

G. Campardo  
R. Micheloni  
D. Novosel

# VLSI-Design of Non-Volatile Memories

 Springer

G. Campardo · R. Micheloni · D. Novosel

---

VLSI-Design of Non-Volatile Memories

Giovanni Campardo · Rino Micheloni · David Novosel

# **VLSI-Design of Non-Volatile Memories**

With 568 Figures

 Springer

Giovanni Campardo  
Rino Micheloni  
STMicroelectronics Srl  
Memory Product Groups  
Flash Division  
Via C. Olivetti 2  
20041 Agrate Brianza (MI)  
Italy

David Novosel  
IMD Intelligent  
Micro Design, Inc.  
Mercer-West Middlesex Road 2456  
West Middlesex, PA 16159  
U.S.A.

ISBN 3-540-20198-X **Springer Berlin Heidelberg New York**

Library of Congress Control Number: 2004116726

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in other ways, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable to prosecution under German Copyright Law.

**Springer is a part of Springer Science+Business Media**

springeronline.com

© Springer-Verlag Berlin Heidelberg 2005  
Printed in Germany

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Data conversion by the author.  
Final processing by PTP-Berlin Protago-TeX-Production GmbH, Germany  
Cover-Design: medionet AG, Berlin  
Printed on acid-free paper 62/3141/Yu - 5 4 3 2 1 0

## 2 Process Aspects

Before starting the design of the device, it is important to know the basic components that are at the designer's disposal to realize the circuitry, i.e. transistors, capacitors, resistors, diodes, bipolars and memory cells.

The characteristics of such components are defined by the process that, in the case of Flash Memories, is CMOS with two layers of poly and two or more of metal, with the recent addition of the triple well. After the process and the components have been defined, the circuit is designed to maximize the electrical performance, while the layout utilizing the available layers is realized in parallel.

The success of a design depends on the correct combination of process, design, and layout.

Before analyzing the components of the circuits, it is worth giving a brief overview of the different steps of fabrication of a typical CMOS process.

### 2.1 Introduction

The fabrication of an integrated circuit resembles the preparation of a *sandwich* in which, on a good slice of bread as a substrate, the sauces diffuse and several layers are harmoniously placed, one on top of the other, to satisfy the customer's taste. All is sealed with the final slice of bread that keeps the whole product together if the underlying layers are planar.

The technological process for the fabrication of integrated circuits is similar to the process described above.

### 2.2 Main Steps of Fabrication for a CMOS Process

The main steps that allow the fabrication of a device in CMOS technology will be described hereafter, with particular attention to the aspects that have the most impact on the design. The following description will show the sequence of process steps for an NMOS and a PMOS *low and high voltage* transistor, and two memory cells. The definition of the different layers is obtained by means of a number of photomasks, chrome images on glass plates that reproduce the polygons of the layout, defining different areas on the wafer through a photolithographic process.

The fabrication starts from a wafer of *p*-type silicon that forms the substrate; during the first operations, two twin-tubs, *n*-well and *p*-well, are realized.

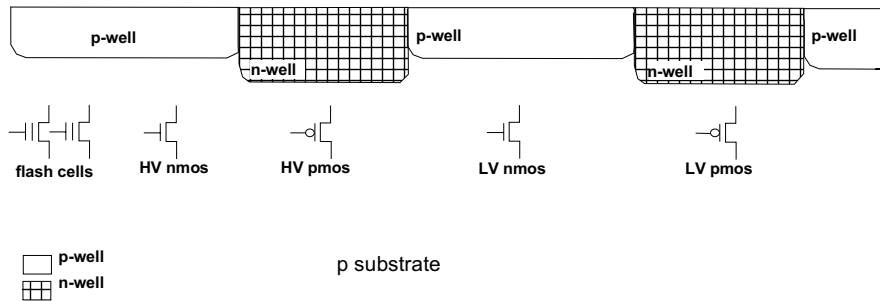


Fig. 2.1. Twin tub diffusion

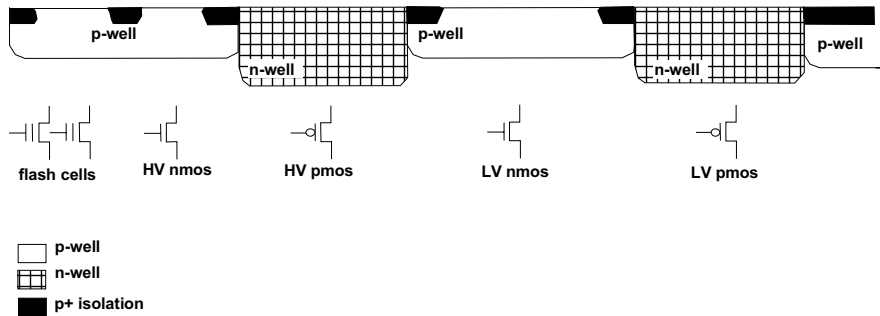


Fig. 2.2. Creation of the p+ insulation areas

The n-well in the *p*-substrate is necessary to fabricate *p*-type transistors, whereas the p-well is necessary to create an *ad-hoc* substrate for the *n*-type transistors (Fig. 2.1).

Subsequently, the insulating diffusions are created, so as to increase the threshold voltage of the parasitic elements that form between the junctions of two contiguous transistors. These parasitic devices are located beneath the thick oxide, also called *field oxide* (around 5000 Å of thickness) that constitutes an insulating barrier between active components (Fig. 2.2). The insulating diffusion, like any other selective step that acts in specific areas of the wafer, is realized by suitably masking the areas that must not be involved in that process step.

Subsequently, the thick oxide is grown on the insulating junctions (Fig. 2.3). The oxide for *high voltage* (HV) and *low voltage* (LV) *p* and *n* transistors are realized (typically around 250 Å and 120 Å, respectively) by means of two masks (Fig. 2.4).

The existence of HV and LV transistors is a characteristic of the memory devices that require voltages higher than VDD to program and erase.

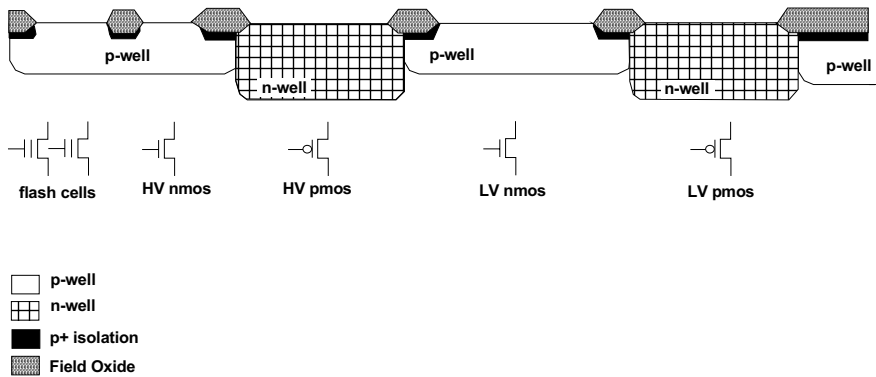


**Fig. 2.3.** Thick oxide growth

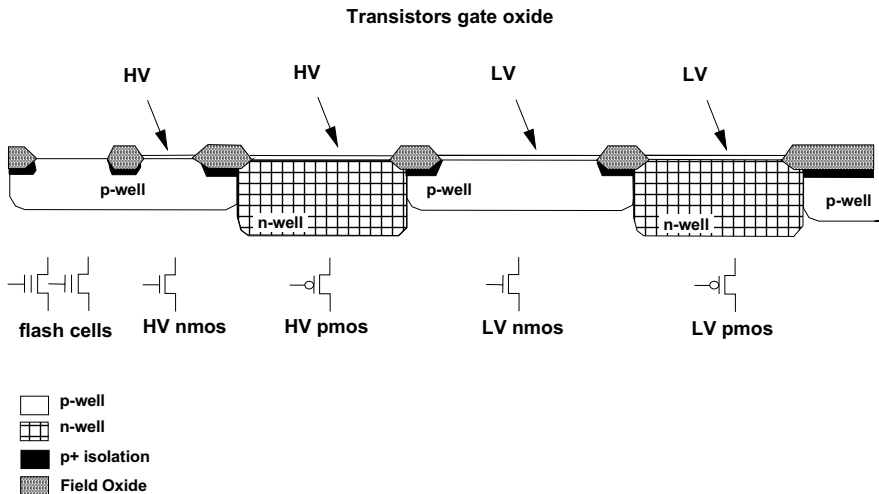


**Fig. 2.4.** Transistor gate oxide growth

The presence of two oxides allows optimizing the low voltage logic circuits, which have to be fast, and the high voltage circuits that, on the other hand, have to be robust. This also improves the functional aspects of the device such as reading speed, standby consumption, etc., since thick oxides limit the electric field at their side to about 4 MV/cm<sup>1</sup> at high voltages. Subsequently, the EPM<sup>2</sup> area is realized so as to implant the channel of the cells of the array. Such implants differentiate the cells from the typical n-channel transistors by setting the threshold voltage and

<sup>1</sup> This value is suggested by experience and reliability measurements.

<sup>2</sup> EPM is the acronym for Enhanced Program Mask.

optimizes the electrical characteristics since the cells are to be programmed and erased (Fig. 2.5).

Afterward, the tunnel oxide (100 Å) is grown directly on the EPM implant that delimits the area of the array. This step is perhaps the most critical since the quality of the tunnel oxide is essential to the device performance and reliability. (Fig. 2.6). The first polysilicon layer, poly1, which will form the floating gate of the cells of the array, is now defined. Typically this poly layer is deposited on all the transistors with the second poly deposited directly on poly1 to form the gate of the standard transistors.

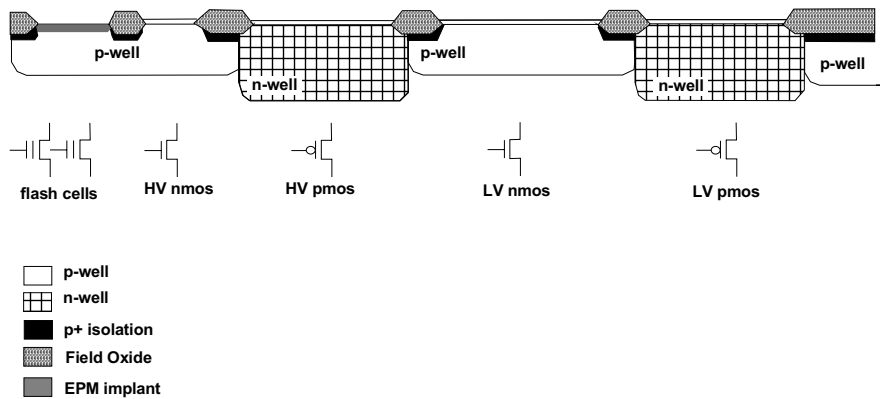


**Fig. 2.5.** Channel diffusion for memory cells

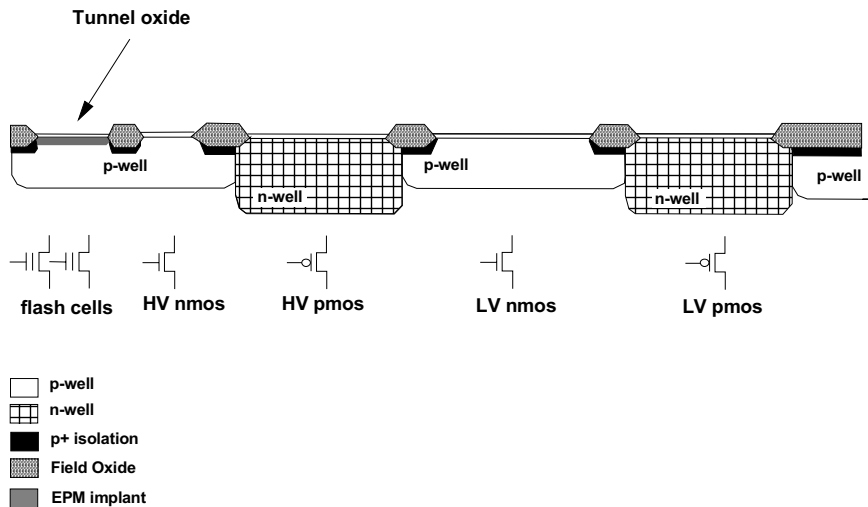


**Fig. 2.6.** Tunnel oxide growth

In this case, the process is called DSCP (Double Short-Circuited Poly) (Fig. 2.7). The layer that separates poly1 from poly2 is called ONO, which is the acronym for the composing materials, i.e. Oxide-Nitride-Oxide. This stacked composition improves the overall quality of the insulation between floating gate and control gate (Fig. 2.8).

Subsequently, the implants for low and high voltage transistors can be defined with another group of masks, to selectively differentiate the threshold voltages of the transistors and, then, deposit the poly2 layer that constitutes the control gate of transistors and cells (Fig. 2.9).

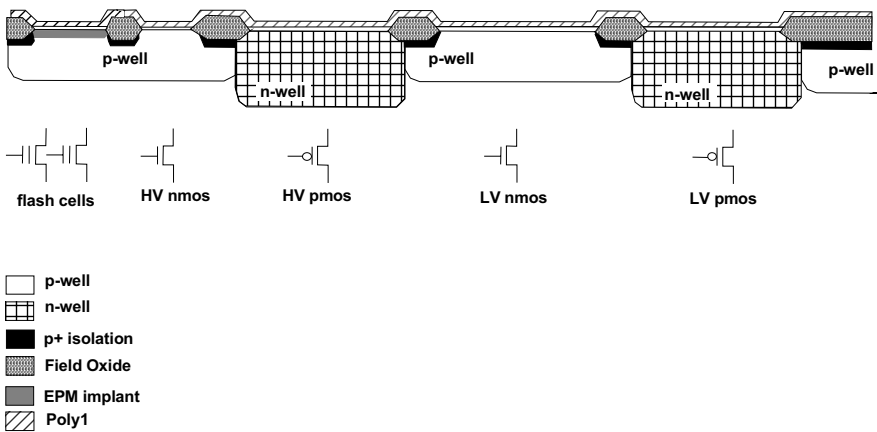


**Fig. 2.7.** Poly1 deposition

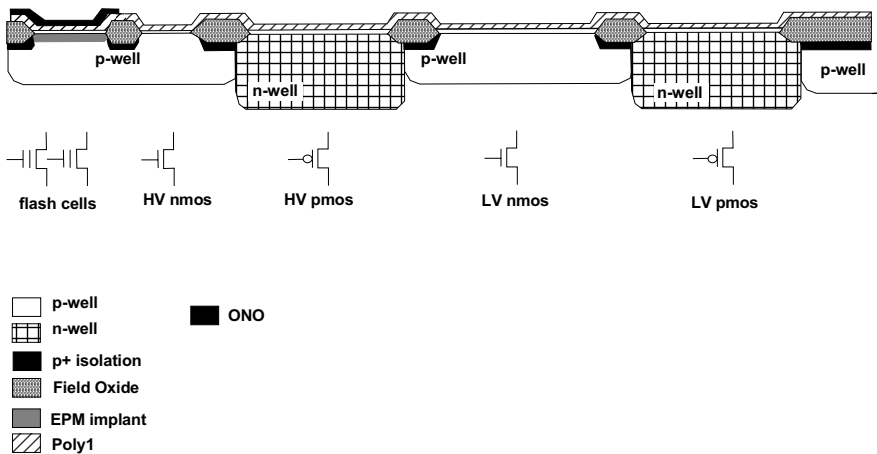


**Fig. 2.8.** Realization of the insulation between poly1 and poly2 (ONO)

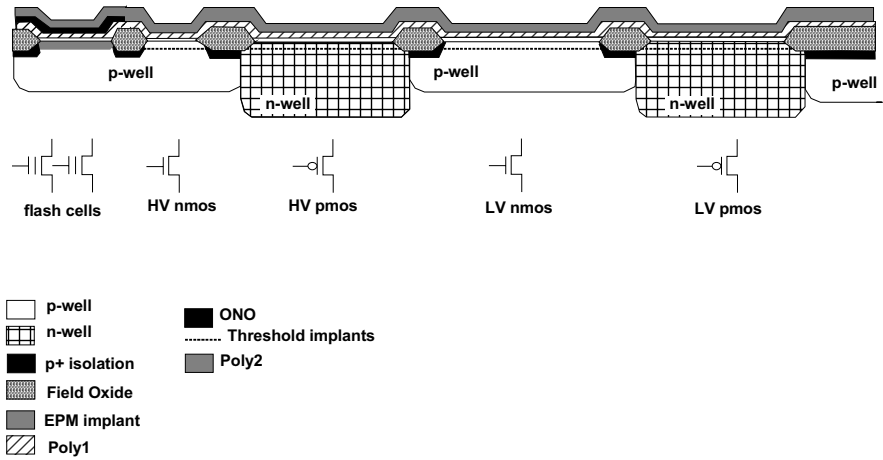


**Fig. 2.9.** Diffusions to define the MOS threshold voltages and deposition of poly2

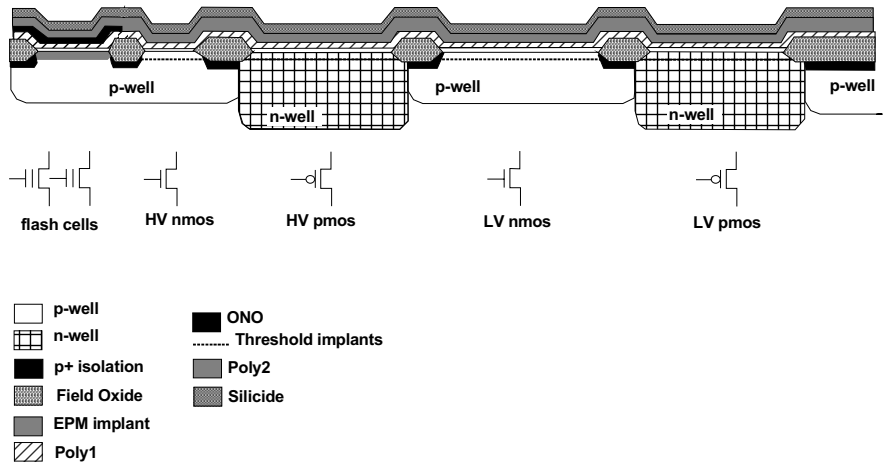


**Fig. 2.10.** Silicide deposition

The implants for the thresholds are realized through the poly1 already deposited.

On top of the poly2 layer, a metallic compound  $WSi_2$ , called silicide, is deposited, which decreases the resistivity of poly2 from around  $50 \Omega/\square$  to  $5 \Omega/\square$ . This step is very important in determining the access time because the rows of the cells are realized with poly2 as a conductor and, hence, their resistivity determines also the RC time constant of the rows (Fig. 2.10).

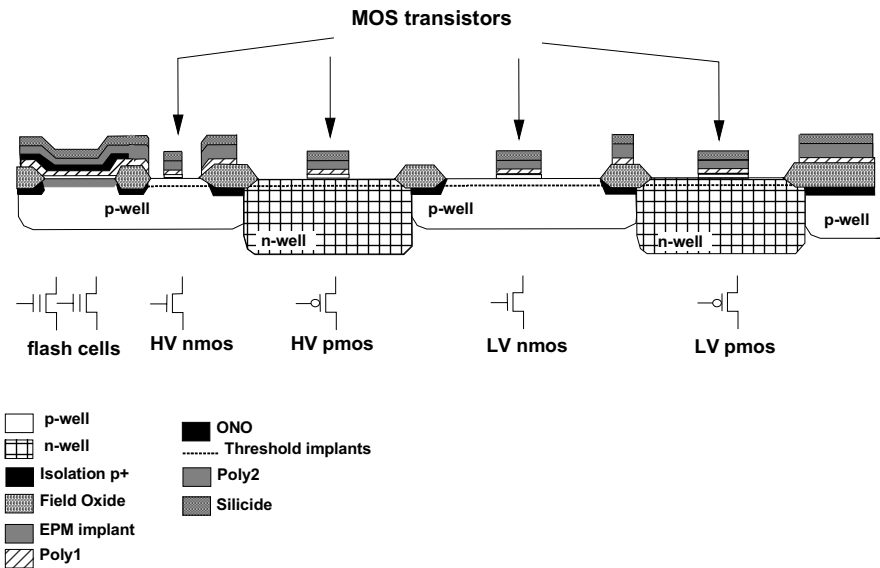


Fig. 2.11. Removal of the layers to define the transistors

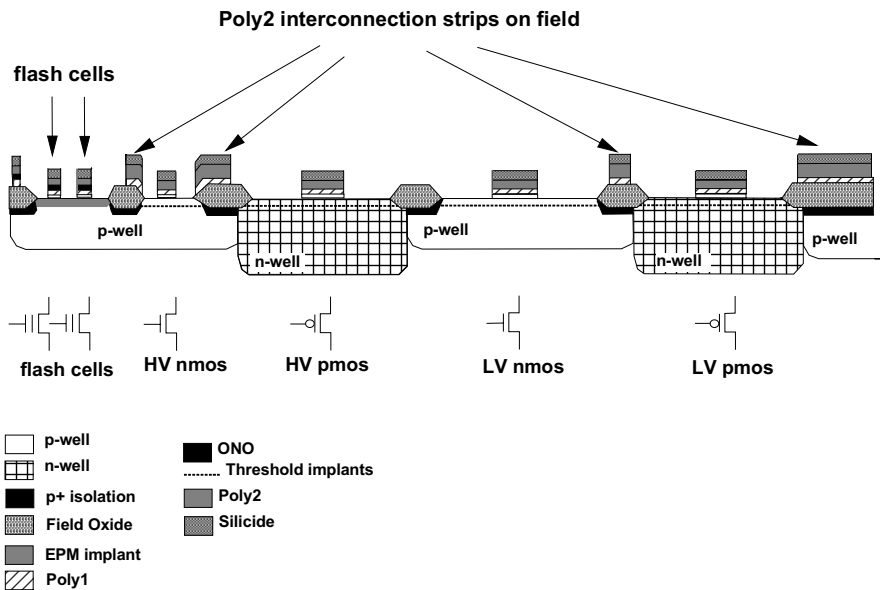


Fig. 2.12. Flash cell definition

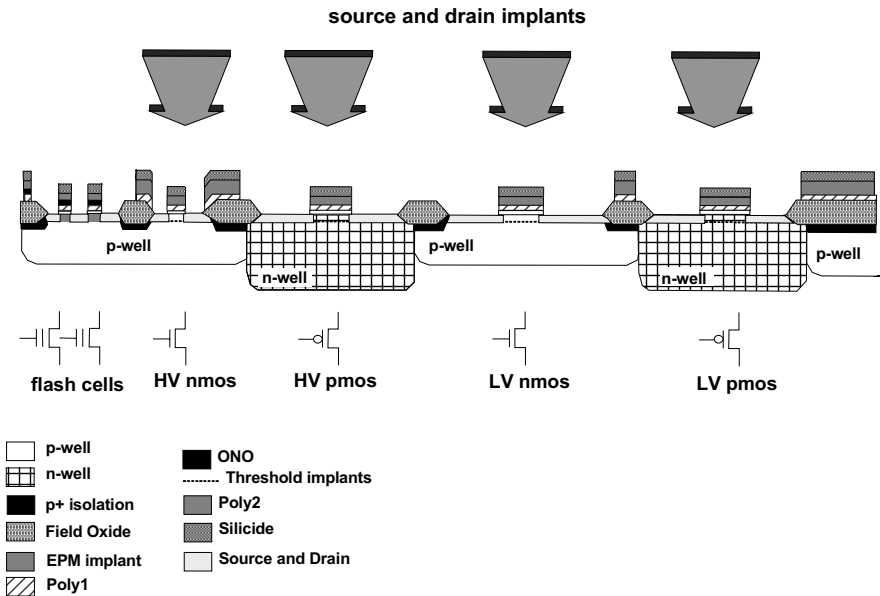


**Fig. 2.13.** Source and drain implants

After the basic layers have been fabricated, it is possible to etch them simultaneously, so as to obtain the structure of MOS transistors and cells as single element.

Transistors and cells are fabricated with different operations. In Fig. 2.11, the transistors have been obtained and, subsequently, the Flash cells are defined (Fig. 2.12). The portions of polysilicon that remain on the field oxide represent the strips of poly2 used as interconnect and are not the gates of the transistors. The source and drain areas for transistors and cells are diffused using other masks, as shown in Fig. 2.13.

The aluminum-silicon alloy connection layer called *Metal* is then deposited. Before that, however, the “holes” that allow the contacts between poly2, active areas, and the different layers of metal, are created. The first step is the insulation of the surface with an intermediate layer of oxide (Fig. 2.14). Afterward, the “holes”, called *contacts* in this case, are cut into the insulator so as to allow the connection between *metal1*, i.e. the first metal layer, and either poly2 or active area (Fig. 2.15).

The *metal1* layer is deposited so as to penetrate into the contacts, thus forming the electrical connection. Metal1 is deposited on the whole wafer (Fig. 2.16), and removed from the areas where it is not needed by means of a masking process. The surface is then completely covered with dielectric materials having the function of both insulating metal1 from the layers that will be deposited thereafter, and planarizing the surface itself. Such materials are called TEOS<sup>3</sup> and SOG<sup>4</sup>.

<sup>3</sup> TEOS is the acronym for TetraEthylOrho Silicate.

<sup>4</sup> SOG is the acronym for Spin-On-Glass.

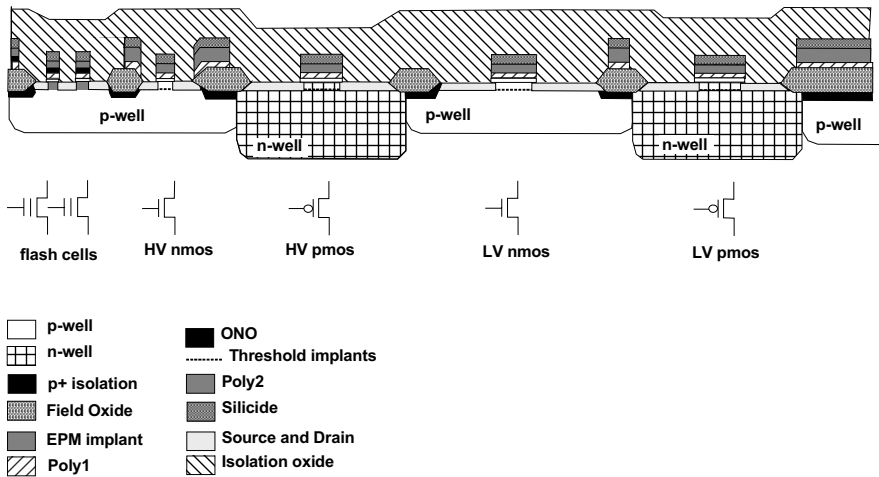


Fig. 2.14. Intermediate insulating oxide

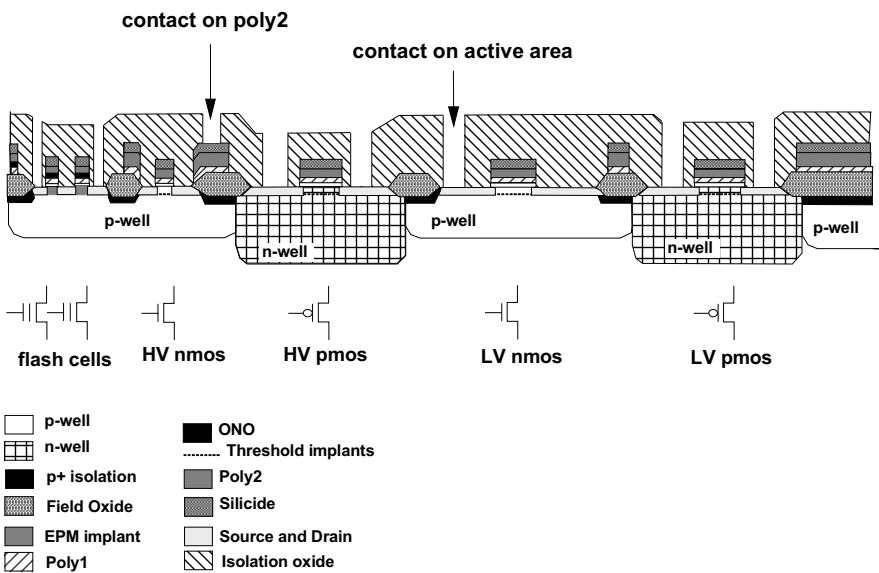


Fig. 2.15. Contact definition

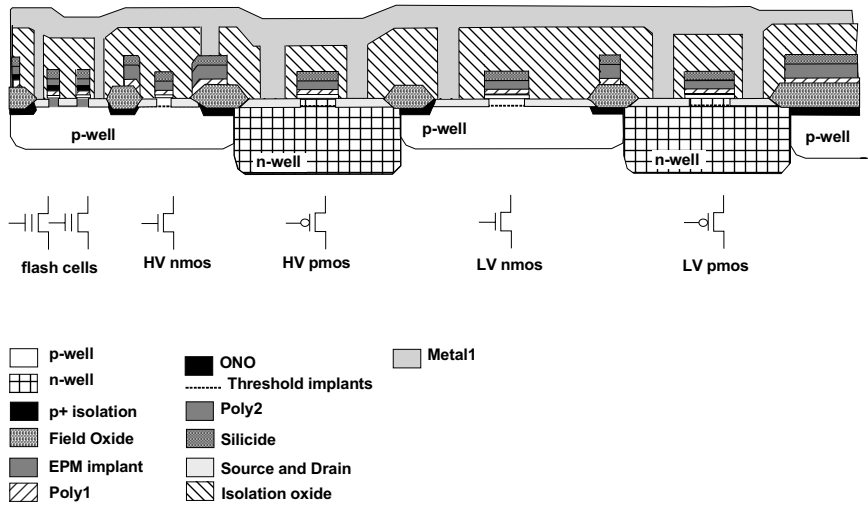


Fig. 2.16. Metal1 deposition

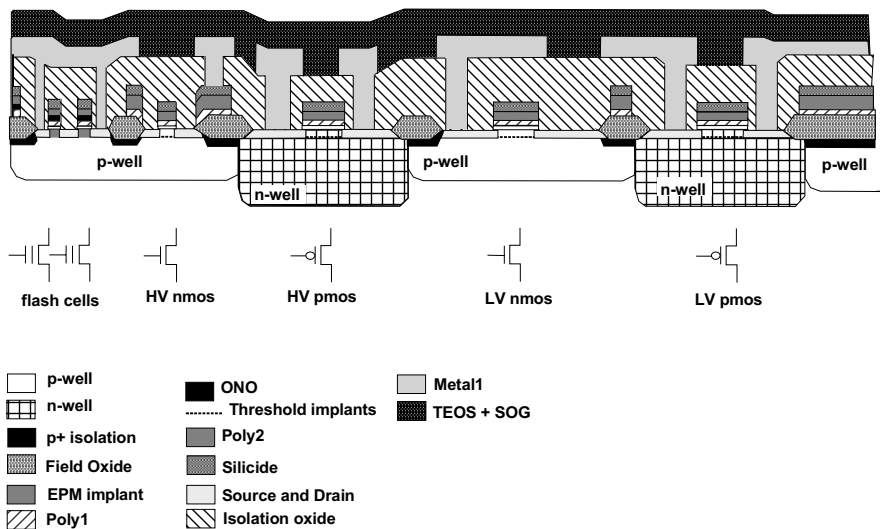


Fig. 2.17. Metal1 definition and intermetal dielectric deposition

Planarity enables the deposition and etching to be more precise and, therefore, smaller dimensions can be defined (Fig. 2.17). Afterward the planarization, the “holes” that allow the connection between metal1 and metal2, called *vias*, are opened (Fig. 2.18). The metal2 is deposited with the same technique used for metal1 and then it is etched and removed from the areas where it is not necessary (Fig. 2.19). Finally, the entire area of the surface is covered with a vitreous layer, or passivation, having the function of sealing and protecting from the external environment.

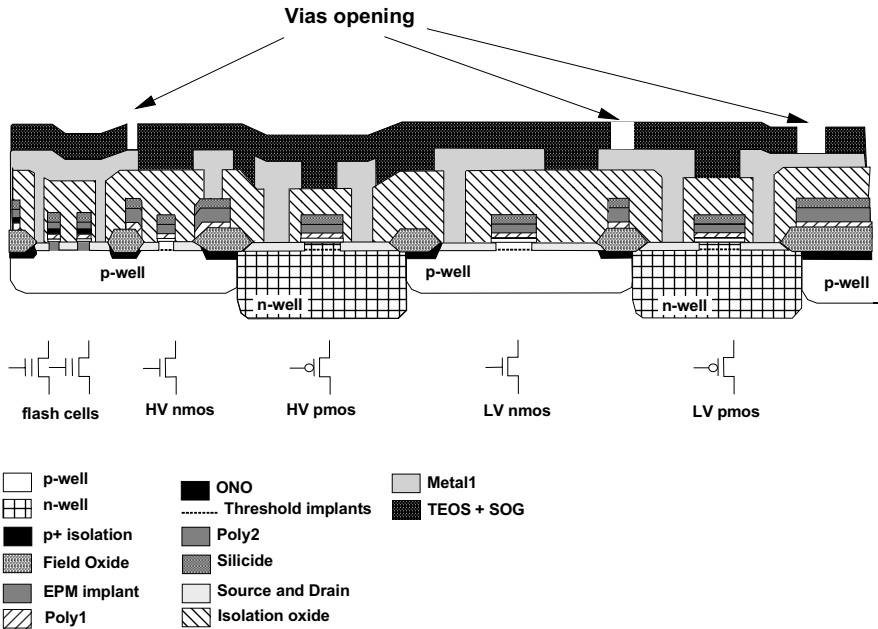


Fig. 2.18. Vias opening

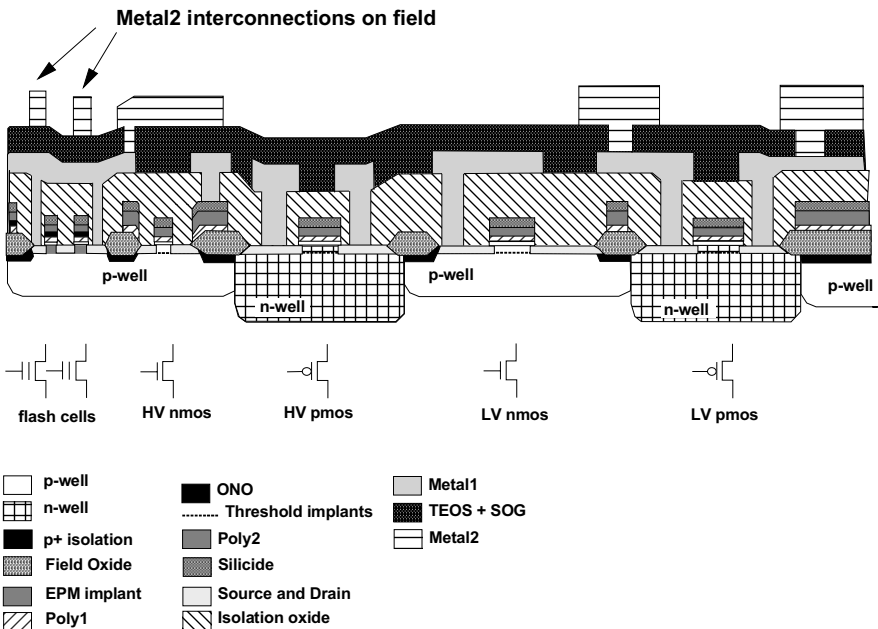


Fig. 2.19. Metal2 deposition

The last operation is the opening of the pad areas in the passivation layer, where the wires are bonded to make connection from the integrated circuit to the pins of the package (Fig. 2.20).

Subsequently, the wafer is lapped, i.e. it is thinned by means of chemical abrasion so as to make it easier to assemble it into the package and, moreover, improve the conduction of heat. The device is then submitted to electrical and functional tests.

The most important steps of a typical CMOS process with two layers of poly and metal for non-volatile cells have been presented. The process is actually constituted of many more steps that can reach 300, the majority of which are cleaning steps. For each major operation, the recipes are defined in terms of temperature environment, gas, atomic species etc. and only the steps that are directly involved in the circuit design have been described in this chapter.

Another characteristic of the process is the *triple well*, which facilitates a simpler solution to erase the Flash memory.

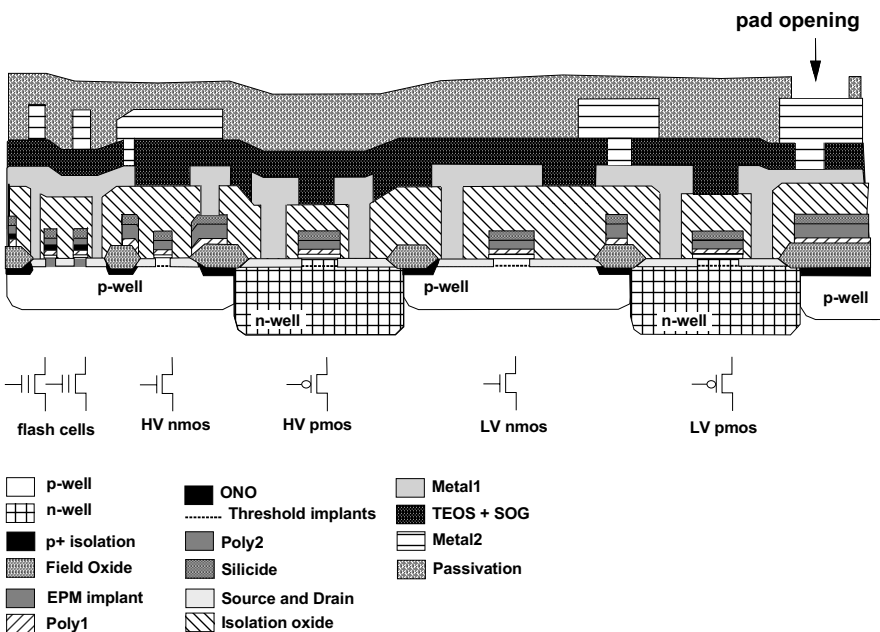


Fig. 2.20. Passivation and pad opening

In order to solve the problem of the erasing Flash cells by use of a negative voltage on the row without increasing the complexity of the circuit, it is useful to create n-channel transistors having substrate that can be biased independently. Areas with n-wells are fabricated through a doped buried layer, thus realizing a p-type tub insulated from the substrate, which is also p-type (Fig. 2.21).

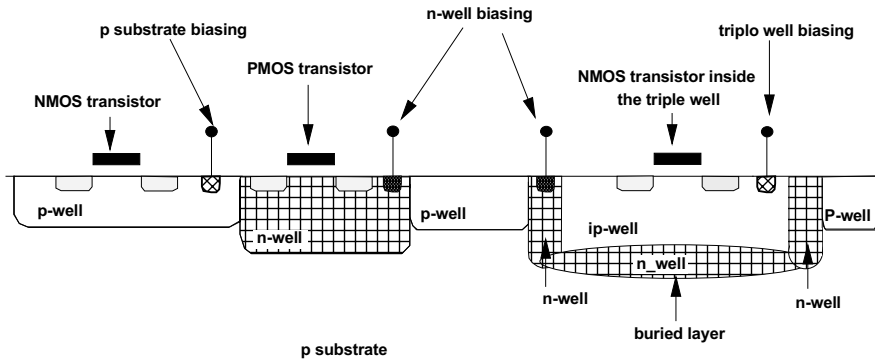


**Fig. 2.21.** Device section that highlights the triple-well transistors with their bias

## Bibliography

- C. Auricchio, R. Bez, A. Lo savio, A. Maurelli, C. Sala, P. Zabberoni G. Baccarani and M. Rudan, "A triple-well architecture for low voltage operation in submicron CMOS devices", (Eds.), Proc. ESSDERC 96, Bologna, Italy, p. 613, (1996).
- T.M. Bloomstein, M. Rothschild, R.R. Kunz, D.E. Hardy, T.B. Goodman, and S.T. Palmacci, "Critical issues in 157 nm lithography", J. Vac. Sci. Technol. B, vol.16, pp. 3154, (1998).
- J.O. Borland, R. Koelsch, "MeV implantation technology: next generation manufacturing with current generation equipment", Solid State Technology, P.1 (1993).
- P. Cappelletti, A. Tutolo, L. Frattin, L. Ravazzi, C. Riva, "The Flash E2PROM cell with Boron P-pocket architecture: advantages and limitations", Proc. Nonvolatile Semiconductor Memory Workshop, (1995). C.T. Gabriel, and J.P. McVittie, "How Plasma Etching Damages Thin Gate Oxides", Sol. State Technology, 81, (June 1992).
- T.Y. Chan, P.K. Ko and C. Hu "Dependence of channel electric field on device scaling" IEEE Electron Dev. Lett., EDL-6, p.551, (Oct. 1985).
- M.L. Chen et al. "Suppression of Hot-Carrier Effects in Submicron CMOS Technology" IEEE Trans. Electron Dev., ED-35, p2210, (Dec. 1988).
- E.G. Cromer, "Mask aligners and stepper for precision microlithography, Solid State Technology, 36, (April 1993).
- J.D. Cuthbert, "Optical projection printing", Solid State Technology, 20, 59, August (1977).
- F.H. Dill, W.P. Hornberger, P.S. Hauge, J.M. Shaw, "Characterization of positive resist", IEEE Trans. Electron. Dev., ED22, 445, (1975).
- G. Dunn, S.S. Scott, "Channel hot-carrier stressing of reoxidized nitrided silicon dioxide", IEEE Trans. On Electron Device, 37, 7, p. 1719, (1990).
- L. Forester, A.L. Butler and G. Schets, "SOG planarization for Polysilicon and First Metal Interconnect in a One Micron CMOS Process" Proceedings of the 6<sup>th</sup> International IEEE VLSI Multilevel Interconnection Conference, p.72, (1989).
- Fowler-Nordheim tunneling on source doping concentration of an n-MOSFET", IEEE Electron Device Letters, 17,11, p.525, (1996).

- G. Ginami, et al., "Survey on Flash technology with specific attention to the critical process parameters related to manufacturing", IEEE Proceeding of the, Vol. 91, No. 4, pp. 503-522, April (2003).
- C. Hill, S. Jones and D. Boys "Rapid Thermal Annealing – Theory and Practice" in R.A. Levy Ed., Reduced Thermal Processing for ULSI, Plenum Press, New York, p. 143, (1989).
- IEEE Standard Department, "IEEE P1005 draft standard for definitions, symbols, and characteristics of floating gate memory arrays", approved 1998.
- E. Kooi, "The invention of LOCOS", IEEE Press, NY, (1991).
- M.D. Levenson, N.S. Viswanathan, R.A. Simpson, "Improving resolution in photolithography with a phase-shifting mask", IEEE Trans. Electron. Dev, ED29, 1828, (1982).
- B.J. Lin, "The optimum numerical aperture for optical projection microlithography", Proc. SPIE, 1463, 42, (1991).
- R.A. Morgan, "Plasma etching in semiconductor fabrication", Elsevier, NY, (1985).
- S. Mori, E. Sakagami, H. Araki, Y. Kaneko, K. Marita, Y. Ohshima, N. Arai, K. Yoshikawa, "ONO interpoly dielectric scaling for non-volatile memories applications", IEEE Trans. On Electron Devices, 38, 2, p. 386-391, (1991).
- E.G. Moore, "Cramming more components onto integrated circuits", Electronics Magazine, vol. 8, pp. 114-117, (April 1965). M. Born, E. Wolf "Principles of optics", Pergamon Press
- A. Nakae, K. Kamon, T. Hanawa, H. Tanabe, "Improvement in optical proximity correction by optimizing second illumination source", Jpn. J. Appl. Phys., Pt.1, vol.35, pp. 6396, (1996).
- J.K. Roberts "Heat and Thermodynamic" 4<sup>th</sup> ed. Blackie and Son, Ltd., Glasgow, (1995).
- K.C. Saraswat, D.L. Brors, J.A. Fair, K.A. Monnig and R.Beyers "Properties os Low Pressure CVD Tungsten Silicide for MOS VLSI Interconnections" IEEE Trans. Electron Dev., ED-30, 1497, (1983).
- K.C. Saraswat, F. Mohammadi and J.D. Meindl "WSi<sub>2</sub> Gate MOS Device" Tech. Dig. IEDM, 462, (1979).
- Semiconductors Industry Association, "International Roadmap for Semiconductors: 2002".
- K. Shibahara et al., "Trench Isolation with V-shaped Buried Oxide for 256-Mbit DRAMs" Tech Dig. IEDM, p. 275, (1992).
- Y. Tang, J. Chen, C. Chang, D. Liu, S. Haddad, Y. Sun, A. Wang, M. Ramskey, M. Kwong, H. Kinoshita, W. Chan, J. Lien, "Different dependence of band-to-band and H. Tsai, C.L. Yu, C.Y. Wu, "A bird's beak technique for LOCOS in VLSI fabrication", IEEE Electron Device Letters, 7, 2, pp. 122-123, (1986).
- K. Tsukamoto et al. "High Energy Ion Implantation for ULSI: Well Engineering and Gettering", Solid State Technology, p.49, (June 1992).
- N.A.H. Wils, P.A. van der Plas, A.H. Montree, "Dimensional characterization of poly buffer LOCOS in comparison with suppressed LOCOS", ESSDERC 90, pp. 535-538, (1990).

## 3 The MOSFET Transistor and the Memory Cell

Although this book addresses topics that require previous basic knowledge of the MOS transistor, an overview of the principles of such component will not be omitted. For a rigorous and complete exposition on the MOS transistor behavior, the reader should refer to specialized literature. This book will focus on certain aspects of the MOS transistor characteristics, and examples are given from the standpoint of the designer, who tends to reason in terms of potential, current, and impedance. Furthermore, the main characteristics of the Flash memory cell will be presented in read, program, and erase mode. Also in this case, the dissertation is reduced to the essential notions to understand the problems of the design. A brief history of the electrical erasing will be outlined in the case of Flash Memories, from the first generation of devices with double bias, to the current devices having a single VDD.

### 3.1 The MOSFET Transistor

The MOSFET transistor (or simply MOS) is defined as a voltage controlled current source. When the behavior of the device with respect to small signals is to be studied, which means small deviations with respect to a fixed steady condition (called bias or operating point), the equations of the representing model can be linearized<sup>1</sup>.

The equivalent circuit of the MOS is represented in Fig. 3.1, in which the gate terminal G is insulated, since it is supposed that no current flows into the gate, and the current generator is controlled only by means of the voltage.

The source terminal S has lower potential than the drain D in case of n-channel transistor.

---

<sup>1</sup> There are several ways to define what is meant by “small signal”. For a MOS transistor, for example, the ratio between bias current in saturation region and signal current can be considered. The small signal condition is then  $v_{gs} \ll (V_{gs} - V_T)$ , where the voltage in lowercase refers to the signal, whereas the bias quantities are indicated in uppercase. Owing to the notion of small signal, it is possible to consider only the variations and, hence, in the analysis of the equivalent circuit, both VDD and ground are considered to be at the same potential, being stationary with respect to the signal variations.

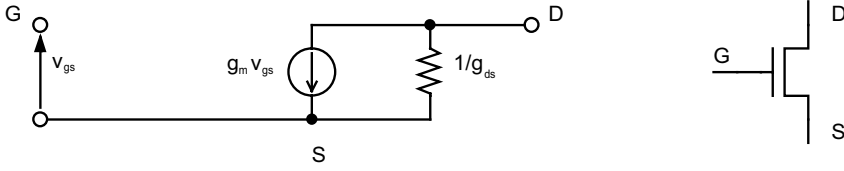


**Fig. 3.1.** Equivalent circuit and graphic symbol of an NMOS transistor

In Fig. 3.1  $g_{ds}$  is the output conductance,  $g_m$  is the transconductance since it links the variation of an input quantity, the voltage  $V_{GS}$ , to an output quantity, the drain current  $I_{DS}$ .

$$g_m = \frac{\partial I_{DS}}{\partial V_{GS}} \quad (3.1)$$

$$g_{ds} = \frac{\partial I_{DS}}{\partial V_{DS}} \quad (3.2)$$

The parameters described are rarely used when designing from scratch, the time when the ideas are converted into circuits. Only the simulation programs can manage them so as to solve the systems of equations and obtain the potentials of all the nodes, and the currents of all the branches. The relevant aspect is that, as it will be shown,  $g_m$  is proportional to the aspect ratio ( $W/L$ ), which is the only parameter that the designer can control.

What is the equivalent resistance  $R_{eq}$  of a PMOS transistor connected to the supply voltage, as shown in Fig. 3.2? If we suppose that the load resistance  $R$  is large enough so that the current drawn from the transistor is less than the capability of supplying current of the transistor itself, the voltage  $V_{OUT}$  is very close to VDD.

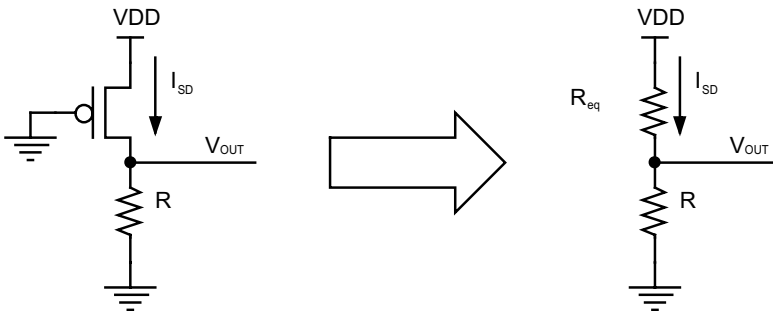


**Fig. 3.2.** Equivalent resistance for a PMOS transistor working as a current source

For example, if  $V_{OUT}$  is 100 mV lower than VDD,  $R_{eq}$  can be calculated by means of the Ohm's law:

$$R_{eq} = \frac{100mV}{I_{SD}} \quad (3.3)$$

For example, with  $I_{DS}$  of 300  $\mu A$ ,  $R_{eq}$  of 333  $\Omega$  is obtained.

A more accurate model that takes into account the effect of bias of the substrate (body effect) is shown in Fig. 3.3.

In this case  $g_{mb}$  is defined as:

$$g_{mb} = \frac{\partial I_{DS}}{\partial V_{BS}} \quad (3.4)$$

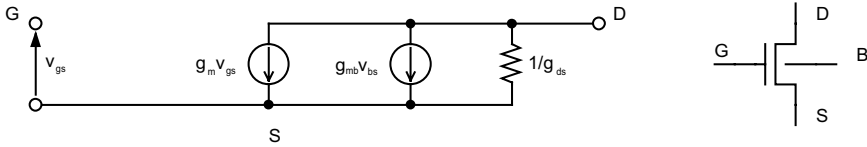


**Fig. 3.3.** Equivalent circuit and graphic symbol for an NMOS transistor where the substrate bias source  $V_{BS}$  is highlighted

The effect of the bias between bulk B and source S for a MOS transistor is quantified by means of the expression of the threshold voltage  $V_T$ . Conventionally, it is assumed that the conduction in a MOS transistor starts when the difference of potential between gate and source exceeds the value  $V_T$  that, in case of n- and p-channel, can be calculated through the following formulae:

$$V_{T,n} = V_{T0} + \gamma \cdot \left( \sqrt{2 \cdot |\Phi_p| + |V_{SB}|} - \sqrt{2 \cdot |\Phi_p|} \right) \quad (3.5)$$

$$V_{T,p} = V_{T0} - \gamma \cdot \left( \sqrt{2 \cdot |\Phi_n| + |V_{SB}|} - \sqrt{2 \cdot |\Phi_n|} \right) \quad (3.6)$$

$$\Phi_p = -\frac{kT}{q} \ln \left( \frac{N_a}{n_i} \right) \quad (3.7)$$

$$\Phi_n = -\frac{kT}{q} \ln \left( \frac{N_d}{n_i} \right) \quad (3.8)$$

$$\gamma = \frac{\sqrt{2 \cdot q \cdot \epsilon_0 \cdot \epsilon_{Si} \cdot N_a}}{C_{ox}} \quad (3.9)$$

|                        |                                                           |                                         |
|------------------------|-----------------------------------------------------------|-----------------------------------------|
| $q$                    | $= 1.602 \cdot 10^{-19} \text{ [C]}$                      | electronic charge                       |
| $k$                    | $= 8.62 \cdot 10^{-5} \text{ [eV/K]}$                     | Boltzmann's constant                    |
| $T$                    | $= \text{[K]}$                                            | absolute temperature                    |
| $\epsilon_0$           | $= 8.854 \cdot 10^{-14} \text{ [F/cm]}$                   | permittivity of vacuum                  |
| $\epsilon_{\text{Si}}$ | $= 11.7$                                                  | relative permittivity of Si             |
| $\epsilon_{\text{ox}}$ | $= 3.9$                                                   | relative permittivity of $\text{SiO}_2$ |
| $N_a$                  | $= \text{[atoms/cm}^3\text{]}$                            | doping density of the p-type substrate  |
| $N_d$                  | $= \text{[atoms/cm}^3\text{]}$                            | doping density of the n-type substrate  |
| $N_i$                  | $= 1.45 \cdot 10^{10} \text{ [cm}^{-3}\text{]}$           | intrinsic carrier concentration         |
| $C_{\text{ox}}$        | $= \epsilon_0 \cdot \epsilon_{\text{ox}} / t_{\text{ox}}$ | oxide capacitance per unit area         |
| $t_{\text{ox}}$        | $= \text{[nm]}$                                           | oxide thickness                         |

$\Phi_p$  is the p-type substrate potential, about 0.6 V, whereas  $V_{T0}$  is the threshold without the contribution of the body (i.e. with  $V_{\text{SB}} = 0$ ). For a rough estimation, it is useful to consider the variation of the threshold as proportional to the square root of the voltage between bulk and source multiplied by the body effect coefficient  $\gamma$ . Such parameter ranges between 0.3 and 1.0  $\text{V}^{1/2}$  in a typical CMOS process.

**Problem 3.1:** Identify some circuits in which the body effect is disadvantageous and others in which, on the contrary, such effect is useful. Discuss the circuits identified.

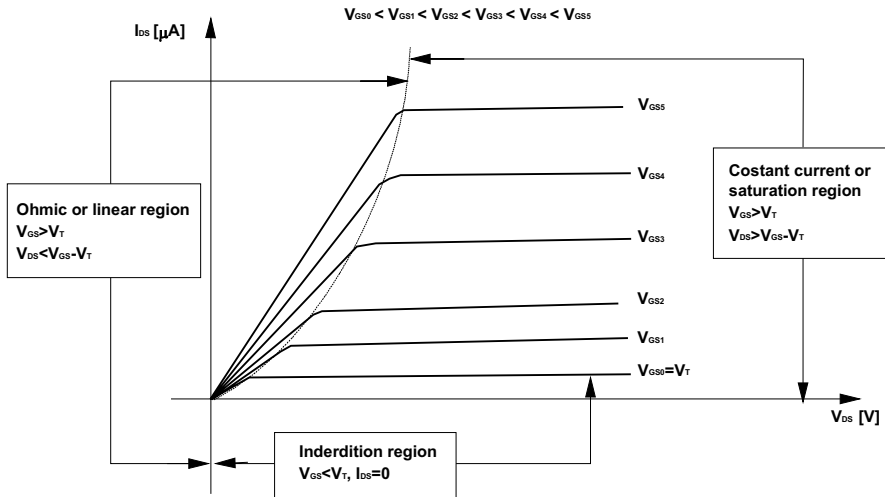


**Fig. 3.4.** Output characteristic for an NMOS transistor

In Fig. 3.4 the output characteristic of an NMOS transistor is reported. In the linear region (or triode) the following relationships exist:

$$V_{\text{DS}} < V_{\text{GS}} - V_T \quad (3.10)$$

$$I_{DS} = \frac{1}{2} \cdot \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot \left[ 2 \cdot (V_{GS} - V_T) \cdot V_{DS} - V_{DS}^2 \right] \quad (3.11)$$

The symbol  $\mu_n$  is the electron mobility that, in silicon, is  $1417 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$ . The W/L ratio is called aspect ratio of the transistor, where W is the channel width and L is the channel length.

In the saturation region in which  $V_{DS} \geq (V_{GS} - V_T)$ , Eq. (3.11) is no longer valid and the drain current can be expressed as:

$$I_{DS} = \frac{1}{2} \cdot \mu_n \cdot C_{ox} \cdot (V_{GS} - V_T)^2 \cdot (1 + \lambda \cdot V_{DS}) \quad (3.12)$$

The term  $(1 + \lambda V_{DS})$  accounts for the fact that in saturation region the transistor characteristic is not parallel to the x-axis but, on the contrary, there is a certain slope (the value of  $\lambda$  ranges between 0.03 and  $0.005 \text{ V}^{-1}$ ). The parameter  $\lambda$  is usually referred to as channel modulation parameter. Finally, the important parameters for frequency analysis are the gate, gate-to-source, and gate-to-drain parasitic capacitance.

At this point, the transconductance and output conductance of the small signal model  $g_m$  and  $g_{ds}$  can be calculated in the two working regions.

In the linear region, using Eq. (3.11) we obtain:

$$g_m = \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot V_{DS} \quad (3.13)$$

$$g_{ds} = \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot (V_{GS} - V_T - V_{DS}) \quad (3.14)$$

Starting from Eq. (3.12), the transconductance in saturation region can be calculated:

$$g_m = \mu_n \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \cdot (V_{GS} - V_T) \cdot (1 + \lambda \cdot V_{DS}) = \frac{2 \cdot I_{DS}}{(V_{GS} - V_T)} \quad (3.15)$$

$$g_{ds} = \frac{\lambda \cdot I_{DS}}{(1 + \lambda \cdot V_{DS})} \quad (3.16)$$

Similar expressions can be obtained for p-channel transistors, keeping in mind that  $V_{T,p}$  is negative and that in silicon  $\mu_p$  equals  $471 \text{ cm}^2\text{V}^{-1}\text{s}^{-1}$ .

**Problem 3.2:** The name of the working regions (linear and saturation) for MOS transistors is inverted with respect to bipolar transistors: find out why.

## 3.2 Transistors Available

The transistors that the designer has at his disposal depend on the kind of process and on the performance required to realize the memory device. For a device oper-

ating at low supply voltage, for example, it is important the use of transistors having different thresholds.

Hereafter a description of the most used types of transistors will be given: p-channel transistors with thin and thick oxide, n-channel transistors fabricated in both common substrate and local triple well p-type tubs, called ip-well for sake of clarity. N-channel transistors can be LVS type (Light Voltage Shift) or NAT (Natural) depending on the kind of threshold required, and, moreover, the LVS can be either HV or LV, depending on the oxide thickness. The transistors with thin oxide have the advantage of a lower threshold voltage than those with thick oxide, but they are not able to manage voltages above VDD, and, typically, they also have shorter minimum channel length. N-channel natural transistors (NAT) have threshold voltage than is even lower than LVS, and, generally, only the low voltage version is fabricated. The HV, LV and NAT versions are also available for p-channel transistors. Table 3.1 summarizes the possible types of transistors with the corresponding symbol that will be found in the schematics of the following chapters. In the table, the values of the threshold voltage at room temperature, the oxide thickness for a typical process, the value of the body effect coefficient  $\gamma$ , and, finally, the value of the current that a square of transistor is able to source or sink in saturation with  $V_{GS}$  and  $V_{DS}$  equals to VDD are reported.<sup>2</sup>

**Table 3.1.** Typical transistors available in a CMOS process and main useful parameters

| Name        | Symbol outside and inside triple well | Threshold typical @ 27 °C [V] | $T_{ox}$ thickness gate oxide [Å] | Body effect coefficient [V] <sup>1/2</sup> | $I_{DS}$ current/square [μA/square] |
|-------------|---------------------------------------|-------------------------------|-----------------------------------|--------------------------------------------|-------------------------------------|
| NMOS LVS LV |                                       | 0.6                           | 120                               | 0.95                                       | 360                                 |
| NMOS LVS HV |                                       | 0.8                           | 250                               | 0.96                                       | 300                                 |
| NMOS NAT LV |                                       | -0.1                          | 120                               | 0.41                                       | 380                                 |
| PMOS LVS LV |                                       | -0.6                          | 120                               | 0.42                                       | 150                                 |
| PMOS LVS HV |                                       | -0.8                          | 250                               | 0.6                                        | 130                                 |
| PMOS NAT LV |                                       | -1.5                          | 120                               | 0.42                                       | 250                                 |
| NMOS DEP LV |                                       | -3.5                          | 120                               | 0.9                                        | 360                                 |

<sup>2</sup> Such parameters are specific to each process and in this case must be considered merely as an example, to provide some figures to discuss.

The circuit design must be based on actual measurements of the transistors from recently processed wafers, to insure the correct dimensioning of the circuitry.

It is important to state that the transistors do not act as ideal current sources. As we know, two well defined operating regions exist, the saturation region, in which the current supplied by the transistor  $I_{DS}$  is nearly constant with respect to the voltage  $V_{DS}$ , and the linear (Ohmic) region, in which the behavior of the transistor is resistive.

Let's consider a circuit like the one in Fig. 3.5 in which M1 is used to charge the capacitor  $C_{LOAD}$ , with  $V_{IN} = VDD$ ; if M1 works as an ideal current source having value of  $I_L$ , we obtain:

$$I_{LOAD} = C_{LOAD} \cdot \frac{dV_{OUT}}{dt} \quad (3.17)$$

$$\Delta V_{OUT} = \frac{I_{LOAD}}{C_{LOAD}} \cdot \Delta t \quad (3.18)$$

with the initial condition of  $V_{OUT} = 0V$  and  $t = 0$ . Thus, the relationship between the voltage of the output node and the charging time is linear.

