智能计算芯片技术及产业趋势和对北京的建议

朱 晶[1,2]
(1.北京国际工程咨询有限公司,北京 100055;2.北京半导体行业协会,北京 100191)

摘 要:智能计算芯片是处理海量数据,体现计算能力的硬件载体,是承载数字经济时代生产力的重要支撑。随着人工智能、5G、大数据、区块链等新一代信息技术的规模化应用,行业数字化转型和产业智能化升级对计算资源的需求呈现指数级增长。加上集成电路的发展进入后摩尔时代,现行计算架构(冯·诺依曼架构)缺陷所导致的瓶颈愈加凸显,发展受到严重挑战。而满足多样化计算需求的新型计算架构和计算范式、满足智能计算需求的先进封装技术、基于新材料新工艺的新型计算器件也进入到创新活跃的阶段,智能计算芯片迎来了快速发展的黄金时代。综合分析了数字经济时代智能计算芯片的战略重要性,梳理了国内外智能计算主要产业格局、新兴关键技术以及发展机遇,并对北京发展智能计算芯片的机遇与路径提出相关建议。

关键词:集成电路;智能计算;计算架构;关键新兴技术;数字经济
中图分类号:F062.9;F49 **文献标识码**:A DOI:10.16157/j.issn.0258-7998.222728

中文引用格式:朱晶. 智能计算芯片技术及产业趋势和对北京的建议[J].电子技术应用,2022,48(11):33-40.
英文引用格式:Zhu Jing. Technologies and industry trends of intelligent computing chips and suggestions for Beijing[J]. Application of Electronic Technique,2022,48(11):33-40.

Technologies and industry trends of intelligent computing chips and suggestions for Beijing

Zhu Jing[1,2]
(1.Beijing International Engineering Consulting Company,Beijing 100055,China;
2.Beijing Semiconductor Industry Association,Beijing 100191,China)

Abstract:Intelligent computing chips are hardware carriers that process massive amounts of data and reflect computing capabilities, and are an important support for productivity in the digital economy era. With the large-scale application of new-generation information technologies such as artificial intelligence, 5G, big data, and blockchain, the demand for computing resources in the digital transformation of the industry and the upgrading of industrial intelligence has increased exponentially. In addition, the development of integrated circuits has entered the post-Moore era, and the bottleneck caused by the defects of the current computing architecture(Von Neumann architecture) has become more and more prominent, and the development has been seriously challenged. The integrated computing architecture, new computing paradigms, and new computing devices based on new materials and new processes that meet diverse computing needs have also entered a stage of active innovation, and intelligent computing chips have ushered in a golden age of rapid development. This paper comprehensively analyzes the strategic importance of intelligent computing chips in the digital economy era, the development paths and trends of domestic and foreign technologies, and sorts out the main industrial patterns and development opportunities of intelligent computing at home and abroad.

Key words:integrated circuit;intelligent computing;computing architecture;key emerging technologies;digital economy

0 引言

随着新一轮科技革命和产业变革的深入发展,算力作为数字经济时代新的生产力,已成为全球战略竞争新焦点。美国于2020年11月发布《引领未来先进计算生态系统战略计划》,计划构建覆盖政产学研的国家级算力体系,巩固本国算力优势。欧盟于2021年3月发布"2030数字指南针"计划,日本和澳大利亚等国也纷纷加大算力建设投入,新型算力基础设施已成为多国的重点关注方向[1]。我国也陆续出台了《全国一体化大数据中心协同创新体系算力枢纽实施方案》等多项政策,正加快启动"东数西算"工程,在京津冀、长三角、粤港澳大湾区、成渝、内蒙古、贵州、甘肃、宁夏8地启动建设国家算力枢纽节点,并规划了10个国家数据中心集群,加速构建以算力为核心的新型基础设施体系。由此可见,算力已成为全社会数智化转型的基石,将直接影响数字经济的发展速度,直接决定社会智能的发展高

度，一个以算力为核心生产力的时代加速到来。而无论是人工智能算法的实现、海量数据的获取和存储还是计算能力的体现都离不开计算芯片。北京在计算芯片领域拥有强大的技术积累、丰富的场景资源以及创新的产业生态，因此，快速推动北京智能计算芯片的技术能级提升和产业集群化发展，对北京促进基于智能算力的社会数字化转型、布局前沿计算新技术赛道、推进多样化算力构筑产业健康生态具有重要意义。

1 数字经济时代智能计算芯片的重要战略意义

1.1 智能计算芯片是突破传统计算模式的重要创新

传统计算芯片长期受益于CMOS器件摩尔定律的高速发展以及冯·诺伊曼(Von Neuman)计算范式下内存与计算分开设计的便利性，通过增加集成度以及降低成本取得了巨大的经济效益，也推动了计算科学的不断进步。然而在数字经济和人工智能时代发展的新阶段，一方面摩尔定律放缓使得传统计算芯片性能增长的边际成本迅速上升，另一方面冯·诺伊曼计算范式在爆发性增长的算力需求挑战下，也出现了访存墙、通信墙、功耗墙、可靠性墙等问题。因此，当前智能计算芯片正处于架构、器件和集成创新异常活跃的阶段。架构层面，存储与计算合为一体的存内计算架构[2]、基于软件定义硬件设计理念的粗粒度可重构结构等芯片架构创新层出不穷。器件层面，晶体管结构和材料也在不断创新，包括新型晶体管结构(如FinFET、纳米片Nanosheet/纳米线Nanowire、CFET等)和新型晶体管材料(如高迁移率沟道(HMC)、新型二维材料、CNT等[3])。封装层面，智能计算芯片带动了异构、异质集成的发展[4]。例如硅光异质集成[5]，即利用光的低延迟、低损耗等优异特性，在硅衬底上将电子电路和光子电路集成，形成光电计算体系，解决传统微电子处理器在高速计算应用上的算力、能耗和输入输出瓶颈问题。此外，2.5D/3D堆叠、芯粒(Chiplet)、系统级封装SiP等创新封装技术也因为智能计算芯片而受到关注[6]。由此可见，智能计算芯片的发展快速推动了器件结构更新，计算架构创新和先进封装路径变革，是当前集成电路领域技术创新的典范。

