Energy-Efficient Cache Memories using a Dual-V_t 4T SRAM Cell with Read-Assist Techniques

Alireza Shafaei and Massoud Pedram
Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089
{shafaeib, pedram}@usc.edu

***Abstract*—In order to improve the energy-efficiency of cache memories, this paper presents a static random access memory (SRAM) cell composed of four transistors using dual-V_t FinFET devices. The proposed 4T SRAM cell is designed by (i) removing pull-down transistors of the standard 6T SRAM, and (ii) using low-leakage high-V_t devices for pull-up transistors and fast low-V_t devices for access transistors. This dual-V_t design simultaneously improves hold and write characteristics, but results in a destructive read operation. Accordingly, read-assist techniques are employed to ensure a non-destructive and robust read operation. A selective row address decoder is also proposed to prevent the undesired write operation in half-selected cells. The 4T SRAM cell compared with the all-single-fin 6T counterpart has a 25% smaller layout area with an aspect ratio closer to one. Furthermore, using 7nm FinFET devices with a nominal supply voltage of 0.45V, the 4T SRAM cell achieves 3.5× lower cell leakage power. Because of these features, the energy consumption of a 32KB L1 (256KB L2) cache memory using 4T SRAM cell compared with its 6T counterpart is reduced by 18% (2×), with 35% (19%) higher cache access frequency.**

I. Introduction

The layout area of a *static random access memory* (SRAM) cell plays an important role in the characteristics of on-chip cache memories. Indeed, reducing the area footprint of the SRAM cell increases the memory density (i.e., the number of bits stored per unit area). At the same time, smaller SRAM cells tend to have shorter wordlines (WLs) and bitlines (BLs), which in turn decreases resistances and capacitances of these lines, and hence faster access latencies and lower access energy consumptions are achieved. Therefore, minimum-size transistors are preferred in SRAM cell designs. In particular, in FinFET technologies, the ideal case is to adopt single-fin devices for all SRAM transistors.

The standard SRAM cell, as shown in Figure 1(a), is composed of six transistors: four transistors (including two pull-up and two pull-down transistors) form two cross-coupled inverters which statically store data, along with two access transistors used for reading from and writing into the memory cell. Read and write operations share access transistors. Hence, for bitlines that are precharged high, the following requirements should be satisfied in order to ensure the proper operation of the 6T SRAM cell. (i) The *read stability* requirement: during a read operation, access transistors should be weaker than pull-down transistors such that access transistors cannot flip (destroy) the stored bit. (ii) The *write-ability* requirement: for a successful write operation, access transistors should be able to change the stored bit, and thus, access transistors should be stronger than pull-up transistors during the write operation.

A major challenge for advanced technology nodes is the increased effect of process variations. This is caused by (i) extremely small geometries where even small deviations may significantly change device properties, and (ii) reduced power supply voltage, V_{dd}, levels which narrow the difference between V_{dd} and the transistor threshold voltage, V_t. Sizing up transistors in the 6T SRAM, or using more robust cells such as the 8T SRAM [1] are effective in mitigating effects of process variations, but both approaches increase the cell area. Accordingly, the all-single-fin 6T SRAM cell equipped with assist techniques has gained attention recently [2], [3], [4]. However, such an SRAM cell still suffers from high leakage power consumption.

In order to further reduce the layout area and leakage power of the 6T SRAM cell, this paper presents a 4T SRAM cell using dual-V_t FinFET devices. The proposed 4T SRAM cell is designed by (i) removing pull-down transistors of the standard 6T SRAM cell, and (ii) using extremely low-leakage ultra-high-V_t (UVT) devices for pull-up transistors and fast low-V_t (LVT) devices for access transistors. This dual-V_t design is essential for the high stability of the hold operation, and is also helpful in improving the write characteristics. However, since access transistors are significantly stronger than pull-up transistors, the cell content is destroyed after a read operation.

For a non-destructive read operation, we take advantage of read-assist techniques. Specifically, we simultaneously apply both wordline underdrive (so as to weaken access transistors) and V_{dd} boost (in order to strengthen pull-up transistors) techniques to achieve a robust and fast read operation. Furthermore, when a cell is accessed, other cells in the same row that share the same WL may be subject to an undesired write operation (this is called the *half-select disturbance*). To resolve this potential serious error, we propose a selective row address decoder which only enables the WL of accessed cells.

The proposed SRAM cell is evaluated using FinFET devices with a physical gate length of 7nm and nominal supply voltage of 0.45V [5]. Monte Carlo simulations are also performed to ensure that noise margins under process variations meet high-yield requirements. Furthermore, FinCACTI tool [6] is used to assess FinFET-based cache memories. The 4T SRAM cell compared with the all-single-fin 6T counterpart has a 25% smaller layout area with an aspect ratio closer to one, and using 7nm FinFET devices under 0.45V, achieves 3.5× lower cell leakage power. Because of these features, the energy consumption of a 32KB L1 (256KB L2) cache memory using 4T SRAM compared with its 6T counterpart is reduced by 18% (2×), with 35% (19%) higher cache access frequency.

