26.5 PVT-Aware Leakage Reduction for On-Die Caches with Improved Read Stability

Chris H. Kim1, Jae-Joon Kim2, Ik-Joon Chang1, Kaushik Roy1

1University of Minnesota, Minneapolis, MN,
2IBM, Yorktown Heights, NY,
3Purdue University, West Lafayette, IN

Leakage control during circuit operation is more challenging than standby mode control due to the short time to deactivate blocks, large overhead energy and run-time leakage variations. This paper proposes circuit techniques that address these challenges to reduce run-time leakage in on-die SRAM caches. A source-biased gated-ground SRAM is proposed; an efficient way to utilize this technique under severe leakage variations exploits the architectural behavior of an L1 cache.

Figure 26.5.1(a) shows the source-biased gated-ground SRAM cell in which the virtual ground voltage (VGN) is raised during sleep mode to generate a negative VGS in the access transistors, reducing BL leakage by orders of magnitude [1]. Sub-threshold and gate leakages are also reduced by the body effect and the lower voltage stress between device terminals. Charging/discharging the VGN of an SRAM block and switching in the extra circuits requires overhead energy. Architectural access patterns of caches reveal that a simple circuit technique can significantly lower the overhead energy by reducing the number of transitions [2]. Data experiences a burst of accesses when it is first entered into an L1 cache. After this live period, a considerably long dead period follows where no access occurs until the data is replaced. Conventional techniques that turn off cache blocks immediately after they are accessed cause the number of sleep and wake-up procedures to increase during the flurry of accesses [2]. The proposed technique, in Fig. 26.5.1(b), periodically sends the cache blocks to sleep so that a once activated cache block is kept in the active mode until the next sleep pulse arrives. Thus, unnecessary wake-up transitions are avoided during the live period while the majority of the cache blocks in the dead period are kept in sleep mode. For best tradeoff between leakage savings and overhead energy under severe leakage variations, TDECAY (the interval between sleep pulses) is determined adaptively. At low leakage conditions, it is optimal to have a long TDECAY that seldom turns off the SRAM. This reduces the dynamic energy overhead which is large compared to the relatively small leakage savings. Alternatively, the block must enter the sleep mode more frequently at high leakage conditions since the amount of leakage that is saved becomes larger than the overhead energy. The designed self-decay-based sleep signal generator is capable of tracking the process and temperature fluctuations during run-time, providing a near-optimal TDECAY. Since the decay rate is determined by the leakage of M2 in Fig. 26.5.1(c), TDECAY becomes shorter at high leakage conditions and vice versa.

Figure 26.5.2 shows organization of a 16kB SRAM for active leakage reduction based on the self-decay scheme. The SRAM is divided into 64 blocks, each having 4x512 cells. The VGN of each block is separated for independent control. Issues of current crowding and ground bounce are resolved by having the sleep transistors distributed across the SRAM block. When the block enters the sleep mode, VGN is raised to VGS via device MP1 (Fig. 26.5.2). VGS is generated from a high-efficiency DC-DC converter. Since entering sleep mode is not time critical, MP1 is sized to be 6% of the total sleep transistor width and is physically placed inside the row decoder block. The SRAM block is activated ahead of time using the predecoder signals so the only delay penalty comes from the bounce in VGN (<50mV) during the read operation and affects the read access time by less than 2% [Fig. 26.5.4(d)]. To validate the effectiveness of the proposed self-decay technique, simulations are carried out using SimpleScalar-3.0 (SPEC2000 benchmarks) and a predictive 70nm technology. Statistical leakage saving results in Fig. 26.5.3 show that the self-decay circuit offers 27% lower leakage compared to a scheme where TDECAY is constant [2].

