北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

金融机构 AI 芯片应用情况专题报告

北京金融科技产业联盟

2023 年 11 月

北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

版权声明

本报告版权属于北京金融科技产业联盟，并受法律保护。转载、编摘或利用其它方式使用本报告文字、图表或观点的，应注明来源。违反上述声明者，将被追究相关法律责任。

北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

编制委员会

编委会成员：

潘润红

编写组成员：

聂丽琴　胡达川　王　硕　纪　钟　罗方华　徐　斌
刘玉海　崔雨萍　伊　纯　方　科　徐梓丞　王静逸
岳永强　郭　贞　陆　俊　原菁菁　杜依迪　解　培
张　彬　李银凤　胡　捷　邓玉洁　徐小芳　王景俊
薛　亮　白　阳　张增金　洪喜如　朱军民　王　勇
武凤霞

参 编 单 位：

北京金融科技产业联盟秘书处

中科可控信息产业有限公司

中国银行股份有限公司

中国建设银行股份有限公司

中国邮政储蓄银行股份有限公司

中国光大银行股份有限公司

华夏银行股份有限公司

华为技术有限公司

北京趋动科技有限公司

北京易道博识科技有限公司

格兰菲智能科技有限公司

北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

目 录

北京金融科技产业联盟
BEIJING FINTECH INDUSTRY ALLIANCE

1 研究意义

习近平总书记在中共中央政治局第九次集体学习上的讲话强调:“人工智能是新一轮科技革命和产业变革的重要驱动力量，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题”。

近年来，人工智能技术在金融领域广泛应用，主要在信贷审核、智能客服、量化交易、金融反欺诈等业务场景应用落地。近期，人工智能现象级应用 ChatGPT 在社会中备受关注，再次引起人工智能、大模型、算法、加速卡等概念的热议。人工智能技术的发展对金融行业具有深远意义。

一是交易模式发生变革。在 2010 年之前客户交易的主要介质为存折、银行卡，以柜台人工服务模式为主；2010 年之后以自助设备替代高柜和低柜，并通过远程视频与自助设备结合实现交易达成；2016 年手机成为了新一代的交易媒介，移动金融成为主流，指纹、人脸等生物识别技术实现了通过人体生物特征信息与金融账户体系的关联，身份核验服务实现密码替代。

二是数据价值充分挖掘。在 2014 年之前，凭证和文件信息都是以图像的方式存储在影像平台，在出现问题的时候通过人工检索的模式来收集交易过程信息和证据信息，且收集到的数据和信息分散，没有充分发挥数据价值。光学字符识别（Optical Character Recognition，OCR）技术的诞生将金融机构大量的影像文件信息进行了识别处理，实现自动化分类整理、内容级搜索、概要信息提取、文件合规性审核、归档、统计分析、知识图谱构

建等，将零散的信息整理形成具有价值的数据。

三是客户服务智能化升级。传统的金融客服平台需要大量人力支撑，存在人员压力大、业务培训难、高峰期人员瓶颈、通话数据价值浪费等问题。人工智能等技术的应用使得金融机构客服坐席通过语音识别（Automatic Speech Recognition, ASR）、语义理解、语音合成（Text-To-Speech, TTS）、语音克隆、智能导航等人工智能的应用，实现了电话营销、电话邀约、智能催收、电话回访、语音通知等各场景的智能化升级。

人工智能应用的成功离不开强大的算力能力支撑，如果说算力是人工智能的“发动机”，那么AI芯片就是人工智能的“火花塞”。当前金融行业AI芯片应用存在难题。

一是AI芯片供需不平衡。一方面，随着生成式人工智能、大模型、隐私计算、大数据等技术的应用逐渐向成熟化和商业化发展，带动了算法公司、应用方等产业各方对AI芯片及服务器投入，尤其是高端芯片的需求不断增长，超出了原有AI芯片的供给能力。另一方面，国内在高端芯片制造方面还存在不足，而一些非市场性因素又限制了国内机构采购国外高端芯片的渠道，导致国内机构面临AI芯片采购难的问题。

二是AI芯片应用成本高。一方面，AI芯片的应用不同于消费级显卡以及零售客户对芯片的需求，金融机构单次采购量少则几十、多则几百，而目前供不应求的市场关系导致AI芯片单价居高不下，大量采购提高了应用成本。另一方面，AI芯片自身也在持续创新和技术进步，随着金融产品和服务的迭代创新，对性

能和效率更高的 AI 芯片更新需求也在不断增加，导致机构持续投入成本上升。

三是异构芯片池化管理不完善。随着人工智能产业的蓬勃发展，不同厂商、不同型号的芯片陆续发布；同时金融机构也开始测试不同芯片性能，开展芯片领域信创工作。目前存在异构芯片的资源池化管理和资源的远程调用能力不完善、AI 算力资源利用率不高等问题。

四是信息安全面临挑战。金融业是数据密集型行业，信息安全不仅关乎金融用户的资产安全和隐私保护，还关系到国家金融系统的安全稳定运行。国产芯片符合我国加密算法相关标准，产品经过安全性测试和认证，与金融机构技术架构开展适配性验证，从硬件、算法等方面保障金融机构信息安全。

综上所述，人工智能的发展对金融行业产生了变革式的影响，极大提升了金融服务实体经济的智能化和数字化水平。研究并解决当前金融机构 AI 应用“卡脖子”、应用成本高、算力资源管理效率低下、信息安全等现实问题，有助于满足金融机构对于 AI 芯片硬件安全可控、供应链可持续、产品高性能等需求，对金融业高质量发展具有重要价值。

2 技术路线

为什么 AI 芯片被视为人工智能应用的重要基础，CPU 是否可以承担相应的工作任务？CPU 作为中央处理器，最擅长的是让各计算指令在串行模式下一条接一条的有序执行，但是在诸如深度学习等人工智能场景下，并不需要太多的程序指令，而是需要海量数据运算，此时 CPU 就无法满足需求。而诸如图形处理器（Graphics Processing Unit, GPU）等 AI 芯片具有高并行结构，在处理图形数据和复杂算法方面拥有比 CPU 更高的效率。

AI 芯片主要包括 GPU、专用集成电路（Application Specific Integrated Circuit , ASIC）、现场可编程门阵列（Field Programmable Gate Array, FPGA）、专用领域架构（Domain Specific Architecture, DSA）处理器等类型。

表格 1 不同 AI 芯片技术路线对比

平台	性能	资源效率	灵活性	软件生态
CPU	单核性能强，但核心数较少	CPU 硬件加速效率最低	指令集最健全，编程灵活	广泛应用，软件生态庞大且成熟
GPGPU	并行计算性能相较 CPU 大幅提升，比 DSA/ASIC 有较大差距	高效集成数千至上万个高效率小核，以及数百个 AI 加速核，资源效率很高	本质是众核并行，指令集丰富，通用性强，AI 算子编写具有很高的灵活性，编程相对简单	1. 由于 GPGPU 有极强通用性，主流的和新出的 AI 算子和框架都基于 GPGPU 实现，AI 生态较健全 2. 硬件产品升级后，计算框架容易向后兼容，GPGPU 软件生态可持续维护
ASIC	针对特定场景定制，无冗余设计，理论上有极致的性能。受限于设计复杂度，难以超大规模设计	理论上最高的资源效率，但不可避免存在功能超集	1. 功能逻辑完全确定，通过配置方式调整功能 2. 不支持 AI 算子编写	不同领域及不同厂家的 ASIC 实现都存在巨大差别，且都需要特定驱动程序，软件生态不完善
FPGA	针对特定场景定制，无冗余设计 AI	AI 算法硬件定制，资源利用效率取决于算法设	灵活性较 ASIC 和 DSA 高，但是 AI 算法设计难度大，	需要针对具体 AI 算法或应用进行定制化设计和开发，软件生态不完善

	算法和应用，计算性能高	计水平和实际需求	需要大规模开发团队	
DSA（如NPU、DSP等）	接近ASIC的性能，由于一定程度的软硬件解耦，能够实现较大规模设计	1.在 AI 算子与 DSA 指令集高度适配情况下，DSA 资源效率很高，甚至接近 ASIC 2.在 AI 算子与 DSA 指令集不适配时，算子性能表现很差，甚至无法适配	支持少量指令，算子编写难度大，需要强大的编译器，把算法映射到特定 DSA 架构；相比 ASIC，具有一定的可编程性	1.软件生态不成熟，主要是自研框架 2.对主流计算框架不支持或支持度十分有限 3.AI 算法及框架层出不穷且更新频繁，DSA 无法及时支持或无法支持，无法构建持续的 AI 软件生态

2.1 硬件层

AI 芯片发展主要路线包括 GPGPU、NPU 等。

2.1.1 GPGPU 路线

GPGPU 是并行计算加速卡技术架构，兼顾并行计算性能和编程通用性，基于该架构的 CUDA（Compute Unified Device Architecture）支撑起了由数百个 AI 软件工具构成的开放 AI 软件生态。

（1） 技术实现

GPGPU 作为运算协处理器，具有高效并行性、高密集运算、超长流水线等技术优势，并可以针对不同场景的需要，增加专用向量、张量、矩阵运算指令，提升浮点运算的精度和性能，整体提升 GPU 的技术性能。

1） 整体架构

典型的 GPGPU 架构的核心部分为可编程多处理器，其核心部分包含了众多可编程多处理器，每个可编程多处理器又包含了多个流处理器，可以支持整型、浮点、特殊函数、矩阵运算等多

种不同类型的计算。GPGPU 通过 PCI-E 总线与 CPU 处理器进行通信，其存在的目的是为了对程序某些模块或者函数进行加速。GPGPU 是原硬件系统的一个扩展，接受 CPU 调度指挥，其硬件构造由计算单元、内存控制器、线程调度器等组成。同时借助特定的互连结构和协议，在一个主机系统 PCI-E 总线上可以使多个并行的 GPGPU 与 CPU 进行互连，这使得一台主机的算力具有可扩展性，合理的组织多个 GPGPU 可以获得更好的加速效果。

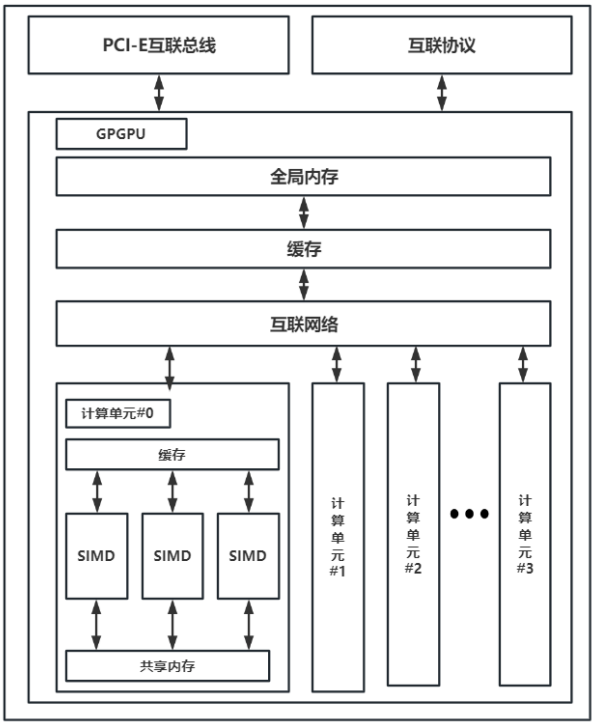

图 1 GPGPU **整体架构**

GPGPU 里有多个计算单元，每个计算单元使用多个单指令多数据（Single Instruction Multiple Data, SIMD）单元，每个 SIMD 单元里又有很多加减乘等流式计算部件，GPGPU 可以有众多流处理单元，因此其吞吐量非常高，如果任务有足够的并行性，GPGPU 可以更快完成。

GPGPU 有相对完整的缓存系统，为数据的重用提供了便利，减少了开发者手动控制的难度，提升了开发效率。缓存系统虽然对用户是编程透明的，但极致性能优化时也有一定的影响，需要考虑数据排布和读取方式等优化方法来提升缓存命中率。GPGPU 有可共享的全局内存单元，用于存储计算单元计算时所要访问的数据。

当 CPU 接受到数据传输指令即将数据拷贝到 GPGPU 时，直接内存访问（Direct Memory Access, DMA）单元来接管这一过程，当 CPU 转交这一个控制权后可以继续执行后续指令。DMA 传输不能保证读取数据时，数据一定是传输完毕的，所以在应用中需要通过查询传输是否完成来保证数据使用的安全性，合理的使用 DMA 可以提高程序的并行度。

①计算核心

SIMD 架构是计算单元的计算核心部件，可以访问寄存器文件，当 GPGPU 进行并行计算时，线程会被分配给计算单元。计算单元包括矢量计算单元和标量计算单元，其中矢量计算单元主要用于复杂计算，而标量计算单元主要用于地址计算、分支跳转等。当多线程同时启动时，会出现同一个计算单元被多个线程共用的情况，在使用时需要考虑资源的分配情况，合理的将线程进行分配。

②存储核心

GPGPU 的寄存器数量相较 CPU 更多，寄存器的访问等待时间比较短，因此在进行密集计算时，可以将常用的数据保存在寄

存器中，减少数据读取的开销。而不同线程之间也可以通过寄存器来进行数据的交互，因此数据读取延迟就可以通过多个线程来回切换进行掩藏。GPGPU 还提供了多种类型的片上存储空间，如 L1 数据缓存、共享内存等。其中，共享内存作用是数据重用，线程可以通过计算单元上的共享内存交换数据或者把一些常用的数据放进共享内存中减少对全局内存的读取次数，提高效率。

2） 节点架构

①GPGPU 与 CPU 互联架构

在 CPU-GPGPU 异构体系结构中，GPGPU 作为 CPU 的协处理器完成图形计算和通用计算，以外部设备的形式，通过 PCI-E 总线和 CPU 通信。常见的异构计算节点体系结构主要由四个部分组成：主存、多核处理器、I/O HUB 和 GPGPU 加速器，图 2 描述了这种体系结构互连关系。

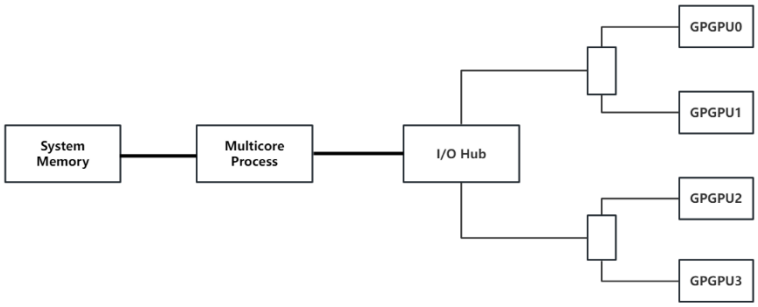

图 2 常见异构计算节点内部互联结构

主存（System Memory）与多核处理器之间通过 Memory Bus 互连，多核处理器借助 I/O HUB 链接多种外部设备，I/O HUB 通过 PCI-E 总线链接，以树状结构链接各种功能设备，在异构计算节点中，主要的设备就是 GPGPU 加速器，在 I/O HUB 与 GPGPU

