Single-Supply 3T Gain-Cell for Low-Voltage Low-Power Applications

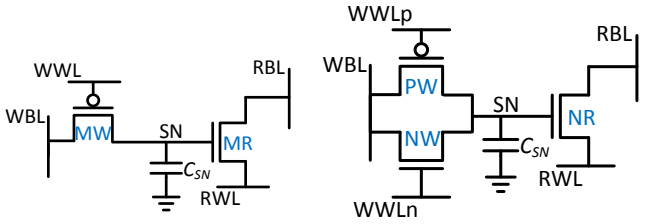
Robert Giterman, *Student Member, IEEE,* Adam Teman, *Member, IEEE,* Pascal Meinerzhagen, *Member, IEEE,* Lior Atias, *Student Member, IEEE,* Andreas Burg, *Member, IEEE,* and Alexander Fish *Member, IEEE*

***Abstract*—Logic compatible gain cell (GC) embedded DRAM (eDRAM) arrays are considered an alternative to SRAM due to their small size, non-ratioed operation, low static leakage, and 2-port functionality. However, traditional GC-eDRAM implementations require boosted control signals in order to write full voltage levels to the cell to reduce the refresh rate and shorten access times. These boosted levels require either an extra power supply or on-chip charge pumps, as well as non-trivial level shifting and toleration of high voltage levels. In this paper, we present a novel, logic compatible, 3T GC-eDRAM bitcell that operates with a single supply voltage and provides superior write capability to conventional GC structures. The proposed circuit is demonstrated with a 2 kb memory macro that was designed and fabricated in a mature 0.18 µm CMOS process, targeted at low-power, energy-efficient applications. The test array is powered with a single supply of 900 mV, showing an 0.8 ms worst-case retention time, a 1.3 ns write-access time, and 2.4 pW/bit of retention power. The proposed topology provides a bitcell area reduction of 43%, as compared to a redrawn 6T SRAM in the same technology, and an overall macro area reduction of 67% including peripherals.**

I. Introduction

In recent years, memories have occupied increasingly large portions of the die area of VLSI systems-on-chip (SoCs), in general, and of microprocessors, in particular, as shown in the latest ITRS edition [1]. This is due to the large 6-Transistor (6T) SRAM bitcell and its area-consuming peripheral circuitry that are the basis for the vast majority of these memories. In addition, the standby power of ultra-low power (ULP) systems, such as biomedical implants and wireless sensor networks, is often dominated by embedded memories which continue to leak during the long retentive standby periods that characterize these systems.

6T SRAM has been the traditional choice for the implementation of embedded memories due to its high-access speed and refresh-free static data retention. However, the 6T bitcell has several drawbacks in modern systems, including its large transistor count, its impeded functionality under voltage scaling, and the aforementioned static leakage currents from the supply voltage (V_{DD}) to GND. One of the interesting alternative implementations that addresses these limitations, while continuing to provide full CMOS logic compatibility, is gain-cell (GC) embedded DRAM (eDRAM), such as the circuit illustrated in Fig. 1(a) [2]–[5]. Most often consisting of two (2T) or three (3T) transistors, GC-eDRAMs provide a reduced silicon footprint, along with inherent 2-port functionality, non-ratioed circuit operation, and very low static leakage currents from V_{DD} to GND. However, as opposed to static memories, such as SRAM, the data retention of GC-eDRAM depends on dynamically stored charge, and thereby requires periodic, power-hungry refresh operations.



(a) 2T Mixed Gain-Cell (b) Proposed 3T Gain-Cell

Fig. 1. Schematic representations of conventional and proposed Gain-Cells.

