11.2 A 1Mb Embedded NOR Flash Memory with 39µW Program Power for mm-Scale High-Temperature Sensor Nodes

Qing Dong[1], Yejoong Kim[1], Inhee Lee[1], Myungjoon Choi[1], Ziyun Li[1], Jingcheng Wang[1], Kaiyuan Yang[1], Yen-Po Chen[1], Junjie Dong[1], Minchang Cho[1], Gyouho Kim[1], Wei-Keng Chang[2], Yun-Sheng Chen[2], Yu-Der Chih[2], David Blaauw[1], Dennis Sylvester[1]

[1]University of Michigan, Ann Arbor, MI
[2]TSMC, Hsinchu, Taiwan

Miniature sensor nodes are ideal for monitoring environmental conditions in emerging applications such as oil exploration. One key requirement for sensor nodes is embedded non-volatile memory for compact and retentive data storage in the event that the sensor power source is exhausted. Non-volatile memory also allows for near-zero standby power modes, which are particularly challenging to achieve at high temperatures when using SRAM in standby due to the exponential rise in leakage with temperature, which rapidly degrades battery life (Fig. 11.2.1). However, traditional NOR flash requires mW-level program and erase power, which cannot be sustained by mm-scale batteries with internal resistances >10kΩ To address this issue, we propose an ultra-low power NOR flash design and demonstrate its integration into a complete sensor system that is specifically designed for environmental monitoring under high temperature conditions: such as when injected into geothermal or oil wells.

The proposed flash design reduces power consumption by using a combined Dickson and ladder pump topology and low-parasitic MIM capacitors to generate 13V with 73% power efficiency, and a cross-sampling current sense amplifier (SA) that doubles the sensing margin. Measured program and erase energy is 49pJ/b and 9.4pJ/b, resulting in a 30× and 22× reduction compared to a standard flash macro. Program power is 39µW and 82µW at 25°C and 125°C, enabling a miniature sensor node system to be powered by only two 8µAh batteries. The complete measured system is 3.88×1.70×1.85mm^3 and consumes only 190nW of power at 125°C in the flash-enabled deep sleep mode, resulting in a 63× power reduction compared to a conventional sensor design.

Embedded split-gate NOR flash requires >10V for hot-carrier-injection-based program and tunneling-based erase. Circuits to generate these high voltages dominate the write power [1,2]. A cross-coupled Dickson charge pump offers the best power efficiency, but the voltage across the flying capacitor increases linearly in each stage, making it necessary to use high-voltage MOS capacitors. However, these HV MOS capacitors have a parasitic/useful capacitance ratio of 46% (NMOS) and 18% (PMOS), lowering pump efficiency [3]. MIM capacitors offer a very low (1%) parasitic loss, but are limited to <3.6V operation. Using a combined Dickson and Cockcroft-Walton ladder pump [3] (Fig. 11.2.2) allows MIM capacitors to be used while maintaining high power efficiency. A single-stage Dickson pump and a four-stage ladder structure generate an output voltage of 13V. A PMOS body switch is used to avoid diode leakage through the N-well, and a regulation loop is used to stabilize the output voltage. The regulation circuit uses a dual-V_{DD} approach (1.2V and 2.5V) to reduce power by 30%. A V_t-based voltage reference generation circuit [4] provides V_{ref} with sub-nW power consumption. Start-up, erase and program operations need a high VCO frequency (>15MHz) to stabilize V_{out}, and hence a high-BW amplifier is required. However, read and standby modes do not require a high-BW amplifier, and therefore amplifier tail current can be lowered in these modes, reducing the total standby power by ~2×. A high-resistance diode-chain divider requires low power, but has a long stabilization time. To address this issue, capacitors are placed in parallel with the diode chain to stabilize the loop within 1µs. Figure 11.2.2 shows the 73% peak-power efficiency of the pump loop. The pump itself reduces power by ~4× compared with the baseline design; the whole loop achieves ~7× power reduction using all of the low-power methods mentioned above.

