ChatGPT: 三大主线，AI算力需求井喷!

AIGC行业深度报告(7)

华西计算机团队

2023年4月19日

分析师：刘泽晶
SAC NO：S1120520020002
邮箱：liuzj1@hx168.com.cn

核心逻辑:

- **政策端与产业端持续发力，算力建设持续提速:** 国家计算力指数与GDP/数字经济的走势呈现出了显著的正相关，根据IDC数据，十五个重点国家的计算力指数平均每提高1点，国家的数字经济和GDP将分别增长 3.5‰和1.8‰；政策端与产业端持续发力，当前我国已进入《新型数据中心发展三年行动计划(2021-2023年)》落地见效的关键年，2023年4月17日国家超算互联网联合体成立，我们认为我国算力建设正式进入持续提速阶段。此外，北京、贵州、上海、惠州、天津等地算力基础设施计划持续落地，彰显我国对算力设施的高度重视。

- **大模型引爆海量算力需求，AI产业量价齐升:** 根据OpenAI数据，模型计算量增长速度远超人工智能硬件算力增长速度，存在万倍差距。运算规模的增长，带动了对AI训练芯片单点算力提升的需求，其中AI芯片、AI服务器、存储是AI基建的重要组成部分。AI硬件竞争升温，芯片价格大涨，根据界面新闻，截至本周英伟达AI旗舰芯片H100售价在多个商铺高至4万美元，相比此前零售商报价3.6万美元，已明显提价，同时，据财联社消息，上游芯片带动服务器价格同步上升，例如闻泰科技等；此外，存储方面，大容量、高速存储需求增加，近期HBM3规格DRAM价格上涨约5倍。

- **全球算力市场持续火热，AI芯片群星闪耀:** 2023年3月23日，英伟达GTC主题演讲开启，英伟达展示全新的芯片和系统、加速库、云服务、AI服务以及助力以助力全球AI生态。我们认为英伟达的目的即快速抢占AI相关市场，从而在科技储备上具备先发优势，同样也侧面反映出全球算力市场具有高需求性和高爆发性。此外，相关科技巨头厂商纷纷加入“算力储备战”，全球算力市场持续火热。例如拟推出人工智能芯片，助力大型语言模型、博通发布用于连接AI超级计算机的Jericho3-AI芯片；我国相关企业同样加速AI芯片布局，例如寒武纪、百度、遂原科技、昆仑芯等。

- **投资建议:** 关注五条投资主线：**1) AI芯片厂商，**相关受益标的为: **赛武纪、海光信息、景嘉微、龙芯中科**等；**2) 存储厂商，**相关受益标的为: **东芯股份、兆易创新、澜起科技、聚辰股份、普再股份、江波龙、佰维存储、恒烁股份**等; **3) 光模块厂商，**相关受益标的为: **新易盛、中际旭创、天孚通信、剑桥科技、源杰科技、联特科技、光迅科技**等；**4)服务器及IDC厂商，**相关受益标的为: **浪潮信息、中科曙光、神州数码、拓维信息、工业富联、润泽科技**等；**5) AI云厂商，**相关受益标的为: **首都在线、云赛智联、青云科技、优刻得、光环新网、新炬网络**等。

- **风险提示:** 核心技术水平升级不及预期的风险、AI伦理风险、政策推进不及预期的风险、中美贸易摩擦升级的风险。

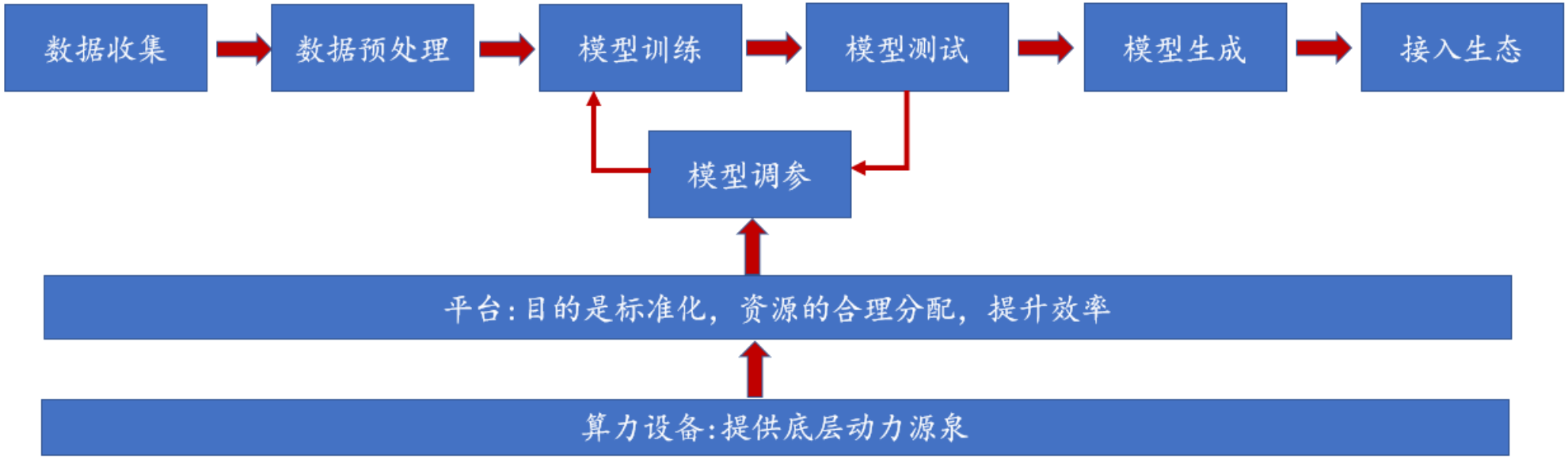
目录

01 全国超算建设启动，AI产业量价齐飞

1.1 重申强调ChatGPT的竞争本质即大模型储备竞赛

- **大模型是人工智能发展的必然趋势**：大模型即“大算力+强算法”结合的产物。大模型通常是在大规模无标注数据上进行训练，学习出一种特征和规则。基于大模型进行应用开发时，将大模型进行微调，如在下游特定任务上的小规模有标注数据进行二次训练，或者不进行微调，就可以完成多个应用场景的任务。
- **大模型是辅助式人工智能向通用性人工智能转变的坚实底座**：大模型增强了人工智能的泛化性、通用性，生产水平得到质的飞跃，过去分散化模型研发下，单一AI应用场景需要多个模型支撑，每个模型需要算法开发、数据处理、模型训练、参数调优等过程。大模型实现了标准化AI研发范式，即简单方式规模化生产，具有“预训练+精调”等功能，显著降低AI开发门槛，即“低成本”和“高效率”。
- **算力是打造大模型生态的必备基础，服务器是算力的载体**：**算力**是训练大模型的底层动力源泉，一个优秀的算力底座在大模型(AI算法)的训练和推理具备效率优势；**平台是大模型和算力之间的“桥梁”**，可针对不同的模型和硬件，实现资源的合理分配，达到软硬件的最优组合，从而大幅提升训练模型的效率。

数据、平台、算力、算法关系示意图



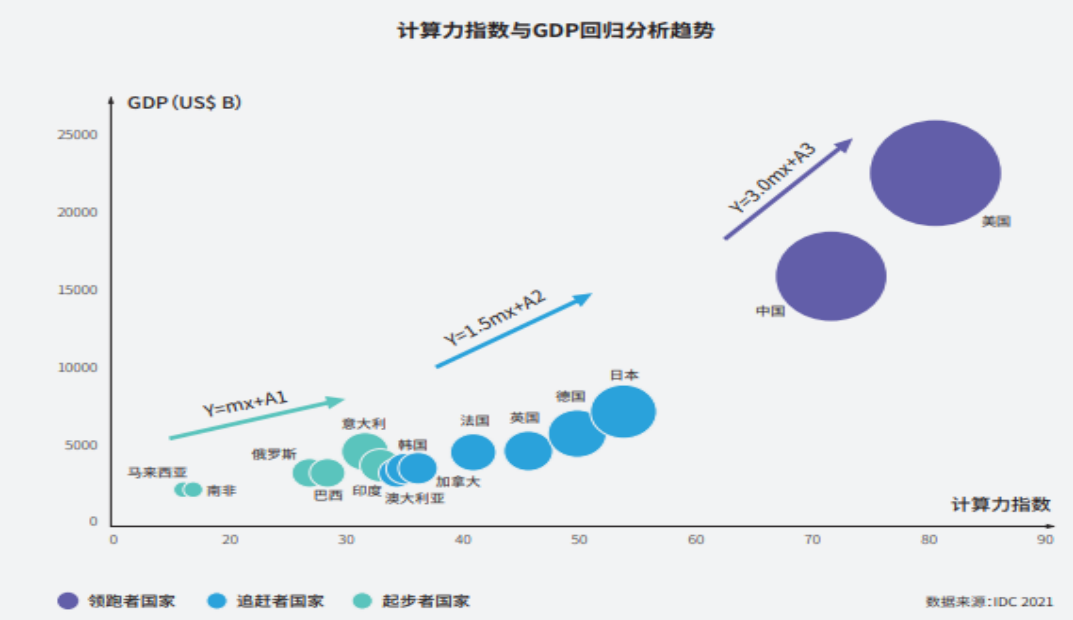
资料来源：IDC，华西证券研究所

1.2.1 国家计算力指数与GDP呈现出了显著的正相关

- **国家计算力指数与GDP/数字经济的走势呈现出了显著的正相关:** 根据IDC数据，十五个重点国家的计算力指数平均每提高1点，国家的数字经济和GDP将分别增长 3.5‰和1.8‰，预计该趋势在2021-2025年将继续保持。此外，当一个国家的计算力指数达到40分以上时，国家的计算力指 数每提升1点，其对于GDP增长的推动力将增加到1.5倍，而当计算力指数达到60分以上时，国家的计算 力指数每提升1点，其对于GDP增长的推动力将提高到3.0倍，对经济的拉动作用变得更加显著。

- **海量应用场景，算力需求高涨**:据华为发布的《计算2030》预测，2030年人类将进入YB数据时代，全球数据每年新增1YB。通用算力将增长10倍到3. 3ZFLOPS 、人工智能算力将增长500倍超过100ZFLOPS。相当于一百万个中国超级计算机神威“太湖之光”的算力总和。

从计算力指数看对经济的增长



计算力对经济的影响

资料来源：IDC《全球算力指数评估报告》，华为《计算2030》华西证券研究所

1.2.1 新型数据中心发展的政策落地推进算力建设持续提速

- **当前我国已进入《新型数据中心发展三年行动计划（2021-2023年）》落地见效的关键年：**《行动计划》主要目标为用3年时间，基本形成布局合理、技术先进、绿色低碳、算力规模与数字经济增长相适应的新型数据中心发展格局。到2023年底，全国数据中心机架规模年均增速保持在20%左右，平均利用率力争提升到60%以上，总算力超过200 EFLOPS，高性能算力占比达到10%。

- 2023年4月17日国家超算互联网联合体成立，**算力建设持续提速：**科技部高新司2023年4月17日在天津组织召开国家超算互联网工作启动会，会议发起成立了国家超算互联网联合体。超算互联网是用互联网思维运营超算，将全国众多超算中心通过算力网络连接起来，构建一体化算力服务平台，解决当前亟待突破的现有单体超算中心运营模式，以应对算力设施分布不均衡、接口不统一、应用软件自主研发和推广不足等问题。

工信部印发《新型数据中心发展三年行动计划（2021-2023年）》

国家超算互联网正式启动

资料来源：中共中央人民政府网、华西证券研究所

1.2.2 地方全力保障数字基础设施建设，积极带动关联产业集聚发展

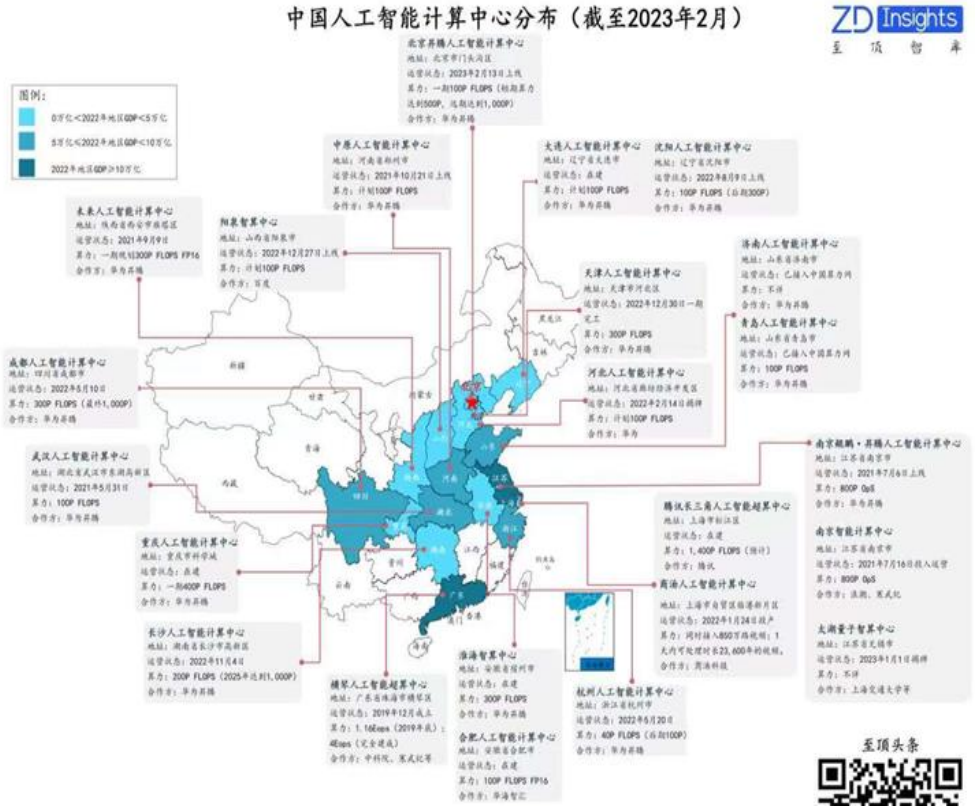
- **北京昇腾人工智能计算中心正式点亮：**北京昇腾人工智能计算中心正式点亮，将推动北京人工智能产业高质量发展。该智能计算中心采用昇腾AI基础软硬件，充分释放硬件算力，加速人工智能企业创新应用和模型孵化。

