Replica Bit-Line Technique for Embedded Multilevel Gain-Cell DRAM

Muhammad Umer Khalid, Pascal Meinerzhagen, and Andreas Burg
Telecommunications Circuits Laboratory, EPFL, Lausanne, Switzerland
Email: engrumer24@gmail.com, pascal.meinerzhagen@epfl.ch, andreas.burg@epfl.ch

***Abstract*—Multilevel gain-cell DRAMs are interesting to improve the area-efficiency of modern fault-tolerant systems-on-chip implemented in deep-submicron CMOS technologies. This paper addresses the problem of long access times in such multilevel gain-cell DRAMs, which are further aggravated by process parameter variations. A replica bit-line (BL) technique, previously proposed for SRAM, is adapted to speed up the multilevel read operation at a negligible area-increase. Moreover, the same replica column is used to improve the write access time. An 8-kb DRAM macro implemented in 90-nm CMOS technology shows that the replica column is able to successfully track die-to-die process, voltage, and temperature variations to generate control signals with optimum delay. Finally, Monte-Carlo simulations show that a small timing margin of 100 ps is sufficient to also cope with within-die process variations.**

I. Introduction

Embedded memories consume an increasingly dominant part of the overall area of system-on-chip (SoC) designs [1]. At the same time, there is a trend to fault-tolerant VLSI systems [2,3] due to increasing process variations and high defect levels in deep-submicron (DSM) CMOS technologies. To ensure area-efficiency in future fault-tolerant VLSI SoCs, it is thus interesting to trade the reliability of embedded memories for higher storage densities. This tradeoff can be achieved by storing many levels (more than one bit) in a single DRAM cell [4].

Replacing conventional 1-transistor 1-capacitor (1T1C) DRAM cells with gain-cells leads to memory macros which are fully compatible with standard digital CMOS technologies [5], reduces cost, and allows for non-destructive read operation. An 8-kb multilevel gain-cell DRAM implemented in 90-nm CMOS technology has roughly half the area of a corresponding single-port SRAM macro, at the cost of a small percentage of read failures, due to within-die (WID) process variation, and limited retention time [5].

Among different multilevel write schemes summarized in [6], charge sharing between bit-line (BL) segments to locally generate many data levels has a small area cost. Moreover, the multilevel read operation is best performed in a sequential fashion to avoid an area-increase due to the need for many parallel sense amplifiers (SAs). However, these area-efficient multilevel write and read schemes result in long access times. This problem is aggravated in DSM CMOS technologies where large timing margins are required due to increasing process variations if reliability is not to be further compromised.

In order to guarantee reliable sense operation and to yet trigger the SA at the earliest possible instant, even in the occurrence of large die-to-die (D2D) process, voltage, and temperature (PVT) variations, different flavors of replica BL techniques have been developed for SRAMs [7–9]. Some of these techniques do also address WID process parameter variation [8,9]. The basic replica BL technique consists of a delay generator (the replica BL) which tracks the delay of the actual BLs across PVT corners [7]. To our knowledge, the replica BL technique has not been exploited yet to improve the access times of multilevel gain-cell DRAMs.

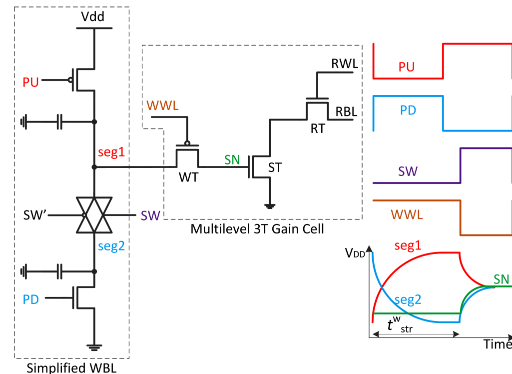


Fig. 1. Multilevel write operation.

