# AI 服务器行业深度：驱动因素、市场机遇、产业链及相关企业深度梳理

AI 服务器主要用于处理深度学习工作负载的海量数据，包括需要大内存容量、高带宽和整体系统内缓存一致性的训练和推断。相较于普通服务器，AI 服务器新增多张高性能加速器（绝大部分为 GPU），拥有更高的计算能力、更快的处理速度和更大的存储空间，以支持高负载和复杂的计算任务。

在当前大模型升级所带来的参数量指数级增加的情况下，AI 训练+推理段模型需求催生了 AI 服务器海量需求。根据 Statista 数据，2021 年全球服务器市场规模达到 831.7 亿美元，同比增长 6.97%，其中 AI 服务器市场达到 156.3 亿美元，同比增长 39.1%。AI 服务器有望成为服务器板块增速最快的细分板块，预计 AI 服务器市场将在 2026 年达到 347.1 亿美元，5 年 CAGR 达到 17.3%。

那么，对于 AI 服务器，当下呈现怎样的市场现状呢？有哪些影响因素驱动着中国市场 AI 服务器进一步向前发展？AI 服务器行业产业链上下游情况如何？市场对于哪些板块有发展机遇？其 AI 服务器行业市场格局如何？相关企业发展情况怎样？整体市场呈现怎样的产业前景？以下内容我们将聚焦这些问题，一起探究 AI 服务器行业的相关问题。

目录





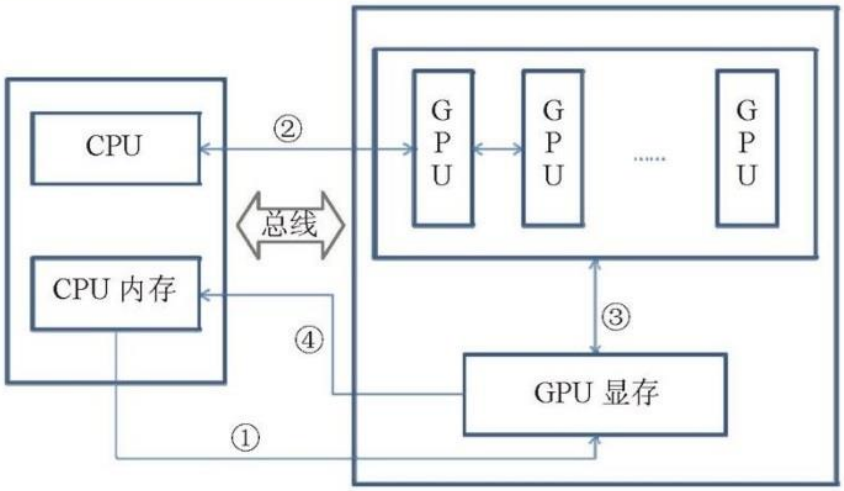
## 一、行业概况

### 1、人工智能服务器为算力支撑，助燃 AI 产业化

#### （1）从通用服务器到 AI 服务器的过渡

**AI 服务器在众多服务器中脱颖而出源于其架构优势**。AI 服务器是承载智慧计算中 AI 计算的核心基础设施，是一种能够提供人工智能的数据服务器，既可以用于支持本地应用程序和网页，也可以为云和本地服务器提供复杂的 AI 模型和服务，通过异构形式适应不同应用范围以及提升服务器的数据处理能力，异构方式包括 CPU+GPU/TPU/ASIC/FPGA。

AI 服务器的 CPU+架构




资料来源：《人工智能服务器技术研究》王峰，天风证券研究所

**AI 服务器的发展脱胎自通用服务器的性能提升需求**。复盘主流服务器的发展历程，随着数据量激增、数据场景复杂化，诞生了适用于不同场景的服务器类型：通用服务器、云计算服务器、边缘计算服务器、AI 服务器。随着大数据、云计算、人工智能及物联网等网络技术的普及，充斥在互联网中的数据呈现几何倍数的增长，使得以 CPU 为主要算力来源的传统服务器承受着越来越大的压力，并且对于目前 CPU 的制程工艺而言，单个 CPU 的核心数已经接近极限，但数据的增加却还在继续，因此服务器数据处理能力必须得到新的提升，在这种环境下，AI 服务器应运而生。面对 ChatGPT 所引出的大规模预训练模型，AI 服务器以其架构优势带来的大吞吐量特点，有望在一众服务器中脱颖而出。

主流服务器类型（不完全统计）

| | 特点 | 配置 | 应用场景 |
|---|---|---|---|
| 通用服务器 | 物理服务器，独立存在，拥有完全管理员权限和独立 IP 地址，安全稳定性高。 | CPU、硬盘、内存等。 | |
| 云计算服务器 | 通过虚拟化技术，将一台/多台服务器虚拟化成一个可以切分的资源池，客户按需灵活配置与扩展，管理便捷，费用相对低廉。 | 按客户需求配置 CPU、内存、数据盘等。 | 适合对业务弹性扩展需求和易用性的需求：电商、IT 行业、教育、移动应用、游戏等。 |
| 边缘计算服务器 | 承担 70%以上的计算任务，需支持 ARM/GPU/NPU 等异构计算，针对不同业务场景开发，远程控制运维。 | | 工业互联网、车联网、医疗保健、AR/VR、智慧城市等。 |
| AI 服务器 | 采用异构形式服务器，承担大量计算；大规模并行运算、多重向量/张量运算、计算效率高。 | GPU/FPGA/ASIC 等加速芯片、CPU、内存等。 | 融合深度学习、机器视觉、知识图谱等人工智能技术的应用：医疗影像智能分析、人脸/语音/指纹识别、安防监控场景等。 |

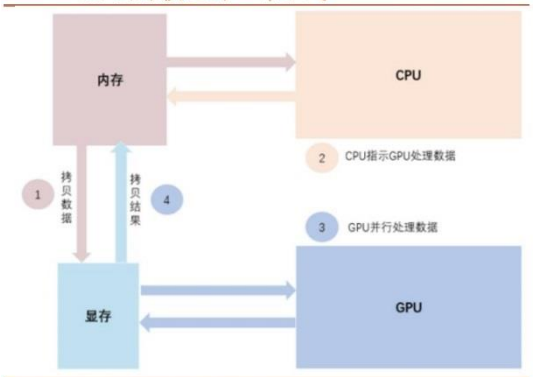
资料来源：人工智能与创新公众号、皖云数科公众号、物联网世界官网、高升数据公众号、机智云物联网公众号、天风证券研究所

