后摩尔时代：半导体芯片创新技术研究进展

张明磊　高子墨　金千翔　张秋钥　陈思宇
辽宁工程技术大学　电子与信息工程学院　辽宁葫芦岛　125100

摘　要：随着器件特征尺寸不断接近物理极限，在强调晶体管微缩的传统摩尔定律步调受阻之后，为了克服功耗、发热、成本与良率等制约因素带来的问题，集成电路与芯片业进入摩尔定律放慢速度后的“后摩尔时代”，进而围绕先进制程的持续推进而展开器件与工艺、系统集成、架构与计算范式的创新与探索。本文对此进行了相关研究整理，针对三种主流创新路径——器件与工艺、系统集成及架构层面上做了简要说明：基于平面MOSFET到FinFET-GAA纳米片/纳米线晶体管及高介电常数金属栅材料、二维材料等新材料方面的结构创新与应用前景；2.5D/3D封装及Chiplet架构如何进行系统级优化以提升系统性能与能效；异构计算、存算一体/近存计算和专用领域架构针对数据密集型任务与高能效任务场景的优势体现；类脑计算、量子计算、自旋电子学等不同方向的特点与挑战。

关键词：后摩尔时代，芯片工艺，技术创新

引言

摩尔定律提出于20世纪60年代，其长期以来一直指导着半导体产业的发展。其核心含义是指集成电路中可容纳的晶体管数目平均每18～24个月会翻一番，同时伴随性能增强和成本降低[1]。但随着器件特征尺寸接近极限，器件功耗大、发热量高、制造成本增加及良率低等问题不断出现，原有的以器件尺度缩减为基点的摩尔定律在上述方面不可避免地出现了缓慢及失效的现象。学界与业界通常将此过程称为“后摩尔时代”。在该时期，集成电路从原来的单纯依靠器件尺寸缩减来实现高集成度的方法转变为依靠新材料、新器件结构、新系统架构以及新的计算范式等多方面并行发展的新模式。基于此，本文归纳了半导体芯片领域的相关研究进展，总结了后摩尔时代的几个关键技术路径。

一、器件与工艺层面的创新

（一）新型晶体管结构

为使CMOS技术体系在后摩尔时代具有延续性，晶体管结构必然是要从2D走向3D、由部分包围走向全包围变化的。传统的平面MOSFET工艺，在沟道长度不断减小的过程中存在着沟道短、漏源电流大、功耗高等一系列缺点，已经无法满足未来更高技术节点下工艺对于器件特性的要求[2]。

鳍式场效应晶体管（Fin Field-Effect Transistor，FinFET）由于引入了三维鳍状沟道结构，利用栅极从多个方向控制沟道载流子来增强电场控制能力，有效地抑制了短沟道效应，FinFET自22 nm节点起逐步取代平面MOSFET，并在14 nm、10 nm及7 nm工艺中成为主流器件结构，为摩尔定律在过去十余年的延续发挥了关键作用。但是随着器件尺寸变小，鳍宽度不能继续缩窄，工艺日益复杂化、难以量产化等问题也随之浮现。

在这个大背景下，GAA晶体管作为FinFET的重要继任者，其设计思想通过使栅极完全包裹沟道，实现对载流子的全方位电场控制，相较FinFET进一步降低了漏电流并提升了亚阈值特性。GAA架构中以基于GAA纳米片（nanosheet）晶体管为主流的发展方向，在垂直方向上采用多个由可调厚度沟道组成的纳米片，使得沟道达到纳米级尺寸，在此情况下，该种结构既能在功耗和功频方面满足要求，又能保证工艺制备的可行性，是一种十分有前景的解决方案。而在局限尺度上，纳米线（nanowire）晶体管可能具有更好的电学控制能力。综合来看，GAA以及其演化出来的其他类型晶体管会被广泛应用于小于等于3nm工艺的器件。

