18.2 A 1.1GHz 12μA/Mb-Leakage SRAM Design in 65nm Ultra-Low-Power CMOS with Integrated Leakage Reduction for Mobile Applications

Y. Wang, H. Ahn, U. Bhattacharya, T. Coan, F. Hamzaoglu, W. Hafez, C.-H. Jan, P. Kolar, S. Kulkarni, J. Lin, Y. Ng, I. Post, L. Wei, Y. Zhang, K. Zhang, M. Bohr

Intel, Hillsboro, OR

Growing demand for high-performance multi-media processor in handheld devices continues to drive the need for large embedded SRAM with low power consumption. The low-power requirement has often been addressed by compromising SRAM performance through the adoption of lower supply voltage and slower performing transistors with low leakage. Supply voltage scaling in large SRAM has not kept pace with technology scaling due to reduced SRAM cell stability and write margin [1]. Many circuit techniques have been proposed [2, 3, 4] to expand design margins for low-voltage operation while reducing leakage current at standby mode. In this work, we report an SRAM design in a 65nm ultra-low-power (ULP) CMOS process [5]. Co-optimization between process technology and circuit design allowed us to achieve fast SRAM access at nominal operating voltage. A large design window in operating voltage along with integrated leakage reduction schemes provides flexibility for both active power management and leakage reduction through dynamic voltage scaling.

The 65nm ULP CMOS process, featuring uni-axial strained silicon transistors, provides a significantly improved trade-off between transistor drive current and leakage [5]. This is essential to achieve fast SRAM access at substantially lower operating voltage and leakage. Standby leakage of the SRAM cell is optimized by simultaneously reducing gate, subthreshold and junction leakage components to sub-100pA/μm at 1.2V, while maintaining competitive cell read current and meeting the retention voltage requirement for a 1Mb SRAM. The low damage junction engineering employed in this technology results in substantial junction leakage reduction while still maintaining good short channel control. Gate oxide thickness optimization and gate nitridation are applied to reduce gate leakage. Well and pocket implants and source/drain spacers are optimized simultaneously to reduce subthreshold leakage. To achieve minimum supply voltage (V_{min}) needed for the target transistor I_{on}/I_{off} performance in SRAM cell and peripheral logic, separate V_t control for N/P transistors in SRAM cells and peripheral circuit is employed.

Figure 18.2.1 shows the configuration and key design features of a 1Mb ULP SRAM macro. To allow SRAM macro running at a lower supply voltage to reduce system power consumption, the macro uses a 0.667μm^2 SRAM cell, which is larger than the minimum that can be supported in this technology. The cell dimension is optimized to achieve high array efficiency (78%) and bit density (115Mb/cm^2) for a 128kb sub-array with improved static noise margin, write margin and read current at low-voltage design point. Each of the eight 128kb sub-arrays in the 1Mb SRAM macro has four banks, with each bank containing a 256-row × 128-column array and one redundant bit. The local read circuitry in column I/O is designed for 4:1 column interleaving. A 2:1 MUX function is built into the sense amplifier and its enable signal is generated by ANDing the read and global column select signals to reduce switching power. Transistor stacking and long-channel transistors are used extensively to save standby leakage in peripheral circuits.

The fully synchronized 128kb sub-array is designed to have single-cycle latency to meet mobile product needs. A fully static circuit design style is adopted to achieve robust read and write design margin for a wide operating voltage window. Figure 18.2.2 shows the schematic of critical read-path circuitry and its timing diagram. The control signals are all triggered by the clock edges. To achieve single-cycle latency, a decoupled sense amplifier is adopted where the inputs and outputs of the sense amplifier are separated. The sense amplifier inputs can be precharged without affecting data at its outputs. A static SR-latch is used to latch the sense amplifier output. A tri-state write driver is used to reduce leakage when the sub-array is in standby mode.

