An ultra-wide program, 122pJ/bit flash memory using charge recycling

Supreet Jeloka, Jeongsup Lee, Ziyun Li, Jinal Shah, Qing Dong, Kaiyuan Yang, Dennis Sylvester, and David Blaauw
University of Michigan, Ann Arbor, MI

Abstract

Embedded flash for low power sensing systems require very low write energy and peak power. This work proposes a 130nm, 1024x260 SONOS flash with an ultra-wide 1Kb program cycle, using efficient FN tunneling based programing and a dedicated, multi-output transition pump with charge sharing and charge recycling. Combined with energy efficient charge pumps, the proposed flash program energy is 122pJ/bit with a 1Mbps throughput.

Introduction

Low power sensing systems have limited energy budgets and typically operate with low activity rates. For such systems, the ability to completely power down and save energy is vital; hence they typically save data in a non-volatile memory such as embedded flash. However, conventional flash technologies consume significant energy during programming [1]. Unlike floating-gate flash, which uses a high current hot carrier injection (HCI) based programming [2], SONOS (Silicon Oxide Nitride Oxide Silicon) flash cells perform both erase and program using Fowler Nordheim (FN) tunneling which requires much smaller current. This give SONOS the potential of low power [3], however FN tunneling is ~10^3× slower and integration of standby power over this long programming cycle typically results in program energy that is similar to flash at several nJ/bit [4].

This work re-architects a SONOS flash array to achieve very low programming power by using: 1) an ultra-wide 1Kb program cycle, enabled by the low FN tunneling current, resulting in higher efficiency by amortizing charge pump overhead over a large number of programmed bits; 2) a dedicated multi-output "transition pump" to support high current draw when transitioning the 1K bitlines and sourcelines; 3) charge sharing and charge recycling in the multi-output transition pump, improves energy efficiency for transitions by 40%; 4) dedicated "DC" charge pumps that efficiently support the low tunneling currents during the long programming phase. Combined, these approaches reduce the SONOS program energy to 122pJ/bit, which is ~10× lower than conventional floating-gate flash and SONOS NV memory, and maintains programming peak power < 300μW. Through wide programming the memory maintains 1Mbps throughput, which is comparable to hot carrier injection based programming, allowing for logging large data sets (e.g., audio) during short wake-up periods.

Fig. 1 shows the required voltages for program and erase of the 130nm SONOS 2T flash cell. Programming requires both the bitline (BL) and sourceline (SL) to be -3.8V for selected cells and 1V for half-selected cells. Once the program voltages are applied, the tunneling takes ~1ms, during which the current is very small. However, as shown in Fig. 1, there is a large transition current at the beginning and end of a program cycle due to the charging/discharging of BLs and SLs. This transition current places a limitation on the number of cells that can be programmed simultaneously.

Proposed SONOS flash

To efficiently implement the ultra-wide program operation, we propose a high current transition charge pump shown in Fig. 2 to support the large transient current when entering program mode. The BL logic selectively connects each bitline to one of four voltage rails: stable -3.8 V, stable 1V, 'Rising Rail', or 'Falling Rail', depending on the current and previous program data values. Data input lines that have matching current and previous values remain static. Bitlines that need to transition from -3.8V to 1V are switched to the 'Rising Rail' which is increased in voltage in several steps by the transition pump after which they are switched to the stable 1V rail. Falling bitlines are similarly transitioned from 1V to -3.8V. The write state machine monitors the comparator output from the charge pump loops to determine voltage stability, for safe entry to/from the transition phase and in between the transition steps.

