INITIAL HIGH LEVEL CACTI ESTIMATES AND PHYSICAL DESIGN IMPLEMENTATION OF FAST SRAM CACHES

Eduardo Conrad Júnior, Eric Ericson Fabris, Sergio Bampi

Federal University of Rio Grande do Sul (UFRGS) - Informatics Institute
P.O. Box 15.064 – Postal Code 91.501-970 – Porto Alegre – Brazil

{econradjr, fabris, bampi}@inf.ufrgs.br

ABSTRACT

A cache memory can be viewed as a simple digital circuit. However, considering that this circuit handles a variety of analog signals, like bit line signals and sensing of charging and discharging nodes with digital behavior, cache memories can be considered as a complex mixed-signal circuit.

This paper addresses the analysis and design of a fast access cache memory in CMOS TSMC 0.18µm technology, from the initial electrical high level model, using the CACTI tool [1], to the final physical design implementation. The design procedure is presented for the cache main building blocks: the memory cell and the sense amplifier circuits. Simulation results are presented in order to validate the first pass of the design.

1. INTRODUCTION

The performance of a computer system is a function of the speed of the individual functional units and the system workload. Memory caches are critical components in the performance of such systems.

This paper addresses the analysis and design of a cache memory, first specifying the design space with the CACTI high level estimation tool [1]. This paper is organized as follows: Section 2 presents a briefly overview of cache memories. Section 3 shows the design methodology of a 32Kbytes CMOS cache memory and the results of a high level analysis of the memory using the CACTI tool. Section 4 shows the design and implementation of the SRAM cell and the sense amplifier, including first design layout. Finally, Section 5 presents our conclusions.

2. CACHE MEMORY OVERVIEW

The performance of a computer system is a function of the speed of the individual functional units, and of the workload presented to the system. In this context, memory caches are critical components in the performance of any computer system cache memory is the simplest cost effective way to achieve high speed memory, with its performance being extremely vital for high performance.

The cache works as a faster memory that store copies of blocks of data, witch can be accessed faster than main memory (high level memory). For these reasons, cache memory are still widely studied nowadays [1-6]. Figure 1 shows the basic architecture of a cache memory with these building blocks.

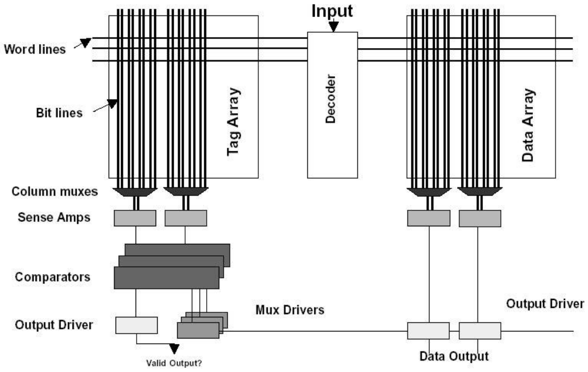

Figure 1– Basic architecture of cache memory with these building blocks [1]

3. CACHE MEMORY DESIGN AND CACTI HIGH LEVEL EVALUATIONS

We intend to follow six basic steps in the design of each memory building block: analysis, specification, implementation, SPICE simulation, layout and validation. Using this basic design steps to obtain initially the building blocks design (transistor level) follows the floorplanning, and, finally, the memory final structure.

In order to illustrate and discuss the design tradeoffs of an implementation of a CMOS Cache Memory, a design of a memory of 32Kbytes capacity, direct mapped, in a TSMC CMOS 0.18µm technology, is presented as a case of study.

In this first design step, the simulation of different physical characteristics sets using a high level estimation tool, the CACTI tool [1], was used in order to obtain a prospecting in area, power consumption and access time. Table 1 and figure 2 shows the CACTI estimation results, for different numbers of memory banks, and the best performance spot, i.e., the best access time with minimum area.

