A Write-Back-Free 2T1D Embedded DRAM with Local Voltage Sensing and a Dual-Row-Access Low Power Mode

Wei Zhang Ki Chul Chun Chris H. Kim
University of Minnesota, Minneapolis, MN 55455

Abstract- **A gain cell embedded DRAM (eDRAM) in a 65nm LP process achieves a 1.0 GHz random access frequency by eliminating the write-back operation. The read bitline swing of the 2T1D cell is improved by employing short local bitlines connected to local voltage sense amplifiers. A low-overhead dual-row access mode improves the worst-case cell retention time by 3X, minimizing refresh power at times when only a fraction of the entire memory is utilized. Measurement results from a 64kb eDRAM test chip in 65nm CMOS demonstrate the effectiveness of the proposed circuit techniques.**

I. INTRODUCTION

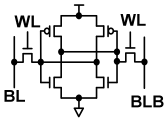
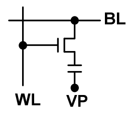
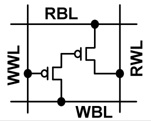
Embedded DRAM (eDRAM) technology has been drawing increasing attention in recent years as an alternative to mainstream 6T SRAM, since it delivers higher bit-cell density and a practical random access time. 1T1C eDRAM has already been adopted for last level caches of high performance server chips [1-2]. Despite successful deployment of 1T1C eDRAM in recent server products, the complicated process steps involved in building the storage capacitor and the special access transistor, coupled with the limited signal swing at low supply voltages, make the scaling of this eDRAM technology unfavorable.

Gain cell eDRAM is considered as a promising embedded memory option with the potential of overcoming the scaling challenges encountered by SRAM and 1T1C eDRAM. It provides decoupled read and write paths which improves low voltage margin while the cell size is approximately 2X denser than that of a 6T SRAM. Moreover, it is logic compatible and the separate read port enables non-destructive read and the capability of driving long bitline loads, making it competitive at low voltages. Table 1 compares the circuit parameters of interest for the three types of embedded memory.

Several recent gain cell eDRAM designs based on 2T or 3T cells have demonstrated practical retention times beyond 100 μsecs [5], SRAM-like performance [6-7], and true logic-compatibility by eliminating boosted voltages [7]. One interesting feature of gain cells that has been largely overlooked in the past is the potential for write-back-free operation by taking advantage of the non-destructive read.

In this work, we have experimentally demonstrated for the first time, a gain cell eDRAM without write-back operation. By removing the write-back from a read operation, the read access speed can be significantly improved. We also apply a local voltage Sense Amplifier (S/A) scheme to overcome the design complexities and variability issues prevalent in the existing current-sensing schemes used for 2T gain cells. Finally, a low-overhead low-power mode based on a dual-row-access scheme extends cell retention time by 3X to save refresh power during periods when only a fraction of the cache memory is being used.

TABLE 1. SRAM VERSUS EDRAM (1T1C AND GAIN CELL)

	6T SRAM [3]	1T1C eDRAM [1]	2T eDRAM [4]
Cell Schematic			
[(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)
[(2)]Redrawn cell size (ratio)	0.575x2.05= 1.179μm^2 (1X)	0.45x0.545= 0.245μm^2 (0.21X)	0.48x0.925= 0.444μm^2 (0.38X)
Low-VDD margin	Poor (ratioed)	Poor (destructive read)	Good (non-ratioed, gain function)
Storage cap.	irrelevant	<10fF	~1fF
Process	Logic compatible	Trench cap. + thick TOX access TR	Logic compatible
Retention time	NA	40μs @105ºC, 99.99%	10μs @25ºC (target) *<1μs @105ºC (Est.)*
Random cycle	1ns[(2)]	2ns	2ns
Static power	1.0X	0.2X	NA

