Memory Modeling for System Synthesis

Sari L. Coumeri and Donald E. Thomas, Jr., *Fellow, IEEE*

***Abstract*—We present our methodology for developing models of on-chip SRAM memory organizations. The models were created to enable the quick evaluation of energy, area, and performance of different memory configurations considered during synthesis. The models are defined in terms of parameters, such as size and mode of operation, which are known at synthesis time. Our methodology does not require knowledge of the underlying memory circuitry and provides models with average percentage errors within 8%. We examine the importance of the different parameters in the models to reduce the time required to develop the models. We found that only ten different memories from a large span of possible memory sizes are needed to obtain reasonably accurate models, with average errors within 15%. In this paper, we present our modeling methodology, discuss the important aspects in developing the models, and examine the parameters necessary in creating accurate models quickly and easily.**

***Index Terms*—Memory, power-consumption model, special low-power99, system-level.**

I. Introduction

POWER consumption of digital systems has become a critical design parameter. Extending battery life in portable applications and reducing cooling requirements in higher transistor density applications make power reduction a crucial consideration during digital system design.

An important class of digital systems include applications, such as video image processing and speech recognition, which are extremely memory-intensive. In such systems, a significant amount of power is consumed during memory accesses. Thus, utilizing low-power memory organizations can greatly reduce the overall power consumption of the system.

This work targets on-chip memories created by memory module generators in which there are many possible memory organizations in terms of size, architecture, technology, etc. To utilize low-power memory configurations during synthesis, we need models to quickly evaluate memory energy, area, and performance. These models need to be in terms of parameters, such as size, organization, and mode of operation, which are known during synthesis time as opposed to lower level parameters such as extracted capacitance and resistance values. This type of model allows us to make predictions during behavioral synthesis and explore a large portion of the design space.

In the past few years, various memory models have been presented. Itoh [1] and Kamble [2] have presented analytical models of memory power. Ko [3] did a measurement-based characterization in which the power of a few different memories were measured. Evans [4] compared five different approaches for modeling the energy of SRAM's and used the models to analyze different internal architectures. Ogawa [5] used circuit reduction techniques for faster characterization of power and delay of SRAM's. Chinosi [6] developed a technique for the automatic characterization of memory power for different modes of operations for a certain-sized memory. Landman [7] used a simulation and model-fitting approach to develop power models in terms of the number of words and the bit width.

Our models were developed to predict energy, delay, and area across the different possible sizes and organizations produced by memory module generators and for different modes of memory operations. Our modeling uses a simulation-based approach which enables the development of black box models. Unlike analytical models, simulation-based approaches do not require detailed knowledge of the underlying circuits, just basic input/output timing information which can be provided by the memory generator. The models are in terms of high-level parameters and can be easily used during synthesis. Our approach is similar to [7], but is generalized enough to handle more complex memory organizations and more modes of memory operation with higher accuracy.

The focus of this paper is on our modeling methodology as well as defining the parameters and simulations necessary in building accurate models which can be used during synthesis quickly and easily. This paper begins by describing our experimental methodology and showing the statistical results of the developed models. Next, we examine the important components of the models. Finally, we discuss additional models developed using the methodology.

II. Experimental Methodology

Our modeling methodology begins with memories which are generated using Duet's Epoch memory module generator (formerly Cascade) [8]. Next, test vectors are automatically generated and SPICE files are modified to prepare them for simulations. Then, Avant!'s Star-Sim, a fast circuit simulator, is used to simulate for energy and delay [9]. From the simulation data, models of memory energy, delay, and area are developed using linear regression with the S-Plus statistical package [10]. Finally, the models are validated to ensure they are statistically sound.

A. Generated Memories

Duet's basic asynchronous SRAM's with chip and output enables were used. To generate a memory in Duet, the number of address lines, the number of words, and the bit width are specified. Additionally, the number of bits per column (BPC),

Manuscript received February 15, 1999. This work was supported in part by the Semiconductor Research Corporation under Contracts 068-048 and 068-073.

S. L. Coumeri is with the Compaq Computer Corporation, Shrewsbury, MA 01545 USA.

D. E. Thomas, Jr. is with the Department of Electrical and Computer Engineering, Carnegie-Mellon University, Pittsburgh, PA 15213 USA (e-mail: thomas@ece.cmu.edu).

Publisher Item Identifier S 1063-8210(00)04348-1.

which gives you control over the aspect ratio of the memory, can be specified. The legal BPC values in Duet are 1, 2, 4, 8, and 16. Therefore, a unique memory size is not defined just by the number of words and bit width. It is defined by the number of rows, number of columns, and bit width, where

$$\text{Rows} = \text{Words/BPC} \tag{1}$$

$$\text{Columns} = \text{Bitwidth} \cdot \text{BPC} \tag{2}$$

$$\text{RowAddressLines} = \text{ceil}(\lg(\text{Rows})) \tag{3}$$

$$\text{ColumnAddressLines} = \lg(\text{BPC}). \tag{4}$$

The number of rows can range from 4 to 256, the number of columns from 1 to 256, and the bit width from 1 to 256. By varying the number of rows, columns, and bit width there are 62 992 different possible legal memory sizes. It is obviously impossible to simulate all possibilities, so a subset must be chosen. Twenty-five different basic Duet SRAM's were generated in .6μ technology with a 3.3-V supply. The largest and smallest memories were included and the rest 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, BPC, and total number of storage bits in the memory.

B. Memory Simulations

After obtaining the SPICE files from the generated memories, Star-Sim simulations were run to measure energy and delay. During these runs data was collected for the various modes of the memories. Since an entire memory was simulated at once, as opposed to separate simulations for the different pieces of the memories, creating the test vectors for the simulations was easy. Knowledge of the memories' internal circuitry was not required, only the I/O timing information supplied by Duet. These simulations are necessary because the delay estimates provided by Duet are overly conservative and the power estimates do not account for different memory modes of operation.

