A Low Active Leakage and High Reliability Phase Change Memory (PCM) Based Non-Volatile FPGA Storage Element

Kejie Huang, *Student Member, IEEE*, Yajun Ha, *Senior Member, IEEE*, Rong Zhao, *Member, IEEE*, Akash Kumar, *Senior Member, IEEE*, and Yong Lian, *Fellow, IEEE*

***Abstract*—The high leakage current has been one of the critical issues in SRAM-based Field Programmable Gate Arrays (FPGAs). In recent works, resistive non-volatile memories (NVMs) have been utilized to tackle the issue with their superior energy efficiency and fast power-on speed. Phase Change Memory (PCM) is one of the most promising resistive NVMs with the advantages of low cost, high density and high resistance ratio. However, most of the reported PCM-based FPGAs have significant active leakage power and reliability issues. This paper presents a low active leakage power and high reliability PCM based non-volatile SRAM (nvSRAM). The low active leakage power and high reliability are achieved by biasing PCM cells at 0 V during FPGA operation. Compared to the state-of-the-art, the proposed nvSRAM based 4-input look up table (LUT) achieves 174 times reduction in active leakage power and 15000 times increase in retention time. In addition, the proposed nvSRAM-based FPGA system significantly accelerates the loading speed to less than 1 ns with 2.54 fJ/cell loading energy.**

***Index Terms*—Active leakage, field programmable gate array (FPGA), low power, multi-context, non-volatile memory (NVM), non-volatile SRAM, phase change memory (PCM), read disturbance.**

I. Introduction

SRAM-BASED Field Programmable Gate Array (FPGA) logic circuits have been under focused development in the past 20 years [1]–[4]. However, they require reprogramming each time when powering on, because SRAMs lose the configuration information after powering down. Moreover, as CMOS technology nodes scale down to 90 nm and below, the leakage power has rapidly become the dominant component of total power dissipation [5], [6]. As a result, SRAM-based FPGAs suffer from slow power-on speed, high power-on power and leakage power. The high power-on power and slow power-on speed limit the power-off opportunities of the FPGA. In other words, it is not possible to power off the FPGA when the idle time between two events is short. Moreover, additional external non-volatile memory (NVM) is required to store the configuration information.

The emerging resistive NVM technologies with the advantages of high density, near zero power-on delay, and superior energy efficiency have provided an excellent platform to advance the FPGA technology. Phase Change Memory (PCM) [7]–[9], Resistive Random Access Memory (RRAM) [10], [11] and Spin-Torque Transfer Magnetoresistive RAM (STT-MRAM) [12], [13] are three candidates in the emerging NVM technologies. PCM could be a universal NVM [14] that provides the benefits of high density [15], high scalability [16], low cost [17] and high resistance ratio [18]. The $4F^2$ small PCM cell size based on 20 nm technology node has been achieved by Samsung [19]. The high resistance ratio between the amorphous $(RESET)$ and poly-crystalline (SET) states increases the read reliability. Moreover, PCM also has the potential to achieve nano-second [20] and sub micro-ampere current switch [21].

A few works have been reported to integrate NVM cells into FPGA circuits in [22]–[26]. However, those works have various drawbacks that limit their applications in FPGAs. For example, the designs in [22], [23] have a write reliability issue due to sneak paths. [24] in essence is the SRAM-based FPGA. Therefore, it still suffers from long configuration time and high configuration power when powering on. [25] and [26] suffer from high active leakage power (the leakage power during normal operation) and low reliability issues due to high dc voltage (VDD) on NVM cells during the FPGA normal operation. The detailed discussion of the related works will be provided in Section II-B.

In this paper, we propose a low active leakage power and high reliability non-volatile SRAM (nvSRAM) storage element with high loading speed. PCM is used in our nvSRAM, but it is worth noting that our nvSRAM cell can be extended to all resistive NVMs. To achieve the low active leakage power and high reliability, PCM cells are only sensed when powering on. In the FPGA operation mode, they are biased at 0 V by pulling both nodes of PCM cells to the ground. Therefore, there is no active leakage power in PCM cells, and the retention time can be greatly improved. As a result, our proposed nvSRAM is able to load configuration information within 1 ns, achieving fast multi-context switching abilities, and 41.8 pW low active leakage power during FPGA operation. The retention can be longer than 10 years. The FPGA system loading speed and energy are 1 ns and 2.54 fJ/cell, respectively.

Manuscript received September 14, 2013; revised January 11, 2014; accepted January 14, 2014. This paper was recommended by Associate Editor B. Amrutur.

K. Huang is with the Department of Electrical and Computer Engineering, National University of Singapore, 119260. He is also with the Department of Engineering Product Design, Singapore University of Technology and Design, Singapore, 138682 (e-mail: a0090886@nus.edu.sg).

Y. Ha, A. Kumar, and Y. Lian are with the Department of Electrical and Computer Engineering, National University of Singapore, 119260 (e-mail: {elehy@nus.edu.sg; eleak@nus.edu.sg; eleliany@nus.edu.sg).

R. Zhao is with the Department of Engineering Product Design, Singapore University of Technology and Design, Singapore, 138682 (e-mail: zhao_rong@sutd.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2014.2312499

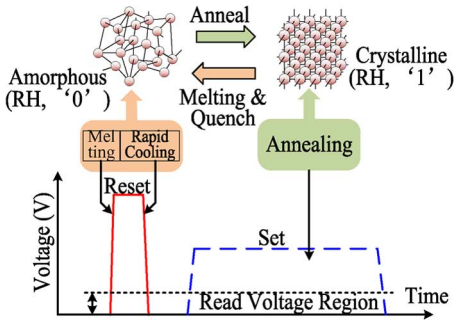


Fig. 1. Phase change materials reversibly switch between amorphous and poly-crystalline states by electrical pulses.

The paper is organized as follows: Section II introduces the background of PCM and the related works. Section III discusses our proposed nvSRAM based FPGAs, and its advantages. Section IV analyzes the proposed PCM based nvSRAM and its operation in different modes, and Section V provides the simulation results of the proposed nvSRAM. Finally, the conclusions are drawn in Section VI.

II. Background

A. Phase Change Memory

The typical PCM structure is a chalcogenide layer (i.e., Ge2Sb2Te5, or GST) sandwiched between a metal contact and a heat electrode. The heat produced by the passage of an electric current through the heating element is used to transform the material between the poly-crystalline and amorphous states. As shown in Fig. 1, if the chalcogenide material is quickly heated (melting) and quenched (rapid cooling), it will be reset to the amorphous state (high resistance state, R_H, binary '0'). On the other hand, if the material is held in its crystallization temperature range for some time (annealing), it will be set to the poly-crystalline state (low resistance state, R_L, binary '1'). The cell resistance between the poly-crystalline and amorphous states may have orders difference. Therefore, as shown in Fig. 1, $RESET$ (quickly heating and quenching) requires short pulse and high voltage, while SET (holding in crystallization temperature) requires long pulse and medium voltage. To avoid unintended write, the read voltage should be much lower than the SET voltage.

