Power Consumption of On-Chip ROMs: Analysis and Modeling

Eike Schmidt[1], Lars Kruse[1], Gerd Jochens[1], Ed Huijbregts[2],
Wouter Nieuweboer[2], Eric Seelen[2], Wolfgang Nebel[1]

[1] OFFIS - Division 1 - Embedded Systems
D - 26121 Oldenburg, Germany
{Eike.Schmidt, Kruse, Jochens}@OFFIS.Uni-oldenburg.DE

[2] Philips Research Laboratories
5656 AA Eindhoven, The Netherlands
{Huijbre,Seelen}@Natlab.Research.Philips.com

Abstract

This paper addresses the problem of modeling the power consumption of on-chip ROMs for gate-level and RT-level power estimations. A route to memory power model development is presented that is also applicable to other memory architectures. The model proposed operates within an error margin of less than 5%.

1 Introduction

Power consumption has become an important dimension in the design space of integrated circuits. It can even be the limiting factor within the design process. Therefore, power estimation before the chip fabrication is often a mandatory task. Gate-level power analysis delivers good results within an error of 5% [1] [2]. In many cases, however, the power consumption of memories is not accounted for in overall calculations. This is often due to the lack of memory power models or dissatisfactory accuracy [3].

Since large portions of the chip area are covered by on-chip memories in today’s system-on-chip designs, memories consume a large fraction of the total dissipated power. 50 to 80% of the overall power consumption in multi-dimensional real-time signal processing applications is due to memory [4]. Therefore, memory power models accurate to the same level as logic gate models are needed.

In [5] a gate-level power simulator is introduced that takes memories into account. They report a maximum memory model error compared to HSpice simulations of about 11% but do not give any details about the model itself.

We present a simulation-based gate-level power model for a specific on-chip ROM architecture from Philips Semiconductors and a method for memory model development that can also be applied to other memory architectures. Our approach models the average current consumption of the memory, however, the model can be extended to deliver time accurate current curves with a user specified time resolution. The model is part of a gate-level power estimator from Philips that is integrated into a Verilog-XL simulator [6]. The characterization of ROM memories is done per instance. This isn’t a problem because the timing characterization is also done for each instance separately. This way of characterization on layout-extracted netlists delivers accurate results.

Section 2 of this paper describes the ROM architecture for deriving our power model. In section 3 the basic steps for discovering and extracting relevant parameters for a model are presented while the model itself is detailed in section 4. Simulation results using our model are given in section 5.

2 ROM architecture

The Philips ROM architecture is organized in a 3-dimensional way (cf. figure 1). The memory consist of a number of memory matrix blocks (dimension Z). Each matrix has several rows (dimension X) as well as columns (dimension Y). All dimensions have a seperate address decoder, the X and Z part being two stages and the Y part one stage. Only the second stages of X and Z part work synchronously, the rest exhibits asynchronous switching. Two bits, one in the X part and one in the Y part, are used for a divided wordline/bitline scheme [7].

A common read bus (CRB) connects all memory blocks to the output. During read cycles, the data word read in the active memory block (w_s) is driven to the CRB, changing the previous state of the bus (w_p) to w_s. An asynchronously operated tristate driver can finally put the last read word to the output (changing it from w_{op} to w_s).

