Energy Consumption Modeling and Optimization for SRAM's

Robert J. Evans, *Member, IEEE,* and Paul D. Franzon, *Member, IEEE*

***Abstract*—The recent trends in portable computing technologies have established the need for energy efficient design strategies. To achieve minimum energy design goals, system designers need a technique to accurately model the energy consumption of their design alternatives without performing a full physical design and full-circuit simulation. This paper presents and compares five approaches for modeling the energy consumption of CMOS circuits. These five modeling approaches have been chosen to represent the various levels of model complexity and accuracy found in the current literature. These modeling approaches are applied to the energy consumption of SRAM's to provide examples of their use and to allow for the comparison of their modeling qualities. It was found that a mixed characterization model—using a CV^2 prediction for digital subsections and fitted simulation results for the analog subsections—is satisfactory (within ±1 process variation) for predicting the absolute energy consumed per cycle. This same model is also very good (within 2%) for predicting an optimum organization for the internal structures of the SRAM. Several common architectures and circuit designs for SRAM's are analyzed with these models. This analysis shows that global, rather than local improvements, produce the largest energy savings.**

I. Introduction

MANY of today's advanced CMOS designs are being applied to low power applications, such as battery powered notebook computers, remote telecommunications equipment, and aerospace applications. Designers of these CMOS systems need fast and reasonably accurate estimations of the energy consumption for their designs. There will usually be several alternatives from which to choose, and the designers will want to evaluate the tradeoffs in time delays and energy consumption for each alternative. This paper deals with the production of energy prediction models for SRAM circuits and their application to minimum energy optimizations.

Several people have presented modeling approaches aimed at estimating the absolute energy consumed by a design without performing a full circuit simulation. One approach to this modeling involves many complicated current calculations, and approaches the complexity of the circuit simulators. Routabi [1] presented such an approach to estimate current waveforms in CMOS circuits. This produced results that were within 10% of the values determined by a full SPICE [2] simulation, while reducing the estimation time by three to four orders of magnitude. Nabavi-Lishi presented a similar which produced a two orders of magnitude savings in estimation time with a similar accuracy [3]. Several researchers have presented other modeling approaches for estimating the energy consumption of CMOS circuits based on the switched internal capacitances of the circuit [4]–[8]. These methods provide a varied level of accuracy, but are much simpler to model than the circuit-simulation type of approaches. Hoppe [9], [10] and Ma [11] have presented modeling techniques to estimate the energy consumption of a circuit with the goal of selecting transistor sizes to optimize the circuit for the minimum energy consumption possible. Burch [12] has discussed the tradeoffs in the accuracy of the energy consumption modeling techniques with their respective estimation times. He also presents the importance of reaching a design decision without spending the time to perform the full circuit simulations.

Almost all of these existing methods of energy estimation are aimed at predicting the absolute amount of energy a given circuit will consume. This leaves open the question of what detail of modeling is required to select the optimum organization for minimized energy consumption from amongst several alternatives.

This paper presents and compares five approaches for modeling the energy consumption of CMOS circuits with the objective of determining the best approaches for absolute and comparative predictions. We also use the models to evaluate a typical set of SRAM architectures and circuit alternatives selected from current literature, with the aim of finding the minimum energy alternatives and the speed-power tradeoffs present in SRAM design. These design alternatives represent variations in the overall global architecture of the device, as with the Divided Word Line (DWL) [13] and Hierarchical Word Line Decoder (HWD) [14]–[16] designs. The effects of bussing or multiplexing the internal address lines amongst the subarrays are evaluated and compared, along with the effects of multibit width word sizes and sequential addressing (synchronous modes). The effects of modifying the circuit designs of the various internal structures are also presented in this work.

Section II of this paper presents the five approaches for modeling energy consumption discussed in this research. Section III presents a comparison of these approaches with respect to their absolute predictive qualities for energy consumption. Section IV evaluates these approaches for their optimum predictive qualities. Section V presents a discussion on the level of detail required for accurate energy modeling, and uses these models to predict the optimum internal organizations

Manuscript received January 11, 1994; revised November 11, 1994. This work was supported in part by the IBM Corporation.

R. J. Evans is with IBM PC Company, Research Triangle Park, NC 27709 USA.

P. D. Franzon is with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC 27607 USA.

IEEE Log Number 9410178.

for minimizing the SRAM energy consumption. The tradeoffs between optimizing a SRAM design for the minimum access time and energy consumption are presented. Section VI presents the conclusions.

II. Approaches to Energy Estimation

In this section we present the methodology which is used to obtain the five energy consumption models for our SRAM circuits. The five energy estimation models are as follows:

a) *Relational Sizing Model*, in which the energy consumption is estimated based on the relative lengths of the internal interconnects;

b) the *Analytical-Based Characterization Model*, in which the energy consumption estimates are based on theoretical calculations of the internal switched capacitance;

c) the *Simulation-Based Characterization Model*, in which the circuits are physically designed and the energy consumption is simulated using a standard design tool;

d) the *Mixed Characterization Model*, a hybrid approach in which the circuit models are mixed between the Analytical- and Simulation-based models, depending upon their analog or digital behavior;

e) the *Measurement-Based Characterization Model*, in which the circuits are fabricated and the actual energy consumption of the devices is measured during operation.

Next we explain the development of these models. As the entire models are too large for inclusion in this paper, we present only the development of one of these models for a typical SRAM subsection. Fig. 1 shows the organization and dependent variables for a typical SRAM architecture. Here, the memory is organized as an array of 2^p memory subarrays, where each subarray stores $2^m \times 2^n$ bits. The number of total address lines is $m + n + p$, with $p = p_0 + p_1 + \cdots$ for each layer of subblocks in the HWD architectures. All five models are applied to the MOSIS 2.0 micron N-Well [17] process to allow for a comparison of the predictive qualities of each model, described in Sections III and IV.