By not transitioning bitlines unnecessarily, the proposed approach improves peak power and energy efficiency by ~50%. In addition, by changing the transition rail voltage in 4 smaller steps instead of one large step, conduction loss in the charge pump is reduced and, step 3 (Fig 2, bottom) is accomplished using charge sharing by shorting the rising and falling transition rails, consuming no energy. Finally, by charging and discharging the transition rails simultaneously out of the same pump (step 4), charge is recycled between the opposite transitioning rails, instead of sourcing or sinking the entire charge from Vdd or Gnd, further improving energy efficiency by ~40%. In total, the charge sharing and charge recycling transitions are 3.3× more energy efficient than a single step transition of BL and SL. Furthermore, stepped transitions using a dedicated transition pump reduces peak power by 14.7× which enables the concurrent programming of 1Kb without collapsing the charge pump voltages, and further reduces programming energy per bit by amortizing fixed programming energy overhead over a much larger number of bits.

After the transition phase, the transition pumps are disconnected and the low programming DC currents are supplied by high efficiency closed-loop charge pumps. Fig. 3 shows the negative charge pump design for the proposed flash array. The negative charge pump consists of three Dickson stages with its clock controlled by an output voltage monitoring loop. To generate a positive intermediate voltage for comparison from the negative charge pump's output, a diode stack divider is placed between a 1.2V reference voltage VREF [5] and the negative pump output. In Fig. 3, as VNEG falls, the voltage divider output falls, allowing the use of a clocked comparator with a positive reference voltage to gate the clock and control VNEG.

As shown in Fig. 4, the clock generator and the on-chip voltage reference generator are shared between all the voltage generation blocks. The clock frequency is trimmed on-chip to reduce charge pump loop power. The flash array performs 64b reads and hence a 16:1 column mux is used before the current sense amplifiers (CSAs). The reference current for the CSA is generated using one erased and one programmed bitcell in the reference column of the selected row. This allows for tracking the soft program and soft erase disturbs, and improves the read margin by 23% in simulation with aged model. Fig. 5 shows the simulated efficiency over load current for the positive & negative pump and the 1V DC-DC converter.

Measured Results

The low power ultra-wide write SONOS flash design with charge recycling was fabricated in 130nm technology. An on-chip BIST tests flash program, erase, and read functionality. In addition, the current from erased and programmed cells are observed for the reference cells, to confirm the effectiveness of program and erase operations. Fig. 6 shows the measured read results; maximum read frequency is 40MHz at nominal Vdd of 1.8V, with read energy of 2.23 pJ/bit. The read operates down to 1.2V at 7MHz, at which point read energy is reduced to 1.2 pJ/bit. In Fig. 7, program energy variation across 0° to 85°C is measured, showing 20% min-max variation. Fig. 8 shows measured distribution of program energy for 21 chips with a mean of 122 pJ/bit. Table 1 compares proposed flash to other tunneling and HCI based NOR flash memories. The measured programming energy for the proposed 1024x260 bit flash is 122 pJ/bit with an average power of 125μW. Block erase consumes 29pJ/bit. By enabling wide-program FN programming, the program throughput is similar to HCI flash, with >10× better program energy efficiency.

References

[1] M.F. Chang et al., JSSC, 2009
[2] R. Sundaram et al., ISSCC, 2005
[3] H. Mitani et al., ISSCC, 2016
[4] K. Ramkumar et al., IMW, 2013
[5] M. Seok et al., ISSCC, 2012
[6] Y. Taito et al., ISSCC, 2015