The data retention time (DRT) of GC-eDRAMs is the time interval from writing a data level into the bitcell to the last moment at which one can still correctly read out the stored information. The DRT is primarily limited by the initial charge stored on the internal bitcell capacitance and the leakage currents that degrade the stored voltage level over time. For traditional 2T and 3T cells, the DRT is significantly affected by the initially degraded voltage level corresponding to data '0' or '1', due to the threshold voltage (V_{T}) drop across the write transistor (MW in Fig. 1(a)). In order to address this problem, a boosted write word line (WWL) voltage is usually employed to pass a full swing level to the storage capacitance. However, this requires the generation of a boosted on-chip voltage, which entails substantial overhead [6]. The magnitude of the voltage boost is set not only to overcome the V_{T} drop, but also to achieve short write-access times, which otherwise are typically longer than for 6T SRAM implementations. Furthermore, charge injection (CI) and clock feedthrough (CF) during WWL signal deassertion cause a voltage step at the storage node, resulting in an initally degraded level at the end of a write access [5]. As this undesired coupling even increases with WWL boost magnitude, a clear tradeoff between write speed, power, and DRT is introduced [5]. In addition, the level-shifting and toleration of higher than nominal voltages can be complex, especially when this boosted voltage is a negative underdrive voltage, as required by implementations employing a PMOS write transistor [2], [4]. The propagation of such a negative voltage can easily lead to voltage drops over device terminals that violate the technology limitations.

Contribution: In this paper, we present a new topology for a 3T gain cell, featuring a complementary transmission gate (TG) in the write port. While the proposed solution is quite straightforward, to our best knowledge, it is novel, and its impact is very high, as shown in the paper. The proposed bitcell provides strong initial data levels (both '1' and '0') for enhanced DRT and robust operation, as well as fast write-access times. This dual advantage is achieved without the need for additional voltages or boosted signals, allowing the use of standard peripheral circuitry for simple SoC integration and small silicon area. In order to demonstrate the functionality of the proposed bitcell, a 2 kb memory macro was designed and fabricated in a mature 0.18 µm CMOS node, which is typically used for ULP applications, such as biomedical sensor

Manuscript received XXX. This work was supported by the Swiss National Science Foundation under project number PP002-119057. A. Teman was supported by a Swiss National Government scholarship and P. Meinerzhagen was supported by an Intel Ph.D fellowship.

R. Giterman and A. Fish are with the Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel (e-mail: robertgi316@gmail.com).

A. Teman, P. Meinerzhagen, and A. Burg are with the Telecommunications Circuits Laboratory (TCL) of the Institute of Electrical Engineering, EPFL, Lausanne, VD, 1015 Switzerland (e-mail: {adam.teman,andreas.burg}@epfl.ch)

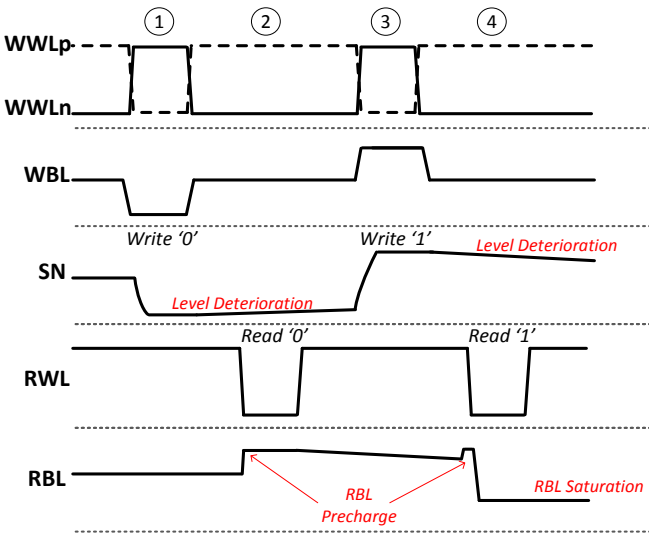


Fig. 2. Timing diagram of subsequent write and read operations: ① Write '0'; ② Read the stored '0'; ③ Write '1'; and ④ Read the stored '1'. Plots extracted from Spectre simulations with nominal parameter values.

nodes and implants. The resulting memory macro consumes only 33% of the area of a single-port 6T SRAM macro of the same size in the same technology node. The manufactured 3T GC-eDRAM macro is shown to be fully functional with a single supply voltage ranging from 600 mV to 1.8 V and a worst-case DRT of 0.8 ms at 900 mV, resulting in 4.9 nW/2kb retention power, which is 17× lower than a previously reported 6T SRAM at this voltage in the same technology [7].

