Power Characterization of RAMs. An Experimental Approach.

Javier Rellán, José L. Ayala, Marisa López-Vallejo
Departamento de Ingeniería Electrónica
Universidad Politécnica de Madrid (Spain)
Email: {jrellan,jayala,marisa}@die.upm.es

***Abstract*— In this paper, we present a flexible methodology for characterizing the power consumption of RAM architectures. The proposed approach overcomes previous limitations by attending to second order effects of internal circuit capacitances and does not rely on technological parameters. The experimental results have been used to provide a set of design guidelines to lead the system designer in the early implementation of RAM hierarchies with low-power constraints. This paper presents the description of the experimental methodology as well as the analysis of some of the collected results. Finally, a typical design case is described.**

I. Introduction

Low power is becoming more and more a concern in microprocessor design. Particularly when designing microprocessors targeting portable embedded systems, energy consumption is an important issue, since it must be supplied by batteries. The power dissipation due to on-chip caches constitutes a significant part of the overall power dissipated by modern microprocessors. For example, the on-chip D-cache of the StrongARM 110, a low-power RISC microprocessor, consumes 16% of its total power [1]. In fact, embedded memories in most of today's SoC designs consume an average of 30 to 50% of the die area [2] and this number is growing annually. Furthermore, the rapidly growing portable electronics market is demanding low power devices, making techniques for energy-efficient caches an important research area [1] [3] [4].

Embedded components require the assistance of design construction, optimization, and analysis tools that produce verifiable memories for SoC designs. These tools attempt to guide the designer with the increasingly complex circuits they produce. Many of these tools rely on static analysis and abstracted memory models [5]. Coarsely characterized memory models with significant timing and/or power guardbanding are enough for physical synthesis and optimization purposes; however, final static analysis with these same models negatively affects system performance. In fact, incorrect models will cause functional failures. Regarding the power consumption, weak characterizations lead to over-estimates and under-estimates of the total energy dissipation, what nowadays cannot be accepted by the integrated system manufacturers.

To reflect the right circuit behavior, memory models require the acquisition of additional data during predesign characterization, where predesign implies analysis prior to instantiation in a design. Also, this characterization has to provide enough information for developing accurate analytical models without technological ties.

The work presented in this paper shows a characterization methodology for memory architectures based on circuit simulations. The proposed approach takes into account second order effects like the effect of internal capacities, not previously considered in analytical models. It also provides a set of design guidelines to lead the system designer on the electronic design with low-power constraints from the very early stages.

The paper structure is as follows. Previous work is summarized in the next section, and a brief theoretical background is given in section III. The experimental methodology is described in section IV, and section V presents the experimental results and analysis. Finally, some conclusions are drawn.

II. Related Work

Developing an accurate model is a time consuming task that requires detailed low level knowledge of circuit design and infrastructure that many researchers do not generally have. Clearly, the accuracy of these models greatly depends on the accuracy of the power models of each functional component and the detail with which component accesses are modeled. In the work by Ghiasi and Grunwald [6], they compare the accuracy of two different power models that rely under distinctly assumptions. They also find out that it is necessary to understand the underlying architecture of both the simulation model and the power model to exercise the appropriate caution in applying the power model to modified architectures.

This necessity of getting accurate power simulations of the underlying circuits has also moved to the search of input vectors for stressing the circuit behavior. In [5], the authors describe methods to automatically generate complete, optimized, and efficient input simulation vectors for the characterization and modeling of embedded memories. The use of accurate models that precisely match silicon behavior will enable correct static timing analysis. These authors do not consider any simulation approach in particular and thus, their approach can be applied to the methodology proposed in our work.

Also, several measuring techniques have been proposed to quantify the parameters involved in the power consumption of integrated circuits. A comparative analysis of these approaches can be found in [7] and some experimental results in [8], while in [9] an efficient way to estimate these parameters in microprocessors implemented in FPGAs is proposed .

