

Nonvolatile Semiconductor Memory Technology

*A Comprehensive Guide
to Understanding and Using
NVSM Devices*

Edited by

William D. Brown
University of Arkansas

Joe E. Brewer
Northrop Grumman Corporation



IEEE
PRESS

IEEE Press Series on Microelectronic Systems
Stu Tewksbury, *Series Editor*
IEEE Solid-State Circuits Council, *Sponsor*
IEEE Components, Packaging, and Manufacturing
Technology Society, *Sponsor*

The Institute of Electrical and Electronics Engineers, Inc., New York

A NOTE TO THE READER

This book has been electronically reproduced from digital information stored at John Wiley & Sons, Inc. We are pleased that the use of this new technology will enable us to keep works of enduring scholarly value in print as long as there is a reasonable demand for them. The content of this book is identical to previous printings.

This book and other books may be purchased at a discount from the publisher when ordered in bulk quantities. Contact:

IEEE Press Marketing
Attn: Special Sales
445 Hoes Lane, P. O. Box 1331
Piscataway, NJ 08855-1331
Fax: (732) 981-9334

For more information on the IEEE Press,
visit the IEEE home page: <http://www.ieee.org/>

© 1998 by the Institute of Electrical and Electronics Engineers, Inc.,
3 Park Avenue, New York, NY 10016-5997

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Sections 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-1470. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Printed in the United States of America
10 9 8 7 6 5 4 3

ISBN 0-7803-1173-6
IEEE Order Number: PC5644

Library of Congress Cataloging-in-Publication Data

Nonvolatile semiconductor memory technology : a comprehensive guide to understanding and using NVSM devices / edited by W. D. Brown, Joe E. Brewer.

p. cm. -- (IEEE Press series on microelectronic systems)

Includes bibliographical references and index.

ISBN 0-7803-1173-6 (cloth)

1. Semiconductor storage devices. I. Brown, W. D. (William D.).

(date) . II. Brewer, Joe (Joe E.) III. Series.

TK7895.M4N634 1997

621.39'732-dc21

97-19691

CIP

List of Contents
List of Acronyms
Foreword

Chapter 1

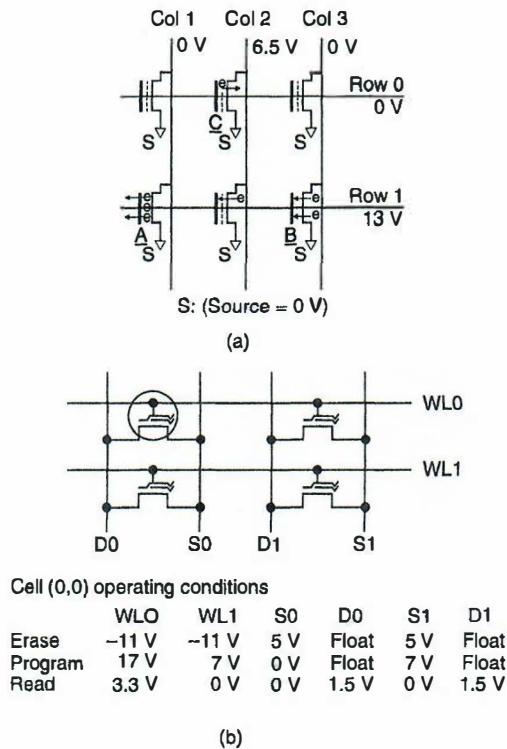


Figure 2.17 (a) Schematic description of the disturb mechanism in single-transistor arrays using the half-voltage scheme: Cell A experiences DC erase disturb, Cell B experiences DC program-disturb, and Cell C experiences program-disturb [2.17], and (b) matrix implementation of the single-transistor memory cell [2.18].

tions are not needed. (A graded junction is required for a high breakdown voltage, but its drawback is that the cell area cannot be scaled easily.)