1) Energy Simulations: The average energy per operation was measured. Read and write energy were treated separately. The energy was also measured while the chip and output enable lines were toggling and for different levels of switching activity on the address lines.

The hierarchical SPICE netlist was instrumented to separately measure the energy of the different memory components. The separate components in the memories are the address transition detection (ATD) logic, memory cells, chip enable multiplexors, row and column decoders, precharge logic, sense amps, and extra buffers.

With a write operation, there were two additional parameters to consider. Once the address changes at the start of a write cycle, the write enable line must remain high for the address setup time. Next, the write enable line is lowered during which time the data is written. Finally, the write enable line is raised for the address hold time before the address changes to again start the next cycle.

TABLE I
PARAMETERS USED IN MODELS

Abbr.	Parameter	Models Used In
Rows	# of Rows	energy, delay, area
Cols	# of Columns	energy, delay, area
BW	Bit Width (width of data word)	energy, delay, area
Addr	# of Address Lines	energy, delay, area
R.Addr	# of Row Address Lines	energy, delay, area
C.Addr	# of Column Address Lines	energy, delay, area
Sw	# of Addr Lines Switching per access	energy
R.Sw	# of Row Addr Lines Switching per access	energy
C.Sw	# of Column Addr Lines Switching per access	energy
CE	Chip Enable signal toggling (0, 1)	energy, delay
OE	Output Enable signal toggling (0,1)	energy
Cap	Capacitive Load (fF)	delay
WTL	Time Write Enable Held Low (ns)	write energy
RW	Extra read before write (0,1)	write energy

Due to static power dissipation, the amount of time the write enable is held low affects the energy. Additionally, the amount of time the write enable signal remains high before it is lowered can impact the energy. Since these are asynchronous memories, address transition detection (ATD) logic is used to detect a change on the address lines and start a memory access. If the write enable line remains high longer than the required address setup time, a memory read will occur before a write, resulting in additional energy.

In synchronous designs there are different ways to generate the write enable signal from the clock, each of which results in different address setup and write enable low times. Therefore, including these parameters in the models of memory energy is important.

2) Delay Simulations: The worst case delay for a memory operation was measured. The read time (the address changing to the data appearing on the output), the write bit time (write enable going low to the data being written to the memory cells), and the write out time (the write enable going low to the data appearing on the output) were measured. Duet specified values for hold and setup times were used.

Delays for when the chip enable is activated and with and without a capacitive load were measured as well. The rise and fall times of the four physical corners of the memory were measured and the worst delay for each was taken.

C. Developing Memory Models

Three categories of memory models were developed from the simulations: area, delay, and energy. All the models are linear equations in terms of parameters known during synthesis. For area, there are width and height models. For delay, there are read, write bit, write out, setup, and hold time models. For energy, there are distinct models for read and write operations.

Each energy model is composed of separate models for the components of the memory (ATD, sense amps, etc.). The sum of the individual component models forms the total energy read and write models. Having separate energy models for the

TABLE II
ACCURACY OF ENERGY, AREA, AND DELAY MODELS

Model	Model-Building Data Set		Validation Data Set			Entire Data Set		
	R^2	Residual St. Error	r^2	$\sqrt{MSPR}$	Avg |%err|	R^2	Residual St. Error	Avg |%err|
Height (μ)	.9998	15.81	.9998	14.99	1.5%	.9999	10.18	1.0%
Width (μ)	.9999	9.59	.9994	16.94	1.8%	.9999	6.70	0.9%
Write Bit Time (s)	.9755	9.78e-11	.8920	1.99e-10	6.1%	.9746	9.17e-11	4.1%
Write Out Time (s)	.9762	1.06e-10	.9318	1.69e-10	5.5%	.9733	1.06e-10	4.2%
Read Time (s)	.9703	1.50e-10	.9059	2.26e-10	3.4%	.9722	1.31e-10	2.5%
Read Energy (J)	.9963	1.57e-11	.9879	2.31e-11	8.4%	.9952	1.43e-11	5.9%
Write Energy (J)	.9990	3.40e-11	.9864	4.77e-11	13%	.9981	3.04e-11	5.8%
Weighted Read Energy (J)	.9953	1.75e-11	.9855	2.68e-11	7.8%	.9936	1.65e-11	5.2%
Weighted Write Energy (J)	.9975	5.32e-11	.9866	4.60e-11	6.1%	.9979	3.22e-11	3.3%
Simple Read Energy (J)	.9831	2.79e-11	.9424	4.76e-11	20%	.9590	3.91e-11	20%
Simple Write Energy (J)	.8340	3.87e-10	.6939	5.91e-10	69%	.6075	4.17e-10	34%
Non-Comp. Read Energy (J)	.9959	1.40e-11	.9810	2.68e-11	9.6%	.9935	1.56e-11	6.2%
Non-Comp. Write Energy (J)	.9963	5.84e-11	.9761	6.13e-11	10%	.9981	2.92e-11	4.4%

