Energy Recovery for the Design of High-Speed, Low-Power Static RAMs

Nestoras Tzartzanis and William C. Athas

{nestoras, athas}@isi.edu

URL: http://www.isi.edu/acmos

University of Southern California – Information Sciences Institute

4676 Admiralty Way, Marina del Rey, California 90292-6695

Abstract

We present a low-power SRAM design based on the theory of energy recovery that reduces the dissipation associated with write operations while operating at high speed. The energy-recovery SRAM was evaluated through SPICE simulations and compared with a standard design. Simulation results of a 256×256 memory configuration indicate that, for successive write operations, energy savings for the different SRAM functions vary from 59% to 76% at 200 MHz operating frequency compared to the conventional design.

1. Introduction

As the density and the operating speed of CMOS VLSI chips increases, power dissipation becomes of utmost importance since it imposes restrictions on chip design and performance. In conventional CMOS circuits, signal energy dissipation is proportional to the signal capacitance. Static RAMs are an important dissipation source in many applications because they contain high-capacitance buses and they are frequently accessed. The major internal capacitive loads of an SRAM array consist of the address lines, the word lines, and the bit lines.

In this paper, we present an SRAM design where power reduction is a result of recovering signal energies [AtTz95] rather than reducing them. We employ a conventional RAM array together with energy-recovery latches/drivers for all major internal loads. At the system level, a resonant driver generates clock signals that synchronize and pipeline the SRAM operations and power the internal capacitance loads through the energy-recovery latches/drivers. The benefits to this approach are threefold: first, high performance is achieved, since data and control signals can have large voltage swings (e.g., from 0 to 5 V); second, dissipation can be reduced for *all* internal high-capacitance buses in the memory; and third, the control signals are gated with the clock pulses to simplify synchronization and speed up the SRAM operations.

An energy-recovery SRAM has been previously proposed [SoYR95]. Although energy savings as high as 85% at 1-2 MHz were reported, the proposed SRAM required a sequence of irregular signals that would likely be difficult to efficiently generate with an energy-recovery driver. In contrast, the SRAM proposed in this paper can operate with the well-known two-phase non-overlapping clocking scheme. Furthermore, a resonant clock driver has been developed and demonstrated for powering circuits with the same design style [AtST96]. HSPICE simulations indicate that the proposed SRAM is more than 100 times faster than the previous design with comparable energy savings.

In conventional SRAM designs, reducing the voltage swing can effectively decrease the dissipation associated with the bit lines, but *only* for read operations. This technique limits the bit-line swing to a small fraction (e.g., 10%) of the supply voltage which is sufficient for the sense amplifiers to produce a full voltage-swing output [AmHo94]. However, this technique cannot be used for *all* buses (address and word lines, and bit lines for write operations) and control signals without speed losses because reducing the voltage swings results in longer access time. In contrast, in the design presented here, energies of all internal buses are recovered without speed loss.

A technique has been proposed where the bit-line voltage swing is significantly reduced for write operations as well [AlAn95]. By using a memory cell with different transistor sizes than the conventional one, writes can be performed by pulling one of the bit lines to 1 V while the other is 0 V and the word line is 5 V. Although the dissipation associated with write operations is reduced, this SRAM design exhibits degraded noise margins compared to the conventional one. Also, the address and word lines are full voltage-swing signals.

The energy-recovery SRAM leverages low voltage swing in the bit lines to reduce dissipation for read opera-

tions. In this paper, we focus only on the write operations which are the novel aspect of the design. Our approach began with a conventional pipelined SRAM organization in which the conventional latches/drivers are replaced with energy-recovery ones (Section 2). We simulated in HSPICE extracted portions of the two SRAMs (i.e., row decoder, one row, one column, and the latches) for a 0.8 µm bulk CMOS technology. The simulations showed that energy recovery resulted in significant energy savings (e.g., 59% to 76%) for the different SRAM parts at 200 MHz (Section 3). The energy recovery and the standard SRAM designs cannot be directly compared because they have different switching behavior. Therefore, we developed models for the total dissipation of successive write operations that took into consideration the different switching behavior of the two SRAMs (Section 4). Finally, we investigated some practical considerations related to the resonant clock driver (Section 5).

