Full-Swing Local Bitline SRAM Architecture Based on the 22-nm FinFET Technology for Low-Voltage Operation

Kyoman Kang, Hanwool Jeong, *Student Member, IEEE*, Younghwi Yang, Juhyun Park, Kiryong Kim, and Seong-Ook Jung, *Senior Member, IEEE*

***Abstract*— The previously proposed average-8T static random access memory (SRAM) has a competitive area and does not require a write-back scheme. In the case of an average-8T SRAM architecture, a full-swing local bitline (BL) that is connected to the gate of the read buffer can be achieved with a boosted wordline (WL) voltage. However, in the case of an average-8T SRAM based on an advanced technology, such as a 22-nm FinFET technology, where the variation in threshold voltage is large, the boosted WL voltage cannot be used, because it degrades the read stability of the SRAM. Thus, a full-swing local BL cannot be achieved, and the gate of the read buffer cannot be driven by the full supply voltage (V_{DD}), resulting in a considerably large read delay. To overcome the above disadvantage, in this paper, a differential SRAM architecture with a full-swing local BL is proposed. In the proposed SRAM architecture, full swing of the local BL is ensured by the use of cross-coupled pMOSs, and the gate of the read buffer is driven by a full V_{DD}, without the need for the boosted WL voltage. Various configurations of the proposed SRAM architecture, which stores multiple bits, are analyzed in terms of the minimum operating voltage and area per bit. The proposed SRAM that stores four bits in one block can achieve a minimum voltage of 0.42 V and a read delay that is 62.6 times lesser than that of the average-8T SRAM based on the 22-nm FinFET technology.**

***Index Terms*— Bit-interleaving, FinFET, low-voltage operation, static random access memory (SRAM).**

I. Introduction

In RECENT years, with the widespread use of battery-powered applications, such as handheld smart devices and implantable medical devices, low-power operation has become a critical issue associated with the system-on-chip (SoC) design. A low-power SoC can be effectively realized with a low-power static random access memory (SRAM) because the SRAM critically affects the total power of the SoC, owing to the fact that it occupies a large portion of the area of the SoC. Further, power reduction can be effectively achieved by decreasing the operating voltage because of the quadratic dependence of power on the operating voltage [1]. However, at a low operating voltage, the adverse effect of the variation in threshold voltage (V_{th}) becomes more significant. It should be noted that an SRAM cell is highly susceptible to variations in V_{th}, given that it is designed with small transistors for high-density integration. Furthermore, in the case of a conventional 6T SRAM cell, a tradeoff exists between the read stability and the write ability, owing to which, it is very challenging to simultaneously achieve sufficient read stability and write ability in a low-voltage region.

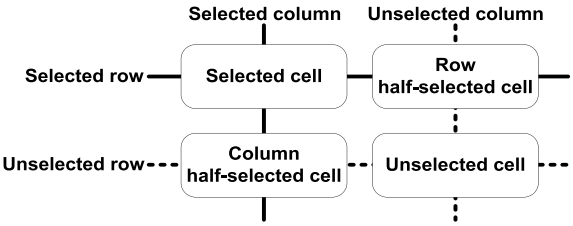


Fig. 1. Selected, half-selected, and unselected cells in a bit-interleaved SRAM array.

Several SRAM cell alternatives with a decoupled read port have been proposed for a low-voltage operation [2]–[6]. The advantage of adding a decoupled read port is that it eliminates the tradeoff between the read stability and the write ability in the SRAM array to which the bit-interleaving is not applied; thus, the read stability and write ability can be optimized separately, facilitating a low-voltage operation. An SRAM cell is also susceptible to soft errors induced by α-particles; to address these errors, it is necessary for the SRAM array to exhibit bit-interleaving [7]. Fig. 1 shows a bit-interleaved SRAM array architecture. In a bit-interleaved SRAM array, the selected cells are the SRAM cells targeted for the read or write operation. The row half-selected cells are the SRAM cells located on the selected row and the unselected column, whereas the column half-selected cells are the SRAM cells located on the unselected row and the selected column. During the write operation, the row half-selected cells are disturbed because of the selection of the word-line (WL) of the row half-selected cells. Thus, the stability

Manuscript received February 15, 2015; revised May 8, 2015; accepted June 9, 2015. Date of publication August 10, 2015; date of current version March 18, 2016. This work was supported in part by the IT Research and Development Program through the Ministry of Knowledge Economy/Korea Evaluation Institute of Technology under Grant 10039174 and in part by the Technology Development of 22-nm Level Foundry Device and PDK.

The authors are with the School of Electrical and Electronic Engineering, Yonsei University, Seoul 120-749, Korea (e-mail: kangkm@yonsei.ac.kr; hanwool87@yonsei.ac.kr; yyhgood@yonsei.ac.kr; pnaynh@yonsei.ac.kr; deservebest@yonsei.ac.kr; sjung@yonsei.ac.kr).

Digital Object Identifier 10.1109/TVLSI.2015.2450500

of the row half-selected cells should also be considered in the SRAM design. This consideration of the stability of row half-selected cells is referred to as a half-select issue. Unfortunately, the aforementioned alternatives do not address the half-select issue without a write-back scheme. The write-back scheme, in particular, ensures the stability of the row half-selected cells by reading the stored data in one cell and then writing back the same data into the same cell; however, this scheme requires additional power, delay, and area. To address the half-select issue without the write-back scheme, a 10T SRAM cell exhibiting a cross-point structure was proposed [8]. This 10T SRAM cell includes vertical and horizontal WLs, both of which need to be selected to access the storage nodes. During the write operation, both the WLs are selected only in the selected cell, owing to which the half-select issue is eliminated. On the other hand, a disadvantage of the 10T SRAM is that it suffers from a large area overhead to accommodate the additional transistors in its architecture. To address this disadvantage, an average-8T SRAM architecture based on a 130-nm technology was proposed; this SRAM architecture is a good alternative to the previously proposed SRAMs in that it addresses the half-select issue with no write-back scheme, and it exhibits a competitive area [9]. However, a drawback of this 8T SRAM is that its read delay increases considerably when it is fabricated using a more advanced technology such as a 22-nm FinFET technology that involves a large variation in V_{th}, because a tradeoff between the read stability and the read delay exists.

In this paper, the drawback of the average-8T SRAM architecture based on an advanced technology is analyzed, and a suitable SRAM architecture that overcomes this drawback is proposed. It should be noted that the proposed differential SRAM architecture can resolve the half-select issue without the need for a write-back scheme, and it exhibits a competitive area; it also exhibits a full-swing local bitline (BL) that enables a considerably smaller read delay than that of an average-8T SRAM architecture.