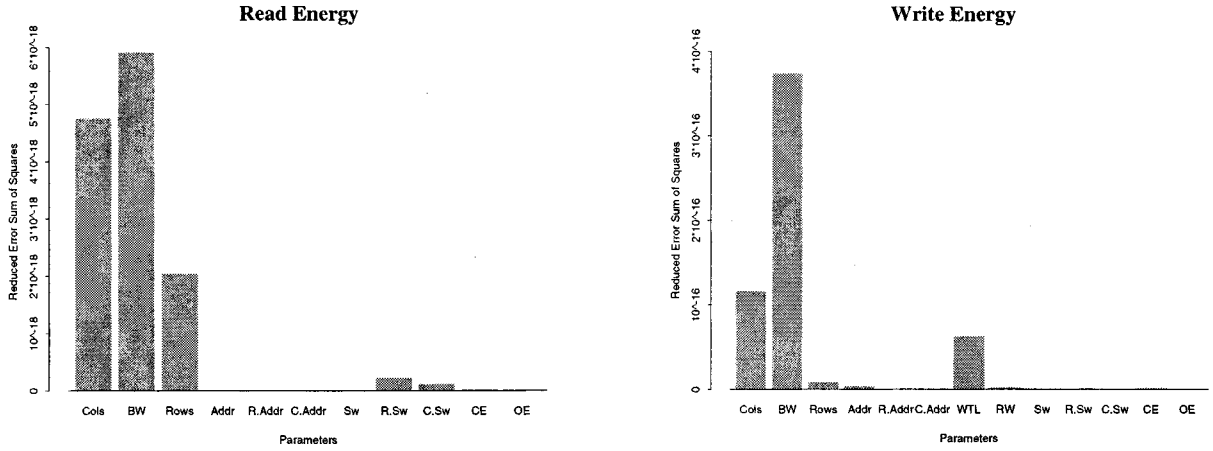

Fig. 1. Reduced error sum of squares.

different components of the memory enables us to develop more accurate models and gain more insight into the energy tradeoffs of the generated memories.

Table I summarizes the parameters used in all of the models. The size parameters are used for all of the models. The other parameters relating to the mode of operation are used for both delay and energy. CE, OE, and RW are all Boolean variables which indicate whether or not the specified action is occurring.

The models were developed using stepwise linear regression in the S-Plus statistical package. The initial models were the specified variables defined in Table I. The stepwise regression improves the initial model by iteratively adding and deleting terms. It can consider multiple interaction terms. For example, since the number of rows and number of columns are both variables specified in the model, it can consider adding the term Rows $\cdot$ Cols to the model. It adds and deletes terms based upon the AIC criteria [11], which tries to improve the coefficient of multiple determination R^2, without overfitting the model. It adds terms which nontrivially contribute to the model and removes useless terms which do not.

Using stepwise regression in our modeling methodology allows us to develop accurate models quickly and easily. It automatically determines which parameters are important to the models and finds the interactions between the independent variables. Without stepwise regression we would have to specify the form of equation, which is difficult to do with a large number of parameters and would require detailed knowledge of the underlying memory circuitry to determine the interaction between the variables.

D. Model Validation

Table II shows the statistical data for the developed models. The second and third columns have the statistics for the model-building data set, which are the coefficient of multiple determination R^2, and the residual standard error for each of the models. The area models had the best fits, followed by the energy and delay models.

Simulations for 25 additional memories were run to build a validation data set. The statistics for this set, shown in columns

Read Energy

Precharge = 1.43e-12 + 5.78e-14*Rows + 1.42e-13*Cols + 2.12e-13*BW
- 1.24e-12*C.Sw + 4.64e-15*Rows*Cols + 1.18e-15*Rows*C.Sw
+ 1.25e-13*Cols*C.Sw + -4.26e-16*Rows*Cols*C.Sw

Sense amps = 1.98e-13 - 2.81e-15*Rows + 2.96e-14*Cols + 6.07e-13*BW
+ 1.70e-14*Sw + 3.66e-15*Rows*BW + 2.41e-14*BW*Sw

ATD =5.42e-13 +3.00e-15*Rows + 4.01e-15*Cols + -2.57e-15*BW
+ 1.70e-12*R.Sw + 1.86e-12*C.Sw + 2.84e-14*Rows*R.Sw
+ 1.99e-14*Cols*C.Sw + 4.27e-14*BW*C.Sw

Buffers = -1.99e-13 + 8.00e-14*Cols + 7.67e-15*BW + 2.15e-13*Addr
- 1.16e-13*RowAddr + 8.11e-13*CE + 2.99e-13*OE + 4.82e-13*Sw
+ 5.09e-14*BW*OE + 2.71e-13*Addr*CE - 6.42e-14*R.Addr*Sw
- 1.17e-16*Cols*BW + 2.66e-14*BW*Addr - 2.40e-14*BW*R.Addr
+ 9.85e-16*Cols*Sw

Row Decoder =5.10e-13 + 3.18e-14*Rows + 2.58e-15*Cols + 9.32e-14*BW
- 8.57e-14*R.Addr + 5.78e-13*C.Addr + 1.82e-13*R.Sw - 9.69e-13*ColSw
- 9.02e-14*CE + 2.02e-13*R.Sw*CE + 7.08e-15*Cols*R.Addr
- 5.30e-15*Rows*C.Addr - 5.30e-16*Cols*BW + 5.74e-14*C.Addr*R.Sw
- 1.42e-13*R.Sw*C.Sw - 1.90e-17*Rows*BW - 8.21e-15*BW*C.Addr
+ 2.08e-14*Cols*C.Sw + 2.49e-17*Rows*Cols - 2.43e-15*Rows*R.Addr
- 4.87e-16*Rows*R.Sw + 1.68e-13*R.Addr*C.Sw - 1.12e-14*BW*R.Addr
- 5.39e-19*Rows*Cols*BW + 6.57e-16*Rows*C.Addr*R.Sw
- 3.27e-15*Cols*R.Addr*C.Sw + 7.63e-17*Cols*BW*R.Addr

Column Decoder = (C.Addr > 1)*(-6.17e-13 - 1.25e-13*Cols + 3.49e-13*BW
+ 1.08e-13*C.Addr + 4.83e-14*R.Sw + 6.61e-13*C.Sw + 2.87e-13*BW*C.Sw
+ 2.78e-14*Cols*C.Addr - 1.36e-14*Cols*C.Sw - 1.07e-13*R.Sw*C.Sw)

CE Muxes =1.63e-14 + 3.70e-17*Rows + 1.43e-17*Cols + 7.87e-17*BW
- 9.66e-16*Addr + 1.91e-14*CE + 3.50e-13*Sw + 5.69e-15*CE*Sw
- 8.15e-16*Cols*CE - 1.02e-17*BW*Sw - 5.63e-18*BW*Addr
- 1.03e-19*Cols*BW + 2.34e-14*Addr*CE - 1.40e-15*BW*CE
- 6.39e-16*Rows*CE + 3.11e-16*BW*CE*Sw + 4.22e-16*BW*Addr*CE
- 6.57e-18*Cols*BW*CE

