Two-Port Low-Power Gain-Cell Storage Array: Voltage Scaling and Retention Time

Rashid Iqbal, Pascal Meinerzhagen, and Andreas Burg
Institute of Electrical Engineering, EPFL, Lausanne, VD, 1015 Switzerland
Email: rasiq992@gmail.com, pascal.meinerzhagen@epfl.ch, andreas.burg@epfl.ch

***Abstract*—The impact of supply voltage scaling on the retention time of a 2-transistor (2T) gain-cell (GC) storage array is investigated, in order to enable low-power/low-voltage data storage. The retention time can be increased when scaling down the supply voltage for a given access statistics and a given write bit-line (WBL) control scheme. Moreover, for a given supply voltage, the retention time can be further increased by controlling the WBL to a voltage level between the supply rails during idle and read states. These two concepts are proved by means of Spectre simulation of a GC-storage array implemented in 180-nm CMOS technology. The proposed 2-kb storage macro is operated at only 40 % of the nominal supply voltage and leverages the GCs to enable two-port operation with a negligible area-increase compared to a single-port implementation.**

I. Introduction

Biomedical implants, wireless sensor networks, and a large variety of other battery-powered handheld devices have a stringent power budget. Low-power processors such as [1,2] are often a key component for such devices. Also Intel has recently presented an experimental near threshold-voltage (NTV) microprocessor [3].

Embedded memories consume an increasingly dominant part of the overall power (and area) of system-on-chip (SoC) designs in general [4] and processors in particular [5]. Supply voltage scaling is an efficient low-power technique which reduces both active energy dissipation and leakage power [6]. However, when gradually scaling down the supply voltage, conventional embedded memory implementations such as 6-transistor (6T)-bitcell SRAM start failing before logic circuits do [6,7]. Embedded memories which operate reliably at scaled supply voltages are therefore key in achieving energy-efficiency in future SoC and microprocessor designs.

Specially designed SRAM operates reliably at scaled supply voltages and even in the subthreshold domain at the price of relatively large 8-transistor (8T) [8] or 10-transistor (10T) [9] bitcells. Latch arrays and flip-flop arrays are a more straightforward approach to reliable low-voltage storage macros but have an even larger area cost for storage capacities higher than a few kb [10].

In conventional 1-transistor-1-capacitor (1T-1C) embedded DRAM (eDRAM), the offset voltage of the sense amplifier limits voltage downscaling, unless dedicated offset cancellation techniques are used [11]. Another major obstacle in low-voltage eDRAM is the degradation of the data retention time, which requires power-consuming refresh operations more frequently [11]. Furthermore, conventional eDRAMs require special process options to build high-density 3D capacitors, which adds cost to standard digital CMOS technologies.

As a further option for building embedded low-voltage storage arrays, gain-cells (GCs) are smaller than any SRAM bitcell, latches, and flip-flops, while they are fully compatible with standard digital CMOS technologies. Recently, various research groups proposed GC-based storage arrays as denser successor of SRAM for on-die caches in processors [5,12]. A dual threshold-voltage (dual-V_{th}) GC storage array [13] is operated at a fraction of the nominal supply voltage; the circuit increases the retention time by using a high threshold-voltage (high-V_{th}) write access transistor (WT). Another storage macro based on a boosted 3-transistor (3T) GC [14] is operable in a supply voltage range from 1.2 down to 0.7 V and uses preferential storage node boosting at the time of reading to increase the retention time (and the read speed).

Previously reported GC storage macros are not clearly classified as either single-port or two-port implementations. Furthermore, while previous work on GC storage arrays targets a given supply voltage (or supply voltage range) and presents dedicated techniques to increase the retention time, the impact of supply voltage scaling on the retention time has not been systematically investigated yet. Moreover, previous publications do not clearly state the assumed write access statistics for the measurement of the retention time, while frequent write accesses may in fact significantly degrade retention time.

Contribution: This work reviews why GCs are inherently suitable for two-port memory implementations with a negligible area-overhead compared to single-port implementations. The fundamental limit to supply voltage scaling in 2-transistor (2T) GC storage arrays in the occurrence of process parameter variation is then discussed. Next, the impact of supply voltage downscaling on the retention time under well-defined access statistics is investigated, allowing for finding the optimum supply voltage for lowest power consumption and highest retention time. Finally, a simple technique to further improve the retention time at any given supply voltage is presented.

