专题: 半导体物理与器件

面向高精度 3 维 NAND 存算一体芯片的多晶硅晶界势垒工艺优化与验证

郑好[1)2)] 刘慧雯[3)] 许克志[3)] 张宝通[3)] 杨远程[3)]
夏志良[3)†] 霍宗亮[1)3)‡]

1) (中国科学院微电子研究所, 存储器实验室, 北京 100029)

2) (中国科学院大学, 北京 100049)

3) (长江存储科技有限责任公司, 武汉 430070)

(2026 年 2 月 4 日收到; 2026 年 3 月 10 日收到修改稿)

随着人工智能与边缘计算的快速发展, 存算一体 (computing-in-memory, CIM) 架构被认为是缓解冯·诺依曼瓶颈的重要技术路径. 基于 3 维 (3D) NAND 的 CIM 方案兼具高存储密度与工艺成熟度, 但在执行矩阵-向量乘法 (matrix-vector multiplication, MVM) 等模拟计算任务时, 串电流分布展宽会导致累加电流偏差, 进而引起计算精度下降问题. 其中 3D NAND 顶部选择 (top select gate, TSG) 晶体管沟道的多晶硅晶界更是会直接影响到串电流的分布. 因此本文采用计算机辅助设计软件 (technology computer-aided design, TCAD) 建立 TSG Deck 器件模型, 分析 TSG 多晶硅沟道晶界陷阱诱发的势垒对开态电流波动的影响规律. 在此基础上, 提出一种通过多晶硅前驱体组合调控实现等效氢钝化窗口优化的工艺方案, 并对不同工艺分组条件下的开态电流分布进行单片晶圆尺度的统计评估. 结果表明, 最优工艺可使位线端电流分布的归一化标准差较优化前减小 50%, 并在 CIM 系统级仿真中使 GPT-2 124M 模型的 INT8 推理中的 MVM 计算误差相对基准工艺降低 14.7%—66.8%. 综上, 本工作为面向高精度 3D NAND CIM 芯片的工艺优化方案提供了可实现的设计依据.

关键词: 3 维 NAND, 存算一体, 顶部选择管, 多晶硅晶界, 电流分布

DOI: 10.7498/aps.75.20260199 **CSTR:** 32037.14.aps.75.20260199

1 引 言

传统冯·诺依曼体系结构由于处理器与存储器物理分离, 导致计算过程中存在频繁的数据搬运, 从而引发带宽受限与能耗开销问题, 成为限制人工智能推理性能提升的关键瓶颈. 存算一体 (computing-in-memory, CIM) 通过在存储阵列内部直接完成乘加/累加等运算, 显著降低访存与数据搬运开销, 被认为是突破“存储墙”的有效途径[1–3].

在多种 CIM 器件实现方案中, 基于 3D NAND 的 CIM 架构依托其超高集成密度、成熟制造平台与非易失性等优势, 具备规模化落地潜力[4,5]. 其基本思想是: 将神经网络权重映射为存储单元阈值电压, 通过在位线 (bit line, BL)、字线 (word line, WL) 或顶部选择管 (top select gate, TSG) 施加输入电压, 使多个 NAND 串产生的电流在位线端自然累加, 从而完成矩阵-向量乘法. 然而, 在大规模并行场景下数千条 NAND 串同时工作, 工艺涨落 (process variation, PV) 引起的器件参数离散, 如阈值电压、接触电阻与栅耦合比等, 会导致串电流

† 通信作者. E-mail: albert_xia@ymtc.com

‡ 通信作者. E-mail: zongliang_huo@ymtc.com

分布展宽, 使累加电流偏离理想线性关系, 最终表现为模型推理精度下降[6–10].

从电流传输路径来看, TSG 位于 3D NAND 串的电流调控上游, 其导通电阻与阈值电压的离散会传递到位线端开态电流, 因此直接决定电流分布的宽窄. 以往的研究与工艺优化多集中于存储阵列沟道本身[11], 但阵列沟道通常膜厚极薄且沟道界面态占比高, 可用于抑制波动的工艺自由度极其有限、调控难度更大. 因此, 本文的核心差异化贡献在于跳出传统阵列沟道的局限, 创新性地将优化焦点转移至对整体串电流分布起决定性作用的 TSG 多晶硅沟道. 通过前驱体组合工艺调控 TSG 晶界势垒, 不仅工程可行性更高, 且对最终 CIM 计算精度的提升效果显著.