- **贵州省大数据局印发《面向全国的算力保障基地建设规划》：**总体目标是到2025年，面向全国的算力保障基地建设任务全面完成，贵州超大规模数据中心集群的地位更加巩固，存算比更加合理，优化基础设施布局、结构、功能和系统集成，数据中心实现集约化、规模化、绿色化发展，网络互联互通、能源安全可靠提高到新的水平，打造具有国际竞争力的数字产业集群。

- **上海市经济信息化委印发《上海市推进算力资源统一调度指导意见》：**主要目标为到2023年底，依托本市人工智能公共算力服务平台，接入并调度4个以上算力基础设施，可调度智能算力达到1,000 PFLOPS（FP16）以上；到2025年，市人工智能公共算力服务平台能级跃升，完善算力交易机制，实现跨地域算力智能调度，通过高效算力调度，推动算力供需均衡，带动产业发展作用显著增强。

- **惠州首个超大规模数据及算力中心力争年内投产：**2023年年初，作为大数据及关联产业发展的重要支撑点的粤港澳大湾区（惠州）数据产业园建设取得明显成效。落户该园区的润泽（惠州）国际信息港一期项目试运行工作进展顺利，预计年内正式投产。其目标是构建具有国际领先技术水平的算力基础设施，带动数据服务及硬件研发制造等关联产业集聚发展。

中国人工智能计算中心分布图（截至2023年2月）



资料来源： 至顶头条，贵州大数据发展管理局，上海经信委，惠州日报，华西证券研究所

1.2.2 地方全力保障数字基础设施建设，积极带动关联产业集聚发展

- **山东首个人工智能计算中心上线运行，竞逐人工智能赛道：**2023年3月17日青岛市人工智能产业园正式开园，同步上线的青岛人工智能计算中心，成为山东首个上线运行的人工智能计算中心。中心首期具备100P算力，相当于5万台高性能PC的算力，将面向青岛乃至胶东地区的企业、高校和科研机构提供普惠公共算力服务。
- **河南省数字化转型战略工作方案出炉，推进郑州、洛阳构建超大型绿色数据中心集群：**2023年3月30日河南省制造强省建设领导小组办公室印发《2023年河南省数字化转型战略工作方案》，目标今年电子信息制造业营业收入力争突破8000亿元，先进计算、软件产业规模均超过500亿元。
- **天津市人工智能计算中心揭牌，加快打造天津数字经济发展新动能：**2023年3月18日，天津市人工智能计算中心正式揭牌上线，助力人工智能产业创新发展。人工智能中心不仅提供基础算力服务，还提供应用创新服务、产业孵化服务等，把算力、算法、数据、应用场景和人才进行5要素的聚集，帮助企业在人工智能科研创新上降本增效。

2023年河南省数字化转型战略目标任务分解

序号	城市	智能工厂/智能车间	贯标升级版/对标升级版（家）	数字化转型项目（个）	企业上云
1	郑州市	22	36/360	116	10740
2	开封市	7	15/150	48	1120
3	洛阳市	12	24/240	80	2080
4	平顶山市	7	15/150	52	1540
5	安阳市	7	15/150	52	460
6	鹤壁市	6	12/120	40	290
7	新乡市	12	21/210	72	1840
8	焦作市	8	18/180	60	440
9	濮阳市	5	12/120	40	570
10	许昌市	9	18/180	60	1610
11	漯河市	6	12/120	40	730
12	三门峡市	5	12/120	40	350
13	南阳市	12	21/210	72	2730
14	商丘市	9	18/180	56	1710
15	信阳市	4	15/150	52	1510
16	周口市	8	15/150	52	1360
17	驻马店市	7	15/150	48	770
18	济源示范区	4	6/150	20	150
合计		150	300/3000	1000	30000

天津市人工智能计算中心内的算力服务器

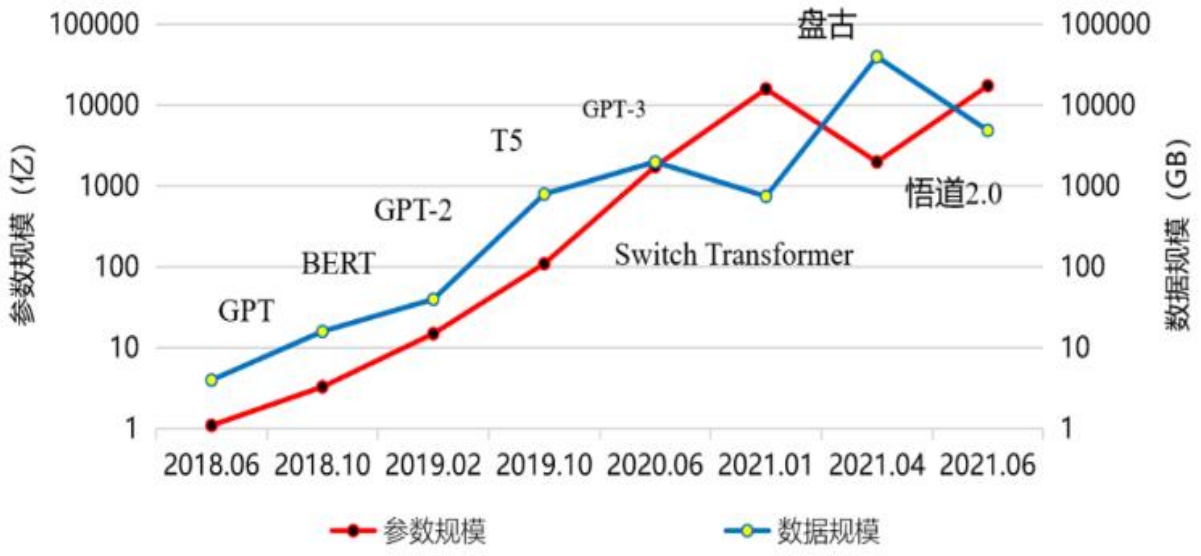
资料来源：山东政府网，河南省工业与信息化厅官网，天津发布，华西证券研究所

1.3.1 ChatGPT开启大模型“军备赛”，算力呈现明显缺口

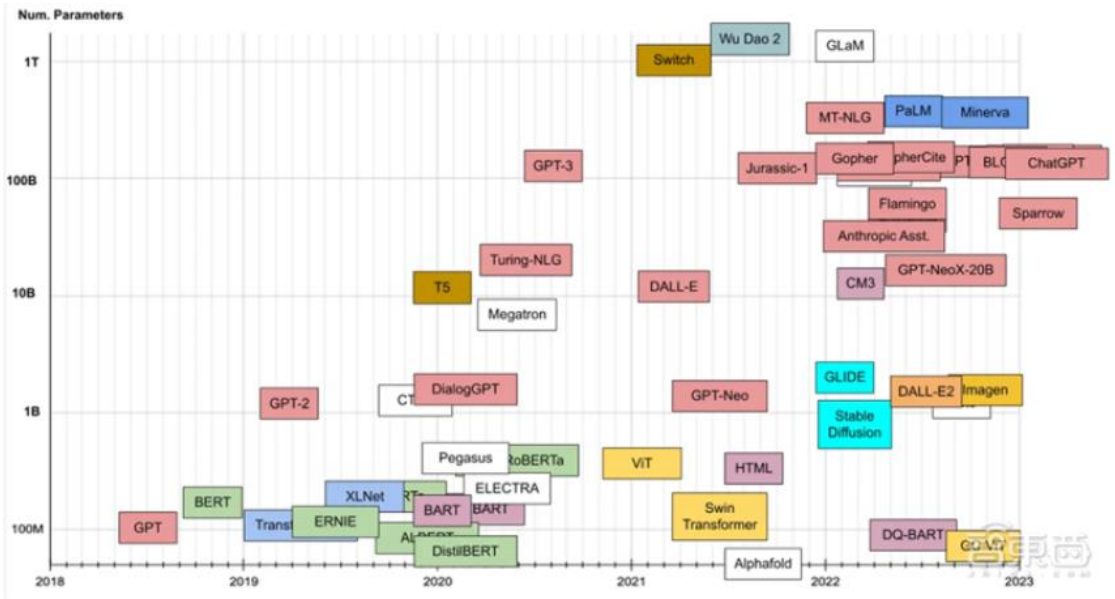
- **ChatGPT开启算力军备赛**：我们已经在《ChatGPT：百度文心一言畅想》中证明数据、平台、算力是打造大模型生态的必备基础，且算力是训练大模型的底层动力源泉，一个优秀的算力底座在大模型(AI算法)的训练和推理具备效率优势；同时，我们在《ChatGPT打响AI算力“军备战”》中证明算力是AI技术角逐“入场券”，其中AI服务器、AI芯片等为核心产品；此外，我们还在《ChatGPT ， 英伟达DGX引爆 AI “核聚变”》中证明以英伟达为代表的科技公司正在快速补足全球AI算力需求，为大模型增添必备“燃料”。

- **大模型参数呈现指数规模，引爆海量算力需求**： 根据财联社和OpenAI数据，ChatGPT浪潮下算力缺口巨大，根据OpenAI数据，模型计算量增长速度远超人工智能硬件算力增长速度，存在万倍差距。运算规模的增长，带动了对AI训练芯片单点算力提升的需求，并对数据传输速度提出了更高的要求。根据智东西数据，过去五年，大模型发展呈现指数级别，部分大模型已达万亿级别，因此对算力需求也随之攀升。

大模型参数数量和训练数据规模快速增长



近年大模型的参数规模增长趋势

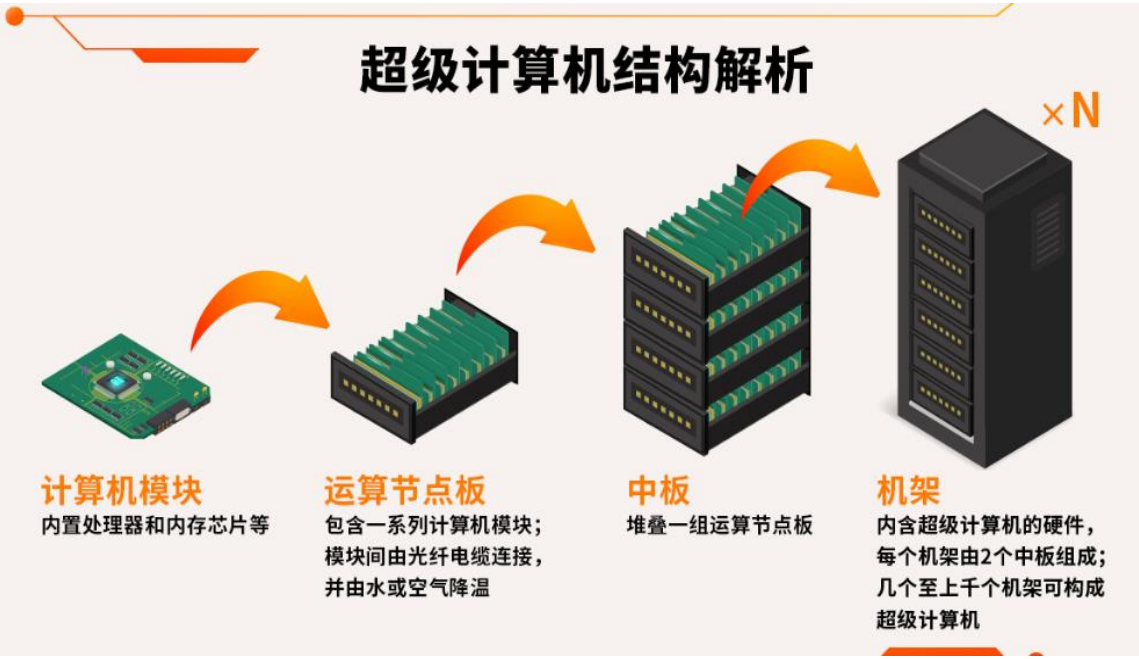


资料来源：新浪，智东西，可创办日报，华西证券研究所

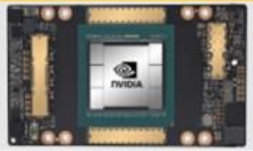
1.3.2 AI芯片、AI服务器为AI基础设施重要组成部分

- **AI芯片是AI算力的“心脏”：**伴随数据海量增长，算法模型趋向复杂，处理对象异构，计算性能要求高，AI 芯片在可高效处理人工智能应用中日渐多样繁杂的计算任务。**其中GPU相较于比CPU更擅长并行计算，**CPU是以低延迟为导向的计算单元，而GPU是以吞吐量为导向的计算单元，转为执行多任务并行。由于微架构的不同导致CPU绝大部分晶体管用于构建控制电路和缓存，只有小部分晶体管用来完成运算工作，GPU则是流处理器和显存控制用于绝大部分晶体管，从而拥有更强大的并行计算能力和浮点计算能力。
- **AI服务器作为超算芯片载体彰显其重要性:** 与通用服务器采用串行架构、以CPU为算力提供者不同的是，AI服务器采取异构架构，如CPU+GPU、CPU+TPU、CPU+其他的加速卡等不同的组合方式，目前广泛使用的是CPU+GPU 。与通用服务器相比，AI服务器拥有更出色的高性能计算能力，未来，随着算力的持续增长，自然语言处理和图像、视频等AI模型的深入发展，AI服务器将被更广泛使用。

