A FPGA friendly approximate computing framework with hybrid Neural networks

Haiyue Song, Xiang Song, Tianjian Li, Naifeng Jing, Xiaoyao Liang and Li Jiang*
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Neural approximate computing with fine-grain quality control is promising to gain energy-efficiency and performance by the trade-off the tolerable errors. Classifier-approximator hybrid architecture, providing fine-grain control of qualities between approximate and accurate execution, is widely used. However, they are not compatible to a heterogeneous computing platform, due to the large communication overhead between the approximate and accurate cores, and the large speed gap between them.

This paper proposes a novel hybrid approximate computing architecture containing a multi-class classifier and multiple approximators (MCMA) with the corresponding iterative co-training methods, which can optimize the invocation of the approxiamtor for higher utilized of the accelerator, and minimiz the communication of redistributing the unsafe-to-approxiamte data. We leverage high-level synthesis tool to generate a microarchitecture with pipelined data path for hiding the communication latency. The experiments on off-the-shelf programable SoC show the superior of the proposed architecture.

摘要：具有细粒度质量控制的，基于神经网络的近似计算是通过调控可接受误差来提高能效，优化性能的好方法。具有在近似计算和精确执行之间进行细粒度质量控制的分类器-近似器结构被广泛采用。但是，由于近似器和精确执行内核之间的通信开销，以及两者之间较大的速度差距，这种方法并不能与异构计算机平台兼容。

本文提出了一种新的混合近似计算结构，该结构使用了迭代式联合训练方法，包含了一个多分类器和多个近似器，可以提高加速器的利用率，降低与精确计算核之间的通信开销。我们利用高级综合分析工具来生成一个具有流水线数据通路的微体系结构来降低通信延迟。在现成的可编程 SoC 上进行的实验表明了我们架构的优越性。

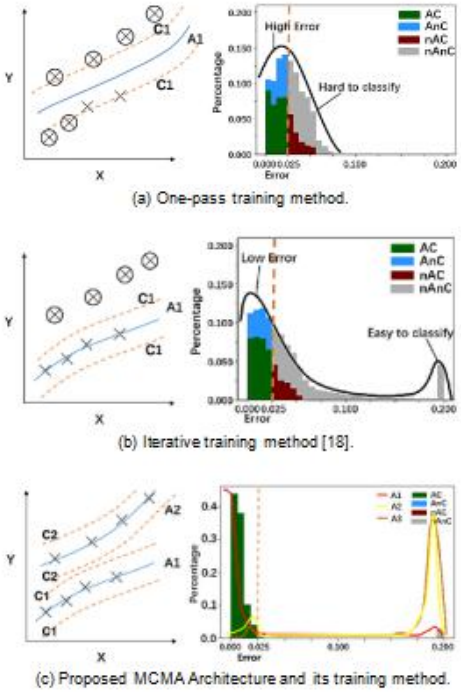


Figure 1: Concept (left column) and the Error Distribution of output from the approximator (right column) for different training methods

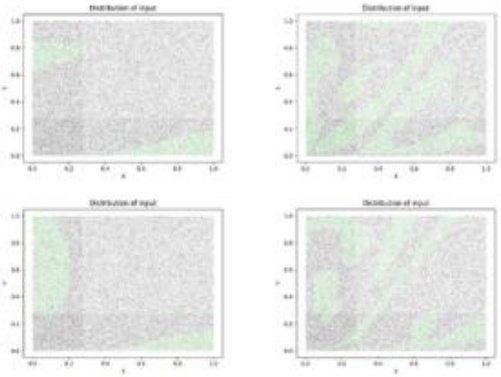


Figure 3: Data distribution using category C (left) and A (right) in two continuous iterations from top to bottom

A Novel Test Method for Metallic CNTs in CNFET-Based SRAMs

Tianjian Li, Feng Xie, Xiaoyao Liang, Qiang Xu, *Member, IEEE*,
Krishnendu Chakrabarty, *Fellow, IEEE*, Naifeng Jing, and Li Jiang, *Member, IEEE*

***Abstract*—Static random access memories (SRAMs) built on carbon nanotube field effect transistors (CNFETs) are promising alternatives to conventional CMOS-based SRAMs, due to their advantages in terms of power consumption and noise immunity. However, the nonideal carbon nanotube (CNT) fabrication process generates metallic-CNTs (m-CNTs) along with semiconductor-CNTs, leading to correlated faulty cells along the growth direction of the m-CNTs. In this paper, we propose a novel low-cost test solution to detect such faults. Instead of using conventional March test to test each and every SRAM cell, we selectively test certain SRAM cells and judiciously skip testing other SRAM cells between the selected cells. To ensure high fault coverage, we propose three jump test algorithms for differ-ent CNFET-SRAM layouts. Moreover, we model m-CNT-induced SRAM faults and characterize their distribution in the SRAM array. Experimental results show that the proposed solutions are able to achieve high fault coverage with low test cost.**

***Index Terms*—Carbon nanotubes (CNTs), integrated circuit testing, static random access memory (SRAM) cells.**

摘要：基于碳纳米晶体管(CNFETs)的静态随机存取存储器(SRAMs)是传统的基于 cmos 的 sram 的替代方案，由于其在功耗和抗噪声方面的优势。然而，非理想碳纳米管(CNT)的晶化过程与半导体-碳纳米管一起产生金属碳纳米管(m-CNTs)，导致沿着 m-CNT 生长方向的故障单元耦合在一起。本文提出了一种新的低成本检测解决方案。我们不使用常规的 March Test 来测试每个 SRAM 单元，而是选择性地测试某些 SRAM 单元，并明智地跳过在选定的单元间测试其他 SRAM 单元。为了确保高故障覆盖率，我们提出了三种不同的 CNFET-SRAM 布局的跳跃测试算法。此外，我们还对 m- cnt 诱导的 SRAM 故障进行了建模，并对其在 SRAM 阵列中的分布进行了表征。实验结果表明，所提出的解决方案能够实现低测试成本的高故障覆盖率

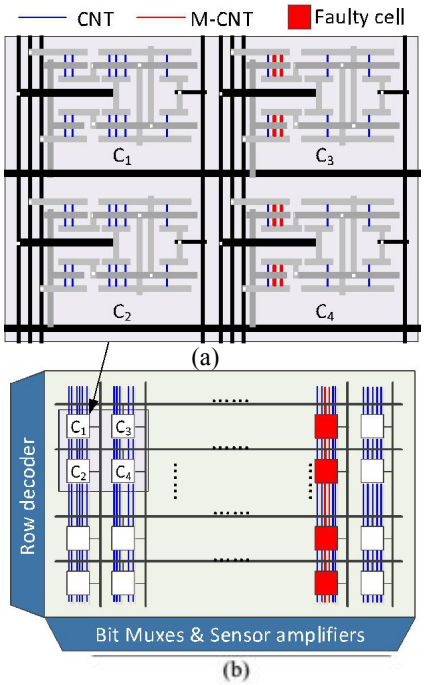


Fig. 1. CNFET-based SRAM layout. (a) Four SRAM cells. (b) SRAM array.

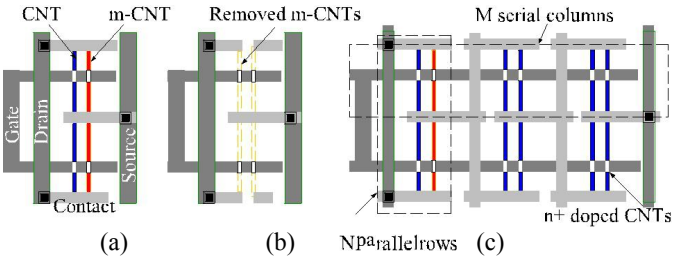


Fig. 2. m-CNT in CNFET. (a) m-CNT leads to short FET. (b) m-CNT removal leads to open FET. (c) Misaligned CNTs.