Each model was developed by dividing the full circuit of interest into functional subsections and individually modeling the energy consumption of each subsection. A *characteristic equation* describing the estimated energy consumed per a given input switching event (a 0-1 or 1-0 transition) is developed for each model and subsection. These characteristic energy equations are dependent on the sizes of the circuit subsections, as represented by the m, n, and p organizational variables. The subsections which contain the possibility of more than one unique switching event (e.g., address decoders) are characterized by averaging the estimated energies resulting from all of the equally probable input events. The energy prediction for the entire circuit is then obtained by summing the individual energy estimates from each subsection into a total energy value for the given series of switching events.

A. Relational Sizing Model

In the relational sizing model, only the major interconnect lines are considered. This model has been adapted from a similar model for RAM's presented by Mead and Conway [18]. This model is relative, and therefore can only be used to predict optimum organizations for delay and energy, and not for absolute energy estimations.

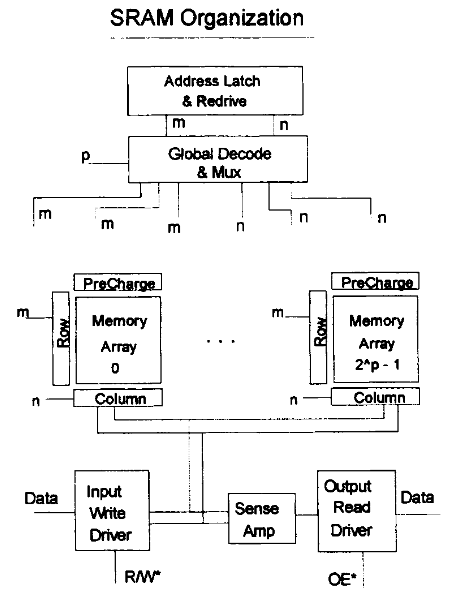

Fig. 1. An SRAM divided into typical subsections, showing the distributions of "m" row address lines, "n" column address lines, and "p" global address lines.

B. Analytical-Based Characterization Model

The energy consumed by a CMOS circuit for a 1-0-1 input switching event may be theoretically modeled by considering the switched internal capacitances and the resulting through currents, using the general formulas discussed by Greggain [19], Chandraksan [20], and Veendrick [21]

$$\text{Charge/Discharge Energy} = \text{Switched Capacitance} * \text{Voltage}^2, \quad (1)$$

and

$$\text{Through Current Energy per 1-0-1 Transition (Max)} = \frac{\beta}{12}(V_{dd} - V_t)^3 \cdot \tau, \quad (2)$$

where

β = transistor beta value;
V_{dd} = supply voltage;
V_t = transistor threshold voltage;
τ = time constant of the input waveform.

Equations (1) and (2) are used to obtain the energy per input switching event for all the internal nodes (stages) in each circuit subsection. For example, the switched capacitance for a typical SRAM Word Line node is

$$C_{\text{Wordline}} = C_{\text{DifDrv}} + 2^n x C_{\text{IntPoly2W}} + 2^n * 2C_{\text{gMin}}, \quad (3)$$

where

n = number of Column Address Lines;
2^n = number of cells per Row/Word Line;
x = horizontal size of a single memory cell;
C_{DifDrv} = Diffusion capacitance of a medium sized word line driver;
$C_{\mathrm{IntPoly2W}}$ = Interconnect capacitance for a Poly line 2λ wide;
C_{gMin} = Average gate capacitance (n or p) in a minimum transistor.

The absolute values for many of the model parameters depend on the fabrication technology. The technology-dependent values are substituted for all of the terms in these equations to obtain the characteristic equations for the subsection in the selected technology.

Certain simplifications must be made in order to reduce the effort required to build the model. The transistor device sizes are limited to three possibilities, minimum-sized, medium-sized, and large. All transistor resistances were assumed to be constantly linear, and all capacitances are assumed to be constant throughout all voltage swings. It was assumed that all nodes undergo full voltage swings, with the exception of the bitlines and internal I/O lines. It was assumed that one of the bitlines for each column would discharge to a value determined by the ratio of the linear resistances of the transistors turned on during the evaluate portion of the cycle, while the other bitline remained fully charged. The offchip capacitance load on the output driver is assumed to be 40 pF. For calculation of the through current energy using (2), the various time constants of the modeled subsections were estimated with a simple first-order RC analysis.

C. Simulation-Based Characterization Model

Selected circuit subsections were physically designed using the MAGIC [22] tool, extracted, and simulated in CAzM [23] using the technology files for the chosen 2.0 micron process. The power supply lines of each subsection were isolated to allow for separate measurement of individual supply currents using the CMOS power meter techniques described by Kang [24] and Yacoub [25]. Input patterns were applied to the circuits through the simulations to obtain a representative set of energy values for each input switching event. A linear regression was performed on these energy values, and the results were used to form the characteristic equations for the simulation-based models.

D. Mixed Characterization Model

The mixed characterization model is a combination of the analytical and simulation-based models. The subsections which contain purely digital behavior with full voltage swings are characterized with the analytical-based model. Those subsections with less than full voltage swings and analog-type behavior are modeled with the simulation-based model to obtain the required accuracy.

E. Measurement-Based Characterization Model

A number of circuit subsection alternatives were fabricated using the MOSIS 2.0 micron process. A selected set of the fabricated circuits is pictured in Fig. 2. The input patterns used in the circuit simulations were applied to the fabricated chips on a HP 82000 Circuit Tester. The supply current consumed by each subsection was measured using a series resistor. The size of this series resistor was chosen to produce a peak differential voltage of about 100 mV, larger than the background noise but small enough to not significantly affect the operation of the circuit being measured. A trapezoidal integration was performed on the measured current waveform over the time period of the chosen input pattern to obtain the energy consumed due to the given transition. The resulting measurements were used to derive the characteristic equations for this section.

