A 462GOPs/J RRAM-Based Nonvolatile Intelligent Processor for Energy Harvesting IoE System Featuring Nonvolatile Logics and Processing-In-Memory

Fang Su[1], Wei-Hao Chen[2], Lixue Xia[1], Chieh-Pu Lo[2], Tianqi Tang[1], Zhibo Wang[1], Kuo-Hsiang Hsu[2], Ming Cheng[1], Jun-Yi Li[2], Yuan Xie[3], Yu Wang[1], Meng-Fan Chang[2], Huazhong Yang[1], and Yongpan Liu[1]

[1]Tsinghua University, Beijing; [2]National Tsing Hua University, Hsinchu; [3]University of California, Santa Barbara

Abstract

An energy-efficient nonvolatile intelligent processor (NIP) is proposed for battery-less energy harvesting system. This NIP employs RRAM-based nonvolatile logics (NVL) with self-write-termination (SWT) scheme and low-power processing-in-memory (PIM) to achieve energy-efficient computing against frequent power-off situations. An NIP test chip was fabricated in 150nm CMOS process using HfO RRAM. This NIP chip achieves 462GOPs/J energy efficiency at 20MHz clock frequency, showing 13× performance improvement over state-of-the-arts. This work presents the first nonvolatile processor capable of general as well as neural network computing in addition to the first integrated chip using RRAM-based PIM.

Introduction

Nonvolatile processor [1-3] has been considered the key component of the Internet of Everything (IoE) edge devices powered by energy harvesting. Upon a power failure, it maintains system states at literally-zero leakage power. As the emerging IoE applications are seeking for more "intelligence" which commonly relies on data-intensive neural networks, nonvolatile processors are required to offer more advanced computing capability. However, considering the already constrained power budget in an energy harvesting scenario, the design should aim at the next order of magnitude improvement in energy efficiency.

This paper presents a nonvolatile intelligent processor (NIP) in 150nm CMOS process with embedded RRAM. Advanced features brought by RRAM are twofold exploited to maximize the energy efficiency: 1) RRAM with data-aware self-write-termination (SWT) is utilized to realize energy-efficient nonvolatile logics (NVL); and 2) the 1T1R-RRAM array enables processing-in-memory (PIM) for fully connected neural networks (FCNNs).

Design of Nonvolatile Intelligent Processor

Fig. 1 shows the NIP block diagram, consisting of an 8051 processor core (CPU), an FCNN-Turbo-Unit (FTU) with four 1T1R RRAM arrays, a 4Kb nonvolatile SRAM (nvSRAM), a power management unit (PMU) and peripheral circuits. The CPU performs general-purpose computation and hosts the communication with off-chip sensors or transceivers. The FTU handles FCNN tasks in recognition and classification applications. The nvSRAM acts as data memory shared by both CPU and FTU. The PMU provides V_{DD} and manages the backup and restore (B/R) decisions for the entire chip during a sudden power failure or an intended power-off. Fig. 2 shows the function of each module in a video surveillance application. The proposed NIP achieves ultra-high energy efficiency in energy harvesting scenarios through the following two key technologies.

A. Energy-efficient NVL. All 2189 flip-flops in the NIP are embedded with a couple of bipolar RRAM device [4] to enable data non-volatility. Fig. 3 shows the schematic of the proposed 2R-nvFF with data-aware SWT. The nvFF has three operation modes: normal, backup and restore. In normal mode, RSWL is set low to save power and avoid voltage/current stress on RRAM devices. Before power supply goes down, the PMU generates a low B/R signal to drive the nvFF into backup mode. The data held in the flip-flop is moved from Q/QB into RRAM devices RL/RR through resistance state switching. A high resistance state (HRS) stands for logic 1 and low resistance state (LRS) for logic 0. The SWT circuit detects the voltage on NX and Q and terminates the switching operation. This helps to reduce the per-bit backup energy and improve RRAM endurance in the case that the data to be backed up matches the current state of RRAM devices [2, 5]. For example, if Q=0 and RL is already in LRS, Q will be pulled high after the SET signal is applied, resulting in a SET-termination (Fig. 4). When power is resumed, the nvFF enters restore mode. According to RRAM state LRS/HRS, Q/QB becomes logic 0/1 so that the processor continues the interrupted operation. Fig. 4 shows the HfO-based bipolar RRAM structure, the operation conditions and captured waveforms of proposed 2R-nvFF with SWT.

B. Low-power PIM. Fig. 5 shows the energy and latency breakdown of a conventional FCNN accelerator [6]. Weight transfer between SRAM and accelerator occupies 62% of total energy consumption and 97% of processing latency. Recently, RRAM has shown great potential to break the "memory-wall", because the matrix-vector-multiplication (MVM) operations can be directly carried out with the weights stored in RRAM array as resistance values. It is also known as PIM [7]. However, conventional RRAM-based PIM architecture faces two challenges in an energy harvesting NIP (Fig. 6): 1) The interfaces between RRAM array and CPU, namely digital-to-analog (D/A) and analog-to-digital (A/D) conversion circuits, become the bottleneck of energy efficiency and chip area; 2) All access transistors are simultaneously turned on by V_{DD} on word-lines (WLs) to perform an MVM operation, resulting in large sneak currents and energy wastes.