## （2）AI 服务器中 CPU+GPU 协同工作，相较普通服务器算力明显提升

**AI 训练模型算力提升速度突破极限，目前英伟达训练型 AI 服务器一般配备 8 个 GPU**。随着以 chatGPT 为代表的 AI 的发展，训练 GPT-3、Megatron-Turing NLG530B 等超大语言模型所要求的算力提升速度已经突破了后摩尔定律算力提升速度的极限，尽管 CPU 不断升级，但 CPU 制程以及单个 CPU 和核心数量接近极限，仅依靠 CPU 无法满足算力需求。CPU 的内核数量大约数十个，但 GPU 具备成千上万个 CUDA 核心，因此 GPU 多个内核决定了其能够在相同的价格和功率范围内，比 CPU 提供更高的指令吞吐量和内存带宽，GPU 能够并行执行成千上万个线程（摊销较慢的单线程性能以实现更大数据吞吐量）。在训练 AI 模型的过程中，需要同时对所有样本数据执行几乎相同的操作，GPU 架构设计能够很好满足 AI 场景需求。AI 服务器相较通用服务器的一个明显差别之一是增加了 GPU，通用服务器一般含有 1-2 个 CPU、不含 GPU，而当前英伟达训练型 AI 服务器一般搭载 8 个 GPU。

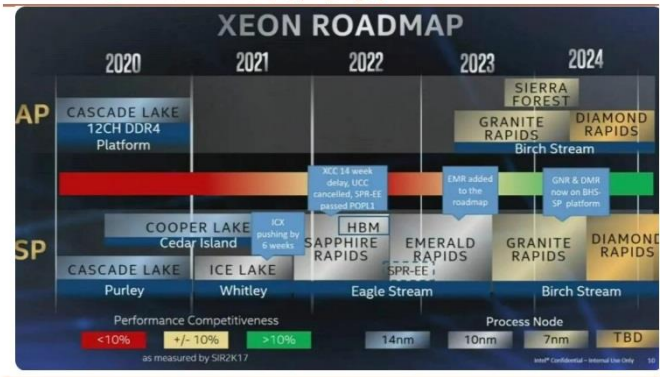
**AI 服务器 GPU 需要 CPU 来进行指令，模型算力提升带动 CPU 核心、主频等提升**。在 AI 服务器的 GPU 模式下，模型训练一般分为 4 步，将输入数据从系统内存拷贝到显存；CPU 指示 GPU 处理数据；GPU 并行完成一系列计算；将计算结果从显存拷贝到内存。虽然 GPU 并行能力优异但无法单独工作，必须由 CPU 进行控制调用，CPU 可以独立工作并直接访问内存数据完成计算。因此在 AI 服务器中，GPU 和 CPU 需要协同工作，训练模型所需算力升级也将带动 CPU 技术升级，例如在英伟达 DGX-2 服务器中，采用的英特尔第三代至强处理器 8168，主频大约 2.7GHz，核心数量为 24 个；在英伟达 DGXH100 服务器中，搭载英特尔第四代至强处理器 8480C，主频提升至最高 3.8GHz，CPU 核心数量提升至大约 56 个。

AI 训练模型的工作方式




资料来源：CSDN、招商证券

Intel XEON CPU 技术路线图




资料来源：Intel、招商证券

**AI 服务器存储器容量伴随 CPU/GPU 的升级而提升，相较传统服务器有数倍提升**。最先进的 AI 服务器尽管增加了大量 GPU 需求，但存储器的数据存储方式、总线连接方式均和普通服务器相近，CPU 的运行数据写入 DRAM 中，CPU 和 GPU 产生的数据共同写入 NAND 中。AI 服务器将提升内存、显存的工作频率和带宽等，带动存储容量明显上升。

**AI 服务器相较于传统服务器算力上大幅跃升**。AI 服务器利用 CPU+的架构模式，CPU 仍作为 CPU 的数据处理主要模块，同时植入并行式计算加速部件，如 ASIC、FPGA、GPU 等，负责人工智能计算负载加速。总而言之，在 CPU+架构下，AI 服务器的技术选型和部件配置针对不同的业务场景做相应的调整优化，通过合理的负载分担实现计算能力的提升。

AI 服务器与普通服务器相比具有更好的技术优势

| | AI 服务器 | 普通服务器 |
|---|---|---|
| 卡的数量 | 以加速卡为主导，基础要求为四块以上的 GPU 卡，甚至需要搭建外部服务器作为支持。 | 以 CPU 为主导，单卡/双卡 CPU。 |
| P2P 通讯 | GPU 卡间需要大量的参数通信，模型越复杂，通信量越大：<br>SXM3 协议下，P2P 带宽高值 300GB/s；<br>SXM2 协议下，P2P 带宽高值 50GB/s；<br>PCI3.0 协议下，P2P 带宽高值 32GB/s。 | 普通 GPU 服务器一般只要求单卡性能。 |
| 特有设计 | 全面考虑对存储、通信、网络等相关领域的技术方案进行合理配置，使之与计算部件的计算能力相匹配，避免出现性能瓶颈。 | - |
| 专有技术 | Purley 平台更大内存带宽；<br>NVlink 提供更大的互联带宽；<br>TensorCore 提供更强的 AI 计算力。 | - |

