A 667 MHz Logic-Compatible Embedded DRAM Featuring an Asymmetric 2T Gain Cell for High Speed On-Die Caches

Ki Chul Chun, Pulkit Jain, Tae-Ho Kim, and Chris H. Kim, *Senior Member, IEEE*

***Abstract*—Circuit techniques for enhancing the retention time and random cycle of logic-compatible embedded DRAMs (eDRAMs) are presented. An asymmetric 2T gain cell utilizes the gate and junction leakages of a PMOS write device to maintain a high data '1' voltage level which enables fast read access using an NMOS read device. A current-mode sense amplifier (C-S/A) featuring a cross-coupled PMOS latch and pseudo-PMOS diode pairs is proposed to overcome the innate problem of small read bit-line (RBL) voltage swing in 2T eDRAMs with improved voltage headroom and better impedance matching under process–voltage–temperature (PVT) variations. A half-swing write bit-line (WBL) scheme is adopted to improve the WBL speed by 33% and reduce its power dissipation by 25% during write-back operation with no effect on retention time. A stepped write word-line (WWL) driver reduces the current drawn from the boosted high and low supplies by 67%. A 192 kb eDRAM test chip with 512 cells-per-BL implemented in a 65 nm low-power (LP) CMOS process shows a random cycle frequency and latency of 667 MHz and 1.65 ns, respectively, at 1.1 V and 85 °C. The measured refresh period at a 99.9% bit yield condition was 110 μs which is comparable to that of recently published 1T1C eDRAM designs.**

***Index Terms*—Cache, logic-compatible eDRAM, random cycle, sense amplifier, 2T gain cell.**

I. Introduction

MULTI-CORE PROCESSORS exploit microarchitecture-level parallelism to deliver higher computing performance while curbing chip power dissipation. The number of cores per socket has increased at a pace of two per year for high end enterprise processors [1]. To fully utilize the multi-core architecture with a larger appetite for data, there needs to be a commensurate increase in the amount of on-die embedded memory [1]–[3]. As a result, in the past decade, the die area devoted to cache memory has grown to approximately 50% in state-of-the-art processors. For example, Intel's 8-core Itanium processor has in total 54 MBs of on-die SRAM memory including a 32 MB Last Level Cache (LLC) [3] while IBM's POWER7 processor has a 32 MB L3 cache built in an embedded DRAM (eDRAM) technology [4], [9]. The need for robust high-density embedded memories is projected to grow as designers continue to seek power-conscious ways to improve multi-core chip performance.

6T SRAMs have been the embedded memory of choice due to their logic compatible bit-cell, fast differential read, and static data retention. However, the relatively large cell size and conflicting requirements for read and write at low operating voltages make aggressive scaling of 6T SRAMs challenging in scaled CMOS technologies. Recently, embedded DRAMs (eDRAMs) have been gaining traction in the research community due to features such as small cell size, low cell leakage, and non-ratioed circuit operation. There have been a number of successful eDRAM designs based on traditional 1T1C DRAM cells as well as logic-compatible gain cells [6]–[14].

A comparison between 6T SRAM, 1T1C eDRAM, and 2T gain cell eDRAM is shown in Table I. For fair comparison, the three memory circuits were evaluated in the same 65 nm process. 1T1C eDRAM has a 4.5X higher bit-cell density and a 5X lower static power dissipation than 6T SRAMs even when the refresh power is included. This enables a smaller chip size, a faster memory access, and a higher memory density which are the most effective ways to improve microprocessor performance under given power constraints. However, 1T1C eDRAMs require a complex capacitor fabrication process as well as an ultra-low leakage access transistor, and also suffer from the destructive read due to the charge sharing operation which makes them less attractive in future technology nodes.

Gain cells are implemented using logic devices allowing them to be built in a standard CMOS process with minimal adjustments. The cell can be implemented using three transistors, or even two transistors when used with delicate read control circuits, achieving a roughly 2X higher bit-cell density than SRAM as recently demonstrated by several industrial designs [5], [10]–[12]. Furthermore, gain cells can have smaller cell leakage current than SRAMs in sleep mode due to the fewer number of devices and the negative-Vgs biasing condition. Therefore, the static power dissipation of gain cell eDRAM including both leakage power and refresh power components can be smaller than that of an SRAM and similar to that of a 1T1C eDRAM [6], [9]. The cell write margin is better than SRAMs since there is no contention between the access device and the cross-coupled latch in a gain cell. Despite these favorable attributes, conventional gain cells suffer from short retention times due to the small storage capacitor and leakage currents that vary exponentially under Process-Voltage-Temperature

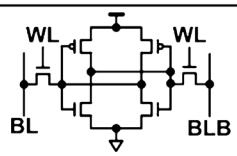
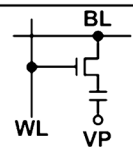
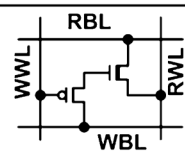
Manuscript received May 08, 2011; revised August 31, 2011; accepted September 01, 2011. Date of publication November 15, 2011; date of current version January 27, 2012. This paper was approved by Associate Editor Stefan Rusu.

The authors are with Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA (e-mail: kichul.chun@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2011.2168729

TABLE I
EMBEDDED MEMORY OPTIONS FOR HIGH DENSITY ON-DIE CACHES

	6T SRAM [5]	1T1C eDRAM [6]	This Work [14]
Cell Schematic			
Process	Logic compatible	+2 (FEOL) +3 (Cap)	Logic compatible
(1)Reported cell size (ratio)	0.46x1.24= 0.5704μm^2 (1X)	0.23x0.55= 0.1265μm^2 (0.22X)	0.475x0.58= 0.2755μm^2 (0.48X) [12]
(2)Redrawn cell size (ratio)	0.575x2.05= 1.179μm^2 (1X)	0.45x0.545= 0.245μm^2 (0.21X)	0.48x0.995= 0.478μm^2 (0.41X)
(2)Redrawn 1Mb macro (ratio)	1.317x1.124= 1.481mm^2 (1X)	0.632x0.739= 0.467mm^2 (0.32X)	1.168x0.638= 0.746mm^2 (0.50X)
Data storage	Latch (Static)	Capacitor (20fF)	MOS gate (<1fF)
Cell access	(+) Differential read (-) Ratioed operation	(-) Destructive read (-) Refresh	(+) Decoupled read and write, (-) Refresh
Random cycle	(2)1GHz	500MHz	(2)667MHz
Static power (ratio)	1X	0.2X	0.19X @500MHz

(1)All designs are in 65nm, (2)Based on the same 65nm low power CMOS process

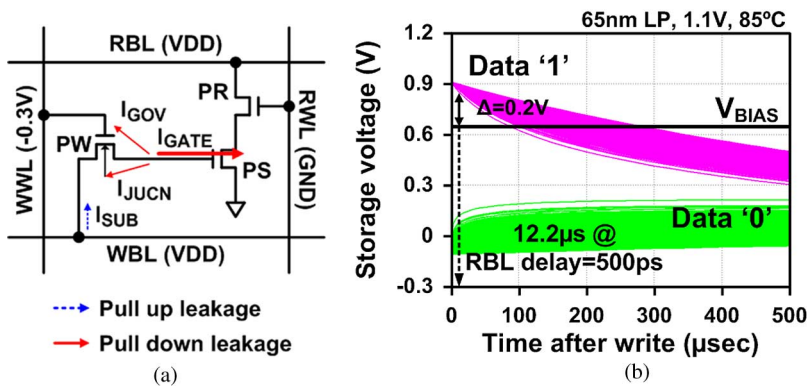


Fig. 1. (a) Leakage components of a 3T NMOS gain cell during data hold mode. (b) Monte-Carlo simulation results of storage node voltage during data hold mode showing 1 Mb macro retention characteristics.

(PVT) variations [10]–[12]. A shorter retention time leads to higher refresh power dissipation and/or smaller read current. The former is a result of the more frequent refresh operation while the latter is due to the faster loss of cell voltage. Frequent refresh operation also reduces memory availability resulting in a degradation in overall system performance. Therefore, attaining practical retention time and improving random access speed remain as key challenges in gain cell eDRAM designs. In this paper, we present circuit techniques for realizing a 1.1 V, 667 MHz random cycle eDRAM with a retention time comparable to that of 1T1C eDRAMs.

The remainder of this paper is organized as follows. Section II presents the proposed circuit techniques to enhance the retention time and improve the performance of a gain cell eDRAM. Section III comprehensively compares macro dimensions, access speeds, and static power dissipations of 6T SRAM and gain cell eDRAM arrays. Section IV describes hardware measurement results from a 65 nm test chip, and conclusions are drawn in Section V.

