Design of 1Mbit RRAM memory to replace eFlash

Wouter Diels and Alexander Standaert
Department of Electrical Engineering, Katholieke Universiteit Leuven
3001 Heverlee, Belgium

***Abstract*—A 1Mbit RRAM memory robust against variations in 45nm technology is presented. The focus lies on read reliability. To overcome variability a tuned reference signal is generated by connecting multiple reference cells in parallel. A bitline load has been designed to obtain maximum bitline voltage difference. Sense amplifier performance has been improved by allowing overlap between passgate-enable and latch-enable signals, this overlap gives rise to a nonlinear phenomenon, the RC-latch-effect. Write operation has not been included in the design and the results are based on circuit simulations.**

I. Introduction

NON-volatile memories such as flash are widely used for mass storage devices, but are also steadily finding their way into the embedded domain. However, as discussed in [1], it is getting difficult to fabricate reliable flash memories in deep-submicron. It is argued that the scaling of flash-memories will not last for more than a few technology nodes. RRAM memories, in which information is stored in the resistive state of a memristor, would be able to scale futher, due to the fact that it is a nanoparticle based memory[2]. Furthermore, the memristor fabrication can easily be integrated in a standard CMOS fabrication process. In this work, a RRAM memory has been designed, armed against intra-die variations. A read access time of 2.3ns and an energy consumption of 0.51pJ per bit access make it faster and less energy greedy than conventional flash memories. This paper is structured as follows: Section II discusses the general memory architecture. In section III, the tuning of the reference voltage is explained. Section IV presents the results of the load analysis, in which an optimal load impedance is chosen for sufficiently large voltage differences for the sense amplifier and sufficiently low voltage drops over the memristor. In section V, some techniques for decreasing the offset voltage of the sense amplifier will be explained. Finally, section VI summarizes the results of the complete memory.

II. General Architecture

The memory size was chosen to be 1 Mbit, with a 1 bit/word readout. Although in practice larger wordsizes are usually implemented, the focus on this work lied primarily in optimizing single bit readout rather than multi bit word architecture optimization.

The general architecture can be seen in figure 1. The memory consists of 512 global blocks (GB). Each GB consists of two local blocks (LB), which in turn consists of 32 bitlines (BL) and sourcelines (SL) and 32 wordlines (WL). A branch is defined as a collection of memory cells connected to one BL and one SL.

A. Branch

In a branch 32 1T1R cells are connected to a BL and a SL. The transistor terminal of the cell is connected to the SL and the memristor terminal of the cell is connected to the BL. The memristors can be either in high or low resistive state (HRS or LRS). Besides these 32 data cells, there is also one reference cell in the branch, its gate is connected to the reference WL. Connected to the BL is a load. The load consists of a single pMOS transistor that is connected to the supply voltage, which is switched on and off by means of digital logic. Also connected to the BL is a nMos transistor that serves as a switch to the ground voltage. An nMOS switch is also placed between the SL and the ground voltage. This switch is needed if a write circuit is implemented so that the current can also flow in reverse direction though the memristor.

B. Local block

A local block (LB) consists of 32 branches, BL & WL decoder and passgates on the BL of each branch (BL mux). The passgates are connected to the output node LBout. To read

