专题: 面向类脑计算的物理电子学

基于 3D-NAND 的神经形态计算

陈阳洋[1)2)3)] 何毓辉[3)4)] 缪向水[3)4)] 杨道虹[1)2)3)†]

1) (华中科技大学, 博士后流动站, 武汉 430074)

2) (武汉新芯集成电路制造有限公司, 博士后工作站, 武汉 430205)

3) (江城实验室, 武汉 430205)

4) (华中科技大学集成电路学院, 武汉 430074)

(2022 年 5 月 16 日收到; 2022 年 9 月 27 日收到修改稿)

神经形态芯片是一种新兴的 AI 芯片. 神经形态芯片基于非冯·诺依曼架构, 模拟人脑的结构和工作方式, 相比冯·诺依曼架构的 AI 芯片, 神经形态芯片在效率和能耗上有显著的优势. 3D-NAND 闪存工艺成熟并且存储密度极高, 基于 3D-NAND 的神经形态芯片受到许多研究者的关注. 然而由于该技术的专利性质, 少有基于 3D-NAND 神经形态计算的硬件实现. 本文综述了用 3D-NAND 实现神经形态计算的工作, 介绍了其中前向传播和反向传播的机制, 并提出了目前 3D NAND 在器件、结构和架构上需要的改进以适用于未来的神经形态计算.

关键词: 神经形态计算, 3D-NAND, 存算一体架构

PACS: 07.05.Mh, 85.35.–p, 84.30.–r, 87.18.Sn **DOI:** 10.7498/aps.71.20220974

1 研究背景

1.1 神经形态计算是未来通用人工智能的重要路径

数据、算法和算力是人工智能 (artificial intelligence, AI) 的三大要素, 未来 AI 的发展将面临算力不足的瓶颈 (如图 1 所示[1]). 一方面, 云端计算中的通用 AI 模型性能强大, 但参数庞大, 通常采用 CPU/GPU 硬件平台进行训练, 训练成本高昂且难以普及. 另一方面, 随着 5G、物联网与工业 4.0 的发展, 越来越多的 AI 应用在边缘端设备中设计和部署, 需要定制化 AI 芯片满足功耗和成本限制下的计算需求, 例如阿里的“含光”、华为的“昇腾”和寒武纪的“思源”等. 通用 CPU/GPU 以及 AI 加速芯片, 均基于传统的冯·诺依曼计算架构, 计算和存储单元分离, “冯·诺依曼瓶颈”不可避免 (如图 2 所示[2–4]): 第一, 数据在计算和存储单元之间不停来回传输, 消耗大部分的计算时间和功耗; 第二, 处理器和存储器之间运算速度的明显差异限制了整体系统的计算效率. 面对这一问题, 存储和计算融合是未来的发展趋势, 新型的计算架构逐渐兴起, 其中包括近存计算 (near-memory computing)、存内计算 (in-memory computing) 以及神经形态计算 (neuromorphic computing). 受人脑智能启发的神经形态芯片引起了学术和工业界的极大兴趣.

人脑在复杂和陌生场景下的学习、推理和决策能力远超过传统计算机. 人脑有超过 10^{11} 个神经元和 10^{15} 个突触, 功耗只有 20 W[5]. 神经形态芯片模拟人脑的结构和工作方式 (如图 3 所示[6]): 在结构上, 用互补金属氧化物半导体 (complementary metal-oxide-semiconductor, CMOS) 或新型器件模拟神经元和突触, 并将神经网络映射到突触阵列

† 通信作者. E-mail: alan_yang@xmcwh.com

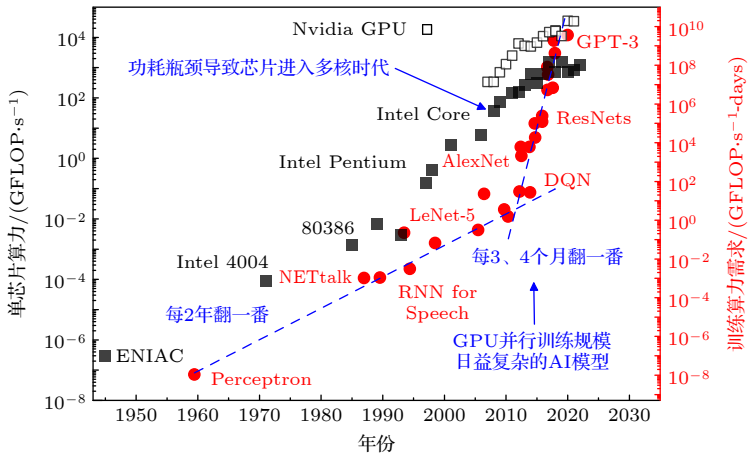


图 1 AI 模型训练的算力需求和单芯片运算性能[1]

Fig. 1. Total amount of computing in AI training and single-core chip performance[1].

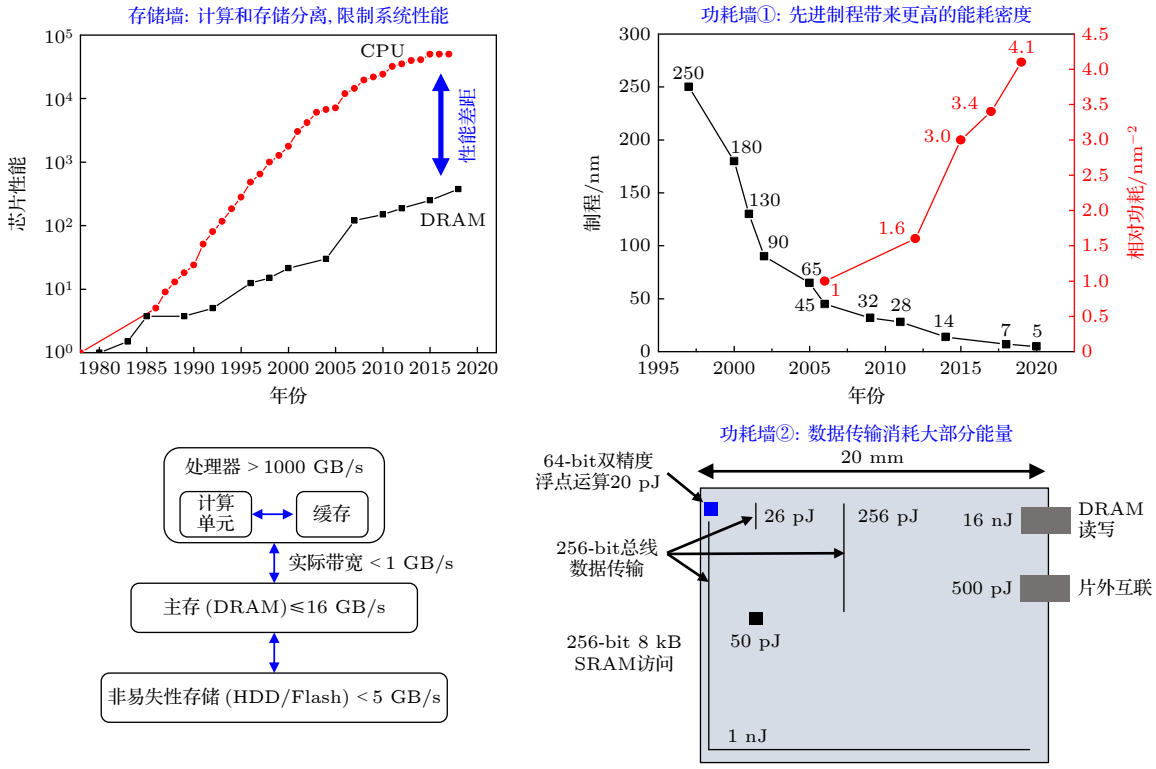


图 2 限制芯片性能提升的冯·诺依曼瓶颈[2–4]

Fig. 2. The von Neumann bottleneck limits chip performance promotion[2–4].

中; 在工作方式上, 用基于权重模拟值计算的人工神经网络 (artificial neural network, ANN) 或基于脉冲计算的脉冲神经网络 (spiking neural network, SNN) 作为算法模型, 输入的事件用电压脉冲编码, 经过突触后转化为电流输入到后级神经元进行积分, 后级神经元达到阈值电压后便向下一级神经元发放电压脉冲. 相比冯·诺依曼架构芯片, 神经形态芯片高并行、低功耗和存算融合的特性, 有望成为未来通用 AI 的理想硬件方案.

1.2 基于存储器的神经形态芯片

神经形态芯片根据实现的器件方案可分为基于传统 CMOS 的神经形态芯片和基于存储器的神经形态芯片两种类型. 基于 CMOS 的神经形态芯片代表性成果包括 TrueNorth, SpiNNaker, BrainScaleS, Loihi, 天机芯和达尔文等. 这些芯片的突触和神经元采用基于 CMOS 的数字电路或者数模混合电路来搭建, 模拟单个神经元或突触行为需要靠多个 CMOS 晶体管组成的电路模块来实现,

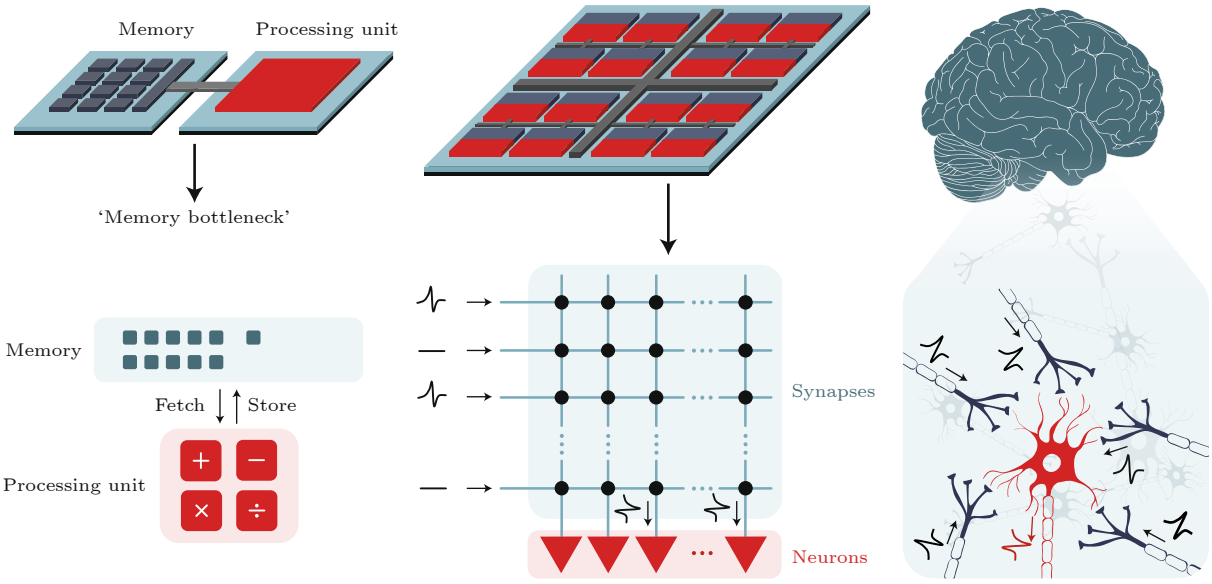


图 3 神经形态计算的原理与芯片架构[6]

Fig. 3. Neuromorphic computing architectures and paradigms[6].

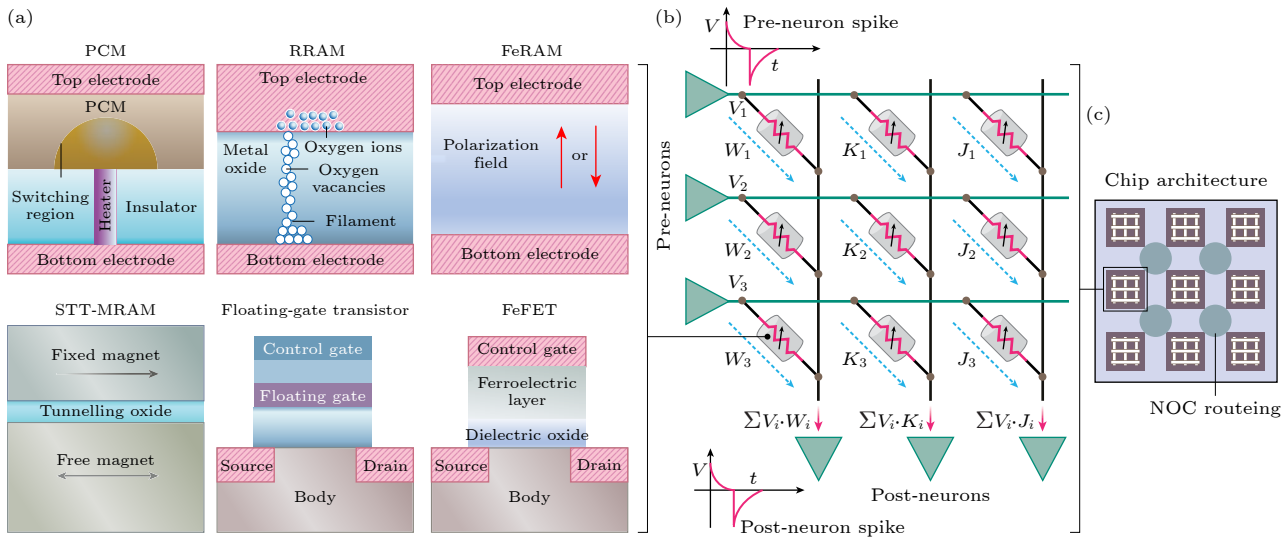


图 4 基于非易失存储器的神经形态计算硬件方案[7] (a) 几种作为突触的非易失性存储器件, 其中包括 PCM, ReRAM, STT-MRAM, FeRAM 和 FeFET; (b) 突触和神经元构成 crossbar 阵列结构用于神经形态计算; (c) 神经形态计算芯片的架构

Fig. 4. Use of non-volatile memory devices as synaptic storage[7]: (a) Non-volatile memory cell as artificial synapse including PCM, ReRAM, STT-MRAM, FeRAM, and FeFET; (b) the implementation of neuromorphic computation on crossbar array consists of artificial synapses and neurons; (c) typical architecture of the neuromorphic chip.

集成度和功耗受到限制, 并且断电后信息无法保存. 基于存储器的神经形态芯片从底层器件仿生的角度出发, 用存储器件模拟神经元和突触, 在功耗和硬件代价上有明显的优势. 近几年, 国内外研究机构展示了众多基于存储器的神经形态计算成果, 硬件方案包括主流的闪存 (NOR/NAND Flash), 以及阻变存储器 (ReRAM)、相变存储器 (PCM)、自旋转移矩磁存储器 (STT-MRAM)、铁电存储器 (FeRAM) 和铁电晶体管存储器 (FeFET) 等新型存储器. 这些存储器的结构和工作原理如图 4 所示[7]. 表 1 列出这几类存储器的性能指标, 其中参数源于已有阵列或芯片实现的研究工作.

1.3 NAND 用于神经形态计算的优点与局限性

突触器件需要具有高集成度、低能耗、耐擦写、CMOS 工艺兼容以及模拟权重调制的特性. 其中模拟权重调制特性需要突触有多个权重、并且

表 1 几种非易失性存储器的性能参数

Table 1. Benchmark table of performance of emerging memories and typical memories.