Figure 11.2.3 shows the block diagram of the 1Mb flash macro, which is separated into two banks, each with its own power gating control. When one bank is active, all of the peripherals in the other banks are power gated. Each bank has two arrays with 256×1024 cells. Current SAs, reference current generation, and high-voltage switches are shared by the two arrays. Page-wise erase mode operates on 8kb, whereas only 8b are programmed or read at a time to lower the instantaneous power; this is consistent with common data resolution in sensor systems (e.g., 8b temperature readings). The program power is reduced by ~5×, by using power-gating and a short word length, compared with the baseline design.

At high temperatures, the read current of an erased cell degrades with reduced mobility, whereas that of a programmed cell increases due to its lowered V_{th}. The read current ratio between the two states thus reduces from 8× to 5× as the temperature increases from 25°C to 125°C, complicating SA design at high temperatures. Figure 11.2.3 shows circuit and timing diagrams of a proposed current SA with cross-sampling that doubles the sensing margin. After precharge, (1) S_1 turns on, integrating I_{ref} on C_L (N_0) and I_{cell} on C_R (N_1). (2) S_2 turns on, and S_1 turns off to maintain the gate voltages stored on C_L and C_R, and I_{ref} and I_{cell}. When SAEN goes high, I_{cell}-I_{ref} flows out from N_2, while I_{cell}-I_{ref} flows into N_3. The current difference (I_{cell}-I_{ref}) is therefore doubled when the latch-based SA is activated, and mismatch between M_0 and M_1 is cancelled with current sampling instead of voltage sampling [5]. The proposed method halves the current offset compared to a conventional method with same transistor sizing (Fig. 11.2.4 top left). The cross-sampling gate capacitors occupy 16% area of the sense amplifier, which is <0.1% of the whole macro.

The flash macro is fabricated in 90nm embedded ESF3 NOR flash technology. A conventional compiled flash using the same bitcell is also fabricated for baseline comparison. Figure 11.2.4 (top right) shows the measured Shmoo plot of the proposed flash macro achieving 11ns access time at 1.2V and 0.75V read V_{DDmin}. Measured average read V_{DDmin} among 10 dies are 0.739V (Fig. 11.2.4 bottom left). V_{DDmin} across –25°C to 125°C is shown in Fig. 11.2.4 (bottom right). Measured erase and program power at 25°C is 15µW and 39µW, which represent a 242× and 87× reduction compared to the baseline design. The erase and program energy is 9.4pJ/b and 49pJ/b, which are 30× and 22× lower than the baseline design. At 125°C, the program and erase power is 82µW and 31µW, enabling reliable function of the battery-powered sensor system. At 11ns cycle time, read energy (power) is 2.2pJ/b (1.618mW). Using a read-cycle time suitable for sensor nodes (1us), the design consumes 25µW and shows better power/frequency scaling than the baseline design due to its lower leakage floor. Standby power is 5.4µW even with active charge pump regulation loops; a 4.5× reduction over the baseline design. Figure 11.2.5 compares the measurement results with the baseline design and other work.

The flash macro is incorporated into a high temperature mm-scale sensor system that consists of multiple chip layers: two batteries, flash, PMU, processor, decoupling capacitors, radio, temperature sensor, energy harvester, and a solar cell layer (photo in Fig. 11.2.6). The FSM in the PMU layer turns on the battery switch (SWn=0) once the wake-up timer reaches N_CYCLE specified in the register file (Fig. 11.2.6, top right). Once the battery switch is on, the PMU boots and provides output voltages to the system using switch-capacitor DC-DC converters. The flash layer then automatically programs the processor layer, and the processor executes the sensor program and stores recorded data in flash. It then writes a 'deep sleep' command to register file (SLP_BIT), triggering the FSM to turn off the battery switch (SWn=1) and power gate all of the blocks except for wake-up timer, FSM, and register file. Figure 11.2.6 (top left) shows the simulated power breakdown in deep sleep mode. Measured active system power drawn from the battery is 25µW and 430µW at 25°C and 125°C. Sleep power is 12nW and 190nW at 25°C and 125°C, representing a 3.3× and 63× reduction compared to a conventional system and greatly extending battery lifetime. The stacked system is fully functional at 125°C in stand-alone operation. Figure 11.2.7 shows die and system photos.