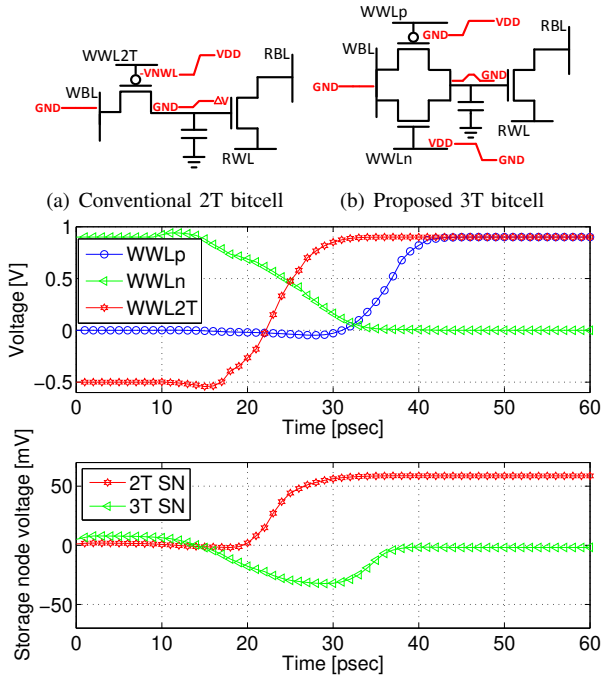
II. Proposed 3T Gain-Cell

A. 3T Gain-Cell Structure

Fig. 1(b) shows the schematic representation of the proposed single-supply 3T gain-cell. The circuit comprises a write port featuring a complementary TG (PW and NW), a read port based on an NMOS device (NR), and a storage node (SN) composed of the parasitic capacitance (C_{SN}) of the three devices and the stacked metal interconnect. The cell is built exclusively from standard V_T transistors and is fully compatible with standard digital CMOS technologies. The gates of PW and NW are connected to the complementary word lines, WWLp and WWLn. A common write bit line (WBL) is used to drive data through the TG during write operations. The full-swing passing capability of the TG enables the propagation of strong levels to the SN without the need for a boosted word line. Read is performed by precharging the read bit line (RBL) and subsequently driving the read word line (RWL) to GND, thereby conditionally discharging the RBL capacitance if the SN is high (data '1') or blocking the discharge path if the SN is low (data '0'). To achieve a reasonable tradeoff between speed, area, power and reliability, a dynamic sense inverter is used on the readout path (see Section III-A). However, other sense amplifiers can be used for improved read performance, such as demonstrated in [3], [8], [9].

B. 3T Gain-Cell Operation

Fig. 2 demonstrates bitcell operation through the application of subsequent write and read operations of both data values with V_{DD}=900 mV. This supply voltage was chosen as a good median voltage between V_{DD} and V_T, as previously shown to be DRT efficient in GC-eDRAM design [5]. Starting with a



(c) Waveform comparison of the two bitcells during write deassertion.

Fig. 3. Effects of charge injection and clock feed-through mechanisms.

charged C_{SN} (①), WBL is driven low and the word lines are asserted (WWLp=0 and WWLn=V_{DD}). As expected, a strong '0' level is passed to the SN, and this level is retained with the deassertion of the word lines due to the opposing CI and CF effects from the PW and NW transistors. During standby, the level on SN deteriorates due to leakage currents, dominated by the sub-V_T leakage of NW and PW in mature CMOS nodes. Therefore, in order to extend the retention time, WBL is driven to V_{DD}/2 during standby and read cycles, thereby significantly reducing the sub-V_T leakage through the TG, for both stored data '0' and '1', compared to the case where WBL is driven to either V_{DD} or GND. The WBL biasing circuitry is described in Section III-A.

During readout (②), the '0' level blocks the discharge path through NR, maintaining the precharged voltage on RBL. During the next write operation (③), WBL is driven high, resulting in a strong '1' stored on the SN. The subsequent read operation (④) provides a strong gate overdrive to transistor NR, thereby discharging RBL to read a '1'. It should be noted that during this operation (Read '1'), bitcells storing '1' and sharing the same column turn on when RBL discharges by more than the V_T of NR, causing it to saturate before it can completely discharge. This phenomenon is common to many GC-eDRAM configurations, as discussed in [5].