The remainder of this paper is organized as follows. Section II describes and analyzes the average-8T SRAM architecture. Section III elucidates the structure and operation of the proposed SRAM architecture. In Section IV, the proposed SRAM architecture is simulated and compared with the average-8T SRAM architecture, based on the 22-nm FinFET technology. Section V summarizes the conclusion of this paper.

II. Average-8T SRAM Architecture

Fig. 2 shows the average-8T SRAM architecture and its operational waveform. A block that stores four bits consists of four pairs of cross-coupled inverters, pass gate transistors (PGL1∼4 and PGR1∼4), block mask transistors (MASK1 and MASK2), write access transistors (WR1 and WR2), and read buffers (RD1∼4). A stacked nMOS structure is used as a read buffer to reduce the read BL (RBL and RBLB) leakage. It is important to note that the block select signal (BLK) and WLs (WL1∼4) are row-based signals, whereas the RBLs and write BLs (WBL and WBLB) are column-based signals. During the hold state, the WLs are held at 0 V to isolate the

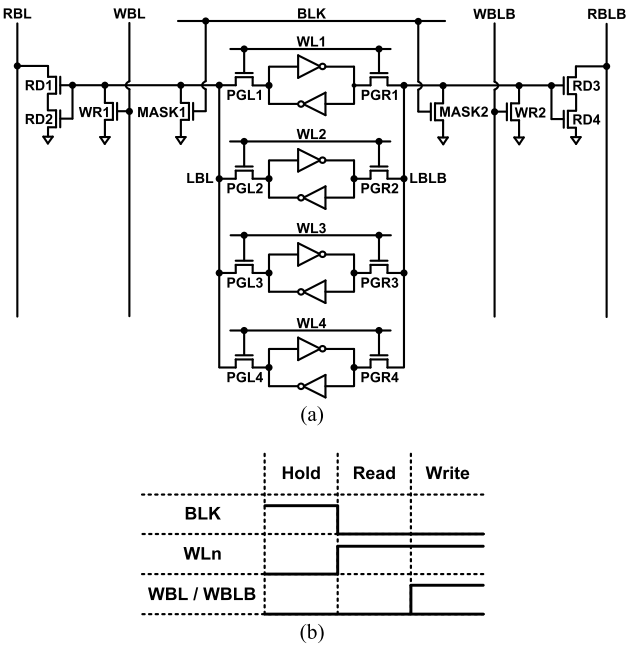


Fig. 2. (a) Average-8T SRAM architecture and (b) its operational waveform.

storage node from the local BLs (LBL and LBLB), and BLK is held at supply voltage (V_{DD}) to discharge the LBLs and to turn OFF the read buffers. The WBLs are held at 0 V, and the RBLs are precharged to V_{DD}.

For the read operation, BLK of the selected block is forced to remain at 0 V to turn OFF the block mask transistors, and the selected WL is enabled to turn ON the pass gate transistors. Thus, the stored data in the selected cell are transferred to the LBLs through the pass gate transistors, and one of RBLs is discharged on the basis of the stored data. In the column half-selected block, which is located on the selected column and the unselected row, during the read operation, the WLs are held at 0 V to turn OFF the pass gate transistors, and the BLK is forced to V_{DD} to turn ON the block mask transistors. In the column half-selected block, which is located on the selected column and the unselected row, during the read operation, the WLs are held at 0 V to turn OFF the pass gate transistors, and the BLK is forced to V_{DD} to turn ON the block mask transistors. In the column half-selected block, the two stacked nMOSs in the read buffer are turned OFF, irrespective of the stored data, because the LBLs are discharged to 0 V by the block mask transistors. Thus, the RBL leakage is said to be data-independent. This property of the RBL leakage facilitates the existence of a large number of cells per RBL. During the read operation, the 1 storage node is disturbed because it is connected to the predischarged LBL via the pass gate transistor. However, a sufficiently high read stability for a low-voltage operation can be achieved because the LBL exhibits a small capacitance. It is important to accurately control the signal timing to achieve robust read stability. If both the WL and the BLK are high, simultaneously, the 1 storage node and the source of the block mask transistor (V_{SS}) will be connected, and the stored data can be flipped. Thus, the WL should rise after the fall of the BLK is completed.

For the write operation, the BLK of the selected block is forced to remain at 0 V, the selected WL is enabled, and one of the WBLs is forced to switch to V_{DD} to turn ON the write access transistor on the basis of the write data. While the read operation is a differential operation, the write operation is a single-ended operation that discharges only the 1 storage node through the pass gate and write access transistors. It is important to note that during the write operation, the row half-selected block is in the same condition as it was in the read operation. The stability of the row half-selected block is ensured owing to the fact that the LBL exhibits a small capacitance, eliminating the need for a write-back scheme. Furthermore, the area overhead is reduced because the additional transistors such as the block mask and the write access transistors, as well as the read buffers, are shared among the four cells.

However, the average-8T SRAM suffers from a drawback associated with the read stability and read delay. During the read operation, the stored 1 cannot be completely delivered to the LBL owing to the V_{th} drop in the nMOS pass gate transistor. Thus, the gate of the read buffer is driven by $V_{DD}-V_{th}$, which results in a considerably large read delay in a low-voltage region. The average-8T SRAM architecture, which was proposed based on the 130-nm technology, alleviates this drawback by boosting the WL voltage, causing the read stability to degrade. Despite the fact that the read stability is degraded by the boosted WL voltage, a sufficiently high read stability is ensured owing to the small capacitance at the LBL. Thus, both a high read stability and a small read delay are ensured in the case of the 130-nm technology. However, in the case of an advanced technology, where the variation in V_{th} is large, it is difficult to achieve a sufficiently high read stability with the boosted WL voltage in spite of the small capacitance at the LBL. Instead, the suppressed WL voltage is required to achieve a sufficiently high read stability in the case of the advanced technology. However, the suppressed WL voltage intensifies the effect of the V_{th} drop and considerably worsens the read delay. In light of this discussion, it can be concluded that the average-8T SRAM architecture exhibits a tradeoff between the read stability and the read delay, and it cannot simultaneously ensure a high read stability and a small read delay in the low-voltage region when it is fabricated using an advanced technology.

