A Forward Body-Biased Low-Leakage SRAM Cache: Device and Architecture Considerations

Chris H. Kim, Jae-Joon Kim, Saibal Mukhopadhyay, and Kaushik Roy
Department of Electrical and Computer Engineering, Purdue University
1285 Electrical Engineering Building, West Lafayette, IN 47907, USA
+1-765-494-3372
{hyungil, jk, sm, kaushik}@ecn.purdue.edu

ABSTRACT

This paper presents a forward body-biasing (FBB) scheme for active leakage power reduction in cache memories. We utilize super high V_T (threshold voltage) devices to suppress the leakage power in unselected portions of a cache while fast operation is achieve by dynamically forward body-biasing the selected SRAM cells. In order to generate a super high V_T device, the 2-D halo doping profile was optimized considering different nanometer regime leakage mechanisms. The transition latency and energy overhead associated with FBB could be minimized by (i) waking up the SRAM cells ahead of the access and (ii) exploiting the cache access pattern. The combined device-circuit-architecture level techniques offer 64% total leakage reduction and 7.3% improvement in bitline delay compared to a previous state-of-the-art low-leakage SRAM technique.

Categories and Subject Descriptors

B.7.1 [**Integrated Circuits**]: Types and Design Styles

General Terms: Performance, Design

Keywords

Forward body-biasing, leakage power, SRAM, super high V_T

1. INTRODUCTION

With the number of transistors on a chip rapidly approaching 1 billion, and the ever-increasing levels of integrated cache, leakage power management is indispensable for cost effective packaging and cooling solutions in high-end microprocessors. Leakage power is also a concern in low-end mobile system-on-chips where the low standby power feature is crucial. Recent energy estimates for 0.13μm process indicate that leakage energy accounts for 30% of L1 cache energy and as much as 80% of L2 cache energy. With device I_{OFF} increasing by a factor of approximately 3 every new technology generation, leakage power consumption in SRAM memories will continue to dominate the total chip leakage.

Forward body-biasing (FBB) has proven to effectively improve performance, suppress short channel effects, and reduce V_T variations [1,2]. V_T roll-off and drain induced barrier lowering (DIBL) which limits scalability of channel length can be resolved by forward body biasing the devices during normal operation [1]. In this paper, a dynamic FBB scheme for low-leakage SRAM caches is presented where leakage power is significantly reduced by utilizing super high V_T transistors. FBB is dynamically applied to only the selected portion of the cache for fast read/write operation. The idea of using a high V_T device and applying a FBB to achieve high drive current in active mode has already been discussed in previous literature [1,3]. They have also mentioned that withdrawing the FBB in standby mode can significantly reduce the leakage power consumption. However, it is not clear how one can optimize the high V_T device for the FBB scheme in scaled technologies where different leakage mechanisms make it non-trivial to obtain a desirable V_T. Moreover, it has not been reported whether a dynamic FBB scheme is profitable enough so that it can be applied in a fine grain manner to reduce the cache leakage during active mode. It is the goal of this paper to explore if such a fine grain dynamic FBB scheme can be useful in reducing active cache leakage and devise a set of solutions to achieve best performance under given leakage power constraints. For this, we have developed combined device-circuit-architecture level techniques and compared the proposed forward body-biased SRAM (FBSRAM) with a prior state-of-the-art low-leakage SRAM technique. This paper makes the following contributions. For the first time, we apply the concept of using a super high V_T device and FBB to dynamically reduce the active leakage in cache memories. We also show optimization of the 2-D halo doping profile for the super high V_T device to achieve total leakage reduction while suppressing the band-to-band tunneling (BTBT) and gate leakage. Circuit techniques and architectural behavior of caches are utilized to reduce the transition latency/energy associated with body-biasing.

ISLPED'03, August 25–27, 2003, Seoul, Korea.