2.2.1 Charge-Pump Circuits

The recent development of 5V-only memories requires that the programming voltage V_{pp} be generated using on-chip charge pumps. The basic voltage multiplier circuit is shown in Fig. 2.18 [2.20]. The nodes of the diode chain are coupled to the inputs via capacitors in parallel so that the capacitors have to withstand the full voltages developed across them. Efficient multiplication can be achieved with relatively high values of stray capacitance, and the current drive capability is independent of the number of multiplier stages. The multiplier operates by pumping charge packets along the diode chains as the coupling capacitors are successively charged and discharged each half-clock cycle. The voltages are not reset after each pumping cycle, so that the average node potentials increase progressively from the input to the output of the diode chain. The circuit implementation is shown in Fig. 2.19, where MOS transistors are used in the diode configuration. The clocks are generated by two inverters driven from an oscillator circuit. Typically, the output is limited by MOS diodes to prevent output voltages from exceeding process limits, such as junc-

tion b
range.

T
based
contro
time =
output
high-v
dence
increa
closed
the ch
loop
capaci
forces
the re
that is

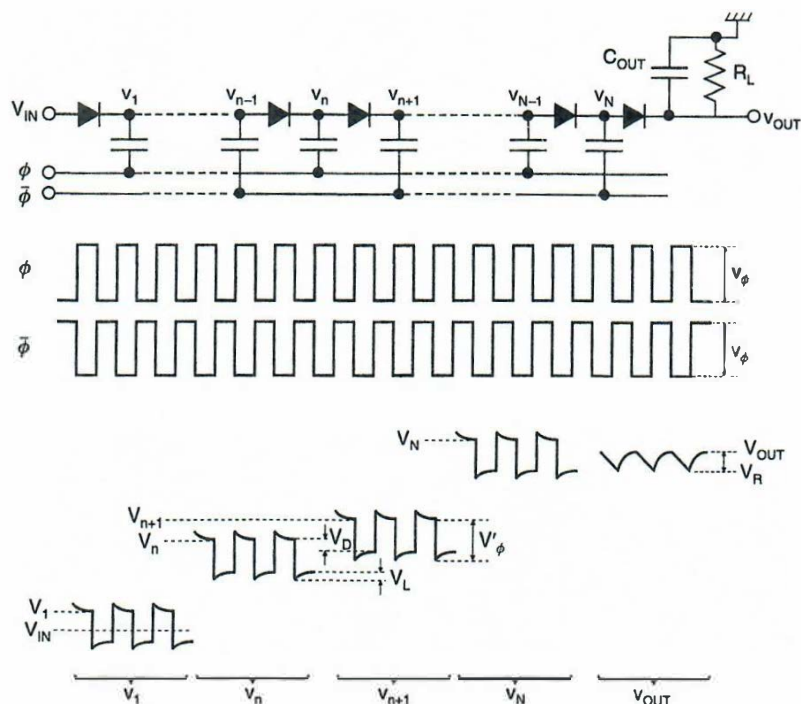


Figure 2.18 Voltage waveforms in an N-stage multiplier showing the voltage relationship between successive nodes of the diode chain [2.20].

tion breakdown voltages, if the supply voltage V_{dd} increases above the specified range.

To reduce the peak fields in the tunnel oxide of cells using electron tunneling-based write/erase operations, the rise time of the programming pulse needs to be well controlled. For an exponential ramp, $RC = 600 \mu\text{sec}$, and for a linear ramp, the rise time $= 40 \mu\text{s}$ [2.16]. This can be achieved by utilizing a charge pump whose current output is low enough to attain the desired rise time using the capacitance on the high-voltage node. However, this does not compensate for the temperature dependence of junction leakage current. As the temperature increases, leakage currents increase, and the charge pump may fail to achieve the desired output voltage. A closed-loop implementation, shown in Fig. 2.20, uses a feedback network to regulate the charge-pump output and control the programming pulse [2.21]. The feedback loop consists of a clock driver (whose output swing is controllable), a diode-capacitor chain, a voltage divider, and a differential amplifier. The negative feedback forces the output voltage to be equal to the voltage applied to node A multiplied by the reciprocal of the divider ratio. The voltage at A is generated by an RC network that is allowed to charge to V_{ref} . By making the time constant of the RC network