II. Gain-Cell Array Architecture

A. Two-Port Implementation

Concurrent read/write access is an effective method for achieving high memory bandwidth [15]. Two-port memories have a separate read and write port to enable such access. In conventional 1T-1C DRAM and conventional SRAM, the same word-lines (WLs) and bit-lines (BLs) are used for both the read and the write operation; enabling two-port operation is non-trivial and requires additional hardware in each cell. As opposed to this, GCs are inherently well suited for two-port operation, as they already have a separate read port consisting of the read-WL (RWL) terminal and the read-BL (RBL) terminal as well as a separate write port consisting of the write-WL (WWL) terminal and the write-BL (WBL) terminal, as shown in Fig. 1. It is therefore straightforward

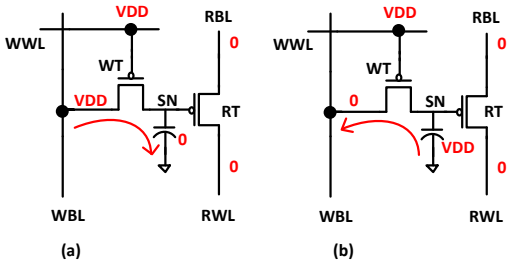

Fig. 1. 2-PMOS Gain-Cell. Worst WBL-state for retention of (a) logic '0' and (b) logic '1'.

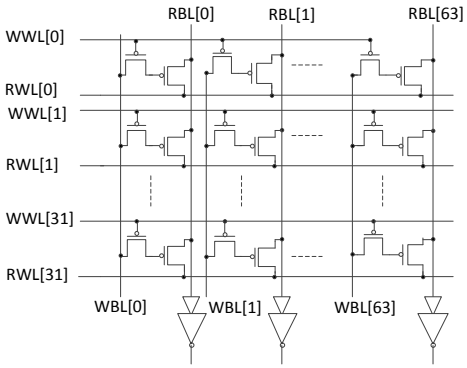

Fig. 2. Storage array with sense inverters.

to enable two-port operation in GC-based storage arrays and benefit from the resulting high memory bandwidth.

In the two-port memory architecture adopted in this work, there are two address decoders: one for the write address, and another one for the read address. A single-port implementation would save one address decoder, but it would require additional logic circuits—comparable in size to a single decoder—to distribute the decoded address to either the write port or the read port, while silencing the other port.

B. Array and Gain-Cell Implementation

Apart from the explicit two-port configuration, the memory architecture serving as a basis for the presented analyses is mostly adopted from [13]. As shown in Fig. 2, the storage array consists of 32 rows and 64 columns. Moreover, the conventional sense amplifiers are replaced with simple sense inverters to improve area-efficiency [13]. To allow for conclusions as general as possible, the basic 2-PMOS GC with regular threshold-voltage (regular-V_{th}) transistors from [12] is adopted in this work, as the high-V_{th} transistors used in [13] might not be available in all technologies. Notice, however, that high-V_{th} transistors may reduce subthreshold conduction by more than 2 orders of magnitude compared to regular-V_{th} transistors [13], and therefore allow for considerably longer retention times.

III. Operation Principle

A. Hold, Write, and Read Operations

In each cell, data is stored in form of charge on the storage node (SN) capacitor, which is formed by the gate capacitance of the storage/read transistor (RT) and junction/wire parasitic capacitance. The parasitic SN capacitor is explicitly shown in Fig. 1.

During a write operation, the WT of the selected GC is turned on to transfer the new data level from the WBL to the SN. To allow the transfer of a clean logic '0', an underdrive voltage of $-500\,\mathrm{mV}$ is applied to the selected WWL.

At the beginning of a read operation, all RBLs are discharged to ground. Next, the selected RWL is pulled high to V_{DD}. If a GC stores a logic '1', its RT remains off and the connected RBL remains at ground. However, if the GC stores a logic '0', the RBL starts to charge through the RT. The sense inverter must switch before RBL is charged to the threshold voltage of RT ($V_{\mathrm{th}}^{\mathrm{RT}}$), as at this time RTs in unselected cells storing logic '0' turn on, which provides a current path to ground and prevents a further voltage rise on the RBL.