加速器之间还会增加 PCI-E Switch，从而扩展 PCI-E 链路链接更多设备。由于部分 AI 应用需要大量的内存复制带宽，英伟达公司开发了 NVLink 技术能够比 PCI-E 提供更高的带宽，对大模型类应用有非常高的性能提升效果，目前芯片出口限制的一个很重要参数就是限制芯片传输速率不超过每秒 600GB，主要就是限制这部分的通道带宽。

②GPGPU 之间的互联

GPGPU 之间通过 PCI-E 互连并完成数据传输，在高级 GPGPU 加速器上，同时还支持互联协议。如今大模型训练的发展趋势就是参数量和神经网络复杂程度增加，往往需要多台服务器搭建集群协同工作进行分布式训练，而极大的参数量传输意味着需要 GPGPU 之间以及服务器之间进行数据传输交互。通过借助特定的互联结构和互联协议，使得 GPGPU 加速器可以获得高带宽、低延时的传输性能，并且可以支持 GPGPU 加速器之间的缓存一致性，从而实现共享显存。

3） 软件栈架构支持

基于 GPGPU 的 AI 芯片生态系统或软件层，是一种轻量级、模块化的软件开发环境，可以提供多种开发工具和运行时环境。同时也拥有丰富的系统关键功能组件支持，每个组件根据其功能自底向上构建出完整的平台功能，可适用于大规模应用程序计算、编译器及程序运行时组件开发。同时，还提供对 OpenCL、SYCL、OpenMP、CUDA 等框架的支持。

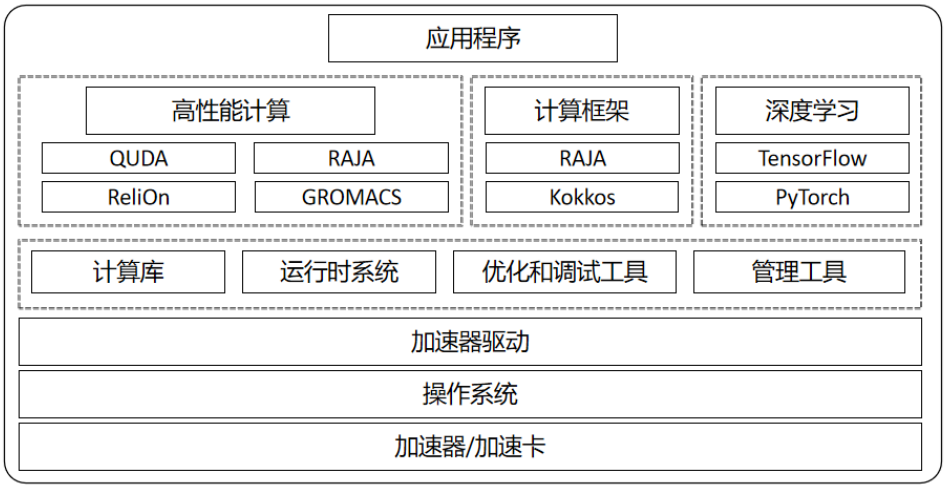

图 3 GPU 生态系统架构

①OpenCL（Open Computing Language，开放设计语言）

OpenCL 是为异构平台编程设计的框架，此异构平台可由 CPU、GPU、FPGA 或其他类型的处理器与硬件加速器组成。OpenCL 由两部分组成，一是用于编写 kernel 的语言，二是用于定义并控制平台的 API。OpenCL 提供了基于任务和基于数据两种并行计算机制，极大地扩展了 GPU 的应用范围，使其可用于处理信号转换、线性代数、分裂检查等没有太多依赖性且计算量大的问题。此外，OpenCL 还具有良好的代码可移植性，为某一个供应商平台编写的 OpenCL 应用程序正常情况下可以在其他供应商平台上运行。

②SYCL

SYCL 作为基于 C++的单源特定于域的嵌入式语言，是 OpenCL 的高级编程模型。它提高了各种加速设备上的编程效率（如高性能计算、机器学习、内嵌计算等），简单来说就是一种跨平台的抽象层，用户不需要关心底层的加速器是什么，按照标

准编写统一的代码就可以在各种平台上运行，大大提高了编写异构计算代码的可移植性和编程效率。SYCL 在目前用于 HPC、AI/ML、自动驾驶汽车、嵌入式或定制设备。

③CUDA

CUDA 是显卡厂商英伟达推出的平行运算平台，也是一种通用并行计算架构，该架构使 GPU 能够解决复杂的计算问题。它包含了 CUDA 指令集架构以及 GPU 内部的并行计算引擎。随着显卡的发展，GPU 越来越强大，而且 GPU 为显示图像做了优化，在计算上已经超越了通用的 CPU。如此强大的芯片如果只是作为显卡就太浪费了。因此英伟达（NVIDIA）推出 CUDA，让显卡可以用于图像计算以外的目的。

（2）技术优势

1）性能优势

GPGPU 具有高度的并行计算能力，支持矩阵计算、大规模数据处理等复杂计算，能够同时处理多个任务和数据，使其在金融行业处理大规模数据时具有快速和高效的优势。同时，GPGPU 支持高精度的计算和数值稳定性更好的算法，可以提高计算准确性，在风险评估和交易决策等场景中有重要意义。此外，使用 GPGPU 可以减少对 CPU 等其他硬件的需求，降低硬件成本。

综上所述，GPGPU 在金融行业中可以提高计算速度、实现更复杂的计算、提高精度和准确性，并且节省硬件成本，这些优势对金融行业的业务运作和决策制定都有着重要的影响。

2）支持场景

AI+大数据分析作为“发动机”，推动金融科技在交互模式变迁、获客留客、风控营销、安全运营等业务发展，从多个维度不断为行业发展注入活力。以下介绍几个 GPGPU 在金融行业的场景化应用。

①人工智能

金融机构正在越来越多地使用人工智能和机器学习技术来进行交易分析和风险管理，GPGPU 可提供预训练大模型、计算机视觉、自然语言处理、语音合成等大模型训练任务以及生物识别、语义识别等特定领域任务。

②GPU 虚拟化

传统的 GPU 虚拟化是指在一个物理 GPU 上创建多个虚拟 GPU 的过程。这种技术可以使多个用户同时使用同一个物理 GPU，从而提高利用率。更进一步，以 GPU 虚拟化为基础，融合了 GPU 共享，资源超分，远程调用，跨物理节点多合一的 GPU 池化技术，可以站在更高的层次解决整个数据中心内 GPU 利用率低、成本高、分配和管理困难的问题。该技术与软件定义网络、软件定义存储的原理类似，以软件定义 GPU 的方式对全局 GPU 进行抽象，软件化后形成一个统一的资源池，方便用户按需对 GPU 资源进行有效调用，无需关注实际物理 GPU 的大小，数量，型号以及安插的物理位置。该技术可以作为一个通用的技术底座，与各种云计算平台进行无缝整合，与各种 AI 异构芯片进行适配对接，形成一池多云、一池多芯的效果。

③大数据分析

大数据分析技术可用于金融行业的风险管理；对大量交易数据进行交易分析，以此预测市场走势和做出投资决策；检测欺诈模式，并预测潜在的欺诈行为；分析客户数据，了解客户需求和偏好，并据此制定更好的产品和服务策略。

④隐私计算

隐私计算可以帮助金融机构在保护客户隐私的同时满足合规性要求、实现数据共享、进行数据计算和分析、训练和测试模型并提高预测准确性，从而提高数据的利用价值。

其中可信执行环境（Trusted Execution Environment , TEE），基于芯片层面机密计算能力，构建运行时安全机制，将存储安全、传输安全与运行时安全想结合，打造基于国密的安全闭环能力。英伟达在 H100 已实现了机密计算能力，国内也已有支持国密的国产 CPU+国产 GPU+区块链的 TEE 方案，对标英特尔 Sgx+英伟达 H100 的模式，为金融机构 AI SaaS 化输出、远程算力安全使用、数据互联互通、内网纵深防御、计算外包等提供落地支持。

3）应用生态

使用 GPGPU 可以加速数据计算速度、降低计算成本、提高数据安全性、支持多种应用场景。以下是 GPGPU 在金融行业的一些应用场景。

①OCR 识别

金融业务中需要处理大量的交易数据和财务报表等文本信息，这些文本信息需要快速准确地识别和处理，利用 GPGPU 加速 OCR 识别可以并行处理 OCR 算法，提高计算精度和处理速

度，从而提高工作效率和准确性。同时，金融业务中需要处理来自不同语言的汇款单据、财务报表等文本信息，GPGPU 可以支持多语言 OCR 识别，更好地满足金融行业需求。

②NLP（自然语言处理）、ASR（语音识别）、TTS（语音合成）

GPGPU 在金融行业的自然语言处理（Natural Language Processing, NLP）方面应用可以加速算法的运行，从而更快地提取关键信息，帮助金融机构更好地了解市场和客户情绪，并做出更好的投资决策。GPGPU 在金融行业语音识别（Automatic Speech Recognition, ASR）方面的应用主要是通过利用 GPU 的并行计算能力来加速语音信号的处理和分析。例如，在交易处理场景，GPGPU 可以加速语音信号的处理和识别算法，更快地识别和处理交易数据；在客户服务场景，GPGPU 可以帮助金融机构加速语音识别算法，更快地识别客户的问题和需求，并提供更好的服务；在身份验证场景，语音识别技术可以用于客户的声纹识别，以识别客户的身份，GPGPU 可以加速语音信号的处理和分析，更快地进行声纹识别。

GPGPU 在金融行业的语音合成（Text To Speech, TTS）方面，主要是通过利用 GPU 的并行计算能力来加速语音合成的过程。例如，在语音报价场景，语音合成技术可以用于生成市场报价的语音信息，大语言模型（Large Language Model, LLM）可接受输入的文本数据，将其转化为自然流畅的人类语音，GPGPU 可以加速语音合成算法，更快地生成市场报价的语音信息；在自动化客户服务场景，LLM 模型支持风格和音色控制，用户可以根

据需要选择不同的音色、声音和说话风格，有助于提高客户满意度，GPGPU 可以帮助金融机构加速语音合成算法，从而更快地生成回复客户请求的语音信息；在智能客服机器人场景，LLM 模型可生成高质量的合成音频，使合成的语音听起来更加自然、清晰和流利，GPGPU 可以加速语音合成算法，更快、更自然地生成机器人的语音输出。

2.1.2 NPU 路线

神经网络处理单元（Neural-network Process Units, NPU）采用 ASIC 技术，通过硬件模拟神经网络的方式克服了 CPU、GPU 在深度学习设计上的先天不足，从而大幅提高了深度学习芯片的运算速度。虽然目前 NPU 主要集中于推理芯片领域，但已经对 GPU 在人工智能领域的地位产生了冲击。NPU 的问世标志着人工智能芯片开始向定制专用化方向发展。

（1） 技术实现

NPU 是专门用于硬件加速人工智能应用的微处理器或计算系统，采用 DSA 的专用技术，特别用于人工神经网络、机器视觉和机器学习等领域。典型应用包括机器人学习、物联网等数据密集型应用或传感器驱动任务。在硬件架构设计时，NPU 专注于人工智能设计，可以通过硬件指令在一个周期内完成 3D Cube、Vector 向量、Scalar 标量的计算。相较于通用处理器，其算力与数据吞吐量之比有数百倍提升，同时功耗维持在较低水平。NPU 主要包括计算单元和数据存储。**计算单元**是核心部件，通常采用矩阵计算、向量计算等方式，可以快速地执行矩阵乘法、卷积等

计算。**数据存储**是另一个关键组成部分。由于神经网络模型通常非常庞大，因此 NPU 需要具备足够的存储容量来存储模型参数和中间计算结果。NPU 的数据存储通常采用高速缓存和显存的结合方式，以便更快地存取和读取数据。

1）**整体架构**[1]

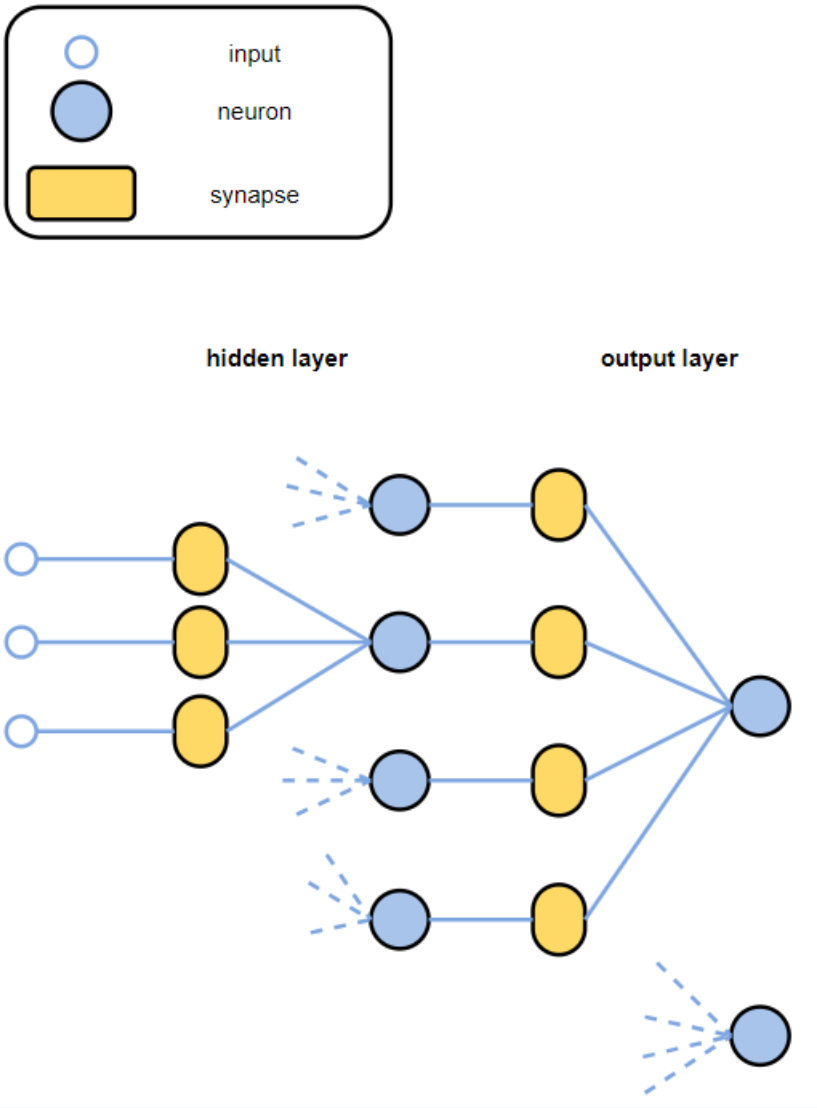

图 4 神经网络模式图

图 4 是神经网络模式图，基于神经网络的人工智能算法，成

[1] 参考来源：Zichao Wang. NPU Development Overview. Scientific Journal of Economics and Management Research, Volume 2 Issue 07, 2020: 133-138.

