Very Fast Recovery Word-line Voltage Regulator for Multilevel Nonvolatile Memories

OSAMA KHOURI*, RINO MICHELONI°, GUIDO TORELLI*

*Department of Electronics
University of Pavia
Via Ferrata 1
27100 Pavia
ITALY

°Memory Product Group
STMicroelectronics
Via C. Olivetti 2
20041 Agrate Brianza (MI)
ITALY

Abstract: - A very fast recovery voltage regulator for large capacitive loads is presented. A feedback cascode technique is used to achieve high recovery speed when a previously discharged capacitance is connected to the regulator output. The design has been optimized for a word-line read voltage regulator for a 4-level (2-bit/cell) 32-Mcell Flash memory (64 Mbits). Computer simulations showed that recovery within a 50-mV ripple is achieved within 20 ns after a word-line is selected, as is required to allow accurate sensing without degrading memory access time. Current consumption is kept within 65 μA.

Keywords: Multilevel nonvolatile memory, Word-line voltage regulator, Capacitive–load voltage regulator.

1. Introduction

Multilevel (ML) nonvolatile (NV) memories [1,2] allow significant increase in storage density for any given fabrication technology. Indeed, in these memories a controlled charge injection into the floating gate allows the cell threshold voltage to be programmed to any of $m = 2^n$ predetermined levels, which leads to the storage of n bits in a single cell.
Sensing ML NV memories is a very challenging task, as the spacing between adjacent programmable threshold levels is reduced with respect to the case of conventional bilevel memories. In particular, the voltage V_{GR} applied to the gate of selected cells must be adequately accurate and stable, so as to allow the programmed cell threshold level to be safely discriminated.
The optimal value of V_{GR} turns out to be a trade-off choice between reliability and design considerations [3]. On the one hand, V_{GR} can not be set too high, to minimize read disturbs (i.e., degradation of the stored charge due to the voltages applied to cell terminals during sensing), while on the other it can not be set too low, to allow a sufficient threshold window (i.e., a sufficient width of the voltage range where all programmed levels have to be allocated). An appropriate value of V_{GR} is in the range of 6 V. This voltage must be generated within the memory chip starting from the supply voltage (V_{DD} = 3 V) by using on-chip voltage multipliers, which are generally based on the charge-pump technique [4].

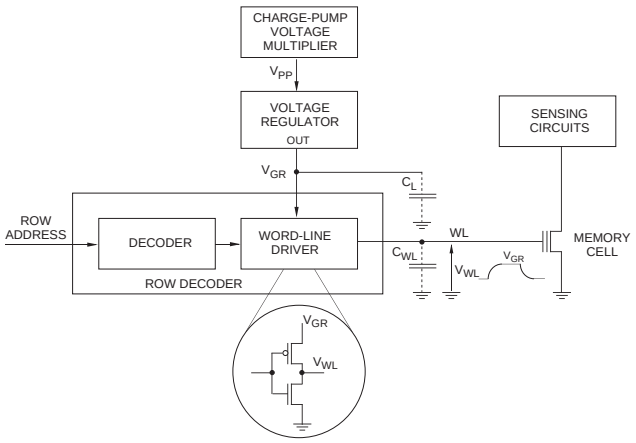

Fig. 1 Block diagram of the reading section of a nonvolatile memory.

The voltage multiplier feeds a voltage regulator, which in turn provides the supply voltage V_{PP} to the final driving stages of the row decoder (Fig. 1). In a multimegabit NV memory, a very large capacitive load C_L is associated to the regulator output due to the parasitic capacitance of all the supplied driving stages.
During sensing, the addressed word-line WL is connected to the regulator output through the pull-up device of the corresponding driving stage, while all other word-lines are grounded. When a word-line is selected in response to a read request, its parasitic capacitance C_{WL} (which was previously discharged) is connected in parallel to C_L (Fig. 2).

This causes a drop ΔV in the regulator output voltage equal to

$$\Delta V = \left(\frac{C_{WL}}{C_L + C_{WL}} \right) V_{GR} \cong \frac{C_{WL}}{C_L} V_{GR} \qquad (1)$$

In order not to degrade memory access time, a key feature required of the regulator is very fast output recovery: when a new word-line is selected, the desired value of V_{GR} has to be provided within a specified ripple in a very short time.

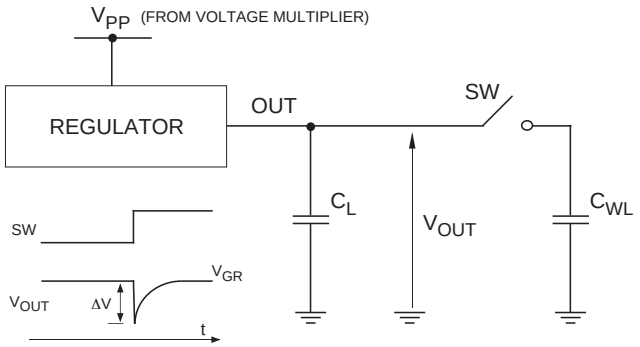

Fig.2 Voltage drop due to charge sharing at the regulator output.