2. LOW LEAKAGE SRAM CELLS

Various SRAM cell leakage reduction techniques have been proposed in the past. They all exploit the fact that the activation portion of a cache is very small, which gives the opportunity to put the large idle portion in a low-leakage sleep mode.

Source biasing scheme raises the source line voltage (V_{SL}) in sleep mode to generate a negative Vgs in the access transistor and reduce the bitline leakage [4,5,6,7]. Sub-threshold and gate leakage in the cell is also reduced due to (i) the lowered signal rail (V_{DD}-V_{SL}) and (ii) body effect in the NMOS transistors. An extra NMOS has to be series connected in the pulldown path in order to cutoff the source line from ground during sleep mode, and this in turn imposes an extra access delay.

Reverse body-biasing (RBB) the NMOS (or PMOS) can reduce sub-threshold leakage via body effect, while not affecting the access time by switching to zero body-biasing (ZBB) in active mode [6,8,9]. A large latency/energy overhead is imposed for the body-bias transition due to the large V_{BODY} swing and substrate capacitance. This scheme becomes less attractive in scaled technologies since the body coefficient decreases with smaller dimensions, and BTBT becomes enhanced by RBB.

Supply voltage is lowered in a dynamic V_{DD} SRAM (DVSRAM) [6,10], which in turn reduces the sub-threshold, gate, and BTBT leakage. Although there is no impact on delay in the active mode, the large V_{DD} swing between sleep and active mode imposes a larger latency/energy transition overhead than the SBSRAM. Moreover, the greatest drawback of the DVSRAM is that it increases the bitline leakage in the sleep mode since the voltage level in the stored node also drops as the V_{DD} is lowered. Therefore, this scheme is not suitable for dual V_T designs where the speed-critical access transistors may already be using low V_T devices with high leakage levels.

A technique that biases the bitlines to an intermediate level has been recently proposed to reduce the access transistor leakage via the DIBL effect [11]. Since only the access transistors benefit from the leakage reduction, the overall leakage savings is marginal. The main limitation comes from the fact that there is a precharge latency whenever a new subarray is accessed. This would mean that an architectural modification is required in order to resolve the multiple hit times in case the precharge instant is not known ahead of time.

3. DEVICE OPTIMIZATION FOR FORWARD BODY-BIASING

We propose a FBSRAM cell architecture, which lowers the standby leakage by means of a super high V_T device enabled by channel doping techniques or work function engineering. To obtain fast read/write operation, the speed critical NMOS transistors are dynamically switched from ZBB to FBB only when the subarray is selected. Next, we will discuss how the collaboration between device engineering and FBB can effectively lower standby leakage and simultaneously achieve high drive current. Throughout this paper, we will divide the various leakage components into 3 main categories; (i) sub-threshold leakage, (ii) direct tunneling gate leakage, and (iii) BTBT currents through the drain-well junction.

3.1 Super-halo 2-D doping profile

A 50nm effective channel length (L_{EFF}) device based upon the super-halo discussion by Taur et al., has been incorporated for the MEDICI simulations [12,13]. The nominal NMOS device has a physical oxide thickness (tox) of 1.5nm and V_T of 270mV. Super-halo uses a non-uniform p+ doping in the source-body and drain-body boundaries to reduce the source-drain depletion width, and to effectively suppress the body punchthrough [12]. V_T roll off and DIBL is also controlled by the 2-D halo doping profile. V_T of a short channel super-halo device can be effectively changed by adjusting the halo doping concentration or varying the halo implant location/angle [14]. Changing the background channel doping is less effective because (i) V_T is less sensitive to channel doping, (ii) DIBL and punchthrough cannot be suppressed as effectively, and (iii) the impact on drive current is more severe [15]. Other general means to raise the device V_T is to have a thicker physical tox or a longer channel length. However, the former will worsen the short channel effect and the latter will increase the load capacitance and area of the FBSRAM.