A testchip is fabricated in a 0.18um, 6M, 1.8V CMOS technology and the die photo is shown in Fig. 26.5.7. The read access frequency of the SRAM is 984MHz and the performance penalty due to the sleep transistors is measured to be less than 2%. The active power consumption is 0.14mW/MHz and the area overhead for the sleep transistors, self-decay circuit and additional peripheral circuits is 6% of the total SRAM area. Figure 26.5.4(a) shows the measured leakage data while sweeping VGN from 0V to 1.2V. BL leakage is virtually zero for a VGN higher than 0.2V leaving just the cell leakage and the junction leakage (Nwell-substrate, drain-body). The measured leakage current of the 16kB SRAM with VGN of 0.9V is 0.42µA. This is only 1/4 of the conventional SRAM leakage. Leakage components illustrated in Fig. 26.5.4(b) show that both BL and cell leakage are significantly reduced by raising VGN. TDECAY is measured from the self-decay circuit at different temperatures and the results are shown in Fig. 26.5.4(c). The dynamic range of TDECAY was 7.6X as the temperature is increased from 20°C to 75°C. As leakage increases with temperature, TDECAY is reduced which, in turn, makes the SRAM enter the sleep state more frequently.

Also implemented is a test circuit to explore the impact of the sleep transistor on the static noise margin (SNM). Interestingly, the measurement results in Fig. 26.5.6(a) show that the read SNM improves with a sleep transistor. For a GSIZE of 1.0 that gives a 2% performance penalty in the designed 16kB SRAM and the improvement in read SNM is 25%. Here, GSIZE indicates the size of the sleep transistor normalized to the size of the NMOS driver in the 6T SRAM cell. To understand this counter-intuitive observation, the voltage transfer characteristics and the virtual ground voltage of a 6T SRAM cell with and without a sleep transistor is shown in Fig. 26.5.6(b). When a sleep transistor is inserted, VGN rises at the right end of the figure causing VGS to rise, (A) in Fig. 26.5.6(b), which has a degrading effect on read SNM. However, the rise in VGN at the center of the figure makes it harder for the SRAM to flip since the source voltage of the NMOS driver transistor also rises. This results in an increase in DC gain as shown in (B) in Fig. 26.5.6(b), improving the overall read SNM by 25%. The reason why the effect of (B) prevails over (A) is because VGN always rises higher at the center than at the right end due to larger pull-up current. Finally, the improvement in SNM for different SRAM cell structures is verified showing that, in general, the principle holds for different SRAM designs. Simulation results in Fig. 26.5.7 show 18-129mV improvement in SNM depending on the beta ratios Q (βPMOS/βPMOS) and R (βPMOS/βPMOS).

Acknowledgments:
This work was performed at Purdue University and has been sponsored by DARPA PAC/C program, Semiconductor Research Corporation (contract #1078.001) and Intel Ph.D. fellowship program.

References:
Figure 26.5.1: (a) Source-biased gate-ground SRAM cell for leakage reduction. (b) Proposed active leakage reduction scheme which periodically shuts off SRAM by a sleep pulse. (c) Proposed self-decay circuit that adaptively changes the interval between sleep pulses for optimal leakage savings under varying conditions.

Figure 26.5.2: Organization of 16kB SRAM with self-decay scheme.

Figure 26.5.3: Statistical leakage reduction (overhead included) of proposed self-decay scheme compared to conventional and fixed decay scheme.

Figure 26.5.4: Measurement data from 16kB SRAM testchip. (a) Reduction of leakage components as VGND is raised. (b) Leakage comparisons between conventional and this work. (c) Self-decay period measured at different temperatures. (d) Read access frequency versus VDD.

Figure 26.5.5: (a) Static noise margin (SNM) versus sleep transistor size measured at 1.8V and room temperature. Sleep transistor for leakage reduction offers 25% higher SRAM read SNM. (b) Measured SRAM butterfly curves and virtual ground voltage for this work (with sleep transistor) and conventional (without sleep transistor) showing improvement in read SNM.
Figure 26.5.6: (a) Static noise margin with and without sleep transistor for different SRAM sizing. (b) Improvement in SNM ranging from 18-129mV.

Figure 26.5.7: Chip microphotograph and details of the 16kB SRAM testchip with self-decay based leakage reduction scheme.