而在 TSG 采用多晶硅沟道时, 晶界处往往富集大量缺陷陷阱, 这些陷阱通过俘获载流子引入局部空间电荷, 在晶界处形成势垒, 从而降低载流子跨晶界传输概率并削弱有效迁移率, 最终表现为开态电流的随机涨落与分布展宽[12,13]. 工业上通常采用氢钝化降低晶界缺陷的电活性, 但该过程对温度、沉积速率、气体配比以及等离子体功率等参数高度敏感[14,15]. 同时, 单纯提高含氢量还可能引发弱键增多与氢聚集等副作用, 使工艺窗口难以通过直接“加大剂量”的方式获得稳定且最优的电流分布[16,17]. 基于上述背景, 本文提出一种面向 3D NAND 的 TSG 多晶硅沟道工艺组合调控方案, 通过在 TSG Deck 架构下独立调整 TSG 多晶硅前驱体沉积与堆叠组合, 间接实现晶界陷阱态与等效氢钝化强度的可控优化.

综上, 本文首先利用计算机辅助设计软件 (technology computer-aided design, TCAD) 仿真建立基于 TSG Deck 架构的多晶硅沟道中晶界势垒与开态电流的对应关系, 随后对不同前驱体组合工艺进行电流分布的统计分析, 得到最优电流分布, 最后将提取的分布参数导入既有 CIM 仿真框架[18], 评估其对计算误差的系统级影响.

2 器件结构和仿真设置

本文采用 Synopsys Sentaurus TCAD 建立 3D NAND TSG Deck 器件模型. 需要特别说明的是, 该 TSG Deck 架构并非为本次研究而对基础结构进行的临时改动, 而是经过大规模产线流片与电性验证的成熟量产架构. 在实际制造中, 该架构完全兼容现有的 3D NAND 标准工艺流, 不仅不会对器件制备产生任何不利影响, 反而能有效提升器件的整体可靠性. 与传统结构不同, 在新 TSG Deck 架构中, TSG 沟道与阵列沟道通过 Plug 结构连接, 因此可将阵列沟道等效为串联导通电阻, 并将优化焦点聚焦于 TSG 沟道本体.

图 1(a) 为完整 TSG Deck 结构示意图, 图 1(b) 为纵剖面. 最下方 Plug 为阵列沟道的重掺杂漏端, 同时也是 TSG 的源端. Plug 上方为 TSG 沟道, 采用 B, P, As 顺序掺杂以确保 TSG 多晶硅沟道反型以及 TSG 漏端欧姆接触. 沟道外侧为氧化层与多晶硅栅, 顶层为后续工艺氮化物阻挡层. 为保证掺杂分布准确, 本文先在 SProcess(Sentaurus process) 中生成掺杂分布并导入 SDE (Sentaurus device editor). 随后在 SDE 中引入基于三维维诺图算法[19]生成的晶界, 用以形成 TSG 多晶硅沟道结构并构建晶界陷阱模型. 为了保证后续晶界势垒仿真产生的电流量级涨落和电流分布的变化趋势的数值具有可信性和参考性, 模型的结构参数主要来自晶圆实测; 而晶界模型, 迁移率模型和复合模型的选择则参考 Zou 等[13,20] 和 Yang 等[21] 的工作; 电学参数如接触电阻/串联电阻同样来自晶圆实测, 具体的关键几何与陷阱态参数如表 1 所示.

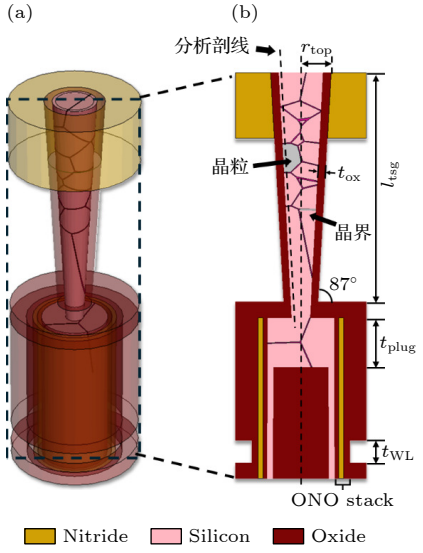

图 1 3D NAND TSG Deck 结构示意图 (a) 完整 TSG Deck 结构; (b) 完整 TSG Deck 结构纵向切面图

Fig. 1. Schematic illustration of the 3D NAND TSG Deck structure: (a) Complete TSG Deck structure; (b) longitudinal cross-sectional view of the complete TSG Deck structure.

表 1 TCAD 仿真参数表

Table 1. Parameter table of TCAD.