Contribution: In this paper, the replica BL technique is applied to multilevel gain-cell DRAM to maintain optimum read access times under PVT variations with a minimum area-overhead. In addition to generating read control signals, the same replica column is also used to generate write control signals with optimum delay.

II. Operation of Multilevel Gain-Cell DRAM

This section reviews the multilevel write and read operation, emphasizing the critical timing delays which are to be tracked over PVT corners.

A. Multilevel Write Operation

The considered multilevel gain-cell DRAM [5] stores 4 levels (2 bits) per gain-cell. These 4 storage levels, as well as 3 reference levels for sensing are generated locally by charge sharing among pre-charged and pre-discharged write bit-line (WBL) segments, connected by transmission-gate switches [10,11]. This level generation technique is particularly area-efficient as it relys on already existing hardware, i.e., the WBL segments. To obtain the correct capacitance ratios for the generation of all levels while maintaining regularity in the storage array, the WBLs are partitioned into 12 equal segments [5].

Fig. 1 exemplifies the multilevel write operation for a given level. First, the switches are set to form two coherent WBL segments with the desired capacitance ratio from the 12 original WBL segments. Shortly after this, the pull-up (PU) and pull-down (PD) devices are enabled to pre-(dis)charge the just formed capacitors to V_{DD} and 0 V, respectively. The delay of this pre-(dis)charge process is the most significant contribution to the write access time. As soon as both capacitors are completely pre-(dis)charged, both segments are detached from the supply rails and the switch, which has been separating

them, is closed to initiate the charge sharing process. At the same time, the write word-line (WWL) signal is enabled to open the write access transistor (WT) of the addressed gain-cell, thereby allowing the generated level to be transferred to its storage node (SN). For an accurate level generation, it is crucial that both capacitors are completely pre-(dis)charged and that the PU/PD devices are completely turned off before the switch between the two capacitors is closed.

For the generation of the highest storage level of 1.1 V, 11 WBL segments need to be pre-charged to V_{DD}, which amounts for the longest possible pre-(dis)charge delay. As indicated on the right-hand side of Fig. 1, $t^{\mathrm{w}}_{\mathrm{str}}$ is defined as the time it takes to charge 11 WBL segments from ground to $0.99\times V_{\mathrm{DD}}$. The replica column will be designed to track $t^{\mathrm{w}}_{\mathrm{str}}$.

B. Multilevel Read Operation

During a multilevel read operation, one data level needs to be compared to different reference levels, which can be done sequentially or in parallel [6]. In this work, a sequential multilevel read scheme is adopted, due to its small area cost. The stored data level is thus compared to 2 different reference levels, where the second reference level is chosen depending on the outcome of the first sense operation.

Fig. 2 shows one gain-cell from the storage array, attached to one side of the SA through the read bit-line (RBL). A reference gain-cell, dedicated to holding the reference levels, is attached to the other side of the SA. Both sense operations (together forming one complete read access) start by writing the corresponding reference level to the reference gain-cell, with the same charge-sharing technique as used for write accesses. Once the reference level is set, the RBLs are pre-charged to V_{DD} and equalized (shorted). After this, the read word-line (RWL) signals of both gain-cells are enabled at exactly the same time, which causes the RBLs to be discharged through the gain-cells. The different voltage levels on the SN of the gain-cells result in unequally strong discharging currents, which eventually develops a voltage difference between the terminals of the SA. The SA is triggered as soon as this voltage difference is big enough to overcome its offset voltage.

It is crucial to trigger the SA at the right time: triggered too early, the voltage difference might be too small to be resolved correctly; triggered too late, both RBLs might already have been discharged completely to ground. Finding a suitable trigger instant is especially difficult since there are many different voltage levels resulting in stronger or weaker discharging currents. The problem is further aggravated by PVT variations.