超级计算机架构



英伟达数据中心GPU类别

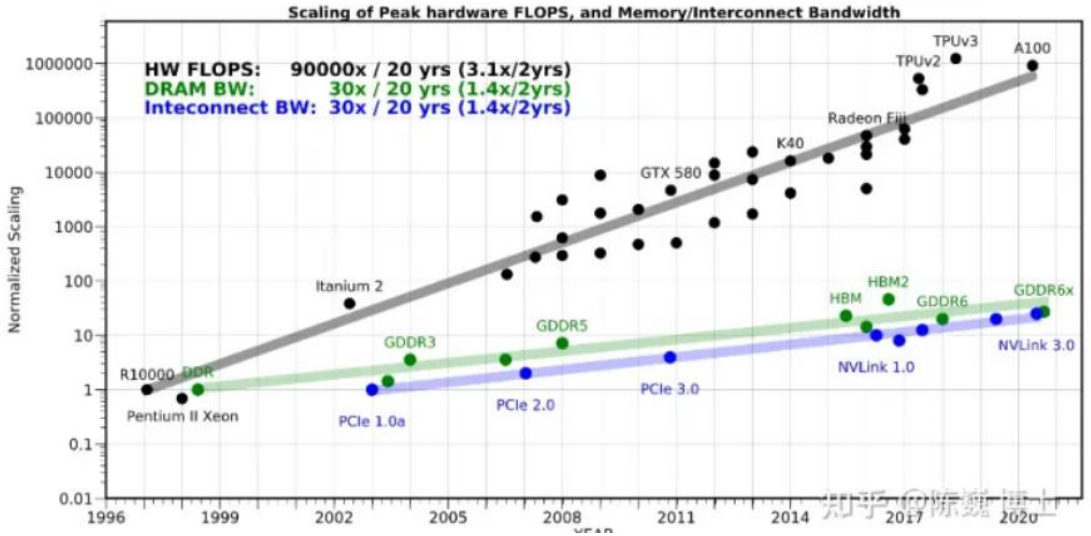
VideoCardz.com	NVIDIA H100	NVIDIA A100	NVIDIA Tesla V100	NVIDIA Tesla P100
Picture				
GPU	GH100	GA100	GV100	GP100
Transistors	80B	54.2B	21.1B	15.3B
Die Size	814 mm²	828 mm²	815 mm²	610 mm²
Architecture	Hopper	Ampere	Volta	Pascal
Fabrication Node	TSMC N4	TSMC N7	12nm FFN	16nm FinFET+
GPU Clusters	132/114*	108	80	56
CUDA Cores	16896/14592*	6912	5120	3584
L2 Cache	50MB	40MB	6MB	4MB
Tensor Cores	528/456*	432	320	-
Memory Bus	5120-bit	5120-bit	4096-bit	4096-bit
Memory Size	80 GB HBM3/HBM2e*	40/80GB HBM2e	16/32 HBM2	16GB HBM2
TDP	700W/350W*	250W/300W/400W	250W/300W/450W	250W/300W
Interface	SXM5/*PCIe Gen5	SXM4/PCIe Gen4	SXM2/PCIe Gen3	SXM/PCIe Gen3
Launch Year	2022	2020	2017	2016

资料来源： CDSN，embeddedcomputing，华西证券研究所

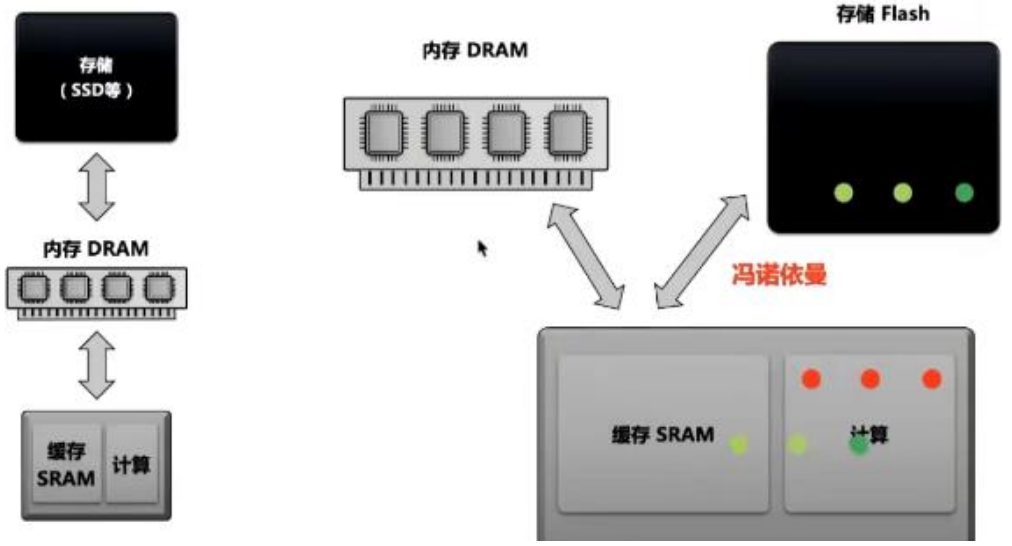
1.3.3 存储同样是计算机的重要组成部分

- **存储是计算机的重要组成结构**：存储器是用来存储程序和数据的部件，对于计算机来说，有了存储器才有记忆功能，才能保证正常工作。存储器按其用途可分为主存储器和辅助存储器，主存储器又称内存储器(简称内存)，辅助存储器又称外存储器(简称外存)。
- XPU、**内存、硬盘组成完整的冯诺依曼体系**：“内存”实为硬盘与CPU之间的中间人，CPU如果直接从硬盘中抓数据，时间会太久。所以“内存”作为中间人，从硬盘里面提取数据，再让CPU直接到内存中拿数据做运算。这样会比直接去硬盘抓数据，快百万倍；CPU里面有一个存储空间Register(寄存器)，运算时，CPU会从内存中把数据载入Register，再让Register中存的数字做运算，运算完再将结果存回内存中，因此运算速度Register > 内存 > 硬盘，速度越快，价格越高，容量越低。
- **算力发展速度远超存储，存储带宽限制计算系统的速度**：在过去二十年，处理器性能以每年大约55%的速度提升，内存性能的提升速度每年只有10%左右。因此，目前的存储速度严重滞后于处理器的计算速度。能耗方面，从处理单元外的存储器提取所需的时间往往是运算时间的成百上千倍，因此能效非常低；“存储墙”成为加速学习时代下的一代挑战，原因是数据在计算单元和存储单元的频繁移动。

算力发展速度远超存储器



数据存储示意图

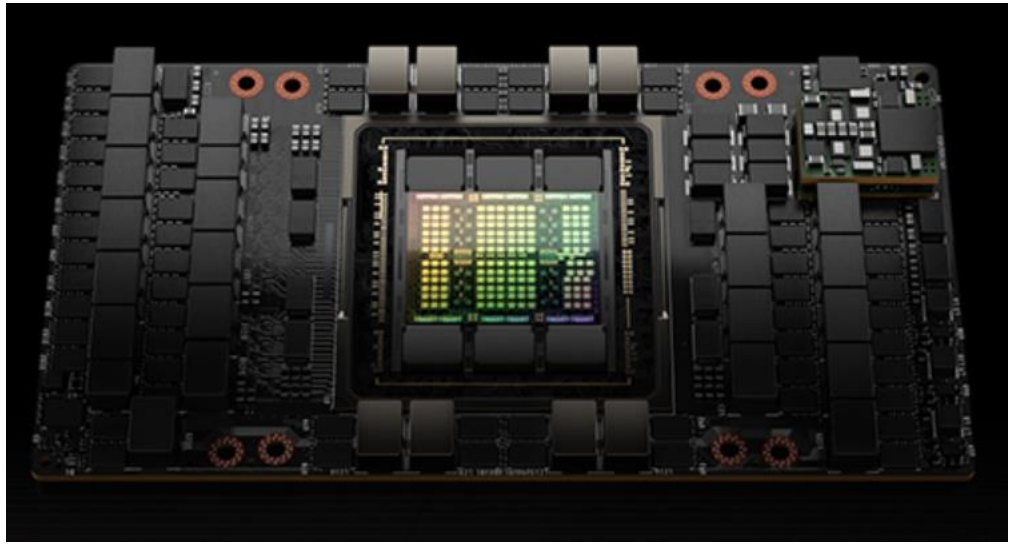


资料来源：知乎@陈巍谈芯，华西证券研究所

1.3.4 AI算力硬件迎来抢购热潮，芯片+服务器价格飞涨

- **AI硬件竞争升温，芯片遭“哄抢”导致价格大涨：**随着ChatGPT带来AI产业大热，相关产业对于AI算力硬件的需求也同步高涨。作为AI算力基础，以英伟达A100和H100 GPU为代表的产品成为抢手货。根据界面新闻，截至本周，英伟达AI旗舰芯片H100售价在多个商铺炒至4万美元，相比此前零售商报价3.6万美元，已明显提价。国内云计算技术人士认为，1万枚英伟达A100芯片是做好AI大模型的算力门槛，而为了支持实际应用、满足服务器需求，OpenAI已使用了约2.5万个英伟达的GPU。我们认为未来随着需求将进一步增加，或将进一步推高高性能AI芯片的价格。
- **上游芯片带动服务器价格同步上行：**一台服务器通常需要4枚-8枚GPU，根据OpenAI训练集群模型估算结果，1746亿参数的GPT-3模型大约需要375-625台8卡DGX A100服务器（对应训练时间10天左右）。一台GPU服务器的成本是普通服务器的10倍以上，GPU价格高涨直接带动服务器价格显著上修。以国产浪潮AI智能服务器为例，根据AI市场报价，其R4900G3规格产品含税价已高达55万元。据财联社消息，闻泰科技同样称其服务器价格呈上涨趋势。

英伟达芯片H100



INSPUR浪潮AI智能服务器

规格	2U Rack
处理器	支持 1 到 2 个英特尔®至强®3100，4100，5100，6100，8100 系列可扩展处理器； 支持 28 核（频率 2.5GHz） 频率 3.6GHz（4 核） 两条 UPI 互连链路，单条链路高速率 10.4GT/s 热设计功率 205W
芯片组	Intel C622/C624
内存	支持 24 根内存，每个处理器支持 6 个内存通道，每个通道 支持 2 个内存插槽，内存 速度可达 2666MT/s 支持 RDIMM 和 LRDIMM，内存保护支持 ECC，内存镜像，内存等级保护
内存最大容量	24 根 DDR4 Registered，LR DIMM，单条 支持 128GB
存储	前置： 8 块或12 块 3.5 英寸硬盘或 25 块 2.5 寸硬盘 内置： 4 块 3.5 英寸硬盘，2 块 M.2 SSD 后置： 4 块 3.5 英寸硬盘+4*2.5 寸硬盘
M.2 & SD	支持两个 M.2 支持两个 MICRO SD
存储控制器	RAID 卡控制器 SAS 3108、3008、9361、PM8060、 SAS 卡控制器 9400 板载形式支持两种模式混插 SATA/NVMe；提供 RAID 0/1/5/6/10/50/60(NVMe 暂不支持) NVMe 需要单独配置 RAID key Software RAID 支持 RAID 0/1/5/
网络接口	1 个 OCP 及 1 个 OCP/PHY 模组提供 1Gb/s,10Gb/s,25Gb/s 支持标准 1GB/10GbE/25Gb/40G/100Gb 网卡

资料来源：AI市场，英伟达官网，财联社，华西证券研究所

1.3.5 AI服务器带来存力硬件需求上行，存储器价格同步高增

- **大容量、高速存储需求增加，HBM扮演重要角色**：HBM（高带宽存储器），是超微半导体和SK海力士发起的一种基于3D堆栈工艺的高性能DRAM。HBM重新调整了内存的功耗效率，能大幅提高数据处理速度，是当下速度最快的DRAM产品，其每瓦带宽比GDDR5高出3倍，且比GDDR5节省了94%的表面积。目前，HBM主要被安装于GPU、AI加速器、超级计算机、高效能服务器等。随着ChatGPT等应用开启AI新时代，加之相关技术演进，预计全球数据生成、储存、处理量将呈等比级数增长，HBM将扮演更重要的角色。
- **第三代HBM报价大涨**：NVIDIA计算卡供不应求，使得HBM3显存出现了严重短缺的情况，由此导致作为HBM3显存供应商的三星及海力士产品报价不断提升，远超平均报价水准。据财联社消息，2023年开年后三星、SK海力士的HBM订单快速增加，近期HBM3规格DRAM价格上涨约5倍。

三星HBM3 Icebolt产品性能

速度突破

· 6.4Gbps 处理速度和高达 819GB/s 的带宽
· 比上一代产品快将近 1.8 倍

芯片堆叠技术

· 采用 12 层 10 纳米级 16 Gb DRAM die堆叠，实现 24GB 的存储容量
· 容量是上一代产品的1.5倍

低功耗

· 处理速度更快，存储容量更大，但能效比上一代产品提高 10%

高可靠性

· 在提供更高速度的情况下，仍提供增强型自我校正功能
· 可校正所有 16 位错误

SK海力士HBM3 DRAM产品性能

HBM3与HBM2E DRAM比较

规格	HBM3	HBM2E	比较
引脚数据速率*	6.4Gbps	3.6Gbps	×1.78
带宽*	819GB/s	460GB/s	
核心Die堆叠*	12 层	8 层	×1.5
堆叠密度*	24GB	16GB	
On-die ECC	O	X	-

