23.4 Nonvolatile 3D-FPGA With Monolithically Stacked RRAM-Based Configuration Memory

Young Yang Liauw, Zhiping Zhang, Wanki Kim, Abbas El Gamal, S. Simon Wong

Stanford University, Stanford, CA

SRAM based configuration memory is the primary contributor to the large area, delay, and power consumption of FPGAs relative to ASICs. In [1] it is estimated that a 3D-FPGA with the configuration memory stacked on top of the FPGA logic and routing can achieve 57% smaller die area than a baseline 2D-FPGA in 65nm CMOS technology. Motivated by these potential performance gains, several programmable logic devices with different monolithically stacked configuration memory technologies have been reported [2-4]. These memory technologies, however, require materials and/or processes that may not be compatible or scalable with CMOS processes. This paper presents the first 3D-FPGA with stacked configuration memory based on the emerging nonvolatile Resistive RAM (RRAM) technology described in [5], which is both compatible and scalable with CMOS.

The architecture of the proposed 3D-FPGA is shown in Fig. 23.4.1. It consists of an array of tiles, row and column decoders and programming drivers for the configuration memory, and configurable I/O blocks (IOBs). Each tile comprises a configurable logic block (CLB) with a cluster of look-up-tables (LUTs), horizontal and vertical segmented routing channels, two connect blocks (CBs) for interfacing the CLB to the channels, and a switch block (SB) for signal routing between channels. The LUTs and routing switches are implemented using pass gate multiplexors controlled by configuration memory cells. The transistors and interconnects are implemented in a standard CMOS process, while the configuration memory is stacked on top of the transistors and interconnects.

The configuration memory architecture and cell design are shown in Fig. 23.4.2. The 1T2R cell consists of two programmable resistors (PRs) and a shared select transistor. The PR is composed of a nitrogen-doped AlO_X based resistance-change film sandwiched between two Al electrodes [5]. The PR has a footprint of $4F^2$, resistance of GΩ in the high-resistance (HR) state, and as low as MΩ in the low-resistance (LR) state. The fabrication process of a PR is back-end compactable and scalable with CMOS. The small footprint of the PR allows for a compact configuration memory cell layout of $24F^2$ (see Fig. 23.4.2). The high on/off resistance ratio allows the two PRs (R_L and R_R) in a memory cell to form a resistive divider with a large margin for the desired logic level. The high resistance in the HR state limits the leakage current through the resistive divider. These characteristics make the 1T2R cell well-suited for FPGA configuration memory.

The operation modes of the configuration memory are illustrated in Fig. 23.4.3. The memory is organized to support the PR's bipolar switching behavior. The left bitline (LBL) and right bitline (RBL) can be connected to programming voltages, V_{DD}, or ground. The select transistor source nodes are connected to a row-based signal (RS), which can be connected to programming voltages or to ground. Because fresh PRs start in the HR state, one of the two PRs must be set to the LR state for the resistive divider to function properly. To write logic 1 into the fresh 1T2R cell at top left, R_L is set to the LR state, as shown in the program diagram in Fig. 23.4.3. After wordline is asserted and RS is connected to ground, a V_{SET} pulse is applied to LBL while RBL is left floating. The set current flows from the top to bottom of R_L. The select transistor limits the current as R_L changes from HR to LR. Similarly, to write 0 into the top right 1T2R cell, R_R is set to the LR state by floating LBL and applying the V_{SET} pulse to RBL. A written cell must be erased and restored to the fresh state before the opposite logic can be written. Resetting the PR that was in the LR state to HR state erases the cell's logic. To reset R_L in the top left 1T2R cell to the HR state, as shown in the erase diagram in Figure 23.4.3, LBL is connected to ground, and RBL is left floating. After wordline is asserted, a V_{RST} pulse is applied to RS. The reset current flows from the bottom to the top of R_L. Similarly, to reset R_R in the top right 1T2R cell, RBL is grounded and LBL is left floating. Both R_L and R_R can be read to verify the memory cell's logic state. The read mode is similar to the erase mode. A V_{RD} pulse ($< V_{RST}$) is applied to RS and the selected bitline is connected to a current mirror based sense amplifier that detects the resistance. During normal FPGA operation, all wordlines and RSs are connected to ground, and LBLs and RBLs are connected to V_{DD} and ground, respectively. This activates the resistive divider logic in each 1T2R cell. The configuration bits that drive LUTs are buffered to prevent accidental programming. Bits that configure routing switches are not buffered, however. The configuration memories are not in the signal paths during normal FPGA operation.