Implemented in a 90-nm CMOS technology, the SA shown in Fig. 2 has an offset voltage of up to 30 mV. As indicated on the right-hand side of Fig. 2, the RBL-discharge delay $t^{\mathrm{r}}_{\mathrm{str}}$ is defined as the required time to discharge a RBL from V_{DD} to $0.45\times V_{\mathrm{DD}}$ through the read path of a gain-cell storing the highest data level. For the highest data and reference levels, a voltage difference between the RBLs of 107 mV is developed within a time $t^{\mathrm{r}}_{\mathrm{str}}$, whereas the voltage difference that develops for the lowest data and reference levels is with 71 mV still high enough for reliable sensing. The replica column will be tracking $t^{\mathrm{r}}_{\mathrm{str}}$, allowing for triggering the SAs at an early yet safe instant for any PVT condition.

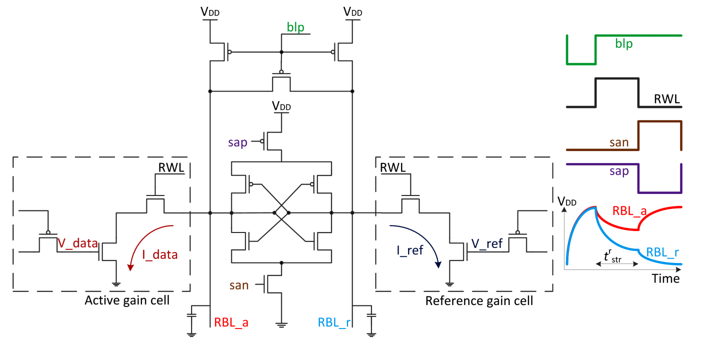


Fig. 2. Multilevel read operation.

III. Replica Column Design

As discussed in the previous section, the multilevel write and read operations consist of a sequence of consecutive events and some of the involved control signals, shown in Fig. 1 and Fig. 2, must be non-overlapping. In a very conservative approach resulting in low performance, the control signals are generated by a finite state machine, allocating an entire clock cycle to each event or to each break between events to avoid overlapping signals [5]. A better approach consists in designing dedicated delay lines, using inverter chains or other conventional delay elements, to generate the required delay for each control signal. However, PVT variations do not affect the delay of such conventional delay elements in the same way as they affect the BL-(dis)charge delays [7]. Due to the poor tracking, conservative timing margins are required which leads to long access times. The replica BL technique improves access times since the much better tracking allows for significantly reduced timing margins, without further compromising the reliability, and at only a small area cost.

To best track the BL-(dis)charge delays in the storage array, the replica column is in principle kept as similar as possible to an actual column in the storage array. However, detecting the previously discussed voltage levels of $0.99\times V_{\mathrm{DD}}$ and $0.45\times V_{\mathrm{DD}}$ would require voltage comparators and voltage references. To minimize the area overhead, the replica column is slightly modified to detect $t^{\mathrm{w}}_{\mathrm{str}}$ and $t^{\mathrm{r}}_{\mathrm{str}}$ using only inverters with a switching threshold of $V_{\mathrm{DD}}/2$.

As shown in Fig. 3, the replica column consists of a replica WBL (rWBL) and a replica RBL (rRBL) and of the associated feedback-based control-signal generation circuits.

A. Replica Write Bit-Line

The rWBL generates the *Rep_wt* signal which disables the PU and PD devices and initiates the charge sharing process in the WBLs of the storage array. As shown in Fig. 4, *Rep_wt* has a delay of $t^{\mathrm{w}}_{\mathrm{rep}}$, which is designed to track $t^{\mathrm{w}}_{\mathrm{str}}$. A positive clock edge starts the charging process of the pre-discharged rWBL. As soon as the voltage on the last rWBL segment becomes larger than the switching threshold of buffer b1, which is engineered to be $V_{\mathrm{DD}}/2$, the *Rep_wt* signal is asserted.

