Cache Design for Low Power and High Yield
Baker Mohammad\textsuperscript{1,2}, Martin Saint-Laurent\textsuperscript{1}, Paul Bassett\textsuperscript{1}, Jacob Abraham\textsuperscript{2}
\textsuperscript{1}QUALCOMM Incorporated \textsuperscript{2}The University of Texas at Austin

ABSTRACT
A novel circuit approach to increase SRAM Static Noise Margin (SNM) and enable lower operating voltage is described. Increasing process variability \cite{1} \cite{2} for new technologies coupled with increased reliability effects like Negative Bias Temperature Instability (NBTI) \cite{3} all contribute to raising the minimum voltage required for stable SRAM. Our strategy is to improve the noise margin of the 6T SRAM cell by reducing the effect of parametric variation of the cell \cite{4}, especially in the low voltage operation mode. This is done using a novel circuit that selectively reduces the voltage swing on the world line and reduces the memory supply voltage during write operation. The proposed design increases the SRAM Static Noise Margin (SNM) and write margin using a single voltage supply and with minimum impact to chip area, complexity, and timing. The technique supports both on-chip corner identification to adapt the SRAM behavior to silicon, and software controllability to tradeoff yield, power, and performance.

I. Introduction
SRAM based memory (cache) is becoming an increasingly important part of embedded processor design because of the impact it has on performance as well as implementation. For example, caches impact area, power, timing, yield, and schedule of the processor. The ever-increasing gap between processor frequencies and DRAM access times has dictated that processors have steadily been using more on-die SRAM to meet performance targets. As a result, SRAM arrays are in over 70\% of the devices and use 50\% of the chip area \cite{5}. A DRAM primary emphasis on density rather than speed makes the performance gap between CPU and main memory even greater. In addition, process scaling with ability to double the number of transistor in each generation makes it possible for on chip memory to almost double in each generation, further expanding the performance gap.

SRAM cell uses the smallest transistors on the chip and form a dense structure with regular patterns. Even though the regularity of the SRAM array may reduce the value of local variation its small geometry size makes any variation a big percentage of the target. The Six Transistor (6T) SRAM cell robustness, small area, and low power are an essential part of the chip optimizations. For sub 90nm process technology pMOS NBTI \cite{6} results on shifting pMOS threshold voltage, thus reducing the SNM of the SRAM and contributing to an increase in the $V_{ddmin}$ in order to compensate for the SNM reduction. SRAM stability and effect on yield often determine the minimum voltage the chip can tolerate with acceptable yield. The direct relationship between voltage supply, yield, and SRAM stability highlight the need for robust and adaptive Cache design styles.

We present a novel approach to employ for cost sensitive chips like consumer electronics. This approach will enable a power efficient product cross different applications through static voltage scaling based on usage models and process corners. The proposal will improve noise margin of the SRAM, hence enhance yield through read/write assist circuits. The technique is based on changing the wordline voltage level of the SRAM to reduce the minimum voltage required to achieve cell stability. This is important especially for cost sensitive mobile application where the chip has to support wide range of performance and power requirements. For example an embedded processor for mobile devices like cell phone needs to support high performance application like H.264 or High Speed Downlink Packet Access (HSDPA), at the same time require to run MP3 players where power efficiency is more important.

The paper is organized as follows: section II review the SRAM cell design principals. Section III discuss number of proposed approaches to address the SNM and low power operation and yield enhancement. Section IV presents our proposed technique to enhance SRAM SNM. Section V describes the detail circuit implementation of generating a reduced voltage swing (RVS) signals to be used for the proposed Cache design. Section VI summarize the new approach results and discusses some limitation.

II. BACKGROUND: SRAM CELL DESIGN
The SRAM 6T cell is the most frequently used cell in any design using on chip memory. Its main functionality is to store data for the program to access provided power is applied. The schematic of 6T cell is shown in Figure 1. The 6T cell design involves complex balancing between a plurality of factors, including, but not limited to:

1. Minimize cell area to achieve high density memory, reduce power, and reduce cost of the chip
2. Cell stability with minimum voltage to prevent yield loss due to data corruption.
3. Good soft error immunity. In systems with high reliability requirement a data error due to soft error can cause catastrophic failures.
4. High cell read current to minimize access time.