Acknowledgements:
This work was supported by TSMC university joint development program and university shuttle program.

References:

[1] M. F. Chang, et al., "A Process Variation Tolerant Embedded Split-Gate Flash Memory Using Pre-Stable Current Sensing Scheme," *IEEE JSSC*, vol. 44, no. 3, pp. 987-994, Mar. 2009.
[2] H. Mitani, et al., "A 90nm Embedded 1T-MONOS Flash Macro for Automotive Applications with 0.07mJ/8kB Rewrite Energy and Endurance Over 100M Cycles Under Tj of 175°C," *ISSCC*, pp. 140-141, Feb. 2016.
[3] T. Ishii, et al., "A 126.6-mm2 AND-Type 512-Mb Flash Memory With 1.8-V Power Supply," *IEEE JSSC*, vol. 36, no. 11, pp. 1707-1712, Nov. 2001.
[4] Q. Dong, et al., "A 114-pW PMOS-Only, Trim-Free Voltage Reference with 0.26% within-Wafer Inaccuracy for nW Systems", *IEEE Symp. VLSI Circuits*, pp. 98-99, June 2016.
[5] M. Jefremow, et al., "Time-Differential Sense Amplifier for Sub-80mV Bitline Voltage Embedded STT-MRAM in 40nm CMOS", *ISSCC*, pp. 216-218, Feb. 2013.

 978-1-5090-3758-2/17/$31.00

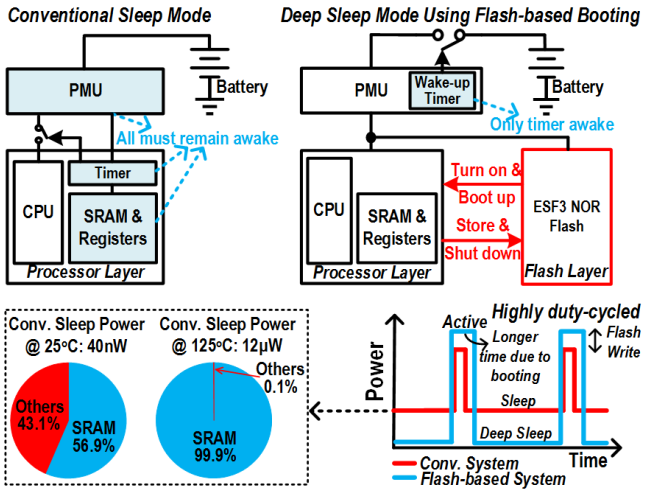


Figure 11.2.1: An SRAM-based sensor system drains a mm-scale battery in minutes at 125°C. A flash-based sensor system can enter deep sleep, reducing power consumption and extending battery life.