*最大数值

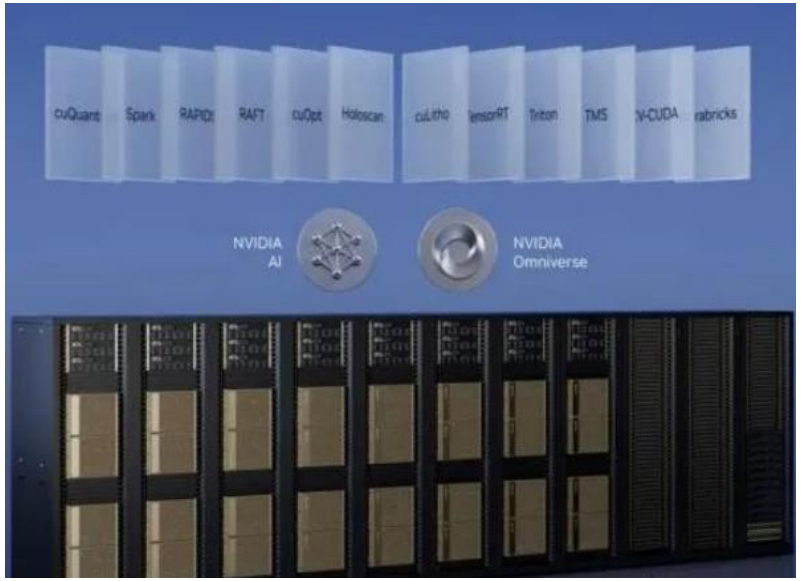
资料来源：三星、海力士公司官网，财联社，华西证券研究所

02 AI芯片群星闪耀

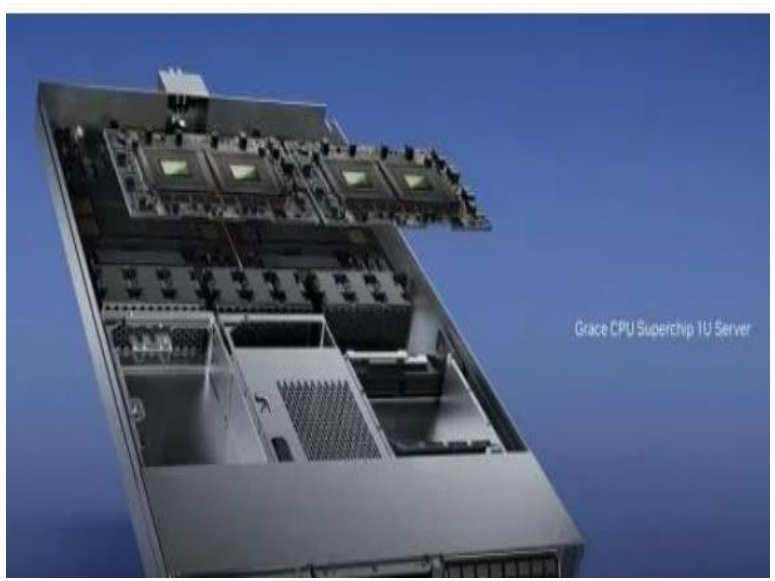
2.1 英伟达DGX引爆AI”核聚变”，全球算力市场持续火热

- ◆ **英伟达持续赋能加速计算AI场景**：2023年3月23日，英伟达GTC会议主题演讲开启，是一场全球的科技盛宴，我们认为其中最重磅的消息，是英伟达展示全新的芯片和系统、加速库、云服务、AI服务，其目的是助力全球AI生态，我们认为此次GTC大会实则为一场全球AI盛宴。
- ✓ **基础软件加速库**：加速库着力解决的是普通计算机无法解决的问题，目前英伟达加速库，已在多个领域持续开辟AI加速的新市场。例如汽车航空空气动力学仿真、量子电路仿真、推荐系统、数字人、优化物流服务、视频处理、基因计算、芯片制造等方面；
- ✓ **硬件架构，全新的数据中心CPU，Grace**：Grace CPU横空出世，其中包含了72个 Arm 核心，由超高速片内可扩展的、缓存一致的网络连接，可提供3.2TB/s 的截面带宽，与以往CPU 不同， 此款CPU可以在云数据中心规模下实现高能效，非常适合云计算应用和科学计算应用。此外，在微服务方面， Grace的速度比最新一代 x86 CPU 的平均速度快 1.3倍， 而在数据处理中则快1.2倍；

英伟达计算加速库示意图



英伟达数据中心GRACE GPU



英伟达数据中心GRACE GPU

资料来源：英伟达官网，华西证券研究所

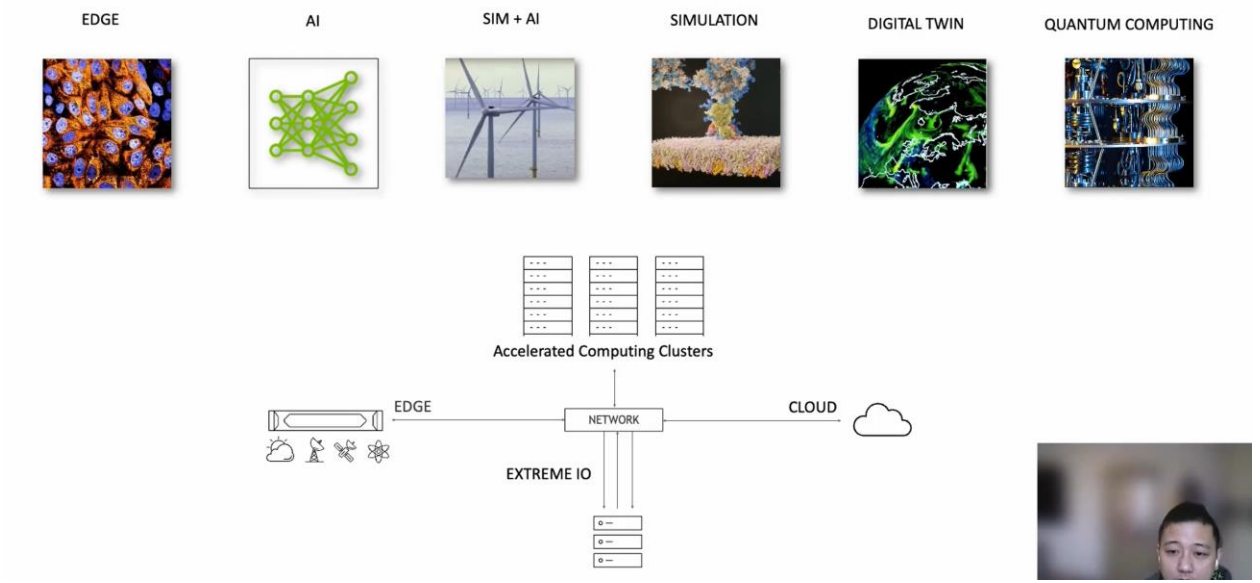
2.1 英伟达DGX引爆AI”核聚变”，全球算力市场持续火热

- ✓ **服务器，英伟达推出DGX(超级计算机)**：DGX配有8个H100 GPU模组，H100配有Transformer引擎，旨在处理大模型，8个H100模组通过NVLINK Switch彼此相连，以实现全面无阻塞通信。8个H100协同工作，类似一个巨型GPU；
- ✓ **云服务，英伟达推出foundations云服务**：企业可以通过在NVIDIA DGX Cloud上的NVIDIA NeMo服务快速采用生成式AI，通过此种云服务能够构建、改进和操作定制的大型语言模型和生成式AI模型，目前已知的服务包括快速定制基础语言模型、加速跨图像、视频和 3D 的模拟和创意设计、生物学云服务等。
- ◆ **英伟达DGX引爆AI”核聚变”，全球算力市场持续火热**：我们认为英伟达重磅发布AI系列产品意义重大，我们认为英伟达的目的即快速抢占AI相关市场，**从而在科技储备上具备先发优势，同样也侧面反映出全球算力市场具有高需求性和高爆发性。**此外，**相关科技巨头厂商纷纷加入“算力储备战”，全球算力市场持续火热。**

英伟达DGX H100相关参数

显卡	8 个 NVIDIA H100 Tensor Core GPU
显存	总计 640GB
表现	32 petaFLOPS FP8
NVIDIA® NVSwitch™	4倍
系统电源使用	~11.3kW 最大值
中央处理器	双 56 核第 4 代英特尔® 至强® 可扩展处理器
系统内存	2TB
联网	4x OSFP 端口服务于 8x 单端口 NVIDIA ConnectX-7 VPI - 400Gb/s InfiniBand/以太网 2x 双端口 NVIDIA ConnectX-7 VPI - 1x 400Gb/s InfiniBand - 1x 200Gb/s 以太网
贮存	操作系统：2 个 1.9TB M.2 NVME 驱动器 内部存储：8 个 3.84TB NVMe U.2 驱动器
管理网络	带 RJ45的 10Gb/s 板载 NIC 带 RJ45 的 50Gb/s 以太网可选 NIC 主机 BMC

英伟达AI加速计算上云架构示意图



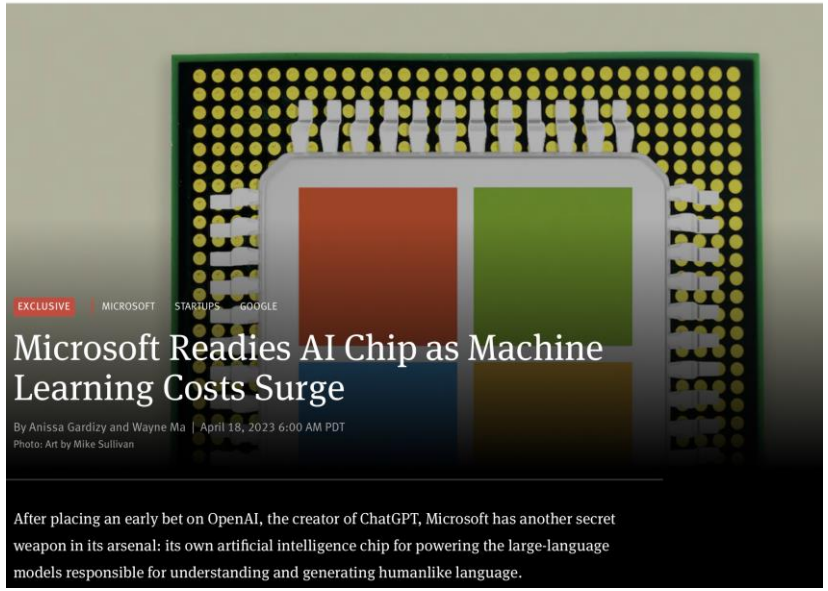
资料来源：英伟达官网，华西证券研究所

2.2 全球厂商开启AI芯片“军备赛”

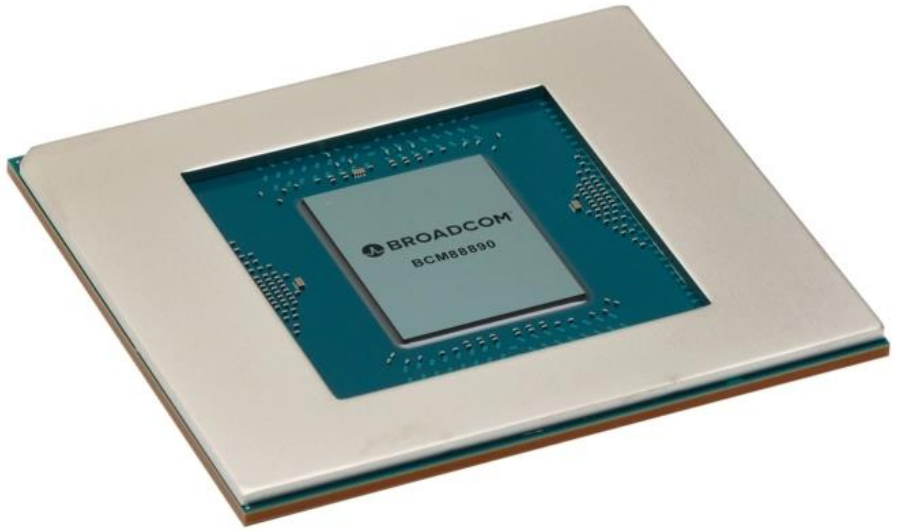
- **微软拟推出人工智能芯片：**根据IT之家4月18日消息，微软准备推出人工智能芯片，代号“雅典娜”，为大型语言模型提供动力。据悉，微软已将芯片提供给一小部分微软与OpenAI员工，他们正测试这项技术。微软希望这种芯片比目前从其他供应商处采购的芯片性能更好。报道指出，**亚马逊、谷歌和Facebook等科技巨头也在开发自家内部芯片。**

- **博通发布用于连接AI超级计算机的Jericho3-AI芯片：**根据财联社4月18日消息，博通公司发布芯片新品Jericho3-AI芯片，用于将超级计算机连接在一起，利用已广泛使用的网络技术进行人工智能工作。据介绍，这款芯片可将多达3.2万个GPU芯片连接在一起，将与另一种名为无限带宽（InfiniBand）的超级计算机网络技术竞争。而目前无限带宽设备的最大制造商为英伟达。