The rest of the paper is organized as follows. The proposed dual-V_t 4T SRAM cell is introduced in Section II. Read-assist techniques and the selective row address decoder are explained in Section III. Simulation results are presented in Section IV. Finally, Section V concludes the paper.

II. Proposed dual-V_t 4T SRAM Cell

The proposed 4T SRAM cell is shown in Figure 1(b). What makes our cell different from prior work (e.g., [7], [8]) is its

978-3-9815370-6-2/DATE16/

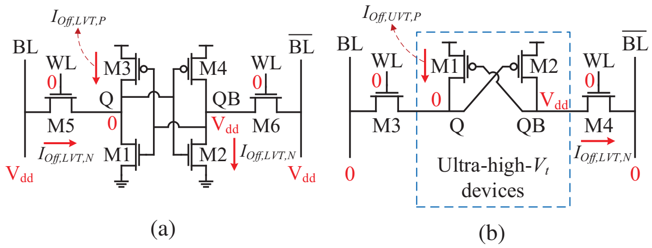
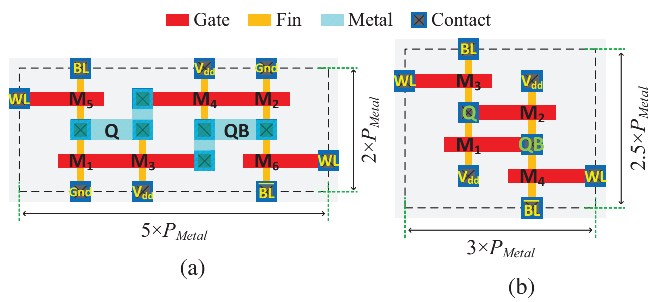
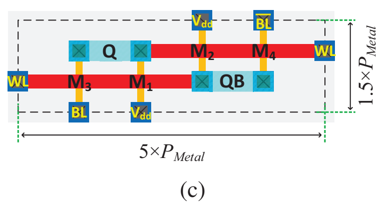


Fig. 1. (a) Standard 6T and (b) the proposed dual-V_t 4T SRAM cells. In our proposed 4T SRAM, access transistors are made of fast LVT devices, whereas pull-up transistors are made of low-leakage UVT devices. Voltage level of each signal and subthreshold leakage paths (arrows) are also shown for an idle SRAM cell storing ‘0’.





Cell	Area $\times(P_{Metal})^2$	AR
(a)	10	0.4
(b)	7.5	0.83
(c)	7.5	0.3

Fig. 2. (a) Layout of 6T SRAM. (b) Our proposed layout, and (c) the layout from [8] for 4T SRAM cell. Area and aspect ratio (AR=height/width) of each layout is reported in the bottom-right table. P_{Metal} denotes the metal pitch.

dual-V_t design which is important for the high stability of hold operation, improving write operation, and reducing the leakage power. Details of this 4T SRAM cell are presented next.

A. *Cell Layout*

Layout of the 6T SRAM cell[1] is shown in Figure 2(a), which is drawn based on the Intel 14nm SRAM cell layout [4]. For 4T SRAM cell, a layout from [8] and our proposed layout are shown in Figure 2(c) and Figure 2(b), respectively. Width and height of each layout is calculated based on the value of the metal-1 pitch, P_{Metal}. Accordingly, while the layout area of the 6T SRAM is $10 \cdot (P_{Metal})^2$, both layouts of the 4T SRAM have an area equal to $7.5 \cdot (P_{Metal})^2$, resulting in 25% smaller area footprint. Another key advantage of our proposed layout for 4T SRAM cell, compared with that of [8], is the aspect ratio which is closer to one. Hence, our proposed layout is closer to a square.

[1]In this paper, 6T refers to an all-single-fin standard 6T SRAM cell.

B. *Hold Operation*

The proposed 4T cell is a semi-static memory. This is because during the hold operation, and depending on the cell content, one storage node is *statically* connected to V_{dd} through one of the pull-up transistors, whereas the other node floats and acts as a *dynamic* storage node. The dynamic node should be kept discharged during the idle mode in order to make sure that data is properly retained. For this purpose, bitlines, BL and $\overline{\text{BL}}$, are pulled to Gnd. On the other hand, by assigning LVT devices to access transistors and high-V_t (HVT) devices to pull-up transistors, access transistors have a higher leakage current than pull-up transistors. Therefore, access transistors are able to keep the dynamic node discharged during idle mode.

