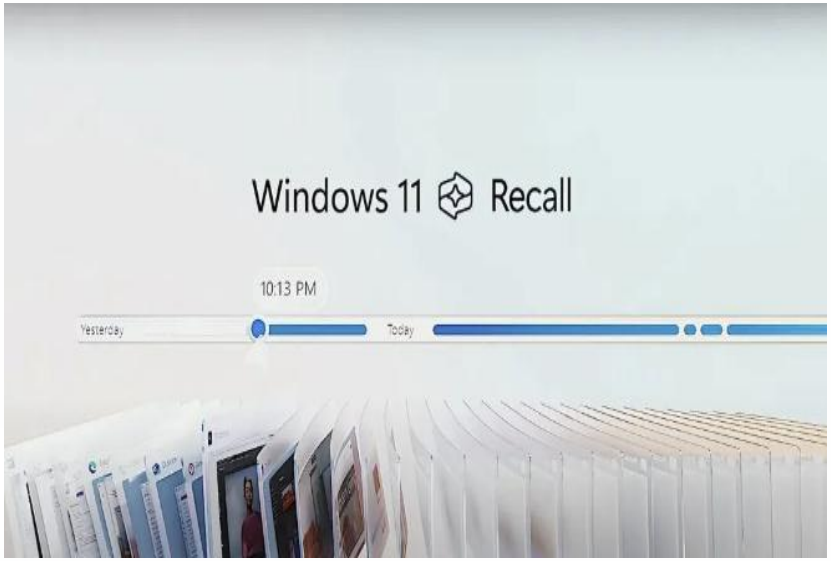
Copilot指导用户进行游戏

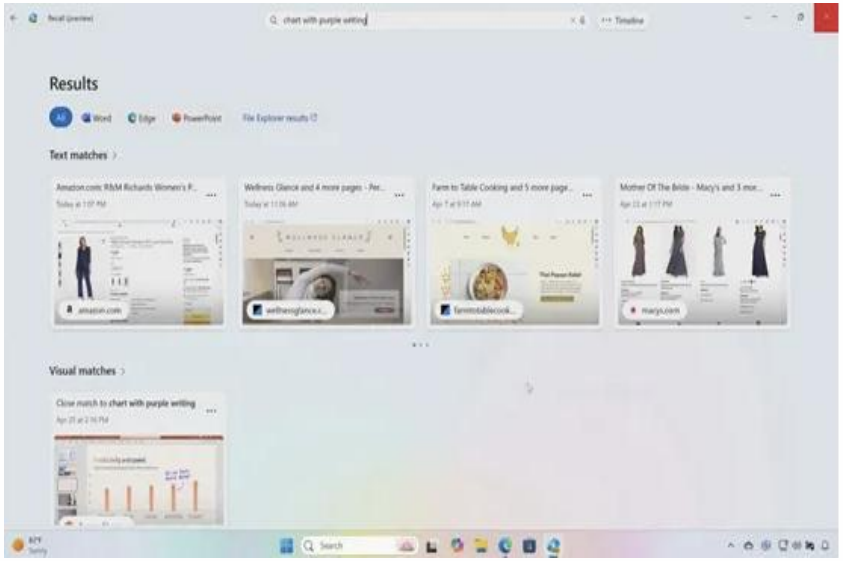
数据来源：微软，西南证券整理

2.3.1 Copilot+PC跨越式升级，端侧AI智能体形态初现

- **最大亮点新增Recall回想功能，对硬件性能提出更高要求**
- 微软创新发布“Recall”功能，可以帮助用户跨越时间线，找到在PC上的任何浏览记录或处理过的任务，包括网页、图片、聊天记录等。
- “Recall”通过每隔几秒钟截取一次活动窗口的屏幕截图来工作，默认情况下记录用户在Windows中的所有操作，最长可达三个月。
- 为满足“Recall”功能的存储需求，硬盘空间至少达到256GB；同时微软在之前也对芯片算力提出40TOPS+的定义，AI正在重塑PC产业链，对硬件性能提出更高要求。

Recall功能可以帮助用户找过任何浏览过的内容





数据来源：微软，西南证券整理

2.3.1 AI重新定义PC处理器标准，关注Arm架构份额提升

- **搭载骁龙X性能超过苹果，Windows on Arm抢占AIPC滩头**
- 微软2012年发布官方电脑Surface系列，2015年推出Windows on Arm，但仅在2021发布的Surface Pro X小众型号上搭载高通8cx Gen 2，其他主力机型全部为Wintel。
- 伴随Copilot+PC对性能、功耗提出更高要求，微软新一代Surface设备全系搭载高通Snapdragon X Elite/Plus处理器，CPU和GPU性能比配备M3处理器的MacBook Air快58%，NPU算力达到45TOPS高于M3芯片的18TOPS，续航时间同样领先。
- 除官方的Surface系列外，联想、宏碁、华硕、戴尔、惠普、三星等PC厂商已经与微软签订了合作协议，6月18日将正式发布20+不同型号的Copilot+ PC，均搭载高通Snapdragon X系列处理器，Windows on Arm在推出9年后借着AIPC的浪潮抢占滩头。

Copilot+PC与苹果M3 Macbook Air对比

	Copilot+PC	苹果M3 Macbook Air
CPU和GPU性能	Copilot+PC 比 M3 Macbook Air 快58%	
NPU性能	45 TOPS	18TOPS
Cinebench 23（单核）	1892	1745
Cinebench 23（多核）	14983	10352
电池容量	90Wh	70Wh
续航时间	16小时56分钟	15小时25分钟

第一批集成骁龙X系列芯片的Copilot+PC

数据来源：微软，西南证券整理

2.3.1 AI重新定义PC处理器标准，关注异构计算

- **AIPC引领桌面处理器从通用计算向异构计算发展，CPU+GPU+NPU成为标配**
- 异构计算（Heterogeneous Computing）主要指使用不同类型指令集和体系架构的计算单元组成系统进行联合计算，合理地将不同类型的计算分配到不同硬件上运算可以获得更优的计算性能和更低的功耗。
- NPU（Neural Process Unit）是专门设计用于加速神经网络计算的处理器，在硬件层面针对AI计算常用的矩阵乘法、卷积等进行优化。
- 为应对散热和能耗瓶颈，芯片厂商纷纷在原有CPU+GPU的基础上引入NPU。

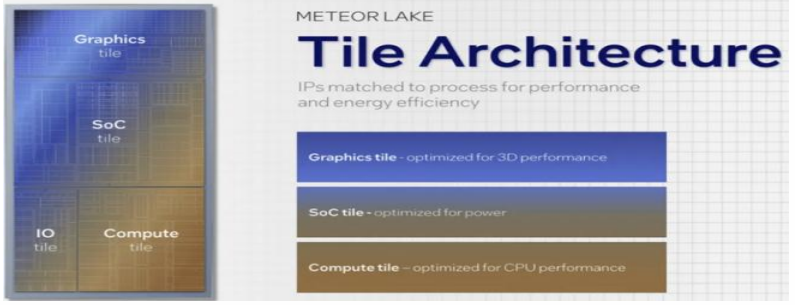
几类处理引擎比较

	CPU	GPU	NPU
定义	中央处理单元	图形处理单元	神经网络处理单元
适用领域	复杂任务，串行指令	逻辑简单、计算密度高的并行任务	神经网络加速，如矩阵乘法、卷积等
硬件抽象	CPU Serial Task	GPU Graphics/ Parallel Task	NPU Recognition/ Parallel Task

通用性

能效比

英特尔、高通处理器架构



数据来源：CSDN，高通，英特尔，西南证券整理

2.3.1 AI重新定义PC处理器标准，关注内存升级

- **新一代处理器或将只支持DDR5/LPDDR5，带动内存容量和速率升级**
- 根据Trendforce和Yole，2023年单台PC的DRAM平均容量大约在9-10GB左右；而根据微软定义，以及当前已经发布的AIPC 内存情况来看，16GB将成为AIPC内存的最低配置。
- 根据Intel，16GB的内存容量仅仅是起点，伴随更多AI应用本地运行，后续32GB以上的大容量内存将成为AIPC的入门级要求。
- 此外，自Snapdragon X Elite、AMD Ryzen 8000、Intel Core Ultra等芯片起，内存仅支持DDR5/LPDDR5，内存容量和速率升级已成确定性趋势。

当前发布AIPC的内存情况

	高通 X Elite	Intel Core Ultra（Meteor Lake）	AMD Ryzen 8000（Hawk point）
最大内存速度	LPDDR5/x-8533	DDR5-5600 LPDDR5/x-7467	DDR5 LPDDR5/x
最大内存容量	64GB	64GB（LP5） 96GB（DDR5）	64GB（LP5） 96GB（DDR5）
已发布机型	Lenovo、Surface等	联想、华硕、惠普等	Thinkpad等
最低内存	最低16GB	最低16GB	最低16GB

数据来源：微软，Yole，Trendforce，Omdia，英特尔，西南证券整理

2.3.1 生态过渡是关键

- **过去Win11对于高通支持仍有欠缺：**根据当前微软官网显示，使用Qualcomm Snapdragon处理器运行Win11，在网络连接、电池续航、开机速度等方面具备明显优势；而劣势基本上体现在应用生态方面：游戏、外设驱动、防病毒软件等支持仍然较为欠缺。
- **过去Arm原生软件较少：**上一代高通8cx Gen3笔记本推出时，原生应用仅有Office套件、Bing浏览器、Photoshop等；大部分常用软件如微信、Chrome、Steam等均需通过转译器，不仅带来性能损耗，且增加功耗；而小部分专业软件如Pr、Ae、CAD等存在完全不兼容的问题。
- **过去硬件厂商很少单独开模：**过去Arm平台的笔电基本上都是根据x86平台的型号修改而来，鲜有OEM厂商为其单独设计开模，难以发挥出Arm平台小封装和无需散热等优点。

微软官网关于高通处理器运行Win11的优劣势说明

运行基于 Qualcomm Snapdragon 处理器的 Windows 11 电脑有哪些优势?

你可以使用配备 Qualcomm Snapdragon 处理器的电脑随时随地进行工作。 你的电脑将:

- **始终连接到 Internet。** 有了手机数据连接，获得手机网络信号就可以上网 - 就像使用手机上网一样。 当你在办公室中、在家中或在你信任的另一个 WLAN 网络附近时，可以连接到 WLAN 以节省手机网络数据流量并继续工作。
- **摆脱电源插座的束缚，电池使用时间长达一整天。** 耗电量比其他电脑少，即使持续使用一个日常工作日或上课日，它的电源也不会耗尽，也不必为寻找插座接通电源而担心。 如果你使用电脑是为了娱乐和休闲，那么在电脑需要进行下一次充电之前，你可以享受长达 20 个小时的本地视频播放时间。
- **开机迅速。** 不使用电脑时，只需像在手机上一样按下电源按钮即可关闭屏幕。 当你取出电脑并重新打开时，它会立即打开。 你可以利用课间、会议期间或其他活动中的琐碎时间来完成你想做的工作，不必浪费时间等待电脑启动。

运行基于 Qualcomm Snapdragon 处理器的 Windows 11 电脑时应注意哪些限制?

无论是否运行处于 S 模式的 Windows 11，运行基于 Qualcomm Snapdragon 处理器的电脑时都存在一些限制:

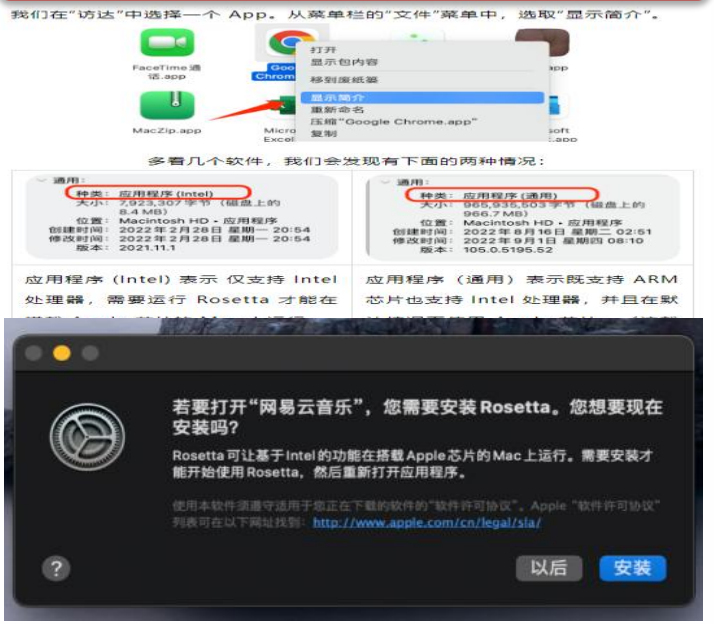
- **硬件、游戏和应用的驱动程序仅适用于基于 Snapdragon 处理器运行的 Windows 11 电脑。** 有关详细信息，请咨询硬件制造商或驱动程序的开发组织。 驱动程序是与硬件设备通信的软件程序 - 通常用于防病毒和反恶意软件、打印或 PDF 软件、辅助技术、CD 和 DVD 实用程序，以及虚拟化软件。 如果某个驱动程序不能运行，依赖该驱动程序的应用或硬件也不会运行（至少不能全功能运行）。 无论是否处于 S 模式，仅当 Windows 中内置了依赖的驱动程序，外设和设备才能运行。
- **部分游戏无法运行。** 使用 1.1 以上 OpenGL 版本的游戏和应用或依赖于"反作弊"驱动程序的游戏和应用无法运行。 可以咨询你的游戏发行商，查看游戏是否可以正常运行。
- **自定义 Windows 体验的应用可能会出现问题。** 包括一些输入法编辑器 (IME)、辅助技术和云存储应用。 开发应用的组织决定应用是否支持基于 Snapdragon 处理器运行的 Windows 11 电脑。
- **不能安装第三方防病毒软件。** 不能在基于 Snapdragon 处理器运行的任意版本的 Windows 11 上安装第三方防病毒软件。 但是，在 Windows 11 设备受支持的生命周期内，Windows Defender 安全中心可以帮助保护电脑的安全。
- **S 模式下不支持客户端 Hyper-V。** 从 Windows 11 开始，禁用 S 模式时，Qualcomm Snapdragon 处理器支持此功能。
- **Windows 传真和扫描不可用。** 此功能不适用于基于 Snapdragon 处理器运行的任何 Windows 11 版本。

数据来源：微软官网，西南证券整理

2.3.1 生态过渡是关键

- **迭代转译软件实现生态过渡，后续期待更多Win On Arm原生软件支持**
- 参照苹果发布M1后的依赖Rosetta 2进行x86至Arm的过渡，针对x86应用在Arm芯片上的模拟/转译，微软也做了新一轮升级，**联合高通打造全新的转译层Prism**，转译后的应用将在Arm硬件上无缝运行，相比过去的转译技术运行速度将提高10%到20%。
- 在2023年5月举办的微软Build开发者大会上，Dropbox、Whatsapp、GoodNotes等一系列应用纷纷宣布了对Windows On Arm的原生支持。而过去的这一年之间，Chrome、Opera、Spotify、Zoom、Blender等应用也纷纷添加了对Arm的原生支持。伴随Copilot+PC的发布，期待更多应用软件厂商宣布将加入Windows On Arm的原生应用开发。

苹果依靠Rosetta2顺利实现生态过渡



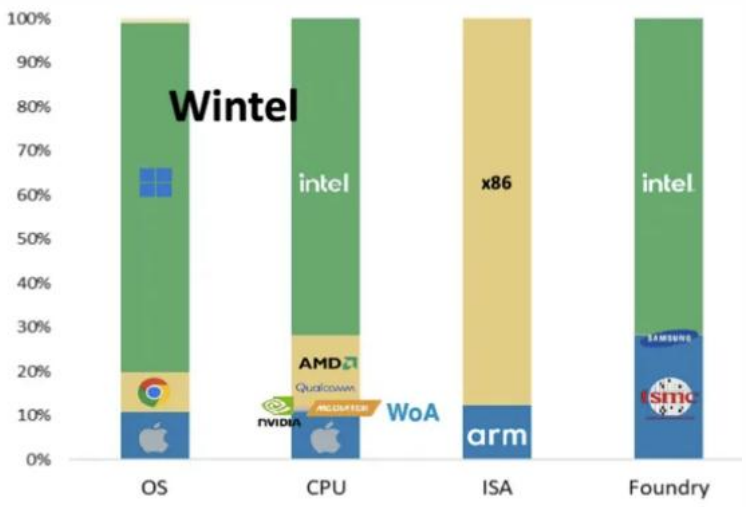
更多应用即将原生支持WoA

数据来源：IT之家，微软，西南证券整理

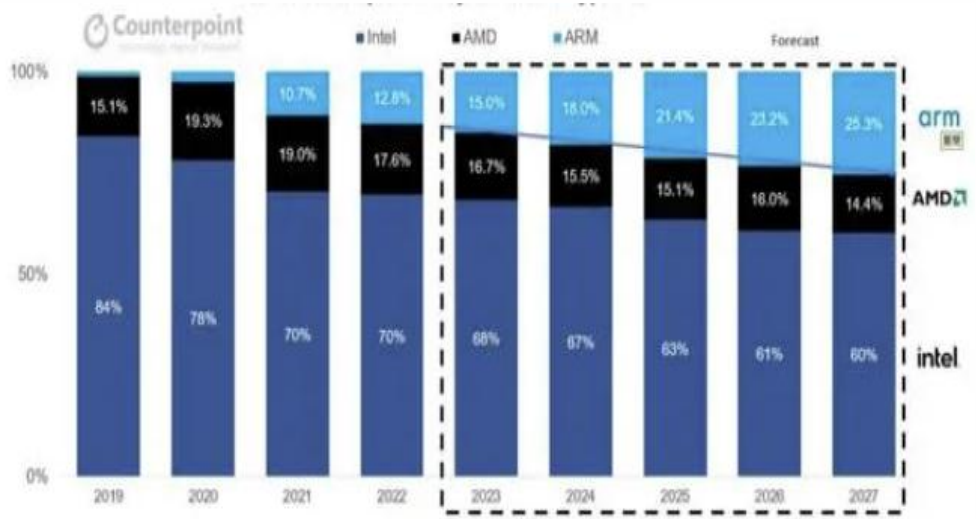
2.3.1 重视ARM份额提升

- 根据IDC，2023年PC市场Arm架构份额大约10%，且以苹果为主；
- 而伴随Copilo+PC对于处理器能效的要求提升，高通代表的ARM阵营已经取得先机；同时，高通与微软的排他性协议将于2025年到期，后续Nvdia、Mediatek等厂商有望加入WoA壮大生态，叠加微软对于软件应用开发生态的扶持力度加强，Arm阵营有望在此轮AIPC浪潮中取得明显份额提升；根据Counterpoint，2027年Arm PC市占率有望超过25%。
- 我们认为，若以苹果作为参考，10%市场份额可以作为生态突破的临界点，足够吸引庞大的开发者生态。**建议重点关注Copilot+PC面世后高通、英特尔、AMD处理器的实际性能和续航表现，以及ARM原生应用的上线情况。**

2023年ARM PC份额仅有10%



2027年Arm PC市占率有望达到25%

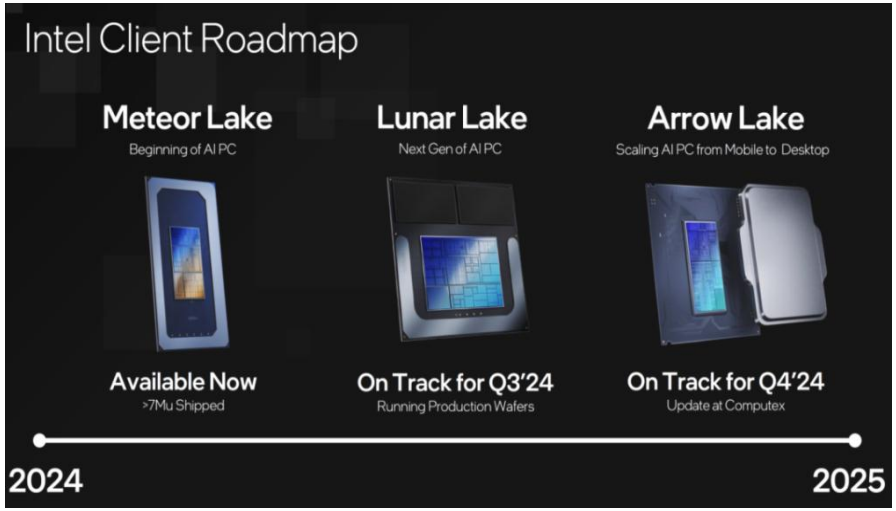


数据来源：IDC，CounterPoint，西南证券整理

2.3.1 x86阵营进行系列架构更新

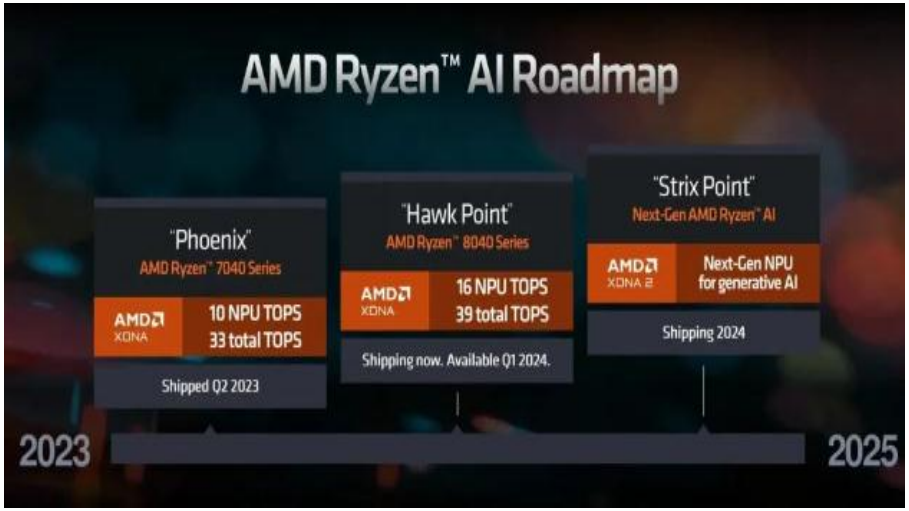
- **面对Arm阵营发起的冲击，Intel、AMD针对功耗和AI算力进行迭代优化**

Intel发布AIPC芯片路线图



- Meteor Lake架构于2023年发布，采用分离式模块架构，首次将NPU集成到PC处理器中，开启AIPC时代的首款尝试
- Lunar Lake架构将于2024年Q3发布，大幅提升性能的同时降低功耗
- Arrow Lake架构将于2024年Q4发布，将AI芯片拓展至桌面台式终端