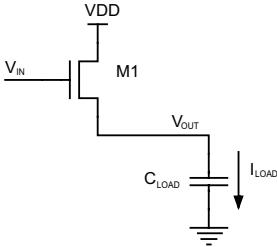


**Fig. 3.5.** N-channel transistor used to charge a capacitor

In reality, the capacitor is initially uncharged, the current starts flowing and the output voltage increases, varying  $V_{GS}$  and  $V_{DS}$  of the transistor according to the load characteristic shown in Fig. 3.6. The charge finishes when M1 switches off, i.e. when  $V_{OUT}$  reaches the value of VDD minus the threshold voltage of M1.

Moreover, the body effect has to be taken into account, since it causes the increase of the threshold voltage of M1 as the output voltage  $V_{OUT}$  increases, slowing down the charging of the output node and stopping it at a value lower than  $VDD - V_{T0}$ .

**Problem 3.3:** Determine the value of  $V_{OUT}$  taking into account the body effect.

Let's analyze how the charge of  $C_{LOAD}$  occurs through a p-channel transistor M2. In this case,  $V_{IN} = GND$ ,  $V_{GS}$  is fixed because it depends on VDD, not on  $V_{OUT}$  like in the case of the n-channel transistor. The capacitor is charged with a constant current until  $V_{DS}$  of M2 is in saturation region.

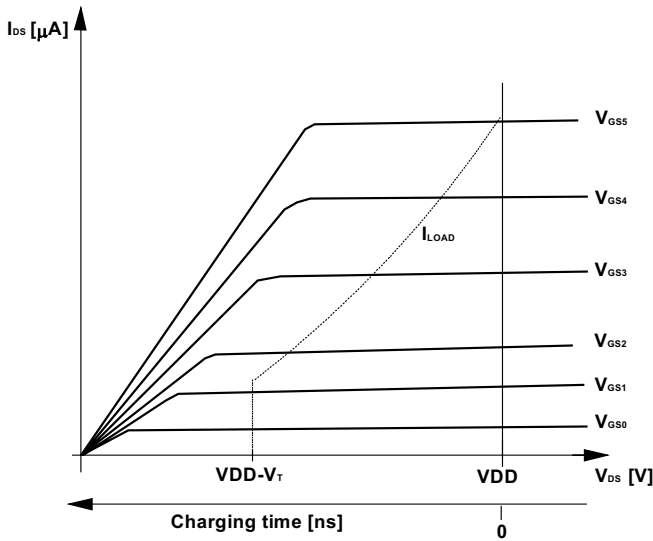


Fig. 3.6. M1 current during  $C_{LOAD}$  charge

After that, in the linear region, the output voltage transient is of exponential type, considering that the transistor dynamic impedance varies for each value of  $V_{DS}$  of M2. In this case,  $I_{LOAD}$  characteristic is represented by the curve named  $V_{GS5}$  in Fig. 3.7.

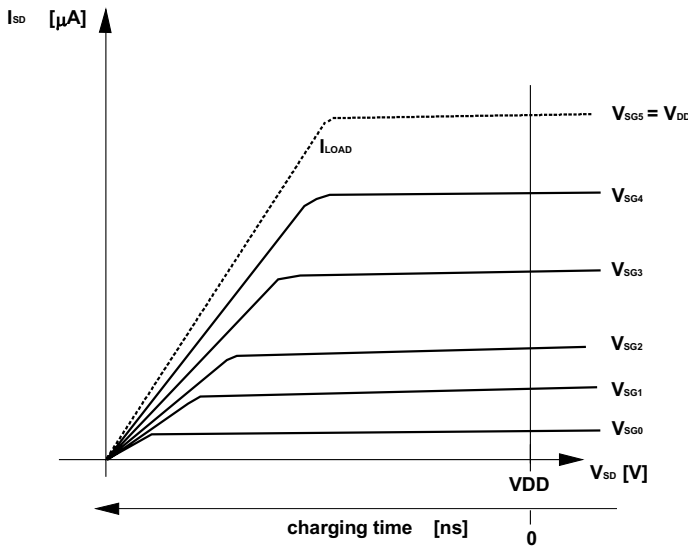


Fig. 3.7. Charging of a capacitor through a PMOS

Let's give another example, considering the circuit in Fig. 3.8, used as voltage reference with  $V_{OUT} < V_{IN}$ .

For M1, the biasing conditions are  $V_{GS} = V_{IN}$  and  $V_{DS} = V_{OUT}$ , while for M2,  $V_{GS} = (V_{IN} - V_{OUT})$  and  $V_{DS} = (VDD - V_{OUT})$ . M1 is in linear region, being:

$$V_{DS} = V_{OUT} < V_{IN} = V_{GS} \quad (3.19)$$

whereas M2 is in saturation

$$V_{DS} = VDD - V_{OUT}; V_{GS} = V_{IN} - V_{OUT}; VDD \geq V_{IN} \quad (3.20)$$

Recalling Eqs. (3.11) and (3.12) the currents that flow in M1 and M2 can be calculated.

$$I_{D2} = \frac{\beta_2}{2} \cdot (V_{IN} - V_{OUT} - V_{T,nat})^2 \quad (3.21)$$

$$I_{D1} = \beta_1 \cdot \left[ (V_{IN} - V_{T,nat}) \cdot V_{OUT} - \frac{1}{2} \cdot V_{OUT}^2 \right] \quad (3.22)$$

$$\beta = \mu \cdot C_{ox} \cdot \left( \frac{W}{L} \right) \quad (3.23)$$

Where  $V_{T,nat}$  is the threshold voltage of the natural transistor considered. Equating the two currents and assuming  $V_{T,nat} = 0$ , after some manipulations the following relation is obtained:

$$V_{OUT} = V_{IN} \cdot \left( 1 - \frac{1}{\sqrt{\rho + 1}} \right); \rho = \frac{\beta_2}{\beta_1} \quad (3.24)$$

Table 3.2 reports simulated and calculated values for  $V_{OUT}$  at different values of  $\rho$ .

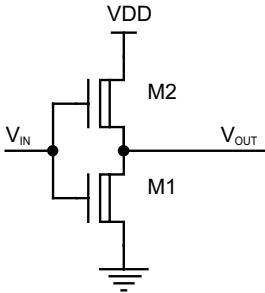


**Fig. 3.8.** Voltage source dependent on  $V_{IN}$  and on the size ratio between M1 and M2

**Problem 3.4:** Remake all the calculations of the theoretical  $V_{OUT}$  taking into account the body effect of M2.

**Table 3.2.** Simulated and calculated values of  $V_{OUT}$  as the size ratio of transistors M1 and M2 of Fig. 3.8 varies.

| W/L M2<br>[μm] | W/L M1<br>[μm] | $\rho$ | $V_{IN}$ [V] | $V_{OUT}^{theo-}$<br>retical [V] | $V_{OUT}^{simulated}$<br>[V] | VDD [V] |
|----------------|----------------|--------|--------------|----------------------------------|------------------------------|---------|
| 20/3           | 20/3           | 1      | 5            | 1.46                             | 1.3                          | 5       |
| 10/3           | 20/3           | 0.5    | 5            | 0.91                             | 0.8                          | 5       |
| 20/3           | 10/3           | 2      | 5            | 2.1                              | 2.1                          | 5       |
| 3/20           | 3/20           | 1      | 5            | 1.46                             | 1.1                          | 5       |
| 3/40           | 3/5            | 0.125  | 4            | 0.2                              | 0.1                          | 5       |
| 3/40           | 3/3            | 0.075  | 3            | 0.1                              | 0.1                          | 5       |
| 40/3           | 3/5            | 22     | 5            | 4                                | 4.1                          | 5       |

### 3.3 The Memory Cell

The memory cell we will deal with is the non-volatile cell, which is able to retain the information even without supply voltage. This is possible due to the insulated gate. Several versions of the non-volatile cell, based on the principle of the floating gate, exist. We will deal in detail with the cell named “T” because of the geometric shape, which constitutes the basic element of the array organization of the so-called NOR-type non-volatile memory. The picture or, better, the layout of the cell is shown in Fig. 3.9 with its basic features. It stores the single bit of information, the reading of several bits in parallel allows obtaining the bytes and the words.

The single cell structure is repeated to create the array that constitutes the bank of memory. Figure 3.10 shows an array of eight cells. As it can be noted, the contact is shared between two cells, which allows reducing the overall area by diminishing the number of contacts and, moreover, increasing the reliability of the device since the contact are critical in the fabrication process. The consequent parallelism of cells is often paid from the electrical point of view as we will see later on.

The cells that are on the same row of poly2 have also the same source junction, which is also shared with the cells of the following row. A source contact at intervals<sup>3</sup> guarantees the connection of the source to ground or to a fixed potential.

<sup>3</sup> Generally every 16 cells.

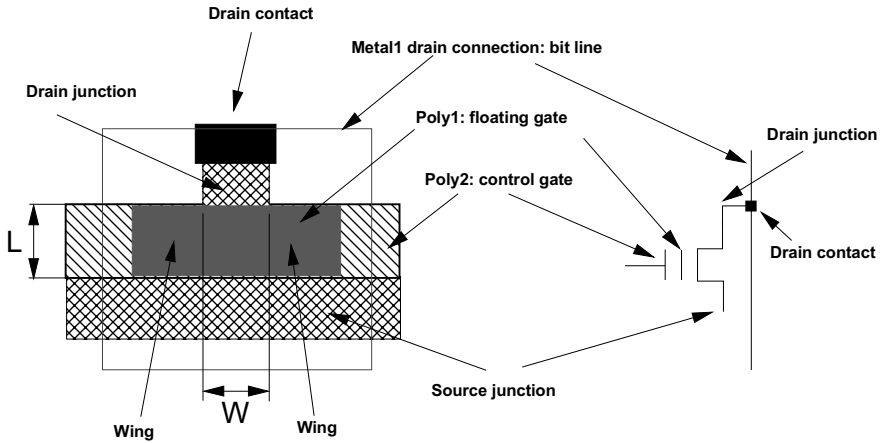


**Fig. 3.9.** Layout of the non-volatile cell and corresponding schematic. The size of L and W are presently less than one micron. The contacts are drawn as square even though they actually result in a round shape on silicon.

Particular attention is then paid to the side cells, the cells that constitutes the rows and columns placed at the edge of the array. As in any repetitive structure, breaking the symmetry at the end of the framework causes non-uniformity that results in memory cells that are different in terms of size and characteristics from those inside the array. Some dummy rows and columns are present but not electrically connected, with the function of reducing the edge effect for the used cells.

**Problem 3.5:** How should side rows and columns be fabricated in an EPROM cell array and in a Flash cell array?

The starting point for the analysis of a memory cell with floating gate is the equivalent capacitance one-dimensional model depicted in Fig. 3.11. In this model, merely electrostatic, the four electric terminals are coupled to the floating one by means of capacitors that can be derived assuming flat and parallel planes. Therefore, the capacitance depends on the thickness of the dielectric and on the area of overlap between floating gate and electrodes.

Considering that a certain charge  $Q$  is present on the floating gate, it is possible to use the relationship among capacitance, charge, and difference of potential at the sides of a capacitor to calculate the potential of the floating gate  $V_{FG}$  with reference to the external potentials:

$$Q = C_{FC} \cdot (V_{FG} - V_{CG}) + C_S \cdot (V_{FG} - V_S) + C_D \cdot (V_{FG} - V_D) + C_B \cdot (V_{FG} - V_B) \quad (3.25)$$

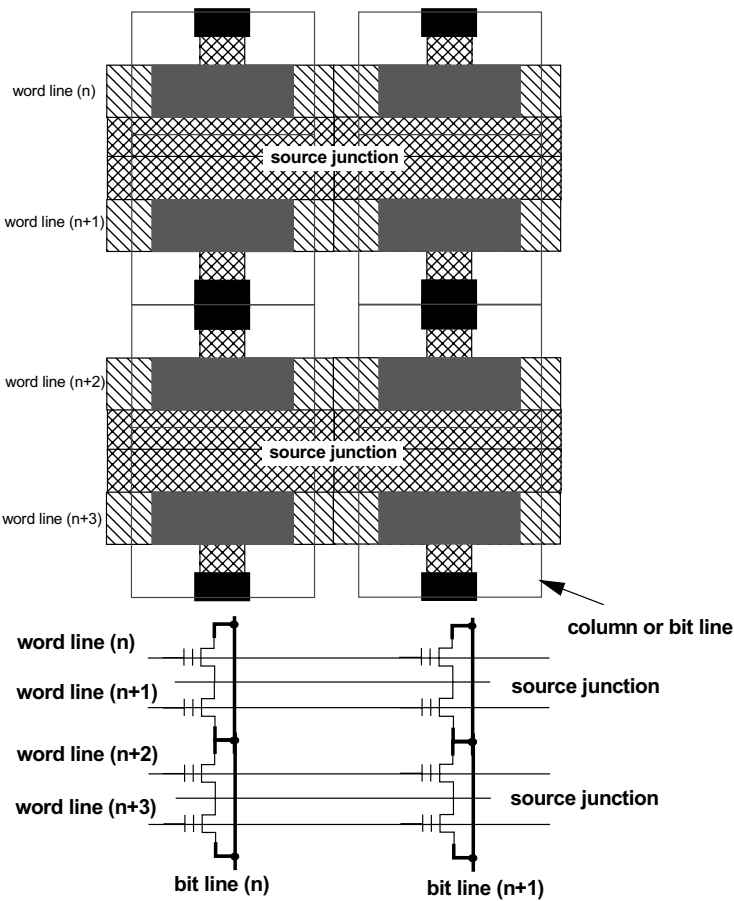


**Fig. 3.10.** Layout of eight cells connected so as to form an NOR-type array, with the source shared between two rows of cells and the drain contact shared between two single cells

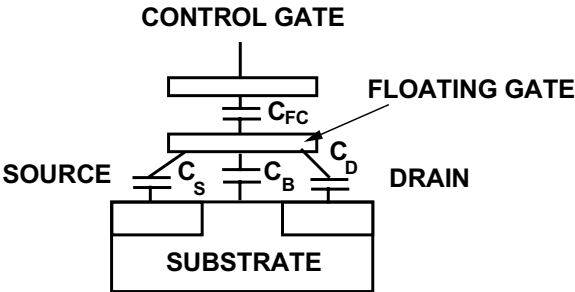


**Fig. 3.11.** One-dimensional model for floating gate transistor

At this point, we can use the following definitions:

$$C_T = C_{FC} + C_D + C_S + C_B \quad (3.26)$$

$$\alpha_S = \frac{C_S}{C_T}; \alpha_D = \frac{C_D}{C_T}; \alpha_B = \frac{C_B}{C_T}; \alpha_G = \frac{C_{FC}}{C_T} \quad (3.27)$$

to obtain:

$$V_{FG} = \alpha_S \cdot V_S + \alpha_D \cdot V_D + \alpha_B \cdot V_B + \alpha_G \cdot V_{CG} + \frac{Q}{C_T} \quad (3.28)$$

It is interesting to point out that the behavior of the cell is identical to the n-channel transistor having the gate terminal at  $V_{FG}$ , i.e. the polarization of the electrodes. This observation is the essential principle of operation of the Flash cell.

Let's analyze the behavior of the cell in the linear region. Substituting the voltage of the floating gate in Eq. (3.11), we get:

$$I_D = \beta^{FG} \cdot \left[ (V_{FG} - V_T^{FG}) \cdot V_D - \frac{V_D^2}{2} \right] \quad (3.29)$$

The superscript FG indicates quantities that refer to the floating gate. With the source and bulk at ground potential, it is possible to calculate the drain current of the cell with reference to the external potentials only. Substituting Eq. (3.28) into Eq. (3.29) we have:

$$I_D = \beta^{FG} \cdot \left[ \left( \alpha_D \cdot V_D + \alpha_G \cdot V_{CG} + \frac{Q}{C_T} - V_T^{FG} \right) \cdot V_D - \frac{V_D^2}{2} \right] \quad (3.30)$$

Now it is possible to define the parameters  $\beta$  and  $V_T$  with respect to the control gate, i.e.:

$$\beta^{CG} = \alpha_G \cdot \beta^{FG}; V_T^{CG} = \frac{1}{\alpha_G} \cdot \left( V_T^{FG} - \frac{Q}{C_T} \right) \quad (3.31)$$

In the case  $Q = 0$ , it is simple to give a circuit representation of the definition of the threshold voltage with respect to the control gate, by considering the capacitive divider formed by  $C_{FC}$  and  $C_T$  and imposing that the voltage of the floating gate equals the threshold voltage. It is now possible to rewrite the drain current of the cell in linear region only:

$$I_D = \beta^{CG} \cdot \left[ (V_{CG} - V_T^{CG}) \cdot V_D + \frac{1}{\alpha_G} \left( \alpha_D - \frac{1}{2} \right) \cdot V_D^2 \right] \quad (3.32)$$

**Problem 3.6:** Write the working equations of the floating gate cell in saturation region. Can the cell switch on also for gate voltage lower than  $V_T$ ?

### 3.4. Reading Characteristics

Due to the array organization, the cells that are on the same column share the bias contact of the drain, whereas the cells placed on the same row share the same gate contact, as depicted in Fig. 3.12<sup>4</sup>.

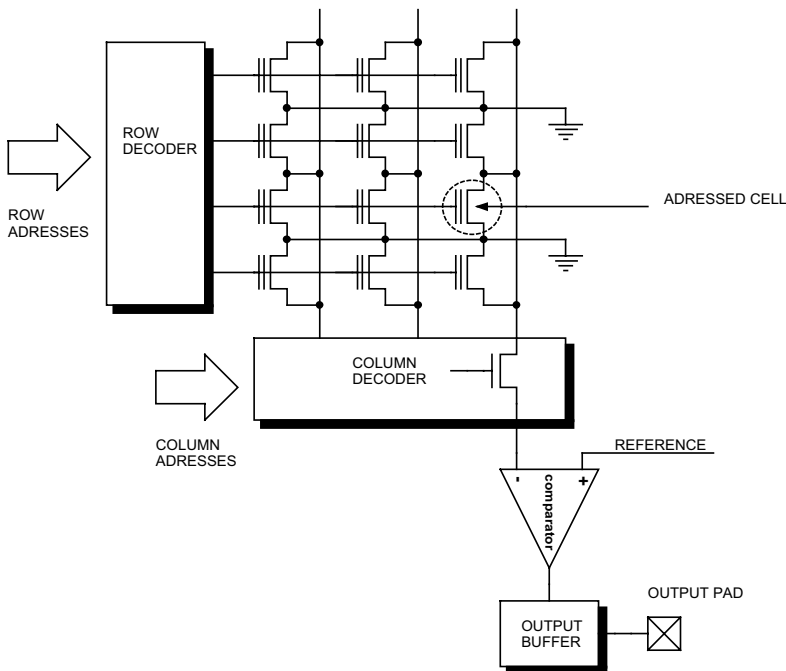


**Fig. 3.12.** Array organization: the cells are connected in the x-direction, i.e. the row, by the gate and in the y-direction, i.e. the column, by the drain contact. The decoding univocally allows addressing each single cell of the array.

Thus, in order to program the cell shown in the figure, the gate must be biased at about 12 V and the drain at about 5 V. Obviously, in such a condition, all the cells on the same row have the gate at 12 V and all those on the same column have the drain at 5 V. Therefore, a cell having gate at ground but belonging to the same column of the cell that is being programmed has the drain at 5 V and suffers the effect of drain stress that tends to erase it. In Fig. 3.13 the impact of such effect on programmed cells for a typical process is shown.

If the columns contain many cells and all of them are to be programmed, in the worst case the maximum programming time allowed, say 250  $\mu$ s, will be neces-

<sup>4</sup> The figure anticipates many issues, but what is important at this point is the organization of the array that allows many cells to be connected to the same drain contact.

sary for each cell. Assuming a column of 1024 cells, the time of drain stress  $t_{ds}$  applied to the first cell of the column equals:

$$t_{ds} = 1023 \cdot 250 \mu s \cong 0.256 s \quad (3.33)$$

This kind of stress should not lead to any problem but, in case the drain voltage were 5 V also during the reading phase there would be a loss of intrinsic charge due not only to the tunnel oxide retention but also to the electric field applied. The characteristic shown in Fig. 3.13 refers to a typical “good” cell, but the situation may be worse by several orders of magnitude for defective cells in terms of threshold shift.

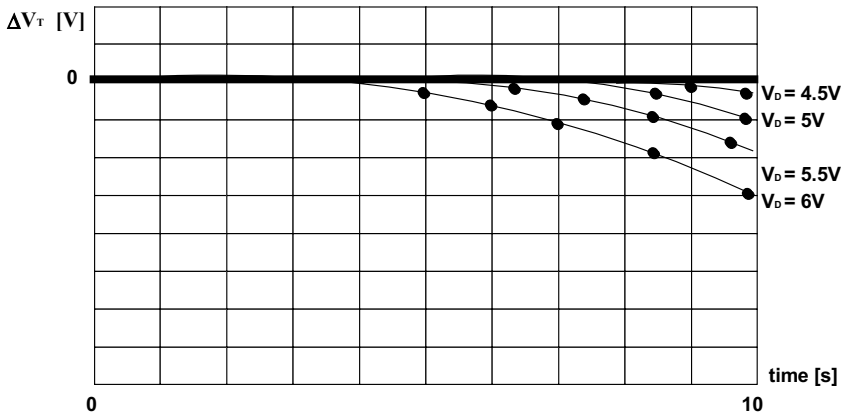


**Fig. 3.13.** Impact of the drain stress on a programmed cell. The deviation from the curve corresponding to zero volts indicates the percentage of charge loss. The family of curves has the drain voltage as parameter.

The effect of the drain stress can potentially occur even at low drain potentials. Thus, it is important to choose values of the drain potential that do not induce stress to the cells, which would lead to loss of information. In practice, the voltage of the drain terminal is driven to 1 V, to the maximum<sup>5</sup>. Considering that the voltage applied between control gate and source during the reading is generally greater than 4 V, this causes the cell to operate in the linear region. It is now possible to calculate the transconductance of the cell in linear region (with constant  $V_D$ ), basing on Eq. (3.32):

$$g_m = \frac{\partial I_{DS}}{\partial V_{CG}} = \beta^{CG} \cdot V_D \quad (3.34)$$

It is possible to observe that the cell transconductance is independent from the content of charge of the floating gate, resulting in the same value for erased or programmed cells.

<sup>5</sup> What is stated here will become more evident after studying the chapter that deals with the techniques to read the cells.

Let's define  $V_{T0}$  as the threshold of the cell with  $Q = 0$ . The step of voltage  $\Delta V_T$  with respect to the control gate, which differentiates the erased cell (logic "1") from the programmed cell (logic "0"), can be calculated starting from Eq. (3.31):

$$\Delta V_T = V_T^{CG} - V_{T0} = -\frac{Q}{C_{FC}} \quad (3.35)$$

Therefore, the drain current of a cell in linear region results to be:

$$I_D = \beta^{CG} \cdot \left[ (V_{CG} - V_{T0} - \Delta V_T) \cdot V_D + \frac{1}{\alpha_G} \cdot \left( \alpha_D - \frac{1}{2} \right) \cdot V_D^2 \right] \quad (3.36)$$

Once the drain voltage used during reading has been fixed, the characteristics of the "1" and "0" cells result to be parallel and separated by a fixed quantity equal to  $\Delta V_T$ . The diagram of the Fig. 3.14 reports the typical quantities for the input characteristics on the axes. The axis of abscissas can be regarded as both  $V_{GS}$  and bias voltage VDD, since it is suitable to drive the gate terminal to the maximum voltage available, which is generally the bias voltage.

For sake of simplicity, hereafter the superscript CG will be omitted and  $V_{CG}$  will be referred to as  $V_{GS}$  considering the source at ground.

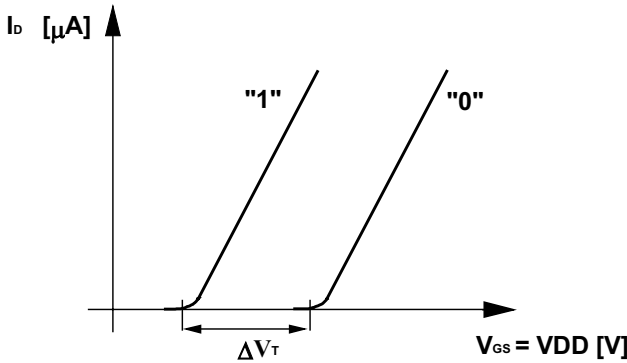


**Fig. 3.14.** The characteristics of programmed or erased cells result to be parallel during the read operation

### 3.5 Programming

The writing operation, or programming for a Flash cell is carried out by means of the so-called hot electrons that are able to overcome the energy barrier corresponding to the thin oxide between the drain area and the insulated gate. Thus, the two terminals of the cell, source and drain, are used in different ways during the two operations: during the erase the electrons flow from the insulated gate to the source, whereas, during the program, the electrons flow from drain to gate. Such specialization of the two terminals allows optimizing the device for the

specialization of the two terminals allows optimizing the device for the two different operations<sup>6</sup>. Let's now examine the program mechanism.

Let's consider an NMOS transistor (the Flash cell is an n-channel transistor) biased with a difference of potential between the gate and source terminals greater than its threshold voltage. In such a condition, minority carriers (electrons) are present all along the channel. If the drain junction is reverse-biased, some minority charges are drawn from the nearby region of channel and, thus, a reduction in the electron density along the channel from source to drain is determined. We define pinch-off voltage the  $V_{DS}$  bias for which complete depletion of the channel is achieved at the drain junction. If the  $V_{DS}$  voltage is further increased, the depleted region expands toward the source and its extremity is called pinch-off point. The presence of such a point defines the saturation condition for a transistor. In the specific case of the program operation, the cell is always in saturation region.

The longitudinal component of the electric field, always pointing from the source to the drain node, has a very intense gradient in the depleted region of the drain. Due to this unhomogeneity, the profile of the energy distribution of the electrons varies significantly along the direction parallel to the channel.

In general, conduction electrons tend to maintain thermal equilibrium with the lattice, yielding the energy that they acquire from the electric field through collisions with acoustic phonons, impurity, or other electrons. As a consequence, their average energy keeps around  $(3/2)KT$ , where  $K$  is the Boltzmann constant and  $T$  is the temperature of the lattice, while their speed varies linearly with the applied electric field. However, increasing the intensity of the field beyond a given limit, some electrons acquire more energy from the field than they can lose (no scatter mechanism is effective). In practice, these carriers are no longer in thermal equilibrium with the lattice and are called hot electrons.

In order to describe the kinetic of the hot electrons, the Fermi-Dirac energy distribution is introduced, similarly to what happens to the charge in thermal equilibrium, but with a specific associated temperature  $T_e > T$ . Silicon lattice, having atomic density of  $5 \cdot 10^{22}$  atoms/cm<sup>3</sup>, is characterized by electrons with covalent bonds (Fig. 3.15). Thermal energy allows some of these electrons to break the atomic bonds and, hence, they are free to move throughout the lattice, passing in conduction band. In a semiconductor, together with the electrons, also holes exists, i.e. absence of electrons, which can be represented as positive charge carriers, that move in the opposite way in the valence band. An important property of the electrons in the lattice is their distribution with respect to the allowed energy states, in the condition of thermal equilibrium. The Fermi-Dirac distribution function,  $f_{FD}$ , provides the probability that an energy status,  $E$ , is occupied by an electron:

$$f_{FD}(E) = \frac{1}{1 + e^{(E-E_F)/kT}} \quad (3.37)$$

where  $E_F$  is the Fermi energy (or level), corresponding to a 50% probability that an electron occupies the related energy status. In an intrinsic semiconductor, i.e. not doped, the number of carriers in the conduction band (electrons) equals the num-

<sup>6</sup> With the introduction of the array in the triple well, the erasing can take place along the entire channel length, making the specialization of the source no longer necessary.

ber of carriers in the valence band (holes), and the Fermi-Dirac function is symmetric with respect to that level. In a doped semiconductor, instead, the Fermi level shifts as a function of the type and quantity of doping present.

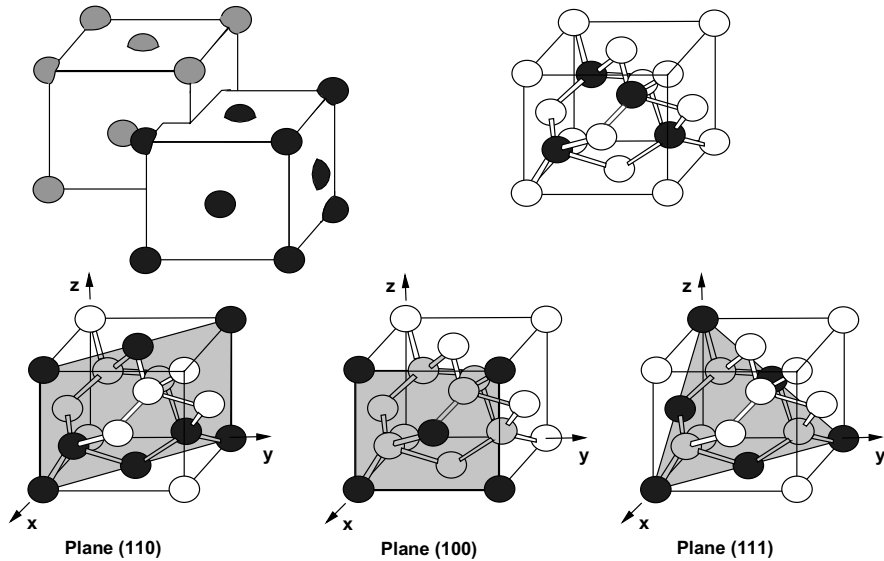


**Fig. 3.15.** Silicon crystallizes in a diamond structure, which is composed of two cubic lattices with centered sides, shifted with respect to each other by a quarter of diagonal. The different planes of the crystal and the respective Miller indexes are reported

The idea of attributing an electronic temperature,  $T_e$ , greater than the lattice one so as to characterize the condition of the hot electrons is not completely correct, since temperature is a concept typical of the thermodynamic equilibrium. However, it is useful to understand the drift of the carriers in presence of strong electric fields.

Hot electrons are responsible for a series of mechanisms that, in various ways, feed the channel, substrate, and gate currents. In Fig. 3.16, the effects that contribute to the gate current are highlighted. Let's now analyze in detail the phenomena of multiplication of the hot electrons in the channel.

Let's considering the phenomenon of the generation of electron-hole pairs due to impact ionization. Impact ionization is a process of coulombian interaction among electrons that activates in presence of high electric field. In practice, it happens that a very energetic electron, impacting against an electron that is in the valence band, yields enough energy for it to pass in the conduction band. At the end of the interaction, both the electrons are in the conduction band while a hole has been generated in the valence band. The energy threshold for which such a

process takes place is nearly equal to  $3/2$  of the silicon energy gap<sup>7</sup>. Considering that the gate current is around some nanoAmperes and the substrate current around microAmperes, it descends that, for the channel electrons, the ionization phenomenon is favored from the energy point of view with respect to their passage across the tunnel oxide. The negative charges,  $e_1$ , that originate due to primary multiplication, increase the population of electrons in the channel, phenomenon also referred to as C.H.E.I.A. (Channel Hot Electron Induced Avalanche). On the other hand, the ionization process provokes loss of energy. The electrons produced are re-accelerated by the longitudinal electric field and their probability of crossing the oxide voltage gap is concentrated in the final part of the channel. In the region of channel where the transverse electric field is direct toward the floating gate, there is a non-null probability that the holes produced by ionizing impacts may cross the oxide, providing a negative contribution to the gate current. It is necessary to bear in mind that they have to pass an energy barrier of 4.7 V, instead of the 3.2 V that is necessary for the electron injection and, furthermore, their effective mass is higher.

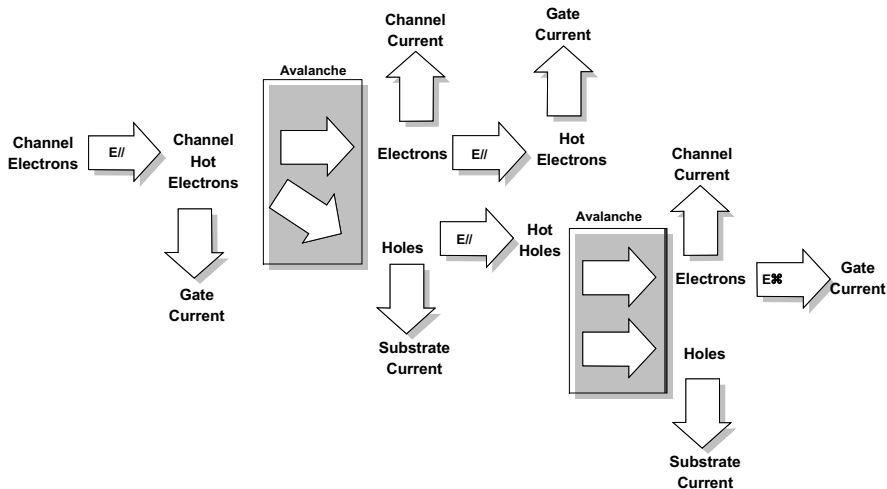


**Fig. 3.16.** Mechanism of charge injection onto the floating gate

As a consequence of the analyzed mechanisms, the amount of electrons (primary or generated by ionization) that are able to cross the tunnel oxide concentrates close to the drain junction. Recently, a physical mechanism known as C.H.I.S.E.L. (Channel Initiated Secondary Electron Injection) has been identified, which is able to produce an electron flow toward the gate, and whose peak of intensity is located far from the drain junction. The holes produced by the first impact ionization acquire enough energy from the electric field direct toward the substrate to start a new multiplication phenomenon. Due to the presence of a suit-

<sup>7</sup> This derives from the momentum and energy conservation equations, assuming that all the carriers have the same final speed and the same mass.

able electric field, the electron generated by this mechanism are heated and attracted by the channel where they can be injected onto the floating gate. Obviously, in the first phase of the programming, this contribution is negligible, since it depends upon the probability that multiplication phenomena occur. However, as the programming operation goes on, the transverse electric field strongly opposes against the electron injection and, thus, the injection efficiency due to the C.H.E. and C.H.E.I.A. loses the most consistent component. The flow of the electrons produced through C.H.I.S.E.L. is reduced by a lower amount, since it is located in a portion of channel closer to the source, where the direction of the electric field is favorable. In this situation, the contribution due to the secondary electrons becomes more important in the expression of the overall gate current. The phenomenon described above can be regulated by means of the substrate bias. Increasing the difference of potential applied to the drain-substrate junction, the intensity of the transverse field below the channel is also increased, which is the fundamental element to the effectiveness of the injection mechanism of secondary electrons.

In conclusion, the ionization phenomenon in the channel causes an increase in the amount of conduction electrons and in the decrease of their average energy.

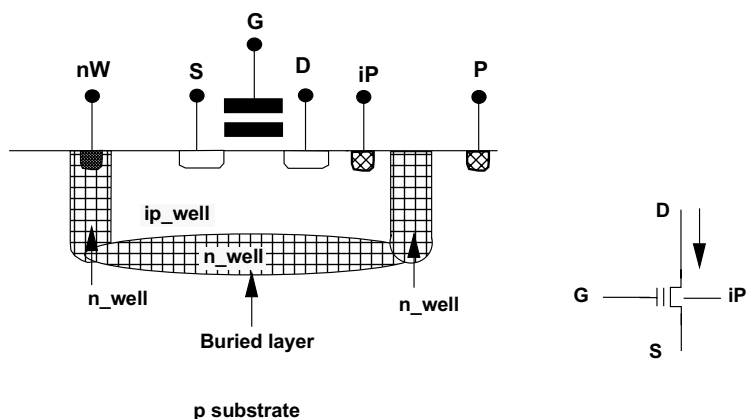
The tail of hot electrons is hence remarkably increased nearby the drain junction and, as a consequence, the charge flow toward the floating gate. During program, when the floating gate voltage reaches the imposed value,  $V_{ds}$ , the transverse electric field dramatically reduces the charge injection in the channel region close to the drain, causing a quick reduction in the gate current.

It is evident that the charge flow across the oxide tightly depends upon the profile of the energy distribution of the hot electrons. Despite the fact that it is very complicated to express the relationship with the voltage applied, it is evident that an increase in the drain voltage acts to increase the percentage of the electrons having energy greater than the oxide energy gap.

From the operating point of view, the minimum speed with which the device is due to accomplish the program operation is determined, and this also defines the voltage that must be applied. Supposing that the threshold voltage is increased by 3 V from an initial  $V_T$  of 2.5 V, the bias configurations that fulfill the requirements can be found. The choice will select the lowest voltages, thus minimizing power consumption and electric stress.

The flow of secondary electrons across the tunnel oxide induced by the negative bias of the substrate is one of the most popular approaches to reduce the voltages applied to the cell terminals. Of course, the technological process must account for the possibility of fabricating the array in triple-well as highlighted in Fig. 3.17. By reducing the voltage of the control gate and the drain, it is possible to reduce the intensity of the electric field across the tunnel oxide, with the consequent improvement in terms of gate and drain stress (see Chap. 20).

In practice, the negative substrate bias reduces the voltage of the pinch-off point, increasing the longitudinal component of the electric field. Moreover, owing to the effect on the threshold voltage (the Flash cell is an NMOS transistor), a decrease in the substrate voltage greatly reduces the overall drain current during program, with the consequent power saving.



| G   | D  | S  | iP  | nW | P  |
|-----|----|----|-----|----|----|
| 12V | 5V | 0V | -5V | 5V | 0V |

**Fig. 3.17.** The substrate of the array (ip-well) is insulated from the substrate of the rest of the circuitry by means of the n-well tub. In this way, it can be biased with a negative voltage during program to increase the writing effectiveness and diminish the drain current of the cell.

In Fig. 3.18 the characteristic of the threshold voltage step,  $\Delta V_T$ , is reported versus the programming time in the case of cell with grounded substrate: this is the so-called program curve. We recall that the threshold voltage of the cells is varied by the gate current by means of charge accumulation on the floating gate.

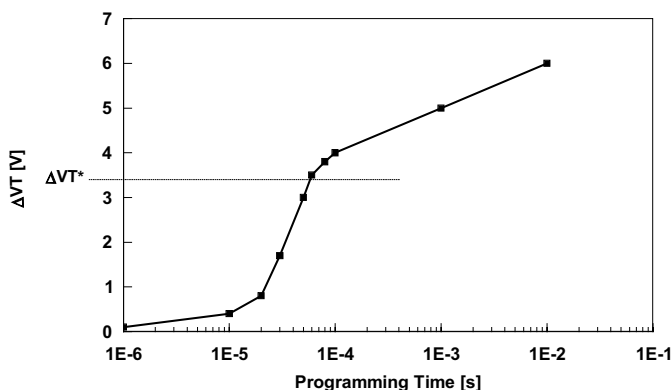


**Fig. 3.18.** Program curve of a Flash cell

In Fig. 3.18 two different zones can clearly be distinguished. The first in which there is a fast increase in  $V_T$  in a short time; the second in which there is a slower increase and a logarithmic function of the time. The point of separation between

the two zones corresponds to the condition in which  $V_{FG}$  equals  $V_{DS}$ , i.e. the voltage of the floating gate equals the voltage of the drain, and, on top of that, it corresponds to the threshold voltage step also known as  $\Delta V_T^*$ .

Let's suppose to apply a  $V_{GS1}$  gate voltage and a  $V_D$  drain voltage to an erased cell. The writing by means of hot electrons will continue until the voltage of the floating gate equals the drain voltage. If, on the contrary, the gate voltage applied is  $V_{GS2} > V_{GS1}$ , we obtain a higher initial gate voltage of the floating gate. As a consequence, in order to reach the  $\Delta V_T^*$  condition, we have to inject more electrons onto the floating gate, which results in a higher threshold voltage of the written cell.

In the case in which  $V_{FG} > V_D$ , the conditions favor the charge injection onto the floating gate, whereas if  $V_{FG} < V_D$ , the charge that has already been injected provokes a reduction in the potential of the floating gate below the value imposed by the drain. The part of the channel closer to the drain diffusion has higher voltage than the floating gate, which does not allow the majority of the hot electrons (generated in this zone of high longitudinal electric field) to cross the oxide<sup>8</sup>. Thus, the slope of the program curve in the zone where  $V_{FG} < V_D$  changes, and the programming speed quickly diminishes. In terms of gate current, we obtain that  $I_G$  is nearly constant when  $V_{FG} > V_D$ , whereas the gate current diminishes exponentially in the region where  $V_{FG} < V_D$ .

Substituting Eq. (3.35) in Eq. (3.28), the potential of the insulated gate can be expressed as:

$$V_{FG} = \alpha_G(V_{CG} - \Delta V_T) + \alpha_D V_D + \alpha_S V_S + \alpha_B V_B \quad (3.38)$$

If we want to calculate the value of  $\Delta V_T^*$ , i.e. the intrinsic voltage step, we have to insert the value of the potentials in the expression above, i.e.  $V_B = V_S = 0$ , and express  $\Delta V_T$  in the hypothesis that  $V_{FG} = V_D$

$$\Delta V_T^* = V_{CG} - \frac{1 - \alpha_D}{\alpha_G} V_D \quad (3.39)$$

The value of  $\Delta V_T^*$  is an important indicator of the programming speed<sup>9</sup>. If the minimum threshold voltage to regard the cell as programmed is greater than  $\Delta V_T^*$ , the cell will follow the knee of the program curve, going in a region where the gate current is very low, with a penalty in terms of programming speed. The problem of the maximization of  $\Delta V_T^*$  through the variation of the physical parameters that impact on the capacitive ratios poses. What we do is increasing the overlap area between control and floating gate by means of the so-called wings. As it can

<sup>8</sup> Near the beginning of the injection process, the inversion layer extends almost all the way to the drain, and the field in the oxide is attractive except for a small portion very near the drain. Current begins to flow through the oxide at the point where the electrons are the hottest and where the oxide field is most favorable. As the floating gate charges up, the floating gate-to-source voltage drops and the drain pinch-off region moves toward the source.

<sup>9</sup> In the cells of the latest generation, such a sharp distinction between the two programming zones no longer exists.

be seen in Fig. 3.9, while the capacitive coupling between control and floating gate increases because of the wings, the coupling with the substrate is nearly constant due to the high thickness of the field oxide underneath. In this way, the  $\alpha_G$  coefficient increases and, taking into account Eq. (3.39), also the value of  $\Delta V_T^*$ .

When operations of modification of the charge content of the floating gate are examined, it is necessary to account for the topological distribution of the cells within the array. For example, the resistive paths that connect the source of the single cells to the ground contact present resistance ranging from a few to several thousands Ohms, depending on the size and the kind of insulation used. This resistance worsens the programming characteristics and further widens the cell distribution. Finally, temperature variations and misalignments of the masks during fabrication worsen the problem.

### 3.6 Program Algorithm

The circuitry necessary to accomplish the programming is very complex. The advent of technologies with more and more reduced dimensions imposes precise control on the potentials to apply to the cells and to timings. In the case of devices with single supply, another problem for the designer is the cell current consumption during this phase. The control circuitry that governs the operation is realized in the same way also in the case of double supply devices. Major attention is paid to the way in which the program voltages are applied. First, the gate voltage must go high in order for the channel to form, and then the drain node must be pulled up with the program pulse. If the drain went up before the gate, unpleasant accidents might occur to all the cells connected to the column, such as spurious erase or program, *snap-back* phenomena (Fig. 3.19) and degradation of the cell performances. Thus, the control circuitry verifies the correct timing application of the signal, their duration and amplitude.

The instant of application of the drain pulse, with the gate already up, causes the cell to sink very high current, a part of which charges the parasitic capacitance of the bit line, whereas the other part crosses the channel producing the necessary hot electrons. This current peak can be estimated around 1 mA in the worst case.

Single supply devices cannot afford charge pumps that, in the hypothesis that we want to program an entire word, are able to source 16 mA during program with a 5 V output voltage. The first step is the adoption of the program on a byte basis instead of on a word basis, even though this increases the programming time<sup>10</sup>.

---

<sup>10</sup> The programming time directly impacts on the cost of the final system since a part of the memory is usually programmed by the manufacturer. The typical programming time ranges from 5 to 10  $\mu$ s to program a word.

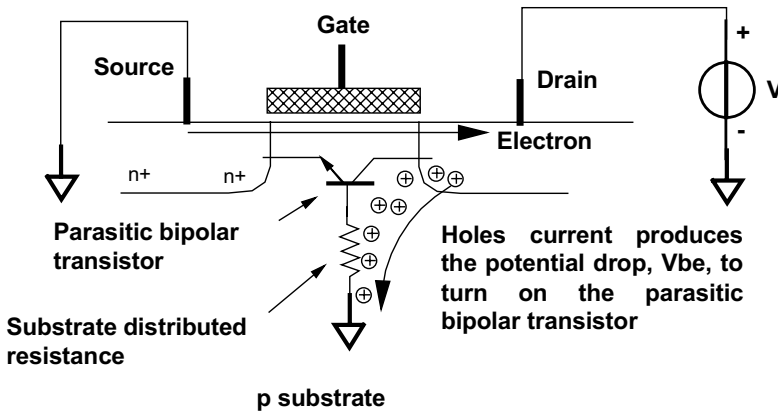


**Fig. 3.19.** Snap-back effect due to the switching on of the parasitic bipolar associated to each MOS. The potential applied to the drain may cause the breakdown of the junction, injecting holes into the substrate and pulling up the voltage of the base of the *npn* transistor indicated. When the bipolar switches on, the current does not flow along the surface, but is sunk into the ground contact of the substrate. Thus, the potential of the drain node decreases, which switches off the bipolar, and so forth. An oscillation is triggered on the drain node

The introduction of a control algorithm that distinguishes the two bytes that compose the word only in the case that the overall number of non-programmed bits is greater than eight may help. Even in this case, the problem of the power consumption is still present. In order to try to reduce its impact, new circuit configurations may be introduced. The first consist in programming by means of a ramp on the gate, as shown in Fig. 3.20.

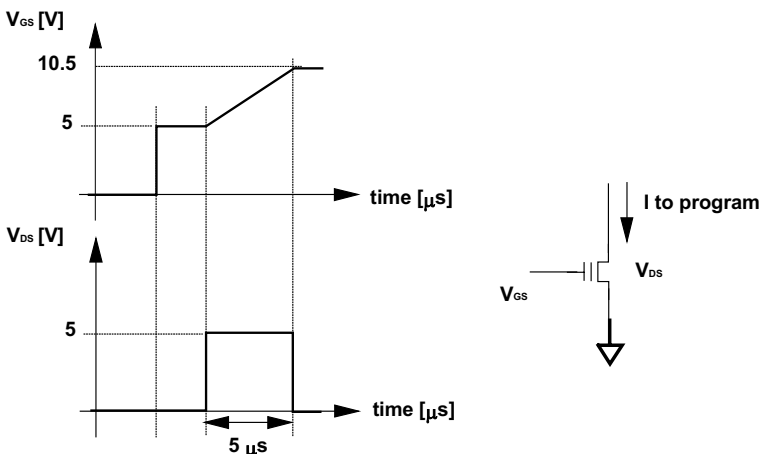


**Fig. 3.20.** Application of the program pulse during the gate ramp to limit the current consumption

During the first phase, the gate is biased at 5 V, which is enough to create the channel and, subsequently it is pulled to the final value of about 10 V by means of a ramp. During the application of the ramp, the drain pulse is applied that, in the case of the example shown, has a minimum duration of 5  $\mu\text{s}$ . In this way, the initial peak of current is reduced by acting on the value of the potential applied.

As already discussed, the placement of the array in triple-well allows biasing the cell substrate, improving the program effectiveness and, thus, reducing the value of the current sunk to about 100  $\mu\text{A}$  for each cell.

### 3.7 Erase Operation

Let's now examine the electrical erase operation in detail. Let's start by considering a gated diode, which is a MOS structure where only the diffusion area beneath the gate is active (Fig. 3.21). In our case, the gated diode represents a system composed of the isolated gate and the cell source junction, where the erase phenomenon takes place.

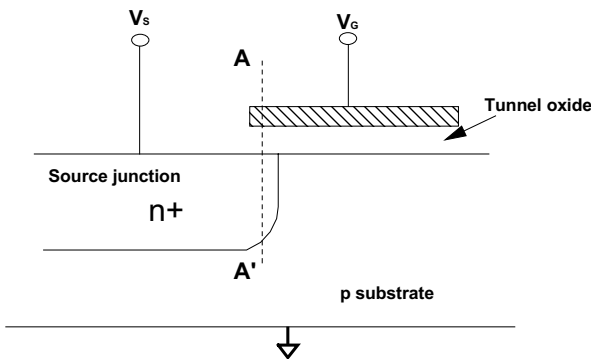


**Fig. 3.21.** Gate diode structure to study the erase operation

If we consider the cross section along the A-A' dashed line and draw the energy band diagram during erase, we obtain Fig. 3.22. The source has positive potential, the gate has negative potential, whereas the body is grounded. Moreover, the cell is programmed, i.e. the floating gate is charged with electrons. Such electrons have a finite probability of crossing the energy barrier related to the gate oxide, due to the Fowler-Nordheim (FN) tunneling. The gate current associated with the FN tunneling phenomenon can be expressed as:

$$I_G = A_{FN} \cdot E_{ox}^2 \cdot \exp\left(-\frac{B_{FN}}{E_{ox}}\right) \quad (3.40)$$

where  $E_{ox}$  is the electric field across the gate oxide (between source and floating gate), while  $A_{FN}$  and  $B_{FN}$  are constants.

Bands' bending is due to the applied potentials and implies that in the valence band there are electrons having the same energy as those in the conduction band (point A in Fig. 3.23).

The possibility of such an interband electron flow depends on the donor concentration in the silicon that is responsible for the band curvature. In the examined case, the  $n^+$  doping concentration of the cell source is indicated in Fig. 3.23.

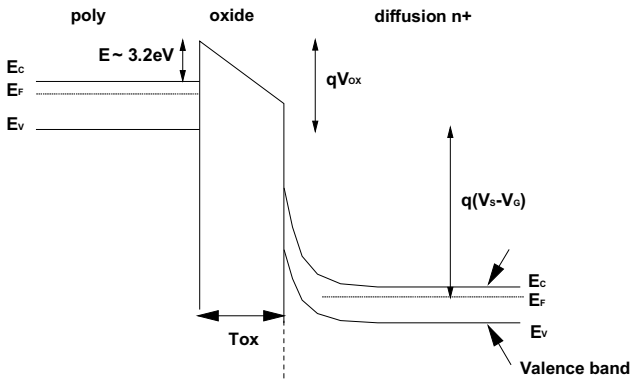


Fig. 3.22. Energy band structure with the applied erase potentials

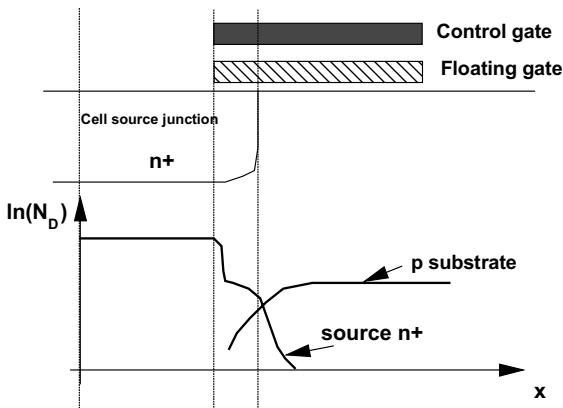


Fig. 3.23. Variation of the  $n^+$  doping concentration along the source-to-gate direction

The voltage drop in the silicon increases as the doping concentration decreases and, thus, an intense electric field is present in the oxide beneath the gate while, on the opposite side, at the source junction end, all the potential drop is located in the

silicon. In Fig. 3.24, the region where the interband tunneling phenomenon due to the electrons that leave holes in the valence band can take place is shown. Such holes flow through the depleted region of the substrate into the ground terminal and are responsible for the current called  $I_{BBT}$  (Band-to-Band Tunnel). The gate current is around some picoAmperes, while the BBT current is three orders of magnitude larger, i.e. around some nanoAmperes per cell. Also  $I_{BBT}$  can be expressed through an expression similar to the one used for the gate current:

$$I_{BBT} = I_S = A_{BB} \cdot E_{Si}^2 \cdot \exp\left(\frac{-B_{BB}}{E_{Si}}\right) \quad (3.41)$$

where  $A_{BB}$  and  $B_{BB}$  are constant and  $E_{Si}$  is the electric field at the silicon surface in the source region underneath the gate.  $E_{ox}$  and  $E_{Si}$  are bound together by the continuity of the displacement vector at the surface.

$$\epsilon_{ox} \cdot E_{ox} = \epsilon_{Si} \cdot E_{Si} \quad (3.42)$$

$$E_{ox} = 3 \cdot E_{Si} \quad (3.43)$$

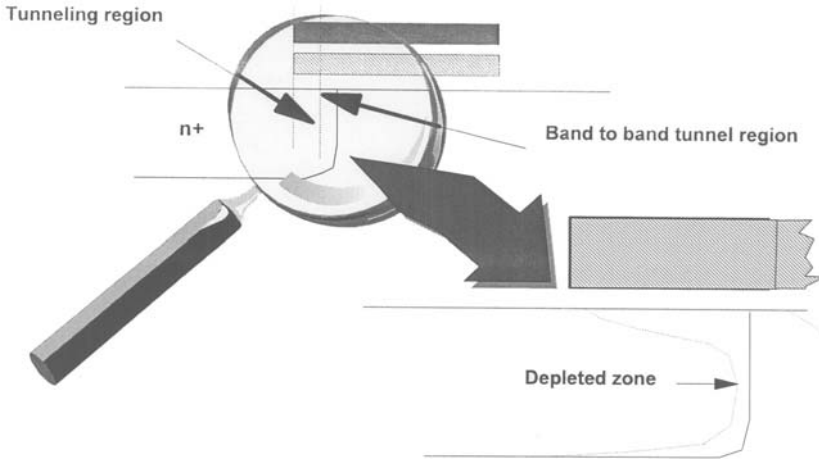


**Fig. 3.24.** The region underneath the gate where the tunneling across the oxide and the interband tunneling can take place

If we draw the logarithmic diagram of the BBT current as a function of the source voltage and regard the gate voltage as a parameter, we obtain Fig. 3.25. By substituting the difference of potential between the source and the gate nodes for the source voltage on the abscissa axis in Fig. 3.25, we obtain the diagram in Fig. 3.26 where the characteristics of the BBT current are overlapped.

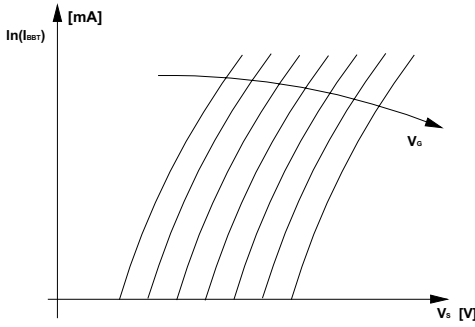


**Fig. 3.25.** Band-to-Band Tunnel current as a function of the source and isolated gate voltage

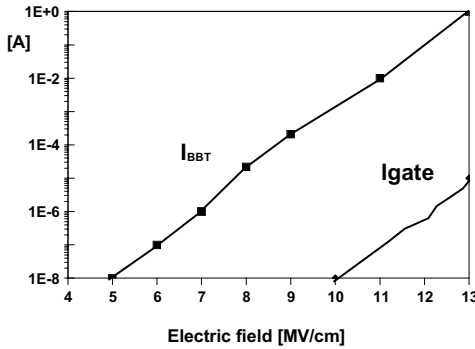


**Fig. 3.26.** BBT and gate current as a function of the difference of potential between the source node and the isolated gate

$V_s - V_G$  is the transverse component of the electric field and  $I_{BBT}$  does not depend on  $V_s$ . Therefore, the current due to the interband tunneling depends only on the transverse component of the electric field.

Let's now consider the memory cell taking into account Eq. (3.25) that associates the voltage of the isolated gate with the source and gate node potentials, the charge trapped, and the capacitive ratios. The drain does not have any impact (at least at a rough estimate) since it is left floating during the erase operation while the substrate is grounded.

$$V_{FG} = \alpha_S \cdot V_S + \alpha_G \cdot V_{CG} + \frac{Q}{C_T} \quad (3.44)$$

If we neglect the voltage drop in the floating gate and in the silicon, the electric field across the tunnel oxide can be calculated as follows:

$$E_{ox} = \frac{|V_{FG} - V_S|}{t_{ox}} = \frac{|(\alpha_S - 1) \cdot V_S + \alpha_G \cdot V_{CG} + \frac{Q}{C_T}|}{t_{ox}} \quad (3.45)$$

### 3.7.1 Erasing at Constant Voltage

In this case the erase operation is carried out by leaving the drain floating, whereas the gate is grounded and the source is pulled up to 10–12 V. The necessary electric field is therefore obtained by controlling only one of the cell terminals. This mode is implemented in flash memories that, besides VDD, have a second bias voltage dedicated to program and erase operations, named VPP. At the initial time, the electric field across the thin oxide depends only upon the charge stored on the floating gate,  $Q$ . The erase operation is carried out by discharging the isolated gate and, at the same time, the electric field and the current diminish in an exponential fashion with the electric field.

The problem is that the initial peak of the electric field, caused by the application of the source voltage, may provoke the breakdown of the source junction. The solution that prevents the junction damage and the cell degradation consists in applying the erase voltage to the source by means of a voltage ramp instead of applying it directly. A simple alternative is the insertion of a resistor,  $R$ , to limit the voltage (Fig. 3.27).

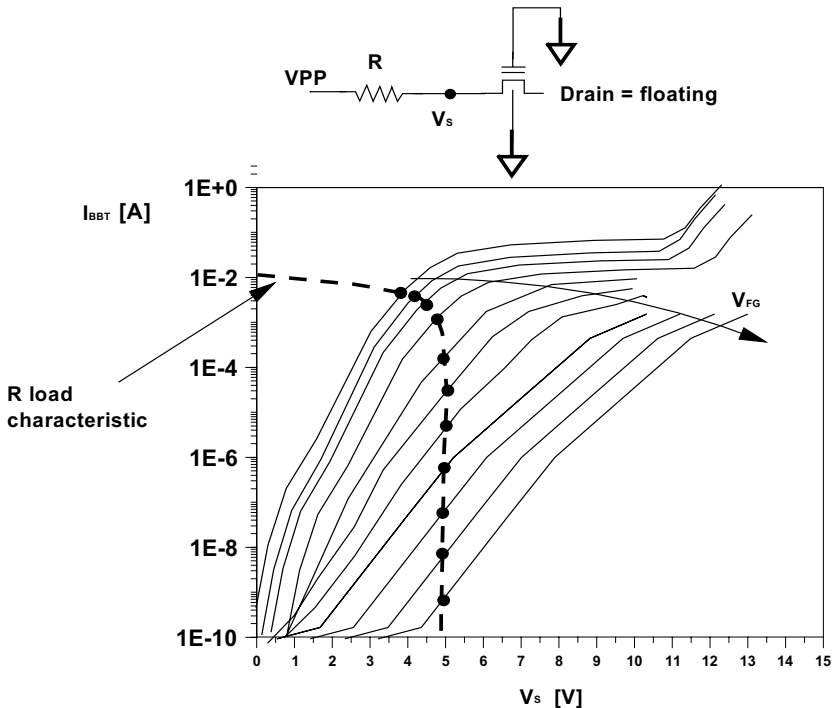


**Fig. 3.27.** Circuit configuration and voltage of the source node in the case of constant erase voltage

The load characteristic of the resistor is overlapped to the  $I_{\text{BBT}}$  characteristics. With the resistor we do not erase through a ramp since the horizontal part of the characteristic, which becomes vertical later, is covered in a short time. Erase can be performed because the voltage drop on the resistor is small and allows obtaining the necessary electric field due to the fact that the gate current is much less than  $I_{\text{BBT}}$ .

The resistor value is determined so as to pull the source potential to a value definitely smaller than the breakdown potential at the beginning of the erase phase, when the BBT current is maximum.

As for program, we can define the erase characteristic as the curve that expresses the threshold voltage shift as a function of time. Figure 3.28 shows as the time that is necessary to reach the  $V_{T,\text{end}}$  threshold voltage is independent of the  $V_T$  voltage of the programmed cells at the beginning of the erase phase. Of course, this is true for a given tunnel oxide thickness.

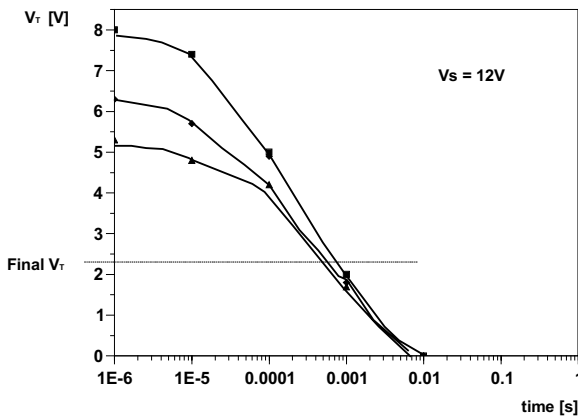


**Fig. 3.28.** Threshold voltage as a function of time

In order to better understand this behavior, let's erase two cells, geometrically and electrically identical, that have different initial charge on the floating gate (Fig. 3.29), i.e. different  $V_T$ . During erase we suppose that the voltage applied to the terminals of the two cells is the same, which means that the two control gates are equipotential as well as the two sources since the local currents are so small that they cannot modify the voltage drop.

At the beginning, the cell having more charge and, therefore, greater threshold voltage, has a more intense electric field and is erased more quickly. The faster cell reaches the amount of charge of the slower cell and, from that moment on, they are erased together since they undergo the same electric field.