Mem Cell =-1.83e-13 - 6.76e-15*Rows + 4.74e-15*Cols + 2.12e-14*BW
- 1.36e-13*Addr + 1.76e-13*R.Addr + 1.09e-13*R.Sw + 6.21e-13*C.Sw
+ 2.94e-14*CE - 6.97e-16*BW*R.Sw - 8.07e-15*BW*CE - 3.31e-14*BW*C.Sw
+ 6.59e-17*Cols*BW - 1.37e-14*R.Addr*R.Sw - 5.47e-15*R.Sw*C.Sw
- 8.94e-16*Cols*R.Sw + 7.09e-16*Rows*Addr - 7.13e-14*R.Addr*C.Sw
+ 5.20e-15*BW*R.Sw*C.Sw - 1.12e-17*Cols*BW*R.Sw

Write Energy

Precharge = 2.23e-12 + 7.26e-14*Rows + 1.04e-13*Cols + -1.25e-13*BW
- 4.54e-13*Addr + 4.73e-12*CE + 4.91e-13*WTL + 9.52e-13*RW
- 2.03e-14*Sw + 1.99e-13*Cols*WTL + 1.94e-15*Rows*Cols
+ 4.35e-14*Cols*RW + 8.39e-13*BW*CE + 9.64e-14*BW*Addr
+ 3.69e-14*Rows*RW + 3.82e-13*BW*RW + 2.43e-14*Cols*Sw
+ 2.35e-15*Rows*Cols*RW

Sense amps = -4.03e-14 - 7.63e-16*Rows + 3.54e-13*BW - 2.07e-14*WTL
+ 9.13e-13*RW + 3.35e-13*BW*WTL + 7.68e-13*BW*RW + 1.38e-15*Rows*BW

ATD = 2.07e-13 + 5.71e-15*Rows + 1.22e-14*Cols - 4.78e-15*BW
+ 1.62e-12*R.Sw + 1.61e-12*C.Sw + 2.83e-14*Rows*R.Sw + 1.32e-13*BW*C.Sw
+ 1.43e-14*Cols*C.Sw - 4.38e-17*Rows*Cols

Buffers = -1.53e-13 + 8.68e-14*Cols + 2.39e-13*BW + 8.72e-14*R.Addr
+ 1.24e-13*C.Addr + 4.32e-13*R.Sw + 1.34e-12*C.Sw + 1.22e-12*CE
+ 2.76e-13*OE + 2.83e-14*WTL + 1.31e-13*RW + 5.04e-14*BW*OE
+ 3.18e-14*BW*C.Addr - 1.71e-13*R.Addr*C.Sw - 1.70e-14*BW*RW
+ 2.11e-13*R.Addr*CE - 2.19e-16*Cols*BW + 4.40e-15*BW*R.Addr
+ 2.59e-13*C.Addr*CE - 1.66e-15*BW*WTL - 5.42e-14*R.Addr*R.Sw
+ 2.46e-15*Cols*OE - 2.29e-16*Cols*WTL + 1.11e-17*Cols*BW*WTL

Row Decoder = 6.57e-13 + 3.77e-15*Rows + 5.08e-14*Cols + 6.10e-15*BW
+ 8.43e-14*Addr + 6.39e-14*R.Sw - 2.73e-13*CE - 5.34e-14*WTL + 4.96e-13*RW
+ 4.84e-14*Cols*RW + 2.49e-14*R.Sw*WTL + 2.81e-13*R.Sw*CE
- 7.27e-16*Cols*R.Sw + 5.16e-17*Rows*BW - 4.24e-17*Cols*BW + 7.17e-14*Addr*RW

Column Decoder = (C.Addr > 1)*(2.69e-13 + 3.61e-14*Cols + 5.32e-13*BW
- 2.24e-14*Addr - 2.66e-14*R.Addr + 2.11e-13*R.Sw - 7.36e-13*C.Sw - 1.21e-14*WTL
+ 2.21e-13*BW*C.Sw - 3.23e-15*Cols*R.Sw + 3.61e-15*BW*WTL
- 2.51e-13*BW*Addr + 2.45e-13*BW*R.Addr - 1.15e-13*R.Sw*C.Sw
+ 1.41e-13*R.Addr*C.Sw)

CE Muxes = 3.75e-13 + 6.56e-15*Rows + 7.26e-15*Cols + 9.12e-15*BW
- 3.09e-14*Addr + 1.90e-14*R.Addr - 1.01e-13*CE + 5.93e-15*WTL + 4.87e-14*RW
+ 3.45e-13*Sw - 5.14e-17*Cols*BW - 1.11e-15*Cols*Addr + 7.08e-18*Rows*BW
+ 1.88e-14*CE*Sw + 1.21e-15*Cols*R.Addr + 1.07e-14*Addr*CE - 1.73e-16*BW*WTL
- 1.79e-17*Cols*WTL - 1.35e-17*Rows*WTL + 7.97e-17*BW*Sw
- 7.36e-16*Rows*R.Addr + 4.81e-17*Rows*Addr - 5.90e-17*Rows*Cols
- 1.85e-16*Rows*RW + 1.12e-18*Cols*BW*WTL + 1.17e-19*Rows*Cols*BW
+ 4.27e-18*Rows*Cols*Addr - 3.87e-19*Rows*BW*WTL