III. Proposed Differential SRAM Architecture

The proposed differential SRAM stores multiple bits in one block, as in the case of an average-8T SRAM. Fig. 3 shows the architecture of the proposed SRAM that stores i bits in one block. The minimum operating voltage and area per bit of the proposed SRAM depend on the number of bits in one block. A configuration that stores four bits in one block is selected as the basic configuration by considering the balance between the minimum operating voltage and the area per bit, which will be described in Section IV. The basic configuration of the proposed SRAM includes four cross-coupled inverter pairs, pass gate transistors (PGL1∼4 and PGR1∼4), block mask transistors (MASK1 and MASK2), write access transistors (WR1 and WR2), read buffers (RD1 and RD2),

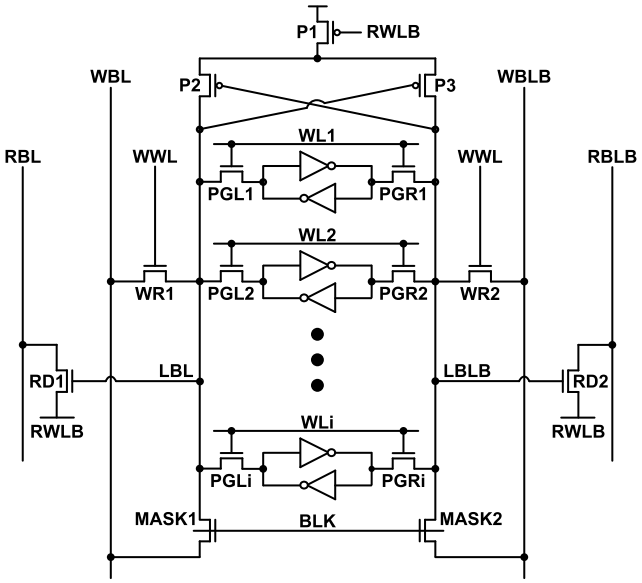


Fig. 3. Proposed SRAM architecture that stores i bits in one block.

a head switch (P1), and cross-coupled pMOSs (P2 and P3). The head switch and cross-coupled pMOSs of the proposed SRAM are notable differences from the average-8T SRAM. WLs (WL1∼4), the block select signal (BLK), and the read WL (RWLB) are row-based signals, whereas the write WL (WWL), write BLs (WBL and WBLB), and read BLs (RBL and RBLB) are column-based signals. During the hold state, WLs, WWL, and WBLs are held at 0 V. BLK is held at V_{DD} to connect the WBLs and the LBLs, so that the LBLs are discharged to 0 V and the read buffers are turned OFF. Further, the RWLB is also held at V_{DD} to turn OFF the head switch and to eliminate the RBL leakage current.

A. Read Operation

The read operation of the proposed SRAM architecture is described in Fig. 4(a). This operation is performed in two phases. During the first phase, BLK of the selected block is forced to remain at 0 V, and the selected WL is enabled. On the basis of the stored data, although the voltage of the LBL that is connected to the 1 storage node becomes high, its value cannot be as high as that of the full V_{DD} because of the V_{th} drop through the pass gate transistor, and the voltage of the other LBL remains low. The read operation in the first phase is similar to that of the average-8T SRAM, except that the RBL is not discharged because the RWLB is high in the first phase. With the assertion of WL, although the 1 storage node is disturbed, the read disturbance is small because of the small capacitance at the LBL. This smaller read disturbance makes the proposed SRAM be able to operate in significantly lower operating voltage compared with 6T SRAM cell. The second phase starts with the falling of the RWLB. The assertion of the RWLB enables not only the discharge of the RBL but also the feedback of cross-coupled pMOSs. Positive feedback of the cross-coupled pMOSs increases the LBL to the value of the full V_{DD}, owing to which the LBL can achieve a full swing, and the gate of the read buffer

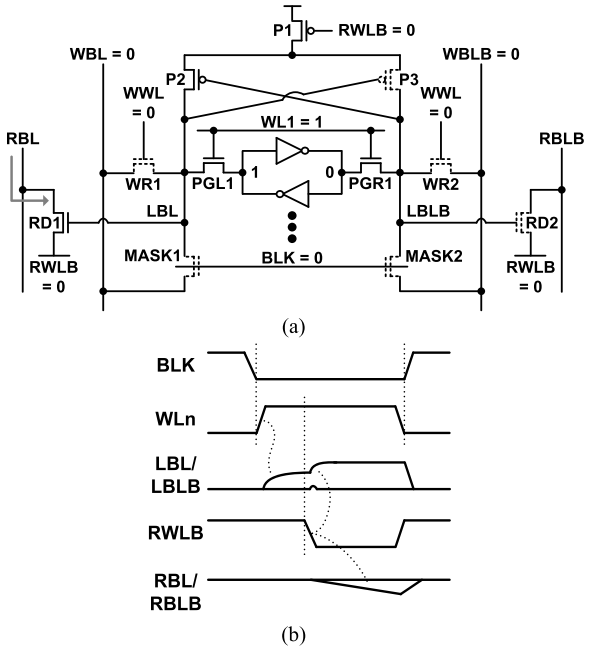


Fig. 4. (a) Read operation and (b) read operational waveform of proposed SRAM architecture.

is driven by the full V_{DD}, without the need for a boosted WL voltage. Thus, in the case of the proposed SRAM based on an advanced technology, the suppressed WL voltage can be used to enhance the read stability, without degrading the read delay. In other words, the advantage of the proposed SRAM architecture is that it eliminates the tradeoff between the read stability and the read delay. The suppressed WL voltage is used to enhance the read stability, and the full-swing LBL minimizes the read delay. In the case of the average-8T SRAM architecture, the read buffer consists of two stacked nMOSs that reduce the RBL leakage. On the other hand, in the proposed SRAM architecture, a single nMOS is used as the read buffer to increase the read current, and the buffer foot is attached to reduce the RBL leakage. The column half-selected block is in the hold state in which the read buffers are turned OFF, so that the RBL discharge is not affected by the column half-selected block.

As in the case of the average-8T SRAM architecture, in the proposed SRAM architecture, it is essential to carefully control the signal timing to avoid the data from flipping, as shown in Fig. 4(b). When both the BLK and the WL are simultaneously high, the 1 storage node and the WBL that is held at 0 V are connected, causing the data to flip. Thus, the WL should increase after the fall in the BLK is completed. For the robustness of the positive feedback of the cross-coupled pMOSs in spite of the variation in V_{th}, a sufficient LBL development is required. Thus, the RWLB should be asserted with a sufficiently large timing margin after the WL is asserted; this requires an additional timing overhead, which does not exist in the average-8T SRAM. An important point to note here is that the total read delay of the proposed SRAM based on an advanced technology such as the 22-nm FinFET technology is considerably lesser than that of the average-8T SRAM because the LBL of the average-8T SRAM architecture