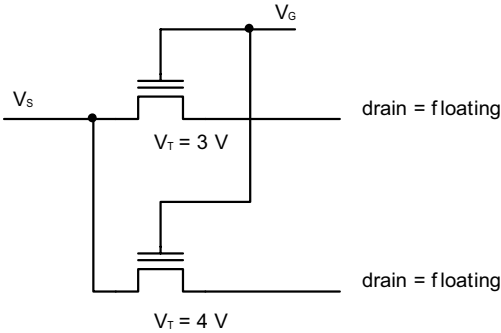


**Fig. 3.29.** Two identical cells with different initial amount of charge are erased simultaneously since they belong to the same sector. The slower cell waits for the faster cell so that they have the same electric field applied

A clarification is now necessary. The tunneling phenomenon that causes the erasing is located in a very narrow region, at least in the channel direction. The polysilicon grains that make up the isolated gate have the same size of the erase region. Besides the geometrical differences, also the possible charge trapped, due to the various process steps that eventually modifies the band structure at the interface<sup>11</sup>, must be taken into account. All the variations can be regarded as an equivalent variation of the oxide thickness. It is important to notice that the shape of the threshold voltage distribution of the erased cells does not depend on the shape of the corresponding distribution of the programmed cells (before erase). It is also possible to regard such a distribution as the overlap of a gaussian distribution and a poissonian tail.

The adoption of the mixed erase instead of the source erase increases the breakdown margin that, we recall, may be a problem only if it involves few cells that undergo source junction degradation caused by the high current flow.

In single bias source devices, high voltages are internally generated by means of charge pumps (see Chap. 15). The limited availability of current sourced by such circuits has led to the negative gate erase.

Taking into account Eq. (3.45) and supposing that:

$$\alpha_S = 0.15; \Delta V_T = 3V; T_{ox} = 10nm; E_{ox} = 10MV/cm \quad (3.46)$$

we obtain

$$V_G = -8V; V_S = 5V \quad (3.47)$$

Notice that, in this case, also the gate capacitive coupling takes place. The advantage in terms of reduction of the voltage applied to the cell source is evident.

<sup>11</sup> The trapped charge is responsible for the phenomenon known as “erratic bit”. During the write/erase cycles, some cells can be found whose  $V_T$  might be negative after the  $n$ -th erase cycle, and might become positive after  $m$  more cycles and eventually negative again after  $n+m$  further cycles.

### 3.7.2 Constant Current Erase

The main drawback of the standard erase method with constant voltages applied is the strong dependence of the erase time on  $V_s$  (typically  $V_{PP}$  or  $V_{DD}$ ) and temperature. Moreover, this approach implies a complete dependence of the electric field on the cell process variation, giving rise to a variable peak of both source and gate current, occurring at the beginning of the erase operation.

On the contrary, the constant current erase method consists in keeping the gate current constant during the whole erase operation. This is obtained by forcing a constant current in the source node in such a way to maintain the electric field always constant. It is based on the observation that, given a certain electric field, the ratio between gate and source current is constant. In fact, as clearly shown in Fig. 3.30, given a determined  $I_s$  and considering Eq. (3.40) and Eq. (3.41),  $E_{Si}$  is fixed and also  $I_G$  is known. The  $I_G/I_s$  ratio is independent of the bias conditions and depends only on the electric field. Moreover, since the gate current is constant, the threshold voltage shift of the cell, proportional to the integral of  $I_G$ , follows a linear time law, whose slope is a function of the chosen source current. The value of the source current can be determined according to the defined erase time and maximum electrical field, chosen according to reliability considerations.

In fact, the charge trapping/generation in the oxide is a strong function of the charge flowing through it and the electric field. The higher the electric field, the greater the oxide degradation. Hence, it is possible to control the erase time degradation, which depends on the oxide damage, by controlling the maximum electric field.

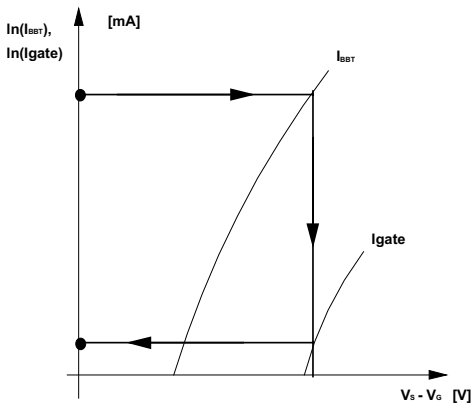


Fig. 3.30. Erase speed defined through the design of the BBT and gate currents

Therefore, using this kind of approach, the threshold voltage variation is constant with time, differently from all the other cases in which the variation is fast at the beginning and slows down progressively. On the other hand, what happens if

$\Delta V_T$  is constant and two cells have different amount of charge? Is the distribution of the erased cells similar to the distribution of the programmed cells?

The answer is negative since the cells tend to behave uniformly, i.e. those that have less charge “wait” for those that are erased more quickly because they have more charge at the beginning. The choice of the current determines the electric field so that the erase speed is the same also in the case of different devices. The source voltage will be tuned in order for the electric field to fulfill the design value.

The main drawback is the precision with which erase is stopped. In the previous cases, the voltage step obtained though a constant pulse at the end of the erasing diminishes and, hence, the precision of the final value of the threshold voltage increases. In this case, instead, the variation of the threshold voltage is the same during all the pulses and, thus, the error with respect to the desired threshold voltage is larger.

3.7.3 Erasing at Negative Gate and Triple-Well Array

The fabrication of the cell array in triple-well (Fig. 3.31) allows biasing the substrate with positive voltage, eliminating the voltage drop responsible for the curvature of the energy bands that causes the spurious band-to-band tunneling current. This result has a fundamental importance for the present memories that work with a single bias voltage that is smaller than the program and erase voltages. The high voltages are generated by means of charge pumps (see Chap. 15) that typically have a limited capability to source current. Increasing the available current means increasing the size of the capacitors of the pumps, with evident repercussions on the overall device area.

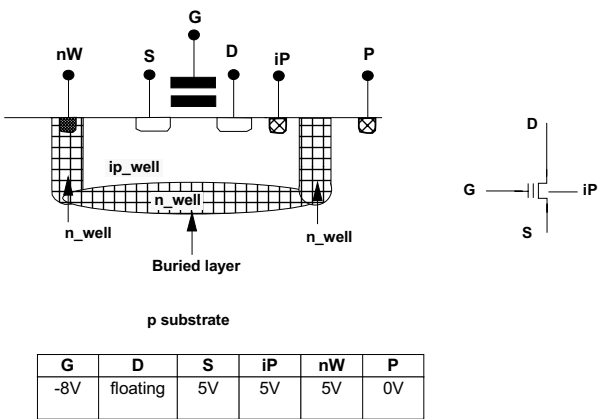


Fig. 3.31. Memory array in triple-well. The array substrate (ip-well) is isolated from the rest of the circuits by an n-well tub

### 3.8 Erase Algorithm

After recognizing the code that the user communicates to the CUI<sup>12</sup>, i.e. the internal memory controller, to erase a sector, the preconditioning operation starts, that is all the cells are programmed. Such a preliminary operation is executed to guarantee uniform aging of the cell population so as to limit the width of the erase distribution. If we erased a sector containing cells with low  $V_T$ , such cells would be over-erased and probable depleted to negative  $V_T$  while decreasing the  $V_T$  of the programmed cells. Figure 3.32 illustrates the failure mode induced by a depleted cell.

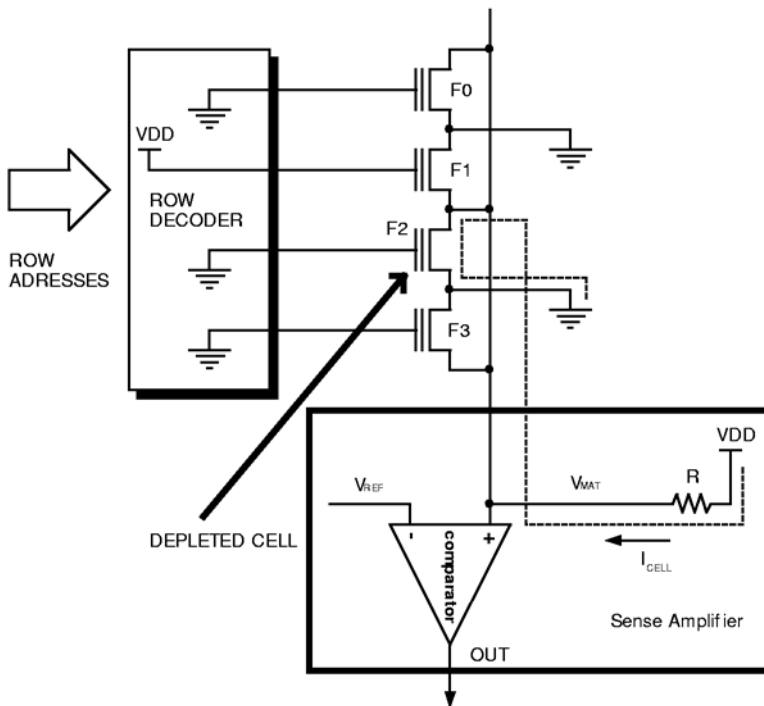


**Fig. 3.32.** Effect of a depleted cell on the reading of a written cell belonging to the same column. The current sunk by the depleted cell simulates, on the sense amplifier point of view, the reading of an erased cell

The row decoder selects the written cell, F1, based on the address applied by the user. This cell does not sink current even when  $V_{GS}$  equals  $V_{DD}$ . The sense amplifier (see Chap. 12) recognizes the cell as written since the potential of node MAT equals  $V_{DD}$ . The presence of a depleted cell, F2, with threshold voltage

<sup>12</sup> It is the acronym for Command User Interface. It is a finite state machine used to decode the commands the user applies to the memory.

equal to  $-2\text{ V}$ , determines a current flow along the bit line. The potential of node MAT decreases and the reading is not correct.

Let's go back to the erase algorithm. At this point, we have a group of programmed cells with gaussian distribution of  $V_T$ . During the actual electrical erase phase, the source and gate voltages of the cells are forced to the proper values with respect to the adopted erase technique (see previous section). The time counter limits the erase pulse to the typical duration of 10 ms. At the end of the pulse, the erase verify phase is carried out. Such an operation is performed through a marginated read that guarantees the correct cell recognition in normal read mode.

The erase pulses continue until the correct verification of all the cells of the sector has been completed. At this point it is necessary to verify that there are no depleted cells that may induce errors during read. It is difficult to detect the depleted cells since the presence of a cell with negative  $V_T$  implies that the reading of all the other cells belonging to the same columns is dominated by the current of the depleted one. Therefore, a search algorithm is applied to verify the presence of leakage current on the columns when all the rows are grounded. When a column with such an anomaly is found, the first cell is addressed and a program pulse is applied with low gate voltage, so as to slightly increase its  $V_T$ , without overcoming the limit used during the erase verify phase. Subsequently, the same algorithm is applied to the second cell of the same column. If no current is drawn, it means that the depleted cell was the previous one that was already recovered with the soft programming pulse; otherwise a pulse is applied to the present cell and so forth, until the end of the column. At the end of the column, the verify phase is repeated and, in the case of presence of current, the procedure is repeated with increased gate voltage, since it means that the program pulse applied to all the cells of the column was not sufficient to recover the depleted cells.

In order to simplify the algorithm controlled by the internal finite state machine, the gate voltage is used as a parameter, whereas the drain voltage is kept to the value used during the normal program phase and the same holds for the pulse duration. As we will see in Chap. 14, the gate voltage control limits the maximum threshold voltage step for a fixed time duration of the pulse.

The erase procedure has also an important feature called *erase suspend* that allows the user to suspend the erase operation for an indefinite time, storing the current state of the procedure, so as to be able to restart after a proper command called erase resume. Such a feature allows the external microprocessor to access a sector of the array to retrieve data, apart from the one that is undergoing erase. In this way, the penalization due to the long erase duration (around 1 s) can be partially overcome.

## Bibliography

- S. Aritome, R. Shirota, G. Hemink, T. Endoh, and F. Masoka, "Reliability issues of Flash memory cells," Proc. IEEE, vol. 81, no. 5, pp. 776-788, (May 1993).
- R. Bez et al., "Depletion Mechanism of Flash cell induced by parasitic drain stress contidion", VLSI Technology Symposium, (1994).

- J. D. Bude, "Gate current by impact ionization feedback in sub-micron MOSFET technologies", in 1995 Symposium VLSI Technology Dig. Tech. Pap., pp. 101–102, (June 1995).
- J. D. Bude, M.R. Pinto and R. K. Smith, "Monte Carlo Simulation of the CHISEL Flash Memory Cell," IEEE Tran. Electron Devices, vol. 47, pp. 1873–1881, (Oct. 2000).
- E. Burstein, S. Lundqvist, "Tunneling Phenomena in Solida", Plenum Press, New-York, (1969).
- E. Camerlenghi, P. Caprara, and G. Crisenza: "A 18  $\mu\text{m}^2$  cell for megabit CMOS EPROM", in Proc. 17th European Solid State Device Research Conf., pp. 765–768, (Sept. 1987).
- John Y. Chen, CMOS devices and technology for VLSI, Prentice Hall, (1990).
- A. Chimenton, P. Pellati, and P. Olivo, "Constant Charge Erasing Scheme for Flash Memories," IEEE Tran. Electron Devices, vol. 49, pp. 613–618, (Apr. 2002).
- A. Chimenton, et al., "Overerase Phenomena: An insight Into Flash Memory reliability", IEEE Proceeding of the, Vol. 91, No. 4, pp. 617–626, (April 2003).
- C. Dunn, C. Kaya, T. Lewis, T. Strauss, J. Schreck, P. Hefley, M. Middendorf, and T. San, "Flash EEPROM disturb mechanisms," in Proc. Int. Rel. Phys. Symp., pp. 299–308, (April 1994).
- B. Eitan and D. Frohman-Bentchkowski, "Hot electron injection into the oxide in n-channel MOS devices", IEEE Trans. Electron Devices, vol. ED-28, pp. 328–340, (March 1981).
- B. Eitan, R. Kazerounian, A. Roy, G. Crisenza, P. Cappelletti, and A. Modelli, "Multilevel Flash cells and their trade-offs", in 1996 IEDM Tech. Dig., pp. 169–172, (Dec. 1996).
- Leo Esaki, "Long Journey into Tunneling", Proceedings of IEEE, vol 62, No 6, pp 825–835, (June 1974).
- D. Frohman-Bentchkowski, "Memory behavior in a floating gate avalanche-injection MOS (FAMOS) structure", Appl. Phys. Lett., vol. 18, pp. 332–334, (1971).
- D. Frohman-Bentchkowski, "FAMOS-A new semiconductor charge storage device", Solid State Electron, vol. 17, pp. 517–520, (1974).
- C. Hu, "Lucky-electron model for channel hot-electron emission", 1979 IEDM Tech. Dig., pp. 22–25, (Dec. 1979).
- C. Hu, "Future CMOS scaling and reliability", Proc. IEEE, vol. 81, pp. 682–689, (May 1993).
- Y. Igura et al., "New Device Degradation Due to "Cold" Carriers Created by Band-to-Band Tunneling", IEEE Electro Device Letters, VOL. 10, NO. 5, MAY (1989).
- C. Kittel, Introduction to Solid State Physic, John Wiley & Sons, New York, (1966).
- M. Lenzlinger and E. H. Snow, "Fowler-Nordheim tunneling into thermally grown  $\text{SiO}_2$ ", J. of Applied Physics, vol. 40, pp. 273–283, (Jan. 1969).
- S. Mahapatra, S. Shukuri, and J. Bude, "CHISEL flash EEPROM—Part I: performance and scaling", IEEE Trans. Electron Devices, vol. ED-49, pp. 1296–1301, (July 2002).
- S. Mahapatra, S. Shukuri, and J. Bude, "CHISEL flash EEPROM—Part I: reliability", IEEE Trans. Electron Devices, vol. ED-49, pp. 1302–1307, (July 2002).
- Yohsuka Mochizucki, "Read-disturb Failure in Flash Memory at low field", Intel reports, Nikkei Electronics Asia, pp. 35–36, (May 1993).
- J. Van Houdt, et al., "The HIMOS Flash technology: The alternative solution for low-cost embedded Mmemory", IEEE Proceeding of the, Vol. 91, No. 4, pp. 627–635, (April 2003).
- Samuel Tuan Wang, "On the I-V characteristics of Floating-Gate Mos transistors", IEEE Transaction on electron devices, Vol ED-26, No 9, September (1979).

## 5 Fundamental Circuit Blocks

### 5.1 Introduction

The first step towards design consists of the study of the most used circuit blocks, from the inverter to boost concepts. A deep analysis of these circuits allows their assembly to get to the desired results. The main circuit blocks are shown always paying attention to the understanding of their functionality. The suggestion is to re-design every circuit to understand the connections between the various elements. If you have a simulator, don't try to understand the behavior of the circuit by simulating it; analyze it on paper and then verify your understanding with simulation.

### 5.2 NMOS and CMOS Inverters

The main structure for logic design is the inverter, which is also the fundamental block to understand the operation of both analog and logic circuits. The easiest usage of a transistor is shown in Fig. 5.1.

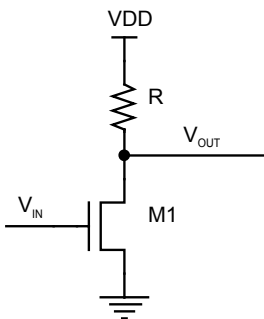


**Fig. 5.1.** Inverter with a resistive load

We know that by modulating the voltage of the input signal we obtain a voltage swing of the output node that can be much bigger than the one on the input. We call this phenomenon “amplification”. An amplifier is a circuit that exploits an impedance variation of a device, M1 in this case, to vary the resistive ratio on the

output. By modulating the channel resistance of M1 we “decide” how much current it can sink and consequently the voltage drop on the load R.

If we consider only the extremes of the range of the input signal  $V_{IN}$ , VDD and ground, we obtain a logic inverter. When  $V_{IN}$  is GND, LVS-type transistor M1 is turned off: no current flows and the voltage of the output  $V_{OUT}$  is equal to the supply voltage VDD. On the other hand when  $V_{IN}$  is VDD, M1 transistor is turned on and  $V_{OUT}$  is near to ground. We cannot say that  $V_{OUT}$  is GND, in this case, because it depends on the partition ratio between the load R and the equivalent resistance of M1.

Output characteristic of M1 transistor are shown in Fig. 5.2. Let's now consider the load line: the value of the short-circuit current ( $V_{OUT} = 0$  V) is equal to  $I_{DD} = VDD/R$  while the open-circuit voltage ( $I_{DS} = 0$ ) is  $V_{OUT} = VDD$ . The possible working points for our circuit are given by the intersections of the characteristics of M1 with the load line. We can see for instance that the minimum value of  $V_{OUT}$  is not ground, but  $V_{DSm}$ , under the assumption that  $V_{GS}$  is equal to VDD.

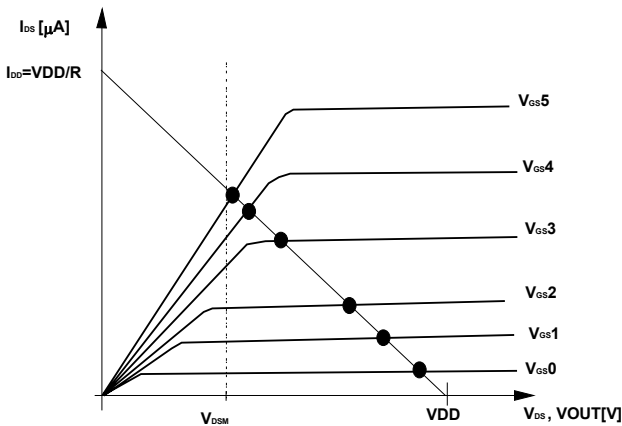


**Fig. 5.2.** Load line where some working points are shown ( $V_{GS0} < V_{GS1} < V_{GS2} < V_{GS3} < V_{GS4} < V_{GS5}$ )

**Problem 5.1:** An interesting exercise to carry out is the calculation of the value to assign to resistance R to be able to satisfy “reasonable” design criteria, for instance charging 1 pF in 2 ns with an associated consumption of few microAmpere. The discussion of the problem shows how inefficient is the use of a resistor to implement an integrated inverter, while it is normal to do it on a board.

The oldest integrated technology, NMOS<sup>1</sup>, has solved this issue by replacing the passive load with an active one. A depletion NMOS transistor, i.e. a MOS whose channel is always present even if  $V_{GS}$  is zero (thanks to the negative threshold voltage) can be used as active load. Figure 5.3 shows the scheme of a NMOS

<sup>1</sup> This technology allows the usage of n-channel MOS transistors only.

inverter where the M2 transistor is diode-connected and has replaced resistor R of Fig. 5.1.

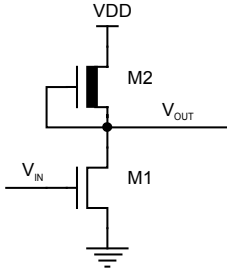


**Fig. 5.3.** NMOS Inverter

The first remark to do is the comparison between the load characteristics of R and M2 (Fig. 5.4). Characteristic of M2 is not linear: for high  $V_{GS}$  values the current sunk by M1 is high and M2 has  $V_{GS} = 0$  V (above  $V_{T,dep}$ ), well turned on and able to provide the required current. The important consequence is that the  $V_{GS}$  of M2 is always zero, while its  $V_{DS}$  is  $VDD - V_{OUT}$ .

If the threshold of M2 is  $V_{T,DEP} = -3.5$  V, then:

- μ  $V_{DS} > V_{GS} - V_{T,DEP}$  is the saturation condition
- μ  $V_{DS} < V_{GS} - V_{T,DEP}$  is the linear zone condition.

Assuming a supply voltage of 5 V and being  $V_{GS} = 0$  and  $V_{DS} = VDD - V_{OUT}$  saturation zone extends between  $V_{OUT} = 0$  V to  $V_{OUT} = 1.5$  V.

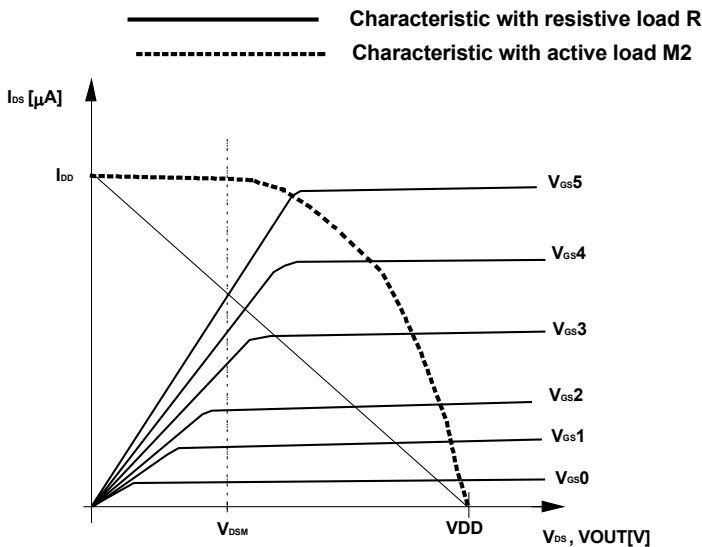


**Fig. 5.4.** Comparison between the characteristic of active and passive load

The advantage with respect to the behavior of the passive resistive load  $R$  is given by the fact that in the saturation region M2 can be compared to an ideal current generator, while in linear zone it behaves like a resistor. The overall effect is that it is able to provide a higher current with much smaller dimensions.

The commutation of a capacitive load is very important and it is always present in MOS devices. In Fig. 5.5 we study the commutation of the inverter composed of M1 and M2 with a capacitive load  $C$  connected to the output. Let the initial state be  $V_{IN} = 0$  V and  $V_{OUT} = V_{DD}$ . The capacitor  $C$  is completely charged and it must be discharged through M1. To achieve a fast discharge of  $C$ , M1 should be big; to be able to charge  $C$ , by turning off M1, in a short time, we must have a big current source, i.e. M2 must be big.

On the other hand, if M2 is too big we are no longer able to discharge  $C$ , because if  $V_{IN}$  is  $V_{DD}$  and we want  $V_{OUT}$  voltage near to ground, M2 must be able to provide less current than the one that M1 can sink. Otherwise the two transistors get to a “ratio” condition and the output node is set to a voltage value proportional to the partition of their conductances.

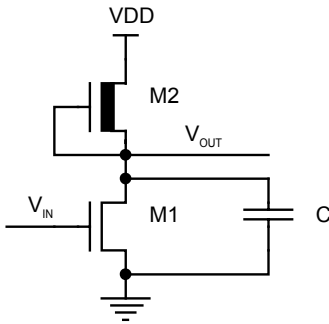


**Fig. 5.5.** NMOS inverter with output capacitive load

Figure 5.6 shows the equivalent circuit for the inverter of Fig. 5.5. Recalling that the equivalent resistance  $R_{eq}$  seen by looking into the source of M2 is equal to

$$R_{eq} = \frac{1}{g_{mb2} + g_{ds2}} \cong \frac{1}{g_{mb2}} \quad (5.1)$$

we can obtain the voltage gain of the amplifier stage

$$\frac{V_{OUT}}{V_{IN}} = -\frac{g_{m1}}{g_{mb2}} \propto -\sqrt{\left(\frac{W}{L}\right)_{M1}} \cdot \left(\frac{L}{W}\right)_{M2} \quad (5.2)$$

Let's recall that the dimensions of M1 and M2 are the only parameters that the designer can set: for instance, for a  $2\text{ }\mu\text{m}$  process, size of M1 can be  $10\text{ }\mu\text{m}/2\text{ }\mu\text{m}$  while size of M2 can be  $4\text{ }\mu\text{m}/8\text{ }\mu\text{m}$ . Transistor M2 is resistive; the mean current provided by a transistor of this kind is  $200\div 300\text{ }\mu\text{A}/\square$ , therefore M2 can provide  $100\div 150\text{ }\mu\text{A}$ . Our inverter has not been designed to drive big capacitances.

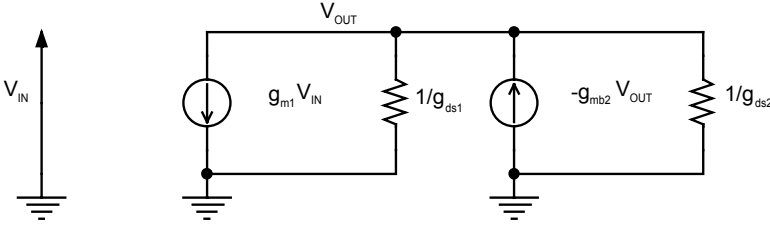


**Fig. 5.6.** Simplified equivalent circuit for the NMOS inverter

Maybe it is better to specify what we mean by big capacitances. The capacitances seen by the integrated circuit towards the external are comprised between 30 pF and 100 pF. Internally, biggest capacitances are given by ground and supply nodes, and they can be as high as some tens of nanoFarad. The capacitive value for a gate is in the order of some tens of femtoFarad. An internal node can have a maximum capacitance of some picoFarad. Our inverter can therefore charge in 1 ns:

$$C = I \frac{\Delta t}{\Delta V} = \frac{150 \cdot 10^{-6} \text{ A} \cdot 10^{-9} \text{ s}}{5 \text{ V}} = 30 \text{ fF} \quad (5.3)$$

that is a gate of some tens of square microns (very small indeed!).

The next step is to move to the CMOS inverter.

The main issue with the NMOS inverter is the current consumption in DC: when  $V_{IN}$  is equal to VDD, both transistors are tuned on and a current is flowing from power supply to ground. If “complementary” p-channel devices are used, the current consumption in DC is no longer present. Figure 5.7 shows the schematic of the CMOS inverter.

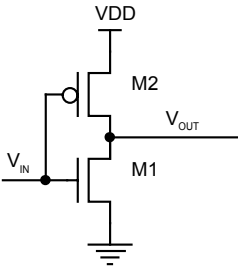


**Fig. 5.7.** CMOS inverter

The two gates are connected to the same input signal  $V_{IN}$ , drain junction of M1 and source junction of M2 are short-circuited to provide the output. When  $V_{IN}$  is at ground, M2 is completely turned on, while M1 is turned off and  $V_{OUT}$  is at VDD. On the other hand, if  $V_{IN}$  is at VDD, M2 is turned off, M1 is turned on and  $V_{OUT}$  is at ground. Contrary to the case of the NMOS inverter,  $V_{OUT}$  always reaches power

supplies. Current consumption occurs only during commutation when, owing to the fact that  $V_{IN}$  moves in a finite time, both transistors are turned on. This current is also known as *crow-bar* current.

**Problem 5.2:** The issue of power consumption is particularly critical for those inverters that implement the output buffers. Design a CMOS inverter with the minimum possible crow-bar current.

In order to discuss the CMOS inverter, it is necessary to recall that the mobility of the holes is approximately 1/3 of the one of the electrons.

The expression for the current is:

$$I_{DS} = \mu C_{ox} \frac{W}{L} f(V_{DS}, V_{GS}, V_T) \quad (5.4)$$

Assuming the same kind of oxide for both the transistors (hypothesis that is always true) a p-channel and a n-channel have the same current when:

$$\left(\frac{W}{L}\right)_{PMOS} = \frac{\mu_n}{\mu_p} \left(\frac{W}{L}\right)_{NMOS} \quad (5.5)$$

A “balanced” inverter in a 0.8  $\mu\text{m}$  process requires, for instance,

$$\left(\frac{W}{L}\right)_{M1} = \frac{10}{0.8}; \left(\frac{W}{L}\right)_{M2} = \frac{30}{0.8} \quad (5.6)$$

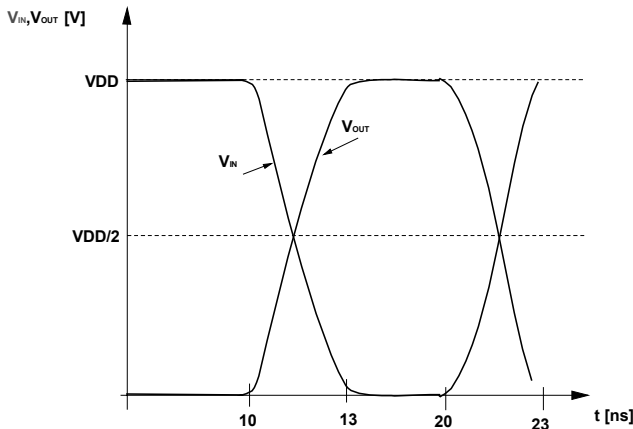


**Fig. 5.8.** Commutation of the balanced CMOS inverter

The result with a capacitive load is shown in Fig. 5.8, where the intersection of the input and output voltages is  $V_{DD}/2$ . If we want to vary the crossing point of the characteristics, we need to act on sizes. If we design  $M1 \gg M2$  it is much easier to pull  $V_{OUT}$  to ground since  $M2$  cannot provide much current and  $M1$  is anyway able to sink it all.

In this case the crossing moves towards the lower end and it becomes lower than  $V_{DD}/2$ . If we design  $M2 \gg M1$ ,  $M1$  is no longer able to sink all the current that  $M2$  provides, and therefore the  $V_{OUT}$  node remains high for a longer time; the crossing point moves towards values higher than  $V_{DD}/2$ . Figure 5.9 illustrates this effect.

**Problem 5.3:** A popular discussion among designer is whether a NMOS inverter has a faster switching time than a CMOS: try to put an end to this argue!

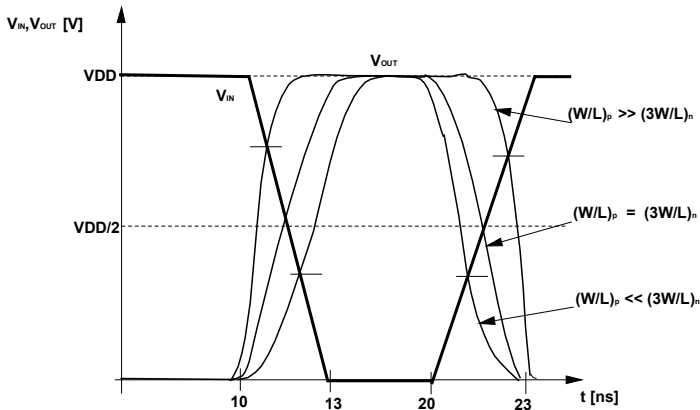


Fig. 5.9. CMOS inverter with different triggering points

## 5.3 The Cascode

The advantage of *cascode* configuration is the reduction of the effect given by Miller amplification of the coupling capacitance between the input and the output signal through, for instance, the gate/drain capacitance. Figure 5.10 shows the circuit where the parasitic capacitance between gate and drain is depicted<sup>2</sup>.

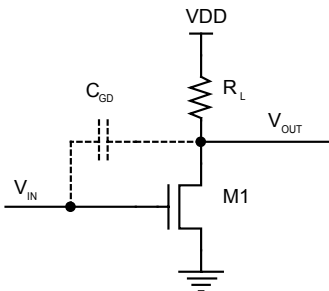


Fig. 5.10. Parasitic capacitance  $C_{GD}$  is amplified by the Miller effect

<sup>2</sup> Remember that between two nodes a parasitic capacitance is always present.

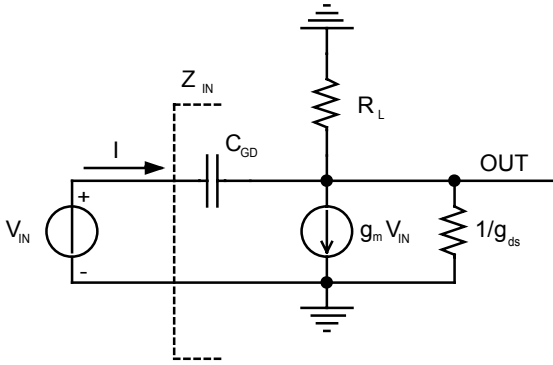


**Fig. 5.11.** Equivalent circuit for the inverter in Fig. 5.10

By substituting the transistor with its equivalent circuit, we get to Fig. 5.11, for which we can write that:

$$I = sC_{GD}(V_{IN} - V_{OUT}) \quad (5.7)$$

Disregarding the feedback contribution given by the current in  $C_{GD}$ , we obtain, for the output voltage:

$$V_{OUT} \cong -g_m R_L V_{IN} \quad (5.8)$$

and therefore:

$$(V_{IN} + g_m R_L V_{IN})sC_{GD} = I \quad (5.9)$$

$$V_{IN}(1 + g_m R_L)sC_{GD} = I \quad (5.10)$$

$$Z_{IN} = \frac{V_{IN}}{I} = \frac{1}{(1 + g_m R_L)sC_{GD}} \quad (5.11)$$

It is as if we had a capacitance equal to  $C_{gd}(1+g_m R_L)$ . If we assume that the gain of the inverter is in the order of several tens, it is easy to figure out how, thanks to Miller amplification, the small gate/drain coupling capacitance can be not so easy to drive for the previous stage. Coupling capacitance between input and output causes a decrease in the input impedance of the circuit, degrading its frequency response; to mitigate this effect, the configuration shown in Fig. 5.12 can be used.

By interposing M2 transistor, biased with a constant gate voltage, the input capacitance is decoupled from the output node. Let's evaluate the input impedance using the equivalent circuit shown in Fig. 5.13:

$$I = sC_{GD}(V_{IN} - V_X) \quad (5.12)$$

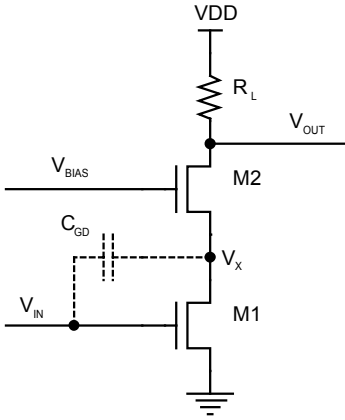


**Fig. 5.12.** Inverter with cascode stage

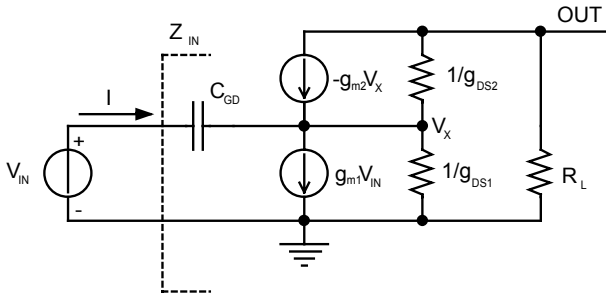


**Fig. 5.13.** Equivalent circuit for the inverter with cascode stage of Fig. 5.12

Disregarding the output resistances of the transistor, we can write that

$$g_{m1}V_{IN} = -g_{m2}V_X \quad (5.13)$$

$$V_{OUT} = g_{m2}V_X R_L \quad (5.14)$$

By combining the last two equations together we get

$$Z_{IN} = \frac{V_{IN}}{I} = \frac{1}{\left(1 + \frac{g_{m1}}{g_{m2}}\right)sC_{GD}} \quad (5.15)$$

In case  $g_{m1} \ll g_{m2}$  we have  $Z_{IN} \sim 1/(sC_{GD})$ . Even if M1 and M2 had similar sizes, Miller amplification factor is a small number, and feedback effect has been eliminated. We can also derive the expression for the voltage gains  $A_{OUT}$  and  $A_X$

$$A_{OUT} = \frac{V_{OUT}}{V_{IN}} = -g_{m1}R_L \quad (5.16)$$

$$A_X = \frac{V_X}{V_{IN}} = -\frac{g_{m1}}{g_{m2}} \quad (5.17)$$

$A_X$  gain must be low to reduce Miller effect, while  $A_{OUT}$  must be high because it is the real gain of the stage.

Figure 5.14 shows a NMOS cascoded configuration where the load resistor has been replaced by an active load (M3). Same considerations as above apply, considering the output resistance of M3 instead of  $R_L$ .

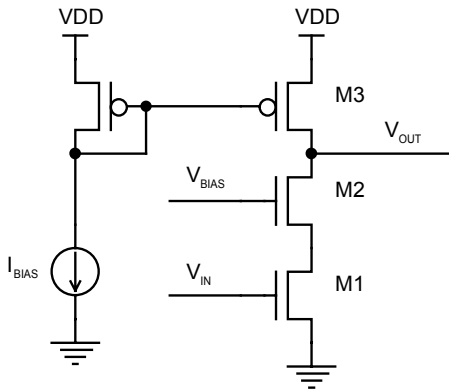


**Fig. 5.14.** Cascoded stage with active load

## 5.4 Differential Stage

In this section the fundamental concepts of one of the most important and widely used circuit configurations, the differential stage, are presented. Figure 5.15 shows the principle scheme of a NMOS differential amplifier. The two loads, M3 and M4, are considered to be equal, as well as M1 and M2;  $I_o$  is an ideal current generator. This circuit provides a difference of the output voltages ( $V_{o1} - V_{o2}$ ) that is greater than the difference of the input voltages ( $V_{i1} - V_{i2}$ ), thus implementing an amplifier.

It is better to recall now what we mean when we talk of differential and common mode gain, the two main parameters that are used to evaluate how good is a differential amplifier.

It is clear that we would like to have the output difference as a sole function of the input difference, and therefore, independent of the value of the VDD, or of the possible “fluctuations” on the ground, or of the size of the load, or also of the  $I_0$  current. Instead, the gain depends on the working point of the circuit; for instance, if supply voltage varies so much that M1 and M2 are outside saturation zone, we cannot expect the gain of the differential stage to remain unchanged. Even if supply had a little variation, thus keeping both M1 and M2 in saturation zone, because of the well-known modulation effect of  $I_{DS}$  current as a function of  $V_{DS}$ , the working point of the input transistor changes, and so does the gain.

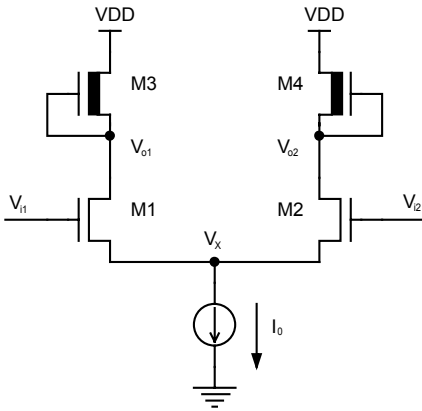


**Fig. 5.15.** Principle scheme for the NMOS differential stage

Let's now assume that  $V_{i1}$  and  $V_{i2}$  have a variation of  $V_{id}/2$ , equal as absolute value but with opposite sign. Then

$$i_{d1} = -i_{d2} \quad (5.18)$$

Since the circuit is completely symmetrical, voltage  $V_x$  cannot vary and the differential stage can be re-drawn as shown in Fig. 5.16, where the current generator  $I_0$  is replaced by a short circuit to ground. Now we can calculate the differential gain  $A_{dm}$

$$A_{dm} = \frac{V_{od}}{V_{id}} = -\frac{g_{m1}}{g_{mb3}} \quad (5.19)$$

Therefore, to have a high differential gain, M1 must be bigger than M3. It is worth noting that the current generator  $I_0$  has no influence on the value of the differential gain.

Let's now analyze the common mode, as in Fig. 5.17, where two possible working points for the differential amplifier of Fig. 5.15 are shown. Biasing 1 and 2 are the values in DC of the inputs  $V_{i1}$  and  $V_{i2}$ ; we would like to have the same output voltage in both cases having equal voltage differences on the input.

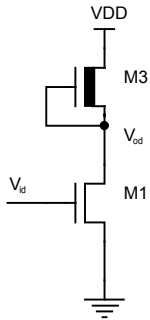


**Fig. 5.16.** Simplified configuration used to calculate  $A_{dm}$

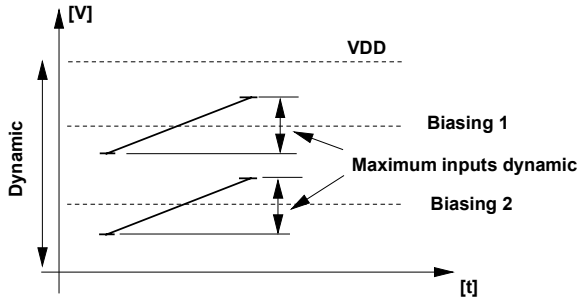


**Fig. 5.17.** Different working points for the differential stage

The contribution of the biasing is defined as “common mode” because it equally affects both sides of the differential stage. Let’s assume to increase the input voltages  $V_{i1}$  and  $V_{i2}$  of the same quantity  $V_{ic}$  (Fig. 5.18):  $V_{o1}$  and  $V_{o2}$  would tend to decrease, but thanks to the current generator  $I_0$ ,  $V_x$  voltage increases preventing the decrease of both  $V_{o1}$  and  $V_{o2}$ .

If the current generator were not present, we would not have any kind of feedback, and the output nodes would change their voltage even if the inputs were short-circuited! Therefore if  $I_0$  is an ideal generator, the common mode gain will be zero because no variation occurs on the outputs when the inputs vary.

$$A_{cm} = \frac{\Delta V_o}{\Delta V_i} = 0 \quad (5.20)$$

The equivalent circuit for the calculation of the common mode gain  $A_{cm}$  is shown in Fig. 5.19, where  $R_G$  is the impedance of  $I_0$  in the real case. If we assume the perfect symmetry of the circuit also in this case, it is possible to analyze just one half of the circuit, provided that we divide the resistance  $R_G$  into two parallel resistors whose value is twice the original one.

$$V_X = g_{m1}(V_{ic} - V_X) \cdot 2 \cdot R_G \quad (5.21)$$

$$A_{cm} = \frac{V_{oc}}{V_{ic}} = \frac{g_{m1} / g_{mb3}}{1 + g_{m1} 2R_G} \quad (5.22)$$

As we have previously said, the common mode gain decreases as resistance  $R_G$  increases.

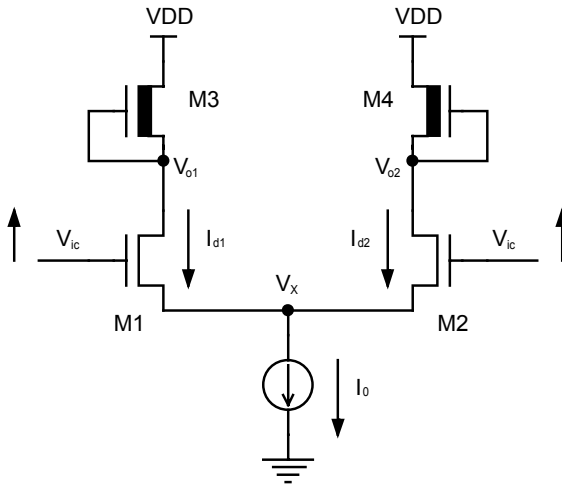


**Fig. 5.18.** Common mode

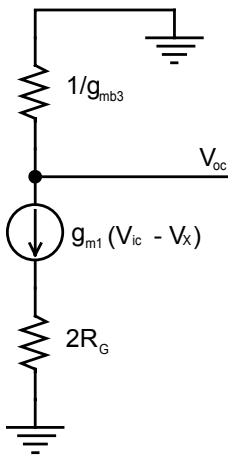


**Fig. 5.19.** Simplified configuration for the calculation of the common mode gain  $A_{cm}$

An important figure of merit is the Common Mode Rejection Ratio (CMRR), ratio between the differential mode gain and the common mode one. A good differential amplifier must therefore have a high CMRR.

In the following, we will also see the use of differentials as comparators.

**Problem 5.4:** Analyze the differences between PMOS-input and NMOS-input differential stages.

## 5.5 The Source Follower<sup>3</sup>

A voltage buffer is a circuit that is able to transfer the voltage signal on its input directly to the output independently of the applied load. One of the most widely used schemes is shown in Fig. 5.20: it is a feedback circuit based on the concept of virtual ground. In this case  $V_{OUT}$  voltage remains equal to  $V_{IN}$  by exploiting the feedback between nodes  $V_{O2}$  and  $V_{OUT}$ . When the circuit is powered up,  $V_{OUT}$  is at ground, therefore only M1 is turned on and  $V_{O2}$  is at VDD. M3 is turned on and it starts flowing current on the load  $R_L$  and to cause  $V_{OUT}$  to increase, which turns on M2, making  $V_{O2}$  decrease, which limits the current of M3 reducing  $V_{OUT}$ . This “loop” goes on until the differential stage has reached its point of equilibrium, i.e. the currents flowing in the two resistors  $R$  are equal to  $I_0/2$ .

Generally the gain stages (for instance the inverter with active load) have output impedance that is too high to drive low resistive or large capacitive loads. In these cases it is necessary to add an output stage that is able to transfer the signal with low impedance. The simplest structure to achieve this result is the source follower that, unlike the structure shown in Fig. 5.20, is also a level shifter. The circuital implementation is shown in Fig. 5.21 where the output is taken on the source terminal onto which the load  $R_L$  is applied.

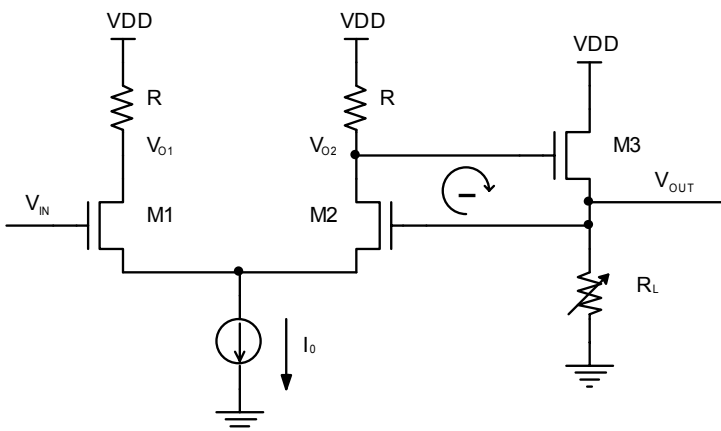


**Fig. 5.20.** The concept of virtual ground:  $V_{OUT}$  “follows” the value of  $V_{IN}$  thanks to the negative feedback that is present in the loop shown.

<sup>3</sup> When dealing with bipolars, it is called emitter follower.

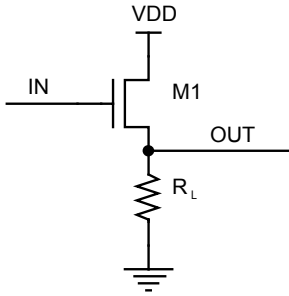


**Fig. 5.21.** Source follower

Figure 5.22 is the equivalent circuit for Fig. 5.21, where the voltage gain, neglecting the parasitic capacitances, is equal to

$$\frac{V_{OUT}}{V_{IN}} = \frac{g_{m1}}{g_{m1} + g_{mb1} + g_{ds1} + (1/R_L)} \quad (5.23)$$

Assuming that M1 is in saturation zone, that  $R$  is big enough to neglect the term that contains it and that the body effect is negligible, we get a unit gain, i.e. the input is equal to the output.

If we replace the load resistance  $R_L$  with an active load, as shown in Fig. 5.23, and we call  $r_o$  and  $C_o$  the total resistance and the total capacitance between the output and ground, we obtain a gain equal to:

$$\frac{V_{OUT}}{V_{IN}} = \frac{sC_{gs} + g_{m1}}{s(C_{gs} + C_o) + g_{m1} + g_{mb1} + g_{ds1} + (1/r_o)} \quad (5.24)$$

The gain in DC always tends to one, while in frequency we have a zero and a pole; if the zero is compensated with the pole, the bandwidth can become very large.

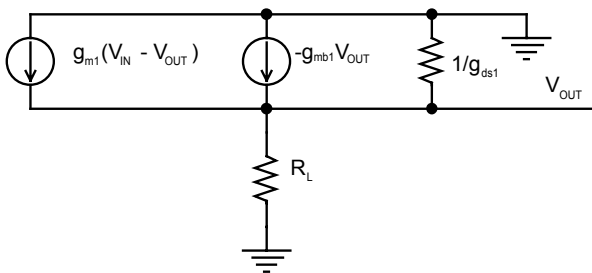


**Fig. 5.22.** Equivalent circuit for the source follower

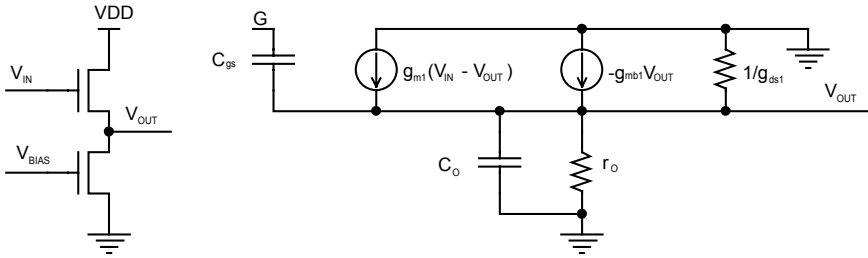


**Fig. 5.23.** Source follower with active load and corresponding equivalent circuit

## 5.6 Voltage References

The simplest implementation of a voltage reference is shown in Fig. 5.24. The partition constituted by R1 and R2 allows the generation of an output voltage that is a fraction of the supply voltage VDD.

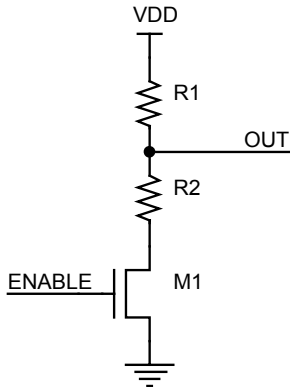


**Fig. 5.24.** Voltage reference based on the partition of the supply voltage

If the equivalent resistance of M1 is not taken into account, the voltage at the node OUT is equal to

$$V_{OUT} = VDD \frac{R2}{R1 + R2} \quad (5.25)$$

Transistor M1 is used to avoid consumption when the partition circuit is not active.

$V_{OUT}$  voltage does not depend neither on the temperature nor on the process, under the assumption that the resistors are fabricated using the same layer. Depend-

ence on supply voltage still occurs, and we will discuss this issue later. We have now to choose which of the available layers can be used to realize the resistors.

Resistivity of poly2 is too low due to the presence of silicide, so resistances that cause low power consumption cannot be fabricated with poly2.<sup>4</sup> The consumption of the partition circuit is in fact

$$I = \frac{VDD}{R1 + R2} \quad (5.26)$$

If VDD is equal to 5 V and the target consumption is 10  $\mu$ A, the value of the resistance must be 5 M $\Omega$ . To get such a value using poly2, it is necessary to use very long stripes that occupy too much space. Similar considerations apply to active area, whose value of resistance is anyway an order of magnitude higher than the one of poly2. Therefore n-well is the remaining choice, whose resistivity is three orders of magnitude higher than the one of poly2. We have already seen how to realize a resistor in n-well and the associated depletion issue. Resistive partition circuits are used anyway thanks to their simple implementation whenever precision is not a key factor, and stabilizing, if necessary, their supply voltage.

Let's now analyze some voltage references implemented using active devices.

### 5.6.1 NMOS

Figure 5.25 shows a reference circuit realized using n-channel transistors; this implementation is stable with respect to variations of the supply voltage. If M1 works in saturation region (high  $V_{DS}$ ) a variation of VDD moves the working point of M1, onto the characteristic  $V_{GS} = 0$ , causing M1 to remain in saturation region. Figure 5.26 shows the graph of both VDD and the output node  $V_{REF}$  for which we can write that

$$V_{REF} = V_{T,M2} \pm |V_{T,M1}| \cdot \sqrt{\beta_1 / \beta_2} \quad (5.27)$$

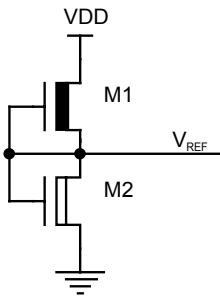


**Fig. 5.25.** Voltage reference stable with respect to VDD variations

<sup>4</sup> Recall that the overall consumption in read mode for a Flash memory device is in the order of tens of mA.

It is possible to increase the stability interval for  $V_{REF}$  by modifying the circuit as shown in Fig. 5.27. Let's observe that the stability range for the circuit depicted in Fig. 5.25, that is between 3.5 V and 5 V, is due to the fact that for low VDD the value of the  $V_{REF}$  node can only decrease, while in the circuit of Fig. 5.27 M4 transistor with its gate to VDD tends to push  $V_{REF}$  upwards for low values of VDD because less current flows across M4.

When VDD is stable, the effect of M4 becomes negligible and the circuit behaves like the one in Fig. 5.25. Transistor M3 acts as an auxiliary current generator when, being VDD low, M1 is not able to provide all the required current to M4. In this way the value of  $V_{REF}$  only depends on the ratio between M1 and M2, while the stability range is regulated by means of M3 and M4.

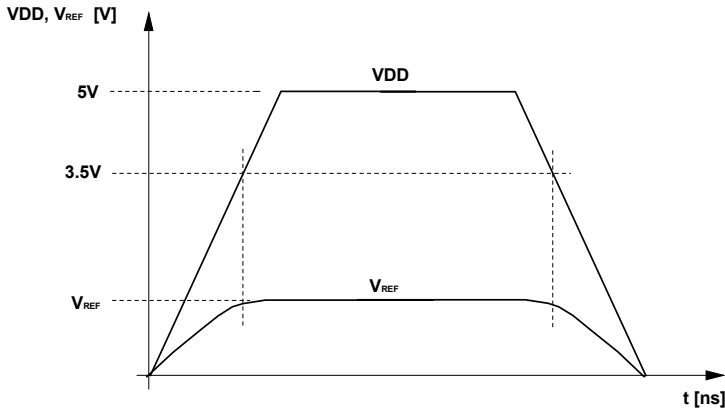


**Fig. 5.26.** Stability of  $V_{REF}$  voltage (Fig. 5.25) for VDD between 3.5 V and 5 V

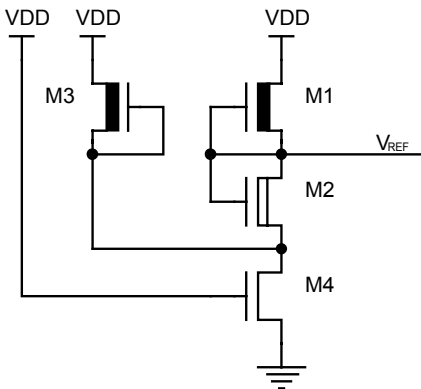


**Fig. 5.27.** Circuit to increase the stability interval of  $V_{REF}$  voltage (Fig. 5.25)

Another voltage reference circuit that can be realized using NMOS transistors only is shown in Fig. 5.28.

Assuming that both the transistors are working in saturation region, we can write that

$$\beta_1(-V_{REF} - V_{T,M1})^2 = \beta_2 \cdot V_{T,M2}^2 \quad (5.28)$$

and therefore

$$V_{REF} = -V_{T,M1} \pm V_{T,M2} \sqrt{\beta_2 / \beta_1} \quad (5.29)$$

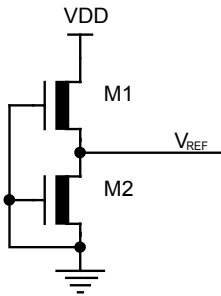


**Fig. 5.28.** Depletion NMOS voltage reference circuit

We must consider the smallest among the two  $V_{REF}$  values, because the maximum possible  $V_{REF}$  is equal to  $-V_{T,M1}$  (the threshold voltage for a depletion NMOS is negative); for higher voltages, M1 is turned off. The threshold voltages for both M1 and M2 are explicitly shown because the body effect must be taken into account for M1.

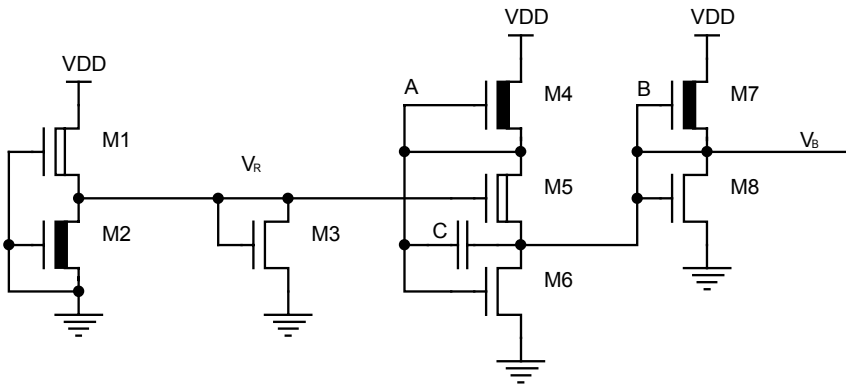


**Fig. 5.29.** Generator for  $V_B$  reference voltage

Another NMOS reference is shown in Fig. 5.29. Transistors M1 and M2 set the required voltage around the threshold of M1; M3 is a protection that limits  $V_R$  voltage to the  $V_T$  of M3, obviously  $V_{T,M3} > V_{T,M1}$ . Transistors M4, M5 e M6 constitute a feedback control stage that is able to stabilize the output node  $V_B$ . If the voltage of  $V_R$  increases, the voltage of node A tends to decrease, cutting off M6 that will bring the voltages of nodes A and B back to the fixed values, i.e.  $2V_{T,M1}$  for the node B. Capacitor C acts as a filter, to absorb the current peaks that might occur when the device is switched on; finally, M7 and M8 act as auxiliary current generators for the output node.

### 5.6.2 CMOS

Let's now analyze some voltage references realized using CMOS transistors. The basic principles are the same previously described, with the additional advantage that p-channel transistors can be used and therefore the contribution of the body effect can be eliminated, since in every CMOS technology the substrate of the PMOS transistors can be independently biased. Figure 5.30 shows an implementation where the VDD is partitioned using two p-channel transistors: in case M1 and M2 have the same size,  $V_{OUT}$  output voltage is equal to  $VDD/2$ .

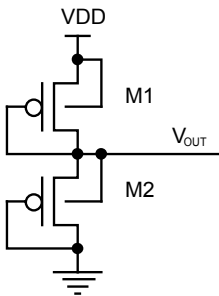


**Fig. 5.30.** PMOS voltage reference where the contribution of the body effect has been eliminated

### 5.6.3 Self-Biased Generator

Figure 5.31 shows a self-biased circuit that is more complex than those studied before; it is made of two current mirrors, one towards VDD, realized with a pair of equally-sized PMOS transistors, and the other towards ground with a degeneration resistance. PMOS mirror forces the same current in the right and in the left branch. M3 and M4 transistors have a different size ratio where  $\beta_3 > \beta_4$ ; since they must sink an equal drain current, their  $V_{GS}$  must be different. All the transistors must work in saturation region to ensure the desired behavior of the circuit.

Transistor M4 is for sure in saturation, since it is diode-connected and  $V_{GS4}$  is equal to

$$V_{GS4} = V_{T,N} + \sqrt{I_{REF} / \beta_4} \quad (5.30)$$

$$\beta_4 = \frac{\mu_n C_{ox} W_4}{2L_4}; K = \frac{W}{L} \quad (5.31)$$

Neglecting the body effect on M3, we can calculate the current  $I$  from the voltage drop on resistance  $R$

$$I_{REF} = \frac{V_{GS4} - V_{GS3}}{R} = \frac{1}{R} \left( \sqrt{I_{REF} / \beta_4} - \sqrt{I_{REF} / \beta_3} \right) \quad (5.32)$$

The equation has two possible solutions

$$I_{REF} = 0; I_{REF} = \frac{2\Psi}{C_{ox}} \cdot \frac{1}{\mu_n R^2} \quad (5.33)$$

where

$$\Psi = \frac{1}{K_3} + \frac{1}{K_4} - \frac{2}{\sqrt{K_3 K_4}} \quad (5.34)$$

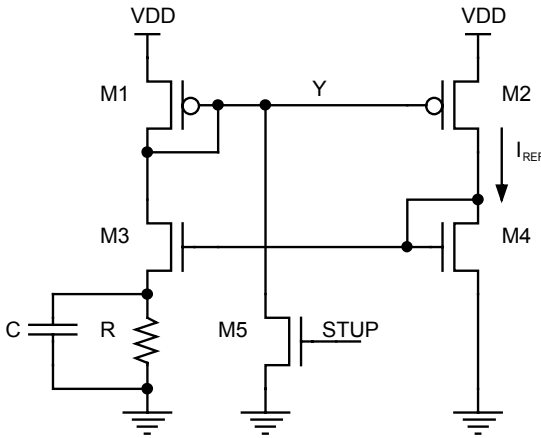


**Fig. 5.31.** Self-biased generator

$\psi$  factor just depends on geometric parameters. The solution where the current is not zero (the only one to take into account) is inversely proportional to two terms that depend on temperature. Since mobility increases when temperature decreases, and the resistance behaves the opposite, the two variations can be partially compensated to reduce the variation range of the current  $I_{REF}$  with temperature. Generally, the best compensation is achieved by realizing the resistance in active

area; using this layer is not always possible because of the associated large area occupation, above all when consumption of few microamperes are required. There is no dependence of the  $I_{\text{REF}}$  current with the supply voltage VDD.