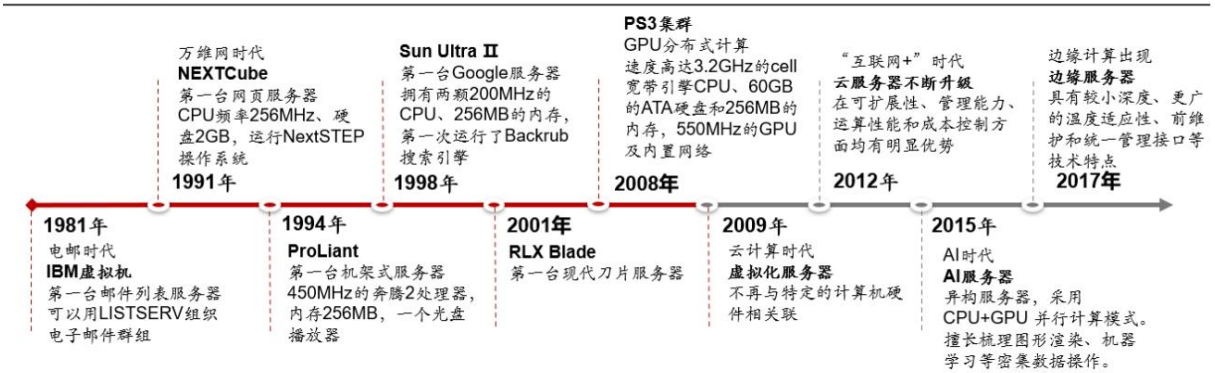
资料来源：南京铌前官网，天风证券研究所

## 2、服务器行业发展历程

**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




## 3、AI 服务器分类

### （1）按应用场景可分为训练和推理，训练对算力要求更高

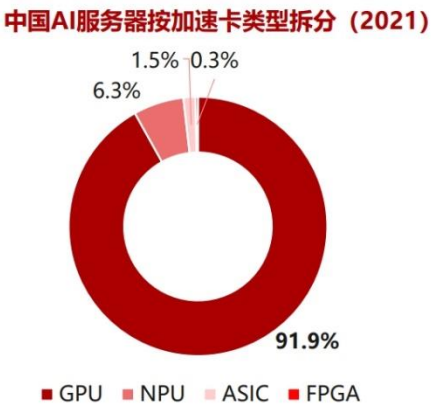
AI 服务器按应用场景可分为训练和推理两种，2021 年中国 AI 服务器推理负载占比约 55.5%，未来有望持续提高。其中训练对芯片算力要求更高，推理对算力的要求偏低。

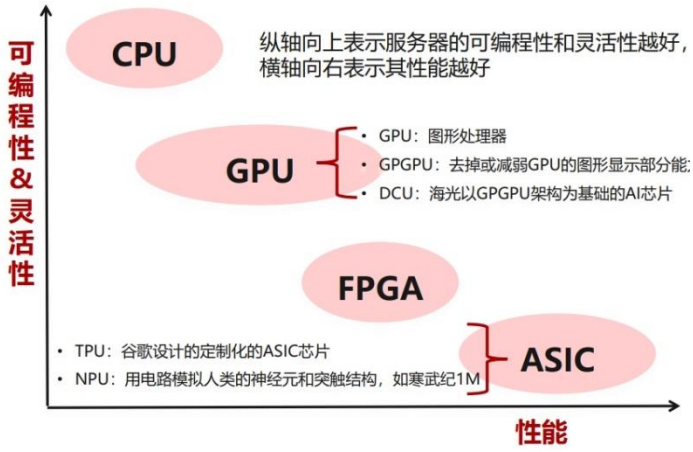
| | 训练 | 推理 |
|---|---|---|
| 概念 | 指借助已有的大量数据样本进行学习，获得诸如更准确的识别和分类等能力的过程； | 对于新的数据，使用经过训练的算法完成特定任务 |
| 算力要求 | 要求训练芯片应具有强大的单芯片计算能力 | 对算力的要求较低 |
| 部署位置 | 训练芯片大多部署于云端 | 推理芯片大多会部署于云端和边缘侧 |


资料来源：浪潮信息官网、IDC、电子发烧友、浙商证券研究所

### （2）按芯片类型可分为 GPU、FPGA、ASIC 等

AI 服务器采用异构形式，按芯片类型可分为 CPU+GPU、CPU+FPGA、CPU+ASIC 等组合。目前 GPU 依然是实现数据中心加速的首选，其他非 GPU 芯片应用逐渐增多，IDC 预计到 2025 年其他非 GPU 芯片占比超过 20%。一般来说，ASIC 的性能最好，但是可编程性和灵活性较弱；在训练或者通用情况下，GPU 则是更好的选择。




资料来源：浪潮信息官网、宽泛科技、CSDN、51CTO、浙商证券研究所

## 4、我国 AI 服务器进入快速增长期

**随着国内数字基础建设数据负载量的需求量不断上升，我国 AI 服务器市场保持较快增速**。根据 IDC 数据，2022 年大陆 AI 服务器出货量达 28.4 万台，预计到 2027 年达到 65 万台，CAGR 为 17.9%，按金额计算，2022 年大陆 AI 服务器销售额为 72.55 亿美元，预计到 2027 年销售额将达到 163.99 亿美元，CAGR 为 17.7%。

图表：2019-2027年大陆AI服务器出货量


图表：2019-2027年大陆AI服务器销售额


来源：IDC,中泰证券研究所



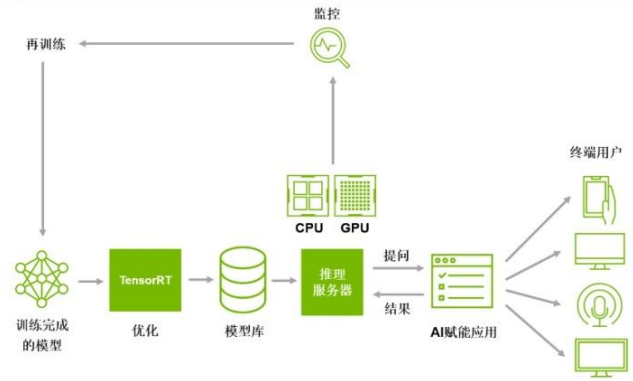
# 二、驱动因素

## 1、ChatGPT 等大模型训练和推理需求激增，驱动 AI 服务器市场高速增长

### （1）AI 模型分为训练和推理两个过程，不断涌现的 AI 大模型推动算力需求激增

**AI 模型主要分为训练和推理过程，训练奠定模型的性能根基，推理是将已有模型应用到具体场景对相应需求做出反馈的过程**。根据英伟达官网给出的示意图，AI 大模型需要利用构建好的算法，在大量的数据库上进行训练，借助大量的算力生成一个对于特定性能指标具有优异表现的模型结果。模型训练好之后在应用端通常称为推理过程，终端用户通过多种方式（包括文字、语音、图片、视频等多模态形式）针对模型提出需求，模型根据自己的理解给出反馈，在推理过程中实现的结果，还可以反过来针对模型进行进一步辅助训练。

