C-10-2

Low-Voltage Embedded Flash-EEPROM in 0.18 μm CMOS

Do Dormans, Dick Boter, Antonio Cacciato, Margriet Diekema, Chantal Dijkstra, Marton Hendriks, Roelof van der Linde, Guoqiao Tao, Hein Valk, Erik van der Vegt and Rob Verhaar
Philips Semiconductors MOS4YOU, Gerstweg 2, 6534 AE Nijmegen, The Netherlands
Tel: +31-24-353-3376, Fax: +31-24-353-5200, Do.Dormans@philips.com

1. Introduction

Adding more functionality in a single device to realise a true system-on-chip puts heavy demands on the process to reach this in a manufacturable and cost-effective way.

Particularly, embedding non-volatile memory for on-chip storage of persistent data, code and content is very challenging because besides highly performing logic blocks and memory, high-voltage (HV) circuitry as well as analogue elements for IO's, sense-amplifiers, charge pumps and stable voltage control are required.

Recently, a 2-transistor (2T) flash memory cell has been described that consists of a stacked gate transistor with an access transistor at the source side [1]. The access transistor allows erasing the memory cells into depletion thus avoiding the over-erasure problems encountered in 1T flash cells. Another advantage is that cells can be read at low voltages thus avoiding the need to boost the control gate (CG) during read [2]. This makes this type of memory explicitly suitable for single-supply low-voltage applications [3]. A 2T flash cell can be programmed and erased by Fowler-Nordheim (FN) tunnelling to and from the channel of the stacked gate transistor. This is not only beneficial for the reliability of the memory but also reduces power consumption and reduces test time, as massive parallel programming is possible.

2. Memory cell and process flow

The layout of the 2T flash cell is presented in Fig. 1. One cell consists of two stacked gates both processed on a tunnel oxide. The bottom poly of the stacked gate at the source (So) act as access gates (AG). Bit-line (BL) contacts and source lines both made in a Local Interconnect Layer of tungsten address cells. A cell area of 0.78 μm^2 is realised using logic design rules (Table I). The memory is formed in an isolated pWell allowing positive and negative voltages for program and erase. Table II gives the operation table for the memory.

The process flow (Table III) is optimised to maintain logic performance. After field isolation and all wells are formed, the logic oxides are grown and covered by poly so that logic wells experience subsequent processing steps only as anneals. By doing so unwanted dipping on STI edges in the logic is also avoided. Succeeding the formation of flash and HV devices the logic process flow is continued. Only the pre-metal dielectric layer that isolates the front-end from the back-end and planarises the wafers prior to the 1st metal deposition was adapted to compensate for the increased topography in the memory array. An overview of all available transistors in this process is given in Table IV.

The compatibility of the flash processing with the baseline process is exemplified in Fig. 2 showing a comparison of delay times of ring-oscillators measured on wafers with and without the flash processing.

3. Flash performance

Fig. 3 shows the program and erase characteristics as function of the total program voltage. With 15V and P/E times of 5 ms (per page) and 100 ms (per sector) respectively, a program window of 4.5V is obtained. An inhibit voltage of 5V is used to prevent cells on un-selected bit-lines from programming. Threshold voltage distributions are typically 100mV/σ without any verification (Fig. 4). Some fast cells are present but they are of no concern in a 2T flash memory. Read currents (Fig. 5) are in the order of 20 μA (Vdd=1.2V) or 30 μA (Vdd=1.8V) at low BL (0.5V) and CG (1.2V) bias. These low voltages allow fast access times and suppress read disturb. Endurance characteristics are shown in Fig. 6. Threshold voltages of both programmed and erased cells increase due to charge trapping in the tunnel oxide and at the oxide interface. Window closure due to electron trapping in the tunnel oxide is relatively small (0.6V after 100k P/E cycles) as expected for FN tunnelling through the full channel of the stacked memory transistor.

A 16Mb flash memory was designed in this process to demonstrate the feasibility of the 2T-flash approach [4]. A photograph of a processed device is shown in Fig. 7. The memory can be read at 1.2V and programmed at 1.5V.

4. Conclusions

A 2T-flash memory suitable for low-voltage (1.2V) and low-power applications is presented. A modular process flow has been developed to embed the memory, high-voltage circuitry and analogue elements into a 5-metal layer 0.18 μm CMOS logic process. Results show the compatibility of this flow with the baseline process and the performance of the 2T flash memory. The feasibility of this memory concept and process flow is demonstrated by the successful manufacturing of an embedded 16Mb flash memory circuit.

References

[1] W.-H. Liu et al., NVSMW 1997, p.4.1
[2] K. Katahashi et al., VLSI 1999, p.21
[3] G. Dormans, et al., NVSMW 2000, p111
[4] T. Ditewig, et al., ISSCC 2001, p. 2.4

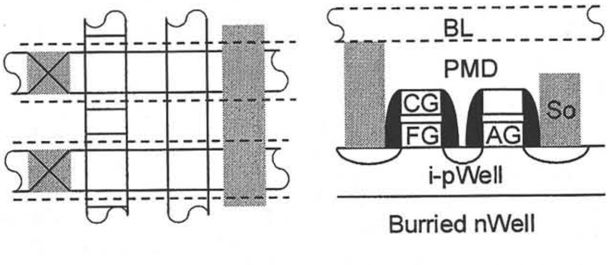

Fig. 1 Schematic layout and cross-section of the 2T-flash cell.