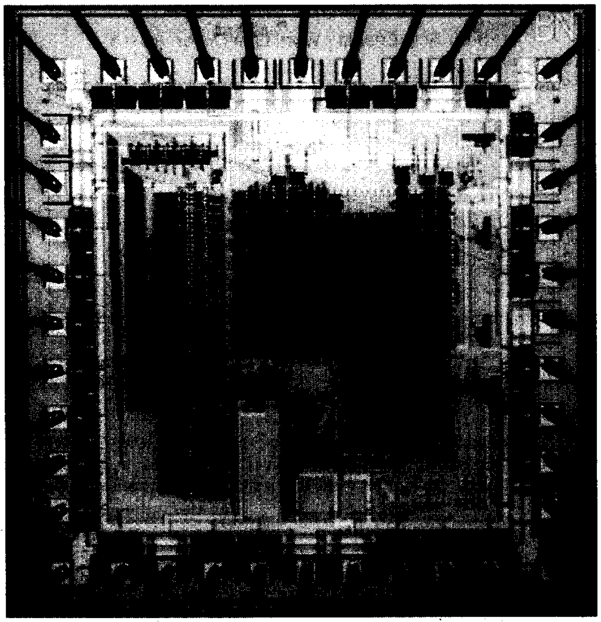

Fig. 2. A photo of one of the four SRAM devices fabricated for verification of the models.

It was possible to obtain satisfactory measurements for 80% of the fabricated circuit structures. If the supply currents were small or the rate of change of current was large, then the inductively induced power-ground line ringing and spiking noise was large, covering the signal of interest. In order to build a complete model for the measurement-based characterization approach, we used the simulation-based characterization equations for the structures for which we could not obtain satisfactory measurements.

In the next section we compare these modeling approaches for their absolute predictive qualities.

III. Comparison of the Models' Absolute Predictive Qualities

In this section we determine the absolute predictive quality of each model by comparing it against the measurement-based model. As the relational sizing model makes no absolute energy predictions, it is not discussed.

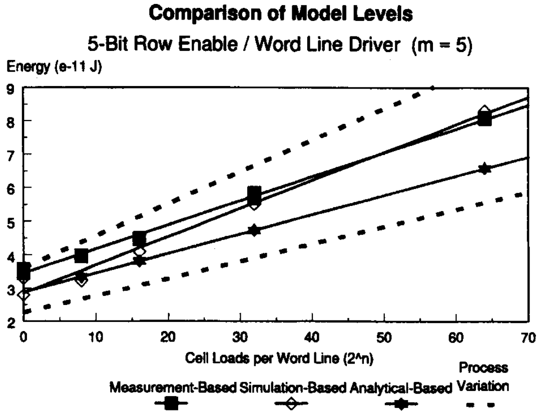

Fig. 3. Comparison of model levels for a 5-b ($m = 5$) Row Enable/Word Line Driver subsection. The energy shown is estimated for a 1-0 transition on the enable line, driving the word line active. This comparison is for the MOSIS 2.0 micron process.

A comparison of the energy estimations from each model for a five-bit ($m = 5$) Row Enable/Word Line Driver subsection is shown in Fig. 3. The values shown are the energy predictions for a 1-0 transition on the input, causing the word line to be driven high. The number of SRAM cells per word line shown on the x-axis is determined by the value of 2^n. The dotted lines on each side of the simulation energy line represent ± one process variation from the mean simulation-based model as obtained by resimulating the physical design using the four-corners technology files provided by MOSIS.

This comparison of models in Fig. 4 shows that the simulation-based energy estimations matched those obtained through the measurement-based characterizations very closely. This was found to be the case in all of the subsections where noise-free measurements were obtainable. In almost all of the subsections it was found that the analytical-based model underestimated the energy predicted by the simulation-based characterizations, with larger structures showing the greatest difference. The analytical-based characterization model is particularly inaccurate for those circuit elements that do not go through a full voltage swing, e.g., analog-type circuits. For the construction of the mixed model, the column lines, IO lines, and the output stage of the write drivers were characterized by their simulation results due to their analog behavior. The remainder of the subsections exhibited digital behavior and were modeled using the analytical-based characterizations.

The subsection energy prediction differences place the analytical-based estimation at the edge of the process variation boundaries. If the acceptance criteria for a model is that it is accurate to within one process variation, then the analytical-based characterization model is acceptable for absolute prediction purposes for full voltage swing (digital) circuits. To completely model the SRAM energy, the mixed model provides sufficiently accurate results.

Table I shows a comparison of the absolute energy predictions for a full memory device, between the analytical-based, mixed, and the simulation-based models. These results show that the analytical-based model underestimates the energy consumption of a full memory circuit by an average of 60%. The mixed model does much better at estimating the total energy consumption, and is within an average of 12% of the prediction of the simulation-based model.

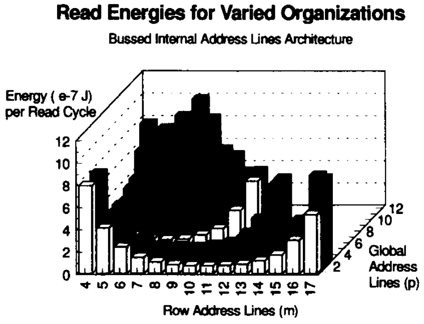

Fig. 4. Read Energy estimations for the bussed internal address line architecture, showing the minimum energy address line organization at $m = 9$, $n = 8$, $p = 5$. These values were estimated using the analytical-based characterization model, for a 4 Mb-×-1 device (22 total address lines), in the MOSIS 2.0 micron technology.