Using high-leakage LVT devices for access transistors and low-leakage HVT devices for pull-up transistors are important for the hold operation of the proposed cell. However, in order to ensure the high stability of the hold operation in the presence of process variations and noises, the dynamic storage node should be kept completely discharged. The turned-off pull-up transistor tries to store charge on the dynamic node through its leakage current. To prevent this undesirable process, leakage current of pull-up transistors should be significantly reduced. This is achieved by adopting UVT devices, which have extremely higher threshold voltages compared with the nominal device.

UVT FinFET Devices: By engineering the work function of the gate material, we are able to aggressively increase the V_t of FinFET devices [9], [10]. In other words, the work function of the FinFET device is tuned during the device optimization in order to achieve UVT devices. An important feature of this approach is that it does not impact the cell layout area.

Leakage Power: Leakage current paths for 6T and 4T SRAM cells storing bit ‘0’ are shown in Figure 1(a) and Figure 1(b), respectively. Due to the symmetric structure of both cells, same leakage paths, but through symmetric transistors, exist when the cell stores bit ‘1’. Therefore, the following discussion is valid for both cases.

Since 6T SRAM is made of LVT devices to meet frequency requirements, the leakage power of the 6T SRAM cell is given by

$$\begin{aligned} P_{leak}(6\text{T}) &= V_{dd} \cdot (I_{Off,LVT,N} \\ &\quad + I_{Off,LVT,N} + I_{Off,LVT,P}) \\ &= (2+r) \cdot V_{dd} \cdot I_{Off,LVT,N}, \end{aligned} \tag{1}$$

where $I_{Off,LVT,N}$ ($I_{Off,LVT,P}$) denotes the OFF current of a single-fin NFET (PFET) LVT device, and r is the PFET to NFET OFF current ratio of the LVT device. On the other hand, the internal cell leakage of the 4T SRAM, because of using UVT devices, is negligible. As a result, the leakage power of the proposed 4T SRAM cell can be calculated as

$$\begin{aligned} P_{leak}(4\text{T}) &= V_{dd} \cdot (I_{Off,LVT,N} + I_{Off,UVT,P}) \\ &\approx V_{dd} \cdot I_{Off,LVT,N}, \end{aligned} \tag{2}$$

where $I_{Off,UVT,P}$ denotes the OFF current of a single-fin PFET UVT device. According to (1) and (2), and depending on the value of r (which is technology dependent), the leakage power of the proposed dual-V_t 4T SRAM cell is at least 2× smaller than that of its 6T counterpart.

C. Write Operation

In order to enhance the write-ability of the proposed SRAM cell, the ON current of the access transistor should be higher than that of the pull-up transistor. In our 4T SRAM cell, access transistors are made of fast LVT devices, whereas very slow UVT devices are used for pull-up transistors. Therefore, this dual-V_t design is not only necessary for ensuring the robustness of the hold operation, but is also important for satisfying the write-ability requirement. The lack of pull-down transistors also helps in improving the write operation. The reason is because the access transistor, when turned on, can easily write into the dynamic storage node. All these features point to a reliable and fast write operation.

III. Challenges and Solutions

Two main challenges of the proposed 4T SRAM cell along with their solutions are discussed in this section.

A. Read Operation using Assist Techniques

Read operation in the 6T SRAM cell is initiated by precharging bitlines to V_{dd}. WL is then activated, and assuming that the cell stores '0', i.e., $V(Q) = 0$, BL is discharged while $\overline{\text{BL}}$ remains unchanged. Also, since pull-down transistors should be stronger than access transistors during the read operation, the content of the cell will not be destroyed. In our 4T SRAM cell, if BL and $\overline{\text{BL}}$ are initially precharged to V_{dd}, when WL is turned on, access transistor can easily write '1' into the dynamic node. This puts the SRAM cell into a metastable state. Hence, read operation in our proposed 4T SRAM cell is initiated by *predischarging* bitlines to 0.

After predischarging bitlines and activating the WL, both dynamic node and the corresponding bitline are '0', and hence, nothing happens at this side. The voltage level of the bitline connected to the static node is increased, which is then sensed by the sense amplifier. However, as shown in Figure 3, pull-up transistor tries to write '1' into the static node, whereas access transistor is trying to write '0'. In our 4T SRAM, since access transistor is stronger than the pull-up transistor, access transistor wins the fight and flips the cell content. Thus, while the dual-V_t design is critical for the hold operation and improving write characteristics, it results in a destructive read operation.

