近存计算架构 AI 芯片中子单粒子效应*

杨卫涛[1)2)†] 胡志良[3)] 何欢[2)] 莫莉华[3)] 赵小红[4)] 宋伍庆[1)] 易天成[3)]
梁天骄[3)] 贺朝会[2)] 李永宏[2)] 王斌[1)] 吴龙胜[1)] 刘欢[5)] 时光[5)]

1) (西安电子科技大学微电子学院, 西安 710071)

2) (西安交通大学核科学与技术学院, 西安 710049)

3) (散裂中子源科学中心, 东莞 523803)

4) (空军工程大学航空工程学院, 西安 710000)

5) (西安电子科技大学空间科学与技术学院, 西安 710071)

(2024 年 3 月 25 日收到; 2024 年 5 月 7 日收到修改稿)

利用中国散裂中子源大气中子辐照谱仪, 对某款 16 nm FinFET 工艺制造的近存计算架构人工智能 AI 芯片进行了大气中子单粒子效应辐照测试研究. 辐照测试中, 在累积中子注量为 1.51×10^{10} n/cm^2 (1 MeV 以上) 情况下, 共探测到 5 类共计 35 个软错误, 尤其是探测到不同于传统冯诺伊曼架构芯片单粒子效应的计算与存储单元同时发生单粒子效应新现象. 基于所探测到的两类功能单元同时单粒子效应新现象, 结合蒙特卡罗仿真模拟, 初步给出了近存计算架构 AI 芯片内物理布局上, 核心功能单元间可降低同时发生单粒子效应的安全间距建议. 该研究为进一步探究非传统冯诺伊曼架构芯片单粒子效应提供了参考与借鉴.

关键词: 近存计算, AI 芯片, 散裂中子源, 大气中子, 单粒子效应

PACS: 85.30.De, 21.60.Ka **DOI:** 10.7498/aps.73.20240430

1 引 言

高可靠的人工智能 (artificial intelligence, AI) 芯片是新一代信息技术的重要组成部分, 对于促进新质生产力健壮发展具有重要意义. 当前, AI 芯片正不断朝着大算力、海量数据等方向持续推进. 同时, 面对摩尔定律逐渐失效以及传统冯诺依曼架构芯片由于数据与存储分离导致的“功耗墙”和“存储墙”问题愈发显著的现实[1–3]. 实现了存储与计算单元紧密融合的存算一体架构 AI 芯片, 正蓬勃发展, 被认为是后摩尔时代最重要的芯片技术方向之一[4].

存算一体芯片根据存储与计算融合程度可分为近存计算架构 (near memory computing, NMC) 和存内计算架构 (in memory computing, IMC) 芯片[5,6]. 图 1(a)—(c) 分别为传统冯诺依曼架构、近存计算架构和存内计算架构芯片结构示意图[2,7]. 对于传统冯诺依曼架构芯片, 运算所需的大量数据主要存储于片外存储器. 如神经网络运算过程中需要的权重和偏差值, 在该架构中就主要存储于此类存储器中, 处理器进行运算时, 需要频繁经过控制线等对片外存储器进行访问, 花费大量时间和功耗. 对于近存计算架构芯片, 其由于计算和存储单元分布在同一芯片内, 且计算所需数据主要存储于计算单元周围. 因此, 在进行相同规模运算时, 相比于冯诺依曼架构而言, 可显著提升运算效率.

* 国家自然科学基金 (批准号: 12275211)、国家自然科学基金青年科学基金 (批准号: 62104260)、陕西省自然科学基础研究计划 (批准号: 2023-JC-QN-0015) 和中央高校基本科研业务费专项资金 (批准号: XJSJ23049) 资助的课题.