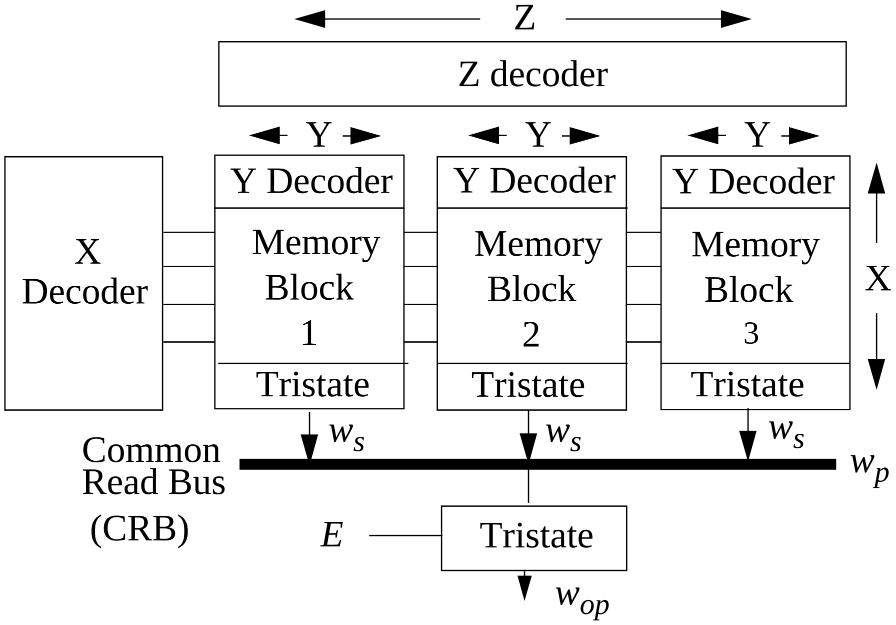

Figure 1. ROM architecture consisting of memory blocks (Z). Each of the blocks in turn consisting of rows (X) and columns (Y). E denotes the output enable signal, w_p is the previous content of the common read bus, w_s is the new word read from the memory and w_{op} is the word visible at the output.

3 Power analysis

Circuit level simulations of complete extracted netlists were conducted to analyse the dependencies of the power consumption on the memory contents and the control and address inputs. The current flow was measured through a set of current monitors, giving a detailed view of the structural distribution of the consumption. Simulation results were investigated using statistical trend analysis. Power dissipation was found to be linked to one of three triggering events: memory activation, address switching and output switching.

In the remaining sections we will use $|w|$ to denote the weight (the number of 1 bits) in word w, and $R(w_1, w_2)$ respectively $F(w_1, w_2)$ to denote the number of rising respectively falling bits when changing from w_1 to w_2.