TABLE I
COMPARISON OF ABSOLUTE ENERGY PREDICTIONS OF ANALYTICAL-BASED AND MIXED MODELS FOR THE FULL ARCHITECTURAL ALTERNATIVES. THE ENERGY CONSUMPTIONS ARE COMPARED AT THE OPTIMUM POINT PREDICTED BY THE ANALYTICAL-BASED MODEL FOR THE MOSIS 2.0 MICRON PROCESS

Architecture & Size	Optimum Energy Organization - Analytical Model *m-n-p* (*p0-p1*)	Optimum Energy Analytical Model (Joules/cycle), Percentage from Sim-Based Energy	Optimum Energy Mixed Model (Joule/cycle), Percentage from Sim-Based Energy	Optimum Energy Sim-Based Model (Joules/cycle)
Single Block (4 kb)	7-5	1.35e-9 (62%)	3.45e-9 (3%)	3.55e-9
Bussed Subarrays	9-8-5	3.00e-8 (59%)	7.27e-8 (1%)	7.23e-8
Multiplexed Subarrays	9-4-9	3.28e-9 (70%)	7.48e-9 (31%)	1.09e-8
DWL	11-5-6	1.25e-8 (55%)	1.89e-8 (32%)	2.80e-8
HWD	10-5-7 (3-4)	1.33e-8 (43%)	2.023-8 (14%)	2.35e-8
DWL Row & Column	10-5-7	8.45e-9 (56%)	1.92e-8 (1%)	1.90e-8

In the next section we present how well these modeling approaches predict an optimum organization for the minimum energy consumption in SRAM's. Section V presents discussions on the use of these models to determine the optimum energy and timing organizations, and to analyze the effects of memory size, scaling, architectural design, and circuit design, on the energy consumption of the SRAM's.

IV. Comparison of the Optimal Predictive Qualities

In this section we present a comparison of how well each model predicts the optimal SRAM organization for minimizing energy consumption. The optimal SRAM organization predictions were obtained by coding the characteristic equations into a program and optimizing over the range of each architectural and circuit alternative. A simple, first-order, RC timing model is also included in this program to allow for evaluations of the tradeoffs between the energy consumptions and the delay times of each subsection. Only the memory read cycle is modeled.

For keeping these comparisons manageable, some assumptions are made about the overall memory architectures and access cycles. The number of address lines (m_Δ, n_Δ, and p_Δ) changing during each memory cycle are assumed to be one-

TABLE II
A COMPARISON OF THE RELATIONAL-SIZING, ANALYTICAL-BASED CHARACTERIZATION, AND MIXED MODEL PREDICTIONS OF THE OPTIMAL ORGANIZATION FOR VARIOUS SRAM ARCHITECTURES

Architecture	Optimum Relat-Sizing Organization & Percent Error	Optimum Analytical-Based Organization & Percent Error	Optimum Mixed-Model Organization & Percent Error	Optimum Simulation-Based Organization
Single 4kb	6/6 (68%)	7/5 (21%)	8/4 (0%)	8/4
Bussed	8/9/5 (65%)	9/8/5 (19%)	--	9/7/6
Muxed	10/3/9 (14%)	9/4/9 (36%)	11/2/9 (0%)	11/2/9
DWL	10/2/10 (86%)	11/5/6 (40%)	11/3/8 (0%)	11/3/8
HWD	--	10/5/7 (23%)	11/3/8 (0%)	11/3/8
DWL Both	9/4/9 (26%)	10/5/7 (22%)	10/3/9 (2%)	11/3/8

The percent error is defined as the difference in the energy predicted by the simulation-based model at each organization. These predictions were obtained for a 2.0 micron MOSIS technology. All memory sizes are assumed to be 4 Mb except where noted.

half of the number of their respective address lines, based on the assumption that each line has 50% probability of undergoing a switching event. A constant evaluate and sense time of 5.0 ns is chosen to simplify the estimations. The purpose in this study was not to evaluate the performance of SRAM sense amps, but to concentrate on the architectural effects on the overall energy consumption. Additional information on SRAM sense amps may be found in [26]–[28].

To evaluate their predictive power, the optimum organization predicted by each model was compared with the results of the simulation-based model. Fig. 4 shows a typical optimum obtained by varying the organizational variables m, n, and p. This particular plot represents the energy optimization for the bussed internal address line architecture, as obtained from the analytical-based model. From this figure, the organization with the minimum energy consumption can be seen to be at $m = 9$, $n = 8$, and $p = 5$. Table II shows a summary of these optimum organization predictions from the various models, for each of the architectural alternatives. The percentage difference in energy consumption between the optimum organization predicted by each model, and the optimum organization predicted by the simulation-based model, is shown for each architecture. This percentage difference in energy is calculated based on the predictions of the simulation-based model at the two organizations.

The relational-sizing model was found to be poor in its prediction of the optimum energy organization. The optimum organizations predicted by this model consume an average of 52% more power than the organizations predicted by the simulation-based model. The main problem in this relational-sizing technique is that it accounts only for the lengths of the main interconnects passing through the memory device.

The analytical-based characterization model shows optimum predictions closer to those of the simulation-based model, as the average error in the energy consumptions between the two organizations is 25%. This error is still too large to consider basing design decisions for mixed-behavior (digital and analog) circuits on this model. However, the analytical-based model is useful in narrowing the choices of design alternatives to be evaluated by the more accurate and involved modeling techniques, even with mixed-behavior circuits such as this memory design.