To achieve a non-destructive read operation, we should weaken the access transistor and/or strengthen the pull-up transistor during the read operation. To do this, we take advantage of assist techniques. Common read-assist techniques include [11]:

- Worldline 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.
- V_{dd} boost (VDDB): Supply voltage level of the cell, denoted by V_{DDC}, is increased above V_{dd}, which subsequently increases the ON current of the pull-up transistor.
- Negative Gnd: Applying a negative voltage to the source terminal of the pull-down transistor results in a drain-to-source voltage greater than V_{dd}, and thus increases the ON current through the pull-down transistor.
- Partial bitline precharge (predischarge): Bitlines are precharged (predischarged) to a voltage level lower than V_{dd} (higher than 0) in order to weaken access transistors.

The negative Gnd technique does not apply to our 4T SRAM cell, and the partial bitline, especially compared with

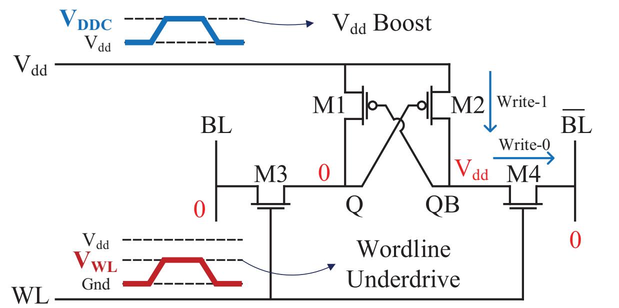


Fig. 3. Read-assist techniques for the proposed 4T SRAM cell. During the read operation, wordline underdrive weakens access transistors, whereas V_{dd} boost strengthens pull-up transistors. Voltage level of each signal for an SRAM cell storing '0' during read operation is also shown.

the WLUD, is not an effective way to weaken the access transistor. Therefore, these two read-assist techniques are not explored in this paper. On the other hand, the WLUD technique, due to weakening the access transistor which subsequently reduces the read current, increases the read latency. Therefore, we simultaneously apply WLUD and VDDB techniques (cf. Figure 3) in order to find a combination that minimizes the energy-delay product of the read operation while read *static noise margin* (SNM) is above a certain level.

The all-single-fin 6T SRAM also requires assist techniques to achieve a non-destructive read operation. The negative Gnd technique needs regulating a negative voltage which is a difficult task [11]. Accordingly, similar to the 4T SRAM, both WLUD and VDDB techniques are applied to the 6T SRAM. Moreover, write operation in the 6T SRAM, especially when process variations are considered, requires assist techniques. Wordline overdrive (WLOD) is adopted for this purpose.

B. Selective Row Address Decoder

One of the main issues of semi-static memories is the low stability of *half-selected cells* (HSCs) [7], [8]. An HSC refers to an idle cell in which the value of a control signal has been changed because of a read or write operation on a different cell. Such cells can be categorized into column or row HSCs which are illustrated in Figure 4 and are explained next.

Column Half-Selected Cells: When a cell is accessed for a write operation, the voltage level of one of the bitlines changes. This change is also observed by all other cells that share the same bitline. Accordingly, a column HSC refers to an idle cell in which one of the bitlines has been flipped because of a write operation on a cell in the same column (cf. cell (c) in Figure 4). This may cause a problem for the dynamic node. However, since access transistors of column HSCs are turned off, and because write operation is very fast in our proposed 4T SRAM cell, the value of the dynamic node cannot be destroyed. Moreover, based on our simulations, the voltage level drop of the dynamic node under column half-select disturbance and for a time period 1000 times longer than the write access latency is less than 1%.

Row Half-Selected Cells: A row HSC refers to an idle cell in which the WL becomes activated due to a read or write operation on a cell in the same row (cf. cell (b) in Figure 4). Since BL and $\overline{\text{BL}}$ are both 0 during the idle mode, activating

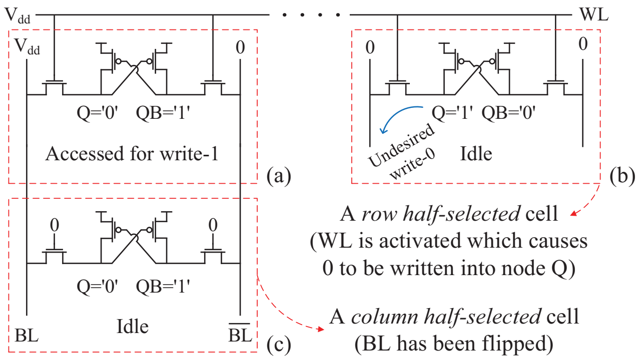


Fig. 4. Half-selected cells: Cell (a) has been accessed for a write-1 operation. Accordingly, (b) and (c) become row and column half-selected cells, respectively.