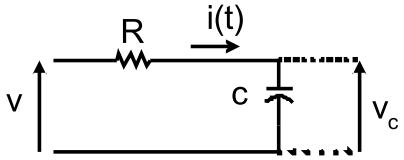


Fig. 1. RC circuit

Once the power models have been devised, these can be used to develop power estimation techniques and power models. One application of simulation-based information to the energy characterization of the cache hierarchy of embedded systems can be found in [10]. Also, several power estimation tools, like Wattch [11] or SimplePower [12], make use of this information for providing power estimations of a more complex architecture. Finally, a similar approach for characterizing memory-based circuits have been followed by models like CACTI [13]. However, these models have not considered the effect of internal capacitances and second order effects which move the equivalent impedances away from the linear scale.

III. Theoretical Background

The power characterization method we follow is based on the following assumptions:

- The source of power dissipation in this model is the charging and discharging of capacitative loads caused by signal transitions. Therefore, only dynamic power consumption is considered, even though this information could be easily extended with the static power information provided by Spice.
- The energy consumption in every transition is proportional to the rising and falling times of the signal. Thus, the experimental method will concentrate on measuring these times.

This second assumption can be explained as follows. Every node in the circuit can be modeled by its RC equivalent circuit (see Figure 1). In this circuit, the evolution in time of the voltage during the charge is

$$v_c(t) = V_{max} \cdot e^{\frac{t}{RC}}$$

If t_1 is the time required for v_c to achieve a 10% of its maximum value, and t_2 is the time required to achieve a 90% of its maximum value, the difference $t_2 - t_1$ is the rising time t_c.

$$v_c(t_1) = V_{max} \cdot e^{\frac{t_1}{RC}} = 0.1 \cdot V_{max}$$

$$v_c(t_2) = V_{max} \cdot e^{\frac{t_2}{RC}} = 0.9 \cdot V_{max}$$

$$\begin{aligned} t_c &= t_2 - t_1 \\ &= -RC \ln 0.1 + RC \ln 0.9 \\ &= RC(\ln 0.9 - \ln 0.1) \\ &= \beta_c \cdot RC \end{aligned} \tag{1}$$

where β is a numeric constant.

Similar expressions can be obtained for the discharging time. Therefore, there is a direct relation between these times and the capacitance of the RC model.

The relation between these capacities (or the charging and discharging times) with the dissipated power is summarized in the following expression:

$$P = 0.5 \cdot C_t \cdot V_{DD}^2 \cdot E(sw) \cdot f_{clk} = \gamma \cdot t \cdot V_{DD}^2 \cdot E(sw) \cdot f_{clk}$$

where C_t is the total equivalent capacity at the output of the node, V_{DD} is the voltage source, $E(sw)$ is a parameter related to the switching activity, and f_{clk} is the clock frequency.

Attending to the previous expressions, it is justified to measure the charging and discharging times of the circuit nodes to estimate the power consumption of the system.

IV. Experimental Method

Our experimental methodology is as follows.

1) First, a set of different memories are generated. In order to evaluate either SRAM and DRAM technologies, both kind of schematic representations have been created.
2) Next, test vectors are also generated to perform all the interesting logical transitions in the circuits to be evaluated.
3) Then, the Spice simulator included within the Cadence environment is used to collect the results of the electrical simulations of such circuits.
4) Finally, and after the careful analysis of these results, a set of guide-lines is established for guiding the circuit and system designer in the efficient design and use of the memory hierarchy when low-power constraints are involved.

A. Generated Memories

Figure 2 shows the basic cell for an SRAM. The memory circuit is composed of *N* rows by *M* columns of memory cells, the corresponding address decoders, the pre-charge circuit, and the sense amplifiers in the output lines. Figure 3 shows this configuration. The number of rows and columns will range from 1 to 16 and, in order to shorten the simulation time, just a representative number out of the 256 configurations will be evaluated. This subset allows us to find the values for the intermediate points by splin-based interpolation, while extrapolation to those near points can also be done.