AI 模型训练和推理应用原理

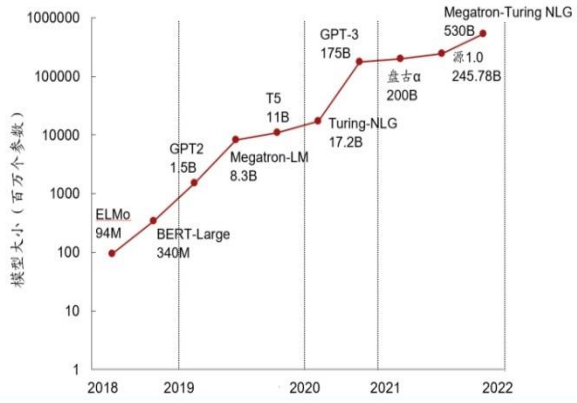



资料来源：英伟达官网，招商证券

**AI 大模型发展过程中，通常伴随着模型参数量增大、训练数据增多的趋势，对于芯片的算力需求持续增长**。根据《AI 算力集群方案设计与优化》总结的过去 4 年全球主要 NLP（自然语言处理）模型，模型的参数量从 ELMo 的 9400 万增长至 Megatron-Turing NLG 的 5300 亿，增长了近 5600 倍。以 GPT-1 到 GPT-3 的发展过程为例，2018 年 6 月 GPT-1 发布，GPT-1 预训练过程是无监督的，采用了 BooksCorpus 数据集，微调过程是有监督的，主要针对语言模型，整个模型参数量达到 1.17 亿，其中预训练数据量达到 5GB。GPT-2 于 2019 年 2 月发布，预训练过程同样是无监督的，采用多任务学习的方式，参数量提升至 15 亿，预训练数据量提升至 40GB。GPT-3 于 2020 年 5 月发布，通过更为海量的参数来进行训练和学习，参数量进一步提升至 1750 亿，预训练数据量提升数个数量级至 45TB。AI 模型的发展在目前阶段来看，更好的性能获取通常意味着更多的参数量和更大的数据集，AI 模型的迅猛发展

与芯片层面的算力进步密不可分，以 GPU 为代表的加速芯片快速迭代发展，为大模型更替奠定了良好的硬件基础。

典型自然语言处理网络模型参数量变化

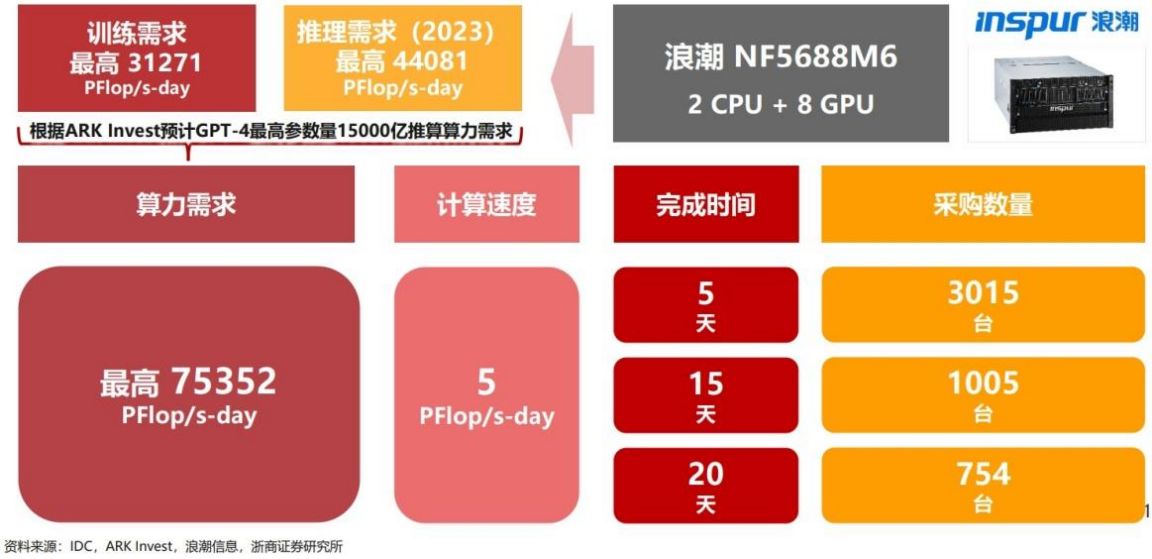



资料来源：《AI 算力集群方案设计与优化》，浪潮信息，招商证券

作为算力的发动机，在市场对算力需求持续增长的前提下，AI 服务器也必将伴随该产业趋势，迎来市场端规模上的快速增长。


资料来源：：《COMPUTE TRENDS ACROSS THREE ERAS OF MACHINE LEARNING》，Jaime等，华福证券研究所

**（2）以 GPT-4 为例，为满足算力需要近千台浪潮 NF5688M6 服务器**



资料来源：IDC，ARK Invest，浪潮信息，浙商证券研究所

## 2、政策支持将推动国内 AI 服务器进入快速增长期

2023 年中共中央、国务院印发《数字中国建设整体布局规划》，明确指出“夯实数字中国建设基础”，“数字基础设施”将拉动大数据中心、超算中心等基建建设，国内 AI 服务器规模有望迎来快速增长。

## 3、美国限制向中国出口先进 GPU，国产厂商崛起势在必行

### （1）美国限制向中国出口先进 GPU，可购买削弱带宽的 A800

**英伟达**推出的三代 GPU 芯片 V100、A100 和 H100 可用于 AI 模型训练和推理，最新一代的 H100 较 A100 计算速度快约 3 倍（67/19.5）。2022 年 8 月，美国要求英伟达停止向中国企业出售 A100 和 H100 两款 GPU 计算芯片，目前中国企业仅能购买特供的 A800 芯片，该芯片较 A100 在互联带宽方面被削弱 1/3，成为当前可行的替代方案。暂时未被禁售的 V100 工艺为 12nm，难以满足目前计算需求。