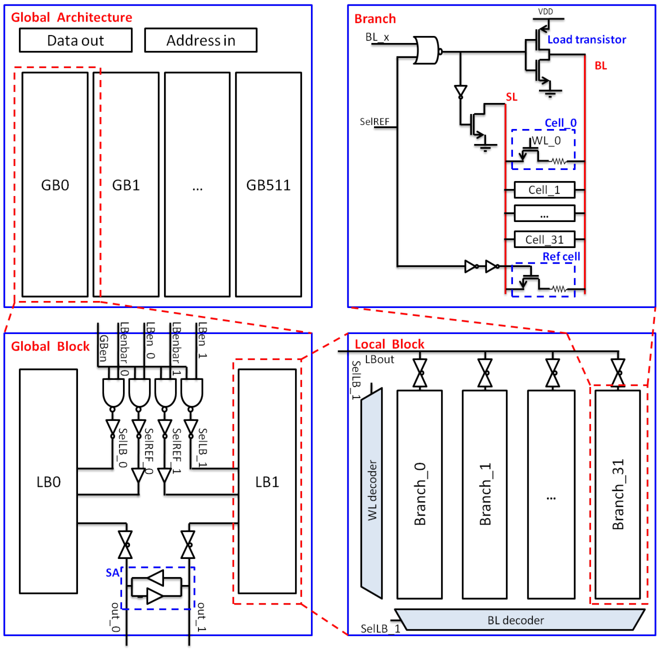

Fig. 1: Overview architecture

out a data cell, the appropriate WL is brought to the supply voltage. The BL-load and SL-transistor of the appropriate BL and SL are switched on. A current will flow from the supply voltage through the load, the cell and SL-switch to the ground voltage and a voltage will appear on the BL node. The passgate of the BL is then turned on and the BL voltage is passed to the output of the LB. Reference signals are generated by activating the reference WL and turning the BL loads and SL switches on. The BLs are then shorted by turning on all the appropriate passgates. This averages the BL voltages to a signal which has a value between the BL voltage of a cell in HRS and LRS. To save energy, not all 32 reference cells in a LB are used to generate the reference signal.

C. Global block

A global block (GB) consists out of two LBs and a sense amplifier (SA). If one LB produces a data signal at its output, the other will produce a reference signal and vice versa. While reading a cell in the memory circuit, only one GB is active. This is guaranteed by the GBen signal. After the SA amplifies the small voltage difference, the result is buffered and transferred out of the memory (not shown on figure).

III. Tuning the reference signal distribution

Due to intra-die variations, the properties of different components in the circuit vary. In this work, variations of the transistor parameters V_T and β are considered and modeled with normal distributions. The variations of the resistive value of the memristor are also considered, normal distributions are used to model both HRS and LRS. Due to these component variations, signals such as the data and reference signal also have a distribution. The distribution of the reference signal however, can be tuned unlike the distributions of the data signal. Recall that the reference signal is generated by shorting active BLs using passgates. Shorting a BL with an addressed HRS reference cell with a BL with a LRS reference cell would suffice for producing a voltage lying between a HRS data voltage and a LS data voltage. By using this shorting technique however the mean of the reference signal PDF would not lie exactly between the means of the HRS data PDF and LRS data PDF. By implementing more than 2 reference cells for the reference signal, and having more HRS (LRS) cells than LRS (HRS) cells, the mean of the reference signal PDF can be shifted. Furthermore, the distribution will have a smaller spread by implementing a bigger amount of reference cells. One should not implement too many reference cells however, since energy consumption (for each active reference cell, current flows through its corresponding bitline) increases drastically. In this design 16 reference cells in a LB are addressed for generating the reference signal, the remaining 16 serve as dummies. Of the 16 active reference cells, 6 are HRS and 10 are LRS. This produces sufficient voltage margins between data and reference signals (figure 3) and even so below nominal voltage supplies as will be shown in section VI.

IV. Load analysis

Due to the aforementioned variations in the circuit, the data and reference signals should be designed in such a way that they are sufficiently far apart. The minimal distance of these two signals directly determines the maximum offset voltage the SA can have. Moreover, the distributions of these signals cannot overlap or a correct reading of the cell is impossible. The load impedance influences the value of the data and reference voltages, but it also determines the settling time of the charging of the BL and the voltage drop over the memristor. Destructive reads might occur because of a too high voltage drop over the memristor. As it turns out, fast settling is not compatible with low memristor voltage drop and large voltage difference. Because the latter are imperative for a reliable memory and the former is not, settling time had no bearing on the final choice of load impedance. Several different types of load were considered [3] but the best performance against variability was achieved with a single transistor load. Figure 2 shows nominal HRS/LRS voltage differences and memristor voltage drop for several pMOS transistor loads with different sizes.

