Minimizing the Energy-Delay Product of SRAM Arrays using a Device-Circuit-Architecture Co-Optimization Framework

Alireza Shafaei Hassan Afzali-Kusha Massoud Pedram

Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089
{shafaeib, afzaliku, pedram}@usc.edu

ABSTRACT

The objective of this paper is to minimize the energy-delay product of static random access memory (SRAM) arrays by using a device-circuit-architecture co-optimization framework. More specifically, at the device-level, high-V_t FinFETs are adopted for the 6T SRAM cell, which significantly reduces the leakage power and improves static noise margins. However, due to the lower ON current, the bitline delay of the read access is increased. Accordingly, at the circuit-level, the voltage level of assist circuits, and at the architecture-level (i.e., the array organization), key parameters of the SRAM array are jointly optimized to derive a design that results in the minimum energy-delay product point. By using the proposed optimization framework, for SRAM array capacities ranging from 1KB to 16KB, on average 59% lower energy-delay product with maximum 12% (and on average 9%) performance penalty is achieved.

CCS Concepts

•Hardware → Static memory; Power estimation and optimization;

Keywords: SRAM array; assist techniques; energy-efficient memory design

1. INTRODUCTION

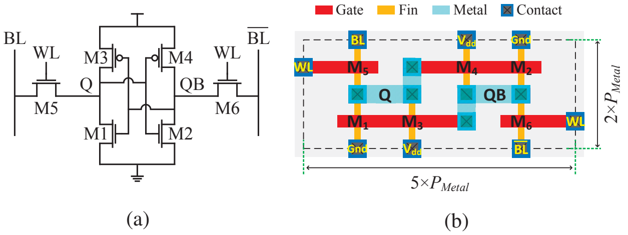
Proper operation of the standard 6T *static random access memory* (SRAM) cell (cf. Figure 1(a)) relies on the relative strength of its underlying transistors. More specifically, for a non-destructive read operation, access transistors should be weaker than pull-down transistors during the read operation such that access transistors cannot flip the cell content. Furthermore, stronger access transistors compared with pull-up transistors are needed during the write operation such that access transistors can successfully write into the SRAM cell. This design can easily fail in advanced technology nodes where the effect of process variations due to small geometries and low supply voltage, V_{dd}, levels are increasing. More robust SRAM cell structures exist (e.g., [3, 2]), but such SRAM cells come at the cost of larger layout area.

In advanced technology nodes, bulk CMOS transistors are replaced with FinFET devices. This is because of the three-dimensional gate control over the channel in FinFETs, which improves the ON/OFF current ratio and increases the immunity of the device to random variations [14, 11, 15]. However, one of the challenges associated with FinFETs at the circuit-level is the *width quantization property*, which dictates the FinFET width to only take discrete values. As a result, fine-grained control over transistor sizing becomes difficult in FinFET technologies. For an SRAM cell, the ideal case in terms of area footprint is to use single-fin devices for all transistors.

To achieve an area and power efficient SRAM cell design, the all-single-fin 6T SRAM cell operating at low voltages has become popular [13, 5, 8]. Degraded stability and performance values due to low voltage operation are then improved by assist circuits. Decreasing V_{dd} reduces leakage, and more importantly, dynamic powers. However, because of negligible *drain induced barrier lowering* (DIBL) effect in FinFETs, and due to the higher contribution of leakage power to the total power consumption, especially for large arrays, power/energy savings are limited by reducing V_{dd}. An alternative and more effective approach, which is adopted in this paper, is to use high-V_t (HVT) devices in SRAM cells. By using these devices, leakage power is significantly reduced (because of lower OFF currents), and noise margins are improved (due to higher ON/OFF current ratios).

The major issue associated with HVT devices is lower ON currents. Performance of the SRAM array is thus degraded which is mainly caused by the reduced read current, resulting in higher bitline (BL) delay. Therefore, we jointly optimize voltage levels of assist techniques (which are also needed to maintain the cell stability) along with key parameters of SRAM array related to the BL (including number of rows, and number of fins of precharger and write buffer) in order to increase the read current and find the array design with the optimal energy-delay product.

For this purpose, various read and write assist techniques and their effect on reliability and performance metrics of the corresponding memory operation are investigated. Based on our analysis, V_{dd} boost and wordline (WL) overdrive techniques are selected to enhance read and write stabilities, respectively, whereas negative Gnd technique is chosen for increasing the read current of the SRAM cell. Furthermore, analytical models for delay and energy consump-



(a) (b)

Figure 1. (a) Schematic and (b) layout [8] of FinFET-based 6T SRAM cell. P_{Metal} denotes the metal pitch.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DAC '16, June 05-09, 2016, Austin, TX, USA
© 2016 ACM. ISBN 978-1-4503-4236-0/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2897937.2898044

tion of an SRAM array considering afore-said read and write assist techniques are presented.

Our device-circuit-architecture co-optimization framework is evaluated using 7nm FinFET devices under nominal V_{dd} level of 450mV [4]. Simulation results are obtained for memory arrays with different capacities (from 128B to 16KB) using HVT and low-V_t (LVT) devices for SRAM cells (in both cases, peripheral circuits are made of LVT devices). For small capacity arrays, because of lower impact of leakage power and small BLs, results of LVT and HVT arrays are close. However, for SRAM array capacities ranging from 1KB to 16KB, on average 59% lower energy-delay product with maximum 12% (and on average 9%) performance penalty is achieved when HVT SRAM cells equipped with negative Gnd technique are used in the array.