Accelerator-friendly Neural-network Training: Learning Variations and Defects in RRAM Crossbar

Lerong Chen[1], Jiawen Li[1], Yiran Chen[2], Qiuping Deng[3] Jiyuan Shen[1] Xiaoyao Liang[1] and Li Jiang[1]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Electrical and Computer Engineering, University of Pittsburgh, PA
[3]Lynmax Research, Beijing, China
Email: {clion0003, myloveys, shenjiyuan, liang-xy, ljiang cs}@sjtu.edu.cn, yiran.chen@pitt.edu, dengqiuping@lynmaxtech.com

Abstract—RRAM crossbar consisting of memristor devices can natu-rally carry out the matrix-vector multiplication; it thereby has gained a great momentum as a highly energy-efficient accelerator for neuro-morphic computing. The resistance variations and stuck-at faults in the memristor devices, however, dramatically degrade not only the chip yield, but also the classification accuracy of the neural-networks running on the RRAM crossbar. Existing hardware-based solutions cause enormous overhead and power consumption, while software-based solutions are less efficient in tolerating stuck-at faults and large variations. In this paper, we propose an accelerator-friendly neural-network training method, by leveraging the inherent self-healing capability of the neural-network, to prevent the large-weight synapses from being mapped to the abnormal memristors based on the fault/variation distribution in the RRAM crossbar. Experimental results show the proposed method can pull the classification accuracy (10%-45% loss in previous works) up close to ideal level with ≤ 1% loss.

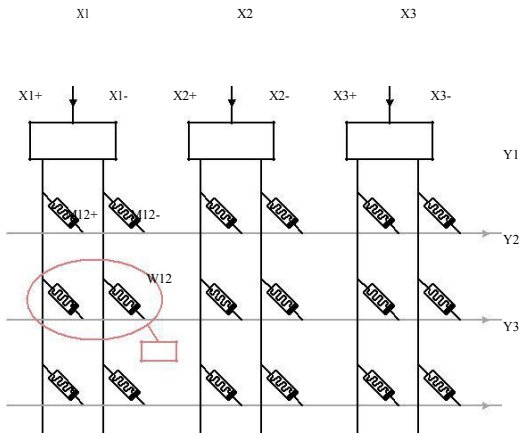


Fig. 1: The structure of a 1R RRAM crossbar.

摘要：

由忆阻器设备组成的 RRAM 交叉开关可以实现矩阵-向量乘法；

因此，它作为一种高度节能的神经-形态计算加速器，获得了巨大的发展动力。

然而，在记忆体器件中，电阻的变化和故障，不仅极大地降低了芯片的产量，而且也大大降低了在 RRAM 交叉开关上运行的神经网络的分类精度。

现有的基于硬件的解决方案带来了巨大的开销和电能消耗，而基于软件的解决方案在容忍故障和大变化方面效率较低。

本文提出了一种利用神经网络固有自愈能力的加速神经网络训练方法。实验结果表明该方法可以把分类精度(10%-45%的损失)到≤1%的损失，接近理想的水平。

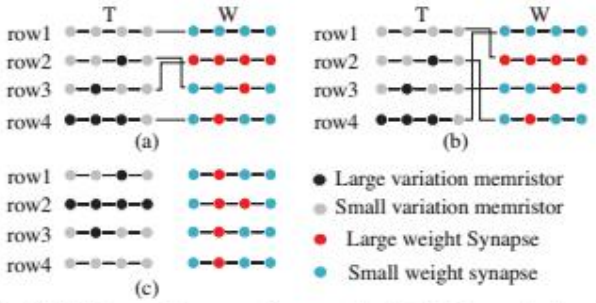


Fig. 2: Weight-memristor mapping examples that (a) the greedy-based method derives; (b) the bipartite-matching method derives; (c) the bipartite-matching method can hardly resolve.

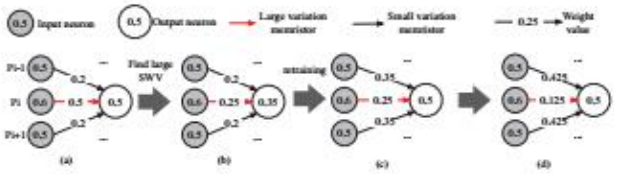


Fig. 3: Weight changing in the neural-network retraining method: (a) pre-trained weight; (b) fixing the weight connection; (c) after retraining; (d) in the next iteration.

This work is partly supported by the National Natural Science Foundation of China (Grant No. 61602300, 61202026 and 61332001), Shanghai Science and Technology Committee (Grant No. 15YF1406000), Program of China National 1000 Young Talent Plan, NSF CNS-1253424, and a funding from Lynmax Research. *Li Jiang is the corresponding author.

CNFET-Based High Throughput Register File Architecture

Tianjian Li[1], Li Jiang[1], Naifeng Jing[2], Nam Sung Kim[3] and Xiaoyao Liang[1]

[1]Department of Computer Science & Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Micro-Nano Electronics, Shanghai Jiao Tong University, Shanghai, China
[3]Department of Electrical and Computer Engineering, University of Illinois, Urbana-Champaign, USA. Email: {ltj2013, ljjang cs, sjtuj, liang-xy}@sjtu.edu.cn, nskim@illinois.edu

ABSTRACT—**A Carbon Nanotube field-effect transistor (CNFET) is a promising alternative to a traditional metal-oxide-semiconductor field-effect transistor (MOSFET) to overcome the "Power Wall" challenge. However, CNFETs are inherently subject to much larger process variation and thereby they can incur a significant design cost to build high-performance processors. Particularly, the large register files (RF) of SIMD GPU-style processors suffer more from such process variations because the number of critical paths are multiplied by the SIMD width and thread count. In this paper, we first show that RF organizations coupled with architectural techniques are critical to RF performance under CNFET-specific variations. Second, we propose several architectural techniques to mitigate the performance degradation, leveraging distinctive characteristics of CNFETs and unique features of SIMD processors. Our experiments demonstrate that the average RF performance is 53% higher than the worst design under variation and only 7% lower than the design with no variation.**

摘要：

碳纳米场效晶体管(CNFET)是一种很有发展前景的，可以替代传统金属-氧化物半导体场效应晶体管(MOSFET)来克服“能量墙”的挑战。

然而，CNFETs 本质上受制于更大的过程变化，因此它们可能需要大量的设计成本来构建高性能的处理器。

特别是，SIMD gpustyle 处理器的大型寄存器文件(RF)在这类过程中会受到更多的影响，因为关键路径的数量是由 SIMD 宽度和线程数相乘的。

在本文中，我们首先展示了射频组织与架构技术相结合，在特定的 cnfet 的变化下对射频性能至关重要。

其次，我们提出了一些架构技术来减轻性能退化，利用了 CNFETs 的特性和 SIMD 处理器的独特性能。

我们的实验表明，平均射频性能比最坏的设计要高 53%，比没有变化的设计要低 7%。

This work is partly supported by Shanghai Science and Technology Com-mittee (Grant No. 15YF1406000), the National Natural Science Foundation of China (Grant No. 61602300, No. 61202026 and No. 61332001), U.S. NSF grant (CNS-1217102), and Program of China National 1000 Young Talent Plan.

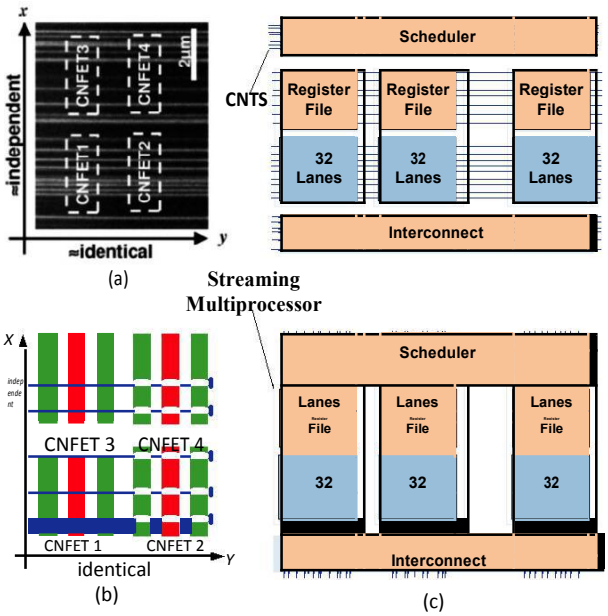


Fig. 1. Asymmetrically-correlated CNT density variation in CNFETs circuits: (a) SEM image [7]; (b) Schematic in circuit level; (c) A conceptual view of a GPU-like SIMD processor with different CNT growth direction.