1.2 智能计算芯片是实现我国双碳战略的必须要素

中国明确提出力争2030年前实现碳达峰、2060年前实现碳中和的目标，受到国际社会的高度关注。而过去十年间，我国数据中心整体用电量以每年超10%的速度增加，是我国为数不多能源消耗占社会总用电量比例持续增长的行业。尽管以整机柜、液冷为代表的能源技术能够大幅降低算力中心能耗，推动算力基础设施的绿色低碳可持续发展，但由服务器、存储和网络通信设备等构成的IT设备系统耗能仍是数据中心耗能的最大来源，占比高达45%。因此智能计算芯片的低碳化、低耗能将极大助力数据中心的绿色转型，实现我国双碳战略的达成。近10年来，近阈值电压(NTV)方法被使用在计算芯片的设计上，不断降低芯片功耗[3]。而超低功耗神经网络则是新兴的主要研究方向，模拟计算、存内计算等创新计算架构，CMOS技术与自旋电子器件的混合集成等创新技术正成为实现高性能、超低功耗和高可靠神经网络的关键技术。此外，自供电计算芯片[3]通过直接在计算芯片上集成自供电电路可以做到对芯片“无电池”供电，例如采用太阳能、无线射频信号、摩擦生电等各类能量采集系统，与计算芯片和外围电路集成，形成自供电计算芯片。而随机计算、近似计算[3]等利用了应用所需的精度水平和系统精度之间的差距来实现优化，满足了节能降耗的需求，被认为是新一代的“绿色”创新计算范式。

1.3 智能计算芯片是强化智能算力能级的战略支撑

在数字经济时代，海量数据处理和各种数字化应用都离不开算力的加工和计算，而算力实现的核心则是由计算机、服务器、高性能计算集群和智能终端承载的各类智能计算芯片。可以说，智能计算芯片是数字经济时代的核心生产力，是支撑数字经济发展的坚实基础。长期以来，我国在芯片领域全球竞争中一直处于相对弱势地位，尤其是在计算、存储等领域的关键芯片上，对外依赖度较高，可以说在传统计算架构主导的产业生态中，“中国芯”很难再有所超越[7]。但是，当前人工智能时代的主流计算范式，有望从根本上改变传统计算架构形成的固有产业生态，从而形成全新的竞争格局。这就意味着，在美国仍然以算法和传统计算体系结构的使用为研究重点的时候，中国有机会以开发智能计算芯片技术来开辟一个新的赛道。此外，我国明确提出布局全国算力网络国家枢纽节点，启动实施“东数西算”工程，构建国家算力网络体系，智能计算芯片作为算力网络的基础支撑势必将发挥更为重要的作用。因此，亟需通过大力发展智能计算芯片，加速实现新一代算力生态的供应链国产化，强化我国智能算力能级的快速跃升。

2 全球智能计算芯片产业发展概况

当前世界正在进入以信息产业为主导的经济发展时期，新一代信息通信技术加速创新突破，数据的爆炸式增长、算法复杂度的不断提高，以及应用场景的日益多元化，对智能计算芯片的需求和要求不断升级，全球算力多样化态势日益凸显，创新步伐进一步加快，智能计算芯片已经成为数字经济新引擎和战略竞争新焦点。

2.1 智能计算芯片的内涵与概念

数字经济社会几乎所有智能化信息处理和运算功能都由高性能的智能计算芯片(CPU、GPU、FPGA、各类ASIC以及它们的各种组合)提供支持。但目前还没有对智能计算芯片的统一定义，通常将面向智能计算专用需求的处理器或加速器芯片称为智能计算芯片。为满足在不同场景下的智能化应用需求，各类专门针对智能应用的设计理念和创新架构不断涌现，智能计算芯片逐渐表

现出专用性、多样性的特点。为了支持多样的智能计算任务和性能要求，理想的智能计算芯片需要具备高度并行的处理能力、低内存延迟和创新的架构，以实现计算元件和内存之间灵活而丰富的连接，除此之外，还要考虑功耗和能效管理水平。

目前智能计算芯片主要有以下两种发展路径：

一种是传统通用计算芯片，以及在 CPU、GPU 等传统通用计算芯片的基础上进行多核、异质集成，加速硬件计算能力。例如 CPU 处理器(X86、ARM、RISC-V 开源架构)、图形处理器(GPU)、在 GPU 基础上进行非图形渲染计算的 GPGPU、现场可编程门阵列(FPGA)、数字信号处理芯片(DSP)、微控制器(MCU)、基于融合异构体系结构的众核处理器以及对某类特定算法或者场景进行加速的专用芯片，主要企业包括英特尔、AMD、英伟达、谷歌、亚马逊等。