One of the important concerns to integrate the NVM in FPGAs is its retention. The NVM may lose its advantage over other volatile memories if the states can only be retained a few seconds. Retention failure of PCM occurs when the phase-change material in the amorphous state is crystallized into the poly-crystalline state. The crystallization process can be accelerated by chip temperature and/or reading bias voltage [27], also named as thermal disturbance and read disturbance, respectively. The bias voltage on PCM cells will heat up phase change material. The crystallization speed of PCM is dependent on the temperature and increases when the temperature is higher. The elevated temperature due to the bias voltage will result in fast crystallization and hence poor retention. This is also one of the reasons to hold the read voltage much lower than SET voltage. Since the read voltage exponentially reduces the retention time [27], we propose to bias PCM cells at 0 V during FPGA operation which could greatly improve their retention performance. The read disturbance not only exists in PCM, but is also one of the major issues in RRAM [28] and STT-MRAM [29], since the read operation shares the same current path as the write operation.

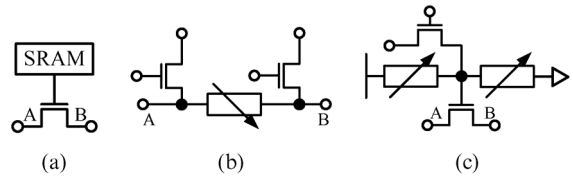


Fig. 2. Conventional (a) SRAM storage element to configure FPGAs; (b) non-volatile storage element to replace the switch transistor and SRAM (2T1R); and (c) non-volatile storage element to configure the switch transistor in FPGAs (1T2R).

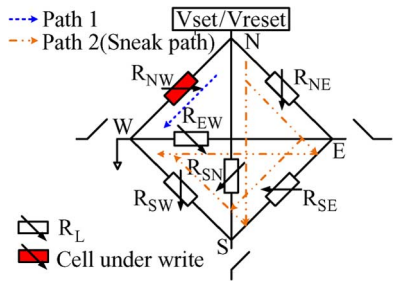


Fig. 3. Conventional 2T1R based non-volatile switch point. The en-dash lines are the paths to program the NVM cells, and dash-dot-dot lines are the sneak paths.

B. Related Works

FGPAs have the opportunity to significantly reduce the area, power and delay with emerging resistive NVMs. We categorize the conventional FPGA configuration memory technologies into three, i.e. SRAM, 2T1R, 1T2R, as shown in Figs. 2(a), 2(b) and 2(c), respectively. The SRAM solution requires long configuration time as well as high power during configuration and standby. The 2T1R solution as shown in Fig. 2(b) was suggested in [22] and [23] to replace the NMOS switch and SRAM cell to achieve high speed and density. Although it addresses some of the issues in SRAM solution, it faces problems such as significant low write reliability and high write power due to the high leakage current in the sneak paths. For example, to program NVM cell R_{NW} between nodes N and W in Fig. 3, the potential on N is at either V_{set} or V_{reset} (where V_{set} and V_{reset} are the SET voltage and $RESET$ voltage, respectively), and W is grounded. However, if R_{NW}, R_{SN} and R_{SW} are at high, low and low states, respectively, most of the current passes through R_{SN} and R_{SW} due to the large difference between R_H and R_L, which can be as high as two orders in resistance. Therefore, the current going through R_{NW} may be too small to switch the state of the cell. The 1T2R solution as shown in Fig. 2(c) was reported in [25] and [26] which has the advantages of instant power-on and zero standby power. Unfortunately, it suffers from high active leakage power and low reliability issues, which limit their application in FPGAs. The high active leakage power and low reliability are caused by the insufficient R_H (only around 2 M Ω), and the low retention of PCM cells with a 1 V bias voltage, respectively.

III. Proposed nvSRAM Based FPGA

The proposed nvSRAM based FPGA, as shown in Fig. 4, has the similar architecture as conventional SRAM-based FPGAs.

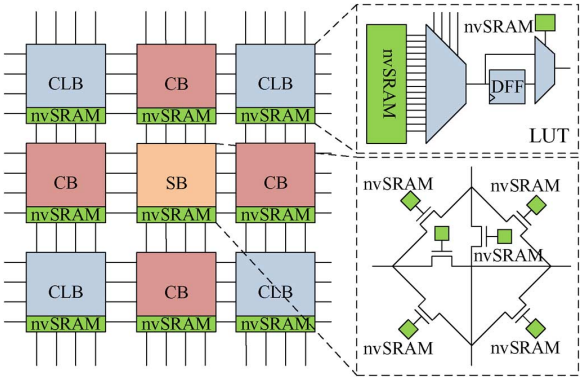
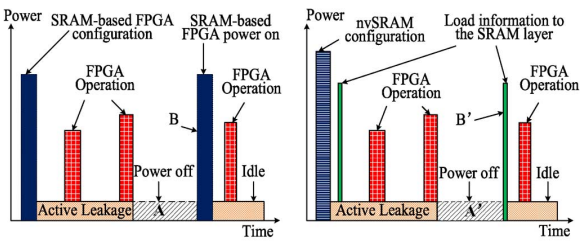


Fig. 4. Proposed nvSRAM based FPGA Architecture. 6T SRAMs are replaced by our proposed nvSRAMs. SB, CB and CLB are switch block, connection block and configurable logic block, respectively.



(a) (b)

Fig. 5. Power consumption of the (a) SRAM-based FPGA and (b) our proposed nvSRAM-based FPGA in different operation modes.

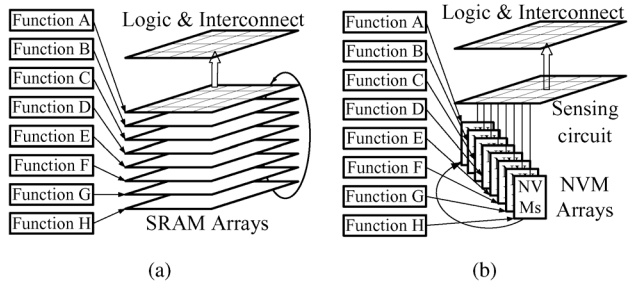
The only difference is that 6T SRAMs are replaced by PCM based nvSRAMs to configure FPGAs.

A. Working Modes and Power Advantage

In the proposed nvSRAM based FPGA, we introduced a loading mode in addition to the traditional sleep mode, configuration mode and normal operation mode. The configuration mode and loading mode of the proposed nvSRAM based FPGA are used to write configuration information to PCM cells, and read configuration information from PCM cells to latches, respectively.

The nvSRAM based FPGAs are only programmed once in the configuration mode. Thereafter, the information stored in PCM cells is sensed in the loading mode to configure the logic and routing in FPGAs. There is only one time loading when FPGAs are powered on. The instant power-on and non-volatile abilities of nvSRAMs reduce the sleep power, power-on time and power-on energy, allowing FPGAs to be powered on/off more frequently to reduce the power consumption.

Fig. 5 explains the power consumption of conventional SRAM-based FPGAs and our nvSRAM-based FPGAs in different modes. As shown in Fig. 5(a), SRAM-based FPGAs have high configuration power and long configuration time. Therefore, SRAM-based FPGAs require significant overhead during power on and off. Break-even point (BEP), which is defined by the time when the reduced sleep energy (area A) equals to the energy required to power on the FPGA (area B), can be used to evaluate power-off possibilities. In other words, only when area A is larger than area B, SRAM-based FPGAs benefit from in powering off in terms of power. Another power off condition is that the sleep time between two events has to be longer than the total width of A and B. As shown in Fig. 5(b), the smaller area B′ of our nvSRAM based FPGA allows area A′ to be much smaller to gain power reduction benefit. Therefore, the width of A′ is much shorter than that of A, and the width of B′ is also much shorter than that of B due to instant power on ability. In other words, our nvSRAM-based FPGAs can be powered off to reduce the FPGA power consumption in a much shorter idle period.