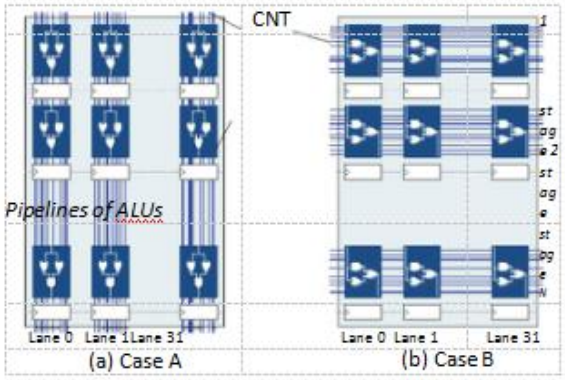


Fig. 3. The impact of CNT growth over the layout of the ALU when stages are (a) along or (b) perpendicular to the CNT growth direction.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCAD.2017.2695899, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems

CNFET-Based High Throughput SIMD Architecture

Li Jiang, Member, IEEE, Tianjian Li, Naifeng Jing, Nam Sung Kim, Fellow, IEEE, Minyi Guo, Senior Member, IEEE, Xiaoyao Liang, Member, IEEE

Abstract—Carbon Nanotube Field Effect Transistor (CNFET), using the carbon nanotubes (CNTs) as the material for conducting, is a promising alternative of CMOS technology to overcome the "power wall" issue. Recently, a microprocessor solely based on CNFETs was fabricated and demonstrated, which is a big step forward to the industrial practice. However, CNFETs are inherently subject to much larger process variation or manufacturing defects; thereby it may cause significant design cost to build high performance processors. This is exacerbated in the large register file (RF) architectures widely used in SIMD architectures, e.g., GPU style processors, where the number of critical paths are multiplied by the SIMD width and thread count. In this paper, we seek cost-effective approaches to address the issues by judiciously exploiting the strong asymmetric spatial correlation in the variation unique to the CNFET fabrication process. This paper presents a microarchitectural model to characterize CNFET delay variation and malfunction, under which we show that the RF organizations coupled with the architectural schemes are critical to the performance and power consumption of the SIMD processor. Therefore, we propose several architectural techniques to mitigate the performance degradation and the impact of CNT metallization, leveraging the distinctive CNFET characteristics and the unique features in the SIMD processors. Experimental results verify the effectiveness of the proposed techniques and demonstrate the great opportunity offered by this new device technology.

Index Terms—CNFET, Asymmetric Spatial Correlation, SIMD Processor, Register File Architecture

摘要:碳纳米晶体管(简称 CNFET)是 CMOS 技术克服"电力墙"问题的一种很有前途的替代方法。

最近，一种基于 CNFETs 的微处理器被制造和演示出来了，这是工业实践的一大步。

然而，CNFETs 天生就受制于更大的过程变化或制造缺陷;

因此，构建高性能处理器可能会造成重大的设计成本。

在 SIMD 体系结构中广泛使用的大型寄存器文件(RF)体系结构中，这种情况更加严重，例如 GPU 风格处理器，其中关键路径的数量是由 SIMD 宽度和线程数相乘的。

在本文中，我们寻求成本效益的方法来解决这些问题，并明智地利用 CNFET 制造过程中独特的变化的强非对称空间相关性。

本文提出了一种微结构模型来描述 CNFET 延迟变化和故障，在此基础上，我们证明了射频组织与建筑方案的耦合对于 SIMD 处理器的性能和功耗是至关重要的。

因此，我们提出了利用一些体系技术来减轻性能退化和 CNT 金属化的影响，利用独特的 CNFET 特征和 SIMD 处理器的独特特性。

实验结果验证了该方法的有效性。

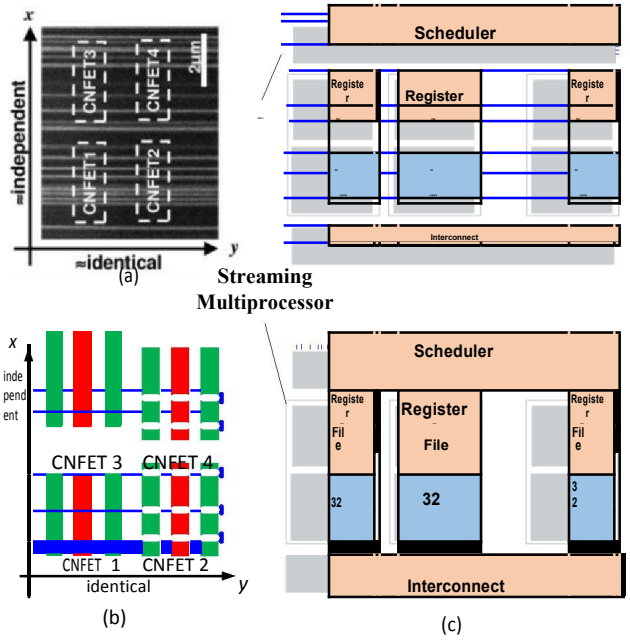


Fig. 1. Asymmetrically-correlated CNT density variation in CNFETs circuits: (a) SEM image [7]; (b) Schematic in circuit level; (c) A conceptual view of a GPU-like SIMD processor with different CNT growth direction.

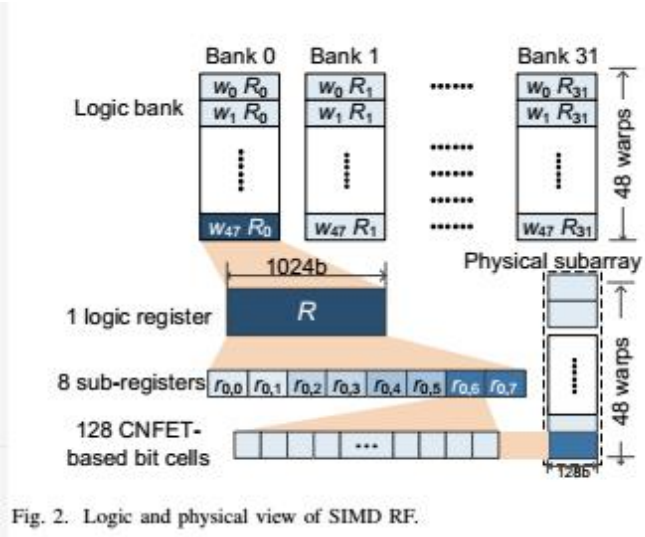


Fig. 2. Logic and physical view of SIMD RF.

Defect Tolerance for CNFET-based SRAMs

Tianjian Li[1], Li Jiang[1], Xiaoyao Liang[1], Qiang Xu[2] and Krishnendu Chakrabarty[3]

[1]Department of Computer Science & Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]Department of Computer Science & Engineering, The Chinese University of Hong Kong, Shartin, N.T., Hong Kong.
[3]Department of Electrical and Computer Engineering, Duke University, Durham, NC.

ABSTRACT—SRAMs based on carbon nanotube field-effect transistors (CNFETs) offer a promising alternative to conventional SRAMs due to their high energy efficiency and low leakage. However, the imperfect CNT fabrication process introduces high defect rates and a unique defect distribution; these problems may offset the power/performance benefits of CNFET-based SRAMs and lead to yield degradation. We propose a redundancy architecture with asymmetrically partitioned column blocks and the sharing of spares among column blocks. We also present a analytical model to characterize the distribution of faults, which can guide the design exploration of the proposed redundancy architecture. Simulation results highlight the accuracy of the proposed model, as well as the efficiency and effectiveness of the redundancy architecture.

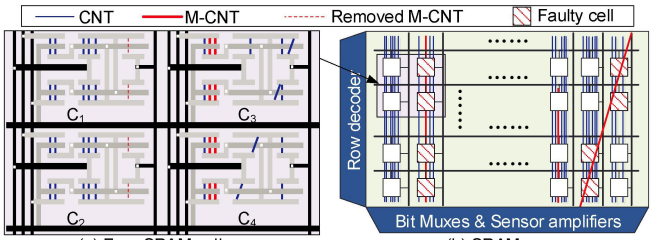


Fig. 1: CNFET-based SRAM layout.