TABLE III
THE CONTRIBUTIONS OF EACH ANALYTICAL-BASED CHARACTERIZATION MODEL ELEMENT TOWARD THE TOTAL ABSOLUTE ENERGY CONSUMPTION PREDICTION

Technology Element	Percentage Contribution		
	2.0 micron	1.2 micron	0.8 micron
Minimum-Sized Gate Capacitances	19 %	14 %	15 %
Non-Minimum-Sized Gate Capacitances	13	11	13
Minimum-Sized Diffusion Capacitances	16	12	14
Non-Minimum-Sized Diffusion Caps.	1	2	1
All Diffusion Caps. - Averaged Value	28	20	22
Interconnect Crossovers Metal2-Metal1	6	9	7
Interconnect - Poly Runs Inside Gates	7	7	6
Interconnect - Decoder Stubs, < 10% of Total Interconnect Lengths	0.5	0.5	0.5
Interconnect Caps.- Long Runs Between Gates	33	43	41
Interconnect - Between Gates, Averaged Value for all Types	56	74	65
Through Current - Minimum-Sized Gates	0.5	1	1
Through Current - Non-Minimum Gates	2	2.5	6

The averaged values for the interconnect and diffusion elements represent the energy contribution when all cases of that particular element are replaced by the single averaged value; e.g., all interconnect capacitances (M1, M2, Poly) are set equal to the average interconnect capacitance value per λ. These results were obtained using the modified analytical-based characterization model.

The mixed model produces the optimum organization predictions which most closely match those of the simulation-based model. The optimum predictions of the mixed model match exactly on several of the architectural alternatives. In the DWL-Row/Column architecture, the difference in energy consumption between the two predicted organizations is only 2%. This model is much more accurate than the analytical-based model, and is easier to develop than the full simulation-based model.

The next section presents discussions on the use of these models to analyze the effects of memory size, scaling, architectural design, and circuit design, on the energy consumption of the SRAM's.

V. DISCUSSION

This section presents discussions on the inner details of the models, and uses them to analyze the effects of memory size, scaling, architectural design, and circuit design, on the energy consumption of the SRAM's. The tradeoffs between optimizing the organization for minimized energy consumption and access time are also discussed.

A. Energy Contributions of the Analytical-Based Model Elements

As described in Section II, the analytical-based model is built from the many individual capacitance and through-current elements contained in the modeled circuit. The contributions of these elements toward the overall absolute and optimum predictions of the model was determined by modifying the modeling programs to allow selective recombination of these elements. Table III shows the relative contributions of these elements toward the absolute energy prediction for a 4-Mb DWL architecture memory. The interconnect (between gates) and gate capacitances are shown to be the two highest

TABLE IV
CONTRIBUTION OF SRAM SUBSECTIONS TOWARD THE TOTAL ENERGY

Subsection	Energy Contribution (Joules/cycle)	Percentage of Total Energy
ATD Address Data Latch	1.62 e-9	21 %
Row Decoder	5.73 e-10	7 %
Row Enable / Word Line Drivers	6.82 e-10	9 %
Column Decode & Enable	6.55 e-11	1 %
Global Decode & Enable	2.44 e-9	31 %
Precharge	2.34 e-9	30 %
Sense Amp	2.52 e-12	--
Output Driver	3.85 e-11	1 %

This data was estimated by the analytical-based characterization model for a DWL Row/Column architecture, with a 10/5/7 organization, for a 4 Mb device in the MOSIS 2.0 micron process.

TABLE V
COMPARISON OF THE ENERGY AND TIME OPTIMIZED GEOMETRIES FOR SELECTED ARCHITECTURES

Architecture & Size	Optimum Energy Organization *m-n-p* (*p0-p1*)	Optimum Timing Organization *m-n-p* (*p0-p1*)	% Energy Savings Available	% Access Time Penalty for Achieving this Savings
Single Block (4 kb)	9-3	6-6	44 %	84 %
Single Block (4 Mb)	13-9	11-11	6 %	159 %
Bussed Subarrays	9-7-6	9-8-5	20 %	2 %
Multiplexed Subarrays	11-2-9	9-8-5	87 %	107 %
DWL	11-3-8	9-8-5	73 %	96 %
HWD	11-3-8 (4-4)	9-7-6 (3-3)	55 %	130 %
DWL Row & Column	11-3-8	9-8-5	74 %	96 %

There are no limits placed on the access time range for this comparison. All values have been obtained using the simulation-based characterization model, with the MOSIS 2.0 micron process. All values represent 4 Mb devices sizes except where noted.

TABLE VI
A COMPARISON OF EFFECTS OF MEMORY DENSITY ON THE OPTIMUM ORGANIZATION AND ENERGY CONSUMPTION

SRAM Density	Optimum Organization	Energy at Optimum Org.
4 Mbit	10/5/7	3.48 e-9 J
16 Mbit	11/5/8	6.58 e-9 J
64 Mbit	12/5/9	1.29 e-8 J
256 Mbit	13/5/10	2.54 e-8 J

The values shown were obtained with the analytical-based characterization model, based on a 0.8 micron technology, and represent the DWL-Row/Column architecture. The energy values represent the total energy consumed by one read access from the memory device.

contributors to this energy prediction, combining to account for 55% to 69% of the energy consumed per access. The through currents, the short interconnect stubs of the decoders, and the diffusion capacitances of nonminimum-sized gates have the least effect on the total energy prediction, as together they account for only 5% of the total energy. The designer can use this information to decide which elements are required to be modeled to meet any given accuracy requirements, and where to direct the efforts in technology-based energy reductions. This table also shows that these relative contribution values do not significantly vary when the technology is scaled to smaller sizes.

Table IV shows the internal breakdown of the energy estimations based on the SRAM subsections. This table shows that the majority of the energy is consumed by the Global Decode and Enable, the Precharge, and the Internal Address Latch/Distribution subsections. This suggests that these subsections will have the dominant effects on the energy consumption and optimizations of the memory devices. The remainder of the subsections consume only a total of 18% of the energy.