C. Comparison to Other Gain-Cell Implementations

A major advantage of the proposed cell over previous 2T and 3T gain cells is the self-dampening effect of CI and CF during write, as demonstrated in Fig. 3. For a 2T cell with a PMOS write transistor (Fig. 3(a)), CI and CF cause a significant positive voltage disturb on the SN during the rising edge of WWL, resulting in a degraded initial '0' level. However, in the proposed 3T cell (Fig. 3(b)), this problem is avoided due to the opposite transitions of the complementary write word lines (WWLn and WWLp), and the opposite polarity of charges injected into C_{SN} from the PMOS and NMOS write

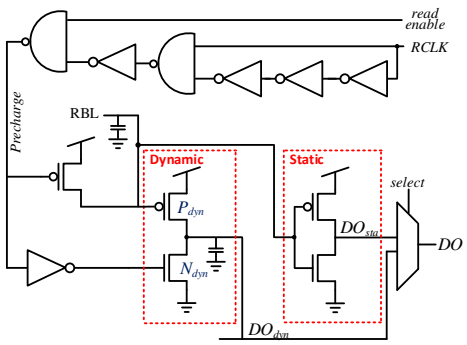


Fig. 4. Dynamic readout architecture.

transistors. This behavior is plotted in Fig. 3(c) for the two configurations with V_{DD}=900 mV and an underdrive voltage of –500 mV for the 2T in order to pass a full '0' level on the falling edge of WWL [5]. Whereas the 2T cell suffers from a degraded level of over 50 mV after WWL deassertion, the CI and CF effects of the complementary devices in the 3T cell essentially negate each other, resulting in a strong 0 V level for this cell. Note that there is still a dip in the 3T SN voltage due to the WWLn signal transitioning earlier than the WWLp signal. This could be avoided with more careful timing control. This strong level leads to advantages in both DRT, as well as read-access time, as the initial data levels are stronger.

III. Memory Macro Peripherals

The single supply voltage required for operation of the proposed topology simplifies the implementation of a full memory macro is a significant advantage over other GC-eDRAMs using conventional bitcells, which require the use of level shifters to create the desired boosted or negative (dependent on the type of write transistors) voltage supply. However, to further improve the array performance, in terms of access time and power consumption, several peripheral techniques were integrated into the designed memory macro. These peripheral circuits and techniques are presented in the following sub-sections.

A. Readout Circuitry

Many previous low voltage embedded memories, targeted at ULP systems, employ a simple sense inverter in order to provide robust, low area and low power data readout. However, such an inverter suffers from a very slow readout, as it requires the RBL to be discharged (charged for a PMOS read device) past the switching threshold of the inverter, which is hard to deviate away from V_{DD}/2. This operation is further impeded by the aforementioned RBL saturation during readout that slows down the discharging (charging) operation, as the RBL voltage decreases (increases). Therefore, the readout path that was integrated into the proposed 3T GC-eDRAM macrocell has two sensing modes: 1) a faster, yet potentially more error-prone, dynamic readout mode; and 2) a slightly slower, yet more reliable, static mode. In all measured prototype chips, both the dynamic (preferred for speed) and static modes were tested successfully. In both sensing modes, LVT PMOS transistors were used in order to allow a faster yet accurate read access time, due to the aforementioned issues. The supply voltage to the readout circuitry is gated with the read enable signal in order to save substantial static power due to the leaky LVT devices. The schematics of the two alternative readout circuits are shown in Fig. 4. The rising edge of the read clock (RCLK) creates a precharge pulse that charges the parasitic capacitance of RBL and discharges the output capacitance of the dynamic sense inverter (DO_{dyn}) through the discharge transistor, N_{dyn}. Subsequently, RBL is conditionally discharged during the read operation, turning on P_{dyn} to flip the output if a '1' is stored in the selected cell. Therefore, an RBL swing of only one V_{T} is required to complete a read operation. Transistor sizes for the dynamic sense inverter and pulse generator were chosen according to post-layout simulations under global and local parametric variations.