另一种是颠覆经典的冯·诺依曼计算架构，由新需求和新架构演进、新材料和新器件催生的新型高性能计算芯片及创新计算范式，例如基于 SNN 的类脑芯片、基于忆阻器的存算一体芯片、硅基光电神经网络芯片、受量子原理启发的计算芯片、自进化计算芯片、自供电计算芯片等。主要企业包括 IBM、Graphcore、SambaNova、Mythic、LightMatter 等。

2.2 全球智能计算芯片产业规模

从全球智能计算芯片市场规模来看，如表 1 所示，根据 Gartner 数据，到 2025 年，用于全球各类智能应用领域的智能计算芯片收入预计将达到 762 亿美元，五年年均复合增长率为 28%。其中，用在手机处理器上的智能计算芯片市场规模最大，到 2025 年将接近 285 亿美元，占比 37.3%。

如表 2 所示，从全球智能计算芯片应用场景来看，根据 Gartner 数据，到 2025 年，智能计算芯片的主要应用场景为移动通信和数据中心，分别占到总规模的 42.5%和 35.8%，合计超过 3/4。而近 5 年年均复合增速最快的则为消费电子和工业应用等市场。

2.3 全球智能计算芯片产业基本格局

2.3.1 美国在传统通用计算芯片领域占据绝对优势

当前全球算力规模正以超过 50%的速度增长，CPU、GPU、FPGA、DPU 等传统通用计算芯片仍承担着最为重要的角色。而英特尔、AMD、英伟达等美国企业在传统通用计算芯片领域占据着绝对垄断优势。英特尔、AMD 在全球服务器 CPU 市场占据超过 90%的市场份额，英伟达在全球 GPU 市场更是占据超过 95%的份额。随着异构计算将通过多种计算单元混合协作模式提升计算并行度和效率，在移动互联网、人工智能、云计算等各类典型应用中占比逐步提高，可以看到美国企业将在 CPU+GPU+FPGA、CPU+GPU+DPU 等芯片异构方案上占据更强优势。而目前我国在 CPU、GPU 等高端通用芯片领域的设计能力与国外先进水平仍然差距较大，尽管华为鲲鹏、飞腾、海光等国内 CPU 芯片目前已实现规模应用，也涌现天数智芯、沐曦、壁仞、摩尔线程等 GPU 初创企业，但我国整体在通用计算芯片自主率仍然较低，信创成为主要市场出口。

表 1 全球智能计算芯片细分产品市场规模和增速

智能计算芯片类型	2022 年规模/百万美元	2025 年规模/百万美元	占比/%	5 年年均复合增长率/%
DSP	32	166	0.2	95.4
独立处理器/多媒体应用处理器	14 727	19 049	25.0	12.3
独立图形处理器(GPU)	5 573	9 495	12.5	28.9
FPGA	438	1 770	2.3	72.8
手机 SoC	20 098	28 420	37.3	31.7
独立通用 CPU	3 411	9 491	12.5	53.2
嵌入式通用 CPU	168	684	0.9	76.9
专用计算芯片	1 244	6 683	8.8	82.0
其他	97	451	0.6	88.3
合计	45 788	76 210	100	28.0

表 2 全球智能计算芯片细分应用场景市场规模和增速

应用场景	2022 规模/百万美元	2025 规模/百万美元	占比/%	5 年年均复合增长率/%
汽车	3 401	8 626	11.3	38.3
移动通信	28 202	32 365	42.5	16.4
数据中心	12 948	27 212	35.8	39.6
消费电子	730	4 769	6.3	125.8
工业应用	489	3 112	4.1	106.7
其他	17	125	0.2	110.2
合计	45 788	76 210	100.0	28.0

2.3.2 智能计算芯片代工领域

当前智能计算芯片的制造需要依赖 14 nm 及以下的先进制程，根据 Gartner 数据，2021 年全球 14 nm 及以下代工产能共 480 万片/月，其中中国台湾地区产能 316 万片/月，韩国产能 80 万片/月，因此两者共同占据全球 14 nm 及以下代工产能超过 80%的市场份额。尤其是在 7 nm/5 nm，韩国及台湾地区企业的产能占比接近 100%。此外，台积电、三星两家企业正在向 3 nm/2 nm 更先进制造积极投入和布局。而我国尽管近年来在芯片代工领域取得了不少进展，连年加大资本支出，成功量产 14 nm FinFET 工艺，但目前我国在智能计算芯片代工需要的更先进制程上依然落后国际领先水平 2 代以上。在 14 nm 及以下代工产能上仅有不到全球 5%的供应能力，远远无法满足国内智能计算芯片市场需求。

2.3.3 互联网和 IT 巨头企业成为智能计算芯片领域关键角色

当前，国内外很多此前只涉及软件和互联网业务的科技巨头都拥有了自主研发的智能计算芯片，并且取得了经市场验证的良好效果。例如，谷歌的 TPU 和自身的 TensorFlow、算力平台共同组成了全世界最好的智能计算技术生态；亚马逊 Nitro 系统架构已发展至第四代，自研芯片已有 3 条产品线，包括基于 ARM 架构的通用计算芯片，以及机器学习训练和推理的 AI 专用芯片，形成的综合计算集群比英伟达 T4 降低了 25%延迟和 30%成本。其他诸如微软、Meta、字节跳动、阿里巴巴等全球领先的互联网和 IT 巨头也正在加速入场。未来，随着数字经济推动科技企业业务边界的持续扩展，数据量的激增，人工智能技术的飞速发展，更多互联网和 IT 巨头将更有动力自研出与自身业务、网络拓扑结构和软件体系有强相关性的芯片。此外，随着全球半导体供应链受政治因素影响不确定性更加凸显，进一步保障关键计算芯片的自主供应也会成为系统级厂商选择自研芯片的主要考虑。