Table 1 – CACTI estimation results for study case

Banks Number	Access Time (ns)	Power Consumption (nJ)	Area (cm^2)
1	1.312430	0.620512	0.043067
2	1.125360	0.943860	0.042780
4	0.987339	1.647050	0.049001
8	0.937489	3.079390	0.059739
16	0.908084	5.840020	0.078761
32	0.946630	12.038000	0.146245
64	0.869381	22.407300	0.223667

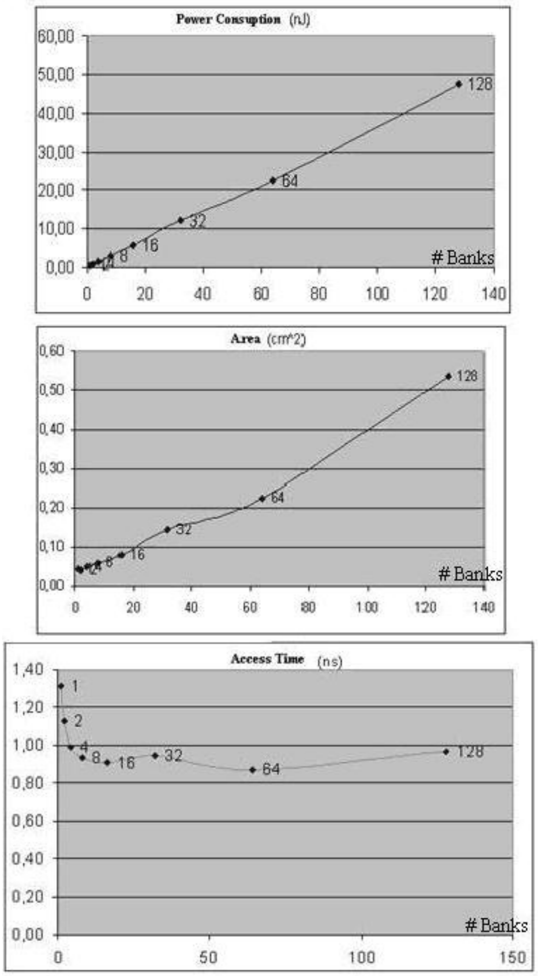

Figure 2 – CACTI estimation results for study case graphed.

In order to obtain the best architecture that provide a minimum point of access time vs. area vs. power consumption, we decide for a four banks architecture because it has about the same area as the two banks architecture, with a significant decrease in the access time.

Thus, the memory is composed of four identical blocks with 512 x 4 bytes x 4 columns in 32 bits interface. The memory banks aspect ratio is 512 x 200 bits (for 16 bytes in one word line). The blocks will be decoded with line and column decoders, composed by one decoder plus tri-state buffers and bit line multiplexers.

4. MEMORY BUILDING BLOCKS DESIGN AND IMPLEMENTATION

4.1. Six Transistors SRAM Cell

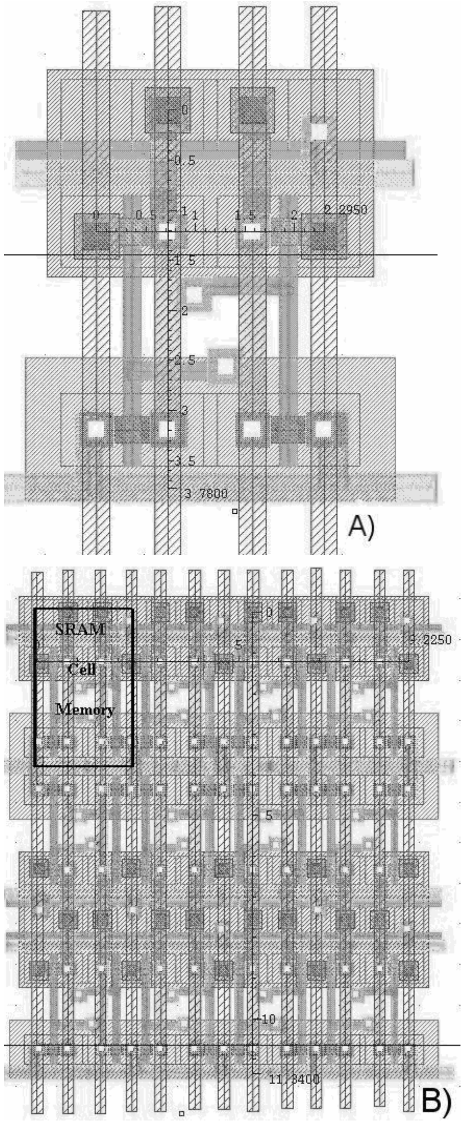

Figure 3 – (A) SRAM bit layout design and (B) SRAM cell array with twelve cells [7].

In TSMC MOSIS 0.18μm technology there are not any angle interconnections, thus we decided by a

Manhattan layout style. Another problem is that the butted contact does not exist in commercial technology, resulting in an increase of the area formed by the distance between metal1/polysilicon contacts in technology rules.