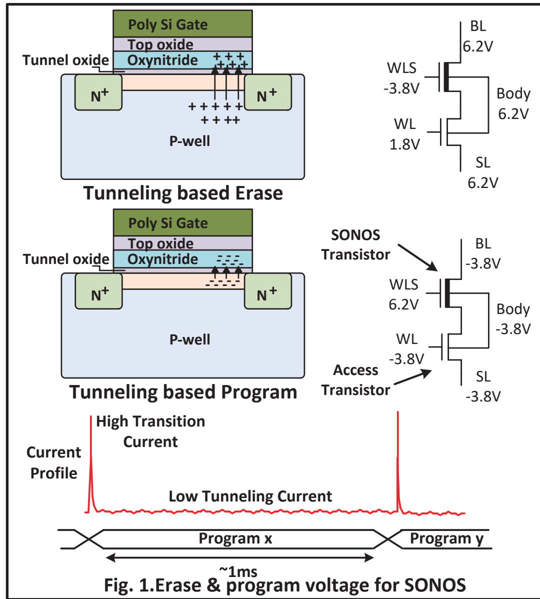

Fig. 1.Erase & program voltage for SONOS

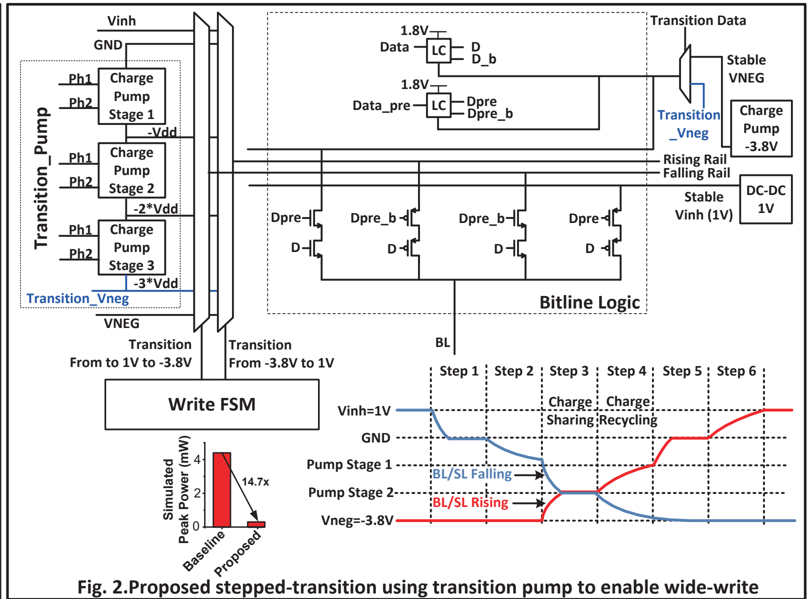

Fig. 2.Proposed stepped-transition using transition pump to enable wide-write

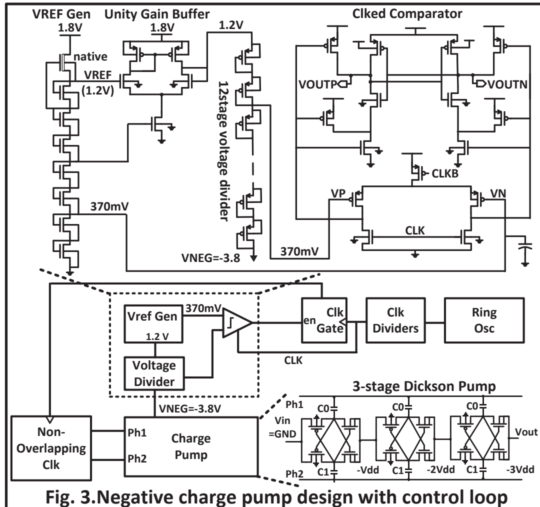

Fig. 3.Negative charge pump design with control loop

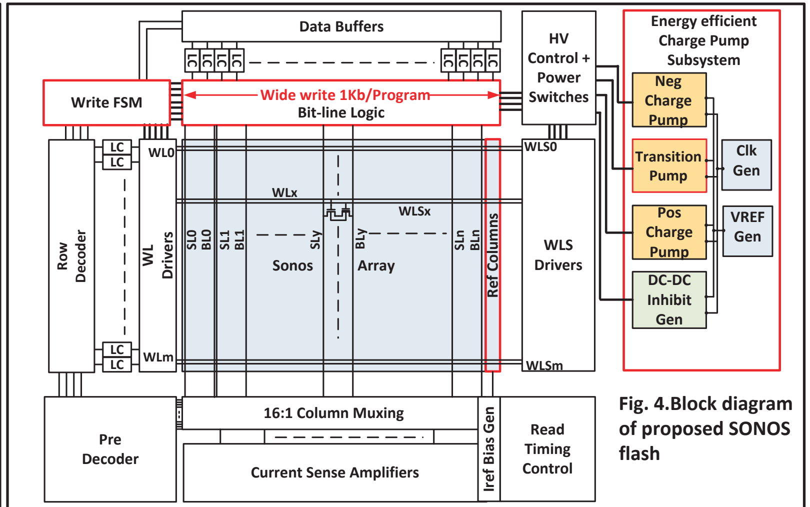

Fig. 4.Block diagram of proposed SONOS flash

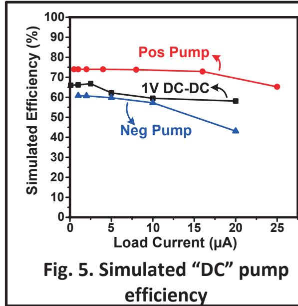

Fig. 5. Simulated "DC" pump efficiency

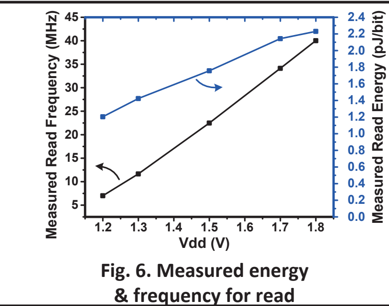

Fig. 6. Measured energy & frequency for read

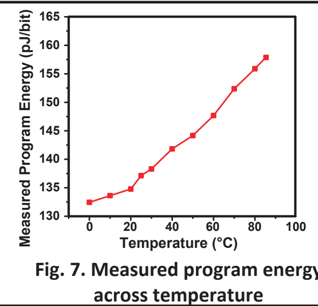

Fig. 7. Measured program energy across temperature

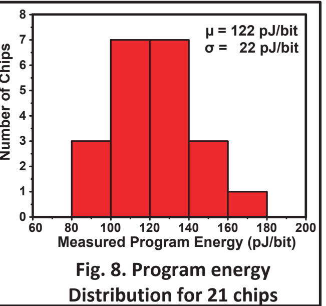