1. Minimize cell area to achieve high density memory, reduce power, and reduce cost of the chip
2. Cell stability with minimum voltage to prevent yield loss due to data corruption.
3. Good soft error immunity. In systems with high reliability requirement a data error due to soft error can cause catastrophic failures.
4. High cell read current to minimize access time.
5. Minimum word line pulse width to save on power (by reducing bitline swing)
6. Low leakage current especially for battery operated system to enable long battery life both during active and standby

Figure 1: SRAM 6T cell schematic

Many interactions between the different factors result in conflict. For example, to obtain good stability, small access time, and good soft error immunity, big transistor sizes may be used which result in big area usage and increased leakage. Also to improve SNM through increasing cell ratio (CR) implies smaller pass transistor (PG) which makes the write margin worse.

$$CR = \frac{W_{PD}}{L_{PD}}$$

A general familiarity of SRAM cell sizing is assumed on the consequent sections or consult one of the reference in [14][15]. For the SRAM to function properly during access at all process corner, voltage and temperature (PVT) the current through PD1 (I1) has to be greater than or equal to the current through pass gate (I0).

$$I1(\text{linear}) \geq I0(\text{saturation})$$

Also during write the current through pass gate (I3) has to be greater than or equal to (I2).

$$\mu nC_m \left( \frac{W_{PD}}{L_{PD}} \right) (V_{ddmem} - V_i - \frac{1}{2}V_n)V_{nl} \geq \frac{\mu nC_m}{2} \left( \frac{W_{PG}}{L_{PG}} \right) (V_{ddw} - V_{nl} - V_i)^z$$

For write completion

$$I3(\text{linear}) \geq I2(\text{saturation})$$

Equation 2 and 3 are used as the base line to find the optimum SRAM cell transistor sizes. Data from actual Silicon are normally used to tune the cell sizes and layout to obtain a robust design. In addition to finding the right balance for transistor sizing and cell area to achieve the design target, many proposals exist to address the minimum voltage requirement and parametric yield loss due to SRAM failure. These proposals can be characterized into four categories: SRAM cell modification, voltage islands, Body/well biasing, and circuit techniques.

III. Present Approaches to Robust SRAM and Cache Design

As previously described, the interaction between the SRAM cell parameter affecting read and write may create optimization difficulties. One proposal suggests modifying the SRAM cell to separate the read and write operation as a way to decouple the read optimization form the write [7]. The proposal adds at least one transistor to the cell which results in bigger cell area (7T cell or 8T cell). One problem with this approach is that it causes a large area increase.

Another approach is the use of voltage island to decouple the memory supply from the logic supply [5][8]. This approach not only enable the supply voltage to run at lower voltage and reduce active power quadratically, it also can be used to reduce leakage power when coupled with different operating models like active, standby, and sleep mode. The disadvantage of this approach is the cost and complexity associated with adding a new power supply. For example, level shifter and isolation circuitry may be required for signals to cross voltage boundaries.

Mukhopadhyay et all in [10] used body bias for NMOS and well bias for PMOS to shift the threshold voltage higher or lower based on the inter-die process corner. Leakage and ring oscillator delay monitoring is used to determines the inter-die process corner. The main purpose of this work was to apply body bias to reduce the number of parametric failures. Since the principle reason for parametric failure is due to random doping fluctuation induced threshold voltage shift so reducing this variant will decrease the probability of the cell to fail. Low vt shift has impact on read and hold failures while High vt shift has impact on access and write failures. Hence, sensing the process corner and specifically the inter die threshold voltage shift can determine which failures are most likely to occur. A circuit to select the proper body bias to minimize the impact of the vt shift is activated.
and the body voltage is applied to form a forward body bias (FBB) for high vt or Reverse Body Bias (RBB) for low vt. This approach does shift all nMOS transistor threshold voltage the same way so its effectiveness in addressing the SNM issue is limited to changing the trip point of the forward inverter inside the 6T cell. It addresses the global variation and can minimize the yield loss due to SRAM parametric failures especially if used along with redundancy. Redundancy can be used to fix limited number of faulty cells in a column or row so adding FFB and RBB increases the chance of passing parts.

Special circuit design techniques such as the ones in [11][12] used to change the voltage applied at the wordline and the vddmem to improve SRAM cell stability and improve yield. Yabuuchai et al in [12] proposed SRAM read/write assist circuits to enlarge the operating margin against wide process and temperature variations with a single supply voltage. The approach used voltage divider to reduce wordline voltage and used dummy bitline capacitance to reduce the vddmem during write. This approach is mainly intended to tune the wordline voltage to increase read stability during memory access. The WL voltage value is reduced through contention which increases the active power. To increase write margin the vddmem is reduced via charge sharing of the vddmem column and dummy metal capacitance, the dummy metal capacitance then get discharge after each access. With process variation balancing capacitance to get a balance voltage is challenging, also with high density circuit wire routing becomes premium and adding dummy metal is less desirable.

Yamaoka et al in [11] used floating vddmem during write to help the write margin of the 6T cell. This approach works good for low frequency application but has limit to help the write margin due to the fact that the vddmem capacitance is comparably big and the discharge path has to go through the pull up transistor PU of the 6T cell which is very small device.