2.3.4 新兴计算范式和先进封装驱动智能计算芯片不断创新

当前以突破摩尔定律发展瓶颈为目的的多种新兴技术不断出现，试图通过改进范式、封装、架构、电路和器件技术等，研制出计算能力有较大突破的新兴智能计算芯片。架构创新是目前智能计算芯片技术创新最为活跃的领域，例如存算一体实现在存储单元中进行计算，运用先进封装技术的近存或数模混合的存内计算，突破冯·诺伊曼体系“存”“算”分离的局限，大幅降低数据交换时间以及计算过程中的数据存取的能耗，目前行业内已有诸多商用探索。此外，还包括模拟计算、自进化架构、可重构计算等领域的创新计算架构、光子计算，以及基于RRAM、自旋电子器件的新型计算范式。我国在基于新架构和新范式的专用计算芯片领域拥有比肩国际水平的技术能力，有望通过在新架构新范式上的积极布局，实现智能计算芯片的“换场竞争”。

3 智能计算芯片的新兴关键技术与挑战

数字经济时代，各类智能化场景对高性能计算需求呈爆发式增长，要求智能计算芯片进行并行处理数据，高内存带宽和低功耗、低延迟操作的能力日益提升。然而摩尔定律发展逐渐趋缓，经典计算体系演进模式已经无法满足需求。因此，围绕着智能计算系统算力和能效的提升，众多智能计算芯片的新兴关键技术开始引起学术界和工业界的关注。

3.1 智能计算芯片的新兴关键技术

3.1.1 异质集成封装

异质集成封装作为超越摩尔定律发展的重要手段之一，已从多种不同材料芯片的二维/三维集成发展到同一衬底上集成多种不同材料、不同结构的器件，并实现了不同工艺器件的一体化互连。应用在智能计算芯片上的异质集成封装主要包括高带宽存储器(High Bandwidth Memory，HBM)技术，混合存储立方(Hybrid Memory Cube，HMC)技术，采用 TSV 工艺发挥高密度、立体化互连对提升存储密度、读取速度和带宽的积极影响[8-9]。再比如神经形态器件也需要多层三维集成，以便形成密集立体互连网络，模拟脑部复杂而高度交连的互连与信息传递结构。而芯粒(Chiplet)则是另一类在智能计算芯片领域应用的异质集成封装方案，通过运用内部互联技术融合多个高密度核心逻辑芯片以及相应的 I/O 单元，突破在延时、设计复杂性、经济性方面的瓶颈。例如对于云端人工智能加速场景，CPU 和智能计算芯片的互联以及多片加速芯片间的互联，目前主要通过 PCIe、NVLink 或者直接用高速 SerDes 等实现。如果采用 Chiplet 技术实现片上互联，带宽、延时和功耗都会有巨大的改善。目前 AMD、英特尔等计算芯片大厂都已经实现了 Chiplet 技术的商用[4,6]。

3.1.2 新计算架构和计算范式

数字经济时代的智能计算不会脱离传统计算，但具有新的计算特点。例如数据处理的过程需要很大的计算量(张量处理等线性代数运算)，但大部分场景对计算精度要求反而不高，需要大量的存储和数据搬移，高带宽、低延时的访存能力，以及计算单元和存储器件间丰富且灵活的连接，因此需要计算架构和计算范式的创新来满足各类智能化场景需求。存算一体架构是被认为可以彻底消除冯·诺伊曼计算架构瓶颈，特别适用于神经网络这种大数据量、大规模并行的应用场景的创新架构之一。该架构直接利用存储器进行数据处理，从而把数据存储与计算融合在同一个芯片当中，目前业内已经出现了 SRAM 存算一体、RRAM/PCM/Flash 多值存算一体、RRAM/PCM/MRAM 二值存算一体等多种基于存算一体的深度神经网络实现[2,9-11]。而将存储器结合微纳传感器工艺构建感知、存储、计算一体的相关技术也成为智能计算芯片