B. Optimum SRAM Organizations for Minimizing Energy Consumption

The optimum organization for minimizing the energy consumption of the SRAM architectures is nonsquare, containing more rows than columns. This organizational weighting is evident in Table V, as all of the architectures show a higher value of the organizational variable m than of n. This nonsquare organization is mainly due to the weighting of the precharge subsections of the memory. The precharge sections are more energy costly in the n direction due to the load of the gate capacitance of the precharge driver transistors associated with each of the 2^n columns. The through-current of the drivers also slightly contributes to this n-direction weighting. For a constant number of cells in a subarray ($m + n =$ constant), the energy associated with precharging the column lines will be constant, and does not weigh the organization toward either extreme of m or n.

We are also interested in the speed-energy tradeoffs. Table V shows that the optimum organization for minimized timing generally has fewer, more square subarrays than does the energy optimum. In the timing estimations the column precharge times will be proportional to only the column lengths, and not the widths. For example, when precharging an $m \times n$ subarray, all 2^n column lines (bitlines), each 2^m cells long, must be precharged. The precharge operation charges all of the 2^n column lines at the same time. Therefore the speed of this precharging of the column lines is only proportional to the length of one column line, or 2^m cells, and will drive the optimum organization toward smaller values of m.

It is assumed that the designer is willing to consider sacrificing some percentage of the device latency time in exchange for an energy savings. A comparison of these unbounded energy-timing tradeoffs for selected architectures is shown in Table V, as estimated with the simulation-based characterization model and the first-order RC timing estimator. These timing-optimized organizations closely match those described in several published SRAM designs [14], [16], [29]–[31]. If the designer is willing to accept a large access time penalty, then very significant energy savings can be obtained. By choosing the optimum energy organization (11/3/8) over the optimum timing organization (9/8/5), the access time would increase by 96% (almost double) to gain a 73% savings in energy. In general, the more complex the memory architecture, the greater difference there is between the energy and timing optimum points.

The designer will not always be willing to sacrifice such large access time penalties to achieve an energy savings. If the designer puts a limit on the range of the access time penalty, a substantial energy savings can still be obtained. For example,

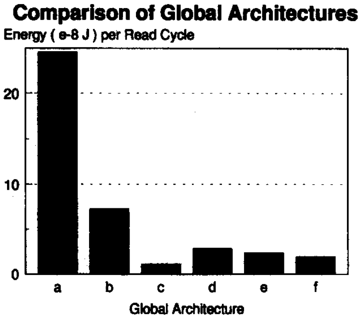

Fig. 5. Comparison of the minimum energy estimations for selected architectures—(a) Single Memory Block, (b) Subarrays with Bussed Address Lines, (c) Subarrays with Multiplexed Address Lines, (d) Divided Word Line Architecture (DWL), (e) Hierarchical Word Decoder Architecture (HWD), and (f) DWL on both Row and Column Decoders. These estimations were performed using the simulation-based characterization model on a 4 Mb-×-1 device, in the MOSIS 2.0 micron process.

with a 20% bounds on the access time, an average of 39% in energy savings can still be obtained. This type of energy and timing tradeoff can be calculated for any range of bounds that might be placed on the access time or energy of the memory device.

The effects on the optimum energy organizations of increasing the total memory size are shown in Table VI. As the size increase, the values of m (rows) and p (subarrays) grow equally, while the number of columns n stays basically the same. The energy cost of adding columns is greater than that of the rows and subarrays, and therefore the number of columns does not increase with the size. This sensitivity to the n (columns) variable is attributed to the energy requirements of the precharge subsections.

C. Comparison of Architectural and Circuit Design Alternatives on SRAM Energy Consumption

The various architectural and circuit alternatives were modeled and the results compared to determine the optimum design features for the minimization of energy consumption. A comparison of the minimum energy consumptions of these architectural alternatives is shown in Fig. 5, as determined with the simulation-based model for unbounded access times. This figure shows the large energy savings of the global architecture changes associated with using subarray-based designs over that of the single array of memory cells design. For a 4-Mb device in our 2.0 micron technology, the best global architecture is the multiplexed internal address line architecture. The optimum organization for minimizing the energy consumption in this architecture is made up of 512 subarrays, each containing 2048 rows and 4 columns ($m = 11$, $n = 2$, $p = 9$). In this architecture the optimum organization is heavily weighted toward a large number of subarrays. With the multiplexed subarrays architecture, the energy consumed by redriving the internal address lines to the many subarrays is much less than in the bussed internal address line architecture. Therefore the energy cost per additional subarray is lower while the cost per row and column remains the same, resulting in a higher optimum number of subarrays.

The optimization of the SRAM designs for minimizing the energy consumptions parallels much of the published work in optimizing the designs for the minimum access times. Most of the multiple subarray architectures analyzed here were originally created to improve the access times of the devices. It is expected that the reductions in switched capacitance obtained by these designs to reduce the access times will also tend to reduce the energy consumed by the devices. This is mostly true up to a point. The Hierarchical Word Line Decoding (HWD) architectures have several additional intermediate stages that allow for a faster overall address decode, but that consume additional energy. A significant energy savings using these architectures will only be realized on very large memories. If the access time is limited to be within 20% of its optimum value, the optimum energies of each architecture will be slightly different. The multiplexed internal bus line architecture is still the best for minimizing the energy consumption of the device. However, the 20% bounds on the access times allows only a $m/n/p = 9/6/7$ organization for this architecture, but still produces a 66% energy savings over the time optimized organization.