B. Write Circuitry

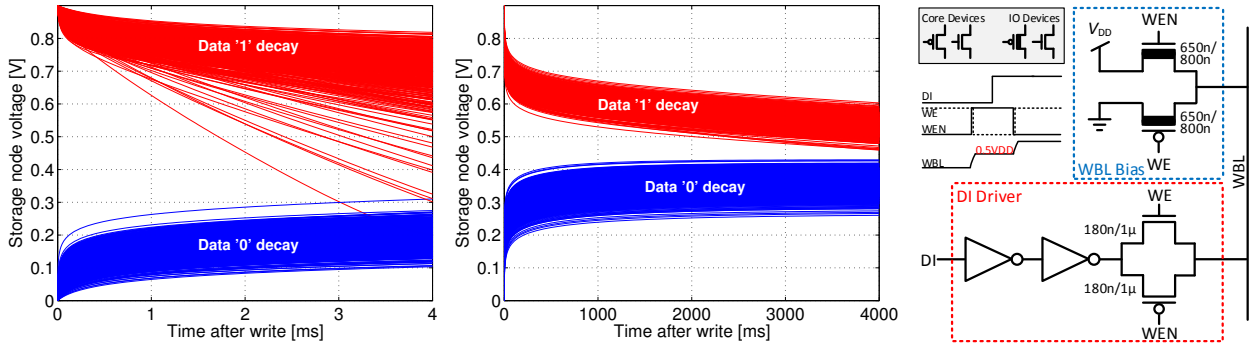
While the proposed single-supply 3T bitcell provides a significant improvement in both write time and initial SN level over standard GC implementations, the dual-transistor write port adds an additional leakage path to/from the SN. The increased aggregated sub-V_{T} current causes faster degradation of the stored charge, leading to reduced DRT, as compared to a reference 2T cell. In addition, several previous works have shown that the data dependent, asymmetric DRT (for data '0' and '1') of standard GCs can be manipulated to overall enhance DRTs by biasing the WBL at the best-case voltage for the weaker data level during standby and read operations [10]. For the proposed 3T configuration, the worst-case DRTs of the '1' and '0' levels are similar, and significant deterioration of the stored levels occurs for both extreme values of WBL bias (V_{DD} and GND). However, a neutral bias of V_{DD}/2 can be applied, thereby greatly reducing the sub-V_{T} current through the TG. Fig. 5 shows the benefit of this measure, displaying the level degradation of stored '1' and '0' data with WBL biases of opposite polarity and V_{DD}/2. With a WBL bias of V_{DD}/2, the DRT can be extended by approximately 1000×.

The write circuitry to implement the median WBL bias during standby and read cycles is shown in Fig. 5(c). A standard inverter chain conditionally drives the Data-in (DI) level on to the WBL through a TG, controlled by complementary write enable signals (WE and WEN). In parallel, a pair of long-channel I/O devices drive WBL during non-write cycles. These devices create a short-circuit path between V_{DD} and GND when WE is low, providing a median potential (V_{DD}/2) at WBL. Due to the thick oxides and long channel-lengths of the I/O devices, process variations are significantly reduced, and the static current is extremely low. Based on the chosen transistor sizes, the static power consumption of the proposed WBL driver during non-write cycles is only a few fW under a 900mV supply voltage, which is negligible compared to the refresh power of the array.

IV. Test Chip and Measurement Results

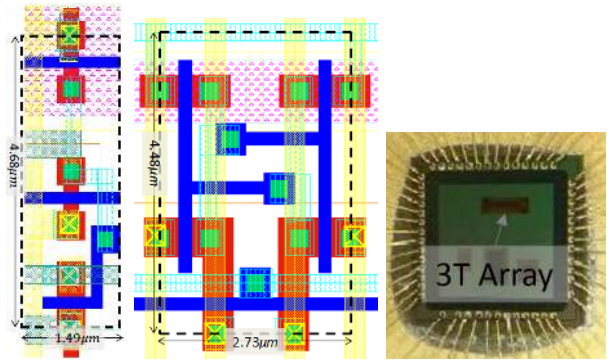
A. Implementation and Testing Procedure

A 64×32 bit (2kb) memory macro based on the proposed 3T gain cell was designed in a standard 0.18 µm CMOS process and integrated into a larger test chip with various test structures. All devices were implemented with standard V_{T} transistors to provide complete logic process compatibility without the need for additional process steps. Minimum sized write transistors were used in order to achieve a small bitcell area, while a slightly wider read transistor was employed in order to improve read-access time. The write word lines were



