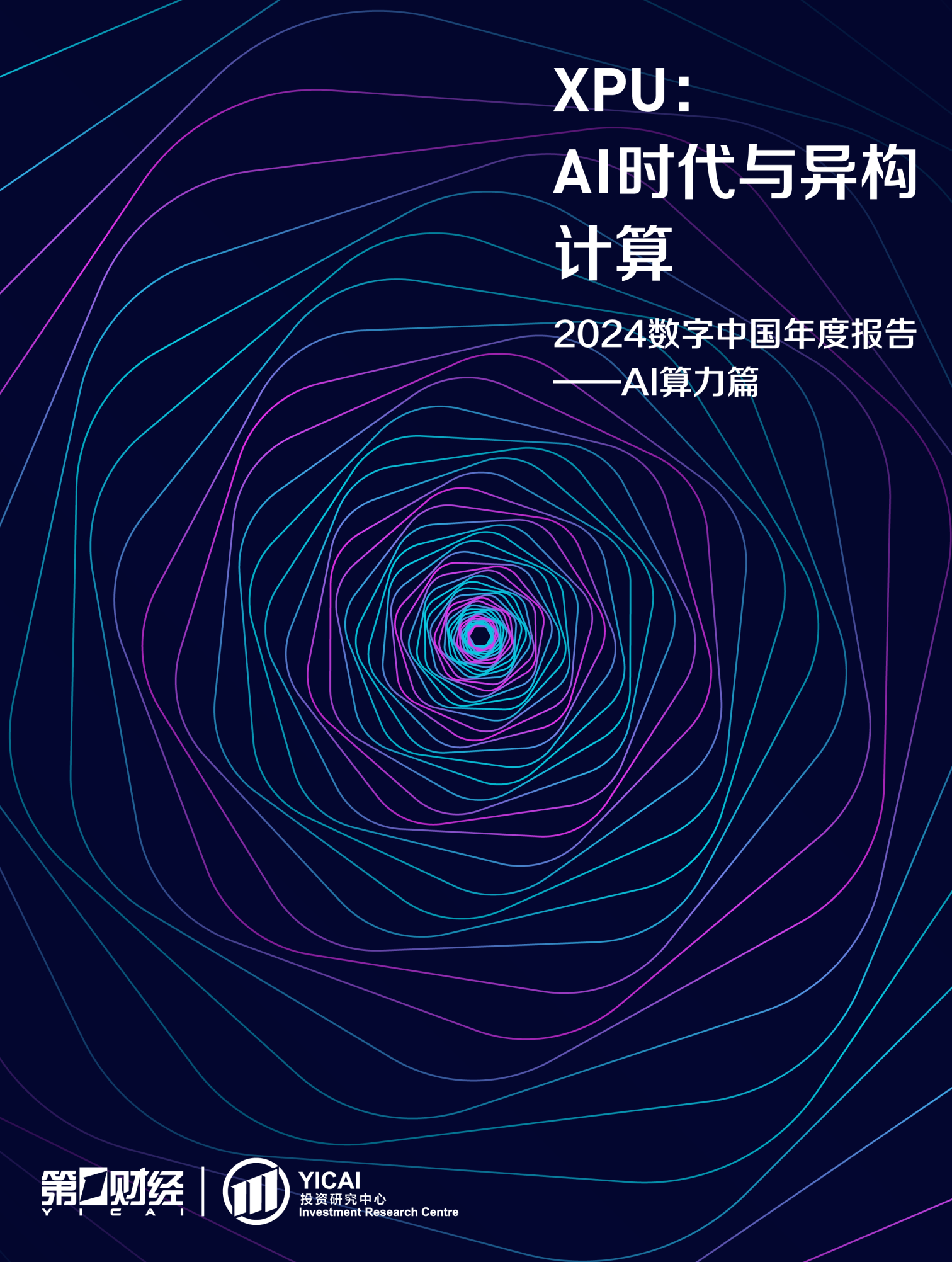
XPU：AI时代与异构计算

2024数字中国年度报告——AI算力篇



第一财经 YICAI

YICAI 投资研究中心 Investment Research Centre

目录 Contents

01
引言

02
CPU

2.1 AI 爆发，推动 CPU 需求提升 06
2.2 CPU 指令集架构之争 07
2.3 主要国产 CPU 厂商 09

03
GPU

3.1 英伟达与 CUDA 生态 11
3.2 国内 GPU 厂商快速发展 13
3.3 兼容与自建生态之路 14

04
FPGA

4.1 低时延与灵活性优势 16
4.2 AI 成增速最快下游应用市场 16
4.3 国产 FPGA 仍处于 0-1 阶段 18

05
ASIC

5.1 TPU：谷歌定制，能效比突出 21
5.2 DPU：CPU 和 GPU 之后的“第三颗主力芯片” 23
5.3 NPU：专为 AI 应用而生 27

06
AI 算力产业趋势展望

07
结语

01 引言

AI 大模型对算力的需求正以指数级速度增长，推动AI 算力平台从单一的单机计算向集群计算转变。

作为算力的主要载体，AI 芯片在AI 时代呈现出强劲的增长势头。

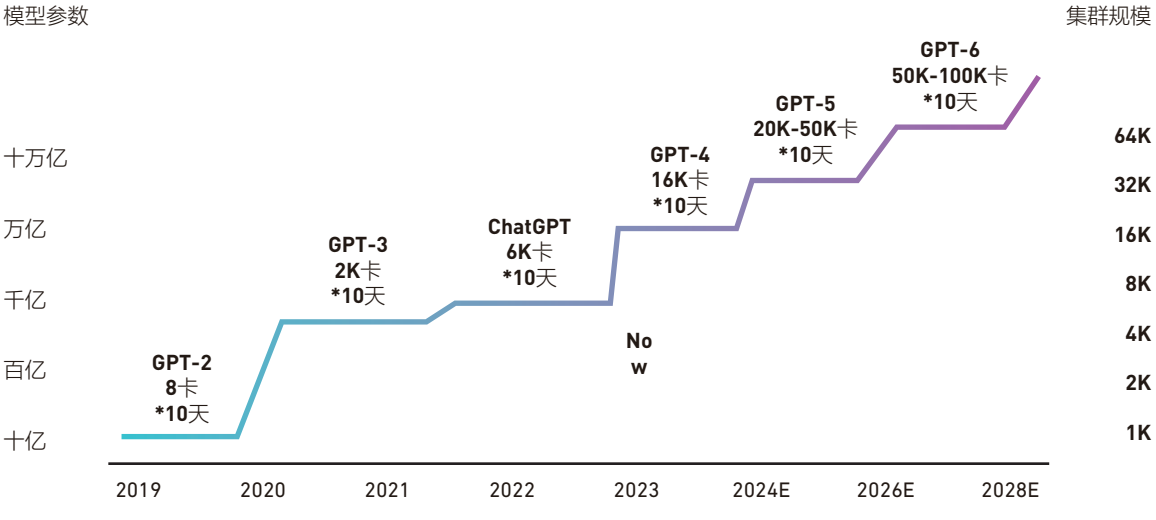
AI 芯片的定义为“专门针对AI 算法做了特殊加速设计的芯片”，按技术架构可以分为通用图形处理器（GPU）、中央处理器（CPU）、专用集成电路芯片（ASIC）以及现场可编程门阵列（FPGA）等，根据场景可以分为云端和端侧。

目前，以GPU 为代表的AI 计算芯片市场规模正快速增长。据Gartner，2023 年全球AI GPU 芯片市场规模约为534 亿美元，预计2024 年同比增速将达25.7%。

由于AI 芯片架构众多，异构计算成为AI 时代丰富算力需求下的必然选择。异构计算是通过在单一系统中利用不同类型的处理器，如CPU、GPU、ASIC、FPGA 等协同工作，执行特定任务，以优化性能和效率，更高效地利用不同类型的计算资源，满足不同的计算需求。

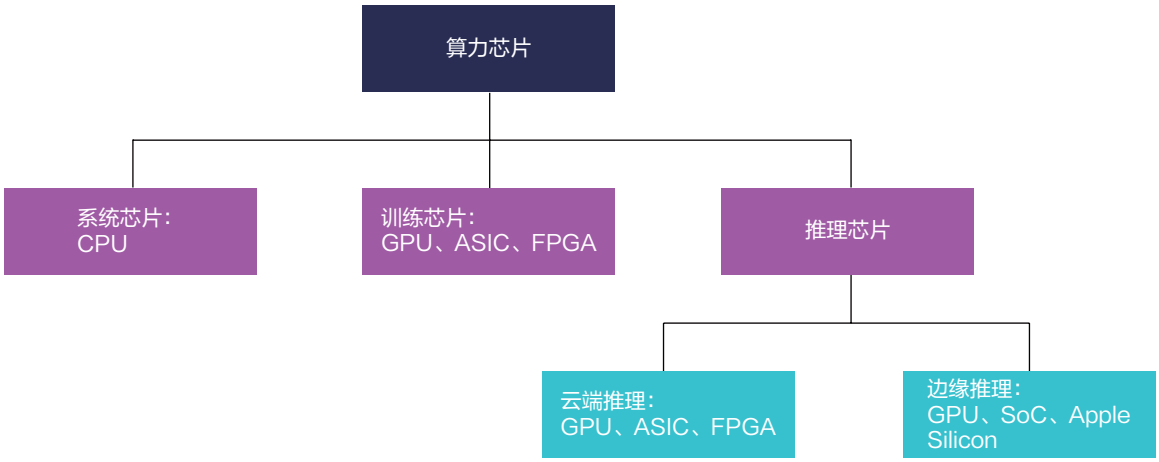
目前，异构计算主要分为三大类：CPU+GPU、CPU+FPGA、CPU+ASIC，其中CPU 可执行通用AI 计算，但其通用架构设计使运行效率受限，因此需要GPU 等协处理器进行大规模并行计算。GPU 是目前商用最广泛的AI 芯片，IDC 数据显示，在中国AI 芯片市场，GPU 占有超过80% 的市场份额。

2019-2028E AI 大模型算力平台参数



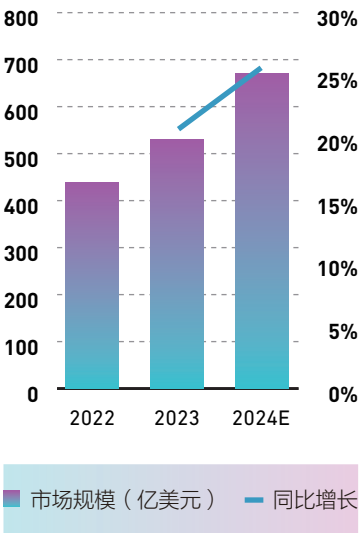
资料来源：华为，长江证券研究所，第一财经整理

算力芯片构成



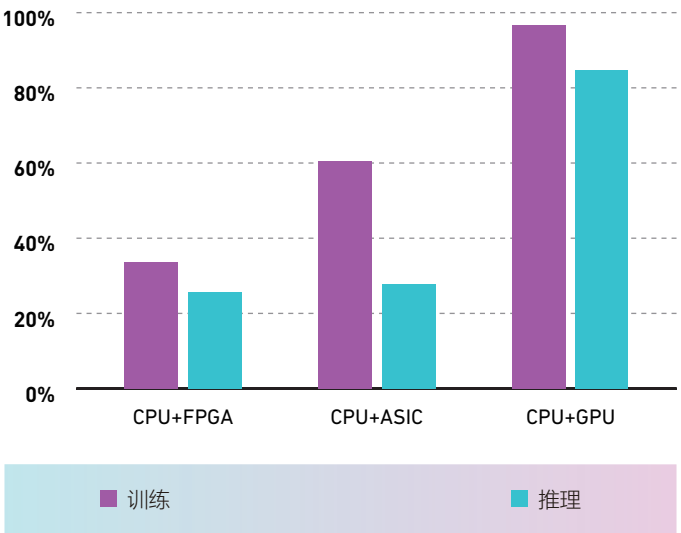
资料来源：华经产业研究院，中航证券研究所，第一财经整理

2022-2024E 全球 AI GPU 市场规模



资料来源：Gartner，华泰研究，第一财经整理

人工智能训练和推理工作负载选用的计算架构



资料来源：IDC，浪潮信息，中航证券研究所，第一财经整理

人工智能训练和推理工作负载选用的计算架构优缺点

	优点	缺点	实例
CPU+FPGA	FPGA灵活性高、无需取指令、译码，执行效率高	总体性价比和效率不占优势。FPGA的大规模开发难度偏高，从业人员相对较少，生态环境不如GPU	微软使用FPGA为Bing搜索智能化进行加速
CPU+ASIC	ASIC体积小、功耗低、计算性能高、计算效率高、芯片出货量越大成本越低	算法固定，一旦算法变化就无法使用。目前人工智能算法远没有到算法平稳期，ASIC专用芯片如何做到适应各种算法是个最大的问题	寒武纪的NPU、地平线的BPU、Google的TPU
CPU+GPU	GPU具有更好的生态环境，可以直接部署在服务器中，方便快速	缺乏灵活性	英伟达、AMD的GPU