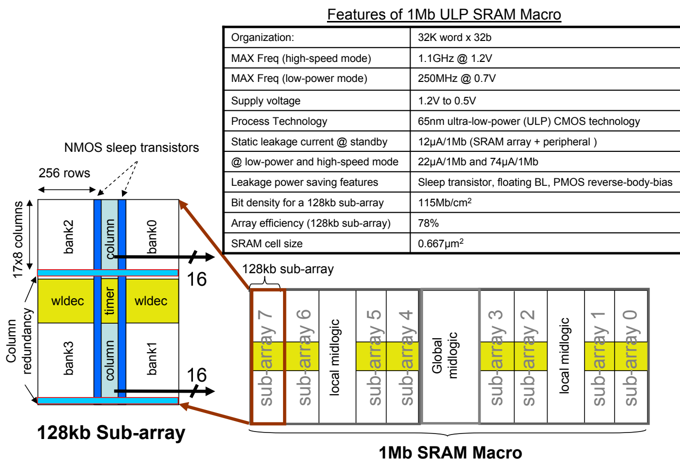
Figure 18.2.3 shows the integrated leakage reduction schemes incorporated in this design, including the programmable NMOS sleep bias transistor [6], floating bitline and SRAM PMOS reverse-body-bias. N-sleep transistor is turned on only in the activated 128kb sub-array, and it is turned off for the rest of unselected sub-arrays in the 1Mb SRAM to reduce array leakage. The virtual ground voltage of the SRAM array in sleep mode is controlled by programmable NMOS bias transistors to achieve optimal voltage control across process skews. The SRAM transistor is optimized with low junction leakage to maximize the benefit of the sleep transistors in reducing subthreshold leakage. The NMOS sleep transistor is distributed along the edges of four SRAM banks. The programmable sleep bias transistor is located in the timer and is shared among four array banks. The bitline precharge devices are turned off during standby mode to float the bitlines. The array leakage is reduced as the voltage level drops at the bitline. The well bias of SRAM PMOS transistors is controlled through a separate power line (V_{NWELL}). During standby mode, the n-well voltage is maintained at the nominal value of 1.2V while the array supply voltage is lowered to 0.5V. The reverse-biased S/D junction of SRAM PMOS transistors results in substantial reduction of subthreshold leakage.

A 44Mb SRAM chip containing the 1Mb ULP SRAM macro with column redundancy, electrical-programmable fuse, and programmable built-in-self-test (PBIST) is built and packaged in an 8-metal layer land-grid-array (LGA) package with flip-chip technology. Figure 18.2.4 shows the measured average cell leakage from an SRAM array with all three leakage reduction schemes turned on. By lowering supply to the retention voltage of 0.5V and maintaining the PMOS n-well voltage at the nominal voltage of 1.2V, the measured average cell leakage is reduced to ~2pA/bit. PMOS reverse-body-bias reduces the cell leakage by ~30% at the retention voltage level, mainly due to the reduction in PMOS subthreshold leakage. Figure 18.2.5 shows the standby leakage of the 1Mb SRAM operating in standby, low-power and high-speed modes. In standby mode, the 1Mb SRAM macro draws ~12μA of leakage current. The leakage current increases to 22μA for low-power operation at 0.7V and 74μA for high-speed operation at 1.2V. For high-speed operation, by turning on the NMOS sleep feature and floating bitlines in unselected sub-arrays, the array leakage current of the 1Mb SRAM macro is reduced by ~70%. Figure 18.2.6 shows the maximum measured frequency of the 1Mb SRAM across a wide range of supply voltages. It achieves 1.1GHz frequency at a nominal voltage of 1.2V and 250MHz at 0.7V, which represents the highest reported frequency for the same class of standby power consumption for mobile applications. The die micrograph of the 44Mb ULP SRAM is shown in Fig. 18.2.7.

Acknowledgements:
The authors are grateful to many members of Advanced Memory Design team and PTD technical staff for their contributions to this work.

References:
[1] T. Kawahara, "Low-Voltage Embedded RAMs in Nanometer Era," *IEEE Conf. Electron Dev. and Solid-State Circuits*, pp. 333-338, Dec., 2005.
[2] M. Yamaoka, N. Maeda, Y Shinozaki, et al., "Low-Power Embedded SRAM Modules with Expanded Margins for Writing," *ISSCC Dig. Tech. Papers*, pp. 480-481, Feb., 2005.
[3] Y. Takeyama, H. Otake, O. Hirabayashi, et al., "A Low Leakage SRAM Macro with Replica Cell Biasing Scheme," *Symp. VLSI Circuits*, pp. 166-167, Jun., 2005.
[4] T. Suzuki, Y. Yamagami, I. Hatanaka, et al., "A Sub-0.5-V Operating Embedded SRAM Featuring a Multi-Bit-Error-Immune Hidden-ECC Scheme," *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 152-160, Jan., 2006.
[5] C.-H. Jan, P. Bai, J. Choi, et al., "A 65nm Ultra Low Power Logic Platform Technology Using Uni-axial Strained Silicon Transistors," *IEDM Tech. Dig.,* pp. 65-66, Dec., 2005.
[6] K. Zhang, U. Bhattacharya, Z. Chen, et al., "SRAM Design on 65-nm CMOS Technology with Dynamic Sleep Transistor for Leakage Reduction," *IEEE J. Solid-State Circuits*, pp. 895-901, Apr., 2005.