II. PROPOSED 2T GAIN CELL eDRAM DESIGN

To aid the understanding of our proposed techniques, we first describe the basic retention characteristics of a conventional 3T gain cell. In the 3T NMOS cell shown in Fig. 1(a), PW denotes the write access device, PS the storage device, and PR the read access device. Unlike 6T SRAMs or 1T1C eDRAMs, gain cells have a decoupled read and write structure—Read Word-Line (RWL) and Read Bit-Line (RBL) are used for read access and

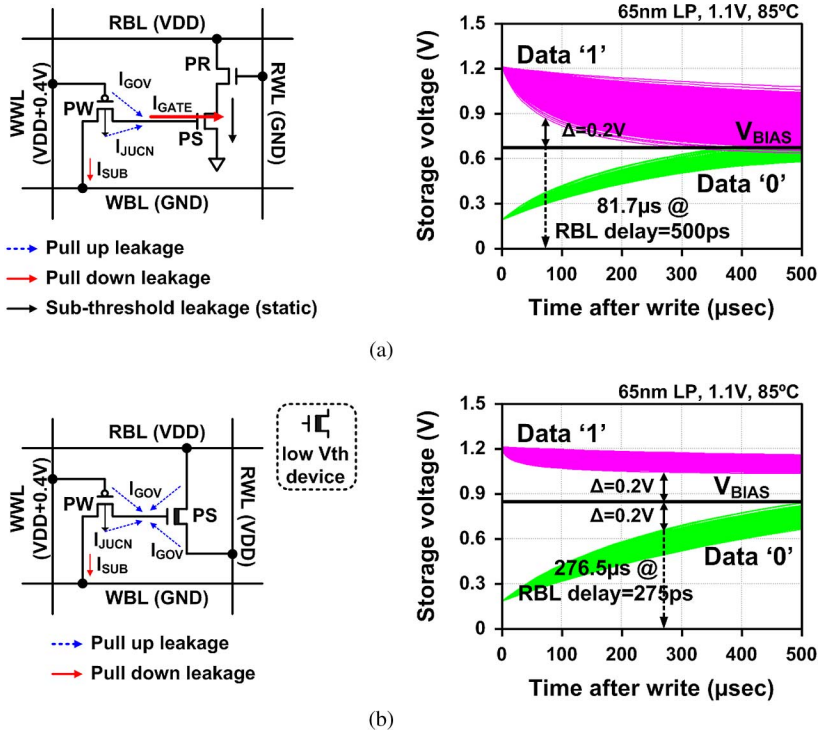


Fig. 2. Circuit diagrams and retention characteristics of (a) a previous asymmetric 3T gain cell [10] and (b) the proposed asymmetric 2T gain cell.

Write Word-Line (WWL) and Write Bit-Line (WBL) are used for write access. This attribute leads to improved read and write margins and flexibility in the bit-cell design—for example, the read and write paths can be optimized separately allowing gain cells to scale favorably in future technology nodes.

In data retention mode, PW and PR are turned off and the storage node is left floating. The sub-threshold, gate, and junction leakages in the surrounding devices cause the floating voltage to change with time as shown in Fig. 1(b). Since the storage node is surrounded by many low supplies in an NMOS only cell, the retention time of data '1' is much shorter than that of data '0'. To make matters worse, the data '1' (not data '0') voltage level is critical for the read access speed as the read port also uses an NMOS. The data retention time depends on the aggregated leakage current flowing into the storage node. Fig. 1(b) shows the cell retention time variations obtained by running 2^{20} Monte-Carlo simulations in HSPICE, which represents the cell-to-cell variation of a 1 Mb memory macro. In this analysis, we define retention time as the time it takes for the cell node voltage to reach a level corresponding to a target RBL delay of 500 ps. The read reference bias level is set as 0.65 V and the data '1' voltage should be higher than this reference voltage by at least 0.2 V to achieve the same read margins as the data '0' case. Results based on our criterion indicate that the retention time of data '1' varies from 12.2 μs to 54.1 μs mainly due to the gate leakage through the inverted channel of the NMOS storage device, while the non-critical data '0' voltage shows a very stable retention characteristic. Note that the WWL coupling after write-back operation results in lower initial storage levels than VDD and GND in case of data '1' and data '0', respectively. This further degrades the retention time of data '1' when a gain cell is implemented only with NMOS devices. The central idea of this work is to maximize the retention time and performance by using a new bit cell that balances the retention characteristics of data '0' and '1'.

A. Asymmetric 2T Gain Cell

PMOS only gain cells were used in recent designs for improving retention time as they have 1–2 orders of magnitude lower gate leakage compared to their NMOS counterpart [12], [13]. However, the pull-up leakage currents of the PMOS devices surrounding the storage node have a negative impact especially on the data '0' level which determines the current through the PMOS read device. In addition, the poor channel mobility of PMOS devices limits the read performance. The new 2T gain cell structure proposed in this work achieves a long retention time without sacrificing read speed by using an NMOS read device driven by RWL for high drive current and a PMOS write device to keep the speed critical data '1' voltage close to VDD [14]. Fig. 2 shows the proposed 2T cell and a previous Asymmetric 3T Cell (ATC) which was chosen for comparison because it also contains both NMOS and PMOS devices, albeit the structure and operating principle are considerably different [10]. In the previous ATC cell, a PMOS device was used for the write access transistor to extend the cell retention time by compensating the NMOS gate leakage with the PMOS gate overlap and junction leakages. However, the leakage compensation effect of this cell is poor under PVT variations because the gate leakage through the inverted channel of the NMOS storage device is dominant for data '1' as shown in Fig. 2(a). In the proposed cell shown in Fig. 2(b), the read access transistor is replaced by the RWL signal whose pre-charge level is VDD. The storage transistor is nominally off making its gate leakage negligible. Since there is no sub-threshold leakage through the read path, a low

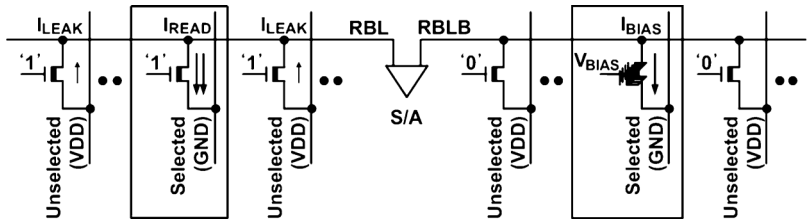


Fig. 3. Illustration of limiting read margin by adjacent cells holding high state in a 2T eDRAM.

Vth transistor can be utilized to further improve read speed. The proposed current sensing scheme described in the next section limits the RBL voltage swing to about 100 mV which eliminates problems associated with the pull-up leakage from the data '1' cells on the same RBL. Fig. 2 (right) shows the simulated retention characteristics of a 1 Mb macro. The WWL coupling after write-back operation boosts data '1' level by 110 mV in a PMOS write device. However, the previous 3T ATC still suffers from a poor data '1' retention time due to the large gate leakage of storage device. The proposed asymmetric 2T gain cell improves worst case retention time by 3.4X while at the same time achieving a 45% shorter RBL delay compared to the previous 3T ATC. An additional benefit of the proposed 2T asymmetric cell is the balanced P and N diffusion densities which makes it more ideal to address Design-For-Manufacturability (DFM) concerns in extremely scaled technologies.

B. Pseudo-PMOS Diode Based Current-Mode Sense Amplifier (C-S/A)

Unlike in 3T cell designs, the RBL of 2T cells must have a limited swing to prevent the leakage current of the unselected cells from causing a read failure as illustrated in Fig. 3. However, a small voltage swing means that the read sensing margin is poor. The proposed asymmetric 2T gain cell worsens this situation since it utilizes a low Vth read device to achieve faster read speed by keeping the speed critical data '1' voltage close to VDD. Simulation results in Fig. 4 show a read failure in the worst case when all unselected cells on the same RBL hold a strong data '1' at a high temperature and fast process corner condition.