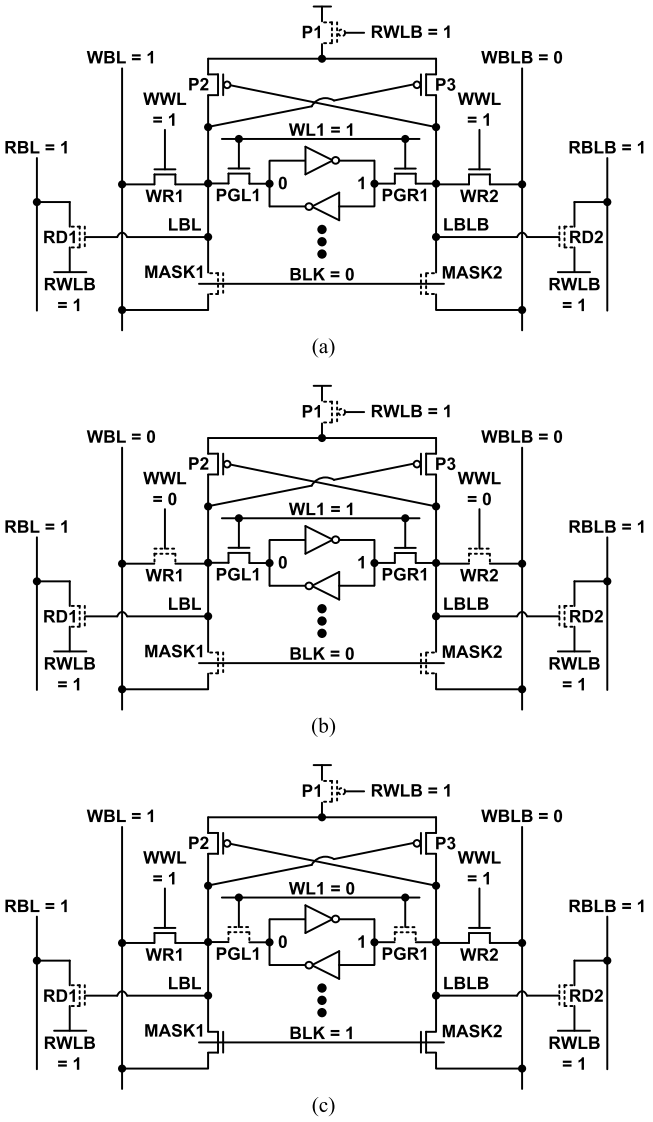


Fig. 5. (a) Selected, (b) row half-selected, and (c) column half-selected blocks of proposed SRAM architecture during write operation.

cannot achieve full swing, and its read buffer consists of two stacked nMOSs.

B. Write Operation

The write operation of the proposed SRAM architecture is shown in Fig. 5(a). As shown in this figure, BLK of the selected block is forced to remain at 0 V, and the selected WL is enabled. Further, the WWL is forced to remain at V_{DD} so that the write access transistors are turned ON, and the WBLs are forced to remain at a certain voltage level on the basis of the write data. Both the storage nodes are connected to the WBLs through pass gate transistors and write access transistors. Thus, the write operation is differential, and the write ability of the proposed SRAM is better than that of the average-8T SRAM, whose write operation is single-ended.

The row half-selected block shown in Fig. 5(b) is in the same condition as it was in the read operation, except that the RWLB is high. Although the storage nodes of the row half-selected blocks are disturbed during the write operation,

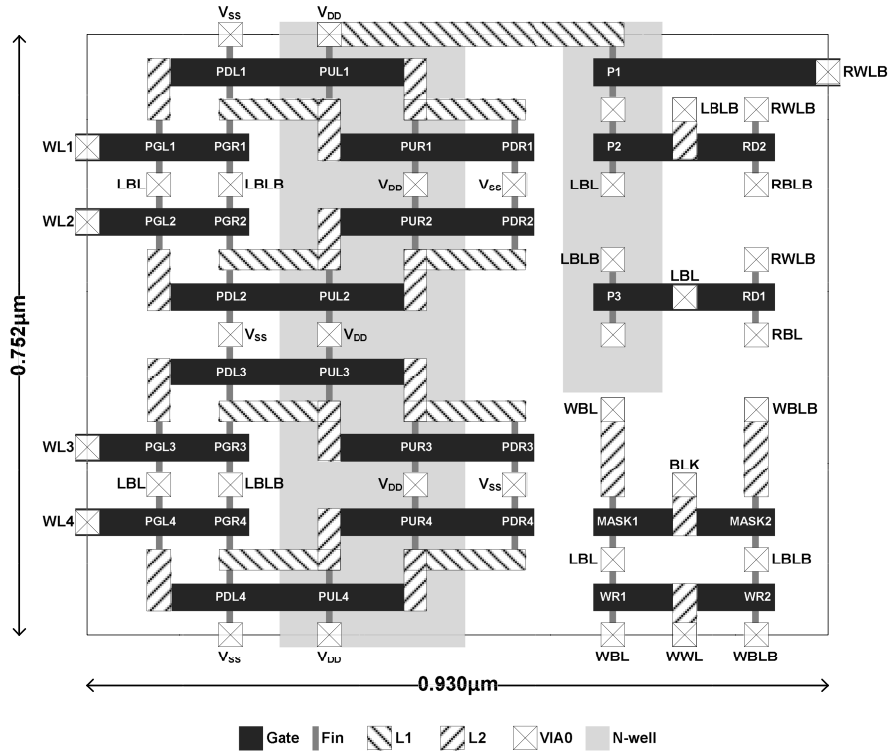


Fig. 6. Layout of the proposed SRAM architecture based on the 22-nm FinFET technology.

the disturbance is small because of the small capacitance at the LBL. Thus, the stability of the row half-selected block is ensured without the need for a write-back scheme. Further, in the case of the average-8T SRAM architecture, during the write operation, the RBLs in the unselected columns are unnecessarily discharged because the row half-selected block is in the same condition as it was in the read operation, resulting in the consumption of a large amount of dynamic power. In this regard, the advantage of the proposed SRAM is that it eliminates the unnecessary RBL discharge by using a buffer foot that is forced to high during the write operation.

Unlike in the case of the average-8T SRAM architecture, in the proposed SRAM architecture, the sources of the block mask transistors are connected to the WBLs, not to V_{SS}, to eliminate a dc current path in the column half-selected block shown in Fig. 5(c). When the sources of the block mask transistors are connected to V_{SS}, the dc current flows from a high WBL to V_{SS} through the write access and block mask transistors in all the column half-selected blocks where BLK and WWL are high, resulting in the consumption of a large amount of static power during the write operation. The dc current path in the column half-selected block is eliminated by connecting the sources of the block mask transistors to the WBLs.

