18.4 A 64Mb SRAM in 22nm SOI Technology Featuring Fine-Granularity Power Gating and Low-Energy Power-Supply-Partition Techniques for 37% Leakage Reduction

Harold Pilo[1], Chad A. Adams[2], Igor Arsovski[1], Robert M. Houle[1], Steven M. Lamphier[1], Michael M. Lee[1], Frank M. Pavlik[1], Sushma N. Sambatur[3], Adnan Seferagic[1], Richard Wu[1], Mohammad I. Younus[4]

[1]IBM, Essex Junction, VT, [2]IBM, Rochester, MN, [3]IBM, Bangalore, India, [4]IBM, Hopewell Junction, NY

A 64Mb SRAM is fabricated in a 22nm high-performance SOI technology [1]. The ever-increasing integration needs of complex SoC are driving the reduction of SRAM leakage power and increase in memory density. While the area and leakage power benefits of eDRAM continue to be leveraged in applications with large contiguous memory blocks [2], SRAM leakage remains a significant portion of the total SoC power. This work describes an SRAM that is optimized for leakage and performance as top priorities over density. The SRAM features a new bitcell (BC) implemented with a fine-granularity power-gating (FGPG) technique to reduce BC leakage by 37%. FGPG improves leakage reduction by 2× compared to bank-based power-gating (PG) techniques [3-4]. Periphery leakage is also reduced by 40% from the previous design [5] with a low-energy power-supply-partition design that leverages higher V_t devices operating at a higher supply voltage. This scheme alone provides an 8% improvement in performance with a small compromise to the AC power.

Figure 18.4.1 shows the 0.128μm² BC configured to support FGPG. Power-supply lines V_VCS1 and V_VCS2 are routed in metal-3 in parallel with the wordline (WL). Larger BC devices are used to improve VCS_{MIN} voltage and performance; the corresponding increase in BC width is leveraged to fit the two required VSS lines in the metal-2 vertical direction in addition to the bitlines (BL) and connections for pull-up (PU1/PU2) and WL contacts. The horizontal power-supply lines enable a small group of rows to be powered independently within the memory bank. To be user-transparent, PG techniques must rely on a local charge reservoir to instantly replenish the charge associated with waking up a bank [3]. Decreasing the memory domain powered up during activation from 128 rows to eight rows not only reduces leakage significantly, but it also lowers the wake-up current and AC power penalty associated with memory wake-up. The improved power efficiency of FGPG enables its use and leakage savings across all memory sizes, which is not feasible with the bank-based approach.

Figure 18.4.2 shows the schematic of the FGPG scheme. Two four-WL groups are shown with PG logic and WL driver. The output of a 4-WL group pre-decode block, GRP0 is shared between WL driver selection and PG device selection (N11/N10 in lower section). GRP0 selects a group of four WL drivers (I1-I4), which are decoded by DEC0-DEC3 (see WL3 in Fig. 18.4.2). GRP0 also selects four PG PFETs, P10-P13 to supply VCS to four V_VCS0 rows. Each row of BCs shares a power supply line with its neighboring row, and every fourth row straddles two different PG domains (WL3 and WL7 in Fig. 18.4.2). When a four-row group is selected, the upper-adjacent four-row power domain is also selected to guarantee full power to every BC within the four-row section. This is highlighted by the turn-on of N11 and N12 in Fig. 18.4.2, which in turn select P10-P13 and P14-P17, respectively. An important consideration of the FGPG technique is the management of the retention voltage. The larger the voltage drop on the V_VCS power domains, the greater the leakage savings; too large of a drop causes data disturbs in the half-powered (HP) rows. HP rows straddle two four-row power domains with unselected WLs. This is illustrated by WL7 in Fig. 18.4.2; V_VCS1 is at VCS, but V_VCS2 is in retention, 150mV below VCS. The unselected BCs connecting to WL7 are biased with PU1 powered to VCS, while PU2 is biased 150mV below VCS. There are several considerations in choosing a four-row PG domain. Combining the strength of four PFET PG devices facilitates the use of FGPG for the WL driver without degrading its performance. The WL pull-up devices (P1-P9 in Fig. 18.4.2) contribute the highest amount of leakage after the BC. Enabling FGPG for the WL driver helps achieve the 40% leakage savings in the periphery. The four-WL group pre-decode circuit is shared between the WL driver and the PG selection logic; this keeps the area overhead of FGPG to 1.5%. Finally, a larger overall PFET decreases the variability and data-dependency of the retention voltage, which is essential in maintaining robust margins for HP rows.