(a) SN degradation with worst-case bias conditions (b) SN degradation with $V_{\mathrm{DD}}/2$ WBL biasing (c) Write circuitry for $V_{\mathrm{DD}}/2$ biasing.

Fig. 5. Improved DRT through $V_{\mathrm{DD}}/2$ WBL biasing and the required circuitry. Plots include 1k Monte Carlo samples at 27°C with V_{DD}=0.9 V.



(a) (b) (c)

Fig. 6. (a) 3T Gain-Cell layout (b) Redrawn 6T SRAM layout (c) Test-chip micrograph

routed with horizontal polysilicon stripes in order to provide a dense bitcell, while other signals were routed on the first (lowest) three metal layers. Higher interconnect layers (metals 4 and 5) were tightly stacked above the bitcell in order to increase C_{SN}. The resulting bitcell, shown in Fig. 6(a), has an area of 6.97 µm^2 (1.49 µm×4.68 µm). This is 43% smaller than a 6T SRAM with the same design rules, shown in Fig. 6(b).

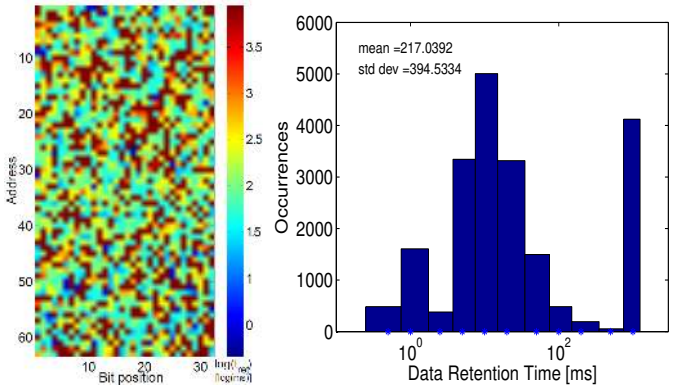
A micrograph of the fabricated test chip is shown in Fig. 6(c). In addition to the 2 kb array, the test chip includes a built-in self-test (BIST) unit for at-speed functionality tests, a 2 kb SRAM macro for storing per-bit data comparison results with expected responses, and several other test structures. The test chip was designed to enable three primary test modes: full, at-speed testing using the BIST; single-operation array control through scan chain configuration; and partial direct array access through external signals. These three test configurations were used to test the functionality of the array and provide the measurement data shown in the next sub-section. The GC array was biased by a separate, low-voltage supply (MV_{DD}), while the digital core supply (V_{DD}) was kept higher to ensure functionality of the BIST and digital control circuits.

The primary test starts by serially loading configuration registers and is followed by launching the BIST to enable at-speed functionality tests of the proposed macrocell with various data values, retention times, and operating frequencies. The general testing procedure of the BIST unit includes writing a predetermined data sequence into the array and then moving into an idle state for a configurable duration. In the idle state, special *write disturb* cycles can be activated to emulate the worst-case retention time biases. This is done by driving the WBLs to the opposite voltage of the data stored in the array, maximizing the sub-V_{T} leakage from/to the SN, emulating writing to different cells in the same column. The percentage of disturb cycles provides a means to measure the DRT under various access statistics, rather than just assuming the non-realistic worst-case of write-accesses during 100% of the time. Following this retention period, the array data is read out, compared with the initially written data, and stored in an on-chip 2 kb SRAM macro. Other test modes enable fine grain debugging, as well as flexible application of test sequences not planned for at design time.