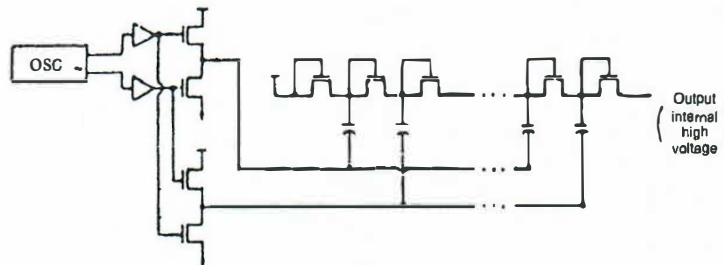


Figure 2.19 Typical circuit implementation in NMOS technology [2.21].

equal to 600 μsec and setting V_{ref} to the optimal programming voltage multiplied by the divider ratio, the desired programming pulse is produced. The feedback loop allows the basic charge pump to be designed to work in the worst case corner condition (high temperature, low V_{cc}) without resulting in overstress in the best case corner (low temperature and high V_{cc}). A nonlocking high-voltage regulation and output shaping circuit is given in Fig. 2.21 [2.22].

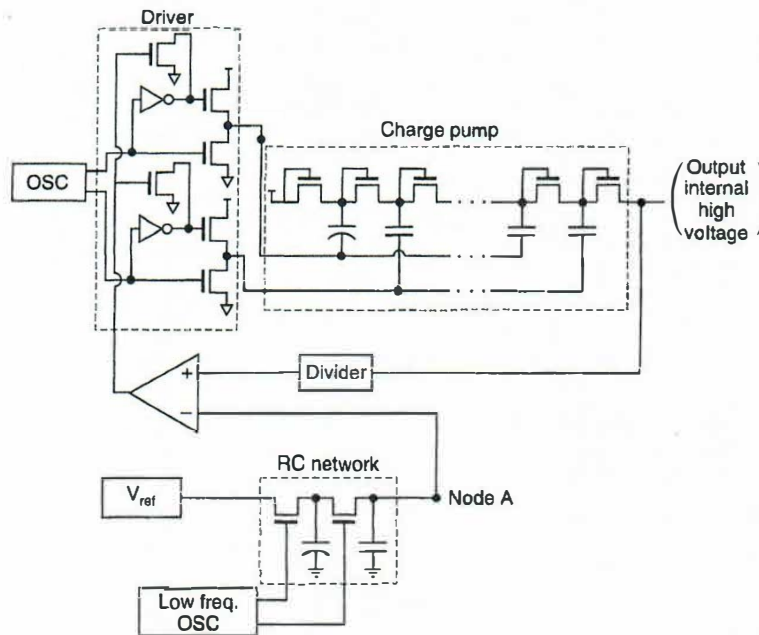


Figure 2.20 Simplified schematic of a high-voltage regulation circuit [2.21].

A
requir
bootstr
several
changin
ber of
large, v
input c
L_g
charge
directio
NMO₅
[2.24].
well in
tive vo
charge:
held of
nected
CD2 c
QD1 is
buffer
When
charge
ment n
the fal

A quick boosting charge-pump circuit has been proposed for use in memories requiring fast programming times [2.23]. Rather than sequentially charging each bootstrap capacitor, as in the above circuits, the basic concept here is to charge several capacitors in parallel and then connect them in series, which is done by changing the number of stages and capacitance used for charge pumping. The number of stages is controlled to be small, while the capacitance for charge pumping is large, which increases the charge-transfer efficiency (ratio of the average output to input current).