### （2）国产 GPU 单卡指标接近英伟达，推理应用更具竞争力

国产算力 GPU 的主要厂商包括**海光信息**、**寒武纪**、**平头哥**、**华为昇腾**、**天数智芯**、**燧原科技**、**摩尔线程**、**壁仞科技**、**沐曦**等公司，部分产品的单卡指标和参数已经与英伟达产品接近或持平。目前国产算力 GPU 芯片在推理场景应用较多且具备一定竞争力，如含光 800、思元 370、MTT S3000 等等。

| | | 寒武纪 | | | 平头哥 | 华为昇腾 | | 天数智芯 | 燧原科技 | 摩尔线程 | 壁仞科技 | | 海光信息 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 思元370 | 思元290 | 思元270 | 含光800 | 昇腾310 | 昇腾910 | 天垓100 | 云燧T20/T21 | MTT S3000 | 壁砺100P | 壁砺104P | DCU |
| 算力指标 | FP64 | | | | | | | | | | | | 10.8TFLOPS |
| | FP32 | 24TFLOPS | | | | | | 37/18.5TFLOPS | 32TFLOPS | 15.2TFLOPS | 240TFLOPS | | |
| | TF32 | | | | | | | | 128TFLOPS | | 480TFLOPS | 256TFLOPS | |
| | FP16 | 96TFLOPS | | | | | 320TFLOPS | 147/37TFLOPS | 128TFLOPS | | | | |
| | BF16 | 96TFLOPS | | | | | | | 128TFLOPS | | 960TFLOPS | 512TFLOPS | |
| | INT16 | 128TOPS | 256 TOPS | 64TOPS | 205TOPS | 8TOPS | 640TOPS | | | | | | |
| | INT8 | 256TOPS | 512 TOPS | 128 TOPS | 825TOPS | 16TOPS | | 295TOPS | 256TOPS | | 1920TOPS | 1024TOPS | |
| 内存容量 | | 24GB LPDDR5 | 32GB HBM2 | 16GB DDR4 | | | | 32GB HBM2 | 32GB HBM2E | 32GB GDDR6 | 64GB HBM2E | 32GB HBM2E | 32GB HBM2 |
| 内存带宽 | | 307.2 GB/s | 1228 GB/s | 102 GB/s | | | | | 1.6TB/s | 448GB/s | 1.64TB/s | 819GB/s | 1TB/s |
| 功耗 | | 150W | 350W | 70w | 276W | 8W | 310W | 250W | 300W | 250W | 450-550W | 300W | 260-350W |

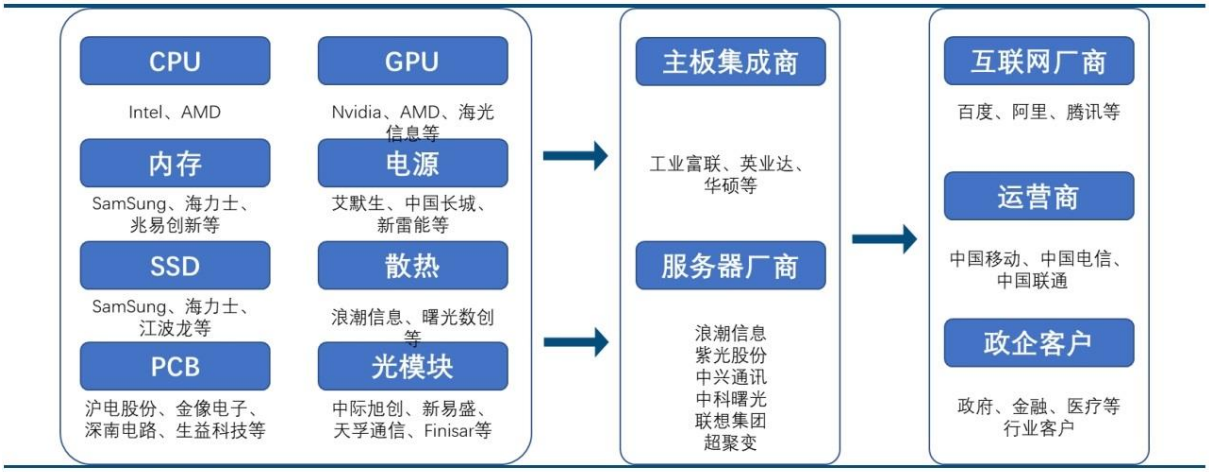
资料来源：各公司官网、海光信息招股书、浙商证券研究所

# 三、产业链分析

## 1、AI 服务器产业链概览

AI 服务器产业链上游主要由服务器元器件生产商组成，其中 CPU、GPU 作为核心组件，主要由 **Intel**、**AMD**、**Nvidia** 供应，国产供应商占比较少，其他部件包括内存、SSD、PCB、光模块、电源等存在更多的国产供应商；产业链中游包括主板集成商和服务器厂商，先由主板集成商将众多芯片集成，再交由服务器厂商装配成整机销售。目前国内企业在服务器厂商中占据重要地位；产业链下游主要包括以 BAT 为首的互联网厂商，移动、电信、联通三大运营商和众多政企客户（主要集中在政府、金融、医疗三大行业，因其最需要 AI 客服等相关产品）。

**AI 服务器产业链概览**




*数据来源：Wind，中信建投*

## 2、上游：芯片和存储是 AI 服务器的主要构成

**计算芯片和存储是服务器的核心构成**。芯片和存储作为 AI 服务器的核心，决定着 AI 服务器的算力和宽带大小。

**AI 服务器 AI 芯片价值量占比提升**。传统的通用型服务器中，售价 10424 美金的 2x Intel Sapphire Rapids Server，CPU 的成本占比约 17.7%，内存和硬盘占比超过 50%。而 AI 服务器，售价为 268495 美金的 Nvidia DGXH100 中，CPU 占比仅 1.9%，GPU 占比高达 72.6%。内存价值量提升，但占比下降至 4.2%左右。AI 服务器较通用服务器价值量提升明显，AI 芯片在 AI 服务器中占有绝对比重。随着 AI 服务器放量，AI 芯片正迎来黄金爆发期。

### （1）CPU：X86 为主，ARM 等其他架构争抢份额

**服务器 CPU 架构包括 X86、ARM、MIPS 和 RISC-V 等**。目前 X86 架构处理器统治着 PC 和服务器市场，Arm 架构处理器统治着移动市场和 IoT 市场，MIPS 是基于 RISC 的衍生架构之一，从工作站、桌面电脑到嵌入式系统再到人工智能，一直在夹缝中求生。近年来 RISC-V 架构则凭借着开源、指令精简、可扩展等优势，在注重能效比的物联网领域大受追捧，并也开始进入更高性能需求的服务器市场。

