ChatGPT: 加速计算服务器时代到来

AIGC行业深度报告(5)

华西计算机团队

2023年3月8日

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

核心逻辑:

◆ **大模型出现有望带动AI服务器需求爆发**

我们认为ChatGPT具备跨时代的意义的本质是AI算法大模型，因此科技巨头已经开始算力“军备赛”，大模型的出现有望带动AI服务器需求爆发。服务器架构随负载量扩张不断优化，已经经历传统单一部署与集群模式，目前正处于分布式模式的转变阶段。CPU、内部存储和外部存储是服务器的核心部件。

◆ **加速计算是服务器成长的核心驱动力**

按照CPU指令集架构的差异，服务器可分为CISC(复杂指令集)、RISC(精简指令集)、VLIM等架构，代表架构为X86。人工智能应用场景下的**加速计算服务器是中国服务器的核心驱动力**，AI服务器相较于通用服务器区别在于硬件架构、加速卡数量与设计方面；我们认为AI服务器众芯片组为服务器的核心，且价值成本占比较高。

◆ **算力时代到来，服务器价值再次凸显**

我们认为服务器是“伴科技类”的硬件产品，随着科技的服务形式和应用方式不断进步，服务器同样在不断迭代升级或更新换代，近年来随着互联网+、云计算、AI+、边缘计算的出现，服务器市场迎来了极大的发展；根据IDC的数据显示，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，而AI服务器作为算力载体为数字经济时代提供广阔动力源泉，更加凸显其重要性。

◆ **投资建议:**

关注两条投资主线：

1)AI服务器生产商，重点推荐中科曙光，其他受益标的为浪潮信息、拓维信息、神州数码；

2)具备算力芯片的厂商，受益标的为寒武纪、海光信息、龙芯中科、景嘉微。

◆ **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。

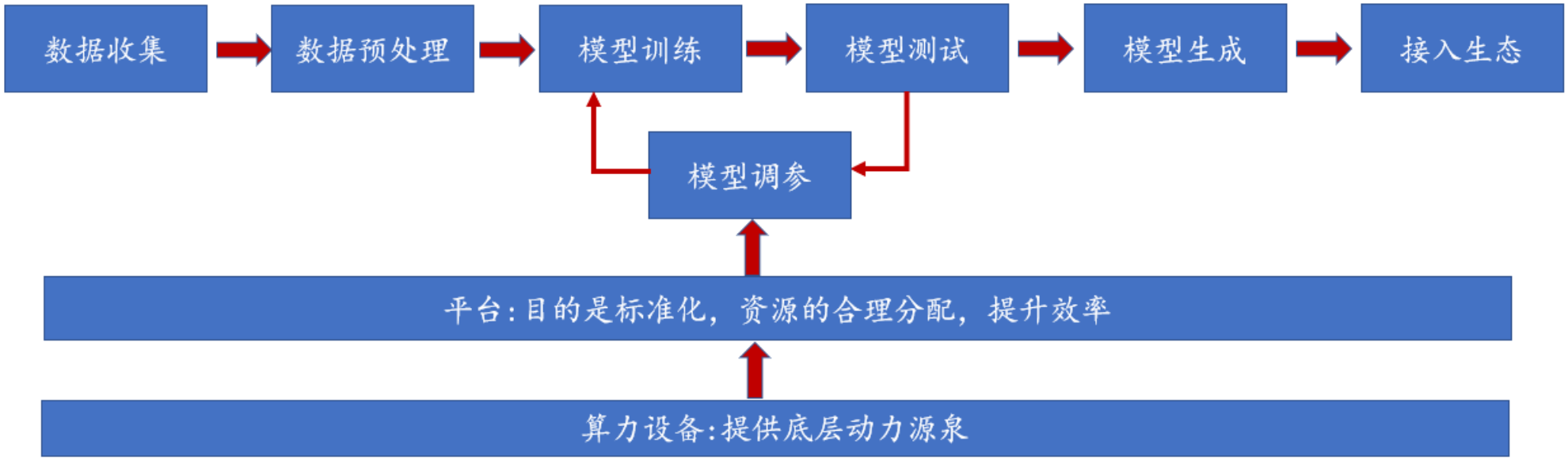
目录

01 AI服务器需求呈现加速状态

1.1 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；**服务器是算力的底层载体**，包含CPU、GPU、内存、硬盘、网卡等，在ChatGPT中具有举足轻重的作用，算力是服务器通过对数据进行处理后实现结果输出的一种能力。

数据、平台、算力、算法关系示意图



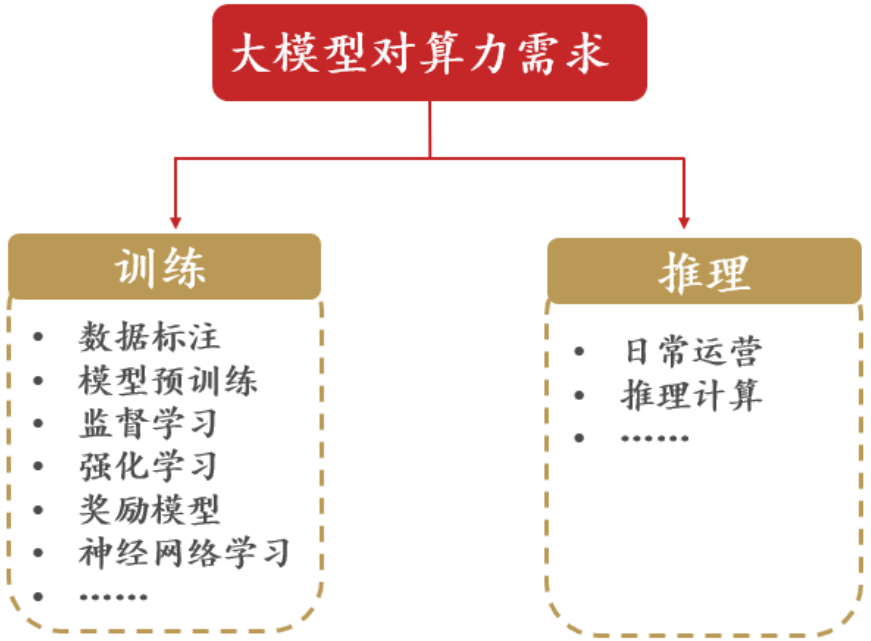
资料来源：IDC，华西证券研究所

1.2 大模型出现带动AI服务器呈现加速状态

- **我们认为大模型的出现有望带动AI服务器需求:** 我们认为除了对低延迟低功耗算力的性能需求，在服务器的种类上也产生了多样化、细分化的场景应用需求。各行业与人工智能技术的深度结合及应用场景的不断成熟与落地，使人工智能芯片朝着多元化的方向发展，为了迎合芯片的多元化，服务器的类型也将越来越丰富，并适用越来越多的行业应用场景。根据IDC的数据，在2021年的统计，预计到2025年中国加速服务器市场规模将达到108.6亿美元，且2023年仍处于中高速增长期，增长率约为20%。
- **AI大模型对算力的需求分别来自训练和推理两个环节。1）训练环节：**通过标记过的数据来训练出一个复杂的神经网络模型，使其能够适应特定的功能，模型具有一定的通用性，以便完成各种各样的学习任务。该环节需要处理海量的数据，注重绝对的计算能力。**2）推理环节：**利用训练好的模型，使用新数据推理出各种结论。借助神经网络模型进行运算，利用输入的新数据来一次性获得正确结论的过程。该环节对算力要求比训练环节略低，但注重综合指标，单位能耗算力、时延、成本等都要考虑。

2021-2025年中国服务器市场规模及增速(亿美元)

AI大模型对于算力(服务器)的需求



资料来源：IDC，华西证券研究所

1.3 服务器架构随负载量扩张而不断优化

- **服务器价值凸显**:计算机的一种，它比普通计算机运行更快、负载更高、价格更贵，主要用于在网络中为其它客户机提供计算或者应用服务。服务器具有高速的CPU运算能力、长时间的可靠运行、强大的I/O外部数据吞吐能力以及更好的扩展性。服务器一般具备承担响应服务请求、承担服务、保障服务的能力。其内部的结构与普通的计算机相差不大，主要包括如：CPU、硬盘、内存，系统、系统总线等，但相较于PC端需考虑几方面，例如可拓展性、易使用性、可用性和易管理性。
- **服务器架构随负载量扩张而不断优化:** 服务器架构经历了从传统单一模式到集群模式，再到分布式架构的优化过程。**传统单一模式**，服务器诞生初期将所有功能汇集在同一个系统，缺点为不便于维护、横向拓展性不佳；因此**集群模式诞生**，这种集群模式将同一项目放在多个服务器上，有效缓解用户访问量大的压力，但由于各个服务器间功能重复却缺乏协同，系统维护成本仍然较高，且增加了用户重复登陆问题，因此**服务器架构进化到分布式模式**。在分布式架构中，整个系统按照不同功能拆分为多个单一功能的子模块，每个模块被放到不同服务器中相互协作，共同组成服务器网络，能够有效解决功能耦合度高等问题且代码复用性高。