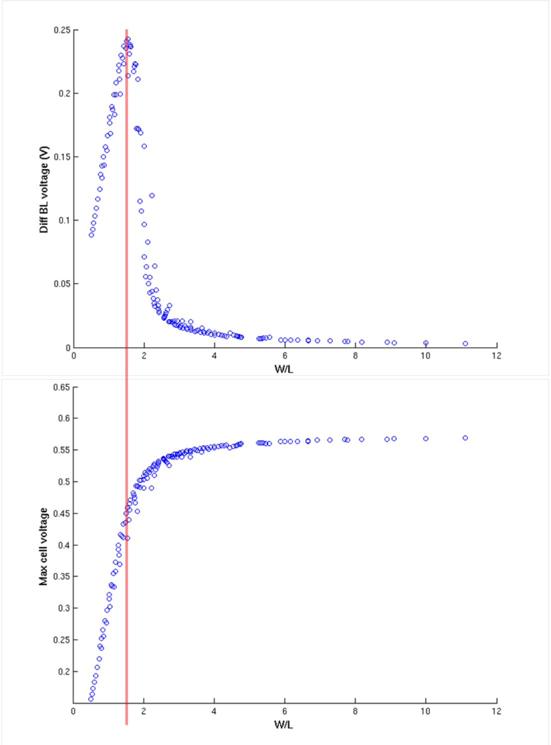

Fig. 2: Nominal HRS/LRS BL voltage difference and voltage drop over memristor for pMOS transistor loads

In the end a transistor with a width of 300nm and a length of 198nm was chosen. In figure 3 the distributions of the data and reference voltages are displayed for this load.

V. Sense amplifier overlap techniques

This design uses the drain-input latch-type SA (see figure 4). Its input/output nodes are connected to the output nodes

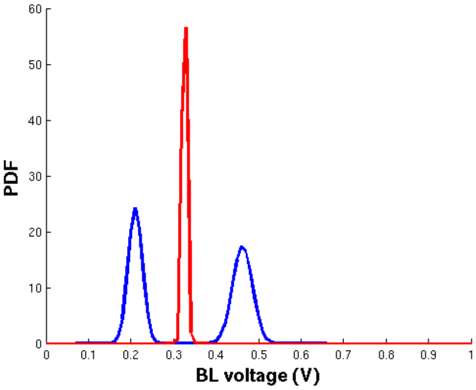

Fig. 3: Distributions of the reference signal and HRS & LRS data signals

of the local block through complementary passgates. There are two ways to implement the latch timing cycle: one could separate the pass operation from the latch operation or could allow overlap between these two operations.

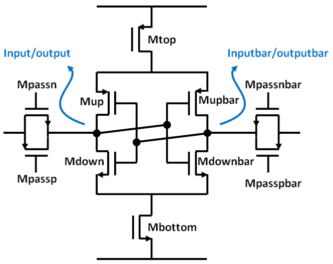

Fig. 4: Drain-input latch-type sense amplifier[4] used in this design

A. No overlap: disconnect inputs before activating SA current source

The SA can be activated after the passgates are disconnected. Using this control scheme, the SA would be separated from the local block when enabled. Offset voltage spread is mainly determined by the Δ_{V_T} and Δ_{β} variations of the differential pairs of the SA. These contributions can be decreased by increasing the sizes of the differential pair transistors. Sizing up the top and bottom transistor has more influence on the latching speed rather than the offset voltage. There is also a slightly surprising contribution on the offset voltage by the passgates. This can be explained by the charge injection of the passgates: when the passgates are turned on, the output voltage of the local block is passed on to the input/output node of the SA almost perfectly - whether there are variations on the transistors or not. When the passgates are turned off, a charge injection occurs on the SA input/output node - distorting the original voltage. The SA operates differentially though, so as long as this charge injection is matched at the two input/output nodes there would be no problem. β mismatch of the passgate transistors however results in charge injection mismatch (see figure 5). Hence the contribution of the passgates to the offset voltage spread. This mismatch can be reduced by sizing up the passgates transistors. More charge is injected when the passgates are turned off, but the distortion difference at the two sides is reduced. This reduces the offset contribution of the passtransistors.