B. Fundamental Limit to Supply Voltage Scaling

The minimum supply voltage is determined by the ability of writing, holding, and reading two distinct data levels. Considering the 2-PMOS GC and avoiding any underdrive voltage, the WT can easily transfer a high voltage level equal to V_{DD}. However, the lowest data level which can be transferred in a reasonable time, i.e., not relying on subthreshold conduction, is equal to the threshold-voltage of WT ($V_{\mathrm{th}}^{\mathrm{WT}}$). When turning off the WT, charge injection and clock feedthrough rise the voltage on the SN (V_{SN}) by ΔV_{SN}, which depends on the SN capacitance, the voltage level being transferred, and many other factors. After writing a logic '0' level, $V_{\mathrm{SN}} = V_{\mathrm{th}}^{\mathrm{WT}} + \Delta V_{\mathrm{SN}}$. Holding a data level on the SN during a small amount of time is possible regardless of V_{DD}. To tell a logic '0' from a logic '1' at the time of reading, V_{SN} must be smaller than $V_{\mathrm{DD}} - V_{\mathrm{th}}^{\mathrm{RT}}$ in order to still be able to turn on the RT:

$$V_{\mathrm{th}}^{\mathrm{WT}} + \Delta V_{\mathrm{SN}} < V_{\mathrm{DD}} - V_{\mathrm{th}}^{\mathrm{RT}} \tag{1}$$

Equation (1) is rearranged to show the lower limit for V_{DD}:

$$V_{\mathrm{th}}^{\mathrm{WT}} + V_{\mathrm{th}}^{\mathrm{RT}} + \Delta V_{\mathrm{SN}} < V_{\mathrm{DD}} \tag{2}$$

To account for process parameter variations (die-to-die and within-die variation), Equation (2) is rewritten as follows, where $\mu(X)$ and $\sigma(X)$ denote the mean and the standard deviation of the random variable X.

$$\left(\mu(V_{\mathrm{th}}^{\mathrm{WT}}) + N\sigma(V_{\mathrm{th}}^{\mathrm{WT}})\right) + \left(\mu(V_{\mathrm{th}}^{\mathrm{RT}}) + N\sigma(V_{\mathrm{th}}^{\mathrm{RT}})\right) + \Delta V_{\mathrm{SN}} < V_{\mathrm{DD}} \tag{3}$$

The parameter N is chosen depending on the desired yield. For small storage arrays of several kb, $N = 3$ is reasonable.

Assuming a WWL underdrive, a clean ground level can be transferred to the SN, and V_{DD} can be further reduced, with its lower limit now given by:

$$\left(\mu(V_{\mathrm{th}}^{\mathrm{RT}}) + N\sigma(V_{\mathrm{th}}^{\mathrm{RT}})\right) + \Delta V_{\mathrm{SN}} < V_{\mathrm{DD}} \tag{4}$$

It is usually beneficial in terms of energy to have a WWL underdrive, as most parts of the circuit can be operated from a lower V_{DD}, while the underdrive voltage is only applied to the write address decoder and the WWL drivers.

In the current case, using an underdrive voltage of $-500\,\mathrm{mV}$, and with $\mu(V_{\mathrm{th}}^{\mathrm{RT}}) = 500\,\mathrm{mV}$, $\sigma(V_{\mathrm{th}}^{\mathrm{RT}}) = 25\,\mathrm{mV}$, $N = 3$,

$\Delta V_{\mathrm{SN}} \approx 100\,\mathrm{mV}$, and a small margin for uncertainty in ΔV_{SN}, the lowest V_{DD} for reliable operation and reasonable yield is 700 mV, which is only 40 % of nominal V_{DD} (1.8 V).