AMD发布AIPC芯片路线图



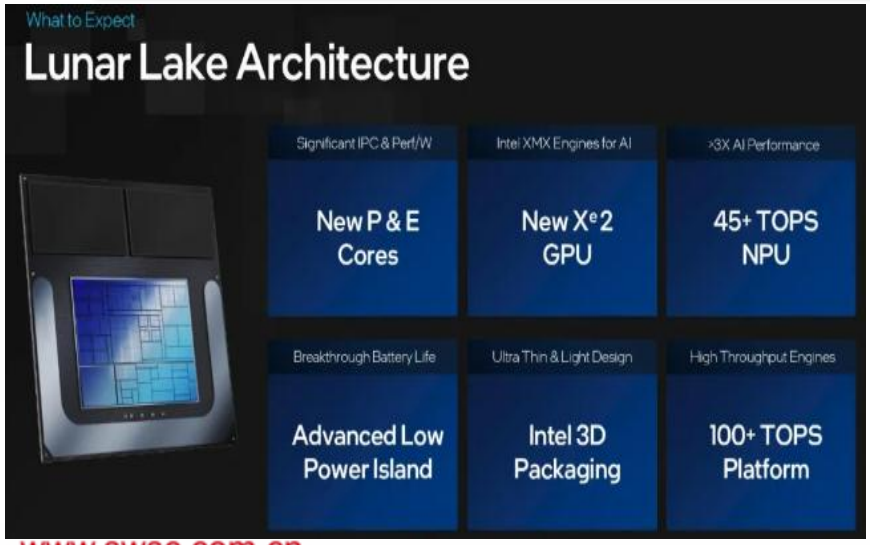
- Ryzen 7040系列于2023年发布，首度集成XDNA AI引擎（NPU），算力达到10TOPS
- Ryzen 8000系列于2023年12月发布，优化NPU提供16TOPS算力
- 下一代 “Strix Point”将于2024年Q3发布，全面升级采用“ZEN5” CPU+ “RDNA3+” GPU+ “XDNA2” NPU架构，实现3倍生成式AI性能

数据来源：Intel，AMD，西南证券整理

2.3.1 Lunar Lake全面架构升级，AI算力与低功耗是核心

- **Lunar Lake预计24Q3推出，围绕AI性能与低功耗方面进行大幅增强**
- 延续分离式模块化设计，但从四个模块简化至两个（计算模块和平台控制器模块），SOC Tile预计预计采用台积电3nm制程而非Intel自家工艺（桌面版Arrow Lake将首先采用Intel 20A）
- Lunar Lake将全面采用全新架构，CPU将升级全新的P&E核，GPU架构将升级为最新的NEW Xe2，NPU架构升级至4.0，最大可提供120TOPS算力
- 首次在处理器内部封装整合内存，可节省40%功耗和250㎜²的主板面积让位给电池等其他部件
- 4+4大小核设计，舍弃LP E-Core而集成了全新的“低功耗岛”架构，不再采用单一物理模块管理节能，将所有可节能的模块纳入统一管理，整体按需开关

Lunar Lake进行全面架构升级



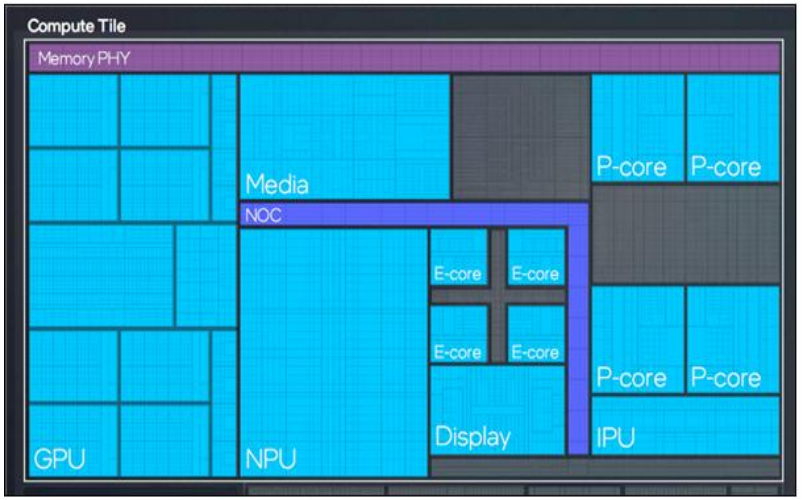
数据来源：Intel，西南证券整理

2.3.1 Lunar Lake全面架构升级，AI算力与低功耗是核心

- **尝试内存封装，降低延迟的同时带来40%功耗节省**
  - Lunar Lake采用全新的MoP（Memory on Package）封装，直接将LPDDR5X-8500内存芯片集成于处理器封装之上，可以大幅加快数据访问速度并降低延迟，最高支持32GB容量与8.5GT/S速率，可节约40%的PHY功耗与250mm²面积，并留出空间让位给电池等其他部件
  - MoP方案类似苹果M系列芯片UMA的设计方式，同时也带来很多争议：32GB内存限制和16bit×4总线或成为性能瓶颈，无法后期进行加装扩容也失去了Intel原有的灵活性优势
- **大小核独立集群，减少静态功耗开销**
  - 4+4大小核心分别位于两个独立族群中，通过NOC总线而非Ringbus总线连接，E-Core的功能与Meteor Lake里面的LP E-core相似，在低负载场景下，P核族群及内部总线无需被激活。

Lunar Lake首次采用内存封装方案

Lunar Lake大小核独立集群



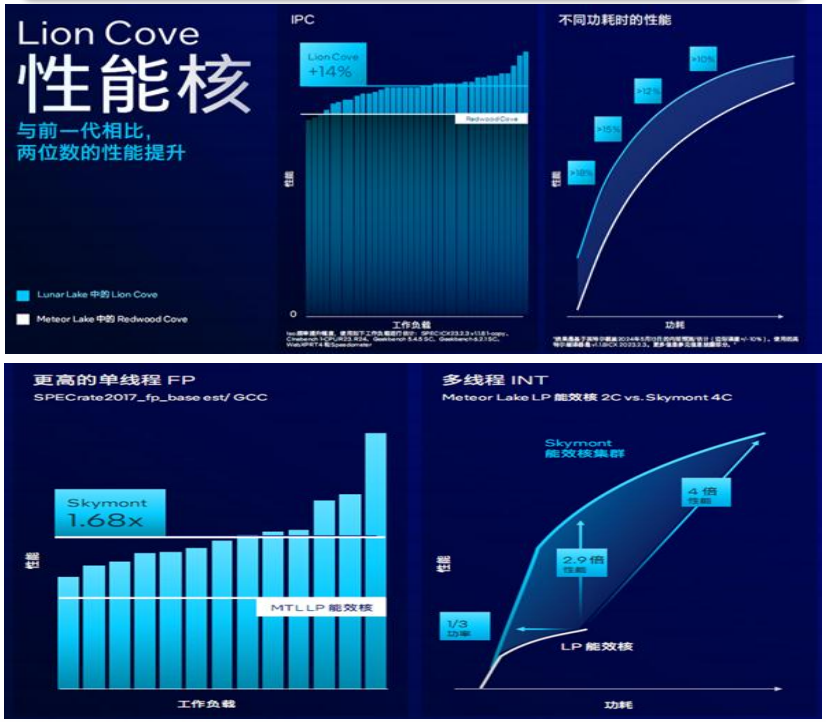
数据来源：Intel，西南证券整理

2.3.1 Lunar Lake全面架构升级，AI算力与低功耗是核心

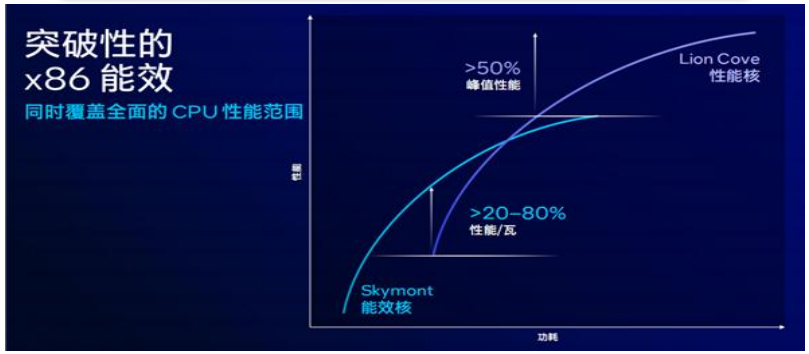
- **全新设计CPU微架构，强调多场景能效最优**

- P核Lion Cove在IPC性能上相比上代平均提升14%，且功耗越低提升越明显；E核在Skymont单线程性能相较上代提升68%，多线程性能相同功耗下提升2.9倍，峰值性能提升4倍
- P、E核组合可带来50%峰值性能的提升，20%-80%的能效提升；叠加软件层面进一步升级Thread Director，丰富动态调整调度策略，设置操作系统隔离区，提供更精细的控制，可将应用能效至多降低35%

P、E核全新升级架构带来性能提升



Lunar Lake电源效率有显著提升

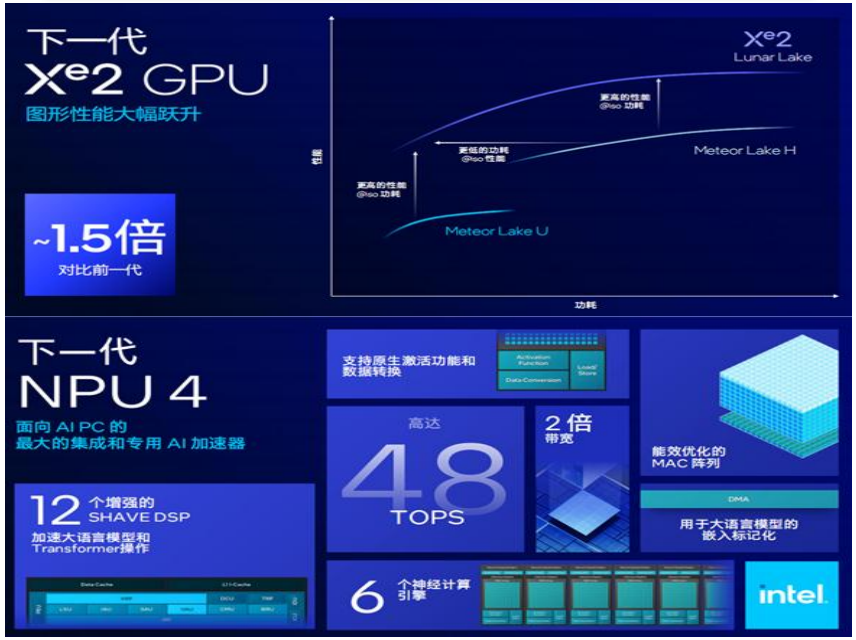


数据来源：Intel，西南证券整理

2.3.1 Lunar Lake全面架构升级，AI算力与低功耗是核心

- **全新GPU提供67TOPS算力，全新NPU提供48TOPS算力**
- GPU升级为Xe2微架构，与Intel即将发布的Battlemage独立显卡一致，并针对低功耗、高能效进行优化，**图形性能相较Meteor Lake GPU的Xe-LPG架构提升50%，AI算力高达67TOPS**
- NPU升级为4.0架构，神经计算引擎从2个增加至6个，12个基于Transformer增强的SHAVE DSP、能效优化的MAC阵列，**带宽相较Meteor Lake提升2倍，AI算力达到48TOPS**
- 尽管NPU峰值算力48TOPS略低于AMD Strix Point的NPU峰值50TOPS，但Lunar lake平台总性能达到120TOPS，高于AMD Strix Point的80TOPS和高通X Elite的75TOPS

GPU&NPU性能大幅提升



Intel近四代芯片对比

	Lunar Lake	Meteor Lake	Raptor Lake	Alder Lake
发布时间	2024H2	2023H2	2023H1	2022H1
CPU制程	TSMC 3nm（预计）	Intel 4	Intel 7	Intel 7
GPU制程	TSMC 3nm（预计）	TSMC 5nm	Intel 7	Intel 7
P-core架构	Lion Cove	Redwood Cove	Raptor Cove	Golden Cove
E-core架构	Skymont	Crestmont	Gracemont	Gracemont
GPU架构	Xe2	Xe-LPG	Iris Xe	Iris Xe
最大内存容量	32GB（处理器封装）	96GB	64GB	64GB
TDP	-	7W-45W	15W-55W	15-55W

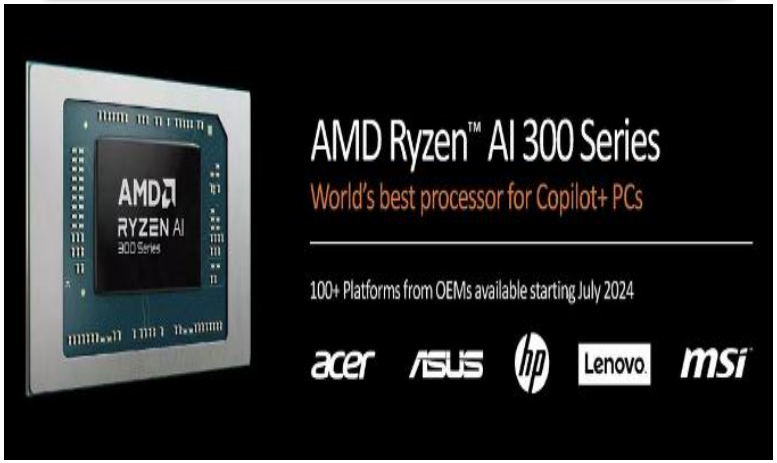
数据来源：Intel，西南证券整理

2.3.1 生态仍是x86阵营的最佳倚仗

- **x86软硬件生态成熟庞大**
  - Wintel生态联盟从1984年建立至今已经发展40年，期间经历了PowerPC、MIPS等冲击，依靠软件开发生态守住市场份额，x86平台如今在PC市场仍有近90%占有率
  - 硬件生态方面，Intel Meteor Lake已经出货800多万颗，产品设计超过230款，遍布48个国家和地区；Lunar Lake已经和超过20家OEM厂商合作，24Q3起有超过80款笔记本陆续上市。同时AMD Strix Point也有100余款机型将从7月份起开始陆续推向市场，对比高通X Elite首发20+款机型仍然有明显优势
  - 软件生态方面，Intel推出AIPC加速计划，目前平台上已有100+ISV厂商提供300多个AI加速功能，优化的大模型也已超过500个
- 我们认为，**由于高通X Elite系列机型正式推出的时间较晚，窗口期仅为一个季度左右。若Arm阵营没能在功耗、创新点上拉开巨大差距，待x86阵营新架构推出，或可重新占据市场主动权。**

Lunar Lake 80+型号电脑将逐步推出

基于Strix Point的100+型号电脑将逐步推出



数据来源：Intel，AMD，西南证券整理

2.3.2 Apple Intelligence发布，带动新一轮创新周期

- **苹果在6月份的WWDC2024发布Apple Intelligence，主要具备以下特点：**
- **系统级AI：**Apple Intelligence融入到苹果的操作系统层面，可在应用中直接调用AI能力，用户无需在单独的AI助手工具与第三方应用之间来回切换，从而提升效率。
- **跨应用的信息整合能力：**在使用过程中，通过为照片、日历、行程和文件等内容创建语义索引，从各种应用中整理和提取信息，找出相关个人数据并提供给AI模型，Apple Intelligence能发现并理解跨应用之间的信息（早期以原生应用为主，后期预计部分第三方亦会支持调用），具备跨平台信息处理能力。
- **采用端云结合的形式。**Apple Intelligence内置苹果自研端侧模型，优先用于本地处理工作；和OpenAI合作外接GPT4o，用于处理更复杂的任务。

实现跨应用整合能力，可从邮件、照片提炼信息回答问题

外接GPT

数据来源：Apple，西南证券整理

2.3.2 苹果的端侧模型布局

- 苹果最初在此轮AI大模型浪潮有所落后。
- 自2010年收购Siri起，苹果前后收购了30多家AI初创公司，但交易金额普遍较低，且几乎集中于图像处理、语音识别、搜索引擎、健康监测等单点功能。
- 苹果一直以来都以消费者使用体验为先，通常会等到技术成熟才正式商用。2023年5月，Tim Cook在苹果的财报电话会议上表示，人工智能有“许多问题需要解决”，重要的是“在开发方法上要深思熟虑”，并计划继续在深思熟虑的基础上将 AI 融入到产品中。
- 苹果过去太注重专利保护，导致相应人才流失。

苹果部分AI领域收购事件

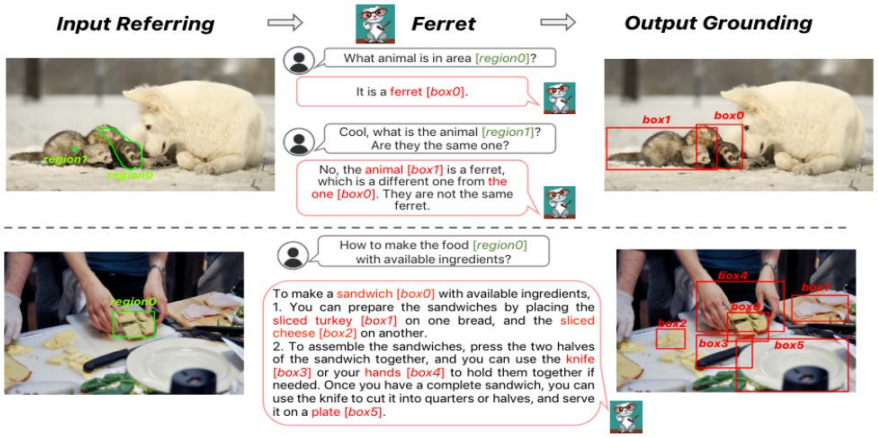
序号	时间	公司	成立时间	收购金额	相关领域
1	2010年4月	Siri	2007年	-	语音识别
2	2010年9月	Polar Rose	2004年	2900万美元	面部识别
3	2013年11月	PrimeSense	2005年	3.6亿美元	深度感知、结构光识别、动感捕捉摄像头
4	2017年2月	RealFace	2014年	200万美元	人脸识别
5	2017年10月	Init.ai	2015年	-	消息助手
6	2018年9月	Shazam	1999年	4亿美元	音乐识别内容推荐
7	2019年8月	Fashwell	2013年	-	图像识别
8	2020年1月	Xnor.ai	2016年	2亿美元	图像识别
9	2020年9月	Vilynx	2011年	5000万美元	AI视频分析
10	2022年2月	AI Music	2016年	3300万美元	AI定制音乐
11	2023年3月	WaveOne	2016年	-	视频压缩AI算法

数据来源：苹果，极客公园，西南证券整理

2.3.2 苹果的端侧模型布局

- **本轮苹果AI功能的更新核心将是围绕Siri的AI大模型改造**
- 根据负责Siri的前工程师：Siri基于笨拙的代码构建，其累赘的设计使得工程师很难添加新功能，如果不进行彻底重写，Siri最终无法成为像ChatGPT那样的人工智能助手。
- 2019年苹果就开始组建专注于对话式AI助手的团队，直接向CEO汇报，但一直没有什么进展、
- 2023年4月开始，苹果不断提高AI研发在内部的优先级，并不断更新招聘需求，主要关于底层大模型研发，以及大模型压缩技术
- 2023年9月，The information爆料，苹果在对话式人工智能的研究上每日投入数百万美元，核心目标之一是让Siri执行多步骤任务
- 2023年10月，苹果发布Ferret开源多模态大语言模型，能够理解和处理图像中任意形状或粒度级别的空间参照，并准确地对开放词汇描述进行定位

苹果Ferret多模态大模型在精确定位的方面取得进步



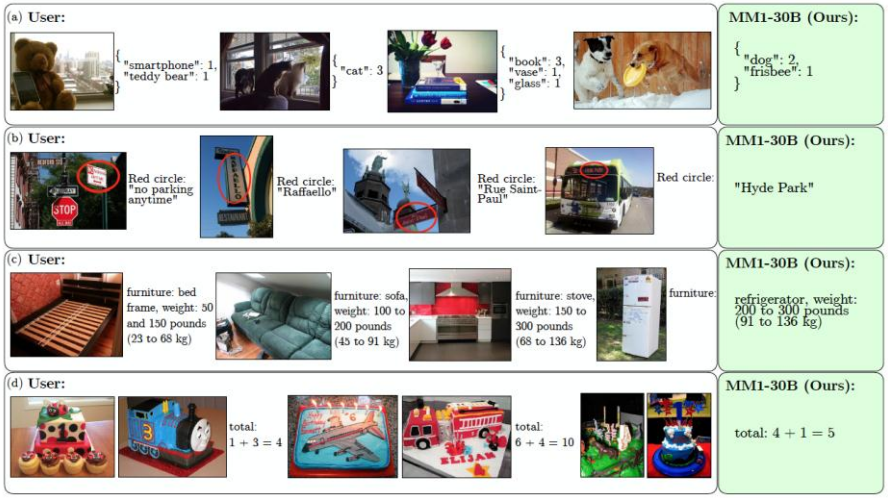
Model	Input Types			Output Grounding	Data Construction			Quantatitive Eval. of Refer/Ground w. Chat
	Point	Box	Free-form		Convention	GPT-Generate	Robustness	
BuboGPT	✗	✗	✗	✓	✓	✗	✗	✗
Vision-LLM	✗	✗	✗	✓	✓	✗	✗	✗
Kosmos-2	✗	✓	✗	✓	✓	✗	✗	✗
Shikra	✓	✓	✗	✓	✓	✓	✗	✗
GPT4-ROI	✗	✓	✗	✗	✓	✗	✗	✗
PVIT	✗	✓	✗	✗	✓	✓	✗	✓
Ferret	✓	✓	✓	✓	✓	✓	✓	✓

数据来源：The information，苹果，量子位，西南证券整理

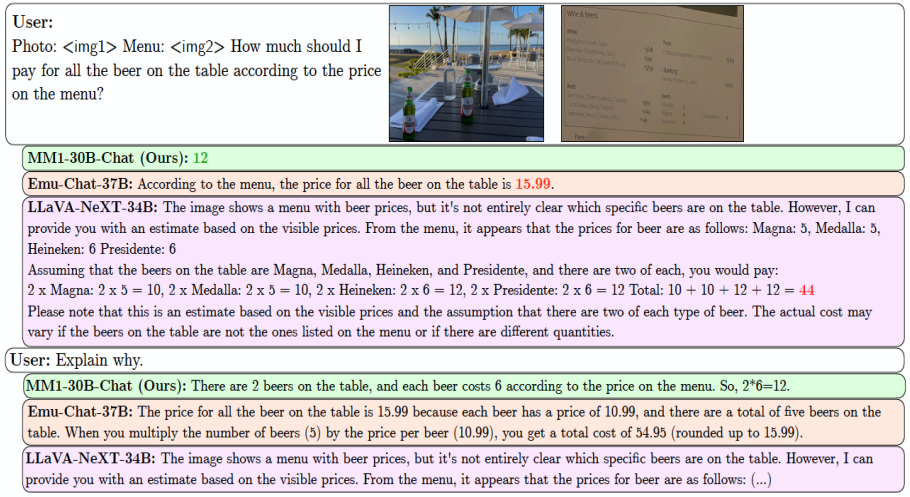
2.3.2 苹果的端侧模型布局——MM1

- 2024年3月苹果发布多模态大模型MM1，MM1由密集模型和混合专家（MoE）变体组成，具有300亿、70亿、30亿参数等多个版本。MM1不仅在预训练指标中实现最优性能SOTA，在一系列已有多模态基准上监督微调后也能保持有竞争力的性能。
- 得益于大规模多模态预训练，MM1模型在上下文预测、多图像和思维链推理等方面表现不错。MM1模型在指令调优后展现出强大的少样本学习能力。在几乎所有基准测试中，苹果的MoE模型比密集模型取得更好的性能，显示MoE进一步扩展的巨大潜力。
- 同时，苹果一改往日封闭的做法，在其技术报告中直接分享了MM1大模型的预训练方法，完整地记载了模型的构建过程。

