卒業論文

Validation of an SRAM Power Model
for Architecture Level Power Estimation

アーキテクチャレベル電力推定のための SRAM モデルの検証

平成　16 年 2 月 12 日提出

指導教官　坂井修一教授

電子工学科

20393 Fransiscus Asisi Doni Januar Nowo Nugroho

Abstract

As power consumption is becoming a more important issue, several power estimation tools have been proposed. However, effectiveness of a power model is not really known without validating it against a real design. In this work, we validate a recently developed architecture level power estimation tool, PRESTO, against our SRAM design. This tool offers flexibility to adapt to various process technology, through the use of SPICE parameters and metal process parameters to calculate circuit's parasitic capacitances. Our validation shows that the model gives reasonable accuracy for relative power estimates, but still shows discouraging results for absolute power estimates.

Acknowledgement

First and foremost, this work is dedicated with whole-hearted gratitude to God in heaven, the source of all wisdom and knowledge, who has given me the ability to learn and the chance to study in Japan.

I would like to thank my advisor, Prof. Shuichi Sakai, for his guidance and continuous support given throughout my research. The opportunity he gave me to work on this research has become an invaluable learning process for me. I am also grateful to Prof. Hidehiko Tanaka, whose advices and insightful comments have helped and encouraged me greatly to become a better researcher.

I am full of gratitude to Chitaka Iwama, who has been very supportive to me in every way, more than I could ever ask. She taught me from the very basic, the things I need to know to do this research. This thesis would not have been completed without her invaluable insights and dedication. I also owe her for the development of the power estimation tool – PRESTO – that I used in this research.

I deeply appreciate Niko Demus Barli for his leadership in NEKO group. His thoughtful advices have helped me so many times, both in my research, and in my daily life. I thank him for always showing genuine interest and concern to everyone in the group.

I am greatly indebted to the lab members who were involved in designing our first SRAM chip – Niko Demus Barli, Chitaka Iwama, Naoya Hatta, Luong Dinh Hung, Yi Ge, Masanori Takada, Takashi Toyoshima – for their dedicated work. A special thank goes to Naoya Hatta, my colleague, who worked on the large part of the SRAM's layout which is used in this research. Without his dedicated work, this research would not have been completed.

I want to express my gratitude also to Naoya Hattori for his invaluable advices in thesis writing and presentation slides making. I also thank the lab members who reviewed this thesis for their thoughtful advices.

Special thanks go to Mr. Shu Shimizu, Mrs. Harumi Yagihara, Mrs. Aya Tamaru, Ms. Akiko Shida, and Ms. Asako Kawanishi for their dedicated works that enable the research in the lab to run smoothly.

Last but not least, I would like to thank my parents and sisters in Indonesia, whose continuous love and encouragements have helped me overcome many things during my study in Japan.

This research is partially funded by Grant-in-Aid for Fundamental Scientific Research B(2) #13480077 from Ministry of Education, Culture, Sports, and Science and Technology

Japan, Semiconductor Technology Academic Research Center (STARC), CREST project of Japan Science and Technology Corporation, and by 21st century of COE project of Japan Society for the Promotion of Science. The VLSI chip in this study has been fabricated in the chip fabrication program of VLSI Design and Education Center(VDEC), the University of Tokyo in collaboration with Hitachi Ltd., Dai Nippon Printing Corporation, Cadence Design System, Inc., and Synopsys, Inc.

Contents

Chapter 1

Introduction

1.1 Background

Power consumption has emerged as a central issue in VLSI circuits. The continuous demand of higher performance, higher reliability, lower cost, and portability of electronic devices cannot be met without taking considerable efforts to optimize power consumption. To devise power reduction methods or make right decisions about some design trade-offs, we need techniques or tools that can estimate power efficiently and accurately.

Power estimation is done at different abstraction levels. Generally, the lower the abstraction level the more accurate the estimation will be. This is because more structural details are available for consideration. But usually lower level power estimation is more complex and takes longer time. Power estimation at architecture level is particularly important, because :

- power optimization at this level will produce large energy saving [1];
- architectural decisions need to be made at early design phase, before circuits are implemented;
- low-level simulation is too time-consuming, hence not practical for large systems.

In [8], Ghiasi and Grunwald compared two architecture level power models, and indicated some issues concerning architecture level power estimation tools. Two of the issues they mentioned are the uncertainty of the models' accuracy against real hardware structures and the need of improving the current accuracy. Developing an efficient power model that can be applied to a wide variety of design with constantly reasonable accuracy is still a challenging, yet important task.

PRESTO [2] is a newly developed architecture level power tool. PRESTO uses low-level parameters to calculate parasitic capacitances, which are then used to calculate dynamic power of the system. This approach offers several advantages, such as flexibility to adapt with future technology. That can be done easily by changing the parameter set for the corresponding technology, which is taken from SPICE model card or metal process

parameter. The use of multi-level power modeling may also ease the modeling of various designs. Unfortunately, the accuracy of this approach in estimating power in real circuit design is still unknown.

1.2 Thesis Contribution

In this thesis, we will validate a power model for SRAM implemented in PRESTO against a real SRAM implementation. The SRAM was laid out on HITACHI 0.18μm process technology. We simulate the extracted netlist of the SRAM using SPICE, an industry standard tool for circuit analysis, and measure power consumed in different parts of the SRAM circuit. The results are used to quantify PRESTO's accuracy and investigate underlying problems.

The results show that given the detailed design information of the underlying circuits, the power model still does not show satisfactory accuracy for absolute value estimation. However, we also see that it gives good predictions for relative power .

The remaining chapters are organized as follows. In Chapter 2, we explain PRESTO's approach in more details and present several related works. In Chapter 3, we explain our validation methodology and the SRAM design we assume for the validation. Chapter 4 presents our validation results, followed with some discussions. We will include discussion on accuracy-determining issues that need to be addressed to improve estimation accuracy. Finally, we conclude our paper in Chapter 5.

Chapter 2

Approaches for Architecture Level Power Estimation

2.1 PRESTO Power Model

Dynamic power consumption can be estimated with the well-known equation $P = \alpha CV^2 f$. With frequency f and voltage swing V known, estimating power consumption becomes a matter of estimating capacitance C and switching activity α.

PRESTO's main feature is the derivation of parasitic capacitances from low-level parameters, i.e. SPICE model parameters and metal process parameters. Given SPICE model cards of a particular process technology, MOS capacitances are calculated using some reduced forms of MOS model equations.

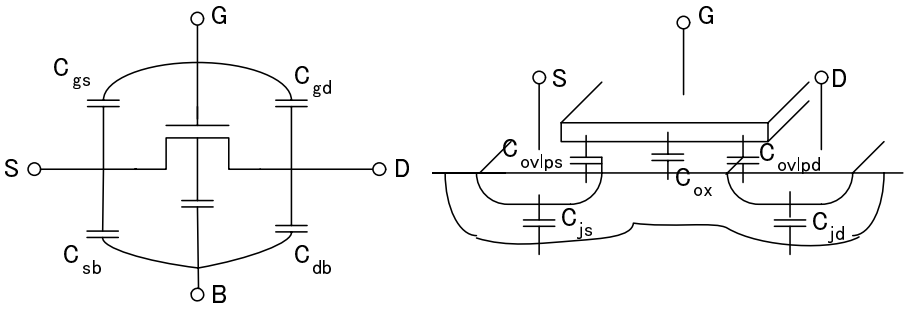

Figure 2.1: MOSFET Parasitic Capacitance Model

Fig. 2.1(left) shows the parasitic capacitance of a MOSFET represented by five capacitors – $C_{gs}, C_{gd}, C_{gb}, C_{sb}, C_{db}$ – connected to MOSFET's four terminals, i.e. gate, source, drain, and bulk. The right figure shows a more physical model of MOS capacitance, where C_{ox} is the gate-oxide capacitance, C_{ovlp} are the gate-overlap capacitances, and C_j are the junction capacitances. The amount and contribution of these capacitances to the terminals

depend on the electrical state of the MOSFET. Consequently, the parasitic capacitance of MOSFET is non-linearly dependent on its terminal voltages.