To overcome this problem, a Current-mode Sense Amplifier (C-S/A) is employed in our design to hold the RBL voltage close to VDD while sensing, allowing a large number of low Vth cells to be connected to a single RBL. The most common C-S/A shown in Fig. 5(a) consists of a PMOS load (P0), a cross-coupled PMOS latch (P1) and an NMOS diode (N1) pair [15]. The PMOS load pair provides currents to the cells and the C-S/A so that RBL can remain close to VDD during read operation. The cross-coupled PMOS latch pair has a negative input impedance and amplifies the input currents. The NMOS diode pair has a positive input impedance and stabilizes the output voltages. The total input impedance of the C-S/A can be expressed as

$$R_{IN} = \frac{g_{m,N1} - g_{m,P1}}{g_{m,N1} g_{m,P1}} \quad (1)$$

indicating that a good matching between the PMOS latch and the NMOS diode pairs is required for a low input impedance. However, in the presence of P/N skew and PVT variations,

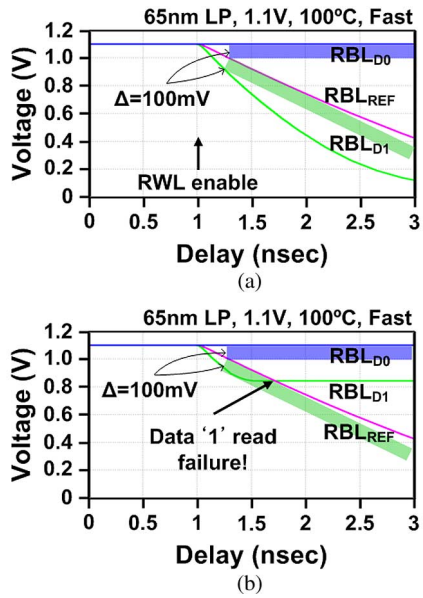


Fig. 4. (a) Simulated RBL sensing waveform when all adjacent cells hold a data '0'. (b) All adjacent cells hold a data '1' indicating a data '1' read failure. The shaded regions denote the $\Delta V_{\mathrm{RBL}} = 100\ \mathrm{mV}$ window between the accessed RBL and the reference.

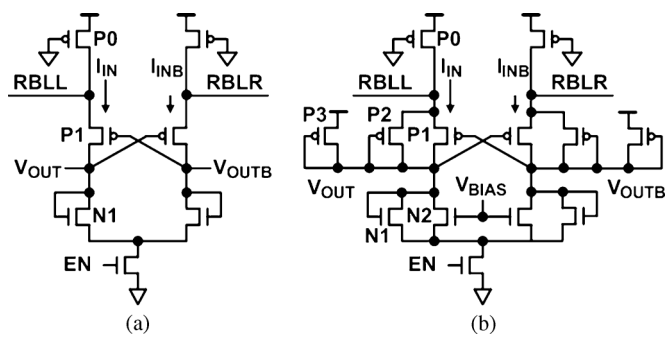


Fig. 5. (a) NP series-stacked C-S/A [15]. (b) Hybrid C-S/A [16].

matching the two impedances becomes difficult. Moreover, this conventional C-S/A suffers from a limited voltage headroom due to the stacked devices between VDD and GND.

An improved circuit shown in Fig. 5(b) consists of two folded PMOS diode pairs (P2 and P3), an NMOS current source (N2), and a cross-coupled PMOS latch pair (P1). N2 is biased using a separate voltage so the voltage headroom is increased by approximately $1 \times \mathrm{Vth}$. Note that the conventional NMOS diode pair (N1) turns on only at a high supply voltage condition to improve the stability of this C-S/A [16]. Despite these advantages, the large number of devices in this circuit makes it impractical

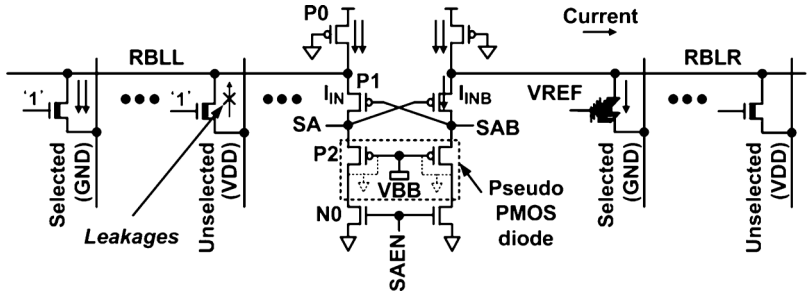


Fig. 6. Proposed pseudo-PMOS diode based C-S/A to overcome the issue of limited RBL voltage swing in a 2T eDRAM with improved voltage headroom and better impedance matching.

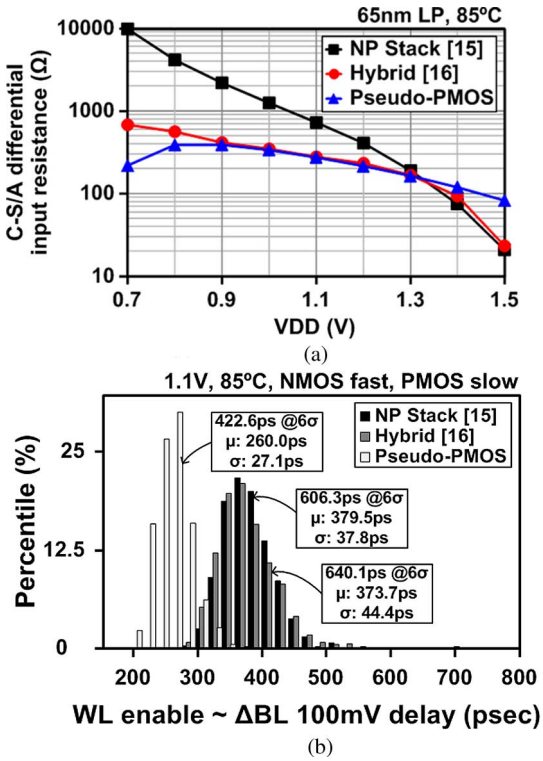


Fig. 7. (a) Simulated input resistance ($\Delta V_{RBL}/\Delta I_{IN}$) vs. VDD. (b) Comparison of RBL sensing delay under PVT variations and mismatches in the C-S/A pairs.

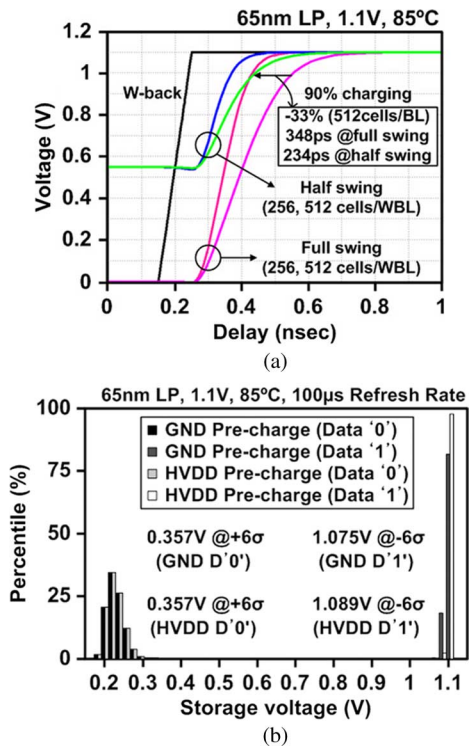


Fig. 8. (a) Simulated waveforms of the WBL charging delay. (b) Simulated storage voltage distributions of a conventional GND pre-discharge (full swing WBL) and the proposed half-VDD pre-charge (half swing WBL) schemes.

for DRAM circuits where every BL should have a dedicated S/A for a row-by-row refresh operation. This results in a large BL-S/A layout overhead in addition to impedance mismatch issues under PVT variations. The input resistance of this hybrid C-S/A is given as

$$R_{IN} = \frac{g_{m,N1} + g_{m,P2} + g_{m,P3} - g_{m,P1}}{(g_{m,P1} + g_{m,P2})(g_{m,N1} + g_{m,P3})}. \quad (2)$$

The proposed C-S/A shown in Fig. 6 consists of a cross-coupled PMOS latch (P1) and a pseudo-PMOS diode (P2) driven by the negative supply VBB which is readily available on the chip for WWL under-driving. Recall that a negative WWL is needed for a PMOS device to write a data '0' into the cell without a threshold voltage loss. Both PMOS pairs (P1 and P2) are in saturation mode which means that they provide better matching. Moreover, voltage headroom is improved compared to previous designs ensuring robust sensing. In order to guarantee that both pairs operate in the saturation region, the following condition should be met.

$$V_{DD} - V_{sd\_P0} - V_{ds\_N0} > V_{DD} + |V_{BB}| - 2 \times |V_{TH}| - V_{sd\_P0} \quad (3)$$