The Information报道微软在研发人工智能芯片



搭载Jericho3AI的BCM88890交换元件

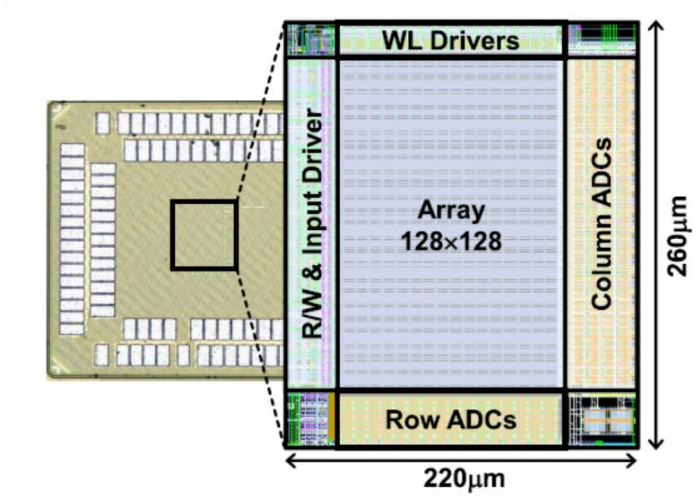


资料来源：IT之家，The Information，博通公司官网，华西证券研究所

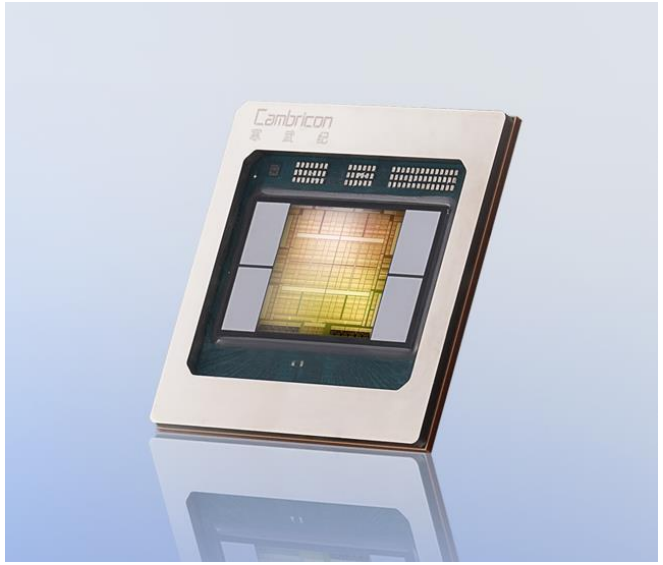
2.3 国内相关企业加速AI芯片布局

- **北京大学人工智能研究院积极布局AI芯片：**根据北京大学新闻网，人工智能研究院在面向边缘AI的可转置存内计算芯片方向研究取得重要进展，该研究团队提出了一种可同时实现高能效模型推断前馈计算与训练反向传播计算的可转置存内计算电路设计；基于28nm标准CMOS工艺完成了可转置存内计算电路的芯片原型验证，芯片在前馈计算时的能效和反向传播计算时的能效均达到世界先进水平。
- **寒武纪第五代智能处理器与第二代高档云端智能芯片正处于研发阶段：**第五代智能处理器架构及第二代高档云端智能芯片正在研发中，属于国际先进水平。公司研发目标为持续提高智能处理器架构的先进性，提高智能处理器 IP 的性能和能效。第二代高档云端智能芯片拟具备充裕的峰值运算能力，支持多芯片间交互，以支持分布式训练；适用于多样化的人工智能训练任务，可应用于互联网、智能计算中心等领域。
- **百度第三代昆仑芯有望在2024年实现量产：**根据科创板日报消息，百度云智一体3.0的AI IaaS层核心的昆仑芯已量产数万片，实现大规模商业化落地，昆仑芯3代将于2024年初量产。

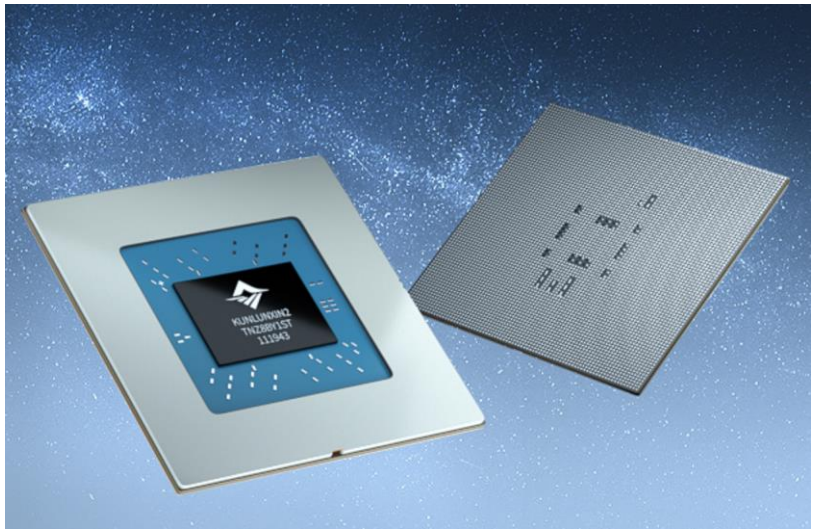
28nm芯片原型显微图



寒武纪寒武纪首颗AI训练芯片思元290



昆仑芯2代AI芯片



资料来源：北京大学新闻网，寒武纪半年报，财联社，昆仑芯官网，华西证券研究所

2.3 国内相关企业加速AI芯片布局

◆ 此外，我国AI芯片已经呈现百柯争流的情况，国产化AI芯片势在必行，相关厂商积极加速推进AI芯片布局，相关产品已经应用在云端训练和推理、机器视觉、自然语言处理、编辑氨基酸、智能驾驶、感存算一体化等场景。

相关重点AI芯片厂商及应用场景

公司名称	主要产品	产品类型	应用场景	竞争优势
寒武纪	思元290/270/370	云端训练和推理	通用型云端训练和边缘/终端推理AI方案	AI核心技术和人才团队优势；同时为云端、边缘端、终端提供全品类系列化智能芯片和处理器产品的能力。
燧原科技	邃思AI训练和推理芯片	云端训练和推理	面向数据中心的高性能云端训练和云端推理应用场景	国内首家同时拥有第二代高性能云端训练和云端推理产品线的公司。
昆仑芯	昆仑芯1代和2代芯片	云端训练和推理	互联网、智慧城市、智算中心、智慧工业、智慧应急、智慧交通、智慧金融等	大规模落地验证、工程化经验积累；深刻的场景理解、全方位的产品视角。
平头哥	含光800 NPU	云端训练和推理	阿里云平台、电商智能搜索	依托阿里平台，为阿里云提供AI计算能力。
沐曦	GPU/AI芯片	云端训练和推理	AI训练、AI 推理、数据中心、科学计算、云游戏和元宇宙等多个前沿领域	拥有顶配全建制团队，丰富GPU量产经验，完整软件生态能力和大量自主创新专利等四大核心竞争优势。
华为海思	Ascend 310/910	边缘计算AI	面向数字中心、边缘、消费终端和IoT场景	Ascend 310采用华为自研达芬奇架构NPU，以高性能3D Cube计算引擎为基础，大幅提高单位功耗下的AI算力。
紫光展锐	T820	边缘计算AI	工业、商业、医疗、家居、教育等	T820可支持AI降噪、AI-HDR、AI-Bokeh景深虚化技术、Face ID解锁、AI人脸检测等丰富AI影像功能
杭州国芯	GX8010	边缘计算AI	智能车载、智能音箱、智能家居、智能穿戴等	针对人工智能与物联网的特点，用于语音信号处理的DSP处理器等模块，兼具图像处理能力，具有高智能、低功耗全集成等特点。
北京君正	AI协处理器T02	视觉AI	深度学习的人形、人脸、车牌的检测和识别	深度学习算法更高效，在复杂环境如遮挡、大角度等场景下更准确，解决CV算法的痛点。
清微智能	TX510	视觉AI	智能安防、智能家居、机器人、航空航天等	具有按需即时重构、高能效、低功耗、通用性特点，可重构计算是后摩尔时代的颠覆性技术之一，清微是第一家将该技术大规模商用的公司。
九天睿芯	ADA 100/200/300	感存算AI	工业、安防、消费品、XR眼动追踪、视觉辨识机器人、自动驾驶	自主创新的“感存算一体”芯片架构是由ASP+ADA两部分组成。ASP可以在模拟信号端直接进行信号的特征分析和提取。
灵汐科技	类脑芯片KA200	感存算AI	脑科学及脑仿真领域	集成30个类脑计算核，各核可独立运行，支持矢量图计算。
千芯科技	可重构存算AI芯片	感存算AI	自然语言处理、医药计算、工业视觉、自动驾驶、智慧城市等	通过自研存算一体技术，可提供能效比超过10-100TOPS/W，优于其他类型AI芯片10-40倍的算力支持。
后摩智能	存算一体AI芯片	感存算AI	智能驾驶、泛机器人、无人车等边缘AI应用场景	与传统冯·诺依曼架构下的大算力芯片相比，后摩智能的存算一体芯片在算力、能效比和成本等方面，都能体现出显著的优势。
云天励飞	DeepEye 2000	安防/人脸识别AI	智能安防、新商业、智慧交通、智能制造智能超算等	芯片具有可编程、高效率、智升级等特点。
地平线	旭日3/征程5	ADAS/自动驾驶AI	汽车ADAS/自动驾驶、AIoT边缘计算	中国唯一实现车规级AI芯片前装量产的企业。征程系列AI芯片出货量已经超过百万，与众多主机厂实现前装量产合作。

资料来源：电子工程专辑，华西证券研究所

03 投资建议: 梳理AIGC相关受益厂商

3.1 投资建议: 梳理AIGC的受益厂商

◆ 我们认为AIGC的出世会产生革命性的影响，同时有望赋能千行百业。**我们梳理了五条路径图，积极的推荐以下五条投资主线:**

✓ **1) AI芯片厂商，**相关受益标的为: **赛武纪、海光信息、景嘉微、龙芯中科**等；

✓ **2) 存储厂商，**相关受益标的为: **东芯股份、兆易创新、澜起科技、聚辰股份、普再股份、江波龙、佰维存储、恒烁股份**等;

✓ **3) 光模块厂商，**相关受益标的为: **新易盛、中际旭创、天孚通信、剑桥科技、源杰科技、联特科技、光迅科技**等；

✓ **4)服务器及IDC厂商，**相关受益标的为: **浪潮信息、中科曙光、神州数码、拓维信息、工业富联、润泽科技**等；

✓ **5) AI云厂商，**相关受益标的为: **首都在线、云赛智联、青云科技、优刻得、光环新网、新炬网络**等。

资料来源：WIND，华西证券研究所

3.1 投资建议：梳理AIGC的受益厂商

AIGC的A股受益标的

公司名称	股票代码	收盘价	市值(亿元)	EPS(元)			PE(倍)		
		2023/4/19	2023/4/19	2021	2022E	2023E	2021	2022E	2023E
寒武纪	688256.SH	221.80	919.63	-2.06	-2.79	-1.80	-	-	-
海光信息	688041.SH	90.55	2104.69	0.16	0.40	0.59	560.3	225.8	152.4
景嘉微	300474.SZ	112.51	512.11	0.97	0.65	0.93	116.0	173.9	120.7
龙芯中科	688047.SH	159.21	638.43	0.66	0.43	0.59	241.2	369.9	270.4
东芯股份	688110.SH	39.03	172.61	0.77	0.79	0.87	50.7	49.3	44.6
兆易创新	603986.SH	130.75	872.14	3.54	3.69	3.95	36.9	35.4	33.1
澜起科技	688008.SH	72.10	819.11	0.73	1.17	1.62	98.8	61.5	44.4
聚辰股份	688123.SH	101.35	122.54	0.90	3.16	4.41	112.6	32.1	23.0
普冉股份	688766.SH	188.50	95.61	9.64	5.82	2.59	19.6	32.4	72.8
江波龙	301308.SZ	102.00	421.12	2.73	0.18	1.00	37.4	570.8	101.9
恒烁股份	688416.SH	72.45	59.87	2.43	0.29	1.00	29.8	249.5	72.1
新易盛	300502.SZ	77.01	390.51	1.31	1.87	2.05	58.8	41.2	37.6
中际旭创	300308.SZ	75.62	605.69	1.21	1.50	1.83	62.5	50.3	41.3
天孚通信	300394.SZ	64.38	253.64	0.79	1.05	1.29	81.7	61.5	49.8
源杰科技	688498.SH	261.96	158.75	2.12	1.81	2.57	123.6	144.9	102.1
联特科技	301205.SZ	102.28	73.72	1.96	1.90	2.39	52.2	53.8	42.9
光迅科技	002281.SZ	28.72	224.87	0.85	0.81	0.94	33.8	35.4	30.6
浪潮信息	000977.SZ	41.30	604.51	1.38	1.65	1.85	30.0	25.0	22.3
中科曙光	603019.SH	49.74	728.20	0.80	1.05	1.37	62.2	47.2	36.4
神州数码	000034.SZ	30.34	202.97	0.37	1.54	1.86	82.3	19.7	16.3
拓维信息	002261.SZ	15.30	192.14	0.07	-0.04	0.15	218.6	-	101.0
工业富联	601138.SH	17.27	3430.32	1.01	1.11	1.19	17.1	15.6	14.5
润泽科技	300442.SZ	69.95	667.23	-0.12	1.30	1.95	-	53.7	35.8
首都在线	300846.SZ	18.34	85.62	0.05	-0.40	0.16	344.7	-	115.7
云赛智联	600602.SH	11.03	130.65	0.19	0.24	0.16	58.7	46.6	68.2
光环新网	300383.SZ	12.64	227.22	0.54	0.36	0.47	23.4	35.3	26.9
新炬网络	605398.SH	37.23	31.01	1.19	0.89	0.85	31.3	41.9	43.7

注：*来自wind一致预测

资料来源：WIND，华西证券研究所

3.2.1 首都在线: AI算力云龙头，AIGC“挖井人”