IV. Impact of Supply Voltage Scaling on Retention Time

Low-voltage low-performance embedded microprocessors are best implemented in older technologies such as 180-nm CMOS to minimize energy dissipation, especially if leakage-reduction techniques such as power gating switches are applied [16]. The considered GC storage array is therefore implemented in a commercial 180-nm CMOS technology. Among many leakage mechanisms, the subthreshold conduction of the WT is clearly the dominant mechanism corrupting the stored data. This subthreshold conduction and consequently the data retention time strongly depend on the voltage level encountered on the WBL, denoted by V_{WBL}.

Assuming that a GC has just been written to and is now holding its data, there are two possible scenarios:

1) Further write operations are performed to GCs on the same WBL, meaning that V_{WBL} is data-dependent and cannot be controlled.
2) The memory remains in idle state (no data accesses) or only read accesses are performed. During idle and read states, V_{WBL} can be controlled to any desired voltage to minimize subthreshold conduction.

Fig. 1 shows the *worst-case access* scenario in terms of retention time where the opposite data level is permanently written to GCs on the same WBL after writing a given data level to the first GC. The *retention mode* scenario presumes an application where a relatively small storage array (few GCs per WBL) is fully written in a negligibly short time, whereafter the memory is kept in idle or read states and the WBL can be controlled to either V_{DD} or ground. Very short write access times, compared to the read access time, may easily be achieved in two-port memories. Under the same access scenario, the potential of controlling the WBL to a voltage level between the supply rails is then evaluated.

A. Worst-Case Access

Assuming the worst-case access scenario where V_{WBL} is permanently opposite to the stored data level, the retention time for a logic '0' ('1'), denoted by t_{ret0} (t_{ret1}), is defined as the time it takes for V_{SN} to rise (fall) to $V_{\mathrm{DD}} - V_{\mathrm{th}}^{\mathrm{RT}}$. At nominal V_{DD}, t_{ret1} is longer than t_{ret0}: the more the logic '1' voltage level decays, the more positive the gate-to-source voltage V_{GS} and the higher the reverse body biasing (RBB) of the WT, both suppressing the subthreshold conduction harder [12].

As shown in Fig. 3, when V_{DD} is gradually scaled down, the storage range for a logic '0', given by $V_{\mathrm{DD}} - V_{\mathrm{th}}^{\mathrm{RT}}$ (if neglecting charge sharing and clock feedthrough for simplicity), becomes smaller, while the storage range for a logic '1', given by $V_{\mathrm{th}}^{\mathrm{RT}}$, remains unchanged. At the same time, when V_{DD} is scaled down, the subthreshold conduction of WT becomes smaller due to its exponential dependence on V_{GS} and the drain-to-source voltage V_{DS}.

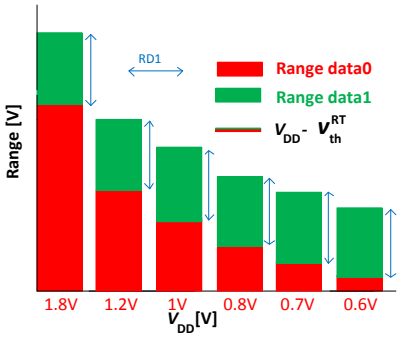

Fig. 3. Storage ranges versus supply voltage V_{DD}.

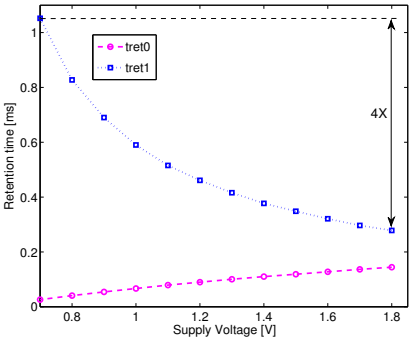

Fig. 4. Retention time versus V_{DD} for always-opposite WBL state.

As a consequence, t_{ret1} increases with decreasing V_{DD}, as shown by the Spectre simulation results in Fig. 4. However, Fig. 4 also shows that t_{ret0} decreases with decreasing V_{DD}, as the always smaller storage range has the higher impact than the decreasing strength of the subthreshold conduction.