苹果MM1大模型



由于其大规模多模态预训练，MM1可进行上下文预测。这使得MM1能够：a）计数对象并遵循自定义格式；b）引用图像的部分并执行OCR；c）展示关于日常物品的常识和词汇知识；d）执行基本的数学函数。

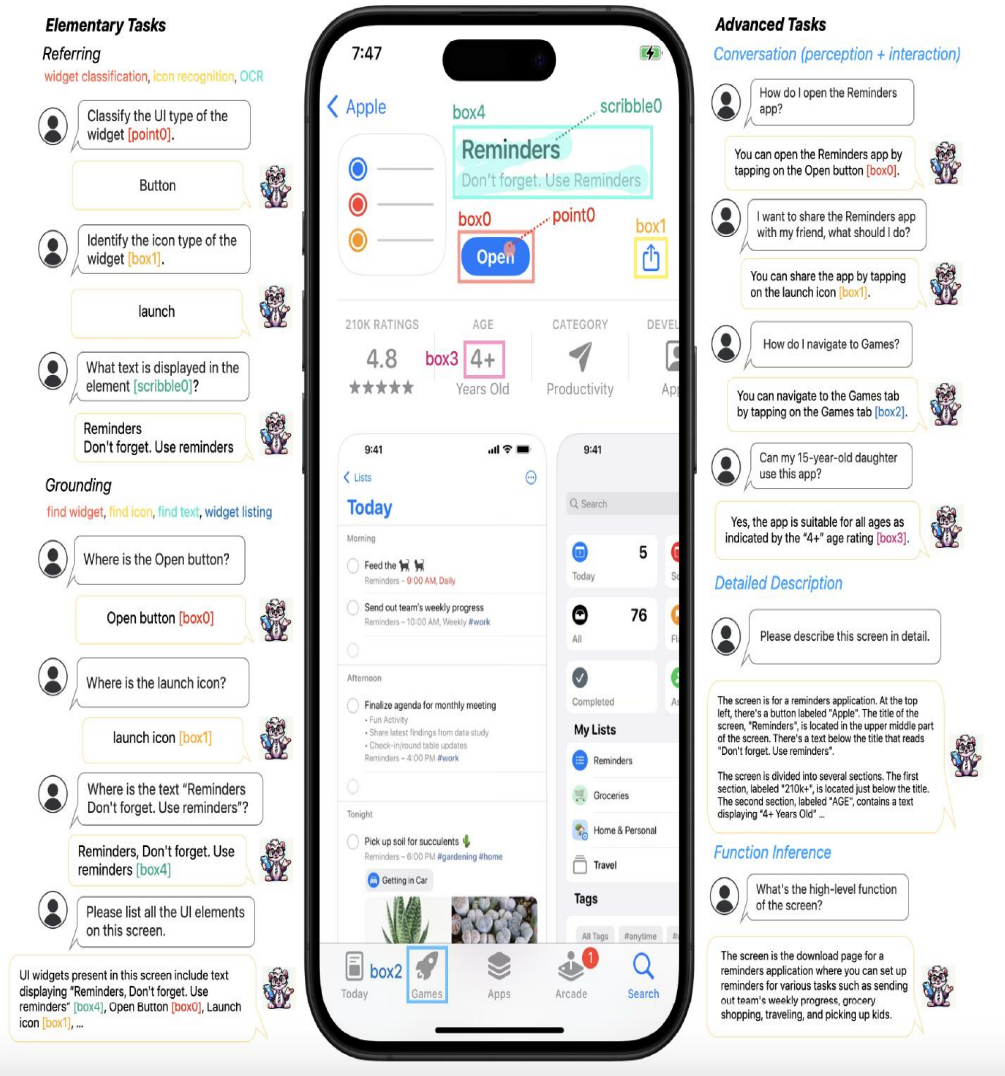


MM1模型可以遵循指令进行跨图像推理。示例和图像来自VILA；当使用思维链提示时，VILA能够正确回答。

数据来源：《MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training》，西南证券整理

2.3.2 苹果的端侧模型布局——Ferret-UI

- 4月8日，苹果展示了多模态模型Ferret-UI。Ferret-UI系统可以理解手机屏幕上的应用程序内容，专为增强对移动端UI屏幕的理解而定制，其配备了引用（referring）、定位（grounding）和推理（reasoning）功能。**Ferret-UI具备了解决现有大部分通用多模态大模型所缺乏的理解用户界面（UI）屏幕并与其有效交互的能力。**Ferret-UI与OpenAI近期发布的GPT-4o、谷歌近期发布的Gemini 1.5 Pro在功能上异曲同工。
- Ferret-UI可执行从基本到复杂任务等11种任务。它能在移动UI屏幕上使用灵活的输入格式（点、框、涂鸦）执行指代任务（例如控件分类、图标识别、OCR）和定位任务（例如查找控件、查找图标、查找文本、控件列表）。这些基本任务为模型提供了丰富的视觉和空间知识，使其能在粗略和精细级别（例如各种图标或文本元素之间）区分UI类型。这些基础知识对于执行更高级的任务至关重要。具体来说，Ferret-UI不仅能在详细描述和感知对话中讨论视觉元素，还可在交互对话中提出面向目标的动作，并通过功能推理推断出屏幕的整体功能。



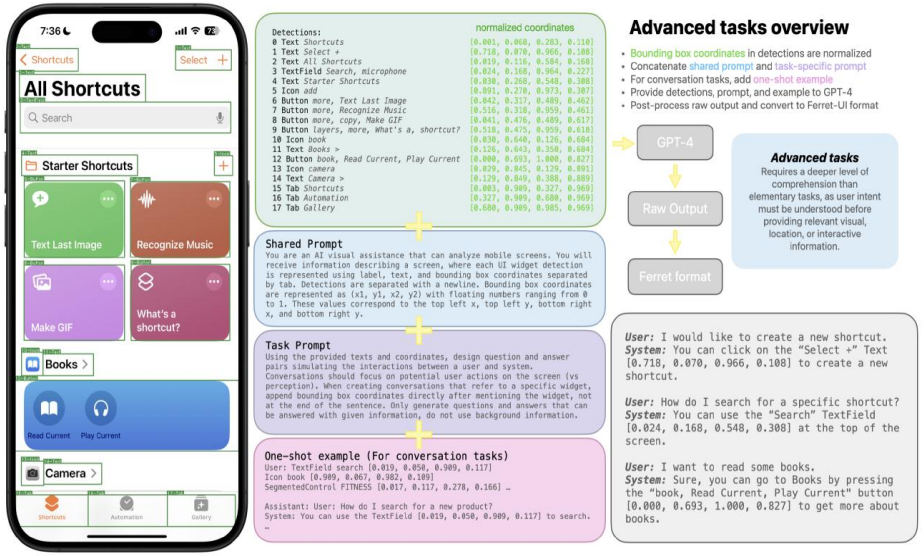
数据来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，西南证券整理

2.3.2 苹果的端侧模型布局——Ferret-UI

- Ferret-UI对基本任务的处理流程：UI检测器输出所有检测到的元素，以及每个元素的类型、文本和边界框。这些检测用于为基本任务创建训练样本。对于定位任务，使用所有元素检测来创建一个用于控件列表的样本，而其余任务一次专注于一个元素。将元素分为图标、文本和非图标/文本控件。对于每种类型，创建一个指代样本和一个定位样本。
- Ferret-UI对复杂任务的处理流程：首先从检测输出中归一化边界框坐标，然后将检测、提示和可选的单次示例发送到GPT-4。对于详细的描述和函数推理，将生成的响应与预先选择的提示配对，以训练Ferret-UI。对于对话任务，直接将 GPT-4输出转换为多回合对话。

Ferret-UI简单任务处理流程

Ferret-UI复杂任务处理流程



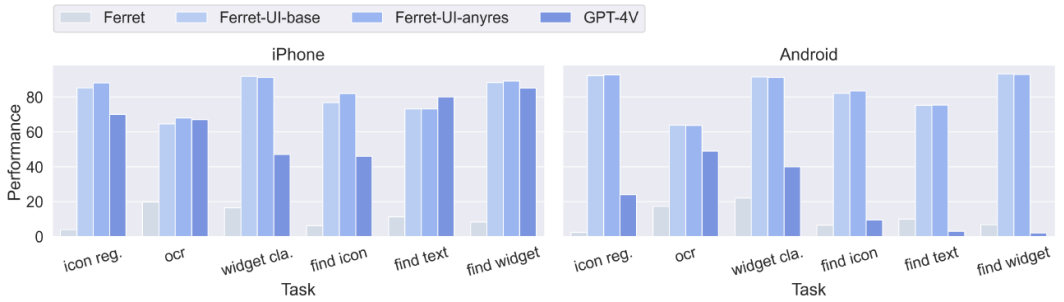
数据来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，西南证券整理

2.3.2 Ferret-UI，理解手机进程并与其交互

- Ferret-UI在简单任务处理上击败了GPT-4V。但在复杂任务处理上还是不如GPT-4V。
- 通过精心设计“任意分辨率”（anyres）以适应各种屏幕宽高比，以及策划包含广泛的基本和高级UI任务的训练样本，**Ferret-UI在引用、定位和推理方面表现出显著的熟练程度。这些增强能力的引入预示着在众多下游UI应用中或将取得重大进步**，从而扩大Ferret-UI在这一领域所能提供的潜在益处。

Ferret-UI性能参数

	Public Benchmark			Elementary Tasks				Advanced Tasks	
	S2W	WiC	TaP	Ref-i	Ref-A	Grd-i	Grd-A	iPhone	Android
Spotlight [30]	106.7	141.8	**88.4**	-	-	-	-	-	-
Ferret [53]	17.6	1.2	46.2	13.3	13.9	8.6	12.9	20.0	20.7
Ferret-UI-base	113.4	**142.0**	78.4	80.5	**82.4**	79.4	83.5	73.4	80.5
Ferret-UI-anyres	**115.6**	140.3	72.9	**82.4**	**82.4**	**81.4**	**83.8**	93.9	71.7
GPT-4V [1]	34.8	23.5	47.6	61.3	37.7	70.3	4.7	**114.3**	**128.2**



	iPhone					Android				
	DetDes	ConvP	ConvI	FuncIn	**Avg**	DetDes	ConvP	ConvI	FuncIn	**Avg**
Ferret [53]	2.5	34.7	23.7	19.1	20.0	2.0	33.9	24.9	21.9	20.7
Fuyu [6]	5.0	24.6	18.8	35.7	21.0	2.0	20.8	44.5	36.1	25.9
CogAgent [20]	53.1	59.7	74.8	71.9	64.9	28.0	58.5	90.1	**90.5**	66.8
Ferret-UI-base	64.5	75.0	77.5	76.5	73.4	90.8	72.8	79.3	79.2	80.5
Ferret-UI-anyres	**97.4**	92.1	91.1	**95.2**	93.9	86.4	70.3	50.2	77.3	70.1
GPT-4V [1]	66.8	**105.6**	**198.5**	86.3	**114.3**	**126.6**	**109.4**	**188.6**	88.3	**128.2**

数据来源：《Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs》，西南证券整理

2.3.2 苹果的端侧模型布局——Open ELM

- 4月26日，苹果宣布了更大的端侧AI推进，推出全新的开源大语言模型OpenELM。OpenELM包含2.7亿、4.5亿、11亿和30亿个参数的四种版本，定位于超小规模模型，运行成本更低，可在手机和笔记本电脑等设备上运行文本生成任务。
- 同时，公司开源了OpenELM模型权重和推理代码、数据集、训练日志、神经网络库CoreNet。
- **OpenELM在不依赖于模型规模的持续膨胀下，通过算法和架构创新达到与GPT-4相当的性能。**OpenELM使用了“分层缩放”策略，有效分配Transformer模型每一层参数，从而提升准确率。在约10亿参数规模下，OpenELM与OLMo相比，准确率提高了2.36%，同时需要的预训练token数量减少了50%。
- OpenELM对 Siri 最显著的强化在于上下文理解的升级，它可以掌握诸如“再次播放那首歌”或“给她打电话”等参考信息，甚至预测用户的需求和偏好，根据过去的行为和上下文理解建议或启动操作。

苹果OpenELM参数和性能对比

Model	Model size	Pretraining tokens	ARC-c	ARC-e	BoolQ	HellaSwag	PIQA	SciQ	WinoGrande	Average	Average w/o SciQ
OpenELM (Ours)	0.27 B	1.5 T	26.45	45.08	53.98	46.71	69.75	84.70	53.91	54.37	49.31
MobiLlama [44]	0.50 B	1.3 T	26.62	46.04	55.72	51.06	71.11	83.60	53.20	55.34	50.63
OpenELM (Ours)	0.45 B	1.5 T	**27.56**	**48.06**	**55.78**	**53.97**	**72.31**	**87.20**	**58.01**	**57.56**	**52.62**
TinyLlama [50]	1.10 B	3.0 T	30.12	55.25	57.83	59.20	73.29	-	59.12	-	55.80
OpenLM [18]	1.00 B	1.6 T	31.00	56.00	**65.00**	61.00	74.00	-	60.00	-	57.83
MobiLlama [44]	0.80 B	1.3 T	28.84	49.62	60.03	52.45	73.18	85.90	55.96	58.00	53.35
MobiLlama [44]	1.26 B	1.3 T	31.91	56.65	60.34	62.18	74.81	89.10	59.27	62.04	57.53
OLMo [17]	1.18 B	3.0 T	31.06	**57.28**	61.74	62.92	75.14	87.00	59.98	62.16	58.02
OpenELM (Ours)	1.08 B	1.5 T	**32.34**	55.43	63.58	**64.81**	**75.57**	**90.60**	**61.72**	**63.44**	**58.91**
OpenELM (Ours)	3.04 B	1.5 T	35.58	59.89	67.40	72.44	78.24	92.70	65.51	67.39	63.18

(a) Results on zero-shot tasks with respect to the standard metrics defined in Tab. 3a.

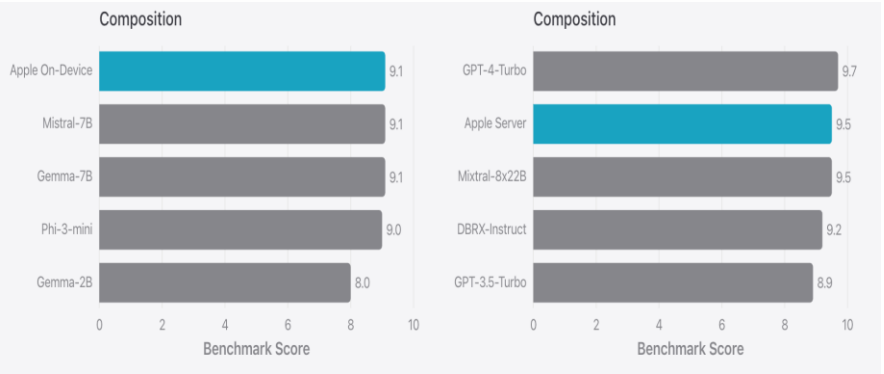
Model	Model size	Pretraining tokens	ARC-c	HellaSwag	MMLU	TruthfulQA-mc2	WinoGrande	Average
Cerebras-GPT [14]	0.26 B	5.1 B	22.01	28.99	26.83	45.98	52.49	35.26
OPT [58]	0.35 B	0.2 T	23.55	36.73	26.02	40.83	52.64	35.95
OpenELM (Ours)	0.27 B	1.5 T	**27.65**	**47.15**	25.72	39.24	**53.83**	**38.72**
Pythia [5]	0.41 B	0.3 T	24.83	41.29	25.99	40.95	54.38	37.49
MobiLlama [44]	0.50 B	1.3 T	29.52	52.75	**26.09**	37.55	56.27	40.44
OpenELM (Ours)	0.45 B	1.5 T	**30.20**	**53.86**	26.01	40.18	**57.22**	**41.50**
MobiLlama [44]	0.80 B	1.3 T	30.63	54.17	25.2	38.41	56.35	40.95
Pythia [5]	1.40 B	0.3 T	32.68	54.96	25.56	**38.66**	57.30	41.83
MobiLlama [44]	1.26 B	1.3 T	34.64	63.27	23.87	35.19	60.77	43.55
OLMo [17]	1.18 B	3.0 T	34.47	63.81	26.16	32.94	60.46	43.57
OpenELM (Ours)	1.08 B	1.5 T	**36.69**	**65.71**	**27.05**	36.98	**63.22**	**45.93**
OpenELM (Ours)	3.04 B	1.5 T	42.24	73.28	26.76	34.98	67.25	48.90

(b) Results on OpenLLM Leaderboard tasks with respect to the metrics defined in Tab. 3b.

Model	Model size	Pretraining tokens	ARC-c	CrowS-Pairs	HellaSwag	MMLU	PIQA	RACE	TruthfulQA	WinoGrande	Average
OpenELM (Ours)	0.27 B	1.5 T	27.65	66.79	47.15	25.72	69.75	30.91	39.24	53.83	45.13
MobiLlama [44]	0.50 B	1.3 T	29.52	65.47	52.75	**26.09**	71.11	32.15	37.55	56.27	46.37
OpenELM (Ours)	0.45 B	1.5 T	**30.20**	**68.63**	**53.86**	26.01	**72.31**	**33.11**	**40.18**	**57.22**	**47.69**
MobiLlama [44]	0.80 B	1.3 T	30.63	66.25	54.17	25.2	73.18	33.68	**38.41**	56.35	47.23
MobiLlama [44]	1.26 B	1.3 T	34.64	70.24	63.27	23.87	74.81	35.02	35.19	60.77	49.73
OLMo [17]	1.18 B	3.0 T	34.47	69.95	63.81	26.16	75.14	**36.75**	32.94	60.46	49.96
OpenELM (Ours)	1.08 B	1.5T	**36.69**	**71.74**	**65.71**	**27.05**	**75.57**	36.46	36.98	**63.22**	**51.68**
OpenELM (Ours)	3.04 B	1.5 T	42.24	73.29	73.28	26.76	78.24	38.76	34.98	67.25	54.35

(c) Results on LLM360 tasks with respect to the metrics defined in Tab. 3c.

苹果端侧模型写作能力测评



数据来源：《OpenELM: An Efficient Language Model Family with Open Training and Inference Framework》，西南证券整理

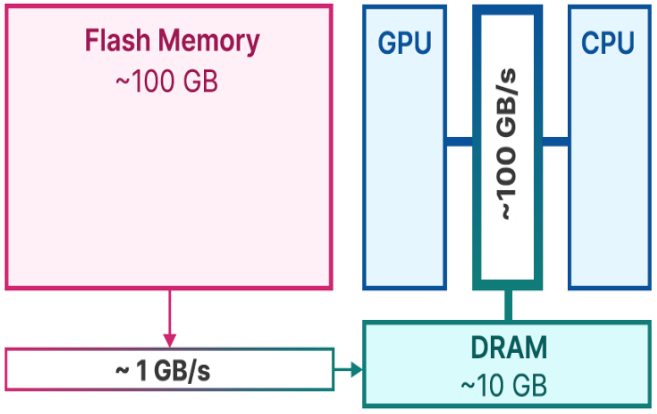
2.3.2 端侧带来硬件更新，芯片和内存是关键

- 苹果在其发表的论文《LLM in a flash: Efficient Large Language Model Inference with Limited Memory》中详细阐述了如何在DRAM容量有限的设备中高效地运行LLMs。
- 苹果将模型参数存储在Flash中，按需将其带入DRAM，从而高效地运行超出可用DRAM容量的LLMs。为此，苹果构建了一个推理成本模型，其通过引入两种主要技术，即“窗口化”（通过重用之前激活的神经元战略性地减少数据传输）、以及“行列打包”（针对闪存的顺序数据访问优势，增加了从闪存读取的数据块大小），减少了从闪存传输的数据量、以及以更大更连续的块读取数据。这些方法使得**能运行的模型大小达到可用DRAM容量的两倍，与CPU和GPU中的简单加载方法相比，推理速度分别提高了4-5倍和20-25倍**。这些方法促成了数据负载的显著减少和内存使用效率的提高，对于端侧部署先进的AI模型尤为关键。

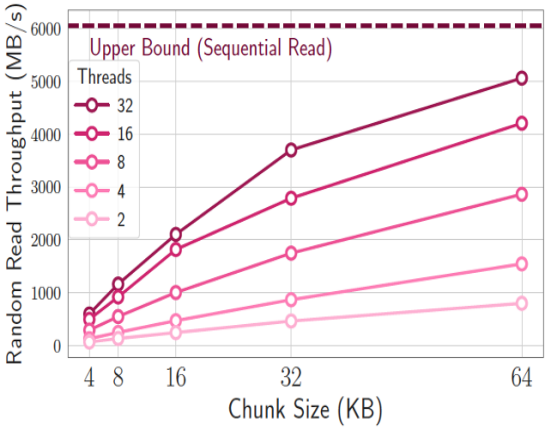
闪存提供更高的容量，但带宽较小

闪存中随机读取的吞吐量随着顺序块大小和线程数量增加而增加

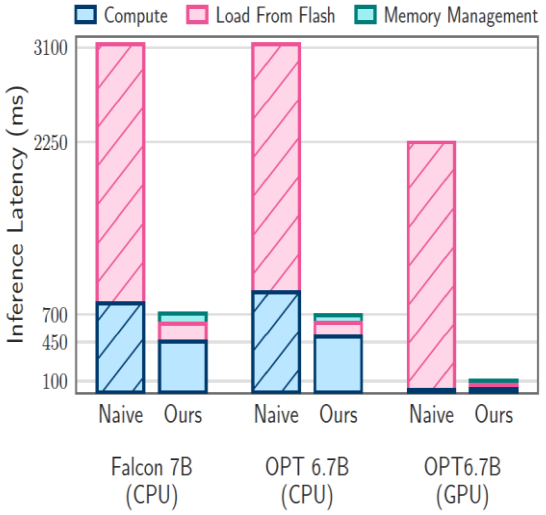
苹果的方法使得单token的推理延迟大幅缩减



(a) Bandwidth in a unified memory architecture



(b) Random read throughput of flash memory



数据来源：苹果，《LLM in a flash: Efficient Large Language Model Inference with Limited Memory》，西南证券整理

2.3.2 端侧带来硬件更新，芯片和内存是关键

- 由于本地运行AI模型对算力和内存有一定的要求，根据苹果官网所示，支持Apple Intelligence端侧运算的芯片和终端要求如下：
  - iPhone目前仅支持A17 Pro芯片，满足要求的仅iPhone 15 Pro/Pro Max；
  - Mac和iPad端支持M1及以上版本的芯片，近几年基于M芯片的苹果Mac、iPad基本都支持
- 由于对旧款硬件产品的支持有限，Apple Intelligence或带动苹果硬件产品线的换机周期。特别是iPhone支持的旧款机型较少，iPhone或存在更强的换机需求。