Transistor M5 provides the required start-up to the circuit, keeping the p-channel turned on for some tens of nanoseconds. In this way the current forced in M1 lets a current flow in M2 that biases M4; this one turns on M3 and at this moment the STUP signal is turned off.

M3 and M4 transistors can also be of different type (for instance M3 natural an M4 LVS); in this case it is not used the overdrive difference between M3 and M4, but their  $V_T$  difference is exploited, which is applied at the terminals of resistance R.

The switching on of this circuit is quite slow as is for all the feedback circuits (in the order of 100 ns in this case) because the settling of  $V_Y$  to its equilibrium value causes a series of variations in the currents of all the transistors. To decrease the switching on transient, it is possible to “help” the nodes to get to their equilibrium value; capacitance C is added, which represents a short-circuit towards ground at power on, thus eliminating the contribution of the resistance, which in turn is the element with the most significant inertia in the power-on chain.

**Problem 5.5:** Add the elements that allow switching off the circuit in such a way that it does not provide current contributions during stand-by phase.

## 5.6.4 Band-Gap Reference

The last voltage reference that we are going to discuss is the *band-gap reference*, whose main feature is to produce an output voltage that is stable with respect to both VDD and temperature variations. This circuit is widely used in the latest generation of non-volatile memory devices, to control the voltages during program and erase operations. First generation EPROM memories, ancestors of the Flash, allowed for instance a variation of the drain voltage around its typical value of some hundreds of millivolt. Nowadays the size of the cells asks for an increased precision for the voltages, above all for drain voltage during program and gate voltage during verify operations. To carry out this requirement, the band-gap reference circuit is used. This circuit exploits the possibility of compensating temperature variation of the  $V_{\text{BE}}$  of a bipolar transistor with a voltage drop proportional to the thermal potential  $V_t$

$$V_t = \frac{kT}{q} \quad (5.35)$$

where  $k$  is Boltzmann’s constant,  $q$  is the charge of the electron and  $T$  is the absolute temperature.

$V_{\text{BE}}$  of a bipolar transistor decreases as the temperature increases, and this behavior can be linearized around a reference temperature, thus defining a first order thermal coefficient

$$\gamma_{V_{BE}} = \left. \frac{\partial V_{BE}}{\partial T} \right|_{T=T_0} \quad (5.36)$$

$$V_{BE} \cong V_{BE0} + \gamma_{V_{BE}} \cdot (T - T_0) \quad (5.37)$$

On the thermal potential, the following relation can be easily derived

$$\gamma_{V_t} = \frac{\partial V_t}{\partial T} = \frac{k}{q} = 0.085 \text{ mV/K} \quad (5.38)$$

Therefore the thermal potential is proportional to temperature, while the thermal coefficient of  $V_{BE}$  depends on the reference temperature that has been chosen to calculate it.

Coefficient  $\gamma$  of  $V_{BE}$  lies between  $-1 \text{ mV/K}$  and  $-2 \text{ mV/K}$ ; therefore  $\gamma$  of  $V_{BE}$  and of  $V_t$  have opposite signs. Now a  $V_{BG}$  voltage can be defined as follows

$$V_{BG} = V_{BE} + G \cdot V_t \quad (5.39)$$

For which the corresponding thermal coefficient can be eliminated by choosing a value of  $G$  such that

$$G = \frac{\left| \gamma_{V_{BE}} \right|}{\gamma_{V_t}} \quad (5.40)$$

Such a compensation technique is extremely simplified and it is known as first order because only the first order thermal coefficient of  $V_{BE}$  is compensated. There are more precise techniques that can take into account non-linear terms too. Furthermore a first order compensation is less precise as the target temperature range increases.

A widely used band-gap reference circuit is shown in Fig. 5.32.

Assuming an ideal situation (infinite gain, no voltage offset, etc.), its input voltages  $V^+$  and  $V^-$  are equal and therefore the voltage at the terminals of R3 resistor can be written as

$$V_{R3} = V_{BE0} - V_{BE1} = V_t \cdot \ln \frac{I_{C0} I_{S1}}{I_{C1} I_{S0}} \quad (5.41)$$

$$I_C = I_S \exp \left( \frac{V_{BE}}{V_t} \right) \quad (5.42)$$

$I_{C0}$  and  $I_{C1}$  are the collector currents of the bipolar transistors Q0 and Q1, while  $I_{S0}$  and  $I_{S1}$  are the corresponding currents of reverse-biasing of the base-emitter junction that depend on physical and geometrical parameters (it is reasonable to assume that these parameters are the same for both bipolar transistors) according to the following equation

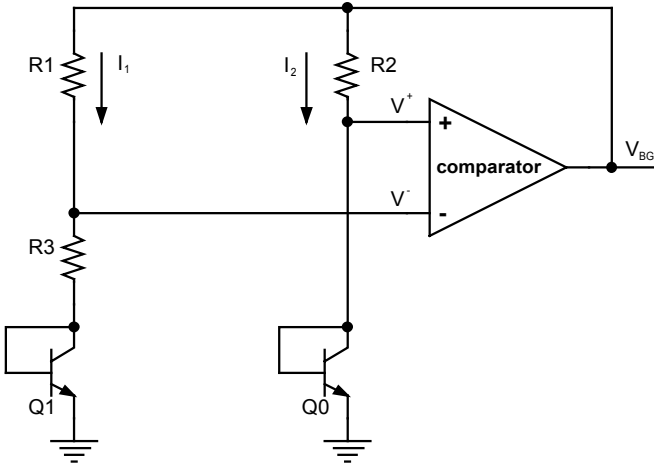


**Fig. 5.32.** Schematic of the band-gap reference circuit

$$I_{S0} = \frac{qD_n n_i^2}{W_B N_A} A_0; I_{S1} = \frac{qD_n n_i^2}{W_B N_A} A_1 \quad (5.43)$$

where

- $A_0$  and  $A_1$  are the emitter areas for the transistors Q0 and Q1;
- $n_i$  is the carrier intrinsic concentration in the semiconductor;
- $W_B$  is the base width;
- $D_n$  is the diffusion constant for the electrons;
- $N_A$  is the base doping concentration.

Always assuming ideal operation, it can be noted that the voltage drop on R1 and R2 is the same and therefore we can write that

$$I_1 \cdot R1 = I_2 \cdot R2 \quad (5.44)$$

$$I_1 = I_{C1} \left( 1 + \frac{1}{\beta} \right); I_2 = I_{C0} \left( 1 + \frac{1}{\beta} \right) \quad (5.45)$$

where  $\beta$  is the current gain of the bipolar transistors.

Therefore by substituting Eqs. (5.43), (5.44) and (5.45) in Eq. (5.41), the latter can be re-written as a function of resistors R1 and R2 and of the area of the bipolars

$$V_{R3} = V_t \ln \frac{R1 \cdot A_1}{R2 \cdot A_0} \quad (5.46)$$

Now, keeping in mind Eq. (5.44), it is possible to determine reference voltage  $V_{BG}$

$$V_{BG} = V_{BE0} + R2 \cdot I_2 = V_{BE0} + R1 \cdot \frac{V_{R3}}{R3} \quad (5.47)$$

By substituting Eq. (5.46) in Eq. (5.47) we get

$$V_{BG} = V_{BE0} + V_t \frac{R1}{R3} \ln \frac{R1 \cdot A_1}{R2 \cdot A_0} \quad (5.48)$$

from which coefficient G can be derived

$$G = \frac{R1}{R3} \ln \frac{R1 \cdot A_1}{R2 \cdot A_0} \quad (5.49)$$

In order to achieve a proper dimensioning of the circuit, the first order thermal coefficient of  $V_{BE}$  must be given and then the value of G must be calculated using Eq. (5.40). Furthermore, design parameters R1, R2, R3,  $A_0$  and  $A_1$  must be chosen accordingly to get the desired value of G. There is not a unique possible set of these parameters, and the choice also depends on target performances like consumption, turn-on time of the circuit etc.

It is worth pointing out that coefficient G is in first approximation independent of the temperature, because we only have resistive ratios (on top of the area ratio). In fact it is possible to model the relationship between temperature and resistance using the following equation

$$R_i(T) = R_i(T_0) \cdot [1 + \gamma(T) \cdot (T - T_0)] \quad (5.50)$$

Therefore considering two resistances with the same thermal coefficient, the term inside the parentheses can be simplified and a constant ratio is achieved. The remaining issue is the non-linear behavior of the resistance with applied voltage variation; in these cases, simulations are needed to determine the right value for the resistances.

A last consideration can be done on the need for a start-up circuit. In fact, condition  $V_{BG} = V^+ = V^- = 0$  is a possible, but unstable, working point where the circuit could be stuck, or anyway from which it could recover very slowly. Therefore an adequate start-up has a twofold aim: avoid that the circuit remains turned off and allow for a faster settling of the reference voltage.

**Problem 5.6:** Design the start-up circuit. Suggestion: put the operational out of balance adequately for a short period of time starting from a proper enable signal for the band-gap reference.

An example graph of the relationship of  $V_{BG}$  voltage with temperature is shown in Fig. 5.33: it is the result of a simulation of a band-gap circuit designed in  $0.15 \mu\text{m}$  technology. It is usually possible to have one or more non-volatile registers inside the device that allow changing factor G acting on the resistive partitioning of Fig. 5.32 without requiring a photomask change. The different curves in

Fig. 5.33 refer to eight different values for the compensation factor  $G$ , assuming that three configuration registers are available.

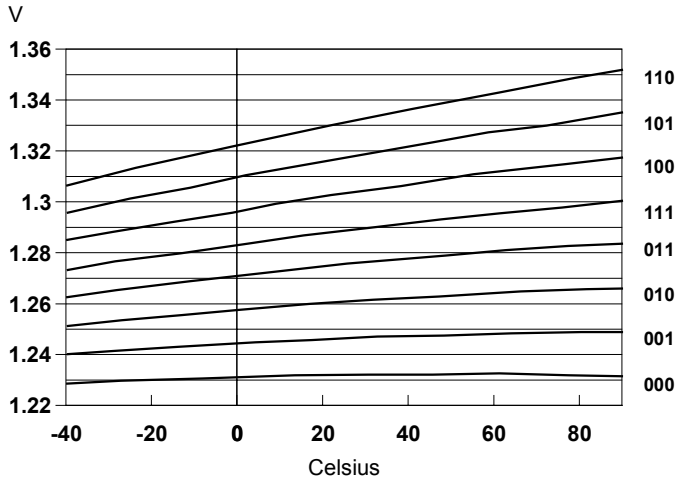


**Fig. 5.33.** Graphical representation of the band-gap voltage versus temperature and  $G$  compensation factor variation

The resulting value for  $V_{BG}$  is around 1.25 V. As we will see in the following, voltage regulation is achieved by multiplying the reference voltage by a proper coefficient. For instance, in order to generate 6 V,  $V_{BG}$  is multiplied by a factor 4.8. This operation has associated issues: variations of the reference voltage are amplified by the regulator and they can cause significant variations around the typical value of the output voltage. Even if  $V_{BG}$  voltage varies of, say, ten millivolt on the whole voltage and temperature range, the final result is that the variation of the regulated voltage is 4.8 times the band-gap one, i.e. some tens of millivolt. That's why it is mandatory to have a very precise and stable  $V_{BG}$ .

## 5.7 Current Mirrors

A widely used circuit configuration is the one known as current mirror. In the following, we will often need to generate a current equal to the reference one, or to a fraction of it, and the current mirror will let us achieve this result. Figure 5.34 shows a current mirror realized with NMOS transistors.

If we assume that M1 works in saturation zone, M1 current  $I_{REF}$ , only depends on its  $V_{GS}$ . Therefore we can write the following equation:

$$I_{REF} = \beta_1 (V_{GS} - V_{T,N})^2 \quad (5.51)$$

Assuming that M2 is working in saturation zone too, we have

$$I = \beta_2 (V_{GS} - V_{T,N})^2 \quad (5.52)$$

Since M1 and M2 are both LVS and they have the same  $V_{GS}$ , then the ratio of their currents is equal to the size ratio:

$$\frac{I_{REF}}{I} = \frac{\beta_1}{\beta_2} \quad (5.53)$$

In case the size of M1 and M2 is the same, then current  $I$  is equal to reference current  $I_{REF}$ .

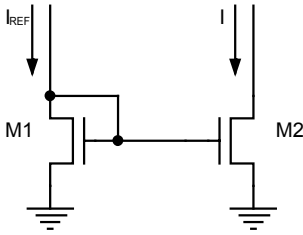


**Fig. 5.34.** NMOS current mirror

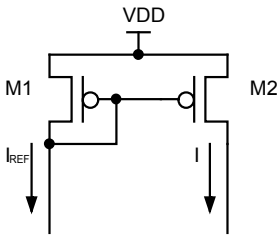


**Fig. 5.35.** PMOS current mirror

In a similar way it is possible to design a current mirror towards supply voltage using two PMOS transistors (Fig. 5.35). Same considerations apply.

In order to evaluate the quality of a mirror as a current generator, it is necessary to analyze its output impedance. In case of NMOS, the output impedance  $z_o$ , based on the small signal model, is equal to

$$z_O = r_{ds2} \propto \frac{L_2}{I_{DS2}} \quad (5.54)$$

If a big output impedance is required, then transistors with a long channel, and therefore with little available current, must be used. A possible solution to increment the output resistance is shown in Fig. 5.36.

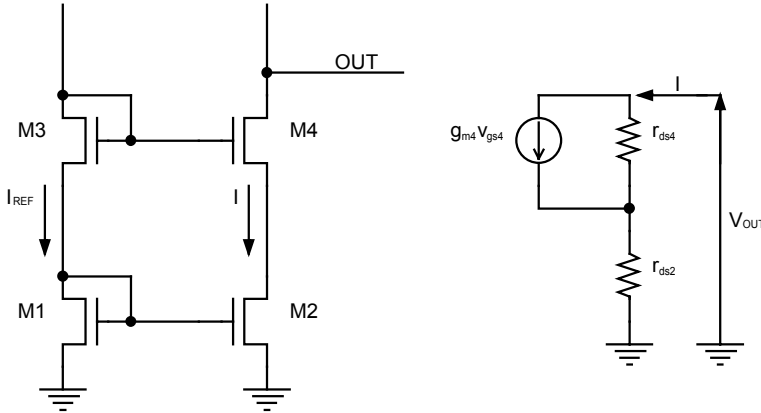


**Fig. 5.36.** Cascoded current mirror

Output impedance in this case is given by

$$z_O = \frac{V_{OUT}}{I} \quad (5.55)$$

$$z_O = r_{ds2} + r_{ds4}(1 + g_{m4}r_{ds2}) \quad (5.56)$$

The drawback of this circuit when compared with the one in Fig. 5.34 is a narrower swing for the output voltage. In fact, minimum  $V_{OUT}$  is equal to:

$$V_{OUT,min} = V_{GS1} + V_{GS3} - V_{GS4} + V_{sat2} \quad (5.57)$$

Recalling that the minimum voltage required between drain and source for saturation can be written as

$$V_{sat} = V_{GS} - V_T \quad (5.58)$$

we get

$$V_{OUT,min} \cong V_{T,N} + 2V_{sat} \quad (5.59)$$

In other words, output voltage can go below gate voltage of M4 of a maximum of a threshold voltage (body effect included). It means that if the gate of M3 is at about 2 V,  $V_{OUT}$  cannot go below 1 V (or less, because of body effect).

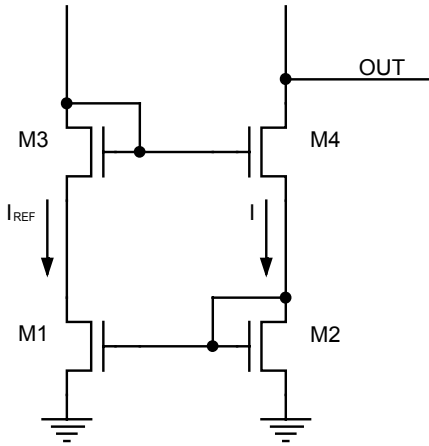


**Fig. 5.37.** Wilson current mirror

**Problem 5.7:** Discuss configuration of Fig. 5.37 comparing it with previous configurations. Design mirrors of Figs. 5.36 and 5.37 using complementary PMOS mirrors and discuss the difference with respect to the solution shown in Fig. 5.35.

## 5.8 NMOS and CMOS Schmitt Trigger

When we have analyzed the inverter, we have seen that it is possible to vary its triggering point modifying the dimensional ratio of the transistors. The Schmitt trigger is an inverter with two triggering points. In other words, this circuit has a high-to-low transition of the output when the input goes above a certain voltage  $V_{TH}^+$  while the opposite transition is inhibited until the input does not go below a voltage  $V_{TH}^-$ . In this way we have an inverter that is able to stabilize the switching against rapid triggering by noise as it passes by triggering point.

Figure 5.38 represents the graph of a possible “asymmetric” transition, where  $V_{IN}$  is the input signal and  $V_{OUT}$  is the output one. The triggering points shown in the figure, i.e. the intersection of the two curves, have different voltage values; the difference between the triggering voltages is called hysteresis.

Let’s analyze first the CMOS circuit that is realized to obtain the characteristic shown in Fig. 5.38, and then the NMOS one.

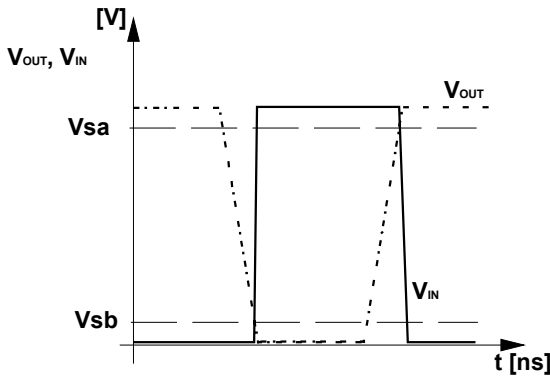


**Fig. 5.38.** Inverter with asymmetric input-output characteristic

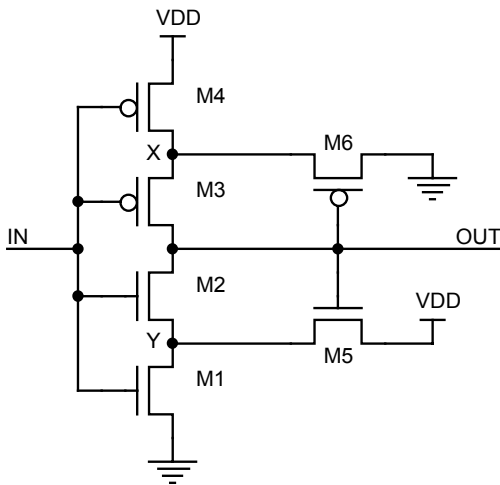


**Fig. 5.39.** CMOS Schmitt Trigger

CMOS version of the Schmitt trigger is shown in Fig. 5.39: hysteresis is achieved by means of the two complementary MOS, M6 and M5. When  $V_{IN}$  is at ground, M3 and M4 are turned on and the output voltage is VDD. Transistor M5 is turned on and it forces  $V_Y$  to  $VDD - V_{T,M5}$  (body effect included). Assuming a supply voltage of 5 V,  $V_Y = 3.5$  V. For sake of simplicity of the analysis, all and only the transistors involved in this phase are depicted in Fig. 5.40.

When the input voltage goes above the threshold voltage of M1 (1 V), M1 starts to conduct while M2 is still turned off. Voltage  $V_Y$  starts to decrease because of the partition between M1 and M5. As  $V_{IN}$  keeps increasing, transistor M1 continues to make  $V_Y$  decrease until when M2 turns on. At this point the circuit starts

to force  $V_{OUT}$  to decrease towards ground. Triggering point of the Schmitt trigger is about 3.5 V for this transition.

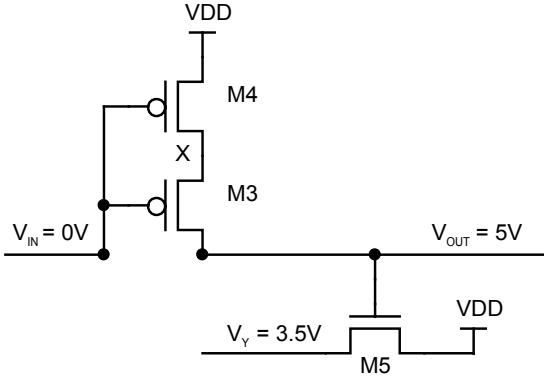


**Fig. 5.40.** CMOS Schmitt trigger: simplified circuit for the analysis of the biasing in case of input voltage equal to ground

The Schmitt trigger, as well as the inverter seen in Sect. 5.2, belongs to the family of the ratioed logics. Static characteristic, and therefore the value of the hysteresis, can be analytically derived by solving the equation that shows the equivalence of the currents of driver and load transistors getting  $V_{OUT}$  as a function of  $V_{IN}$ .

Assuming that voltage  $V_Y$  is able to bias M1 in its saturation region, we can calculate the triggering point for the high-low transition of the output.

$$V_{IN} = V_Y + V_{T,M4} \quad (5.60)$$

$$\frac{\beta_1}{2} (V_{IN} - V_{T,M1})^2 = \frac{\beta_5}{2} (VDD - V_Y - V_{T,M6})^2 \quad (5.61)$$

$$V_{IN} = \frac{VDD + \sqrt{\beta_1 / \beta_5} \cdot V_{T,M1}}{1 + \sqrt{\beta_1 / \beta_5}} \quad (5.62)$$

Let's now consider the case of input voltage equal to VDD (Fig. 5.41). Transistors M1 and M2 are turned on and the output is low. Node X is floating and it is forced by M6 to reach its own threshold voltage (1.5 V). When  $V_{IN}$  goes one threshold voltage  $V_{T,P}$  below VDD, M4 turns on and starts to compete with M6. Owing to the progressive turning on of M4, the voltage of node X starts to increase until a value that lets M3 to turn on causing the transition of the output node towards VDD. In this case the triggering point of the circuit is about 1.5 V.

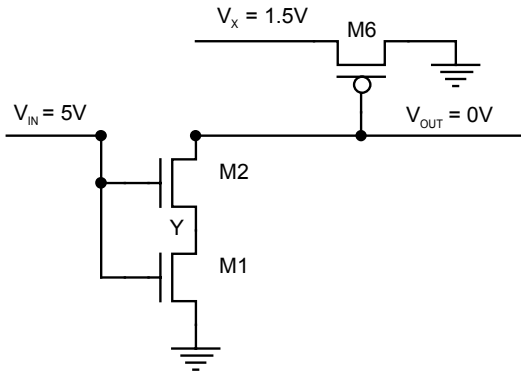


**Fig. 5.41.** CMOS Schmitt Trigger: simplified circuit for the analysis of the biasing in case of input voltage equal to VDD

The scheme of the Schmitt trigger realized using n-channel transistors only is shown in Fig. 5.42.

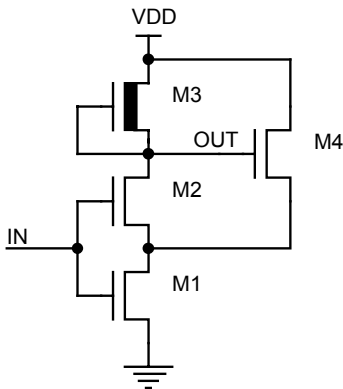


**Fig. 5.42.** NMOS Schmitt Trigger

Analyzing at Fig. 5.42, it is clear that in case  $V_{OUT}$  is equal to zero, transistor M4 is turned off; therefore the current of M4 is present only during one of the two transitions. In particular, when  $V_{IN}$  increases, M1 and M2 conduct and, as soon as  $V_{IN}$  goes above  $V_{T,N}$ , M4 is turned off. During the high-low transition of the  $IN$  signal (Fig. 5.43), M4 turns on after M2 turns off and only when the voltage at the node  $OUT$  is higher than  $V_{T,N}$  (body effect included).

**Problem 5.8:** Calculate the value of the hysteresis for the circuits of Figs. 5.39 and 5.42.

Another circuitual scheme that can be used to get a hysteresis is shown in Fig. 5.44. It is based on two inverters that are connected in positive feedback loop. INV1 is the driver of the circuit under investigation. Let's assume that signal IN is low: M1 and M4 are turned on while M2 and M3 are turned off. If M4 were not present, low-high transition of the input voltage would immediately drive the output to low; being M4 present and turned on, it follows that  $V_{IN}$  must increase more than in the case of the driving of a normal inverter, in order to be able to push  $V_{OUT}$  low enough (so that it can turn off M4). The drawback of this circuit is the consumption of the driver.

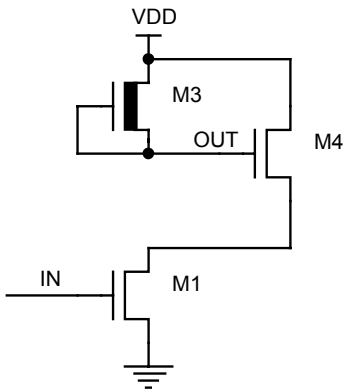


**Fig. 5.43.** NMOS: Schmitt trigger: simplified circuit for the analysis of the biasing in case of high-low transition of the input signal

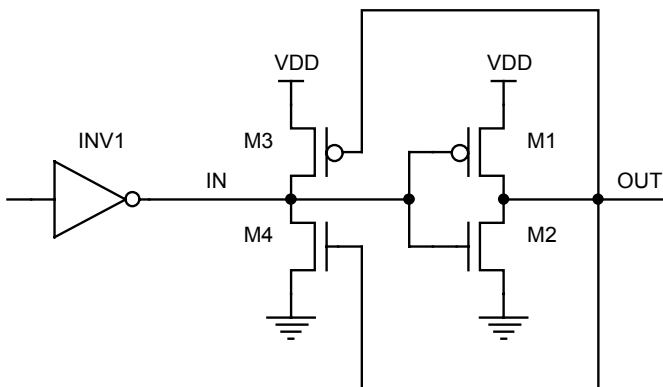


**Fig. 5.44.** Using a latch as an inverter with different triggering points

## 5.9 Voltage Level Shifter Latch

Inside a Non-Volatile memory device different voltages are present, which are used for program, erase, read, verify etc. It is therefore required to shift various signals from the VDD/GND to the VPC/GND range, where  $V_{PC} > V_{DD}$ ; the circuit used for this task is shown in Fig. 5.45a. VPC supply voltage can be either VDD (for example 5 V) or 12 V, during the different operating modes, while the inverter that drives M2 is always supplied at VDD. The voltage swing of the IN signal is in the VDD/GND range. When M2 is turned on, the voltage of the output node  $V_{OUT}$  is ground, whereas it is VPC when M2 is turned off and M1 is on thanks to the positive reaction due to M3 and M4 transistors. To decrease the voltages applied to the transistors, thus increasing reliability of the whole structure, the circuit depicted in Fig. 5.45a is modified as shown in Fig. 5.45b. Aim of M5 and M6 transistors is to reduce  $V_{DG}$  on M1 and M2 transistors.

Positive reaction can be used to design negative level shifter as well, as shown in Fig. 5.46. HVNEG value can be either GND or negative. M7 and M8 are NMOS realized in the triple well to be able to transfer negative voltages. Power supply for the inverter and for transistors M9 and M10 can be reduced to a value lower than VDD to limit the voltage difference applied to both the junctions and the oxides.

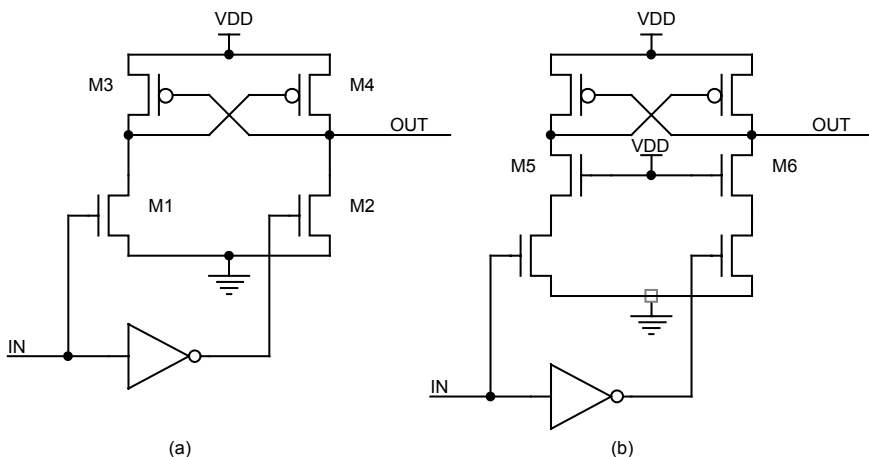


**Fig. 5.45.** Voltage Level Shifter Latch

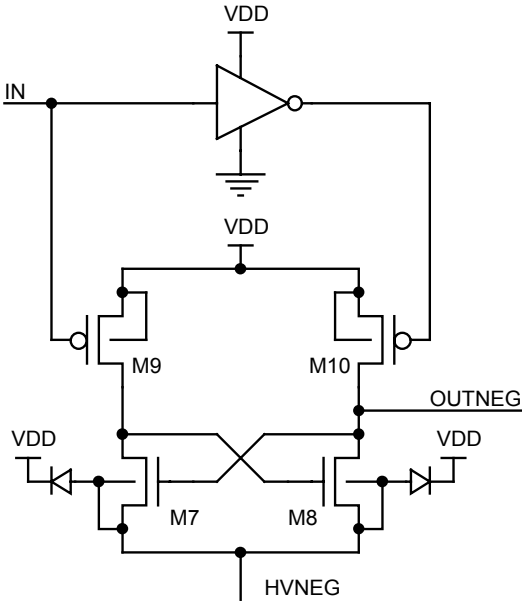


**Fig. 5.46.** “Cascode” solutions for the level shifter latch

## 5.10 Power On Reset Circuits

A problem shared by nearly all digital devices is their correct initialization (*Reset*) as soon as supply voltage is applied, i.e. during *Power On*. A proper circuit to generate the so-called *Power On Reset* (POR) is required, whose rationale is the comparison between the ramp of the supply voltage and the value of a node inside the circuit used as a reference. To achieve correct behavior, it is mandatory that the voltage on this node reaches a stable value before the supply voltage.

In an NMOS device, having an active signal during the ramp of the supply voltage is not an issue, since power consumption during inactive phase of the POR circuit itself can be disregarded. An example is shown in Fig. 5.47: the circuit is composed of a positive reaction block, prototype of the set/reset latch, and by two other blocks, comprised of M1 and M2 (which we already know and which acts as a voltage reference) and of the series of M3, M4 and M5, whose sizes are chosen in such a way that the node “6” is equal to the supply voltage minus a fixed offset. The circuit is really simple; more complex versions can be designed in order to improve switching speed, but the rationale is always the same. Figure 5.48 shows the behavior of the main nodes; depending on the result of the comparison of the input nodes, the latch changes its state or not.

State-of-the-art devices are very demanding as far as initialization of the circuitry is concerned. The POR circuit must be very fast, must work at low voltage,

i.e. the threshold is set around 1.5 V, because such memories operate at a minimum voltage of 1.8 V (and even less than that, in a near future) and therefore, at this VDD value, the circuitry must already be initialized. A certain amount of hysteresis is required as well, because the security margin is greatly reduced by the decrease of both the supply voltage and the POR triggering value.

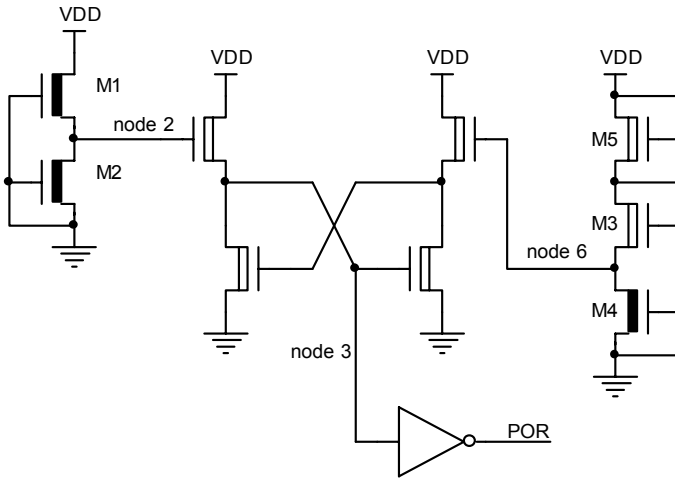


Fig. 5.47. NMOS-based POR circuit

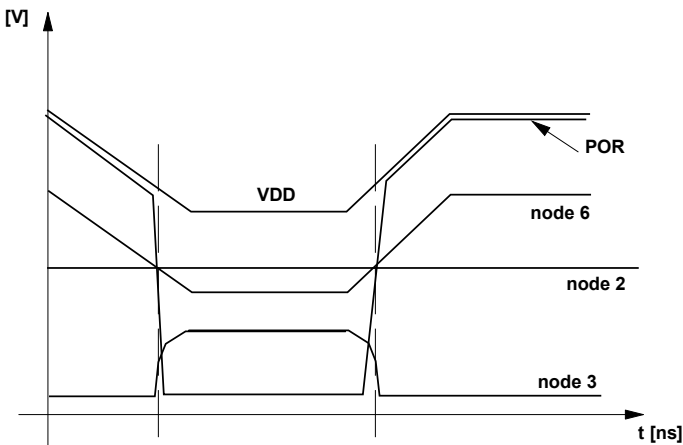


Fig. 5.48. Behavior of the nodes of the POR circuit of Fig. 5.47

Finally power consumption requirements account for a few microamperes, because the specification for power consumption in stand-by has become a top priority whenever the memory is used in a portable device.

Let's see a CMOS circuit that satisfies all these requirements.

In this case a comparator whose output, called **INTPOR**, directly influences the **POR** signal by making the comparison between the two signals, the reference and the supply follower. The scheme shown in Fig. 5.49 represents the main parts of the system. The task of the block called **VDDDIV** is to produce an output voltage with the same linear pattern of the **VDD**, but whose angular coefficient is scaled by a factor  $m$ ; this parameter is indeed a partition ratio, whose stability must be guaranteed by the circuit implementation.

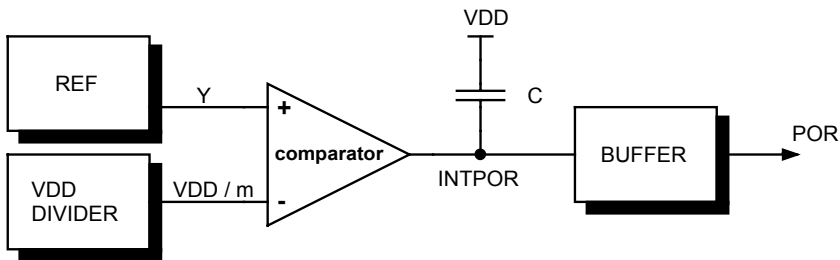


**Fig. 5.49.** Block scheme of the usual Power On Reset circuit

The function of the capacitor **C** is to allow the coupling of the **INTPOR** signal to the supply voltage during the ramp-up transient. The aim of the **BUFFER** is to de-couple the dynamic behavior of the output signal, whose range is equal to the entire **VDD** value, from the dynamic behavior of **INTPOR**, which can undergo a smaller variation; this improves the speed of the circuit. The main issue of a **POR** circuit realized as described above is consumption; in fact the scheme requires all the circuits to remain in the active state, to be able to re-generate the **POR** signal in case of a voltage drop, and this requires a non-zero current. Such a behavior is not acceptable in low-power devices, such as the single-supply memories, where a stand-by current virtually equal to zero is specified. To overcome this problem, the following requirements must be met:

1. Specific enable inputs must be introduced, so that several parts of the circuit can be switched off as soon as the **POR** signal has been generated;
2. The system must be able to autonomously exit the **OFF** state in case of unexpected supply voltage drop.

The main modifications required to achieve a zero-consumption condition are shown in Fig. 5.50. The schematics that follow depict the circuit implementation of the single blocks discussed before.

Aim of the scheme depicted in Fig. 5.51 is to provide a voltage level, tapped at the **Y** node, which is stable enough with respect to the possible variations in temperature, process parameters and supply voltage.

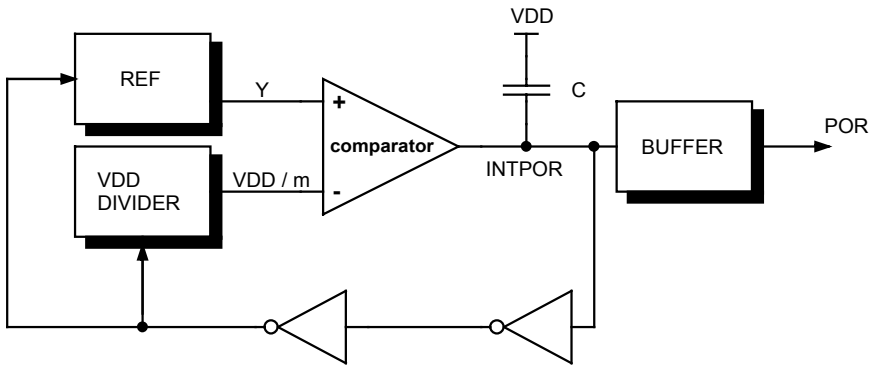


**Fig. 5.50.** Modified scheme of the POR to eliminate power consumption

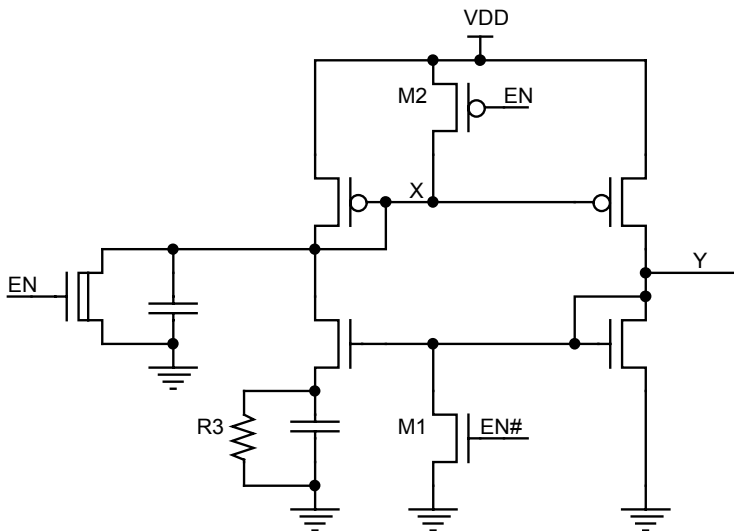


**Fig. 5.51.** Reference voltage generator

The circuit that partitions the VDD is realized using a resistive partition (Fig. 5.52); the two inverters are used to short-circuit the two resistors R1 and R2 to ground when the Y signal is low.

In Fig. 5.53 the circuit that compares the partition of the VDD and the reference value at the Y node is shown: the implementation is based on a differential stage whose output is suitably amplified by the two inverters I2 and I3 that can be brought to a zero-consumption stage by M3 as soon as the steady state has been reached.

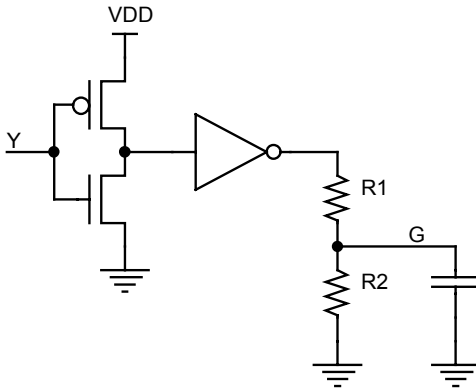


**Fig. 5.52.** VDD voltage divider

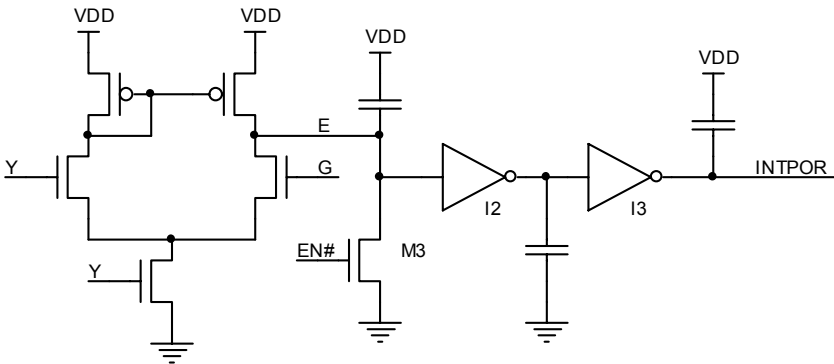


**Fig. 5.53.** The comparator

## 5.11 Analog Switch

The title indicates that the following text discusses the usage of the MOS transistor as a pass transistor (Fig. 5.54), and also the issues related to the coupling between gate, source and drain, which becomes evident in high frequency.

These subjects are dealt with in many books, so we would like to discuss the usage of the transistors as a switch in DC conditions by means of an example.

Dual voltage Non-Volatile memories (Flash and EPROM) have two pins for two different power supplies: one is the usual VDD, the other, known as VPP, must be present when the device is either programmed or erased. In any other condition it can be assigned any value between zero and a maximum as stated in

the specification. The problem is that the two supplies must converge inside the device to a common node, as the supply for some circuits (Fig. 5.55).

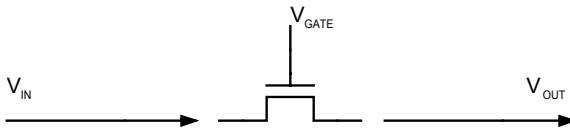


**Fig. 5.54.** MOS transistor used as a switch or pass transistor

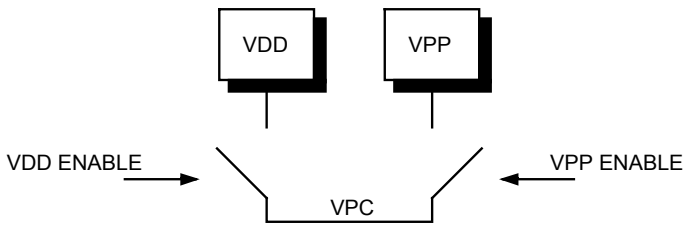


**Fig. 5.55.** The ideal VDD, VPP switch

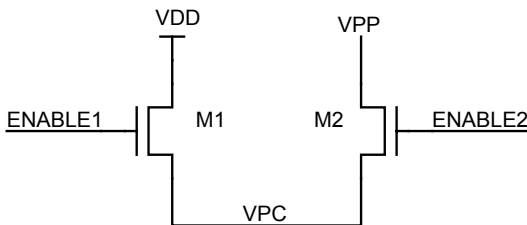


**Fig. 5.56.** Analog switch realized using NMOS transistors

This task would be easy if relay switches were integrated in silicon, which is not the case: therefore MOSFET must be used. In case of an NMOS process, n-channel transistors are the only choice (Fig. 5.56). In case M1 is on and M2 is off, VPC node cannot reach a value higher than power supply minus the threshold of the transistor (body effect included). In this way it does not work properly, therefore a circuit to boost the gates of M1 and M2 is needed; the section on the bootstrap describes how to achieve both VDD and VPP levels on the VPC node. In the case where p-channel transistors are available, a configuration as shown in Fig. 5.57 can be used. The question is: which is the right connection for the substrates of M1 and M2?

Remember that the PMOS transistor is inside a n-well and that both the source and drain junctions are of type  $p^+$ ; to avoid forward biasing of the diode towards the substrate, the potential of the latter must always be the highest voltage present.

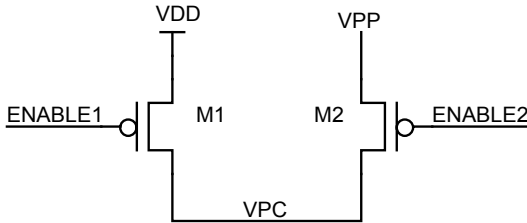


**Fig. 5.57.** Analog switch realized using PMOS transistors

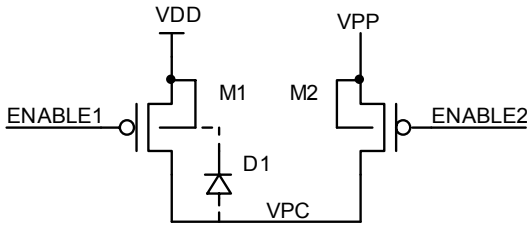


**Fig. 5.58.** The problem of forward biasing of the junctions in a PMOS switch

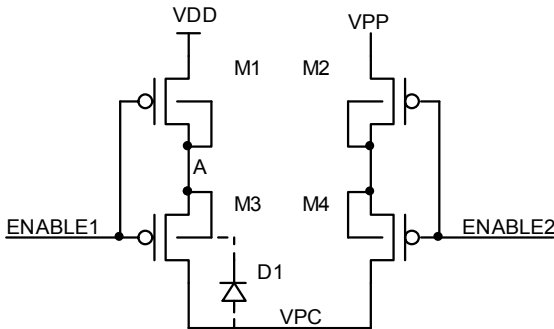


**Fig. 5.59.** A first attempt to isolate the n-wells of the PMOS

Suppose the substrate is connected as shown in Fig. 5.58: if VPC is equal to VDD, everything works fine; if VPC is equal to VPP, D1 diode is forward biased and nothing works. If the two n-wells are connected towards VPC node, the opposite situation occurs: in case of VPC equal to VDD, the junction of M2 towards VPP is forward biased.

Let's see if a proper control of the n-wells can solve the issue, modifying the structure as shown in Fig. 5.59. When both M1 and M3 are off while both M2 and M4 are on, ENABLE1 is at VPP and ENABLE2 is at ground, then the situation is under control for M2 and M4, while the n-wells of M1 and M3 (node A) are floating. The potential of node A, to which a parasitic capacitance is certainly associated, is initially at VDD (i.e. when ENABLE signals are in the opposite state). Parasitic diode D1 forces  $V_A$  to VPP minus the threshold voltage of the diode. The risk is related to charge injection into the substrate that could trigger latch-up effects. As soon as VPC is no longer needed at VPP, M2 and M4 are turned off while M1 and M3 are turned on. When M1 is on, node A (floating) is connected to VDD and the remaining charge stored in the parasitic capacitance is discharged, but the transient condition is difficult to control, and also the two voltages are externally controlled by the user. For example, voltage on VPC might be higher than VDD, even for a short time, when the read circuitry is on and the modify circuitry is off, thus stressing the matrix cells. The configuration shown in Fig. 5.59 is used without problems in the most recent Flash memories, where the VPP is not provided by the user, but rather it is internally generated by charge pumps. In this way the transient is controlled by the internal logic, and the user cannot inadvertently cause any improper biasing conditions.

**Problem 5.9:** What happens if the n-wells of M1, M2, M3 and M4 are connected to the opposite side?

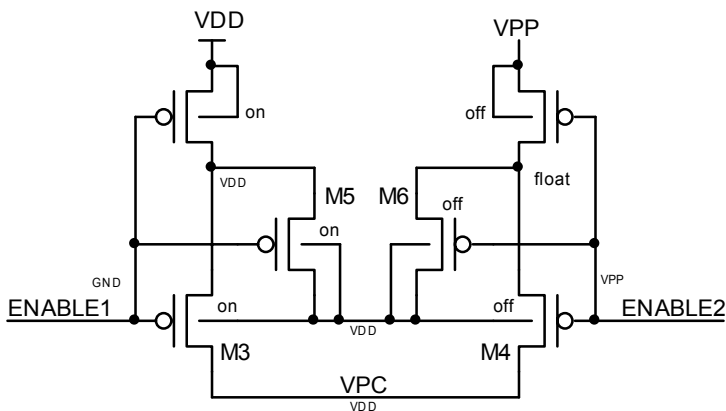


**Fig. 5.60.** PMOS switch when VPC = VDD. The states of the various transistors and the voltage of the n-wells are shown

The configuration shown on Fig. 5.60 solves the issue by controlling the substrate of the M3 and M4 p-channels, which present forward biased diodes and which could therefore charge the nodes of the switch to undesired potentials.

In Figs. 5.60 and 5.61 the state of every transistor and the voltages of the various nodes are shown for the case VPC = VDD and VPC = VPP respectively.

Thanks to the introduction of M5 and M6, no floating n-well is present. This solution avoids the issues of control on the transistors without decreasing the care posed in the layout of the solution, since floating nodes might inadvertently trigger a latch-up.

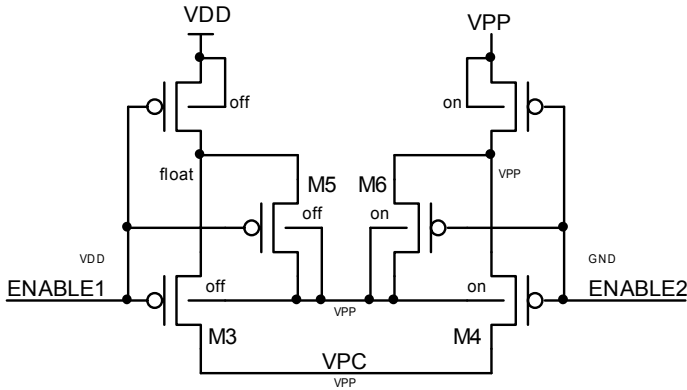


**Fig. 5.61.** PMOS switch when  $V_{PC} = V_{PP}$ . The states of the various transistors and the voltage of the n-wells are shown

The swing of the ENABLE signals is in the range between ground and VPP. Level shifter circuits similar to the one described in Sect. 5.9 can generate these signals.

## 5.12 Bootstrap

The word bootstrap literally means “to pull yourself up by your boots”, that is to be able to rise from the ground by grasping the straps of the boots! In order to better understand this new concept, which is widely used in NMOS technology, let’s recall how a capacitor is charged when a NMOS buffer is used (Fig. 5.62). A NMOS buffer cannot be realized using a depletion transistor as pull-up, because it would not be possible to turn it off.

**Problem 5.10.** Better elaborate the previous statement.

If  $V_{PD}$  is GND (M2 turned off) and  $V_{PU}$  is at VDD, then M1 is turned on. At the beginning,  $C_L$  capacitor is discharged and the potential of the OUT node is at ground. Initial  $V_{GS}$  of M1 is VDD as well as its  $V_{DS}$ .

The main issue is that the charge of the output capacitor obviously increases the voltage of the OUT node, thus progressively decreasing both  $V_{GS}$  and  $V_{DS}$  of transistor M1. The final outcome is that  $V_{OUT}$  does not reach the value of the power supply and that the charge of the capacitor does not occur at constant current, since the current decreases as time elapses.

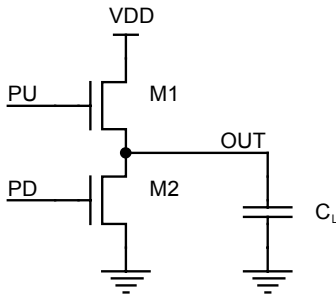
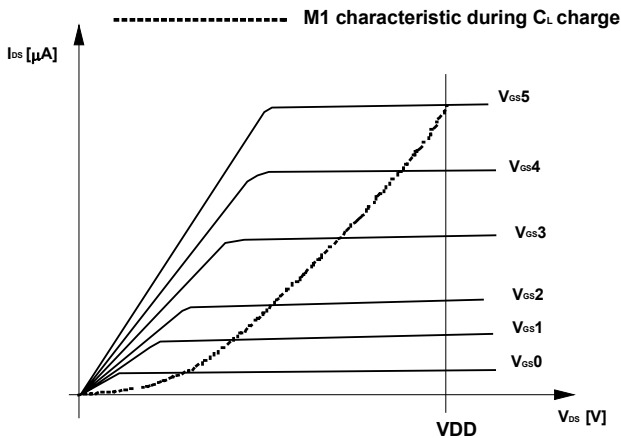


Fig. 5.62. NMOS output stage

Fig. 5.63.  $C_L$  charge characteristic

We have already seen in Fig. 3.7 (see Fig. 5.63 here reported for simplicity) the characteristics of the transistor M1 and the load curve for the charge of the capacitances. This kind of buffer does not allow having a large output swing and at the same time it limits the speed of the charge of the capacitance because current decreases with time: here comes the concept of bootstrap.

Figure 5.64 shows only the pull-up of Fig. 5.62, and capacitor  $C_B$  has been added between the gate and the source of M1.

Let's assume to have "precharged" node Y to the value of the supply voltage and then to have disconnected it from a conductive path, so that  $C_B$  cannot be discharged. Now let's see what happens when M1 starts to charge the output capacitance  $C_L$ . At the beginning we will have, as is with the previous case,  $V_{GS} = V_{DS} = VDD$ , being  $C_L$  discharged. M1 provides current,  $V_{OUT}$  increases,  $C_B$  maintains the voltage difference at its nodes (VDD) and therefore the voltage of the floating node Y tends to increase. The  $V_{GS}$  of M1 does not decrease while the output ca-

capacitor is charging, remaining, at least as a first approximation, constant. In reality, the characteristic of M1 is the dashed line in Fig. 5.65, where it is clear that we are not moving on a constant characteristic of  $V_{GS}$  because of the parasitic capacitor of node Y.

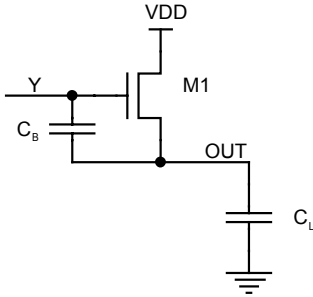


**Fig. 5.64.** NMOS output stage with bootstrap capacitor

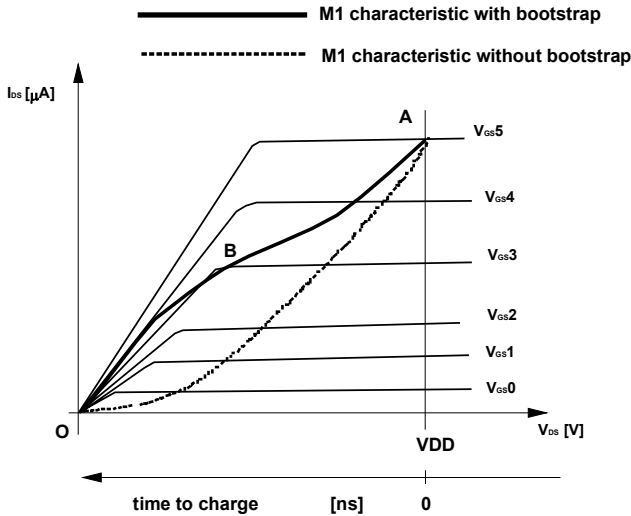


**Fig. 5.65.** Characteristic of M1 during the charge transient of  $C_L$

The characteristic shown in Fig. 5.65 is broken into two distinct sections: AB, which represents the part where M1 is in saturation and B0, where M1 is in linear zone. On top of that, there is also the effect of the progressive decrease of  $V_{DS}$  voltage during the charge; moreover, the threshold of M1 varies due to body effect, therefore its current further decreases as time goes by. The proper dimension-

ing of the bootstrap capacitance  $C_B$  is vital to the correct operation of the circuit; before discussing the criteria for the dimensioning, let's see an example of complete bootstrap circuit. Figure 5.66 shows a driver of a row decoder for a NMOS memory. We recommend again redrawing the circuit on a sheet: redrawing helps understanding connections and it is the first step in understanding the circuit itself.

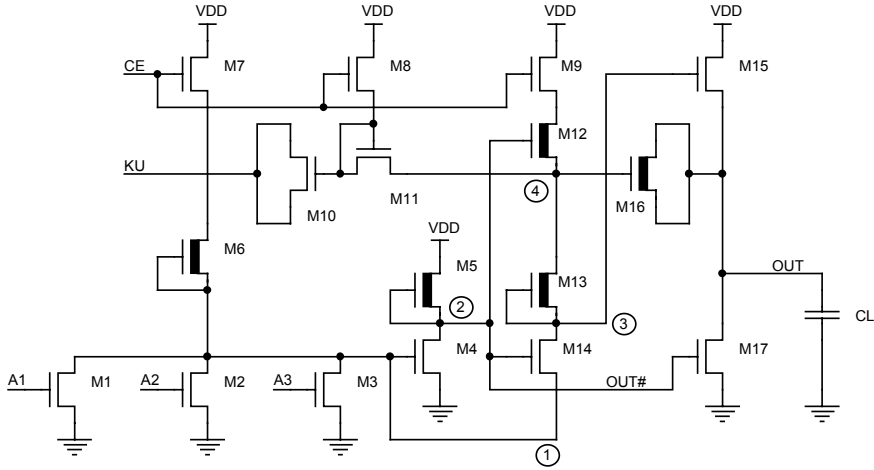


**Fig. 5.66.** An example of NMOS row decoding circuit where bootstrap is implemented

Aim of the circuit is to set the output signal OUT to high when A1, A2 and A3 are low. First of all, let's observe that the task of the transistors M7, M8 and M9 is just to enable circuit operation through the CE# signal, while M6 is the load of the NOR composed by M1, M2 and M3.

Let's assume that A1 be high and the circuit enabled (CE# low). M14 is turned on, the voltage of the node 1 is low and the one of the node 2 is at VDD because M4 is turned off. Therefore  $V_3$  is at ground while  $V_4$  is high but not at VDD because of the partition between M12 and M13. M15 is turned off and M17 is turned on, therefore  $V_{OUT}$  is at ground. If the condition  $A1 = A2 = A3 = 0$  occurs,  $V_3$  increases,  $V_2$  decreases and M14 is turned off. Figure 5.67 shows the final stage of the circuit for a better understanding of the biasing conditions.

Node 3 increases in voltage as node 4, because M13 behaves as a resistor. Transistor M15 has the gate at a positive voltage and it starts providing current, charging  $C_L$ . Transistor M16 acts as bootstrap capacitance, transferring the voltage difference acquired by  $C_L$  to node 4 and, automatically, to node 3. Node 3 is floating as long as following relation holds

$$V_4 > V_2 - V_{T,M12} \quad (5.63)$$

The task of M9 is to deliver to node 4 the charge required during the idle state of the circuit, when  $V_{OUT}$  is at ground.

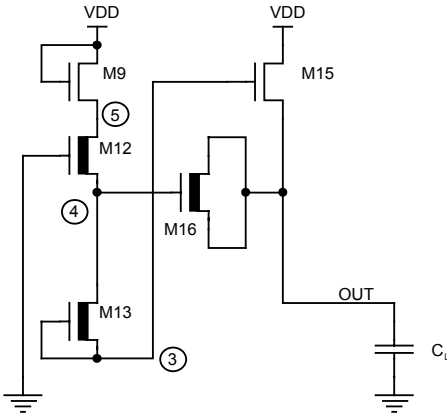


**Fig. 5.67.** The final stage of the circuit of Fig. 5.66 where the biasing during bootstrap phase is highlighted

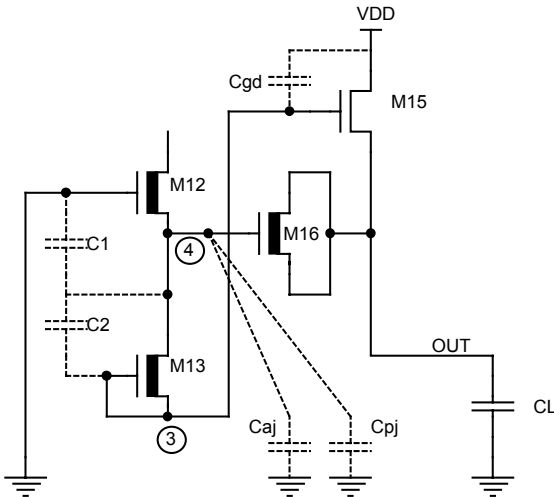


**Fig. 5.68.** The final stage of the circuit shown in Fig. 5.66 where parasitic capacitances are highlighted

Figure 5.68 shows the parasitic capacitances at node 4:  $C_1$  and  $C_2$  are the gate/source capacitances of M12 and gate/drain capacitances of M13 respectively, while capacitances  $C_{aj}$  and  $C_{pj}$  are area and perimeter capacitances of node 4, given by the diffusions of M12 and M13. The dimensioning of the bootstrap capacitance, i.e. of transistor M16 that is capacitor-connected here, is a function of the parasitic capacitances.

