A yield centric statistical design method for optimization of the SRAM active column

A yield centric statistical design method for optimization of the SRAM active column

T.S. Doorn[1], J.A. Croon[1], E.J.W. ter Maten[1], A. Di Bucchianico[2]

[1]NXP Semiconductors, Eindhoven, the Netherlands, toby.doorn@nxp.com
[2]Eindhoven University of Technology, Eindhoven, the Netherlands

***Abstract*—For robust design of SRAM memories, it is not sufficient to guarantee good statistical margins on the SRAM cell parameters. The sense amplifier needs sufficient input signal before it can reliably sense the data, while the SRAM cell requires sufficient time to develop that input signal. This paper presents a new statistical method that allows optimization of the access time of an SRAM memory, while guaranteeing a yield target set by the designer. Using this method, the access time of a high performance advanced CMOS SRAM has been improved 6%, while simultaneously reducing the sense amplifier size.**

I. INTRODUCTION

As transistor dimensions continue to decrease, variability of device parameters has an ever larger impact on system performance. In particular mismatch (local variation) has increased dramatically for minimum sized transistors over the years as mismatch is proportional to $A_{vt}/\sqrt{area}$, while mismatch parameter A_{vt} has not scaled for recent technologies. Semiconductor memories make use of the smallest possible devices for density reasons, and these devices are particularly vulnerable to mismatch.

For system-on-chips (SoCs), Static Random Access Memory (SRAM) is the most widely used memory type. With technology scaling, an increasing amount of memory per SoC and the drive for low power and low voltage systems, SRAM variability and how to deal with variability in designs have rightfully received a large amount of attention. A lot of recent effort has been put into dealing with the impact of mismatch on SRAM cell parameters like static noise margin (SNM) [1]-[3]. However, the impact of mismatch does not stop at the memory cell level. After proper cell design, the operation of the memory system has to be statistically guaranteed. This paper shows how to deal with mismatch in the part of the SRAM memory that is most affected by it: the active column (also referred to as data path).

Whether or not an SRAM cell is properly read depends mostly on the memory cell, the sense amplifier and the time at which the sense amp is activated. Because the inputs of a latch type sense amp are precharged to the supply voltage, the minimum input signal required by such a sense amp is mainly determined by the mismatch of its pulldown NMOS pair. This mismatch can therefore be seen as the sense amp offset V_{SAo}. The SRAM cell needs to develop a signal ΔV_{bl} that is larger than this offset. This paper deals with intra-die V_{th} mismatch, which is independent stochastic variation. As V_{th} mismatch is mainly responsible for intra-die variations in ΔV_{bl} and V_{SAo}, they are modeled as independent stochastic parameters as well. The data in a cell is only incorrectly read if the cell is connected to a poor sense amp. The data will be correctly read if a cell is connected to a sufficiently good sense amp.

This paper presents a new statistical method to calculate when the sense amp can be activated to guarantee with a predetermined probability that ΔV_{bl} is large enough. A sense amp can sense data faster if it is presented with a larger input signal, but the memory cell would require more time to develop that input signal. The method presented here exploits this trade-off to optimize the access time of the memory and the size of the sense amps for any yield target.

The impact of mismatch on the memory active column (Figure 1) has so far been analyzed in only a few papers [4], [5]. [4] does not show how to statistically optimize the access time of the memory, which is explained in detail in sections III, IV and V of this work. In [5], the failure probability of the memory cell/sense amp combination is calculated, given an access time target. It is our opinion that the problem should be approached from the opposite direction. Given a yield target, one would like to optimize the access time and sense amplifier size. Only in this way is it possible to fairly compare different design choices, as to us yield is a parameter that should not be compromised on. Finding the optimal solution for a yield target of our choosing is therefore a priority for us. In addition, this work reduces the complexity of the problem by breaking it in two parts (memory cell and sense amp) using probability theory. The result is an intuitive method that provides insight in the contributions of different parts of the active column to the access time of the memory. Given a yield (loss) target, the optimal access time and sense amp size can be determined for any memory configuration.

This paper is organized as follows. Section II presents the probability theory needed to calculate and choose the failure probabilities for the memory cell and sense amp. Section III describes how to apply this method to the SRAM active

column. Section IV presents the results from the analysis. Section V outlines how to optimize the sense amp size and access time. Finally, we present our conclusions in section VI.