资料来源：IDC，浪潮信息，中航证券研究所，第一财经整理

本报告聚焦构成AI算力底座的核心芯片，将从CPU、GPU、FPGA和TPU、NPU、DPU等AISC芯片分别展开，梳理国内AI芯片产业的发展现状并探寻各细分领域投资机会。

02
CPU

CPU，即中央处理器，常由控制单元、算术逻辑单元（ALU）和寄存器组成，负责指令读取、译码与执行，对研发技术和生态构建具有很高要求，对计算机的性能和运行效率具有重要影响。

作为计算机的运算和控制中心，在AI 发展浪潮下，我国计算机CPU 行业发展进入“快车道”：智算中心的建设拉动了服务器需求，AIPC 的更新换代也在推动需求增长。

2.1 AI 爆发，推动 CPU 需求提升

CPU 作为服务器的核心部件，其性能和功耗直接影响着服务器的整体表现。AI 大模型快速发展背景下下，智能算力需求爆发，AI 服务器出货量的增长将带动云端CPU 需求。

根据TrendForce 数据，2023 年全球AI 服务器（包含搭载GPU、FPGA、ASIC 等）出货量近120 万台，年增38.4%，占整体服务器出货量近9%，预计至2026 年将占15%，2022~2026 年全球AI 服务器（包含搭载AI 训练、推论用的GPU、FPGA、ASIC 等加速芯片）出货量年复合增长率预计达29%。

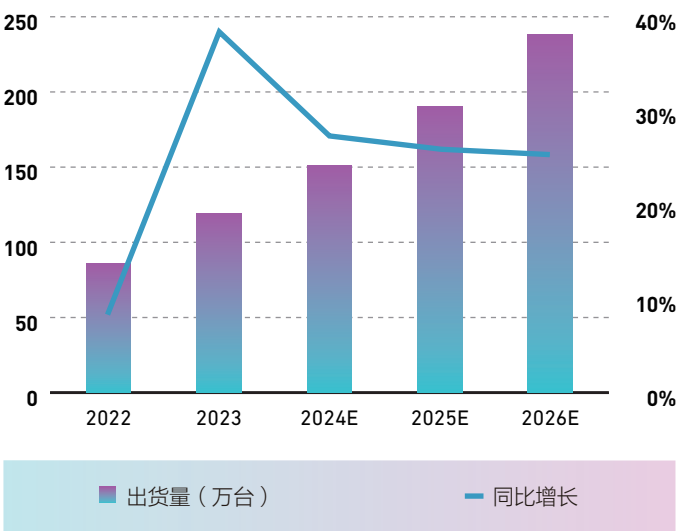
在PC 市场，2024 年是公认的AIPC 元年，AIPC 有望带动新一轮换机潮。业内认为，当前AIPC 的发展，类似早期Windows 系统的发展，未来有望复制Windows 系统的成功。

Canalys 预计，2024 年全球PC 出货量为2.67 亿台，同比增长7.6%，其中AIPC 出货量在2000 万台左右；2027 年，将有60% 的电脑具备AI 处理能力，AIPC 出货量将超过1.75 亿台。

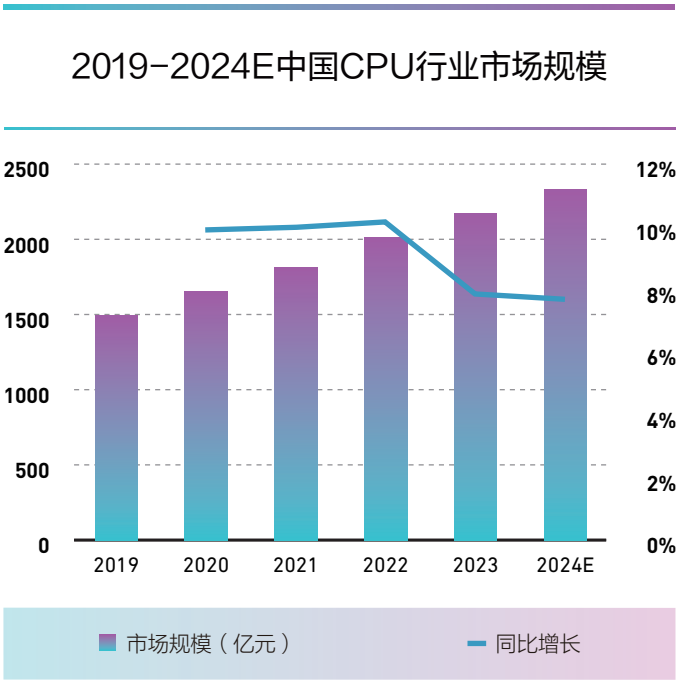
根据中商产业研究院数据，2024 年中国CPU 市场规模约为2326 亿元。而大模型掀起的AI 浪潮，不仅拉动了CPU 的需求增长，也对CPU 的功能提出了更高的要求。

“很多AI 应用其实可以用不同的XPU 处理单元进行计算，CPU 的编程更灵活，算子加速可以通过通用的编译器处理，相比NPU 更通用、更灵活。”此芯科技生态战略总经理周杰此前告诉第一财经，

2022-2026E全球AI服务器出货量



资料来源：TrendForce，民生证券研究院，第一财经整理



资料来源：中商产业研究院，中泰证券研究所，第一财经整理

此芯科技的Armv9 CPU包含向量扩展指令加速，能够更好地帮助在端侧运行大语言模型，最新的iPhone 16系列也对CPU进行了升级，包含了矩阵运算指令加速，以更好地支持那些基于CPU的Apple Intelligence应用。”

2.2 CPU指令集架构之争

芯片架构是设计和构建微处理器或其他集成电路的基础，定义了芯片的组织结构、功能单元、数据流以及指令集等关键特性。指令集架构则是芯片架构中最顶层的设计，定义了计算机硬件能够执行的所有指令的集合，是计算机系统设计中的核心部分，对上层软件和下层硬件都具有重要影响。

根据设计原则的不同，指令集架构主要分为两大阵营：复杂指令集计算（CISC）和精简指令集计算（RISC），并由此衍生出x86、ARM、RISC-V等指令集架构。

X86由Intel推出，属于CISC设计，拥有丰富的指令集，支持多种寻址模式和复杂的操作。目前搭载x86架构指令集的处理器主要应用在PC和服务器上。由于Intel严格把控专利授权，因此x86市场基本被Intel垄断。

在服务器市场，x86是CPU的主流架构。根据IDC数据，2023年x86架构服务器市场份额为88%，中国x86服务器市场出货量为362万台，预计2024年还将增长5.7%。

Arm架构基于RISC设计，Arm内核微架构IP选择多样、设计精简可靠、在低功耗领域表现优异。Arm的商业模式是专注处理器内核架构的授权，而不直接设计、制造芯片，中立的地位使得Arm通过授权广泛推广了生态系统。据悉，目前Arm架构在以移动终端芯片（手机平板等）、机顶盒、视频监控等为代表的移动智能领域获得广泛应用。

整体看，目前云端服务器多采用X86架构，边缘侧采用Arm架构，但近年来格局也在发生改变，Arm架构逐渐向服务器领域渗透。“现在全球很多互联网厂商都开始用Arm架构来设计自己的服务器芯片，比如亚马逊、谷歌、微软等，以及国内的阿里。”周杰对第一财经表示。此外，AIPC搭载NPU，Arm架构或更具优势。Arm CEO Rene Haas在COMPUTEX 2024上表示，Arm预计将五年内拿下Windows PC市场50%以上的份额。

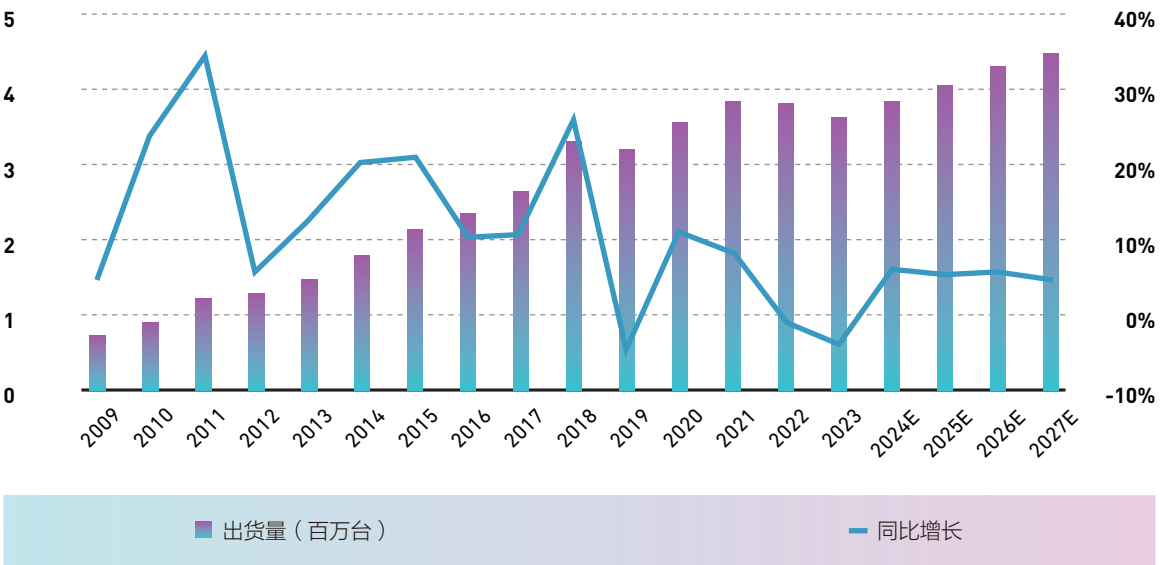
开放精简指令集架构RISC-V近年来也开始兴起。RISC-V完全开源，采用宽松的BSD协议，企业可以完全自由免费使用，同时也允许企业添加自有指令集，而不必开放共享，以支持差异化发展；同时架构简单，短小精悍，不同

复杂指令集CISC与精简指令集RISC的对比

比较内容	CISC	RISC
特点	指令多，一条指令执行多个功能	指令少，复杂任务由多个精简指令组合完成
优点	特定功能执行效率高，例如多媒体处理	常用工作执行效率高，功耗低
缺点	系统设计复杂，执行效率低	部分复杂任务处理效率偏低，例如多媒体处理
典型架构	X86	ARM、Power、Alpha、MIPS、LoongArch

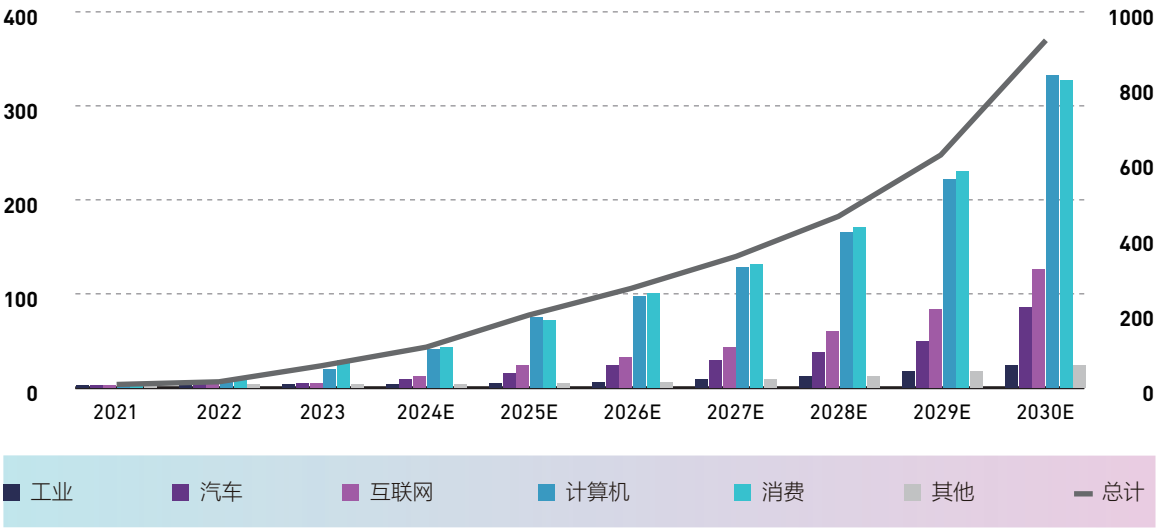
资料来源：海光信息招股说明书，华创证券，第一财经整理

2009-2027E中国x86服务器出货量



资料来源：IDC，海光信息2023年年报，第一财经整理

2021-2030E基于RISC-V的SoC营收情况（亿美元）



资料来源：The SHD Group，第一财经整理

的部分还能以模块化的方式组成在一起，从而通过一套统一架构满足各种不同应用场景。据了解，目前国内多家初创企业开始用RISC-V进行芯片设计。SHD Group预计，2030年基于RISC-V的SoC营收有望高达920亿美元，2021-2030年复合年增长率高达47%。

