Emerging Non-volatile Memory Technologies for Reconfigurable Architectures

Elaine Ou, Philip Leong

School of Electrical and Information Engineering,
The University of Sydney, Sydney, Australia 2006
{elaine.ou, philip.leong}@sydney.edu.au

***Abstract*— This work evaluates the potential application of emerging non-volatile memory technologies to reconfigurable architectures based on hybrid CMOS/resistive-switching FPGAs. The non-volatility of these devices lends them well to designs requiring low power consumption and reduced configuration time at power up. These memory technologies are assessed based on their effectiveness for use as interconnect routing switches in terms of programming power, reliability, scalability, and fabrication cost. The feasibility of architectural integration and innovations in reconfigurable architecture for non-volatile memories are also discussed. With sufficient redundancy and defect-tolerance, hybrid FPGA architectures may facilitate the integration of emerging non-volatile memory technologies with reconfigurable logic.**

I. Introduction

SRAM-based field-programmable gate arrays (FPGAs) have been the standard focus of reconfigurable computing in the last two decades. Some FPGAs use flash memory cells as configuration devices so that the FPGA can retain its configured state when the power is off, but flash memory has its drawbacks in terms of cost, speed, and write power requirements. Furthermore, flash is expected to reach the limits of scalability by 20 nm technology nodes [1].

Alternatives to flash memory have been widely explored for non-volatile memory applications as scaling limitations become imminent. A key difference between flash and the newer technologies is that a flash memory cell is a three-terminal active device, whereas the emerging memories discussed in this section are two-terminal passive devices that exhibit switching behavior by a change in resistance. A transistor-free cell structure best enables future scalability for high-density memory because the cell size can potentially be fabricated as small as the minimum feature size in a process technology.

Recent developments in non-volatile memory technologies may make it possible to increase the flexibility of reconfigurable devices without the limitations of flash memory cells. This paper serves as a review of emerging memory technologies and projected future developments toward their integration into reconfigurable computing architectures.

II. Architectural Considerations for Non-volatile Memory Technologies

Hybrid reconfigurable architectures employ a combination of a CMOS transistor stack and multiple levels of wires with resistance-switching memory devices serving as routing configuration at the crosspoints of nanowires or interconnect [2]. These structures typically consist of a sea-of-gates connected with wires and configurable non-volatile memory elements at the junctions, as depicted in Figure 1. At the substrate level, the field-programmable nanowire interconnect architecture implements a structure consisting of logic gates, buffers, and flip-flops to form a ”hypercell” logic structure that is analogous to a configurable logic block in an FPGA. Nanowires form interconnect above the cells [3]. More recently, logic fabrics consisting of multiplexers have also been explored [4]. In this work, we assume the fabrication of standard CMOS process metal wires laid out in the minimum metal width and pitch. The emerging memory materials discussed here can be deposited between multiple memory layers, removing the need for high-density nanowires on a single 2-dimensional plane.

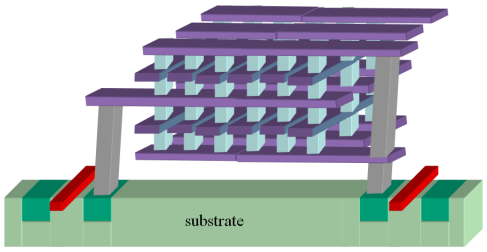

Fig. 1. Multiple levels of interconnect fabricated above a CMOS transistor stack with non-volatile configuration devices at the crosspoints.

A. Target Switch Characteristics

Signal integrity and latency are closely dependent on memory cell material characteristics in transistor-free interconnect architectures, as there are no other rectifying devices or buffers at the switchpoints.

A memory device optimized for non-volatile memory may not possess the optimal switch characteristics for a routing fabric. A non-volatile memory device is generally designed to have low-power, high-speed read and program operations. These characteristics are certainly desirable for an interconnect switch, but unlike a memory cell, an interconnect switch is not augmented by a sense amplifier. A switching device in the routing fabric should have conductive and capacitive properties that resemble those of a MOSFET in saturation and cut-off. The low-resistance state (”on” state) and cell capacitance

determine the speed of signal propagation through a switch. A low on-resistance is also critical to maintain signal integrity. The low-resistance state should be as conductive as a transistor in saturation, otherwise, additional buffers would be needed for signals passing through multiple switches. The high-resistance state ("off" state) of the memory device determines the amount of leakage current that may pass through open switches.

Devices used for non-volatile memory products need to support high endurance and long retention times, meaning they need to sustain over 10^6 write and erase cycles and maintain their data for over 10 years. These properties are not as critical in reconfigurable computing usage, because an FPGA is not usually reprogrammed over a million times in its lifetime. Low write/erase latency is not as crucial, but programming power requirements must still be taken into account. High-voltage or wide transistors for programming circuitry would increase area overhead and slow down the chip.

Defect density and device matching remain important issues when considering non-volatile switching technologies. Nanowire FPGA architectures have been demonstrated to tolerate up to a 10% defect density with some optimization [5], but variations in memory cell resistance values may make it difficult to quantify discrete errors.

III. Emerging Memory Technologies

It is not possible to completely cover the multitude of memory technologies currently pursued in active research, but we will briefly discuss the most popular memory technologies - those based on magnetic switching, resistance-change, and phase-change materials. Other emerging technologies, such as ferroelectric RAM and carbon nanotube memory, also have the potential to serve in non-volatile applications.

A. Magnetic Switching Technologies

Magnetic random-access memory, or MRAM, uses two ferromagnetic plates separated by a thin insulating layer for a memory cell. The lower plate is set to a fixed polarity, while the polarity of the upper plate is free and can be switched during a write operation. The electrical resistance of the memory cell changes depending on whether or not the polarity is aligned between the two plates. The switching accuracy is a function of the applied voltage. The write current has to be high enough to ensure devices are switched to desired states, while still remain low enough to avoid junction barrier breakdown.

Newer MRAM memories have been demonstrated to have higher-speed performance than flash, better reliability and endurance, and can be fabricated with only four additional postprocessing masks [6]. The integration of magnetic materials into a standard CMOS logic process is still a major challenge.

Magnetic tunneling junctions have been demonstrated to have a resistance-area value of approximately 5 $\Omega\mu\text{m}^2$ and up to 1056% tunneling magnetoresistance [7]. These respective resistances would increase with scaling as the memory cell area is reduced, making the switches less conductive and possibly unsuitable in the "on" state.

B. Resistance-change Memory Technologies

Resistance-change memories (RRAM) are designed for detectable changes in conductivity, similar to an MRAM cell, but typically employ a structure consisting of metal-insulator-metal to serve as the memory element. The resistance-switching is caused by the creation and destruction of conductive filaments in the material. The low resistance generally stays constant with cell size, while the high resistance increases with decreased cell size.

Metal oxides are highly compatible with modern CMOS processes and require as few as two additional mask steps for fabrication. Furthermore, they have demonstrated high-speed and low-power switching abilities [8].

A big challenge still facing resistance-change memory technology is that of wide resistance-switching distributions, as shown in Figure 2. In a resistance-switching metal-oxide, the position of the conductive filament formation may change each programming cycle, resulting in some resistance variation.

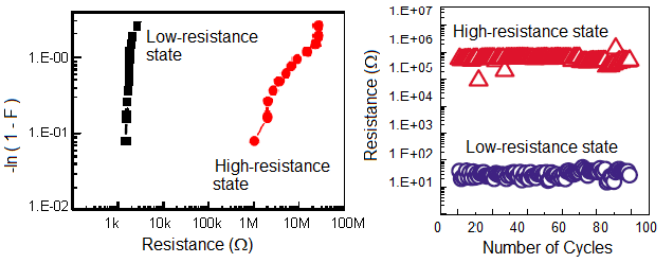

Fig. 2. Left: Weibull plot showing probability density of high and low resistance ranges [8]. Right: Erasing (setting a device to a high-resistance state) yield as a function of the select transistor gate voltage to control current compliance for Cu_2O memory [9].

C. Phase-change Memory Technologies

Phase-change memory operates by using a resistive heater to melt and recrystallize a chalcogenide memory cell in either a crystalline (low-resistance) or amorphous (high-resistance) state. The material is melted using current pulses, and the programmed state of the memory cell depends on the rate at which the current decreases in the falling edge of the pulse.

Recent work has demonstrated the successful integration of phase-change material in 90nm standard CMOS technology, with a factor of 25 difference in conductivity between the high- and low-resistance cells [10]. Phase-change memory cells scale favorably, with reduced program power consumption and increased operation speed with smaller sizes. The stability of the programmed state is expected to deteriorate at the nanoscale, but successful implementations have been demonstrated down to 45 nm [11].

Phase-change memories have shown sufficient endurance and reliability to serve in commercial products as random-access non-volatile memory. It has received much attention as a future replacement for NOR flash because of its scalability. However, thermal disturbance effects become a significant consideration in densely packed arrays, where using a heater

to program adjacent cells could cause a change in the high-resistance state as well as the threshold switching voltage [12]

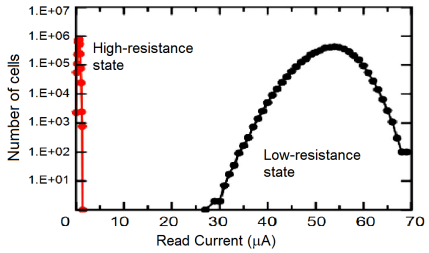

Fig. 3. Read current distributions indicating the spread of low-resistance and high-resistance cells collected over a 4Mb array. [10].

D. Memory Technology Summary

Representative memory cell attributes for each of the memory technologies are summarized in Table I. Even within the three emerging memory technologies surveyed, there is a wide range of materials and structures under development, and the parameters in the table only serve as a sample of the memory devices that may be suitable for reconfigurable computing.

TABLE I
DEMONSTRATED MEMORY MATERIAL CHARACTERISTICS.

Switch Technology	R_{ON}	R_{OFF}	Prog. Current
65nm NMOS	40 Ω	1 MΩ	n/a
90 nm STT MRAM [7]	6 kΩ	60 kΩ	200 μA
180 nm HfO_2 RRAM [8]	1 kΩ	1 MΩ	25 μA
1 μm CuO RRAM [9]	30 Ω	600 kΩ	1 mA
90 nm Phase-change [10]	3 kΩ	900 kΩ	250 μA

Metal-oxide resistance change materials and phase-change materials are particularly attractive due to their scalability and adaptability to CMOS logic processes. Metal oxides in particular have the lowest programming current requirements and thus transistors for write circuitry can be sized more conservatively. Furthermore, their resistance values come closest to matching an actual transistor switch.

IV. DISCUSSION

Aside from memory cell material characteristics, the overall feasibility of integrating a memory material into a hybrid FPGA architecture needs to be evaluated. Table II summarizes the higher-level attributes associated with each memory technology.

A. Defect Tolerance and Reliability

Defects in non-volatile memory switches may be exhibited as a non-programmable cross-point defect, or an incompletely programmed switch that either causes an excessive voltage drop across an "on" switch or allows signal leakage through an "off" switch. Non-programmable defects can be dealt with simply by provisioning adequate spares and implementing a matching strategy at configuration time [5]. Incompletely

TABLE II
SUMMARY OF MEMORY TECHNOLOGY ATTRIBUTES [1].

Switch Technology	Scalability	Endurance	Retention	Add'l Masks
SRAM	8 nm	∞	power-off	0
MRAM	16 nm	10^{16}	10 years	4
Resistance-change	<10nm	10^{6}	10 years	2
Phase-change	<10nm	10^{9}	300 years	3

programmed switches are more difficult to quantify in terms of bit-error rate. Device-level simulations will need to be performed to verify the tolerability of varying resistance levels. For the most part, metal-oxide resistance-change materials are most desirable because they have the lowest low-resistance states, and variations tend to be evidenced as higher resistance values in the "off" state.

Phase-change memory and MRAM have demonstrated sufficient reliability to successfully serve in commercial non-volatile memory applications. Each memory technology has particular susceptibility to different failure mechanisms that could be more apparent in a densely-packed hybrid architecture. Phase-change and resistance-change memory are susceptible to incomplete programming due to current and voltage variations, which could be problematic as programming individual cells in a cross-point array can cause disturbances to adjacent cells. Temperature-induced variation is also a significant issue for phase-change memory because switches are placed close together in a hybrid architecture. MRAM is more likely to see complete device failure and switches can be disturbed by stray fields.

Given the switching mechanisms described for each of the non-volatile memory technologies, we do not expect lifetime defects to become an issue. The non-volatile memory technologies described in the previous section all have demonstrated retention times of 10 years or longer and can endure at least 10^6 write/erase cycles.

B. Area, Cost, and Scalability

Not only can the alternative memory technologies be fabricated in a much smaller area than SRAM switches due to their lack of transistors, but a hybrid reconfigurable architecture provides more efficient use of silicon by moving routing and interconnect resources from the area between logic modules to the metal layers above the substrate. Thus, logic blocks can be immediately adjacent, resulting in a much smaller die size. The main cost benefit comes when considering the fact that SRAM-based FPGAs do not retain their data and configuration data must be reloaded to the SRAM through an external non-volatile memory device. External memory is no longer necessary when implementing integrated non-volatile memory throughout the FPGA. Absolute fabrication costs are difficult to estimate as the emerging non-volatile memory technologies are not yet in mass production. But because they are not built

in the substrate, they can be combined with back-end-of-the-line processing. Resistance-change metal oxides require the fewest additional mask steps for fabrication.

The write voltage of MRAM, phase-change, and resistance-change memories are all expected to scale down with more advanced technology processes. While SRAM may face reliability challenges with scaling due to the increasing rate of single-event upsets (SEUs) [14], the emerging memory technologies do not use electron-based storage mechanisms and are not susceptible to radiation-induced errors. Phase-change and metal-oxide resistance-change memories are expected to be functional beyond 10 nm process technologies [1].

C. Performance Considerations

The biggest factor affecting on-chip latency will be the interconnect capacitance. The small size of the non-volatile memory switches can lead to dense interconnect and high parasitic capacitance. In a 65-nm process technology, the lateral parasitic capacitance between two metal lines fabricated at the minimum pitch is 0.148 fF/μm, and the vertical parasitic capacitance between interconnect is 0.0214 fF/μm. Segmentation of routing resources can be used to minimize the effects of interconnect capacitance by allowing for unused routing wires to be disconnected [13]. Alternately, it may not be necessary or even desirable to lay out interconnect at the minimum metal pitch.

D. Tools and Architectural Development

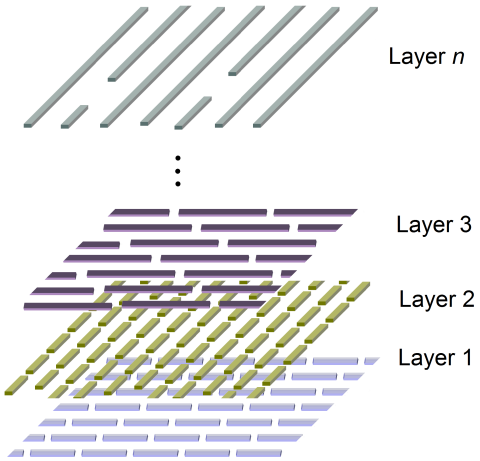

Fig. 4. Hierarchical 3D interconnect structure.

When considering the multilayer interconnect of a hybrid 3D FPGA, a number of architectural modifications can be performed to optimize defect tolerance. Current non-volatile memory materials do not yet exhibit ideal characteristics for implementation in reconfigurable interconnect, but designing hierarchical interconnect where upper layer wires have sparser arrays can provide longer connections and improve routing efficiency. Lower interconnect layers should be partitioned into electrically isolated segments as they will have denser switch placement, and short wire lengths will avoid leakage current problems presented by insufficiently high resistance when bypassing "off" switches. This is depicted in Figure 4.