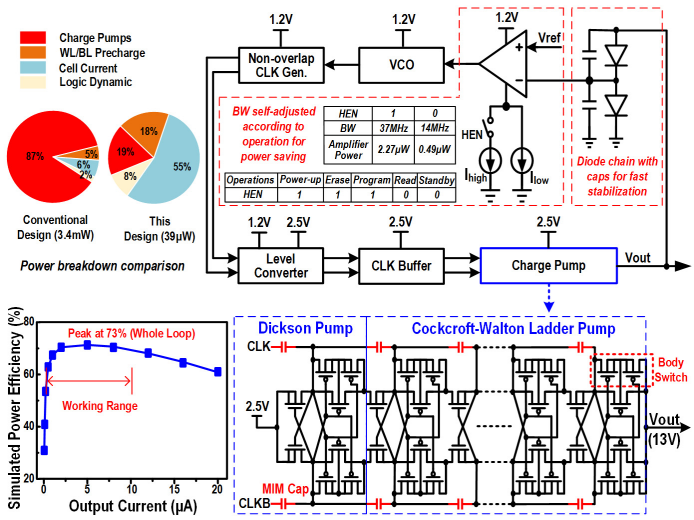


Figure 11.2.2: Combined Dickson and ladder pump topology, and self-adjusted regulation loop, which achieves a 73% peak efficiency.

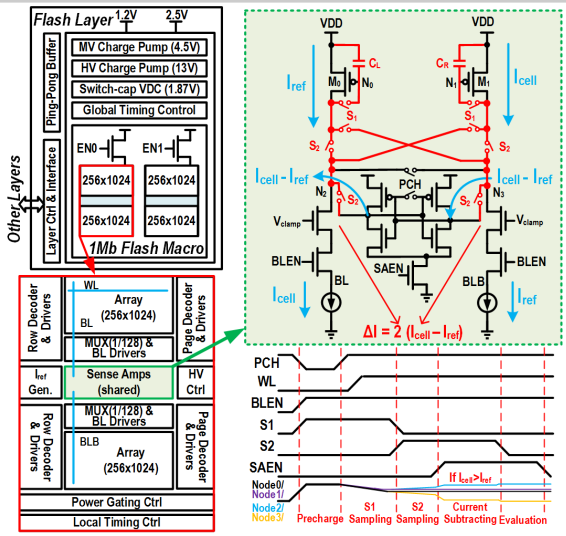


Figure 11.2.3: 1Mb Flash macro organization. Cross-sampling is used to double the sense margin of the current sense amplifier.

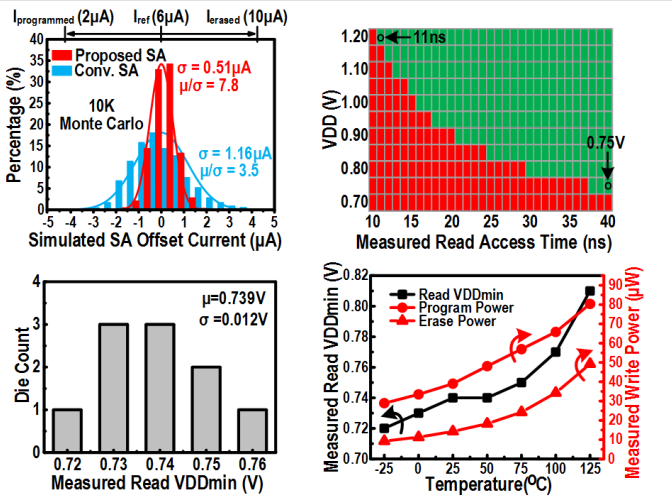
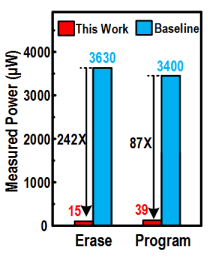
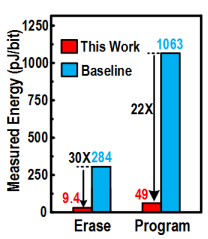


Figure 11.2.4: The proposed SA halves offset. Measured access time is 11ns at 1.2V. Average read V_{DDmin} among 10 dies is 0.739V. Read V_{DDmin} and write power across temperature shown at bottom right.