II. YIELD CALCULATIONS USING PROBABILITY THEORY

When designing a memory, one would like to design for a yield (loss) target. Even if redundancy would be used in the memory to repair tail bits, a yield target is required. In this section, the mathematics are described that allow calculation of the fastest possible access time t_a of an SRAM with yield loss target *YL* for total amount of memory on a SoC of *N* bits, *n* bits per sense amp and *m=N/n* sense amps per SoC.

A memory instance consists of *m* blocks *B* of *n* cells and 1 sense amp per block. Yield loss *YL* can be calculated from the probability that one or more blocks have a faulty cell/sense amp combination. As a block is either faulty or not, each block has a binomial distribution with probability *P(B)* for failing. This is described in equation 1. The approximation is valid for $0 \le P(B) < 1/m$, which is generally the case.

$$\begin{aligned} YL \equiv \quad P(\#B \ge 1) &= 1 - P(\#B = 0) \\ &= 1 - \binom{m}{0} P(B)^0 (1 - P(B))^m \\ &\approx m \cdot P(B) \end{aligned} \tag{1}$$

Rewriting equation 1, the required P(B) can be obtained, which can be simulated, as is shown later in this paper.

$$P(B) = \frac{n \cdot YL}{N} \tag{2}$$

Equation (2) shows that for a memory amount of $N=10^7$ bits and $n=1024$ bits per block a yield loss of $YL=10^{-3}$ can be obtained by designing for a block failure probability of $P(B) \approx 10^{-7}$.

Now *P(B)* has to be calculated from the probability that a worst case cell, generating worst case differential voltage $\Delta V_{bl,wc}$, is connected to a sense amp with an offset V_{SAo}.

$$\begin{aligned} P(B) &= P(\Delta V_{bl,wc} \le V_{SAo}) \\ &= \int_{-\infty}^{\infty} f_{SAo}(v) F_{\Delta Vbl,wc}(v) dv \end{aligned} \tag{3}$$

with f_{SAo} the probability density function of V_{SAo}. The distribution function $F_{\Delta Vbl,wc}$ of the voltage $\Delta V_{bl,wc}$ can be estimated from the distribution function of the voltage ΔV_{bl} that a single cell generates. A similar line of reasoning applies as was used for equation 1.

$$F_{\Delta Vbl,wc}(v) = 1 - (1 - F_{\Delta Vbl}(v))^n \tag{4}$$

$$F_{\Delta Vbl}(v) \approx \frac{F_{\Delta Vbl,wc}(v)}{n} \tag{5}$$

Like equation 1, equation 5 is valid for $0 \le F_{\Delta Vbl}(v) < 1/n$. By using equations 3 and 4, the impact of mismatch on the sense amp and memory cell can be separately simulated, after which the failure probability of a block and of the entire memory instance can be calculated.

III. SIMULATION METHOD AND CIRCUIT

The derivation shown in section II has been used to evaluate the impact of the method on the worst-case (slow-n, slow-p process, $V_{dd}=0.99V$, $T=125°C$) access time of a fast 8 kbit advanced CMOS SRAM design, having 1024 bitcells per sense amp and 8 to 1 column-multiplexing. As this paper deals with uncorrelated random V_{th} variation, the yield for 10Mbit of memory can be calculated by assuming that it consists of many instances of the 8kbit memory. Any other memory size can be used in the analysis. We design for $YL=10^{-3}$ for 10 Mbit of 8kbit memory instances, as was used as an example in section II. The result is a block failure probability $P(B) \approx 10^{-7}$. As timings are crucial for the SRAM active column, transient simulations were done. For this, an accurate active column model was required, of which a simplified version is shown in Figure 1.