C. Layout and Area

In the layout of the average-8T SRAM architecture, the four cells are located at each corner of the layout, and not in one line, and the additional transistors are placed at the center of the layout. This folded column layout configuration contributes toward a decrease in the RBL capacitance of the average-8T SRAM architecture. However, this folded column configuration is not applied to the proposed SRAM architecture to reduce its area even though the RBL capacitance increases. Fig. 6 shows the layout of the basic configuration of the proposed SRAM architecture based on the 22-nm FinFET technology, designed with the smallest transistors. The local interconnect in the middle of line is employed to reduce the number of metal layers [10]. V_{DD} and V_{SS} are routed in metal 1; the LBLs are routed in metal 2; the BLK and RWLB are routed in metal 3; the RBLs, WBLs, and WWL are routed in metal 4; and the WLs are routed in metal 5. Fig. 7 shows a comparison between the areas per bit of the previous and proposed SRAMs. With an increase in the number of bits in a block, the area per bit of the proposed

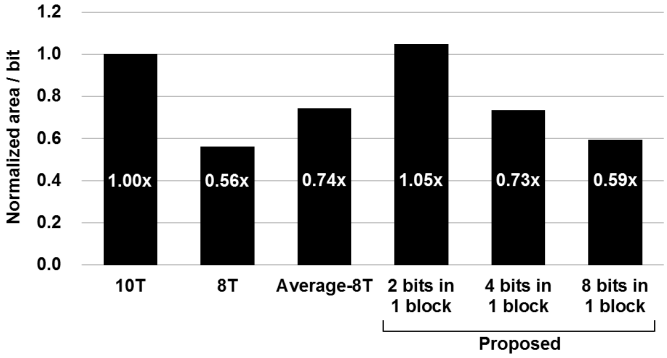


Fig. 7. Area comparison with previous SRAMs.

TABLE I
TECHNOLOGY PARAMETERS FOR $V_{DD} = 0.8$ V

Parameter	NMOS	PMOS
Gate length	34 nm	34 nm
Equivalent oxide thickness	0.9 nm	0.9 nm
Fin thickness	8 nm	8 nm
Fin height	34 nm	34 nm
On current	880 μA/μm	780 μA/μm
Off current	1 nA/μm	1 nA/μm
Sub-threshold swing	69 mV/dec	72 mV/dec
DIBL	46 mV/V	50 mV/V
Threshold voltage (V_{th})	230 mV (Sat.) 264 mV (Lin.)	245 mV (Sat.) 283 mV (Lin.)

V_{th} in the saturation and linear modes are measured as the V_{GS} when I_{DS} per effective width is 10^{-5} A/μm with $|V_{DS}| = V_{DD}$ and $|V_{DS}| = 0.05$V [12].

SRAM architecture decreases because a greater number of cells share the additional transistors. The area of the proposed SRAM that stores two bits in one block is greater than that of the 10T SRAM cell. However, the area per bit of the proposed SRAM that stores four bits in one block is 27% lesser than that of the 10T SRAM cell and slightly lesser than that of the average-8T SRAM, despite the fact that a greater number of transistors are attributed to the use of the head switch and cross-coupled pMOSs.

IV. SIMULATION RESULTS AND COMPARISON

The architecture of the proposed SRAM is verified by the HSPICE Monte Carlo simulation using a BSIM-CMG FinFET model [11]. The characteristics of this model are fitted to those of a commercial low-power device based on the 22-nm FinFET technology [12]. Table I lists the technology parameters. For a statistical analysis, it is assumed that the variation in $V_{\rm th}$ of each transistor follows a Gaussian distribution, whose standard deviation ($\sigma_{V_{\rm th}}$) is expressed by:

$$\sigma_{V_{\rm th}} = \frac{A_{Vt}}{\sqrt{\text{Length} \times \text{Width}}} \tag{1}$$

where an A_{Vt} of 1.5 mV $\cdot$ μm is assumed according to [13]. It is crucial to select suitable metrics to precisely measure the read stability and write ability. The read stability of the proposed SRAM is affected by the capacitance at the LBL. Therefore, it is important to measure the dynamic read noise margin. Transient simulation is performed with noise voltages inserted between the storage nodes, and the minimum noise voltage that causes the data to flip is considered as the dynamic read noise margin. Furthermore, both the drain and gate capacitances of the transistors and the wire capacitance are considered for the capacitance at the LBL [14]. Fig. 8 shows the dynamic read noise margins of the proposed SRAM at different process, voltage, and temperature (PVT) corners. The dynamic read noise margin is larger in the SF (slow nMOS and fast pMOS) corner than in the TT (typical nMOS, typical pMOS) or FS (fast nMOS, slow pMOS) corner. This is because the 1 storage node is disturbed when it is connected to the predischarged LBL via the nMOS pass gate transistors during read operation. Thus, the fast pMOS in the cross-coupled inverters and the slow nMOS pass gate transistors enhance the read noise margin. Finally, a BL write trip voltage is considered as a write-ability metric [15].

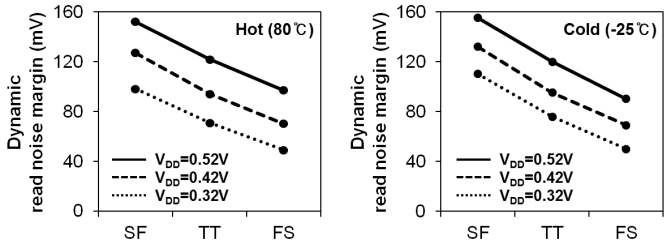


Fig. 8. Dynamic read noise margins of proposed SRAM at different PVT corners.

A. Minimum Operating Voltage

An assist circuit contributes toward a decrease in the minimum operating voltage of an SRAM. In this regard, several types of assist circuits exist. For example, boosted cell supply voltage ($V_{\rm CELL}$), negative $V_{\rm SS}$, suppressed WL, and suppressed BL read assist circuits improve the read stability of an SRAM. On the other hand, suppressed $V_{\rm CELL}$, boosted $V_{\rm SS}$, boosted WL, and negative WBL write assist circuits enhance the write ability of an SRAM. The boosted $V_{\rm CELL}$, negative $V_{\rm SS}$, and suppressed BL read assist circuits are column-based, while the suppressed WL read assist circuit is row-based. If the boosted $V_{\rm CELL}$ or negative $V_{\rm SS}$ read assist circuit is used in the proposed SRAM architecture, during the write operation, read assist will be applied to all the unselected columns because the row half-selected blocks are in the same condition as they were in the read operation, resulting in the consumption of a large amount of power. The suppressed BL read assist circuit cannot be applied to the proposed SRAM architecture, because it is essential for the LBLs of the proposed SRAM to be predischarged to 0 V to turn OFF the read buffer in the column half-selected block. Thus, the suppressed WL read assist circuit is applied to the proposed SRAM architecture.