Parameters		This Work		Baseline (Compiler Flash)		ISSCC 16 [2]
Process (nm)		90		90		90
Cell Type		ESF3		ESF3		MONOS
Capacity		1Mb		2Mb		1Mb
Macro Area (mm²)		0.73		0.91		2.26
Power Supply Range (V)		2.2~2.8 / 0.75~1.4		1.6~3.6 / 0.95~1.4		3 ~ 3.6
Operation Unit		Erase:8kb Program:8b Read:8b		Erase:64kb Program:32b Read:32b		Erase:16kb Program:1kb Read:32b
Erase	Cycle (ms)	5		5		5
	Power (µW)	15(25°C)	49(125°C)	3630		NA
	Energy (pJ/bit)	9.4(25°C)	31(125°C)	284		NA
Program	Cycle (µs)	10		10		3000
	Power (µW)	39(25°C)	82(125°C)	3400		323(175°C)
	Energy (pJ/bit)	49(25°C)	92(125°C)	1063		946(175°C)
Read	Cycle (ns)	11(best)	1000	23(best)	1000	20
	Power (µW)	1618	25	5364	157	NA
	Energy (pJ/bit)	2.2	3.1	3.9	19.6	NA
	VCCmin(V)	0.75		0.95		NA
Standby Power (µW)		5.4		24.3		NA
Sleep Power (µW)		0.003		1.53		NA
Endurance		>10k		>10k		>100M
Retention		>10 years		>10 years		NA

Figure 11.2.5: Measurement results comparison with baseline design and other work.

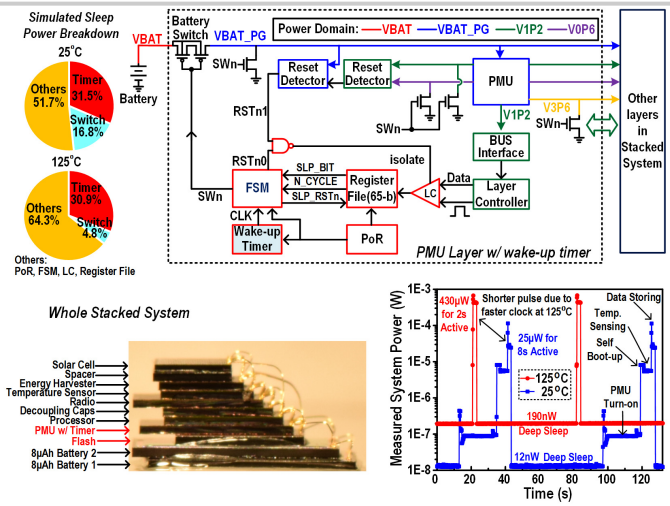


Figure 11.2.6: Diagram of PMU layer with wake-up timer and measured stacked system power. Deep sleep power at 25°C and 125°C are 12nW and 190nW, representing a 3.3× and 63× power reduction.

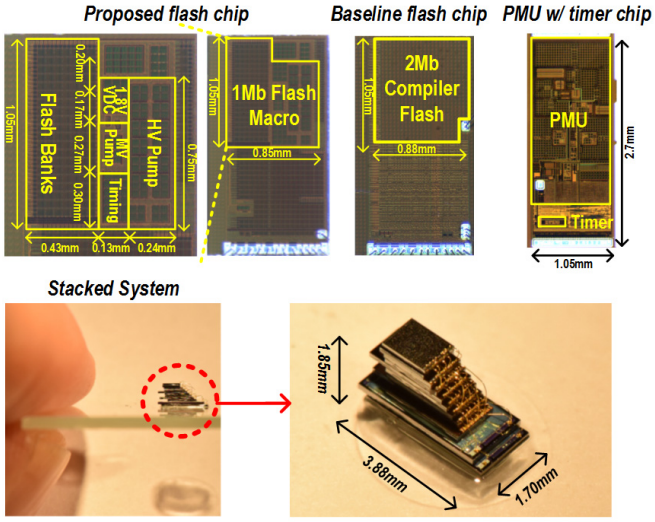


Figure 11.2.7: Photos of flash chips (90nm eFlash technology), PMU w/ timer (180nm technology) and the whole stacked sensor system.

978-1-5090-3758-2/17/$31.00