中国科学: 信息科学 2024 年 第 54 卷 第 1 期: 34–47
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
集成电路未来发展及关键问题观点专题

高能效高安全新兴计算芯片: 现状、挑战与展望

刘伟强[1,2*†], 陈珂[1,2†], 吴比[1,2], 邓尔雅[1,2], 王佑[1,2], 龚宇[1,2], 崔益军[1,2], 王成华[1,2]

1. 南京航空航天大学集成电路学院, 南京 211106
2. 空天集成电路与微系统工业和信息化部重点实验室, 南京 211106
* 通信作者. E-mail: liuweiqiang@nuaa.edu.cn
† 同等贡献

收稿日期: 2023–10–23; 接受日期: 2023–12–07; 网络出版日期: 2024–01–03

国家自然科学基金 (批准号: 92364201, 62022041, 62101252, 62371226, 62201254, 62304107, 62104107) 资助项目

摘要 智能信息化社会对算力的需求日益增长, 高能效和高安全性的计算芯片已经成为支撑科技创新和社会进步不可或缺的基础设施. 新兴计算范式作为提升算力的创新技术, 近年来在理论和技术方面取得了重要突破, 引起了学术界和工业界广泛关注. 本文从电路设计方法、新型芯片架构以及脑启发算法等多个角度介绍和分析了新兴计算芯片的相关前沿技术, 同时讨论了各项技术的阶段性特征以及所面临的设计挑战和安全可信挑战, 最后展望了新兴计算芯片技术的未来发展, 并阐述了其发展的重点方向.

关键词 新兴计算范式, 安全可信, 近似计算, 随机计算, 存内计算, 脑启发式计算

1 引言

计算能力已成为人类文明进步的关键引擎之一, 加速了科技和工程的发展, 改变了人们的生活、商业和社会互动方式, 为解决复杂的全球挑战提供了重要的工具和资源. 在 2023 年 10 月, 工业和信息化部等六部门联合印发了《算力基础设施高质量发展行动计划》, 该计划认为 “集信息计算力、网络运载力、数据存储力为一体的算力作为新型生产力, 呈现多元泛在、智能敏捷、安全可靠、绿色低碳等特征, 对于助推产业转型升级、赋能科技创新进步、满足人民美好生活需要和实现社会高效能治理具有重要意义”.

伴随信息技术的不断进步, 以大型预训练模型 ChatGPT 为代表的新兴应用发展对算力需求日益增大, 有研究表明人类社会在 2020 年需要处理的信息已达 10^{37} bit, 在 2050 年预计将达到 10^{46} bit [1].

引用格式: 刘伟强, 陈珂, 吴比, 等. 高能效高安全新兴计算芯片: 现状、挑战与展望. 中国科学: 信息科学, 2024, 54: 34–47, doi: 10.1360/SSI-2023-0316
Liu W Q, Chen K, Wu B, et al. High-efficiency and high-security emerging computing chips: development, challenges, and prospects (in Chinese). Sci Sin Inform, 2024, 54: 34–47, doi: 10.1360/SSI-2023-0316

据国际数据公司 IDC 估计, 2022 年中国智能算力规模达到 268.0 EFLOPS (每秒百亿亿次浮点运算), 超过通用算力规模, 预计到 2026 年智能算力规模将达到 1271.4 EFLOPS. 2021∼2026 年, 预计中国智能算力规模年复合增长率达 52.3%, 同期通用算力规模年复合增长率为 18.5%. 华为预测未来 10 年人工智能算力需求将会增长 500 倍以上 [2].

面对这一趋势, 传统基于工艺红利的设计方法已经无法满足迅猛增长的算力需求. 此外, 集成电路产业正面临着包括工艺演进趋缓, 国际形势变化导致产业链波动等因素造成的一系列复杂挑战. 根据美国半导体研究公司发布的《半导体十年计划》报告, 全球通用计算设备的能源消耗正以指数级增长, 而全球能源生产仅以每年 2% 的速率增长. 在应对气候变化和我国的 “双碳” 政策的大背景下, 不断增长的计算能源需求正在带来新的风险. 此外, 高度互联互通的信息系统与应用对硬件的安全性提出了更高的要求. 物联网、社交平台、无人设备等新兴技术的大规模部署必须基于安全可信的信息通信与处理, 但令人担忧的是针对网络与计算系统的各类攻击在时时刻刻威胁着数据信息安全, 危害国家安全和社会经济稳定.

面临一系列风险挑战, 新兴计算范式作为设计方法学的演进产物, 在传统数字 – 模拟计算的基础上, 通过新的算法、架构以及电路设计范式进一步挖掘现有半导体器件的性能潜力, 成为提高计算能效重要的技术手段. 如图 1 所示, 本文结合目前学术界与工业界在新兴计算领域的最新研究, 从基本电路、芯片架构和算法设计等方面对新兴计算芯片设计的发展现状、设计与安全可信挑战及未来趋势等方面进行介绍和讨论.

2 高能效新兴计算的发展现状

随着 Dennard 缩放定律的失效及后摩尔时代的来临, 计算芯片难以从 CMOS 先进制程中得到更多的算力与能效提升. 然而新型半导体器件的研发仍然不够成熟, 难以满足当前工业界的需求. 面临这一现状, 如何从相对成熟的半导体技术中进一步发掘其算力承载潜力是解决现阶段算力发展瓶颈的重要发展方向. 本节从电路设计、芯片架构以及算法技术等方面介绍相关技术与其发展现状.

2.1 新兴可容错计算电路

在以大数据和智能计算为代表的新兴应用中, 通常不需要唯一精确的 “黄金” 结果, 一个能够被应用接受得足够好的结果即可满足系统要求. 另外在经典的 DSP 场景中, 往往包含容错与纠错机制. 因此在电路设计层面, 可以通过引入计算误差进行可容错计算, 从而降低电路的能耗, 加快处理速度. 其中近似计算 (approximate computing) 与随机计算 (stochastic computing) 是这一类型的新兴电路设计范式的典型代表.

2.1.1 近似计算

大数据、人工智能等计算场景中存在大量的数据与计算冗余. 相关文献统计发现当前在数据挖掘、计算机视觉、模式识别、智能通信等典型的 12 类大算力需求场景中, 有着平均 80% 的可容错执行时间比例. 因此相较于 “完全精确”, “算得快” 及 “足够准” 成为大规模计算系统的主要追求目标. 如图 2(a) 所示, 近似计算区别于传统的计算范式, 采用非精确的系统设计和电路结构, 以可容忍的计算精度损失换取硬件性能、能效以及面积效率等指标的大幅提升, 成为大算力需求下重要的范式之一. 合理部署近似计算, 既可以大幅减少算法运算规模及存储需求, 又可以维持足够好的计算精度, 以满足实际应用的需求 [3]. 此外, 大算力所面临的 “存储墙” 的能耗, 也有大量工作通过近阈值存储结构提高