† 通信作者. E-mail: yangweitao01@xidian.edu.cn

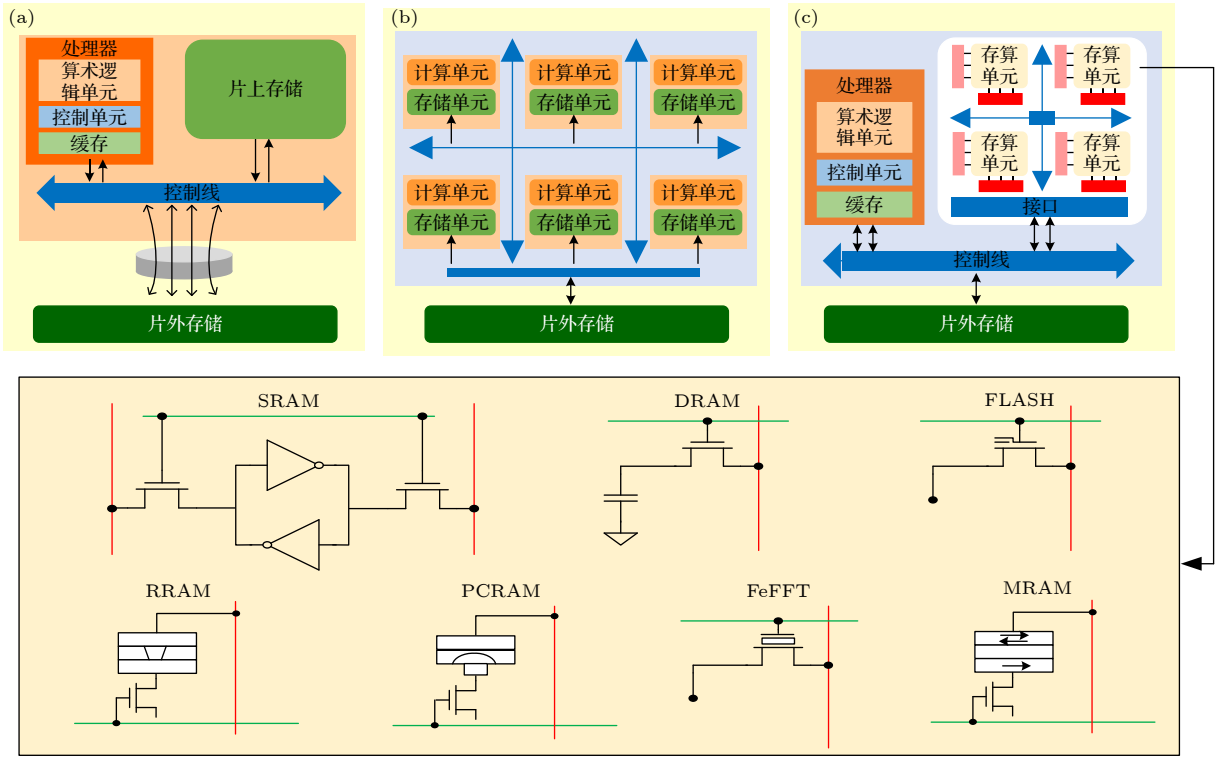


图 1 不同架构芯片结构示意图[2,7] (a) 冯诺依曼; (b) 近存计算; (c) 存内计算

Fig. 1. Different chip architectures[2,7]: (a) Von Neumann; (b) near memory computing; (c) in memory computing.

对于存内计算架构芯片, 其在存算单元内的存储器上可直接进行计算.

根据存储单元结构的不同, 可进一步细分为不同类型的存内计算架构芯片, 如静态随机存储器型 (static random access memory, SRAM)、动态随机存储器型 (dynamic random access memory, DRAM)、浮栅型 (FLASH)、阻变型 (resistive random access memory, RRAM)、相变型 (phase change random access memory, PCRAM)、铁电型 (ferroelectric field effect transistor, FeFET) 以及磁存储器型 (magnetoresistive random access memory, MRAM)[2].

未来, 随着先进制造工艺生产的不同架构存算一体 AI 芯片在各场景中的广泛部署与应用, 大气环境下, 不同能量中子入射导致的单粒子效应 (single event effect, SEE) 威胁将是存算一体芯片不得不考虑的问题. 例如, 当近存计算架构芯片应用于智能武器系统时、应用于无人机作业时、应用于智能驾驶时, 都可能遭受大气中子诱发的单粒子效应威胁, 甚至可能造成难以估量的后果. 因此, 本研究在国内率先瞄准大气中子导致的非传统架构芯片单粒子效应问题.

大气环境充斥着不同能量的中子, 其能量范围从 meV 到 GeV[8–10]. 不同能量中子能够与半导体中硅原子核发生核反应, 产生次级高能粒子进而诱发单粒子效应, 并导致芯片工作异常, 甚至烧毁[11]. 辐照实验是探究芯片大气中子单粒子效应敏感性的最直接手段. 不过, 自然环境中的中子通量低, 因此, 采用与大气中子能谱相近的散裂中子源进行辐照测试就成了最理想的选择[12]. 中国散裂中子源的投入运行为国内开展先进工艺 AI 芯片大气中子单粒子效应辐照测试研究提供了关键平台[13].

对包括 AI 芯片在内的不同对象的大气中子单粒子效应辐照测试研究, 已有报道主要还集中在传统冯诺依曼架构芯片, 尚缺乏对先进制造工艺的近存计算或者存内计算架构芯片的深入研究[9,14–17]. 而这些架构的变化是否会带来新的单粒子效应问题, 亟待深入探究. 比如, 对近存计算架构芯片而言, 其在单芯片内实现存储与计算单元紧密分布的同时, 是否会遭受单个高能粒子入射导致两者同时发生单粒子效应的问题.

为了探究新型架构芯片可能出现的单粒子效应新问题, 本研究以某款 16 nm FinFET 工艺近存

计算架构 AI 芯片为研究对象, 借助中国散裂中子源大气中子辐照谱仪, 对其进行初步的单粒子效应辐照测试. 与此同时, 结合蒙特卡罗模拟分析方法对实验中探测到的部分单粒子效应进行了分析. 通过本研究, 旨在为国内进一步开展非传统冯诺依曼架构芯片单粒子效应研究提供借鉴与参考.