Large negative voltages can be generated (from a 5 V supply) using negative charge pumps. Operating principles are similar to positive charge pumps, except the direction of the diode chain (see Fig. 2.18) is reversed. Typical implementation, in NMOS technology, used for negative substrate bias generation is shown in Fig. 2.22 [2.24]. Node V_x is kept below ground using bootstrap techniques. Deep p-well and n-well implants would be needed to isolate the n^+ junctions connected to large negative voltages. When the oscillator output is high, the charge-pump capacitor CD2 charges to $V_{dd} - V_{t,QD2}$. The enhancement pulldown transistor connected to CD1 is held off by the output of the internal stage of the inverting push-pull buffer connected to CD1. When the oscillator output becomes low, the charge-pump capacitor CD2 couples node B below ground, to $-(V_{dd} - V_t)$. So, the source node of device QD1 is not at ground, but is coupled to $-(V_{dd} - V_t)$. The output of the push-pull buffer now goes high. The voltage across capacitor CD1 charges to $2(V_{dd} - V_t)$. When the oscillator output returns to the high level, the charging circuit for the charge-pump capacitor CD2 is isolated from the rest of the generator by enhancement mode diode QD1, and the voltage at node A is coupled by capacitor CD1 and the falling output voltage of the push-pull buffer to approximately $-2(V_{dd} - V_t)$.

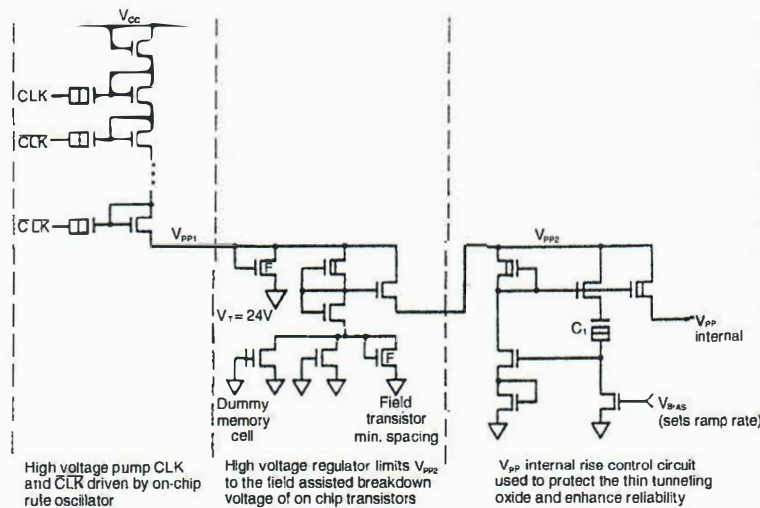


Figure 2.21 High-voltage (V_{pp}) generator for 5V-only operation [2.22].

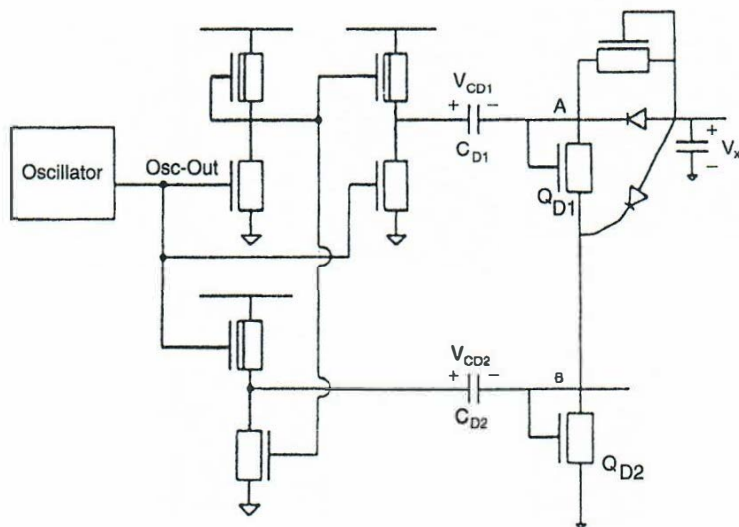


Figure 2.22 NMOS negative charge pump circuit [2.32].