服务器架构演变过程示意图

资料来源：百度百科，IDC，华西证券研究所

1.4 深度拆解服务器核心硬件组成部分

◆ **服务器的组成**：服务器主要由主板、内存、CPU、磁盘、网卡、显卡、电源、主机箱等硬件设备组成；其中CPU、内部存储和外部存储是组成核心部件。

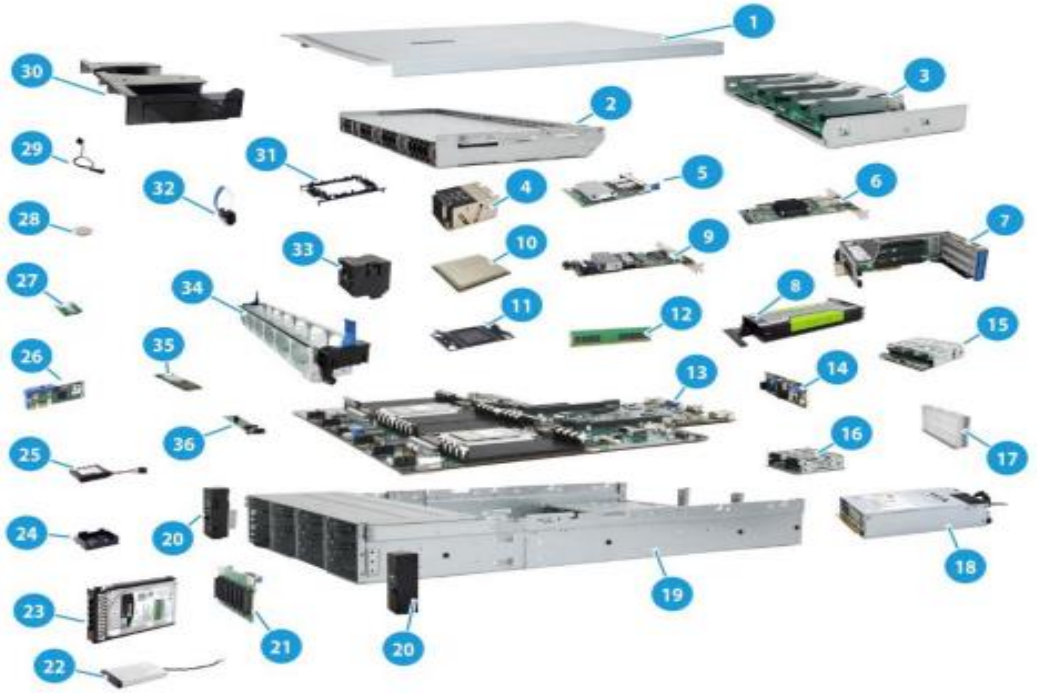
✓ CPU**处理器**：负责整个服务器的运算与控制，相当于人的大脑，是直接影响到服务器性能的核心部件。单台服务器可由多个CPU组成，一般服务器CPU个数多为2-4颗，也可有单颗的；虚拟化主机CPU有4-8颗的。CPU越多服务器性能越高。CPU的核数一般都是四核。

✓ **内部存储**：是CPU和硬盘之间的缓冲设备，是临时存储器（作用是临时存放数据），程序在运行的时候，都会调度到内存中运行，服务器关闭或程序关闭之后数据将自动从内存中释放掉。

✓ **外部存储**：永久存放数据的存储器，其中常用的硬盘有300GB，500GB，1TB，3TB，4TB等。硬盘类型分机械硬盘，固态硬盘两种。

◆ **硬件成本构成**：我们认为，以一台通用服务器为例，CPU(主板或芯片组)占比最高，大约占成本50%以上，内存(内部存储+外部存储)占比约为20%。

H3C UniServer R4900 G5服务器硬件结构拆解



H3C UniServer R4900 G5服务器硬件结构注释

编号	名称	编号	名称
3	中置GPU模块	12	内存
5、6	网卡	13	主板
7	Riser卡	18	电源模块
8	GPU卡	23	硬盘
9	存储控制卡	25	超级电容
10	CPU	27	加密模块
12	内存	28	系统电池

资料来源：H3C UniServer R4900 G5 技术白皮书，华西证券研究所

1.5 服务器的分类:按机箱结构分类

- ◆ **服务器按照机箱结构可分为:塔式服务器、机架式服务器、机柜式服务器、刀片式服务器。**
- ✓ **塔式服务器:** 采用台式机箱结构，常见的入门级和工作组级服务器基本上都采用这一服务器结构类型。**优点**：对放置空间要求较小，拓展性高，应用范围广泛，成本较低；**缺点**：升级扩张有限，独立性强；
- ✓ **机架式服务器**：设计宗旨主要是为了尽可能减少服务器空间的占用，例如专业网络设备。**优点**：比塔式服务器对空间的要求更小。可扩展性强，扩展操作便利；**缺点**：拓展和散热受到一定限制，因此无法实现完美的设备扩张，单机性能有限；
- ✓ **机柜式服务器**：应用于企业端，内部设备较多或不同设备单元放置在一个机柜中。**优点**:功能模块与支撑模块彻底分离，可靠高效。灵活架构，允许网络、计算、存储有机共存、维护简便，**缺点**：投入成本较高、能耗高、内部拓展性有限。
- ✓ **刀片式服务器:** 专为特殊应用行业和高密度计算机环境而生，每一片“刀片”即模板，类似独立服务器，在集群模式下，具备高速网络环境、资源共享等领域，广泛应用于数码媒体、医学、航天、军事等领域，性能较高，可实现轻松替换且便于维护，但是价格成本较高。

塔式服务器示意图

机架式服务器示意图

刀片式服务器示意图

机柜式服务器

资料来源：IDC，华西证券研究所

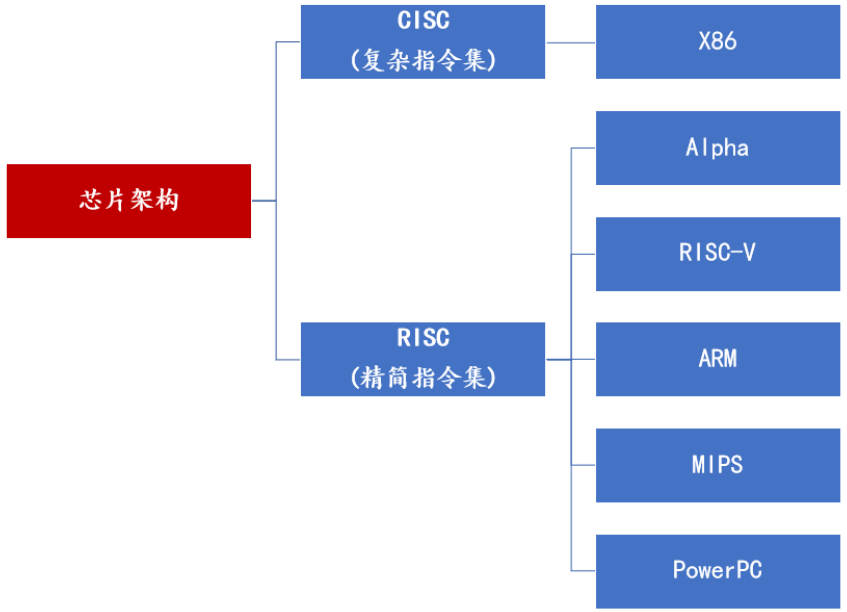
1.6.1 服务器的分类方式: 按照CPU架构分类

- ◆ **伴随应用需求不断扩张，不同架构服务器百花齐放**：按照CPU指令集架构的差异，服务器可分为CISC、RISC、VLIM等架构。
- ✓ **CISC(复杂指令集):庞大复杂的指令数目，**常见CISC微指令集主要集中在：AMD、Intel、VIA等IA-32、X86架构的CPU产品；优点在于能够有效缩短新指令的微代码设计时间，允许设计师实现CISC体系机器的向上相容，指令丰富且功能强大，而缺点指令使用率不均衡、不利于采用先进结构提高性能等。
- ✓ **RISC(精简指令集): 对指令数目和寻址方式都做了精简。**包含了简单、基本的指令，透过这些简单、基本的指令，就可以组合成复杂指令，常见RISC微指令集主要集中在：DECAlpha、ARC、ARM、AVR、MIPS、PA-RISC、PowerPC、RISC-V中，优点在于指令执行效率高，原因是90%指令由硬件直接完成，10%的指令是由软件以组合的方式完成；缺点在于指令数较少，功能不及CISC强大。
- ✓ **VLIM(超长指令集架构)：**采用多个独立的功能部件,指令调度是由编译器静态调度完成,因此指令可同时流出数目越大，超长指令的性能就明显；优点在于结构简单且价格低廉，缺点在于编译器负担较重，且需要更多内存，目前微处理器有Intel的IA-64和AMD的x86-64。