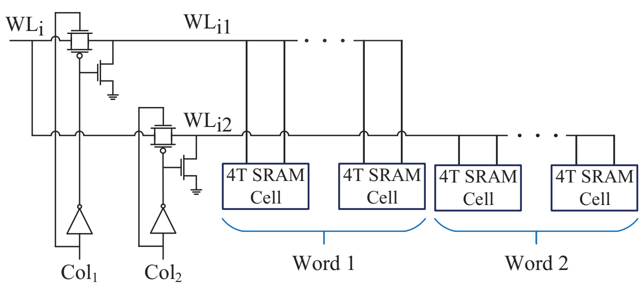


Fig. 5. Selective row decoder: Col$_j$ (WL$_i$) is the output of the column (row) decoder, which denotes word j (row i). WL$_{ij}$ is the wordline of cells in row i and word j. A word is a group of cells that are read or written in the same cycle.

access transistors causes a write-0 into the static node of row HSCs, which in turn puts these cells into a metastable state.

To avoid this undesired write in row HSCs, we modify the row address decoder such that only the WL of accessed cells is activated. The circuit of the proposed selective row address decoder is shown in Figure 5, which also receives inputs from the column decoder. A word in Figure 5 refers to a group of cells which will be read or written in the same cycle. If the SRAM array has R bits (i.e., SRAM cells) in each row, and w bits are read or written in each cycle, then $n_w = \lceil R/w \rceil$ words exist per row. For $n_w = 1$, there is no row HSC in the memory, and thus, the selective row decoder is not needed.

IV. Simulation Results

A. Simulation Setup

FinFET Devices: Simulation results are obtained using FinFET devices with a physical gate length of 7nm and a nominal V_{dd} of 0.45V [5]. The adopted 7nm FinFET library includes LVT, HVT, and UVT devices. For the proposed 4T SRAM cell, we use LVT and UVT devices for access and pull-up transistors, respectively. The 4T SRAM cell is compared with the all-single-fin 6T SRAM cell. LVT devices are used for all transistors in the 6T SRAM cell.

SRAM Cell Characteristics: For each SRAM cell, leakage power consumption as well as hold, read, and write noise margins are measured using HSpice simulations. Leakage power is the total power dissipation during the idle mode. Hold and read SNMs are measured based on butterfly curves

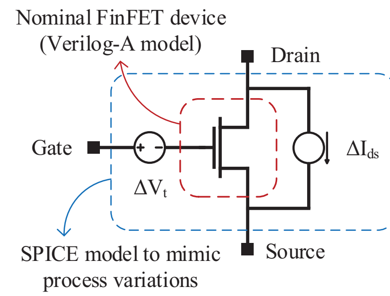


Fig. 6. Modeling the effect of process variations in look-up table-based Verilog-A models [14].

[12]. Write margin is defined as the difference between the V_{dd} and the minimum wordline voltage that is needed to flip the cell content [13]. For assist techniques, we allow voltage levels to increase or decrease up to 50% from the nominal V_{dd}. Moreover, to ensure that SRAM cells satisfy the high-yield requirement, we perform Monte Carlo simulations with 2000 samples. For this purpose, the mean, μ, and standard deviation, σ, of hold, read, and write noise margins are measured. A high-yield SRAM cell requires a $\mu/\sigma \geq 6$ for each operation.

Process Variations: The adopted 7nm FinFET devices are lookup table-based Verilog-A models, which are generated for nominal conditions. Variations of fin length, fin width, work function, and doping concentration are then modeled by variations on the threshold voltage and drain-to-source current. More precisely, each transistor of the SRAM cell is modeled as the circuit shown in Figure 6 [14]. In other words, for each transistor (i) a voltage source is inserted on the gate terminal in order to inject variations on the threshold voltage, and (ii) a current source is added between drain and source terminals in order to introduce variations on the saturation current. Following [13], the V_t variation from the nominal value for transistor M$_i$ during j^{th} Monte Carlo run, denoted by $\Delta V_{t,ij}$, is calculated as follows:

$$\Delta V_{t,ij} = \Delta V_{t,j}^{global} + \Delta V_{t,ij}^{local}, \tag{3}$$

where $\Delta V_{t,j}^{global}$ captures the global variations and is the same value for all SRAM transistors in each Monte Carlo run, whereas local variations are captured by $\Delta V_{t,ij}^{local}$ which is a unique value for each SRAM transistor in each Monte Carlo run. Based on TCAD simulations, we use 8% global and 5% local variations. Similarly, the drain-to-source current variation, $\Delta I_{ds,ij}$, is measured using the following equation:

$$\Delta I_{ds,ij} = \Delta I_{ds,j}^{global} + \Delta I_{ds,ij}^{local}. \tag{4}$$

Cache Memories: We use the FinCACTI tool [6] to derive the characteristics of FinFET-based cache memories. Support for assist techniques and other considerations for the 4T SRAM cell are also added to this tool. For this paper, we adopt L1 data (L1-D) and L1 instruction (L1-I) cache memories, both 16KB and 2-way set-associative, resulting in a 32KB L1 cache, and a 256KB, 8-way set-associative L2 cache memory. Total power consumption, P_{total}, and energy consumption per cycle, E_{cycle}, of each cache memory are calculated as follows:

$$\rho = \frac{\text{number of cache accesses}}{\text{total number of instructions}} \tag{5}$$

TABLE I. . Noise margins of 6T and 4T SRAM cells under 7nm FinFET devices and V_{dd} = 450mV.