A系列芯片参数

型号	CPU性能	GPU性能	NPU性能	对应RAM	工艺制程	搭载机型
A10	2×大核，2.34GHz 2×小核，1.09GHz	PowerVR GT7600 Plus 900MHz 345.6 GFLOPS	——	64-bit单通道 LPDDR4，1600MHz 25.6 GB/s	TSMC 16nm	iPhone 7 iPhone 7 Plus
A11	2×大核，2.39GHz 4×小核，1.19GHz	3核GPU自研 1066MHz 408 GFLPOS	2×神经网络引擎 600 BOPS	64-bit单通道 LPDDR4，2133MHz 34.1 GB/s	TSMC 10nm	iPhone 8/8 Plus iPhone X
A12	2×大核，2.49GHz 4×小核，1.59GHz	4核GPU自研 1125MHz 576 GFLOPS	8×神经网络引擎 5 TOPS	64-bit单通道 LPDDR4，2133MHz 34.1 GB/s	TSMC N7 7nm	iPhone XS/XS Max iPhone XR
A13	2×大核，2.65GHz 4×小核，1.80GHz	4核GPU自研 1575MHz 806 GFLOPS	8×神经网络引擎 5.5 TOPS	64-bit单通道 LPDDR4X，2133MHz 34.1 GB/s	TSMC N7P 7nm	iPhone 11 iPhone 11 Pro/Pro Max iPhone SE
A14	2×大核，2.99GHz 4×小核，1.82GHz	4核GPU自研 1700MHz 870 GFLOPS	8×神经网络引擎 11 TOPS	64-bit单通道 LPDDR4X，2133MHz 34.1 GB/s	TSMC N5 5nm	iPhone 12/12 Mini iPhone 12 Pro/Pro Max
A15	2×大核，2.93-3.23GHz 4×小核，1.82GHz	4/5核GPU自研 1175 GFLOPS	16×神经网络引擎 15.8 TOPS	LPDDR4X，6GB 3200MHz 42.7 GB/s	TSMC N5P 5nm	iPhone 13/13 Mini iPhone 13 Pro/Pro Max iPhone SE3 iPhone 14/14 Plus
A16	2×大核，3.46GHz 4×小核，2.02GHz	5核GPU自研 1468 GFLOPS	16×神经网络引擎 17 TOPS	LPDDR5，6GB	TSMC N4 4nm	iPhone 14 Pro/Pro Max iPhone 15/15 Plus
A17 Pro	**2×大核，3.70GHz 4×小核，2.02GHz**	**6核GPU自研**	**16×神经网络引擎 35 TOPS**	**LPDDR5，8GB**	**TSMC 3nm**	**iPhone 15 Pro iPhone 15 Pro Max**

数据来源：苹果，西南证券整理

2.3.2 AI倒逼硬件更新，内存是关键

- 苹果当前对于端侧AI的NPU算力要求并没有那么高，**而更关键在于至少8GB运行内存**：
  - M1芯片只提供11TOPS算力，而A16提供17TOPS算力；
  - M1芯片提供16GB运存，A16最高提供6GB运存。
- 同样，参考三星Galaxy S24系列搭载Gemini nano（1.8B和3.25B），内存下限也为8GB；而其他主流安卓厂商的模型大小基本在7B，根据OPPO、Vivo等技术负责人采访，当前70亿参数大模型经过压缩和轻量化之后可压缩至4GB左右，再加上OS占用和APP保活，最低内存要求在12GB左右。
- 根据Yole，2023年高端智能手机平均DRAM为9GB，伴随后期AI手机渗透率提升以及端侧模型可能的参数增加，预期2024年平均DRAM增长至10GB，2026年增长至26GB。

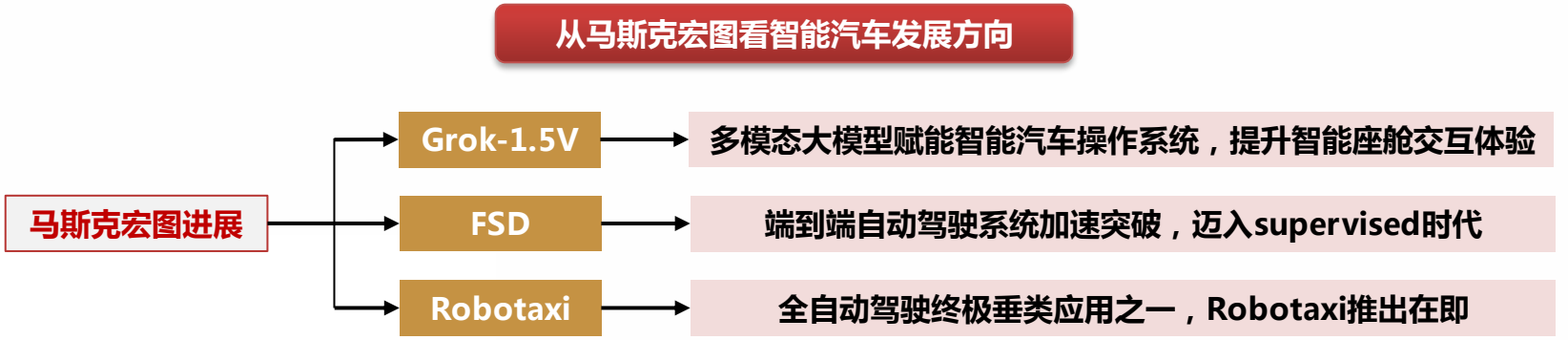
M系列芯片参数

型号	M1	M2 Max	M2 Ultra	M3	M3 Pro	M3 Max	M4
CPU	8核中央处理器，4×性能核心 +4×能效核心	12核中央处理器，8×性能核心 +4×能效核心	24核中央处理器，16×性能核心 +8×能效核心	8核，4×性能核心 +4×能效核心	12核，6×性能核心 +6×能效核心	16核，12×性能核心+4×能效核心	10核，4×性能核心 +6×能效核心
GPU	8核图形处理器，2.6TFLOPS（FP32）	38核图形处理器，13.49 TFLOPS（FP32）	60/76核图形处理器，27.2 TFLOPS（FP32）	10核，硬件加速光线追踪	18核，硬件加速光线追踪	40核，硬件加速光线追踪 17.04 TFLOPS（FP32）	10 核，硬件加速光线追踪
NPU	16核神经网络引擎，11TOPS	16 核神经网络引擎，16TOPS	32 核神经网络引擎，31.6TOPS	16 核神经网络引擎，18TOPS	16 核神经网络引擎，18TOPS	16 核神经网络引擎，18TOPS	16 核神经网络引擎，38TOPS
支持内存	最高16GB	最高96GB	最高192GB	最高24GB	最高36GB	最高128GB	最高24GB
内存带宽	68.25GB/s	400GB/s	800GB/s	100GB/s	150GB/s	400GB/s	120GB/s
工艺制程	TSMC 5nm	TSMC 5nm	TSMC5nm	TSMC 第一代3nm	TSMC 第一代3nm	TSMC 第一代3nm	TSMC 第二代3nm

数据来源：苹果，Yole，西南证券整理

2.3.3 大模型赋能智能汽车

- **Musk的三大AI布局宏图：1）Grok-1.5V：Grok系列模型有望成为24年全球黑马，加速追赶第一梯队，重塑开源竞争格局。**xAI于2024年4月12日发布其首款多模态大模型Grok-1.5V，除文本功能外，模型还可以处理包括文档、图表、图片在内的各种视觉信息，并能进行多学科推理。我们认为，Grok多模态大模型有望反哺特斯拉FSD系统的多模态推理和人机交互体验，加速大模型自动驾驶的发展。**2）FSD v12.3.3：自动驾驶技术迈上新台阶，端到端能力持续领先。**2024年3月31日，特斯拉发布FSD v12.3.3版本，并开始向美国用户陆续推送，宣布全美用户可免费试用1个月，并将FSD的Beta标识改为Supervised标志，从此开启“有监督”时代。此外，特斯拉官方媒体账号于4月13日宣布，FSD（Supervised）的月度订阅费用从199美元降至99美元，同时正式在加拿大开启月度订阅模式，费用为99加元（约72美元）。我们认为，FSD系统的升级迭代以及订阅费用的下降，有望进一步推动FSD渗透率提升，提升汽车软件价值。**3）Tesla Robotaxi：复用FSD能力，引领共享化时代。**2024年4月5日，马斯克在其官方媒体账号上宣布，特斯拉将于2024年8月8日推出Robotaxi，Robotaxi有望依托FSD系统实现全自动驾驶，打造B端终极垂类应用。
- 以马斯克宏图为鉴，我们认为智能汽车领域在2024年主要有三大发展方向，**一是大模型上车，赋能汽车操作系统和智能座舱人机交互体验；二是端到端自动驾驶系统持续升级，推动搭载渗透率和付费渗透率提升；三是关注Robotaxi推广，把握主要玩家推进节奏。**



数据来源：Tesla，西南证券整理

2.3.3 大模型赋能智能汽车

- **FSD Beta时代：**FSD的Beta时代从2020年10月21日的开始，共历时3年零5个月。**从软件系统的改进方向来看，最初集中于智能座舱领域**（如遥控召唤、UI界面、车载娱乐等功能），**而后逐步发展至智能驾驶**（前期以辅助自动驾驶为主，如辅助自动转向/变道/泊车/导航等）**和动力领域**（如充电、制热等问题）。后续，**Beta v9**从底层架构对**感知**算法进行重写，采用**纯视觉方案**；**Beta v10**对神经网络进行**重新训练**；**Beta v11**对大量基础性软件进行**改写**，迈入神经网络**强化**阶段；**Beta v12引入端到端神经网络**，致力于感知决策一体化。
- **FSDsupervised时代：**2024年3月31日起，特斯拉宣布FSD从Beta进入Supervised时代，标志着其端到端算法已具备泛用性和安全性，自动驾驶在大模型的赋能下已经具备可推广性。

特斯拉 FSD Beta 版本迭代时间线

21H1及以前		21H2		22H1		22H2		23H1		23H2		24H1	
2020.40.8.10	2020.10.21	v9.0	2021.07.10	v10.8.1	2022.01.07	v10.13	2022.07.18	v10.69.25.2	2023.01.16	v11.4.5	2023.07.08	v12.1.1	2024.01.12
2020.40.8.11	2020.10.24	v9.1	2021.07.31	v10.9	2022.01.16	v10.69	2022.08.21	v11.3	2023.02.14	v11.4.6	2023.07.21	v12.1.2	2024.01.21
2020.40.8.12	2020.10.31	v9.2	2021.08.15	v10.10	2022.02.01	v10.69.1	2022.08.30	v11.3.1	2023.03.07	v11.4.7	2023.08.24	v12.2	2024.02.10
2020.44.10.2	2020.11.14	v10.0	2021.09.11	v10.10.1	2022.02.15	v10.69.1.1	2022.09.04	v11.3.2	2023.03.19	v11.4.7.1	2023.09.19	v12.2.1	2024.02.19
v6.0	2020.11.27	v10.0.1	2021.09.17	v10.10.2	2022.02.18	v10.69.2	2022.09.10	v11.3.3	2023.03.25	v11.4.7.2	2023.09.24	v12.3	2024.03.12
v7.0	2020.12.14	v10.1	2021.09.27	v10.11	2022.03.13	v10.69.2.1	2022.09.17	v11.3.4	2023.04.01	v11.4.7.3	2023.10.19	v12.3.1	2024.03.24
v8.0	2020.12.30	v10.2	2021.10.11	v10.11.1	2022.03.26	v10.69.2.2	2022.09.19	v11.3.5	2023.04.01	v11.4.8	2023.11.18	v12.3.2	2024.03.26
v8.1	2021.01.26	v10.3	2021.10.24	v10.11.2	2022.04.03	v10.69.2.3	2022.10.03	v11.3.6	2023.04.08	v11.4.8.1	2023.11.30	v12.3.2.1	2024.03.28
v8.2	2021.03.04	v10.3.1	2021.10.25	v10.12	2022.05.18	v10.69.2.4	2022.10.31	v11.4	2023.04.17	v11.4.9	2023.12.15	v12.3.3	2024.03.31
		v10.4	2021.11.06	v10.12.1	2022.05.20	v10.69.3	2022.11.01	v11.4.1	2023.05.11	v12	2023.11.24	v12.3.4	2024.04.16
		v10.5	2021.11.21	v10.12.2	2022.05.27	v10.69.3.1	2022.11.18	v11.4.2	2023.05.26	v12.1	2023.12.22		
		v10.6	2021.12.04			v10.69.3.2	2022.12.05	v11.4.3	2023.06.09				
		v10.6.1	2021.12.11			v10.69.3.3	2022.12.11	v11.4.4	2023.06.19				
		v10.7	skip			v10.69.25	2022.12.19						
		v10.8	2021.12.24			v10.69.25.1	2022.12.23						
						v11	2022.11.11						
						v11.2	2022.12.25						

重写感知算法（对应HydraNet）

开启FSD Beta时代

对神经网络进行重新训练

强化

引入端到端

迈入Supervised时代

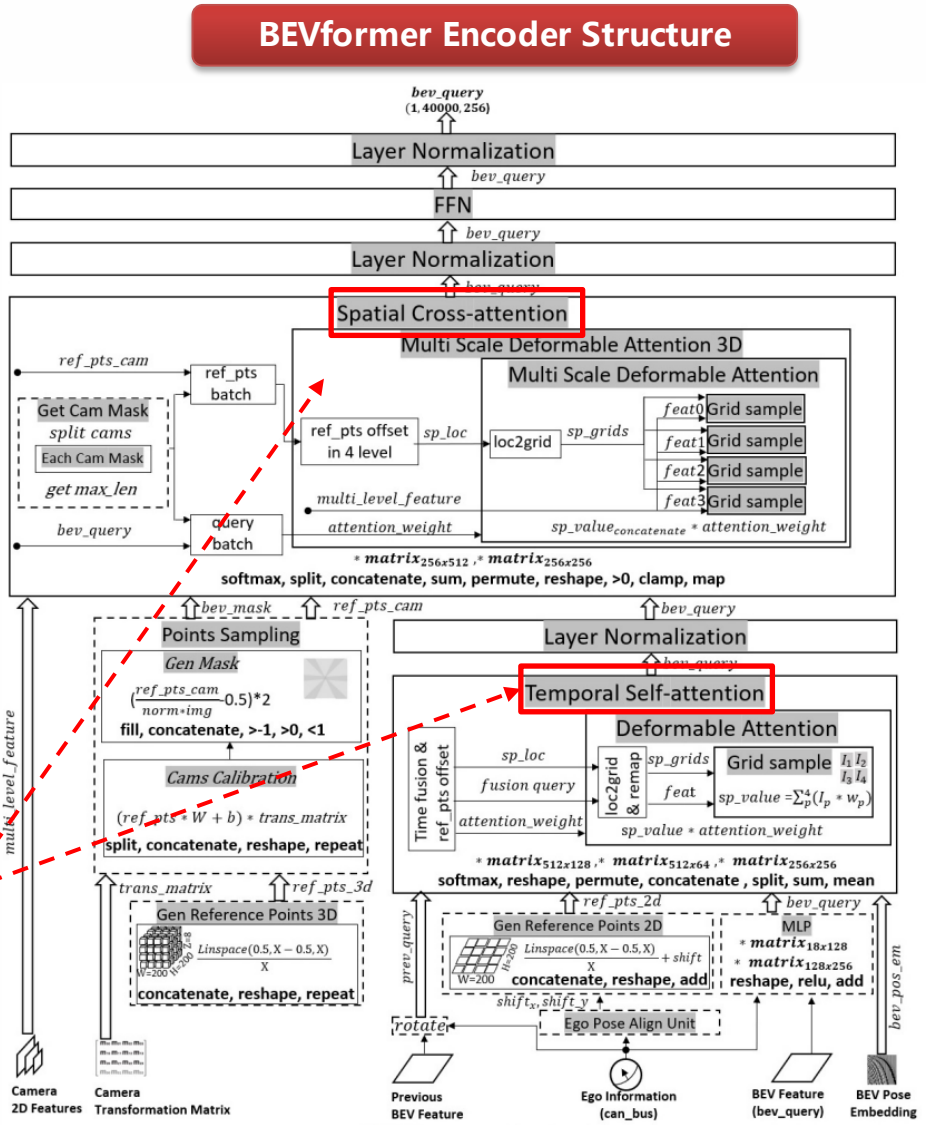
数据来源：Tslachan社交媒体账号，西南证券整理

2.3.3 技术路径：引入Transformer，大幅提升信息处理能力

- 2021年特斯拉于AI Day首次在算法层面引入Transformer取代CNN-head；后续小鹏汽车等国内车企也积极引入Transformer架构，并在短时间内完成了架构的重写。
- **Transformer架构在自动驾驶系统的感知环节中的运用优势：**①Transformer在自然语言处理领域和计算机视觉感知领域均能发挥作用。②Transformer在处理大规模数据量场景上具备优势，较神经网络可以更好地在海量图像数据中识别数据间的关联关系，更有利于构建向量空间。③Transformer网络架构**引入注意力机制**，关注重要信息而非全部信息，在时间性方面具有更高的**并行计算效率**，在空间性能方面具有更强的**泛化能力**。

①时间自注意机制：通过自我信息校准对由previous BEV feature和current BEV feature初始化的bev query执行可变形注意(deformable attention)。

②空间交叉注意机制：从2D摄像头特征中提取BEV特征，且同样运用可变形注意机制，采用多摄像头query，增加两大模块，一是摄像头掩模模块，可生成BEV空间中的每个摄像头掩模，另一个是多级偏移模块，可获得4个级别的参考点偏移。



数据来源：《Transformer-based models and hardware acceleration analysis in autonomous driving: A survey》，西南证券整理

2.3.3 技术路径：BEV实现动态还原，占用网络展现4D泛化世界

- **AI算法在视觉呈现上赋予感知系统“脑补”能力，感知系统逐步具备实时性、更稳定、更精准。**自动驾驶感知系统形成的视觉表达从透视图逐渐发展到鸟瞰图和占用网络，道路还原从2D空间扩展为3D、4D空间，使车辆在动态运动的过程中能够实时构建现实地图，在多颗摄像头的感知下迅速追踪物体的距离和速度、发现被遮挡的物体，并增强现实世界的还原细节和精度，让系统的感知呈现更加符合人类驾驶的需求。

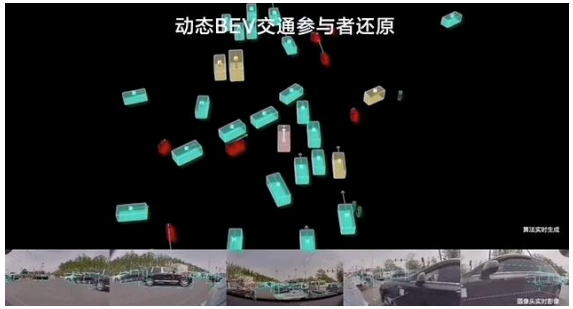
- **透视图（Perspective View）：**即人眼通常看到的2D视图。在人类视觉中，难以看到被遮挡的物体，但在实际驾驶过程中，人类驾驶员可以凭借经验和记忆对可能存在遮挡情况的风险进行规避，但自动驾驶系统如果是基于透视图的视觉进行感知和预测，车辆则很难做到提前预警和规避。

- **鸟瞰图（Birds View）：**即自上而下的视图，具备上帝视角。鸟瞰图感知方案可以在3D空间上分离所有对象，解决透视图中视野被遮挡的问题，减少对自动驾驶对高精地图的依赖，但在高度检测上效果不够理想。

- **占用网络（Occupancy Network）：**占用网络通过算法对物理世界进行**数据化和泛化建模**，在3D空间上测出不同物体的**高度**，呈现**4D视觉。**例如，识别道路上的垃圾桶、临时施工牌等障碍物。

鸟瞰图和占用网络的视觉呈现对比

- 静态BEV网络通过感知还原道路结构，减少对高精地图的依赖。



- 可以解决视野被遮挡的问题，并实时动态还原现实道路的情况

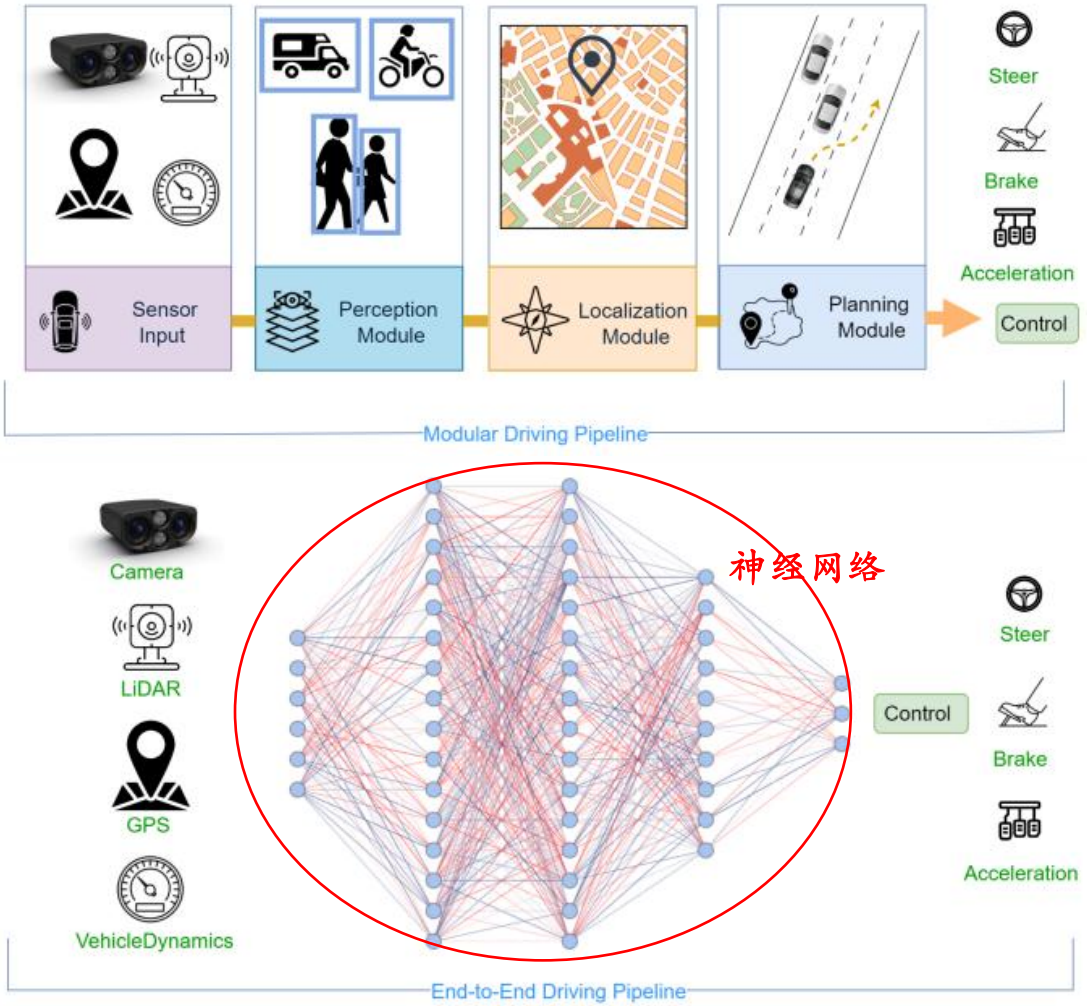
- 可测量出障碍物的高度，识别细节物体

数据来源：Tesla，西南证券整理

2.3.3 技术路径：端到端取代规则模块，泛化能力增强应对corner case

- 通过对比右侧的模块化和端到端两大技术**路径示意图**，我们更能直观地理解两者的区别：模块化方案由众多子模块组成，每个子模块对应特定的任务和功能（Rule-based）；端到端方案则是输入感知信息并直接生成控制信号，中间通过统一的神经网络进行处理（End-to--End）。
- **端到端有望突破性能天花板，找到近似最优解。**对比分而治之和端到端两种解决办法，分而治之可以在有限的精力内快速实现性能的提升、并形成解决方案，但该方法容易陷入局部最优解，导致性能上限仅为80%。而端到端解决方案通过反复多次、集中优化一系列组件，从而不断突破性能天花板，直至实现完全的端到端解决方案，从而摆脱局部最优解的痛点，找到近似全局的最优解。