Moreover, it should be pointed out that a charge-pump based voltage multiplier shows a rather large output ripple at the charge-pump operation frequency. Adequate power supply rejection ratio (PSRR) must therefore be provided at this frequency so as to minimize fluctuations in the regulated voltage and, hence, in the read voltage of the selected word-line. In addition, at device power-up, a significant time is needed by the voltage multiplier to provide a high output voltage [5]. The regulator must therefore ensure limited voltage drop-out to guarantee correct circuit operation (including reading) within the shortest time. Finally, as charge-pump voltage multipliers have limited output current drive capability, the regulator must consume a reduced amount of power. This also helps for fast set-up of the voltage multiplier.

This paper presents a voltage regulator designed to meet the above requirements. More specifically, the regulator was optimized for a 2-bit/cell 32-Mcell Flash memory (64-Mbits) in 0.2-μm CMOS fabrication process (minimum channel length = 0.6 μm in the periphery). In our design, V_{GR} = 6 V, C_L = 230 pF and C_{WL}= 6.5 pF and, therefore, the output voltage drop ΔV when a new word-line is connected to the regulator output, is ~170 mV. The target specification for the regulator was a recovery time less than 25 ns within a ripple of 50 mV with a current consumption less than 70 μA.

2 – Circuit description

Fig. 3 illustrates the basic topology of the designed low drop-out voltage regulator [6]. This is made up by an error amplifier (differential amplifier A), a pass device (p-channel transistor MP), a feedback network (resistors R_1 and R_2), and a reference voltage V_{BG}. The use of a p-channel pass device ensures limited drop-out and large output current sourcing capability during transients. Block COMP provides frequency compensation, as shown in detail below.

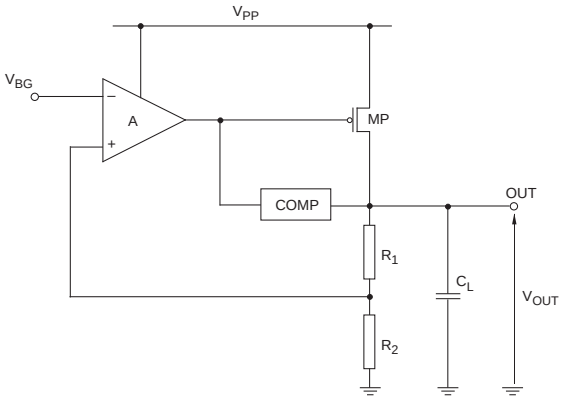

Fig. 3 Low drop-out voltage regulator.

Assuming ideal components (namely, no offset and sufficiently high loop gain), the regulator output voltage is equal to

$$V_{GR} = V_{BG}\left(1 + \frac{R_1}{R_2}\right) \qquad (2)$$

A band-gap based voltage reference V_{BG} and a pair of matched resistors R_1 and R_2 ensure an accurate value of V_{GR}.

The regulator loop includes two gain stages and, hence, adequate frequency compensation is required to prevent excessive overshoots and ringing in the output voltage during transients. Conventional Miller compensation using a simple capacitor C_C can not be used to implement the compensation block COMP because:

i) at high frequencies, the feedforward path through C_C effectively shorts the drain of MP to its gate, thus degrading the PSRR performance of the regulator;
ii) in the presence of a large capacitive load C_L, a large capacitor C_C is needed; this severely limits the slew rate (SR = I_0/C_C) of the differential amplifier, as its biasing current I_0 must meet the above low-power consumption requirements.

To avoid these drawbacks and achieve the desired fast recovery performance, the cascode feedback compensation technique [7] was used, as shown in the detailed circuit scheme of Fig. 4.

The input differential stage, made up by transistors M1 to M7, includes common-gate devices M6 and M7 to improve PSRR performance. A compensation capacitor C_C is connected to the source of the common-gate transistor M13 (node A). In such a way, a low-impedance path is ensured from the left plate of C_C to the gate of the pass device MP (node B), thus producing a dominant pole. By contrast, the feedforward path from node B to the output through the compensation capacitor is removed, thus preventing PSRR degradation. M9 and M11 are two equal current sources, which provide the common gate transistor M13 with the necessary bias current.