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

II. WRITE-BACK-FREE READ OPERATION

A write-back-free read operation provides us a system speedup opportunity which has been neglected in previous gain cell designs [5-7]. In conventional 1T1C eDRAMs, the read operation relies on charge sharing between the storage capacitance and the bitline capacitance. Due to the destructive read nature, a write-back is needed to reinforce the cell data after the charge sharing operation. Gain cells on the other hand, have a non-destructive read, eliminating the need for a write-back. Fig. 1 compares the cycle time between a 6T SRAM and various 2T eDRAMs implemented in the same 65nm low-power process. It shows that, by eliminating the write-back delay that accounts for approximately 30% of the read cycle, a significant improvement in both operating frequency and active power can be achieved. Although these benefits are only applicable to the read cycle, it is sufficient to improve the overall system performance significantly, because in general there are more reads than writes in a processor cache, and a read may stall the system and therefore degrade the system level performance, while a write will not.

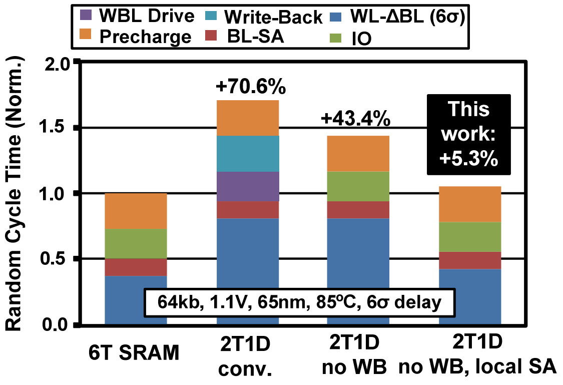


Fig.1. Random cycle time comparison between SRAM and eDRAM.

A 2T1D cell, shown in Fig. 2, was used in this work to demonstrate the write-back-free operation. The gain cell design is different from the previous 2T1C cell [7] in that the P-type coupling device shared between adjacent cells is replaced by a separate N-type diode in each cell. This provides the same beneficial coupling up effect [5] during read while minimizing any coupling noise from the adjacent cells. The PCOU signal preferentially couples up the cell node voltage for increased read current and preferentially couples down the cell voltage after sensing to restore the full voltage levels without requiring a boosted negative voltage. Further details on the circuit operation of the 2T1C cell can be found in [7]. Another major difference from [7] is that PCOU switching occurs within a single read cycle to allow consecutive reads without a write-back phase. Even though the write cycle contains a read operation, which is required due to the row-wise write nature of eDRAMs to avoid overwriting the data of unaccessed cells on the same row, we kept the write cycle timing intact so that the beneficial write feature described in [7] can be preserved.

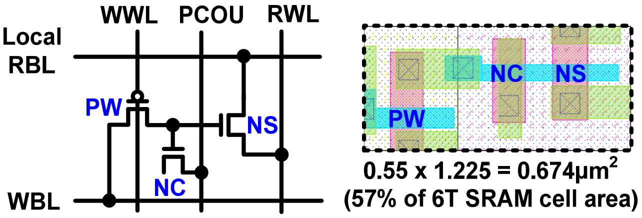


Fig.2. 2T1D gain cell with preferential boosting [5].

Simulation results in Fig. 3 examine the impact of write-back-free reads for access rates from 0.01% to 10%. The voltage window between data '1' and data '0' remains unchanged for access rates up to 1%. For access rates greater than 1%, the cell voltage window slightly improves due to the minute difference between the couple up and couple down voltages that gets accumulated over a long retention period (e.g. 200 µsec), which compensates for the pull-up leakages. Fig. 4 shows that for practical access rates, getting rid of the write-back does not have an adverse effect on the data '1' and data '0' levels. Our test chip design focuses on experimentally verifying the impact of write-back-free reads on overall eDRAM performance for practical access rates.