模块化&端到端 自动驾驶技术路径示意图



数据来源：《Recent Advancements in End-to-End Autonomous Driving using Deep Learning: A Survey》，西南证券整理

2.3.3 Grok1.5V增强真实世界理解能力

- **Grok系列模型推进节奏紧凑，强调真实世界理解能力。**

- **发展时间线：1）**2023年11月3日，Grok-1大语言模型初始版本发布，该模型基于Transformer架构，拥有3140亿参数，能够处理8192个tokens的上下文长度；**2）**2024年3月17日，xAI宣布正式开源Grok-1模型，包括其权重和网络架构；**3）**2024年3月28日，xAI宣布**Grok-1.5**版本的开发，且该版本在推理能力和长文本能力（128K token，较Grok-1增加16倍）方面取得显著提升；**4）**2024年4月12日，xAI宣布**Grok-1.5V**多模态模型，Grok大模型在短时间内取得显著进展，正加速追赶第一梯队，并加入开源阵营，有望成为年内大模型黑马。

- **模型能力：**Grok-1.5V基于多方数据集与GPT-4V、Claude 3 Sonnet、Claude 3 Opus、Gemini Pro 1.5以上第一梯队多模态大模型进行对比，其中，**Grok-1.5V在数学能力、文本阅读能力和真实世界理解能力三方面超越同类模型**，在多学科推理、图表理解、文件处理能力方面略逊于其他模型，整体上看，Grok-1.5V可与以上第一梯队模型比肩。**未来，随着开源社区的参与和后续的技术迭代，Grok模型的潜力和应用范围有望进一步扩大。**

Grok-1.5V在多个基准测试中的得分

Benchmark		Grok-1.5V	GPT-4V	Claude 3 Sonnet	Claude 3 Opus	Gemini Pro 1.5
多学科推理能力	MMMU	53.6%	56.8%	53.1%	**59.4%**	58.5%
数学能力	Mathvista	**52.8%**	49.9%	47.9%	50.5%	52.1%
科学图表/示意图能力	AI2D	88.3%	78.2%	**88.7%**	88.1%	80.3%
文本阅读能力	TextVQA	**78.1%**	78.0%	-	-	73.5%
图表能力	ChartQA	76.1%	78.5%	81.1%	80.8%	**81.3%**
文件处理能力	DocVQA	85.6%	88.4%	**89.5%**	89.3%	86.5%
对真实世界的理解能力	RealWorldQA	**68.7%**	61.4%	51.9%	49.8%	67.5%

数据来源：X.ai官网，西南证券整理

2.3.3 Grok1.5V增强真实世界理解能力

- **真实世界理解能力强大，有望赋能自动驾驶多模态推理。**为评估多模态模型对基本现实世界空间的理解能力，Grok团队推出RealWorldQA新基准，当前，RealWorldQA数据集包含700多张图片，每张图片都对应一个问题和易于验证的答案，Grok-1.5V模型凭着其强大的多模态理解能力和真实世界理解能力，在RealWorldQA基准测试中，表现超过同类产品，取得68.7%的得分，高于GPT-4V的61.4%、Claude 3 Sonnet的51.9%和Gemini Pro 1.5的67.5%。在官网示例中，可以看到该数据集中包含众多现实世界案例，尤其在自动驾驶领域，**X.ai有望依托特斯拉的大量优质真实世界数据和成熟数据管线，构建能够更好地理解世界的多模态大模型**，较同类产品拉开更大差距。与此同时，**Grok多模态大模型的进展也有望反哺特斯拉旗下车型的智能座舱人机交互体验和FSD系统的多模态推理**，未来FSD V13有望具备理解语言Tokens的能力，加速大模型自动驾驶发展。

Grok-1.5V模型真实世界理解能力示例

Q：披萨刀和剪刀谁更大？

A. 披萨刀更大

B. 剪刀较大

C. 它们的大小大致相同

Q：从当前车道我们可以去哪里？

A. 左转　B. 直走

C. 左转并直行　D. 右转

Q：鉴于轿车的前置摄像头拍摄的画面，是否有足够的空间围绕前面的灰色汽车行驶？

A. 是的　B. 不

Q：根据图片，恐龙面向哪个基本方向？

A. 北　B. 南

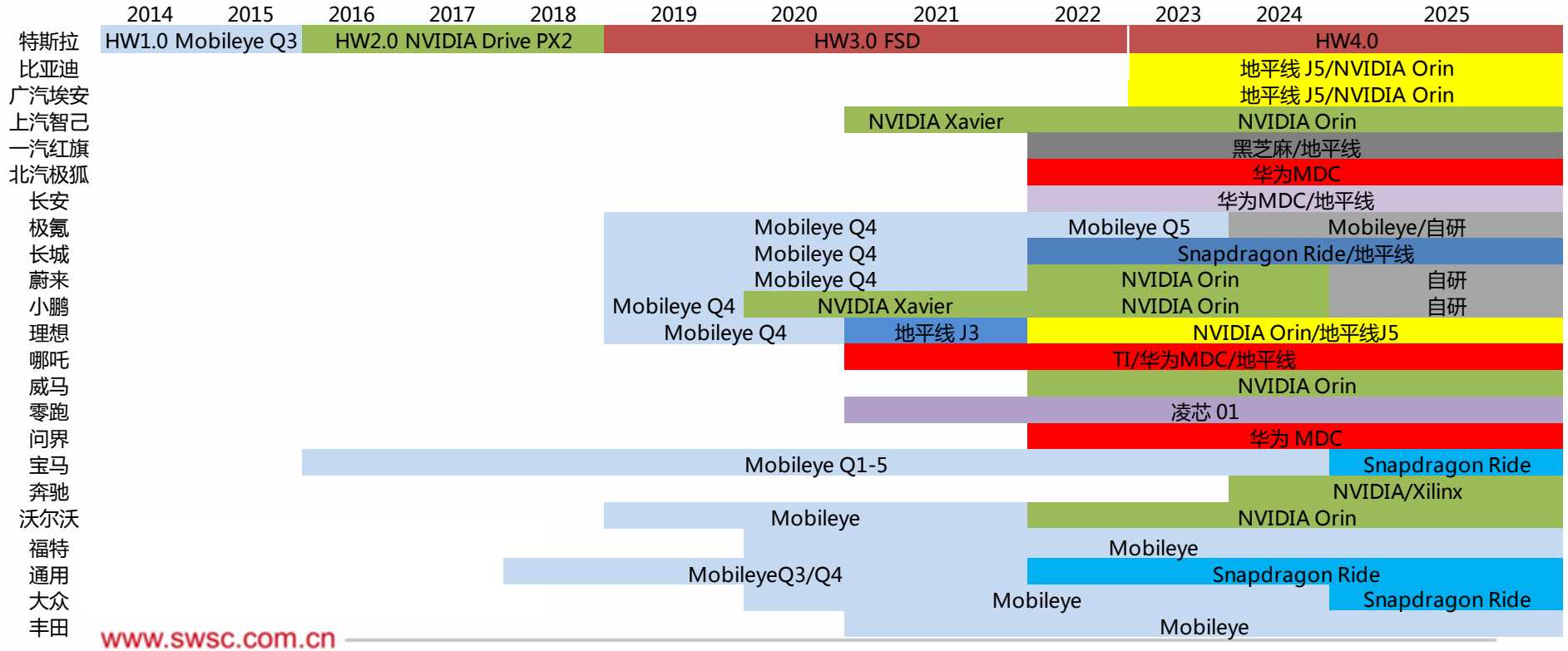
C. 东　D. 西

数据来源：X.ai官网，西南证券整理

2.3.3 大模型或同样倒逼智能汽车芯片革新

- 从各主机厂的自动驾驶的上车时间线看，2021年前Mobileye凭借软硬件黑盒集成的模式，帮助Tier1和主机厂降低研发成本和时间，提升经济效益，从而快速占据L2级别以下主要市场份额。
- 2022年成为明显转折点，伴随整车电子架构的更替，以及车企加大自研力度，以英伟达为代表的大算力平台开始规模化量产，高通、华为、地平线等“新秀”亦不断拿下整车厂定点，Mobileye也开始妥协采取更加开放的方案，当前赛道呈现出“一超多强”的格局。

各主机厂自动驾驶芯片量产时间线梳理

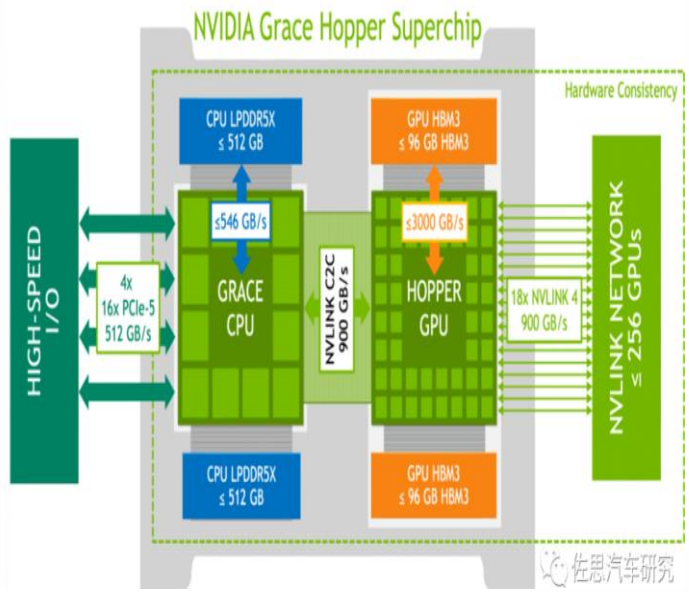


数据来源：各车企官网，车东西，佐思汽研，九章智驾，高工智能汽车等，西南证券整理

2.3.3 大模型或同样倒逼智能汽车芯片革新

- 再往后看，为适应大模型在车内的部署，叠加汽车EE架构向中央计算平台演进，汽车芯片有望针对性进行革新，行业格局有望迎来新一轮重塑。
- 英伟达于2022年发布下一代面向中央计算架构的汽车芯片Thor，预计从Ampere架构升级为Blackwell架构，最高可实现2000TOPS的算力，预计2023年Q4推出工程样片，2024年Q3开始量产。
- 高通同样于2022年发布Snapdragon Ride Flex，包括Mid、High、Premium三个级别。最高级的Ride Flex Premium SoC再加上外挂的AI加速器，其综合AI算力同样能够达到2000TOPS

英伟达下一代大算力汽车芯片Thor



高通下一代大算力汽车芯片Ride Flex

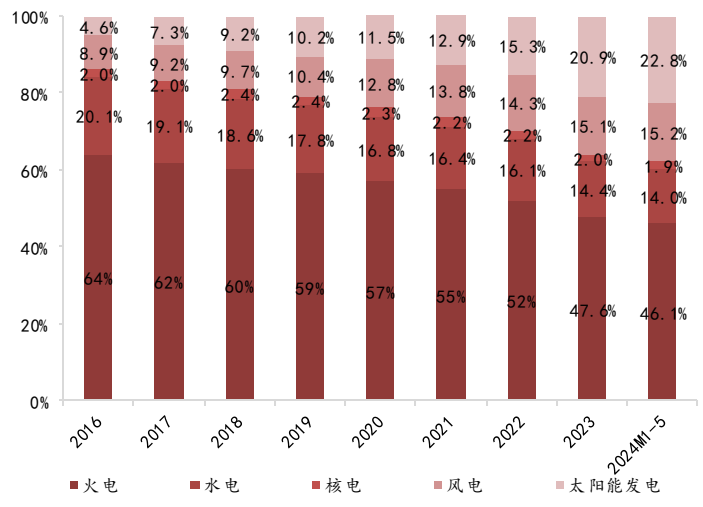
数据来源：英伟达，高通，西南证券整理

目 录

3.1 新能源消纳红线放开至90%，装机量将迎来快速增长

- **双碳背景下，风电和光伏发电逐步渗透成为国内主力能源。**截至2024年5月底，太阳能发电装机容量约6.9亿千瓦，占比从2016年末的4.6%提升至22.8%；风电装机容量约4.6亿千瓦，占比由2016年末的8.9%提升至15.2%，风光新能源装机容量占能源装机比重达38%。
- **95%的消纳红线下，新能源快速渗透消纳压力显著提升。**2018年10月，国家发布《清洁能源消纳行动计划（2018-2020年）》，首次明确要求自2018年起，要确保弃风、弃光电量连年下降，到2020年时，**光伏发电利用率要高于95%**。2024年1-5月，全国光伏发电利用率为96.7%，风电利用率为95.9%。但随着新能源渗透率的快速提高，新能源电力的消纳压力逐渐增加，尤其对于本地电力消费量较低、输送能力不足、缺乏调节手段的省份地区而言，仅靠现有资源维持较低的弃风弃光率、满足95%的新能源消纳红线压力较大。

新能源新增装机渗透率达38%



2024年1-5月风电消纳比例

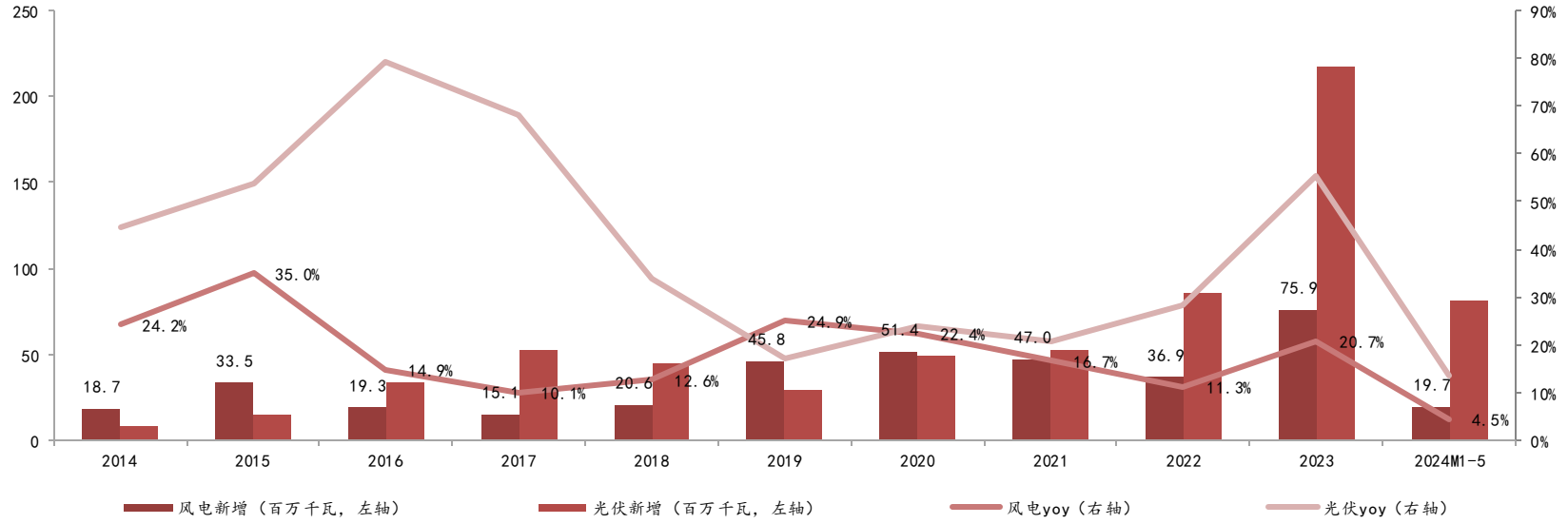
排名	地区	光伏利用率	风电利用率	排名	地区	光伏利用率	风电利用率
1	上海	100.0%	100.0%	18	蒙东	97.1%	92.0%
2	浙江	100.0%	100.0%	19	云南	96.8%	99.0%
3	福建	100.0%	100.0%	20	全国	96.7%	95.9%
4	重庆	100.0%	100.0%	21	宁夏	96.6%	98.0%
5	广西	100.0%	100.0%	22	河南	96.5%	94.8%
6	江苏	99.8%	99.9%	23	湖北	96.2%	98.8%
7	安徽	99.8%	99.9%	24	吉林	96.0%	91.2%
8	广东	99.8%	99.4%	25	陕西	95.2%	95.6%
9	四川	99.6%	100.0%	26	黑龙江	95.2%	94.0%
10	北京	99.5%	97.8%	27	辽宁	95.1%	92.8%
11	海南	99.4%	99.9%	28	新疆	95.0%	94.1%
12	湖南	99.0%	96.3%	29	河北	94.9%	93.2%
13	贵州	98.8%	99.4%	30	蒙西	93.5%	93.6%
14	天津	98.8%	99.3%	31	甘肃	91.6%	93.5%
15	山西	98.1%	98.7%	32	青海	91.3%	93.0%
16	江西	97.9%	99.5%	33	西藏	73.2%	97.4%
17	山东	97.8%	95.5%				

数据来源：国家能源局，公众号全国新能源消纳预警中心，西南证券整理

3.1 新能源消纳红线放开至90%，装机量将迎来快速增长

- **2024年5月23日，国务院关于印发《2024—2025年节能降碳行动方案》的通知**，文件提出：在保证经济性前提下，资源条件较好地区的新能源利用率可降低至90%。新能源消纳红线放宽，新能源装机有望迎来持续大规模发展。
- **随着新能源渗透比例持续提升，电力市场化改革和新型电力系统建设也在持续推进，以解决大规模新能源并网带来的消纳问题：**5月28日，国家能源局印发《关于做好新能源消纳工作保障新能源高质量发展的通知》，旨在解决新能源大规模发展的同时保持合理利用水平问题，主要提出4点任务：1）加快推进新能源配套电网项目建设；2）积极推进系统调节能力提升和网源协调发展；3）充分发挥电网资源配置平台作用；4）科学优化新能源利用率目标。

2014-2024M1~5风光新增装机及yoy



数据来源：国家能源局，西南证券整理

3.1 电力市场化持续推进，源网荷储相关建设有望加速

➢ **政策频发，新能源消纳倒逼电力市场化加速改革。**

◆ 2024年2月，国家发改委与能源局联合发布《关于建立健全电力辅助服务市场价格机制的通知》提出，持续推进电力辅助服务市场建设，科学确定辅助服务市场需求， 合理设置有偿辅助服务品种，规范辅助服务计价等市场规则，从国家层面统一建立健全电力辅助服务市场价格机制，相关政策自2024年3月1日起实施 。

◆ 2024年4月和5月，《电力市场监管办法》与《电力市场运行基本规则》接连落地，电力市场化改革加速推进。

电力市场化政策频发

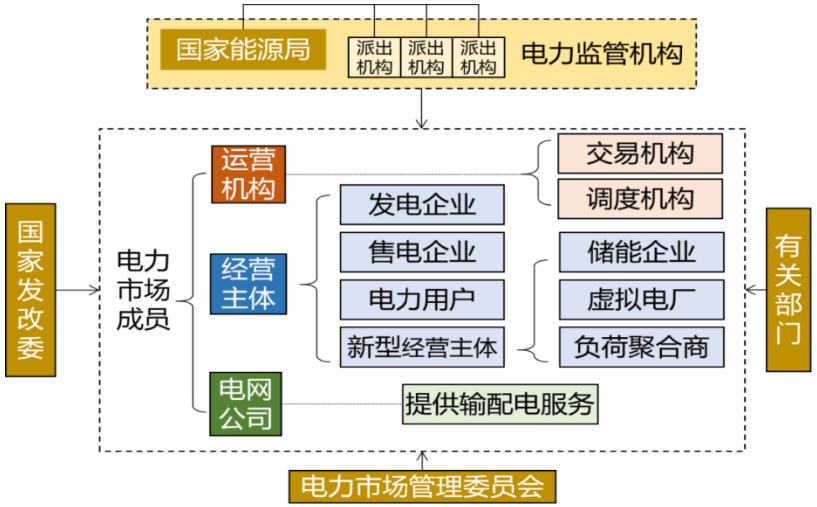
日期	部门	政策	主要内容
2022.1	国家发改委、国家能源局	《关于加快建设全国统一电力市场体系的指导意见》	2025年全国统一电力市场体系初步建成，2030年全国统一电力市场体系基本建成。
2022.3	国家发改委、国家能源局	《关于加快推进电力现货市场建设工作的通知》	支持具备条件的现货试点不间断运行，尽快形成长期稳定运行的现货市场。
2023.7	中央深改委	《关于深化电力体制改革加快构建新型电力系统的指导意见》	强调要深化电力体制改革，加快构建清洁低碳、安全充裕、经济高效、供需协同、灵活智能的新型电力系统。
2023.9	国家发改委、国家能源局	《电力现货市场基本规则(试行)》	在国家层面首次出台电力现货市场规则性文件，为推动电力现货市场从试点走向全国打好基础。
2023.10	国家发改委、国家能源局	《关于进一步加快电力现货市场建设工作的通知》	明确浙江2024年6月前启动现货市场连续结算试运行，辽宁、江苏、安徽等力争在2023年底前开展长周期结算试运行等，有序扩大现货市场建设范围。
2024.2	国家发改委、国家能源局	《关于建立健全电力辅助服务市场价格机制的通知》	加强辅助服务市场与电能量市场的有效衔接，通过采取合理设置有偿辅助服务品种，规范市场计价规则，完善价格形成机制，推动费用规范有序传导等措施，进一步完善我国电力辅助服务市场建设。
2024.4	国家发改委	《电力市场监管办法》	《办法》是强化电力市场成员行为监管，维护电力市场秩序的重要保障。此次办法修订进一步完善电力市场监管对象、调整了监管内容，有助于更好推进全国统一电力市场体系建设。
2024.5	国家发改委	《电力市场运行基本规则》	明确了包括经营主体、电力市场运营机构和提供输配电服务的电网企业等在内的市场成员范围，引入了电力中长期交易、电力现货交易、电力辅助服务交易、容量交易等新的交易类型。
2024.6	国家能源局	《电力市场注册基本规则(征求意见稿)》	统一电力市场注册机制，加强和规范电力市场注册工作，维护电力市场秩序和各类经营主体合法权益。