Therefore, twenty-five different basic SRAMs were generated in .8u technology with a 4.5V supply. This technology is obviously quit old compared with the current industrial

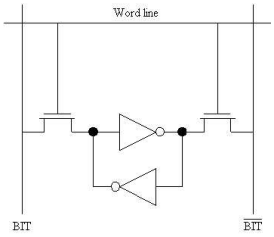


Fig. 2. SRAM cell

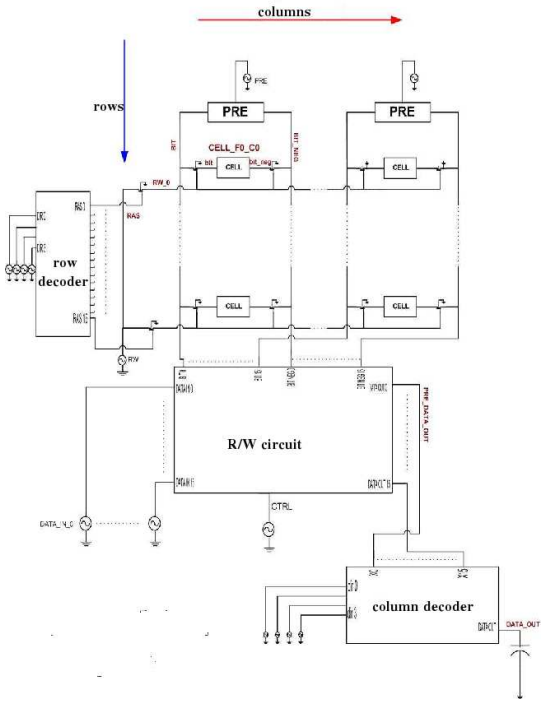


Fig. 3. SRAM circuit

standards, however, the results and methodology presented are easily adapted to newer technologies. The largest and smallest size memories were included and the others were chosen randomly. The subset of chosen memories was examined to ensure a good variation in the number of rows, number of columns, bit width, number of row and column address lines, and total number of storage bits in the memory.

Figure 4 shows the basic cell for a DRAM. Similarly to the SRAM configuration, the DRAM circuit is composed of *N* rows by *M* columns of memory cells, the address decoders and the sense amplifiers in the output lines. No pre-charge circuit can be found in this case, however, a more complex control logic for re-writing the memory contents is needed.

B. Generated Test Vectors

The waveforms used during the circuit simulation reflect the normal behavior of the memory devices (read and write operations, pre-charge, re-writing cycles, etc.) as well as strongly push the logic transitions in every node of interest. This last constraint allows us to calculate the equivalent capacity (and thus, to calculate the associated energy consumption) in every

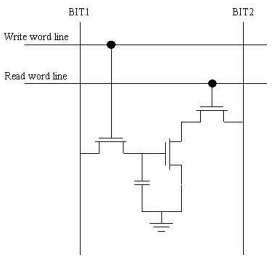


Fig. 4. DRAM cell

node, and for every line and type of access in the circuit. Opposite to previous approaches that limit the analysis to the address and bus lines and to the read/write operation, the proposed methodology performs the analysis in a bigger representative set of experimental variables.

Therefore, the generated waveforms do not only match the normal functioning of the device (addressing, data write cycles, pre-charge cycles, data read cycles, data re-writing if needed, etc), but also take into account the bit-switch effect in the address and data buses (’0’ after ’1’, ’1’ after ’0’, ’1’ after ’1’ and ’0’ after ’0’). Moreover, the addressing policy also considers the effect of correlated and uncorrelated addresses.

The accurate choice of the test vectors is a key-factor to obtain correct simulations and meaningful results.

C. Circuit-Level Simulations

A representative subset of circuit configurations for both SRAM and DRAM technologies has been implemented and simulated with the Spice simulator within the Cadence environment.

The simulations have used the test vectors previously built, and they have been run up to completion. In every simulation we have monitorized both the right functioning of the circuit (read operation, data memorization, write operation, etc), and also the charge (rising) and discharge (falling) times in every representative point.