(a) (b)

Fig. 6. (a) Conventional SRAM-based multi-context FPGA; (b) proposed nvSRAM based multi-context FPGA.

B. Multi-Context FPGA and Area Advantage

One solution to reduce the chip area and power consumption is through runtime reconfiguration (RTR) by increasing the hardware utilization [30]. RTR is the ability to modify or change the functional configuration of the device during operation. It can reduce the hardware components (area) and power consumption by reusing the same FPGA for several functions. As it involves reconfiguration during program execution, fast configuration is very important for RTR. However, the traditional single-context FPGA structure only allows one full-chip configuration to be loaded at a time results in very slow reconfiguration. Therefore, SRAM-based multi-context FPGA has been proposed [31]. A key advantage of the multi-context FPGA over a single-context architecture is that it allows the nanoseconds context switch, whereas the single-context may take milliseconds or more to be reprogrammed [31].

However, due to the volatile nature of the SRAM, SRAM-based multi-context FPGAs still suffer from several fundamental drawbacks, including long configuration loading time (need to reload the configuration from the external NVM array every time when powering on), excessive active leakage power (have to always power on all context layers), large configuration memory area (large size of SRAM), low standby possibility and etc.

We propose using NVMs to replace SRAMs to form an NVM-based multi-context FPGA. The NVMs are used to store the FPGA configuration information. Fig. 6(a) illustrates the N-layer multi-context architecture for conventional SRAM-based multi-context FPGAs. N is set to 8 in this example for illustration, but not limited to 8. It can be seen that there are eight context layers of SRAMs. Each SRAM layer contains the configuration information for a different function. Based on the application, different SRAM layer is selected. The switching among these configuration layers can be achieved during execution. The multiple configuration layers can be combined to emulate a single large function. Fig. 6(b) shows the proposed nvSRAM based multi-context FPGA. The main difference is that the eight SRAM layers are replaced by eight NVM layers. Each NVM layer contains different function. It has the same operation scheme as the conventional SRAM-based one. A shared sensing circuit is

designed to control the NVM layers. Because the cell size of NVM is only about 3% of that of SRAM [32], the chip area of FPGA could thus be significantly reduced.

IV. Proposed Storage Element

To reduce the active leakage power and increase the reliability, we follow three design principles. The first principle is to bias PCM cells at 0 V during the FPGA normal operation. Hence there is no active leakage current on PCM cells, and their states will not be disturbed. The second principle is to quickly load the configuration information from PCM cells to latches with low read power, thus allows the FPGA to be powered on/off more frequently, and switch between contexts much faster. The last principle is to remove the high voltage inside the nvSRAM during PCM cell programming, thus low VDD devices can be used to achieve high density. With these principles, we propose both single-context nvSRAM and multi-context nvSRAM in the following.

A. Single Context nvSRAM

The proposed PCM based single-context nvSRAM storage element is shown in Fig. 7. As discussed in Section III, our proposed nvSRAM has three modes besides the sleep mode, the detailed description of each mode is provided as follows:

a). In the configuration (write) mode, read enable signal (REb) is high to turn off the equalization transistor MP_2, thus the four transistors (MP_0, MP_1, MN_0 and MN_1) formed latch isolates FPGA operation supply voltage (VDD) from nodes SL_p and SL_n. This results in no dc path between VDD and the write voltages (V_{set} and V_{reset}) of the PCM cells. Meanwhile, the control signal S_1 is high to pull nodes SL_p and SL_n to the ground. The nodes BL_p and BL_n are driven by the SET voltage (V_{set}) and RESET voltage (V_{reset}) pulses according to the configuration information. For example, if the configuration information is "0", R_0 and R_1 are under RESET and SET operations, respectively. It is worth noting that the high write voltage is not connected to SL_p or SL_n as reported in [33]. This avoids the use of thick oxide transistors in the latch. After configuration, R_0 is at high resistance state (R_H), and R_1 is at low resistance state (R_L). The simplified schematic of the proposed nvSRAM to write the PCM cells is shown in Fig. 8(a).

b). In the loading (read) mode, as shown in Fig. 8(b), BL_p and BL_n are pulled to the ground, and S_1 is low to disconnect SL_p and SL_n from the ground. Meanwhile, REb is also low to equalize SL_p and SL_n to $VDD-V_{thp}-V_{thn}$, where V_{thp} and V_{thn} are the threshold voltages of PMOS and NMOS transistors, respectively. Due to pre-configured information on R_0 and R_1, the nvSRAM forms two asymmetric current paths. For example, when $R_0 = R_H$, $R_1 = R_L$, the current on R_1 is much larger than that on R_0. Therefore, the output node Q_p is pulled down, thus pulls up Q_n. The asymmetry of current paths forms a third current path in MP_2 from Q_n to Q_p. Once REb is high, the latch pulls Q_n to VDD and Q_p to the ground.

c). In the FPGA normal operation mode, BL_p and BL_n are still at the ground, and REb is high. Moreover, S_1 is turned on to pull SL_p and SL_n to the ground and thus bias PCM cells at 0 V, resulting in zero active leakage power and long retention time. The nvSRAM works like a convectional SRAM to configure the logic and routing in the FPGA. Fig. 8(c) shows the simplified SRAM-like schematic of the nvSRAM during the FPGA normal operation mode.

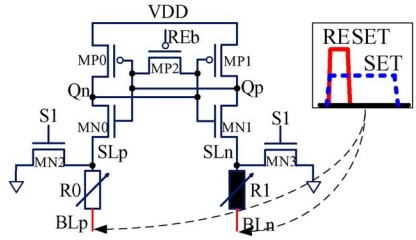


Fig. 7. Proposed single-context nvSRAM. The signals BL_p and BL_n are shared with other nvSRAMs in the same column.

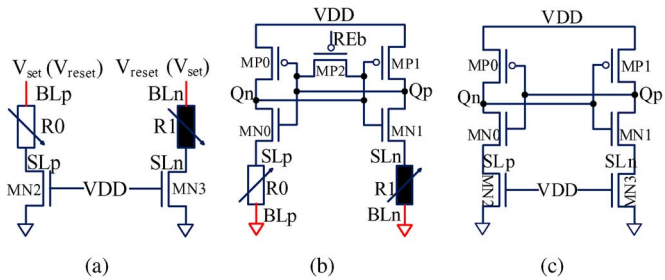


Fig. 8. Proposed single context in the (a) write mode, (b) read mode, and (d) FPGA execution mode.

TABLE I
Control Logic Information of Our Proposed nvSRAM in Different Operation Modes

Modes	REb	S_1	BL_p	BL_n
Write (1)	1	1	V_{set}	V_{reset}
Write (0)	1	1	V_{reset}	V_{set}
Read	Negative Pulse	0	0	0
Normal operation	1	1	0	0

The control logic information of our proposed nvSRAM in different operation modes is tabulated in Table I. The proposed nvSRAM contains 7 transistors, one more than the conventional 6T SRAM. During writing, the drain of transistors MN_2 and MN_3 are pulled to the ground, and the high write voltage is isolated by the PCM cells. As a result, thin oxide transistors can be used in the nvSRAM, leading to significant reduction in nvSRAM size.