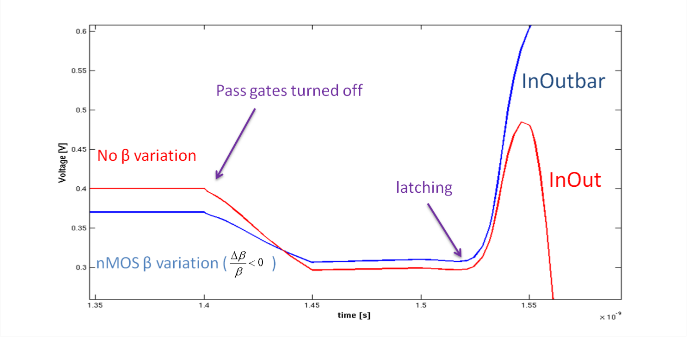

Fig. 5: Charge injection mismatch due to β variations of passgate transistors causes incorrect latching

B. Overlap: inputs are connected when SA current source is already activated

If the passgates are still turned on when the SA starts latching, the voltage difference of the input/output nodes has not experienced charge injection mismatch. When the voltage difference on the SA internal nodes has been sufficiently amplified, the passgates are turned off. Charge injection mismatch will occur, but it will not change the outcome of the latching anymore. During the overlap of the pass and latching operation, the passgates can be modeled by resistors. Neglecting the capacitance of the output node (LBout) of the local block, the situation can be depicted as in figure 6. CL is the bitline capacitance.

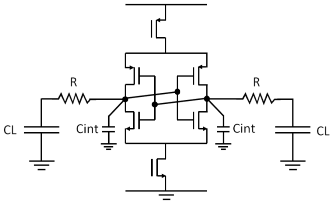

Fig. 6: Simplified circuit when overlap between pass enable and latch enable is applied

One could suspect that a large BL capacitance would significantly increase the latching time. While this is true for small values of R, as can be seen in figure 7, for greater values the latching goes through two phases: during the first phase, it appears as if the big capacitance is decoupled from the SA. After this fast phase, settling is much slower.

This RC-latch effect arises when the RC-product of the pass-gate resistance and the BL capacitance is large. When it occurs, the capacitance behaves as a short-circuit and current flows through the resistor and a corresponding voltage drop

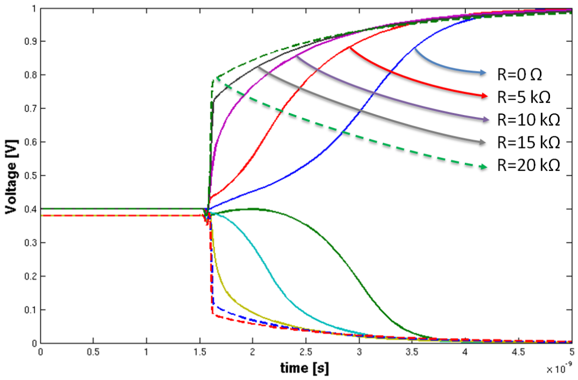

Fig. 7: Transient simulation for the schematic of figure 6 for different values of R, CL = 46fF, SA is minimal, no variations are included in simulation

over the resistor builds up. Afterwards the load capacitance charges itself at a much lower time constant.

The overlap between the enabling of the passgates and the SA thus needn't be this great: a load-less SA latching delay suffices. For this situation V_T and β mismatch of the passgate transistors results in mismatch of the resistors. The interaction between these resistors and the SA is strongly non-linear. Depending on the precharged values of the load capacitances, this R mismatch can result in incorrect latching. There is thus still a contribution of the passgates to the offset voltage spread. The interaction between R and the SA reduces the contributions of the differential pair mismatch however. For a minimal SA, the offset voltage spread is smaller with overlap than without. By sizing all the transistors, this spread can be reduced to the desired level. Speed can be improved by using the RC-latch effect which implies using small pass-gate transistors, or by sizing the rest of the SA. This allows the SA to deliver a large current to charge the BL capacitance.

VI. Read Throughput Results

A read access time simulation test was performed on the complete circuit (Figure 8). In this test the supply voltage was decreased and the maximum read throughput was determined. The test was performed at a (SPICE environment) temperature of 30°C. At each point in the shmoo plot, 100 Monte Carlo simulations were performed. At 1V Vdd, the circuit achieves a read access time of 2.3ns. during the read cycle, its energy consumption is 0.51pJ. Most of the energy (65%) is consumed by the bitlines. 25% of the energy consumption is due to the logic, while the buffers and SA together take the remaining 10%. As the supply voltage goes down, the read throughput also goes down. This is caused by a combination of phenomena. First of all the logic operates slower. This causes a timing issue in which the bitline load transistor is activated before the cell. This timing issue results in a rapid increase in bitline voltage which causes a longer settling time of the bitline when the cell is eventually selected. A second reason for the decrease in read throughput at a lower supply voltage, is the slower latching of the SA.