A PU device with reduced drive strength in the rWBL and an increased load of all 12 rWBL segments (instead of the worst case of only 11 WBL segments in the storage array) ensure that $t^{\mathrm{w}}_{\mathrm{rep}}$ always exceeds $t^{\mathrm{w}}_{\mathrm{str}}$ by a small safety margin, even with buffer b1 already switching at $V_{\mathrm{DD}}/2$.

In order to prepare the rWBL for the next write operation, it must be discharged again, which is achieved in a short time

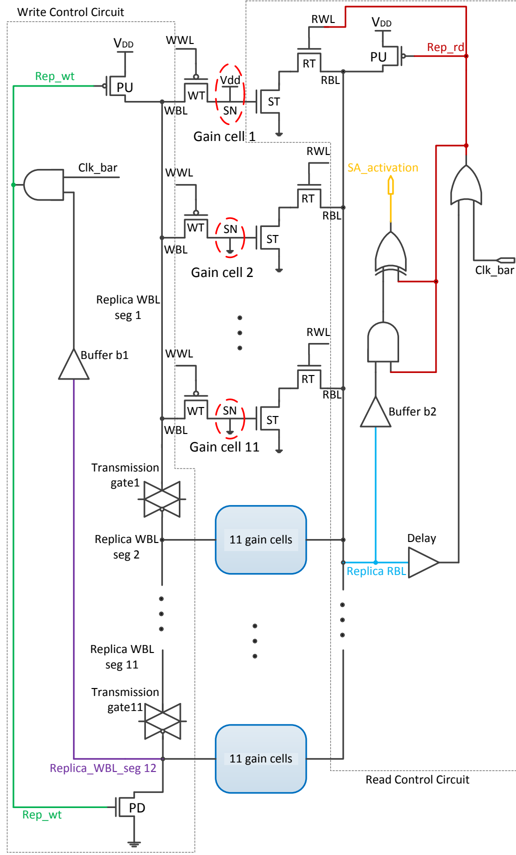


Fig. 3. Schematic of the replica column.

by using a wider PD transistor and a wider nMOS transistor in the corresponding transmission-gates, compared to the storage array.

B. Replica Read Bit-Line

The rRBL generates the control signal *SA_activation*, from which the signals *san* and *sap* (cf. Fig. 2) activating the SAs are derived. As shown in Fig. 4, *SA_activation* has a delay $t^{\mathrm{r}}_{\mathrm{rep}}$, which is designed to track $t^{\mathrm{r}}_{\mathrm{str}}$. Both $t^{\mathrm{r}}_{\mathrm{rep}}$ and $t^{\mathrm{r}}_{\mathrm{str}}$ are referred to the instant where the RBLs and the rRBL start discharging.

At the beginning of the clock period, the rRBL is rapidly precharged through a strong PU device together with the RBLs in the storage array. Once the rRBL reaches the trip-point ($V_{\mathrm{DD}}/2$) of the first connected feedback path, and after an additional short delay ensuring that all RBLs (rRBL and actual RBLs) have reached V_{DD}, the control signal *Rep_rd* is asserted, indicating that the array is ready for evaluation.

Upon assertion of *Rep_rd*, the rRBL and all RBLs in the storage array start to discharge at exactly the same time. The rRBL is discharged through only one gain-cell in the replica column, referred to as the replica gain-cell. As highlighted in Fig. 3, the SN of this replica gain-cell is connected to V_{DD}, which results in a slightly stronger discharging current than the strongest one in the storage array (the highest voltage level is 1.1 V, while V_{DD} is 1.2 V). To compensate for this, and to continue working with trip-points of $V_{\mathrm{DD}}/2$, the transistors in the read path of the replica gain-cell are made slightly weaker than in the actual gain-cells. By means of a second feedback loop, *SA_activation* is asserted as soon as the voltage on the

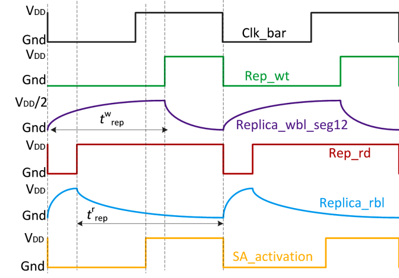


Fig. 4. Waveforms of main control signals.

rRBL reaches the switching threshold ($V_{\mathrm{DD}}/2$) of buffer b2.