**Problem 5.11:** Compare the different types of capacitors that can be used and compare their characteristics.

Let's call  $C_p$  the capacitance equivalent to the network of the parasitic ones of node 4 and  $C_B$  the boost capacitance (M16). Since node 4 is isolated, the initial charge  $Q_i$  that is present on the capacitors is preserved after bootstrap occurs. Assuming that node 4 has been precharged to a value equal to  $V_{DD}/n$ , we can write:

$$Q_i = (C_p + C_B) \frac{V_{DD}}{n} \quad (5.64)$$

$$Q_f = C_p V_f + C_B (V_f - V_{DD}) \quad (5.65)$$

$$Q_i = Q_f \quad (5.66)$$

$$V_f = \frac{V_{DD}}{n} + \frac{C_B}{C_p + C_B} \cdot V_{DD} \quad (5.67)$$

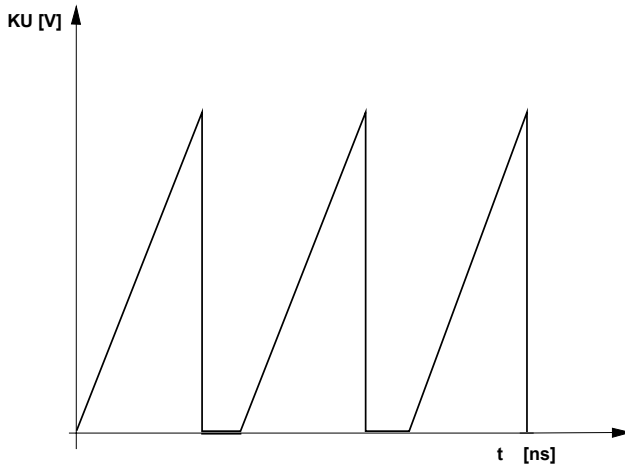


Fig. 5.69. KU signal

We made the assumption that the output node OUT can get to the power supply at the end of the bootstrap phase; if we want the voltage difference on the capacitor to be integrally preserved at the end of the bootstrap, we must have  $C_p = 0$ .

Let's assume that we want to preserve, after the bootstrap, the 90% of the voltage difference on M16. Starting with node 4 precharged at 3 V, we can write that

$$V_f = 5V + 3V \cdot 0.9 = 7.7V \quad (5.68)$$

Using this value of  $V_f$  in Eq. (5.67), it results that  $C_B$  must be approximately 16 times larger than  $C_p$ .

The only transistors that we have not mentioned yet in Fig. 5.66 are M10 and M11. Several junctions are connected to node 4, which can be discharged by their leakage currents; moreover, in case of EPROM devices, the external light passing through the window can help discharging the floating nodes.

The task of signal KU and of transistors M10 and M11 is to “supply” node 4 with charge. In order to achieve a good efficiency, the oscillator that generated KU has a long rise time and a very short fall time (Fig. 5.69). M10 is used as a capacitor, while M11 is diode-connected, to prevent node 4 from discharging during the falling phase of KU.

Let’s now examine the most common types of bootstrap.

### 5.12.1 PUSH-PULL Bootstrap

Precharge reaches node 3 during the time interval between the rise of node 1 and the rise of the OUT node; the delay introduced by the inverter placed between the input and the output is exploited to bias node 2. The drawback of this kind of circuit is that in case of a glitch on the input signal, occurring while the OUT node is rising, the bootstrap capacitor could be partially discharged, thus reducing the charge efficiency on  $C_L$ . Following schematic can be used to prevent such situation from happening.

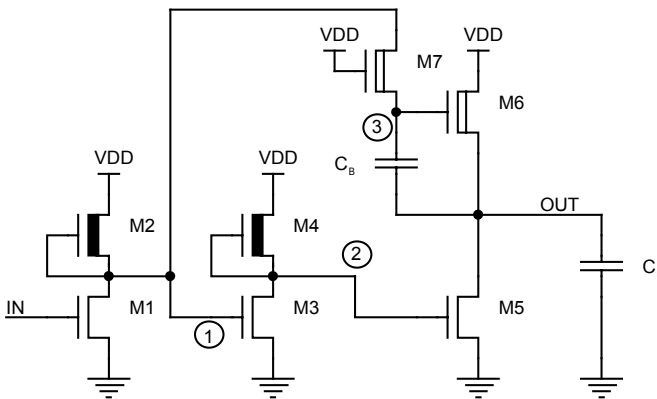


Fig. 5.70. PUSH-PULL Bootstrap

### 5.12.2 PUSH-PULL Bootstrap with Anti-Glitch

Precharge of node 6 takes place in the time interval between the rise of node 1 and the fall of node 4. Therefore as soon as  $V_{OUT}$  starts rising, the precharge pass transistor M1 is for sure OFF, thus preventing an incidental discharge of  $C_B$  in case of glitch on the input signal.

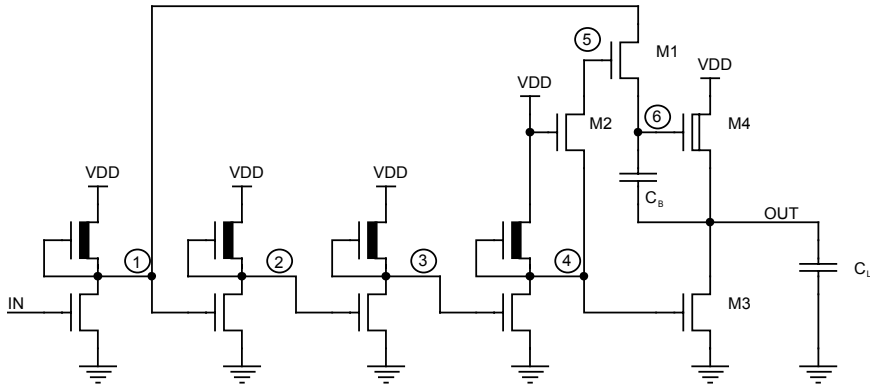


**Fig. 5.71.** PUSH-PULL Bootstrap with anti-glitch

### 5.12.3 PUSH-PULL Bootstrap for a Large Load

In case the output load is large, it is better to separate the output stage into two parts to speed up the rise of the bootstrapped node without the need of waiting the output node, thus allowing a higher charge current (Fig. 5.72).

The initial scheme of Fig. 5.66 is not immune to discharge of  $C_B$  in case of glitch on the input and, with respect to the versions of Fig. 5.70 and 5.71, it exploits the precharge for all the time when the voltage of node 1 is low, instead of a very short time equivalent to the delay of one or two stages of inverters.

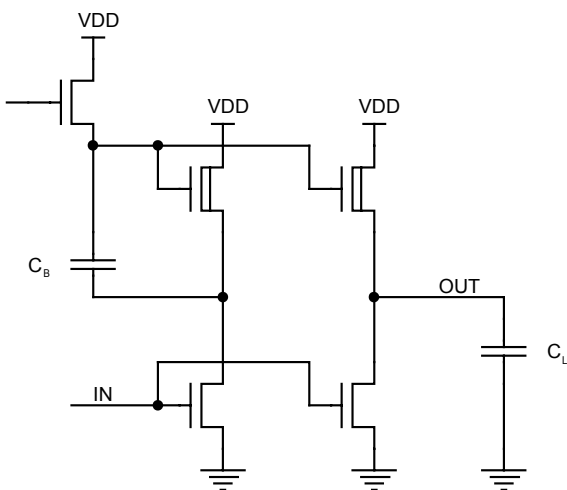


**Fig. 5.72.** PUSH-PULL Bootstrap configured for a large load

## 5.13 Oscillators

In electronic systems design, it is often necessary to generate periodic waveforms, triangular, square-shaped, sinusoidal etc. The circuits used to generate such waveforms are called oscillators; some practical implementations are shown in this chapter.

Ring oscillator is one of the well-known schemes. It is composed of a feedback chain of an odd number of inverters, as shown in Fig. 5.73. A simple analysis of the schematic shows that this circuit has a single equilibrium point, corresponding to the situation where all the inverters are biased on the trigger threshold. Under these conditions, every single inverter is biased in a region of high voltage gain, thus determining a high loop gain. As a consequence, the equilibrium point is unstable and, even just because of noise, the circuit in Fig. 5.73 starts oscillating producing a periodic waveform.

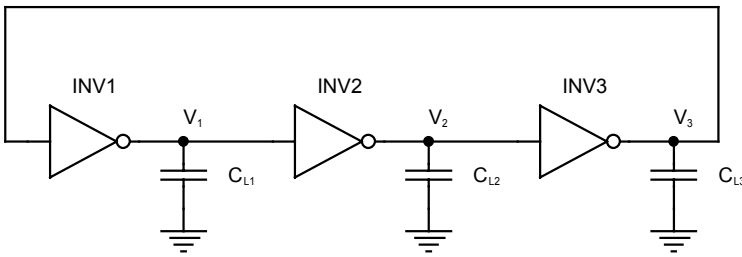


**Fig. 5.73.** Ring oscillator

Let's now analyze the behavior of the circuit gate by gate. When the output voltage  $V_1$  of the first inverter toggles from the logic state low to high, the output of the second inverter goes low. The time difference between the moments when  $V_2$  is  $V_{DD}/2$  and when  $V_1$  had the same value is the propagation delay  $\tau_2$  of INV2. When the output of the second inverter goes low, the output of INV3 goes high with a propagation delay  $\tau_3$ . Summing up, each inverter of the chain triggers the commutation of the following one, and the last inverter acts on the first one thus guaranteeing the continuity of the oscillation. The oscillation period  $T$  can be calculated using the propagation delays of the single inverters. Assuming the load capacitances is equal and all the inverters are identical, therefore having the same propagation delay  $\tau_D$ , we can write that

$$T = \tau_1 + \tau_2 + \tau_3 + \tau_1 + \tau_2 + \tau_3 = 6\tau_D \quad (5.69)$$

Given a generic number  $p$  of inverters that compose the ring oscillators, oscillation frequency is therefore equal to

$$f = \frac{1}{T} = \frac{1}{2p\tau_D} \quad (5.70)$$

Owing to this feature, ring oscillators are often used to benchmark the speed of a given technology. In fact, it is enough to integrate an oscillator of this type and accurately measure the oscillation frequency; dividing by the number of inverters in the chain, it is possible to determine the mean propagation delay of a given technological process. Because of the high speed of the single gates in present technologies, it is necessary to implement many stages in the chain to get frequency that are low enough to be measured; otherwise, it is possible to decrease the frequency dimensioning the load capacitors  $C_L$  differently.

Let's now analyze a NMOS oscillator, used to re-integrate the charge lost by the bootstrapped node because of the leakage: this circuit is responsible for the generation of the KU signal shown in Fig. 5.69. The main feature of this oscillator is that the output voltage must have a slow rise time and a very fast fall time.

Figure 5.74 shows the schematic of the circuit and related block diagram.

It is immediately evident that the last two stages are supplied by a  $V_{PP}$  voltage, greater than  $V_{DD}$ , because the output signal will be then applied to other circuits that are supplied by  $V_{PP}$  as well (during program operation). The block diagram shows a high voltage detector, a Schmitt trigger and some inverters; of course, in order to allow the oscillation, an odd number of inversions must be present. Last inverter acts as a buffer to provide the required output current.

High voltage detector compares  $V_{DD}$ , applied to the gate of M1, with the gate voltage of M2: comparison is done in current, carefully selecting the size of the two transistors.

**Problem 5.12:** Find out the right size for both M1 and M2.

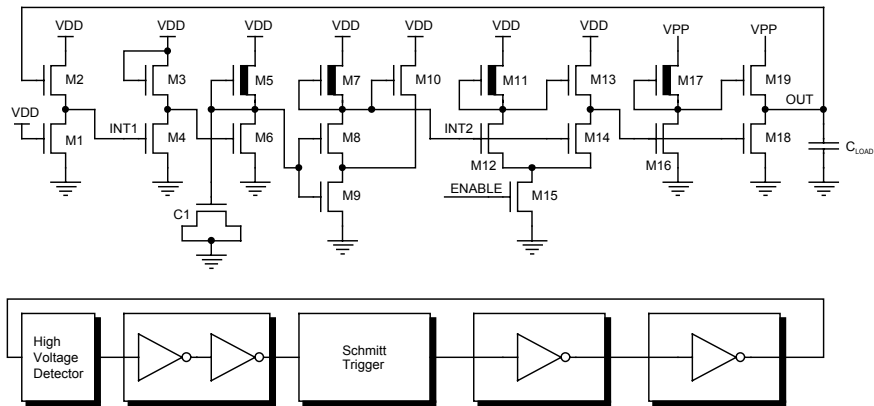


**Fig. 5.74.** An example of NMOS oscillator. The shape of the output signal is a voltage ramp, featuring a slow rise time and a fast fall time.

In order to generate the ramp on the output node,  $C_{LOAD}$  is charged using a constant current. When M19 is turned on, it is virtually diode-connected; working in saturation region, it is equivalent to a current generator. Generation of the output

ramp is then achieved by keeping M19 turned on for the chosen time, paying attention that the output voltage does not reach a value high enough to turn off M19, thus modifying the waveform of the output signal.

$V_{OUT}$  voltage is fed back to the input, where the comparison takes place: when the output voltage has reached the desired value, node INT1 turns on M4, that allows the charge of C1 through M5. At this point the trigger toggles, node INT2 goes low and the fast discharge of the output node through M18 is triggered: M18 is dimensioned in such a way that it can sink the discharge current completely.

Once the voltage of the output node has dropped down to ground, M1 is turned off. Before a new charge of  $C_{LOAD}$  can take place again, re-charge of C1 and the toggle of the trigger must occur. The latter has the advantage that it filters out undesired toggling induced by the noise on power supplies, since it introduces a hysteresis. Finally, M15 is an enable transistor that, if turned off, keeps the output node low. Figure 5.69 shows the qualitative behavior of the output node.

**Problem 5.13:** Why the enable transistor M15 is not inserted in the last stage, but it is in the previous one? Every time a choice is possible, options are never equivalent: one is always the best for our aims.

The other NMOS oscillator in Fig. 5.75 produces an output waveform that is not exactly squared, but whose edges are steep. Unlike the previous one, output load capacitance is smaller and therefore it is possible to exploit bootstrap techniques to obtain an oscillator over the full power supply swing. The study of this circuit is left to the reader. Enjoy!

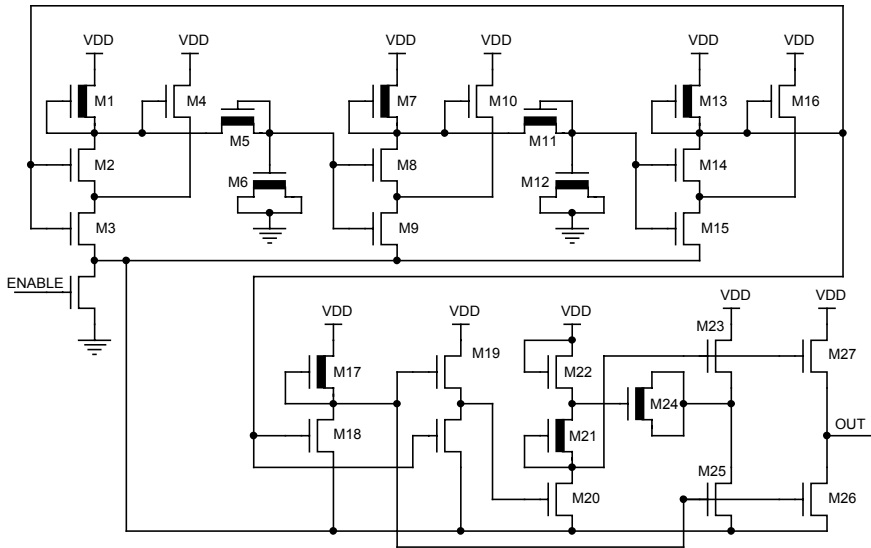


**Fig. 5.75.** NMOS oscillator with squared-waveform output



## 5.14 Circuits to Detect Third Level Signals

Communication between the internal and the external of the device takes place through the pads that are driven by structures composed of inverters, that are therefore able to accept logic values equal to either ground or VDD (CMOS interface).

How can we do, for instance, to use the addresses without triggering a new read, when any logic combination we choose maps to an address in the memory space? It is possible to design a buffer that is capable of detecting three voltage levels: ground, i.e. CMOS logic “0”, power supply, i.e. CMOS logic “1” and then a voltage level higher than power supply, that we call third level.

For a device operating with a VDD equal to 5 V, the value of the third level is equal to VPP, while for 3 V-operating devices, third levels allows for lower values, around 10 V. Third levels are used by the manufacturer to place the device in those test modes that are precluded to the user. Anyway the customer has some user modes, described in the specification (like byte identifier read, temporary sector un-protection etc.) that allow for the use of third levels. The voltage value depends on both the device and the technological process used.

When a third level is detected, i.e. an input whose voltage value is much higher than power supply, a path different from the normal input buffer is activated.

The main item to take into account when designing a third level circuit is that it must not cause power consumption, or leakage towards ground, otherwise it can invalidate the normal behavior of the input buffer into which it is inserted, and whose pad it shares. A second item is related to robustness with respect to ElectroStatic Discharge, that we will cover later, according to which it is not recommended to connect drains of n-channel transistors directly to power supplies.

A possible approach for a NMOS technology is shown in Fig. 5.77, where we can recognize the same input stages of the oscillator shown in Fig. 5.74, used for the same purpose. The advantage of this kind of third level buffer is that the input is on the gate, so that we can neglect all the issues related to both ESD protection and latch-up, in case of CMOS. Unfortunately, since the circuit is based on the ratio between two currents, power consumption is not zero.

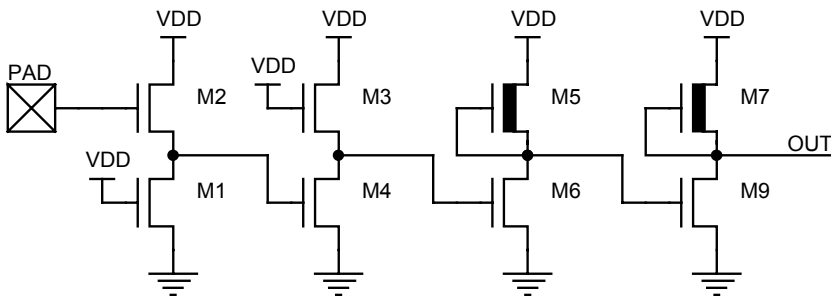


Fig. 5.77. NMOS third level buffer

A typical buffer for a CMOS device is shown in Fig. 5.78: the chain of p-channel diodes provides the triggering threshold.

The voltage of PAD must be at least equal to the supply voltage plus  $|V_{T,P}|$  so that M1 can be turned on. But at this point the diode M2 drops a threshold and therefore the supply of the inverter does not allow M3 to win over M4. On the contrary: to have M3 winning over M4, apart from the dimensions that must be chosen for this aim, i.e. M4 must be resistive and M3 conductive, the voltage on the source of M3 must be at least equal to VDD. Thus the voltage on the pad must be at least equal to  $VDD + 2|V_{T,P}|$  plus the contribution of the given body effect plus a quantity that depends on the size ratio between M3 and M4. Assuming that VDD is equal to 5 V, the triggering threshold is around 8 V. If power supply increases, a greater voltage must be applied on PAD in order to turn the circuit on.

Finally, Fig. 5.79 shows the scheme of the third level buffer used for a CMOS device, composed of a chain of p-channel diodes that is directly connected to the pad and that is clamped to ground by a n-channel diode. Voltage on the gate of the first inverter is usually too low to cause the overall output to go high.

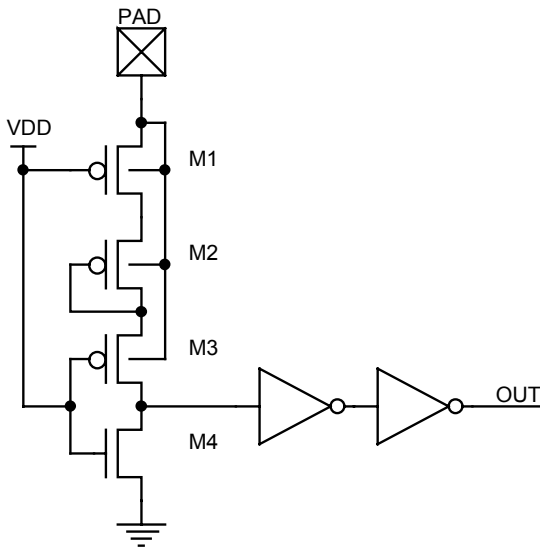


**Fig. 5.78.** CMOS third level buffer whose triggering threshold depends on the supply voltage

When the voltage value on PAD is greater than five p-channel  $|V_{T,P}|$  plus one n-channel threshold voltage, node F can go to ground turning on the third level. In this case the value of the voltage required on the pad to turn on the third level is independent from VDD power supply.

The presence of two n-channel transistors, M6 LVS and N7 natural, as pull-down of the chain, is necessary to guarantee that the following inverter is turned off when the third level is not applied. In fact the n-channel of the inverter INV1 is

of the same kind of M6 and, therefore, without N7 the input of INV1 would be exactly at the triggering threshold of the inverter itself, and power consumption might take place.

This case would be almost equal to a floating node condition, i.e. un-driven, at the mercy of spurious couplings.

**Problem 5.15:** Analyze the last two circuits illustrated above and decide whether it is better that the value of the triggering threshold for a third level buffer be a function of VDD power supply or not.

**Problem 5.16:** Analyze the nature of a floating node, define it and evaluate which damages it could cause.

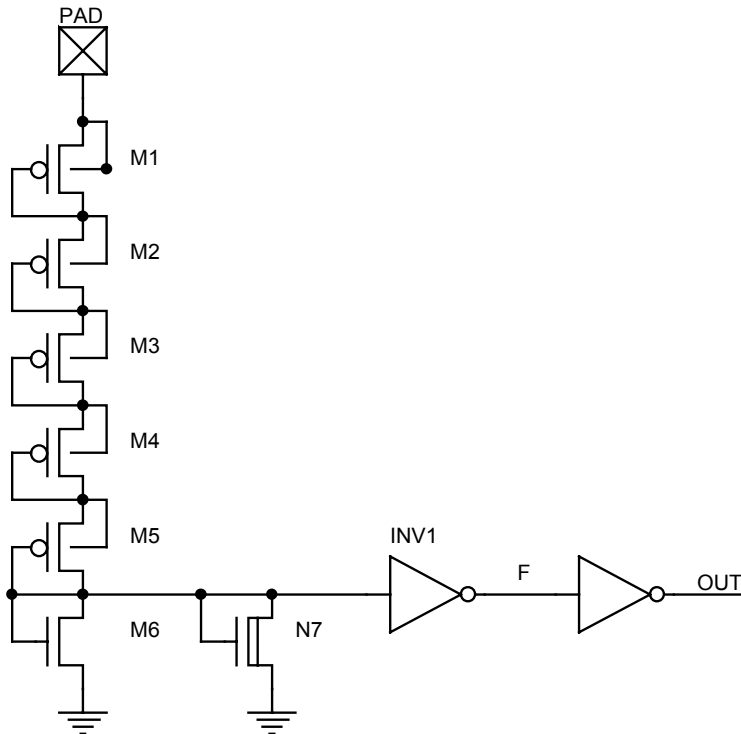


Fig. 5.79. CMOS third level buffer independent from supply voltage

## 5.15 VDD Low Detector

Sensibility of the flash cells calls for a careful control of the voltages applied to the matrix, in order to ensure that improper voltages be applied at the wrong times, thus damaging the cells or changing their threshold voltage. For this reason it is important to control the value of the power supply during modify operations.

If, for instance, we want to program, then program path to the matrix must be enabled and the pumped voltages are generated starting from power supply; during the operation, supply voltage must remain at the right level, otherwise program operation might not perform correctly.

For this reason, inside the device there is a circuit that controls the value of power supply and, in case it falls below a given value, a signal that stops the charge pumps and inhibits the most dangerous paths towards the matrix is activated.

The circuit can be realized using a resistive partition, whose task is to follow VDD, and a comparator whose non-inverting input is connected to the output signal of the band-gap. The output of the comparator is then sent to a Schmitt trigger to filter out spurious commutations.

**Problem 5.17:** Realize the circuit that has been just suggested.

## Bibliography

- B.K. Ahuja, "An improved frequency compensation technique for CMOS operational amplifiers", IEEE J. Solid-State Circuits, vol. SC-18, pp. 629-633, (Dec. 1993).
- L.A. Akerst, "An analytical expression for the threshold voltage of a small geometry MOSFET", Solid State Electronics, vol. 24, pp. 621-627, (1981).
- M. Annaratone, Digital CMOS Circuit Design. Kluwer Academic Publishers.
- J.C. Bertails, "Low frequency noise considerations for MOS amplifiers design", IEEE Journal of Solid State Circuits, Vol. SC-14, No. 4, pp. 773-776, (August 1979).
- J.Y. Chen "CMOS The emerging VLSI technology", IEEE Circuits and Device Magazine, pp. 16-31, (March 1986).
- E.J. Dickes, D.E. Carlton, "A Graphical analysis of the Schmitt trigger circuit", IEEE Journal of Solid State Circuits, vol. SC-17, No. 6, pp. 1194-1197, (December 1982).
- Edit by P.R. Gray, D.A. Hodges, R. W. Brodersen, "Analog MOS integrated circuits", IEEE Press, (1980).
- P.R. Gray, R.G. Meyer, "MOS Operational Amplifier Design- A Tutorial Overview", IEEE Journal of Solid State Circuits, Vol. SC-17, No.6, pp. 969-982, (December 1982).
- R. Gregorian, G.C. Temes, Analog Mos integrated circuits for signal processing, J. Wiley & Sons, (1986).
- T.P. Haraszti, CMOS Memory Circuits. Boston, MA: Kluwer Academic Publishers, ch. 5, (2000).
- L.G. Heller and W.R. Griffin, "Cascode voltage switch logic: a differential CMOS logic family", in 1984 IEEE Int. Solid-State Circuits Conf. Dig. Tech. Pap., pp. 16-17, (Feb.1984).
- A. Kuma, "CMOS Circuit optimization", Solid State Electronics, vol. 26, No. 1, pp. 47-58, (1983).
- C.M. Lee, Ellen W. Szeto, "Zipper CMOS", IEEE Custom Integrated Circuits Conference, pp. 236-239, (1986).
- T. Mano, J. Yamada, J. Inoue, and S. Nakajima, "Circuit techniques for a VLSI memory," IEEE J. Solid-State Circuits, vol. 18, no. 5, pp. 463-469, (Oct. 1983).
- W.L. Martino et al, "An on-chip back-bias generator for MOS dynamic memory", IEEE Journal of Solid State Circuits, Vol. SC-15, No. 5, pp. 820-826, (October 1980).

- 
- K. Natori, "Sensitivity of dynamic MOS Flip-Flop sense amplifiers", IEEE Transaction on Electron Devices, Vol. ED-33, No. 4, pp.482-488, (April 1986).
- D. Senderowicz, J.H. Huggins, "A low noise NMOS operational amplifier", IEEE Journal of Solid State Circuits, Vol. SC-17, No. 6, pp. 999-1008, (December 1982).
- H. Taub, D. Shilling, Digital Integrated Electronics, McGraw-Hill, 1997.
- U. Tietze, Ch. Schenk, "Advanced electronic circuit", Springer-Verlag Berlin Heidelberg New York, (1978)
- Y.P. Tsividis, "Technique for increasing the gain-bandwidth product of N-M.O.S. and P-M.O.S. integrated inverters", Electronics Letters, vol. 13, No. 14, 7th, pp.421-422, (July 1977).
- Y.P. Tsividis, "Design consideration in single-channel MOS Analog Integrated Circuits- A Tutorial", IEEE Journal of Solid State Circuits, Vol. SC-13, No.3, pp. 383-391, (June 1978).
- E. Vittoz et al., A Collection of CSEM Papers, Electronic Design Books, Penton Publishing, (1995).

## 6 Layout

The aim of this short chapter is to present the main concepts that must be understood in order to read a layout, i.e. the translation of an electrical schematic into polygons. Learning to read a layout is a long and tedious process, which might take years to become proficient, because of the many clever methods used in developing the layout itself. Undoubtedly, it can be said that the success of a device lies in the combined skill of both the designer and the layout engineer.

### 6.1 Custom Layout

Designing and manufacturing integrated circuits is a very complex activity; despite the powerful machines used and the thousands of lines of software written to make such machines work, making of a chip is nevertheless a craftsman work.

Designers, layout engineers and technologists do not rely on automatic, error-free procedures; on the contrary, they become more and more similar to Renaissance Masters with their recipes and secrets orally handed down to their favorite pupils. The art of custom layout, i.e. hand-made, transistor after transistor, is probably the most astonishing activity for both the novice who observes a layout for the first time, discovering geometric arabesques and the cunning designer who cannot help being pleased observing the work to which he has contributed.

Layout artists or draftsmen translate transistors, resistors, capacitors, cells and everything on the schematics into polygons that, at the end, will constitute the photomasks to realize the device. Making of the layout requires a deep knowledge of the rules of the technological process; every layer maintains distance rules towards the other layers that are complied with and checked, by the draftsmen themselves, by means of a program known as DRC (Drawing Rules Check). At the end, another checking tool, known as LVS (Layout Versus Schematic), verifies the matching of the overall schematic with the layout.

Let's see some simple circuit and the corresponding layout.

### 6.2 A Three-Inputs NAND

Let's start and analyze the three-inputs NAND shown in Fig. 6.1. As we have already seen in the chapter on the process, the first layer that is diffused is the n-

well<sup>1</sup>; then there is the definition of the areas where a thin oxide is present, as is the case with the polysilicon gate or where no oxide is present because a contact with a junction must be done. The contact is realized making a hole in the oxide in order to reach the junction. The wires to propagate the signal are<sup>2</sup> metal1 and metal2, connected through the vias.

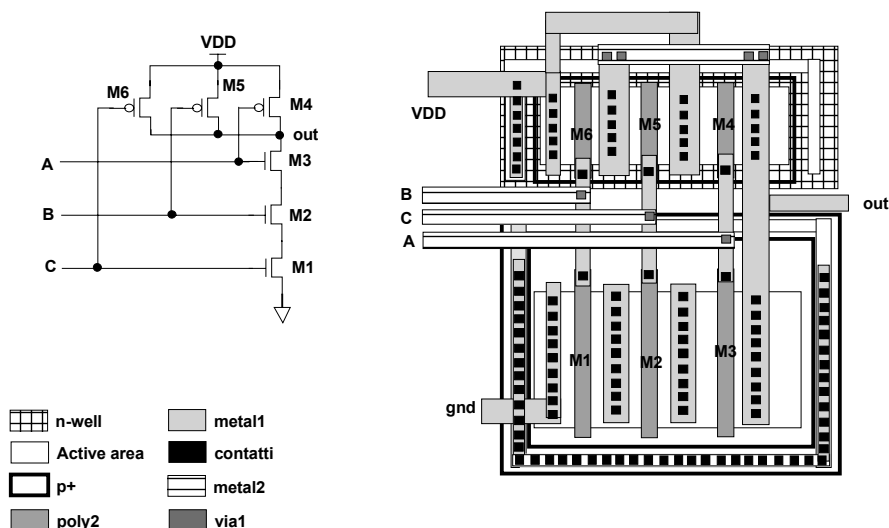


**Fig. 6.1.** Electric scheme and layout for a three-inputs NAND

**Problem 6.1:** Try and redo all the layouts shown in the present chapter using a single layer of metal.

Analysis shows the gates of the p-channels connected to those of the n-channels through a metal1 wire. It is not possible to connect the gates, for instance M1 and M6, in poly2, because the n-channel transistors are surrounded by a ring of active area that will be implanted  $p^+$ , in order to have a protection ring against the latch-up, a parasitic effect that we will describe afterwards. Poly2 cannot cross any active area; otherwise it would constitute a MOS transistor.

In the case of p-channel transistors, the ring for biasing and protection is interrupted on the side facing the n-channels. This allows additional degrees of freedom, but on the other hand it increases the risk of latch-up. The two situations must be evaluated case-by-case depending on the different requirements.

<sup>1</sup> It is important to remind that there is not a one-to-one correspondence between a layout layer and an operation (diffusion, deposition or attack). Some layers are obtained as logic combination of two or more layout layers.

<sup>2</sup> Throughout the layout examples, a two metal layer process is considered. Cheaper processes exist, where a single metal layer is available; on the other hand, the processes used for the most powerful microprocessors can have five or more metal layers

Among NMOS transistors, M1, M2 and M3, the active areas are contacted by metal1 polygons; the minimum distance between two poly strips is usually adopted, to decrease active area resistance. A special consideration must be done about contacts: in advanced processes, they are realized as equally sized squares<sup>3</sup>, to overcome lithography and filling issues. When the hole that will allow metal1 to get in contact with the active area is done, the contacts are first filled with a metallic compound, and then metal1 is deposited.

Figure 6.2 shows the issues related to contact filling. We can see in Fig. 6.2a the metal1 incorrectly filling the contact, since it is thinner near the edges, so that it could break because of either mechanical or thermal stress when current flows. A more robust procedure is to fill the contact, as in Fig. 6.2b, and then have the metal1 contacting the filling material, thus ensuring a more reliable connection.

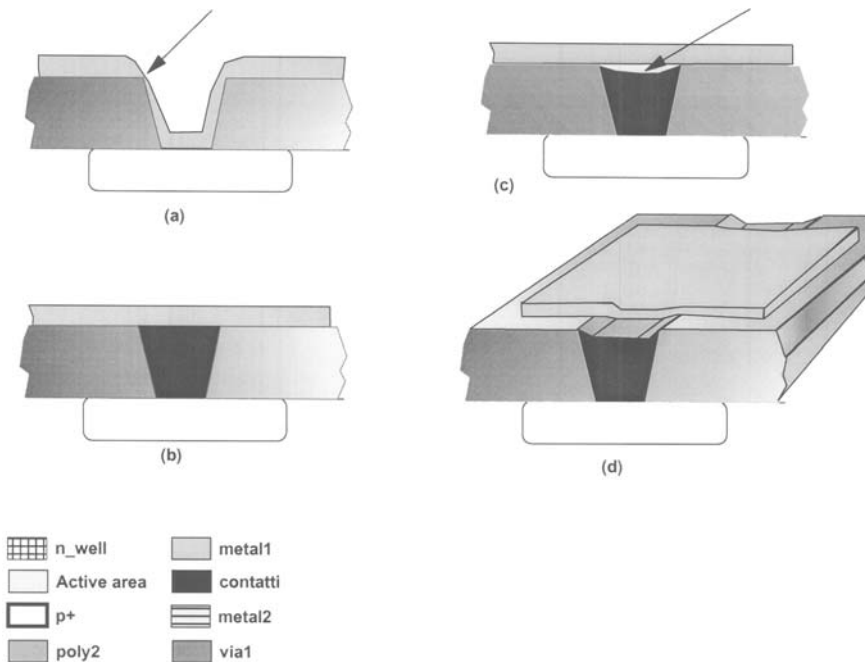


**Fig. 6.2.** Issues related to contact filling

The issue sketched in Fig. 6.2c is the filling of the contact that might leave a small depression in the center, thus worsening the contact with metal1. The usage of long, differently sized contacts would increase the probability of incurring a depression over the filling material and it would increase the difficulty of control during lithographic exposure (Fig. 6.2d).

<sup>3</sup> On silicon, they result in a round hole.

### 6.3 A Three-Inputs NOR

Figure 6.3 shows the layout of a three-inputs NOR. Same consideration as in the previous paragraph applies. For the sake of simplicity, the size of the transistors is not shown.

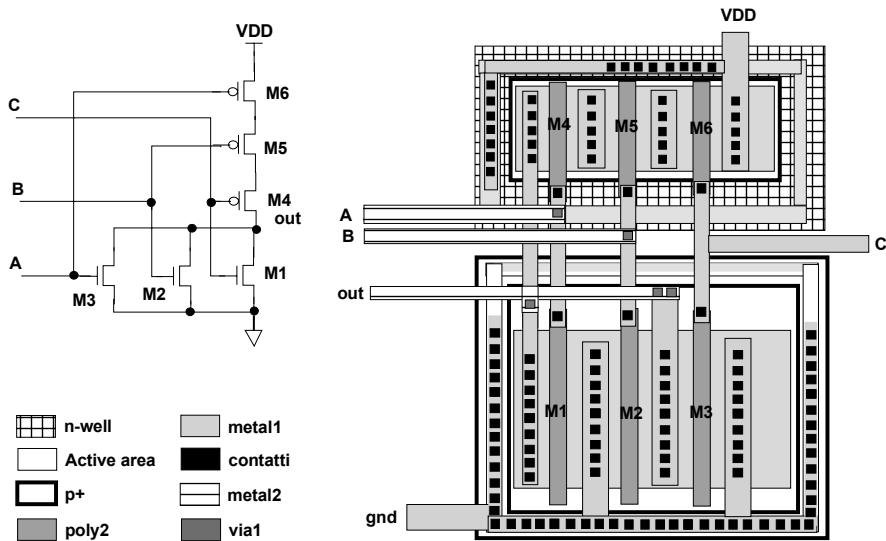


**Fig. 6.3.** Electric scheme and layout for a three-inputs NOR

**Problem 6.2:** Redraw the layout of Fig. 6.3 in such a way that every PMOS has its n-well connected to its own source, thus eliminating the corresponding body effect.

**Problem 6.3:** Analyze the pros and cons of the solution to Problem 6.2. Estimate the costs to eliminate body effect for each NOR and NAND.

### 6.4 An Interdigitized Inverter and a Capacitor

In several occasions it is required to draw transistors whose  $W$  is very large. In the case of output buffers, for instance, size might be greater than 1 mm. In these situations, it is not possible to draw a single, long polygon in poly, for both area occupation and defect-related reasons. The solution is to split the gate into different “fingers”, whose maximum length is given by process rules: such pieces of poly are then connected in parallel to form the final transistor (and the resulting structure resembles a comb).

Figure 6.4 shows an inverter where both the PMOS and the NMOS have been split into three different, parallel branches.

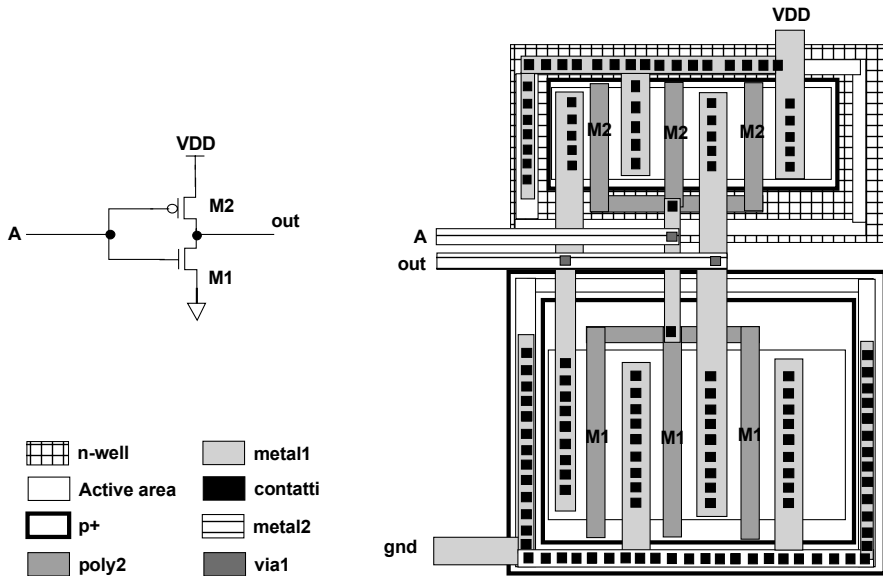


**Fig. 6.4.** Electric scheme and layout for an interdigitized inverter

Finally, Fig. 6.5 depicts the layout of a poly2/n-well capacitor; the upper plate is the poly2 layer, contacted through metal1 to the node B, while the lower plate A is the n-well, which is contacted through a portion of active area.

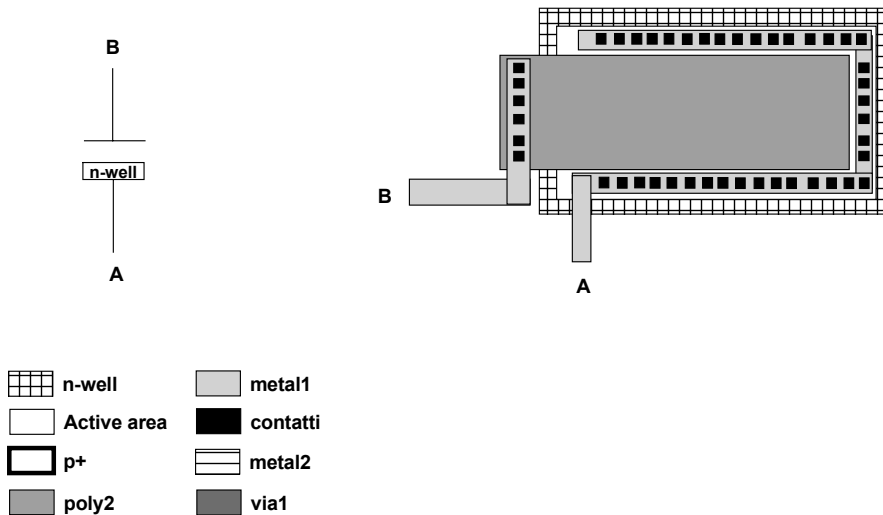


**Fig. 6.5.** Poly/n-well capacitor

## 6.5 Area and Perimeter Parasitic Capacitances

For every process, the mutual capacitances of all the different layers are calculated. Figure 6.6 shows an inverter, realized using a single “finger”, and the cross-section of the NMOS transistor. Signal nodes, i.e. those connected to metal1, are affected by parasitic capacitances since they touch source and drain junctions. Such capacitances result from the sum of two contributions: junction area capacitance and perimeter capacitance; the latter is caused by the fact that at the end of every junction the field oxide starts again, and below the field oxide a  $p^+$  junction is diffused, to isolate two active areas separated by a field oxide drawn with minimum size.

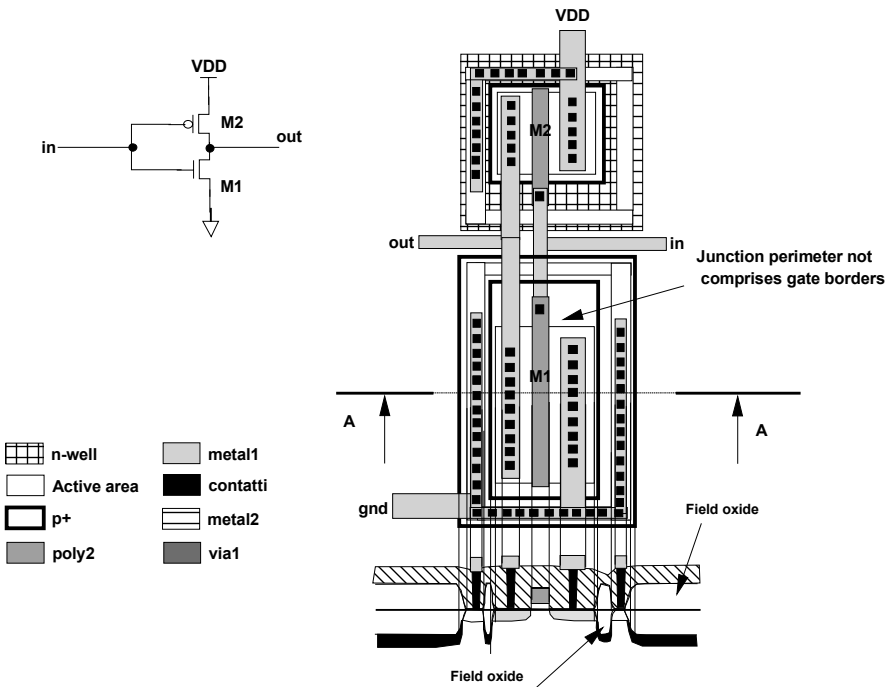


**Fig. 6.6.** Calculation of perimeter capacitance

Therefore the calculation of the perimeter for both a source and a drain junction is limited to the edges of the fields, and it excludes the edge between the gate and the junction itself, since  $p^+$  junction is not present there. The perimeter to be considered is shown in Fig. 6.6. The calculation for a PMOS transistor is similar, just recalling that the capacitances of n-well with respect to p-type substrate should be taken into account also. There is a layout technique that allows eliminating the perimeter capacitance of one of the two junctions (caused, as we said, by the  $p^+$  diffusion beneath the field). These transistors are used when breakdown value must

be increased, and therefore they are called “field-less”. Figure 6.7 shows one of these transistors; the gate is not a simple finger, but it surrounds source (or drain) area; in this way the field oxide at the end of the active area is not present.

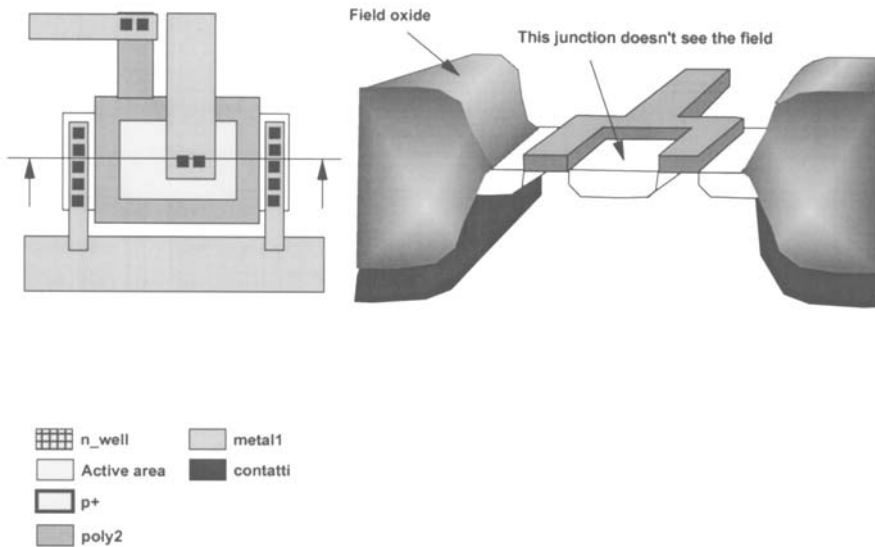


**Fig. 6.7.** “Field-less” transistor: the cross-section, without metal1, shows the central region of  $n^+$  that does not face  $p^+$  diffusions that are present below the field.

Precise calculation of parasitic capacitances is obviously very important to generate an accurate simulation of the device. Many errors can be imputed by either under or over estimated capacitances, because the resulting electrical performances of the circuit are indeed incorrect.

The big challenge is to be able to estimate the capacitance before having completed the layout: this skill makes the difference between a designer and a *DESIGNER*.

## 6.6 Automatic Layout

A custom layout allows a high degree of both area and performance optimization, because the draftsman has direct control on the basic components. It is possible, for instance, to pay more attention to the transmission line of certain signals depending on how critical they are with respect to both timing and noise immunity. However this approach has a high cost in terms of both time and human resources required for its implementation. In fact, the layout of a chip usually goes in parallel with the design of the schematics; therefore, a heavy modification of a sche-

matic can bring to a complete remake of the layout, and an associated waste of time is evident.

Increased complexity of internal algorithms and the growing number of functionality supported in today's Flash memories has caused a dramatic growth of the associated logic circuitry. To compensate this effect, a standard cell based approach is more often used.

A standard cell is a circuit that execute a given logic function: three-inputs NAND, D-type Flip Flops, Multiplexers, NOR, etc. These cells constitute a standard library and they are characterized by a layout that follows a similar structure: same height, aligned power supplies, similar position for inputs and outputs.

Layout is done in such a way that all the cells have, in one direction, say Y, the same height, so that the complete circuit can be realized as an ordered sequence of cells placed onto rows to constitute a matrix, as sketched in Fig. 6.8. Of course, X size is directly proportional to the size and the number of cells on the row. Among adjacent rows of standard cells, an interconnection channel is left, to be able to connect the various logic cells as appropriate.

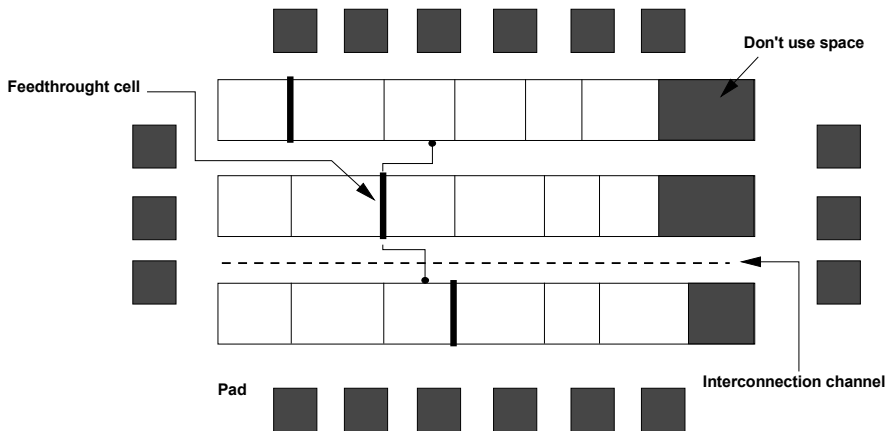


**Fig. 6.8.** Standard cell layout floorplan

Thanks to the availability of the standard cells library, digital design can use several software tools. The designer describes the logic network by means of a Hardware Description Language (HDL); then applying the “synthesis” operation with the proper timing and area constraints, the schematic is automatically generated. Of course, synthesis is technology dependent, since it maps the desired functionality onto the cells belonging to a library designed for the specific technological process used to design the device.

At this point, a Place&Route software starts from the synthesized schematic and picks up the predefined layout cells corresponding to the standard cells found in the schematic itself (these layout cells are in a standard library too, specific for the process as well), properly connecting them together.

“Placement” defines the physical position of every element of the circuit on the surface of the chip.

“Routing” is the following phase and defines the paths of the wires that connect electrically equivalent nodes. First automatic routing algorithm were used to resolve the trace layout in PCBs; later on, they have been adapted to solve the much more complex issues that can be found in integrated circuits. A very common concept in routing algorithms is to consider the layout as a labyrinth: finding a path between two pins on the same net is just the same as finding a way out of a maze. Anyway this is not enough, since routing must fulfill at least two main constraints: technological, because process rule must be followed (minimum distances, widths etc.); design constraints, because some signals are more critical than others from a timing point of view, or because the connection area should be as small as possible, to minimize area and thus the cost.

The number of connective layers depends of course on the type of process: in case of a two metal layer process, for instance, metal1 and poly might be used for connections inside the cells, while metal2 is reserved for cell-to-cell interconnections.

Parasitic capacitances due to interconnections are then extracted from the automatic layout, so that the designer can simulate the network again to a higher degree of accuracy, since the real loads have been *back-annotated*. In case of problems, either the schematic or the logic flow is modified and the synthesis operation is run again, so that a new layout is generated and so on, until all the defined parameters converge on the required specifications.

The most evident advantage of standard cell approach is design speed, which allows the designer to take care of the whole system instead of the single component, thus reducing time to market. Once the necessary logic functions are defined, proper CAD tools are used to automatically generate the layout using only the components that are part of the standard library.

The investigation of this topic alone would require, for sure, a dedicated book.

## Bibliography

- E. Charbon, R. Gharpurey, P. Miliozzi, R.G. Meyer, A. Sangiovanni-Vincentelli, *SUBSTRATE NOISE – Analysis and Optimization for IC design*. Kluwer Academic Publishers, (2001).
- J.M. Cohn, et al., “Analog Device-Level Layout Automation”, Norwell, MA: Kluwer, (1994).
- R.L.M. Dang and N. Shigyo, “Coupling capacitance for two-dimensional wires”, *IEEE Electron Device Letters*, Vol. EDL-2, No. 8, pp. 196-197, (August 1981).
- M. I. Elmasry, “Capacitance calculations in mOSFET VLSI”, *IEEE Electron Device Letters*, Vol. EDL-3, No. 1, pp. 6-7, (January 1982).
- F. Maloberti, “Analog Design for CMOS VLSI System”, Kluwer Academic Publishers, Boston, (2001).
- W. Maly, *Atlas of I.C. Technology*, The Benjamin Cummings Publishing Company, (1987).
- K. Ming-Dou, W. Chung-Yu, W. Tain-Shun, “Area-Efficient Layout Design for CMOS Output Transistors”, *IEEE, Trans. On Electron Devices*, vol. 44, no. 4, (April 1997).

- T. Sakurai and K. Tamaru, "Single formulas for two-and three-dimensional capacitances", IEEE Transactions on Electron Devices, Vol. ED-30, No.2, pp. 183-185, (February 1983).
- N.P. Van Der Meiji, J.T. Fokkema, "VLSI circuit reconstruction from work topology", North Holland INTEGRATION, the VLSI Journal vol. 2, pp. 85-119, (1984).

# 13 Multilevel Read

## 13.1 Multilevel Storage

The necessity of non-volatile memories to have higher and higher densities leads to the consideration of multilevel memories. In this case, it is possible to generate more than two logic levels by controlling the amount of charge stored on the floating gate.

Figure 13.1 shows the characteristics ( $V_{GS}$ ,  $I_{DS}$ ) of a bi-level Flash cell in reading condition. The characteristic with the  $V_{TV}$  threshold voltage, generally ranging between 0.5 and 2.5 V, is assigned to logic value “1”, whereas the characteristic with  $V_{TW}$ , typically greater than 5 V, is assigned to value “0”. As stated in Chap. 12, reading consists of the conversion of the current drawn by the cell at a defined  $V_{GS}$  into a voltage that is then translated into the corresponding logic level.

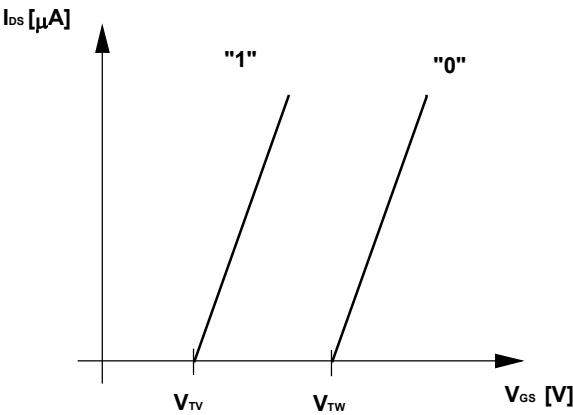


**Fig. 13.1.** Bi-level cell

In the case of multilevel cells, the ( $V_{GS}$ ,  $I_{DS}$ ) plane is divided by several characteristics. For example, in the case of 2 bits/cell, there are four resulting characteristics (Fig. 13.2) to which the 11, 10, 01, and 00 logic values are associated. In reality, we obviously deal with threshold voltage distributions spaced with respect to each other in a way that allows the determination of the logic value by means of the sense amplifiers.

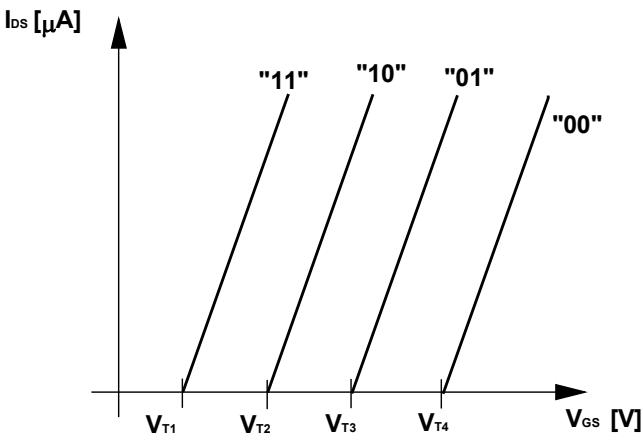


**Fig. 13.2.** Binary encoding of the threshold voltage in the multilevel case

One of the problems in reading the cell is related to the  $V_{GS}$  voltage. Since reading is a current-to-voltage conversion, the voltage that is applied to the cell gate must be at least greater than  $V_{T3}$ . Figure 13.3 shows one possible placement of the distribution; for sake of simplicity, the distributions are reported in the same scale (except the 11 value which has a larger  $V_{GS}$  range) and equally spaced. The read voltage, for example, can be fixed to 6 V.

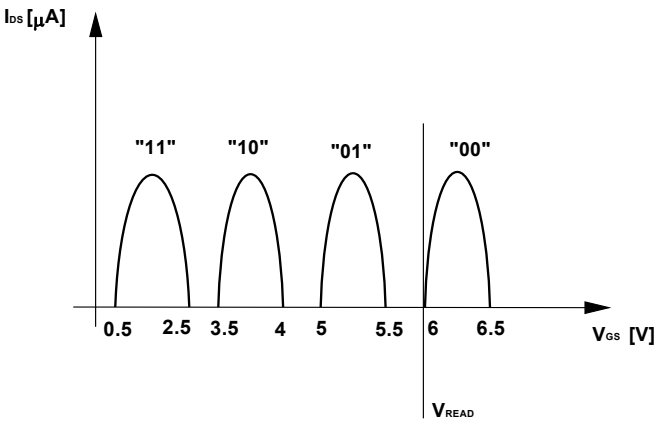


**Fig. 13.3.** Position of the threshold voltages for a cell containing two bits

## 13.2 Current Sensing Method

One way to discriminate the cell current is the usage of three reference cells, with  $V_T$  programmable during the testing phase and located between the various characteristics (Fig. 13.4). For sake of simplicity, also in this case the three references are centered between the distributions of cell currents representing the logic states.

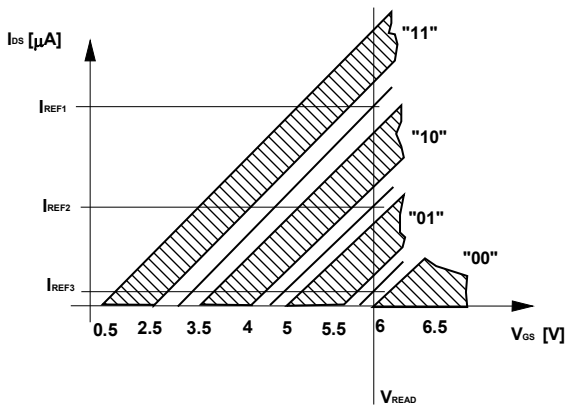


**Fig. 13.4.** Current sensing approach: the cell is read at constant  $V_{GS}$ , exploiting the three different reference cells

The read operation is accomplished by applying a gate voltage to the memory cells ( $V_{READ}$ ) and comparing the current drawn by the cell to the three reference currents,  $I_{REF1}$ ,  $I_{REF2}$ ,  $I_{REF3}$ . A circuit structure, composed of several I-V converters, compares the conversion results and provides the logic value to the output, as shown in Fig. 13.5. The minimum working window of the threshold voltage is determined by reliability requirements rather than the number of bits that the cell stores. It is then evident that the multilevel structure causes a reduction in the distance between the various  $V_T$  distributions. This reduction increases the difficulty of sensing, since there is less differential current between logic states. Due to the limited gain of the cells (between 20 and 25  $\mu A/V$ ), the sense circuitry must discriminate currents on the order of 10  $\mu A$ , also in the case of only four levels.

Figure 13.6 shows the real position of the three reference cells. The characteristics of the first two, R1 and R2, are parallel to the characteristics of the array cell. The threshold voltage of the third cell, R3, is instead set to the upper value of the "01" threshold voltage distribution. To understand the particular position of R3 it is necessary to recall that, while  $V_{READ}$  remains practically fixed as bias voltage and temperature vary (it is derived from a bandgap reference), the cell threshold voltage drifts with the temperature. Thus, there is the risk of a too low reference current when the temperature decreases. The unbalanced read technique is just for lowering the threshold voltage of R3.

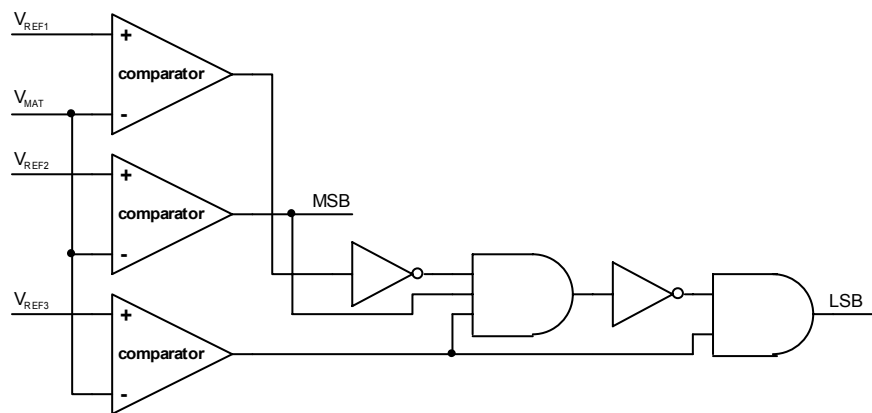


Fig. 13.5. Analog-to-digital converter for parallel multilevel sensing

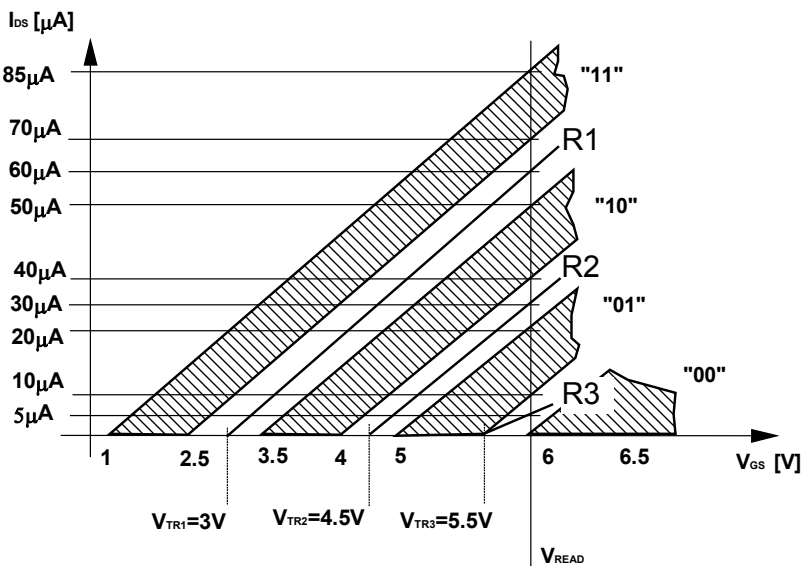


Fig. 13.6. Actual position of the reference cells: R3 is used in the unbalanced configuration

In order to fix the problem of the limited available current, it is possible to use the amplified I-V converter, already analyzed in Sect. 12.9. It is preferable to amplify the current of the cell corresponding to the "01" level because this is the one that sinks less current during read due to the high  $V_{tr}$ . For example we could decide to multiply the current of the "01" cell by three; while the "11" and "10" distributions are only doubled so as to limit power consumption. Figure 13.7 summarizes the ratios defined between the currents of the various distributions.