B. Measurement Results

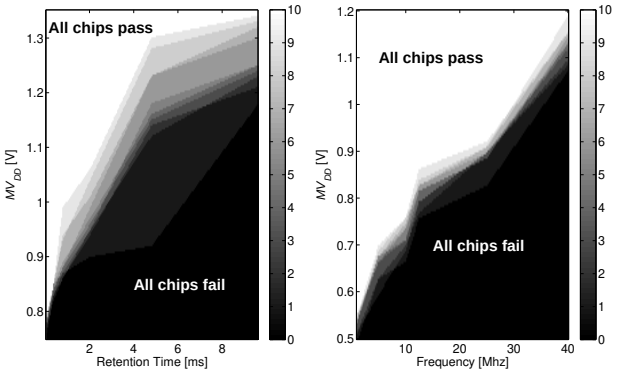
The test chips were packaged and measured using the test procedures described above. All 10 packaged chips were fully operational across the complete range of supply voltages (MV_{DD}) from 600 mV to 1.8 V. In order to study per-bit DRT, the entire array was written to a single value and then put into standby for a given time period (t_{ret}). Following this retention period, the array was read out and compared to the written data values. This procedure was repeated for several t_{ret} values and the measured bit DRT was calculated as the first t_{ret} that caused a read failure for each particular bit. In order to better emulate a typical operating scenario, a 5% write disturb was applied to the array. While the DRTs were measured for both all-'0' and all-'1' data levels, only all-'1' measurements are shown, since it was found to be the worst case of the two. This corresponds with Fig. 5, showing that data '1' degrades faster than data '0', while also requiring a longer sensing period for the chosen read circuitry. The resulting retention map for one of the measured chips is shown in Fig. 7(a). All measurements were taken at room temperature, which is considered a typical temperature for ULP applications that do not suffer from self-heating due to low computational complexity. DRT is shown on a $\log_{10}$ scale to better visualize the difference between cells, as it varies over several orders-of-magnitude. The lack of a systematic pattern shows that the difference between cells is primarily due to local process variations. The wide distribution of DRT is shown in Fig. 7(b) for all bitcells of ten measured dies (a total of 20,480 cells). The minimum and maximum DRTs were found to be 0.8 ms and 1978 ms respectively. The large spread and lack of systematic pattern correspond with previous studies [11], [12].

Fig. 8(a) shows a Shmoo plot of MV_{DD} vs. the measured DRT for 10 different chips. The grayscale map corresponds to the number of chips that were fully functional for the respective V_{DD} and targeted DRT. The gray levels are a result of global (die-to-die) variations, affecting the DRT of the array



(a) (b)

Fig. 7. (a) DRT distribution of 20,480 3T gain cells. (b) Retention time map of a 2 kb 3T GC array with MV_{DD}=900mV



(a) Min. MV_{DD} vs. DRT (b) Min. MV_{DD} vs. frequency

Fig. 8. Minimum memory supply voltage Shmoo plots for 10 measured chips.

for a given MV_{DD}. Even though lowering MV_{DD} results in lower leakage currents from/to the SN, the margin between '1' and '0' levels decreases, resulting in a lower DRT. We note that if this range of DRTs is insufficient, previously reported techniques, such as body biasing [11] can be used.

The functionality of the memories at different frequencies is demonstrated in Fig. 8(b) for various MV_{DD} voltages, as measured for 10 different chips. The measurement was conducted by writing '1' to all bits of the array and reading them out after the previously measured maximum DRT. The operating frequency was swept for every MV_{DD} to provide the Shmoo plot. The array showed full functionality for all frequencies up to 40Mhz with a supply voltage of 1.2V. Measurements at higher frequencies were impossible due to limitations of the test setup, which was designed for low frequency ULP applications.

The memory frequency is limited by the read-access time, since during readout, RBL needs to discharge in order to flip the sense inverter. This operation strongly depends on the parasitic capacitance of the RBL, as well as the data held in the unselected cells sharing the same column. If any of these cells stores a '1', the RBL discharge saturates at around V_{DD}-V_{T}, since at that point the unselected NRs start conducting and counteract the discharging efforts of the selected bit. In addition, due to the fact that readout was performed after the maximum DRT for each V_{DD}, a worst-case scenario was measured, since the data in the cells was already deteriorated, making the drive current much lower compared to the case where strong data levels are stored in the cells. Raising V_{DD} clearly improves the read-access time due to a stronger drive current, which makes the sense inverters switch faster.