Table I. Flash cell characteristics

Stacked gate transistor	W/L = 0.24 µm / 0.24 µm
Access transistor	W/L = 0.24 µm / 0.24 µm
Cell size	0.78 µm^2
Tunnel oxide	8.5 nm
Inter-poly dielectric	ONO 6/6/6 nm

Table II. Memory operation scheme

	BL	AG	CG	So	pWell
Read	0.5/ 0	Vdd / 0	+1.2	0	0
Program	-5 / 0	-5	+10 / 0	Float	-5
Erase	Float	+10	-5 / +10	Float	+10

Table III. Schematic process flow

- Field isolation (STI)
- Well formation (logic, triple, high-voltage)
- Logic oxides (3.2 & 7.5 nm) and gate deposition
- Tunnel oxidation and floating gate formation
- ONO formation
- High-voltage oxidation and control-gate formation
- Gate patterning and S/D implantations
- Pre-metal-dielectric & local interconnect formation
- Standard 5-metal back-end

Table IV. Logic, analogue and HV device characteristics

Device	Oxide (nm)	Vt (mV)	Isat (µA/µm)
High-performance nMOS	3.2	470	600
High-performance pMOS	3.2	480	290
Low-leakage nMOS	3.2	590	500
Low-leakage pMOS	3.2	570	250
"Analogue" nMOS	7.5	700	530
"Analogue" pMOS	7.5	740	250
High-voltage nMOS	20	600	300
High-Voltage pMOS	20	600	180

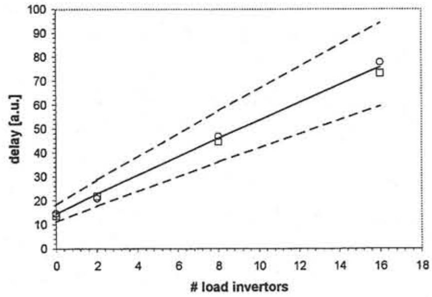

Fig. 2 Delay time of ring-oscillators (201 inverters) measured on wafers with (circles) or without (squares) flash processing and simulated (fast, nominal, slow) using baseline parameters.

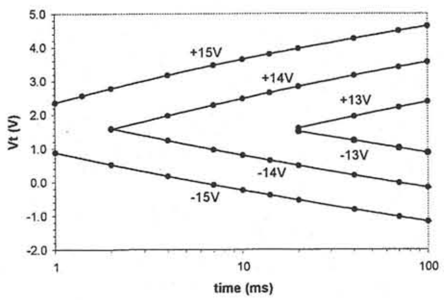

Fig. 3 P/E curves as function of total P/E voltage.

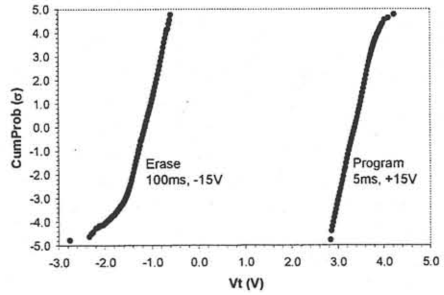

Fig. 4 Cumulative threshold voltage distributions.

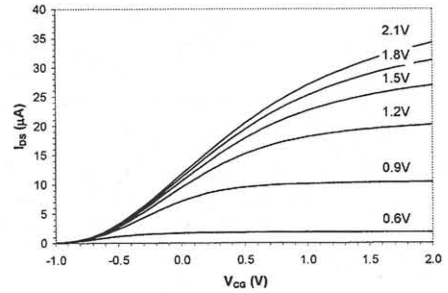

Fig. 5 Read current (@Vds=0.5V) as function of AG voltage.

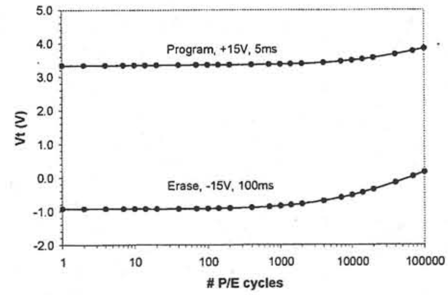

Fig. 6 Endurance curve.

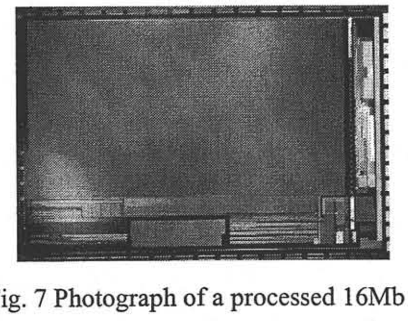

Fig. 7 Photograph of a processed 16Mb flash memory.