（二）新材料的引入

基于硅基CMOS工艺面临物理瓶颈效应，单凭器件结构优化提升芯片性能的技术路线日渐式微，研究新材料是解决芯片性能不足的有效方案，也是超越摩尔时代的主流方向。硅材料电子迁移率低、功耗高、难以进一

步微缩等问题开始显露，由此带来能否找到比硅性能更好的、具有成本优势的新材料成为摆在中国及世界CMOS集成电路领域科学家面前的重大问题。

在栅极材料与介质方面，高介电常数（High-k）介质与金属栅（Metal Gate）技术已在先进制程中得到成熟应用。通过使用高介电常数材料替代传统SiO_2作为栅介质，可在保持较高栅电容的同时显著降低漏电流，从而有效缓解功耗问题。这一技术路线的成功应用，为后摩尔时代器件性能优化提供了重要支撑。

与此同时，二维材料因其原子级厚度和天然抑制短沟道效应的优势，成为后摩尔时代极具潜力的研究方向。理论上具有极佳电学、力学特性的石墨烯、TMDC（过渡金属硫化物：如MoS_2、WS_2）材料是制作超微缩器件的绝佳材料，但由于带隙调控困难、接触电阻大及无法保证可重复制造等问题，离真正的产业应用还需时日。

（三）先进封装与集成工艺

在器件尺寸微缩受限，制程成本急剧增加的形势下，器件单芯片性能增加的边际效用越来越小，而基于先进封装和集成工艺可为器件提供更高的系统级性能。相较于传统的封装方式，先进封装更侧重于从系统层面去考虑高密度、高带宽、低功耗的互连方法[3]。

2.5D和3D封装是采用硅中介层（interposer）或垂直堆叠的方式把多个芯片或功能模块，在空间上靠得很近的封装形式，极大缩减了互连长度，减小了信号延时和功耗。

Chiplet架构是后摩尔时代的系统级设计理念，基于此把大的系统拆分成一些功能比较明确，可以复用的芯片，用高速互联的方式去组合起来，既提升了设计的灵活度和良率，同时还能做到研发、制造的成本的降低。

二、架构层面的创新

（一）异构计算架构

迈进后摩尔时代之后，应用呈现出极端多样、无比复杂等特点，传统的以通用CPU为核心的传统计算架构已经难以做到通过扩展性能去满足性能要求和提升能效的目的，一方面通用CPU在做超大规模并行、高吞吐场景的应用时效率较差，另一方面单纯靠增加主频或者增加晶体管数量在功耗受限场景下不能保证能够有持续性能增长的趋势，因此一种新型的更有利于软件优化利用、并且有助于算力发挥的异构计算架构逐渐成为解决此类问题的方法之一。

异构计算是在同一台计算机上将不同的处理单元比如CPU、GPU、FPGA以及ASIC等组装在一起，然后依据具体的业务场景分配任务，在同一台机器上选择适合的任务分配到合适的硬件上去，达到以合适的硬件执行合适的任务的目的。对于人工智能、科学计算、大数据处理等场景而言，异构计算架构都具有明显的优势。比如，在深度学习的训练与推理场景中，这些任务常常需要做很多矩阵运算，其计算性质非常适合于GPU以及AI加速器。

（二）存算一体与近存计算

传统的冯·诺依曼计算体系架构中将计算单元和存储单元分开，当计算规模较小的时候，这种设计非常灵活和易于扩展。但是在大数据的应用下，需要经常拷贝数据的运算大量增加，这时，就会给系统的运行速度以及功耗带来一些瓶颈，也出现了所谓的“存储墙”、“功耗墙”，目前只通过提高存储带宽或者增加缓存层级都很难彻底解决这样的问题。

存算一体（Processing-in-Memory，PIM）与近存计算（Near-Memory Computing）就是把一部分计算移动到存储器的附近或者里面去实现，把数据计算的部分工作放在数据所在的存储器附近的节点上，而不是先把数据挪到CPU里然后再做运算。这样可以有效降低数据传输带来的能耗和时延，在诸如矩阵运算、向量计算、模式匹配的数据密集型业务中存在天然的优势。