3.2 Super high V_T device design

A new super high V_T (V_T=350mV) doping profile for the FBSRAM is generated by adjusting the peak halo doping concentration. One of the concerns with increasing the peak halo doping is the exponential increase of BTBT current due to the high field effect near the drain-body junction. We were able to generate a super high V_T device that gives equivalent I_{OFF} of an optimized SBSRAM cell [4,5,6,7] while keeping the BTBT current to be less than 3% of the total leakage consumption.

Fig. 1 shows the drain current of the super high V_T device compared to the nominal V_T device. Under a ZBB, device I_{OFF} is reduced by 4X offering a low standby leakage. By applying a 500mV FBB to the super high V_T device, the I_{ON} is improved by 17%, offering a 3% higher drive current compared to the nominal V_T device. The total leakage of a 6TSRAM cell shown in Fig. 2 (top) indicates that the reduction is mainly due to the improvement in sub-threshold leakage dominating the I_{OFF} in high temperatures.

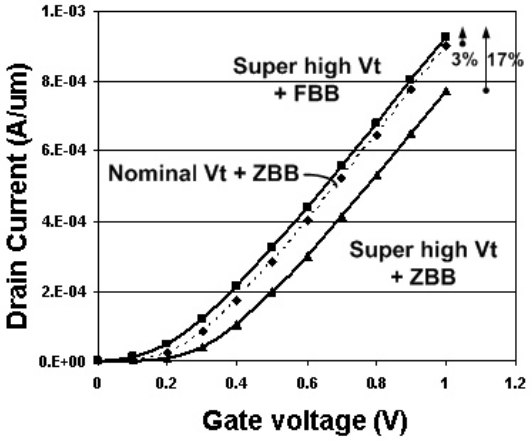

Fig. 1. NMOS drain current with and without forward body-biasing for a nominal V_T device and a super high V_T device (50nm L_{EFF}, 1.0V, T=110°C)

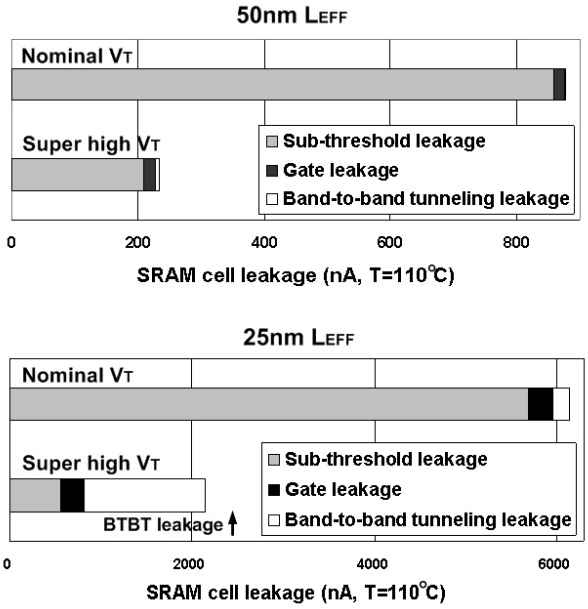

Fig. 2. SRAM cell leakage of nominal V_T vs. super high V_T. The super high V_T device is generated using channel engineering.

3.3 Scaling Trends

The super high V_T accomplished by raising the halo doping concentration gives rise to a noticeable BTBT component (Fig. 2, top). This becomes more evident in future process generations where

a higher baseline halo concentration is needed to suppress the worsening V_T roll off and DIBL with device scaling (Fig. 2, bottom). In technologies where one cannot afford a higher halo doping, a super high V_T device can be built by using a gate material with a higher work function [15]. Fig. 3 shows that a super high V_T device can be realized using gate work function engineering, without impacting the BTBT leakage.

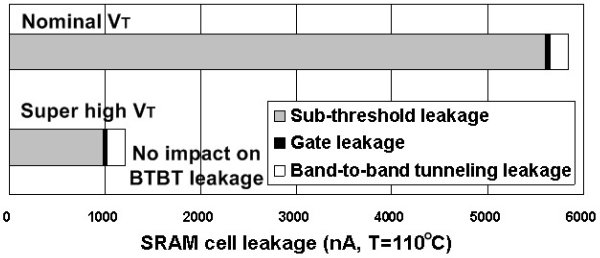

Fig. 3. SRAM cell leakage of nominal V_T vs. super high V_T. The super high V_T device is generated using work function engineering.