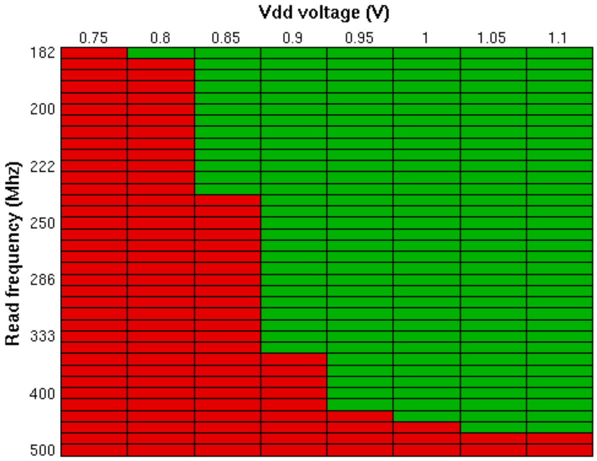

Fig. 8: Results read access time test. Green succesful simulation. Red unsuccesful simulation

On figure 8 it can clearly be seen that the circuit can only operate down to a supply voltage of $0.8V$. An explanation for this is given in figure 9. As vdd is decreased, the bitline voltage distributions of the reference signal and HRS & LRS data signals also decrease. Eventually the bitline voltage distributions overlap and there is a high probability of circuit failure.

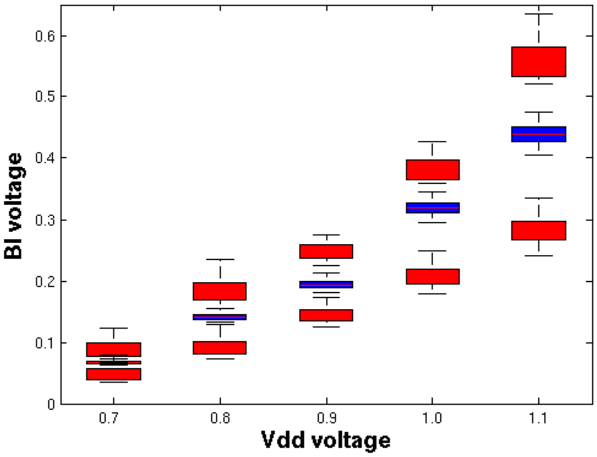

Fig. 9: BL voltage distribution for reference signal and HRS & LRS signals in function of vdd

VII. Conclusion

A 1Mb RRAM memory has been designed and presented. This memory design employs several techniques to improve performance and reliability. A good reference voltage distribution was achieved by averaging the signal from 10LRS and 6HRS cells. A load impedance was carefully chosen to maximize HRS & LRS voltage differences while keeping the memristor voltage drop sufficiently low to avoid resistive switching and thus destructive reading. Overlap between passgates and sense amplifier operation reduces offset voltage spread as it avoids differential charge injection. Although keeping the passgates enabled while activating the SA does mean current flows from SA to BL, the impact of this on SA speed and energy is small, as Rpassgate*CBL is much larger than the overlap time.

Acknowledgment

The authors would like to thank Prof. Wim Dehaene for his guidance of this thesis, Stefan Cosemans for his assistence and Prof. Patrick Reynaert, Prof. Michiel Steyaert and Prof. Filip Tavernier for the opportunity to do a PhD at the KULeuven.

References

[1] K. Prall and K. Parat, “25nm 64gb mlc nand technology and scaling challenges invited paper,” in *Electron Devices Meeting (IEDM), 2010 IEEE International*, Dec 2010, pp. 5.2.1–5.2.4.

[2] H. Nazarian, “Crossbar resistive memory: The future technology for nand flash,” *white paper from crossbar*, 2013.

[3] F. Ren, H. Park, R. Dorrance, Y. Toriyama, C.-K. Yang, and D. Markovic, “A body-voltage-sensing-based short pulse reading circuit for spin-torque transfer rams (stt-rams),” in *Quality Electronic Design (ISQED), 2012 13th International Symposium on*, March 2012, pp. 275–282.

[4] S. Cosemans, “Variability-aware design of low power sram memories,” Ph.D. dissertation, KULeuven, 2010.