B. Multi-Context nvSRAM

We further propose an nvSRAM with multiple layers of programming bits (multi-context nvSRAM), where each layer can be activated at a different time point. Our proposed multi-context nvSRAM shows a great potential in run-time reconfiguration applications, since it only needs less than 1 ns to switch between different contexts.

The proposed multi-context nvSRAM, as shown in Fig. 9, not only has the non-volatile and instant power-on advantages, but also helps to reduce the area by sharing the latch. Compared to the SRAM-based multi-context FPGA, the area, standby power, power-on time and power-on energy could be significantly reduced. In Fig. 9, the context select transistor pairs $MN_4\langle N-1:0\rangle$ and $MN_5\langle N-1:0\rangle$ are inserted between the latch and PCM cells. The context select transistors are controlled by the context

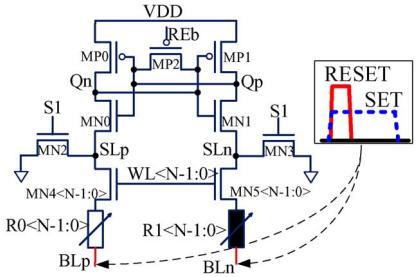


Fig. 9. Proposed multi-context nvSRAM. The signals BL_p and BL_n are shared with other nvSRAMs in the same column.

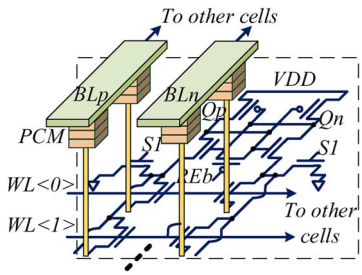


Fig. 10. Schematic of the nvSRAM 3D integration. The phase change material is deposited in the format of thin-film on the top of the CMOS transistors.

select address $WL\langle N-1:0\rangle$. The N-context requires N bits context selected address, N pairs of select transistors and N pairs of PCM cells.

The multi-context nvSRAM has four operation modes in addition to the sleep mode: the configuration mode, the loading mode, the multi-context switch mode and the FPGA normal operation mode. These modes are similar to the single-context nvSRAM except the context switch mode. The context switching mode is for run-time reconfiguration, which performs almost the same as the read operation. The only difference is that it first changes the context address to the targeted layer before sensing the configuration information from the selected layer to the latch.

A 3D integration schematic of the CMOS circuits and PCM cells is shown in Fig. 10. The phase change material is deposited in the format of thin-film on the top of the CMOS circuits, thus no additional area is required for PCM cells. The latch is shared by different context layers, resulting smaller area of the multi-context nvSRAM than the multi-context SRAM. Fig. 10 shows an example of 2-context nvSRAM, where all PCM cells are placed in the same layer.

The multi-context nvSRAM also allows dynamic reconfiguration during the FPGA normal operation when required logic function is not pre-configured in PCM cells. The FPGA operation is not interrupted when writing new information to the PCM cells. During dynamic reconfiguration, S_1 is high to pull the nodes SL_p and SL_n to the ground. Therefore, the configuration information is still latched by MP_0, MP_1 and MN_0 to MN_3. Then a normal write operation is performed to the selected PCM cells. The new states of the PCM cells could be sensed at any time when required by the FPGA systems. The FPGA systems are interrupted in a very short time period since the sensing speed is less than 1 ns.

V. Simulation Results

In this section, we first evaluate the power and delay performance of the proposed single-context nvSRAM based 4-input LUT, and another three 4-input LUT architectures. After that, we analyze the retention of PCM cells to be integrated in three different schemes. In the second part of this section, we compare the power, delay, loading energy and area among these four multi-context 4-input LUTs.

TABLE II
Parameters of the PCM Used in the Simulation

PCM	Parameter
Technology node	$20nm$
SET/RESET pulse width	$200ns/20ns$
SET/RESET voltage	$1.2V/1.7V$
SET/RESET current	$60\mu A/100\mu A$
Low/High Resistance	$20K\Omega/2M\Omega$

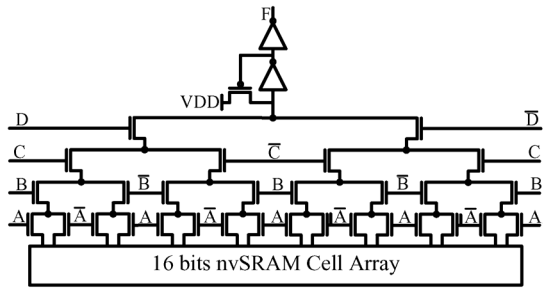


Fig. 11. The 4-input LUT structure used to evaluate the proposed nvSRAM.

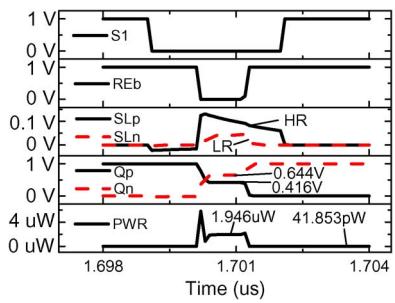


Fig. 12. Power and delay simulation results of the proposed nvSRAM when loading the states from PCM cells to the latch.

To evaluate the proposed nvSRAM, test benches were built based on a 45 nm CMOS process node. GST based PCM is used in our simulation. Our PCM model uses the same resistance value and pulse width as [25]. The high resistance (R_H) and the low resistance (R_L) are 2 MΩ and 20 KΩ, respectively. The SET and $RESET$ pulse widths of the PCM model are 200 ns and 20 ns, respectively. Our default SET and $RESET$ voltages are 1.2 V and 1.7 V, respectively. The detailed PCM parameters are tabulated in Table II. We built a read disturbance model according to the data provided by [34] to compare the data retention.

A. Single Context Simulation Results

The power and delay simulation results given in Fig. 12 shows that our proposed nvSRAM achieves a 41.8 pW low active leakage power and a within 1 ns high sensing speed. The low active leakage power is due to zero bias voltage on PCM cells by pulling SL_p and SL_n to the ground. The reading power of nvSRAM cell is only around 1.95 uW, hence the time and energy consumed by reading are shorter and lower than configuration of the SRAM cell when FPGAs are powered on.