SPICE is able to simulate many circuit characteristics, including the non-linearity of parasitic capacitances, and is useful to do many circuit analyses very accurately. However, depending on circuit's complexity, the simulations can be too time-consuming and impractical for architecture level estimation. PRESTO aims for a faster power estimation with reasonable accuracy by using some simplified equations, based on SPICE parameters.

In PRESTO, it is assumed that the gate-oxide capacitance is distributed equally to the source and drain. Source and drain capacitances to bulk are considered to be equal to their respective junction capacitances. Hence, we use the following equations to calculate C_{gd} and C_{db}.

$$C_{gd} = C_{ovlpd} + C_{ox}/2 \tag{2.1}$$

$$C_{db} = C_{jd} \tag{2.2}$$

where C_{ovlpd}, C_{ox} is the capacitance due to gate-drain overlap and the gate-oxide capacitance respectively.

The gate-to-channel capacitance lumped the source region C_{gs} can be calculated from gate-source overlap capacitance C_{ovlps} and C_{ox} in the same manner. Gate-bulk capacitance C_{gb} is considered to be zero. Therefore, gate capacitance and drain capacitance can be calculated as follows :

$$C_{gate} = C_{ovlpd} + C_{ovlps} + C_{ox} \tag{2.3}$$

$$C_{drain} = C_{ovlpd} + C_{jd} + C_{ox}/2 \tag{2.4}$$

$C_{ox}, C_{ovlpd}, C_{ovlps}$, and C_{jd} are calculated using BSIM3 model [15]. Appendix A gives more details of PRESTO's capacitance calculation.

One important feature of the model is the use of low-level knowledge – such as MOS parameters, the circuit design used – to estimate power at higher levels. For example, SRAM is built of some subcircuits, such as SRAM cells, writer units, and sense amplifiers (SA) etc., which in turn are built of some logic gates or MOS transistors. At the bottom level, PRESTO uses SPICE parameters to calculate device parasitics and metal geometry parameters to calculate interconnect capacitances.

Another important concept is the estimation of switching activity (AF : activity factor). It is defined as the probability of a component to make a power-dissipating transition. In PRESTO, AF can be assigned separately for each node, which can be a logic gate or a transistor. Switching activity of a node is dependent on its input signals. To calculate the signal activity at intermediate nodes in combinational circuits, we adopt the probabilistic technique as described in [14].

2.2 Related Work

A number of high-level power estimation techniques have been proposed. Those techniques can be classified into two groups : top-down and bottom-up [4]. In top-down techniques,

internal structural details of the circuit are unknown. Conversely, bottom-up methods use the knowledge of internal structural details of the circuits to estimate power. We can also classify the methods as analytical methods and empirical methods. Analytical methods attempt to relate power consumption with fundamental quantities, such as physical capacitances and signal activity of a design. Conversely, the strategy in empirical methods is to measure the power consumption of some existing implementations and use the measurement results to build power models. Some techniques combine those two methods. In [1], Evans found that using CV^2 prediction for digital subsections, and fitted simulation results for analog subsections was a good approach for predicting absolute energy consumption and optimum organization of SRAMs. Refer to [10], for instance, for a survey on high-level power estimation methods.

In [9], Coumeri and Thomas use simulation data and linear regression to develop memory models for energy, delay, and area. They showed that using the simulation results of only ten different sized memories was sufficient to obtain reasonably accurate estimations for a large span of possible memory sizes. They claimed that the average errors are within 15%. This approach can be useful to investigate memory power in connection with size parameters change. Another advantage is that detailed knowledge of the underlying circuit is not necessary (top-down technique). Schmidt *et al.* [3] also use similar approach, but they also extend the model to also reflect nonlinear dependencies.

Hezavei *et al.* [5] developed a primary input oriented high-level power estimation technique. They use simulations to investigate the power consumption caused by all possible input transitions of relatively small circuits. They also proposed the use of those results, combined with structural information to estimate the power consumption of large circuits. However, in general, power models that are built based on simulation results do not facilitate adaptation for different process technology. Thus, when process technology changed, we need to redo the simulations for the targeted technology and rebuild the models.

A popular example of tools that use analytical power models is CACTI [6]. CACTI was originally developed to analyze cache access time, but later extended to also model cache area and power consumption. Several other tools, such as Wattch [11], and PRESTO itself use models based on or similar to CACTI. CACTI uses several embedded parameters measured in $0.8\mu m$ technology, and assumes linear scaling for both power and delay with technology feature size. This assumption could induce significant error in deep submicron technology [2], thus eliminated in PRESTO.

Modeling switching activity is an important factor in modeling power consumption at architecture level. Discussions on this can be found for example in [7]. In [4], Gupta and Najm use a low-level (typically gate-level) description to build a four-dimentional table-based model. The power model takes into account circuit input switching activity and the input correlation to estimate power dissipation of combinational logic circuits.

Chapter 3

Validation Methodology

In this chapter we will explain the methodology we used to validate the power model, including the SRAM design we used and the validation flow.
We designed a 4-kBytes SRAM using HITACHI 0.18 μm process technology. Table 3.1 shows some details of the design. Each four columns are multiplexed into a one-bit I/O circuit. For bit line's sensing we use two-stage sense amplifiers. The SRAM organization is depicted in Fig. 3.1.
For validation, we measured power dissipation of our SRAM design using three different ways. First, we ran HSPICE circuit simulations on the implemented schematic. The circuit schematic was implemented on the same design parameters as the layout. Diffusion area (AD) and diffusion periphery (PD) were also taken into account in schematic simulations. These two parameters are useful for calculating device parasitics. Wire capacitances were calculated from wire dimensions and metal process parameters, and represented as their equivalent capacitors. However, we cannot observe the effect of metal's fringing and coupling capacitances only using extractions at this level.

Second, we extracted the circuit netlist of each component from our layout, and performed similar analysis. The SPICE model that is used to calculate MOS capacitance is the same as the one in schematic simulation. But, as all device parameters, including AD and PD, are obtained directly from netlist extraction, we can expect more accurate value for device parasitic estimates. In addition, metal capacitances are also extracted directly

Table 3.1: Specification of the SRAM design used for validation

Size (rows × cols)	256 × 128
No. of Port	1
Bits per word	32
Sub-banking	None
Freq (simulated)	500MHz
V_{dd}	1.8 V

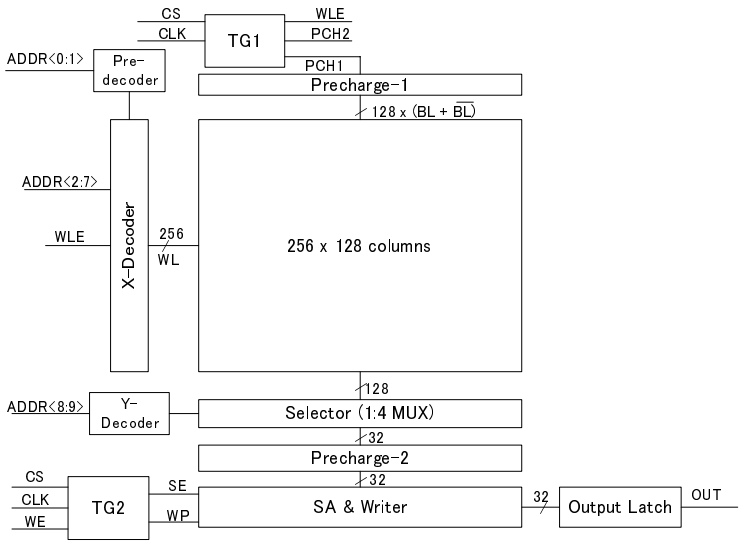

Figure 3.1: SRAM Organization

from layout. Thus, the effect of metal's fringing and coupling capacitances can also be measured. Therefore, in this work, we hold that layout simulation gives the most accurate power estimation, and use it as the target for validation.