CISC与RISC比较

	CISC	RISC
指令系统	指令系统丰富，有专用指令来完成特定的功能，处理特殊任务效率较高、指令长度不同	保留简单高效的常用指令，复杂指令通过简单指令组合，实现特殊功能效率低，可通过流水技术弥补，指令长度相同
存储操作	存储器操作指令多，可直接操作内存和寄存器，数据流控制复杂	对存储器操作有限制，运算基本都限于寄存器间，控制简单。
程序	汇编语言程序编程相对简单，科学计算及复杂操作的程序设计相对容易，效率较高	汇编语言程序一般需要较大的内存空间，实现特殊功能时程序复杂
指令执行时间	很多复杂指令都通过CPU内的微码来完成，微码比较复杂的指令需要多个时钟周期才能完成，指令不等长周期增加了指令流水线优化的难度	大部分的指令都可以在一个时钟周期内完成降低了指令流水线设计的复杂度
中断	CISC计算机是在一条指令执行结束后响应中断	RISC计算机在一条指令执行的适当地方可以响应中断，但是相比CISC指令执行的时间短，所以中断响应及时
CPU	CPU包含有丰富的电路单元，因而功能强、面积大、功耗大	CPU包含有较少的单元电路，因而面积小、功耗低
设计周期	CISC微处理器结构复杂，设计周期长，采用微程序可适当降低复杂性	RISC微处理器结构简单，布局紧凑，设计周期短，且易于采用最新技术

芯片根据指令集分类

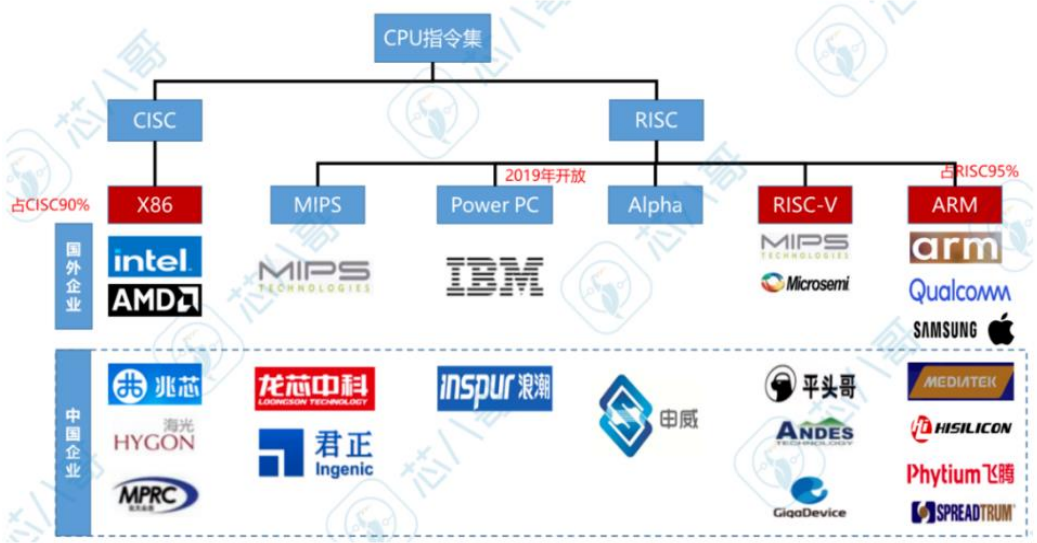


资料来源：CSDN，华西证券研究所

1.6.2 X86和ARM各具优势，国产生态迎新机遇

- **X86架构服务器仍占绝对优势，ARM架构服务器潜力巨大：**根据市场应用占比把服务器分为X86服务器和非X86服务器，目前使用X86架构的服务器CPU仍然占据绝对优势。根据芯八哥数据，按照2021年统计数据，X86架构市场占比高达97%，ARM占比仅为2.07%，Power BI 占比为0.27% ，但以ARM为代表的RISC结构近年来增长迅猛，尤其国内诞生了以华为海思、阿里平头哥为代表芯片企业。
- **X86和ARM各具优势:** ARM体积小、低功耗、低成本、执行更加高效、指令长度固定，然而在性能上不及X86，如果ARM要在性能上接近X86，就需要极高的频率，从而带来较高能耗；X86单条指令功能强大且指令数相对较小、带宽要求低，然而缺点在于寻址范围小、部分计算机利用率不高、执行速度慢。
- **ARM加速迭代，国产生态迎新机遇:** 根据TrendForce数据预测，随着云数据中心增长，预计到2025年，ARM架构在数据中心服务器市场渗透率将达到22%；ARM在服务器的市场崭露头角，早在2008年高通、博通、微软、华为、飞腾等，也陆续开发了各自的ARM服务器CPU，2019年，随着着ARM的Neoverse平台路线图的推出，服务器市场份额渗透率得到质的提升；国产生态迎新机遇，X86生态依然被AMD和英特尔垄断，而ARM架构随着国产生态和技术逐渐成熟，迎来国产替代的新机遇。

CPU指令集生态



ARM和X86服务器对比

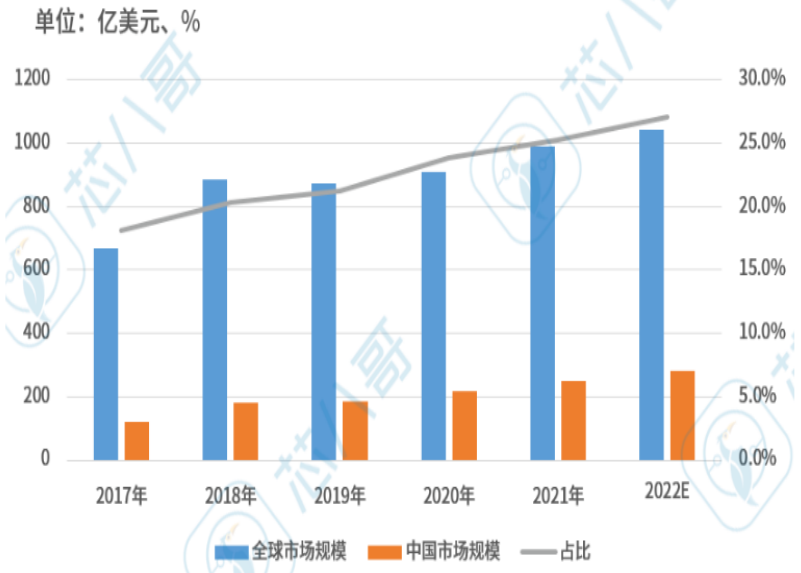
	ARM	X86
架构	华为，飞腾，Ampere，Marvell	Intel/AMD
特点	众核架构，适合高并发、高带宽的计算场景；	高主频、高功耗，覆盖高性能和通用计算场景
价值	提升计算效率，节能、省空间。高效能计算带来高性价比	驱动性能增长的工艺改进边际成本激增，摩尔定律难以为继
生态	IP 授权商业模式，生态开放和融合，数据中心应用生态逐步完善	数据中心应用生态完善，但产业被垄断、把控，无法合作共赢

资料来源：芯八哥，CSDN，OFweek 电子工程网，华西证券研究所

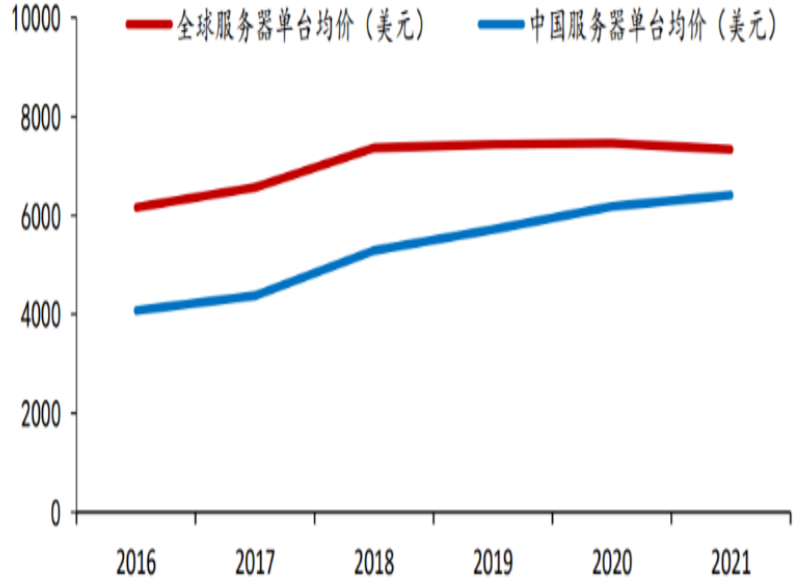
1.6.3 我国服务器占比逐年攀升，呈现快速增长态势

- **我国服务器占比逐年攀升，云有望成为重要抓手：**根据IDC数据，2021年全球服务器市场规模为992亿美元，同比增长9.01%，中国市场规模约251亿美元，近年来占全球比重呈现快速上升趋势，已成为全球最主要的服务器增长市场。此外，公有云作为国内外数据流量的重要抓手，服务器同样彰显其重要算力底座，我们认为其存在巨大成长空间。
- **我国服务器单台均价接近全球均价：**2021年全球服务器平均单价高达7328 美元/台，我国市场也达到了6415 美元/台，我国服务器价格呈现上升状态，并且接近全球服务器平均价格。
- **人工智能应用场景下的加速计算服务器是中国服务器的核心驱动力：**根据IDC的数据，随着智能应用正不断深入，从碎片化过渡到深度融合的一体化，从单点转换为多元化的应用场景，在金融、制造、能源和公共事业等行业体现尤为显著。2021年上半年，中国加速计算服务器市场达到24亿美元，同比增长85.1%，此外，中国2021年H1服务器排名前五的厂商分别为浪潮、新华三、华为、戴尔、联想。此外，根据IDC的预测，未来中国整体服务器的复合增长率为12.7%，2025年中国服务器市场规模预计将达到424.7亿美元。