A 4-input LUT in Fig. 11 is used to evaluate the performance of the four LUTs based on the proposed nvSRAM, SRAM, and

TABLE III
RESULTS COMPARISON AMONG THE SRAM, PROPOSED nvSRAM, [25] AND [26]

	This work	[25]	[26]	SRAM
Non-volatile	Yes	Yes	Yes	No
4-input LUT Active Leakage Power	$1.19nW$	$207nW$	$2.15\mu W$	$1.17nW$
4-input LUT Switching Energy	$2.58fJ$	$3fJ$	$2.2fJ$	$2.5fJ$
4-input LUT Pull-down Delay	$280ps$	$310ps$	$316ps$	$270ps$
4-input LUT Pull-up Delay	$250ps$	$220ps$	$186ps$	$220ps$
FPGA Power-on Speed	<1ns (∼300ps)	$90ps$	$90ps$	milliseconds
FPGA Power-on Energy	$2.54fJ/bit$	$2.16fJ/bit$	$3.07fJ/bit$	∼50fJ/bit [35]
Data Retention	>10 years	$250\mu s$	$250\mu s$	Preserved so long as voltage is applied

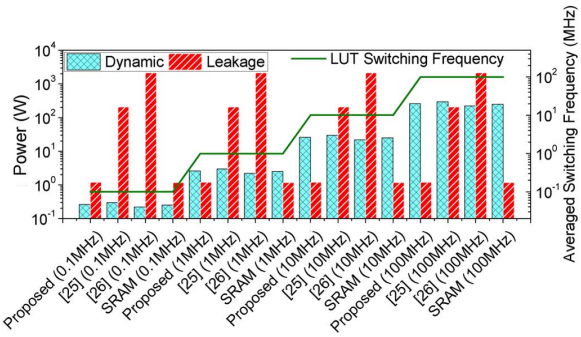


Fig. 13. Power consumption comparison among different LUT architectures.

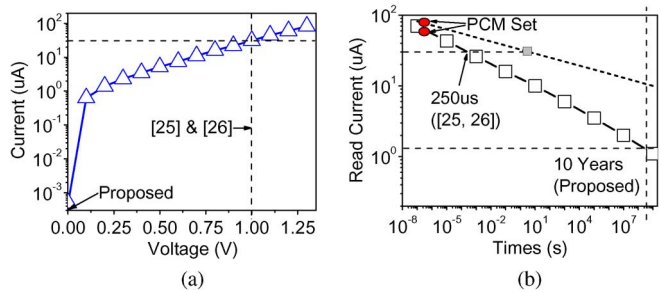


Fig. 14. (a) *I–V* curve of the PCM cell in the amorphous state. (b) the PCM retention of the designs in [25], [26], and our proposed nvSRAM.

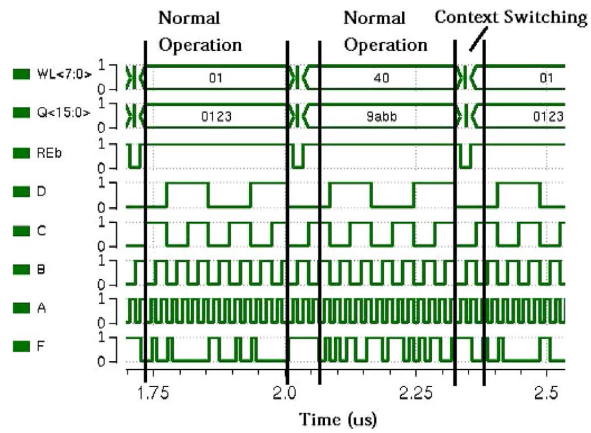


Fig. 15. RTR simulation results of the proposed 8-context nvSRAM based 4-input LUT.

those in [25] and [26]. The LUT in [25] is extended to the same four inputs. The SRAM based LUTs use the same structure as in Fig. 11 by replacing nvSRAM cells with 6T SRAMs. The resistance of the pull-down resistor in [26] is set to the logarithmic middle point of R_H and R_L (200 KΩ).

The power and delay comparison among the four 4-input LUTs is tabulated in Table III. The delay is measured from input A to output F. As shown in Table III, the proposed nvSRAM based 4-input LUT achieves the similar speed performance as the conventional schemes. The 1.19 nW active leakage power is similar to the SRAM-based LUT, but much smaller than [25] and [26]. The active leakage power of [25] and [26] is about 174 times and 1810 times higher than that of the proposed structure, respectively. Based on the 4-input LUT simulation results, our nvSRAM-based LUT could be powered off to reduce the leakage power when the sleep time is longer than 34.5 μs.

As illustrated in Fig. 13, the dynamic power and active leakage power of the four LUTs are compared at different operating frequencies. At low frequency (i.e., 0.1 MHz), the active leakage power of [25] and [26] are 2–4 orders higher than the dynamic power. Only when the averaged switching frequency is higher than 100 MHz, the active leakage power in [25] gets lower than the dynamic power. However, the active leakage power in [26] is still more than 10 times higher than its dynamic power. In contrast, even at 1 MHz low switching frequency, the active leakage power of the LUT with our proposed nvSRAM is already lower than the dynamic power.

The retention time of PCM cells with our proposed nvSRAM, and the circuits in [25] and [26] are evaluated based on the data reported in [34]. As shown in Fig. 14, the reading current is exponentially increased with the reading voltage, and the crystallization time of PCM cells is exponentially reduced with reading current increased, which is because of the higher temperature inside PCM cells at higher reading current. Therefore, when the cells are biased at 1 V, the high reading current (30 μA) leads to much shorter data retention time (crystallized in 250 μs). In our proposed design, the retention time could be longer than 10 years, since the sensing energy is low and there is no bias current in PCM cells during FPGA normal operations. The results are summarized in Table III. The retention time of PCM may be improved by using different materials (i.e., GeTe) [36], [37]. However, the SET voltage/current may be increased due to the different materials. Moreover, the low retention problem may not be fully addressed due to the high dc biased voltage, i.e., the short-dash line shown in Fig. 14(b).

B. Multi-Context Simulation Results

The multi-context 4-input LUTs use the same structure as the single-context 4-input LUTs. Fig. 15 shows the run time reconfiguration of the 4-input LUT with 8-context nvSRAM. At the first read cycle, the multi-context nvSRAM address $8'h01$ is selected. This address sets the LUT to $16'h0123$ to have the logic function of $F = \bar{A}\bar{B}\bar{C} + A\bar{B}\bar{D}$. When the read operation is finished, the states of the PCM cells $(16'h0123)$ are sensed and latched at the output $Q\langle 15:0\rangle$. The inputs of the LUT are swept from $4'$b0000 to $4'$b1111, and the sequence of the output signal F is . . . 1100_0100_1000_0000 . . ., which agrees well with the

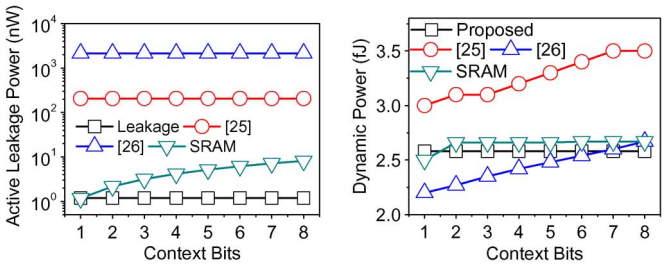


Fig. 16. 4-input LUT (a) active leakage power and (b) dynamic power comparison among the 6T SRAM, the designs in [25], [26], and the proposed nvSRAM.

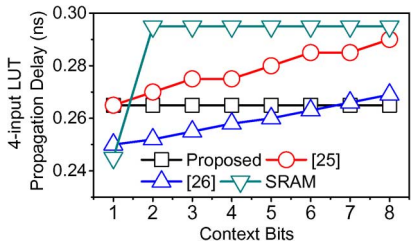


Fig. 17. 4-input LUT active propagation delay comparison among the 6T SRAM, the designs in [25], [26], and the proposed nvSRAM.

states of the PCM cells. At around 2 us, another read cycle selects $8'h40$ as the context address of the nvSRAM which sets the LUT logic function to $F = AB + A\bar{C} + \bar{B}C + \bar{B}\bar{D}$. When the read operation is completed, the states of the data $Q\langle 15:0\rangle$ have been changed to $16'h9abb$. The switch between different context could be accomplished in less than 1 ns.