- **公司绑定英伟达、燧原，AI云开启第二波成长曲线。**公司已摆脱单一的IaaS 公有云，重点转向AI算力云转型，有望借助底层英伟达GPU算力储备，以AI云为抓手，开启第二波成长曲线，我们认为算力网络以及边缘节点是公司AI云的核心壁垒之一。公司首云星图云算力平台已经震撼发布，深度绑定英伟达，算力平台采用A100、A40、A5000，为全球数字世界多场景提供澎湃算力。同时，公司携手燧原科技，开启AIGC芯征程，重点针对大模型 MaaS开展联合攻关，正式推出云燧i20支撑的AIGC实时推理应用。
- **海外游戏具备竞争优势，AI算力云赋能千行百业。**我们认为公司AI云平台产品发布与公司底层算力储备密不可分，借助通过算力、网络、存储等核心能力构建“云-网-数”一体的边缘计算平台，就近为高算力业务场景如**云游戏、AI、XR、数字人、数字孪生、元宇宙、智能制造**等各领域提供了算力支持。其中云游戏方面，我们判断云游戏市场处于“技术成熟走向商业可行”与“商业可行走向商业腾飞”的交替阶段，公司坐拥算力和算力网络双重竞争优势。此外，公司传统IDC和云业务积极布局海外，也将会是公司另一个业绩爆发点。

公司星图云底层算例示意图

公司与遂原科技合作示意图

资料来源：公司官网，招股说明书，华西证券研究所

3.2.2 浪潮信息：中国服务器/AI服务器市占率稳居榜首

- **浪潮信息是全球领先的新型IT基础架构产品、方案及服务提供商**：公司是全球领先的 AI 基础设施供应商，拥有业内最全的人工智能计算全堆栈解决方案，涉及训练、推理、边缘等全栈 AI 场景，构建起领先的 AI 算法模型、AI 框架优化、AI 开发管理和应用优化等全栈 AI 能力，为智慧时代提供坚实的基础设施支撑。
- **公司算力技术壁垒浓厚**：生产算力方面，公司拥有业内最强最全的 AI 计算产品阵列，业界性能最好的Transformer 训练服务器 NF5488、全球首个 AI 开放加速计算系统 MX1、自研 AI 大模型计算框架 LMS。聚合算力层面，公司针对高并发训练推理集群进行架构优化，构建了高性能的NVMe 存储池，深度优化了软件栈，性能提升 3.5 倍以上。调度算力层面，浪潮信息 AIstation 计算资源平台可支持 AI 训练和推理，是业界功能最全的 AI 管理平台；同时，浪潮信息还有自动机器学习平台 AutoML Suite，可实现自动建模，加速产业化应用。

浪潮信息智算中心

浪潮信息智算中心

资料来源：公司官网，华西证券研究所

3.2.3 中科曙光：我国高性能计算、智能计算领军企业

- **中科曙光作我国核心信息基础设施领军企业**：在高端计算、存储、安全、数据中心等领域拥有深厚的技术积淀和领先的市场份额，并充分发挥高端计算优势，布局智能计算、云计算、大数据等领域的技术研发，打造计算产业生态，为科研探索创新、行业信息化建设、产业转型升级、数字经济发展提供了坚实可信的支撑。

- **依托先进计算领域的先发优势和技术细节，中科曙光全面布局智能计算**：完成了包括AI核心组件、人工智能服务器、人工智能管理平台、软件等多项创新，构建了完整的AI计算服务体系。并积极响应时代需求，在智能计算中心建设浪潮下，形成了5A级智能计算中心整体方案。目前，曙光5A智能计算中心已在广东、安徽、浙江等地建成，江苏、湖北、湖南等地已进入建设阶段，其他地区也在紧张筹备和规划中。

中科曙光主要产品

通用服务器
机架式服务器
高密度服务器
刀片服务器
核心应用服务器

智能计算服务器
深度学习训练
智能应用推理

终端&工作站
微型计算机
工作站

高性能计算机
通用高性能计算机
高性能计算机系统组件
高性能计算机的服务支撑

机房冷却设施
微模块产品
液冷基础设施产品

存储产品
分布式统一存储
多控统一存储
高密度存储服务器
备份一体机

网络安全产品
数据中心安全产品
汇聚分流设备
智能加速卡
网络内容识别分析系统
网络态势感知系统

大数据平台软件
大数据智能引擎系列
数据工程服务系列
视频智能分析系列
大数据与人工智能实训平台

云计算平台软件
云计算操作系统
超融合一体机
云桌面
云容灾

计算服务
弹性计算服务
混合计算服务
专有计算服务
API
托管、运营

云计算服务
云服务器 ECS
裸金属 BMS
对象存储 OSS
云容器实例 CCI
人工智能服务
数据开发 DDS
数据治理中心 DGS
数据服务 DSS
数据可视化 DAV
数据集成 Data Integration

城市云
智慧城市
国资云
交通云
医疗云

5A级智算中心

中科曙光硅立方液体相变冷却计算机

资料来源：公司官网，华西证券研究所

3.2.4 神州数码: 华为生态核心践行者

- **神州数码领先的数字化转型**：神州数码围绕企业数字化转型的关键要素，开创性的提出“数云融合”战略和技术体系框架，着力在云原生、数字原生、数云融合关键技术和信创产业上架构产品和服务能力，为处在不同数字化转型阶段的快消零售、汽车、金融、医疗、政企、教育、运营商等行业客户提供泛在的敏捷IT能力和融合的数据驱动能力。
- **神州数码为华为生态核心践行者**：公司旗下的神州鲲泰基于华为鲲鹏处理器多款不同种类的服务器产品，包括1、单路服务器：R222、R224；2、双路服务器：R522、R524、R722、R724、R2240、R2260、R2280。3、四路服务器：R822。此外，公司基于华为鲲鹏920处理器与昇腾Atlas AI加速卡，神州数码开发了采用ARM架构的一系列AI服务器。

神州数码服务器及相关参数

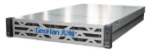
名称	示意图	形态	处理器	内存支持	AI加速卡/AI处理器	AI算力
KunTai A222		2U单路边缘机架式服务器	1*鲲鹏920处理器，24核，主频2.6GHz	4个DDR4 RDIMM，最高速率3200MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能；单根内存条容量支持16GB/32GB/64GB/128GB	最大支持3张Atlas 300V 视频解析卡或Atlas 300I Pro 推理卡或Atlas 300V Pro 视频解析卡	最大420 TOPS INT8
KunTai A722		2U 双路推理型 AI 机架式服务器	2*鲲鹏920处理器，支持32、48、64核可选，主频2.6GHz	16个或32个DDR4 RDIMM，最高速率2933MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能；单根内存条容量支持16GB/32GB/64GB/128GB	最大支持8张，Atlas 300V 视频解析卡或Atlas 300I Pro 推理卡或Atlas 300V Pro 视频解析卡	最大1120 TOPS INT8
KunTai A924		4U四路训练型AI机架式服务器	4*鲲鹏920处理器，支持48核，主频 2.6GHz	支持32个DDR4内存插槽，速率最高2933MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能；单根内存条容量支持32GB/64GB/128GB	8*昇腾910，支持直出100G RoCE网络接口	最大512Tops Int8或256Tops FP16

资料来源：公司官网，华西证券研究所

3.2.5 拓维信息: 华为生态重要参与者

- **拓维信息是领先的软硬一体化解决方案提供商:** 公司1996年成立，业务涵盖政企数字化、智能计算、鸿蒙生态，覆盖全国31个省级行政区、海外10+国家，聚焦数字政府、运营商、考试、交通、制造、教育等重点领域和行业，服务超过1500家政企客户，为其提供全栈国产数字化解决方案和一站式全生命周期的综合服务。
- **拓维信息为华为生态重要参与者**：**“兆瀚”系列通用服务器**是基于ARM架构，搭载鲲鹏920处理器设计开发的机架式型服务器，拥有高的性能、可靠性、高效环保、兼容性强等特点；**“兆瀚”系列AI服务器**能够满足当前各类主流AI场景与AI大模型的训练需求，已经在国内多个区域人工智能计算中心、城市人工智能中枢、通用AI服务器场景中得到了应用，已经在国内多家头部互联网企业开展适配测试。

拓维信息旗下“兆瀚”系列服务器产品介绍

种类	名称	示意图	形态	处理器	内存支持	AI加速卡/AI处理器	AI算力
通用服务器	兆瀚RH220系列		2U双路机架	支持两颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率180w。	最多支持32个DDR4内存DIMM插槽，最高速率2933MT/s	/	/
	兆瀚RH520系列		4U机架服务器	支持两颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率180w。	最多支持32个DDR4内存DIMM插槽，最高速率2933MT/s	/	/
AI服务器	兆瀚RA2300-A		2U推理服务器	支持两颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率180w。	最多支持32个DDR4内存DIMM插槽，最高速率2933MT/s	支持Atlas 300I Pro推理卡和Atlas 300V Pro视频解析卡	最大1.12 POPS INT8；最大560 TFLOPS PF16
	兆瀚SA300		2U智能边缘服务器	支持一颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率181w。	最多支持4个DDR4内存DIMM插槽，最高速率2934MT/s	支持Atlas 300I Pro推理卡/Atlas 300V Pro 视频解析卡	最大420 TOPS INT8 或 384路1080P 30 FPS视频解析（硬件解码能力）
	兆瀚RA5900-A		4U训练服务器	支持四颗华为鲲鹏920处理器，CPU主频2.6GHz。单CPU最多64个内核，最大功率182w。	最多32个DDR4内存插槽，支持RDIMM。单根内存条容量支持32 GB/64GB	8*昇腾910	/
	兆瀚RA2302-B		2U AI 服务器	2*64核青松处理器	32个DDR4内存插槽，最高3200 MT/s，支持ECC	最大支持4个Atlas 300I/V Pro	最大560 TPOS INT8

资料来源：公司官网，华西证券研究所

3.2.6 海光信息：支持全精度，GPU实现规模量产

- **海光信息主要从事高端处理器、加速器等计算芯片产品和系统的研究、开发，主要产品包括海光CPU和海光DCU**：2018年10月，公司启动深算一号DCU产品设计，海光8100采用先进的FinFET工艺，典型应用场景下性能指标可以达到国际同类型高端产品的同期水平。2020年1月，公司启动DCU深算二号的产品研发。

- **海光DCU性能强大**：海光DCU基于大规模并行计算微结构进行设计，不但具备强大的双精度浮点计算能力，同时在单精度、半精度、整型计算方面表现同样优异，是一款计算性能强大、能效比较高的通用协处理器。海光DCU集成片上高带宽内存芯片，可以在大规模数据计算过程中提供优异的数据处理能力。

海光信息主要产品

系列	7000系列CPU	5000系列CPU	3000系列CPU
核心规格	最大32个物理核心	最大16个物理核心	最大8个物理核心
应用领域	高端通用服务器、先进计算系统	通用服务器	个人工作站、工控设备等终端产品

系列	8000系列DCU
核心规格	60-64个深度计算单元
应用领域	先进计算系统、人工智能

海光深算一号性能达到国际同类产品水平

项目	海光	NVIDIA	AMD
品牌	深算一号	Ampere 100	MI100
生产工艺	7nm FinFET	7nm FinFET	7nm FinFET
核心数量	4096 (64CUs)	2560 CUDA processors 640 Tensor processors	120CUs
内核频率	Up to 1.5GHz（FP64） Up to 1.7Ghz（FP32）	Up to 1.53Ghz	Up to 1.5GHz（FP64） Up to 1.7Ghz （FP32）
显存容量	32GB HBM2	80GB HBM2e	32GB HBM2
显存位宽	4096 bit	5120 bit	4096bit
显存频率	2.0 GHz	3.2 GHz	2.4 GHz
显存带宽	1024 GB/s	2039 GB/s	1228 GB/s
TDP	350 W	400 W	300 W
CPU to GPU 互联	PCIe Gen4 x 16	PCIe Gen4 x 16	PCIe GEN4 x 16
GPU to GPU 互联	xGMI x 2, Up to 184 GB/s	NVLink up to 600 GB/s	Infinity Fabric x 3， up to 276 GB/s

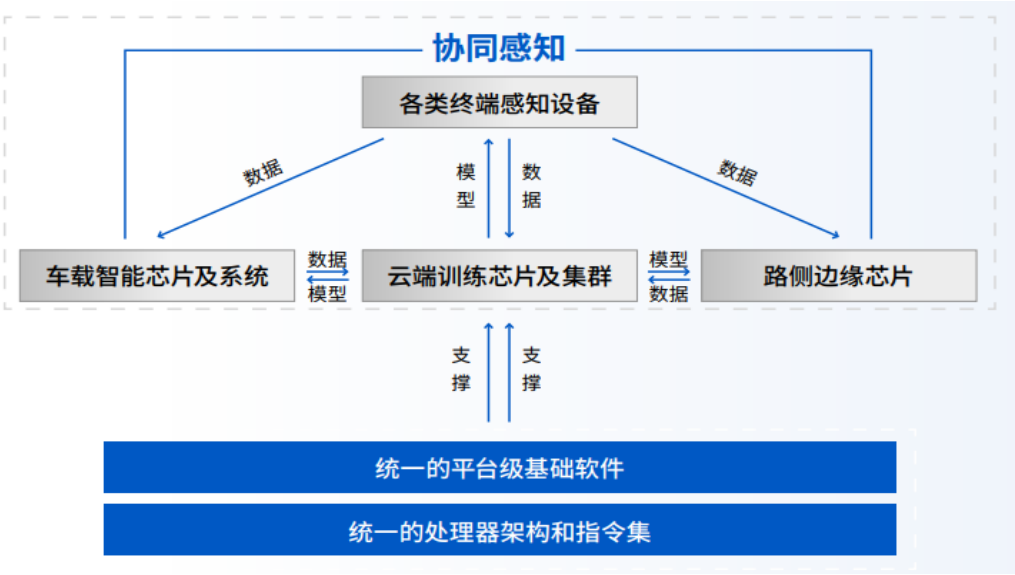
资料来源：公司公告，华西证券研究所

3.2.7 寒武纪：少数全面掌握AI芯片技术的企业之一

- **寒武纪是目前国际上少数几家全面系统掌握了通用型智能芯片及其基础系统软件研发和产品化核心技术的企业之一**：寒武纪主营业务是应用于各类云服务器、边缘计算设备、终端设备中人工智能核心芯片的研发和销售。公司的主要产品包括终端智能处理器IP、云端智能芯片及加速卡、边缘智能芯片及加速卡以及与上述产品配套的基础系统软件平台。