Mem Cell = 1.18e-13 + 2.76e-17*Rows - 1.28e-15*Cols + 5.17e-14*BW - 1.71e-14*Addr
- 3.16e-14*WTL + 4.85e-13*RW + 6.75e-14*Sw + 2.91e-14*BW*RW + 2.31e-15*BW*WTL
+ 3.63e-16*Cols*BW + 6.81e-17*Cols*Sw + 8.78e-15*Cols*RW + 5.31e-15*Rows*RW
- 1.35e-13*Addr*RW + 3.28e-17*Rows*BW + 3.40e-15*BW*Sw + 9.48e-16*Cols*WTL
- 1.24e-14*WTL*RW - 2.12e-15*WTL*Sw + 2.12e-16*Rows*WTL - 3.33e-16*Rows*Sw
+ 6.29e-16*Rows*BW*RW - 4.13e-17*Cols*BW*Sw - 3.18e-15*BW*WTL*RW
- 1.15e-16*Cols*WTL*Sw

Fig. 2. Energy models based upon entire data set.

four and five, include the square of the correlation between the measured and predicted values r^2 and the square root of the mean-squared prediction error $\sqrt{\text{MSPR}}$. These values can be compared to the R^2 and the residual standard error of the model-building data set to measure our models' predictive ability. The predictions for the energy and area models are very accurate. The accuracy drops slightly for the delay models.

The sixth column in the table shows the average absolute percentage error for all of the simulated memories. This is calculated by

$$|\%\text{err}| = |\text{MeasuredEnergy} - \text{PredictedEnergy}|/\text{MeasuredEnergy}. \quad (5)$$

The average percentage error is fairly low, but jumps to 13% for the write energy. The problem occurs because there is more than a 500 $\times$ difference in write energy between the largest and smallest data points. The extremely small memories have energy values smaller than the standard error of the equations and therefore can end up with percentage errors larger than 100%. To account for this problem, each data point i was given the following weight:

$$\text{Weight}_i = \max(\text{Energy})/\text{Energy}_i \quad (6)$$

where $\max(\text{Energy})$ is the maximum energy for all the data points and Energy_i is the energy for data point i. A weighted stepwise regression was done for the read and write energy. Rows 8 and 9 show the weighted regression results. This weighting boosts the importance of the smaller energy data points and improves the average absolute percentage error. The improvement was less than 1% for the read energy. However, the write energy absolute percentage error was cut in half.

The last set of columns is for models developed using the entire set of data. Since these models were developed with more than double the data, the error for almost all these models is improved over the original models developed from the model-building data set.

Rows 10 and 11 show results for simplified models. These models were developed doing a weighted linear regression using the equation from [7] as opposed to using stepwise regression. This equation, shown below, does not account for different aspect ratios within the memory or for different modes of operation

$$\text{Energy} = C_0 + C_1\text{Words} + C_2\text{BW} + C_3\text{Words} \cdot \text{BW}. \quad (7)$$

The simple read model had fits and standard errors slightly worse than our model. However, the simple write model was inaccurate with residual and predicted errors approximately an order of magnitude larger. The average percentage error was considerably larger for both the read and write models.

The last two rows of the table are the results for models created doing weighted stepwise regression for the total energy as opposed to separate componentized models for each portion of the memory (sense amps, ATD, etc.). The fits and standard errors were comparable for these models. However, the average

TABLE III
COMPONENTS OF ENERGY

	Read Energy		Write Energy	
Component	Variables in Model (Abbr. are defined in Table)	Avg. % of Energy (± St. Dev.)	Variables in Model	Avg. % of Energy (± St. Dev.)
Precharge	Cols, Rows, BW, C.Sw	43±9%	Cols, WTL, Rows, BW, Addr, RW, CE, Sw	46±7%
Sense amps	BW, Cols, Rows, Sw	34±14%	BW, WTL, RW, Rows	24±10%
ATD	R.Sw, Rows, C.Sw, Cols, BW	12±10%	R.Sw, Rows, C.Sw, Cols, BW	12±11%
Buffers	Cols, CE, OE, BW, Sw, Addr, R.Addr	5±2%	BW, Cols, CE, OE, C.Sw, R.Addr, WTL, R.Sw, C.Addr, RW	10±2%
Row Decoder	Cols, R.Sw, Rows, CE, R.Addr, C.Sw, C.Addr, BW	4±2%	RW, Cols, R.Sw, Rows, BW, Addr, CE, WTL	3±1%
Column Decoder	C.Sw, BW, R.Sw, Cols, C.Addr	0.8±2%	C.Sw, WTL, BW, Cols, Addr, R.Addr, R.Sw	0.8±2%
Chip Enable Muxes	Sw, CE, Cols, Addr, BW, Rows	0.8±0.8%	Sw, BW, Cols, Addr, R.Addr, CE, RW, Rows, WTL	1±1%
Memory Cell	BW, Cols, R.Sw, Rows, Addr, R.Addr, C.Sw	0.7±0.5%	BW, RW, WTL, Sw, Addr, Rows, Cols	2±1%
Entire Model	BW, Cols, Rows, R.Sw, C.Sw, CE, OE, Sw, R.Addr, Addr, C.Addr	------	BW, Cols, WTL, Rows, Addr, RW, CE, R.Sw, Sw, C.Sw, OE, R.Addr, C.Addr	------

percentage errors were worse. The read and write energy models developed from the entire data set are shown in Fig. 2 at the end of the paper.

III. IMPORTANT FACTORS OF MODELS

Using our methodology, very accurate models of energy, area, and delay were created. However, running many memory simulations can be CPU intensive. The simulations ranged from a few minutes to a few days of CPU time, depending on the size of the memory. Therefore, to create accurate models quickly and easily, it is necessary to determine which factors are most important while developing the model.

A. Parameters of Models

Table III shows the independent parameters used in each of the energy models. Type III ANOVA (analysis of variance) tables [12] were examined to see how much each independent variable reduces the sum of square error in the model. The ANOVA tables were formed from the models based upon the entire set of data. The variables in the table are listed in order of importance (from highest to lowest variance).