D. Processing of the Results

Once the circuit simulations have been finished, the collected results are analyzed to explain the delay and capacitance dependence with the underlying architecture. The increase or decrease of the charge and discharge times (and, consequently, of the equivalent capacitances) are related to the architectural modifications (number of rows and columns) and then summarized in a set of guide rules for the system designer.

The results for the finite set of simulated points are interpolated to the rest of intermediate points in the test space. The collection of points are approximated in 2-dimensions by third-order splin lines (the lowest-order lines that best fit the results), one for the results when the number of rows are ranged, and the other one for the results when the number of columns are ranged (see Figure 5). The analyzed space is wide enough for allowing the extrapolation of the results to near points.

In this way, a system designer interested on establishing a memory hierarchy with low-power constraints, can automati-

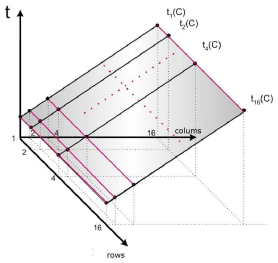


Fig. 5. 2-D interpolation

cally decide the best configuration for rows, columns or bit-width in terms of energy dissipation. The decision is made from the very early design stages, and no turn-back design cycles are needed.

V. Lessons Learned

This section presents some of the collected results for both SRAM and DRAM configurations. In particular, just a few of the collected measures will be outlined and carefully analyzed.

Finally, an example of system level design guided by the rules and conclusions extracted from the previous experiments will be described.

A. SRAM during Pre-Charge Cycle

Figure 6 shows the set of waveforms used to simulate the SRAM behavior. As was previously commented, the test waves are designed to stress all the circuit points during the normal functioning of the memory.

Every circuit point of interest has been studied and characterized in terms of charging and discharging times. In particular, we will describe the results obtained for the bit line before the pass transistor (outside the memory cell) when a pre-charging cycle is applied.

Figures 7a and 7b show the rising and falling times respectively for the bit-line outside the memory cell during the pre-charge cycle. For the analysis of these results, either the dependence with the number of columns and rows should be distinguished. Attending to the dependence with the number of columns, it has to be observed how every memory cell is independent from the nearests during the pre-charge cycle, and this explains the independence of both rising a falling times with the number of columns. However, when increasing the number of rows, it also increases the equivalent impedance seen by the bit-line, increasing the rising time in this line. The discharge of the pass transistors in the column happens through the read/write control circuit (see Figure 8), and the current through this circuit increases with the number of rows (number of current sources). Since the voltage value in the line remains constant but the current increases, the equivalent impedance seen in the bit-line decreases with the number of rows.

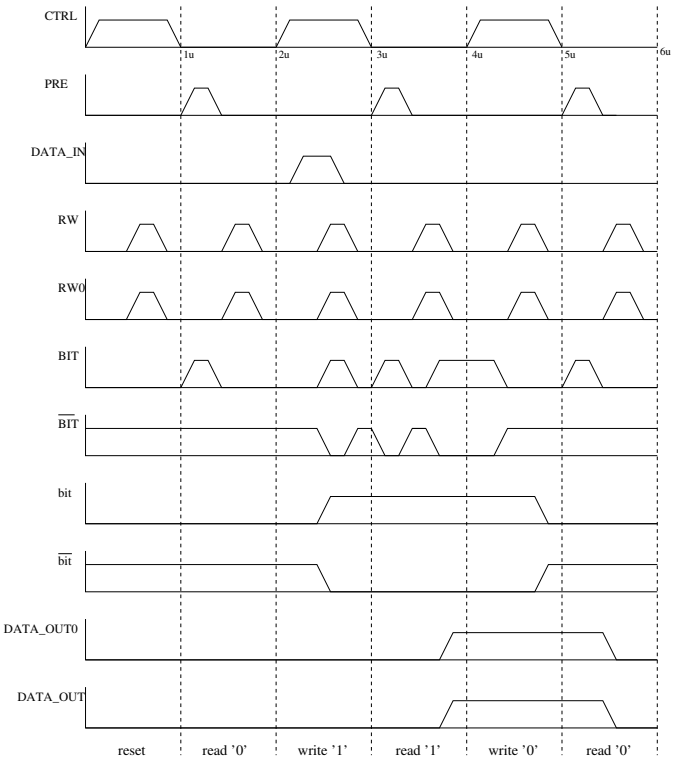
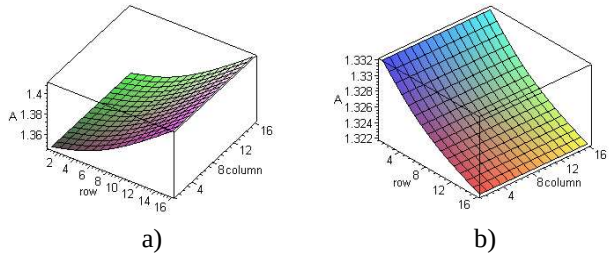