参数名	值	参数名	值
沟道顶部半径 r_{top} /nm	40	晶界尾态类受主陷阱浓度 (峰值在导带处)/($\text{cm}^{-3}\cdot\text{eV}^{-1}$)	$1,2,3\times10^{21}$
沟道倾角 θ /(°)	87	晶界深能级类受主陷阱浓度 (峰值在禁带中央上0.1 eV)/($\text{cm}^{-3}\cdot\text{eV}^{-1}$)	1×10^{19}
沟道氧化层厚度 t_{ox} /nm	10	晶界尾态类受主陷阱分布σ/eV	0.05
Plug厚度 t_{plug} /nm	60	晶界深能级类受主陷阱分布σ/eV	0.1
TSG栅长 l_{tsg} /nm	200	Si/O_x界面尾态类受主陷阱浓度 (峰值在导带下0.1 eV)/($\text{cm}^{-2}\cdot\text{eV}^{-1}$)	2×10^{13}
Punch深度 t_{punch} /nm	20	Si/O_x界面尾态 类受主陷阱分布σ/eV	0.035
Plug掺杂 $N_{\text{D}}^{\text{Plug}}$ / cm^{-3}	4×10^{17}	晶粒尺寸/nm	50

图 2 给出了 $V_{\text{d}}=1.5$ V 偏置条件下的 I_{d}-V_{g} 仿真曲线与实验统计趋势对比, 其中实验数据是在单片晶圆上测试的 100 个芯片的开态电流统计分布. 可见, 在亚阈值与强反型区, 仿真曲线与实验曲线大致相同, 说明所构建的结构与陷阱模型能够保证后续仿真中开态电流变化的量级的正确性.

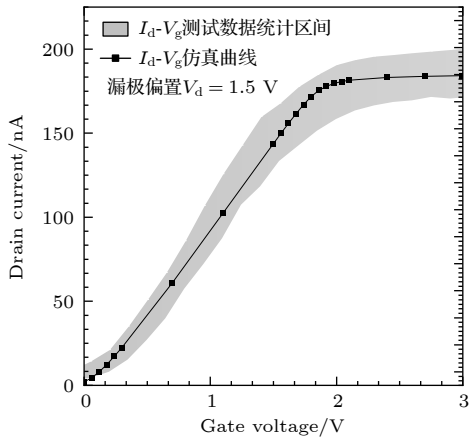

图 2 $V_{\text{d}}=1.5$ V 时 I_{d}-V_{g} 转移特性测试数据与仿真结果对比

Fig. 2. Comparison between the measured I_{d}-V_{g} transfer characteristics and the simulation results at $V_{\text{d}}=1.5$ V.

为进一步指导工艺优化方向, 本文在多晶硅晶界区域引入不同密度的浅能级与深能级类受主陷阱, 用以分析陷阱浓度对开态电流分布的影响, 具体参数如表 1 所示, 其中陷阱密度范围由 Zou 等[13,20] 和 Yang 等[21] 的工作结合工程经验得到. 晶界陷阱对载流子的俘获将改变局部电荷中性条件并引入空间电荷积累, 从而在晶界附近形成额外的静电势起伏, 并表现为导带能级的局部上抬, 即晶界势垒. 该势垒提高了载流子跨晶界传输的能量门槛, 降低跨晶界输运概率与等效迁移率, 使沟道等效电阻增大. 同时, 由于晶界缺陷态具有显著的随机性与工艺敏感性, 上述效应会进一步放大器件间差异, 最终在统计意义上体现为开态电流分布的展宽.

为了定量描述这一过程, Seto[22] 建立了经典的多晶硅载流子输运模型用于描述晶界势垒高度 E_{b} 、面陷阱密度 N_{t} 与有效迁移率 μ_{eff} 之间的解析关系. 根据 Seto 模型, 多晶硅晶界处因陷阱俘获载流子而形成的耗尽区势垒高度 E_{b} 可表示为 $E_{\text{b}}=qN_{\text{t}}^2/(8\varepsilon_{\text{Si}}N)$ (晶粒部分耗尽), 其中, q 为元电荷量, ε_{Si} 为硅的介电常数, N 为沟道内的有效载流子浓度. 该势垒的存在使得载流子跨越晶界的输运主要依赖于热发射机制和隧穿机制, 从而导致多晶硅的有效迁移率 μ_{eff} 受势垒高度和温度影响[23], 且呈指数级变化, 即 $\mu_{\text{eff}}=\mu_0\exp\left[-E_{\text{b}}/(kT)\right]$. 其中 μ_0 为晶粒体内的本征迁移率, k 为玻尔兹曼常数, T 为绝对温度.

将所提取到的 TSG 沟道附近的面陷阱密度 3×10^{12} cm^{-2} 和有效载流子浓度 1.2×10^{18} cm^{-3} 代入公式计算理论势能高度, 结果为 0.048 eV , 与图 3(a) 中 TCAD 仿真得到的 0.05 eV 的势垒高度基本一致.