The ANOVA tables for all of the components show that the most important variables to the read model are the size parameters, followed by the address switching parameters, followed by the chip and output enable toggling. The most important variables to the write model are the size parameters, followed by the write mode of operation parameters, followed by the switching and toggling parameters. Fig. 1 shows a plot of how much each parameter reduces the error sum of squares in the read and write energy models. The plot shows how much more important the size parameters are compared with the mode of operation parameters.

Table III also shows the average percentage of energy consumed in each of the memory components. This was calculated by using the entire data set models to make predictions on the 62 992 different Duet memories. The precharge logic and sense amps consumed the largest average percentage of energy, 43% and 34% of the read energy and 46% and 24% of the write energy. The standard deviations for these averages are quite high. Therefore, the distribution of the energy and the effects of the different parameters vary throughout the memory design space.

Both the precharge and sense amp models are mainly dependent on size parameters and the write mode of operation parameters. The switching parameters are important for the ATD and the chip and output enable toggling parameters are important for the buffer models which consume much lower average percentages of energy. The switching parameters are significant in memory configurations with a large ratio of number of address lines to total bits of storage. Chip and output enable toggling parameters are important in memories with a low number of storage bits where the energy of the buffers is not overshadowed by the precharge and sense amp energy.

To further determine the importance of the different mode of operation parameters, models were created in which certain variables were removed. The impact of removing the switching parameters and removing the chip enable and output enable toggling parameters from the read models was examined. Removing the write time low (WTL), the read followed by a write (RW), the switching, and the chip and output enable toggling parameters from the write models was also investigated.

Each model was created with a subset of the model-building data. The rest of the model-building data plus the validation data set data were used as new validation data. For example, in the models created without chip and output enable toggling parameters, the subset of the model-building data where the chip and output enable were not toggling were used to develop the models. The toggling parameters (CE and OE) were left out of the initial models and a weighted stepwise linear regression was performed. The rest of the data from the model-building data set (where the chip and output enables were toggling) and the entire validation set were used as validation data. These newly developed models were used to make predictions on the new validation data set. The correlation between the measured and predicted values r^2, the average square root of the mean-squared prediction error $\sqrt{\text{MSPR}}$, and the average absolute percentage error are shown in Table IV.

TABLE IV
EFFECTS OF PARAMETERS ON THE MODELS

Energy Model	r^2	$\sqrt{MSPR}$	Avg |%err|
Weighted Read Energy	.9855	2.68e-11	7.8%
Weighted Read w/out Switching Parameters	.9719	3.35e-11	14%
Weighted Read w/out CE or OE Parameters	.9855	2.39e-11	8.8%
Weighted Write Energy	.9866	4.60e-11	6.1%
Weighted Write w/out WTL Parameter	.5343	4.90e-10	50%
Weighted Write w/out RW Parameter	.9793	1.27e-10	9.6%
Weighted Write w/out Switching Parameters	.9906	5.88e-11	11%
Weighted Write w/out CE or OE Parameters	.9855	5.63e-11	7.0%

The WTL parameter for the write energy has the most significant effect of all the mode of operation parameters. There is almost an order of magnitude difference in the error when this parameter is not included in the models. Including the RW parameter improves the write energy model by approximately 3%. The improvements from including the switching parameters in the models was approximately 6% for the read models and 5% for the write models. Including the chip and output enable toggling parameters offers minimal improvements. There was a 1% improvement in the average absolute percentage error by including these parameters. In fact, the $\sqrt{\text{MSPR}}$ actually improves in the read energy models when these parameters are removed.

B. Number of Memories in Data Set

Since the size parameters are the most important to the models, the next question to answer is how many different sized memories are needed to get accurate models? An experiment was conducted in which different energy models were developed from subsets of the 25 model-building data set memories. For a certain-sized subset, a weighted stepwise linear regression was run, and the rest of the data from the 50 simulated memories (model-building data set plus validation data set) were used as validation data.

Table V shows the statistical results of the subset models. There were four different sized subsets: 5, 10, 15, and 20. The sizes of the validation data sets for each of these were 45, 40, 35, and 30, respectively. For each of the subset sizes, 20 samples were run. Each sample was chosen randomly. The first group of columns shows the statistics for the read energy subset models and the second group shows the statistics for the write energy subset models. These statistics include the average square of the correlation between the measured and predicted values r^2, the average square root of the mean-squared prediction error $\sqrt{\text{MSPR}}$, and the average of the average absolute percent error.

The average predictions of the models based upon five memories are very poor. But the predictions improve significantly with ten memories in the data set. The predictions drop slightly for the sample size of 15 in the read energy models. This was due to the fact that one of the samples was really poor with an average absolute percentage error of 150%. If the outlier is removed from the samples, the predictions improve over the ten-memory sample size. (This sample did not cause a problem with the write models.) The predictions of the 20 size samples improves even further.

With just ten memories in the data set, fairly accurate models of read energy can be developed. Since the subsets of memories were chosen randomly, a 15% average absolute percentage error is an expected value. However, if some care is taken to ensure that the parameters of the memories are well distributed, poor samples can be avoided. Not having memories from a certain portion of the design space or having too many memories from one part of the design space can bias the models. In the outlier sample for the read energy, there were no memories with a small number of rows and large bit width. Therefore, the developed models were unable to predict accurately in this region of the memory space.

Subset experiments were conducted for the delay and area models as well. The delay models had similar results in that models based upon five memories were very poor, but models based upon ten memories had average absolute percentage errors within 15%. Accurate area models could be developed with errors within 5% with just five memories in the data set.

By reducing the number of memories in the data set and eliminating some of the mode of operation parameters, such as the chip and output enable toggling parameters, the CPU time due to simulations can be reduced. This enables the quick development of accurate energy, area, and delay memory models.

IV. ALTERNATE MODELS

Alternate memory models were developed using our methodology. In the first set of alternate models, the supply voltage was included as a parameter. Since voltage scaling is a technique commonly used to reduce power, the supply voltage is a useful parameter to include in the models. The second set of models was developed from memories with modified internal circuitry. These models were developed to show how the methodology is able to handle memories with a significantly different energy/delay design space.