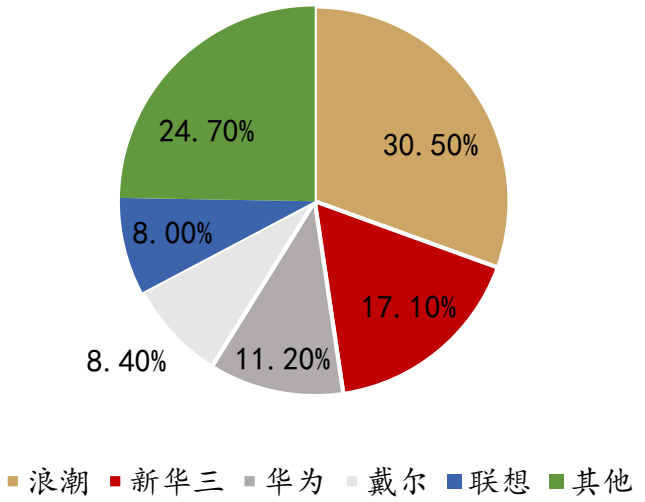
全球服务器与中国服务器市场规模及占比



全球服务器价格及中国服务器价格比较



2021年上半年中国服务器市场份额

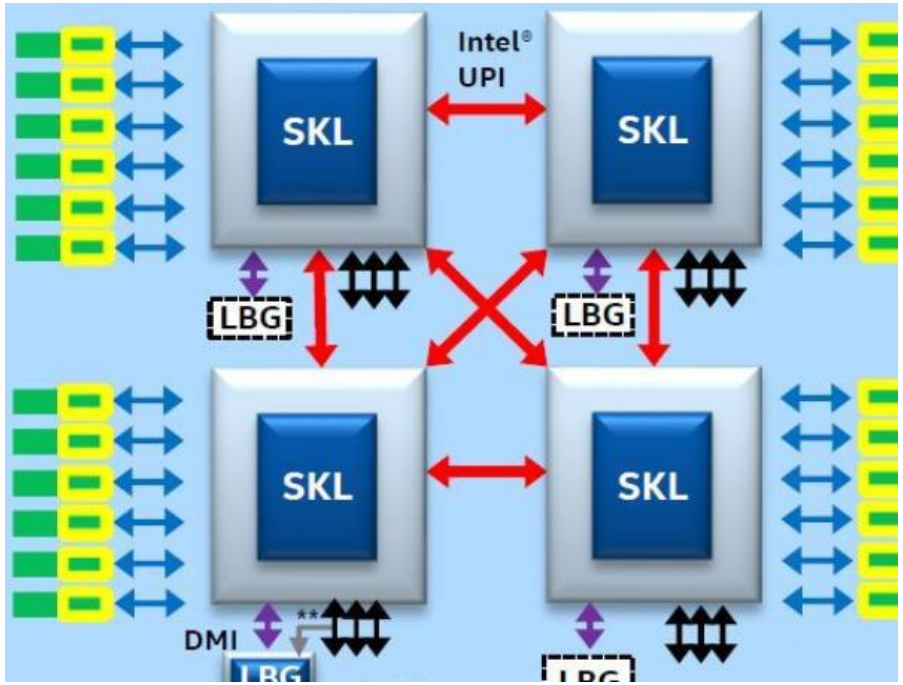


资料来源：芯八哥，IDC，华西证券研究所

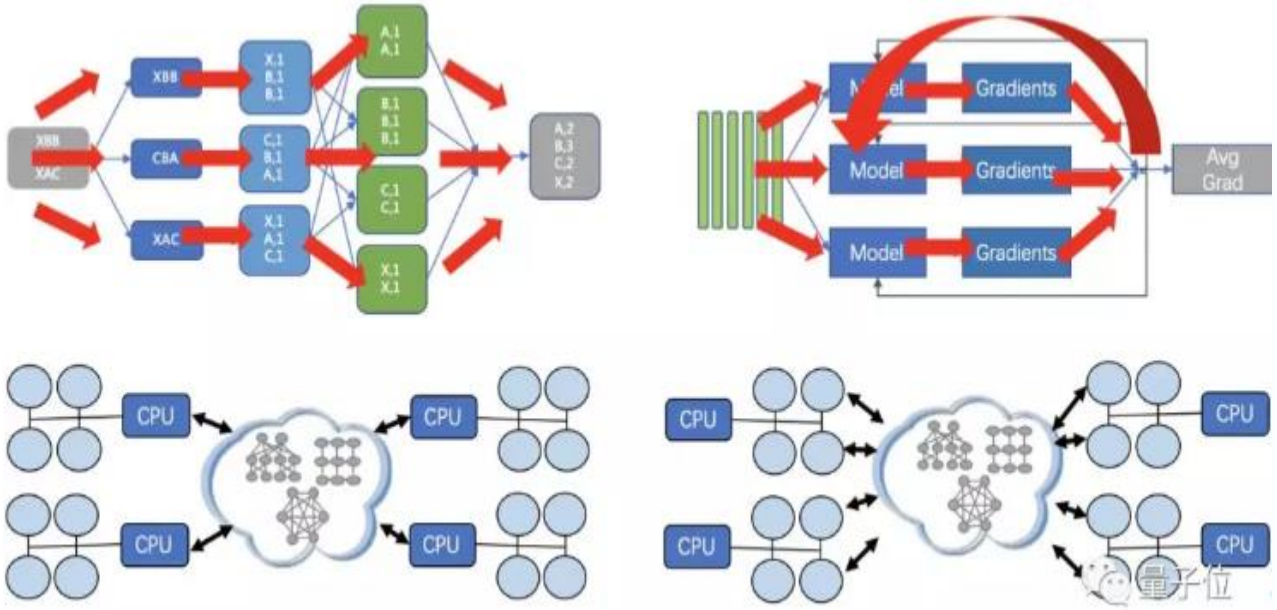
1.7.1 AI服务器: 加速计算服务器是服务器成长的核心驱动力

- **服务器同样可以按照CPU数量进行分类:** 可以分为单路服务器、双路服务器、四路服务器和多路服务器。“路”指的是服务器物理CPU的数量，也就是服务器主板上CPU插槽的数量。单路指服务器支持1个CPU；双路指服务器支持2个CPU；四路指服务器支持4个CPU；以此类推。一般CPU数量越多，即拥有更强的性能，同时能显著降低性能的功耗比。
- **AI服务器价值凸显:** 随着大数据、云计算、人工智能等技术的成熟与在各行各业的应用，AI服务器价值凸显；1、**硬件架构**，相较于通用服务器，AI服务器是采用异构形式的服务器，在异构方式上可以根据应用的范围采用不同的组合方式，如CPU+GPU、CPU+TPU、CPU+其他的加速卡等；2、**加速卡数量**：通用服务器一般是单路或多路CPU架构，而AI服务器需要承担大量的AI运算，一般配置四块及以上加速卡；3、**独特设计**，AI服务器由于对加速卡的独特需求，需要针对性的对于系统结构、散热等做专门的设计，才能满足AI服务器需求。