Operation	SRAM Cell	Assist Techniques V_{DDC} ($\times V_{dd}$)	V_{WL} ($\times V_{dd}$)	Noise Margin from Monte Carlo Simulations μ (mV)	σ (mV)	μ/σ
Hold	6T	N/A	N/A	168.01	15.35	10.95
	4T	N/A	N/A	190.07	18.63	10.21
Write	6T	1	1	121.23	23.40	5.18
	6T	1	1.1 (∗)	166.84	22.40	7.45
	6T	1	1.5 (∗)	346.64	22.07	15.71
	4T	1	1	335.50	28.99	11.57
Read	6T	1	1	31.22	25.89	1.21
	6T	1.5 (⋄)	0.9 (⋆)	173.42	14.90	11.64
	4T	1	1	0 †	—	—
	4T	1.5 (⋄)	0.58 (⋆)	170.04	26.53	6.41

† Content of the proposed 4T SRAM cell without assist techniques is immediately destroyed after a read operation.
Write-assist technique: (∗) Wordline Overdrive
Read-assist techniques: (⋄) V_{dd} Boost, (⋆) Wordline Underdrive

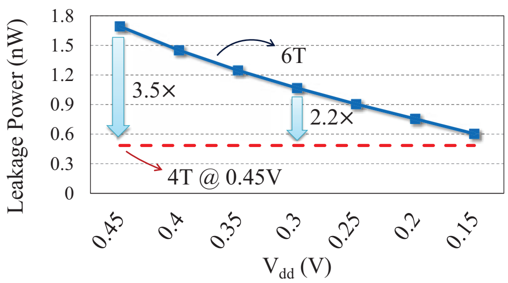
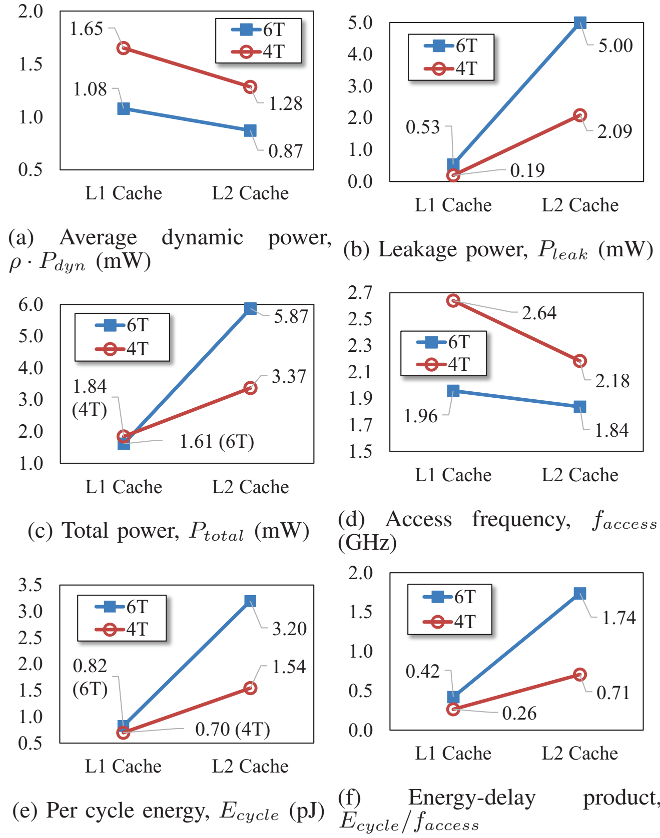


Fig. 7. Leakage power results of 6T (shown over different V_{dd} values) and 4T (shown only for V_{dd}=0.45V) SRAM cells.



(a) Average dynamic power, $\rho \cdot P_{dyn}$ (mW)
(b) Leakage power, P_{leak} (mW)
(c) Total power, P_{total} (mW)
(d) Access frequency, f_{access} (GHz)
(e) Per cycle energy, E_{cycle} (pJ)
(f) Energy-delay product, E_{cycle}/f_{access}

Fig. 8. Results of L1 (32KB, 2-way) and L2 (256KB, 8-way) cache memories using 6T and 4T SRAM cells.

$$P_{total} = \rho \cdot P_{dyn} + P_{leak} \quad (6)$$

$$E_{cycle} = P_{total}/f_{access} \quad (7)$$

where ρ, P_{dyn}, P_{leak} and f_{access} denote the access ratio, dynamic power, leakage power, and access frequency of the cache memory, respectively. Based on our simulations using the Sniper tool [15] on SPLASH-2 [16] and PARSEC [17] benchmarks, average cache ratios of L1-I, L1-D, and L2 are 12%, 34%, and 2%, respectively. L1-I and L1-D results are summed up and shown as L1 in this section.