功模拟了人类大脑内部神经元的结构。图中的 neuron 代表的就是单个神经元，synapse 代表神经元的突触，hidden layer 是神经网络中的隐含层，output layer 是输出层，input 是神经网络输入。图 5 为 NPU 处理器的内部结构。

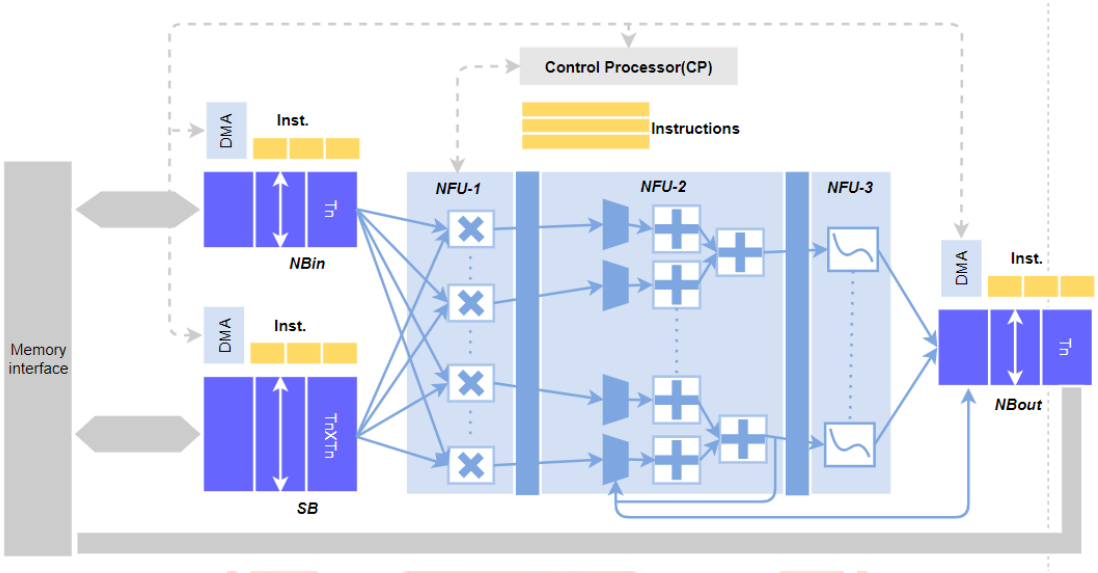

图 5 NPU 内部结构

蓝色区域是硬件逻辑模拟的神经网络结构，称为 NFU(Neural Functional Units)。从左到右分为三个部分，NFU-1、NFU-2、NFU-3。NFU-1 是乘法单元，有多个乘法器; NFU-2 是加法树，有多个加法树，每个加法树由多个加法器组成; NFU-3 是激活单元，有多个激活单元。总的来说 NFU 将资源分为若干份，每一份包括 NFU-1 的多个乘法器，NFU-2 的一个加法树和 NFU-3 的一个激活函数运算器，运算时一份资源中乘法器同时运行输出多个结果，送加法树，加法树运算后得出一个结果送激活函数，激活函数运算判断是否激活。除此之外还有三个缓冲区，一个存储输入的数据(NBin)，一个存储运算的权值(SB)，一个存储结果(NBout)。使

用 NPU 架构的深度学习神经网络芯片性能得到了很大的提高，运算速度远超 GPU 和 CPU。

2）软件栈架构支持

NPU 的软件栈提供了控制和访问 NPU 硬件的接口，支持神经网络框架的集成，以及模型转换和优化工具，还包括调试和性能分析功能，帮助开发者更好地利用 NPU 的计算能力，快速部署和优化神经网络模型，并进行调试和性能优化。目前，通过中间表示（Intermediate Representation, IR）模型构建可以支持 TNN、MNN、Paddle-Lite 等深度学习框架。

IR 模型构建主要通过 3 项关键技术实现。一是子图分割，即找出 NPU 不支持的算子，以不支持算子为界将模型切分为多个模型；二是算子转换（桥接，bridge），即取出 IR 模型的超参数和参数，填入 NPU 算子生成器；三是离线调用 IR 生成专用模型，即将多个算子整合为模型并优化。[2]

①TNN

TNN 最新版本支持服务端 X86 和 NVIDIA 硬件，采用了 OpenVINO 和 TensorRT 的集成方式，能够快速获取硬件厂商的最新优化成果，并添加自定义实现来达到性能极致。考虑到桌面端应用的安装包大小限制，TNN 通过 JIT 和手工优化的方式实现了轻量级的 X86 后端，整体库大小仅为 5MB 左右。此外，TNN 还扩展了对 CV 类模型的支持，包括 3D-CNN、LSTM 和 BERT 等模型结构，并新增了 19 个算子，例如 LSTM、GridSample、

[2] 参考来源：主流手机 NPU 软件栈调研（2021 Q2），https://zhuanlan.zhihu.com/p/380317994。

Histogram 等，总共支持 107 个算子。[3]

②MNN

MNN 是一种轻量级的深度学习端侧推理引擎，旨在解决深度神经网络模型在端侧推理运行时的问题，包括深度神经网络模型的优化、转换和推理。它支持主流模型格式，如 TensorFlow、Caffe、ONNX 等，涵盖了常用网络，如 CNN、RNN 和 GAN 等。此外，MNN 还提供转换、可视化和调试工具，并且能够方便地部署到移动设备和各种嵌入式设备中。[4]

③Paddle-Lite

Paddle Lite 是一个高性能、轻量级、具有灵活性和易于扩展的深度学习推理框架，旨在支持包括移动端、嵌入式设备和边缘端在内的多种硬件平台。支持 PaddlePaddle、TensorFlow、Caffe、ONNX 模型的推理部署。[5][6]

（2） 技术优势

1）性能优势

高效性。NPU 的设计目的就是为了进行深度学习计算，因此具有非常高的计算效率和能耗效率，能够在短时间内完成大规模的神经网络计算任务。

低延迟。由于 NPU 的设计中将计算和存储密集的任务分离开来，避免了 CPU 和 GPU 中计算和存储竞争的问题，因此 NPU

[3] 参考来源：腾讯发布推理框架 TNN 全平台版本，同时支持移动端、桌面端和服务端，https://cloud.tencent.com/developer/article/1819147?areaSource=102001.13&traceId=HDLu9TPBTtI6eFfa4tJTw

[4] 参考来源：AI 平台-MNN【推理引擎】，https://developer.aliyun.com/article/701406。

[5] 参考来源 1：PaddlePaddle/Paddle-Lite: PaddlePaddle High Performance Deep Learning Inference Engine for Mobile and Edge (飞桨高性能深度学习端侧推理引擎)，https://github.com/PaddlePaddle/Paddle-Lite。

6 参考来源 2：Paddle-Lite 文档，https://paddlepaddle.github.io/Paddle-Lite/。

可以在很短的时间内响应计算请求，执行计算任务。此外，NPU 的计算单元也被优化，采用了更加高效的矩阵计算和向量计算方式，可以快速地执行大规模矩阵乘法、卷积等操作。

稳定性。由于 NPU 的计算单元和数据存储经过精心设计和测试，可以在长时间的运行中保持高效、稳定的性能，因此 NPU 通常具有很好的容错性和可靠性，即使在高负载、复杂计算任务的情况下，也能够保持稳定的计算性能。

可编程性。与传统的 ASIC 芯片不同，NPU 通常具有一定的可编程性，可以通过软件调整参数和配置来适应不同的计算任务。这意味着 NPU 不仅适用于固定的深度学习模型，还可以适应不同的算法、框架和数据集。

2）支持场景

NPU 拥有对矩阵和卷积运算的高效性，主要得益于其架构所采用的定制化硬件设计，专门用来加速矩阵运算。同时 NPU 架构采用了较多的分布式缓存与计算单元进行配合，能有效减少数据到计算单元的搬运，在提升计算力的同时还减小了数据传输的功耗，使得 NPU 在深度卷积神经网络这一应用领域具有一定优势，可以广泛应用于各类 AI 领域。例如，计算机视觉（CV）：高清视频内容处理、智能驾驶辅助、无人机和机器人等；自然语言处理（NLP）：机器翻译、文本语义理解、文本分类等；人工智能内容生成（AIGC）：以文生图，文章生成、人机对话、智能搜索等；语音识别（ASR）、语音合成（TTS）、基于用户画像的内容推荐、广告推荐等。

①手机 AI 的核心载体

NPU 是专门负责实现 AI 运算和应用的处理器，相较于 CPU 和 GPU，实现数量级提升，具有更优异的能效。它为场景优化器提供了动力，使得识别照片中的内容更加准确，并能够提示相机调整为适合主体的理想设置。NPU 还可以模糊自拍照的背景并创建散景效果，利用 AI 场景识别来拍摄照片，并进行运算修图。此外，NPU 还能判断光源和暗光细节合成超级夜景。

②自动驾驶

NPU 能够快速地处理大量的传感器数据，在自动驾驶系统中扮演着至关重要的角色。例如，NPU 可以用于处理摄像头、雷达、激光雷达等传感器数据，从而实现对车辆的实时控制和决策。

③人脸识别

人脸识别是一个需要高效计算的应用场景，NPU 可以为人脸识别系统提供快速的计算能力，支持人脸检测、特征提取和匹配等操作，并能够快速处理大量的图像数据，从而实现快速准确的人脸识别。

④智能语音

智能语音是一个需要实时高效计算的应用场景，NPU 可以为智能语音系统提供高效的计算能力，支持语音识别、语音合成等操作。智能语音系统能够通过 NPU 快速地处理用户的语音指令，实现对音箱的实时控制和管理。

⑤视频监控

视频监控是一个需要处理大量视频数据的应用场景，NPU 可

以为视频监控系统提供高效的计算能力，支持视频数据的处理、分析和识别等操作。

⑥金融应用

NPU 的引入使 AI 建模的时效从小时级提升到了分钟级，为金融企业大幅降低了 AI 开发和运维门槛，为智能风控、智慧营销等场景智能化提供了有力支撑。

3）未来展望

多样化。随着人工智能应用场景的不断扩展和多样化，NPU 需要不断适应不同的应用场景和算法模型。未来，NPU 的设计和开发将更加注重多样性，以支持更广泛的应用场景和算法模型。

集成化。目前，许多 NPU 需要与其他芯片（如 CPU、GPU 等）配合使用才能完成整个计算任务。未来，NPU 的设计和开发将更加注重集成化，以使 NPU 能够更加自主地完成计算任务，减少对其他芯片的依赖。

模块化。NPU 可以根据不同应用场景和算法模型的需要，选择不同的计算模块来进行计算，从而提高计算效率和灵活性。

高性能。随着人工智能应用场景的不断拓展和算法模型的不断复杂化，NPU 需要具备更高的计算能力和更快的响应速度，以适应未来的发展需要。

2.1.3 路线对比

GPGPU 架构是通用图形处理器，适用于需要进行大量并行计算的任务，如科学计算、计算机视觉和机器学习等。GPGPU 架构不仅兼顾硬件性能，还与软件生态兼容，可通过 AI 软件生态

获得更多的优势，而无需自研 AI 软件工具。GPGPU 具有强大的计算能力和更为成熟的编程框架。虽然 GPU 在并行计算能力方面具有优势，但不能单独工作，需要 CPU 的协同处理。此外，高性能的 GPU 也存在着功耗高、体积大等问题，对于一些小型设备和移动设备来说可能会存在使用上的限制。

NPU 是一种专用芯片，具有小体积、低功耗、高计算性能和高计算效率等特点。它是针对特定范围的 AI 算子设计的 Tensor 和 ASIC 加速单元，没有明确的领域定义。NPU 是受生物神经网络启发而构建的，相比于 CPU 和 GPU 处理器需要使用数千条指令完成的神经元处理，NPU 只需要一条或几条指令就能完成，因此在深度学习的处理效率方面具有明显的优势。

NPU 与 GPU 设计思路不同。GPU 考虑到计算的通用性，在提升算力的同时要考虑到数据吞吐量的提升，NPU 针对特定领域设计，无需考虑通用应用对于内存带宽的需求。但是，与通用性强的 CPU 和 GPU 相比，NPU 的软件生态兼容度较差，需要自研 AI 软件工具和算法模型。它不适合用于大量样本的训练，更适合于预测和推理等场景。

2.2 开发平台

国际上主流的并行计算平台和编程模型以英伟达公司的 CUDA 平台和 AMD 公司的 ROCm 平台（Radeon Open Computing platform）为主。其中英伟达公司的用户数量具有较大优势。

国内主要分为 GPGPU 路线和 NPU 路线。其中 GPGPU 路线大都在以 CUDA 为标准开发相关的异构并行计算平台，主要有

中科海光、天数智芯、摩尔线程等。以中科海光 DCU 为例，与 CUDA 和 ROCm 的对比如下。

表格 2 开发平台对比

软件生态	英伟达	AMD	海光 DCU
并行计算平台	CUDA	ROCm	DTK
框架支持	TensorFlow、PyTorch、PaddlePaddle 等全系列	TensorFlow、PyTorch、PaddlePaddle 等全系列	TensorFlow、PyTorch、PaddlePaddle 等全系列
推理	TensorRT	MIGraphX	MIGraphX
训练	全精度支持	全精度支持	全精度支持
ONNX Runtime	支持	支持	支持
通信库	NCCL	RCCL	RCCL
算力支持	矢量计算、矩阵计算/Tensor Core 张量计算核心	矢量计算、矩阵计算	矢量计算、矩阵计算
算法模型识别精度	主流使用英伟达 GPU 进行模型的开发。	与英伟达卡一致，无差别	与英伟达卡一致，无差别