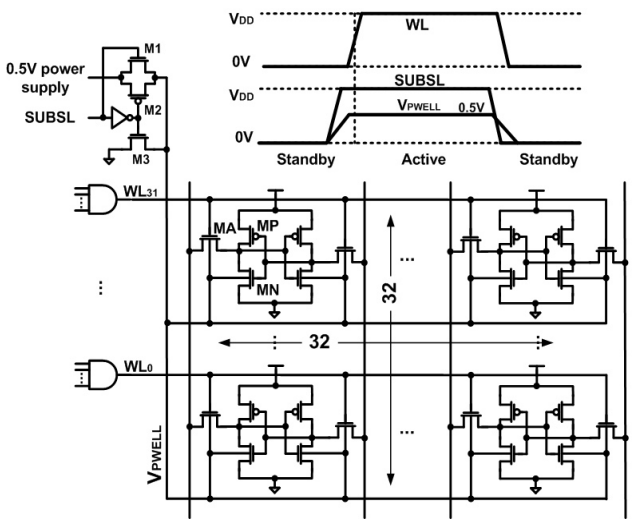

Fig. 4. A 32x32 FBSRAM subarray with body-bias driver. (W_{MA}=0.2µm, W_{MP}=0.3µm, W_{MN}=0.3µm, L_{EFF}=50nm)

4. FBSRAM INSTRUCTION CACHE DESIGN

4.1 Dynamic Leakage Control Scheme

Fig. 4 shows the circuit diagram of a 32bx32b FBSRAM subarray with body-bias drivers M1-M3. Each subarray has a SUBSL signal input which is generated by the decoder circuit. When there is an access to a particular subarray, the SUBSL signal is fired and M1 and M2 are turned on. This switches the NMOS body-bias (V_{PWELL}) to a 0.5V FBB, increasing the drive current and achieving a fast read/write operation. When a subarray is not accessed on the other hand, the SUBSL signal stays low and a ZBB is applied to the super high V_T devices via M3. This substantially reduces the total I_{OFF} during the inactive periods. Triple well technology is required to isolate the p-substrate of different subarrays. The V_{PWELL} line can be routed using an upper metal layer, so the only area overhead comes from the boundary of each subarray where design rule requires an area margin for well isolation. This area overhead, however, is significantly less than previous row-by-row body-biasing techniques where each cache line is isolated from the adjacent ones [8,9].

4.2 Transition Latency Hiding

The subarray select signal SUBSL is used to trigger the ZBB to FBB transition even before the wordline signal is fired. This is shown in Fig. 4 where the charging of V_{PWELL} is completed before the word line signal arrives. Only the most significant address bits A[5:N] are needed to generate the SUBSL signal, allowing extra time to complete the body-bias transition before the wordline arrives. This reduces the increase in access time due to the body bias transition for a subarray based memory architecture with independent body bias control.

4.3 Transition Energy Reduction

The SUBSL signal in Fig. 4 is generated from the most significant address bits (A[5:N]) and thus it will not toggle unless the most significant address bits are switched. Although hiding the transition latency enables a fast body-bias transition, the transition energy remains unchanged. Observation of the cache access pattern reveals that the number of body-bias transitions is significantly less than the worst case. When data is first brought into a cache, it experiences a burst of accesses. After the flurry of accesses, there is a considerably long period of time between the last access and the point when the data is replaced referred to as the "dead period" [16]. This implies that there is (i) a high chance that a subarray in access will be accessed again in the next cycle and that (ii) a subarray in sleep mode is more likely to stay in sleep mode throughout the dead period. This behavior is visualized in Fig. 5 for a 32KByte, 4 way, 128B block, 32bx32b subarray based L1 instruction cache simulated using SimpleScalar-3.0 for 500 million instruction cycles. On average, the percentage of accesses hitting the same subarray for consecutive cycles is 93% for SPEC2000 benchmark applications. Therefore, the ZBB to FBB transition happens only in 7% (=100%−93%) of the total accesses, significantly lowering the overhead energy for a subarray based cache architecture.