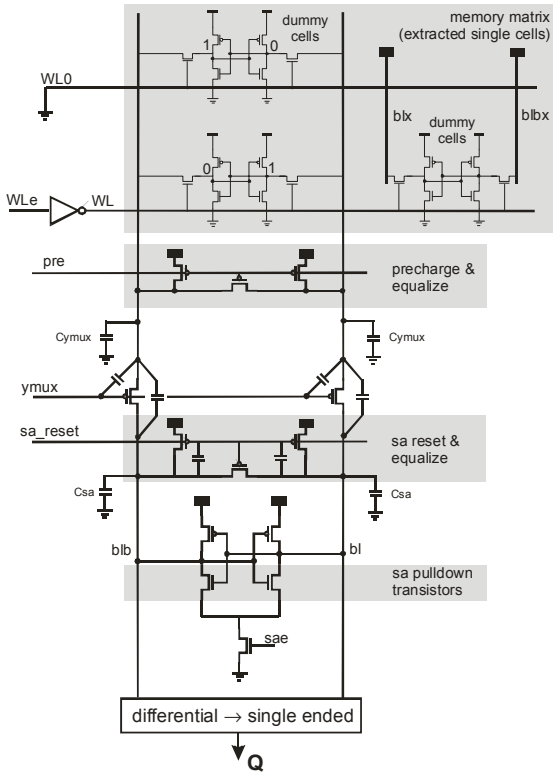

Figure 1: Simplified schematic of the SRAM active column. Note that the cell differential voltage is evaluated at the sense amp inputs, which is why these inputs are marked *bl* and *blb*.

The access time of an SRAM roughly consists of three parts: the address decoding time t_{dec}, the time Δt_{cell} the memory cell needs to discharge the bitlines and the time Δt_{sa} the sense amps requires to make a decision. The latter two are the subject of this work, as by far the larger part of variability is present there. Figure 2 schematically shows the timing of the most relevant signals in the SRAM active column. As soon as wordline *WL* goes high, the cell starts to discharge one of the

bitlines and a differential voltage across those bitlines $\Delta V_{bl}=V_{bl}-V_{blb}$ developes. All non-accessed cells in the column store the inverse data compared to the accessed cell in order to maximize the impact of leakage currents on ΔV_{bl}. As soon as ΔV_{bl} is sufficiently high, the sense amp can be activated using signal *sae*. The required magnitude of ΔV_{bl} depends on the sense amp offset, which is determined by the mismatch of mainly its two pulldown transistors. This minimum ΔV_{bl} is determined by doing Monte-Carlo (MC) simulations on the circuit in Figure 1. The time that is required by the cell to develop ΔV_{bl} is referred to as Δt_{cell}. As soon as *sae* is activated, some time passes before the data is stable at output *Q*. The time required for the data to appear at *Q* after *sae* is activated is referred to as Δt_{sa} (Figure 2). The column multiplexer is de-activated 50ps after *sae* is enabled.

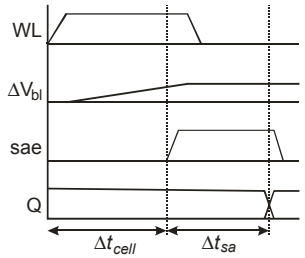

Figure 2: Timing diagram for the most important signals in the SRAM active column.

The simulations aim to minimize $\Delta t_{cell}+\Delta t_{sa}$ for a given yield loss target. As was shown in section II, the yield loss target can be translated in a target for the failure probability *P(B)* for a block of 1 sense amp and *n* memory cells.

Both the required values of Δt_{cell} and of Δt_{sa} to achieve $P(B)=10^{-7}$ have to be determined using Monte-Carlo simulations. The parameters that link Δt_{cell} and Δt_{sa} are the bitline differential voltage ΔV_{bl} and the sense amp offset V_{SAo}. By simulating the ΔV_{bl} and V_{SAo} distributions for different values of Δt_{cell} and Δt_{sa} respectively, equation 3 can be solved and block failure probability *P(B)* calculated.

The distribution for ΔV_{bl}, $F_{\Delta Vbl}$, is slightly non-normal and therefore its distribution cannot be extrapolated and has to be simulated entirely. As simulating an entire distribution requires simulation of very small probabilities, these simulations cannot be done with standard MC, as this would require too many MC trials. Using Importance Sampling (IS) MC [1], these small probabilities can be simulated and this technique is therefore applied using 50k trials, which is not a high number for estimation of probabilities in the range of 10^{-7}. The distribution for ΔV_{SAo} is normal and standard MC simulations (20k trials) can be used to estimate mean and standard deviation and to construct extrapolated distributions.

IV. Simulation Results

Figure 3 shows the ΔV_{bl} distributions $F_{\Delta Vbl}$ for different values of Δt_{cell}. The distrbutions are steep for small Δt_{cell} and become wider as Δt_{cell} increases. This is because the memory cells act as current sources and as time progresses, ΔV_{bl} of the fastest cell diverges ever further from ΔV_{bl} of the slowest cell.