Finally, we measured the SRAM power consumption using a power model implemented in PRESTO. The MOS capacitance in PRESTO is calculated from SPICE parameters with some simplifications as described in previous chapter. Inaccuracy in this capacitance estimation is one factor that will cause some difference between the simulation and the calculation results. Metal capacitance is calculated in the same way as that of the schematic level simulation. Fig. 3.2 shows the flow of our validation method.

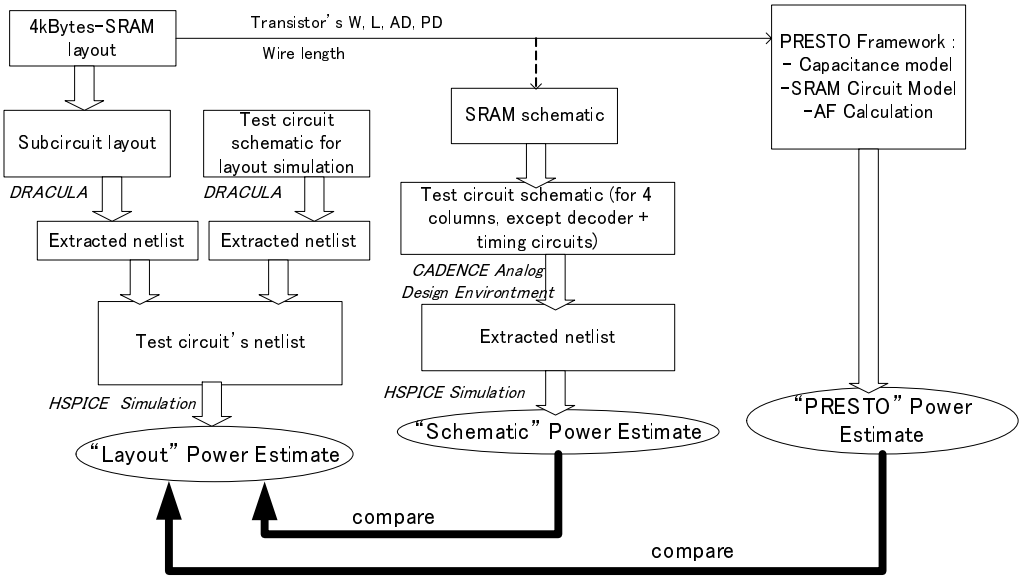

Figure 3.2: Validation flow of the SRAM power model

Chapter 4

SRAM Power Consumption

In this chapter, we present some power measurement results obtained using the methods described in the previous section and use them to validate the power model implemented in PRESTO. First, we will confirm the power model for each component, then we will also discuss the accuracy against overall power consumption.

4.1 Bit Line Circuits

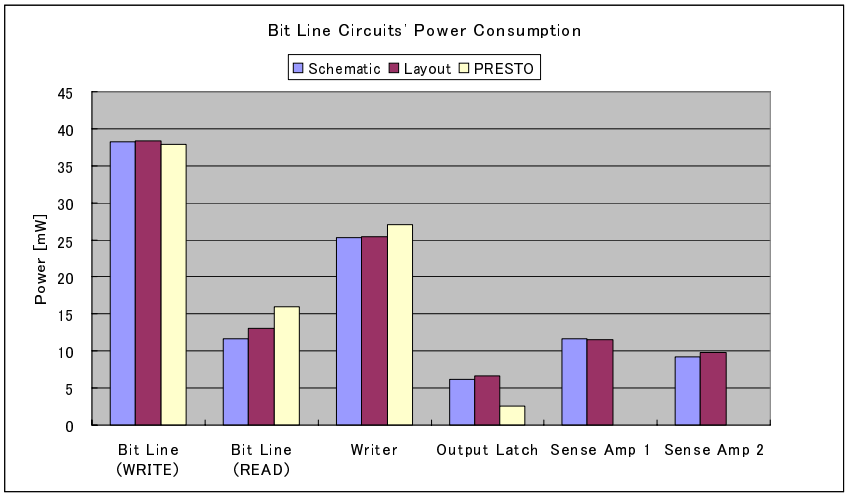

Figure 4.1: Power Consumption in Bit Line Circuits

Fig. 4.1 shows the power consumed in bit line circuits. Measurement results according to schematic simulation, layout simulation, and PRESTO's estimates are presented from left to right. Power consumption of the components that show data dependence is presented as its average. However, the power consumption of sense amplifiers varies greatly

according to the design, and its modeling requires expertise in analog circuits. Thus, it is not included in this thesis. When it is needed for comparison, we will use the simulation result for sense amplifiers' power consumption.

In a read operation, schematic simulation results differ from layout simulation results by 0.6% to about 11%, with an average of approximately 4%. However, in a write operation the difference is small, i.e. within 1% for all components. We hold that those differences come mainly from the difference in metal capacitance model. We observe that the largest difference was seen in bit line, as the effect of wire capacitance is the largest. Unless the analysis on wire capacitance is very essential, schematic simulation can still give useful information for analyzing power.

On the other hand, PRESTO's estimates differ from layout simulation results by the average of 31% for a read operation, and 12% for a write operation. The summary of the model of each component and possible sources of error are discussed below.

4.1.1 Bit Line

Fig. 4.2 shows the schematic diagram of bit line. A read operation in the SRAM design we assume starts with precharging all bit lines (and their counterparts, say -bit lines)high and enabling the wordline decoders. As one wordline is asserted, each RAM cell in that line will attempt to pull down either bit or -bit line depending on the data stored in it. This causes a small voltage difference between the bit line pair, which will be amplified by sense amplifiers. Then, the data bits are sent out through output latches. Therefore, in PRESTO, bit lines energy consumption in a read access can be modeled as :

$$BL_{(read)}Energy = C_{bit} \times V_{bitr} \times V_{dd} \times N_{col} \tag{4.1}$$

where C_{bit} is the bit line's capacitance, V_{bitr} is the voltage swing in a read operation, V_{dd} is the supply voltage, and N_{col} is the number of columns, which is equal to 256 in our design.

In a write operation, according to the data to be written, one side of the SRAM cell in the selected bit lines have to be pulled down under the RAM-cell inverter threshold, and the other side has to be pulled up. The pulling down is done through the pull-down transistor in write-drivers. However, unselected bit lines are also pulled down through access transistors and the n-channel pull-down on SRAM cells, as in a read operation. Accordingly, bit lines energy consumption in a write cycle can be modeled as :

$$BL_{(write)}Energy = C_{bit} \times (1/4 \times V_{bitw} + 3/4 \times V_{bitr}) \times V_{dd} \times 256 \tag{4.2}$$

where V_{bitw} is the voltage swing in a write operation and the other terms are as in eq. 4.1. V_{bitw} can sometimes be well approximated as V_{dd}. The factors 1/4 and 3/4 appear because one out of four bit line pairs is selected in each one-bit I/O circuit.

