A 40nm 1.0Mb Pipeline 6T SRAM with Variation-Tolerant
Step-Up Word-Line and Adaptive Data-Aware Write-Assist

Chi-Shin Chang, Hao-I Yang, Yi-Wei Lin, Wei-Nan Liao, Yi-Wei Lin, Wei Hwang, Shyh-Jye Jou
Department of Electronics Engineering
National Chiao Tung University, Hsinchu, Taiwan, R.O.C.
cthuang@mail.nctu.edu.tw

Ming-Hsien Tu, Huan-Shun Huang, Yong-Jyun Hu, Paul-Sen Kan, Cheng-Yo Cheng, Wei-Chang Wang, Jian-Hao Wang, Kuen-Di Lee, Chia-Cheng Chen, Wei-Chiang Shih
Faraday Technology Corporation
Hsinchu, Taiwan, R.O.C.

Abstract—We present a 1.0Mb pipeline 6T SRAM in 40nm Low-Power CMOS technology. The design employs a variation-tolerant Step-Up Word-Line (SUWL) to improve the Read Static Noise Margin (RSNM) without compromising the Read performance and Write-ability. The Write-ability is further enhanced by an Adaptive Data-Aware Write-Assist (ADAWA) scheme. The 1.0Mb test chip operates from 1.5V to 0.7V, with operating frequency of 800MHz@1.2V and 25°C. The measured power consumption is 23.21mW (Active)/2.42mW (Leakage) at 1.2V, TT, 25°C; and 6.01mW (Active)/0.35mW (Leakage) at 0.7V, TT, 25°C.

I. INTRODUCTION

In memory-rich SoC, Static Random Access Memory (SRAM) dominates the performance, power consumption, area and minimum operating voltage (V_{MIN}). The ever stringent power consumption requirements of portable and hand-held devices dictate the use of low-power low-leakage technology. The low-power low-leakage transistors inherently have high V_{T}, thus having difficulty meeting the relentless demand for higher performance. Pipeline SRAM designs, that trade the access latency for higher operating frequency, have been widely used in the cache memory of high-performance microprocessors [1, 2]. The combination of deep sub-100nm low-power low-leakage technology with pipeline SRAM design offers high density, low-power and high operating frequency that would otherwise be impossible for cost-performance SRAM compiler applications.

In this work, we present a 1.0Mb 6T pipeline SRAM design in 40nm Low-Power (40LP) CMOS technology. The design employs a variation-tolerant Step-Up Word-Line (SUWL) to mitigate Read disturb and Half-Select disturb to improve the Read Static Noise Margin (RSNM) without compromising the Read performance and Write-ability. The Write-ability is further enhanced by an Adaptive Data-Aware Write-Assist (ADAWA) scheme with simple virtual cell supply control logic. Section II describes the pipeline structure. Section III presents the SUWL scheme. Section IV discusses the ADAWA scheme. Section V describes the 1.0Mb test chip implementation 40LP CMOS technology. Section VI presents and analyzes the measurement results.

Fig. 1 shows the two-stage pipeline structure [1, 2]. The conventional L1-L2 (Master-Slave) latches with non-overlapping clock are used for the Input Latch and Output Latch. L1 is “Active Low” and used to capture data of the preceding stage, whereas L2 is “Active High” and used to launch the captured data to the succeeding stage. The MiddleLatch consists of L1 only, and is followed by an AND gate [2] (AND with the L2 clock) which drives the Word-Line (WL). The first “Positive Half-Cycle” (i.e. 1st L2 Cycle) is used for Decode. The second “Positive Half-Cycle” (2nd L2 cycle) is used for WL activation, data-sensing through Local Bit-Line (LBL), Global Bit-Line (GBL), and data latching into GBL Latch and Data-Out (DO) Latch. As such, functions are performed during the “Positive Half-Cycles” and the “Negative Half-Cycles” are used for capturing data and precharge. The second “Positive Half-Cycle” is the cycle-time gating period. To minimize the clock skew and jitter among the latches, H-tree clock distribution commonly used in processor designs is employed.

II. PIPELINE STRUCTURE

III. STEP-UP WORD-LINE (SUWL)

Word-Line Under-Drive (WLUD) (also called Suppressed Word-Line (SWL)) is a commonly used technique to mitigate Read disturb and improve RSNM for 6T SRAM [3-5]. The improvement in RSNM, however, comes at the expense of
degraded Read performance and Write-ability. Noting that the maximum Read disturb occurs at the beginning of Read operation when the LBL voltage is high, various multi-step WL control techniques [6], where the under-driven WL voltage level steps up when the LBL voltage drops to a lower level, have been developed to mitigate Read disturb while containing and minimizing the degradation of Read performance and Write-ability.

