（报告出品方/作者：长江证券，杨洋、钟智铧）

ChatGPT：生成式AI引爆技术奇点

AIGC全称为AI-Generated Content，指基于生成对抗网络GAN、大型预训练模型等人工智能技术，通过已有数据寻找规律，并通过适当的泛化能力生成相关内容的技术。与之相类似的概念还包括Synthetic media，合成式媒体，主要指基于AI生成的文字、图像、音频等。

2020年，1750亿参数的GPT-3在问答、摘要、翻译、续写等语言类任务上均展现出了优秀的通用能力，证明了“大力出奇迹”在语言类模型上的可行性。自此 之后，海量数据、更多参数、多元的数据采集渠道等成为国内清华大学、智源研究院、达摩院、华为、北京大学、百度等参与者的关注点。

2022年12月，ChatGPT 3.5令人惊艳的使用体验引爆社会热潮，搜索热度和用户增长都出现了极为明显的提升。

目前，大型文本预训练模型作为底层工具，商业变现能力逐渐清晰。以GPT-3为例，其文本生成能力已被直接应用于Writesonic、 Conversion.ai、 Snazzy AI、Copysmith、 Copy.ai、 Headlime等文本写作/编辑工具中。同时也被作为部分文本内容的提供方，服务于AI dungeon等文本具有重要意义的延展应用领域。

过去传统的人工智能偏向于分析能力，即通过分析一组数据，发现其中的规律和模式并用于其他多种用途，比如应用最为广泛的个性化推荐算法。而现在人工智 能正在生成新的东西，而不是仅仅局限于分析已经存在的东西，实现了人工智能从感知理解世界到生成创造世界的跃迁。因此，从这个意义上来看，广义的 AIGC 可以看作是像人类一样具备生成创造能力的AI技术，即生成式AI它可以基于训练数据和生成算法模型，自主生成创造新的文本、图像、音乐、视频、3D 交 互内容(如虚拟化身、虚拟物品、虚拟环境)等各种形式的内容和数据，以及包括开启科学新发现创造新的价值和意义等。因此，AIGC 已经加速成为了AI 领域的新疆域，推动人工智能迎来下一个时代。

人工智能带来的生产力变革风声已近

追求生产力的提升和生产关系的优化，是人类社会发展的根源动力和核心目标，而生产力升级的最本质目标就是效率提升和成本降低。从人类社会四次工业/科技 革命来看，第一次工业革命的核心成果是以蒸汽机为代表的机械替代人力，第二次工业革命是以电力、燃油为代表的能源突破，第三次是以计算机及信息技术为 代表的信息结构性变革和自动化生产，其共同的特征就是生产规模的不断扩大、生产方式

上科技应用不断地向工业和社会的更高层结构渗透。底层的、低技术含 量的、规模庞大的生产模块不断被机器替代，人力生产持续向高层的、复杂的、尖端的生产方式和技术模块演进，是一个不变的趋势。

人工智能三要素逐步成熟，推动行业进入爆发期

AIGC的本质是内容与场景，其发展需要AI与后端基建，算法、算据和算力三要素耦合共振。AIGC的三大发展阶段是： 模型赋智阶段(从现实生成数字）：AIGC利用AI技术构建模拟现实世界的数字孪生模型；
认知交互阶段(从数字生成数字）：A能够学习并创作更丰富的内容；空间赋能阶段(从数字生成现实）：AIGC基于物联网，多模态技术获取多维信息，实现更加智能的人与机器互动。

市场规模：2021 年，全球人工智能市场收支规模(含硬件、软件及服务)达 850 亿美元。IDC 预测，2022 年该市场规模将同比增长约 20%至1017亿美元，并将于 2025 年突破 2000 亿美元大关，CAGR 达24.5%，显示出强劲的产业化增长势头。2021年，中国人工智能市场收支规模达到 82亿美元，占全球市场规模的 9.6%，在全球人工智能产业化地区中仅次于美国及欧盟，位居全球第三。IDC 预测，2022 年该市场规模将同比增长约 24%至 102 亿美元，并将于 2025 年突破 160 亿美元。