 1-4244-0852-0/07/$25.00 ©2007 IEEE.



Features of 1Mb ULP SRAM Macro

Organization:	32K word x 32b
MAX Freq (high-speed mode)	1.1GHz @ 1.2V
MAX Freq (low-power mode)	250MHz @ 0.7V
Supply voltage	1.2V to 0.5V
Process Technology	65nm ultra-low-power (ULP) CMOS technology
Static leakage current @ standby	12μA/1Mb (SRAM array + peripheral)
@ low-power and high-speed mode	22μA/1Mb and 74μA/1Mb
Leakage power saving features	Sleep transistor, floating BL, PMOS reverse-body-bias
Bit density for a 128kb sub-array	115Mb/cm²
Array efficiency (128kb sub-array)	78%
SRAM cell size	0.667μm²

Figure 18.2.1: Configuration and features of 1Mb SRAM with integrated sleep transistor.

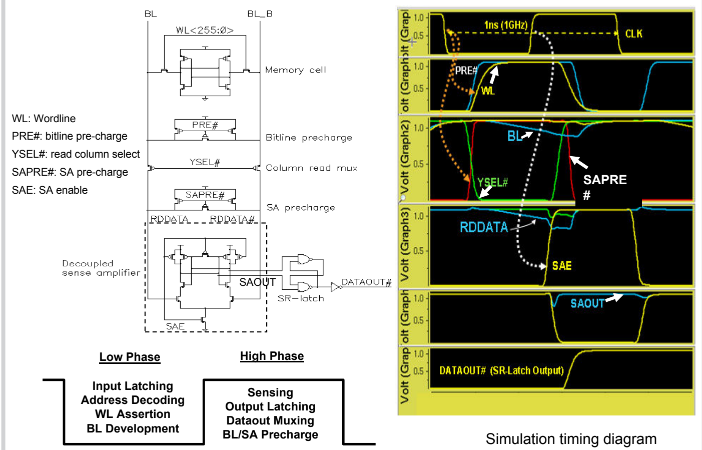


Figure 18.2.2: Read path schematic, simulation timing diagram, and single-cycle throughput pipeline of 128kb sub-array.

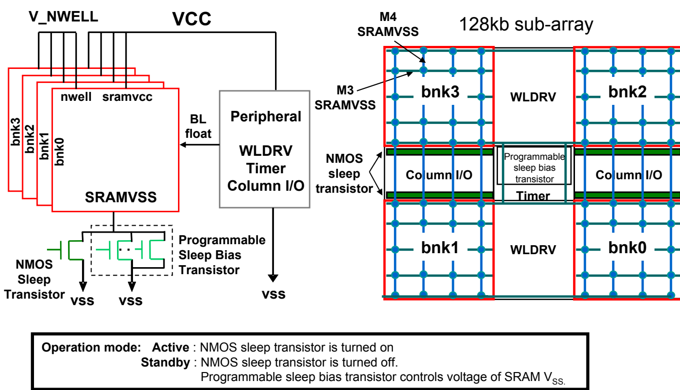
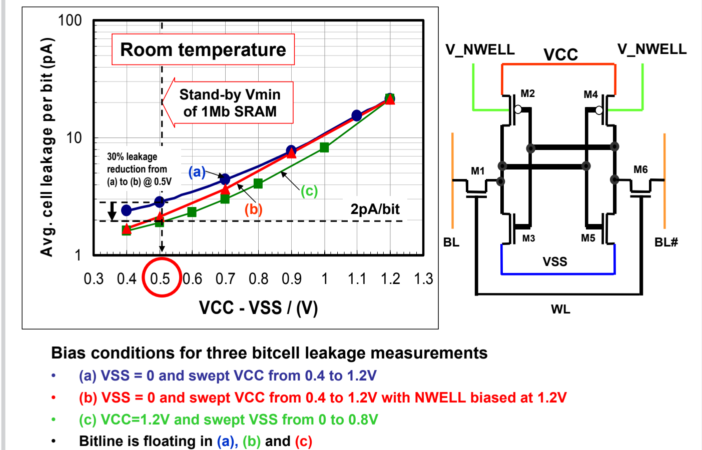


Figure 18.2.3: SRAM array leakage reduction features, and designs of sleep, bias transistors and power lines in a 128kb sub-array.