Due to the intrinsic lower ON current of HVT devices compared with their LVT counterpart, a performance penalty is inevitable; however, the main purpose of our optimization framework is to minimize the performance gap between HVT and LVT arrays. Accordingly, we show that the negative Gnd technique is an effective solution for reducing the BL delay of the SRAM array.

The rest of the paper is organized as follows. Benefits and challenges of using HVT devices in SRAM cells are discussed in Section 2. Assist techniques and the SRAM array model are presented in Section 3 and Section 4, respectively. Simulation results are reported in Section 5, and finally, Section 6 concludes the paper.

2. 6T SRAM CELL WITH HVT DEVICES

In this paper, a 7nm FinFET library with a nominal supply voltage of 450mV is adopted [4]. This library includes LVT and HVT devices, where HVT devices compared with their LVT counterparts have 2× lower ON current, 20× lower OFF current, and 10× higher ON/OFF current ratio. In our SRAM arrays, for performance considerations, peripheral circuits are made of LVT devices, but SRAM cell transistors are either LVT or HVT. The 6T SRAM cell made of LVT (HVT) devices will be referred to as 6T-LVT (6T-HVT). To achieve an area efficient cell footprint, single-fin device are used for all transistors of 6T-LVT and 6T-HVT SRAM cells.

The lower OFF current of HVT devices significantly reduces the leakage power of the SRAM cell. Figure 2(a) shows leakage powers of 6T-LVT and 6T-HVT SRAMs under scaled voltages. By reducing the V_{dd} from the nominal value to 100mV, V_{dd} is decreased by 4× (3×) in 6T-LVT (6T-HVT) SRAM. However, under the nominal V_{dd} operation, 20× leakage power reduction is gained by adopting HVT devices. Moreover, leakage power of 6T-LVT at 100mV (which is difficult to realize due to the increased susceptibility to noises and process variations under such ultra-low voltage) is still 5× higher than that of the 6T-HVT at 450mV. As a result, HVT devices are quite effective in substantial leakage power reduction in SRAMs.

Another advantage of HVT devices is the higher ON/OFF current ratio, which helps in increasing the *static noise margin* (SNM) of the cell. Hold SNMs (HSNMs) of 6T-LVT and 6T-HVT SRAMs, which are measured based on butterfly curves [12], are shown in Figure 2(b) for different V_{dd} values. Based on our Monte Carlo analysis, noise margins of the 6T SRAM cell using the adopted 7nm FinFETs should be greater than 35% of V_{dd} in order to achieve a high-yield SRAM cell. Accordingly, as shown in Figure 2(b), while 6T-LVT cannot meet the yield requirements under 250mV, 6T-HVT can reliably hold data for all shown V_{dd} levels. Also, read SNM (RSNM) of 6T-HVT under nominal V_{dd} is 1.9× larger than that of the 6T-LVT (cf. Figure 3(a)), but it is still lower than 35% of V_{dd}.

Lower ON current is the major drawback of HVT devices, which reduces the read current of the SRAM cell. More precisely, for the read operation, BLs are initially precharged to V_{dd}, and then WL of the accessed cell is activated. Next, the BL, which is connected to the SRAM node that stores ‘0’, is discharged through corresponding pull-down and access transistors. When the voltage level of BL is dropped by a certain value, called sensing voltage and denoted by ΔV_S, sense amplifier is enabled. As a result, by using HVT devices, pull-down and access transistors are weakened, which in turn reduces the read current. Consequently, BL delay (or, sensing delay defined as the time that is needed to discharge BL to $V_{dd} - \Delta V_S$) is increased.

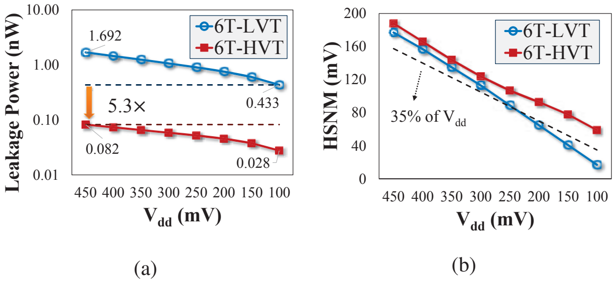


Figure 2. Comparison of (a) HSNMs and (b) leakage powers of 6T-LVT and 6T-HVT SRAMs under different V_{dd} values. Vertical axis in (a) is in logarithmic (base 10) scale.

Considering the $C{\cdot}V/I$ equation, the BL delay in 6T-HVT SRAM array may be decreased by following three approaches. (i) *Reducing* ΔV_S, which is difficult to do especially in advanced technology nodes with increased effect of process variations. (ii) *Increasing the read current of the SRAM cell*, which can be done by assist circuits. Such circuits are also needed to improve the RSNM and write margin of the all-single-fin 6T SRAM cell. (iii) *Decreasing the BL capacitance*, which is possible to achieve by decreasing the number of rows (or the number of SRAM cells in each column) of the array. Accordingly, we co-optimize voltage level of assist techniques along with the key parameters of the SRAM array related to BL to minimize the energy-delay product of the 6T-HVT array.