To demonstrate the proposed 3-D FPGA, we designed and fabricated several small prototypes in 0.18μm CMOS technology with 5 metal layers. The PRs are subsequently integrated on the CMOS wafer and a 6th metal layer is deposited to connect the PRs to the bitlines. A micrograph of a 3-D FPGA prototype with 289 logic tiles and 21Kb configuration memory is shown in Fig. 23.4.7. The FPGA core occupies 2.56mm^2. The CLB in each tile contains a 4-input LUT and each routing channel contains 16 segmented routing tracks. The top and cross-section views show the PRs on top of CMOS.

The prototype has been fully tested and characterized. Figure 23.4.4 plots the measured resistances of R_L and R_R in the memory cell demonstrating that the cell remains stable. To read the resistance after each program and erase cycle, V_{RD} is set to 1.0V. The >10 resistance ratio is sufficient for proper logic function. A lower LR state and hence a higher resistance ratio can be achieved with a larger programming current, and an endurance of 10^5 programming cycles has been demonstrated [5]. The measured resistances show that the leakage current through the resistive divider is below 0.5nA per memory cell. The V_{SET} and V_{RST} voltages are 3.0V and 1.8V respectively, and the measured I_{SET} and I_{RST} currents are <1uA. V_{SET} and V_{RST} can be reduced for an advanced technology node by thinning the resistive film. To demonstrate the FPGA operation, a 4-bit adder and a 4-bit LFSR are manually mapped into two 3D-FPGA prototypes. The adder utilizes 6 CLBs, and the LFSR utilizes 1 CLB to implement XOR logic and 4 CLBs for shift registers. The placements of both circuits and their measured delay are shown in Fig. 23.4.5. The measured maximum operating frequency of the LFSR is 250MHz while consuming 39.6mW at 1.8V supply.

To evaluate the performance gains of the 3D-FPGA in a scaled CMOS process, we designed and laid out a 2D-FPGA tile and a corresponding 3D-FPGA tile with stacked RRAM configuration memory in 65nm CMOS with 7 metal layers. Both tile designs implement a cluster of 8 LUTs per CLB and 96-track-wide segmented routing channels, which are similar to commercial FPGAs [6]. Figure 23.4.6 compares the layout area and simulated energy-delay product (EDP) of the two designs. The 3D-FPGA achieves 40% smaller area and 28% lower EDP over the baseline 2D-FPGA. When considering the compatibility and scalability of the RRAM technology, the proposed 3D-FPGA has the potential of realizing the promise of 3D integration in most advanced CMOS technologies.

Acknowledgement:
This work was supported by DARPA/SPAWAR grant N66001-04-1-8916.

References:

[1] M. Lin, et al., "Performance benefits of monolithically stacked 3D-FPGA," *Int. Symposium on Field-Programmable Gate Arrays Dig. Tech. Paper*, pp. 113-122, Feb., 2006.

[2] M. Miyamura, et al., "Programmable Cell Array Using Rewritable Solid-Electrolyte Switch Integrated in 90nm CMOS," *ISSCC Dig. Tech. Papers*, pp. 228-229, Feb., 2011.

[3] T. Naito, et al., "World's first monolithic 3D-FPGA with TFT SRAM over 90nm 9 layer Cu CMOS, " *IEEE Symposium on VLSI Technology*, pp. 219-220, June, 2010.

[4] M. Sekikawa, et al., "A Novel SPRAM (SPin-transfer torque RAM)-based Reconfigurable Logic Block for 3D-Stacked reconfigurable Spin Processor," *Intl. Electron Device Meeting Tech. Dig.*, pp. 936-937, Dec., 2008.