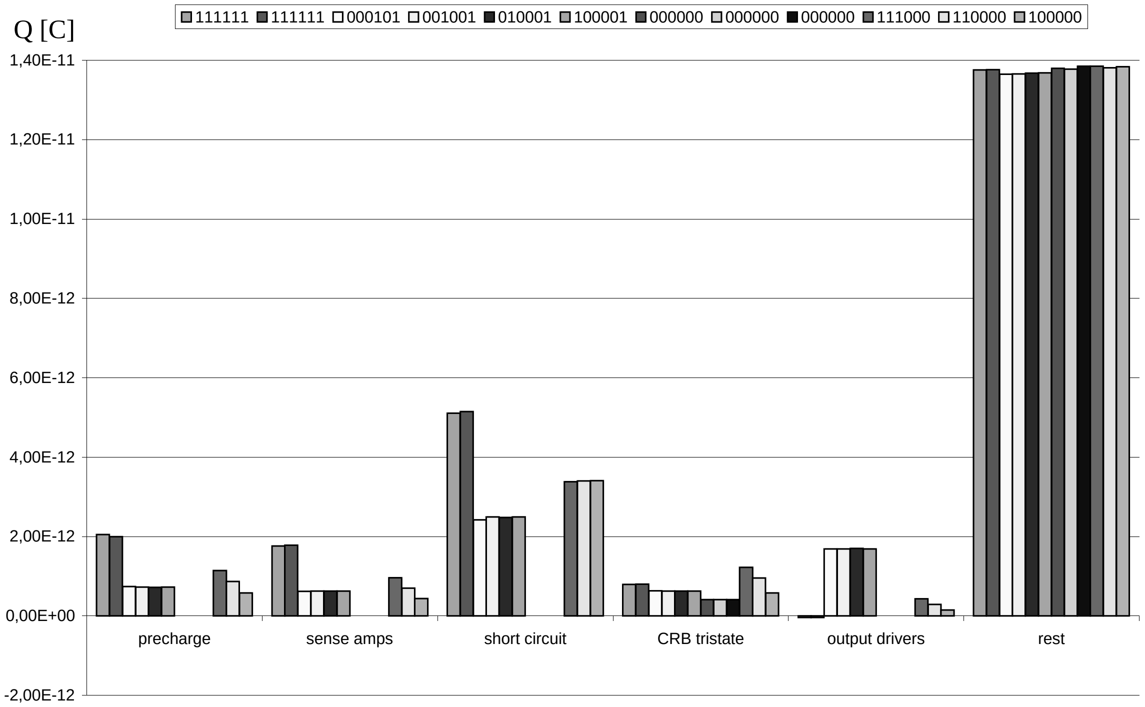

Figure 2. Power consumption of the memory activation for 12 cycles. The structural blocks exhibiting the most variance are the precharge circuitry, the sense amps, the short circuit part (see below), the common read bus tristate and the output drivers. The complete rest has nearly constant dissipation.

3.1 Memory activation power

The power linked to the memory activation was discovered to be independent from the accessed address. It is however data dependent. Figure 2 illustrates the data dependency for 12 different cycles. For the first six measurements the preceding word read was 000000 and for the last six 111111. The word read during the measured cycles is depicted. The figures and tables presented in this paper stem from a ROM instance with 16 words of 6 bits in the X dimension, 4 words in the Y dimension and only one memory block (Z dimension). The figure shows clearly that the influence of the data is restricted to only a few structural components of the memory. These components are:

1. The sense amplifiers, which amplify the voltage differences in the memory matrix during the read to produce correct logic levels.
2. The short circuit current of the accidentally sensed word. This is a current specific to the analyzed ROM architecture (see below).
3. The precharge of the memory bit lines.

4. The tristate driver of the common read bus.
5. The output tristate drivers (output was unactivated during measurements).

The variances of power consumption in the just mentioned structural components had to be quantified and linked to the concrete data for modeling. A statistical trend analysis rendered the following results:

X category	slope [C/weight]	intercept [C]	correlation
$\lvert w_s \rvert$	6.030E-13	1.372E-13	0.996
$\lvert w_a \rvert$	8.345E-13	6.130E-14	0.998
$\lvert w_p \rvert$	6.519E-14	4.117E-13	0.998
$R(w_p,w_s)$	2.720E-13	4.077E-13	1.000
$F(w_p,w_s)$	-4.392E-14	8.066E-13	-0.995

Table 1. Linear coefficients for power dissipation of memory activation.

1. In the memory blocks the power dissipation is linearly dependent on the weight of the read word w_s (cf. table 2 for linear coefficients and the standard correlation coefficients).
2. The edges on the common read bus during the memory cycle determine the power of both, the common read bus and the output tristate drivers. The trends are again strongly linear. With activated output during the cycle (E = 1) a different set of coefficients is necessary than for the cycle with unactivated output (E = 0). This reflects the different activity on the output for these two situations.
3. In the Philips memories a second data word is accidentally read internally during each cycle. The address of this word has a fixed relationship to the address really targeted. This accidentally read word w_a has no impact on the memory functionality, but has an associated power consumption that grows linearly with the weight of the word.

The remaining power is completely unvarying between cycles (cf. figure 2).

3.2 Output switching power

The analysis results for the output activation and deactivation are typical for tristate driver situations. The power consumption has two separate parts. The first part consists of the power associated with the switching of the tristate driver logic, which grows linearly with the weight of w_s for both the activation and deactivation. The second part consists of the power associated with driving the output load when the output changes from w_{op} to w_s, and is a linear function of the number of rising and falling edges on the output. See table 2 for the trends.

X category	slope [C/weight]	intercept [C]	correlation
$\|w_s\|$ (activation)	4.071E-14	1.238E-12	0.998
$\|w_s\|$ (deactivation)	-4.490E-15	7.638E-13	-0.989
$R(w_{op},w_s)$	7.750E-14	1.786E-15	0.999
$F(w_{op},w_s)$	1.393E-14	8.107E-14	1.000

Table 2. Linear coefficients for power dissipation of output (de-)/activation.