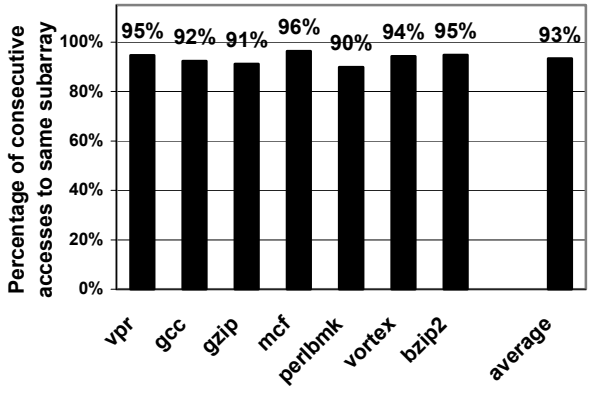

Fig. 5. Percentage of hitting the same subarray in consecutive cycles for SPEC2000 benchmark applications (32KByte, 4 way, 128B block, L1 instruction cache).

5. SIMULATION RESULTS

Total leakage power savings and performance improvement of the FBSRAM with super high V_T transistors were derived through extensive MEDICI simulations. All simulations use 50nm L_{EFF} devices, supply voltage of 1.0V and threshold voltage of 270/350mV (nominal V_T/super high V_T). We determine the L1 instruction cache geometry to be the same as the one used for the SimpleScalar

simulations in the previous section (32KByte, 4 way, 128B block, 32bx32b subarray).

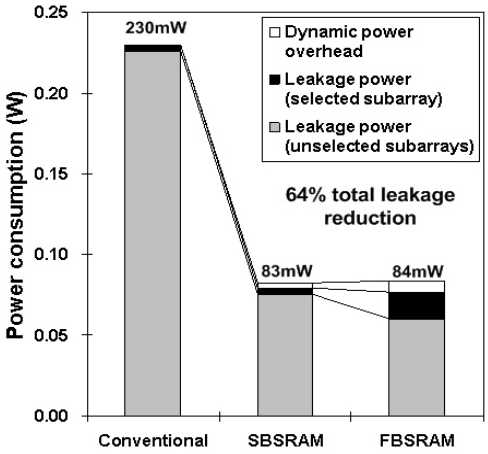

Fig. 6. Leakage power and dynamic power overhead of FBSRAM compared to prior SRAM techniques (50nm L_{EFF}, 1.0V, T=110˚C).

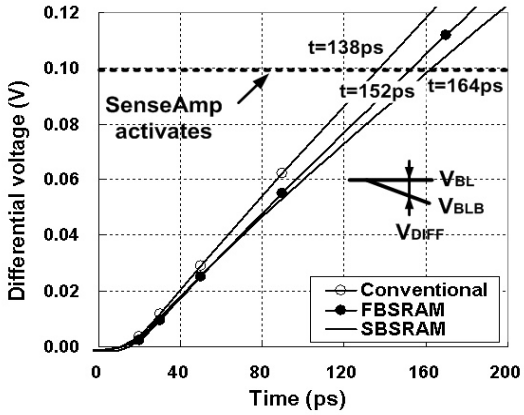

Fig. 7. Differential voltage (=V_{BL}-V_{BLB}) of FBSRAM compared to conventional and SBSRAM (50nm L_{EFF}, 1.0V, T=110˚C).