2.3 算力服务

在 AI 场景下，从算力供给的角度看，还存在以下问题。

一是存在算力孤岛。算力供给侧通常会基于多个厂商的多类 AI 服务器进行建设，不同厂商不同 AI 芯片架构的服务器资源互相独立，生态隔离，形成算力孤岛。

二是算力资源配置不均。运维人员在资源规划时需要结合多

种硬件上预估的应用规模进行采购，很难实现精准配比，同时由于上层应用与底层硬件紧绑定关系，后期也较难实现迁移调整，因此可能会出现某些厂商的硬件资源不够而另外一些厂商的硬件资源闲置的情况。

三是算力资源利用率低。当前 AI 芯片虚拟化能力存在局限性，物理资源只能以独占式的分配方法提供给用户实例使用，无法实现动态调整和灵活调度，导致底层资源无法被充分利用。

综上，如何将金融机构内广泛部署的异构多元算力资源与多种专业算法模型间进行有效协同，驱使业务应用能平滑的在各级算力资源上进行流转运行，充分利用巨量算力资源，是金融业数字化转型并实现业务创新的关键点之一。

算力服务层的目标是深度适配多元异构算力资源，形成多厂商、多架构的异构智能算力混合资源池，实现从传统的以硬件资源为单位、静态分配使用算力的方式，转变为以计算能力为单位对算力资源进行动态、灵活地配给，应用无需关注智能算力的位置、数量和类型。**算力服务层包括算力池化、算力调度和算力运维。**

2.3.1 算力池化

算力池化通过构建底层异构硬件的统一抽象模型，并对应用调用底层算力资源的请求进行重定向和再调度，从而实现各类硬件资源的一体池化。同时为应对智算业务的潮汐效应，算力池化可根据业务需求及算力负载情况提供算力资源弹性扩缩容的能力。池化能力需包括：①算力与显存的细颗粒度切分，这是最基

本的功能要素。②远程调用，即 CPU 与 GPU 分离解耦，在一台普通 CPU 机器上部署 AI 任务，先调用本机 CPU 进行数据预处理，再通过网络远程调用 GPU 进行加速。③资源聚合，把多机多卡快速聚合给到一个任务，免去复杂的调度过程与模型拆分过程，快速交付。④算力超分，允许算力超分，通过资源超分实现多种业务场景的 GPU 资源共享与复用。⑤虚拟显存，也叫显存扩展，调用内存补显存。⑥随需应变，资源动态伸缩，无需重启虚机/容器，所有虚拟 GPU 资源的分配与回收都是动态的，而且可以按需调整大小，无需重启。

2.3.2 算力调度

由于上述池化能力中，已将算力进行抽象，同时具备了 CPU 与 GPU 分离解耦的远程调用 GPU 的能力，因此，需要通过有效算力调度，提高算力芯片的利用率，降低算力闲置，更好地管理人工智能算力的使用情况；对用户行为进行分析和监管，优化算力设备布局规划，提升业务部署效能，使应用具有更好的稳定性和扩展性。算力调度包括：①紧凑型、均衡型等智能化自动化调度策略，目的在于让算力分布更加科学合理，避免算力碎片或算力热点。②允许特殊指定节点，指定 AI 芯片型号。③有效纳管多种异构算力芯片，并智能化匹配多种 AI 业务需求，实现自动调度。④池化调度自带任务级别的排队功能。⑤对排队任务进行优先级设定，允许高优先级任务插队。⑥双资源池，即支持 GPU 芯片在物理 GPU 与虚拟 GPU 之间切换，两者同时存在，同时管理。

2.3.3 算力运维

使用算力池化技术，可以实现 AI 算力资源的在线秒级分配和回收，结合丰富的管理功能，大大降低运维难度，提高运维人员的工作效率。运维管理需包括：①组件高可用，支持管理节点多副本，无单点故障。②节点管理，支持图形化显示 AI 芯片资源状态，资源利用率等信息。③GPU 热插拔，资源池平滑扩（缩）容，应用无感知的前提下添加和删除 GPU 节点，或者物理 GPU 卡。④动态回收，分配出去的资源在闲时能够自动回收，一方面提供利用率，避免独占，另一方面可以降低管理难度。⑤灰度升级，平滑升级，不影响业务。⑥能实现多租户管理，并对其进行配额管理。⑦能实现 AI 芯片的健康监控，包括温度、风扇、时间、任务、负载等多种维度。⑧配套运维工具、日志工具与监控工具。

3 产业分析

3.1 产业概览

人工智能芯片产业链涵盖从芯片设计、芯片制造到人工智能芯片应用等多个环节。其中芯片设计是芯片的“灵魂”，决定了芯片将能够实现什么样的功能，也是进行芯片制造的前提和基础；芯片制造是晶圆加工厂商根据设计要求生产芯片的核心环节，也是整个芯片产业链上对综合技术能力要求最高的一步。

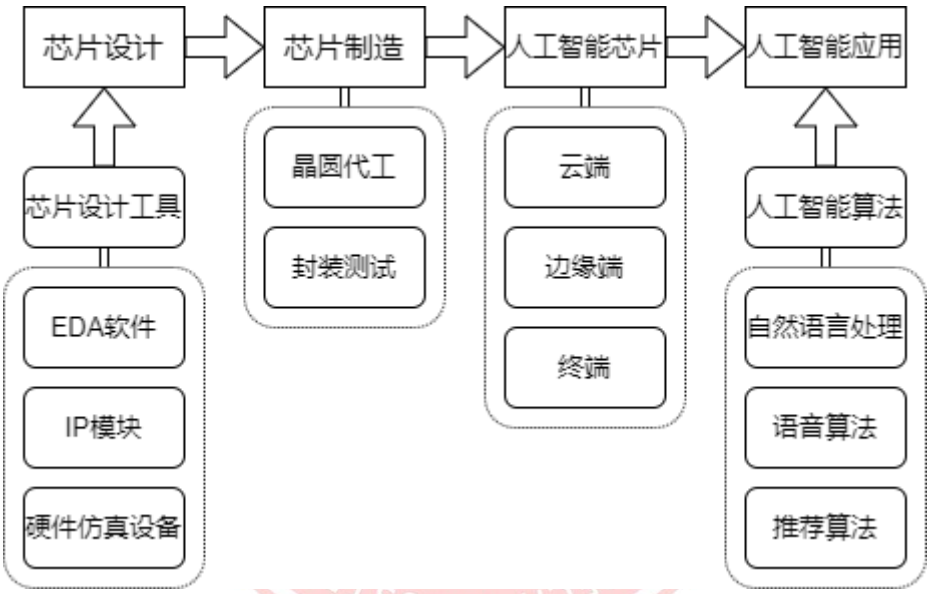

图 6 芯片产业链

芯片设计：人工智能芯片行业的核心环节之一，包括利用 EDA 工具进行系统设计、IP 模块授权应用、仿真模拟验证等过程。只有经过了严谨的芯片设计过程，才有机会生产出包含预期功能、满足预期性能的人工智能芯片。国际知名的人工智能芯片设计公司有英伟达、英特尔、AMD、三星、谷歌等。国内芯片设计公司在近年也实现了快速发展，在各细分领域取得了不错的成绩，并在国际上也开始崭露头角，如海光、华为海思等。

芯片制造：人工智能芯片行业的核心环节，为人工智能芯片行业提供产业支撑，主要包括晶圆代工和封装测试两部分。从晶圆到芯片需要经历晶圆加工、氧化、光刻等一系列复杂流程，有着极高的资本和技术壁垒，先进制程占比逐年提高。国际上晶圆代工的代表企业包括台积电、三星、格芯等，国内对于晶圆代工的重视程度也在不断提高，处于追赶阶段，代表企业有中芯国际、华虹半导体等。

人工智能芯片：算力需求大大促进了人工智能芯片市场的发展，是人工智能能够持续演进和提升的基石。从数据中心云端训练和推理任务负载，再到边缘端和终端的推理应用都离不开人工智能芯片的使用。目前国内人工智能芯片的使用仍然以国际厂商的产品为主。近年来，国内人工智能芯片公司生态体系初步完善，涌现了一大批创业公司，产品成熟度不断提高，如海光、寒武纪、燧原、壁仞科技、登临科技等。

人工智能应用：随着人工智能算法和芯片的不断发展，人工智能应用的广度和深度也不断丰富，如金融、云计算、消费电子、各类智慧场景业务等，算法准确性和效率提升明显。市场对自动驾驶热点的持续关注，也使其正在成为人工智能应用的重要发展方向。此外，国家政策也在持续推动人工智能芯片在保障民生等公共领域的应用落地。

人工智能算法：人工智能通过机器模拟人类意识，学习人类思维方法处理问题，算法是人工智能的灵魂，也是实现批量化解决人工智能问题的核心手段。当前对于人工智能算法的研究及应

用的主流方向主要为计算机视觉和自然语言处理。计算机视觉算法常见的有 ResNet50、YoloV3、SegNet 等，涉及图像分类、目标检测、图像分割等多种应用场景。自然语言处理算法的目标是将自然语言转换为计算机可识别的指令，搭建机器与人沟通的桥梁，常见的算法有著名的 Transformer、Bert、LSTM 等。人工智能算法研究机构主要为各大高校和研究机构，以及世界知名互联网企业和 AI 算法公司，如谷歌、亚马逊等，国内的主要有百度、字节跳动、科大讯飞、商汤科技、旷视科技等。

3.2 国际情况

3.2.1 政策层面

人工智能领域现已成为世界科技创新的关键领域，对数字经济发展具有重要意义。自 2016 年起，先后有 40 余个国家和地区把推动人工智能发展上升到国家战略级高度。近年来，特别是新冠疫情的冲击下，越来越多的国家认识到，人工智能对于提升全球竞争力具有关键作用，进一步加强对人工智能领域的攻关及应用力度。

美国：2022 年 9 月签署《芯片和科学法案》（以下简称芯片法案），计划为美国半导体产业提供高达 527 亿美元的政府补贴，鼓励企业在美国研发和制造芯片，在未来几年提供约 2000 亿美元的科研经费支持，重点支持人工智能、机器人技术、量子计算等前沿科技。美国在芯片法案中加入“中国护栏”条款，禁止获得联邦资金的公司在中国大幅增产先进制程芯片，期限 10 年，旨在使各国的芯片产业转移到美国，同时限制中国半导体芯片产业

发展。

欧盟：2022 年 2 月推出《欧盟芯片法案》，该法案基本上是美国版法案的框架，旨在加强欧盟芯片生态系统的技术能力和创新并确保欧洲在芯片研究和创新方面处于领先地位。向半导体行业投入超过 430 亿欧元公共和私有资金。用于加强现有研究、开发和创新，以确保部署先进的半导体工具以及用于原型设计、测试的试验生产线等。目标是在 2030 年将欧洲半导体市场份额从当前的 10%提升至 20%。

日本：2022 年 3 月实施“半导体援助法”“特定高度情报通信技术活用系统开发供给导入促进法（5G 法）”，为持续提供半导体供应、在国内兴建半导体的企业提供补助。日本 2021 财年预算修正案显示，约在半导体行业投入 7740 亿日元（约合人民币 423 亿元）。提出 5 年时间赶上新一代产品的尖端制造技术，但不会在规模上追赶台积电和三星。

韩国：2021 年 5 月实施“K 半导体战略”，旨在扩大扶持力度，本土企业自强。未来十年，包括三星电子和 SK 海力士在内的 153 家企业将在本土半导体业务上投入 4510 亿美元。对研发和设施投资将分别减免 40%至 50%和 10%至 20%的税金。在 2021 年下半年至 2024 年期间，将对从事半导体等“关键战略技术”的大型企业的资本支出税收优惠从目前的最高 3%提高到 6%。计划在 2030 年将韩国发展成为综合半导体强国，主导全球供应链。

3.2.2 产业机构

（1） 英伟达

英伟达是人工智能主力芯片供应商，是 GPGPU 领域的龙头企业，技术上保持着绝对领先，平均每两年推出新一代芯片架构，每代产品性能始终能够保持稳定的提升和强大的产品竞争力。目前，英伟达的产品矩阵已覆盖数据中心、专业图形图像、消费级游戏和汽车等多业务场景。其中，面向数据中心场景，英伟达先后发布了 V100、A100 以及基于最新 Hopper 架构的 H100 芯片，结合成熟的 CUDA 生态以及不断迭代发布的 NVLink 互联技术，构成了英伟达技术和市场领先的“护城河”，可以提供当前世界上最强大的 AI 算力产品，引领 GPGPU 领域的技术发展方向。同时，英伟达还通过 NVIDIA AI Enterprise、NVIDIA DGX Cloud 等服务和解决方案，进一步释放 AI 潜力，占据主流市场。

1）主要产品

英伟达数据中心旗舰 GPU 产品长期占据市场主导地位，主要产品包括 V100、A100 以及 H100。

V100：基于英伟达先进的 Volta 架构，是英伟达在数据中心和 AI 计算领域取得成功过程中，里程碑式的重要产品。Volta 架构加持下，V100 不仅在计算单元数量上有显著提升，还搭载了 HBM2 显存，提供更高的显存带宽，并通过 NVLink 2.0 技术，实现了更高效的集群拓展能力。最重要的是，V100 上首次集成了第一代 Tensor Core 单元，利用混合精度加速深度学习重点矩阵和卷积运算，大幅度提高了深度学习的训练和推理效率。

A100：当前市场最主流需求的英伟达数据中心 GPU 产品，

也是 V100 停产后的升级替代款型。A100 基于英伟达 Ampere 架构，采用 7nm 制程，与 V100 相比，计算单元数量进一步增加，且支持 NVLink 3.0 技术。A100 还采用了第三代 Tensor Core 单元，支持 TF32、BF16 等更多数据格式，并支持通过稀疏化方式进一步提升深度学习计算效率。新增的 MIG 技术，可以实现 GPU 资源的硬件隔离，满足在云计算环境下资源的灵活分配需求。

H100：英伟达最新 Hopper 架构下的旗舰产品，采用台积电 4N 工艺制程，计算单元数量相比于 A100 实现翻倍增长。第四代 Tensor Core 单元不仅在相同时钟频率下实现性能 2 倍提升，还增加了 Transformer 引擎，通过硬件实现对 GPT 等 Transformer 类深度学习算法模型计算速度的显著提升，进一步释放 AI 算力。此外，H100 还支持 MIG 2.0、机密计算等新特性，是英伟达目前为止性能最强大的 GPGPU 产品。

2）开发平台

英伟达开发平台 CUDA 是在统一架构 GPU 上面向通用计算的并行编程开发平台。CUDA 开发平台中包含了不同层次的调用接口，其中包括一组运行时 API、一组设备驱动 API 以及 CUDA 提供的函数库 API。