B. Cell-Level Results

Table I reports hold, read, and write noise margins of 6T and 4T SRAM cells under 0.45V operation. Both SRAM cells have a very robust hold operation. However, the proposed 4T SRAM because of having a dynamic node needs a higher hold SNM to satisfy the high-yield requirement, which is achieved by adopting extremely low-leakage UVT devices for pull-up transistors. Furthermore, 4T SRAM without assist techniques has a very robust write operation. For 6T cell, we use WLOD write-assist technique, which increases V_{WL} in order to make access transistors stronger than pull-up transistors. Strengthening access transistors during write operation also increases the write current, and hence a faster write operation is obtained. While 10% increase in V_{WL} is sufficient for the 6T SRAM to meet the high-yield requirement, 50% increase results in a very robust and fast write operation. For cache-level results, WLOD with 10% increase is assumed for the 6T cell.

Without read-assist techniques, 6T has a very poor read stability, and 4T immediately loses its data after a read operation. Accordingly, as we mentioned earlier, both WLUD and VDDB read-assist techniques are applied. More specifically, we sweep V_{DDC} from V_{dd} to $1.5 \times V_{dd}$, and V_{WL} from V_{dd} to $0.5 \times V_{dd}$, and report a (V_{DDC}, V_{WL}) pair that minimizes the energy-delay product of read access and has a high read stability. Based on our simulations, we derived $(V_{DDC} = 1.5 \times V_{dd}, V_{WL} = 0.9 \times V_{dd})$ and $(V_{DDC} = 1.5 \times V_{dd}, V_{WL} = 0.58 \times V_{dd})$ for 6T and 4T SRAM cells, respectively. Using these values, both cells meet the high-yield requirement for read operation. However, 6T has 80% higher μ/σ than that of the 4T SRAM cell.

Under V_{dd} = 0.45V, leakage power of 6T SRAM is 1.692nW, whereas that of the proposed 4T SRAM is 0.485nW, resulting in 3.5× lower leakage power. Figure 7 shows the leakage power of 6T SRAM cell for different V_{dd} values, compared with the leakage power of 4T SRAM at the nominal V_{dd}. Even at 0.15V, the leakage power of 6T is 25% higher than that of 4T at 0.45V. This shows the effectiveness of the proposed 4T SRAM cell design in reducing the leakage power which is especially crucial for high-capacity cache memories.

C. Cache-Level Results

Results of the 32KB L1 and 256KB L2 cache memories using 6T and 4T SRAM cells are shown in Figure 8. Cell width of the proposed 4T SRAM is 40% smaller than that of the 6T counterpart, which causes a significant reduction in the wordline delay. On the other hand, 4T SRAM, because of larger

TABLE II. Predicted values of metal pitch (P_{Metal}) for future FinFET technologies based on Intel 22nm and 14nm values.

	Scaling Factor [18]	**22nm Node** [18]	**14nm Node** [18]	**10nm Node**	**7nm Node**
P_{Metal}	0.78	90nm	70nm	55nm	43nm

TABLE III. . Area components of a 256×256 memory subarray made of 6T and 4T SRAM cells.

	Width (μm)	**Height** (μm)	**Area** (μm^2)
6T SRAM	62.17	21.81	1,355.55
4T SRAM	37.64	27.26	1,025.82
Improvement (%)	39%	-25%	24%

cell height and more importantly due to lower read current (as a result of using lower V_{WL}), has a higher bitline delay. Overall, since the wordline delay is the main component of cache access latency, 35% and 19% higher access frequencies for L1 and L2 caches, respectively, are achieved by using the proposed 4T SRAM cell.

Higher cache access frequency yields to higher dynamic power. However, 3.5× lower cell leakage power of the 4T SRAM significantly decreases the cache leakage power consumption. As a result, the energy consumption per cycle and energy-delay product of L1 (L2) using the proposed 4T SRAM compared with the 6T counterpart are reduced by 18% (2×) and 59% (2.5×), respectively. Low activity which results in long idle cycles, and large number of SRAM cells make the leakage power of L2 the main component of the total cache power consumption. Therefore, the effect of leakage power reduction by using the 4T SRAM is more noticeable in L2, and hence, higher improvements in the energy consumption and energy-delay product are observed.

