Design of 1Mbit RRAM memory for NV applications

Diels Wouter, Standaert Alexander

***Abstract*—A 1Mbit RRAM memory is presented. The focus lies on read operation, write operation has not been included in the design. Reference signal distribution can be modified by connecting reference cells in parallel. A bitline load has been chosen for 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, christened the RC-latch-effect. All Spectre simulations have been performed with 45nm PTM transistor models.**

May 20, 2014

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 DSM. 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 indefinitely for now. 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 which could degrade performance. First the general architecture, the way the 1T1R cells are put together, will be described. In the following section, load analysis results will be presented, 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. Afterwards, the tuning of the reference voltage will be explained. Finally, some techniques for decreasing the offset voltage of the sense amplifier will be explained.

II. General architecture

The general architecture can be seen in figure 1. The memory consist of 512 global blocks (GB). Each GB consists of two local blocks (LB), in which 32 bitlines (BL) and sourcelines (SL) and 32 wordlines (WL) are embedded.

A. Branch

In a branch 32 WLs are connected to as many 1T1R cells through the transistorgates. The memristors can be either HRS or LRS. The remaining memristor terminal is connected to a BL and the source of the transistor is connected to a SL. Besides these 32 data cells connected to 32 WL, there is also one reference cell in the branch, its gate is connected to the reference WL. At the top of the BL, the drain of a pMOS transistor is connected, its source is connected to the supply voltage. This transistor serves as a switchable load impedance. At the bottom of the BL, an nMOS transistor serves as switch to the ground voltage. An nMOS switch is also placed between the SL and the ground voltage.

B. Local block

A local block (LB) consists of 32 branches combined as well as a BL & WL decoder and passgates on the BL of each branch. The passgates are connected to the output node of the LB. To read out a data cell in a LB, the appropriate WL is brought to the supply voltage and the cell's BL-load/switch and SL-switch are turned on. A current will flow from the supply voltage through the load and cell (and SL-switch) to the ground voltage and a voltage will appear on the BL node. This voltage is passed to the output of the LB by turning on the passgate of that BL. Reference signals are generated by bringing the reference WL to the supply voltage and turning certain BL-load/switches and SL-switches on. The BLs are then shorted by turning on all the appropriate passgates.

C. Global block

Two local blocks are brought together with a sense amplifier and its sample-and-hold switch in a global block (GB). If one LB produces a data signal at its output, the other will produce a reference signal and vice versa.

III. Tuning the reference signal distribution

It is assumed that certain elementary variables of components in the circuit have a normal distribution due to intra-die variations. These variables include the Δ_{V_T} and Δ_β parameters of transistors and the Δ_R parameter for the memristor. Due to these elementary variations, signals such as the data and reference signal also have a distribution. The distribution of the reference signal however, can be tuned. 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

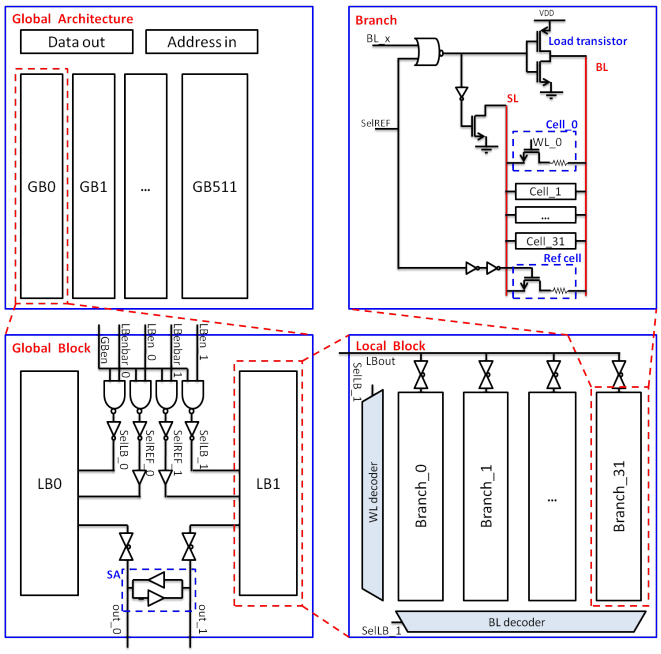

Fig. 1: Overview architecture

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.

IV. Load analysis

Due to the aforementioned variations in the circuit, the data and reference signals should be designed as such that they are sufficiently far apart. After all the sense amplifier will have an offset voltage because of these same variations and the difference of its inputs must be larger than this offset in order for the SA to latch correctly. Besides designing the SA to have a small offset spread to realize this, the difference of its inputs can be widened by choosing a good load impedance. The load impedance not only influences the value of the data and reference voltages, it also determines the settling time of the charging of the BL and the voltage drop over the memristor. This drop can not be too high, destructive reads might occur because of this. 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 functional memory and the former is not, settling time had no bearing on the final choice of load impedance.

V. Sense amplifier overlap techniques

The sense amplifier used in this design is the drain-input latch-type SA (see figure 2). Its input/output nodes are connected to the output nodes 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.

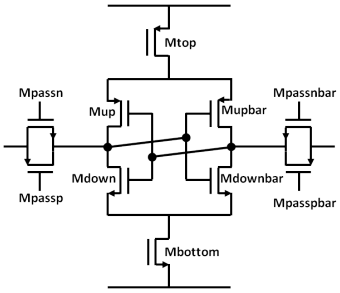

Fig. 2: Drain-input latch-type sense amplifier used in this design

A. No overlap

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 not more influence on the latching speed rather than the offset voltage. There is also a maybe 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 3). 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 difference at the two sides is reduced.