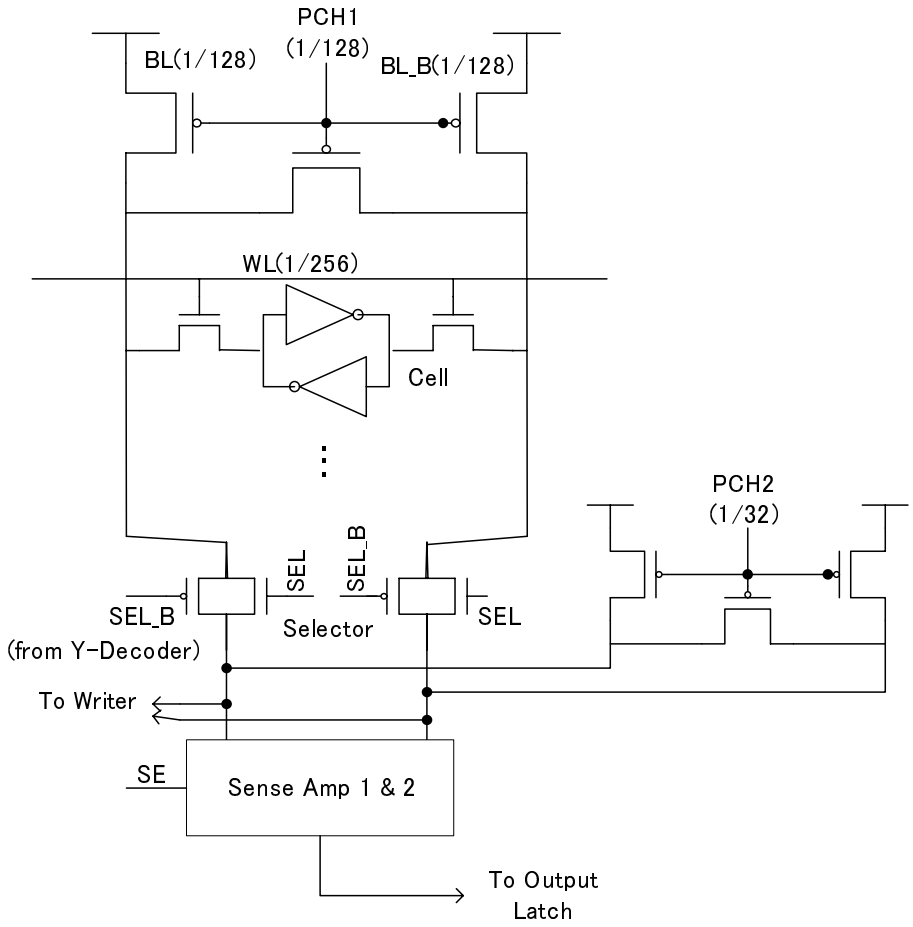

Figure 4.2: The schematic diagram of bit line

With these assumptions, our model overestimates bit line's power by 22% on a read operation, but underestimates it slightly by 1.3% on a write operation. Some possible sources of error in bit line's power estimates are listed as follows:

- *Drain capacitance model of pass transistors.*
  As there is only one word line asserted in every access, most of the pass transistors operate in the 'cut-off' region. Thus, the drain capacitance of pass transistors is approximated by C_{ovlp}. The drain capacitance of pass transistors is the most dominant component in bit line capacitance. It accounts for almost half of bit line's total capacitance in our SRAM design. Inaccuracy in the modeling of pass transistor's drain capacitances may have been the most significant error source in bit line's power estimate.

- *Variation in voltage swings.*
  In PRESTO we assume that voltage swings of all bit lines are uniform. However, waveform observation of bit line's voltage shows that the voltage of selected bit lines and unselected bit lines swing differently. This is probably because, the selected bit lines are also pulled up by the secondary precharge circuits (PCH2), thus make their voltage swings smaller than the unselected bit lines. We observed that the variation can be as much as 40% of the maximum voltage swings. Currently, as we used the value of selected bit line's voltage swing for calculation, the effect of voltage swing may have been underestimated. In write operation, the difference is smaller, because selected bit lines make (almost) full voltage swings.

- *Error in estimating the drain capacitance of selector's pass gates.*
  It is because the MOS transistors of selector's pass gates operate in different region. In the selected bit lines, as SE (sense-amp enable signal) is asserted, pass gate transistors will become 'on', that is moving its operation region from 'cut-off' to linear or saturated. However, the pass gate transistors in the other columns will remain 'off'. Hence, the effective drain capacitances will vary with the operating region. However, as this fact was ignored in PRESTO, current model tends to overestimate these capacitances.

- *Parasitic capacitances of write-drivers' internal nodes and sense amplifiers' pull-down transistors.* Error in these capacitance's estimates tend to overestimate the secondary precharger's capacitive loads.

- *Fringing and coupling capacitances.*
  Their effect is still not considered in current model. However, comparison between layout and schematic simulation results has shown that their effect is not as large as device capacitances' for the process technology we used. In the future, as wire cannot scale down in the same manner as transistor, their effect will become larger and thus metal capacitance modeling will become more significant.

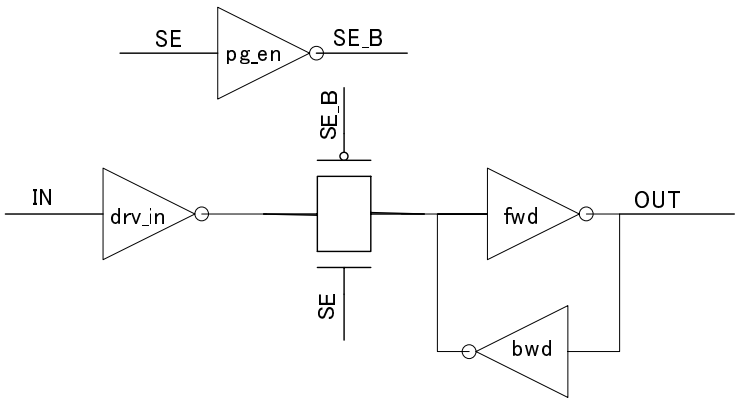

Figure 4.3: Output Latch

4.1.2 Latch

The power consumption of output latches in Fig. 4.1 was calculated in PRESTO with the assumption that written data – '0' or '1' – have equal probability, and the input signals are fully digital (similar to clock signal).Thus, dynamic power of an output latch will be dissipated by the three inverters in Fig. 4.3, drv_{in}, fwd, bwd, with the switching probability of 0.25.

However, simulation results show that these assumptions resulted in large errors of more than 100% on average power consumption. Fig. 4.4 presents latch's energy consumption for some cases, according to SPICE simulation and PRESTO. The results show that for all cases, the power model largely underestimated the power consumption. We also see that for three out of four cases we investigated, the value is between 6 mW and 7 mW.

From waveform observation we noticed that the activity factor 0.25 was not a proper estimation, because the input of *drv_in* which comes from sense amplifier's output is not a regular pulse signal, with its value fluctuating at intermediate voltages between 0 and V_{dd}. This signal affects the power consumption of *drv_in* directly. We observed that the power consumption of latch is very dependent on data. The power consumed in output latch for each read data pattern is shown in Fig. 4.5.