	新型存储器					主流存储器	
	PCM	ReRAM	STT-MRAM	FeRAM	FeFET	NOR flash	NAND flash
器件结构	1T-1R	1T-1R	1T-1R	1T-1C/ 1T-1FTJ	1T	1T	1T
器件面积F^2	4—40	6—26	9—75	30—40	10—30	10	4
工艺节点/nm	14[8,9]	14[10]	22[11]	130[12]	22[13]	40	15(2D) 80(3D)[14]
芯片容量	16—64 GB (Intel optane)	1 MB—32 GB[10,15]	16 kB—4 GB[16,17]	16 kB—128 MB[18,19]	64 kB—32 MB[13,20]	1 MB—2 GB (Micron/Infineon/Macronix product manuals)	1 TB[21]
写电压/V	< 3	< 1	< 2	< 3	~1.5	10—15	15—25
写时间/ns	75[9]	1—10[22]	3—14[23,24]	1—10[25]	1—10[25]	10000[26]	100000[27]
写功耗/(pJ·bit^{-1})	1[28]	~10[29]	4.5[30]	0.1[25]	1—10 fJ[25]	49[26]	~1000[31]
读时间	12—130 ns[8,32]	2.9—21 ns[33,34]	2—26 ns[30,35]	1—25 ns	1—10 ns	11 ns	3 μs[27]
读功耗/(pJ·bit^{-1})	2.47[36]	1.76[29]	0.7[30]	9.8—19.2[37]	0.27 fJ/bit[38]	2.2[26]	~100[31]
开关比	$> 10^4$	$> 10^3$	> 2	> 2	$> 10^8$	$> 10^8$	$> 10^8$
擦写次数	$> 10^{10}$ [22]	$> 10^{12}$	$> 10^{12}$	$> 10^{12}$	10^5—10^9	$< 10^6$	$< 10^5$
保持时间/a	> 10	> 10	> 10	> 10	> 10	> 10	> 10
存储比特	2	6[39]	1	3—4[39]	2—3	2	4

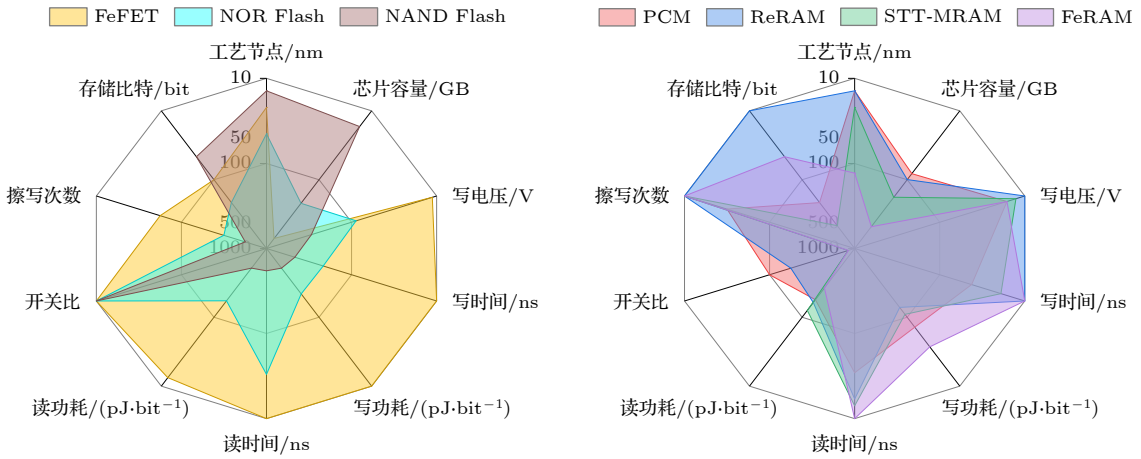


图 5 各种存储器的性能参数对比 (数据来源于表 1)

Fig. 5. Performance comparison of various memories (data was extracted from Table 1 and plotted into a radar diagram).

权重对称线性变化. 从图 5 可以看到, 新型存储器 ReRAM, PCM, STT-MRAM, FeRAM 和 FeFET 在读写速度、读写电压、擦写次数和功耗方面具有突出的优势. 然而 NAND/NOR Flash 的器件特性使得其在众多硬件方案中有不可替代的优势: 1) ReRAM, PCM, FeRAM 等二端结构器件为了防止潜行通路 (sneak path) 的产生需要集成用于选通的晶体管, NAND/NOR 的基本存储单元是三端的 MOSFET, 栅电极自带选通功能, 硬件代价小; 2) NOR/NAND flash 工艺成熟, 器件的阈值电压分布稳定, 即权重的分布稳定, 并且单元之间的性能高度一致, 适合高精度的数值运算; 3) 得益于 NAND/NOR 存储单元的栅控机制, 即阈值电压由器件单元俘获层中俘获电荷数量决定, 通过增量步进脉冲写入技术 (incremental step pulse program, ISPP) 可获得稳定的阈值电压分布, 权重 (电导率) 的模拟权重调制特性更容易实现.

NOR 和 NAND 的存储单元分别为浮栅型闪存 (floating gate flash, FGF) 和电荷俘获型闪存 (charge trapping flash, CTF), 二者均利用栅介质

中的存储电荷调控阈值电压, 主要区别在于电荷存储层的材料、电荷俘获机制和阵列的结构, 如图 6 所示. FGF 的浮栅层通常为掺杂的多晶硅, 利用热电子注入效应实现电荷的存储: 在源、漏电极和栅、源电极之间施加高电压, 电子在沟道中被源漏电场加速后, 被栅源之间的电场吸引, 穿过隧穿氧化层注入到浮栅层中. CTF 的电荷俘获层通常采用氮化硅材料, 利用 F-N 隧穿效应 (Fowler-Nordheim tunneling) 实现电荷的存储: 源、漏接地, 栅极施加高电压, 源极电子通过 F-N 隧穿效应穿过隧穿氧化层进入浮栅. 在阵列结构方面, NOR 的布局采用并行结构, 每个存储单元均有源线 (source line, SL) 和字线 (word line, WL) 引出, 随机读写的速度快, 但过多的布线和较大的器件尺寸 ($10F^2$) 使得其存储密度难以进一步提升. 相比之下 NAND 采用串行结构, 每个存储单元的源极无需单独引线, 因此具有更小的单元特征尺寸 ($4F^2$) 和更高的存储密度. 利用三维集成技术, NAND 的存储密度提高至 TB 量级. 2022 年 5 月, 国内厂商长江存储量产的 3D-NAND 已达到 192 层, 2022 年 8 月 SK Hynix 已宣布量产 238 层 3D-NAND, 如此高的集成密度碾压 NOR 以及其他新型存储器.

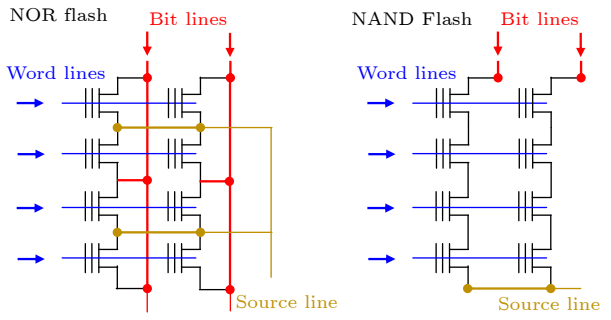


图 6 NOR 和 NAND 的电路结构

Fig. 6. Circuit structure diagram of NOR flash and NAND flash.

NAND 的高存储密度为神经形态计算提供了充足的硬件资源、更高的存储密度, 意味着芯片能分配更多的硬件资源用于映射更大规模、更多数量和种类的神经网络, 芯片性能也越强. 然而 3D-NAND 的存储单元, 电荷俘获型晶体管 (transistor with charge trap layer, CTL) 的擦写次数不高 (如表 1 所示, 次数 $< 10^5$), 因此基于 3D-NAND 的 AI 芯片适用于权重更新不太频繁的场景.

AI 系统通常涉及训练 (training) 和推断 (inference) 过程, 训练过程中需要输入大量的样本数据, 并且根据输出反馈不断调节芯片中权重的分布直至输出达到预期的精度, 使芯片具备学习能力, 这个过程涉及频繁的权重更新. 而推断过程是在已训练好的芯片上输入新的数据, 完成指定任务的过程, 权重更新频率低. 因此, 3D-NAND 芯片在执行推断任务的场景中, 得益于其超高的存储密度, 相对其他种类的存储器芯片具有显著的优势.

2 基于 3D-NAND 神经形态计算的研究进展

2.1 3D-NAND 的结构和突触特性

随着 2D-NAND 工艺微缩到 14 nm 节点, 每个单元只能容纳少量的电子, 并且单元之间电子的串扰问题使得尺寸继续微缩变得愈加困难且不够经济. 在不降低工艺节点的前提下提高存储密度和降低成本, 3D-NAND 技术成为了必然的选择. 铠侠 (原东芝)、镁光、海力士和旺宏均提出了各自的 3D-NAND 技术方案, 如图 7 所示 [40]. 这些技术方案从 3D 堆叠方式上可分为两种: 一种是栅极堆叠 (gate stack) 结构, 沟道为垂直方向; 另一种是沟道堆叠 (channel stack) 结构, 栅极为垂直方向, 如图 8 所示 [41,42].

其中存储层一般采用浮栅 (floating gate, FG) 或者电荷俘获层 (charge trap layer, CTL). 一般来说, FG 采用掺杂的多晶硅, 存储单元尺寸较大, 存储电荷量较多, 阈值电压窗口较大, 保持特性较好. CTL 材料为氮化硅, 存储电荷量相对较少, 存储单元的阈值电压窗口和保持特性略差, 但 CTL 器件的尺寸小, 集成度更高. 目前三星量产的 96 层 V-NAND 技术便是基于栅极堆叠和 CTL 层的 TCAT 结构.

神经形态计算芯片中突触是基本的结构单元, 在 3D-NAND 中用 CTL 器件作为突触 (图 9(a))[43]. 突触的权重 (电导) 非易失且连续可调, 即模拟权重调制特性. 存储用 3D-NAND 中 CTL 器件的阈值电压通过增量步进脉冲写入 (incremental step pulse program, ISPP) 机制进行调节 (图 9(b))[44]. 用于神经形态计算的 3D-NAND 中, CTL 不仅可用 ISPP 机制进行调节, 而且可用多次同脉冲写入 (multiple identical pulses program, MIPP) 机制进行调节以模拟权重调制特性 (图 9(c))[45].

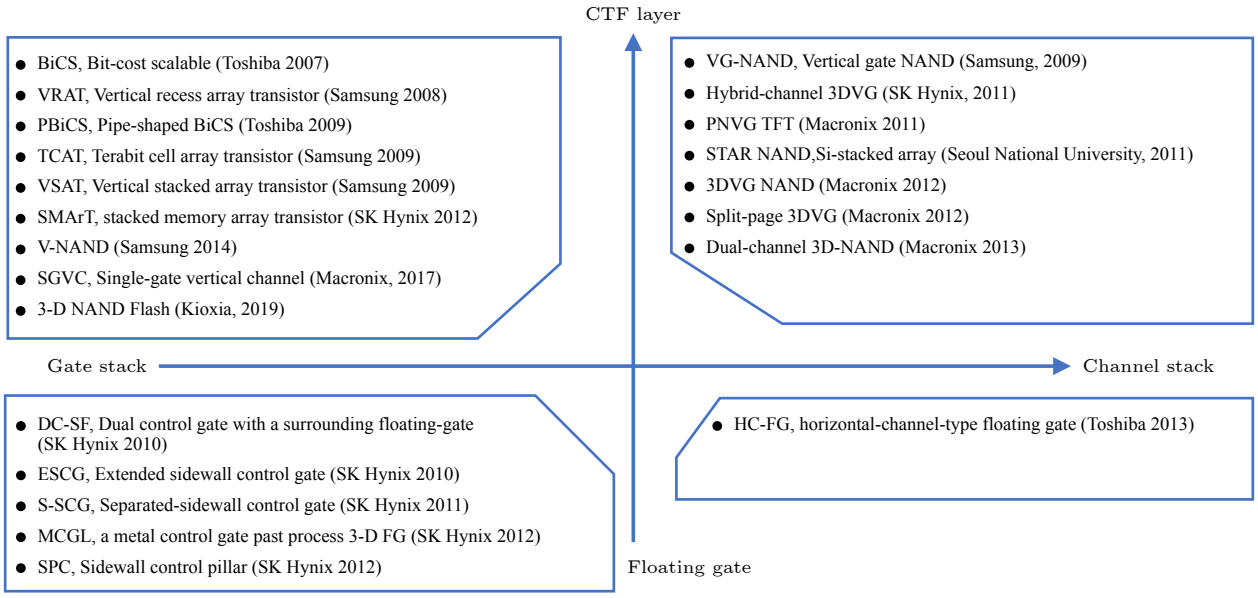


图 7 各种 3D-NAND 的技术方案[40]

Fig. 7. A summary of 3D-NAND technologies[40].

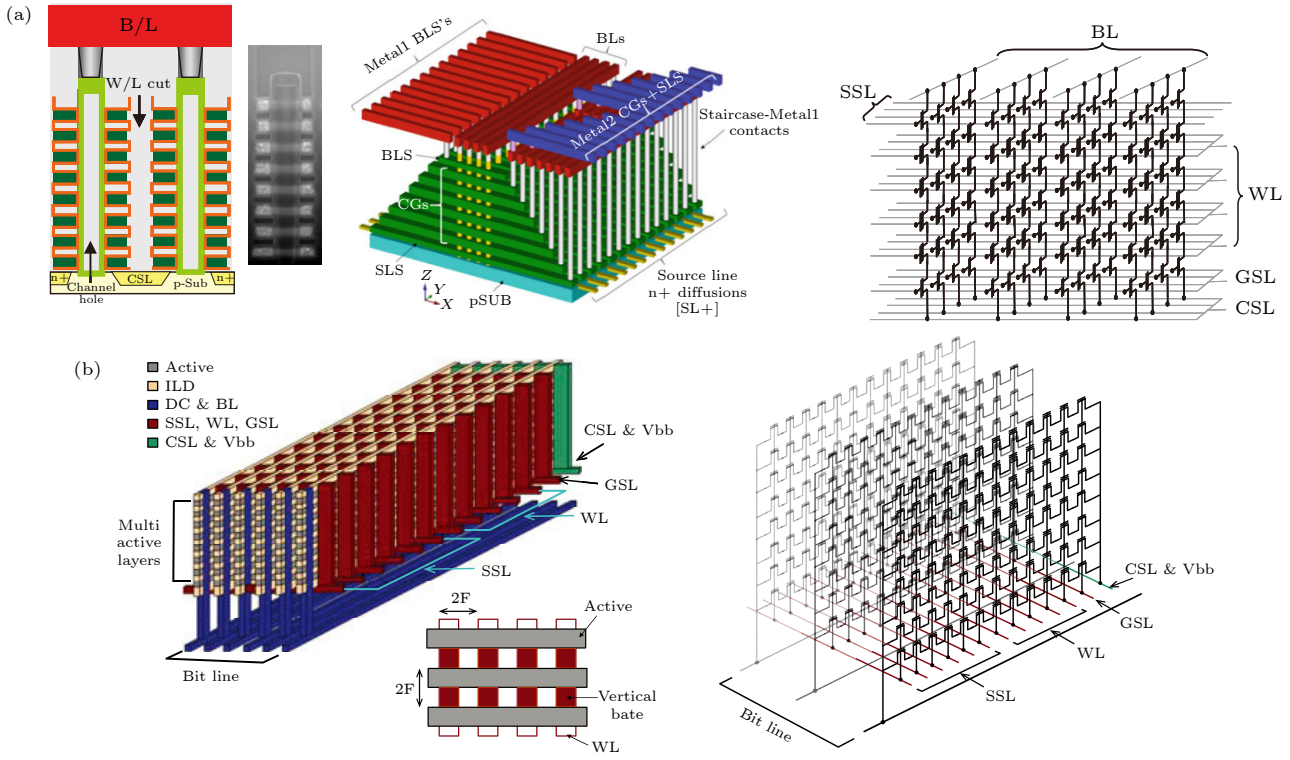


图 8 (a) 栅极堆叠的 TCAT 技术[41]; (b) 沟道堆叠的 VG-NAND 技术[42]

Fig. 8. (a) TCAT technology with gate stack architecture[41]; (b) VG-NAND technology with channel stack architecture[42].

2.2 基于 3D-NAND 神经形态计算的相关工作

神经形态计算包含 3 个过程: 前向传播、反向传播和权重更新. 前向传播, 即前级神经元发放的电压脉冲经过突触阵列转化为电流脉冲传递给后级神经元. 把前级神经元发放的输入电压脉冲和输入到后级神经元的电流脉冲视为向量, 突触的权重构成矩阵, 那么前向传播过程等价为输入信号和突触权重进行乘加运算 (multiply-and-accumulation, MAC) 的累计, 即向量矩阵相乘 (vector matrix multiplication, VMM). 反向传播过程中计算各层

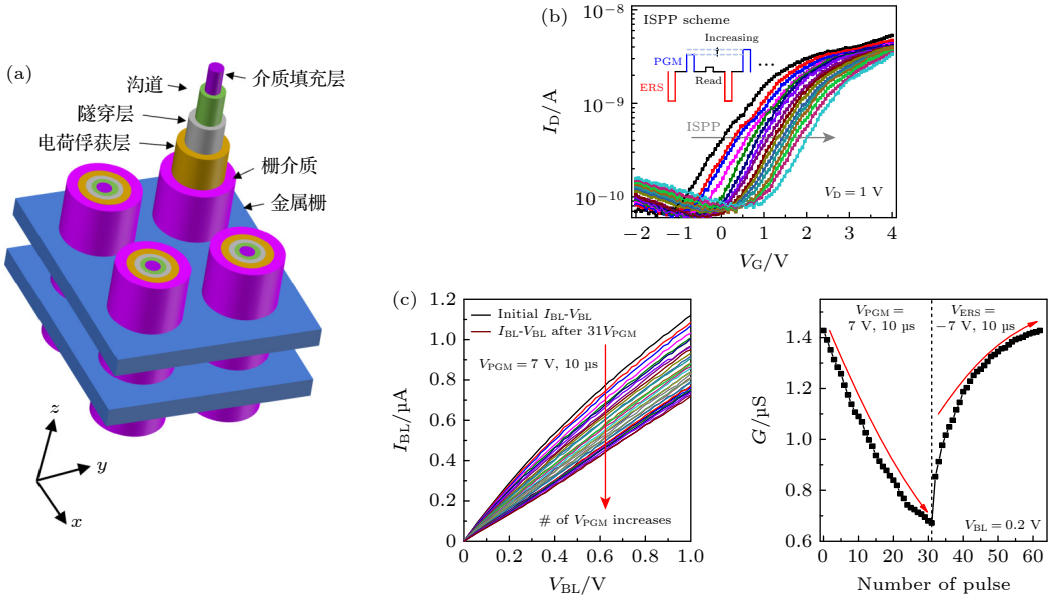


图 9 (a) CTL 单元的结构示意图 (以栅极堆叠结构为例)[43]; (b) 采用 ISPP 机制调控 3D-NAND 中 CTL 器件的阈值电压[44]; (c) 3D-NAND 中 CTL 器件的模拟电导特性, 即模拟权重调制特性[45]

Fig. 9. (a) Illustration of typical gate-stack type 3D-NAND[43]; (b) ISPP modulation of threshold voltage in CTL device[44]; (c) analog conductivity characteristics of CTL devices in 3D-NAND[45].