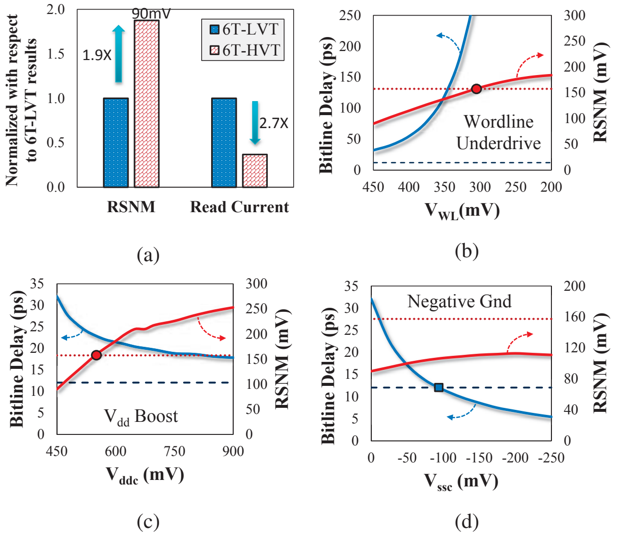
3. ASSIST CIRCUITS FOR 6T SRAM

The purpose of assist techniques is to increase the reliability and performance metrics of read and write operations. This is generally achieved by increasing/decreasing the voltage level of WL, BL, cell V_{dd}, or cell V_{ss} from their nominal values during read and write operations. In the rest of this section, different read and write assist techniques are discussed, and their impact on noise margins and access delays of 6T-HVT SRAM are analyzed.

3.1 Read-Assist Circuits

To enhance the read stability of the 6T SRAM cell, one can strengthen the pull-down transistor or weaken the access transistor. Accordingly, widely-used read-assist techniques are as follows [18]: (i) **WL underdrive (WLUD):** Voltage of WL (denoted by V_{WL}), which is applied to the gate terminal of the access transistor, is set to a voltage level lower than V_{dd}. Thus, access transistor is weakly turned on. (ii) **V_{dd} boost:** Supply voltage level of the cell, represented as V_{DDC}, is increased above V_{dd}. (iii) **Negative Gnd:** A negative voltage, denoted by V_{SSC}, is applied to the source terminal of the pull-down transistor.

Effects of the afore-said assist techniques on reliability (RSNM) and performance (BL delay, assuming that 64 cells are in each column) are shown in Figure 3(b)-(d). The WLUD technique, because of weakening the access transistor, increases the RSNM and for V_{WL}=300mV (33% lower than the nominal value) can meet the yield requirement. However, weakening the access transistor also reduces the read current, a result of which is increase in the BL delay. Increasing V_{WL} during read operation has the opposite effect, i.e., higher read current but reduced RSNM. Therefore, because of degrading either RSNM or read current, we opted not to utilize the WLUD technique.



(a) (b) (c) (d)

Figure 3. (a) Comparison of RSNM and read current values of 6T-LVT and 6T-HVT SRAMs (results are normalized with respect to 6T-LVT values). Effect of (a) V_{dd} boost, (b) negative Gnd, and (c) wordline underdrive read-assist techniques on BL delay and RSNM of 6T-HVT SRAM. For bitline delay, a column with 64 SRAM cells is assumed. Top (bottom) vertical line in (b)-(d) denotes the minimum acceptable RSNM level (BL delay of 6T-LVT with no assist), and a circle (square) is used to indicate the cross point.

Voltage levels of SRAM signals during read operation, for an SRAM cell storing '0' and assuming that both V_{dd} boost and negative Gnd techniques are applied, are shown in Figure 4. As a result of V_{dd} boost, voltage level of gate terminal of pull-down transistor becomes V_{DDC}, which significantly strengthens the pull-down transistor, resulting in higher RSNMs. As can be seen in Figure 3(c), for V_{DDC}=550mV (22% higher than the nominal value), yield requirement is satisfied. Pull-down transistor then writes V_{SSC} on node Q, which subsequently increases the drain-to-source voltage of access transistor, and thus read current is significantly increased. Because of this, with $V_{SSC} = -100$mV (whose absolute value is 22% of V_{dd}), the BL delay of 6T-HVT array becomes same as the BL delay of 6T-LVT array with no assist technique. However, since negative Gnd makes both access and pull-down transistors stronger, the influence on RSNM is less powerful.

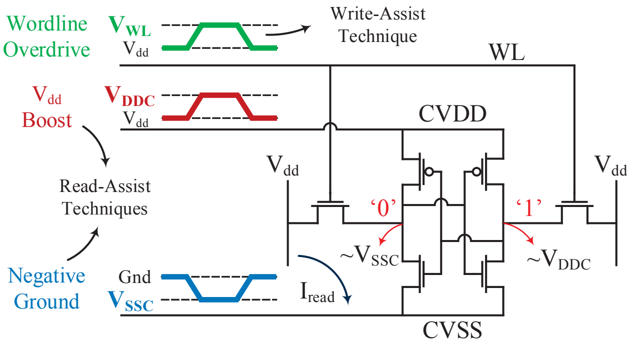
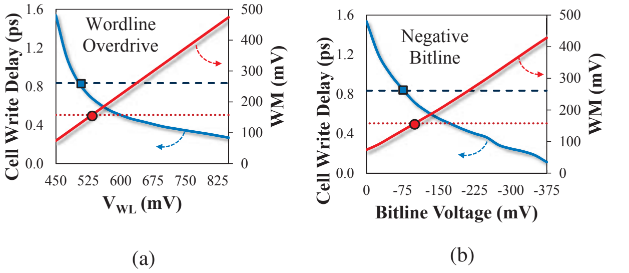


Figure 4. Read and write assist techniques adopted in this paper. Voltage level of each signal for an SRAM cell storing bit '0' during read operation is also shown.



(a) (b)

Figure 5. Effect of (a) wordline overdrive, and (b) negative bitline write-assist techniques on cell write delay and WM of 6T-HVT SRAM. Top (bottom) vertical line denotes the cell write delay of 6T-LVT with no assist (minimum acceptable WM level), and a square (circle) is used to indicate the cross point.