Likewise, the suppressed $V_{\rm CELL}$, boosted $V_{\rm SS}$, and negative WBL write assist circuits are column-based, and the boosted WL write assist circuit is row-based. If the suppressed $V_{\rm CELL}$ or boosted $V_{\rm SS}$ write assist circuit is applied, not only the wire capacitance but also the storage node capacitances in all the column half-selected cells will be charged and discharged, resulting in the consumption of a large amount of dynamic power. The boosted WL write assist circuit cannot be applied to the proposed SRAM because the WL is already used as the suppressed WL read assist circuit. Thus, the negative WBL write assist circuit is considered to be the most suitable write assist circuit for the proposed SRAM architecture.

While the WBLs in the proposed SRAM architecture are connected to the storage nodes via write access and pass gate transistors, the WBLs in the average-8T SRAM architecture are connected to the gates of the write access transistors. The function of the WBLs in the average-8T SRAM architecture is to turn ON or OFF the write access transistors, not driving the storage nodes. To apply the negative BL write assist in the average-8T SRAM architecture, $V_{\rm SS}$ at the sources of the WR1 and WR2 are replaced with other signal lines,

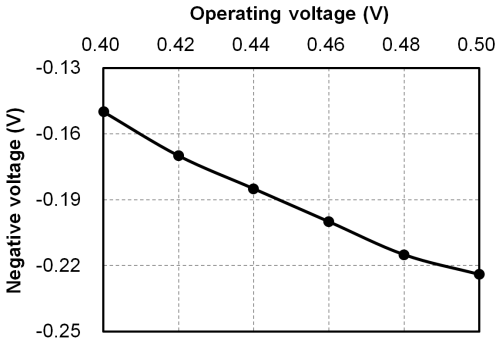


Fig. 9. Maximum negative voltage for write assist circuit while 5σ hold stability yield is ensured in the proposed and average-8T SRAM in common.

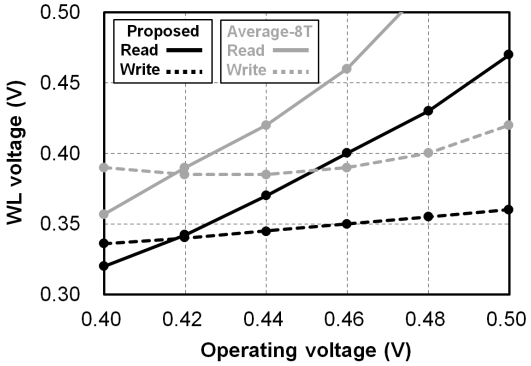


Fig. 10. Maximum and minimum WL voltages for 5σ read stability and write ability yields, respectively, in proposed and average-8T SRAMs.

WVSS and WVSSB, respectively, which can be controlled separately with VSS and can be decreased to the negative voltage for write assist. It has the same effect as the negative WBL write assist circuit in the proposed SRAM structure.

When the negative BL write assist is applied, the negative voltage which WBL can have is limited. As soon as the WBL voltage is lowered to a negative, the leakage currents are supplied from the column half-selected cells to the WBL. The magnitude of these leakage components is exponentially increased as the WBL voltage is decreased to a negative value, and increased leakage current prevents WBL from being further lowered. At the same time, the increased leakage current from the column half-selected cells degrades the hold stability of the column half-selected cells. Thus, the negative voltage for write assist is limited for the hold stability. Fig. 9 shows the maximum negative voltage of the write assist circuit based on various operating voltages, ensuring the 5σ hold stability yield of the column half-selected cell. In the cases of the proposed SRAM and average-8T SRAM, the maximum negative voltages are same because the effects of the negative voltage on the column half-selected cells are same in both architectures.

In the case of the average-8T and proposed SRAM architectures, the read stability is inversely proportional to the WL voltage, and the write ability is directly proportional to the WL voltage. Fig. 10 shows the maximum WL voltage for achieving a 5σ read stability yield (solid line) and the minimum WL voltage for achieving a 5σ write ability yield (dotted line) when the negative voltage of the write assist circuit is applied, as shown in Fig. 9, for the

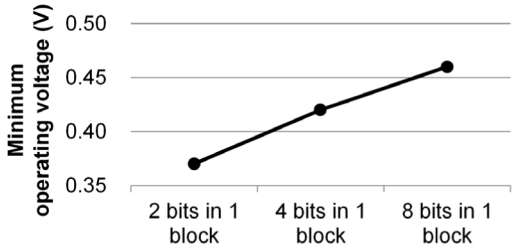


Fig. 11. Minimum operating voltages of proposed SRAM architecture based on various configurations.

average-8T SRAM and proposed SRAM architectures that store four bits in one block. It is crucial that the SRAM be operated in the region below the solid line and above the dotted line to ensure its robust operation, because both the read stability yield and write ability yield are satisfied ($>5\sigma$) in the overlapping region. At the same operating voltage, it is found that the read stability of the proposed SRAM is lesser than that of the average-8T SRAM because of the fact that the greater number of transistors connected to the LBL increase the LBL capacitance and degrade the read stability. In light of this fact, the maximum WL voltage for achieving the 5σ read stability should be decreased in the proposed SRAM. On the other hand, the write ability of the proposed SRAM is better than that of the average-8T SRAM because the write operation of the proposed SRAM is differential and that of the average-8T SRAM is single-ended. Thus, the minimum WL voltage for achieving the 5σ write ability can be lowered in the proposed SRAM. Consequently, the minimum operating voltage is almost the same, i.e., 0.42 V, in both these SRAM architectures.

Fig. 11 shows the minimum operating voltages of the proposed SRAM based on the various configurations. In the case of the proposed SRAM, the read stability improves with a decrease in the number of bits in one block, because of the smaller capacitance at the LBL. Thus, the minimum operating voltage decreases with the number of bits in one block. However, although the configuration that stores two bits in one block exhibits the lowest minimum operating voltage, it has a larger area per bit than the 10T SRAM cell. On the other hand, although the configuration that stores eight bits in one block has the smallest area per bit, it exhibits a somewhat higher minimum operating voltage than the other configurations. Given that the configuration that stores four bits in one block can operate in a low-voltage region with a smaller area per bit than the 10T SRAM cell, this configuration is selected as the basic configuration of the proposed SRAM architecture, which is used for the following simulations and comparisons.