For a dynamic memory, the relevant metric for static power consumption is retention power, composed of the sum of leakage and refresh power. Retention power was measured for storing an entire array of '0' and '1' under supply voltages ranging from 600 mV to 1.2 V. Leakage power mainly consists of sub-V_{T} currents from the SN to the WBL, which is biased at MV_{DD}/2 during standby periods. The measured retention power was found to be 4.9 nW for MV_{DD}=900 mV and 1.01 µW for MV_{DD}=1.2 V, which is 17× lower than a previously reported 6T SRAM cell [7], operated at 1.2V in the same technology node (consuming 26 µW standby power with V_{DD}=1.8V). The refresh frequency for every MV_{DD} was selected according to previously measured minimum DRTs.

V. Conclusions

This paper proposes a novel 3T Gain-Cell eDRAM macro-cell targeted at ULP systems and providing high storage density. The proposed GC is operated from a single supply voltage, eliminating the need for boosted voltages, commonly found in prior-art implementations. The proposed cell exhibits faster write-access than conventional GC circuits, while minimizing CI and CF through effects, thereby increasing DRTs and reducing refresh power consumption. The cell area is only 57% of a redrawn 6T SRAM in the same technology, making it a suitable alternative to SRAM for low-power memories. A test-chip containing a 2 kb memory macro based on the proposed 3T GC was fabricated in a mature 0.18 µm CMOS technology and several chips were tested. Measurement results show full functionality at voltages ranging from 600 mv to 1.8 V with retention power as much as 17× lower than a previously reported 6T SRAM in the same technology node.

References

[1] "International technology roadmap for semiconductors - 2013 edition," 2013. [Online]. Available: http://www.itrs.net

[2] D. Somasekhar *et al.*, "2 GHz 2 Mb 2T gain cell memory macro with 128 GBytes/sec bandwidth in a 65 nm logic process technology," *IEEE JSSC*, vol. 44, no. 1, pp. 174–185, 2009.

[3] K. Chun *et al.*, "A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches," *IEEE JSSC*, 2012.

[4] Y. Lee *et al.*, "A 5.4nW/kB retention power logic-compatible embedded DRAM with 2T dual-VT gain cell for low power sensing applicaions," in *Proc. IEEE A-SSCC*, 2010.

[5] P. Meinerzhagen *et al.*, "Exploration of sub-VT and near-VT 2T gain-cell memories for ultra-low power applications under technology scaling," *MDPI JLPEA*, vol. 3, no. 2, pp. 54–72, 2013.

[6] A. Raychowdhury *et al.*, "PVT-and-aging adaptive wordline boosting for 8T SRAM power reduction," in *Proc. IEEE ISSCC*, 2010, pp. 352–353.

[7] G. Sreerama Reddy *et al.*, "Design and implementation of 8k-bits low power SRAM in 180nm technology," in *IMCECS*, vol. 2, 2009.

[8] K. C. Chun *et al.*, "A 3T gain cell eDRAM utilizing preferential boosting for high density and low power on-die caches," *JSSC*, 2011.

[9] K. C. Chun, W. Zhang *et al.*, "A 2T1C embedded DRAM macro with no boosted supplies featuring a 7T SRAM based repair and a cell storage monitor," *JSSC*, vol. 47, no. 10, pp. 2517–2526, 2012.

[10] A. Teman *et al.*, "Replica technique for adaptive refresh timing of gain-cell-eDRAM," *TCAS-II*, vol. 61, no. 4, pp. 259–263, April 2014.

[11] P. Meinerzhagen *et al.*, "Impact of body biasing on the retention time of gain-cell memories," *IET JoE*, vol. 1, no. 1, 2013.

[12] W. Zhang *et al.*, "A write-back-free 2T1D eDRAM with local voltage sensing and a dual-row-access LP mode," *TCAS-I*, vol. 60, no. 8, pp. 2030–2038, 2013.