Figure 18.4.3 shows a detailed schematic of the FGPG driver for four rows. During active mode (PG_ENABLE=VCS, DS=VSS) the four-WL decode output, GRP turns on N11 to switch V_VCSN to VSS and charge V_VCS network to VCS. GRP_NEXT is connected to the next driver to select the adjacent V_VCS. During PG mode, GRP is VSS and P1/P2 are on to connect V_VCSN to V_VCS. This biases P10-P13 to a diode configuration to allow V_VCS to discharge (by the BC leakage) to 150mV below VCS. The circuit also has provisions for turning PG off during stress tests (PG_ENABLE=VSS) or enabling deep sleep (DS=VCS; array data is lost). Figure 18.4.4 shows the test results for an 8Mb section of the test chip at 1.1V and 85°C. Leakage savings during deep sleep (X) and FGPG (Z) are compared to full power (O). An 80% reduction in leakage is measured for <50% of the die. While 2× greater than the model expectation, it is attributed to very slow logic PFETs in the hardware. Improved leakage savings is also observed as leakage increases. The bottom-right section of Fig. 18.4.4 also shows simulation waveforms of FGPG during WL selection.

SRAMs have unique requirements that must be considered when choosing the voltage operating point and the dominant V_t-type of the constituent transistors. Most SRAMs today have dual supply voltages; a higher BC voltage enables the logic supply voltage to continue to scale. By judiciously choosing where to partition these domains and the device V_t types, the optimum power-performance product can be achieved. Figure 18.4.5 compares the relative performance and leakage of normal-V_t devices (NV_t) and high-V_t devices (HV_t) at several voltages. The gate delay at the slow performance corner (left graph) is 8% faster for HV_t logic operating at 100mV higher than NV_t logic. The graph on the right shows that the worst-case leakage is 35% lower for an HV_t device operating at 100mV higher than that of the NV_t device. To minimize total power consumption and improve performance, the voltage-domain partitioning of Fig. 18.4.6 is developed. Low-energy circuits (light boxes) are placed on the higher VCS domain employing the HV_t devices. These circuits fall into two categories, 1) lightly loaded and switching every cycle (e.g., control timing and address-decode circuits) and 2) large loads but operating at a very low duty cycle (e.g., WL driver). Using the higher supply voltage for the WL driver as well as signals controlling the column circuitry is essential in maintaining timing relationships when operating at high frequencies. The dark boxes on the right of the diagram show circuitry that consumes a large fraction of the overall AC power; these are assigned to the lower voltage domain (VDD) and include the internal 128b I/O bus. Careful consideration is placed on minimizing area relating to level-shift circuitry (cross-hatched) that translate from the VDD domain to the higher VCS domain. The WL-redundancy compare logic is kept at VDD to avoid large numbers of level-shift circuits.

Figure 18.4.7 shows the die micrograph of the 64Mb SRAM with one 512Kb instance magnified. To illustrate the leakage improvement of FGPG, a 128-WL × 560-BL bank is highlighted, which coincides with the active bank of the bank-based approach [3]. This is compared to FGPG where only eight rows are activated. The 512Kb macro features are also summarized in Fig. 18.4.7.

References:

[1] Narasimha, S., et al., "22nm High-Performance SOI Technology Featuring Dual-Embedded Stressors, Epi-Plate High-K Deep-Trench Embedded DRAM and Self-Aligned Via 15LM BEOL", *IEDM*, Dec. 2012.

[2] Barth, J., et al., "A 45nm SOI Embedded DRAM macro for POWER7™ 32MB On-Chip L3 Cache", *ISSCC*, pp. 342-343, Feb. 2010.

[3] Pilo, H., et al., "A 450ps Access-Time SRAM Macro in 45nm SOI Featuring a Two-Stage Sensing-Scheme and Dynamic Power Management," *ISSCC*, pp. 378-379, Feb. 2008.

[4] Hamzaoglu, F., et al., "A 153Mb-SRAM Design with Dynamic Stability Enhancement and Leakage Reduction in 45nm High-K Metal-Gate CMOS Technology", *ISSCC*, pp. 376-3779, Feb. 2008.