- **公司AI技术积累浓厚**：能提供云边端一体、软硬件协同、训练推理融合、具备统一生态的系列化智能芯片产品和平台化基础系统软件。2022年3月，寒武纪正式发布了新款训练加速卡“MLU370-X8”，搭载双芯片四芯粒封装的思元370，集成寒武纪MLU-Link多芯互联技术，主要面向AI训练任务。

寒武纪“云边端车”协同



寒武纪产品技术图谱

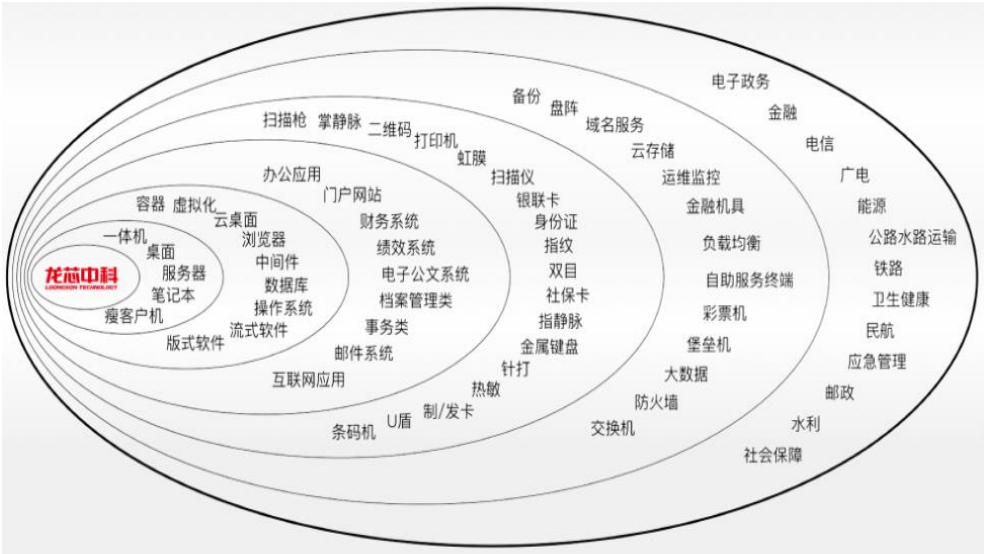
资料来源：公司官网，华西证券研究所

3.2.8 龙芯中科：2K2000系列集成自主GPU

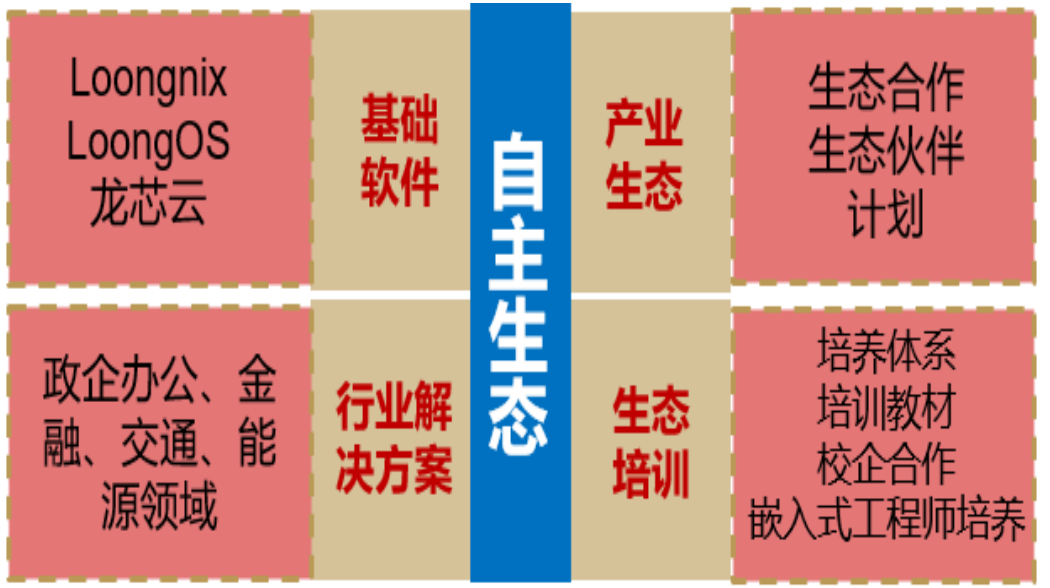
- **龙芯中科主营业务为处理器及配套芯片的研制、销售及服务：** 主要产品与服务包括处理器及配套芯片产品与基础软硬件解决方案业务。公司基于信息系统和工控系统两条主线开展产业生态建设，面向网络安全、办公与业务信息化、工控及物联网等领域与合作伙伴保持全面合作，产品在电子政务、能源、交通、金融、电信、教育等行业领域已获得广泛应用。

- **公司自主研发2K200系列GPU：**2022年12月，龙芯2K2000完成了初步功能调试及性能测试，达到其设计目标，2023年将推出试用。龙芯2K2000集成了两个LA364处理器核，典型工作频率为1.5GHz，共享2MB的L2缓存，SPEC2006INT（base）单核定/浮点分值达到13.5/14.9分。龙芯2K2000芯片集成了龙芯自主研发的GPU，并优化了图形算法和性能。

龙芯中科生态合作示意图



龙芯中科自主生态



资料来源：公司官网，华西证券研究所

3.2.9 景嘉微：新一代JM9系列有望打开商用市场

- **国产GPU龙头企业:** 公司成立于2006年，主要从事军用电子产品的研发、生产、销售，目前形成了三大业务板块分别是图形线控模块、小型专用雷达和芯片业务。GPU方面，2014年首推JM5400实现了军用GPU的国产替代；第二款芯片JM7200于2018年研发成功，具备了PC端的功能；日前，公司9系列芯片研发成功，具备高性能计算能力。

- **新一代JM9系列有望打开商用市场**：日前，公司JM9系列图形处理芯片已顺利发布，应用领域涵盖地理信息系统、媒体处理、CAD辅助设计、游戏、虚拟化等高性能显示和人工智能计算领域。目前，信创市场为公司提供了新的业务增长点，JM9系列图形处理芯片的成功发布将为公司未来进一步拓展通用市场提供强有力的产品支撑。

景嘉微GPU系列产品

景嘉微7系列GPU示意图

资料来源： 公司官网，华西证券研究所

3.2.10 东芯股份：多类别存储芯片助力企业具备先发优势

- **东芯股份拥有独立自主的知识产权，努力打造中国领先的存储设计企业：**公司是一家Fabless（无晶圆厂）芯片企业，主要从事芯片设计和销售业务，将晶圆制造、封装测试等生产环节委托第三方完成，拥有独立自主的知识产权，聚焦于中小容量NAND/NOR/DRAM芯片的研发、设计和销售，是目前国内少数可以同时提供NAND/NOR/DRAM设计工艺和产品方案的存储芯片研发设计公司。公司愿景是成为中国领先的存储设计企业，使命为提供可靠高效的存储产品及设计方案。

- **担当本土存储“芯”使命，研发前瞻性产品—存算一体化芯片&DTR NAND：**东芯半导体持续研发和优化产品性能，致力于研发1×nm NAND Flash芯片，聚焦于高附加值产品，研发前瞻性产品：存算一体化芯片&DTR NAND。东芯半导体在已有的多类别存储技术的基础上，叠加新的研发方案，在存算一体布局中具有很高的先发优势。

存储产品之间差异性

比较项目	非易失性		易失性
	NAND Flash	NOR Flash	DRAM
存储原理	浮栅型	浮栅型/电子俘获型	电容充放电型
读取速度	较慢	较快	极快
擦除/写入速度	快	较慢	极快
存储容量	高（Gb/Tb）	中（Mb/Gb）	中（Mb/Gb）
擦写次数	十万级别	十万级别	-

公司产品工艺流程

资料来源：公司官网，公司招股书，华西证券研究所

3.2.11 兆易创新：身处“集成电路设计”行业的IC设计企业

- **在整个产业链中处于重要地位并拥有核心竞争力：**兆易创新成立于2005年4月，主要业务为存储器、微控制器和传感器的研发、技术支持和销售。公司产品广泛应用于工业、消费类电子、汽车、物联网、计算、移动应用以及网络和电信行业等各个领域，助力社会智能化升级。公司作为 IC 设计企业，自成立以来一直采取 Fabless 模式，专注于集成电路设计、销售和客户服务环节，将晶圆制造、封装和测试等环节外包给专门的晶圆代工、封装及测试厂商。

- **作为全球化芯片设计公司，兆易创新致力于存储器、控制器及周边产品的设计研发。**公司存储器产品包括：闪存芯片（NOR Flash、NAND Flash）和动态随机存取存储器（DRAM），公司以存储为主，控制器及周边产品为辅，多赛道多产品线的组合布局多元化布局助力穿越周期影响，技术和产品优势不断增强。

存储器主要参数

Model	Length	Width	Height(Max)	Pitch
USON6 1.2*1.2 mm	1.20mm	1.20mm	0.40mm	0.40mm
USON8 1.5*1.5 mm	1.50 mm	1.50 mm	0.50 mm	0.40 mm
USON8 3*2 mm (0.45mm)	3.00 mm	2.00 mm	0.50 mm	0.50 mm
USON8 3*4 mm	3.00 mm	4.00 mm	0.60 mm	0.80 mm
USON8 4*4 mm	4.00 mm	4.00 mm	0.50 mm	0.80 mm

存储器产品介绍

英文名称	中文名称	介绍	应用
NOR Flash	代码型闪存芯片	主要用来存储代码及少量数据	公司 NOR Flash 产品广泛 应用于物联网、工业及汽车电子、穿戴式设备、人工智能、网络通信、安防监控产品、PC 主板、 移动设备、数字机顶盒、路由器、家庭网关等领域
NAND Flash	数据型闪存芯片	分为两大类：大容量 NAND Flash 主要为 MLC、TLC 2D NAND或 3D NAND，擦写次数从几百次至数千次，多应用于大容量数据存储；小容量 NAND Flash 主要是 SLC 2D NAND，可靠性更高，擦写次数达到数万次以上。	公司 NAND Flash 产品属 于SLC NAND，为移动设备、机顶盒、数据卡、电视、汽车电子等设备的多媒体数据存储应用提供所必需的大容量存储。
DRAM	动态随机存取存储器	是当前市场中最为重要的系统内存，在计算系统中占 据核心位置，广泛应用于服务器、移动设备、PC、消费电子等领域。因极高的技术和资金壁垒，DRAM 领域市场处于高度集中甚至垄断态势。	公司首款自有品牌 DRAM 产品已于 2021 年 6 月推 出，实现了从设计、流片，到封测、验证的全国产化，在满足消费类市场强劲需求的同时，助力 国产自主供应生态圈的发展构建。该产品主要面向消费类、工业控制类及汽车类等市场领域，应 用于机顶盒、电视、监控、网络通信、智慧家庭、平板电脑、车载影音系统等诸多领域

资料来源：公司官网，华西证券研究所

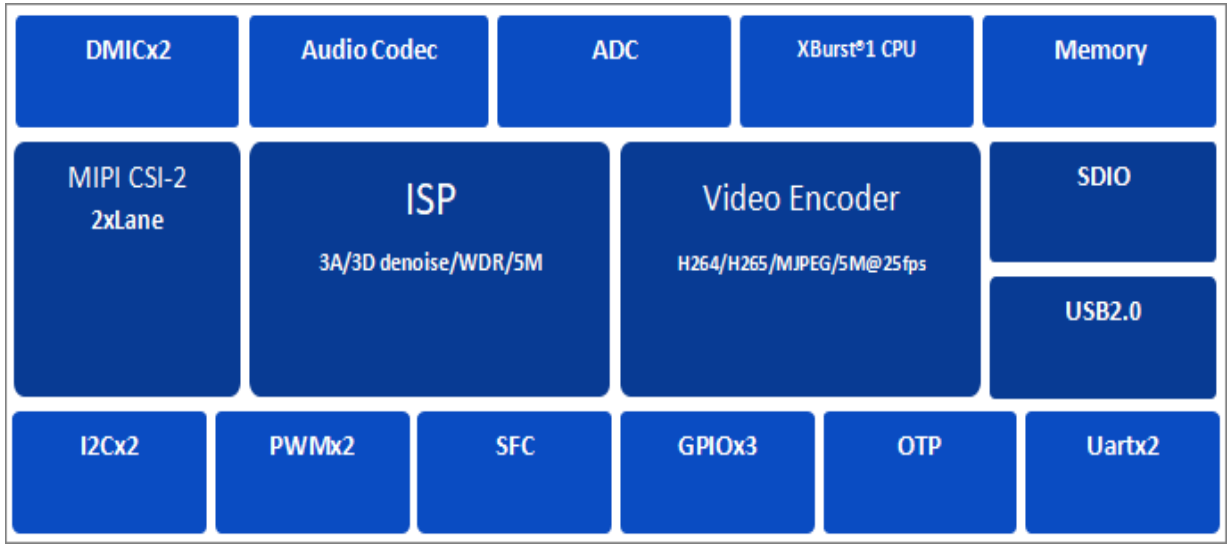
3.2.12 北京君正：多年深耕自主创新

- **公司拥有较强的自主创新能力**：多年来在自主创新CPU技术、视频编解码技术、图像和声音信号处理技术、SoC芯片技术、软件平台技术等多个领域形成多项核心技术。公司已形成可持续发展的梯队化产品布局，基于自主创新的XBurst CPU和视频编解码等核心技术，公司推出了一系列具有高性价比的微处理器芯片产品和智能视频芯片产品，各类别的芯片产品分别面向不同的市场领域。
- **2020年，君正完成对北京矽成（ISSI）及其下属子品牌Lumissil的收购，并拥有其100%股份。**其中，ISSI存储部门有高速低功耗SRAM，低中密度DRAM，NOR/NAND Flash，嵌入式Flash pFusion，及eMMC等芯片产品。模拟和互联部门Lumissil有LED驱动、触控传感、音频驱动、微处理器、电源管理和互联等芯片产品。