**IV. Proposed Reduced Voltage Swing (RVS) Wordline based memory with write assist**

We present our approach to increase SNM of SRAM through reducing the wordline voltage. We illustrate the effectiveness of the approach mathematically and then confirm using HSpice simulation. First, an equation for the SNM is derived. It shows that reducing the WL voltage improves the SNM. Then confirm result with Spice simulation using 45nm foundry SRAM cell. Solving equation 2 for Vn1 using data from the low leakage 45nm process technology with the following values: Vtn of 0.35v, CR of 4.4, and Vtp=0.4v yield equation 4

\[
V_{n1} = 0.17(4V_{ccmem} + 2V_{ddw} - 2.1 - 2.82V_{ccmem}^2 + 2V_{ccmem}V_{ddw} - 2.1V_{ccmem} - V_{ddw}^2 + 0.37)
\]

\[
2.82V_{ccmem}^2 + 2V_{ccmem}V_{ddw} - 2.1V_{ccmem} - V_{ddw}^2 + 0.37
\]

The inverter trip point is shown in equation 5.

\[
V_{\text{trip}} = \frac{1}{1 + \frac{1}{K_R}} \left( V_{\text{dd}} + V_{\text{tp}} \right)
\]

\[
K_R = \frac{\mu_n C_{ox} W_n}{L_n} \frac{W_p}{L_p}
\]

Calculating Vn1 and Vtn using fast nMOS slow pMOS transistor and subtracting Vtn from Vn1 will give the SNM. Figure 3 plots the Vn1 and SNM for different values of wordline voltage and vdd.

---

**Figure 2: SRAM Internal node voltage (Vn1) and SNM with different wordline voltages**

The analytical model predicts that reducing the wordline voltage increases the SNM in both the nominal and low voltage supply. We also confirm this finding by
simulating using Hspice and a foundry approved SRAM cell on 45nm bulk process.

Figure 3: SRAM internal node voltage for different wordline voltage value

Figure 3 shows the waveform result of Hspice simulation. Since the approach depends on reducing the wordline voltage based on PVT corner the write margin gets worse as the pass gate becomes slower with less voltage. To address this we implement a selective reduction in the memory supply for the columns that are Figure 4: Traditional and expected V\textsubscript{ddmin} for Write completion chosen for write. To show that reducing memory supply improves write we solve equation 3 for Vn2

\[ V_{n2} = 0.33(1.8 - 2.45) \sqrt{1.5V_{ddw} - 1.2V_{ddw} - V_{ddmem}^2 + 0.7V_{ddmem} + 0.12} \]

Figure 4 shows the plot of Vn2 voltage and Vth versus different Vddmem. For write completion the Vn2 voltage needs to be less than or equal the Vth. The plot shows the intersection of the Vth with the Vn2 for 2 values of the WL voltage. It is obvious from the plot the effectiveness of lowering the V\textsubscript{ddmem} on reducing the minimum voltage.

V. Circuit to Generate RVS WL and Memory Supply

A detail schematic of the circuit used to generate the RVS wordline is show in Figure 5. The Power\_mode\_wl signal is used to enable or disable the RVS circuit based on the PVT corner or program control. When Power\_mode\_wl is set to logic 1 then the new proposed circuit will function the same as the traditional one and wordline will swing full swing. The Cn[n:0] signals are used to adjust the level of the WL voltage. This is done through adjusting the size of Mp3 transistor and hence changing how fast the charge will be transferred from WL node into pk0 node. The Mn1 nMOS device is used to pre condition pk0 node during normal operation mode to enable the pull up path. Both Mn1 and Mn1 are minimum size transistor and they don’t have any impact on timing.

The same concept used in the above circuit is used to lower the Vddmem supply during write operation. Figure 6 shows the detail circuits and the expected waveform.

Figure 5: Circuit producing reduce voltage swing WL

The same concept used in the above circuit is used to lower the Vddmem supply during write operation. Figure 6 shows the detail circuits and the expected waveform.
VI. Summary and Limitation of the Proposed Solution

We showed both mathematically and through Hspice simulation using data from 45nm low power process a technique that increased the SNM of SRAM cell by reducing the wordline voltage. We also improved the write timing of the cell by selectively reducing the memory supply. For both cases the $V_{ddmin}$ is reduced. For our case study the $V_{ddmin}$ is reduced from 1v to 0.8v. This reduction of 0.2 v result in 36% of active power. We showed a low power circuit implementation to generate the reduced voltage on wordline and the memory supply using single voltage supply. The new design increased the SNM and the write margin of the SRAM cell hence increased yield. The design offers full controllability and programmability of the RVS circuit it can be used to tune the SRAM behavior to match silicon via on chip control circuit. A simple ring oscillator delay or leakage monitor [10] can be used to tune the circuit based on transistors parameter.

Both RVS Wl and RVS Vccmem circuits use a delay elements to tune the value of the new voltage level, hence its granularity and level is limited by the delay element speed. Also the speed of the memory access will be reduced to Wl voltage reduction. The timing impact can be reduce if the control signal enables the RVS control only on the fast corners where the SNM is most likely to effect the cell. Since the spread between the slow corner and fast corner is big and the circuit is designed to meet timing at the slow corner then the timing impact might not be an issue.

References