大模型参数量快速提升，算力需求大幅增加

大模型主要由各大龙头企业推动，在国内科技公司中，阿里巴巴达摩院在2020年推出了M6大模型，百度在2021年推出了文心大模型，腾讯在2022年推出了混元AI大模型。 大模型最核心的除了算法外主要是参数的设置，其中参数量(Params)形容模型的大小程度，类似于算法中的空间复杂度,往往参数量越大(复杂程度越高)的神经网 络模型对算力的需求程度更高，复杂的神经网络模型的算法参数量约千亿级别甚至万亿级别，与已知应用级别的呈现指数级别的差异。这些模型不仅在参数量上 达到了千亿级别，而且数据集规模也高达TB级别，想要完成这些大模型的训练，就至少需要投入超过1000PetaFlop/s-day的计算资源。

大模型主要由各大龙头企业推动，在国内科技公司中，阿里巴巴达摩院在2020年推出了M6大模型，百度在2021年推出了文心大模型，腾讯在2022年推出了混元AI大模型。 大模型最核心的除了算法外主要是参数的设置，其中参数量(Params)形容模型的大小程度，类似于算法中的空间复杂度,往往参数量越大(复杂程度越高)的神经网 络模型对算力的需求程度更高，复杂的神经网络模型的算法参数量约千亿级别甚至万亿级别，与已知应用级别的呈现指数级别的差异。这些模型不仅在参数量上 达到了千亿级别，而且数据集规模也高达TB级别，想要完成这些大模型的训

练，就至少需要投入超过1000PetaFlop/s-day的计算资源。

人工智能技术将全面赋能各行各业。预计到2025年，人工智能涉及的场景规模将达到2,081亿美金，并在无人驾驶、智慧金融、智慧医疗、智慧零售、文娱等领 域大显身手。人工智能技术对于算力的核心拉动点在于未来各应用场景内单设备芯片算力的增长和人工智能技术的行业渗透率的进一步提升，带动对云计算中心、边缘设备和终端NPU的巨大需求。整体预计在 2030年，人工智能相关领域对于算力的需求将达到~16,000EFLOPS，相当于1,600亿颗高通骁龙855内置的人工 智能芯片所能提供的算力。

AI+Chiplet： 信息革命的基石

应用-软件-硬件循环向上，AI芯片发展多元变化

以人工智能芯片为例，目前主要有两种发展路径：一种是延续传统计算架构，加速硬件计算能力，主要以CPU、GPU、FPGA、ASIC为代表。当前阶段，GPU配 合CPU是AI芯片的主流，而后随着视觉、语音、深度学习的算法在FPGA以及ASIC芯片上的不断优化，此两者也将逐步占有更多的市场份额，从而与GPU达成长期共存的局面。 深度神经网络算法是大型多层的网络模型，典型的有循环神经网络和卷积神经网络，模型单次推断通常需要数十亿甚至上百亿次的运算，对芯片的计算力提出了 更高要求，同时对器件的体积、功耗还有一定的约束。

GPU：并行运算带来对AI应用的高度适配

在架构上GPU由数以千计的更小、更高效的核心（类似于CPU中的ALU）组成，这些核心专为同时处理多任务而设计。现在的CPU，一般是多核（multi-core）结构；而 GPU 一般是众核（many-core）结构。

为充分利用GPU的计算能力，NVIDIA在2006年推出了CUDA（Computer Unifie Device Architecture，统一计算设备架构）这一编程架构。该架构使GPU能够 解决复杂的计算问题。它包含了CUDA指令集架构（ISA）以及GPU内部的并行计算引擎。英伟达在GPU里面增加了Tensor Core为AI服务，它的并行力度就从基本的数据点进化到以小矩阵快来进行计算。所以Tensor Core最基本的并行单元是一个4×4的矩阵块，能够在一个时钟周期里面算出一个4×4矩阵和另一个4×4矩 阵相乘的结果。原来用数据点来并行的话，它需要16次这样的计算，才能算出一个4×4的矩阵。相比之下Tensor Core的算力比原来的GPU要高，等价的功耗等比原来GPU的要低，这就是Tensor Core用来做矩阵一个显著的进化。

英伟达Ampere GA100是迄今为止 设计的最大的7nm GPU。GPU完全

针对HPC市场而设计，具有科学研究， 人工智能，深度神经网络和AI推理等应用程序。NVIDIA A100基于7nm Ampere GA100 GPU，具有6912 CUDA内核和 432 Tensor Core ， 540亿个晶体管数，108个流式多处理器。采用第三代NVLINK，GPU和 服务器双向带宽为4.8 TB/s，GPU间的互连速度为600 GB/s。另外， Tesla A100在5120条内存总线上的HBM2内存可达40GB。