“RISC-V架构目前更多应用在一些固定场景，比如IoT设备等，但它对于通用生态的支持，比如PC、服务器领域，需要更多的资源投入和研发支持。”周杰告诉第一财经。

此外，也有部分企业自研指令集。龙芯中科（688047.SH）2023年在MIPS架构的基础上，宣布独立研发出具有完全自主产权的Loong Arch架构，申威基于Alpha架构研发出SW64架构。

“生态是一个很重要的问题，X86的生态无疑是最成熟的，Arm架构生态也越来越成熟。”周杰表示，“自研架构的生态相比X86和Arm架构生态，会有一定的差距。如果需要支持更通用的生态，整个产业链上下游就需要很多投入，比如各个软件厂商需要去适配自研架构，并进行相关软件优化。”

2.3 主要国产CPU厂商

市场格局来看，目前Intel、AMD领跑通用CPU市场。

根据CPU市场追踪机构Mercury Research的报告，2024年第二季度，Intel在全球PC市场占据78.9%的市场份额，AMD市占率为占据21.1%；在服务器CPU市场，Intel市占率高达75.9%，AMD为24.1%。

国产CPU厂商分为三大流派：一是自研架构，如前文所述的龙芯和申威；二是使用Arm架构，如中国长城（000066.SZ）子公司飞腾、华为海思；三是使用X86架构，上海兆芯和海光信息（688041.SH）均使用X86架构。

与GPU相比，国内CPU的发展则更为成熟。根据头

主要国产CPU厂商及产品

芯片类别	厂商	产品系列	详情
CPU	海光信息	3000系列、5000系	国产先进微处理器产业推动者，海光CPU兼容市场主流的x86指令集
	龙芯	龙芯一号、龙芯二号、龙芯三号	国内自主CPU引领者，自主生态构建者，具有自主指令系统龙架构，成为与X86/ARM并列的顶层开源生态系统
	华为	鲲鹏920	全球领先的Fabless半导体与器件设计公司，鲲鹏920采用7nm制造工艺，基于ARM架构授权
	飞腾	腾云系列、腾锐系列、腾珑系列	自主核心芯片提供商，致力于飞腾系列国产高性能、低功耗通用计算微处理器的设计研发和产业化推广
	兆芯	开先系列、开胜系列	掌握自主通用处理器及其系统平台芯片研发设计的核心技术，全面覆盖其微架构与实现技术
	中威	申威221、申威411、中威421、申威421M、中威1621	提供申威处理器芯片内核、封装设计、技术支持服务及销售，小型计算机研发、测试、销售、服务及核心部件生产

资料来源：Wind，各公司官网，华创证券，第一财经整理

豹研究院数据，2022年中国CPU国产化率为30%-40%。同时，中国CPU国产化率还在不断提升，在最近的运营商集采中，国产CPU比例接近70%。

随着信创的持续推进，国产CPU或将获得较大的增长空间。Canalys数据显示，2023年中国政府及教育部门的PC采购量达到272万台，占到全国出货量的6%。2024年3月，中央政府采购网发布公告称，在采购计算机时应当将CPU、操作系统符合安全可靠测评要求纳入采购需求。Canalys认为，上述六家都属于符合该公告中国产CPU条件的品牌。

“产业正在努力推进先进制造工艺和上游EDA工具发展，这些都对芯片产品的竞争力产生很大影响。”周杰称，“结合国内目前的产业现状，设计出一款芯片问题不大，关键是设计出来之后，一是怎么生产，二是如何建立一个好的生态，帮助开发者和终端用户比较好地用起来，这一点尤为重要，也是公司核心能力的体现。”

03 GPU

GPU，即图形处理单元，是计算机显卡的核心。

与CPU 相比，GPU 的逻辑运算单元较少，单个运算单元（ALU）处理能力更弱，但能够实现多个ALU 并行计算。同样运行3000 次的简单运算，CPU 由于串行计算，需要3000 个时钟周期，而配有3000 个ALU 的GPU 运行只需要1 个时钟周期。

不过，GPU 处理并行计算并不是作为一个独立的计算平台，而是与CPU 通过PCIe 总线连接在一起来协同工作，可视为CPU 的协处理器。

作为计算机的图形处理以及并行计算内核，GPU 最基本的功能是图形显示和分担CPU 的计算量，主要可以分为图形图像渲染计算GPU 和运算协作处理器GPGPU（通用计算图形处理器），后者去掉或减弱GPU 的图形显示能力，将其余部分全部投入通用计算，实现处理人工智能、专业计算等加速应用。本报告主要讨论的也是后者。

由于GPU 拥有数千个ALU，能够并行执行数百万个数学运算，因此GPU 与深度学习技术完美契合，使用GPU 做辅助计算，能够更快地提高AI 的性能。CPU+GPU 成为了目前应用最广泛的算力底座。

根据中商产业研究院数据，2023 年中国GPU 市场规模为807 亿元，同比增长32.78%，预计2024 年将增至1073 亿元，2020-2024 年复合增长率（CAGR）达32.8%。而据智研咨询，全球人工智能GPU 市场2020-2024 年CAGR 也将达到30.73%。

3.1 英伟达与 CUDA 生态

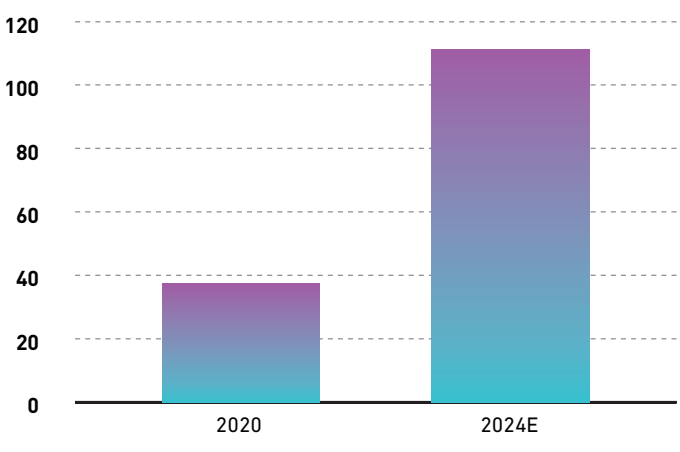
“GPU 的核心竞争力在于架构等因素决定的性能先进性和计算生态壁垒。”华安嘉业相关负责人此前告诉第一财经。

一方面，性能先进性体现在高精度浮点计算能力。训练需要密集的计算得到模型，没有训练，就不可能会有推理。而训练需要更高的精度，一般来说需要float 型，如FP32，32 位的浮点型来处理数据。

另一方面，生态也是GPU 发展需要解决的问题。

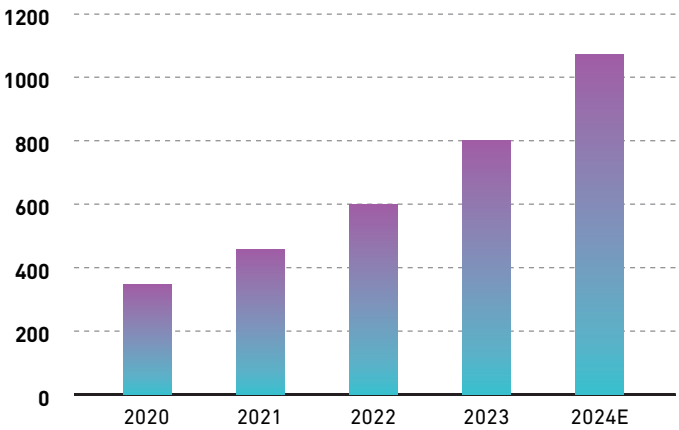
根据调研机构TechInsights 数据，2023 年全球数据中心GPU 总出货量达到了385 万颗，相比2022

2020及2024E全球人工智能GPU市场规模（亿美元）



资料来源：智研咨询，华创证券，第一财经整理

2020-2024E中国GPU市场规模（亿元）



资料来源：中商产业研究院，华创证券，第一财经整理

年的267 万颗增长了44.2%。其中，英伟达以98% 的市场份额稳居第一，出货量达376 万块，销售额同比增长超42%。

众所周知，英伟达凭借先发优势，以及大幅降低开发门槛的CUDA 架构，稳稳圈住了大批用户，不仅使GPU 在通用计算中逐渐成为主角，也成就了自身的护城河。

2007 年，英伟达首次推出通用并行计算架构CUDA(Compute Unified Device Architecture，统一计算设备架构)，使GPU 成为通用并行数据处理加速器，即GPGPU。CUDA 支持Windows、Linux、MacOS 三种主流操作系统，支持CUDA C 语言和OpenCL 及CUDA Fortran 语言。

CUDA 架构不用再像过去GPU 架构那样将通用计算映射到图形API（应用程序编程接口）中，大大降低了CUDA 的开发门槛。因此，CUDA 推出后发展迅速，广泛应用于石油勘测、天文计算、流体力学模拟、分子动力学仿真、生物计算、图像处理、音视频编解码等领域。

这为英伟达拿下GPU 过半市场份额奠定了基础。此后，英伟达通用计算架构持续升级迭代，2010 年发布Fermi 架

CUDA核心优势

优势	描述
高性能计算	CUDA能够显著提升计算性能，特别适用于需要大量数值计算和科学计算的任务。
易用性	CUDA提供了类似于C语言的编程接口，使得开发者能够更容易地上手并进行高效的GPU编程。
广泛的应用支持	CUDA已应用于多个NVIDIA的GPU系列，并在多个领域得到了广泛的应用。
完整的工具链	CUDA提供了包括性能分析工具、调试器以及样例代码和教程在内的完整工具链，为开发者提供了全面的支持环境。

资料来源：CSDN，长江证券研究所，第一财经整理

构，2012 年发布Kepler 架构，GPU 在通用计算中逐渐成为主角。

AI 芯片应用生态非长期研发和迭代更新不能解决。英伟达早在CUDA 问世之初就开始生态建设，AMD 和Intel 也推出了自研生态ROCm 和one API，但CUDA 凭借先发优势早已站稳脚跟。为解决应用问题，AMD 和Intel 通过工具将CUDA 代码转换成自己的编程模型，从而实现针对 CUDA 环境的代码编译。

但中信证券表示，由于CUDA 的闭源特性，以及快速的更新，后来者很难通过指令翻译等方式完美兼容，即使部分兼容也会有较大的性能损失，导致在性价比上持续落后英伟达。同时，CUDA 毕竟是英伟达的专属软件栈，包含了许多英伟达GPU 硬件的专有特性，这部分在其他厂商的芯片上并不能得到体现。

这正是国内厂商面临的困境。2024 年初，英伟达宣布禁止在其他GPU 上通过转译层运行CUDA 软件的消息引起业内广泛讨论。“在工具链层面兼容CUDA 的GPU 厂商会受到影响，但影响本身在技术层面还是比较复杂的。英伟达其实发出了一个非常强烈的信号，就是他正在扎紧自己生态的篱笆。”某GPU 业内人士对第一财经表示。