We had assumed that when the data read was not inverted, *drv_in* does not make any power dissipating transition ($0 \rightarrow 1$ transition). However, as mentioned above, *drv_in*'s input signal does fluctuate continuously. This is due to precharging activities and write operations. Hence, the actual power consumption is larger than our average estimation by far.

If we assume that *drv_in* makes a power dissipating transition every cycle (the case for Maximum in Fig. 4.4), we get a much better estimation with 7.5% of error against layout estimate on average power. Fig. 4.4 may also imply that when the data do not strongly incline to one value, this assumption can be used to estimate latch's power.

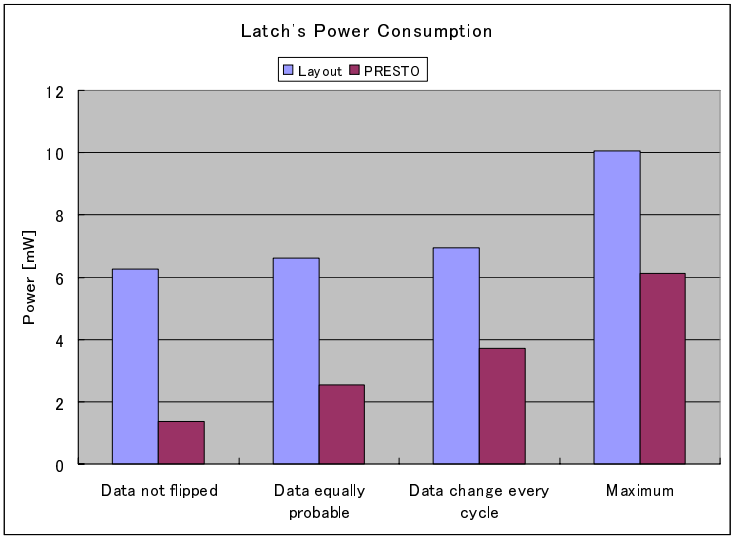

Figure 4.4: Latch's Power Consumption

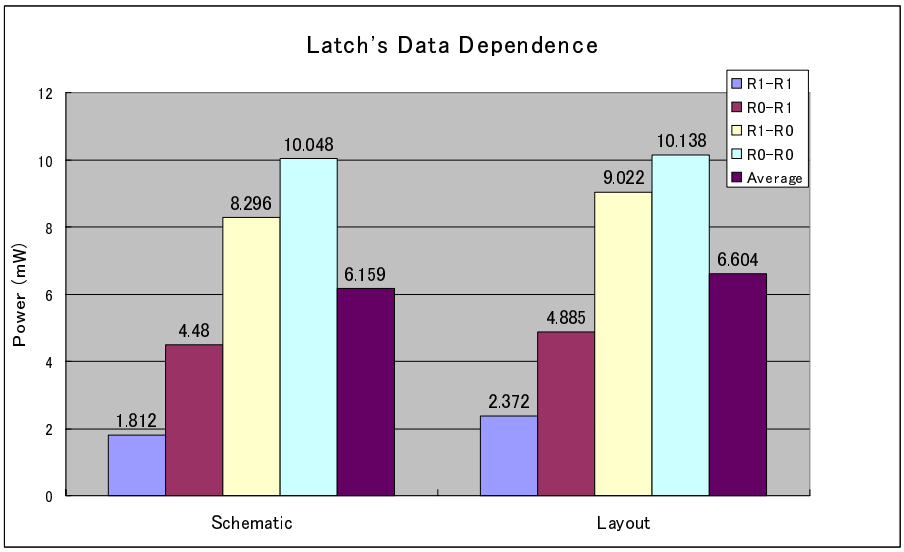

Figure 4.5: Data Dependence of Latch's Power Consumption

However, to explain the data dependence of latch adequately, further investigation is needed. Another option is using empirical approach in modeling components that involves analog signals, as proposed by Evans [1].

4.1.3 Writer

Fig. 4.6 shows the writer circuit used in our design. First, the average power consumption of the writer unit was estimated assuming equal probability of data. Thus, each inverter in data drivers switches with the activity factor of '0.25'. In write-enable drivers, the input signal (WP), moves along with clock signal during a write operation. Thus, as the inverters here make one power dissipating transition every cycle, they are given the activity factor 1. The write driver consists of two tri-state buffers, with one connected to bit line with the input *data_in*, and the other one connected to their respective complements. At each write cycle, one side of the bit line pair will be pulled down through write driver's pull-down transistor while the other side be kept high. When the data to write flipped, the load capacitances of one of the tri-state buffers will be charged while the other will be discharged. Therefore, as the probability of data to flip is '0.5', we set the activity factor of each tri-state buffer to '0.25'. Using these assumptions, PRESTO overestimates the average power consumption by 7%.

Simulation results confirmed that the power consumption of the writer unit is dependent on whether the data to write flips or not. Fig. 4.7 shows the schematic simulation results to investigate the data dependence of writer and bit line. The results show that in two successive write operation, the power consumed when data was flipped is larger – 1.1 times for bit line, and almost 5.0 times for writer unit – compared with the power consumed when data was constant. The strong data dependence seen in writer unit is attributed to the activity of data-drivers and write-drivers (tri-state buffers in Fig. 4.6. When the data does not flip, no switching activity happens in data-drivers. The power consumed in these drivers accounts for almost 30% of total power consumed in writer unit. We also confirmed that the power consumed in tri-state buffers is also negligibly small when the data does not flip. When the data change, these drivers account for about 40% of total writer's power consumption.

4.2 Timing Circuits

Our SRAM design employs two timing generators. The first timing generator (TG1) controls the word line enabling and precharging activity. The second timing generator (TG2) controls the sense amplifiers (SE signal) and write enabling (WP signal).

As input signals usually make full voltage swings, and only have a few patterns, it is relatively easy to estimate the gates' switching activity. For example, if we assume that the chip-select (CS) signal is always high, precharging and word enabling signals are determined by the clock signal. For TG2, we assume an equal probability for SE and WP. Therefore, we believe that the only significant source of error in timing circuits' power estimates is the inaccuracy in capacitance estimation. Fig. 4.8 shows the power consumed