**英特尔是服务器市场的龙头企业，2022 年仍占据全球服务器市场 70.77%的份额**。**AMD** 主要产品是 EPYC（霄龙）系列 CPU，2022 年占据 19.84%的市场份额。**海光信息**通过与 AMD 成立合资公司成都

海光集成的方式，变相拥有了 X86 架构的授权，不过海光信息仅获得 AMD 第一代 EPYC 的 Zen 架构，没有获得 Zen2、Zen3 系列架构授权。**上海兆芯**通过引入台湾**威盛控股**获得了部分 X86 架构授权，不过盛威与英特尔的 X86 合同已于 2018 年 4 月到期，不能使用英特尔新的 X86 专利，只能在旧 X86 架构下继续研发。**海思半导体**的鲲鹏处理器和**天津飞腾**处理器兼容 ARM 指令集；**龙芯中科**处理器采用 LoongArch 指令集，主要产品与服务涵盖处理器及配套芯片产品；**成都申威**处理器采用 SW-64 指令集，主要应用于服务器、桌面计算机等设备。

服务器CPU主要架构参与者

| | 参与者 | 说明 |
|---|---|---|
| X86 | 国外 intel、AMD | Intel：2022年占据全球服务器芯片市场70.77%左右的份额，Xeon(至强)系列 |
| | | AMD：致力于X86高端服务器，拥有Zen系列 |
| | 国内海光信息、上海兆芯、成都申威 | 海光信息：海光7000高端系列、5000中端系列和海光3000低端系列 |
| | | 上海兆芯：KH-40000系列、30000系列、20000系列 |
| | | 成都申威：申威411、421、1621等系列 |
| Arm | 国外 高通、Cavium、Amazon | 高通：基于ARM架构定制的微处理器内核 |
| | | Cavium：多核MIPS和ARM处理器提供商，处理器广泛应用于网络、通讯等领域的安全产品 |
| | | Amazon：收购以色列Annapurna Labs，推出自研Graviton2处理器 |
| | 国内 华为海思、飞腾 | 华为海思：基于ARM架构，已推出7nm产品 |
| | | 飞腾（中国长城）：源自国防科大，专注ARM架构，服务于“天河”系列计算机，达到世界主流水准 |
| MIPS | 龙芯中科 | 源于中科院计算所，立足党政军市场 |
| RISC-V | 算能科技 | 基于RISC-V 的64核服务器级CPU |

资料来源：各公司官网，Counterpoint，国际电子商情，电子信息技术，全球半导体观察，快科技，龙芯中科招股说明书，华福证券研究所

各服务器CPU代表芯片

| | intel | AMD | 海光 | 兆芯 | 海思 | 飞腾 | 龙芯 | 申威 |
|---|---|---|---|---|---|---|---|---|
| 品牌 | Xeon 8592 | EPYC9654 | 海光7285 | KH-40000系列 | 鲲鹏920-7260 | S2500 | 企业级3C5000L | 申威1621 |
| 指令集 | X86 | X86 | X86 | X86 | ARM | ARM | LoongArch | X86 |
| 核心数 | 64 | 96 | 32 | 32 | 64 | 64 | 16 | 16 |
| 超线程 | 128 | 192 | 64 | 32 | 不支持 | 不支持 | 不支持 | 不支持 |
| 主频 | 3.9GHZ | 3.7GHZ | 2.0GHZ | 2.7GHz | GHZ | GHZ | GHZ | GHZ |
| 内存类型 | DDR5 | DDR5 | DDR4 | DDR4 | DDR4 | DDR4 | DDR4 | DDR3 |
| 内存通道数 | 8 | 12 | 8 | 8 | 8 | 8 | 4 | 8 |
| 最高内存频率 | 5600MHZ | 4800MHZ | 2666MHZ | 3200MHZ | MHZ | MHZ | MHZ | MHZ |
| PCIe通道数 | 128 | 128 | 128 | 128 | 40 | 17 | 32 | 16 |

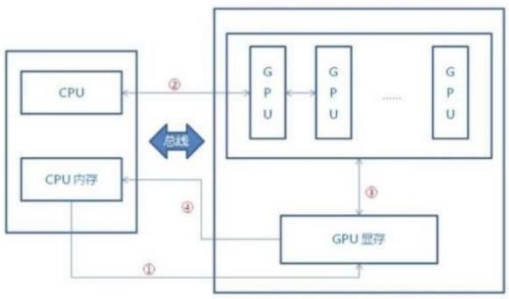
资料来源：各公司官网，海光信息招股说明书，华福证券研究所

## （2）AI 芯片是 AI 服务器的大脑

**AI 服务器采取 GPU 架构，适合大规模并行计算**。AI 服务器由传统服务器演变发展而来。相比于通用服务器，AI 服务器为异构服务器，可以多种组合方式，搭载多个 GPU、CPU 以及大算力 AI 芯片，极大程度解决传统服务器算力不足的缺点。AI 服务器采用 GPU 架构，GPU 具有众多计算单元和长流水线，简单控制逻辑，省去 Cache。面对类型统一、相互无依赖的大规模数据，处于无需中断的计算环境。相较之下，CPU 被 Cache 和复杂控制逻辑占据，通用性导致复杂的内部结构，处理不同数据类型引入分支和中断。

**AI 芯片是 AI 服务器算力的核心**。AI 芯片是 AI 服务器算力的核心，也被称为 AI 加速器或计算卡，专门用于处理人工智能应用中的大量计算任务。按技术架构分类，AI 芯片可分为 GPU、FPGA、ASIC 和 NPU 等。GPU 是一种通用型芯片，ASIC 是一种专用型芯片，而 FPGA 则处于两者之间，具有半定制化的特点。按照功能分类，可分为训练和推理芯片。按照应用场景分类，可分为云端和边缘端芯片。随着 AIPC、AIPIN、AIPHONE 等更多应用场景出现，AI 芯片的空间有望进一步打开。