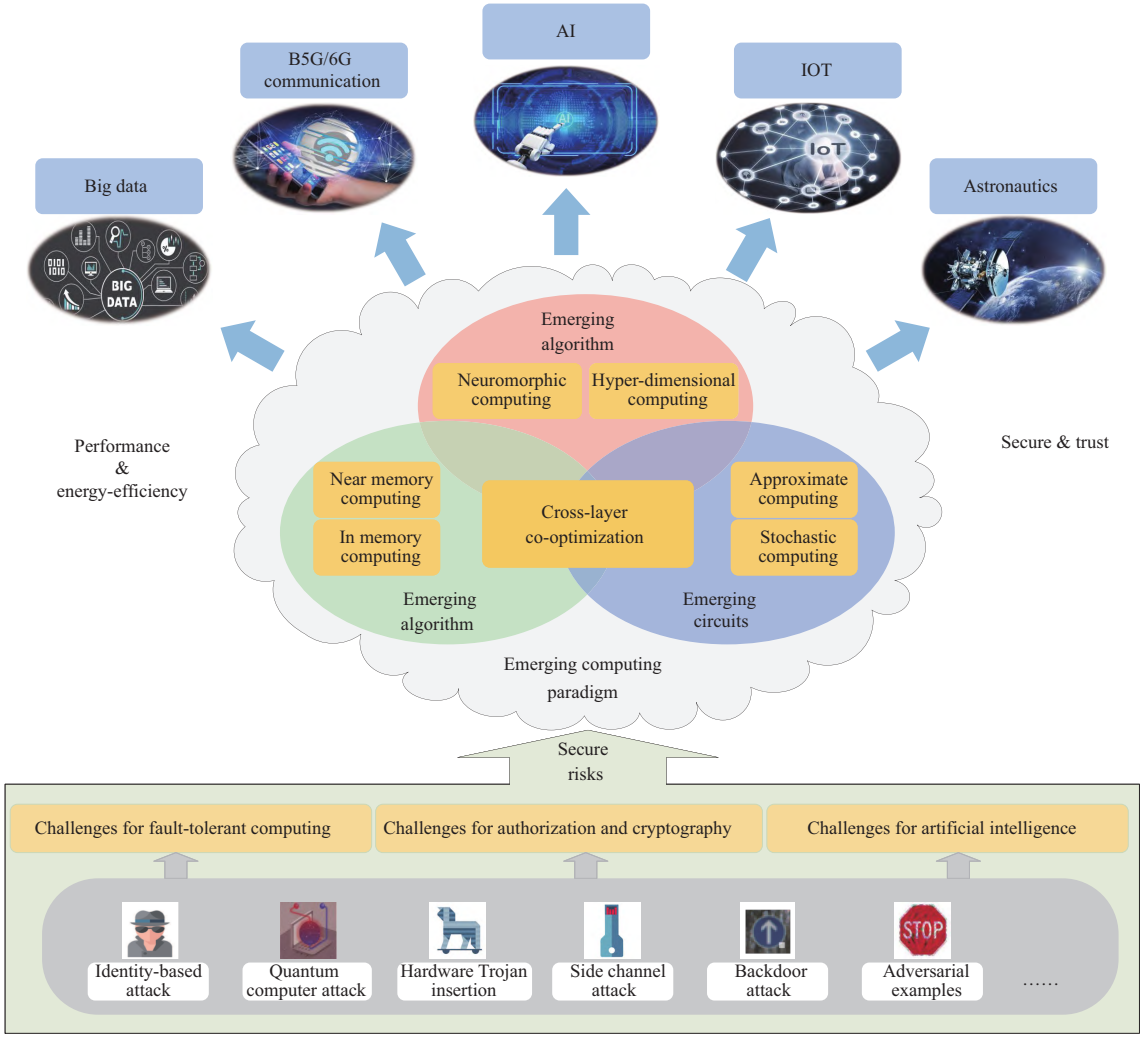

图 1　新兴计算范式及其安全挑战

Figure 1　Emerging computing paradigms and their secure challenges

能效, 使 SRAM 工作在近阈值电压, 存取 “几乎相同” 的数据, 提供 “可接受的” 数据处理精度. 此外在存内计算嵌入近似电路等逻辑, 也成为扩充存内计算场景的重要实现范式 [4], 这也为极低功耗的边缘计算芯片扩展了新的设计思路.

当前, 从专用计算到通用计算都可以看到近似计算的应用范例. 智能处理器中大量采用精度可调的近似设计, 如 TPU、TrueNorth、Cambric、含光 800 系列处理器等, 通过采用多位宽可控或模拟计算等方式, 使其计算精度损失在可控范围的同时大幅提高能效 [5]; 在 CPU, GPGPU 等通用计算架构中, 近似计算也已成为提高 Cache 命中率的重要手段, 通过近似预测技术可以大幅提高 Cache 命中率、指令复用率 [6], 此外, 在微架构中引入低电压计算, 平均减少 30% 能耗, 实现同工艺下的能效的代际提升 [7]. 另外在通信领域, 近似计算技术可使数字信号处理电路得到大幅的能效提升 [8].

为获得较好的电路实现结果, 在计算系统设计中引入近似计算往往需要权衡多个设计层次. 从算法设计、架构优化、单元设计、存储结构、器件工艺等层面均有不同的近似计算实现方案, 同时对近似计算系统的设计方法也需要相应的创新优化 [9], 以在增加设计维度的同时, 提高设计效率.