3.3 Address switching

It was already mentioned that a part of the address decoders work synchronously. Their power consumption is consequently linked to the memory activation and therefore already accounted for in the preceding sections.

The remaining asynchronous power dissipation of the address lines is about 1% of the power consumed during a memory cycle. Although this may seem neglicable, multiple address changes may occur without the occurence of a single memory activation and a considerable amount of energy might be dissipated. Hence, correct modeling of the asynchronous part is necessary.

The modeling of this asynchronous part is complicated by the spatial correlation that exist between subsets of address bits. Figure 3 shows an example of this correlation in a simple 2-to-4 decoder. The power dissipation in the uncorrelated left part can be captured for each address bit separately. However, the power dissipa-

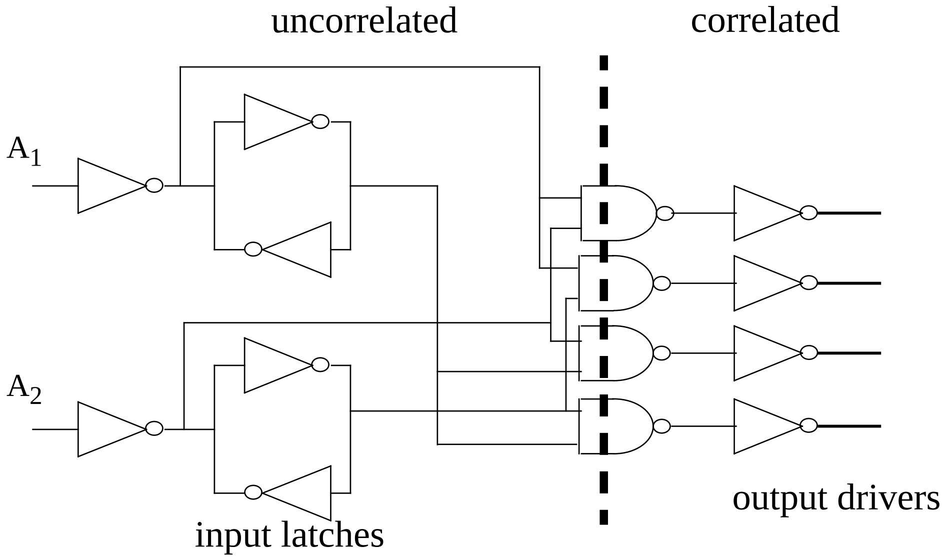

Figure 3. Example of a 2-to-4 address predecoder block with a cut between uncorrelated and correlated parts of the logic.

tion in the correlated right part depends on the values of all address bits. Clearly, the amount of different values grows exponentially with the size of the decoders. Fortunately, the maximum amount of correlated address bits is bounded as described in section 4.2.

4 ROM power model

The dependencies described were captured in a complex set of structurally motivated model equations. For the integration of this model into an event driven simulator, two major questions have to be answered: (1) how are the dependencies reflected in an easy to use simulation model and 2) what event triggers what part of the power consumption? In this section we will answer these questions.

From the analysis performed, we can divide the ROM power into four parts

$$P_{ROM} = P_m + P_a + P_o + P_s \tag{1}$$

where P_m represents the memory activation power, P_a represents the power associated with asynchronous address changes, P_o represents the output power and P_s is the static (constant) power.

In the model we will assume linear dependencies only. In the following sections we will use c_i to denote constant values (slopes or offsets).

4.1 Memory activation power P_m

The memory activation power in itself consists of three parts viz. (1) the sensing related precharge power, precharge short circuit power and sense amp power, (2) the CRB tristate driver power and finally (3) the synchronous part of the output driver (which is dependent on the state of E), and can be written as

$$P_m = c_1|w_s| + c_2|w_p| + c_3|w_a| + c_4 R(w_p, w_s) + c_5 F(w_p, w_s) + c_6 \qquad (2)$$

The coefficients c_i for this and the following equations are obtained from characterization. Note that two sets of coefficients $c_1, \ldots, c_6$ exist, one set for $E = 1$ when the output is driven and one set for $E = 0$ when only the tristate input is driven. The power P_m is dissipated each time the ROM is activated.