Based on the above discussions, both V_{dd} boost and negative Gnd techniques are effective in increasing the RSNM and read current of the SRAM cell, respectively. Thus, we simultaneously apply both techniques to the 6T-HVT SRAM. However, larger values of V_{DDC} and V_{SSC} also increase the energy consumption of the SRAM array. Therefore, to minimize the energy-delay product, optimal values of V_{DDC} and V_{SSC} should be derived.

3.2 Write-Assist Circuits

Characteristics of an SRAM cell related to the write operation include the write margin (WM) and cell-level write delay. WM is measured as the difference between the V_{dd} and the minimum WL voltage that is needed to flip the cell content [9], and cell-level write delay is defined as the time WL reaches 50% of V_{dd} until Q and QB reach the same value.

Both WM and cell write delay rely on the relative strength of the access transistor to that of the pull-up transistor. Hence, by strengthening the access transistor both WM and cell write delay can be improved. For this purpose, the following write-assist techniques can be used [18]: (i) **WL overdrive (WLOD):** V_{WL} is set to a voltage level higher than V_{dd} to strongly turn on the access transistor. (ii) **Negative BL:** Write operation into the SRAM cell conventionally (without write-assist) occurs from a BL that is 0. By using a negative voltage for that BL, the gate-to-source voltage becomes larger, more strongly turning on the access transistor.

Figure 5 shows the effect of WLOD and negative BL techniques on the WM and cell write delay of the 6T-HVT SRAM. Negative BL has a higher impact on reducing the cell write delay, but this delay even without using write-assists (which is 1.5ps) is much lower than WL and BL delays. In terms of WM, WLOD and negative BL meet the write yield requirements at V_{WL}=540mV (20% higher than the nominal value) and $V_{BL} = -100$mV (22% of V_{dd}), respectively. These results show that WLOD is slightly more effective in improving the WM, which can also be concluded from the definition of WM. Therefore,we opted to utilize the WLOD technique as the write-assist technique for the 6T-HVT SRAM (cf. Figure 4). Same as V_{DDC} and V_{SSC}, the value of V_{WL} should be optimized.

4. SRAM ARRAY MODEL

In this section, an analytical model for the SRAM array, which considers effects of various adopted assist techniques is presented. Please note that our analytical model is different from that of [6] because our model accounts for the effects of adopted read/write assist techniques. An SRAM array organized with n_r rows and n_c columns, where n_r and n_c are powers of two, is assumed. Hence, the array contains $M = n_r \cdot n_c$ bits. Peripheral circuits are modeled as shown in Figure 6 (the figure shows only one SRAM cell). For assist circuits, same as [16] and [17], we use multiplexers to dynamically switch between appropriate voltage levels for cell V_{dd} (CVDD), cell V_{ss} (CVSS), and WL rails. V_{DDC}, V_{SSC}, and V_{WL} are provided by

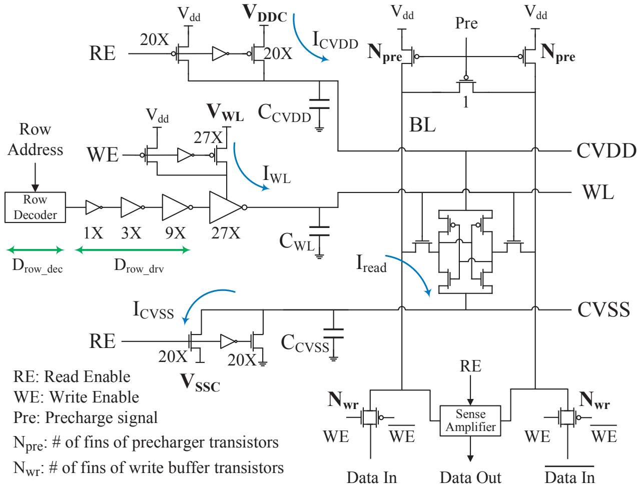


Figure 6. SRAM array model. Optimization variables are boldfaced. Peripheral circuits needed for the column decoder are not shown in this figure. V_{DDC} and V_{SSC} are shared among all SRAM cells in each row. V_{WL} is shared among all last stage inverters of WL drivers [17].

Table 1. Interconnect (wire) capacitances in the SRAM array. COL is the output of the column decoder. C_{dn} (C_{dp}) and C_{gn} (C_{gp}) denote the drain and gate capacitances of single-fin n-channel (p-channel) FinFET device. N_x parameters are the number of fins of the corresponding transistor which are defined in Figure 6.

Wire	Parameter	Equation
CVDD	C_{CVDD}	$n_c(C_{width} + 2C_{dp}) + 2 \times 20 \times C_{dp}$
CVSS	C_{CVSS}	$n_c(C_{width} + 2C_{dn}) + 2 \times 20 \times C_{dn}$
WL	C_{WL}	$n_c(C_{width} + 2C_{gn}) + 27(C_{dn} + C_{dp})$
COL	C_{COL}	0, if $n_C \leq W$ $n_c C_{width} + 27(C_{dn} + C_{dp})$ $+2WN_{wr}(C_{gn} + C_{gp})$, if $n_C > W$
BL	C_{BL}	$n_r(C_{height} + C_{dn}) + (N_{pre} + 1)C_{dp}$ $+N_{wr}(C_{dn} + C_{dp}) + C_{dp}$, if $n_C \leq W$ $n_r(C_{height} + C_{dn}) + (N_{pre} + 1)C_{dp}$ $+2N_{wr}(C_{dn} + C_{dp})$, if $n_C > W$

an external source or an on-die DC-DC converter [16, 17]. Moreover, each output of row decoder is connected to a driver. The design of this driver (superbuffer) is derived analytically and verified by SPICE simulations for the adopted FinFETs. To avoid large area overheads, four inverter stages are used.