四路服务器示意图



阿里云多路AI集群服务器示意图

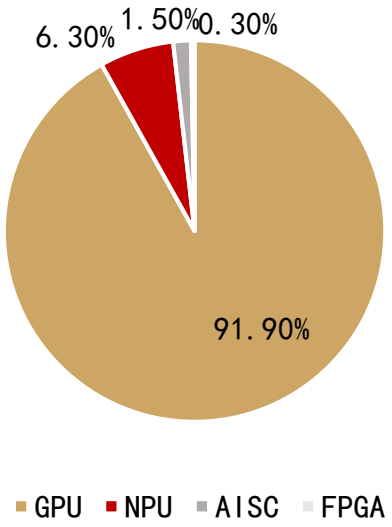


资料来源：搜狐，博客，阿里云官网，华西证券研究所

1.7.2 AI服务器: GPU为主流“加速卡”，正在大放异彩

- **AI芯片是AI算力的“心脏”，GPU价值凸显:** 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 芯片在人工智能的算法和应用上做针对性设计，可高效处理人工智能应用中日渐多样繁杂的计算任务。在人工智能不断扩大渗透的数字时代，芯片多元化展现出广阔的应用前景，通过不断演进的架构，为下一代计算提供源源不断的动力源泉。
- **GPU作为AI芯片的主力军，正在大放异彩**: AI芯片主要包括图形处理器(GPU)、现场可编程门阵列(FPGA)、专用集成电路(ASIC)、神经拟态芯片(NPU)等。人工智能深度学习需要异常强大的并行处理能力，GPU相比于CPU更擅长于并行计算能力，正在大放异彩。根据IDC的数据，2021年H1中国人工智能芯片，GPU占比最多为91.90%。
- **GPU服务器优势显著:** GPU服务器超强的计算功能可应用于海量数据处理方面的运算，如搜索、大数据推荐、智能输入法等，相较于通用服务器，在数据量和计算量方面具有成倍的效率优势。此外，**GPU可作为深度学习的训练平台**，优势在于1、GPU 服务器可直接加速计算服务，亦可直接与外界连接通信；2、GPU服务器和云服务器搭配使用，云服务器为主，GPU服务器负责提供计算平台；3、对象存储 COS 可以为 GPU 服务器提供大数据量的云存储服务。

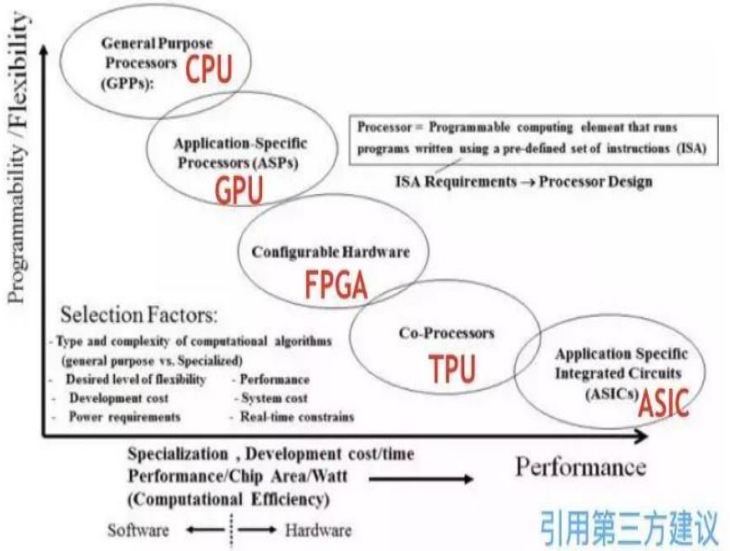
2021年H1中国人工智能芯片占比



GPU、FPGA、ASIC对比

AI芯片	释意
GPU	显卡的核心单元，是单指令、多数据处理器。GPU采用数量众多的计算单元和超长的流水线，在图型领域的加速方面具有技术优势
FPGA	集成了大量的基本门电路及存储器，利用门电路直接运算、速度较快。用户可以自由定义这些门电路和存储器之间的布线，改变执行方案，从而调整到最佳运行效果，相较于GPU灵活度更高、功耗更低；
ASIC	为特定目的、面向特定用户需求设计的定制芯片，具备体积小、功耗低、可靠性更高等有点。在大规模量产的情况下，具备成本低的特点。

GPU、FPGA、ASIC对比(纵轴代表灵活性、横轴代表性能)

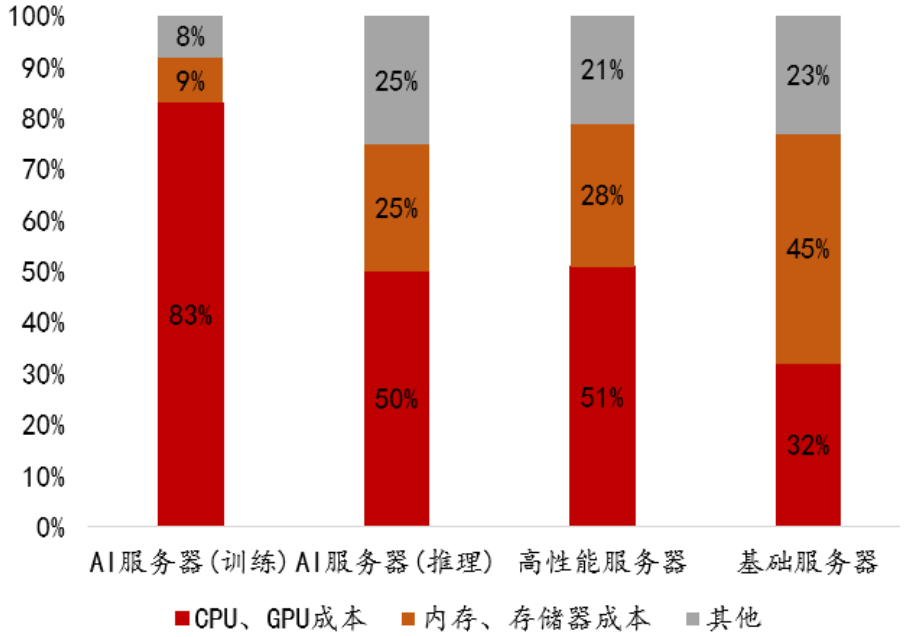


资料来源：IDC，CSDN，华西证券研究所

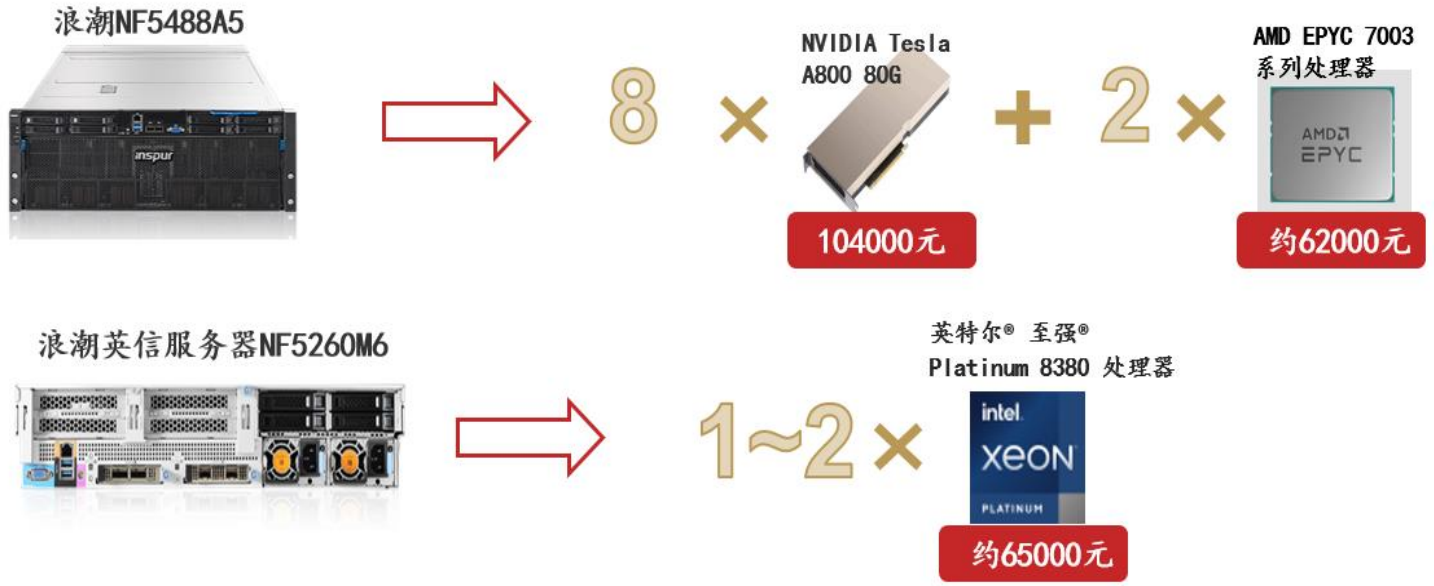
1.7.3 AI服务器: 芯片组(CPU+GPU)价值成本凸显

- ◆ **AI服务器芯片组价值成本凸显:** 根据Wind及芯语的数据，AI服务器相较于高性能服务器、基础服务器在芯片组(CPU+GPU)的价格往往更高，AI服务器(训练)芯片组的成本占比高达83%、AI服务器(推理)芯片组占比为50%，远远高于通用服务器芯片组的占比。
- ✓ 浪潮通用服务器浪潮英信服务器NF5260M6搭载第三代英特尔®至强®可扩展处理器的一款2U双路机架式服务器，可支持1-2个支持1到2个英特尔® 至强® 第三代可扩展处理器，根据Intel官网数据，此款处理器建议零售价为9359美元，折合人民币约65000元。
- ✓ 浪潮AI处理器浪潮NF5488A5是一款浪潮自研的具有超强算力的AI服务器，性能领先。在4U空间内支持8颗第三代NVLink的NVIDIA A800 GPU，搭载2颗支持PCIe4.0的AMD EPYC 7002/7003 处理器，可提供极致训练性能和超高数据吞吐，广泛适用于图像、视频、语音识别、金融分析、智能客服等典型AI应用场景，根据天极网和中关村在线网数据，该款AMD(CPU)售价为8880美元，折合人民币约62000元，该款GPU售价为104000元。