Fig. 16(a) shows the multi-context 4-input LUT leakage power comparison among the 6T SRAM, the designs in [25], [26], and our proposed nvSRAM. Since the designs in [25], [26], and our proposed nvSRAM are using non-volatile memory technologies, the unselected context bits could be turned off, thus the active leakage power increases little at the wide span of context bits. However, the SRAM based LUT has to power on the unselected SRAM cells, thus higher context bits LUT draws higher active leakage power. Our nvSRAM based 8-context LUT reduces active power by 8, 174 and 1810 times, respectively, compared to the 8-context LUTs using 6T SRAM, the designs in [25] and [26].

Fig. 16(b) shows the 4-input LUT dynamic power comparison among four techniques. The SRAM and our proposed nvSRAM based LUTs have the similar dynamic power due to the same LUT structure is used. The dynamic power of [25], [26] gets higher with larger context bits is due to the parasitic capacitance from the other PCM select transistors.

Fig. 17 shows the propagation delay of four techniques. The additional context select switches are inserted between the multi-context SRAM and LUT switch matrix, resulting a longer delay in the multi-context SRAM based LUT compared to the single-context LUT. The parasitic capacitance of the design in [25], [26] gets larger at higher context bits, thus the total propagation delay is proportional to the context bits. The speed of our nvSRAM is determined by the latch. Therefore, its propagation delay is not affected by the increase of context bits.

Fig. 18 gives the loading energy per information bit comparison between the proposed nvSRAM, and the designs in [25] and [26]. The loading power of our nvSRAM has little dependence on context bits, which are 2.54 fJ and 2.58 fJ for the single-context and 8-context, respectively. However, from single-context to 8-context, the loading power increases about 40% and 10%, respectively, in designs of [25] and [26].

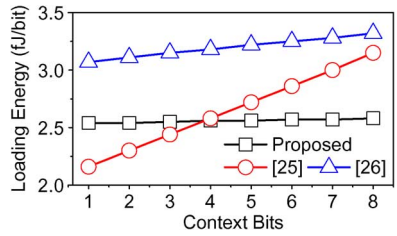


Fig. 18. 4-input LUT loading power comparison among the 6T SRAM, the designs in [25], [26], and the proposed nvSRAM.

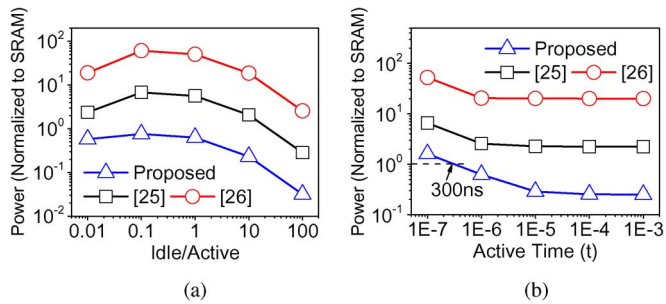


Fig. 19. The 8-context 4-input LUT power comparison among the designs in [25], [26], and the proposed nvSRAM. All of the results are normalized to the SRAM based 8-context 4-input LUT under the same conditions. The average LUT switching frequency is set to 10 MHz. (a) The power consumption versus the ratio of idle time and active time. The active time is set to 1 ms. (b) The power consumption versus the active time. The ratio of idle time and active time is 0.9.

We further estimated the power consumption of the 8-context 4-input LUTs involving both idle/sleep time and active time as shown in Fig. 19. The SRAM-based LUT is still powered on during idle time, and its results are used as the baseline to compare the designs in [25], [26], and the proposed nvSRAM. As shown in Fig. 19(a), our nvSRAM based LUT has much lower power consumption than the SRAM-based LUT regardless of idle and active ratio. In contrast, the designs of [25] and [26] based LUTs start to outperform SRAM-based LUT in terms of power consumption only if the idle and active ratio is higher than 25 and 300, respectively. This is mainly due to high active leakage power. As shown in Fig. 19(b), our nvSRAM-based LUT consumes less power than the SRAM-based LUT when active time is longer than 300 ns. Unfortunately, the designs of [25] and [26] based LUTs have more than 2 and 20 times higher power consumption than the SRAM-based LUT, respectively.

The area of the proposed multi-context nvSRAM can be derived from $AREA = AREA_1 + N * AREA_2$, where $AREA_1$ is the area of the latch plus the area of MN_2, MN_3 and the equalization transistor, $AREA_2$ is the area of single memory select pair. $AREA_1$ approximately equals to the area of the single context nvSRAM which is only 0.84 um^2 based on the 45 nm CMOS process node. The area comparison in Fig. 20 is based on the layout and the data provided in [25], which has been normalized to 45 nm after dividing it by 4. The cell size of the single context 6T SRAM in Fig. 20 is normalized to 1. Because of the thick oxide transistors, the normalized area of the PCM cell in [25] is more than 5 times larger than the proposed nvSRAM.

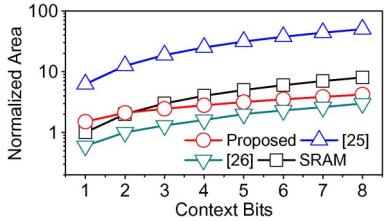


Fig. 20. Area comparison among the 6T SRAM, the design in [25] and our proposed nvSRAM. The area is normalized to the single context 6T SRAM.

The area of our nvSRAM gets smaller than 6T SRAM when the context bits are larger than 2.

VI. Conclusion

In this paper, we have proposed a PCM based non-volatile SRAM, which greatly reduces the active leakage power, and enhances the reliability of PCM cells by biasing PCM cells at 0 V during the FPGA normal operation. The results have shown that the 4-input LUT with our nvSRAM has only 1.19 nW active leakage power while producing 1 ns fast loading speed. These features allow the system to be powered on/off to reduce the leakage power when standby time is longer than 34.5 us. The analysis also shown that the retention of the PCM cells can be longer than 10 years. The results suggest that our proposed nvSRAM is a promising solution for low power and high reliability FPGAs.

References

[1] S. Brown, J. Rose, and Z. Vranesic, "A detailed router for field-programmable gate arrays," in *Proc. ICCAD-90*, Nov. 1990, pp. 382–385.

[2] P. Chow, S. Seo, J. Rose, K. Chung, G. Páez-Monzón, and I. Rahardja, "The design of an sram-based field-programmable gate array. i. architecture," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 7, no. 2, pp. 191–197, 1999.

[3] I. Kuon, R. Tessier, and J. Rose, "Fpga architecture: Survey and challenges," *Foundations and Trends in Electronic Design Automation*, vol. 2, no. 2, pp. 135–253, 2008.

[4] T. Lin, W. Zhang, and N. Jha, "SRAM-based nature: A dynamically reconfigurable fpga based on 10t low-power srams," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. 20, no. 11, pp. 2151–2156, 2012.