突触的误差: 对输出神经元的结果与目标结果进行比对, 从输出神经元往输入神经元方向逐层计算各层突触权重的误差. 最后根据计算结果更新各层突触的权重. 由于最近几年才被学术界关注, 3D-NAND 用于神经形态计算的相关报道并不多, 应用多集中在前向传播和反向传播方面.

2.2.1 3D-NAND 用于前向传播

3D NAND 是多层 2D NAND 的堆叠. 用于神经形态计算时, 3D NAND 比 2D NAND 多了层间选通的操作. 为了使读者更方便地理解输入编码、器件选通、差分对突触和权重转置的概念, 首先介绍 Lee 等[45,46] 基于 2D NAND 的神经形态计算的工作, 如图 10 所示. 图 10(a) 为 2D NAND 的神经形态计算法则, 具体地:

1) 前级神经元 X_i 与后级神经元 Y_j 之间的权重 W_{ij} 用差分对实现, 即 $W_{ij}=G_{ij}^{+}-G_{ij}^{-}$, 其中 G_{ij}^{+} 为正权重, G_{ij}^{-} 为负权重. 假设第 l –1 层有 N 个神经元 $X_{i=1,2,\cdots,N}$, X_i 输出信号 a_i^{l-1}, 第 l 层具有 M 个神经元 $Y_{j=1,2,\cdots,M}$.

2) 前向传播. 后级神经元 Y_j 接收到的电流信号为 $\sum_i^N W_{ij}a_i^{(l-1)}$, 硬件上用电压 V_i^{l-1} 作为输入施加在 NAND string 的漏端上, 即 Y_j 接收到的电流信号为 $\sum_i^N (G_{ij}^{+}-G_{ij}^{-})V_i^{(l-1)}$.

3) 反向传播. 前向传播中得到的输出与预期的输出之间往往存在误差, 为了使神经网络达到预期的识别率, 需要计算出每层突触的误差然后进行权重更新. 反向传播中将误差信号从最后一层神经元往第一层神经元传递, 实现每层突触误差的计算. 第 $l-1$ 层中神经元 X_i 接收的误差信号为

$$\delta_j^{l-1}=\sum_j^M W_{ij}\delta_j^l f'\left(s_i^{l-1}\right),$$

其中 $f'\left(s_i^{l-1}\right)$ 为神经元 X_i 激活函数的梯度. 硬件上将 δ_j^l 转化为对应的电压 V_j^l, 神经元激活函数多采用 ReLu 函数 (一阶梯度值为 1), 那么

$$\delta_j^{l-1}=\sum_j^M (G_{ji}^{+}-G_{ji}^{-})V_j^l f'\left(V_i^{l-1}\right)=\sum_j^M (G_{ji}^{+}-G_{ji}^{-})V_j^l.$$

4) 权重更新. 根据已经得到的误差, 进行权重更新, 权重的更新量 $\Delta W_{ij}=-\eta\delta_j^l f\left(s_i^{l-1}\right)$, 其中 η 为学习率. 硬件上, CTL 器件的阈值电压随写入脉冲的增大而上升, 器件电导 (权重) 减小. 因此, 当 $\Delta W_{ij}>0$ 时, 需要增大突触的权重, 那么在负突触上施加写入脉冲, 负突触权重减小 $\left|\Delta G_{ij}^{-}\right|$, 突触对权重增加 $\left|\Delta G_{ij}^{-}\right|$. 当 $\Delta W_{ij}<0$ 时, 需要减小突触对的权重, 那么在正突触上施加写入脉冲, 正突触权重减小 $\left|\Delta G_{ij}^{+}\right|$, 突触对权重减小 $\left|\Delta G_{ij}^{+}\right|$.

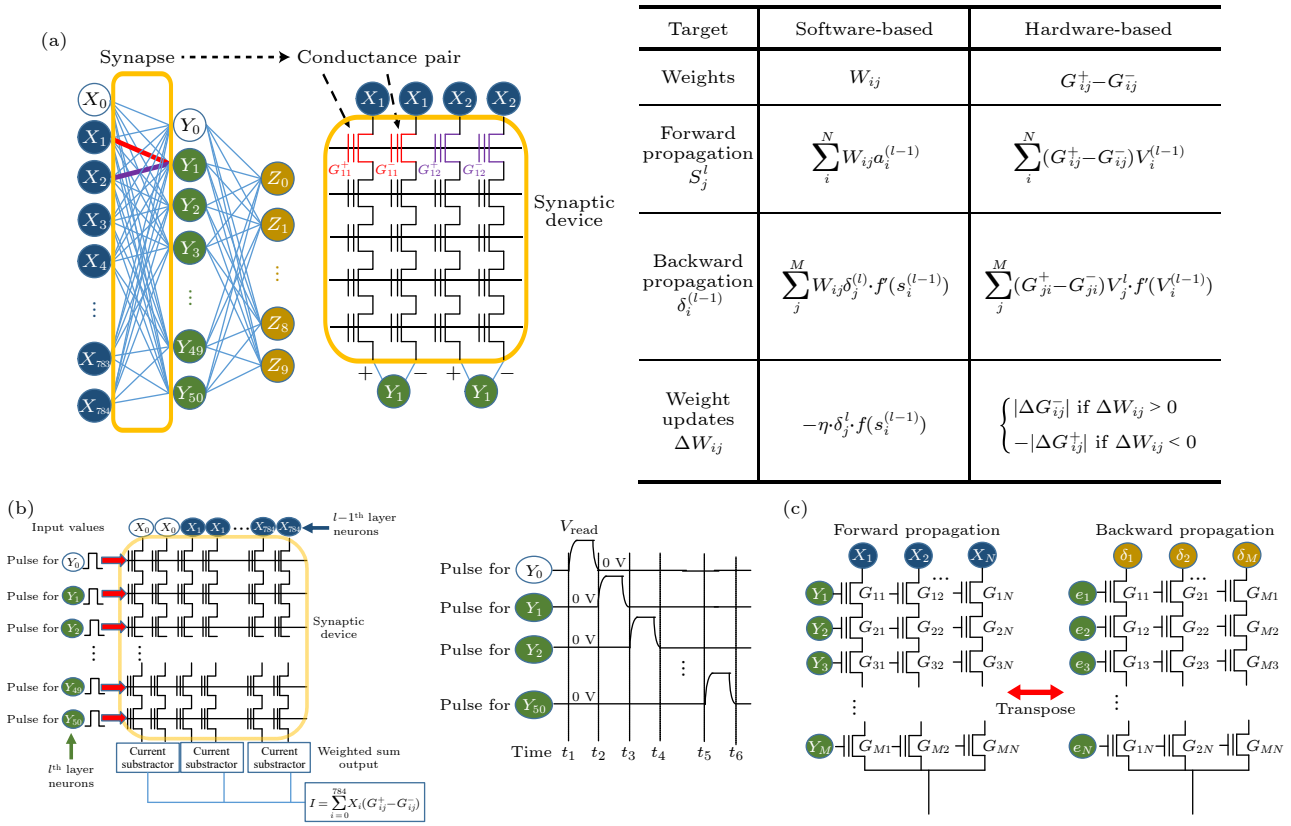


图 10 (a) 基于 NAND 的神经形态计算法则[46]; (b), (c) 前向传播和反向传播过程中 NAND 的操作方法[45]

Fig. 10. (a) Learning rule of software-based and NAND-based neural network[46]; 2D-NAND operation method in (b) forward and (c) backward propagation[45].

前向传播和反向传播的具体操作方法如图 10(b) 和图 10(c) 所示. 前向传播: 1) 输入 X_i 采用幅值编码, 施加在 BL 上. 2) NAND 的每个 page 等同于神经元 (X_0, X_1, ···, X_{785}) 对一个后级神经元 (Y_i) 进行全连接的突触. 对目标 page 施加 $V_{\rm read}$, 其他 page 施加 $V_{\rm pass}$. 其中 $V_{\rm read} < V_{\rm pass} < V_{\rm program}$, 施加 $V_{\rm pass}$ 的 CTL 器件处于导通状态, 可视为导线. 输入的电压脉冲, 经过目标 page 后转化为电流在 SL 上相加, 完成一次 VMM 过程. 3) 重复上述操作, 依次读取每个 page 的电流, 完成一次前向传播过程. 前向传播中, page 中 CTL 器件为 785 × 2 个, 一共有 51 个 page. 由于 NAND 的串行结构无法将误差信号从 SL 端输入, 反向传播过程中误差信号 (δ_1, δ_2, ···, δ_{51}) 仍从 BL 端输入, 但此时一个 page 中的 CTL 器件数为 51 × 2 个, 共有 785 个 page. 因此反向传播过程要另外选取硬件资源, 映射权重时突触阵列的配置与前向传播时的阵列互为转置, 如图 10(c) 所示.

2019 年佐治亚理工的余诗孟和清华大学的钱鹤等[47] 提出了一种基于 3D-NAND 的 VMM 方案, 如图 11(a) 所示. 需要指出的是, 通常情况下 3D-NAND 中一个 WL 控制一个平面的器件, 但图 11(a) 中每一层器件沿 Y 方向均有独立的 WL 作为输入端. 具体的 VMM 的过程为: 1) 输入信号采用二值编码, 通过地址解码器和传输门电路实现不同层的选通 (如图 11(b) 所示). 左侧传输门的输入端施加导通电压 $V_{\rm pass}$, 地址解码器连接在传输门的 PMOS 栅上. 右侧传输门的输入端施加选通电压 $V_{\rm sel}$, 地址解码器连接在传输门的 NMOS 栅上. 当输入信号为 0 时, 左侧传输门开启, 右侧传输门关闭, WL 上施加 $V_{\rm pass}$ 电压. 当输入信号为 1 时, 左侧传输门关闭, 右侧传输门打开, WL 上施加 $V_{\rm sel}$ 电压. 2) SL 上施加读电压 $V_{\rm read}$, 经过目标层突触转化为电流在 BL 上读取. 从 XY 平面上看, CTL 器件以经典的 crossbar 形式排列.

余诗孟、钱鹤等[47] 提出的基于 3D-NAND 的 VMM 方案, 并未用于神经形态计算. 如果沿用此方案进行神经形态计算, 其优势在于同一平面内的 CTL 单元为 crossbar 结构, 反向传播可采用相同的操作方法, 误差信号 δ_i 从 WL 输入, SL 上施加

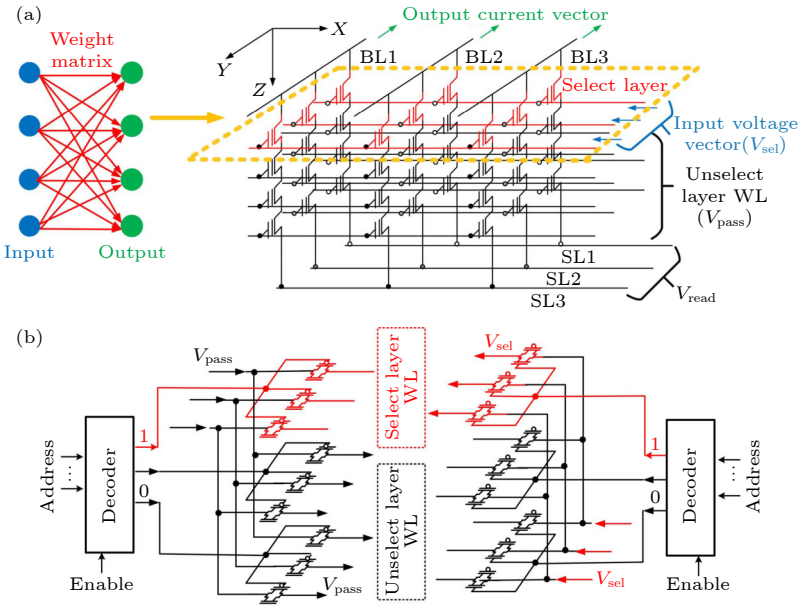


图 11 一种基于 3D-NAND 的 VMM 方案[47] (a) 3D-NAND 的电路结构和 VMM 的操作方法; (b) 用于 3D-NAND 层间选通的外围电路结构

Fig. 11. A case of using 3D-NAND for VMM operation[47]: (a) Circuit diagram and bias scheme of 3D-NAND array architecture for VMM; (b) peripheral circuitry for layer-to-layer selection.

V_{read}, BL 上读取电流, 无需另外选用硬件资源. 并且后续的权重更新可通过 WL 上施加电压脉冲实现. 但进一步的工作需要考虑两个方面: 1) 目前不存在这种结构的 3D-NAND. 同一平面上制备独立的 WL 将大大增加工艺难度、单元尺寸和引线的复杂度, 硬件实现难度大; 2) 栅压有无作为输入 (二值编码), 那么所选通的器件必须工作在饱和区, CTL 只能有 1 bit 的存储态, 如果用一个 CTL 代表一个突触, 那么神经网络的精度只有 1 bit. 要进一步提高神经网络的精度, 可以用多个器件等效为一个具有多比特精度的突触, 虽然硬件代价成倍增加, 但 3D-NAND 的大容量可轻松满足其需求.

2019 年 Lee 等[48] 研究了如何用 2D NAND 实现二值神经网络 (binary neural network, BNN) 的前向传播, 如图 12 所示. 二值神经网络的输入、权重和输出均只有两种状态, 运算过程较为简单, 适合对精度要求不高的推断 (inference) 过程. BNN 的运行只需要器件有两个稳定可区分的阻态, 对硬件要求不高. 图 12(a) 展示了用 SLC NAND 进行二值运算的原理 (single level cell, SLC 即存储态为 1 bit 的 CTL 器件), 输入信号施加在源端用于开关的晶体管的 WL 上 (即 bit line selector, BLS 或者 select gate at drain side, SGD). 同一个 page 上两个相邻的 SLC 器件构成一个突触对, 两个器件中有且只有一个为写入状态. 左侧的 string 上的 BLS 有电压, 右侧无电压时, 输入标记为+1, 反之标记为–1. 突触对左侧 SLC 器件为未写入状态 ($V_{th,\,low}$), 右侧为写入状态 ($V_{th,\,high}$) 时, 突触权重标记为 $W=+1$, 反之 $W=-1$. 输入信号为+1, 当 $W=+1$ 时, SL 端输出电流 I_{SL}, 如果将 I_{ref} 设为 $0.5\times I_{SL}$时, 电流经过差分放大器, 输出 $+0.5\times I_{SL}$, 转化为正电压, 标记为+1. 反之当 $W=-1$ 时, SL 端无电流输出, 电路输出 $-0.5\times I_{SL}$, 读出负电压, 标记为–1. 同理, 当输入信号为–1 时, 权重分别为+1 和–1 时, 输出分别为–1 和+1. 这种输入和权重状态相同才有电流输出的过程等效为 XNOR 逻辑运算.

BNN 前向传播的原理如图 12(b) 所示. 其中一个 SLC 器件依次对应一个前级神经元对后级多个神经元的突触. 前级神经元的输入视为一组向量, 同时输入到 NAND 的 BLS 上, 如图 12(c) 所示. 基于这种硬件方案, Lee 等[48] 设计了二值全连接神经网络以及二值卷积神经网络分别用于 MNIST 和 CIFAR-10 图片数据的识别, 网络的训练次数与识别率的关系如图 12(d) 所示, 识别率分别达到 98.12% 和 87.11%.