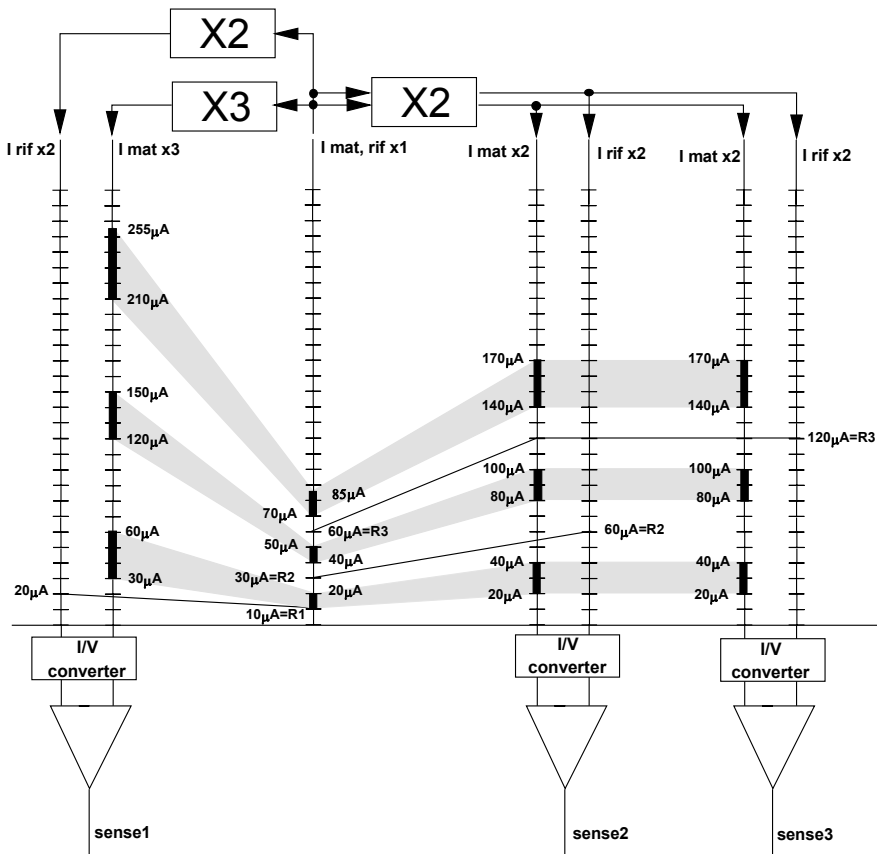


**Fig. 13.7.** Amplified I/V converter for multilevel reading

The vertical axes represent matrix and reference currents. The “Imat, rif x1” axis indicates the current values referred to the tails of the different distributions, as shown in Fig. 13.4. The first distribution is in the range between 100 and 70  $\mu\text{A}$ , the second between 50 and 40  $\mu\text{A}$ , the third between 20 and 10  $\mu\text{A}$ . The three references sink 60  $\mu\text{A}$  (R1), 30  $\mu\text{A}$  (R2), and 10  $\mu\text{A}$  (R3), respectively. The different currents are collected by means of current mirrors (the blocks located above) and transferred to the current/voltage converters.

The current of the array cell is sensed and multiplied, in one case by three, in the other cases by two. The reference currents are always multiplied by two and, thus, in the case of the converters connected to sense2 and sense3, the overall effect is the doubling of the dynamic signal range. In the case of sense1, with the lowest reference current R3, we must introduce a different multiplication factor for the array, to be able to distinguish the array from the reference. In this specific case, R3 must always be lower than the current drawn by the cell and, hence, the cell current must have a higher multiplication factor. Figure 13.7 shows the differ-

ent currents that reach the current/voltage converters: for each reference, the different vertical axes represent the possible position of the current of the various distributions. Basically, the three reference currents are collected by the small reference array, amplified by a factor 2 and presented as inputs of the respective I/V converters. The cell current is instead collected and multiplied by two or three. The current multiplied by two is input to the converters of sense2 and sense3, while the current amplified by three is input to sense1. In this way, we obtain the desired result of increasing the dynamic range of the available current, limiting, at the same time, the threshold voltage of reference R3 at the upper value of the “01” distribution.

### 13.3 Multilevel Programming

The precision required in the multilevel case is not limited to the read operation; it is also necessary to design a write mechanism to program the cells within the different distributions. The write mechanism usually employed in Flash memories with NOR architecture is the hot electron technique presented in Chap. 3: by applying a voltage of about 10 V on the control gate, a voltage of about 5 V to the drain and leaving the source connected to ground, it is possible to generate an intense electric field between isolated gate and channel and, at the same time, a high current between source and drain. This permits the electrons flowing between source and drain to acquire high kinetic energy and reach the floating gate from the channel, overcoming the energy barrier to cross the tunnel oxide.

The advantages offered by the implementation of a multilevel memory are accompanied by formidable problems related to the reduction of the difference between the different threshold voltages corresponding to the various charge quantities stored on the floating gate and, hence, between the different cell current flow.

The programming of a multilevel memory cell requires very precise control of the charge stored on the floating gate.

It is possible to define a linear relationship between the variation of the voltage applied to the gate during the program phase,  $\Delta V_G$ , and the threshold voltage step,  $\Delta V_T$ , obtained at fixed  $V_D$  and  $V_S$ , as shown in Fig. 13.8.

The memory cell is, therefore, programmed by applying a series of linearly increasing step voltages to the gate: each program pulse has to differ from the previous and the subsequent one by a constant value,  $\Delta V_G$ . The program gate voltage has therefore the appearance of a staircase with constant step. In particular, a significant result obtained by the use of this method is that, neglecting all the effects related to retention, temperature variation and so forth, the width of the threshold voltage distribution equals the step of the staircase voltage applied.

The method described above is inherently slow since it requires the application of a sequence of pulses to the cell gate. In order to obtain a single byte program time comparable with a standard bi-level cell it is therefore necessary to program several cells in parallel. Let's suppose that the single byte program time is 6  $\mu\text{s}$  in the bi-level case; if we need 200  $\mu\text{s}$  to apply the entire program stair, it is neces-

sary to program 256 bits simultaneously to obtain a program time of  $6\ \mu\text{s}$  for each single byte.

The reader can refer to the next chapter for a more detailed description of the entire program algorithm.

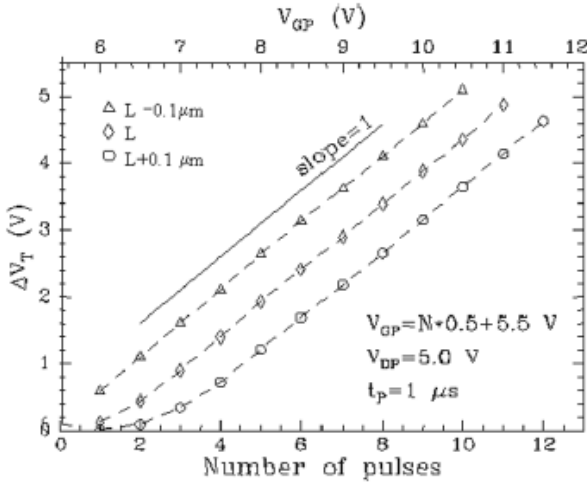


**Fig. 13.8.** Staircase programming:  $\Delta V_T$  as a function of the number of pulses for three different channel lengths. The upper axis shows the gate voltage at each program step

## 13.4 Current/Voltage Reference Network

As stated in Sect. 12.15, the number of reference cells typically used in Flash memories is lower than the number of sense amplifiers present; as a consequence, the  $V_{\text{REF}}$  reference voltage is derived from one cell and shared among several sense amplifiers. Generally, two or four reference voltages can be obtained from one reference cell, and each of them is shared among a group of sense amplifiers, as sketched in Fig. 13.9. For example, the  $V_{\text{REF}}$  reference voltage can be the  $V_{\text{MIRROR}}$  voltage in Fig. 12.40; by replying  $I_{\text{REF}}$  several times (M1, M2 mirror), several  $V_{\text{MIRROR}}$  can be obtained.

The conversion of array current into voltage is carried out locally inside the sense amplifier. The reference current is instead converted into a voltage in a location away from the sense amplifiers since the reference cells belong to the dedicated small array. The reference voltages are obviously referred to a local ground of their I/V converters and, hence, it is important that the offset in the ground connections be minimized between the reference circuitry and the sense amplifiers.

In Fig. 13.10 the layout of the power grids inside a generic device is shown. The ground potential is routed by means of a ring of metal lines, whose width is the result of a tradeoff between area occupation and parasitic resistance. The volt-

age drop on the ground lines depends on the line layout while the current consumption of the sense amplifiers depends on the current of the array cells that are read. Thus, the current consumption and the consequent voltage drop on the ground lines depend on the pattern that is read.

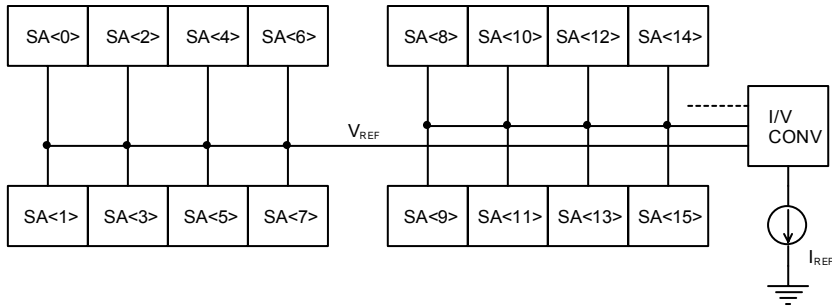


**Fig. 13.9.**  $V_{REF}$  reference voltage shared among several sense amplifiers

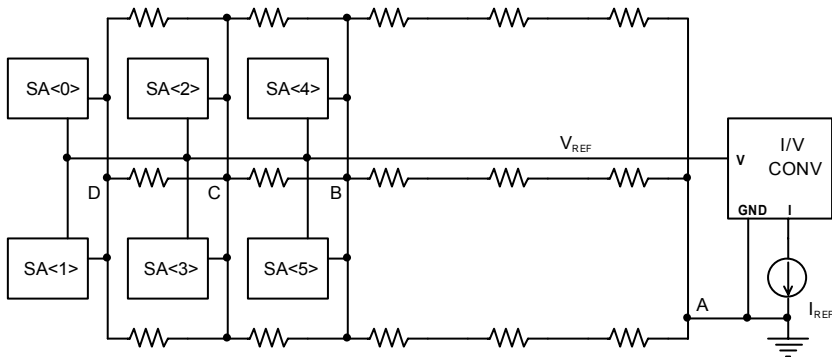


**Fig. 13.10.** Layout of the ground lines and  $V_{REF}$  in the area of the sense amplifiers

In order to better understand the importance of the effect described above, let's consider the case of a multilevel memory.

Programming is performed by increasing the cell threshold voltage by small steps (typically 300 mV for a multilevel memory with 2 bits/cell). Starting from the distribution of erased cells ("11"), a verify operation is performed at the end of each single step, i.e. an actual read operation is performed. The cells that have reached the required distribution are not subjected to further programming pulses. During the various verifications, the current that flows in the sense amplifiers depends on several cell currents, which are changing as the cells are programmed. The verifications are thus performed with different ground currents and can have

variation in the ground offset as the cells are programmed to different threshold voltages.

The problem arises when one or more cells reach the required distribution. Although they are not involved (because they have been verified correctly), further program pulses will be applied to the other cells. In practice, at the end of the program phase, the voltage drop on the ground lines does not equal the drop during the first cell verification.

Let's examine Fig. 13.10 and suppose that we want to program the cell connected to SA<0> to level "11" and all the others to "00".

The I/V conversion takes place according to the local ground of the converter. For the reference it is node A in Fig. 13.10, for the SA<0> sense amplifier, it is node D, for SA<2> it is node C and so forth.

The first verification is successful for SA<0>, while the other cells have not yet reached the programmed state so these cells do not verify correctly. The first verification is executed with the sense amplifiers connected to the cells that sink the highest current and, therefore, the voltage drop between nodes D and A is the highest.

As the threshold voltage of the cells is increased, the current consumption diminishes; and therefore the voltage drop between D and A decreases. This variation of  $V_{DA}$  is transformed into a differential voltage input signal for the voltage comparators of the sense amplifiers. This means that the distributions are enlarged and consequentially the device has lower noise immunity and increased access time.

**Problem 13.1:** Due to the effect described above, do the distributions of the threshold voltage enlarge only in one direction or do both tails move?

The solution generally adopted in such a case is to bring the generation of  $V_{REF}$  closer to the group of sense amplifiers that share the reference, as indicated in Fig. 13.11. In this case, the reference signal transmitted inside the memory is the  $I_{REF}$  current which implies advantages in terms of noise immunity.

**Problem 13.2:** What may happen when a current is reproduced by means of transistors in mirror configuration that are located in very different areas of the chip? Recalling the PMOS case, consider that a small voltage drop takes place between the sources of the two p-channels.

In Fig. 13.11, the ground for the I/V conversion of the reference current is node B. In this way we eliminate the contribution of  $V_{BA}$  because this drop is now a common mode signal for the voltage comparator of the sense amplifiers. However, the variation of the voltage drops has not been overcome since the problem of  $V_{DB}$  is still present.

A good solution in such cases is the adoption of a different ground line for each I/V converter. The various ground lines are connected to a single node (node B in Fig. 13.11) to which the I/V conversion of the reference current must be referred. In this way, the voltage drop between the ground of the single converter and node B is no longer influenced by the current variations of the other sense amplifiers.

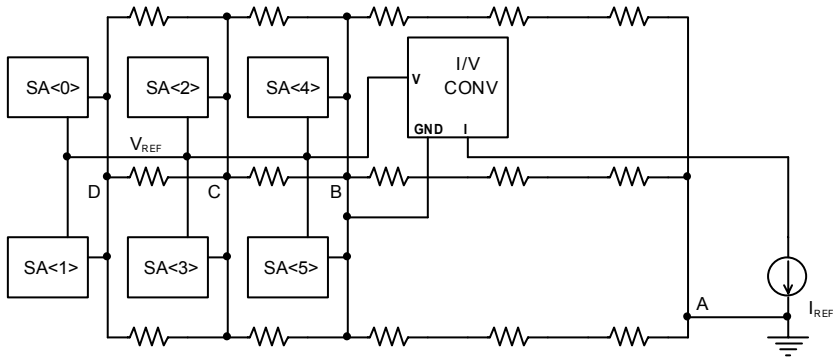


**Fig. 13.11.** The reference signal is the  $I_{REF}$  current and the I/V conversion is performed locally, close to the group of sense amplifiers that share the same  $V_{REF}$

**Problem 13.3:** Draw the layout of ground and reference voltages in the case of dedicated ground line for each sense amplifier.

**Problem 13.4:** Evaluate the impact of the realized layout in terms of area occupation.

## 13.5 Voltage Sensing Method

Basically, the reading of a non-volatile cell can be accomplished in two ways: with either constant gate voltage or constant drain current, as shown in Fig. 13.12. In both cases, the drain and body voltages are kept constant. If the gate voltage is kept fixed, we talk about current sensing, since the value of the current indicates the program status. On the other hand, if the drain current is kept constant, voltage sensing is utilized since  $V_{GS}$  provides the means to determine the data stored. While the current sensing method and its variations are the conventional techniques used to read the traditional bi-level and four-level memories, the voltage sensing method has been developed to suitably read the more complex multilevel devices.

Figure 13.13 shows the eight logic levels that are necessary to store three bits in a single cell. In this case the voltage sensing is appropriate. Instead of biasing the word line directly to 6 V, the  $V_{GS}$  voltage of the cells increases in a staircase fashion. For example, the first and the final value of such a staircase could be 2 and 6 V respectively. The number of steps of the row voltage equals the number of levels minus one.

The operating principle of this technique is based upon the comparison between the current drawn by the cell and an  $I_{REF}$  current generator. In Fig. 13.14 the “101” distribution is shown and the reference current is 10  $\mu A$ . The transient of the word line voltage is reported in Fig. 13.15. At each step, the sense amplifier verifies whether the cell sinks more current than 10  $\mu A$ . With reference to the case of the “101” level, the comparator of the sense amplifier toggles during the third step of

the  $V_{GS}$  voltage. This operation requires a single sense amplifier for each cell, with remarkable saving in terms of area since the result of the comparison is “1” or “0”, like in the case of bi-level reading.

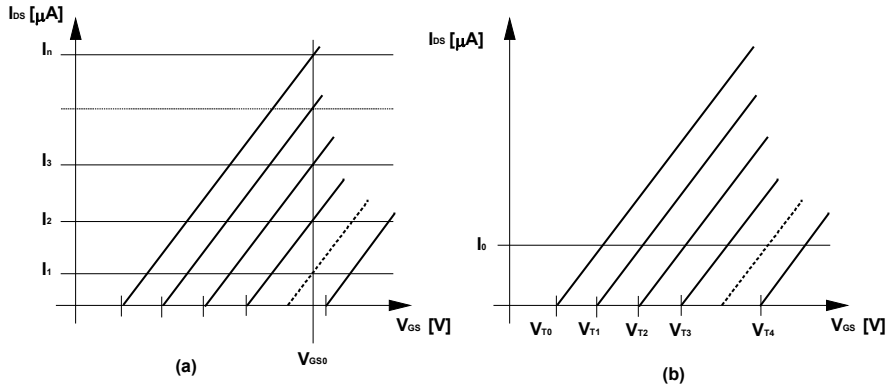


**Fig. 13.12.** Comparison of read methods: (a) constant gate voltage ( $V_{GS0}$ ), (b) constant drain current ( $I_0$ )

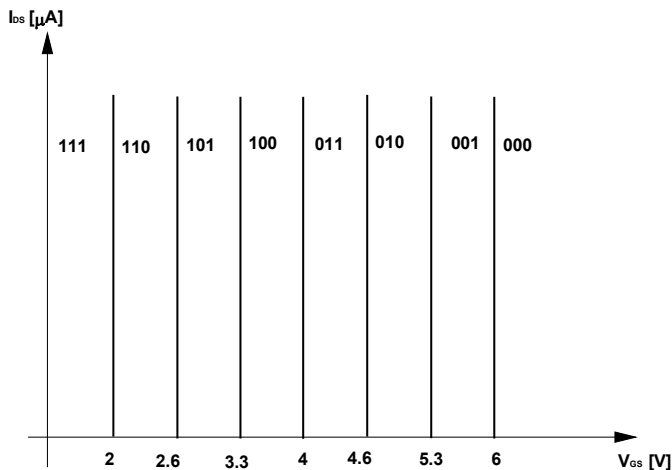


**Fig. 13.13.** The eight distributions that are necessary to store 3 bits in a single cell

The steps of the row voltage can have a 15 ns time duration, leading the overall time to charge the row to 105 ns. This delay significantly impacts the access time. By reading more cells in parallel and exploiting the fact that each cell stores three bits, it is possible to improve the read throughput by means of an architecture that allows burst readings (see Chap. 11).

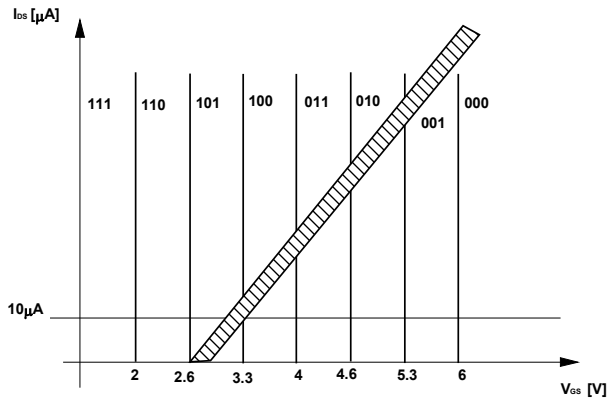


**Fig. 13.14.** Current/voltage characteristics of the cells corresponding to the “101” logic level

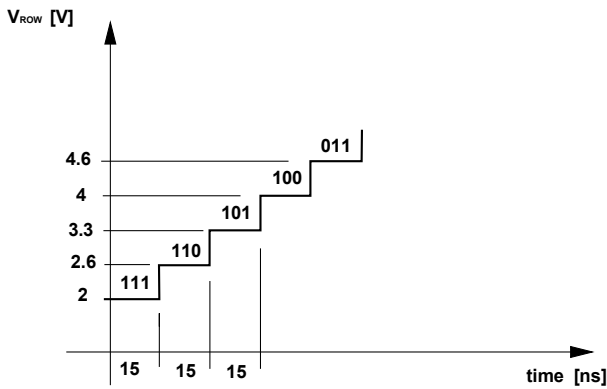


**Fig. 13.15.** Word line staircase voltage

The comparison between the cell current and the current of the reference generator can be carried out with a classic sense amplifier structure, as sketched in Fig. 13.16. The I/V converter has to pull down the  $V_{\text{MAT}}$  potential so as to be lower than  $V_{\text{REF}}$  when the cell sinks more than  $10 \mu\text{A}$ . The toggling of the comparator, together with the voltage value the stairs has reached identifies the set of bits stored in the memory cell.

**Problem 13.5:** Considering what discussed in Chap. 12, define the details of the sense amplifier in Fig. 13.16.

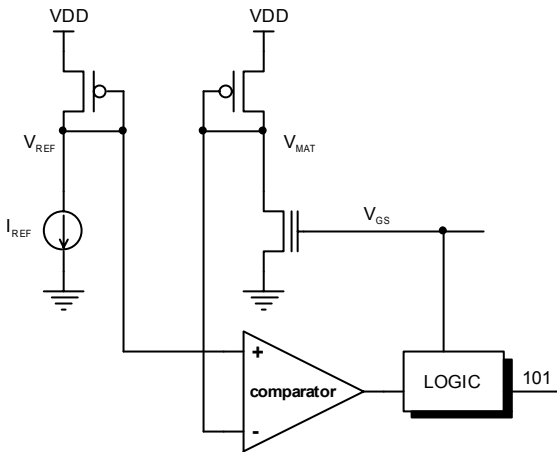


**Fig. 13.16.** Schematic of the sense amplifier that performs the comparison between the cell current and the reference generator (for example 10  $\mu\text{A}$ )

## 13.6 Sample & Hold Sense Amplifier

The foregoing solution is not easily employed when more than three bits are stored in the same cell. In fact, the necessity of higher and higher number of steps for the word line voltage has an increasingly negative impact on the access time. In this section, a sense amplifier that performs a sort of sample & hold operation and also holds the value of the row voltage that caused the toggling is presented.

As we saw in Chap. 9, row and column decoders allow the unambiguous selection of the cells to read or program. The byte (or word) is made of cells that are located on the same row but on different columns. In this way, the current drawn by each cell is converted into a voltage. This kind of architecture does not permit independent behavior of the word line for each cell unless a serial read is carried out.

In order to separate the eight cells that compose the byte it is possible to adopt a hierarchical row decoding. In Fig. 13.17 the typical structure of a hierarchical selection of the word line is indicated. The local decoders separate the sectors, i.e. the portions of the array that are simultaneously erased.

Figure 13.18 shows a different kind of hierarchical organization, in which the same sector is further divided into blocks by means of local row decoders. In this way, bits belonging to different bytes (or words) are grouped in the same block and, therefore, each cell of the byte has its own local row. Using such a technique, it is easy to design the sample & hold circuit which stores the row voltage. Figure 13.19 shows the sample & hold circuit of two cells belonging to the same byte but located on different rows. For the sake of simplicity, in the representation of the local decoder only the PMOS transistor (PY0, PY1) used to transfer the read voltage has been highlighted. The reader can refer to Sect. 9.12 for further details on the structure of the local decoders.

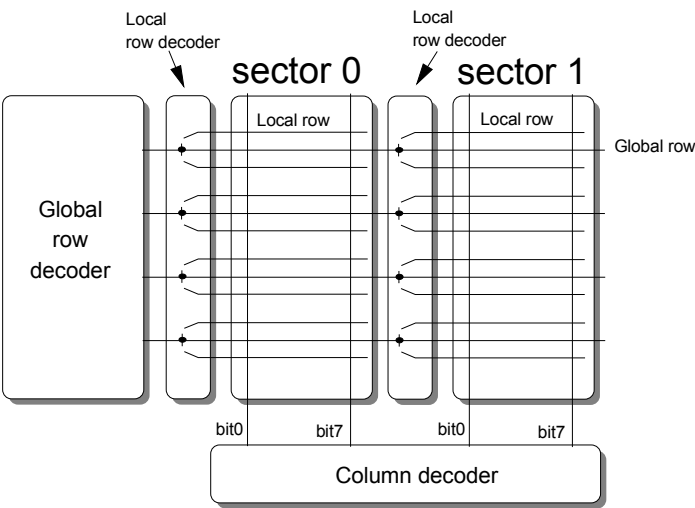


**Fig. 13.17.** Hierarchical row decoding

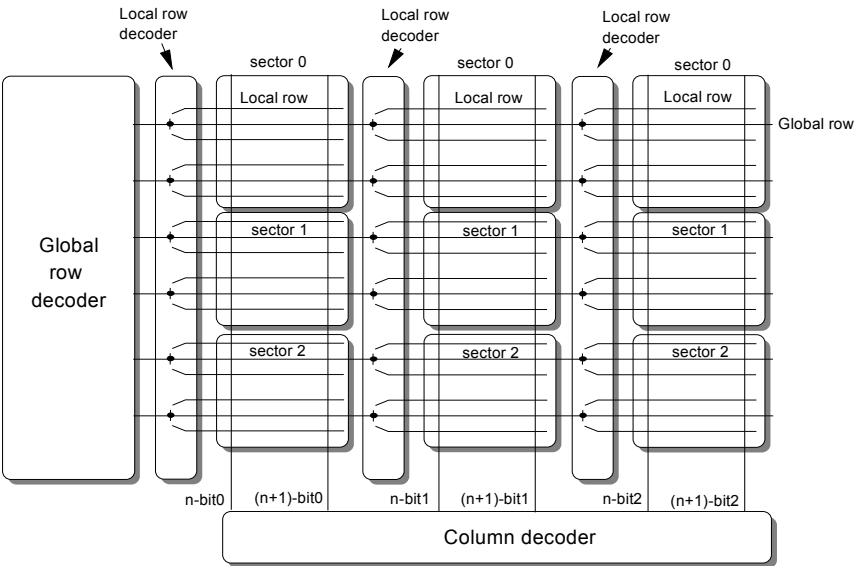


**Fig. 13.18.** Row decoder for sample & hold sense amplifiers

The voltage of the global rows increases linearly with time. When the comparator signals that the cell sinks more than  $10\ \mu\text{A}$  ( $I_{\text{REF}}$ ), the P1 p-channel transistor is switched off. The local row voltage is then stored on the parasitic capacitance (0.5

$\div 1$  pF), and is approximately the voltage reached by the main word line (MWL) at the time the comparator commutes. Obviously, each local row (LWL) is able to reach a voltage value independently of the other rows. We could imagine a voltage ramp applied to the global row having  $100 \div 150$  ns time duration, depending on the number of levels to discriminate. The principle is to take the access time of a standard 3 V bi-level flash memory as a reference, i.e. 70 ns typically. In the case of 16 levels, 8 cells produce 32 output bits, which means 4 times the number of bits of the bi-level memory that works by bytes.

The access time to read 4 bytes of a bi-level memory is therefore 280 ns, which is the figure we refer to as a benchmark.

Let's now determine how to read the voltage of the word line. In Fig. 13.20 the circuitry of two cells is shown: let's concentrate on the cell marked with the circle. The voltage ramp is applied to MWL<1>, while YO<0> is active. In such conditions, M1, M3, M5, and M7 are on while M4 and M8 are off. The only local row that is high, as a result of the combination of the high main word line and YO<0>, is the local row to which the indicated cell belongs. The MWL starts going up, driven by the global row decoder (not indicated in the figure). When this cell sinks 10  $\mu$ A, the comparator related to cell 0 triggers, turning M1 off and, hence, sampling the local row. Meanwhile, M2 is on and M6 is off. The voltage of the addressed local row is transferred to the A/D converter through the wl-bit line. The conversions of all the other cells that form the byte (word) are performed in parallel. The local rows have transistor M2 in common, which is also connected to M6.

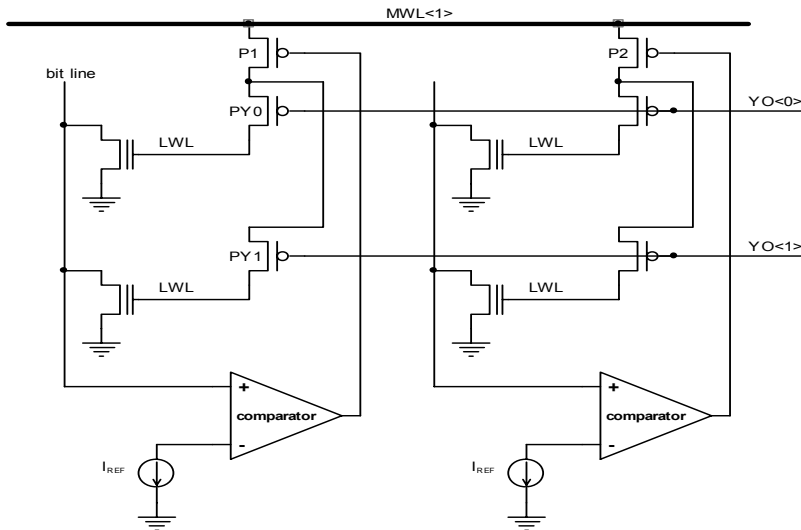
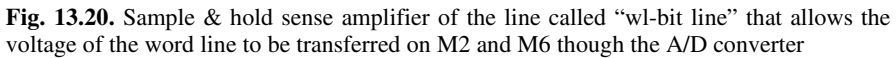


**Fig. 13.19.** Sample & hold sense amplifier for two bits: the voltage value reached by the word line is hold in the parasitic capacitor of the row


$$V_{SH} \cdot C_{WL} + (V_{SH} - V_{T,M2}) \cdot C_B = V_{AD} \cdot (C_{WL} + C_B) \quad (13.1)$$

$$V_{AD} = V_{SH} - V_{T,M2} \frac{C_B}{C_{WI} + C_R} \quad (13.2)$$

Considering the case in which  $C_B \gg C_{WL}$ , we obtain:

$$V_{AD} = V_{SH} - V_{T,M2} \quad (13.3)$$

The above relationship assumes the body effect is negligible. In order to actually eliminate the contribution due to body effect, transistors M2 and M6 must be fabricated in a triple well structure. Then the voltage that is input to the converter is that of the word line minus a constant term.

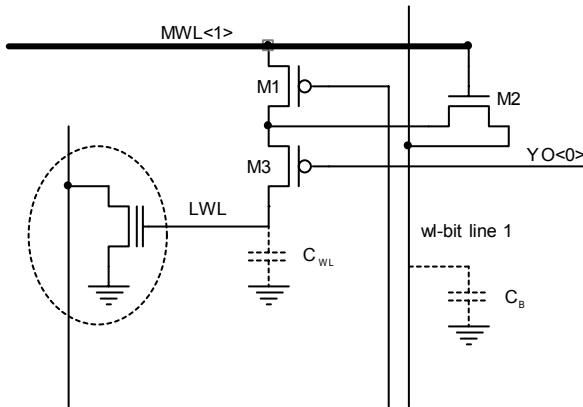


**Fig. 13.21.** The parasitic capacitor limits and modifies the voltage transferred from the local word line to the A/D converter

The solution proposed above suffers from the delay introduced by the circuit used for the feedback (switching P1 off in Fig. 13.19). If the signal that drives the loop opening does not operate properly,  $V_{GS}$  will be incorrect since it will contain a  $\Delta V$  term proportional to the delay. In fact, the ramp applied to the word line continues regardless whether the local word line is connected. An error due to the loop delay will be more significant if the ramp rises quickly. The  $\Delta V$  term can be regarded as an offset, and therefore eliminated if it is independent of operating conditions and of cell status. Another limit is due to the nature of the sample circuits: once the loop is open, the  $V_{SH}$  voltage is fixed and cannot be changed. Any disturb that causes the comparator to toggle will result in an incorrect reading of the cell contents.

## 13.7 Closed-Loop Voltage Sensing

Another way to implement voltage sensing is to use a real closed-loop feedback. The difference between the current drawn by the selected and the reference cell drives (more or less directly) a voltage source connected to the cell gate. Equilib-

rium is reached when the two currents are equal. The gate is regulated to the voltage value that maintains a state of equilibrium. The block diagram of the closed-loop voltage sensing is reported in Fig. 13.22.

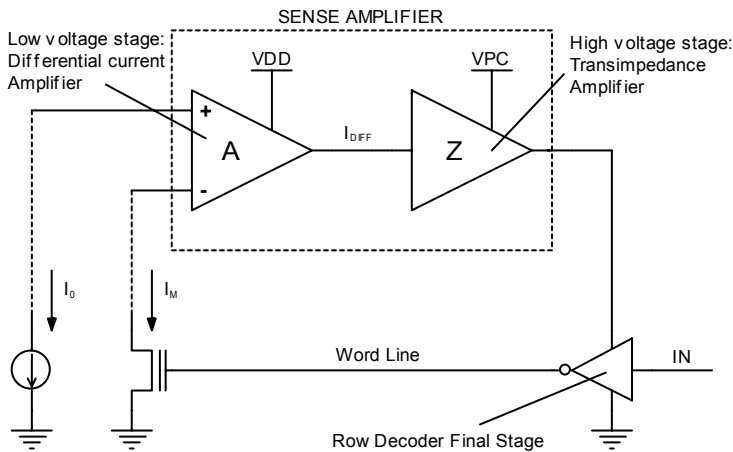


**Fig. 13.22.** Closed-loop voltage sensing approach

The advantage of this solution is the fact that the  $V_{GS}$  voltage settles only when the current of the cell reaches the required value. Possible delays in the signal propagation do not impact the final value but only the settling time. The main problem of the closed-loop reading method is that high speed reading increases power consumption. High speed means that the voltages involved have to vary quickly, and the parasitic capacitors of the system have to be charged and discharged with high current.

In order to realize the circuitry, it is possible to use a class AB amplifier with two stages, as indicated in Fig. 13.23.

The class AB amplifier has good performances during both the word line charging and discharging phase. This solution also minimizes the overshoots caused by delays in the feedback. The choice to use a two-stage structure is mainly due to the power consumption of the charge pumps. In fact, as one can see in Fig. 13.23, only the second stage is biased by means of the  $V_{PCX}$  ( $> V_{DD}$ ) high voltage.

Let's now move on to the analysis of the circuit. The bit line limiter is a conventional common gate device, with M1 biased by inverter INV. The cell current is mirrored by the first stage and compared with the reference current source. The resulting current is amplified and buffered by the second stage. The bias circuit of the HV stage, made up of transistors M4, M5, M10, and M11, is worth noting. The operation is based on the principle of the current mirror, with a mirroring factor of unity. The input branch is made of two diode connected MOS's (M10 and M11), together with the  $I_{HV}$  source so as to control the current that flows in the

branch. The voltage drop at the diode terminals is the same as the one between the gates of M4 and M5. If a current is injected into the input branch, it is mirrored in the output branch. Thus, due to the symmetry, node B has the same potential as node F. In this way, the DC currents and voltages of the high voltage stage can be controlled, without limiting the dynamic range.

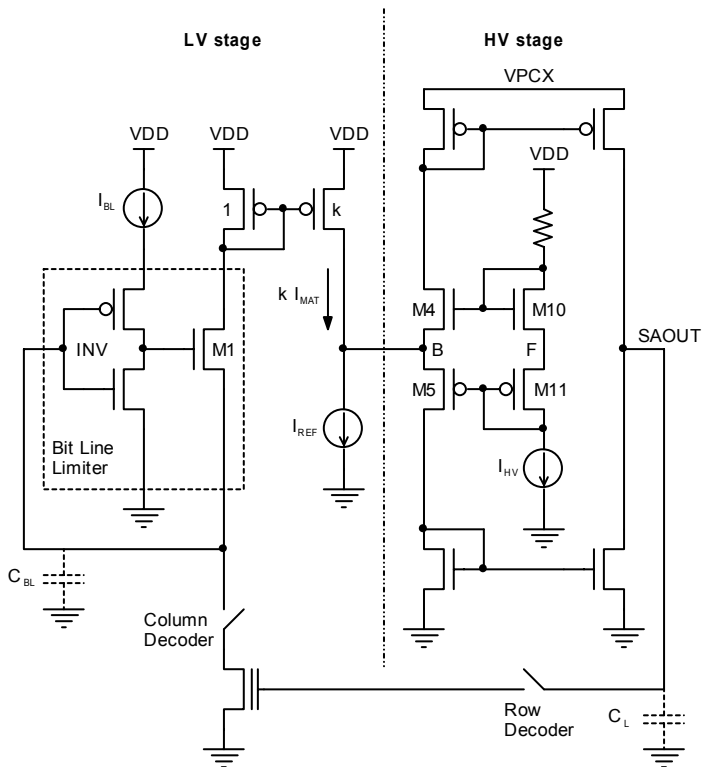


**Fig. 13.23.** Sense amplifier for closed loop voltage sensing

In order to speed up the sensing operation, the SAOUT output node can be set to an intermediate potential during the precharging of the bit line. When the loop is enabled, two transients can be obtained as shown in Fig. 13.24. If  $V_{SAOUT}$  reaches a final value greater than the starting one, we have a constant current charging until the cell starts sinking current, then a settling exponential behavior follows. In the case when the steady-state potential value of the word line is lower than the precharging voltage, we have an exponential discharge.

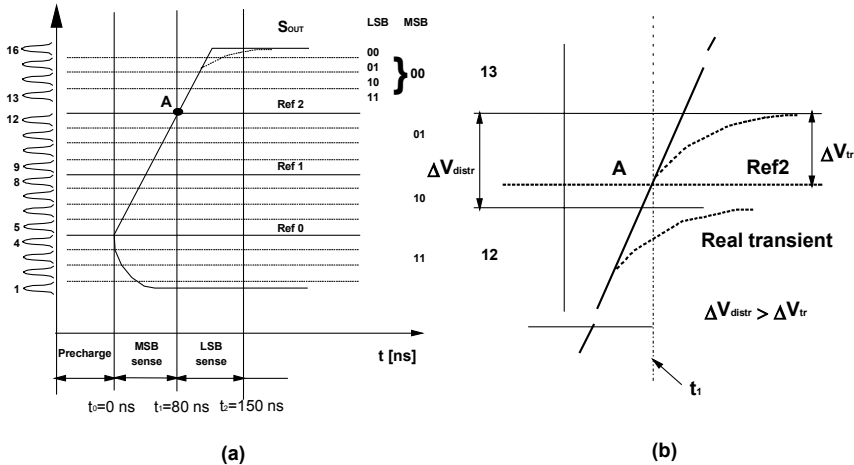


**Fig. 13.24.** (a) Transient of the sensing loop output voltage (the reference levels required for the A/D conversion are also shown; solid lines: MSB references; dashed lines: LSB references); (b) detail of the transient around point A for level 12 and 13

With this kind of read operation, it is possible to obtain access times around 200 ns in the case of 4-bits/cell memory. The consumption in terms of current sourced by node VPCX is around some tens of microAmperes for each sense amplifier.

## 13.8 Hierarchical Row Decoding for Multiple Sensing Loops

In order to close the loop it is necessary to find a way to apply the output voltage of the regulator to the cell gate. The structure of the decoder is still the one shown in Fig. 13.18, where a single sensing circuit has to be used for all the array blocks that are vertically arranged. Before delving into the circuit details, it is important to understand what kind of voltages we have to apply to the cell gate. During the read phase, the minimum voltage that is read is typically around 1.5 V, and the maximum voltage is around 7 V. In reality, the feedback loop represents a direct path to also transfer the programming voltages to the cell gate, which typically vary from 1 V to 9 V.

The method to allow both modes of operation in the same row circuitry is accomplished by the insertion of an inverter in the schematic of the hierarchical decoder in Fig. 9.34, thus obtaining the schematic in Fig. 13.25. Such an inverter must be driven by the MWL and, on the other hand, has to drive a metal row that is realized within the block but performs the same functions as the MWL in the hierarchical row decoder. For this reason it is named SUBMWL.

As for the hierarchical row decoder, SUBMWL is connected to 4 local word lines. At this point, it is necessary to clarify that the  $V_{SAOUT}$  voltage represents the sense amplifier output voltage during the read operation, whereas, during the program phase it represents the output voltage of the program voltage stairs. Therefore, the M3 p-channel, operating as a pass transistor, closes the loop. Transistor M5 is necessary to prevent non-addressed SUBMWL from being in a floating condition.

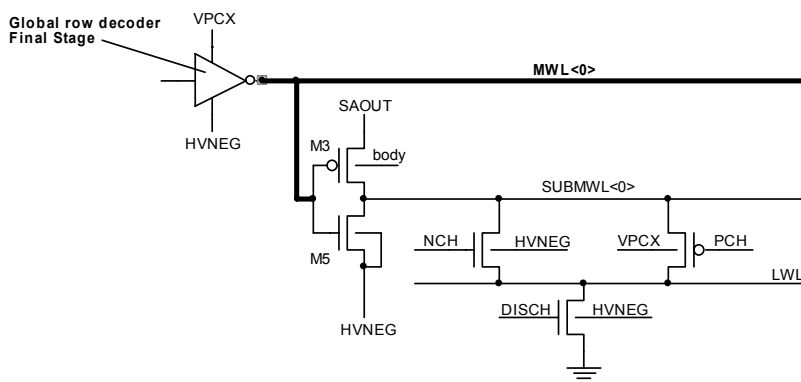


**Fig. 13.25.** Introduction of SUBMWL in the hierarchical decoder

Actually, in this implementation, the structure has some drawbacks. In fact, it is important to note that it is necessary to drive transistor M3 with low voltage in order to transfer the output voltage of the sense amplifier to the cell gate, whereas transistor M3 must be driven with high voltage to tie SUBMWL to ground. While during read and program, the value of  $V_{SAOUT}$  is not necessarily known, and M3 can be switched off only by applying to its gate the highest available voltage during each specific operation. This means that only the selected MWL must be tied to ground while all the unselected MWL's must be at high voltage.

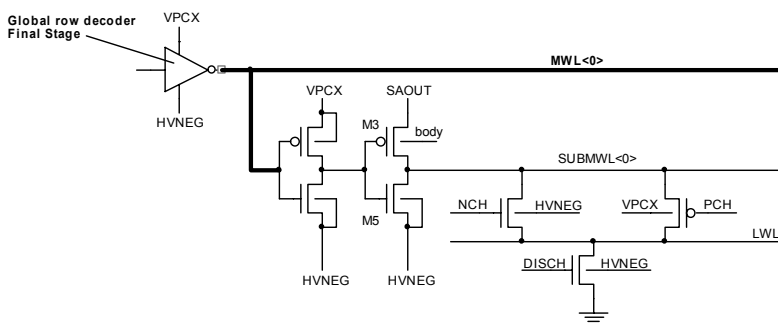


**Fig. 13.26.** Introduction of a second inverter not to have selected MWL's at low voltage and unselected MWL's at high voltage

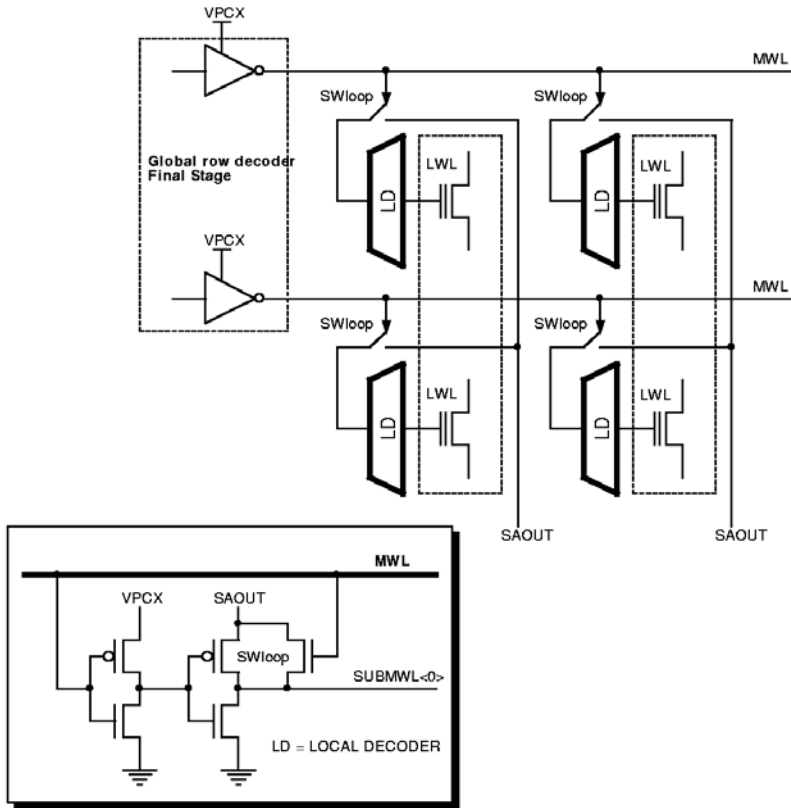


**Fig. 13.27.** Hierarchical row selection with separate addresses and analog voltage transmission functions

The problem is that the high voltages are generated internally by means of charge pumps and the charging of considerable capacitance may significantly impact the time needed to reach the required voltage. Thus, it would be better to be in the opposite situation: a single MWL selected at high voltage and all the unselected word lines tied to ground. The solution is the local insertion of an inverter biased between VPCX and HVNEG, as indicated in Fig. 13.26.

At this point, we still have the problem of where to connect the bulk of M3. Connecting it to the sense amplifier output would slow down the settling of the output voltage since node SAOUT would have a very high capacitance especially if a single sensing circuit is used for a row of vertical sectors, as previously described. In fact, all the parasitic capacitors between the n-well and the p-substrate of these transistors should be biased at  $V_{SAOUT}$ .

Therefore, we have to connect the bulk of M3 to the VPCX potential. However, in this way, due to the body effect, the threshold voltage of M3 (9 V in the case of VPCX) reaches 1.8 V and, hence, it is not sufficient to use only this transistor to

transfer the sense output voltage to the SUBMWL's, since it would not be possible to transfer voltages around 1 V. It is then necessary to add an n-channel transistor in parallel to M3, directly driven by the MWL.

With such a structure, the area occupation increases even though all the transistors introduced have minimum size, but they still have to fit in the row pitch.

In conclusion the memory is organized as in Fig. 13.27.

## 13.9 A/D Conversion

The analog-to-digital conversion phase is the one in which the analog value of a cell operating parameter (gate voltage or drain current) is translated into a digital value.

The voltage (or the current sourced by the cell) is compared with a set of references and, depending on the result of this comparison, it is possible to obtain the value of the bit stored. Hereafter, some methods used to accomplish such a task are proposed.

The first technique, directly derived from the bi-level case, is the comparison of the analog value with respect to all the references. The comparison can be carried out in parallel, in series, or in a hybrid way.

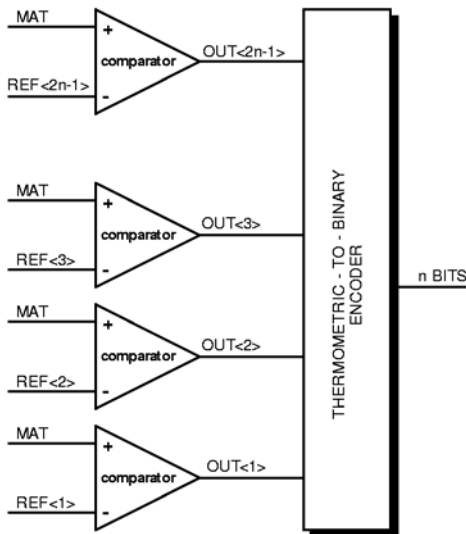


**Fig. 13.28.** Parallel sensing architecture

The parallel method basically uses a flash-type A/D converter as shown in Fig. 13.28. The final result is obtained during the time of a single evaluation. The drawback is that one comparator is used for each level. Such a method is therefore

expensive in terms of both area and power consumption, making it practically unfeasible in the case of 16 levels. Furthermore, we would also have the problem of reproducing the analog value derived from the array cell a large number of times, with the resultant limitations in terms of precision. The decision of not using the same array signal for all the converters is chiefly due to the need for limiting the kick-back effect induced by the simultaneous toggling of a large number of comparators.

A method that does not introduce these disadvantages is the serial one: a single comparator is used with obvious saving in terms of area and power consumption. The voltage of the memory cell is applied to one of the comparator inputs, while the references are applied to the other input, one at a time. According to the reference that has caused the comparator to toggle, the data are decoded. Two implementations are possible. In the standard serial method, the references are applied to the comparator in a monotonic way. For example, starting from the lowest reference value and continuing in increasing order, the digital conversion is carried out based on the number of levels compared (Fig. 13.29). Unfortunately, the disadvantage is that the conversion time depends on the data stored. In the case of 16 levels, 15 comparisons are necessary before accomplishing the decoding of data corresponding to the last level.

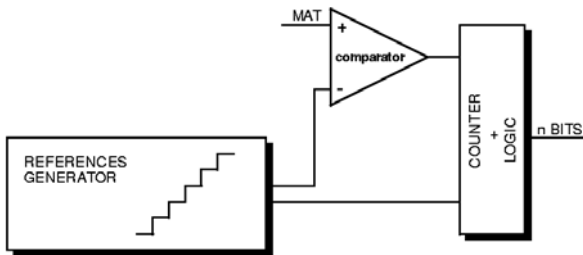


**Fig. 13.29.** Classic serial conversion method

If we design a more complicated control circuitry, the successive approximation method can be implemented. The read operation starts by comparing the analog value of the cell parameter with the value of the center reference level (the eighth in the case of 16 levels). Analyzing the result of this operation, it is possible to determine which half of the range the stored values belong to. At this point, the middle value of the selected range is chosen to perform the second comparison. This operation is repeated until the data are decoded (Fig. 13.30). The number of comparisons equals the number of bits stored in the single cell. Although faster than the classic one, also this method is quite slow. It is also necessary to pay attention to the area required for the additional logic so as not to waste the savings due to the serial method.

To overcome the problems related to the methods previously described (power consumption and area occupation for the parallel approach, long decoding time for the serial one), a mixed approach has been developed, which represents a hybrid

solution between serial and parallel. Theoretically, it represents a generalization of the dichotomic serial method. The logic of the successive approximations is still used but several comparisons are performed in parallel. The dichotomic method allows dividing the references in two groups. If the number of comparisons simultaneously performed is  $p$ , the references will be divided in  $(p + 1)$  groups. The operation is then repeated on one of the  $(p + 1)$  groups, and so forth. In the case of 16 levels, it is possible to carry out the conversion in only two steps, decoding two bits at a time, as shown in Fig. 13.31. In general, assuming that the same number of bit ( $z$ ) is decoded at each step,  $n/z$  steps are necessary to sense  $n$  levels. The number of comparators that have to be used is  $(2^z - 1)$ . This method represents a very good compromise between area consumption, power consumption, and timing performance.

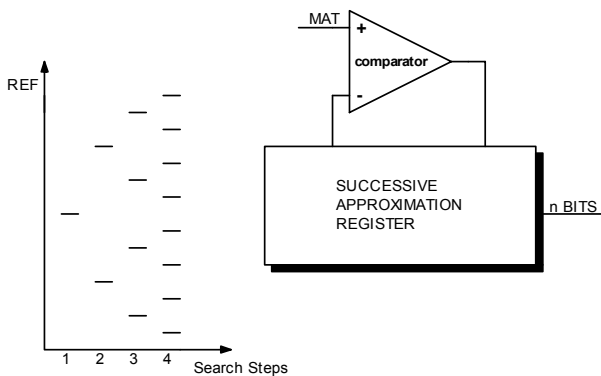


**Fig. 13.30.** Dichotomic serial conversion method

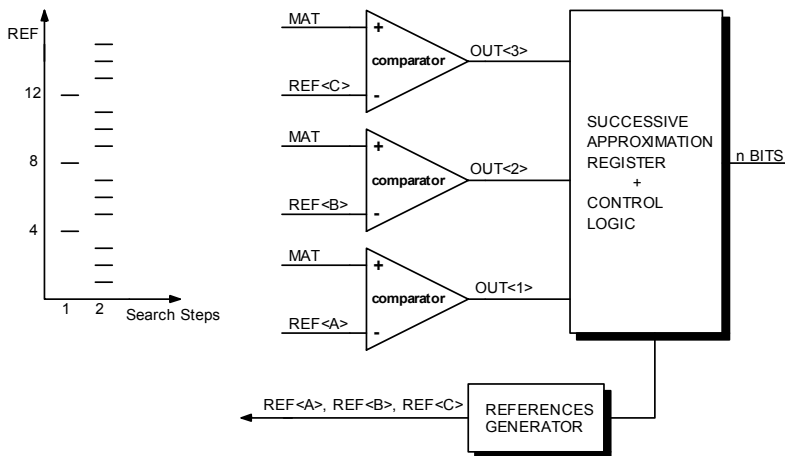


**Fig. 13.31.** Mixed parallel-serial sensing technique. During the first conversion step: REF<A> = REF<4>, REF<B> = REF<8>, REF<C> = REF<12>

### 13.10 Low Power Comparator

It should be noted that the voltage obtained as a result of signal production in a voltage mode approach is equal to the programmed threshold voltage plus an overdrive voltage  $\Delta V$  which depends on the reference current drawn by the cell. Signal recognition has therefore to process voltage signals placed in a range equal to the available  $V_T$  window shifted by  $\Delta V$ . This implies a high input common mode range (CMR) of the comparator, in the range of  $V_{PP}$  ( $>V_{DD}$ ). As the power efficiency of charge pumps is low and high reading parallelism has to be provided, signal production and recognition must be carried out with limited power consumption.

As a consequence, the comparator design aims at power consumption optimization, while still ensuring the above input CMR as well as required comparison time and accuracy. To achieve adequate sensitivity and high operational speed, the comparator is made up by an input buffer followed by a regenerative stage. The required accuracy is obtained by adopting an output offset storage technique.

Figure 13.32 shows the block diagram of the comparator. Blocks  $G_{m1}$  and  $G_{m2}$  represent two fully differential stages, which share their load resistors  $R$  and  $R'$ . The first stage (powered by  $V_{PP}$ ) is devoted to input signal buffering, providing a low voltage gain ( $A_v = 2$ ) in order to minimize the current drawn from  $V_{PP}$ .

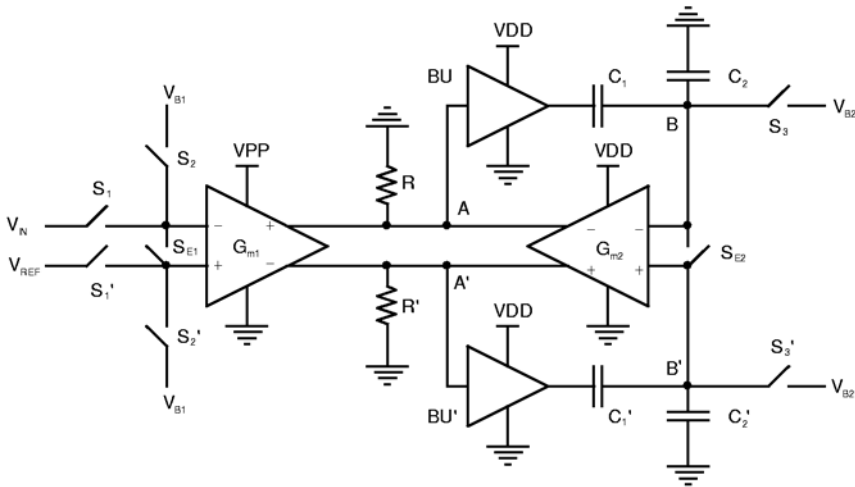


**Fig. 13.32.** Simplified comparator block diagram

The second stage (powered by  $V_{DD}$ ), together with feedback-connected capacitors  $C_1$  and  $C_1'$ , provides the regenerative action required to achieve high gain at high speed.

The two feedback capacitors are driven by unity-gain voltage buffers BU and BU' (which are also powered by VDD) to minimize the capacitive load of output nodes A and A'.

Offset cancellation is carried out on both differential stages, so as to minimize residual input-referred offset. Additional capacitors  $C_2$  and  $C_2'$  are used to reduce clock feed-through effects generated when turning switches  $S_3$ ,  $S_3'$  on and  $S_{E2}$  off.  $V_{B2}$  is a bias voltage.

To reduce total power dissipation, a two-phase offset cancellation technique can be used. Indeed, this approach allows the time duration of the working phases to be relaxed as compared to a three-phase approach with the same comparison time.

In conventional two-phase offset cancellation techniques, the first working phase provides input equalization and output offset storage, while input signal amplification and output signal regeneration are carried out simultaneously in the second working phase.

However, the comparator operation at the beginning of the second phase (that is, when starting the regenerative action) is affected by disturbs coming when opening the equalization switches of the input stage ( $S_2$ ,  $S_2'$  and  $S_{E1}$ ) and closing input switches ( $S_1$  and  $S_1'$ ). Indeed capacitive coupling with control phases, charge injection, and charge sharing between input lines and the comparator inputs take place. Moreover, the inverting input of the first differential stage suffers by larger disturb effects as compared with its non-inverting counterpart, because the input signal  $V_{IN}$  is common to all comparators in the sensing circuit. This can lead to transient inversion in the input signal polarity, which triggers the regenerative stage in the wrong direction. This effects turns out to be particularly severe in the presence of a non negligible impedance of the input signal generator.

To overcome this drawback, a different working phase sequence can be adopted. During the first phase, the input signal ( $V_{IN} - V_{REF}$ ) is applied to the input stage ( $S_1$ ,  $S_1'$  on;  $S_2$ ,  $S_2'$ ,  $S_{E1}$  off), while the input nodes of the regenerative stage are equalized ( $S_3$ ,  $S_3'$ ,  $S_{E2}$  on). This way, the input signal as well as offset contributions by both stage are amplified by a predetermined small factor, and the resulting voltage is stored on capacitors  $C_1$  and  $C_1'$ .

During the second phase, the input nodes of the first stage are equalized ( $S_1$ ,  $S_1'$  off;  $S_2$ ,  $S_2'$ ,  $S_{E1}$  on), while the positive feedback around the second stage is enabled ( $S_3$ ,  $S_3'$ ,  $S_{E2}$  off). As a consequence of the input node equalization, a voltage drop equal to  $G_{m1}R(V_{IN} - V_{REF})$  takes place across the output terminals A and A', which, in turn, causes a voltage imbalance across the input terminals of the regenerative stage (nodes B and B'):

$$V_B - V_{B'} = G_{m1} \cdot R \cdot \frac{C_1}{C_1 + C_2} (V_{IN} - V_{REF}) \quad (13.4)$$

The above voltage imbalance, which is offset compensated, starts the regenerative action of the second stage, thus providing the desired high-speed large output swing.

As the input terminals of the comparator are equalized by  $S_{E1}$  during the regeneration phase, any disturb effect at the comparator input can be regarded as a common mode contribution and is therefore rejected by the input differential

stage. Residual offset is mainly due to charge injection caused by the switches connected to the input nodes of the regenerative stage, which are turned off at the beginning of the second phase. To minimize the contribution due to mismatches between  $S_3$  and  $S_3'$ , a small delay is provided before switching  $S_{E2}$  off. Moreover, charge injection effects caused by the latter switch, are reduced by additional capacitors  $C_2$  and  $C_2'$ .

**Problem 13.6:** Transform the block diagram in Fig. 13.32 in a circuit schematic.

With a comparator such as the one described above, a comparison time of about 20 nanoseconds can be achieved, assuming an input signal of 10 mV. The VPP consumption is around 10  $\mu$ A.