Xburst系列CPU Core参数

	XBurst1	XBurst2
Base ISA	MIPS32 R5	MIPS32 R5
SIMD Extension	MXU2.0 - 128bits SIMD	MSA128 - 128bits SIMD MXA128 - 128bits SIMD MXU3.0- 512bits SIMD
Micro-Architecture	9 stage pipeline Single issue	Dual-Issue In-Order 2 Hardware Threads SMT
Coremark	2.3	3.6 (single thread)
Power Consumption	0.07mW/MHz, 40nm	0.13mW/MHz, 28nm

芯片框图



资料来源：IDC，华西证券研究所

3.2.13 澜起股份: 业界领先的集成电路设计公司

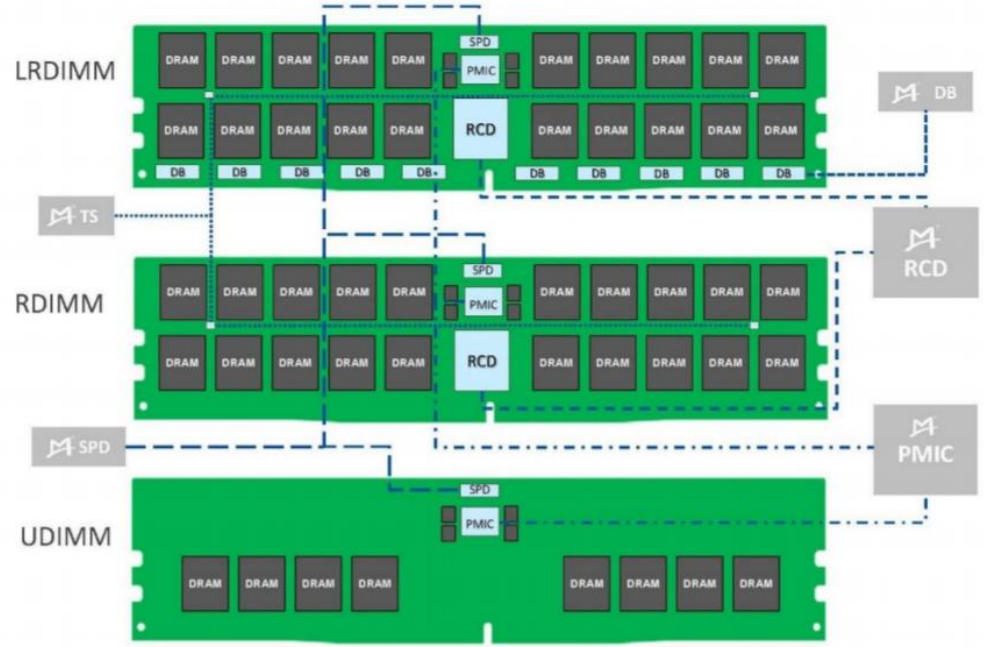
- **业界领先的集成电路设计公司，澜起股份为全球仅有的3家内存接口芯片供应商之一。**主要经营模式为Fabless模式，公司的两大产品线为互连类芯片产品线（主要包括内存接口芯片、内存模组配套芯片、PCIe Retimer 芯 片、MXC 芯片等）和津逮®服务器平台产品线（包括津逮®CPU 和混合安全内存模组（HSDIMM®））。同时，公司正在研发基于“近内存计算架构”的 AI 芯片。

- **公司发明的DDR4全缓冲“1+9”架构被采纳为国际标准。**现已成为全球可提供从DDR2到DDR4内存全缓冲/半缓冲完整解决方案的主要供应商之一，在内存接口芯片市场位列全球前二及内存模组配套芯片。

公司DDR4内存接口芯片自带产品及其应用情况

DDR4 内存接口芯片产品	应用
Gen1.0 DDR4 RCD 芯片	DDR4 RDIMM 和 LRDIMM，支持速率达 DDR4-2133
Gen1.0 DDR4 DB 芯片	DDR4 LRDIMM，支持速率达 DDR4-2133
Gen1.5 DDR4 RCD 芯片	DDR4 RDIMM 和 LRDIMM，支持速率达 DDR4-2400
Gen1.5 DDR4 DB 芯片	DDR4 LRDIMM，支持速率达 DDR4-2400
Gen2 DDR4 RCD 芯片	DDR4 RDIMM 和 LRDIMM，支持速率达 DDR4-2666
Gen2 DDR4 DB 芯片	DDR4 LRDIMM，支持速率达 DDR4-2666
Gen2 Plus DDR4 RCD 芯片	DDR4 RDIMM、LRDIMM 和 NVDIMM，支持速率达 DDR4-3200
Gen2 Plus DDR4 DB 芯片	DDR4 LRDIMM，支持速率达 DDR4-3200

公司DDR5内存接口芯片及内存模组配套芯片

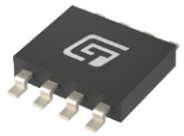


资料来源：公司年报，华西证券研究所

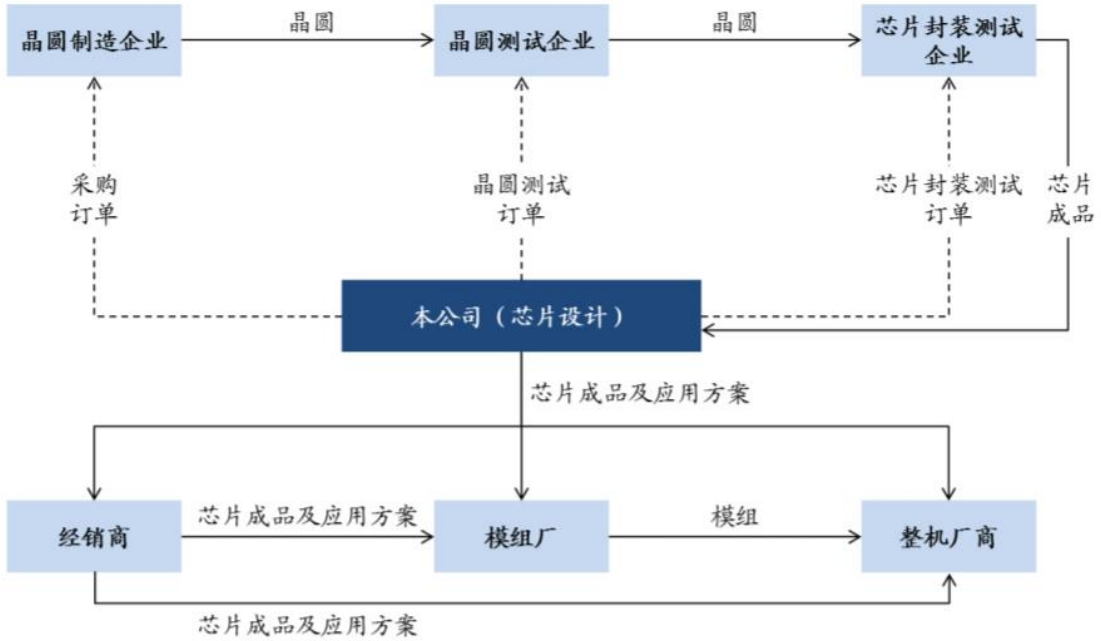
3.2.14 聚辰股份：十年“创芯”发展，精准洞悉市场新需求

- **作为一家全球化的芯片设计高新技术企业，**聚辰半导体长期致力于为客户提供存储、模拟和混合信号集成电路产品并提供应用解决方案和技术支持服务。**公司目前拥有EEPROM、音圈马达驱动芯片和智能卡芯片三条主要产品线**，产品广泛应用于智能手机、液晶面板、蓝牙模块、通讯、计算机及周边、医疗仪器、白色家电、汽车电子、工业控制等众多领域。
- **公司已在智能手机摄像头、液晶面板、 计算机及周边等细分领域奠定了领先优势**，未来公司将持续以市场需求为导向，以自主创新为驱动，对EEPROM、音圈马达驱动芯片、智能卡芯片等现有产品线进行完善和升级，并积极开拓NOR Flash、电机驱动芯片等新产品领域。

聚辰股份主要产品一览

产品种类	细分种类	型号	特性	产品示意图
EEPROM	传统应用领域	I2C系列	EEPROM（电可擦除可编程只读存储器）是一类通用型的非易失性存储芯片，在断电情况下仍能保留所存储的数据信息，可以在计算机或专用设备上擦除已有信息重新编程，耐擦写性能至少100万次。聚辰的EEPROM产品具有高可靠性、宽电压、高兼容性、低功耗等特点。公司EEPROM产品线包括I2C、SPI和Microwire等标准接口的系列EEPROM产品，以及主要应用于计算机和服务器内存条的SPD/SPD+TS（温度传感器）系列EEPROM产品。	
		SPI系列		
		Microwire系列		
		SPD/SPD+TS系列		
	车规级	A2	具备高可靠性和低失效率等优势，擦写次数最高可达400万次以上，其温度适应能力强，数据可存储100年。新产品SPI NOR Flash也按照车规等级125℃设计。产品主要应用于车载摄像头、液晶显示、娱乐系统等外围部件，并逐渐延伸到新能源车三电系统、车身控制、底盘传动及微电机、智能座舱等核心部件。	
		A1		
NOR Flash		1.65~3.6V	NOR Flash以其合适的容量、灵活的存取操作、及其非易失性产品特性，非常适合作为智能设备的指令程序存储器。随着5G、IOT、AMOLED、TDDI、TWS及汽车电子等应用市场快速发展，NOR Flash的需求保持持续的增长动力。聚辰半导体以其领先的存储器设计技术，推出SPI NOR Flash产品，可以覆盖从消费级，到工业级，直至汽车级的所有应用，产品在可靠性，功耗，温度和速度等关键性能指标方面的技术领先性。	
		2.7~3.6V		
		1.65~1.95V		
		1.1~2.0V		

聚辰股份整体业务流程



资料来源：IDC，华西证券研究所

04 风险提示

风险提示

- 核心技术水平升级不及预期的风险: AIGC相关产业技术壁垒较高，公司核心技术难以突破，进程低于预期，影响整体进度。
- AI伦理风险: AI可能会生产违反道德、常规、法律等内容。
- 政策推进不及预期的风险: 受到宏观经济、财政、疫情影响，政策推进节奏不及预期。
- 中美贸易摩擦升级的风险: 供应链存在部分海外提供商，容易受到美国”卡脖子”制裁，导致产品研发不及预期。

免责声明

分析师与研究助理简介

刘泽晶（首席分析师）2014-2015年新财富计算机行业团队第三、第五名，水晶球第三名，10年证券从业经验。

分析师承诺

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

评级说明

公司评级标准	投资评级	说明
以报告发布日后的6个月内公司股价相对上证指数的涨跌幅为基准。	买入	分析师预测在此期间股价相对强于上证指数达到或超过15%
	增持	分析师预测在此期间股价相对强于上证指数在5%—15%之间
	中性	分析师预测在此期间股价相对上证指数在-5%—5%之间
	减持	分析师预测在此期间股价相对弱于上证指数5%—15%之间
	卖出	分析师预测在此期间股价相对弱于上证指数达到或超过15%
行业评级标准		
以报告发布日后的6个月内行业指数的涨跌幅为基准。	推荐	分析师预测在此期间行业指数相对强于上证指数达到或超过10%
	中性	分析师预测在此期间行业指数相对上证指数在-10%—10%之间
	回避	分析师预测在此期间行业指数相对弱于上证指数达到或超过10%

华西证券研究所：

地址：北京市西城区太平桥大街丰汇园11号丰汇时代大厦南座5层

网址：http://www.hx168.com.cn/hxzq/hxindex.html

免责声明

华西证券股份有限公司（以下简称“本公司”）具备证券投资咨询业务资格。本报告仅供本公司签约客户使用。本公司不会因接收人收到或者经由其他渠道转发收到本报告而直接视其为本公司客户。

本报告基于本公司研究所及其研究人员认为的已经公开的资料或者研究人员的实地调研资料，但本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载资料、意见以及推测仅于本报告发布当日的判断，且这种判断受到研究方法、研究依据等多方面的制约。在不同时期，本公司可发出与本报告所载资料、意见及预测不一致的报告。本公司不保证本报告所含信息始终保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者需自行关注相应更新或修改。

在任何情况下，本报告仅提供给签约客户参考使用，任何信息或所表述的意见绝不构成对任何人的投资建议。市场有风险，投资需谨慎。投资者不应将本报告视为做出投资决策的惟一参考因素，亦不应认为本报告可以取代自己的判断。在任何情况下，本报告均未考虑到个别客户的特殊投资目标、财务状况或需求，不能作为客户进行客户买卖、认购证券或者其他金融工具的保证或邀请。在任何情况下，本公司、本公司员工或者其他关联方均不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告而导致的任何可能损失负有任何责任。投资者因使用本公司研究报告做出的任何投资决策均是独立行为，与本公司、本公司员工及其他关联方无关。

本公司建立起信息隔离墙制度、跨墙制度来规范管理跨部门、跨关联机构之间的信息流动。务请投资者注意，在法律许可的前提下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的前提下，本公司的董事、高级职员或员工可能担任本报告所提到的公司的董事。