2019 年 Lue 等[49] 提出一种用基于 SLC 3D-NAND 的卷积核映射和前向传播方案, 用多个 BL 输入和 SLC 器件实现 4 bit 精度的输入和 4 bit

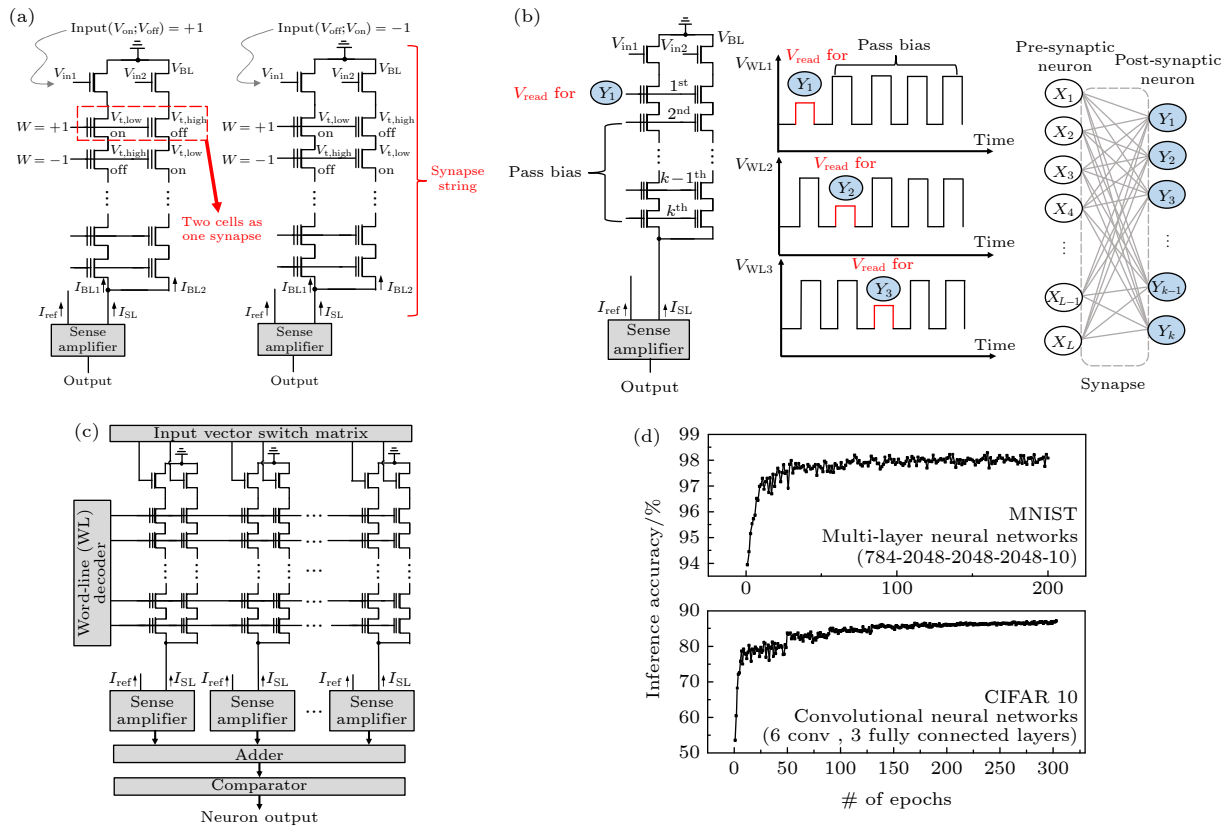
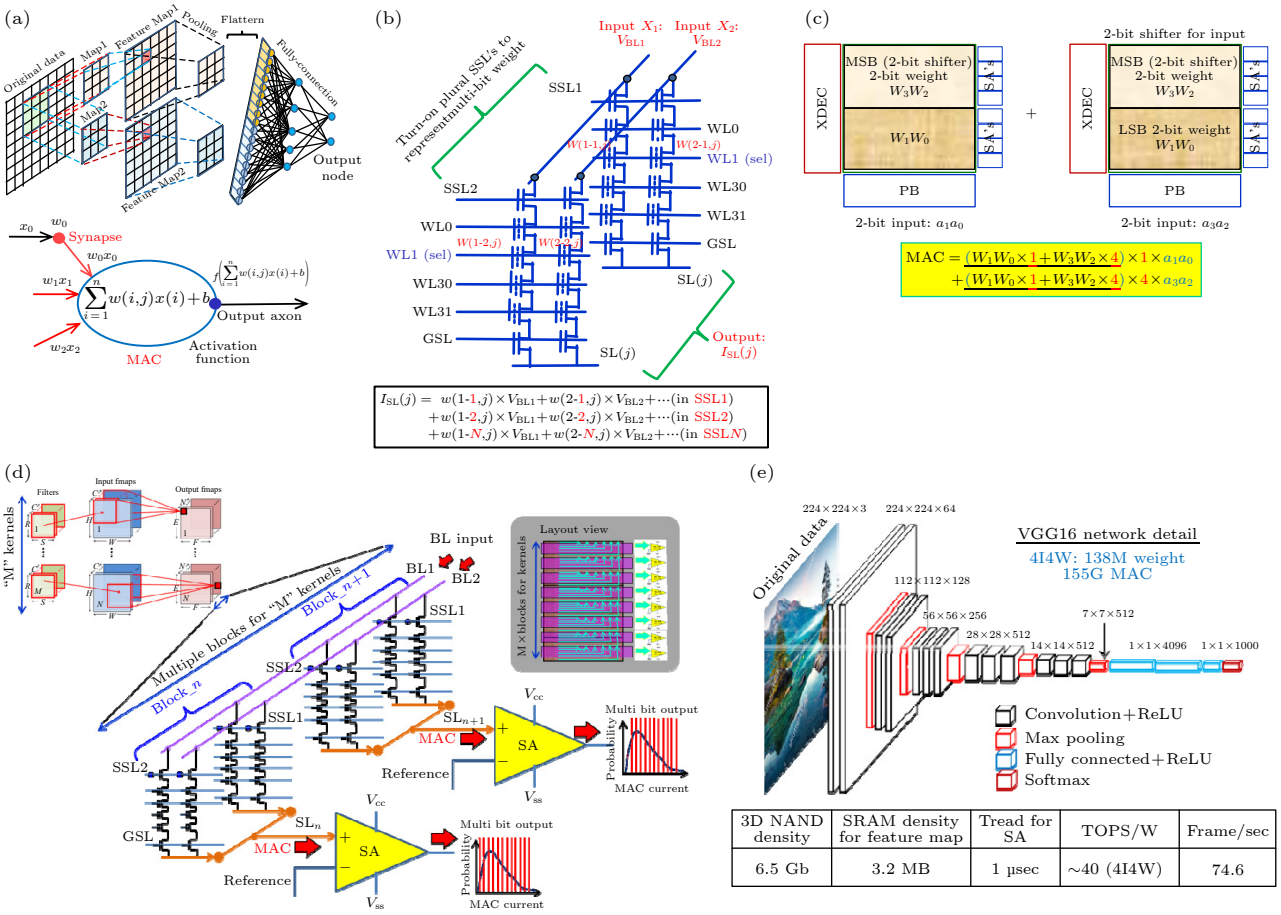


图 12 基于 2D NAND 的二值神经网络 BNN[48] (a) 相邻的两个 CTL 器件组成差分对形式的突触, 与输入的信号进行同或运算 (XNOR); (b), (c) BNN 前向传播中 NAND 的操作方法以及 NAND 电路示意图; (d) 采用二值全连接神经网络和二值卷积神经网络分别用于 MNIST 和 CIFAR-10 图像库的识别性能

Fig. 12. A synaptic architecture based on 2D NAND for binary neural network (BNN) [48]: (a) NAND string structure for XNOR operation, in which two neighboring CTL device constructs a differential pair as one synapse; (b) operation scheme for forward propagation; (c) schematic diagram of synaptic array architecture; (d) the performance of binarized multi-layer and convolutional neural networks for MNIST and CIFAR-10 database recognition task respectively.

精度的权重, 如图 13 所示. 卷积神经网络 CNN 往往采用多个卷积核, 对图片进行卷积操作需要进行大量的 MAC 运算, 原理如图 13(a) 所示. 图片中每个像素对应一个输入信号, 一次卷积的过程等同于像素输入与卷积核中对应的权重进行 MAC 运算 (即前向传播). 卷积核中的权重具有多 bit 精度, 可以将权重拆分为高位和低位, 存放在多个 SLC 中, 例如 2 个 SLC 可以实现 2 bit 存储状态. 2 bit 精度的输入和 2 bit 精度的权重的 MAC 过程 (2 bit input & 2 bit weight, 2I2W) 如图 13(b) 所示. 输入信号的两位 $X1\langle 0\rangle$和 $X1\langle 1\rangle$先后施加在 BL1 上, 权重的两位分别存储在 2 个 SLC 器件上, 即 $W(1\text{-}1,\ j)$ 和 $W(1\text{-}2,\ j)$. MAC 过程的得到的总电流为 $X_1\langle 0\rangle \times [W(1\text{-}1,j)+2\times W(1\text{-}2,j)]+2X_1\times [W(1\text{-}1,j)+2\times W(1\text{-}2,j)]$, 分 4 次相乘后移位相加获得. 依次类推, 4 bit 精度的 MAC 过程 (4I4W) 可拆分为 2 组 2 bit 的输入和 2 组 2 bit 的权重进行相乘后移位相加, 工作原理如图 13(c)所示. 图 13(d) 展示了卷积核电路的工作原理. 图中沿 BL 方向划分了 M 个 block, 每个 block 代表不同的通道, 每个 block 中的每一层代表一个卷积核. 输入信号从 BL 进入, 3 根 BL 构成一个 2 bit的输入, 一个 4 bit 精度的输入信号需要 6 根 BL(图中只绘出一根). 3 个 SLC 器件构成一个 2 bit 精度的权重, 4 bit 精度的权重需要 6 个 SLC, 对应地需要 6 根 SSL(图中只绘出 2 根), 因此 4I4W 过程需要 36 个 SLC. 对于更高精度的输入和权重, 以及更多的图片像素 (文中不考虑输入端口复用), 需要更多的 BL 和 SSL, 阵列具有相当的规模. Lue 等[49] 并未从硬件上实现卷积核功能, 他们测试了 64 GB SGVC NAND 中单元的电性能. 最后基于 4 bit 精度的 3D-NAND 卷积核电路, 运行 VGG-16 卷积神经网络对 CIFAR-10 数据库进行识别, 识别率达到 90%.



3D NAND density	SRAM density for feature map	Tread for SA	TOPS/W	Frame/sec
6.5 Gb	3.2 MB	1 μsec	~40 (4I4W)	74.6

图 13 采用 SLC 3D-NAND 实现 4 bit 精度的卷积神经网络方案[49] (a) 卷积神经网络的工作原理示意图 (上), 涉及大量的 MAC 过程 (下); (b) 对于多 bit 权重的 MAC 过程, 用多个 SLC 器件构成一个多 bit 权重; (c) 4 bit 精度输入与 4 bit 精度权重的 MAC 原理; (d) 卷积核电路的工作原理; (e) VGG 16 神经网络的结构图以及用所设计的 3D-NAND 加速 VGG 16 的性能

Fig. 13. A case of SLC 3D-NAND for convolution neural network (CNN) with 4-bit resolution[49]: (a) Flow schematic of CNN; (b) in a MAC array, plural SSLs to stand for multi-bit weight; (c) the arithmetic principle of MAC with 4-bit input and 4-bit weight (4I4W); (d) convolutional core circuit and working principle diagram; (e) schematic diagram of VGG 16 CNN and the simulated performance of 3D-NAND hardware implementation.

2019 年, Kim 等[50] 提出了一种基于 3D-NAND 的卷积核映射方案, 用 4 个 MLC (multi-level cell, 2 bit 存储态的 CTL 器件) 构成了 8 bit 存储态的突触, 并实现了 8 bit 精度的 CNN 卷积运算, 如图 14 所示. 图 14(a) 中将 BiCS 结构的 3D-NAND 中的每个 block 按 BL 方向展开成 2D-NAND, 每个 SGD 线 (select gate at drain side, SGD) 连接一个前级神经元接收输入信号. 具体的卷积核映射方法如图 14(b) 所示. 一个 8 bit 的权重映射到一个 string 上相邻的 4 个 MLC 中 (P_{15}/N_{15}, P_{14}/N_{14}, P_{13}/N_{13}, P_{12}/N_{12}). 两个相邻的 block 中同一个 SGD 线控制的两个 string 上的权重分别标记为正、负, 构成差分对. 差分对中权重为正时, 负权重设为 0, 反之权重为负时, 正权重设为 0. Kim 等[50] 在 2021 年的报道中增加了对卷积运算具体过程的阐述, 如图 14(c) 所示. 其过程为: 1) 两个输入信号的第一位和权重的前两位进行乘加运算后得到 4, 下一步两个输入信号的第一位和权重的下两位进行乘加运算得到结果 3, 两次运算结果进行移位相加, 即 $3\times2^2+4\times2^0$; 2) 图中有 24 根 SGG 线, 支持 24 个神经元信号同时输入进行乘加运算, 乘加运算 32 次后移位相加, 得到 24 个前级神经元对同一个后级神经元的输出. 同一个 string 上有 16 个 MLC, 可存储 4 个 8 bit 权重, 图 14(b) 的电路可以映射 24 个前级神经元和 4 个后级神经元之间全连接的突触.

由于 NAND 厂商禁止开放 WL 和 SGD 端口的控制, Kim 等[50] 用自研的 16 层 die 堆叠的 eNAND 等效 3D-NAND(具体的 eNAND 结构和工作原理不在此赘述). 图 14(d) 中展示了基于 eNAND 的

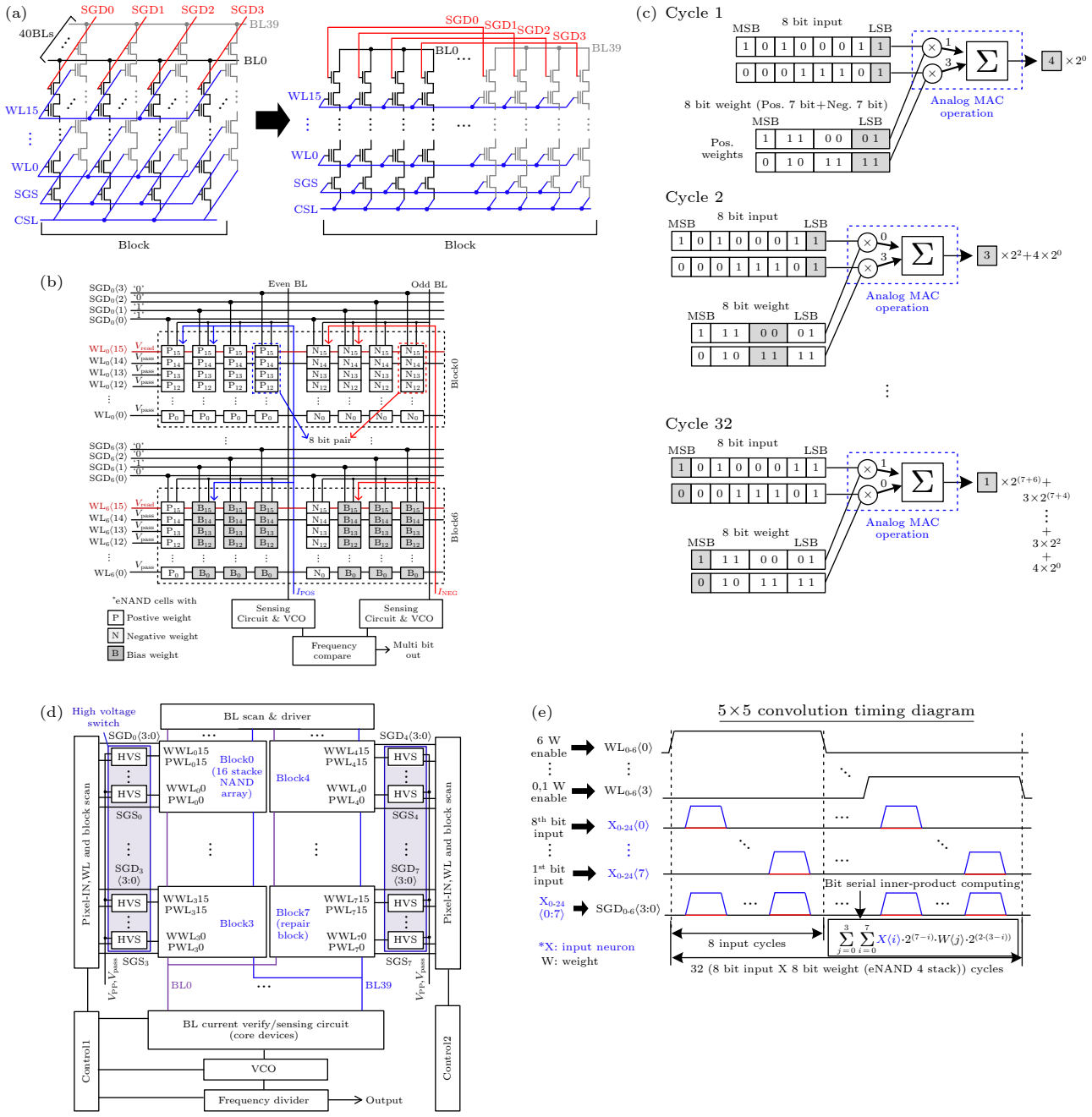


图 14 基于 MLC 3D-NAND 的 8-bit 精度卷积方案[50,51] (a) 基于 BiCS 结构的 3D-NAND 电路图[51]; (b) 权重的映射方式, 正、负权重存储在相邻的两个 block 中[51]; (c) 2 个 8 bit 精度的输入信号和 2 个 8 bit 精度的权重的乘加运算过程[51]; (d) 基于 eNAND 的卷积核电路, 有 7 个 block, 28 个输入端口, 满足 5 × 5 卷积核的功能[51]; (e) 卷积过程的信号时序图[51]

Fig. 14. A case of MLC 3D-NAND for CNN with 8-bit resolution[50,51]: (a) Circuit diagram of BiCS type 3D-NAND, the 3D structure can be flattened into a 2D structure[51]; (b) weight mapping method, positive and negative weight stored in two neighboring blocks[51]; (c) MAC operation principle of 2 inputs and two weights with 8-bit resolution[51]; (d) convolutional core circuit diagram with 7 blocks and 28 input ports can be used for 5 × 5 convolution operation[51]; (e) timing diagram of 5 × 5 convolution operation with 8-bit data and 8-bit weights[51].