（三）专用领域架构（DSA）

随着应用场景日趋专业化，对通用计算架构的“平均最优”性能与能耗指标已经越来越无法满足某些专业的最优化需求。于是人们从对一类应用或算法设计专用架构入手，基于对这些应用或者算法本身的性质而对其进行硬件结构、指令集、数据通路等方面的专门化设计来实现性能及能效方面的超越。

在AI领域中，由于深度学习推理和训练任务都对矩阵运算以及数据复用有着规律性，所以加速了AI专用加速器的出现和发展。比如谷歌的张量处理单元（Tensor Processing Unit，TPU），就专门用于神经网络计算，从功能上来说其实是DPU的一种表现形式，相比于FPGA/DSP架构，能效和吞吐率实现了较大的提高。同样也有一些AI推理的芯片应用于边缘计算、智能终端或者数据中心中。

除了靠硬件创新驱动外，专用领域架构的研发，还需要软件和算法相互匹配的配合优化才能充分发挥硬核

能力，发挥后摩尔时代的“软硬件协同设计”能力。从深层上讲，在编译器、运行时、算法层面都具有很好的硬件特性，则能更好的发挥D.A.S的效用。总而言之，目前DSA是软硬件协同推动半导体架构演进、带动应用场景性能提升不可或缺的一股力量。

三、新型计算范式探索

（一）类脑计算

类脑计算（neuromorphic computing）是一种受到人脑信息处理机制启发的新范式计算方法，核心是通过软硬件来模仿生物神经系统的神经元和突触结构、工作机理等实现高并行、低功耗、可适应性信息处理。不同于基于冯诺依曼架构的传统数字计算方式所特有的数据流驱动和顺序执行，类脑计算是以事件驱动和分布式并行计算以及计算存储紧密耦合为主要特征的一种计算模式，该计算模式适合于完成感知、学习和决策等任务[4]。

在算法层面来说，脉冲神经网络是类脑计算的重要理论基础。SNN将离散的脉冲信号映射到生物神经元的发射行为，并且可以通过时间顺序的方式去表示和存储信息，相比传统的SNN算法具备功耗更低、更加实时的特点，同时SNN在如何训练SNN模型、保证SNN模型的稳定性及准确率方面仍有待进一步完善。

从硬件角度讲，由于具有与突触类似的特性，忆阻器（Memristor）等新型器件得到了更多的关注。它们可以在非易失的状态下存储权重的信息，在物理层面上实现了加权求和的运算，从而可以实现高密度和低功耗的类脑计算芯片的构建。基于忆阻阵列的类脑计算架构已经取得了明显的进展，对于如智能感知、模式识别以及自适应学习等很多方面，都表现出了十分乐观的应用前景。虽然当前的类脑计算依旧处于发展阶段，但它是后摩尔时代突破能效瓶颈、扩展计算范式的重要途径之一。