Fig. 6. SRAM chronogram



a) b)

Fig. 7. SRAM: Bit-line times during pre-charge. a) Rising time; b) Falling time

B. DRAM during Write Cycle

Figure 9 shows the set of waveforms used to simulate the DRAM behavior. As was previously commented, the test waves are designed to stress all the circuit points during the normal functioning of the memory.

Every interesting circuit point has been studied and characterized in terms of charging and discharging times. In particular, we will describe the results obtained for the bit line after the pass transistor (inside the memory cell) during a write cycle.

Figures 10a and 10b show the rising and falling times respectively for the bit-line inside the memory cell. For the analysis of these results, either the dependence with the number of columns and rows should be distinguished. Attending to the dependence with the number of columns, it has to be observed how every memory cell is independent from the nearests because the read line is not asserted. Therefore, the

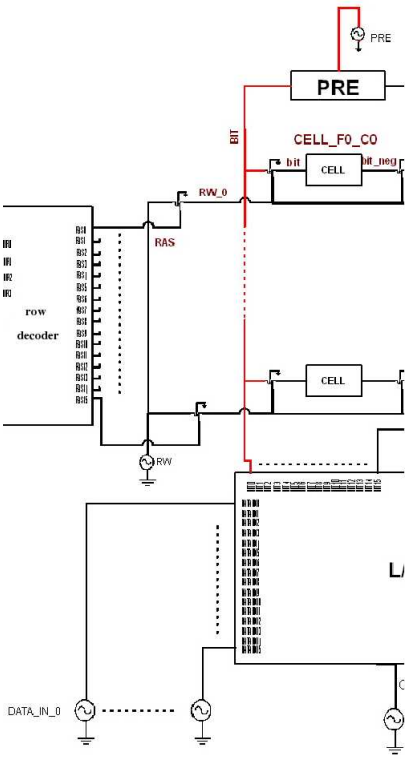


Fig. 8. Measures in SRAM bit-line during pre-charge (dependence with rows)

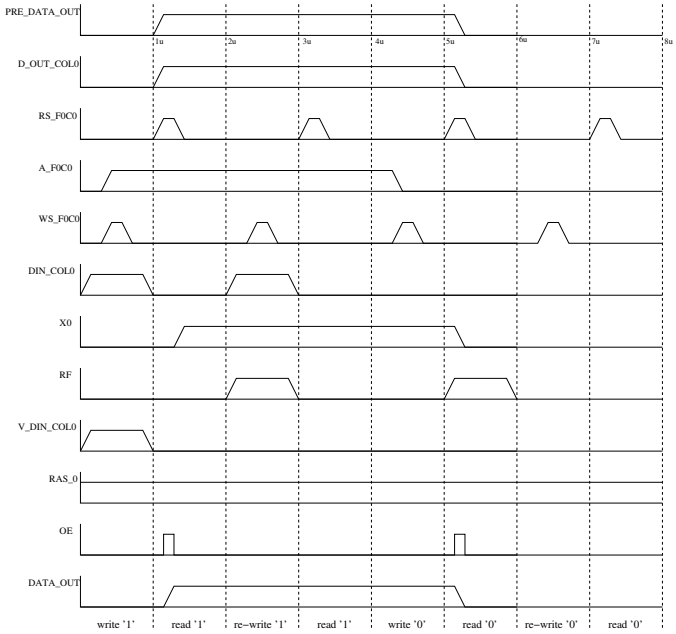
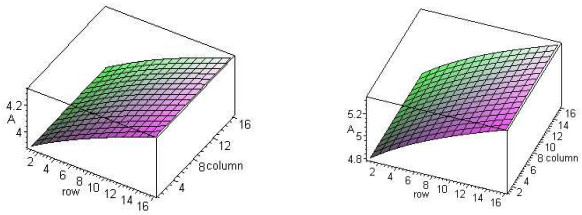