卷积核电路, 有 7 个卷积用的 Block 和一个用于修复的 Block. 此电路具有 28 个输入端口, 可以映射 5 × 5 大小的卷积核, 卷积的脉冲时序如图 14(e) 所示. 基于此硬件方案, Kim 等[50] 设计了 LetNet-5 进行 MNIST 手写数字识别, 识别率达到 98.5%.

2021 年, Kim 课题组[52] 提出了一种基于 3D NAND 的卷积核映射方案, 用 1 bit 的 MAC 算子进行分部相乘 (partial multiplication) 并采用 Booth 编码映射正负权重, 显著地降低了 VMM 过程中的电流, 提高了芯片的能效, 文中将这种芯片架构

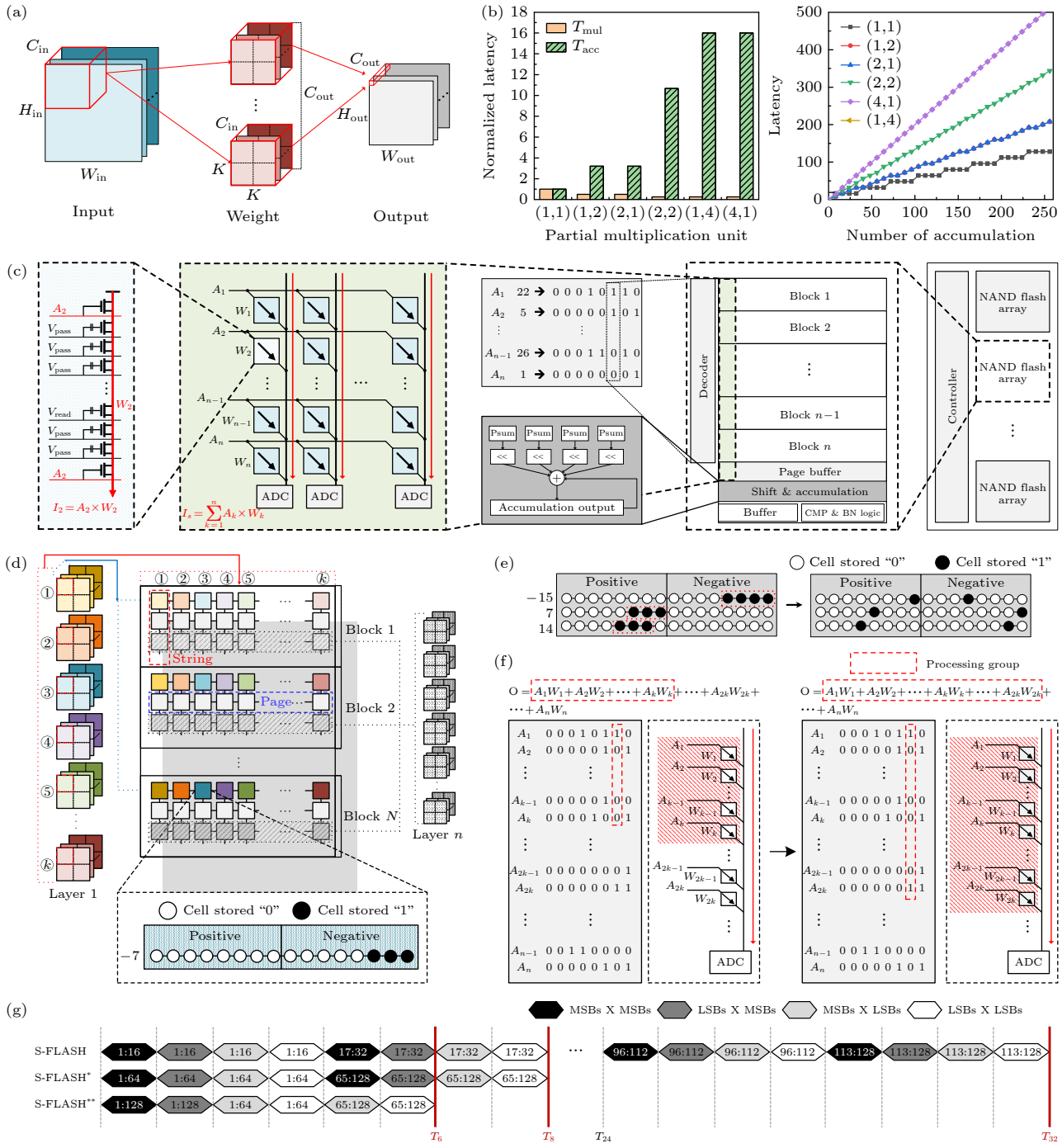


图 15 基于 3D NAND 的 S-Flash 芯片用于卷积神经网络加速[52] (a) 卷积过程示意图; (b) MAC 算子的比特对乘、加运算延迟时间的影响 (左) 和 MAC 算子的比特对累加运算延迟时间的影响 (右); (c) S-Flash 芯片的架构; (d) S-Flash 中权重分布的示意图, 用 16 个 SLC 构成一个差分结构的突触; (e) 通过 Booth 编码分配权重; (f), (g) 同时操作的 SSL 和 BL 增大 1 倍, MAC 次数缩减了 1/4

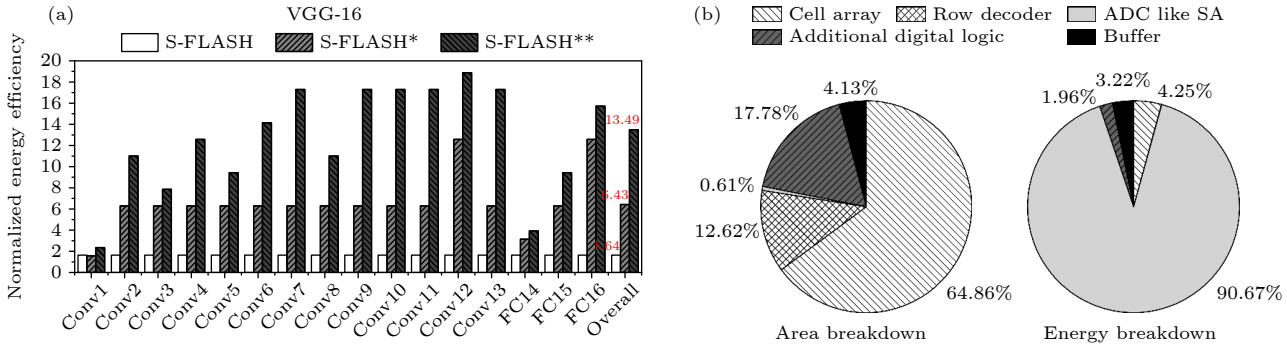
Fig. 15. 3D NAND-based CNN accelerator named as S-Flash[52]: (a) Convolutional operation of CNN; (b) normalized latency for multiplication, accumulation (left) and MAC operation in various multiplication units (right); (c) overall S-FLASH architecture; (d) overall weight data layout, in which a differential synapse constructed with 16 SLC; (e) weight allocation by Booth coding; (f), (g) double the concurrently operated BLs and SSLs, 4 times faster the MAC operation speed.

命名为 S-Flash. 图 15(a) 为卷积运算的过程, 一个卷积层由 k 个尺寸为 $K \times K \times N$ 大小的卷积核构成. 输入信号和卷积核的精度通常为 8 bit, 那么卷积核在图像上进行一次滑动将产生 $k \times K^2 \times N$ (8 bit×8 bit) 次 MAC 运算. 如何优化 8 bit×8 bit 计算过程是文献 [52] 研究的重点. 文中将输入信号和权重拆分为低比特的算子进行分部相乘. 将输入信号和权重的比特分别记为 (B_a, B_w), 当 $(B_a, B_w) =$

(1, 1) 时, 进行 64 次乘法运算. 当 (B_a, B_w) = (1, 2) 或 (2, 1) 时, 进行 32 次乘法运算. 图 15(b) 展示了 (B_a, B_w) 对乘法和积分对电路延时的影响, 随 (B_a, B_w) 增大, 乘法运算的延时略有降低, 但电流积分的延时显著增大. 这是因为 MAC 运算的速率受限于 ADC 精度, ADC 精度越高, 外围电路的开销、延时和能耗越大. 采用一般精度的 ADC(如 3-bit/4-bit), (B_a, B_w) 越大, 电流积分的次数越多, 电流积分时间显著延长. 文中采用 (B_a, B_w) = (1, 1) 作为 MAC 运算的基本操作单位. 图 15(c) 展示了 S-Flash 的架构, 神经网络的每一层突触映射到沿 WL 的每一层 SLC 中, 输入信号 A_1—A_n 的 8 位信号依次施加在 BL 上, MAC 电流用 SSL 收集, n 为卷积核的通道数, 用一个 block 映射一个卷积核的一个通道, n 个 block 构成一个 array 映射一个卷积核, 多个 array 映射一个卷积层. 图 15(d) 展示了卷积核的映射过程, 其中突触权重的精度为 8 bit, 16 个 SLC 构成一个具有正、负权重的突触对. 图 15(e) 展示了通过用 Booth 编码正、负权重, 以增加权重中的“0”位, 提高 MAC 过程中的稀疏性, 从而降低 MAC 过程中的总电流. 为了进一步提高计算效率, 将 $A_k \times W_k$ 对应的 SLC 阵列中 BL 和 SSL 增大一倍, SLC 单元增大 2 倍, 原先 8 bit × 8 bit 的 MAC 次数从 64 缩减为 16, 如图 15(f) 和图 15(g) 所示, 将 MAC 过程中具有 4 倍 SLC 规模但权重未经过 Booth 编码的 S-Flash 定义为 S-Flash*, 具有 4 倍 SLC 规模并且权重经过 Booth 编码的 S-Flash 定义为 S-Flash**.

最后通过电路仿真研究了 S-Flash 芯片在运行卷积神经网络时的性能参数. 分别用 S-Flash, S-Flash*和 S-Flash**运行 VGG-16 神经网络, 计算了前向传播中每层神经网络的能效, 并以 GPU 方案时的能效为标准进行归一化, 结果如图 16(a) 所示. 从图 16(a) 可以看出, S-Flash, S-Flash*和 S-Flash** 的能效分别是 GPU 的 1.64, 6.43 和 13.49 倍, S-Flash*和 S-Flash**的能效分别是 S-Flash 的 3.9 倍和 8.2 倍, 所以正、负权重的分配对芯片能效的影响最大. S-Flash**中各电路模块的面积和能耗对比如图 16(b) 所示, 可以看到 ADC 面积占芯片面积的 0.61%, 但能耗占 90.67%. 图 16(c) 中表格列出了 S-Flash**和其他芯片的性能参数, 包括存储密度、单位面积的峰值吞吐量和能效, 可以看到 S-Flash**在这 3 个方面均有明显优势.

2021 年, 余诗孟课题组 [53] 提出了一种基于 3D NAND 的神经形态芯片架构, 用于卷积神经网络加速. 图 17(a) 所示为在 3D NAND 的一个



(c) Comparison with other platforms

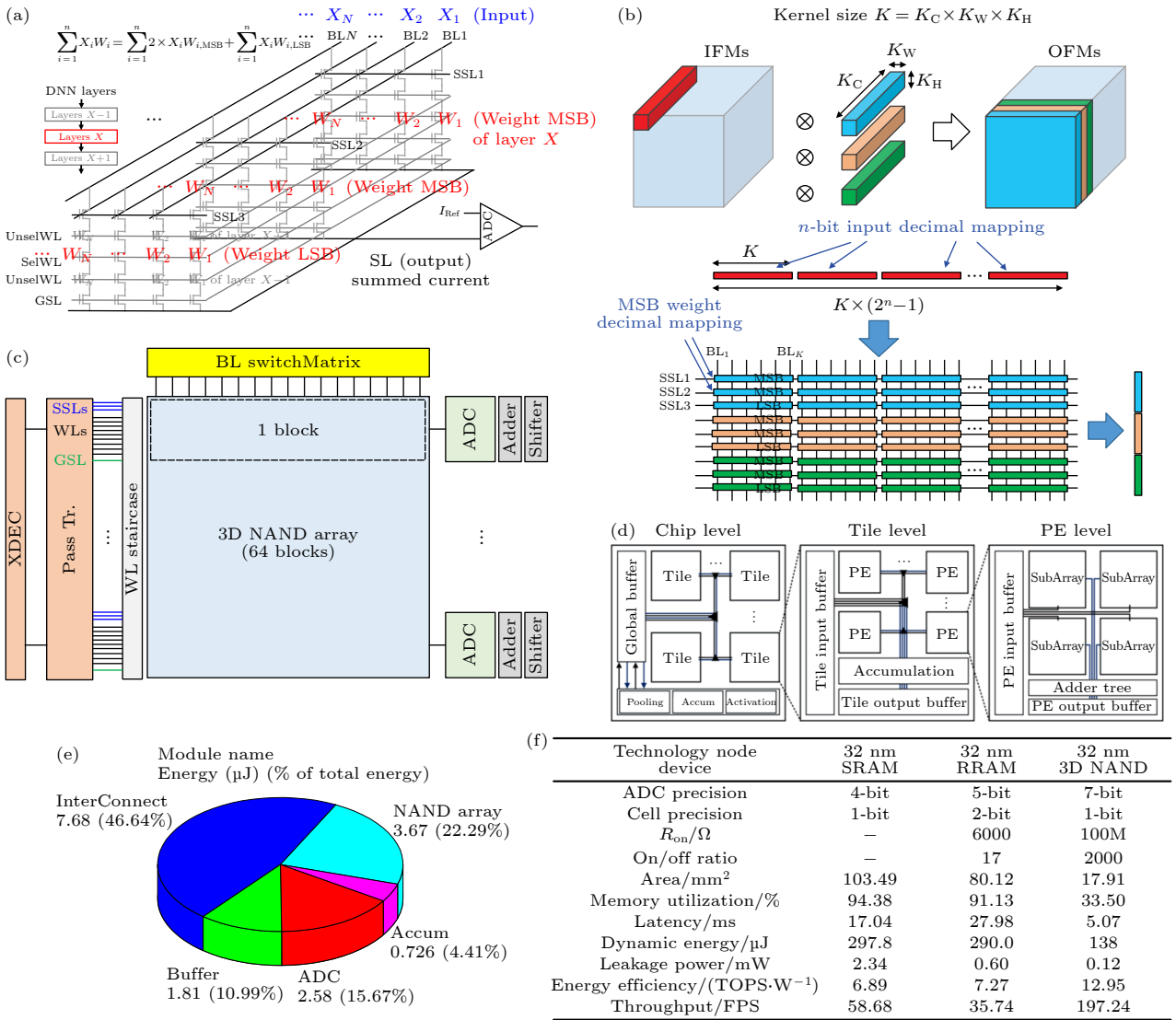
	Titan-x [28]	Eyeriss [29]	DaDianNao [30]	Neural Cache[3]	DRISA [4]	SRAM with PA[31]	eDRAM PIM[19]	ReRAM PIM[7]	ReRAM PIM[15]	NOR flash PIM[10]	S-FLASH** (our work)
Type/platform	GPU	ASIC		Digital PIM		Analog PIM					
				SRAM	DRAM	SRAM	eDRAM	ReRAM		NOR flash	NAND flash
Memory density/ (MB·mm^{-2})	7.27	0.072	0.82	0.55	3.97	0.98	2.91	1.57	0.04	4.36	13.96
Peakth roughput per area/ (GOPS·mm^{-2})	3.98	53.3	126.92	191.57	6.39	N/A	N/A	3185	68.3	43.68	41.87
Energy efficiency/ (GOPS·W^{-1})	26.76	92.44	286.40	~70.44	~16.5	125.86	531.25	1972	3793	169.63	313.60

† The process node of all platforms is scaled to 22 nm
‡ The MAC operation is normalized to INT8 operation

图 16 S-Flash 电路仿真的结果 [52] (a) S-Flash/S-Flash*/S-Flash**运行 VGG-16 卷积神经网络时的能效; (b) S-Flash**电路模块的面积对比和运行 VGG-16 时各电路模块的能耗对比; (c) S-Flash**与其他芯片的性能参数对比