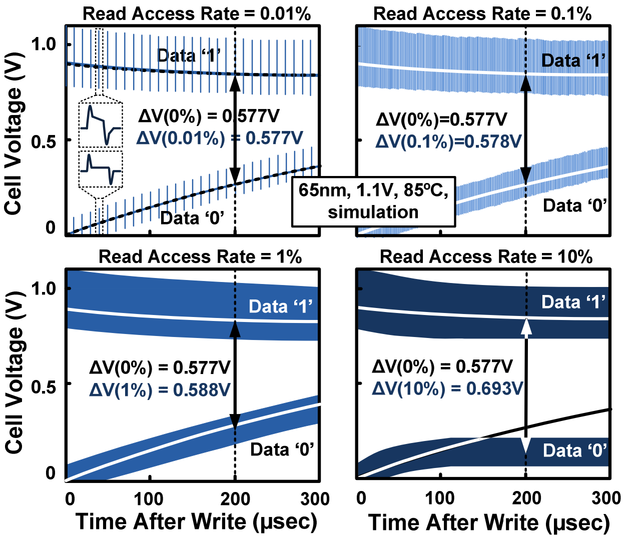


Fig.3. Cell retention characteristics without write-back for different read access rates. Coupling strength increases with cell voltage.

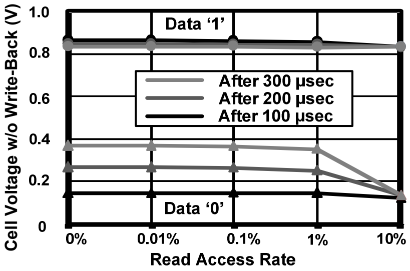


Fig.4. Cell voltage w/o write-back vs. read access rate.

III. LOCAL-SENSE-AMPLIFIER ARCHITECTURE

For 2T-based gain cell, despite the fast speed and compact cell size compared to its 3T counterpart, read disturbance has been a common issue [6] as shown in Fig. 5 (above). Current from the unselected cells storing data '1' limits the RBL swing to ~0.2V which is not sufficient enough for reliable voltage sensing. Common practice has been to implement a current-sensing scheme, which keeps the RBL level close to VDD during read [6-7]. However, a current-sensing scheme suffers

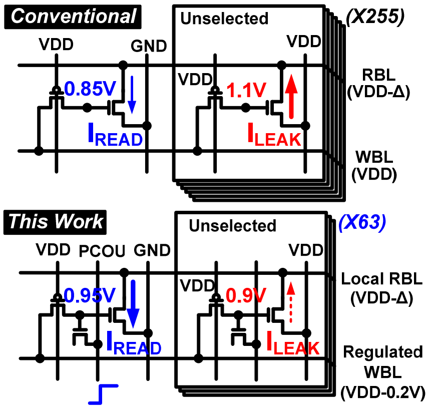


Fig.5. Read disturbance mitigation of 2T1D cell using (i) beneficial PCOU coupling, (ii) regulated WBL, and (iii) short local read bitlines.

from variation issues in the dummy cell due to the single-ended sensing and cannot utilize dummy averaging techniques because of the small input impedance requirement. This will introduce significant design overhead and scalability challenges in future technology nodes.

To get around this problem, we apply three circuit techniques that allow a more robust voltage-sensing scheme to be used. First, the beneficial PCOU coupling in the accessed 2T1D cell provides stronger pull-down read current. Second, a regulated WBL scheme [7] lowers the fresh data '1' voltage by 0.2V (1.1V→0.9V), significantly reducing the disturbance current. This has proven to have little impact on data retention since data '1' quickly stabilizes to around 0.85V due to the cell leakage profile [6]. Finally, a Local-Sense-Amplifier (LSA) scheme with short read bitlines [1] limits the maximum number of unselected cells to 63 which in turn reduces the worst case read disturbance current and provides a sufficient signal margin for reliable voltage sensing.