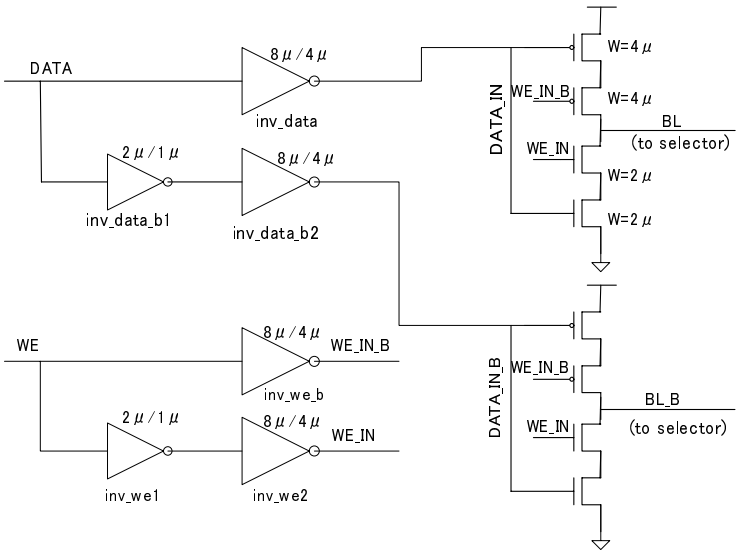

Figure 4.6: Writer Circuit

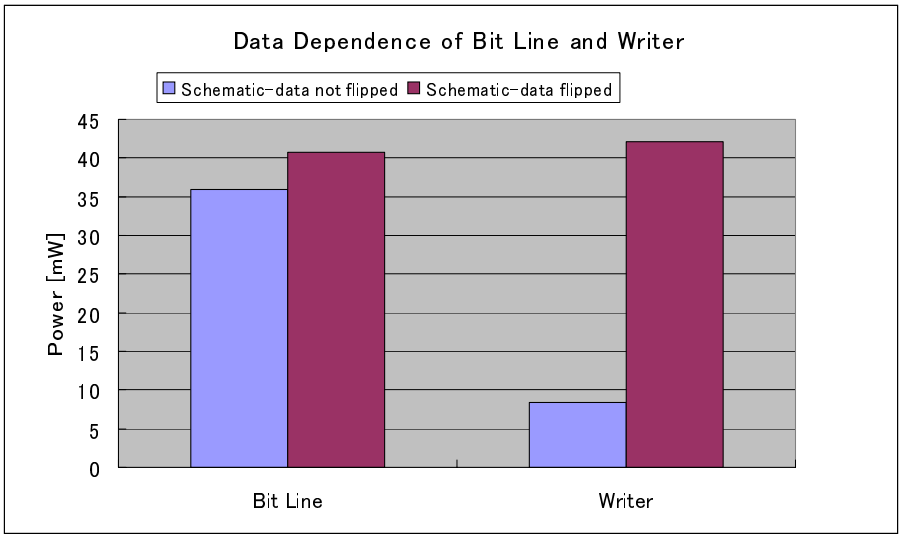

Figure 4.7: Data Dependence of Bit Line and Writer Power Consumption

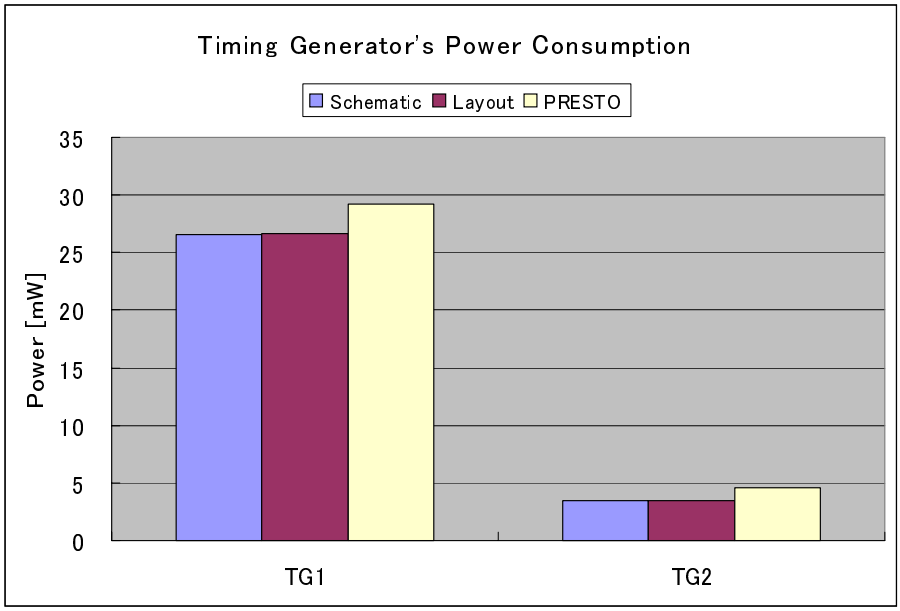

Figure 4.8: Timing Generator's Power Consumption

in the timing circuits. Validation results show less than 10% of error on TG1's power estimation and about 31% of error on TG2's power estimation. Control circuits, such as timing generators usually need to drive large capacitive loads, that they use large inverters. For a circuit like these, the inaccuracy in parasitic capacitance models can result in large errors. We believe that the largest source of error was the inaccuracy in estimating the load capacitance of the final drivers. In case of TG1, the power consumed by the final inverters occupies almost 40% of its total power consumption.

4.3 Decoder

The decoder we use in our SRAM design consists of three components: Predecoder, X-Decoder, and Y-Decoder. Combination of Predecoder and X-Decoder will select one out of 256 word lines. Y-Decoder will select one out of four multiplexed columns. Fig. 4.9 shows the Y-Decoder(Predecoder) circuit. X-Decoder circuit is shown in Fig. 4.10

4.3.1 Our Model

We investigated PRESTO's accuracy to estimate minimum and maximum power consumption of decoder. We also calculated the average power when address inputs are random. In case of minimum power consumption we assumed that one address input to X-decoder flipped, the case in which minimum switching power is consumed according to simulation

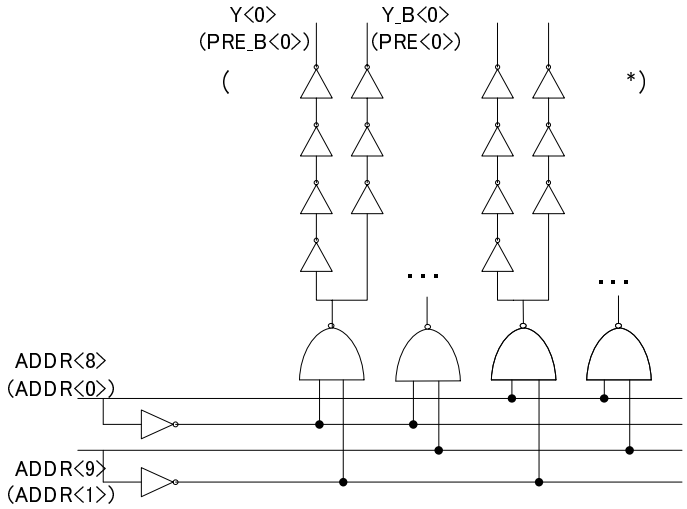

Figure 4.9: Y-Decoder(Predecoder) Circuit Schematic

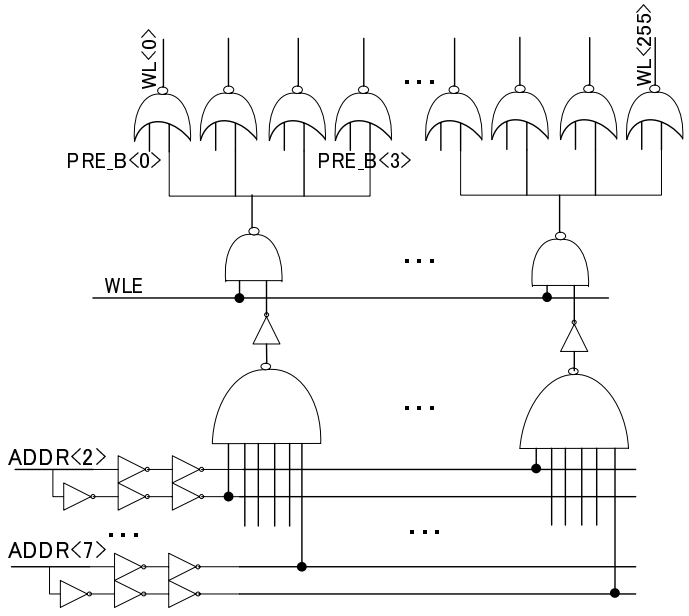

Figure 4.10: X-Decoder Circuit Schematic

result. In case of maximum power consumption, we assumed the transition pattern that causes the most transitions in decoder's gates, that is when all address inputs make $1 \rightarrow 0$ transition simultaneously while WLE(word-line enable) signal is kept high. For average power estimation, we assumed equal probability, thus set the activity factor of all addresss inputs to "0.5", then calculate the activity of each intermediate node.