CUDA 驱动 API 能够直接控制底层硬件结构，CUDA 运行时 API 则是对驱动 API 的封装。程序设计过程中既可以调用 CUDA 驱动 API 实现对硬件更高效的控制，也可以使用 CUDA 运行时 API 更便捷、更直观地实现 CUDA 开发平台提供的计算模式。

开发库是基于 CUDA 技术所提供的应用开发库。其中，CUDA 包含了两个重要的标准数学运算库——CUDA 快速傅立叶变换（CUDA Fast Fourier Transform, cuFFT）和基础线性代数子程序库（CUDA Basic Linear Algebra Subprograms, cuBLAS）。这两个数学运算库所解决的是典型的大规模的并行计算问题，也是在密集数据计算中常见的计算类型。

运行时环境提供了应用开发接口和运行时组件，包括基本数据类型的定义和各类计算、类型转换、内存管理、设备访问和执行调度等函数。基于 CUDA 开发的程序代码在实际执行中分为两种，一种是运行在 CPU 上的宿主代码（Host Code），一种是运行在 GPU 上的设备代码（Device Code）。不同类型的代码由于其运行的物理位置不同，能够访问到的资源不同，因此对应的运行期组件也分为公共组件、宿主组件和设备组件三个部分，基本囊括了所有在 GPGPU 开发中所需要的功能和能够使用到的资源接口，开发人员可以通过运行时环境的编程接口实现各种类型的计算。CUDA 提供的运行时环境通过驱动来实现各种功能。

CUDA 支持 Windows、Linux、MacOS 三种主流操作系统，CUDA C 语言和 OpenCL 及 CUDA Fortran 语言。无论使用何种语言或接口，指令最终都会被驱动程序转换成并行线程执行（Parallel Thread Execution, PTX, CUDA 架构中的指令集，类似于汇编语言）代码，交由显示核心计算。

（2） AMD

AMD 是高性能计算（High Performance Computing, HPC）主

力芯片供应商，全球领先的半导体技术提供商，产品覆盖 GPU、APU（Accelerated Processing Unit）及 FPGA 等多个领域。AMD EPYC（霄龙）处理器面向云计算、HPC 等高性能计算工作场景，凭借多核优势以及优异性能，在服务器领域始终保持着较高的市场份额。同时，AMD 也是少数可以和英伟达可以在全球范围内 GPGPU 领域展开竞争的企业，AMD Instinct 系列加速器结合 ROCm 生态，可以满足 Exascale 级（百亿亿次级）工作负载需求，加速大规模 HPC 和 AI 训练任务。

1）主要产品

当前 AMD 面向数据中心和 AI 计算场景的 GPU 产品主要为 Instinct MI200 系列，基于 CDNA2 架构。MI200 系列共包含 3 款产品，分别为 MI250X、MI250 以及 MI210。

MI250X 整卡采用 OAM（开放加速模组）形态设计，是首个采用 MCM（多芯片模组）封装，也是首个支持 128GB HBM2e 超大显存的 GPU。MI250X 通过第二代矩阵核心进一步加速 FP64 和 FP32 矩阵运算，可以提供 383 TFLOPS 的 FP16 理论性能，是第一款百亿亿次级的加速卡产品。MI250 则是 MI250X 的精简款型，屏蔽部分核心计算单元，各项性能指标小幅下调，其余参数规格完全不变。MI210 为 PCIe 扩展卡形态，同样基于 CDNA2 架构，采用单芯片设计。相较于 MI250X/250，MI210 性能有所下降，但 PCIe 形态扩展了 MI200 系列的应用场景，降低了使用条件，且 MI210 拥有基于 AMD Matrix Core 技术的广泛混合精度能力，可以为加速深度学习训练提供一个强大解决方案。

2）开发平台

AMD 的开发平台 Radeon 开放计算平台（Radeon Open Computing platform, ROCm）是基于开源项目的 AMD GPU 计算生态。ROCm 之于 AMD GPU 相当于 CUDA 之于 NVIDIA GPU。

3.3 国内情况

3.3.1 政策层面

近年来，我国采取了一系列政策推进集成电路和人工智能产业的发展，包括提升集成电路设计水平、发展高端芯片和人工智能技术、促进新型基础设施建设等。此外，还有税收优惠政策和标准化管理措施。未来，我国将继续加快基础设施建设，并聚焦于高端芯片、人工智能算法等领域的研发与应用突破，同时加强前沿技术如量子计算和神经网络芯片的布局。

表格 3 我国人工智能芯片重点政策汇总[7]

时间	政策	政策内容
2015.05	《中国制造 2025》	将“集成电路及专用设备”作为新一代信息技术产业的重点突破口列在需要大力推动的重点领域之首。提出着力提升集成电路设计水平,不断丰富知识产权(IP)核和设计工具,突破关系国家信息与网络安全及电子整机产业发展的核心通用芯片,提升国产芯片的应用适配能力的发展要求。
2015.06	《“互联网+”行动指导意见》	支持发展核心芯片、高端服务器研发和云计算、大数据应用。

[7] 参考来源：我国人工智能芯片 PEST 分析：政策利好技术突破带动行业发展

2016.05	《“互联网+”人工智能三年行动实施方案》	对人工智能芯片发展方案提出多项要求,并促进智能终端、可穿戴设备的推广落地。
2016.08	《“十三五”国家科技创新规划》	国家科技重大专项包括多个涉及芯片设计、制造的重大专项，要求整体创新能力进入世界先进行列;面向 2030 年体现国家战略意图的重大科技项目中，类脑计算的开发是重点之一;新一代信息技术里重点提到了微纳电子与系统集成技术、高性能计算和人工智能等技术。
2016.07	《国家信息化发展战略纲要》	制定国家信息领域核心技术设备发展战略纲要,以体系化思维弥补单点弱势,打造国际先进、安全可控的核心技术体系，带动集成电路、基础软件、核心元器件等薄弱环节实现根本性突破。
2016.11	《产业技术创新能力发展规划 2016-2020 年)》	着力提升集成电路设计水平,发展高端芯片,不断丰富知识产权 IP 核和设计工具,推动先进制造和特色制造工艺发展,提升封装测试产业的发展水平，形成关键制造装备和关键材料供货能力,加紧布局超越摩尔相关领域。
2016.12	《“十三五”国家战略性新兴产业发展规划》	提升核心基础硬件供给能力。提升关键芯片设计水平,发展面向新应用的芯片。加强类脑芯片、超导芯片、石墨烯存储、非易失存储、忆阻器等新原理组件研发,推进后摩尔定律时代微电子技术开发与应用。

2017.07	《新一代人工智能发展规划》	到 2020 年人工智能总体技术和应用与世界先进水平同步,人工智能产业成为新的重要经济增长点,人工智能技术应用成为改善民生的新途径。
2017.12	《促进新一代人工智能产业发展三年行动计划(2018-2020)》	从推动产业发展角度出发,结合“中国制造2025”,对《新一代人工智能发展规划》相关任务进行细化和落实,以信息技术与制造技术深度融合为主线,以新一代人工智能技术的产业化和集成应用为重点,推动人工智能与实体经济相融合。
2017.05	《大数据产业发展规划(2016-2020 年)》	结合行业应用，研发大数据分析、理解、预测及决策支持与知识服务等智能数据应用技术。突破面向大数据的新型计算、存储、传感、通信等芯片及融合架构、内存计算、亿级并发、EB 级存储、绿色计算等技术,推动软硬件协同发展。
2018.01	《人工智能标准化白皮书(2018 版)》	宣布成立国家人工智能标准化总体组、专家咨询组,负责全面统筹规划和协调管理我国人工智能标准化工作。
2018.08	《扩大和升级信息消费三年行动计划(2018-2020 年)》	利用物联网、大数据、云计算、人工智能等技术推动电子产品智能化升级,提升手机、计算机、彩色电视机、音响等各类终端产品的中高端供给体系质量,推进智能可穿戴设备、虚/增强现实、超高清终端设备、消费类无人机等产品的研发及产业化。

2018.03	《关于集成电路生产企业有关企业所得税政策问题的通知》	对满足要求的集成电路生产企业实行税收优惠减免政策,符合条件的集成电路生产企业可享受前五年免征企业所得税,第六年至第十年按照 25%的法定税率减半征收企业所得税,并享受至期满为止的优惠政策。
2019.05	《关于集成电路设计和软件产业企业所得税政策的公告》	依法成立且符合条件的集成电路设计企业和软件企业,在 2018 年 12 月 31 日前自获利年度起计算优惠期,第一年至第二年免征企业所得税,第三年至第五年按照 25%的法定税率减半征收企业所得税,并享受至期满为止。
2019.11	《工业和信息化部关于加快培育共享制造新模式新业态促进制造业高质量发展的指导意见》	推动新型基础设施建设，加强 5G、人工智能、工业互联网、物联网等新型基础设施建设,扩大高速率、大容量、低延时网络覆盖范围,鼓励制造企业通过内网改造升级实现人、机、物互联,为共享制造提供信息网络支撑。
2020.04	中共中央政治局常务委员会会议	加快 5G 网络、数据中心等新型基础设施建设进度。
2021.03	《“十四五”规划》	聚焦高端芯片、操作系统、人工智能关键算法、传感器等关键领域，加快推进基础理论、基础算法、装备材料等研发突破与迭代应用。加快布局量子计算、量子通信、神经芯片、DNA 存储等前沿技术。

3.3.2 产业机构

（1） 中科海光

海光 DCU 是中国 HPC 超算、科学计算的主要芯片之一，全国已有 20 万片的 DCU 算力在网服务，用于基因工程研究、航空航天、医药研发、气象分析、地质勘测等领域。海光 DCU 采用与英伟达、AMD 一致的 GPGPU 路线，实现了算法开发和硬件应用的解耦。其异构软件开发平台 DTK（DCU Tool Kit）对标 CUDA，兼容 ROCm 软件生态。

经权威机构测试，算法引擎从英伟达 GPU 迁移到海光 DCU 后无识别精度下降问题，而一般算法引擎从英伟达 GPU 迁移到 NPU 会出现算法/引擎的识别准确率下降，造成产品无法使用。海光 DCU 支持标量计算和矢量计算，适用于各种人工智能场景，适配难度小、周期短、成本低。

在大模型领域，海光 DCU 已支撑国内顶级大模型应用。中科院自动化所的“紫东太初”三模态大模型，实现了图文音语义统一表达，海光基于 DCU 平台开展了多模态预训练模型架构设计与优化，在 100 亿参数下，达到 1024 个节点（4096 张 DCU 卡）近线性扩展，实现多模态理解与生成的多任务统一建模。

北京智源人工智能研究院的“悟道 2.0”，是双语跨模态大模型。海光与智源合作，对 2000 亿参数的 GLM-XXLarge 模型进行训练，基于 PyTorch 框架、Apex 混合精度计算框架、DeepSpeed 和 Megatron 分布式框架，通过模型并行、数据并行和流水线并行调优，最高实现 6000 个节点（24000 张 DCU 卡）扩展，节点扩展能力接近线性。海光 DCU 支持清华智谱、讯飞星火、百度文心一言、LLAMA、百川智能等主流大模型引擎和应用。

海光国产 X86 路线具备 CPU 逻辑运算与 DCU 异构计算的综合服务能力，其中海光 CPU 已在金融行业规模化应用。

（2）华为昇腾

1）芯片特点

昇腾 AI 芯片本质是一个片上系统（System on Chip，SoC），主要可以应用在和图像、视频、语音、文字处理相关的应用场景。其主要的架构组成部件包括特制的计算单元、大容量的存储单元和相应的控制单元。该芯片大致可以划分为：芯片系统控制 CPU（Control CPU），AI 计算引擎（包括 AI Core 和 AI CPU），多层级的片上系统缓存（Cache）或缓冲区（Buffer），数字视觉预处理模块（Digital Vision Pre-Processing，DVPP）等。

昇腾 AI 芯片集成了 CPU 核，每个核心都有独立的 L1 和 L2 缓存，所有核心共享一个片上 L3 缓存。集成的 CPU 核按照功能可以划分为专用于控制芯片整体运行的主控 CPU 和专用于承担非矩阵类复杂计算的 AI CPU。两类任务占用的 CPU 核数可由软件根据系统实际运行情况动态分配。

除了 CPU 之外，昇腾 AI 芯片真正的算力担当是采用了达芬奇架构的 AI Core。这些 AI Core 通过特别设计的架构和电路实现了高通量、大算力和低功耗，特别适合处理深度学习中神经网络必须的常用计算，如矩阵相乘等。由于采用了模块化的设计，可以很方便的通过叠加模块的方法提高后续芯片的计算力。

2）CANN 异构计算架构

芯片使能软件是 AI 处理器的使能引擎，支撑开发者自定义

算子开发，更好的应用和挖掘 AI 处理器的能力。通过自定义子开发，可以丰富和繁荣 AI 处理器的算子，来支撑更多的网络模型，从而更好的支撑行业应用的开发。

为了获取更高的芯片运行效率，提升 AI 算力硬件系统的吞吐率，帮助开发者提升开发和训练效率。芯片使能软件包括芯片编程接口、计算编译引擎、计算执行引擎、计算库以及计算调优引擎等功能单元。

3）AI 框架 MindSpore

AI 框架的主要目的是把程序员从繁琐细致的具体编程工作中解放出来，从而可以将主要精力集中在人工智能算法的调优和改进上。由于深度学习的算法发展很快，同时支持深度学习算力的硬件众多，一个框架的好坏往往取决于对上层算法和底层硬件的广泛兼容和适配能力。在硬件基础设施上提供 AI 框架支持，支持 MindSpore、TensorFlow 开发框架。具有自动微分、自动并行、动静态图结合、全场景部署协同、全栈协同加速、科学计算工具包等特性。

同时，针对 AI 科学计算工具包 MindScience，提供场景应用创新，拓展 MindSpore 的边界，MindScience 是基于昇思 MindSpore 融合架构打造的科学计算行业套件，包含了业界领先的数据集、基础模型、预置高精度模型和前后处理工具，加速了科学行业应用开发。目前已推出面向电子信息行业的 MindElec 套件和面向生命科学行业的 MindSPONGE 套件，分别实现了电磁仿真性能提升 10 倍和生物制药化合物模拟效率提升 50%。

（3） 寒武纪[8]

寒武纪是智能芯片领域全球知名的新兴公司，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。寒武纪不直接从事人工智能最终应用产品的开发和销售，而是通过对各类人工智能算法和应用场景有着深入的研究和理解，研发和销售符合市场需求的、性能优越、能效出色、易于使用的智能芯片及配套系统软件产品。寒武纪成立于 2016 年，成立之初以 IP 授权的形式进入市场，推出商用终端智能处理器 IP 产品寒武纪 1A。随后寒武纪迅速拓展云端业务，2018 年推出第一代云端 AI 芯片思元 MLU100。寒武纪不断拓展业务范围，陆续推出云端芯片以及边缘 AI 系列芯片。