2. Energy-Recovery SRAM

The energy-recovery SRAM is based on the well-known organization (Fig. 1a). No extra circuitry is necessary. It includes the address latches, the row decoder, the word-line latches, the data input latches, the memory array, the sense amplifiers, and the data output latches. The RAM array is made from the "classic" 6-transistor memory cell (Fig. 1b).

The row decoder consists of precharged NAND gates. Precharged NOR gates would have been faster and would not suffer from charge redistribution, but they are worse in terms of dissipation. In precharged NOR gates, transistors are connected in parallel and all gates but one discharge for each memory access. Furthermore, the discharged node of a NOR gate has more parasitic capacitance than the one of a NAND gate since the transistors are connected in parallel.

The proposed SRAM employs energy-recovery techniques for low-power dissipation. All the internal bus and control signals are clock-powered. A simple resonant driver can generate the clock signals [AtST96]. Circuit-wise, the main difference between the energy-recovery SRAM and the conventional one is the latch/driver (Fig. 2a). The energy-recovery latch/driver operates as a filter that either passes the clock pulses to the output or clamps the output to ground depending on the stored datum.

The operation of the energy-recovery latch/driver relies on the bootstrapping effect [GlDo85]. The signals φ_1 and φ_2 are two-phase non-overlapping clock phases that swing from 0 to V_{ph}. V_{ddH} is a dc supply at voltage V_{ph} that is connected only to pass transistor gates and dissipates no power. The two inverters are powered from a lower voltage dc supply, i.e., $V_{ph}-V_{th}$, where V_{th} is the nFET threshold voltage. During φ_1, D_{in} is stored on the gate capacitance of M_3 (i.e., node D_{isol}). If D_{in} is low, then M_4 clamps D_{out} to ground. If D_{in} is high (Fig. 2b), then D_{isol} is charged to $V_{ph}-V_{th}$. When the positive edge of φ_2 occurs, the voltage of the isolation node
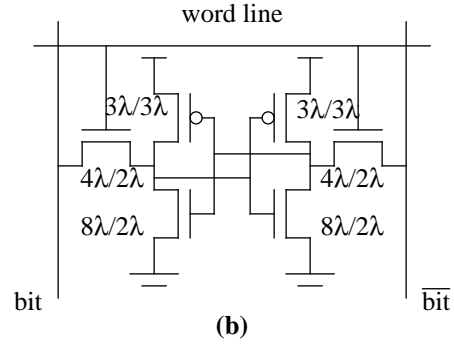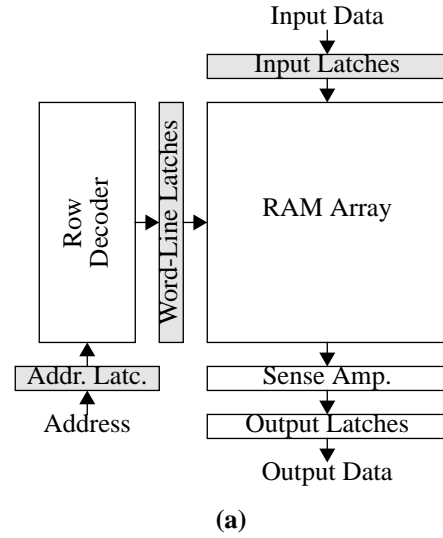

(a)

(b)

Figure 1: Energy-recovery SRAM organization (a) and cell (b).

D_{isol} is bootstrapped to well above V_{ph} due to the gate-to-channel capacitance of M_3, and the output is charged to V_{ph}. Therefore, although the input signal, D_{in}, need not swing to more than $V_{ph}-V_{th}$, a full voltage-swing output D_{out} is generated.

The dc supply V_{ddH} is introduced so that the transistor M_3 is always actively driven. Phase φ_1 could be used instead of V_{ddH} to drive the transistor M_2. If this was the case, when φ_2 occurred and D_{isol} was at 0 V, the voltage of D_{isol} would bootstrap to above 0 V and there would be a short-circuit current drawn from φ_2 to ground through the transistors M_3 and M_4.