AI服务器由CPU+GPU架构组成




资料来源：中国通信标准化协会，华福证券研究所

主要AI芯片对比

| | GPU | FPGA | ASIC | NPU |
|---|---|---|---|---|
| 特点 | 通用型 | 半定制化 | 专用型 | 模拟人脑 |
| 芯片架构 | 叠加大量计算单元和高速内存，逻辑控制单元简单 | 具备可重构数字门电路和存储器，根据应用定制 | 电路结构可根据特定领域应用和特定算法定制 | - |
| 擅长领域 | 3D图像处理，密集型并行运算 | 算法更新频繁或者市场规模较小的专用领域 | 市场需求量大的专用领域 | 适用于各种具体行业 |
| 优点 | 计算能力强，通用性强开发周期短，难度小，风险低 | 功能可修改，高性能、功耗远低于GPU，一次性成本低本低 | 专业性强、性能高于FPGA、功耗低、量产成本低 | 最低功耗；通信效率高；认知能力强 |
| 缺点 | 价格贵，功耗高 | 编程门槛高、量产成本高 | 开发周期长，难度大，风险高，一次性成本高 | 处于探索阶段 |
| 代表企业 | 英伟达、AMD、景嘉微、海光信息 | 赛灵思、英特尔、百度 | 谷歌、寒武纪 | 英特尔 |

资料来源：亿欧智库，智能计算芯世界、各公司官网，华福证券研究所

## （3）存储：内存容量大幅提升，HBM 成 AI 服务器标配

### 1）HBM 相较 GDDR 更适用于 AI 服务器

**HBM 芯片适用于高性能要求的 AI 训练计算**。处理器的性能以每年大约 55%速度快速提升，而内存性能的提升速度则只有每年 10%左右。不均衡的发展速度造成了当前内存的存取速度严重滞后于处理器的计算速度。高性能处理器难以发挥出应有的功效。GDDR5 作为通用内存，容量较小、位宽低且远离 CPU 或 SoC，由于无法跟上 GPU 性能的增长速度及不断上升的功耗，已经无法满足高性能计算场景对带宽的需要。**HBM（High-Bandwidth Memory）本质还是一种内存产品**，可以理解为与 CPU 或 SoC 对应的内存层级，将原本在 PCB 板上的 DDR 和 GPU 芯片同时集成到 SiP 封装中，使内存更加靠近 GPU，使用 HBM 可以将 DRAM 和处理器（CPU，GPU 以及其他 ASIC）之间的通信带宽大大提升，从而缓解这些处理器的内存墙问题。目前 HBM 已经成为高端 GPU 的标配，同时也应用于部分针对云端处理的 AI 芯片（例如谷歌的 TPU）中，未来有望拓展至更多应用场景。

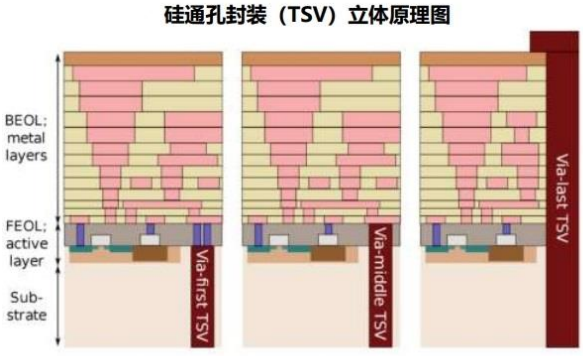
**HBM 更高速、更高带宽、更高位宽等优异性能将引爆市场需求**。凭借独特的 TSV 信号纵向连接技术，HBM 内部将数个 DRAM 芯片在缓冲芯片上进行立体堆叠，其内部堆叠的 DDR 层数可达 4 层、8 层以至 12 层，从而形成大容量、高位宽的 DDR 组合阵列

### 2）HBM 的优异性能离不开硅通孔技术与 CoWos 封装技术的发展

**TSV 技术具有高密度集成、高电性能、多功能集成和低制造成本等优势**。HBM 通过 SIP 和 TSV 技术将数个 DRAM 裸片像楼层一样垂直堆叠。**台积电** CoWoS-S 通过硅中介层承载处理器和 HBM。HBM 与处理器“组装”在一起需要借助硅中介层。HBM 通过 CoWoS 等 2.5D 封装工艺，和 CPU/GPU 等并排铺设在硅中介基板上，CPU/GPU 等逻辑 die 采用倒片封装（FC）形式和硅中介基板连接，存储器和 GPU 等逻辑芯片之间通过中介层实现通信，然后将内插器和有源硅连接到包含要放置在系统 PCB 上的 I/O 的封装基板。

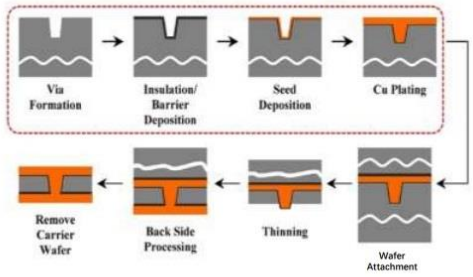
**HBM 工艺流程包括 TSV 形成、绝缘层全气绝缘、阻挡层、种子层沉积、电镀填充、CMP 抛光等步骤**。TSV 为 HBM 核心工艺，中文全称是硅通孔技术。TSV 技术通过铜、钨、多晶硅等导电物质的填充，实现硅通孔的垂直电气互联。HBM 对封装高度、散热性能提出更高要求，3D 封装关键原材料成为核心。在 3D 封装关键原材料方面，颗粒状环氧塑封料（GMC）HBM 由于 3D 堆叠导致芯片厚度较高，因此需要用特殊的颗粒状环氧塑封料（GMC）封装。为了解决 HBM 封装厚度增大和散热需求大的问题，GMC 需要大量添加核心材料 low-α 球硅和 low-α 球铝。

硅通孔封装（TSV）立体原理图




资料来源：未来半导体，华福证券研究所

硅通孔封装（TSV）的工艺流程




资料来源：艾邦半导体，华福证券研究所

### （4）上游供应危机尚未解除，国产替代提上日程

**人工智能芯片搭载率将持续增高，GPU 仍为主流方案**。据 IDC 调研显示，当前每台人工智能服务器上配备 2 个 GPU、3 个 FPGA 或 3 个 ASIC 的比例最高，未来 18 个月，比例最高的服务器有望配备 4 个

GPU、7 个 FPGA 或 5 个 ASIC，普遍搭载率均呈上升趋势。再看中国市场，目前在国内市场主要是用以 GPU 为主实现数据中心计算加速，市场占有率近 90%，这主要是因为 GPU 可以较好支持高度并行的工作负载。ASIC、FPGA、NPU 等非 GPU 芯片市场占有率超过 10%，得益于近年智慧城市建设、无人驾驶载具、智慧医疗系统构建、智能家居等成为热门领域，应用于该类领域的非 GPU 芯片也得到发展。未来面对需求的多元增长，AI 芯片将呈现百花齐放的发展空间。