服务器成本购成



浪潮服务器成本对比



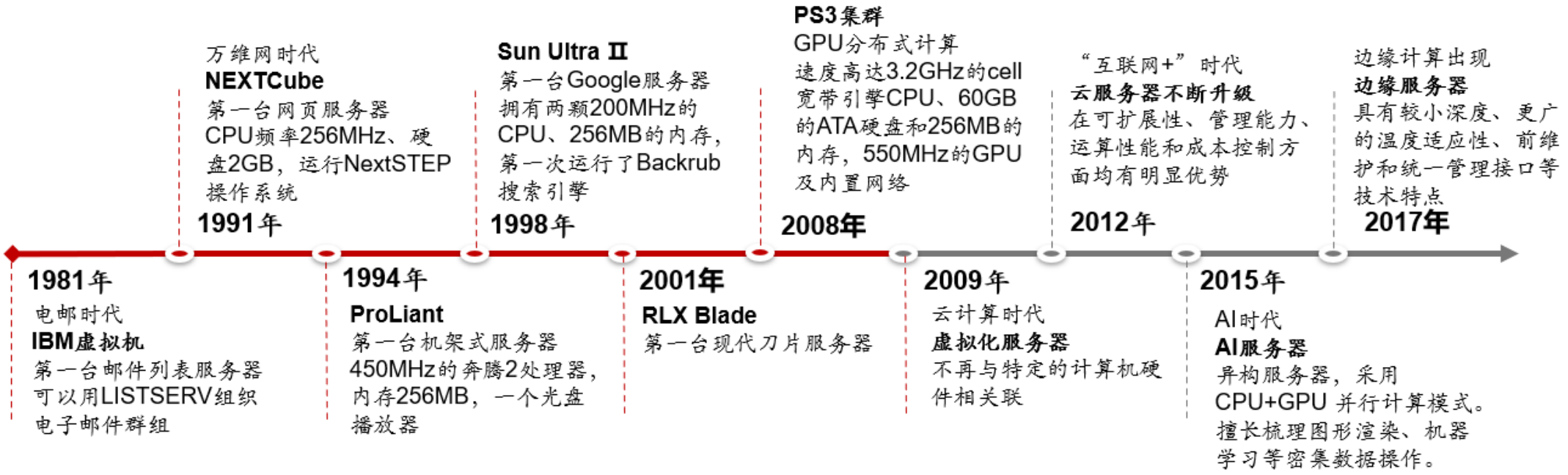
资料来源：WIND，芯语，浪潮官网，华西证券研究所

02 拥抱AI服务器的星辰大海

2.1 服务器发展路径：“伴科技类”升级产品

- **我们认为服务器是“伴科技类”的硬件产品，随着科技的服务形式和应用方式不断进步，服务器同样在不断迭代升级或更新换代**：世界上最早的服务器可以追溯到1981年IBM大型机上的BITNET电子邮件群组，是第一台邮件列表服务器。此后，随着万维网的出现和搜索引擎等互联网迭代升级，技术不断迭代。
- **近年，随着互联网+、云计算、AI+、边缘计算的出现，服务器市场迎来了极大的发展：**2009年左右，随着虚拟化技术不断成熟，**云计算**的服务模式被大众广泛接受，云数据中心对服务器的需求旺盛；2012年左右，我国进入**“互联网+”时代**，云计算服务模式叠加电子商务的需求，拓展性、运算性能、数据存储容量等需求凸显，服务器需求不断增加；2015年左右，全球进入”**AI+时代**”，以人工智能、深度学习、神经网络的训练和推理等赋能千行百业，AI服务器价值凸显，其具备图形渲染和海量数据的并行运算等优势，市场需求旺盛；2017年左右，随着**边缘计算**、“物联网+”的兴起，叠加AI等需求，服务器市场依旧火热。

服务器的本质是伴随

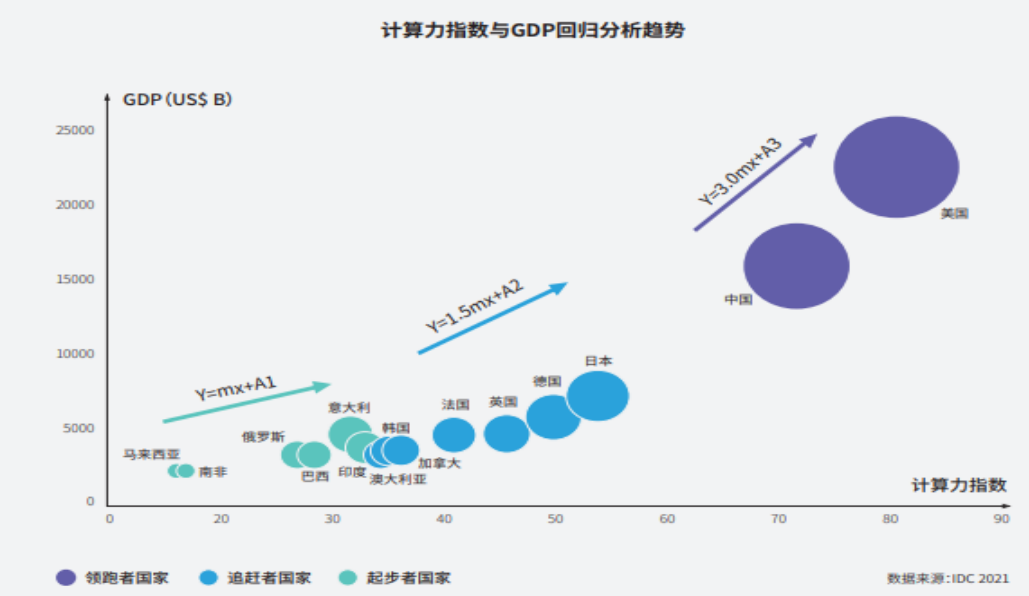


资料来源：公开资料整理，腾讯云官网，资产信息网，新睿云，华西证券研究所

2.2 算力时代到来，服务器价值凸显

- **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。
- AI**服务器作为算力载体为数字经济时代提供广阔动力源泉**：不同于通用服务器，AI服务器更专精于海量数据处理和运算方面，我们认为其可以为人工智能、深度学习、神经网络、大模型等场景提供广阔的动力源泉，并广泛应用于医学、材料、金融、科技等千行百业。