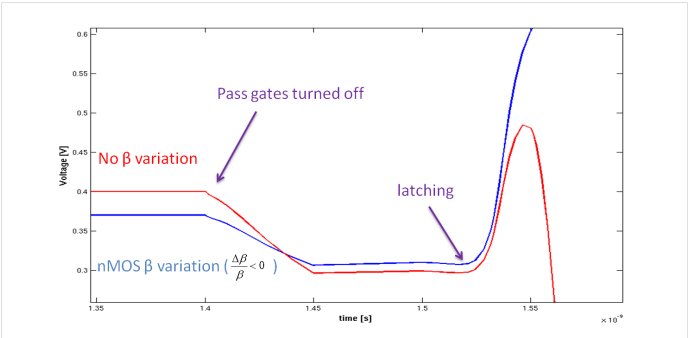

Fig. 3: Charge injection mismatch due to β variations of passgate transistors

B. Overlap

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 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 resistance of the passgates in the local blocks, the situation can be depicted as in figure 4. CL is the bitline capacitance.

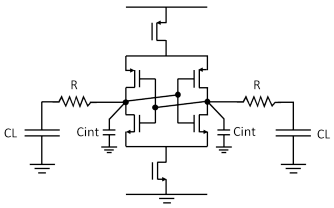

Fig. 4: Simplified circuit when overlap between pass enable and latch enable is applied

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 5, 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.

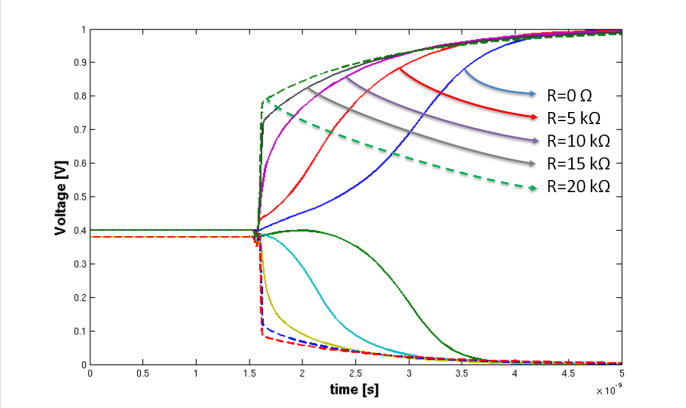

Fig. 5: Transient simulation for the schematic of figure 4 for different values of R, CL = 46fF, SA is minimal, no variations are included in simulation

This RC-latch effect arises when the RC-product is large. When it occurs, the capacitance behaves as a short-circuit and current flows through the resistor and a corresponding voltage drop over the resistor builds up. Afterwards the load capacitance charges itself at a much lower time constant.

a) : 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 arbitrarily reduced. The resistance of the passgates mustn't be too small however, or slow settling might occur. Then again, a large SA would probably charge the load capacitance effortlessly anyway.

VI. Read Throughput Results

A read access time simulation test was performed on the finished circuit (Figure 6). In this test the supply voltage was decreased and the maximum read throughput was measured. The test was performed at a (SPICE) temperature of 30°C. At each point in the shmoo plot, 100 MonteCarlo simulations were performed. The circuit has a read access time of 2.3ns at a supply voltage of 1V. 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, the buffers and SA 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 the cell is. 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.

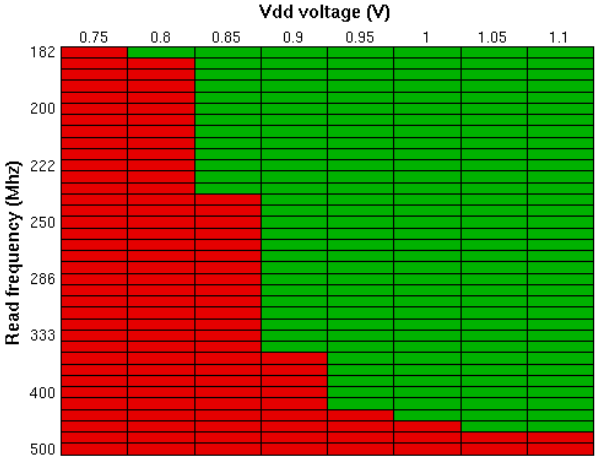

Fig. 6: Results read access time test

On figure 6 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 7. 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.

VII. Conclusion

[**?**]

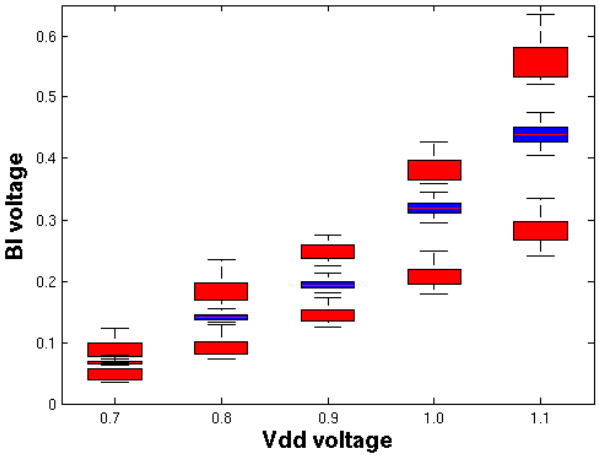

Fig. 7: BL voltage distribution for reference signal and HRS & LRS signals in function of vdd

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.