[5] Pilo, H., et al., "A 64Mb SRAM in 32nm High-k Metal-Gate SOI Technology with 0.7V Operation Enabled by Stability, Write-Ability and Read-Ability Enhancements," *ISSCC*, pp. 254-255, Feb. 2011.

 978-1-4673-4516-3/13/$31.00

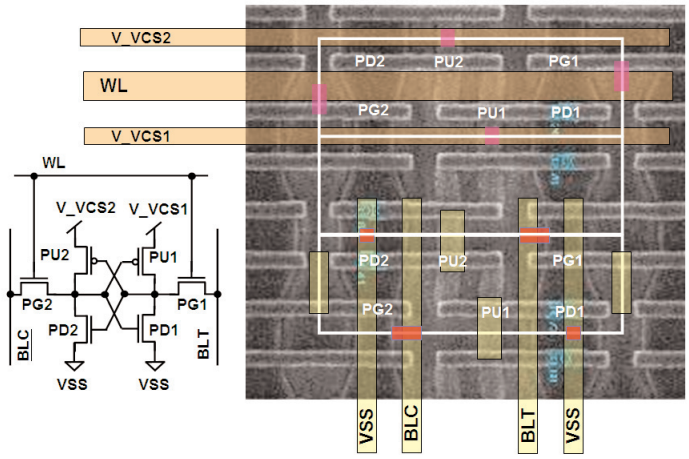


Figure 18.4.1: 22nm Bitcell with horizontal power lines.

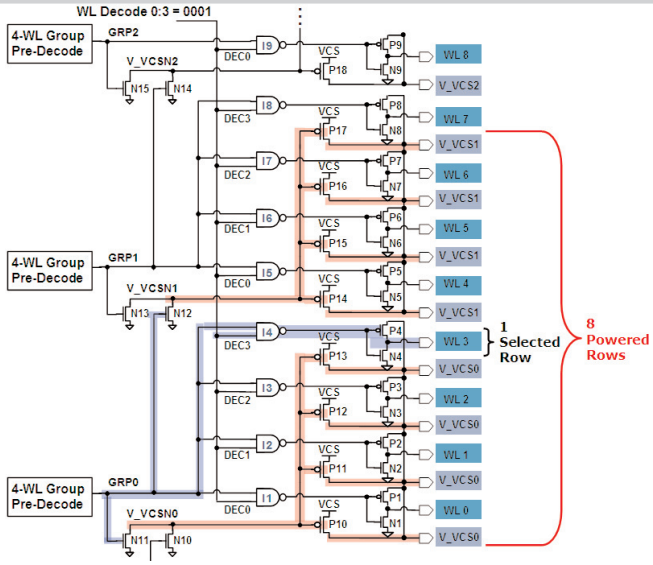


Figure 18.4.2: Fine-granularity power gating.

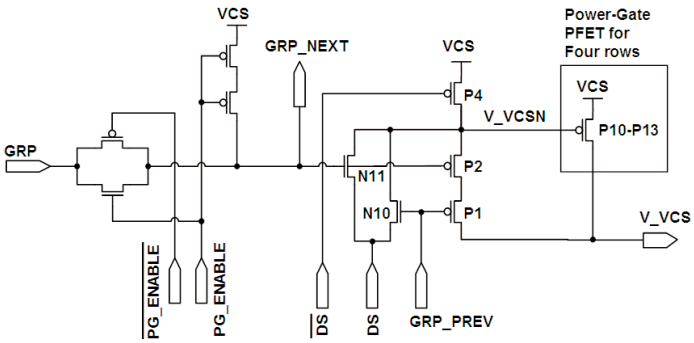


Figure 18.4.3: Power-gate device driver.