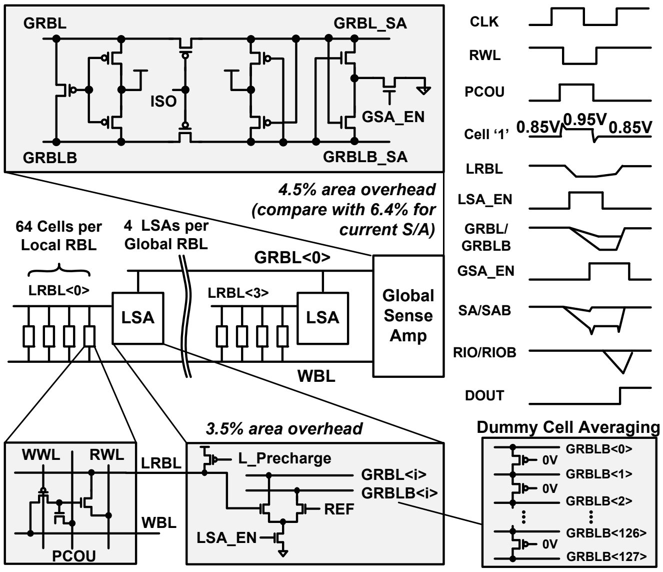


Fig.6. Schematic and timing of local and global sense amplifier architecture.

Fig. 6 shows the detailed architecture with local and global voltage S/A's. The reduced load of the local RBLs ensures a fast voltage development while a simple global voltage S/A further speeds up the signal propagation. Dummy cell averaging which was not possible in previous current-sensing schemes, can now be implemented to enhance the robustness under PVT variations. The area overhead compared to a conventional current-sensing scheme is only 1.6% due to the simpler voltage S/A circuit that partially offsets the LSA area overhead.

IV. DUAL-ROW-ACCESS LOW POWER MODE

For applications that do not utilize the entire cache memory space, shutting down parts of the array is a practical way to save power. For gain cell eDRAMs, however, activating multiple rows at the same time can yield greater power savings because the refresh power is determined by the worst case retention time of the tail cells. Fig. 7 shows our dual-row-access mode, where two wordlines with respective LSAs are enabled at the same time without changing the read reference circuit. The weak cell is thus repaired by a stronger one according to spatial randomness, and in addition, mismatch between the LSAs themselves gets averaged out. Moreover, the effective sensing current of the global read bitline (GRBL) is doubled, which improves the retention time even further under the same sensing window requirement and timing constraints. Therefore, the worst retention time can be improved by more than 2X using a dual-row-access scheme, while a simple powering-down approach may still suffer from the tail cell's retention time. Note that activating more than two rows at a time (e.g. 4-row-access, 8-row-access) has diminishing returns while incurring significant design overhead in the form of dedicated timing and reference circuitry. Therefore, we propose to use a simple dual-row-access mode.

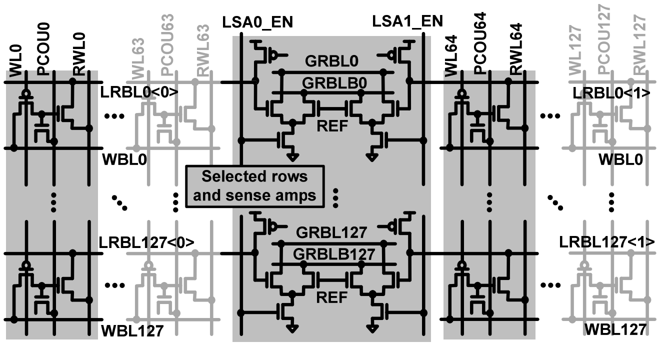


Fig.7. Dual-row access mode illustration (WL0 and WL64 selected).