This can be further simplified as:

$$|V_{BB}| < 2 \times |V_{TH}| - V_{ds\_N0} \quad (4)$$

For VBB $= -0.5$ V, $V_{ds\_N0} = 0.05$ V, and Vth $= -0.315$ V (85 °C) which are the values used in our design, the above inequality indicates that the P1 and P2 pairs will operate in the saturation region under PVT variations while enhancing the voltage headroom by 0.5 V. Similar to the conventional C-S/A, the input resistance of the proposed C-S/A can be expressed as:

$$R_{IN} = \frac{g_{m,P2} - g_{m,P1}}{g_{m,P1} g_{m,P2}}. \quad (5)$$

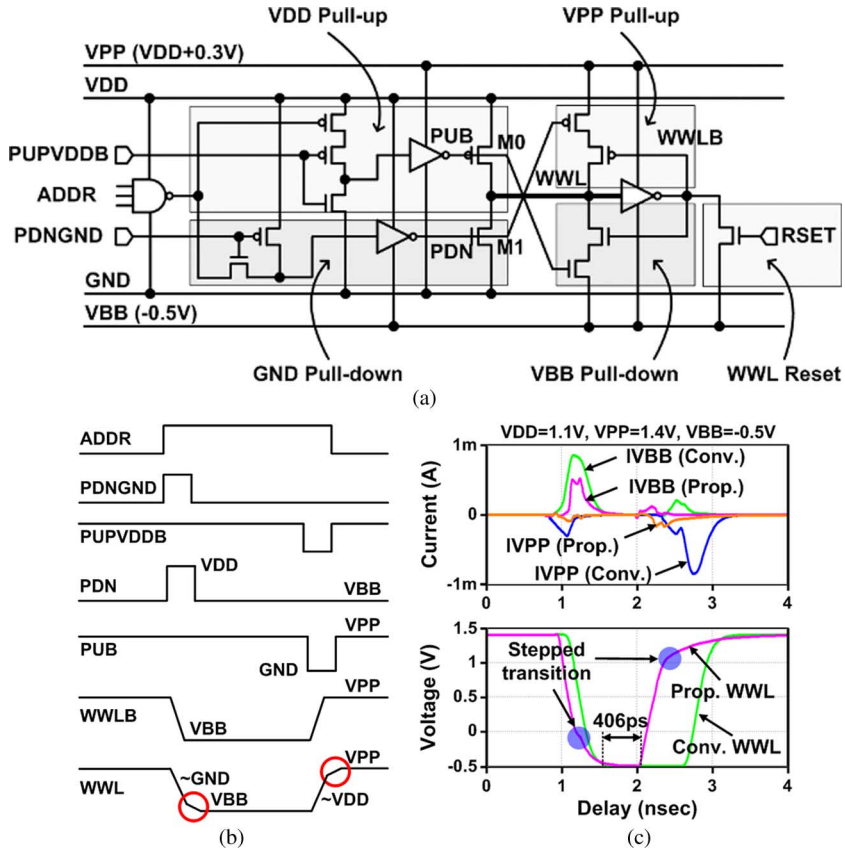


Fig. 9. Proposed stepped WWL driver. (a) Schematic. (b) Timing diagram. (c) Simulated boosted current consumptions and WWL waveforms during transition.

Fig. 7(a) shows simulated differential input resistances of the three C-S/A's at different VDDs. For this comparison, the C-S/A pairs were designed to have a minimum input resistance at the high VDD corner to ensure good stability [16]. The previous NP stack structure suffers from large input resistance at low operating voltage conditions leading to a considerable signal loss for the current sensing scheme. When this C-S/A operates in the sub-threshold region, the transconductances of the two pairs decrease. The denominator of (1) is the product of the two transconductances, while the numerator is the sum. This results in a rapid increase in input resistance at lower supply voltages as shown in Fig. 7(a). Input resistance of the previous hybrid C-S/A and the proposed pseudo-PMOS C-S/A show a stable response down to 0.9 V and 0.7 V, respectively. The maximum input resistance allowed in this design is 500 Ω which corresponds to a 10% signal loss during current sensing. Unlike the previous hybrid C-S/A, the improvement of low voltage margin in the proposed design depends on the voltage difference between the VBB (-0.5 V) and the threshold voltage (-0.315 V). Fig. 7(b) shows simulation results of RBL sensing delay for the NP stack, hybrid, and proposed C-S/A's. Each distribution represents the delay variation of the proposed gain cells from a 1 Mb macro with a refresh period of 100 μs. These Monte-Carlo results include cell leakage variations as well as device variations in the read path and C-S/A pairs. Although the hybrid C-S/A has a smaller input resistance than the NP stack C-S/A at 1.1 V, ensuring good matching between the large number of device pairs is difficult and results in a poor overall performance. The proposed C-S/A utilizing a pseudo-PMOS diode enhances the RBL sensing delay by 30.3% (6-sigma point) due to the improved impedance matching and better low VDD margin.

C. Half Swing Write Bit-Line Scheme

With the improved read bit-line sensing speed and increased number of cells per BL, WBL switching speed becomes the performance bottleneck. Similar to the half-VDD pre-charge technique employed in standard 1T1C DRAMs, a half swing WBL scheme can be applied to gain cell eDRAMs. By using a half swing WBL scheme with a tri-state buffer, the write speed is improved by 33% and the average WBL charging current is reduced by 25% without affecting the retention characteristics of the proposed 2T cell. Fig. 8(a) shows simulated waveforms for a conventional GND pre-discharge scheme (full swing) and the half-VDD pre-charge scheme (half swing) indicating a 33% improvement in WBL charging speed. Retention characteristics of the GND pre-discharge scheme and the half-VDD pre-charge scheme are similar as shown in Fig. 8(b) since the sub-threshold leakage through the write device during data hold mode is negligible owing to the WWL over drive (VDD + α, where α = 0.3 V in this design). Moreover, sub-threshold leakage through the write device can be effectively cut off during the data '1' write-back operation of a cell sharing the same WBL. The half swing WBL scheme is implemented as a part of the write-back circuit as shown in Fig. 10.

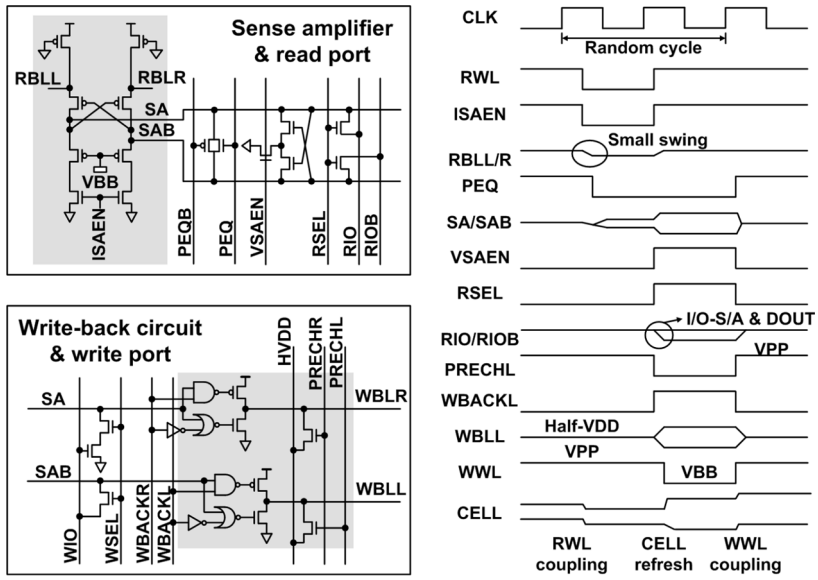


Fig. 10. (a) Circuit diagram of the proposed Sense Amplifier (S/A) with read port, write port, and write-back circuits. (b) Two-stage read and write-back timing diagram.

D. Stepped Write Word-Line Driver

DRAMs require a positive boosted voltage (VPP) to suppress the sub-threshold leakage in the write access device as well as a negative boosted voltage (VBB) to write data into the cell without a Vth drop (PMOS write device case). In order to reduce the power and area overhead of charge pumps during fast chip operation, we adopted a stepped WWL control scheme which minimizes the current drawn from the boosted VPP and VBB voltages by utilizing the main VDD and GND supplies for most of the WWL transition. The proposed WWL scheme consists of a nominal VDD/GND driver including tri-state control circuits, a boosted VPP/VBB driver with an inverted signal, and a reset device as shown in Fig. 9(a). Before the cell access, PUB and PDN nodes in Fig. 9(a) are set to VPP and VBB, respectively. This deactivates the VDD/GND driver by cutting off the short circuit current path from VPP to VDD and from GND to VBB. The RSET signal is switched to VPP ensuring that all WWL's are pre-charged to the desired VPP level. Except during the initialization phase, the RSET signal stays at VBB. At the beginning of the write-back operation, decoded address signals and a short pulsed signal of PDNGND enable the GND pull-down path in Fig. 9(a). This drives the selected WWL towards GND. As the selected WWL is discharged, WWLB switches and enables the VBB pull-down path which drives the WWL to VBB. The pulse duration has to be carefully controlled to guarantee proper circuit operation while saving the WWL switching power. If the pulse duration is too short, the VBB pull-down path will not be enabled whereas if it is too long, there will be short circuit current between VBB and GND. In this design, we chose a pulse duration of 375 ps which gave sufficient timing margin at a slight increase in the current drawn from the boosted supply. The operating principle of the opposite high-to-low WWL transition is similar to what we described above and the waveforms are shown in Fig. 9(b).