It would be expected that multibit word memory architectures would be more energy efficient per each bit read, as a single address specifies several bits. The overhead energy of the row, column, and global decoders/enables for each additional bit read is saved during each cycle. Several word sizes were compared for the energy consumption per bit acquired. The energy per bit read is found to decrease rapidly as the word size increases. In this analysis, the use of a 64-b wide memory device saves 85% of the energy of using 64 1-b wide devices. This suggests the use of "wider" SRAM's, placing more bits of the same address location on one physical device. A similar comparison of the energy consumptions for a sequentially-addressed memory was also performed. This comparison shows that the energy savings in the SRAM due to this sequential-access design are minimal. The reasons for this are straightforward. Although the external address lines do not change for each additional word accessed, there is still considerable switching activity inside the memory devices. However, sequential addressing will produce up to a 50% energy savings at the higher level of the system design, as the CPU section does not have to repeatedly drive the system-level interconnect and input loads of each memory chip for each data word in the desired memory block.

In general, the circuit design alternatives analyzed produce a local, rather than global, improvement in the energy consumption of the memory. These local improvements produce an incremental, and often less dramatic, improvement in energy consumption amongst the various subarray-based architectures. For example, internal changes in the row and column decoder designs produce an approximately 38% savings in the energy consumed by the decoder subsections, but the overall savings is less than 3% of the total device energy consumption. The details of the energy consumption and speed-power tradeoff analyses on several circuit-level design alternatives are found in [32].

The next section presents a summary of the results and conclusions.

VI. Summary and Conclusion

Five approaches to predicting the power consumption of different SRAM structures were developed and compared. It was found that the mixed characterization model—using a CV^2 prediction for digital subsections and fitted simulation results for the analog subsections—was satisfactory (within ±1 process variation) for predicting the absolute energy consumed per cycle and very good (within 2%) for predicting the optimum organization.

The absolute energy and optimum energy organization estimations of the analytical-based model are built on the contributions of several internal elements. The between-gate interconnect capacitances, gate capacitances, and minimum-sized diffusion capacitances account for approximately 80% of the absolute energy consumption.

The optimum organizations for minimized energy consumption in our SRAM designs are found to be nonsquare, containing more rows than columns, and becoming less square with increasing size. The optimum organizations for minimizing the memory access time are more square (more equal number of rows and columns) than those for minimizing the energy consumption. With some compromise in access time, the memory can be reorganized to provide a significant energy savings.

An analysis of several common architectures and circuit designs for SRAM's shows that global, rather than local circuit improvements, produce the largest savings in energy consumptions. The best architecture determined for our process is a multiplexed internal address line design, with an organization of 512 subarrays, each containing 2048 rows and 4 columns. Multibit word organizations produce a savings in the energy per bit read, suggesting that the energy consumption can be further optimized by placing as many bits per word accessed in the same physical memory device. Sequential-address memories do not themselves provide a significant energy savings over singly-addressed devices, although their use can save up to 50% of the memory access energy at the system level.

Acknowledgment

The authors wish to thank the Microelectronic Center of North Carolina (MCNC) and the IBM Corporation for the use of their test equipment.

References

[1] F. Rouatbi, B. Haroun, and A. J. Al-Khalili, "Power estimation tool for sub-micron CMOS VLSI circuits," in *Proc. IEEE ICCAD Conf.*, Nov. 1992, pp. 204–209.

[2] L. W. Nagel, "SPICE2: A computer program to simulate semiconductor circuits," Electronics Research Labs., University of California, Berkeley, CA, ERL Memo. ERL-520, May 1975.

[3] A. Nabavi-Lishi and N. Rumin, "Delay and bus current evaluation in CMOS logic circuits," in *Proc. IEEE ICCAD Conf.*, Nov. 1992, pp. 198–203.

[4] L. Greggain and B. White, "Predicting and scaling power consumption in CMOS ASIC's," in *Proc. Second Annu. IEEE ASIC Seminar and Exhibit*, IEEE Catalog #89TH0280-8, pp. 8-6.1–8-6.4, 1989.

[5] D. Liu and C. Svensson, "Trading speed for low power by choice of supply and threshold voltages," *IEEE J. Solid-State Circuits*, vol. 28, no. 1, pp. 10–17, Jan. 1993.

[6] P. Vanoostende, P. Six, J. Vandewalle, and H. J. DeMan, "Estimation of typical power of synchronous CMOS circuits using a hierarchy of simulators," *IEEE J. Solid-State Circuits*, vol. 28, no. 1, pp. 26–39, Jan. 1993.

[7] S. Devadas, "Estimation of power dissipation in CMOS combinational circuits using Boolean function mapping," *IEEE Trans. Computer-Aided Design*, vol. 11, no. 3, pp. 373–383, Mar. 1992.

[8] A. Tyagi, "Energy consumption in multielective and boundary VLSI computations," *IEEE J. Solid-State Circuits*, vol. 26, no. 9, pp. 1240–1248, Sept. 1991.

[9] B. Hoppe *et al.*, "Optimization of high-speed CMOS logic circuits with analytical models for signal delay, chip area, and dynamic power dissipation," *IEEE Trans. Computer-Aided Design*, vol. 22, no. 3, Mar. 1990.

[10] B. Hoppe, G. Neuendorf, and D. Schmitt-Lansiedel, "Automatic transistor sizing in high performance CMOS logic circuits," *Proc. IEEE VLSI Comput. Peripherals*, IEEE Cat. no. 89CH2704-5, pp. 5/25–27, Hamburg, W. Germany, May 1989.

[11] S. Ma and P. Franzon, "Energy control and accurate delay estimation in the design of CMOS buffers," *IEEE J. Solid-State Devices*, vol. 29, no. 9, Sept. 1994.

[12] R. Burch, F. Najm, P. Yang, and T. Trick, "McPower: A Monte Carlo approach to power estimation," in *Proc. IEEE ICCAD Conf.*, Nov. 1992, pp. 90–96.

[13] M. Yoshimoto *et al.*, "A divided word-line structure in the static RAM and its application to a 64 k full CMOS RAM," *IEEE J. Solid-State Circuits*, Vol. SC-18, no. 5, pp. 479–485, Oct. 1983.

