26.3 A 65nm 1Gb 2b/Cell NOR Flash with 2.25MB/s Program Throughput and 400MB/s DDR interface

Corrado Villa, Daniele Vimercati, Stefan Schippers, Salvatore Polizzi, Andrea Scavuzzo, Maurizio Perroni, Maurizio Gaibotti, Mauro Luigi Sali

STMicroelectronics, Palermo, Italy

NOR flash architectures continue to improve to cope with wireless application requirements of faster XIP (eXecute In Place) performance as well as faster programming throughput. A 1.8V 65nm 2b/cell 1Gb NOR flash memory [1,2] based on time-domain voltage-ramp reading concept, flexible read-while-write (RWW is presented. This paper describes the program method, architecture and algorithm able to reach 2.25MB/s with internal charge pumps, as well as the read concept and RWW [3,4] management, allowing 70ns random access time and an experimental DDR support for 400MB/s read throughput.

Figure 26.3.1 shows the die floor-plan. The 1Gb array is split into 8 logical banks of 64 erasable blocks each. A block contains 2Mb organized in 8K pages of 16 words, 16 bits each. A bank is the logically independent unit to be considered in RWW operation. Each bank contains 128 2b/cell sense amplifiers, producing 256 bits and able to read out simultaneously an entire page of 16 words in a typical 70ns access time. To manage RWW operation, logical and analog multiplexers are placed in each bank to drive local row and column decoders. To handle the voltages required and minimize area, all multiplexers are made with high voltage MOS.

To speed-up programming performance a pipelined program/verify logic and algorithm are introduced. The program control logic, shown in Fig. 26.3.2, includes program verify logic, a double 512-word RAM buffer, master-slave program load and sense amplifiers. Programming operation starts by loading the data to be programmed into the user buffer RAM (UBRAM). Once the loading is complete, the buffer content is moved into the operation RAM (ORAM) to start the programming. During the verify phase, the ORAM contents are updated by clearing the bits that have been successfully programmed. These steps are repeated until the full contenta are programmed and all bits in the ORAM are cleared.

To improve write performance, the bitline selection is synchronized with the program loads clock during the programming phase, as shown in Fig. 26.3.3). The master content is transferred into the slave starting the program pulse. During the program pulse the ORAM content is transferred to the master latch preparing the new value for the next bit line to be pulsed. Following the address increment, the column is switched to the next bitline and the related data are loaded into the slave. This guarantees that there is virtually no time loss between program pulses.

To optimize the programming efficiency, for every step of the wordline staircase voltage the bit-line pulse is applied (if required) in sequence to all the cells corresponding to the entire ORAM buffer, followed by the verification of the whole buffer, before moving to the next wordline staircase step. The internal programming parallelism is 64 cells. To complete the operation, multiple programming cycles are performed. Figure 26.3.4a shows how each programming level is built by splitting the programming operation in two phases: a preliminary phase presets the cells to a coarse level, using a first reference cell set, then a fine phase adjusts the cells to the final level using a second reference cell set. This algorithm [5] is minimizes the crosstalk associated with floating-gate coupling in 65nm process.

Using the architecture and techniques reported above, the 1Gb flash NOR is able to program a 1KB write buffer in 435µs, equivalent to 2.25MB/s sustained programming throughput (Fig. 26.3.4b).

The reading is based on the voltage ramp time domain concept [6]. Three clock edges (REF1, REF2 and REF3) are generated when the 3 reference cells, whose gates are driven by the fast linear ramp, reach a fixed current level. The switching of the array sense amplifier clocks the latching of the REF signals, which represent one of the possible four states of a sensed cell.

A source of noise, potentially disturbing the RWW operation, is the switching of a bank of sense amplifiers. This supply noise can disturb another bank already in read operation. To avoid such event, during RWW operation, a synchronization circuit prevents switching on or off a sense bank in verify status during read evaluation of another sense bank. This mechanism prevents the start of a verify operation during the most critical phase of the read (protected read). Figure 26.3.5 shows how the logic circuit delays verify operation (switching on the sense bank) until the critical read phase is completed with a negligible effect on the verify cycle (<30ns). The duration of the protected read phase is small enough to assure that the delayed verify is not affecting the next read cycle.

Simulation verifies the 16b DDR interface read architecture operates up to 400MB/s read throughput. Figure 26.3.6 shows the programming throughput and the read throughput demonstrated on silicon of 512Mb, same technology, specification and architecture, which was the first product of the same family to be produced. Figure 26.3.7 summarizes the chip features.

Acknowledgements:
We gratefully acknowledge the work of the whole Design team and CAD support, of the Engineering team, and of the R&D Flash Technology Development. Thanks to G. Casagrande, M. Dallabora and G. Crisenza for support and helpful comments.