B. Read Delay

A comparison is drawn between the read delays of the average-8T and the proposed SRAM architectures at a minimum operating voltage of 0.42 V, while using assist circuits for achieving 5σ read stability and write ability yields. In this paper, an SRAM array with 256 rows and 128 columns is assumed. The read delay is defined as the delay between the fall of the BLK to the development of

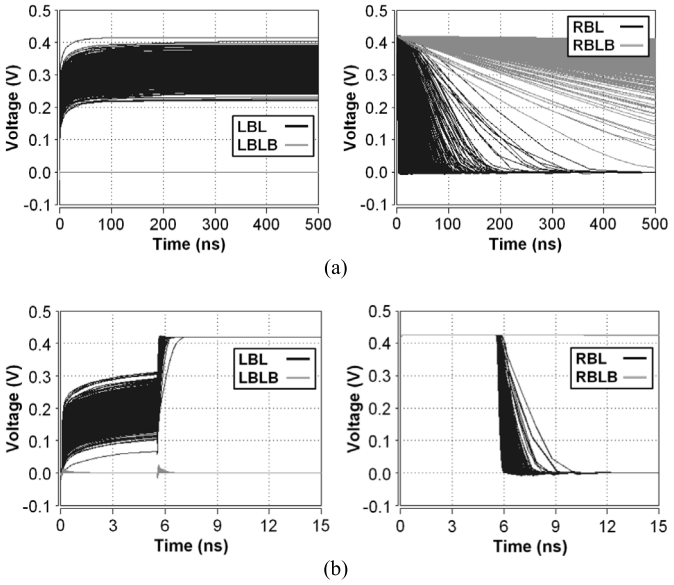


Fig. 12. Waveforms of LBLs and RBLs during read operation in (a) average-8T and (b) proposed SRAM architectures.

the 120-mV RBL. As mentioned in Section III, a suitable timing margin is required between the WL and the RWLB for obtaining robustness of the positive feedback of the cross-coupled pMOSs in the proposed SRAM architecture. This robustness of the positive feedback is disturbed when the strengths of the pass gate transistors are decreased or when the strength of the pMOS whose gate is connected to a high LBL is increased and the strength of the other pMOS is decreased with a variation in $V_{\rm th}$. From the simulation result, a timing margin of 5.5 ns is sufficient to achieve the 5σ yield of the feedback. Fig. 12 shows the waveforms of the LBLs and the RBLs during the read operation, obtained from the Monte Carlo simulation. As analyzed, in the case of the average-8T SRAM architecture, the LBL cannot attain the value of full $V_{\rm DD}$; thus, the gate of the read buffer cannot be driven by the full $V_{\rm DD}$. Moreover, the read buffer of the average-8T SRAM architecture consists of two stacked nMOSs, owing to which a large delay is caused in the development of the RBL in the low-voltage region. On the other hand, the full-swing LBL and the single nMOS read buffer of the proposed SRAM architecture speed up the development of the RBL. In the proposed and average-8T SRAMs, the 5σ worst read delays are 8.32 and 520.99 ns, respectively. Thus, the read delay of the proposed SRAM is 62.6 times lesser than that of the average-8T SRAM at a minimum operating voltage of 0.42 V.

C. Energy Consumption and Standby Power

The average energy consumption per operation is measured in a 256 rows × 128 columns SRAM array with 4-to-1 bit-interleaving at a minimum operating voltage of 0.42 V. Fig. 13 shows the simulated read and write energy consumptions. Despite the fact that the proposed SRAM exhibits a higher RBL capacitance than the average-8T SRAM, as mentioned in Section III, the proposed SRAM consumes a considerably smaller amount of read energy than that consumed by the average-8T SRAM because the exceedingly long

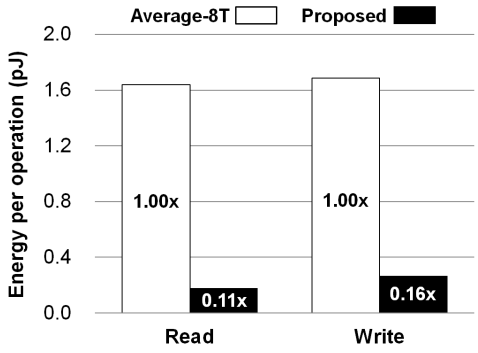


Fig. 13. Read and write operation energies in average-8T and proposed SRAMs.

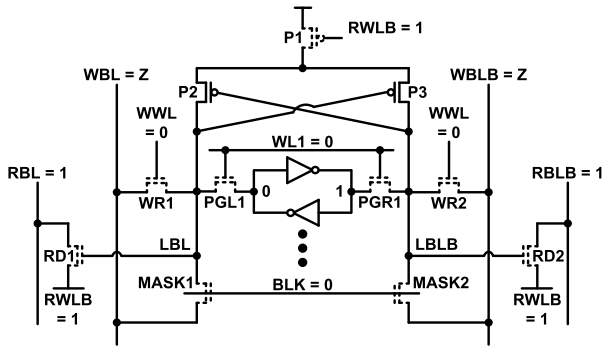


Fig. 14. Standby mode of proposed SRAM architecture.

read delay of the average-8T SRAM causes a large active leakage. Further, the write energy of the proposed SRAM is considerably lesser than that of the average-8T SRAM because the unnecessary RBL discharges in the unselected columns are eliminated in the proposed SRAM.

The standby power is measured at a minimum data retention voltage, which ensures the 5σ hold stability yield. The minimum data retention voltages of the average-8T and the proposed SRAMs are identical at 0.24 V. In the standby mode of the average-8T SRAM architecture, the BLK, WLs, and WBLs are held at 0 V, whereas the RBLs are set to a high-impedance mode. Fig. 14 shows the standby mode of the proposed SRAM architecture, where the BLK, WLs, and WWL are held at 0 V, and the RBLs and RWLB are held at $V_{\rm DD}$, whereas the WBLs are set to a high-impedance mode. Unlike in the case of the average-8T SRAM architecture, in the proposed SRAM architecture, although the cross-coupled pMOSs form additional leakage paths, the leakage paths through the write access transistors are eliminated. Consequently, from the result of the simulation, the average standby power of a 256 rows × 128 columns SRAM array is approximately identical at 1.13 μW for both the SRAM architectures.