A. Including Supply Voltage Parameter

When generating memories in Duet, the user can specify the supply voltage. Specifying different voltages will create memories with different transistor sizings. The set of memories generated for our models was designed to work with a 3.3-V supply. Instead of generating additional memories for different supply voltages, simulations with different supply voltages were run on the previously generated set. Since the memories were designed for 3.3 V, there was a limit to how far the voltage could be scaled before the memories stopped functioning properly.

Using the original set of 25 generated memories, additional energy and delay simulations were run with a supply voltage of 2.7 V. The voltage models were created with just the 3.3- and 2.7-V supply simulation results. With the validation data set, additional simulations were run with a supply of 3.0 and 2.7 V. All three voltage supply simulation results were used to validate the models. This is not a large range of voltage supplies; however, it is sufficient to see if a voltage parameter could be easily added to the models.

TABLE V
MODELS FROM SUBSETS OF DATA

Subset Size	# of Samples	Read Energy			Write Energy		
		Avg. r^2 ($\pm$ St. Dev.)	Avg. $\sqrt{MSPR}$ ($\pm$ St. Dev.)	Avg Avg |%Err| ($\pm$ St. Dev.)	Avg. r^2 ($\pm$ St. Dev.)	Avg. $\sqrt{MSPR}$ ($\pm$ St. Dev.)	Avg Avg |%Err| ($\pm$ St. Dev.)
5	20	.53$\pm$.31	6.60e-10$\pm$1.88e-09	310$\pm$616%	.90$\pm$.20	3.91e-10$\pm$1.15e-09	86.3$\pm$187%
10	20	.97$\pm$.02	3.52e-11$\pm$1.57e-11	15.3$\pm$10.3%	.98$\pm$.01	8.80e-11$\pm$2.51e-11	15.9$\pm$4.6%
15	20	.94$\pm$.16	4.60e-11$\pm$6.47e-11	17.1$\pm$31.5%	.99$\pm$.01	6.08e-11$\pm$2.05e-11	8.0$\pm$3.4%
15	19	.98$\pm$.01	3.17e-11$\pm$9.14e-12	10.1$\pm$3.7%	---	---	---
20	20	.98$\pm$.02	3.14e-11$\pm$1.04e-11	10.4$\pm$4.8%	.99$\pm$.002	4.72e-11$\pm$4.54e-12	6.5$\pm$1.1%

TABLE VI
ACCURACY OF VOLTAGE MODELS

Model	Model-Building Data Set		Validation Data Set		Avg |%err|	Entire Data Set		Avg |%err|
	R^2	Residual St. Error	r^2	$\sqrt{MSPR}$		R^2	Residual St. Error	
Write Bit Time (s)	.9832	2.58e-10	.8605	2.50e-10	6.3$\pm$6.6	.9793	2.61e-10	3.2$\pm$2.6
Write Out Time (s)	.9726	3.67e-10	.9294	2.05e-10	5.9$\pm$5.2	.9704	3.67e-10	4.1$\pm$3.3
Read Time (s)	.9777	4.11e-10	.7581	4.42e-10	5.1$\pm$7.4	.9791	3.57e-10	2.0$\pm$1.6
Read Energy (J)	.9947	1.62e-11	.9875	1.90e-11	7.9$\pm$8.5	.9935	1.50e-11	5.0$\pm$6.0
Write Energy (J)	.9989	3.09e-11	.9888	3.85e-11	5.2$\pm$5.2	.9980	2.70e-11	3.5$\pm$3.8

TABLE VII
ACCURACY OF MODELS WITH MODIFIED CIRCUITRY

Model	Model-Building Data Set		Validation Data Set		Avg |%err|	Entire Data Set		Avg |%err|
	R^2	Residual St. Error	r^2	$\sqrt{MSPR}$		R^2	Residual St. Error	
Write Bit Time (s)	.9503	1.60e-10	.7917	2.79e-10	5.5$\pm$6.0	.9247	1.84e-10	5.1$\pm$4.0
Write Out Time (s)	.8844	5.20e-10	.5790	1.02e-9	27$\pm$27	.8823	5.24e-10	13$\pm$14
Read Time (s)	.9506	1.68e-10	.8057	6.57e-10	7.8$\pm$11	.9584	1.76e-10	3.1$\pm$2.5
Read Energy (J)	.9975	7.41e-12	.9893	3.08e-11	9.3$\pm$10	.9894	1.73e-11	6.8$\pm$5.8
Write Energy (J)	.9888	2.19e-11	.9771	3.94e-11	8.6$\pm$8.0	.9808	3.28e-11	7.1$\pm$6.2

Since dynamic power dissipation is proportional to $V_{\rm dd}^2$ and static power dissipation is proportional to $V_{\rm dd}$, the voltage energy equations were forced to take this form [13]. The parameter $1/V_{\rm dd}$ was included in the initial models and the dependent variable for the regression was Energy$/V_{\rm dd}^2$. The developed models were multiplied by $V_{\rm dd}^2$ to get the energy equations. This forces each term in the energy models to include either $V_{\rm dd}$ or $V_{\rm dd}^2$.

Likewise, in first-order delay equations, the gate delay is inversely proportional to the supply voltage and the interconnect delay does not depend on the supply voltage [13]. The parameter $1/V_{\rm dd}$ was included in the initial delay models. Therefore, some terms in the delay models include $1/V_{\rm dd}$ and some terms do not.

The statistical results for the delay and energy voltage models are shown in Table VI. Using the methodology accurate energy models in which the supply voltage was taken into account were developed. The delay models were fairly accurate as well, although the r^2 values are lower for the write bit time and read time. All of the models had average absolute percentage errors within 8%.