新的研究方向。此外，创新的计算架构和范式还包括基于RRAM的近似计算芯片、基于自然仿生算法的计算芯片、自进化AI芯片等，但大部分还处于实验室研究阶段[2]。

3.1.3 新材料和新器件

随着摩尔定律的放缓，由于基础物理原理限制和经济的原因，CMOS工艺和器件持续提高集成密度变得越来越困难。因此，为了提高智能计算芯片的性能和成品率、确保工艺继续向前推进，需要引入新的材料和器件结构。相对于传统的硅(Si)、砷化镓(GaAs)等材料，原子尺度的一维或二维半导体材料呈现出高迁移率、能带可调等优异的物理特性，展现出在新型逻辑器件方面巨大的应用潜力。碳纳米管(CNTs)、黑磷(B-P)和过渡族金属硫属化合物(TMDCs)等新型低维半导体材料，已经被用于半导体晶体管器件的沟道材料并构造了多种新型纳电子器件[12-13]。此外，还包括自旋基逻辑、隧道FET和新材料FET等可替代传统CMOS开关的器件和芯片。光子器件(带宽大、速度快)适用于人工神经网络的优势引起了业界较大关注度。光子计算利用光学器件折射、干涉等特性进行运算，在处理深度学习中大量的矩阵计算的乘加任务时，由于在光子领域实现矩阵运算的基础乘积累加运算(MAC)并不会在本质上消耗能量，因此拥有更高的处理速度和更低的能耗，从而有利于深度学习中的人工神经网络计算速度和性能的提升[14-15]。而基于忆阻器(Memristor)、相变存储器(Phase Change Memory)、铁电器件(Ferroelectric Device)、磁隧道结(Magnetic Tunnel-Junction)、离子栅控晶体管等新型器件的神经形态芯片，从底层器件仿生的角度出发，在器件层面即开始模拟生物的基本信息处理单元——神经元和突触，在功耗、硬件代价等方面也具有显著优势，目前还处于实验室探索和商用早期阶段[16]。

3.1.4 应用驱动的专用加速

在当前智能化应用各领域的算法和应用还处在高速发展和快速迭代的阶段，针对特定领域而不针对特定应用的设计，将是智能计算芯片设计的一个指导原则。专用数据处理器(DPU)就是应用驱动技术路线下的重要产物。DPU最直接的作用是作为数据中心场景中CPU的卸载引擎，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[17]。而基于领域定制加速器(Domain-Specific Accelerator，DSA)，软件定义芯片(Software Defined Chip)理念的可重构计算则被认为是能够根据应用场景和产品需求改变功能，实现了高灵活性的芯片设计。例如粗粒度(粒度是指可重构计算处理器数据通路中运算单元的数据位宽度)动态可重构处理器(CGRA)采用粗计算颗粒度计算单元以及较为精简的互连结构，使得芯片上的功能单元具有可重构能力，能够实现大量应用类型的算法到可重构计算引擎的映射，在通用计算架构的高灵活性和专用架构的高能效性之间取得良好的折中。清华大学微电子学研究所设计的Thinker，即为基于CGRA的可实现计算阵列重构、存储带宽重构、数据位宽重构的智能计算芯片[7,18-19]。

3.2 智能计算芯片新兴关键技术面临的主要挑战

近年来，围绕国内外智能计算芯片领域的关键技术创新极为活跃，但由于上述新兴技术前瞻性强，涉及多学科交叉融合，因此仍然需要克服诸多挑战才能实现大规模的应用。

3.2.1 新兴计算架构、器件和集成创新均不够成熟

尽管当前基于各种新兴关键技术的智能计算芯片创新活跃，但相比于传统通用计算芯片，发展仍然处于初级阶段，在功能、生态完善、工具支持方面都不够成熟。在功能方面，传统通用计算芯片对计算机科学领域特性覆盖较强，主要支持当前主流的各类神经网络实现，而智能计算芯片的覆盖能力相对较弱。在生态方面，传统计算芯片生态发展较为完善，包括丰富的开发框架(如TensorFlow、PyTorch等)、工具链等支持，支持扩展为超大规模算力平台。而很多基于新架构、新材料和器件的智能计算芯片生态发展处于起步阶段，开发框架和工具链正在兴起，目前正在逐渐提升算法部署便捷性和用户友好性等指标。例如Chiplet技术目前在数据互联标准、EDA工具、封装和测试技术上还有待持续完善和优化。异质集成在器件级、芯片晶圆级和子系统级的设计、仿真、加工和验证方面，也需要一套完整的、EDA软件支持的解决方案。

3.2.2 跨产业、跨学科的协同创新及资源组织挑战

异质集成、新计算机构和计算范式、新材料和新器件还有DSA加速等技术领域往往需要跨学科、跨产业链的协同，因此对创新资源组织和合作提出了更高的要求。例如存算一体芯片涉及器件-芯片-算法-应用等多层次的跨层协同，特别是基于新型存储介质的存算一体技术，器件物理原理、行为特性、集成工艺都不尽相同，需要跨层协同来实现性能(精度、功耗、时延等)与成本的最优。再比如Chiplet与2.5/3D封装结合，其内部各个芯粒可能采用的是不同的制程工艺、不同架构，同时还需要加入高速互联总线、接口IP、HBM内存，各个模块可能还需要用到不同的材料进行互联。因此，Chiplet设计的时候，就需要实现设计、封装、工艺和材料各部分的协同，将内部封装的各个模块看成一个整体的系统，需要一开始就要考虑到整个系统层级的设计和优化。

3.2.3 新兴关键技术还存在有待攻克的瓶颈

目前基于各种新兴关键技术的智能计算芯片还存在一些技术限制和技术难点，导致整体发展进度低于预期，并未形成大规模的应用。例如异质集成的工艺还不完善，尤其是单片异质集成，由于异质集成涉及的工艺步骤多且复杂，对工艺精度要求较高，同时对能够异质集成的器件也有一定的约束，因此虽然异质集成目前已