The energy-recovery latch/driver is small in area. M_2 has to be small to minimize the parasitic capacitance of node D_{isol}. M_4 can be small since it only clamps D_{out} to ground to avoid coupling of φ_2 to D_{out} when D_{isol} is 0 V, but it does *not* discharge the load capacitance. On the other hand, the size of the device M_3 is critical. There are two criteria for sizing M_3. First, the ratio of the gate capacitance of M_3 to the parasitic capacitance of the node D_{isol} should be large enough to allow the voltage of D_{isol} to bootstrap to $V_{ph}+V_{th}$. This criterion applies for small capacitance loads and/or slow systems. Second, the transistor M_3 should be large enough to meet the
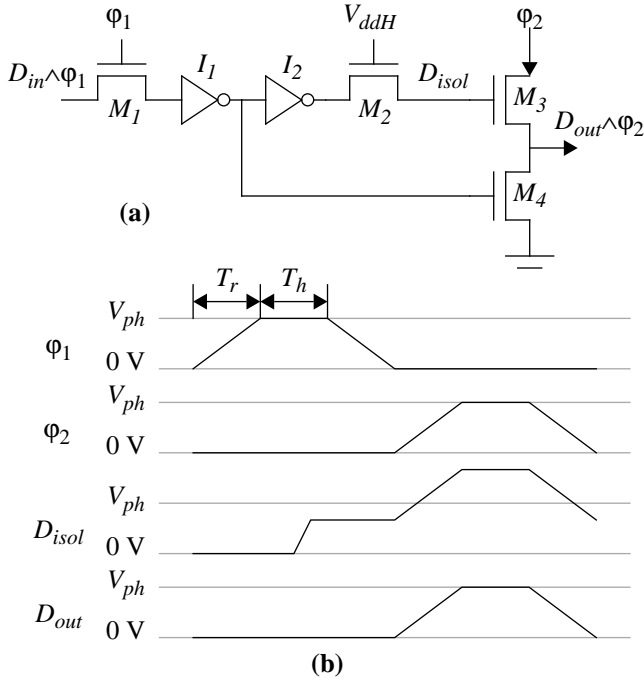
(a)

(b)

Figure 2: Energy-recovery latch/driver (a) and timing diagram when D_{in} is high (b).

system frequency specification. The second criterion is important for a high-speed SRAM with high-capacitance buses.

To provide guidelines for sizing the transistor M_3, we make some assumptions regarding the phases and the operation of M_3. First, the phases are assumed to be variable voltage sources that are ramped from 0 to V_{ph} during time T_r and remain at V_{ph} for time T_h (Fig. 2b). Second, M_3 is modeled as a resistance R. If C is the load capacitance, then the output voltage V_{out} at time t ($T_r \le t \le T_r + T_h$) is given by:

$$V_{out}(t) = V_{ph}\left(1 - \frac{RC}{T_r}\left(1 - e^{-\frac{T_r}{RC}}\right)e^{-\frac{t - T_r}{RC}}\right) \quad (1)$$

Using the above equation, the resistance of M_3 can be estimated for a desired output voltage V_{out} at time t. Assuming that M_3 operates in the linear region, its on-resistance, R, can be modeled as [AtTz95]:

$$R = \frac{L^2}{\mu C_g(\alpha V_{ph} - 2V_{th})} \quad ; \quad C_g = L \cdot W \cdot C_{ox} \quad (2)$$

where C_g, C_{ox}, L, W, and μ denote the gate capacitance of M_3, the oxide capacitance per unit area, the channel length, the channel width, and the carrier mobility, respectively. The parameter α models the reduction of the bootstrapping effect due to the parasitic capacitance. Therefore, for a certain technology and for a given on-resistance R, the size of M_3 can be approximated from Eq. 2. However, Eq. 2 does not take into account effects such as threshold voltage variation and gate-

to-source voltage variation. Simulations are necessary to adequately estimate the on-resistance of M_3.

In SRAMs, most of the high-capacitance signals (e.g., the address and bit lines) are in dual-rail form. The energy-recovery latch/driver can generate dual-rail output by duplicating *only* the transistors M_2, M_3, and M_4.

The energy dissipated in the energy-recovery latch/driver can be partitioned into internal energy and output energy. The internal energy is dissipated in driving the internal latch/driver capacitances; the output energy is the energy dissipated in driving the load capacitance through transistor M_3. The energy injected in the internal capacitance of the latch/driver is not recovered, and hence, the ratio of the load capacitance to the internal capacitance is an energy-efficiency metric for the energy-recovery latch/driver. In large SRAMs, the load capacitance is ten or more times greater than the internal latch/driver capacitance.