从计算力指数看对经济的增长



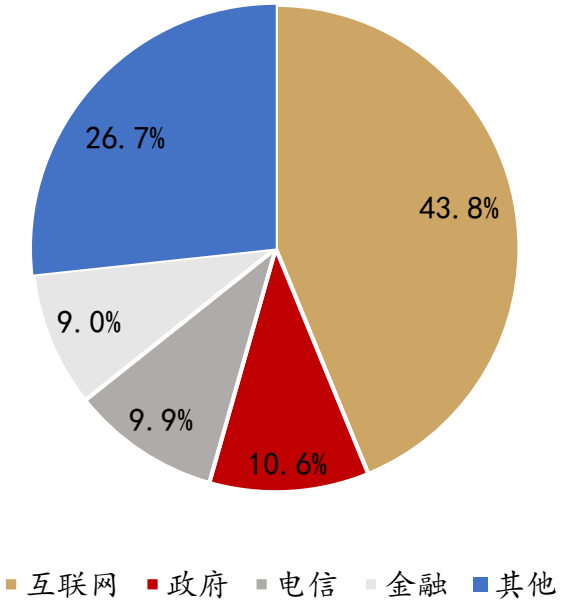
计算力对经济的影响

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

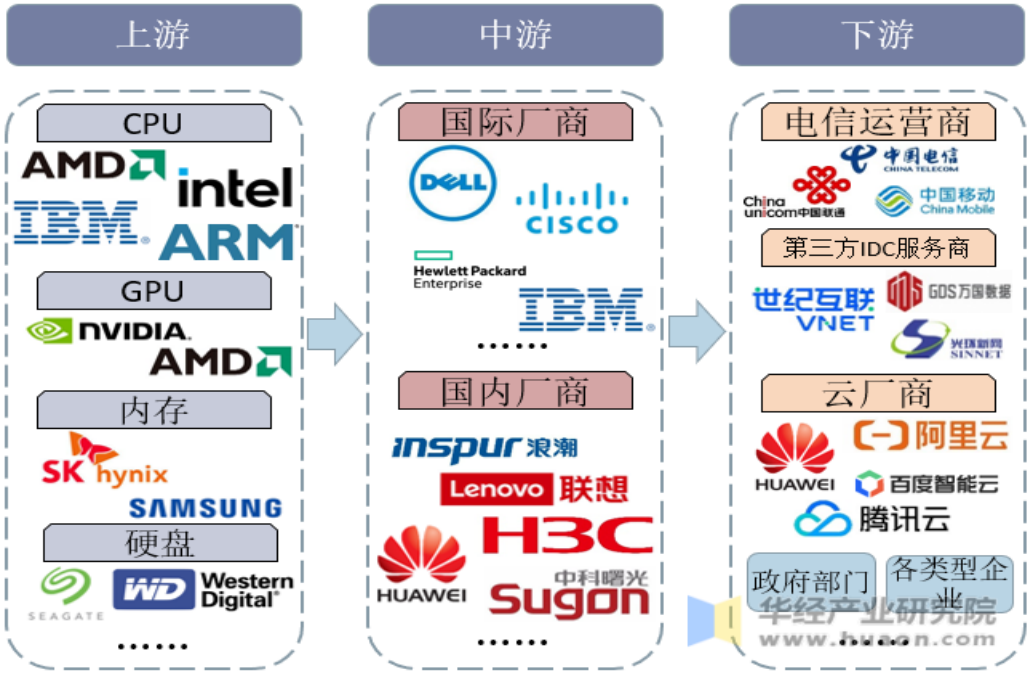
2.3 服务器产业链梳理，下游赋能千行百业

- **服务器产业链梳理，关注产业链中上游:** 服务器行业产业链上游为CPU、GPU、内存、硬盘、RAID控制器、电源、软件系统等原材料为主；中游为服务器行业；下游客户群体有互联网云服务商、电信运营商、第三方IDC服务商、政府部门、各类型企业等。我们认为在算力和数字时代的大背景下，AI服务器作为算力载体为数字经济时代提供广阔动力源泉，更加彰显其重要性。
- **服务器赋能千行百业:** 根据IDC数据，服务器赋能千行百业，实则为数字经济的底层基础设施；其中，互联网行业占比最多，为43.8%，广泛应用于电子商务、电子邮件、电子游戏等领域；电信行业占比9.9%，应用场景为通讯网络、云平台建设；金融占比约为9.0%，广泛应用于商业业务系统、银行系统等场景；政府领域占比为10.6%，主要应用于数字政务、办公系统等领域。

21年服务器下游占比



服务器产业链



资料来源：华经产业研究院，IDC，华西证券研究所

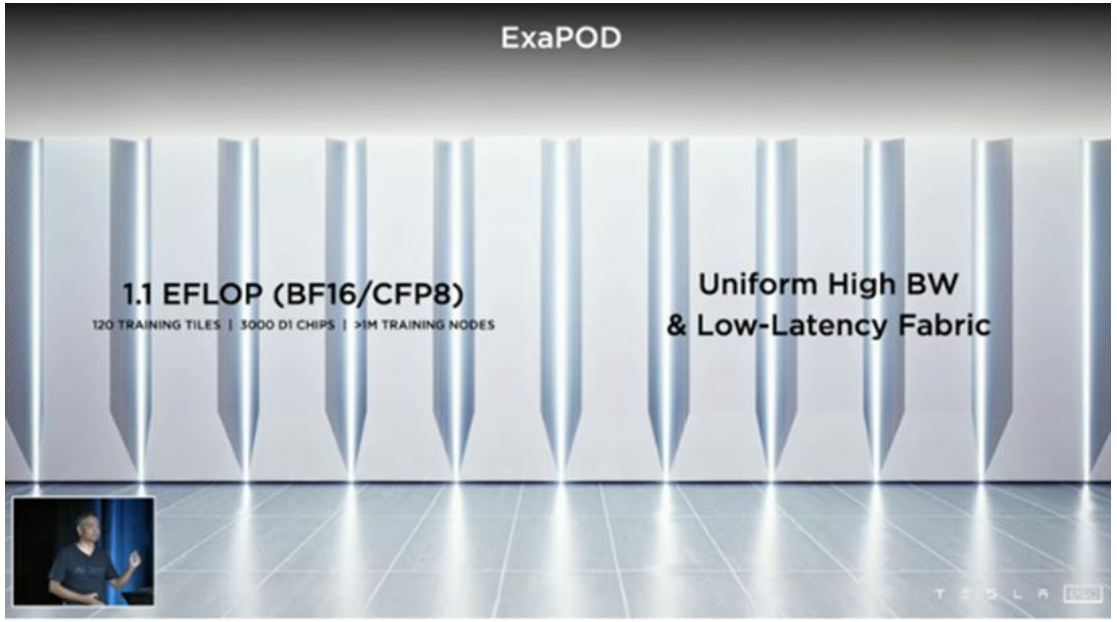
2.4 数字经济时代，服务器应用前景广阔

- **互联网，积极拥抱新兴技术，领先全球算力水平**： 根据IDC数据统计，2021年互联网企业采购的IT基础架构中，超过九成被应用于云计算部署方式。此外，互联网与人工智能、大数据等新兴技术的结合也催生了对海量计算能力的需求。目前，从互联网数据中心的体量来看，中、美仍处在第一梯队，中美两国占全球整体服务器保有量六成以上。近年来互联网行业在亚太区的增长颇为突出，这主要源于疫情之后在线需求的增加，以 及亚太地区经济的复苏。此外，中国持续加大数据中心的部署，更多企业采取云服务方式。

- **电信，利用算力投入优化内部管理、赋能业务创新:** 内部，随着5G、云计算等技术的落地，电信运营商对内面临着业务增长压力；外部，智慧交通、智慧零售、车联网、游戏娱乐、AR/VR应用等增值业务等算力需求逐步增加。海量创新业务增长对数据快速访问价值凸显，要求电信数据厂商承担数据高并发、低延迟传输、保证业务永续的能力。

IBM超级计算机示意图

特斯拉超级计算机示意图

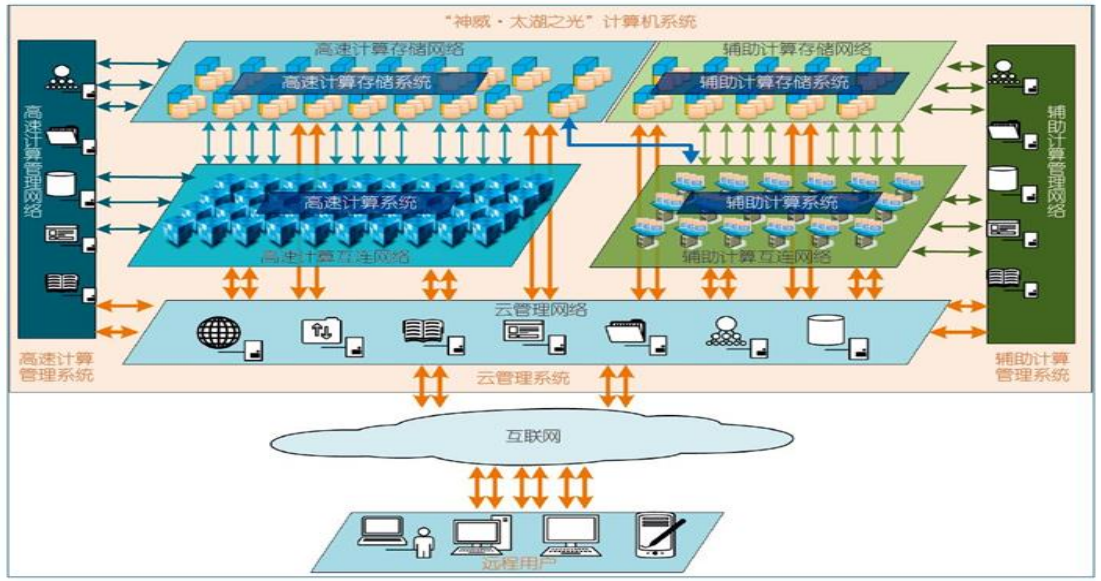


资料来源： IDC《全球算力指数评估报告》，华西证券研究所

2.4 数字经济时代，服务器应用前景广阔

- **金融，智能化加速，有力支撑金融业务创新发展**：随着移动互联网场景的普及，金融行业（包含银行、保险和证券）的数字化业务迅猛发展，呈现出线上化、智能化、无接触等特征，此外，数字银行、个人财富管理、数字化借贷、 全渠道支付等新兴金融场景层出不穷。金融行业对业务的及时性相应要求极高，移动互联业务由于其高并发、高峰值场景需求，稳定、安全、高效、弹性的基础设施成为首选。
- **制造， 实现智能制造，推动数字工厂建设:** 制造业是实体经济发展的核心支撑力量，也是全球算力水平最高的传统行业之一，2021年算力支出占全球12%，其中包括大型ERP系统运转、物联网、传感器的应用。此外，在人工智能、大数据、区块链等新兴技术使用上，制造业也领 先于大部分传统行业。根据IDC的预测，到2025年，中国制造业IT相关支 出占全球市场将达到20%左右。
- **医疗，算力投入有望推动信息化平台建设:** 随着医疗信息化等领域的高投入，初步形成以计算平台为核心的综合信息系统，在医院范围内形成数据互联互通、区域协同、分级诊疗的体系。随着AI等技术发展，大数据赋能医疗行业智能化升级将是下一个发展目标。