摘要：

基于碳纳米场效应晶体管(CNFETs)的 SRAMs，由于其高的能量效率和低的泄漏率，为传统的 sram 提供了一种很好的替代选择。

然而，不完善的 CNT 制造过程引入了高缺陷率和独特的缺陷分布；

这些问题可能会抵消基于 cnfet 的 SRAMs 的性能和性能优势，并导致产量下降。

我们提出了一种冗余架构，它具有不对称的分块的列块，以及在列块之间共享备件。

我们还提出了一种分析故障分布特征的分析模型，可以指导冗余结构设计。

仿真结果表明了该模型的正确性，以及冗余架构的有效性和有效性。

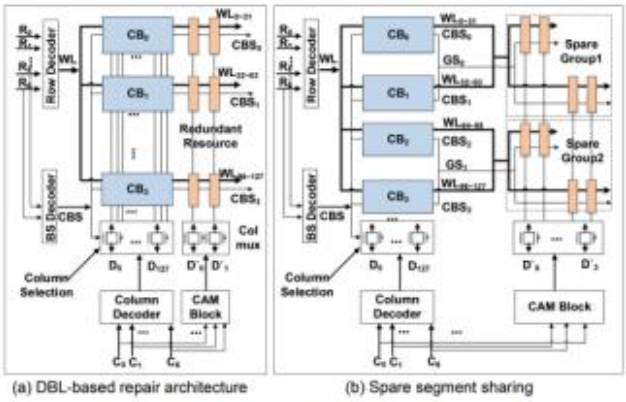


Fig. 5: Proposed redundancy architecture.

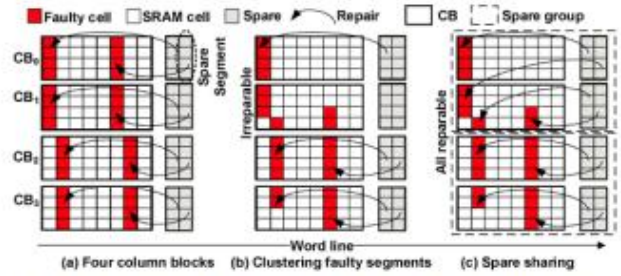


Fig. 4: A conceptional example of different repairing mechanisms.

978-1-4673-8773-6/16/$31.00

Fault Clustering Technique for 3D Memory BISR

Tianjian Li[1], Yan Han[1], Xiaoyao Liang[1], Hsien-Hsin S. Lee[2] and Li Jiang[1]*
[1] Department of Computer Science & Engineering, Shanghai Jiao Tong University
[2]Taiwan Semiconductor Manufacturing Company, Ltd.
Email: {ltj2013, hy123321, liang-xy, ljiang_cs}@sjtu.edu.cn, hhleeq@tsmc.com

Abstract—Three Dimensional (3D) memory has gained a great momen- tum because of its large storage capacity, bandwidth and etc. A critical challenge for 3D memory is the significant yield loss due to the disruptive integration process: any memory die that cannot be successfully repaired leads to the failure of the whole stack. The repair ratio of each die must be as high as possible to guarantee the overall yield. Existing memory repair methods, however, follow the traditional way of using redundancies: a redundant row/column replaces a row/column containing few or even one faulty cell. We propose a novel technique specifically in 3D memory that can overcome this limitation. It can cluster faulty cells across layers to the same row/column in the same memory array so that each redundant row/column can repair more "faults". Moreover, it can be applied to the existing repair algorithms. We design the BIST and BISR modules to implement the proposed repair technique. Experimental results show more than 71% enhancement of the repair ratio over the global 3D GESP solution and 80% redundancy-cost reduction, respectively.

摘要：
三维(3 d)内存获得非常好的增长势头,因为它存储容量大,带宽等。3 d内存的面临的一个关键挑战是重要的产量损失破坏性了一体化的进程;任何内存死,不能成功地修复,导致整体失败。每个模具的修复率必须尽可能高，以保证总收率。然而，现有的内存修复方法遵循了使用冗余的传统方法:冗余的行/列取代了包含很少甚至一个错误单元的行/列。我们提出了一种新的技术，特别是在三维内存中，可以克服这一局限性。它可以在同一内存数组中，将有缺陷的单元跨层排列到同一行/列，这样每个冗余行/列可以修复更多的“故障”。此外，它还可以应用于现有的修复算法。我们设计了BIST和BISR模块来实现提出的这种修复技术。实验结果表明，在全球3D GESP解决方法中，这个方法的修复率提高了71%以上，减少了80%的冗余成本。

This work is partly supported by the National Natural Science Foundation of China (Grant No. 61602300, Grant No. 61202026 and No. 61332001), Shanghai Science and Technology Committee (Grant No. 15YF1406000), and Program of China National 1000 Young Talent Plan. *Li Jiang is the corresponding author.

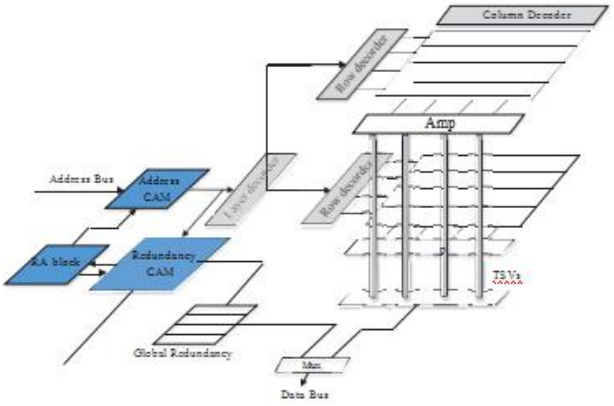


Fig. 3: Overview of 3D global architecture.

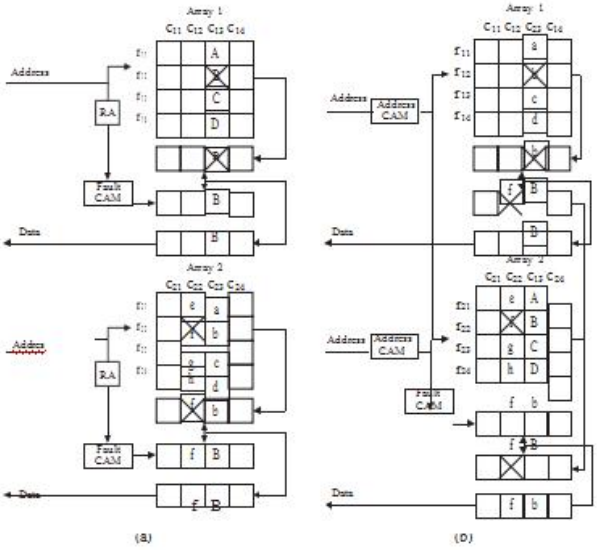


Fig. 4: The principle of fault clustering.

978-3-9815370-9-3/DATE17/©2017 EDAA

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPDS.2017.2773516, IEEE Transactions on Parallel and Distributed Systems

IBOM: An Integrated and Balanced On-Chip Memory for High Performance GPGPUs

Jianfei Wang, Qin Wang, Li Jiang, Chao Li, Xiaoyao Liang, Naifeng Jing

Abstract—GPGPU accelerated computing has revolutionized a broad range of applications. To serve between the ever-growing computing capability and external memory, the on-chip memory is becoming increasingly important to GPGPU performance for general-purpose computing. Inherited from the traditional CPUs, however, the contemporary GPGPU on-chip memory design is suboptimal to the SIMT (single instruction, multiple threads) execution. In particular, the on-chip first-level data (L1D) cache thrashing, resulting from insufficient capacity and imbalanced usage, leads to a low hit rate and limits the overall performance.
In this study, we reform the contemporary on-chip memory design and propose an integrated and balanced on-chip memory (IBOM) architecture for high-performance GPGPUs. It first virtually enlarges the L1D cache size by an integrated architecture that exploits the under-utilized register file (RF) with lightweight ISA, compiler and microarchitecture supports. Then with sufficient capacity, it is able to improve the cache usage by a set balancing technique that exploits the under-utilized set resources. In our proposed IBOM design, the register and cache accesses are amenable to normal pipeline operations with simple changes. It adequately exploits the size inversion in GPGPU on-chip memory, and enables optimized utilization of the precious resources for higher performance and energy efficiency with even smaller on-chip memory size. The experiment results demonstrate that the proposed IBOM design can offer an average of 29.6% increase in L1D hit rate and in turn 3X performance improvement for the cache-sensitive applications.

Index Terms—GPGPU, Cache Thrashing, Register File, Integrated Memory, Set Balancing, Compiler, High Performance.