The energy-recovery latch/driver operates straightforwardly with precharged logic, since the output is active for only one phase. The fundamental overhead of energy-recovery techniques is the extra time required to recover the energy. In memory designs, this overhead is hidden. Typically, the cycle time is determined by the latency to drive the word line. The word line, however, should be switched high and low within the same phase, which is the intrinsic behavior of energy-recovery circuits with resonant clock drivers.

Memory accesses are pipelined with the following sequence for write operations (Fig. 3): In φ_1, the address is latched and the decoder is precharged. In φ_2, the decoder is driven and the word lines are evaluated and latched. In the same phase, the input data are latched. Finally, in the successive φ_1, the word and bit lines are enabled and the write transaction happens.
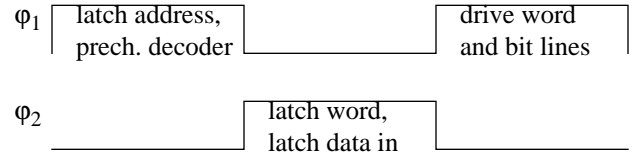

Figure 3: SRAM timing diagram for writes.

3. Simulation Results

In this Section, we describe the simulation experiment and present the simulation results. The target memory size is 256×256. Simulations were performed in HSPICE for an n-well 0.8 µm technology offered through MOSIS. To get meaningful results from HSPICE (i.e., including all the parasitic capacitance in the long metal lines) within a reasonable time, only a portion of the memory was simulated. This portion included a complete row and a complete column from the memory array, the precharge transistors, the decoder, all

the address latches, one word line latch, and the input and output latches for one bit.

The write transaction was simulated at a 200 MHz operating frequency for $V_{ph} = 5$V. Each phase was 2.5 ns (Fig. 4). To increase the charging time and perform adiabatic charging [Atha96], the phase rising and falling edges were 1 ns each, leaving 0.5 ns for each phase to be held at 5 V. There was no gap between the two phases. The outputs of the energy-recovery latches/drivers varied from 4.3 V to 4.9 V. Device M_3 (Fig. 2) was 32 μm wide and its on-resistance during the charging time was 192 Ω which agrees with Eq. 1. However, Eq. 2 indicates a 50% smaller resistance due to the variations of the threshold voltage and the gate-to-source voltage of M_3 during the charging time.
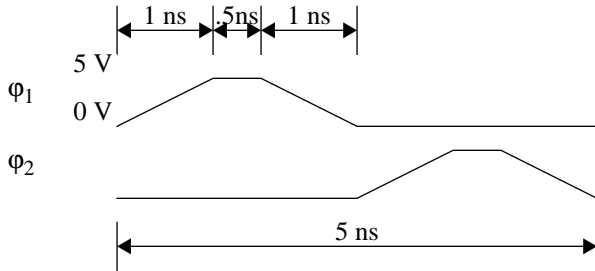

Figure 4: Phases for the energy-recovery SRAM.

The energy dissipation of the various blocks of the SRAM was evaluated separately (Table 1) assuming all returned energy was recycled. The "bit-line driver" row includes the energy dissipation of a single bit column.

	Phases	Supply	Total
Address drivers and Decoder	89.44 pJ	10.94 pJ	100.38 pJ (E_{decER})
Word-line driver	15.99 pJ	2.30 pJ	18.29 pJ (E_{wrdER})
Bit-line driver	11.04 pJ (E_{bitDER})	1.14 pJ (E_{bitIER})	12.18 pJ
Memory cell	N/A	0.38 pJ	0.38 pJ

TABLE 1. Energy dissipation break down for the energy-recovery SRAM.

For comparison purposes, we replaced the energy-recovery latches/drivers with conventional ones and simulated the SRAM at the same speed (e.g., 200 MHz). To minimize the dissipation of the conventional SRAM while maintaining 200 MHz operating frequency, we reduced the supply voltage to 4 V. We determined the energy dissipation in the same parts of the memory (Table 2). In this case, the buses are powered from the dc power supply line and the control signals are used solely to control the operation.