神威·太湖之光超级计算机系统架构



天河三号部署示意图

资料来源： IDC《全球算力指数评估报告》，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。**我们梳理了两条路径图，积极的推荐以下两条投资主线:**
- ✓ **1)具备服务器能力的厂商，重点推荐中科曙光，其他受益标的为浪潮信息、拓维信息、神州数码**
- ✓ **2)具备算力芯片的厂商，受益标的为寒武纪、海光信息、龙芯中科、景嘉微**

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)		
		2023/3/8	2023/3/8	2021	2022E	2023E	2021	2022E	2023E
寒武纪*	688256.SH	90.01	360.77	-2.06	-2.79	-1.79	-	-	-
拓维信息*	002261.SZ	11.28	141.66	0.07	-0.04	0.15	161.1	-	74.5
神州数码*	000034.SZ	26.59	177.85	0.37	1.56	1.89	72.1	17.0	14.1
龙芯中科*	688047.SH	115.15	461.75	0.66	0.43	0.72	174.5	267.5	160.1
浪潮信息*	000977.SZ	35.27	516.25	1.38	1.67	2.02	25.6	21.2	17.5
景嘉微*	300474.SZ	74.71	339.99	0.97	0.64	0.93	77.0	117.6	80.3
中科曙光	603019.SH	33.00	483.12	0.80	1.03	1.47	41.3	32.0	22.4
海光信息	688041.SH	53.16	1235.62	0.16	0.35	0.64	329.0	151.9	83.1

注：*来自wind一致预测

资料来源：WIND，华西证券研究所

3.2.1 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

3.2.2 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

3.2.3 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

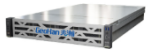
名称	示意图	形态	处理器	内存支持	AI加速卡/AI处理器	AI算力
KunTai A222		2U单路边缘机架式服务器	1*鲲鹏920处理器，24核，主频2.6GHz	4个DDR4 RDIMM，最高速率3200MT/s内存保护支持ECC、SEC/DED、SDDC、Patrol scrubbing功能；单根内存条容量支持16GB/32GB/64GB/128GB	最大支持3张Atlas 300V 视频解析卡或Atlas 300I Pro 推理卡或Atlas 300V Pro 视频解析卡	最大420 TOPS INT8
KunTai A722		2U 双路推理型 AI 机架式服务器	2*鲲鹏920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

3.2.4 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

3.3.1 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

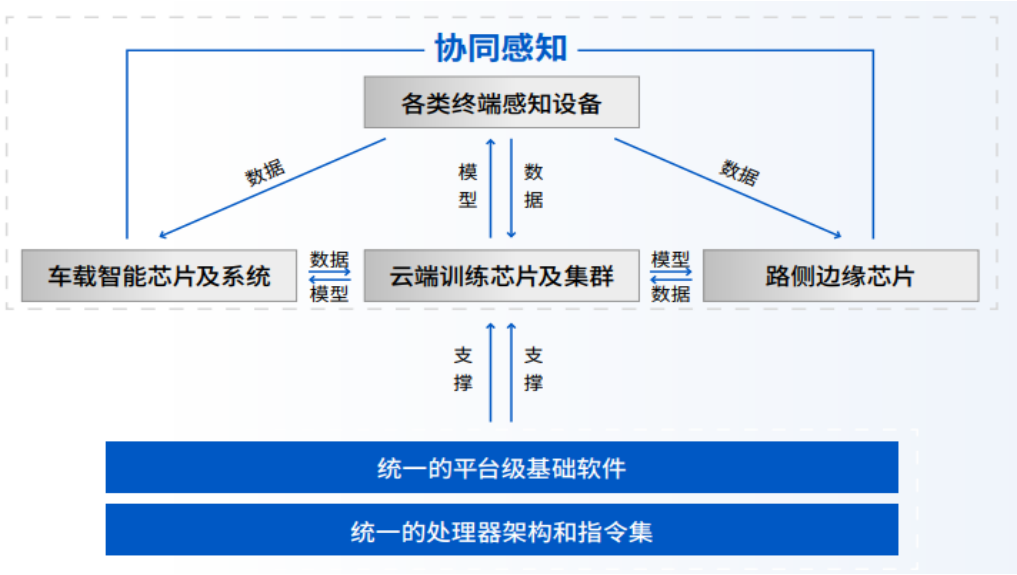
资料来源：公司公告，华西证券研究所

3.3.2 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

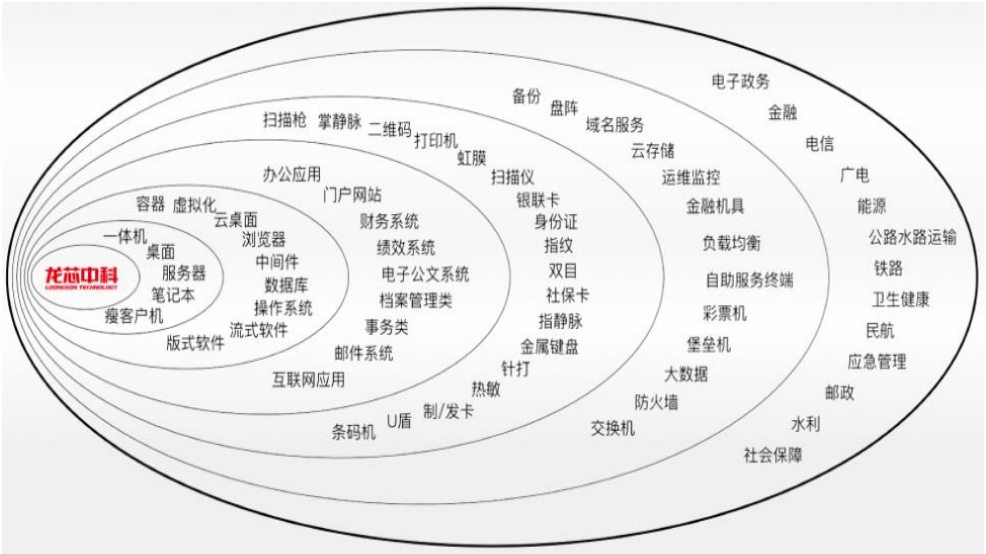
资料来源：公司官网，华西证券研究所

3.3.3 龙芯中科：2K2000系列集成自主GPU

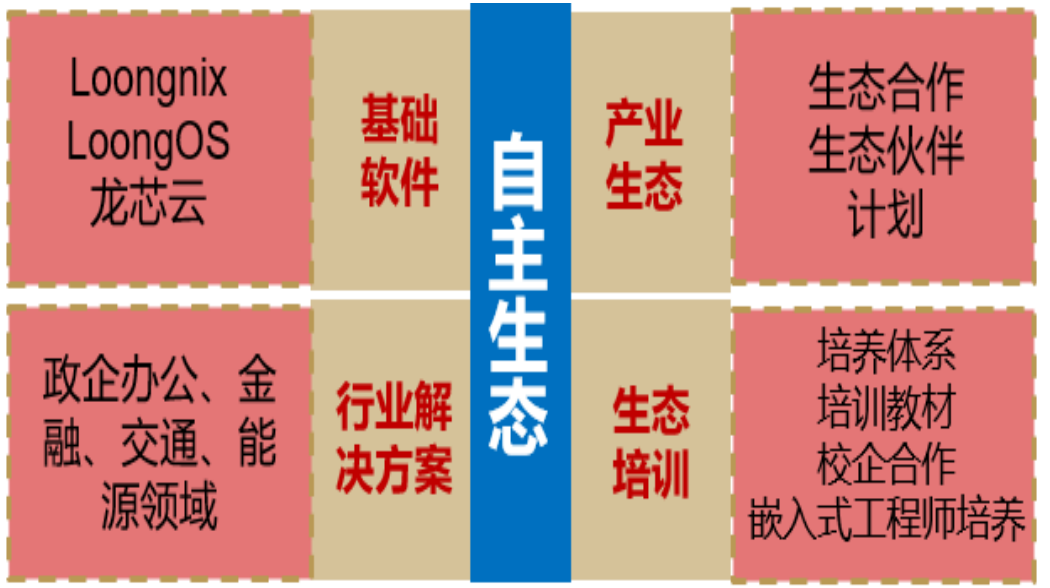
- **龙芯中科主营业务为处理器及配套芯片的研制、销售及服务：** 主要产品与服务包括处理器及配套芯片产品与基础软硬件解决方案业务。公司基于信息系统和工控系统两条主线开展产业生态建设，面向网络安全、办公与业务信息化、工控及物联网等领域与合作伙伴保持全面合作，产品在电子政务、能源、交通、金融、电信、教育等行业领域已获得广泛应用。

- **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

3.3.4 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

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。

THANKS