B. Retention Mode

1) WBL Control to Ground: If the access scenario is now changed, assuming only idle and read states after initially writing the entire storage array, V_{WBL} can be controlled to ground, in order to avoid the decay of a logic '0'. In this case, the data retention time of the storage array is given by t_{ret1}. When scaling V_{DD} from its nominal value of 1.8 V down to 700 mV, the data retention time increases by 4× (cf. Fig. 4). At the same time, the power consumption is considerably reduced, due to 1) lower V_{DD}, and 2) fewer required refresh cycles.

2) WBL Control for Enhanced Retention Time: Still presuming the retention mode scenario, but now considering that V_{WBL} can be controlled to any desired voltage level between the supply rails to reduce subthreshold conduction, the retention time for any V_{DD} can be further increased compared to the WBL-discharge control.

Fig. 5 shows t_{ret1} and t_{ret0} as a function of V_{WBL}, for different values of V_{DD}. Clearly, t_{ret0} increases with decreasing V_{WBL} for any considered V_{DD}, due to a constant storage range and decreasing strength of the subthreshold conduction. For the same reasons, t_{ret1} increases with increasing V_{WBL}. The highest retention times are reached when V_{WBL} approaches $V_{\mathrm{DD}} - V_{\mathrm{th}}^{\mathrm{RT}}$, and t_{ret1} (t_{ret0}) becomes infinitely long for V_{WBL} higher (lower) than $V_{\mathrm{DD}} - V_{\mathrm{th}}^{\mathrm{RT}}$. However, the slopes in this region are very steep, so that any noise on V_{WBL} considerably degrades the retention time. At $V_{\mathrm{DD}} = 700\,\mathrm{mV}$,

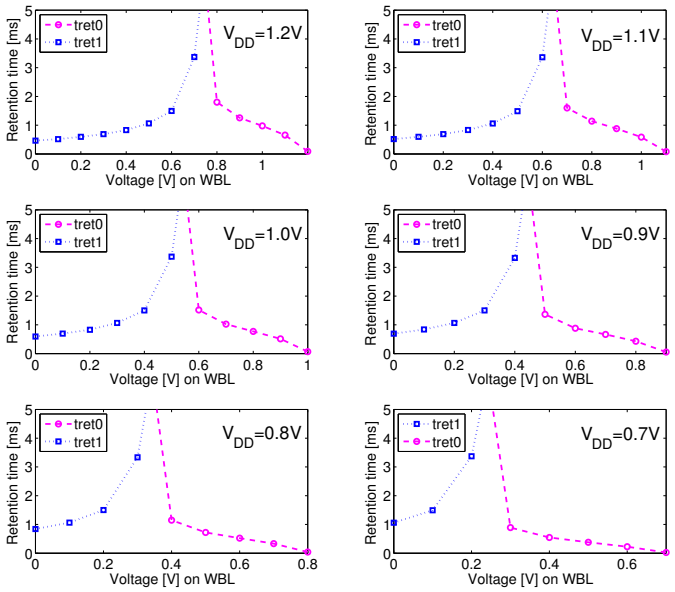

Fig. 5. WBL control for enhanced retention time.

choosing $V_{\mathrm{WBL}} = 200\,\mathrm{mV}$, a retention time of 3.3 ms is achieved, corresponding to a 3.3× improvement compared the case where V_{WBL} is controlled to ground.

V. Implementation Results

In the retention mode, an overall improvement of 13.2× in retention time and a considerable reduction in power consumption are obtained by supply voltage scaling and the controlled WBL technique. The active refresh power of the presented 2-kb macro is 10.8 pW/bit, while the leakage power is 1.1 pW/bit, amounting to a total refresh power of 11.9 pW/bit.

Table I compares this work to a selection of GC storage arrays in literature [12,13,17]. All retention time and refresh power values are given for a temperature of 25 °C, unless otherwise stated.

For the same technology node (180 nm), Table I shows the effectiveness of a high-V_{th} WT [13] (if available) to improve the retention time by around 100×. For smaller technology nodes (65 nm), [17] manages to keep a good retention time using a low-leakage process (and circuit-level techniques); however, in a native 65-nm logic process [12] (design optimized for high bandwidth), the retention time is degraded by around 100×.