3.2 国内 GPU 厂商加速发展

主要国产GPU厂商及产品

芯片类别	厂商	产品系列	详情
GPU	华为	腾系列	提供AI推理、视频分析等功能，支持检索聚类、OCR识别、语音分析、视频分析等场景。
	海光信息	DCU系列	以 GPGPU 架构为基础，兼容通用的“类 CUDA”环境，可广泛应用于大数据处理、人工智能、商业计算等应用领域。
	寒武纪	思元370、思元290系列	寒武纪思元290芯片，采用创新性的MLUv02扩展架构，思元370是寒武纪首款采用chiplet（芯粒）技术的AI芯片，都基于7nm工艺。
	壁仞科技	壁砺系列	能够为数据中心广泛应用的PCIe形态GPU服务器提供灵活部署的强大的通用算力。
	摩尔线程	MTT S 系列	基于MUSA架构和自研MTLink技术，MTT S4000可以支持多卡互联，支持千卡集群基础设施建设，加速千亿参数大语言模型计算。
	燧原科技	燧原S60、云燧系列等	燧原S60是面向数据中心大规模部署的新一代人工智能推理加速卡，满足大语言模型、搜广推及传统模型的需求。
	天数智芯	天垓、智铠系列	天垓系列产品采用通用GPU架构，兼容国际主流GPU通用计算模型，支持国内外主流AI生态和深度学习框架及原生算子；智铠系列采用通用GPU架构及通用指令集，可以灵活支持各种算法模型，便于客户自定义开发。
	沐曦集成电路	曦思N系列、曦云C系列	曦思N系列是面向云端应用的智算推理产品，采用高带宽内存，提供强大的算力和领先的视频编解码能力；曦云C系列通用GPU(GPGPU)芯片是针对智算及通用计算的完美解决方案，沐曦自主知识产权架构提供强大高精度及多精度混合算力，可广泛应用于智算以及通用计算、教育和科研等场景。

资料来源：公司官网，第一财经整理

当前国内GPU 厂商纷纷大力投入研发迭代架构，谋求构建自主软硬件生态。

华为可谓国内GPU 产业发展的引领者，近年来华为在昇腾AI 系列芯片迭代、自主可控产能提升方面取得了长足进展。根据《2023 智能算力发展白皮书》，假设到2025 年我国智能算力规模至少要达到900EFlops，预计昇腾市占率将达到30%，昇腾910B 算力相当于A100（FP16 约300Tflops），平均价格为12 万元，开源证券保守估计昇腾910B 累计市场空间为1080 亿元。

海光信息（688041.SH）的DCU 也属于GPU 的一种，其DCU 协处理器全面兼容ROCm GPU 计算生态。据悉，ROCm 和CUDA 在生态、编程环境等方面高度相似，CUDA 用户可以以较低代价快速迁移至ROCm 平台，因此ROCm 也被称为“类CUDA”，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。

根据官网描述，寒武纪（688256.SH）首款云端训练芯片思元290 实现了INT8 算力512TOPS，而在研的第五代智能处理器微架构对推荐系统和大语言模型训练推理场景进行了重点优化，在编程灵活性、能效、功耗、面积等方面大幅提升了产品竞争力。

半导体初创企业中，燧原科技、芯瞳半导体、芯动科技、摩尔线程、天数智芯、壁仞科技等均已陆续推出产品。据悉，2020 年开始，国内GPU 行业融资环境有较大改善，初创公司遍地开花。

摩尔线程成立于2020 年，是一家以全功能GPU 芯片设计为主的集成电路企业。该公司推出了全面对标CUDA 的MUSA 架构，用户能够将CUDA 写的应用程序通过摩尔线程的编译器重新编译成MUSA 的应用，从而实现接近零成本迁移，同时也能够通过标准编程语言开发新的应用。“MUSA 本身是一套独立自主的生态，同时也是个开放的、可以吸收现有生态的全新生态。”摩尔线程CTO 张钰勃告诉第一财经。

燧原科技专注于人工智能领域云端和边缘算力产品，致力于为通用人工智能打造算力底座，提供原始创新、具备自主知识产权的AI 加速卡、系统集群和软硬件解决方案。

3.3 兼容与自建生态之路

国产算力芯片如果从“能用”变成“好用”，仍需投入大量研发成本和时间。华安嘉业上述负责人对第一财经表示，国产GPU 在起步阶段兼容现有生态更容易发展，但长期还是要摆脱兼容思路，发展自有核心技术。

“我们经常讲兼容，但兼容不代表要和英伟达做得完全一样，而是说你做的东西可以去承载所有技术的生态，可以把英伟达的生态吸收过来、直接利用。但要做功能全面对标英伟达的GPU 芯片难度很大，目前大多数厂商采取的策略是仅实现英伟达GPU 人工智能加速的部分功能。”张钰勃表示。

“独立自主和开放兼容并不矛盾。我们一方面可以独立自主发展实现可控，一方面也可以开放兼容现有CUDA 优势。”张钰勃告诉第一财经，“只有硬件功能完全对标，才能够有效地把CUDA 生态的应用吸收过来。如果没有办法吸收现有生态，另建一个新生态，真要建成也是十几二十年的事。”

客户迁移成本是推动国产GPU 厂商加速生态建设的重要因素之一。目前，国内也存在一些坚持“难而正确”理念的厂商，选择了自建生态、不兼容的道路，燧原科技就是其一。

对于这类算力厂商来说，客户迁移成本始终存在，因此需要寻找志同道合的客户。“燧原希望与产业合作伙伴一起构建一个开放开源的生态系统，我们的客户也愿意跟那些有长期主义的合作伙伴一起去打磨产品。”燧原科技首席生态官李星宇此前告诉第一财经。

据悉，随着技术的发展，国内厂商的自建生态之路有望越走越宽。

“技术生态的范式转移，给像燧原这样的初创公司自建生态带来一个新的契机。”李星宇认为，随着大模型时代的到来，模型的架构底座趋向于一

致，即Transformer，这收敛了对于硬件的需求，让硬件设计的方向更加聚焦和明确，减轻了碎片化程度；与此同时，越来越流行的开源框架和编程语言，让芯片公司有更好的基础去适配不同模型，让开发者更容易在开发工具层面去适配不同的硬件。

“客户的迁移成本取决于很多因素，但整体的趋势是越来越便捷。”李星宇表示，“比如说我们兼容PyTorch 的主流算子，采用这些主流算子的模型理论上可以直接迁移而不需要改源代码。同时未来我们也会支持更多主流的开源编程语言，让客户开发新的模型时，也会变得更加容易。”

虽然目前国内有多家AI 芯片厂商选择自建生态，但并未形成统一生态，各家正处于跑马圈地、各自发展的时期。诚然，在技术发展的早期和技术快速迭代的时期，很难制定一套统一的标准。正如海外GPU发展的早期，行业存在四十多家企业，但大浪淘沙后，仅留存几家企业做大做强。在快速变化的技术趋势面前，每个人都有自己不同的理解，让市场去选择，让客户去选择，可能是一个更好的方式。

“技术的提升最终是靠市场和客户的需求牵引，中国真正的优势在于拥有全球最大的市场，以及众多开发者愿意去拥抱新技术。”李星宇表示。

04 FPGA

FPGA，即现场可编程门阵列，是在硅片上预先设计实现的具有可编程特性的集成电路，用户在使用过程中可以通过软件重新配置芯片内部的资源实现不同功能，因此具有优越的灵活性，能够满足不同场景的应用需求。

4.1 低时延与灵活性优势

相比CPU，FPGA 具备两大性能优势，一是优越的灵活性，二是低时延处理。

FPGA 芯片类似于集成电路中的积木，用户可根据各自的需求和想法，将其拼搭成不同的功能、特性的电路结构，以满足不同场景的应用需求。GPU 在设计完成后无法改动硬件资源，而 FPGA 根据特定应用对硬件进行编程，更具灵活性。机器学习使用多条指令平行处理单一数据，FPGA 的定制化能力更能满足精确度较低、分散、非常规深度神经网络计算需求。

时延方面，CPU 为了保证最大程度的通用性和复杂任务的处理，引入了指令集和对应的取指译码操作，而FPGA在设计时就相当于预先指定了指令，无需像CPU 一样进行Fetch（取指）-Decode（译码），可以直接进入相当于CPU 的Excecute（执行）的环节。同时，FPGA 采用高并行架构，数十万个 CLB 可以同步执行。现代CPU 虽然有多个ALU（核心计算单元）以完成并行计算，但在并行度上仍然不如有数十万个CLB 的FPGA。

此外，对于现代CPU 为了提升并行度增加的模块，比如BranchPrediction（分支预测），Out-of-orderExecution（乱序执行），调度（Scheduler），FPGA 都不需要，因此FPGA 完成指令所花费的时钟周期要远小于CPU。

在高并行计算中，FPGA 的时延优势更突出。比如，在完成雷达波束赋形这一高并行算法时，使用XilinxVirtex7（FPGA）时延仅需3.3ms，而且时钟频率仅需125MHz，系统功耗仅为75W，而使用ARMA9（CPU）在667MHz 的时钟频率下依然需要250ms 才能完成，而且系统功耗高至1400W。

4.2 AI 成增速最快下游应用市场

低延时与灵活性优势造就了FPGA 广阔的下游市场。东兴证券研报称，2022 年FPGA 全球市场空间超80 亿美元，其中大部分需求来自于电信、工业、数据中心&AI、国防& 航空航天四大领域，2028 年将增长至接近200 亿美元，2022 年~2028 年CAGR 超15%，由中国市场引领增长。

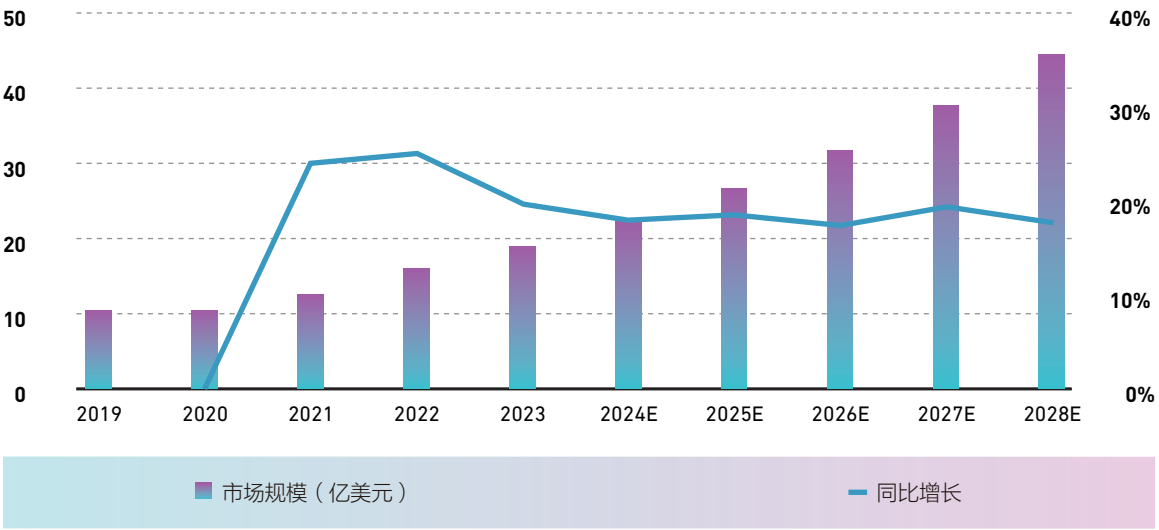
据Marketsandmarkets 数据，2022 年中国FPGA 市场规模约为16 亿美元，随着AI& 数据中心、电信、国防& 航空航天、汽车市场对FPGA 的需求日益增长，预计2028 年规模约为45 亿美元，2022 年~2028 年复合增速18%，高于全球其他地区。

FPGA 在数据中心&AI 领域是重要的计算芯片类别，可以作为异构计算的关键一环，起到加速计算的作用。根据东兴证券的测算，2028 年FPGA 在全球AI 领域应用的市场规模将达到35 亿美元，占比逾15%，2022 年~2028 年CAGR 为18%，是增速最快的细分领域。

具体来看，FPGA 在数据中心&AI 的增长驱动力主要来自于低时延推理的需求。

FPGA 的高并行、低时延、低功耗的特性特别适合需要实时推理的场景。例如，

2019-2028E中国FPGA市场规模



资料来源：Marketsandmarkets，东兴证券研究所，第一财经整理

YOLO（Youonlylookonce）是目前最重要的实时目标检测算法，使用赛灵思（Xilinx）的ZynqUltraScale+MPSoC（某个FPGA方案），可以在约18ms的时延完成YOLOv5（第五代YOLO算法），而使用Zynq7100（另一种FPGA方案）则可实现压缩后的YOLOv7模型（第七代YOLO算法），在30帧率下实现小于33ms(15ms)的低时延推理。