目前，寒武纪的主要产品线包括云端产品线、边缘产品线和 IP 授权及软件三部分。云端产品线目前包括云端智能芯片、加速卡及训练整机。其中，云端智能芯片及加速卡是云服务器、数据中心等进行人工智能处理的核心器件，能为云计算和数据中心场景下的人工智能应用程序提供高计算密度、高能效的硬件计算资源，支撑该类场景下复杂度和数据吞吐量高速增长的人工智能处理任务。

（4） 昆仑芯[9]

昆仑芯成立于 2011 年，前身为百度智能芯片及架构部，团队成员多数成员来自百度、高通、Marvell、Tesla 等行业头部公

[8] 参考来源：产业趋势！AI 芯片行业发展政策、市场供需、竞争格局及未来前景分析，智研咨询，http://t.10jqka.com.cn/pid_277705567.shtml

[9] 参考来源：中国"芯"势力-昆仑芯，https://zhuanlan.zhihu.com/p/608006865?utm_id=0

司。昆仑芯新品 AI 芯片 R200 于 2022 智算峰会上正式发布，基于新一代昆仑芯自研架构 XPU-R，通用性和性能显著提升，采用 7nm 先进工艺，算力可达 256TOPS。配合百度飞桨平台，获得更友好开发的环境。

昆仑芯目前有 2 个系列的产品，分别是昆仑芯 1 代芯片 K 系列和昆仑芯 2 代芯片 R 系列。昆仑芯 1 代芯片采用 XPU-K 架构，制程为 14nm 工艺，主要应用于云数据中心和智能边缘，支持全 AI 算法。第二代云端通用人工智能计算处理器，采用新一代昆仑芯 XPU-R 架构，通用性和性能显著提升，制程采用 7nm 先进工艺，GDDR6 高性能显存，支持虚拟化，芯片间互联和视频编解码。

（5） 趋动科技

趋动科技成立于 2019 年，专注于为企业用户构建数据中心级 AI 算力资源池和 AI 开发平台。

1）GPU 资源池化技术的演进

趋动科技认为 GPU 资源池化技术从初期的简单虚拟化，到资源池化，经历了四个技术演进阶段。**简单虚拟化：**将物理 GPU 按照 2 的 N 次方，切分成多个固定大小的 vGPU（Virtual GPU，虚拟 GPU），每个 vGPU 的算力和显存相等。实践证明，不同的 AI 模型对于算力、显存资源的需求不同，所以这样的切分方式并不能满足 AI 模型多样化的需求。**任意虚拟化：**将物理 GPU 按照算力和显存两个维度，自定义切分，获得满足 AI 应用个性化需求的 vGPU。**远程调用：**AI 应用与物理 GPU 服务器分离部署，

允许通过高性能网络远程调用 GPU 资源。这样可以实现 AI 应用与物理 GPU 资源剥离，AI 应用可以部署在私有云的任意位置，只需要网络可达，即可调用 GPU 资源。**资源池化：**形成 GPU 资源池后，需要统一的管理面来实现管理、监控、资源调度和资源回收等功能。同时，也需要提供北向 API，与数据中心级的资源调度平台对接，让用户在单一界面，就可以调度包括 vGPU 在内的数据中心内的各类资源。

趋动科技的 OrionX 猎户座 AI 算力资源池化解决方案已经能做到第四个阶段，能够帮助用户提高资源利用率和降低总体拥有成本，提高算法工程师的工作效率。

2）异构 AI 算力资源池

面对国内外的政策形式，国内 AI 芯片的迅速崛起，趋动科技在技术方面积极创新，OrionX 现在已经能兼容除英伟达以外的国产卡，包括海光、寒武纪，实现异构资源池化管理。

使用海光、寒武纪、英伟达等异构算力加速卡构建 AI 算力加速资源池，资源池内各类硬件加速卡可通过趋动科技 OrionX 进行算力抽象，软件化后形成统一的 AI 加速算力资源提供给上层应用使用。异构算力池化解决方案可在实现多厂商 AI 算力硬件统一管理、统一调度、统一使用的同时，结合软件定义异构算力技术实现 AI 算力统筹分配、AI 算力资源池化、高效 AI 算力保障、围绕 AI 算力运维管理。

（6） 格兰菲

格兰菲拥有图形图像以及 AMOLED 驱动产品的核心技术，

掌握图形图像产品架构，从架构到算法再到实现，具备自主知识产权，全程安全、可靠、可控。解决方案提供的产品和服务包括独显芯片和图形图像产品的 IP 设计服务，主要面向显卡及其它桌面或移动端计算市场。

Arise1 系列独立显卡能胜任办公、网页、基础设计软件、多屏（2-4）显示的应用场景。通用计算性能胜任生物特征识别、物体分类等 AI 终端应用。视频编解码性能可应用于各种视频宣传的柜面窗口。

Arise1 系列独立显卡有 Arise-GT-10C0/2G、Arise-GT-10C0/4G、Arise-GT-10C0/6G、Arise1020/2G 等型号产品，主要应用于桌面、商业显示以及通用计算等中高端应用场景，芯片基于 28 纳米工艺制造，内置完全独立自主研发的新一代图形图像处理引擎，兼容银河麒麟 KOS、统信软件 UOS、Windows 等主流操作系统，同时可在 X86、ARM、MIPS、Loongarch 等主流 CPU 平台操作运行，支持多种图形图像的 API 接口标准，如 DirectX11、OpenGL4.5、OpenCL1.2 等。

格兰菲显卡采用的 IMR 桌面级渲染架构（非 TBR 移动级渲染架构），经过 5 代产品迭代，已在信创市场真替真用。产品主要与国内整机产商合作，以整机预安装的方式给到终端客户。目前终端客户已经应用于党政办公、金融、能源、教育等领域。产品适用于需要高性能图像处理能力的应用场景，如 3D 建模、动画制作、虚拟现实、增强现实、科学计算、柜面终端等。

（7）摩尔线程

1）公司简介

摩尔线程成立于2020年10月，是一家以GPU芯片设计为主的集成电路高科技公司。公司致力于创新面向元计算应用的新一代GPU，构建融合视觉计算、3D图形计算、科学计算及人工智能计算的综合计算平台，建立基于云原生GPU计算的生态系统，助力驱动数字经济发展。

2）主要产品

摩尔线程专注于研发设计全功能GPU芯片及相关产品，支持3D图形渲染、AI训练与推理加速、超高清视频编解码、物理仿真与科学计算等多种组合工作负载，兼顾算力与算效，能够为中国科技生态合作伙伴提供强大的计算加速能力，广泛赋能数字经济多个领域。摩尔线程目前已与国产CPU飞腾、鲲鹏、龙芯、海光、兆芯，以及国产操作系统厂商麒麟、统信均完成兼容性互认，加速国产化数字办公应用的同时，也为国产工业设计软件、GIS软件等提供流畅稳定的使用体验。

MTT S50基于摩尔线程先进的MUSA架构打造，是一款8GB显存的工作站级显卡。凭借强大的三维渲染性能与丰富的图形框架特性支持，MTT S50支持osgEarth、超图、中地数码、苍穹GIS和易智睿等多个主流GIS厂商应用，可以在主流平台上流畅运行CAD 3D、BIM等工业软件。MTT S50提供对DBNet、CRNN、Yolo、Restnet50/101等主流AI模型，以及对PyTorch、TensorFlow、PaddlePaddle框架的支持。

MTT S3000基于摩尔线程MUSA架构，同时也是第一款基

于“春晓”的全功能服务器 GPU 产品。基于 MUSA 软件栈，现有算法可实现向 MTT S3000 的无缝迁移；能够支持包含单机单卡、单机多卡、多机多卡在内的多种部署模式。MTT S3000 兼容 PyTorch、TensorFlow、百度飞桨（PaddlePaddle）、计图（Jittor）等多种主流深度学习框架，并实现了对 Transformer、CNN、RNN 等数十类 AI 模型的加速。摩尔线程提供了系列基于摩尔线程创新性 MT Mesh 2.0 的 GPU 云原生方案。MT Mesh 2.0 可以根据云端中心应用负载，自动化分配 GPU 计算和显存资源，实现 GPU 算力弹性伸缩。

摩尔线程元宇宙平台提供从硬件集群、软件基础架构到 SDK 工具链的全栈式解决方案，涵盖元宇宙的多个核心要素。MTVERSE 以摩尔线程 MUSA GPU 集群为算力基础，为用户提供计算基础架构及服务，包括大数据、AI 训练与推理、图形渲染和物理仿真三大平台，提供从硬件集群、软件基础架构到 SDK 工具链的全栈式解决方案，涵盖元宇宙中人、场景和内容等多个核心要素。

面向大模型应用领域，摩尔线程提供大模型训练、部署和运营的完整软硬件一体数据中心 GPU 计算集群交钥匙解决方案。面向我国企业客户，提供用于训练的数据处理方案、建设大模型分布式训练，训练模型部署和迭代优化，下游场景推理的专业系统方案、低成本算力设施、业务落地部署的运维技术。基于 KUAE 内的预训练模型，提供 finetune 和 promot tuning 的方案，支持行业合作伙伴和行业客户定制模型，大规模降低用户研究和使用大

模型的成本。

3）开发平台

MUSA(Moore Threads Unified System Architecture)是摩尔线程推出的通用并行计算架构，该架构使 GPU 能够解决复杂的计算问题。它包含了 MUSA 指令集架构（ISA）以及 GPU 内部的并行计算引擎。MUSA 带有一个软件环境，开发人员可以使用 C/C++语言来为 MUSA 架构编写程序，所编写出的程序可以在支持 MUSA 的 GPU 处理器上以超高性能运行。同时 MUSA 软件栈提供了各个领域的加速库，供开发者直接使用，应用在各种高性能计算的场景中。

完整的 MUSA 开发环境提供 compiler、profiler、porter、debugger 等完整的开发调试工具集，包含 AI、多媒体、信号处理、通信等多个加速库，各个底层依赖的库和运行时，并且维护了各个组件的版本依赖关系，提升开发应用的效率。MUSA 兼容 CUDA 编程接口，提供了编译器，可以编译 CUDA 程序，CUDA 开发者可以继续采用熟悉的方式做开发，帮助 CUDA 开发者更快了解 MUSA 架构，更好地发挥国产化平台的算力。

4 金融应用情况

4.1 应用场景

金融行业是人工智能重要的应用领域，人工智能在金融行业的应用提升了金融业服务能力水平。人工智能在金融领域中的应用按照技术复杂度和应用成熟度等可分为传统场景和创新场景。

4.1.1 传统场景

（1） 金融开户

金融开户是每个客户进入金融服务必须经历的第一步，金融开户无论在移动端开户还是柜台开户，无论是个人开户还是机构开户，无论是银行、证券还是保险等行业，都是标准步骤。金融开户涉及人脸识别、图像分类、OCR 识别等 AI 技术，开户流程中需要做人脸验证、需要对客户提供的资料进行分类，对于重要的证件进行 OCR 识别和结构化提取。

通过人脸识别技术对用户的身份进行远程核验，机构将身份证信息和用户的现场照片发到身份核验系统，由核验系统返回身份信息的确认或拒绝，以此实现用户注册和开卡的集中管控和备案，并留存关键证据信息。

OCR 识别技术将用户信息、签约合同等内容实现统一管理。结构化支持文件分类、内容检索、要素信息抽取、文件内容合规性审核、文件概要抽取等。与人脸识别技术结合，实现了身份信息与证件等资料信息的绑定。

（2） 智能客服

通过电话客服渠道、网上客服、APP 以及智能机器人终端与

客户进行语音或文本的互动交流，理解客户业务需求，语音回复客户提出的业务咨询，并能根据客户语音导航至指定业务模块。对传统按键式菜单进行改造，用户使用自然语音与系统交互，实现菜单扁平化，提升用户满意度，减轻人工服务压力，降低运营成本。电话客服不再受限于菜单，可开展全业务的语音导航服务。该应用场景涉及语音识别（ASR）和语音合成（TTS）技术，同时需要自然语言处理技术理解客户的问题，并且给出准确的回复，如 ChatGPT 就是目前典型的 NLP 技术产品。

（3） 财务报销

企业对员工的财务报销，以前是人工进行贴票、手工计算、最后进行报销。采用 OCR 辅助的报销方式，只需要将报销的发票、单据用手机或者扫描仪进行扫描，系统获取单据图像之后，自动进行 OCR 识别，并进行分类，根据设定规则进行汇总，最后进入 OA 系统，通过 OA 审核就可以发送到财务部门进行支付。可以有效降低报销环节的人力成本，提高报销的效率，降低报销中存在的风险。

（4） 信贷审核

信贷是金融机构的主要业务，为了降低信贷风险，需要对贷款对象进行风险排查。个人客户需要提供银行流水，经过 OCR 识别后分析是否有异常行为，并及时给出提醒；机构客户提供财务报表等信息，经过 OCR 对财务报表进行识别和分析，得出企业经营数据和信用情况，发现风险点并及时给出审核意见。

（5） 视频监控与人脸识别

金融机构网点是重点监控区域，一方面提前识别可疑人员和客户，另一方面通过行为识别还可以监控金融机构员工行为是否合规、安全等。运用图形视频处理技术，实时监控员工在规定动作以外的行为，提醒后台人员注意，识别并标记视频监控中发现的员工可疑行为录像片段，提示后台人员查看。集中运营中心、机房、保险柜、金库等重要场所可采用人脸门禁提高内部安全控制，通过人脸识别的验证方式，实现银行内部安全管理，有效防范不法分子的非法入侵；同时进行多人的人脸识别，实现智能识别，达到安全防范的目标。

（6） 金融反欺诈

金融欺诈是指借款人用虚构数据、隐瞒事实的方式来骗取贷款，且在申请贷款后主观上没有还款意愿，或客观上没有偿还能力，可能造成出借人资金损失的行为。常见的金融欺诈类型有虚假用户注册、企业欺诈、金融钓鱼网站、病毒木马程序、账户隐私窃取、融资套现、他人冒用等。利用人工智能技术，利用多维数据（支付、助贷、征信等）建立反欺诈模型，制定精准的反欺诈规则，识别各种欺诈行为，有效降低金融风险。

（7） 量化交易

传统的量化投资策略是通过建立各种数学模型，在各种金融数据中试图找出市场规律并加以利用。无论是根据人的经验判断，还是通过经典的数学模型，力所能及的模式都是有限的。从探寻股票市场的全局来看，人类积累经验的研究可以接近某一个局部的最优，而真正全局的“最优解”或许超出了目前传统量化力所