References:

[1] M. Bauer, R. Alexis, G. Atwood, et al., “A Multilevel-Cell, 32Mb Flash Memory,” *ISSCC Dig. Tech. Papers*, pp. 132-135, Feb., 1995.

[2] G. Campardo, R. Micheloni, S. Commodaro, et al. “40-mm^2 3-V-Only 50-MHz 64-Mbit 2-b/cell CHE NOR Flash Memory,” *IEEE J. Solid-State Circuits*, vol. 35, no. 11, pp. 1655-1667, Nov., 2000.

[3] H.A. Castro, K. Augustine, S. Balasubrahmanyam, et al., “A 125MHz Burst Mode 0.18µm 128Mbit 2 Bits per Cell Flash Memory,” pp. 304-307, *Symp. VLSI Circuits*, June., 2002.

[4] D. Elmhurst, R. Bains, T. Bressi, et al., “A 1.8V 128Mb 125MHz Multi-Level Cell Flash Memory with Flexible Read While Write,” *ISSCC Dig. Tech. Papers*, pp. 286-287, Feb., 2003.

[5] M. Taub, R. Bains, G. Barkley, et al., “A 90nm 512Mb 166MHz Multilevel Cell Flash Memory with 1.5MByte/s Programming,” *ISSCC Dig. Tech. Papers*, pp. 54-55, Feb., 2005.

[6] C. Villa, D. Vimercati, S. Schippers, et al., “A 125MHz Burst-Mode Flexible Read-While-Write 256Mbit 2b/c 1.8V NOR Flash Memory,” *ISSCC Dig. Tech. Papers*, pp. 52-53, Feb., 2005.

1-4244-0852-0/07/$25.00 ©2007 IEEE.

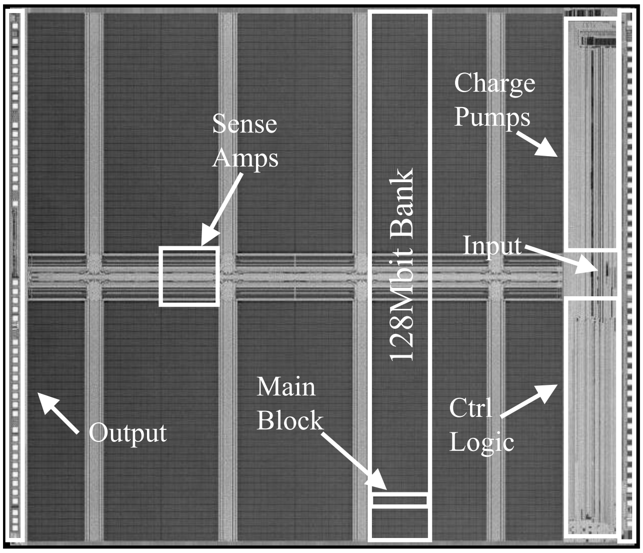


Figure 26.3.1: 1Gb die floor plan.

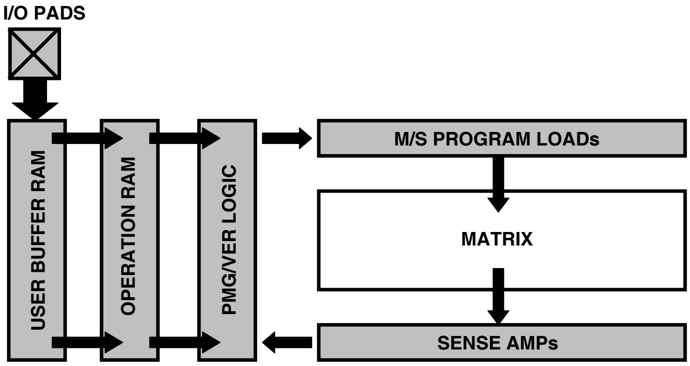


Figure 26.3.2: Program block diagram.

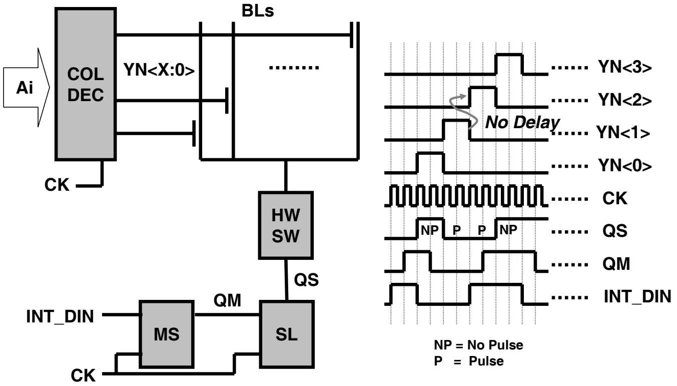


Figure 26.3.3: Synchronized load and BL switching.