We have chosen to compare the proposed FBSRAM technique with only the SBSRAM since other SRAM techniques such as reverse body biasing or dynamic V_{DD} have fundamental design problems which may make these techniques not likely to be used in commercial systems. Fig. 6 compares the leakage power dissipation of FBSRAM with SBSRAM. A nominal V_T of 270mV with 0.2V source biasing voltage was used for the SBSRAM, which gives equivalent leakage savings as the FBSRAM having super high V_T devices. By raising the source line voltage from 0V to 0.2V, the SBSRAM was able to reduce the total leakage power by 64% (including overhead). The overhead dynamic power and the leakage from the selected SRAM cells account for 8% of the total leakage power. The FBSRAM was able to achieve iso-leakage power consumption as the SBSRAM by applying ZBB to the unselected portion of the cache.

Bitline delay which is the time for the differential voltage (Fig. 7) to reach 100mV, turned out to be smaller in the FBSRAM (152ps) by 7.3% compared to the SBSRAM (164ps) for iso-leakage conditions. This is because the SBSRAM, though using a lower V_T device, contains a 3-transistor stack in the pulldown path aggravating the bitline performance. The bitline delay of the SBSRAM can be improved by increasing the width of the sleep transistor. However, this will impact the leakage savings, increase the transition energy overhead, and worsen the impact of the sleep transistor on SRAM area [4]. We have shown in Fig. 1 that the drive current of a super high V_T device with FBB is higher than a nominal V_T device with ZBB by 3%. The junction capacitance of the super high V_T device however, is larger due to the FBB and increased halo doping. This results in a 10.1% (138ps→152ps) slower bitline delay for the FBSRAM compared to a conventional 6TSRAM using nominal V_T transistors.

6. CONCLUSIONS

Previous low-leakage SRAM architectures have inherent limitations due to either delay overhead, increased bitline leakage, multiple hit times, or impact on cell stability. Even state-of-the-art low-leakage SRAM cell techniques such as the source biasing scheme has issues such as large performance penalty and degradation in SER. This paper starts from a simple initiative of utilizing super high V_T devices for low-leakage and forward body-biasing them for high performance. We have looked at different levels of the design to reduce leakage power at low expenses and achieve high performance. At the device level, the super high V_T doping profile was optimized to improve the DIBL and V_T roll-off while suppressing the BTBT leakage component. Transition latency associated with FBB was hidden by modifying the decoder circuit to give an early notice to the subarray that is to be waken up. At the architectural level, the general cache access pattern is exploited to lower the body-bias transition energy. As a result, the combined device-circuit-architecture level techniques achieve 64% total cache leakage reduction (overhead included) with 7.3% improved bitline delay compared to the SBSRAM.

7. ACKNOWLEDGMENTS

This research has been funded by the DARPA PACC program.

8. REFERENCES

[1] A. Keshavarzi et al., VLSI Circuits Symp., pp. 312-315, 2002
[2] M. Miyazaki et al., ISSCC, pp. 420-421, 2000
[3] S. Narendra, et al., ISSCC, pp. 270-271, 2002
[4] A. Agarawal et al., DAC, pp. 473-478, 2002
[5] H. Yamauchi et al., VLSI Circuits Symp., pp. 126-127, 1996
[6] A. J. Bhavnagarwala et al., ASIC/SOC Conference, pp. 359-363, 2000
[7] K. Osada et al., ISSCC, pp. 302-303, 2003
[8] H. Kawaguchi et al., VLSI Circuits Symp., pp. 140-141, 1998
[9] C. H. Kim and K. Roy, ISLPED, pp. 251-254, 2002
[10] K. Flautner et al., ISCA, pp. 148-157, 2002
[11] S. Heo et al., ISCA, pp. 137-147, 2002
[12] Y. Taur et al., IEDM, pp. 789-792, 1998
[13] http://www-mtl.mit.edu/Well/
[14] B. Yu et al., IEDM, pp. 653-656, 1999
[15] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Devices*, Cambridge University Press, Nov. 1998
[16] S. Kaxiras et al., ISCA, pp. 240-251, 2001