IV. Robustness Analysis

This section verifies the effectiveness of the replica column in firmly tracking $t^{\mathrm{w}}_{\mathrm{str}}$ and $t^{\mathrm{r}}_{\mathrm{str}}$ over P(D2D)VT corners and that the small remaining timing margins are sufficient to cope with WID variations.

A. Immunity to PVT Variations

As shown in Fig. 5, $t^{\mathrm{w}}_{\mathrm{rep}}$ (initiation of charge sharing) closely tracks $t^{\mathrm{w}}_{\mathrm{str}}$ (storage WBLs precharged to $0.99\times V_{\mathrm{DD}}$) over a) D2D process parameter variations, b) supply voltage variations, and c) temperature variations. For D2D process variations, 5 different corners are considered, corresponding to different combinations of slow (S), typical (T), and fast (F) nMOS (N) and pMOS (P) devices. The supply voltage is varied from -10% to $+10\%$ of its nominal value (1.2 V). Finally, temperature variations from -25 to 125 °C are applied. Fig. 5 also shows that the small safety margin between $t^{\mathrm{w}}_{\mathrm{rep}}$ and $t^{\mathrm{w}}_{\mathrm{str}}$ is always around 100 ps.

Fig. 6 shows that an excellent delay tracking is also achieved for the control signals involved in the read operation: $t^{\mathrm{r}}_{\mathrm{rep}}$ (activation of SA) closely tracks $t^{\mathrm{r}}_{\mathrm{str}}$ (storage RBLs discharged to $0.45\times V_{\mathrm{DD}}$) over the aforementioned large P(D2D)VT variations. Moreover, the separation between $t^{\mathrm{r}}_{\mathrm{rep}}$ and $t^{\mathrm{r}}_{\mathrm{str}}$ is always around 50 ps.

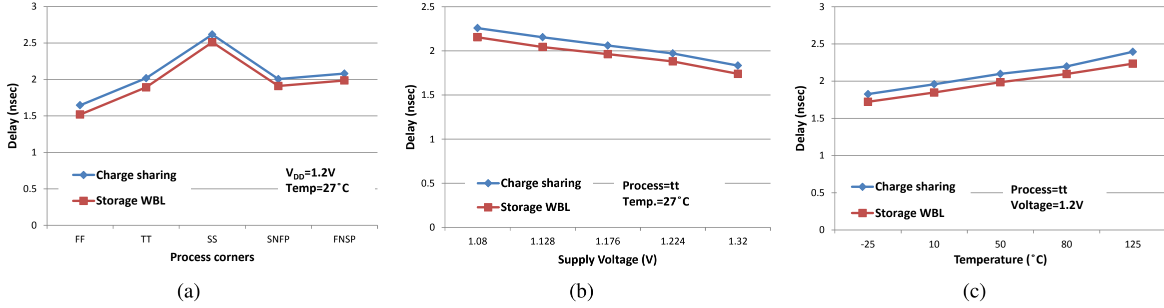
B. Mismatch

While the proposed replica BL technique has been proven to be immune to P(D2D)VT variations, it is still affected by WID process parameter variations (or parameter mismatch between transistors in the storage array and in the replica column). For an accurate storage and reference level generation, $t^{\mathrm{w}}_{\mathrm{rep}}$ should be longer than $t^{\mathrm{w}}_{\mathrm{str}}$. Fig. 7(a) shows the statistical distribution of $\Delta t_{\mathrm{w}} = t^{\mathrm{w}}_{\mathrm{rep}} - t^{\mathrm{w}}_{\mathrm{str}}$ in the occurence of WID variation, resulting from 1000 Monte-Carlo trials, assuming typical P(D2D) conditions. The mean and standard deviation of Δt_{w} are around 100 and 50 ps, showing that a very small safety margin is sufficient for a reasonably high yield and only very few negative occurrences of Δt_{w}.