[14] T. Hirose, H. Kuriyama, S. Murakami, K. Yuzuriha, T. Mukai, K. Tsutsumi, Y. Nishimura, Y. Kohno, and K. Anami, "A 20-ns 4-Mb CMOS SRAM with hierarchical word decoding architecture," *IEEE J. Solid-State Circuits*, vol. 25, no. 5, pp. 1068–1074, Oct. 1990.

[15] H. Goto, H. Ohkubo, K. Kondou, M. Ohkawa, H. Mitani, S. Horiba, M. Soeda, F. Hayashi, Y. Hachiya, T. Shimizu, M. Ando, and Z. Matsuda, "A 3.3-V 12-ns 16-Mb CMOS SRAM," *IEEE J. Solid-State Circuits*, vol. 27, no. 11, pp. 1490–1496, Nov. 1992.

[16] S. Murakami *et al.*, "A 21 mW 4 Mb CMOS SRAM for battery operation," in *Proc. IEEE Int. Solid-State Circuits Conf.*, Feb. 1991, pp. 46–47.

[17] *MOSIS Implementation System User's Manual*, Release 3.0, Information Sciences Instititute, Univ. of Southern California—Marina del Rey, CA, 1990.

[18] C. Mead and L. Conway, *Introduction to VLSI Systems*, 2nd ed. Reading, MA: Addison-Wesley, 1980.

[19] L. Greggain and B. White, "Predicting and scaling power consumption in CMOS ASIC's," *Proc. Second Annu. IEEE ASIC Seminar and Exhibit*, IEEE Catalog #89TH0280-8, pp. 8/6.1–4, 1989.

[20] A. Chandrakasan, S. Sheng, and R. W. Broderson, "Low-power CMOS digital design," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 473–484, Apr. 1992.

[21] H. J. M. Veendrick, "Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits," *IEEE J. Solid-State Circuits*, vol. SC-19, no. 4, pp. 468–473, Aug. 1984.

[22] R. Hamachi *et al.*, *MAGIC User's Manual*, Berkeley CAD Tools, Univ. of California at Berkeley, 1986.

[23] D. Erdman *et al.*, *CAZM—Circuit Analyzer with Macromodeling, User's Manual*, Release 4.1, Microelectronics Center of North Carolina and Duke University, June 1990.

[24] S. M. Kang, "Accurate simulation of power dissipation in VLSI circuits," *IEEE J. Solid-State Circuits*, vol. SC-21, no. 5, pp. 889–891, Oct. 1986.

[25] G. Yacoub and W. Ku, "An enhanced technique for simulating short-circuit power dissipation," *IEEE J. Solid-State Circuits*, vol. 24, no. 3, pp. 844–847, June 1989.

[26] T. N. Blalock and R. Jaeger, "A high-speed clamped bit-line current-mode sense amplifier," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 542–548, Apr. 1991.

[27] H. Nambu *et al.*, "High-speed sensing techniques for ultrahigh-speed SRAM's," *IEEE J. Solid-State Circuits*, vol. 27, no. 4, pp. 632–640, Apr. 1992.

[28] E. Seevinck, P. van Beers, and H. Ontrop, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 525–535, Apr. 1991.

[29] S. Aizaki *et al.*, "A 15-ns 4-Mb CMOS SRAM," *IEEE J. Solid-State Circuits*, vol. 25, no. 5, pp. 1063–1067, Oct. 1990.

[30] T. Ootani *et al.*, "A 4-Mb CMOS SRAM with a PMOS thin-film transistor load cell," *IEEE J. Solid-State Circuits*, vol. 25, no. 5, pp. 1082–1092, Oct. 1990.

[31] K. Sasaki *et al.*, "A 23-ns 4-Mb CMOS SRAM with 0.2-μA standby current," *IEEE J. Solid-State Circuits*, vol. 25, no. 5, pp. 1075–1081, Oct. 1990.

[32] R. J. Evans, "Energy consumption modeling and optimization for SRAM's, Ph.D. dissertation, Dept. of Electrical and Computer Engineering, North Carolina State Univ., Raleigh, NC, July 1993.

Robert J. Evans (S'77–M'82–S'89–M'93) received the B.Eng. degree from Vanderbilt University, Nashville, TN, in 1980, the M.Eng. degree from The University of Virginia, Charlottesville, VA, in 1982, and the Ph.D. degree in Computer Engineering from North Carolina State University, Raleigh, in 1993.

He has worked as a medical instrumentation engineer for The University of Virginia Medical Center from 1982–1985. In 1985 he joined IBM at Charlotte, NC, moving in 1987 to IBM Research Triangle Park, NC. He is currently a member of the Mobile Products Development Group at IBM-RTP, working on the development of IBM ThinkPad computers. His research interests are in the area of low-energy computing and advanced electronic packaging.

Dr. Evans is a member IEPS, NSPE, and is a registered Professional Engineer.

Paul D. Franzon (S'86–M'89) received the Ph.D. degree in electrical engineering from the University of Adelaide, Adelaide, Australia, in 1989.

Before completing his Ph.D. he had worked at AT&T Bell Laboratories in Holmdel, NJ, in 1986 and 1987, and with the Australian Defence Science and Technology Organization. In these positions he worked on problems in wafer scale integration, IC yield modeling, and VLSI design. He is currently an Associate Professor at North Carolina State University. His current research interests include the design sciences for high-speed packaging and interconnect, particularly signal integrity management and optimized MCM system design. Other interests include low-power electronics and applications of MEMS. His teaching interests focus on microelectronic system building, including design for signal integrity, packaging system optimization, circuit design, and processor design. He has written and edited a book on multichip modules.

Dr. Franzon is a member of ISHM. He is on the steering committees for the IEEE MCM Conference and IEEE Topical Meeting on Electrical Performance of Electronic Packaging. In 1993, he received a National Science Foundation Young Investigators Award.