Bias conditions for three bitcell leakage measurements

- (a) VSS = 0 and swept VCC from 0.4 to 1.2V
- (b) VSS = 0 and swept VCC from 0.4 to 1.2V with NWELL biased at 1.2V
- (c) VCC=1.2V and swept VSS from 0 to 0.8V
- Bitline is floating in (a), (b) and (c)

Figure 18.2.4: Measured SRAM cell leakage at various leakage reduction schemes.

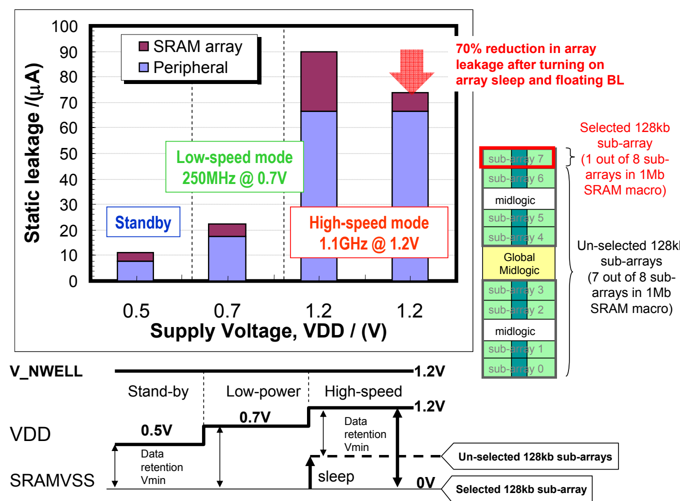


Figure 18.2.5: Standby leakage current of 1Mb SRAM macro.

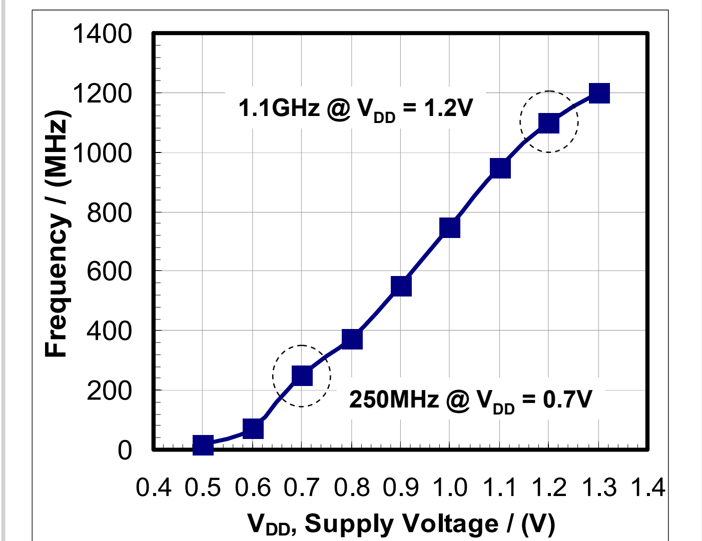


Figure 18.2.6: Maximum access frequency vs. supply voltage of 1MbULP SRAM macro.

Continued on Page 606

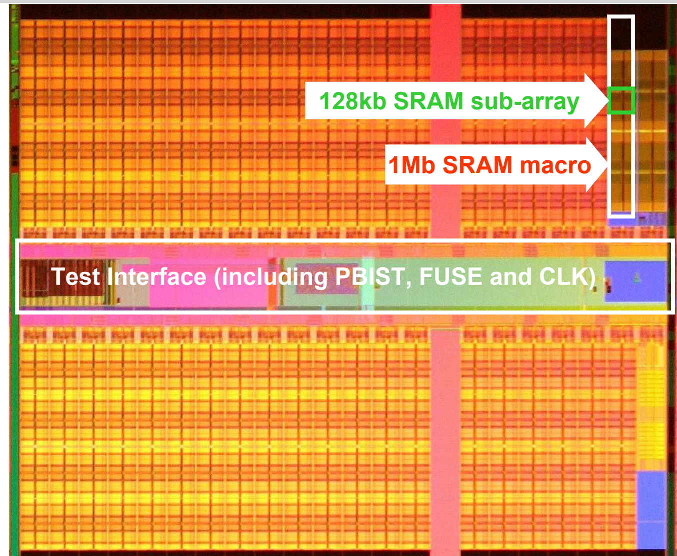


Figure 18.2.7 Die micrograph of 44Mb ultra-low-power SRAM.

 1-4244-0852-0/07/$25.00 ©2007 IEEE.