B. Circuit Modifications

The methodology was also applied to developing models of memories with different internal circuitry. This was done to see how well the methodology works for memories with a different energy and delay design space. Work done in [14] took the Duet memories and modified the internal circuitry to improve the power dissipation. The modifications were made to the precharge logic, the sense amps, and the ATD logic. Pullup transistors were removed from the precharge logic, the differential amplifiers were removed from the sense amps, and the asynchronous ATD logic was replaced with a clock signal. These changes improve the power dissipation but increase the delay of the memories.

These circuit modifications were performed on the SPICE netlist produced by Duet and were straight forward, allowing the modification of the netlists to be automated. Fifteen memories from the original 25 generated were modified. Energy and delay simulations were run on these modified netlists and energy and delay models were developed from the results. The validation

set consisted of simulations from a subset of 15 memories from the 25 memory validation set. Subsets of size 15 were chosen for these models to decrease the amount of simulation CPU time.

The statistics for this new set of energy and delay models are shown in Table VII. Since only 15 memories were used in developing the models, the average absolute percentage errors were expected to be around 10%. The energy models and the write bit and read time delay models have errors within 10%. However, the write out time model's average absolute percentage error was high at 27%. It is likely that a different set of 15 memories could improve this model.

V. Conclusion

We have presented our modeling methodology for memory energy, area, and delay. Our methodology provides an easy and accurate way to develop memory models which can be used for synthesis without detailed knowledge of the underlying circuitry. The models developed using our technique had average percentage errors within 8%. Using a weighted stepwise linear regression technique to determine the form of the models reduced the standard error over an order of magnitude from a simplified model approach. We showed that the size parameters were the most important to consider while developing the models and that it is only necessary to simulate ten different sized memories to obtain models with average errors within 15%.

Using our methodology, we were able to develop accurate alternate memory models. There were additional models which included voltage as a parameter and models of memories with modified internal circuitry. Our methodology is generalized so memory models can be quickly developed for different module generators and then be easily used within a synthesis environment.

Although the models were developed for on-chip SRAM's, the methodology could be applied to DRAM's and special DRAM architectures such as synchronous, Rambus, and video DRAM's. This would require adding more parameters or models to capture the additional modes of operation.

Acknowledgment

The authors would like to thank M. Posner for his statistical advice and Prof. Herman Schmit for his insightful discussions.

References

[1] K. Itoh, K. Sasaki, and Y. Nakagome, "Trends in low-power RAM circuit technologies," *Proc. IEEE*, vol. 83, pp. 524–543, Apr. 1995.
[2] M. B. Kamble and K. Ghose, "Analytical energy dissipation models for low power caches," in *Int. Symp. Low Power Design*, 1997, pp. 143–148.
[3] U. Ko and P. T. Balsara, "Characterization and design of a low-power, high-performance cache architecture," in *Int. Symp. VLSI Technology, Systems, and Applications*, 1995, pp. 235–238.
[4] R. J. Evans and P. D. Franzon, "Energy consumption modeling and optimization for SRAM's," *IEEE J. Solid-State Circuits*, vol. 30, pp. 571–579, May 1995.
[5] K. Ogawa, "PASTEL: A parametrized memory characterization system," in *Proc. DATE'98*, Mar. 1998.
[6] M. Chinosi, R. Zafalon, and C. Guardiani, "Automatic characterization and modeling of power consumption in static RAM's," in *Int. Symp. Low-Power Electronics and Design*, Aug. 1998.
[7] P. E. Landman and J. M. Rabaey, "Architectural power analysis: The dual bit type method," *IEEE Trans. VLSI Syst.*, vol. 3, pp. 173–187, June 1995.
[8] "Epoch User's Manual," Cascade Design Automation Corp., Bellevue, WA, 1996.
[9] "Star-Sim User's Guide," Avant! Corp., Fremont, CA, 1997.
[10] "S-PLUS User's Manual," StatSci, Seattle, WA, 1993.
[11] J. M. Chambers and T. J. Hastie, *Statistical Models in S*. Pacific Grove, CA: Wadsworth and Brooks, 1992.
[12] J. Neter, W. Wasserman, and M. H. Kutner, *Applied Linear Regression Models*. Homewood, IL: Irwin, 1989.
[13] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design*. Reading, MA: Addison-Wesley, 1985.
[14] M. Berty, "Exploring low power memory design," Master's thesis, Carnegie-Mellon Univ., Pittsburgh, PA, May 1998.
[15] S. L. Coumeri and D. E. Thomas, "Memory modeling for system synthesis," in *Int. Symp. Low-Power Electronics and Design*, Aug. 1998.

Sari L. Coumeri received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, in 1993 and the M.S. and Ph.D. degrees in electrical and computer engineering from Carnegie-Mellon University, Pittsburgh, PA, in 1995 and 1999, respectively.

She is currently working on power analysis and estimation tools in the Alpha CAD group at Compaq Computer Corporation, Shrewsbury, MA. Her research interests include low-power design, behavioral synthesis, and hardware–software codesign.

Donald E. Thomas, Jr. (S'74–M'77–SM'86–F'89) received the Ph.D. degree in 1977 from Carnegie-Mellon University, Pittsburgh, PA.

He is currently Professor of Electrical and Computer Engineering at Carnegie-Mellon University, working in the area of behavior-based design of digital systems, including such areas as hardware/software codesign, architectural partitioning, and synthesis of low-power systems. From 1985 to 1986, he was a Visiting Scientist at the IBM T. J. Watson Research Center, Yorktown Heights, NY. He has been on the editorial board of the IEEE Design and Test Magazine and an Associate Editor for the IEEE Transactions on Very Large Scale Integration (VLSI) Systems. He is a coauthor of *The Verilog Hardware Description Language Fourth Edition*.

Dr. Thomas is a member of the IEEE Computer Society and the ACM. He was Chair of the 1989 Design Automation Conference. He was on the IEEE Computer Society Board of Governors and was Chair of the CODES/CASHE-96 Workshop on Hardware/Software Co-Design. He was elected Fellow of the IEEE "for contributions to automatic design of integrated circuits and systems, and to education in computer engineering."