For the read operation, it does not matter if $t^{\mathrm{r}}_{\mathrm{rep}}$ or $t^{\mathrm{r}}_{\mathrm{str}}$ is longer. However, to trigger the SAs at a safe instant, the standard deviation of $\Delta t_{\mathrm{r}} = t^{\mathrm{r}}_{\mathrm{rep}} - t^{\mathrm{r}}_{\mathrm{str}}$ should be small. Fig. 7(b) shows the 1k-point statistical distribution of Δt_{r}, for WID variation under typical P(D2D)VT conditions. With a standard deviation of 12 ps, Δt_{r} is always reasonably small compared to the RBL-discharge delays of more than 1 ns.

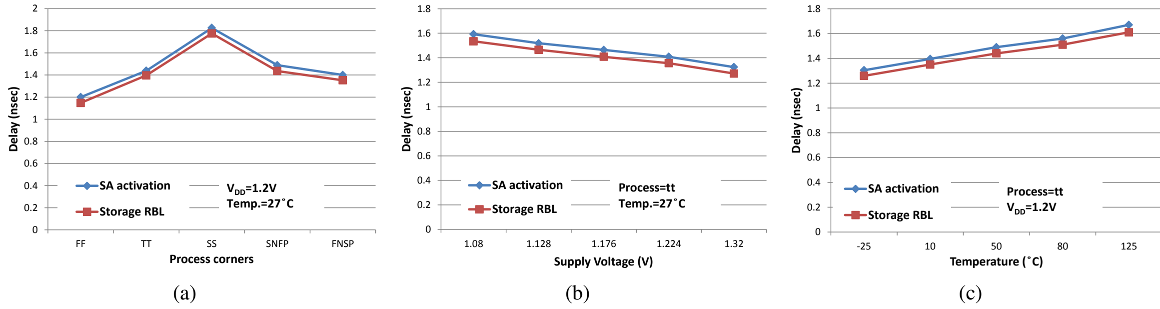
V. Implementation Results

Table I shows the total write access time t_{write} and the total read access time t_{read}, including the time required for address decoding, for fast (FF, 1.32 V, 10 °C), typical (TT,



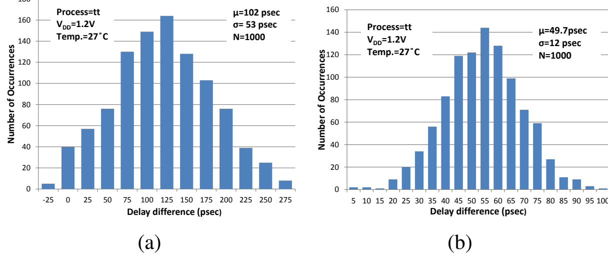
(a) (b) (c)

Fig. 5. Write operation: delay $t^{\mathrm{w}}_{\mathrm{str}}$ (storage WBLs precharged to $0.99 \times V_{\mathrm{DD}}$) variability caused by (a) die-to-die process, (b) supply voltage, and (c) temperature variation and its tracking $t^{\mathrm{w}}_{\mathrm{rep}}$ (initiation of charge sharing).



(a) (b) (c)

Fig. 6. Read operation: delay $t^{\mathrm{r}}_{\mathrm{str}}$ (storage RBLs discharged to $0.45 \times V_{\mathrm{DD}}$) variability caused by (a) die-to-die process, (b) supply voltage, and (c) temperature variation and its tracking $t^{\mathrm{r}}_{\mathrm{rep}}$ (activation of SA).