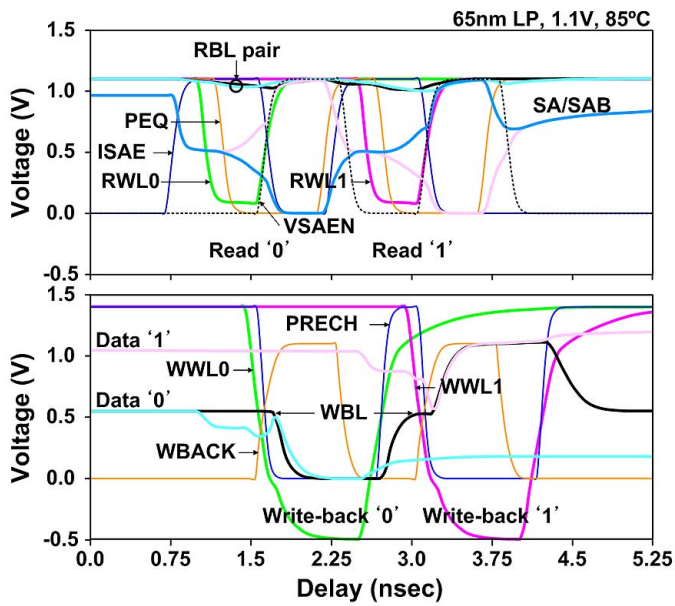


Fig. 11. Simulated waveforms of back-to-back read and write-back operations for a 1.5 ns cycle time.

Fig. 9(c) shows the simulated waveforms of the current consumption and WWL transition for the conventional and proposed schemes. With a stepped WWL control scheme, 67% of the boosted supply current and 4.3% of the total chip area can be saved with two additional peripheral control signals and four more transistors in the WWL control circuit compared to conventional two-stage level shifters. Note that during a step transition of WWL, the effective pulse width is decreased. Nevertheless, a WWL pulse width of 406 ps can be achieved at a 1.5 ns cycle time which is significantly longer than the required pulse width of 210 ps. Further details on the macro level timing will be given in Section III.

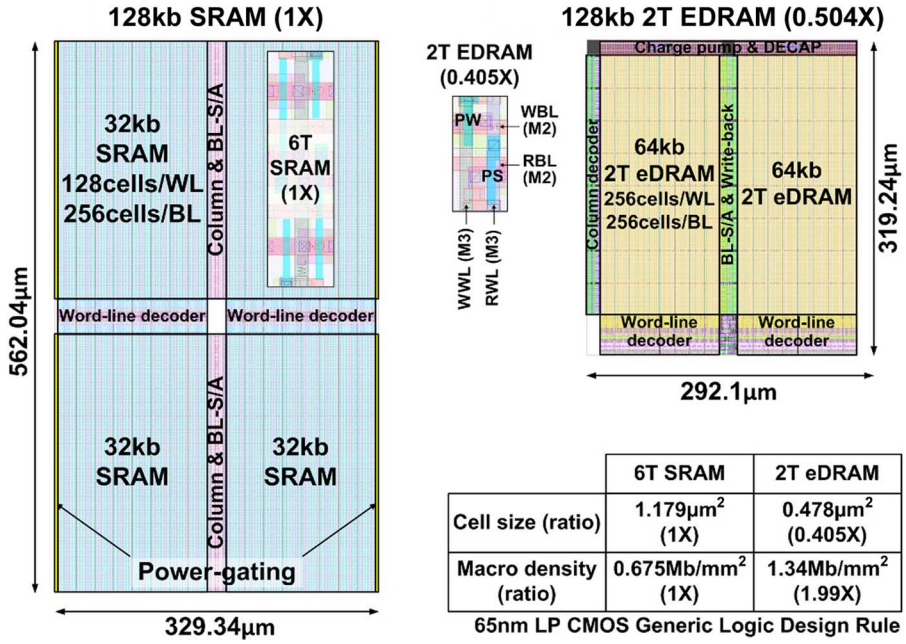


Fig. 12. Comparison of bit-cell and 128 kb sub-array layout between 6T SRAM and 2T eDRAM.

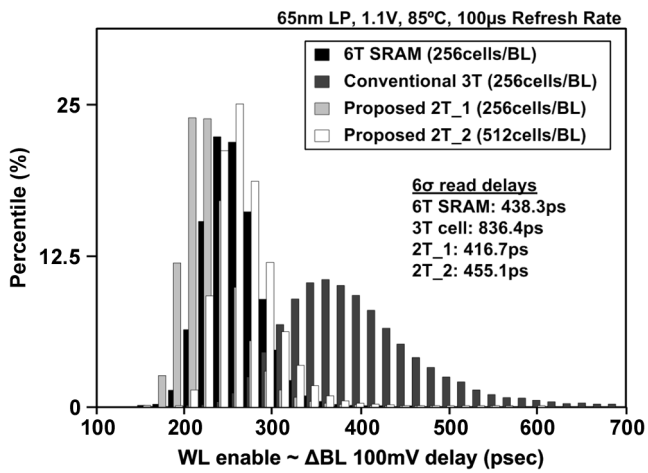


Fig. 13. RBL sensing delay distributions of SRAM and gain cell eDRAMs each with a 1 Mb macro density.

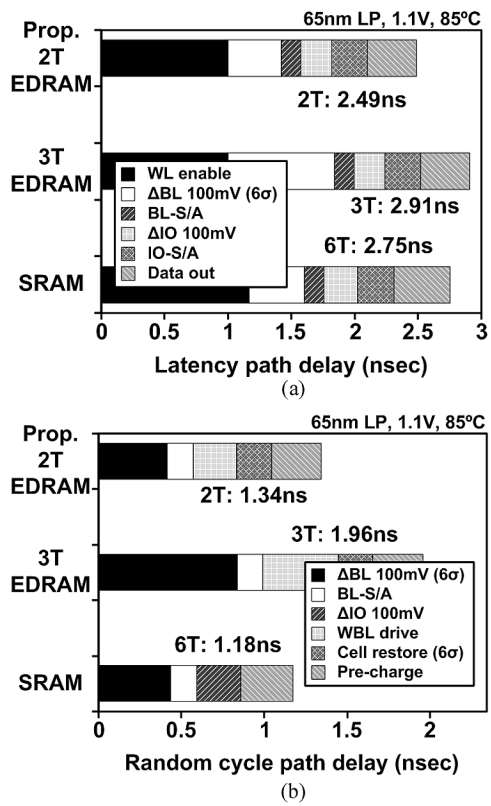


Fig. 14. Performance comparison of 1 Mb macros using SRAM and gain cell eDRAMs. (a) Latency. (b) Random cycle.

E. Sense Amplifier and Write-Back Circuit Design

Fig. 10 shows the complete schematic and timing diagram of the proposed S/A, read port, write-back, and write port. A two-stage full pipeline structure was implemented to control the read and write-back operations. In the first clock cycle, the RWL is selected. When the C-S/A control signal (ISAEN) is enabled, the C-S/A amplifies the input signals to analog voltage signals while the RBL held close to VDD. Once a recognizable voltage difference is developed, the voltage S/A control signal (VSAEN) is fired. In the second clock cycle, read-out and write-back operations follow. After the write-back, WBLs are pre-charged back to half-VDD using the boosted supply VPP control signal (PRECHL/R). A stepped PRECH control scheme can be also adopted to further minimize the current drawn from the boosted supply VPP.

Fig. 11 shows post-layout simulation waveforms of the proposed 2T eDRAM. This includes the proposed asymmetric 2T gain cell, the pseudo-PMOS diode based C-S/A, a half-swing WBL scheme, and a stepped WWL driver. The memory array with 192 cells-per-WL and 512-cells-per-BL can operate at a random cycle time of 1.5 ns for a test sequence of data '0' read and write-back followed by data '1' read and write-back.

III. Comparison Between SRAM and Gain Cell eDRAM

In order to demonstrate the advantages of the proposed 2T eDRAM over conventional 3T eDRAM or 6T SRAM, this section presents macro level layout and performance comparisons. Static power comparisons are detailed in Section IV. Extensive Monte-Carlo simulations were performed on megabit density SRAM and eDRAM arrays to estimate their performance in a practical scenario [13], [17]. Our analysis includes process variation in the memory cells and the C-S/A as well as realistic fluctuations for the reference biases and boosted supplies.

A. Macro Layout Comparison

Fig. 12 shows the bit-cell and 128 kb sub-array layouts of a 6T SRAM and the proposed 2T eDRAM in a generic 65 nm LP CMOS process. Dense bit-cell design rules were not available to the authors but for area comparison purposes, using a logic design rule is a generally accepted practice. The 6T SRAM used for the comparison has the following transistor dimensions: $W_{PU} = W_{min}$, $W_{PD} = 2 \times W_{min}$, and $W_{ACCESS} = W_{min}$, with all devices using a minimum channel length. This is the most general sizing scheme and extensive Monte Carlo simulations were performed to verify good read and write margins. The bit cell area of the proposed 2T gain cell is 59.5% smaller (or 2.47X denser) than that of a 6T SRAM resulting in a 49.6% smaller area for a 128 kb sub-array. It is worth mentioning that layout of the 128 kb 2T eDRAM sub-array includes a BL-S/A and write-back driver in each BL, full RWL and WWL decoders, and charge pumps for generating boosted high and low supplies. The unit 128 kb sub-array can be tiled to build a larger memory macro.