[5] A. Tajalli and Y. Leblebici, "Design trade-offs in ultra-low-power digital nanoscale cmos," *IEEE Trans. Circuits Syst.–I: Reg. Papers*, vol. 58, no. 9, pp. 2189–2200, 2011.

[6] M. Henry and L. Nazhandali, "Nems-based functional unit power-gating: Design, analysis, optimization," *IEEE Trans. Circuits Syst.–I:Reg. Papers*, vol. 60, no. 2, pp. 290–302, 2013.

[7] S. Lai, "Current status of the phase change memory and its future," in *IEDM Tech Dig.*, 2003, pp. 10–11.

[8] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, "Scalable high performance main memory system using phase-change memory technology," *ACM SIGARCH Computer Architecture News*, vol. 37, no. 3, pp. 24–33, 2009.

[9] G. Close, U. Frey, J. Morrish, R. Jordan, S. Lewis, T. Maffitt, M. BrightSky, C. Hagleitner, C. Lam, and E. Eleftheriou, "A 256-mcell phase-change memory chip operating at 2 + bit/cell," *IEEE Trans. Circuits Syst.–I:Reg. Papers*, vol. 60, no. 6, pp. 1521–1533, 2013.

[10] D.-H. Kwon, K. M. Kim, J. H. Jang, J. M. Jeon, M. H. Lee, G. H. Kim, X.-S. Li, G.-S. Park, B. Lee, and S. Han *et al.*, "Atomic structure of conducting nanofilaments in tio2 resistive switching memory," *Nature Nanotechnol.*, vol. 5, no. 2, pp. 148–153, 2010.

[11] S.-Y. Kim, J.-M. Baek, D.-J. Seo, J.-K. Park, J.-H. Chun, and K.-W. Kwon, "Power-efficient fast write and hidden refresh of reram using an adc-based sense amplifier," *IEEE Trans. Circuits Syst.–II:Express Briefs*, vol. 60, no. 11, pp. 776–780, 2013.

[12] K. Huang, N. Ning, and Y. Lian, "Optimization scheme to minimize reference resistance distribution of spin-transfer-torque mram," *IEEE Trans. Very Large Scale Integration (VLSI) Syst.*, vol. PP, no. 99, pp. 1–1, 2013.

[13] K. Huang and Y. Lian, "A low-power low-vdd nonvolatile latch using spin transfer torque mram," *IEEE Trans. Nanotechnol.*, vol. 12, no. 6, pp. 1094–1103, 2013.

[14] M. Wuttig, "Phase-change materials: Towards a universal memory?," *Nature Materials*, vol. 4, no. 4, pp. 265–266, 2005.

[15] H. Hamann, M. O'Boyle, Y. Martin, M. Rooks, and H. Wickramasinghe, "Ultra-high-density phase-change storage and memory," *Nature Materials*, vol. 5, no. 5, pp. 383–387, 2006.

[16] S. Lee, Y. Jung, and R. Agarwal, "Highly scalable non-volatile and ultra-low-power phase-change nanowire memory," *Nature Nanotechnol.*, vol. 2, no. 10, pp. 626–630, 2007.

[17] M. Lankhorst, B. Ketelaars, and R. Wolters, "Low-cost and nanoscale non-volatile memory concept for future silicon chips," *Nature Materials*, vol. 4, no. 4, pp. 347–352, 2005.

[18] G. Servalli, "A 45 nm generation phase change memory technology," in *IEDM Tech. Dig.*, 2009, pp. 1–4, IEEE.

[19] Y. Choi, I. Song, M. Park, H. Chung, S. Chang, B. Cho, J. Kim, Y. Oh, D. Kwon, and J. Sunwoo *et al.*, "A 20 nm 1.8 v 8 gb pram with 40 mb/s program bandwidth," in *Proc. ISSCC*, 2012, pp. 46–48.

[20] D. Loke, T. Lee, W. Wang, L. Shi, R. Zhao, Y. Yeo, T. Chong, and S. Elliott, "Breaking the speed limits of phase-change memory," *Science*, vol. 336, no. 6088, pp. 1566–1569, 2012.

[21] F. Xiong, A. Liao, D. Estrada, and E. Pop, "Low-power switching of phase-change materials with carbon nanotube electrodes," *Science*, vol. 332, no. 6029, pp. 568–570, 2011.

[22] J. Cong and B. Xiao, "mrfpga: A novel fpga architecture with memristor-based reconfiguration," in *Proc. NANOARCH*, 2011, pp. 1–8.

[23] P.-E. Gaillardon, M. Ben-Jamaa, G. Beneventi, F. Clermidy, and L. Perniola, "Emerging memory technologies for reconfigurable routing in fpga architecture," in *Proc. ICECS*, Dec. 2010, pp. 62–65.

[24] Y. Chen, J. Zhao, and Y. Xie, "3d-nonfar: Three-dimensional non-volatile fpga architecture using phase change memory," in *Proc. 16th ACM/IEEE Int. Symp. Low Power Electron. Design*, 2010, pp. 55–60.

[25] C. Wen, J. Li, S. Kim, M. Breitwisch, C. Lam, J. Paramesh, and L. Pileggi, "A non-volatile look-up table design using pcm (phase-change memory) cells," in *Proc. VLSIC*, 2011, pp. 302–303.

[26] P. Gaillardon, D. Sacchetto, G. Beneventi, M. Ben Jamaa, L. Perniola, F. Clermidy, I. O'Connor, and G. De Micheli, "Design and architectural assessment of 3-d resistive memory technologies in fpgas," *IEEE Trans. Nanotechnol.*, vol. 12, no. 1, pp. 40–50, 2013.

[27] S. Kim, B. Lee, M. Asheghi, G. Hurkx, J. Reifenberg, K. Goodson, and H. Wong, "Thermal disturbance and its impact on reliability of phase-change memory studied by the micro-thermal stage," in *Proc. IRPS*, 2010, pp. 99–103, IEEE.

[28] Y. Chen, H. Lee, P. Chen, P. Gu, C. Chen, W. Lin, W. Liu, Y. Hsu, S. Sheu, and P. Chiang *et al.*, "Highly scalable hafnium oxide memory with improvements of resistive distribution and read disturb immunity," in *IEDM Tech. Digest*, 2009, pp. 1–4.

[29] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, and M. Liu *et al.*, "45 nm low power cmos logic compatible embedded stt mram utilizing a reverse-connection 1t/1mtj cell," in *IEDM Tech. Dig.*, 2009, pp. 1–4.

[30] K. Compton and S. Hauck, "Reconfigurable computing: A survey of systems and software," *ACM Computing Surveys (csuR)*, vol. 34, no. 2, pp. 171–210, 2002.

[31] K. Compton, S. Hauck, and K. Compton, "An introduction to reconfigurable computing," *IEEE Comput.*, 2000.

[32] "International technology roadmap for semiconductors—Emerging research devices (erd)," [Online]. Available: http://www.itrs.net/Links/2011ITRS/Home2011.htm 2011

[33] N. Bruchon, L. Torres, G. Sassatelli, and G. Cambon, "New nonvolatile fpga concept using magnetic tunneling junction," in *IEEE Comput. Society Annual Symp. Emerging VLSI Technol. Architectures,*, 2006, 6-pp.