Cache Area: Area components of an 8KB memory subarray made of 6T and 4T SRAM cells are measured by FinCACTI, and reported in Table III. The value of P_{Metal} for 7nm FinFET technology, which is needed for SRAM cell area calculations, is obtained from the scaling factor of Intel 14nm FinFET with respect to Intel 22nm FinFET [18] (cf. Table II). The memory subarray includes a 256×256 array of SRAM cells along with peripheral circuits such as row and column address decoders, wordline drivers, bitline prechargers, column multiplexers, and sense amplifiers. The smaller cell width of 4T SRAM compared with its 6T counterpart not only reduces the width of the SRAM array, but also decreases the transistor sizing of wordline drivers (since the WL capacitance has been reduced) which compensates for the area overhead of the selective row address decoder. By using the proposed 4T SRAM, the area of the aforesaid memory subarray is decreased by 24%.

V. Conclusion

We presented a dual-V_t 4T SRAM cell, and showed its robust operation under a 7nm FinFET technology operating at 0.45V. The key idea is to use extremely low-leakage UVT devices for pull-up transistors, and fast LVT devices for access transistors. This dual-V_t design is essential for the high stability of hold operation, and is also helpful in improving the write characteristics. Non-destructive read operation is then ensured by using read-assist techniques, and the undesired write operation in row half-selected cells is prevented by a selective row address decoder. Because of the 25% smaller layout area, and 3.5× lower cell leakage power of 4T SRAM compared with the all-single-fin 6T counterpart, higher energy-efficient cache memories are gained by using the proposed 4T SRAM cell. This 4T SRAM design because of its semi-static nature may not satisfy the high-yield requirements under low voltage operation, which is needed to further reduce the leakage power. Using error-correcting codes to relax the yield requirements of the SRAM cell may be useful for this purpose.

Acknowledgment

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] L. Chang *et al.*, "Stable SRAM Cell Design for the 32 nm Node and Beyond," in *Symposium on VLSI Technology*, June 2005, pp. 128–129.
[2] 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, pp. 232–233.
[3] 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)*, Feb 2014, pp. 238–239.
[4] 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)*, Feb 2015, pp. 1–3.
[5] S. Chen *et al.*, "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.
[6] A. Shafaei *et al.*, "FinCACTI: Architectural Analysis and Modeling of Caches with Deeply-Scaled FinFET Devices," in *IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, July 2014, pp. 290–295.
[7] Z. Guo *et al.*, "FinFET-based SRAM Design," in *International Symposium on Low Power Electronics and Design (ISLPED)*, Aug 2005, pp. 2–7.
[8] M.-L. Fan *et al.*, "Comparison of 4T and 6T FinFET SRAM Cells for Subthreshold Operation Considering Variability—A Model-Based Approach," *IEEE Transactions on Electron Devices*, vol. 58, no. 3, pp. 609–616, March 2011.
[9] Y.-K. Choi *et al.*, "FinFET Process Refinements for Improved Mobility and Gate Work Function Engineering," in *International Electron Devices Meeting (IEDM)*, Dec 2002, pp. 259–262.
[10] A. Veloso *et al.*, "Highly Scalable Effective Work Function Engineering Approach for Multi-VT Modulation of Planar and FinFET-based RMG High-K Last Devices for (Sub-)22nm Nodes," in *Symposium on VLSI Technology (VLSIT)*, June 2013, pp. T194–T195.
[11] 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*, vol. 59, no. 12, pp. 853–857, 2012.
[12] E. Seevinck, F. List, and J. Lohstroh, "Static-Noise Margin Analysis of MOS SRAM Cells," *IEEE Journal of Solid-State Circuits*, vol. 22, no. 5, pp. 748–754, Oct 1987.
[13] D. Lu *et al.*, "Compact Modeling of Variation in FinFET SRAM Cells," *IEEE Design & Test of Computers*, vol. 27, no. 2, pp. 44–50, 2010.
[14] P. Royer and M. Lopez-Vallejo, "Using pMOS Pass-Gates to Boost SRAM Performance by Exploiting Strain Effects in Sub-20-nm FinFET Technologies," *IEEE Transactions on Nanotechnology*, vol. 13, no. 6, pp. 1226–1233, Nov 2014.
[15] T. Carlson, W. Heirman, and L. Eeckhout, "Sniper: Exploring the level of abstraction for scalable and accurate parallel multi-core simulation," in *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, Nov 2011, pp. 1–12.
[16] S. C. Woo *et al.*, "The SPLASH-2 Programs: Characterization and Methodological Considerations," in *Annual International Symposium on Computer Architecture (ISCA)*, 1995, pp. 24–36.
[17] C. Bienia and K. Li, "PARSEC 2.0: A New Benchmark Suite for Chip-Multiprocessors," in *5th Annual Workshop on Modeling, Benchmarking and Simulation*, June 2009.
[18] 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)*, Dec 2014, pp. 3.7.1–3.7.3.