## Bibliography

- M. Bauer, R. Alexis, G. Atwood, B. Baltar, A. Fazio, K. Frary, M. Hensel, M. Ishac, J. Javanifard, M. Landgraf, D. Leak, K. Loe, D. Mills, P. Ruby, R. Rozman, S. Sweha, S. Talreja, and K. Wojciechowski, "A multilevel-cell 32Mb Flash memory", in 1995 IEEE Int. Solid-State Circuits Conf. Dig. Tech. Pap., pp. 132–133, (Feb. 1995).
- G. Campardo et al., "40mm2 3V Only 50 MHz 64 Mb 2-b/cell CHE NOR Flash Memory", J. of Solid State Circuits, 35, 1665, (2000).
- G. Campardo and R. Micheloni "Architecture of non volatile memory with multi-bit cells" Elsevier Science, Microelectronic Engineering, Volume 59, Issue 1-4, page. 173-181, (November 2001).
- Y.-J. Choi, K.-D. Suh, Y.-N. Koh, J.-W. Park, K.-J. Lee, Y.-J. Cho, and B.-H. Suh, "A high speed programming scheme for multi-level NAND Flash memory", in 1996 Symp. VLSI Circuits Dig. Tech. Papers., pp. 170–171, (June 1996).
- C. de Graaf, P. Young, and D. Hulsbos, "Feasibility of multilevel storage in Flash EEPROM cells", in Proc. 25th European Solid State Device Research Conf., pp. 213–216, (Sept. 1995).
- M. Grossi, et al., "Program schemes for multilevel Flash Memories", IEEE Proceeding of the, Vol. 91, No. 4, pp. 594–601, (April 2003).
- M. Horiguchi, M. Aoki, Y. Nakagome, S. Ikenaga, and K. Shimohigashi: "An experimental large-capacity semiconductor file memory using 16-levels/cell storage," IEEE J. Solid-State Circuits, vol. 23, no. 1, pp. 27–33, (Feb. 1988).
- T.S. Jung, Y.-J. Choi, K.-D. Suh, B.-H. Suh, J.-K. Kim, Y.-H. Lim, Y.-N. Koh, J.-W. Park, K.-J. Lee, J.-H. Park, K.-T. Park, J.-R. Kim, J.-H. Lee, H.-K. Lim, "A 117-mm<sup>2</sup> 3.3V only 128-Mb multilevel NAND Flash memory for mass storage applications", IEEE J. Solid-State Circuits, vol. SC-31, pp. 1575–1583, (November 1996).
- R. Micheloni, O. Khouri, S. Gregori, A. Cabrini, G. Campardo, L. Fratin, and G. Torelli, "A 0.13- $\mu$ m CMOS NOR Flash memory experimental chip for 4-b/cell storage," ESSCIRC 28th Proc. European Solid-State Circuit Conf., pp. 131–134, (Sept. 2002).
- R. Micheloni, I. Motta, O. Khouri, and G. Torelli, "Stand-by low-power architecture in a 3-V only 2-bit/cell 64-Mbit Flash memory", in Proc. 8th IEEE Int. Conf. Electronics, Circuits, and Systems, vol. II, pp. 929–932, (Sept. 2001).
- R. Micheloni et al., "Method for reading a multilevel nonvolatile memory and multilevel nonvolatile memory", USA patent No. 6,301,149, (October 9, 2001).
- R. Micheloni et al., "Read circuit for a nonvolatile memory", USA patent No. 6,327,184, (December 4, 2000).

- 
- A. Modelli, R. Bez, A. Visconti "Multi-level Flash Memory Technology", 2001 International Conference on Solid State Devices and Materials (SSDM), Tokyo, Extended Abstract, pp. 516-517, (2001).
- A. Modelli, A. Manstretta, and G. Torelli, "Basic Feasibility Constraints for Multilevel CHE-Programmed Flash Memories", IEEE Trans. Electron Devices, ED-48, NO. 9, pp. 2032-2042, (September 2001).
- M. Ohkawa et al., "A 9.8mm<sup>2</sup> Die size 3.3V 64Mb Flash memory with FN-NOR type Four level cell", IEEE, Journal of Solid-State Circuit, vol. 31, no. 11, p. 1584, (Nov. 96).
- B. Riccò, G. Torelli, M. Lanzoni, A. Manstretta, H. E. Maes, D. Montanari, and A. Modelli, "Nonvolatile multilevel memories for digital applications", Proc. IEEE, vol. 86, pp. 2399-2421, (Dec. 1998).
- P. L. Rolandi, et al.: "1M-cell 6b/cell analog flash memory for digital storage", IEEE ISSCC Dig. Tech. Papers, pp. 334-335, (Feb. 1998).
- P.L. Rolandi et al., "A 32Mb-4b/cell analog Flash Memory Supporting Variable Density with Only Supply and serial I/O", 25th ESSCIRC '99.
- R. Shirota, G.J. Hemink, K. Takeuchi, H. Nakamura, and S. Aritome, "A new programming method and cell architecture for multi-level NAND Flash memories", presented at the 14th IEEE Non-Volatile Semiconductor Memory Workshop, Monterey, CA, paper 2.7, (Aug. 1995).

# 15 Circuits Used in Program and Erase Operations

Nowadays, every Flash memory device contains integrated charge pump circuits: this feature is mandatory because power supplies have been reduced but at the same time internal high voltages are required by the memory to get the electric fields needed by both program and erase operations.

## 15.1 Introduction

In the past few years, the use of portable devices has considerably increased; the key feature for these applications is reduced power consumption, which calls for chips that operate from a single power supply. This is the main cause for the transition from dual voltage (VPP and VDD where VPP > VDD) to single voltage (VDD) memory design. Moreover supply voltage tends to decrease in order to limit power consumption.

One of the main issues related to voltage scaling is to insure the proper operation of the device, i.e. the correct execution of all those operation which call for voltages greater than power supply, such as erase and program of the Flash memory (Table 15.1). In fact, the physical phenomena and the electrical fields that must take place when these operations are executed are not possible with present supply voltages (from 1.8 V to 3 V), unless the gate oxide thickness is reduced. Such a solution cannot be implemented, since an excessive reduction of the oxide thickness would negatively affect the ability of the cell to retain information, thus compromising its reliability.

**Table 15.1.** Biasing of the terminals of the memory cell during the different operations

| Operation | N-well | Ip-well   | Source | Gate      | Drain    |
|-----------|--------|-----------|--------|-----------|----------|
| Read      | 3 V    | 0 V       | 0 V    | 2 ... 6 V | 0.8 V    |
| Program   | 3 V    | - 1/- 2 V | 0 V    | 8 V       | 4.5 V    |
| Erase     | 8 V    | 8 V       | 8 V    | -8 V      | Floating |

To summarize, the single voltage supply memories require circuits which are able to generate the (high) voltages required by the various operations starting from the (low) power supply. Such circuits are usually based on the charge pump principle, and therefore they are often referred to as “charge pumps”.

## 15.2 Dual Voltage Devices

In the case of dual voltage devices, read, write and erase operations are carried out using a high voltage, VPP, supplied to the flash memory. It is therefore unnecessary to have on-chip high voltage generators, since all the voltages greater than VDD can be obtained from VPP through an appropriate voltage partitioning.

In addition, for devices realized using technologies without p-channel transistors, the main issue is how to completely transfer such a voltage.

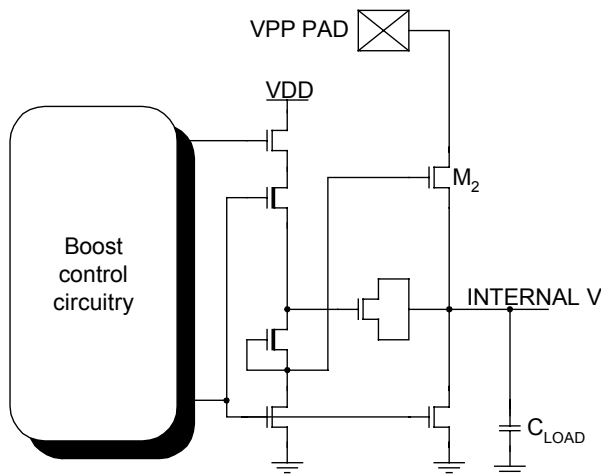


**Fig. 15.1** VPP switch for an NMOS device

A circuit that can be used for this purpose is shown in Fig. 15.1. It is a circuit where a bootstrap is used to bias the gate of the  $M_2$  pass transistor at a value greater than VPP to prevent the internal voltage from being limited by the  $V_T$  of the pass itself. The only drawback of this circuit is related to the  $W$  of  $M_2$ : this transistor can be as wide as 1 mm if the load capacitances connected (to the output) are significant (e.g. both row and column capacitances) and it is necessary to charge them in a reasonable amount of time.

The externally provided voltage can be used to bias the gate or the drain of the cells during program; depending on its usage, the circuit must be properly designed and/or modified.

If the VPP voltage is used to bias the gate of the cells, then the node called INTERNAL V, i.e. the node of the circuit shown in Fig. 15.1, is connected to the supply node of the row decoder.

The switch described above can be used to transfer into the chip both VPP and VDD voltages as shown in Fig. 15.2. Again, the main issue is the size of  $M_2$ ; in order to minimize this device, the circuit that requires the VDD supply during the read operation must be carefully selected.

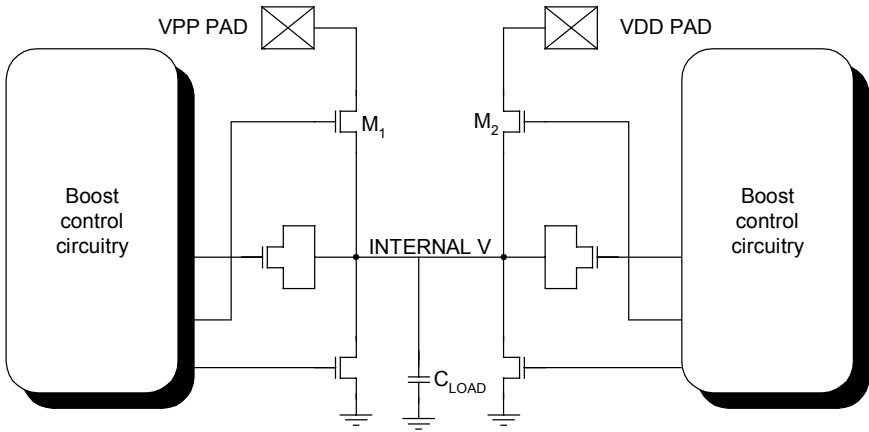


**Fig. 15.2.** VPC switch for the first NMOS devices

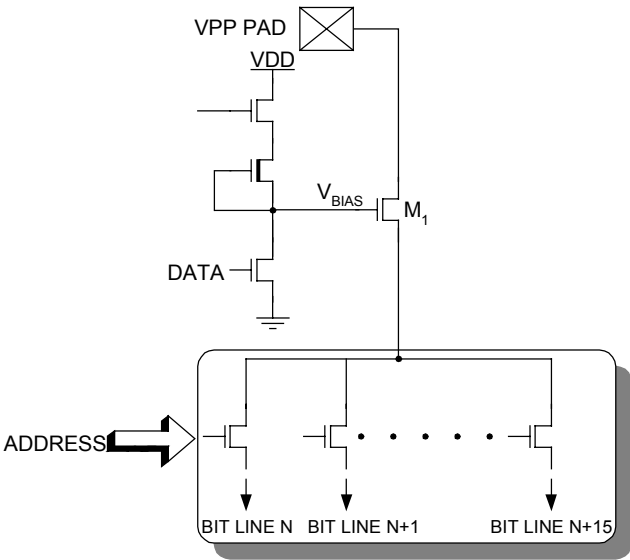


**Fig. 15.3.** Drain voltage without regulation

Whenever the voltage is used to bias the drain of the cell, it is necessary to employ a feedback circuit to regulate the voltage to the proper range, rather than let the voltage be determined only by the size of  $M_1$  and by the value of  $V_{BIAS}$  of its gate. In fact,  $V_{BIAS}$  is determined by the inverter that drives transistor  $M_1$  (Fig. 15.3). Variations of  $V_{PP}$ , temperature and current sunk by the cell can therefore cause a wide variation on the voltage applied to the drain.

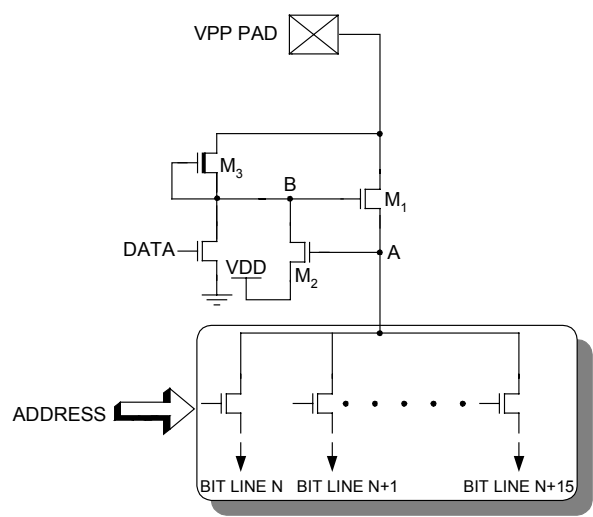


**Fig. 15.4.** Regulated drain voltage

The variation above described can be avoided by inserting a simple regulator as shown in Fig. 15.4. When the circuit is switched on, the voltage of node A increases until it turns on  $M_2$ . The gate of  $M_1$  is biased at a voltage value that is determined by the drive ratio between  $M_3$  and  $M_2$ . At this point, if the voltage of node A tried to increase,  $M_2$  would absorb more current from VDD, thus causing the gate voltage of  $M_1$  to decrease. Of course, the opposite would happen in case the voltage of node A tended to decrease.

## 15.3 Charge Pumps

Figure 15.5 shows the basic scheme of a charge pump: it is composed of a series of diodes and capacitors. Diodes are required to establish the direction of the current flow, while the main task of the capacitors is to accumulate the charge that is then transferred from one capacitor to the next capacitor, by driving the capacitors with the alternate phases CK and CK# (Fig. 15.6). A storage capacitor is placed at the output of the pump, to hold the final voltage.

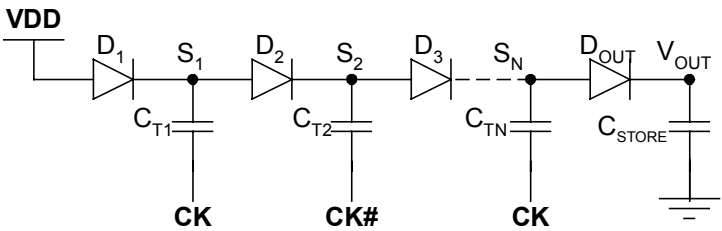


Fig. 15.5. Basic scheme of a charge pump

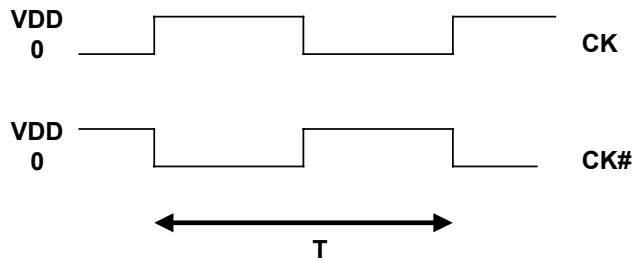


Fig. 15.6. CK and CK# phases; T is the clock period

Two important physical phenomena are the bases for the operation of a charge pump. The former is related to the basic capacitor characteristic: the voltage across the capacitor cannot change instantaneously, as shown in Fig. 15.7; the latter is the so called charge sharing phenomenon, which takes place when two capacitors precharged at different voltages are short-circuited. From Fig. 15.8 it follows that the final voltage of the short-circuited node depends on the initial voltages of the two capacitors as well as on their relative dimensions.

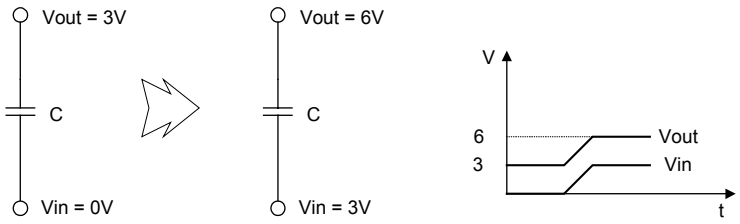


Fig. 15.7. Voltage variation at the terminals of a capacitor

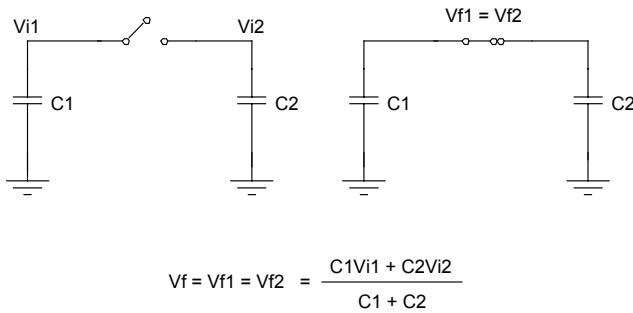


**Fig. 15.8.** Charge sharing principle

To understand how a charge pump works it is helpful to assume that:

- Diodes have a null threshold voltage;
- All transfer capacitors  $C_T$  are of the same size;
- CK and CK# are square waveforms (where CK# is the negated version of CK), initially at ground;
- $S_i$  nodes are initially precharged to VDD

At the power up of the CK and CK# signals, charge storage and transfer begin:

1. when CK goes from ground to VDD,  $S_1$  becomes equal to 2 VDD (the initial value (VDD) + VDD). At this point, a portion of the charge previously stored in capacitor  $C_{T1}$  is transferred to capacitor  $C_{T2}$ . Transfer stops when the voltage of node  $S_2$  is equal to the voltage of node  $S_1$ . Diode  $D_1$  prevents the charge from going back to the supply of the circuit, forcing the charge to go into capacitor  $C_{T2}$ ;
2. when CK goes from VDD to ground CK# rises from ground to VDD and capacitor  $C_{T2}$  then transfers a portion of its stored charge to capacitor  $C_{T3}$ . At the same time capacitor  $C_{T1}$  that partially discharged during the previous phase is recharged to the value of VDD.

The voltage of the output node continues to increase by an increment that becomes smaller and smaller until the voltage  $V_{OUT}$  reaches the value of  $VDD + n VDD$ , where  $n$  is equal to the number of stages of the pump, while the internal nodes voltage reaches a maximum value of  $V(S_i) = VDD + iVDD$ . As soon as these voltages are reached, no further charge transfer across the diodes takes place.

Following an example shows how the voltages at nodes  $S_i$  and  $V_{OUT}$  of an ideal, three-stages pump vary (Fig. 15.9). A power supply of 3 V and a storage capacitor are assumed.

The initial conditions for the circuit are those previously described: CK and CK# are at ground while the intermediate stages  $S_i$  and the output node  $V_{OUT}$  are precharged at VDD.

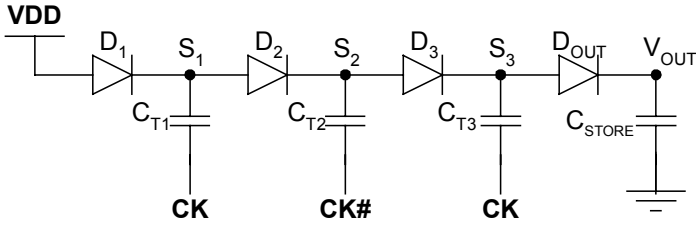


**Fig. 15.9.** Three-stages charge pump

As soon as the circuit is activated, CK goes to VDD, while CK# remains at ground.  $S_1$  and  $S_3$ , thanks to the boost effect, will instantaneously charge at a voltage equal to 6 V and they will start to transfer charge to capacitor  $C_{T2}$  and to the storage capacitor  $C_{STORE}$  respectively. In this way the voltages at the terminals of  $C_{T1}$  and  $C_{T3}$  decrease, while the voltages at the terminals of  $C_{T2}$  and  $C_{STORE}$  increase. Assuming that all the capacitors have the same size and that the voltage drop on diodes is zero, it follows that the charge transfer stops as soon as the voltages at the terminals of the two pairs of capacitors are equal, i.e. when  $S_1 = S_2 = 4.5$  V and  $S_3 = V_{OUT} = 4.5$  V.

$S_3$  is initially greater than  $S_2$  owing to the way the capacitances have been driven, i.e. by means of alternate signals CK and CK#. Diode  $D_3$  is therefore reverse-biased and charge moves exclusively from  $C_{T1}$  to  $C_{T2}$ .

On the other hand, a boost is applied to  $C_{T1}$  and  $C_{T3}$ , and when CK goes to GND,  $S_1$  and  $S_3$  go to 1.5 V ( $4.5$  V – VDD). Meanwhile CK# has reached VDD and therefore voltage  $S_2$  becomes equal to 7.5 V ( $4.5$  V + VDD). At this point  $C_{T2}$  starts to transfer charge to  $C_{T3}$ . But now  $C_{T3}$  is precharged at 1.5 V, therefore the final voltage on nodes  $S_2$  and  $S_3$  is equal to 4.5 V. At the same time capacitor  $C_{T1}$  charges to VDD through diode  $D_1$  that allows charge flow from VDD to  $C_{T1}$  as it is forward-biased.

Summing up, at the end of a clock period we have:

$$\begin{aligned} S_1 &= 3 \text{ V,} \\ S_2 &= 1.5 \text{ V,} \\ S_3 &= 4.5 \text{ V,} \\ V_{OUT} &= 4.5 \text{ V} \end{aligned}$$

At the end of the second clock period we get:

$$\begin{aligned} S_1 &= 3 \text{ V,} \\ S_2 &= 2.25 \text{ V,} \\ S_3 &= 5.25 \text{ V,} \\ V_{OUT} &= 5.25 \text{ V} \end{aligned}$$

Although, at a first glance, it seems that  $V_{OUT}$  is the only voltage tending to increase, it is easy to realize that the voltage of each single node increases as the cycles continue.

Maximum output voltage that can be achievable when no current load is connected to the output is:

$$V_{MAX} = VDD + nVDD \quad (15.1)$$

In the previous example  $V_{MAX}$  is equal to 12 V (3 V + 3 · 3 V)  
 Since the charge transferred at each clock cycle of either CK or CK# is:

$$Q = C_T \cdot \Delta V = C_T \cdot VDD \quad (15.2)$$

the current provided on the output in a clock period T is equal to:

$$I = \frac{Q}{T} = \frac{C_T \cdot VDD}{T} \quad (15.3)$$

From Eq. (15.1) and Eq. (15.3) it is easy to obtain the output resistance of the pump:

$$R_P = \frac{V_{MAX} - V_{SUPPLY}}{I} = \frac{(n+1) \cdot VDD - VDD}{\frac{Q}{T}} = n \frac{T}{C_T} \quad (15.4)$$

As an example, Fig. 15.10 shows the plot of the output resistance of a pump implemented in a device.

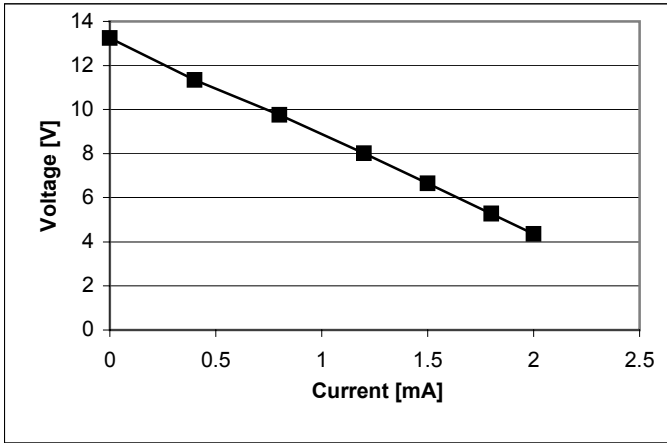


**Fig. 15.10.** Output resistance of a 3 V supplied charge pump made up of 4 stages.

Anyway it is necessary to point out that this equation holds true if and only if there is a complete transfer of the stored charges between a capacitor ( $C_{Ti}$ ) and the subsequent one ( $C_{Ti+1}$ ).

As far as we have seen until now, charge pump can be considered as a  $V_{MAX}$  voltage generator with a series resistor  $R_p$ ; therefore if a load sinking a constant current  $I_0$  were connected to the output, the general value for the voltage  $V_{OUT}$  would be:

$$V_{OUT} = V_{MAX} - n \frac{I_0 T}{C_T} = V_{MAX} - R_p I_0 \quad (15.5)$$

In this case,  $C_{Ti}$  capacitors must provide a charge equal to  $q = I_0 \cdot T$  to the output at each clock period  $T$ . For this reason a reduction of the voltage value occurs for both the corresponding nodes  $S_i$  and for the output node.

Now that the operating principle of the pump should be clear, it is possible to get nearer to the “real” model, taking into account the fact that the inserted diodes have a threshold voltage not equal to zero, which we will refer to as  $V_D$  in the following.

In this case the output voltage, as well as the voltage of the intermediate stages, is smaller than the voltage we have considered so far. In fact the charge transfer from capacitor  $C_{Ti}$  to capacitor  $C_{T+1}$  ends as soon as the voltage of the  $i$ -th stage is equal to the voltage of the  $(i + 1)$ -th minus  $V_D$ , simply because at this point the diode between the two stages becomes an open circuit. Resulting  $V_{MAX}$  voltage is therefore:

$$V_{MAX} = VDD - V_D + n \cdot (VDD - V_D) \quad (15.6)$$

It is important to point out that the size of the diodes must be carefully chosen: if they are too big, the associated area, which equates to capacitance, is no longer negligible, while if they are too small then the resistance of the pump  $R_p$  becomes high because of the series resistance represented by the diodes themselves.

What happens if we increase the number of stages? Considering Eq. (15.4) and Eq. (15.6), two opposite effects take place: on one hand, output voltage  $V_{MAX}$  increases; on the other hand, resistance  $R_p$  increases as well. The latter result is far from being desirable, because whenever a current absorption occurs, a high value of resistance implies a considerable decrease of the voltage  $V_{OUT}$  (which, in reality, is the available voltage) with respect to  $V_{MAX}$ .

Observing Eq. (15.4) we could think about reducing the resistance of the pump  $R_p$  by modifying either the size of the transfer capacitors or the duration of the clock period. Both solutions are not recommended.

Indeed it is a fact that there is usually a limited range of clock periods (often just one clock period) that allow us to get the maximum output voltage; such a period is often referred to as optimum. Figure 15.11 shows voltage  $V_{OUT}$  of a pump as a function of different clock periods. It is worth noting that as the current absorbed by the pump varies,  $V_{OUT}$  changes as is expected, while the optimum period does not vary.

The reason behind the existence of a limited range of values is related to the non-ideality of the components used to realize the pump: in fact, in spite of what happens in the ideal pump, charge transfer from one capacitor to the following one does not take place instantaneously, but it depends on several factors like, for instance, the threshold voltage of the diodes, the driving capability of CK and CK# phases generators, typical leakage currents of the capacitors, parasitic capacitances, etc.

To summarize, reducing the clock period while maintaining the size of the capacitors could prevent a complete transfer of the stored charge from happening: the driving of the capacitor would be less effective and the output voltage would not be at its maximum; the same undesired result would be achieved if the size of the capacitors were increased without changing clock period.

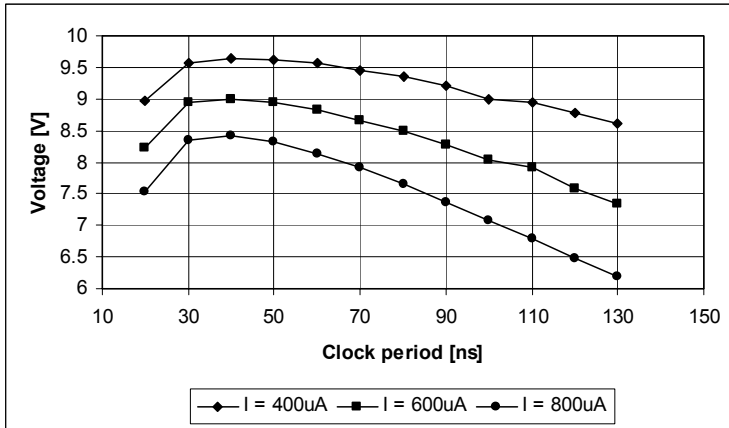


**Fig. 15.11.** Optimum period for a charge pump

Among the non-ideality previously listed, parasitic capacitances deserve further discussion. By “parasitic capacitances” we mean all those capacitances connected (directly or indirectly) to the transfer capacitors. They are charged during one half-cycle, and in the following half-cycle they transfer the accumulated charge towards ground, instead of transferring it to the storage capacitors.

The charge of these capacitors is therefore necessary for the proper operation of the circuit, since their presence can not be avoided but it is counterproductive to generating the output voltage because part of the charge is “stolen” from the transfer capacitors. Going further into detail, the parasitic capacitance  $C_p$  consists of:

1. capacitance between the lower plate of the capacitor and ground;
2. capacitance between the upper plate of the capacitor and ground;
3. capacitance between the anode-substrate and the cathode-substrate typical for the diode.

Resulting voltage  $V_{MAX}$  taking this effect into account is:

$$V_{MAX} = VDD - V_D + n \cdot \left( \frac{C_T}{C_T + C_p} VDD - V_D \right) \quad (15.7)$$

where  $C_T$  is the value of the transfer capacitors.

Designing a charge pump, therefore, means to explore solutions to realize an ideal voltage generator by overcoming the non ideal characteristics of components, which tend to limit the output voltage below the ideal value.

## 15.4 Different Types of Charge Pumps

Thanks to the continuous improvements of technologies to realize memory devices and the evolution of the electronic circuits design techniques, novel designs of charge pumps have been devised, which have improved performance, reduced  $R_p$  resistance and increased output voltage.

### 15.4.1 Dickson Pump Based on Bipolar Diodes

The Dickson pump is the one of the oldest type of pump, and from a structural point of view it is identical to the pump shown in Fig. 15.5: an alternating series of diodes and transfer capacitors driven by two opposite phases (CK and CK#), which in turn do not need delays or phase shift.

Even if both the pump and the phase generator are quite easy from a circuitual point of view, such a structure cannot be effectively realized in those technologies where triple-well is not available. In fact, in order to design a working circuit, it is advantageous to use bipolar diodes, which are typically realized by means of n-diffusion in a p-well tub.

### 15.4.2 Dickson Pump Based on Transistor-Based Diodes

In order to make the charge pump independent of a given technology, it is possible to substitute the bipolar diodes of the pump previously described with diodes realized using properly connected MOS transistors as shown in Fig. 15.12.

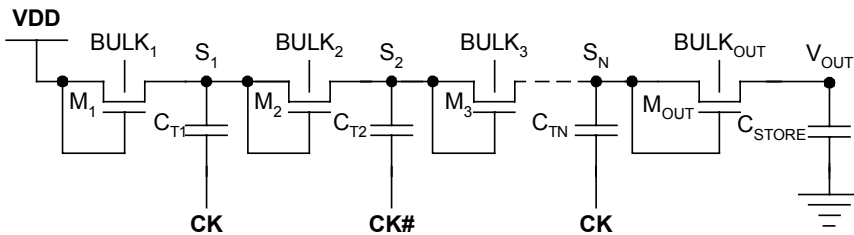


Fig. 15.12 Diodes realized using MOS transistors

One of the advantages of this solution is that the latch-up phenomenon, which is typical of pumps based on bipolar diodes, can be avoided. The first issue to cope with when designing this pump is the choice of where to connect the body terminal of the inserted transistors. The only possible solution is to connect the bulk of the NMOS (diode-connected) to ground if we want to use NMOS transistors without triple well. Indeed, in this way we are guaranteed that the body is not only connected to the lowest voltage of the circuit, but also that it is stable. Even if such a solution leads to a relatively simple circuit, the drawback is that the output voltage is considerably reduced because of body effect. In fact, the threshold voltage of the transistors tends to increase stage after stage until it reaches its maximum value on the output transistor. If we decide to replace the diodes with PMOS transistors instead of NMOS, the issue of where to connect the body must be considered, since it is difficult to say which voltage best fits, considering the behavior of the voltages at the terminals of the transistors; indeed, it is not possible to find a voltage between drain and source which is constantly higher.

There are simple solutions that overcome the body effect issue; the drawback is that in the case of an NMOS it is again necessary to use a triple well technology in order to be able to connect the bulk to a voltage that is different from the substrate voltage.

Figure 15.13 shows an example of a bulk bias circuit: it is a voltage divider circuit composed of diodes: the voltage of the intermediate nodes can be used to bias the bulk of each charge pump transistor. Diodes are used instead of resistors in order to save area with the same power consumption; moreover since the current is provided by the pump, only a minimum current can be used, without compromising the performance of the pump itself. From Fig. 15.13 it is possible to see that the intermediate nodes of the divider ( $P_i$ ) are connected to the bulk through a NMOS diode-connected transistor and a capacitor. These elements are used to filter the voltage of nodes  $P_i$  since these nodes are not fixed, but they have an oscillation proportional to the output voltage oscillation (which in turn is related to the current required on the output).

Using this structure, called a bulk biaser, it is possible to reduce the body effect by connecting the bulk nodes of the pump transistors to a voltage slightly lower than the corresponding drain and source nodes.

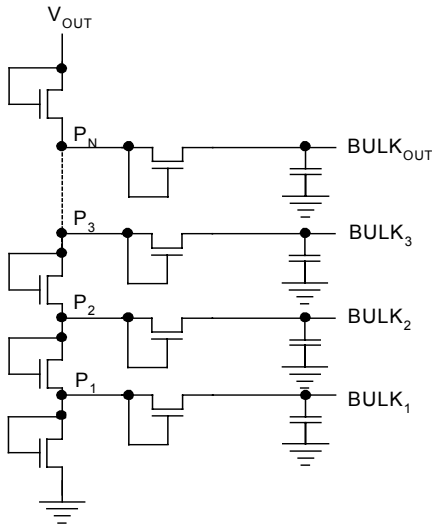


**Fig. 15.13.** Bulk biaser

Alternatively, a dynamic biasing of the bulk nodes could be devised. The structure shown in Fig. 15.14 allows biasing the bulk nodes of the pump transistors to the lowest voltage between drain and source of the transistor (or the highest one in case the same structure would have been realized using PMOS transistors) exploiting the stages of the pump itself. The disadvantage of this solution is that the bulk nodes of the pump might be not biased during the transient between one semi-

period and the following one, because of the delay between the turning on of the two transistors  $M_1$ ,  $M_2$ . This risk must be carefully evaluated in the design phase taking into account the parasitic capacitances that influence the stages, the tubs of the stages themselves, the commutation time of the phases, etc.

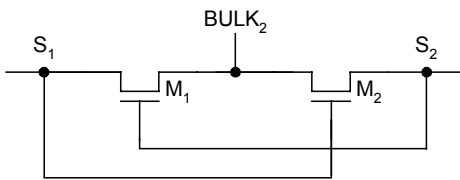


**Fig. 15.14.** Dynamic biasing of the bulk nodes

**Problem 15.1:** Analyze in detail the behavior of the circuit in Fig. 15.14.

The performance of the pump that uses the diode-connected transistors instead of the bipolar diodes is influenced by two effects:

1. parasitic resistance of the transistor, that effects the optimum period;
2. the threshold voltage of the transistors, which is typically higher than the turn on voltage of a bipolar diode, which directly effects the output voltage  $V_{OUT}$ .

Since the two solutions (NMOS/PMOS) share the same issues, it is reasonable to wonder if one is better than the other, and why.

The pumps realized with PMOS are for sure simple and reliable architectures. Anyway, in order to have a negligible parasitic resistance introduced by the transistor (in particular if supply voltage is very low, e.g. 1.8 V), the size of the transistor must be quite big, and area occupation becomes an issue.

On the other hand, the use of NMOS transistors realized in triple well is the only way to reduce the body effect. In this case it is mandatory to control the biasing of each well of the transistor, in order to prevent any forward biasing of the junctions.

### 15.4.3 Charge Pump Based on Pass Transistors

The diodes inside the pump are needed to establish the current direction and to prevent any charge from flowing towards the previous stage. The main drawback of the use of such components, apart from the technological complexity to realize them, is their high threshold voltage, i.e. the need of having a consistent voltage difference between drain and source, in order to have a charge transfer. It is clear that under these conditions the charge transfer between one stage and the following one cannot be complete.

The alternative to the diodes to obtain both charge transfer and node connections is represented by transistors used as switches. MOS behavior is involved, i.e.



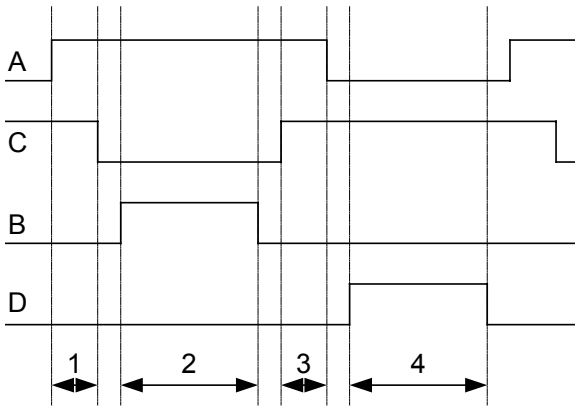


**Fig. 15.16.** Timing diagram of the phases for the pass-transistor pump: 1) precharge of capacitor  $C_p$  of transistor  $M_1$ , 2) charge transfer between capacitors  $C_{T-1}$  and  $C_T$ , 3) precharge of capacitor  $C_p$  of transistor  $M_3$ , 4) charge transfer between capacitors  $C_T$  and  $C_{T+1}$

If we could use N-channel transistors in triple-well, we might think about connecting the body of the switch to its drain.

In this way, not only do we save area on silicon, because structure like bulk bi-aser or dynamic bulk bi-aser are no longer required, but we also have a positive side effect, i.e. a bipolar diode is created inside the triple well transistor (Fig. 15.17), which starts transferring charge as soon as the voltage difference between drain and source is equal to its  $V_T$ . This fact has several consequences. On one hand it is possible to start the charge transfer from one capacitor to the following one earlier and the resistance for the NMOS transistor is reduced, since the charge transfer takes place through two paths: one is the channel when the MOS is activated, the other one is the bipolar (at least until the bipolar diode is active). On the other hand, the pre-charge of capacitors  $C_p$  might be still incomplete, therefore the boost of the gate of the transistors might become inefficient. Another issue due to the presence of the bipolar is the possibility that latch-up phenomena occur.

To transfer the charges from the last transfer capacitor to the storage capacitor on the output a diode-connected transistor or a further pass-transistor can be used. Both solutions have pros and cons.

A diode-connected transistor has the advantage of being easily designed, while the disadvantage is that it requires a voltage difference at its terminal in order to operate correctly; as a consequence the voltage transferred to the output is equal to the voltage of the last stage minus the threshold voltage of the diode. On the contrary, the pass-transistor allows transferring all the voltage to the output, but it requires a boost structure for its gate (identical to the one required by the internal stages of the pump) to be properly driven, i.e. to be able to correctly transfer the voltage. Power consumption increases when this pump structure is compared to the pumps previously seen. In fact the phases must be created “ad hoc”: delays must be carefully introduced so that under every operating condition of the pump simultaneous boost of the pass-transistors or of capacitors belonging to different



As soon as CK phase toggles from GND to VDD, the value of node A becomes VDD, activating transistor  $M_{N2}$ . Since phase CK# remains at GND, the charge starts flowing from power supply to capacitor  $C_{T2}$  until B reaches a value equal to  $VDD - V_{TH'N2}$ . When phase CK goes to GND again, it brings node A along, thus turning off transistor  $M_{N2}$ . Now phase CK# gets to the value of VDD, therefore the value of node B becomes  $VDD - V_{TH'N2} + VDD$  turning on transistor  $M_{N1}$ , through which  $C_{T1}$  is charged until it reaches VDD. Of course when CK# goes to GND again, node B is biased at a voltage equal to  $VDD - V_{TH'N2}$ . At this point the phases repeat. It is worth noting that each capacitance is boosted first, then charged, changing its value from VDD to 2 VDD. Since these operations are performed alternately on each half of the circuit, during a whole period either node A or node B is at 2 VDD (thus the reason for calling this circuit a “voltage doubler”). If, at this point, it is possible to design a circuit that can transmit on its output the 2 VDD voltage, independently of which node (either A or B) is biased at that value, then passing from the voltage doubler to the charge pump is relatively simple: in fact, it is sufficient to repeat such a structure where the output of a stage is used as the supply voltage for the following stage.

A PMOS output stage properly connected to nodes A and B, as the one shown in Fig. 15.19, is the simplest circuit to obtain an output voltage of 2 VDD: when CK is at VDD, A is at 2 VDD, while B is at VDD. Transistor  $M_{N1}$  is therefore turned off, while  $M_{P1}$  is active, transferring the voltage of node A to node OUT. In the meanwhile  $M_{P2}$  is turned off and  $M_{N2}$  is turned on, charging capacitor  $C_{T2}$ . When CK goes back to GND and CK# becomes VDD, then the circuit behaves in the opposite way:  $M_{N1}$  and  $M_{P2}$  are active (the former charges capacitor  $C_{T1}$ , the latter transfers the voltage of node B to the output) while  $M_{N2}$  and  $M_{P1}$  are turned off. The way the circuit is conceived, there are no direct paths between VDD and node OUT: these paths would jeopardize the functionality of the circuit itself.

Also in this case, the issue of the bulk biasing of the transistors exists (of course under the assumption that triple-well NMOS are used).

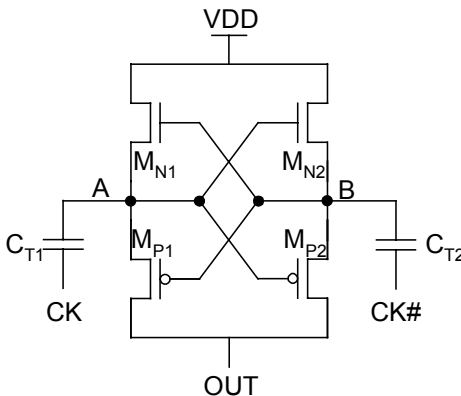


Fig. 15.19. Voltage doubler with PMOS output stage

Again, the solutions described in previous sections apply, and the same pros and cons hold true: one solution is to connect the bulk terminal to the power supply for the n-channel transistor and to the output node for the p-channel transistor, another one is the “dynamic biasing” of the bulk nodes. Figure 15.20 represents a possible circuit solution that allows keeping the body of the NMOS transistors to the lowest voltage, and the body of the PMOS transistors to the highest voltage, under any circumstances.

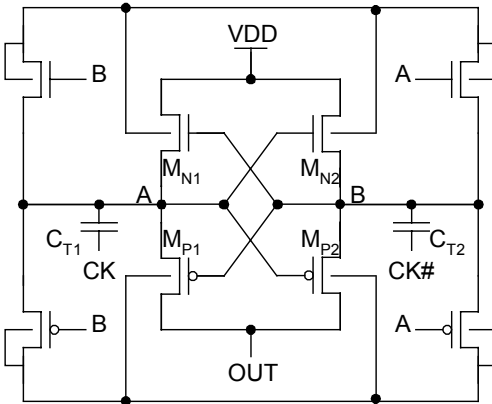


**Fig. 15.20.** Biasing of the bulk nodes of the transistors

A small specification is required at this point: when more stages are cascaded to realize a pump as shown in Fig. 15.21, we can decide to drive the capacitances of the stage with a signal whose swing is either GND/VDD or GND/supply of the stage. In the former case, the gain of each stage is about VDD; in the latter the gain is about 2 VDD. In any case, from an implementation point of view, the former solution is preferable.

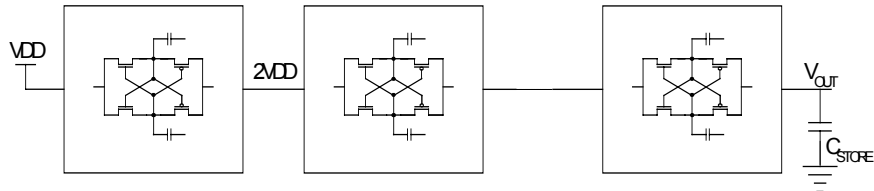


**Fig. 15.21** Implementation of a charge pump by cascading a series of voltage doubler circuits;  $V_{OUT}$  output voltage is equal to  $(n + 1) VDD$  if CK and CK# signals have a 0 – VDD swing, while it is equal to  $2^n VDD$  if the swing of CK and CK# for the  $i$ -th stage is 0 – supply of the  $i$ -th stage

This structure has several advantages:

- the possibility of driving the capacitances using only two complemented phases. This implies a relatively simple phase generator circuit with respect to the one used in the pumps described in the previous section;
- the possibility of realizing the circuit using “low voltage” transistors (i.e. transistors that can bear a voltage difference at their terminal lower than power supply) since the voltage difference at the terminals of each transistor does not go above VDD in any operating phase of the circuit;
- no boost capacitors are required for the pass transistors (since we can use the one implemented for the charge transfer), thus achieving a reduction of the area occupation and limiting the dissipated power to the one really needed to produce the desired voltage;
- a decrease of the ripple that naturally comes from the pump.

In fact, due to the operation of the charge pump and of the capacitors in general, the output of the pump is more or less noisy depending on the phase that is toggling at a given time. Let's refer to Fig. 15.5: when the clock CK goes high, the charge stored in the capacitor  $C_{T3}$  is transferred to the storage capacitor.  $V_{OUT}$  instantaneously charges to a value equal to the previous one plus the supply voltage; the corresponding waveform will show at that time an upward peak; then  $V_{OUT}$  discharges until CK goes back to GND and CK# goes to VDD. The waveform of  $V_{OUT}$  will show an upward peak again, but its height is less than the previous case, since the charge is now shared between two capacitors (i.e.  $C_{T3}$  and  $C_{STORE}$ ). Now the discharge of  $V_{OUT}$  occurs again until the new phase comes in.

When talking about “ripple”, we generally refer to the height of the “peaks” that can be found in the output node waveform.

Analyzing the output of the voltage doubler as a function of different frequencies it can be seen that it is better if the clock is as short as possible. This fact can be a limitation for the circuit, since in these kinds of applications, the clock generators are often the result of a trade-off between area occupation and precision: in other words the risk is that it may be difficult to implement structures that occupy a small area and at the same time guarantee the same clock period as voltage and temperature conditions vary.

### 15.4.5 Voltage Tripler

This circuit is derived from the voltage doubler and provides an output voltage whose value is three times the supply voltage value.

Figure 15.22 shows the schematic of such a stage.

Since the structure is symmetric, it is possible to understand the circuit behavior by studying just one half of it. Let's start again from the condition where both CK and CK# are forced to GND, i.e. all the transistors are turned off. When CK goes to VDD, CK# remains at GND. Under these conditions, B is biased at VDD, A is biased at 2 VDD and since  $M_{N5}$  is turned on, D is biased at VDD. On the contrary, F is driven to GND by  $M_{N6}$  and transistor  $M_{P4}$  is turned off.

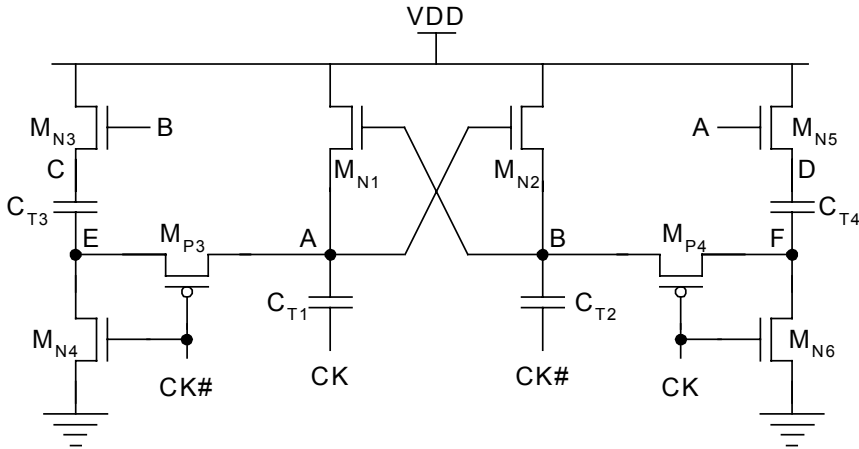


**Fig. 15.22.** Voltage Tripler

When CK# goes from GND to VDD, then B goes to 2 VDD while A goes to VDD, thus turning off  $M_{N5}$ .  $M_{N6}$  is turned on instead, since it is driven by the signal CK that is at GND.  $M_{P4}$  is turned on instead, and it short-circuits nodes B and F. Therefore F reaches a voltage of 2 VDD and, due to the boost effect, D is biased at 3 VDD.

The behavior of the portion of the circuit not analyzed yet is essentially complementary. In fact, when CK# is at VDD, capacitor  $C_{T3}$  is precharged at VDD: node C is kept at VDD, while E is driven to GND by  $M_{N4}$ ; when CK# is at GND, then E is biased at 2 VDD and C, due to the boost effect, is biased at 3 VDD.

At this point, we just need a structure that is able to output the 3 VDD voltage at each clock cycle. To achieve this result, we can use again the structure shown in Fig. 15.19.

Also in this case we need to cascade several stages in order to build a pump, and the output of a stage must be used as the power supply for the following one. If N is the number of cascaded stages, then the maximum voltage that can be achieved is equal to:

- $V_{MAX} = (N + 2) VDD$  if the phases that drive the stages have a GND/VDD swing (it is worth noting that the  $(n + 2)$  factor is due to the fact that for all the stages there is a voltage increment equal to VDD except for the first one, whose increment is 2 VDD);
- $V_{MAX} = 3^N VDD$ , if the phases that drive the stages have a swing between ground and the voltage of the driven stage.

Contrary to the two-phases pump analyzed above, this pump cannot be implemented using low voltage transistors only, since the voltage drop at the terminals of the transistors can go above VDD under certain operating conditions.

## 15.5 High Voltage Limiter

The voltage generated by the charge pump linearly varies as a function of the supply voltage. Taking into consideration that in certain applications the difference between the minimum and maximum supply voltage can be about a Volt (e.g. the devices that work with a power supply equal to 3 V must work at both 2.7 V and 3.6 V), the following two situations might both occur if we implement any sort of charge pump structure with a given and predefined number of stages:

1. get the voltage required by the operation that we want to perform on the cell, using a low supply voltage;
2. get a voltage that is above the maximum limit for the oxides of the cells, if we use high supply voltages.

Therefore it would be useful to cascade the charge pump with a structure that, acting on the pump itself, is able to limit the output voltage  $V_{OUT}$  to a value that cannot damage the memory cells, and acts independently from the supply voltage.

There are mainly two ways to achieve this result, i.e. applying either a continuous or an ON/OFF regulation. If the clock of the pump is realized using an oscillator (a cascade of an odd number of closed-loop inverters), it is quite simple to achieve a continuous regulation: it is sufficient to vary the number of inverters that compose the oscillator, thus obtaining a variation of the clock frequency. On the other hand, an ON/OFF regulation does not structurally change the clock generator, but the different phases that drive the capacitors of the pump are kept at ground when the pump itself has to be stopped. In general it is preferable to act on the phase generator if this structure is available, otherwise it is the clock generator that is turned on and off.

A typical circuit used for the ON/OFF regulation is the high regulator shown in Fig. 15.23.

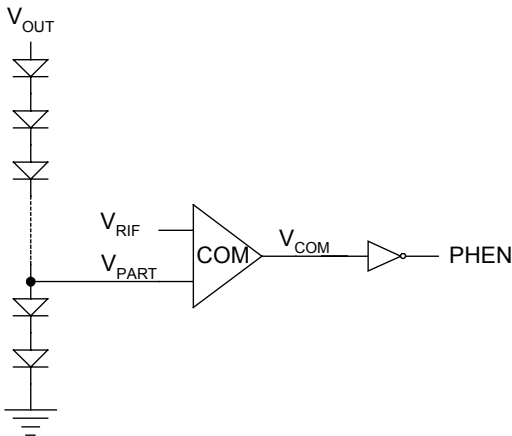


**Fig. 15.23.** High Regulator for the ON/OFF regulation

The output voltage of the pump,  $V_{OUT}$ , is sent to a divider circuit composed of low-conductivity diodes in order to limit as much as possible the current consumption of the charge pump. Voltage  $V_{PART}$ , whose behavior is the same as  $V_{OUT}$  in a smaller scale, is tapped from the divider circuit and compared with a reference voltage,  $V_{RIF}$ . As soon as  $V_{PART}$  becomes greater than  $V_{RIF}$ ,  $V_{COM}$  becomes VDD and PHEN goes to '0': the charge pump phases are immediately forced to GND, so that no charge transfer can occur between the capacitors; in this case,  $V_{OUT}$  voltage can only decrease since the charge absorbed by the circuits connected to  $V_{OUT}$  is not "replaced". Therefore  $V_{PART}$  decreases as well and as soon as it becomes smaller than  $V_{RIF}$ ,  $V_{COM}$  goes to GND and PHEN goes to VDD, thus turning on again the charge pump.

The effect of the limitation of the pump output voltage,  $V_{OUT}$ , to a precise value, is evident if the plot of the output resistance is analyzed: as shown in Fig. 15.24, the output voltage of the pump is about 8 V when the current absorption is between 0 and 1.2 mA; in the case of higher absorption, the regulator remains inactive.

From the description of the behavior of the high regulator, it should be clear that the value of the voltage which  $V_{OUT}$  is limited to can be easily changed by varying the reference voltage  $V_{RIF}$ .

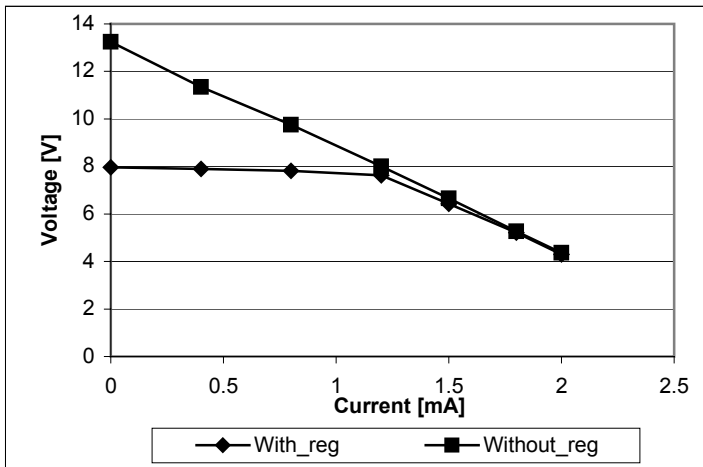


**Fig. 15.24.** Comparison between the output voltages of the same pump obtained with and without the use of a high regulator

## 15.6 Charge Pumps for Negative Voltages

Apart from the positive voltages, it is necessary to generate negative voltages inside the memory in order to bias the gate of the cells during the erase operation. Usually the negative pump is dimensioned by considering the time that is targeted to bias the gates of all the cells to be erased and therefore not taking into account a specific current demand (since no current is absorbed through the gate).

The operating principle of a negative pump is the charge transfer between one capacitor and the next one, as in the positive pumps case. The difference is that the transfer takes place from the storage capacitor toward ground, so that the output node can reach negative voltages.

In order to ease the charge transfer between one capacitor and the next one, the boost technique for the transfer capacitors is used, and either two or four phases are implemented depending on the desired type of pump.

The principle scheme of the negative charge pump is shown in Fig. 15.25.

If the triple-well technology is not available, the diodes of the pump must be implemented using P-channel transistors. The main drawback is again related to the body effect, which has a major influence primarily on the last stages. As we have already seen for the positive pumps, a possible solution to mitigate this negative effect is to use a pump realized using PMOS with dynamic biasing of the bulk nodes.

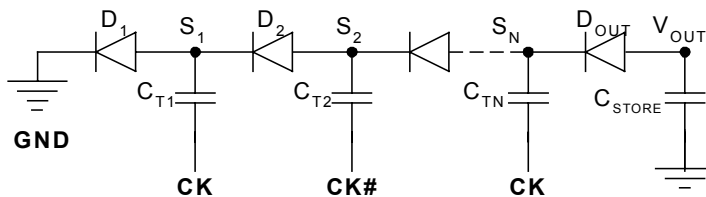


**Fig. 15.25.** Principle scheme of a negative charge pump

Since the output voltage of the pump depends on the number of stages, on the supply voltage and on the voltage drop on the transistors used for the charge transfer from a capacitor to the next one, the problem of guaranteeing a voltage that allows a proper erase operation (i.e. without damaging the cells) arises. It is therefore mandatory to use a high regulator that limits the output voltage  $V_{OUT}$ . The operating principle is similar to the one described in the previous section: the output voltage is compared with a reference one and, depending on the result of the comparison, the phase generator of the pump is turned either on or off. The main difference is related to the “reaction time” of this regulator: since it is only used during the erase operation, the comparator can be slower and the diodes can be smaller, thus calling for a smaller current consumption of the pump. The main drawback is that in this case the ripple on the output voltage of the pump can become large.

## 15.7 Voltage Regulation Principles

As it should be clear from the previous sections, the memory cells cannot be biased directly with the charge pump output voltage, which usually needs to be “filtered” to exclude its characteristic ripple and to limit it to a stable value. The circuit that performs these functions is the so-called voltage regulator<sup>1</sup>.

The simplest voltage regulator consists of an operational amplifier, which is composed by a differential input stage, an output stage and an appropriate feedback net; the chosen topology for each of these blocks depends on the required performances. For example, depending on applications, settling time and power consumption could be important parameters; some regulators have to supply a stable output voltage with zero dc current, others have to regulate the output voltage while sourcing a huge current to the load.

Also frequency compensation is a major parameter to be taken into account: frequency compensation is defined as the way to get stability when the operational amplifier is operated in closed loop with negative feedback.

In the following sections, an overview of problems and solutions related to the principles of voltage regulation is presented, specifically referring to the gate and drain voltage. Nevertheless, many other high voltages are required to perform Flash operations; refer to Chap. 16 to have a complete overview of the high voltages required in last generation Flash memories.

## 15.8 Gate Voltage Regulation

### 15.8.1 Circuit Structure

In the earlier Flash devices, the gate of the cell was biased with high voltages during modify only, while the core power supply level was sufficient for reading. Nowadays, as the core power supply scales down to 3 V (or 1.8 V), the VDD level is not effective to perform accurate reads, i.e. to distinguish whether a cell is programmed or erased: also high-level gate voltage is required to perform readings.

Because this voltage is produced on-chip by mean of a charge pump, it needs appropriate regulation.

Since reading the cells is the most frequent operation in a Flash memory, the circuits devoted to the read mode must have a reduced static current consumption: the gate voltage regulator has to be designed to minimize its load on the charge pump, thus reducing the power consumption from the core power supply (this requirement becomes more and more important for Flash devices targeted for portable equipment).

---

<sup>1</sup> The charge pump output voltage regulation is only an aspect of a general problem: in a Flash memory, stable voltage values less than VDD are required that have to be precise even if the core power supply varies. Also in this case voltage regulators are required, which are supplied by VDD instead of a charge pump output.

Furthermore, a reduced steady-state current allows the read charge pump to reduce its setup time at power-up.

A further requirement is fast settling time of the output voltage to its steady-state value  $V_{\text{READ}}$  as a consequence of the wordline (WL) addressing.

In fact, when a WL is addressed, the output voltage  $V_{\text{GROUT}}$  shows a drop with respect to  $V_{\text{READ}}$  because of the charge sharing phenomenon between the WL parasitic capacitance,  $C_{\text{WL}}$ , and the output capacitive load of the regulator,  $C_L$ , as schematically sketched in Fig. 15.26.

The resultant voltage drop  $\Delta V$  is very fast and can be excessive, i.e. it can move  $V_{\text{GROUT}}$  out of the limits that permit a proper read of the memory cell. To avoid incorrect read results, the output voltage recovery has to be fast with respect to the RC of the addressed WL, so as not to degrade the memory access time.

Finally, as the charge pump output voltage that supplies the regulator shows a ripple at its operative clock, it is necessary to take care of the positive Power Supply Rejection Ratio (PSRR), so as to minimize the amount of power supply noise injected into the output voltage.

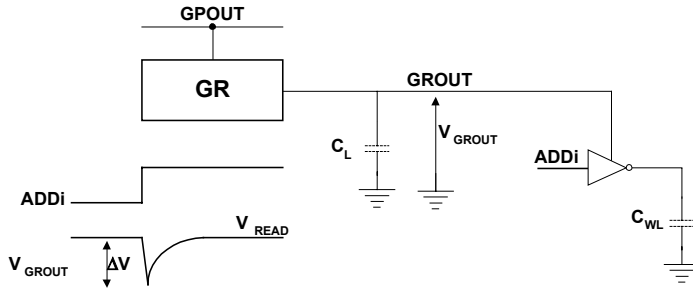


**Fig. 15.26.** Charge sharing between the regulator output and the addressed wordline

All of the above requirements taken into account, the most suitable topology for the read gate voltage regulator is the one depicted in Fig. 15.27. It is composed of an input differential amplifier and a single-ended PMOS output stage; a resistive divider ( $R_1 + R_2$ ) acts as a feedback net. The inverting input of the differential amplifier is driven by the reference voltage  $V_{\text{BG}}$ , generated by a band-gap circuit, while the non-inverting one is connected to the regulator output by the feedback net.

Assuming the circuit consists of ideal components (namely, no offset and sufficiently high loop gain), the regulator output voltage is:

$$V_{\text{GROUT}} = V_{\text{BG}} \left( 1 + \frac{R_1}{R_2} \right) \quad (15.8)$$

Since integrated resistors can be made with good matching, the precision of  $V_{\text{GROUT}}$  depends essentially on the precision of  $V_{\text{BG}}$  (refer to Chap. 5).

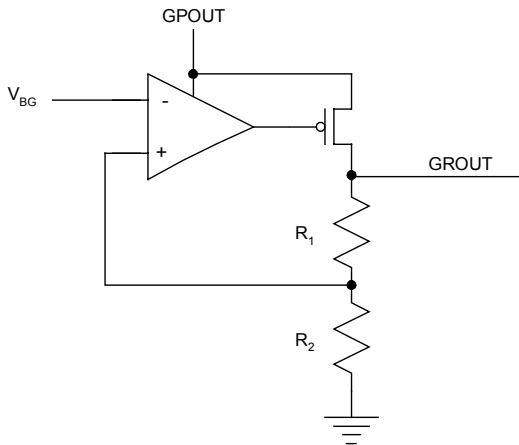


**Fig. 15.27.** Principle scheme of the regulator GR

The resistive feedback, which acts as a pull-down for the second gain stage, allows the regulator dc current to be kept low if  $R_1$  and  $R_2$  are large; in other words, the output stage steady-state current is:

$$I_{out,DC} = \frac{V_{GROUT}}{R_1 + R_2} \quad (15.9)$$

If the aspect ratio of the output p-channel transistor is kept high, the choice of a resistive feedback makes the recovery from undershoot easier. On the other hand, the recovery from an overshoot can be performed by mean of the RC network composed of the series of  $R_1$  and  $R_2$ , and the load capacitance.