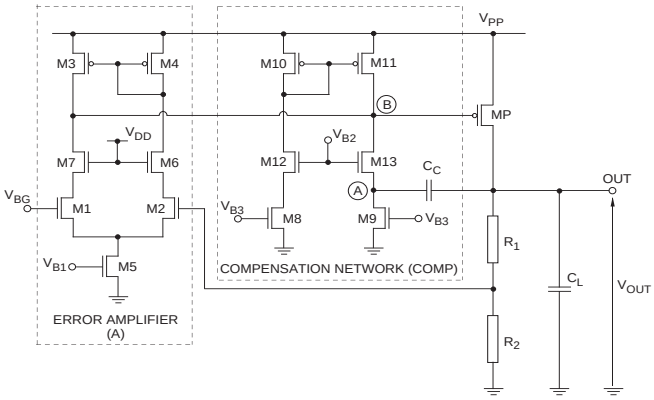

Fig. 4 Detailed circuit schematic of the voltage regulator

Routine analysis shows that the dominant pole of the open-loop structure is

$$P_1 \cong \frac{1}{(g_{mp}R_{OUT})C_C R_B} \tag{3}$$

where g_{mp} is the transconductance of MP and C_B and R_B are the effective capacitance and the small-signal resistance, respectively, at the output of the differential amplifier (node B).
The frequency of the secondary pole is equal to

$$P_2 \cong \frac{g_{mp}C_C}{C_B(C_C + C_L)} \tag{4}$$

The used compensation technique offers an improvement by a factor of C_C/C_B in high frequency shifting of the secondary pole as compared to conventional Miller compensation.
For frequency stability, the unity–gain frequency of the loop the must be placed well below the secondary pole, i.e.

$$\frac{g_{mp}C_C}{C_B(C_C + C_L)} > \frac{g_{m1}}{C_C} \tag{5}$$

and, hence,

$$\frac{g_{m1}}{g_{mp}} < \frac{C_C^{\,2}}{C_B(C_C + C_L)} \tag{6}$$

The reduced power budget limits the values of g_{m1} and g_{mp}, even though the latter must have an adequate value to provide the regulator with adequate output current driving capability. For any given values of g_{m1}/g_{mp}, C_L and C_B, the compensation capacitor C_C must be chosen so as to satisfy eq. (6), in order to ensure sufficient phase margin. The value of C_C must also allow adequate amplifier slew rate. As a more effective compensation is provided with respect to the case of conventional Miller technique, a small C_C can be used, which results in better slew rate performance for any given value of I_0.
The reduced value of g_{m1} limits the gain-bandwidth product of the feedback loop which, therefore, can not achieve the required fast output voltage recovery in response to an output voltage drop. However, in the used circuit topology, an additional internal feedback path is provided from node OUT to the gate of the pass device MP through the path (C_C, M13). Thus, a drop ΔV in the output voltage gives rise to a voltage decrease in the gate of MP, which in turn produces a fast increase in the output current and, therefore, speeds up output voltage recovery.
This additional feedback path is basically made up by a high-pass structure (C_C and impedance at node A) followed by a low-pass one (common-gate device M13 and impedance at node B). The transfer functions of the two structures ($F_H(s)$ and $F_L(s)$, respectively) are easily calculated as

$$F_H(s) = \frac{V_A(s)}{V_{OUT}(s)} = \left(\frac{1}{1 + (g_{m13} + sC_A)/sC_C}\right) =$$

$$\cong \frac{sC_C/g_{m13}}{1 + sC_C/g_{m13}} \tag{7}$$

$$F_L(s) = \frac{V_B(s)}{V_A(s)} = \left(\frac{g_{m13}R_B}{1 + sR_BC_B}\right) \tag{8}$$

where g_{m13} is the transconductance of M13 and C_A is the parasitic capacitance at node A, assumed much smaller than C_C. The cascade of $F_H(s)$ and $F_L(s)$ gives obviously rise to an overall band-pass transfer function $F_T(s) = V_B(s)/V_{OUT}(s)$:

$$F_T(s) = \left(\frac{sC_C/g_{m13}}{1 + sC_C/g_{m13}}\right)\left(\frac{g_{m13}R_B}{1 + sR_BC_B}\right) \tag{9}$$

In practice, $R_BC_B > C_C/g_{m13}$ and, therefore, the pole of the low pass structure is located below the pole of the high-pass structure The magnitudes of transfer functions $F_H(s)$, $F_L(s)$ and $F_T(s)$ are plotted in Fig. 5.

The pass-band gain turns out to be equal to C_C/C_B, and the upper cut-off frequency is g_{m13}/C_C. To achieve the required fast output voltage recovery, we must set $C_C >> C_B$ and choose a sufficiently large value of g_{m13}.