2 辐照测试

2.1 待测芯片

本研究中, 辐照测试所用芯片为某款存储、计算和控制单元紧密融合于同一芯片内、且具有典型近存计算架构特征的 AI 芯片, 其生产制造工艺为 16 nm FinFET 工艺, 可应用于智能感知与识别、边缘计算、智能驾驶等领域. 其具有 26 TOPS 算力性能, 支持 8 位和 16 位的 ONNX 和 TensorFlow 框架. 芯片内集成有由 8 个神经网络运算集群组成的神经网络处理单元, 通用处理器单元和图像处理单元, 以及可配置且可通过滑动窗口访问的存储器, 神经网络处理单元互联控制器等. 此外, 芯片内还集成有多种接口[18,19]. 图 2 为该款 AI 芯片神经网络数据流架构及分类识别 AI 应用示意图. 图 2 中, 顶部神经网络图中的 Layer1, Layer2 和 Layer3 分别为神经网络卷积层示意图, 中间资源映射图为每个卷积层所需的芯片资源映射情况示意图, 底部网络配置图为近存计算架构芯片内神经网络设计具体资源配置示意图. 对于该芯片而言, 其可根据待测神经网络规模大小, 映射成不同配置的存储、控制和计算单元. 虽然中子辐照时可以不对芯片进行开盖, 但为了更准确地掌握待测芯片的尺寸信息, 以及辐照过程中保证束斑中心位置与裸片中心位置重合, 辐照测试前, 对芯片亦进行了开盖处理. 图 3 为开盖后的芯片照片, 开盖后的裸芯片大小为 0.8 cm×1.0 cm.

该 AI 芯片通过 M.2 接口与主控开发板相连接. 主控开发板负责将编译后的可执行文件传输到目标芯片, 后者根据接收到的可执行文件进行资源布局、配置和实现. 一旦启动测试后, 近存计算架构 AI 芯片即可独立运行. 通过摄像头捕获目标, 再利用部署的神经网络对捕获的目标进行识别探测, 并将检测结果通过 HDMI 接口进行实时输出显示.

2.2 辐照实验

辐照实验在中国散裂中子源大气中子辐照谱仪进行. 图 4 为实验中所用能谱图, 其中 ANIS 能谱数据由蒙特卡罗粒子输运程序 (MCNPX) 根据中国散裂中子源靶站-谱仪工程模型计算所得. 而 JEDEC 数据为纽约户外海平面基于 bonner 球测量得到 1 MeV 以上中子能谱[20]. 实验时对应 I5—II3

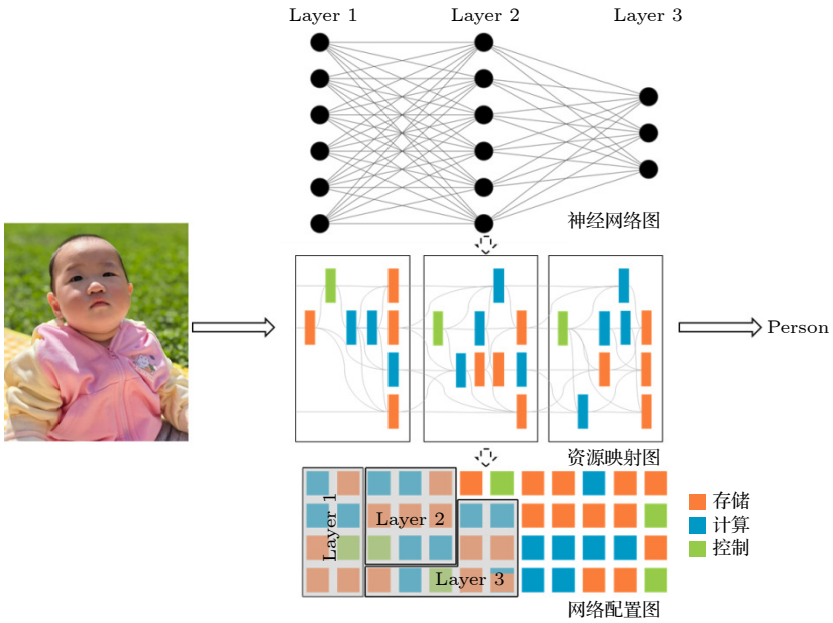


图 2 待测芯片数据流架构及 AI 应用示意图

Fig. 2. Diagram of data flow architecture and AI applications for the test chip.

工况, 样品位置距离靶心 20.8 m, 加速器运行功率为 140 kW, 其 1 MeV 以上中子注量率约为真实大气中子注量率的 2.56×10^8 倍.

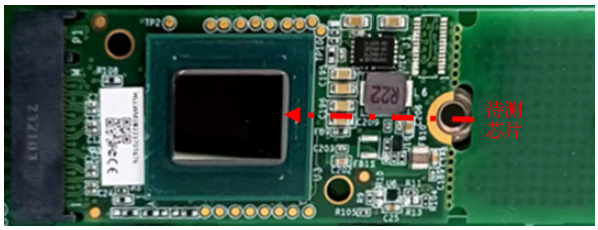


图 3 开盖后的待测芯片照片

Fig. 3. Photo of the de-capped test chip.