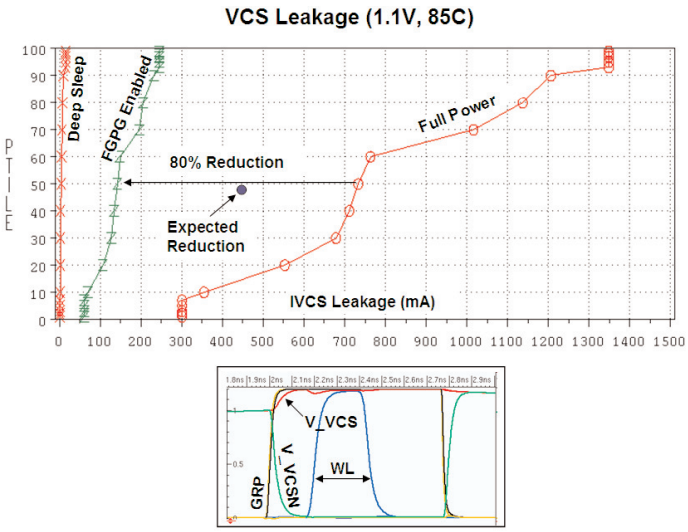
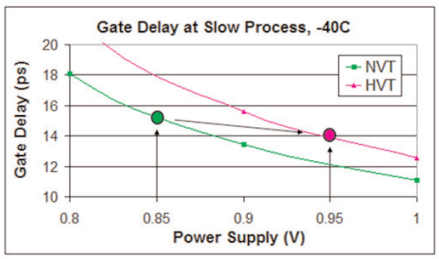


Figure 18.4.4: FGPG hardware results and waveforms.



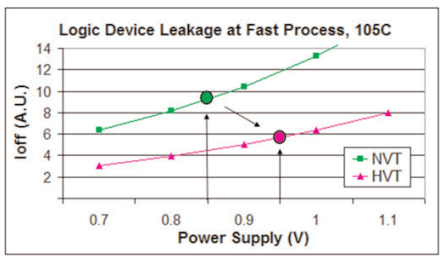


Figure 18.4.5: Low-energy power-supply partition: rationale for high-V_t (HV_t) use.

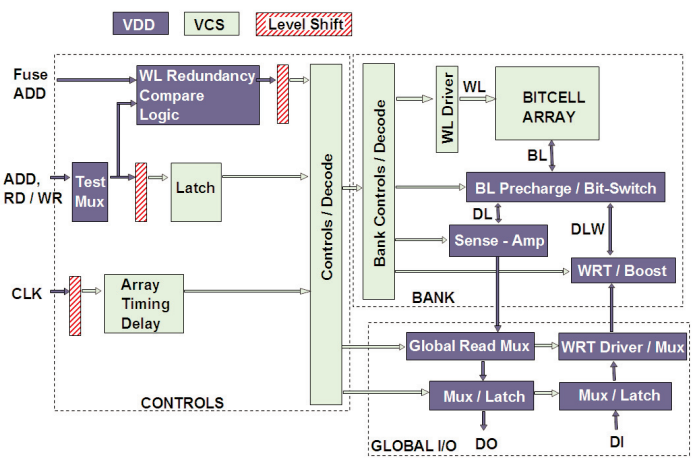
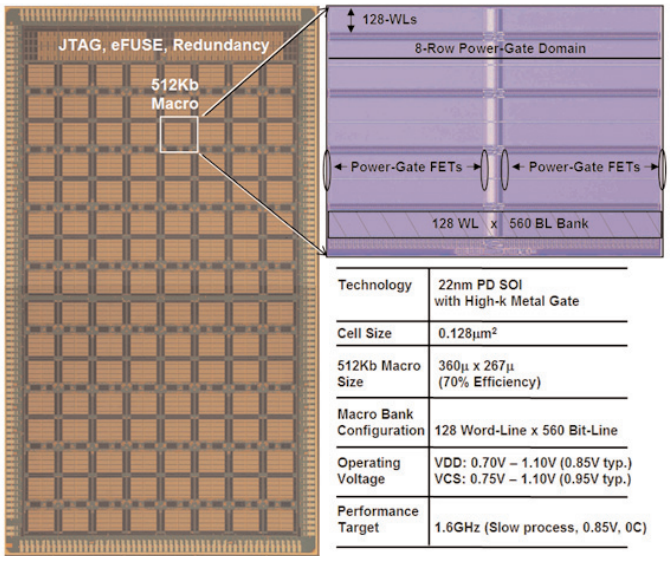


Figure 18.4.6: Low-energy power-supply partition.

18



Technology	22nm PD SOI with High-k Metal Gate
Cell Size	0.128μm²
512Kb Macro Size	360μ x 267μ (70% Efficiency)
Macro Bank Configuration	128 Word-Line x 560 Bit-Line
Operating Voltage	VDD: 0.70V – 1.10V (0.85V typ.) VCS: 0.75V – 1.10V (0.95V typ.)
Performance Target	1.6GHz (Slow process, 0.85V, 0C)

Figure 18.4.7: Micrograph of 64Mb test-chip and features.

978-1-4673-4516-3/13/$31.00