能实现，但尚未进行大规模量产，并且量产后的产品良率还有待进一步验证。还需要对材料的性能、退化和失效机理作进一步研究，以建立有效的、多种物理、多尺度模型来准确预测失效的发生，提高可靠性。另外，忆阻器大规模集成是 RRAM 存算一体芯片应用的前提，但目前制约集成规模的关键基础问题是忆阻器阵列中的串扰和忆阻器阵列的制备工艺。帮助解决忆阻器阵列中串扰问题的高性能选通器件，目前还无成熟的解决方案。而忆阻器阵列制备工艺的均一性、稳定性及其与 CMOS 制造技术的兼容性，目前仍需要相关技术的进步来提升[20]。

4 北京智能计算芯片产业基础、优势与挑战

北京在智能计算芯片领域具有雄厚的产业基础，强大的科研创新积累和前沿技术布局，人才、场景和创新环境都是北京发展智能计算芯片的优势，但同时也存在着诸多发展瓶颈与挑战，例如智能计算芯片新兴技术赛道增量企业不多，缺乏精准的政策支持，与场景的配套衔接有待深入等。

4.1 北京智能计算芯片产业发展基础与优势

4.1.1 技术创新优势

北京在智能计算芯片领域拥有众多前沿技术创新资源，拥有北京大学、北京脑科学与类脑研究中心黄如团队，清华大学类脑计算研究中心施路平、张悠慧、李国齐团队，清华大学魏少军、尹首一团队，清华大学吴华强团队，中科院微电子所刘琦团队，中科院半导体所鲁华祥团队，北京航空航天大学赵巍胜团队等。研究学者数量和研究成果均居于全国领先，在基础理论、关键技术、计算系统、企业布局方面已积累了一定量级的底层储备。并且北京几乎在专用深度学习处理器、存算一体芯片、可重构计算架构芯片、开源 RISC-V 架构芯片、硅基光子计算芯片等所有智能计算芯片的新兴技术路线上都有科研资源进行布局，很多北京智能计算芯片企业的产品直接来源于高校成果转化，例如清微智能的 Thinker 芯片，即为清华微电子所可重构计算架构的产业化成果，此外，灵汐、湃方、知存、光子算数等企业也都孵化于国内外高校及院所的创新成果，因此北京智能计算芯片基于前沿技术创新的群体式跃升条件初步形成。

4.1.2 产业基础优势

北京是国内在智能计算芯片产业领域布局最早的城市，国内首家科创板上市的人工智能芯片企业寒武纪、国内通用处理器 CPU 方面的领军企业龙芯中科、北京君正都诞生于北京。目前北京智能计算芯片产业总体销售收入接近 50 亿元左右，企业数量超过 50 家。此外北京还包括北京智源人工智能研究院、北京微芯区块链与边缘计算研究院等新型研发机构，推动北京成为全球人工智能学术思想、基础理论、顶尖人才、企业创新和发展政策的源头，率先成为国际领先的人工智能创新高地。北京在智能计算芯片领域的产业优势还体现在拥有全面完备的人工智能产业链基础，根据智源研究院数据，北京人工智能相关企业数量超过 1 500 家，占国内人工智能企业总量的 20%，居全国首位。同时北京还是国内拥有开源深度学习框架最多的地区以及中国最大的开发者聚集地，开发者数量约占全国的 20%。北京快速构建具有全球影响力的人工智能产业生态体系将大力推动北京智能计算芯片的发展。

4.1.3 场景应用优势

智能计算芯片是数字经济时代众多智能化场景和应用的基础硬件载体，北京是全球新经济企业最多的城市，也是企业场景创新最活跃的城市，而应用场景建设也是北京推动全国科技创新中心建设的重要举措。

2019 年北京出台了《加快应用场景建设推进首都高质量发展的工作方案》，成立加快应用场景建设统筹联席会议。近两年北京已发布三批共 90 项应用场景建设项目，总投资超过 200 亿元，聚焦医疗健康、城市治理、科技冬奥、政务、交通、教育等重点领域。通过应用场景建设，推动人工智能技术与实体经济深度融合，可以加速孵化培育一批技术领先型的智能计算芯片公司。此外北京集聚了一大批央企国企，互联网和整机公司，都是场景的定义者和芯片的需求方，近年来也纷纷通过自研、控股或投资等方式，布局智能计算芯片这个赛道。

4.2 北京智能计算芯片产业面临的挑战

4.2.1 创业新赛道企业数量减少，“独角兽”明显断档

近年来北京在计算类芯片领域新增优质企业数量不多，依然依赖寒武纪、地平线等存量“独角兽”企业。尤其在新的智能计算芯片赛道，例如基于 RISC-V 开源架构的计算芯片，GPU 图形处理器，DPU、ARM CPU，存算一体芯片等，自去年起，资本加速向这些领域集聚，国内相继涌现出一众玩家，单轮融资纪录不断被刷新，但基本总部都在上海，而选择在北京设立子公司。目前在 ARM CPU 和 GPU 等计算“大芯片”赛道，总部设立在北京的企业均只有一家，其他如天数智芯、壁仞科技、沐曦集成电路、登临科技等企业，总部均在上海。DPU、存算一体芯片领域的新创企业也都选择在京外设立总部，由此可见，北京在智能计算领域对企业的吸引力严重弱化。

4.2.2 缺乏精准政策支持

智能计算芯片属于技术人才密集和资金密集的领域。进入到人工智能、区块链等要求高算力的场景时代，算法迭代速度加快，传统计算架构出现瓶颈，基于新架构新范式的计算芯片创新空前活跃，因此对企业在技术和人才领域的要求极高。此外，智能计算芯片的制造工艺节点普遍在 7 nm 以下，所需要的设计研发支出和流片费用都较大(7 nm FinFET 工艺流片费用约 3 000 万美元，合 2 亿元人民币)，同时由于计算类芯片的迭代周期较快，为保持技术前瞻性、领先性和核心竞争优势，企业必须持续进行研发投入。因此，“毛毛雨”般的补贴额度