B. Macro Performance Comparison

Fig. 13 shows read bit-line delay distributions for the following four memory arrays; a 1 Mb SRAM with 256 cells-per-BL, a 1 Mb conventional 3T eDRAM with 256 cells-per-BL, and a 1 Mb proposed 2T eDRAM with 256 and 512 cells-per-BL. The single-ended sensing nature and the gradual loss in the storage node voltage of the conventional 3T eDRAM result in a 6-sigma read bit-line delay that is 1.9 times longer than a 6T SRAM as shown in Fig. 13. The proposed 2T eDRAM makes up for this performance shortfall, achieving a bit-line sensing speed comparable to that of a 6T SRAM with 256 cells-per-BL. For an array with 512 cells-per-BL, the proposed 2T eDRAM shows only a 4% longer RBL sensing delay than a 6T SRAM that has half the number of cells-per BL. The performance improvement is attributed to the following three factors: excellent data '1' retention, low Vth device in the decoupled read path, and the proposed C-S/A which makes the read speed more or less independent of the RBL capacitance. For cache sizes of 1 Mb or larger, the proposed 2T eDRAM achieves a faster access time owing to the shorter global interconnect delay made possible by the smaller bit-cell size as shown Fig. 14(a). Therefore, a 512 cells-per-BL architecture was chosen for this 2T eDRAM design in order to verify our proposed schemes under extreme cases and to reduce the array layout overhead stemming from the complicated BL-S/A and write-back circuits.

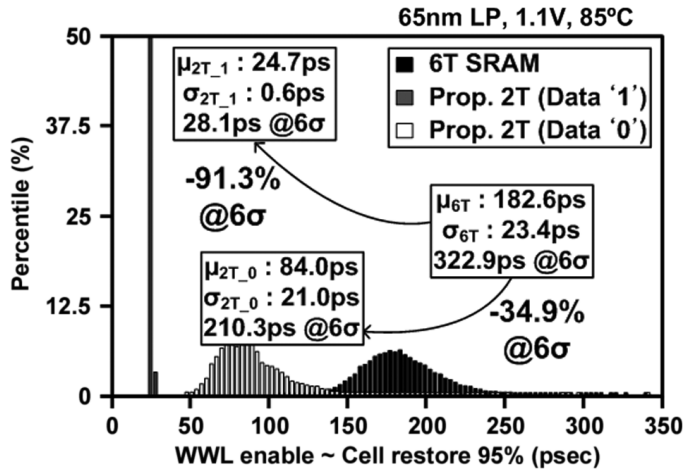


Fig. 15. Write delay distributions of SRAM and gain cell eDRAM each with a 1 Mb macro density.

Embedded DRAMs require a write-back operation after the read operation to restore the cell data. This results in a 66.5% slower random cycle time for a conventional 3T eDRAM compared to a 6T SRAM as shown in Fig. 14(b). The proposed 2T eDRAM improves the random cycle time by 31.6% compared to a conventional 3T eDRAM that has half the number of cells per BL.

Fig. 15 shows the 1 Mb write delay distributions of a 6T SRAM array and the proposed 2T eDRAM array. Here, the write delay is defined as the WL activation to the time when the cell node reaches 95% of the full voltage swing. The write speed of the gain cell is faster than the 6T SRAM since the latter is based on a ratioed operation. For the speed critical data '1' case, the proposed 2T eDRAM achieves an 11.5X faster write-back (6-sigma point performance) compared to the 6T SRAM as shown in Fig. 15. Note that the WWL of the gain cell must be sufficiently negative in order for the PMOS write devices to pass a good data '0' level. For a WWL under-drive voltage of -0.5 V, the 1 Mb Monte-Carlo simulations show a write speedup of 35% (6-sigma point) for data '0'.

IV. Test Chip Implementation and Measurements

A 192 kb eDRAM test chip was implemented in a 1.2 V, 65 nm Low-Power (LP) logic CMOS process to demonstrate the proposed circuit techniques. The detailed array architecture is shown in Fig. 16 consisting of two 96 kb blocks sharing BL-S/A and write-back circuits located at the center of the array. The dummy memory cells in each block are 4X larger than the regular cells to minimize random device mismatch. RWL pull-down drivers are inserted every 64 WL's in order to minimize the RWL ground noise during read access. Fig. 17 shows the chip microphotograph and a feature summary table of the 192 kb eDRAM test chip. For a 99.9% bit yield at 1.1 V and 85 °C, our design achieves a random cycle frequency of 667 MHz and 500 MHz using a refresh period of 110 μs and 1200 μs, respectively. By increasing the VPP level from 1.5 V to 1.6 V, a 100 μs retention time can be achieved under a 99.99% bit yield condition. To put this into perspective, the target retention time of a previous 2T gain cell eDRAM design was 10 μs [12] while the

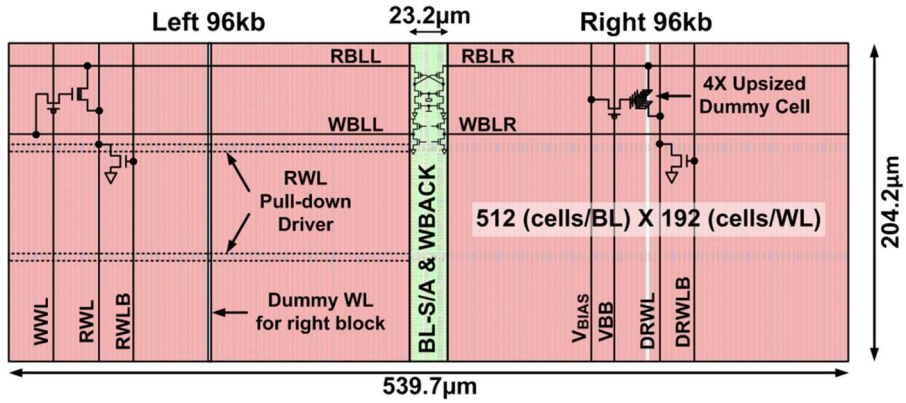
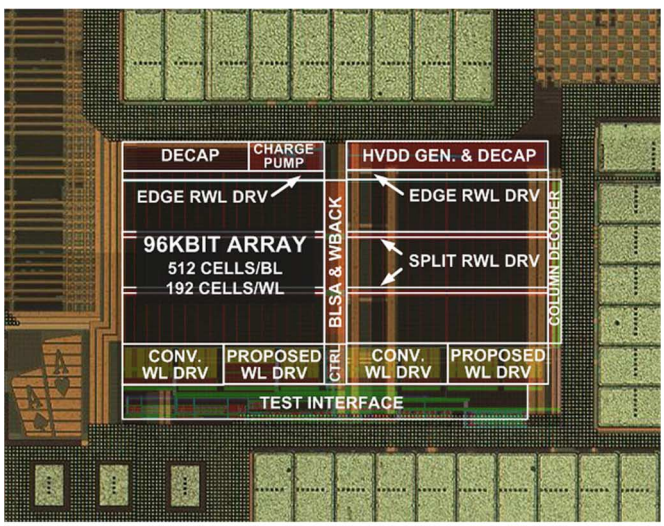


Fig. 16. A 192 kb test array architecture with 192 cells-per-WL and 512 cells-per-BL.



(a)

Process	65nm LP CMOS
Circuit dimension	555.8x297.8µm^2
Array size	192kb (192 WLs, 2x512 BLs)
Cell size	41% of 6T SRAM
* Retention time @ 1.1V, 85ºC	110µs @ 667MHz 1200µs @ 500MHz
Latency	1.39ns @ 1.2V 1.65ns @ 1.1V
** Refresh power @ 1.2V, 85ºC	1.16mW/Mb @ 667MHz 108.84µW/Mb @ 500MHz

* 99.9% bit yield condition
** 110µsec refresh rate for 667MHz
1200µsec refresh rate for 500MHz

(b)

Fig. 17. (a) Microphotograph of the 65 nm eDRAM test chip. (b) Chip feature summary.