In this work, we propose a variation-tolerant SUWL technique based on our previously proposed dual-tracking WLUD [5] as shown in Fig. 2. The core dual-tracking WLUD consists of PMOS M3 which tracks the WL driver pull-up PMOS and the cell holding PMOS, and NMOS M4 which tracks the cell access NMOS. The scheme exhibits excellent process corner tracking capability, and is resistor-less and area-efficient. In the proposed SUWL scheme, a replica-based LBL tracking circuit is added to control the gate voltage of PMOS M3. At the beginning of Read operation, the replica-LBL voltage is high, so PMOS M3 is “ON” and WLUD is enabled. As the voltage of the replica-LBL drops below the trip voltage of the inverter driving node S2, PMOS M3 starts to turn off, and WLUD is disabled. S0 is an external pin that can be used to completely disable WLUD if desired. Fig. 3 shows the simulated waveforms for SUWL with different LBL length (Leading edge rise time = 100 ps, VDD = 1.1 V, PTNT, 25°C).

Since SUWL deals primarily with dynamic stability, its effectiveness can only be assessed using dynamic metrics [7]. Fig. 4 shows the Monte Carlo simulation results for Read Margin (RM) vs. VDD with 3σ local random variation at PSNF corner and 125°C where Read Stability is the worst. The RM is a dynamic metric defined as

\[ \text{RM} = V_{\text{Trip}} - V_{\text{Disturb}} \]

where \( V_{\text{Trip}} \) is the trip voltage of the cell inverter, and \( V_{\text{Disturb}} \) is the maximum Read disturb voltage obtained from the transient Monte Carlo simulations. Due to the limitation on computational resource and prohibitively long simulation time for transient Monte Carlo simulations, the simulations are performed up to 3σ local random variation. It can be seen from Fig. 4 that the proposed SUWL technique improves the RM by 17mV (33mV) at 1.1V (0.7V).

IV. ADAPTIVE DATA-AWARE WRITE-ASSIST (ADAWA)

The Write-ability is further enhanced by the ADAWA scheme shown in Fig. 5. The column-based cell supply is split into 2 virtual supply lines, one for the left-half cells (VCSBLB), and the other for the right-half cells (VCSBL). In our previously proposed DAWA scheme [5], each of the virtual supply lines is controlled by a large PMOS power-switch and a small PMOS keeper. However, sizing of the keeper to effectively contain the low-going floating virtual supply across all PVT corners becomes extremely difficult, especially in deep sub-100nm technology. In the present scheme (Fig. 5(a)), each virtual supply line is controlled by two power-switches. The inner power-switch (M1/M2) is controlled by WE (Write Enable). Thus, M1 and M2 will be turned off by WE during Write operation to weaken the virtual supply. The pulse width of WE, determined by the “AND” of LWE (Local Write Enable) and delayed LWEB (Local Write Enable Bar), is programmable as shown in Fig. 5(b) by the 4:1 pulse generator.
multiplexer. The outer power-switch (M3/M4) is controlled by the BLB/BL through an inverter. Depending on Data-in at the BL pair, the low-going bit-line turns off either M3 or M4, causing the corresponding virtual supply node (VCSBLB or VCSBL) to drop, thus reducing the Vgs and contention of the corresponding cell holding PMOS to enhance Write-ability and WM. The opposite half-cell inverter is unaffected (with its virtual supply held by its outer power-switch) and maintains its strength and feedback action. The timing of outer supply switching is initiated directly from the low-going BL, thus tolerant to PVT variations and VT scatter. Data-aware switching of half-cell supplies reduces dynamic supply switching power and noise to half, and improves supply switching speed [5, 8, 9]. The scheme adds minimum loading to BL (only 1 extra Cgate). In UMC 40nm Low Power CMOS, splitting the vertical cell supply line into VCSBLB and VCSBL does not increase the 6T cell size. Notice that in Read mode, when either BL or BLB is pulled-down by cell access NMOS and pull-down NMOS, one of the outer power-switch (M3/M4) may turn off. This does not appear to be a problem since the corresponding virtual supply (VCSBLB/VCSBL) is still held by the inner power-switch (M1/M2), so the RSNM and Read performance are not affected.

![ADAWA Control Circuit](image)

Fig. 5. (a) Adaptive Data-Aware Write-Assist (ADAWA), (b) WE pulse width generation and control, and (c) Timing diagram for ADAWA.