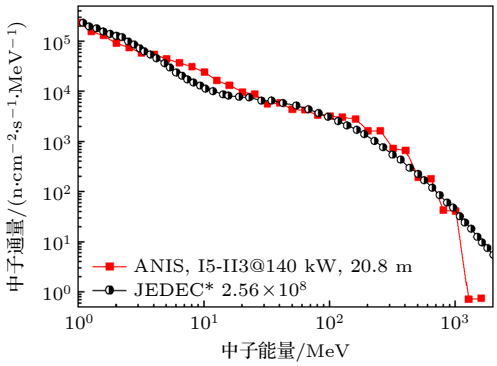


图 4 辐照实验所用中子能谱, 其中 ANIS 为实验所用能谱, JEDEC 为参考能谱

Fig. 4. Neutron spectrum applied in the irradiation test, ANIS with the utilized spectrum during irradiation test, and the JEDEC with the referred.

辐照实验中, 1 MeV 以上中子注量率为 1.41×10^6 n·cm^{-2}·s^{-1}, 累积注量为 1.51×10^{10} n·cm^{-2}. 辐照过程中, 所使用的束斑大小为 2 cm×2 cm, 束斑均匀性为±10%, 该束斑可覆盖整个待测芯片. 辐照期间, 该芯片实时运行部署于其中的 YOLOV5 神经网络模型, 所测试的网络共包含 3 个卷积层, 对应神经元数目分别为 94, 85 和 75, 所需的可通过滑动窗口访问的最大窗口对应存储器容量为 5.875 MB. 基于运行的网络模型, 在芯片辐照期间对辐照室内所放置的目标物进行实时分类识别检测, 目标物共包括: 鼠标、键盘和行李箱. 图 5 为辐照现场照片.

辐照测试过程中, 实时监测并输出计算、存储和控制单元的出错信息. 如果某一时刻只输出有存储单元出错信息, 则表明此时只有存储单元发生单粒子效应. 如果只输出有计算单元出错信息, 则意味着此时只有计算单元发生单粒子效应. 若同时输出计算单元和存储单元出错信息, 则表明两者同时发生了单粒子效应.

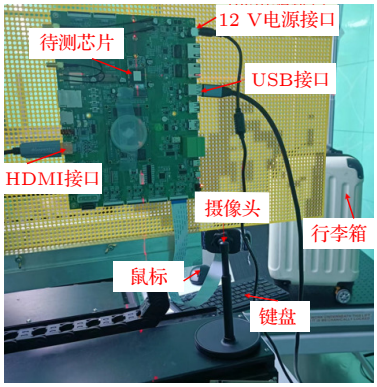


图 5 辐照实验现场照片

Fig. 5. Photo of the irradiation worksite.

2.3 辐照测试结果

在 1.51×10^{10} n·cm^{-2} 的累积中子注量下, 共探测到 5 类, 合计 35 个单粒子效应软错误, 分别为存储单元单粒子翻转、计算单元单粒子翻转、存储与计算单元同时单粒子翻转、超时、进程终止. 各软错误的具体描述如下.

存储单元单粒子翻转 (SEU/MEM): 粒子入射导致存储单元发生数据翻转.

计算单元单粒子翻转 (SEU/COMP): 粒子入射导致计算单元内发生数据翻转.

存储与计算单元同时单粒子翻转 (SEU/(MEM+COMP)): 粒子入射导致存储和计算单元内同时发生了数据翻转.

超时 (Timeout): 粒子入射导致神经网络处理过程超过了最大设定时间.

进程终止 (Process-killed): 粒子入射导致 Linux 进程中止, 神经网络停止运行.

表 1 列出了探测到的不同类型单粒子效应发生情况. 进一步, 存储单元单粒子翻转发生情况见表 2.

表 1 探测到的不同类型单粒子效应

Table 1. Detected kinds of single event effect.

软错误	数量
SEU/MEM	30
SEU/COMP	2
SEU/MEM+COMP	1
Timeout	1
Process-killed	1

表 2 存储单元单粒子效应

Table 2. Single event effect in memory cell.

翻转单元	数量	翻转单元	数量
1	3	10	8
2	5	11	1
4	2	13	1
8	10		

3 分析与讨论

近存计算架构 AI 芯片实现了存储与计算单元的紧密融合, 使得其在运行需要大量运算的人工智能应用时, 可直接在芯片内部存储单元读取对应权重数值以进行快速的运算处理. 然而, 通过辐照测试可以看出, 这种紧密融合确实又带来了新的单粒子效应问题, 如表 1 中所探测到的存储与计算单元同时单粒子翻转问题.

在传统冯诺依曼架构的 AI 芯片中, 权重和偏差数值通常存储于外部的双倍速率同步动态随机存储器, 即 DDR 中, 而运算操作则在计算单元内进行. 这种情况下, 虽然数据传输会导致额外的功耗和延迟开销, 但是由于存储与计算单元分别分布于电路板上的不同位置, 且两者封装于不同芯片内. 因此, 一般不会出现存储单元与计算单元同时发生单粒子翻转的情况.

针对本文所研究的近存计算架构 AI 芯片, 由图 2 可知, 其实现了存储、计算和控制单元数据流在单芯片内的传输与处理, 且这 3 种单元紧密分布, 这就给多单元同时发生单粒子效应提供了可能. 接下来, 结合蒙特卡罗模拟对大气中子入射导致的近存计算架构 AI 芯片单粒子效应进行更详细的分析.