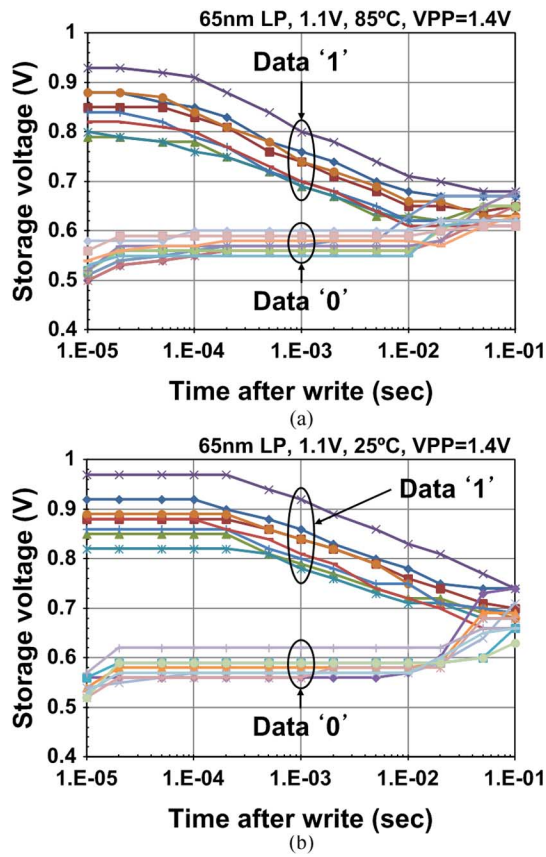


Fig. 18. Measured storage node voltage at different retention times at (a) 85 °C and (b) 25 °C.

measured retention time of a commercial 1T1C eDRAM was 40 μs at 105 °C with a 99.99% bit yield [6] each with a random cycle of 500 MHz.

By externally adjusting the read reference voltage (VDUM), we can indirectly and noninvasively measure the storage node voltage at different data retention times [13]. For example, read failure will happen for data '1' if the VDUM level is higher than the storage node voltage so the storage voltage can be measured by sweeping the VDUM voltage and measuring the point of failure. It is worth mentioning that the storage node voltage measured using this method includes effects such as process variation or transient noise (e.g. coupling noise or supply noise) providing us with the "effective" cell node voltage. The measured storage node voltage of the proposed 2T eDRAM in Fig. 18 shows that retention times even longer than 1 ms can be achieved.

Adjusting the VPP level modulates the gate overlap and gate-induced drain leakages and hence allows us to achieve an

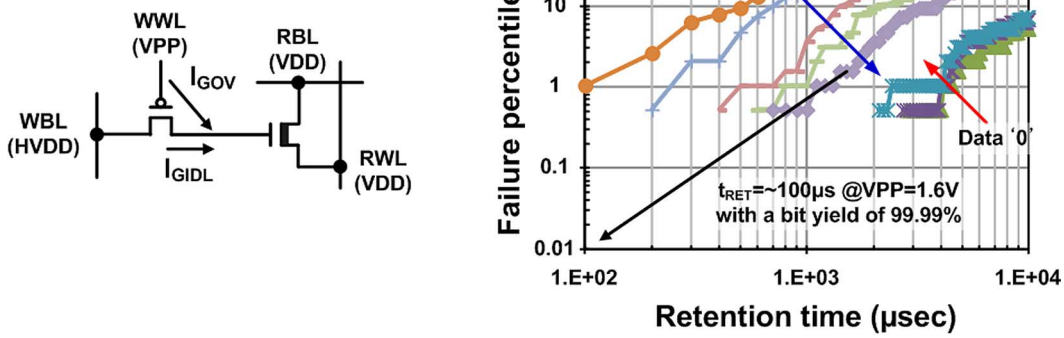


Fig. 19. Measured retention time distribution vs. boosted high supply (VPP) level.

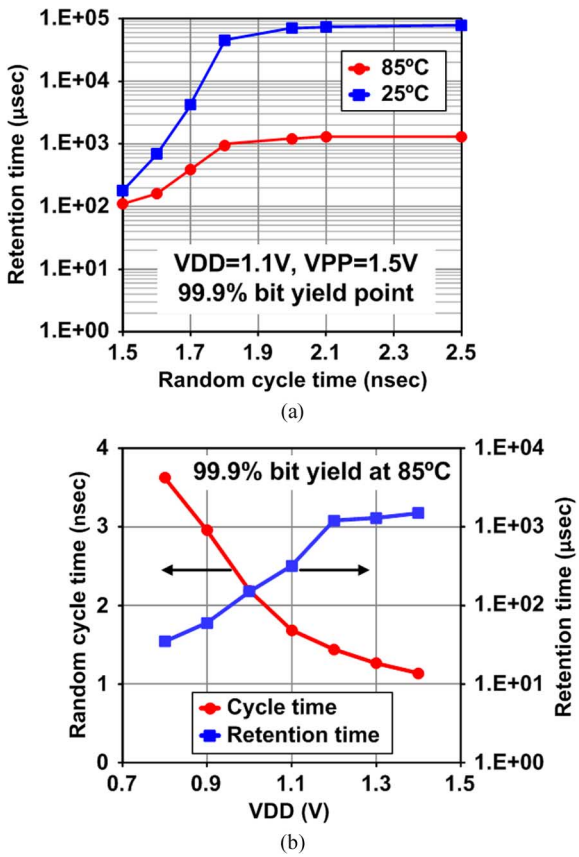


Fig. 20. (a) Measured random cycle time vs. retention time. (b) Measured VDD shmoo of random cycle time and corresponding retention time.

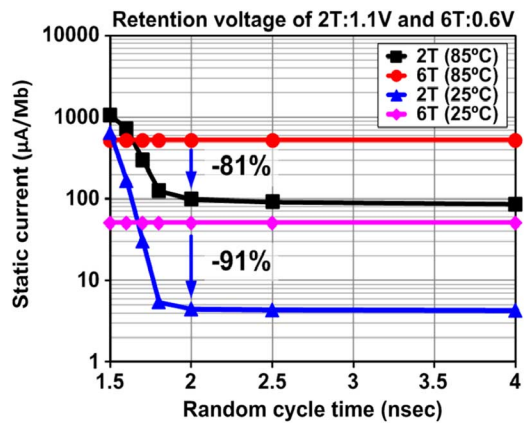


Fig. 21. Static power comparison between 6T SRAM and the proposed 2T eDRAM with varying random cycle time at 85 °C and 25 °C.

optimal retention time with the consideration of both data '1' and data '0' cases as shown in Fig. 19. This dependency can be further exploited for post-fabrication trimming to cope with die-to-die variations.

The retention time of a 2T eDRAM can be extended at the expense of a longer random cycle time as shown in Fig. 20(a). We can utilize this trade-off to enhance access speed, and at the same time minimize refresh power dissipation of the 2T eDRAM. During memory access, the S/A enable signal was triggered as early as possible after the RWL activation to achieve a high random cycle frequency as high as 667 MHz. Moreover, a delayed S/A enable signal extends the retention time resulting in significant refresh power savings. The measured refresh power at a random cycle of 667 MHz and 500 MHz were 1.16 mW/Mb and 109 μW/Mb, respectively at 1.1 V and 85 °C. The flexibility in the cycle time offers further opportunities to reduce refresh power depending on the system level workload and frequency requirements. For a 1 Mb macro with 1024 WL's, only 1.40% of the total operating time is spent on refresh for a 1.5 ns random cycle and a 110 μs refresh period. The refresh overhead reduces to 0.17% for a 2.0 ns random cycle and a 1200 μs refresh period. The measured VDD shmoo of cycle time and the corresponding retention time in Fig. 20(b) shows a wide operating voltage range from 1.4 V down to 0.8 V.

Fig. 21 shows the static current consumption of a 6T SRAM and the proposed 2T eDRAM for different random cycle times. We assume a power-gated SRAM with a data retention voltage of 0.6 V. Supply voltage of the 2T eDRAM is assumed to be 1.1 V during hold mode. For very short random cycles (e.g. 1.5 ns), the static current of the proposed 2T eDRAM is much larger than that of the 6T SRAM due the frequent refresh operation required to maintain a good cell node voltage. However, for longer random cycle times, the RBL sensing margin of the 2T

eDRAM improves significantly which increases the retention time. For a 2.0 ns random cycle time, the proposed 2T eDRAM has an 81% and 91% smaller static current consumption than a power-gated SRAM [18] at 85 °C and 25 °C, respectively. The retention time of the proposed 2T eDRAM cannot be improved further for cycle times longer than 2.0 ns cycle time as shown in Fig. 20(a). The maximum achievable retention time is set by the data window shown in Fig. 18 and the variability in the bit-cells and BL-S/A's. The random cycle and retention time of eDRAMs are highly dependent on the number of cells-per-BL. The proposed 2T eDRAM has 16 times more cells on the same RBL than previous 1T1C eDRAMs [6], [7] and 4 times more cells than a previous 2T PMOS eDRAM [12]. The measured random cycle time with 512 cells-per-BL was 1.5 ns (667 MHz) which is a 33.4% improvement compared to previous eDRAM designs while achieving a retention time similar to 1T1C eDRAMs. For a random cycle of 500 MHz, the measured retention time is >120X longer than a previous 2T PMOS eDRAM and around 12X longer than a 1T1C eDRAM.