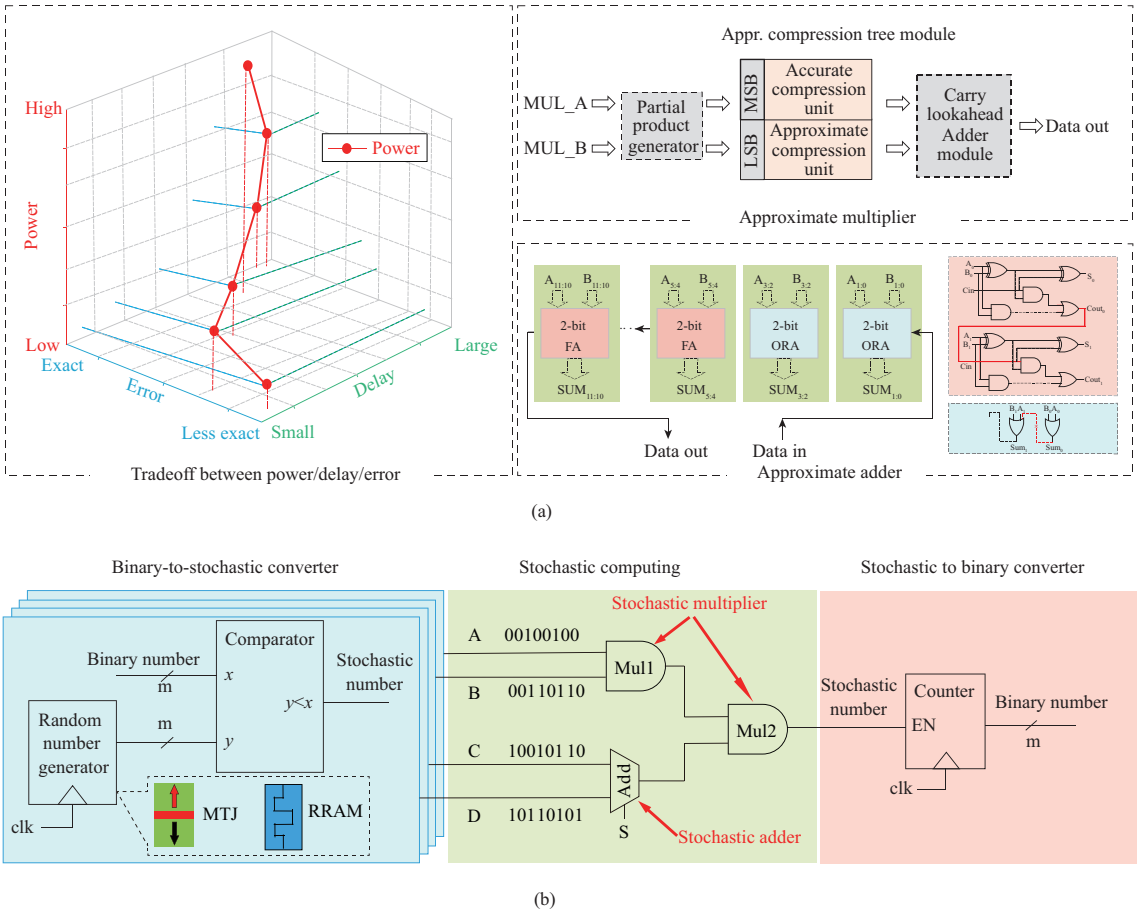

图 2 新兴可容错计算电路示例. (a) 近似计算电路; (b) 随机计算电路

Figure 2 Emerging fault-tolerant computing circuits. (a) Approximate computing circuits; (b) stochastic computing circuits

2.1.2 随机计算

随着晶体管尺寸的不断微缩, 因工艺偏差、热噪声等引起的数字集成电路计算出错的概率日益增大, 探索发展具有高容错性的新型计算范式成为了当前集成电路设计领域的前沿热点研究 [10]. 与传统的二进制计算相比, 随机计算以数字化概率的形式表示和处理信息, 从而简化运算单元, 具有低硬件开销、低功耗和高容错性等优点. 如图 2(b) 所示, 经典的随机计算电路实现方案中, 需要先把输入数据通过一个随机序列生成电路转变为随机数序列, 然后输入随机计算单元进行相应的计算. 最后, 表示结果的随机序列通过一个计数器转换为二进制数后再输出. 其中, 随机计算单元采用随机序列代表数据进行计算, 单比特数据的跳变对计算结果的影响几乎可以忽略, 并且 0 到 1 的跳变和 1 到 0 的跳变概率是相等的, 可以相互抵消, 因此容错性强, 非常适合用于超高温、超低温和强辐射等极端环境中, 在错误校验码、图像处理、量子计算和神经网络等前沿领域具有良好的应用前景 [11,12]. 另外, 随机计算使用非常低复杂度的算术单元, 如加法电路只需要一个多路选择器即可实现, 乘法电路只需要一个与门即可实现, 因此在硬件开销和功耗方面具有一定的优势.

相对于随机计算中简单的算术单元, 其随机序列生成单元占用了高达 80% 的电路面积, 弱化了随机计算的面积优势 [13]. 因此, 设计实现低硬件成本的随机序列生成电路成为该研究领域内面临的一

个关键核心问题. 随机计算的计算精度可以通过增加输入随机序列长度来提高, 但是随着序列长度的增加, 在不增加随机序列生成单元的前提下, 电路系统的延迟和功耗都会增加. 因此, 设计实现高吞吐率的随机数发生器成为了提高随机计算精度的主要途径. 衡量随机数发生器的性能一般用吞吐率、硬件开销和能耗 3 个指标. 目前大多数随机计算的电路设计主要集中在随机数产生机制的创新上. 特别是最近大量涌现的低功耗存储器件, 如自旋电子器件、忆阻器等阻性器件, 其自身固有的随机性可用于产生高质量的随机数, 而且其固有的非易失性又能降低电路能耗, 从而全面提高随机数发生电路的整体性能, 已经成为 "后摩尔时代" 集成电路的重要发展方向之一 [14, 15]. 基于新型低功耗器件的随机计算有效地提高了随机数转换效率, 同时极大地降低了能耗和硬件开销.

由于随机计算在加法和乘法方面的优势, 特别适合用于主要由加法和乘法组成的深度学习硬件实现中. 目前, 已有大量研究探索随机计算在神经网络计算中的应用 [16, 17], 在相同学习效率和精度下展现出比传统二进制逻辑更优的面积开销、延迟和功耗性能. 随机计算在人工智能的应用仍有待深入探索.

2.2 新兴存内计算芯片架构

一直以来, 计算机内存在延迟和能耗方面一直跟不上处理器技术的发展, 即所谓的 "存储墙". 一直以来, 研究人员试图通过引入多层次缓存架构、拓宽存储访问接口等方式来减轻访问片外动态随机存储器 (dynamic random access memory, DRAM) 带来的巨大开销. 然而, 一方面处理器片上可承载的缓存容量无法进一步提升; 另一方面, 人工智能等数据密集型应用带来了更加频繁的数据搬移需求, 迫使研究人员不得不把目光转向对 CPU 和存储器逻辑关系的思考. 存内计算 (computing in memory) 突破冯 · 诺依曼 (von Neumann) 架构, 直接在存储器中进行计算操作. 由于数据不再需要被传输到存储器外, 大大减少了数据搬运产生的开销, 能有效解决访存带来的能效问题. 此外, 基于存储阵列的结构特点, 可实现大规模并行计算以提高算力与能效. 如图 3 所示, 目前的存内计算架构研究分为两种主要的技术路径, 即近存计算与模拟存算.

2.2.1 近存计算架构

在存内计算领域, 最直观的想法就是把存储器放置在 CPU 可以快速访问的位置, 随之衍生出近存计算 (near memory computing). 近存计算将 CPU 与大容量 DRAM 存储直接进行一体化封装, 通过内部总线互联的形式实现更快速的访存. 这种思路为存储体系结构的演进提供了新的方案, 结合 DRAM 的密度优势以及先进封装工艺, 在一定程度上缓解了 "存储墙" 问题. 然而, 随着存储需求的增长, 容量、面积和性能之间的矛盾很快使这一思路也遇到了瓶颈. 而随着 2.5D/3D 集成技术的发展, 近存计算也转向了在垂直方向上进行拓展的思路. 比较典型的 2.5D/3D 集成案例就是将存储切块垂直堆叠在计算单元层之上, 通过穿硅通孔技术进行垂直相连. 这样一来, 底层的计算单元就可以更快的速度访问到上层的存储模块. 同时, 垂直方向上的堆叠所带来的互联线长度增长速率远远低于平面上的存储扩展方案. 目前, 三星集团提出的 PIM-HBM 已经正式商业化, 其在 2.5D 堆叠技术基础上将多层 DRAM 芯片进行堆叠, 并在部分 DRAM 芯片的存储子阵列 (bank) 级 I/O (input/output) 灵敏放大器处集成浮点乘加运算单元, 利用 Bank 级并行激活提高计算吞吐量.

但从本质上看, 近存计算并未改变数据存储和处理之间物理隔离的现状, 是对应用导向的 "存储墙" 和 "功耗墙" 问题的弱化, 其计算吞吐量虽然在一定程度上得到提高, 但由于受到集成的运算单元数据的位宽限制, 最终需要面对的还是性能和容量之间的鸿沟. 所以, 近存计算架构可以认为是冯 · 诺依曼架构到存内计算架构之间的一种过渡技术.