To mitigate those problems, a low-power MVM engine is designed with binary interface and input-controlled access transistors. The binary input vector is directly connected to the WLs, and the outputs are obtained through 1-to-3-bit adaptive sense amplifiers (SAs) at the end of bit-lines (BLs). As Fig. 7 shows, the proposed structure has two benefits, especially for networks with binary weights and input/output. Firstly, the A/D and D/A overheads are eliminated, which brings 44% and 95% savings in energy and area, respectively. Secondly, the input-controlled access transistors remain OFF when the row input is zero. This yields 64% energy saving compared to conventional architecture with all transistors on.

Fig. 8 shows the high level architecture of the FTU, including four designed MVM engines of size 32×32. Two arrays share the same inputs and store positive and negative weights, and two FCNN layers can be mapped to the NIP simultaneously.

Fabrication and Measurement Results

The NIP is fabricated in 150nm CMOS process using HfO RRAM. Fig. 9 shows the microphotograph and specifications of the chip. The backup/restore operation of NVL (including nvFF and nvSRAM) consumes 58nJ/0.5nJ, and can be finished within two/one cycle(s). The proposed NIP achieves 462GOPs/J peak energy efficiency at 20MHz clock frequency with 3.3V supply voltage. A small footprint of 1.9×1.9 mm^2 makes it suitable for "smart dust" devices in IoE. Table I compares the performance of the NIP with prior work. It achieves 13× improvement in energy efficiency over state-of-the-arts. Thanks to the energy-efficient NVL and low-power PIM techniques, the NIP shows 50× and 17× savings on processing time and energy in a real-world application (Fig. 10).

References

[1] Y. Wang, *ESSCIRC*, 2012.
[2] Y. Liu, *ISSCC*, 2016.
[3] T. Onuki, *VLSIC*, 2016.
[4] S. S. Sheu, *ISSCC*, 2011.
[5] C. P. Lo, *IEDM*, 2016.
[6] J. Oh, *ISSCC*, 2011.
[7] P. Chi, *ISCA*, 2016.

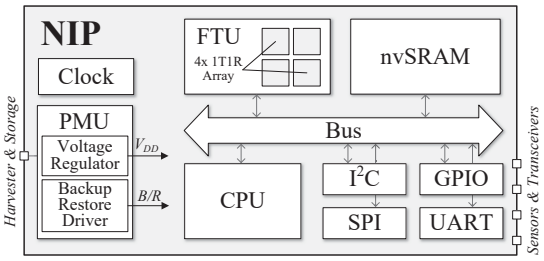

Fig. 1 Top-level architecture of proposed NIP.

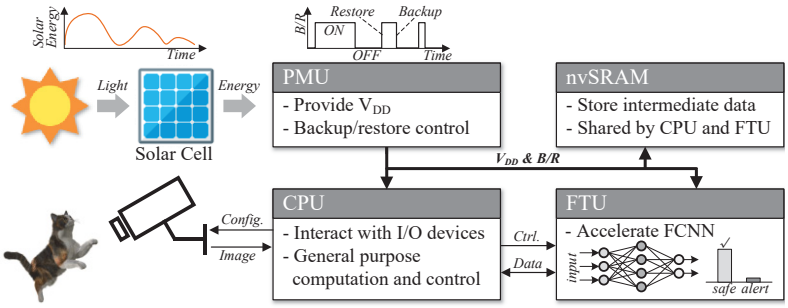

Fig. 2 Module functions in an energy harvesting video surveillance system.

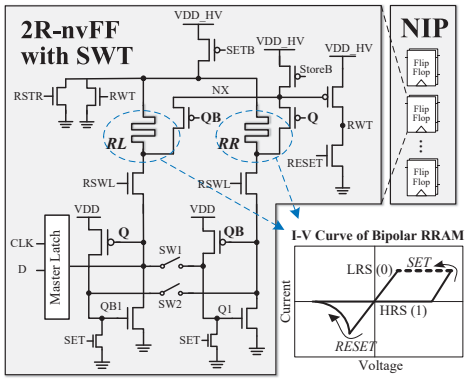

Fig. 3 Circuit schematic of proposed nvFF.

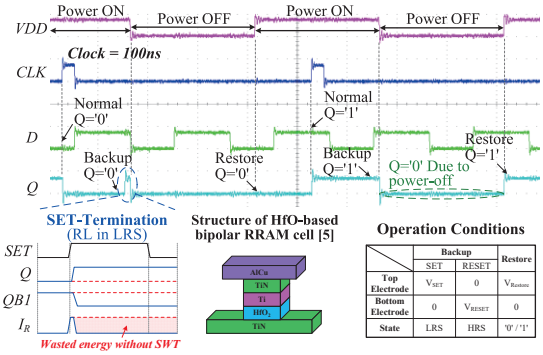

Fig. 4 Waveforms and operation conditions of nvFF.