此外，FPGA在数据中心常用于计算加速，比如云计算的加速实例、金融的高频/低时延的交易系统处理。FPGA还广泛应用于数据中心的互联、存储控制系统。

FPGA各大下游全球市场规模、主要应用场景和未来增长驱动力

	市场规模（亿美元）				CAGR	为什么需要FPGA	增长驱动力
	2019A	2022A	2025E	2028E			
电信	14 占比：25%	21	35	53	17%	低时延的处理能力（空口、基带等的信号处理） 高度灵活的可编程性（通信协议更新） 高并行&低功耗的方案（空口、回传）	通信技术迭代带来的速率需求提升→高速处理需求增长→FPGA价值量提升 通信网络安全需求提升→高速处理需求增长→FPGA渗透率提升

FPGA各大下游全球市场规模、主要应用场景和未来增长驱动力

	市场规模（亿美元）				CAGR	为什么需要FPGA	增长驱动力
	2019A	2022A	2025E	2028E			
汽车	69 占比：11%		14	21	14%	低时延的处理能力（ADAS、控制） 高度灵活的可编程性（EV电力系统）	ADAS等汽车智能化需求增长→低时延处理的需求的增长→FPGA渗透率&价值量提升
工业	10 占比：16%	13	20	28	13%	高并行低时延处理（工业视觉、伺服控制、仪器仪表） 高度灵活和低成本的方案（协议桥接）	工业领域高精度控制需求提升→低时延信号处理的需求增长→FPGA渗透率提升
医疗	45 占比：5%		69		11%	高并行低时延处理（医疗影像、医疗机器人） 高灵活、高并行、低功耗的方案（便携式医疗设备）	医疗诊断和外科手术复杂化→更高清画质、更低时延的图像处理→FPGA价值量提升
消费电子	45 占比：6%		71	0	12%	高并行低时延（商业显示LED信号控制） 高度灵活、低成本和低功耗的方案（小批量消费电子产品、DVD等）	高清显示&接口标准的复杂化→10数量要求增长&桥接需求增长→FPGA渗透率提升
国防&航空航天	71 占比：13%	11	62	4	15%	高并行低时延处理（制导、发动机火工、目标识别、通信） 动态可重构性带来的高可靠性（航天电子系统）	精确武器、无人机、SAR等高速雷达愈发广泛的使用、军用网络的复杂化→信号链的性能要求提升→高并行、低时延处理需求→FPGA渗透率&价值量提升
多媒体&广播	57 占比：8%		91	2	11%	高并行低时延处理（广播流媒体实时处理）	4k/8k等高清晰度在视频、体育、新闻直播→实时、低时延处理数据需求提升→FPGA渗透率&价值量提升
数据中心&AI	81 占比：15%	32	23	5	18%	高并行低时延处理（加速计算、存储控制、边缘A1） 高度灵活的方案（板上功能扩展&桥接）	数据中心速率升级→高带宽处理需求增长→FPGA价值量提升 边缘AI场景拓展→低时延推理的需求增长→FPGA渗透率&价值量提升
全球	57	83	130	192	15%	数据中心&AI、电信、国防&航空航天、汽车市场引领增长	

注：CAGR为2022A-2028E的6年复合增速预测，渗透率提升即发生FPGA替代竞品方案的情况（竞品包括MCU、GPU等），价值量提升即对更高端FPGA的需求增长。

资料来源：Xilinx，Altera，Lattice，Marketsandmarket，东兴证券研究所，第一财经整理

4.3 国产FPGA仍处于0-1阶段

目前，全球FPGA市场按制程可大致分为三类：90nm以上，主要用于高可靠的航天应用；20-90 nm，主要用于国防、航空航天、汽车、消费电子等领域；≤16nm以下。其中，14/16nm FPGA主要应用在电信领域的基带单元和有线网络，汽车领域的激光雷达，工业中的安防和仪器仪表等，7nm FPGA则主要应用在数据中心加速计算。

制程竞争是贯穿FPGA发展历史的竞争主线，制程领先后则具备先发优势，占据更多市场份额和盈利空间，是FPGA市场份额最直接的决定

全球FPGA制程及典型应用领域

制程	典型应用领域
0.13/0.15μm	A&D：航天应用
28/45/65/90nm	A&D：制导、目标识别、姿态控制、通信、航天
28/45/65nm	汽车：ADAS、车身
28/45nm	消费：显示、桥接
	工业：工业控制、激光加工、机器视觉、通信
28nm	电信：空口侧
14/16nm	电信：基带侧、有线网络
	汽车：激光雷达
	工业：安防、仪器仪表
	ASIC验证、医疗影像、多媒体&广播、A&D
7/10nm	数据中心&AI、计算加速

资料来源：第一财经整理

因素。因此FPGA 市场高度集中，龙头赛灵思（Xilinx）占据过半份额，前四名玩家合计份额超90%。在数据中心和AI加速计算领域，赛灵思占据绝对性份额。

而国内厂商在应用于数据中心&AI 领域的高端FPGA 市场仍处在从0 到1 阶段。国内FPGA 厂商不仅需要硬件架构的创新，还需要EDA 软件和自研IP 能力的提升，以及更完善的国产应用生态培养。

05 ASIC

ASIC，Application Specific Integrated Circuit，即专用集成电路芯片，是一种为了专门目的或算法而设计的芯片。

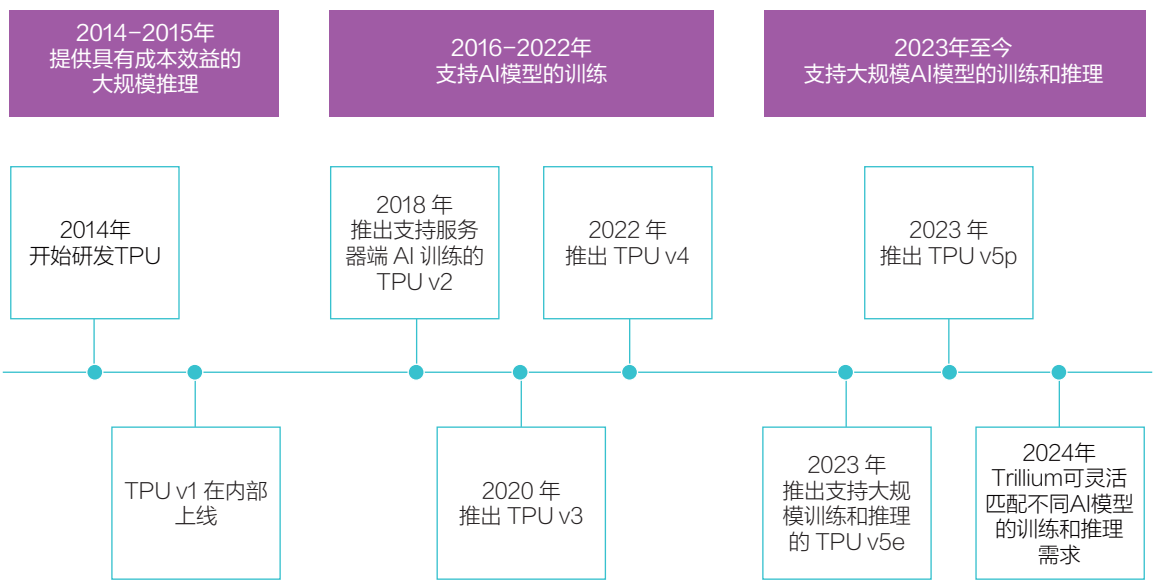
ASIC 芯片的架构并不固定，既有较为简单的网卡芯片，用于控制网络流量，满足防火墙需求等，也有类似于谷歌TPU 等的顶尖AI 芯片。只要是为了某一类算法或某一类用户需求而去专门设计的芯片，都可以称之为ASIC。

相比能够运行各种应用程序的GPU 和能够在制造后重新编程以执行不同任务的FPGA，ASIC 需要定制设计，灵活性较差。但由于ASIC 是为了某一类需求和算法而设计的芯片，因此其在特定应用中表现出色，性能明显优于其他芯片。

2023 年，数据中心定制加速计算芯片（ASIC）规模约66 亿美元，在AI 加速计算芯片市场占有率较低，为16%。业内预计，AI ASIC 芯片成长空间广阔，未来增速有望超过通用加速计算芯片。Marvell 称，2028 年定制芯片规模有望超400 亿美元，CAGR 达45%，而通用加速计算芯片2028 年预计达到1716 亿美元市场规模，CAGR 为32%。

目前，ASIC 芯片根据运算类型主要分为TPU、DPU 和NPU，分别对应不同的基础计

谷歌TPU发展史



资料来源：Google，华福证券研究所，第一财经整理

算功能。

TPU 即为谷歌发明的AI处理器，主要支持张量计算，DPU 则是用于数据中心内部的加速计算，NPU 则是对应了上一轮AI 热潮中的CNN 神经卷积算法，后被大量集成进了边缘设备的处理芯片中。

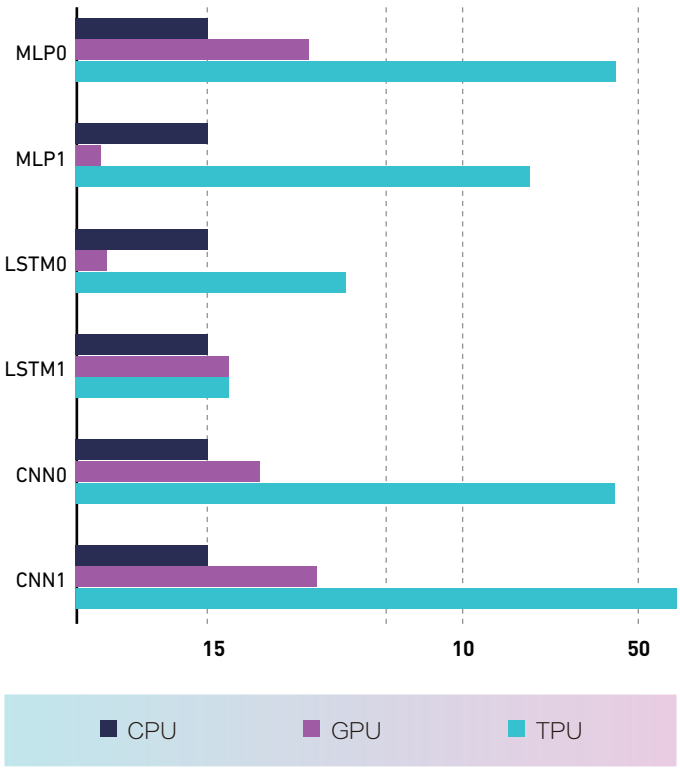
5.1 TPU：谷歌定制，能效比突出

TPU，即张量处理单元，属于ASIC 的一种，是谷歌专门为加速深层神经网络运算能力而研发的一款芯片，为机器学习领域而定制。

与传统CPU、GPU 架构不同，TPU 的MXU 设计采用了脉动阵列（systolic array）架构，数据流动呈现出周期性的脉冲模式，类似于心脏跳动的供血方式。

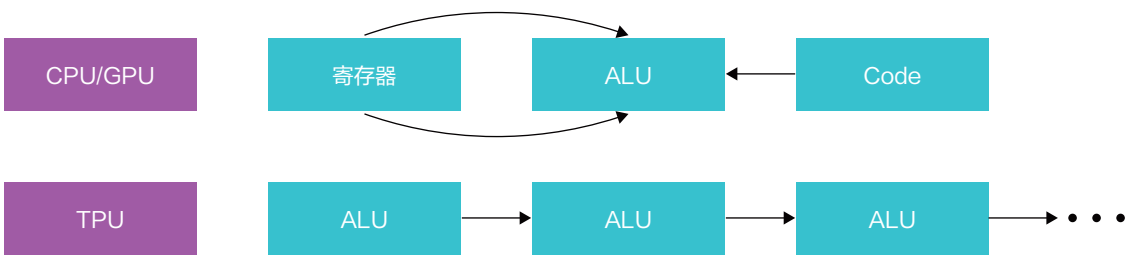
CPU 与GPU 在每次运算中需要从多个寄存器中进行存

TPU在神经网络上性能最大可达CPU的71倍



资料来源：Google Cloud官网，量子位公众号，天风证券研究所，第一财经整理

TPU与CPU/GPU运算方式对比



资料来源：天风证券，第一财经整理

TPU和GPGPU对比图

Google TPU								
芯片	TPU v1	TPU v2	TPU v3	TPU v4i	TPU v4	TPU v5e	TPU v5p	Trillium
发布时间	2015	2017	2018	2020	2021	2023	2023	2024
制程	28	16	16	7	7	5	5	4
BF16 TFLOPS	/	46	123	137.5	69	197	459	926
INT8 TFLOPS	92	/	/	138	275	394	918	1852
显存类型	DDR3	HBM	/	/	HBM2	HBM2	HBM3	HBM3或HBM3e
内存（GB）	8	16	32	8	32	16	95	32