摘要：GPGPU 加速计算革新了广泛的应用。为了在不断增长的计算能力和外部存储器之间服务，芯片内存对通用计算的 GPGPU 性能变得越来越重要。然而，从传统的 cpu 继承下来，当前的 GPGPU 芯片内存设计对 SIMT(单指令、多线程)执行是次优的。特别是，芯片上的一级数据(L1D)缓存由于容量不足和不平衡的使用而导致了低命中率，并限制了整体性能。在本研究中，我们对当代的芯片内存设计进行了改革，并提出了一种集成的、平衡的芯片内存(IBOM)架构，用于高性能的 gpgpu。它首先通过一个集成的体系结构将 L1D 缓存大小扩大，利用轻量级 ISA、编译器和微架构支持的未充分利用的寄存器文件(RF)。然后，有了足够的容量，它就能够通过一种集合资源的设置平衡技术来提高缓存的使用。在我们所提议的 IBOM 设计中，寄存器和缓存访问可以通过简单的更改来满足正常的管道操作。它充分利用了 GPGPU 芯片内存中的大小倒置，并利用更小的芯片内存，优化了资源的利用率，提高了性能和能源效率。实验结果表明，IBOM 的设计可以使 L1D 命中率平均提高 29.6%，对缓存敏感的应用程序的性能提高 3 倍。

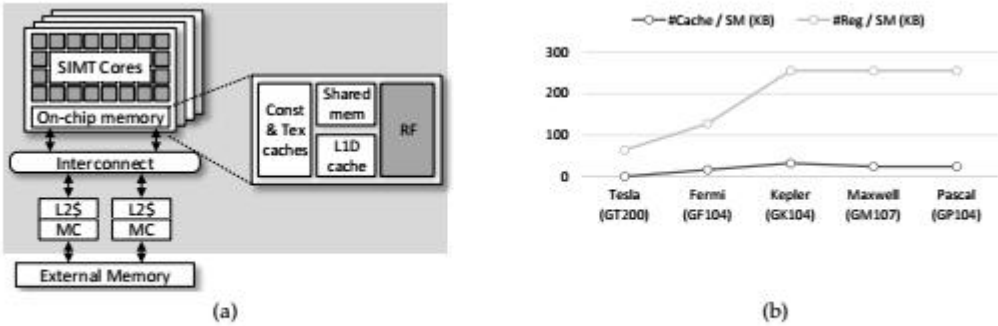


Fig. 1: (a) General GPGPU architecture and the on-chip memory, (b) the trend of RF and L1D capacity per SM.

In-growth Test for Monolithic 3D Integrated SRAM

Pu Pang[1], Yixun Zhang[1], Tianjian Li[1], Sung Kyu Lim[2], Quan Chen[1], Xiaoyao Liang[1] and Li Jiang[1]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
[2]School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, Georgia, USA

Abstract—Monolithic three-dimensional integration (M3I) directly fabricates tiers of integrated circuits upon each other and provides millions of vertical interconnections with inter-layer vias (ILVs). It thus brings higher integration density and communication capability compared with three-dimensional stacked integration (3D-SI). However, the Known-Good-Die prob-lem haunting 3D-SI–a faulty tier causes the failure of the entire stack–also occurs in M3I. Lack of efficient test methodologies such as the pre-bond testing in 3D-SI, M3I may have a more significant yield drop and thus its cost may be unacceptable for main-stream adoption. This paper introduces a novel In-growth test method for M3I SRAM. We propose a novel Design-for-Test (DfT) methodology to enable the proposed In-growth test on cell-level partitioned incomplete SRAM cells. We also build a statistical model of cost and discover a prospective judgement to determine whether or not to stop the fabrication, in order to prevent from raising the cost of fabricating more tiers upon the irreparable tiers. We find that a “sweet point” exists in the judgement, which can minimize the overall cost. Experimental results show the effectiveness of our proposed test methodology.

摘要：
单体三维集成电路（M3I）直接在已有集成电路层上建造集成电路层，并提供数百万与层间过孔（ILV）的垂直互连。因此它具有比三维堆叠集成（3D-SI）更高的集成密度和通信能力。然而，困扰 3D-SI 的已知良品问题（一个错误的电路层导致整个芯片的故障）也发生在 M3I 中。缺乏有效的测试方法，例如 3D-SI 中的预结合测试，M3I 可能会有较大的产量下降，因此它的成本不被主流所接受。本文介绍了一种用于 M3I SRAM 的新型生长中测试方法。我们提出了一种新颖的设计测试（DfT）方法，以实现针对单元级分区不完整 SRAM 单元的生长中测试。我们还建立了一个成本统计模型，并提出了一个预期判断，以确定是否停止制造，以防止在不可修复的电路层上制造更多层级的成本。我们发现判断中存在一个“甜点”，可以将整体成本降至最低。实验结果显示了我们提出的测试方法的有效性。

This research was partially supported by National Natural Science Foundation of China (Grant No. 61602300) and Shanghai Science and Technology Committee (Grant No. 15YF1406000). Corresponding author is Li Jiang.

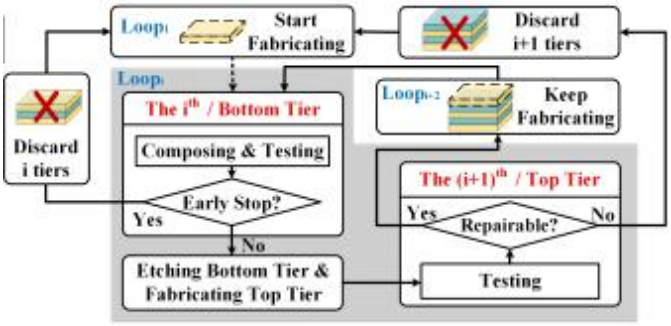


Fig. 1: The flow of the fabrication process integrated with In-growth test.

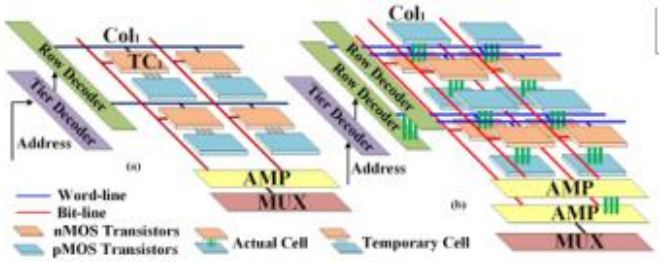


Fig. 2: The DfT method of incomplete cells. (a) Temporary cells and the layout of the bottom tier. (b) Actual cells and the layout oftwo tiers.

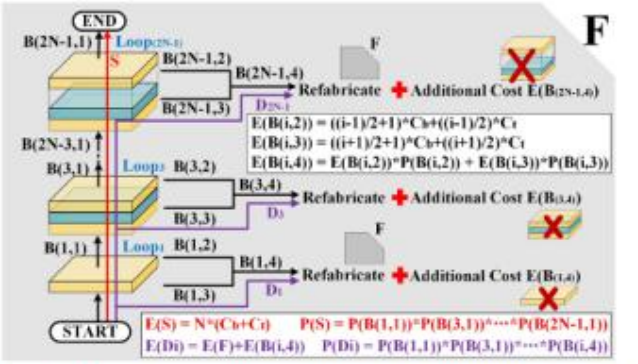


Fig. 3: The diagram of the 2N-tier M3I SRAM fabrication process integrated with In-growth test.

Jump Test for Metallic CNTs in CNFET-Based SRAM

Feng Xie[y], Xiaoyao Liang[y], Qiang Xu[z], Krishnendu Chakrabarty[x], Naifeng Jing[y] and Li Jiang[yr]

[y] Dept. CS&E, Shanghai Jiao Tong University, Shanghai, China
[z] Department of CS&E, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
[x]Deptartment of ECE, Duke University, Durham, NC
[r] State Key Laboratory of High-end Server & Storage Technology, China

ABSTRACT

SRAMs built on Carbon Nanotube Field Transistors (CN-FET) are promising alternatives to conventional CMOS-based SRAMs, due to their advantages in terms of both power con-sumption and noise margin. However, non-ideal Carbon Nan-otube (CNT) fabrication process generates metallic-CNTs (m-CNTs) along with semiconductor-CNTs (s-CNTs), rendering correlated faulty cells along the growth direction of the m-CNTs. Based on this phenomenon, we propose a novel testing algorithm for detecting m-CNTs, wherein consecutive write and read operations jump over multiple cells rather than march-ing through each and every cell, thereby signi cantly reducing the testing cost. The proposed jump test can be invoked before the march test to screen out those CNFET-SRAMs doomed to failure, and this can reduce the subsequent test overhead. Experimental results show that the proposed solution is able to achieve a high fault coverage with much less testing cost.