[34] A. Pirovano, A. Redaelli, F. Pellizzer, F. Ottogalli, M. Tosi, D. Ielmini, A. Lacaita, and R. Bez, "Reliability study of phase-change nonvolatile memories," *IEEE Trans. Device Materials Reliabil.*, vol. 4, no. 3, pp. 422–427, 2004.

[35] M. Wieckowski, G. K. Chen, D. Kim, D. Blaauw, and D. Sylvester, "A 128 kb high density portless sram using hierarchical bitlines and thyristor sense amplifiers," in *Proc. ISQED*, 2011, pp. 1–4.

[36] L. Perniola, V. Sousa, A. Fantini, E. Arbaoui, A. Bastard, M. Armand, A. Fargeix, C. Jahan, J.-F. Nodin, and A. Persico *et al.*, "Electrical behavior of phase-change memory cells based on gete," *IEEE Electron Device Lett.*, vol. 31, no. 5, pp. 488–490, 2010.

[37] G. Betti Beneventi, L. Perniola, V. Sousa, E. Gourvest, S. Maitrejean, J. Bastien, A. Bastard, B. Hyot, A. Fargeix, and C. Jahan *et al.*, "Carbon-doped gete: A promising material for phase-change memories," *Solid-State Electron.*, vol. 65, pp. 197–204, 2011.

Kejie Huang (S'13) received the B.S. and M.S. degrees from the College of Information Science and Engineering from Zhejiang University, China in 2003 and 2006, respectively. He is currently working towards the Ph.D. degree in the Department of Electrical and Computer Engineering, National University of Singapore.

He is a Research Engineer in the Department of Engineering Product Design, Singapore University of Technology and Design. His research interests include architecture and circuit optimization for reconfigurable computing systems, emerging embedded memory design, and neuromorphic circuit and system design.

Yajun Ha (SM'09) received the B.S. degree from Zhejiang University, China, in 1996, the M.Eng. degree from the National University of Singapore in 1999, and the Ph.D. degree from Katholieke Universiteit Leuven, Belgium, in 2004, all in electrical engineering.

He is currently a Research Scientist at the Institute for Infocomm Research, Singapore. Prior to this, he was an Assistant Professor with the Department of Electrical and Computer Engineering at the National University of Singapore, and a Researcher at the Inter-University MicroElectronics Center (IMEC) in Leuven, Belgium. His research interests include the general area of embedded computing (VLSI) architecture and design methodologies, with a focus on reconfigurable computing. He has published around 70 internationally peer-reviewed journal/conference papers on these topics.

Dr. Ha has served in a number of positions in the professional communities. He serves as the Associate Editor for the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION (VLSI)SYSTEMS (since 2013), the Associate Editor for the IEEE TRANSACTIONS CIRCUITS AND SYSTEM– II: EXPRESS BRIEFS (2011–2013) and the *Journal of Low Power Electronics* (since 2009). He serves as the General Co-Chair of ASP-DAC 2014; Program Co-Chair for FPT 2010 and FPT 2013; Chair of the Singapore Chapter of the IEEE Circuits and Systems (CAS) Society (2011 and 2012); member of the ASP-DAC Steering Committee; member of the IEEE CAS VLSI and Applications Technical Committee. He is a program committee member for a number of well-known conferences in the fields of embedded systems and FPGAs, such as DATE, ASP-DAC, FPL and FPT.

Rong Zhao (M'06) received the Ph.D. degree in electrical and computer engineering from National University of Singapore in 1999.

Currently she is an Associate Professor in the pillar of Engineering Product Development, Singapore University of Technology and Design (SUTD). Prior to joining SUTD, she was a Senior Scientist, Principle Investigator, and Assistant Division Manager at the Data Storage Institute, A*STAR. Her main research interests include non-volatile memories (phase-change memory and resistive memory) and reconfigurable devices covering from material synthesis, device design/fabrication, to chip design/prototyping. More recently, she has broadened her activities, entering the field of artificial cognitive memory and energy harvesters. She is the author or co-author of about 80 publications in international journals and international conference proceedings.

Dr. Zhao is the Co-Chair of the technical committee of the IEEE Non-Volatile Memory Technology Symposium (2012–2014), and the lead organizer of Material Research Society (2011 and 2012) spring meetings for the Phase Change Symposium.

Akash Kumar (M'09–SM'14) received the B.S. in computer engineering from the National University of Singapore (NUS), Singapore, in 2002. the joint Master of Technological Design degree in embedded systems from NUS and the Eindhoven University of Technology (TUe), Eindhoven, The Netherlands, in 2004, and the joint Ph.D. in electrical engineering in the area of embedded systems from TUe and NUS, in 2009.

Since 2009, he has been with the Department of Electrical and Computer Engineering, NUS. Currently, he is an Assistant Professor there. His research interests include analysis, architectures, design methodologies, and resource management of embedded multiprocessor systems. He has published over 60 papers in leading international electronic design automation journals and conferences.

Dr. Kumar is a member of various technical program committee of design automation and FPGA conferences like DAC, DATE, FPL, FPT.

Yong Lian (M'90–SM'99–F'09) received the B.Sc. degree from Shanghai Jiao Tong University, China, in 1984, and the Ph.D. degree from the National University of Singapore, in 1994.

He worked in industry for nine years before joining the National University of Singapore in 1996 where he is currently a Professor and Area Director in Integrated Circuits and Embedded Systems in the Department of Electrical and Computer Engineering. His research interests include digital filter design, biomedical instrumentation, wireless and wearable biomedical devices, low power IC design, RF IC design, and e-learning tools for large class teaching. He is the author or coauthor of over 150 scientific publications in peer reviewed journals, conference proceedings, and chapters in books.

Dr Lian is the recipient of the 1996 IEEE Circuits and Systems Society's Guillemin-Cauer Award for the best paper published in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–II: EXPRESS BRIEFS, the Best Student Paper Award (as Advisor) in the IEEE 2007 International Conference on Multimedia & Expo (ICME07), the 2008 Best Paper Award from the IEEE Communications Society for the paper published in the IEEE TRANSACTIONS ON MULTIMEDIA, the winner of 47th DAC/ISSCC Student Design Contest (as Advisor). He currently serves as the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–II: EXPRESS BRIEFS. He has been involved in various IEEE activities. He serves/served as the Vice President for Asia Pacific Region of the IEEE Circuits and Systems (CAS) Society; IEEE CAS Society Representative to the IEEE Biometrics Council, IEEE CAS Society Representative to the Biotechnology Council, Chair of the Biomedical Circuits and Systems Technical Committee of IEEE CAS Society; Chair of the Digital Signal Processing Technical Committee of IEEE CAS Society; and steering committee member of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS. He was the Distinguished Lecturer of the IEEE CAS Society. He serves/served as Associate Editors for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I AND II (2002–2009), Associate Editor for the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS (2007–now), Associate Editor for the Journal of Circuits Systems and Signal Processing (2000–2009), Guest Editor of Special Issues in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–I: REGULAR PAPERS (2010), IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS (2008) and in the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS–I: REGULAR PAPERS (2005), Guest Editor of the *Journal of Circuits Systems and Signal Processing* for the Special Issues in 2003, 2005, and 2010. He is the founder of several international conferences including the IEEE International Conference on Biomedical Circuits and Systems (BioCAS), International Conference on Green Circuits and Systems (ICGCS), and Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia).