ABSTRACT

Advances in silicon technology have enabled the semiconductor industry to shrink the area of a transistor by roughly a factor of two with each successive technology node. This trend has continued unabated for the past five decades and has made computing devices ubiquitous in modern culture. Due to continuous advances in CMOS technology fueled by a growing and fiercely competitive market, there have been sustained innovations in the field of memory devices and circuits. This has led to a very advanced and hybrid computer memory hierarchy. Leading technology reports have shown the growing importance of memory in overall system architecture. Therefore, development in optimizing memory designs and schemes has been a primary criterion for computer architecture engineers. SRAM is typically used to realize memory due to its high speed of operation. Unlike DRAM, where the data is stored as a charge on the capacitor and needs constant refreshing, SRAM uses digital logic to store the data in its internal node. The requirement of a refresh scheme gives rise to additional circuitry and timing requirements, leading to complicated modules for DRAM memory. On the other hand, SRAM modules are more straightforward compared to DRAM, which makes it easier for design engineers to create an interface to access the memory.

However, almost all implementations of conventional SRAM circuits in literature are volatile. SRAM circuits lose data during the power-down state and thus need a hold voltage to maintain the memory state. This results in significant leakage power and energy dissipation. Various schemes, such as power gating, has been proposed to mitigate such issues. In power gating, a low voltage (hold voltage) is used for volatile memory to retain data while all logic circuits are turned off. However, even maintenance of this hold voltage (during power-down mode) in high-performance processing units leads to colossal power dissipation. Moreover, such schemes add more on-chip area and the supply scaling affects the stability performance of the bit-cells. Even worse, during unexpected power failures, the data in volatile memory may be lost, and computation tasks restart. Thus, it is essential to introduce non-volatility in these memories. Traditional non-volatile schemes use a backup memory block (Flash), where data is off-loaded to it from SRAM block when a power-down signal is applied. However, this technique degrades memory performance, as more data transfers are required. Multiple circuits have been proposed in the literature to backup data from on-chip memory (SRAM), FFs and register to off-chip non-

volatile memory (NVM) such as PCM, OxRAM, MRAM, thus preserving the system state in case of power failures. OxRAM devices have several advantages over other existing non-volatile memory solutions, such as low fabrication cost, fabrication at the back-end, manufacturing in via along with its property of being non-volatile.

All the NVSRAM designs proposed in the literature are based *‘last bit non-volatility’* and require a controlled power-down signal to manage the data off-loading. This increase routing congestion and sophisticated control signal mechanisms. In this thesis, we present a `real-time non-volatile' SRAM (NVSRAM) cell based on oxide-based random access memory (OxRAM) devices. This thesis explores the programming scheme of NVSRAM, optimizing the cell for faster and power-efficient performance, analyzing the stability of the cell which will determine the reliability of the memory design, advantages when using the proposed NVSRAM at a system level, and studies the implications on performance parameters when NVSRAM is used in last level cache. The thesis also covers the electrical characterization and modeling of various OxRAM devices to understand the device behavior under different operating conditions.