3.1 蒙特卡罗模拟

本文所研究的待测芯片为 16 nm FinFET 工艺, 其采用倒封装形式. 大气中子入射造成的单粒子效应主要来源于不同能量中子与半导体中硅原子核发生核反应产生的次级粒子. 因此, 次级粒子的线性能量转移值 (linear energy transfer, LET) 和硅中射程对不同类型单粒子效应的发生有重要影响.

参照所测试的 AI 芯片的纵向结构信息, 如图 6 所示, 利用 Geant4 仿真软件[21,22], 构建了 0.08 cm×0.1 cm×789 μm 的探测器靶. 其中, 纵向的 789 μm 结构中, 上层硅衬底厚度为 784.43 μm, 衬底下方 10 层金属合计厚度为 3.07 μm, SiO_2 层合计厚度为 1.50 μm. 仿真中, 粒子源为与散裂中子源大气中子辐照谱仪相匹配的中子能谱, 共进行了 8×10^8 个粒子的仿真, 对应的注量为 10^{11} n·cm^{-2}, 所采用的物理模型为 FTFP_BERT_HP. 期间, 重点统计了产生的次级粒子类型和能量.

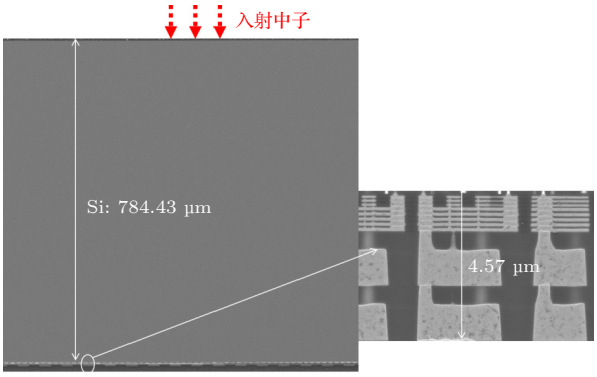


图 6 所测试的 AI 芯片纵向结构信息

Fig. 6. Vertical structure of the tested AI chip.

中子不带电, 不同能量的中子需要与硅原子核发生核反应产生次级高能重离子, 才有可能导致单粒子效应的发生. 大气中子包含不同能量的中子, 其能够与硅原子核发生 (n, α), (n, p), (n, d), (n, n-α) 等反应, 可产生射向不同方向的次级重离子或碎片化离子, 这些离子就有可能在芯片内穿过多个单元并沉积能量, 进而造成多单元的同时翻转. 图 7 为仿真分析模型示意图, 次级粒子导致两种不同单元同时发生单粒子效应, 需要满足以下两个条件:

1) LET 值大于阈值;

2) 硅中射程大于两单元最小间距.

由图 7 可以看出, 当次级粒子在满足 LET 数值大于阈值时, 一旦硅中射程超过了两单元最小间

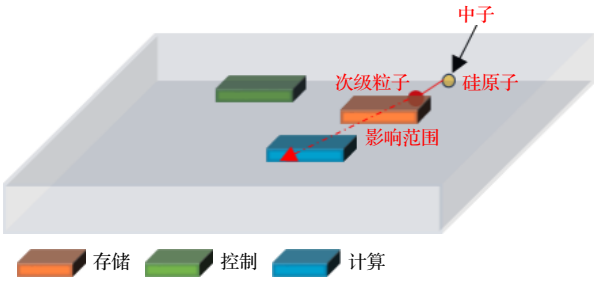


图 7 次级粒子影响多个单元示意图

Fig. 7. Diagram of affected cells by secondary particle.

距, 就有可能导致两者同时发生单粒子效应的后果. 虽然, 受辐照机时、累积注量等因素影响, 当前辐照测试中探测只到了存储与计算单元的同时翻转, 但这并不意味着其他情况下的多单元同时单粒子效应现象不会出现. 比如存储单元与控制单元同时发生单粒子效应, 甚至存储单元、计算单元与控制单元同时发生单粒子效应情况. 这些现象都对近存计算架构 AI 芯片的单粒子效应加固设计提出了新的要求.

图 8 为蒙特卡罗模拟所产生的主要次级粒子对应的 LET 和硅中射程. 由图 8 可知, 大气中子入射产生的次级粒子主要有 Ne, Si, Al, Mg, He 等, 对应的 LET 范围为 0.30—8.70 MeV·cm^2/mg, 对应的硅中射程最大可到百微米级. 对于 14/16 nm FinFET 工艺芯片而言, 据报道, 其单粒子效应 LET 阈值约为 0.10 MeV·cm^2/mg[23]. 因此, 图 8 所示重离子足以诱发不同单元单粒子效应 (满足条件 1). 与此同时, 如果以获得的最大硅中射程为参考, 那么可以推断, 该近存计算架构 AI 芯片内, 部分存储与计算单元的物理间距应该在百微米以内 (满足条件 2). 这也意味着, 在近存计算架构 AI 芯片设计中, 保证性能的前提下, 如有可能, 将关键单元物理布局之间的间距设置为不少于百微米时, 有可能降低地面应用场景中由大气中子导致的两类功能单元同时发生单粒子效应的风险.