考虑到开态导通时电流主要沿氧化层/沟道界面附近形成的反型层传输, 本文在距氧化层界面向沟道内偏移 1 nm 处分析了不同晶界陷阱密度的导带能量分布和电流密度分布情况, 剖面位置如图 1(b) 所示. 结果如图 3(a) 所示, 在晶界位置导带能量出现明显抬升, 同一位置的电流密度呈现约 1 个数量级的突变. 这是由于在热发射/隧穿共同主导的跨晶界输运条件下, 势垒高度的微小变化即可引起跨晶界电导的显著变化. 进一步地, 在不

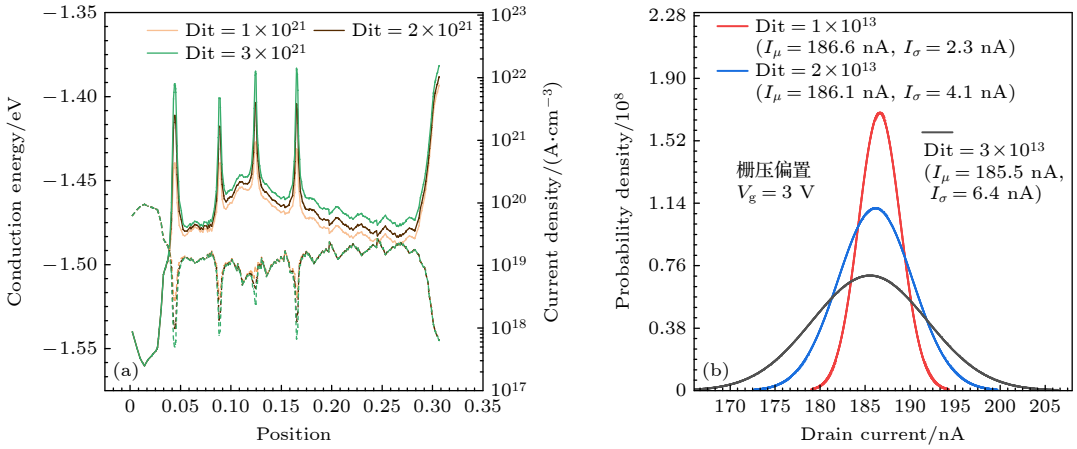

图 3 TCAD 仿真结果 (a) 不同晶界陷阱密度条件下, 氧化层/沟道界面处向沟道偏移 1 nm 位置的导带能量和电流密度图, 其中实线为导带能量, 虚线为电流密度; (b) 不同晶界陷阱密度条件下, 电流分布的变化情况

Fig. 3. TCAD simulation results: (a) Conduction band energy and current density profiles at a position 1 nm from the oxide/channel interface into the channel under different grain boundary trap densities, the solid lines represent conduction band energy, and the dashed lines represent current density; (b) variation of the current distribution under different grain-boundary trap densities.

同晶界陷阱密度条件下, 如图 3(b) 所示, 可以观察到在 $V_{\rm g}=3$ V 的强反型条件下, 开态电流分布宽度随晶界陷阱密度增加呈上升的趋势. 综上, 导带抬升与电流密度降低在空间位置上的一致性表明通过调控晶界陷阱密度, 从而调控势垒高度及其波动可有效收敛 TSG 开态电流的统计分布.

3 工艺调控和结果分析

基于上述机理, 本文采用 3 种 TSG 多晶硅沟道前驱体 (以 NS, MS, DS 代称) 及其组合进行工艺调控, 3 种前驱体基本定性特性如表 2 所示.

表 2 不同前驱体特性对照表

Table 2. Comparison table of different precursor characteristics.

前驱体	分解	含氢量	粗糙度	成本
NS	—	—	好	低
MS	4H	多	差	低
DS	3H	少	较好	高

不同前驱体在成核行为、晶粒尺寸与晶界形貌、以及氢相关活性基团引入与保留能力等方面存在差异, 从而改变晶界缺陷的产生与可钝化比例, 最终体现为晶界陷阱态密度、势垒分布及其统计波动变化.

具体而言, NS 前驱体具有极佳的表面形貌控制能力, 能够形成平整的氧化层/多晶硅界面. 然而, 其分子结构导致沉积过程中的氢保留率极低, 难以对悬挂键形成有效钝化; MS 前驱体作为低成本的主体填充材料, 其分子中含有较多氢原子, 理论上能提供富氢环境. 但在实际工艺中, MS 的成核过程较为剧烈, 极易导致界面粗糙度恶化, 引入额外的表面散射中心, 从而降低载流子迁移率; DS 前驱体分解活化能较低, 成膜质量与表面平整度优于 MS, 但成本最高, 大量沉积不存在可行性.

传统的单一前驱体工艺面临鱼与熊掌不可兼得的困境. 从表面化学反应动力学角度来看, 硅源前驱体在反应腔室内的热分解与吸附过程主导了薄膜的最终微观形貌. NS 及其类似的高阶硅烷前驱体分子量较大, 在气相中分解所需活化能较高. 当其到达氧化硅表面时, 倾向于形成极其均匀的成核点, 从而沉积出均方根粗糙度极低的多晶硅基层. 然而, NS 分子在完全分解成膜后, 副产物迅速解吸附, 几乎不会在晶格内部残留游离氢. 这意味着虽然界面平整, 但后续生长的晶粒边界处悬挂键处于完全暴露状态. 若单用 MS, 虽然 MS 沉积速率快、成本低, 且在 500—600 ℃ 的典型低气压化学气相沉积 (low pressure chemical vapor deposition, LPCVD) 窗口下分解反应异常剧烈, 但这种剧烈的反应导致晶粒呈随机的三维岛状生长, 不同取向的晶粒在挤压碰撞时会产生大量不规则的空隙与高应力晶界. 尽管 MS 体系中氢气浓度高, 但由于快速的薄膜堆叠, 氢原子难以充分扩散至晶界深处与悬挂键形成稳定的 Si—H 键.