Fig. 6 shows the Monte Carlo transient simulation results for AC Write Margin (AC WM, defined as the WM with WL pulse width of 1.0 ns) versus VDD at PFNS corner (where Write-ability is the worst) with 3σ local random variation. The improvement due to ADAWA can be clearly seen. The Write Vmin, defined here as the VDD where AC WM becomes 0, improves from 0.7V to 0.6V. Fig. 7 shows the Monte Carlo simulation results of the Write Time (Time-to-Write) at various VDD with 3σ local random variation at PFNS corner. The Write time improves by 18% at VDD = 0.8V.

![Monte Carlo simulation results of AC WM vs. V DD with 3σ local random variation at PFNS corner.](image)

V. 1.0Mb TEST CHIP IMPLEMENTATION

A 1.0Mb test chip is implemented in UMC 40nm LP CMOS. The 6T SRAM cell size is 0.303 μm². The 1.0Mb array is organized into 8192 Word x 128 bits. There are 2048 Columns with inter-leaving 16 architecture. The data I/O width is 128 bit. The Local Word-Line (LWL) length is 128 bit and LBL length is 8 bit. Single-ended large-signal sensing with a power-gated inverter is used.

![Monte Carlo simulation results of Write Time with and without ADAWA considering 3σ local random variation at PFNS corner.](image)

Fig. 8 shows the Read path from the Middle Latch to the Output Latch. The GBL is partition into GLBU and GLBD, each 256 bit long. The 256 bit is further divided into 32 local banks, so the LBL length is 8 bit. The Most Significant Bit (MSB) of address is used to control M6 and M9 for selection of GLBU or GLBD. If GLBU is selected, MSB will disable M6 to render GLBU node floating (for sensing). The Selected Bar (SELEB) signal will enable M4 of the selected local bank to allow passing of the sensed signal to GLBU. MSB_B will enable M9 to allow passing of GLBU data to GLB Latch. Low-VT devices are used in Decode path to reduce the Decode time. “Cycle Stealing” is used in the Decode path and on GBL Latch to improve the cycle time. The 2nd L2 cycle (from WL activation to DO Latch, see Fig. 1) is the cycle time limiting period.

Fig. 9 shows the layout view and die photo of the 1.0Mb test chip. The core chip area is 849.48 μm x 2900.48 μm. The area overhead for SUWL and ADAWA are 6.5% and 14.2%, respectively.
VI. 1.0Mb Test Chip Measurement Results

Fig. 10 shows the measured frequency Shmoo of the test chip. Fig. 11 shows the measured Bit Failure Rate (BFR) with SUWL and ADAWA. The characteristics of the chip is summarized in Table-I. The test chip operates from 1.5V to 0.7V, with operating frequency of 800MHz@1.2V and 25°C. With SUWL and ADAWA, the Read failure dominates BFR for VDD < 0.7V. The measured power consumption is 23.21mW (Active)/2.42mW (Leakage) at 1.2V, TT, 25°C; and 6.01mW (Active)/0.35mW (Leakage) at 0.7V, TT, 25°C.

Table-I: Characteristics of 1.0Mb Test Chip

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity</td>
<td>1.0M bits (8192 x 128 bits)</td>
</tr>
<tr>
<td>Technology</td>
<td>40nm Low Power (LP) CMOS</td>
</tr>
<tr>
<td>6T Cell Size</td>
<td>0.303 mm²</td>
</tr>
<tr>
<td>Core Area</td>
<td>849.48 mm² x 2900.48 mm²</td>
</tr>
<tr>
<td>Die/DO</td>
<td>128-bit</td>
</tr>
<tr>
<td>Bit Interleaving</td>
<td>16-bit</td>
</tr>
<tr>
<td>Local BL Length;  Local WL Length</td>
<td>8-bit; 128-bit</td>
</tr>
<tr>
<td>Measured Max. Freq. (1.2V, 25°C)</td>
<td>800 MHz</td>
</tr>
<tr>
<td>Measured Power (1.2V, TT, 25°C)</td>
<td>Active: 23.21mW (50% R/W)</td>
</tr>
<tr>
<td></td>
<td>Leakage: 2.42 mW</td>
</tr>
<tr>
<td>Measured Power (0.7V, TT, 25°C)</td>
<td>Active: 6.01mW (50% R/W)</td>
</tr>
<tr>
<td></td>
<td>Leakage: 0.35 mW</td>
</tr>
<tr>
<td>VDD (V)</td>
<td>0.70V</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENT

This work was supported by the Ministry of Education in Taiwan under ATU Program and the Ministry of Economic Affairs in Taiwan under Contract 100-EC-17-A-01-S1-124.

REFERENCES