Fig. 16. Simulation result of S-Flash[52]: (a) Energy efficiency evaluation result of each VGG-16 layer accelerated by S-Flash/S-Flash*/S-Flash**; (b) area and energy breakdown of S-FLASH**; (c) comparison with the other platform.

block 中进行 VMM 操作的方法: 沿 WL 方向的每一层 SLC 器件对应神经网络的每一层突触, 输入信号 X_N 施加在 BL 上, 权重 W_i 精度为 2 bit, 用 3 个 SLC 代表一个 W_i, 3 根 SSL 用于突触选通. X_N 采用十进制编码, 即精度为 n bit 的 X_N, 可用 2^n–1 根 BL 来表示. 用十进制编码的优势在于 block 中所有 string 上的电流可直接相加后进行模数转换, 而无需移位操作. 图 17(b) 中展示了 2 bit 权重精度的卷积核映射方案, 卷积核尺寸为 $K = K_C \times K_W \times K_H$, X_N 精度为 n bit, 卷积核中一个单元的权重映射到 $3 \times (2^n - 1)$ 个 SLC 中. 图 17(c) 为多个 block 构成的 subarray, 用于映射一个卷积层. VGG-8 卷积神经网络中, 最大的卷积层中有 16 个卷积核, W_i 精度为 2 bit, X_N 精度为 8 bit. 用一个 block 映射精度为 2 bit 的卷积核, block 的大小为 $521 \times 3 \times 3 \times (2^2 - 1) \times (2^2 - 1) \times 32$ WL = 1.27 Mb. 考虑到 VGG-8 中权重的精度为 8 bit, 采用 4 个 2 bit 精度的 block 通过分部相乘实现, 那么 1 个 subarray 的大小为 16 × 4 × 1.27 Mb = 81 Mb. 图 17(d) 为芯片架构的示意图, 芯片中有 4 个 tile, 每个 tile 有 4 个 PE(process element), 每个 PE 有 4 个 subarray. VGG-8 神经网络需要 110 Mb SLC, 所设计的芯片容量完全满足需求. 用 HSPICE 计算了 3D NAND 运行 VGG-8



(f)

Technology node device	32 nm SRAM	32 nm RRAM	32 nm 3D NAND
ADC precision	4-bit	5-bit	7-bit
Cell precision	1-bit	2-bit	1-bit
R_{on}/Ω	–	6000	100M
On/off ratio	–	17	2000
Area/mm^2	103.49	80.12	17.91
Memory utilization/%	94.38	91.13	33.50
Latency/ms	17.04	27.98	5.07
Dynamic energy/μJ	297.8	290.0	138
Leakage power/mW	2.34	0.60	0.12
Energy efficiency/(TOPS·W^{-1})	6.89	7.27	12.95
Throughput/FPS	58.68	35.74	197.24

图 17 输入信号采用十进制编码的 3D NAND 芯片用于卷积神经网络加速[53] (a) 用 3D NAND 做 VMM 的操作方法; (b) 卷积核映射的方案; (c) subarray 的结构示意图; (d) 芯片的架构示意图; (e) 用 3D NAND 芯片运行 VGG-8 神经网络用于 CIFAR-10 图片库识别时, 各电路模块的能耗对比; (f) 3D NAND 与其他芯片的性能对比

Fig. 17. A 3D NAND CNN accelerator with decimal input coding[53] (a) VMM operation method by using 3D NAND; (b) the mapping method of a CNN kernel; (c) designed subarray configuration; (d) hierarchy of the 3D NAND-based neuromorphic chip architecture; (e) energy breakdown of 3D NAND-based chip on VGG-8 network for the CIFAR-10 dataset; (f) comparison with other chips.

网络用于 CIFAR-10 图像识别任务的性能, 电路中各模块的能耗占比如图 17(e) 所示. 图 17(f) 列出了 3D NAND 的性能参数, 并与 RRAM 和 SRAM 做比较, 可以看到所设计的 3D NAND 在各项性能指标上均有明显的优势.

2022 年, 霍宗亮课题组[54] 利用 3D NAND 中 CTL 的模拟权重调制特性, 设计了两层全连接神经网络, 用 Winner-takes-all(WTA) 非监督学习算法实现 ZVN 图像的识别. 图 18(a) 为 3D NAND 的操作示意图, 不同于其他文献, 文中阵列的输入信号施加在 WL 上, 输入信号为 $V_{\rm read}$ 和 $V_{\rm pass}$, 分别代表输入像素点的明暗两种状态, 即 1 和 0. 突触映射在 WL 平面的 CTL 阵列中, 采用差分对结构, 输出电流在 BL 端收集并通过 TSG (top select gate transistor) 控制. 图 18(b) 中为 CTL 器件的模拟权重调制特性, 可以看到 CTL 器件的长时程增强 (long term potentiation, LTP) 和长时程抑制 (long term depression, LTD) 过程具有良好的线性度. 图 18(c) 展示了用 3D NAND 实现非监督学习的训练过程. 首先以相同的概率输入标准的 Z, V, N 图像然后以 WTA 法则进行权重更新. WTA 的权重更新法则: 以图像“Z”为例, 当输入

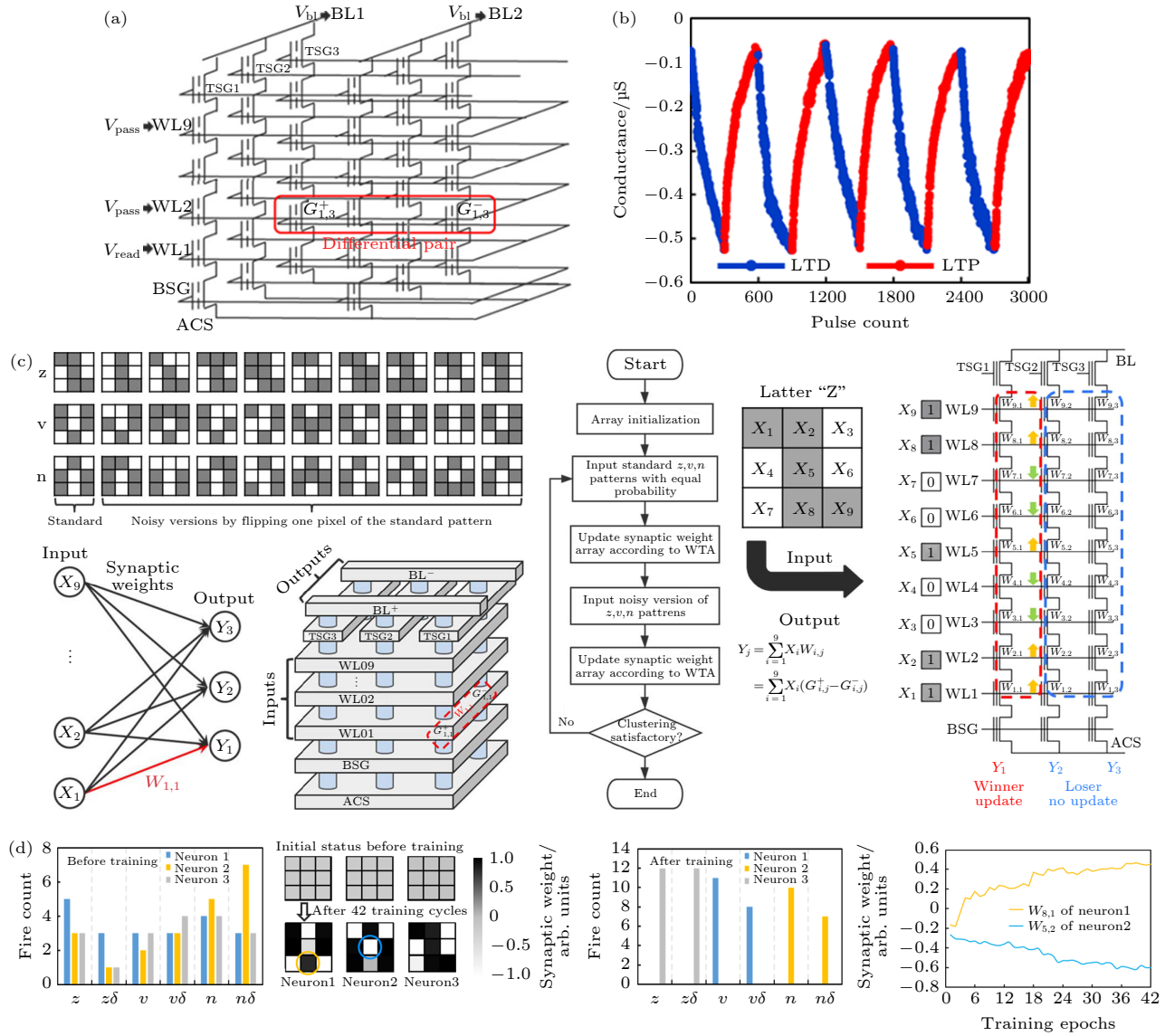


图 18 基于 3D NAND 的 WTA 神经网络用于非监督学习[54] (a) 具有差分对结构的 3D NAND; (b) CTL 器件的模拟权重调制特性; (c) 用 3D NAND 训练 WTA 神经网络的过程; (d) WTA 神经网络的训练结果

Fig. 18. A 3D NAND-based WTA neural network for unsupervised learning[54]: (a) Schematic of the differential pair in 3D NAND flash array; (b) analog weight modulation of measured CTL device; (c) the training procedure of WTA neural network using 3D NAND array; (d) stylized letter clustering results before and after training.

X_i 为 1 时, 增强对应的 3 个权重 $G_{i,\ j=1,\ 2,\ 3}$, 即增大 $G^+_{i,j}$ 值, 减小 $G^+_{i,j}$ 值. 然后将输出最大的后级神经元标记为 winner, 增强连接在 winner 神经元的突触权重, 即增大 $G^+_{i,j}$ 值, 减小 $G^-_{i,j}$ 值, 其他突触权重不变. 在训练了标准图像后, 训练带有噪声的 Z, V, N 图像, 即有一个像素点反转的图片, 以提高神经网络的容错率. 测试时用一组随机图像做推断, 评估神经网络的性能. 图 18(d) 中展示了训练之前和训练之后, 3 个神经元的响应情况以及权重的分部. 经过 42 个训练周期后, 3 个神经元能准确识别 Z, V, N 图像, 并且输入像素的权重随训练次数的增大, 往正确的预测方向增大/减小.

上述关于用于前向传播的 3D NAND 均采用二值或者幅值编码. 2020 年 Lee 等[55] 设计了一种基于脉宽编码的操作方案, 在 3D-NAND 中实现了前向传播过程, 如图 19 所示. 图 19(a) 展示了前向传播的原理和操作方式: 1) 前级神经元发放的脉冲输入信号采用脉宽编码, 通过脉宽调制电路 (pulse width modulation, PWM) 发放施加在 SSL, 神经网络中每一层突触映射到 3D-NAND 中每一层 CTL 器件上, 并且采用差分对结构用两个 CTL 器件分别存储正、负权重; 2) SL 上施加驱动电压 $V_{\rm BL}$, 选通层的 WL 上加较小的选通电压 $V_{\rm read}$, 未选通层的 WL 上加较大的导通电压 $V_{\rm pass}$, $V_{\rm BL}$ 经过选通层的正、负权重转化为电流进入后级神经元电路; 3) 神经元电路的结构如图 19(b) 所示, 输出的电流经过电流镜构成的差分电路, 相减后对电容进行充电得到后级神经元的电压 $V_{\rm c}$, 其中 SSL 上输入信号的脉宽决定了 SSL 上晶体管的开启时间, 即电流的积分时间. Lee 等[55] 根据 3D-NAND 的硬件方案设计了 3 层全连接的卷积神经网络, 对 CIFAR-10 图片数据库进行识别任务, 并且比较了权重精度为 1 bit 和 4 bit 时网络的性能表现, 如图 19(c) 所示.

2.2.2 3D-NAND 用于反向传播

由于 CTL 器件擦写次数有限, 因此 3D-NAND 通常用于权重更新不太频繁的应用场景. 通常将已训练好的权重映射到 3D-NAND 中, 执行推断任

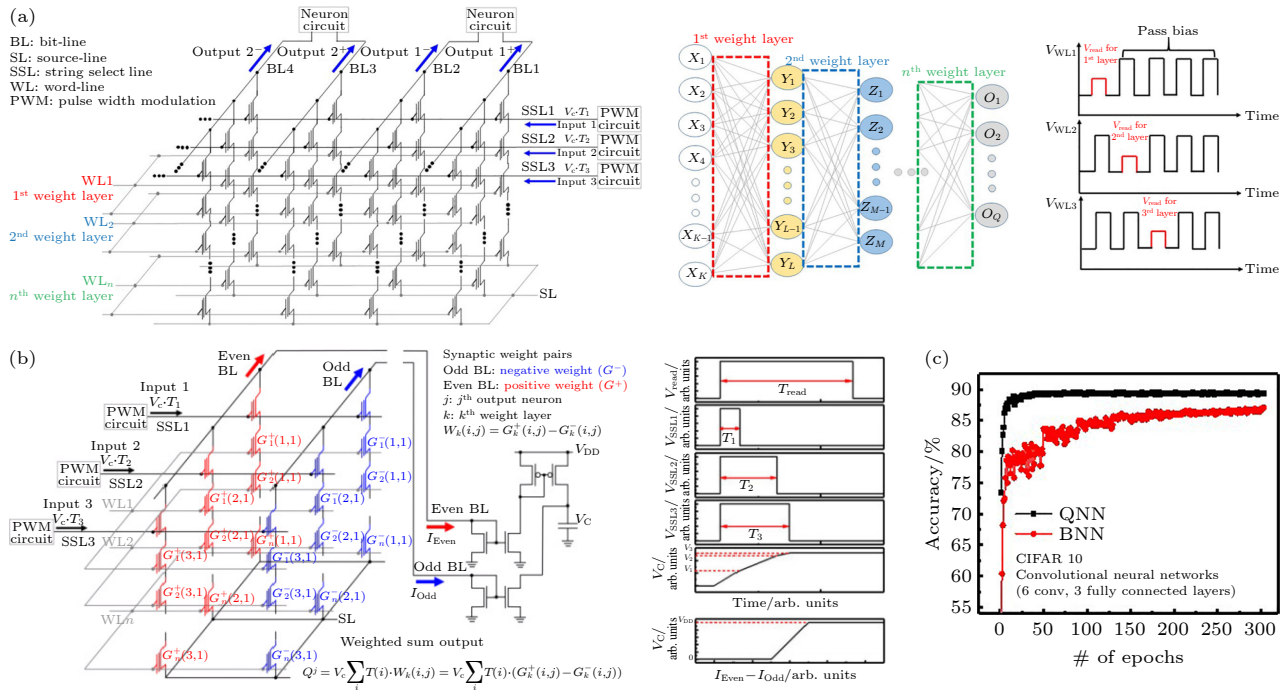


图 19 3D-NAND 中采用脉宽编码实现前向传播[55] (a) 前向传播的操作方式、神经网络示意图和 WL 上读电压的时序图; (b) 左: 前向传播的工作原理示示意图, 即两个标记为正、负权重的突触对构成一个突触, 输出电流相减后通过电容积分转化为电压 $V_{\rm c}$; 右: 读电压 $V_{\rm read}$、SSL 上输入脉冲 $V_{\rm SSL}$ 和 $V_{\rm c}$ 的时序图; (c) 3 层全连接 CNN 网络在权重精度为 4 bit(QNN) 和 1 bit (BNN) 条件下对 CIFAR-10 图片数据库的识别性能

Fig. 19. Forward propagation using 3D-NAND with pulse width modulation (PWM) scheme[55]: (a) Operation scheme of forward propagation, schematic diagram of neural networks, the timing diagram of pulses applied to WLs; (b) Left: schematic diagram of synaptic string array consisting of synapses with positive weight (G^+) and synapses with negative weight (G^-); Right: timing diagram of $V_{\rm read}$, $V_{\rm SSL}$, and $V_{\rm c}$; (c) simulated classification accuracy of 4-bit QNN and BNN for CIFAR-10 images.

务, 2.2.1 节中介绍的工作均基于此应用场景. 但在硬件上直接训练权重, 显然更智能且经济. 除了有限的器件擦写次数, 在 3D-NAND 中做反向传播面临两方面的挑战. 第一, 反向传播中突触矩阵和前向传播中互为转置, 另选硬件资源映射转置后的突触矩阵做反向传播 (如图 10(c) 所示), 训练时间代价较高, 也容易出错. 第二, 突触往往采用差分对结构, NAND 的串行结构天然不适合反向传播. 因此如何设计反向传播的操作方案是关键.