因此, 本文提出引入 DS 中间层的方案. 考虑

到 DS 的热分解温度显著低于 MS, 这意味着在相同的工艺温度下, DS 在硅表面的黏附概率和表面迁移率更高[24], 更倾向于层状生长. 更重要的是, DS 分解过程中会产生大量高活性的含氢自由基. 当我们在 NS 成核层之上、MS 主体填充之前插入 DS 层时, 这层缓冲带不仅继承了 NS 层的平整度, 其释放的活性氢基团能够在 MS 高速填充引发的热应力作用下, 通过晶界间隙向外扩散, 与多晶硅晶界以及 Si/SiO_2 界面的缺陷进行原位化学结合. 这种原位钝化可以避免由于后续器件制造中多次热循环导致的氢逸出效应, 使得最终固化在晶界处的 Si—H 键浓度最大化, 从而有效填平了禁带中央的深能级陷阱.

为此, 如图 4 所示, 本文设计了一套“三明治”式的沉积方案进行验证. 实验在标准 3D NAND 产线上进行, 保持温度、压力等宏观参数一致, 通过改变沉积配方设置了 4 个对照组.

基准组 (Group A, NS+MS): 采用传统工艺, 底层沉积薄层 NS 以确立良好的界面形貌, 随后直接使用 MS 进行主体填充. 该方案虽然保证了物理界面的完整性, 但缺乏针对近界面处晶界深能级陷阱的有效钝化手段.

优化组 (Group B, NS+9 nm DS+MS): 在 NS 成核层与 MS 填充层之间, 创新性地插入了 9 nm 厚的 DS 中间层. 这一设计巧妙利用了 DS 前驱体的特性, 在载流子输运最密集的沟道浅表层构建了一个“低缺陷、高有序”的缓冲带, 实现了原位氢钝化与低界面散射的平衡.

扩展组 (Group C/D): 进一步尝试在底部 Plug 接触区引入 DS(Group C) 或增大 DS 层厚度至 18 nm(Group D).

在相同测试偏置条件下, 本文对单片晶圆内 400—500 个 Die 的 TSG 开态电流进行统计采样, 并使用电流均值和归一化标准差指标对电流分布的集中程度与形态进行表征.

统计结果如图 5 所示, 实验数据的统计拟合采用了与后续仿真相同的偏态 t 分布, 以确保两者具备可比性, 其中 d 表示自由度, 控制分布尾部的厚度, d 越小, 表示离群值越多; l 表示位置, 即分布在 x 轴上的中心位置, 不代表电流均值; v 表示方差, 控制分布的离散程度, v 越小代表分布越集中, 反之越离散; s 表示偏度, 偏度越负表示数据越向左偏, 反之越向右偏. 从统计结果看, 基准工艺 A 组的中心位置电流最低、归一化标准差最高, 表明该工艺下晶界势垒网络及其随机涨落更为显著. B 组在电流均值与离散度两方面同时表现最优: 中心位置电流最高、归一化标准差最低, 说明该前驱体组合能够在保证导通能力的同时有效抑制器件间波动. C 与 D 组相较 A 组有一定改善, 但整体仍劣于 B 组. 结合工艺过程分析推测, Plug 区域额外引入 DS 可能改变局部掺杂活化/扩散轮廓或界面反应, 从而影响串联电阻组成与统计分布, 导致整体指标未进一步改善, 该推测仍需通过方块电阻/接触电阻等表征验证.

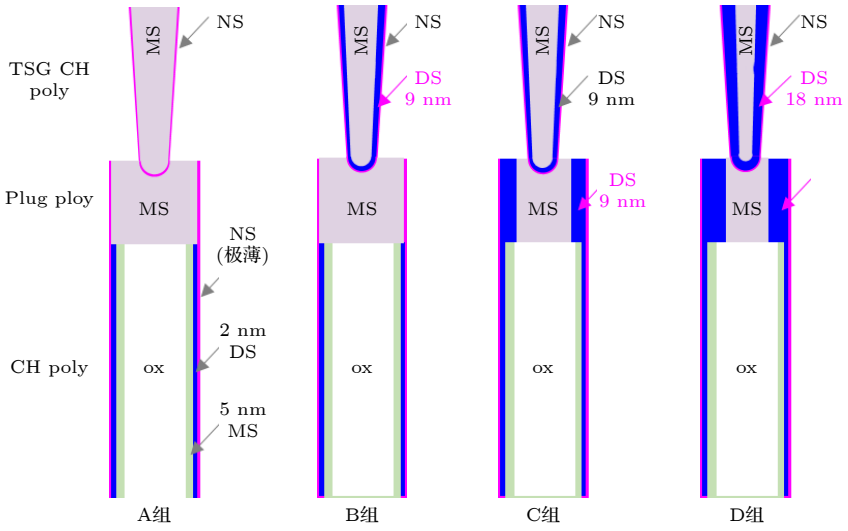

图 4 不同工艺条件分组实验示意图

Fig. 4. Schematic diagram of the experimental grouping under different process conditions.

