

Inside NAND Flash Memories



Michelsoni · Crippa · Marelli

Rino Micheloni
Luca Crippa
Alessia Marelli

Inside NAND Flash Memories

 Springer

Inside NAND Flash Memories

Rino Micheloni • Luca Crippa • Alessia Marelli

Inside NAND Flash Memories

 Springer

Rino Micheloni
Integrated Device Technology
Agrate Brianza
Italy
rino.micheloni@ieee.org

Luca Crippa
Forward Insights
North York
Canada
luca.crippa@ieee.org

Alessia Marelli
Integrated Device Technology
Agrate Brianza
Italy
alessiamarelli@gmail.com

ISBN 978-94-007-9834-2 ISBN 978-90-481-9431-5 (eBook)
DOI 10.1007/978-90-481-9431-5
Springer Dordrecht Heidelberg London New York

© Springer Science+Business Media B.V. 2010
Softcover re-print of the Hardcover 1st edition 2010
No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Cover design: eStudio Calamar S.L.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

2 NAND overview: from memory to systems

R. Micheloni¹, A. Marelli² and S. Commodaro³

2.1 Introduction

It was in 1965, just after the invention of the bipolar transistor by W. Shockley, W. Brattain and J. Bardeen, that Gordon Moore, co-founder of Intel, observed that the number of transistors per square centimeter in a microchip doubled every year. Moore thought that such trend would have proven true for the years to come as well, and indeed in the following years the density of active components in an integrated circuit kept on doubling every 18 months. For example, in the 18 months that elapsed between the Pentium processor 1.3 and the Pentium-4, the number of transistors grew from 28 to 55 million.

Today, a standard desktop PC has processors whose operating frequency is in the order of some gigahertz, while its memory can store as much information as terabytes.

In this scenario, a meaningful portion of the devices produced is represented by memories, one of the key components of any electronic systems.

Semiconductor memories can be divided into two major categories: RAM, acronym for *Random Access Memories*, and ROM, acronym for *Read Only Memories*: RAM loses its content when power supply is switched off, while ROM virtually holds it forever. A third category lies in between, i.e. NVM, acronym for *Non-Volatile Memories*, whose content can be electrically altered but it is also preserved when power supply is switched off. These are more flexible than the original ROM, whose content is defined during manufacturing and cannot be changed by the consumer anymore.

The history of non-volatile memories began in the 1970s, with the introduction of the first EPROM memory (*Erasable Programmable Read Only Memory*). Since then, non-volatile memories have always been considered one of the most important families of semiconductors and, up to the 1990s, their interest was tied up more to their role as a product of development for new technologies than to their economic value. Since the early 1990s, with the introduction of non-volatile Flash memories into portable products like mobile phones, palmtop, camcorders, digital cameras and so on, the market of these memories has experienced a stunning increase.

¹ Integrated Device Technology, rino.micheloni@icce.org

² Integrated Device Technology, alessiamarelli@gmail.com

³ Pegasus MicroDesign, stefano_commodaro@yahoo.it

The most popular Flash memory cell is based on the *Floating Gate* (FG) technology, whose cross section is shown in Fig. 2.1. A MOS transistor is built with two overlapping gates rather than a single one: the first one is completely surrounded by oxide, while the second one is contacted to form the gate terminal. The isolated gate constitutes an excellent “trap” for electrons, which guarantees charge retention for years. The operations performed to inject and remove electrons from the isolated gate are called program and erase, respectively. These operations modify the threshold voltage V_{TH} of the memory cell, which is a special type of MOS transistor. Applying a fixed voltage to cell’s terminals, it is then possible to discriminate two storage levels: when the gate voltage is higher than the cell’s V_{TH} , the cell is on (“1”), otherwise it is off (“0”).

It is worth mentioning that, due to floating gate scalability reasons, charge trap memories are gaining more and more attention and they are described in Chap. 5, together with their 3D evolution.

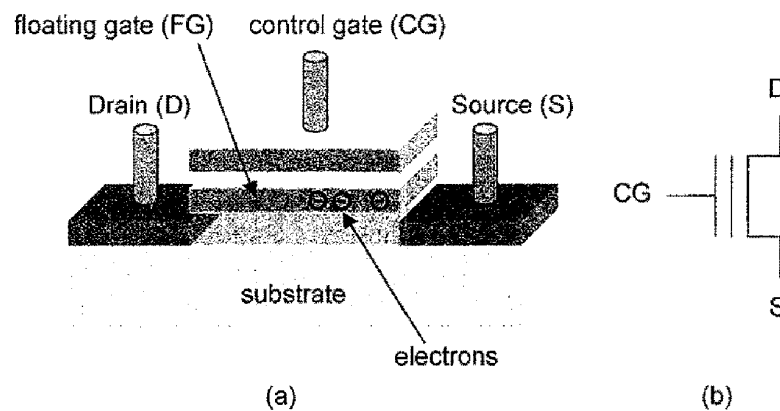


Fig. 2.1. (a) Floating gate memory cell and (b) its schematic symbol

2.2 NAND memory

2.2.1 Array

The memory cells are packed to form a matrix in order to optimize silicon area occupation. Depending on how the cells are organized in the matrix, it is possible to distinguish between NAND and NOR Flash memories. This book is about NAND memories as they are the most widespread in the storage systems. NOR architecture is described in great details in [1].

In the NAND string, the cells are connected in series, in groups of 32 or 64, as shown in Fig. 2.2. Two selection transistors are placed at the edges of the string, to ensure the connections to the source line (through M_{SL}) and to the bitline (through M_{DL}). Each NAND string shares the bitline contact with another string. Control gates are connected through wordlines (WLs).