NVIDIA GPU					
芯片	V100	A100	H100	H200	GB200
发布时间	2017	2020	2022	2023	2024
制程	12	7	4	4	4
FP64 TFLOPS	7.8	9.7	33.5	34	/
FP32 TFLOPS	15.7	19.5	67	67	/
FP16 Tensor TFLOPS	125	623.8	989	1979	5000
INT8/FP8 Tensor TFLOPS	/	/	1978	3958	10000
显存类型	HBM2	HBM2e	HBM3	HBM3e	HBM3e
内存（GB）	16	80	80	141	384

注：第五代和第六代TPU制程为The Next Platform推测
资料来源：The Next Platform，NVIDIA官网，芯东西，量子位，新智元，AI时代前沿，IT之家，AI科技评论，华福证券研究所，第一财经整理

取；而TPU 的脉动阵列将多个ALU 串联在一起，复用从一个寄存器中读取的结果。

相比于CPU、GPU，TPU 在机器学习任务中因高能效脱颖而出，其中TPU v1 在神经网络性能上最大可达同时期CPU 的71 倍、GPU 的2.7 倍。

华福证券认为，与英伟达GPU 相比，在算力上，谷歌TPU 目前暂时落后一代，在性能功耗比上谷歌优势显著。

根据Capvision，谷歌TPU70%-80% 的算力用于内部业务场景使用，剩余20%-30% 以租赁方式供外部使用。随着TPUv4 于2021 年推出和大型语言模型的出现，谷歌芯片业务的规模显著增加，2023 年TPU 出货量已突破200 万颗量级。

5.2 DPU：CPU 和 GPU 之后的“第三颗主力芯片”

DPU，即数据处理器，具备强大网络处理能力，以及安全、存储与网络卸载功能，可释放CPU 算力，能够完成CPU 所不擅长的网络协议处理、数据加解密、数据压缩等数据处理任务，并对各类资源分别管理、扩容、调度，实现数据中心降本提效。即处理“CPU 做不好，GPU 做不了”的任务。

在AI 时代，智算中心需要处理的数据量井喷，DPU 能够释放智算中心的有效算力，能够解决基础设施的降本增效问题，重要性和渗透率正逐渐提升。中国信通院预计，未来全球DPU 市场规模仍将保持30% 的复合增速，2025 年全球DPU 市场规模将接近150 亿美元。

5.2.1 三U 一体，更适配智算时代的解决方案

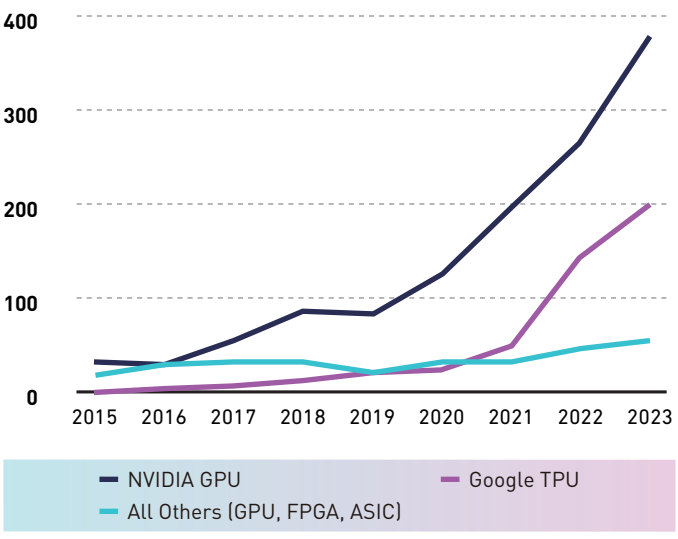
“DPU 这个概念是四年前被英伟达炒作起来的。在收购了以色列公司Mellanox 后，英伟达一跃成为业界首个既有CPU、GPU，也有DPU 的数据中心完整解决方案的供应商。”芯启源创办人卢笙在接受第一财经专访时表示，芯启源是国内最早一批从事DPU 研发的厂商之一，可以追溯到2018 年，那个时候还称为智能网卡Smartnic。

“过去承载网络传输功能的是传统网卡，后来诞生了智能网卡，四年前逐步演化为DPU。”专注于智能计算芯片研发设计的中科驭数（北京）科技有限公司高级副总裁张宇告诉第一财经。

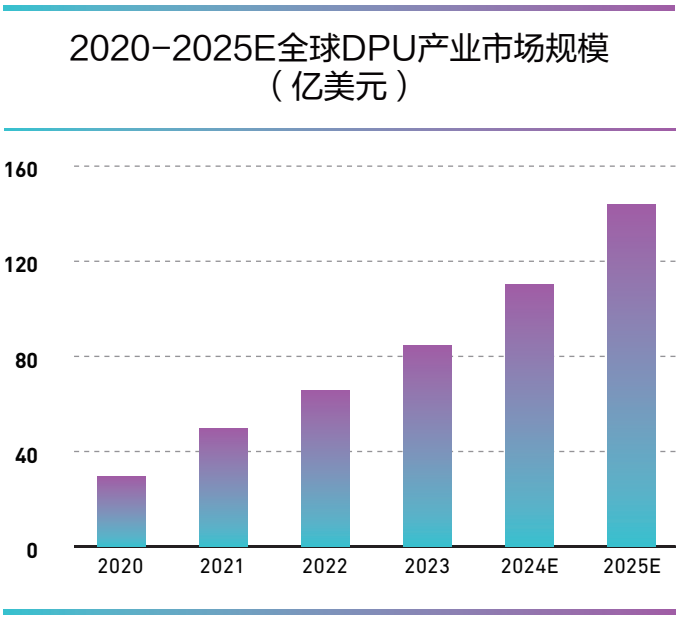
2020 年，英伟达发布的DPU 产品战略中将其定位为数据中心继CPU 和GPU 之后的“第三颗主力芯片”，自此引爆了DPU 概念。

如今，DPU 已成为数据中心内新兴的专用处理器，专门

2015-2023年全球数据中心加速器年出货量（万颗）



资料来源：TechInsights，华福证券研究所，第一财经整理



资料来源：中国信息通信研究院华东分院，赛迪研究院，第一财经整理

设计用于加速数据中心中的安全、网络和存储任务，针对高带宽、低延迟的数据密集型计算场景提供动力。DPU 的核心作用是接管原本由CPU 处理的网络、存储、安全和管理等任务，从而释放CPU 资源，并加强数据安全与隐私保护。

“英伟达所做的智算中心的方案，其实都是三U 一体的。英伟达三年前的DGX A100 服务器、后面的DGX GH200 等一系列，都是包含CPU、GPU 和DPU 的。当然DPU 里还有像RDMA 这种智能网卡，这些其实都可以归为DPU，它们本质上是一个东西。所以从这个角度看，当前行业内引领的，或者说大家公认的方向，是在智算中心里CPU 加GPU 加DPU 三者协同。”张宇表示，通用数据中心的方案则更多是CPU 加存储加网络，一些云原生场景里对低时延高吞吐的数据网络处理也是刚需，智算场景对网络处理性能要求更高。

“如果把CPU 比喻为大脑，用于整体控制，那么GPU 则更像是肌肉，用于提供坚实的充沛的并行计算的算力，而DPU 则更像是血管和神经，将GPU 需要算的数据，通过DPU 运输到服务器中，完成控制指令交换和协议转换。”张宇称。

“多PU 的配合实际上是整体计算架构的升级，从过去以通用CPU 为主的架构走向加速器为主的计算架构，通过CPU、GPU、DPU、NPU 等配合来提升整体计算方案的性价比。”张宇表示，“目前在技术方面，DPU 已经逐步趋向成熟，边界也比较成熟。网络安全加解密、零信任、网络卸载，已经基本上成为了DPU 稳定承载的功能。”

5.2.2 减少capex 投入和能耗，具有一定性价比

作为CPU 的卸载引擎，DPU 最直接的作用是接管网络虚拟化、硬件资源池化等基础设施层服务，释放CPU 的算力到上层应用，因此能够有效释放智算中心的算力，提升能效比。

“英伟达此前承认其上一代生成式AI 服务器的算力芯片的效率只有设计能力40%，我们测下来只有30% 多，这意味着大部分算力是被闲置的，究其原因主要是在集群间等待计算产生的中间变量完成数据同步，网络通路的能力限制了算力底座的上限，而这恰恰是DPU 的真正价值所在。”卢笙表示，这使得DPU 又被推到风口浪尖。

在数据量爆炸的AI 时代，DPU 不仅能够协助构建兼具低时延、大带宽、高速数据通路的新型算力底座，还能够安全高效地调度、管理、联通这些分布式CPU、GPU 资源，从而释放智算中心的有效算力。因此，DPU 的部署能够减少数据中心的一次性capex（资本性支出）投入。Cisco（思科）的数据显示，通过虚拟化技术，企业可以减少高达40% 的服务器数量，同时提高资源

利用率。

另一方面，DPU 通过专用硬件加速网络、安全和存储任务，提高了数据中心的能效。

卢笙介绍，以中国移动在浙江省SD-WAN 这一个应用场景为例，“通过芯启源DPU网卡打造的软硬件一体化解决方案，实现了网络安全业务卸载，相较于传统纯软件SD-WAN 网络方案，单机效率提升了6-8 倍，整体项目也节约了80% 的服务器部署投入和每年的软件费用等，极大降低了CAPEX 投入；此外，由于机器部署减少，数据中心的能耗降低，经测算每年可以节省超300 万度电，同时极大降低了数据中心的运营成本。”

成本方面，第一财经了解到，DPU 的研发和生产成本相对较高，特别是使用先进工艺时，因而价格较高，但由于部署DPU 解决方案，既能够减少服务器设备数量，也能在后续运行计算过程中节省能耗，因此整体系统成本上仍然具备一定性价比，但也要根据具体场景和应用情况来讨论。

5.2.3 未来三年是商用落地的关键时期

不过，目前DPU 的渗透率提升仍面临阻力。

专注于硬科技领域早期投资的创投机构中科创星相关人士告诉第一财经，DPU 作为软硬件协同的虚拟化架构，需要与CPU 中运行的虚拟化软件栈进行有效对接，同时，DPU 的硬件设计必须考虑到与现有系统的兼容性和集成性；其次，DPU 的架构和接口尚未形成统一标准，不同厂商的产品存在差异，这给用户在使用、维护和升级时带来挑战；此外，软件生态尚未成熟，缺乏完善的开发工具、驱动程序和操作系统支持，“但目前已经有公司在做”。

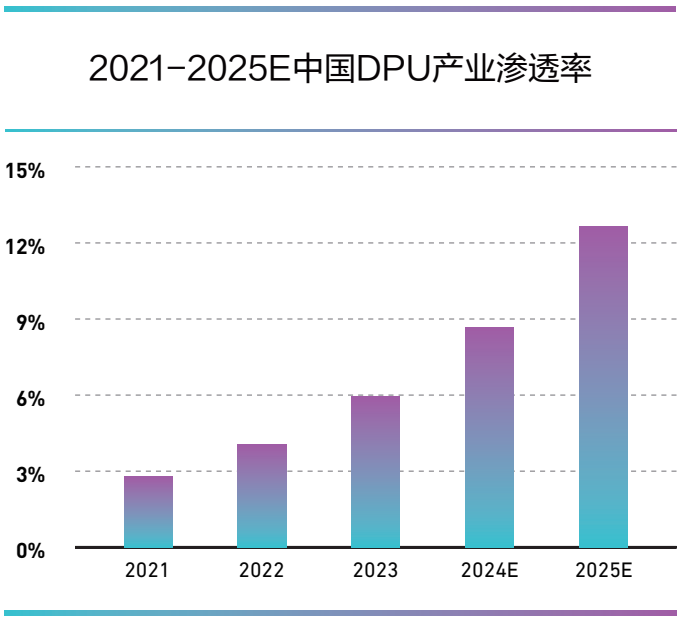
卢笙表示，DPU 需要专用的高效指令集，这也是其核心竞争力所在，剩下三分之二的工作则是围绕指令集打造生态，生态建设是DPU 行业的核心壁垒，生态建设的成熟度决定了产品商业化落地的速度。