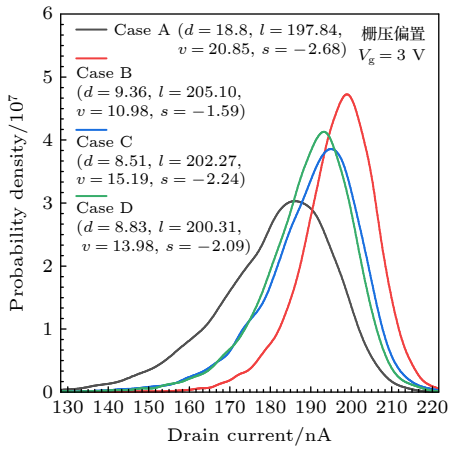

图 5 不同工艺条件下的电流分布对比

Fig. 5. Current distributions under the different process condition.

进一步, 我们测试了同一种工艺下, 两片不同晶圆的统计电流分布, 如图 6 所示, 可以观察到各工艺在各自不同晶圆下的统计电流分布也近似相同, 说明所使用的多晶硅沟道沉积工艺稳定, 具有量产能力.

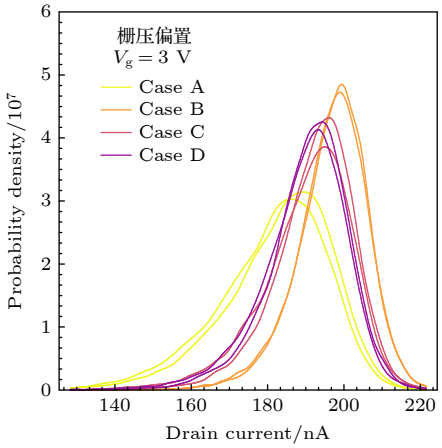

图 6 相同工艺条件下, 不同晶圆的电流分布对比

Fig. 6. Wafer-to-wafer variation of current distributions under the same process condition.

4 系统仿真与计算误差分析

在获取了最优工作条件后, 本文将使用前期构建的 3D NAND 存算一体仿真框架, 将实验数据统计得到的偏态 t 分布参数在矩阵-向量乘法 (matrix-vectormultiplication, MVM) 中以乘法噪声形式注入到每列电流累加结果, 噪声幅度由归一化标准差参数化. 平台以 Python 为开发语言, 当前版本聚焦 GPT-2 系列模型推理过程的软硬件协同评估. 其总体结构建立在 PyTorch 生态之上. 在保留原生张量表达、自动求导与模块封装接口的同时, 新增硬件模拟子层, 用以替换模型中若干具有高度矩阵乘-加特征的算子 (例如多头注意力中的线性投影与前馈网络的全连接层), 将其计算负载映射到前文工作中所定义的 3D NAND-SS 阵列中, 实现对存算一体原位矩阵向量乘过程的可控复现与评估.

在表 3 所示的 3D NAND-SS 架构参数和模型推理参数下进行仿真, 在 INT8 量化设置下对 GPT-2 124M 模型进行推理仿真, 得到不同电流分布噪声对应的前六层解码器中前馈神经网络的 MVM 计算误差对比结果. 如图 7 所示, 相较于基准工艺, 采用 B 组最优电流分布后 MVM 计算误差最小降低 14.7%, 最大降低 66.8%, 下降比取决于所在计算层参与的权重维度, 权重矩阵越大, 同比误差下降越多.

表 3 3D NAND 架构参数和模型推理参数

Table 3. 3D NAND architecture parameters and model inference parameters.

3D NAND架构参数	值	模型推理参数	值
Plane数每芯片	4	噪声分布选择	Skew-t分布
Block数每Plane	216	TOP-K	50
TSG数每Block	10	推理温度	1
BL数每Plane	131072	量化数	8
Layer数每芯片	32	量化模式	非对称静态量化
纵向切分数	216	模型	GPT-2 124M
横向切分数	1024		
ADC最大分辨率	128		

注: 缩减层数用于简化仿真, 实际产品为128层.