摘要：

基于碳纳米场晶体管(CN-FET)上的 SRAMs，由于其在功率消耗和噪声边界方面的优势，可以成为传统的基于 cmos 的 SRAMs 的替代方案，然而，非理想碳纳米管(CNT)的制造过程与半导体-碳纳米管(s-CNTs)一起产生金属碳纳米管(m-CNTs)，在 m-CNTs 的生长方向上呈现出相关的缺陷。基于这一现象，我们提出了一种新的 m-CNTs 检测算法，其中连续写入和读取操作跳过多个单元，而不是通过每个单元进行标记，从而显著降低了测试成本。在行进式测试之前，可以调用建议的跳跃式测试，以筛选出那些注定要失败的 CNFET-SRAMs，这可以减少后续的测试开销。实验结果表明，所提出的解决方案能够以较低的测试成本实现较高的故障覆盖率。

DAC'15, June 07-11 2015, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2744769.2744864.

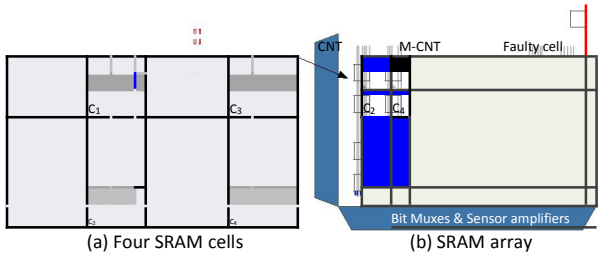


Figure 1: CNFET-based SRAM layout.

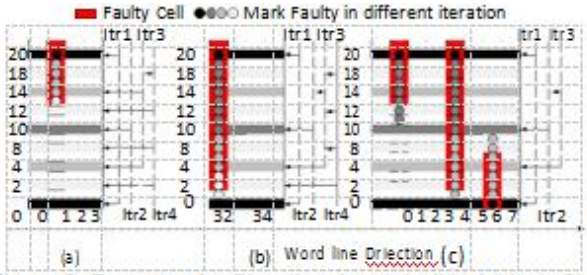


Figure 5: Examples for Bisection Jump: (a)(b) single m-CNT in a word range; (c) double m-CNTs in a word range

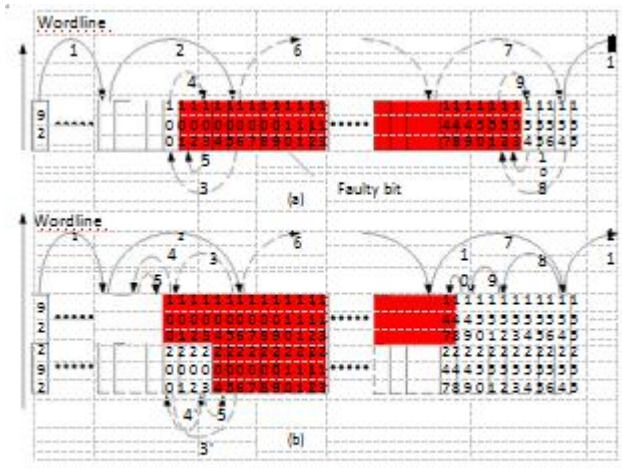


Figure 4: Examples illustrating recursive jump, word-line is in vertical direction: (a) one column of cells; (b) two columns of cells.

On Microarchitectural Modeling for CNFET-based Circuits

Tianjian Li, Hao Chen, Weikang Qian, Xiaoyao Liang and Li Jiang
Shanghai Jiao Tong University, Shanghai, China
Email: {ltj2013, ljiang cs}@sjtu.edu.cn

***Abstract*—**Carbon Nanotube Field-Effect-Transistors (CNFETs) show great promise to be an alternative to traditional CMOS technology, due to their extremely high energy efficiency. Unfortunately, the lack of control over the Carbon NanoTube (CNT) growth process causes CNFET circuits to suffer from the CNT count variation, which degrades the CNFET circuit performance. Compared to the CMOS process variation, the CNT count variation exhibits asymmetric spatial correlation. In this work, we propose an analytic model that integrates the impact of the asymmetric spatial correlation into the key microarchitectural blocks. We use this model to evaluate the variations in circuit performance for different layout styles and microarchitectural parameters. We further explore the opportunity of leveraging the asymmetric spatial correlation for performance enhancement. Experimental results based on SPICE simulation and architec-tural simulations showed the accuracy and effectiveness of the proposed model.

摘要：

碳纳米管场效应晶体管(CN-FETs)由于其极高的能源效率，极有可能成为传统 CMOS 技术的替代品。然而，碳纳米管(CNT)生长过程的缺乏控制使得 CNFET 电路受到 CNT 计数变化的影响，这降低了 CNFET 电路的性能。与 CMOS 工艺变化相比，CNT 计数变化呈现出不对称的空间相关性。在这一工作中，我们提出了一种分析模型，将非对称空间关联的影响整合到关键的微架构块中。我们用这个模型来评估不同布局样式和微观结构参数在电路性能上的变化。我们进一步探讨利用非对称空间相关性进行性能增强的机会。基于 SPICE 模拟和结构模拟的实验结果表明了该模型的准确性和有效性。

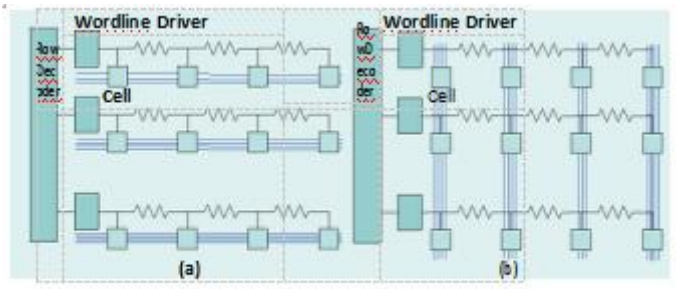


Fig. 3. Circuit of Wordline:(a) wordline parallel with CNT growth (b)wordline perpendicular to CNT growth

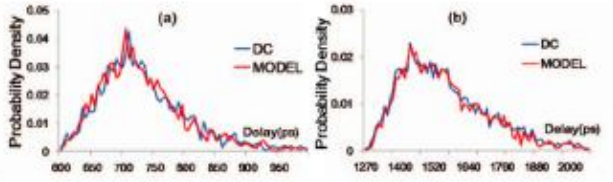


Fig. 4. Validating the logic model: (a) 16-bit multiplier and (b) M1 CPU

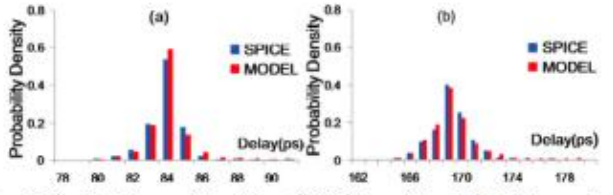


Fig. 5. Result of the wordline delay: (a) 64-bit wordline; (b) 128-bit wordline

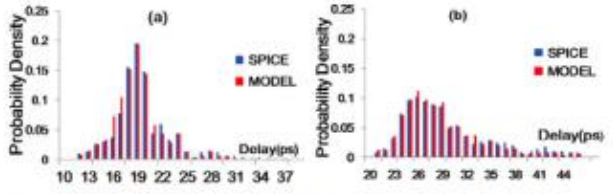


Fig. 6. Result of the bitline delay: (a) 64-bit bitline; (b) 128-bit bitline

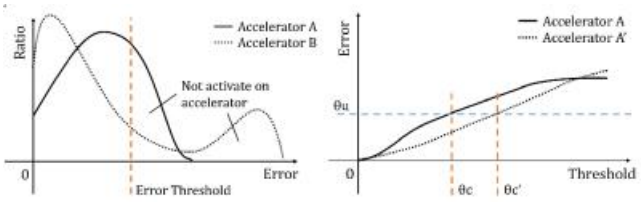
On Quality Trade-off Control for Approximate Computing Using Iterative Training

Chengwen Xu[1], Xiangyu Wu[1], Wenqi Yin[3], Qiang Xu[5], Naifeng Jing[3], Xiaoyao Liang[1, 2, 4] and Li Jiang[1, 2, 4, 6 * 1]
Dept. of CSE, [2] Brain Science and Technology Research Center, [3] Dept. of MICRO/NANO Electronics
[4]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering Shanghai Jiao Tong University
[5] Dept. of CSE, The Chinese University of Hong Kong; [6] Lynmax Research

ABSTRACT

Quality control plays a key role in approximate computing to save the energy and guarantee that the quality of the computation outcome satisfies users' requirement. Previous works proposed a hybrid architecture, composed of a classier for error prediction and an approximate accelerator for approximate computing using well trained neural-networks. Only inputs predicted to meet the quality are executed by the accelerator. However, the design of this hybrid architecture, relying on one-pass training process, has not been fully explored. In this paper, we propose a novel optimization framework. It advocates an iteratively training process to coordinate the train-ing of the classier and the accelerator with a judicious selection of training data. It integrates a dynamic threshold tuning algorithm to maximize the invocation of the accelerator (i.e., energy-efficiency) under the quality requirement. At last, we propose an efficient algorithm to explore the topologies of the accelerator and the classier comprehensively. Experimental results shows significant improvement on the quality and the energy-efficiency compared to the conventional one-pass training method.