整体来看，DPU 产业目前仍以国外企业为主导，三大巨头英伟达、博通和英特尔的份额占比较高，亚马逊、微软这些科技公司也在跟进。国内方面，中国移动、阿里等大企业也在研发专用的DPU，初创企业如芯启源、中科驭数、大禹智芯等也取得了相应的成果或进步。

“国内外的DPU 技术发展处于同一阶段，但国外企业积累更深厚一些。在我看来，DPU 产业其实已经逐步走向成熟和快速落地的阶段。国外可能比国内可能走得更早更快一点。”张宇表示。

在DPU 商业化落地方面，目前国内仅有华为、阿里、中兴等大型云厂商，以及芯启源、中科驭数等少数DPU 新势力已实现商用。信通院预计，2025 年我国数据中心DPU 渗透率可达到12.7%。

张宇认为，DPU 走到当前阶段，更重要是与云在IaaS 这一层的深度融合，尤其是如何给客户提供全面、便捷、透明



资料来源：中国信息通信研究院华东分院，第一财经整理

的纯软件的IaaS 方案，使他们能够平滑地迁移到用DPU 来支撑这种高能效比的云方案。

“这块的迁移需要行业的共同努力，并且需要持续很长时间，甚至要以年来计。”张宇称，“亚马逊云就走得比较快，他们研发实力比较强，已经完成了IaaS on DPU 的转换，但对国内大部分企业来说，步伐不会迈得太大，可能会从最痛的几个点先透明地用起来，比如OVS 卸载、网络升级等。”

“DPU 的商业化不仅仅依靠于传统数据中心的IaaS 领域，也包括网络安全、高性能存储、集群通信等诸多的行业和领域。”卢笙表示，芯启源多年来深耕“DPU for Security”方向，将DPU 应用于诸如防火墙、安全网关等产品，目前已进入深信服网安产品线，成为标配扩展卡，解决了诸如Intel CPU 大象流处理能力不足等业界难题。

“从目前的产业发展趋势来看，如果技术发展符合预期，大概2025-2027 年会有一个爆发。”上述中科创星相关人士表示，原因在于，随着数字经济、AI 和云计算产业的发展，服务器市场会迎来一个增长，尤其是在金融、政府和电力用户领域，不仅需要大量的DPU 来处理数据，提升计算效率，还需要DPU 发挥安全性的优势。

“DPU 芯片确实已经大规模应用了，目前的增速在每年20%-30%。但DPU 的行业特质就是需要保持稳定性，需要稳定在集群上运行几个月再扩集群。”张宇表示，更重要的一点，结合国内信创行业的发展来看，这两三年会是非常关键的时期，是每家DPU 厂商都需要把握好的关键时间窗口。

主要国产DPU厂商及产品

芯片类别	厂商	产品系列	详情
DPU	阿里云	CIPU2.0	为阿里云基础设施量身打造的处理器，这一系列新产品涵盖了通用计算、加速计算和容器计算等多个领域。
	百度智能云	太行DPU2.0	基于百度太行DPU2.0实现了计算架构的全面升级，实现了全核售卖的虚拟机能力、可预期的存储和网络I/O性能，网络带宽升级到200Gbps，网络转发性能达到5000万PPS。
	北中网芯	鲸鲨NE系列	专注于数据处理及网络传输，完成诸如网络协议处理，交换路由，加密解密，安全检测等对性能及效率有更高要求的工作，成功解决25G和100G网络接入问题。
	大禹智芯	Parayus系列	Paratus2.0可广泛应用于公有云，边缘云，企事业内部私有云，及其他需要高性能网络支撑或复杂网络流量处理等场景。
	芯启源	Agilio智能网卡	Agilio智能网卡的通用计算网络解决方案，为服务提供商和企业云网络中的计算和服务节点应用提供了高性价比和可扩展性。
	中科驭数	K2-Pro	K2-Pro是专为解决大规模数据中心所面临的网络带宽、网络规模、传输效率、部署管理等一系列挑战，而设计的高效能网络处理芯片。该芯片采用纯自研KPU架构，融合了网络卸载、存储卸载、安全卸载、计算卸载等功能。
	移动云	大云磐石DPU	带宽达到400Gbps，将广泛应用于中国移动数据中心建设，涵盖通用计算、智能计算等业务场景。
	天翼云	紫金DPU系列	采用SOC+FPGA架构，应用于数据中心、云计算、边缘计算领域。

资料来源：公司官网，第一财经整理

“DPU 还不是一个标准化的产品，商业化上量的过程需要结合市场的需求和不同应用场景的深入打磨，需要上下游厂家协同，从几百片小规模试点到上万片的大规模部署需要积跬步至千里。”卢笙表示，DPU 的商业化需要全行业伙伴共同努力，加强彼此的生态互信与合作，在国产化 CPU+GPU+DPU 的3U 一体商业化之路携手前行。

5.2.4 国内DPU 产业蓄势待发

“好消息是中国厂商和国际厂商都在同一起跑线上。”卢笙表示，随着需求增长，中国市场本身体量很大，同时国家大力推进数字经济，这都给中国厂商发展提供了很好的土壤。

卢笙认为，中国厂商在 DPU 领域能够脱颖而出的最大优势在于中国数字经济的应用场景走在世界前列，比如随处可见的移动支付等。

国内DPU 产业亦蓄势待发，除了云厂商外，芯启源、中科驭数、星云智联、大禹智芯等企业纷纷入局。

5.3 NPU：专为 AI 应用而生

NPU，即神经网络处理单元，用于高效执行神经网络的计算，通常具有优化的硬件架构，如向量处理单元、矩阵乘法单元、卷积单元和激活函数单元等，能够在硬件级别上执行大规模矩阵运算和卷积运算，以提高神经网络计算效率。

当前各类AI 算法主要利用深度神经网络等算法模拟人类神经元和突触，NPU 能够实现更高效率、更低能耗处理人工神经网络、随机森林等机器学习算法和深度学习模型。如今，多家手机厂商已搭载NPU，AIPC 也将通过“CPU+NPU+GPU”打造本地混合计算。

5.3.1 NPU 高效能、低功耗、更擅长推理

“相比于CPU 和GPU，NPU 的优势在于高效能、低功耗、易于编程、降低了开发门槛，同时支持多种语言和框架方便开发者进行模型开发和部署。”IDC 中国分析师杜昀龙告诉第一财经。

NPU 采用“数据驱动并行计算”的架构，在电路层模拟人类神经元和突触，特别擅长处理视频、图像类的海量多媒体数据。区别于CPU、GPU 所遵循的冯诺依曼架构，NPU 能够通过突触权重实现存储计算一体化，提高运行效率，因此比GPU 更擅长推理。且 NPU 芯片设计逻辑更为简单，在处理推理工作负载时具有显著的能耗节约优势。

杜昀龙认为，缺少像英伟达GPU 那样完善的生态环境是目前NPU 渗透率提升最大的瓶颈。据IDC 数据，2022 年中国实现数据中心计算加速仍以GPU 为主，NPU 占比仅12%，但较以往明显增长。

5.3.2 NPU 更适合端侧、边缘侧部署

如今，大模型已进入轻量化时代，端侧AI 应用正加速落地，商汤（00020.HK）曾在2023 年年报中表示，2024 年将是端侧大模型应用的爆发之年。商汤联合创始人、首席科学家王晓刚解释称，成本、数据传输延迟、数据安全与隐私等几个重要问题，都可以通过端侧AI 或云端结合来解决。

与云侧不同的是，端侧对于功耗更加敏感，对低功耗芯片的需求更明显。因此，随着人工智能应用场景陆续落地，NPU 易开发、高效能、低功耗等优势逐渐突显。业内普遍认为，在大算力需求爆发下，云侧的算力需求将传递至端侧。目前，实现智能终端算力的最常用方式是在SoC 芯片中内置 NPU 模块。

“NPU 是专门为AI 应用而设计的芯片，目前看NPU 通常用于边缘侧和端侧场景更多，比如人脸识别、人脸解锁、影像处理等。”杜昀龙表示。

“大模型部署到端侧时，对芯片的算力、存储、数据传输都有明确要求，要求这些芯片能更好地支持Transformer、Diffusion 等架构。”王晓刚对第一财经表示。

相比于在云端用GPU 部署 Transformer 大模型，在边缘侧、端侧部署Transformer 的最大挑战来自于功耗。因此在

端侧和边缘侧，GPU 并非最合适的架构。

“Transformer 是现在所有大模型的基础，它本身的基础架构也是固定的，在这种基础最原子层架构固定的情况下，DSA（专用领域架构处理器）架构是比GPU 更优的架构。对于AI 推理芯片，尤其是在边缘侧、端侧部署的推理芯片来说，DSA 架构芯片的运算效率、能量消耗优于GPU。”国内AI 算力公司爱芯元智创始人、董事长仇肖莘表示。

NPU 就是基于DSA 领域专用架构技术的处理器，同时比GPU 更擅长推理，且芯片设计逻辑更为简单，具有能耗节约优势，因此NPU 可以作为在AI 计算效率和功耗之间的平衡方案。

“NPU 跟CPU、GPU 一样，都是整个计算架构的一部分，是计算架构里边异构的处理器，用以实现不同的功能。GPU 原来是专门为图形渲染做的，更像是张量计算使用的一个处理器，而NPU 更像是一个原生的AI 计算处理器，CPU 则是另外一个处理器，这三个处理器加起来构成了AI 的一个计算底座。”仇肖莘在接受第一财经专访时表示，“实际上，NPU 在GPU 的基础上更近了一步，它的灵活性不如GPU，但它的好处是能效比非常高，更适合边缘侧、端侧。”

5.3.3 NPU 渗透率提升，成本是关键

NPU 适合广泛部署在端侧、边缘侧的重要原因，也在于端侧与边缘侧需要性价比，需要在成本可控的前提下提供AI 能力。仇肖莘认为，如何以低成本的NPU 去实现AI 的价值，是实现AI 普惠的关键。

“现在自动驾驶很火热，但一个显而易见的问题是，到底有多少车能够用得起这个功能？如果一套解决方案花费在1 万元人民币，可能10 万元以下的车就用不起；如果能够把它的整个成本降到5000 元以下，或者3000 元，占整车成本的比例就会低很多，普及也会更快一些。”仇肖莘称，现在有一些高端手机和新出的AIPC 已经搭载了NPU，但价格确实比较贵，要把NPU 的成本降下

主要国产NPU厂商及产品

芯片类别	厂商	产品系列	详情
NPU	华为	NPU	采用达芬奇架构，集成在麒麟芯片上。
	瑞芯微	RK3576系列等，其中内置自研NPU	RK3576采用先进制程设计，搭载公司自研6TOPs算力的最新一代NPU，支持Transformer模型架构相关算子。
	爱芯元智	爱芯通元混合精度NPU	支持混合精度算法设计，natively支持Transformer网络结构，为大模型在边缘侧、端侧的应用提供良好的基础。
	寒武纪	思元220	基于寒武纪MLUv02架构，可以实现终端设备和边缘端设备的AI赋能方案。
	平头哥	含光800	含光800是一颗高性能人工智能推理芯片，基于12nm工艺，采用平头哥自研架构，通过软硬件协同设计实现性能突破，目前已成功应用在数据中心、边缘服务器等场景。
	国科微	AINPU	用于车载领域，可满足辅助驾驶、视觉感知的算力需求。

资料来源：公司官网，第一财经整理

来，需要达成规模效应，现在量没有起来确实很难降价。

仇肖莘认为，从今年开始，NPU 的需求，或者说端侧AI芯片的需求会是一个快速上涨的趋势。“从我们自己客户的需求来看，他们的预测都会比原来更高。”

5.3.4 国产芯片厂商大力布局

目前，国内芯片厂商正奋力自研NPU，以迎接AI 浪潮。以阿里平头哥为代表的芯片公司已推出面向数据中心AI 应用的人工智能推理芯片，其NPU含光800 已成功应用在数据中心、边缘服务器等场景。

NPU IP 方面，芯原股份（688521.SH）2016 年通过对图芯美国的收购，获得了图形处理器（GPU）IP，在此基础上自主开发出了NPU IP。芯原股份此前告诉第一财经，目前，在AIoT 领域，公司用于人工智能的神经网络处理器IP已经被50 多家客户的100 多款芯片所采用，被用在物联网、可穿戴设备、安防监控、服务器、汽车电子等10 个应用领域。