4.3.2 Simulation Results

The power consumption of a decoder shows strong dependence on input patterns. However, considering all possible input patterns is too complex and time consuming. Thus, in our analysis, we assumed that all address inputs that are connected to similar components are symmetrical. For instance, ADDR$\langle 0 \rangle$ and ADDR$\langle 1 \rangle$ are symmetrical because they both are connected to X-Decoder. We investigated the effect of address change for each component by measuring the power consumption when a number of address transition occured while other address inputs to different components are fixed. Fig. 4.11 summarizes our measurement results.

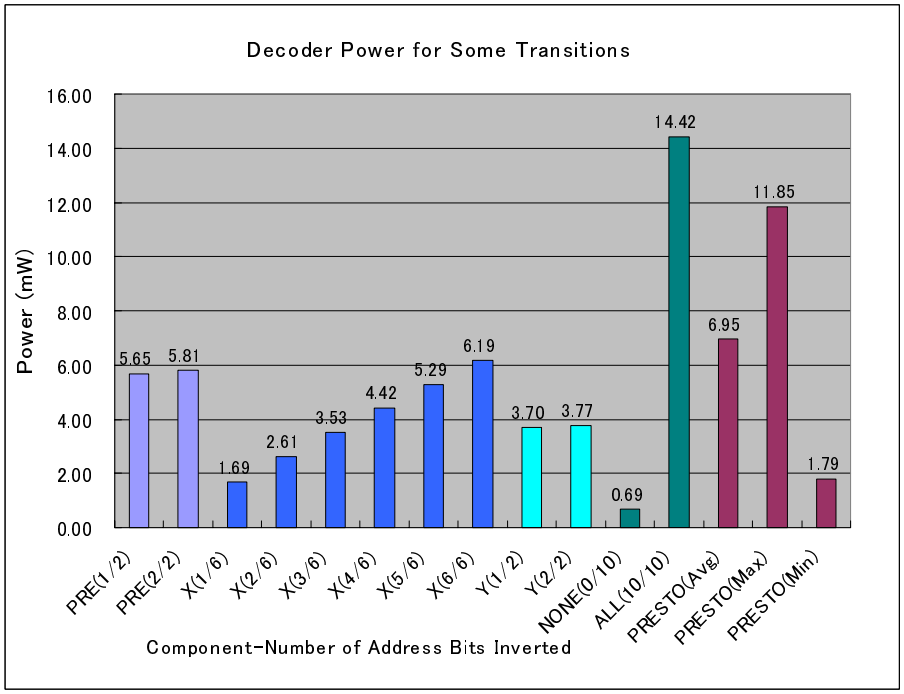

Figure 4.11: Decoder Power Consumption for Several Transitions and PRESTO's Estimation

The X-axis shows the place and number of address transition, except for the last three bars which represent PRESTO's estimates. For example, $PRE(1/2)$ in the graph represents the case when one out of two address inputs in Predecoder makes transition.

Maximum switching power is consumed when all address inputs changed simultaneously ("ALL(10/10)" in Fig. 4.11). On the other hand, minimum switching power is consumed when only one address input to X-Decoder changed ("X(1/2)" in Fig. 4.11). Simulation results show a wide range of decoder's power consumption with the maximum power is around 8.5 times greater than the minimum power.

Our model overestimates the minimum power consumption of decoder by 6%, but underestimates the maximum by 18%. The sources of error will be discussed below.

When there is no address changed, switching power is not dissipated either in Predecoder or Y-Decoder. In X-decoder, however, as one of its input, i.e. WLE, moves along with clock signal, switchings still happen in the selected line's drivers (one NAND and one NOR gate). We noted that as there is only one word line selected at one time, there are only a few possible switching patterns for the gates after the first NAND gates of each component. Hence, the variation in decoder's power consumption comes mainly from the switching power of address drivers. Because all final address drivers are connected to NAND gates' inputs, the model of NAND's gate capacitance seems to be one determining factor of the power estimation's accuracy. Another source of error is the estimation of word line's capacitance, which includes 256 access transistors' gate capacitances and the word line's wire capacitance.

4.4 Total Power Consumption

Fig. 4.12 presents SRAM's total power consumption in a read and write operation, assuming minimum or maximum decoder power. We observe that PRESTO overestimates the total power consumption with less than 6% of error for all cases. PRESTO can also be used to estimate relative power consumption of each SRAM component. Table 4.1 summarizes the relative contribution of each component to SRAM's power consumption, according to layout simulation and PRESTO. We see reasonable accuracy, except for latch, where PRESTO underestimates its relative power by 60%. We also find that PRESTO predicts accurately the WRITE-READ power ratio of our SRAM design.

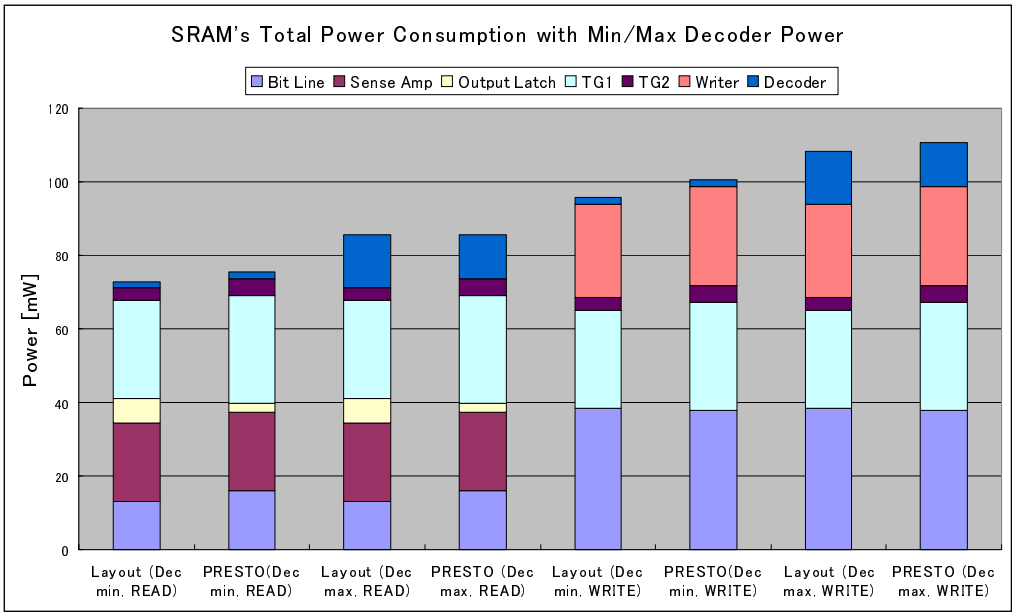

Figure 4.12: SRAM's Total Power Consumption

Table 4.1: SPICE v.s. PRESTO for relative power consumption.
[For components other than decoder, decoder power is not included in the percentage calculation.]