Current placement and routing tools can be adapted to map logic as a sea-of-gates onto a hybrid architecture while implementing redundancy and signal buffering as needed. Because current non-volatile memory technologies cannot yet match an actual transistor in terms of resistance-switching characteristics, mapping optimization should minimize the routing of signals past switches in the high-resistance state and avoid segmentation. Tools for mapping logic to accommodate non-ideal switch characteristics are currently under development. Future work will also involve the derivation of analogous circuit design parameters to provide metrics for comparing hybrid architectures with traditional island-style layouts.

V. Conclusion

Different non-volatile memory technologies have been evaluated in terms of their potential application to a hybrid reconfigurable architecture. Hybrid FPGA architectures can be cost-effective as the routing interconnect and configuration switches can be fabricated above the logic blocks. The lower endurance requirements and higher defect tolerance can allow reconfigurable computing architectures to support emerging memory technologies that may not yet be well-developed enough for commercial memory applications.

References

[1] *ITRS Roadmap.* Available: http://www.itrs.net/

[2] A. Dehon. "Nanowire-Based Programmable Architectures," *ACM Journ. on Emerging Tech. in Computing Systems*, vol. 1, pp. 109-162, July 2005.

[3] G. Snider and R. S. Williams, "Nano/CMOS architectures using a field-programmable nanowire interconnect", *Nanotechnology*, vol. 18, art. 035204, 2007.

[4] D. Strukov and A. Mishchenko, "Monolithically Stackable Hybrid FPGA," *Proc. Conf. on Design, Automation, Test Europe*, vol. 16, pp. 888-900, 2010.

[5] A. DeHon and H. Naeimi, "Seven strategies for tolerating highly defective fabrication," *IEEE Design & Test of Computers*, vol. 22, pp. 306-315, 2005.

[6] W. Gallagher and S. Parkin, "Development of the magnetic tunnel junction MRAM at IBM: from first junctions to a 16-Mb MRAM demonstrator chip," *IBM J. of Research and Development*, vol. 50, pp. 5-23, 2006.

[7] T. Min, et al. "A Study of Write Margin of Spin Torque Transfer Magnetic Random Access Memory," *IEEE Trans. on Magnetics*, vol. 46, pp.2322-2327, 2010.

[8] H. Y. Lee, et al. "Low power and high speed bipolar switching with a thin reactive Ti buffer layer in robust HfO2 based RRAM," *Proc. Intl Electron Devices Meeting*, pp. 1-4, 2008.

[9] C. H. Kim, et al. "Observation of bistable resistance memory switching in CuO thin films," *Appl. Phys. Letters*, vol. 94, article 102107, 2009.

[10] G. De Sandre, et al. "A 4Mb LV MOS-Selected Embedded Phase Change Memory in 90 nm Standard CMOS Technology," *IEEE J. of Solid-State Circuits*, vol. 46, pp. 52-63, 2011.

[11] H.-S. Wong, et al. "Recent Progress of Phase Change Memory (PCM) and Resistive Switching Random Access Memory (RRAM)," *IEEE Intl. Conf. on Solid-State and Integrated Circuit Tech.*, pp.1055-1060, 2010.

[12] S.-B. Kim, et al. "Thermal Disturbance and its Impact on Reliability of Phase-Change Memory Studied by the Micro-Thermal Stage," *IEEE Intl Reliability Phys. Symp*, pp. 99-103, 2010.

[13] V. Betz and J. Rose, "FPGA Routing Architecture: Segmentation and Buffering to Optimize Speed and Density," *ACM/SIGDA Intl Symp on Field Programmable Gate Arrays*, pp. 59 - 68,1999.

[14] R. C. Baumann, "Radiation-induced soft errors in advanced semiconductor technologies," *IEEE Trans. Device and Materials Reliability*, vol. 5, pp. 305-316, 2005.