The polysilicon high resistance in the world line is a drawback compared to others consulted layout techniques. Our layout has a world line with polysilicon and metal1, based in [8]. This characteristic helps in to achieve the smallest memory access time, since it decreases the RC load in each world line. Figures 3a and 3b shows a SRAM bit layout and a test memory array, respectively. An important layout characteristic is the regular array by interconnection of mirrors basic cells. The Vss metal3 interconnections are used by neighbor cells. The same thing occurs in the bit line connection between two mirror cells in the layout. Figure 3b shows a neighbors cells interconnections and regularity using our SRAM cell design.

All transistor dimensions in the memory cell are 0.27 x 0.18 µm. The SRAM cell has 2.295 x 3.78 µm2 area and the whole array (with 12 SRAM cells) have 9.225 x 11.34 µm2.

The bit flip test is necessary in order to test the access influence in the bit state for the bit line capacitances. Using a bit-line capacitance estimation, we simulated the access to a bit using a SPICE simulator, showed in Figure 4, showing that the bit flip does not occurs in the working cycle.

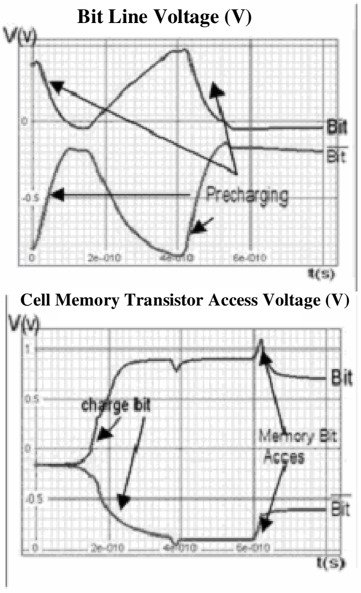

Figure 4 – Bit flip test results: (A) Bit Line Voltage and (B) Cell Memory Transistor Access Voltage.

4.2. Sense Amplifier

The design goal is an approximately 1ns access time. A considerable time in the reading cycle in the memory is the sense amplifier delay, what requires a high performance sense amplifier.

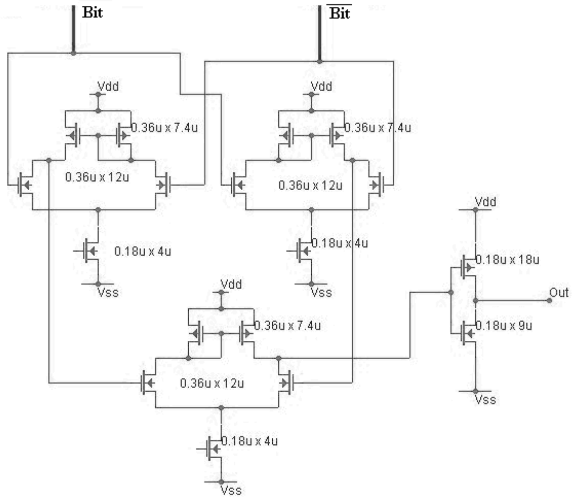

Figure 5 – Sense Amplifier based in three differential pairs

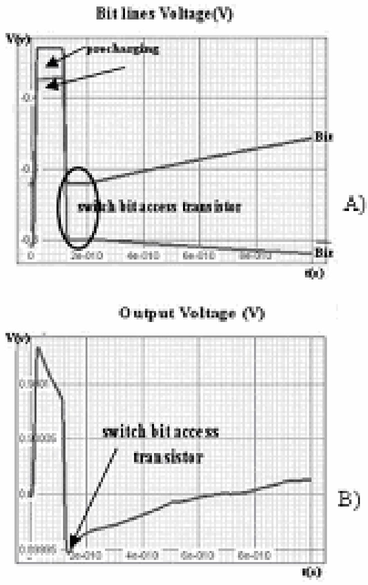

Figure 6 – Sense amplifier test results: (A)Bit Line Voltage and (B)Output Voltage

Several sense amplifier models are presented in publications [9-11], but a large current dissipation is necessary to decrease the time delay. Figure 5 shows a sense amplifier based in three differential pairs. This sense amplifier was designed and tested using SPICE simulations. Figure 6 shows a sense amplifier test to detect a memory position bit 1. The bit line pre-charge and the switch of an access bit transistor are the steps to read a position memory after the world line and column decoder.

For the implementation, in order to obtain a functional layout, the common centroid layout technique must be used for best layout matching. As future work we intend to obtain a new layout to our sense amplifier by evaluations to choose between voltage or current based sense amplifier. A sense amplifier first pass layout is shows in figure 6 [6].