2021 年, Lee 等[56] 在 2020 年报道的方案上[55] 做出改进, 可以在 3D-NAND 同一个 CTL 阵列中同时实现神经形态计算中的前向传播和反向传播过程, 如图 20 所示. 根据神经网络的工作原理, 反向传播过程中误差信号从最后一层神经元输入并依次向前级神经元传递. 对于 crossbar 结构的突触阵列, 反向传播时误差信号从原输出端输入, 从矩阵相乘角度看, 前向传播和反向传播过程中突触阵列的矩阵互为转置的关系, 如图 20(a) 所示. 但如果硬件采用差分对结构, 相邻的器件分别为正、负权重, 前向传播和反向传播时突触的矩阵则为非转置关系, 如图 20(b) 所示. 为了实现反向传播, Lee 等[56] 将所有突触的正、负权重用两个阵列分开配置, 如图 20(c) 所示.

Lee 等[56] 用 3D-NAND 实现前向和反向传播的原理如图 21 所示. 前向传播过程如图 21(a) 所示: 1) 前级神经元的输入信号用脉宽编码, 通过 PWM 模块发生并施加到 BL 上, SSL 控制 NAND string 的开关, 即控制流入到对应的各个后级神经元的电流. 2) 3D-NAND 中每一层突触器件对应神经元中的每一层全连接突触, 如图 21(c) 所示, 信号前向传播到第 i 层时, 在 3D-NAND 第 i 层的 WL 上施加选通电压 $V_{\rm read}$, 在其他层施加导通电压 $V_{\rm pass}$; 3) 所有前级神经元输入的脉冲信号分别经过正、负权重, 转化为电流在输出神经元电路中进行差分后积分转化为输出电压. 反向传播的过程如图 21(b) 所示: 1) 反向传播的误差信号同样采用脉宽编码, 通过 PWM 电路发生并施加在 SSL 上. 由于读操作中 CTL 单元的电流受栅-源电压控制, 即 $V_{\rm read}$ 减去 CTL 单元源端的电位, 而源极的电位与 pass 状态的 CTL 单元的分布情况有关. 如果反向传播过程的误差信号施加在 SL 端, 那么 CTL 的源极电位与前向传播时不一致, 即 CTL 栅源电

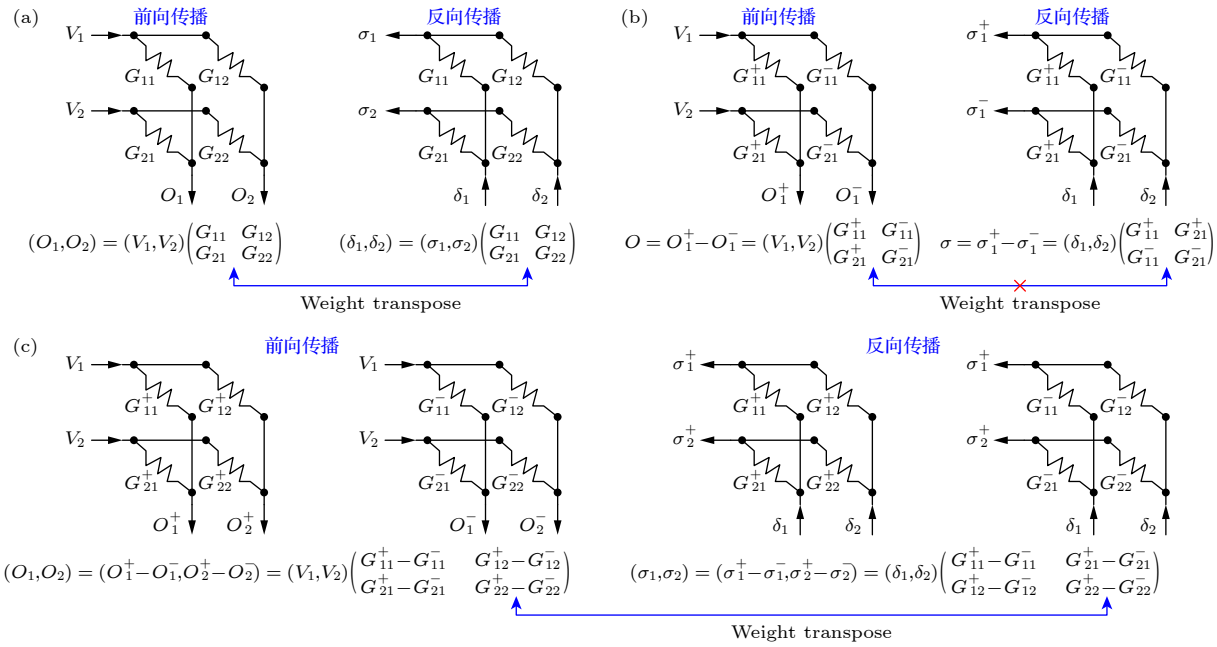


图 20 差分对突触阵列中将正、负权重分开放置可实现反向传播[56] (a) 前向传播和反向传播过程中对应的突触阵列, 从矩阵运算角度上看互为转置结构; (b) 通常情况下, 差分对结构中的正、负权重在同一个阵列中, 突触阵列与前向传播过程中并非转置的关系, 无法实现反向传播功能; (c) 将正、负权重分开置于不同的阵列中, 可以实现反向传播过程

Fig. 20. Backward propagation can be implemented using a differential synaptic array where positive and negative weights are separated[56]: (a) The matrix of synapse weight in forward and backward propagation are transposed; (b) synaptic array architecture consisting of two adjacent cells representing G^+ and G^-, the weights cannot be transposed; (c) synaptic array architecture where G^+ and G^- weights are separated in different arrays.

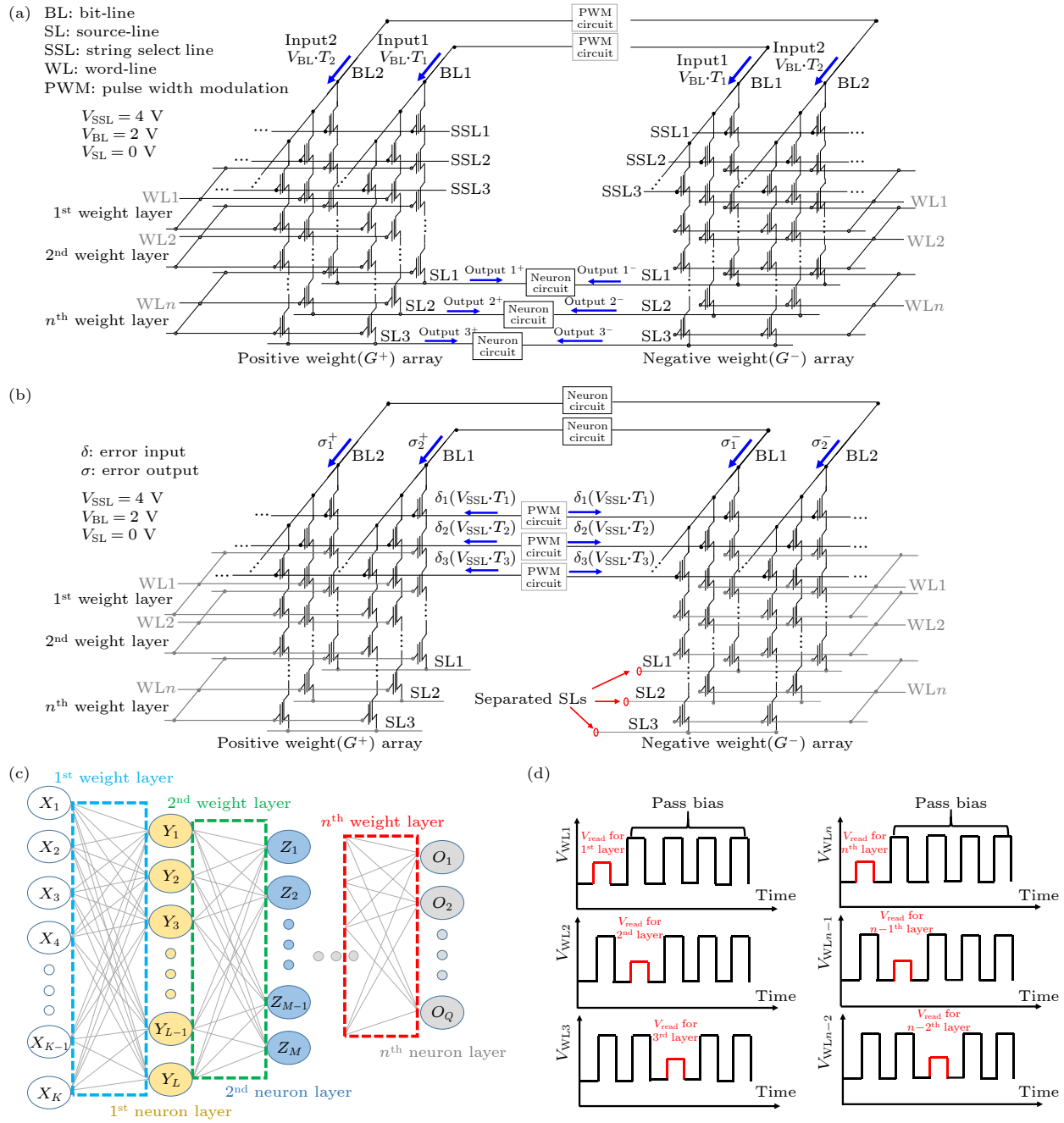


图 21 基于 3D-NAND 的前向传播和反向传播的工作原理[56] (a) 前向传播中 3D-NAND 的操作方法; (b) 反向传播中 3D-NAND 的操作方法; (c) 具有 n 层全连接突触的神经网络结构; (d) 前向传播和反向传播中 WL 上选通电压 V_{read} 和导通电压 V_{pass} 的时序

Fig. 21. Forward and backward propagation using 3D-NAND with PWM scheme[56]: (a) Synaptic array architecture based on NAND flash memory for forwarding propagation operation; (b) synaptic array architecture based on NAND flash memory for backward propagation operation; (c) schematic of neural networks consisting of n weight layers; (d) timing diagram of V_{read} and V_{pass} in forwarding propagation and backward propagation.

压不一致, 读电流将会产生较大的误差. 因此 Lee 等[56]将误差信号从 SSL 端输入, 既能保证权重的转置也能避免 pass 单元造成的读误差. 2) V_{read} 和 V_{pass} 依次施加各层 WL 上, 施加的顺序与前向传播相反. 3) BL 连接前级神经元电路, SL 上施加驱动电压 V_{BL}, 每个 NAND string 上的产生电流在神经元电路中做差分后积分得到输出电压, 通过 SSL 上误差信号的脉宽决定了电流在神经元电路中的积分时间得到对应的电压.

CTL 器件通过写脉冲可具有 32 个不同的 V_{th}, 即 5 bit 的权重态, 如图 22(a) 所示. 图 22(b) 中用不同的读电压 V_1—V_5 可得到不同的权重态分布,

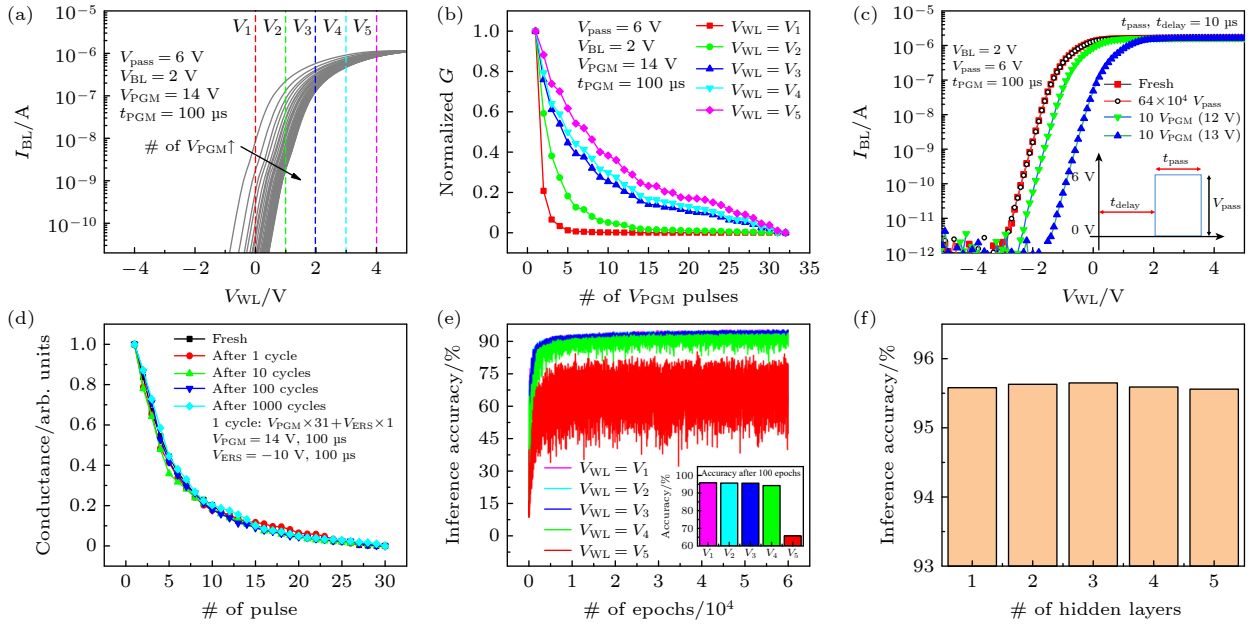


图 22 器件特性和神经网络性能[56] (a) 器件的 I_{BL}-V_{BL} 特性随写入脉冲数量的变化; (b) 图 (a) 中器件的归一化电导随写脉冲数量的变化; (c) 导通电压 V_{pass} 和写电压 V_{PGM} 对阈值电压的影响; (d) 初始状态的器件和经历过擦写循环的器件对写入脉冲的响应; (e) 三层全连接神经网络的识别率; (f) 隐藏层数量对识别率的影响

Fig. 22. Device characteristics and neural network performance[56]: (a) I_{BL}-V_{BL} curves with an increasing number of program pulses; (b) normalized conductance responses measured in (a); (c) I_{BL}-V_{WL} curves measured in a fresh, V_{pass} disturbed, and programmed cell; (d) conductance response of fresh and cycled cell; (e) recognition accuracy of 3-layer neural network; (f) recognition accuracy with the number of hidden layers.

表 2 基于 3 D-NAND 的神经形态计算的各项工作比对

Table 2. A comparison of reviewed works.

	文献工作									
	[45]	[47]	[48]	[49]	[50]	[52]	[53]	[54]	[55]	[56]
技术节点	N.A.	N.A.	N.A.	N.A.	65 nm	N.A.	32 nm	N.A.	26 nm	26 nm
芯片容量	N.A.	N.A.	N.A.	64 GB	N.A.	N.A.	1.13 GB	N.A.	N.A.	N.A.
器件类型	N.A.	SLC	SLC	SLC	SLC	SLC	SLC	Analog	MLC	PLC
输入端口	位线	字线	漏端选通管的字线	位线	漏端选通管的字线	位线	位线	字线	位线	位线
输入编码	脉冲幅值	N.A.	二值编码	数字编码	数字编码	数字编码	数字编码	二值编码	脉宽编码	脉宽编码
输入信号精度	模拟	N.A.	1 bit	4 bit	8 bit	8 bit	8 bit	1 bit	模拟	模拟
突触精度	N.A.	1 bit	1 bit	4 bit	8 bit	8 bit	8 bit	N.A.	4 bit/2 bit	6 bit
反向传播方式	另选转置的突触阵列, 误差信号在位线上输入	N.A.	N.A.	N.A.	N.A.	N.A.	N.A.	N.A	N.A.	误差信号施加在源极选通管位线上
神经网络	3层全连接网络	N.A.	4层全连网络、6层卷积 + 3层全连接	VGG16	LeNet-5	VGG-16	VGG-8	两层全连接网络	6层卷积 + 3层全连接	2—7层全连接网络
神经网络的突触数量	0.3 MB	N.A.	N.A.	138 MB	N.A.	N.A.	110 MB	27	N.A.	N.A.
数据集	MNIST	N.A.	MNIST, CIFAR-10	CIFAR-10	MNIST	N.A.	CIFAR-10	ZVN	CIFAR-10	MNIST
识别率	94.5%	N.A.	98.12%, 87.11%	90%	98.5%	N.A.	N.A.	N.A.	89.38% (4 bit权重)、87.1% (2 bit权重)	95.65%
能效/(TOPS·W^{-1})	N.A.	N.A.	N.A.	~40	N.A.	0.3	12.95	N.A	N.A.	N.A.

其中读电压越大, 权重变化范围越小, 但分布越线性. $V_{\rm pass}$ 对 $V_{\rm th}$ 基本没有影响, 如图 22(c) 所示. 写/擦脉冲循环对权重态的分布也基本无影响, 如图 22(d) 所示. CLT 器件权重的精度为 5 bit, 差分对构成的突触精度达到 6 bit. 基于测试得到的几种权重态分布和单元之间的权重误差分布, Lee 等[56] 设计了全连接神经网络用于 MNIST 图片识别任务, 网络的性能表现如图 22(e) 所示. 其中权重态分布越线性, 网络识别率越高. 将网络中隐藏层数量增加到 3 层, 可以略微提高网络识别率, 如图 22(f) 所示.