We also assume that in each access cycle, W bits are read or written. Accordingly, when $n_c > W$, additional circuits for the column multiplexer are needed, which are not shown in Figure 6. For such cases, a column decoder and associated drivers (similar to the row decoder) are included [7]. Further, a transmission gate is added between the end of each BL and write buffer/sense amplifier. The input data bit is thus written through two transmission gates in series. In addition to n_r and n_c, N_{pre} and N_{wr} (i.e., number of fins of precharger and write buffer transistors, respectively) are also included as array optimization variables. This is because by increasing N_{pre} and N_{wr}, precharge and BL write delays decrease, respectively, but at the same time, BL capacitance increases which may increase the read delay.

Capacitance equations for all array interconnects are reported in Table 1. Wire capacitance has been considered in this paper as follows. Based on the layout geometries of 6T SRAM cell (cf. Figure 1(b)), wire capacitance of cell width (height) is given by $C_{width} = 5 \times P_{Metal} \times C_w$ ($C_{height} = 0.4 \times C_{width}$), where P_{Metal} and C_w denote metal pitch and wire capacitance per μm, respectively. The delay, D, and switching enegy consumption, E_{sw}, of an interconnect are then calculated using the following equations:

$$D = \frac{C \cdot \Delta V}{I}, \qquad E_{sw} = V \cdot I \cdot D = C \cdot V \cdot \Delta V, \tag{1}$$

where, C and ΔV are the total capacitance and voltage change of the interconnect, respectively, and V and I denote the supply voltage and average current values of the interconnect driver, respectively. Values of C, V, ΔV, and I are shown in Table 2 for different interconnect-related delay/energy components of the array.

Equations for delay and switching energy consumption of read and write operations are given in Table 3 for $n_c > W$. If $n_c \leq W$, then all components associated with the column multiplexer become 0. These equations are written for the cell at the top-right corner of the array (the worst-case). For read access, WL should arrive at the cell, then BL starts discharging. At the same time, column decoder should activate the last column multiplexer. Hence, the maximum of these two delays is added to the sense amplifier delay and precharge time. Moreover, to ensure a robust read operation, CVDD and CVSS should arrive to their final value before WL reaches 50% of V_{dd}. For this purpose, number of fins of the PFET (NFET) device that drives CVDD (CVSS) is set to 20 (which is obtained for $n_c = 1024$). For write access, WL and input data bit simultaneously move toward the cell, then we should wait for the cell write and precharge delays.

The total delay and energy consumption of the array, represented by D_{array} and E_{array}, respectively, are obtained as follows:

$$\begin{aligned} D_{array} &= \max(D_{rd}, D_{wr}), && (2)\\ E_{array,sw} &= \beta \cdot E_{sw,rd} + (1-\beta) \cdot E_{sw,wr}, && (3)\\ E_{array,leak} &= M \cdot P_{leak,sram} \cdot D_{array}, && (4)\\ E_{array} &= \alpha \cdot E_{array,sw} + E_{array,leak}, && (5) \end{aligned}$$

where β denotes the ratio of read accesses to the total accesses, α is the array activity factor (defined as the probability of accessing the array in a cycle), $P_{leak,sram}$ is the leakage power of the SRAM cell, and $E_{array,sw}$ ($E_{array,leak}$) denotes the switching (leakage) component of the array energy consumption. Leakage power of peripheral circuits is very small compared with the leakage power of SRAM cells, especially when large number of cells exist. Hence, only leakage power of SRAM cells is taken into account in Equation (4). Furthermore, for definitions of D_{rd}, D_{wr}, $E_{sw,rd}$, and $E_{sw,wr}$ please refer to Table 3.

Finally, the optimization problem is defined as follows. **Given** M (i.e., the memory capacity in bits), **find** the values of V_{DDC}, V_{SSC}, V_{WL}, n_r ($n_c = M/n_r$), N_{wr}, and N_{pre}, **such that** $E_{array} \times D_{array}$ is minimized, **while** yield requirements of the SRAM cell are satisfied. An accurate way to analytically express the constraint is: $\min((\mu - k\sigma)_{\text{HSNM}}, (\mu - k\sigma)_{\text{RSNM}}, (\mu - k\sigma)_{\text{WM}}) \geq 0$, where $1 \leq k \leq 6$ depending on the yield requirements. However, for simplicity, the following constraint will be used in this paper: $\min(\text{HSNM}, \text{RSNM}, \text{WM}) \geq \delta$, where δ is the minimum acceptable noise margin level.

5. SIMULATION RESULTS

Simulation results are presented in this section, which are obtained using the following values: $V_{dd} = 450$mV (nominal supply voltage), $\beta = 0.5$, $\alpha = 0.5$, $\delta = 0.35 \times V_{dd} = 158$mV, $W = 64$ bits, $\Delta V_S = 120$mV, $P_{Metal} = 43$nm (obtained for 7nm FinFET from the scaling factor of Intel 14nm FinFET with respect to Intel 22nm FinFET [10]), and $C_W = 0.17$fF (calculated based on ITRS 2012 report [1] for 7nm node). Gate and drain capacitances, currents in Table 2, as well as delays and energy consumptions of decoder, driver, sense amplifier, and cell-level write are measured by SPICE simulations, and those with dependencies on a variable

Table 2. C, V, ΔV, and I values needed to derive the delay and switching energy consumption of different components of the array. $I_{ON,PFET}$ and $I_{ON,TG}$ denote the ON current of a single-fin PFET and transmission gate, respectively. For definitions of other currents please refer to Fingure 6. Coefficients used for each I are obtained for adopted FinFET devices to fit the model with SPICE simulations.