2022年，NVIDIA推出了具有采用全 新Hopper架构的, 800亿个晶体管的H100，这是首款支持Pcle5.0标准的 GPU，单个H100就支持40Tb/s的IO 带宽。

英伟达：三重壁垒构造AI时代软硬件一体化龙头

第一层壁垒：硬件层。GPU奠定图形渲染和AI算力基础，英伟达硬件层的三芯战略已逐步成型：GPU解决AI大规模并行运算痛点，DPU解决AI训练推理中设备网络通信与CPU负荷问题，CPU填上三芯战略最后一块拼图，GPU强 耦 合 设 计 构 造 完 整 AI 解决方案 ， NVlink+NVSwitch+ConnectX突破芯片直连和设备网络连接限制，GPUDirect Storage 技术实现高性能存储和数据 访问

第二层壁垒：软件层。CUDA释放GPU潜力引航AI发展，DOCA、Omniverse等软件层进一步填充生态，增强AI行业对英伟达的粘性。CUDA从底层代码出发发挥GPU并行运 算优势 ， 奠 定 近 十 年 人 工 智 能 发 展 基 础 ， DOCA 为 BlueField DPU量身定做软件开发平台，复刻GPU+CUDA的强耦合成功路径，Omniverse初试工业共享虚拟空间，从硬件→软件→云上社区，在强劲软硬件基础上打造系统级 AI生态圈，NVIDIA AI Enterprise加速AI模型开发，未来或 有望助力实现以AI开发AI。

第三层壁垒：应用层。游戏显卡、数据中心、自动驾驶、元宇宙先后接力，十年成长曲线浪潮叠加。

景嘉微：构造图形GPU国产化基础

景嘉微在图形处理芯片领域经过多年的技术钻研，成功自主研发了一系列具有自主知识产权的GPU芯片，是公司图形显控模块产品的核心部件并以此在行业内 形成了核心技术优势。公司以JM5400研发成功为起点，不断研发更为先进且适用更为广泛的一系列GPU芯片，随着公司JM7200和JM9系列图形处理芯片的 成功研发，公司联合国内主要CPU、整机厂商、操作系统、行业应用厂商等开展适配与调试工作，共同构建国产化计算机应用生态，在通用领域成功实现广泛 应用。2022年5月，

公司JM9系列第二款图形处理芯片成功研发，可以满足地理信息系统、媒体处理、CAD辅助设计、游戏、虚拟化等高性能显示需求和人工 智能计算需求，可广泛应用于用于台式机、笔记本、一体机、服务器、工控机、自助终端等设备。

沐曦：顶级团队布局全栈解决方案

沐曦2020年9月成立于上海，拥有技术完备、设计和产业化经验丰富的团队，核心成员平均拥有近20年高性能GPU产品端到端研发经验，曾主导过十多款世界主流高性能GPU产品研发，包括GPU架构定义、GPU IP设计、GPU SoC设计及GPU系统解决方案的量产交付全流程。

打造全栈GPU芯片产品，推出MXN系列GPU（曦思）用于AI推理，MXC系列GPU（曦云）用于AI训练及通用计算，以及MXG系列GPU（曦彩）用于图形渲 染，满足数据中心对“高能效”和“高通用性”的算力需求。沐曦产品均采用完全自主研发的GPU IP，拥有完全自主知识产权的指令集和架构，配以兼容主 流GPU生态的完整软件栈（MXMACA），具备高能效和高通用性的天然优势，能够为客户构建软硬件一体的全面生态解决方案，是“双碳”背景下推动数据中心建设和产业数字化、智能化转型升级的算力基石。

AI还可以买什么？

服务器：AI驱动的硬件军备竞赛

目前，人工智能商业价值在全球范围内获得广泛认可，行业用户对于AI价值的认知、技术供应商在AI落地的方法论与实践方面日趋成熟。随着人工智能产业化应用的加速发展，全球AI基础设施支出持续呈现高增长态势。 据TrendForce，截至2022 年，预计搭载 GPGPU（General Purpose GPU） 的 AI 服务器年出货量占整体服务器比重近 1%；2023 年预计在 ChatBot 相关应用加持下，预估出货量同比增长可达 8%；2022-2026 年复合增长率将达 10.8%。 据 IDC ， 2026 年 预 计 全 球 AI 服务器市场规模将达 347 亿 美 元 ，2020~2026E年间复合增速达17.3%。