06
AI 算力产业趋势展望

本报告所讨论的CPU、GPU、FPGA、ASIC 芯片，提供了当前阶段的主要AI 算力。但AI 大算力底座的构成并不局限于上述芯片，还包括存储器等存力芯片、接口芯片等运力芯片、服务器等硬件集成、交换机等硬件设备、数据中心运维、通信网络传输等，总之，AI 算力发展需要全产业链的共同进步。

第一财经在与产业人士、机构投资者、第三方分析机构等交流探讨上述AI 芯片的发展现状过程中，发现了部分行业趋势，主要有以下几点：

第一，AI 推理环节，FPGA 和ASIC 或能争夺更多市场。

异构计算趋势下，GPU 仍是当前除CPU 外商用最广泛的AI 芯片。这主要是由于CPU 难以高效率地胜任多个复杂的并行计算任务，需要GPU 等AI 芯片辅助完成部分任务。GPU 作为CPU 的协处理器，至今已能够在性能、成本等方面具备优势，更重要的是，英伟达打造的CUDA 生态，让其GPU 产品拥有较为完善的生态环境，这些都推动了GPU 的商业化。

相比GPU，FPGA、ASIC 在性能和效率上更具优势，但成本更高，尤其是高端FPGA 更依赖于先进制程，因此广泛商用仍需一定时日。

值得注意的是，CPU 和GPU 都属于冯诺依曼架构，计算和存储功能分别由芯片和存储器完成，数据从存储器中获取，处理完毕后再回到存储器，从处理单元外的存储器提取数据所需的时间往往是运算时间的成百上千倍，即存在“存储墙”限制。而FPGA 和ASIC 却没有这样的限制。

因此，GPU 不适合用于AI 推理。当前动辄千亿参数的AI 大模型需要千卡、万卡集群算力进行训练，而在后续的AI 大模型应用阶段，AI 推理将成主战场，FPGA 和ASIC 或许能获得更多市场。

第二，存算一体或成为AI 时代新选择。

FPGA、ASIC 类似于GPU 的大规模商用仍待时日，在此期间，配合GPU 的广泛应用，存力与运力也在快速发展。

2024 年初，搭载在高端AI 服务器配合GPU 的HBM（High Band width Memory，高带宽存储器）爆火。HBM 拥有超高带宽，主要缘于其将原本在PCB 板上的DDR 内存颗粒和GPU 芯片同时集成到SiP 封装中，使内存更加靠近GPU，即“近存计算”。随着存算技术的发展，未来存内计算、存内逻辑，即“存算一体”，有望成为AI 时代的新选择。

存算一体是在存储器中嵌入计算能力，以新的运算架构进行二维和三维矩阵乘法/ 加法运算。其中存内计算的计算操作由位于存储芯片内部的独立计算单元完成，存储和计算可以是模拟的也可以是数字的，一般用于算法固定的场景算法计算；存内逻辑通过在内部存储中添加计算逻辑，直接在内部存储执行数据计算，这种架构数据传输路径最短，同时能满足大模型的计算精度要求。

此外，AI 服务器内存和硬盘也在迭代升级，已开始搭配最先进的DDR5、LPDDR5、SSD 等存储。

第三，异构计算时代，先进封装战略地位将突显。

摩尔定律已逼近物理极限，先进封装成为系统级解决方案，战略地位将不断凸显。

AI 芯片就融合了多种先进封装技术，HBM 应用TSV 堆叠技术获得超高带宽，而为了将HBM 和GPU 集成，

CoWoS 封装技术被深度开发。CoWoS 是英伟达选择的主流封装技术，能够以合理的成本提供最高的互连密度和最大的封装尺寸。人工智能加速芯片可能用到的Chiplet 技术也依赖于先进封装。

半导体产业链中，我国封测环节发展已较为成熟，长电科技（600584.SH）、通富微电（002156.SZ）等企业近年来不断加深对先进封装的布局。

第四，国产AI 芯片厂商亟待构筑护城河。

技术不断迭代且流片成本高昂的芯片行业存在明显的“二八定律”。

“芯片行业有一个非常明显的特点，它会有一个体量巨大的老大和一个还活得不错的老二，剩下的全都会死。”某AI 算法上市公司高管此前告诉第一财经，“因为芯片的主要原材料是沙子，难的是工艺。换句话说，流水线一旦启动，造一颗芯片和造100 万颗芯片的价格不会相差很多，所以出货量大的企业才能盈利。很多芯片创业公司的经营状况不好，这是符合市场规律的，他们需要在残酷的市场竞争中最少变成老二。”

因此，先发优势是芯片厂商构筑核心壁垒的关键因素之一，在某个细分领域站稳脚跟后迭代下一代产品，凭借生态和口碑建立起用户习惯，后来者分羹的难度会越来越大，英伟达就是最好的案例。

近年我国支持自主可控的氛围浓厚，无论是政策端还是产业下游客户，都在积极配合半导体产业链上游和中游，国内厂商亟待借力政策和产业东风，加速构筑自身护城河。

第五，国内AI 芯片生态构建将加速。

正如英伟达凭借CUDA 占据全球GPU 的绝对份额，英特尔凭借X86 的成熟生态占据服务器CPU 的绝对份额，国内AI 芯片厂商也在加速自研和生态构建之路，如兼容CUDA 的海光信息、摩尔线程等，自研架构搭建生态的华为、燧原科技等。

其中，华为正在通过CANN 和AscendCL 构建国产GPU 软件生态。CANN 是华为为AI 场景设计的异构计算架构，旨在为开发者提供高效、灵活且易于使用的工具和服务，以便在华为昇腾AI 硬件上开发和部署各种人工智能应用。CANN 对上支持多种AI 框架，对下服务AI 处理器与编程，这意味着无论使用哪种AI 框架，CANN 都能提供支持。

软件生态建设非一日之功。随着政策支持下国产硬件数量的增加，软件开发商将更积极地为这些平台开发应用，从而丰富软件生态，形成硬件发展与促进软件生态的良性循环。

07
结语

整体来看，我国AI 芯片产业发展仍将面临多重阻力，一方面，上游EDA 工具、IP、晶圆代工厂的制程进度等，给AI 芯片设计厂商带来制约；另一方面，国外企业凭借先发优势、技术优势、生态配套及用户习惯已占据大部分市场份额，国内厂商在打开下游市场时面临诸多挑战，在产品研发上也需要考虑兼容路径。

不过，中国市场规模庞大，技术与生态百花齐放，且政策扶持力度大，产业链配合积极度高，未来随着先进封装、存算一体等技术的发展，推理芯片的大规模运用，中国AI 芯片厂商有望走出自己的发展壮大之路。

本报告参考文献

[1] AI 智算时代已至，算力芯片加速升级- 中航证券
[2] 电子行业“FPGA 五问五答”系列报告三：FPGA 在各行业究竟用在哪里？未来哪个下游最有机会？- 东兴证券
[3] TPU：为更专用的AI 计算而生- 华福证券
[4] 计算机行业：重视AI 芯片配套的软件生态- 国泰君安
[5] FPGA 国产化龙头，凤凰天地阔，涅槃终有时- 德邦证券
[6] 计算机行业算力知识普惠系列一：AI 芯片的基础关键参数- 天风证券
[7] 海外科技行业：算力需求高增，AI ASIC 突围在即- 国泰君安
[8] 计算机行业深度研究报告：业绩拐点已至，安全可控进入新发展周期- 华创证券

数据说明

数据丨案例丨观点来源

如无特殊说明，报告中数据和内容均来自第一财经的调研、采访及公开资料。

版权声明

本报告页面内容、页面设计的所有内容（包括但不限于文字、图片、图表、标志、标识、商标、商号等）版权均归上海第一财经传媒有限公司（以下简称“我司”）所有。凡未经我司书面授权，任何单位或个人不得复制、转载、重制、修改、展示；不得以任何形式提供给第三方使用本报告的部分或全部内容。任何单位或个人违反前述规定的，均属于侵犯我司知识产权的行为，我司将追究其法律责任，并根据实际情况追究侵权者赔偿责任。

免责声明

本报告中所载的内容、资料及相关数据来源，均被视为最初发布当日作者的判断，并不保证本报告中的内容及观点在将来不会发生任何变更。我们力求但不保证本报告所涉及信息的准确性和完整性。报告中所表述的观点、信息，在任何情况下、对任何人不构成投资建议。在任何情况下，任何人因使用本报告中的内容所引致的后果应自行承担责任。

出品
第一财经丨第一财经投研中心

主编
钱焜

主笔
冯丽君

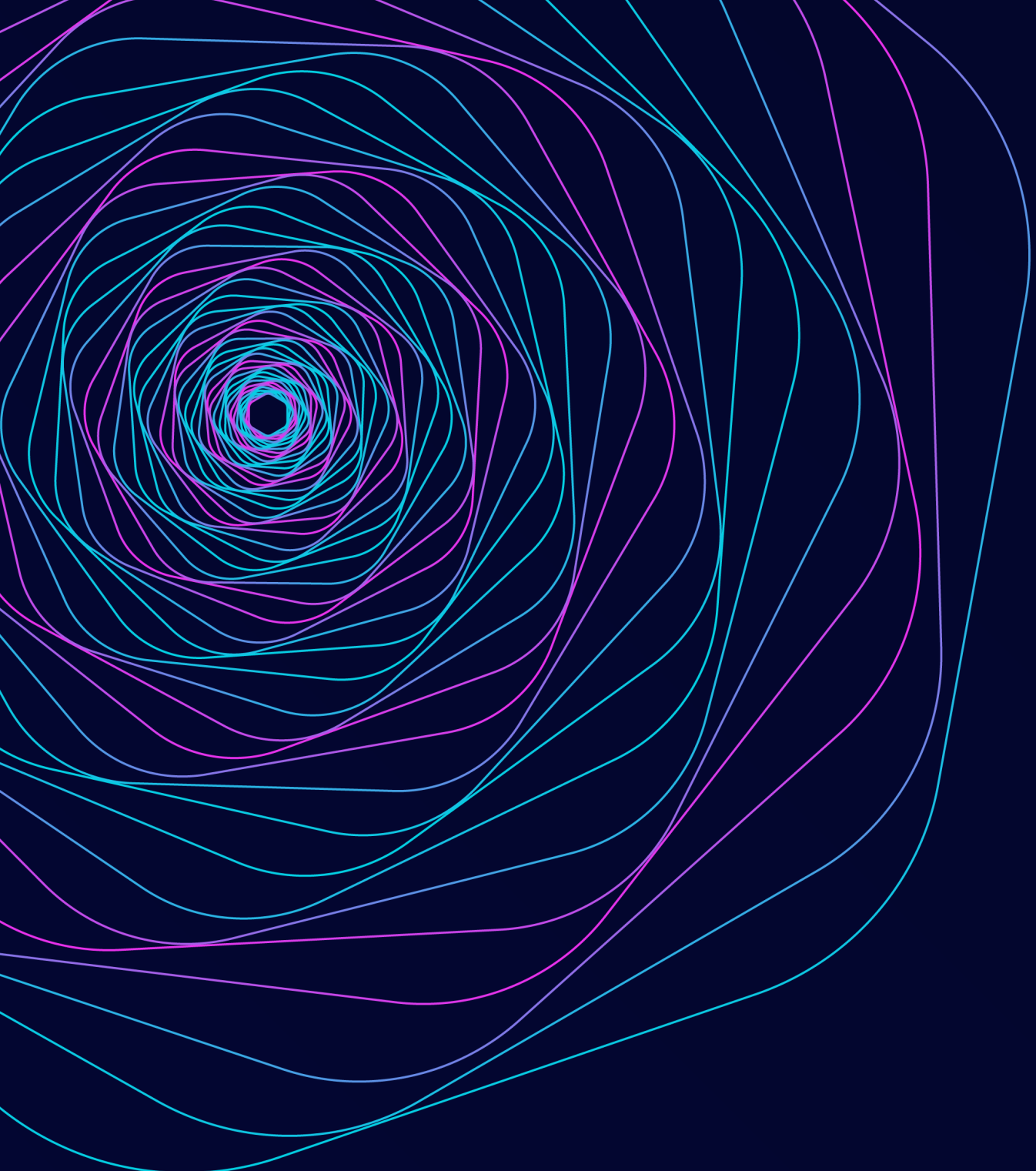
编辑
黄宇

审定
钱焜 黄宇

视觉
符乐乐

统筹
朱国泉 周瑾

联系人
冯丽君 fenglijun@yicai.com



第一财经 YICAI

YICAI
投资研究中心
Investment Research Centre