Parameter	**Subscript**	C	V	ΔV	I
Cell V_{dd} rail	$CVDD$	C_{CVDD}	V_{dd}	$V_{DDC} - Vdd$	$0.30 \times 20 \times I_{CVDD}(V_{DDC})$
Cell V_{ss} rail	$CVSS$	C_{CVSS}	V_{dd}	$\lvert V_{SSC} \rvert$	$0.15 \times 20 \times I_{CVSS}(V_{SSC})$
WL during read	WL, rd	C_{WL}	V_{dd}	V_{dd}	$0.25 \times 27 \times I_{ON,PFET}$
WL during write	WL, wr	C_{WL}	V_{dd}	V_{WL}	$0.18 \times 27 \times I_{WL}(V_{WL})$
Column decoder (COL)	COL	C_{COL}	V_{dd}	V_{dd}	$0.33 \times 27 \times I_{ON,PFET}$
BL during read	BL, rd	C_{BL}	$V_{DDC} - V_{SSC}$	ΔV_S	$I_{read}(V_{DDC}, V_{SSC})$
BL during write	BL, wr	C_{BL}	V_{dd}	V_{dd}	$0.50 \times N_{wr} \times I_{ON,TG}$
Precharge after read	PRE, rd	C_{BL}	V_{dd}	ΔV_S	$0.50 \times N_{pre} \times I_{ON,PFET}$
Precharge after write	PRE, wr	C_{BL}	V_{dd}	V_{dd}	$0.50 \times N_{pre} \times I_{ON,PFET}$

Table 3. Equations for delay and switching energy consumption of read and write operations. D_{row_dec} (D_{col_dec}) denotes the propagation delay of the row (column) decoder which is a function of number of rows (number of words in a row). D_{row_drv} (D_{col_drv}) is the propagation delay of first three stages of the WL (COL) driver. D_{sense_amp} and D_{write_sram} are the sense amplifier and cell write delays, respectively, where cell write delay is a function of V_{WL}. Other delay components are defined in Table 2. Energy components are defined similarly.

Parameter	**Operation**	**Equation**
Delay	Read	$D_{rd} = \max(D_{row\_dec}(\log(n_r)) + D_{row\_drv} + D_{WL,rd} + D_{BL,rd},\ D_{col\_dec}(\log(n_c/W)) + D_{col\_drv} + D_{COL}) + D_{sense\_amp} + D_{PRE,rd}$
Delay	Write	$D_{wr} = \max(D_{row\_dec}(\log(n_r)) + D_{row\_drv} + D_{WL,wr},\ D_{col\_dec}(\log(n_c/W)) + D_{col\_drv} + D_{COL} + D_{BL,wr}) + D_{write\_sram}(V_{WL}) + D_{PRE,wr}$
Energy	Read	$E_{sw,rd} = E_{row\_dec}(\log(n_r)) + E_{row\_drv} + E_{WL,rd} + E_{BL,rd} + E_{col\_dec}(\log(n_c/W)) + E_{col\_drv} + E_{COL} + E_{sense\_amp} + E_{PRE,rd} + E_{CVDD} + E_{CVSS}$
Energy	Write	$E_{sw,wr} = E_{row\_dec}(\log(n_r)) + E_{row\_drv} + E_{WL,wr} + E_{col\_dec}(\log(n_c/W)) + E_{col\_drv} + E_{COL} + E_{BL,wr} + E_{write\_sram}(V_{WL}) + E_{PRE,wr}$

are stored in look-up tables. Energy consumptions of assist circuits are multiplied by a scaling factor to account for inefficiency of DC-DC converters. Also, the leakage power of 6T-LVT (6T-HVT) is 1.692nW (0.082nW) (cf. Figure 2(a)).

Among optimization variables, V_{DDC} and V_{WL} are set to the minimum voltage levels that meet yield requirements of RSNM and WM, respectively. This is because increasing V_{DDC} increases the read energy consumption, but has no impact on read delay (see $E_{sw,rd}$ and D_{rd} in Table 3). On the other hand, increasing V_{WL} increases both write energy consumption and WL delay, while the contribution of cell write delay (which is decreased by larger V_{WL} values) to the overall delay is negligible. Hence, the purpose of optimizing V_{DDC} and V_{WL} is to maintain noise margins above the minimum acceptable level. Based on SPICE simulations, we have V_{DDC}=640mV (550mV) and V_{WL}=490mV (540mV) for 6T-LVT (6T-HVT). HSNM in both SRAMs at 450mV is above δ.

For other optimization variables, the following ranges are assumed: $V_{SSC} = \{0, -10mV, \cdots, -240mV\}$ (since below -240mV RSNM degrades), $n_r = \{2^1, 2^2, \cdots, 2^{10}\}$, $N_{pre} = \{1, 2, \cdots, 50\}$, and $N_{wr} = \{1, 2, \cdots, 20\}$. Because only four variables with relatively small ranges are left, we can derive the minimum energy-delay product point of the array using an exhaustive search. All simulation results in this section are obtained in less than two minutes using a server machine with Intel E7-8837 processor and 64GB memory running Debian 8.2.