数据来源：国家发改委、国家能源局，西南证券整理

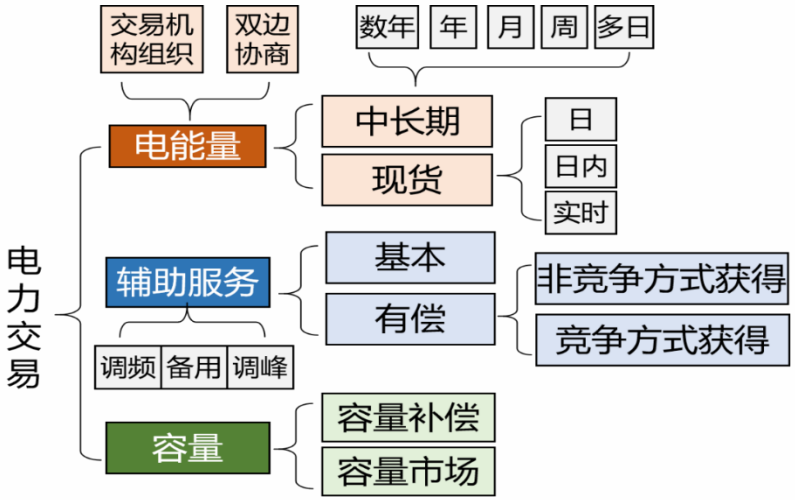
3.1 电力市场化持续推进，源网荷储相关建设有望加速

- **2024年7月1日，《电力市场运行基本规则》正式施行，2005年10月13日发布的《电力市场运营基本规则》（原国家电力监管委员会令第10号）同时废止。主要变化包括：**
- **1）经营主体扩展：**2024年的规则新增售电企业、电力用户和新型经营主体（含储能企业、虚拟电厂、负荷聚合商等）；
- **2）交易类型增加：**2024年的规则首次将“容量交易”纳入电力市场交易范畴，为新型储能和发电企业的成本回收提供新的途径；
- **3）进入与退出机制：**电力市场实行注册制度；
- **4）市场范围和运营机构：**推进全国统一电力市场建设，相比2005年的区域电力市场，2024年的规则更强调市场的统一性和完整性。

电力市场新规则下电力市场成员



电力市场新规则下交易类型和交易方式

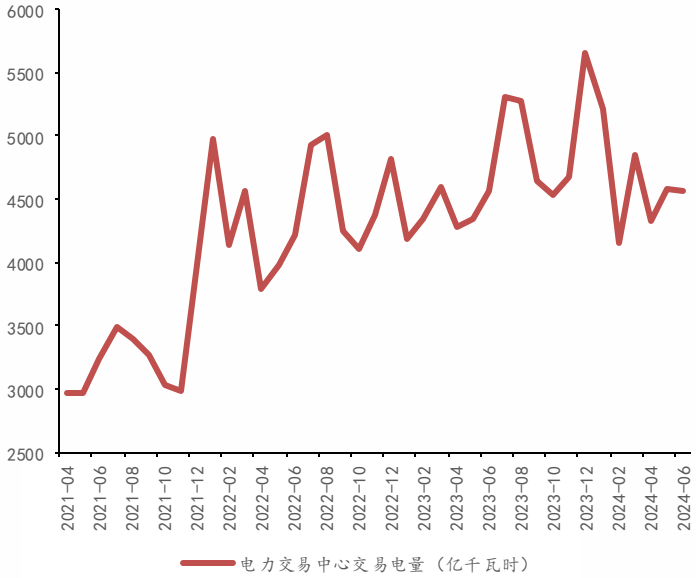


数据来源：国家发改委，国家能源局，西南证券整理

3.1 电力市场化持续推进，源网荷储相关建设有望加速

- **电力交易量持续增长，长周期运行试点城市逐步扩容。**随着电力市场体系不断完善，通过市场交易的电量持续增长，2024年6月占全社会用电比重为58.8%，建设电力现货市场是加快构建新型电力系统的重要举措。
- 截至2023年底，全国共有29个地区开展电力现货市场(试)运行：**1）第一批试点8个地区中**，山西、广东电力现货市场于23年年末先后转入正式运行，省间电力现货市场启动整年连续结算试运行；**2）第二批6个试点地区**，其中江苏、安徽、辽宁、湖北、河南这5个地区2023年共完成9次结算试运行，运行时间合计230天。**3）非试点地区**，2023年6月20日江西率先完成全国首个非试点地区电力现货市场结算试运行。

电力交易中心电力交易量呈现增长趋势



各地区电力现货市场运行进度

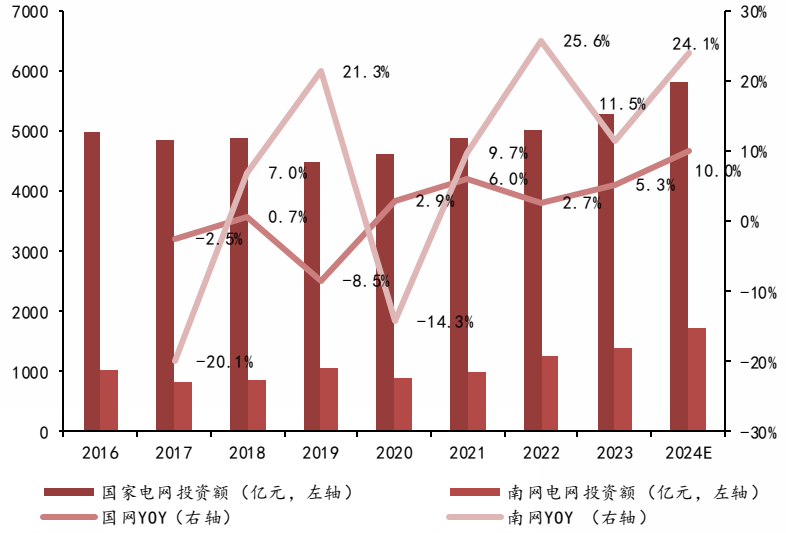
	地区	市场进展	运行时间		地区	市场进展	运行时间
第一批试点	南方（以广东为起步）	由连续结算试运行转入正式运行	自12月28日起	非试点地区	河北南部	结算试运行	分2次，共13天
	蒙西	连续结算试运行	全年		吉林	模拟试运行	20天
	浙江	调电试运行	分2次，共6天		黑龙江	调电试运行	1天
	山西	由连续结算试运行转入正式运行	自12月22日起		江西	结算试运行	7天
	山东	连续结算试运行	全年		湖南	结算试运行	3天
	福建	结算试运行	第一阶段全年，第二阶段15天		广西	结算试运行	1天
	四川	结算试运行	8个月（枯水期）		云南	结算试运行	1天
	甘肃	连续结算试运行	全年		陕西	结算试运行	7天
第二批试点	上海	调电试运行	分2次，共16天		青海	调电试运行	1天
	江苏	结算试运行	1个月		宁夏	结算试运行	3天
	安徽	结算试运行	分4次，共47天		新疆	调电试运行	分2次，共2天
	辽宁	结算试运行	分2次，共33天				
	河南	结算试运行	1个月				
	湖北	结算试运行	3个月				

数据来源：国家能源局，中电联，西南证券整理

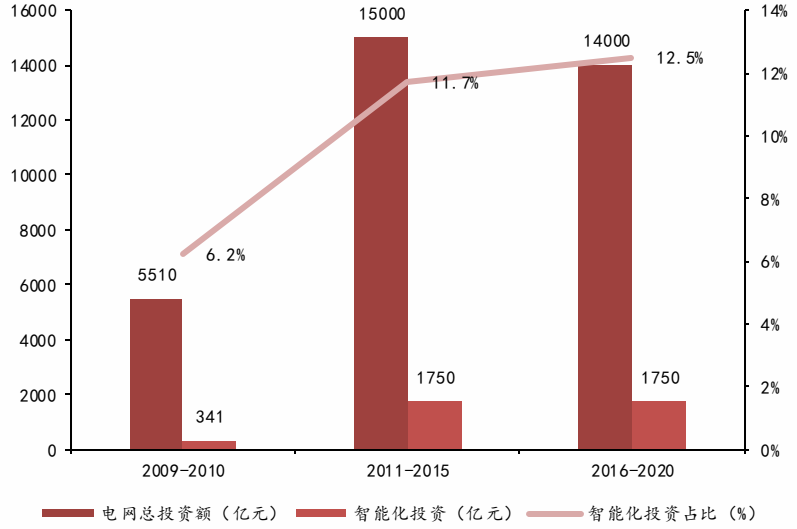
3.1 电力市场化持续推进，源网荷储相关建设有望加速

- **电网投资具备韧性，智能化投资占比稳健增长。**2023年，国家电网和南方电网投资额分别同比+5.3%/+11.5%至5275/1394亿元。从智能化投资占比来看，2016-2020年电网智能化投资占比为12.5%，较2009-2010年的6.2%增长明显。
- **电网投资重点从输电网走向配电网，重视配电网智能化产业链发展机会。**建立安全、高效的智能配电网，是提升新能源消纳比例，构建新型电力系统的关键环节。2024年3月1日，国家发改委、国家能源局发布《关于新形势下配电网高质量发展的指导意见》，助力构建经济高效、供需协同、灵活智能的新型电力系统。随着新能源渗透比例持续增长，后续配电网投资有望继续提升，**主要增加的投资方向包括：**1）配电网改造扩容；2）配网智能化相关提升运行灵活性。

2016-2024E国家电网和南方电网投资额



电网智能化投资占比逐步提升



数据来源：国家电网，南方电网，《国家电网智能化规划总报告》，西南证券整理

3.1 电力市场化持续推进，源网荷储相关建设有望加速

➢ **智能配电投资额规模测算：**

- **假设1：**“十四五”期间国网计划完成约2.4万亿投资，观察前三年情况有望超预期，国家电网预计2024年电网投资额有望超5000亿元，假设2024/2025年国网投资额分别为5100/5000亿元。
- **假设2：**“十四五”期间南网投资约6700亿元，对应到2024-2025年投资额同比增速年均约10%。
- **假设3**：根据国网规划，“十四五”期间配电网投资额需达到50%。由于“十四五”前期主干网特高压建设需求迫切，需解决新能源大基地消纳等问题，故配电侧投资比例较弱。伴随近年来特高压相关线路陆续核准开工投运，我们认为后续电网投资有望向配电网倾斜，预计2024-2025年配电网投资占比有望分别提升至45%/50%。
- **假设4:** 配电网智能化占比从“十三五”的10%提升至2021年的15%，伴随分布式新能源大规模接入，配用电智能化升价加快，预计2024-2025年智能电网规模占比分别为25%/30%。

	2016	2017	2018	2019	2020	2021	2022	2023	2024E	2025E
国网投资额（亿元）	4747.8	4853.7	4889.6	4473.0	4605.0	4882.0	5094.0	5275	5100	5000
南网投资额（亿元）	775.0	817.0	874.0	1060.0	907.0	995.0	1094.5	1375	1512.5	1663.75
配电网投资占比	20%	20%	20%	20%	20%	30%	35%	40%	45%	50%
配电网投资额（亿元）	1104.6	1134.1	1152.7	1106.6	1102.4	1763.1	2166.0	2660	2975.6	3331.9
配电网智能化投资占比	10%	10%	10%	10%	10%	15%	17%	20%	25%	30%
智能配网投资额（亿元）	110.5	113.4	115.3	110.7	110.2	264.5	368.2	532	743.9	999.6

数据来源：西南电新外发报告友讯达，国家电网，南方电网，西南证券整理

3.1 电力市场化持续推进，源网荷储相关建设有望加速

- **电力市场化改革和新型电力系统建设推进的背景下，主要受益板块：**

1）源：火电灵活性改造。火电灵活发电主要是指增加火电机组的出力变化范围，提升响应负荷变化或调度指令的能力。在新能源消纳压力下，火电灵活发电的价值日益提升：1）源端火电灵活性改造经济成本较低；2）未来随着容量电价机制的完善以及电力市场化的推进，运营商进行火电灵活性改造的积极性有望持续提升。

2）网：智能配电网。新型电力系统建设需要适应高比例新能源的配电网进行配套，后续配电网投资有望继续提升。后续配电网需要增加的投资方向包括：1）配电网改造扩容；2）配网智能化相关提升运行灵活性。

3）荷：虚拟电厂。虚拟电厂是将不同位置的可调负荷、储能、充电桩等可控资源聚合起来的一种智慧能源系统。2024年6月1日起正式施行的《电力市场监管办法》明确新增虚拟电厂作为电力交易主体，为可控负荷、新型储能、分布式新能源等灵活性资源提供进入市场的机会，充分激发和释放用户侧灵活调节能力，促进电力市场的多元化和效率提升。

4）储：储能。在储能设施中，目前使用最为广泛、成熟且经济的是抽水蓄能电站，但其对于地理条件要求较高，建设周期长，难以灵活布局。2021年4月，国家发展改革委、国家能源局提出：明确新型储能独立经营主体地位；健全新型储能价格机制；健全“新能源+储能”项目激励机制。2022年初，国家发展改革委、国家能源局联合印发《“十四五”新型储能发展实施方案》，明确到2025年，新型储能由商业化初期步入规模化发展阶段，具备大规模商业化应用条件。政策利好下，新型储能多元化高质量发展取得显著成效。据国家能源局的数据显示，截至2023年底，全国已经建成投运新型储能项目累计装机规模达3139万千瓦/6687万千瓦时，平均储能时长2.1小时。

- **相关标的：**科远智慧、国网信通、国能日新、朗新集团、南网科技等。

数据来源：国家能源局，人民网，西南证券整理

3.2 “低空”+“车路云”拉动新型基建，泛交通领域订单频现

➢ **低空经济顶层设计落地，进入战略发展阶段。**

- 2021年，低空经济被写入《国家综合立体交通网规划纲要》；2023年11月2日，国家空中交通管理委员会办公室发布了《中华人民共和国空域管理条例(征求意见稿)》，非管制空域权限放开；2023年12月，中央经济工作会议提出“打造生物制造、商业航天、低空经济等若干战略性新兴产业”；，2024年1月1日起《无人驾驶航空器飞行管理暂行条例》正式施行，我国无人驾驶航空器产业进入有法可依的规范化发展新阶段。
- 2024年3月，低空经济被首次写入《政府工作报告》，将低空经济定位新兴产业和未来产业，打造成为经济发展新增长引擎；同月，工信部联合三部门发布《通用航空装备创新应用实施方案（2024-2030年）》，提出到2027年，通用航空公共服务装备体系基本完善，以无人化、电动化、智能化为技术特征的新型通用航空装备在城市空运、物流配送、应急救援等领域实现商业应用。到2030年，通用航空装备全面融入人民生产生活各领域，成为低空经济增长的强大推动力，形成万亿级市场规模。
- 当前阶段，近20个省份政府工作报告中提及低空经济，各地低空经济相关的发展条例、实施方案、行动计划等陆续出台，对基础设施建设、应用场景拓展、产业培育等方面提出具体目标和规划。

各地区低空经济相关政策

时间	地区	政策	主要内容
2024.2	苏州	《苏州市低空经济高质量发展实施方案（2024-2026年）》	到2026年，力争聚集产业链相关企业500家，产业规模达600亿元。建成1～2个通用机场和200个以上垂直起降点，开通至周边机场3～5条通用航空短途运输航线、100条以上无人机航线，无人机商业飞行取得突破性进展
2024.3	深圳市龙华区	《龙华区低空经济试验区2024年度建设方案》	到2024年底，初步建成低空经济先导区。建设一批地面基础设施，新增40个以上低空飞行器起降平台及末端配送设施。力争开通35条以上区内无人机航线，载货无人机商业飞行突破30万架次/年。引导10个以上低空产业项目在龙华落地，加快产业集聚发展。
2024.4	安徽	《安徽省加快培育发展低空经济实施方案（2024-2027年）及若干措施》	到2025年，低空经济规模力争达到600亿元，规模以上企业达到180家左右，其中，培育生态主导型企业1—2家。到2027年，低空经济规模力争达到800亿元，规模以上企业力争达到240家左右，其中，生态主导型企业3—5家。
2024.5	南京	《南京市促进低空经济高质量发展实施方案（2024-2026年）》	全市低空经济产业规模超500亿元。建成240个以上低空航空器起降场及配套的信息化基础设施；建成3个以上试飞测试场和操控员培训点；规划建设1—2个通用机场；开通120条以上低空航线；全市低空经济领域高新技术企业超120家；建成15个省级以上创新平台；培育30个以上具备示范效应的创新应用场景

数据来源：中国政府网，民航局，交通运输部，各地政府官网，西南证券整理

3.2 “低空”+“车路云”拉动新型基建，泛交通领域订单频现

- 车路云一体化历经多年布局，进入大规模试点阶段。
- 2020年，发改委等11部门印发《智能汽车创新发展战略》，提出构建先进完备的智能汽车基础设施体系，推进智能化道路基础设施规划建设；2021年3月，车路协同被列入“十四五规划”。
- 2022-2023年，期间一系列配套政策接连推出，进一步规范完善车联网建设标准体系，开展准入试点，推动智能网联汽车与新能源、智能交通、智慧城市等的融合发展。
- 2024年1月，工信部等五部委联合发布了《智能网联汽车车路云一体化应用试点工作的通知（2024-2026年）》，部署开展智能网联汽车“车路云一体化”应用试点工作，将建成一批架构相同、标准统一、业务互通、安全可靠的城市级应用试点项目，标志着“车路云一体化”正式进入大规模铺开试点阶段。

各部门车路云一体化相关重要文件政策梳理

时间	部门	政策	主要内容
2020.2	发改委等11部门	《智能汽车创新发展战略》	提出构建先进完备的智能汽车基础设施体系，推进智能化道路基础设施规划建设、建设广泛覆盖的车用无线通信网络、建设覆盖全国的车用高精度时空基准服务能力、建设覆盖全国路网的道路交通地理信息系统、建设国家智能汽车大数据云控基础平台等成为重点建设任务
2023.7	工信部	《国家车联网产业标准体系建设指南(智能网联汽车)(2023版)》	到 2025 年，系统形成能够支撑组合驾驶辅助和自动驾驶通用功能的智能网联汽车标准体系；到 2030 年，全面形成能够支撑实现单车智能和网联赋能协同发展的智能网联汽车标准体系。
2024.1	工信部	《关于开展智能网联汽车“车路云一体化”应用试点工作的通知》	推动智能化路侧基础设施和云控基础平台建设，提升车载终端装配率，开展智能网联汽车“车路云一体化”系统架构设计和多种场景应用，形成统一的车路协同技术标准与测试评价体系，健全道路交通安全保障能力，促进规模化示范应用和新型商业模式探索，大力推动智能网联汽车产业化发展。
2024.5	财政部、交通运输部	《关于支持引导公路水路交通运输基础设施数字化转型升级的通知》	自2024年起，在3年左右时间支持30个左右的示范区域，打造一批线网一体化的示范通道及网络，力争推动85%左右的繁忙国家高速公路、25%左右的繁忙普通国道和70%左右的重要国家高等级航道实现数字化转型升级。在智慧扩容方面实现示范通道通行效率提升20%左右;在安全增效方面实现突发事件应急响应效率提升30%左右;在融合创新方面凝练总结一批具有较高推广价值的应用场景和关键技术等。

数据来源：发改委，工信部，交通运输部，西南证券整理

3.2 “低空”+“车路云”拉动新型基建，泛交通领域订单频现

- 当前“低空经济”和“车路云一体化”均迈入从政策发布-订单落地转化的阶段，基础设施建设为体量较大、前期重要性较高的环节，**建议关注“硬基建”的通导监设备和路端设备，“软基建”的云控平台和飞行服务平台及空管系统。**
- 相关标的：莱斯信息、纳瑞雷达、深城交、千方科技、万集科技等。

低空经济相关订单

时间	地区	项目	金额	进展	主要内容
2024.1	安徽	安徽省新技术融合应用低空飞行服务平台	955.43万元	开工	新技术融合应用低空飞行服务平台一套，包含系统设计、硬件设备购置、开发、调试、集成测试、安装部署、培训、售后和运维服务
2024.1	无锡	无锡市低空经济发展规划及实施方案（二次）	298万元	中标	编制无锡市低空航空器起降设施布局规划及低空空域精细化划设方案，形成无锡低空经济发展“1+1+1+N”指导文件，助力无锡抢占全球低空经济科技创新和产业发展高地
2024.4	四川	低空智联网示范项目	377.54万元	开工	本次试点总体按照“1+N”思路开展项目建设，即1套低空基础支撑设施，N个示范应用，其中低空基础支撑设施中，新建试点范围空域图，并在现有办公场地基础上升级改造，建成示范区低空管控服务中心，在示范应用方面，本次重点开展试点范围无人机综合巡检示范应用和西南财经大学天府学院成都校区（东区）物流配送示范应用。
2024.6	深圳	南山区低空航空器起降设施布局规划及低空空域精细化设方案	202.5万元	中标	南山区低空航空器起降设施布局规划及低空空域精细化划设方案

车路云相关订单

时间	地区	项目	金额	进展	主要内容
2024.4	鄂尔多斯	鄂尔多斯市新能源智能网联汽车车路云一体化应用示范项目	1.05亿元	中标	覆盖康巴什核心区以及康巴什北区，新建智慧化路口数量36个，新建智慧化路段点位49个，道路单向总里程约30公里
2024.5	北京	北京市车路云一体化新型基础设施建设项目	99.39亿元	招标	在通州区、顺义区、朝阳区、昌平区、密云区、怀柔区、海淀区、石景山区、丰台区、门头沟区、房山区、大兴区、亦庄经开区共选取2324平方公里范围内约6050个道路路口开展建设，以及除上述道路路口外本项目双智专网网络中心的建设和改造。
2024.6	武汉	武汉市智能网联新能源汽车“车路云”一体化重大示范项目	170.84亿元	审批完成	建设全市统一的智能网联汽车服务平台、1.5万个智慧泊位、5.578km智慧道路（经开区）改造，16万方智能网联汽车产业研发基地（东湖高新区）、车规级芯片产业园、无人驾驶产业园。推动城市级智慧道路覆盖率及车载终端装配率的显著提升。

数据来源：各地政府官网，招标网，西南证券整理

3.3 谋划新一轮税改，财税IT迎景气上行周期

- **新一轮税改方向有望明确，财税IT将迎来景气上行周期。**
  - 2023年12月，中央经济工作会议明确提出“深化重点领域改革”，要“谋划新一轮财税体制改革，落实金融体制改革”。
  - 2024年3月《关于2023年中央和地方预算执行情况与2024年中央和地方预算草案的报告》中提到要“坚持目标导向、问题导向，谋划新一轮财税体制改革，建立健全与中国式现代化相适应的现代财政制度”；《报告》还提及优化税制结构、完善财政转移支付体系、稳步推进省以下财政体制改革等举措。