In Figure 4, the sense amp offset distributions are added for different Δt_{sa}. The dotted and solid lines show ($\Delta t_{sa}=300ps$) that the distribution F_{SAo} is well approximated by a normal distribution. This simplifies calculatation of pdf f_{SAo} that is to be used for solving equation 3. Equation 3 can now be used with the different $F_{\Delta Vbl}(\Delta t_{cell})$ and $f_{SAo}(\Delta t_{sa})$ curves to calculate *P(B)* as a function of Δt_{cell} and Δt_{sa} (Figure 5).

The dashed line in Figure 5 shows *P(B)* versus Δt_{cell} for a relatively long $\Delta t_{sa}=300ps$, calculated from the $F_{\Delta Vbl}$ curves at the Δt_{cell} shown in Figure 4. This represents the best case Δt_{cell}, as the sense amp gets a long time to decide on what data is in the cell. Up to a certain input level, a sense amp can make a decision faster if its input voltage is higher. For very high input voltages, this no longer holds, as the sense amp decision time is then limited by its intrinsic speed.

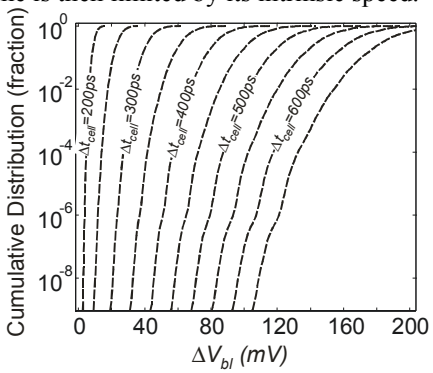

Figure 3: $F_{\Delta Vbl}$ for different Δt_{cell} (stepsize 50ps). Using equation 5, $F_{\Delta Vbl,wc}$ can be calculated from $F_{\Delta Vbl}$.

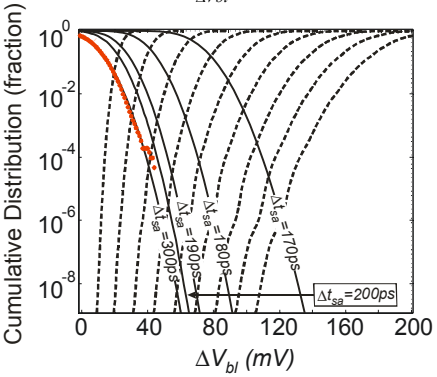

Figure 4: $1-F_{SAo}$ for different Δt_{sa} plotted together with $F_{\Delta Vbl}$. To solve equation 3, f_{SAo} is required. Nevertheless $1-F_{SAo}$ is plotted, as this presents the concept clearer than plotting f_{SAo}.

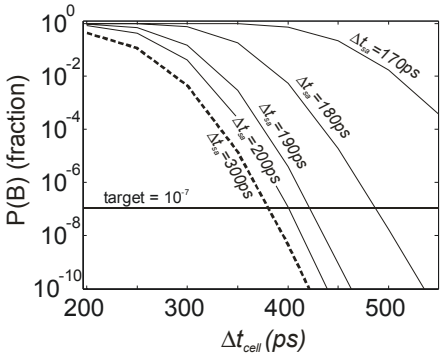

Figure 5: Block failure probability P(B) versus Δt_{cell} for different Δt_{sa}. For this example, the target is $P(B)=10^{-7}$.

To generate the higher input voltage, the memory cell requires a longer time. Δt_{cell} and Δt_{sa} can therefore be traded off in order to find the optimal $\Delta t_{cell} + \Delta t_{sa}$ for this memory instance. By restricting Δt_{sa}, the sense amp requires a higher input signal to evaluate the memory data on time and the sense amp offset distribution is shifted to a higher ΔV_{bl}, as shown in Figure 4. Figure 5 shows that restricting Δt_{sa} results in the *P(B)* curves shifting to larger values of Δt_{cell}, as higher input voltages are required and the memory cell needs more time to generate that voltage.

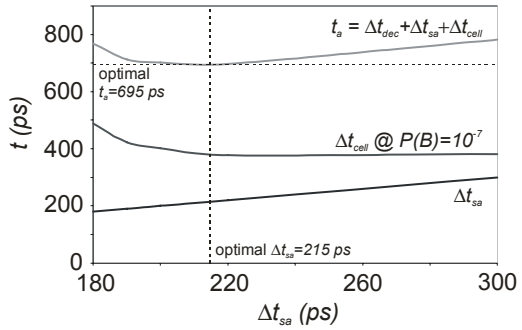

Figure 6: By plotting Δt_{cell} versus Δt_{sa}, the access time of the memory t_a can be optimized.