In our simulations, we consider two methods: (i) **M1:** only one extra voltage level (a high voltage) other than V_{dd} is available, whose value is set to $\max(V_{DDC}, V_{WL})$, i.e., 640mV (550mV) for 6T-LVT (6T-HVT). (ii) **M2:** No restriction on the number of voltage levels is considered. Hence, we have three extra pins for LVT-based array (V_{DDC}=640mV, V_{WL}=490mV, and V_{SSC}). However, since V_{DDC} and V_{WL} are very close in 6T-HVT, only two pins are used for HVT-based array ($V_{DDC} = V_{WL}$=550mV, and V_{SSC}).

Delay, energy, and energy-delay product values of different SRAM arrays with different capacities are shown in Figures 7(a), 7(b), and 7(c), respectively. Design parameters of SRAM arrays are also reported in Table 4. Since read delay is typically greater than write delay, a slack for write delay is available. Therefore, smaller N_{wr} values are used which reduce the BL capacitance and allows N_{pre} to have larger values (increasing N_{pre} is important to reduce the precharge delay).

A very effective approach for BL delay reduction is the negative Gnd technique. The read current can be expressed analytically as $I_{read} = b \cdot (V_{DDC} - V_{SSC} - V_t)^a$, where based on our fitting results, a=1.3, b=0.000095A/V$^{1.3}$, and V_t=335mV for HVT devices. For 6T-HVT, by using $V_{SSC} = -240$mV instead of 0, when V_{DDC}=550mV, 4.3× increase in I_{read} is gained which directly affects the BL delay. As for the aspect ratio of the array, since width of 6T SRAM cell is 2.5× larger than its height, smaller number of columns is usually preferred. This can effectively happen when read current of SRAM cell, due to the negative Gnd technique, is significantly increased.

Based on Figure 7(a), 6T-HVT-M1[1] has the highest delay, mainly due to low read currents. However, as indicated in Figure 7(d), BL delay and hence the total delay are significantly reduced in 6T-HVT-M2 (on average 3.3× for BL delay and 1.8× for total delay), which points to the effectiveness of negative Gnd technique in reducing the overall delay. As expected, 6T-LVT-M2 has the lowest delay, and 6T-HVT-M2 compared with 6T-LVT-M2 has on average 9%

[1] We use this notation to refer to an array which uses 6T-HVT SRAM cell and implements method M1. Similar notation is used for other cases.

Table 4. SRAM array design parameters for the minimum energy-delay point. Voltages are reported in mV.

M	SRAM	n_r	n_c	N_{pre}	N_{wr}	V_{DDC}	V_{SSC}	V_{WL}
128B	6T-LVT-M1	64	16	7	1	640	0	640
	6T-HVT-M1	32	32	4	1	550	0	550
	6T-LVT-M2	64	16	8	1	640	-210	490
	6T-HVT-M2	64	16	7	1	550	-240	550
256B	6T-LVT-M1	64	32	7	1	640	0	640
	6T-HVT-M1	64	32	5	1	550	0	550
	6T-LVT-M2	64	32	9	1	640	-180	490
	6T-HVT-M2	64	32	8	1	550	-230	550
1KB	6T-LVT-M1	128	64	12	1	640	0	640
	6T-HVT-M1	128	64	7	1	550	0	550
	6T-LVT-M2	128	64	16	2	640	-240	490
	6T-HVT-M2	128	64	12	2	550	-240	550
4KB	6T-LVT-M1	256	128	18	4	640	0	640
	6T-HVT-M1	256	128	11	2	550	0	550
	6T-LVT-M2	512	64	37	3	640	-240	490
	6T-HVT-M2	512	64	25	3	550	-240	550
16KB	6T-LVT-M1	512	256	26	4	640	0	640
	6T-HVT-M1	512	256	16	2	550	0	550
	6T-LVT-M2	512	256	40	8	640	-240	490
	6T-HVT-M2	512	256	30	6	550	-240	550

(4%) performance penalty for arrays larger (smaller) than 1KB, with the maximum performance penalty of 12% for the 4KB array.

Because of 20× smaller leakage power of 6T-HVT SRAM compared with 6T-LVT counterpart, energy consumptions of HVT-based arrays are significantly lower (cf. Figure 7(b)), especially for large arrays. Accordingly, due to significantly lower energy consumptions and because of at most 12% performance penalty, energy-delay product is improved in 6T-HVT-M2 arrays. More precisely, energy-delay product of 6T-HVT-M2 compared with 6T-LVT-M2 on average is 59% (14%) smaller for arrays larger (smaller) than 1KB, and 78% lower with 8% performance penalty for the 16KB array.

6. CONCLUSIONS

A device-circuit-architecture co-optimization framework is presented in this paper for minimizing the energy-delay product of SRAM arrays. The key idea is to adopt HVT FinFET devices for the SRAM cell, which significantly reduces the leakage power and enhances HSNM and RSNM. The side effect is lower ON current, resulting in lower read current and hence performance degradation. Accordingly, the performance degradation is mitigated by jointly optimizing voltage levels of assist circuits and key parameters of SRAM array. Different read and write assist techniques were evaluated, and analytical models for delay and energy consumption of SRAM array were proposed. By using the proposed optimization framework, for SRAM array capacities ranging from 1KB to 16KB, on average 59% lower energy-delay product with maximum 12% (and on average 9%) performance penalty is achieved.

7. ACKNOWLEDGMENTS

This work is supported in part by grants from the PERFECT program of the Defense Advanced Research Projects Agency, and the Software and Hardware Foundations of the National Science Foundation.

References

[1] International Technology Roadmap for Semiconductors (ITRS). [Online]. Available: http://www.itrs.net/.