The output load of the regulator GR is a huge parasitic capacitance  $C_L$  (hundreds of pF), since  $V_{GROUT}$  biases the final stages of the row decoders, i.e. the circuits that charge the WLs to the read voltage. The value of the output load increases with the size of the memory device (more row decoders), making the design of a fast regulator difficult.

To perform the read operation, the addressed WL is connected to the regulator output through the row decoder, while the other WLs are kept at GND: in such a way, the WL parasitic capacitance is connected in parallel to the regulator output load,  $C_L$ . As the charge sharing is almost instantaneous, the output voltage drop,  $\Delta V$ , is equal to:

$$\Delta V = \left( \frac{C_{WL}}{C_L + C_{WL}} \right) \cdot V_{GROUT} \cong \frac{C_{WL}}{C_L} V_{GROUT} \quad (15.10)$$

Let us make an example to determine the values: supposing that the target device is a 64Mbit Flash memory, typical values are:

$$C_L = 200 \text{ pF}, C_{WL} = 4 \text{ pF}, V_{GR} = 6 \text{ V}, V_{BG} = 1.3 \text{ V}, I_{out,dc} = 50 \text{ }\mu\text{A}$$

The resistance values to be chosen to satisfy the current requirements are:

$$R_1 \sim 90 \text{ k}\Omega \quad R_2 \sim 30 \text{ k}\Omega$$

and the voltage drop is:

$$\Delta V = 120 \text{ mV}$$

The circuit scheme of GR is shown in Fig. 15.28, where it is not difficult to find the blocks we talked about in the past few pages.

The n-channel transistors M1 and M2 act as a differential input stage, whose current mirror is represented by the PMOS transistors M5 and M6. M3 and M4 are inserted both to disable the regulator (when the regulator enable signal GREN is at “0”), and to decouple the internal nodes of the regulator from  $V_{BG}$  when GREN = “1”, acting as cascode-connected transistors (they have to be biased in saturation region).

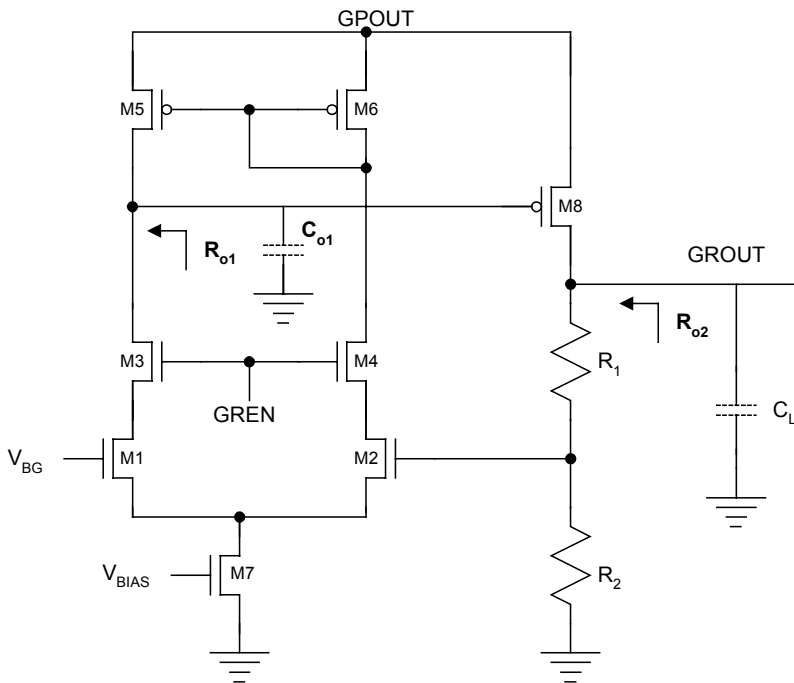


**Fig. 15.28.** The voltage regulator GR with its parasitic capacitances,  $C_{o1}$  and  $C_L$ .

Without M3, the  $C_{gdM1}$  parasitic capacitance would couple  $V_{BG}$  with the drain of M5, i.e. the differential stage output. Since this node is a “moving” one (it is subject not only to the differential output voltage swing, but also to the charge pump ripple), M3 prevents the noise from being transferred to the bandgap output.

The differential stage current generator is represented by M7, biased in saturation region by mean of suitable gate voltage  $V_{BIAS}$  that produces the steady-state differential stage current:

$$I_{first,DC} = I_{DM7} = \frac{\mu_N C_{OX}}{2} \left( \frac{W}{L} \right)_{M7} (V_{BIAS} - V_{thM7})^2 \quad (15.11)$$

where  $\mu_N$ ,  $C_{OX}$  e  $V_{thM7}$  are carriers mobility, gate oxide specific capacitance and threshold voltage of the NMOS M7, respectively.

The value of  $I_{first,DC}$  cannot be set too high because GR has to meet low power consumption requirements; this poses limitations to the dynamic behaviour of the regulator, especially with respect to slew rate performance.

The output gain stage is a p-channel transistor (M8 in Fig. 15.28), which is designed with a high aspect ratio in order to minimize the time penalty when charging the parasitic capacitance of the addressed WL.

### 15.8.2 Frequency Compensation

Operational amplifiers have usually multiple poles: for this reason they need frequency compensation, i.e. their open-loop transfer function has to be modified for the closed-loop circuit to be stable.

Referring to Fig. 15.29, where  $A = V_{out}/V_{in}$  is the open-loop gain and  $\beta$  the feedback factor (supposed to be frequency independent), the loop gain function  $G_{loop}$  of a negative feedback system is defined as:

$$G_{loop} = \frac{Y(s)}{X(s)} = \frac{A(s)}{1 + A(s)\beta}, s = j\omega \quad (15.12)$$

The poles locations of the closed-loop transfer function can be found by solving  $1 + A(s)\beta = 0$ .

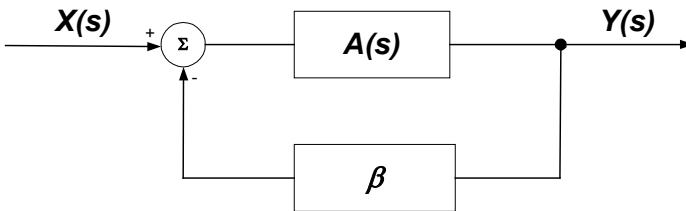


**Fig. 15.29.** Negative feedback schematic representation

The regulator GR needs both fast and highly stable response: so the  $G_{loop}$  function must have a high gain-bandwidth product (GB) and at least  $60^\circ$  phase margin.

The major obstacle to these objectives is the high value of the output parasitic capacitance  $C_L$ , as shown in the following.

Let us consider the regulator GR as shown in Fig. 15.28; its steady-state open-loop gain  $A_{dc}$  is:

$$A_{dc} = g_{mM1} R_{o1} g_{mM8} R_{o2} \quad (15.13)$$

$g_{mM1}$  e  $g_{mM8}$  are the transconductances of M1 and M8;  $R_{o1}$  e  $R_{o2}$  are the output resistances of the differential stage and of the PMOS output stage, respectively (the output resistances are schematically sketched in Fig. 15.28).  $R_{o2}$  is equal to the feedback network resistive series ( $R_1 + R_2$ ), in parallel with the output resistance of M8.  $R_{o1}$  is the parallel of M5 and M3 output resistances; since M3 is connected in common-gate configuration, it shows very high output resistance:

$$R_{cascode} = r_{dsM1} g_{mM3} r_{dsM3} \quad (15.14)$$

For this reason,  $R_{o1}$  is essentially equal to  $r_{dsM5}$ .

In order to evaluate the stability of GR, let us calculate the position of poles and zeroes.

In presence of huge output capacitive load and feedback resistance (as in our case), the dominant pole is the one associated with the regulator output:

$$|p_1| = \frac{1}{R_{o2} C_L} \quad (15.15)$$

In a first-order analysis, the regulator is a two-poles system and the resistance and parasitic capacitance of the first stage determine the second pole. The parasitic capacitance of the differential stage output is the sum of M3 and M5 drain junction capacitances (let it be called  $C_1$ ), and the gate-drain capacitance of M8. In any case, this latter capacitance is placed across a gain stage, thus exploiting the so-called Miller effect: a capacitance  $C$  placed between the input and the output of a gain stage, with gain  $A$ , can be represented as a capacitance  $AC$  connected between input and GND of said gain stage, and a capacitance  $C$  connected between its output and GND, as depicted in Fig. 15.30.

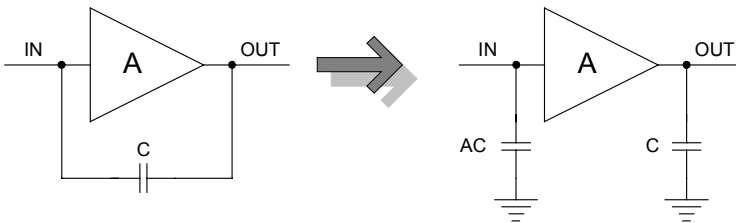


**Fig. 15.30.** Miller effect

For this reason, a capacitance  $C_{MILL} = C_{gdM8} R_{o2} g_{mM8}$  is placed in parallel to  $C_1$  (let us name the sum of them  $C_{o1}$ ), while the output load is increased by  $C_{gdM8}$  (anyway, since  $C_{gdM8} \ll C_L$ , the position of the dominant pole does not change).

The parasitic capacitance  $C_{o1}$  leads to a pole and a zero respectively at the frequencies:

$$|p_2| = \frac{1}{R_{o1}(C_1 + g_{mM8}R_{o2}C_{gdM8})} \cong \frac{1}{R_{o1}g_{mM8}R_{o2}C_{gdM8}} \quad (15.16)$$

$$z = \frac{g_{mM8}}{C_{gdM8}} \quad (15.17)$$

How poles and zero are placed with respect to each other depends on the parameters of the regulator. Usually, it is quite hard to design the regulator in order to have a good phase margin without a settling time penalty.

A possible solution to compensate a second pole placed in the bandwidth is to place a half-plane zero exactly at the frequency of the second pole: in such a way, the regulator acts ideally as a single-pole system (the dominant one). Anyway, this kind of compensation is usually not reliable, because a well-placed zero at a given operating condition may not “follow” the pole that it has to compensate when the operating condition varies (for example, as a consequence of temperature variations).

Another way to compensate is to use the Miller effect by placing a suitable compensating capacitor  $C_c$  across the second gain stage, as depicted in Fig. 15.31.

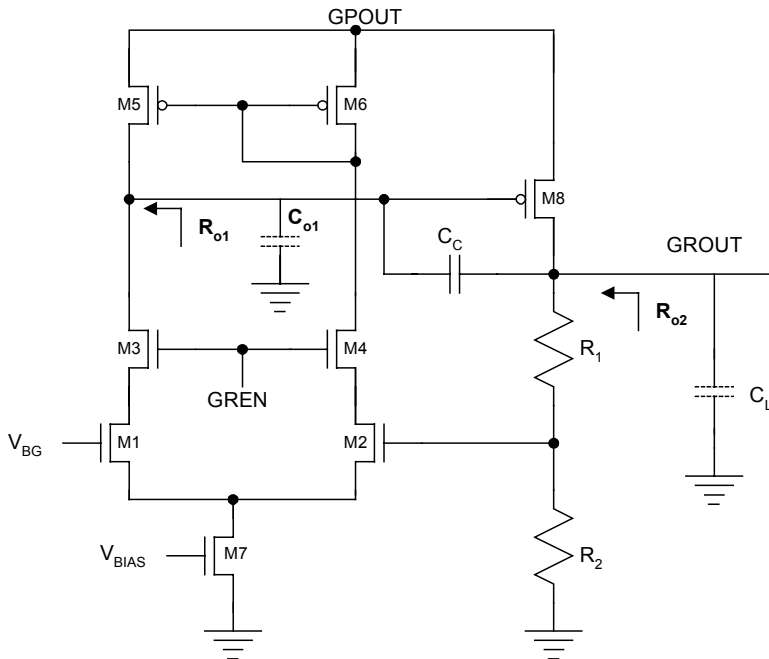


Fig. 15.31. GR with Miller compensation

Concerning the position of the poles,  $C_c$  acts in two ways: first of all, the capacitance  $C_{o1}$  is increased approximately by  $g_{mM8}R_{o2}C_c$ , thus reducing the frequency of the pole  $p_2$  related to the differential stage output. Furthermore, the pole  $p_1$  moves to higher frequency thanks to the negative feedback, which reduces the output stage resistance: this phenomenon is referred to as pole splitting.

By the effect of Miller compensation, the poles  $p_2$  e  $p_1$  move respectively to the frequencies:

$$\begin{aligned} |p_2| &\cong \frac{1}{C_{o1}R_{o1} + C_L R_{o2} + C_c(R_{o1} + R_{o2} + g_{mM8}R_{o1}R_{o2})} \cong \\ &\cong \frac{1}{g_{mM8}R_{o1}R_{o2}C_c} \end{aligned} \quad (15.18)$$

$$|p_1| \cong \frac{C_c g_{mM8}}{C_{o1}C_c + C_L C_c + C_{o1}C_L} \cong \frac{g_{mM8}}{C_L} \quad (15.19)$$

The approximations in Eqs. (15.18) and (15.19) are valid when  $C_L > C_c > C_{o1}$  and under the reasonable hypothesis that the second stage has an adequate gain.  $C_{o1}$  can be reduced by mean of accurate design and layout of the first stage. Depending on the value of  $C_c$ , it is possible to invert the position of the poles with respect to the scheme of Fig. 15.28, thus making the pole  $p_2$  of the first stage dominant.

Because of the unity-gain feedback between input and output of the second stage at high frequency,  $C_c$  leads also to a right-plane zero located at the frequency:

$$z_1 = \frac{g_{mM8}}{C_c} \quad (15.20)$$

The phase shift introduced by a right-plane zero is negative as the one of a left-plane pole: for this reason, this zero degrades the phase margin, and it may become a problem if it is placed near to the unity-gain frequency.

Remembering that the unity-gain frequency in a dominant-pole approximation is given by:

$$GB = A(0) \cdot |p_{dom}| = \frac{g_{mM1}}{C_c} \quad (15.21)$$

to obtain 60° phase margin, it is necessary that:

$$\pm 180^\circ - \tan^{-1}\left(\frac{\omega}{|p_1|}\right) - \tan^{-1}\left(\frac{\omega}{|p_2|}\right) - \tan^{-1}\left(\frac{\omega}{z}\right) = 60^\circ \quad (15.22)$$

To satisfy Eq. (15.22), it is necessary that  $z > 10GB$ ; in this case we obtain  $|p_2| > 2.2GB$ . As a result, the following relationships hold:

$$\frac{g_{mM8}}{C_c} > \frac{10 \cdot g_{mM1}}{C_c} \Rightarrow g_{mM8} > 10 \cdot g_{mM1} \quad (15.23)$$

$$\frac{g_{mM8}}{C_L} > \frac{2.2 \cdot g_{mM1}}{C_C} \Rightarrow C_C > 0.22 C_L \quad (15.24)$$

Referring to Eq. (15.23), the zero is far away from the unity gain frequency if the transconductance of the second stage is much higher than the one of the first stage. In a CMOS design, it is difficult to achieve large differences between  $g_{mM1}$  and  $g_{mM8}$ , since the transconductance increases with the square root of the current and the aspect ratio. If the zero is near the unity-gain frequency, it modifies significantly the frequency response near to the “zero dB crossing”, thus degrading the stability conditions.

Furthermore, as shown in Eq. (15.24), a huge output load  $C_L$  requires a big  $C_C$ , thus degrading the slew rate of the differential stage ( $SR = I_{\text{first,DC}}/C_C$ ), since its biasing current  $I_{\text{first,DC}}$  must meet low power consumption requirements. However, we cannot use a smaller compensation capacitance: if the phase margin is less than  $60^\circ$  (that is, contained between  $45^\circ$  and  $60^\circ$ ), the frequency response of the system leads to damped oscillations. As a consequence of how we designed GR, it is crucial that its output voltage recovers to the steady-state value without any overshoot which could lead to an excessive increase of the gate voltage of M8, thus cutting the output stage off. If this happens, the recovery to the regulated value  $V_{\text{READ}}$  can be performed only by discharging  $C_L$  through the resistive divider  $R_1 + R_2$ , but the time constant of this discharge is very high by design, because the output load  $C_L$  is huge, and the sum  $(R_1 + R_2)$  is great in order to reduce the static output current. Referring to the previous 64 Mbit Flash memory example, the discharge time constant of the output node is:

$$\tau_{\text{disch}} = (R_1 + R_2) * C_L = 120 \text{ k } \Omega * 200 \text{ pF} = 24 \text{ } \mu\text{s} \quad (15.25)$$

In practice, we cannot leave the circuit as it is, but we must find some remedy to the unsatisfactory phase margin.

The right-plane zero problem (and the related feedforward) can be solved by introducing a unity-gain buffer between GROUT and  $C_C$ , as depicted in Fig. 15.32.

By mean of the unity-gain buffer, the right plane zero is cancelled, while the position of the poles is kept unchanged with respect to Miller compensation.

**Problem 15.2:** Demonstrate it by mean of the small signals equivalent circuit depicted in Fig. 15.33.

If the voltage buffer has a finite output resistance,  $R_0$ , the zero is not cancelled perfectly, but a pole and a left-plane zero replace it.

With suitable modifications to the small signal equivalent circuit, one can demonstrate that the new pole and zero are located respectively at:

$$|p_3| \cong \frac{1}{R_0 C_{o1}} \quad (15.26)$$

$$z_1 \cong -\frac{1}{R_0 C_C} \quad (15.27)$$

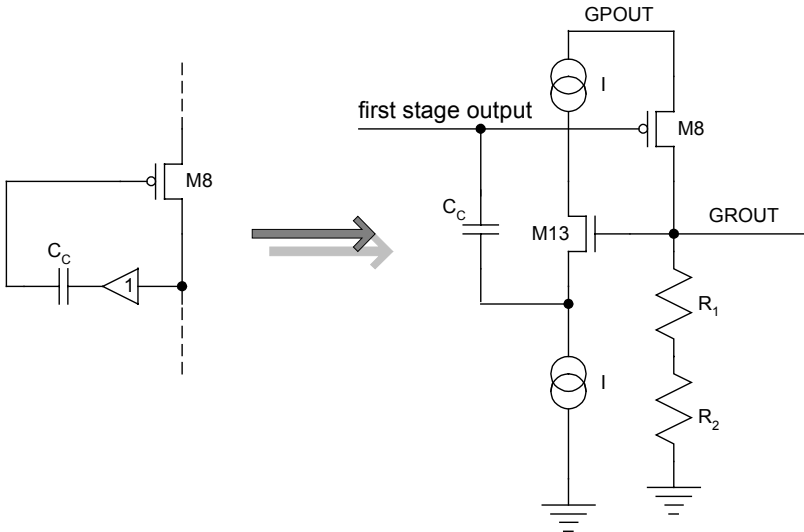


**Fig. 15.32.** Unity-gain buffer frequency compensation (Follower Feedback Compensation, FFC)

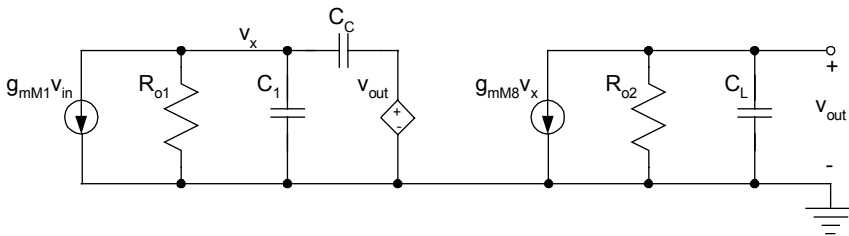


**Fig. 15.33.** Small signal equivalent circuit of GR with FFC

To obtain a unity-gain buffer, it is convenient to use a n-channel transistor in source follower configuration (Follower Feedback Compensation, FFC), as M13 in Fig. 15.32. The follower has to be biased with an appropriate dc current generator  $I$ ; for the follower to be working properly,  $V_{\text{GROUT}}$  must never be less than the sum of the voltages required to bias in saturation region both M13 and the current generator connected towards GND. This poses limitations to the use of this kind of compensation if  $V_{\text{GROUT}}$  may assume low levels (for example  $<2\text{V}$ ), as it will be explained in Sect. 15.8.4.

Instead of using a voltage buffer, we can use a current buffer in series with the compensating capacitor. Since the easiest way to obtain a current buffer is to use a transistor biased in common-gate configuration (cascode), as shown in Fig. 15.34 we can name this kind of compensation Cascode Feedback Compensation (CFC).

$V_{BCASC}$  is chosen to keep M23 in saturation, i.e. to make it work as a cascode. The current generators can be designed as in Fig. 15.34; for this purpose, another biasing voltage  $V_{BIAS}$  is shown, not necessarily different from  $V_{BCASC}$ .

In steady state, the compensating capacitor  $C_C$  is charged to  $(V_{GROUT} - V_{sM23})$  and the current generators provide the current  $I$ , which flows in M23. When  $C_C$  has to be charged as a consequence of an output voltage drop, the additional current  $\Delta I$  is provided by the differential stage and a current  $I + \Delta I$  flows in M23. The current  $\Delta I$  keeps the compensating capacitor charged in such a way that the source terminal of M23 acts as a virtual ground.

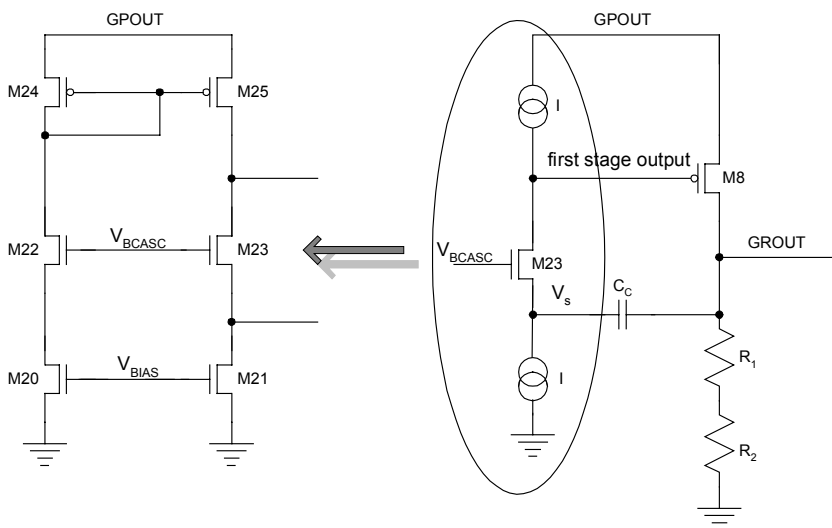


**Fig. 15.34.** Cascode Feedback Compensation

It is not trivial to determine the transfer function of the system; anyway, starting from the small signal equivalent circuit, with some algebra we obtain the open-loop dominant pole:

$$|p_2| \cong \frac{1}{g_{mM8} R_{o1} R_{o2} C_C} \quad (15.28)$$

Note that the expression of the dominant pole is the same as Eq. (15.25); however, since  $R_{o1}$  is smaller than in Miller compensation, the dominant pole appears at higher frequency; in fact,  $R_{o1}$  now is:

$$R_{o1} = r_{dsM3} // r_{dsM5} // r_{dsM23} // r_{dsM25} \cong r_{dsM5} // r_{dsM25} \quad (15.29)$$

$g_{mM23}$  is the transconductance of M23,  $r_{dsM25}$  e  $r_{dsM23}$  the channel conductances of M25 and M23, respectively. Since M23 is biased in common-gate configuration, its channel conductance is very high and can be neglected.

The second pole frequency is:

$$|p_1| \cong \frac{g_{mM8} C_C}{C_{o1}(C_C + C_L)} \quad (15.30)$$

As in FFC, the system shows a third pole and a left-plane zero, located respectively at:

$$|p_3| \cong \frac{g_{mM23} C_L}{C_C^2} \quad (15.31)$$

$$z_1 = -\frac{g_{mM23}}{C_C} \quad (15.32)$$

Usually (i.e. when  $C_L$  is the biggest capacitance in the system), the feedback cascode compensation allows the second pole high frequency shifting to be  $\sim C_C/C_{o1}$  times more effective than in Miller or feedback follower compensation.

Assuming that  $C_{o1} \ll C_C$  or  $C_{o1} \ll C_L$ , the unity gain frequency is still given by  $g_{mM1}/C_C$ ; to be sure that the second pole is well beside the unity gain bandwidth, in FFC and CFC we have to satisfy, respectively:

$$\frac{g_{mM1}}{g_{mM8}} < \frac{C_C}{C_L} \quad (15.33)$$

$$\frac{g_{mM1}}{g_{mM8}} < \frac{C_C^2}{C_{o1}(C_C + C_L)} \quad (15.34)$$

The reduced power budget limits  $g_{mM1}$  and  $g_{mM8}$ , even if the latter is chosen to provide the regulator with the required current capability. To ensure an adequate phase margin, for each set of  $g_{mM1}/g_{mM8}$ ,  $C_L$ , and  $C_{o1}$ , the compensating  $C_C$  has to be chosen to satisfy either Eq. (15.33) or Eq. (15.34). Anyway, since both FFC and CFC allow more effective compensation with respect to conventional Miller technique, a smaller  $C_C$  can be used, thus achieving better SR performances for any given value of  $I_{first,DC}$ .

Furthermore, even if the reduced value of  $g_{mM1}$  limits the feedback loop gain-bandwidth product, an additional feedback loop is created from GROUT to the gate of M8: through the follower M13 and  $C_C$  in case of FFC, through  $C_C$  and the cascode M23 in case of CFC. By mean of these new loops, an output voltage drop  $\Delta V$  is reflected directly in a fast decrease in M8 gate voltage, which in turn leads to a fast increase in the output current. Nonetheless, the FFC additional loop is not as fast as the CFC one. When compensating with the feedback follower technique, the parasitic capacitance seen at the differential stage output is the sum of  $C_{o1}$  and  $C_C$ , while it is just  $C_{o1}$  when CFC is used. Referring to the 64Mbit example,  $C_{o1}$  is on the order of pF, while  $C_C$  is about ten pF: in case of cascode feedback compensation, the additional feedback loop will be faster.

### 15.8.3 Positive Power Supply Rejection Ratio (PSRR)

Good power supply rejection ratio performance is mandatory in a regulated system, especially when the regulator is supplied with a voltage produced by a charge pump instead of VDD. A characteristic of a good positive PSRR is when the noise on the output of the charge pump is not transferred to the regulated voltage; in our case we define the positive PSRR as:

$$PSRR^+ = \frac{(V_{GROUT} / V_{in})_{V_{GPOUT}=0}}{(V_{GROUT} / V_{GPOUT})_{V_{in}=0}} \quad (15.35)$$

The PSRR performance depends on the compensation strategy used in the system; the exhaustive analysis of the positive PSRR can be performed only if we calculate the PSRR transfer function by mean of the small signals equivalent circuits. Nonetheless, a few qualitative considerations can be deduced directly from the schematics.

From the gate of M8 (the differential stage output), the charge pump noise can be transferred through  $C_{gdM8}$ , which is usually small and can therefore transfer only high-frequency noise.

If we adopt the conventional Miller compensation, the concepts remain the same:  $C_c (>> C_{gdM8})$  couples the positive supply directly with the regulator output, however it may happen at lower frequencies, depending on the  $C_L$  value (remember Eq. (15.24)). Also lower frequency noise can be transferred to the output and for this reason, this type of compensation should be avoided.

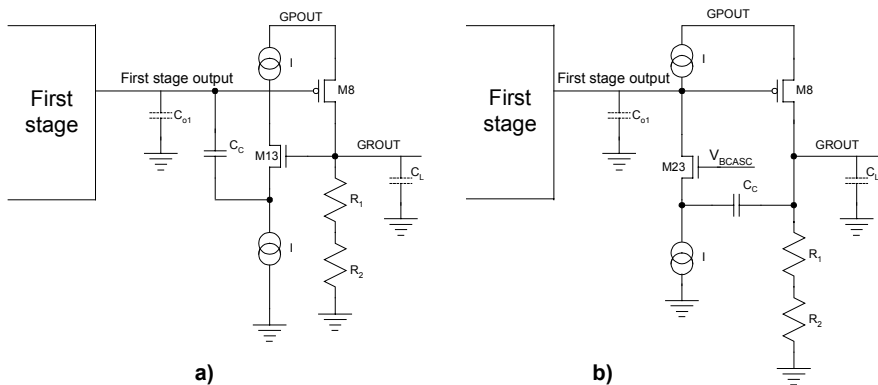


Fig. 15.35. FCC (a) and CFC (b) positive PSRR

When compensating with FFC or CFC, the direct M8 gate-to-drain connection via  $C_c$  is eliminated (of course, the connection via  $C_{gdM8}$  is still present, but typically it acts well above the charge pump noise frequency). The noise transmission mechanism is therefore different, and depends on the biasing of M8 (Fig. 15.35).

As a general rule, the differential stage output filters any signal higher than its cut-off frequency, i.e. the gate of M8 cannot “follow” a noise beyond this frequency. In this case, the noise signal  $V_{sgM8}$  is approximately equal to the noise on GPOUT ( $V_{sM8}$ ), thus modulating the biasing of M8: for this reason, the drain terminal of M8, that is GROUT, varies as well, since it is forced to source a static current due to its biasing by  $R_1 + R_2$ .

As discussed in the previous section, since the first stage output has a heavier capacitive load in FFC than in CFC, its cut-off happens at a lower frequency, thus transferring the charge pump noise to the output (if the charge pump operating frequency is located in the pass band). Therefore, the larger CFC noise pass band allows better positive PSRR performances.

### 15.8.4 Program Gate Voltage

In order to perform effective and precise program operations, the best method for biasing the WLs is with a linear “staircase” voltage; in such a way, if the drain terminal is kept to a suitable and constant voltage value  $V_{DP}$ , after an initial transient a 1:1 relation exists between gate voltage increment,  $\Delta V_{GP}$ , and voltage shift after each program pulse,  $\Delta V_T$ .

This method requires a small amount of current from the drain side (in single supply devices the drain program voltage is produced by a charge pump, as will be explained in the next section), since the cells are not biased with an excessive overdrive, and this permits an increase in the amount of program parallelism.

To save silicon area and power consumption, it is convenient to use the same regulator GR that produces the read voltage to provide the gate staircase program voltage as well; furthermore, this approach assures that in a Program & Verify scheme (refer to Chap. 14) the verify and the read voltages match. Ultimately, the gate voltage regulator should act as an Analog-to-Digital converter.

To obtain “programmable” gate voltage values, we can divide the feedback resistive partition into a series of smaller resistances, each one having a switch in parallel.

When adopting this solution, two major problems have to be taken into account: the first concerns the linearity of the output voltage as a function of the selection command. Figure 15.36 shows an example of the ideal staircase voltage (dashed line) with respect to the real one (continuous line). The differential linearity error  $\varepsilon_d$  is given by:

$$\varepsilon_d = \frac{|V_{ID}(n) - V_{RE}(n)|}{\Delta V_{GP}} \quad (15.36)$$

where  $V_{ID}(i)$  is the ideal voltage at the  $i$ -th step,  $V_{RE}(i)$  the real one  $\Delta V_{GP}$  the program step.

The second problem concerns the charge injection related to the commutation of the switches, which leads to voltage overshoot on the feedback node. In order to have a fast settling time on the program voltage, these overshoots have to be reduced as much as possible.

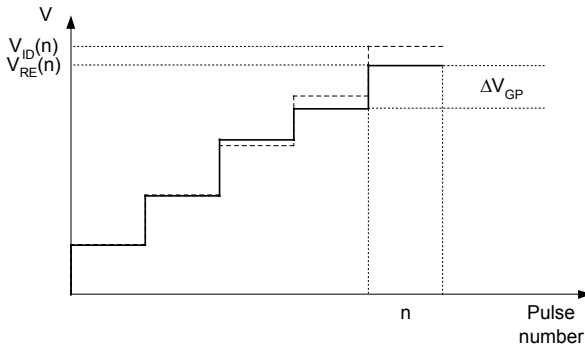


**Fig. 15.36.** Linearity error

To satisfy linearity requirements, a good solution consists in a “programmable”  $R_2$  ( $R_2$  is the resistor connected between node E and GND), composed of a parallel combination of  $n$  binary weighted resistors, a certain number of which are active at the same time as a function of the program voltage, as sketched in Fig. 15.37a. In this case, this solution is not feasible because the resistor divider has to assure low power consumption: the total value of the resistances is unacceptable. Furthermore, the output stage current variation, due to  $R_2$ , leads to an offset voltage in the regulator (since the differential stage current does not change).

One can choose the solution in Fig. 15.37b, where  $R_1$  is made programmable by mean of a series of binary weighted resistors, each one with a selection switch in parallel; in this case, the number of switches simultaneously open depends on the selected program voltage level, giving rise to serious linearity problems (as much as  $\pm 15\%$  of the voltage step). Furthermore, when many switches commute simultaneously, the node E is subject to high charge injection, especially if the switches are designed with high aspect ratio to minimize their parasitic resistance. The best solution from the linearity point of view consists in dividing the resistor  $R_1$  in a series of  $w$  equal resistors of value  $\Delta R$ , while  $R_2$  is kept constant (Fig. 15.37c).

The voltage step  $\Delta V_{GP}$  is achieved at each step by increasing the value of  $R_1$  by a fixed amount  $\Delta R$ ; at the  $p$ -th program step the gate program voltage  $V_{GP}$  is equal to:

$$V_{GP} = V_{BG} \left( 1 + \frac{R_1^* + p\Delta R}{R_2} \right) \quad (15.37)$$

where  $R_1^*$  is the value that  $R_1$  assumes at the first program step. The program step voltage  $\Delta V_{GP}$  is:

$$\Delta V_{GP} = V_{BG} \frac{\Delta R}{R_2} \quad (15.38)$$

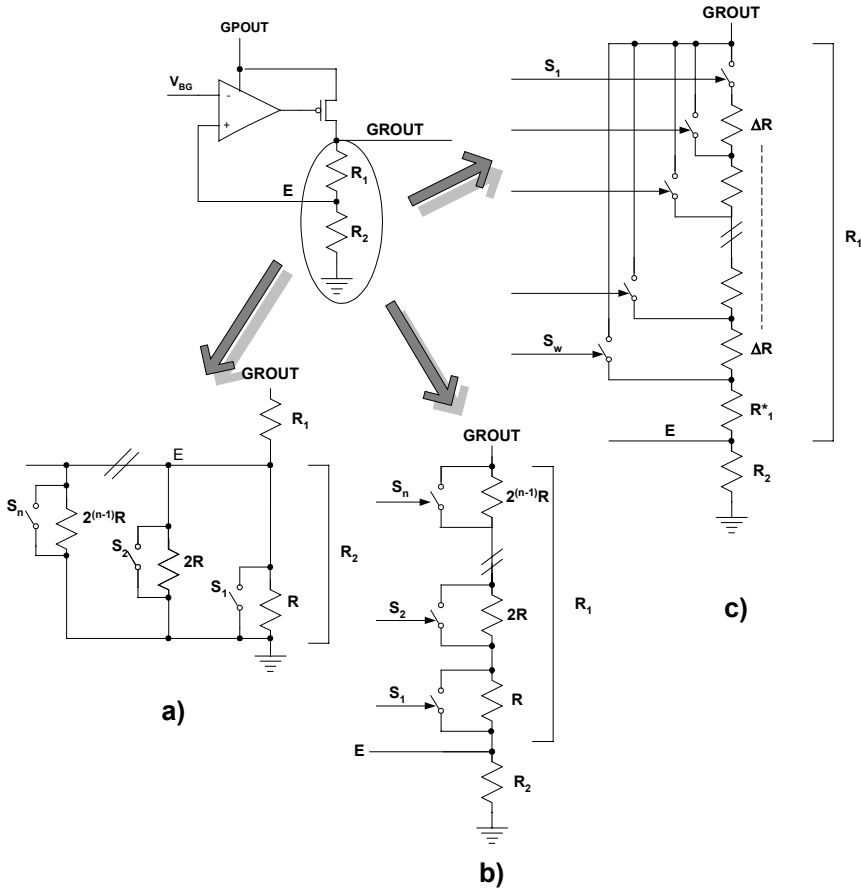


**Fig. 15.37.** Three different ways to make the feedback net programmable

At each  $p$ -th program pulse, a suitable value of resistor  $\Delta R$  is shorted by means of the control logic that produces the  $w$  signals to drive the switches from  $S_1$  to  $S_w$ , thus achieving  $R_1 = R^*_1 + p\Delta R$ . To avoid introducing parasitic resistances that may vary from one step to the next, only one switch is active during each program pulse, while the others are off.

Usually the gate voltage ramp starts at  $\sim 1.5$  V and continues until 9.5 V; the voltage step  $\Delta V_{GP}$  depends on the program accuracy.

After each program pulse the cells are verified, i.e. they are read to check whether the desired threshold voltage has been reached or not. During verify, the gates of the cells are biased with the verify voltage  $V_{PV}$  (usually the verify gate voltage is equal to the read voltage,  $V_{PV} = V_{GR}$ ).

To verify after each program pulse means that a highly capacitive node, i.e. the output node GROUT, must be driven to different voltage levels; as depicted in Fig. 15.38, two different situations may occur, depending on whether  $V_{GP} < V_{PV}$  or  $V_{GP} > V_{PV}$ . In the former case, at the end of the  $k$ -th program pulse  $C_L$  has to be charged to  $V_{PV}$ , while at the beginning of the  $(k+1)$ -th pulse  $C_L$  has to be discharged to  $V_{GP}$ ; in the latter case at the end of the  $k$ -th program pulse  $C_L$  has to be discharged to  $V_{PV}$ , while at the beginning of the  $(k+1)$ -th pulse  $C_L$  has to be charged to  $V_{GP}$ .

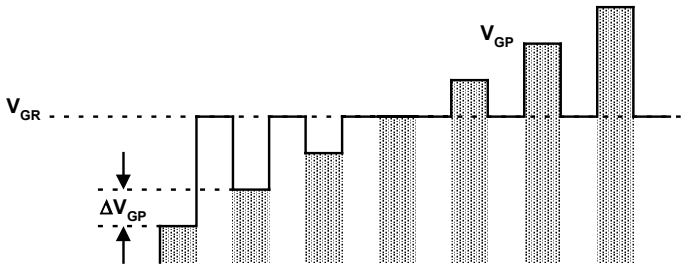


**Fig. 15.38.** Gate voltage qualitative behaviour during Program & Verify algorithm

As explained in the previous section, it is not difficult to charge  $C_L$  to a higher voltage level if the pull-up transistor M8 is designed with a high aspect ratio. On the contrary, the discharge operation is slow because the resistive feedback net limits the pull-down current. To get a faster discharge, the topology shown in Fig. 15.39 can be used.

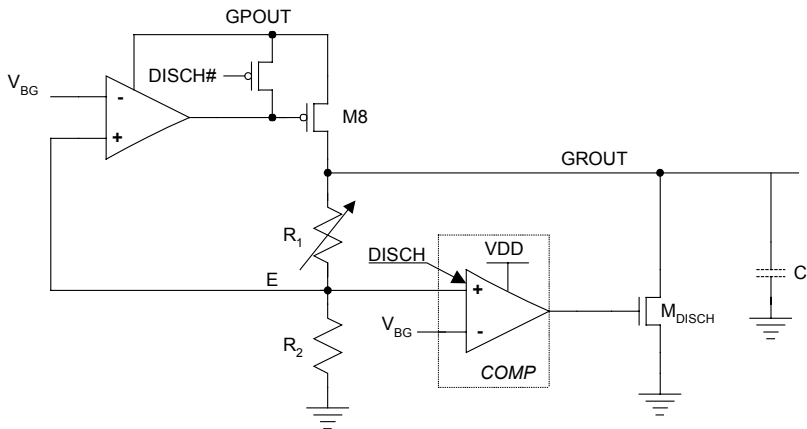


**Fig. 15.39.** Program gate voltage regulator with discharge circuit

When the regulator output has to be discharged from the actual value  $V_{\text{PRES}}$  to  $V_{\text{NEXT}} < V_{\text{PRES}}$ , the comparator COMP (that can be designed as a combination of a differential stage and an inverter) is enabled by a suitable signal DISCH, while the pull-up transistor M8 is disabled by DISCH# not to short GPOUT to GND. Meanwhile, the resistor  $R_1$  is programmed to obtain  $V_{\text{NEXT}}$ . In such a way, since the new value of  $R_1$  is less than the previous one and the feedback loop of the regulator is interrupted, the voltage  $V_E$  at node E increases, thus turning  $M_{\text{DISCH}}$  on and enabling the discharge path from GROUT to GND. This path remains active until  $V_E > V_{\text{BG}}$ , i.e. until  $V_{\text{GROUT}} > V_{\text{NEXT}}$ . As DISCH is switched to the low logic level and M8 is turned on, the next output voltage level  $V_{\text{NEXT}}$  can be reached quickly. This solution allows the reduction of the current requirements of the charge pump GP, thus achieving a faster settling time and reducing the total programming time.

## 15.9 Drain Voltage Regulation and Temperature Dependence

Now we will explore the problems related to drain voltage regulation, which is a crucial parameter affecting the programming operation. Figure 15.40 shows the program high voltages scheme in a dual supply Flash memory.

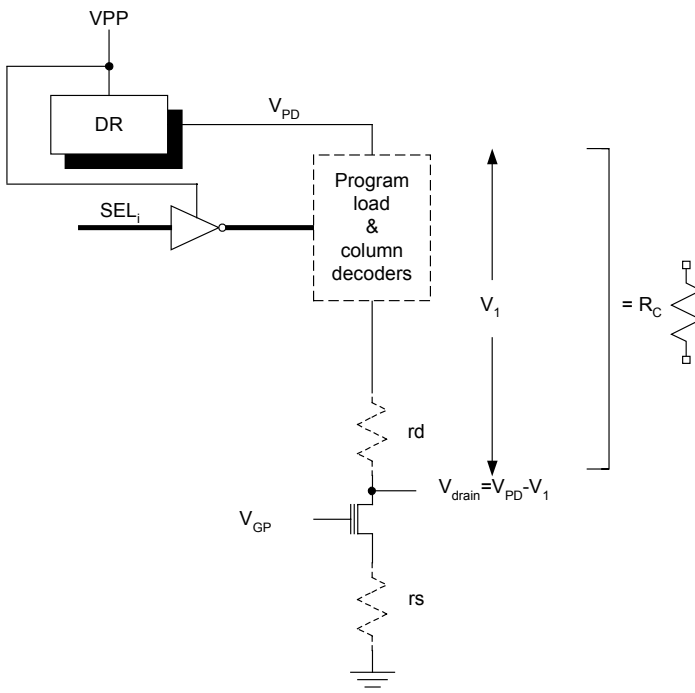


Fig. 15.40. Drain voltage regulation ( $V_{PD}$ ) in a dual supply Flash memory

The drain voltage is obtained by mean of a regulator, DR, that regulates the voltage  $V_{PD}$  before the program load (refer to Sect. 16.5) and the column decoders. In such a way, the voltage on the drain is not  $V_{PD}$ , but  $V_{PD}-V_1$ , where  $V_1$  is the voltage drop across the bit line and pass transistors parasitic resistance. In order to limit this voltage drop, the column decoder pass transistors must be properly biased so they operate in linear region.

Figure 15.41 shows an example of a program load regulator, implemented in dual power supply Flash memory devices, which compensates for the voltage drop across the column decoder.

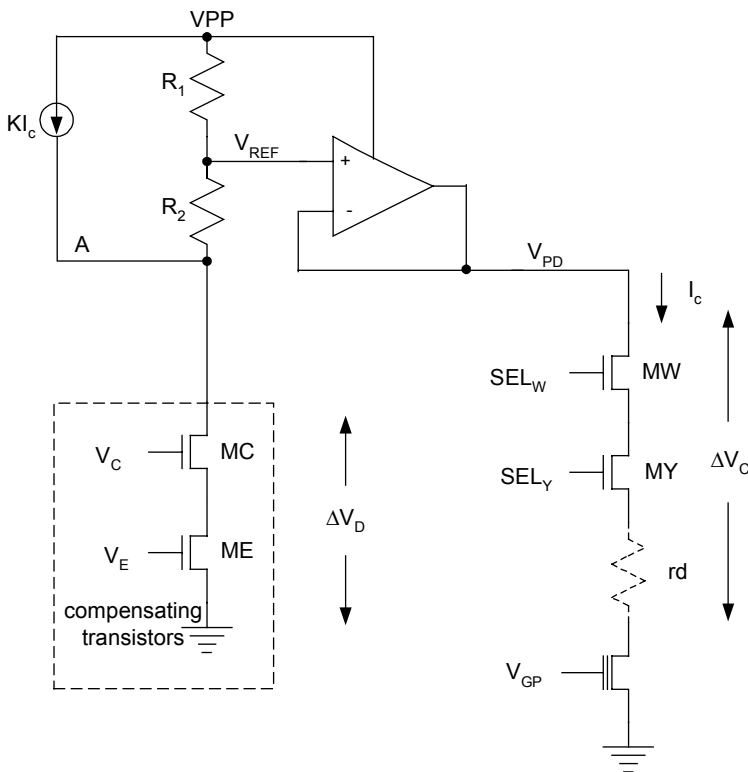


**Fig. 15.41.** Adaptive  $V_{PD}$  regulation

A suitable voltage  $V_{PD} \sim V_{REF}$  is applied to the load by mean of an operational amplifier connected in a unity-gain non-inverting configuration.  $V_{REF}$  is obtained with a resistive divider made up by the resistors  $R_1$  and  $R_2$ , and two series-connected MOS transistors operating in the triode region (MC and ME). These are matched with the bitline select transistors MW and MY, respectively, with a suitable scaling factor, and are driven by the same control signals. Since MC and ME do not perform any selection, they will be referred to as compensating transistors.

A current-controlled current source injects a current  $KI_c$  into the node A, so as to provide an additional contribution  $\Delta V_D$  to  $V_{REF}$  with respect to its quiescent value at  $I_c=0$ ,  $V_{R0}$ .

$$V_{REF} = V_{R0} + \Delta V_D \quad (15.39)$$

The scaling factor  $K$  is chosen so as to make  $\Delta V_D$  equal to the drop  $\Delta V_C = I_c R_{on,s}$  across the bit line select devices ( $R_{on,s} = R_{on,MW} + R_{on,MY}$ ). Thanks to the adaptive feedback loop, the voltage  $V_{PD}$  on the selected bitline is ideally equal to  $V_{R0}$  during the whole program operation, regardless of the variation of the fabrication process or the programming current. This architecture is also suitable for parallel programming, for example on byte or word basis. In this case, the current fed to the load is equal to the sum of the currents drawn by the  $n$  cells being programmed,  $nI_c$ . Each cell is addressed through its corresponding pair of bitline select transistors. For correct compensation of the voltage drop across these devices, the current  $KnI_c$  must be delivered to  $n$  identical branches  $MC_i$ ,  $ME_i$  placed in parallel, as shown in Fig. 15.42. The enable devices  $ME_i$  activate the correct number of compensation branches, according to the data to be stored.

It is not easy to extend this solution to single supply voltage devices, especially when high program parallelism is required. In these devices the drain program voltage is provide by an on-chip charge pump, DP, which should substain not only the worst case total drain program current (and the current consumption of the regulator), but also its replica, multiplied by the scaling factor  $K$ .

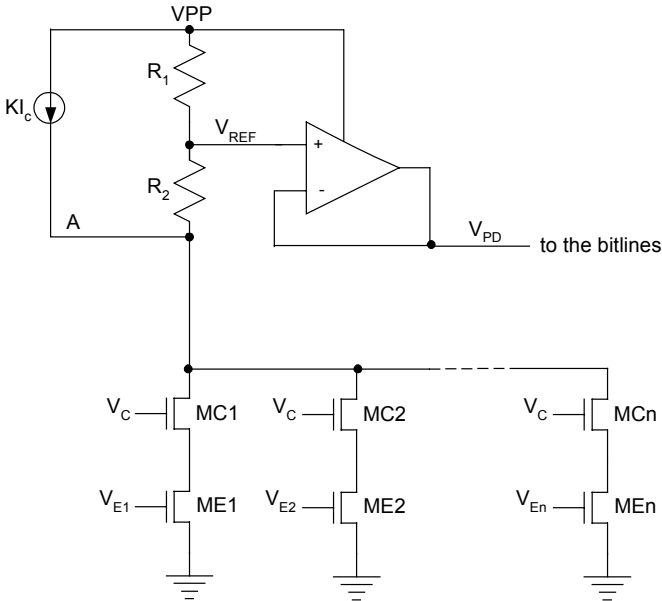


**Fig. 15.42.** Adaptive program load regulator for parallel programming

In single supply Flash devices we forego the  $\Delta V_c$  drop compensation, and we design a regulation scheme like the one shown in Fig. 15.43, where  $V_{PD}$  is obtained from a band-gap reference. The column decoder parasitic resistance,  $R_c$ , increases with temperature, thus decreasing the program voltage applied to the drain (assuming the program current is constant). For example, if the program current is  $400\ \mu\text{A}$  per cell, the voltage drop  $\Delta V_c$  may vary from  $200\ \text{mV}$  at  $-40^\circ\text{C}$  to  $350\ \text{mV}$  at  $120^\circ\text{C}$ . It should be clear that the efficiency of the program circuits, which must force a sufficiently high drain voltage, is reduced as the temperature increases.

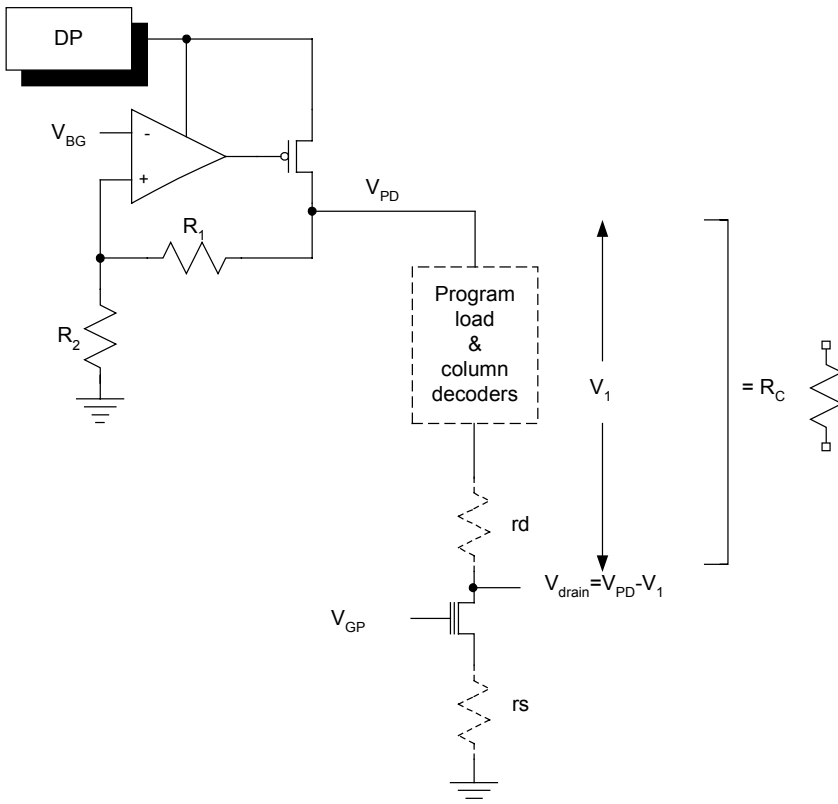


**Fig. 15.43.**  $V_{PD}$  regulation in single supply Flash devices (the voltage drop across the column decoder pass transistors is not compensated)

We can modify the circuit in Fig. 15.43 by substituting the input voltage on the inverting input of the differential amplifier with a suitable voltage, whose mean value is equal to the band-gap voltage  $V_{BG}$ , and linearly increases with temperature. Such a voltage can be produced by the circuit represented in Fig. 15.44. Three operational amplifiers make up the thermal tracker: the first reproduces  $V_{BG}$ ,

while the last one performs the difference between this voltage and a suitable amplification  $V_x$  of the emitter-base voltage of a pnp transistor.

Since the forward-biased diode voltage shows a linear decreasing of  $-2\text{mV}/^\circ\text{C}$ ,  $V_{OUT}$  in Fig. 15.44 increases linearly with temperature.

The following relationships hold:

$$V_{OUT} = V_{BG} + (V_{BG} - V_{BE}) \frac{R_B}{R_A} - \frac{R_B}{R_A} \frac{R_1}{R_2} V_{BE} \quad (15.40)$$

$$\frac{\delta V_{OUT}}{\delta T} = - \left( 1 + \frac{R_1}{R_2} \right) \frac{R_B}{R_A} \frac{\delta V_{BE}}{\delta T} \quad (15.41)$$

If  $R_1 = R_2$ , from Eq. (15.40) we obtain:

$$V_{OUT} = 2V_{BG} - 2V_{EB} \frac{R_B}{R_A} \quad (15.42)$$

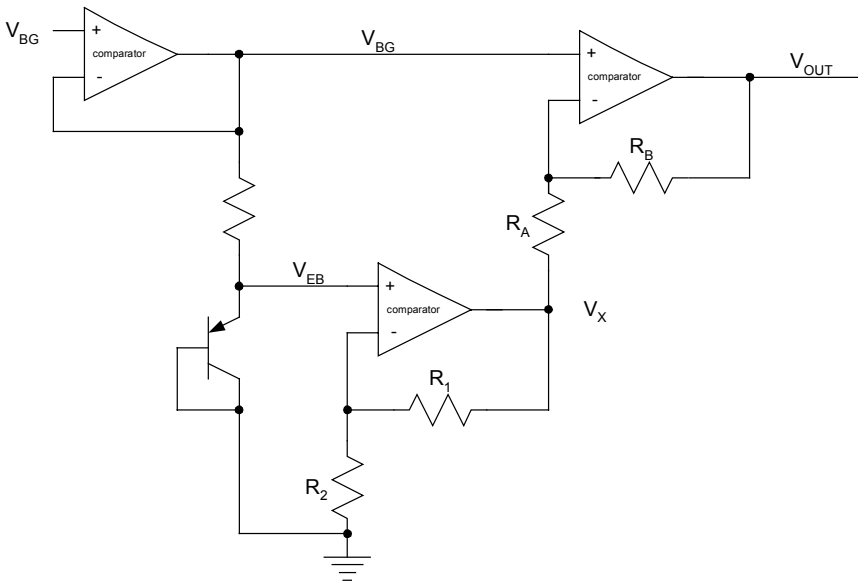


**Fig. 15.44.** Thermal tracker

Furthermore, Eq. (15.41) can be rewritten as:

$$\frac{\delta V_{OUT}}{\delta T} = -2 \frac{R_B}{R_A} \frac{\delta V_{EB}}{\delta T} \quad (15.43)$$

i.e. the thermal derivative of  $V_{OUT}$  depends on  $R_B/R_A$  ratio.

To summarize,  $V_{OUT}$  can be applied to each regulator that uses  $V_{BG}$  without any other structural modification. Substituting  $V_{BG}$  by  $V_{OUT}$  in Fig. 15.43, we notice that  $V_{PD} = K \cdot V_{OUT}$  increases with temperature, as  $V_{OUT}$  increases. Since the program voltage applied on the cell drain is:

$$V_{drain} = V_{PD} - \Delta V_C \quad (15.44)$$

it is possible to compensate for the increase in  $\Delta V_C$  by an offset in  $V_{PD}$ , so as to assure the thermal tracking of the drain voltage.

## Bibliography

- B.K. Ahuja, An Improved Frequency Compensation Technique for CMOS Operational Amplifiers, *Journal of Solid-State Circuits*, vol. SC-18, pp. 629-633, (Nov. 1983).
- P.E. Allen, D.R. Holberg, *CMOS Analog Circuit Design*, Holt, Rinehart and Winston
- C. S. Bill el al., High Voltage charge pumps with series capacitances, U.S. patent No. 5,059,815, (Oct. 22, 1991).
- G. Di Cataldo and G.Palumbo, "Double and triple charge pump for power IC dynamic models which take parasitic effects into account", *IEEE Trans. Circuits Syst.*, vol. CAS-40, pp. 92-101, (Feb. 1993).
- 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, pp. 374-378, (Jun. 1976).
- P. Favrat, P. Deval, and M. J. Declercq, "A high-efficiency CMOS voltage doubler", *IEEE J. Solid State Circuits*, vol. SC-33, pp. 410-416, (Mar. 1998).
- A. Ghilardelli, G. Campardo, J. Mulatti, "Bidirectional charge pump generating either a positive or negative voltage", USA patent No. 6,184,741, (Feb. 6, 2001).
- P.R. Gray, P.J. Hurst, S.H. Lewis, and R.G. Meyer, *Analysis and Design of Analog Integrated Circuits*. New York, NY: Kluwer John Wiley & Sons, Inc., ch. 4, (2001).
- S.S. Haddadi et al., Flash E<sup>2</sup>PROM Array with Negative Gate Voltage Erase Operation, U.S. patent n. 5,077,691, (Oct. 23, 1989).
- T. Kawahama et al., Bit line clamped sensing Multiplex and accurate high voltage generator for quarter micron Flash memory, *Journal of Solid State Circuit*, Vol 31, No 11, p. 1590, (Nov. 96).
- T. Kawahara, T. Kobayashi, Y. Jyouno, S. Sacki, N. Miyamoto, T. Adachi, M. Kato, A. Sato, J. Yugami, H. Kume, and K. Kimura, "Bit-line clamped sensing multiplex and accurate high-voltage generator for 0.25 $\mu$ m flash memories", in 1996 IEEE Int. Solid-State Circuits Conf. Dig. Tech. Pap., pp. 38-39, (Feb. 1996).
- O. Khouri, S. Gregori, R. Micheloni, D. Soltesz, and G. Torelli, "Low output resistance charge pump for Flash memory programming", 2001 IEEE Proc. Int. Workshop on Memory Technology, Design and Testing, San Jose, CA (USA), pp. 99-104, (Aug. 2001).
- O. Khouri, S. Gregori, A. Cabrini, R. Micheloni, and G. Torelli, "Improved charge pump for Flash Memory applications in triple-well CMOS technology", in 2002 IEEE Proc. Int. Symposium on Industrial Electronics, L'Aquila (Italy), pp. 1322-1326, (Jul. 2002).

- 
- O. Khouri, R. Micheloni, and G. Torelli, "Very fast recovery word-line voltage regulator for multilevel non-volatile memories", in Proc. Third IMACS/IEEE Int. Multiconference Circuits, Communications and Computers, Athens, Greece, pp. 3781–3784, (Jun. 1999).
- O. Khouri, R. Micheloni, I. Motta, A. Sacco, G. Torelli, "Capacitive boosting circuit for the regulation of the word line reading voltage in non-volatile memories", U.S. Patent No. 6.259.635, (Jul. 10, 2001).
- O. Khouri, R. Micheloni, I. Motta, A. Sacco, G. Torelli, "Capacitive compensation circuit for the regulation of the wordline reading voltage in non-volatile memories", U.S. Patent No. 6.259.632, (Jul. 10, 2001).
- O. Khouri, R. Micheloni, I. Motta, and G. Torelli, "Voltage regulating circuit for a capacitive load", U.S. Patent No. 6.249.112, (Jun. 19, 2001).
- O. Khouri, R. Micheloni, A. Sacco, G. Campardo, and G. Torelli, "Program word-line voltage generator for multilevel Flash memories", in Proc. 7th IEEE Int. Conf. on Electronics, Circuits, and Systems, vol. II, pp. 1030–1033, (Dec. 2000).
- O. Khouri, R. Micheloni, S. Gregori, and G. Torelli, "Fast Voltage Regulator for Multilevel Flash Memories", in Records 2000 IEEE Int. Workshop on Memory Technology, Design and Testing, pp. 34–38, (Aug. 2000).
- M. Maccarrone et al, "Program load adaptive voltage regulator for Flash memories", Journal of Solid State Circuit, Vol. 32, No 1, p. 100, (Jan. 1997). F. Maloberti, Analog Design for CMOS VLSI Systems, 2001 Kluwer Academic Publishers, Boston
- M. Mihara, Y. Terada, and M. Yamada, "Negative heap pump for low voltage operation flash memory", in 1996 Symposium VLSI Circuits Dig. Tech. Pap., pp. 76–77, (Jun. 1996).
- D.B. Ribner, M.A. Copeland, Design Techniques for Cascode CMOS Op Amps with Improved PSRR and Common Mode Input Range, IEEE Journal of Solid-State Circuits, vol. SC-19, N. 6, pp. 919-925, (Dec. 1984)
- 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, pp. 36–44, (Jan. 1998).
- T. Tanzawa and T. Tanaka, "A dynamic analysis of the Dickson charge pump circuit", IEEE J. Solid-State Circuits, vol. SC-32, no. 8, pp.1231-1240, (Aug. 1997).
- J.S. Witters, G. Groeseneken, and H. Maes, "Analysis and modeling of on-chip high-voltage generator circuits for use in EEPROM circuits", IEEE J. Solid-State Circuits, vol. SC-24, pp. 1372–1380, (Oct. 1989).
- C.C Wang and J. Wu, "Efficiency improvement in charge pump circuits", IEEE J. Solid-State Circuits, vol. SC-32, pp. 852–860, (Jun. 1997).
- R.J. Widlar, "New developments in IC voltage regulators", IEEE J. Solid-State Circuits, vol. SC-20, pp. 816–818, (Feb. 1971).
- J.T. Wu and K.L. Chang, "MOS charge pumps for low-voltage operation", IEEE J. Solid-State Circuits, vol. 33, pp.592-597, (Apr. 1998).
- M. Zhang, N. Llaaser, and F. Devos, "Improved voltage tripler structure with symmetrical stacking charge pump", El. Letters, vol. 37, pp. 668–669, (May 2001).