Fig. 8. Program energy Distribution for 21 chips

	This Work	ISSCC'16 [3]	ISSCC'15 [6]	ISSCC'05 [2]
Technology	130nm 2T SONOS	90nm 1T MONOS	28nm SG-MONOS	180nm
Memory Size	1024×260	128kB	Code =2MB Data = 64kB	16MB
Program Technique	FN Tunneling	FN Tunneling	HCI	HCI
#Bit per program	1024	1024	32	32
Program Throughput	1Mbps	341Kbps	2Mbps	3Mbps
Program Energy	122pJ/bit	1.07nJ/bit	n.a.	12.6nJ/bit
Erase Energy	29pJ/bit	1.07nJ/bit	n.a.	n.a.
Read Energy	2.23 pJ/bit (1.8V) 1.20 pJ/bit (1.2V)	n.a.	n.a.	n.a.
Read Frequency	40MHz (1.8V) 7MHz (1.2V)	52MHz	Code = 200MHz Data = 10MHz	n.a.

Table 1.Comparison with previous works

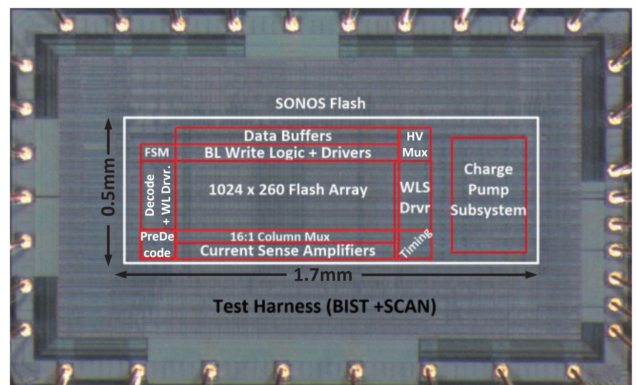

Fig. 9. Proposed SONOS flash die photo