对于计算芯片企业而言非常“不解渴”。尽管我市自去年开始在流片补贴上的力度逐步加大，但政策的精准性依然不强，政策的特色化依然不足。目前北京数十家计算芯片企业中在北京享受政策补贴的不超过 10 家，寒武纪、地平线等头部企业多年来一直不在北京享受流片补贴政策，而转到上海、南京等地。

4.2.3 智能计算软硬件企业合作和资源对接力度不足

万物智联时代将会带来越来越多的计算下沉到边缘和终端，场景将会更复杂，计算需要更高效，响应需要更快速，就需要智能计算芯片在保障低成本的前提下，能够支持多模态，保障低时延、高能效，并且具备较高的安全性。因此智能计算芯片必须从场景出发，通过系统了解行业需求，探索新型的计算架构，设计定制化的芯片架构，在大幅提升性能的同时，降低功耗和成本，同时满足人工智能算力以及跨设备形态的需求，并通过服务赋能其他行业。因此智能计算芯片企业若是和软件及拥有场景的整机企业合作，将能加速提升产品的竞争力。目前北京的众多智能计算芯片企业鲜少与本地的人工智能算法、互联网、整机厂商进行创新协同合作，不利于北京在智能计算芯片领域形成持续化的引领优势。

4.2.4 先进工艺与封装环节仍被“卡脖子”

随着智能计算芯片的逐步演进和迭代，先进工艺和先进封装成为持续优化芯片性能和成本的关键创新路径。尤其是 2.5D/3D、SiP 系统级封装、芯粒 Chiplet 等先进封装和异质集成技术已成为智能计算芯片兼顾更高性能和更高灵活性的必要选择。但目前北京在 14 nm 以下先进工艺产能和先进封装上的供给仍然缺失，无法支撑在京智能计算芯片企业的供应链自主创新，随时面临“卡脖子”风险。尤其是在封装领域，北京目前缺乏在先进封装积极布局的头部企业，相比长三角地区明显弱势，不利于北京在智能计算芯片领域构建价值闭环，从而形成协同联动的完整产业生态体系。

5 结论

智能计算芯片在数字经济领域和“东数西算”工程中的战略和基础地位启示我们，北京应该充分发挥自身在前沿创新、高端人才集聚、算法和应用场景资源众多等优势，找准在智能计算芯片领域的着力点和突破点，积极布局关键产品和新兴技术赛道，突破更多智能计算芯片领域的技术短板，建立起全产业链先进产能供应能力，实现北京智能计算产业整体跃升。

一是建议设立北京“智能计算芯片”资金专项，支持企业在智能计算芯片的体系架构和新兴计算范式、先进工艺制造、先进封装、芯片与算法、场景的联动合作等方面取得变革性突破，探索与产业需求和实际应用相结合的支持方式和组织模式，形成北京全面且长期的智能计算芯片发展战略。

二是建议抢抓技术变革的制高点，在京加快布局可能改变“竞争赛道”的新兴技术赛道。重点支持存内/近存计算、可重构计算、材料和器件创新、异质集成创新、应用驱动的专用加速芯片等关键技术领域，引导高校科研院所与企业面向前沿技术联合攻关，以加强新兴技术赛道在关键技术、验证、工艺、测试等核心环节的全面储备，推动这些前沿技术尽快与现有市场需求对接，加快产业化进程，促进新技术和新生态的培育壮大。

三是强化北京智能计算芯片产业链自主创新和与场景深度衔接的协同创新。重点支持满足智能计算芯片需求的国产 EDA、先进工艺制造、先进封装企业在京布局，强化智能计算芯片供应链的自主创新。结合“东数西算”工程的实施，鼓励在京互联网和系统厂商与智能计算芯片企业结成应用场景“联合体”，为推动企业特别是中小企业技术创新应用提供更多“高含金量”场景条件，加强新技术应用示范，推动创新资源聚合，带动产业深度融合发展。

参考文献

[1] 中国移动通信集团有限公司.中国移动算力网络白皮书[R].2021.
[2] 李雅琪，温晓君.存算一体化的发展现状挑战与对策建议[J].互联网经济，2020(4)：15-17.
[3] 张臣雄.AI 芯片：前沿技术与创新未来[M].北京：人民邮电出版社，2021.
[4] 缪旻，金玉丰.微系统集成全新阶段——IC 芯片与电子集成封装的融合发展[J].微电子学与计算机，2021(1)：1-6.
[5] 杨晓丽，夏瑾，张欢，等.深度剖析光电子技术和产业现状[EB/OL].(2019-05-27)[2022-03-10].http://www.elec-fans.com/d/942690.html.
[6] 张墅野，李振锋，何鹏.微系统三维异质异构集成研究进展[J].电子与封装，2021(10)：73-83.
[7] 尹首一，郭珩，魏少军.人工智能芯片发展的现状及趋势[J].科技导报，2018，36(17)：45-51.
[8] 姚玉良，钱宇，施得君.HBM 和 HMC 技术研究[J].高性能计算技术，2015(2)：8-12.
[9] 李双辰，谢源.计算存储一体化智能芯片[J].中国计算机学会通讯，2018(2)：16-19.
[10] 刘琦.存算一体:超越“存储墙”的计算架构突破[J].前沿科学，2018(4)：66-70.
[11] 郭昕婕，王绍迪.端侧智能存算一体芯片概述[J].微纳电子与智能制造，2019(2)：72-82.
[12] 秦敬凯，甄良，徐成彦.后摩尔时代晶体管:新兴材料与尺寸极限[J].自然杂志，2020(3)：221-230.
[13] 朱进宇，闫峥，苑乔，等.集成电路技术领域最新进展及新技术展望[J].微电子学，2020(2)：219-226.
[14] 白冰，裴丽，左晓燕.用于人工智能的硅基光电子芯片[J].中兴通讯技术，2021(1)：77-82.
[15] 周治平，徐鹏飞，董晓文.硅基光电计算[J].中国激光，2020，47(6)：1-15.