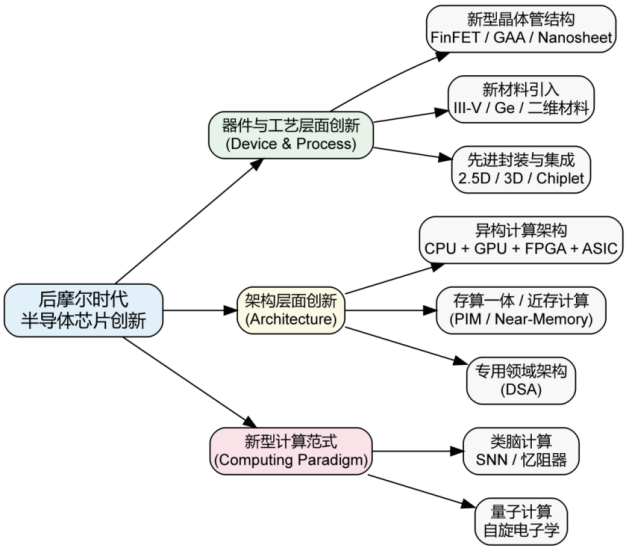
（二）量子计算与自旋电子学

量子计算是一种全新的计算范式，由于其量子比特叠加与纠缠所带来的指数级加速优势，故在特定问题上远超经典计算机，尤其是大数分解、量子模拟以及组合优化等领域，量子算法已经得到大量理论验证；加之在传统半导体技术路线逐步走向尽头的大背景下，人们寄予厚望的量子计算或将是突破计算能力增长瓶颈的最优解。

现阶段量子计算刚刚走出实验室步入工程化应用，量子比特数量少、退相干严重及容错计算等都是量子计算发展的困难；与此同时，以超导量子比特为代表的众多研究方向以及离子阱、半导体自旋量子比特等方面都不断有新的进展，为我们以后的可扩展量子计算机的实现打下良好的基础。在后摩尔时代的发展规划中，长久来说，量子计算可能会成为经典计算的一个很好的补充，而不会完全代替它[5]。

与此同时，自旋电子学（Spintronics）采用电子自旋自由度来完成信息存储和处理的功能，在低功耗、非易失逻辑电路中寻找到了新的途径，相对传统的基于电荷的电子器件，其功耗小、信息保持时间长、抗辐射能力强等优点，从而给纳米级交叉的磁场效应提供了很大的想象空间；如磁阻随机存储器（MRAM）是自旋电子学最典型的产物，如今已经实现了规模量产，并且被广泛应用在嵌入式系统及高可靠计算领域当中。

自旋转移力矩型和自旋轨道力矩型器件结构具有良好的可编程性，能实现多种功能逻辑门电路；从逻辑计算角度考虑，是设计新型低功耗逻辑电路的一种可能。尽管自旋电子学在后续逻辑集成及系统架构上还存在不少问题，但对于实现后摩尔时代的存算一体以及降低能耗仍具有重要的研究价值。



四、挑战与发展趋势

尽管现在的后摩尔时代的技术路线多，在后摩尔时代依然存在制造工艺变得更加复杂、设计的成本越来越高、软硬件的结合难度大等问题。以后半导体技术会从以前单个的点做得越来越深逐步转向做一些系统级的优化和多学科的交叉。通过材料、器件、架构、算法这些方面的协同创新把我们的计算能力、计算效率不断地提高上去。

结论

后摩尔时代不是半导体技术创新终止的地方，而是新的技术创新范式出发的地。通过各种新的器件结构、新材料、先进封装、异构架构及新的计算范式相结合的应用，将给半导体产业带来性能、能效和功能多样性的突破性进展。往后摩尔时代的方向上看，多元发展将是半导体芯片技术的主要方向，以几种不同的技术路线同时进行开发。

参考文献

[1] 张志勇.碳纳米管CMOS电子学－从晶体管到中等规模集成电路[C]// 中国化学会.中国化学会第30届学术年会摘要集－第一分会：表面物理化学.北京大学纳米器件物理与化学教育部重点实验室，北京大学电子学系；北京华碳元芯电子科技有限公司，2016：44.

[2] 王奕琛.环栅型纳米片晶体管器件建模与参数优化加速技术研究[D].华中科技大学，2024.

[3] 李蛇宏，高集成半导体的关键封测技术研究及应用.四川省，四川明泰微电子科技股份有限公司，2023-08-08.

[4] 贡以纯，明建宇，吴思齐，等.面向类脑计算的低电压忆阻器研究进展[J].物理学报，2024，73（20）：233-257.

[5] 郑超，周琪，崔东岳，等.自旋光电效应与有机自旋半导体材料开发[J].化学进展，2025，37（10）：1410-1427.