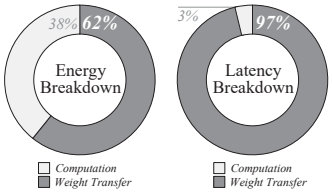

Fig. 5 Energy/latency breakdown of conventional FCNN accelerator where weight transfer from SRAM becomes the bottleneck.

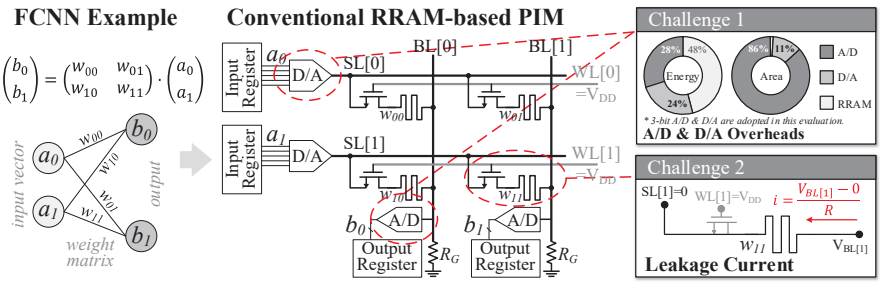

Fig. 6 Conventional RRAM-based PIM and its challenges.

Table I
Performance comparison with prior work

NVP	ISSCC 16 [2]	VLSI 16 [3]	This Work	Improvement
Process (nm)	65	65	150	-
Nonvolatile technology	RRAM	IGZO	RRAM	-
Area (mm²)	4.46	1.05	3.69	-
Efficiency (GOPs/J)	30.3	35.5	462.1	13.0–15.3×
Standby power (nW)	0	9.0	0	100%
Restore speed (clock cycle)	1	3	1	3×
On-chip FCNN layer	0	0	2	100%

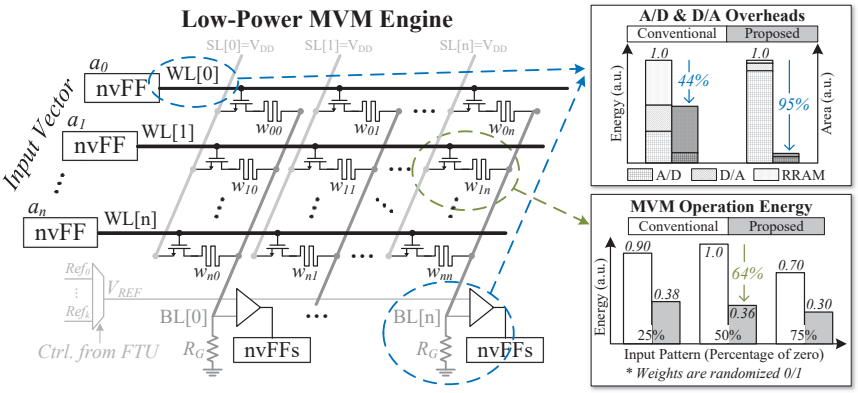

Fig. 7 Proposed low-power MVM engine and its performance improvements.

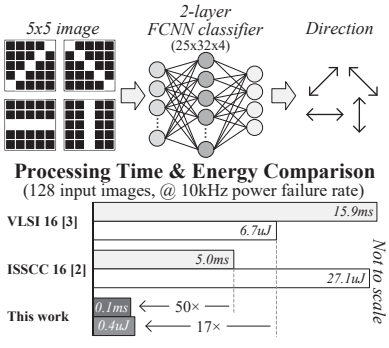

Fig. 10 Processing time and energy comparison in real-world application.

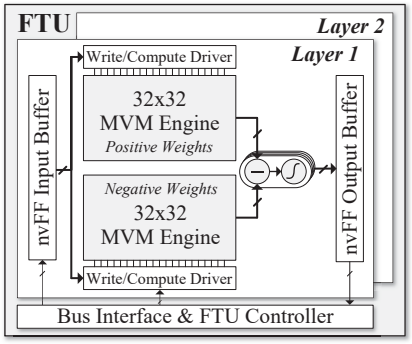

Fig. 8 High-level structure of proposed FTU.

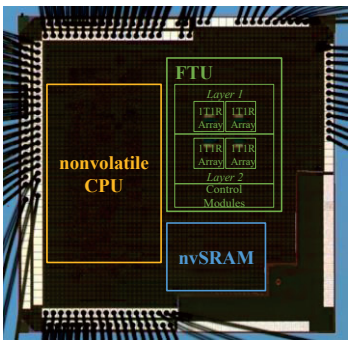

Fig. 9 Chip photograph and summary.

Process technology	HfO RRAM + 150nm CMOS
Area	1920um × 1920um
# of pins	206
Supply voltage	1.8V (core) / 3.3V (IO)
Clock frequency	20MHz
Power consumption	22.2mW
Energy efficiency	462.1GOPs/J
Backup/restore speed	2 cycles / 1 cycle
Backup/restore energy	57.96nJ / 0.51nJ
RRAM utilization	2189 nvFF 4Kb nvSRAM Four 32x32 MVM Engine
RRAM area	nvFF: 10.98% nvSRAM: 4.23% MVM Engine: 22.45%