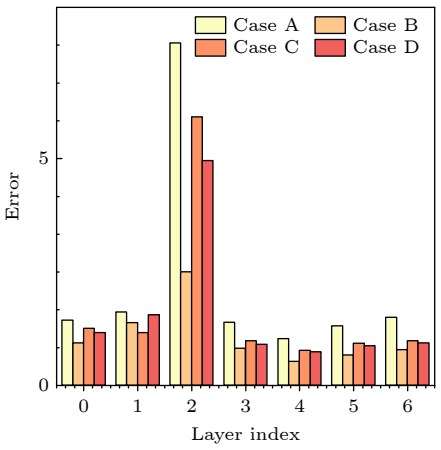

图 7 不同工艺电流分布输入下的 MVM 计算误差对比

Fig. 7. Comparative analysis of MVM calculation errors under current-distribution inputs across different fabrication processes.

5 结 论

本文面向 3D NAND 存算一体应用中串电流分布展宽导致计算误差上升的关键问题, 围绕 TSG 多晶硅沟道晶界势垒机制开展了仿真与实验研究. TCAD 结果表明晶界陷阱诱发的导带势垒会导致电流密度出现数量级波动, 从而成为开态电流分布展宽的重要物理来源. 基于此结果, 本文提出并验证了通过多晶硅前驱体组合实现等效氢钝化窗口优化的工艺方案, 在最优工艺方案条件下电流分布的归一化标准差较基准工艺下降了 50%. 进一步的 CIM 系统级仿真表明, 该最优电流分布可将 GPT-2 124M 的 INT8 量化条件下的 MVM 计算误差降低 14.7%—66.8%. 该结果为高性能 3D NAND CIM 芯片的工艺设计提供了重要参考.

参考文献

[1] Yu K, Kim S, Choi J R 2024 *IEEE Access* **12** 186679

[2] Milojicic D, Bresniker K, Campbell G, Faraboschi P, Strachan J P, Williams S 2018 *IEEE 38th International Conference on Distributed Computing Systems* Vienna, Austria, July 02–06, 2018 pp1300–1309

[3] Ma Y F, Du Y, Du L, Lin J, Wang Z F 2020 *Proceedings of the 2020 on Great Lakes Symposium on VLSI* New York, NY, USA, September 7–9, 2020 pp265–270

[4] Kim M, Liu M, Everson L R, Kim C H 2022 *IEEE J. Solid-State Circuits* **57** 625

[5] Shim W, Jiang H, Peng X, Yu S 2021 *Proceedings of the International Symposium on Memory Systems* Washington DC, USA, September 28-October 1, 2021 pp77–85

[6] Shim W, Yu S 2021 *IEEE J. Explor. Solid-State Comput. Devices Circuits* **7** 61

[7] Kang M, Kim H, Shin H, Sim J, Kim K, Kim L S 2022 *IEEE Trans. Comput.* **71** 1291

[8] Lin Y Y, Lee F M, Du P Y, Lin C C, Hsieh C C, Lee M H 2025 *IEEE Trans. Electron Devices* **72** 4837

[9] Shim W 2022 *J. Semicond. Technol. Sci.* **22** 341

[10] Lee S T, Lee J H 2020 *Front. Neurosci.* **14** 571292

[11] Nam K, Park C, Yoon J S, Yang G, Park M S, Baek R H 2022 *IEEE Trans. Electron Devices* **69** 3681

[12] Wang Y Q, Zhao Y L, Yu C X, Zhang J 2025 *Acta Phys. Sin.* **74** 186301 (in Chinese) [王禹齐, 赵耀林, 喻晨曦, 张俊 2025 物理学报 **74** 186301]

[13] Zou X Q, Jin L, Yan L, Zhang Y, Ai D, Zhao C L, Xu F, Li C L, Huo Z L 2019 *Solid-State Electron.* **153** 67

[14] Magramene A, Moumene M, Hadjoudja H, Zaidi B, Gagui S, Hadjoudja B, Chouial B, Chibani A 2023 *Int. J. Adv. Manuf. Technol.* **128** 4331

[15] Jahan I, Arellano J D, Shi Z 2025 *J. Mater. Chem. C* **13** 23675

[16] Scheller L P, Weizman M, Simon P, Fehr M, Nickel N H 2012 *J. Appl. Phys.* **112** 063711

[17] Kang D, Sio H C, Stuckelberger J, Liu R, Yan D, Zhang X, Macdonald D 2021 *ACS Appl. Mater. Interfaces* **13** 55164

[18] Zheng H, Liu H W, Fang Y X, Fan D Y, Han Y H, Hou C Y, Liu W, Xia Z L, Huo Z L 2025 *Acta Phys. Sin.* **74** 248502 (in Chinese) [郑好, 刘慧雯, 方语萱, 范冬宇, 韩玉辉, 侯春源, 刘威, 夏志良, 霍宗亮 2025 物理学报 **74** 248502]

[19] Boada I, Coll N, Madern N, Antoni Sellarès J 2008 *Int. J. Comput. Math.* **85** 1003