The last column of Table 2 shows the ratio of the total energy dissipation in the energy-recovery SRAM and the conventional one. The energy-recovery SRAM exhibits

	Control	Supply	Total	Ratio
Address drivers and Decoder	47.13 pJ	313.54 pJ	360.67 pJ (E_{decCon})	0.28
Word-line driver	1.37 pJ	42.50 pJ	43.87 pJ (E_{wrdCon})	0.41
Bit-line driver	12.92 pJ	37.64 pJ	50.56 pJ (E_{bitCon})	0.24
Memory cell	N/A	0.30 pJ	0.30 pJ	1.26

TABLE 2. Energy dissipation break down for the conventional SRAM.

lower dissipation in driving the high-capacitance signals. Only flipping the cell requires more energy in the energy-recovery SRAM than in the conventional one. However, this energy is insignificant compared to the energy required from the other parts of the memory.

The difference in dissipation between the energy-recovery SRAM and its conventional counterpart are due to the following reasons:

- In the energy-recovery SRAM, energy is recovered from both the high-capacitance data signals and the control signals. The latter are signals required to synchronize the operation of the SRAM.

- Conventional drivers consist of nFETs and pFETs. The latter become very large (i.e., more than 100 μm wide) in order for the conventional SRAM to operate at high frequencies, and hence, they contribute a significant fraction of the energy dissipation. Conventional driver transistors are 1.5-4 times wider than their counterparts in energy-recovery drivers.

- The energy-recovery drivers have negligible short-circuit power dissipation. In contrast, the large conventional drivers have short-circuit current during their transitions.

4. Energy Dissipation Models

In this Section, we provide some guidelines for estimating the average energy dissipation of the SRAM for write operations. This is difficult because it depends on the dynamic sequence of the memory operations. Since this paper addresses only write operations, we further investigate the average energy dissipation of a write operation that follows another write operation.

There is a key difference between the behavior of the conventional and the energy-recovery SRAMs. Let n denote the number of columns of the RAM array, i.e., the RAM contains a total of $2 \cdot n$ bit lines. In the conventional case, when a write completes, n out of the $2 \cdot n$ bit lines remain charged up. The energy dissipated for a subsequent write operation depends on the number of bit lines that need to be switched. In the energy-recovery SRAM, when a write completes, all

the bit lines are at 0 V. To perform a successive write, n out of the $2 \cdot n$ bit lines need to be pulsed. Therefore, the energy dissipated in the bit lines does *not* depend on the switching activity.

In the following, we model the energy dissipated in the conventional and the energy-recovery SRAMs for a write operation that follows another write. In both cases, we neglect the energy dissipated inside the cell when it is flipped because it is insignificant compared to the energy dissipated in the other parts of the memory.

Let E_{decCon}, E_{wrdCon}, and E_{bitCon} denote the energy dissipation in the row decoder (including the address lines), the word line, and each bit line pair, respectively, for the conventional SRAM. Let a_{sw} be the switching activity factor; a_{sw} denotes the fraction of bits that need to be switched in regards to the previous write operation ($0 \leq a_{sw} \leq 1$). Then, the total energy dissipation for the write operation, E_{con}, is:

$$E_{con} = E_{decCon} + E_{wrdCon} + a_{sw} \cdot n \cdot E_{bitCon} \quad (3)$$

Let E_{decER} and E_{wrdER} denote the energy dissipation in the row decoder (including the address lines) and the word line, respectively, for the energy-recovery SRAM. The dissipation of the bit lines is partitioned into the energy dissipated internally to the latch (E_{bitIER}) and the energy dissipated for driving the bit line (E_{bitDER}). The former is dissipated from the dc power supply only when the current bit is different than the one stored for the previous write and the latter is dissipated from the two phases. The total energy dissipation for the write operation, E_{ER}, is:

$$E_{ER} = E_{decER} + E_{wrdER} + n\,(E_{bitDER} + a_{sw}E_{bitIER}) \quad (4)$$