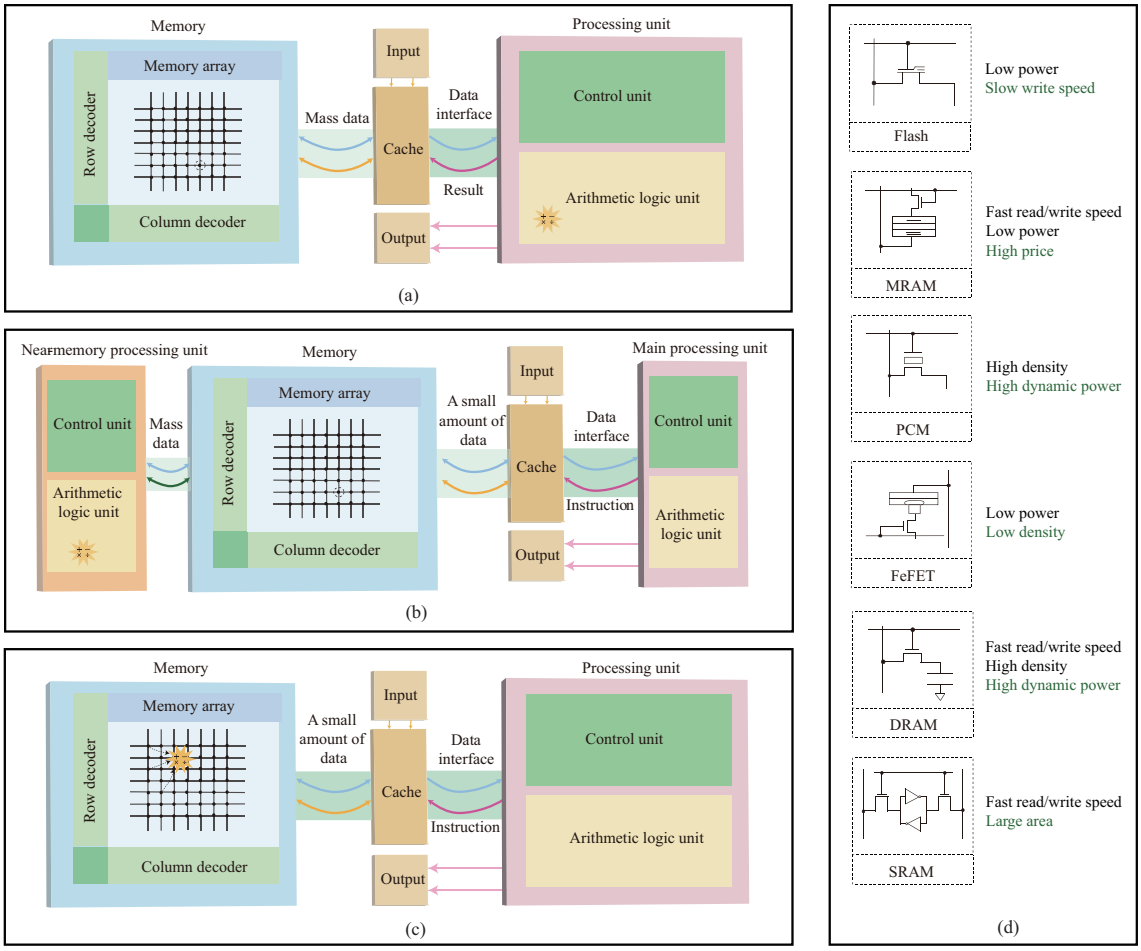

图 3 存内计算. (a) 传统冯 · 诺依曼架构示意图; (b) 近存计算架构示意图; (c) 模拟存内计算架构示意图; (d) 各种器件构成的存内计算单元

Figure 3 In-memory computing. (a) Traditional von Neumann architecture; (b) near memory architecture; (c) analog in-memory computing; (d) memory cells with different devices for INC

2.2.2 模拟存内计算架构

基于交叉存储阵列的模拟存内计算架构主要基于物理定律 (欧姆定律 (Ohm's law) 和基尔霍夫定律 (Kirchhoff laws)), 在存算阵列上实现乘加运算. 结合人工智能场景, 卷积神经网络成为存内计算研究的核心, 从能效、性能、可靠性、安全等方面展开了大量的研究. 此外, 存内计算在机器学习、语音识别、图神经网络、深度学习训练等人工智能应用上, 也表现出诸多优势. 为了提高计算能效, 研究人员从存内计算阵列设计、ADC 外围电路优化、面向存内计算的模型结构设计展开探索. 然而, 存内计算的工艺和器件依旧处于研究阶段, 在大规模集成电路生产上面临较严重的工艺偏差问题. 针对此问题, 从噪声感知的神经网络训练算法、编码、译码、冗余计算等方面开展研究.

近年来, 基于各类存储介质的存内计算芯片得到学术界和产业界的广泛关注和重点研发. 2018 年, DARPA 发布了一个 "电子复兴计划", 目标是研究未来 5~10 年内颠覆性微电子技术, 其中专门成立了一个 "智能存储与存内计算" 中心, 并重点支持了两个基于自旋电子器件的存内计算芯片项目 (共 1420 万美金). 2021 年, 三星、东芝、台积电、IBM、格罗方德、欧洲微电子研究中心 (Interuniversity

Microelectronics Centre, IMEC) 等国际半导体巨头以及 Mythic、Syntiant、知存科技等初创企业, 均发布存内计算原型器件和芯片, 但离大规模量产还有一定距离. 总体来说, 存内计算架构有望打破冯·诺依曼架构瓶颈, 可为物联网、大数据和人工智能等具有海量数据特征的智能应用场景提供高能效硬件解决方案.

根据掉电后能否保存数据, 当前存内计算单元主要分为易失性存储器 (如 DRAM, SRAM 等) 和非易失性存储器 (如阻变随机存储器 ReRAM、磁性随机存储器 MRAM 等). SRAM 工艺成熟、存储数据无须刷新、存取速度快、扩展性好; 但是属于易失性存储器 (掉电数据丢失), 且集成度低、成本较高、功耗较大, 难以通过较低成本实现大规模大算力存内计算芯片 [18]. DRAM 工艺成熟、单元面积较小; 但同样属于易失性存储器, 需定期刷新, 且存在漏电问题, 难以实现高精度存内计算芯片 [19]. Flash 是非易失性存储器; 但读写延迟较大, 难以实现高性能计算, 且扩展性较差 (难以微缩至 28 nm 以下). ReRAM 是新型非易失性存储器, 且能够实现大规模交叉点阵列, 是未来实现存内计算芯片的潜力介质之一; 但是目前的工艺还不太成熟, 一致性较差. PCM 是非易失性存储器, 也能够实现大规模交叉点阵列; 但是功耗较大、速度较慢、耐久性较差 [20]. 铁电晶体管 FeFET 可实现非易失性存储, 也能实现交叉点阵列; 但是目前的工艺也不太成熟. MRAM 是非易失性存储器, 具有高耐久性、高速度、低功耗等优点, 工艺相对较成熟, 扩展性较好, 目前的第一代 (Toggle-MRAM) 和第二代 (STT-MRAM) 在国外已实现量产, 但是器件的阻值 (约几千欧姆) 与高低阻值比率 (约 250%) 相对较小, 在实现多比特存内计算芯片方面具有一定挑战 [21, 22].