V. EXPERIMENTAL DATA

A 64kb eDRAM test chip was implemented in a 1.2V, 65nm logic CMOS process to demonstrate the proposed circuit techniques. The chip achieves a 99.9% retention time of 325μsec and a refresh power of 234.1μW/Mb at 1.1V, 85ºC. Figs. 8 and 9 show the failure percentile and retention map from a 1kb sub-array, respectively, for a worst case read disturbance pattern and a 1.0 ns read cycle time. No noticeable changes in the retention time were observed across different access rates. Although our measurement setup supports only up to a 1% access rate, this is sufficient for real-world applications as indicated in Fig. 5. For a 99.99% bit yield, the

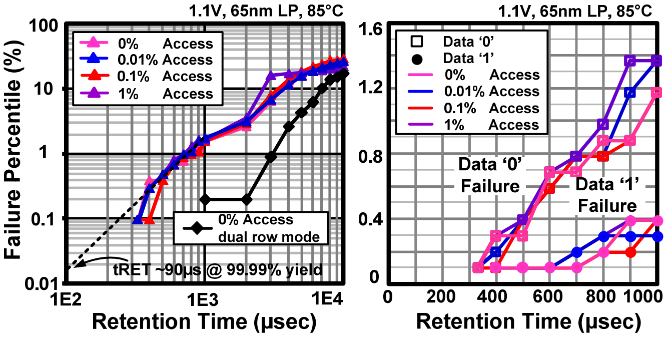


Fig.8. Failure percentiles for a 1kb sub-array (left) and detailed view of tail cells (right).

extrapolated retention time was 90µsec. Due to the dummy averaging feature of our proposed design, the measured retention map in Fig. 9 shows no significant signs of bitline dependency in the failure pattern which is in contrast to the results presented in our prior work [7].

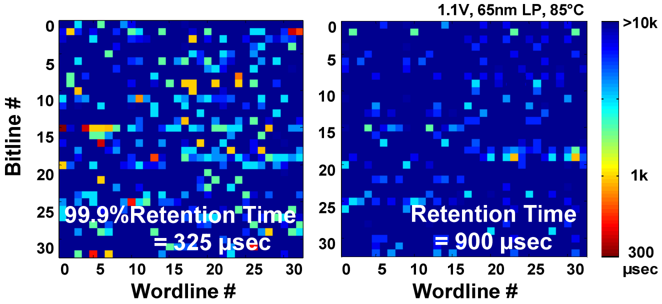


Fig. 9. Measured retention maps for single (left) and dual (right) row access modes. No significant bitline dependence is observed in either case.

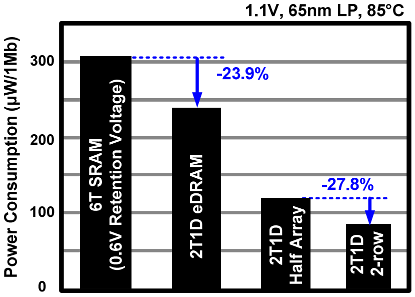


Fig.10. Power consumption comparison between a power gated SRAM with a 0.6V retention voltage and various 2T1D eDRAM power down modes.

Fig. 9 also shows that the worst case retention time improves from 325µsec to 900µsec using the dual-row-access mode. Fig. 10 compares the power dissipation of SRAM and various eDRAM configurations. The proposed 2T1D design (single row access) achieves a 23.9% power saving compared to that of a power gated 6T SRAM with a 0.6V retention voltage. Compared to a simple power down mode where half the eDRAM is unused, a 27.8% power reduction was achieved using the dual-row-access mode.

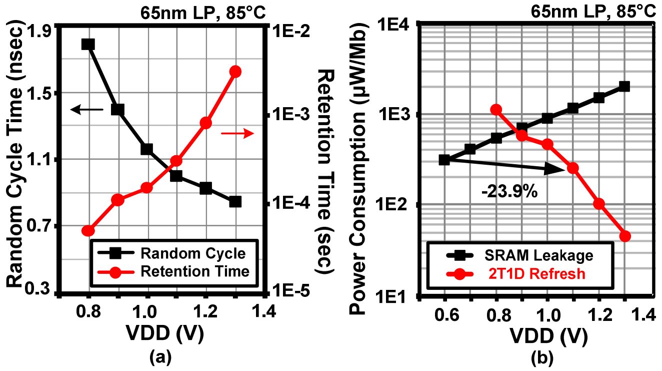
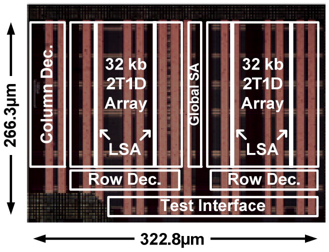


Fig.11. (a) Measured VDD shmoo and (b) static power comparison.