摘要：质量控制在近似计算中起着关键的作用，以节约能源，保证计算的质量并满足用户的需求。之前的作品提出了一种混合架构，由一个用于错误预测的应用程序，以及一个使用训练有素的神经网络的近似计算的近似加速器组成。只有预测符合质量的输入才会被加速器执行。然而，这种混合结构的设计，依赖于一遍通过的训练过程，还没有得到充分的探索。本文提出了一种新的优化框架。它提倡一种迭代的训练过程，通过明智的选择训练数据来协调对加速器的训练。它集成了一个动态阈值调优算法，以最大限度地调用加速器（例如在质量要求下，能源效率）。最后，我们提出了一种有效的算法，以全面地探索加速器的拓扑结构。实验结果表明，与传统的单传训练方法相比，在质量和能源节省方面有了显著的改进。



(a) Data distribution ratio on the (b) Approximation error VS. clas- approximation error. si er threshold.

Figure 1: A conceptual example.

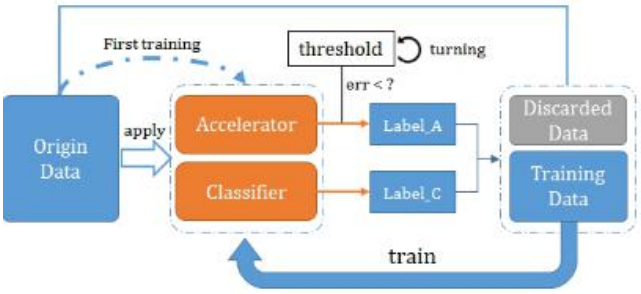


Figure 2: The process of iterative training.

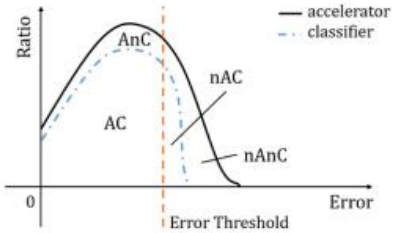


Figure 3: Distribution of the four categories of the original data derived from the accelerator and the classier.

* The corresponding author is Li Jiang, jiangli@cs.sjtu.edu.cn

DAC'17, Austin, TX, USA

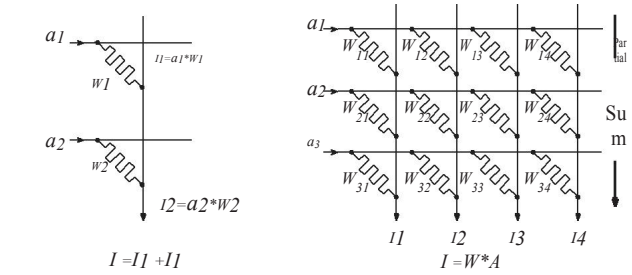
DOI: http://dx.doi.org/10.1145/3061639.3062294

RECOM: An Efficient Resistive Accelerator for Compressed Deep Neural Networks

Houxiang Ji †, Linghao Song‡, Li Jiang†§, Hai(Halen) Li‡ and Yiran Chen‡§
Zhiyuan College, Shanghai Jiao Tong University, Shanghai, China
†Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
‡Department of Electrical and Computer Engineering, Duke University, Durham NC, U.S.A Email: {jihouxiang, ljiang cs}@sjtu.edu.cn, {linghao.song, hai.li, yiran.chen}@duke.edu

Abstract— Deep Neural Networks (DNNs) play a key role in prevailing machine learning applications. Resistive random-access memory (ReRAM) is capable of both computation and storage, contributing to the acceleration on DNNs by processing in memory. Besides, a significant amount of zero weights is observed in DNNs, providing a space to reduce computation cost further by skipping ineffectual calculations associated with them. However, the irregular distribution of zero weights in DNNs makes it difficult for resistive accelerators to take advantage of the sparsity as expected efficiently, because of its high reliance on regular matrix-vector multiplication in ReRAM. In this work, we propose ReCom, the first resistive accelerator to support sparse DNN processing. ReCom is an efficient resistive accelerator for compressed deep neural networks, where DNN weights are structurally compressed to eliminate zero parameters and become hardware-friendly. Zero DNN activation is also considered at the same time. Two technologies, Structurally-compressed Weight Oriented Fetching (SWOF) and In-layer Pipeline for Memory and Computation (IPMC), are particularly proposed. In our evaluation, ReCom can achieve 3.37x speedup and 2.41x energy efficiency compared to a state-of-the-art resistive accelerator.

摘要：
深度神经网络（DNN）在现在普遍流行的机器学习应用中起着关键作用。阻变式随机访问存储器（ReRAM）可以同时处理计算和存储，通过内存内处理加速深度神经网络的处理。另外，研究人员发现在 DNN 中存在着显著数目的零值权值，通过跳过与之相关联的无用计算为进一步降低计算成本提供了探索的空间。然而，因为阻变性加速器对于规则化的矩阵向量乘法有着高度的依赖性，DNN 中零值权值的不规则分布使得阻变性加速器利用起来难以达到预期的高效率。在这篇文章中，我们提出了 ReCom，第一个支持稀疏 DNN 处理的忆阻性加速器。ReCom 是一种高效的压缩深度神经网络忆阻性加速器，在 ReCom 中 DNN 的权值被规则化压缩来消除零值参数并且整个权值矩阵变得更加硬件友好。零值的 DNN 输入值也同时被考虑了进来。我们着重提出了两种技术 Structurally-compressed Weight Oriented Fetching（SWOF）和 In-layer Pipeline for Memory and Computation（IPMC）。在测试中，和最前沿的忆阻性加速器相比，ReCom 能达到 3.37 倍的加速和 2.41 倍的能量节省。



(a) Multiply and Accumulation (b) Matrix-Vector Multiplication

Fig. 1. Matrix-Vector Multiplication on ReRAM

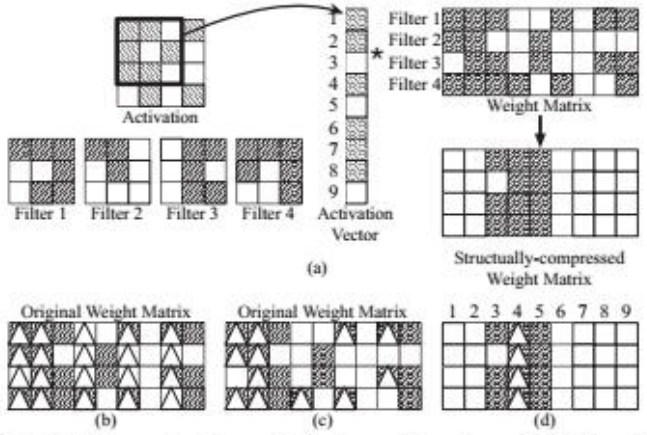


Fig. 2. (a) Conversion of a convolution to a matrix-vector multiplication and comparison of non-zeros (triangles) in computation between (b) [8], (c) [9] and (d) RECOM

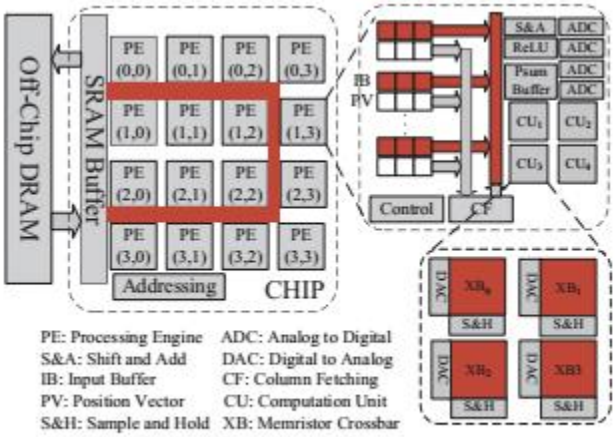


Fig. 3. RECOM Top-level Architecture

Sneak-Path Based Test and Diagnosis for 1 R RRAM Crossbar Using Voltage Bias Technique

Tianjian Lit, Xiangyu Bit, NaifengJingt, Xiaoyao Liangt and Lijiangt§ •

t Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

ABSTRACT

Metal-oxide resistive random access memories with a single memristor device at the crosspoint (IR RRAM) is a promising alternative to next generation storage technology due to their high density, scalability, non-volatility and low power consumption. However, the imperfect fabrication process introduces high defect rates of the nanoscale memristor devices and leads to yield degradation. In addition, sneak-paths occur in lR RRAM crossbar that can jeaperdize the normal read/write operation. Previous work proposes voltage bias technique to eliminate the sneak-paths. Instead, in the paper, we leverage voltage bias to manipulate various distribution of sneak-paths that can screen one or multiple faults out of a 4 x 4 region of memristors at once, and consequently diagnose the exact location of each faulty memristor within three write-read operations. The SPICE simulation results highlight the effectiveness and efficiency of the proposed test method.