To further investigate the role of the switching activity factor a_{sw} in the energy dissipation of the conventional (E_{con}) and the energy-recovery (E_{ER}) SRAMs, we plot the ratio E_{ER} / E_{con} as a function of a_{sw} (Fig. 5).
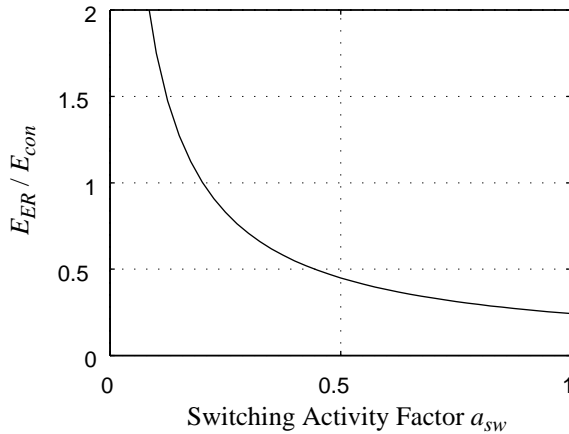

Figure 5: E_{ER} / E_{con} vs. switching activity factor

The energy savings are increased as the switching activity factor increases. Significant energy savings are possible (55%) when half of the bit lines switch. Both SRAMs dissipate the same energy when 20% of the bit lines switch. For less than 20% switching activity, the energy-recovery SRAM dissipates more energy for the write operation compared to its conventional counterpart.

The two SRAM designs would have the same switching behavior if stepwise charging [SvAW96] was employed instead of resonant charging. However, stepwise charging is less energy efficient than resonant charging.

5. Clock-Driver Issues

In this Section, we give a brief description of a resonant clock driver circuit that can be used in conjunction with the energy-recovery latches/drivers (Fig. 2), we discuss issues associated with the resonant clock driver, and, finally, we provide HSPICE simulation results of the portion of the SRAM that was discussed in Section 3 when it operates with the resonant clock driver.

The resonant clock driver (Fig. 6a), which is described in detail elsewhere [AtST96], consists of two inductors (L_1, L_2) and two nFETs (M_1, M_2). The two capacitors (C_{L1}, C_{L2}) indicate the capacitance loads driven by the two phases. The clock driver generates two *almost* non-overlapping phases (Fig. 6b). Despite the small overlap at the edges of the two clock phases, these phases are compatible with the energy-recovery latches/drivers.
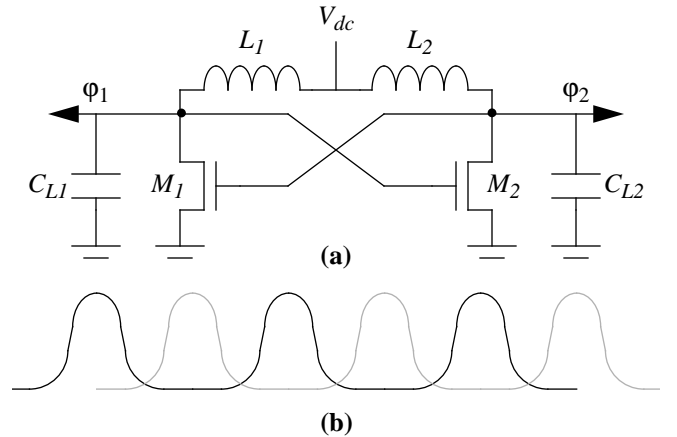

(a)

(b)

Figure 6: Resonant clock driver (a), and phase waveforms (b).

There are some practical issues regarding the operation of the SRAM in conjunction with the resonant clock driver. The most important issue is that the load capacitances driven by the two phases should be balanced. Relatively small capacitance imbalances do not affect the operation of the clock driver. However, in the case of the SRAM organization shown in Fig. 1, there is a large difference between the load capacitances of the two phases. Assuming a $n \times n$ RAM array, φ_1 drives n bit lines and the word line, whereas φ_2 drives only $\log_2 n$ address lines. The phase load capacitances can be

perfectly balanced if the RAM array is split into two sub-arrays with $n/2$ columns, each of which driven in different phases. The overhead of this RAM organization is: (a) one more row decoder is required, and (b) the memory operation latency increases by one phase. The row decoder space is negligible compared to the size of the RAM array. The latency overhead does not affect the memory throughput since it is still possible to complete one operation per cycle.