V. Conclusion

An advantage of the average-8T SRAM architecture is that it does not require a write-back scheme for bit-interleaving, and it exhibits a competitive area. However, in the case of an average-8T SRAM architecture based on an advanced technology such as a 22-nm FinFET technology, full-swing LBL cannot be achieved owing to the tradeoff between the read

stability and the read delay. Thus, the gate of the read buffer cannot be driven by a full V_{DD}, resulting in a considerably large read delay in a low-voltage region. Further, the RBLs in the unselected columns are unnecessarily discharged during the write operation, resulting in the consumption of a large amount of dynamic power in the write operation. In the proposed differential SRAM, the tradeoff between the read stability and the read delay is eliminated. A full-swing LBL is achieved using cross-coupled pMOSs; thus, the gate of the read buffer is driven by a full V_{DD}, while a suppressed WL read assist circuit is applied to enhance read stability. Further, the single nMOS read buffer contributes toward improving the read delay. In addition, the unnecessary RBL discharge during the write operation is eliminated by using the read buffer with a buffer foot, resulting in the saving of power during the write operation. Consequently, it can be concluded that the proposed SRAM based on the 22-nm FinFET technology exhibits a considerably smaller read delay and consumes less energy with a slightly smaller area than the average-8T SRAM.

REFERENCES

[1] B. H. Calhoun, J. F. Ryan, S. Khanna, M. Putic, and J. Lach, "Flexible circuits and architectures for ultralow power," *Proc. IEEE*, vol. 98, no. 2, pp. 267–282, Feb. 2010.

[2] L. Chang *et al.*, "An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 956–963, Apr. 2008.

[3] N. Verma and A. P. Chandrakasan, "A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, Jan. 2008.

[4] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, "A 0.2 V, 480 kb subthreshold SRAM with 1 k cells per bitline for ultra-low-voltage computing," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 518–529, Feb. 2008.

[5] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, Mar. 2007.

[6] Q. Li, B. Wang, and T. T. Kim, "A 5.61 pJ, 16 kb 9T SRAM with single-ended equalized bitlines and fast local write-back for cell stability improvement," in *Proc. Eur. Solid-State Device Res. Conf.*, Sep. 2012, pp. 201–204.

[7] J. Maiz, S. Hareland, K. Zhang, and P. Armstrong, "Characterization of multi-bit soft error events in advanced SRAMs," in *IEDM Tech. Dig.*, Dec. 2003, pp. 21.4.1–21.4.4.

[8] 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 J. Solid-State Circuits*, vol. 44, no. 2, pp. 650–658, Feb. 2009.

[9] M. Khayatzadeh and Y. Lian, "Average-8T differential-sensing subthreshold SRAM with bit interleaving and 1k bits per bitline," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 22, no. 5, pp. 971–982, May 2014.

[10] K. Ronse *et al.*, "Opportunities and challenges in device scaling by the introduction of EUV lithography," in *IEDM Tech. Dig.*, Dec. 2012, pp. 18.5.1–18.5.4.

[11] (Jul. 2013). *BSIM-CMG 107.0.0 Multi-Gate MOSFET Compact Model*. [Online]. http://wwwdevice.eecs.berkeley.edu/bsim/?page=BSIMCMG_LR

[12] C. Auth *et al.*, "A 22 nm high performance and low-power CMOS technology featuring fully-depleted tri-gate transistors, self-aligned contacts and high density MIM capacitors," in *Proc. Symp. VLSI Technol.*, Jun. 2012, pp. 131–132.

[13] C.-H. Lin *et al.*, "Channel doping impact on FinFETs for 22 nm and beyond," in *Proc. Symp. VLSI Technol.*, Jun. 2012, pp. 15–16.

[14] D. Ingerly *et al.*, "Low-k interconnect stack with metal-insulator-metal capacitors for 22 nm high volume manufacturing," in *Proc. IEEE Int. Interconnect Technol. Conf.*, Jun. 2012 pp. 1–3.

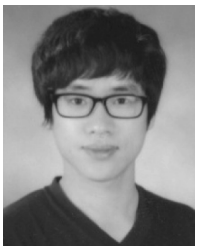
[15] Z. Guo, A. Carlson, L.-T. Pang, K. T. Duong, T.-J. K. Liu, and B. Nikolic, "Large-scale SRAM variability characterization in 45 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3174–3192, Nov. 2009.

Kyoman Kang was born in Gunpo, Korea, in 1987. He received the B.S. and M.S. degrees in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2013 and 2015, respectively.

He joined the Memory Division, Samsung Electronics, Hwaseong, Korea, in 2015. His current research interests include FinFET-based low power and high performance SRAM, and the circuit design flash memory.

Hanwool Jeong (S'12) was born in Seoul, Korea, in 1987. He received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, in 2012, where he is currently pursuing the Ph.D. degree in electrical and electronic engineering.

His current research interests include FinFET-based SRAM bitcell design and SRAM assist-circuit design.



Younghwi Yang was born in Seoul, Korea, in 1988. He received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, in 2011, where he is currently pursuing the Ph.D. degree in electrical and electronic engineering.

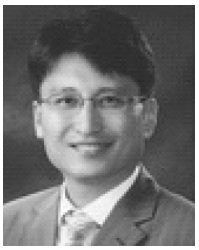
His current research interests include ETSOI-based SRAM bitcell design, and FinFET based near- and sub-V_{th} SRAM.

Juhyun Park was born in Incheon, Korea, in 1988. He received the B.S. degree in electronic and electrical engineering from Hongik University, Seoul, Korea, in 2012. He is currently pursuing the Ph.D. degree in electrical and electronic engineering from Yonsei University, Seoul.

His current research interests include FinFET-based near- and sub-V_{th} SRAM.

Kiryong Kim was born in Yeongdong, Korea, in 1989. He received the B.S. degree in electrical and electronic engineering from Yonsei University, Seoul, Korea, in 2014, where he is currently pursuing the Ph.D. degree in electrical and electronic engineering.

His current research interests include FinFET-based SRAM bitcell design and SRAM assist circuit design.



Seong-Ook Jung (M'00–SM'03) received the B.S. and M.S. degrees in electronic engineering from Yonsei University, Seoul, Korea, in 1987 and 1989, respectively, and the Ph.D. degree in electrical engineering from the University of Illinois at Urbana–Champaign, Urbana, IL, USA, in 2002.

He was with Samsung Electronics Company, Ltd., Hwasung, Korea, from 1989 to 1998, where he was involved in the specialty memories, such as video RAM, graphic RAM, and window RAM, and merged memory-logic. From 2001 to 2003, he was with T-RAM Inc., Mountain View, CA, USA, where he was the Leader of Thyristor-Based Memory Circuit Design Team. From 2003 to 2006, he was with Qualcomm Inc., San Diego, CA, USA, where he was involved in the high-performance low-power embedded memories, process variation tolerant circuit design, and low-power circuit techniques. Since 2006, he has been a Professor with Yonsei University. His current research interests include process variation tolerant circuit design, low-power circuit design, mixed-mode circuit design, and future-generation memory and technology.

Dr. Jung is currently a Board Member of the IEEE SSCS Seoul Chapter.