能及的范畴。

一方面，对于市场中蕴藏的复杂的非线性规律，很难通过传统数学模型进行挖掘；另一方面，对于海量数据的挖掘，困于计算机运算能力的限制，如果不利用数据挖掘算法，往往需要耗费大量时间。人工智能能够提供非线性关系的模糊处理，弥补了人脑逻辑思维模式的单一性，同时，如果加以利用相关算法，可以大幅提高规律的搜索效率。人工智能的引入也使得投资策略更加丰富，如 AI 算法对于非线性模式的因子挖掘在多因子领域比传统线性多因子模型更加敏锐。

4.1.2 创新场景

（1） 数字人

数字人（Digital Human/Meta Human），是运用数字技术创造出来的、与人类形象接近的数字化人物形象。为金融行业客户提供更亲切、智能、高效的金融数字人解决方案，满足客户营销、培训等多样化金融业务场景。

数字人将人脸识别、计算机视觉图像分析、语音识别、自然语言处理、语言合成等多种人工智能技术结合应用。在营业厅大屏助手、手机银行营销助手、自助终端虚拟形象等已广泛应用。在大模型的加持下，数字人实现了形象自动生成、声音拟人化、多轮对话、自主学习等各类能力的全面升级。

（2） 交互式客服

在新时代，智能客服平台在通用认知智能大模型的技术突破下，在文本生成、语言理解、知识问答、逻辑推理、数学能力等

方面有着跨越式发展，能够超越人工客服的服务能力和知识理解范围，成为金融机构与用户互联互通的重要纽带。

（3）AI 双录

全场景 AI 赋能，解决质检滞后、工作时段固化等双录问题，极大降低人力时间成本，保障双录过程完整、严谨、合规，改善双录体验，用户随时随地办理业务，实现自助双录。

通过综合计算机视觉图像处理、生物识别、OCR 识别、语言识别、语言质检、声纹识别、自然语言处理、语音合成等技术，对理财售卖、贷款受理、信贷开户、保险签约等各类需要强监管的场景，实现即时前督和批量后督处理。

（4）流程自动化机器人

机器人流程自动化（Robotic Process Automation, RPA）通过特定的“机器人软件”，模拟人在计算机上的操作，按规则自动执行流程任务，将日常重复性工作、规则与逻辑明确、跨系统数据集成、数据搜集、检索和汇总等场景通过替代鼠标和键盘的动作，实现自动化完成办公任务。包含但不限于财务管理、人力资源、运营管理、供应商筛选、IT 运维管理、数字法务等。

在大模型的加持下，基于自然语言自动生成 RPA 业务流程和脚本，实现端到端的自动化。

（5）知识图谱

知识图谱是一种大规模语义网络，它以结构化形式描述客观世界中的概念、实体及其关系。实现了智能检索、智能推荐、智能问答、故障诊断、质量控制等场景。包括知识管理、知识交互、

知识存储、知识构建、知识应用、知识服务等环节。

（6） 机器学习平台

机器学习平台为金融、地产、交通和互联网等多行业客户提供数据分析能力、实时能力和 AI 能力的建设。提供机器学习分析和实时计算能力，帮助数据分析师和数据科学家快速协同开发，实现模型管理和应用支持，在科技创新、人工智能等前沿领域为客户业务创造更大价值。

机器学习平台可以实现指标加工、交易反欺诈、实时精准营销、实时授信、实时交易、实时监控等业务的支撑。

4.2 机构实践

4.2.1 建设银行

（1） 案例背景

中国建设银行应用人工智能技术构建智慧型银行，结合业务场景，通过产品化建设和模型能力定制研发，服务于总分行多个业务部门，支撑多个业务场景，算法能力在银行各个领域得到了实践验证，产品化 AI 能力敏捷供给水平实现突破，在同业达到第一梯队水平。并在计算机视觉、自然语言处理、智能语音、智能推荐与决策、知识图谱等 5 个技术领域深耕研发，敏捷支持业务降本、增效与提质。

在计算机视觉领域，建设了自主可控的文字识别能力和视频分析基础能力，支撑所有业务部门的 AI 相关业务需求工作，通过产品化创新，敏捷支持行内众多场景，提升用户体验。在影像文字识别方向，支撑了多种影像识别需求。通过零代码产品化，

将票据相关业务缩短研发实施周期，显著提升了能力释放效率。在视频识别方向，支撑人体分析、动作行为识别、车辆车牌识别等视频分析需求，提升云、边侧视觉服务支持能力。

在自然语言处理领域，提升融合多模态的信息抽取能力，适配行内业务需求。在智能审单场景，实现多类凭证的自动识别、多类要素信息智能采集及规则检核，节约了大部分的人工审核成本，有效解决业务场景中凭证类型多、数量大、风险高的业务痛点；在货币市场交易场景，缩短交易处理时长，形成同业较好实践；支撑风险预警项目，优化风险事件模型推理性能，对每天资讯等文本进行分析，智能识别有风险的事件，提升每秒可处理文本数量，降低了业务分析压力。

在智能语音领域，研究语音识别和语音合成能力自研技术，相较外部黑盒引入能力，自研能力可基于行内场景和数据进行模型迭代和效果提升，对语音识别能力近场中文普通话识别准确率达 95%，业务效果显著优于原行内产品能力和第三方产品通用识别能力。基于自研语音能力，逐步启动原语音识别能力的全面替换工作，批量提升行内客服效果和智能化水平；语音合成能力可支持数万字级别长文本合成，音色可个性化定制，漏字多字率低，长静音覆盖错误率低，应用于建行语音播报场景中。

在智能推荐与决策领域，支撑智能缴费推荐需求，针对千万级 C 端用户，优化缴费体验、提升 C 端留存率，促进缴费支付的转化率；完成内容推荐算法和工程的定制化初版研发，并试点部分城市；支撑相关业务的智能推荐初版研发，提升流量分发效

率以及内容曝光率；在内容推荐方面，提升用户粘性和页面停留时长，提高离线自评相似推荐结果准确率。已启动自动训练迭代的多模态推荐引擎建设。量化方向深耕业务场景提升服务能力，通过多类服务模型研发，进一步加强对智能投资场景的赋能。

（2） 技术方案

人工智能技术作为中国建设银行 TOP+金融科技战略的重要支撑技术之一来进行建设。目前，已形成了“1+5+N”的人工智能技术应用体系，建成了人工智能平台，构建了计算机视觉、自然语言处理、智能语音、智能推荐与决策、知识图谱等五个技术领域的能力，释放了丰富的人工智能应用场景。在研发能力纵向建设方面，目前各领域 AI 能力技术影响力逐步提升，在十六届文档分析与识别国际会议（International Conference on Document Analysis and Recognition，ICDAR）上，核心赛道中超越平安科技、联想 AI、德国 AI 研究中心等战队，荣获第二名；参加首届全国 AI 安全大赛，夺得人脸攻防核心赛道冠军。连续三年作为中国计算机学会（CCF）国际 AIOps 挑战赛主办方。

为推进基于国产化 GPU 芯片的应用工作，中国建设银行建立了实现同异构 GPU 芯片的统一管理、算力资源调度和应用的人工智能技术方案，如图 7 所示：

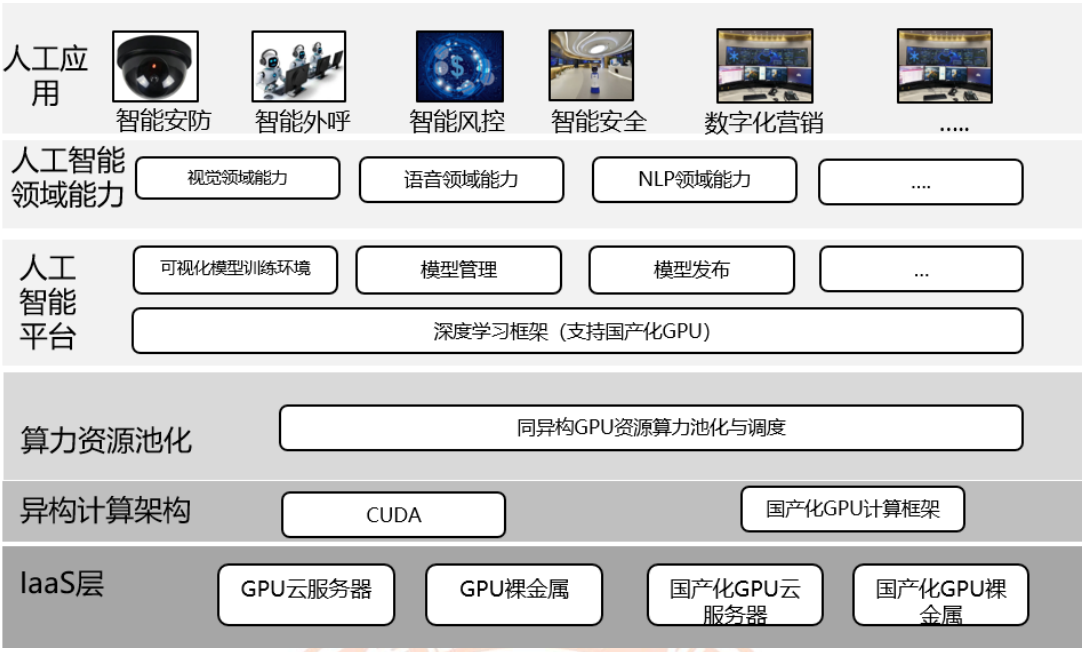

图 7 同异构 GPU 芯片的人工智能技术架构

IaaS 层：提供国外 GPU 芯片算力和国产化 GPU 芯片的算力云服务。**算力资源池化层：**实现同异构 GPU 资源的容器化、池化管理与调度。**人工智能平台：**实现深度学习框架对同异构 GPU 芯片适配。**人工智能领域能力：**提供人工智能各个领域的技术服务。

（3） 痛点、风险及挑战

随着人工智能应用领域和应用场景的不断扩展，尤其是基于大模型的应用，将出现突发式、爆发式的对 GPU 算力供给需求，需要解决底层 GPU 算力快速供给的痛点。

目前主要依靠国外 GPU 芯片，国内 GPU 芯片的适配处于起步阶段，面临着国外 GPU 芯片断供的挑战，国产化 GPU 芯片算力和算子弱于国外 GPU 芯片而带来的应用改造挑战，以及适配国产化 GPU 芯片的一些技术生态挑战等。

4.2.2 中国银行

（1） 案例背景

前期，中国 IT 底层标准、架构、产品、生态大多数都由国外 IT 商业公司来制定，由此存在诸多的底层技术、信息安全、数据保存方式被限制的风险，因此我国需逐步建立基于自己的 IT 底层架构和标准，近几年国产芯片厂家在芯片研究领域取得了巨大突破，国产化芯片正逐步应用于金融行业的人工智能领域。从人工智能的角度出发，研究国产 GPU 架构和中国银行目前人工智能产品使用的深度学习框架的适配程度，从而为后续将中国银行人工智能产品的算力替换为国产芯片做好准备；研究国产 GPU 在基础任务上的性能指标，研究国产 GPU 在基础图像识别、语音识别、自然语言处理等任务上的性能指标，包括响应时间、准确率、并发量等，输出同等配置下国产 GPU 与 CPU、非国产 GPU 的性能对比报告。得出国产芯片所需的资源配置，防止后续在国产芯片的使用和替代上，因为性能的差异，导致满足不了业务场景需要。在金融行业的人工智能领域中，研究 AI 芯片国芯人工智能服务器在图像分类、图像识别等场景等高频人工智能业务场景中的相关能力，验证其能否满足业务迁移、扩容的系列需求，目的是在日益复杂的国际形势下，率先开展 AI 国产化替代实践，为 AI 业务连续性增添保证。

（2） 技术方案

在测试环境中，搭载了某国产服务器、并配置国产推理卡，安装某国产操作系统。在 PyTorch 框架下，研究 ResNet50 模型

的吞吐率、准确率、并发量等性能信息是否满足场景需求。确认国产 GPU 的响应时间的效率需达到仅使用 CPU 识别的 5 倍以上，是同等价格国际主流显卡的 80%以上。确认国产 GPU 的并发量的效率需达到仅使用 CPU 识别的 3 倍以上，是同等价格国际主流显卡的 70%以上。确认国产 GPU 的准确率与同等价格过偶记主流显卡一致。从而实现验证在金融场景中，国产 AI 芯片在图像目标检测的可用性。

（3） 发展规划

输出相关测试的训练性能验证报告和推理性能验证报告。在训练上，相比于同等价位国际主流芯片，精度损失了 1.2%，训练性能提升了 65.6%。在推理上，相比于同等价位国际主流芯片，推理性能提升了 29.3%。在金融场景中，验证了国产 AI 芯片在图像目标检测可用的基础上，进一步去评估国产 GPU 性能，统计并比较国产 GPU 与国产 CPU、同业非国产 CPU、同业非国产 GPU 在图像模型识别、训练上的性能。从而验证了智能图像识别模型迁移、同步的策略可行性。

在 ChatGPT 热潮席卷全球以及大模型日益广泛推广的背景下，生成式人工智能（AIGC）将会拉动芯片产业量价齐升。在量上，AIGC 带来的全新场景+原场景流量大幅提高；在价上，对高端芯片的需求将拉动芯片均价上升。因此，针对大模型算力研究主要分为三部分：一是研究包括开源的、商业性私有化部署的大模型的算力要求；二是研究大模型在训练过程中不同阶段的算力要求；三是在不同芯片上运行各类大模型，研究当前市面上开源

大模型对各类芯片的支持程度。

下一步将围绕以下方面开展工作。一是在智能客服领域，进一步研究大模型赋能业务的相关场景，并继续基于中国银行与国产厂商的联合创新实验室机制，借助国产 GPU 进行大模型的相关场景实践。二是推动 AI 芯片适配工作复杂、虚拟化精度不足、生态体系不成熟和 AI 算力集群建设成本高等问题的解决。三是采用开源、产学研联合创新、商用产品采购等技术路线，探索适合的金融应用场景。

4.2.3 邮储银行

随着大数据、云计算、人工智能等技术快速发展，金融行业对技术的理解愈加深入，金融科技开始深度应用于获客、风控、贷后管理、客户服务等环节，在业务中的应用与渗透逐渐加深，成为业务发展核心力量。

面对新的形势，邮储银行适时提出了“创新驱动、科技引领”的战略目标，通过搭建协同、共享的金融科技创新体系推动全行联动创新。邮储银行创新实验室积极开展机器学习、智能语音、生物识别、知识图谱、智慧物联、智能决策等新技术的研究，积极探索“人工智能平台+应用”的模式赋能智能营销、智能服务、智能风控、智能运营等业务场景，发挥科技创新驱动作用。