[2] I. J. Chang, J.-J. Kim, S. P. Park, and K. Roy. A 32 kb 10T Sub-Threshold SRAM Array With Bit-Interleaving and Differential Read Scheme in 90 nm CMOS. *IEEE Journal of Solid-State Circuits*, 44(2):650–658, 2009.

[3] L. Chang *et al.*, Stable SRAM Cell Design for the 32 nm Node and Beyond. In *Symposium on VLSI Technology*, pages 128–129, 2005.

[4] S. Chen, Y. Wang, X. Lin, Q. Xie, and M. Pedram. Performance prediction for multiple-threshold 7nm-FinFET-based circuits operating in multiple voltage regimes using a cross-layer simulation framework. In *IEEE SOI-3D-Subthreshold Microelectronics Technology Unified Conference (S3S)*, Oct. 2014.

[5] Y.-H. Chen *et al.*, A 16nm 128Mb SRAM in high-k metal-gate FinFET technology with write-assist circuitry for low-VMIN applications. In *IEEE International Solid-State Circuits Conference (ISSCC)*, pages 238–239, Feb 2014.

[6] A. Garg and T.-H. Kim. Sram array structures for energy efficiency enhancement. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 60(6):351–355, 2013.

[7] S. M. Kang, Y. Leblebici, and C. Kim. *CMOS Digital Integrated Circuits: Analysis and Design*. McGraw-Hill, 2015.

[8] E. Karl *et al.*, A 0.6V 1.5GHz 84Mb SRAM Design in 14nm FinFET CMOS Technology. In *IEEE International Solid-State Circuits Conference (ISSCC)*, pages 1–3, Feb 2015.

[9] D. Lu, C.-H. Lin, A. Niknejad, and C. Hu. Compact Modeling of Variation in FinFET SRAM Cells. *IEEE Design & Test of Computers*, 27(2):44–50, 2010.

[10] S. Natarajan *et al.*, A 14nm Logic Technology Featuring 2nd-Generation FinFET, Air-Gapped Interconnects, Self-Aligned Double Patterning and a 0.0588μm^2 SRAM Cell Size. In *IEEE International Electron Devices Meeting (IEDM)*, pages 3.7.1–3.7.3, Dec 2014.

[11] E. Nowak *et al.*, Turning silicon on its edge [double gate cmos/finfet technology]. *IEEE Circuits and Devices Magazine*, 20(1):20–31, 2004.

[12] E. Seevinck, F. List, and J. Lohstroh. Static-Noise Margin Analysis of MOS SRAM Cells. *IEEE Journal of Solid-State Circuits*, Oct. 1987.

[13] T. Song *et al.*, A 14nm FinFET 128Mb 6T SRAM with VMIN-enhancement techniques for low-power applications. In *IEEE International Solid-State Circuits Conference (ISSCC)*, Feb 2014.

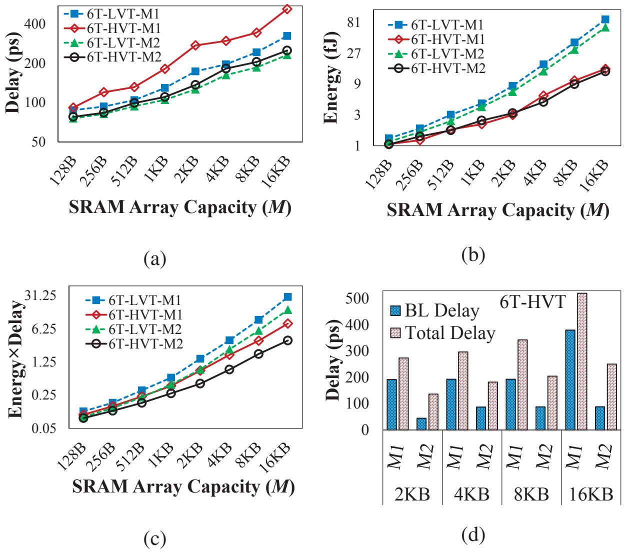
[14] S. Tang *et al.*, FinFET - A Quasi-Planar Double-Gate MOSFET. In *IEEE International Solid-State Circuits Conference (ISSCC)*, 2001.

[15] X. Wang, A. Brown, B. Cheng, and A. Asenov. Statistical variability and reliability in nanoscale finfets. In *IEEE International Electron Devices Meeting (IEDM)*, pages 5.4.1–5.4.4, Dec 2011.

[16] M. Yabuuchi *et al.*, 16 nm FinFET High-k/Metal-gate 256-kbit 6T SRAM macros with Wordline Overdriven Assist. In *IEEE International Electron Devices Meeting (IEDM)*, pages 3.3.1–3.3.3, Dec 2014.

[17] M. Yabuuchi, Y. Tsukamoto, M. Morimoto, M. Tanaka, and K. Nii. 20nm High-Density Single-Port and Dual-Port SRAMs with Wordline-Voltage-Adjustment System for Read/Write Assists. In *IEEE International Solid-State Circuits Conference (ISSCC)*, pages 234–235, 2014.

[18] B. Zimmer *et al.*, SRAM Assist Techniques for Operation in a Wide Voltage Range in 28-nm CMOS. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 59(12):853–857, 2012.



(a)

(b)

(c)

(d)

Figure 7. Simulation results: (a) Delay, (b) energy consumption, and (c) energy-delay product of different SRAM arrays. (d) BL delay vs. total delay in 6T-HVT-M1 and 6T-HVT-M2 arrays. Vertical axes in (a)-(c) are in logarithmic (base 2) scale.