Fig. 9. DRAM chronogram



a) b)

Fig. 10. DRAM: Bit-line times. a) Rising time; b) Falling time

rising and falling times are completely independent with the number of columns (Figure 11 clarifies this independence between close adjacent cells. It can be observed that there is not any communication path between adjacent columns, therefore, the situation is the one presented in the Figure).

With respect to the number of rows, it can be observed how all the memory cells in the same column share the write line DATA_IN_COL0 (see Figure 12), and how the voltage source has to provide the current for every equivalent impedance. When increasing the number of rows, the equivalent impedance seen by the source also increases and, since the current provided by the voltage source is constant, the rising and falling times increase with the number of rows.

C. Design Example

In this section we will show how the knowledge acquired during the simulation phase can be used by the system designer to select the best memory configuration that satisfies the power constraints. The example will be shown for a very simple architecture (16-bits memory), but it can be extended to any other configuration.

For the memory of interest, five different configurations can be selected from the simulated design space (1x16, 2x8, 4x4, 8x2 and 16x1), each one with different energy behavior. For these configurations, the dynamic dissipated energy can be expressed as

$$P_D(M,N) = \gamma_1 \cdot \sum K_n \Delta t_n V_n^2 = \gamma_2 \cdot \sum K_n \Delta t_n \alpha_{V_n}$$

where $P_D(M,N)$ is the dynamic power consumption for the $M \times N$ memory configuration, and it is obtained after adding the contribution of every voltage switch (which has been proved to be proportional to t_n through a factor γ). K_n is related to the switching activity of the circuit for the input test vectors. The time Δt_n is the charging or discharging time for the node n. Finally, a weighting factor for the voltage has been introduced, α_{V_n}. This factor reflects the contribution of the different voltages at the circuit nodes in the total energy dissipation. Therefore, it can be defined as

$$\alpha_{V_j} = (\frac{V_j}{V_{nMAX}})^2$$

After simulating the SRAM and DRAM topologies with the method proposed in previous sections, and after applying the previous expressions, the following results are obtained (technological values have been omitted to obtain a technology-independent implementation)

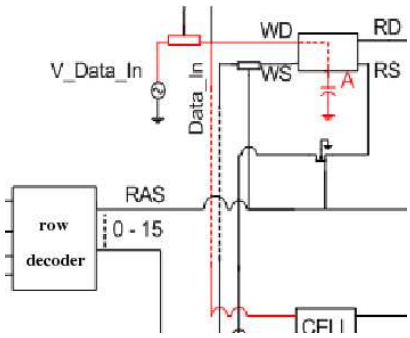


Fig. 11. Measures in DRAM bit-line (dependence with columns)

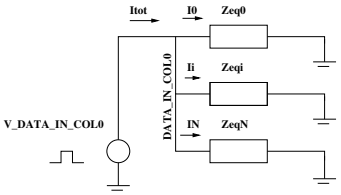


Fig. 12. Measures in DRAM bit-line (dependence with rows)

This table shows the energy savings obtained for different memory configurations without applying any additional low-power policy. The column stating the total values states the result of applying the expression of energy consumption after removing the technology-dependent parameters (therefore, the results cannot be expressed in energy units). Finally, the values presented in the table are normalized with respect to the most-hungry configuration for the SRAM and the DRAM (16x1, the configuration with more rows).