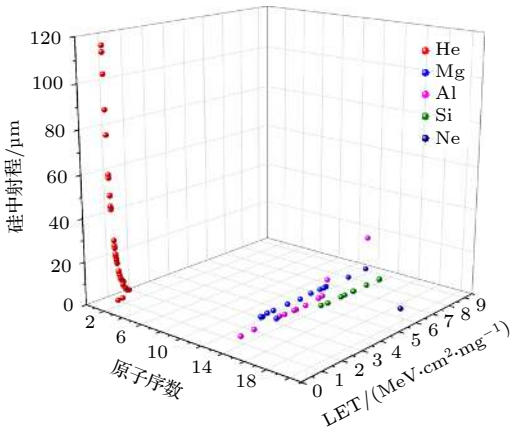


图 8 大气中子入射硅半导体所产生的主要次级粒子

Fig. 8. Secondary particles of atmospheric neutron striking silicon.

3.2 软错误敏感性

辐照测试过程中, 在 1 MeV 以上中子累积注量为 1.51×10^{10} n·cm^{-2} 的情况下, 如表 1 所示, 存储单元单独发生单粒子效应次数为 30 次, 计算单元单独发生单粒子效应次数为 2 次, 两者同时发生单粒子效应次数为 1 次. 为了量化分析不同功能单元单粒子效应软错误敏感性, 将同时发生的 1 次单粒子效应分别对计算和存储单元各计算 1 次, 则存储和计算单元分别发生单粒子效应为 31 次和 3 次. 控制单元数据翻转会导致超时和进程终止现象的出现, 因此, 在量化分析时, 所探测到的超时和进程终止软错误记为控制单元的单粒子效应.

对于存储、计算和控制单元, 根据 (1) 式和 (2) 式, 可计算其单粒子效应截面和软错误率. 表 3 分别为各功能单元单粒子效应截面和北京地面软错误率具体数值 (其中软错误计算参考北京地面 1 MeV 以上大气中子通量: 14.80 n·cm^{-2}·h^{-1})[9,20]. 由表 3 可知, 对于该近存计算架构 AI 芯片而言, 存储单元软错误率分别为计算和控制单元的 10.34 倍和 15.51 倍. 导致这一现象的主要原因是近存计算架构 AI 芯片中, 存储单元在芯片内的面积占比大于计算和控制单元, 导致离子入射轰击存储单元的概率大于其他单元.

$$\sigma = n/F, \tag{1}$$

$$\varphi = 10^9 \cdot \sigma \cdot \eta, \tag{2}$$

式中, σ 单粒子效应截面 (cm^2), n 为软错误个数, F 为中子注量 (n·cm^{-2}), ϕ 为软错误率 (FIT), η 为真实环境下的中子通量 (n·cm^{-2}·h^{-1}).

表 3 不同单元效应截面和软错误率

Table 3. Cross section and soft error rate of different cells.

单元	单粒子效应截面/(10^{-10} cm^2)	软错误率/FIT
存储	20.5	30.40
计算	1.99	2.94
控制	1.32	1.96

针对发生次数最多的存储单元单粒子翻转, 表 2 中统计了存储单元具体的单粒子效应信息. 可以看出, 8 单元翻转占比最大, 为 33.33%, 10 单元翻转占比次之, 为 26.67%. 两者之和超过了 50%, 这也表明对于 16 nm 及更小尺寸 AI 芯片单粒子效应, 多单元翻转也是近存计算架构 AI 芯片单粒子效应加固需要重点考虑的问题. Yang 等[24] 报道了 14 MeV 中子入射导致的 14 nm FinFET 工艺 SRAM 单粒子效应, 指出应考虑多单元翻转的影

响. Takashi 等[25] 报道了 1—300 MeV 散裂中子导致的 12 nm FinFET 工艺 SRAM 单粒子效应, 重点关注了散裂中子导致的多单元翻转问题. 这些都表明对于更小尺寸电子系统, 多单元翻转的影响不容忽视.

同时需要说明的是, 相比于真实情况下的大气中子, 实验所用大气中子辐照谱仪高能中子部分偏软. 这可能使得根据当前辐照测试结果所估计的软错误相较真实情况存在保守估计的可能性. 此外, 当前研究主要关注了架构变化带来的单粒子效应新问题, 未涉及不同能量中子贡献. 而大气中子能量范围从 meV 到 GeV, 后续还需要结合所发现的单粒子效应新问题, 进一步开展更多实验以探究不同能段中子的影响.

4 结 论

针对一款 16 nm FinFET 工艺制造的典型近存计算架构人工智能芯片, 在国内首次利用中国散裂中子源大气中子辐照谱仪开展了大气中子单粒子效应辐照测试研究. 在中子累积注量为 1.51×10^{10} n·cm^{-2} (1 MeV 以上) 下, 共探测到软错误 35 次, 包括存储单元单粒子翻转、计算单元单粒子翻转, 以及超时和进程终止, 尤其是探测到了存储与计算单元同时发生单粒子效应的新现象. 此外, 发现存储单元单粒子效应分别为计算和控制单元的 10.34 倍和 15.51 倍. 研究指出, 未来对于近存计算架构存算一体芯片单粒子效应加固, 既要考虑多类功能单元同时发生单粒子效应的新问题, 还要考虑存储单元多单元翻转问题. 在近存计算架构 AI 芯片设计中, 在保证性能的前提下, 若可将核心单元的物理分布间距设置在百微米以上, 则可以降低多单元同时遭受大气中子单粒子效应的风险.