**GPU 海外寡头垄断格局+禁运风险或为国产 AI 服务器的主要瓶颈**。GPU 主要分为独立 GPU 和集成 GPU，前者用于 AI 服务器、高性能电脑中，后者则主要用于移动端设备。目前 **Nvidia** 和 **AMD** 垄断独立 GPU 市场，其中 Nvidia 优势更为明显，2021Q1 市占率达到 83%。同时，据电子发烧网，Nvidia 的 GPU 芯片是 AI 大模型的关键，在大模型训练市场的市占比近 100%，而 GPT-3.5 大模型需要高达 2 万枚 GPU，未来商业化后或将超过 3 万枚。同时国内如**浪潮**、**宁畅**等国内品牌厂商的 AI 服务器中同样配置 Nvidia 的芯片。受到中美脱钩的持续影响，部分供应 AI 服务器的 GPU 成为限制出口的产品，直接影响国内 AI 服务器的出货量。根据美国商务部工业与安全局宣布的针对中国出口先进芯片的管制新规声明，凡输入/输出（I/O）双向传输速度高于 600GB/s，同时每次操作的比特长度乘以 TOPS 计算出的处理性能大于或等于 4800 的产品，将无法出口至中国，英伟达的 A100 即属于限制范围之内。从 AI 芯片行业投融资来看，目前国内 AI 芯片产业热度持续高涨，根据 IT 桔子的数据，2022 年中国 AI 芯片行业投融资额达到 179.5 亿元，资本的持续进入有望加速国内 GPU 国产化进程，逐步切入 AI 服务器的供应链中。

**国内 AI 服务器所搭载的 GPU 厂商主要以英伟达为主**

| 品牌厂商 | 型号 | GPU |
|---|---|---|
| 浪潮信息 | NF5688M6 | 8 个 NVIDIA 最新的 NVSwitch 全互联 500W Ampere 架构 GPU |
| 宁畅 | X660 G45 LP | 8 个 NVIDIA 的 A800 |
| 拓维信息 | 兆瀚 AI 推理服务器 | 8 张 Atlas300I（国产芯片） |
| 宝德科技 | PR4910W | 8 个 NVIDIA A800/10 个 NVIDIA 的 A40/30 |

资料来源：各公司官网，华为计算公众号，天风证券研究所

**国内厂商正在陆续推出 GPU 产品进行市场检验，国产替代提上日程**。伴随资本和政策的持续加码，一批国内 GPU 厂商逐渐崭露头角。然而，我们仍需看到在芯片设计制造领域，我国缺乏设计软件、先进制程及设备，与世界领先水平之间尚有差距，该领域部分产品及装备仍十分依赖进口，国产 GPU 之路仍是路漫漫其修远兮。

国内生产供应服务器的 GPU 厂商

| 公司 | 进展 | 性能 |
| --- | --- | --- |
| 景嘉微（300474.SZ） | 第三代 GPU 芯片 JM9 系列的两款产品分别于 2021 年 11 月 16 日和 2022 年 6 月 28 日完成阶段性测试工作 | 用于地理信息系统、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。 |
| 龙芯中科（688047.SH） | 2022 年推出 16 核芯片产品 3C5000，32 核 3D5000 研制成功 | 主要用于存储服务器，取得石油、石化等客户的突破； |
| 海光信息（688041.SH） | 海光深算一号 DCU 实现商业化应用，深算二号正在研发中 | 广泛用于大数据处理、人工智能、商业计算等计算密集类应用领域，主要部署在服务器集群或数据中心，为应用程序提供高性能、高能效比的算力，支撑高复杂度和高吞吐量的数据处理任务。 |
| 芯原股份（688521.SH） | 推出 Vivante3DGPGPUIP | 提供从低功耗嵌入式设备到高性能服务器的计算能力，满足广泛的 AI 计算需求。 |
| 天数智芯 | 2021 年 3 月发布了首款通用 GPU 训练芯片天垓 100 | 广泛支持传统机器学习、数学运算、加解密及数字信号处理等领域； |
| | 2022 年 12 月发布通用 GPU 智铠 100 | 支持国内外主流深度学习框架，拥有丰富编程接口拓展和高性能函数库，广泛适用于智慧城市、智慧港口、智慧交通等众多领域。 |
| 壁仞科技 | 2022 年 9 月发布首款通用 GPU 芯片 BR100 | 算力创下全球记录，率先采用 Chiplet 技术。 |

资料来源：Wind，电子发烧网公众号，天天 IC 公众号，天数智芯公众号，天风证券研究所

## 3、中游：当前 AI 服务器厂商在手订单充分，AI 服务器市场高增长确定性较强

自去年 ChatGPT 带动的大模型浪潮以来，国内外头部互联网厂商纷纷加入 AI 算力的军备竞赛，加大对于 AI 算力侧的资源投入。AI 算力的高景气带动 AI 服务器需求端爆发式增长，并体现在 AI 服务器厂商订单端。

全球 AI 服务器出货金额排名第一位的龙头厂商**浪潮信息**，提到一季度以来 AI 服务器市场迎来明显增长，客户关注点由价格转向能否及时满足自身需求。此外，据**紫光股份**于投资者互动平台的回复，其 AI 服务器订单今年一季度有很大提升，产能满足市场需求不存在问题，针对 GPT 场景优化的 GPU 服务器已经完成开发，预计今年二季度全面上市。作为全球 ICT 设备龙头企业的**联想集团**，根据其最新公布的财报数据，ISG（基础设施解决方案业务集团）在 2023 年 1-3 月实现营收同比增长 56.2%，全财年营收同比增长 36.6%，主要受益于海外 AI 服务器需求爆发以及存储业务的高速增长，公司预期新财年 AI 服务器收入增速将显著快于通用服务器，带动 ISG 部门营收增长超市场平均水平 20%以上。**中科曙光**深度布局算力领域，包括上游芯片、中游服务器解决方案、液冷技术、以及下游算力调度等业务，公司于投资者互动平台多次回复，会根据用户需求提供通用算力和智能算力产品及服务，随着我国算力需求的增