2.2.3 各项工作的比对

基于 3D-NAND 的神经形态计算的各项工作比对如表 2 所列.

3 总结与展望

过去几年, 具有存算一体特性的 AI 芯片不断涌现, 工艺节点涵盖了 14—180 nm, 计算架构包括了近存计算、存内计算和神经形态计算, 应用场景覆盖了边缘端到云端设备. 在各种硬件方案中, 基于 3D-NAND 的神经形态芯片在芯片容量, CMOS 工艺兼容性和成本方面极具优势. 本文首先介绍了 3D-NAND 的基本结构和原理, 以及用于神经形态计算的优势和不足. 然后详细梳理了近几年关于 NAND 和 3D-NAND 用于神经形态计算的代表性工作, 重点介绍了其中的编码方式、前向传播原理和反向传播过程.

基于现有的工作, 考虑到 3D-NAND 的优势与不足, 如用于未来的神经形态计算, 3D-NAND 需要做的调整如下:

1) 器件层面. 用于数据存储的 3D NAND, 器件采用电荷俘获型晶体管 (CTL), 通过在栅极施加高幅值和长时程的脉宽 (>10 V, >100 μs), 利用 Fowler-Nordheim 隧穿效应, 在电荷俘获层中注入或擦除电子以改变阈值电压 ($V_{\rm th}$), 实现存储功能. 随擦写次数的增加, 隧穿绝缘层的晶格会被破坏甚至失效, 因此 CTL 的擦写次数有限. 低功耗是神经形态计算的特点, CTL 器件的操作功耗需要进一步优化. 目前国内外的一些研究机构, 探索了将氧化铪基铁电材料替代传统的氮化硅电荷俘获层[57,58], 利用铁电效应实现了器件的存储功能. 如果能将铁电技术成功地应用到 3D NAND 中, 能大幅提高器件的擦写次数, 并且降低操作功耗.

2) 结构层面. 1) CTL 晶体管是 3D NAND 的基本单元, 多个 CTL 器件组成一个 NAND string, 多个 string 组成一个 block, 多个 block 组成 3D NAND 结构. 在神经形态计算中, 突触和神经元是神经网络的基本单元. 2) 突触可由一个或多个 CTL 器件构成. 对于低精度的计算, 可采用幅值或者脉宽编码, 输入/输出均为模拟信号, 单个 CTL 突触即可满足模拟计算的需求, 电路结构简单原理直观. 对于高精度的计算, 则采用二值编码, 用多个 SLC 构建一个多 bit 精度的突触, 采用二进制计算方式. 3) 突触多采用差分对结构 $G=G^{+}-G^{-}$, 为了避免正、负突触阈值电压达到最大而无法进一步更新权重, 3D NAND 中通常需要定期进行块擦除并重新赋予突触权重值. 2021 年, 首尔大学和 SK Hynix 合作开发了适用于神经形态计算的单个 CTL 器件的擦除方案, 避免了定期的块擦除[59].

3) 架构层面. 存储用途的 3D NAND 只涉及读、写、擦操作, 计算由外部的 CPU 负责. 读写按 block→string→CTL 的顺序串行操作. 区别于存储用途, 在用于神经形态计算的 3D NAND 中, 读操作增加了 MAC 运算, 外围电路需要配置大量的 ADC/DAC 和移位加法器等单元. 并且读写操作按神经网络的映射规则执行, 不一定按 block→string→CTL 的顺序.

最后, 由于 3D-NAND 的专利特性, 厂商并未开放用户对芯片颗粒端口的权限. 目前的工作中, 前向传播过程和反向传播过程并未做硬件实现, 多数是基于厂商样片测得的存储单元特性以及读误差分布, 通过电路和软件层面上仿真得到的结果. 未来的工作应该考虑与厂商有更深入的交流合作, 在硬件层面执行前向传播、反向传播和权重更新, 更直接地展示 3D-NAND 在神经形态计算方面的应用潜力.

参考文献

[1] Amodei D, Hernandez D, Sastry G, Clark J, Brockman G, Sutskever I https://openai.com/blog/ai-and-compute/ [2022-4-11]

[2] Patterson D A, Hennessy J L 2021 *Computer Organization and Design RISC-V Edition: the Hardware Software Interface* (6th Ed.) (Amsterdam: Morgan Kaufmann) p44

[3] Gai S https://pensando.io/dennard-scaling-and-other-power-considerations/ [2022-4-11]

[4] Dally B https://www.cs.colostate.edu/~cs575dl/Sp2015/Lec

tures/Dally2015.pdf [2022-4-11]
[5] Drachman D A 2005 *Neurology* **64** 2004
[6] Zhang W, Gao B, Tang J, Yao P, Yu S, Chang M F, Yoo H J, Qian H, Wu H 2020 *Nat. Electron.* **3** 371
[7] Roy K, Jaiswal A, Panda P 2019 *Nature* **575** 607
[8] Khaddam-Aljameh R, Stanisavljevic M, Mas J F, et al. 2021 *2021 Symposium on VLSI Technology* Kyoto, Japan, June 13–19, 2021 p1
[9] Narayanan P, Ambrogio S, Okazaki A, et al. 2021 *2021 Symposium on VLSI Technology* Kyoto, Japan, June 13–19, 2021 p1
[10] Yang J, Xue X, Xu X, Lv H, Zhang F, Zeng X, Chang M F, Liu M 2020 *2020 IEEE Symposium on VLSI Circuits* Honolulu, HI, USA, June 16–19, 2020 p1
[11] Chih Y D, Shih Y C, Lee C F, et al.2020 *2020 IEEE International Solid-State Circuits Conference* San Francisco, CA, USA, Feb 16–20, 2020 p222
[12] Liu Y, Su F, Yang Y, Wang Z, Wang Y, Li Z, Li X, Yoshimura R, Naiki T, Tsuwa T, Saito T, Wang Z, Taniuchi K, Yang H 2019 *IEEE J. Solid-State Circuits* **54** 885
[13] Dünkel S, Trentzsch M, Richter R, et al. 2017 *2017 IEEE International Electron Devices Meeting* San Francisco, California, USA, Dec 2–6, 2017, p19.17. 11
[14] IRDS™ 2021 *International Roadmap for Devices and Systems* (2021 Ed.) (IEEE) from https://irds.ieee.org/images/files/pdf/2021/2021IRDS_MM_Tables.xlsx [2022-4-11]
[15] Liu T Y, Yan T H, Scheuerlein R, et al. 2013 *2013 IEEE International Solid-State Circuits Conference* San Francisco, CA, USA, Feb. 17–21, 2013 p210
[16] Chen Y, Li H, Wang X, Zhu W, Xu W, Zhang T 2012 *IEEE J. Solid-State Circuits* **47** 560
[17] Rho K, Tsuchida K, Kim D, et al. 2017 *2017 IEEE International Solid-State Circuits Conference* San Franciso, CA, USA, Feb. 5–9, 2017 p396
[18] Zwerg M, Baumann A, Kuhn R, et al. 2011 *2011 IEEE International Solid-State Circuits Conference* San Francisco, CA, USA, Feb. 20–24, 2011 p334
[19] Takashima D, Nagadomi Y, Ozaki T 2011 *IEEE J. Solid-State Circuits* **46** 681
[20] Trentzsch M, Flachowsky S, Richter R, et al. 2016 *2016 IEEE International Electron Devices Meeting* San Francisco, CA, USA, December 3–7, 2016 p11.15.11
[21] Lee J W, Na D, Kavala A, et al. 2020 *2020 IEEE Symposium on VLSI Circuits* Honolulu, HI, United States, June 16–19, 2020 p1
[22] Mulaosmanovic H, Breyer E T, Dünkel S, Beyer S, Mikolajick T, Slesazeck S 2021 *Nanotechnology* **32** 502002
[23] Noguchi H, Ikegami K, Kushida K, et al. 2015 *2015 IEEE International Solid-State Circuits Conference* San Franciso, CA, USA, February 22–26, 2015 p1
[24] Sato H, Honjo H, Watanabe T, et al. 2018 *2018 IEEE International Electron Devices Meeting* San Francisco, CA, USA, December 1–5, 2018 p27.22.21
[25] Khan A I, Keshavarzi A, Datta S 2020 *Nat. Electron.* **3** 588
[26] Dong Q, Kim Y, Lee I, et al. 2017 *2017 IEEE International Solid-State Circuits Conference* San Francisco, CA, USA, February 5–9 2017 p198
[27] Cheong W, Yoon C, Woo S, et al. 2018 *2018 IEEE International Solid-State Circuits Conference* San Francisco, California, USA, February 11–15, 2018 p338
[28] Liang J, Jeyasingh R G D, Chen H, Wong H P 2011 *2011 Symposium on VLSI Technology* Kyoto, Japan, June 14–16, 2011 p100
[29] Wu T F, Le B Q, Radway R, et al. 2019 *2019 IEEE International Solid-State Circuits Conference* San Francisco, CA, USA, February 17–21, 2019 p226
[30] Dong Q, Wang Z, Lim J, Zhang Y, Sinangil M E, Shih Y C, Chih Y D, Chang J, Blaauw D, Sylvester D 2019 *IEEE J. Solid-State Circuits* **54** 231
[31] Yoon S, Youn Y, Kim S 2015 *2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems* New York, NY, USA, August 24–26, 2015 p1045
[32] Sandre G D, Bettini L, Pirola A, et al. 2010 *2010 IEEE International Solid-State Circuits Conference* San Francisco, California, USA, February 7–11, 2010 p268
[33] Lo C P, Lin W Z, Lin W Y, Lin H T, Yang T H, Chiang Y N, King Y C, Lin C J, Chih Y D, Chang T Y J, Chang M F 2019 *IEEE J. Solid-State Circuits* **54** 584
[34] Xue X Y, Jian W X, Yang J G, Xiao F J, Chen G, Xu X L, Xie Y F, Lin Y Y, Huang R, Zhou Q T, Wu J G 2012 *2012 Symposium on VLSI Circuits* Honolulu, HI, USA, June 13–15, 2012 p42
[35] Chang M, Shen S, Liu C, Wu C, Lin Y, King Y, Lin C, Liao H, Chih Y, Yamauchi H 2013 *IEEE J. Solid-State Circuits* **48** 864
[36] Chen J, Chiang R C, Huang H H, Venkataramani G 2012 *SIGOPS Oper. Syst. Rev.* **45** 48
[37] Qazi M, Clinton M, Bartling S, Chandrakasan A P 2012 *IEEE J. Solid-State Circuits* **47** 141
[38] Breyer E T, Mulaosmanovic H, Trommer J, Melde T, Dünkel S, Trentzsch M, Beyer S, Slesazeck S, Mikolajick T 2020 *IEEE J. Electron Devices Soc.* **8** 748
[39] Wang Z, Wu H, Burr G W, Hwang C S, Wang K L, Xia Q, Yang J J 2020 *Nat. Rev. Mater.* **5** 173
[40] Lee G H, Hwang S, Yu J, Kim H 2021 *Appl. Sci.* **11** 6703
[41] Jang J, Kim H S, Cho W, et al.2009 *2009 Symposium on VLSI Technology* Kyoto, Japan, June 15–17, 2009 p192
[42] Wonjoo K, Sangmoo C, Junghun S, Taehee L, Park C, Hyoungsoo K, Juhwan J, Inkyong Y, Park Y 2009 *2009 Symposium on VLSI Technology* Kyoto, Japan, June 15–17, 2009 p188
[43] Micheloni R 2016 *3D Flash Memories* (Dordrecht: Springer Netherlands) p89
[44] Seo Y T, Kwon D, Noh Y, Lee S, Park M K, Woo S Y, Park B G, Lee J H 2021 *IEEE Trans. Electron Devices* **68** 3801
[45] Lee S T, Lim S, Choi N, Bae J H, Kim C H, Lee S, Lee D H, Lee T, Chung S, Park B G, Lee J H 2008 *2018 IEEE Symposium on VLSI Technology* Honolulu, HI, USA, June 18–22, 2018 p169
[46] Lee S T, Lim S, Choi N Y, Bae J H, Kwon D, Park B G, Lee J H 2019 *IEEE J. Electron Devices Soc.* **7** 1085
[47] Wang P, Xu F, Wang B, Gao B, Wu H, Qian H, Yu S 2019 *IEEE Trans. Very Large Scale Integr. VLSI Syst.* **27** 988
[48] Lee S T, Kim H, Bae J H, Yoo H, Choi N Y, Kwon D, Lim S, Park B G, Lee J H 2019 *2019 IEEE International Electron Devices Meeting* San Francisco, CA, USA, December 7–11, 2019 p38.34.31
[49] Lue H T, Hsu P K, Wei M L, Yeh T H, Du P Y, Chen W C, Wang K C, Lu C Y 2019 *2019 IEEE International Electron Devices Meeting* San Francisco, CA, USA, December 7–11, 2019 p38.31.31
[50] Kim M, Liu M, Everson L, Park G, Jeon Y, Kim S, Lee S, Song S, Kim C H 2019 *2019 IEEE International Electron*

Devices Meeting San Francisco, CA, USA, December 7–11, 2019 p38.33.31
[51] Kim M, Liu M, Everson L R, Kim C H 2022 *IEEE J. Solid-State Circuits* **57** 625
[52] Kang M, Kim H, Shin H, Sim J, Kim K, Kim L S 2022 *IEEE Trans. Comput.* **71** 1291
[53] Hsu P K, Du P Y, Lo C R, Lue H T, Chen W C, Hsu T H, Yeh T H, Hsieh C C, Wei M L, Wang K C, Lu C Y 2020 *2020 IEEE International Memory Workshop* Dresden, Germany, May 17–20, 2020 p1
[54] Zhou W, Jin L, Jia X, Wang T, Xu P, Zhang A, Huo Z 2022 *IEEE Electron Device Lett.* **43** 374
[55] Lee S T, Lee J H 2020 *Front. Neurosci.* **14** 517292
[56] Lee S T, Yeom G, Yoo H, Kim H S, Lim S, Bae J H, Park B G, Lee J H 2021 *IEEE Trans. Electron Devices* **68** 3365
[57] Kim M K, Kim I J, Lee J S 2021 *Sci. Adv.* **7** 1341
[58] Yoon S, Hong S I, Choi G, Kim D, Kim I, Jeon S M, Kim C, Min K 2022 *2022 IEEE International Memory Workshop* Dresden, Germany, May 15–18, 2022 p 1
[59] Yoo H N, Back J W, Kim N H, Kwon D, Park B G, Lee J H 2022 *2022 IEEE Symposium on VLSI Technology and Circuits* Honolulu, HI, USA, June 12–17, 2022 p304

SPECIAL TOPIC—Physical electronics for brain-inspired computing

3D-NAND flash memory based neuromorphic computing

Chen Yang-Yang [1)2)3)] He Yu-Hui [3)4)]
Miao Xiang-Shui [3)4)] Yang Dao-Hong [1)2)3)†]

1) (*Post-doctoral Mobile Station, Huazhong University of Science and Technology, Wuhan* 430074, *China*)
2) (*Post-doctoral Work Station, Wuhan Xinxin Semiconductor Manufacturing Co., Ltd., Wuhan* 430205, *China*)
3) (*Hubei Yangtze Memory Laboratories, Wuhan* 430205, *China*)
4) (*School of Integrated Circuit, Huazhong University of Science and Technology, Wuhan* 430074, *China*)

(Received 16 May 2022; revised manuscript received 27 September 2022)

Abstract

A neuromorphic chip is an emerging AI chip. The neuromorphic chip is based on non-Von Neumann architecture, and it simulates the structure and working principle of the human brain. Compared with non-Von Neumann architecture AI chips, the neuromorphic chips have significant improvement of efficiency and energy consumption advantages. The 3D-NAND flash memory has the merits of a mature process and ultra-high storage density, and recently it attracted many researchers' attention. However, owing to the proprietary nature of the technology, there are few hardware implementations. This paper reviews the present research status of neuromorphic computing by using the 3D-NAND flash memory, introduces the forward propagation and backward propagation schemes, and proposes several improvements on the device, structure, and architecture of 3D NAND for neuromorphic computing.

Keywords: neuromorphic computing, 3D-NAND, in-memory computing architecture

PACS: 07.05.Mh, 85.35.–p, 84.30.–r, 87.18.Sn **DOI:** 10.7498/aps.71.20220974

† Corresponding author. E-mail: alan_yang@xmcwh.com

物理学报 Acta Physica Sinica
Chinese Physical Society
Institute of Physics, CAS

基于3D-NAND的神经形态计算

陈阳洋 何毓辉 缪向水 杨道虹

3D-NAND flash memory based neuromorphic computing

Chen Yang-Yang He Yu-Hui Miao Xiang-Shui Yang Dao-Hong

引用信息 Citation: Acta Physica Sinica, 71, 210702 (2022) DOI: 10.7498/aps.71.20220974
在线阅读 View online: https://doi.org/10.7498/aps.71.20220974
当期内容 View table of contents: http://wulixb.iphy.ac.cn