参考文献

[1] Zhou Z, Huang P, Kang J F 2022 *Acta Phys. Sin.* **71** 148507 (in Chinese) [周正, 黄鹏, 康晋锋 2022 物理学报 **71** 148507]

[2] Guo X J, Wang G Y, Wang S D 2023 *J. Electron. Inf. Technol.* **45** 1888 (in Chinese) [郭昕婕, 王光耀, 王绍迪 2023 电子与信息学报 **45** 1888]

[3] Sun Z, Kvatinsky S, Si X, Mehonic A, Cai Y, Huang R 2023 *Nat. Electron.* **6** 823

[4] Kang W, Kou J, Zhao W S 2024 *Sci. Sin. Inf.* **54** 16 (in Chinese) [康旺, 寇竞, 赵巍胜 2024 中国科学: 信息科学 **54** 16]

[5] Kamil K, Sudeep P, Ryan G K 2020 *J. Low Power Electron. Appl.* **10** 30

[6] Liu W Q, Chen K, Wu B, Deng E Y, Wang Y, Gong Y, Cui Y J, Wang C H 2024 *Sci. Sin. Inf.* **54** 34 (in Chinese) [刘伟强, 陈珂, 吴比, 邓尔雅, 王佑, 龚宇, 崔益军, 王成华 2024 中国科学: 信息科学 **54** 34]

[7] Wilfried H, Anand R, Kaushik R, Bhaswar C, Charudatta M P, Cheng W, Supratik G 2023 *Adv. Mater.* **35** 2204944

[8] Hu Z L, Yang W T, Li Y H, Li Y, He C H, Wang S L, Zhou B, Yu Q Z, He H, Xie F, Bai Y R, Liang T J 2019 *Acta Phys. Sin.* **68** 238502 (in Chinese) [胡志良, 杨卫涛, 李永宏, 李洋, 贺朝会, 王松林, 周斌, 于全芝, 何欢, 谢飞, 白雨蓉, 梁天骄 2019 物理学报 **68** 238502]

[9] Yang W T, Li Y H, Li Y, Hu Z L, Xie F, He C H, Wang S L, Zhou B, He H, Khan W, Liang T J 2019 *Microelec. Reliab.* **99** 119

[10] Hu Z L, Yang W T, Zhou B, Liu Y N, He C H, Wang S L, Yu Q Z, Liang T J 2023 *J. Nucl. Sci. Technol.* **60** 473

[11] Wang X, Zhang F Q, Chen W, Guo X Q, Ding L L, Luo Y H 2020 *Acta Phys. Sin.* **69** 162901 (in Chinese) [王勋, 张凤祁, 陈伟, 郭晓强, 丁李利, 罗尹虹 2020 物理学报 **69** 162901]

[12] Wang X, Zhang F Q, Chen W, Guo X Q, Ding L L, Luo Y H 2019 *Acta Phys. Sin.* **68** 052901 (in Chinese) [王勋, 张凤祁, 陈伟, 郭晓强, 丁李利, 罗尹虹 2019 物理学报 **68** 052901]

[13] Cao S, Yin W, Zhou B, Hu Z L, Shen F, Yi T C, Wang S L, Liang T J 2024 *Acta Phys. Sin.* **73** 092501 (in Chinese) [曹嵩, 殷雯, 周斌, 胡志良, 沈飞, 易天成, 王松林, 梁天骄 2024 物理学报 **73** 092501]

[14] Wang H B, Wang Y S, Xiao J H, Wang S L, Liang T J 2021 *IEEE Trans. Nucl. Sci.* **68** 394

[15] Dimitris A, Nikos F, Aitzan S, Vasileios V, Ioanna S, Mihalis P, Ye R, John G, Mikel L, Maria K, Carlo C, Chris F 2024 *IEEE Trans. Reliab.* **73** 771

[16] Rubens L R J, Sujit M, Carlo C, Maria K, Manon L, Christopher F, Paolo R 2022 *IEEE Trans. Nucl. Sci.* **69** 567

[17] Jordan D A, Jennings C L, Michael J W 2018 *IEEE Radiation Effects Data Workshop (REDW) Waikoloa*, HI, USA

[18] Avi B, Givat S, Or D, Kiryat O, Daniel C, Ramat G, Gilad N, Modiin-Maccabim R 2023 US Patent 11551028 B2

[19] Hailo-8 AI Accelerator. https://hailo.ai/products/ai-accelerators/hailo-8-ai-accelerator/. [2023-10-1]

[20] Measurement and Reporting of Alpha Particle and Terrestrial Cosmic Ray-induced Soft Errors in Semiconductor Devices. https://www.jedec.org/document_search?search_api_views_fulltext=JESD89A. [2024-2-11]

[21] Allison J, Amako K, Apostolakis J, et al. 2006 *IEEE Trans. Nucl. Sci* **53** 270