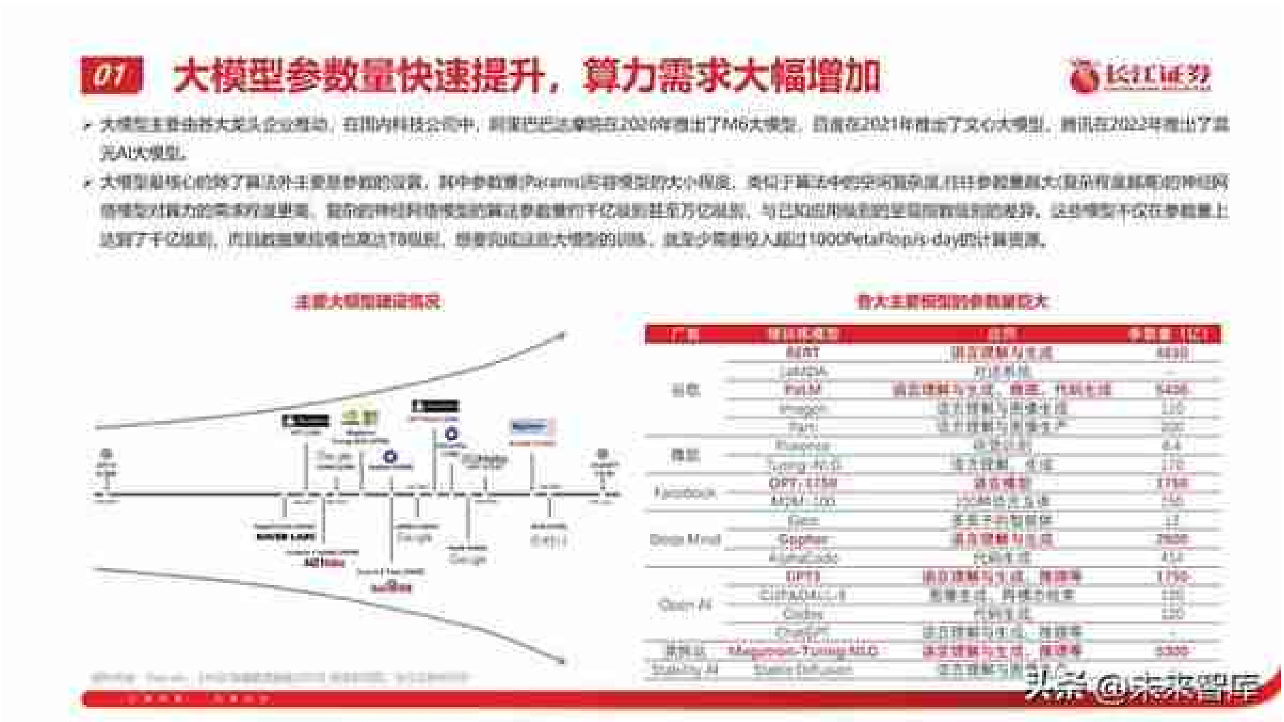
算力芯片以外的服务器投资方向梳理

服务器的硬件主要包括：处理器、内存、芯片组、I/O（RAID卡、网卡、HBA卡）、硬盘、机箱（电源、风扇）。 在硬件的成本构成上，CPU及芯片组、内存、外部存储是大头。以一台普通的服务器生产成本为例，CPU及芯片组大致占比50%左右，内存大致占比 15% 左 右，外部存储大致占比10%左右，其他硬件占比25%左右。AI服务器中GPU的占比则远较其他成本高。