2.3 新兴脑启发式计算算法

对比传统的半导体数字 – 模拟计算模式, 基于神经元的生物智能的工作模式截然不同. 虽然传统的数字 – 模拟计算在算术运算等方面其性能已远超人脑, 但在感知、识别、分析、推演等场景中人脑的能力依然优秀. 另外, 生物智能经过漫长的进化与学习训练过程, 其能量效率有目前半导体技术计算系统不可比拟的巨大优势. 因此, 生物智能的工作模式是启发电路设计者探索高能效计算的一个重要方向.

2.3.1 神经拟态计算与脉冲神经网络

随着人工神经网络 (artificial neural network, ANN) 结构复杂度不断增加, 训练神经网络所需的大量标注数据和严苛的算力要求在一定程度上限制了人工神经网络的持续发展. 因此, 受脑科学启发的类脑智能研究成为了有望推动下一代人工智能技术和新型信息产业的发展的新方向. 而其中的脉冲神经网络 (spiking neural network, SNN) 则是类脑智能研究的核心, 它是一种使用基于事件驱动的稀疏计算来进行信息处理的人工神经网络. 区别于传统 ANN, SNN 除了神经元和突触状态之外, 还将时间概念纳入了其操作之中: 通过将网络的输入信息编码为脉冲序列信号, 并保持脉冲的时序关系输入神经元, 依据神经元接受到的膜电位与阈值电位的数值关系来决定神经信号的发射, 从而使计算过程更加接近真实的生物神经系统 [23], 其主要工作模式如图 4(a) 所示. 相较于 ANN, 以事件驱动的 SNN 计算功耗主要集中在神经元发放脉冲的瞬间, 而不是像 ANN 网络一样持续地高电平工作, 从而有效降低了神经元在未激活时的能耗. 此外, 由于时间序列脉冲的离散特性, SNN 对于输入数据的小变化和扰动具有一定的容忍度, 使其在面向处理实际环境中的不确定性和噪声时具备更高的鲁棒性. 结合 SNN 超低功耗的优势, IBM、英特尔和清华大学等已经相继发布多款 SNN 芯片 [24~26]. 同时, 中国科学院微电子研究所尚德龙团队于 2023 年 10 月正式发布了 “问天 I” 类脑计算机, 实现了 5 亿神经元 2500 亿突触智能规模, 计算能效较现有计算体系提升 10 倍以上.

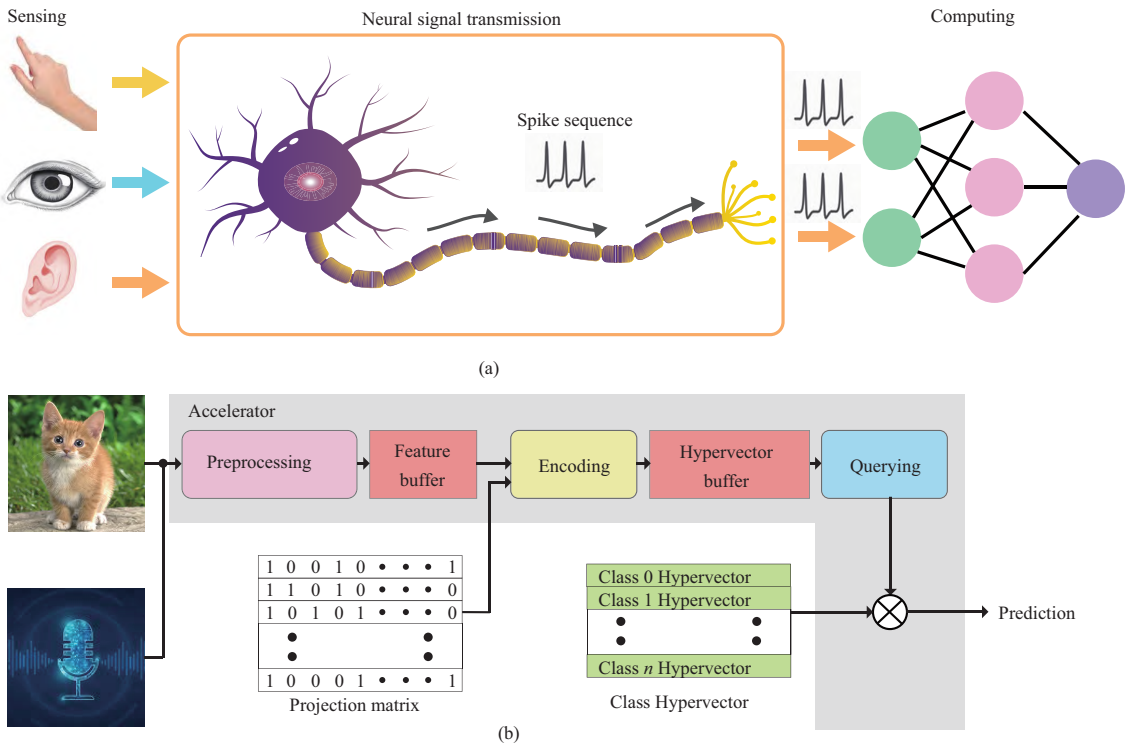

图 4 脑启发式新兴智能算法. (a) 神经形态计算; (b) 超维矢量计算

Figure 4 Brain-inspired algorithms. (a) Neuromorphic computing; (b) hyper-dimensional computing

另外结合新型非易失存储器件的存内硬件设计也是当前主流的 SNN 设计思路之一. 其中, 结合 ReRAM 的阻值可调性、权值存储和模拟计算能力的脉冲型神经网络硬件吸引了众多研究人员的注意力. 如, 密歇根大学的 Cai 等 [27] 设计实现了一款集成了 54×108 规模 1R 忆阻器阵列的类脑芯片, 可实现感知网络、稀疏编码、基于多层网络的主成分分析等应用. 然而, 具备多值特性的 RRAM 所固有的阻值偏差和阻值漂移也同样制约着 SNN 网络的计算精度及网络规模. 因此, FeRAM, MRAM 等其他新型非易失存储技术也同样是 SNN 硬件的重要研究方向. 尤其是具备随机计算能力以及稳定二值状态的 MRAM 器件, 已经成为 SNN 训练和计算的重要实现途径 [28].

2.3.2 超维矢量计算 (hyperdimensional computing, HDC)

区别于现有类脑神经网络, 超维矢量计算凭借其低复杂度、可解释性和鲁棒性, 逐渐成为另一种广受关注的神经网络新型计算范式, 尤其是对于边缘侧的应用场景 [29,30]. 超维矢量计算受生物学理论基础启发, 模仿人脑将感官系统接收外界刺激后产生的输入信号映射为超高维度的神经信号, 并在此超维空间中并行地、高效地实现信号的查询、比对等处理, 如图 4(b) 所示. 凭借超维矢量空间中存在的正交性和全息性, 融合了输入特征的超维神经信号与对应类别矢量的距离将显著低于非对应分类矢量, 从而将传统神经网络中的复杂计算过程转化为简单的数据比对/检索操作, 并可以以极低的检索开销来实现认知任务 (如分类、聚类等). 而其中的类别超向量通常由启发式的方法得到 [31], 或者进一步通过高效的重训练方法进行微调以提升性能 [32]. 在这种情况下, 极低的训练开销成为超维计算的另一优势, 其使得极低算力的边缘侧设备具有了在线训练的能力. 因此, 基于超维矢量计算范式的新型人工智能芯片将有可能成为现有神经网络芯片突破的新方向.