V. Conclusion

Several circuit techniques have been presented for improving data retention time and enhancing performance of gain cell eDRAMs. The proposed asymmetric 2T gain cell keeps the critical data '1' level close to VDD to improve memory performance and reduce static power dissipation. The proposed pseudo-PMOS diode based C-S/A eliminates the RBL leakage, provides better impedance matching, and offers more voltage headroom than previous designs. The half swing WBL scheme with a tri-state buffer achieves a 33% faster write speed and a 25% smaller WBL charging current without affecting the retention characteristics. Finally, a stepped WWL control scheme reduces the current drawn from the boosted supply by 67% which results in a 4.3% reduction in memory array area due to the smaller charge pump circuit and decoupling capacitors. Measurement results show a 667 MHz random cycle using a 110 μs refresh period for a 99.9% bit yield at 1.1 V, 85 °C. The static power dissipation including refresh currents and cell leakages was 109 μW/Mb at 500 MHz, 1.1 V, 85 °C which is 81% smaller than a power gated SRAM under a data retention voltage of 0.6 V.

References

[1] S. Rusu, S. Tam, H. Muljono, J. Stinson, and D. Ayers *et al.*, "A 45 nm 8-core enterprise Xeon processor," *IEEE J. Solid-State Circuits*, vol. 45, no. 1, pp. 7–14, Jan. 2010.

[2] S. Rusu, S. Tam, H. Muljono, D. Ayers, and J. Chang *et al.*, "A 65-nm dual-core multithreaded Xeon processor with 16-MB L3 cache," *IEEE J. Solid-State Circuits*, vol. 42, no. 1, pp. 17–25, Jan. 2007.

[3] R. J. Riedlinger, R. Bhatia, L. Biro, B. Bowhill, and E. Fetzer *et al.*, "A 32 nm 3.1 billion transistor 12-wide-issue Itanium processor for mission-critical servers," in *IEEE ISSCC Dig. Tech. Papers*, 2011, pp. 84–85.

[4] R. Kalla, B. Sinharoy, W. J. Starke, and M. Floyd, "POWER7: IBM's next-generation server processor," *IEEE Micro*, vol. 30, no. 2, pp. 7–15, Mar.–Apr. 2010.

[5] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, and D. Murray *et al.*, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 895–901, Apr. 2005.

[6] J. Barth, W. R. Reohr, P. Parries, G. Fredeman, and J. Golz *et al.*, "A 500 MHz random cycle, 1.5 ns latency, SOI embedded DRAM macro featuring a three-transistor micro sense amplifier," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 86–95, Jan. 2008.

[7] S. Romanovsky, A. Katoch, A. Achyuthan, C. O'Connell, and S. Natarajan *et al.*, "A 500 MHz random-access embedded 1 Mb DRAM macro in bulk CMOS," in *IEEE ISSCC Dig. Tech. Papers*, 2008, pp. 270–271.

[8] P. J. Klim, J. Barth, W. R. Reohr, D. Dick, and G. Fredeman *et al.*, "A 1 MB cache subsystem prototype with 1.8 ns embedded DRAMs in 45 nm SOI CMOS," *IEEE J. Solid-State Circuits*, vol. 44, no. 4, pp. 1216–1226, Apr. 2009.

[9] J. Barth, D. Plass, E. Nelson, C. Hwang, and G. Fredeman *et al.*, "A 45 nm SOI embedded DRAM macro for the POWER processor 32 MByte on-chip L3 cache," *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 64–75, Jan. 2011.

[10] M. Ichihashi, H. Toda, Y. Itoh, and K. Ishibashi, "0.5 V asymmetric three-Tr. cell (ATC) DRAM using 90 nm generic CMOS logic process," in *Proc. VLSI Circuits Symp.*, 2005, pp. 366–369.

[11] W. K. Luk, J. Cai, R. H. Dennard, M. J. Immediato, and S. V. Kosonocky, "A 3-transistor DRAM cell with gated diode for enhanced speed and retention time," in *Proc. VLSI Circuits Symp.*, 2006, pp. 184–185.

[12] D. Somasekhar, Y. D. Ye, P. Aseron, S. L. Lu, and M. M. Khellah *et al.*, "2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 174–185, Jan. 2009.

[13] K. Chun, P. Jain, J. Lee, and C. H. Kim, "A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches," *IEEE J. Solid-State Circuits*, vol. 46, no. 6, pp. 1495–1505, Jun. 2011.

[14] K. Chun, P. Jain, T. Kim, and C. H. Kim, "A 1.1 V, 667 MHz random cycle, asymmetric 2T gain cell embedded DRAM with a 99.9 percentile retention time of 110 μsec," in *Proc. VLSI Circuits Symp.*, 2010, pp. 191–192.

[15] E. Seevinck, P. J. van Beers, and H. Ontrop, "Current-mode techniques for high-speed VLSI circuits with application to current sense amplifier for CMOS SRAM's," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 525–536, Apr. 1991.

[16] J. Sim, H. Yoon, K. Chun, H. Lee, and S. Hong *et al.*, "A 1.8-V 128-Mb mobile DRAM with double boosting pump, hybrid current sense amplifier, and dual-referenced adjustment scheme for temperature sensor," *IEEE J. Solid-State Circuits*, vol. 38, no. 4, pp. 631–640, Apr. 2003.

[17] K. Agarwal and S. Nassif, "The impact of random device variation on SRAM cell stability in sub-90-nm CMOS technologies," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 16, no. 1, pp. 86–97, Jan. 2008.

[18] C. H. Kim, J. Kim, I. Chang, and K. Roy, "PVT-aware leakage reduction for on-die caches with improved read stability," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 170–178, Jan. 2006.

Ki Chul Chun received the B.S. degree in electronics engineering from Yonsei University, Seoul, Korea, in 1998 and the M.S. degree in electrical engineering from KAIST, Daejeon, Korea, in 2000. Since 2007, he has been working towards the Ph.D. degree in the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis.

He joined the Memory Division, Samsung Electronics, Hwasung, Korea, in 2000, where he has been involved in DRAM circuit design. His research interests include digital, mixed-signal and memory circuit designs with special focus on DRAM, PRAM, and STT-MRAM in scaled technologies.

Mr. Chun is the recipient of a Samsung Ph.D. Scholarship for outstanding employees and an ISLPED 2009 Low Power Design Contest Award.

Pulkit Jain received the B.Tech degree in electrical engineering from the Indian Institute of Technology (IIT) Kanpur in 2007. For the M.S, his research involved power delivery issues in three-dimensional integrated circuits. He is currently a Ph.D. student at the Department of Electrical Engineering, University of Minnesota, Minneapolis, where he is working on circuit techniques to monitor aging and variation in circuit design.

Mr. Jain is the recipient of an IBM scholarship award and 10+ authored/coauthored journal and conference papers.

Tae-Ho Kim received the B.S. degree in computer engineering and the M.S. degree in electrical engineering from the University of Minnesota, Minneapolis, MN, in 2007 and 2010, respectively.

He joined the Future Device R&D Lab., LG Electronics, Seoul, Korea, in 2010, where he has been engaged in the algorithm development and its FPGA implementation regarding on 3D display. His research interests include circuit design for pen touch application and FPGA implementation for the algorithm related with gesture recognition and 3D touch.

Chris H. Kim (M'04–SM'10) received the B.S. and M.S. degrees from Seoul National University, Seoul, Korea, and the Ph.D. degree from Purdue University, West Lafayette, IN.

He spent a year at Intel Corporation where he performed research on variation-tolerant circuits, on-die leakage sensor design and crosstalk noise analysis. In 2004, he joined the electrical and computer engineering faculty at the University of Minnesota, Minneapolis, MN, where he is currently an Associate Professor.

Prof. Kim is the recipient of an NSF CAREER Award, a Mcknight Foundation Land-Grant Professorship, a 3M Non-Tenured Faculty Award, DAC/ISSCC Student Design Contest Awards, IBM Faculty Partnership Awards, an IEEE Circuits and Systems Society Outstanding Young Author Award, ISLPED Low Power Design Contest Awards, an Intel Ph.D. Fellowship, and the Magoon's Award for Excellence in Teaching. He is an author/coauthor of 100+ journal and conference papers and has served as a technical program committee member for numerous circuit design conferences. He was the technical program committee chair for the 2010 International Symposium on Low Power Electronics and Design (ISLPED) and the guest editor for a special issue of the *IEEE Design and Test Magazine*. His research interests include digital, mixed-signal, and memory circuit design in silicon and non-silicon (organic and magnetic) technologies.