报告节选：


01 ChatGPT：生成式AI引爆技术奇点
长江证券
ChatGPT搜索热度显著下降
各应用1亿用户达成时长（月）
头条@未来智库


01 人工智能带来的生产力变革风声已近
长江证券
历次工业革命
中国人工智能场景发展程度
头条@未来智库


01 大模型参数量快速提升，算力需求大幅增加
长江证券
大模型主要由各大龙头企业推动，在国内科技公司中，阿里巴巴达摩院在2020年推出了M6大模型，百度在2021年推出了文心大模型，腾讯在2022年推出了混元AI大模型。
大模型最核心的除了算法外主要是参数的设置，其中参数量(Params)是指模型的大小程度，类似于算法中的空间复杂度，往往参数量越大(复杂程度越高)的神经网络模型对算力的需求就越更高。现在的神经网络模型的算法参数量有千亿级别甚至万亿级别，与已知应用领域的呈指数级别的趋势。这些模型不仅在参数量上达到了千亿级别，而且数据集规模也高达TB级别，想要完成这些大模型的训练，就至少需要投入超过1000PetaFlop/s-day的计算资源。
各大主要模型的参数量巨大
谷歌 BERT 语言理解与生成 4810
LaMDA 对话系统 -
PaLM 语言理解与生成、推理、代码生成 5400
Imagen 语言理解与图像生成 110
Parti 语言理解与图像生产 200
微软 Florence 视觉识别 6.4
Turing-NLG 语言理解、生成 170
Facebook OPT-175B 语言模型 1750
M2M-100 100种语言互译 150
Deep Mind Gato 多面手的智能体 12
Gopher 语言理解与生成 2800
AlphaCode 代码生成 414
Open AI GPT3 语言理解与生成、推理等 1750
CLIP&DALL-E 图像生成、跨模态检索 120
Codex 代码生成 120
ChatGPT 语言理解与生成、推理等 -
英伟达 Megatron-Turing NLG 语言理解与生成、推理等 5300
Stability AI Stable Diffusion 语言理解与图像生成
头条 @未来智库


01 大模型参数量快速提升，算力需求大幅增加
长江证券
1、算力总需求量=参数量*词条长度（单个词语计算次数，单精度）*训练词数
2、GPU总需求量=算力总需求量/单颗加速卡算力/计算用时
按照175B的参数规模测算，若训练时长为1个月，则需求A100 GPU需要数超6000张，A100加速卡成本1.56亿美元；若参数量提升至481B，则加速卡成本上升至4.28亿美元。
主要大模型算力加速卡需求情况
头条 @未来智库

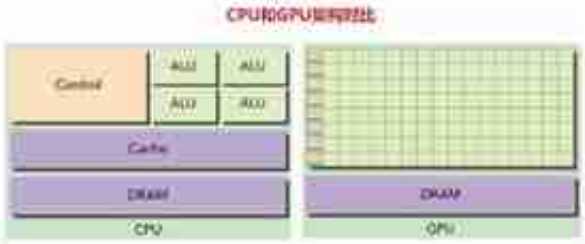

长江证券
02
AI+Chiplet:
信息革命的基石
头条@未来智库

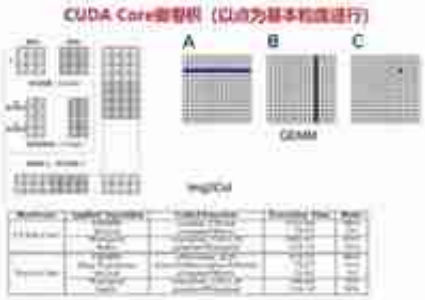

02 应用-软件-硬件循环向上，AI芯片发展多元变化
长江证券
头条@未来智库

02 GPU：并行运算带来对AI应用的高度适配

长江证券

- 在架构上GPU由数以千计的更小、更高效的核心（专为同时处理多任务而设计）组成。现在的CPU，一般是多核（multi-core）结构；而GPU一般是众核（many-core）结构。
- 为充分利用GPU的计算能力，NVIDIA在2006年推出了CUDA（Computer Unifie Device Architecture，统一计算设备架构）这一编程架构。该架构使GPU能够解决复杂的计算问题。它包含了CUDA指令集架构（ISA）以及GPU内部的并行计算引擎。英伟达在GPU里面增加了Tensor Core为AI服务，它的并行力量就从基本的数据点进化到以小矩阵快速进行计算。所以Tensor Core最基本的并行单元是一个4×4的矩阵块，即能在一个时钟周期里面算出一个4×4矩阵和另一个4×4矩阵相乘的结果。原来用数据点来并行的话，它需要16次这样的计算，才能算出一个4×4的矩阵，相比之下Tensor Core的算力比原来的GPU要高。等价的功耗等比原来GPU的要低。这就是Tensor Core用来做矩阵一个很棒的进化。