Another practical issue is the clock driver implementation. At least one clock driver is necessary for the entire system. Since it is expensive to implement on-chip high-Q inductors, the most practical way to implement the clock driver is to use off-chip high-Q inductors. Assume the SRAM organization with balanced capacitance loads for each phase and let C and L denote the load capacitance and inductance respectively. Then the operating frequency f is approximately:

$$f \approx \frac{1}{2\pi\sqrt{LC}} \qquad (5)$$

For a 256×256 SRAM structure with two RAM sub-arrays (i.e., 256×128 each one), the capacitance load for each phase is roughly 160 pF for the 0.8 μm technology. For a 200 MHz operating frequency, each inductor should be 4 nH.

For the sake of comparison, we simulated in HSPICE the portion of the SRAM that was used for the experiments of Section 3 with the resonant clock driver at 200 MHz. Each inductor was 60 nH since only a portion of the RAM was simulated. Each one of M_1 and M_2 was 400 μm wide for the same 0.8 μm technology. The energy dissipation of the phases is 127.21 pJ (Table 3) as opposed to 116.47 pJ when the phases are linear ramps (Section 3).

Energy recovery	Phases	Supply	Total
Resonant phases	127.21 pJ	12.89 pJ	140.10 pJ
Linear ramps	116.47 pJ	14.38 pJ	130.85 pJ
Conventional	**Control**	**Supply**	**Total**
Conventional drivers	61.42 pJ	450.10 pJ	511.52 pJ

TABLE 3. Summary of energy dissipation simulations.

6. Conclusions

In this paper, we presented a low-power, high-speed SRAM based on energy-recovery techniques. The energy-recovery SRAM can be readily derived from a conventional design by replacing the conventional latches/drivers with bootstrapped energy-recovery ones. In addition to recovering energy, the energy-recovery latches drive the high-capacitance lines through nFETs. That reduces the dissipation inside the latches and makes them overall smaller in size.

Energy recovery results in significant energy savings (e.g., from 59% to 76%) in the simulated SRAM parts for write operations compared to a conventional design at 200 MHz. A direct comparison between the two designs is possible only when the dynamic SRAM behavior is known because the two SRAM designs exhibit different switching behavior. However, based on the simulation results, if on average 50% of the bit lines switch in successive write operations, the energy-recovery SRAM would dissipate 55% less energy than the conventional design.

Energy-recovery latches/drivers might also be attractive in dynamic 4-transistor memory structures, where the bit signals must be full voltage-swing signals for write operations.

Acknowledgments

References

[AlAn95]   J. Alowersson, P. Andersson, *SRAM Cells for Low-Power Write in Buffer Memories*, Proc. of the 1995 Symposium on Low Power Electronics, San Jose, CA, October 9-11, 1995.

[AmHo94]   B.S. Amrutur, M. Horowitz, *Techniques to Reduce Power in Fast Wide Memories*, Proc. of the 1994 Symposium on Low Power Electronics, San Diego, CA, October 1994.

[Atha96]   W.C. Athas, *Energy-Recovery CMOS*, in *Low Power Design Methodologies*, edited by J. Rabaey, M. Pedram, Kluwer Academic Publishers, 1996.

[AtST96]   W.C. Athas, L. "J." Svensson, N. Tzartzanis, *A Resonant Signal Driver For Two-Phase, Almost-Non-Overlapping Clocks*, Proc. of the 1996 International Symposium on Circuits and Systems, Atlanta, GA, May 12-15, 1996.

[AtTz95]   W.C. Athas, N. Tzartzanis, *Energy Recovery for Low-Power CMOS*, Proc. of the 1995 Chapel Hill Conference on VLSI, Chapel Hill, NC, March 27-29, 1995.

[GlDo85]   L.A. Glasser, D.W. Dobberpuhl, *The Design and Analysis of VLSI Circuits*, Addison-Wesley, Reading, MA, 1985.

[SoYR95]   D. Somasekhar, Y. Ye, K. Roy, *An Energy Recovery Static RAM Memory Core*, Proc. of the 1995 Symposium on Low Power Electronics, San Jose, CA, October 9-11, 1995.

[SvAW96]   L. "J." Svensson, W.C. Athas, R.S.-C. Wen, *A sub-CV^2 pad driver with 10 ns transition time*, Proc. of the 1996 ISLPED, Monterey, CA, August 12-14, 1996.