	SPICE	PRESTO
Read Power Breakdown		
Timing Circuits	43%	46%
Decoder	2%-20%	2%-16%
Bit Line	18%	22%
Output Latch	9%	3%
Write Power Breakdown		
Timing Circuits	32%	35%
Decoder	2%-13%	2%-11%
Bit Line	41%	38%
Writer	27%	27%
Write Power v.s. Read Power Ratio		
	1.4	1.4

Chapter 5

Conclusions

In this research, we validated a power model implemented in PRESTO framework against SPICE. To estimate power consumption, the model calculates device parasitic capacitances directly from SPICE parameters. It also estimates interconnect capacitances from metal process parameters. This approach offers flexibility to adapt to future device and metal process technology. In this work we used an SRAM design we implemented using 0.18 μm process technology to validate the power model and investigate the model's approach.

The approach shows variations in power estimation's accuracy of different SRAM components, from about 62% of underestimation for output latch to about 31% of overestimation for the second timing generator. The model shows reasonable accuracy, within 10%, in power estimates for writer unit and the first timing generator, that controls word-line enabling and precharging activity. But, it shows discouraging results in absolute power estimation for other components.

However, the power model model gives good estimations on relative power consumption. Thus, it can be used to predict the relative contribution of each component of SRAM, as well as READ/WRITE power ratio.

We noticed that using this approach, transistor's capacitance modeling accuracy seems to have the most impact on power estimation accuracy. We observed that large errors are especially seen in the components with large drivers and large capacitive loads. In circuits like these, the modeling of the final driver's capacitive loads is really significant because the largest part of power may be consumed in the final drivers. Another large error is seen in output latch, due to the nature of its input. We conclude that the current model is not suited to estimate the power consumption of circuits with analog input signals.

In this work, we discussed some possible error sources for each component, which mainly comes from voltage-dependence of parasitic capacitances. The next step of this research will be to develop a simple model to reflect the voltage dependence of transistor's parasitic capacitances. Validation on different SRAM designs or other structures is also needed. But, when we still focus on device capacitance modeling, this can be done without mask-layout implementation. In our validation some parameters we used for calculation are still obtained from schematic/layout implementation, such as voltage swing

and metal lengths. However, to use the power model in architecture level, we need to include some methods to estimate the value of those parameters. Inclusion of multi-layer metal capacitance model is also one important step to cope with the future technology.

Bibliography

[1] R.J. Evans and P.D. Franzon, "Energy Consumption Modeling and Optimization for SRAM's," IEEE Journal of Solid-State Circuits, vol. 30 pages 571-579, May 1995.

[2] C. Iwama, L.D. Hung, N.D. Barli, S. Sakai, and H. Tanaka,"The Design of PRESTO: A Framework for Architecture Level Power Estimation," 情報処理学会研究報告 2003-ARC-154 (SWoPP 松江 2003), pages 103-108, August 2003.

[3] E. Schmidt, G. Jochens, L. Kruse, F. Theeuwen, and W. Nebel,"Automatic Nonlinear Memory Power Modelling," Procedings of Conference on Design, Automation and Test in Europe(DATE), March 2001

[4] S. Gupta and F.N Najm,"Power Modeling for High Level Power Estimation," IEEE Transactions on VLSI Systems, Vol. 1, February 2000.

[5] J. Hezavei, Vijaykrishnan N., M.J Irwin, M. Kandemir and D. Duarte,"Input Sensitive High-level Power Analysis," Proceedings of the 2001 IEEE Workshop on Signal Processing Systems (SiPS), pp. 149-156, Antwerp - Belgium, September 2001.

[6] P. Shivakumar and N.P. Jouppi,"CACTI 3.0: An Integrated Cache Timing, Power, and Area Model," Technical Report 2001/2, Digital Western Research Lab, 2001

[7] F.N. Najm,"Improved Estimation of the Switching Activity for Reliability Prediction in VLSI Circuits," IEEE Custom Integrated Circuits Conference, 1994.

[8] S. Ghiasi and D. Grunwald,"A Comparison of Two Architectural Power Models," In Workshop on Power-Aware Computer Systems. Cambridge, MA, Nov 2000.

[9] Sari L. Coumeri, Donald E. Thomas,"Memory Modeling for System Synthesis," In Proceedings of the 1998 international symposium on Low power electronics and design, August 1998.

[10] P. Landman,"High-level power estimation," in ISLPED'96: ACM/IEEE Int. Symp. Low-Power Electronics and Design, Monterey, CA, Aug. 1996, pp. 29–35.

[11] D. Brooks, V. Tiwari V., and M. Martonosi,"Wattch: A Framework for Architectural-Level Power Analysis and Optimizations", in Proceedings of the 27th International Symposium on Computer Architecture, ISCA'00, June 2000, pp. 83-94.

[12] J.M Rabaey *et al. Digital Integrated Circuits : A Design Perspective, 2nd Ed.* Prentice Hall, 2003.

[13] N. Weste and K. Eshraghian. *Principles of CMOS VLSI Design : A System Perspective, 2nd. Ed.* Addisson-Wesley, 1993.

[14] K. Roy and S. Prasad. *Low-Power CMOS VLSI Circuit Design.* John-Wiley & Sons, Inc., 2000.

[15] J. Huang, Z. Liu, M. Jeng, K. Hui, M. Chan, P. Ko and C. Hu. *BSIM3 Manual (version 2).* University of California, Berkeley, 1994

Appendix A

Capacitance Model

This appendix describes how PRESTO calculates C_{ox}, C_{ovlp}, and C_j, using SPICE parameters. In all equations, SPICE parameters are written in bold. C_{ox} is calculated as parallel-plate gate oxide capacitance, and is proportional to gate area.

$$C_{ox} = \mathbf{COX} \cdot W_{eff} L_{eff} = \frac{\varepsilon}{\mathbf{TOX}} \cdot W_{eff} L_{eff} \tag{A.1}$$

C_{ovlpd} is the gate parasitic due to the gate-drain overlap.

$$C_{ovlpd} = (\mathbf{COX} \cdot \mathbf{LD} + \mathbf{CGDO}) \cdot W_{eff} \tag{A.2}$$

The gate-source overlap C_{ovlps}, is also calculated with the same equation.

Junction capacitance C_{jd} is composed of bottom wall capacitance C_{jw} and sidewall capacitance C_{jsw}.

$$C_{jd} = C_{jw}\mathbf{AD} + C_{jsw}\mathbf{PD} \tag{A.3}$$

$$C_{js} = C_{jw}\mathbf{AS} + C_{jsw}\mathbf{PS} \tag{A.4}$$

AD (**AS**) and **PD** (**PS**) is area and perimeter of the drain(source) diffusion respectively.

In SPICE, C_{jw} and C_{jsw} are given as a function of reverse bias voltage V_R between diffusion and substrate.

$$C_{jw}(V_R) = \mathbf{CJ}(1 + \frac{V_R}{\mathbf{PB}})^{-\mathbf{MJ}} \tag{A.5}$$

$$C_{jsw}(V_R) = \mathbf{CJSW}(1 + \frac{V_R}{\mathbf{PBSW}})^{-\mathbf{MJSW}} \tag{A.6}$$

However, in PRESTO, they are represented by the average values at $0 \leq V_R \leq V_{dd}$:

$$C_{jw} = \mathbf{CJ} \cdot \frac{\mathbf{PB}}{V_{dd}(1 - \mathbf{MJ})}((1 + \frac{V_{dd}}{\mathbf{PB}})^{1-\mathbf{MJ}} - 1) \tag{A.7}$$

$$C_{jsw} = \mathbf{CJSW} \cdot \frac{\mathbf{PBSW}}{V_{dd}(1 - \mathbf{MJSW})}((1 + \frac{V_{dd}}{\mathbf{PBSW}})^{1-\mathbf{MJSW}} - 1) \tag{A.8}$$