[5] W. Kim, et al., "Forming-Free Nitrogen-Doped AlO_X RRAM with Sub-uA Programming Current," *IEEE Symposium on VLSI Technology*, pp. 22-23, June, 2011.

[6] Xilinx "Virtex-II Pro Data Sheet," Accessed on Sept., 2011, <*www.xilinx.com/support/documentation/data_sheets/ds083.pdf*>.

978-1-4673-0377-4/12/$31.00 ©2012 IEEE

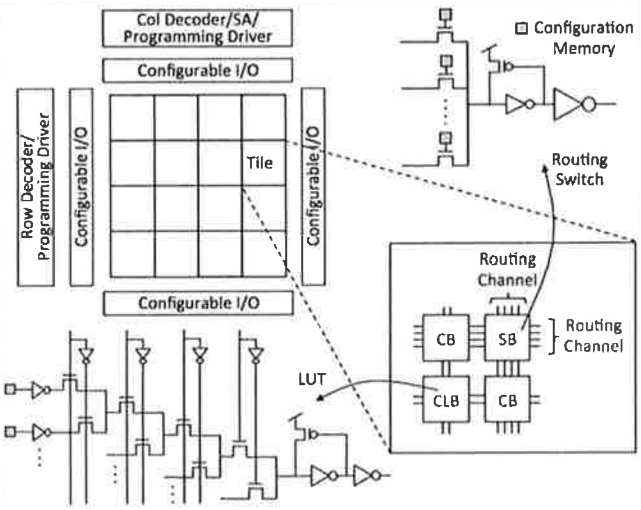


Figure 23.4.1: 3D-FPGA architecture, LUT and routing switch circuits.

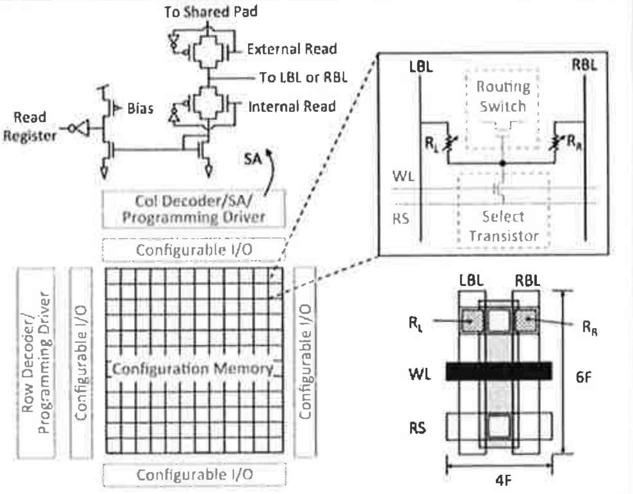


Figure 23.4.2: RRAM-based configuration memory architecture, cell design and layout. Gray schematics refer to circuits in the CMOS layer.

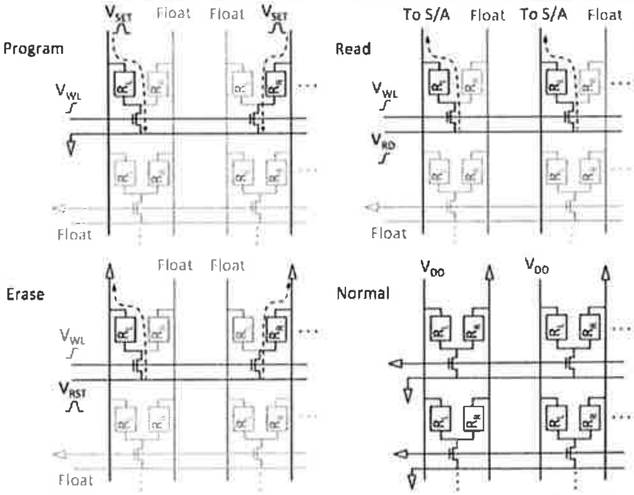


Figure 23.4.3: Schematic of four operation modes of the configuration memory.