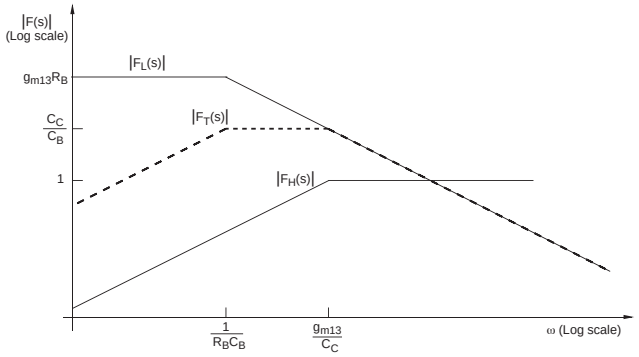

Fig. 5 Bode plots of $F_H(s)$, $F_L(s)$ and $F_T(s)$.

3. Simulation results

The described regulator was designed for the above 64-Mbit Flash memory. Optimization was carried out with ELDO (V_{DD} = 3 V ± 10). Resistors R_1 and R_2 were set equal to 159 kOhm and 41 kOhm, respectively. This gives V_{GR} = 6.015 V when using a bandgap reference V_{BG} equal to 1.233 V. C_C and g_{m13} were set to 5 pF and 470 μA/V, respectively.

According to simulations, the regulator loop has a a unity-gain frequency of 1.2 MHz with a phase margin of 75°.

The simulated transient in Fig. 6 illustrates that the regulator achieves output voltage recovery within the specified ripple of 50 mV in 20 ns. Target specifications were also completely met in worst-case simulations. The PSRR of the regulator at the charge-pump operation frequency (20 MHz) is −53 dB, which is adequate for our application. The total DC current consumption is 65 μA.

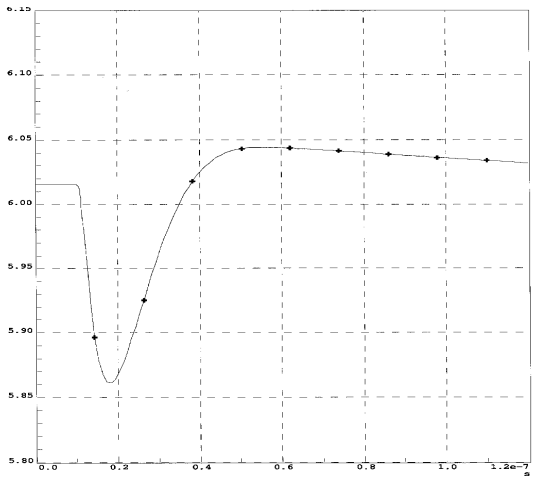

Fig. 6 Recovery transient of the regulator.

4. Conclusion

This paper has presented a very fast recovery word-line voltage regulator for ML NV memories. Circuit design was optimized for a 2-bit/cell 64-Mbit Flash memory chip in 0.2-μm CMOS technology. Although the design was optimized for the above purpose, the voltage regulator described is suited for use in any application where an accurate and stable voltage must be provided to a very large capacitive load and some discharged capacitance is connected to the regulator output during circuit operation, thereby causing an instantaneous voltage drop (e.g in switched-capacitor applications).

Acknowledgements

This work was partially supported by Italian MURST (Cofunded Project). The authors would like to thank I. Motta, J. Mulatti and A. Sacco for fruitful discussions.

References:

[1] M. Bauer, et al., A multilevel-cell 32Mb Flash memory, *Dig. Tech. Papers IEEE ISSCC 1995*, Feb. 1995, pp. 132-133.

[2] B. Riccò, et al., Nonvolatile multilevel memories for digital applications, *Proceedings of the IEEE*, vol. 86, no. 12, Dec. 1998., pp. 2399-2421.

[3] J. Dickson, On-chip high voltage generation in MNOS integrated circuits using an improved voltage multiplier technique, *IEEE J. Solid-State Circuits*, vol. SC-11, no. 3, June 1976, pp. 374-378.

[4] C. Calligaro, A. Manstretta, A. Modelli, and G. Torelli, Technological and design constraints for multilevel flash memories, *Third IEEE Int. Conf. on Electronics, Circuits, and Systems* (ICECS), Oct. 1996, pp. 1005-1008..

[5] T. Tanzawa and T. Tanaka, A dynamic analysis of the Dickson charge pump circuit, *IEEE J. Solid-State Circuits*, vol. SC-32, no 8, Aug. 1997, pp. 1231-1240,.

[6] G. A. Rincon-Mora and P. E. Allen, A low-voltage, low quiescent current, low drop-out regulator, *IEEE J.Solid-State Circuits*, vol. SC-33, Jan. 1998, pp 36-44..

[7] B. K. Ahuja, An improved frequency compensation technique for CMOS operational amplifiers, *IEEE J. Solid-State Circuits*, vol. SC-18, Dec. 1993, pp 629-633.