The measured VDD shmoo for retention time and cycle time is plotted in Fig. 11 (a), and a static power comparison between 6T SRAM and the proposed eDRAM design for different supply voltages is shown in Fig. 11(b). The longer retention time at higher supply voltages makes the eDRAM refresh power to be lower than the static power of an SRAM. Note that the optimal supply voltage of an eDRAM is usually higher than that of an SRAM due to the refresh power dominating the overall static power consumption. Fig. 12 shows the die microphotograph and summarizes the key features.



Process	65nm LP CMOS
Main feature	Write-back-free read
Cell size	57% of 6T SRAM
Array size	64 kb (512WLx128BL)
99.9% retention time	325µsec @ 1.1V, 85ºC
Rd. Cycle	1.0ns @ 1.1V, 85ºC
Wr./Ref. Cycle	1.5ns @ 1.1V, 85ºC
Cell Availability	99.5% @ 1Mb, 1.1V, 85ºC, 325 µsec refresh
Static Power	234.1 µW/Mb @ 1.1V, 85ºC, 325 µsec refresh

Fig. 12. Microphotograph and summary of test chip characteristics.

VI. CONCLUSION

We have presented several circuit techniques to enhance the performance and robustness of gain cell eDRAM. The proposed design was the first to experimentally verify write-back-free read operation in gain cells. No noticeable retention time difference was observed across a wide range of access rates. We also proposed various circuit techniques for mitigating read disturbance issues including a local-sense-amplifier scheme. In addition, a dual-row-access low power mode was introduced to further reduce static power in scenarios where less than half the cache is being utilized. Test chip measurements were presented from a 64kb eDRAM array implemented in a 1.2V, 65nm CMOS process.

REFERENCES

[1] J. Barth, et al., “A 500 MHz random cycle, 1.5 ns latency, SOI embedded DRAM macro featuring a three-transistor micro sense amplifier,” *IEEE Journal of Solid-State Circuits*, Vol. 43, Issue 1, pp. 86-95, 2008.

[2] J. Barth, et al., “A 45 nm SOI embedded DRAM macro for POWER7^{TM} processor 32 MB on-chip L3 cache,” *IEEE Journal of Solid-State Circuits*, Vol. 46, Issue 1, pp. 64-75, 2011.

[3] K. Zhang, et al., “SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction,” *IEEE Journal of Solid-State Circuits*, Vol. 40, Issue 4, pp. 895-901, 2005.

[4] 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 Journal of Solid-State Circuits*, Vol. 44, Issue 1, pp. 174-185, 2009.

[5] W. Luk, J. Cai, R. Dennard, M. Immediato, S. Kosonocky, “A 3-transistor DRAM cell with gated diode for enhanced speed and retention time,” *VLSI Circuits Symposium*, pp. 184-185, 2006.

[6] K. Chun, P. Jain, T. Kim, C. Kim, “A 1.1V, 667MHz random cycle, asymmetric 2T gain cell embedded DRAM with a 99.9 percentile retention time of 110µsec,” *VLSI Circuits Symposium*, pp. 191-192, 2010.

[7] K. Chun, W. Zhang, P. Jain, C. Kim, “A 700MHz 2T1C embedded DRAM macro in a generic logic process with no boosted supplies,” *International Solid-State Circuits Conference*, pp. 506-507, 2011.