[20] Zou X Q, Xia Z L, Jin L, Zhang Y, Jiang D D, Li D H, Xu Q, Hong P Z, Zeng M, Gao J, Tang Z Y, Mei S N, Huo Z L 2016 *13th IEEE International Conference on Solid-State and Integrated Circuit Technology* Hangzhou, China, October 25–28, 2016 pp1122–1124

[21] Yang T, Xia Z L, Shi D D, Ouyang Y J, Huo Z L 2020 *IEEE J. Electron Devices Soc.* **8** 140

[22] Seto J Y W 1975 *J. Appl. Phys.* **46** 5247

[23] Mathur P C, Sharma R P, Shrivastava R, Saxena P, Kotnala R K 1983 *J. Appl. Phys.* **54** 3913

[24] Buss R J, Ho P, Breiland W G, Coltrin M E 1988 *J. Appl. Phys.* **63** 2808

SPECIAL TOPIC — Semiconductor physics and devices

Process optimization and validation of the polysilicon grain-boundary barrier for high-accuracy 3-dimensional NAND compute-in-memory chips

ZHENG Hao [1)2)] LIU Huiwen [3)] XU Kezhi [3)] ZHANG Baotong [3)]
YANG Yuancheng [3)] XIA Zhiliang [3)†] HUO Zongliang [1)3)‡]

1) (*Institute of Microelectronics of Chinese Academy of Sciences*, *Beijing* 100029, *China*)

2) (*University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

3) (*Yangtze Memory Technology Corp*, *Wuhan* 430070, *China*)

(Received 4 February 2026; revised manuscript received 10 March 2026)

Abstract

With the rapid development of artificial intelligence and edge computing, the computing-in-memory (CIM) architecture is considered a crucial technical path to alleviate the von Neumann bottleneck. While 3-dimensional (3D) NAND-based CIM schemes offer distinct advantages in high storage density and process maturity, their execution of analog computing tasks, such as matrix-vector multiplication (MVM), suffers from calculation accuracy degradation. This is primarily caused by the broadening of the string current distribution, which leads to accumulated current deviations. In particular, the polysilicon grain boundaries (GBs) within the top select gate (TSG) channel of 3D NAND strings play a decisive role in determining this current distribution.

To address this challenge, this study utilizes technology computer-aided design (TCAD) to construct a mature TSG Deck device model, analyzing the influence mechanism of potential barriers induced by GB traps on on-state current fluctuations. Simulation results demonstrate that acceptor-like traps at grain boundaries induce local potential barriers, and the variance of these barriers is the dominant physical source of on-state current instability. Guided by these physical insights, a novel process optimization strategy is proposed to modulate the equivalent hydrogen passivation window by combining polysilicon precursors with distinct nucleation and hydrogen-content characteristics (denoted as NS, MS, and DS). Specifically, innovatively inserting a 9 nm DS precursor interlayer between the NS nucleation layer and the MS bulk-fill layer creates a low-defect buffer zone, achieving in-situ hydrogen passivation of deep-level traps without compromising interface smoothness.

Wafer-scale statistical analysis of on-state current distributions across different process splits confirms that the optimal precursor combination reduces the normalized standard deviation of the bit-line current by 50% compared to the baseline process. Furthermore, to evaluate the system-level impact, the measured current distributions were fitted to a skew-t distribution and injected as multiplication noise into a custom CIM simulation framework. System-level simulations of INT8 quantization inference for the GPT-2 124M model indicate that the optimized device characteristics significantly reduce MVM calculation errors by 14.7% to 66.8%, depending on the weight matrix dimensions. In conclusion, this work bridges device-level process optimization with system-level performance, providing a highly manufacturable design basis for high-precision 3D NAND CIM chips.

Keywords: 3-dimensional NAND, computing-in-memory, top select gate (TSG), polysilicon grain boundary, current distribution

DOI: 10.7498/aps.75.20260199 **CSTR:** 32037.14.aps.75.20260199

† Corresponding author. E-mail: albert_xia@ymtc.com

‡ Corresponding author. E-mail: zongliang_huo@ymtc.com

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

面向高精度3维NAND存算一体芯片的多晶硅晶界势垒工艺优化与验证

郑好 刘慧雯 许克志 张宝通 杨远程 夏志良 霍宗亮

Process optimization and validation of the polysilicon grain-boundary barrier for high-accuracy 3-dimensional NAND compute-in-memory chips

ZHENG Hao LIU Huiwen XU Kezhi ZHANG Baotong YANG Yuancheng XIA Zhiliang HUO Zongliang

引用信息 Citation: Acta Physica Sinica, 75, 080804 (2026) DOI: 10.7498/aps.75.20260199
CSTR: 32037.14.aps.75.20260199
在线阅读 View online: https://doi.org/10.7498/aps.75.20260199
当期内容 View table of contents: http://wulixb.iphy.ac.cn