4.2 Address change power P_a

The asynchronous address change power can be modeled using two complementary constructions. Firstly, a power contribution $c_{A_i r}$ is associated with each rising event of address bit A_i and similarly a power contribution $c_{A_i f}$ is associated with each falling event of address bit A_i.

The second type of contribution is more complex due to spatial correlation between subsets of address bits in the address decoders (cf. section 3.3). As described before the address consists of an X, Y and a Z part. The bits in the Z part are uncorrelated, and their associated power dissipation is already fully captured by the contributions described above. However, this is not true for the X and Y part. The bits in the X respectively Y part can be partitioned such that all bits in a subpartition are correlated and no two subpartitions are correlated. This partition is uniquely determined by the address decoding hardware of the ROM. For each subpartition we will have an additional power contribution, the value of which depends on the actual bit pattern. Since the maximum amount of bits in any subpartition is 3 (minimum is 2 since, by definition, single-bit subpartitions indicate uncorrelated bits), at maximum 8 values have to be stored per subpartition and hence this poses no complexity problem.

Let $d_1, d_2, \ldots, d_8$ denote the dissipation values for a given subpartition. When the address bits in the subpartition change state from a to b, an amount of power d_b is consumed. If the state changes via an intermediate value, say from a to b via x, the power d_x is only dissipated if the transient state x exists sufficiently long. Indeed we need to ensure that each value is held for a predefined amount of time before adding the associated power offset.

4.3 Output related power P_o

The output related power shows a typical tristate behavior and can be written as

$$P_o = c_7|w_s| + c_8R(w_{op}, w_s) + c_9F(w_{op}, w_s) + c_{10} \tag{3}$$

Note that two sets of coefficients $c_7, \ldots, c_{10}$ exist, one set for rising E and one set for falling E. Obviously, power P_o is dissipated each time the E signal has a rising or falling event.

5 Conclusions and Future Plans

This paper presents a structure based analysis of the power consumption of on-chip ROMs and the derivation of a model which is applicable for accurate gate-level simulations. It was shown clearly, that the data-dependency of the power consumption can be approximated by a linear function of few parameters with good accuracy. The model error is below 5% for synchronous memory cycles and below 15% for asynchronous events on the input address bits. Although this paper concentrates on one ROM family only, the results can be seen as typical in most respects. This is mainly due to the high regularity of ROM structures, symmetry properties of memory acccesses and the small influence of parasetic effects. Good understanding of the circuit structure is however necessary for the description of the predecoder architecture and other design specific features. The analysis strategy presented here serves well at capturing these individual aspects.

It has to be mentioned critically, that the estimation accuracy for large memories cannot be assessed by simulation. Thus for large ROMs physical measurements will have to show the accuracy of our model.

Future work is directed towards the investigation of other memory types, bigger memories and modeling on higher levels of abstraction.

6 Literature

[1] D. Rabe, G. Jochens, L. Kruse and W. Nebel, ”Power-Simulation of Cell Based ASICS: Accuracy- and Performance Trade-Offs”, DATE 1998.

[2] W. Roethig, A.M. Zarkesh and M. Andrews, ”Power and Timing Modeling for ASIC Designs”, DATE 1998.

[3] C. Svensson and D. Liu, ”A Power Estimation Tool and Prospects of Power Savings in CMOS VLSI Chips”, IWLPD 1994.

[4] S. Wuytack, F. Catthoor, F. Franssen, L. Nachtergaele and H. De Man, ”Global communication and memory optimizing transformations for low power systems”, IWLPD 1994.

[5] T. Xanthopoulos, Y. Yaoi and A. Chandrakasan, ”Architectural Exploration Using Verilog-Based Power Estimation: A Case Study of the IDCT”, DAC 1997.

[6] Diesel User Guide, Philips ED&T, V1.1.0.

[7] B. Prince and G. Due-Gundersen, ’’Semiconductor Memories“, Wiley, New York, 1983.