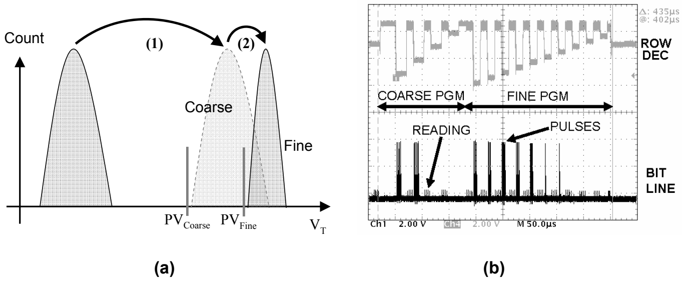


Figure 26.3.4: Program phases and results.

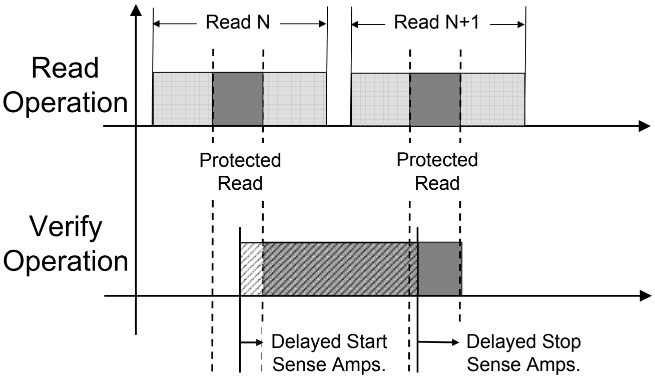


Figure 26.3.5: Synchronization scheme.

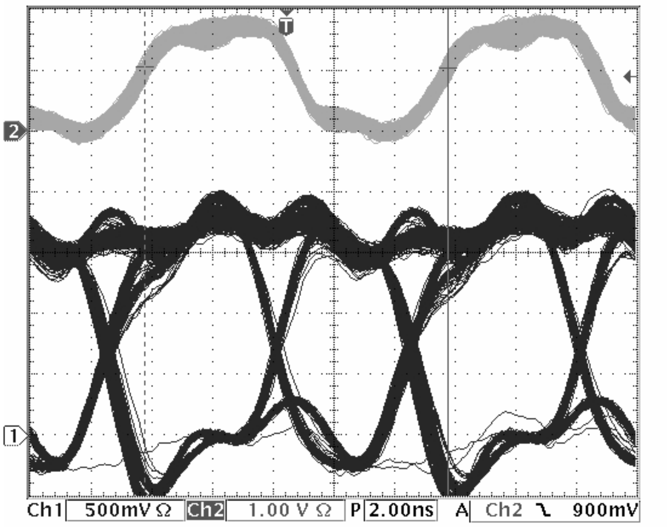
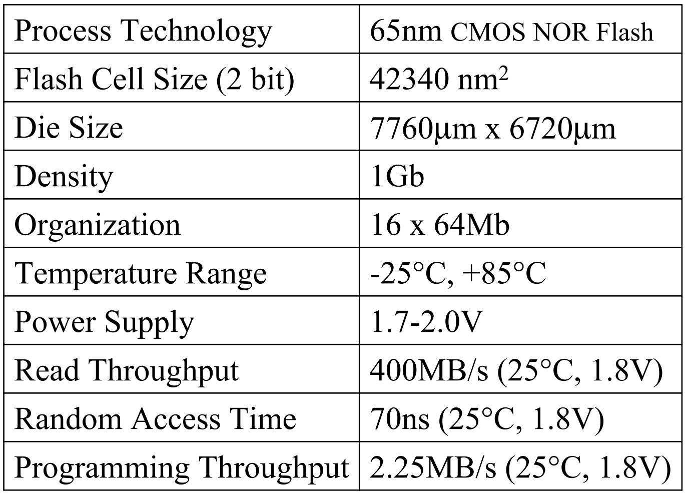


Figure 26.3.6: 400MB/s read throughput.

Continued on Page 616



Process Technology	65nm CMOS NOR Flash
Flash Cell Size (2 bit)	42340 nm^2
Die Size	7760µm x 6720µm
Density	1Gb
Organization	16 x 64Mb
Temperature Range	-25°C, +85°C
Power Supply	1.7-2.0V
Read Throughput	400MB/s (25°C, 1.8V)
Random Access Time	70ns (25°C, 1.8V)
Programming Throughput	2.25MB/s (25°C, 1.8V)

Figure 26.3.7: Key features table.