（1） 案例背景

近年来，伴随数字经济的突飞猛进，银行业务快速发展、各类交易场景日益丰富，金融数字化转型全面提速，金融机构逐年增加科技投入，对于 AI 芯片、服务器等人工智能基础建设不断

扩大。

在数字中国建设的大背景下，信创作为我国全面推动科技自立自强的重要举措，在数字中国安全建设方面发挥着重要作用。金融信创作为信创产业及生态的关键部分，也是我国科技自立自强的重要驱动力。金融机构逐渐采纳自主研发、自主可控的国内技术路线，加快国产基础软硬件等转型升级，已成为大势所趋。

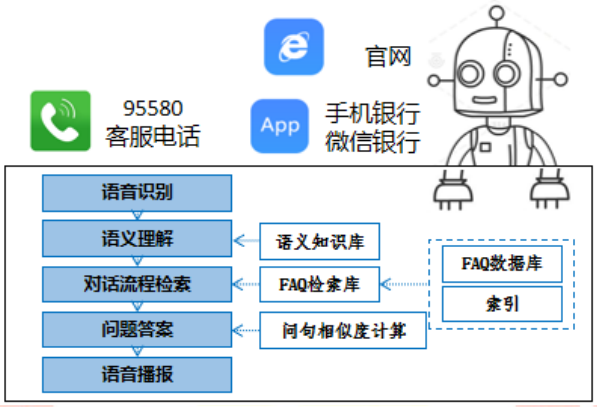

图 8 邮储大脑智能语言语音系统示意

邮储大脑智能语言语音系统集成语音识别、语义理解、语音合成、对话流程管理等技术，可让机器人理解客户说的内容，让用户“动动嘴”就能通过 WEB\APP\电话等多渠道获得 7*24 小时的金融服务，复杂的业务还可一键切换至人工服务，让视障同胞、老年同胞也能够享受到便捷的服务体验，大大提升了银行客服的服务效率和覆盖范围，降低了运营成本。智能外呼则使用智能语音机器人与客户完成全流程自动交互，针对不同场景可以配置差异化人机耦合策略，有标准化服务、覆盖速率高、整体成本低、规模扩展快等特点，可以实现外呼效率最大化，助力银行降本增效。

（2） 国产化适配测试情况

为了推进智能化技术的规模化应用，邮储银行积极响应国产化自主可控的国家战略方针以及金融机构整体信创要求，布局国产化 AI 芯片测试和适配工作，并进行国产芯片的测试。邮储银行选取典型、成熟、广泛应用的智能语音场景作为试点，其中实时语音识别场景测试了某国产 GPU 芯片性能，基本达到国际主流加速卡性能的 10%左右，满足可用要求。后续将针对目前存在的某 GPU 部分模型对于动态输入响应速度较差的问题进行进一步定位和优化，提高语音识别引擎所使用的底层算子对于动态输入的响应，减少同类算子的编译及执行时间。完成优化后，预计提升某 GPU 性能至国际主流显卡的 40%左右。

4.2.4 光大银行

随着智能检索、生物识别、计算机视觉、机器学习、OCR 字符识别、自然语言处理、RPA 机器人、AR、VR 等人工智能技术在金融应用领域的迅速发展，这些关键技术在给重构和优化传统金融业务流程带来新机遇的同时，也在落地的过程中提出了新的挑战。技术创新已成为传统金融机构转型升级的重要手段，金融机构开始挖掘以用户为中心的大数据资源，并将数据成本转化为数据价值。

在人工智能产业蓬勃发展的时代背景下，光大银行积极开展生物识别、自然语言处理、计算机视觉、知识图谱等新兴技术的研究，并在智能客服、智能风控、智能运营等实际金融业务场景融入人工智能应用，发挥科技创新驱动作用。

（1） 智能客服场景

光大银行通过将语音识别和自然语言理解技术集成于智能客服系统，实现了客服系统“自助+智能+人工”三层的服务模式，全天候提供服务，在提升客户体验的同时，也大大降低了光大银行的运营成本。其中智能语音项目通过语音识别和自然语言理解技术的应用，实现了光大银行语音客服系统的智能化升级，通过机器人完成客服系统的语音导航、语音交互、语音咨询功能，为客户提供服务。智能文字项目基于场景和业务模型开发上下文关联模型，应用于光大银行网站、网银、微信、百度知道等互联网渠道，为客户提供基于文字的智能客服服务，机器人回答准确率高，大部分文字客服由智能文字机器人来完成，只有极少数的请求交由人工处理。

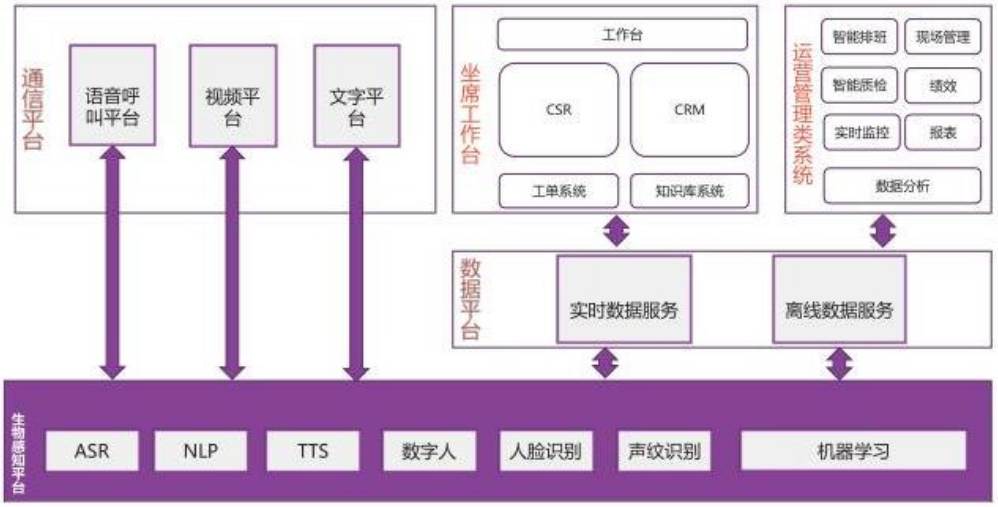

图 9 光大银行智能客服场景

（2） 智能营销场景

智能营销是通过人工智能和大数据分析技术，根据客户的基础属性、风险偏好、业务需求及业务倾向等信息对客户群体进行

细分，挖掘客户潜在需求，进行客户行为预测，从而开展针对性的营销活动，实现业务营销从传统大众营销向智能营销的转换。通常智能营销由客户画像、客户行为预测和营销自动化组成。在KYC方面，光大银行完成了电子银行KYC及行为分析系统的建设，在传统客户标签的基础上，引入客户在电子渠道的行为采集技术，通过客户实时行为的捕获和分析，支持面向客户的一对一实时营销推荐，提高营销成功率。在客户行为预测方面，光大银行利用数据挖掘建模技术，预测客户个体的行为变换，提高产品和客户营销的精准度。

4.2.5 华夏银行

（1） 案例背景

1） 解决场景

一是移动营销系统调用证件识别场景：客户进件手工填写身份证件信息，通过投产系统拍照完成客户证件信息录入，提高了客户录入效率与用户体验。二是PAD系统进件名片识别场景：名片识别模块可快速识别并录入客户名片。三是历史进件证件识别场景：针对卡中心存量历史进件进行跑批处理，从中挑出身份证影像并识别证件信息。

2） 服务器端

身份证识别、军官证识别、护照识别、港澳台居民往来内地通行证识别、内地往来港澳台居民通行证识别。

3） 移动端

身份证识别、银行卡识别、名片识别。

4） 接入系统

移动营销系统、PAD 申请系统。

5） 业务量统计

从 2019 年上线至今，各业务系统识别量为 100 万户/年。

（2） 技术方案

整体采用深度神经网络 OCR 识别技术，仅用三个工序即可完成识别过程：文字区域定位、整行识别、结构化输出。定位与识别均采用卷积神经网络 CNN、循环神经网络 RNN、长短期记忆网络 LSTM 技术实现，可在灰度图像上实现文字区域的自动定位和整行文字的 OCR 识别。尤其是后者，彻底弥补了传统 OCR 技术中单字识别技术无法借助上下文来判断形似字的问题。

基于深度学习的 OCR 识别过程共分为四个主要的步骤。一是方向识别并矫正：对证件在图片内的方向进行识别，然后根据需要进行旋转；二是文字定位：对图片上的文字区域进行定位；三是文字识别：对文字区域进行整行识别；四是结构化输出：对识别结果进行结构化分析，输出为结构化的数据；深度学习的 OCR 识别过程，会定位出所有文字并进行识别，最后进行结构化分析并输出。

功能亮点：深度学习的 OCR 识别过程，可针对清楚的单个证件的单独识别，也可以实现一张图有多张不清晰、褶皱的卡证实现自动切割、定位、分类、识别最后进行结构化分析并输出整个流程。

1） 阶段非信创方案

2019 年采购多品牌服务器 4 台，不同系列国际主流显卡 10 张，可满足识别效率为 8 张/秒。

2） 阶段半信创方案

2022 年基于目前市场上有可替代信创服务器，新增信创服务器用于 OCR 系统，主要采用国产服务器 10 台，30 张某国际主流显卡。

3） 阶段全信创方案

2023 年市场上信创服务器及国产芯片趋于成熟，卡中心与金融信创生态实验室启动并完成信创硬件适配验证，新增 4 台某国产服务器及 8 张某国产显卡。

（3） 发展规划

满足业务需求扩容：根据往年数据统计业务量增长情况，目前采购的 OCR 许可授权可满足卡中心未来 5 年的业务使用。

适配其他国产服务器：邀请厂商适配其他国产服务器避免服务器性能单一硬件风险。

行方能力提供平台：在全栈信创情况下提供更丰富产品，后期计划建设标注训练平台承接全行 OCR 能力建设任务，打造自主可控的平台，节省采购成本。

5 后续工作建议

人工智能技术已广泛应用在金融业各业务场景中，成为金融业变革交易模式、释放数据价值、实现数字化转型的技术基础。AI 芯片作为人工智能技术的硬件底座，是实现大算力、大模型的“发动机”和“加速器”。打造自主可控、安全稳定、性能优良的 AI 芯片产品，不仅推动人工智能技术发展，更进一步赋能金融业务实现高质量发展。为进一步降低金融机构 AI 芯片应用成本、实现金融业 AI 应用场景突破，需产业各方凝聚共识、发挥合力、紧密协作，共同打造基于国产化 AI 芯片的完整应用生态体系，本文提出以下六个工作方向，为下一步工作提供指引。

5.1 形成一批具有金融行业特色的应用系列标准

一是筑牢金融业 AI 芯片应用标准底座，围绕参考架构、应用分类等进行基础共性标准研制。**二是**规范 AI 芯片应用的测试方法、测试要求、测试评价等相关内容和过程，围绕 AI 芯片在不同技术路线、应用场景下的性能、功耗、兼容性、利用率、异构适配等方面编制测试标准，研究适合金融业的测试平台和工具。**三是**围绕数据安全、隐私保护、算法可解释性、合规性等方面编制安全保障标准。**四是**加强标准的宣贯解读、培训工作，推进标准实施，促进标准化成果应用落地。

5.2 推动一批金融行业普遍关注的课题攻关研究

通过面向金融机构、产业机构、科研院所等征集一批金融行业普遍关注的研究课题开展联合攻关。重点围绕国产芯片适配验证、异构芯片算力池化、AI 芯片与信息安全、基于国芯的场景化

验证等方向，充分发挥行业协会、产业联盟的社会团体力量，组织产学研用各方开展课题联合攻关。根据研究成果，建立相应的产业示范性项目，在金融行业试验和推广，为更多金融机构提供参考借鉴。沉淀相关研究和推广成果经验，并适时研究推进形成具有行业共识的行业报告、白皮书等。

5.3 征集一批适宜开展适配验证的金融应用案例

面向金融机构，按照银行业、证券业、保险业及其他类别，围绕金融开户、智能客服、信贷审核、金融反诈、量化交易、流程自动化机器人、AI 双录等金融行业 AI 芯片应用的重点场景，征集一批已自行或与芯片、算法公司等适配验证过的成熟应用案例，或拟开展适配验证的应用案例，形成金融业 AI 芯片应用适配验证案例集，展现行业应用潜力和示范作用的优秀经验，推动产业快速发展。

5.4 研发一批面向中小金融机构的易用产品服务

鼓励产业机构为中小金融机构开发模块化、可定制、安全合规的 AI 方面产品服务，以功能简单易用、界面简洁明了、操作直观友好、产品开箱即用为原则，使金融机构能以较低的成本、较短的周期引进相关服务能力，提升中小金融机构的业务效率和竞争力。产业机构建立迅速响应的支持团队，为金融机构提供后续培训和支持服务，帮助中小金融机构顺利上手并充分利用产品功能。持续与金融机构建立紧密合作关系，收集金融机构反馈和需求，并及时进行产品优化升级。

5.5 制定一批可用性强的国产芯片产品服务目录

金融机构依托自身业务实际，各自开展芯片适配验证工作，不仅适配工作繁杂，占用机构较多人力、物力以及算力资源，还不能形成良好的行业适配经验，产生适配工作的孤岛效应。鼓励和支持检测机构提供芯片测评服务，对芯片在不同场景下的应用提供测试验证。基于金融机构主要应用的生物识别技术、OCR 识别技术、智能客服、AI 双录、大模型训练和推理等推广 AI 能力与国产芯片的适配效果，形成一批可用性强的国产 AI 芯片产品目录，不仅有助于金融机构根据不同业务场景、不同技术路线来选择合适的芯片产品，更能有效促进国产芯片在研发和应用两方面的正向反馈和良性循环。

5.6 打造覆盖芯片应用全产业链的创新生态系统

一是邀请大型金融机构、有实力的 AI 算法公司专家，组建专家团队和工作组，基于金融行业的适配验证工作，为金融机构在国产化芯片适配、产品选型、AI 应用解决方案等方面提供咨询工作。**二是**组织大型金融机构和 AI 芯片厂商、算力公司等开展场景化落地验证工作。**三是**组织开展不同细分领域竞赛。例如，围绕 AI 芯片计算性能、功耗和性能平衡、模型准确率和精度、异构芯片适配能力、算法优化等方面开展竞赛，全方位提升 AI 芯片领域技术进步与金融行业创新应用。**四是**搭建产金对接平台，针对 AI 芯片适配、异构芯片池化、大模型等主题建立技术对接洽谈或应用展览活动，促进金融机构与产业机构合作和交流。