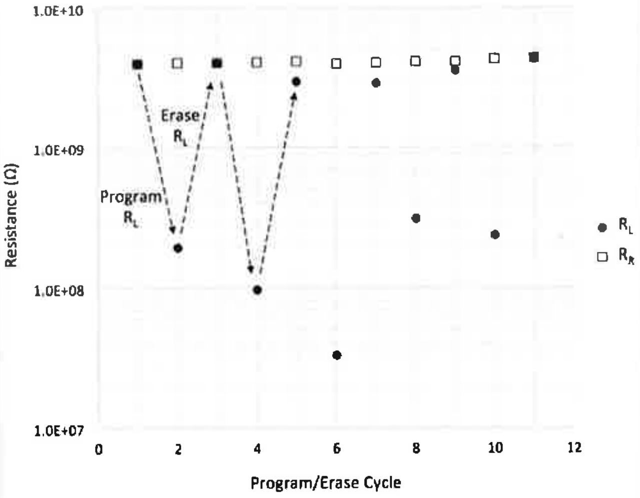


Figure 23.4.4: Measured resistance of PRs in the configuration memory.

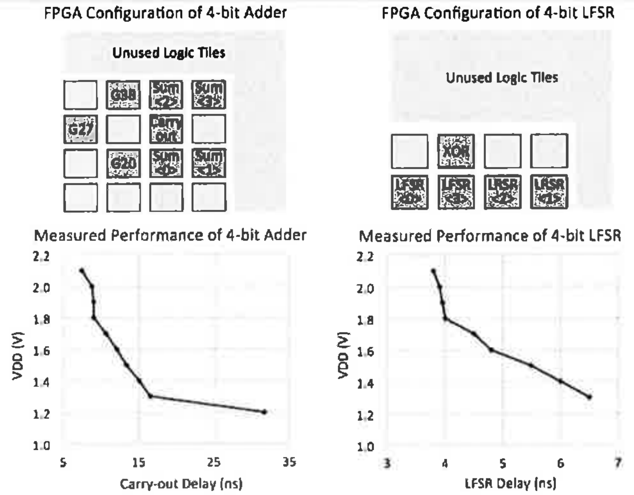
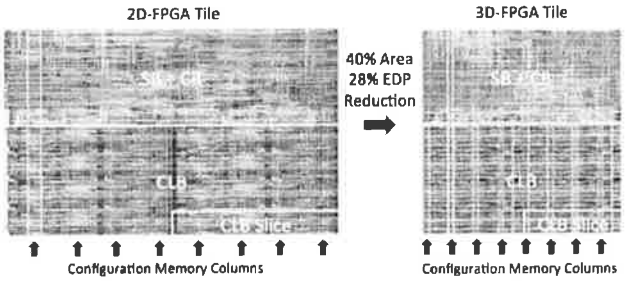


Figure 23.4.5: FPGA implementation and measured performance of a 4-bit adder and 4-bit LFSR.



	SRAM based 2D-FPGA	RRAM based 3D-FPGA
Tile Area (um²)	4x10⁴	2.4x10⁴
Tile EDP (pJ*nS)	42.9	30.8

Figure 23.4.6: Layout and simulated performance comparison between 3D-FPGA tile and baseline 2D-FPGA tile in 65nm CMOS.

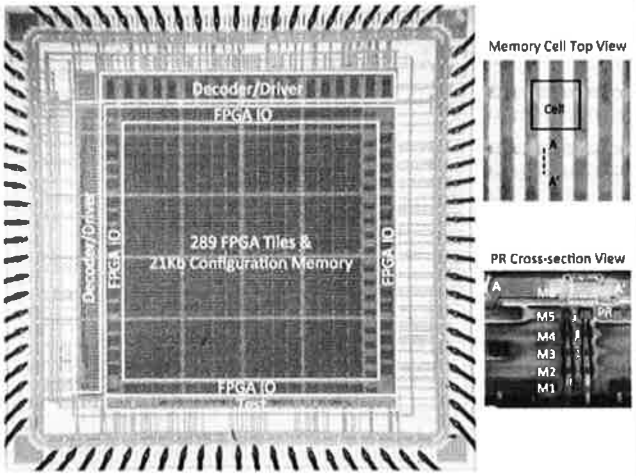


Figure 23.4.7: Chip micrographs and SEM image of a FPGA prototype in 0.18μm CMOS with PRs stacked on top.