目前, 超维矢量计算是一个软件算法和硬件架构设计方法高度耦合的研究方向. 在算法方面, 统一的、高效的编码方法始终是研究热点. 现有的多种编码方法能够有效表征输入信号的特征, 但是它们的复杂度都相对较高 (尤其是在输入为图像时). 针对不同种类的输入信号也需要选取特定的编码方法 (如图像使用 Level-Id 或 Random Projection [33], 而文本需使用 N-gram [34]), 否则会导致准确率的严重下降. 最新的 GENERIC [35] 虽然通过将 Random Projection 和 N-gram 结合, 实现了一种能够适用于不同种类输入的统一编码方法, 但其有着更高的复杂度. 因此, 忽略面向高能效数据比对硬件的专用数据编码电路设计需求, 而单纯考虑算法的优化将使得超维矢量计算在数据分类过程中低开销优势逐渐被抵消掉.

3 新兴计算的挑战

对于上述提到的几种新兴计算, 能从不同的设计层面提升能量效率, 但也面临诸多挑战. 本节从设计角度和安全可信角度阐述各种新兴计算芯片面临的挑战.

3.1 高能效新兴计算的设计挑战

在面向可容错计算电路方面, 主要面临的设计挑战如下:

• 近似计算技术的精度评估难度大, 系统级误差分析管理、跨层设计与协同优化尚难解决; 此外, 融合新工艺、新器件特点的近似计算对 EDA 设计工具、芯片的可靠性等都提出了重大的技术挑战.

• 随机计算中随机数转换所需要的时间大大高于计算单元所需的时间, 严重影响了随机计算电路的速度, 限制了随机计算的应用范围, 使得随机计算较难在高性能计算中展开应用.

在存储中心计算架构方面, 主要面临的设计挑战如下:

• 近存计算受其承担的工作负载及数据局部性影响, 导致芯片底层局部散热更加困难, 而传统加在芯片顶层的散热栅技术也逐渐无法满足需求, 因此在芯片散热和电路可靠性方面都存在设计瓶颈.

• 存内架构的计算精度依赖于器件本身的物理特性, 对工艺偏差极其敏感, 导致其目前计算精度较传统数字方案不占优势. 另外, 存内计算架构也面临电路稳定性和适用范围小等多方面的问题和挑战.

在面向脑启发式算法方面, 主要面临的设计挑战如下:

• 脉冲神经网络的现有主要训练方法带来的多梯度消失或梯度爆炸问题, 而预训练 ANN 转换 SNN 中的算法复杂度代价较大. 此外, 现有 SNN 芯片复杂的脉冲编码转化电路设计以及 EDA 工具库的缺失也同样影响了 SNN 芯片的开发.

• 超维计算在面向 10000 维以上的向量时, 基于 CMOS 方案的编码和查找电路在能效方面有着不可忽视的劣势, 而在与非易失性器件结合的技术路线上在工艺层面上仍未成熟. 另外, 超维计算的算术运算能力依然处于初步探索阶段.

综上所述, 表 1 对比了本文介绍的几种主要的新兴计算范式的优缺点. 针对不同的应用需求, 结合各种新兴计算范式的特点, 才能发挥出新兴计算范式的最大效用.

3.2 高能效新兴计算芯片的安全可信挑战

新兴计算虽然带来了能效优势, 但也同时带来了新的安全可信问题, 其中主要有以下几种挑战:

• 计算芯片在能耗受限的条件下如何实现安全可信认证和数据的机密性是一个重要的难题. 面对量子攻击威胁, 需要在新兴计算芯片中部署后量子密码加速引擎, 而后量子密码的复杂数学算法使得

表 1 不同新兴计算范式的对比

Table 1 Comparison for different emerging computing paradigms

Emerging computing paradigm	Primary features	Advantages	Disadvantages	Computing power & energy-efficiency
Approximate computing	Sacrificing computing accuracy for circuit performance	High reliability & process maturity	Lack error management mechanism & design methodology	Google TPU 86 TOPS 2.15 TOPS/W
Stochastic computing	Represent number by bitstream with probability	Extreme low circuit complexity	Low computing accuracy & huge cost for RNG	CiM-BNN [36] 1.342 μJ/image 3 ns read latency
Computing in memory	Process data near or in memory	Reduce memory access high parallelism	Sensitive to PVT variation & low accuracy	IBM HERMES 63.1 TOPS 9.76 TOPS/W
Neuromorphic computing	Exploit pulse instead of bits	Event-driven design & high error resilience	Unfriendly for training process	SpiNNaker2 2.24 TOPS 2.1 TOPS/W
Hyper-dimensional computing	Represent number by hyper-dimensional vector	High energy-efficiency & online learning friendly	Unfriendly for arithmetic operation & large cost for encoding	tiny-hd [34] 400 TOPS 6.56 GOPS/W

在高能效芯片上实现硬件加速变得极具挑战性.

• 可容错计算电路会引入新的安全漏洞. 其在执行过程中误差的不可预测性, 使得其计算结果是否被恶意修改难以判断. 此外, 容错电路比精确电路更容易被插入硬件木马. 同时, 容错电路的概念也会导致逆向工程重建近似的电路, 增加了被攻击的风险.

• 深度神经网络模型也面临着对抗样本攻击和后门攻击等安全威胁. 对抗样本攻击中, 攻击者可以通过修改训练数据集、操纵输入特征或数据标签等方式进行攻击. 后门攻击中, 攻击者可通过向目标神经网络嵌入木马神经元, 或修改深度学习模型的权重值来实施后门攻击.

为了应对上述所提到的安全威胁, 多种安全计算架构被提出, 如图 5 所示. 为了应对量子计算的攻击, 基于存内计算的后量子加密技术架构设计被提出, 其使用向量矩阵乘法在数论变换 (number theoretic transform, NTT) 和逆 NTT 操作期间实现最大的并行性 [37]. 为了防止互连和近似函数篡改, 包含输入完整性检查和基于排他逻辑的攻击检测方法的近似计算安全架构被提出 [38]. 在深度神经网络防止对抗样本攻击方面, 能够在现实物理条件下生效的鲁棒且自然的物理对抗样本生成方法被提出 [39], 其能够生成类似自然老化样式的对抗样本, 与原始图像非常相似, 不会引起人类的察觉, 同时能够使得 Faster R-CNN Inception v2, SSD Inception v2, YOLO v2 等模型判错.

4 新兴计算芯片的展望

新兴计算范式作为提高算力的重要解决途径, 已经投入了大量的研究工作. 本节从作者观点出发, 给出了新兴计算技术在中短期和中长期的技术展望.