新一轮税改潜在重点方向

	新一轮税改重点可能方向	可能措施
收入端	优化税制结构	消费税征收从生产环节后移至消费环节并稳步下划地方
		个税税率调整
		适当扩大综合所得征税范围
		扩大增值税抵扣范围
支出端	完善预算管理制度	深化预算绩效管理、完善基本支出标准、完善国债收益率曲线、完善财政资金直达机制、推动央地财政系统信息贯通、加强预决算公开等方式
	进一步理清央地财政关系	中央和地方财政始终存在着财权和事权不匹配的问题，2022年地方财政收入占总财政收入的比重为53.4%，然而支出占比却高达86.4%。近年来，事权上移的趋势逐渐明朗；未来财政体制改革的方向可能仍是强化上级事权，减少共同事权，并避免“层层下压”的现象，中央政府加杠杆、地方政府压杠杆的趋势有望进一步持续

数据来源：人民网，西南证券整理

3.3 谋划新一轮税改，财税IT迎景气上行周期

- **税务线：**
  - **新一轮税改：1）G端，**优化税制结构带来相关系统更新改造需求；**2）B端，**税收“颗粒归仓，应收尽收”背景下，伴随税务信息化和智能化逐步推进，税务信息企业对合规税优、风险控制等功能需求持续提升。
  - **金税四期：**税务信息化当前建设重点是21年启动的金税四期，目前金税四期在省级已基本上线，市场格局基本稳固。金税四期的全面推广核心抓手在于数电票，数电票推动预算单位和企业财务处理流程从开具、报销、入账、档案、存储等环节全部电子化，目前仍处于试点推广阶段。随着数电票全面开放以及金税四期相关系统建设完善，合规税优、风险控制等需求未来将迎来增长。

- **相关标的：**税友股份、中软国际、航天信息、百望股份等。

数据来源：税友股份投资者交流，西南证券整理

3.3 谋划新一轮税改，财税IT迎景气上行周期

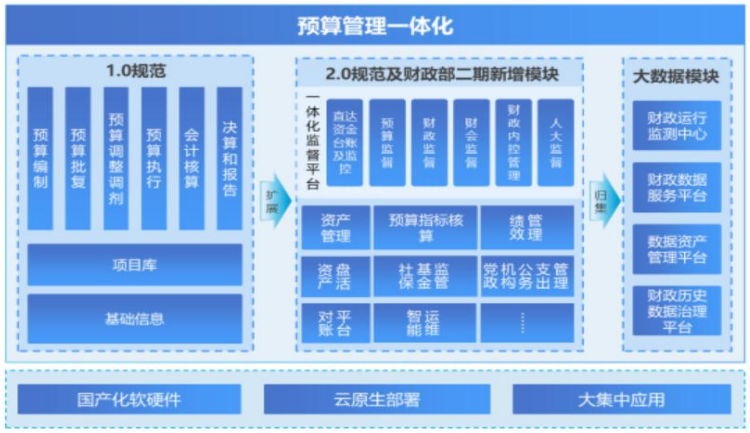
➢ **财政线：**

◆ **新一轮税改：**本轮税改支出端改造潜在重点之一是完善预算管理制度，包括但不限于深化预算绩效管理、完善基本支出标准、完善国债收益率曲线、完善财政资金直达机制、推动央地财政系统信息贯通、加强预决算公开等方式。实现央地财政系统信息贯通途径是通过预算管理一体化系统及预算单位运营平台实现转移支付的监控和规范，预计新一轮税改将带来财政部和预算单位对转移支付的监督管理等模块需求增加。

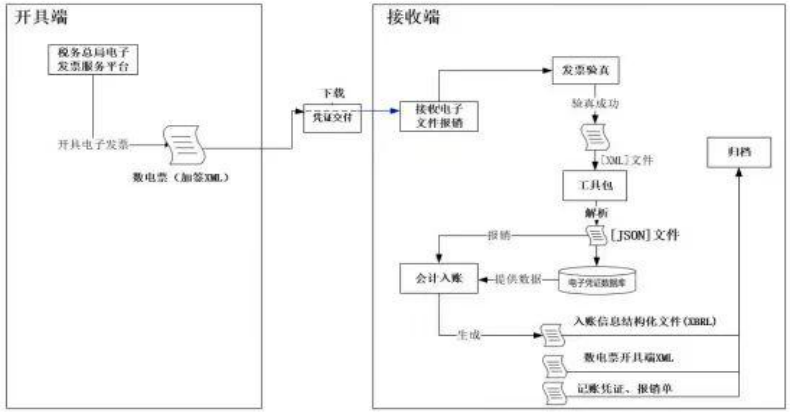
◆ **电子凭证推广：**2023年4月，财政部联合有关部门开展电子凭证会计数据标准制定试点推广工作，旨在通过统一技术规范、统一结构化数据标准，提升财务管理水平和效率；推动会计工作向数字化、标准化方向发展，为国家经济管理和决策提供可靠的数据支持。电子凭证推广的基础下，财政部可通过通过电子票据实现对预算单位支出的监控跟踪。新背景下，电子凭证接收端改造有望迎来加速推进。

➢ **相关标的：**博思软件、中科江南等。

预算管理一体化2.0提升财政监督效率



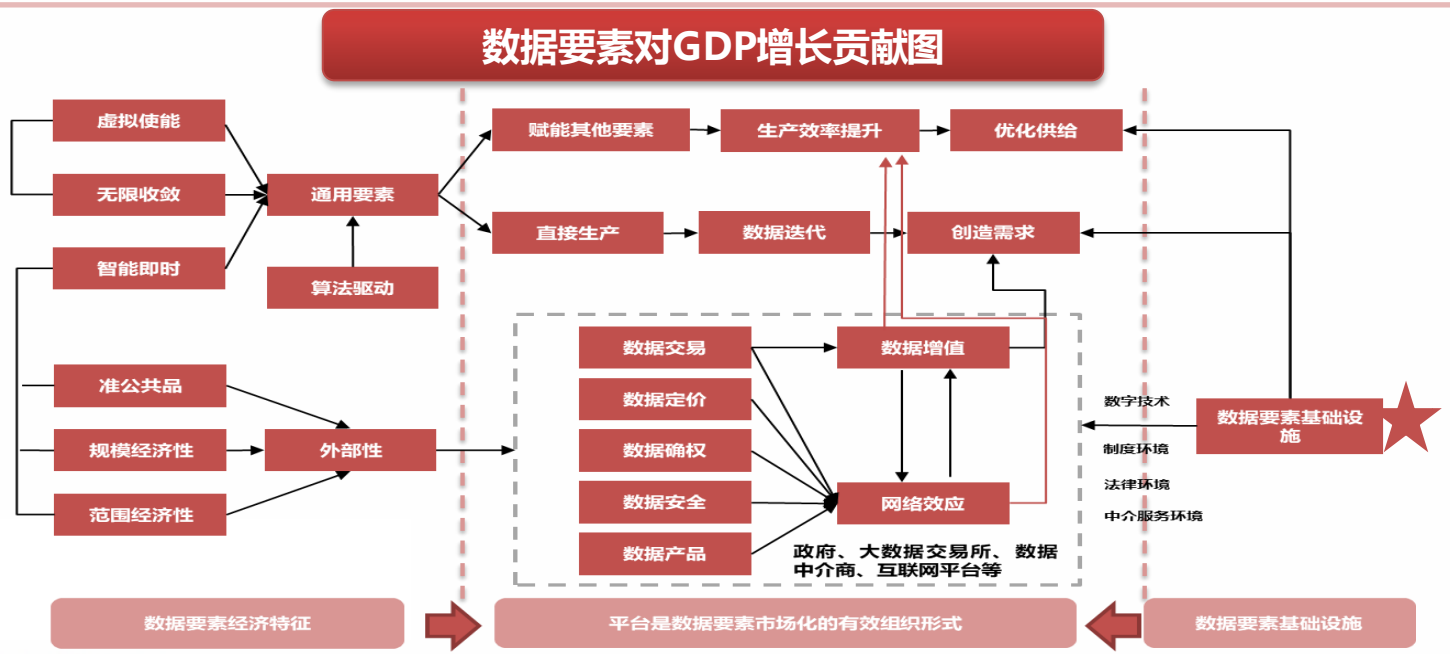
电子凭证的开具和接收流程图



数据来源：财政部，博思软件投资者交流，西南证券整理

3.4 数据要素大厦即将落成，亟待制度文件封顶

- 国家数据局局长刘烈宏在全球数字经济大会上表示，**国家数据局2024年将以制度建设为主线，今年陆续推出数据产权、数据流通、收益分配、安全治理、公共数据开发利用、企业数据开发利用、数字经济高质量发展、数据基础设施建设指引等8项制度文件。**
- 我们认为，七月会后国家数据局即将发布的数据产权、数据流通、收益分配等系列制度文件有望为确权、定价、流通体系制度不完善等关键堵点难点做出规范，通过多场景、多主体的复用，加速产业落地节奏，**产业链有望迎来密集政策催化。**
- 相关标的：1）地方公共数据运营：广电运通、云赛智联、太极股份等；2）医保数据运营：国新健康、久远银海、博思软件等；3）数据资产入表：拓尔思、卓创资讯、上海钢联等。



数据来源：《中国数据要素发展报告》，西南证券整理

3.4 因地制宜，积极探索公共数据开放运营

- 当前地方性公共数据授权的运营模式，初步可分为特许开发模式、独家经营模式、市场供应链模式、主题牌照模式等。
  - 特许开发：以杭州为代表，将数据运营权授权给特定的加工利用方
  - 独家经营：以青岛为代表，运营单位负责建设唯一公共数据运营平台
  - 市场供应链：以长沙为代表，拆解运营与加工职能，构建多级授权
  - 主题牌照：以北京为代表，按不同主题（领域、区域、综合等）授予牌照

四类模式对比分析

	特许开发模式	独家经营模式	市场供应链模式	主题牌照模式
代表城市	杭州	青岛	长沙	北京
授权方	市数据资源管理局	市大数据主管部门	市数据资源局	非统一授权方，经北京市人民政府审定的公共数据专区授权运营申请单位
运营方	加工使用主体	运营单位	数据运营主体，并额外单独列示数据加工主体	专区运营单位
最终服务对象	泛社会	公共数据应用单位，需具备一定数据安全保障能力有明确的应用场景需求	数据使用方（未明确定义）	按专区分类，有不同的目的和对象侧重

数据来源：国家法律法规数据库，开放数据中国，西南证券整理

3.4 地方成立数据集团，探索“管运分离”

- 同时，上海、河南、湖北等地开始纷纷成立数据集团，探索“管运分离”模式：政府作为公共数据持有方和授权方，引入数据集团国有企业作为公共数据的运营方，开展数据资产运营、数字产业投资和数据交易服务等工作，有望与产业界开展积极合作，完善数据的供给、配置以及市场化的开发利用。

地方数据集团	成立时间	注册资本
省级平台		
福建省大数据集团	2022年8月	100亿元
上海数据集团	2022年9月	50亿元
河南数据集团	2023年1月	10亿元
湖北数据集团	2023年6月	50亿元
数字湖南	2023年6月	5亿元
数字新疆集团	2023年8月	15亿元
市级平台		
苏州市大数据集团	2022年9月	20亿元
福州数据集团	2022年11月	10亿元
无锡大数据集团	2023年4月	2.9亿元
武汉数据集团	2023年5月	20亿元
拉萨数字经济产业集团	2023年5月	1亿元
成都数据集团	2023年7月	40亿元
扬州大数据集团	2023年9月	1亿元
央企平台		
中国电子数据产业集团	2023年4月	30亿元

数据来源：各省市政府官网，企查查，西南证券整理

3.4 数据变现案例涌现，商业模式逐步清晰

- 数据资产化层面，当前国内各主体对融资、证券化等创新模式做出积极尝试。
  - 2022年10月，全国首笔千万元**数据资产质押融资贷款**，佳华科技用其两个大气环境质量检测的数据资产做质押，获得北京银行城市副中心分行1000万元融资贷款。
  - 2023年3月，全国首笔**无质押数据资产增信贷款**，微言科技凭借在深圳数据交易所上架的标的，获得光大银行深圳分行的无质押数据资产增信贷款1000万元。
  - 2023年4月，全国首单**数字资产保险**，由中国人保西安分公司承保，为10家企业的数字资产提供了总额为1000万元的报账。
  - 2023年7月，全国首单**数据信托产品交易**，广西电网作为委托人，以信托形式将部分电力数据委托给中航信托，广西电网能源科技公司作为共同受托人对数据产品进行专业开发。
  - 2023年7月，全国首个**数据知识产权证券**化产品，杭州高新金控以12家企业的145件知识产权作为质押物，帮助企业获得定向资产支持票据（ABN）证券化融资1.02亿元，票面利率2.80%，发行期限358天。
  - 2023年8月，全国首例**数据资产作价入股**，青岛海通智能科技研究院等三家公司进行数据资产作价入股的签约仪式。

数据来源：数据要素社，西南证券整理

3.4 数据变现案例涌现，商业模式逐步清晰

- 数据交易层面，当前场内交易活跃度持续攀升。
- 以上海数据交易所为例，2021年11月成立当日，挂牌产品20个，当月交易额仅为30万元；2022年全年，挂牌产品超过800个，交易额突破1亿元；2023年，挂牌产品突破2000个，8月单月交易额突破1亿元，全年交易额有望突破10亿元。
- 对比2022年，详情还是不可查询的状态，当前的挂牌产品已详细列示内容说明、使用说明、来源描述、资质信息、产品价格等，数据交易的规范性得到大幅度提升。

数据产品示例

收钱吧

商家商户关键特征

通过输入实际经营者信息，查询商户的关键经营数据、商户可持续性数据等。

产品挂牌代码：1239626802-DBKⅢCQ6301

基本信息

产品名称：商家商户关键特征

系列名称：商户和经营统计类数据

供方名称：上海收钱吧互联网科技股份有限公司

应用板块：金融,综合,贸易

数据主题：商户属性、经营

产品类型：数据服务

产品描述：通过输入实际经营者信息，查询商户的关键经营数据、商户可持续性数据等。

关键词：经营，小微，流水

更新频率：动态(每日更新)

覆盖范围：全国31个省、自治区、直辖市

存储大小：193 MB

增量存储大小：10 MB/日

底层数据维度：244

内容说明

输入字段：

序号	字段名称	字段描述	示例值
1	md5处理的身份证号	参数身份证号、手机号和营业执照号至少传一个	42b62cd8067****ac1b9d92158479a6
2	md5处理的手机号	md5和sha256处理的手机号选择传一个，无手机号可皆不传	eed70585e64373****f9d82b4dbcb5cf
3	sha256处理的手机号	md5和sha256处理的手机号选择传一个，无手机号可皆不传	a916c3160c9073****76b2b8161b12b 34b8d7ad7e191a3e16a61f87ed2ac0cd4
4	营业执照号（统一社会信用代码）	参数身份证号、手机号和营业执照号至少传一个	913607****35JJBTX5
5	用户是否授权	0:已授权 1:未授权，非必填	0
6	场景	01：金融 02：消费 03：地图	02：消费

输出字段：

序号	字段名称	字段描述	示例值
1	商户id	商户唯一id	39d00000-0000-****-a906-4ed3800004b0
2	入网日期	入网日期	2018/8/15 19:44
3	商户短期流店概率	商户近期交易降低到极低水平的概率模型结果	0.42
4	商户经营评价	商户经营评价	58
5	反广告客户短期可持续性指数	可用于协助短期可持续性评分，0-700，值越大，短期可持续性越好	450
6	广告客户长期可持续性指数	可用于协助可持续性评分，0-700，值越大，可持续性越好	450
7	广告价值客户类别	可用于协助商户可持续性评级，1-10，值越小，客户可持续性越好	8

数据来源：上海数据交易所，西南证券整理

3.4 数据变现案例涌现，商业模式逐步清晰

- 2023年12月1日，北京国际大数据交易所正式上线“数据授权平台-微信小程序版”并开启公开测试，在探索数据收益分配机制的道路上更近一步。
- 通过在平台上实名注册，每个主体都能了解自己名下的数据目录，并通过对数据使用的逐笔逐场景授权操作，替代当下普遍存在的“一揽子授权”现象。
- 通过该授权操作，企业及个人都可以获得一定比例的数据交易收益。

数据授权平台主页

可授权信息类型

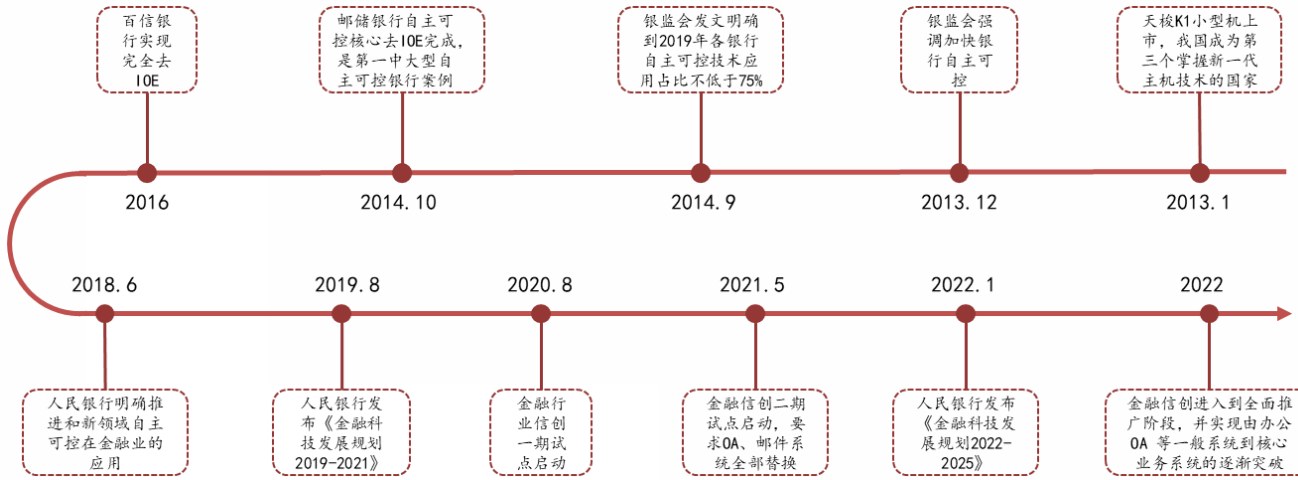
可授权使用场景

数据来源：北京国际大数据交易所，西南证券整理

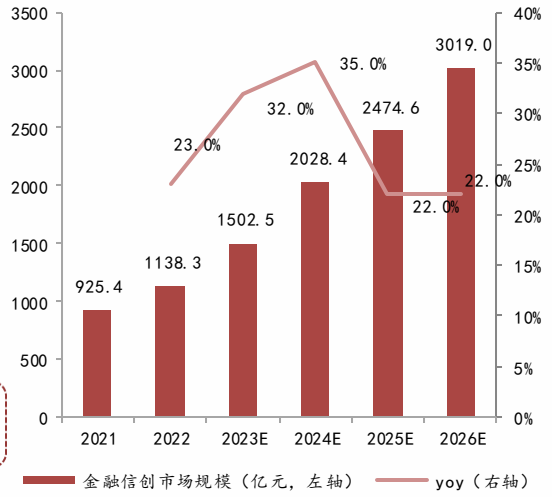
3.5 金融信创加速推进，证券新一代核心迎来落地高峰期

- **金融信创深度进一步扩大，市场规模快速增长。**金融信创以“先试点，后全面”的推广方式进行，2020年共有47家试点机构，主要包括头部银行、保险、证券、一行两会和交易所，项目多以OA办公系统为主，要求信创基础软硬件采购额占IT外采的5-8%；2021年试点机构扩容至198家，范围扩展到终端机具+管理软件+一般业务+关键业务，信创基础软硬件在IT外采中占比要求提高至10-15%；2022年进入规模推广阶段，广度、深度都进一步扩大，实现从办公OA等一般系统到核心业务系统的逐渐突破，要求信创软硬件投入占行业全年IT支出的30%。
- 基于金融信创有望进入全面推广阶段的判断，我们认为金融信创基础软硬件以及上层应用解决方案将迎来高景气发展。从市场规模来看，2022年我国金融信创行业市场规模为1138.3亿元，预计2026年将增长至3019.0亿元，2022-2026年均复合增长率为26.7%。

金融信创发展历程



2021-2026E金融信创市场规模



数据来源：壹零智库，亿欧，西南证券整理

3.5 金融信创加速推进，证券新一代核心迎来落地高峰期

- **金融信创加速，推动行业分布式架构转型进程。**近年来，以IBM服务器、Oracle数据库、EMC储存为基础的IOE集中式架构，逐渐难以适应高并发、高迁移性、高兼容性的数字金融新业态。以x86计算机和分布式数据库搭建的分布式方案，凭借其良好的拓展性、低廉的边际成本以及强大的数据处理能力逐渐成为金融IT的发展转型方向。
- 从技术上来看，相比集中式架构对于海外软硬件厂商的依赖，分布式架构在硬件层面能够实现国产化替代（普通X86服务器，各类国产自研的分布式数据库）。因此，分布式作为金融IT新一轮周期的技术支撑，与信创产业的国家战略共同驱动金融机构数字化转型升级和国产化进程加速，推动新一轮金融IT市场空间释放。

分布式架构性能显著提升

	集中式架构	分布式架构	提升
技术特点	大型机/小型机+数据库+集中存储	标准服务器+高带宽低时延网络	垂直升级变为水平扩展
交易速度	几十~几百毫秒	几~几十微妙	快1000+倍
处理能力	几千~几万笔/秒	几万~几十万笔/秒	提高10+倍
可靠性/可用性	单活高可用，分钟级切换，切换过程中可能有数据丢失	双活高可用，秒级切换，切换过程中零数据丢失	大幅提高
技术来源/可控性	单一设备供应商，进口封闭平台	多个设备供应商，开放平台，国产化成为可能	自主可控能力大幅提高
硬件成本	进口主机价格昂贵	标准服务器价格低廉	大幅降低2/3

数据来源：华锐技术，西南证券整理

3.5 金融信创加速推进，证券新一代核心迎来落地高峰期

- **从金融核心业务系统信创进度来看：**
- 1）银行核心业务系统信创推进较快：六大国有行最早开始银行业核心业务系统的信创替换，根据产业信息，截至目前六大国有行和部分股份制银行已基本实现核心业务系统信创版本的建设；
- 2）证券业新一代核心业务系统正处于信创密集建设期：政策要求关键单位27年做完整的信创切换，部分头部券商要时间节点要求更早（25年），考虑到出方案、招投标、落地实施周期和共存阶段，整体需要1-2年，大型券商在23-24年陆续开始核心业务系统的招投标。
- **相关标的：**顶点软件、恒生电子、金证股份等。