[22] Zhang Z G, Lei Z F, Tong T, Li X H, Wang S L, Liang T J, Xi K, Peng C, He Y J, Huang Y, En Y F 2020 *Acta Phys. Sin.* **69** 056101 (in Chinese) [张战刚, 雷志锋, 童腾, 李晓辉, 王松林, 梁天骄, 习凯, 彭超, 何玉娟, 黄云, 恩云飞 2020 物理学报 **69** 056101]

[23] Mo L H, Ye B, Liu J, Zhang Z G, Tong T, Sun Y M, Luo J 2021 *Nucl. Phys. Rev.* **38** 327

[24] Yang S H, Zhang Z Z, Lei Z F, Tong T, Li X H, Xi K, Wu F G 2022 *Appl. Sci.* **12** 9685

[25] Takashi K, Masanori H, Hideya M 2020 *IEEE Trans. Nucl. Sci.* **67** 1485

Neutron induced single event effects on near-memory computing architecture AI chips*

Yang Wei-Tao [1)2)†] Hu Zhi-Liang [3)] He Huan [2)] Mo Li-Hua [3)]
Zhao Xiao-Hong [4)] Song Wu-Qing [1)] Yi Tian-Cheng [3)] Liang Tian-Jiao [3)]
He Chao-Hui [2)] Li Yong-Hong [2)] Wang Bin [1)] Wu Long-Sheng [1)]
Liu Huan [5)] Shi Guang [5)]

1) (*School of Microelectronics, Xidian University, Xi'an* 710071, *China*)

2) (*School of Nuclear Science and Technology, Xi'an Jiaotong University, Xi'an* 710049, *China*)

3) (*Spallation Neutron Source Science Center, Dongguan* 523803, *China*)

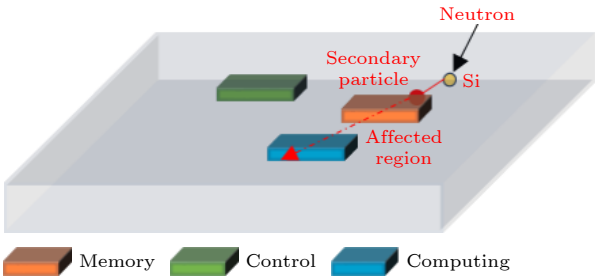
4) (*Aviation Engineering Institute, Air Force Engineering University, Xi'an* 710000, *China*)

5) (*School of Aerospace Science and Technology, Xidian University, Xi'an* 710071, *China*)

(Received 25 March 2024; revised manuscript received 7 May 2024)

Abstract

For the near-memory computing architecture AI chip manufactured by using 16 nm FinFET technology, atmospheric neutron single event effect irradiation tests are conducted for the first time in China by using the atmospheric neutron irradiation spectrometer (ANIS) at the China Spallation Neutron Source. During the irradiation, the YOLOV5 algorithm neural network running on the AI chip is used for real-time detection of target objects, including mice, keyboard, and luggage. The purpose of the test is to investigate the new single event effect that may occur on near-memory computing architecture AI chip. Finally, at an accumulated neutron fluence of 1.51×10^{10} n·cm^{-2} (above 1 MeV), a total of 35 soft errors are detected in 5 categories. Particularly noteworthy is the observation of a new finding, where both computing and memory units experience single event effects simultaneously, which is different from the traditional von Neumann architecture chips. Based on the single event effects that occur simultaneously in these two units, combined with Monte Carlo simulation, a preliminary estimation is made of the physical layout distance between the computing unit and the memory unit on the chip. Furthermore, suggestions are proposed to simultaneously reduce the risk of single event effect in multi cells. This study provides valuable reference and insights for further exploring the single event effects in non-traditional von Neumann architecture chips.



Keywords: near memory computing, AI chip, spallation neutron source, atmospheric neutron, single event effect

PACS: 85.30.De, 21.60.Ka **DOI:** 10.7498/aps.73.20240430

* Project supported by the National Natural Science Foundation of China (Grant No. 12275211), the National Natural Science Foundation of China Young Scientists Fund (Grant No. 62104260), the Natural Science Basic Research Plan of Shaanxi Province, China (Grant No. 2023-JC-QN-0015), and the Fundamental Research Funds for the Central Universities, China (Grant No. XJSJ23049).

† Corresponding author. E-mail: yangweitao01@xidian.edu.cn

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

近存计算架构AI芯片中子单粒子效应

杨卫涛 胡志良 何欢 莫莉华 赵小红 宋伍庆 易天成 梁天骄 贺朝会 李永宏 王斌 吴龙胜 刘欢 时光

Neutron induced single event effects on near-memory computing architecture AI chips

Yang Wei-Tao Hu Zhi-Liang He Huan Mo Li-Hua Zhao Xiao-Hong Song Wu-Qing Yi Tian-Cheng
Liang Tian-Jiao He Chao-Hui Li Yong-Hong Wang Bin Wu Long-Sheng Liu Huan Shi Guang

引用信息 Citation: Acta Physica Sinica, 73, 138502 (2024) DOI: 10.7498/aps.73.20240430
在线阅读 View online: https://doi.org/10.7498/aps.73.20240430
当期内容 View table of contents: http://wulixb.iphy.ac.cn