Now $\Delta t_{cell}+\Delta t_{sa}$ can be optimized by plotting Δt_{cell} versus Δt_{sa}, for a specified *YL* (Figure 6). This optimization of $\Delta t_{cell}+\Delta t_{sa}$ results in optimization of the access time t_a by adding a fixed value t_{dec} for the address decoding time: $t_a = t_{dec}+\Delta t_{cell}+\Delta t_{sa}$. For this memory design, an access time of t_a=*695ps* can be statistically guaranteed for *YL*=10^{-3} and 10 Mbit of memory.

V. Sense Amplifier and Access Time Optimization

Using the techniques described in the previous sections, the possibility now exists to statistically optimize the sense amplifier. A sense amp with smaller transistors has a larger offset voltage and hence for the same input signal need more time to sense the data. A sense amp with smaller transistors however results in a lower bitline capacitance, resulting in a faster discharge of the bitline by the memory cell. A lower internal node capacitance can also lead to the sense amp becoming faster, as discharging these nodes can be done faster. The trade-off therefore is between mismatch and capacitance.

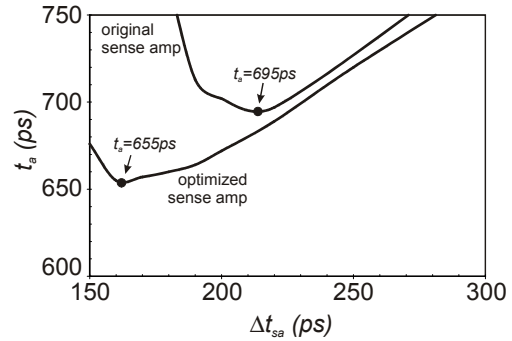

Figure 7: Access time of the original sense amp compared to that of the optimized sense amp.

In particular for memories with a low number of memory cells per column does this optimization make sense, as the sense amp capacitance can be of the same order as that of the bitlines. For long bitlines, the bitline capacitance dominates and optimization of the sense amp size yields lower gains.

Figure 7 shows the access time of the original sense amp and that of the optimized sense amp. Using this new statistical method, the optimization of the sense amp lead to a reduction in access time of 6%, while simultaneously reducing the silicon area. A further reduction in sense amplifier size leads to excessively high offset values and an increase in access time.

VI. Conclusions

In this work, a new statistical approach to analyze the active column of an SRAM is presented. Using this statistical method, both the access time and the sense amplifier size can be optimized for any pre-determined yield target.

Using probability theory, the optimization of the SRAM active column can be achieved by separately analyzing the sense amp- and memory cell failure distributions. This is an advantage, as the failure distributions can be constructed from simulating individual sense amps and memory cells, which can be done with relatively simple simulations.

A sense amp can make a faster decision on the data being read if it is presented with a higher input voltage. The memory cell that has to generate this input voltage then requires more time to generate that voltage. The sense amp decision time can therefore be traded off with the memory cell signal generation time. This new method exploits this trade-off for optimization of both the time at which the sense amp can be enabled and of the access time of the memory.

Finally, using this method, the access time of a high performance advanced CMOS SRAM has been improved by 6%, while simultaneously reducing the size of the sense amp.

Acknowledgments

The inputs from Roelof Salters, Patrick van de Steeg, Erik van Bussel, Baki Tezel en Erwin Tielemans (all NXP Semiconductors) on SRAM design and architecture have been indispensible for the quality of this work.

References

[1] T.S. Doorn, E.J.W. ter Maten, J. Croon, A. Di Bucchianico, O. Wittich, *Importance sampling Monte-Carlo simulations for accurate estimation of SRAM yield*, Proceedings of the ESSCIRC, pp. 230-233, 2008.

[2] T. Merelle et al., *First observation of FinFET specific mismatch behavior and optimization guidelines for SRAM scaling*, International Electron Devices Meeting Technical Digest, 2008

[3] C. Maufront, R. Ferrant, *Advanced statistical methodology for 6T-SRAM design*, Proceedings of the ICECS, pp. 756-758, 2007

[4] R. Aitken, S. Idunji, *Worst-case design and margin for embedded SRAM*, Proceedings of DATE, 2007

[5] R.M. Houle, *Simple statistical analysis techniques to determine optimum sense amp set times*, IEEE Journal of Solid State Circuits, vol. 43, no. 8, pp. 1816-1825, 2008.