(a) (b)

Fig. 7. Distribution of (a) $\Delta t_{\mathrm{w}} = t^{\mathrm{w}}_{\mathrm{rep}} - t^{\mathrm{w}}_{\mathrm{str}}$, and (b) $\Delta t_{\mathrm{r}} = t^{\mathrm{r}}_{\mathrm{rep}} - t^{\mathrm{r}}_{\mathrm{str}}$.

1.20 V, 27 °C), and slow (SS, 1.08 V, 80 °C) PVT conditions. Note that $t_{\mathrm{read}} = 4 \times t_{\mathrm{write}}$, as a read access consists of two write accesses to a reference gain-cell (1 clock cycle each), each write access being followed by a sense operation (1 clock cycle each).

TABLE I
TOTAL ACCESS TIMES FOR DIFFERENT PVT CONDITIONS

PVT condition	t_{write} [ns]	t_{read} [ns]
Fast	2.3	9.2
Typical	3.0	12.0
Slow	5.0	20.0

The implemented replica BL technique provides savings in the write access time of 2.7 ns for fast PVT conditions, and 2.0 ns for typical PVT conditions, compared to a design with fixed timing margins which guarantee accurate level generation even for slow PVT conditions. Similarly, the savings in read access time are 10.8 ns and 8.0 ns for fast and typical PVT conditions, respectively. More importantly, the replica BL technique is much safer than using fixed timing margins, as it finds an appropriate SA trigger instant for each PVT condition.

VI. CONCLUSIONS

Area-efficient multilevel gain-cell DRAMs implemented in DSM CMOS technologies have high access times, due to large timing margins required to deal with PVT variations. In order to improve both the read and the write access times, the proposed replica bit-line technique generates optimum control signals for the multilevel write and read operations by firmly tracking bit-line charge/discharge delays in the storage array over large PVT variations.

Monte Carlo simulations show that the remaining small timing margins of 100 ps are sufficient to deal with mismatch. The write and read access times are improved up to 2.7 and 10.8 ns compared to a design with fixed timing margins. Delay tracking is also essential to trigger the sense amplifiers at a safe instant of time for any given PVT condition.

ACKNOWLEDGMENT

This work was kindly supported by the Swiss National Science Foundation under the project number PP002-119057.

REFERENCES

[1] "ITRS," 2009. [Online]. Available: http://www.itrs.net
[2] M. Breuer, "Let's think analog," in *Proc. IEEE CSAS on VLSI*, May 2005.
[3] S. Ghosh and K. Roy, "Parameter variation tolerance and error resiliency: New design paradigm for the nanoscale era," *Proc. IEEE*, Oct. 2010.
[4] Y. Xiang *et al.*, "Design of a multilevel DRAM with adjustable cell capacity," in *Proc. IEEE CCECE*, 2001.
[5] P. A. Meinerzhagen *et al.*, "Design and failure analysis of logic-compatible multilevel gain-cell-based DRAM for fault-tolerant VLSI systems," in *Proc. IEEE/ACM GLSVLSI*, 2011.
[6] J. C. Koob *et al.*, "Design and characterization of a multilevel DRAM," *IEEE JVLSIS*, Sep. 2011.
[7] B. Amrutur and M. Horowitz, "A replica technique for wordline and sense control in low-power SRAM's," *IEEE JSSC*, Aug. 1998.
[8] U. Arslan *et al.*, "Variation-tolerant SRAM sense-amplifier timing using configurable replica bitlines," in *Proc. IEEE CICC*, Sep. 2008.
[9] S. Komatsu *et al.*, "A 40-nm low-power SRAM with multi-stage replica-bitline technique for reducing timing variation," in *Proc. IEEE CICC*, Sep. 2009.
[10] B. Cockburn *et al.*, "A multilevel DRAM with hierarchical bitlines and serial sensing," in *Proc. IEEE MTDT*, Jul. 2003.
[11] P. Gillingham, "A sense and restore technique for multilevel DRAM," *IEEE TCSII*, Jul. 1996.