• **中短期 – 挖掘设计潜力.** 根据当前半导体技术的发展情况, 目前主流的芯片仍然基于硅基晶体管技术, 通过数字逻辑或模拟电信号来实现计算功能. 以当前备受瞩目的大型预训练模型为例, 其训练过程依赖于庞大的 GPU 集群. 因此, 从中短期来看, 探索基于相对成熟工艺的新兴计算范式, 将成为近若干年高能效计算芯片的发展机遇和产业化方向. 结合人工智能、下一代通信、泛物联网等相

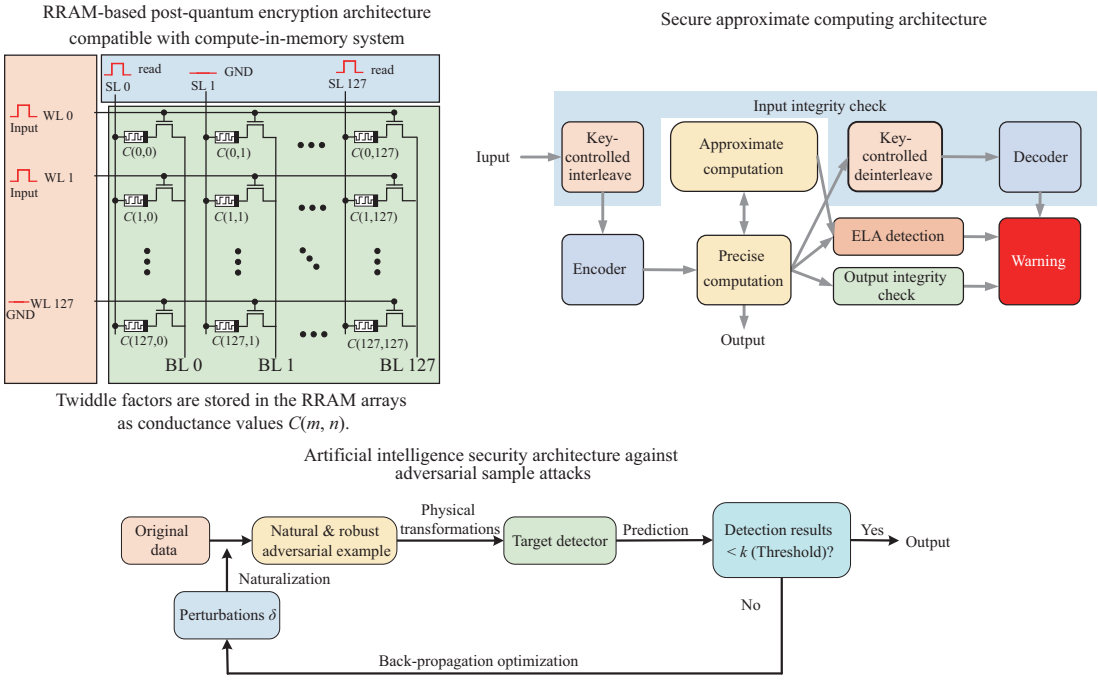

图 5　针对特定攻击威胁的安全新兴计算架构

Figure 5　Secure emerging computing architecture against specific attack threats

关应用领域, 脑启发式的智能算法以及近似计算、随机计算等新兴电路设计范式, 可在成熟的 CMOS 技术和相对成熟的非易失性器件上部署, 有望大幅提高芯片的能效, 弥补先进制程技术的不足, 部分解决了计算芯片技术的 “卡脖子” 问题. 同时, 与这些新兴计算范式相关的设计理论、方法和自动设计工具的研究与开发将为新兴计算范式的大规模应用提供支持. 有助于设计人员充分挖掘现有半导体工艺的计算潜力. 另外, 如何实现不同的新兴计算范式的高效异构集成以及 “芯粒” 领域的先进封装与互连关键技术也将成为重要研究方向, 有望为计算系统提供更高的性能和效率, 推动计算技术的进一步发展. 与此同时, 新兴计算范式所面临的安全可信风险也是在其大范围应用前所需解决的问题. 从另一角度来看, 新兴计算范式也可能从新的设计维度帮助实现新的硬件安全机制, 这些问题有较大的研究空间.

• **中长期 – 突破器件制约.** 从中长期展望而言, 当前的微纳电子技术正在持续演进, 研究人员积极探索不仅局限于硅基晶体管技术, 还包括新型半导体二维材料器件技术. 这些探索助推着半导体领域的前进. 随着技术工艺的不断成熟, 本文所介绍的计算范式有望借助新型半导体器件, 进一步接近微纳电子器件的兰道尔物理极限. 这一趋势将推动计算能力的持续提升, 不仅可提高计算速度和效率, 还可能改变传统半导体技术的局面, 为未来的科技进步开辟新的道路. 另外, 人类势必会在寻找新型计算载体方面取得突破性进展. 当前尚处于初探阶段的技术, 如量子计算、光子计算、生物计算等, 都具备潜力成为实现计算能力发展的关键技术途径, 有望推动计算功能的飞跃性发展. 然而, 这些领域仍然需要深入研究和不断的创新, 以充分发挥其潜力, 实现科技的长期进步.

5　总结

人类正在经历一场万物互联和信息智能化的浪潮, 计算技术的演进在这一革命中扮演着尤为重要

的角色, 成为科技创新不可或缺的基础支持. 然而, 从当前的技术进展来看, 无论是计算能力的硬件设计方法还是相关领域仍然存在广泛的探索空间. 本文综述了当前一系列新兴计算范式, 其中包括电路设计方式、芯片架构以及类脑智慧算法. 同时, 也对这些新兴计算技术所面临的挑战和未来可能的发展方向进行了深入探讨. 我们期望这些信息能为相关领域的从业者提供有益的参考, 积极探索高效能、高安全性芯片的理论和技术创新, 以此为未来智慧社会的发展注入新的动力.