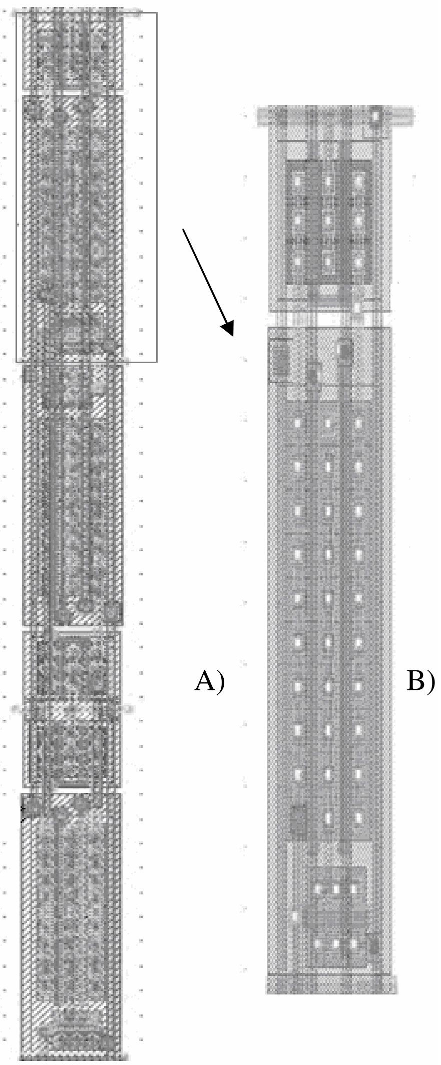

Figure 6 – (A) sense amplifier first design view and (B) Current mirror design view.

5. CONCLUSIONS

The design methodology used in a 32Kbytes CMOS cache memory design, including preliminary layout and simulation results, was presented on this paper. These results show the design space evolution, including design challenges and methodology. The cache electrical design is simplified considering the best solution for each block individually; if one block has a large delay, it affects the whole system performance.

As future work we intent to test several blocks implementations or architectures to choose the best solution aiming fast SRAM. We intent to finish all cache memory building blocks implementation and test. From this point we intent to obtain a fully integrated memory. A future prototype test chip should validate the design methodology and implementation.

6. REFERENCES

[01] N. P. Jouppi, and P. Shivakumar, "An Integrated Cache Timing, Power and Area Model," Western Research Laboratory, USA, 2001.

[02] T. Wada, R. Suresh, and S. A. Przybylski, "An Enhanced Access and Cycle Time Model for On-Chip Cache Memories," IEEE Journal of Solid-State Circuits, vol. 27, Nº 8, August, 1992.

[03] B. Wicht, S. Paul, and D. Schmitt-Landsiedel, "Analysis and Compensation of the Bitline Multiplexer in SRAM Current Sense Amplifiers," IEEE Journal of Solid-State Circuits, Vol. 36, Nº 11, November, 2001.

[04] K.-S. Yeo, Z.-H. Goh, Q.-X. Zhang, and W.-G. Yeo, "High-Performance Low Power Current Sense Amplifier Using a Cross-Coupled Current-Mirror Configuration," Proc.-Circuits Devices Syst., Vol. 149, Nº 516, October/November, 2002.

[05] B. Wicht, S. Paul, D. Schimitt-Landsiel, "Analysis and Compensation of the Bitline Multiplexer in SRAM Current Sense Amplifiers", IEEE Journal of Solid-State Circuits, V. 36, Nº 11 , Nov 2001.

[06] M. Wieckowski, M. Margala, "A 32KB SRAM Cache Using Current Mode Operation And Asynchronous Wave-pipelined Descoders", IEEE SOC Conference, Proceedings. IEEE International, 2004.

[07] E. Conrad J., "Memória Cache – Da Estimativa CACTI a Implementação Física," Departamento de Engenharia Elétrica da UFRGS, Porto Alegre / Brazil, 2005.

[08] N. Okazaki, F. Miyaii, Y. Harada, J. Ayoama, and T. Shimada, "A 30ns 256K Full CMOS SRAM," ISSCC, February, 1986.

[09] SHIBATA, N. Current Sense Amplifier for Low-Voltage Memories, IEICE Trans. Electron, V. E79-C, Nº 8, Aug 1996.

[10] Y. Tsiatouhas, et al., "New Memory Sense Amplifier Designs in CMOS Technology", The 7th IEEE International Conference on Electronics, Circuits and Systems, 2000.

[11] M. Margala, "Low-Power Circuits Design", The 7th IEEE International Workshop on Memory Technology, Design, and Testing, Proceedings p.115, 1999.