02 英伟达：三重壁垒构造AI时代软硬件一体化龙头

长江证券

- **第一层壁垒：硬件层。**GPU高效图形渲染和AI算力基础。英伟达硬件层的三芯战略已逐步成型：GPU解决AI大规模并行运算痛点，DPU解决AI训练推理中设备网络通信与CPU负荷问题，CPU填上三芯战略最后一块拼图，GPU强耦合设计构造完整AI解决方案，NVlink+NVSwitch+ConnectX突破芯片直连和设备网络连接瓶颈，GPUDirect Storage技术实现高性能存储和数据访问。
- **第二层壁垒：软件层。**CUDA释放GPU算力引航AI发展，DOCA、Omniverse等软件层进一步填充生态，增强AI行业对英伟达的粘性。CUDA从面世代码出发发挥GPU并行运算优势，奠定近十年人工智能发展基础。DOCA为BlueField DPU量身定制软件开发平台，复刻GPU+CUDA的强耦合成功路径。Omniverse初试工业共享虚拟空间，从硬件→软件→云上社区，在强悍软硬件基础上打造系统级AI生态圈。NVIDIA AI Enterprise加速AI模型开发，未来或有望助力实现以AI开发AI
- **第三层壁垒：应用层。**游戏显卡、数据中心、自动驾驶、元宇宙先发优力，十年成长曲线测量加。


02 沐曦：顶级团队布局全栈解决方案
长江证券
曦思N100
产品兼容性证明
头条@未来智库


02 芯原股份：GPU+NPU的IP卖水人
长江证券
头条@未来智库

02 国产FPGA替代先锋——紫光同创

长江证券

- 紫光国微控股子公司深圳市紫光同创电子有限公司，紫光同创是中国FPGA领导厂商，总投资超过20亿元，主要从事可编程逻辑器件（FPGA、CPLD等）的研发与销售，EDA设计工具的开发，产品覆盖通信、工业控制、视频监控、消费电子、数据中心等应用领域。近期推进高中低端全系列FPGA产品的研制开发工作。公司研发实力深厚，职工人数超过500人，研发人员占比为65%，拥有专利超300项，核心专利占比超过80%。紫光同创2020年已实现3.16亿元销售收入，同比大幅增长超210%，未来有望进一步提升公司FPGA产品及IP、软件等解决方案盈利能力，公司还可提供国产FPGA器件的系统解决方案、板卡外包设计服务以及IP模块解决方案，正在逐步从FPGA产品提供商向提供FPGA平台和系统解决方案的平台型企业发展。

紫光同创发展历程

紫光同创FPGA开发软件平台

02 Chiplet：芯片集成的新形态，规避限制的关键技术

长江证券

- 目前，**主流系统级单芯片（SoC）都是将多个负责不同类型计算任务的计算单元，通过光刻的形式制作到同一块基板上。**比如，目前搭载的智能手机的SoC芯片上，基本都集成了CPU、GPU、DSP、ISP、NPU、Modem等众多的不同功能的计算单元，以及诸多的接口IP，其中涉及的是高端的集成化。利用先进制程对于所有的单元进行全面的提升，缺点则是设计时间漫长、成本高昂。
- 而Chiplet技术是SoC集成发展到一定程度之后的一种新的芯片设计方式，它通过**将SoC分成较小的裸片（Die），然后每个单元选择最适合的工艺制程进行制造，再将这些模块化的小芯片（裸片）互联起来，采用新型封装技术，将不同功能不同工艺制造的小芯片封装在一起，成为一个异构集成芯片。**

基于Chiplet的异构架构应用处理器的示意图

Chiplet可灵活实现集成各类功能的芯片设计


02 下游需求加速扩围，产业链价值重构
长江证券
全球基于Chiplet的器件市场规模（台下游应用，百万美元）
2018 2019 2020 2021 2022 2023 2024
AMD
EPYC（Naples）
EPYC（Rome）
Ryzen（Matisse）
Intel
Altera Stratix 10 FPGA
Lakefield SoC
Sapphire Rapids
Agilex FPGA
Apple
M1 Max
Xilinx
Marvell
Intel
头条@未来智库


02 先进封装：Chiplet下封装技术价值凸显，国内OSAT强者恒强
长江证券
全球先进封装市场需求迅速爆发。由于AI芯片组需要运算速度更快的内核，更小巧的外形以及高能效，AI市场的不断扩张推动先进封装行业的增长。同时，5G技术普及也增加了先进封装需求，5G芯片组较依赖先进封装技术，来实现高性能、小尺寸和低功耗。据Yole，2021年，全球先进封装市场规模超300亿美元，预计2027年可达近600亿美元。
2021-2027E全球先进封装市场规模（百万美元）
2021年全球先进封装企业营收排名（Top15+企业，百万美元）
头条@未来智库


03 服务器：AI驱动的硬件军备竞赛
长江证券
全球AI服务器2020-2026E市场规模及预测（十亿美元）
全球AI服务器2020-2026E市场规模及预测（十亿美元）
头条@未来智库