参考文献

1 Semiconductor research corporation. The Decadal Plan for Semiconductors. 2021. https://www.src.org/about/decadal-plan/
2 国际数据公司 IDC, 浪潮信息, 清华大学全球产业研究院. 2022–2023 全球计算力指数评估报告. 2023. https://www.igi.tsinghua.edu.cn/info/1019/1321.htm
3 Liu W, Lombardi F, Shulte M. A retrospective and prospective view of approximate computing. Proc IEEE, 2020, 108: 394–399
4 Kang M, Gonugondla S K, Shanbhag N R. Deep in-memory architectures in SRAM: an analog approach to approximate computing. Proc IEEE, 2020, 108: 2251–2275
5 Armeniakos G, Zervakis G, Soudris D, et al. Hardware approximate techniques for deep neural network accelerators: a survey. ACM Comput Surv, 2022, 55: 1–36
6 Zhao W, Feng D, Tong W, et al. APPcache+: an STT-MRAM-based approximate cache system with low power and long lifetime. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2023, 42: 3840–3853
7 Zhang H, Putic M, Lach J. Low power GPGPU computation with imprecise hardware. In: Proceedings of the 51st Annual Design Automation Conference, 2014. 1–6
8 Liu W, Liao Q, Qiao F, et al. Approximate designs for fast Fourier transform (FFT) with application to speech recognition. IEEE Trans Circuits Syst I, 2019, 66: 4727–4739
9 Liu W Q, Lombardi F. Approximate Computing. Cham: Springer, 2022
10 Zhang Y W, Wang R S, Jiang X B, et al. Design guidelines of stochastic computing based on FinFET: a technology-circuit perspective. In: Proceedings of IEEE International Electron Devices Meeting (IEDM), San Francisco, 2017
11 Liu Y, Liu S, Wang Y, et al. A survey of stochastic computing neural networks for machine learning applications. IEEE Trans Neural Netw Learn Syst, 2021, 32: 2809–2824
12 Alawad M, Lin M. Survey of stochastic-based computation paradigms. IEEE Trans Emerg Top Comput, 2019, 7: 98–114
13 Alaghi A, Hayes J P. Survey of stochastic computing. ACM Trans Embed Comput Syst, 2013, 12: 1–19
14 Hu J, Li B, Ma C, et al. Spin-Hall-Effect-based stochastic number generator for parallel stochastic computing. IEEE Trans Electron Devices, 2019, 66: 3620–3627
15 Lammie C, Eshraghian J K, Lu W D, et al. Memristive stochastic computing for deep learning parameter optimization. IEEE Trans Circuits Syst II, 2021, 68: 1650–1654
16 Romaszkan W, Li T, Garg R, et al. A 4.4-75-TOPS/W 14-nm programmable, performance- and precision-tunable all-digital stochastic computing neural network inference accelerator. IEEE Solid-State Circuits Lett, 2022, 5: 206–209
17 Chen Z, Ma Y, Wang Z. Hybrid stochastic-binary computing for low-latency and high-precision inference of CNNs. IEEE Trans Circuits Syst I, 2022, 69: 2707–2720
18 Xue C X, Hung J M, Kao H Y, et al. A 22 nm 4 Mb 8b-precision ReRAM computing in-memory macro with 11.91 to 195.7 TOPS/W for tiny AI edge devices. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 246–248
19 Kwon Y C, Lee S H, Lee J, et al. A 20 nm 6 GB function-in-memory DRAM, based on HBM2 with a 1.2 TFLOPS programmable computing unit using bank-level parallelism, for machine learning applications. In: Proceedings of IEEE International Solid-State Circuits Conference (ISSCC), 2021. 350–352
20 Sun X, Khwa W S, Chen Y S, et al. PCM-based analog compute-in-memory: impact of device non-idealities on inference accuracy. IEEE Trans Electron Devices, 2021, 68: 5585–5591
21 Engel B N, Akerman J, Butcher B, et al. A 4-Mb toggle MRAM based on a novel bit and switching method. IEEE

Trans Magn, 2005, 41: 132–136
22 Kawahara T, Takemura R, Miura K, et al. 2 Mb SPRAM (SPin-transfer torque RAM) with bit-by-bit bi-directional current write and parallelizing-direction current read. IEEE J Solid-State Circuits, 2008, 43: 109–120
23 Rathi N, Agrawal A, Lee C, et al. Exploring spike-based learning for neuromorphic computing: prospects and perspectives. In: Proceedings of Design, Automation & Test in Europe Conference & Exhibition (DATE), Grenoble, 2021. 902–907
24 Khodagholy D, Gelinas J N, Thesen T, et al. NeuroGrid: recording action potentials from the surface of the brain. Nat Neurosci, 2015, 18: 310–315
25 Akopyan F, Sawada J, Cassidy A, et al. TrueNorth: design and tool flow of a 65 mW 1 million neuron programmable neurosynaptic chip. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2015, 34: 1537–1557
26 Painkras E, Plana L A, Garside J, et al. SpiNNaker: a 1-W 18-core system-on-chip for massively-parallel neural network simulation. IEEE J Solid-State Circuits, 2013, 48: 1943–1953
27 Cai F, Correll J M, Lee S H, et al. A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations. Nat Electron, 2019, 2: 290–299
28 Zhou P, Smith J A, Deremo L, et al. Synchronous unsupervised STDP learning with stochastic STT-MRAM switching. 2021. ArXiv:2112.05707
29 Kanerva P. Hyperdimensional computing: an introduction to computing in distributed representation with high-dimensional random vectors. Cogn Comput, 2009, 1: 139–159
30 Amrouch H, Imani M, Jiao X, et al. Brain-inspired hyperdimensional computing for ultra-efficient edge AI. In: Proceedings of International Conference on Hardware/Software Codesign and System Synthesis, 2022. 25–34
31 Ge L, Parhi K K. Classification using hyperdimensional computing: a review. IEEE Circuits Syst Mag, 2020, 20: 30–47
32 Imani M, Bosch S, Datta S, et al. QuantHD: a quantization framework for hyperdimensional computing. IEEE Trans Comput-Aided Des Integr Circuits Syst, 2020, 39: 2268–2278
33 Khaleghi B, Xu H, Morris J, et al. Tiny-HD: ultraefficient hyperdimensional computing engine for IoT applications. In: Proceedings of IEEE Design, Automation & Test in Europe Conference & Exhibition (DATE), 2021. 408–413
34 Rahimi A, Kanerva P, Rabaey J M. A robust and energy-efficient classifier using brain-inspired hyperdimensional computing. In: Proceedings of IEEE International Symposium on Low Power Electronics and Design, 2016. 64–69
35 Khaleghi B, Kang J, Xu H, et al. GENERIC: highly efficient learning engine on edge using hyperdimensional computing. In: Proceedings of the 59th ACM/IEEE Design Automation Conference (DAC), 2022. 1117–1122
36 Gu H, Jia X, Liu Y, et al. CiM-BNN: computing-in-MRAM architecture for stochastic computing based Bayesian neural network. IEEE Trans Emerg Top Comput, 2023. doi: 10.1109/TETC.2023.3317136
37 Park Y, Wang Z, Yoo S, et al. RM-NTT: an RRAM-based compute-in-memory number theoretic transform accelerator. IEEE J Explor Solid-State Comput Devices Circuits, 2022, 8: 93–101
38 Yellu P, Monjur M R, Kammerer T, et al. Security threats and countermeasures for approximate arithmetic computing. In: Proceedings of the 25th Asia and South Pacific Design Automation Conference (ASP-DAC), Beijing, 2020. 259–264
39 Xue M, Yuan C, He C, et al. NaturalAE: natural and robust physical adversarial examples for object detectors. J Inf Security Appl, 2021, 57: 102694

High-efficiency and high-security emerging computing chips: development, challenges, and prospects

Weiqiang LIU[1,2*†], Ke CHEN[1,2†], Bi WU[1,2], Erya DENG[1,2], You WANG[1,2], Yu GONG[1,2], Yijun CUI[1,2] & Chenghua WANG[1,2]

1. *College of Integrated Circuits, Nanjing University of Aeronautics and Astronautics, Nanjing* 211106*, China*;
2. *Key Laboratory of Aerospace Integrated Circuits and Microsystem, Ministry of Industry and Information Technology, Nanjing* 211106*, China*
* Corresponding author. E-mail: liuweiqiang@nuaa.edu.cn
† Equal contribution

Abstract The demand for computational power in the era of intelligent informatization is steadily increasing, and high-efficiency and high-security computing chips have become indispensable infrastructures that support technological innovation and societal progress. As an innovative technology for enhancing computational power, emerging computing paradigms have made significant breakthroughs in both theory and technology in recent years, attracting widespread attention from both academia and industry. This paper provides an introduction to and analysis of the cutting-edge technologies related to emerging computing chips from various perspectives, including circuit design methods, novel chip architectures, and brain-inspired algorithms. In addition, it discusses the phase-specific characteristics of each technology and the design and security challenges they face. Finally, the paper presents prospects for the future development of emerging computing chip technology and outlines the key directions for its advancement.

Keywords emerging computing paradigm, secure and trust, approximate computing, stochastic computing, in-memory computing, brain-inspired computing