证券核心交易系统信创进展

供应商	核心业务系统信创进展
恒生电子	截至2023年底，公司所有主产品已经完成信创适配，助力70多家金融机构实现核心业务系统自主创新升级;UF3.0在深圳头部券商完成交易信创全链路试点上线工作，进入信创批量上线阶段,内存两融交易系统的试点上线，为全面实现做好了技术准备;O45完成全栈信创研发，新签53家新客户，完成24家客户竣工，其中华宝基金O45项目等关键项目的竣工，为全面批量交付做好了准备;PB信创在头部券商上线竣工。
顶点软件	2023年在东吴证券A5完成了100%国产化设备与基础系统软件的国产化替代，成为业内首个信创化单轨运行，100%完成国产替代的案例。
金证股份	2023年公司全面推广分布式、低时延的信创版新一代证券综合业务平台 FS2.0，推动在券商客户端逐步上线：新一代证券业务综合服务平台FS2.0订单系统在平安证券成功上线，完成全部客户迁移；FS2.0信创版底座在中信建投、中金财富证券投产；FS2.0信创版认证系统在广发证券部署
华锐技术	22年在国君实现新一代核心业务系统全业务上线

数据来源：各公司公告，西南证券整理

目 录

科远智慧（002380）：火电国产改造高景气，订单加速进行时

□ 投资逻辑：

1）火电DCS国产化改造大年，公司为行业龙头深度受益：电力生产作为一个国家的重要经济命脉，DCS作为火电厂的“大脑”，自主可控迫在眉睫。公司作为国内DCS的领先厂商，在百万机组、燃机控制系统等领域实现行业国产化首突破，标杆项目示范效应明显，有望深度受益本轮国产化替代周期。**2）重点项目捷报频传，订单加速进行时：**2024年以来，公司在陆续中标多个百万机组DCS系统自主可控改造项目以及新建扩建机组自主可控DCS系统项目，订单呈现进一步加速趋势。**3）轻装上阵成效显著，盈利重回上行通道：**公司2022年将生物质发电项目进行停产处置后轻装上阵，同时员工人数保持相对稳定，人效提升明显，利润弹性有望快速释放。

□ 业绩预测与投资建议：

预计公司2024-2026年归母净利润分别为2.4亿元、3.1亿元、4.0亿元，对应PE分别为18倍、14倍、11倍，给予“买入”评级。

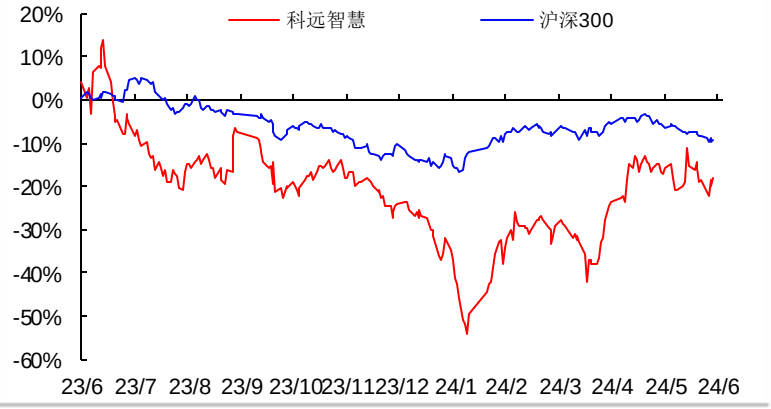
□ 风险提示：

政策推进力度不及预期、行业竞争加剧等风险。

业绩预测和估值指标

指标	2023A	2024E	2025E	2026E
营业收入（百万元）	1407.10	1900.03	2444.14	3048.28
营业收入增长率	21.90%	35.03%	28.64%	24.72%
归母净利润（百万元）	160.77	241.00	311.81	397.29
净利润增长率	137.07%	49.91%	29.38%	27.41%
EPS（元）	0.67	1.00	1.30	1.66
P/E	27.32	18.22	14.08	11.05

股价表现



数据来源：Wind，西南证券

数据来源：Wind，西南证券整理

道通科技（688208）：进入业绩兑现期，新能源业务有望再加速

□ 投资逻辑：

1）逐步走过转型阵痛期，盈利能力修复明显：公司2023年存在1.85亿非经常性损失，主要系此前与诉讼达成和解后支付的一次性费用，当前公司已和第三方数据公司签订合约，不利因素得到出清。费用端看，公司已逐步走过转型阵痛期，2023年公司研发费用率为16.5%，同比下降8.32pp，控费成效开始显现，盈利能力得到明显修复。**2）充电桩实现销售突破，传统业务需求回暖。**公司新能源充电桩业务开始放量，2023年全年实现收入5.7亿元，同比增长493.2%；2024Q1实现收入1.6亿元，同比增长103.3%。传统业务方面，海外需求也开始回暖，其中ADAS业务高速增长。**3）体系搭建完善，美国工厂已经投产：**2023年年底，公司美国工厂已经正式建成投产，后续有望降低《基础设施法案》带来的供应链风险，迎来产品规模化放量拐点，实现爆发式增长。

□ 业绩预测与投资建议：

预计公司2024-2026年归母净利润分别为5.4亿元、6.8亿元、8.4亿元，对应PE分别为22倍、18倍、14倍，给予“买入”评级。

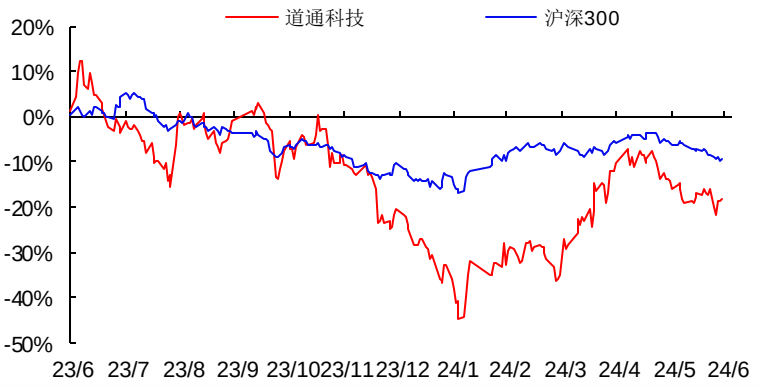
□ 风险提示：

国际贸易形势恶化风险；汇率波动风险；原材料价格上涨风险；供应链风险；行业竞争加剧风险等。

业绩预测和估值指标

指标	2023A	2024E	2025E	2026E
营业收入（百万元）	3251.15	4103.27	5189.47	6425.33
营业收入增长率	43.50%	26.21%	26.47%	23.81%
归母净利润（百万元）	179.23	541.73	676.21	838.80
净利润增长率	75.66%	202.25%	24.83%	24.04%
EPS（元）	0.40	1.20	1.50	1.86
P/E	67.22	22.33	17.90	14.43

股价表现



数据来源：Wind，西南证券

数据来源：Wind，西南证券整理

海光信息（688041）：AI与信创双轮驱动，国芯领军持续迭代

□ 投资逻辑：

1）深算二号重磅推出，性能翻倍提升：在国内高性能计算芯片受限加大的背景下，公司深算二号在2023年9月发布，性能比起上一代产品翻倍提升，并且兼容“类CUDA”环境，软硬件生态丰富，当前已在各地智算中心落地应用，对部分英伟达产品实现了良好替代，有望深度受益于下游AI算力的需求爆发。**2）CPU持续迭代，进一步缩小海外差距：**公司推出海光四号CPU产品，在性能、生态、自主可控程度等各方面进一步扩大优势，伴随信创的持续推进，后续业绩有望快速增长。**3）后续产品推出路线清晰，加快追赶脚步：**当前海光五号CPU产品、深算三号DCU产品研发进展顺利，其中深算三号针对AI训练硬件性能有望大幅提升，后续有望逐步打开互联网客户市场。

□ 业绩预测与投资建议：

预计公司2024-2026年归母净利润分别为16.9亿元、23.8亿元、31.8亿元，对应PE分别为97倍、69倍、51倍，给予“买入”评级。

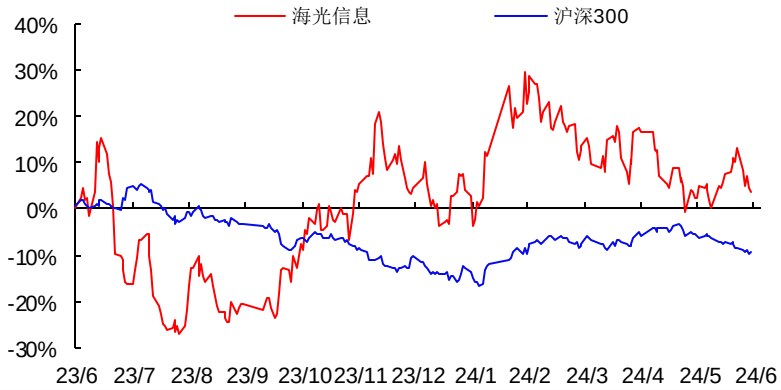
□ 风险提示：

研发进度或不及预期；信创推进或不及预期；供应链风险；行业竞争加剧风险等。

业绩预测和估值指标

指标	2023A	2024E	2025E	2026E
营业收入（百万元）	6012.00	8470.06	11961.00	15928.51
营业收入增长率	17.30%	40.89%	41.21%	33.17%
归母净利润（百万元）	1263.18	1688.88	2384.48	3184.46
净利润增长率	57.20%	33.70%	41.19%	33.55%
EPS（元）	0.54	0.73	1.03	1.37
P/E	129.39	96.78	68.55	51.33

股价表现



数据来源：Wind，西南证券

数据来源：Wind，西南证券整理

神州数码（000034）：业绩稳健增长，自主业务加速扩张

□ 投资逻辑：

1）AI服务器占比有望提升：公司打造了鲲鹏+昇腾为核心的AI服务器产品，可适用于多类场景需求。伴随大模型带动下游需求爆发，高价值的AI服务器占比有望继续提升，公司毛利结构或将进一步改善。**2）可转债落地加码云与信创布局，业务版图持续延伸：**公司可转债落地，募资超过13亿元，后续公司自主品牌产能将得到大幅扩张，同时基于DPU架构的混合算力平台已经完成第一期研发工作。**3）发布新一期员工持股计划，锚定长远发展：**2024年3月，公司发布最新一期员工持股计划（草案），拟筹集资金不超过4.63亿元，约占公司股本总额的2.39%。自2022年起，公司每年滚动进行员工持股或股权激励的计划，有利于公司在行业变革期绑定吸引优质人才，体现对未来发展坚定信心。

□ 业绩预测与投资建议：

预计公司2024-2026年归母净利润分别为14.1亿元、16.9亿元、20.0亿元，对应PE分别为11倍、9倍、8倍，给予“买入”评级。

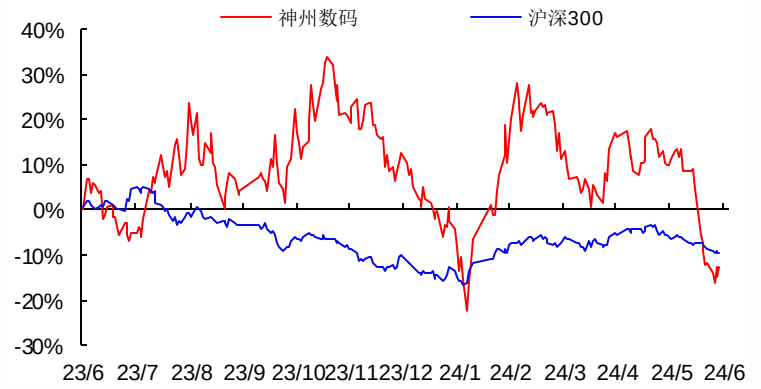
□ 风险提示：

信创产业推进不及预期、市场竞争加剧、产品研发进度不及预期、新业务拓展不及预期等风险。

业绩预测和估值指标

指标	2023A	2024E	2025E	2026E
营业收入（百万元）	119623.89	126177.59	134824.11	145326.74
营业收入增长率	3.23%	5.48%	6.85%	7.79%
归母净利润（百万元）	1171.78	1408.65	1693.27	2003.07
净利润增长率	16.66%	20.21%	20.20%	18.30%
EPS（元）	1.75	2.10	2.53	2.99
P/E	13.16	10.93	9.09	7.68

股价表现



数据来源：Wind，西南证券

数据来源：Wind，西南证券整理

金山办公（688111）：双订阅稳步增长，AI商业化元年可期

□ 投资逻辑：

1）WPS AI产品公测，有望拉动公司实现量价齐升。WPS AI全家桶已经开启公测，同时发布面向企业的一站式AI办公平台WPS365，其生产力和用户吸引力发生质变，产品定价与付费渗透率对标海外产品均有极大提升潜力，中期估值空间有望明显扩张。**2）订阅业务维持高速增长，业绩有支撑：**2024年Q1，公司C端订阅增速为24.8%、B端订阅增速为13.6%，PC端MAU为2.7亿，同比+7.14%，客户粘性较强。**3）国产替代核心受益标的，**伴随信创产业深化推进，公司业绩或将迎来明显加速。

□ 业绩预测与投资建议：

预计公司2024-2026年归母净利润分别为16.8亿元、21.6亿元、28.4亿元，对应PE分别为51倍、41倍、31倍，给予“买入”评级。

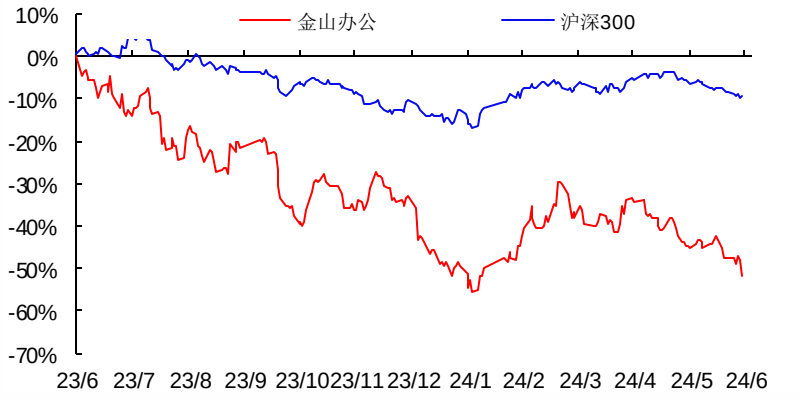
□ 风险提示：

WPS AI研发进度不及预期、信创产业推进不及预期、行业竞争加剧等风险。

业绩预测和估值指标

指标	2023A	2024E	2025E	2026E
营业收入（百万元）	4555.97	5793.59	7489.72	9661.96
营业收入增长率	17.27%	27.16%	29.28%	29.00%
归母净利润（百万元）	1317.74	1679.51	2158.22	2844.59
净利润增长率	17.92%	27.45%	28.50%	31.80%
EPS（元）	2.85	3.64	4.67	6.16
P/E	67.01	51.40	40.92	31.05

股价表现



数据来源：Wind，西南证券

数据来源：Wind，西南证券整理

新国都（300130）：业绩高速增长，海外布局完善

□ 投资逻辑：

1）收单流水已重回增势，后续行业格局有望向龙头集中：随着23Q4嘉联支付收单产品线开始大批量出货，23年12月份月度流水实现环比增长；伴随行业监管政策的逐步明晰，行业格局有望迎来进一步出清，公司后续或重新调整经营拓客策略，预计2024年收单业务流水重回增长趋势，具备较大利润弹性。**2）跨境支付与海外收单布局完善：**公司大力推动支付服务出海战略，跨境事业群团队组建完毕，并推出跨境支付产品Paykka支持10+全球主流币种的收款业务以及全球150+币种的收单业务，形成完整的“终端+支付”产品体系，后续海外业务有望成为公司新的增长极。**3）AIGC创造额外成长动力。**公司的AI业务在技术研发和商业化运作已取得初步成绩，国内亦成立子公司开展AI Agent技术研发。

业绩预测与投资建议：

预计公司2024-2026年归母净利润分别为9.4亿元、11.8亿元、14.2亿元，对应PE分别为10倍、8倍、6倍，给予“买入”评级。

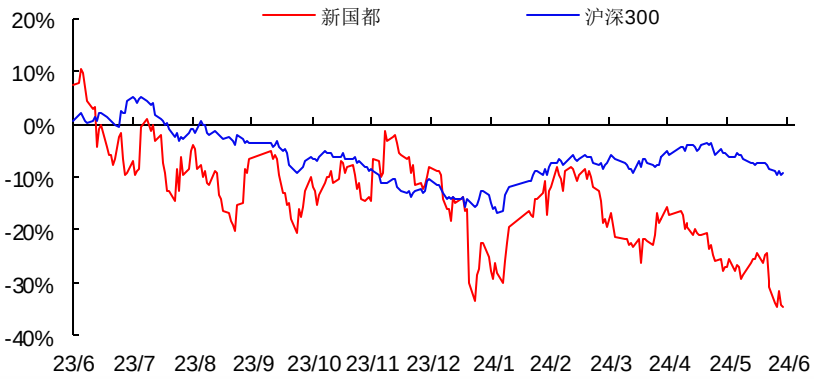
□ 风险提示：

线下消费复苏不及预期；商誉减值风险；汇率波动风险；行业竞争加剧等。

业绩预测和估值指标

指标	2023A	2024E	2025E	2026E
营业收入（百万元）	3801.26	4683.85	5750.71	6786.80
营业收入增长率	-11.94%	23.22%	22.78%	18.02%
归母净利润（百万元）	755.04	942.77	1180.54	1421.59
净利润增长率	1586.11%	24.86%	25.22%	20.42%
EPS（元）	1.36	1.69	2.12	2.55
P/E	11.92	9.55	7.62	6.33

股价表现



数据来源：Wind，西南证券

数据来源：Wind，西南证券整理

重点公司估值表

代码	公司	股价（元）	EPS（元）				PE（倍）				投资评级
			2023A	2024E	2025E	2026E	2023A	2024E	2025E	2026E	
002380	科远智慧	18.30	0.67	1.00	1.30	1.66	27	18	14	11	买入
688208	道通科技	24.14	0.40	1.20	1.50	1.86	67	22	18	14	买入
688041	海光信息	70.32	0.54	0.73	1.03	1.37	129	97	69	51	买入
000034	神州数码	22.89	1.75	2.10	2.53	2.99	13	11	9	8	买入
688111	金山办公	227.50	2.85	3.64	4.67	6.16	67	51	41	31	买入
300130	新国都	16.64	1.36	1.69	2.12	2.55	12	10	8	6	买入

数据来源：Wind，西南证券（时间截止2024/06/28）

目 录

风险提示

- 宏观经济承压；
- AI应用落地不及预期；
- 信创进度不及预期；
- 贸易摩擦加剧；
- 原材料价格上涨；
- 汇率波动风险；
- 板块政策发生重大变化；
- 研发进度不及预期等。

分析师：王湘杰
执业证号：S1250521120002
电话：0755-26671517
邮箱：wxj@swsc.com.cn

分析师：邓文鑫
执业证号：S1250523070002
邮箱：dwx@swsc.com.cn

分析师：罗紫莹
执业证号：S1250524070003
邮箱：lzyyf@swsc.com.cn

西南证券 SOUTHWEST SECURITIES

西南证券研究发展中心

西南证券投资评级说明

报告中投资建议所涉及的评级分为公司评级和行业评级（另有说明的除外）。评级标准为报告发布日后6个月内的相对市场表现，即：以报告发布日后6个月内公司股价（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；美国市场以纳斯达克综合指数或标普500指数为基准。

公司评级	买入：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在20%以上
	持有：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于10%与20%之间
	中性：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-10%与10%之间
	回避：未来6个月内，个股相对同期相关证券市场代表性指数涨幅介于-20%与-10%之间
	卖出：未来6个月内，个股相对同期相关证券市场代表性指数涨幅在-20%以下
行业评级	强于大市：未来6个月内，行业整体回报高于同期相关证券市场代表性指数5%以上
	跟随大市：未来6个月内，行业整体回报介于同期相关证券市场代表性指数-5%与5%之间
	弱于大市：未来6个月内，行业整体回报低于同期相关证券市场代表性指数-5%以下

分析师承诺

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注册为证券分析师，报告所采用的数据均来自合法合规渠道，分析逻辑基于分析师的职业理解，通过合理判断得出结论，独立、客观地出具本报告。分析师承诺不曾因，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获取任何形式的补偿。

重要声明

西南证券 SOUTHWEST SECURITIES

西南证券研究发展中心

西南证券研究发展中心

上海

地址：上海市浦东新区陆家嘴21世纪大厦10楼

邮编：200120

北京

地址：北京市西城区金融大街35号国际企业大厦A座8楼

邮编：100033

深圳

地址：深圳市福田区益田路6001号太平金融大厦22楼

邮编：518038

重庆

地址：重庆市江北区金沙门路32号西南证券总部大楼21楼

邮编：400025

西南证券机构销售团队

区域	姓名	职务	手机	邮箱	姓名	职务	手机	邮箱
上海	蒋诗烽	总经理助理/销售总监	18621310081	jsf@swsc.com.cn	魏晓阳	销售经理	15026480118	wxyang@swsc.com.cn
	崔露文	销售副总监	15642960315	clw@swsc.com.cn	欧若诗	销售经理	18223769969	ors@swsc.com.cn
	谭世泽	高级销售经理	13122900886	tsz@swsc.com.cn	李嘉隆	销售经理	15800507223	ljlong@swsc.com.cn
	李煜	高级销售经理	18801732511	yfliyu@swsc.com.cn	龚怡芸	销售经理	13524211935	gongyy@swsc.com.cn
	卞黎旸	高级销售经理	13262983309	bly@swsc.com.cn	孙启迪	销售经理	19946297109	sqdi@swsc.com.cn
	田婧雯	高级销售经理	18817337408	tjw@swsc.com.cn	蒋宇洁	销售经理	15905851569	jyj@swsc.com.c
	张玉梅	销售经理	18957157330	zymyf@swsc.com.cn				
北京	李杨	销售总监	18601139362	yfly@swsc.com.cn	王一菲	销售经理	18040060359	wyf@swsc.com.cn
	张岚	销售副总监	18601241803	zhanglan@swsc.com.cn	王宇飞	销售经理	18500981866	wangyuf@swsc.com
	杨薇	资深销售经理	15652285702	yangwei@swsc.com.cn	路漫天	销售经理	18610741553	lmtyf@swsc.com.cn
	姚航	高级销售经理	15652026677	yhang@swsc.com.cn	马冰竹	销售经理	13126590325	mbz@swsc.com.cn
	张鑫	高级销售经理	15981953220	zhxin@swsc.com.cn				
广深	郑龑	广深销售负责人	18825189744	zhengyan@swsc.com.cn	丁凡	销售经理	15559989681	dingfyf@swsc.com.cn
	杨新意	广深销售联席负责人	17628609919	yxy@swsc.com.cn	陈紫琳	销售经理	13266723634	chzlyf@swsc.com.cn
	张文锋	高级销售经理	13642639789	zwf@swsc.com.cn	陈韵然	销售经理	18208801355	cyryf@swsc.com.cn
	龚之涵	销售经理	15808001926	gongzh@swsc.com.cn	林哲睿	销售经理	15602268757	lzr@swsc.com.cn