In CMOS implementation, negative charge pumps are typically implemented with PMOS transistors, as shown in Fig. 2.23 [2.18, 2.25, 2.32]. The charge-pump circuit has six cascaded voltage multiplier stages, and each stage consists of one PMOS transistor, one resistor, and two capacitors. The capacitor is formed with a gate capacitor of a PMOS depletion transistor. P7 is a depletion-type PMOS and is used to provide the ground level at V_{ERA} , when the negative pump is not used. With node A initially at ground, both capacitors connected to node A will be charged to $\approx V_{dd}$, that is, 5 V. When PA1 goes low, because node A is isolated, it tracks PA1 and will initially go to about -5 V. This turns P1 on, causing node B to be discharged. The final node A and node B voltages will depend on the capacitor values and the PMOS threshold voltage because the charge on the two capacitors connected to PA1 and PA2 is shared between the capacitors connected to node A and node B. This charge sharing reduces the gate voltage on P1, making it more resistive. Next, PA2 becomes low, P1 is turned on hard since the gate node of P1 becomes lower than node A due to the RC network between node A and PA2, and the node B voltage is lowered further. Each stage discharges the previous stage, and the output node V_{ERA} becomes -13 V.

2.2.2 High-Voltage Load Circuits

Since the amount of current that can be generated by an on-chip charge pump is limited, it is essential that the programming currents be minimized. The memory transistor and decoding circuitry are the main parts of the chip that must be connected to the charge pump. During write, electrons tunnel to the floating gate, and

the load
During a
pump (v
tance an
the arra
which is
Unfortu
the nece
devices c
circuit u

Pric
the trans
onto the
at a thre
of thresl
limits its
the perij
drawn fi
Thi
routing

4.8.3 Charge-Pump Circuits

Charge-pump circuits have been discussed in previous chapters. The concept of a quick boosting charge-pump circuit for high-density and low-voltage Flash memories [4.102] is shown in Fig. 4.114.

For the conventional charge-pump circuit, the number of stages connected in series between the power supply and the output of the charge-pump circuit is fixed. This fixed number of stages is designed so as to generate a certain high voltage required for programming. However, the charge-transfer efficiency, I_{PP}/I_{CC} , for the charge-pump circuit is given by $1/(n+1)$, where I_{CC} and I_{PP} are the mean input and output currents for the charge-pump circuit, respectively, and n is the number of stages connected in series between the power supply and the output of the charge-pump circuit. As a result, many stages connected in series are so redundant that the conventional charge-pump circuit not only consumes too much power, but also lowers the charge-transfer efficiency, while the boosted voltage is not much higher than the power supply voltage at the beginning of the operation. Thus, it takes a longer time for the conventional charge-pump circuit to generate a desired voltage for programming when a higher density EEPROM increases the load capacitance for the charge-pump circuit and the power supply voltage is lowered.

The charge-pump scheme conceptually presented in Fig. 4.114 can change the number of stages and the capacitance used for charge pumping. While the output voltage is not much higher than the power supply voltage in the beginning, the number of stages is controlled to be small and the capacitance for charge pumping is controlled to be large in order to increase the charge-transfer efficiency. Accordingly, as the output voltage is boosted, the number of stages is increased and the capacitance for charge pumping is decreased step by step. As a result, the charge-pump scheme shown in Fig. 4.114 offers a high charge-transfer efficiency and reduces the rising time of the output voltage.

The switches SW1 and SW2 are controlled by a set of complementary voltages of a high and low voltage converted by a level shifter circuit (LS).

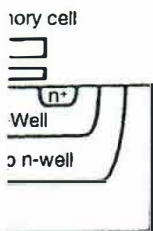
4.9. FLASH APPLICATIONS

In this section, the current and potential applications of Flash memory are discussed and a brief discussion on DRAM versus Flash is presented.

4.9.1 EPROM Replacement

EPROM was important in the product development phase but was replaced by Mask ROM as the product went into volume production. On the other hand, with Flash, product designers and end-users can continue to update the product more easily and in situ. Another advantage of Flash over EPROM is a smaller footprint, thus saving board space. In embedded microcontroller applications, Flash is replacing EPROM, as well as EEPROM, Boot RAM, and battery RAM. Low-voltage

gram of negative



mes 5 V. Then,
at NMOS tran-
sistors at 5 V. At the
signals (BERm)
therefore, only the
the other hand,
me 5 V and Z -
ected row-main
sector erase and
the row decoder.

1, compared to
the cell type. For
are connected
operation and
to switch V_{pp} .
X-decoders or