These results show how the simple election of the memory topology (a degree of freedom available in many designs) can be used to efficiently reduce the power consumption from the design phase without any additional requirement.

Summarizing, there are some important rules that the system designer has to remember when planning the memory hierarchy:

- The underlying topology (rows and columns) has a strong impact on the power consumption of the device;
- For equal memory size, memory rows containing more than one data word can reduce energy consumption;
- Memories that split the data word between more than one row can show lower energy consumption. However, in those cases, the energy dissipated in the bus during the extra access has to be evaluated;
- The memory architecture has to be considered from the very early design stages to make feasible the energy reductions and the architecture adaptation;

VI. Conclusions

In this paper, a flexible and versatile methodology for the power characterization of memory hierarchies has been presented. The proposed methodology does not rely on technological parameters and can be applied to several implementation models. This power characterization, on the contrary to previous circuit characterizations, takes into account second order effects of internal circuit capacitances.

	SRAM		DRAM	
	TOTAL	**REDUCTION**	**TOTAL**	**REDUCTION**
16x1	65,037	0%	54,735	0%
1x16	47,229	27.4%	54,508	0.4%
8x2	53,831	17.3%	52,643	3.9%
2x8	47,027	**27.7%**	53,824	3.5%
4x4	48,524	25.4%	52,573	**4%**

Several circuit schemes have been studied, and numerous results have been analyzed. The collected conclusions have been summarized in a set of design rules for the system designer interested on low-power constraints.

Acknowledgements

This work was supported by the Spanish Ministry of Science and Technology under contract TIC2003-07036.

References

[1] J. Kin, M. Gupta, and W. H. Mangione-Smith, "Filtering memory references to increase energy efficiency," *IEEE Trans. on Computers*, vol. 49, no. 1, January 2000.

[2] E. Hall and G. Costakis, "Developing a design methodology for embedded memories," *Integrated System Design*, January 2000.

[3] M. Kamble and K. Ghose, "Energy-efficiency of VLSI caches: a comparative study," in *Int. Conf. on VLSI Design*, 1997.

[4] H.-H. S. Lee and G. S. Tyson, "Region-based caching: An efficient memory architecture for embedded processors," in *CASES*, 2000.

[5] J. Abraham, "Directed dynamic simulation methods for embedded memories in nanometer processes," in *System-on-Chip and ASIC Design Conference*, 2003.

[6] S. Ghiasi and D. Grunwald, "A comparison of two architectural power models," in *International Workshop on Power-Aware Computer Systems*, 2000.

[7] J. Rius, A. Peidro, S. Manich, R. Rodríguez, and E. Boemo, "Measuring power and energy of CMOS circuits: A comparative analysis," in *Design of Circuits and Integrated Systems Conference*, 2003.

[8] J. Rius, A. Peidro, S. Manich, and R. Rodriguez-Sánchez, "Power and energy consumption of CMOS circuits: Measurement methods and experimental results," in *International Workshop on Integrated Circuit and System Design, Power and Timing Modeling, Optimization and Simulation*, 2003.

[9] S. López-Buedo and E. Boemo, "Making visible the thermal behaviour of embedded microprocessors on FPGAs. A progress report," in *International Symposium on Field-Programmable Gate Arrays*, 2004.

[10] J. L. Ayala and M. López-Vallejo, "A unified framework for power-aware design of embedded systems," in *International Workshop on Integrated Circuit and System Design, Power and Timing Modeling, Optimization and Simulation*, 2003.

[11] D. Brooks, V. Tiwari, and M. Martonosi, "Wattch: A framework for architectural-level power analysis and optimizations," in *International Symposum on Computer Architecture*, 2000.

[12] W. Ye, N. Vijaykrishnan, M. Kandemir, and M. Irwin, "The design and use of SimplePower: A cycle-accurate energy estimation tool," in *Design Automation Conference*, 2000.

[13] G. Reinman and N. Jouppi, "An integrated cache timing and power model," COMPAQ-Western Research Lab, Tech. Rep., 1999.