摘要：阻变式存储器（RRAM）有希望成为下一代存储器技术，因为它存在存储密度高、非易失性以及低功耗等特性。然而，不完备的制造工艺引起很高的故障率，进而降低存储器的良品率。另外，存储器阵列中的旁路电流会影响阵列的正常读写操作。其他研究人员提出了添加偏置电压的方法去抑制旁路电流的影响。在这篇论文中，我们使用偏置电压控制旁路电流去得到测试区域中的一个或者多个故障单元。SPICE 实验仿真结果验证了本文测试算法的效率和准确性。

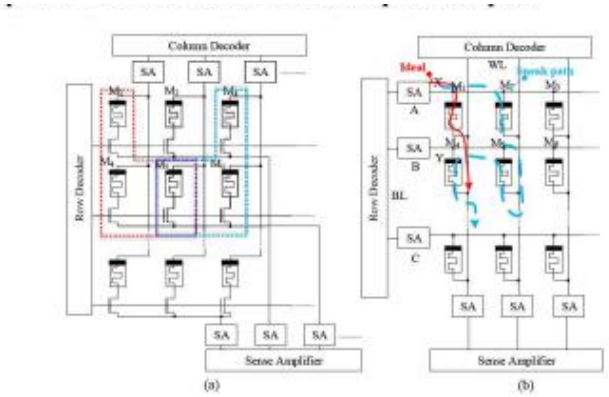


Figure 1: (a) An example 1T1R RRAM crossbar; (b) An example 1R RRAM crossbar and the example sneak-path.

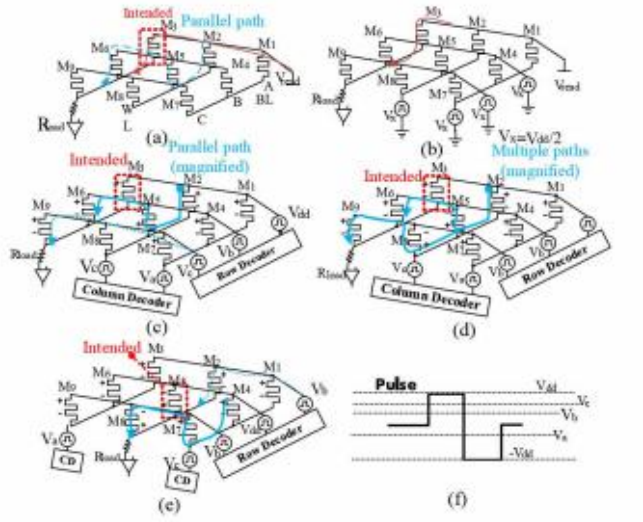


Figure 2: An example for sneak-path control: (a) A read operation affected by sneak-path; (b) Sneak-path elimination with an uniform level of voltage bias; (c) one and (d) two magnified sneak-paths; (e) Another example of magnified sneak-paths; (f) Patterns are programmed in the form of pulse with different voltage levels, distributed from -Vdd to Vdd.

'The corresponding author is Li Jiang, jiangli@cs.sjtu.edu.cn

DAC'17, Austin, TX, USA

DOI: http:/ /dx.doi.org/10.1145/3061639.3062318

Timing-Driven Placement for Carbon Nanotube Circuits

Chen Wang[1], Li Jiang[1], Shiyan Hu[2], Tianjian Li[1], Xiaoyao Liang[1], Naifeng Jing[1] and Weikang Qian[1]
[1] Shanghai Jiao Tong University, Shanghai, China
[2] Michigan Technological University, Houghton, Michigan, USA
Email: [1]{wangchen 2011, ljiang cs, ltj2013, liang-xy, sjtuj, qianwk}@sjtu.edu.cn, [2]shiyan@mtu.edu

***Abstract*—Carbon nanotube field effect transistors (CNFETs), which use carbon nanotubes (CNTs) as the transistor channel, are promising substitution of conventional CMOS technology. However, due to the stochastic assembly process of CNTs, the number of CNTs in each CNFET has a large variation, resulting in a vast circuit delay variation and timing yield degradation. To overcome it, we propose a timing-driven placement method for CNFET circuits. It exploits a unique feature of CNFET circuits, namely, asymmetric spatial correlation: CNFETs that lie along the CNT growth direction are highly correlated in terms of their electrical properties. Our method distributes CNFETs of the same critical paths to different rows perpendicular to the CNT growth direction during both global and detailed placement phases, while optimizing the timing of these critical paths. Experimental results demonstrated that our approach reduces both the mean and the variance of circuit delay, leading to an improvement in timing yield.**

摘要：

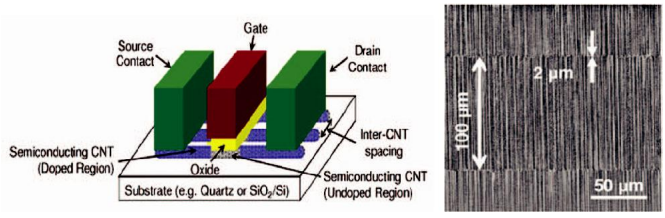
碳纳米管场效应晶体管(简称 CNFETs)利用碳纳米管作为晶体管通道，有望取代传统的 CMOS 技术。

然而，由于 CNTs 的随机装配过程，每个 CNFET 中 CNTs 的数量有很大的变化，导致了一个巨大的电路延迟变化和时间收益率下降。

为了克服这一问题，我们提出了一种基于时间驱动的 CNFET 电路布局方法。它利用了 CNFET 电路的一个独特的特性，即不对称的空间相关性。

我们的方法在全局和详细的放置阶段将相同关键路径的 CNFETs 分布到与 CNT 生长方向垂直的不同行，同时优化这些关键路径的时间。

实验结果表明，我们的方法降低了电路延迟的均值和方差，从而提高了时间收益率。



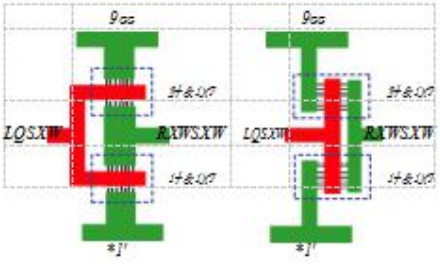


Fig. 2: Two standard cell layout styles with different CNT growth directions with regard to the Vdd/GND rails. On the left, the CNT growth direction is perpendicular to the Vdd/GND rails. On the right, the CNT growth direction is parallel to the Vdd/GND rails.

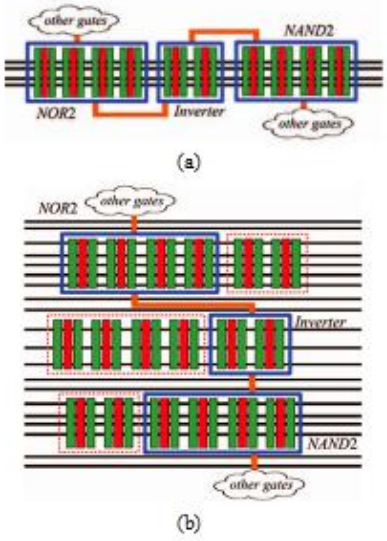


Fig. 3: Gates of the chosen critical path allocated to the same and the different CNT rows.