[16] 王宗巍,杨玉超,蔡一茂,等.面向神经形态计算的智能芯片与器件技术[J].中国科学基金,2019(6):656-662.
[17] 鄢贵海.DPU:以数据为中心的专用处理器[J].中国计算机学会通讯,2021,17(10).
[18] 辛思达.面向粗粒度动态可重构处理器的通用领域算法实现与优化[D].长沙:国防科学技术大学,2015.
[19] 魏少军,李兆石,朱建峰,等.可重构计算:软件可定义的计算引擎[J].中国科学(信息科学),2020,50(9):1407-1426.
[20] 李锟,曹荣荣,孙毅,等.基于忆阻器的感存算一体技术研究进展[J].微纳电子与智能制造,2019(4):87-102.

(收稿日期:2022-03-10)

作者简介:
朱晶(1984-),女,硕士,高级经济师,主要研究方向:集成电路、新一代信息技术、技术经济学。

扫码下载电子文档

(上接第 32 页)

Recognition,2019.
[20] STANDARD A P I.Welded tanks for oil storage[S].2013.
[21] He Yihui,Zhang Xiangyu,Sun Jian.Channel pruning for accelerating very deep neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2017.
[22] ARTHUR D,VASSILVITSKII S.K-means++:the advantages of careful seeding[C]//SODA'07,2006.
[23] KRISHNA K,MURTY M N.Genetic K-means algorithm[J].IEEE Transactions on Systems,Man,and Cybernetics,Part B(Cybernetics),1999,29(3):433-439.
[24] REZATOFIGHI H,TSOI N,GWAK J,et al.Generalized intersection over union:a metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019.
[25] Zheng Zhaohui,Wang Ping,Liu Wei,et al.Distance-IoU loss:faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence,2020.
[26] VAN ETTEN A.Satellite imagery multiscale rapid detection with windowed networks[C]//2019 IEEE Winter Conference on Applications of Computer Vision(WACV),IEEE,2019.
[27] GARY B,KAEHLER A.Learning OpenCV:computer vision with the OpenCV library[M].O'Reilly Media,2008.
[28] CHENG G,HAN J,ZHOU P,et al.Multi-class geospatial object detection and geographic image classification based on collection of part detectors[J].ISPRS Journal of Photogrammetry and Remote Sensing,2014,98(dec.):119-132.
[29] CHENG G,HAN J.A survey on object detection in optical remote sensing images[J].ISPRS Journal of Photogrammetry & Remote Sensing,2016,117:11-28.
[30] CHENG G,ZHOU P,HAN J.Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images[J].IEEE Transactions on Geoence and Remote Sensing,2016,54(12):7405-7415.
[31] Xia Guisong,Bai Xiang,Ding Jian,et al.DOTA:a large-scale dataset for object detection in aerial images[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2018.
[32] Xia Guisong,Bai Xiang,Ding Jian,et al.DOTA:a lorge-scale dataset for object detection in aerial images[C]//The IEEE Conference on Computer Vision and Pattern Recognition(CVPR),2018.
[33] LONG Y,GONG Y,XIAO Z,et al.Accurate object localization in remote sensing images based on convolutional neural networks[J].IEEE Transactions on Geoence and Remote Sensing,2017(5):1-13.
[34] XIAO Z,LIU Q,TANG G,et al.Elliptic fourier transformation-based histograms of oriented gradients for rotationally invariant object detection in remote-sensing images[J].International Journal of Remote Sensing,2015,36(2):618-644.
[35] Xia Guisong,Hu Jingwen,Hu Fan,et al.AID:a benchmark data set for performance evaluation of aerial scene classification[J].IEEE Transactions on Geoscience and Remote Sensing,2017,55(7):3965-3981.
[36] Tan Mingxing,Pang Ruoming,LE Q V.Efficientdet:scalable and efficient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2020.
[37] He Yihui,Zhang Xiangyu,Sun Jian.Channel pruning for accelerating very deep neural networks[C]//Proceedings of the IEEE International Conference on Computer Vision,2017.

(收稿日期:2022-07-15)

作者简介:
李想(1992-),男,硕士研究生,工程师,主要研究方向:深度学习、图像处理。
特日根(1987-),男,博士研究生,副研究员,主要研究方向:深度学习、大数据分析。
赵宇恒(1992-),男,硕士研究生,实习研究员,主要研究方向:深度学习、图像处理。

扫码下载电子文档

版权声明