In the presented study relying on a commercial 180-nm CMOS technology, the active refresh power is clearly dominant compared to the leakage power, meaning that any effort to increase the retention time also significantly reduces the total refresh power (cf. Table I). Reference [17] reports higher refresh power in 65-nm CMOS, but also uses a slightly higher supply voltage and measures at a temperature of 85 °C.

VI. Conclusions

Gain-cell storage arrays are an interesting alternative to SRAM macros in low-power/low-voltage SoCs and microprocessors. Gain-cells are inherently suitable for building two-port memories (as opposed to SRAM and conventional eDRAM). 2-transistor gain-cell storage arrays can be reliably operated at low supply voltages close to the threshold-voltage.

The data retention time improves by 4× when scaling down the supply voltage from 1.8 to 0.7 V, provided that write access is unfrequent and short. In addition to this, another 3.3× improvement in retention time is achieved by controlling the voltage on the write bit-line to a value between the supply rails during idle and read states. This overall improvement in retention time of 13.2× combined with operation at less than 40 % of nominal V_{DD} leads to a refresh power of 11.9 pW/bit.

TABLE I
Comparison of Gain-Cell Storage Arrays

Publication	[12]	[13]	[17]	This
Technology node [nm]	65	180	65	180
V_{DD} [V]	1.1	0.75	0.9	0.7
Retention Time [ms]	0.01	306[a]	1.25[b]	3.3
Refresh Power [pW/bit]	-	0.662	87.1 (85 °C)	11.9

[a] High-V_{th} transistor reduces leakage by more than 2 orders of magnitude [13]
[b] Low-leakage CMOS technology

Acknowledgment

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

References

[1] S. Hanson *et al.*, "A low-voltage processor for sensing applications with picowatt standby mode," *IEEE JSSC*, April 2009.
[2] S. Seo *et al.*, "Diet SODA: A power-efficient processor for digital cameras," in *Proc. ACM/IEEE ISLPED*, 2010.
[3] "A solar powered IA core? no way!" Intel Developer Forum, Sept. 2011. [Online]. Available: http://blogs.intel.com/research/2011/09/ntvp.php
[4] "International technology roadmap for semiconductors," 2009. [Online]. Available: http://www.itrs.net
[5] K. C. Chun *et al.*, "A 3T gain cell embedded DRAM utilizing preferential boosting for high density and low power on-die caches," *IEEE JSSC*, 2011.
[6] F. J. Kurdahi *et al.*, "Low-power multimedia system design by aggressive voltage scaling," *IEEE TVLSI*, 2010.
[7] J. Chen *et al.*, "An ultra-low-power memory with a subthreshold power supply voltage," *IEEE JSSC*, 2006.
[8] Y.-W. Chiu *et al.*, "8T single-ended sub-threshold SRAM with cross-point data-aware write operation," in *Proc. IEEE ISLPED*, 2011.
[9] B. H. Calhoun and A. P. Chandrakasan, "A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation," *IEEE JSSC*, 2007.
[10] P. Meinerzhagen *et al.*, "Benchmarking of standard-cell based memories in the sub-VT domain in 65-nm CMOS technology," *IEEE JETCAS*, 2011.
[11] S. Hong *et al.*, "Low-votage DRAM sensing scheme with offset-cancellation sense amplifier," *IEEE JSSC*, 2002.
[12] D. Somasekhar *et al.*, "2GHz 2Mb 2T gain-cell memory macro with 128GB/s bandwidth in a 65nm logic process," in *Proc. IEEE ISSCC*, 2008.
[13] 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.
[14] K. C. Chun *et al.*, "Logic-compatible embedded DRAM design for memory intensive low power systems," in *Proc. IEEE ISCAS*, 2010.
[15] M. Kaku *et al.*, "An 833MHz pseudo-two-port embedded DRAM for graphics applications," in *Proc. IEEE ISSCC*, 2008.
[16] M. Seok *et al.*, "Optimal technology selection for minimizing energy and variability in low voltage applications," in *Proc. ACM/IEEE ISLPED*, 2008.
[17] K. C. Chun *et al.*, "A sub-0.9V logic-compatible embedded DRAM with boosted 3T gain cell, regulated bit-line write scheme and PVT-tracking read reference bias," in *Proc. IEEE Symposium on VLSI Circuits*, 2009.